Tools for sequence assembly and annotation

Chapter · February 2018

CITATIONS
2

READS
1,624

2 authors, including:

Georrge John. J
Christ College, Rajkot
70 PUBLICATIONS **296** CITATIONS

SEE PROFILE

TOOLS FOR SEQUENCE ASSEMBLY AND ANNOTATION

Mishal John and **John J Georrge***

Department of Bioinformatics, Christ College, Rajkot, Gujarat
*E-mail ID: mishaljohnvincent@gmail.com, johnjgeorrge@gmail.com**

ABSTRACT: Due to the development of sequencing techniques, several sequences are being sequenced rapidly. But are mostly being left as it is, either not knowing what can be done further, or not knowing how to analyze, etc. In such circumstances, a guidance to go further can be of great help. Sequence assembly is one way of utilizing the sequence, where it merges the fragmented sequences to form a complete genome. The obtained sequence is further annotated to find the gene locations and all the coding regions in that particular sequence. The sequence is compared for a thorough study of the sequence, so as to find similarities and uses of the sequences, finding the gene responsible for a disease, protein family and domain identification, etc. Now that the possible ways of dealing with a sequence is known, it is also important to choose an appropriate tool. A software analysis of Genome assembly and annotation is specified in this chapter, which might be helpful for the researchers working in this field.

Keywords: Assembly, Annotation

1. INTRODUCTION

The entire set of genetic information present in the genetic material of an organism (DNA, RNA), including all the coding and non-coding regions is none as the Genome of that organism[1]. The human genome is made up of approximately 3 billion nucleotides. The DNA is composed mainly of four components (nucleotides)i.e, Adenine, Thymine, Guanine and Cytosine (A, T, G, and C). Determining the order in which these components are arranged is known as Genome/DNA sequencing. Various techniques starting from the like Sanger sequencing, leading to the introduction of 454 pyrosequencing, followed by Solexa, SOLiD and Helicos, and also the development of Nanopore technology are used for this purpose[2].Sequences are sequenced independently to a particular length, as per the technology used. Certain computer algorithms are helpful in aligning and merging these fragments (sequence reads) to form contigs (longer continuous stretches) in the process of *de novo assembly*[3]. Marking specific features of the Genome sequence, briefing out information about its structure and function, is a highly significant process which is known as *Annotation*. Structural annotation identifies structural elements or segments in the genome using only the characteristics of the sequence relying on its pattern recognition[4]. Collecting information about genes and recounting their biological identity, molecular function, biological role, subcellular location and their expression domains within the organism is done by Functional annotation[5].An assembly with annotation, in the past was considered as *build*[6].The assessment of similarities and differences (DNA sequence, genes, gene order, regulatory sequences) between the genomes of diverse organisms reveals the relationship between the individuals[7]. Researchers use various computational tools in order carefully compare the characteristics that define various organisms, thus pinpointing the regions of similarity and differences. Comparative genomics distinguishes conserved regions from divergent and functional from non-functional DNA, and also, contributes to the identification of the general functional class of particular DNA segments, like coding regions and non-coding regions and some gene regulatory regions[8-10].

2. *DE NOVO* ASSEMBLY

As it is not easy or possible for the sequencing techniques to sequence a whole genome continuously, the genome is broken down into several fragments in order to make it simpler to sequence. But after sequencing, these have to be put back to make the complete original sequence. Thus there is a need to align and merge these fragments, for which the process called “Assembly” is used[11]. In other words, the reconstruction of the unknown contiguous DNA sequence correctly by inferring it with the help of a number of fragments is said to be Assembly[12].Adapter trimming, quality filtering, error correction, creation of contigs, and verification of contigs by mapping reads to the assembly and the creation/verification of scaffolds are the basic steps involved[13].It is already evident that sequencing is a highly delicate work. Similarly, it is even harder to assemble. The chief difficulty in assembling is the *genomic repeats*. The struggle of assembly relies on the number of reads that are being assembled. Computational algorithms are being developed to master over such issues, but still the assembly is not close to be the complete solution. In biological research, assembly softwaremust familiarize to the recent applications of the DNA sequencing. There are certain challenges by the sequencing technology as well that impact the assembly which are mentioned below[2].

- The presence of short reads and the absence of mate-pair cause difficulty in assembling the repeats.
- Rising of new types of errors which demands for modification of the existing software and incorporate technology specific features in assembly software.
- Repetitive property of DNA, leading to the fault-tolerant and alternative seeking algorithms

- Huge amount of data leads to the difficulty in the efficiency and requirement of parallel implementations or specialized hardware when practiced in large genomes.

None of the assembly approaches available now can reconstruct a genome completely from read data alone.

2.1. Features of Genome Assemblers

None of the assembly approaches available now can reconstruct a genome completely from read data alone. As there are more than 200 tools available for Assembling, the top ten Assemblers which are widely being used are described below table 1.

SL. NO.	NAME	FEATURES	ADVANTAGES	LIMITATIONS
1	Mira	Allows hybrid assemblies of Sanger, 454, Solexa, IonTorrent and PacBio (CCS &ecCLR) data Can use paired-end and / or unpaired data Supports ancillary data in TRACEINFO format (from NCBI) Marks places of interest with tags so that these can be found quickly in finishing programs Has an SNP analysis pipeline for sequencing data of viruses and prokaryotes. Available at http://www.chevreux.org/mira_downloads.html	Contains a comparable combination of algorithms For both Genomic and Transcript data[14] Works very well on bacterial genomes Produces large contigs Produces contigs containing reads	May not be advisable for large genomes. Use of preprocessed data incorrectly leads to case-error probabilities and functions to detect the resolve possible misassemblies[14].
2	RS_HGAP_Assembler.3 Hierarchical Genome Assembly Process	Three steps by first preassembling reads, assembling the preassembled reads using Celera Assembler and finally polish using Quiver. Can support up to 100 Mb from SMRT Portal	Incorporates a 10-fold speed improvement for microbial assembly. Consists of pre-assembly, *de novo* assembly with PacBio'sAssembleUnitig, and assembly polishing with Quiver[15]. Does not require accurate raw reads to correct errors.	Mapping and trimming parameters might need to be optimized.
3	A5 miseq Pipeline	Has a 5-step procedure Substantially revises steps As compared to A5 pipeline, instead of discarding reads in the initial step, only the contaminated portion of the read gets trimmed. In many instances, A5-miseq assemblies have had higher NGA50 values, fewer misassemblies and fewer base-calling errors than A5 pipeline[13].	Automated adapter trimming, more full-length genes assembled, NCBI-ready outputs and production of base-call quality scores A5-miseq should be particularly useful for researchers with limited bioinformatics experience or computing resources	Following its publication, assembly pipelines might be inadvertently tuned to produce high scores specifically on that dataset. This could result in artificially high scores that do not accurately reflect the expected performance on other datasets[13].
4	Assembly By Short Sequencing (ABySS)	Pioneer of the Representation of a de Bruijn graph[16]. Highly contiguous genome assemblies f long reads were obtained from human and other organisms whose contiguity ranges in megabase[17]. Could assemble 3.5 billion paired-end reads. ~2.76 million contigs (>100 bps) were created, which	Assembles very large datasets produces by sequencing human genome. Able to parallelize the assembly of billions of short reads over a cluster of commodity hardware[16]	To correct assembly assessing breakpoint metric has to be manually done. Contig length range is still limited to tens of kb which is shorter than the megabases obtained by the long-read sequences[17].

		represented the reference human genome by 68%[16].		
5	ALL-PATHS	2 concepts – Finding all paths across given read pair And localization Presented as a graph retaining ambiguity Available at ftp://ftp.broadinstitute.org/pub/crd/ALL-PATHS/Release-LG/	All the assemblies are highly complete and contiguous. Coverage >96% in all cases good accuracy, short-range contiguity, long-range connectivity, and coverage of the genome[18].	Terminate at worst case, so to fail to return any terminations at all. Leading to holes in the final assembly, thus prone to error[19].
6	CAP 3	Uses forward – reverse constraints to correct errors and join contigs In construction of consensus sequences, CAP3 uses redundant coverage Freely available at huang@mtu.edu.	Often produces less errors as compared with PHRAP Scaffold construction is easier than PHRAP Poor regions and false overlaps are identified and removed	With the use of forward – reverse constraint as some of them are not correct due to the lane tracking and cloning errors[20]
7	Ray Meta	Assembly based on distributed MPI Longer contigs as compared with many others like velvet Accurate in assembling and profiling a 3 billion red metagenomic experiment on bacterial genome, in 15 hours with 1,024 processor cores, by using only 1.5GB per core[21]. Open source available on http://denovoassembler.sf.net.	Produces longer contigs and more bases Widespread application Provide precise taxonomic profusions Can run on multiple computers[22] Can be written in C++ and can run in parallel on numerous interconnected computers.	Loss of information due to the construction of graph Produce excessive number of misassemblies[23] Processing of large and complex datasets can be facilitated by the software.
8	Meta Velvet-SL	Uses supervised machine learning to improve performance Classifies every node from the graph For very short read (25-50 bp) datasets and high coverage, it is highly preferred[24]. https://www.ebi.ac.uk/~zerbino/velvet/s	Generate assemblies with higher N50 scores and higher quality[25]. Fast Widely used	Velvet scaffolding is error-prone Exclusive access to a computer with large amount of available memory for a single MetaVelvet assembly is required (minimum 128 GB)[22]. Produces only consensus sequences May not work as well on gappy error models (e.g. 454, IonTorrent, PacBio)
9	TruSPAdes TSLA (TruSeq Synthetic Long Reads)	Long and accurate virtual reads are generated from an assembly of barcoded pools of short reads. Instead of the entire metagenome, it inherits the repeat structure of a TSLR barcode from an individual 6.2Gb, 28M reads, 2x100bp, insert size ~ 215bp (stdE.coli isolate) 6.3 Gb, 29M reads, 2x100bp, insert size ~ 270bp (E.coli single cell)[26].	Produces long and accurate contigs Works with many data types (e.g. NanoporeMinION, PacBio) Supports RNA and metagenome data [23].	Multiple input libraries are not supported[27]. Produces only consensus sequences Doesn't work with low coverage Not designed for large genomes

		Can be downloaded from http://cab.spbu.ru/soft-ware/spades/.		
10	PHRAP	Uses a banded version of Smith-Waterman-Gotoh algorithm to compare the input sequences pairwise. PHRED quality score information is used to create accurate and high quality contig sequences Cloning vectors which would interfere with the read alignments in the assembly are masked by Cross match/Swat algorithm	If the sequence on both the strands are covered or not, if the reads are sequenced by more than one chemistry and the quality values of the base in each read, factors like this contribute the calculation of the PHRAP score Able to balance between the discrepancies and preventing of stacking repeat sequences[28].	More error rate as compared with CAP3[20].

Table. 1. Properties of Widely used Assemblers in brief

Given below certain other tools for genome assembly-

1. TriMetAss 1.2
   The Trinity-based Iterative Metagenomics Assembler. Only select regions surrounding interesting features in metagenomic data can be assembled. Used in very common and well-conserved genes[29].
2. OMWare 1.0
   Efficient Assembly of Genome-wide Physical Map. Aims to help scientists in using optical map data. As they exist in a range of formats, it summarizes the optical maps with their most common manipulations[30].
3. LightAssembler
   Lightweight Resources Assembly Algorithm. For high-throughput assembly reads[31].
4. QUAST 4.1
   Quality Assessment Tool for Genome Assemblies. Can work irrespective of the presence of reference genome[32].
5. DNA Baser 4.36
   DNA Sequence Assembly & Analysis. Revolutionary in automatic DNA sequence assembly, DNA sequence analysis, contig editing, file format conversion and mutation detection[33, 34].
6. COCACOLA
   A general outline of Binning MetagenomicContigs using Sequence COmposition, Read CoverAge, CO-alignment, and Paired-end Read LinkAge. Is able to construct species from highly complicated environmental samples besides handling strain-level variations[35].
7. MaxBin 2.2
   Binning Assembled Metagenomic Sequences. Bins assembled metagenomic sequences based on Expectation-Maximization algorithm[36].
8. GAML 0.1
   Genome Assembly by Maximum Likelihood. A prototype genome assembly tools based on maximum likelihood of the assembly. Covers error rate, insert length and other features of individual sequencing technologies[37, 38].
9. NanoMark
   DNA Assembly Benchmark for Nanopore long reads. Based on third generation sequencing[39].
10. ARC 1.1.4-beta
    Assembly by Reduced Complexity. Pipeline that facilitates iterative, reference guided de novo assemblies. Capable of breaking large, complex problems in to smaller manageable chunks. [http://ibest.github.io/ARC/]
11. TransPS 1.1.0
    Transcriptome Post Scaffolding. Pipeline to post-process the pre-assembled transcriptomes with the help of reference-based method. An align-layout-consensus consisting of 3 major stages is applied[40].
12. assemblyManager
    Computing the Robotic Commands for 2ab Assembly[41].
13. BinPacker 1.1
    Packing-Based De Novo Transcriptome Assembly from RNA-seq Data. Novel de novo assembler. Transcriptome assembly problem is modeled as tracking a set of trajectories representing coverage based on sizes of their corresponding isoforms and solves a series of bib-packing problems[42].

14. FermiKit 0.13
    De novo Assembly based Variant Calling pipeline for Illumina Short Reads. Variant calling pipeline for deep Illuminaresequencing data based on de novo assembly. The assembly retains long deletion, novel sequence insertions, translocations and copy number besides encoding SNPs and short INDELs. It is also considered a better long insertion caller[43].
15. REPdenovo
    A tool to Construct Repeats directly from Raw Reads. Designed to construct repeats directly from the sequence reads. Provides much functionality and can generate much longer repeats. Its main functionalities are Assembly and Scaffolding[44].
16. Xander
    Gene-targeted Metagenomic Assembler. Novel method to target assembly of specific protein-coding genes with the help of a graph structure with a combination of Bruijn graph and HMMs[45].
17. SWAP-Assembler 2
    A scalable and fully parallelized Genome Assembler. Intended for massive sequencing data. A multi-step bi-directed graph is adopted with which the standard genome assembly becomes equivalent to the edge merging operations in a semi-group[46].
18. TGNet
    Visualization and Quality Assessment of de novo Genome Assemblies. A visualization and quality assessment of de novo genome assemblies based on a Cytoscape. It is capable of detecting inconsistencies between a genome assembly and an independently derived transcriptome assembly[47].
19. misFinder v0.4.05.05
    Identify Mis-assemblies in an unbiased manner using Reference and Paired-end Reads. Destines to identify the assembly errors with high accuracy in an unbiased way. Their mis-assembled positions are corrected to improve the assembly accuracy for downstream analysis[48].
20. Scaffold_builder v2.2
    Contigs generated by draft sequencing along a reference sequence are ordered by this software. N's help in filling the gaps and Muscle help in aligning he small overlaps. It is possible not only to assemble but also annotate the genomes[49].
21. Rnnotator 3.5.0
    Pipeline which generates models by de novo assembly. Full-length transcripts are reconstructed in the absence of a complete reference genome.producese highly accurate contigs[50].
22. SATRAP 0.2
    SOLiD Assembler TRAnslation Program. It adopts an efficient strategy to translate into bases the color space assembly. It can also be used as a stand –alone program so to perform color space translation[51].
23. Bandage v0.7.1
    Main purpose is to visualize de novo assembly graph. It opens up new possibilities for analyzing de novo assemblies by displaying connections which are not present in the contig file[52].
24. HapCol
    Haplotype Assembly from Long Gapless Reads. For each single nucleotide polymorphism position it is exponential in the maximum number of corrections, hence reducing the overall error-correction score. Fast and efficient in memory for haplotypes from long reads that are gapless[53].
25. REAGO 1.1
    REconstruct 16S ribosomal RNA Genes from MetagenOmic data. Approaches the challenges by with a combination of secondary structure – aware homology search, properties of zRNA genes and de novo assembly[54].
26. FGAP 1.8.1
    Automated Gap Closing tool
    It merges alternative assemblies or incorporates alternative data in order to improve the genome sequences, analyses the gap region to indicate the best sequence to close the gap[55].
27. DETONATE 1.10
    DE novo TranscriptOmerNa-seq Assembly with or without the Truth Evaluation. consists of two component packages, RSEM-EVAL and REF-EVAL. Both packages are mainly intended to be used to evaluate de novo transcriptome assemblies, although REF-EVAL can be used to compare sets of any kinds of genomic sequences[56].
28. Trinity 2.1.1
    Trinity represents a novel method for the efficient and robust de novo reconstruction of transcriptomes from RNA-Seq data. Trinity combines three independent software modules: Inchworm, Chrysalis, and Butterfly, applied sequentially to process large volumes of RNA-Seq reads. Trinity partitions the sequence data into many individual de Bruijn graphs, each representing the transcriptional complexity at at a given gene or locus, and then processes each graph independently to extract full-length splicing isoforms and to tease apart transcripts derived from paralogous genes[57].

29. IsoSCM 2.0.11
    IsoSCM (Isoform Structural Change Model) is a new method for transcript assembly that incorporates change-point analysis to improve the 3′ UTR annotation process[58].
30. IVA 1.0.3 – Iterative Virus Assembler
    IVA is a de novo assembler designed to assemble virus genomes that have no repeat sequences, using Illumina read pairs sequenced from mixed populations at extremely high and variable depth[59].

3. GENOME ANNOTATION

Gene annotation is been widely misunderstood as gene prediction. But in reality, a gene prediction is a prediction of the intron–exon structure of a gene based on a mathematical model, while gene annotation is the synthesis of intron–exon structure from multiple lines of evidence including gene prediction, expression data (often in the form of mRNA-seq data), protein homology, and repetitive elements[60] involving two steps i.e., evidence generation and synthesis[61].Evidence generation combines repeat masking, transcript and protein alignments, gene predictions, and whole genome alignment of closely related species are commonly used to provide evidence in genome annotation[62]. Once the evidence is generated, the annotator begins its next step of synthesizing the information into gene annotations[63].

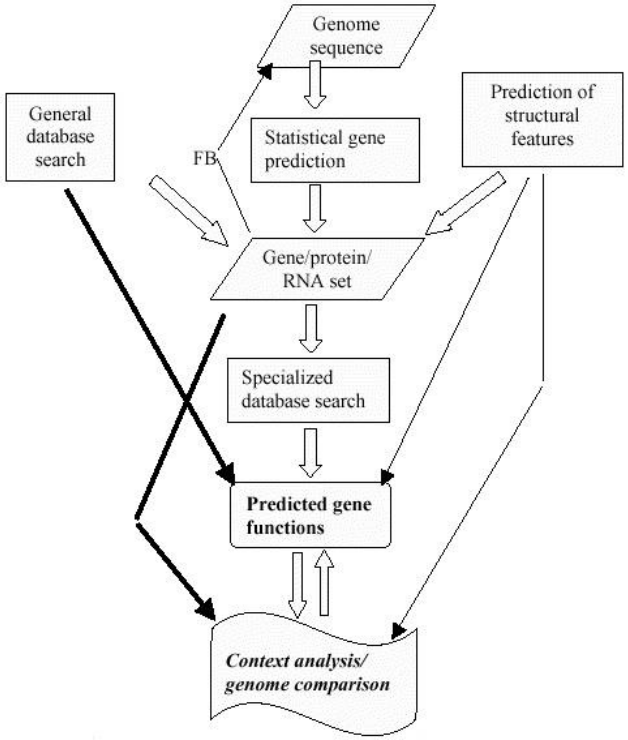

Figure 1: generalized flow chart of genome annotation[64].

Table2: Certain widely used tools for genome Annotation is listed.

Sl. No.	Name	Description	Advantages	Limitation
1	Blast2GO	A universal tool for annotation, visualization and analysis in functional genomics research	Accuracy is 65–70% It has successful in extracting relevant functional features of the sequences based on theuse of the predicted annotation[65].	Accuracy Difficulty in analysis of poorly characterized organisms
2	Rast	Rapid Annotations using Subsystems Technology.	Able to identify protein - encoding. rRNA and tRNA genes. Assigns functions to the genes Predicts which subsystems are represented in the genome. Reconstructs the metabolic network[66].	Time consuming Accuracy not satisfactory.

3	KAAS server	An automatic genome annotation and pathwayreconstruction server.	Gives three views in output Gives the pathways involved[67].	Time and accuracy
4	BASYs	a web server for automated bacterial genome annotation	Compares favorably with other automated systems. Permits high throughput, detailed and fully automated annotation of bacterial genomes[68].	Accuracy does not exceed 60%

Annotation tools rarely being used

Sl No	Description
1.	Prokka Rapid prokaryotic genome annotation Decreases running time in multi-core computers. Requires certain features like High multicore computers, multiple single CPU threads, etc.,
2.	GenDB Open source genome annotation system for prokaryote genomes Its flexible and extensible
3.	SVM Support Vector Machine. One advantage of this is its scalability.
4.	eggNO-Mapper Fast Genome-Wide Functional Annotation through Orthology Assignment. Based on fast Orthology mapping.
5.	CDART A protein homology by based ob Domain Architecture.
6.	FTG A web-server for analyzing nucleotide sequences to predict the genes using Fourier transform techniques.
7.	EGPred Prediction of Eukaryotic Genes Using Ab Initio Methods After Combining with Sequence Similarity Approaches.
8.	SOBA Sequence Ontology BI analysis. and includes both a bug tracker and a feature request tracker for continued development and maintenance of the tool.
9.	VIGOR An annotation program for small viral genomes. User friendly and is used for five different virus gene prediction.
10.	FLAN Short for FLu ANnotation. A web server for influenza virus genome annotation.

4. CONCLUSION

Although there are many tools for both Assembly and Annotation, based on the three-parameter analysis, out of the tools available for Assembly, the best can be Mira and RS_HGAP, and in the case of Annotation, it can be Blast2GO, K AAS server and Rast server.

5. REFERENCES

1. Brosius, J., *The fragmented gene.* Annals of the New York Academy of Sciences, 2009. **1178**(1): p. 186-193.
2. Pop, M., *Genome assembly reborn: recent computational challenges.* Briefings in bioinformatics, 2009. **10**(4): p. 354-366.
3. Ekblom, R. and J.B. Wolf, *A field guide to whole-genome sequencing, assembly and annotation.* Evolutionary applications, 2014. **7**(9): p. 1026-1042.
4. Hudaiberdiev, S., *Computational analysis of quorum sensing systems in bacterial genomes: developing automated annotiation tools*. 2014, Univerza v Novi Gorici, Fakulteta za podiplomski študij.
5. Berardini, T.Z., et al., *Functional annotation of the Arabidopsis genome using controlled vocabularies.*

Plant physiology, 2004. **135**(2): p. 745-755.

6. Kitts, P., *Genome assembly and annotation process.* McEntyre J, Ostell Jeditors. The NCBI Handbook. Bethesda: National Center for Biotechnology Information, 2002.
7. Touchman, J., *Comparative Genomics.* Nature Education Knowledge, 2010. **3**(10): p. 13.
8. Miller, W., et al., *Comparative genomics.* Annu. Rev. Genomics Hum. Genet., 2004. **5**: p. 15-56.
9. Atman Vaidya, V.S.N., John J. Georrge, Singh S. P, *Comparative Analysis of Thermophilic Proteases.* Research Journal of Life Sciences, Bioinformatics, Pharmaceutical and Chemical Sciences, 2018. **04**(06): p. 65-91.
10. Varun S. Nair, J.J.G. *Earthworm's Genomics and Toxicogenomics*. in *Proceedings of International Science Symposium on Recent Trends in Science and Technology (ISBN: 9788193347553)*. 2017. Bharti Publications, New Delhi.
11. El-Metwally, S., et al., *Next-generation sequence assembly: four stages of data processing and computational challenges.* PLoS computational biology, 2013. **9**(12): p. e1003345.
12. Chevreux, B., T. Wetter, and S. Suhai. *Genome sequence assembly using trace signals and additional sequence information*. in *German conference on bioinformatics*. 1999. Citeseer.
13. Coil, D., G. Jospin, and A.E. Darling, *A5-miseq: an updated pipeline to assemble microbial genomes from Illumina MiSeq data.* Bioinformatics, 2014. **31**(4): p. 587-589.
14. Chevreux, B., et al., *Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs.* Genome research, 2004. **14**(6): p. 1147-1159.
15. Chin, C.-S., et al., *Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data.* Nature Methods, 2013. **10**: p. 563.
16. Simpson, J.T., et al., *ABySS: a parallel assembler for short read sequence data.* Genome research, 2009. **19**(6): p. 1117-1123.
17. Jackman, S.D., et al., *ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter.* Genome research, 2017. **27**(5): p. 768-777.
18. Gnerre, S., et al., *High-quality draft assemblies of mammalian genomes from massively parallel sequence data.* Proceedings of the National Academy of Sciences, 2011. **108**(4): p. 1513-1518.
19. Butler, J., et al., *ALLPATHS: de novo assembly of whole-genome shotgun microreads.* Genome research, 2008. **18**(5): p. 810-820.
20. Huang, X. and A. Madan, *CAP3: A DNA sequence assembly program.* Genome research, 1999. **9**(9): p. 868-877.
21. Liu, L., et al., *Comparison of next-generation sequencing systems.* BioMed Research International, 2012. **2012**.
22. Boisvert, S., et al., *Ray Meta: scalable de novo metagenome assembly and profiling.* Genome biology, 2012. **13**(12): p. R122.
23. Bankevich, A. and P.A. Pevzner, *TruSPAdes: barcode assembly of TruSeq synthetic long reads.* Nature methods, 2016. **13**(3): p. 248.
24. Zerbino, D.R. and E. Birney, *Velvet: algorithms for de novo short read assembly using de Bruijn graphs.* Genome research, 2008. **18**(5): p. 821-829.
25. Sato, K. and Y. Sakakibara, *MetaVelvet-SL: an extension of the Velvet assembler to a de novo metagenomic assembler utilizing supervised learning.* DNA research, 2014. **22**(1): p. 69-77.
26. Nurk, S., et al., *metaSPAdes: a new versatile metagenomic assembler.* Genome research, 2017. **27**(5): p. 824-834.
27. Vollmers, J., S. Wiegand, and A.-K. Kaster, *Comparing and evaluating metagenome assembly tools from a microbiologist's perspective-Not only size matters!* PloS one, 2017. **12**(1): p. e0169662.
28. de la Bastide, M. and W.R. McCombie, *Assembling genomic DNA sequences with PHRAP.* Current Protocols in Bioinformatics, 2007: p. 11.4. 1-11.4. 15.
29. Bengtsson-Palme, J., et al., *Shotgun metagenomics reveals a wide array of antibiotic resistance genes and mobile elements in a polluted lake in India.* Frontiers in Microbiology, 2014. **5**: p. 648.
30. Sharp, A.R. and J.A. Udall, *OMWare: a tool for efficient assembly of genome-wide physical maps.* BMC bioinformatics, 2016. **17**(7): p. 241.
31. El-Metwally, S., M. Zakaria, and T. Hamza, *LightAssembler: Fast and memory-efficient assembly algorithm for high-throughput sequencing reads.* Bioinformatics, 2016. **32**(21): p. 3215-3223.
32. Gurevich, A., et al., *QUAST: quality assessment tool for genome assemblies.* Bioinformatics, 2013. **29**(8): p. 1072-1075.

33. Tarchini, R., et al., *The complete sequence of 340 kb of DNA around the rice Adh1–Adh2 region reveals interrupted colinearity with maize chromosome 4.* The Plant Cell, 2000. **12**(3): p. 381-391.
34. Hiom, K., M. Melek, and M. Gellert, *DNA transposition by the RAG1 and RAG2 proteins: a possible source of oncogenic translocations.* Cell, 1998. **94**(4): p. 463-470.
35. Lu, Y.Y., et al., *COCACOLA: binning metagenomic contigs using sequence COmposition, read CoverAge, CO-alignment and paired-end read LinkAge.* Bioinformatics, 2017. **33**(6): p. 791-798.
36. Wu, Y.-W., et al., *MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm.* Microbiome, 2014. **2**(1): p. 26.
37. Medvedev, P. and M. Brudno, *Maximum likelihood genome assembly.* Journal of computational Biology, 2009. **16**(8): p. 1101-1116.
38. Bož̌a, V., B. Brejová, and T. Vinař, *GAML: genome assembly by maximum likelihood.* Algorithms for Molecular Biology, 2015. **10**(1): p. 18.
39. Sović, I., *Algorithms for de novo genome assembly from third generation sequencing data*. 2016.
40. Liu, M., et al., *A transcriptome post-Scaffolding method for assembling high quality contigs.* Computational biology journal, 2014. **2014**.
41. Leguia, M., et al., *Automated assembly of standard biological parts*, in *Methods in enzymology*. 2011, Elsevier. p. 363-397.
42. Liu, J., et al., *TransComb: genome-guided transcriptome assembly via combing junctions in splicing graphs.* Genome biology, 2016. **17**(1): p. 213.
43. Li, H., *FermiKit: assembly-based variant calling for Illumina resequencing data.* Bioinformatics, 2015. **31**(22): p. 3694-3696.
44. Chu, C., R. Nielsen, and Y. Wu, *REPdenovo: inferring de novo repeat motifs from short sequence reads.* PloS one, 2016. **11**(3): p. e0150719.
45. Wang, Q., et al., *Xander: employing a novel method for efficient gene-targeted metagenomic assembly.* Microbiome, 2015. **3**(1): p. 32.
46. Meng, J., et al., *SWAP-Assembler: scalable and efficient genome assembly towards thousands of cores.* BMC Bioinformatics, 2014. **15**(9): p. S2.
47. Riba-Grognuz, O., et al., *Visualization and quality assessment of de novo genome assemblies.* Bioinformatics, 2011. **27**(24): p. 3425-3426.
48. Zhu, X., et al., *misFinder: identify mis-assemblies in an unbiased manner using reference and paired-end reads.* BMC bioinformatics, 2015. **16**(1): p. 386.
49. Silva, G.G., et al., *Combining de novo and reference-guided assembly with scaffold_builder.* Source code for biology and medicine, 2013. **8**(1): p. 23.
50. Martin, J., et al., *Rnnotator: an automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads.* BMC genomics, 2010. **11**(1): p. 663.
51. Campagna, D., et al., *SATRAP: SOLiD Assembler TRAnslation Program.* PloS one, 2015. **10**(9): p. e0137436.
52. Wick, R.R., et al., *Bandage: interactive visualization of de novo genome assemblies.* Bioinformatics, 2015. **31**(20): p. 3350-3352.
53. Pirola, Y., et al., *HapCol: accurate and memory-efficient haplotype assembly from long reads.* Bioinformatics, 2015. **32**(11): p. 1610-1617.
54. Yuan, C., et al., *Reconstructing 16S rRNA genes in metagenomic data.* Bioinformatics, 2015. **31**(12): p. i35-i43.
55. Piro, V.C., et al., *FGAP: an automated gap closing tool.* BMC research notes, 2014. **7**(1): p. 371.
56. Li, G., et al., *Identification of putative olfactory genes from the oriental fruit moth Grapholita molesta via an antennal transcriptome analysis.* PloS one, 2015. **10**(11): p. e0142193.
57. Davidson, N.M. and A. Oshlack, *Corset: enabling differential gene expression analysis for de novo assembled transcriptomes.* Genome biology, 2014. **15**(7): p. 410.
58. Son, H.G., et al., *RNA surveillance via nonsense-mediated mRNA decay is crucial for longevity in daf-2/insulin/IGF-1 mutant C. elegans.* Nature communications, 2017. **8**: p. 14749.
59. Radja, A., et al., *Pollen Patterns Form from Modulated Phases.* arXiv preprint arXiv:1803.03643, 2018.
60. Korf, I., *Gene finding in novel genomes.* BMC bioinformatics, 2004. **5**(1): p. 59.
61. Holt, C. and M. Yandell, *MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects.* BMC bioinformatics, 2011. **12**(1): p. 491.

62. Campbell, M.S. and M. Yandell, *An Introduction to Genome Annotation.* Current protocols in bioinformatics, 2015: p. 4.1. 1-4.1. 17.
63. Adams, M.D., et al., *The genome sequence of Drosophila melanogaster.* Science, 2000. **287**(5461): p. 2185-2195.
64. Koonin, E.V. and M.Y. Galperin, *Principles and methods of sequence analysis*, in *Sequence—Evolution—Function*. 2003, Springer. p. 111-192.
65. Conesa, A., et al., *Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research.* Bioinformatics, 2005. **21**(18): p. 3674-3676.
66. Aziz, R.K., et al., *The RAST Server: rapid annotations using subsystems technology.* BMC genomics, 2008. **9**(1): p. 75.
67. Moriya, Y., et al., *KAAS: an automatic genome annotation and pathway reconstruction server.* Nucleic acids research, 2007. **35**(suppl_2): p. W182-W185.
68. Van Domselaar, G.H., et al., *BASys: a web server for automated bacterial genome annotation.* Nucleic acids research, 2005. **33**(suppl_2): p. W455-W459.

How to cite this Book Chapter?

APA Style
Mishal John and John J. Georrge (2018). Tools for sequence assembly and annotation. *Proceedings of 10th National Science Symposium on Recent Trends in Science and Technology* (pp. 87-96). ISBN: 9788192952130. Rajkot, Gujarat, India: Christ Publications

MLA Style
Mishal John and John J. Georrge. "Tools for sequence assembly and annotation". *Proceedings of 10th National Science Symposium on Recent Trends in Science and Technology (ISBN: 9788192952130).* Rajkot, Gujarat, India: Christ Publications, 2018. pp. 87-96.

Chicago Style
Mishal John and John J. Georrge. "Tools for sequence assembly and annotation". In *proceedings of 10th National Science Symposium on Recent Trends in Science and Technology (ISBN: 9788192952130*), pp. 87-96. Rajkot, Gujarat, India: Christ Publications, 2018.