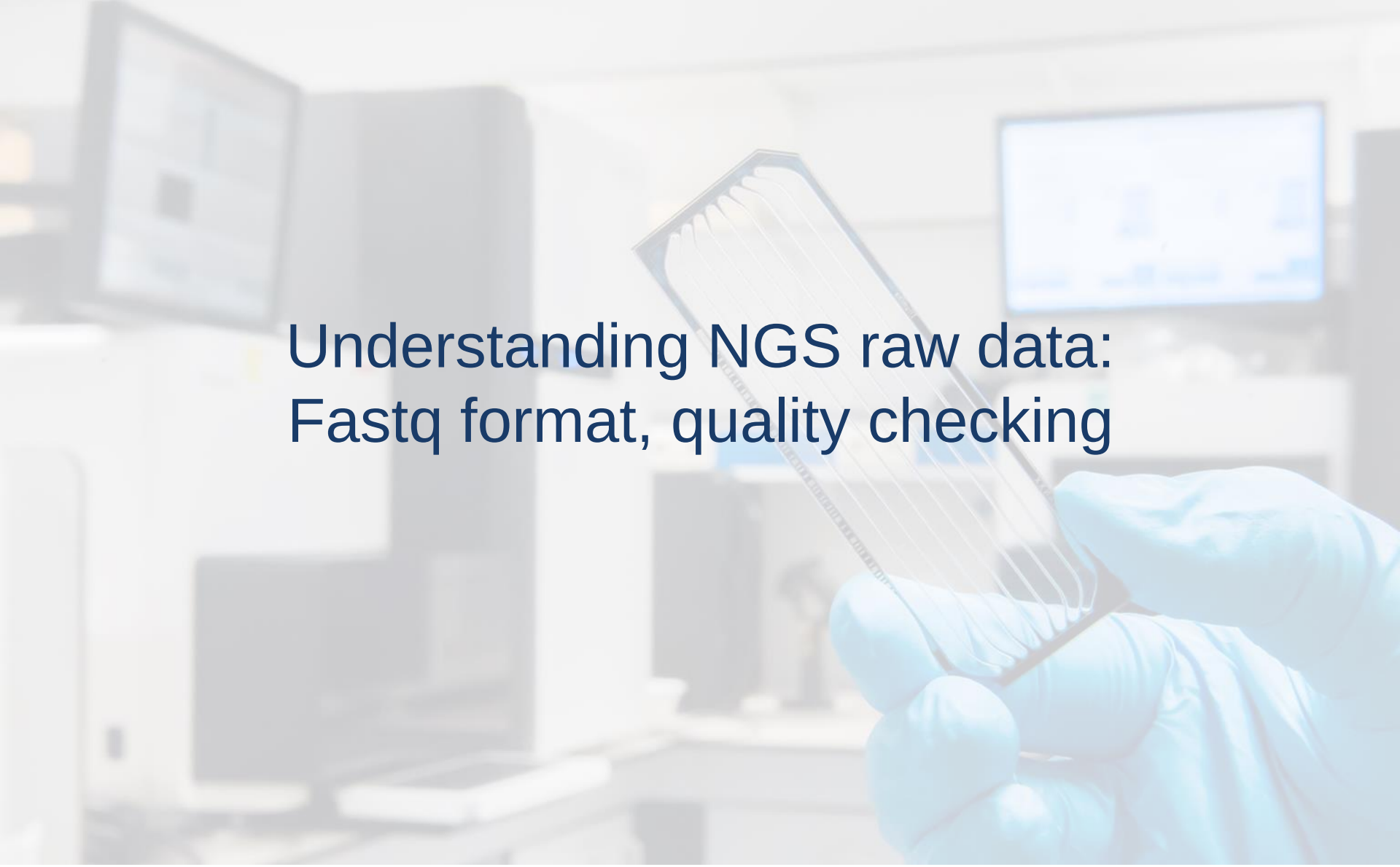


Understanding NGS raw data: Fastq format, quality checking

GENOMICS CORELEUVEN

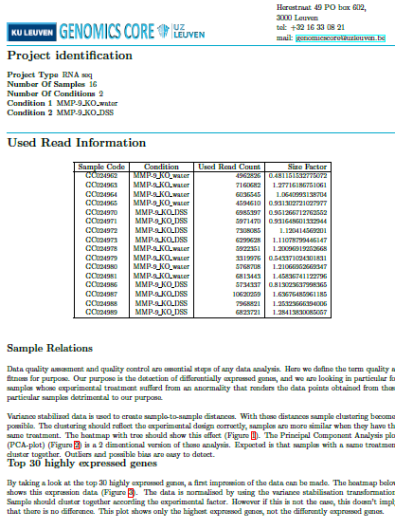
Álvaro Cortés Calabuig
Erika Souche

Overview

- Why this bioinformatics session?
- Basic NGS terminology
- NGS Pipelines
- Fastq format
- Fastq quality control

Why this session?

A typical RNA-Seq analysis at the Genomics Core delivers:



KU LEUVEN GENOMICS CORE UZ LEUVEN

Herestraat 49 PO box 602,
3000 Leuven
tel: +32 16 33 08 21
mail: genomicscore@uzleuven.be

Project identification

Project Type RNA seq
Number Of Samples 16
Number Of Conditions 2
Condition 1 MMP-9_KO_water
Condition 2 MMP-9_KO_DSS

Used Read Information

Sample Code	Condition	Used Read Count	Size Factor
GC024962	MMP-9_KO_water	4962826	0.4811513227775072
GC024963	MMP-9_KO_water	7160682	1.377161867511061
GC024964	MMP-9_KO_water	6036545	1.0645993138704
GC024965	MMP-9_KO_water	4594610	0.9313027211027977
GC024970	MMP-9_KO_DSS	6985397	0.8812667127622552
GC024971	MMP-9_KO_DSS	5971470	0.9316480133239944
GC024972	MMP-9_KO_DSS	7308085	1.12041456920l
GC024973	MMP-9_KO_DSS	6299628	1.1107879946147
GC024978	MMP-9_KO_water	5922351	1.2009691955668
GC024979	MMP-9_KO_water	3319976	0.5433702430l831
GC024980	MMP-9_KO_water	5768708	1.2106052669547
GC024981	MMP-9_KO_water	6813443	1.4583674112279[illegible]
GC024986	MMP-9_KO_DSS	5734337	0.8130263799836[illegible]
GC024987	MMP-9_KO_DSS	10639359	1.4367648596118[illegible]
GC024988	MMP-9_KO_DSS	7968821	1.3532666394606
GC024989	MMP-9_KO_DSS	6825721	1.2841383008[illegible]

Sample Relations

Data quality assesment and quality control are essential steps of any data analysis. Here we define the term quality a fitness for purpose. Our purpose is the detection of differentially expressed genes, and we are looking in particular fo samples whose experimental treatment suffered from an anormality that renders the data points obtained from thes particular samples detrimental to our purpose.

Variance stabilised data is used to create sample-to-sample distances. With these distances sample clustering become possible. The clustering should reflect the experimental design correctly, samples are more similar when they have th same treatment. The heatmap with true should show this effect (Figure 1). The Principal Component Analysis plo (PCA-plot) (Figure 2) is a 2 dimentional version of these analysis. Expected is that samples with a same treatmen cluster together. Outliers and possible bias are easy to detect.

Top 30 highly expressed genes

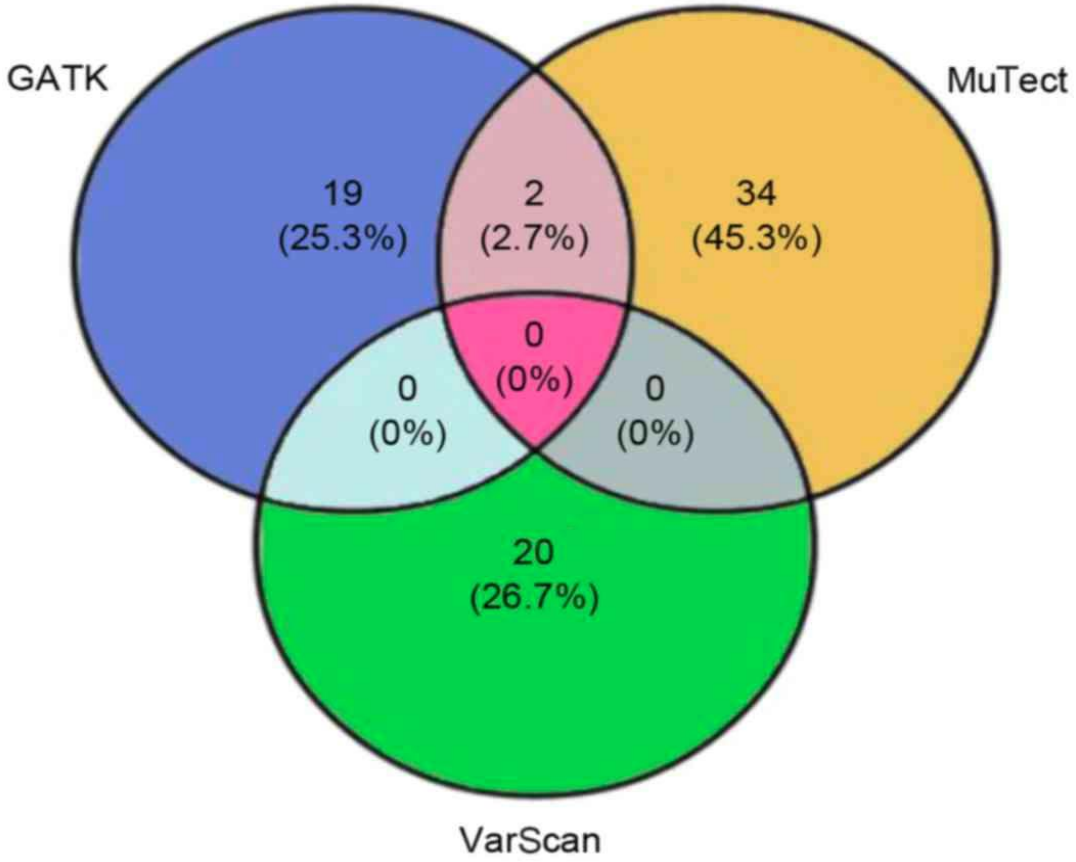
By taking a look at the top 30 highly expressed genes, a first impression of the data can be made. The heatmap belo shows this expression data (Figure 3). The data is normalised by using the variance stabilisation transformatio Samples should cluster together according the experimental factor. However if this is not the case, this doesn't impl that there is no difference. This plot shows only the highest expressed genes, not the differently expressed genes.

- 10 report files: differentially expressed genes, reads quality report, counting report,…

3GB of raw and analyzed data: Fastq, BAM, counts, normalized counts, gene expression files, heatmaps…

1. Results Interpretation & Analysis

2. Data Re-analysis

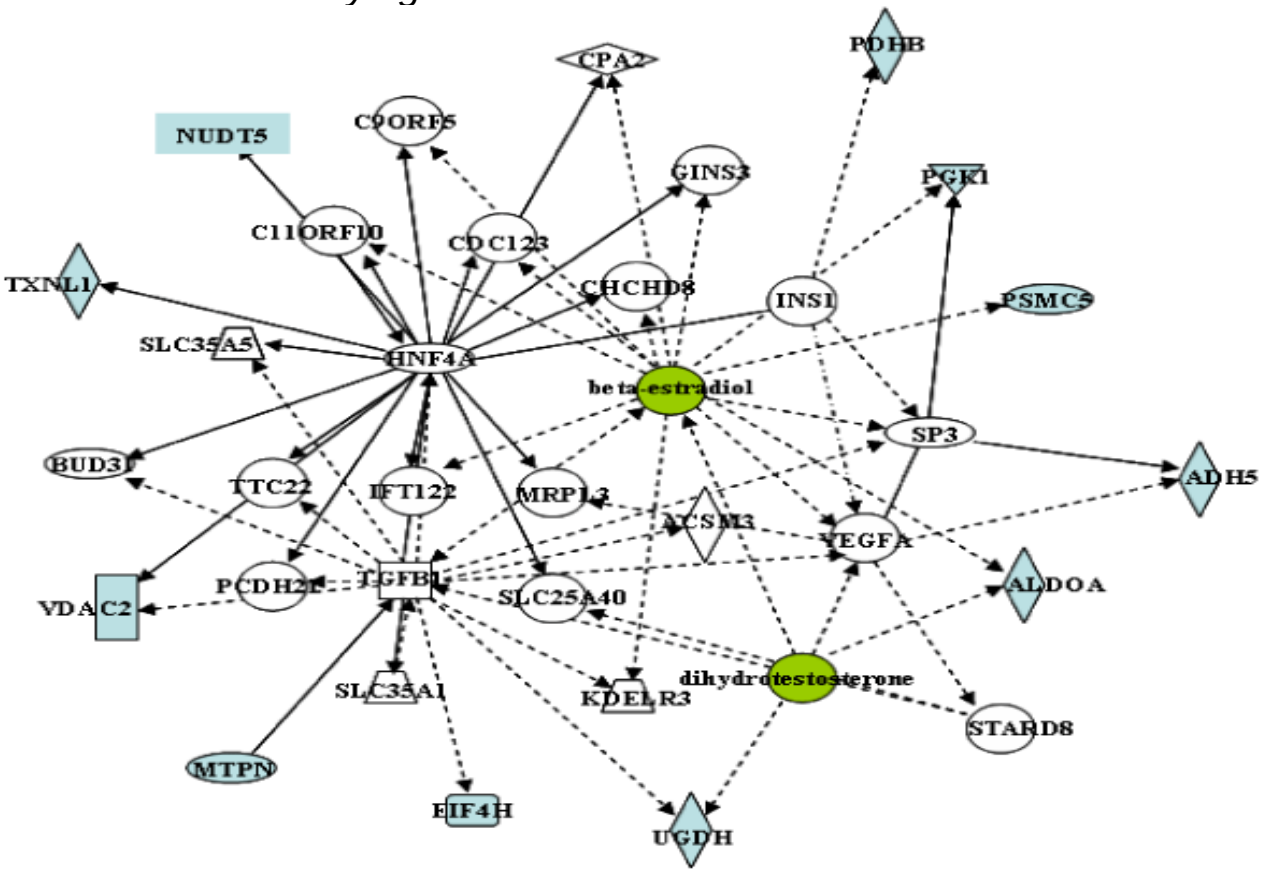


A three-caller pipeline for variant analysis of cancer whole-exome sequencing data

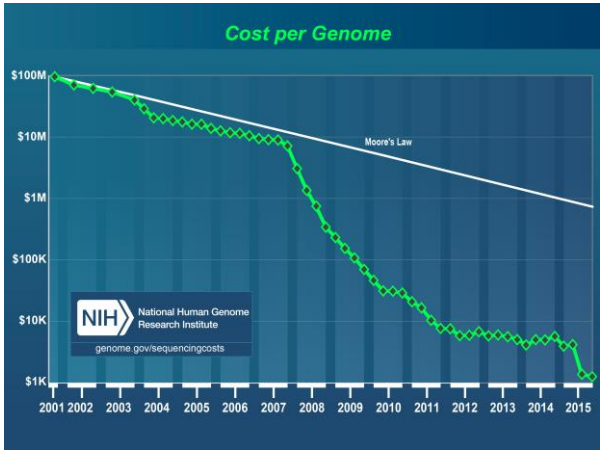
Authors: Ze-Kun Liu, Yu-Kui Shang, ✉ Zhi-Nan Chen, ✉ Huijie Bian

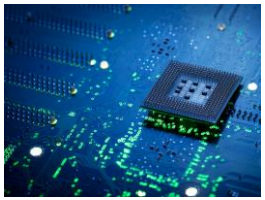
3. Downstream Analysis

Reviewer: "The transcriptional and proteomic profiling experiments propose some interesting followup pathways for further analysis that may shed light on the mechanisms underlying these conditions. "



4. Costs Saving





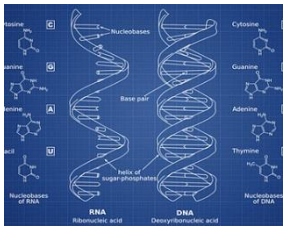
- $1.000 sequencing and $5.000 bioinformatics costs?!?!

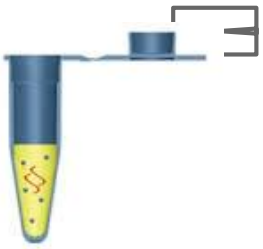
Overview

- Why this bioinformatics session?
- Basic NGS terminology
- NGS Pipelines
- Fastq format
- Fastq quality control

NGS Bioinformatics

- **Experimental Design**





- **Library Preparation**

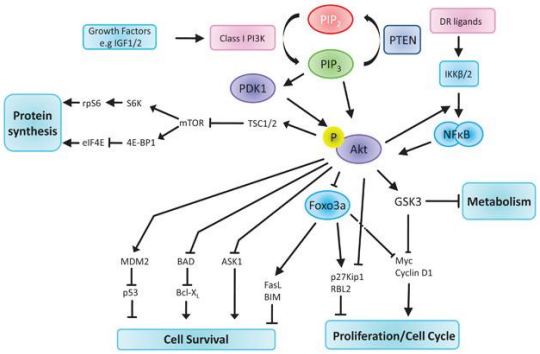
- **Sequencing**

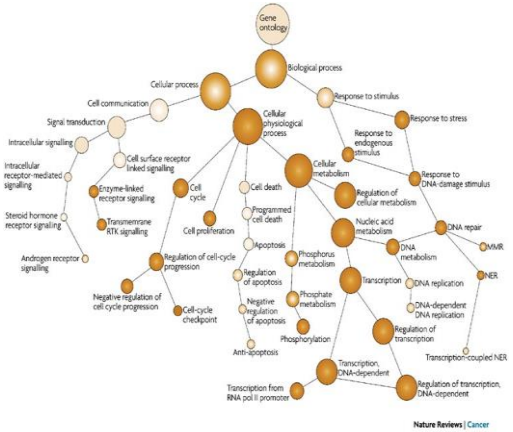
NGS Bioinformatics

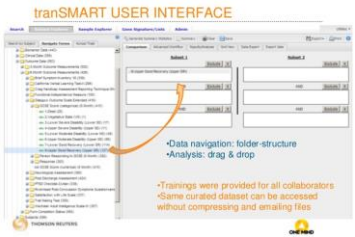
- Follow up and support

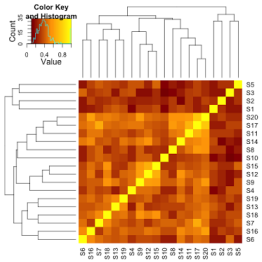
Bioinformatics...

...is a broad field

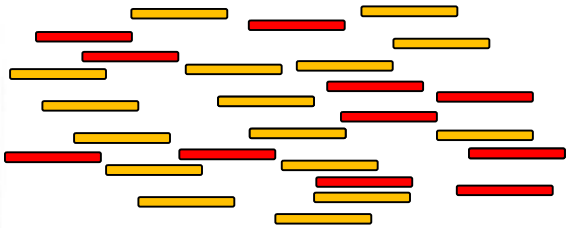








NGS Bioinformatics



NGS bioinformatics: interpretation and analysis of NGS data using informatics tools

A day in NGS Bioinformatics

- Poly-A tail
- Fastq
- Ontology
- Fragments cluster
- Hadoop
- Xenograft Murine infiltration
- Flow Cell
- Demultiplexing
- VCF File
- Ortholog gene
- BAM
- Negative binomial distribution
- Cluster
- snRNA
- Bonferroni Correction
- Barcode

Computer Cluster

NGS data means big data...means big computing power…by now.

Whole Human Genome:
300Gb, gzip file

Exome data:
6GB

RNA-Seq
1GB

NGS data is usually analyzed on a supercomputer or cluster:

- UZ Leuven: Hydra, Google Cloud
- KU Leuven: VSC Flemish Super Computer

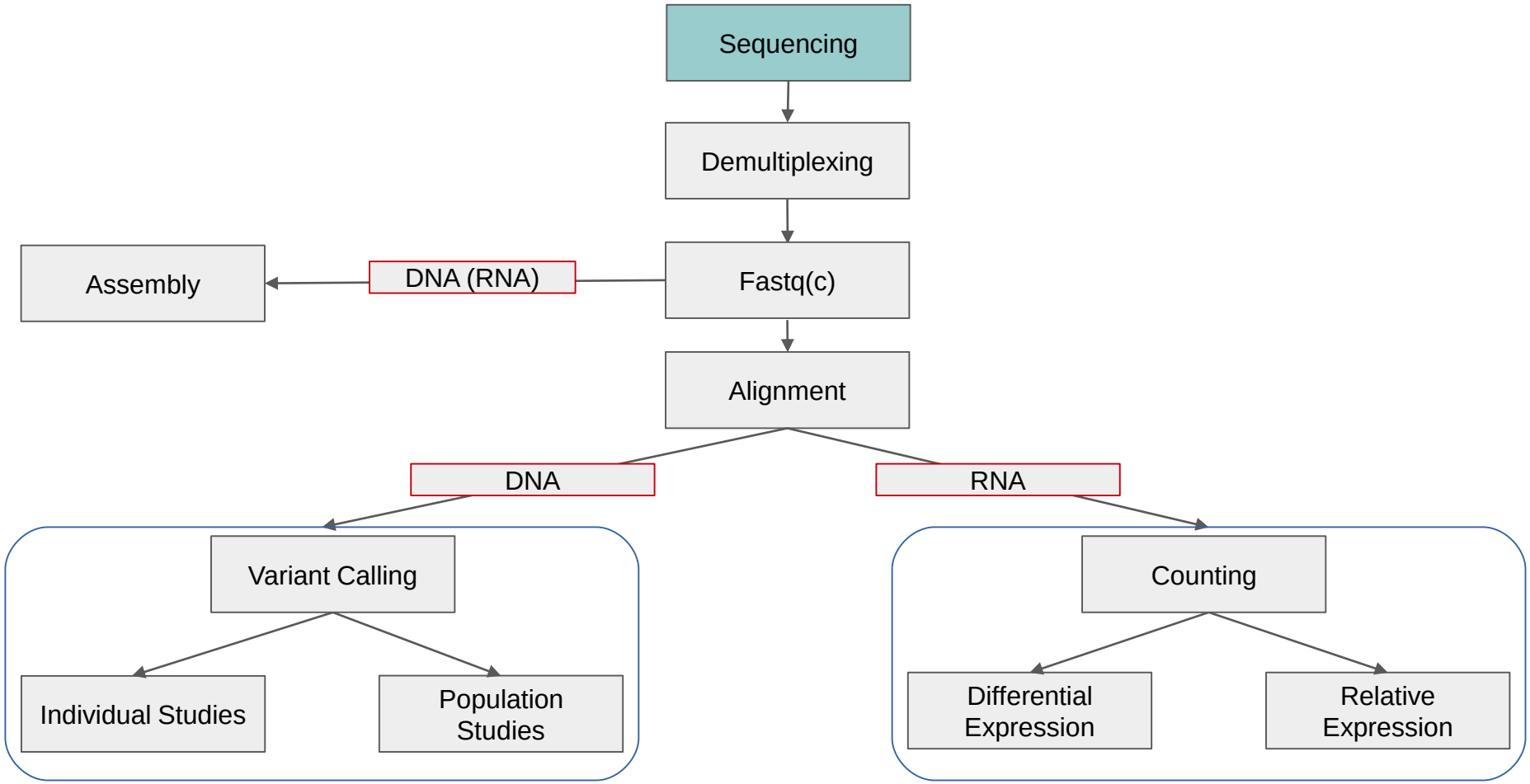
Overview

- Why this bioinformatics session?
- Basic NGS terminology
- NGS Pipelines
- Fastq format
- Fastq quality control

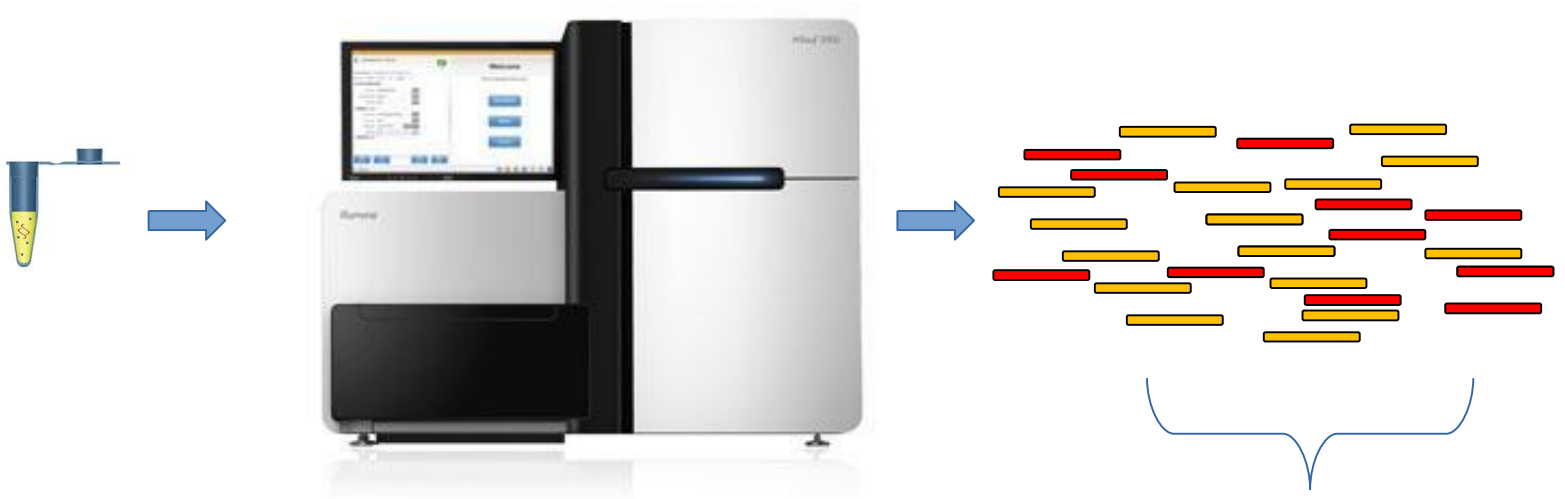
NGS Common Pipelines

- Computational steps to transform input data into processed output information in the context of biological data

Sequencing


Sequencing
Demultiplexing
Fastq(c)
DNA (RNA)
Assembly
Alignment
DNA
RNA
Variant Calling
Individual Studies
Population Studies
Counting
Differential Expression
Relative Expression

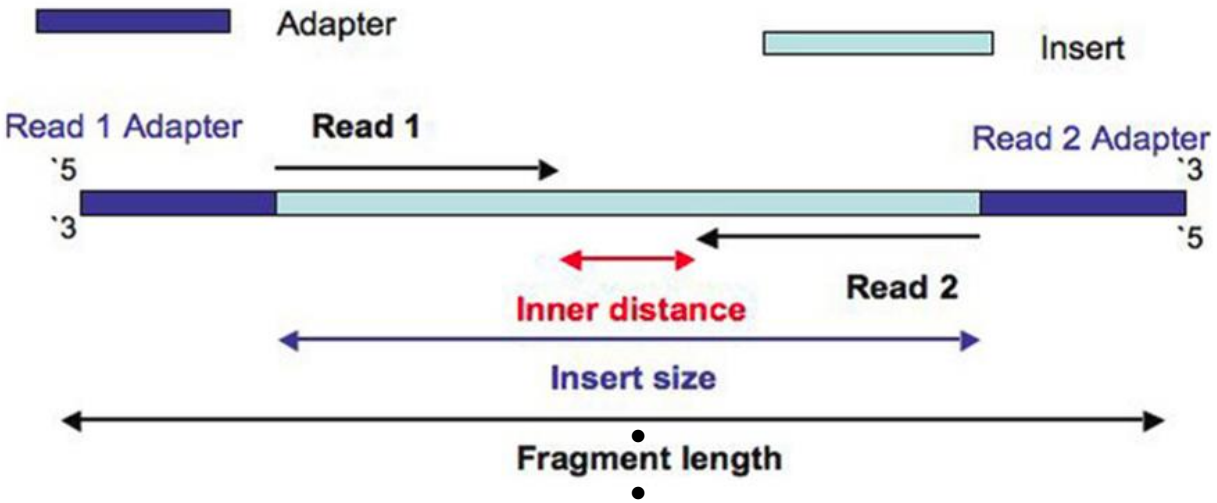
From DNA fragments to Digital reads



DNA Library

Reads

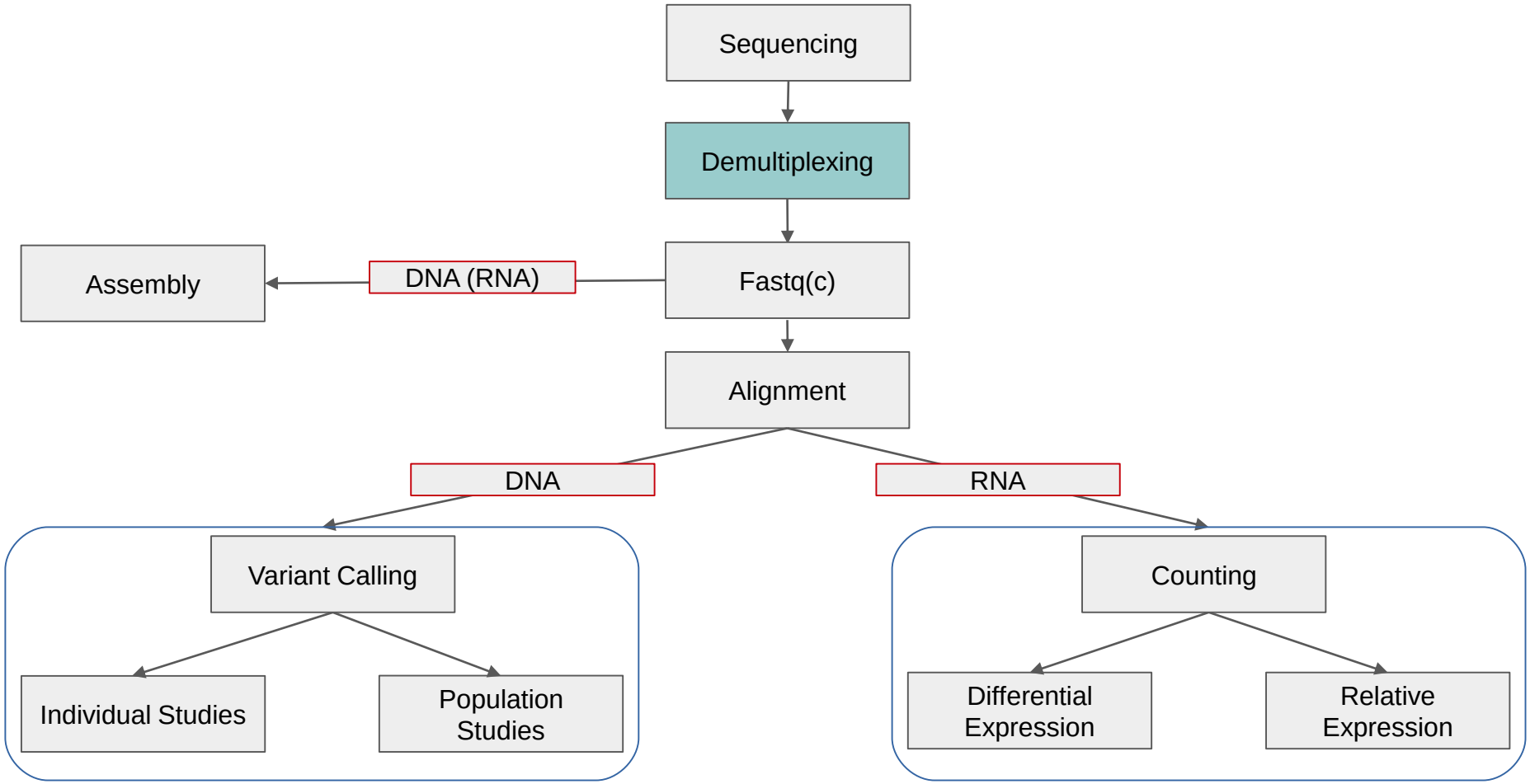
(c)DNA Fragments and Reads



Fragment: biological entity
Read: bioinformatics concept

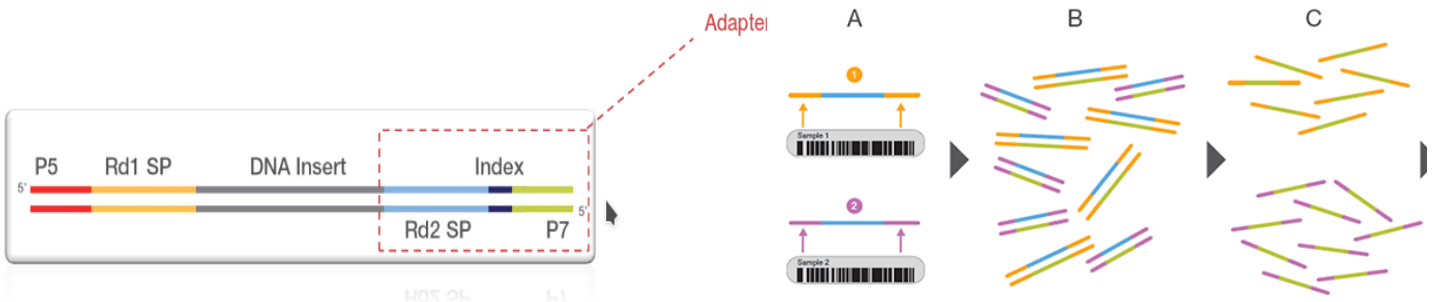
- Fragment: the DNA template + adapters that were loaded on the sequencing machine (is not completely sequenced)
- Read: a raw sequence originating from a sequencing machine
- Single Read: Sequencing only from one end
- Paired-end: Sequencing starting from both ends of the insert

Demultiplexing

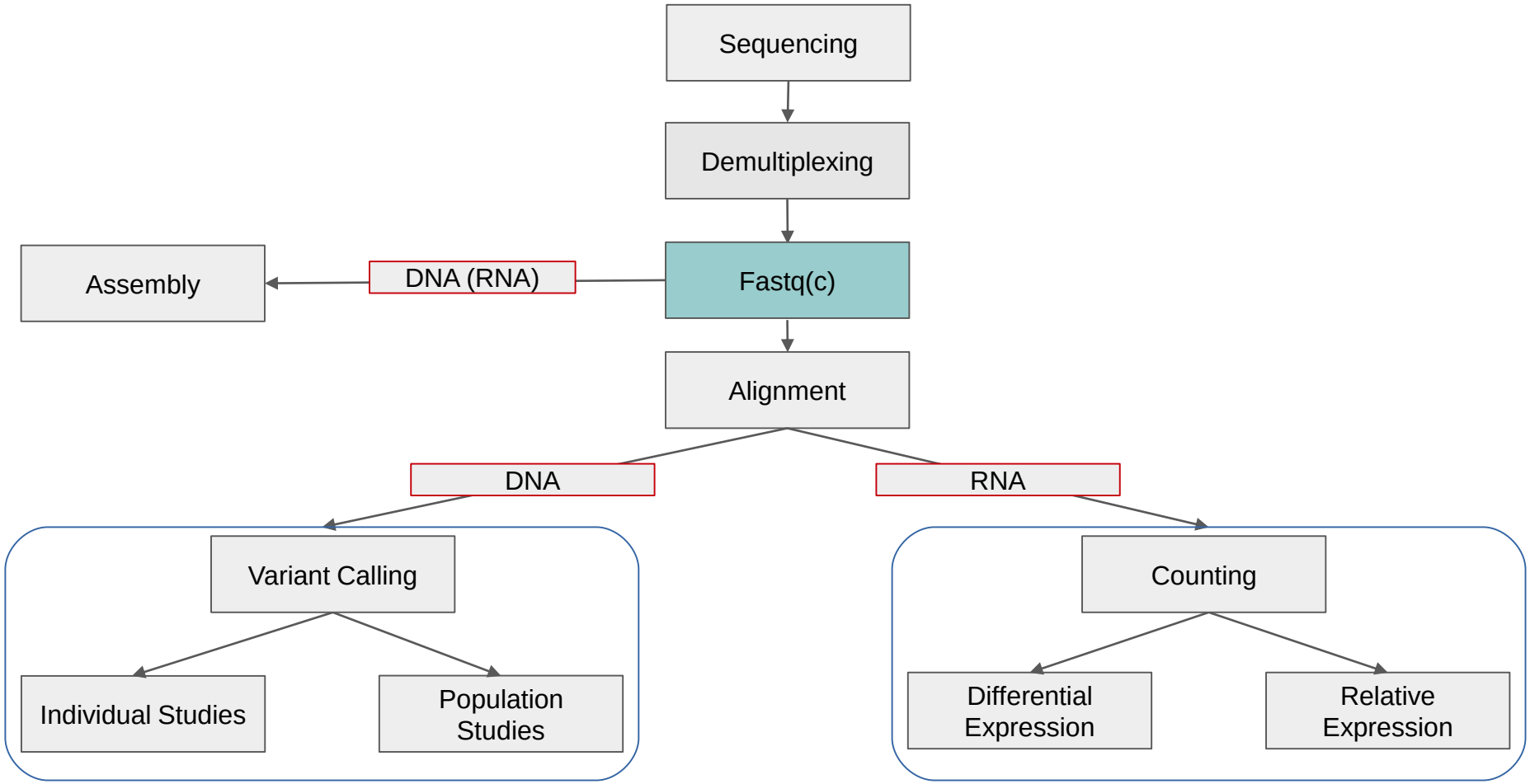


(De)Multiplexing

Multiple samples can be *pooled together or multiplexed* into one or more flowcells



Fastq Files


Sequencing
Demultiplexing
Assembly
DNA (RNA)
Fastq(c)
Alignment
DNA
RNA
Variant Calling
Counting
Individual Studies
Population Studies
Differential Expression
Relative Expression

Fastq Files

The result of demultiplexing is one or more fastq files containing raw reads

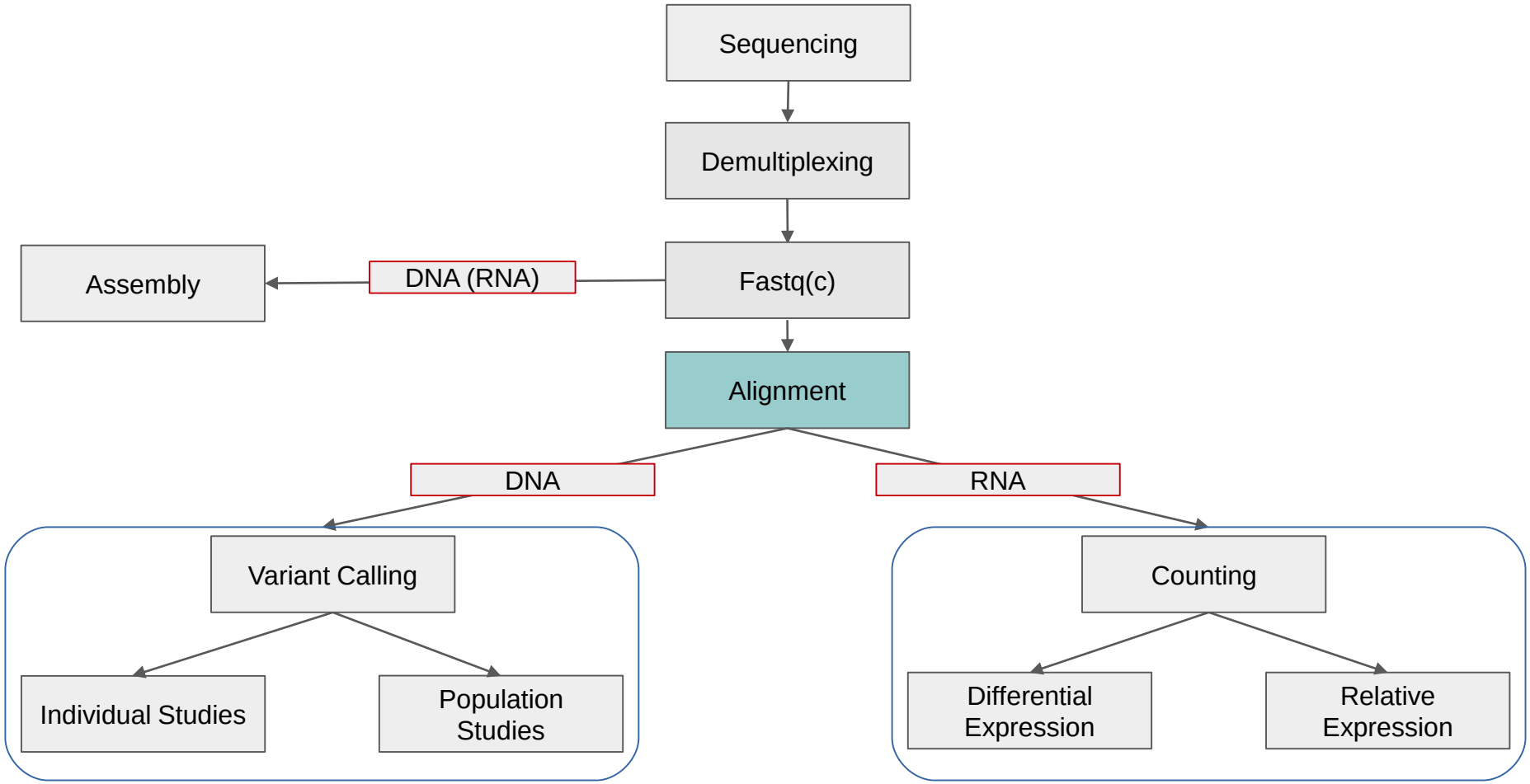
Fastq files are:

- Human readable (not binary) text files.
- Referred as raw data.
- Real diamonds NGS bioinformatics project
- Not ordered wrt originating (c)DNA
- Often compressed using gzip

```
-rwxr-xr-x  1 vsc31439 lp_biogenomics  32G 17 janv. 15:05 GC036462.R1.fastq.gz
-rwxr-xr-x  1 vsc31439 lp_biogenomics  39G 17 janv. 15:49 GC036462.R2.fastq.gz
-rwxr-xr-x  1 vsc31439 lp_biogenomics  31G 17 janv. 16:25 GC036463.R1.fastq.gz
-rwxr-xr-x  1 vsc31439 lp_biogenomics  37G 17 janv. 17:07 GC036463.R2.fastq.gz
-rwxr-xr-x  1 vsc31439 lp_biogenomics  31G 17 janv. 17:41 GC036464.R1.fastq.gz
-rwxr-xr-x  1 vsc31439 vsc31439        37G 17 janv. 18:21 GC036464.R2.fastq.gz
```

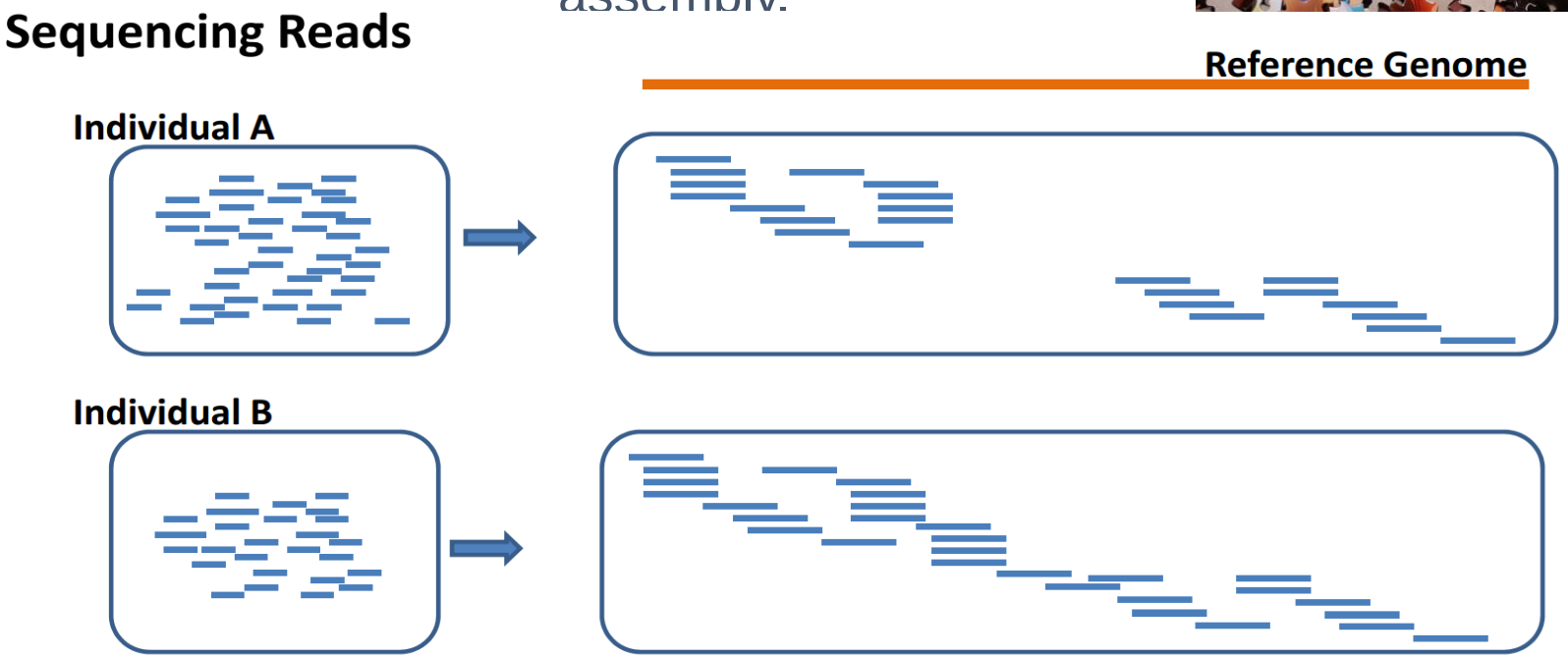
```
: vsc31420@hpc-p-login-1 /staging/leuven/stg_00019/full_genomes/test_ws 11:48 $ ls -lha  GC036463.R1.fastq
-rwxr-xr-x 1 vsc31420 vsc31420 134G 29 mars  11:24 GC036463.R1.fastq
```

Mapping and Alignment



Mapping and Alignment to Reference

Mapping refers to the process of aligning short reads to a reference sequence, whether the reference is a complete genome, transcriptome, or *de novo* assembly.



Reference Sequence in Fasta Format

- genome.fa human-readable nucleotide sequence
- Species dependent: Mouse genome: 2.6GB
- Evolves as it is updated

- 

Mapping vs Alignment

- Reference Ch1:

  1234567890123456
  ATGGTTACACCATT

- Read:

  GGTTCA

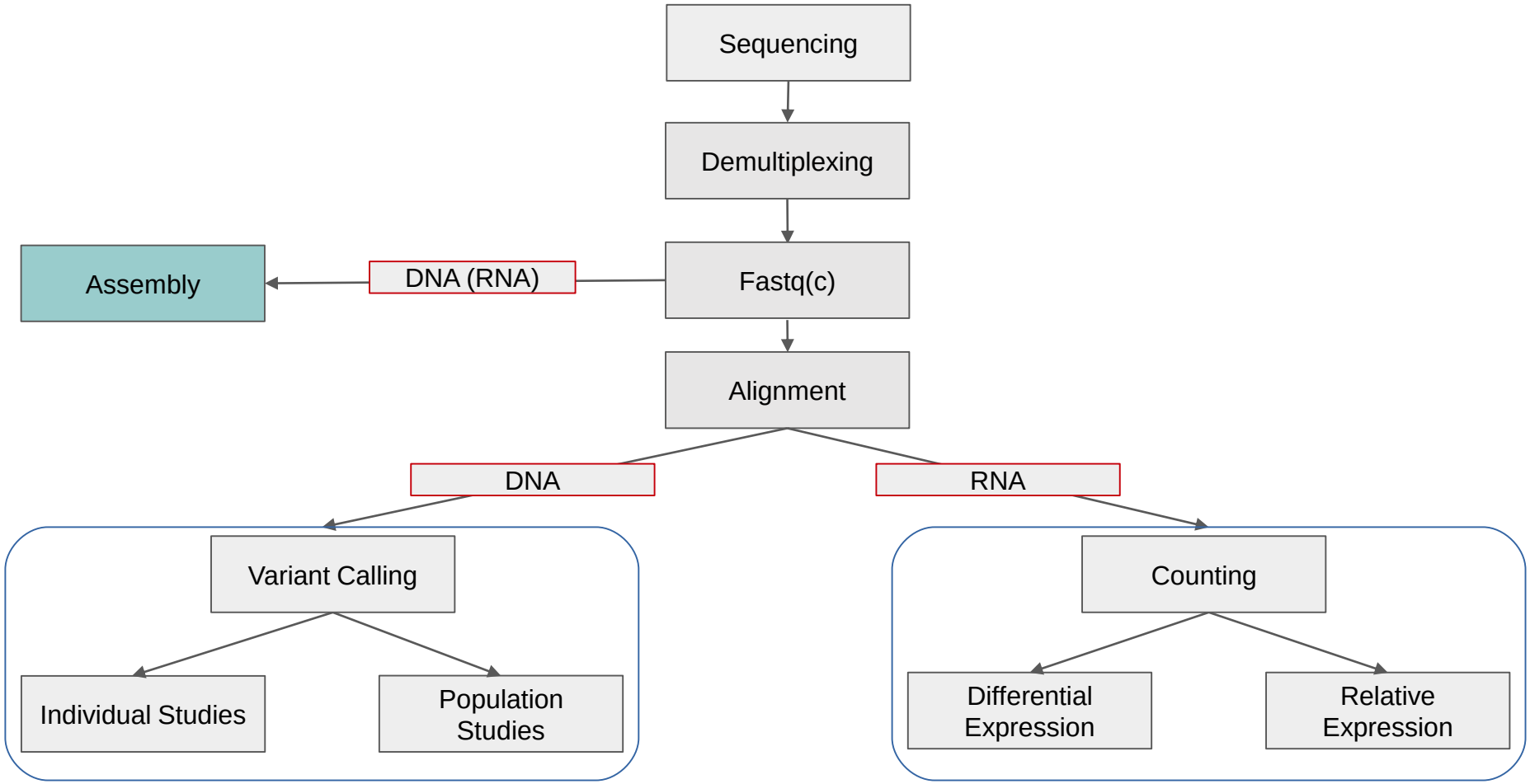
- Possible alignment:

  ATGGTTACACCATT
  GGTT-CA

- Mapping: Ch1-pos3

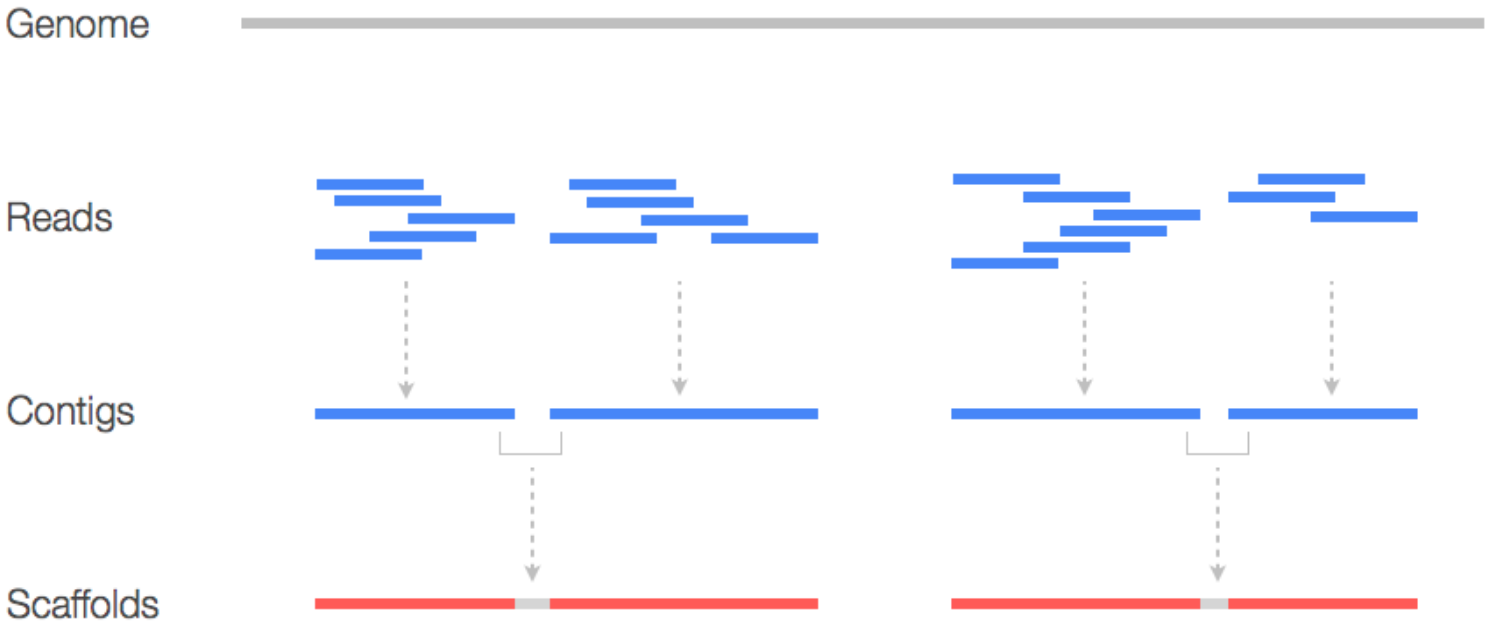
- *"Also of note is that by this time the terms "read alignment" and "read mapping" had become interchangeable. The BWA and Bowtie papers both used both terms, as did many other papers."*
- https://liorpachter.wordpress.com/2015/11/01/what-is-a-read-mapping

Assembly


Sequencing
Demultiplexing
Fastq(c)
DNA (RNA)
Assembly
Alignment
DNA
RNA
Variant Calling
Individual Studies
Population Studies
Counting
Differential Expression
Relative Expression

Assembly

- The generation of a reference, from scratch (*de novo*) or reference assisted.
- Overlapping reads are merged to contigs (smallest unitable unit without unknown bases)
- Contigs that belong together, but where the connecting sequence is unknown, can be connected to scaffolds, inserting N's for the unknown bases



Overview

- Why this bioinformatics session?
- Basic NGS terminology
- NGS Pipelines
- Fastq format
- Fastq quality control

Fastq Format and Reads

- What is a “read”?
  - A raw sequence (ordered collection) of nucleotides names A,C,G,T, or N.
  - Read length is experiment-dependent
    - mRNA differential expression: 51bp SE
    - Whole-Genome sequencing: 150bp PE
- Fastq file?
  - Plain-text file, where each read and complementary information occupies 4 consecutive lines

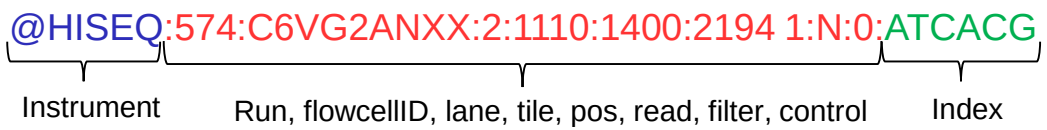
```
@HISEQ:574:C6VG2ANXX:2:1110:1400:2194 1:N:0:ATCACG
GGGGGATTCTCACTAGGTCTCAAGGTCTCTCACTCTCGGTAGTGTTCCCAG
+
CCCCCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
@HISEQ:574:C6VG2ANXX:2:1110:1560:2177 1:N:0:ATCACG
ATGGTCCAGCAAGGGGTATGCTGAGAAGGGGAGCAGTTCAGAACCCATCAG
+
CCCCCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
@HISEQ:574:C6VG2ANXX:2:1110:1583:2223 1:N:0:ATCACG
CTACCTTCACTATCAACATAGCAAACACACCTTTAGCTCCAGCTATTAACA
+
CCCCCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
@HISEQ:574:C6VG2ANXX:2:1110:1609:2245 1:N:0:ATCACG
AGCTTAAGAGGCACTAGACACAGCCAGCTTCTTCAGGTGATCCATGAACAC
```

Raw Reads

```
1 @HISEQ:574:C6VG2ANXX:2:1110:1400:2194 1:N:0:ATCACG
2 GGGGGATTCTCACTAGGTCTCAAGGTCTCTCACTCTCGGTAGTGTTCCCAG
3 +
4 CCCCCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
```

Line 1: Read identifier and is followed by a sequence that is

Unique for each read, platform dependent
Begins with a '@' character



Line 2: Raw sequence of nucleotides
Line 3: begins with a '+' character and is optionally followed by the same sequence identifier.

Line 4: Quality values for the sequence in Line 2

Raw Reads: Base Calling Quality

```
@HISEQ:574:C6VG2ANXX:2:1110:1400:2194 1:N:0:ATCACG
GGGGGATTCTCACTAGGTCTCAAGGTCTCTCACTCTCGGTAGTGTTCCCAG
+
CCCCCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
```

- A quality score (Q-score) is a prediction of the probability of an error in base calling.

- It serves as a compact way to communicate very small error probabilities
  - P = 10^(–Q/10)
  - Q = –10 log10(P)

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			

Raw Reads: Base Calling Quality

```
@HISEQ:574:C6VG2ANXX:2:1110:1400:2194 1:N:0:ATCACG
GGGGCATTCTCACTACGTCTCAAGGTCTCTCACTCTCGGTAGTGTTCCCAG
+
CCCCCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
```

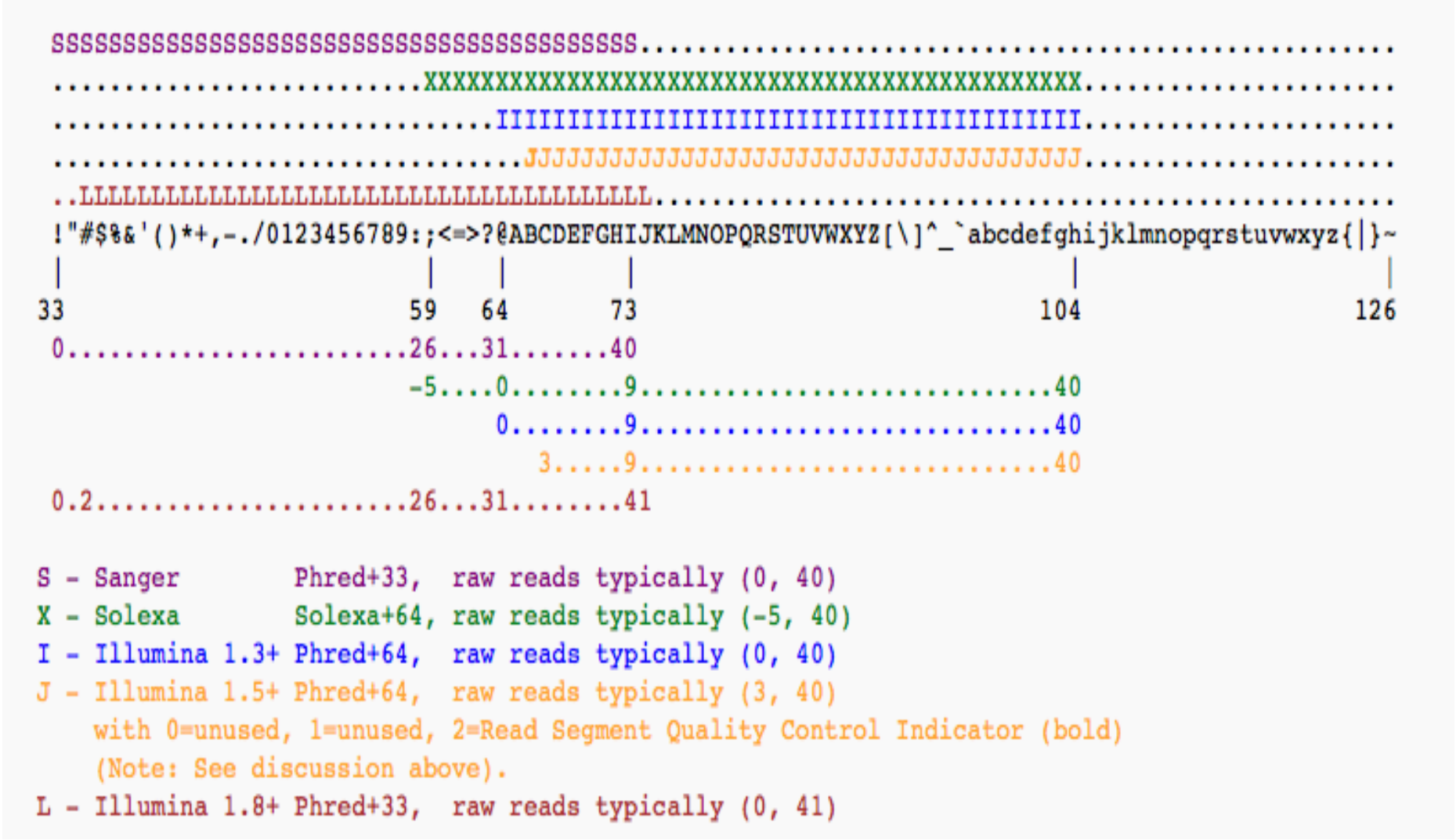
- Base: G
- Quality: C
- ASCII: 67
- Q: 34
- P: 0.00040

- Base: G
- Quality: G
- ASCII: 71
- Q: 38
- P: 0.00016

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			

Raw Reads: Base Calling Quality



Raw Reads: Base Calling Quality

Checking individual base calling manually is

rarely done

Fastq file quality control

- How many reads do I have ?
- Is that enough ?
- How good is the quality of the data ?
- Did I sequence what I wanted to sequence ?
- Is pre-processing needed ?
- ….

Example 1 – targeted sequencing

- Finding the genetic cause of a disease
  - ~ 6,000 genes
  - Illumina PE 126bp

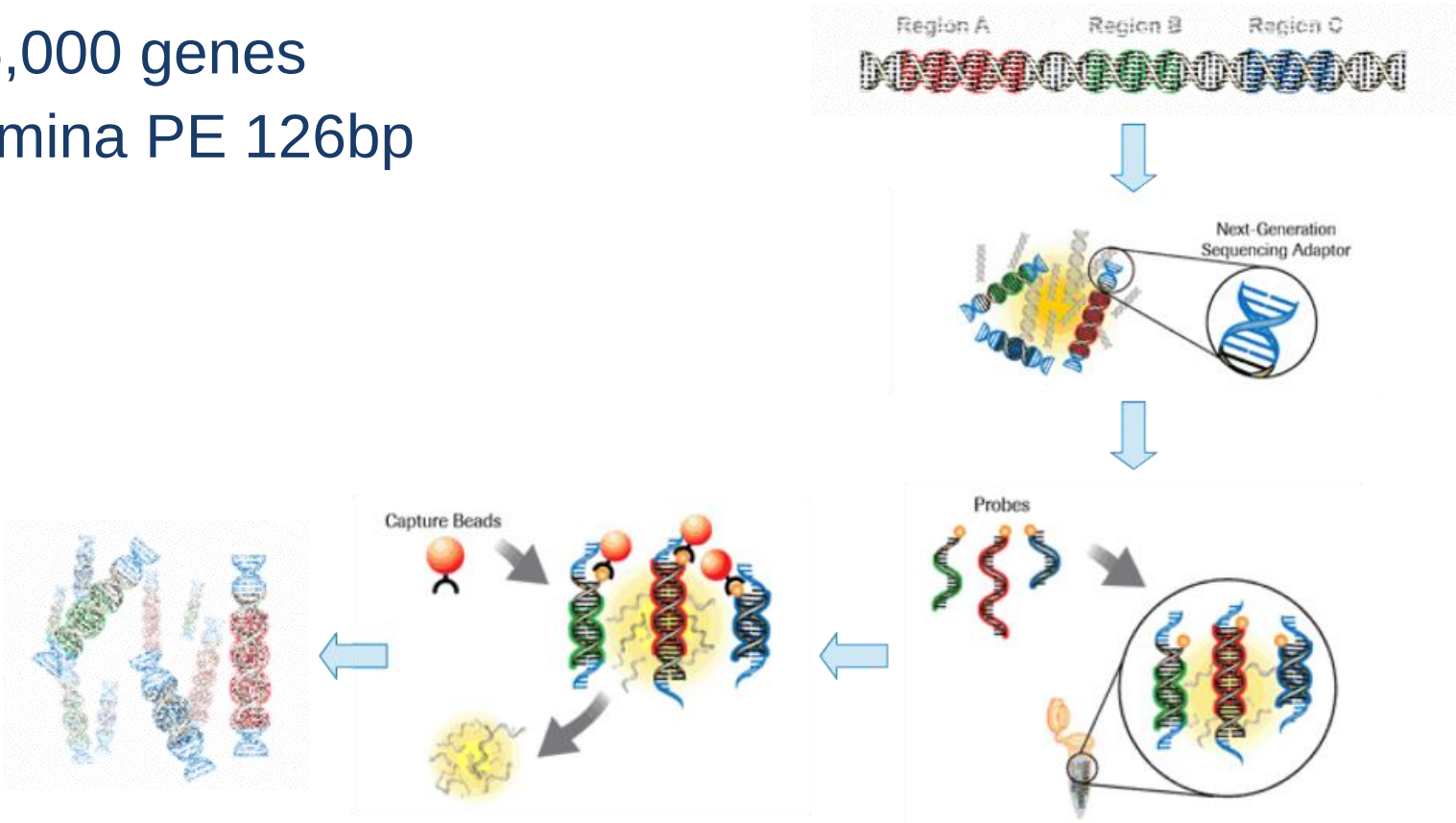
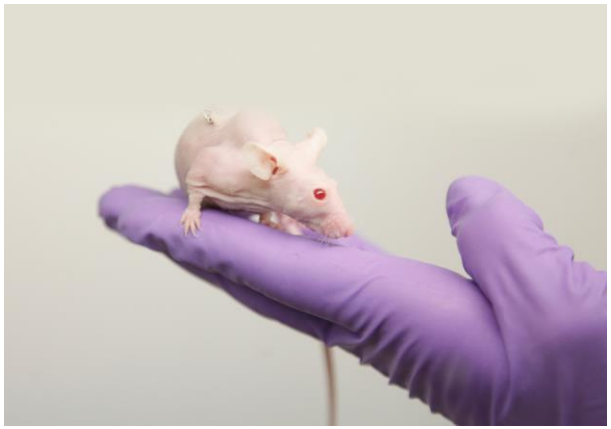


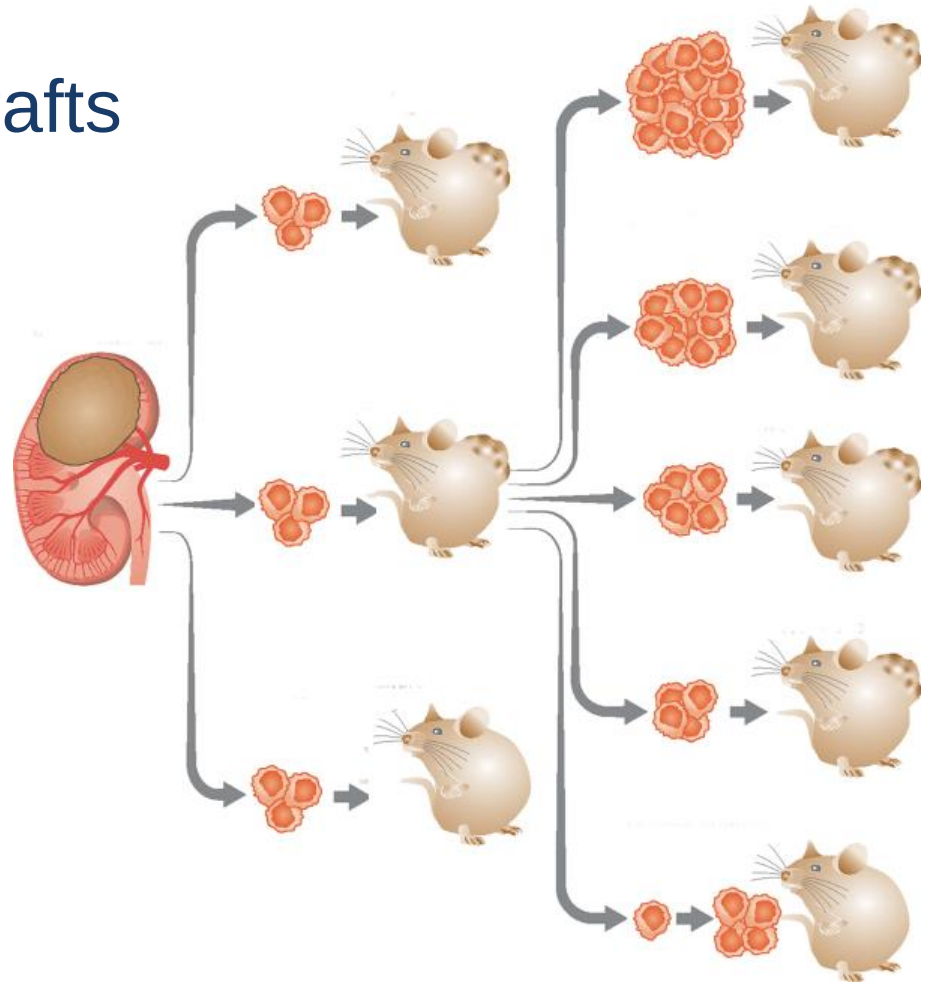
Figure adapted from Nimblegen

Example 2 – amplicon sequencing

- Fingerprinting of xenocrafts
  - 31 SNPs
  - Illumina PE 151bp



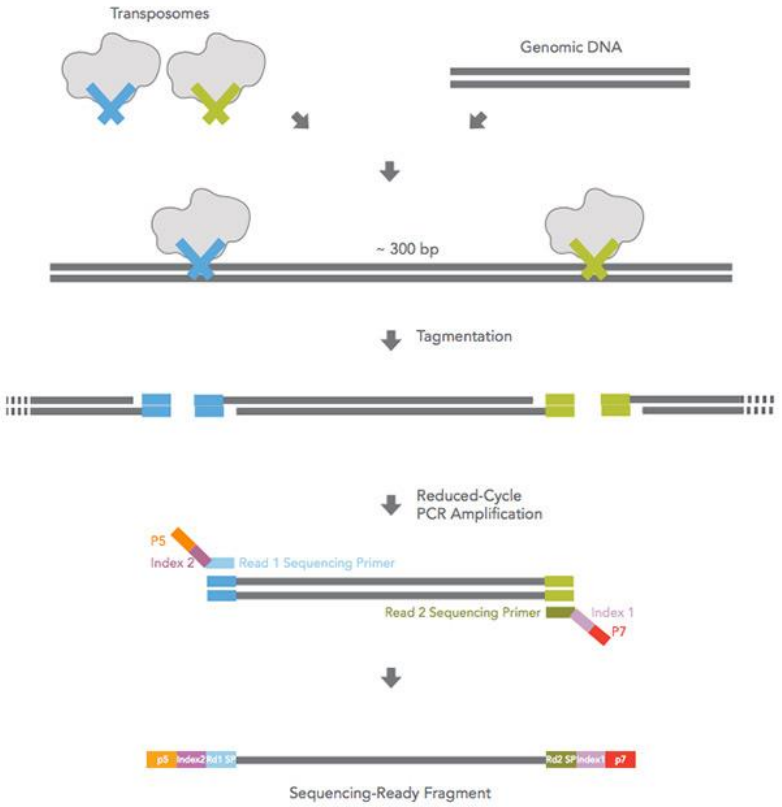
Trace
Patient Derived Tumor Xenograft Platform



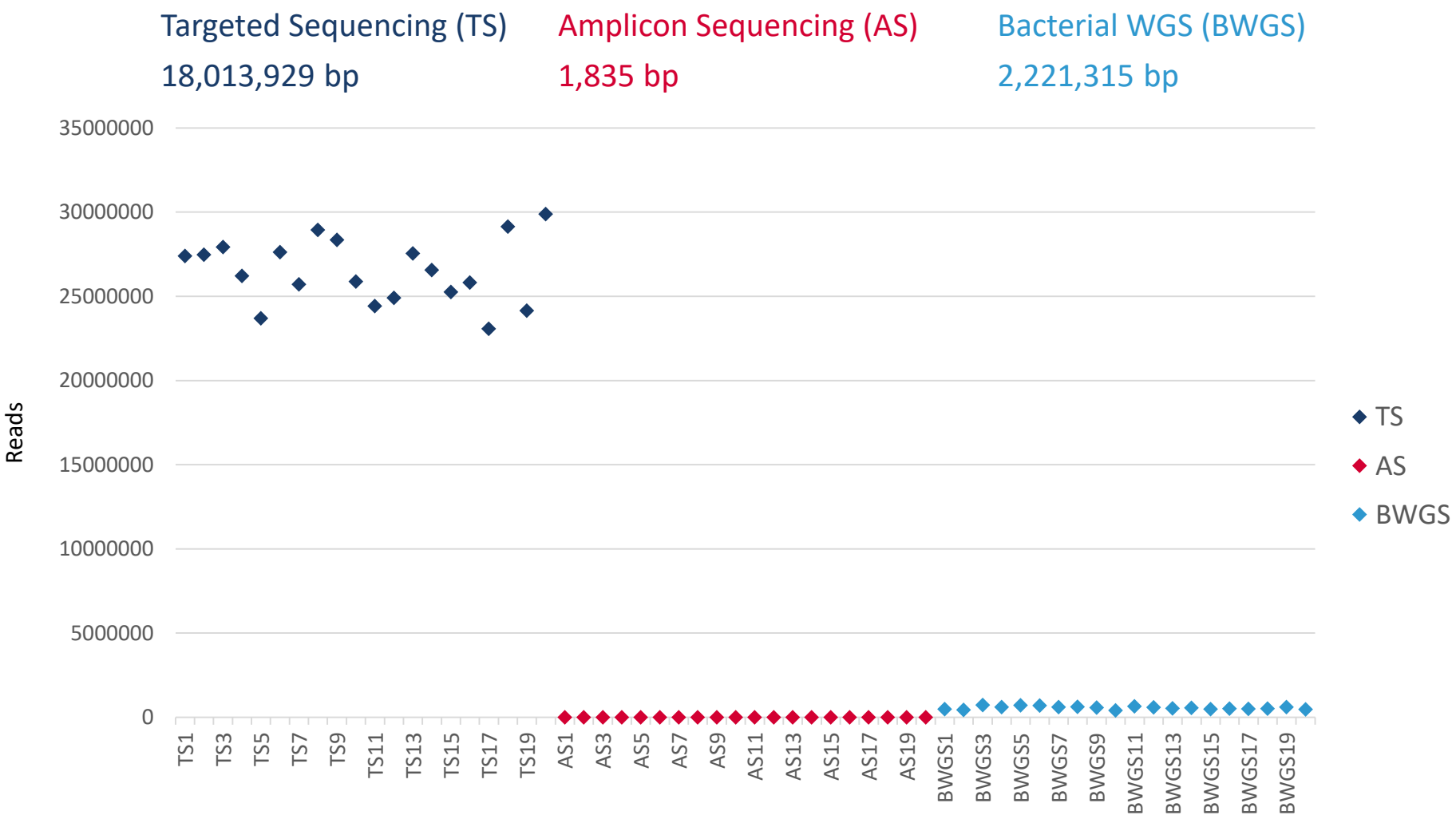
Trace platform, http://www.uzleuven-kuleuven.be/lki/trace/
Figure adapted from Peter Hohenstein, EMBO Molecular Medicine: 5 (1), 2013

Example 3 – whole genome sequencing

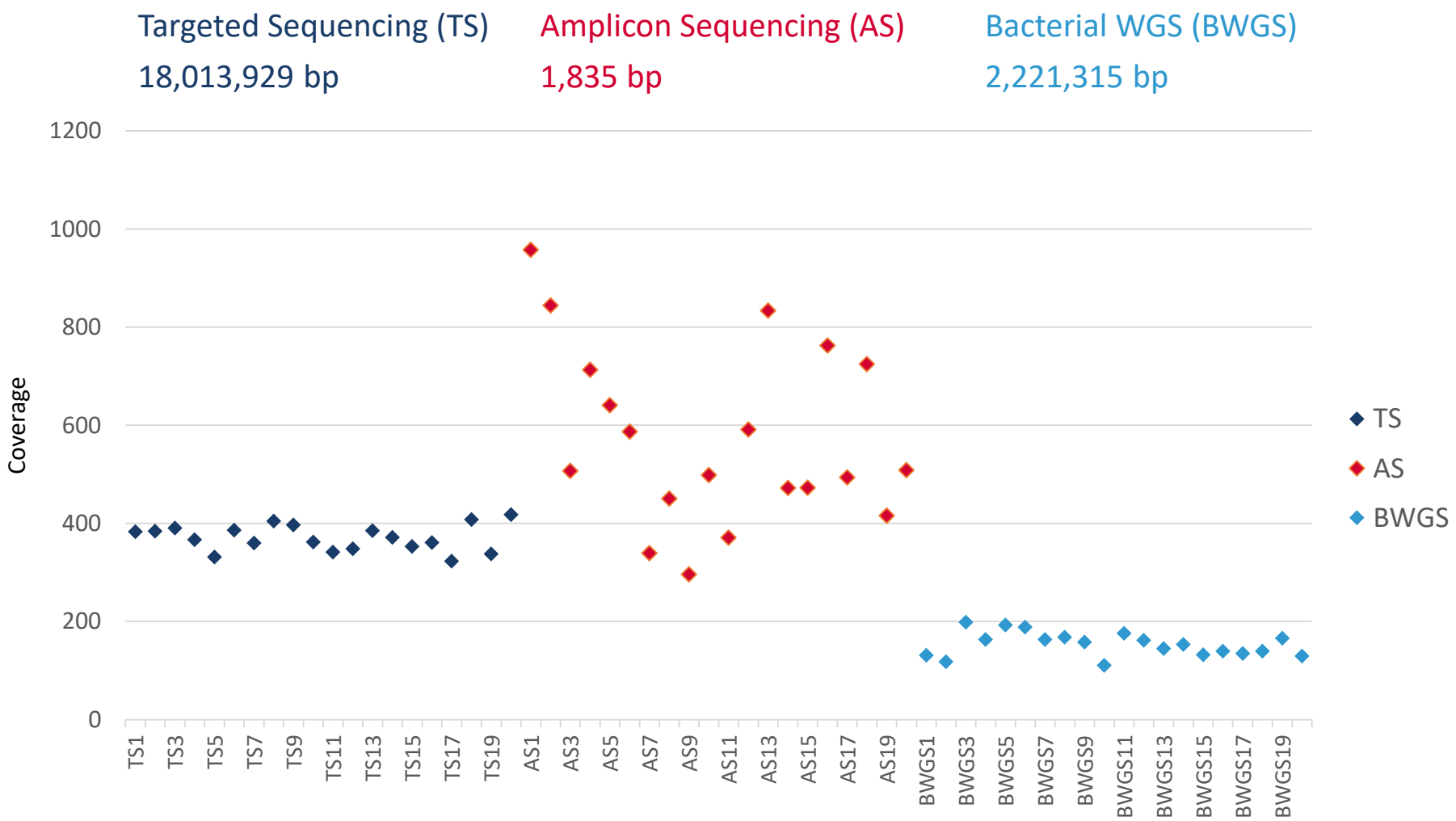
- Predicting bacterial resistance
  - Whole Genome Sequencing (WGS)
    - *Streptococcus pneumoniae*
    - *Mycobacterium tuberculosis*
  - Illumina PE 301 bp



Number of reads



Estimated coverage



Fastq files – QC

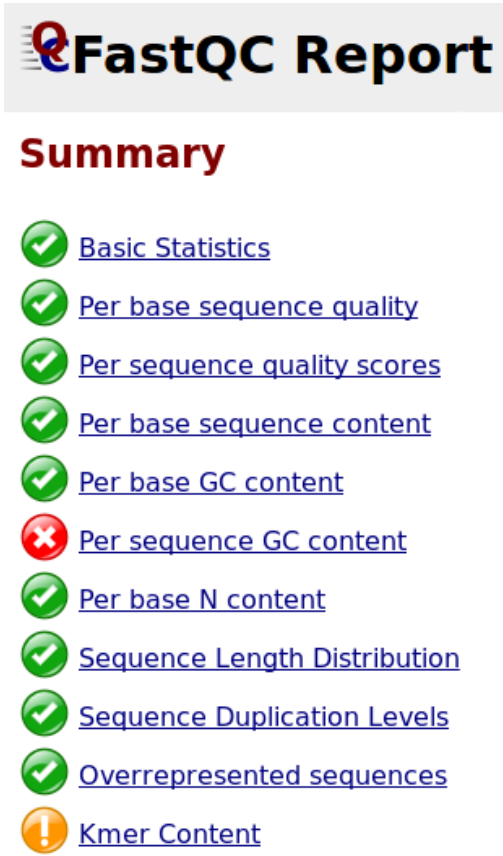
- FastQC
  - Check Phred quality scores
  - Check GC content
  - Check read content
  - …

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
https://multiqc.info

FastQC

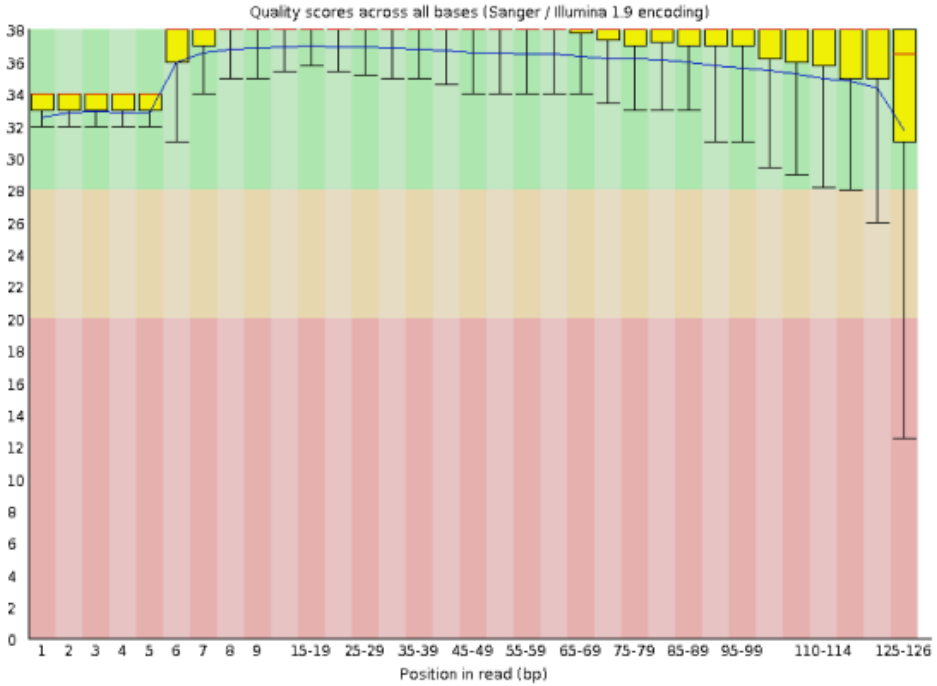
- Summary report

```
fastqc -o result sample.R1.fastq.gz
fastqc -o result sample.R2.fastq.gz
```

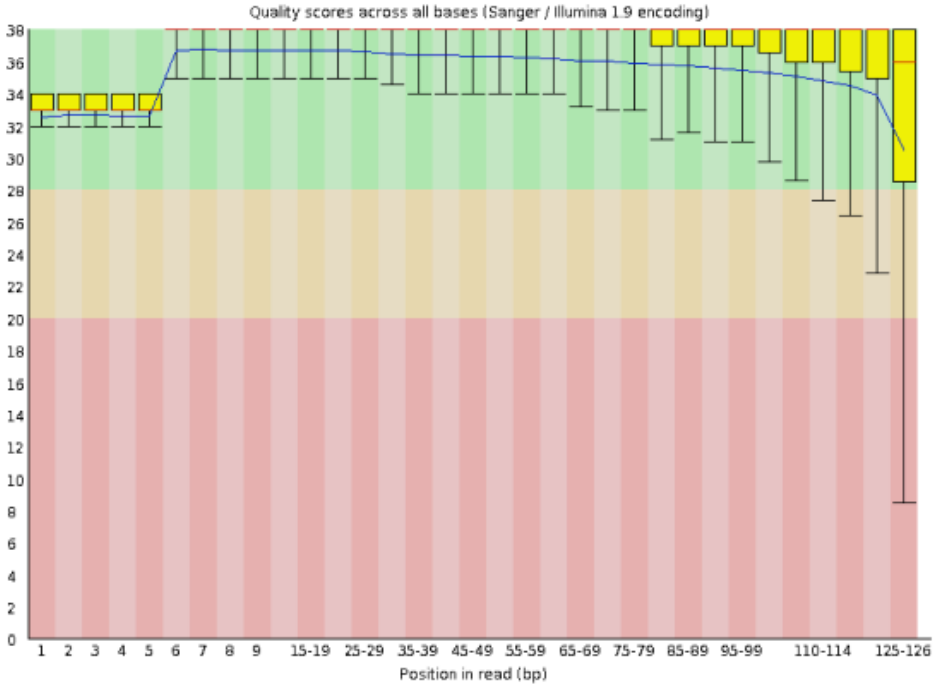


FastQC – Phred quality score by position

- Example 1 – targeted sequencing



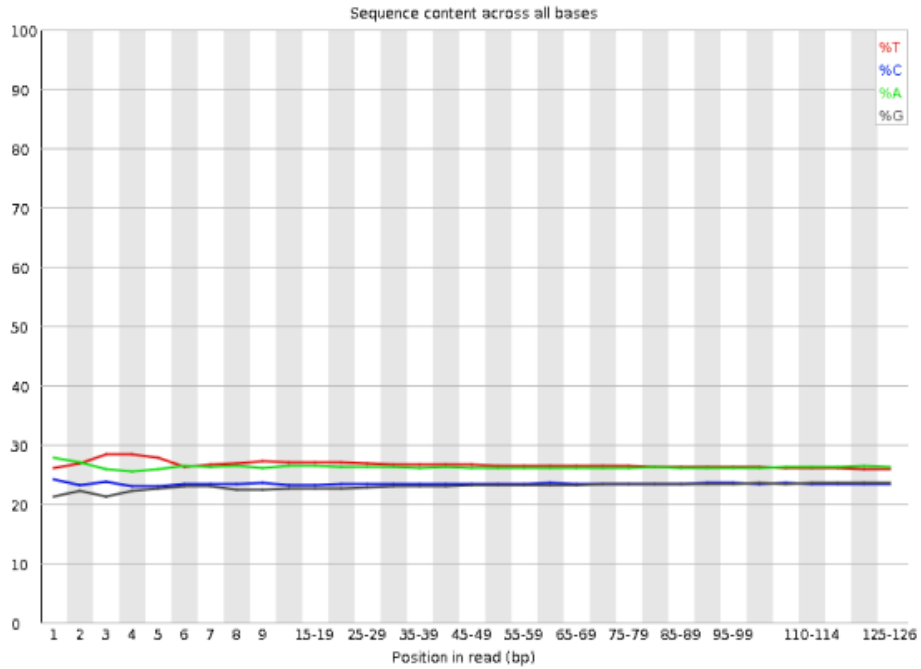
R1



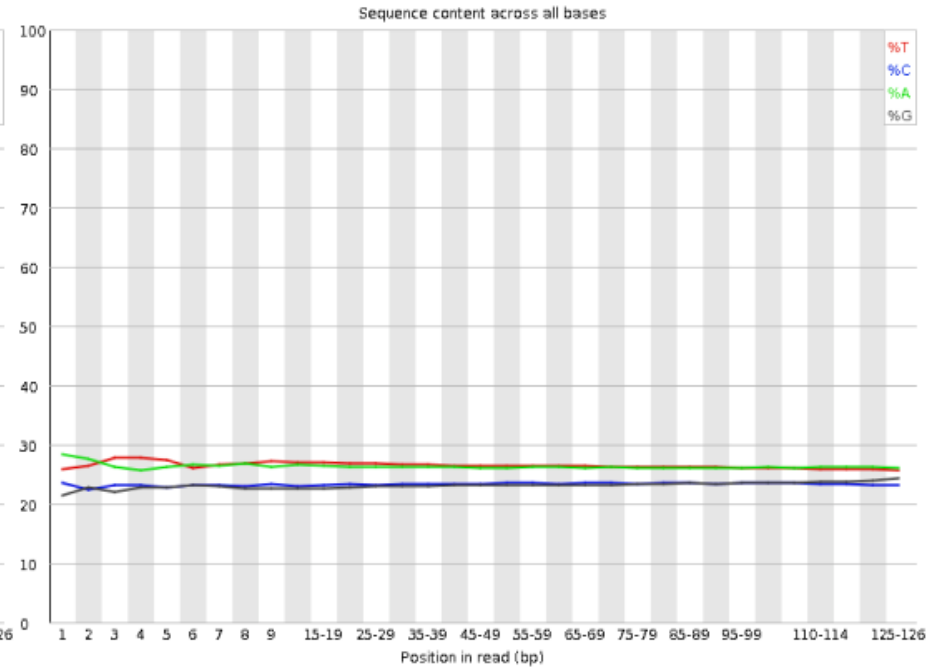
R2

FastQC – Base content by position

- Example 1 – targeted sequencing
  - G-C 25-26%
  - A-T 24-25%



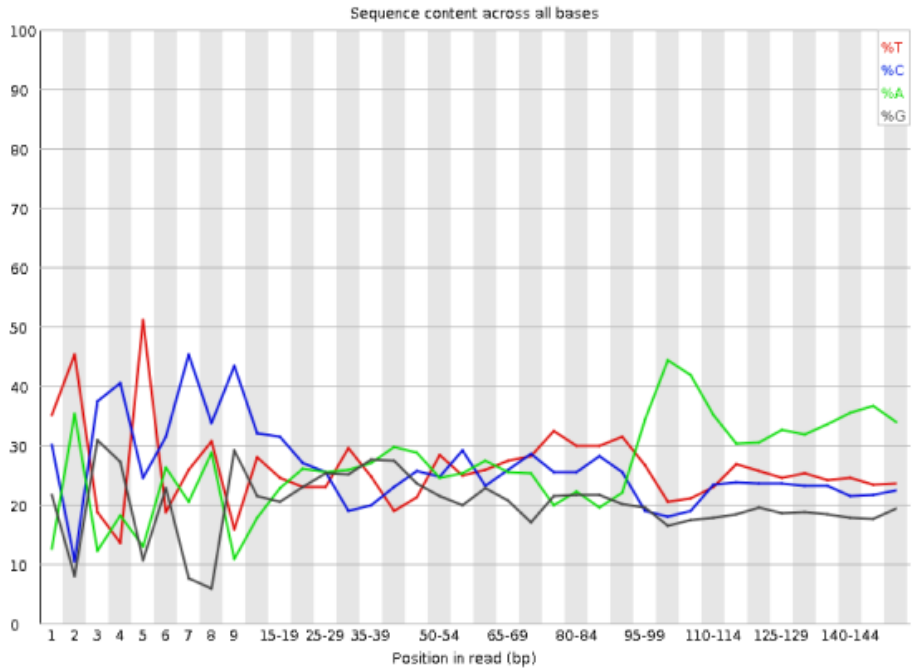
R1



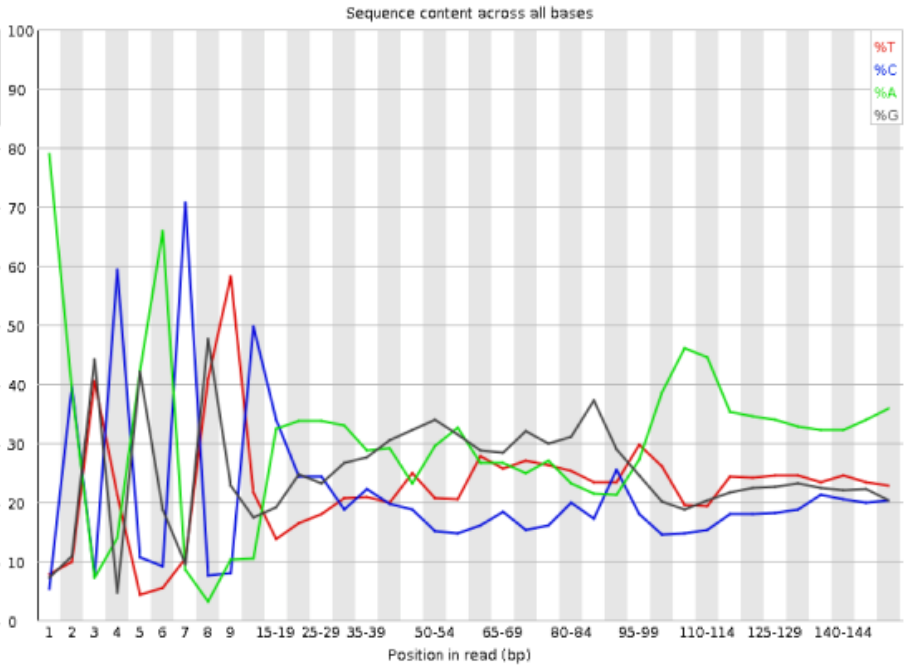
R2

FastQC – Base content by position

- Example 2 – amplicon sequencing
  - G-C 25-27%
  - A-T 24-24%



R1



R2

FastQC – Over-represented sequences

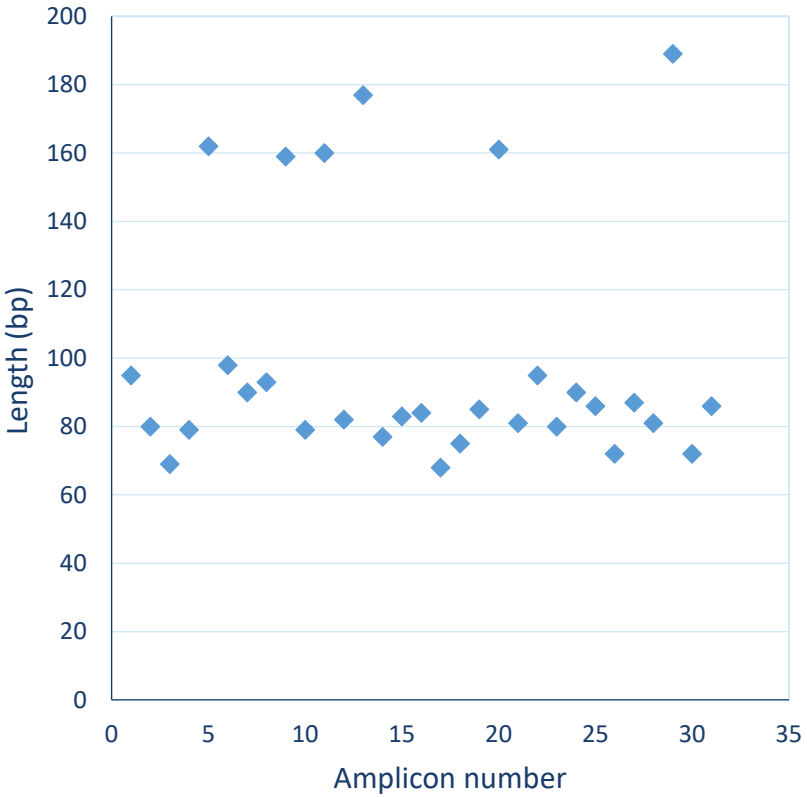
- Example 2 – amplicon sequencing

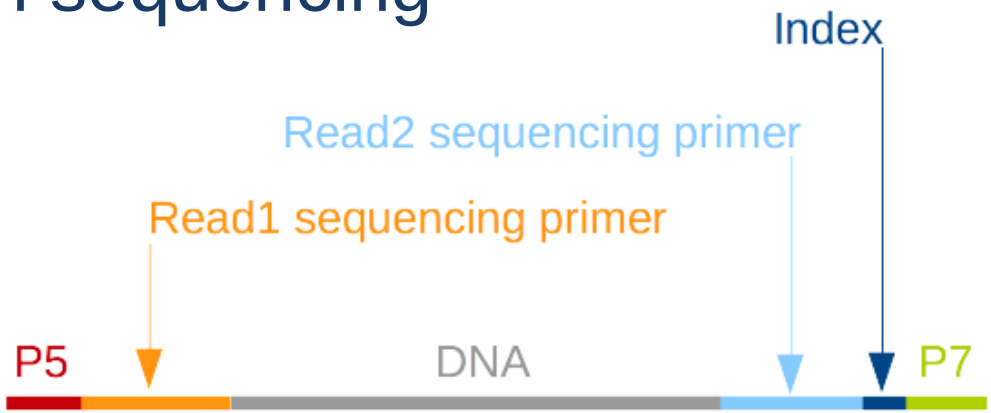
Overrepresented sequences

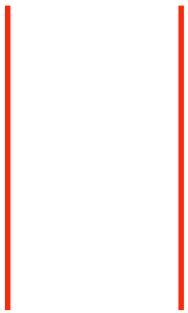
Sequence	Count	Percentage	Possible Source
GTGATCTCCAACTTTGACCTGACCGTCGCTTAGATCGGAAGAGCACACGT	92	5.2421652421652425	No Hit
CAGTGACACTAGTCTGCAACAAACGCCACTTAGATCGGAAGAGCACACGT	88	5.014245014245014	No Hit
TTCCTCACCTGCCCTGACCGTCGCTTAGATCGGAAGAGCACACGTCTGAA	52	2.9629629629629632	Illumina Multiplexing PCR Primer 2.01 (100% over 24bp)
TTCCTCACCTGCCCTGCAACAAACGCCACTTAGATCGGAAGAGCACACGT	49	2.792022792022792	No Hit
CTCACATCAGCCTGACCTGACCGTCGCTTAGATCGGAAGAGCACACGTCT	31	1.7663817663817662	Illumina Multiplexing PCR Primer 2.01 (100% over 21bp)
CAGCCTCTGCTCTCACCTGACCTGACCGTCGCTTAGATCGGAAGAGCACA	26	1.4814814814814816	No Hit
TTCCTCACCTGCCCTGCACTCAATCATCGTCTCCTAGATCGGAAGAGCAC	25	1.4245014245014245	No Hit
TTCCTCACCTGCCCTGCACTCTCCTCACCTCCACCCTGCACTCTCCTCAC	22	1.2535612535612535	No Hit
CTCACATCAGCCTGACACAACTTAGGACCACTTGAATAGAGAGCCTCAGT	22	1.2535612535612535	No Hit
TTCCTCACCTGCCCTGCACTCTCCTCACCTCCTCCCTGCACTCAATCATC	21	1.1965811965811968	No Hit
GACCAGAAGAACCTGACCTGACCGTCGCTTAGATCGGAAGAGCACACGTC	21	1.1965811965811968	No Hit
CTCACATCAGCCTGACCGTCGCTTAGATCGGAAGAGCACACGTCTGAACT	19	1.0826210826210827	Illumina Multiplexing PCR Primer 2.01 (100% over 26bp)
TCCTCCCTCTTGATGTGACCGTCGCTTAGATCGGAAGAGCACACGTCTGA	14	0.7977207977207977	Illumina Multiplexing PCR Primer 2.01 (100% over 23bp)
GTACAGCTGCACTGTGAAGATCGGAAGAGCACACGTCTGAACTCCAGTCA	14	0.7977207977207977	Illumina Multiplexing PCR Primer 2.01 (100% over 33bp)
TTCCTCACCTGCCCTGCACCATGAATGTTTTTTATAAAAAGGCTGTTGGC	12	0.6837606837606838	No Hit
AGGTAAGTGACAGTTTGCTCATGGGAAAGGAGATAGATCGGAAGAGCACA	12	0.6837606837606838	No Hit
GTGATCTCCAACTTTGACCGTCGCTTAGATCGGAAGAGCACACGTCTGAA	11	0.6267806267806267	Illumina Multiplexing PCR Primer 2.01 (100% over 24bp)
CAAGAGCTCAGAGGAGGAAGCTGTCAGAGATCGGAAGAGCACACGTCTGA	11	0.6267806267806267	Illumina Multiplexing PCR Primer 2.01 (100% over 23bp)
TTTGTACTTGTACCTGACCGTCGCTTAGATCGGAAGAGCACACGTCTGAA	11	0.6267806267806267	Illumina Multiplexing PCR Primer 2.01 (100% over 24bp)
CTCACATCAGCCTGACACTTTAAGTCGGGAGTCAGAAAGTACCCAAGGAG	9	0.5128205128205128	No Hit
TTGGTGTACATGTGTTTGTGTGTGTGTGGGGGGAAGTTGAGTAGATCG	9	0.5128205128205128	No Hit
TTCCTCACCTGCCCTGCACTCGATAATTCAATACATAATATTCAATAATT	9	0.5128205128205128	No Hit
GTACAGCTGGTACAAGAACCAGATCGGAAGAGCACACGTCTGAACTCCAG	9	0.5128205128205128	Illumina Multiplexing PCR Primer 2.01 (100% over 30bp)
TTCCTCACCTGACCTGACCGTCGCTTAGATCGGAAGAGCACACGTCTGAA	8	0.4558404558404558	Illumina Multiplexing PCR Primer 2.01 (100% over 24bp)
TTGGTGTACATGTGTTTGTGTGTGTGGGGGGAAGTTGAGTAGATCGGAAG	8	0.4558404558404558	No Hit
CATCTGCATGGTGATCCTGGGCTCTGTAGTGGTGGCTGCAAAGAGGTGCT	8	0.4558404558404558	No Hit
CATTTCCATTGCCAACCGAGTCCATTGTGCACAGTATGAAGACAGCACAT	8	0.4558404558404558	No Hit
GATGTTCAGGCATTCCCAGTTAGGTGAGTAAACCCTTGATCAGTCACTAT	7	0.39886039886039887	No Hit
AGGTAAGTGACAGTTTGCTCAGGGAAAGTGTGAGATTGGATTCTTTAAAC	7	0.39886039886039887	No Hit
TGGCCTTGACAAACAGATCGGAAGAGCACACGTCTGAACTCCAGTCACAG	7	0.39886039886039887	TruSeq Adapter, Index 2 (97% over 35bp)
TTCCTCACCTGCCCTGCACTCTCCTCACCTCCACTCTCATCACCTCCACC	7	0.39886039886039887	No Hit
ACACTGGGCTAGACACTCGTATGGTTGTATGGGGTTTCTCTTCTTAGAGA	7	0.39886039886039887	No Hit
TTTGTACTTGTACCTGGGCGCATCGTTCATTTTTCAGTTGTGGATAGCAC	7	0.39886039886039887	No Hit
AAGAGCCTGCCTGACCGTCGCTTAGATCGGAAGAGCACACGTCTGAACTC	6	0.3418803418803419	Illumina Multiplexing PCR Primer 2.01 (100% over 27bp)

FastQC – Over-represented sequences

- Example 2 – amplicon sequencing







Contamination check

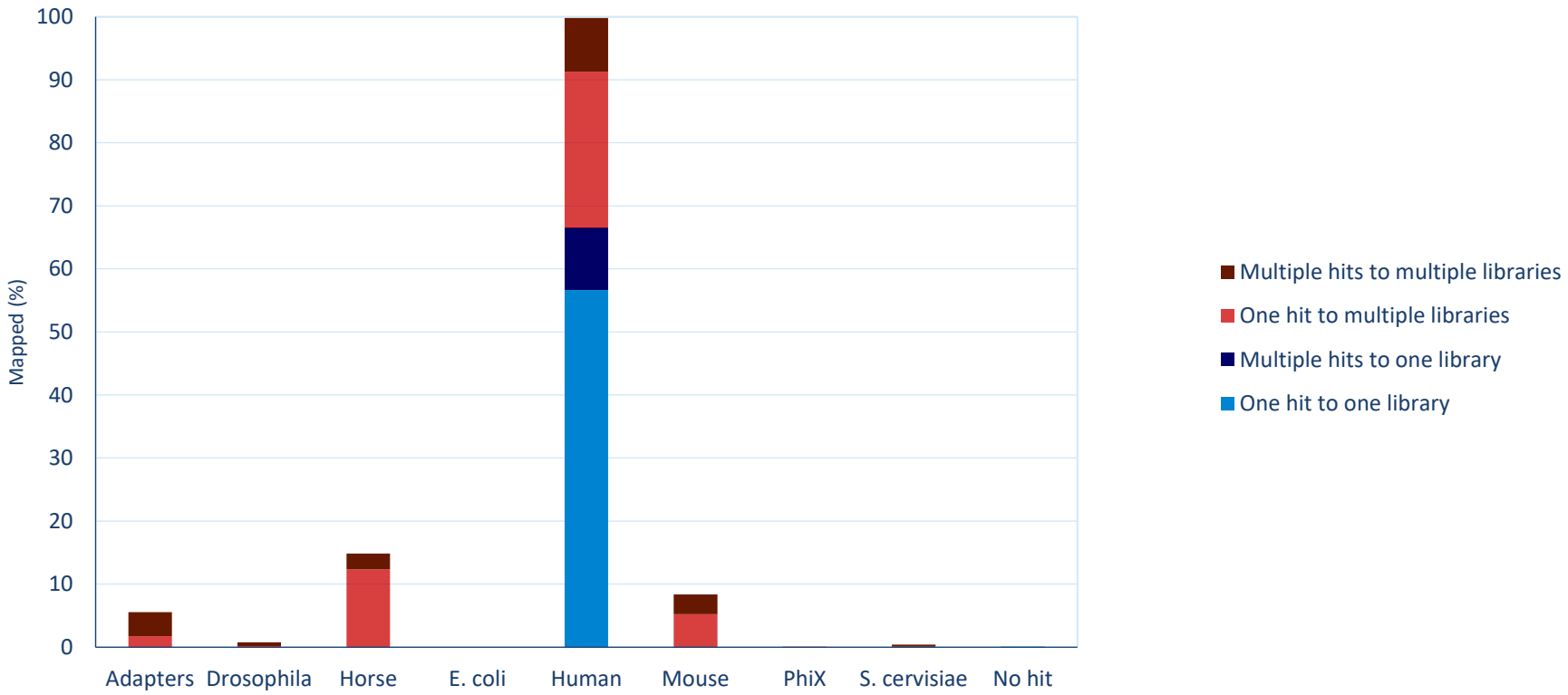
- FastQ Screen
  - Compare reads to various libraries
  - Any library can be searched against
  - Output proportion of reads with
    - One hit to one library
    - Multiple hits to one library
    - One hit to multiple libraries
    - Multiple hits to multiple libraries

```
fastq_screen --subset 100000 --conf fastq_screen.conf --aligner
bowtie2 --outdir result --nohits sample.R1.fastq.gz
```

https://www.bioinformatics.babraham.ac.uk/projects/fastq_screen/

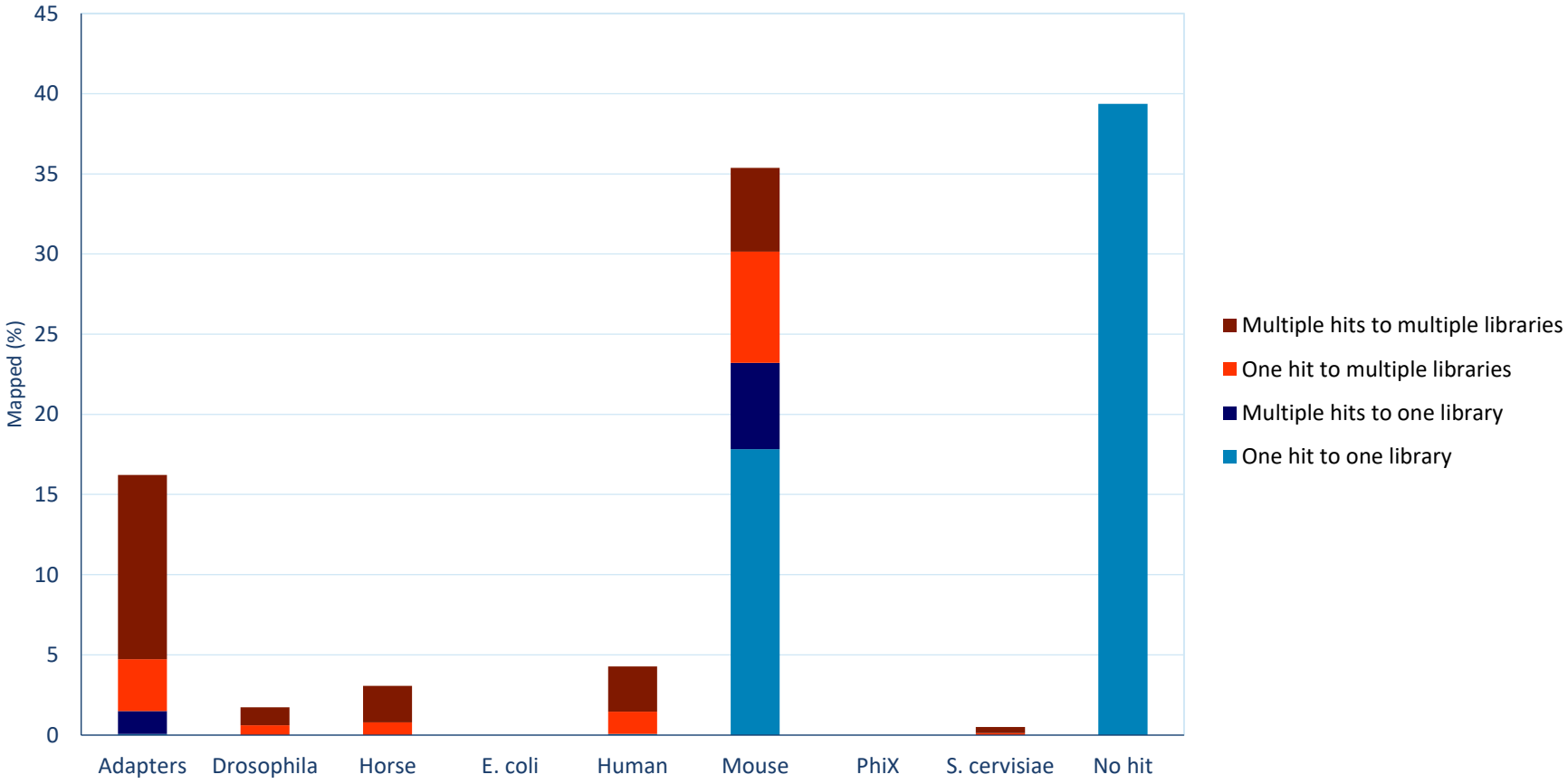
Contamination check – FastQ Screen

- Example 1 – targeted sequencing



Contamination check – FastQ Screen

- Example 2 – amplicon sequencing



Contamination check – kraken2

- Taxonomic sequence classification system
  - Build (custom) database
  - Compare k-mer from reads to database

```
kraken2-build --download-taxonomy --db database
kraken2-build --download-library library --db database
kraken2-build --build --db database --minimizer-spaces 0
kraken2 --db database --paired sample.R1.fastq.gz
sample.R2.fastq.gz –report kraken2Report.txt --use-names >
kraken2.output.txt
```

https://ccb.jhu.edu/software/kraken2/

Contamination check – kraken2

- Example 3 – bacterial WGS

Fragments covered by clade (%)	Fragments covered by clade	Fragments assigned to taxon	Rank code	NCBI taxonomic ID	scientific name
0.04	243	243	U	0	unclassified
99.96	632527	0	R	1	root
99.96	632527	0	R1	131567	cellular organisms
99.96	632491	121	D	2	Bacteria
99.55	629954	1	D1	1783272	Terrabacteria group
99.53	629777	0	P	201174	Actinobacteria
99.53	629777	17	C	1760	Actinobacteria
99.52	629754	20	O	85007	Corynebacteriales
99.52	629733	74	F	1762	Mycobacteriaceae
99.51	629657	5911	G	1763	Mycobacterium
98.56	623663	5295	G1	77643	Mycobacterium tuberculosis complex
97.69	618141	613920	S	1773	Mycobacterium tuberculosis
0.04	227	78	S	78331	Mycobacterium canettii
0.01	37	35	S	1768	Mycobacterium kansasii
0.03	176	0	P	1239	Firmicutes
0.03	176	0	C	91061	Bacilli
0.03	172	0	O	186826	Lactobacillales
0.03	172	0	F	1300	Streptococcaceae
0.03	172	7	G	1301	Streptococcus
0.03	161	155	S	1313	Streptococcus pneumoniae
0	2	1	S	28037	Streptococcus mitis
0	1	0	S	257758	Streptococcus pseudopneumoniae

Pre-processing ?

- Process fastq files prior to further analysis
  - Remove reads from other species
  - Trim adapters
  - Clip low quality bases
  - Merge overlapping reads from same fragment
  - ...

Adapter clipping & trimming

- FastqMcf
  - Detect & remove sequencing adapters and primers
  - Detect & clip poor quality at the ends of reads
  - Remove low complexity reads
  - Detect & remove Ns from ends of reads
  - Keep PE reads in right order

```
fastq-mcf -H -X  -o sample_filtered.R1.fastq.gz -o
/sample_filtered.R2.fastq.gz adapters.fa sample.R1.fastq
sample.R2.fastq
```

https://github.com/ExpressionAnalysis/ea-utils/blob/wiki/FastqMcf.md

Adapter clipping & trimming – FastqMcf

- Example 2 – amplicon sequencing
  - Input
    - 2 fastq files of 1,834 reads each
  - Outputs
    - 2 fastq files of 1,801 reads each
    - List of adapter found

Adapter TruSeq_Adapter,_Index_1 : counted **1038** at the 'end' of 'sample.R1.fastq'

…

Adapter Illumina_Single_End_Sequencing_Primer_3p : counted **1046** at the 'end' of 'sample.R2.fastq'

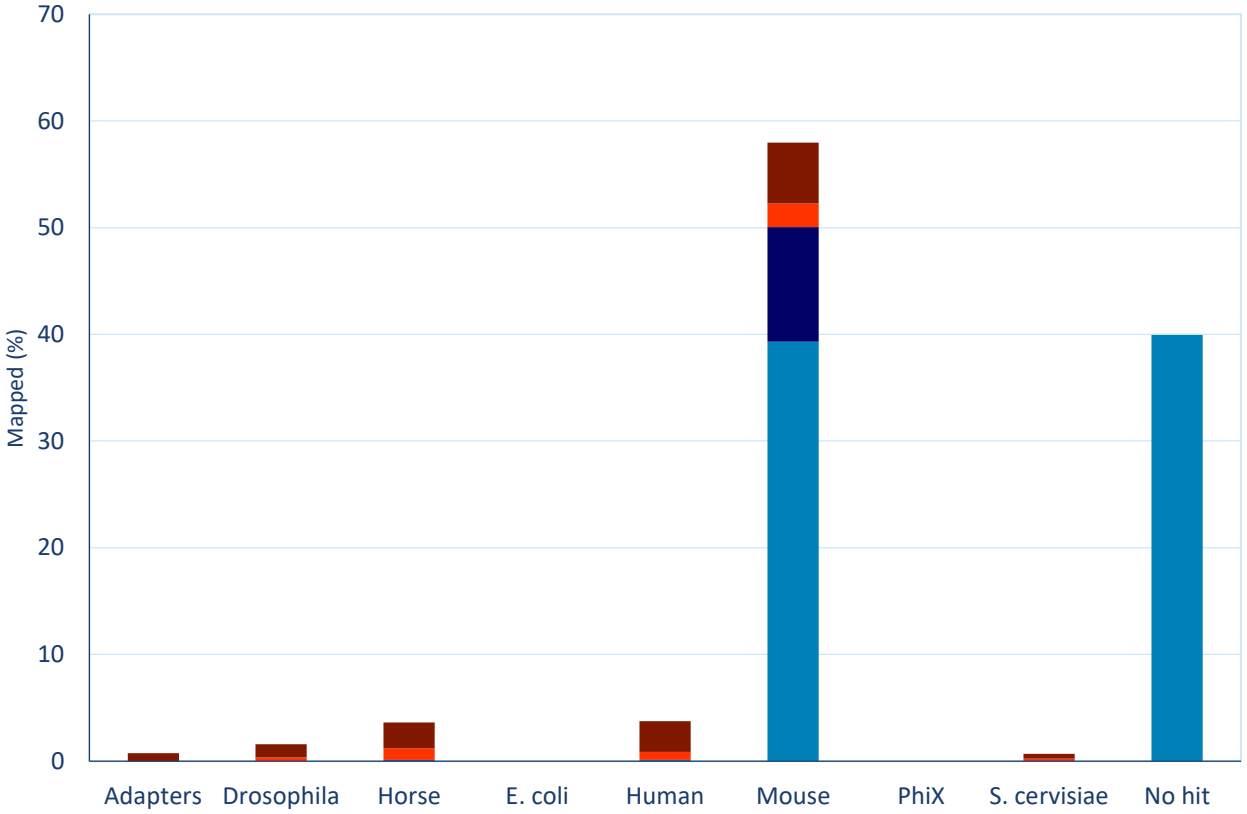
…

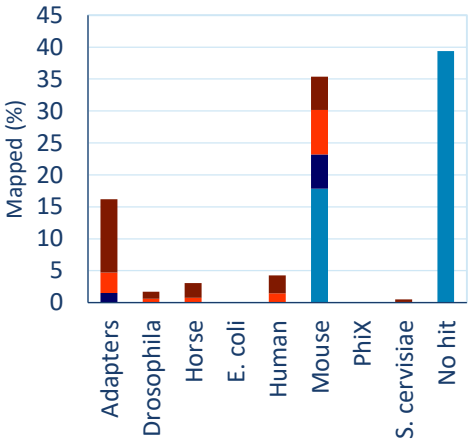
Total reads: 1801

Too short after clip: 33

Adapter clipping & trimming – FastqMcf

- Example 2 – amplicon sequencing





Multiple hits to multiple libraries
One hit to multiple libraries
Multiple hits to one library
One hit to one library

Read selection – Kraken 2 & seqtk

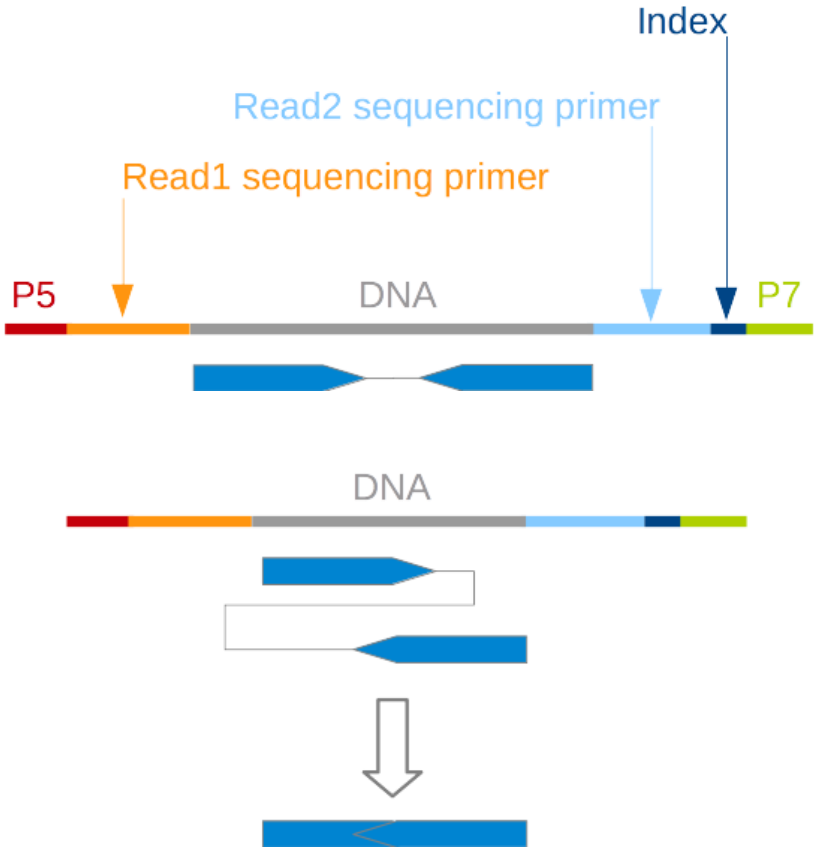
- Select reads from sequenced organism
  - Kraken2 outpout
  - Seqtk: toolkit for processing sequences in FASTA/Q formats

```
grep ”organism” kraken2.output.txt | cut –f2 > reads.list
seqtk subseq sample.R1.fastq.gz reads.list | gzip - >
sample.selected.R1.fastq.gz
seqtk subseq sample.R2.fastq.gz reads.list | gzip - >
sample.selected.R2.fastq.gz
```

https://ccb.jhu.edu/software/kraken2/
https://github.com/lh3/seqtk

Merge overlapping reads – FLASH2

- FLASH (Fast Length Adjustment of SHort reads)
  - Merge paired-end reads
  - Keep DNA fragment only

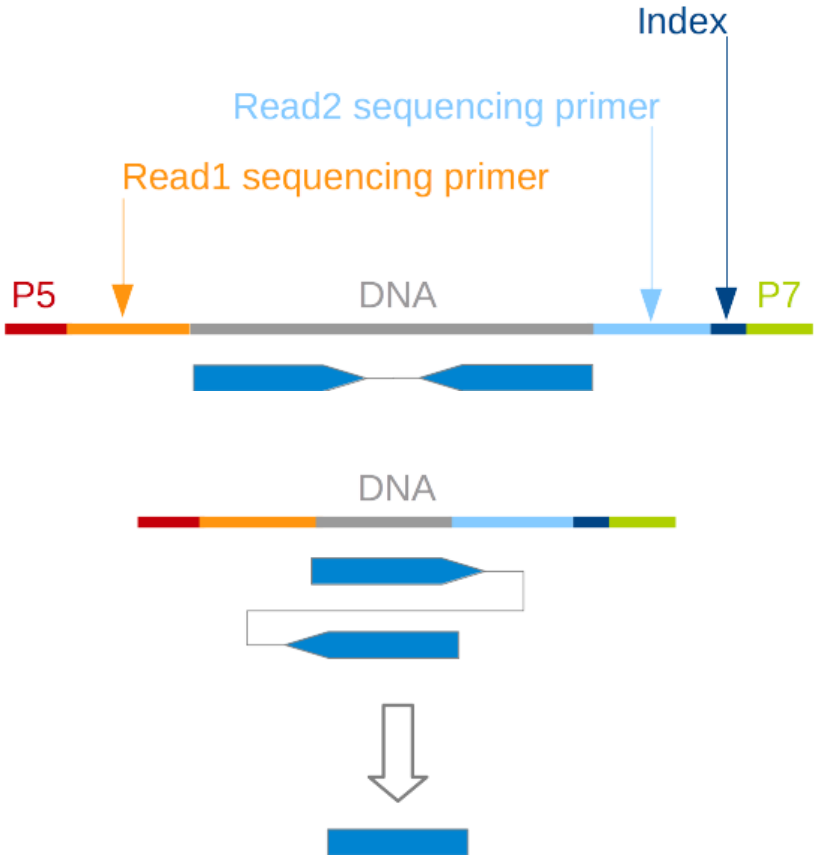



```
flash2 --max-overlap=250 --min-overlap=20 --allow-outies -d result -o
sample.flashed sample.R1.fastq.gz sample.R2.fastq.gz > flash.log
```

https://github.com/dstreett/FLASH2

Merge overlapping reads – FLASH2

- FLASH (Fast Length Adjustment of SHort reads)
  - Merge paired-end reads
  - Keep DNA fragment only




```
flash2 --max-overlap=250 --min-overlap=20 --allow-outies -d result -o
sample.flashed sample.R1.fastq.gz sample.R2.fastq.gz > flash.log
```

https://github.com/dstreett/FLASH2

Questions?

GENOMICS CORELEUVEN

Álvaro Cortés Calabuig
Erika Souche