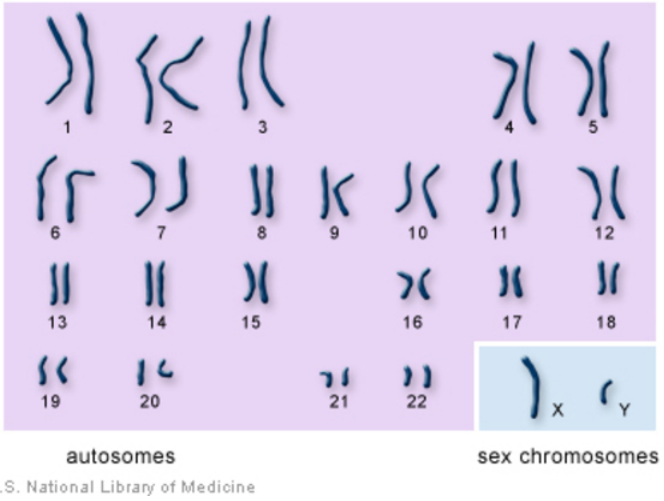
De novo whole genome assembly

Qi Sun

Bioinformatics Facility
Cornell University

When someone says “I sequenced a genome”, technically it could mean very different things.

Expectation



Human is diploid. A perfect genome assembly should give us 46 chromosomal sequences.

(23 maternal + 23 paternal)

Different levels of finishes

De novo Assembly

- **Diploid assembly:** phased assemblies with 46 sequences;
- **Chromosomal level assembly:** 23 sequences, collapsed paternal/maternal chromosomes;
- **Contig level assembly:** hundreds or thousands of contigs;

Reference guided variants calling

- **Genome re-sequencing:** SNP and indels anchored to a reference genome

Inferred from haplotypes

- Re-constructed genome from haplotypes.

Every individual's genome is different. But *de novo* assembly for every genome is too expensive.

The Concept of Reference Genome

>personA_chr1-paternal

GATGGGATTGGGGTTTTCCCCTCCCATGTGCTCAAGACTGGCGCTAAAAGTTTTGAGCTTCTCAAAAGTC
TAGAGCCACCGTCCAGGGAGCAGGTAGCTGCTGGGCTCCGGGGACACTTTGCGTTCGGGCTGGGAGCGTG
CTTTCCACGACGGTGACACGCTTCCCTGGATTGGCAGCCAGACTGCCTTCCGGGTCACTGCCATGGAGGA
GCCGCAGTCAGATCCTAGCGTCGAGCCCCCTCTGAGTCAGGAAACATTTTCAGACCTATGGAAACTACTT
CCTGAAAACAACGTTCTGTCCCCCTTGCCGTCCCAAGCAATGGATGATTTGATGCTGTCCCCGGACGATA
TTGAACAATGGTTCACTGAAGACCCAGGTCCAGATGAAGCTCCCAGAATGCCAGAGGCTGCTCCCCCCGT
GGCCCCTGCACCAGCAGCTCCTACACCGGCGGCCCCTGCACCAGCCCCCTCCTGGCCCCTGTCATCTTCT

>personA_chr1-maternal

GATGGGATTGGGGTTTTCCCCTCCCATGTGCTCAAGACTGGCGCTAAAAGTTTTGAGCTTCTCAAAAGTC
TAGAGCCACCGTCCAGGGAGCAGGTAGCTGCTGGGCTCCGGGGACACTTTGCGTTCGGGCTGGGAGCGTG
CTTTCCACGACGGTGACACGCTTCCCTGGATTGGCAGCCAGACTGCCTTCCGGGTCACTGCCATGGAGGA
GCCGCAGTCAGATCCTAGCGTCGAGCCCCCTCTGAGTCAGGAAACATTTTCAGACCTATGGAAACTACTT
CCTGAAAACAACGTTCTGTCCCCCTTGCCGTCCCAAGCAATGGATGATTTGATGCTGTCCCCGGACGATA
TTGAACAATGGTTCACTGAAGACCCAGGTCCAGATGAAGCTCCCAGAATGCCAGAGGCTGCTCCCCCCGT
GGCCCCTGCACCAGCAGCTCCTACACCGGCGGCCCCTGCACCAGCCCCCTCCTGGCCCCTGTCATCTTCT

>personB_chr1-paternal

GATGGGATTGGGGTTTTCCCCTCCCATGTGCTCAAGACTGGCGCTAAAAGTTTTGAGCTTCTCAAAAGTC
TAGAGCCACCGTCCAGGGAGCAGGTAGCTGCTGGGCTCCGGGGACACTTTGCGTTCGGGCTGGGAGCGTG
CTTTCCACGACGGTGACACGCTTCCCTGGATTGGCAGCCAGACTGCCTTCCGGGTCACTGCCATGGAGGA
GCCGCAGTCAGATCCTAGCGTCGAGCCCCCTCTGAGTCAGGAAACATTTTCAGACCTATGGAAACTACTT
CCTGAAAACAACGTTCTGTCCCCCTTGCCGTCCCAAGCAATGGATGATTTGATGCTGTCCCCGGACGATA
TTGAACAATGGTTCACTGAAGACCCAGGTCCAGATGAAGCTCCCAGAATGCCAGAGGCTGCTCCCCCCGT
GGCCCCTGCACCAGCAGCTCCTACACCGGCGGCCCCTGCACCAGCCCCCTCCTGGCCCCTGTCATCTTCT

>personB_chr1-maternal

GATGGGATTGGGGTTTTCCCCTCCCATGTGCTCAAGACTGGCGCTAAAAGTTTTGAGCTTCTCAAAAGTC
TAGAGCCACCGTCCAGGGAGCAGGTAGCTGCTGGGCTCCGGGGACACTTTGCGTTCGGGCTGGGAGCGTG
CTTTCCACGACGGTGACACGCTTCCCTGGATTGGCAGCCAGACTGCCTTCCGGGTCACTGCCATGGAGGA
GCCGCAGTCAGATCCTAGCGTCGAGCCCCCTCTGAGTCAGGAAACATTTTCAGACCTATGGAAACTACTT
CCTGAAAACAACGTTCTGTCCCCCTTGCCGTCCCAAGCAATGGATGATTTGATGCTGTCCCCGGACGATA
TTGAACAATGGTTCACTGAAGACCCAGGTCCAGATGAAGCTCCCAGAATGCCAGAGGCTGCTCCCCCCGT
GGCCCCTGCACCAGCAGCTCCTACACCGGCGGCCCCTGCACCAGCCCCCTCCTGGCCCCTGTCATCTTCT

Reference

GATGGGATTGGGGTTTGAGCTTCTCAAAAGTC

personA ……T/A……..

personB ……T/T……..

personC ……T/A……..

personD ……T/A……..

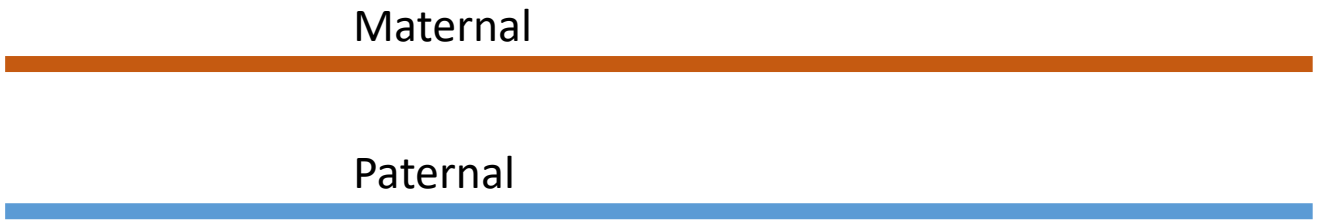
Why do we need a reference genome:

- It is a community standard, enabling comparison between genomes;
- Reference guided re-sequencing is a lot cheaper and easier to do with today's technology.

Cornell professor Dr. Doug Antczak with Twilight, DNA donor for the horse reference genome.

Reference genome is a mosaic of paternal and maternal genomes from one individual

Diploid genome



Reference

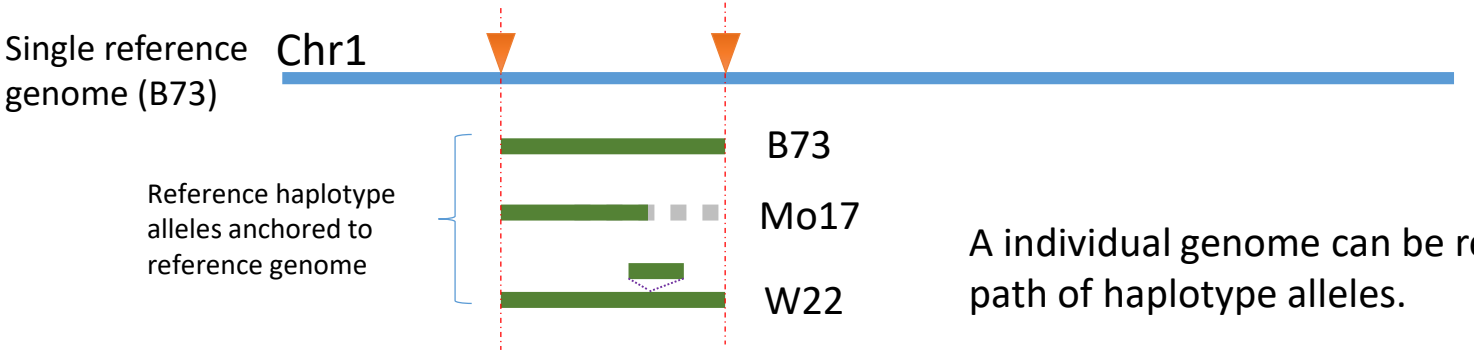
The human reference is a composite genome from multiple anonymous individuals

Single or multiple reference genomes per species?

Current reference genome B73 does not represent all the haplotypes in maize. Now there are >26 maize genome assemblies.

Venkatesh, etl al. *Journal of Agricultural and Food Chemistry* 2015, 63, 21, 5282-5295

Practical Haplotype Graph (PHG) by E. Buckler



A individual genome can be represented as a path of haplotype alleles.

Sequencing technologies (before 2019)

- **Short reads** (150bp)
  - Illumina

0.1% Error

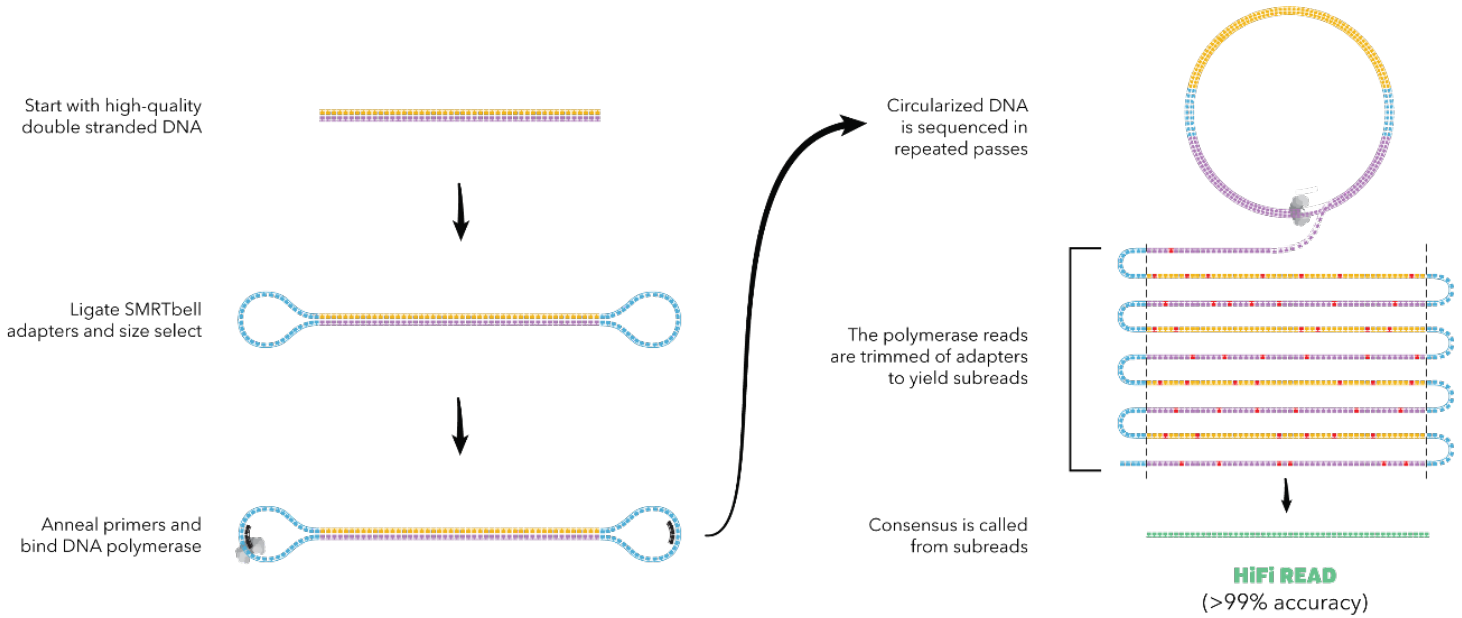
- **Long reads** (>10kb)
  - PacBio
  - Oxford Nanopore

10% Error

Lastest development: **PacBio HiFi**

- **~ 20 kb sequence reads;**
- **>99.9% accuracy**

(Circular consensus sequencing)



(Most remaining errors are mostly indels in homopolymer regions)

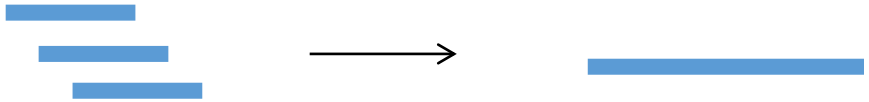
Bioinformatics: steps in genome assembly

Raw data: reads

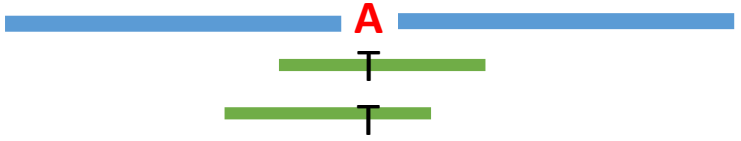
Preprocessing clean up reads

e.g. trim low quality part of the read

Contiging reads to contigs



Polishing Error correction



Scaffolding longer pieces

Preprocessing: Read error correction

Short reads

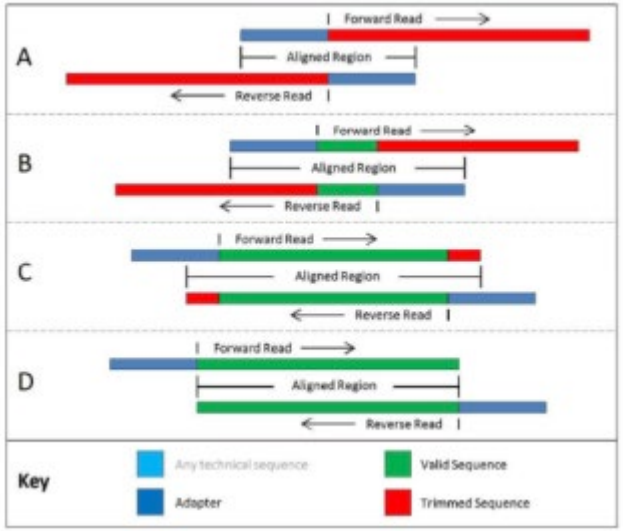
Trimming 3’ adapters and low quality sequences

Software:

- bbduk (part of bbtools)
- Trimmomatic

Palindrome mode:
“adapter sequence” +
“overlapping in PE reads”

Trimmomatic: default
Bbduk: parameter “tbo”



Long reads

PacBio CLR (continuous long reads) & Nanopore

Error correction between molecules

- FALCON
- MECAT
- CANU

PacBio HiFi

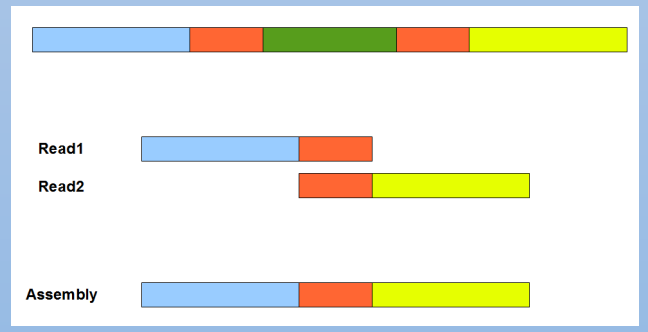
Error correction through CCS (single-**molecule** consensus)

- PacBio SMRT Analysis

Contiging: two major types of software

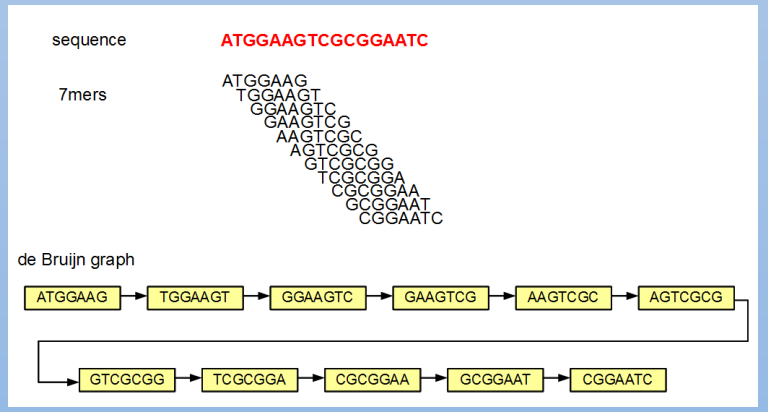
Long reads (>10kb)

overlap–layout–consensus



Short reads (2x150bp)

de-bruijn-graph



source: http://www.homolog.us/Tutorials/index.php

Canu, Falson, Flye, et al.

Spades, Abyss, et al

de-bruijn-graph

K=17

TACGGGAAATTTAGATC

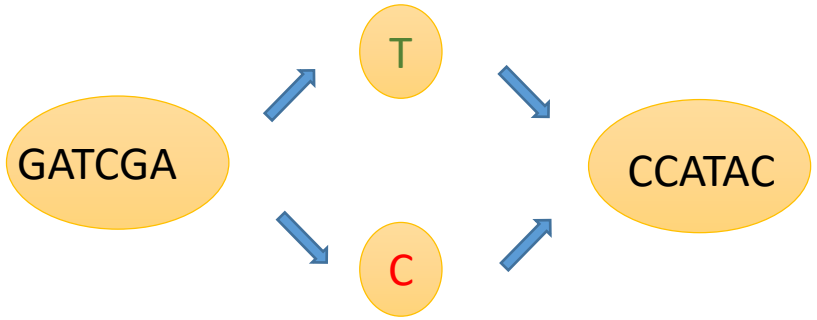
ACGGGAAATTTAGATCG

CGGGAAATTTAGATCGA

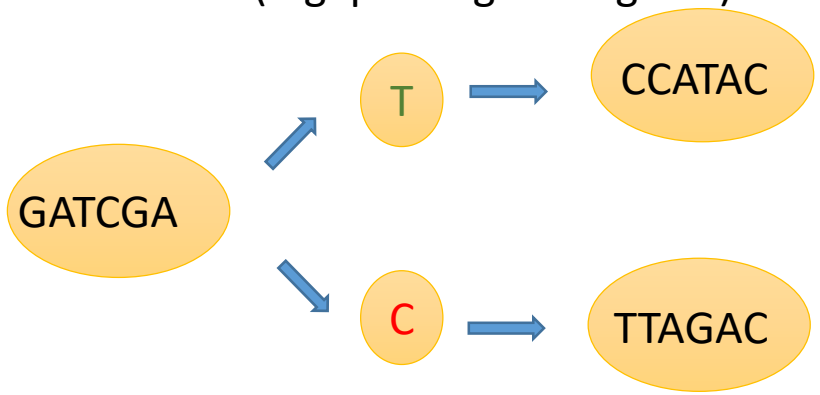
GGGAAATTTAGATCGAT

GGGAAATTTAGATCGAC

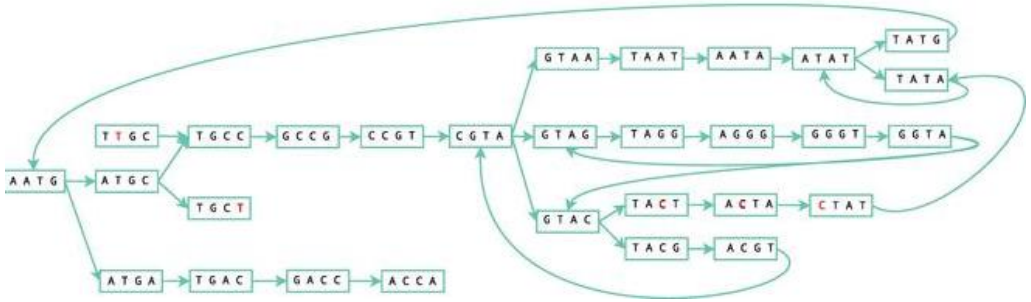
Bubble (e.g. sequencing errors)



Crosslinks (e.g. paralogous regions)

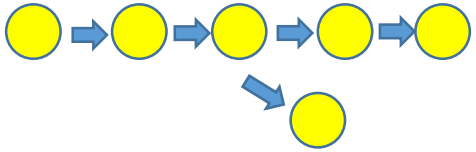


de-bruijn-graph network and how to solve

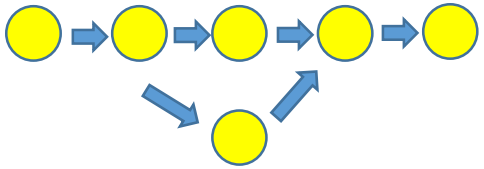


Tips, bubbles and crosslinks

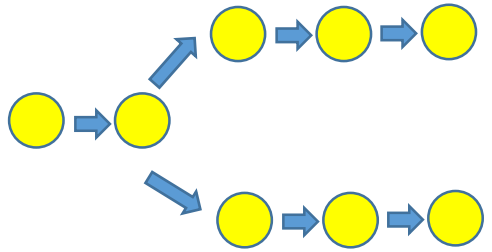
Tips



Bubbles



Crosslinks



Determine optimal Kmer size

Short K-mers → more branching

TACGGG

ACGGGC

ACGGGT

ACGGGA

Long K-mers → more gaps

TACGGGACTATATAGACTACTAGAGCTAG

ACGGGACTATATAGACTACTAGAGCTAGA

CGGGACTATATAGACTACTAGAGCTAGAT

Missing kmer

N50 with different kmer (kb)

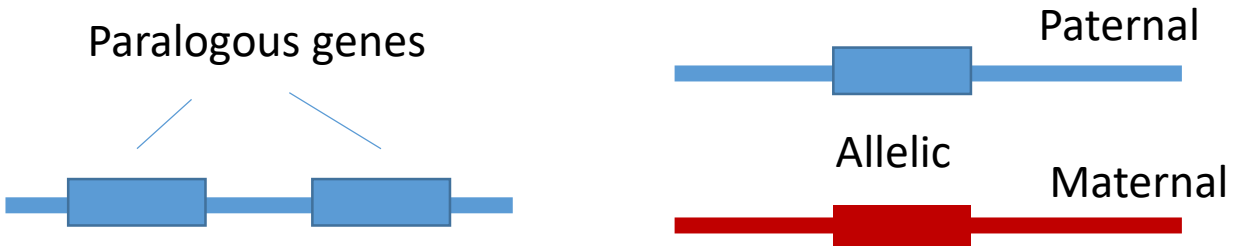
Kmer size	75	95	105	115
contig N50	268	476	476	268
scaffold N50	543	543	543	268

Read length: 199 bp Coverage: ~100 x

SPAdes: use a series of kmers.

To solve the de bruijn network, the software is tuned to collapse allelic genes, but not paralogous genes;

(software is normally tuned for human genome)



<u>De bruijn method not suitable for</u>

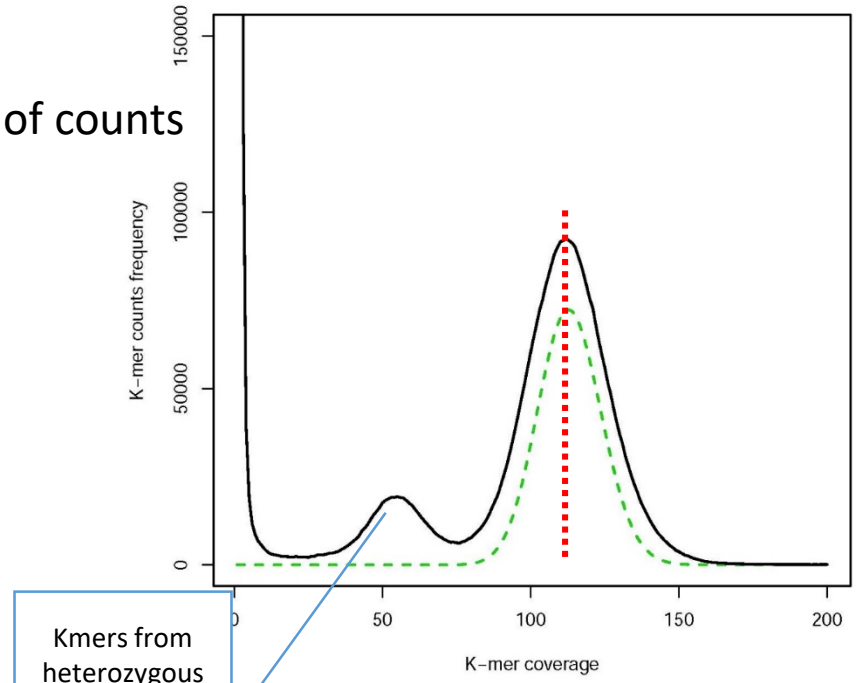
- Highly heterozysgous;
- Polyploidy;
- Highly repetitive genomes;

Kmer distribution can be used to estimate genome size, heterozygosity, and repeat contents

Sequencing data: 20 GB
Coverage: 10x
Genome size = 2GB

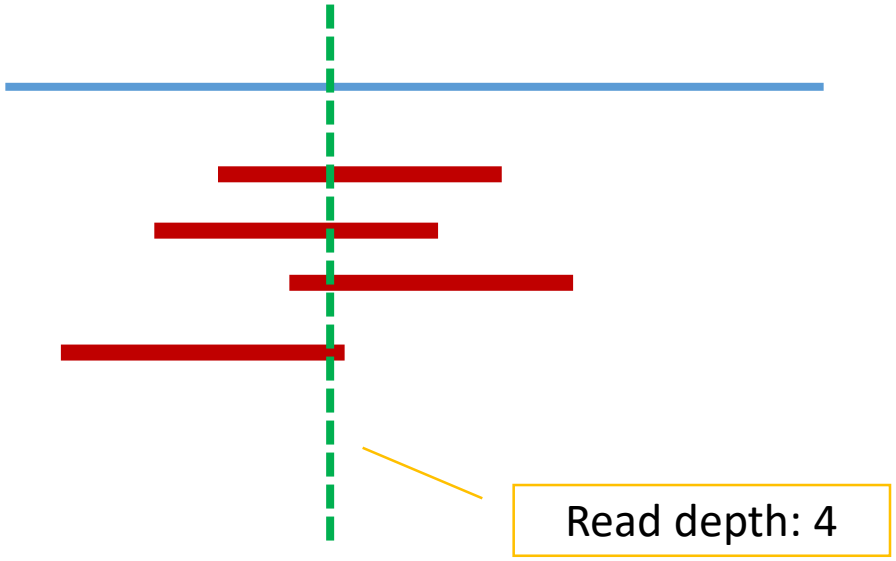
$$\text{Genome size} = \frac{\text{Total base pairs}}{\text{Coverage}}$$

Calculated by modeling kmer count distribution

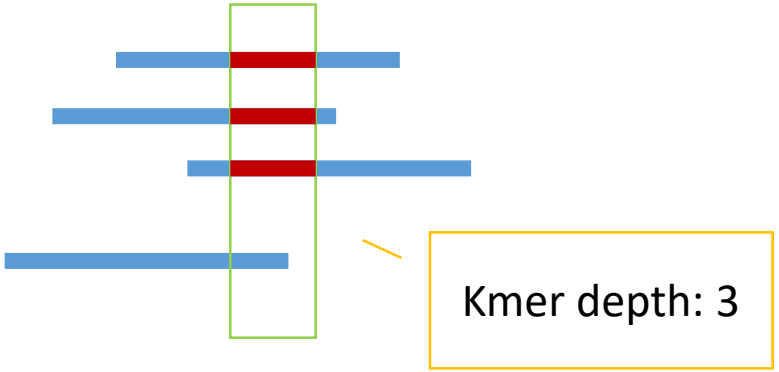


Read depth vs Kmer depth

Read depth: number of reads at a genome position

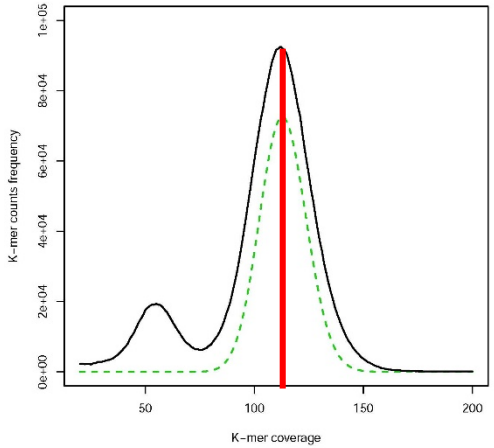


Kmer depth: number of occurrence of each kmer



$$N = M^* L/(L-K+1)$$

M: kmer depth = 112
L: read length = 101 bp
K: Kmer size =21 bp
N: read depth =140



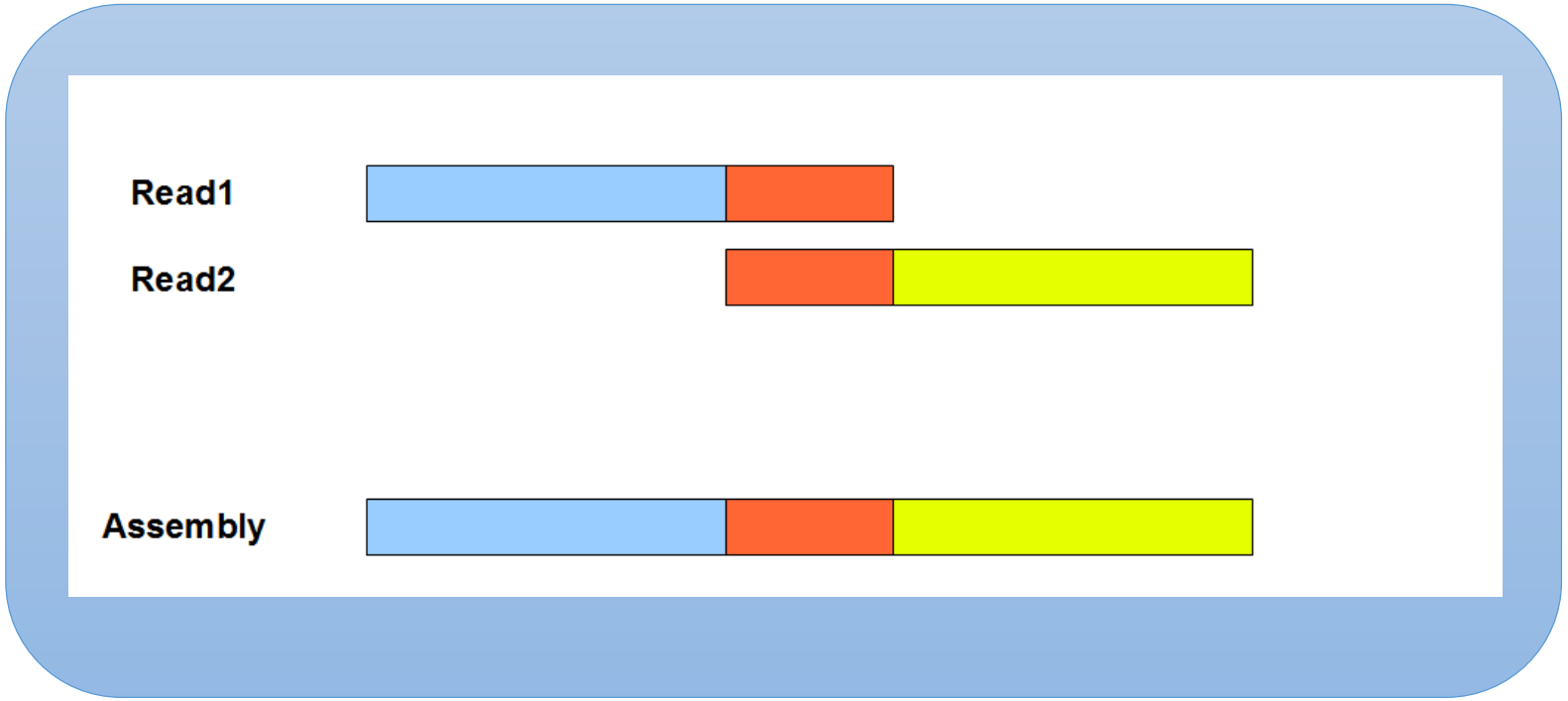
Software to estimate genome size:

ErrorCorrectReads.pl (from ALLPATHS-LG)

```
ErrorCorrectReads.pl \
PAIRED_READS_A_IN=R1.fastq.gz \
PAIRED_READS_B_IN=R2.fastq.gz  \
KEEP_KMER_SPECTRA=1  \
PHRED_ENCODING=33 \
PLOIDY=1  \
READS_OUT=corrected_out \
>& report.log &
```

Long reads

Overlap–Layout–Consensus (OLC)



- PacBio
- Oxford Nanopore

Long-read Sequencing Platform: **PacBio SMRT**

CLR: 5 - 50kb, 10% error rate

- For genomes <10 mb (multiplexed);
- High coverage (>40x) required;
- DNA Methylation detection.

HiFi: 15 – 25 kb, 0.1% error rate

- For large genomes;
- Low coverage ok (20x).

Long-read Sequencing Platform: Oxford Nanopore

Minion

Advantage:

Portable and cheap sequencer;

Disadvantage:

- High error rate (indel in homopolymer regions);
- Require Illumina reads for error correction.

* Good for reference guided analysis

Technical challenge: High Molecular Weight DNA extraction

DNA fragment length

vs

Sequencing read length

Long reads assembly software:

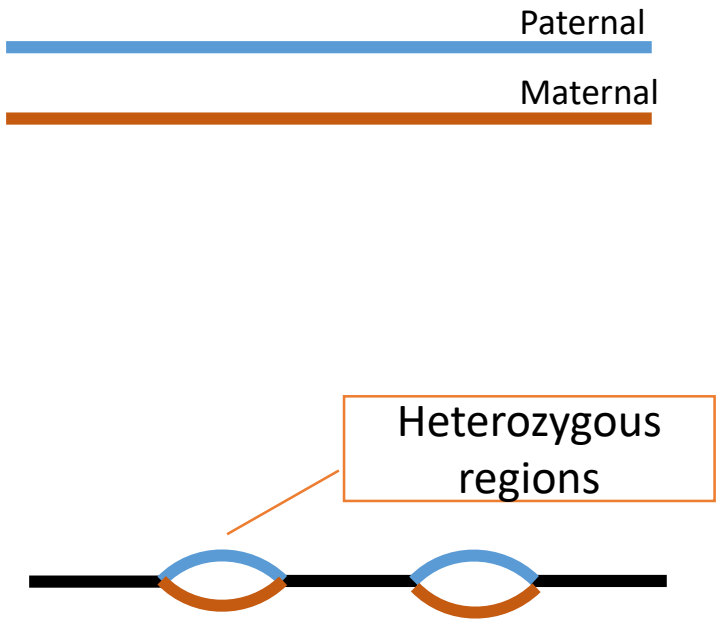
Error correction: FALCON, MECAT, CANU, et al

Assembly: CANU, HiCANU (open source), FALCON (by PacBio)

* Error correction could take weeks, and generates very large temporary files, could be >20TB for a large genome

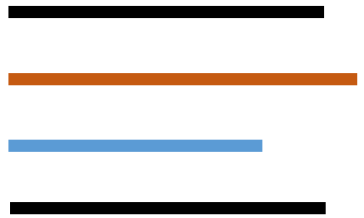
Assembly results – concept of haplotigs

A diploid genome



Assembly results

Haplotigs

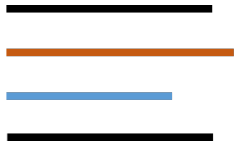


Random phased assembly

CANU output files

<prefix>.contigs.fasta

Primary assembly.



<prefix>.unitigs.fasta

All haplotigs

<prefix>.unassembled.fasta

Contigs with poor read support (>50% region with <3X coverage)

Preprocessing

Contiging

Polishing

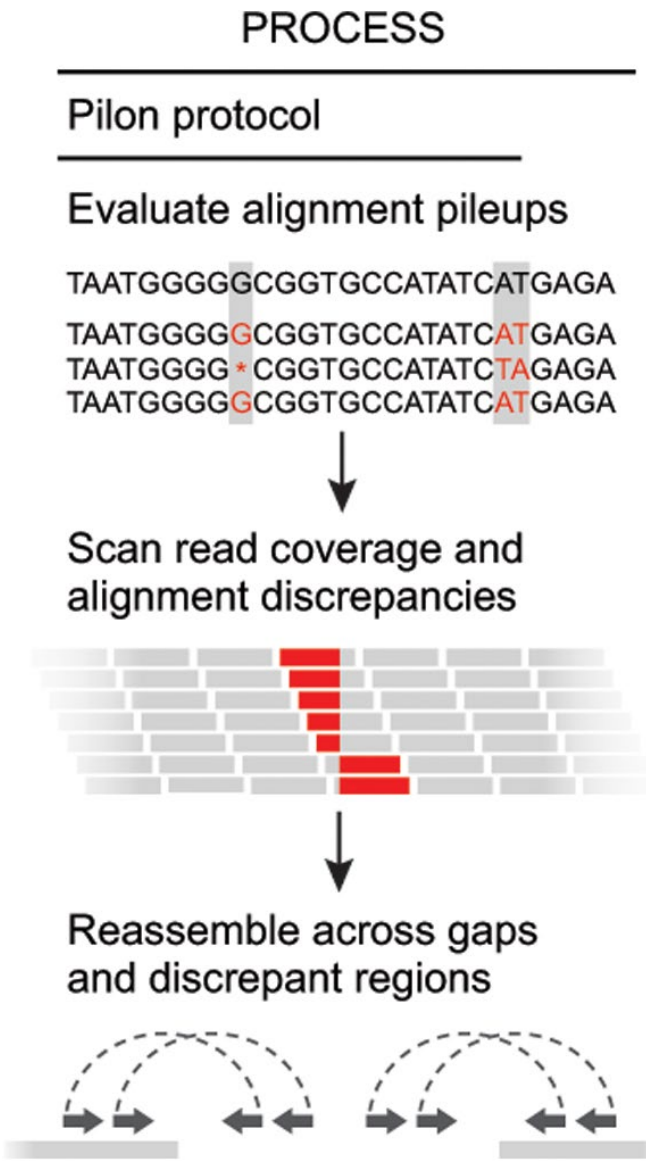
Scaffolding

Polishing with pilon

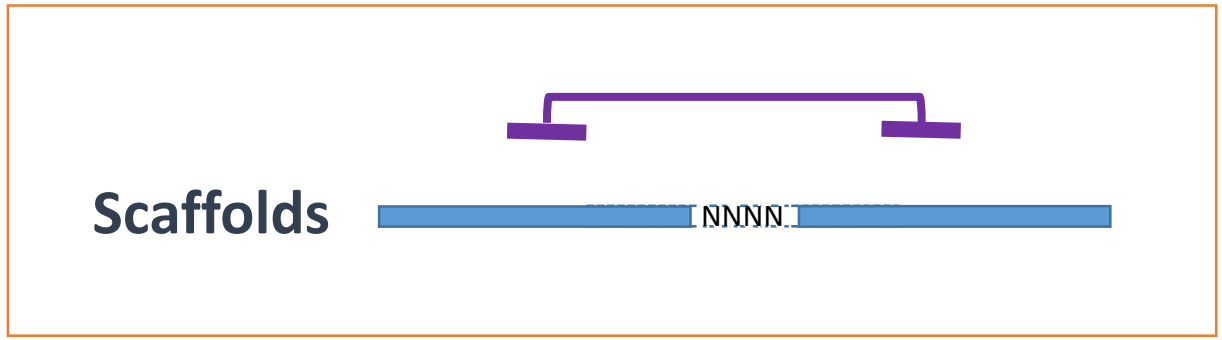
Polishing software:

- pilon (polishing with short reads)
- Recon (polishing with long/short reads)

SPAdes has a “--careful” option that does error correction with read alignment

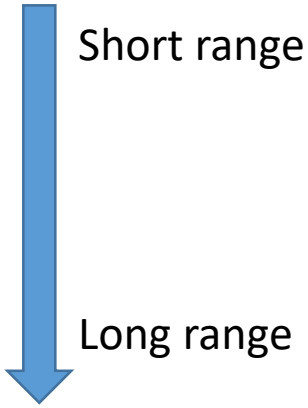


Scaffolding contigs



Technologies

- **Long-read:** PacBio or Nanopore
- **BioNano:** Optical Mapping:
- **Hi-C:** Dovetail; Phase Genomics

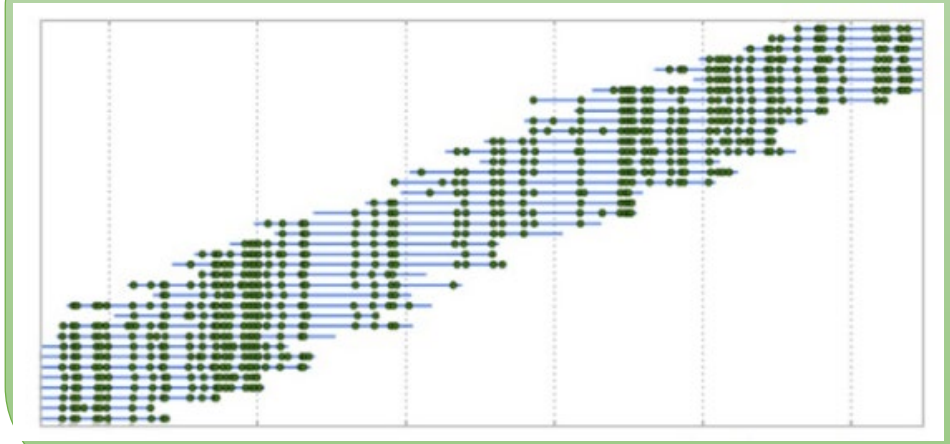


Scaffolding strategies: Physical maps

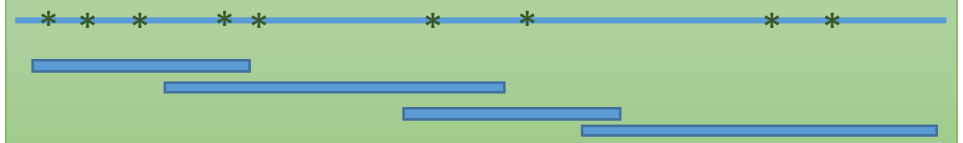
BioNano optical map

- Label 7-mer nickase recognition sites;
- Measure fragment length

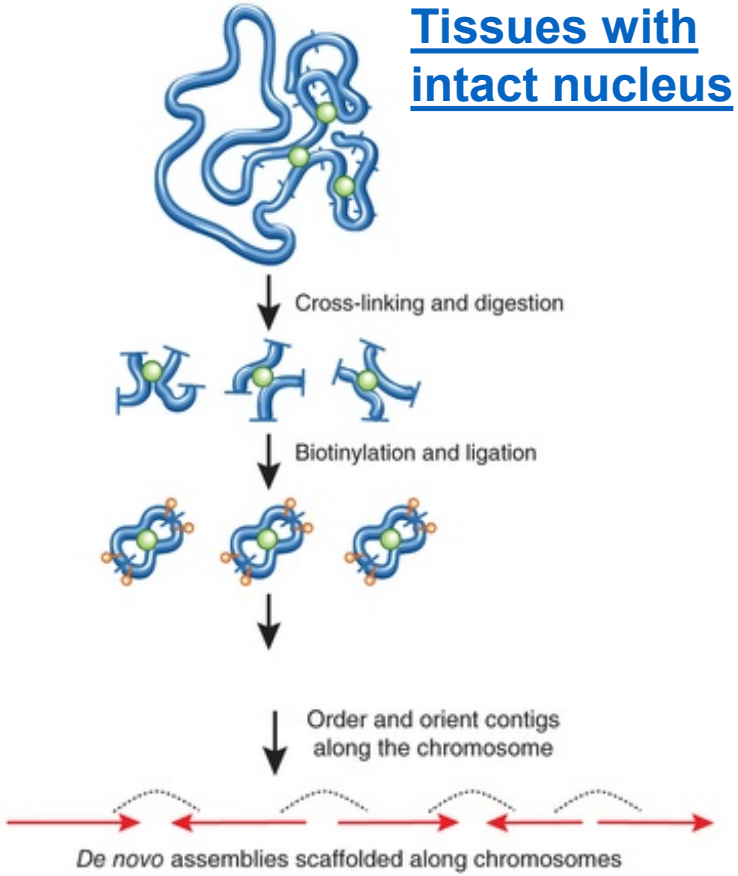
Assemble the optical map



Place contig to the optical map

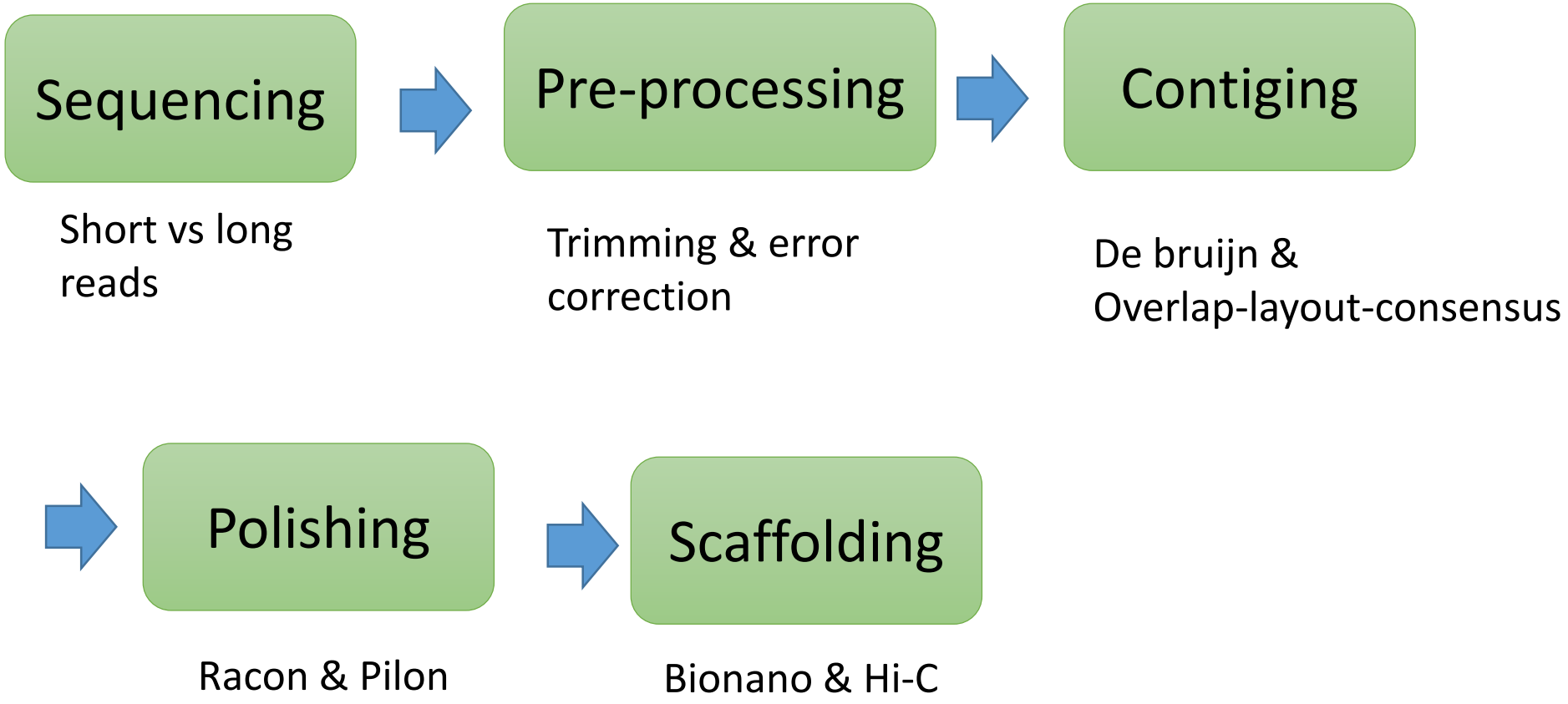


Hi-C:



Commercial services:
Dovetail & Phased Genomics

Summary of genome assembly



What is trending this year:

- PacBio HiFi (20kb long reads with 99.9% accuracy);
- Chromosomal level finish;

Assessment of assembled genome

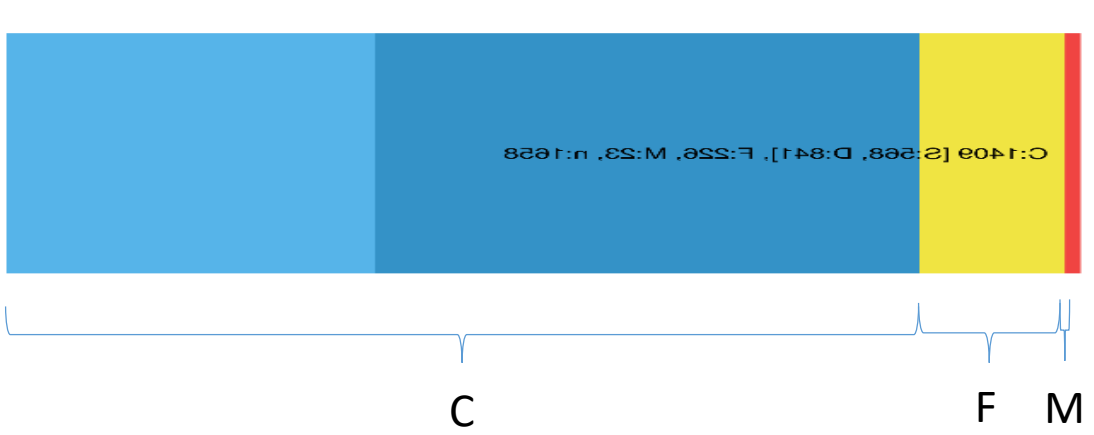
- Completeness
  - Compare assembly size to estimated genome size;
  - Gene space completeness: **BUSCO**
- Contiguity
  - N50
- Accuracy of contig/scaffold
  - Chimerics;
  - Collapsed paralogs;

Concept of BUSCO

Evgeny M Zdobnov lab, University of Geneva
https://busco.ezlab.org/

BUSCO Score

C: 85% [S:34%,D:51%],
F: 14%,
M: 1%,
(n: 1658)



C: Complete
- S: single copy;
- D: duplicated;

F: Fragmented;
M: Missing;
n: Total groups;

busco --list-datasets

BUSCO “Core” gene database

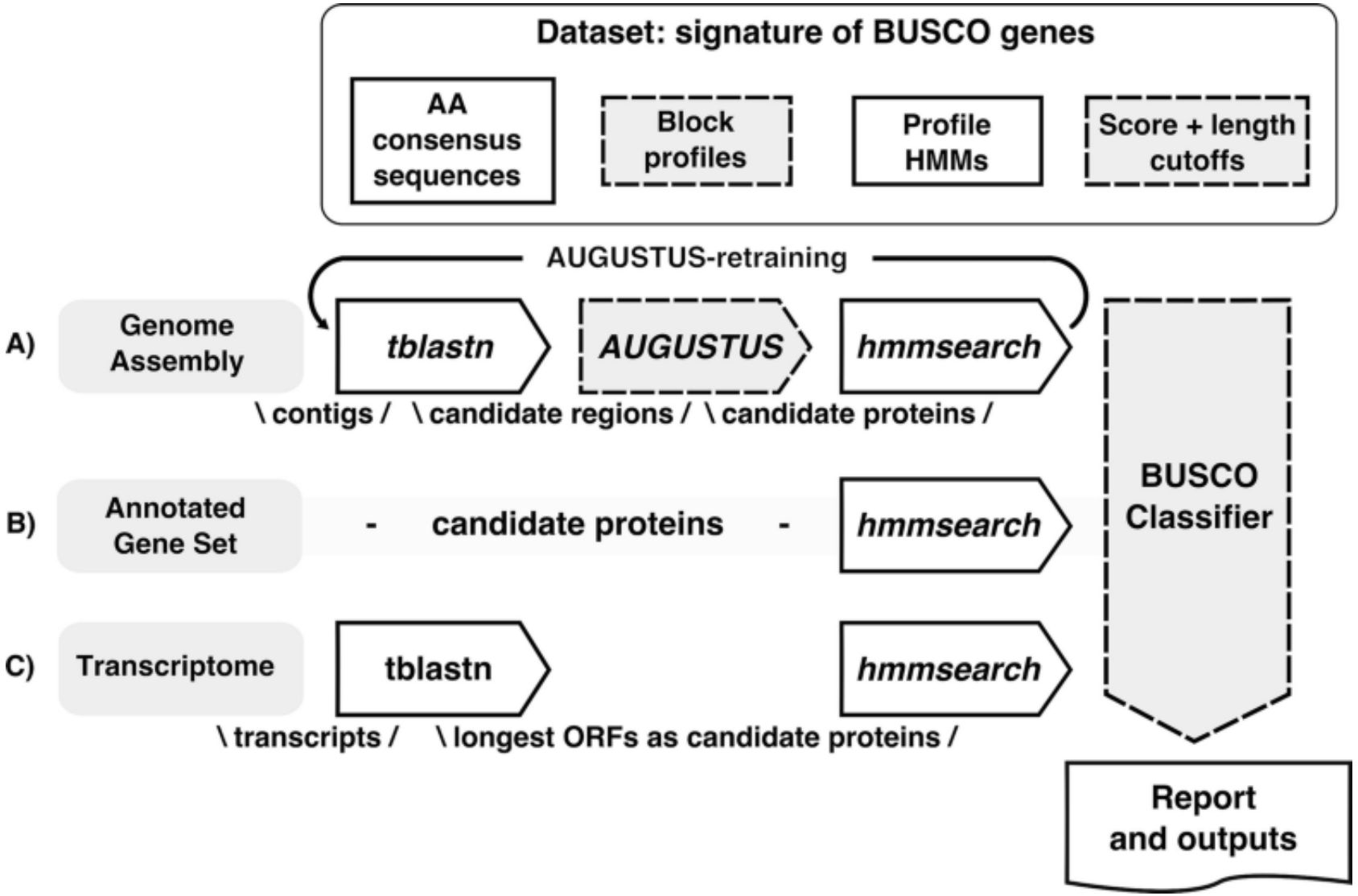
Vertebrates:

- vertebrata_odb10
  - actinopterygii_odb10
    - cyprinodontiformes_odb10
  - tetrapoda_odb10
    - mammalia_odb10
      - eutheria_odb10
        - euarchontoglires_odb10
          - glires_odb10
          - primates_odb10
        - laurasiatheria_odb10
          - carnivora_odb10
          - cetartiodactyla_odb10
    - sauropsida_odb10
      - aves_odb10
        - passeriformes_odb10

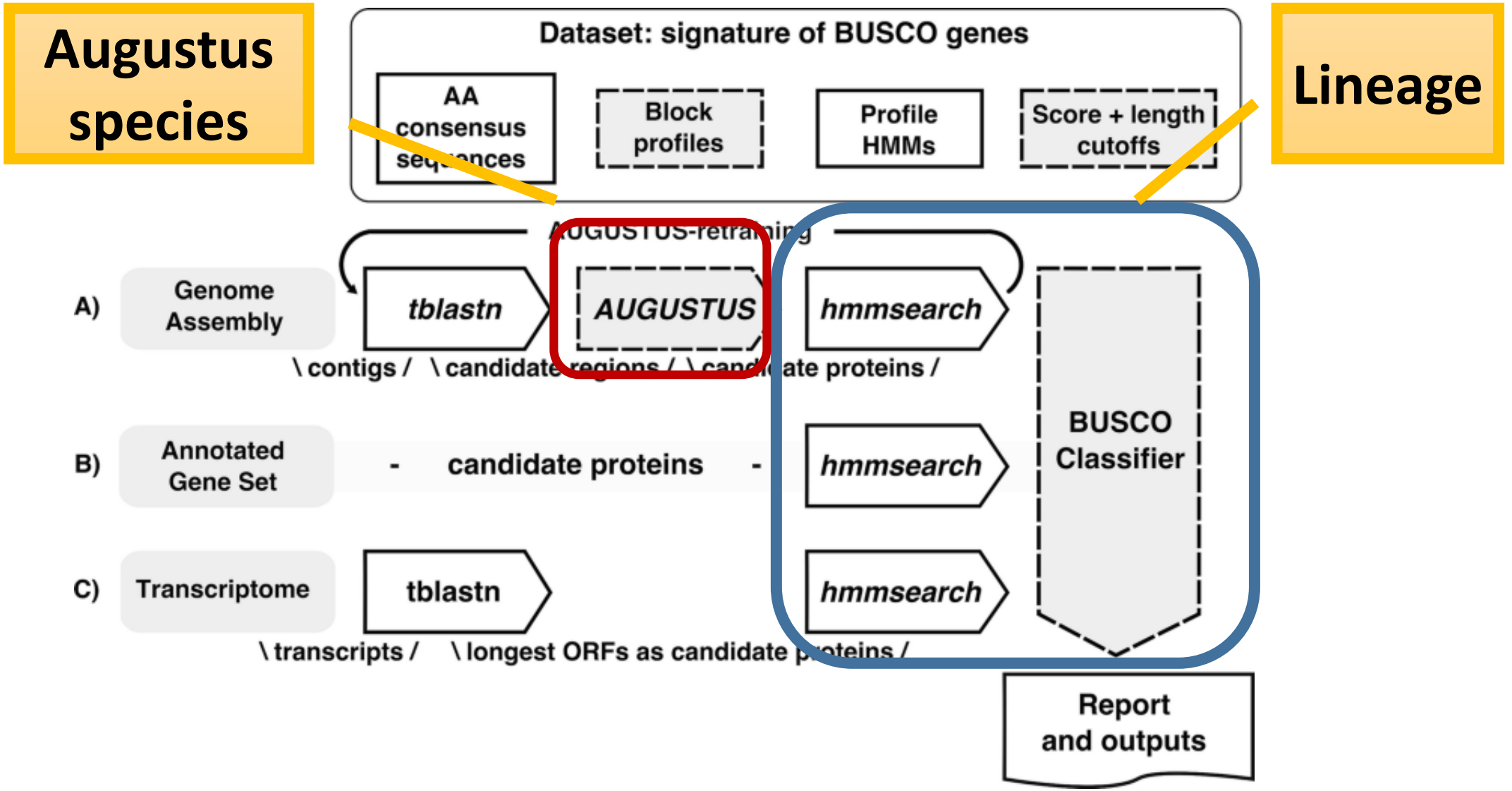
For example, for Leemar genome, any one of these lineages is valid:

- vertebrata_odb10
- tetrapoda_odb10
- mammalia_odb10
- eutheria_odb10
- euarchontoglires_odb10
- primates_odb10

BUSCO pipeline



BUSCO pipeline



Run BUSCO:

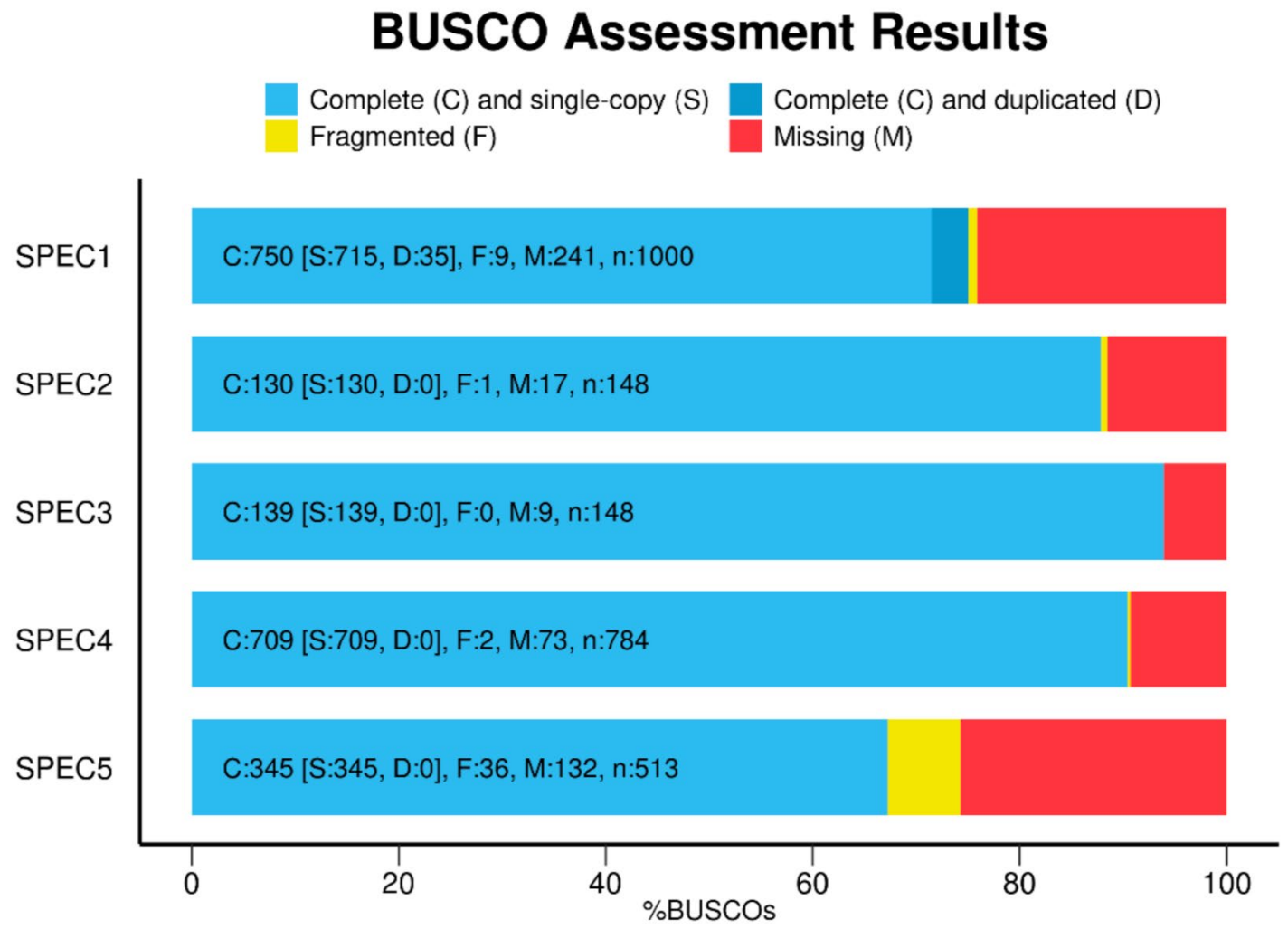
```
busco -i input.fasta \
        -o outputfile \
        -l [LINEAGE]  \   # e.g. vertebrate, plant
        -m [MODE]         #genome, transcriptome, protein
```

Other parameters:

- --augustus_species species model for gene prediction. Default species is based on lineage setting. E.g. default for vertebrate is “human”.

(When run BUSCO on a shark genome, set augustus_species to human is much better than “elephant shark”, probably due to high quality human model)

A good genome should have BUSCO score 90 or higher
(blue regions in the plot)



Evaluation of Genome assembly: Contiguity

N50 and L50 *

N50 50% (base pairs) of the assemblies are contigs above this size.

L50 Number of contigs greater than the N50 length.

NG50 and LG50

N50 is calculated based on assembly size. NG50 is calculated based on estimated genome size.

Genome Accuracy: Chimerics and Collapsed paralogs

- Physical map: Bionano & Hi-C
- Genetic linkage

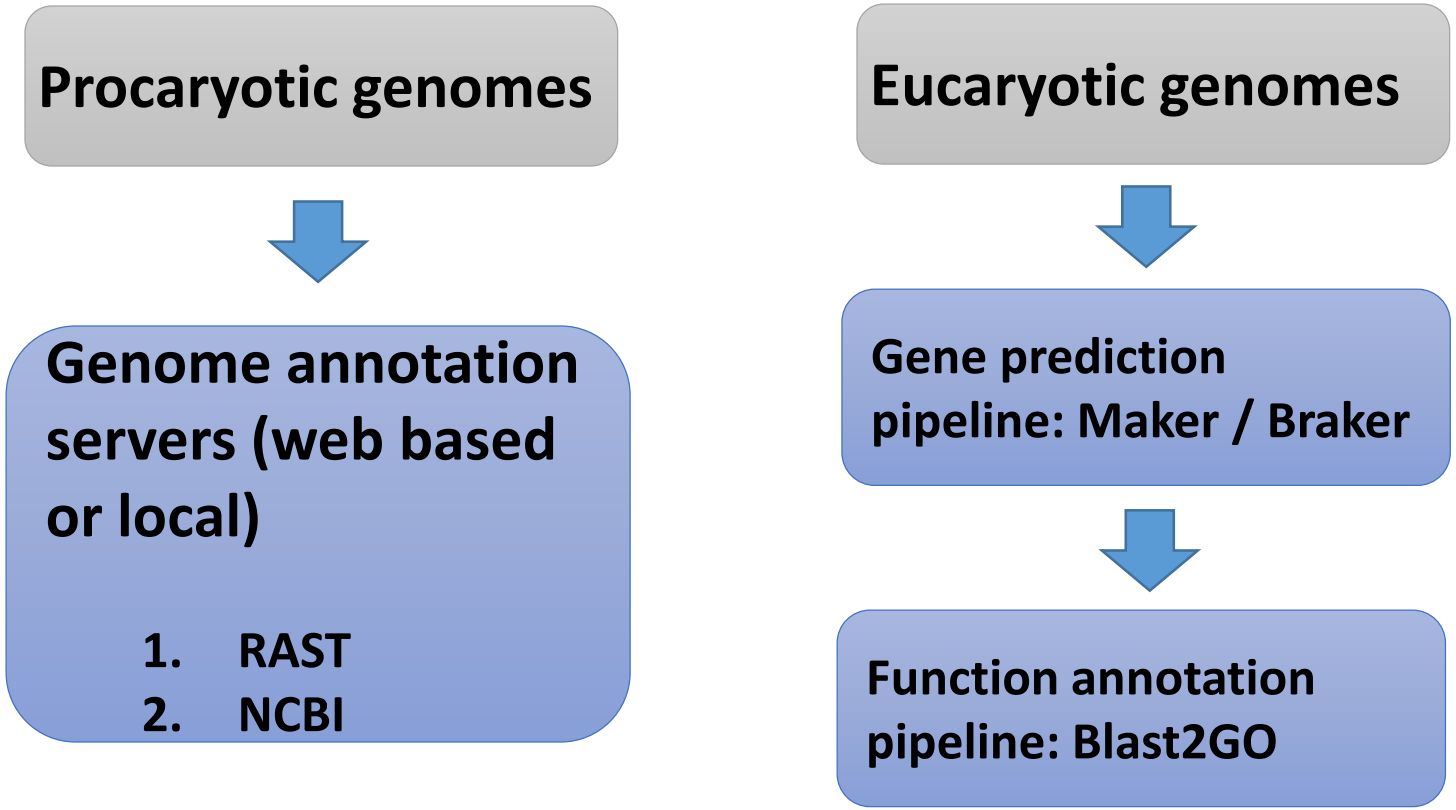
How to correct?

BioNano & Hi-C

Summary of assembly evaluation:

- Genome size (estimated vs assembled);
- N50 size;
- BUSCO score;
- Accuracy (optional for now)

From assembled genome to annotated genome



What software Cornell people are using (2019 - 2020)

SPAdes	54.8%
supernova	30.1%
canu	17.4%
centroFlye	8.5%
velvet	2.7%
Unicycler	2.4%
MaSuRCA	1.7%
Flye	1.4%
SOAPdenovo2	1.2%
SKESA	1.1%
platanus	0.6%
abyss	0.2%
necat	0.2%
mccortex	0.1%
FALCON	0.0%

(long read assembler in brown)

* Percentage of time at least one job in BioHPC is running this software)