An Automated Workflow for Mitochondrial DNA Extraction and Analysis from High Throughput Sequencing Data

Alisha Mechtley
Advisor: Dr. Alan Lemmon
Florida State University

Abstract Next-generation sequencing data are rich in information and contain many off-target sequences (reads), including mitochondrial reads, that are often ignored but which may be biologically relevant. Mitochondrial DNA (mtDNA) is now providing new perspectives on the tree of life and the etiology of the common complex diseases. The mtDNA codes for important bioenergetic genes, has a very high mutation rate, and can be present in thousands of copies per cell. The mechanisms by which new mtDNA mutations arise among thousands of other mtDNAs (called heteroplasmies) is poorly understood, and is complicated by the presence of nuclear mitochondrial insertions (NUMTs). My research utilizes current de-novo and referenced based methods of mitochondrial genome extraction from high throughput sequencing data. I implement a workflow to map reads from a popular next generation platform (Illumina) to custom-built reference genomes and extract the NUMTs using an open-source genome analysis platform, Galaxy. Workflows in Galaxy can be shared and published via the web, improving repeatability and data sharing among scientists. I discuss how to extend this workflow to include NUMT insertion rate detection in gene trees and heteroplasmy variant detection and annotation.

Introduction

Illumina technology is currently the most successful and widely adopted next-generation sequencing technology. Illumina can be used to sequence single-end or paired-end reads. Paired-end reads are two ends of the same DNA molecule where one end is sequenced, flipped around, and the other end is sequenced on the opposite strand. In this pipeline, single or paired-end reads can be bioinformatically manipulated in order to extract the DNA that maps to a mitochondrial reference genome of the same species or of a closely related species. Those reads that map to both the nuclear and mitochondrial genomes are identified as mitochondrial sequences that have been inserted into the nuclear genome (NUMTs). The flanking regions of NUMTs are extracted from the nuclear genome and a gene tree of nuclear mitochondrial insertions is created for several species. From this tree, the insertion rate of mitochondrial sequences into the nuclear genome can be detected. The remaining mitochondrial reads are run through a separate part of the pipeline for heteroplasmy detection and heteroplasmy variant quantification.

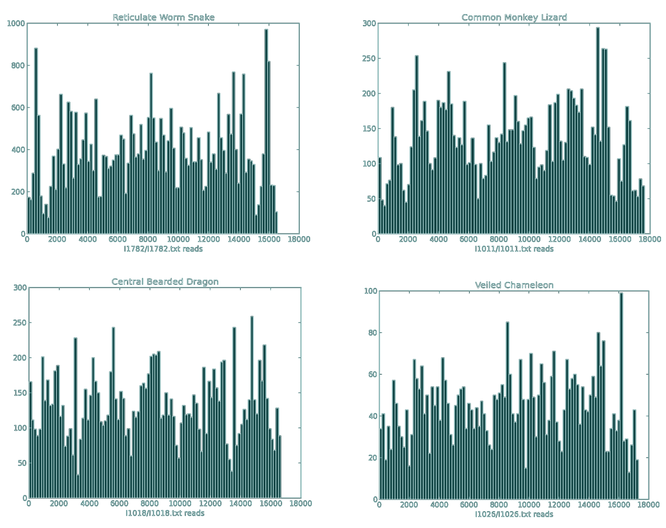

Figure 1: The number of mitochondrial reads versus the mapped location (in base pairs) for four vertebrates in anchored phylogenomic studies

Workflow

The workflow starts with raw reads and a mitochondrial reference genome from the same species or from a closely related species. Reads that do not map to the reference using a Burrow-Wheelers (Heng and Durbin, 2009) algorithm are filtered out, the remaining reads are mapped to the nuclear genome of that species using Bowtie 2 (Langmead and Salzberg, 2012). Reads that map to both the nuclear and mitochondrial genomes are considered NUMTs, the rest are assembled into the mitochondrial genome for that sample or screened for heteroplasmies in a separate workflow. The flanking regions (hundreds of base pairs upstream and downstream of NUMTs) are extracted from the nuclear DNA reference so that multiple mitochondrial insertions of the same sequence into different locations in the nuclear genome can be identified. The reads are assembled into contigs (combined where they overlap) using Velvet. Once the NUMTs of several species have been collected, they will be aligned to each other and a phylogenetic tree will be constructed. The age of insertion of a NUMT is calculated by aligning each sequence to a previously aligned ancestral and modern mitochondrial sequence. The total number of sites in the aligned region that differ in the ancestral and modern mitochondrial sequences are tabulated, and the same is done for the NUMT and modern sequences to create an allele matching ratio. This ratio is used as an estimate for the point of insertion (Bensasson et. al 2003 and Dayama et al. 2014).

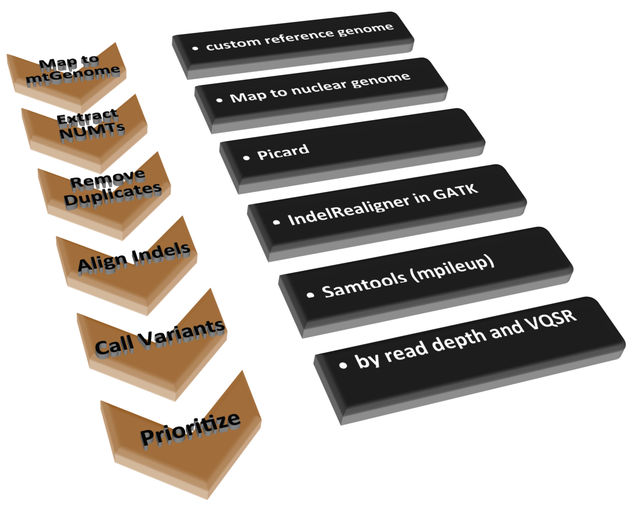

Figure 2: Method of mitochondrial DNA and NUMT extraction based on the MToolBox pipeline.

Example

Heteroplasmy detection can be performed if the DNA sequences were isolated from separate tissues (e.g., blood, lung, or cheek). The 1000 Genomes project sequenced and published the genomes of a large number of people (2,577 individuals from around the world including over 26 different ancestry populations) and much of the data was analyzed for heteroplasmy by Diroma et al (2014). Although only blood samples were taken, the study segregated samples into Lymphoblastoid Cell Lines (LCLs) and blood subsets on the basis of Epstein-Barr virus (EBV) coverage values provided by the 1000 Genomes Consortium and a mitochondrial extraction and heteroplasmy pipeline was created. That pipeline was later implemented in a program called Mtoolbox (Calabrese et al. 2014), which internally references a human mitochondrial and nuclear genome for mapping and is only available to use on human sequence data as a result. Since many of the algorithms and tools used in their pipeline (e.g. BWA, GATK, mpileup, SAMtools, and various filters) are available in Galaxy, I created a workflow utilizing these tools that can also be applied to other species, including non-model organisms whose genomes have not been published. Although software already exists for mitochondrial DNA extraction and annotation (Hahn, Bachmann, and Chevreux, 2013), it currently does not offer options for NUMT extraction, heteroplasmy detection, or visualization of the results. Having such options available for multiple species makes it possible to study the evolution of NUMTs and herteroplasmies which can shed light on the mechanisms underlying mitochondrial dysfunction and its role in a wide range of metabolic and degenerative diseases, cancer, and aging.

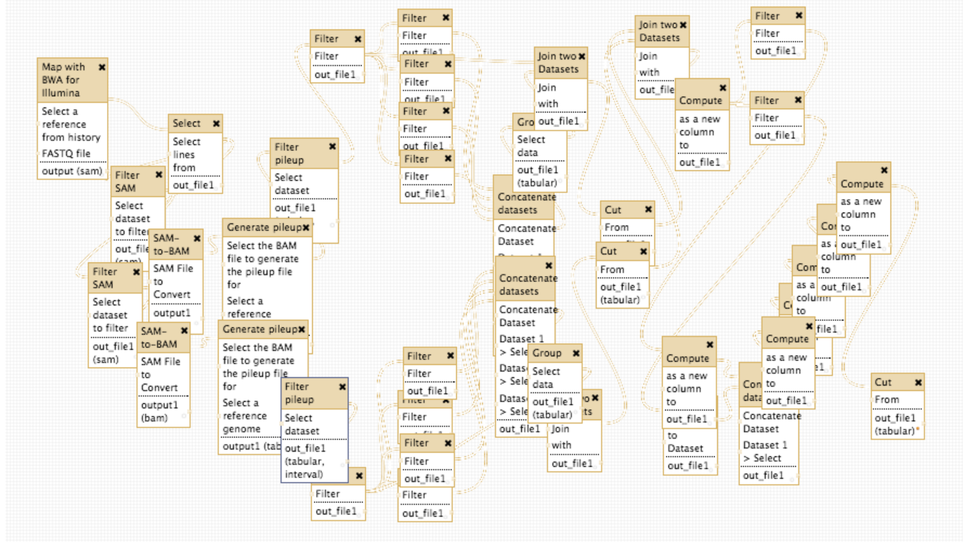

Figure 3: Example workflow for heteroplasmy detection in Galaxy

References

- DNA Sequencing, *www.illumina.com* N.p., n.d. Web. 20 Mar. 2014
- Yakovchuk P, Protozanova E, Frank-Kamenetskii MD 2006, *Base-stacking and base-pairing contributions into thermal stability of the DNA double helix*, Nucleic Acids Res. 34 (2): 56474.
- Li, Heng, and Richard Durbin. "Fast and accurate short read alignment with BurrowsWheeler transform." *Bioinformatics* 25, no. 14 (2009): 1754-1760.
- Langmead, Ben, and Steven L. Salzberg. "Fast gapped-read alignment with Bowtie 2." *Nature methods* 9, no. 4 (2012): 357-359.
- Dayama, Gargi, Sarah B. Emery, Jeffrey M. Kidd, and Ryan E. Mills. "The genomic landscape of polymorphic human nuclear mitochondrial insertions." *Nucleic acids research* 42, no. 20 (2014): 12640-12649.
- Bensasson, Douda, Marcus W. Feldman, and Dmitri A. Petrov. "Rates of DNA duplication and mitochondrial DNA insertion in the human genome." *Journal of Molecular Evolution* 57, no. 3 (2003): 343-354.
- Diroma, Maria Angela, Claudia Calabrese, Domenico Simone, Mariangela Santorsola, Francesco Maria Calabrese, Giuseppe Gasparre, and Marcella Attimonelli. "Extraction and annotation of human mitochondrial genomes from 1000 Genomes Whole Exome Sequencing data." *BMC genomics* 15, no. Suppl 3 (2014): S2.
- Calabrese, Claudia, Domenico Simone, Maria Angela Diroma, Mariangela Santorsola, Cristiano Gutt, Giuseppe Gasparre, Ernesto Picardi, Graziano Pesole, and Marcella Attimonelli. "MToolBox: a highly automated pipeline for heteroplasmy annotation and prioritization analysis of human mitochondrial variants in high-throughput sequencing." *Bioinformatics* 30, no. 21 (2014): 3115-3117.
- Hahn, Christoph, Lutz Bachmann, and Bastien Chevreux. "Reconstructing mitochondrial genomes directly from genomic next-generation sequencing readsa baiting and iterative mapping approach." *Nucleic acids research* (2013): gkt371.