# Statistical Mitogenome Assembly with Repeats

Fahad Alqahtani
*Department of Computer Science and Engineering*
*University of Connecticut*
Storrs, CT, USA
fahad.alqahtani@uconn.edu

Ion I. Măndoiu
*Department of Computer Science and Engineering*
*University of Connecticut*
Storrs, CT, USA
ion@engr.uconn.edu

***Abstract*—In this abstract we introduce an automated pipeline that can assemble *de novo* and annotate complete circular mitochondrial sequences from whole genome sequencing data.**

***Index Terms*—mitochondrial genome assembly, bioinformatics pipeline**

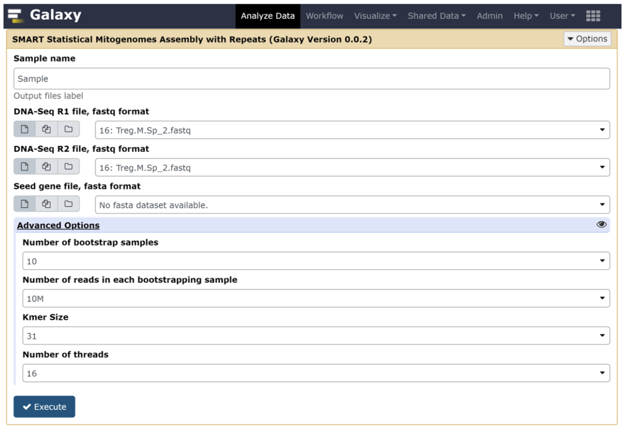
## I. Introduction

The mitochondria are cellular organelles often called the cell powerhouses due to their key role in the production of adenosine triphosphate (ATP), the energy currency of the cells. Found in most eukaryotic organisms, the mitochondria have their own circular genomes. They are inherited maternally in most animals, and typically present in thousands of copies in the cytoplasm of each cell, although the copy number varies between cells of different organs [1].

Single nucleotide polymorphisms in mitochondrial genomes have long been used for tracking human migrations. Mitochondrial DNA mutations and *heteroplasmy* (simultaneous presence of multiple mitochondrial sequences in a cell) have also been associated with human diseases. [2]. Finally, mitochondrial genome sequences can be used for evolutionary studies of less studied species for which nuclear genomes have not yet been assembled [3].

Next-generation sequencing technologies have made it possible to quickly and inexpensively generate large numbers of relatively short reads from both the nuclear and mitochondrial DNA of the cells. Unfortunately, assembling such *whole-genome sequencing (WGS)* data with standard *de novo* assemblers often fails to generate high quality mitochondrial genome sequences due to the large difference in copy number (and hence sequencing depth) between the mitochondrial and nuclear genomes. Assembly of complete mitochondrial genome sequences is further complicated by the fact that many *de novo* assemblers are not designed to assemble circular genomes, and by the presence of repeats in the mitochondrial genomes of some species. In this abstract we introduce the *Statistical Mitogenome Assembly with Repeats (SMART)* pipeline that can assemble *de novo* and annotate complete circular mitochondrial genome sequence from whole genome sequencing data even in the presence of repeats. The SMART pipeline is available as a web-based Galaxy tool at https://neo.engr.uconn.edu/?tool_id=SMART (see Fig. 1).

## II. Pipeline Workflow



SMART Statistical Mitogenomes Assembly with Repeats (Galaxy Version 0.0.2)

Sample name: Sample
Output files label

DNA-Seq R1 file, fastq format: 16: Treg.M.Sp_2.fastq

DNA-Seq R2 file, fastq format: 16: Treg.M.Sp_2.fastq

Seed gene file, fasta format: No fasta dataset available.

Advanced Options

Number of bootstrap samples: 10

Number of reads in each bootstrapping sample: 10M

Kmer Size: 31

Number of threads: 16

Fig. 1. Galaxy interface for the SMART pipeline.

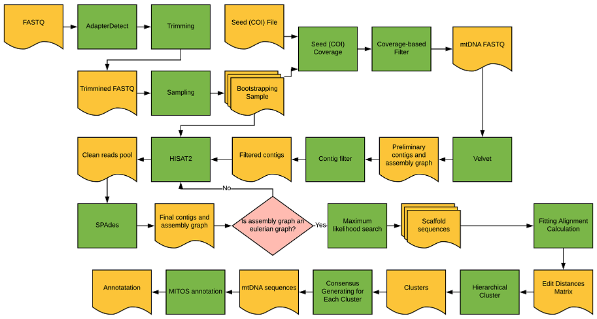


Fig. 2. SMART pipeline workflow.

The main stages of the SMART pipeline (see Fig. 2) are as follows. (1) Automatic adapter detection and trimming. (2) Coverage-based read filtering based on available seed sequences such as the cytochrome c oxidase I (COI) gene. (3) Preliminary contig assembly using Velvet of reads passing the coverage filter. (4) Filtering of preliminary contigs by BLAST searches against a comprehensive mitochondrial database. (5) Mitochondrial read selection by alignment to preliminary contigs and secondary assembly using SPAdes. (6) Iterative scaffolding and gap filling based on maximum likelihood. The above steps are repeated multiple times on sets of reads generated by bootstrapping. In the final stage (7) circular sequences assembled from each bootstrap sample are aligned, clustered, and a consensus sequence is generated and annotated for each cluster.

## References

[1] K. Veltri, E. Myrna, and S. Gurmit. “Distinct genomic copy number in mitochondria of different mammalian organs.” Journal of cellular physiology 143.1 (1990): 160-164.

[2] J. Stewart, and C. Patrick. ”The dynamics of mitochondrial DNA heteroplasmy: implications for human health and disease.” Nature Reviews Genetics 16.9 (2015): 530.

[3] A. Kurabayashi, and S. Masayuki. ”Afrobatrachian mitochondrial genomes: genome reorganization, gene rearrangement mechanisms, and evolutionary trends of duplicated and rearranged genes.” BMC genomics 14.1 (2013): 633.

This work was supported in part by a KACST Ph.D. Scholarship to FA and NSF Award 1564936 to IIM.

978-1-5386-8520-4/18/$31.00 ©2018 IEEE