Chapter 9
The Draft Genome of the MD-2 Pineapple

Raimi M. Redwan, Akzam Saidin, and Subbiah V. Kumar

Introduction

The main challenge in assembling plant genome is its ploidy level, repeats content, and polymorphism. The second-generation sequencing delivered the throughput and the accuracy that is crucial to whole-genome sequencing but insufficient and remained challenging for some plant species. It is known that genomes produced by next-generation sequencing produced small contigs that would inflate the number of annotated genes (Varshney et al. 2011) and missed on the transposable elements that are abundant in plant genome due to their repetitive nature (Michael and Jackson 2013).

In assembling plant genomes, many reported the unresolved part of the genome, that is, the heterochromatin region that was left unassembled in the final draft (Cheung et al. 2006; Tuskan et al. 2006; Ming et al. 2008; Wang et al. 2012a, b, 2014). This region is tightly packed in the centric and subtelomeric regions of the chromosome, and is highly repetitive, making the sequences difficult for sequencing and assembly (Hoskins et al. 2002). However, the complexity of the regions does not make the region any less important to be decoded as the regions also contained genes and important regulatory elements for euchromatic genes (He et al. 2012). The task to resolve the heterochromatic region in whole-genome sequencing

R. M. Redwan
Faculty of Agro-Based Industry, Universiti Malaysia Kelantan, Jeli, Kelantan, Malaysia

Biotechnology Research Institute, Universiti Malaysia Sabah,
Kota Kinabalu, Sabah, Malaysia

A. Saidin
Novocraft Technology Sdn. Bhd, Petaling Jaya, Selangor, Malaysia

S. V. Kumar (✉)
Biotechnology Research Institute, Universiti Malaysia Sabah,
Kota Kinabalu, Sabah, Malaysia
e-mail: vijay@ums.edu.my

© Springer Nature Switzerland AG 2018

R. Ming (ed.), *Genetics and Genomics of Pineapple*, Plant Genetics and
Genomics: Crops and Models 22, https://doi.org/10.1007/978-3-030-00614-3_9

project especially the one using shotgun strategies was only performed as a subsequent improvement of the genome draft using concise physical mapping for targeted transposons resequencing (Devine et al. 1997; Hoskins et al. 2002). This sort of information may not be available for non-model plants, and the intrinsic solution to improve the resolution of repetitive reads of the heterochromatic region is longer reads that can span through the elements.

The use of long reads from the third-generation sequencing is not directly useful neither to the feasibility of complete de novo whole-genome sequencing. High accuracy reads of 99.99% of PacBio reads can only be achieved as consensus reads, for the random errors to be resolved by consensus calling. At single pass, PacBio reads contain high error rate, and due to this independent use of the reads requires error correction. This is because errors in reads will cause failure for the assembler to establish overlap-layout path between reads in order to merge them. Error correction can be performed either by using the PacBio reads itself or by adopting the high accuracy reads from the second-generation sequencing. Self-correction module of PacBio reads required redundant coverage of at least 50 of the targeted genome to generate an accurate consensus (Chin et al. 2013) and for pineapple whole-genome sequencing which has estimated genome size of 526 Mb, this is translated to 26.3 Gb of data, in equality of 58 sequencing SMRT cells at output of 450 Mb per cell.

In addition, the cost for PacBio sequencing data per base pair was not cheap as compared to second-generation sequencing. It is preferable that the long reads performed self-error correction in order to eliminate transmission of inherent error profile from another sequencing platform and to reduce length trimming due to lack of reads coverage from the other reads pool (that may suffer sequencing bias). The strategy may be the best options for any future de novo sequencing of genome, but at its current price, generating 50-fold coverage for large size eukaryotic genome can be difficult for many researchers, especially in developing countries.

Nevertheless, the potential of PacBio long reads to finish assembly of genome into finished, single contig by shotgun sequencing is undisputable and has been proven (Koren et al. 2012a; Chin et al. 2013; Huddleston et al. 2014). But all these were limited only on bacterial genome with size range of 2–6 Mb, which enable deep sequencing with just few SMRT cells run on PacBio platform. For complex plant genome, this would require many SMRT cells to achieve sufficient coverage. Alternative to this is by using hybrid sequencing technology to borrow the high accuracy from the second-generation sequencing technology in improving the long reads of PacBio. In addition, many sequencing genome projects have started using the second-generation sequencing. This data could not possibly be wasted and should be utilized for what it is best for, and that is the accuracy. Recently, the method has deemed successful with complete assembly of several genomes (Koren et al. 2012b; Ribeiro et al. 2012; Pendleton et al. 2015) and to a lesser extent to improve the contiguity of complex genome such as orchid (Yan et al. 2015).

In the motivation to sequence the pineapple genome, the main challenge relies on its heterozygosity and recalcitrant to self-pollinate. The innate parthenocarpic nature of the plant prevents the development of in-breed lines to facilitate its sequencing project. The presence of high number of multi-alleles in the genome complicates the assembly process especially at the contigging process as it caused the formation of “bubble” structures due to the mismatch. In the assembly of pineapple genome of hybrid F153, the problem of heterozygosity is reduced by using the haplotype phasing methods to eliminate one of the haploid copies to reduce the complexity of the assembly (Ming et al. 2015).

In the assembly of the commercially important MD-2 pineapple, long sequencing read technology is used to tackle the problem of repetitive and complex multi-allelic regions of the genomes. However, due to the high random error that is innate at low coverage of the PacBio long reads, the sequence reads demand accuracy improvements prior to its direct use in whole-genome sequencing assembly. The approach used in this project is to combine the two leading-edge sequencers (i.e., Illumina and PacBio) in a hybrid assembly to construct a draft for MD-2 pineapple genome.

Three different strategies were tested to find the most optimal pipeline that can produce an assembly that is complete as defined by the assembly size, accurate as defined by the content of gene predicted, and contiguous as defined by the scaffold size and N50. In the first strategy, de novo assembly of short-insert reads was improved by using PBJelly to perform gap-filling and scaffolding by applying the PacBio subreads (i.e., uncorrected). Secondly, the contigs from the short reads assembly were used as anchor in assembling the uncorrected PacBio long reads using the newly developed DBG2OLC software (Ye et al. 2016). Finally, following error correction of the PacBio long reads by using the Illumina short reads through novoLR package (Hercus 2015), the error-corrected PacBio reads were de novo assembled using traditional overlap-layout-based assembler, Celera (Myers et al. 2000). Assemblies from the three strategies were then selected based on the basic assembly metrics, the number of pineapple’s transcripts mapped to the genome, and the number of core eukaryotic gene found in the genome through assessment using CEGMA.

Sample Materials

The MD-2 pineapple was obtained from Malaysia Pineapple Industry Board and was maintained at Biotechnology Research Institute, UMS, for pineapple laboratory work. In this study, all genomic DNA extraction was performed on the pineapple leaves from a single plant.

De Bruijn-Based Assembly Using Only Short Reads

In finding the most optimal assemblers for the high-heterozygous genome of pineapple, three different assemblers were chosen based on its known credibility and specialty to handle complex genome. The result of the quality assessment for the three assemblies by using Assemblathon (Earl et al. 2011) was tabulated in Table 9.1.

Table 9.1 Summary of assembly metrics across three different pineapple draft genomes produced using the respective assembly software

	Platanus	SOAPdenovo	ABySS
Total base	4.01E + 08	3.34E + 08	6.18E + 08
Number of reads	133,557	430,236	1,710,293
Max length	128,428	131,768	124,558
Mean length	3004.41	776.86	361.48
Median length	775	138	93
Min length	91	100	64
N50	10,670	6122	3678
N75	4273	1259	199
N90	1482	175	93
N95	781	119	70

The basic statistic of the assembly metrics was summarized in plot for comparison (Fig. 9.1). In comparison, Platanus produced assembly with the highest N50, followed by SOAPdenovo and then ABySS. However, SOAPdenovo produced assembly with the longest scaffold, followed by Platanus and ABySS. Assembly by ABySS achieved total assembly size larger than the estimated genome size (526 Mb) and with the most number of scaffolds. This and the lack of its contiguity indicated the failure of ABySS to collapse the haplotypes leading to assembly with intermixed homologous sequences within the assembly.

Meanwhile, even though SOAPdenovo produced the longest scaffold, most of the remaining scaffolds were small in size causing the N50 to be low. In addition, the genome coverage from SOAPdenovo only achieved 63.5% of the estimated genome size, and this is the result of collapsed assembly, most probably at the repetitive region. On the other hand, even though Platanus failed to produce the longest scaffold, its N50 was the highest, which indicated that most of the scaffolds produced were with larger length than the other two assemblies. Its genome coverage was also higher than SOAPdenovo, by 12.7%. Overall, based on the genome coverage and contiguity, Platanus assembler was the most optimal to assemble the heterozygous genome of pineapple. The assembly was then selected and improved using PBJelly software (English et al. 2012) for gap-filling and scaffolding.

Following gap-filling and scaffolding by PBJelly, the N50 of the initial Platanus assembly increased almost threefold, and also the number of scaffolds reduced to 56,179, which was less than half of the initial number. In fact, all of the assembly metrics improved considerably after processing using PBJelly (Table 9.2).

Notably, there were over 60% of gaps that were filled and another 15% that were extended (Table 9.3). Nevertheless, from the gap statistics, it can be implied that much of the gaps filled by PBJelly were from the small-sized gap, as judging by the smaller improvement of N50 as compared to the N95 of the gap size. This is most probably due to the mean read length of PacBio reads which was 7669 bp. Thus, many of the gaps that can be resolved were among the small-sized gaps. In addition,

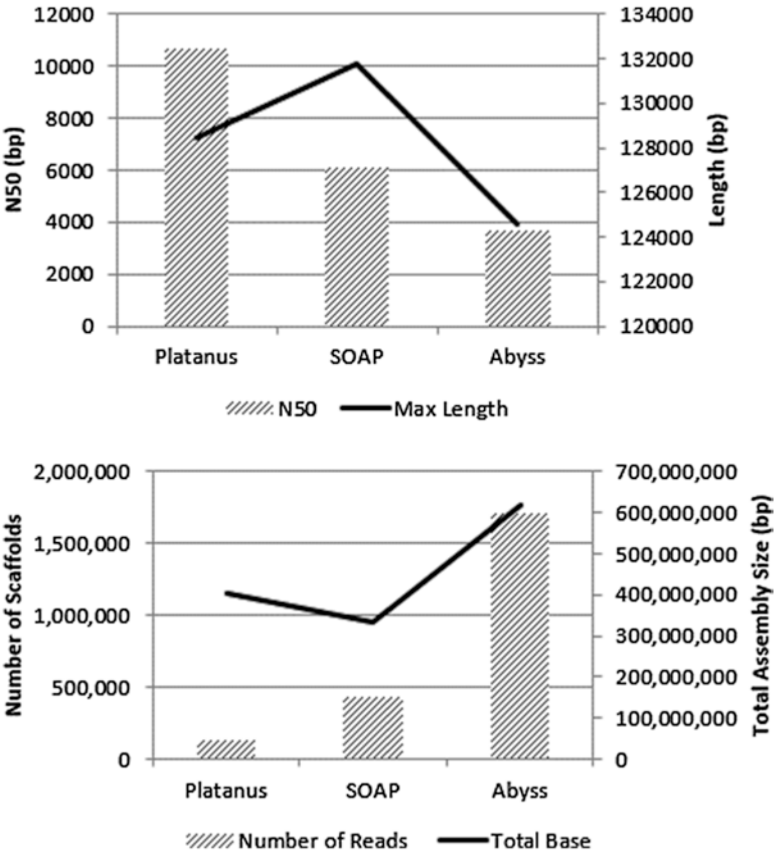

Fig. 9.1 Comparison of the assembly metrics of three different short-read assemblies using de Bruijn-based method

Table 9.2 Summary of the assembly metrics of the Platanus's assembly before and after processing using PBJelly for gap-filling and scaffolding

	Platanus	PBJelly	Improvements
N50 (bp)	10,670	30,811	3×
N75 (bp)	4273	11,390	3×
N90 (bp)	1482	2935	2×
N95 (bp)	781	1390	2×
Total assembly size (bp)	401,259,391	427,459,878	6.52% increment
Number of scaffolds	133,557	56,179	42% decrement
Max length (bp)	128,428	633,806	5×
Mean length (bp)	3004	7609	3×
Median length (bp)	775	1584	2×
Min length (bp)	91	234	3×
Mode length (bp)	100	535	5×

Table 9.3 Gap fill statistics for Platanus assembly after PBJelly

	Platanus	Platanus-PBJelly	Improvement
Gap count	33,738	14,301	2.3×
Gap N50 (bp)	217	243	1.1×
Gap N95 (bp)	46	72	1.6×
Total gap size	4,477,681	2,487,278	1.8×

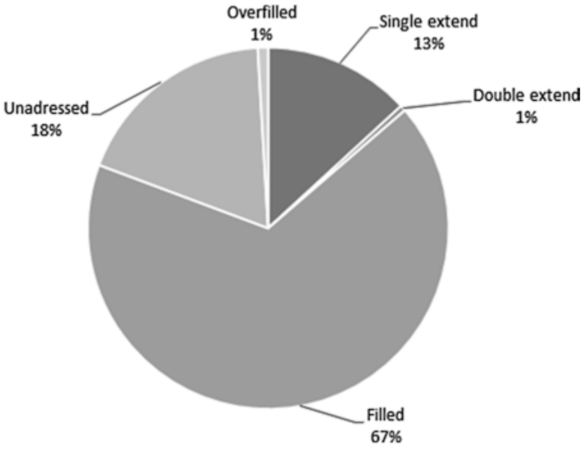

Fig. 9.2 Chart pie of gap improvement performed by PBJelly. The chart depicts the different proportion of the gaps that were treated by PBJelly either by filling, double extension, or single extension. Overfilled represent the gap that could not be improved due to inconsistent length of gap and reads that match the gap region. Unaddressed is the gaps that have no reads match to address the gap

it is also probable that with the longer reads and higher error rate of the PacBio reads, finding match to bridge large-sized gap was deemed difficult.

In improving genome's contiguity, several strategies were undertaken by PBJelly. Over half of the changes performed were by filling the gap (i.e., connect flanking sequences and fill the gap) and only 14% were by extending either one or both of the flanking sequences into the gap (Fig. 9.2). The 18% unaddressed gap were the gaps with "nofillmetric" status which indicated gaps that were unable to create any consensus sequences from PacBio reads to fill the gap, and the 1% "overfilled" status gaps were the gaps that have unmatched predicted gaps' size after the correction being performed by PBJelly. The fact that large proportion of the gap improvement performed was by filling the gap supports the previous observation that much of the gaps that have been improved were from the small-sized gap.

In comparison to other short-read-based assemblies, the strategy of using the PacBio long reads for scaffolding was shown to be feasible as it produced comparable contiguity as with assemblies that use mate-pair technology for scaffolding. The N50 achieved were similar to several other short-read assemblies such as hop (Natsume et al. 2014), sweet potato (Hirakawa et al. 2015), and horseweed (Peng et al. 2014). Nonetheless, there were also several other short-read-only assemblies that outperformed the draft genome by contiguity, achieving N50 of more than 100 kb. This includes the chickpea genome (Varshney et al. 2013), the pigeon pea genome (Varshney et al. 2011), and the genome of cotton (Wang et al. 2012a, b).

Table 9.4 Number of pineapple transcripts mapped to pineapple draft genome assembled using Platanus and PBJelly

Item	Counts
Transcripts mapped (114,077)	113,520
Transcripts mapped more than or equal to 80%	100,427
Transcripts mapped more than or equal to 90%	94,459

Number in parenthesis is the total number of pineapple transcripts used in mapping to the assembled contigs to assess its accuracy

Table 9.5 CEGMA assessment of the pineapple draft genome assembled using Platanus and PBJelly

Item	Counts
Number of CEGs mapped in complete	221
Number of CEGs mapped in partial	245
Percentage of paralogy	30.36%
Total number of KOGs found	449

It is worthwhile to note that all of these superior assemblies contained ultra-large insert size libraries, constructed using fosmid, and bacterial artificial chromosome system. In assembling plant genome, there were many factors that contribute to their superiority in contiguity. Factor such as the genome complexity which includes the level of heterozygosity, ploidy level, and presence of duplicated regions plays a major role in the assembly process (Schatz et al. 2012). In addition, it is also important that the assembly includes various insert size libraries of the sequencing data as this may facilitate in scaffolding the contigs to improve its contiguity.

The accuracy of the pineapple assembly was evident by the number of high-quality mapping of 114,077 pineapple transcripts to the assembly (Table 9.4). Only 0.5% of the transcripts were not mapped to the draft genome. Furthermore, 82% of the mapped transcripts were mapped in complete with more than 90% alignment length to the subject (i.e., the pineapple transcripts). About 11.48% of the transcripts were mapped but with poor query coverage (i.e., transcripts coverage of less than 80%). Transcripts that were mapped in poor coverage suggested that the assembly contains missing exon or misassembled region leading to incomplete mapping of the transcripts. Alternatively, since some of the transcripts originated from RNASeq transcriptome assembly, there were also chances that the missing transcripts were by themselves misassembled.

In addition, the genome was also evaluated using CEGMA to identify 248 highly conservative core eukaryotic genes (CEGs) within the draft genome sequences. Table 9.5 shows the number of CEGs identified either in complete or in partial. Gene found in complete indicates gene that has alignment length of more than 70% to the genes and in partial for less than 70%. The draft genome assembled using Platanus and PBJelly reached completeness of 89% based on complete alignment but 98% based on partial alignment. In this context, the completeness refers to the complete set of core eukaryotic genes the genome contained. From the result, it can be implied that the genome is almost complete by the presence of the core genes it encoded. However, some of the genes were incomplete and contained missing coding region, leading to partial mapping of the core genes. The number of orthologous genes set was moderate with only 30% of the genes found contained more than one ortholog. This is expected of plant genome that usually contained duplicated set of genes. A total of 449 out of 458 KOGs (euKaryotic Orthologous Group) were identified within the genome. This number is lower when compared to what has been identified within the chickpea (Varshney et al. 2013) and pigeon pea genome (Varshney et al. 2011). It is important to note that both of the legume genomes achieved much higher N50 as compared to the pineapple draft genome produced using Platanus and PBJelly. In comparison to the hop genome, which had scaffold N50 similar to the Platanus-PBJelly assembly, the number of CEGs found was higher by 7.75% (Natsume et al. 2014).

De Novo Assembly of Error-Corrected Long Reads by Mapping

In the second strategy of using DBG2OLC software, error-corrected long reads was assembled using overlap-layout graph method but with assistance of the contigs from Platanus. The first stage of assembly using DBG2OLC produced contigs (i.e., no Ns or gap) from the assembled long reads as no scaffolding was performed by the assembler. Table 9.6 showed the summary of assembly metrics produced by Assemblathon in quality assessment of the DBG2OLC's assembly.

The assembly produced by DBG2OLC showed impressive contiguity even at the contig level with 5771 numbers of contig and N50 of 162,783 bp. In addition, the assembly also produced 14 contigs with size of more than one million bp, and the longest contigs were almost two million bp. Contigs at this size could be the bases to build the genome from the contig to the chromosome level. Nevertheless, similar to the short-read assembly, the draft assembled by DBG2OLC also suffered from collapsed assembly size with only 82.2% coverage of the estimated pineapple genome size. The reason behind this is probably due to the collapsed Platanus assembly that was used to anchor the long reads prior to their assembly. Thus, even though DBG2OLC was able to assemble the long reads by using compressed long-read data, the assembly was inevitably limited to the inherent disadvantage of short-read assembly and that is the collapsed assembly size.

Subsequently, the draft was further improved by scaffolding using all available sequencing data from short reads to the transcriptomic data. The summary of the assembly's metrics was shown in Table 9.7. The scaffolding process improved the genome's contiguity significantly. The scaffold N50 was increased by twofold, and the number of sequences was reduced by 42% than the initial assembly. Impressively, the number of large scaffold with size of more than one million bp had increased from 14 to 42 sequences, and the largest scaffold achieved length of more than two million bp. The level of contiguity that this draft showcased were comparable to the moso bamboo draft genome assembled by using intensive BAC and fosmid system

Table 9.6 Assembly metrics of contigs from pineapple draft genome assembled using DBG2OLC

	Contigs
Number of sequences	5771
Total size (bp)	432,500,402
Longest sequences (bp)	1,963,534
Shortest sequences (bp)	1680
Number of sequences >500 nt	5771 (100.00%)
Number of sequences >1 K nt	5771 (100.00%)
Number of sequences >10 K nt	5293 (91.70%)
Number of sequences >100 K nt	1125 (19.50%)
Number of sequences >1 M nt	14 (0.20%)
Mean sequence length (bp)	74,944
Median sequence length (bp)	32,958
N50 sequence length (bp)	162,783
L50 sequence count	627
N75	65,338
N90	29,206
N95	19,692
Sequences %A	31
Sequences %C	19
Sequences %G	19
Sequences %T	31

Number in parenthesis corresponds to the percentage of sequence counts within the respective length limit

to provide for large mate-paired data (Peng et al. 2013). This highlights the possibility of the long-read third-generation sequencing technology to replace the traditional time-consuming and laborious BAC and fosmid cloning system in increasing genome's contiguity.

Nevertheless, the inherent problem of collapsed assembly size still remained unsolved with an increase of only 2% of the total size after scaffolding. The final genome only covered 84% of the estimated genome size of pineapple. This is the average genome coverage observed with draft genome assembled using short reads such as strawberry (Shulaev et al. 2011), flax (Wang et al. 2012a, b), and apple (Velasco et al. 2010). Its incompleteness was also indicated by the number of unmapped reads upon mapping of the short reads onto the draft assembly (Table 9.8).

The draft assembly's accuracy assessment by transcriptome mapping revealed its lower completeness as compared to the previous PBJelly draft assembly (Table 9.9). The assembly contained 804 missing transcripts as compared to 503 in PBJelly draft assembly. Consequently, the genome also contained less number of perfect transcript mapping than the PBJelly draft assembly.

In addition, the CEGMA assessment also highlighted its lower completeness as it contained lesser number of CEGs than the previous short-read assembly. The genome marked completeness of 98.39% for all the CEGs that had been identified within the genome (Table 9.10). The genome encoded one less ultra-conserved

Table 9.7 Assembly metrics of assembly from pineapple draft genome assembled using DBG2OLC at contig and scaffold level

	Contigs	Scaffolds
Number of sequences	5771	3325
Total size (bp)	432,500,402	444,262,876
Longest sequences (bp)	1,963,534	2,208,934
Shortest sequences (bp)	1680	1680
Number of sequences >500 nt	5771 (100.00%)	3325 (100.00%)
Number of sequences >1 K nt	5771 (100.00%)	3325 (100.00%)
Number of sequences >10 K nt	5293 (91.70%)	3104 (93.40%)
Number of sequences >100 K nt	1125 (19.50%)	1141 (34.30%)
Number of sequences >1 M nt	14 (0.20%)	42 (1.30%)
Mean sequence length (bp)	74,944	133,613
Median sequence length (bp)	32,958	53,683
N50 sequence length (bp)	162,783	326,628
L50 sequence count	627	360
Sequences %A	31.0	29.84
Sequences %C	19.0	18.83
Sequences %G	19.0	18.86
Sequences %T	31.0	29.83
N75	65,338	144,165
N90	9206	58,670
N95	19,692	32,808

Number in parenthesis corresponds to the percentage of sequence counts within the respective length limit

Table 9.8 The number of short reads mapped to the DBG2OLC draft assembly

	350 bp	550 bp	750 bp
Percentage of unmapped reads	1.15645	1.56049	2.57544
Percentage of sub-par quality mappings	14.13	16.98218	11.43325
Number of proper paired reads	353,649,810	342,696,820	27,701,719
Percentage of proper pairs	7.61E + 01	70.49757	74.35671

CEGs as compared to the previous assembly. Nevertheless, the genome still contained more number of identified CEGs in complete than the later assembly. This probably attributed to its higher genome contiguity than the PBJelly assembly.

Overall, the result showed that with increased contiguity the DBG2OLC had lost some part of the genome. Most probably the assembly avoided the complex region that enabled its increased contiguity. This phenomenon was previously observed in Assemblathon 1 (Earl et al. 2011). In the study, known simulated sequencing data upon testing with several assemblers produced draft genome with different level of contiguity and accuracy, and usually there were a trade-off of accuracy when the contiguity was superior (El-Metwally et al. 2014). Most importantly, there is also a concern of misassembled genome that similarly will also lead to missing or rather

Table 9.9 Number of transcripts mapped to the draft genome assembled by DBG2OLC

Item	Count
Number of transcripts mapped (114,077)	113,273
Number of transcripts mapped more than or equal to 80%	100,154
Number of transcripts mapped more than or equal to 90%	93,628

Table 9.10 CEGMA assessment of the pineapple draft genome assembled using DBG2OLC

Item	Counts
Number of CEGs mapped in complete	231
Number of CEGs mapped in partial	244
Percentage of paralogy	44.6%
Total number of KOGs found	447

poor mapping of the transcripts. After several tens of the genomes had been published, a group of researchers inspected several of the published drafts and alarmingly found more than hundreds of misassemblies (Salzberg and Yorke 2005). Thus, in assembling reference genomes, accuracy should be of top priority to ensure that the most precise data are being delivered to the public database especially for further downstream genome analysis.

De Novo Assembly of Error-Corrected Long Reads

Another assembly using the error-corrected long reads was attempted without the assistance of the short reads. This was inspired by the traditional strategy of performing whole-genome shotgun methods during the first-generation sequencing data. Celera assembler was designed to assemble long Sanger sequencing reads and thus can efficiently handle long-read sequences from PacBio. In the second strategy, 15.8× of error-corrected PacBio reads which accounted for 3,334,620 reads and 8.3 Gb of high accuracy long-read sequence data were assembled using Celera software. Because of the lack of mate-pair reads, the assembly process by Celera was only up to "untigging" process, and no scaffolding was performed in the run. The assembly by Celera produced contigs with assembly metrics summarized in Table 9.11.

The assembly produced 46,036 contigs with 50% of the assembly contained within 5773 contigs with size of at least 25,277 bp or larger. This contigs contained no ambiguous base and were produced after consensus calling performed within the Celera assembly run. Only one contig reached sequence length of more than one million bp, and majority of the contigs were sized less than 10,000 bp. Even though the N50 of the assembly was lesser as compared to the short-read-only assembly, the number of scaffolds produced was 22% less than the previous draft. This implied that most of the contigs in the assembly were longer in length, as confirmed by its

Table 9.11 Assembly metrics of contigs from pineapple draft genome assembled from error-corrected PacBio long reads using Celera

Item	Count
Sequence counts	46,036
Total size (bp)	780,569,372
Longest length (bp)	1,217,037
Shortest length (bp)	1079
Number of sequence >500 nt	46,036 (100%)
Number of sequence >1 K nt	46,036 (100%)
Number of sequence >10 K nt	22,615 (49.10%)
Number of sequence >100 K nt	868 (1.90%)
Number of sequence >1 M nt	1 (0%)
Mean sequence size (bp)	16,956
Median sequence size (bp)	9871
N50 sequence length (bp)	25,277
L50 sequence count	5773
Sequence %A	29.98
Sequence %C	20.03
Sequence %G	20.04
Sequence %T	29.95
N75	11,929
N90	7678
N95	5700

Number in parenthesis corresponds to the percentage of sequence counts within the respective length limit

Table 9.12 The number of unmapped transcripts in the draft genome of Celera at contig level

	Counts
RNASeq without hit (39,859)	81
EST without hit (5941)	174
Long PacBio RNA sequencing (68,277)	114
Total of all pineapple's transcript with no hit	369

Number in parenthesis refers to the total number of available transcripts that were mapped

much larger N90 as compared to the previous short-read assembly. Nonetheless, the contiguity of the contigs produced was far fragmented as compared to DBG2OLC's draft assembly.

Despite that, the assembly was still being considered because by far it contained the most number of transcripts mapped, which depicts its highest accuracy (Table 9.12). The draft encoded 99% of the RNASeq assembled transcripts and the pineapple EST obtained from public database.

On the contrary to the previous drafts, this assembly suffered from an inflated assembly size, having a total size that was 48.4% larger than the estimated pineapple haploid genome size. A similar observation of inflated assembly caused by

Table 9.13 The number of short reads mapped to the contigs from Celera draft assembly

	350 bp	550 bp	750 bp
Percentage of unmapped reads	0.18787	0.4217	0.8881
Percentage of sub-par quality mappings	62.82377	61.1894	58.32061
Number of proper paired reads	3.9E + 08	3.96E + 08	31,133,759
Percentage of proper pairs	36.74439	38.02799	39.3169

heterozygous genome was also observed previously in assembling the polymorphic genome of *Ciona savignyi* (Vinson et al. 2005). In addition, it is assertive that the problem is caused by the high level of heterozygosity of pineapple as indicated by the low number of proper paired reads mapped to the assembled contigs as shown in Table 9.13. The table showed the number of short reads that were mapped back to the contigs using Novoalign, and the alignment file produced was then assessed using QATools (https://github.com/CosteaPaul/qaTools). The software reports the number of reads mapped in paired and uniquely (i.e., exactly once). Only the reads mapped in paired and uniquely were considered as proper pairs, and the sub-par mapped reads were the low-quality mapped reads (caused by multi-mapped). The low percentage of proper pairs as shown in the table was evidential to the redundancy caused by variance between the two homologous copies of the diploid pineapple genome. Due to high heterozygosity, the diploid allelic copies of the genome were unsuccessfully collapsed into one reference to produce a single haploid reference draft genome.

In assembling genomes, the heterozygous loci can cause the emergence of “bubble” along the assembly path. This bubble appeared in the presence of heterozygous loci between two homozygous loci within an assembly path, and to resolve the problem, the bubble was popped to produce one linear assembly path. The Celera assembler, with the option of “utgBubblePopping” turned on, can collapse the paths within the bubble into sequence alignment and perform consensus calling to represent both. Even though the options were enabled for the assembly, the redundancy caused by allelic copy of the genome was still apparent.

Thus, to produce only a single haploid representation of the genome, it is imperative that the similar contigs need to be removed, and the longest representation of the allelic contigs should be chosen in the final draft. The first-stage redundancy removal by global similarity search, to remove short contigs (i.e., length of below 25,000 bp) with similarity of more than 80% to the longer contigs, was successful to remove 31% of the original contigs. However the total assembly size was only been reduced by 11% as most of the contigs that have been removed are of the short size. At this point, other allelic copy of the contigs could not be removed because of the large allelic differences that occurred within one contig. In addition, higher variant between the haplotypes in the genome also complicated the assembly process causing the assembler to build multiple composite of polymorphic paths that could be the real haplotype or else just spurious assembly error. These multiple composite assembly paths eventually would be long and significantly different among each other as the assembly graph traverses further. Hence, simple redundancy removal by

Table 9.14 The number of short reads mapped to the contigs from Celera draft assembly after redundancy removal

	Libraries		
	350 bp	550 bp	750 bp
Percentage of unmapped reads (%)	0.66688	1.31595	2.67409
Percentage of sub-par quality mappings (%)	6.87299	6.51996	5.42978
Number of proper paired reads	318,304,899	322,077,592	26,068,927
Percentage of proper pairs (%)	91.43317	90.65051	88.20062

global similarity search as above would not be sufficient to discard redundant allelic contigs of the assembly.

In order to rigorously remove the redundant contigs, the assembly was first split at the region where there was weak short-read support. The region that was with weak short-read support was recognized by low-quality mapping (Qscore below 10) within the alignment file produced after mapping back the short reads onto the contigs using Novoalign. The second-stage redundancy removal performed on the split contigs successfully removed 35% of the original assembly bringing the total assembly size to 508 Mbp, which was 96.5% of the pineapple haploid genome size. At this point the number of contigs was reduced to 30,585, and the N50 was slightly increased to 26,588 bp.

Most importantly, the number of reads mapped back to the draft assembly at this point had improved significantly (Table 9.14). This implied that much of the redundancy has been improved as most the short read were mapped in proper pair at exactly once. Furthermore, the low percentage of unmapped reads signified that much of the reads were assembled and included in the draft genome assembly and that the uses of the available reads thus far were saturated.

Subsequently, after the split and reduced contigs were merged back by multiple scaffolding process using all available sequencing data (including transcriptome), the draft assembly was improved significantly with only 18% of the number of contigs and more than six times better N50 than the initial assembly. The whole assembly and scaffolding processes with its milestone are summarized in Fig. 9.3, and the summary of the assembly statistics is given in Table 9.15.

The accuracy of the draft assembly by Celera was evident as only 348 of the 114,077 pineapple transcripts were not found within the genome, and 87% of the transcripts were mapped with more than 90% coverage (Table 9.16). Moreover, only 8.27% of the mapped transcripts were mapped in coverage of less than 80%, and the other 4.18% were mapped in coverage above 80%. The high-quality transcript alignment coupled with high mappability implied its accuracy and completeness.

Moreover, the draft assembly also achieved completeness of 98.79%, as 245 out of the 248 ultra-conserved CEGs were identified within the genome (Table 9.17). Two hundred thirty-one of the identified CEGs were in complete. In parallel to their larger assembly size, the draft also contained higher percentage of paralogy.

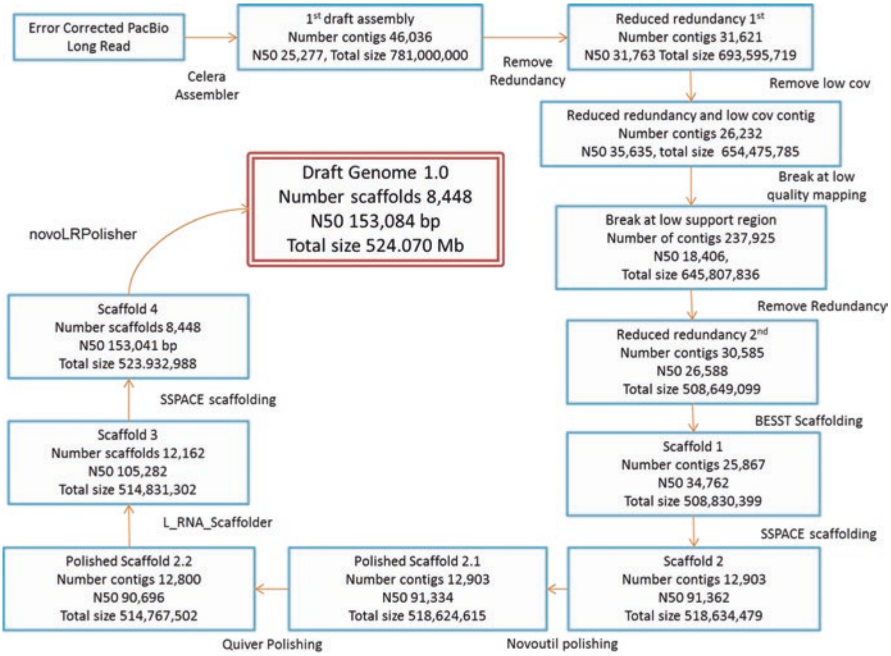
Error Corrected PacBio Long Read
Celera Assembler
1st draft assembly Number contigs 46,036 N50 25,277, Total size 781,000,000
Remove Redundancy
Reduced redundancy 1st Number contigs 31,621 N50 31,763 Total size 693,595,719
Remove low cov
Reduced redundancy and low cov contig Number contigs 26,232 N50 35,635, total size 654,475,785
Break at low quality mapping
Break at low support region Number of contigs 237,925 N50 18,406, Total size 645,807,836
Remove Redundancy
Reduced redundancy 2nd Number contigs 30,585 N50 26,588 Total size 508,649,099
BESST Scaffolding
Scaffold 1 Number contigs 25,867 N50 34,762 Total size 508,830,399
SSPACE scaffolding
Scaffold 2 Number contigs 12,903 N50 91,362 Total size 518,634,479
Novoutil polishing
Polished Scaffold 2.1 Number contigs 12,903 N50 91,334 Total size 518,624,615
Quiver Polishing
Polished Scaffold 2.2 Number contigs 12,800 N50 90,696 Total size 514,767,502
L_RNA_Scaffolder
Scaffold 3 Number scaffolds 12,162 N50 105,282 Total size 514,831,302
SSPACE scaffolding
Scaffold 4 Number scaffolds 8,448 N50 153,041 bp Total size 523.932,988
novoLRPolisher
Draft Genome 1.0 Number scaffolds 8,448 N50 153,084 bp Total size 524.070 Mb

Fig. 9.3 Methods of scaffolding and polishing the Celera assembly with respective milestone of assembly improvement after each process

Table 9.15 Summary of assembly statistics of Celera assembly of error-corrected PacBio reads before and after improvements

	Initial contig	Final contig	Final scaffold
Number of sequences	46,036	18,127	8448
Total size (bp)	780,569,372	509,962,048	524,069,662
Longest sequences (bp)	1,217,037	1,227,022	1,287,057
Shortest sequences (bp)	1079	1	1002
Number of sequences >500 nt	46,036 (100%)	17,782 (98.1%)	8448 (100%)
Number of sequences >1 K nt	46,036 (100%)	17,774 (98.1%)	8448 (100%)
Number of sequences >10 K nt	22,615 (49.1%)	11,245 (62%)	6372 (75.4%)
Number of sequences >100 K nt	868 (1.9%)	930 (5.1%)	1521 (18%)
Number of sequences >1 M nt	1 (0%)	1 (0%)	6 (0.1%)
Mean sequence length (bp)	16,956	28,133	62,035
Median sequence length (bp)	9871	13,557	24,886
N50 sequence length (bp)	25,277	58,665	153,084
L50 sequence count	5773	1987	901
Sequences %A	30	31	30
Sequences %C	20	19	19
Sequences %G	20	19	19
Sequences %T	30	31	30
N75	11,929	n/a	67,283
N90	7678	n/a	27,416
N95	5700	n/a	16,741

Table 9.16 Number of pineapple transcripts mapped to pineapple draft genome assembled using Celera

Item	Counts
Transcripts mapped (114,077)	113,729
Transcripts mapped more than or equal to 80%	104,297
Transcripts mapped more than or equal to 90%	99,532

Number in parenthesis is the total number of pineapple transcripts used in mapping

Table 9.17 CEGMA assessment of the pineapple draft genome assembled using Celera

Item	Counts
Number of CEGs mapped in complete	231
Number of CEGs mapped in partial	245
Percentage of paralogy	51.8%
Total number of KOGs found	447

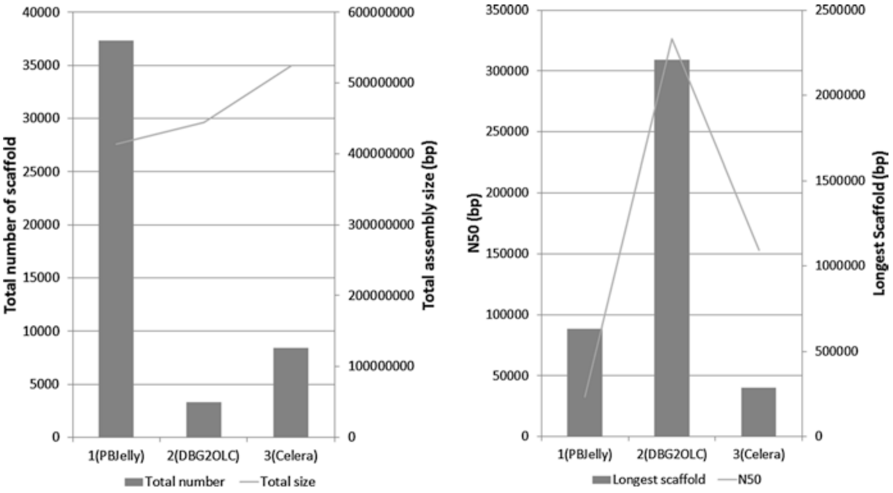

Fig. 9.4 Plot to compare the assembly metrics between drafts produced using three different strategies

Draft of MD-2 Pineapple Genome

In choosing the most optimal assembly, accuracy and completeness are the top priorities. This is crucial to ensure that the references when it serves as information gateway for further downstream genomic application can deliver the most accurate information. However, it is also important that the draft assembly achieved enough contiguity for it to provide significant genetic information (Fierst, 2015).

When comparing the three strategies used, the assembly performed by the DBG2OLC software was the most superior in terms of contiguity (Fig. 9.4). It has the least number of final scaffold numbers, but similar to the short reads, the draft

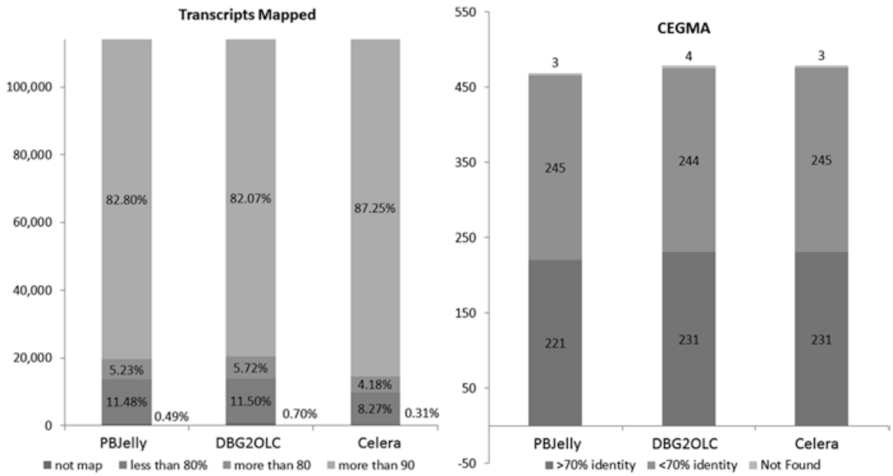

Fig. 9.5 Plot representing the comparison of the number of mapped transcripts (above) and CEGMA (below) among the draft assembled by the three strategies

assembled also fell short in genome coverage (i.e., collapsed total assembly size). The assembly produced using short reads and improved by long reads to scaffold was the most fragmented with an increment of the scaffold numbers of more than fourfold than the other two strategies. The contiguity of error-corrected long reads assembled using Celera was more than twofold in inferiority compared to DBG2OLC as judged by the number of scaffolds, N50, and the length of the longest scaffold. Despite that, the total assembly size of draft assembly produced by Celera was the closest to the estimated haploid genome size of pineapple. Both the other two strategies produced total assembly size of less than 90%.

Most importantly, the Celera assembly achieved the highest accuracy as assessed by transcript mapping and CEGMA (Fig. 9.5). Interestingly, the most contiguous assembly (i.e., DBG2OLC draft assembly) had the least number of transcripts mapped, and overall it had the highest percentage of poorly mapped transcripts. This result supports the previous observation of compensation between contiguity and accuracy (Fierst 2015), especially after the emergence of second-generation sequencing where large repeats are usually collapsed. In addition, previous study had shown that with the short-read technology, several regions particularly the GC-rich region had escaped sequencing and, thus, were not present in the final draft (Chen et al. 2013). Even though DBG2OLC used the error-corrected long reads, the genome was assembled by using the short-read assembly as the foundation in order to simplify the assembly process. Thus, the assembly would include only what is present within the short-read assembly and disregards what is not present. The process was certainly effective in producing contiguous assembly. However it eliminates the advantage of the PacBio long reads which are known not to have any sequencing bias (Ferrarini et al. 2013). Hence, as proven accurate with decent contiguity, the assembly produced using Celera was chosen to be considered as the final genome draft of pineapple.

In comparison with other draft of plant genome sequences, the assembly of pineapple genome draft scored fairly well in terms of contiguity and much better than other drafts in terms of genome coverage (Table 9.18). Despite of the lower contiguity, the genome coverage of the draft was much better than the majority of vascular plants that have been sequenced thus far. This is contributed to the use of long

Table 9.18 Comparison of the assembly metrics of the available draft genomes of plant species

Plant	Platform	Number of scaffold	N50 (kb)	Percent coverage	References
Strawberry	454 Illumina SOLiD	3200	1300	87.0	Shulaev et al. (2011)
Pigeon pea	Illumina Sanger	137,542	516	72.7	Varshney et al. (2011)
Flax	Illumina	88,384	693	81.0	Wang et al. (2012a, b)
Chickpea	Illumina Sanger	7163	39,900	72.0	Varshney et al. (2013)
Bamboo	Illumina Sanger	277,278	328	97.7	Gui et al. (2007)
Apple	Sanger 454	122,146	16	81.3	Velasco et al. (2010)
Horseweed	Illumina 454 PacBio	13,966	33	92.3	Peng et al. (2014)
Hop	Illumina	132,476	37	80.0	Natsume et al. (2014)
Pear	454	142,083	88,114	96.0	Chagné et al. (2014)
Adzuki bean	Illumina 454	3883	703	75.0	Kang et al. (2015)
Common bean	Illumina 454 Sanger	708	5000	80.5	Schmutz et al. (2014)
Sweet orange	Illumina	4811	1690	87.3	Xu et al. (2013)
Cacao	Illumina 454 Sanger	4792	473	76.0	Argout et al. (2011)
Date palm	454 SOLiD Sanger	82,354	329	90.0	Al-Mssallem et al. (2013)
Oak	454 Illumina	1468	260	50.0	Plomion et al. (2015)
Pineapple	Illumina 454 Moleculo PacBio BAC	3133	11,800	72.6	Ming et al. (2015)
Pineapple	PacBio	8448	153	99.6	This study

sequencing read technology which enables the construction of large repeat, which otherwise collapsed with short-read sequencing technology.

The final genome draft was named as ACMD2, and its annotation was uploaded to DDBJ/ENA/GenBank database. This version described can be accessed under the accession number LSRQ00000000.

Acknowledgment We thank Hydayaty Yusoff and the Pineapple Board of Malaysia for the pineapple sample, Caroline Chan from Pacific Biosciences (Asia Pacific) and Dana Chow from TreeCode Sdn Bhd for assistance with the Pacific Biosciences RSII, and Novocraft Sdn. Bhd. for the computing facility used in this project. This project is funded by the Ministry of Education and the Ministry of Science, Technology and Innovation, Malaysia, through the Fundamental Research Grant Scheme (FRG0319-SG-2013) and Science Fund (SCF0087-BIO-2013), respectively.

References

Al-Mssallem IS, Hu S, Zhang X, Lin Q, Liu W, Tan J, Yu X et al (2013) Genome sequence of the date palm *Phoenix dactylifera* L. Nat Commun 4:1–9

Argout X, Salse J, Aury J-M, Guiltinan MJ, Droc G, Gouzy J, Allegre M et al (2011) The genome of *Theobroma cacao*. Nat Genet 43:101–108

Chagné D, Crowhurst RN, Pindo M, Thrimawithana A, Deng C, Ireland H, Fiers M et al (2014) The draft genome sequence of European pear (*Pyrus communis* L. "Bartlett"). PLoS One 9:e92644

Chen YC, Liu T, Yu CH, Chiang TY, Hwang CC (2013) Effects of GC Bias in next-generation-sequencing data on de novo genome assembly. PLoS One 8(4):e62856

Cheung F, Haas BJ, Goldberg S, May GD, Xiao Y, Town CD (2006) Sequencing *Medicago truncatula* expressed sequenced tags using 454 life sciences technology. BMC Genomics 7:1–10

Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A et al (2013) Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods 10:563–569

Devine SE, Chissoe SL, Eby Y, Wilson RK, Boeke JD (1997) A transposon-based strategy for sequencing repetitive DNA in eukaryotic genomes. Genome Res 7:551–563

Earl D, Bradnam K, John JS, Darling A, Lin D, Fass J, On H et al (2011) Assemblathon 1: a competitive assessment of de novo short read assembly methods. Genome Res 21:2224–2241

El-Metwally S, Ouda OM, Helmy M (2014) Assessment of next-generation sequence assembly. In: Next generation sequencing technologies and challenges in sequence assembly. Springer, New York, NY, pp 95–101

English AC, Richards S, Han Y, Wang M, Vee V, Qu J, Qin X et al (2012) Mind the gap: upgrading genomes with Pacific biosciences RS long-read sequencing technology. PLoS One 7:e47768

Ferrarini M, Moretto M, Ward J a, Šurbanovski N, Stevanović V, Giongo L, Viola R et al (2013) An evaluation of the PacBio RS platform for sequencing and de novo assembly of a chloroplast genome. BMC Genomics 14:670

Fierst JL (2015) Using linkage maps to correct and scaffold de novo genome assemblies: methods, challenges, and computational tools. Front Genet 6:1–8

Gui Y, Wang S, Quan L, Zhou C, Long S, Zheng H, Jin L, Zhang X, Ma N, Fan L (2007) Genome size and sequence composition of moso bamboo: a comparative study. Sci China C Life Sci 50:700–705

He B, Caudy A, Parsons L, Rosebrock A, Pane A, Raj S, Wieschaus E (2012) Mapping the pericentric heterochromatin by comparative genomic hybridization analysis and chromosome deletions in *Drosophila melanogaster*. Genome Res 22:2507–2519

Hercus C (2015) novoLR package. In: Novocraft Technologies Sdn. Bhd. Kuala Lumpur, Malaysia

Hirakawa H, Okada Y, Tabuchi H, Shirasawa K, Watanabe A, Tsuruoka H, Minami C et al (2015) Survey of genome sequences in a wild sweet potato, *Ipomoea trifida* (H. B. K.) G. Don. DNA Res 22:171–179

Hoskins RA, Smith CD, Carlson JW, Bernardo A, Halpern A, Kaminker JS, Kennedy C et al (2002) Heterochromatic sequences in a *Drosophila* whole-genome shotgun assembly. Genome Biol 3:1–16

Huddleston J, Ranade S, Malig M, Antonacci F, Chaisson M, Hon L, Sudmant PH et al (2014) Reconstructing complex regions of genomes using long-read sequencing technology. Genome Res 24:688–696

Kang YJ, Satyawan D, Shim S, Lee T, Lee J, Hwang WJ, Kim SK et al (2015) Draft genome sequence of adzuki bean, *Vigna angularis*. Sci Rep 5:8069

Koren S, Harhay GP, Smith TPL, Bono JL, Harhay DM, Mcvey S, Radune D, Bergman NH, Phillippy AM (2012a) Reducing assembly complexity of microbial genomes with single-molecule sequencing. Nat Biotechnol 30:693–700

Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, Wang Z et al (2012b) Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nat Biotechnol 30:693–700

Michael TP, Jackson S (2013) The first 50 plant genomes. Plant Genome 6:1–7

Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH, Senin P et al (2008) The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya Linnaeus*). Nature 452:991–996

Ming R, VanBuren R, Wai CM, Tang H, Schatz MC, Bowers JE, Lyons E et al (2015) The pineapple genome and the evolution of CAM photosynthesis. Nat Genet 47:1435–1442

Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA et al (2000) A whole-genome assembly of *Drosophila*. Science 287:2196–2204

Natsume S, Takagi H, Shiraishi A, Murata J, Toyonaga H, Patzak J, Takagi M et al (2014) The draft genome of Hop (*Humulus lupulus*), an essence for brewing. Plant Cell Physiol 56(3):428–441

Pendleton M, Sebra R, Pang AWC, Ummat A, Franzen O, Rausch T, Stütz AM et al (2015) Assembly and diploid architecture of an individual human genome via single-molecule technologies. Nat Methods 12:780–786

Peng Y, Lai Z, Lane T, Nageswara-Rao M, Okada M, Jasieniuk M, O'Geen H et al (2014) De novo genome assembly of the economically important weed horseweed using integrated data from multiple sequencing platforms. Plant Physiol 166:1241–1254

Peng Z, Lu Y, Li L, Zhao Q, Feng Q, Gao Z, Lu H et al (2013) The draft genome of the fast-growing non-timber forest species moso bamboo (*Phyllostachys heterocycla*). Nat Genet 45:456–461

Plomion C, Aury J-M, Amselem J, Alaeitabar T, Barbe V, Belser C, Bergès H et al (2016) Decoding the oak genome: public release of sequence data, assembly, annotation and publication strategies. Mol Ecol Resour 16(1):254–265

Ribeiro FJ, Przybylski D, Yin S, Sharpe T, Gnerre S, Abouelleil A, Berlin AM et al (2012) Finished bacterial genomes from shotgun sequence data. Genome Res 22:2270–2277

Salzberg SL, Yorke JA (2005) Beware of mis-assembled genomes. Bioinformatics 21:4320–4321

Schatz MC, Witkowski J, McCombie WR (2012) Current challenges in de novo plant genome sequencing and assembly. Genome Biol 13:243

Schmutz J, McClean PE, Mamidi S, Wu GA, Cannon SB, Grimwood J, Jenkins J et al (2014) A reference genome for common bean and genome-wide analysis of dual domestications. Nat Genet 46:707–713

Shulaev V, Sargent DJ, Crowhurst RN, Mockler TC, Folkerts O, Delcher AL, Jaiswal P et al (2011) The genome of woodland strawberry (*Fragaria vesca*). Nat Genet 43:109–116

Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N et al (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & gray). Science 313:1596–1604

Varshney RK, Chen W, Li Y, Bharti AK, Saxena RK, Schlueter JA, Donoghue MTA et al (2011) Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. Nat Biotechnol 30:83–89

Varshney RK, Song C, Saxena RK, Azam S, Yu S, Sharpe AG, Cannon S et al (2013) Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. Nat Biotechnol 31:240–246

Velasco R, Zharkikh A, Affourtit J, Dhingra A, Cestaro A, Kalyanaraman A, Fontana P et al (2010) The genome of the domesticated apple (Malus × domestica Borkh.). Nat Genet 42:833–839

Vinson JP, Jaffe DB, O'Neill K, Karlsson EK, Stange-Thomann N, Anderson S, Mesirov JP et al (2005) Assembly of polymorphic genomes: algorithms and application to *Ciona savignyi*. Genome Res 15:1127–1135

Wang W, Feng B, Xiao J, Xia Z, Zhou X, Li P, Zhang W et al (2014) Cassava genome from a wild ancestor to cultivated varieties. Nat Commun 5:5110

Wang Z, Hobson N, Galindo L, Zhu S, Shi D, McDill J, Yang L et al (2012a) The genome of flax (*Linum usitatissimum*) assembled de novo from short shotgun sequence reads. Plant J 72:461–473

Wang K, Wang Z, Li F, Ye W, Wang J, Song G, Yue Z et al (2012b) The draft genome of a diploid cotton *Gossypium raimondii*. Nat Genet 44:1098–1103

Xu Q, Chen L-L, Ruan X, Chen D, Zhu A, Chen C, Bertrand D et al (2013) The draft genome of sweet orange (*Citrus sinensis*). Nat Genet 45:59–66

Yan L, Wang X, Liu H, Tian Y, Lian J, Yang R, Hao S et al (2015) The genome of Dendrobium officinale illuminates the biology of the important traditional Chinese orchid herb. Mol Plant 8:922–934

Ye C, Hill CM, Wu S, Ruan J, Ma Z (2016) DBG2OLC: efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. Sci Rep 6:31900