MitoFree a user-friendly and lightweight mitogenomics pipeline for public data

Gabriel A. Vieira[1]; Francisco Prosdocimi[1]

[1] *Laboratório de Genômica e Biodiversidade. Instituto de Bioquímica Médica Leopoldo de Meis,*
Universidade Federal do Rio de Janeiro, 21941-902, Rio de Janeiro-RJ, Brasil.
E-mail: gabriel.vieira@bioqmed.ufrj.br

Introduction

Sequence Read Archive (SRA) is the largest public database of raw sequencing data, featuring WGS, RNA-Seq, Exome and other sequencing datasets. These data can be used to obtain partial and complete mitochondrial genomes (mitogenomes) useful in population genetics, evolutionary and phylogeographic studies. A large number of species have public data available in SRA but lack complete sequences of mitogenomes. In order to make mitogenomics accessible to researchers without access to robust servers and/or bioinformatics expertise, we are developing **MitoFree software**. MitoFree is a lightweight script that aims to automate mitogenome assembly, annotation and phylogenomic analyses based on public sequencing data.

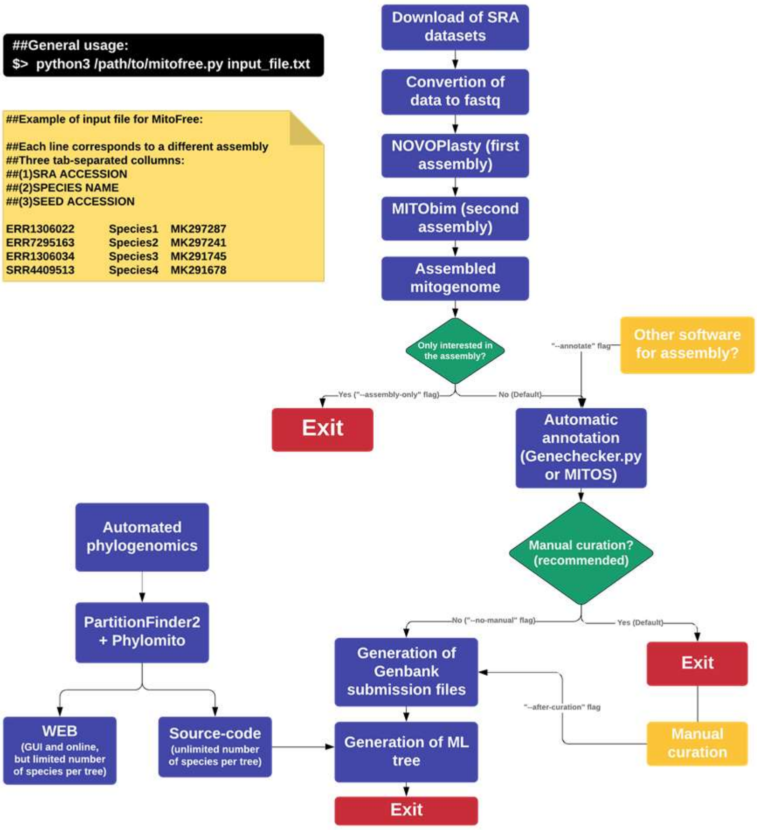
Material and Methods

MitoFree is being developed in Python 3 and uses the Biopython module and several other published algorithms/software. In its full implementation, MitoFree's algorithm will automatically download the SRA dataset and convert it to fastq using "fastq-dump" script (sratoolkit package). Then, it will perform an initial assembly of the mitogenome using NOVOPlasty and second assembly round using MITObim. The mitogenome annotation will be performed by either MITOS Web Server or GeneChecker and will automatically provide the input files required for Genbank submission. Finally, a phylogenomic tree will be constructed using the concatamer of all 13 mitochondrial protein-coding genes through Phylomito and PartitionFinder2.

Results and Discussion

A beta version of MitoFree can be downloaded at https://github.com/gavieira/mitofree. This version automatically downloads SRA datasets and uses them to assemble mitogenomes. The assemblers used by MitoFree are very efficient in RAM usage, allowing this pipeline to be run on standard personal computers.

In order to run, the program needs an input file with three columns per line: (i) SRA run accession number; (ii) sample identifier; and (iii) GenBank accession number of a mitochondrial sequence from a related organism. MitoFree's command-line usage, input file and workflow are depicted in Figure 1.



```
##General usage:
$>  python3 /path/to/mitofree.py input_file.txt
```

```
##Example of input file for MitoFree:

##Each line corresponds to a different assembly
##Three tab-separated collumns:
##(1)SRA ACCESSION
##(2)SPECIES NAME
##(3)SEED ACCESSION

ERR1306022        Species1   MK297287
ERR7295163        Species2   MK297241
ERR1306034        Species3   MK291745
SRR4409513        Species4   MK291678
```

Download of SRA datasets
Convertion of data to fastq
NOVOPlasty (first assembly)
MITObim (second assembly)
Assembled mitogenome
Only interested in the assembly?
Yes ("--assembly-only" flag)
No (Default)
"--annotate" flag
Other software for assembly?
Exit
Automatic annotation (Genechecker.py or MITOS)
Manual curation? (recommended)
No ("--no-manual" flag)
Yes (Default)
Exit
Manual curation
"--after-curation" flag
Generation of Genbank submission files
Generation of ML tree
Exit
Automated phylogenomics
PartitionFinder2 + Phylomito
WEB (GUI and online, but limited number of species per tree)
Source-code (unlimited number of species per tree)

Figure 1. MitoFree's workflow and usage.

Conclusion

This software will hopefully contribute to making mitogenomic studies as widespread and relevant as possible, fostering and expediting developments on the study of numerous clades.

Acknowledgements

We thank CNPq and FAPERJ for funding our work.