RESEARCH ARTICLE

Zdravka Ivanova
Evelina Daskalova
Georgi Minkov
Vesselin Baev

Chloroplast genome assembly approaches from NGS data

Authors' addresses:
University of Plovdiv "Paisii Hilendarski", Department of Plant Physiology and Molecular Biology, 24 Tzar Asen Str., Plovdiv 4000, Bulgaria.

Correspondence:
Vesselin Baev
University of Plovdiv "Paisii Hilendarski", Department of Plant Physiology and Molecular Biology, 24 Tzar Asen Str., Plovdiv 4000, Bulgaria.
Tel.: +359 32 629 495
e-mail: vebaev@plantgene.eu

Article info:
Received: 5 April 2017
In revised form: 18 April 2017
Accepted: 20 April 2017

ABSTRACT
The advent of Next Generation Sequencing platforms has led to significant research innovation in the field of whole genome assembly algorithms and software. The Illumina platforms produce large amounts of distinctive sequencing data with a shorter read length and higher coverage in comparison with Sanger Sequencing. In response to this, several new assemblers were developed specifically for *de novo* assembly of next generation sequencing data. This study compares software assembly packages named Edena, SPAdes, ABySS and analyzes the results delivered by *de novo* assembly experiments. We showed that a plastid genome assembly can be completed on a 32 bit Linux OS with 4 GB RAM, indicating that an experiment including *de novo* chloroplast assembly of millions very short reads can be executed successfully and quickly on a desktop computer.

Key words: Next-generation sequencing, Genome assembly, assembly algorithms, chloroplast genome

Introduction

The recent innovations in Next-generation sequencing technologies (Buermans & Den Dunnen, 2014), such as Illumina/Solexa, have dramatically reduced the cost of sequencing and increased the volume of short reads with high genome coverage generated from new target genomes. Managing this much data and accounting for sequencing errors are main challenges in the field of bioinformatics in general and *de novo* genome assembly in particular. Due to the short read lengths (shorter than even the shortest genome) and the large volume of data produced by NGS, Whole Genome Sequencing assembly is of the most complicated problems in bioinformatics. Nevertheless, rapid progress has been observed in the development of genome assembly methods and software tools recently.

Based on different approaches, all sequencers produce data in the form of short reads. These can be single end (SE) or paired end (PE) reads, where SE refers to only one end of a sequence, whereas PE refers to both ends of a sequence fragment. Nowadays, this approach provides a new solution for existing problems of *de novo* assembly. Genome assembly is a process of grouping (merging) short reads into contigs and then grouping those into scaffolds. Respectively, the output of assemblers is a set of contigs or scaffolds, which can be evaluated statistically in order to assess their accuracy and quality.

The assessment of assemblies can be performed by estimating the size and accuracy of their contigs or scaffolds. Several software tools and statistical metrics exist for this purpose. Assembly quality is defined by statistical measures such as maximum length of contigs, average length of contigs, and – most importantly – N50. N50 for a given assembly is the length of the smallest contig in the set of the largest contigs whose combined length represents at least 50% of the assembly. To estimate the accuracy of an assembly, the number and size of gaps obtained in it are used.

In this study, we evaluated chloroplast genome assemblies obtained from Edena, SPAdes and ABySS, then the results were compared and the final decision about the reliability and accuracy of these three assemblers was made.

Materials and Methods

NGS data from sequencing

The data source for this study is the recently sequenced, assembled and annotated 153099 bp chloroplast genome of the resurrection plant *Haberlea rhodopensis Friv*. (Ivanova et al. 2017). In the process of generating different assemblies, combinations of *de novo* and reference-based assembly methods were used. The chloroplast genome of the closely related resurrection species *Boea hygrometrica* was used as a reference (Zhang et al. 2012).

Library preparation and sequencing were performed at BGI-Shenzhen, China. Two biological replicates of

8,365,536 and 7,129,508 paired-end sequences, 100 bp in length with 170 bp library insert-size were generated by Illumina Technology.

FastQC

FastQC is a bioinformatics tool which aims is to provide a simple way for quality control of Next generation sequencing data. It generates summary graphs and tables in order to quickly assess NGS data.

Edena

Edena is a *de novo* short read assembler, developed for small genomes (Hernandez et al. 2008). It is standalone software, distributed freely only for non-profit academic use and available under the General Public License (GPLv3) at www.genomic.ch/edena.php. The assembler is based on the methods of overlap layout assembly framework. Edena requires all reads to have the same lengths, with a maximum of 128 nucleotides. It can accept both unpaired and paired short read in fasta and fastq formats. The software works in a two-step process – overlapping and assembling. The process begins by processing the short reads in order to remove redundant information. The second stage includes computing all overlaps and constructing overlap graphs. Finally, all contigs with a minimum size represented in the graphs are provided as an output.

SPAdes

SPAdes (St. Petersburg genome assembler) is a free assembler, designed for small genomes (Bankevich et al. 2012). It requires 64-bit Linux or Mac OS system and Python programming language. The assembler was released under GPLv2 and can be downloaded from http://cab.spbu.ru/software/spades/. SPAdes is based on Brujin graphs and generates single-cell and standard (multi-cell) assemblies. The current version of SPAdes works with Illumina or Ion torrent short reads of various types: paired-end reads, unpaired reads, mate-pair reads. It takes as input fasta and fastq files and supports unpaired reads in BAM format. SPAdes is used for generating hybrid assemblies using PacBio, Oxford nanopore and Sanger reads, and was recently integrated into Galaxy pipelines.

ABySS

ABySS is a de-novo parallel sequence assembler (Simpson et. 2009). It is developed for short paired-end reads, as well as small and large genomes. The assembler is an open-source, free software and can be downloaded from http://www.bcgsc.ca/platform/bioinfo/software/abyss. The algorithm is based on the Brujin graphs. It can parallelize the assembly of billions of short reads on either a dedicated server or on commodity hardware. The assembly is performed in two stages. First, the contigs are extended without using the paired-end information. Second, the paired-end information is used to merge them into contigs. The assembler works with single-end data and takes as an input fasta, fastq and BAM files.

QUAST

QUAST is an open-source, free software, available on http://bioinf.spbau.ru/quast (Gurevich et al. 2013). It is designed as a quality assessment tool for evaluating and comparing genome assemblies and can evaluate assemblies with and without a reference genome. QUAST performs comparison and produces statistical information in summary tables and plots. It is fast, parallelized and can be effectively run on multi-processor machines.

Mummer

MUMmer is an open source, free software, hosted at http://mummer.sourceforge.net/ (Kurtz et al. 2004). The package is developed for rapidly aligning entire genomes in a complete or draft form. In addition, the MUMmer package produces plots in order to visualize alignments.

Abacas

ABACAS (Algorithm Based Automatic Contiguation of Assembled Sequences) is developed for aligning, ordering and orientating assembled contigs based on a reference genome (Assefa

et al. 2009). The software package uses MUMmer to find the positions of contigs or scaffolds against a reference genome. It generates as output the pseudomolecule of the investigated genome and a comparison file which can be used for visualizing ordered contigs. The software is freely available and can be downloaded from http://abacas.sourceforge.net/.

FLASH

FLASH is a very fast tool to merge paired-end reads from fragments that are shorter than twice the length of reads (Magoc & Salzberg, 2011). The extended length of reads has a positive impact on the improvement of genome assemblies.

Results

Data pre-processing

The first stage of our analysis was the preprocessing of data from Next-generation sequencing. We merged short reads from NGS into longer reads using the FLASH software tool (Magoc & Salzberg, 2011). Next, we analyzed the quality of data obtained from FastQC. Quality controls are shown on Figure 1 and Figure 2.

Genome assembly using Edena, SPAdes and ABySS

Genome assemblies using three different assemblers Edena, SPAdes, ABySS were generated with the following setting: m=60, k-mers =, 75 and 53 respectively.

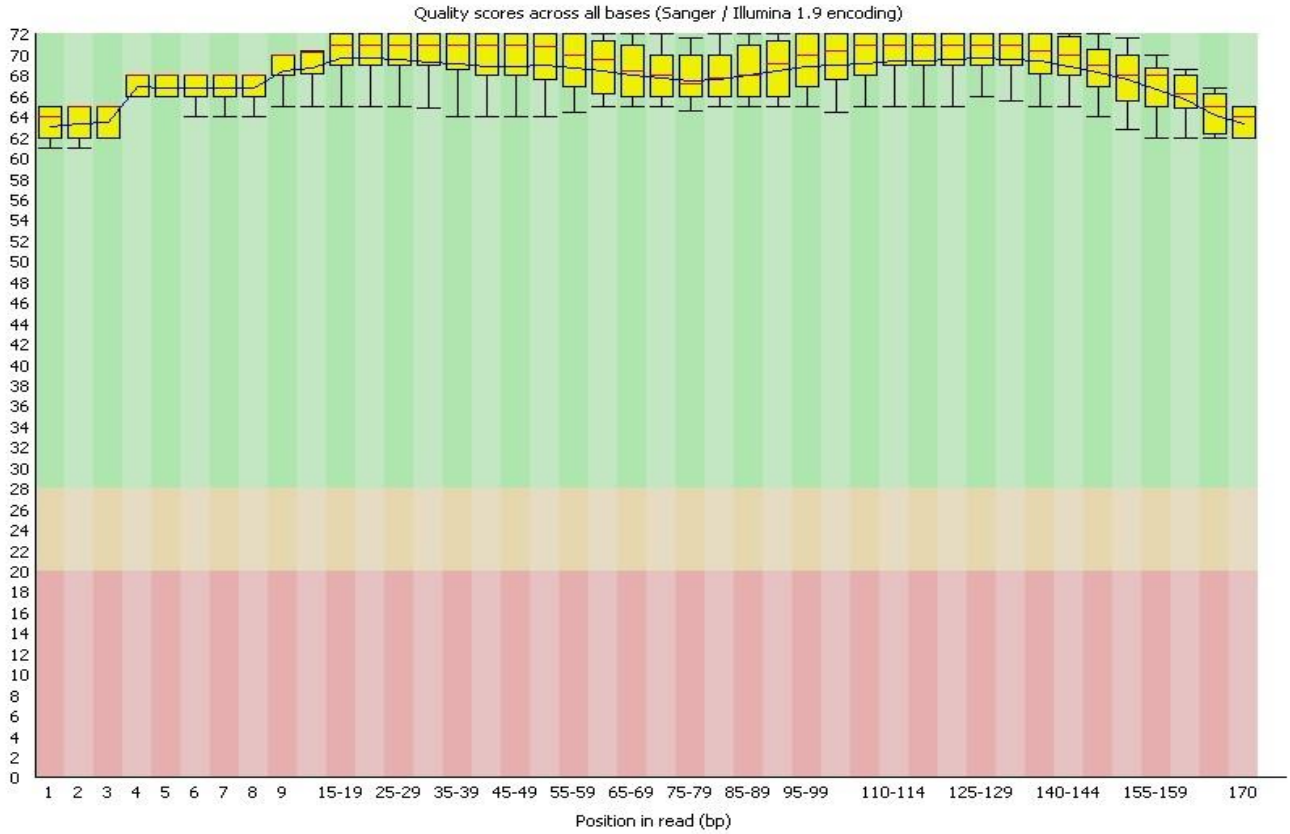

Figure 1. *Quality control of replicates 1 data.*

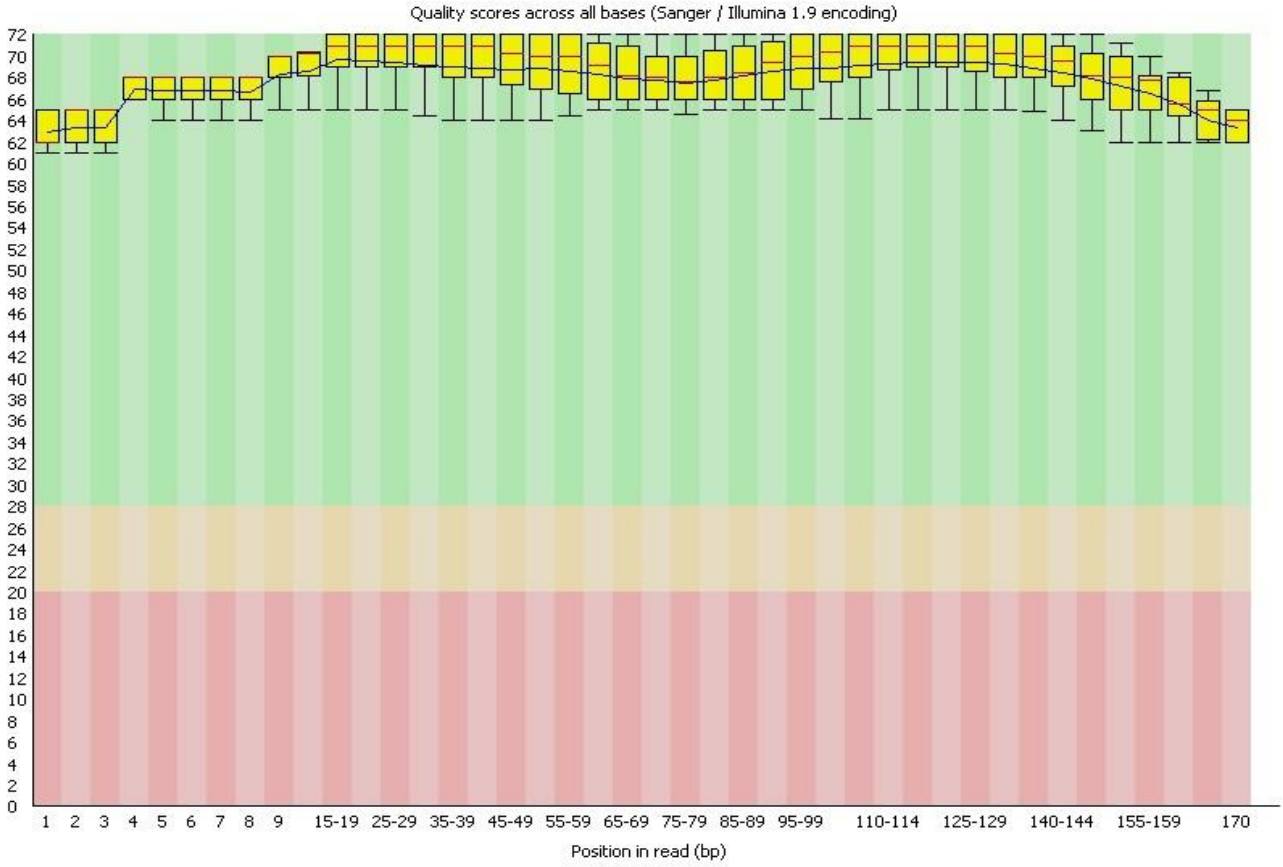

Figure 2. *Quality control of replicates 2 data.*

Statistical analysis of assemblies

Using the QUAST software package, we analyzed assemblies obtained from Edena, SPAdes and ABySS assemblers. The reports about the most important statistical parameters are shown in Table1.

Table 1. *Results from QUAST.*

	Edena	SPAdes	ABySS
contigs number	56	129	85
largest contig	11555	12657	27642
GC%	38.13	37.67	37.0
N50	2656	3922	17662
N75	1449	1645	9831

Analysis of gaps

Genome assemblies generated with Next generation sequencing short reads usually contain a number of gaps. We performed analysis of gaps with the ABACAS software tool. The results are summarized and presented in Table 2.

Table 2. *Gaps in assemblies from Edena, SPAdes and ABySS.*

	Edena	SPAdes	ABySS
gaps number	55	70	18
largest gap (nt)	15018	12279	7487
total gaps (nt)	48401	35739	25787

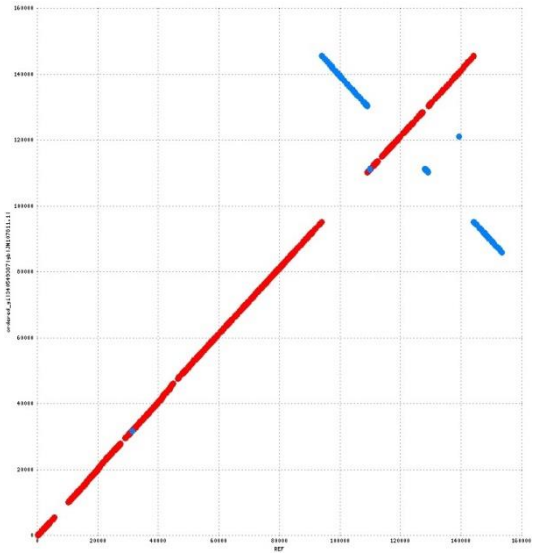

Figure 3. *MUMmer plot of Edena assembly.*

MUMmer plots of assemblies

In order to visualize the alignment between the draft genome of *Haberlea rhodopensis* and reference the genome of *Boea hygrometrica*, we used the software aligner MUMmer. Generated graphics are shown on Figures 3, 4 and 5.

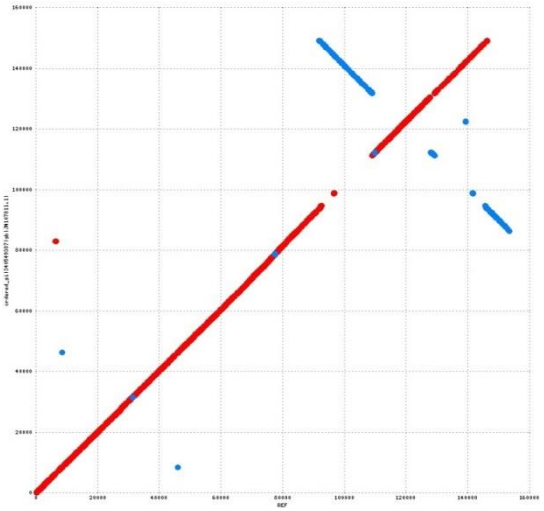

Figure 4. *MUMmer plot of SPAdes assembly.*

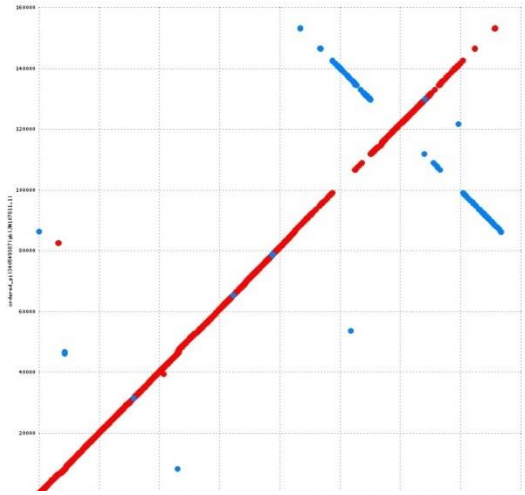

Figure 5. *MUMmer plot of ABySS assembly.*

Discussion

Preprocessing, overview of tested assemblers and assembly

Preprocessing and quality control is constitute major steps in the genome assembly pipeline (Chen et al. 2014). All merged NGS data show in Figure 1 and Figure 2 is located in the green region. On this base, we can conclude that data quality is excellent and sufficient to continue to the genome assembly process.

In recent years, many software tools have been developed for assembling whole-genome sequencing data. We can divide the most popular NGS graph-based methods into two groups based on their algorithms: The Brujin Graphs (DBG) and the Overlap-Layout-Consensus (OLC) methods (Pevzner et al. 2001; Compeau et al. 2011). For our experiment, we chose to use three assemblers: the OLC-based Edena and the DBG based SPAdes and ABySS. All three tools are designed to assemble small genomes, although ABySS is also capable of assembling large genomes, as well. It is worth mentioning

that, during our work with these assemblers, the time for assembly generation was very short – approximately 15 minutes or less per run (Table 3). The use of assemblers is relatively easy, requiring only the Linux command line.

Table 3. *The overview of tested assemblers.*

Assembler	Algorithm	Est. time per run	Genomes assembled
Edena	Overlap-Layout-Consensus	~ 15 min	bacterial genomes
SPAdes	De Bruijn Graph	~ 15 min	small genomes
ABySS	De Bruijn Graph	~ 15 min	small and large genomes

While different assemblers can produce a result of completely different qualities, they are still valid for the parameters we used in the process of genome assembly. We tested the assemblers based on DBG graphs in a wide range of mandatory k-mers in order to obtain sufficient quality of assemblies. We compared results and identified that k-mer = 75 for SPAdes and k-mer = 53 for ABySS give the best results for each assembler. The best results from Edena were obtained by using an overlap value of m = 60.

Analysis of assemblies – statistical estimation, gaps and visualization

We compared currently available and widely-used assemblers for Illumina reads with reads length of 100 bp and library insert size of 170bp. Table 1 shows a comparison between these three assemblers based on results from QUAST. We can see that Edena, SPAdes and ABySS produce different outputs.

In genome assembly, it is important to produce a small number of contigs with a high N50 value. We observed that Edena gives the smallest number of contigs, but ABySS generates contigs with the highest N50 value. Another important parameter in genome assembly is the size of the largest contig (Table 1).

One of the fundamental problems in genome assembly is the generation of fragmented assemblies due to gaps in draft genome sequences. Based on the number and size of gaps in the assembly, we can draw conclusions about the quality of the assembled short reads.

We analyzed gaps from our 3 experiments with Edena, SPAdes and AbySS respectively. Table 2 shows the comparison of gaps between assemblies generated by three assemblers. We observed the smallest number of gaps in the ABySS assembly and the highest number in the Edena assembly. The results in Table 2 show that both the size of the largest gap and the total size of gaps are also highest in Edena, while AbySS has the lowest number of gaps and the lowest total size of gaps. We conclude that the Edena assembly is more fragmented than other two while ABySS gives much better assemblies in terms of contigs size, N50 value, the size of gaps and the total size of gaps. This is confirmed by the MUMmer plots on Figures 3, 4 and 5.

Conclusion

In this paper, we presented three assemblers based on DBG and OLC algorithms and compared the results of their work. We observed that assemblies built from NGS reads are far from perfect. They are fragmented with many gaps and inconsistent statistical parameters such as N50 and contigs size. However, some assemblers performed much better than others, showing that improvement in development of assemblers is still possible. WGS assembly is an active area in bioinformatics with rapid progress in the design of new software tools, thus we hope that fundamental problems with current assemblers will be solved eventually.

References

Assefa S, Keane TM, Otto TD, Newbold C, Berriman M. 2009. ABACAS: algorithm-based automatic contiguation of assembled sequences. Bioinformatics, 25(15): 1968-1969.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahni N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol., 19(5): 455-477.

Buermans HP, Den Dunnen JT. 2014. Next generation sequencing technology: advances and applications. Biochim. Biophys. Acta, 1842(10): 1932-1941.

Chen C, Khaleel SS, Huang H, Wu CH. 2014. Software for pre-processing Illumina next-generation sequencing short read sequences. Source Code Biol. Med., 9: 8.

Compeau PE, Pevzner PA, Tesler G. 2011. How to apply de Bruijn graphs to genome assembly. Nat. Biotechnol., 29(11): 987-991.

Hernandez D, Francois P, Farinelli L, Osteras M, Schrenzel J. 2008. De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. Genome Res., 18(5): 802-809.

Ivanova Z, Sablok G, Daskalova E, Zahmanova G, Apostolova E, Yahubyan G, Baev V. 2017. Chloroplast genome analysis of resurrection tertiary relict Haberlea rhodopensis highlights genes important for desiccation stress response. 2017. 8: 204.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. Genome Biol., 5(2): R12.

Magoc T, Salzberg SL. 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics, 27(21): 2957-2963.

Pevzner PA, Tang H, Waterman MS. 2001. An Eulerian path approach to DNA fragment assembly. Proc. Natl. Sci. USA., 98(17): 9748-9753.

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(6): 1117-1123.

Zhang T, Fang Y, Wang X, Deng X, Zhang X, Hu S, Yu J. 2012. The complete chloroplast and mitochondrial genome sequences of Boea hygrometrica: insights into the evolution of plant organellar genomes. PLoS One, 7(1): e30531.