wikia2 Documentation

Release latest

February 22, 2017

Elixir wiki

This website is a wiki that aims to provide a platform for bioinformatics knowledge and good practices in genome assembly and annotation.

__TOC__

= Genome assembly =

== Illumina technology == ===Denovo assembly=== *[[Ray]] *[[MaSuRCA]] *[[IDBA]] *[[SOAPdenovo2]] *[[Spades]] *[[Abyss]] *[[Velvet]] *[[Minia]]

===Metagenome assembly=== *[[Ray Meta]] *[[MetaVelvet]] *[[Meta-IDBA]] *[[Omega]]

== Pacbio and nanopore sequencing technology ==

== Long reads assembly ==

Third technology of sequencing brought by Pacbioscience and Oxford nanopore generate average reads length of more than 10,000bp and thus can advantageously be used to improve the genome assembly.

In fact, long reads span more repetitive elements and thus produce more contiguous reconstruction of the genome.

However, long reads have raw error rate of 10% to 15%, requiring a preliminary phase of read correction before the assembly process.

There are two main families of assemblers based on long reads: *Long Reads Only assembler (LRO) *Short and Long Reads assembler (SLR)

LRO Assemblers take only long reads as inputs. SLR Assemblers ask for long reads and short reads.

Some LRO assemblers need corrected long reads as input. Several software to correct long reads, based in two strategies, are available. The first strategy consist in aligning long reads against themselves. The second one need short reads as well as long reads. As the error rate of illumina reads is significantly lower, short reads are used to correct long reads.

=== Denovo assembly ===

In general, the strategy of these assemblers is based on Overlap-Layout-Consensus (OLC) algorithm. It produces alignments between long reads, calculates the best overlap graph and then generates the consensus sequences of contigs from the graph.

The assembly will be more efficient if long reads have a low error rate.

Up to now, 7 denovo assemblers have been listed: *[[Celera Assembler]]<ref>Sergey Koren et al., Hybrid error correction and de novo assembly of single-molecule sequencing reads, Nature Biotechnology 30(7):693-700, 2012</ref> *[[Falcon]]<ref>Falcon Genome Assembly Tool Kit Manual, Jason Chin, https://github.com/PacificBiosciences/FALCON/wiki/Manual, Jan. 2016</ref> *[[Miniasm]]<ref>Heng Li, Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences, arXiv:1512.01801., 2015</ref> *[[Newbler]]<ref>Genome sequencing in microfabricated high-density picolitre reactors., Margulies M et al., Nature 2005, 437: 376–380., 2005</ref> *[[Smartdenovo]]<ref> Ultra-fast de novo assembler using long noisy reads, https://github.com/ruanjue/smartdenovo</ref> *[[Abruijn]]<ref>Assembly of Long Error-Prone Reads Using de Bruijn Graphs, Yu Lin et al., http://dx.doi.org/10.1101/048413, 2016</ref> *[[Ra]]<ref>Fast and sensitive mapping of nanopore sequencing reads with GraphMap, Ivan Sović et al., Nature Communications 7,Article number:11307;doi:10.1038/ncomms11307, 2016</ref>

=== Hybrid assembly ===

Up to now, 4 hybrid assemblers have been listed: *[[DBG2OLC]]<ref>DBG2OLC: Efficient Assembly of Large Genomes Using the Compressed Overlap Graph, Chengxi Ye et al., arXiv:1410.2801., 2015</ref> *[[Spades]]<ref>SPAdes: A New Genome Assembly Algorithm and its Applications to Single-Cell Sequencing, Anton Bankevich at al., Journal of Computational Biology. 19(5)., 2012</ref> *[[Cerulean]]<ref>Cerulean: A hybrid

assembly using high throughput short and long reads, Viraj Deshpande et al., arXiv:1307.7933., 2013</ref> *[[Unicycler]]<ref>Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads, Ryan R. Wick et al, doi: https://doi.org/10.1101/096412, 22 dec. 2016</ref>

Schematically, assembly pipelines that use long reads and short reads initiate a pre-assembly (production of contigs) from short readings, then long readings are used to improve the pre-assembly by closing gaps, resolving repetitive regions,...

== Long read correction ==

The available correction software are mainly based on two strategies:

*Hybrid correction: it needs long reads and short reads as input *Denovo correction: it takes only long reads

The first one uses short reads, such as Illumina, which have a much lower error rate, to correct long reads. The other strategy consist in aligning long reads against themselves.

==== Hybrid correction ====

Some hybrid correctors gave the possibility of recovering only the corrected regions ('trim' function) as well as the corrected and non-corrected regions of long reads after correction (untrimmed reads).

8 evaluated hybrid correctors are listed:

*[[LSC-2]]<ref>Kin Fai Au et al., Improving PacBio Long Read Accuracy by Short Read Alignment, http://dx.doi.org/10.1371/journal.pone.0046679, 2012</ref> *[[Pacbiotoca]]<ref>Sergey Koren et al., Hybrid error correction and de novo assembly of single-molecule sequencing reads, Nature Biotechnology 30(7):693-700, 2012</ref> *[[Ectools]]<ref>Hayan Lee et al., Error correction and assembly complexity of single molecule sequencing reads, http://dx.doi.org/10.1101/006395, 2014 </ref> *[[Proovread]]<ref>Hackl T., proovread: large-scale high-accuracy PacBio correction through iterative short read consensus, Bioinformatics. 30(21):3004-11. doi: 10.1093/bioinformatics/btu392, 2014</ref> *[[Lordec]]<ref>LoRDEC: accurate and efficient long read error correction, Salmela L, Rivals E, Bioinformatics. 30(24):3506-14. doi: 10.1093/bioinformatics/btu538, 2014</ref> *[[Nanocorr]]<ref>Oxford Nanopore Sequencing and de novo Assembly of a Eukaryotic Genome, Sara Goodwin et al., Genome Research doi: 10.1101/gr.191395.115, 2015</ref> *[[Nas]]<ref>Genome assembly using Nanopore-guided long and error-free DNA reads, Mohammed-Amin Madoui et al., BMC Genomics 16:327; doi: 10.1186/s12864-015-1519-z, 2015</ref> *[[Jabba]]<ref>Jabba: hybrid error correction for long sequencing reads, Giles Miclotte at al., Algorithms for Molecular Biology 11:10; doi: 10.1186/s13015-016-0075-7, 2016</ref>

==== Denovo correction ====

3 Denovo correctors are listed:

*[[Pacbiotoca]]<ref>Sergey Koren et al., Hybrid error correction and de novo assembly of single-molecule sequencing reads, Nature Biotechnology 30(7):693-700, 2012</ref> *[[MHAP (CANU)]]<ref>Assembling Large Genomes with Single-Molecule Sequencing and Locality Sensitive Hashing, Konstantin Berlin at al., Nature Biotechnology doi: 10.1038/nbt.3238, 2015</ref> *[[Lorma]]<ref>Accurate selfcorrection of errors in long reads using de Bruijn graphs, Leena Salmela et al., arXiv:1604.02233, 2016</ref>

== Assembly Boosting == In the case of an existing assembly, long reads can be used to join contigs or fill the internal gaps inside the scaffolds.

===== Gap filling ===== *[[PBJelly]] *[[Pilon]]

===== Scaffolding ===== *[[SPACELongReads]] *[[PBJelly]] *[[AHA]]

= Genome annotation =

== Structural annotation == ===Transposable elements detection and annotation=== *[[The REPET package]]<ref>Hadi Quesneville et al., Combined Evidence Annotation of Transposable Elements in Genome Sequences, http://dx.doi.org/10.1371/journal.pcbi.0010022, July 29, 2005</ref>

===Gene prediction=== *[[PASA]] *[[Eugene]] *[[Augustus]] *[[Maker]] *[[GlimmerM]] *[[Twinscan]]

===Genome viewer and manual annotation=== *[[WebApollo]] *[[Artemis]] *[[GenomeView]] *[[Orcae]] *[[myGenomeBrowser]]

== Functional annotation == *[[InterProScan]] *[[Gene Ontology]] *[[KASS]]

= About Elixir = The goal of ELIXIR [https://www.elixir-europe.org/] is to orchestrate the collection, quality control and archiving of large amounts of biological data produced by life science researchers.

== References == Consult the [//meta.wikimedia.org/wiki/Help:Contents User's Guide] for information on using the wiki software. * [//www.mediawiki.org/wiki/Special:MyLanguage/Manual:Configuration_settings Configuration settings list] * [//www.mediawiki.org/wiki/Special:MyLanguage/Manual:FAQ MediaWiki FAQ] * [https://lists.wikimedia.org/mailman/listinfo/mediawiki-announce MediaWiki release mailing list] * [//www.mediawiki.org/wiki/Special:MyLanguage/Localisation#Translation_resources Localise MediaWiki for your language]