# *De novo* genome assembly

Computational Genomics | Lecture 14

Tom Harrop

melbournebioinformatics.org.au

# How good is my assembly?

**Assembly qualities** (jargon!):

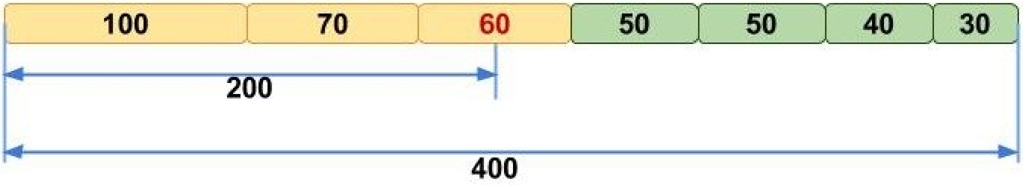
- draft
- reference
- chromosome-level
- complete/closed, telomere to telomere (T2T)

**High-quality genomes** have[1]:

- contig $N_{50}$ length ≥ 1 Mb
- scaffold $N_{50}$ length ≥ 10 Mb
- ≥ 90% of sequence assigned to chromosomes
- error rate ≤ 0.01% (1 error in 10,000 bases)

[1] according to the VGP

# $N_{50}$ is a contiguity statistic



***The sequence length of the shortest contig at 50% of the total genome length***

Is this $N_{50}$ ($N$ → number)?
Or $L_{50}$ ($L$ → length)?
***$N_{50}$ length***?
🤷

The Molecular Ecologist

# Naïve $N_{50}$ in python3

```
> #!/usr/bin/env python3
+
+ import numpy as np
+
+ contig_lengths = [
+   40, 100, 70, 50, 60, 50, 30]
+ sorted_lengths = sorted(
+   contig_lengths,
+   reverse=True)
+
+ genome_size = sum(sorted_lengths)
+ print(genome_size)
```

```
400
```

```
> i = [
+   x >= genome_size * 0.5
+   for x in np.cumsum(sorted_lengths)
+   ].index(True)
+
+ print(i)
```

```
2
```

```
> sorted_lengths[i]
```

```
60
```

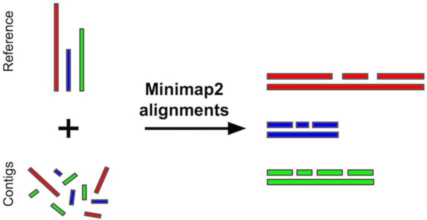
# How do we know if we have a good assembly?

| Metric | Draft | Target | VGP | Finished |
|---|---|---|---|---|
| Contig $N_{50}$ | > 10 Kb | > 1 Mb | 1–25 Mb | Chr |
| Scaffold $N_{50}$ | > 100 Kb | > 10 Mb | 23–480 Mb | Chr |
| Gaps | < 10,000 | < 1000 | 75–1500 | None |
| Completeness | > 80% | > 90% | 87–98% | 100% |
| Genes (BUSCO) | > 80% | > 90% | 82–98% | > 98% |
| Mappability | > 70% | > 80% | 96% | 98% |

$N_{50}$ describes the distribution of contig or scaffold sizes.
Higher $N_{50}$ means the assembly is in bigger chunks.

VPG assembly metrics, Rhie *et al.*, 2020

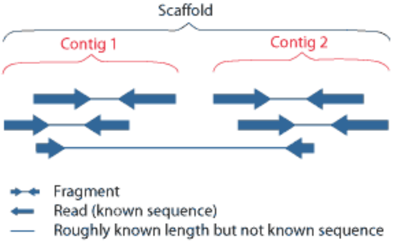
# Scaffolding



- **reference genome**
- mate pairs
- long reads with a short-read assembly
- proximity ligation
- optical mapping

Alonge *et al.*, 2019

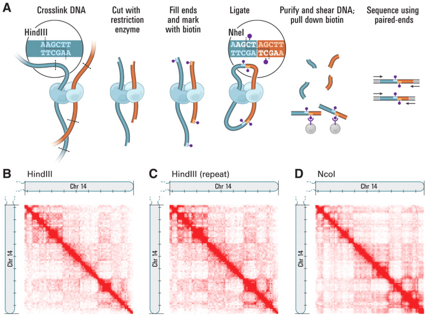
# Scaffolding



University of California via Wikimedia Commons

- reference genome
- **mate pairs**
- **long reads with a short-read assembly**
- proximity ligation
- optical mapping
- linkage mapping

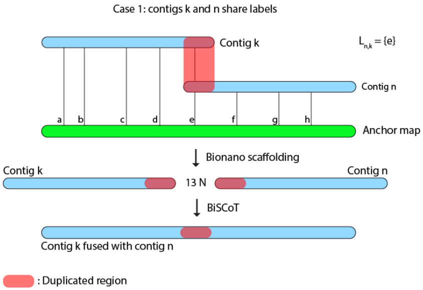
# Scaffolding



- reference genome
- mate pairs
- long reads with a short-read assembly
- **proximity ligation**
- optical mapping
- linkage mapping

Lieberman-Aiden *et al.*, 2009

# Scaffolding



- reference genome
- mate pairs
- long reads with a short-read assembly
- proximity ligation
- **optical mapping**
- **linkage mapping**

Istace *et al.*, 2020

# Summary

**Two major approaches to *de novo* assembly**:

**Short read**

- prokaryotes, simple genomes
- relatively cheap
- high accuracy assemblies, but struggles with complex genomes
- assembly algorithms are fast and efficient

**Long read**

- resolves complicated genomes
- expensive, relatively tricky to generate data
- accuracy a little lower
- assembly is slow and memory-hungry

Scaffolding improves the contiguity in many cases

Hybrid approaches are currently popular