

Genome Assembly

Clyde Dapat



WHO Collaborating Centre
for Reference and
Research on Influenza
VIDRL



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

Objective

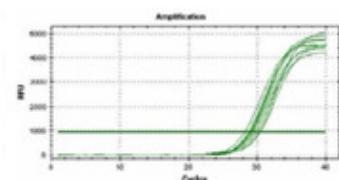
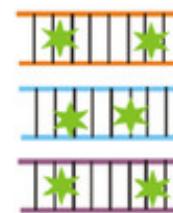
- To learn the basic principle of genome assembly

Viral sequencing review

1. Clinical sample

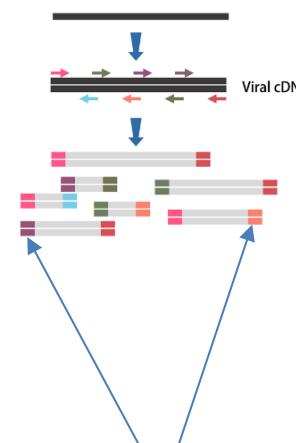


2. Diagnostic PCR



3. Amplify viral RNA

Tiled amplicon PCR



PCR
primers

4. Sequencing

Short-read Illumina sequencing

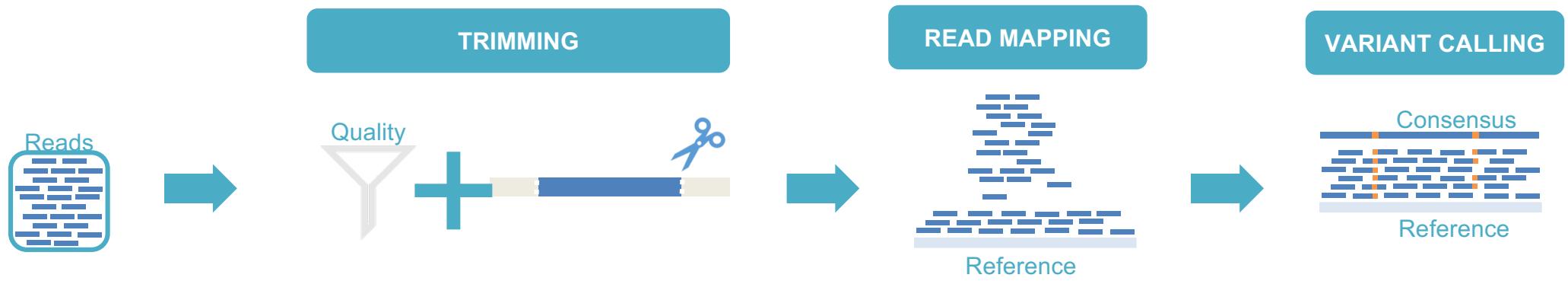


OR

Long-read Nanopore sequencing



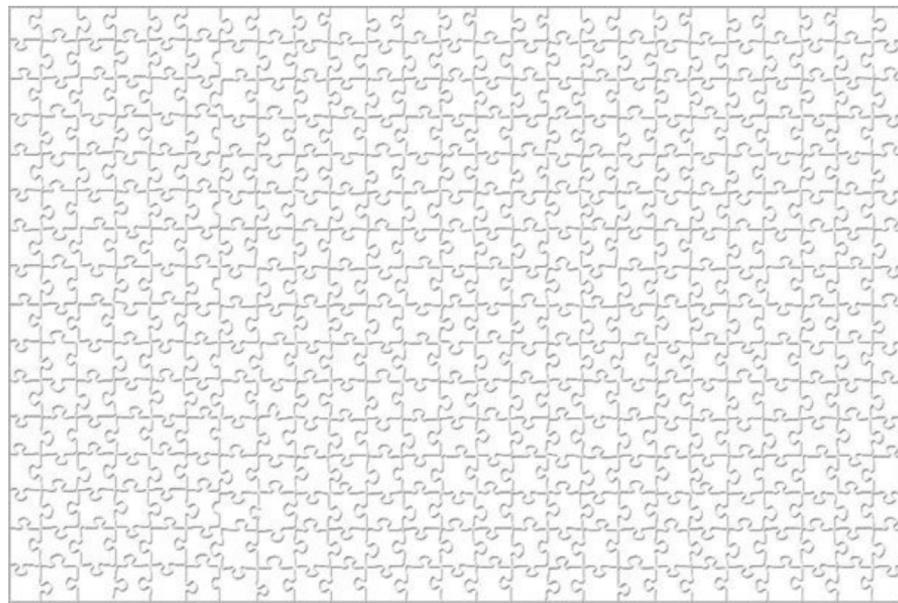
Processing amplicon sequencing data



Types of assembly



REFERENCE-BASED



DE NOVO

De novo assembly

- Latin for “from the new”
- Reference sequence is not needed
- Results in contiguous sequences (contigs)
- Applicable for metagenomics and assembling genomes that have no reference

AHICS IS
NFORMATIC
S IS COOL!
BIOINFO

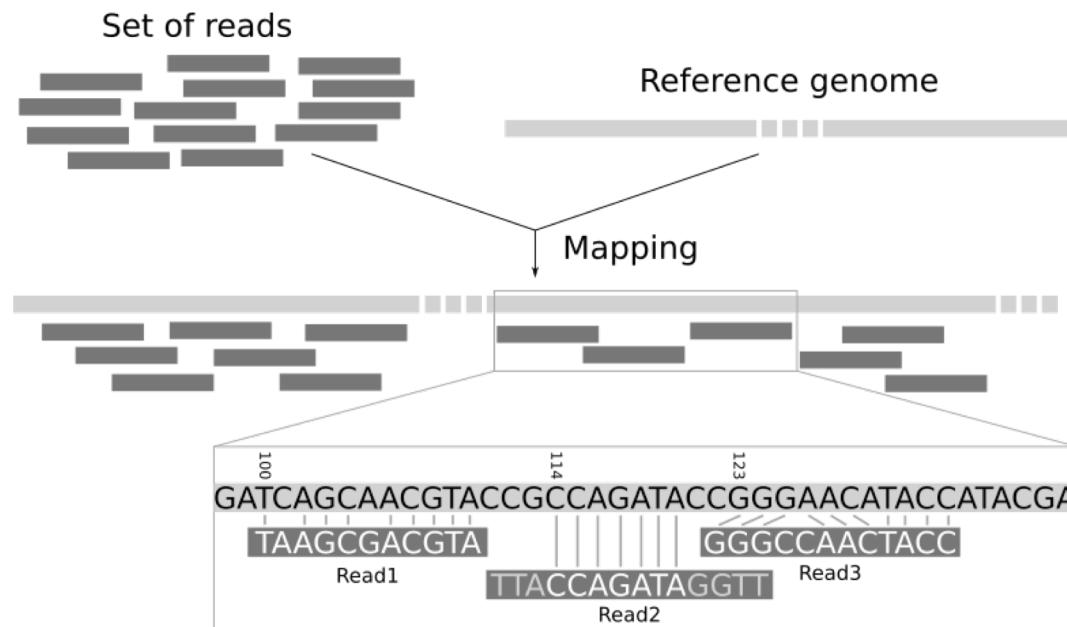
BIOINFO
NFORMATIC
AHICS IS
S IS COOL!

BIOINFORMA
T
ICS IS COOL!
H

BIOINFORMATICS IS COOL!

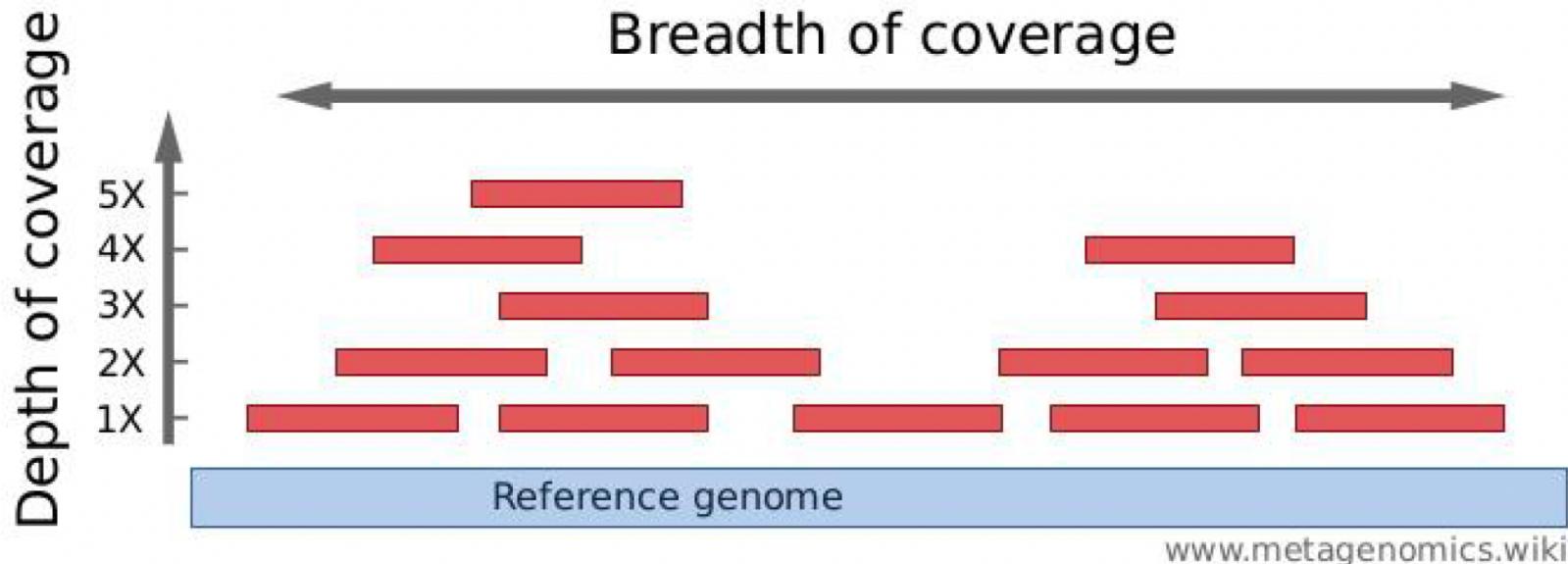
Read mapping

- Read mapping is the process of aligning reads to a reference genome



<https://training.galaxyproject.org/training-material/topics/sequence-analysis/images/mapping/mapping.png>

Depth vs breadth of coverage



- Depth of coverage – number of times a specific position in a genome is sequenced
- e.g. 50x depth of coverage means that each base in the genome has been sequenced 50 times on average
- Breadth of coverage – proportion of the genome that has been sequenced
- e.g. 95% coverage means 95% of the length of the genome has been sequenced

Read mapping tools

Minimizers – collects unique kmers that are used as seeds to search for longer matches.
eg: minimap2

Position	1	2	3	4	5	6	7	1	2	3	4	5	6	7	8	9	10	11	12
Sequence	2	3	1	0	3	4	3	4	2	6	4	7	2	8	1	4	7	5	1
<i>k</i> -mers with minimizer in bold	2	3	1					4	2	6	4	7	2	8					
		3	1	0					2	6	4	7	2	8	1				
			1	0	3					6	4	7	2	8	1	4			
				0	3	4					4	7	2	8	1	4	7		
						3	4	3					7	2	8	1	4	7	5
	(a)							(b)						2	8	1	4	7	5

Read mapping tools

Burrows-Wheeler Transformation (BWT)

- maps genomic sequences using a data compression technique that rearranges characters in a string to group identical ones together, facilitating efficient indexing and alignment

e.g. bwa or Bowtie2

Burrows-Wheeler Transform

Text transform that is useful for compression & search.

banana

banana\$

anana\$b

nana\$ba

ana\$ban

na\$bana

a\$banan

\$banana

sort →

\$banana

a\$bana

ana\$ba

anana\$b

banana\$

nana\$ba

na\$bana

BWT(banana) =
annb\$aa

Tends to put runs of the same character together.

Makes compression work well.

“bzip” is based on this.

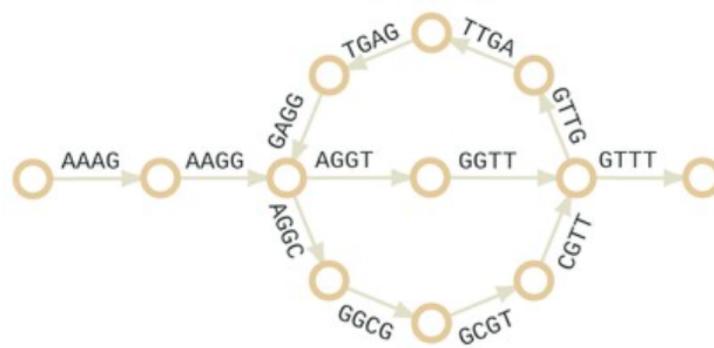
Read mapping tools

A. Short read to k -mers ($k=4$)

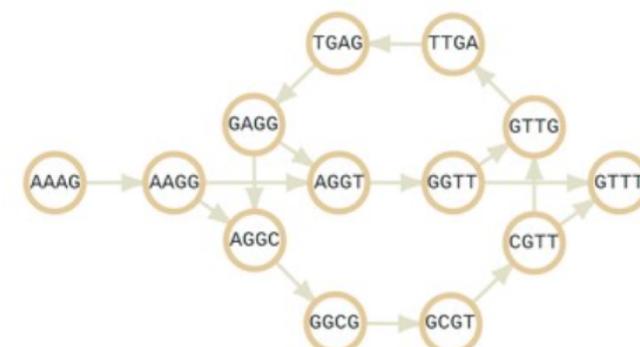
AAAGGCGTTGAGGTT

AAAG
AAGG
AGGC
GGC
GCGT
CGTT
GTTG
TTGA
TGAG
GAGG
AGGT
GGTT

B. Eulerian de Bruijn graph



C. Hamiltonian de Bruijn graph



Sequence Alignment/Map (SAM)

- SAM is a file format to store information about the alignment of reads to a sequencing genome
- Human-readable text files consist of tab-separated columns, with each line representing a single read alignment and each column providing specific information of the alignment

SAM file format

©HD VN:1.5 SO:coordinate											Header section
©SQ SN:ref LN:45											Alignment section
r001 99 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGGATACTG *											
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *											
r003 0 ref 9 30 5S6M * 0 0 GCCTAAGCTAA * SA:Z:ref,29,-,6H5M,17,0;											
r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC *											
r003 2064 ref 29 17 6H5M * 0 0 TAGGC * SA:Z:ref,9,+,5S6M,30,1;											
r001 147 ref 37 30 9M = 7 -39 CAGCGGCAT * NM:i:1											

Optional fields in the format of TAG:TYPE:VALUE

QUAL: read quality; * meaning such information is not available

SEQ: read sequence

TLEN: the number of bases covered by the reads from the same fragment. Plus/minus means the current read is the leftmost/rightmost read. E.g. compare first and last lines.

PNEXT: Position of the primary alignment of the NEXT read in the template. Set as 0 when the information is unavailable. It corresponds to POS column.

RNEXT: reference sequence name of the primary alignment of the NEXT read. For paired-end sequencing, NEXT read is the paired read, corresponding to the RNAME column.

CIGAR: summary of alignment, e.g. insertion, deletion

MAPQ: mapping quality

POS: 1-based position

RNAME: reference sequence name, e.g. chromosome/transcript id

FLAG: indicates alignment information about the read, e.g. paired, aligned, etc.

QNAME: query template name, aka. read ID

<https://www.samformat.info/sam-format-flag>

Binary Alignment/Map (BAM)

- same as SAM but encoded in binary (e.g. 10110)
- more compact and efficient in terms of storage and processing

Thank you !

clyde.dapat@influenzacentre.org



WHO Collaborating Centre
for Reference and
Research on Influenza
VIDRL



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

