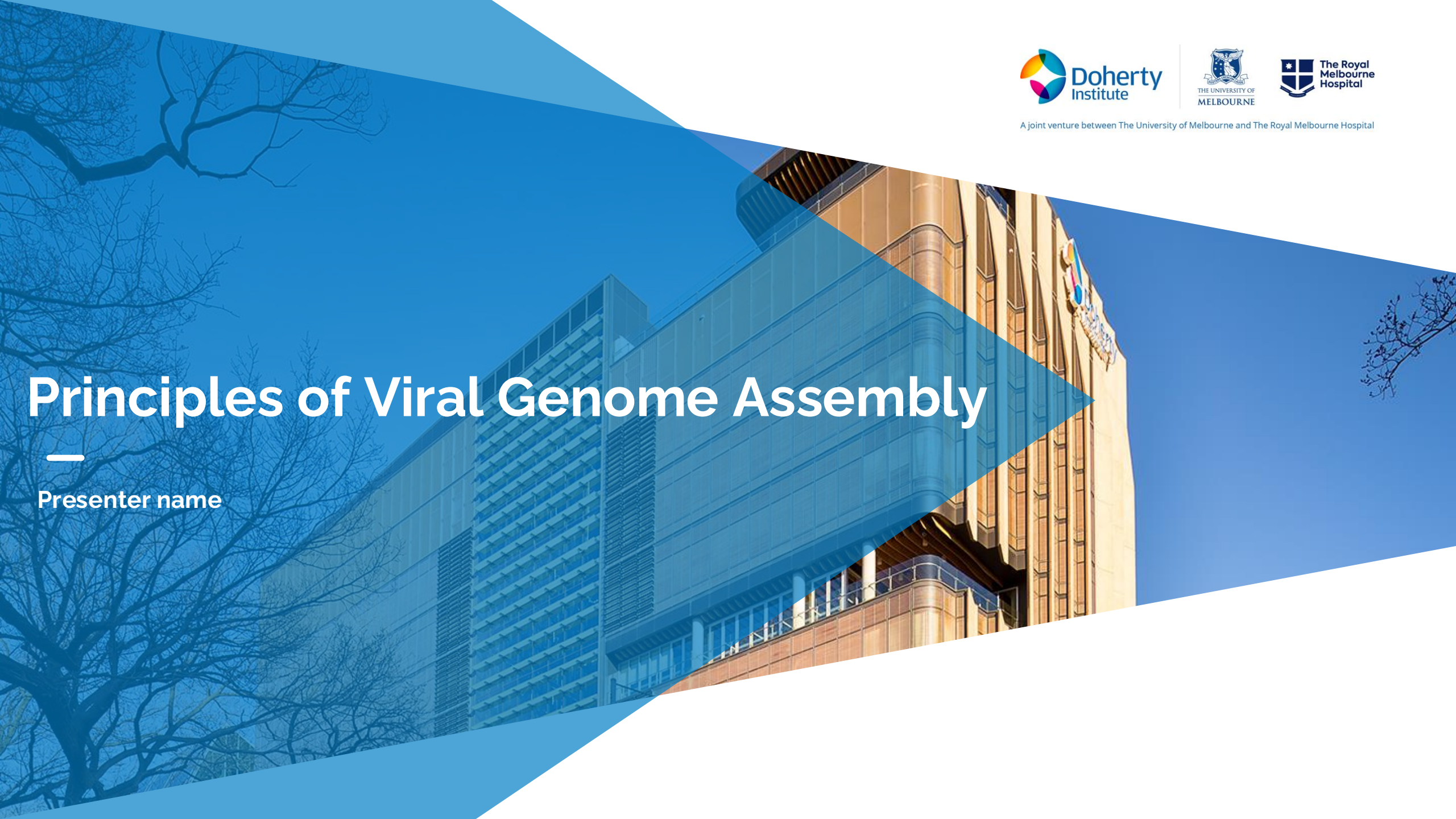


Principles of Viral Genome Assembly

Presenter name

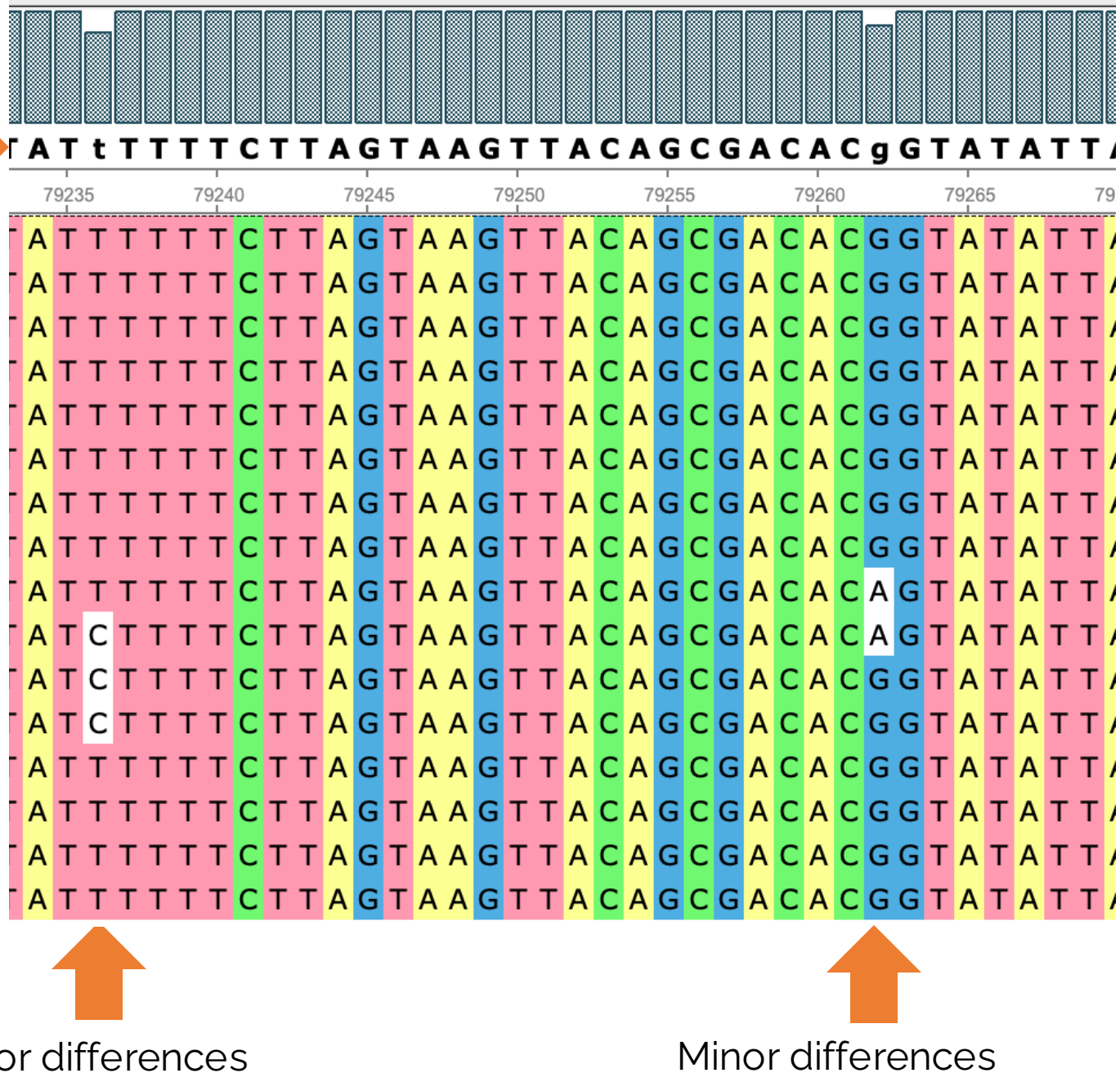
A joint venture between The University of Melbourne and The Royal Melbourne Hospital

Objectives

- Learn the two main methods for creating a virus genome
- Understand the theory of :
  - Reference based assembly
  - de novo assembly
- Understand why trimming primers is important for tiled amplicon experiments

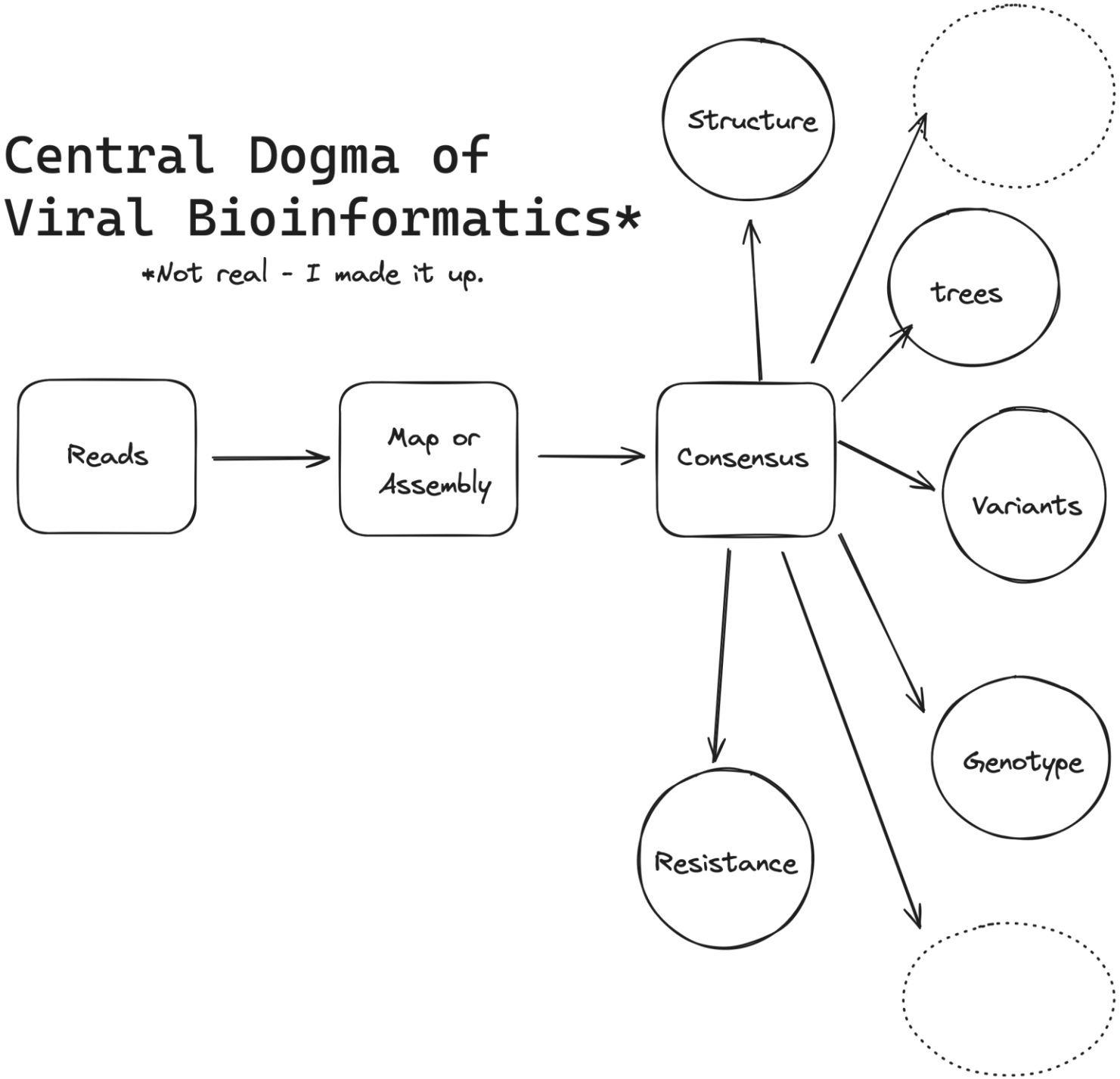
Consensus sequence

Majority nucleotide
"CONSENSUS"



- Calculated sequence of most frequent residues, either nucleotide or amino acid, found at each position in a sequence alignment.

Why create a consensus sequence?

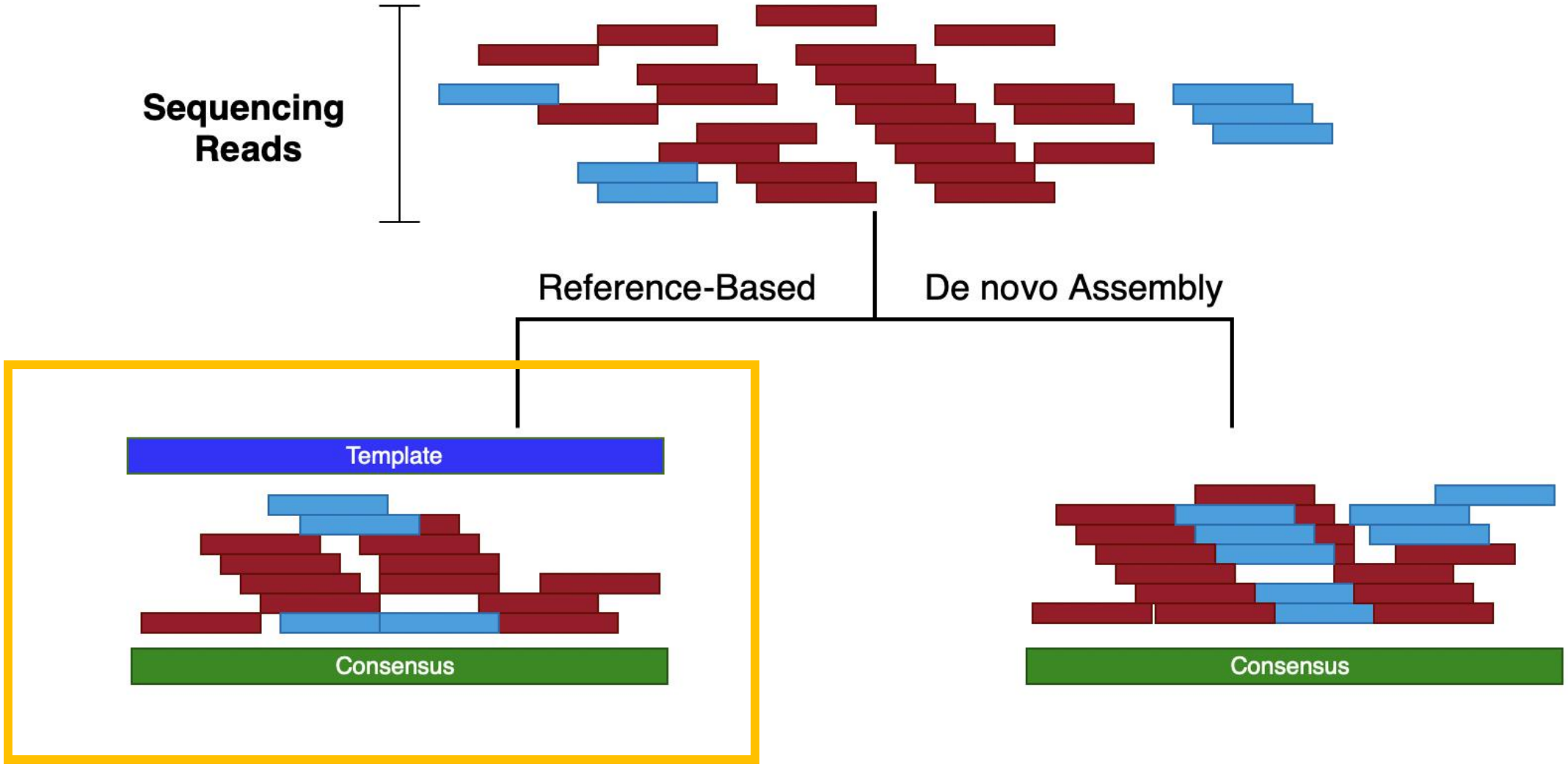


How is a consensus created?

Two techniques for creating a consensus sequence

- *de novo* (latin: of new)
- Reference based assembly

How is a consensus created?



How is a consensus created? Method 1

de novo (from new) assembly

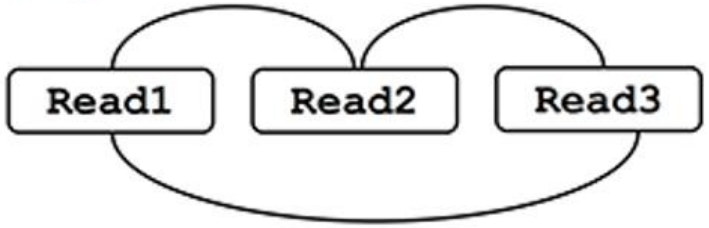
- assembles short(er) nucleotide sequences into longer ones without the use of a reference genome
- Pros: No information on sample required - good for virus discovery, complex samples, not biased by reference
- Cons: Computational and resource expensive, requires high technical expertise,

Bacterial genome assembly is usually performed by de novo method

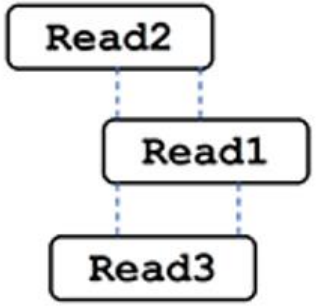
de novo assembly

(a) Overlap, Layout, Consensus assembly

(i) Find overlaps



(ii) Layout reads



(iii) Build consensus

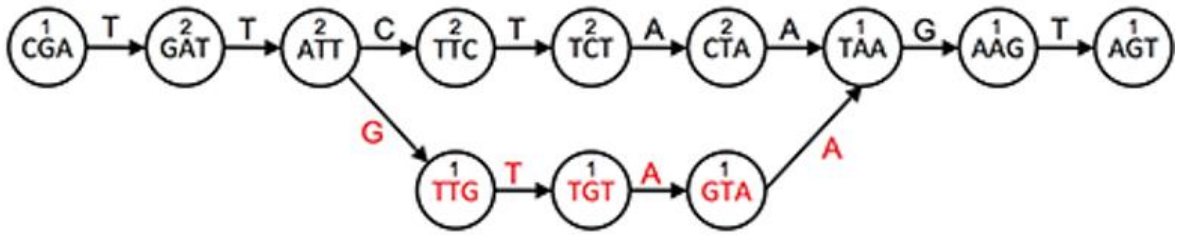
```
CGATTCTA
   TTCTAAGT
 GATTGTAA
-----------
CGATTCTAAGT
```

(b) De Bruijn graph assembly

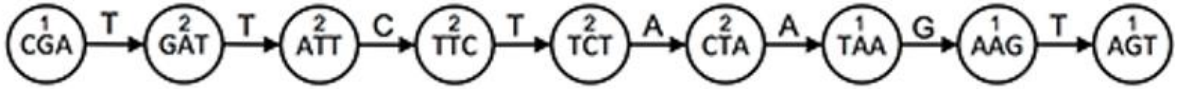
(i) Make kmers

```
Read1: TTCTAAGT     Read2: CGATTCTA     Read3: GATTGTAA
Kmers: TTC          Kmers: CGA          Kmers: GAT
        TCT                 GAT                 ATT
         CTA                 ATT                 TTG
          TAA                 TTC                 TGT
           AAG                 TCT                 GTA
            AGT                 CTA                 TAA
```

(ii) Build graph



(iii) Walk graph and output contigs



CGATTCTAAGT

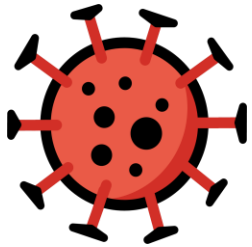
How is a consensus created? - Method 2

Reference based assembly

Using a known high-quality genome as a reference to map or align reads.

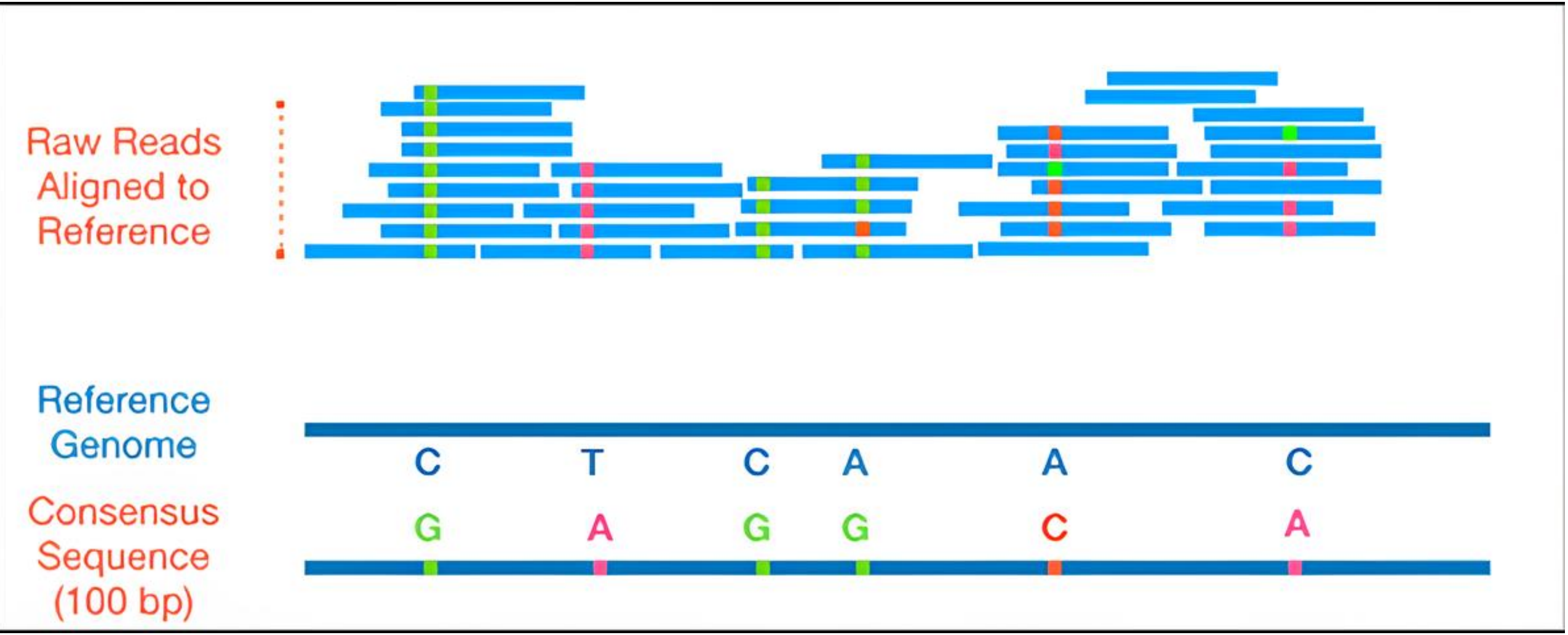
Pros: Computational efficient, simplified workflow, suited for amplicon experiments

Cons: Needs prior knowledge of sample composition, must choose appropriate reference, large structural variants will be missed



Virus genome assembly is generally performed using reference based methods.

Reference based assembly



Choosing a reference

!Extremely important! select an appropriate reference.

Choosing a reference

!Extremely important! select an appropriate reference.

- How do you select an appropriate reference?

Choosing a reference

!Extremely important! select an appropriate reference.

- How do you select an appropriate reference?
- What is the definition of an appropriate reference?

Choosing a reference

!Extremely important! select an appropriate reference.

- How do you select an appropriate reference?
- What is the definition of an appropriate reference?
- What is the effect of choosing the wrong reference?

What is an appropriate reference genome?

Simply put: A genome that is as close as possible to the specimen sequenced.

Reference genomes are (usually):

- annotated with gene/protein information
- high quality – assembled using accurate and high depth data
- verified and trusted – produced by reputable labs, from cultured material or by material providers (eg ACTC)

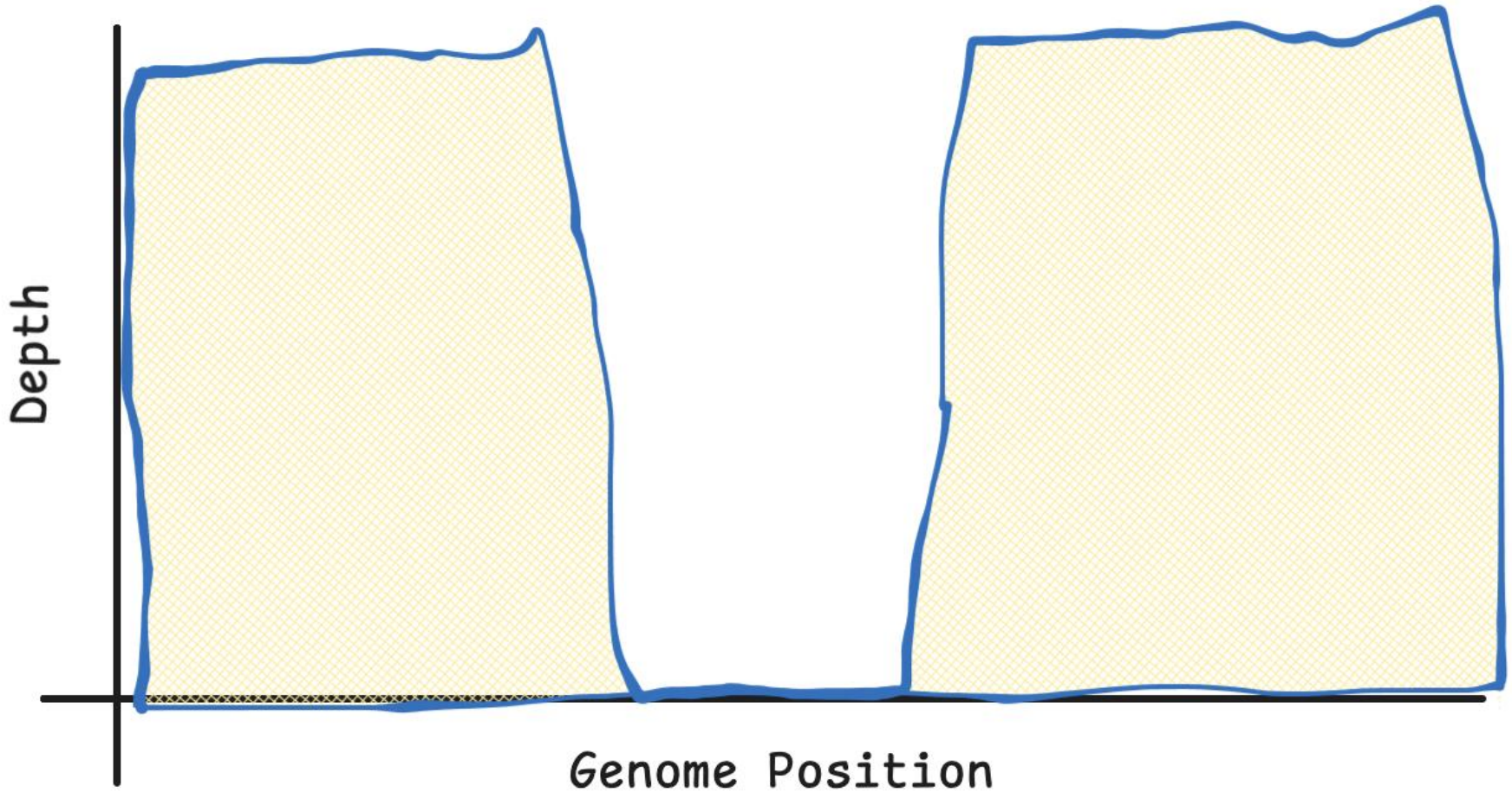
General Rules for selecting a reference genome

1. Read the literature – what are others lab using?
2. Search ICTV website ictv.global - information is at the genus level
3. Ask a domain expert – someone who has worked on the disease/virus
4. When in doubt: Choose one from Refseq* (NC_XXXXXX)

*Refseq is a curated database similar to Genbank but contains high quality genomes

The effect of a wrong genome

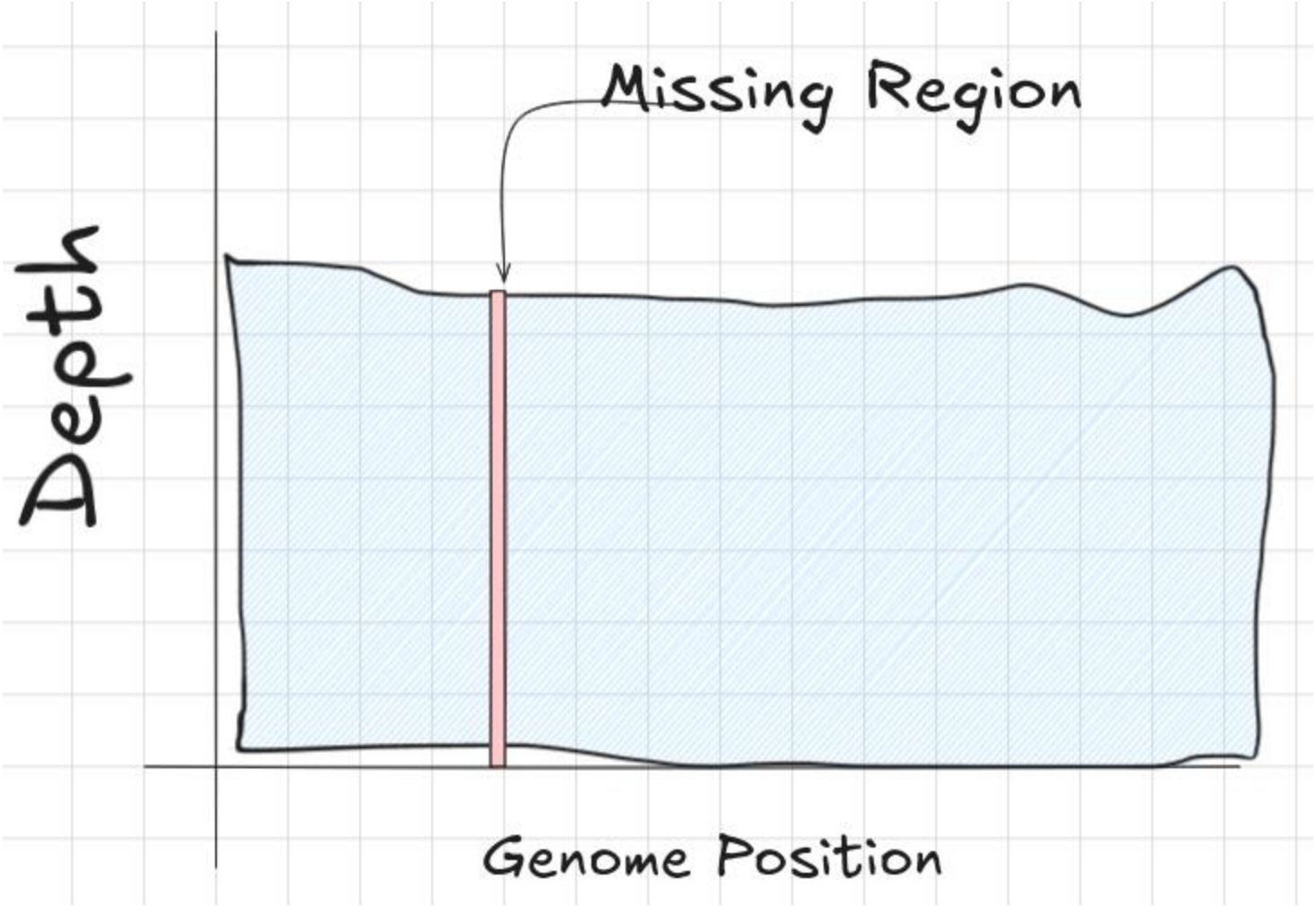
Effect: Incomplete or incorrect genome assembly



Appears to be a gap. But it is actually highly different region, reads did not map here.

The effect of a wrong genome

Effect: Appears fine but region missing



Gene/region not present in reference.

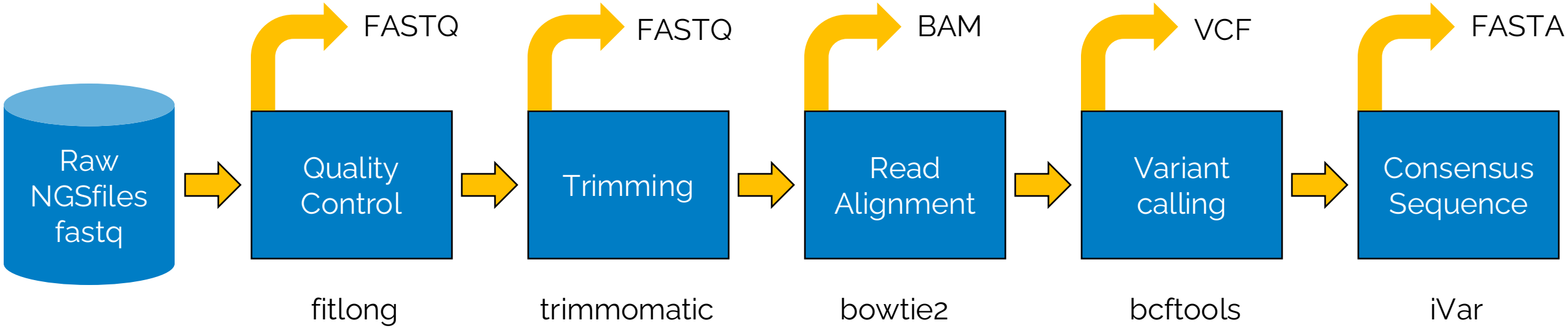
Reads either not aligned or truncated/trimmed.

Reference genomes for tiled amplicon

For tiled amplicon schemes, the primers and the reference genome are a pair. There is no need to choose a reference genome. **Use the provided primers/genome.**

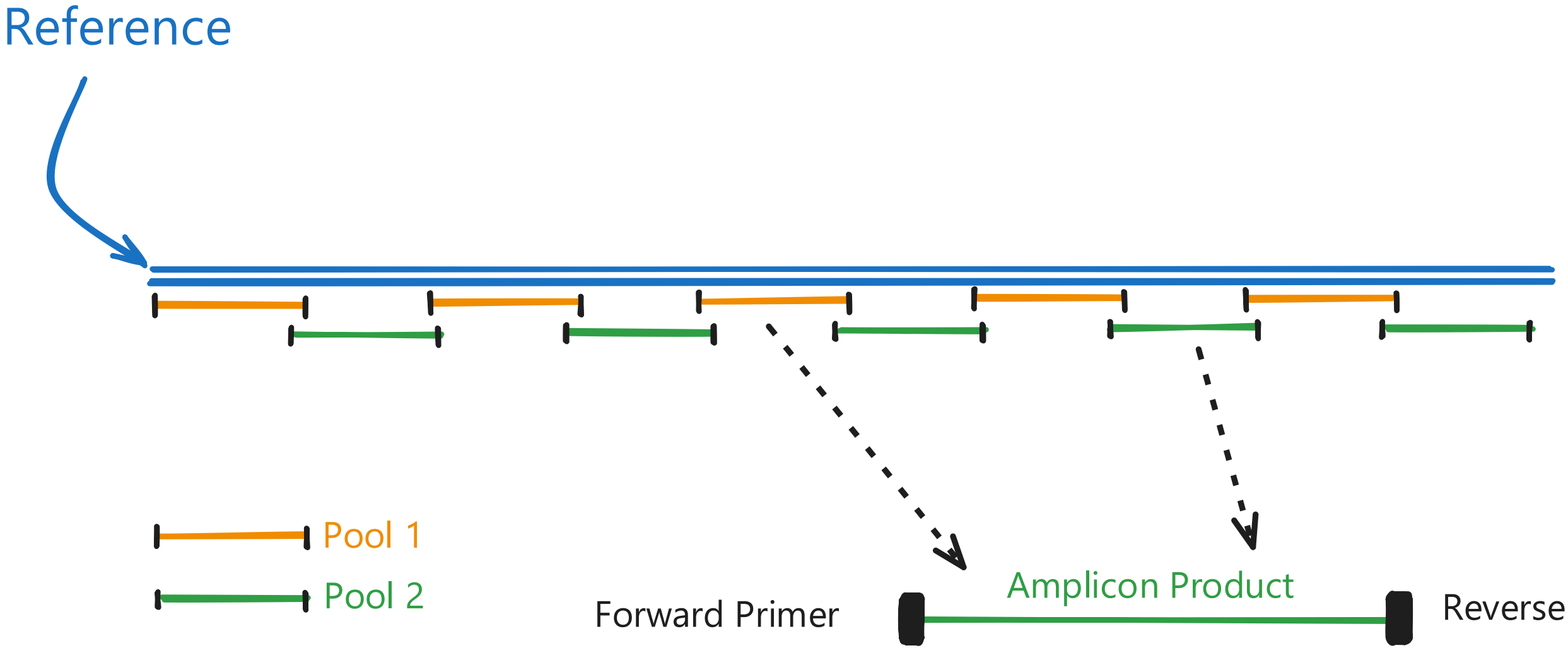
Do not change these unless your primer scheme changes

Reference based pipeline - ONT

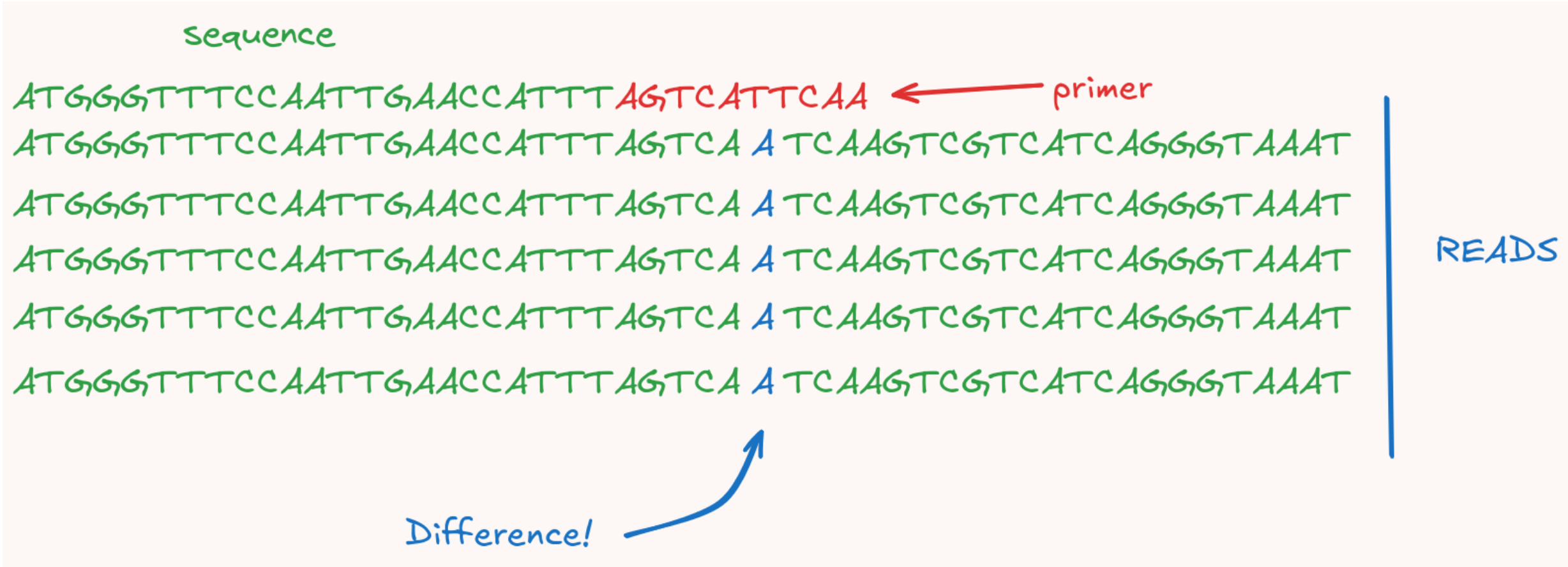


More details on MPXV pipeline in later lecture

Amplicon Sequencing – Reference based



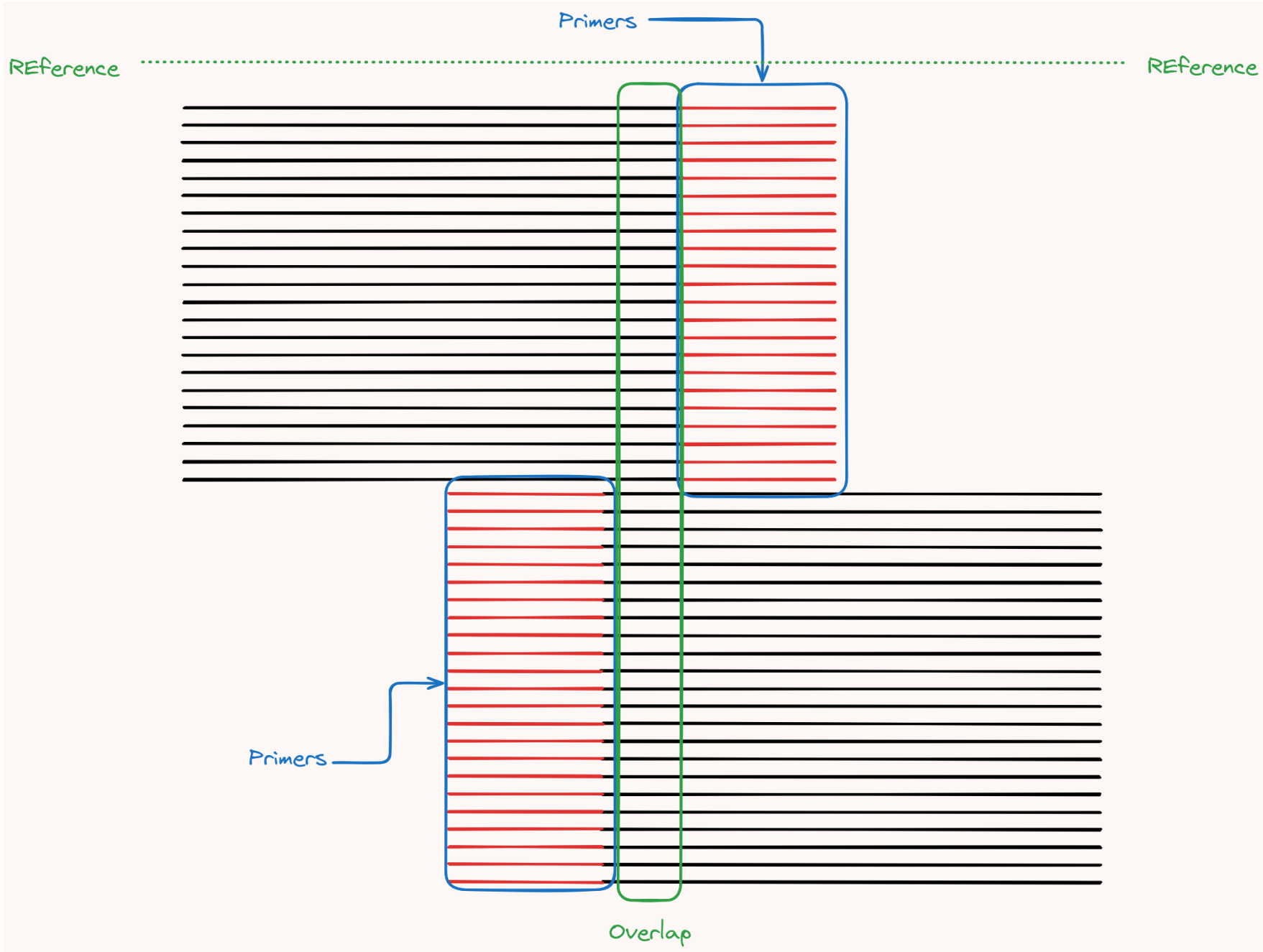
Why trim primers?



Primer sequence can **not** be incorporated into consensus sequence

Primers overlap

No information loss because of amplicon design.



How is a consensus created?

- Two methods – essentially the same:
  1. **Use a dedicated variant caller to identify mutations (SNPs, INDELs), then apply mutations to reference***
  2. Use the samtools mpileup format – take the majority call
     - iVar uses the mpileup format

Sequence	Position	Reference Base	Read Count	Read Results	Quality
seq1	272	T	24	,. $.....,,.,.,.,.,.,.,.^+.	<<<+;<<<<<<<<<<=<;<;7<&
seq1	273	T	23	,.....,.,.,.,.,.,.,..A	<<<;<<<<<<<<<3<=<<<;<<+
seq1	274	T	23	,.$.....,.,.,.,.,.,.,...	7<7;<;<<<<<<<<<=<;<;<<6
seq1	275	A	23	,$.....,.,.,.,.,.,.,.,...^l.	<+;9*<<<<<<<<<=<<:;<<<<
seq1	276	G	22	...T,,.,.,.,.,.,.,.,.,...	33;+<<7=7<<7<&<<1;<<6<
seq1	277	T	22	.....,.,.,.C.,.,.,.,.,.G.	+7<;<<<<<<<&<=<<:;<<&<
seq1	278	G	23	.....,.,.,.,.,.,.,.,.,.,...^k.	%38*<<;<7<<7<=<<<;<<<<<
seq1	279	C	23	A..T,,.,.,.,.,.,.,.,.,.,...	75&<<<<<<<<<=<<<9<<:<<<

https://en.wikipedia.org/wiki/Pileup_format

Let's play a game – align two sequences

- Go to:
  - http://teacheng.illinois.edu/SequenceAlignment/
- Finish levels 1 -3

Questions? + Resources

- GTN Course – debrujin graph (de novo) assembly
  - https://training.galaxyproject.org/training-material/topics/assembly/tutorials/debruijn-graph-assembly/tutorial.html

  GTN Course – Deeper look into assembly algorithms (advanced):
  - https://training.galaxyproject.org/training-material/topics/assembly/tutorials/algorithms-introduction/slides.html#1

  SIB Videos:
  - https://www.youtube.com/watch?v=VhJ2JKE4fBs
  - Read Mapping https://www.youtube.com/watch?v=552Rv-HrV6Q

  NCBI documentation on reference selection:
    - https://www.ncbi.nlm.nih.gov/datasets/docs/v2/policies