De novo genome assembly

Computational Genomics | Lecture 14

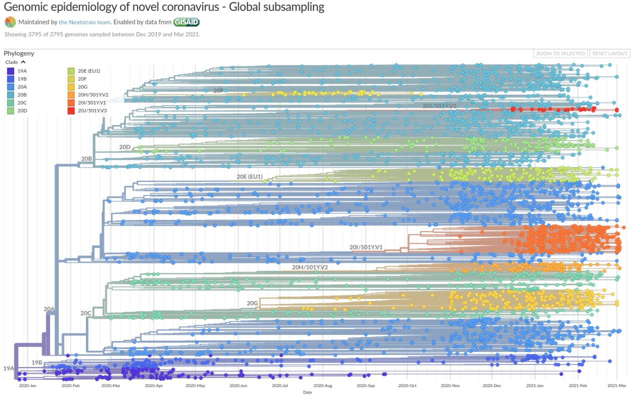
Tom Harrop

melbournebioinformatics.org.au

De novo genome assembly

1. Introduction to genome assemblies
2. Assembly software and algorithms
3. Assembly metrics and scaffolding
4. Live session: an example hybrid assembly

Why sequence genomes?



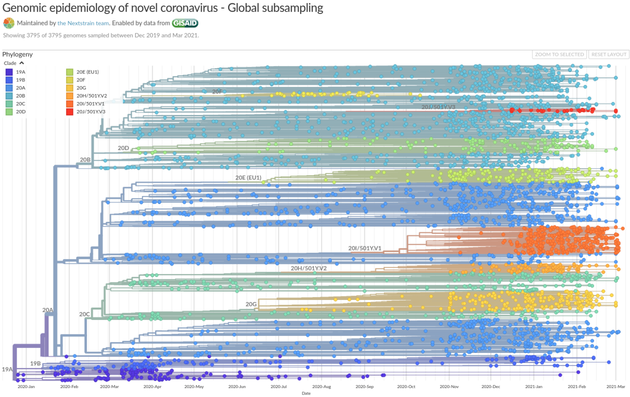
Genomic epidemiology of novel coronavirus - Global subsampling

Maintained by the Nextstrain team. Enabled by data from GISAID

Showing 3795 of 3795 genomes sampled between Dec 2019 and Mar 2021.

- compare differences between species

Why sequence genomes?



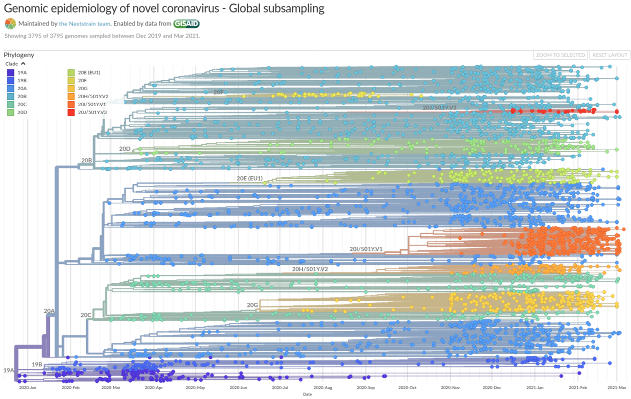
Genomic epidemiology of novel coronavirus - Global subsampling

Maintained by the Nextstrain team. Enabled by data from GISAID

Showing 3795 of 3795 genomes sampled between Dec 2019 and Mar 2021.

- compare differences between species
- compare variants of a species or population

Why sequence genomes?



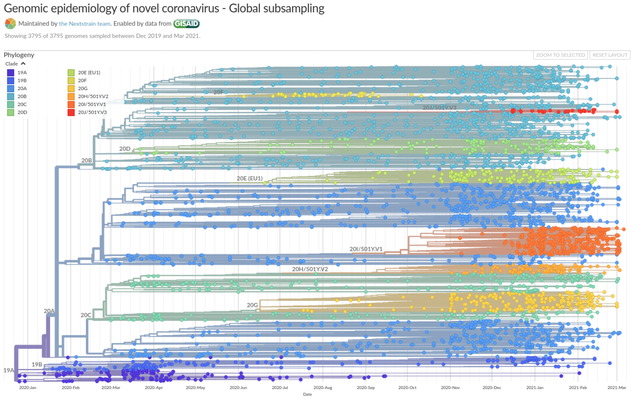
Genomic epidemiology of novel coronavirus - Global subsampling

Maintained by the Nextstrain team. Enabled by data from GISAID

Showing 3795 of 3795 genomes sampled between Dec 2019 and Mar 2021.

- compare differences between species
- compare variants of a species or population
- research diseases

Why sequence genomes?



Genomic epidemiology of novel coronavirus - Global subsampling

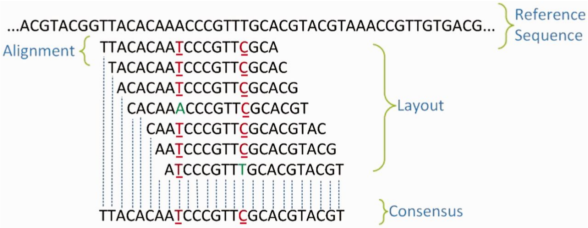
Maintained by the Nextstrain team. Enabled by data from GISAID

Showing 3795 of 3795 genomes sampled between Dec 2019 and Mar 2021.

- compare differences between species
- compare variants of a species or population
- research diseases
- provide a reference for gene expression analysis

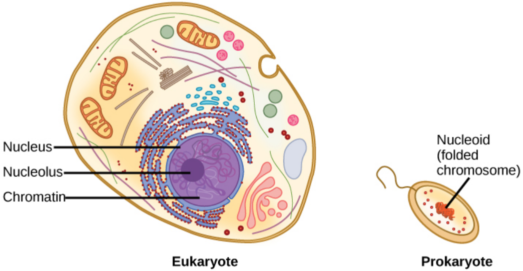
How is *de novo* assembly different to multiple sequence alignment?

- ***de novo***: from scratch, without a reference
- literally: anew, over again from the beginning
- sequencing reads ± structural information → genome sequence



Wajid & Serpedin, 2016

Prokaryotic and eukaryotic genomes



CNX OpenStax via Wikimedia Commons

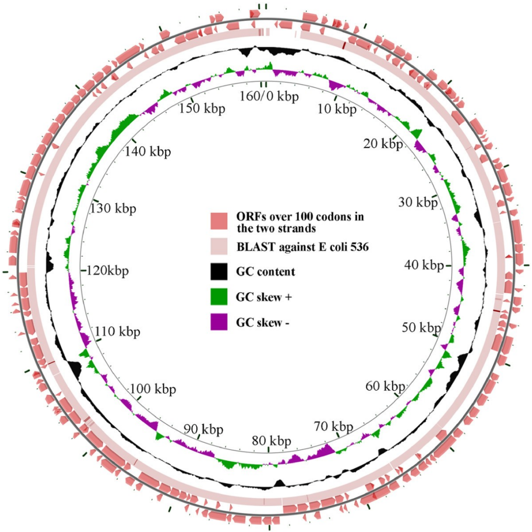
In prokaryotes:

- chromosomes (usually one)
  - genes rarely have introns
  - *coding dense*
- plasmids
- bacteriophage

In eukaryotes:

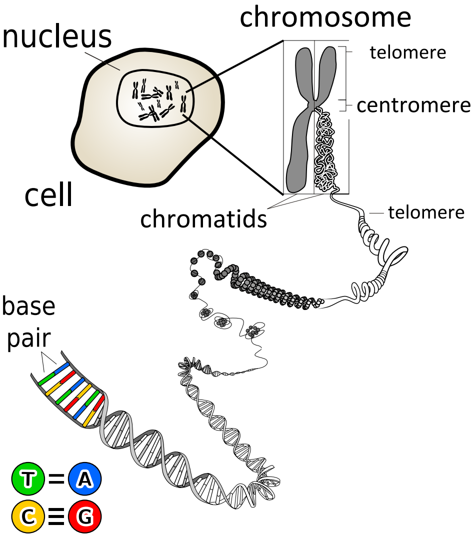
- nuclear genome (chromosomes)
  - genes ± introns
  - non-coding elements
  - mobile elements
  - centromeres
  - telomeres
- mitochondria
- chloroplasts

Bacterial genomes can be dense



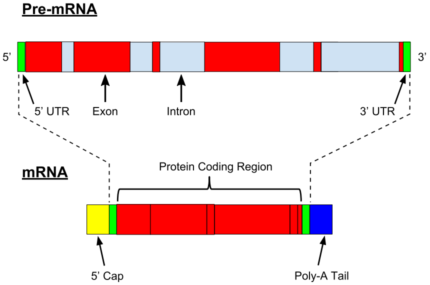
Darling *et al.*, 2014

The nuclear genome of eukaryotes



KES47 via Wikimedia Commons

The nuclear genome of eukaryotes

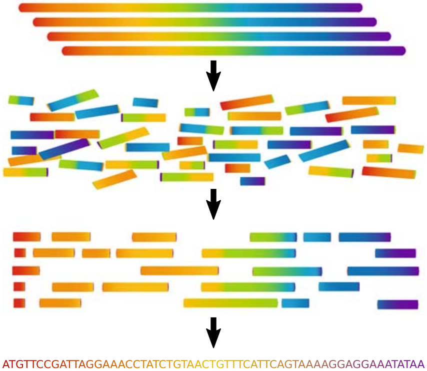


Non-coding sequences

- Telomeres, centromeres
- Introns and untranslated regions
- Regulatory elements
- Pseudogenes
- Repetitive sequences *e.g.* mobile elements

Nastypatty via Wikimedia Commons

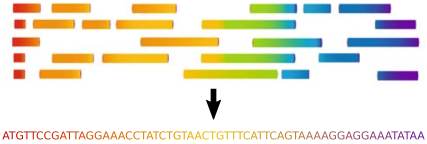
Genome assembly concepts



Commins *et al*., 2009

- The genome is fragmented for sequencing
- The sequencing ***reads*** might be
  - 100-350 b long (Illumina)
  - ~20 kb long (PacBio HiFi)
  - up to a few hundred thousand bases long (Nanopore)
- ***Assembly*** is the process of reconstructing the genome from the sequenced reads
- It’s not always possible to assemble the complete sequence

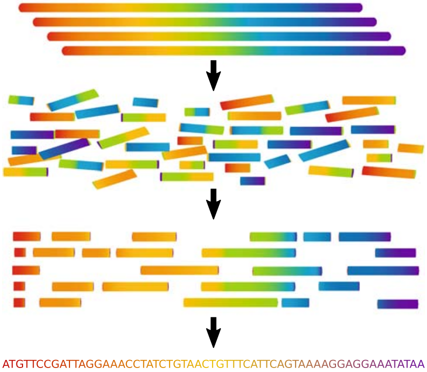
Sequencing coverage



- aim to cover each base > 30 times
- final sequence is the ***consensus*** of all the reads covering that base
- $1\ \text{Gb} \times 30\times\ \text{coverage} = 30\ \text{Gb}$
- $\frac{30\ \text{Gb}}{150\ \text{b}} = 200$ million reads
- using PacBio reads, with an average length of 20 kb?

Commins *et al.*, 2009

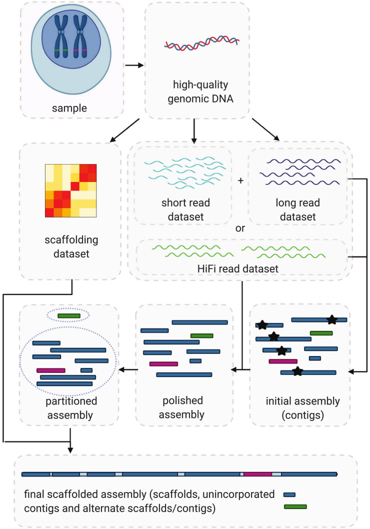
Sequencing strategies for genome assembly



Commins *et al.*, 2009

- ***Hierarchical shotgun*** Sanger sequencing
- Short read, Illumina sequencing
  - 100–350 b
  - sometimes called ***high-throughput***, ***next-generation*** (!) or ***2nd-generation sequencing***
  - good for draft assemblies of eukaryote genomes
- Long read (***third-generation***) sequencing
  - PacBio: ~ 20 kb reads
  - Nanpore: up to 100s of kb, read N_{50} usually > 20 kb
  - expect much better contiguity, but can have accuracy issues

Hybrid genome assembly



Whibley *et al.*, 2020

- ***hybrid assemblies*** combine long and short reads
- scaffolding the hybrid assembly can generate chromosome-level assemblies