Understanding NGS raw data: FASTQ format, quality checking

Erika Souche

Outline

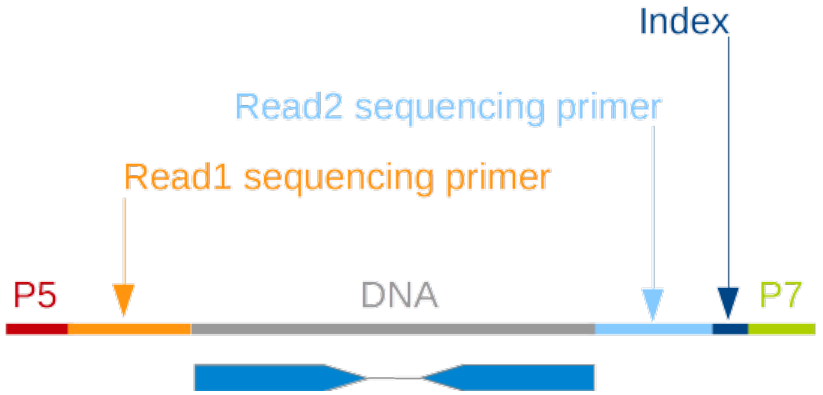
- Definitions
- FASTQ format
- Quality control
- Pre-processing

*Focus on Illumina short read sequencing

Outline

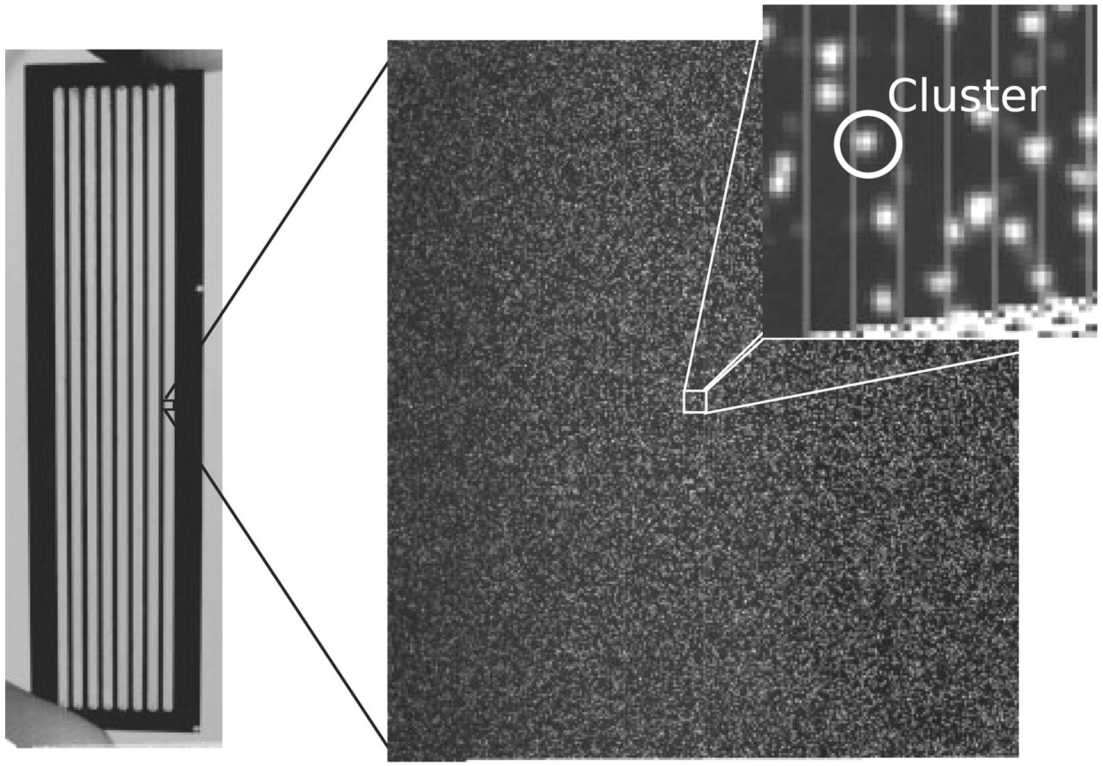
- Definitions
- FASTQ format
- Quality control
- Pre-processing

Library



- DNA fragment
- Index/barcode
- Single End (SE) sequencing
- Paired End (PE) sequencing
- Insert size (=DNA fragment size)

Illumina flowcell

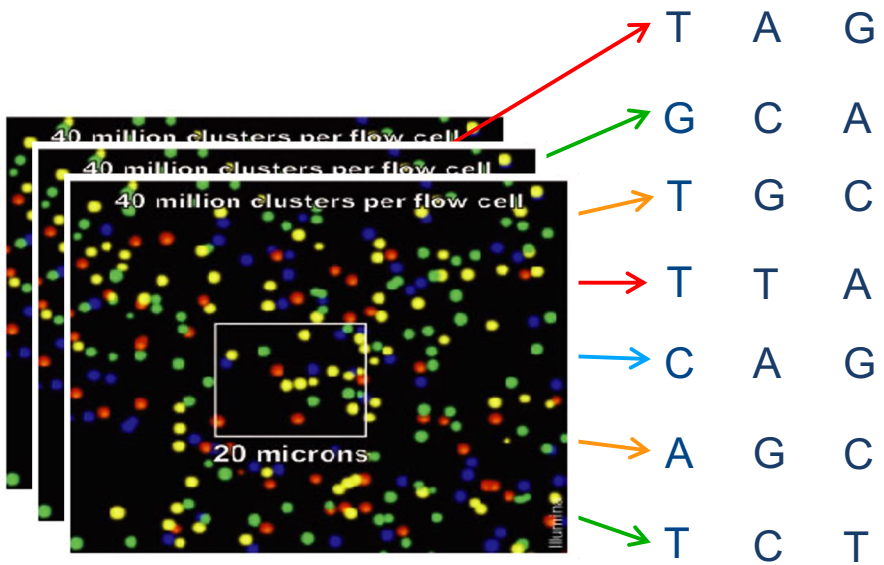


- Cluster

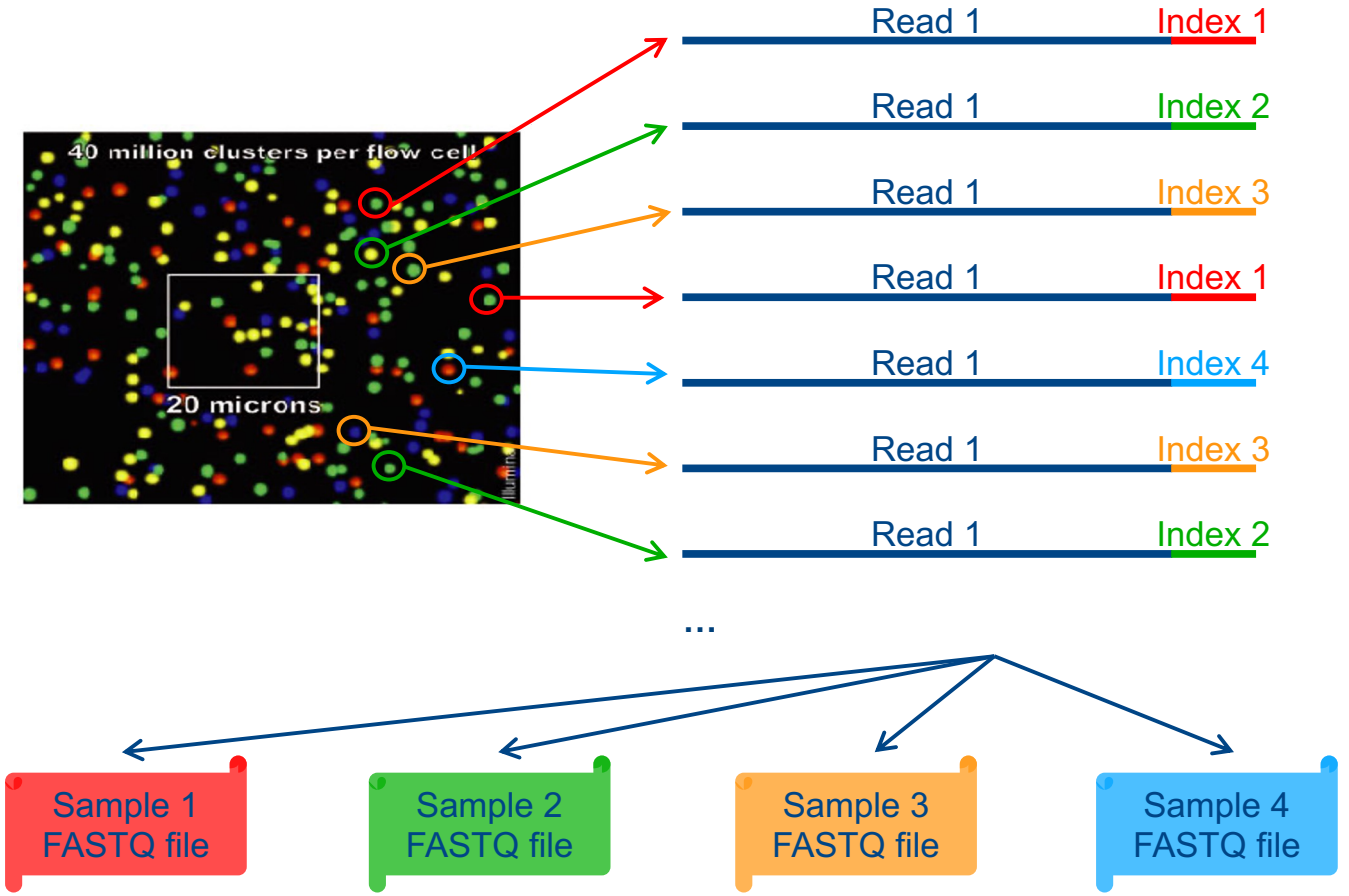
  Bright spot on an image

  Represents 1,000s of copies of the same DNA fragment

Sequencing


40 million clusters per flow cell
40 million clusters per flow cell
40 million clusters per flow cell
20 microns
T A G
G C A
T G C
T T A
C A G
A G C
T C T
...

Demultiplexing



Outline

- Definitions
- FASTQ format
- Quality control
- Pre-processing

FASTA file format

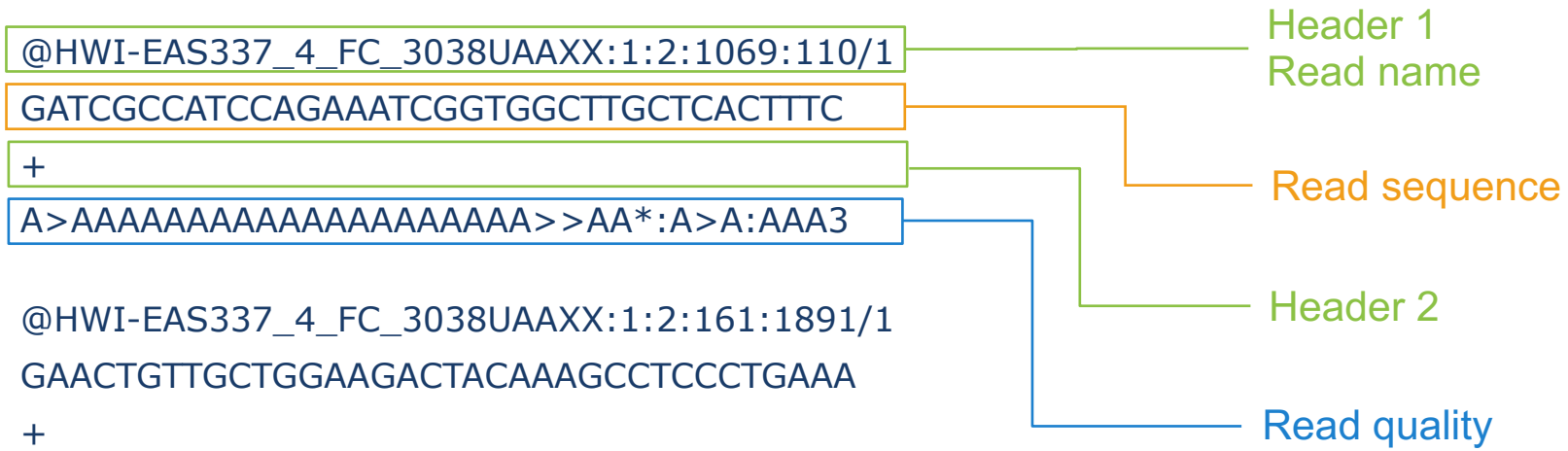
```
>1 dna:chromosome chromosome:GRCh38:1:1:248956422:1 REF
```

Header 1
Contig name

```
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
...
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCTAACCCAACCCTAACCCTAACCCTAACCCTAACCCTAA
CCCTAACCCCTAACCCTAACCCTAACCCTAACCCTAACCTAACCCTAACCCTAACCCTAA
CCCTAACCCTAACCCTAACCCTAACCCTAACCCCTAACCCTAACCCTAAACCCTAAACCC
TAACCCTAACCCTAACCCTAACCCTAACCCCAACCCCAACCCCAACCCCAACCCCAACCC
CAACCCTAACCCCTAACCCTAACCCTAACCCTACCCTAACCCTAACCCTAACCCTAACCC
TAACCCTAACCCCTAACCCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAAC
CCCTAACCCTAACCCTAACCCTAACCCTCGCGGTACCCTCAGCCGGCCCGCCCGCCCGGG
TCTGACCTGAGGAGAACTGTGCTCCGCCTTCAGAGTACCACCGAAATCTGTGCAGAGGAC
AACGCAGCTCCGCCCTCGCGGTGCTCTCCGGGTCTGTGCTGAGGAGAACGCAACTCCGCC
GTTGCAAAGGCGCGCCGCGCCGGCGCAGGCGCAGAGAGGCGCGCCGCGCCGGCGCAGGCG
CAGAGAGGCGCGCCGCGCCGGCGCAGGCGCAGAGAGGCGCGCCGCGCCGGCGCAGGCGCA
GAGAGGCGCGCCGCGCCGGCGCAGGCGCAGAGAGGCGCGCCGCGCCGGCGCAGGCGCAGA
CACATGCTAGCGCGTCGGGGTGGAGGCGTGGCGCAGGCGCAGAGAGGCGCGCCGCGCCGG
CGCAGGCGCAGAGACACATGCTACCGCGTCCAGGGGTGGAGGCGTGGCGCAGGCGCAGAG
AGGCGCACCGCGCCGGCGCAGGCGCAGAGACACATGCTAGCGCGTCCAGGGGTGGAGGCG
TGGCGCAGGCGCAGAGACGCAAGCCTACGGGCGGGGGTTGGGGGGGCGTGTGTTGCAGGA
...
```

Contig sequence

FASTQ file format



Read sequence & quality

@HWI-EAS337_4_FC_3038UAAXX:1:2:1069:110/1

GATCGCCATCCAGAAATCGGTGGCTTGCTCACTTTC

Read sequence

+

A>AAAAAAAAAAAAAAAAAAA>>AA*:A>A:AAA3

Read quality

Probability of incorrect base call (P)

Phred Q =−10 log10 P

1 ASCII character per base

Phred Q = dec(ASCII) – 33

Regular ASCII Chart (character codes 0 - 127)

000		(nul)	016	►	(dle)	032	sp	048	0	064	@	080	P	096	`	112	p
001	☺	(soh)	017	◄	(dc1)	033	!	049	1	065	A	081	Q	097	a	113	q
002	☻	(stx)	018	↕	(dc2)	034	"	050	2	066	B	082	R	098	b	114	r
003	♥	(etx)	019	‼	(dc3)	035	#	051	3	067	C	083	S	099	c	115	s
004	♦	(eot)	020	¶	(dc4)	036	$	052	4	068	D	084	T	100	d	116	t
005	♣	(enq)	021	§	(nak)	037	%	053	5	069	E	085	U	101	e	117	u
006	♠	(ack)	022	▬	(syn)	038	&	054	6	070	F	086	V	102	f	118	v
007	•	(bel)	023	↨	(etb)	039	'	055	7	071	G	087	W	103	g	119	w
008	◘	(bs)	024	↑	(can)	040	(	056	8	072	H	088	X	104	h	120	x
009		(tab)	025	↓	(em)	041	)	057	9	073	I	089	Y	105	i	121	y
010		(lf)	026		(eof)	042	*	058	:	074	J	090	Z	106	j	122	z
011	♂	(vt)	027	←	(esc)	043	+	059	;	075	K	091	[	107	k	123	{
012	♀	(np)	028	∟	(fs)	044	,	060	<	076	L	092	\	108	l	124	|
013		(cr)	029	↔	(gs)	045	-	061	=	077	M	093	]	109	m	125	}
014	♫	(so)	030	▲	(rs)	046	.	062	>	078	N	094	^	110	n	126	~
015	☼	(si)	031	▼	(us)	047	/	063	?	079	O	095	_	111	o	127	⌂

Read sequence & quality

@HWI-EAS337_4_FC_3038UAAXX:1:2:1069:110/1

GATCGCCATCCAGAAATCGGTGGCTTGCTCACTTTC

+

A>AAAAAAAAAAAAAAAAAAA>>AA*:A>A:AAA3

Read sequence

Read quality

Probability of incorrect base call (P)

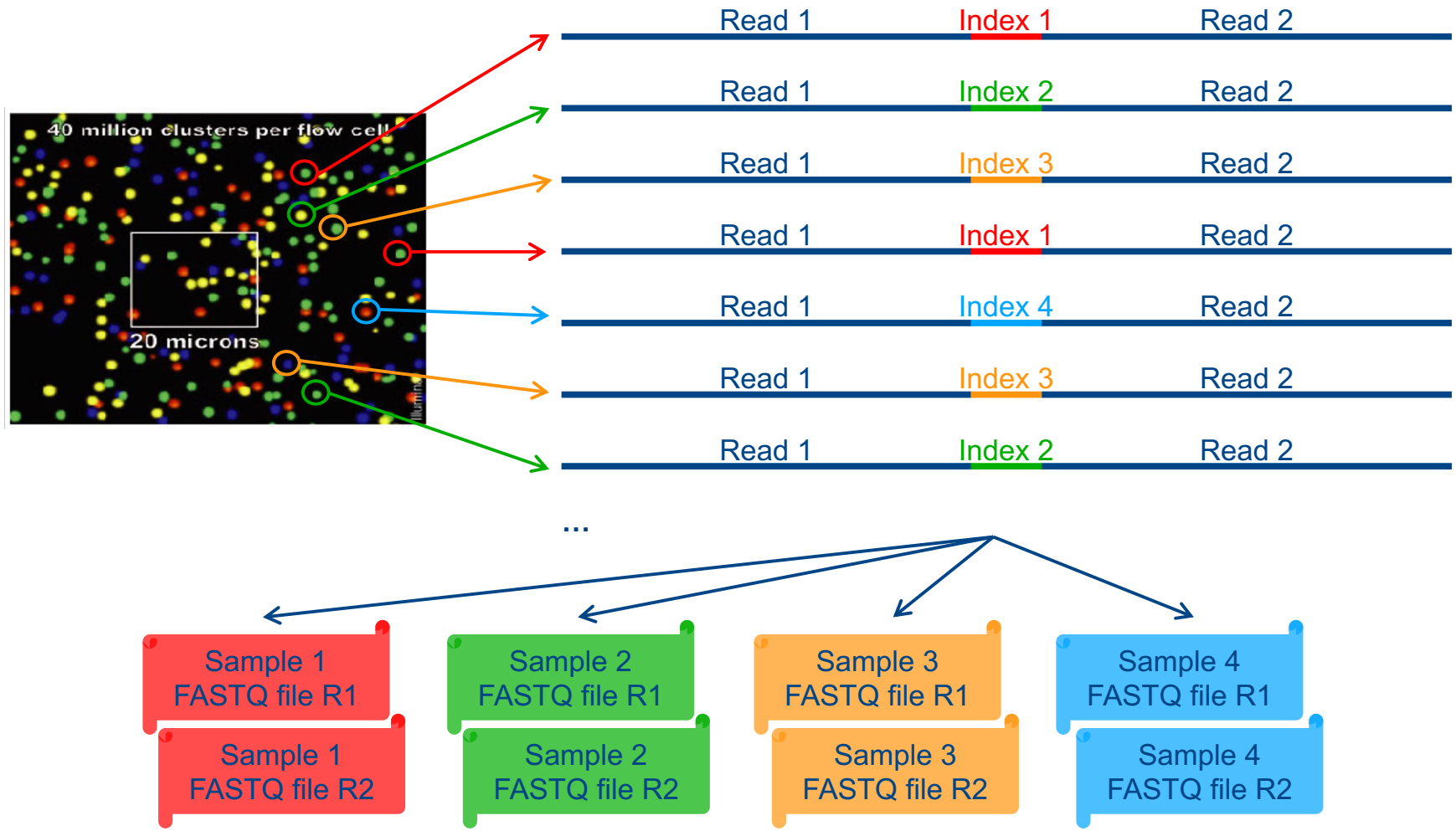
Phred Q =−10 log10 P

1 ASCII character per base

Phred Q = dec(ASCII) – 33

Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

Demultiplexing – PE sequencing



FASTQ file format – PE sequencing

Read 1

@HWI-EAS337_4_FC_3038UAAXX:1:2:1069:110**/1**
GATCGCCATCCAGAAATCGGTGGCTTGCTCACTTTC
+
A>AAAAAAAAAAAAAAAAAAA>>AA*:A>A:AAA3

@HWI-EAS337_4_FC_3038UAAXX:1:2:161:1891**/1**
GAACTGTTGCTGGAAGACTACAAAGCCTCCCTGAAA
+
A?A4?A>A>>A?<A>AA?AA=????;??6?????=<

@HWI-EAS337_4_FC_3038UAAXX:1:2:1041:124**/1**
GGTGATTTCAGCCAACAGGCCAGATTGAGCGTAGCC
+
A:AAAAAAAAAAAAA=AAA=AAAAAAAA<AA=AA9=

@HWI-EAS337_4_FC_3038UAAXX:1:2:1114:113**/1**
GCTTTGTTCGTGAAGCGTTATCGCTGGTGACATGGG
+
AAAAA?AAAAA?AA?A?AAAA>>AA>A:A>>7A>>A

…

Read 2

@HWI-EAS337_4_FC_3038UAAXX:1:2:1069:110**/2**
AACGAAGACCGCGTCGTATTGTTCCAAAAGCGAATC
+
AAAAAAAAAAAAAAAACAAAAAAAAAAAAA>>>>><

@HWI-EAS337_4_FC_3038UAAXX:1:2:161:1891**/2**
CAATATTTTACGTTGCTAATGACAGTGAACAGACTT
+
AAAAA>;AA6AAAA?AAAA?A&AAA<?;AA><<5;<

@HWI-EAS337_4_FC_3038UAAXX:1:2:1041:124**/2**
CCGGAATTAAAGTCACCGTTGAGCATCCGGATAAAC
+
AAAA;AAAAAAAAAA?AAA=A?AAAAAA6=>=><

@HWI-EAS337_4_FC_3038UAAXX:1:2:1114:113**/2**
CCGAAACAGACGCCCAGCACCCGATCGGTGCCTGAC
+
AAAAAA>AAAAAAAAAAAAAAAAAAAA(A==6=6<

…

Demultiplexing statistics

RUN	MACHINE	FLOWCELL	LANE	PROJECT	SAMPLE	BARCODE	CLUSTER.COUNT	BARCODE.COUNT	BARCODE.0.MISMATCH.COUNT	BARCODE.1.MISMATCH.COUNT	BARCODE.0.MISMATCH.PERC	BARCODE.1.MISMATCH.PERC	CLUSTER.COUNT.SUM.LANE	BASECALL.COUNT	YIELD.Q30.SUM	QSUM.SUM	BASECALL.QUAL.AVG	BASECALL.QUAL.ABOVE30.PERC	CLUSTER.COUNT.PERC
191004	HiSeq2000	FCA	1	Project	GC085522	CGCATGAT+TCAGGCTT	5,552,039	5,552,039	5,552,039	NA	1	NA	151,682,423	1,121,511,878	1,076,542,969	42,813,220,820	38.17	0.96	0.04
191004	HiSeq2000	FCA	1	Project	GC085522_ID	CTGAAGCT+TATAGCCT	9,777	9,777	9,777	NA	1	NA	151,682,423	1,974,954	1,913,612	75,745,127	38.35	0.97	0
191004	HiSeq2000	FCA	1	Project	GC085523	CTTAGGAC+GTAGGAGT	5,814,208	5,814,208	5,814,208	NA	1	NA	151,682,423	1,174,470,016	1,129,817,062	44,881,213,892	38.21	0.96	0.04
191004	HiSeq2000	FCA	1	Project	GC085523_ID	CTGAAGCT+ATAGAGGC	17,079	17,079	17,079	NA	1	NA	151,682,423	3,449,958	3,316,987	131,549,660	38.13	0.96	0
191004	HiSeq2000	FCA	1	Project	GC085524	ATCCGGTA+TATCGGTC	6,290,933	6,290,933	6,290,933	NA	1	NA	151,682,423	1,270,768,466	1,221,984,994	48,547,800,901	38.20	0.96	0.04
191004	HiSeq2000	FCA	1	Project	GC085524_ID	CTGAAGCT+CCTATCCT	14,085	14,085	14,085	NA	1	NA	151,682,423	2,845,170	2,757,522	109,110,362	38.35	0.97	0
191004	HiSeq2000	FCA	1	default	Undetermined	unknown	11,130,696	712,364,544	712,364,544	NA	1	NA	151,682,423	2,248,400,592	2,030,960,115	82,104,595,067	36.52	0.9	0.07

SAMPLE	BARCODE	CLUSTER.COUNT	BARCODE.COUNT	BARCODE.0.MISMATCH.COUNT	BARCODE.1.MISMATCH.COUNT	BARCODE.0.MISMATCH.PERC	BARCODE.1.MISMATCH.PERC
GC085522	CGCATGAT+TCAGGCTT	5,552,039	5,552,039	5,552,039	NA	1	NA
GC085522_ID	CTGAAGCT+TATAGCCT	9,777	9,777	9,777	NA	1	NA
GC085523	CTTAGGAC+GTAGGAGT	5,814,208	5,814,208	5,814,208	NA	1	NA
GC085523_ID	CTGAAGCT+ATAGAGGC	17,079	17,079	17,079	NA	1	NA
GC085524	ATCCGGTA+TATCGGTC	6,290,933	6,290,933	6,290,933	NA	1	NA
GC085524_ID	CTGAAGCT+CCTATCCT	14,085	14,085	14,085	NA	1	NA
Undetermined	unknown	11,130,696	712,364,544	712,364,544	NA	1	NA

Demultiplexing statistics

RUN	MACHINE	FLOWCELL	LANE	PROJECT	SAMPLE	BARCODE	CLUSTER.COUNT	BARCODE.COUNT	BARCODE.0.MISMATCH.COUNT	BARCODE.1.MISMATCH.COUNT	BARCODE.0.MISMATCH.PERC	BARCODE.1.MISMATCH.PERC	CLUSTER.COUNT.SUM.LANE	BASECALL.COUNT	YIELD.Q30.SUM	QSUM.SUM	BASECALL.QUAL.AVG	BASECALL.QUAL.ABOVE30.PERC	CLUSTER.COUNT.PERC
191004	HiSeq2000	FCA	1	Project	GC085522	CGCATGAT+TCAGGCTT	5,552,039	5,552,039	5,552,039	NA	1	NA	151,682,423	1,121,511,878	1,076,542,969	42,813,220,820	38.17	0.96	0.04
191004	HiSeq2000	FCA	1	Project	GC085522_ID	CTGAAGCT+TATAGCCT	9,777	9,777	9,777	NA	1	NA	151,682,423	1,974,954	1,913,612	75,745,127	38.35	0.97	0
191004	HiSeq2000	FCA	1	Project	GC085523	CTTAGGAC+GTAGGAGT	5,814,208	5,814,208	5,814,208	NA	1	NA	151,682,423	1,174,470,016	1,129,817,062	44,881,213,892	38.21	0.96	0.04
191004	HiSeq2000	FCA	1	Project	GC085523_ID	CTGAAGCT+ATAGAGGC	17,079	17,079	17,079	NA	1	NA	151,682,423	3,449,958	3,316,987	131,549,660	38.13	0.96	0
191004	HiSeq2000	FCA	1	Project	GC085524	ATCCGGTA+TATCGGTC	6,290,933	6,290,933	6,290,933	NA	1	NA	151,682,423	1,270,768,466	1,221,984,994	48,547,800,901	38.20	0.96	0.04
191004	HiSeq2000	FCA	1	Project	GC085524_ID	CTGAAGCT+CCTATCCT	14,085	14,085	14,085	NA	1	NA	151,682,423	2,845,170	2,757,522	109,110,362	38.35	0.97	0
191004	HiSeq2000	FCA	1	default	Undetermined	unknown	11,130,696	712,364,544	712,364,544	NA	1	NA	151,682,423	2,248,400,592	2,030,960,115	82,104,595,067	36.52	0.9	0.07

SAMPLE	CLUSTER.COUNT.SUM.LANE	BASECALL.COUNT	YIELD.Q30.SUM	QSUM.SUM	BASECALL.QUAL.AVG	BASECALL.QUAL.ABOVE30.PERC	CLUSTER.COUNT.PERC
GC085522	151,682,423	1,121,511,878	1,076,542,969	42,813,220,820	38.17	0.96	0.04
GC085522_ID	151,682,423	1,974,954	1,913,612	75,745,127	38.35	0.97	0
GC085523	151,682,423	1,174,470,016	1,129,817,062	44,881,213,892	38.21	0.96	0.04
GC085523_ID	151,682,423	3,449,958	3,316,987	131,549,660	38.13	0.96	0
GC085524	151,682,423	1,270,768,466	1,221,984,994	48,547,800,901	38.20	0.96	0.04
GC085524_ID	151,682,423	2,845,170	2,757,522	109,110,362	38.35	0.97	0
Undetermined	151,682,423	2,248,400,592	2,030,960,115	82,104,595,067	36.52	0.9	0.07

Outline

- Definitions
- FASTQ format
- Quality control
- Pre-processing

Example 1 – targeted capture sequencing

- Finding the genetic cause of a disease
  - ~ 6,000 genes
  - Illumina PE 125 bp

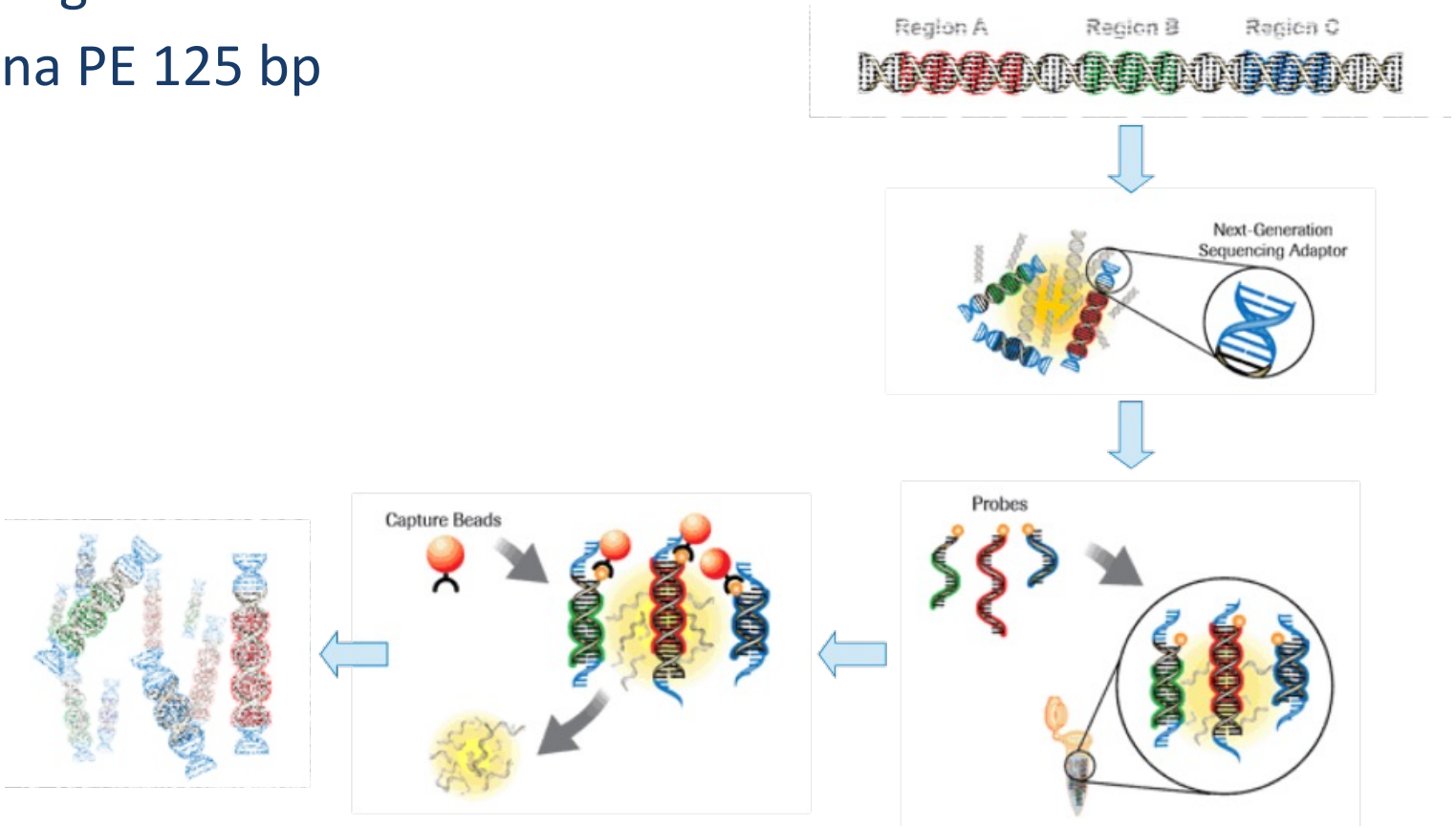
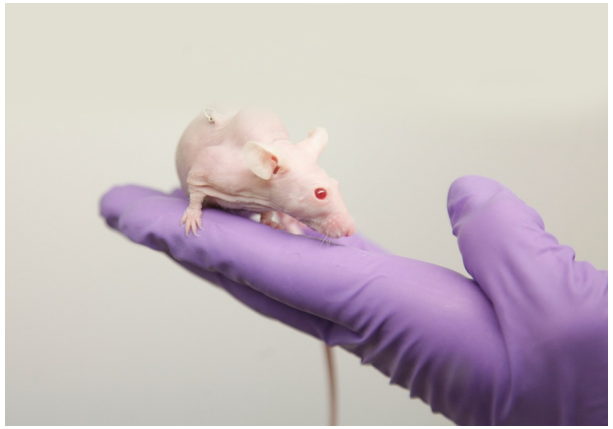


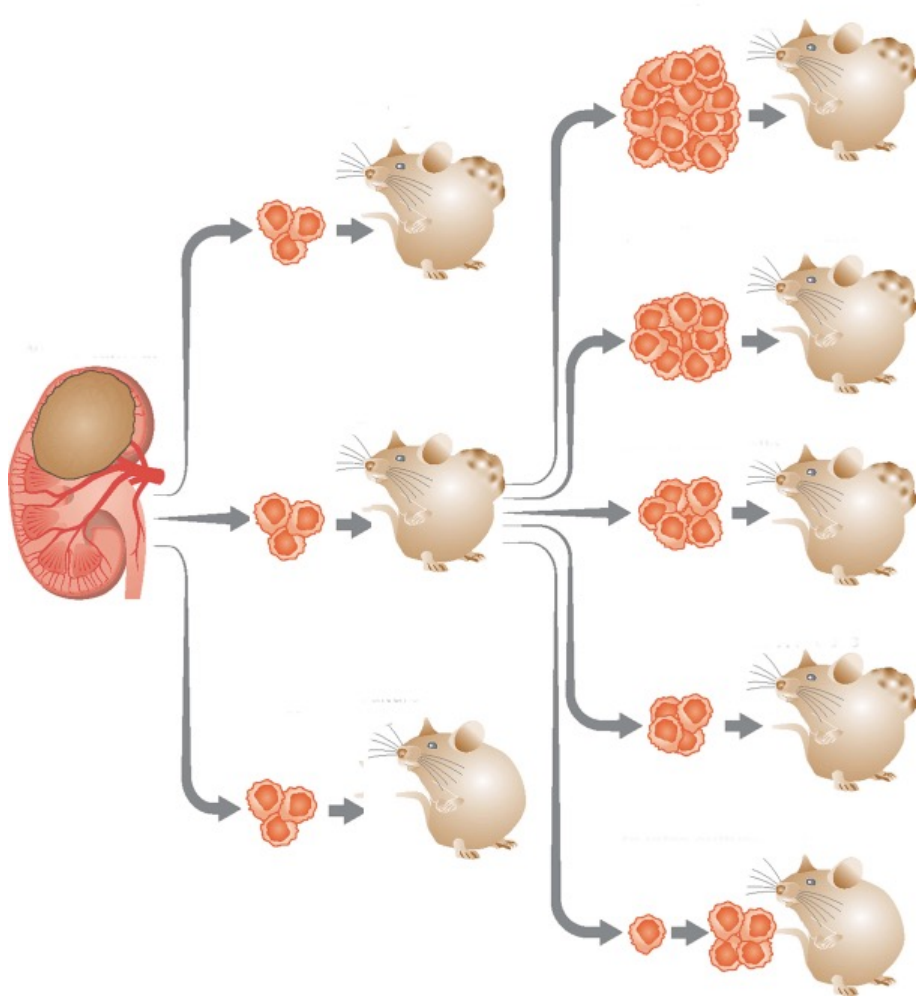
Figure adapted from Nimblegen

Example 2 – targeted amplicon sequencing

- Fingerprinting of xenocrafts
  - 31 SNPs
  - Illumina PE 150 bp



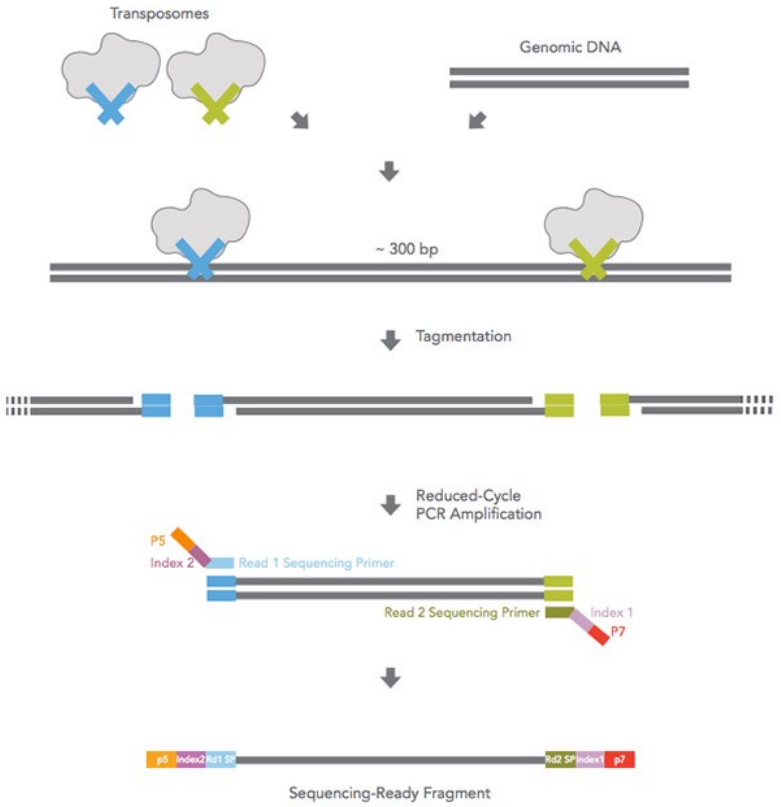
Trace
Patient Derived Tumor Xenograft
Platform



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

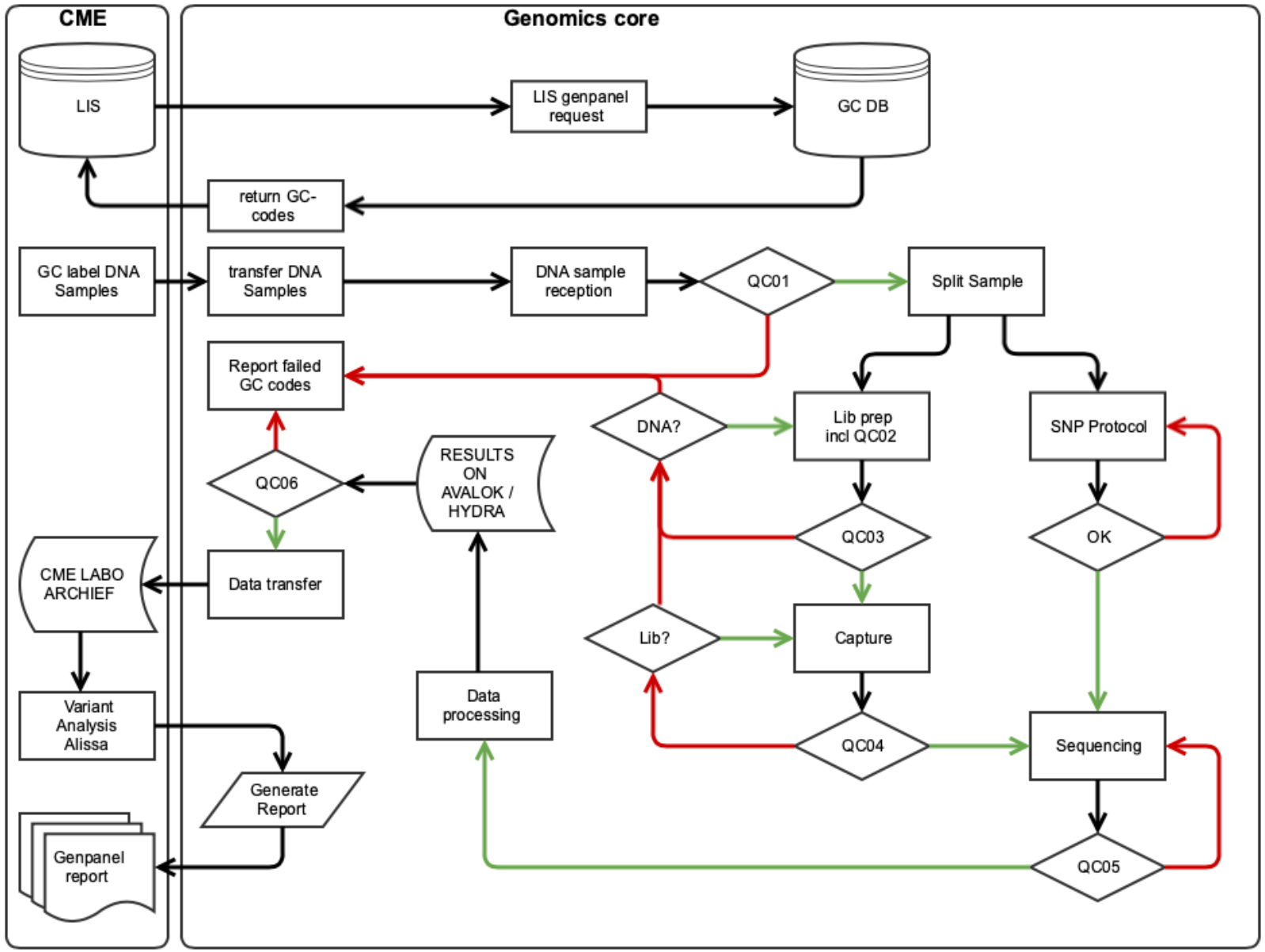
Example 3 – whole genome sequencing

- Predicting bacterial resistance
  - Whole Genome Sequencing (WGS)
    - *Mycobacterium tuberculosis (BWGS)*
  - Illumina PE 300 bp



NGS workflow

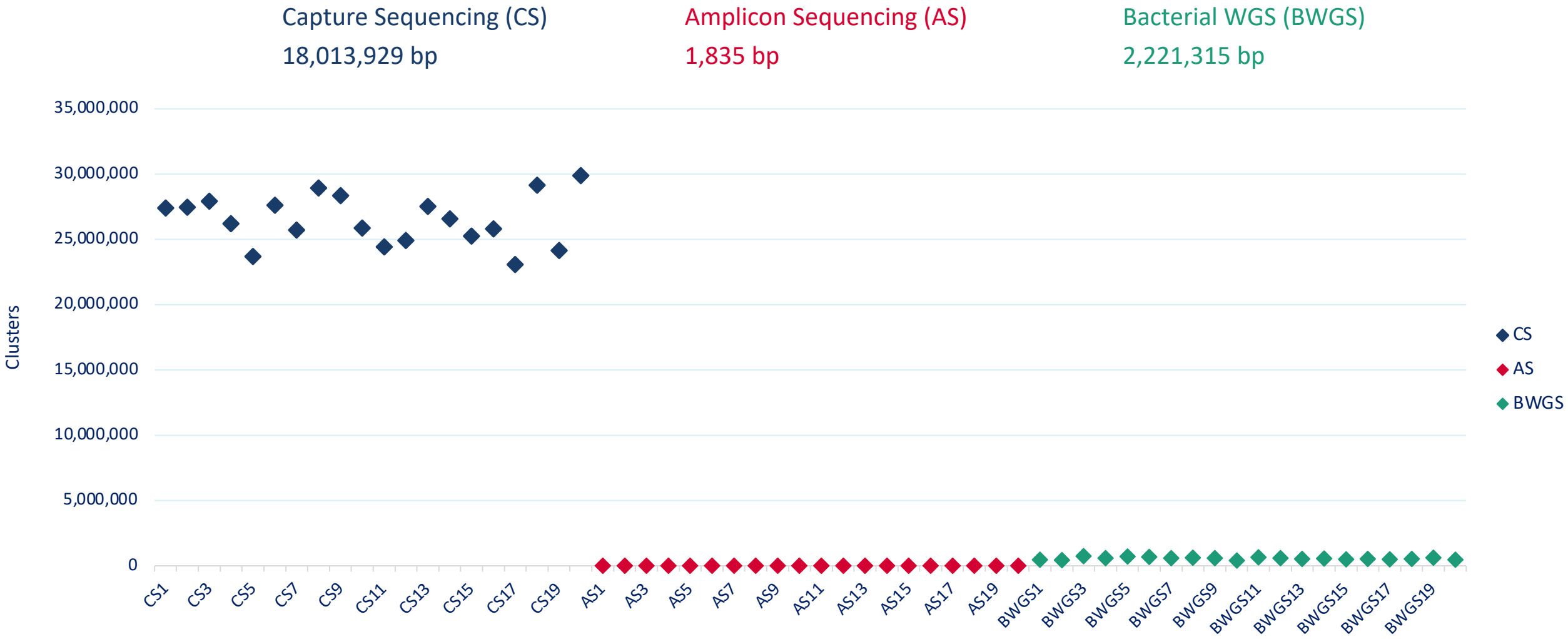
- Example of targeted (capture) test



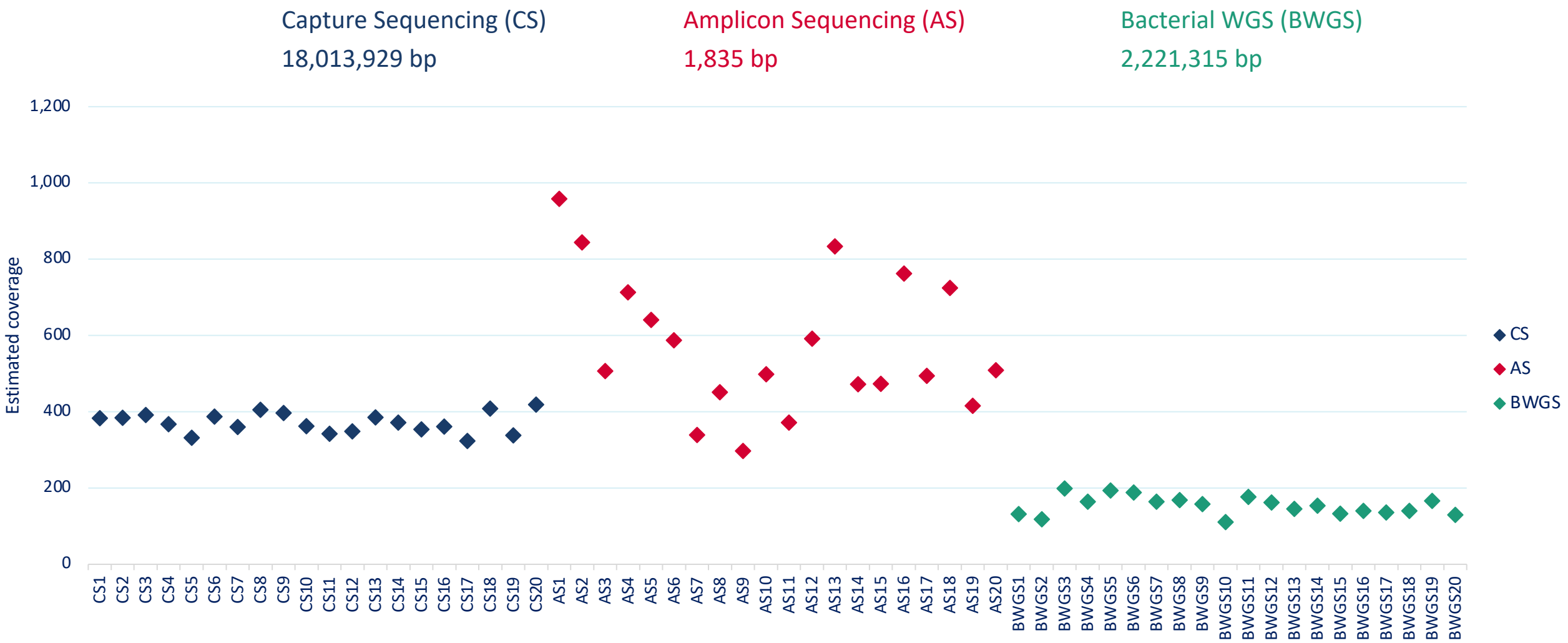
Need for quality control

- Complex workflow
- Cost efficient assay
- Ensure sequenced data is OK
  - Enough data ?
  - Correct content ?
  - Good quality data ?

Number of reads



Estimated mean coverage



FASTQ files – QC

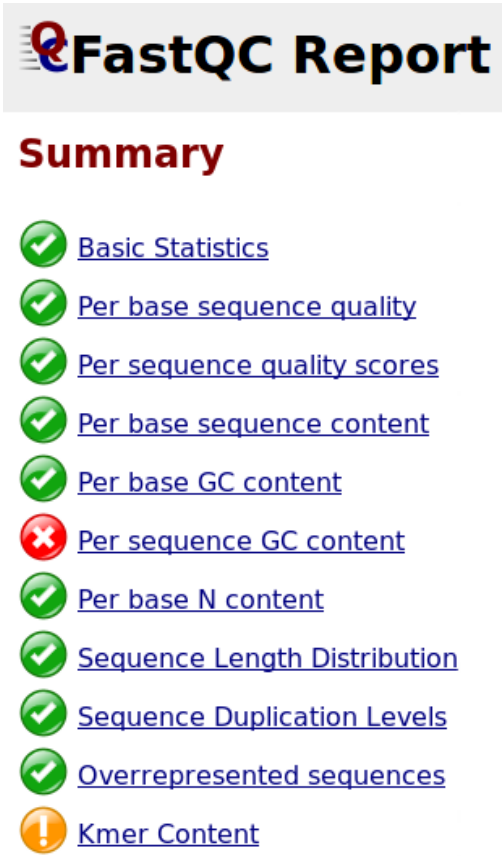
- FastQC
  - Check Phred quality scores
  - Check GC content
  - Check read content
  - …

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

FastQC

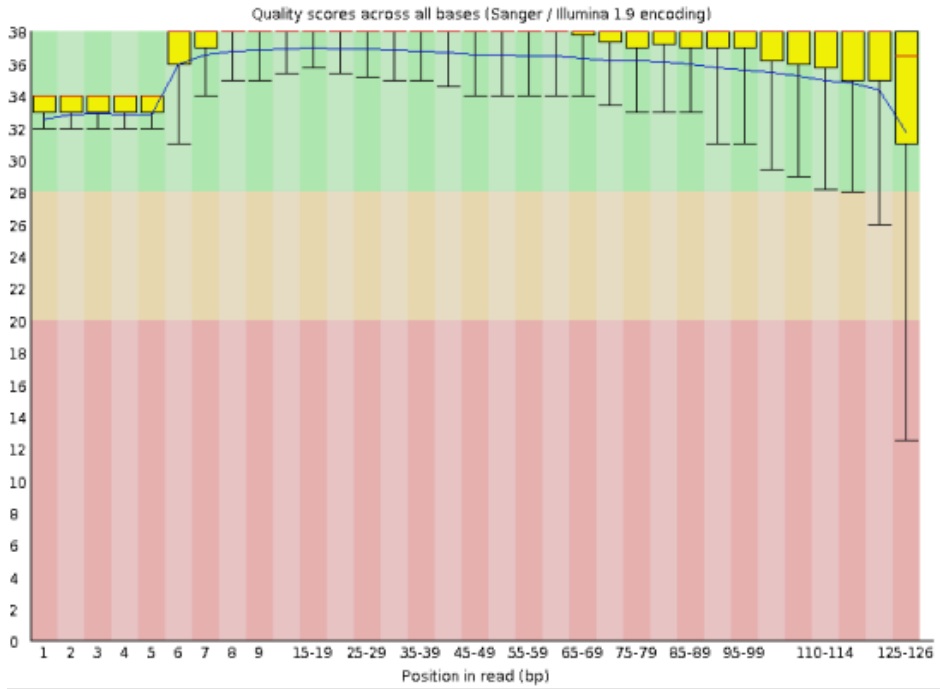
- Summary report

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

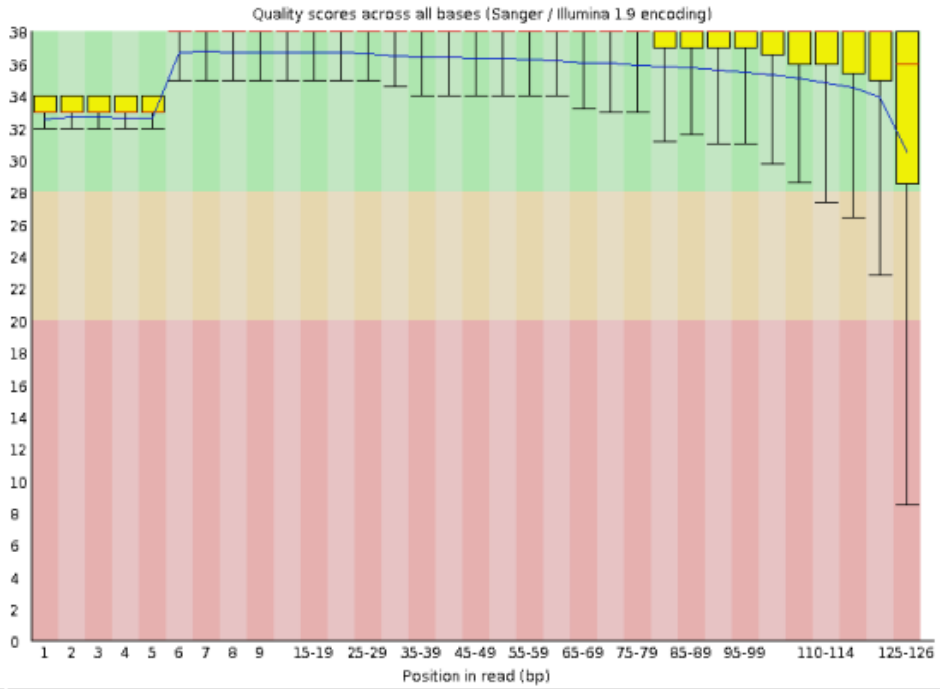


FastQC – Phred quality score by position

- Example 1 – targeted capture sequencing



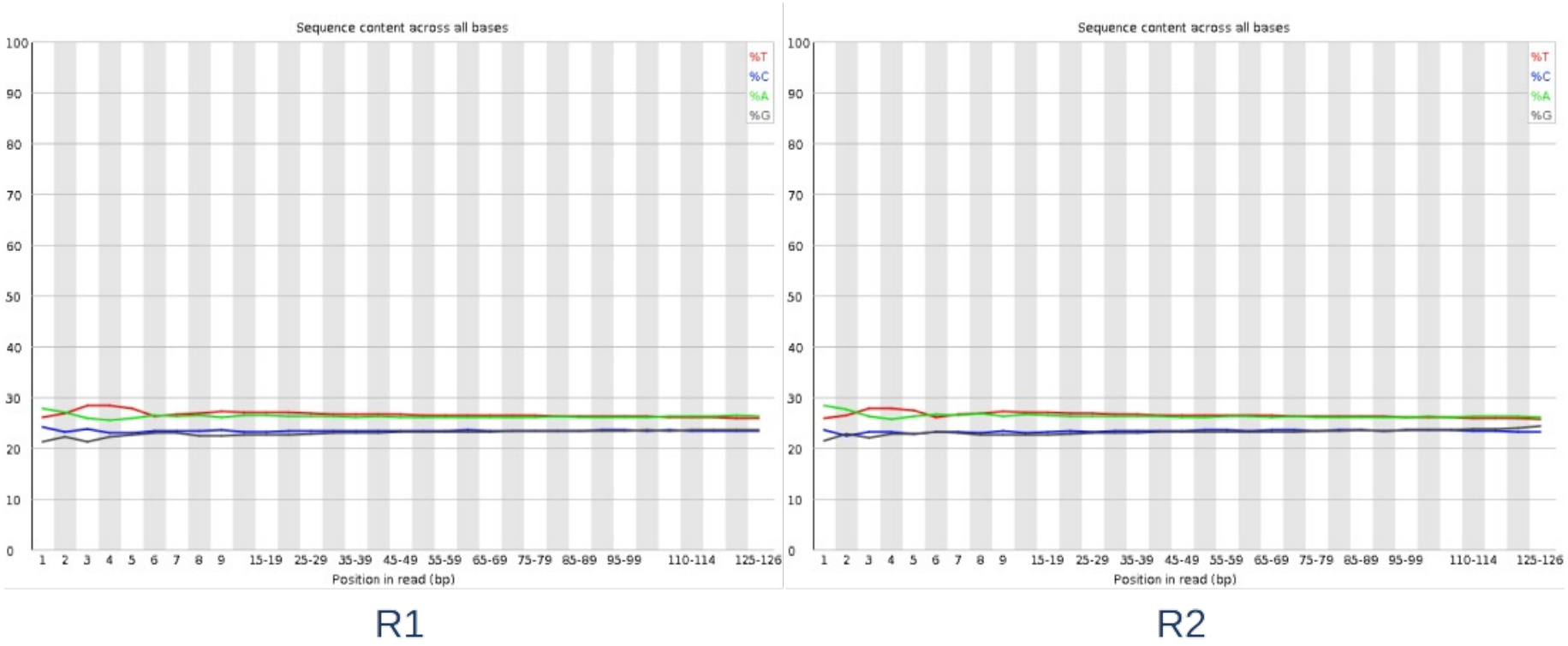
R1



R2

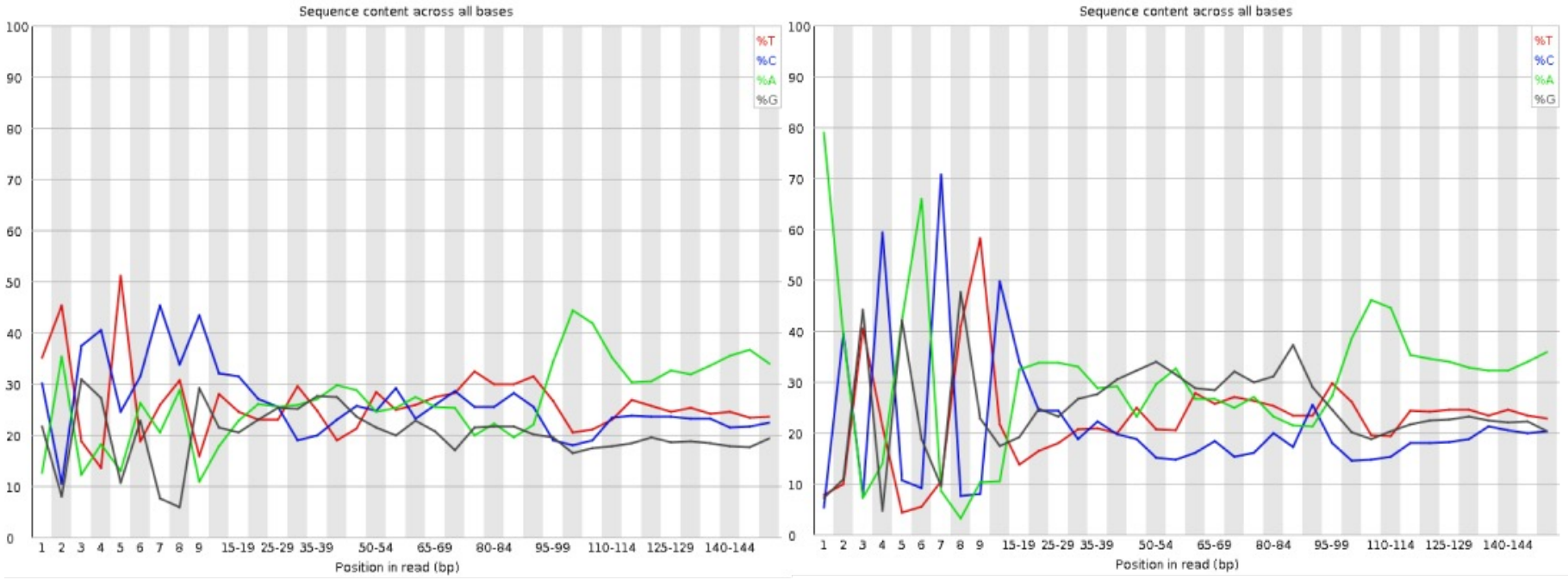
FastQC – Base content by position

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



FastQC – Base content by position

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



R1 R2

FastQC – Over-represented sequences

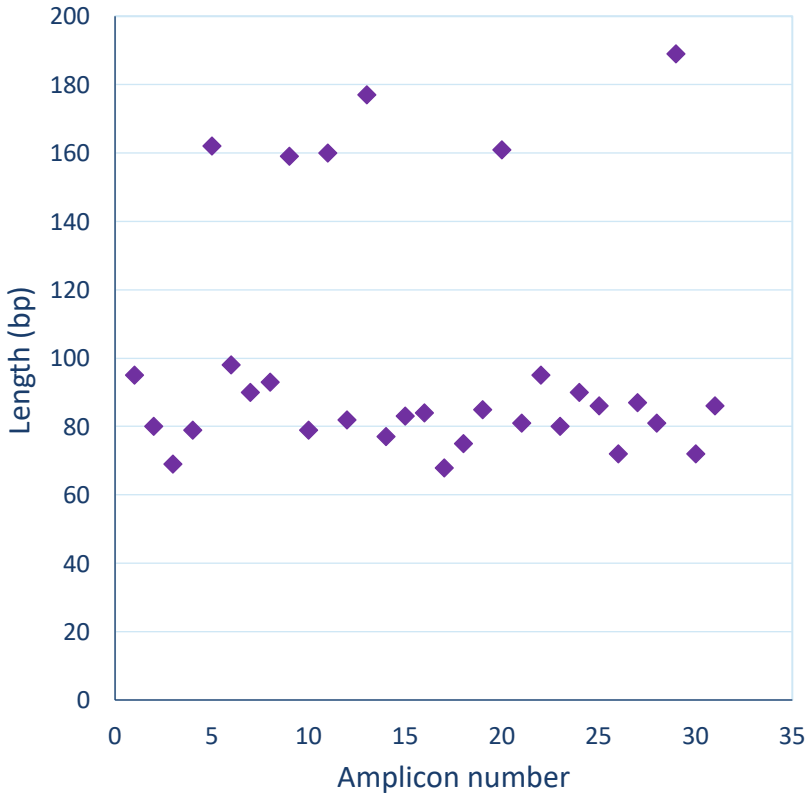
- Example 2 – targeted amplicon sequencing

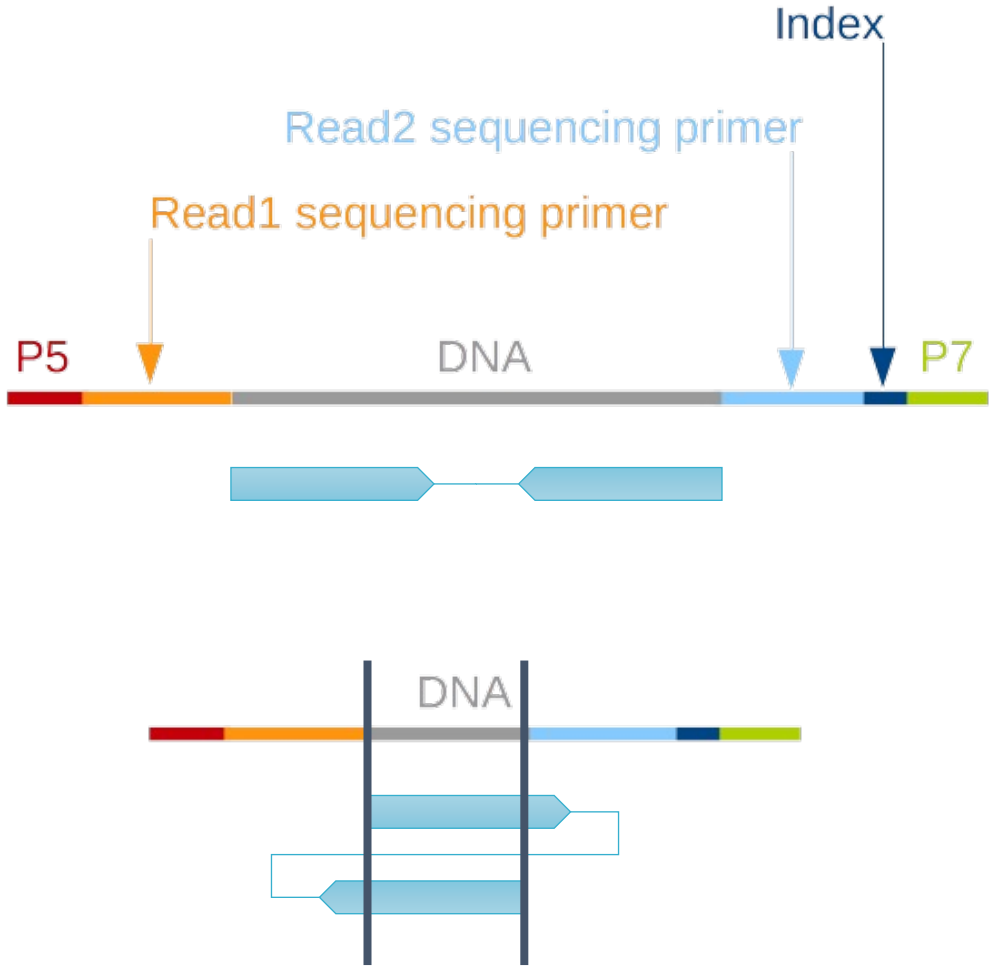
Overrepresented sequences

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 – targeted amplicon sequencing





Contamination check

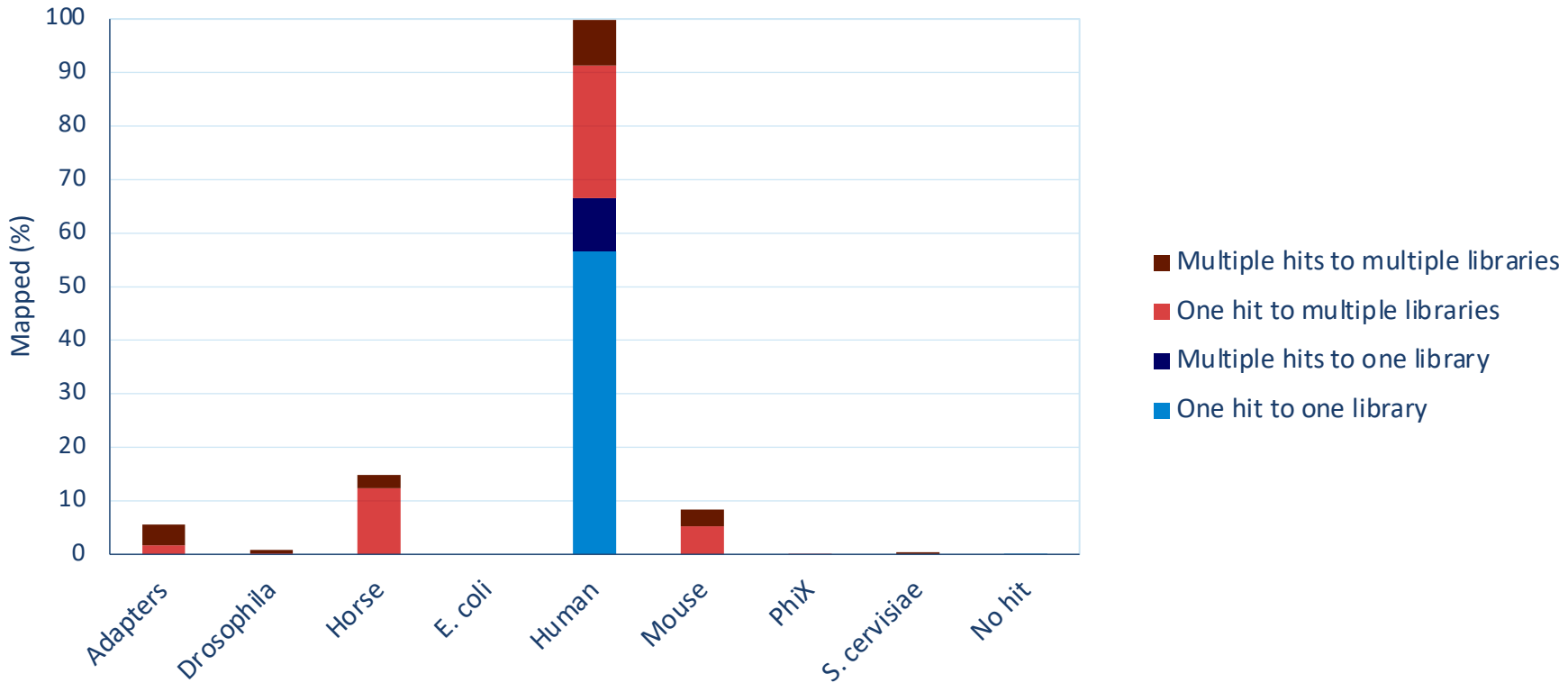
- Compare reads to various libraries
- Any library can be searched against
- Output proportion of reads with hit(s) to libraries

```
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/

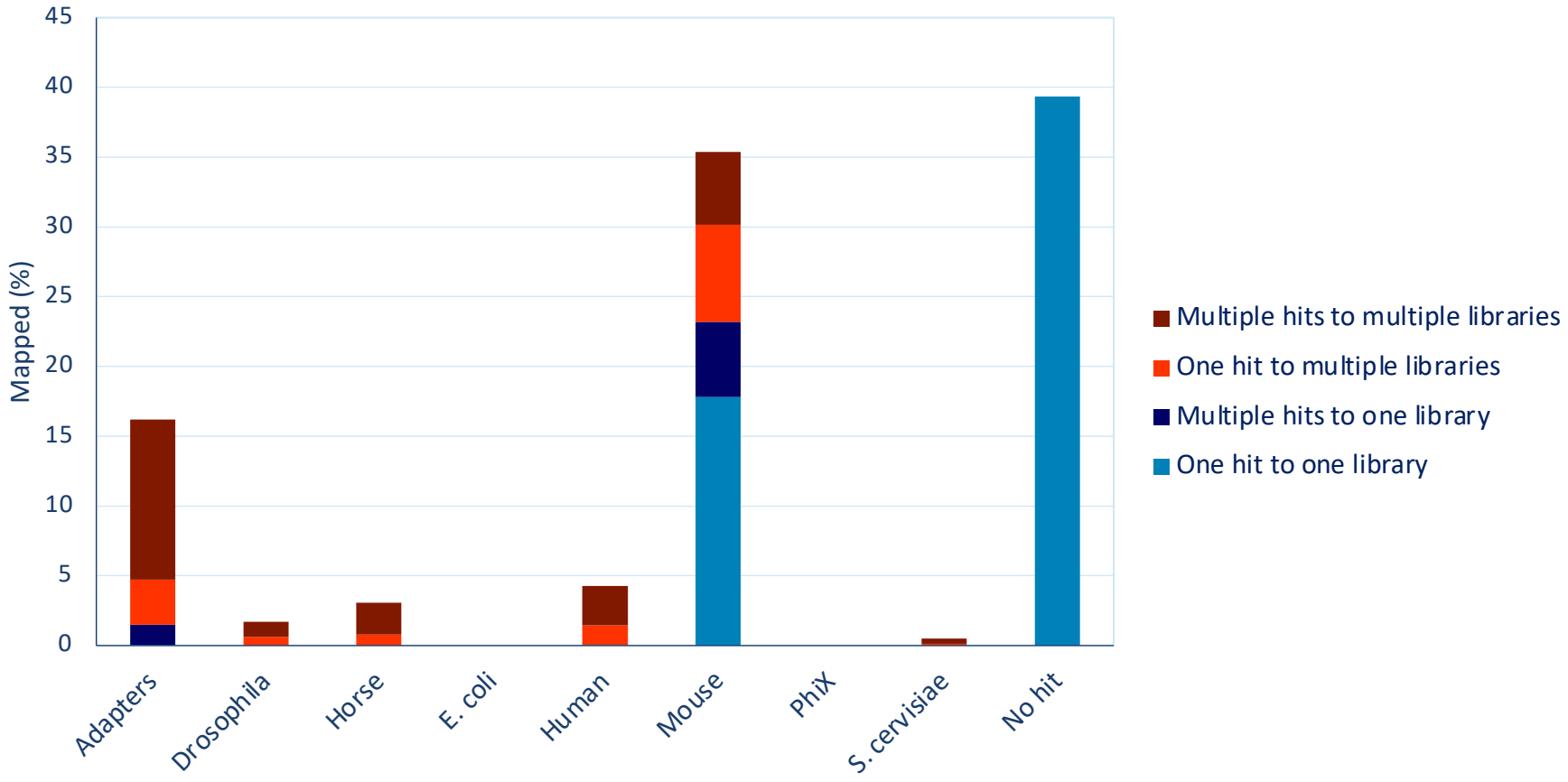
Contamination check

- Example 1 – targeted capture sequencing



Contamination check

- Example 2 – targeted amplicon sequencing



Contamination check

- 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

- 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

Outline

- Definitions
- FASTQ format
- Quality control
- Pre-processing

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 DNA fragment
  - ...

Clipping & trimming

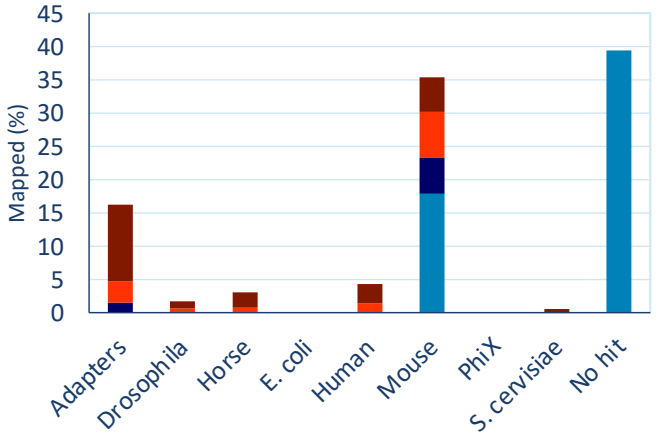
- Sequencing adapters and primers
- Poor quality bases at the ends of reads
- Ns from ends of reads
- Remove low complexity reads

```
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

- Example 2 – targeted 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'

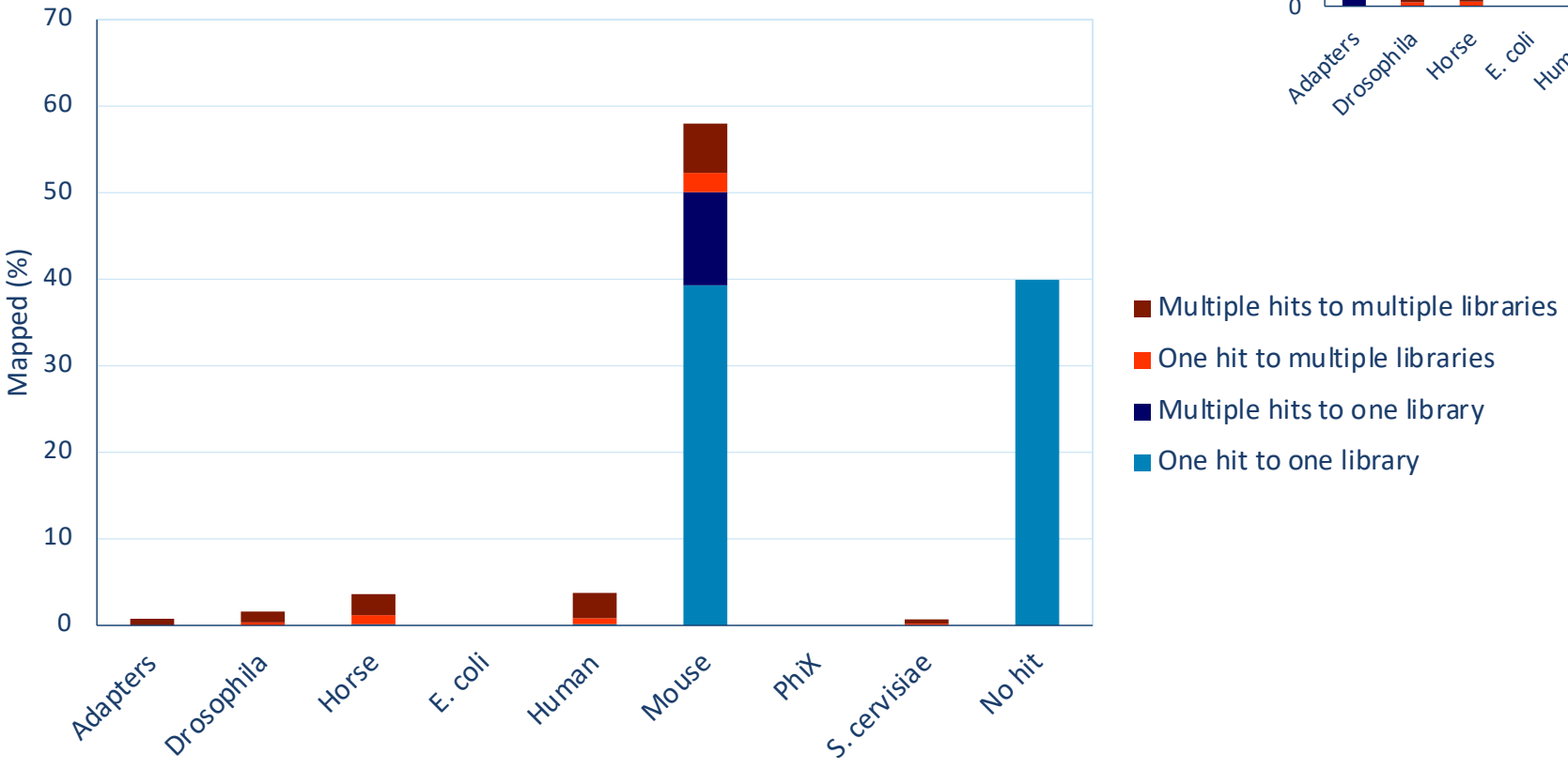
…

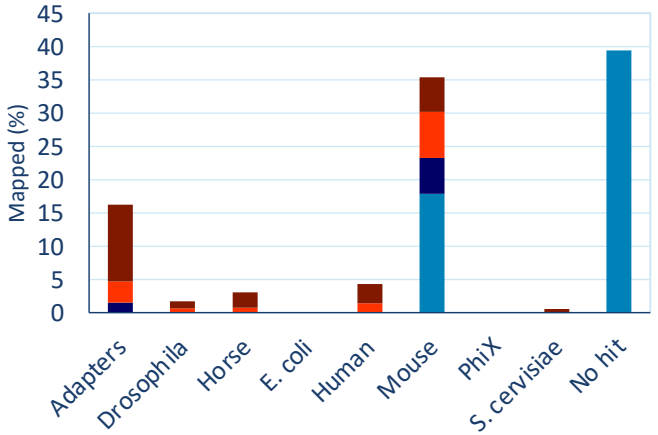
Total reads: 1801

Too short after clip: 33

Adapter clipping & trimming

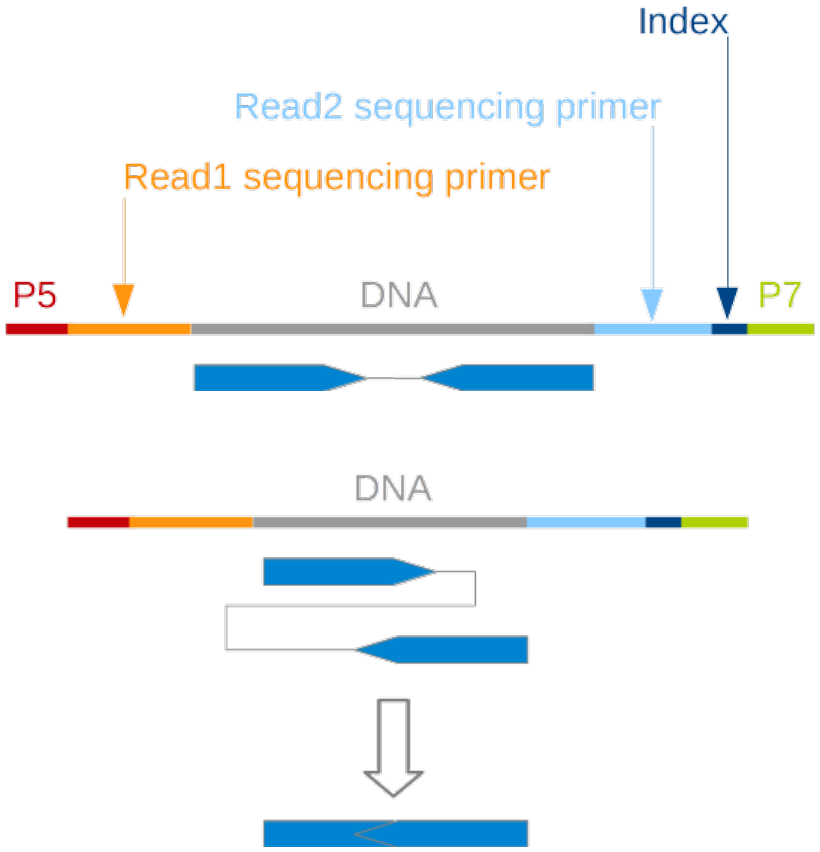
- Example 2 – targeted amplicon sequencing





Merge overlapping reads

- Merge paired-end reads
- Keep DNA insert only

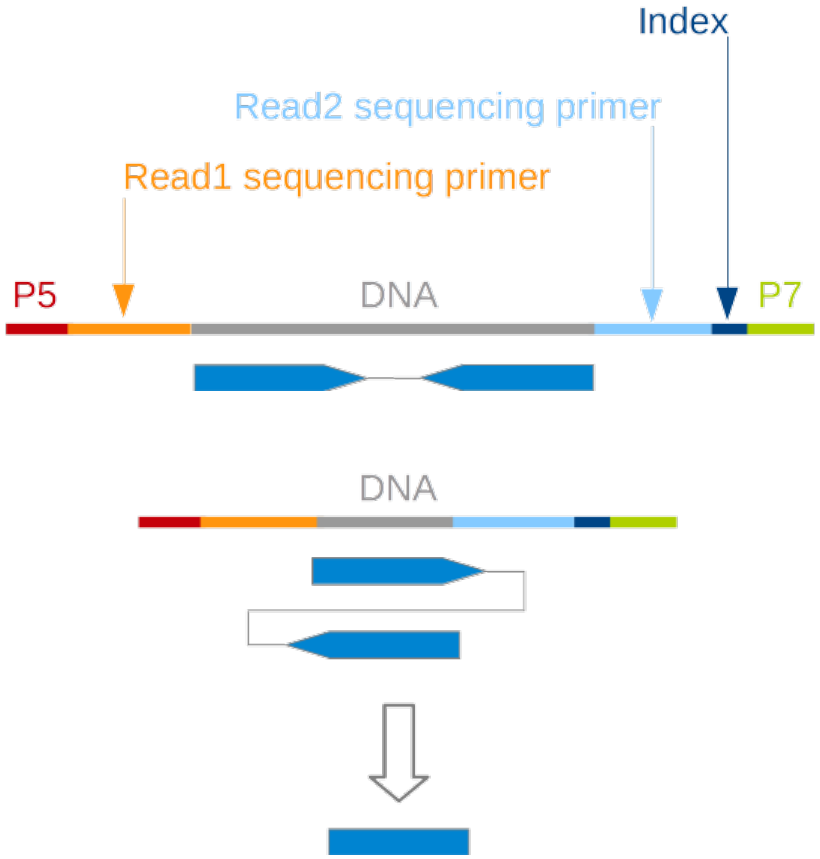



```
flash2 --max-overlap=250 --min-overlap=20 --allow-outies -d result -o
sample.flashed sample.R1.fastq.gz sample.R2.fastq.gz > flash.log
```

https://github.com/dstreett/FLASH2

Merge overlapping reads

- Merge paired-end reads
- Keep DNA insert 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

Read selection

- Select reads from sequenced organism

Fragments covered by clade (%)	Fragments covered by clade	Fragments assigned to taxon	Rank code	NCBI taxonomic ID	scientific name
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

```
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

Summary

- Quality control can be done at the raw data level
- Pipeline has to be tailored for analysis
- Pre-processing might be required depending on the analysis

Questions?

Erika Souche