



UK Health  
Security  
Agency

# *De novo assembly*

Dr. David Bibby,  
Genomics and Clinical Virology,  
22<sup>nd</sup> February 2024

# Overview

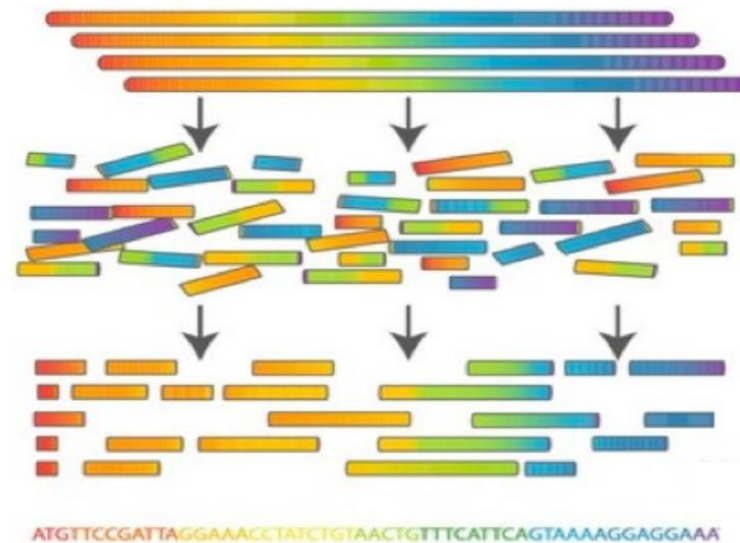
- What is de novo assembly?
  - How does it differ from reference mapping?
- When might it be used?
- How does it work?
- Pitfalls & difficulties with virus data
- Outputs

# What is *de novo* assembly?

Basic definition:

**“The process of reconstructing sample sequence(s) without any guide reference(s)”**

*De novo*: “from the beginning”



**Sample sequences**

**Sequencing reads**

**Assembly**

**Sequence!**

Commins *et al.* Bio. Proc. Online (2009) 11(1)

# What is *de novo* assembly?

“Trying to solve a huge jigsaw puzzle where you don’t have the picture”

But there are also a number of additional confounding issues:

1. All the pieces are blank
2. Some are missing
3. Some are damaged
4. Many are duplicates
5. None are edges
6. Some belong to other puzzles



# When might it be used?

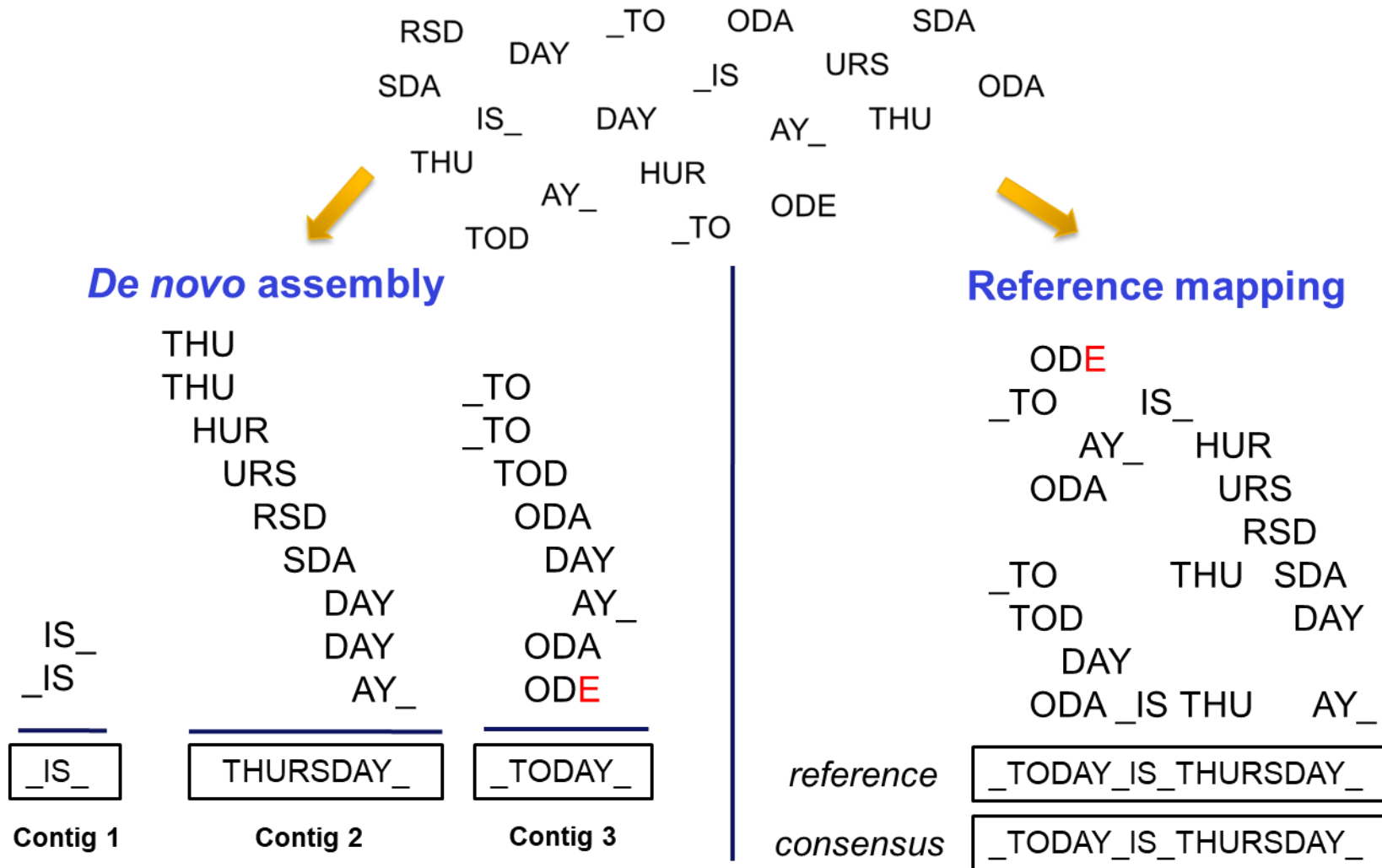
So why not use a reference?

- No suitable references for this virus
- Poorly studied
- Hard to sequence
- Unreliable sequence data
- Highly divergent species (e.g. many RNA viruses)

Target not specified

- Metagenomic analysis
- Syndromic testing
- Unknown aetiological agent of disease

# How does it work?



# How does it work?

## Principles

1. Find overlaps between reads
2. Build a graph of overlaps
3. Correct errors
4. Traverse graph
5. Deliver contigs

## Methods

1. Greedy Approach
2. Overlap-Consensus-Layout
3. De Bruijn graphs
4. Other (e.g. VICUNA)

# How does it work – Greedy Approach?

1. Iterate through pairs of reads
2. If an overlap exists, merge the two reads into one
  - Overlap detection criteria can be parameterized
3. Repeat until no further overlaps are possible

**Slow**

**Inaccurate**

**Easily confounded by repeats**

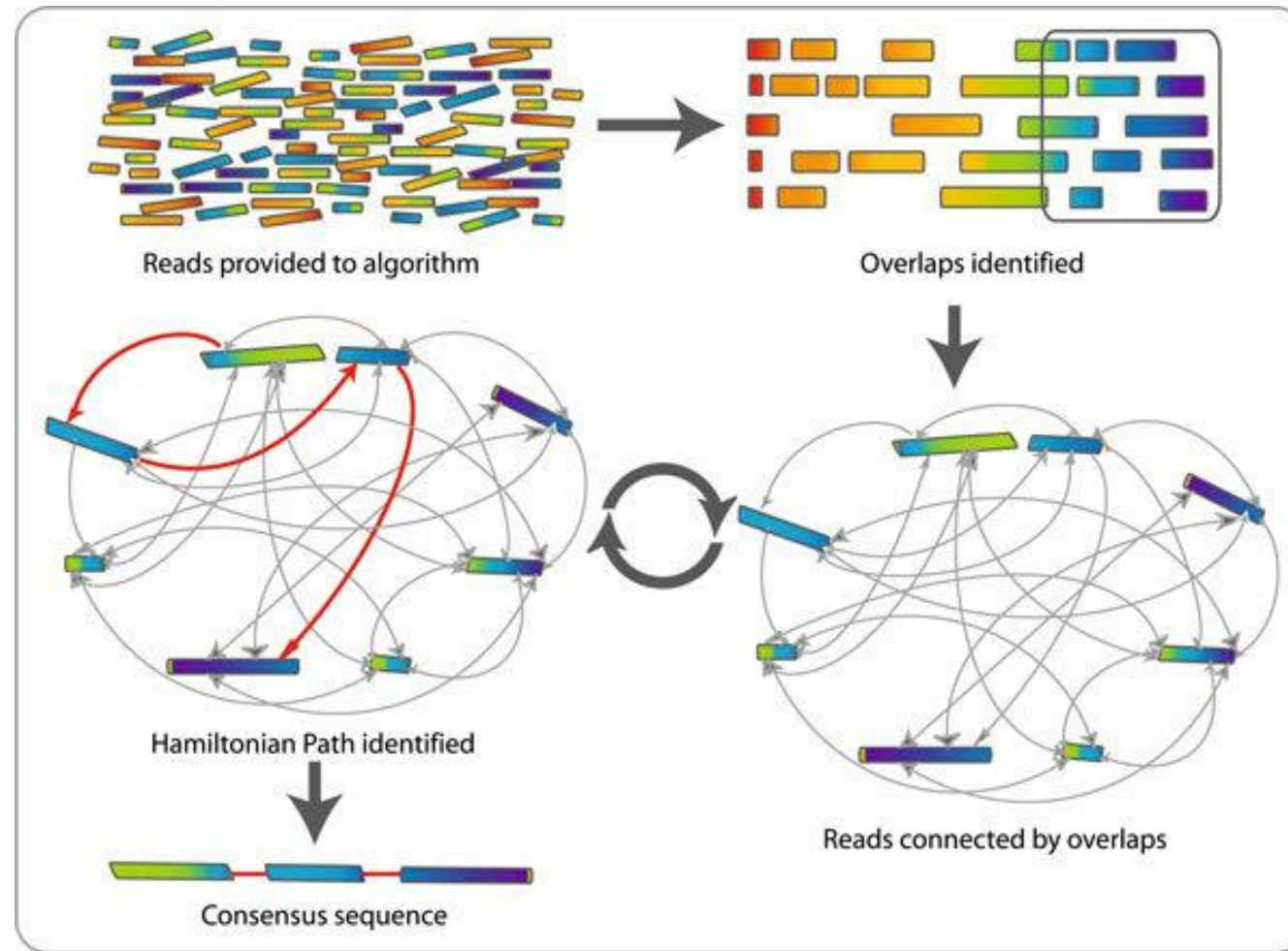


# How does it work – Overlap-Layout-Consensus?

1. Find all overlaps between reads
  - Again, overlap detection criteria can be parameterized
2. Create a graph of all overlaps
3. Traverse graph to find an unambiguous path
  - “Hamiltonian path” (each *vertex* only once)

**Arachne**  
**PCAP**  
**Newbler**  
**Celera Assembler**

# How does it work – Overlap-Layout-Consensus?



Commins *et al.* Bio. Proc. Online (2009) 11(1)

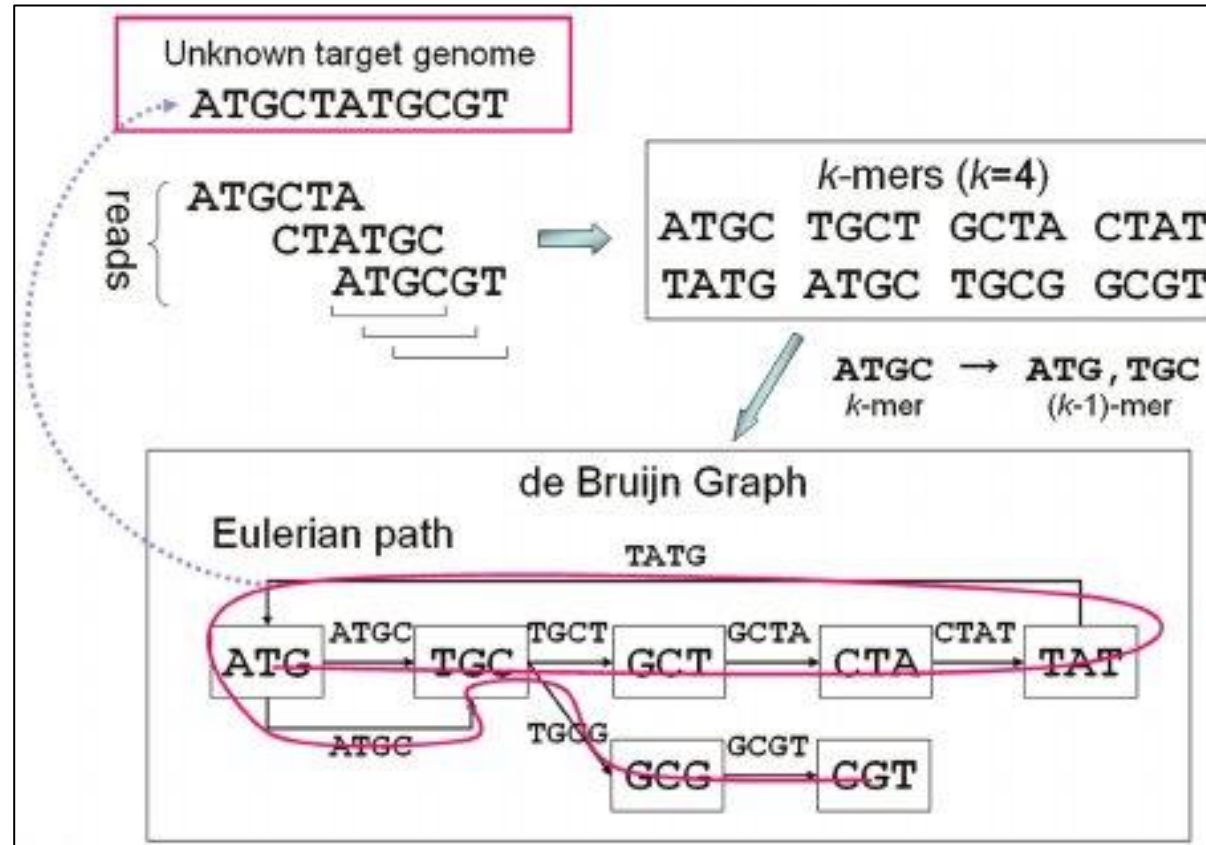
# How does it work – De Bruijn graphs?

1. Derive  $k$ mers from sequence reads
  - Subsequences of length  $k$
2. Create overlap graph using  $[k-1]$ mers
3. Traverse graph to reconstruct likely sequence
  - “Eulerian path” (each *edge* only once)
4. More memory-efficient than OLC

ABYSS  
SOAPdenovo  
Trinity

Ben Langmead, Johns Hopkins  
“De Bruijn Graph assembly” pdf

# How does it work – De Bruijn graphs?



Namiki *et al.* Nuc. Acids. Res. (2012) 40:e155

# Viral data - challenges

## **High levels of variation in the data set**

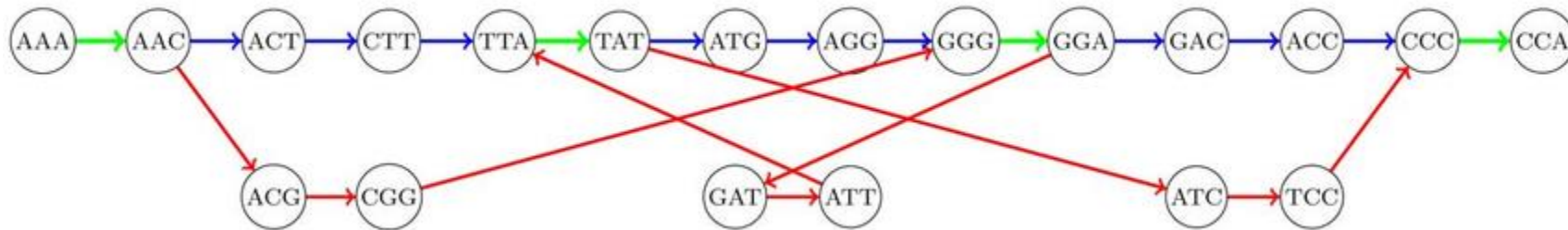
- Errors from sequencing
- Errors from RT-PCR
- Coverage variation from library prep
- Variable quantities of off-target sequence data

## **High levels of variation in the source virus sample**

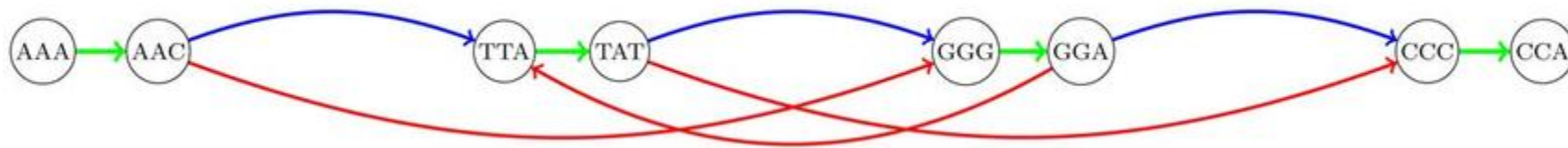
- Continuous quasispecies variation at the nucleotide level
- Length polymorphisms
- “High polyploidy”

# Viral data - challenges

OLC & De Bruijn graphs have many “bubbles” & “branches”



Whilst these can be simplified, they can be impossible to resolve



# Viral data - challenges

- Traditional assemblers often generate multiple contigs across a single region
- Virus-specific assemblers try to allow for low-level nucleotide variation

Velvet

IVA

SPAdes

VICUNA

- Some tools use the variation to reconstruct haplotypes
  - Linking reads into longer structures
  - HaploClique, Savage

# Viral data - challenges

## Ways to improve assemblies

1. Trim reads – QC
  - The termini comprise the overlaps; hence critical for good assembly
  - Often of poor quality
  - Errors may be statistically distinguishable from true variation
2. Reduce volume of duplicate information
  - Many near-identical reads can confound even the best assemblers
  - Reduce redundant reads through normalisation (see practical)
3. Eliminate off-target reads
  - Contaminating host/bacterial reads etc.
  - Can be mis-incorporated into graphs and hence contigs
  - Run a mapping against e.g. human/mouse and retain unmapped reads



# *De novo* outputs

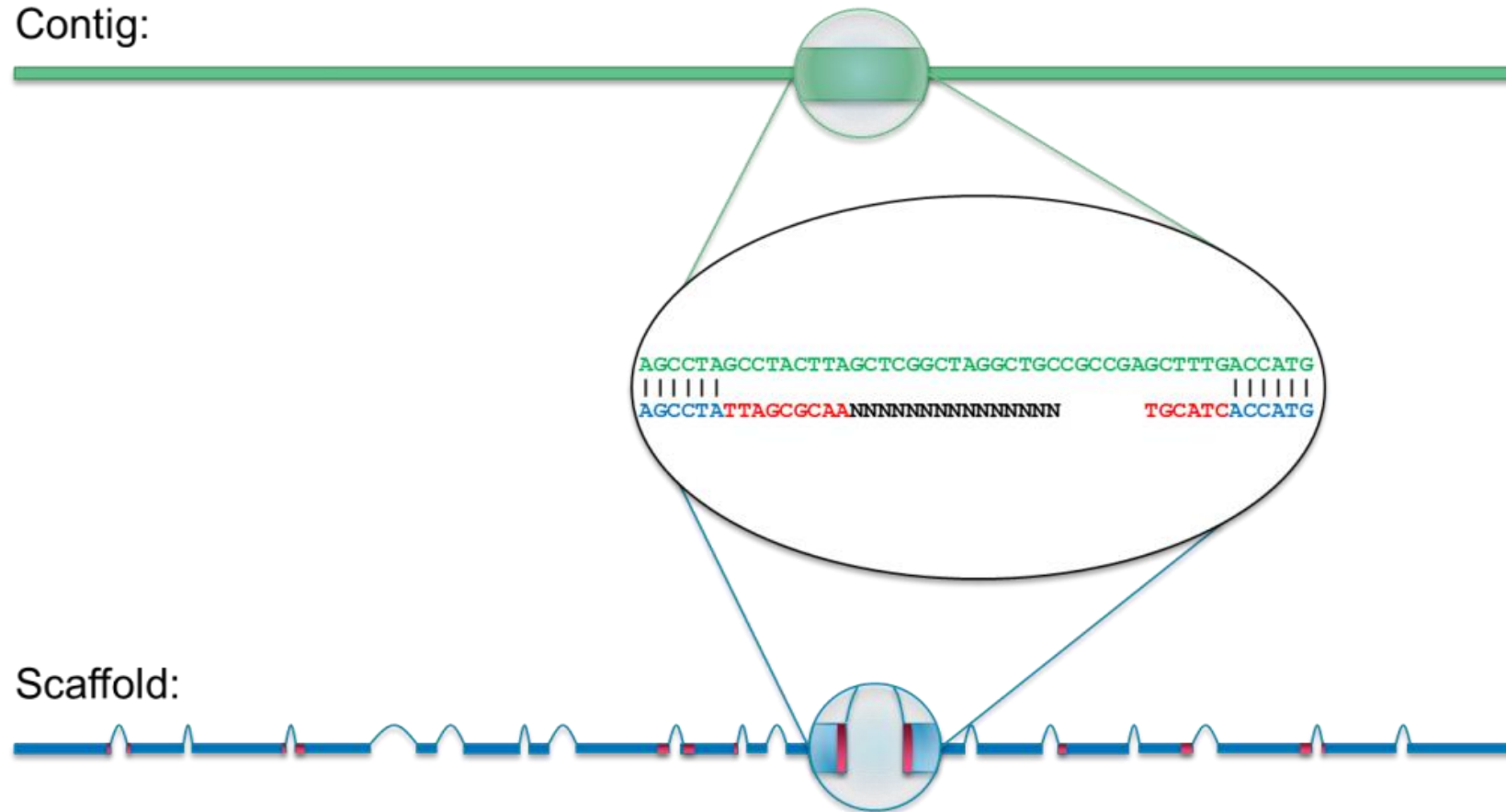
## **Contigs**

- Continuous stretches of sequence containing only A, C, G, or T without gaps.
- These can be interrogated for similarity to known targets or used as reference sequences for mapping

## **Scaffolds**

- Created by joining contigs together using additional information regarding the relative position and orientation of the contigs with reference to a genome
- The contigs within a scaffold are separated by gaps, which are denoted by a variable number of 'N' letters.
- In viral genomes, these can represent repeat regions, or unresolved areas of structural uncertainty

## De novo outputs



PacBio: [ow.ly/fAUJ304re2V](https://ow.ly/fAUJ304re2V)

# Practical

## Three parts

1. SPAdes, QUAST, & BLAST
  2. Trimmomatic
  3. Normalize-by-median
- Please `cd` to `/de_novo_assembly/files` for the practical data
  - The practical guide is in `/de_novo_assembly`