 doi:10.20944/preprints201808.0423.v1

mitoMaker:

a pipeline for automatic assembly and annotation of animal mitochondria using raw NGS data

Alex Schomaker-Bastos[1π] and Francisco Prosdocimi[1*]

[1] Laboratório Multidisciplinar para Análise de Dados (LAMPADA), Instituto de Bioquímica Médica Leopoldo de Meis, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil.

*Correspondence: prosdocimi@bioqmed.ufrj.br

π ***In memoriam (08/01/2015)***

 doi:10.20944/preprints201808.0423.v1

Abstract

Next-generation sequencing is now a mature technology, allowing partial animal genomes to be produced for many clades. Though many software exist for genome assembly and annotation, a simple pipeline that allows researchers to input raw sequencing reads in *fastq* format and allow the retrieval of a completely assembled and annotated mitochondrial genome is still missing. mitoMaker 1.0 is a pipeline developed in python that implements (i) recursive *de novo* assembly of mitochondrial genomes using a set of increasing k-mers; (ii) search for the best matching result to a target mitogenome and; (iii) performs iterative reference-based strategies to optimize the assembly. After (iv) checking for circularization and (v) positioning tRNA-Phe at the beginning, (vi) *geneChecker.py* module performs a complete annotation of the mitochondrial genome and provides a GenBank formatted file as output.

Availability and implementation: The software runs in UNIX terminal-based interface. The source code is freely available under the MIT License at http://mitomaker.sourceforge.net/

Keywords: mitochondrial DNA, mitochondrial genome, genome assembly, genome annotation, next generation sequencing, animal genomics, partial genomics, bioinformatics.

 doi:10.20944/preprints201808.0423.v1

1. Introduction

With more than a decade of existence, next-generation sequencing is a mature technology and has proven its revolutionary impact to XXI century biology (van Dijk et al., 2014; Koboldt et al., 2013). It has been used worldwide to allow the quick production of high and low coverage genomic sequences from a broad spectrum of organisms (Zhao et al., 2013; Pérez-Enciso et al., 2010). However, researchers are still overwhelmed with the enormous amount of data generated even in partial genomic projects that makes use of sequencing reads provided by 454, Illumina, Ion Torrent and SOLiD DNA sequencers. On the other hand, the advance of bioinformatics tools used to analyze these data automatically is also developing fast, allowing easy processing and annotation of genomic information (Lee et al., 2012). Here we suggest that whole mitochondrial genome assembly and annotation is an interesting starting point for a partial genome analysis of a given organism whose genome has been sequenced in low coverage. Therefore, we provide a completely automatic pipeline developed to assemble and annotate the whole mitogenome using an open-source, user friendly and thorough script developed in python and called via LINUX command line.

After sequencing, assembling and annotating more than 20 metazoan mitogenomes (Prosdocimi et al., 2012; Perini et al., 2014; Souto et al., 2016; Prosdocimi et al., 2016; Uliano-Silva et al., 2016) we observed that, although many software exist to

 doi:10.20944/preprints201808.0423.v1

facilitate the assembly and annotation, a straightforward pipeline allowing an easy reconstruction and annotation of mitochondrial genomes based on raw genomic reads is still missing. This fact motivated us to develop mitoMaker, a pipeline for automatic assembly and annotation of whole mitochondrial genomes based on raw data generated from highperformance DNA sequencers.

mitoMaker is a python script that performs a number of steps, including: (i) de novo assembly of a partial animal genome; (ii) reference-based assembly of the best assembled mitochondrion obtained in the de-novo step; (iii) circularization check to confirm circular genome; (iv) annotation of all protein coding, tRNA and rRNA genes based on a closely related mitochondrial genome, either available in the package or provided by the user. The software produces an output containing the best version of the mitogenome assembled completely annotated in GenBank and FASTA files, providing also a .tbl file ready for database submission to Sequin/GenBank, besides other file formats and statistics.

2. Methods

mitoMaker v1.0 is an open source package that includes: Biopython library (Cock et al., 2009), SOAPdenovo and SOAPdenovo-Trans (Luo et al., 2012), MIRA v4.0 (Chevreux et al., 1999), MITObim v1.6 and v1.7 with minor modifications (Hahn et al., 2013), BLAST+ (Alstchul et al., 1990; Gish et al., 1993), and tRNAscan-SE (Lowe & Eddy, 1997).

The software initially starts a de novo assembly round using SOAPdenovo, SOAPdenovo-trans or MIRA software with the entire set of reads provided by the user. SOAP-denovo-trans

 doi:10.20944/preprints201808.0423.v1

version for transcriptome assembly is used by default since mitochondrial reads are frequently found in higher amounts than average genomic reads and they are often interpreted as repeats by genome assemblers (Rubinstein et al., 2013). In a single lane of Illumina sequencing, even when multiplexed for dozens of different samples using a Nextera kit, for example, millions of sequencing reads are generated, frequently allowing a whole mitochondrial assembly with high coverage (>100-fold). For non-illumina sequencers we suggest the usage of MIRA software, turned on with an appropriate flag, for initial de novo assembly of raw reads (please check the mitoMaker manual provided in the *Supplementary Information*).

Once the de-novo assembly is finished, BLAST+ is used to find the best contig or scaffold matching the target mitogenome provided by the user or selected from the mitoMaker package. If this best match is not close enough in size (default: < 92.5% of size coverage) to the target, other contigs are searched using blastn and all matching contigs are filtered and concatenated according to their relative position of the reference.

The assembly result is checked for circularization, looking for matching regions at the start and end of the provisory mitogenome. Afterwards, mitoMaker checks for the presence of each and every feature present in the reference GenBank file through the *geneChecker* module, which calls upon blastn and blastx, such as the software DOGMA does (Wyman et al., 2004). If circularization is not confirmed and at least one gene feature is missing, a new de novo assembly step is started with a different k-mer size. By default, mitoMaker performs the assembly using 5 different k-mers and allows the user to specify different values in the command line.

mitoMaker stores the very best mitochondrial assembly found during the de novo steps run with different k-mers and then executes the software MIRA and MITObim (Hahn et al., 2013) iteratively in the Mapping mode, using as reference the very

best assembly obtained in the de novo steps. This procedure allows gap closures and increases the build, improving the number of features found and facilitating circularization.

After exhaustively performing de novo and mapping assemblies, the mitogenome is normally built using raw fastq files generated from pair-end sequences containing ~3 gigabytes of genomic information each or ~10 million pair-end reads of 100bp. We verified that mitochondrial reads frequently represent about 0.1%-1% of the entire sequencing reads dataset, although this percentage is different from organism to organism. The mitogenome is normally assembled with high coverage.

As an output, mitoMaker provides a Genbank file containing all the annotated mitochondrial genes ordered with tRNA-Phe at the beginning, with FASTA files corresponding to the annotated Genbank, SEQUIN formatted file (TBL), CAF/MAF final assembly results used for read coverage checks, a PNG image for easy visualization of the annotation results as well as a STATS file presenting general information about the mitogenome assembled.

If the appropriate GenBank reference file is provided, mitoMaker can also be used to assemble and annotate other organellar genomes, such as chloroplasts and other plastids.

We strongly recommend that the user double-checks each feature provided by mitoMaker by visual inspection using Artemis (Carver et al., 2012) or any other genome visualization software in order to resolve inconsistencies that may be present in the start and end positions of genes.

3. Conclusion

mitoMaker is an easy-to-use pipeline for animal mitogenome assembly and annotation. It uses well-known software and internal algorithms, automatically integrating applications in order to provide thoroughly analysed mitogenomes. mitoMaker reduces considerably the amount of expert time and effort required to perform the assembly and annotation of whole mitochondria based on shotgun, genomic NGS data.

4. Acknowledgements

We thank Helena Magarinos Souto, Marcela Uliano-Silva, Violeta Perini, Igor Costa and Nicholas Lima for trying beta versions of mitoMaker and suggesting improvements. The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

The current manuscript was produced in September 2014 and submitted to the specialized journal in bioinformatics field. The editor refused the publication because we did not compare the results to other software: "I must formally decline the manuscript, due to lack of comparisons." We did not performed comparisons at the time because there was not any other program that performed the same features as mitoMaker. Until today, I am not aware about any other program specifically design to assemble and annotate mitochondrial genomes.

Two months after the refusal, the undergraduate student Alex Schomaker Bastos that developed the software and conducted most assemblies and annotations tests was shot dead in a violent robbery happened in Rio de Janeiro on January, 2015. Alex was a brilliant, dedicated and creative student. He was very much interested in Bioinformatics, algorithm development and he was also great teacher, with a genuine interest in educate the younger students. We were very much shocked with his decease and this manuscript has been in drawer since them. With the current possibility of publishing this work as a preprint, I decided to send it for publishing as homage to Alex's memory.

Mitomaker have been extensively used in our laboratory to assemble and annotate dozens of complete mitochondrial genomes from different animals. Although it keeps working, it is nowadays out of date because newer versions of genome assemblers have been produced. Nevertheless, the annotator script originally developed by Alex from the mitoMaker package and named *GeneChecker.py* (it is available at the sourceforge link provided in the abstract) is still in usage and it is probably the most efficient and quick annotator for mitochondrial genomes available worldwide.

5. References

1. Altschul et al. (1990) Basic local alignment search tool. J. Mol. Biol., 215, 403-410.

2. Carver T et al. (2012) Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. Bioinformatics, 28, 464-469.

3. Chevreux B et al. (1999) Genome sequence assembly using trace signals and additional sequence information. Computer Science and Biology: Proceedings of the German Conference on Bioinformatics, 99, 45-56.

4. Cock PJ et al. (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics, 25(11), 1422-3.

5. Gish, W et al. (1993) Identification of protein coding regions by database similarity search. Nature Genet., 3, 266-272

6. Hahn C et al. (2013) Reconstructing mitochondrial genomes directly from genomic next generation sequencing reads - a baiting and iterative mapping approach. Nucleic Acids Res., 1–9.

7. Koboldt DC et al. (2013) The next-generation sequencing revolution and its impact on genomics. Cell., 155(1), 27-38.

8. Lee HC et al. (2012) Bioinformatics tools and databases for analysis of nextgeneration sequence data. Brief Funct. Genomics, 11(1), 12-24.

9. Lowe TM et al. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res., 25, 955–964.

10. Luo et al. (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience, 1, 18.

11. Pérez-Enciso M et al. (2010) Massive parallel sequencing in animal genetics: wherefroms and wheretos. Anim Genet., 41(6), 561-9.

12. Perini et al. (2014) The complete mitochondrial genome of the southern purplespotted gudgeon Mogurnda adspersa (Perciformes: Eleotridae) through pyrosequencing. Mitochondrial DNA, Mar 24, 1-3.

13. Prosdocimi F et al. (2012) The complete mitochondrial genome of two recently derived species of the fish genus Nannoperca (Perciformes, Percichthyidae). Mol. Biol. Rep., 39, 2767–2772.

14. Prosdocimi F, Souto HM, Ruschi PA, Furtado C, Jennings WB. Complete mitochondrial genome of the versicoloured emerald hummingbird Amazilia versicolor, a polymorphic species. Mitochondrial DNA A DNA Mapp Seq Anal. 2016 Sep;27(5):3214-5. doi: 10.3109/19401736.2015.1007352. Epub 2015 Mar 11. PubMed PMID: 25758043.

15. Rubinstein ND et. al. (2013) Deep sequencing of mixed total DNA without barcodes allow efficient assembly of highly plastic ascidian mitochondrial genomes. Genome Biol. Evol., 5(6), 1185-1199.

16. Souto HM, Ruschi PA, Furtado C, Jennings WB, Prosdocimi F. The complete mitochondrial genome of the ruby-topaz hummingbird Chrysolampis mosquitus through Illumina sequencing. Mitochondrial DNA A DNA Mapp Seq Anal. 2016;27(1):769-70. doi: 10.3109/19401736.2014.915533. Epub 2014 May 19. PubMed PMID: 24841437.

17. Uliano-Silva M, Americo J, Bastos AS, Furtado C, Rebelo MF, Prosdocimi F. Complete mitochondrial genome of the brown mussel Perna

perna (Bivalve, Mytilidae). Mitochondrial DNA A DNA Mapp Seq Anal. 2016 Nov;27(6):3955-3956. Epub 2015 Jan 28. PubMed PMID: 25627319.

18. van Dijk EL et al. (2014) Ten years of next-generation sequencing technology. Trends Genet., 30(9), 418-426.

19. Wyman SK et al. (2004) Automatic annotation of organellar genomes with DOGMA. Bioinformatics, 20(17), 3252-3255.

20. Zhao YQ et al. (2013) Comparative systems biology between human and animal models based on next-generation sequencing methods. Dongwuxue Yanjiu, 34(E2), E35-41.