Algorithms for Circular Organelle Genome Assembly

Fahad Alqahtani

Supervisor: Dr. Ion Măndoiu
Associate Advisors: Dr. Mukul Bansal & Dr. Derek Aguiar

Computer Science & Engineering Department
University of Connecticut

Outline

- Background/Motivation
- Related Work
- Statistical Mitogenome Assembly with Repeats (SMART)
    - The pipeline
    - Results
- Conclusion & Ongoing and Future Work

Outline

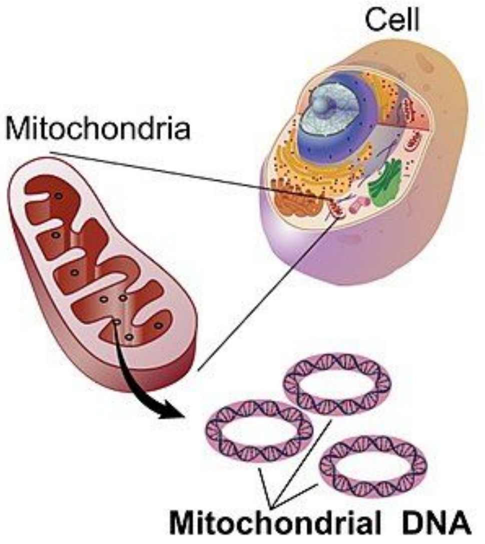
- **Background/Motivation**
- Related Work
- Statistical Mitogenome Assembly with Repeats (SMART)
  - The pipeline
  - Results
- Conclusion & Ongoing and Future Work

Organelle

- It is subunit within a cell
  - has a specific function
- Some types of organelles have own genomes
  - Mitochondria
  - Chloroplasts

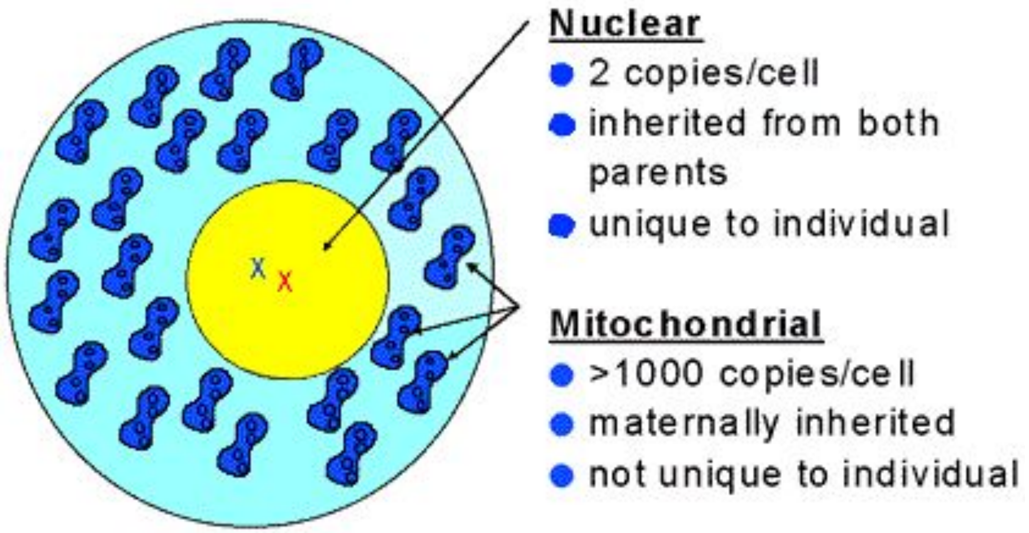
Mitochondria

- Cellular organelles within eukaryotic cells
  - Convert chemical energy from food into adenosine triphosphate (ATP)
  - The popular term "powerhouse of the cell" was coined by Philip Siekevitz in 1957



Mitochondrial DNA - Wikipedia

Nuclear Genome vs. Mitochondrial Genome

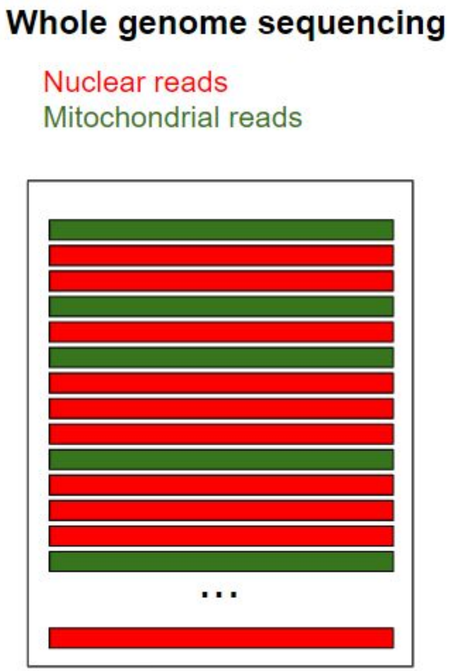


Nuclear

- 2 copies/cell
- inherited from both parents
- unique to individual

Mitochondrial

- >1000 copies/cell
- maternally inherited
- not unique to individual



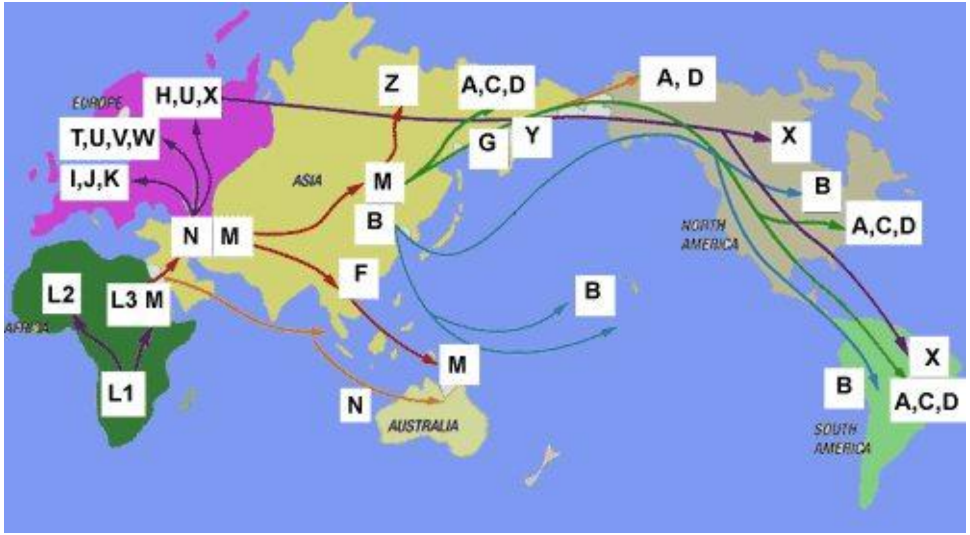
Whole genome sequencing

Nuclear reads
Mitochondrial reads

Source:https://www.fbi.gov/about-us/lab/forensic-science-communications/fsc/july1999/dnalist.htm/dnaf1.htm

Why sequence the mitogenome?

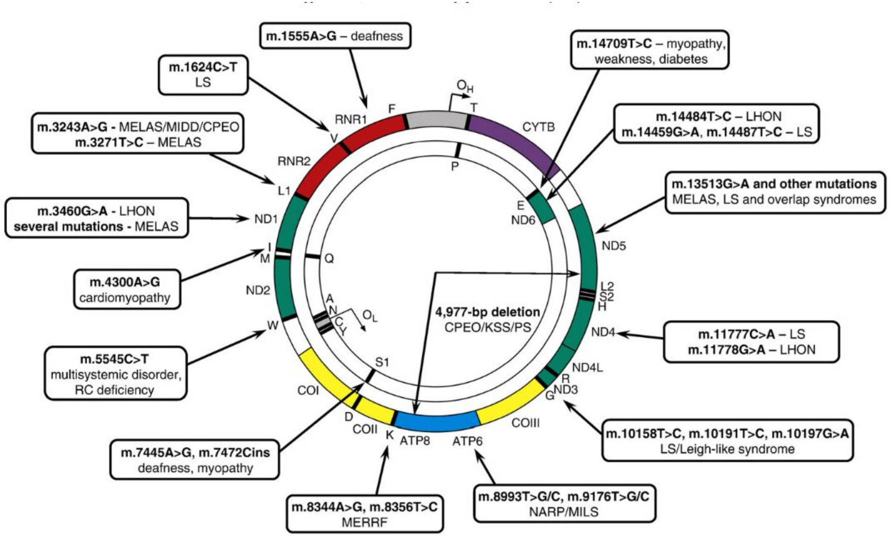
- Inferring human population migrations
  - Single nucleotide polymorphisms in mitochondrial genome have long been used for tracking human migration



https://blog.23andme.com/ancestry/haplogroups-explained/

Why sequence the mitogenome?

- Plays Important role in disease
  - Mitochondrial DNA mutations have also been associated with human diseases



m.1555A>G – deafness

m.1624C>T LS

m.3243A>G - MELAS/MIDD/CPEO
m.3271T>C – MELAS

m.3460G>A - LHON
several mutations - MELAS

m.4300A>G cardiomyopathy

m.5545C>T multisystemic disorder, RC deficiency

m.7445A>G, m.7472Cins deafness, myopathy

m.8344A>G, m.8356T>C MERRF

m.8993T>G/C, m.9176T>G/C NARP/MILS

m.10158T>C, m.10191T>C, m.10197G>A LS/Leigh-like syndrome

m.11777C>A – LS
m.11778G>A – LHON

m.13513G>A and other mutations MELAS, LS and overlap syndromes

m.14484T>C – LHON
m.14459G>A, m.14487T>C – LS

m.14709T>C – myopathy, weakness, diabetes

4,977-bp deletion CPEO/KSS/PS

Tuppen, Helen AL, et al. "Mitochondrial DNA mutations and human disease." Biochimica et Biophysica Acta (BBA)-Bioenergetics 1797.2 (2010): 113-128.

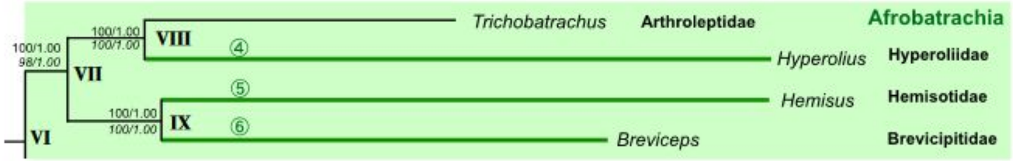
Why sequence the mitogenome?

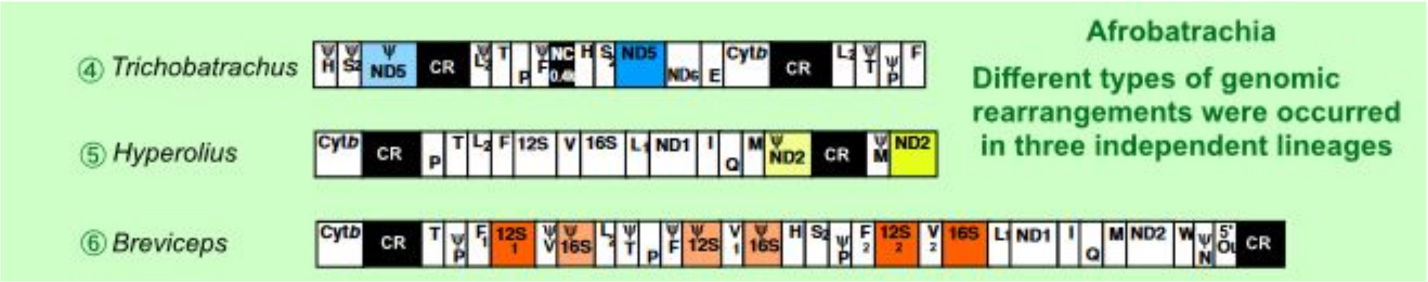
- Useful tool in forensic sciences
  - Mitochondrial DNA analysis can be a useful tool in forensics, especially when a crime scene sample contains degraded DNA not suitable for nuclear DNA tests

https://dps.mn.gov/divisions/bca/bca-divisions/forensic-science/Pages/dna-mitochondrial.aspx

Why sequence the mitogenome?

- Species tree reconstruction
  - Mitochondrial genome sequences can be used for evolutionary studies of non-model species for which nuclear genomes are not yet available




④ Trichobatrachus
ψ H | ψ S2 | ψ ND5 | CR | ψ L2 | T | P | ψ F | NC 0.48 | H | S2 | ND5 | ND6 | E | Cytb | CR | L2 | ψ T | ψ P | F
⑤ Hyperolius
Cytb | CR | P | T | L2 | F | 12S | V | 16S | L1 | ND1 | I | Q | M | ψ ND2 | CR | ψ M | ND2
⑥ Breviceps
Cytb | CR | T | ψ P | F1 | 12S 1 | ψ V | ψ 16S | L2 | ψ T | P | ψ F | ψ 12S | V1 | ψ 16S | H | S2 | ψ P | F2 | 12S 2 | V2 | 16S | L1 | ND1 | I | Q | M | ND2 | W | ψ N | 5' OL | CR
Afrobatrachia
Different types of genomic rearrangements were occurred in three independent lineages

Kurabayashi, Atsushi, and Masayuki Sumida. "Afrobatrachian mitochondrial genomes: genome reorganization, gene rearrangement mechanisms, and evolutionary trends of duplicated and rearranged genes." BMC genomics 14.1 (2013): 633.

Outline

- Background/Motivation
- **Related Work**
- Statistical Mitogenome Assembly with Repeats (SMART)
  - The pipeline
  - Results
- Conclusion & Ongoing and Future Work

Mitochondrial DNA Isolation

- Mitochondrial DNA can be experimentally separated from the nuclear DNA and sequenced independently
    - protocols are laborious.

Long-read WGS Data

- Organelle_PBA [Soorni et al 2017]
    - High coverage required (> 50x) & relatively high cost of long-read sequencing make this approach uncommon

Off-the-shelf de Novo Genome Assembly Tools

- Use short reads
    - Most abundant type WGS data
- Fail to generate high quality mitochondrial genome sequences
    - A large difference in copy number (and hence sequencing depth) between the mitochondrial and nuclear genomes
- Recent example:
    - Pyxicephalus adspersus (African bullfrog)

wikipedia

Most Mitogenome Assembly Tools

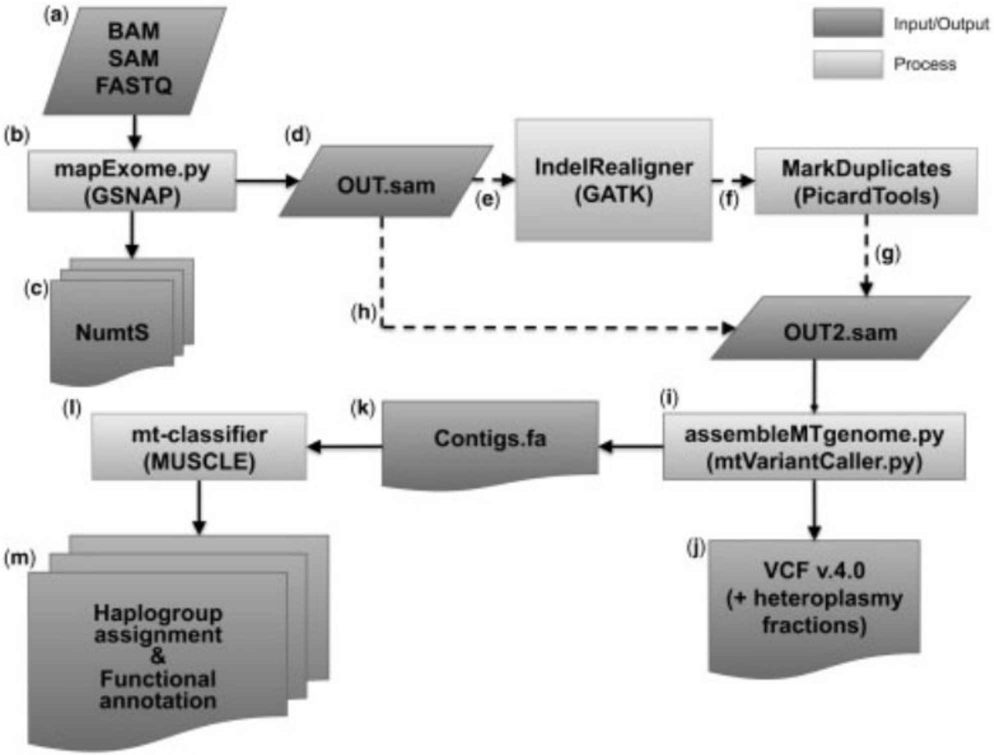
- Categories:
  - Reference-based
    - MToolBox [Calabrese, et al 2014]
  - Seed-and-extend
    - MITObim [Hahn at el 2013] and NOVOPlasty [Dierckxsens at el 2017]
  - De Novo
    - plasmidSPAdes [Antipov et al 2016] and Norgal [Al-Nakeeb et al 2017]

MToolBox

input:

1. Raw data or prealigned reads
2. A mitogenome reference genome
3. A nuclear reference genome

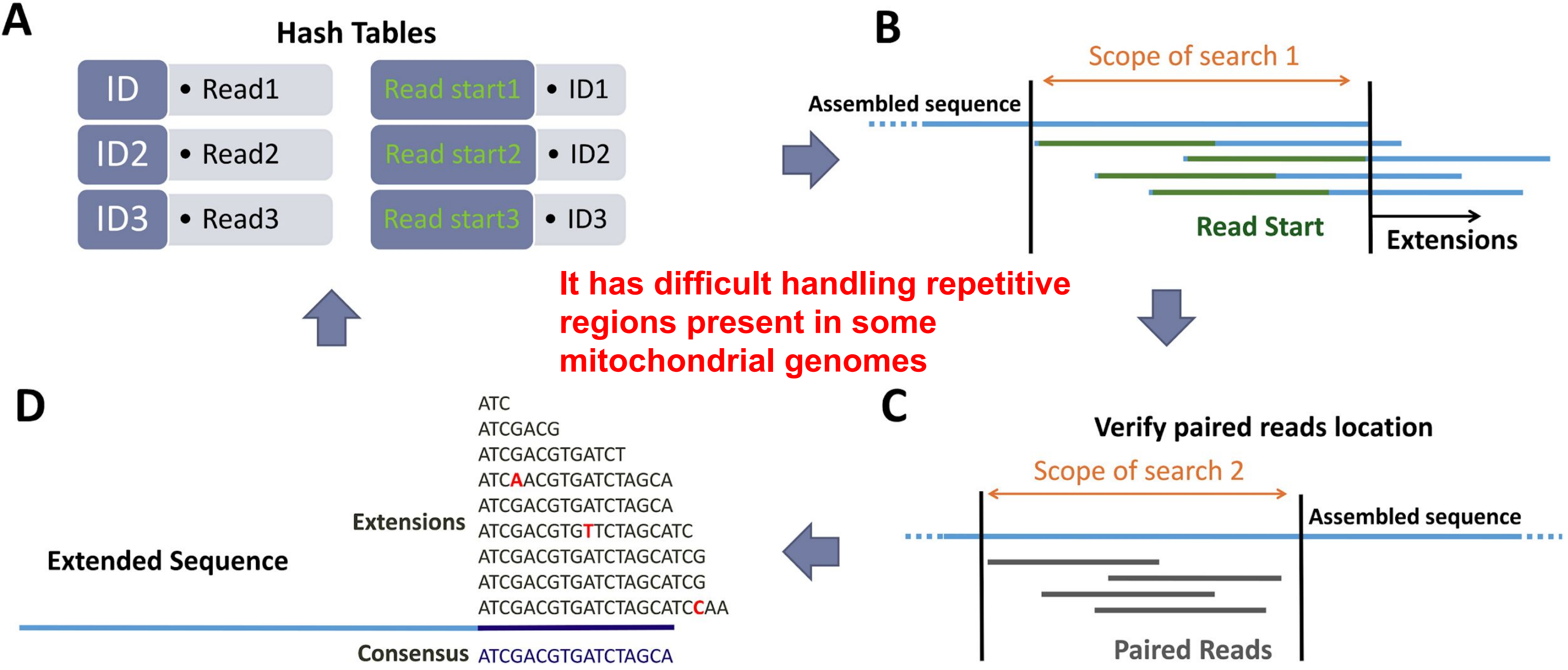
It cannot be used for non-model organisms



NOVOPlasty

Input: 1)Raw reads 2) insert size 3) read length 4) mitogenome size range

5) a seed sequence (coi gene)

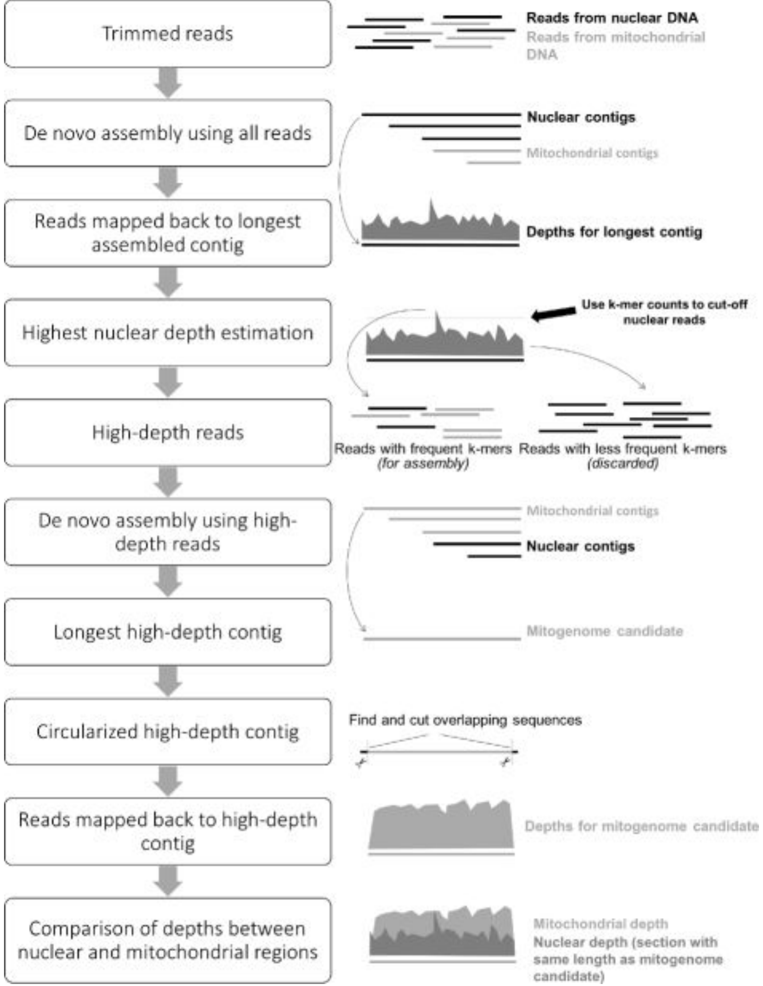


It has difficult handling repetitive regions present in some mitochondrial genomes

Norgal

Input:

Raw reads



Trimmed reads

De novo assembly using all reads

Reads mapped back to longest assembled contig

Highest nuclear depth estimation

High-depth reads

De novo assembly using high-depth reads

Longest high-depth contig

Circularized high-depth contig

Reads mapped back to high-depth contig

Comparison of depths between nuclear and mitochondrial regions

Reads from nuclear DNA
Reads from mitochondrial DNA
Nuclear contigs
Mitochondrial contigs
Depths for longest contig
Use k-mer counts to cut-off nuclear reads
Reads with frequent k-mers (for assembly)
Reads with less frequent k-mers (discarded)
Mitochondrial contigs
Nuclear contigs
Mitogenome candidate
Find and cut overlapping sequences
Depths for mitogenome candidate
Mitochondrial depth
Nuclear depth (section with same length as mitogenome candidate)

It can have prohibitive running times and may still fail to reconstruct complete mitogenomes particularly in the presence of repeats shared between the nuclear and organelle genomes

Outline

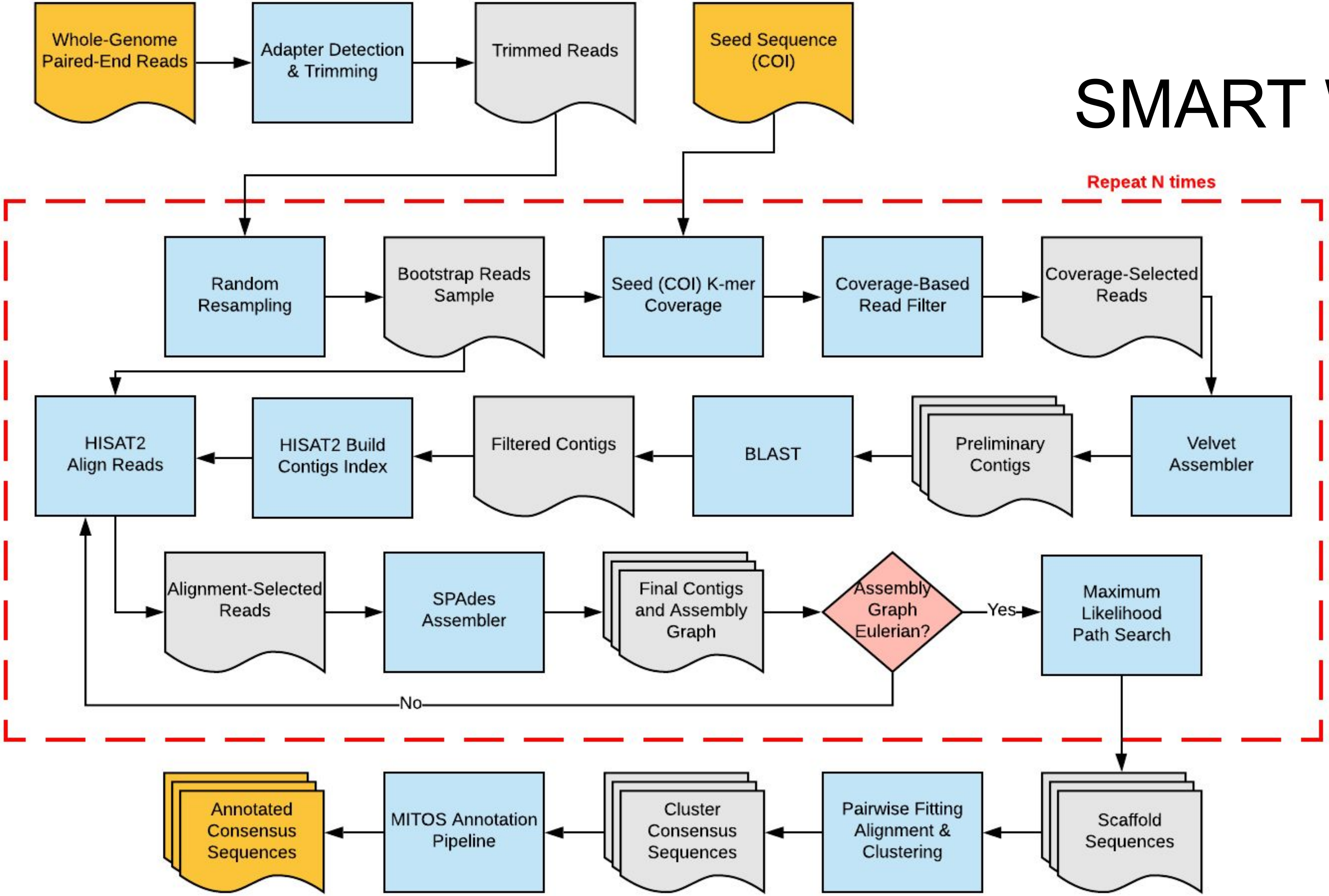
- Background/Motivation
- Related Work
- **Statistical Mitogenome Assembly with Repeats (SMART)**
    - **The pipeline**
    - Results
- Conclusion & Ongoing and Future Work

SMART

Statistical **M**itogenome **A**ssembly with **R**epea**T**s

- Input:
  1. Paired-end WGS reads
  2. Seed sequence (COI gene)
- Output:
  - Complete/circular mitogenome (or largest scaffold)

SMART Workflow



Whole-Genome Paired-End Reads

Adapter Detection & Trimming

Trimmed Reads

Seed Sequence (COI)

Repeat N times

Random Resampling

Bootstrap Reads Sample

Seed (COI) K-mer Coverage

Coverage-Based Read Filter

Coverage-Selected Reads

Velvet Assembler

Preliminary Contigs

BLAST

Filtered Contigs

HISAT2 Build Contigs Index

HISAT2 Align Reads

Alignment-Selected Reads

SPAdes Assembler

Final Contigs and Assembly Graph

Assembly Graph Eulerian?

Yes

No

Maximum Likelihood Path Search

Scaffold Sequences

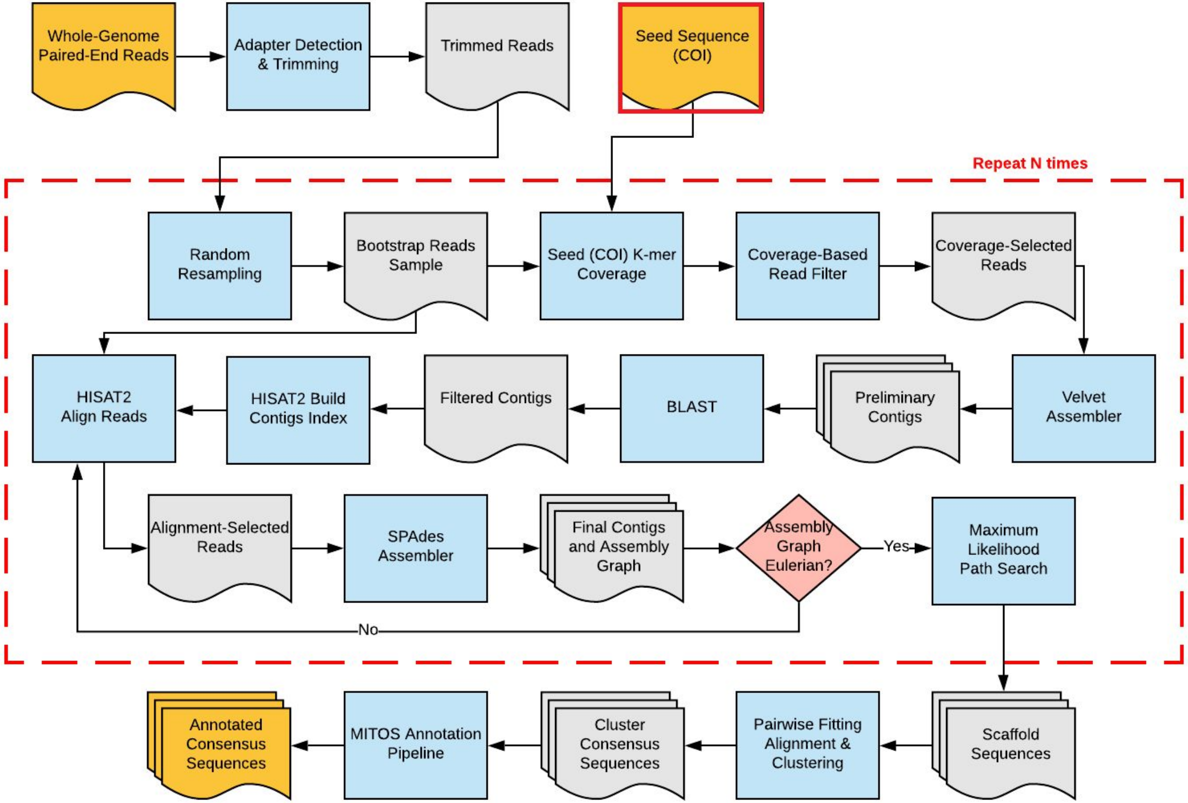
Pairwise Fitting Alignment & Clustering

Cluster Consensus Sequences

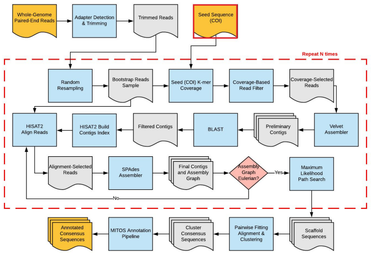
MITOS Annotation Pipeline

Annotated Consensus Sequences

Seed Selection


Whole-Genome Paired-End Reads
Adapter Detection & Trimming
Trimmed Reads
Seed Sequence (COI)
Repeat N times
Random Resampling
Bootstrap Reads Sample
Seed (COI) K-mer Coverage
Coverage-Based Read Filter
Coverage-Selected Reads
HISAT2 Align Reads
HISAT2 Build Contigs Index
Filtered Contigs
BLAST
Preliminary Contigs
Velvet Assembler
Alignment-Selected Reads
SPAdes Assembler
Final Contigs and Assembly Graph
Assembly Graph Eulerian?
Yes
Maximum Likelihood Path Search
No
Annotated Consensus Sequences
MITOS Annotation Pipeline
Cluster Consensus Sequences
Pairwise Fitting Alignment & Clustering
Scaffold Sequences

Seed Selection



Whole-Genome Paired-End Reads

Adapter Detection & Trimming

Trimmed Reads

Seed Sequence (COI)

Repeat N times

Random Resampling

Bootstrap Reads Sample

Seed (COI) K-mer Coverage

Coverage-Based Read Filter

Coverage-Selected Reads

Velvet Assembler

Preliminary Contigs

BLAST

Filtered Contigs

HISAT2 Build Contigs Index

HISAT2 Align Reads

Alignment-Selected Reads

SPAdes Assembler

Final Contigs and Assembly Graph

Assembly Graph Eulerian?

Yes

No

Maximum Likelihood Path Search

Scaffold Sequences

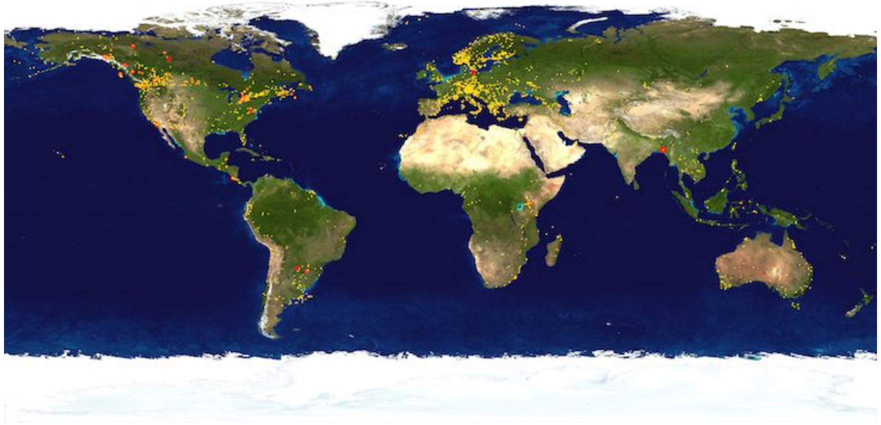
Pairwise Fitting Alignment & Clustering

Cluster Consensus Sequences

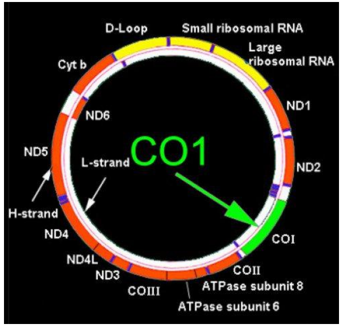
MITOS Annotation Pipeline

Annotated Consensus Sequences

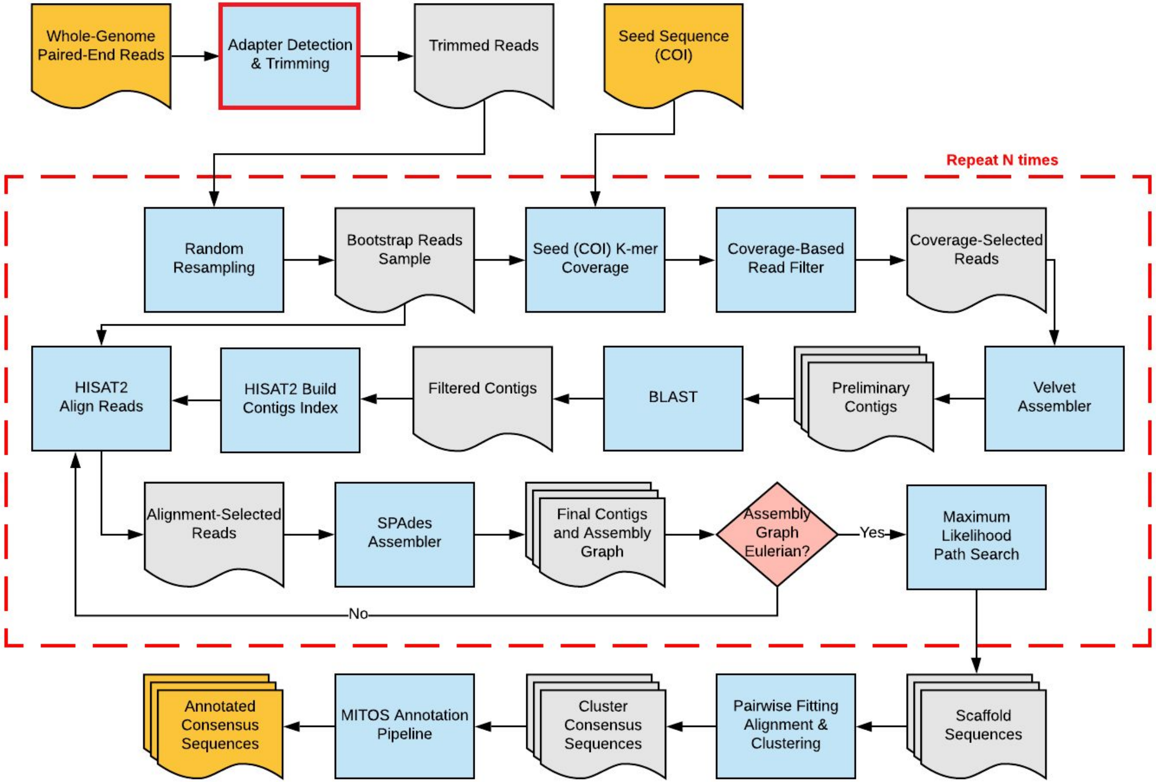
- Cytochrome c oxidase subunit 1 (COI) gene has been selected as a “DNA barcode” for taxonomic classification
- Barcode of Life Datasystem (BOLD) has > 1.4M public barcodes from 118,358K animal species



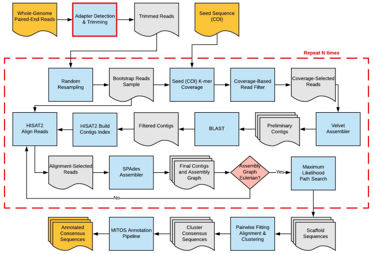
http://www.boldsystems.org/



Adapter Detection and Trimming


Whole-Genome Paired-End Reads
Adapter Detection & Trimming
Trimmed Reads
Seed Sequence (COI)
Repeat N times
Random Resampling
Bootstrap Reads Sample
Seed (COI) K-mer Coverage
Coverage-Based Read Filter
Coverage-Selected Reads
HISAT2 Align Reads
HISAT2 Build Contigs Index
Filtered Contigs
BLAST
Preliminary Contigs
Velvet Assembler
Alignment-Selected Reads
SPAdes Assembler
Final Contigs and Assembly Graph
Assembly Graph Eulerian?
Yes
Maximum Likelihood Path Search
No
Annotated Consensus Sequences
MITOS Annotation Pipeline
Cluster Consensus Sequences
Pairwise Fitting Alignment & Clustering
Scaffold Sequences

Adapter Detection and Trimming



Whole-Genome Paired-End Reads

Adapter Detection & Trimming

Trimmed Reads

Seed Sequence (COI)

Repeat N times

Random Resampling

Bootstrap Reads Sample

Seed (COI) K-mer Coverage

Coverage-Based Read Filter

Coverage-Selected Reads

Velvet Assembler

Preliminary Contigs

BLAST

Filtered Contigs

HISAT2 Build Contigs Index

HISAT2 Align Reads

Alignment-Selected Reads

SPAdes Assembler

Final Contigs and Assembly Graph

Assembly Graph Eulerian?

Yes

No

Maximum Likelihood Path Search

Scaffold Sequences

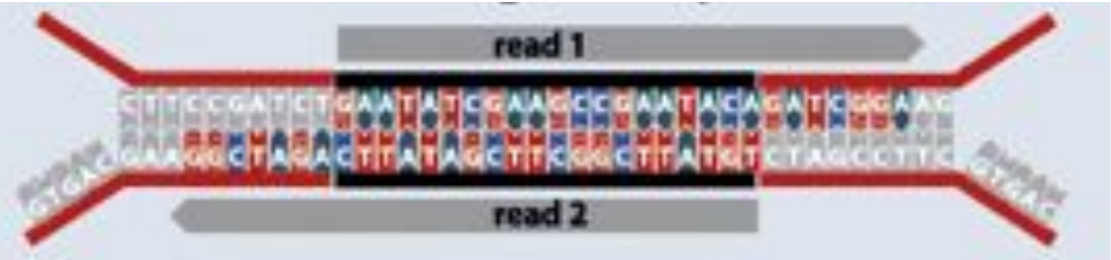
Pairwise Fitting Alignment & Clustering

Cluster Consensus Sequences

MITOS Annotation Pipeline

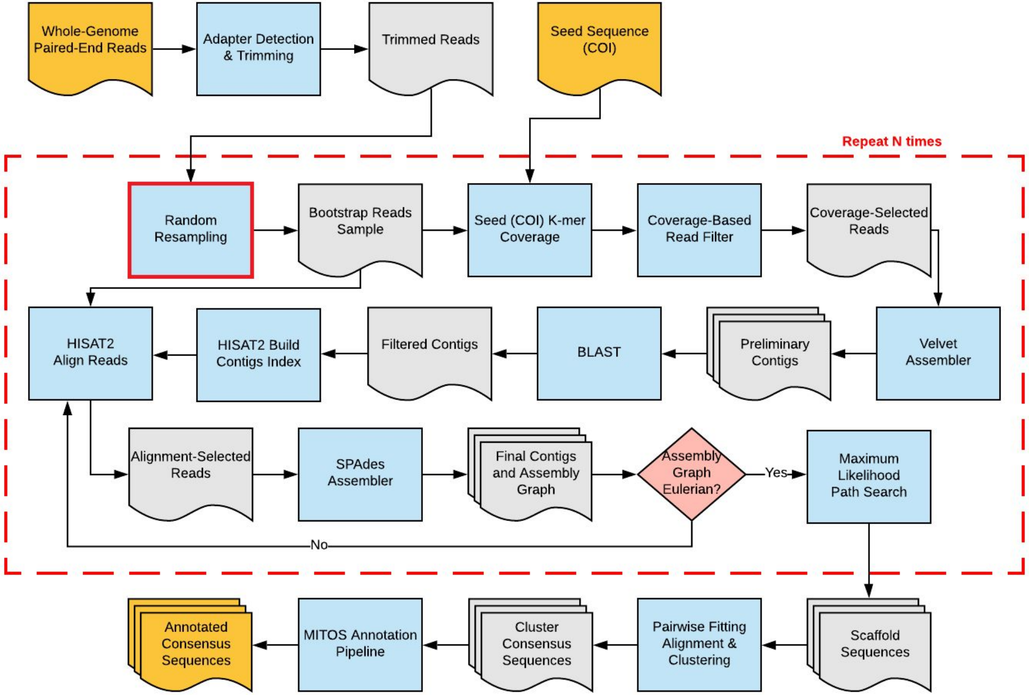
Annotated Consensus Sequences

- Automatic detection of adaptors and trimming using Perl/C++ modules from the IRFinder package

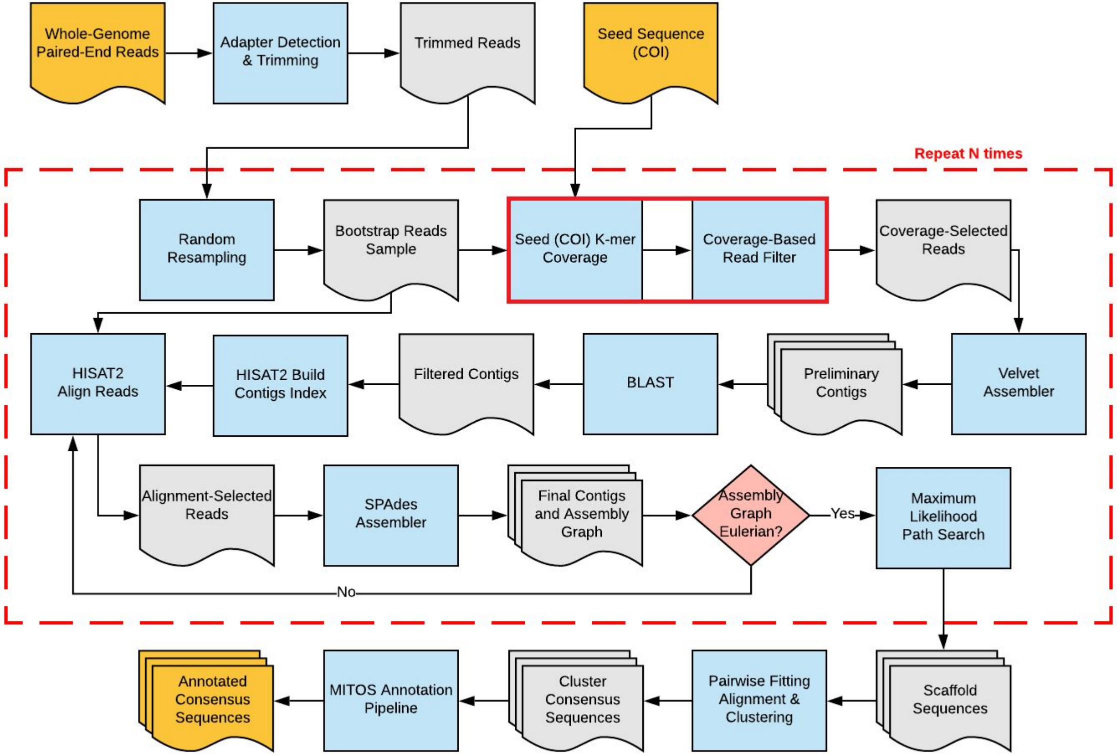


Middleton, Robert, et al. "IRFinder: assessing the impact of intron retention on mammalian gene expression." Genome biology 18.1 (2017): 51.

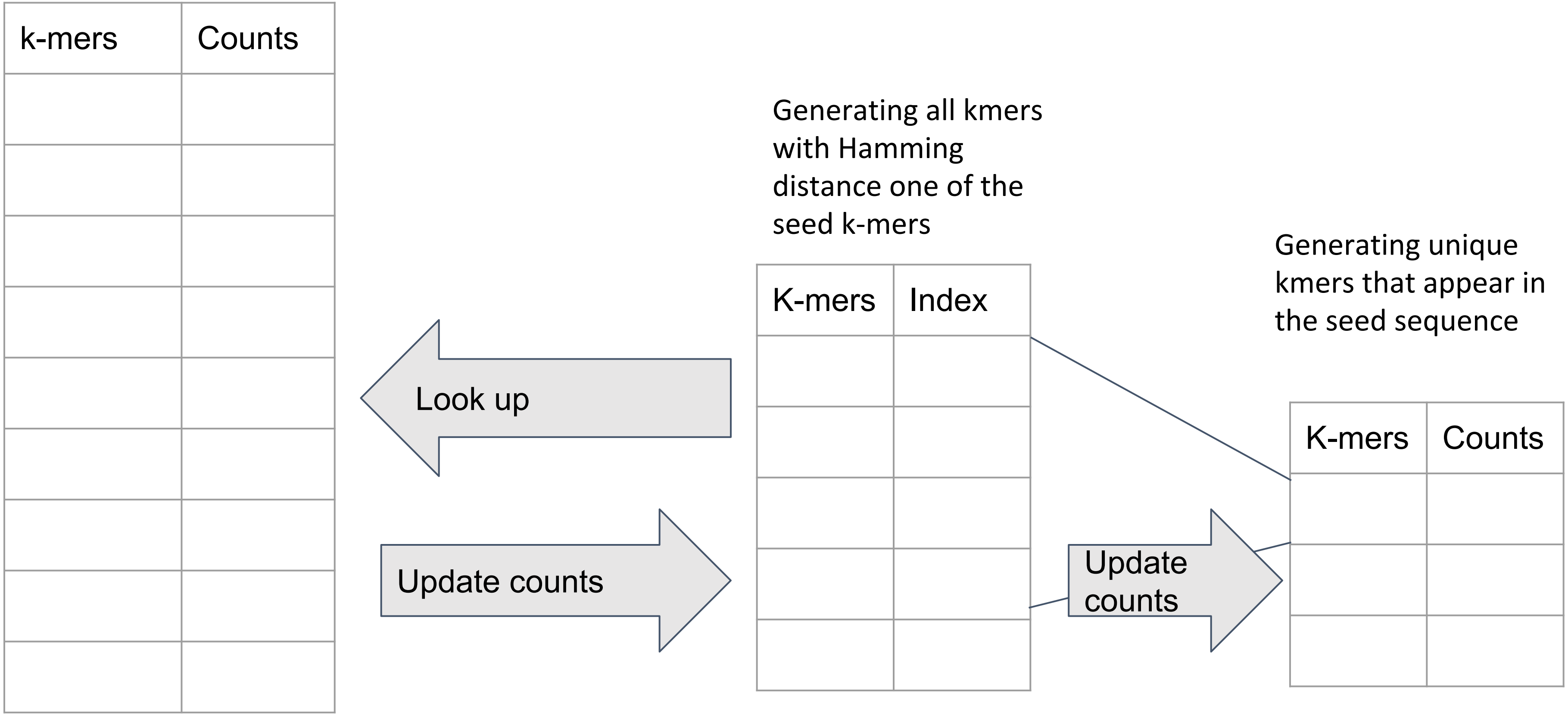
Random Read Re-sampling


Whole-Genome Paired-End Reads
Adapter Detection & Trimming
Trimmed Reads
Seed Sequence (COI)
Repeat N times
Random Resampling
Bootstrap Reads Sample
Seed (COI) K-mer Coverage
Coverage-Based Read Filter
Coverage-Selected Reads
Velvet Assembler
Preliminary Contigs
BLAST
Filtered Contigs
HISAT2 Build Contigs Index
HISAT2 Align Reads
Alignment-Selected Reads
SPAdes Assembler
Final Contigs and Assembly Graph
Assembly Graph Eulerian?
Yes
No
Maximum Likelihood Path Search
Scaffold Sequences
Pairwise Fitting Alignment & Clustering
Cluster Consensus Sequences
MITOS Annotation Pipeline
Annotated Consensus Sequences

Coverage-based Read Filtering


Whole-Genome Paired-End Reads
Adapter Detection & Trimming
Trimmed Reads
Seed Sequence (COI)
Repeat N times
Random Resampling
Bootstrap Reads Sample
Seed (COI) K-mer Coverage
Coverage-Based Read Filter
Coverage-Selected Reads
HISAT2 Align Reads
HISAT2 Build Contigs Index
Filtered Contigs
BLAST
Preliminary Contigs
Velvet Assembler
Alignment-Selected Reads
SPAdes Assembler
Final Contigs and Assembly Graph
Assembly Graph Eulerian?
Yes
No
Maximum Likelihood Path Search
Annotated Consensus Sequences
MITOS Annotation Pipeline
Cluster Consensus Sequences
Pairwise Fitting Alignment & Clustering
Scaffold Sequences

Counting number of times unique kmers appear in Bootstrap sample



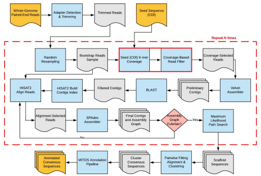
k-mers	Counts

Generating all kmers with Hamming distance one of the seed k-mers

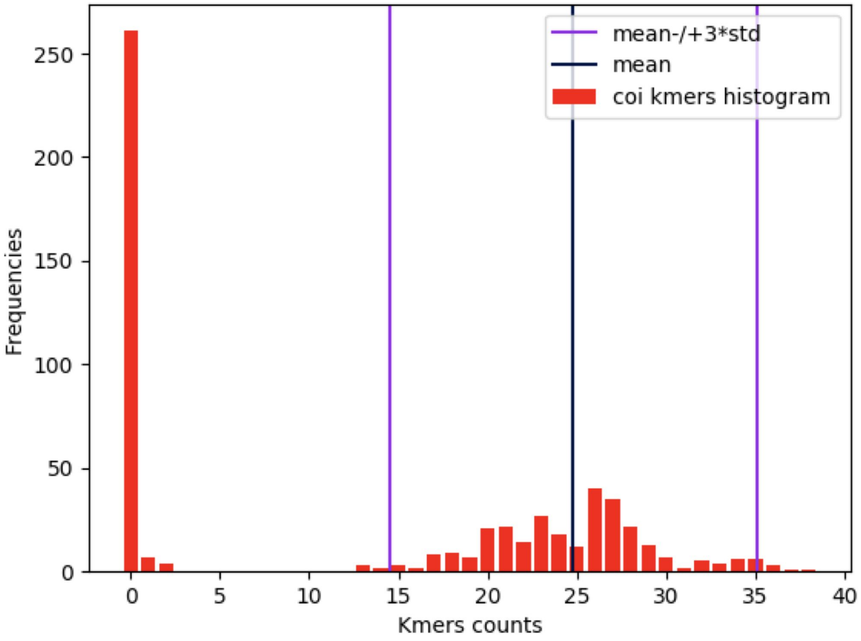
K-mers	Index

Generating unique kmers that appear in the seed sequence

K-mers	Counts


Whole-Genome Paired-End Reads
Adapter Detection & Trimming
Trimmed Reads
Seed Sequence (COI)
Repeat N times
Random Resampling
Bootstrap Reads Sample
Seed (COI) K-mer Coverage
Coverage-Based Read Filter
Coverage-Selected Reads
HISAT2 Align Reads
HISAT2 Build Contigs Index
Filtered Contigs
BLAST
Preliminary Contigs
Velvet Assembler
Alignment-Selected Reads
SPAdes Assembler
Final Contigs and Assembly Graph
Assembly Graph Eulerian?
Yes
No
Maximum Likelihood Path Search
Annotated Consensus Sequences
MITOS Annotation Pipeline
Cluster Consensus Sequences
Pairwise Fitting Alignment & Clustering
Scaffold Sequences

COI K-mers Counts Distribution



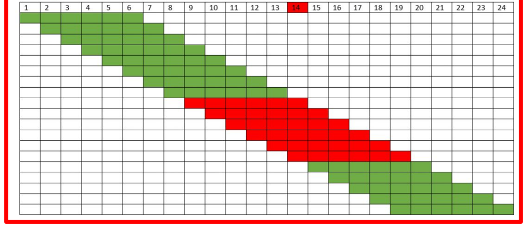
Two-component Gaussian mixture model to the one-dimensional distribution

Unique kmers appear
in Bootstrap sample

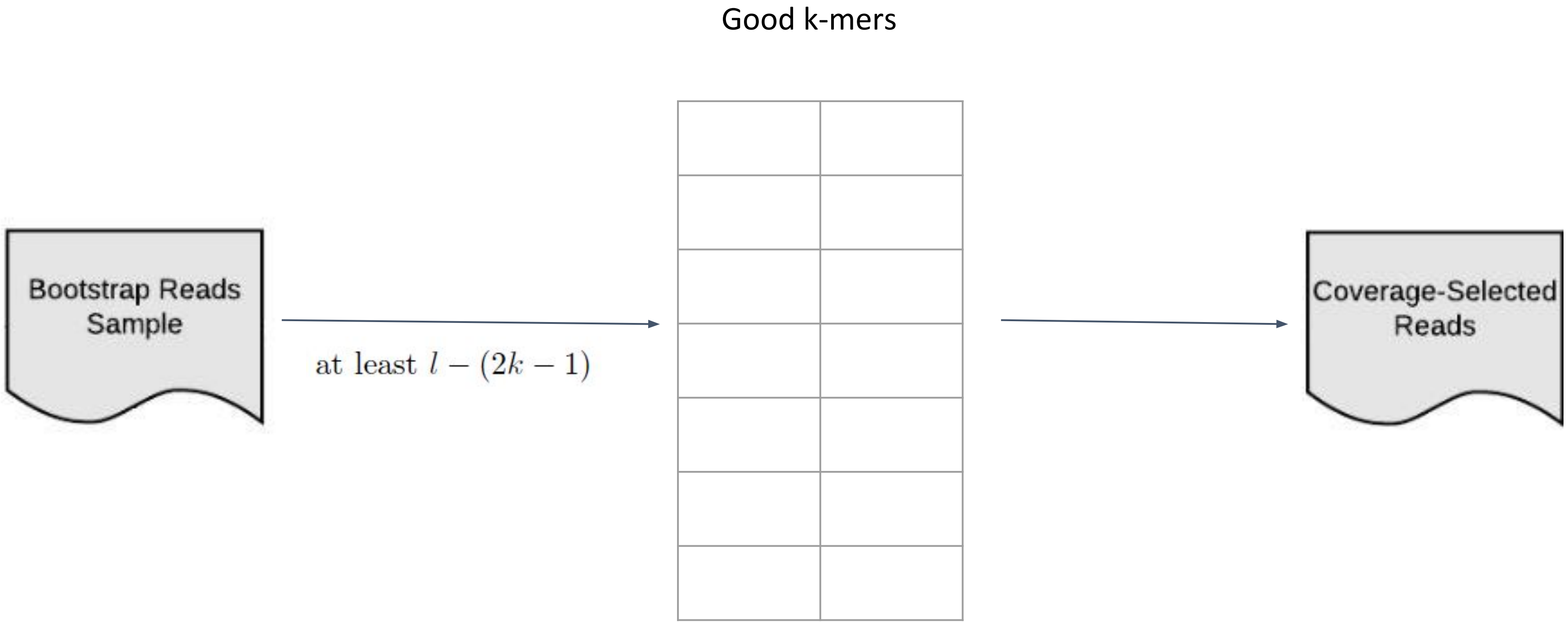
k-mers	Counts

$$|count(x) - \mu| \leq 3\sigma$$

Good k-mers



Reads with one sequencing error are kept



Good k-mers

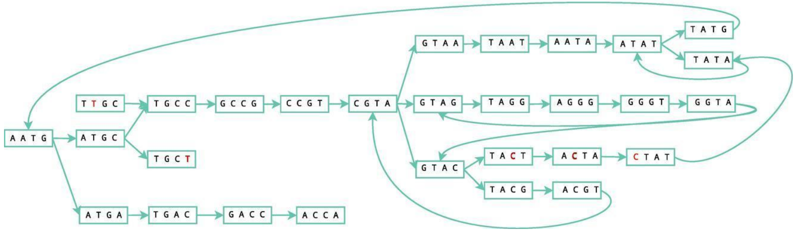
Bootstrap Reads Sample

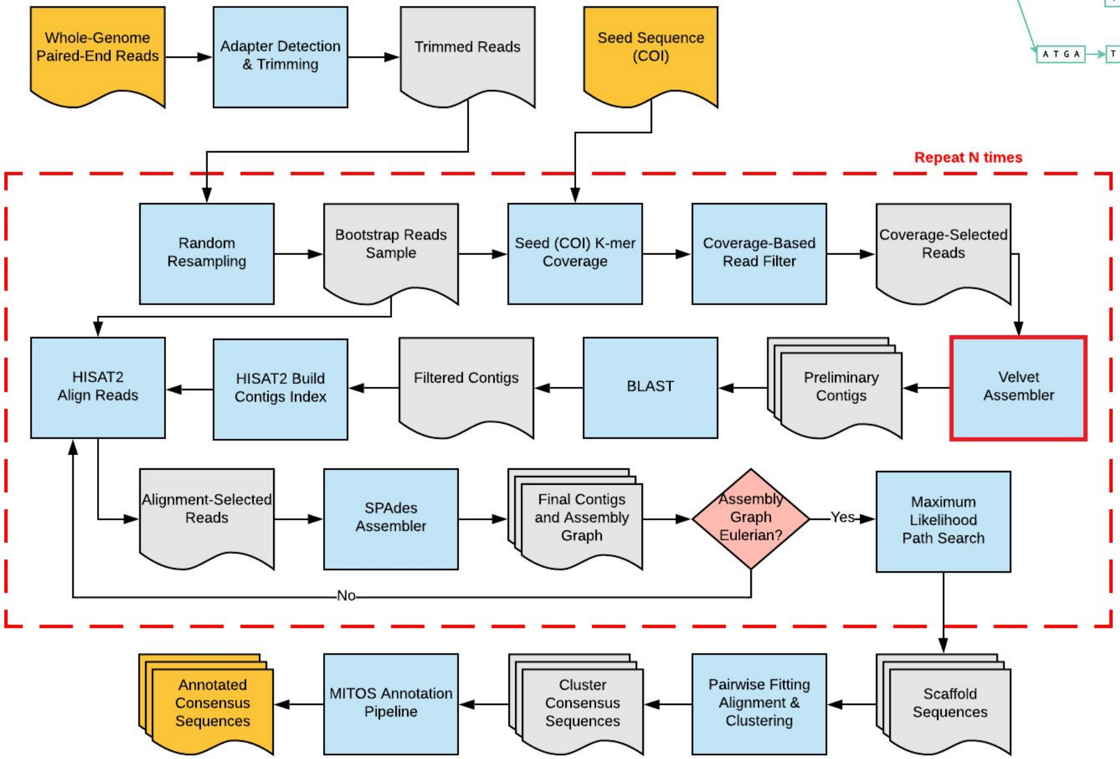
at least $l - (2k - 1)$

Coverage-Selected Reads

Preliminary Assembly

https://en.wikipedia.org/wiki/Velvet_assembler





Whole-Genome Paired-End Reads

Adapter Detection & Trimming

Trimmed Reads

Seed Sequence (COI)

Repeat N times

Random Resampling

Bootstrap Reads Sample

Seed (COI) K-mer Coverage

Coverage-Based Read Filter

Coverage-Selected Reads

Velvet Assembler

Preliminary Contigs

BLAST

Filtered Contigs

HISAT2 Build Contigs Index

HISAT2 Align Reads

Alignment-Selected Reads

SPAdes Assembler

Final Contigs and Assembly Graph

Assembly Graph Eulerian?

Yes

No

Maximum Likelihood Path Search

Scaffold Sequences

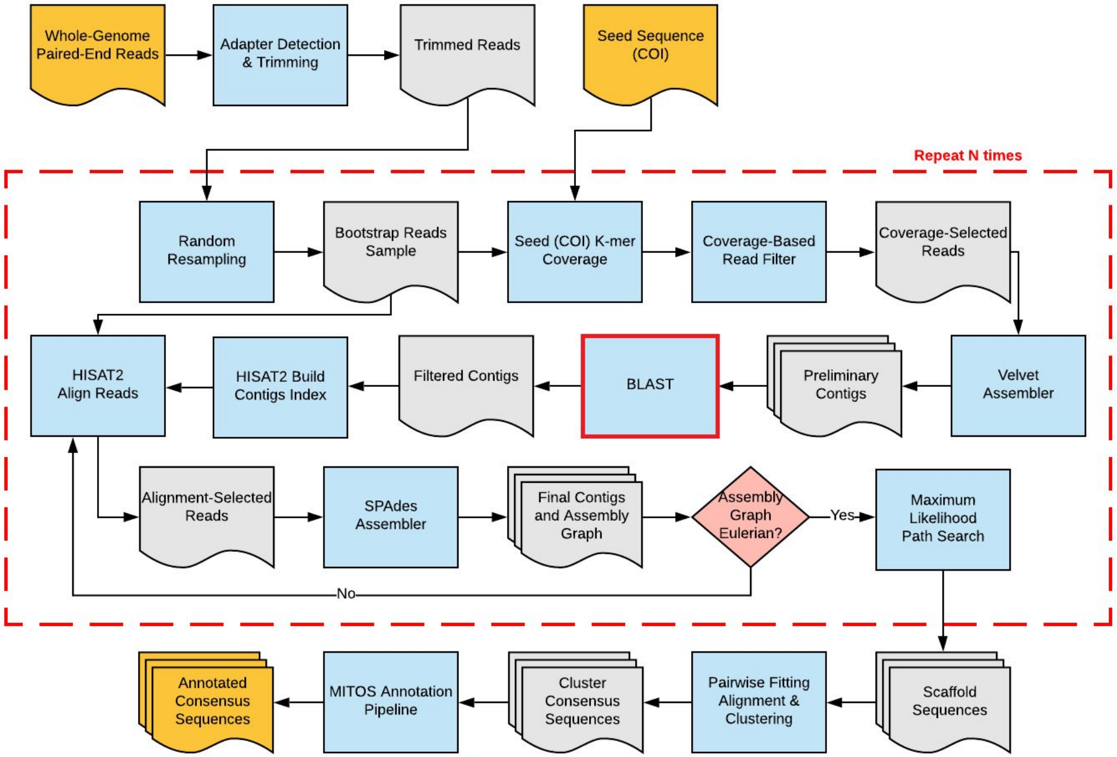
Pairwise Fitting Alignment & Clustering

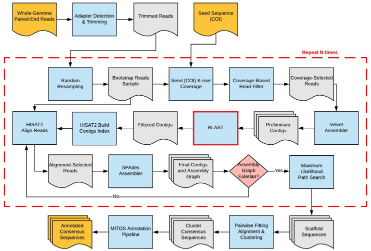
Cluster Consensus Sequences

MITOS Annotation Pipeline

Annotated Consensus Sequences

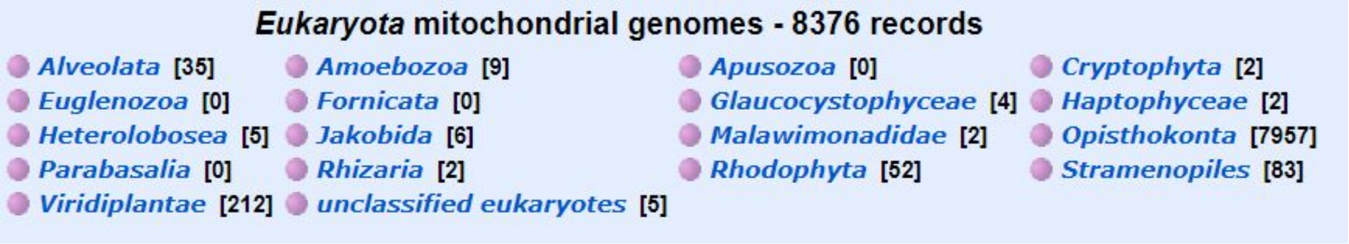
Preliminary Contig Filtering


Whole-Genome Paired-End Reads
Adapter Detection & Trimming
Trimmed Reads
Seed Sequence (COI)
Repeat N times
Random Resampling
Bootstrap Reads Sample
Seed (COI) K-mer Coverage
Coverage-Based Read Filter
Coverage-Selected Reads
HISAT2 Align Reads
HISAT2 Build Contigs Index
Filtered Contigs
BLAST
Preliminary Contigs
Velvet Assembler
Alignment-Selected Reads
SPAdes Assembler
Final Contigs and Assembly Graph
Assembly Graph Eulerian?
Yes
Maximum Likelihood Path Search
No
Annotated Consensus Sequences
MITOS Annotation Pipeline
Cluster Consensus Sequences
Pairwise Fitting Alignment & Clustering
Scaffold Sequences


Whole-Genome Paired-End Reads
Adapter Detection & Trimming
Trimmed Reads
Seed Sequence (COI)
Repeat N times
Random Resampling
Bootstrap Reads Sample
Seed (COI) K-mer Coverage
Coverage-Based Read Filter
Coverage-Selected Reads
HISAT2 Align Reads
HISAT2 Build Contigs Index
Filtered Contigs
BLAST
Preliminary Contigs
Velvet Assembler
Alignment-Selected Reads
SPAdes Assembler
Final Contigs and Assembly Graph
Assembly Graph Eulerian?
Yes
Maximum Likelihood Path Search
No
Annotated Consensus Sequences
MITOS Annotation Pipeline
Cluster Consensus Sequences
Pairwise Fitting Alignment & Clustering
Scaffold Sequences

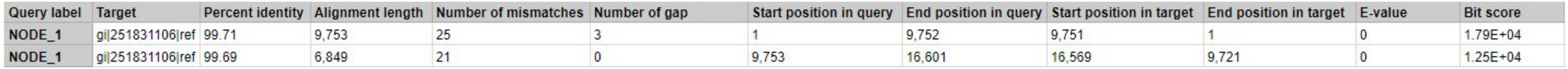
Preliminary Contig Filtering

- Contigs aligned against a local database eukaryotic mitogenomes using nucleotide-nucleotide BLAST
    - Keep contigs that have hits with E-value of 10^{-10} or less



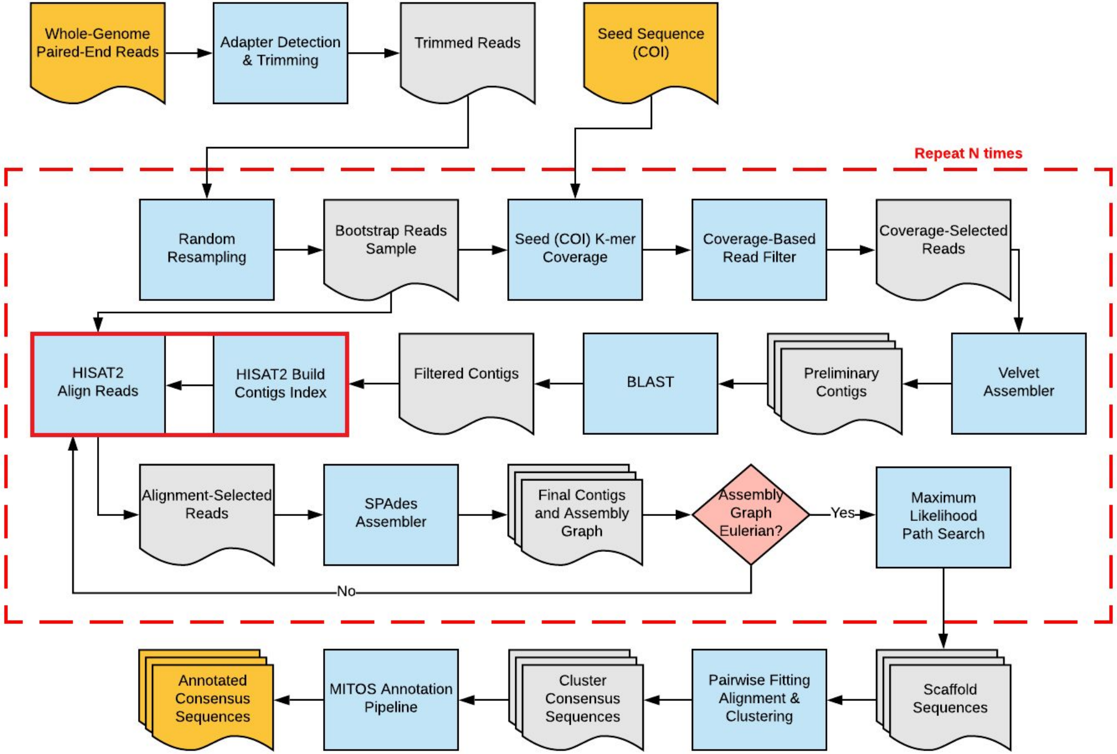
***Eukaryota* mitochondrial genomes - 8376 records**

● *Alveolata* [35]	● *Amoebozoa* [9]	● *Apusozoa* [0]	● *Cryptophyta* [2]
● *Euglenozoa* [0]	● *Fornicata* [0]	● *Glaucocystophyceae* [4]	● *Haptophyceae* [2]
● *Heterolobosea* [5]	● *Jakobida* [6]	● *Malawimonadidae* [2]	● *Opisthokonta* [7957]
● *Parabasalia* [0]	● *Rhizaria* [2]	● *Rhodophyta* [52]	● *Stramenopiles* [83]
● *Viridiplantae* [212]	● *unclassified eukaryotes* [5]		



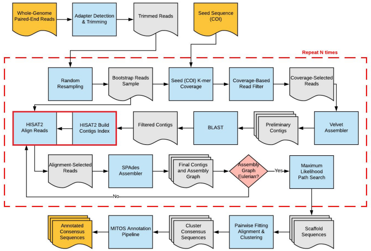
Query label	Target	Percent identity	Alignment length	Number of mismatches	Number of gap	Start position in query	End position in query	Start position in target	End position in target	E-value	Bit score
NODE_1	gi|251831106|ref	99.71	9,753	25	3	1	9,752	9,751	1	0	1.79E+04
NODE_1	gi|251831106|ref	99.69	6,849	21	0	9,753	16,601	16,569	9,721	0	1.25E+04

Alignment-based Read Filtering

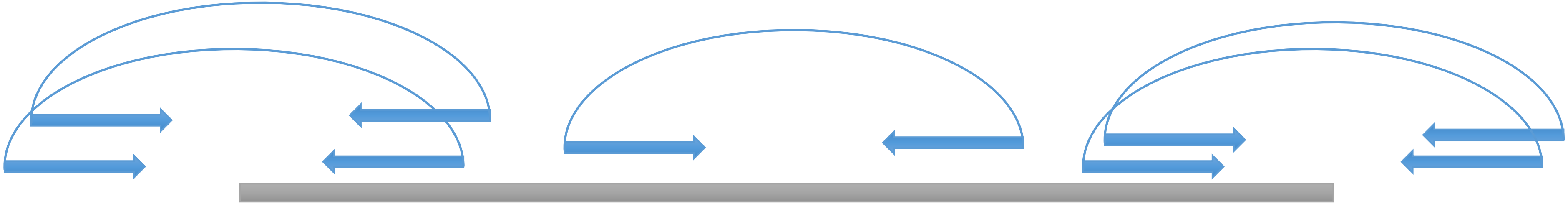


Whole-Genome Paired-End Reads
Adapter Detection & Trimming
Trimmed Reads
Seed Sequence (COI)
Repeat N times
Random Resampling
Bootstrap Reads Sample
Seed (COI) K-mer Coverage
Coverage-Based Read Filter
Coverage-Selected Reads
HISAT2 Align Reads
HISAT2 Build Contigs Index
Filtered Contigs
BLAST
Preliminary Contigs
Velvet Assembler
Alignment-Selected Reads
SPAdes Assembler
Final Contigs and Assembly Graph
Assembly Graph Eulerian?
Yes
No
Maximum Likelihood Path Search
Annotated Consensus Sequences
MITOS Annotation Pipeline
Cluster Consensus Sequences
Pairwise Fitting Alignment & Clustering
Scaffold Sequences

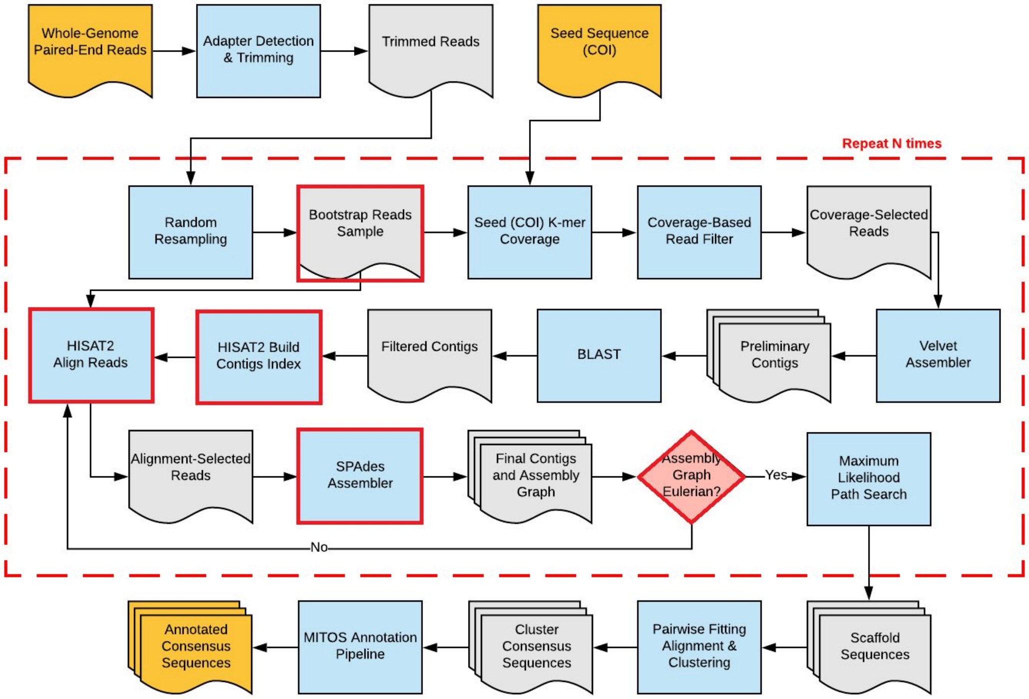
Alignment-based Read Filtering


Whole-Genome Paired-End Reads
Adapter Detection & Trimming
Trimmed Reads
Seed Sequence (COI)
Repeat N times
Random Resampling
Bootstrap Reads Sample
Seed (COI) K-mer Coverage
Coverage-Based Read Filter
Coverage-Selected Reads
HISAT2 Align Reads
HISAT2 Build Contigs Index
Filtered Contigs
BLAST
Preliminary Contigs
Velvet Assembler
Alignment-Selected Reads
SPAdes Assembler
Final Contigs and Assembly Graph
Assembly Graph Eulerian?
Yes
No
Maximum Likelihood Path Search
Annotated Consensus Sequences
MITOS Annotation Pipeline
Cluster Consensus Sequences
Pairwise Fitting Alignment & Clustering
Scaffold Sequences

- Using HISAT2
  - Fast and sensitive aligner for NGS reads
- Pulls out the read pairs that have at least one of the reads aligned



Secondary Assembly



Whole-Genome Paired-End Reads

Adapter Detection & Trimming

Trimmed Reads

Seed Sequence (COI)

Repeat N times

Random Resampling

Bootstrap Reads Sample

Seed (COI) K-mer Coverage

Coverage-Based Read Filter

Coverage-Selected Reads

HISAT2 Align Reads

HISAT2 Build Contigs Index

Filtered Contigs

BLAST

Preliminary Contigs

Velvet Assembler

Alignment-Selected Reads

SPAdes Assembler

Final Contigs and Assembly Graph

Assembly Graph Eulerian?

Yes

No

Maximum Likelihood Path Search

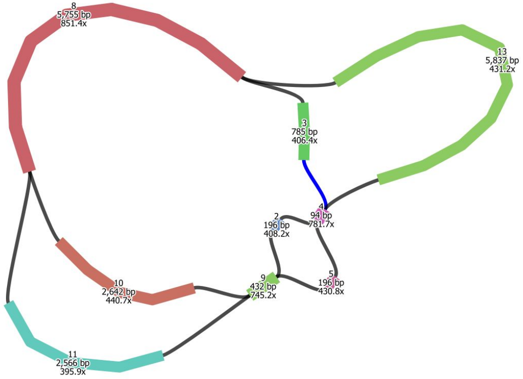
Annotated Consensus Sequences

MITOS Annotation Pipeline

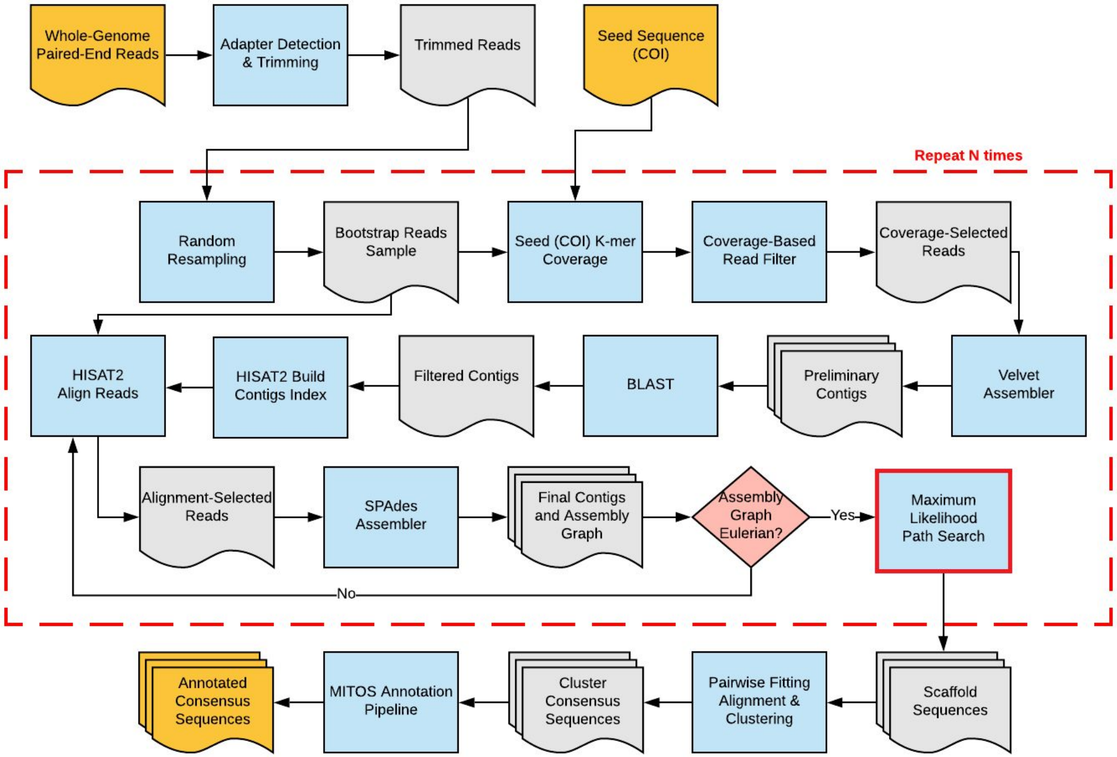
Cluster Consensus Sequences

Pairwise Fitting Alignment & Clustering

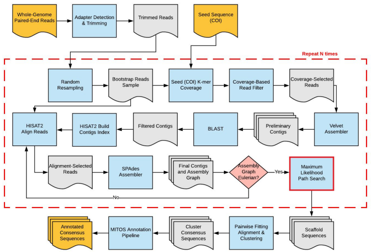
Scaffold Sequences



Scaffolding


Whole-Genome Paired-End Reads
Adapter Detection & Trimming
Trimmed Reads
Seed Sequence (COI)
Repeat N times
Random Resampling
Bootstrap Reads Sample
Seed (COI) K-mer Coverage
Coverage-Based Read Filter
Coverage-Selected Reads
Velvet Assembler
Preliminary Contigs
BLAST
Filtered Contigs
HISAT2 Build Contigs Index
HISAT2 Align Reads
Alignment-Selected Reads
SPAdes Assembler
Final Contigs and Assembly Graph
Assembly Graph Eulerian?
Yes
No
Maximum Likelihood Path Search
Scaffold Sequences
Pairwise Fitting Alignment & Clustering
Cluster Consensus Sequences
MITOS Annotation Pipeline
Annotated Consensus Sequences

Scaffolding



Whole-Genome Paired-End Reads

Adapter Detection & Trimming

Trimmed Reads

Seed Sequence (COI)

Repeat N times

Random Resampling

Bootstrap Reads Sample

Seed (COI) K-mer Coverage

Coverage-Based Read Filter

Coverage-Selected Reads

Velvet Assembler

Preliminary Contigs

BLAST

Filtered Contigs

HISAT2 Build Contigs Index

HISAT2 Align Reads

Alignment-Selected Reads

SPAdes Assembler

Final Contigs and Assembly Graph

Assembly Graph Eulerian?

Yes

No

Maximum Likelihood Path Search

Scaffold Sequences

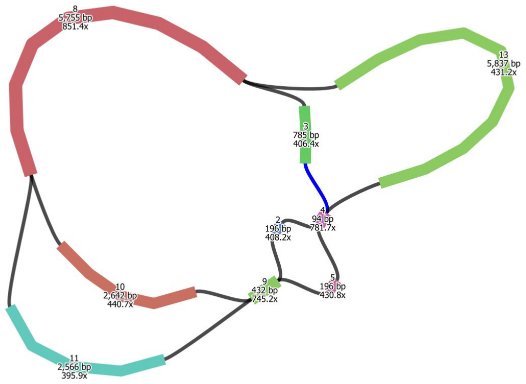
Pairwise Fitting Alignment & Clustering

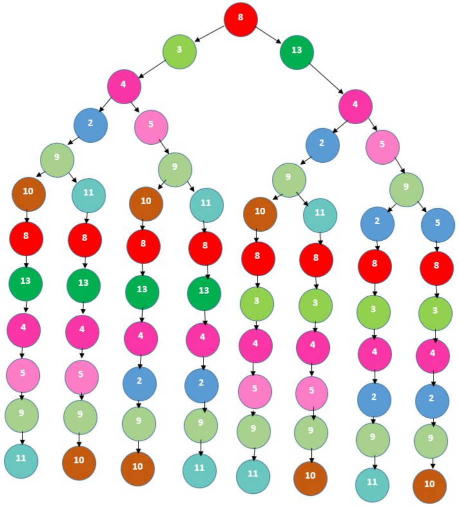
Cluster Consensus Sequences

MITOS Annotation Pipeline

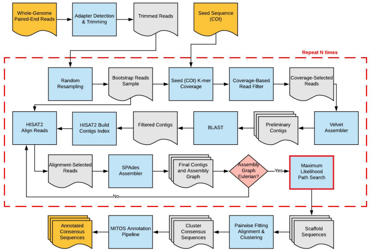
Annotated Consensus Sequences

- Eulerian paths evaluated using likelihood model implemented in ALE [Clark et al 2013]





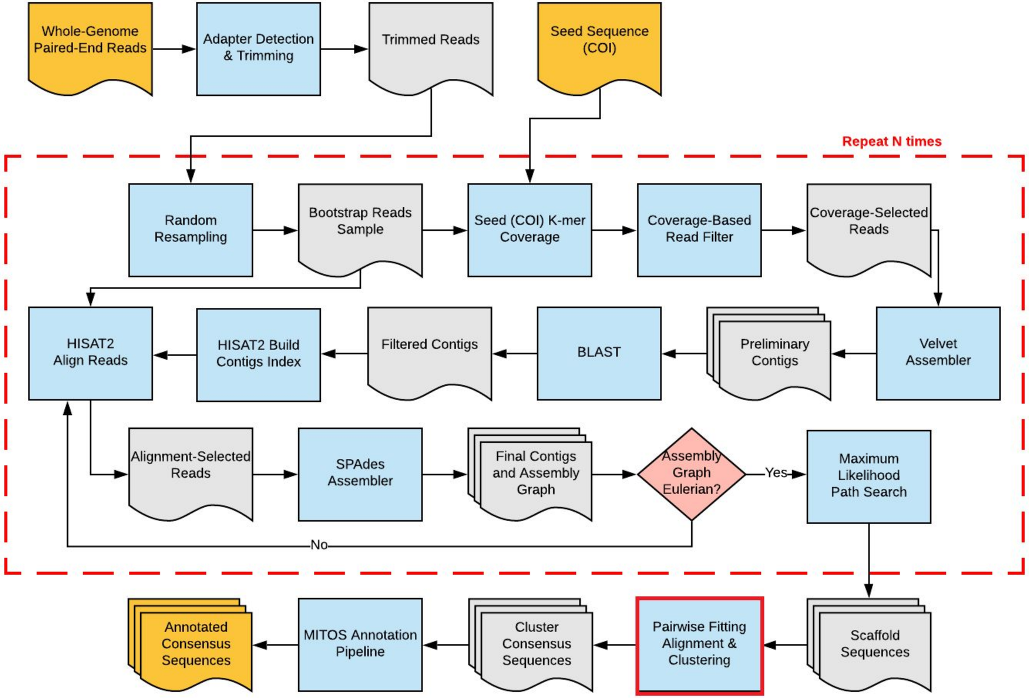
ALE likelihood

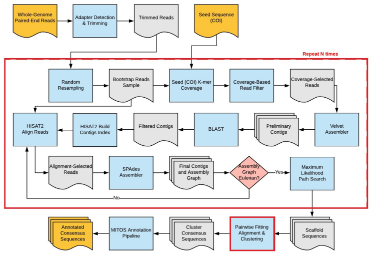

Whole-Genome Paired-End Reads
Adapter Detection & Trimming
Trimmed Reads
Seed Sequence (COI)
Repeat N times
Random Resampling
Bootstrap Reads Sample
Seed (COI) K-mer Coverage
Coverage-Based Read Filter
Coverage-Selected Reads
HISAT2 Align Reads
HISAT2 Build Contigs Index
Filtered Contigs
BLAST
Preliminary Contigs
Velvet Assembler
Alignment-Selected Reads
SPAdes Assembler
Final Contigs and Assembly Graph
Assembly Graph Eulerian?
Yes
Maximum Likelihood Path Search
No
Annotated Consensus Sequences
MITOS Annotation Pipeline
Cluster Consensus Sequences
Pairwise Fitting Alignment & Clustering
Scaffold Sequences

- **Placement scoring:**
  - How well read sequences agree with the assembly
- **Insert scoring:**
  - How well PE insert lengths match those we would expect
- **Depth scoring:**
  - How well depth at each location agrees with depth expected after GC-bias correction
- **K-mer scoring:**
  - How well k-mer counts of each contig match multinomial distribution estimated from entire assembly

https://academic.oup.com/bioinformatics/article/29/4/435/199222

Clustering

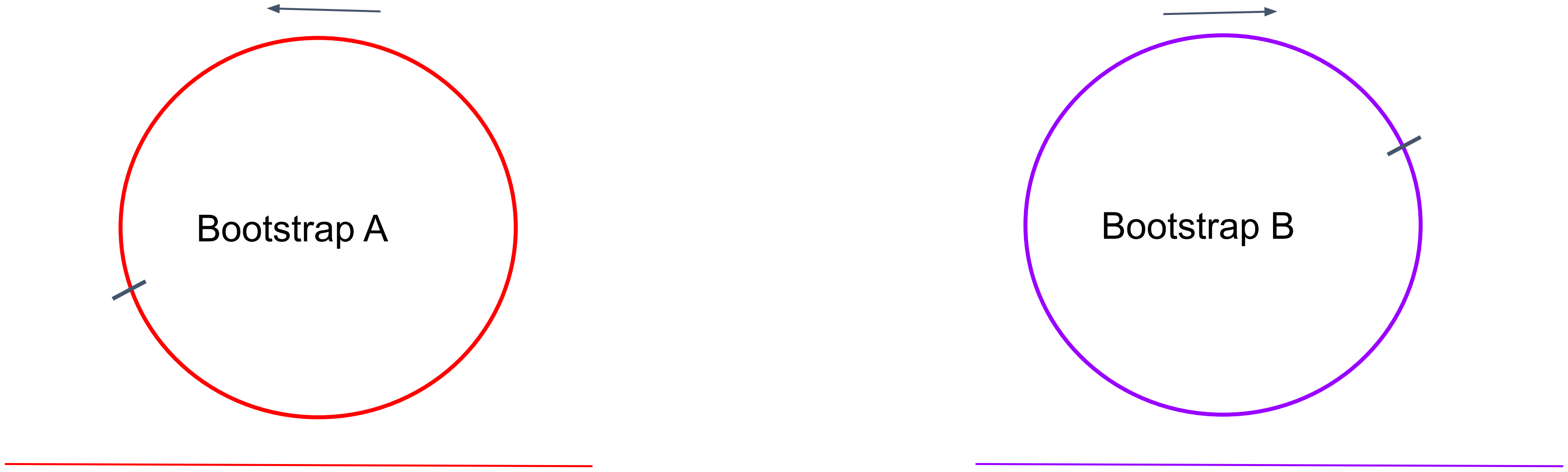

Whole-Genome Paired-End Reads
Adapter Detection & Trimming
Trimmed Reads
Seed Sequence (COI)
Repeat N times
Random Resampling
Bootstrap Reads Sample
Seed (COI) K-mer Coverage
Coverage-Based Read Filter
Coverage-Selected Reads
HISAT2 Align Reads
HISAT2 Build Contigs Index
Filtered Contigs
BLAST
Preliminary Contigs
Velvet Assembler
Alignment-Selected Reads
SPAdes Assembler
Final Contigs and Assembly Graph
Assembly Graph Eulerian?
Yes
Maximum Likelihood Path Search
No
Annotated Consensus Sequences
MITOS Annotation Pipeline
Cluster Consensus Sequences
Pairwise Fitting Alignment & Clustering
Scaffold Sequences



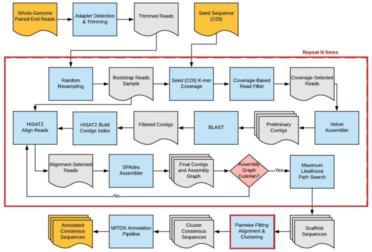
Whole-Genome Paired-End Reads
Adapter Detection & Trimming
Trimmed Reads
Seed Sequence (COI)
Repeat N times
Random Resampling
Bootstrap Reads Sample
Seed (COI) K-mer Coverage
Coverage-Based Read Filter
Coverage-Selected Reads
Velvet Assembler
Preliminary Contigs
BLAST
Filtered Contigs
HISAT2 Build Contigs Index
HISAT2 Align Reads
Alignment-Selected Reads
SPAdes Assembler
Final Contigs and Assembly Graph
Assembly Graph Eulerian?
Yes
No
Maximum Likelihood Path Search
Scaffold Sequences
Pairwise Fitting Alignment & Clustering
Cluster Consensus Sequences
MITOS Annotation Pipeline
Annotated Consensus Sequences

Clustering

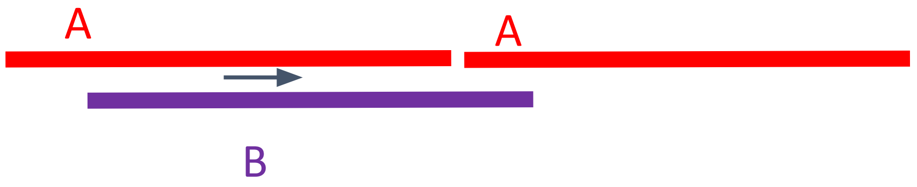
- Process repeated for n bootstrap samples
  - Pairwise distances computed using fitting alignment
    - Rotation invariant
    - Direction invariant

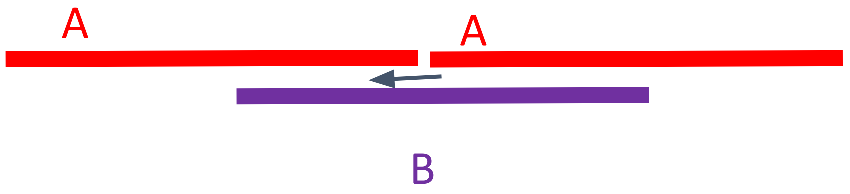


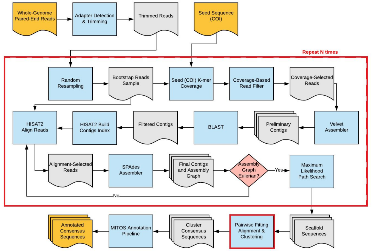
Clustering


Whole-Genome Paired-End Reads
Adapter Detection & Trimming
Trimmed Reads
Seed Sequence (COI)
Repeat N times
Random Resampling
Bootstrap Reads Sample
Seed (COI) K-mer Coverage
Coverage-Based Read Filter
Coverage-Selected Reads
HISAT2 Align Reads
HISAT2 Build Contigs Index
Filtered Contigs
BLAST
Preliminary Contigs
Velvet Assembler
Alignment-Selected Reads
SPAdes Assembler
Final Contigs and Assembly Graph
Assembly Graph Eulerian?
Yes
No
Maximum Likelihood Path Search
Annotated Consensus Sequences
MITOS Annotation Pipeline
Cluster Consensus Sequences
Pairwise Fitting Alignment & Clustering
Scaffold Sequences

- If bootstrap A is longer than bootstrap B, we duplicate the longest sequence.
- Use the both shortest sequence and its Watson-Crick complement

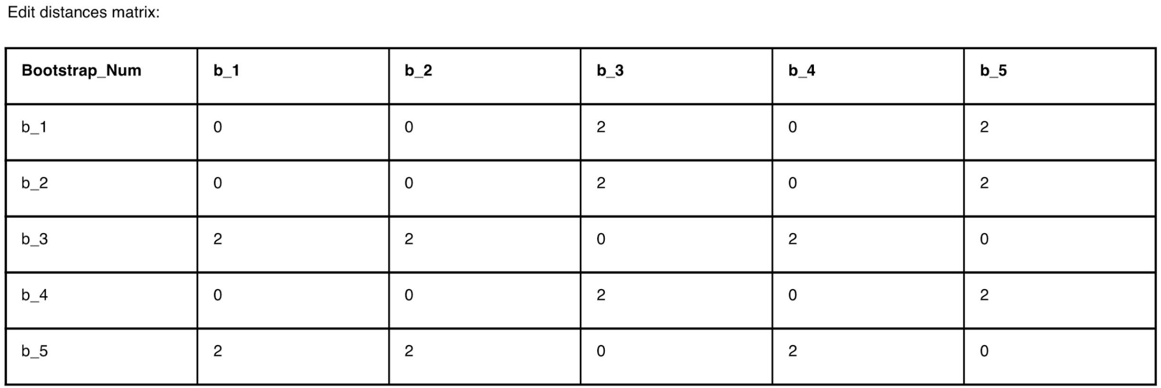





Whole-Genome Paired-End Reads
Adapter Detection & Trimming
Trimmed Reads
Seed Sequence (COI)
Repeat N times
Random Resampling
Bootstrap Reads Sample
Seed (COI) K-mer Coverage
Coverage-Based Read Filter
Coverage-Selected Reads
HISAT2 Align Reads
HISAT2 Build Contigs Index
Filtered Contigs
BLAST
Preliminary Contigs
Velvet Assembler
Alignment-Selected Reads
SPAdes Assembler
Final Contigs and Assembly Graph
Assembly Graph Eulerian?
Yes
No
Maximum Likelihood Path Search
Annotated Consensus Sequences
MITOS Annotation Pipeline
Cluster Consensus Sequences
Pairwise Fitting Alignment & Clustering
Scaffold Sequences

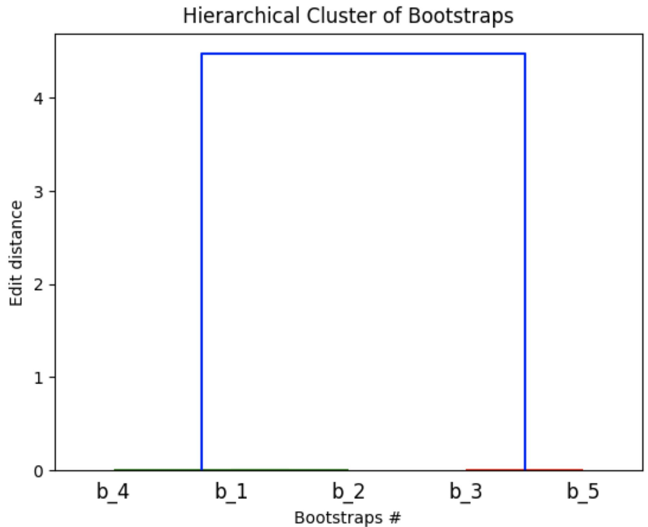
Clustering

- Using hierarchical clustering on the edit distance matrix
- A consensus sequences is generated for each cluster

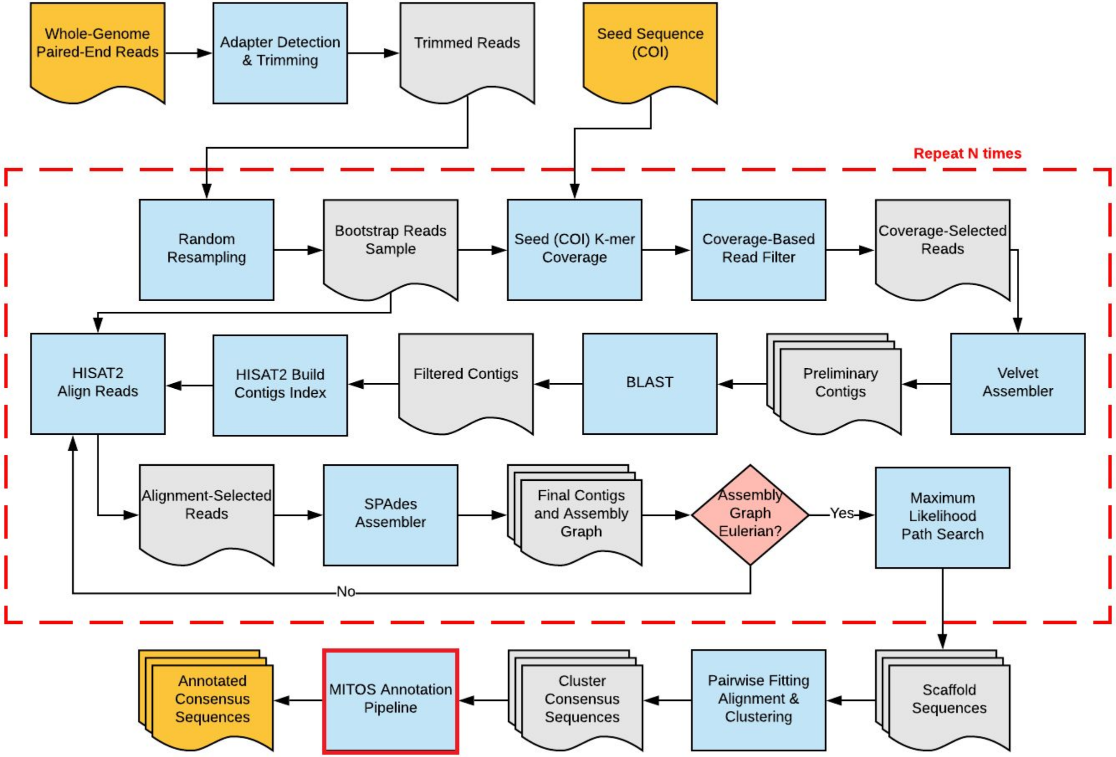


Edit distances matrix:

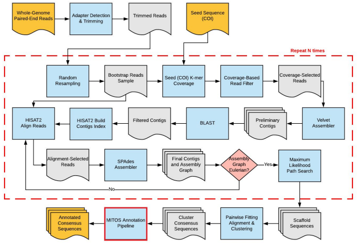
Bootstrap_Num	b_1	b_2	b_3	b_4	b_5
b_1	0	0	2	0	2
b_2	0	0	2	0	2
b_3	2	2	0	2	0
b_4	0	0	2	0	2
b_5	2	2	0	2	0

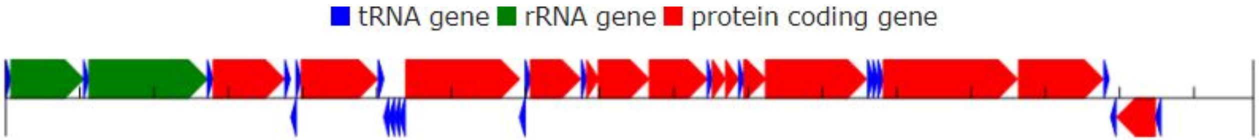


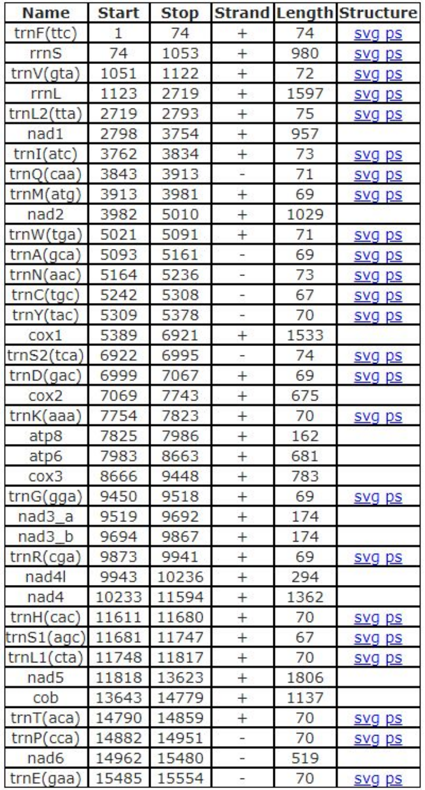
Annotation


Whole-Genome Paired-End Reads
Adapter Detection & Trimming
Trimmed Reads
Seed Sequence (COI)
Repeat N times
Random Resampling
Bootstrap Reads Sample
Seed (COI) K-mer Coverage
Coverage-Based Read Filter
Coverage-Selected Reads
HISAT2 Align Reads
HISAT2 Build Contigs Index
Filtered Contigs
BLAST
Preliminary Contigs
Velvet Assembler
Alignment-Selected Reads
SPAdes Assembler
Final Contigs and Assembly Graph
Assembly Graph Eulerian?
Yes
Maximum Likelihood Path Search
No
Annotated Consensus Sequences
MITOS Annotation Pipeline
Cluster Consensus Sequences
Pairwise Fitting Alignment & Clustering
Scaffold Sequences

MITOS annotation

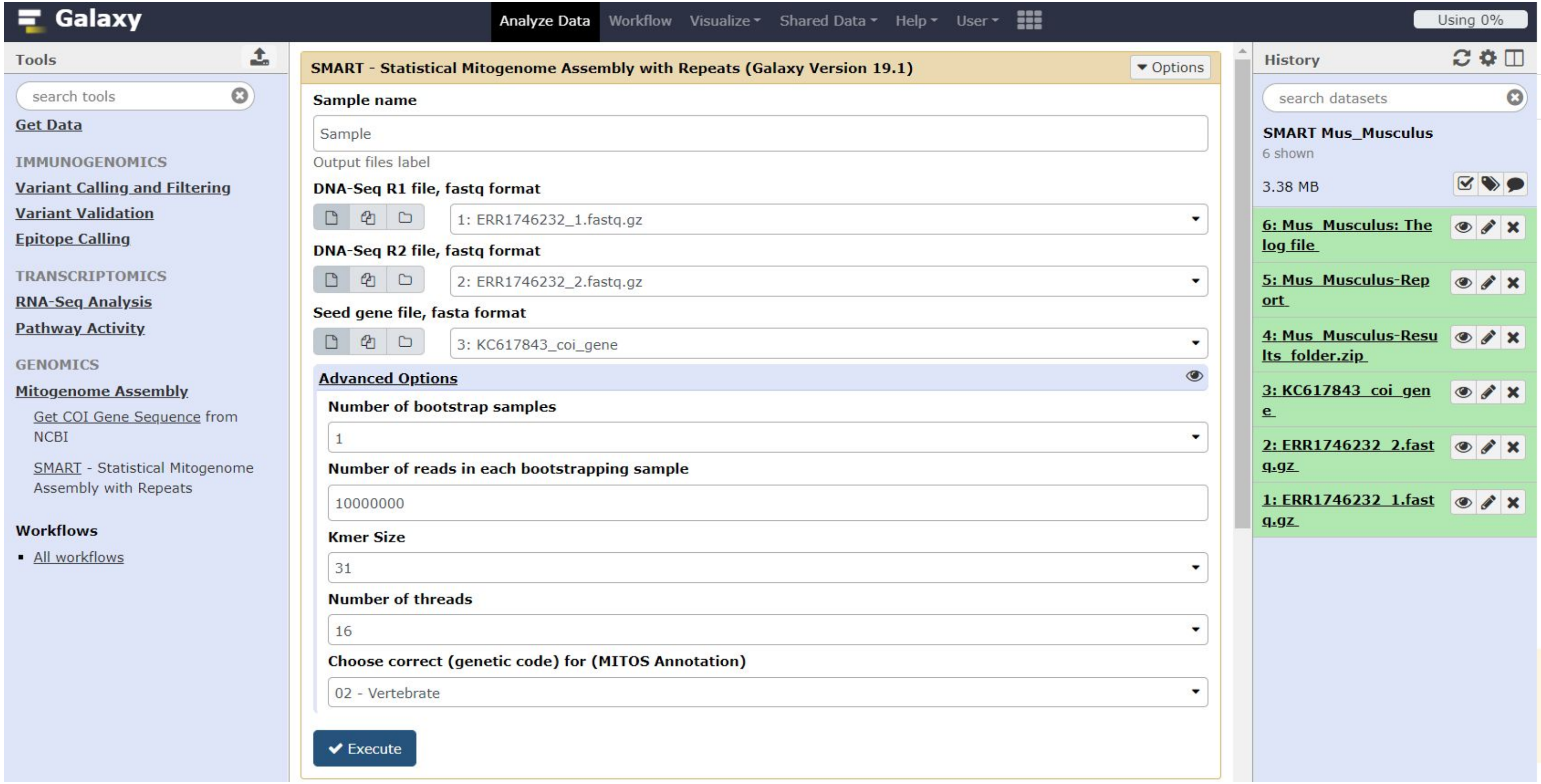

Whole-Genome Paired-End Reads
Adapter Detection & Trimming
Trimmed Reads
Seed Sequence (COI)
Repeat N times
Random Resampling
Bootstrap Reads Sample
Seed (COI) K-mer Coverage
Coverage-Based Read Filter
Coverage-Selected Reads
Velvet Assembler
Preliminary Contigs
BLAST
Filtered Contigs
HISAT2 Build Contigs Index
HISAT2 Align Reads
Alignment-Selected Reads
SPAdes Assembler
Final Contigs and Assembly Graph
Assembly Graph Eulerian?
Yes
No
Maximum Likelihood Path Search
Scaffold Sequences
Pairwise Fitting Alignment & Clustering
Cluster Consensus Sequences
MITOS Annotation Pipeline
Annotated Consensus Sequences





Name	Start	Stop	Strand	Length	Structure
trnF(ttc)	1	74	+	74	svg ps
rrnS	74	1053	+	980	svg ps
trnV(gta)	1051	1122	+	72	svg ps
rrnL	1123	2719	+	1597	svg ps
trnL2(tta)	2719	2793	+	75	svg ps
nad1	2798	3754	+	957	
trnI(atc)	3762	3834	+	73	svg ps
trnQ(caa)	3843	3913	-	71	svg ps
trnM(atg)	3913	3981	+	69	svg ps
nad2	3982	5010	+	1029	
trnW(tga)	5021	5091	+	71	svg ps
trnA(gca)	5093	5161	-	69	svg ps
trnN(aac)	5164	5236	-	73	svg ps
trnC(tgc)	5242	5308	-	67	svg ps
trnY(tac)	5309	5378	-	70	svg ps
cox1	5389	6921	+	1533	
trnS2(tca)	6922	6995	-	74	svg ps
trnD(gac)	6999	7067	+	69	svg ps
cox2	7069	7743	+	675	
trnK(aaa)	7754	7823	+	70	svg ps
atp8	7825	7986	+	162	
atp6	7983	8663	+	681	
cox3	8666	9448	+	783	
trnG(gga)	9450	9518	+	69	svg ps
nad3_a	9519	9692	+	174	
nad3_b	9694	9867	+	174	
trnR(cga)	9873	9941	+	69	svg ps
nad4l	9943	10236	+	294	
nad4	10233	11594	+	1362	
trnH(cac)	11611	11680	+	70	svg ps
trnS1(agc)	11681	11747	+	67	svg ps
trnL1(cta)	11748	11817	+	70	svg ps
nad5	11818	13623	+	1806	
cob	13643	14779	+	1137	
trnT(aca)	14790	14859	+	70	svg ps
trnP(cca)	14882	14951	-	70	svg ps
nad6	14962	15480	-	519	
trnE(gaa)	15485	15554	-	70	svg ps

Galaxy Interface @ neo.engr.uconn.edu/?toolid=SMART



Get Data

IMMUNOGENOMICS

Variant Calling and Filtering

Variant Validation

Epitope Calling

TRANSCRIPTOMICS

RNA-Seq Analysis

Pathway Activity

GENOMICS

Mitogenome Assembly

Get COI Gene Sequence from NCBI

SMART - Statistical Mitogenome Assembly with Repeats

Workflows

- All workflows

SMART - Statistical Mitogenome Assembly with Repeats (Galaxy Version 19.1)

Sample name: Sample

Output files label

DNA-Seq R1 file, fastq format: 1: ERR1746232_1.fastq.gz

DNA-Seq R2 file, fastq format: 2: ERR1746232_2.fastq.gz

Seed gene file, fasta format: 3: KC617843_coi_gene

Advanced Options

Number of bootstrap samples: 1

Number of reads in each bootstrapping sample: 10000000

Kmer Size: 31

Number of threads: 16

Choose correct (genetic code) for (MITOS Annotation): 02 - Vertebrate

SMART Mus_Musculus

6 shown

3.38 MB

6: Mus_Musculus: The log file

5: Mus_Musculus-Report

4: Mus_Musculus-Results_folder.zip

3: KC617843_coi_gene

2: ERR1746232_2.fastq.gz

1: ERR1746232_1.fastq.gz

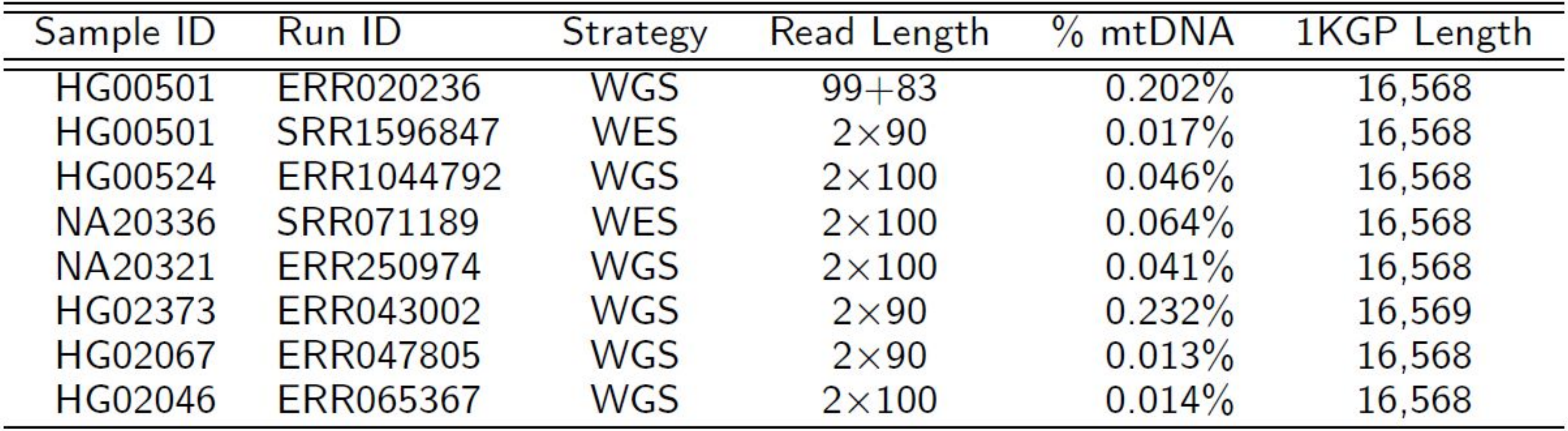
Outline

- Background/Motivation
- Related Work
- Statistical Mitogenome Assembly with Repeats (SMART)
  - The pipeline
  - **Results**
- Conclusion & ongoing Work

Datasets

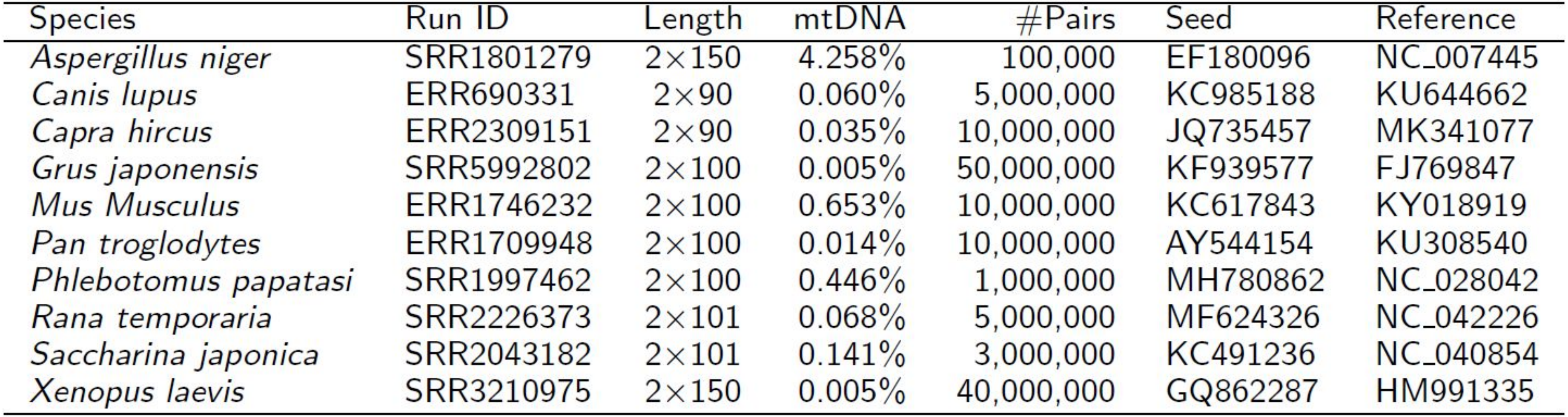
- Human datasets
- Non-Human datasets

Human WGS and WES datasets



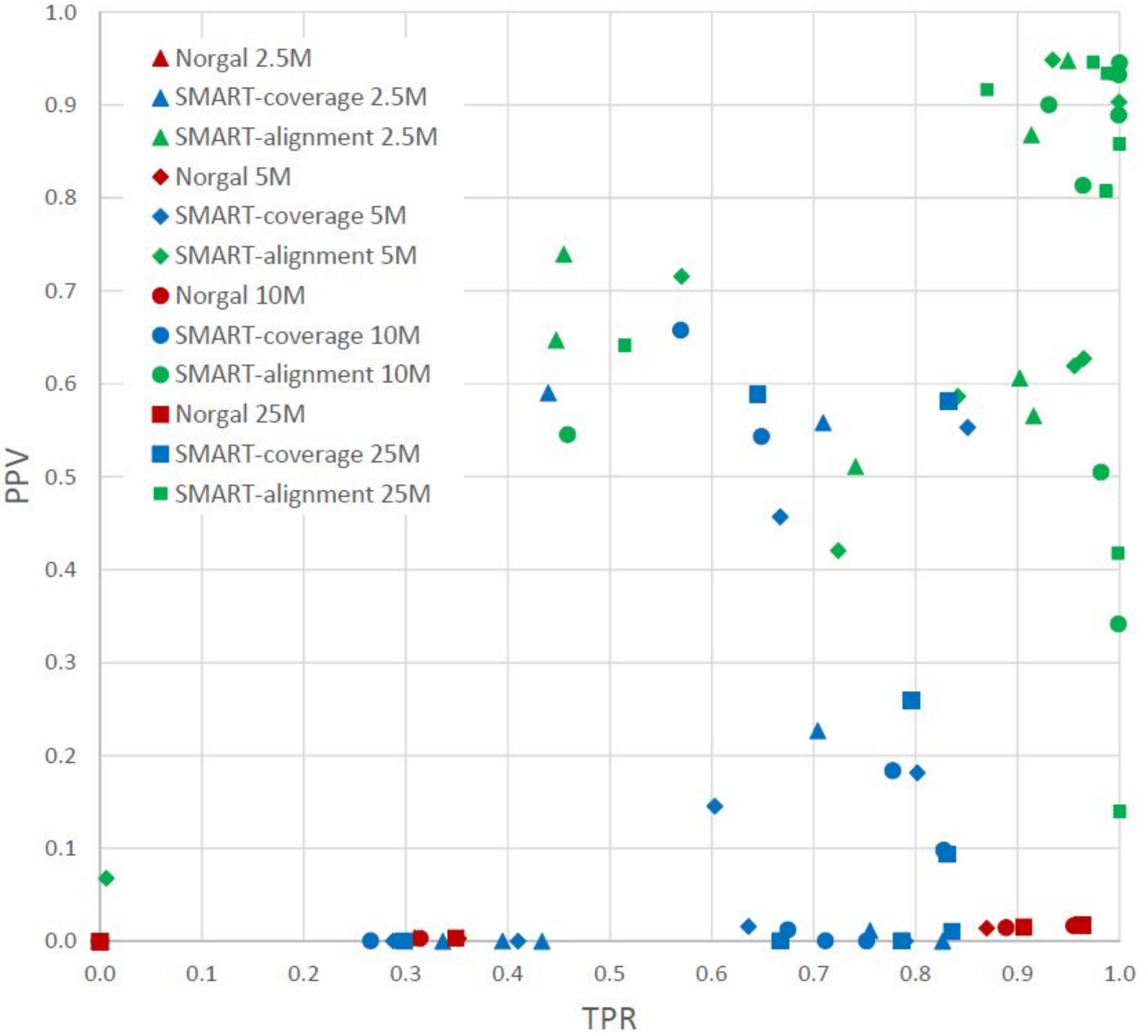
Sample ID	Run ID	Strategy	Read Length	% mtDNA	1KGP Length
HG00501	ERR020236	WGS	99+83	0.202%	16,568
HG00501	SRR1596847	WES	2×90	0.017%	16,568
HG00524	ERR1044792	WGS	2×100	0.046%	16,568
NA20336	SRR071189	WES	2×100	0.064%	16,568
NA20321	ERR250974	WGS	2×100	0.041%	16,568
HG02373	ERR043002	WGS	2×90	0.232%	16,569
HG02067	ERR047805	WGS	2×90	0.013%	16,568
HG02046	ERR065367	WGS	2×100	0.014%	16,568

Non-Human WGS datasets

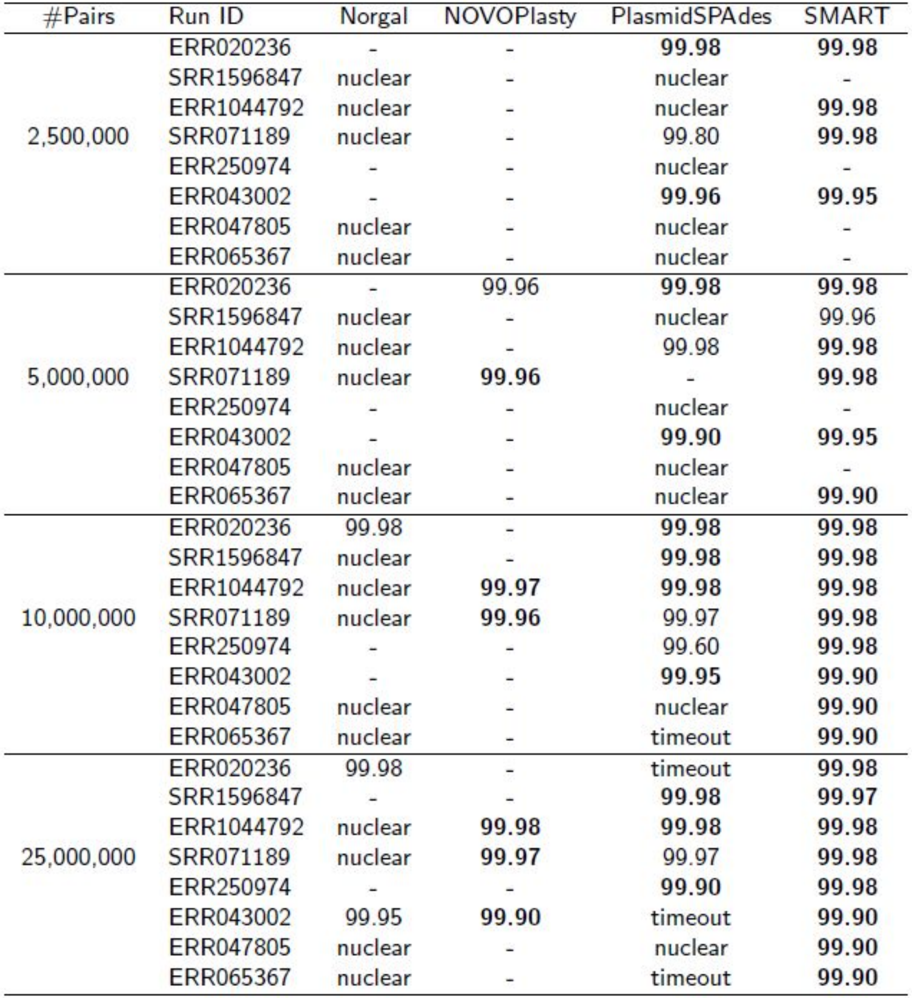


Species	Run ID	Length	mtDNA	#Pairs	Seed	Reference
Aspergillus niger	SRR1801279	2×150	4.258%	100,000	EF180096	NC_007445
Canis lupus	ERR690331	2×90	0.060%	5,000,000	KC985188	KU644662
Capra hircus	ERR2309151	2×90	0.035%	10,000,000	JQ735457	MK341077
Grus japonensis	SRR5992802	2×100	0.005%	50,000,000	KF939577	FJ769847
Mus Musculus	ERR1746232	2×100	0.653%	10,000,000	KC617843	KY018919
Pan troglodytes	ERR1709948	2×100	0.014%	10,000,000	AY544154	KU308540
Phlebotomus papatasi	SRR1997462	2×100	0.446%	1,000,000	MH780862	NC_028042
Rana temporaria	SRR2226373	2×101	0.068%	5,000,000	MF624326	NC_042226
Saccharina japonica	SRR2043182	2×101	0.141%	3,000,000	KC491236	NC_040854
Xenopus laevis	SRR3210975	2×150	0.005%	40,000,000	GQ862287	HM991335

Assessment of read filtering accuracy for human datasets with 2.5-25M read pairs



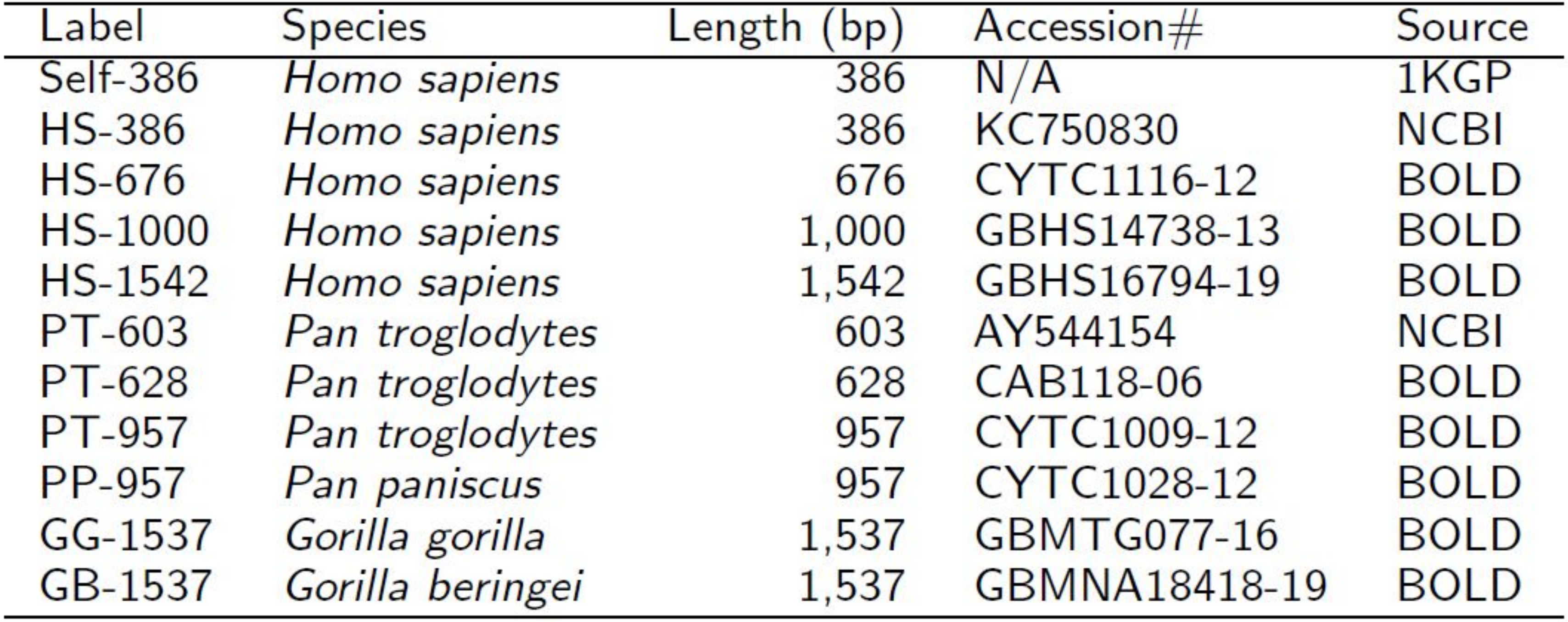
Assembly accuracy comparison on human datasets



#Pairs	Run ID	Norgal	NOVOPlasty	PlasmidSPAdes	SMART
2,500,000	ERR020236	-	-	**99.98**	**99.98**
	SRR1596847	nuclear	-	nuclear	-
	ERR1044792	nuclear	-	nuclear	**99.98**
	SRR071189	nuclear	-	99.80	**99.98**
	ERR250974	-	-	nuclear	-
	ERR043002	-	-	**99.96**	**99.95**
	ERR047805	nuclear	-	nuclear	-
	ERR065367	nuclear	-	nuclear	-
5,000,000	ERR020236	-	99.96	**99.98**	**99.98**
	SRR1596847	nuclear	-	nuclear	99.96
	ERR1044792	nuclear	-	99.98	**99.98**
	SRR071189	nuclear	**99.96**	-	**99.98**
	ERR250974	-	-	nuclear	-
	ERR043002	-	-	**99.90**	**99.95**
	ERR047805	nuclear	-	nuclear	-
	ERR065367	nuclear	-	nuclear	**99.90**
10,000,000	ERR020236	99.98	-	**99.98**	**99.98**
	SRR1596847	nuclear	-	**99.98**	**99.98**
	ERR1044792	nuclear	**99.97**	**99.98**	**99.98**
	SRR071189	nuclear	**99.96**	99.97	**99.98**
	ERR250974	-	-	99.60	**99.98**
	ERR043002	-	-	**99.95**	**99.90**
	ERR047805	nuclear	-	nuclear	**99.90**
	ERR065367	nuclear	-	timeout	**99.90**
25,000,000	ERR020236	99.98	-	timeout	**99.98**
	SRR1596847	-	-	**99.98**	**99.97**
	ERR1044792	nuclear	**99.98**	**99.98**	**99.98**
	SRR071189	nuclear	**99.97**	99.97	**99.98**
	ERR250974	-	-	**99.90**	**99.98**
	ERR043002	99.95	**99.90**	timeout	**99.90**
	ERR047805	nuclear	-	nuclear	**99.90**
	ERR065367	nuclear	-	timeout	**99.90**

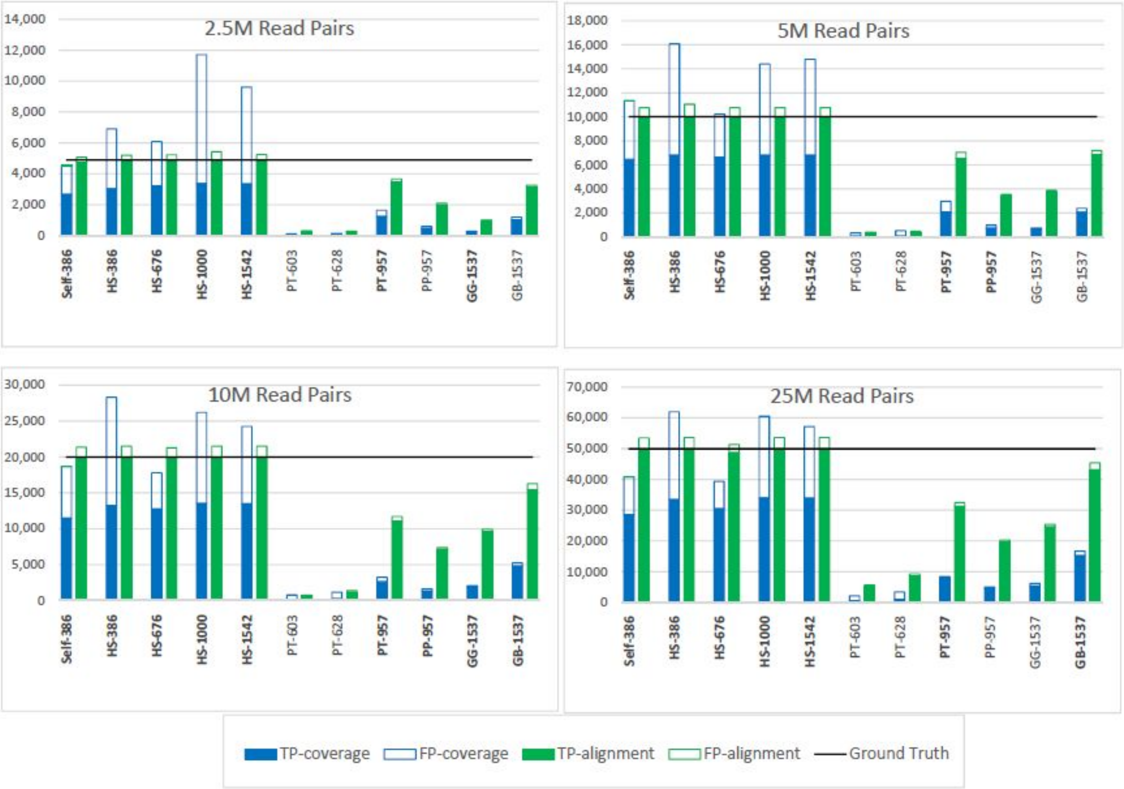
The percentage identity is typeset in **bold** if the reconstructed sequence was a complete circular genome.

Effect of the seed Length and Similarity on Read Filtering Accuracy and Assembly



Label	Species	Length (bp)	Accession#	Source
Self-386	*Homo sapiens*	386	N/A	1KGP
HS-386	*Homo sapiens*	386	KC750830	NCBI
HS-676	*Homo sapiens*	676	CYTC1116-12	BOLD
HS-1000	*Homo sapiens*	1,000	GBHS14738-13	BOLD
HS-1542	*Homo sapiens*	1,542	GBHS16794-19	BOLD
PT-603	*Pan troglodytes*	603	AY544154	NCBI
PT-628	*Pan troglodytes*	628	CAB118-06	BOLD
PT-957	*Pan troglodytes*	957	CYTC1009-12	BOLD
PP-957	*Pan paniscus*	957	CYTC1028-12	BOLD
GG-1537	*Gorilla gorilla*	1,537	GBMTG077-16	BOLD
GB-1537	*Gorilla beringei*	1,537	GBMNA18418-19	BOLD

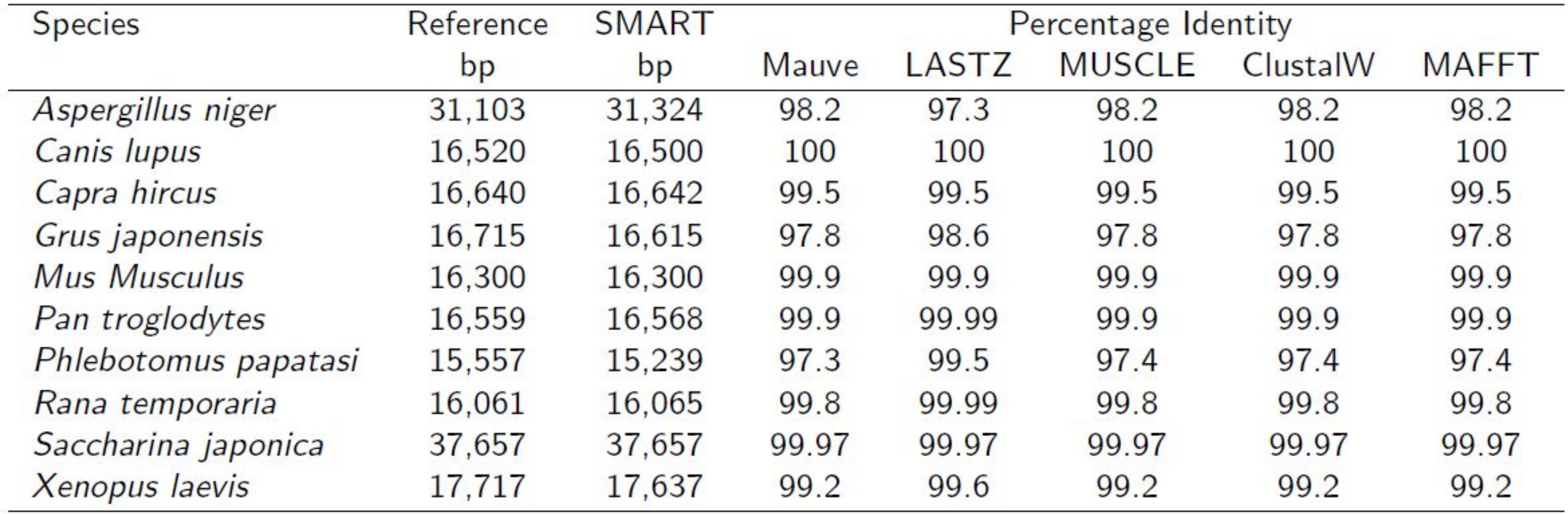
Effect of the seed Length and Similarity on Read Filtering Accuracy and Assembly



datasets with 2.5M-25M read pairs randomly selected from WGS run ERR020236

The percentage identity is typeset in **bold** if the reconstructed sequence was a complete circular genome.

SMART assembly accuracy for non-human datasets



Species	Reference	SMART	Percentage Identity				
	bp	bp	Mauve	LASTZ	MUSCLE	ClustalW	MAFFT
Aspergillus niger	31,103	31,324	98.2	97.3	98.2	98.2	98.2
Canis lupus	16,520	16,500	100	100	100	100	100
Capra hircus	16,640	16,642	99.5	99.5	99.5	99.5	99.5
Grus japonensis	16,715	16,615	97.8	98.6	97.8	97.8	97.8
Mus Musculus	16,300	16,300	99.9	99.9	99.9	99.9	99.9
Pan troglodytes	16,559	16,568	99.9	99.99	99.9	99.9	99.9
Phlebotomus papatasi	15,557	15,239	97.3	99.5	97.4	97.4	97.4
Rana temporaria	16,061	16,065	99.8	99.99	99.8	99.8	99.8
Saccharina japonica	37,657	37,657	99.97	99.97	99.97	99.97	99.97
Xenopus laevis	17,717	17,637	99.2	99.6	99.2	99.2	99.2

All are circular except Rana temporaria

Outline

- Background/Motivation
- Related Work
- Statistical Mitogenome Assembly with Repeats (SMART)
  - The pipeline
  - Results
- **Conclusion & Ongoing and Future Work**

Conclusions

- SMART is an automated pipeline for de novo mitogenome assembly from WGS reads
- Based on statistical framework
  - Probabilistic read classifier based on coverage
  - Likelihood maximization for resolving ambiguities in assembly graph
  - Assembly confidence estimated by bootstrapping
- Produces complete/circular assemblies even from low-coverage WGS data
- Available via galaxy interface at neo.engr.uconn.edu/?toolid=SMART
- SMART paper is under review

Ongoing and Future Work

- Improving Mitogenomes assembly
  - Multi-sample Coverage based Read Filter & Assembly
  - Codon Usage Bias Read Filter
  - Orphan Mitogenomes Project
- Mitochondrial DNA Forensics
- Plants organelles Assembly

Ongoing and Future Work

- **Improving Mitogenomes assembly**
  - **Multi-sample Coverage based Read Filter & Assembly**
  - Codon Usage Bias Read Filter
  - Orphan Mitogenomes Project
- Mitochondrial DNA Forensics
- Plants organelles Assembly

Multi-sample Coverage based Read Filter

- Using more than one samples in a SMART run

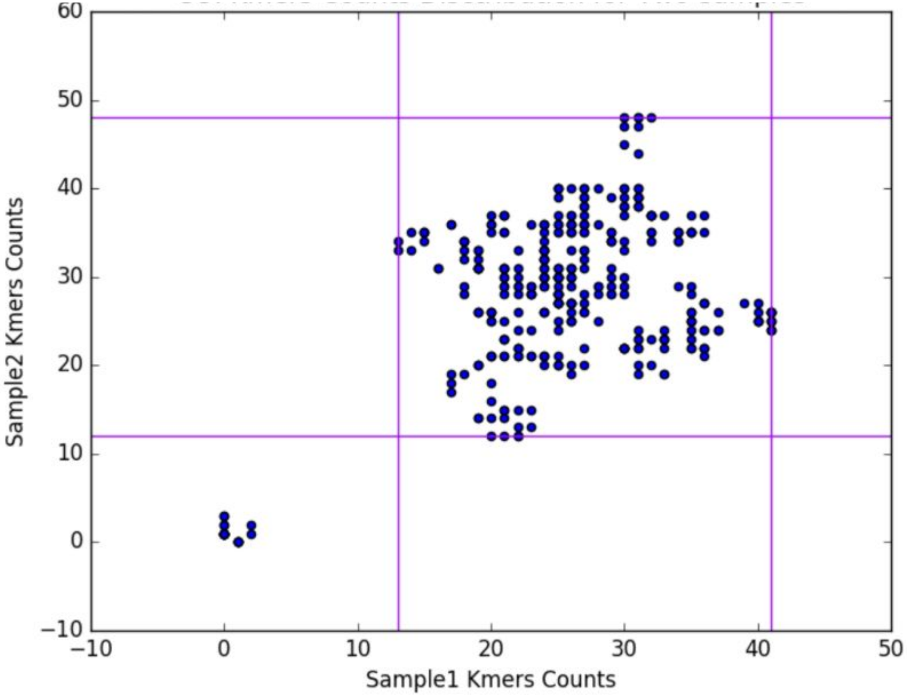
Multi-sample Coverage based Read Filter

- 1KGP “HG00675” has two runs
  - SRR593357
  - SRR593484

Run ID	Read Count	Library Layout	Library Strategy	Library Source	Library Selection
SRR593357	2,137,487	Paired	WGS	Genomic	Random
SRR593484	2,132,673	Paired	WGS	Genomic	Random

Multi-sample Coverage based Read Filter

COI K-mers Counts Distribution of Two Samples



Multi-sample Coverage based Read Filter

Run ID	Num of reads pairs	Filter	TPR	PPV	F-Score
SRR593357	2,137,487	Coverage-based	0.426	0.408	0.263
		Multi-sample	**1**	0.386	0.386
SRR593484	2,132,673	Coverage-based	0.770	0.505	0.439
		Multi-sample	**1**	0.399	0.399

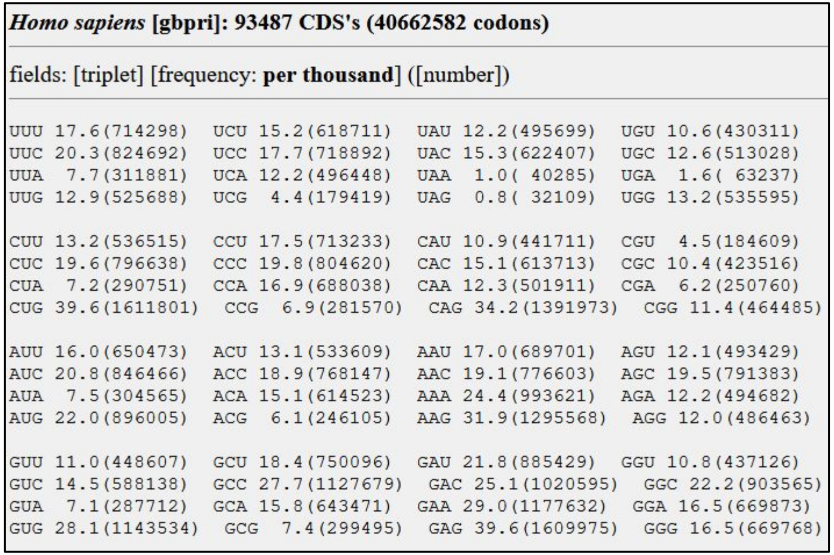
Multi-sample Mitogenome Assembly

- We plan to use ALLPATHS-LG assembler
- We plan to generate mitogenome sequence of Pyxicephalus adspersus (African bullfrog):
    - We plan to use two WGS libraries with two different insert sizes: 180 bp, and 550 bp.

Ongoing and Future Work

- Improving Mitogenomes assembly
  - Multi-sample Coverage based Read Filter & Assembly
  - **Codon Usage Bias Read Filter**
  - Orphan Mitogenomes Project
- Mitochondrial DNA Forensics
- Plants organelles Assembly

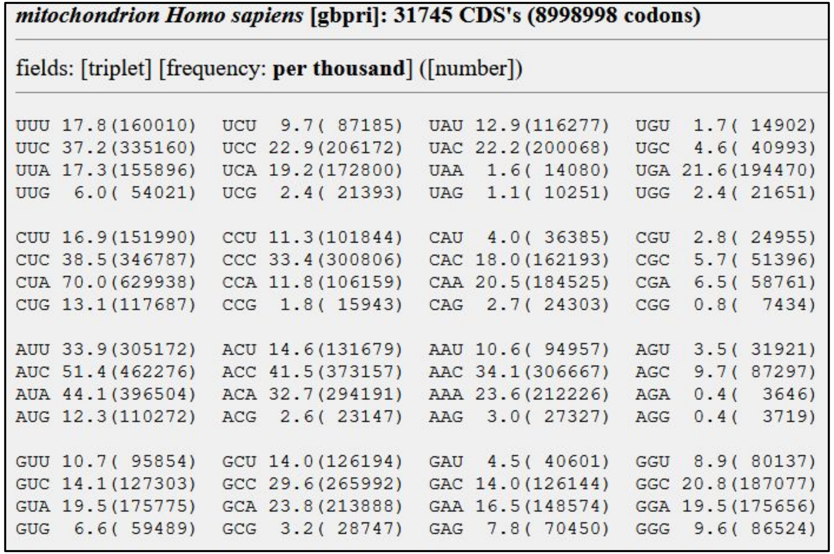
Codon Usage Bias Read Filter



***Homo sapiens* [gbpri]: 93487 CDS's (40662582 codons)**

fields: [triplet] [frequency: **per thousand**] ([number])

```
UUU 17.6(714298)  UCU 15.2(618711)  UAU 12.2(495699)  UGU 10.6(430311)
UUC 20.3(824692)  UCC 17.7(718892)  UAC 15.3(622407)  UGC 12.6(513028)
UUA  7.7(311881)  UCA 12.2(496448)  UAA  1.0( 40285)  UGA  1.6( 63237)
UUG 12.9(525688)  UCG  4.4(179419)  UAG  0.8( 32109)  UGG 13.2(535595)

CUU 13.2(536515)  CCU 17.5(713233)  CAU 10.9(441711)  CGU  4.5(184609)
CUC 19.6(796638)  CCC 19.8(804620)  CAC 15.1(613713)  CGC 10.4(423516)
CUA  7.2(290751)  CCA 16.9(688038)  CAA 12.3(501911)  CGA  6.2(250760)
CUG 39.6(1611801)  CCG  6.9(281570)  CAG 34.2(1391973)  CGG 11.4(464485)

AUU 16.0(650473)  ACU 13.1(533609)  AAU 17.0(689701)  AGU 12.1(493429)
AUC 20.8(846466)  ACC 18.9(768147)  AAC 19.1(776603)  AGC 19.5(791383)
AUA  7.5(304565)  ACA 15.1(614523)  AAA 24.4(993621)  AGA 12.2(494682)
AUG 22.0(896005)  ACG  6.1(246105)  AAG 31.9(1295568)  AGG 12.0(486463)

GUU 11.0(448607)  GCU 18.4(750096)  GAU 21.8(885429)  GGU 10.8(437126)
GUC 14.5(588138)  GCC 27.7(1127679)  GAC 25.1(1020595)  GGC 22.2(903565)
GUA  7.1(287712)  GCA 15.8(643471)  GAA 29.0(1177632)  GGA 16.5(669873)
GUG 28.1(1143534)  GCG  7.4(299495)  GAG 39.6(1609975)  GGG 16.5(669768)
```



***mitochondrion Homo sapiens* [gbpri]: 31745 CDS's (8998998 codons)**

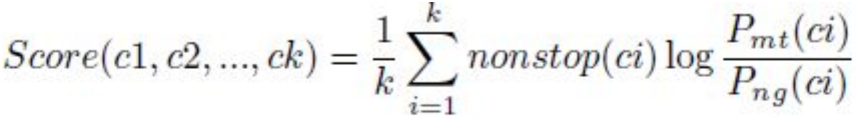
fields: [triplet] [frequency: **per thousand**] ([number])

```
UUU 17.8(160010)  UCU  9.7( 87185)  UAU 12.9(116277)  UGU  1.7( 14902)
UUC 37.2(335160)  UCC 22.9(206172)  UAC 22.2(200068)  UGC  4.6( 40993)
UUA 17.3(155896)  UCA 19.2(172800)  UAA  1.6( 14080)  UGA 21.6(194470)
UUG  6.0( 54021)  UCG  2.4( 21393)  UAG  1.1( 10251)  UGG  2.4( 21651)

CUU 16.9(151990)  CCU 11.3(101844)  CAU  4.0( 36385)  CGU  2.8( 24955)
CUC 38.5(346787)  CCC 33.4(300806)  CAC 18.0(162193)  CGC  5.7( 51396)
CUA 70.0(629938)  CCA 11.8(106159)  CAA 20.5(184525)  CGA  6.5( 58761)
CUG 13.1(117687)  CCG  1.8( 15943)  CAG  2.7( 24303)  CGG  0.8(  7434)

AUU 33.9(305172)  ACU 14.6(131679)  AAU 10.6( 94957)  AGU  3.5( 31921)
AUC 51.4(462276)  ACC 41.5(373157)  AAC 34.1(306667)  AGC  9.7( 87297)
AUA 44.1(396504)  ACA 32.7(294191)  AAA 23.6(212226)  AGA  0.4(  3646)
AUG 12.3(110272)  ACG  2.6( 23147)  AAG  3.0( 27327)  AGG  0.4(  3719)

GUU 10.7( 95854)  GCU 14.0(126194)  GAU  4.5( 40601)  GGU  8.9( 80137)
GUC 14.1(127303)  GCC 29.6(265992)  GAC 14.0(126144)  GGC 20.8(187077)
GUA 19.5(175775)  GCA 23.8(213888)  GAA 16.5(148574)  GGA 19.5(175656)
GUG  6.6( 59489)  GCG  3.2( 28747)  GAG  7.8( 70450)  GGG  9.6( 86524)
```

Codon Usage Database:https://www.kazusa.or.jp/codon/

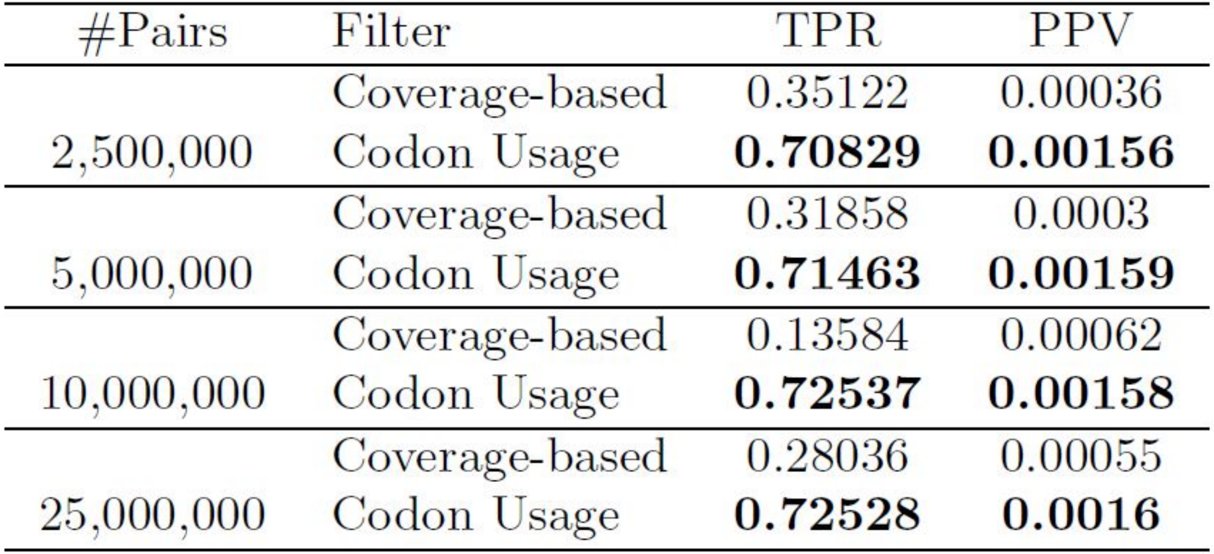
Codon Usage Bias Read Filter

- For each read
    - Calculating score for each (Open reading frames) ORF using bellow equation if codon is non stop; otherwise, a large negative number will be given to stop codon.
    - We normalizing the score by k number of codons in each ORF



$$Score(c1, c2, ..., ck) = \frac{1}{k}\sum_{i=1}^{k} nonstop(ci)\log\frac{P_{mt}(ci)}{P_{ng}(ci)}$$

Preliminary results of read filter accuracy comparison on datasets with 2.5M-25M read pairs randomly selected from WGS run ERR020236



#Pairs	Filter	TPR	PPV
2,500,000	Coverage-based	0.35122	0.00036
	Codon Usage	**0.70829**	**0.00156**
5,000,000	Coverage-based	0.31858	0.0003
	Codon Usage	**0.71463**	**0.00159**
10,000,000	Coverage-based	0.13584	0.00062
	Codon Usage	**0.72537**	**0.00158**
25,000,000	Coverage-based	0.28036	0.00055
	Codon Usage	**0.72528**	**0.0016**

Ongoing and Future Work

- Improving Mitogenomes assembly
    - Multi-sample Coverage based Read Filter & Assembly
    - Codon Usage Bias Read Filter
    - **Orphan Mitogenomes Project**
- Mitochondrial DNA Forensics
- Plants organelles Assembly

Orphan Mitogenomes Project

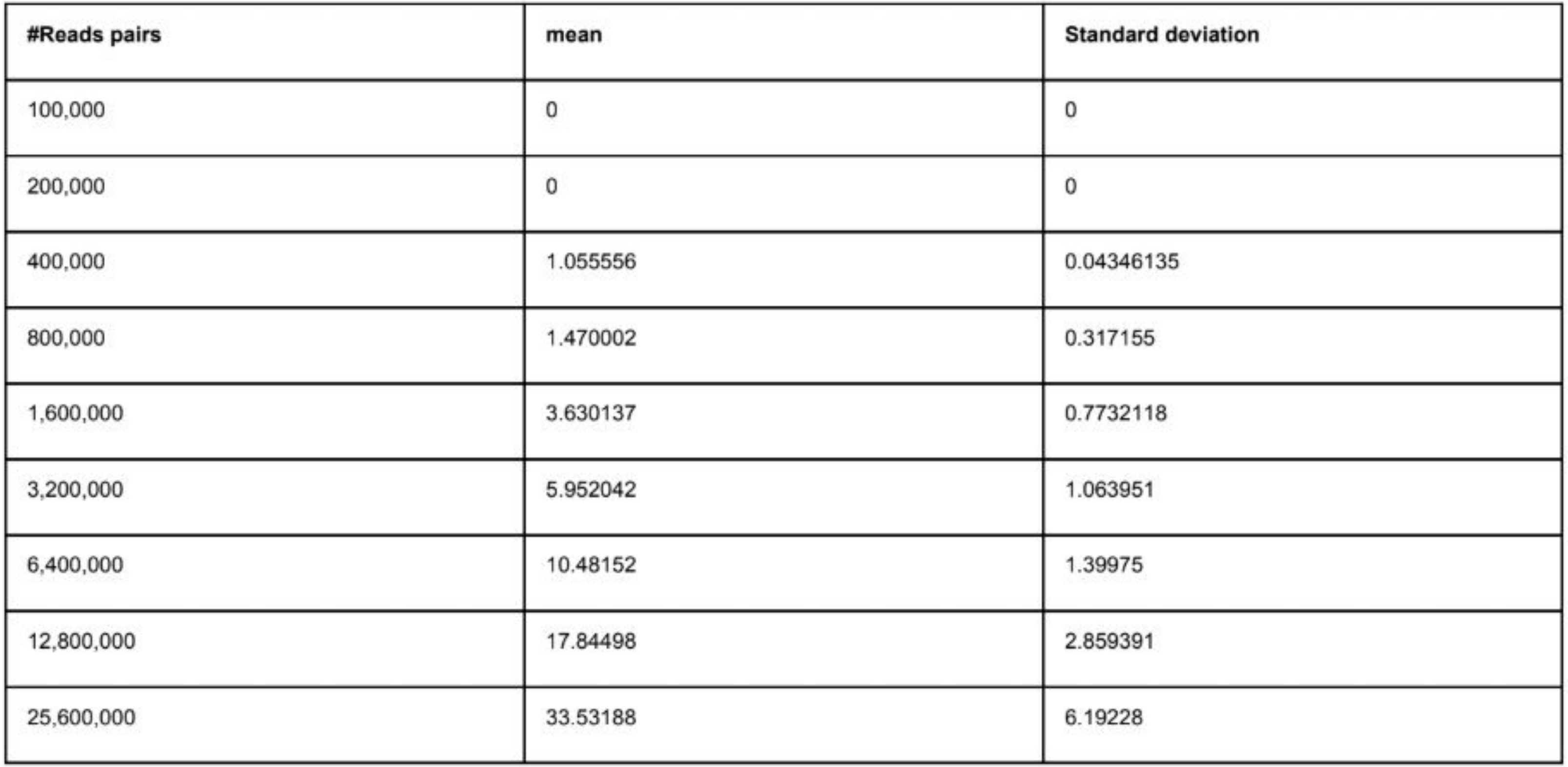
- Some of NCBI organisms have neither complete nor partial mitogenomes in NCBI
- Example,
  - 235 paired-end WGS data of mammals in NCBI have no complete/partial mitogenomes.

Automatically Choosing Needed # Read Pairs

- Pipistrellus pipistrellus

Run ID	#Read Pairs	Library Layout	Library Strategy	Library Source	Library Selection
ERR3316150	120,578,311	Paired	WGS	Genomic	Random

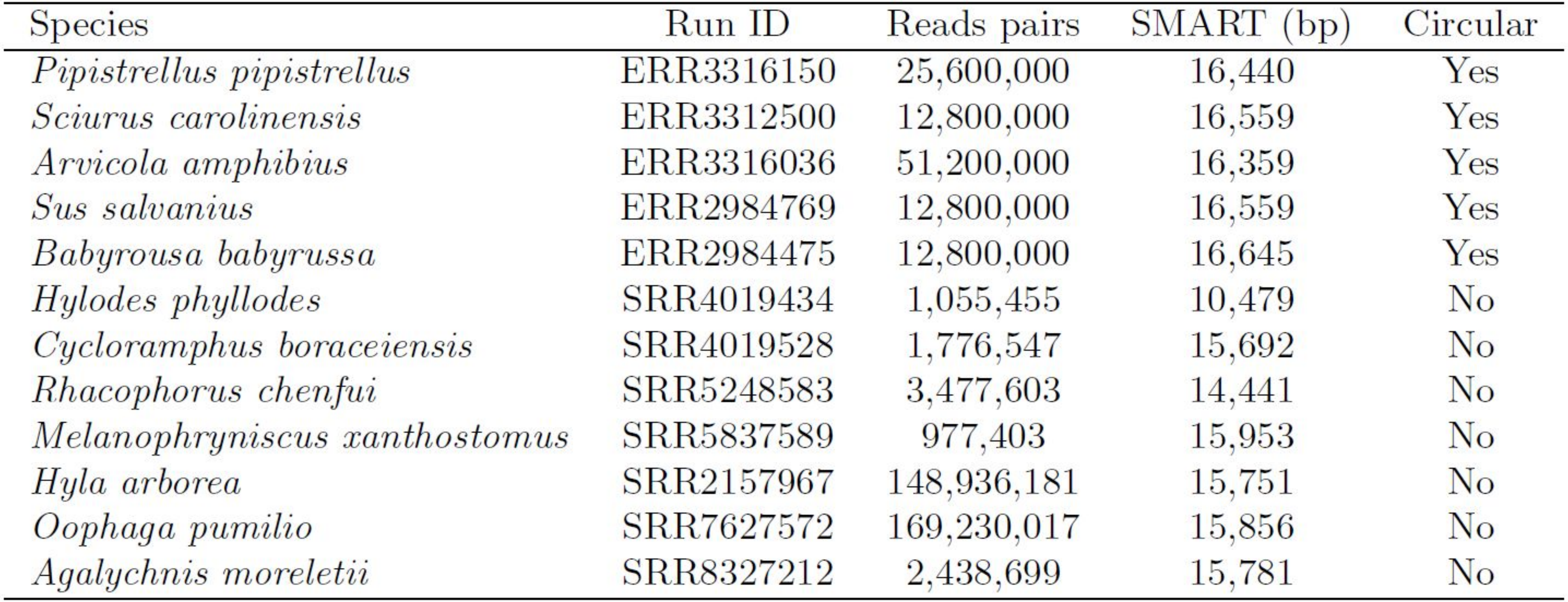
Automatically Choosing Needed # Read Pairs



#Reads pairs	mean	Standard deviation
100,000	0	0
200,000	0	0
400,000	1.055556	0.04346135
800,000	1.470002	0.317155
1,600,000	3.630137	0.7732118
3,200,000	5.952042	1.063951
6,400,000	10.48152	1.39975
12,800,000	17.84498	2.859391
25,600,000	33.53188	6.19228

Results

We plan to submit these mitogenomes to GenBank as Third Party Annotation (TPA) sequence.



Species	Run ID	Reads pairs	SMART (bp)	Circular
Pipistrellus pipistrellus	ERR3316150	25,600,000	16,440	Yes
Sciurus carolinensis	ERR3312500	12,800,000	16,559	Yes
Arvicola amphibius	ERR3316036	51,200,000	16,359	Yes
Sus salvanius	ERR2984769	12,800,000	16,559	Yes
Babyrousa babyrussa	ERR2984475	12,800,000	16,645	Yes
Hylodes phyllodes	SRR4019434	1,055,455	10,479	No
Cycloramphus boraceiensis	SRR4019528	1,776,547	15,692	No
Rhacophorus chenfui	SRR5248583	3,477,603	14,441	No
Melanophryniscus xanthostomus	SRR5837589	977,403	15,953	No
Hyla arborea	SRR2157967	148,936,181	15,751	No
Oophaga pumilio	SRR7627572	169,230,017	15,856	No
Agalychnis moreletii	SRR8327212	2,438,699	15,781	No

Ongoing and Future Work

- Improving Mitogenomes assembly
  - Multi-sample Coverage based Read Filter & Assembly
  - Orphan Mitogenomes Project
  - Codon Usage Bias Read Filter
- **Mitochondrial DNA Forensics**
- Plants organelles Assembly

Mitochondrial DNA Forensics

- Reconstructing mitochondrial DNA sequences from heterogeneous samples
- The more contributors are in mixture DNA sample, the more complicated analyzing of the mixture will be

Steps

1. Aligning mixture DNA fastq file to the reference mtDNA sequence
2. SNV calling
3. Generating incompatibility conflict graph
4. Run at tool (ReFHap) for Max-cut
5. Generating one consensus mtDNA sequence from each partition.

Ongoing and Future Work

- Improving Mitogenomes assembly
  - Multi-sample Coverage based Filter & Assembly
  - Orphan Mitogenomes Project
  - Codon Usage Bias Filter
- Mitochondrial DNA Forensics
- **Plants organelles Assembly**

Plants organelles Assembly

- Circular organelles in Plants:
  - Mitochondria
  - Chloroplasts
- Plants organelle genomes are much larger than in animals
- Mitochondrial genome sizes in plants are between 200,000 and 2,000,000 bp
- 90% of these larger plants mitochondrial DNA sequences are introns and repeated sequences
- Chloroplasts genomes range size is between 120,000 and 170,000

Thank You for Your Attention

Any questions?