Chapter 10
Assessment of Next-Generation Sequence Assembly

Abstract Although there are different measures to evaluate assembler performance and assembly quality, developing assessment tools that incorporate present measures and defining new ones for the various assembly types (genomic, transcriptomic, and metagenomic) still remain a major challenge in the next-generation environment. In this chapter, we will introduce different approaches for assembly assessment as well as discuss upcoming assembly evaluation studies/tools.

10.1 Introduction to Assembly Assessment

The assessment of the assembly process is mainly performed from two perspectives. The first perspective is assembly quality, which evaluates the contiguity, consistency, and accuracy of the assembled genomes using different approaches [1–4]. The second perspective is the performance and usability of the assembler, which includes numerous issues such as hardware and software requirements, ease of installation and execution, user-friendly interfaces, run time per analysis, required memory per 1 GB of data, and the speed of responsiveness to user commands [5–8].

10.2 Contiguity and Consistency Measures

10.2.1 Contiguity Assessment

Statistics metrics are usually used to assess the contiguity of the assembled contigs/scaffolds. These metrics include the distribution of their lengths, their maximum, minimum and average lengths, the number of resulting contigs/scaffolds, the total sum of the assembled contigs/scaffolds, the total length of their short reads, and the N_x score. N_{50} and N_{75} represent the most important metrics for measuring contig/

S. El-Metwally et al., *Next Generation Sequencing Technologies and Challenges in Sequence Assembly*, SpringerBriefs in Systems Biology 7,
DOI 10.1007/978-1-4939-0715-1_10, © The Authors 2014

scaffold contiguity. They are defined as the length of the contig/scaffold such that 50 %/75 % of its bases are in contigs of greater or equal length [1–4, 9–12]. Although a large value of the N_x score indicates more contiguity in the assembled contigs/scaffolds, the misassembly of contig/scaffold sequences may also increase the score [13].

10.2.2 Consistency Assessment

Due to the presence of abundant information in paired-end libraries, including the estimation of insert size among each pair of reads and their orientation, approaches assessing consistency can utilize this information in the evaluation process. Following the completion of the assembly process, read pairs can be located in the draft sequence. In this case, a comparison of the assembly process with the annotated information of the read pairs (such as separation distance or orientation) can occur. Based on the number of satisfying constraints, we can infer the validity of the assembled sequence [14]. A recently introduced metric also utilizes the idea of aligning the paired-end reads to the assembled genome in generating Feature-Response Curves (FRC) to overcome the available tradeoff between the contiguity and accuracy of the assembly results [15, 16]. Other consistency methods target the type of sequence being assembled (such as haplotype sequences) [3] as well as the constraints imposed by the read coverage to assess the assembled sequences [17] or optical maps [18].

10.3 Accuracy Measures

Comparing the draft sequence assemblies to ones that have been completed represents the most important metric in evaluating the assembly quality [3, 9]. This reference can be an assembled genome of the same species or of a closely related species. The comparative process takes different perspectives such as aligning the two sequences using one of the available alignment tools (i.e., tools that were mentioned in Chap. 2) that report the percentage covered by the assembled sequence [5, 19], the long-range contiguity of the assembled contigs/scaffolds [20], their accuracy and the introduction of modification patterns in the assembled sequences such as insertions, deletions, and substitutions [21]. Furthermore, the comparison process assists in the identification of core genetic components and novel genes [22]. The number of misassembled contigs/scaffolds (i.e., breaks) and the number of misaligned bases (i.e., mis-calls) are also used as accuracy metrics in the context of alignment to a reference sequence [23]. Another perspective for assessment occurs during the unavailability of the reference genome. In this case, the comparative process requires independent genetic material from a public database. These genetic components (such as mRNA or cloned genes) can only be utilized if they and the assembled sequences belong to the same type of organism. When this criterion cannot be met, the accuracy approaches enlist components from closely related organisms or conserved sequences [1, 22].

10.4 Assembler's Performance Measures

The runtime and memory usage of an assembler are the most important criteria for the usability measure. Depending on the available computational resources, current assemblers used in next-generation environments are classified into two categories. In the first category, the assemblers run on a single machine with very large memory requirements, e.g., to assemble human and mammalian genomes [19, 24]. In the other category, assemblers are run on tightly coupled cluster machines [25]. The high-throughput nature of next-generation sequencing technologies and the presence of short-read sequences and their quality scores imposes a major constraint on the system memory available. To ensure efficient memory savings, most assemblers formulate the assembly problem as a set of graph nodes and rely on efficient data structures to accommodate these nodes. These different graph models were discussed earlier (see Sect. 9.3), including their advantages and disadvantages with respect to computational resources and several studies that reformulated their representations to ensure efficient storage in memory. However, no memory-efficient solution is presently available for next-generation sequence assemblers, creating a need for new tools and algorithms in this area.

10.5 Assessment Tools and Evaluation Studies for Assessing Assembly Quality

There are several studies for evaluating assembly quality based on combining the approaches that we have discussed previously or defining novel strategies. Furthermore, there are tools that are especially designed for the assessment of the sequence assembly quality. However, the generation of assessment tools that consider the complexity of the data sets being assembled, the assembly algorithms, different parameter settings, and the nature of sequencing experiments are still lacking [21, 26]. It is also important to note that there is always a tradeoff between the different quality measures. For instance, trying to maximize the value of one measure (i.e., improve contigs/scaffolds connectivity) may decrease the value of another (i.e., contigs/scaffolds accuracy). Here, we will mention some studies that attempted to design assessment approaches and metrics that are applicable to wide range of next-generation sequence assembly techniques. Then, we will review the available assembly assessment tools.

10.5.1 Evaluation Studies for Assessing Assembly Quality

Assemblathon [27] is one of the studies that defined its own statistical metrics in addition to existing ones. It uses the haplotype sequences as reference measures to newly defined metrics such as NG_{50}, which is the same as N_{50} but uses an average length of haplotype sequences instead of contig/scaffold lengths during its computation. Similarly, CPNG$_{50}$/SPNG$_{50}$ denotes the average length of contigs/scaffolds consistent with the haplotype sequences, while CC50 measures the connectivity between any two randomly chosen points in the assembled genomes. The recently published version of the Assemblathon [28] addressed some practical issues during assembly evaluation, including the consideration of diverse assembly results from various assemblers with different parameter settings, the choice of assemblers based on metrics of interest and overlooking contiguity metrics when studying the genetic components of the assembled sequences.

E-size is yet another statistical metric introduced in GAGE [13]. E-size measures the expectation that a certain point (or base), which is chosen randomly from a reference genome, is located in the assembled contigs/scaffolds in terms of their lengths. Additionally, GAGE also discussed the different factors that can affect the evaluation process, such as the complexity of the genome being assembled and the employed assembler. It also reported that various statistical measures cannot be used alone in indicating assembly quality due to inefficiencies in representing the contiguity and accuracy of the assembled sequences. A more recent version of this study is called GAGE-B [29]. GAGE-B evaluated different bacterial genome assemblers using libraries with high coverage reads and studied the effect of the coverage and read lengths on the assembly quality.

Additionally, Haiminen et al. [30] reported that the assessment process can be affected by the nature of sequencing experiments, such as the average length of short reads, their coverage, and the rate of sequencing errors. Furthermore, they give a different score for each mis-call base according to diverse-modified operations, such as substitutions, insertions, deletions, reordering, redundancy, and relocations. The accuracy of the assembled sequence is determined by gathering these scoring values.

10.5.2 Assembly Assessment Tools

QUAST [31] is an assessment tool that uses a combination of metrics which consider the presence or absence of the reference genomes. It uses N_{50}, NG$_{50}$, NA$_{50}$, and NGA$_{50}$ in measuring the assembly quality in terms of aligned blocks rather than aligned contigs/scaffolds. QUAST also combines other discussed metrics such as the total number of misassembled contigs/scaffolds and genetic components. Moreover, it provides a full set of functionality to generate different statistical reports supplemented with plots and figures.

Computing Genome Assembly Likelihoods (CGAL) [32] introduced the likelihood metric during de novo assembly evaluation based on the uniformity of the read coverage, errors in the sequenced reads, the distribution of insert sizes, and the size of unassembled reads.

REAPR [33] is another reference-free assessment tool that identifies errors in the assembled sequences using paired-end reads and provides useful information to the end users that reflects the quality of the algorithm used in the assembly process.

10.6 Assessment of Transcriptome and Metagenomes Assembly Quality

The assessment of assembled transcripts also represents a challenge in the next-generation environment since it relies on the abundance of reference transcripts, its length, its different splicing isoforms, and the existence of novel transcripts. Martin and Wang proposed different metrics for assessing transcriptome assembly at different levels of complexity in the context of the abundance of reference transcripts that are well expressed and originate from the same transcriptome sequences [34]. These metrics include accuracy, completeness, contiguity, chimerism, and variant resolution. Although these metrics measure the assembled transcripts according to a set of reference transcripts, they provide useful insight regarding the correct number of assembled bases, the percentage of coverage with respect to reference transcripts, the number of chimeric transcripts that are introduced during the assembly process, and the percentage of resulting variations in the assembled transcripts [34, 35]. If the reference transcripts are not available, other complementary approaches may be utilized instead. This includes examining the encoding of full-length ORFs in different isoforms and performing subsequent validation through the use of proteomic assays [36].

The evaluation of metagenomic sequence assemblies is another formidable challenge in the next-generation sequencing environment due to the presence of a variety of genetic materials from different microbial communities. Mende and colleagues [37] proposed a number of metrics for evaluative purposes, including the number of chimeric contigs, the accuracy of contigs based on their defined scoring scheme, and the variety of genetic components in the resulting assembly sequences.

Charuvake and Rangwala [38] presented the entropy metric to measure the degree of chimerism in contig sequences. Furthermore, they exploited the paired-end reads and sequence coverage to measure the assembly quality. Recently, Assembly Likelihood Evaluation (ALE) [39] announced a reference independent framework for assessing metagenomic and single-cell assemblies. ALE utilizes statistical methods that rely on different informational sources such as paired-end constraints and relevant factors during sequencing experiments (i.e., coverage, errors, and length). In addition, it reports various assembly errors such as base-call errors, misassembled chimeric sequences, as well as genome rearrangements that are a result of indel operations.

References

1. Church DM, Goodstadt L, Hillier LW, Zody MC, Goldstein S et al. (2009) Lineage-specific biology revealed by a finished genome assembly of the mouse. PLoS Biol 7 (5):e1000112. doi:10.1371/journal.pbio.1000112
2. Colbourne JK, Pfrender ME, Gilbert D, Thomas WK, Tucker A et al. (2011) The ecoresponsive genome of Daphnia pulex. Science 331 (6017):555-561. doi:10.1126/science.1197761
3. Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB et al. (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438 (7069):803-819. doi:nature04338

4. Locke DP, Hillier LW, Warren WC, Worley KC, Nazareth LV et al. (2011) Comparative and demographic analysis of orang-utan genomes. Nature 469 (7331):529-533. doi:10.1038/nature09687
5. Zhang W, Chen J, Yang Y, Tang Y, Shang J et al. (2011) A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies. PLoS One 6 (3):e17915. doi:10.1371/journal.pone.0017915
6. Alkan C, Sajjadian S, Eichler EE (2011) Limitations of next-generation genome sequence assembly. Nat Methods 8 (1):61-65. doi:10.1038/nmeth.1527
7. Golovko G, Khanipov K, Rojas M, Martinez-Alcantara A, Howard JJ et al. (2012) Slim-Filter: an interactive windows-based application for illumina genome analyzer data assessment and manipulation. BMC bioinformatics 13:166. doi:10.1186/1471-2105-13-166
8. Powell DR, Seemann T (2013) VAGUE: a graphical user interface for the Velvet assembler. Bioinformatics 29 (2):264-265. doi:10.1093/bioinformatics/bts664
9. Li R, Fan W, Tian G, Zhu H, He L et al. (2010) The sequence and de novo assembly of the giant panda genome. Nature 463 (7279):311-317. doi:10.1038/nature08696
10. Lin Y, Li J, Shen H, Zhang L, Papasian CJ et al. (2011) Comparative studies of de novo assembly tools for next-generation sequencing technologies. Bioinformatics 27 (15):2031-2037. doi:10.1093/bioinformatics/btr319
11. Liu Y, Qin X, Song XZ, Jiang H, Shen Y et al. (2009) Bos taurus genome assembly. BMC genomics 10:180. doi:10.1186/1471-2164-10-180
12. Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A et al. (2008) The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus). Nature 452 (7190):991-996. doi:10.1038/nature06856
13. Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T et al. (2012) GAGE: A critical evaluation of genome assemblies and assembly algorithms. Genome research 22 (3):557-567. doi:10.1101/gr.131383.111
14. Huson DH, Halpern AL, Lai Z, Myers EW, Reinert K et al. Comparing Assemblies Using Fragments and Mate-Pairs. In: WABI ‘01 Proceedings of the First International Workshop on Algorithms in Bioinformatics Århus, Denmark, 2001. Springer Berlin Heidelberg, pp 294-306
15. Narzisi G, Mishra B (2011) Comparing de novo genome assembly: the long and short of it. PLoS One 6 (4):e19175. doi:10.1371/journal.pone.0019175
16. Vezzi F, Narzisi G, Mishra B (2012) Feature-by-feature—evaluating de novo sequence assembly. PLoS One 7 (2):e31002. doi:10.1371/journal.pone.0031002
17. Phillippy AM, Schatz MC, Pop M (2008) Genome assembly forensics: finding the elusive misassembly. Genome Biol 9 (3):R55. doi:10.1186/gb-2008-9-3-r55
18. Zhou S, Bechner MC, Place M, Churas CP, Pape L et al. (2007) Validation of rice genome sequence by optical mapping. BMC genomics 8:278. doi:1471-2164-8-278
19. 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 research 20 (2):265-272. doi:10.1101/gr.097261.109
20. Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN et al. (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proceedings of the National Academy of Sciences of the United States of America 108 (4):1513-1518. doi:10.1073/pnas.1017351108
21. Meader S, Hillier LW, Locke D, Ponting CP, Lunter G (2010) Genome assembly quality: assessment and improvement using the neutral indel model. Genome research 20 (5):675-684. doi:10.1101/gr.096966.109
22. Parra G, Bradnam K, Ning Z, Keane T, Korf I (2009) Assessing the gene space in draft genomes. Nucleic acids research 37 (1):289-297. doi:10.1093/nar/gkn916
23. Hubisz MJ, Lin MF, Kellis M, Siepel A (2011) Error and error mitigation in low-coverage genome assemblies. PLoS One 6 (2):e17034. doi:10.1371/journal.pone.0017034
24. Li H (2012) Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly. Bioinformatics 28 (14):1838-1844. doi:10.1093/bioinformatics/bts280

25. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ et al. (2009) ABySS: a parallel assembler for short read sequence data. Genome research 19 (6):1117-1123. doi:10.1101/gr.089532.108
26. Salzberg SL, Yorke JA (2005) Beware of mis-assembled genomes. Bioinformatics 21 (24):4320-4321. doi: 10.1093/bioinformatics/bti769
27. Earl D, Bradnam K, St John J, Darling A, Lin D et al. (2011) Assemblathon 1: a competitive assessment of de novo short read assembly methods. Genome research 21 (12):2224-2241. doi:10.1101/gr.126599.111
28. Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M et al. (2013) Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. Gigascience 2 (1):10. doi:2047-217X-2-10
29. Magoc T, Pabinger S, Canzar S, Liu XY, Su Q et al. (2013) GAGE-B: an evaluation of genome assemblers for bacterial organisms. Bioinformatics 29 (14):1718-1725. doi:10.1093/bioinformatics/btt273
30. Haiminen N, Kuhn DN, Parida L, Rigoutsos I (2011) Evaluation of methods for de novo genome assembly from high-throughput sequencing reads reveals dependencies that affect the quality of the results. PLoS One 6 (9):e24182. doi:10.1371/journal.pone.0024182
31. Gurevich A, Saveliev V, Vyahhi N, Tesler G (2013) QUAST: quality assessment tool for genome assemblies. Bioinformatics 29 (8):1072-1075. doi:10.1093/bioinformatics/btt086
32. Rahman A, Pachter L (2013) CGAL: computing genome assembly likelihoods. Genome Biol 14 (1). doi: 10.1186/Gb-2013-14-1-R8
33. Hunt M, Kikuchi T, Sanders M, Newbold C, Berriman M et al. (2013) REAPR: a universal tool for genome assembly evaluation. Genome Biol 14 (5):R47. doi:gb-2013-14-5-r47
34. Martin JA, Wang Z (2011) Next-generation transcriptome assembly. Nature reviews Genetics 12 (10):671-682. doi:10.1038/nrg3068
35. Martin J, Bruno VM, Fang Z, Meng X, Blow M et al. (2010) Rnnotator: an automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads. BMC genomics 11:663. doi:10.1186/1471-2164-11-663
36. Adamidi C, Wang Y, Gruen D, Mastrobuoni G, You X et al. (2011) De novo assembly and validation of planaria transcriptome by massive parallel sequencing and shotgun proteomics. Genome research 21 (7):1193-1200. doi:10.1101/gr.113779.110
37. Mende DR, Waller AS, Sunagawa S, Jarvelin AI, Chan MM et al. (2012) Assessment of metagenomic assembly using simulated next generation sequencing data. PLoS One 7 (2):e31386. doi:10.1371/journal.pone.0031386
38. Charuvaka A, Rangwala H (2011) Evaluation of short read metagenomic assembly. BMC genomics 12 Suppl 2:S8. doi:10.1186/1471-2164-12-S2-S8
39. Clark SC, Egan R, Frazier PI, Wang Z (2013) ALE: a generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies. Bioinformatics 29 (4):435-443. doi:10.1093/bioinformatics/bts723