

# MitoFree

## AN AUTOMATED PIPELINE FOR MITOGENOME ASSEMBLY, ANNOTATION AND PHYLOGENOMIC ANALYSIS USING PUBLIC DATA

Gabriel A. Vieira<sup>1</sup>; Francisco Prosdocimi<sup>1</sup>

<sup>1</sup> 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](mailto: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. SRA presents hundreds of datasets containing sequencing data from ants that covers more than 17 subfamilies. Nevertheless, GenBank present only 35 complete mitogenomes for the clade, barely encompassing six subfamilies. Thus, the use of public data for mitogenomics might bring light in studies focusing ant diversity, habitat colonization, dispersion and evolution. 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 (sra toolkit 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 concatenation 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 be used, the program needs an inputfile 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.

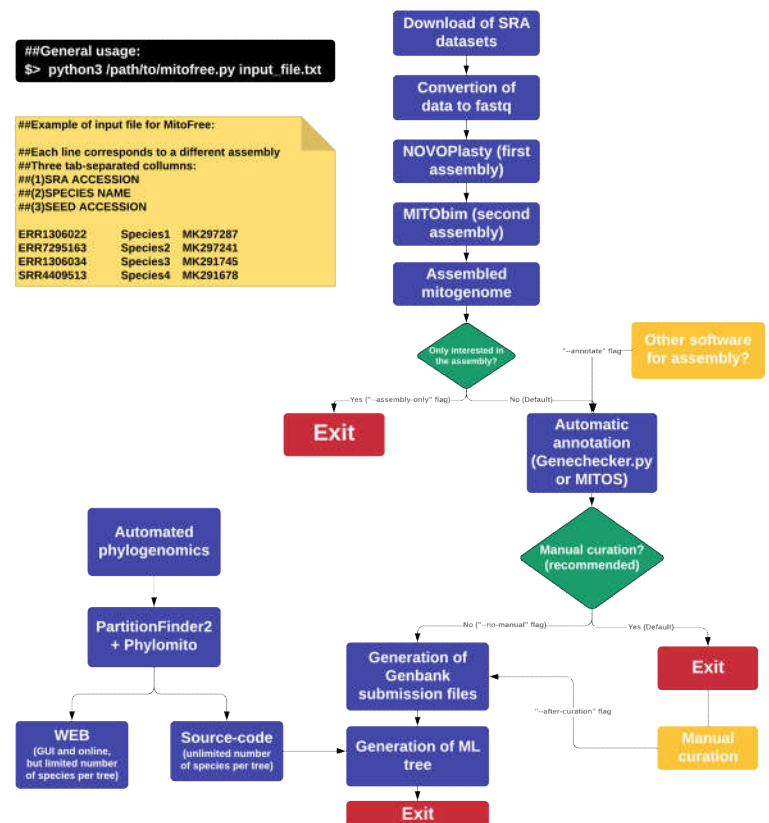


Figure 1. MitoFree's workflow and usage.

### Conclusion

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