Assembly of Large Genomes with High Ploidy

Ekaterina Esaulova
Advisor: Yana Safonova, Algorithmic Biology Lab, SPbAU RAS

Bioinformatics
Institute

Introduction

Ploidy is the number of sets of chromosomes in the nucleus of a cell.

Genome is large if the total number of DNA base pairs in one copy of a haploid genome is greater than 0,5 Gpb.

Introduction

… assembly?

- dipSPAdes
- ABySS
- ALLPATHS
- SOAPdenovo
- Platanus

Introduction

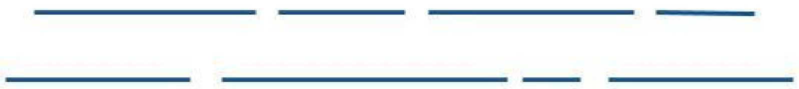
… assembly?

- dipSPAdes
- ABySS
- ALLPATHS
- SOAPdenovo

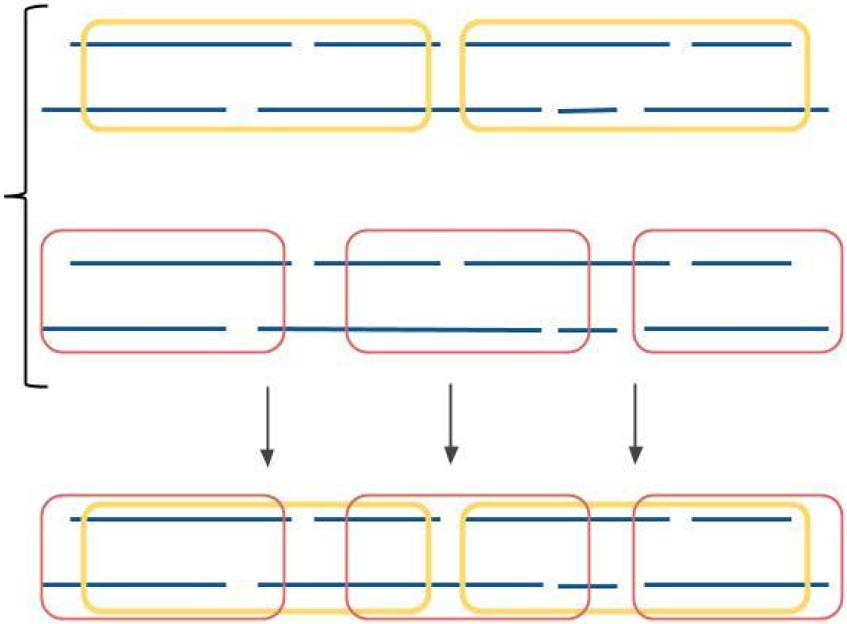
Large genome, high ploidy → troubles with de Bruijn graph→ short contigs

Idea

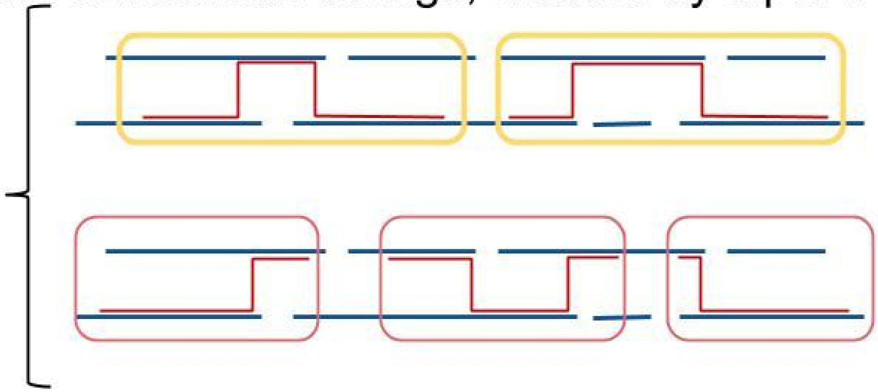
1. Set of haplocontigs



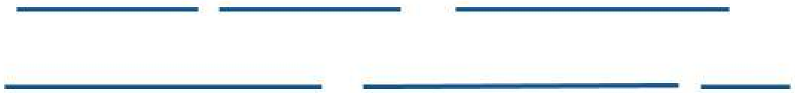
2. Overlapping decompositions of contigs



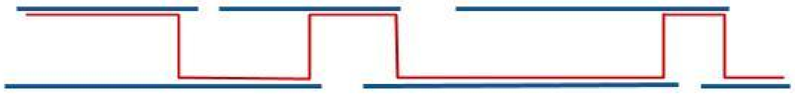
3. Consensus contigs, created by dipSPAdes



4. Set of overlapping consensus contigs



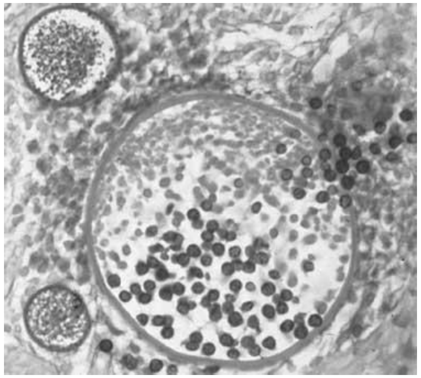
5. Final consensus



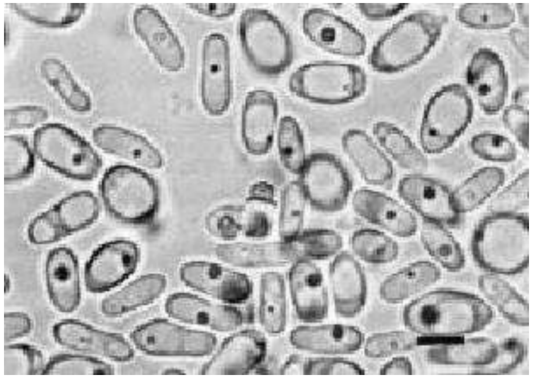
Datasets

- Amoeboaphelidium protococcarum (precomputed contigs)
- Cyberlindnera jadinii (Illumina, 150x2, IS = 265)
- Ciona Savignyi (Illumina, 265x2, IS = 624)

A. protococcarum

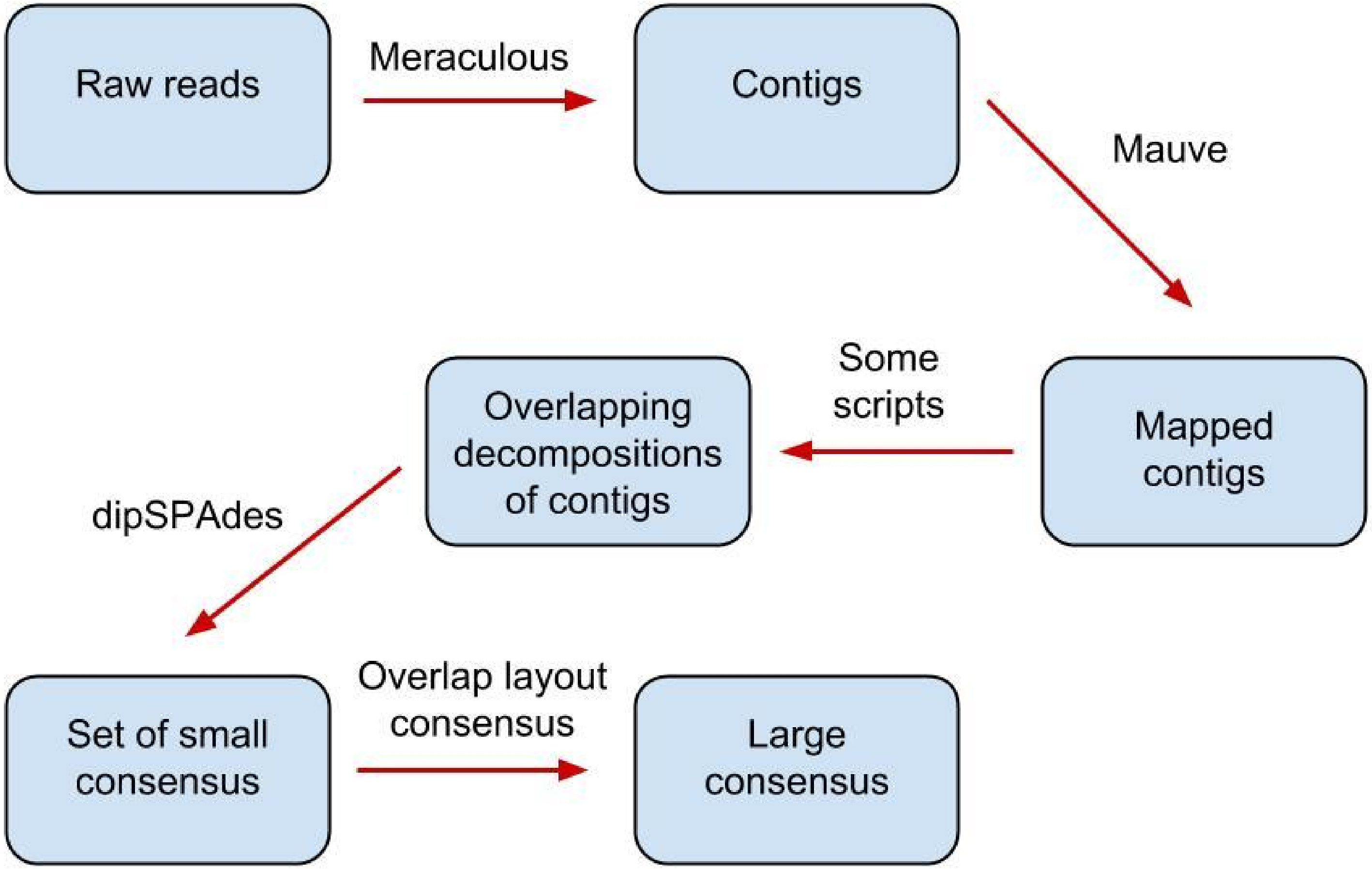


Cyberlindnera jadinii



Ciona Savignyi

Pipeline


Raw reads
Meraculous
Contigs
Mauve
Mapped contigs
Some scripts
Overlapping decompositions of contigs
dipSPAdes
Set of small consensus
Overlap layout consensus
Large consensus

Results: Meraculous, haplocontigs

Features:

- a lot of required options
- run time: 1-2 days (on 5 processors)
- need to set k and other important parameters manually

Problems:

- very fragmented assembly

Results: Meraculous, haplocontigs

Ciona	#1	#2	#3	Expect
k	45	75	99	
Largest contig	10,589	13,013	4,283	
Total length	49,276,453	50,291,820	1,928,777	180,000,000*2
N50	723	679	597	

Cyberlindnera	#1	#2	#3	Expect
k	55	75	99	
Largest contig	79,136	79,156	102,929	
Total length	4,734,456	7,010,481	10,757,812	12,000,000*2
N50	2,684	1,222	1,043	

Results: Mauve, map of contigs

Features:

- a need to filter contigs before running Mauve
- non-trivial output
- doesn’t align contig on set of contigs
- Not greater than one occurrence is found for every part of contig in a set of contigs
- run time: for set of ~5.000 contigs - 0.5-3 min for a contig

Overlapping decompositions

Problem: want to find sets of overlapping contigs.

Solution:

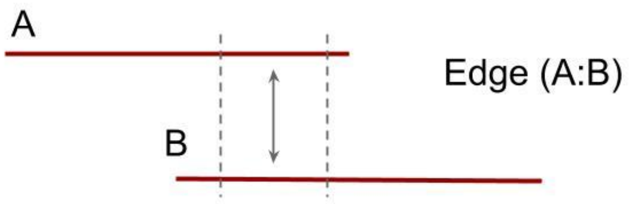
- A set of directed weighted graphs:

  vertices = {subset of contigs}

  edges = {(A:B) | contigs A and B overlap}

  edge's weight = {# of bases in overlapping region}

- Finding the longest path in graphs.



Project goals

1. Computing preliminary fragmented haplocontigs (using Meraculous)
2. Alignment by Mauve and construction of the map of contigs
3. Construction of overlapping decompositions

---

4. Construction of a small consensus by dipSPAdes
5. Construction of a large consensus by the overlap layout consensus
6. Quality assessment of the constructed consensus contigs

Results

Bash and python scripts for:

- running Meraculous on our data
- preparing data for mapping contigs
- mapping contigs by Mauve on server
- proccessing Mauve’s output

… creation of contig’s chains

+ experience in Linux, bash, assemblers, genome alignments

Problems and plans

- finish creation of overlapping decompositions
- run dipSPAdes and get consensus
- finish pipeline for Ciona and Cyberlindnera data (we run Mauve only on Amoeboaphelidium’s contigs)
- find alternative to Mauve or tune Mauve for mapping contigs

Thank you! Questions?