NGS Genome Assembly Workflow

Quality control to consensus

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

History of NGS

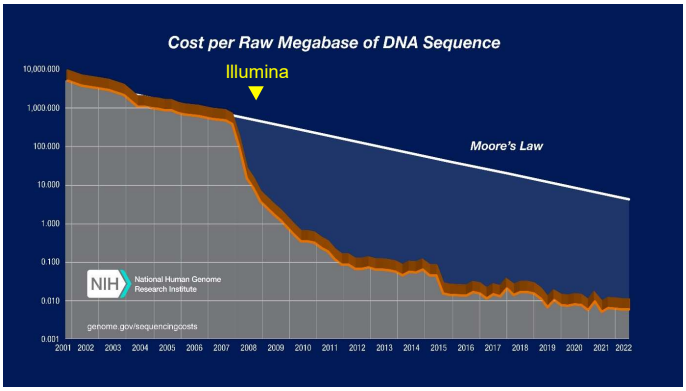
LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

1st gen
- Sanger sequencing.
- Based on dNTPs.
- 300 to 750 bp.

Next Generation Sequencing

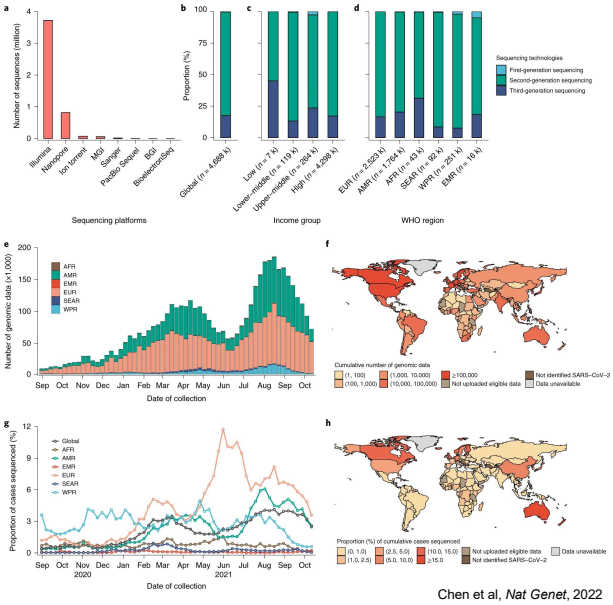
2nd gen
- I.e. Illumina, 454. Ion Torrent.
- "Sequencing by synthesis"
- Shorter read length (75-300 bp)
- High throughput.

3rd gen
- ONT, pacbio.
- Long read
- High error rate (~5%).



Current Situation

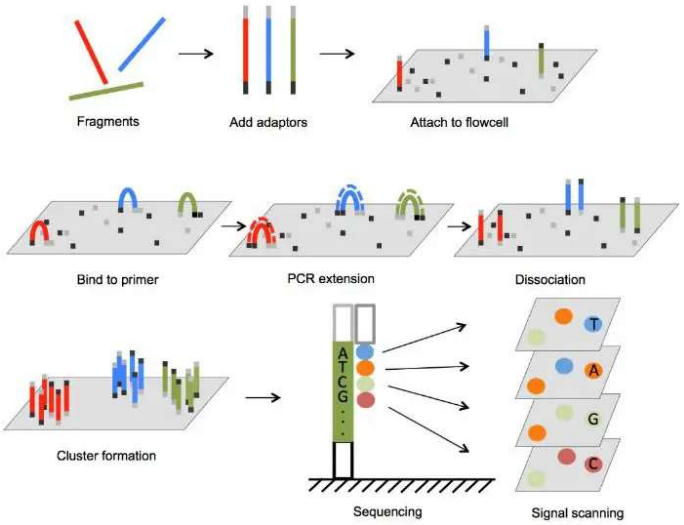
- This data is specific for SARS-CoV-2
- Illumina is still widely used, followed by ONT/Pacbio.
- MGI sequencing – new player?
- Discrepancy between European/American and Asia/Africa/Pacific.



Chen et al, *Nat Genet*, 2022

How does it work? – Short Read

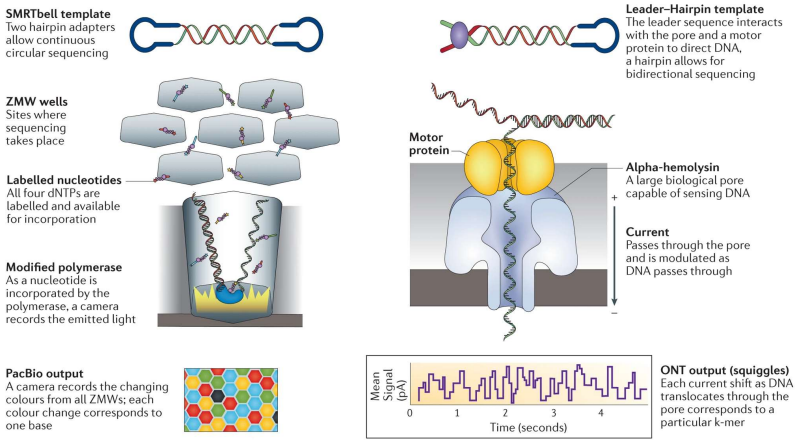
- Most popular – Illumina sequencing.
- High throughput and high accuracy.
- Paired end – sequence the insert from both sides.
- Short read – from 75 bp to 300 bp.
- Real time analysis is not possible.



Kulsi et al, Next Generation Sequencing: Advances, Applications and Challenges, 2016

How does it work? – Long read

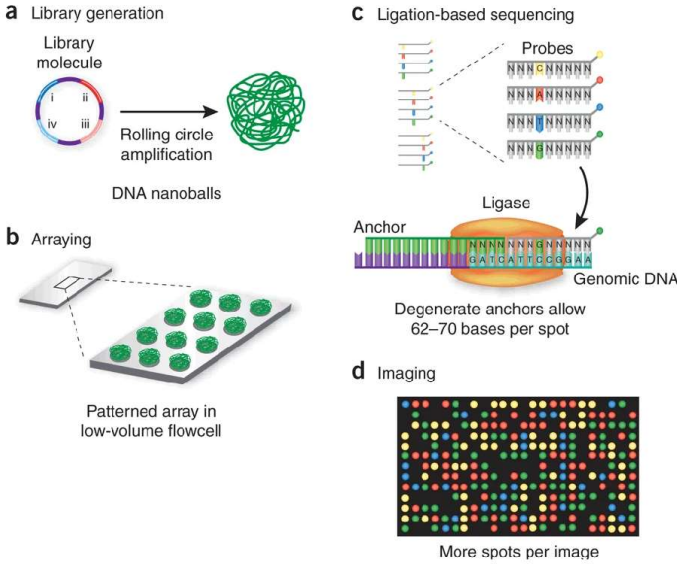
- Pacific BioSciences.
  - —HiFi sequencing.
- Read length can be up to 10-15 kb.
- Oxford Nanopore Technologies.
  - —Use nanopores.
  - —Ultra long read.
  - —Recently introduced **duplex reads**.
  - —Real time data analysis.
- Instrument size: PacBio >>> ONT



McCombie et al, *Nat Rev Genet*, 2016

Nanoball sequencing

- A trending approach.
  - —Lower cost;
  - —High throughput;
  - —High accuracy (1:100,000 error rate).
- Library preparation – forming a nanoballs.
- Sequencing is based on combinatorial probe–anchor synthesis (cPAS).
  - —Insert length 100-170.
  - —Possible for paired end.



Porreca GJ, *Nature Biotechnology*, 2010

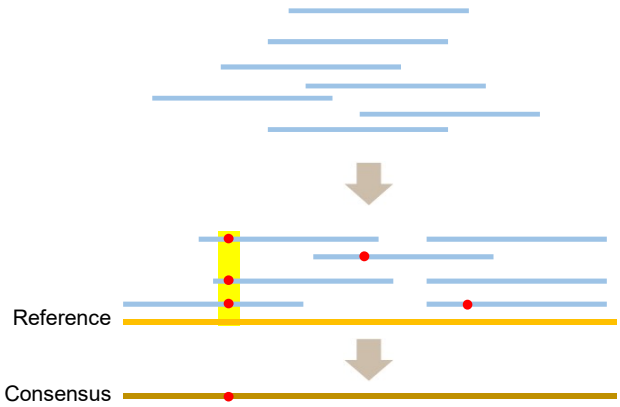
Does size matters?

Source: https://www.illumina.com/systems/sequencing-platforms/iseq.html, https://www.pacb.com/blog/introducing-the-sequel-system-the-scalable-platform-for-smrt-sequencing/, https://nanoporetech.com/resource-centre/establishment-and-cryptic-transmission-zika-virus-brazil-and-americas

Genome assembly

Reference assembly

- Required a reference in which NGS reads can be mapped into.
- The reference must be similar to the organism from which the NGS data is generated.
  —NCBI Refseq is a good start.
- Once reads are mapped, variants can be identified.
- This variants will then be applied to the reference, producing a consensus.



De novo assembly

- Trying to complete a jigsaw puzzle that you bought secondhand.
  —Missing pieces – some parts cannot be sequenced.
  —Pieces from other puzzles – contaminants.
  —Missing reference picture – no reference to align.

Assembly – De novo assembly

Reads:

OUER THE LAZY DOG
RHWN FOX JUM
THE QUICK BRO
CK BROWN FOX
JUMPS OVER
HE LAZY DOG
MPS OVER THE

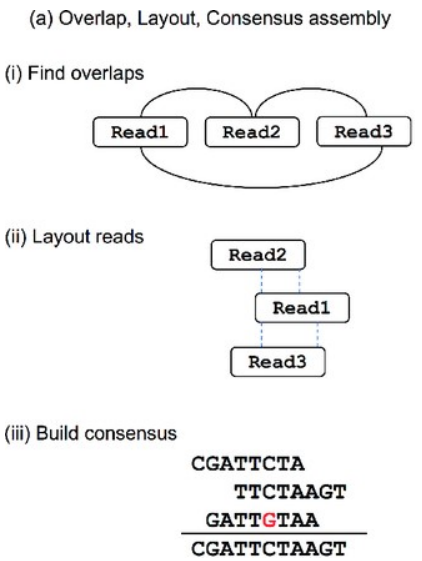
THE QUICK BROWN FOX JUMPS OVER THE LAZY DOG

THE QUICK BRO
CK BROWN FOX
RHWN FOX JUM
JUMPS OVER
MPS OUER THE
OUER THE LAZY DOG
HE LAZY DOG

THE QUICK BROWN FOX JUMPS OUER THE LAZY DOG

De novo assembly

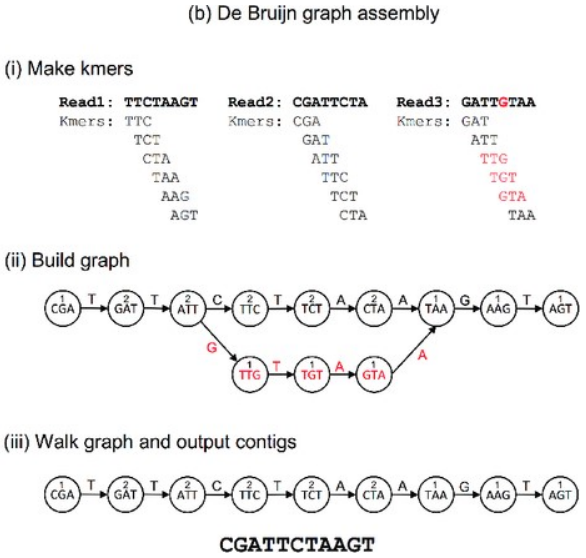
- Two main approach: **overlap and graph-based assembly**.
- Overlap, layout, consensus method:
  - —Finding overlap among reads in a pairwise manner.
  - —The overlapped reads then laid into contigs.
  - —Calling consensus.
  - —NGS data has million of reads – pairwise comparison is computationally expensive!



Ayling et al., *Briefings in Bioinformatics*, 2019

De novo assembly

- Graph-based assembly:
  - —Reads are divide into kmers (a sequence length of k).
  - —Find overlap with a length of k-1 between kmers.
  - —Build the graph.
  - —Find best path – remove branches.



Ayling et al., *Briefings in Bioinformatics*, 2019

De novo assembly

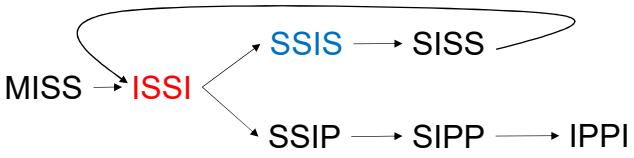
Challenges in repeats

MISSISSIPPI

Read 1: MISSIS
MISS
ISSI
SSIS

Read 2: SSISSI
SSIS
SISS
ISSI

Read 3: SSIPPI
SSIP
SIPP
IPPI



MISSISSIPPI?
MISSISSISSISSIPPI?
…

"It is impossible to resolve a repeat of a length of N without having reads longer than N."

De novo assembly

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

Challenges

- Computational resources – analyzing million reads.
- Highly similar sequences – repeats.
- Ploidy or quasi-species.
- Long read vs short read:
  - Hybrid assembly is possible.

Common

General flowchart or reference assembly

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

Basecalling
- Signals to ATCG

Preprocessing
- Quality check and filtering

Assembly
- Making database
- Getting references
- Aligning reads
- SAM to BAM
- Visualization
- Calculating coverage depth and breadth.

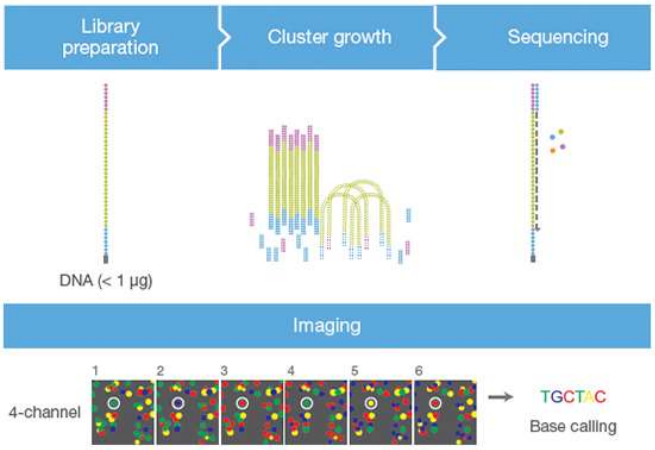
Consensus
- Collect information.
- Call variants and filtering.

Polishing
- Correction

Common

Basecalling – signal to ATCG

- Conversion of certain signal into nucleotide sequences → Basecalling.
  - —Sanger: chromatogram peak.
  - —Illumina: images.
  - —Nanopore: electrical current – squiggles.

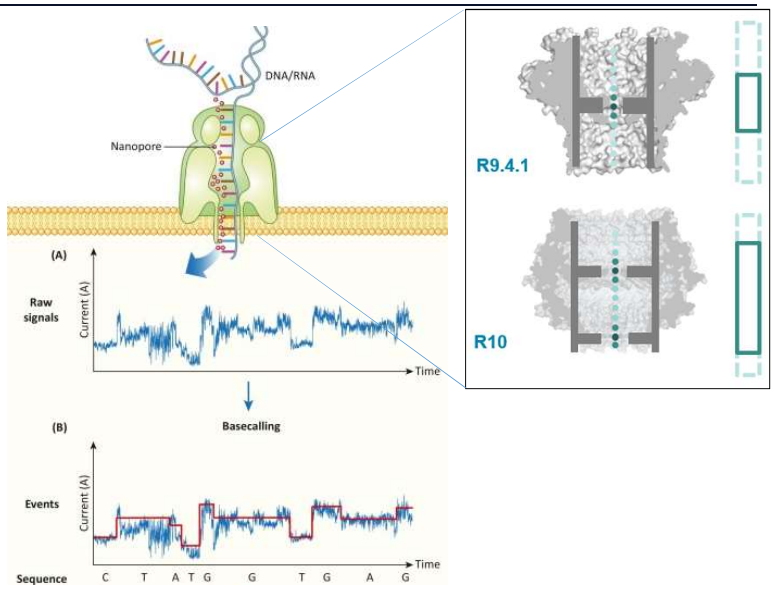


https://www.illumina.com/science/technology/next-generation-sequencing/sequencing-technology/2-channel-sbs.html

Long read

Basecalling – signal to ATCG

- **Fast5** – squiggle data.
- Available tools: Guppy, Bonito, **Dorado**.
- Uses neural network to translate the squiggle.
  - —Model of the basecaller is an important factor.
- Update: Fast5 file type is replaced by **Pod5**.



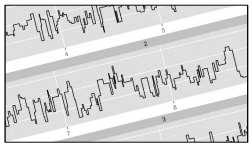
Wan KY *et al.*, Trends in Genetic, 2021.

Common

Pre-processing – Overview of fastq file

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

Signal



Sequence – in FASTQ format.

```
@e15b15ab-a0a0-4bfb-a0ca-25f8db169a5e runid=7
ACTGGAGCTAGGATGAGTTCCAATGGTCCTCATCGCCTGCACCA
+
-----..(*-./;=<9442>{73.-+*-<337::<44488<==?32
```

Read ID
Sequence
Separator
Quality

- A text-based file format to store sequence.

Phred quality scores

ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 $	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

Quality Score	Probability of Incorrect Base Call	Inferred Base Call Accuracy
10 (Q10)	1 in 10	90%
20 (Q20)	1 in 100	99%
30 (Q30)	1 in 1000	99.90%

Common

Pre-processing – Overview of fastq file

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

Which read has a higher quality?

```
@e15b15ab-a0a0-4bfb-a0ca-25f8db169a5e runid=78212cb5574a3734e3db6194
ACTGGAGCTAGGATGAGTTCCAATGGTCCTCATCGCCTGCACCATCTGCCTAGCCTGACAACGCCTCC
+
-----..(*-./;=<9442>{73.-+*-<337::<44488<==?323>=:45144*((%##$&##&&20
```

```
@A01619:134:HWMLMDSX3:2:1101:13476:1000 1:N:0:NCGTAACT+GGTTGAAC
CGCAGGCTCCACTCCTGGTGGTGCCCTTCCGTCAATTCCTTTAAGTTTCAGCTTTGCAACCATACTCC
+
::FFFFFFFFFFFFFFFFFF::F:FFFFFFFFF:F,FFFFFFFF:FFFF,FFFFF:FFFFFF:FFFFF
```

ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

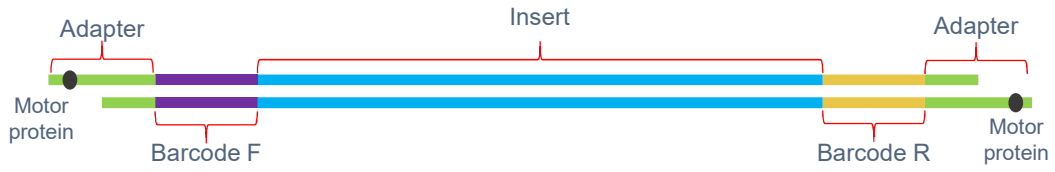
Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 $	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

Common

Pre-processing – anatomy of an NGS library

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

Typical Nanopore library



Typical Illumina library

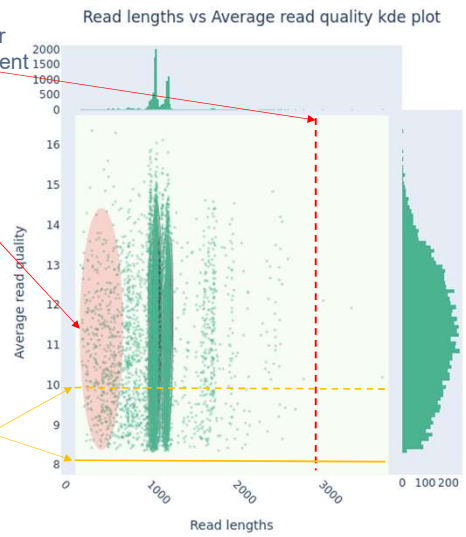
Long read

Pre-processing – Quality check and filtering

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

- For long read quality check, **NanoPlot** (https://github.com/wdecoster/NanoPlot) can be used.
- Adapter trimming and low quality filter can be applied during basecalling with Guppy.
- Additionally, several tools can be used to filter the reads:
  - —Porechop (https://github.com/rrwick/Porechop)
  - —Fastp (https://github.com/OpenGene/fastp)

- Some reads are longer than the longest segment (>3 Kb).
- Many reads are shorter than the shortest gene (<500 bp).
- Threshold for quality seemed to be set at 8. Let's increase to 10.



Long read

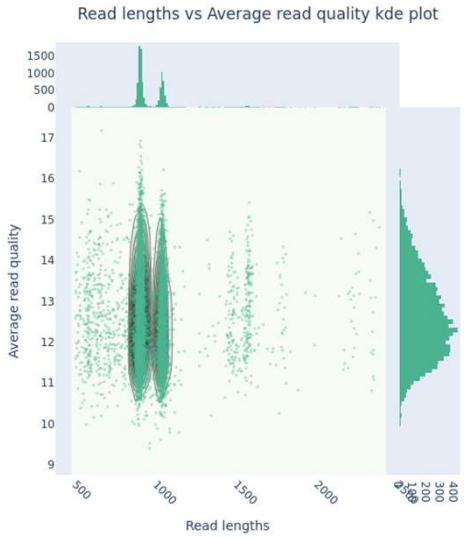
Pre-processing – Quality check and filtering

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

- After trimming

Summary statistics

General summary	
Mean read length	960.2
Mean read quality	12.0
Median read length	908.0
Median read quality	12.5
Number of reads	20,938.0
Read length N50	913.0
STDEV read length	167.2
Total bases	20,105,208.0
Number, percentage and megabases of reads above quality cutoffs	
>Q5	20938 (100.0%) 20.1Mb
>Q7	20938 (100.0%) 20.1Mb
>Q10	20920 (99.9%) 20.1Mb
>Q12	14474 (69.1%) 13.9Mb
>Q15	449 (2.1%) 0.4Mb
Top 5 highest mean basecall quality scores and their read lengths	
1	17.2 (659)
2	17.0 (906)
3	16.9 (903)
4	16.9 (664)
5	16.8 (897)
Top 5 longest reads and their mean basecall quality score	
1	2785 (11.0)
2	2396 (13.2)
3	2385 (14.8)
4	2365 (14.5)
5	2365 (14.3)



Short read

Pre-processing – Quality check and filtering

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

- Quality check for short-read usually is conducted using FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/).
- Produces summary of the reads.

Summary

- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

Basic Statistics

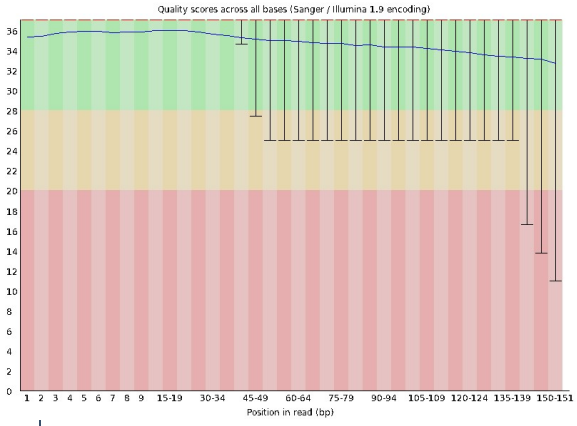
Measure	Value
Filename	NG-31476_1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	11289361
Total Bases	1.7 Gbp
Sequences flagged as poor quality	0
Sequence length	151
%GC	56

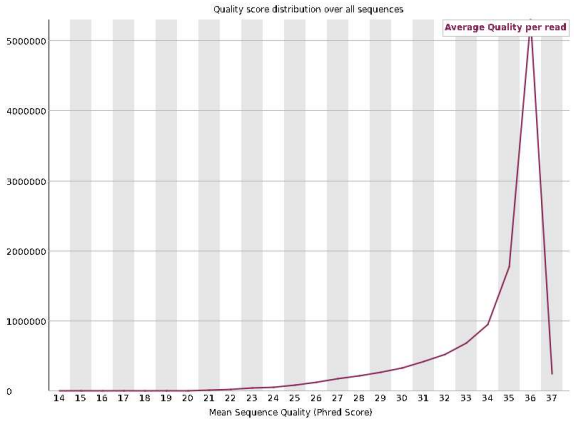
Short read

Pre-processing – Quality check and filtering

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

- Two important graph is the per base quality and per sequence quality.
- Ideally quality should be above 20 (~99% error rate).
- Quality of the reads reduced towards the 3' end.
- Trim the low quality bases with Trimmomatic.





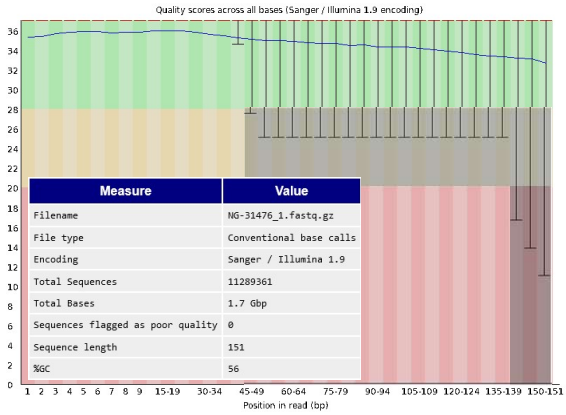
Short read

Pre-processing – Quality check and filtering

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

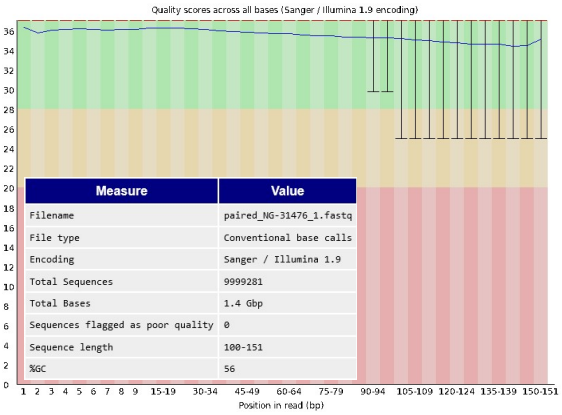
- Quality filtering with Trimmomatic.
- Good practice – perform quality check before and after filtering.

Before



Measure	Value
Filename	NG-31476_1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	11289361
Total Bases	1.7 Gbp
Sequences flagged as poor quality	0
Sequence length	151
%GC	56

After



Measure	Value
Filename	paired_NG-31476_1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	9999281
Total Bases	1.4 Gbp
Sequences flagged as poor quality	0
Sequence length	100-151
%GC	56

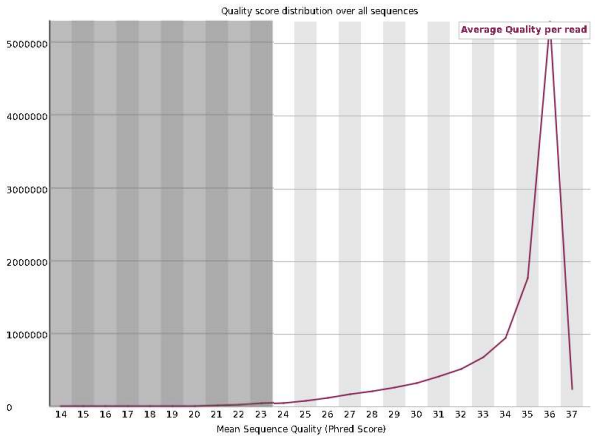
Short read

Pre-processing – Quality check and filtering

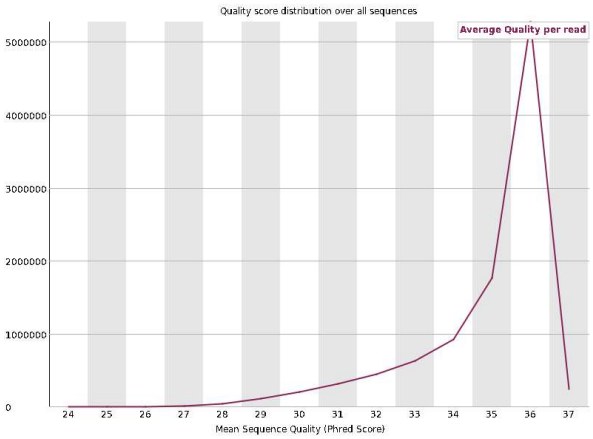
LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

- Quality filtering with Trimmomatic – filter by read length and quality, trim low quality bases.
- Good practice – perform quality check before and after filtering.

Before



After



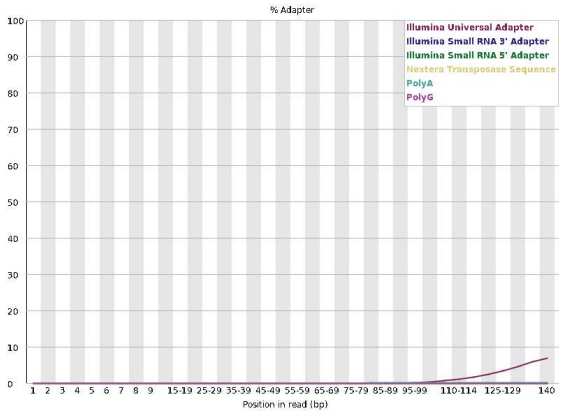
Short read

Pre-processing – Quality check and filtering

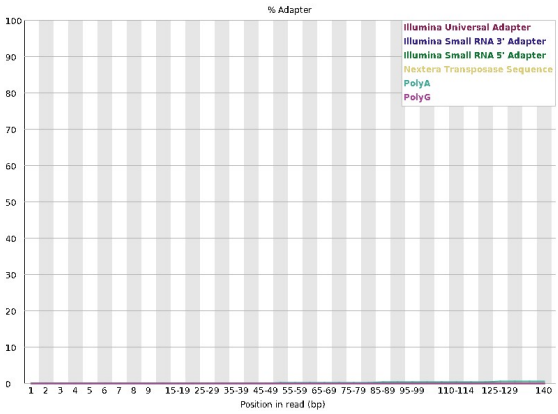
LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

- In case of adapter contamination, the adapter sequences can be removed using CutAdapt (https://cutadapt.readthedocs.io/en/stable/) or trimmomatic.
  —Information on kits and library preparation are important.

Before



After



Common

Sequence complexity

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

- The complexity is defined as the percentage of base that is different from its next base (base[i] != base[i+1]).
- Example:
  - —A 51-bp sequence, with 3 bases that is different from its next base.
  - —Seq: 'AAA**AT**TTTTTTTTTTTTTTTTTTTT**TG**GGGGGGGGGGGGGGGGGGGGG**GC**CCC'
  - —complexity = 3/(51-1) = 6%
- Low complexity sequences are less useful for taxonomic analysis (i.e. Kraken) and can be a problem for assembler

Common

Assembly – finding references

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

- Reference based assembly highly depends on the reference.

"Good reference = good consensus"

- Generally it is advised to use sequences listed in NCBI RefSeq.
- Some viruses, like FLUAV is unique: they are **highly diverse** as a combination of error prone RDRP and reassorment event.
  - —Each sample need a unique reference set.
- Perform **alignment search** to find the best possible reference in the available database.

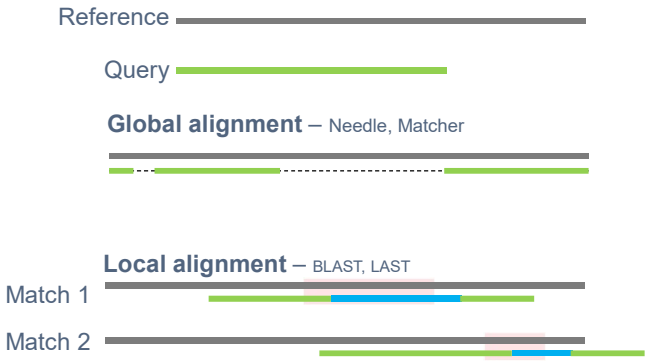
Common

Assembly – finding references

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

Sequence aligment

- Identifying similar region between reference and query – pairwise alignment.
- Can be local or global.




```
Seq1:  WHEREISWALTERNOW       (16aa)
Seq2:  HEWASHEREBUTNOWISHERE  (21aa)
```

Global

```
Seq1:  1                W--HEREISWALTERNOW 16
                        W  HERE
Seq2:  1  HEWASHEREBUTNOWISHERE            21
```

Local

```
Seq1: 1  W--HERE 5        Seq1:  1 W--HERE  5
         W  HERE                   W  HERE
Seq2: 3  WASHERE 9        Seq2: 15 WISHERE 21
```

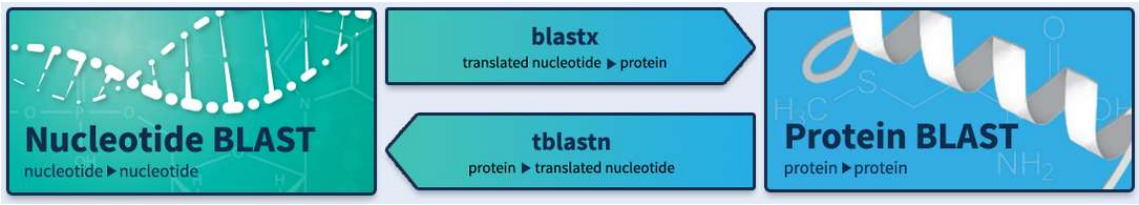
Common

Assembly – finding references

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

Understanding BLAST

- BLAST stands for Basic Local Alignment Search Tool.
- A **heuristic** approach to find similarity between the input/query and the reference/subject.
  —Query or subject can be nucleotide or protein sequences.



- Uses **FASTA** file as an input.

Header →
Sequence →

```
>bff10bf5-4bfd-4faf-8c24-f6365ee9cf0f runid=1d100d4986d3546e81ed8f5b8ce0b2df2affeba0
CACGCCAGCAAAAGCAGGTACTGATTCAAAAATGGAAGACTTTGTTACGTAATGCTTCAATCCAATGATTGTCGAGCTTGCGGA
>beeb80e7-0584-4f4e-9010-1de54464a60b runid=1d100d4986d3546e81ed8f5b8ce0b2df2affeba0
TACGCGCCAGTAGAAACAAAGGGTGTTTTTATCATTAAATAAGCTGAAACGAGAAAGTTTATCTCTTGCTCCACTTCAAGCAAT
```

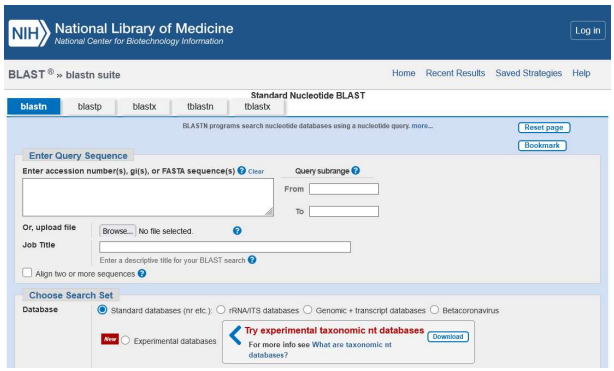
Common

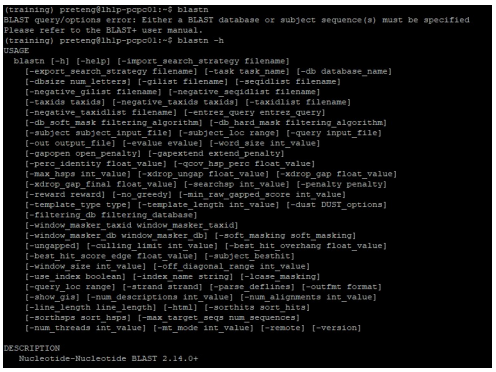
Assembly – finding references

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

Web or Command Line Interface (CLI)

- Web interface – extensive database, but not so practical for large datasets.
- Command Line interface – computationally expensive, customizable database, practical for large dataset.





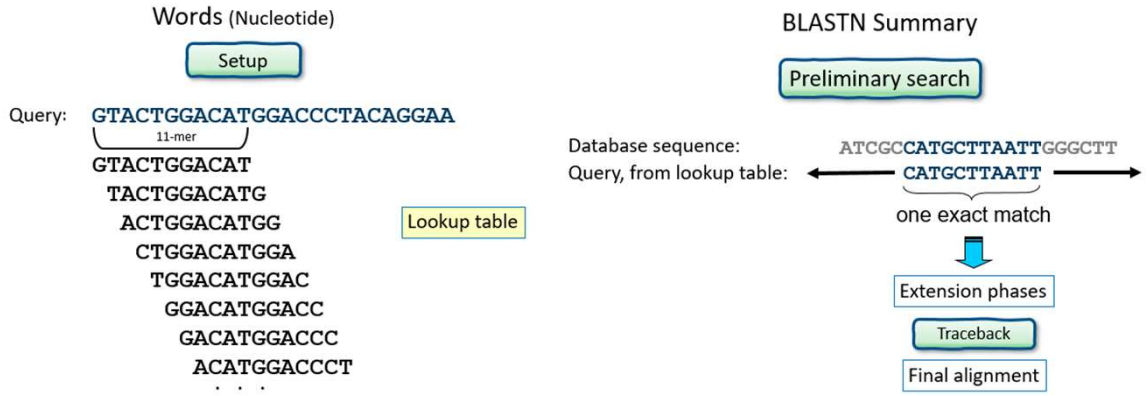
```
(training) preteng@lhlp-pcpc01:~$ blastn
BLAST query/options error: Either a BLAST database or subject sequence(s) must be specified
Please refer to the BLAST+ user manual.
(training) preteng@lhlp-pcpc01:~$ blastn -h
USAGE
  blastn [-h] [-help] [-import_search_strategy filename]
    [-export_search_strategy filename] [-task task_name] [-db database_name]
    [-dbsize num_letters] [-gilist filename] [-seqidlist filename]
    [-negative_gilist filename] [-negative_seqidlist filename]
    [-taxids taxids] [-negative_taxids taxids] [-taxidlist filename]
    [-negative_taxidlist filename] [-entrez_query entrez_query]
    [-db_soft_mask filtering_algorithm] [-db_hard_mask filtering_algorithm]
    [-subject subject_input_file] [-subject_loc range] [-query input_file]
    [-out output_file] [-evalue evalue] [-word_size int_value]
    [-gapopen open_penalty] [-gapextend extend_penalty]
    [-perc_identity float_value] [-qcov_hsp_perc float_value]
    [-max_hsps int_value] [-xdrop_ungap float_value] [-xdrop_gap float_value]
    [-xdrop_gap_final float_value] [-searchsp int_value] [-penalty penalty]
    [-reward reward] [-no_greedy] [-min_raw_gapped_score int_value]
    [-template_type type] [-template_length int_value] [-dust DUST_options]
    [-filtering_db filtering_database]
    [-window_masker_taxid window_masker_taxid]
    [-window_masker_db window_masker_db] [-soft_masking soft_masking]
    [-ungapped] [-culling_limit int_value] [-best_hit_overhang float_value]
    [-best_hit_score_edge float_value] [-subject_besthit]
    [-window_size int_value] [-off_diagonal_range int_value]
    [-use_index boolean] [-index_name string] [-lcase_masking]
    [-query_loc range] [-strand strand] [-parse_deflines] [-outfmt format]
    [-show_gis] [-num_descriptions int_value] [-num_alignments int_value]
    [-line_length line_length] [-html] [-sorthits sort_hits]
    [-sorthsps sort_hsps] [-max_target_seqs num_sequences]
    [-num_threads int_value] [-mt_mode int_value] [-remote] [-version]

DESCRIPTION
  Nucleotide-Nucleotide BLAST 2.14.0+
```


Common

Assembly – finding references

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

How does BLAST work?



Source: NLM, NCBI

Common

Assembly – finding references

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

Understanding the tabular BLAST output

qseqid	sseqid	pident	length	mismatch	gapopen	qstart	qend	sstart	ssend	evalue	bitscore
23c3a299-f8d1-44b7-8ee6-7d809e74fb5a	2247312	95.56	901	9	23	14	895	4	892	0.00E+00	1413
23c3a299-f8d1-44b7-8ee6-7d809e74fb5a	2247304	95.56	901	9	23	14	895	4	892	0.00E+00	1413
23c3a299-f8d1-44b7-8ee6-7d809e74fb5a	2247296	95.56	901	9	23	14	895	4	892	0.00E+00	1413
23c3a299-f8d1-44b7-8ee6-7d809e74fb5a	2247288	95.56	901	9	23	14	895	4	892	0.00E+00	1413
23c3a299-f8d1-44b7-8ee6-7d809e74fb5a	2247192	95.56	901	9	23	14	895	4	892	0.00E+00	1413

- A summarized version of the result, without the alignment.
  - **qseqid**: sequence name.
  - **sseqid**: name of the subject (reference sequence).
  - **pident**: percent identity.
  - **length**: alignment length, not always equal to read length.
  - **mismatch**: number of mismatch.
  - **gapopen**: number of gap openings.
  - **qstart**: start position of the alignment in query.
  - **qend**: end position of the alignment in query.
  - **sstart**: start position of the alignment in subject.
  - **send**: end position of the alignment in subject.
  - **evalue**: expect value, in short how likely to find the alignment by chance.
  - **bitscore**: alignment score (in bit).

Common

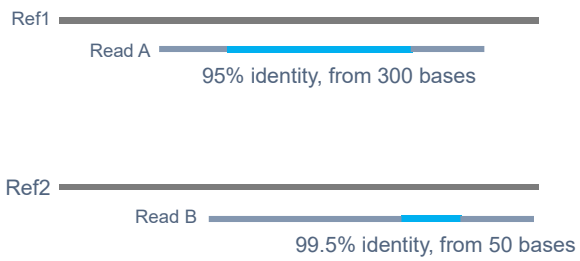
Assembly – finding references

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

Understanding the tabular BLAST output

qseqid	sseqid	pident	length	mismatch	gapopen	qstart	qend	sstart	ssend	evalue	bitscore
23c3a299-f8d1-44b7-8ee6-7d809e74fb5a	2247312	95.56	901	9	23	14	895	4	892	0.00E+00	1413
23c3a299-f8d1-44b7-8ee6-7d809e74fb5a	2247304	95.56	901	9	23	14	895	4	892	0.00E+00	1413
23c3a299-f8d1-44b7-8ee6-7d809e74fb5a	2247296	95.56	901	9	23	14	895	4	892	0.00E+00	1413
23c3a299-f8d1-44b7-8ee6-7d809e74fb5a	2247288	95.56	901	9	23	14	895	4	892	0.00E+00	1413
23c3a299-f8d1-44b7-8ee6-7d809e74fb5a	2247192	95.56	901	9	23	14	895	4	892	0.00E+00	1413

- Sorting the result based on percent identity?



- Important parameter to sort the results:
  - E-value – bit score corrected for database size.
  - Bit score.
  - Percent identity.
- Thresholds can be set for pident and evalue.

Common

Assembly – finding references

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

How to optimize BLAST search

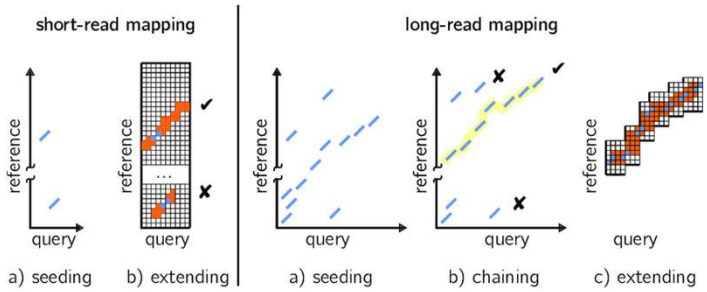
- Database:
  - —Smaller database – faster.
  - —Completeness.
- As a **heuristic process**, it is difficult to define “best hit” from the search.
  - —Avoid limiting the result to less than five hits.
- Word size – define the length of the initial sequence to look for (default is 28 for megablast).
  - —Smaller word size – more accurate, but takes more time.
- BLAST has several algorithms, or tasks:
  - —megablast (default): for very similar sequences.
  - —dc-megablast: discontigous megablast, for more dissimilar sequence.
  - —blastn: finding related sequence from other species – find homologous.
  - —blastn-short: short sequences, <30 bp.

Common

Assembly – Mapping

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

- Mapping → simply aligning the reads to the reference genome.
- Mapping tools is specific to long read, or short read.
  - —Difference in algorithm used.
  - —Popular tools for **short read**: BWA, BWA-MEM, Bowtie2.
  - —Popular tools for **long read**: minimap2, BWA-MEM.
  - —Splice aware mapping for RNA-seq: STAR, Tophat.

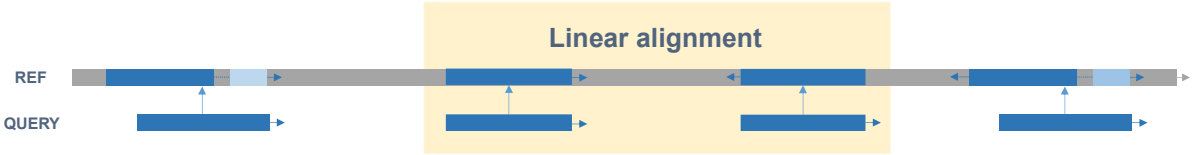


Sahlin et al., Genome Biology, 2023.

Common

Assembly – Mapping

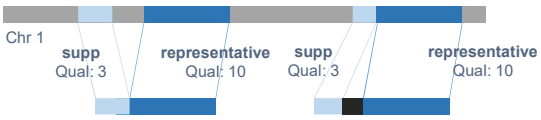
Primary, secondary, and supplementary



Multiple mapping

- i.e. tandem repeats.

Chimeric alignment



- Structural variants – insertion, deletion, inversion

Common

Assembly – SAM and BAM

- Mapping results are stored as a **SAM (Sequence Alignment Map)** or in compressed form as **BAM (Binary Alignment Map)**.
- A specialized tool called **samtools** is available to manipulate SAM/BAM files (http://www.htslib.org/doc/samtools.html).
- BAM file can be visualized:
  - —Available tools: **IGV**, Tablet, **Ugene**, Geneious, etc.
  - —The reference files must be the same reference used for mapping.
  - —The index file is required.

Common

Assembly – SAM and BAM

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

SAM format

```
@HD VN:1.6 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
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
 1    2   3   4  5      6      7   8         9      10           11
```

Header section

1. QNAME : query name
2. **FLAG : alignment information**
3. RNAME : reference name
4. POS : start position of the alignment.
5. MAPQ : mapping quality
6. CIGAR : alignment summary
7. Mate pair information, RNEXT: ref sequence name of mate pair and PNEXT: position of mate pair.
8. TLEN: number of bases covered by the reads from the same fragment. Plus/minus sign marks of the read is on rightmost or leftmost.
9. SEQ : read sequence.
10. QUAL : read quality.
11. Optional field

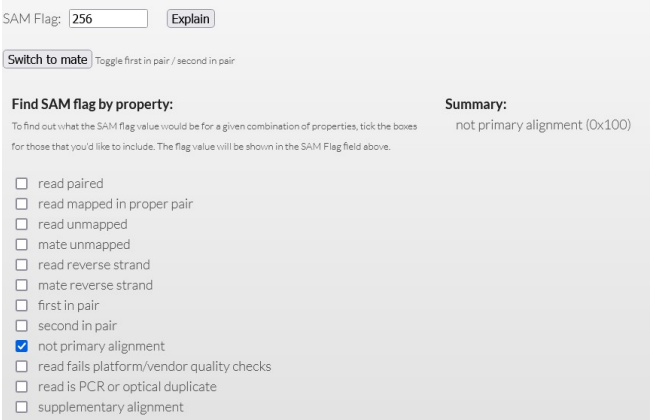
Common

Assembly – SAM and BAM

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

SAM Flags

- Each read will be assigned a “flag” which sign us what kind of aligment the read has.
  - —Flag 4 is for unmapped reads.
  - —Flag 2048 is for supplementary reads.
- Can be used to filter the BAM/SAM file.



https://broadinstitute.github.io/picard/explain-flags.html

Common

Assembly – SAM and BAM

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

CIGAR string

- Stands for Compact Idiosyncratic Gapped Alignment Report.
- How alignment is written in a sam file.
- Operators
  - —M = Match
  - —N = gap
  - —D = deletion
  - —I = insertion

Examples

Ref : ATGCGTCG TAAGCGTG
Query : CGTCGTTAAGCG
Cigar : 5M1I6M

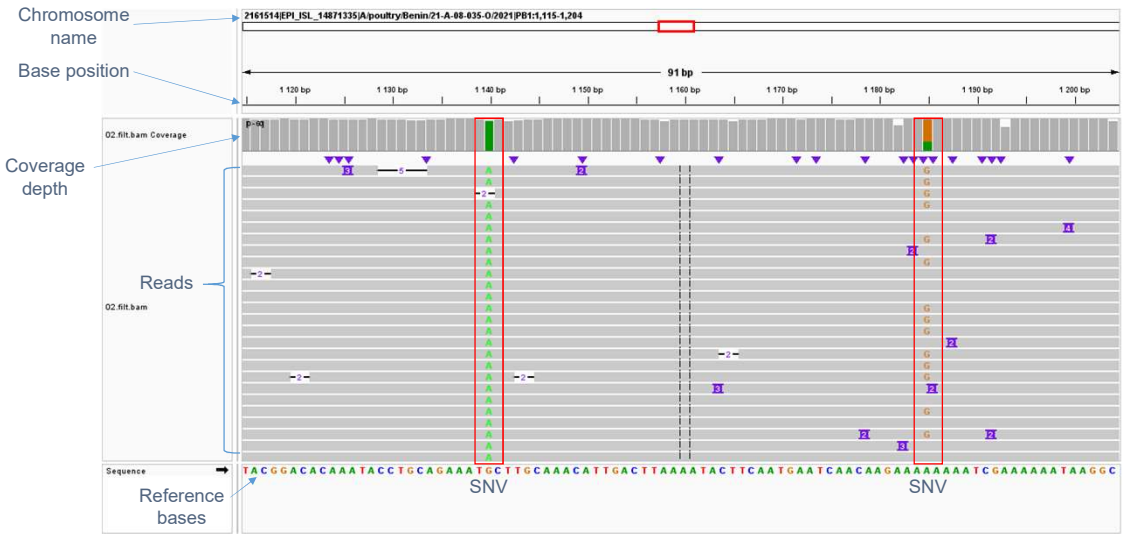
Ref : ATGCGTCGTTAA CGTG
Query : CGTCG AAGCG
Cigar : 5M2D2M1I2M

Common

Assembly – Visualizing mapping result

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

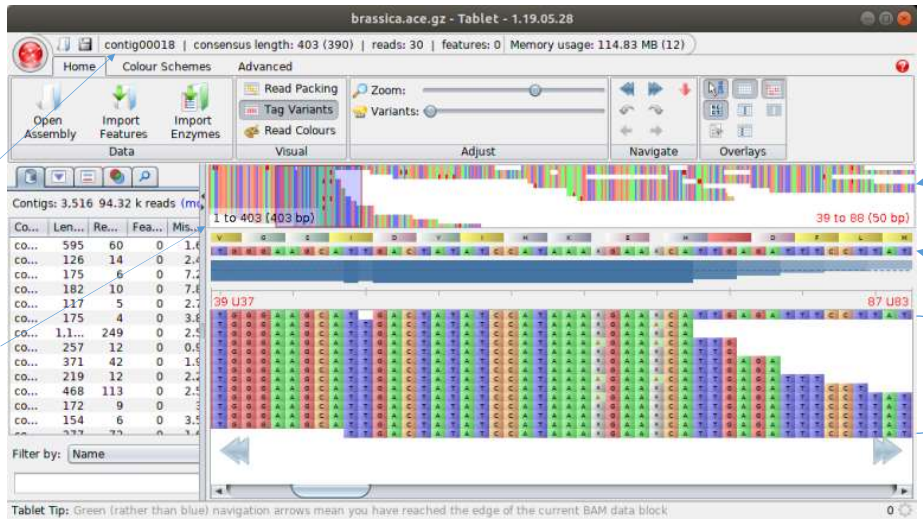
Visualization - IGV



Common

Assembly – Visualizing mapping result

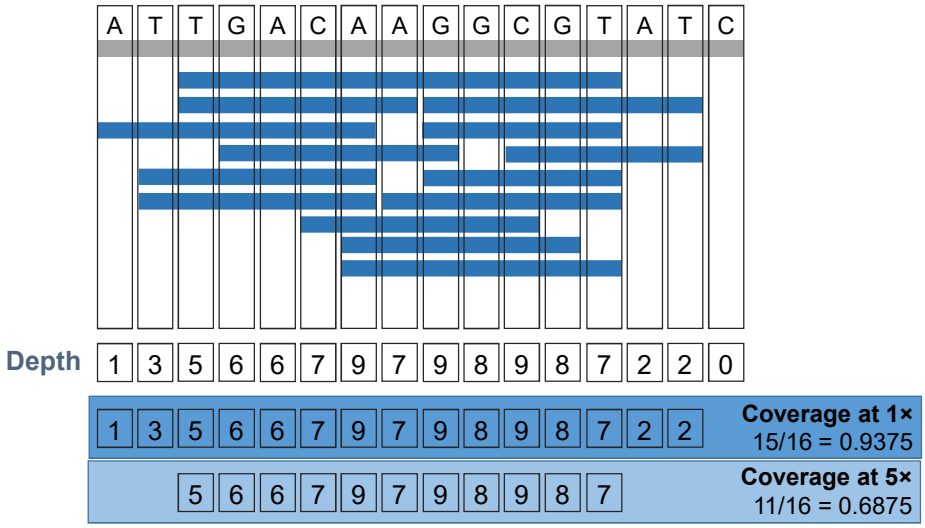
Visualization - Tablet



Common

Assembly – Genome coverage

Breadth (or simply coverage) and depth



Common

Assembly – Genome coverage

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

Statistics

- Get alignment statistics – *samtools coverage*.
- Calculate the coverage and mean depth.
  - —Important – *samtools coverage* calculates read with depth more or equal to one.
  - —Calculate the depth at certain × manually using output from *samtools depth*.
  - —The depth threshold varies; 10× to 30× is suggested for Illumina read.

#rname	startpos	endpos	numreads	covbases	coverage	meandepth	meanbaseq	meanmapq
1963319|EPI_ISL_9012572|NA	1	1458	691	1458	100	426.525	19.5	59.6
2130820|EPI_ISL_14533193|MP	1	1027	12841	1027	100	11814.3	19.4	59.8
2247243|EPI_ISL_16189978|PA	1	2227	257	2227	100	94.2685	20.3	59.2
2247289|EPI_ISL_16190394|PB2	1	2345	370	2345	100	97.449	19.9	58.6
2247293|EPI_ISL_16190394|NP	1	1569	856	1569	100	662.576	20.3	59.9
2247300|EPI_ISL_16190395|HA	1	1780	485	1780	100	226.219	19.5	59.4
2247306|EPI_ISL_16190578|PB1	1	2342	174	2342	100	55.1763	20.2	59.3
2247312|EPI_ISL_16190578|NS	1	892	21749	892	100	20203.7	19.6	59.9

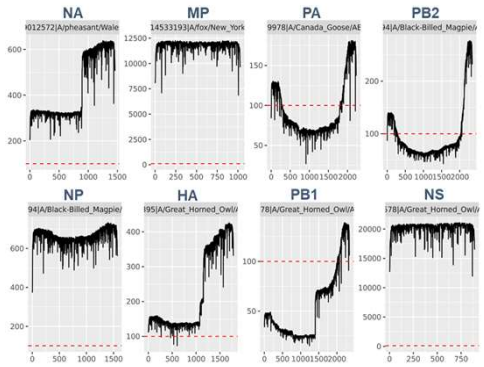
Common

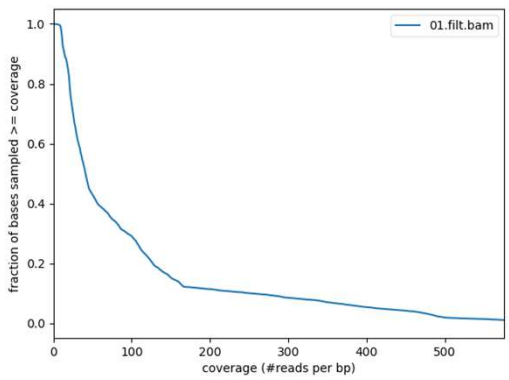
Assembly – Genome coverage

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

Statistics

- Visualization of depth across the genome is useful to identify the low coverage region.
  - —Use samtools depth to get per base depth, which then can be plot using R.
- Other useful tools is deepTools (https://deeptools.readthedocs.io/en/develop/index.html), which gives coverage (y axis) on different depth (x axis).





Common

Assembly – Consensus calling

Collecting information

- Generate a “pileup” file.
  - —Per-base summary of the alignment file.
  - —**Transforming information from aligned read to base/position information**.
  - —Important to take into account: base quality and mapping quality.
  - —Samtools and BCFtools can both generate pileup file, but BCFtools is the preferred approach.

Example of a pileup file

Ref ID	Pos	Ref	Depth	Bases	Qual
1963319|EPI_ISL_9012572|A/pheasant/Wales/385129/2021|NA	1	A	10	^].^].^].^].^].^].^].^].	>44291:>*$
1963319|EPI_ISL_9012572|A/pheasant/Wales/385129/2021|NA	2	G	11	.....^].	<5366,2:,(.
1963319|EPI_ISL_9012572|A/pheasant/Wales/385129/2021|NA	3	C	12	.....^].	<7684,2:.+.&
1963319|EPI_ISL_9012572|A/pheasant/Wales/385129/2021|NA	4	A	13	GGG.....G..^].	26997-28/+3(8

Check https://doi.org/10.1093/bioinformatics/btp352 for more explanation.

Common

Assembly – Consensus calling

Calling variants and apply to the reference

- The pileup file will be processed through BCFtools to call for the variant.
  - —Single nucleotide variants, insertion, deletion, and low coverage region.
- BCFtools is a tool to manipulate BCF/VCF files
  - —BCF – binary variant calling file.
  - —VCF – variant calling file.
- Apply the variants called into the reference sequence.

Pileup file

Pos	Ref	Depth	Bases
4	A	13	GGG.....G..^].
6	G	8	…T..^].

Variants Filter position with depth <10

Pos	Ref	Alt	Sample01.bam
4	A	G	DP=13
6	G	.	DP=8

Reference
ATTACG…

Consensus
ATTGCN…

General flowchart for de-novo assembly

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

Preprocessing	Assembly	Polishing	Evaluation
• Quality check and filtering	• De-novo assembly.	• Error correction	• Re-align reads to contigs/scaffold. • Get coverage breadth and depth. • Genome annotation.

Common

Assembly – De novo assembly

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

- Available tools for de novo assembly for short reads:
  - —Velvet
  - —SOAPdenovo
  - —Forge
  - —ABySS
  - —Megahit (metagenome assembler).
- Available tools for de novo assembly for long reads:
  - —Canu
  - —Flye
  - —MetaFlye (metagenome assembler)

Polishing – error correction

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

Polishing tools for ONT

- Medaka (https://github.com/nanoporetech/medaka)
  - —Taking a draft consensus, remapping the read and apply **neural network** to correct potential error based on a model.
- Nanopolish (https://github.com/jts/nanopolish)
  - —Error correction based on the raw squiggles signals.
- Homopolish (https://github.com/ythuang0522/homopolish)
  - —Use **support vector machine** to distinguish a systematic error or strain variation using homologous sequences.

Polishing tools for Illumina

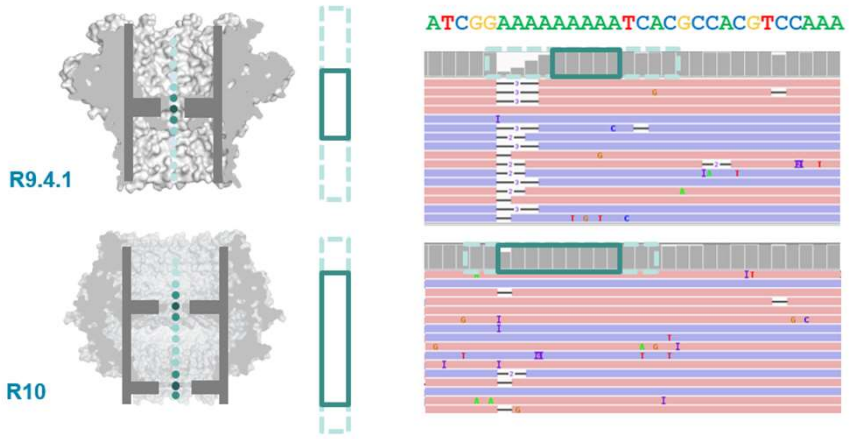
- Pilon (https://github.com/broadinstitute/pilon)
  - —Error correction and gap filling of the consensus by inspecting the alignment/pileup.

Polishing

LUXEMBOURG INSTITUTE OF HEALTH
RESEARCH DEDICATED TO LIFE

Minimize sequencing error

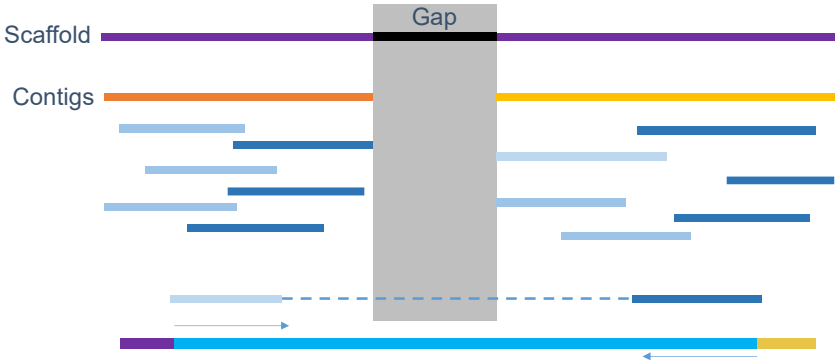
- Long read sequencing is prone to indel error and homopolymer error.
  - —Depends on the speed of DNA strand passing the pore: fast = deletion, slow = insertion.
  - —Difficulty in estimating the true length of the homopolymer.



Common

Evaluation

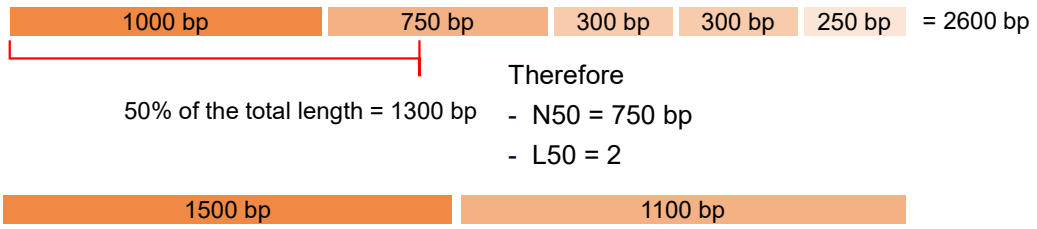
- Terminology:
  - —Contig (from contiguous)
  - —Scaffold
- What do we want?
  - —Complete – length is equal or less to the expected size.
  - —Contiguous – fewer, longer contigs.
  - —Correct contigs.



Common

Evaluation

- Measure of contiguity:
  - —N50: the length of the **shortest contig** in the **group of longest sequences** that together represent (at least) 50% of the nucleotides in the set of sequences.
  - —L50: count of the **smallest number of contigs** (arranged from longest to shortest) that represent (at least) 50% of the nucleotides in the set of sequences.
  - —Rule of thumb = assembly with good contiguity will have longer contigs and less contigs (high N50 and low L50).



Common

Evaluation

- Other parameter:
  - Assembly size compared to expected genome size (can be a proxy to completeness).
  - BUSCO (Benchmarking Universal Single-Copy Orthologs) – evaluation of the genome annotation.

Summary

	Long read	Short read
Basecalling	**Dorado**, Guppy.	Bcl2fastq, Illumina-provided software
Quality check	**Nanoplot**, NanoQC, Fastp, Qualimap	**Fastqc**, Fastp, Qualimap
Demultiplexing	**Dorado**, Guppy, Porechop	
Quality filtering	**Fastp**, Porechop	**Trimmomatic**, Fastp, CutAdapt
Mapping	**Minimap2**, BWA-MEM (up to 1 kb)	BWA-mem, BOWTIE2
De-novo assembly	Canu, **Flye**, Shasta, MetaFlye	**Megahit**, SPAdes, IDBA-UD, ABySS
Hybrid de-novo assembly	MaSuRCA, SPAdes, Unicycler	
Variant calling	iVar, Samtools, BCFtools, Freebayes	Samtools, BCFtools, iVar, Freebayes
Polishing	Medaka, Homopolish, Nanopolish	Pilon