Quality Control for Sequencing Experiments

v2023-04

Simon Andrews, Sarah Inglesfield

simon.andrews@babraham.ac.uk

sarah.inglesfield@babraham.ac.uk

Babraham Bioinformatics

Babraham Bioinformatics

- Support service for bioinformatics
  - Academic – Babraham Institute
  - Commercial – Consultancy

- Support BI Sequencing Facility
  - MiSeq/ HiSeq/ NextSeq-based sequencing service
  - Data Management / Processing / Analysis

Interests in QC

QC packages

FastQC

Librarian

Fast Screen

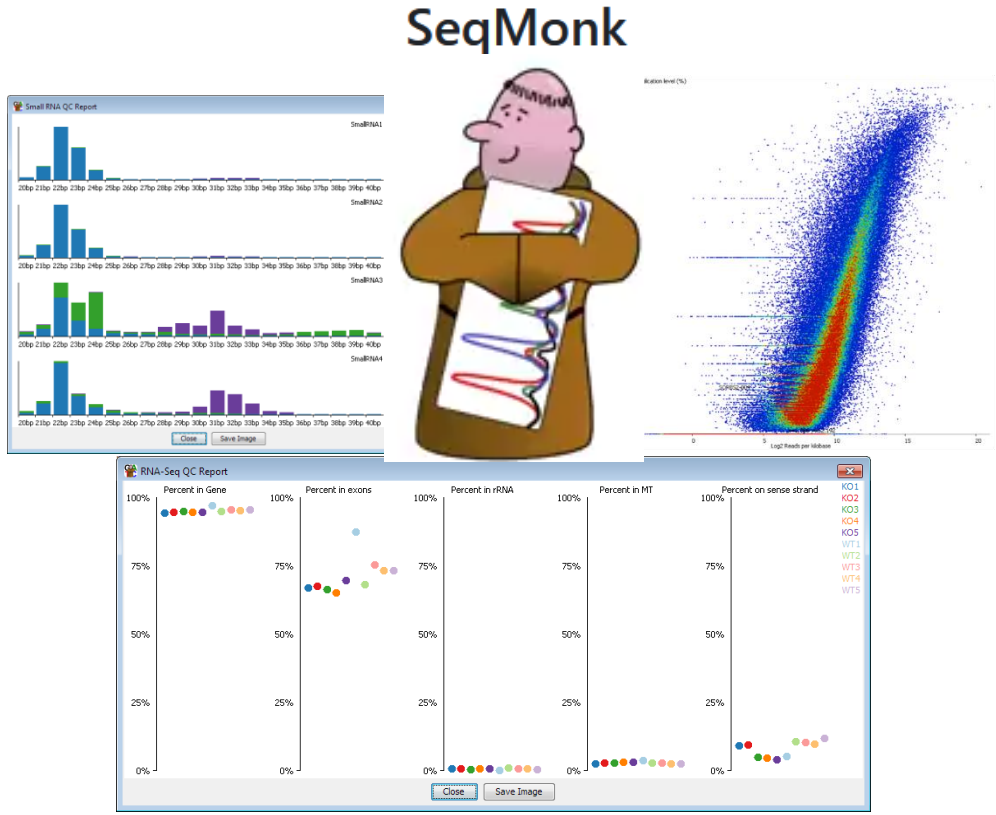
Application specific QC

Bismark

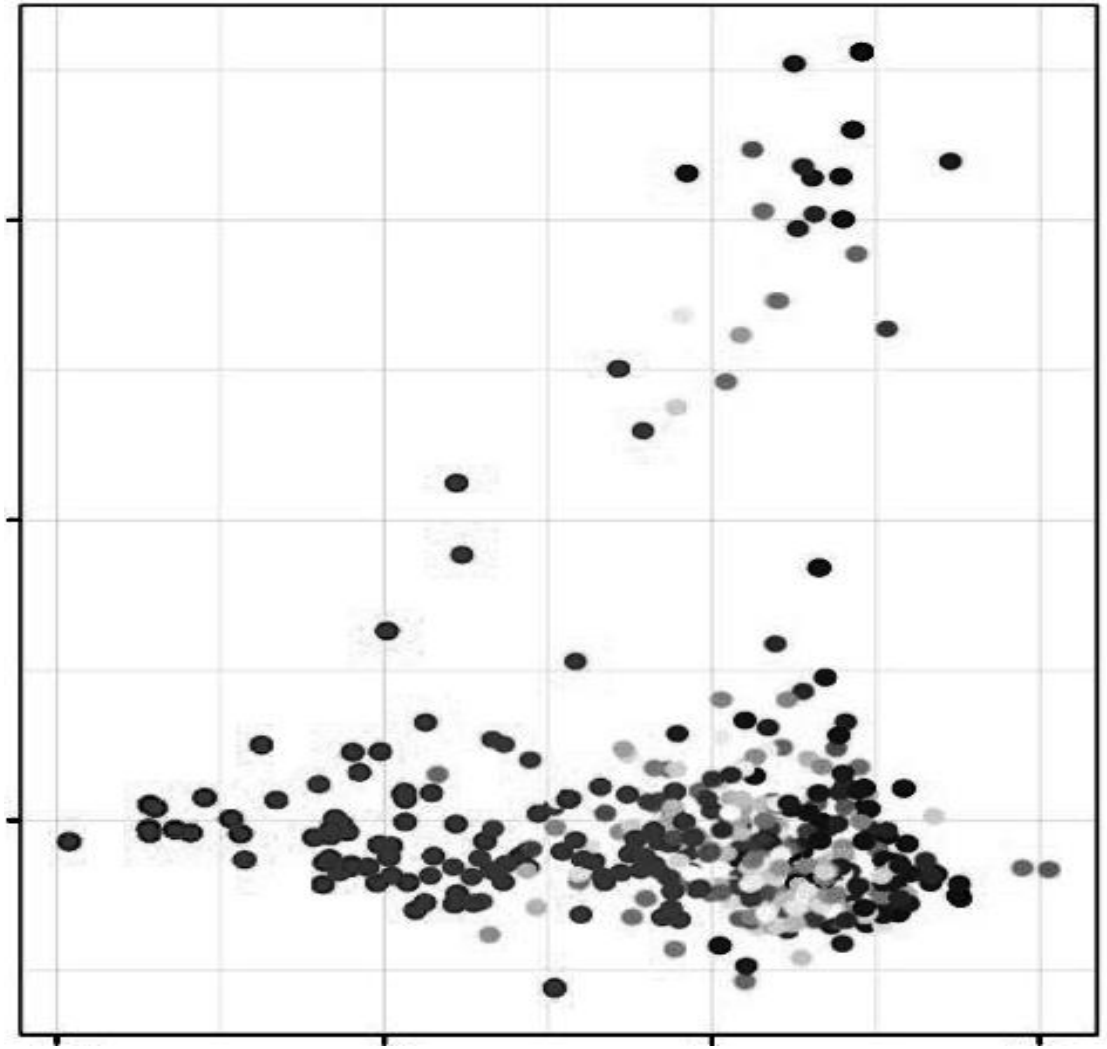
HiCUP

Data visualisation QC

SeqMonk



An example of why QC is important…

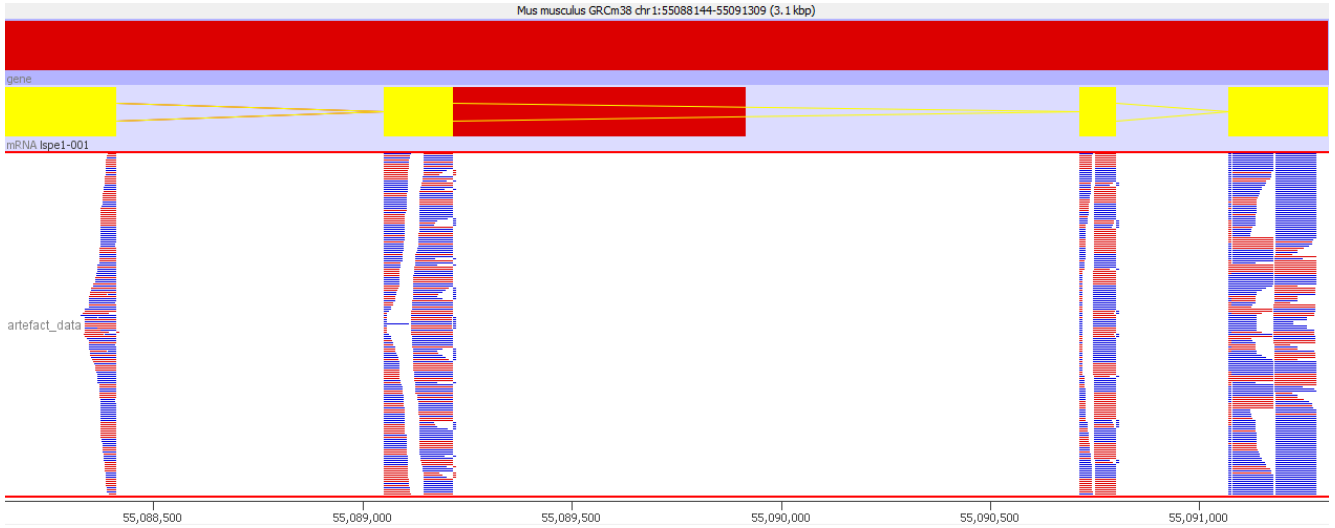


- Single Cell RNA-Seq
  - Each dot is a cell
  - An outgroup is clearly visible
  - What is it?

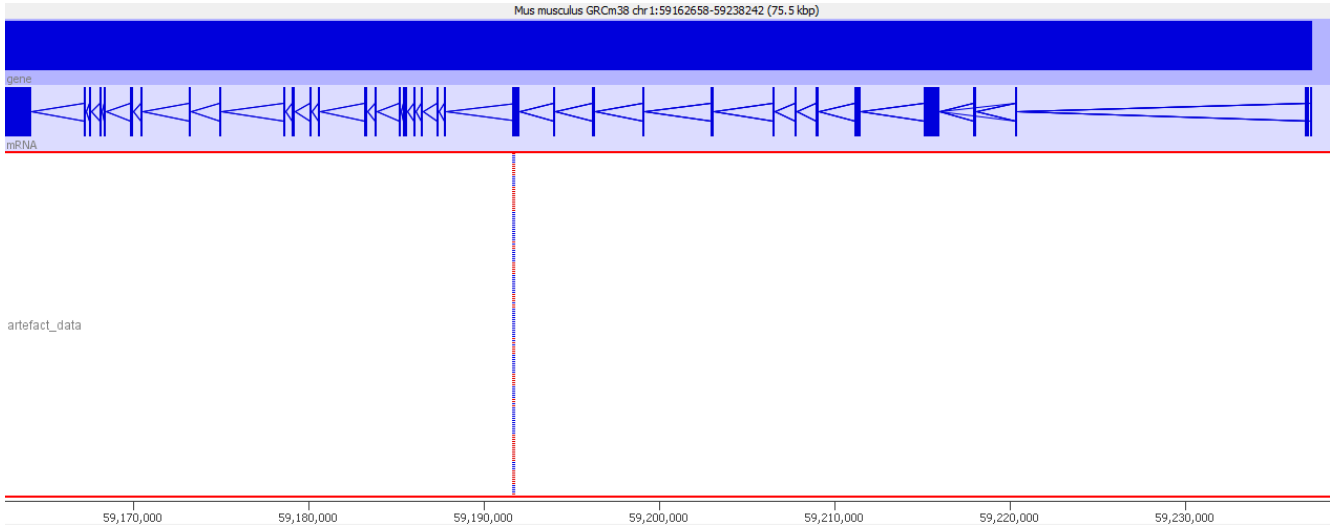
Genes for PC2 (85 total)

Gene	Description
Arhgef4	Rho guanine nucleotide exchange factor (GEF) 4
Cflar	CASP8 and FADD-like apoptosis regulator
Als2	amyotrophic lateral sclerosis 2 (juvenile) homolog (human)
Cxcr2	chemokine (C-X-C motif) receptor 2
Col4a3	collagen, type IV, alpha 3
Sag	retinal S-antigen
Gpr35	G protein-coupled receptor 35
Acmsd	amino carboxymuconate semialdehyde decarboxylase
Qsox1	quiescin Q6 sulfhydryl oxidase 1
9430070O13Rik	RIKEN cDNA 9430070O13 gene
Mrps14	mitochondrial ribosomal protein S14
Scyl3	SCY1-like 3 (S. cerevisiae)
Ildr2	immunoglobulin-like domain containing receptor 2
Atp1a2	ATPase, Na+/K+ transporting, alpha 2 polypeptide
Slamf8	SLAM family member 8
Wdr38	WD repeat domain 38
Exd1	exonuclease 3'-5' domain containing 1
Serf2	small EDRK-rich factor 2

Coverage of Raw Data

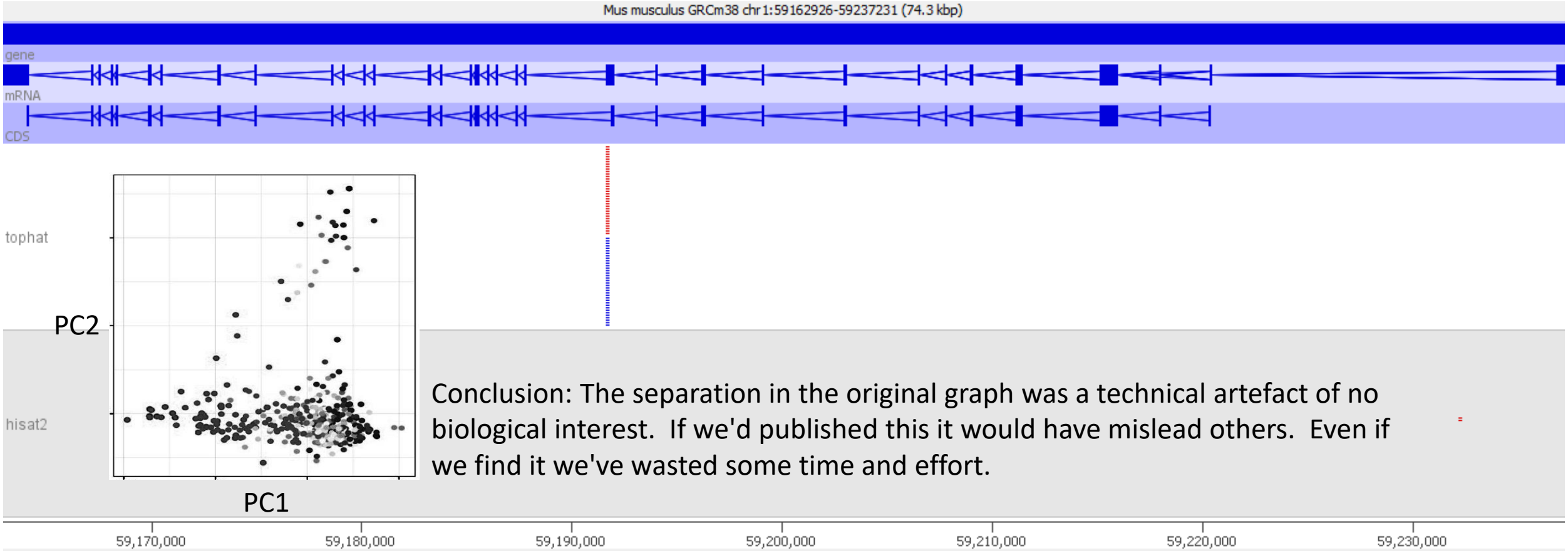


Normal Gene



PCA Gene

Using a different read mapper…



Conclusion: The separation in the original graph was a technical artefact of no biological interest. If we'd published this it would have mislead others. Even if we find it we've wasted some time and effort.

What is the point of QC?

- Technical problems don't cause pipelines to fail
- Technical problems don't prevent hits being generated
- Technical hits often look biologically real
- Unexpected, interesting effects can easily be missed
- Finding problems through follow-on work is slow and expensive!

- QC saves you time and effort! (and money)

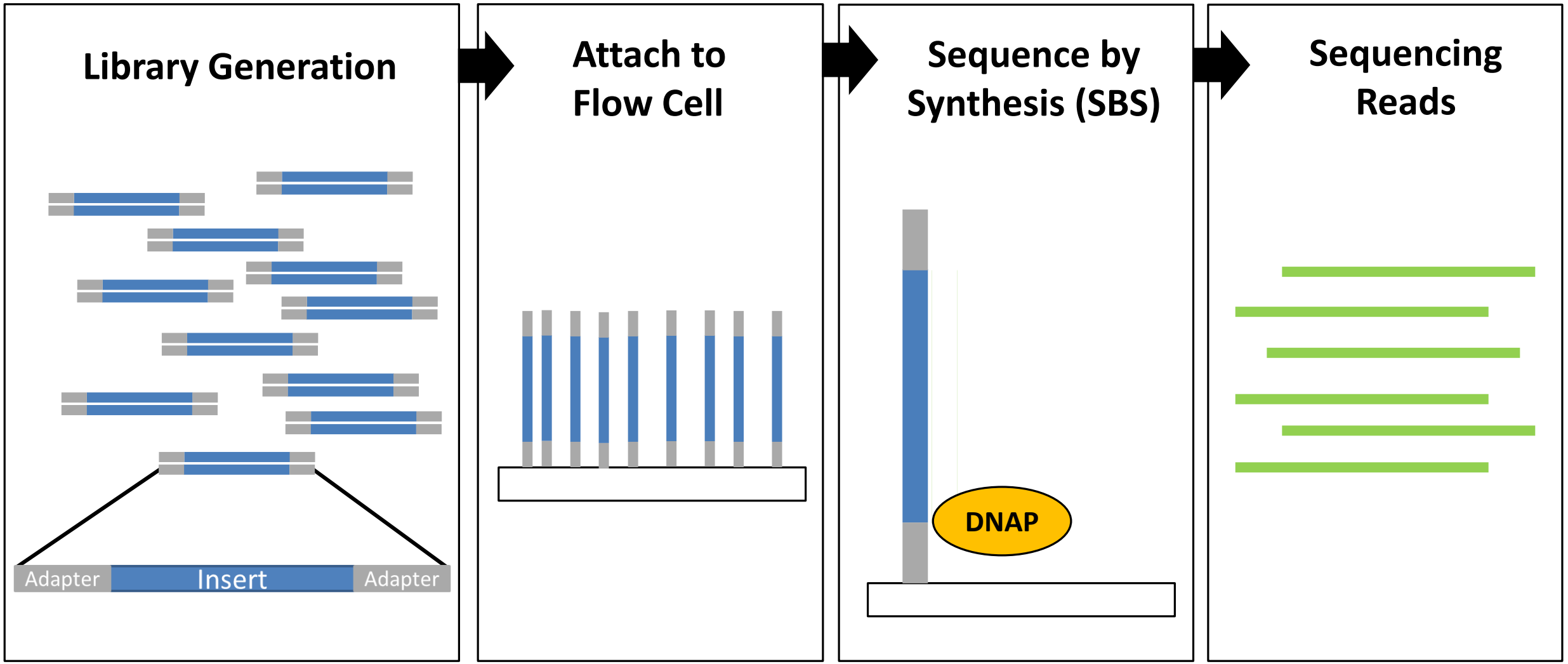
Course Structure

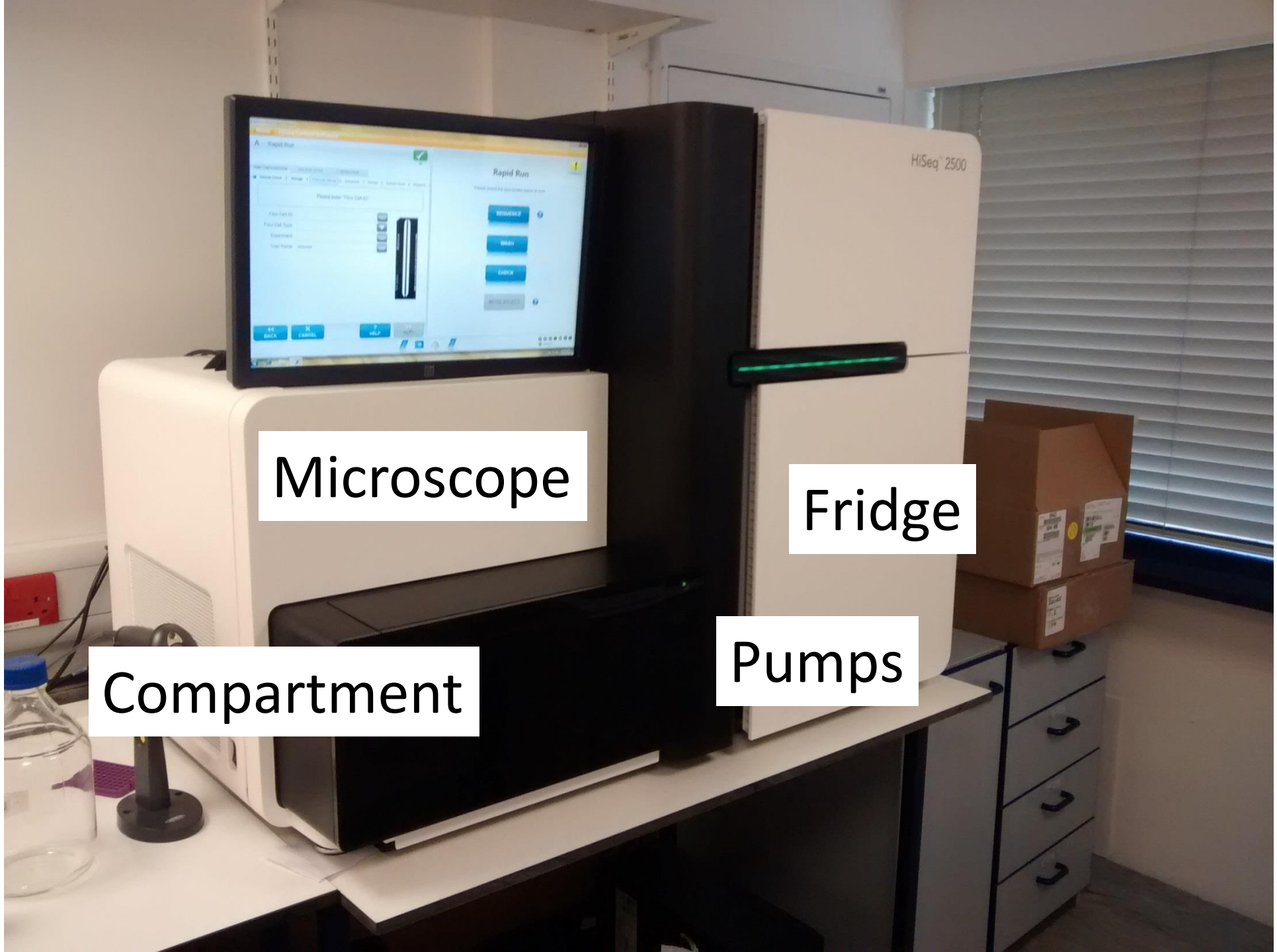
- How does Illumina Sequencing work?
- What can QC tell us?
  - QC Software
  - Universal metrics
  - Library Dependent metrics
  - Consistency
- Putting QC into Practice

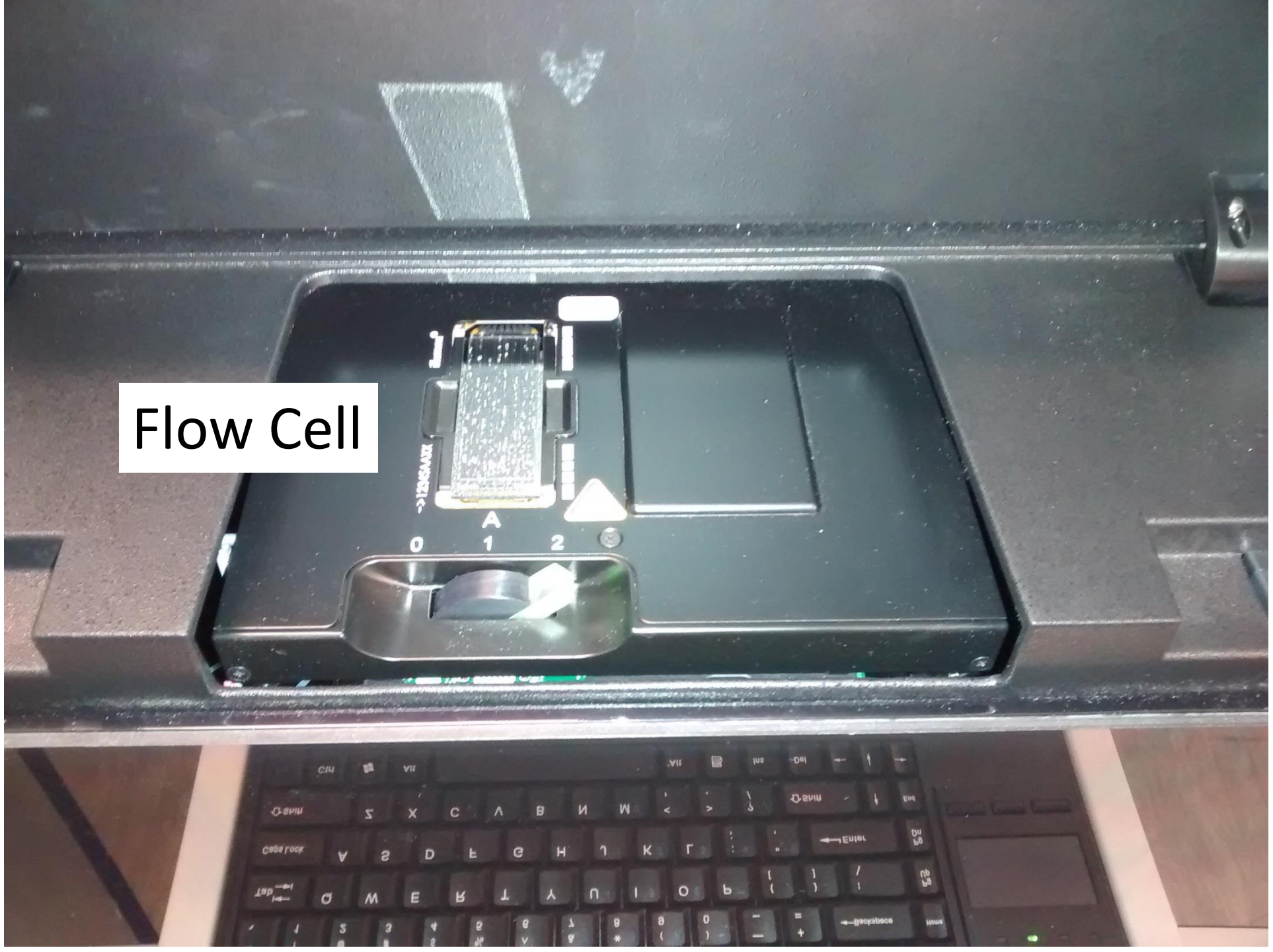
How Does Illumina Sequencing Work?

Babraham Bioinformatics

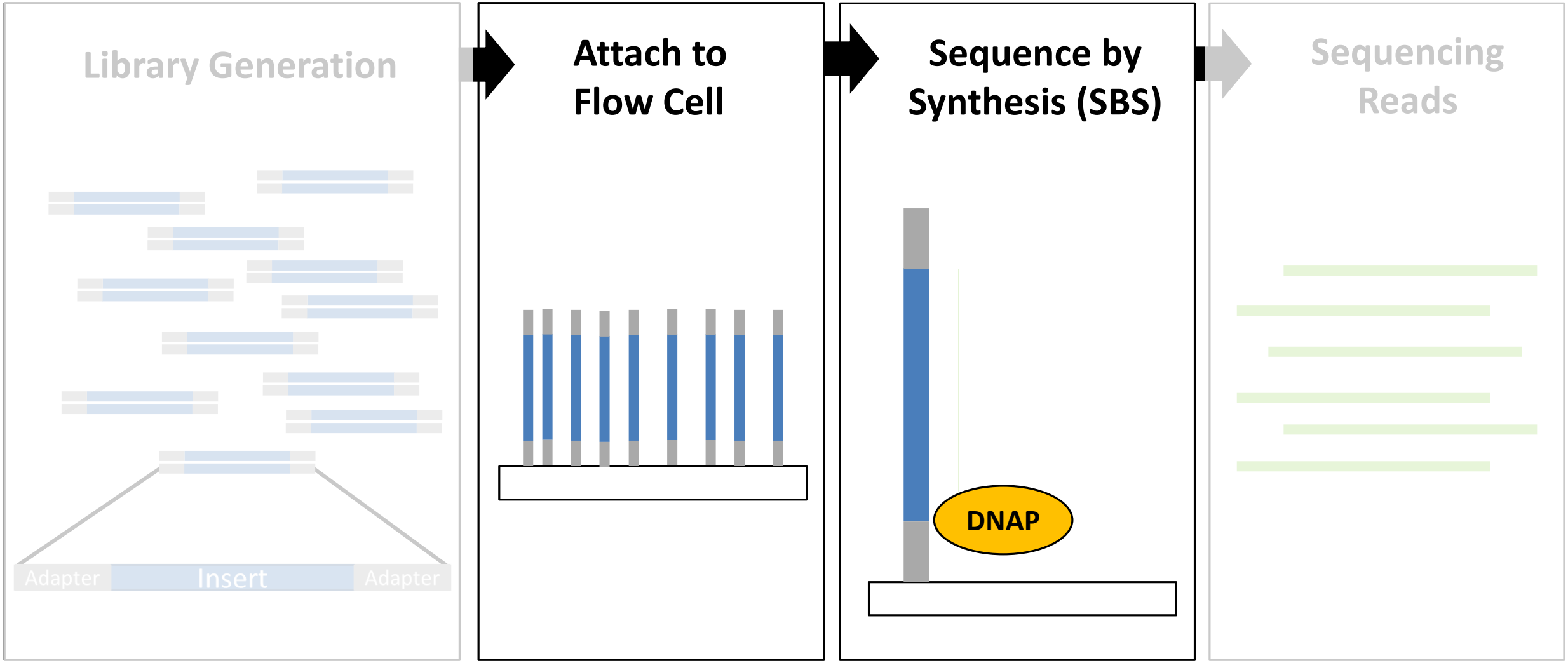
Illumina Sequencing: An Overview



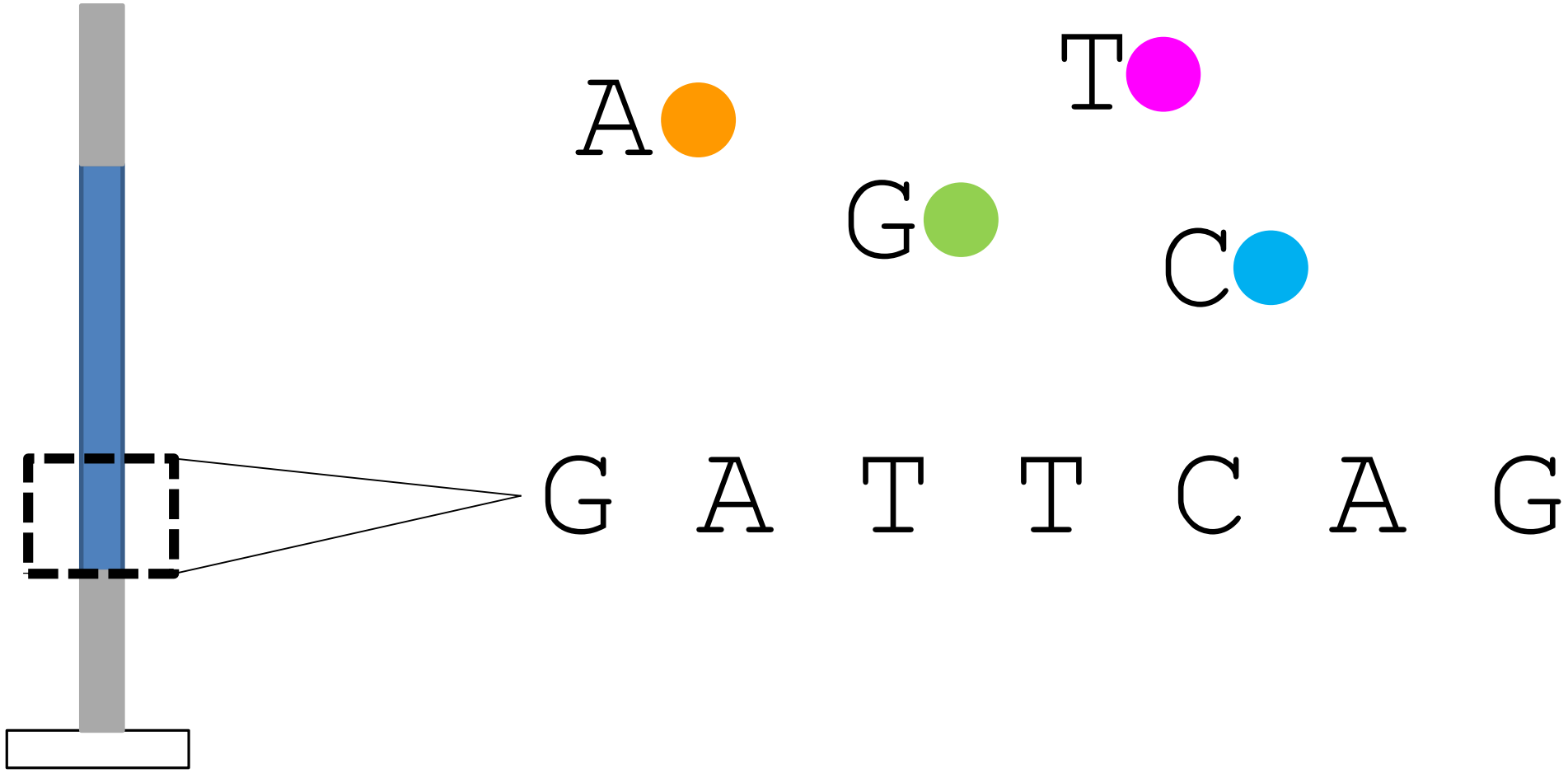

Microscope
Fridge
Compartment
Pumps
HiSeq 2500
Rapid Run


Flow Cell
A
0
1
2

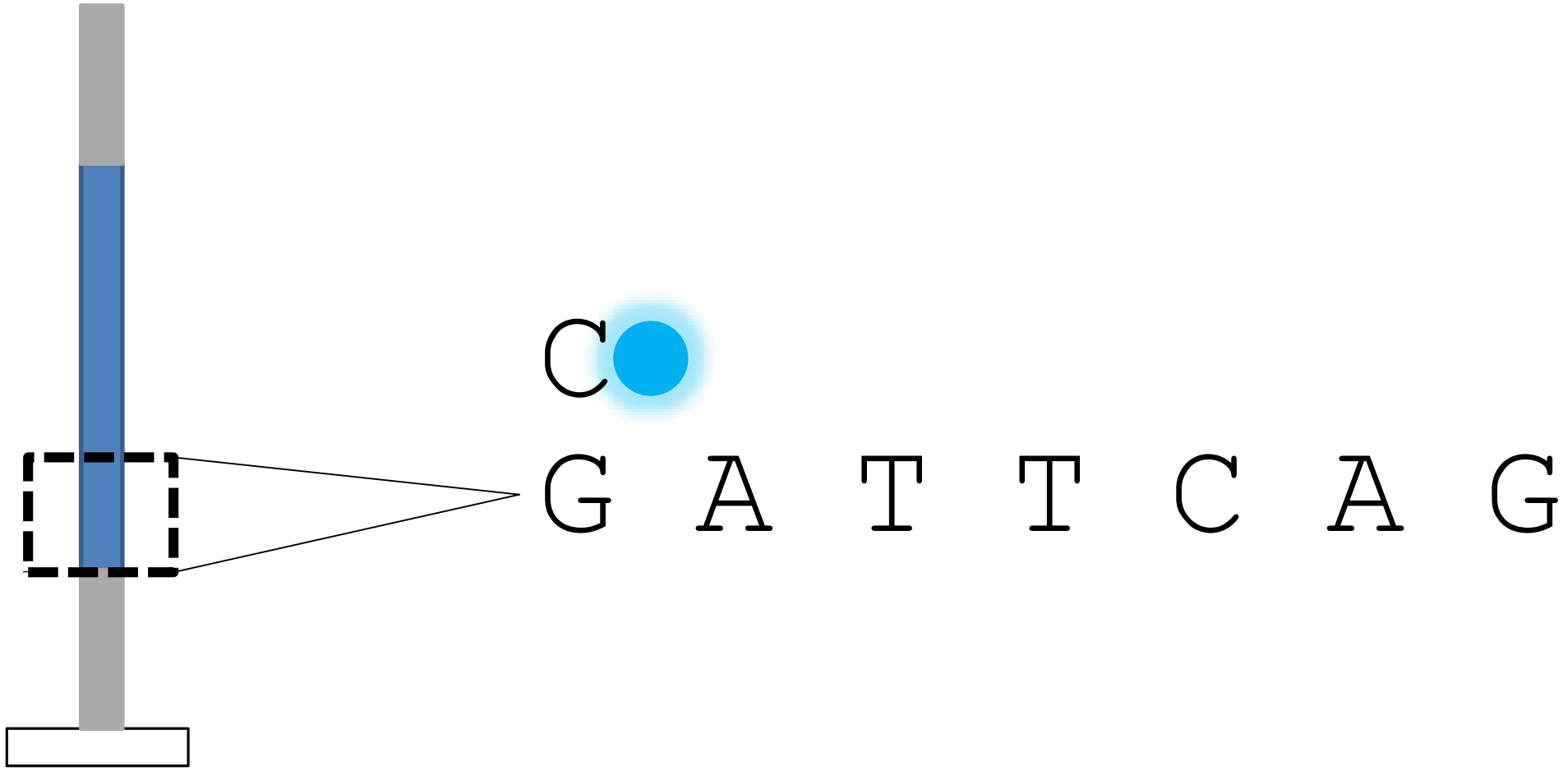
Illumina Sequencing: An Overview



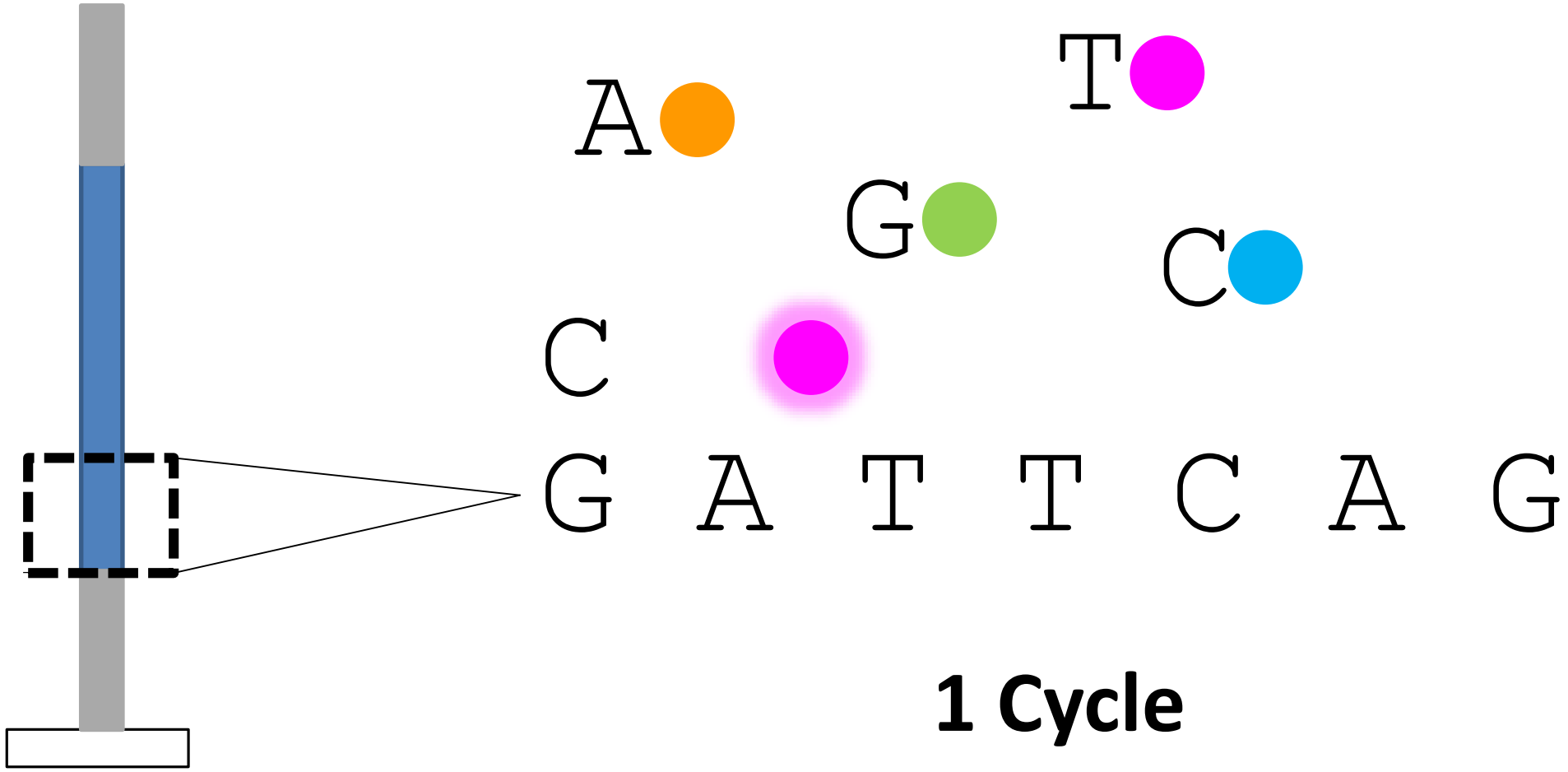
SBS For a Single Molecule



SBS For a Single Molecule

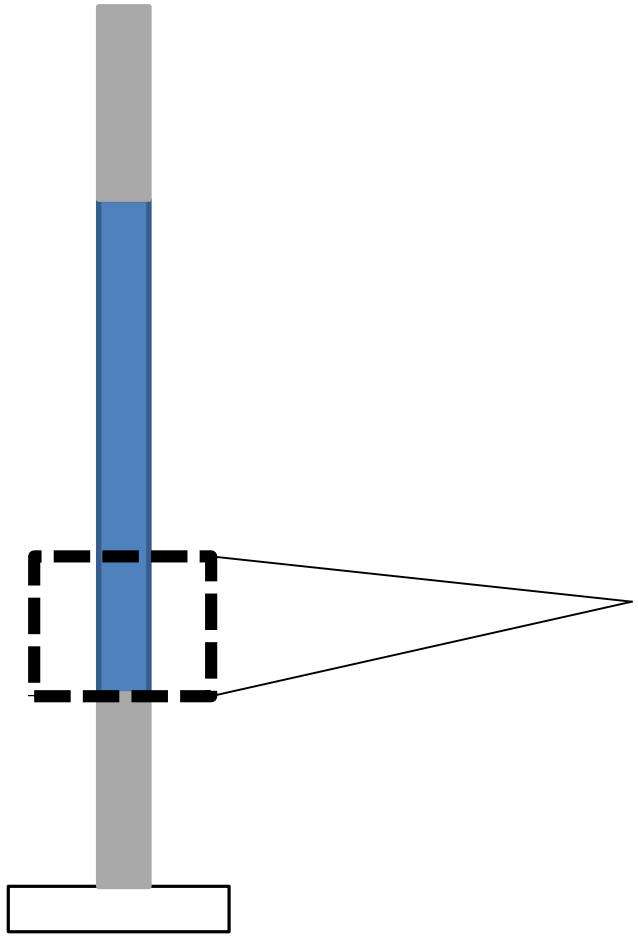


SBS For a Single Molecule



1 Cycle

SBS For a Single Molecule



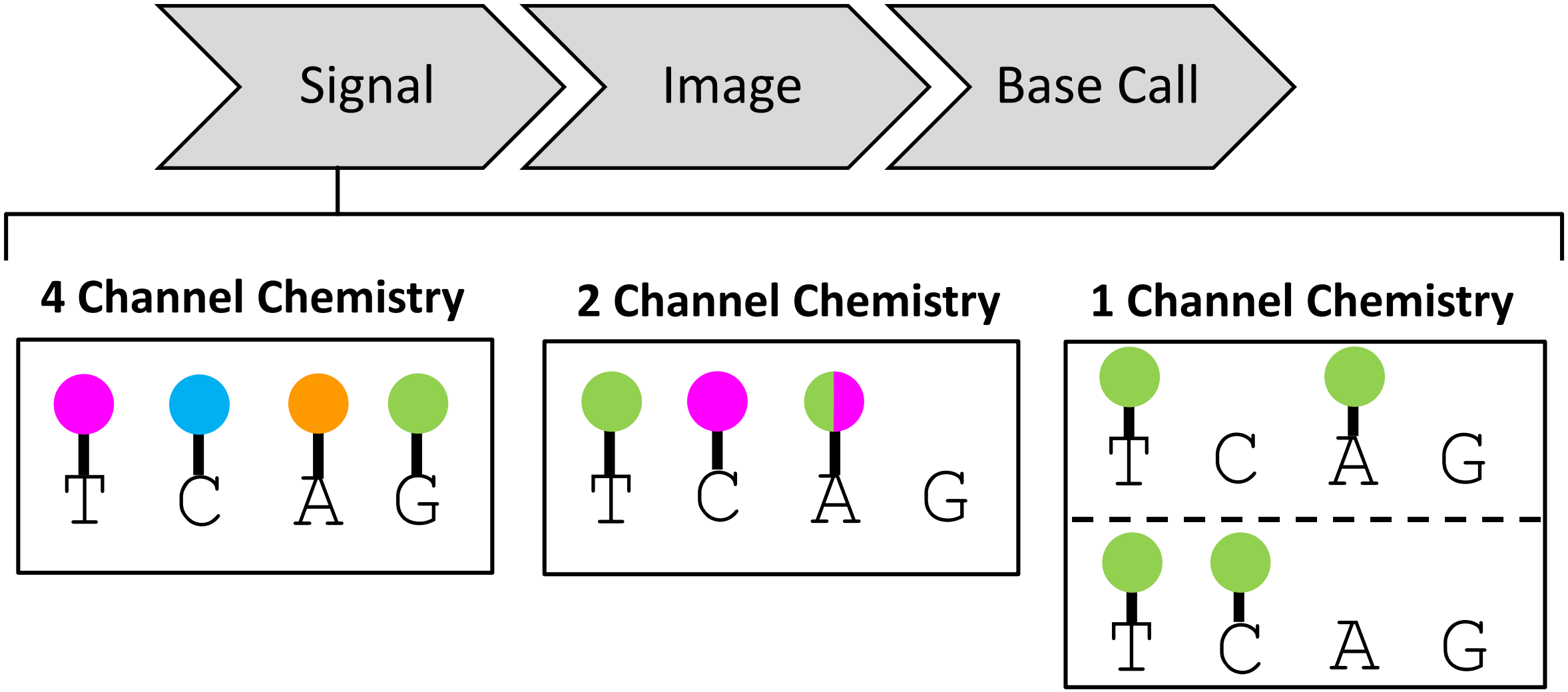
C T A A G T C

G A T T C A G

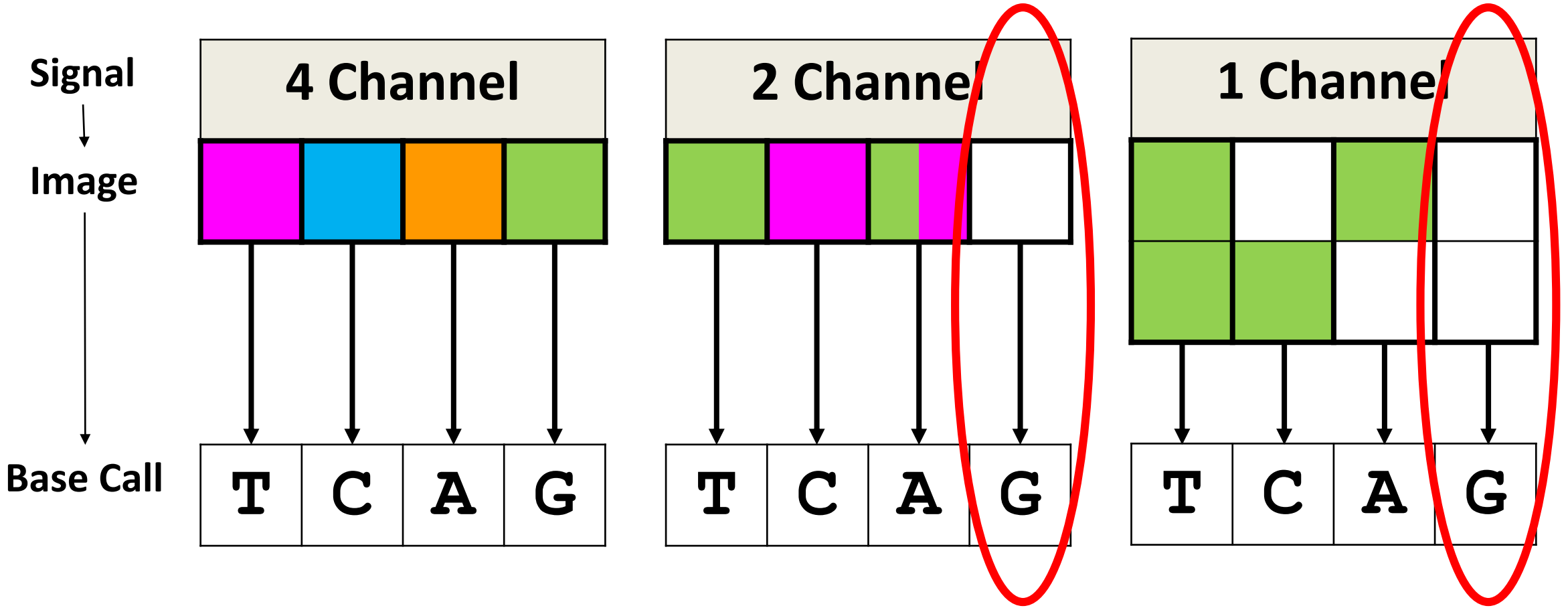
ect…

7 Cycles … Read length

Comparing Chemistry

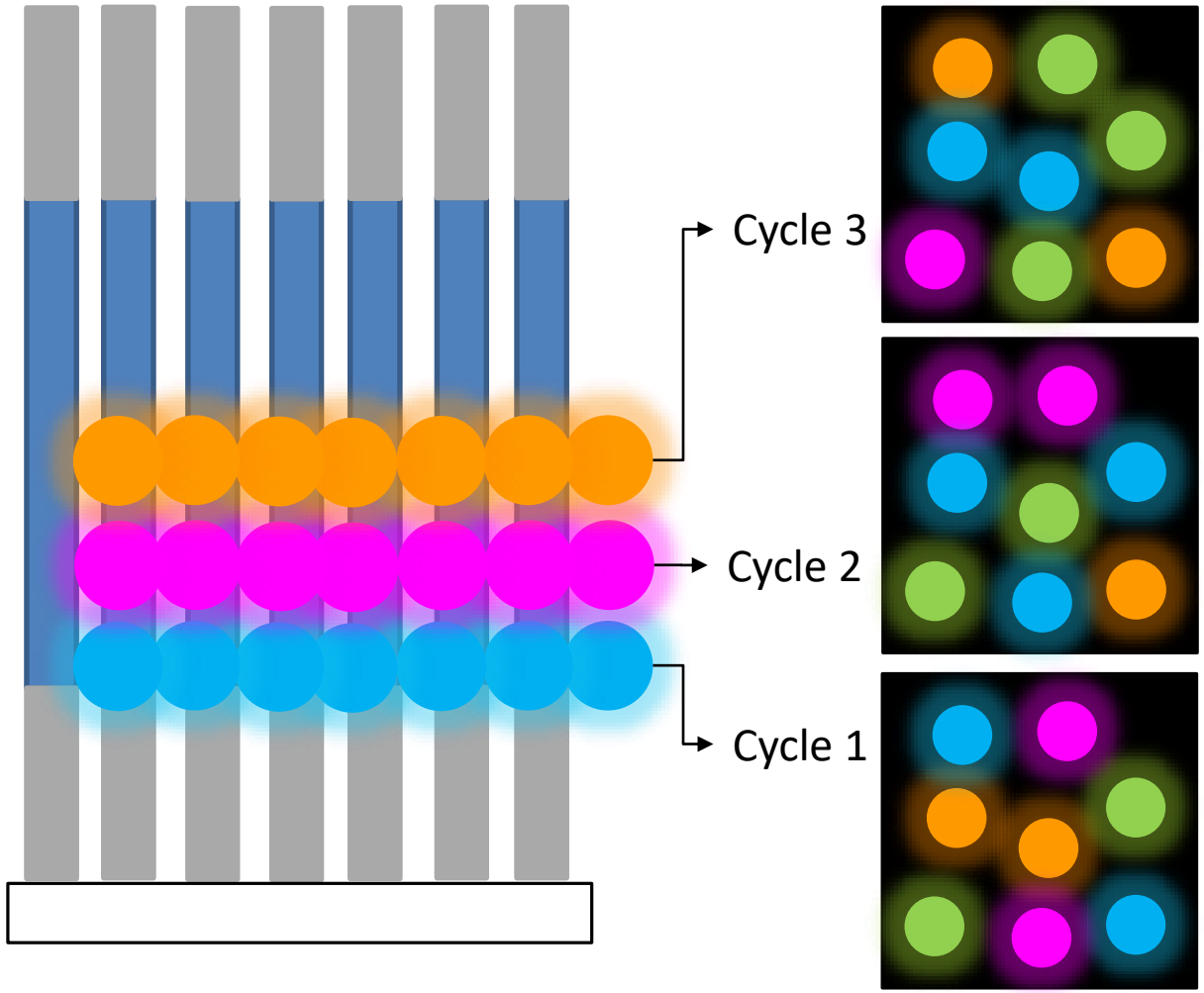


Comparing Chemistry



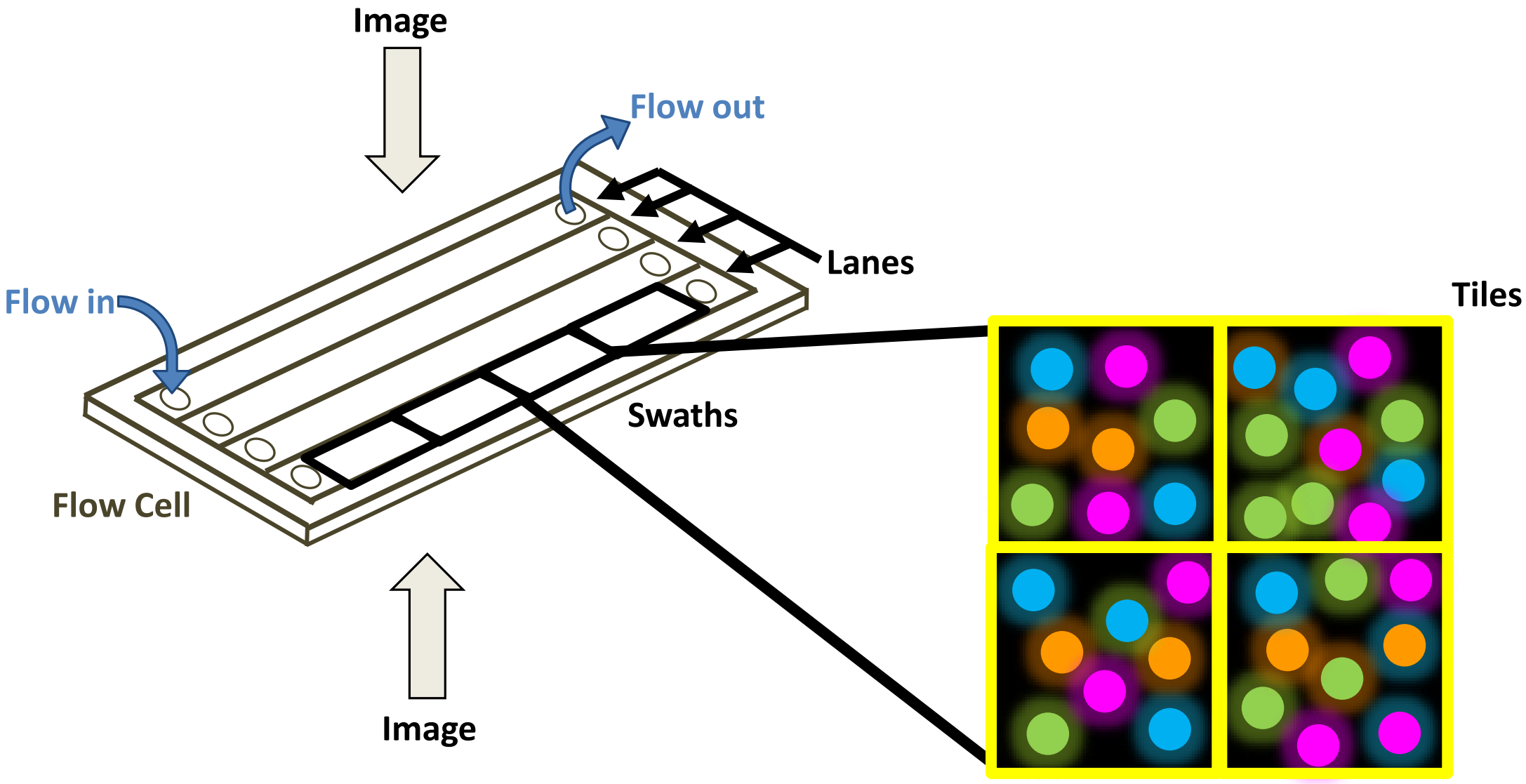
No Signal is interpreted as G

Detecting a Signal

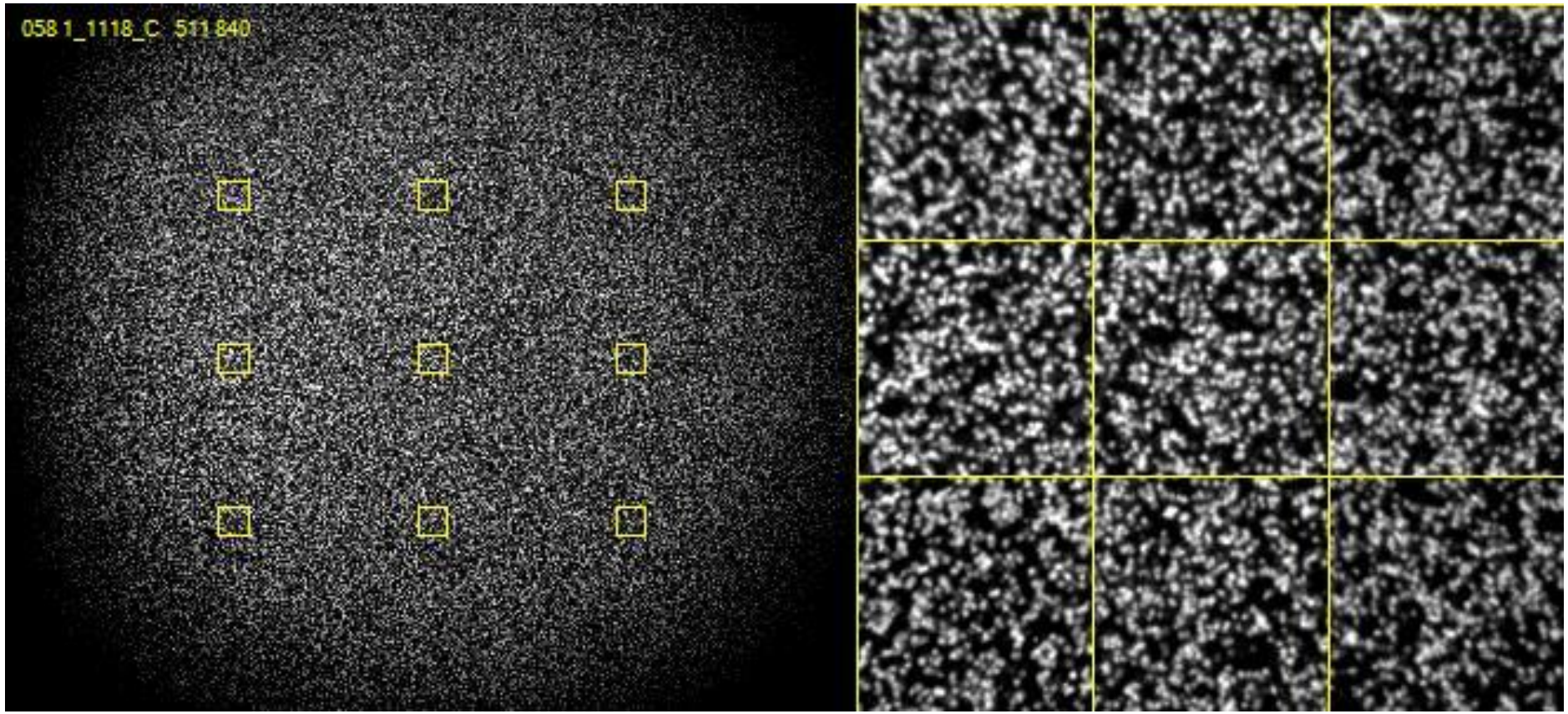


- One Molecule isn’t Enough
- Amplify to generate Cluster
- Multiple Clusters on a Flow Cell

Flow Cell Imaging

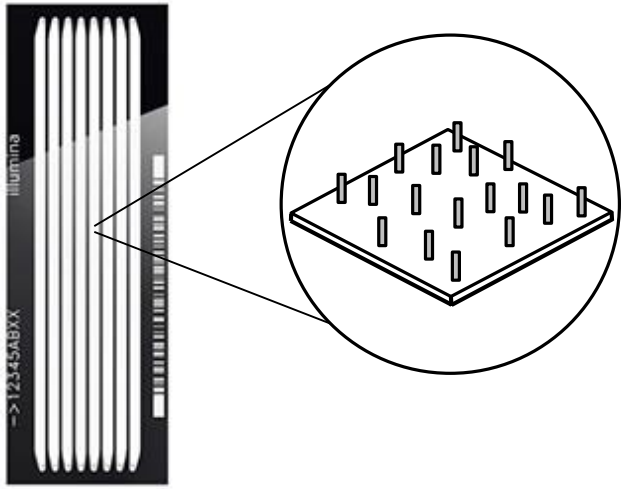


Real Illumina Sequence Data

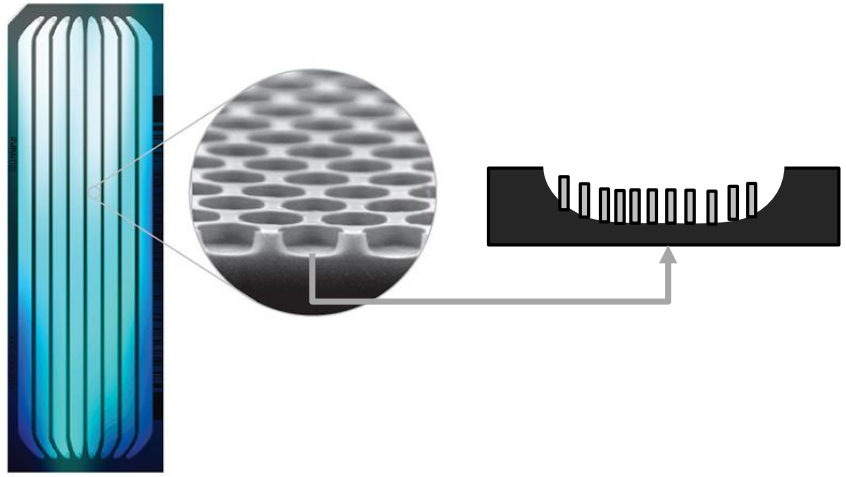


Creating Clusters

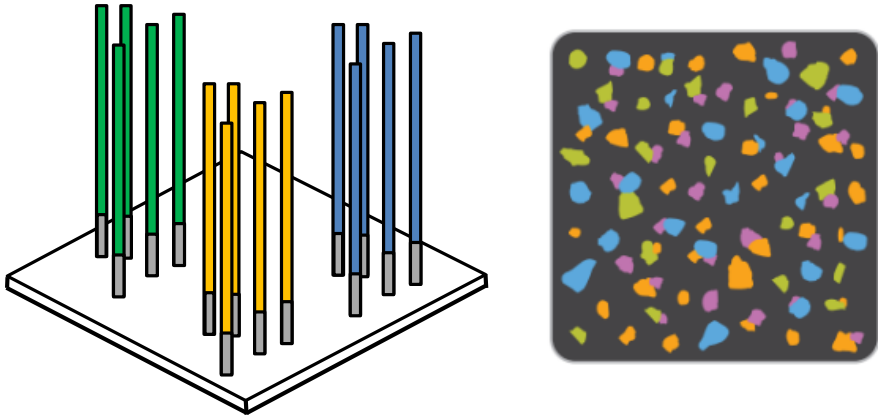
Non Patterned Flow Cell



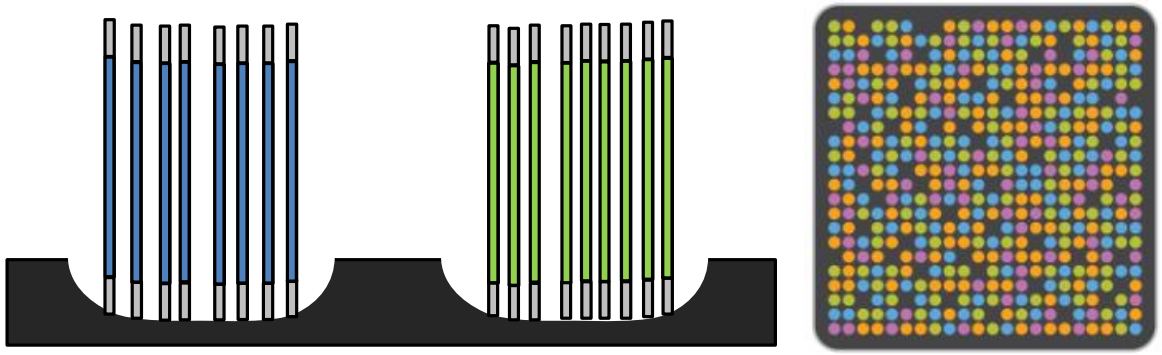
Patterned Flow Cell



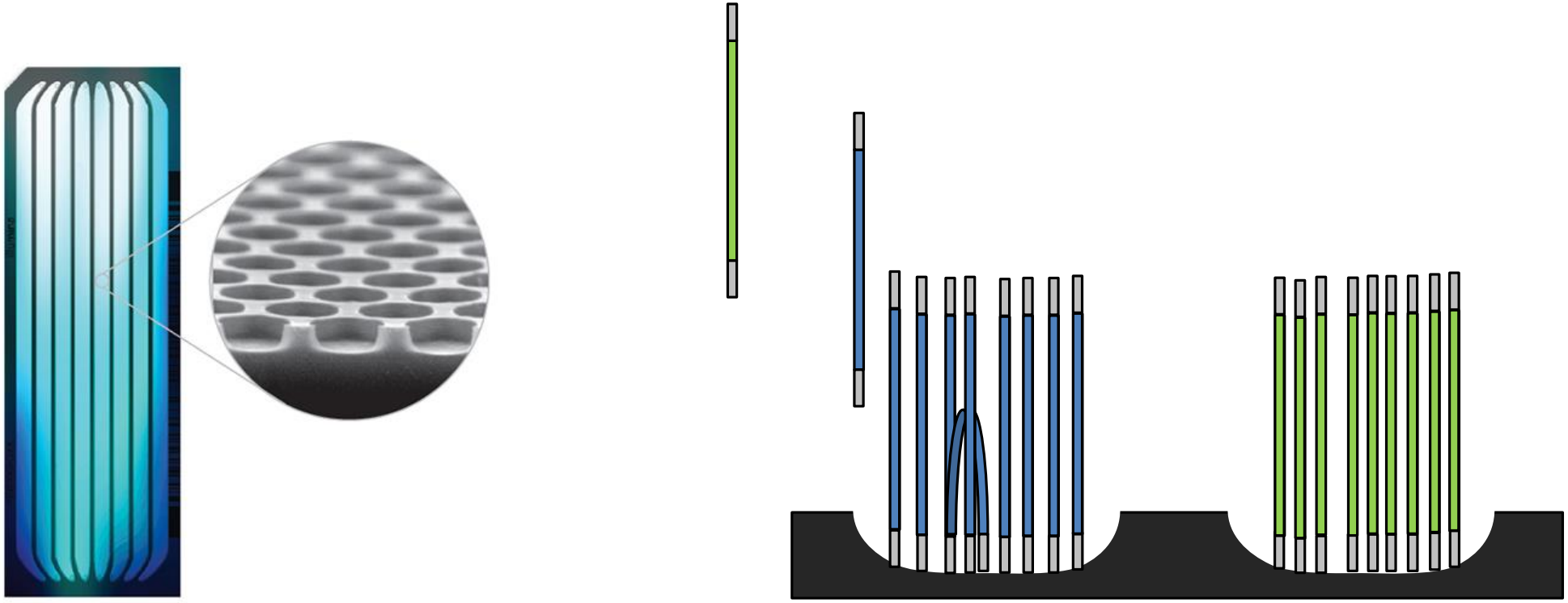
Random Clusters



Pre-defined Clusters



Creating Clusters: Patterned Flow Cells



Single molecule attaches in a nano well → Recombinase Polymerase Bridge Amplification → Cluster of identical (ish) molecules created in a nanowell

Amplification is faster than seeding

Good and Bad things about clusters

Good

- Generates large signal
- Is robust to random mistakes
- Needs a small amount of starting material

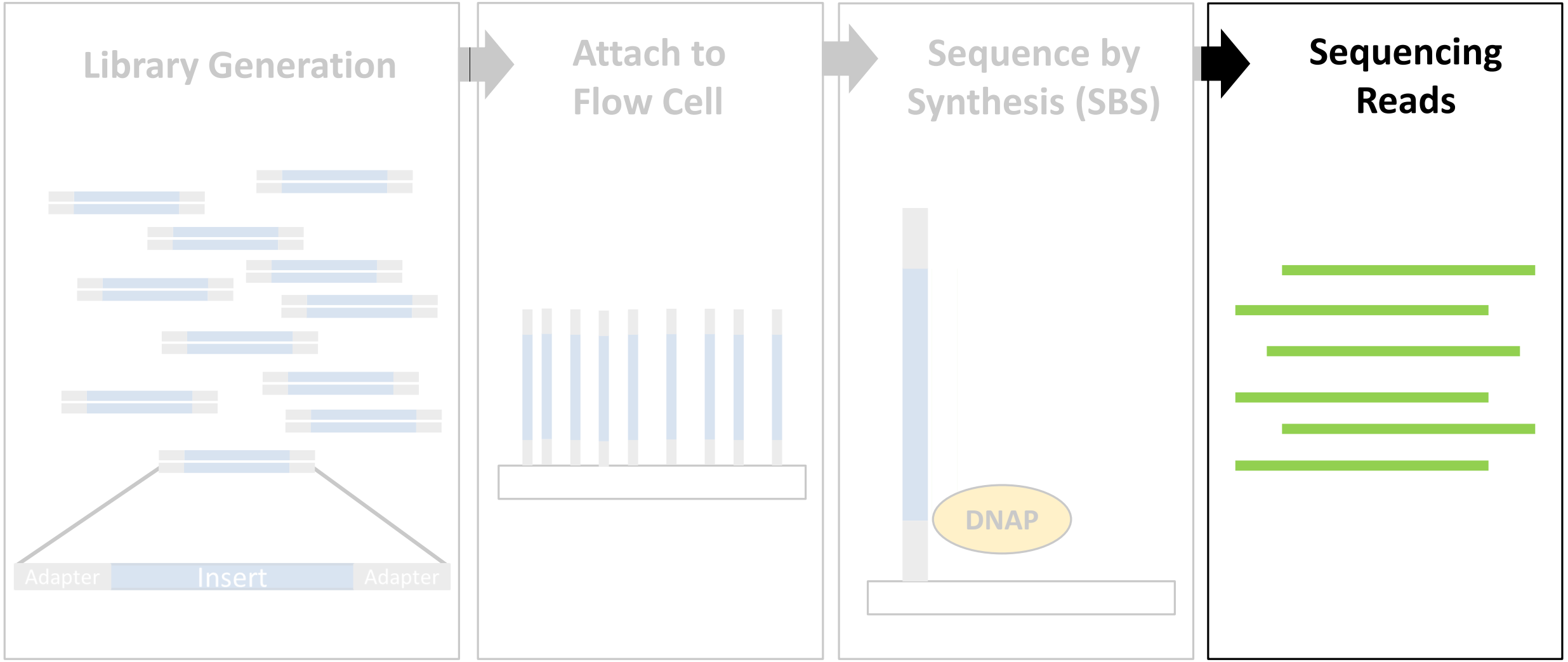
Bad

- Bridging limits length
- Molecules in a cluster get out of sync
  - 2 bases added
  - No bases added
  - Reaction stalls
- Can get mixed signals if clusters overlap (non-patterned)
- Can get re-seeding (patterned)
- Can get index hopping (patterned)

Different sequencers, same chemistry

Sequencer	Number of lanes	Reads per lane	Max read length	Dyes
iSeq 100	1	~4 million	150bp	1
MiniSeq	1	~7 million	150bp	2
MiSeq	1	~20 million	300bp	4
NextSeq	1	~400 million	150bp	2
HiSeq 2xxx	16	~200 million	150bp	4
HiSeq 4xxx	16	~300 million	150bp	4
NovaSeq	8	~2.5 billion	150bp	2

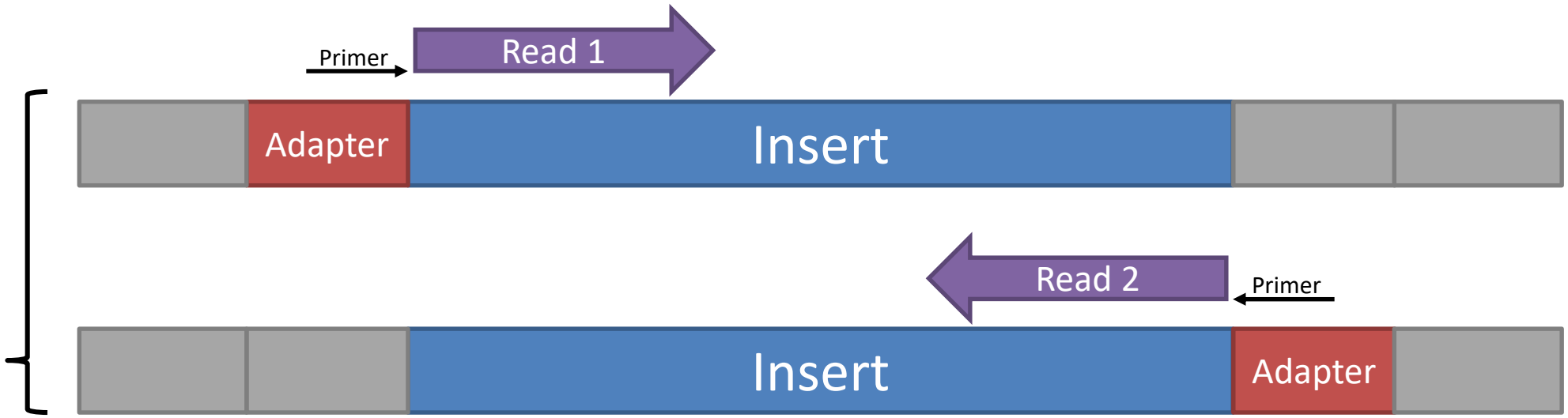
Illumina Sequencing: An Overview



What Reads Do You Get

Single End Run
(one fastq file)

Paired End Run
(two fastq files)



Used to separate libraries – often don't see the fastq file

Barcode Read

FastQ Format Data

```
@HWUSI-EAS611:34:6669YAAXX:1:1:5069:1159 1:N:0:
TCGATAATACCGTTTTTTTCCGTTTGATGTTGATACCATT
+
IIHIIHIIIIIIIIIIIIIIIIIIIIIIIIIHIIIIHIIIII
@HWUSI-EAS611:34:6669YAAXX:1:1:5243:1158 1:N:0:
TATCTGTAGATTTCACAGACTCAAATGTAAATATGCAGAG
+
DF=DBD<BBFGGGGGGGBD@GGGD4@CA3CGG>DDD:D,B
@HWUSI-EAS611:34:6669YAAXX:1:1:5266:1162 1:N:0:
GGAGGAAGTATCACTTCCTTGCCTGCCTCCTCTGGGGCCT
+
:GBGGGGGGGGGDGGDEDGGDGGGGDHHDHGHHGBGG:GG
```

A single FastQ Entry

```
@HWUSI-EAS611:34:6669YAAXX:1:1:5266:1162 1:N:0:
GGAGGAAGTATCACTTCCTTGCCTGCCTCCTCTGGGGCCT
+
:GBGGGGGGGGGDGGDEDGGDGGGGDHHDHGHHGBGG:GG
```

1. Header - starts with @
2. Base calls (can include N or IUPAC codes)
3. Mid-line - starts with + usually empty
4. Quality scores (= Phred Scores)

Illumina Header Sections

```
@HWUSI-EAS611:34:6669YAAXX:5:1:5069:1159 1:N:0:
```

- Starts with @ (required by fastq spec)
- Instrument ID (HWUSI-EAS611)
- Run number (34)
- Flowcell ID (6669YAAXX)
- Lane (5)
- Tile (1)
- X-position (5069)
- Y-position (1159)
- [space]
- Read number (1)
- Was filtered (Y/N) (N) - You wouldn't normally see the Ys
- Control number (0 = no control)
- Sample number (only if demultiplexed using Illumina's software)

Phred Scores

- Start from (p) - the probability that the reported call is incorrect
- Initial transformation to a Phred score - positive integer from floating point
- Phred = $-10 * (\text{int})\log_{10}(p)$
  - p=0.1 Phred = 10
  - p=0.01 Phred = 20
  - p=0.001 Phred = 30

Phred Score Encoding

- Translation of Phred score to single ASCII letter
- Based on standard ASCII table

- Can't translate directly
  - low values are non-printing

- Encode with Sanger System*
  - Phred+33

*Historically also had Illumina = Phred+64

0	NUL	17	C1	33	!	50	2	67	C
1	SOH	18	DC2	34	“	51	3	68	D
2	STX	19	DC3	35	#	52	4	69	E
3	ETX	20	DC4	36	$	53	5	70	F
4	EOT	21	NAK	37	%	54	6	71	G
5	ENQ	22	SYN	38	&	55	7	72	H
6	ACK	23	ETB	39	'	56	8	73	I
7	BEL	24	CAN	40	(	57	9	74	J
8	BS	25	EM	41	)	58	:	75	K
9	HT	26	SUB	42	*	59	;	76	L
10	LF	27	ESC	43	+	60	<	77	M
11	VT	28	FS	44	,	61	=	78	N
12	FF	29	GS	45	-	62	>	79	O
13	CR	30	RS	46	.	63	?	80	P
14	SO	31	US	47	/	64	@	81	Q
15	SI	32	(SPACE)	48	0	65	A	82	R
16	DLE			49	1	66	B	83	S

Phred score encoding

:GBGGGGGGGGGDGGDEDGGDGGGGDHHDHGHHGBGG:GG

: = ASCII 58

Phred33 encoding so Phred = 25

$p = 10^{(25/-10)}$

$p = 0.003$

032	{ }	052	4	072	H	092	\	112	p
033	!	053	5	073	I	093	]	113	q
034	"	054	6	074	J	094	^	114	r
035	#	055	7	075	K	095	_	115	s
036	$	056	8	076	L	096	`	116	t
037	%	057	9	077	M	097	a	117	u
038	&	058	:	078	N	098	b	118	v
039	'	059	;	079	O	099	c	119	w
040	(	060	<	080	P	100	d	120	x
041	)	061	=	081	Q	101	e	121	y
042	*	062	>	082	R	102	f	122	z
043	+	063	?	083	S	103	g	123	{
044	,	064	@	084	T	104	h	124	|
045	-	065	A	085	U	105	i	125	}
046	.	066	B	086	V	106	j	126	~
047	/	067	C	087	W	107	k	127	{ }
048	0	068	D	088	X	108	l	128	Ç
049	1	069	E	089	Y	109	m	129	ü
050	2	070	F	090	Z	110	n	130	é
051	3	071	G	091	[	111	o	131	â

Phred score encoding

:GBGGGGGGGGGDGGDEDGGDGGGGDHHDHGHHGBGG:GG

Symbol	ASCII	Phred	Probability of miscall
:	58	25	$p = 10^{\wedge(25/-10)} = 0.003$
G	71	?	?

BETTER

or

WORSE

032	{ }	052	4	072	H	092	\	112	p
033	!	053	5	073	I	093	]	113	q
034	"	054	6	074	J	094	^	114	r
035	#	055	7	075	K	095	_	115	s
036	$	056	8	076	L	096	`	116	t
037	%	057	9	077	M	097	a	117	u
038	&	058	:	078	N	098	b	118	v
039	'	059	;	079	O	099	c	119	w
040	(	060	<	080	P	100	d	120	x
041	)	061	=	081	Q	101	e	121	y
042	*	062	>	082	R	102	f	122	z
043	+	063	?	083	S	103	g	123	{
044	,	064	@	084	T	104	h	124	|
045	-	065	A	085	U	105	i	125	}
046	.	066	B	086	V	106	j	126	~
047	/	067	C	087	W	107	k	127	{ }
048	0	068	D	088	X	108	l	128	Ç
049	1	069	E	089	Y	109	m	129	ü
050	2	070	F	090	Z	110	n	130	é
051	3	071	G	091	[	111	o	131	â

Phred score encoding

:GBGGGGGGGGGDGGDEDGGDGGGGDHHDHGHHGBGG:GG

Symbol	ASCII	Phred	Probability of miscall
:	58	25	$p = 10^{\wedge(25/-10)} = 0.003$
G	71	38	$p = 10^{\wedge(38/-10)} = 0.00016$

BETTER

032	{ }	052	4	072	H	092	\	112	p
033	!	053	5	073	I	093	]	113	q
034	"	054	6	074	J	094	^	114	r
035	#	055	7	075	K	095	_	115	s
036	$	056	8	076	L	096	`	116	t
037	%	057	9	077	M	097	a	117	u
038	&	058	:	078	N	098	b	118	v
039	'	059	;	079	O	099	c	119	w
040	(	060	<	080	P	100	d	120	x
041	)	061	=	081	Q	101	e	121	y
042	*	062	>	082	R	102	f	122	z
043	+	063	?	083	S	103	g	123	{
044	,	064	@	084	T	104	h	124	|
045	-	065	A	085	U	105	i	125	}
046	.	066	B	086	V	106	j	126	~
047	/	067	C	087	W	107	k	127	{ }
048	0	068	D	088	X	108	l	128	Ç
049	1	069	E	089	Y	109	m	129	ü
050	2	070	F	090	Z	110	n	130	é
051	3	071	G	091	[	111	o	131	â

Phred score encoding

:GBGGGGGGGGGDGGDEDGGDGGGGDHHDHGHHGBGG:GG

Symbol	ASCII	Phred	Probability of miscall
G	71	38	$p = 10^{\wedge(38/-10)} = 0.00016$
B	66	?	?

BETTER

or

WORSE

032	{ }	052	4	072	H	092	\	112	p
033	!	053	5	073	I	093	]	113	q
034	"	054	6	074	J	094	^	114	r
035	#	055	7	075	K	095	_	115	s
036	$	056	8	076	L	096	`	116	t
037	%	057	9	077	M	097	a	117	u
038	&	058	:	078	N	098	b	118	v
039	'	059	;	079	O	099	c	119	w
040	(	060	<	080	P	100	d	120	x
041	)	061	=	081	Q	101	e	121	y
042	*	062	>	082	R	102	f	122	z
043	+	063	?	083	S	103	g	123	{
044	,	064	@	084	T	104	h	124	|
045	-	065	A	085	U	105	i	125	}
046	.	066	B	086	V	106	j	126	~
047	/	067	C	087	W	107	k	127	{ }
048	0	068	D	088	X	108	l	128	Ç
049	1	069	E	089	Y	109	m	129	ü
050	2	070	F	090	Z	110	n	130	é
051	3	071	G	091	[	111	o	131	â

Phred score encoding

:GBGGGGGGGGGDGGDEDGGDGGGGDHHDHGHHGBGG:GG

Symbol	ASCII	Phred	Probability of miscall
G	71	38	$p = 10^{\wedge(38/-10)} = 0.00016$
B	66	33	$p = 10^{\wedge(33/-10)} = 0.0005$

WORSE

032	{ }	052	4	072	H	092	\	112	p
033	!	053	5	073	I	093	]	113	q
034	"	054	6	074	J	094	^	114	r
035	#	055	7	075	K	095	_	115	s
036	$	056	8	076	L	096	`	116	t
037	%	057	9	077	M	097	a	117	u
038	&	058	:	078	N	098	b	118	v
039	'	059	;	079	O	099	c	119	w
040	(	060	<	080	P	100	d	120	x
041	)	061	=	081	Q	101	e	121	y
042	*	062	>	082	R	102	f	122	z
043	+	063	?	083	S	103	g	123	{
044	,	064	@	084	T	104	h	124	|
045	-	065	A	085	U	105	i	125	}
046	.	066	B	086	V	106	j	126	~
047	/	067	C	087	W	107	k	127	{ }
048	0	068	D	088	X	108	l	128	Ç
049	1	069	E	089	Y	109	m	129	ü
050	2	070	F	090	Z	110	n	130	é
051	3	071	G	091	[	111	o	131	â

Aligned Data – BAM files

- Expanded file containing alignment data as well as everything in the fastq file

- Two sections
  - Header (list of reference seqs and how the file was created)
  - Alignments (details of the alignments and sequences)

- Need special programs to read, normally 'samtools'

BAM Header

```
[andrewss@headstone Sample_lane1]$ samtools view -H lane1000_TTAGGC_test_L001_R1_GRCm38_hisat2.bam


@HD     VN:1.0  SO:unsorted
@SQ     SN:1    LN:195471971
@SQ     SN:10   LN:130694993
@SQ     SN:11   LN:122082543
@SQ     SN:12   LN:120129022
@SQ     SN:13   LN:120421639
…etc…

@PG     ID:hisat2       PN:hisat2       VN:2.1.0        CL:"/bi/apps/hisat2/2.1.0/hisat2-align-s --
wrapper basic-0 --dta --sp 1000,1000 -p 7 -t --phred33-quals -x
/bi/scratch/Genomes/Mouse/GRCm38/Mus_musculus.GRCm38 --known-splicesite-infile
/bi/scratch/Genomes/Mouse/GRCm38/Mus_musculus.GRCm38.90.hisat2_splices.txt -U /tmp/17469.unp"
```

BAM Alignments

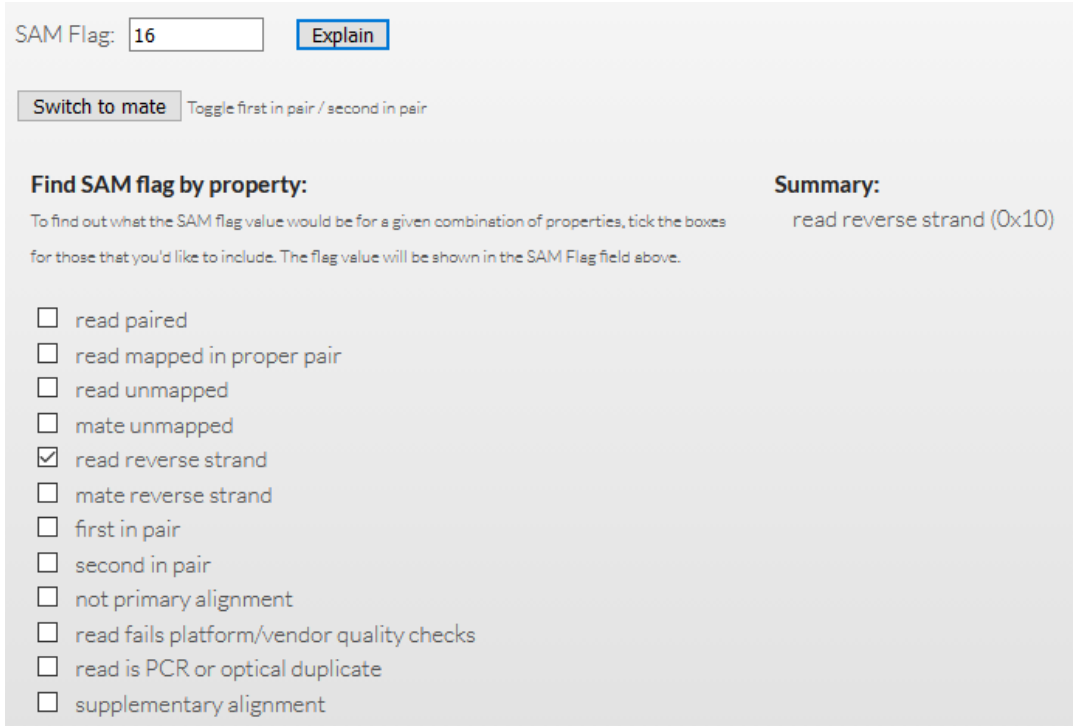
```
[andrewss@headstone Sample_lane1]$ samtools view
lane6045_TTAGGC_Col4_contr_L001_R1_trimmed_GRCm38_hisat2.bam | less

HWI-D00436:394:CBGLBANXX:1:1101:1222:1861       16      chr18      57944851        60      50M     *
0       0       AAAAGATCTCTTGATTTAGAATTTTCTCTCAAATGTGAGGGACTTTTATN
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGBA<=#      AS:i:-2 XN:i:0  XM:i:1  XO:i:0  XG:i:0
NM:i:1  MD:Z:49G0       YT:Z:UU NH:i:1
```

- Sections

1.	Sequence name	HWI-D00436:394:CBGLBANXX:1:1101:1222:1861
2.	Alignment Flags	16
3.	Reference sequence name	chr18
4.	Start position	5794485
5.	Mapping Quality (Phred)	60
6.	Alignment (CIGAR) string	50M
7.	Paired sequence name	*
8.	Paired sequence position	0
9.	Total insert length	0
10.	Called Bases	AAAAGATCTCTTGATTTAGAATTTTCTCTCAAATGTGAGGGACTTTTATN
11.	Base Quality String (Phred 33)	GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGBA<=#
12.	Other Tags	

BAM Flags

- A 12-bit binary number with a set of TRUE/FALSE values
  1. Sequence is paired end
  2. All reads from this template are aligned
  3. This read didn't align
  4. The paired read didn't align
  5. The read aligned in the reverse orientation
  6. The paired read aligned in the reverse orientiation
  7. This is the first read
  8. This is the second read
  9. This is not the best alignment for this read
  10. This read failed upstream QC
  11. This read is a duplicate
  12. This is part of a chimeric alignment



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

What can QC tell us?

Babraham Bioinformatics

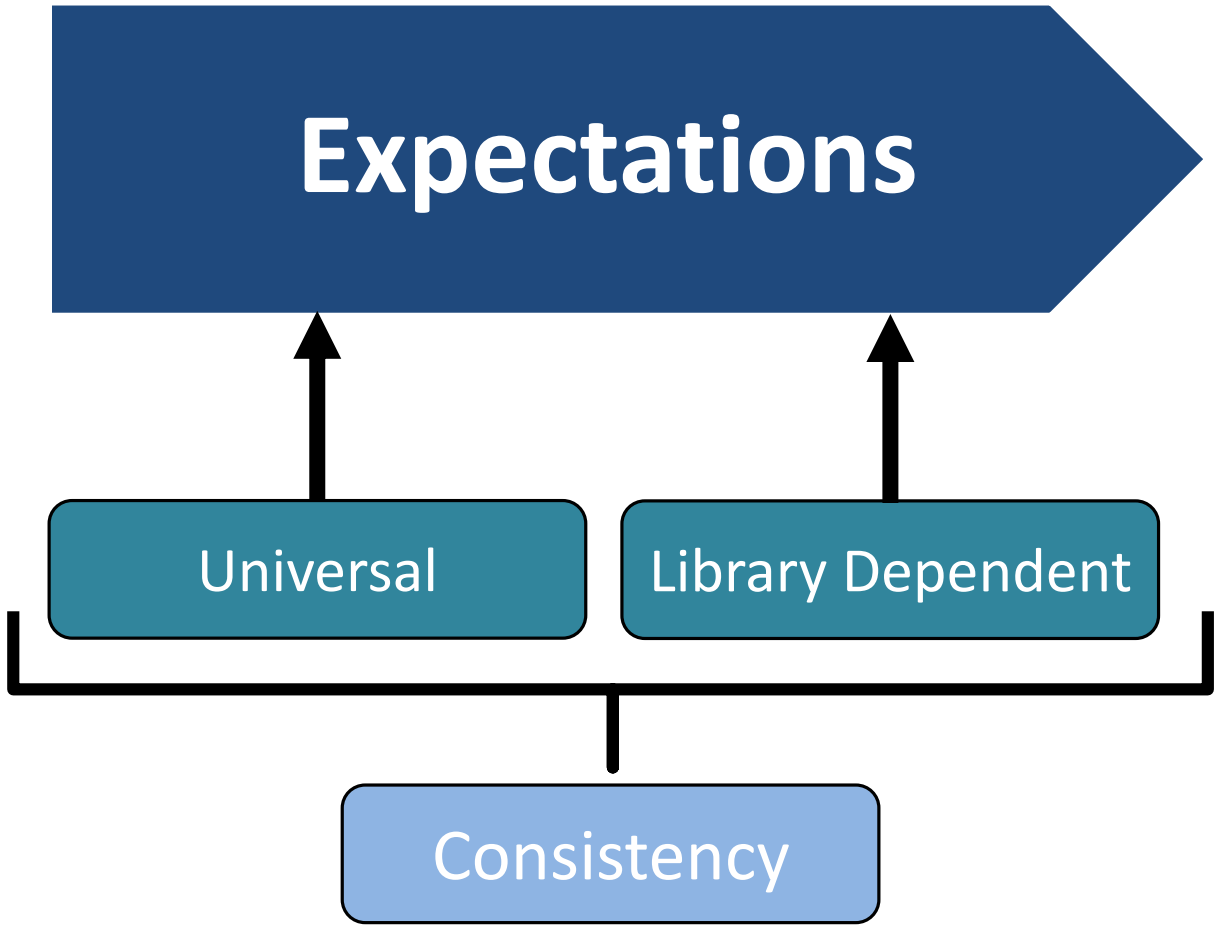
QC metrics can we work with

- Per Base Quality Scores
- Library composition
  - Base level
  - Sequence Level
  - Known sequences
- Mapping statistics
- Downstream Quantitation Values

Context is Key for QC

QC should be about what you expect and what you see

Context is Key for QC



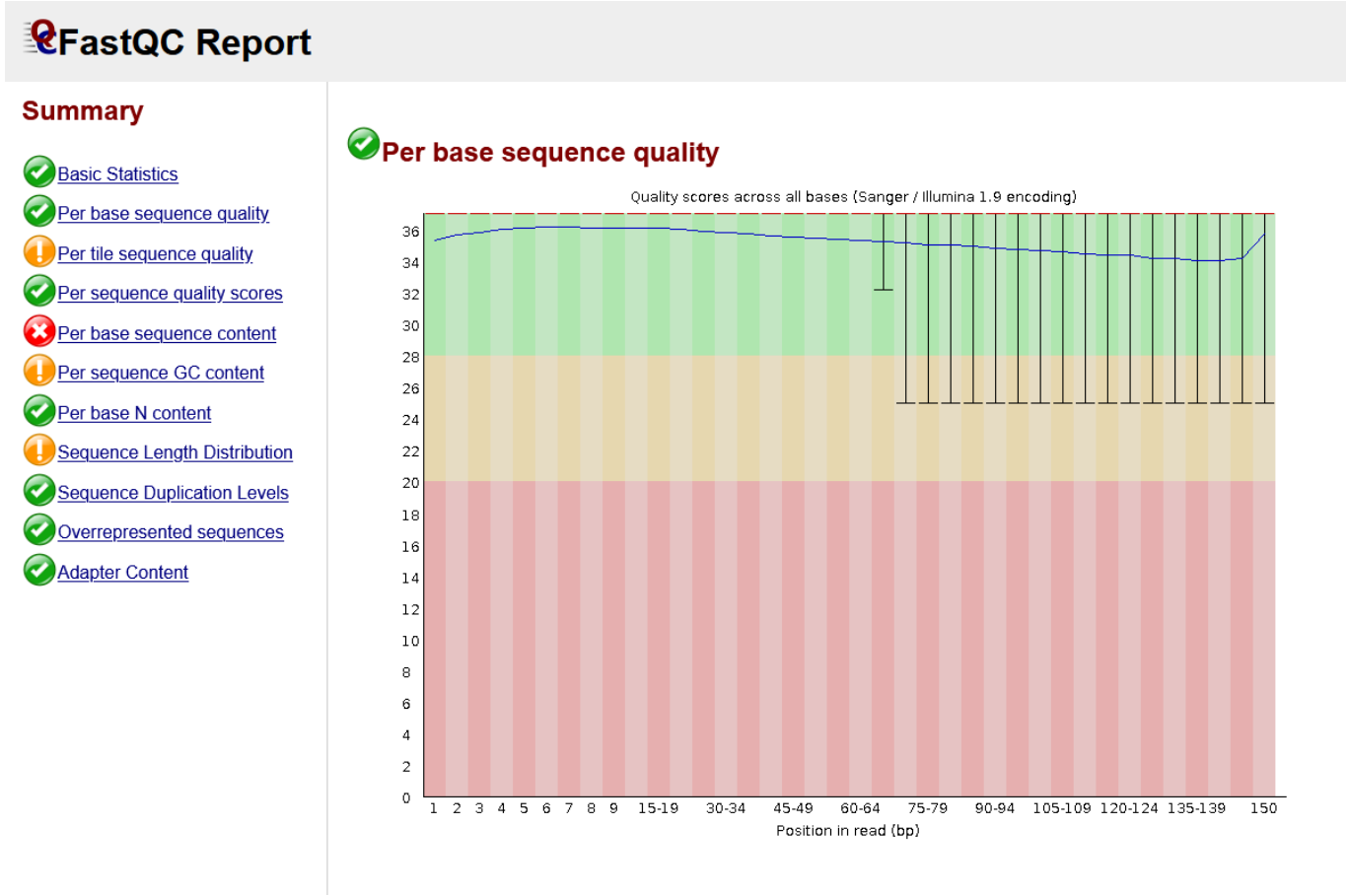
Individual Library:

Replicate Libraries:

Some Software Packages for Sequence QC

Babraham Bioinformatics

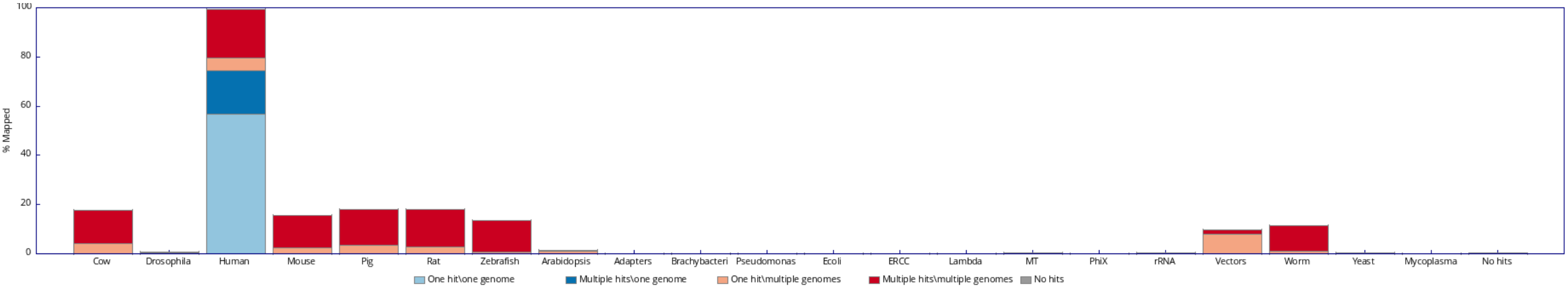
FastQC



- Reads raw fastq files
- Performs multiple checks
  - Pass/warn/fail
  - Compares to genomic library
- HTML Report

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

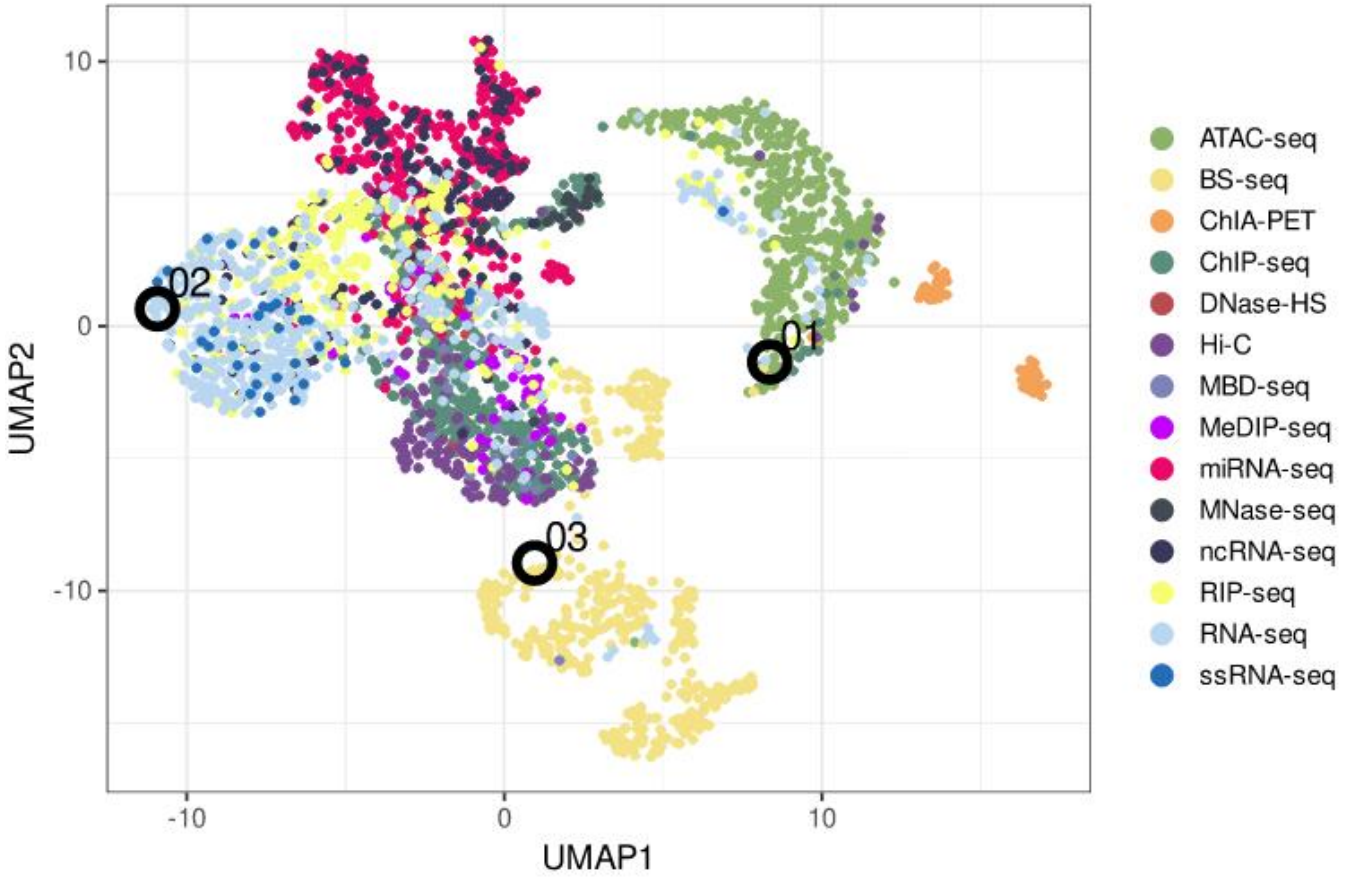
FastQ Screen



- Reads fastq files
- Maps against a range of species / contaminants
- Identifies unexpected sequences in your library

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

Librarian

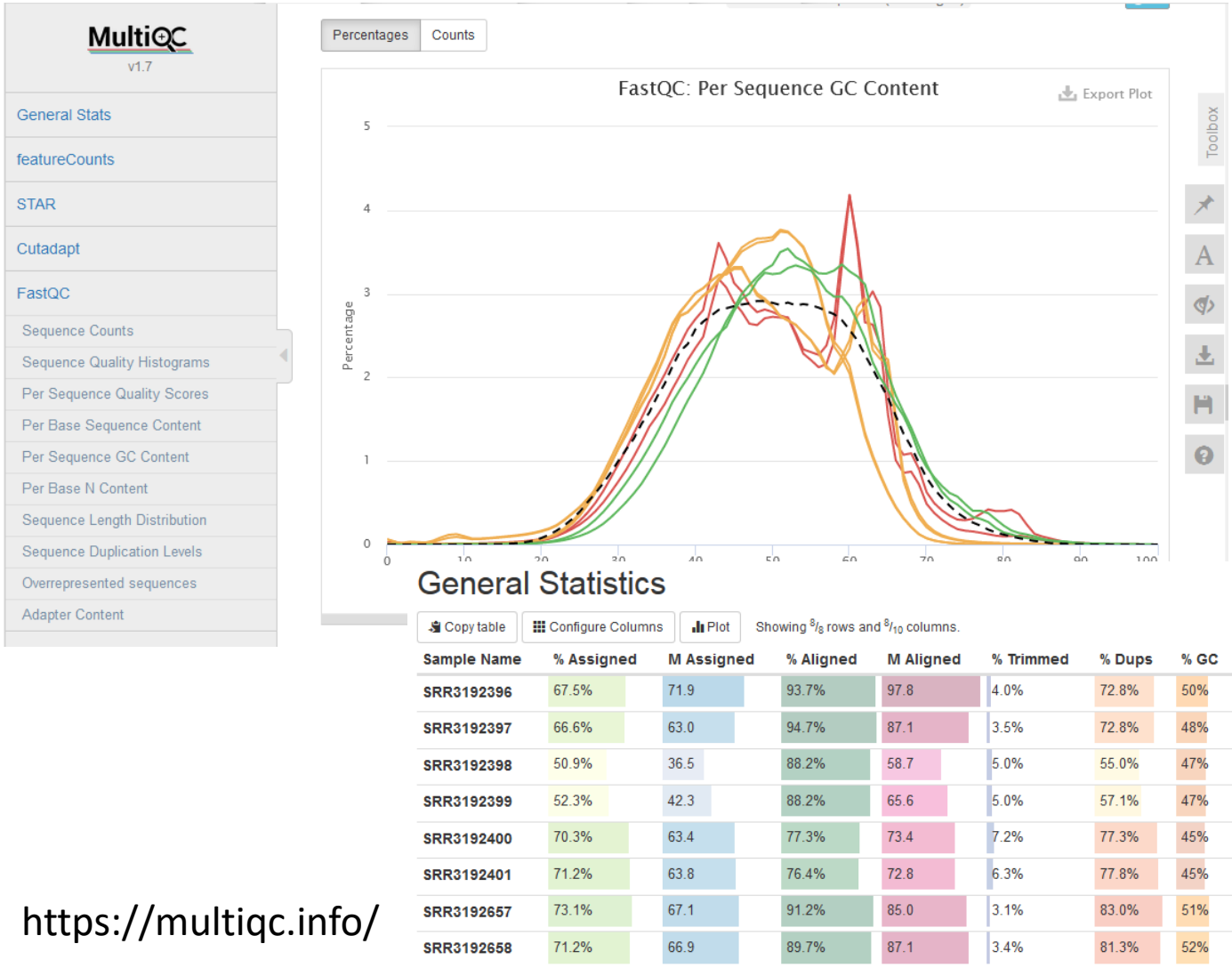


- Reads FastQ files
- Looks at the base composition
- Predict library type

Under development

https://github.com/DesmondWillowbrook/Librarian

MultiQC



- Aggregates QC information from multiple samples
- Large number of programs supported
- Combined HTML report

https://multiqc.info/

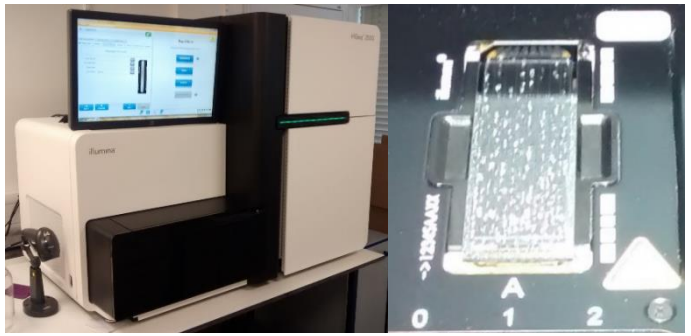
Assessing Universal Metrics

Babraham Bioinformatics

Universal QC Metrics

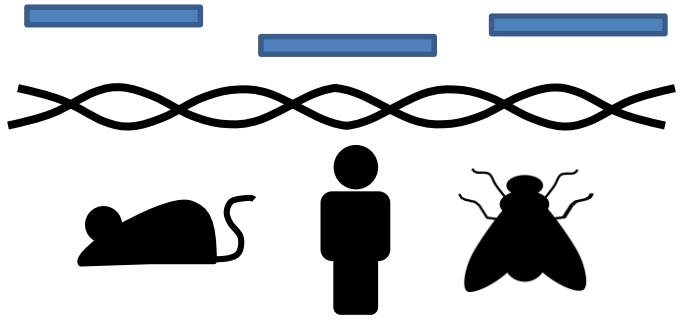
- Demultiplexing

- Base Call Quality

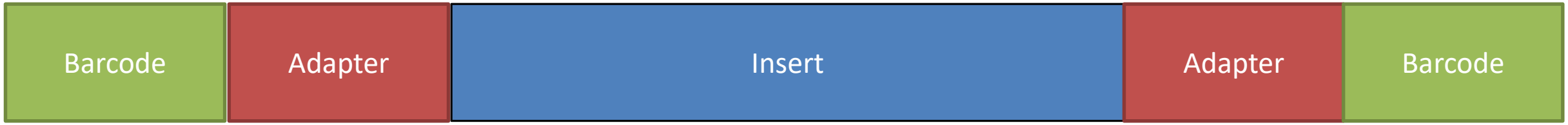


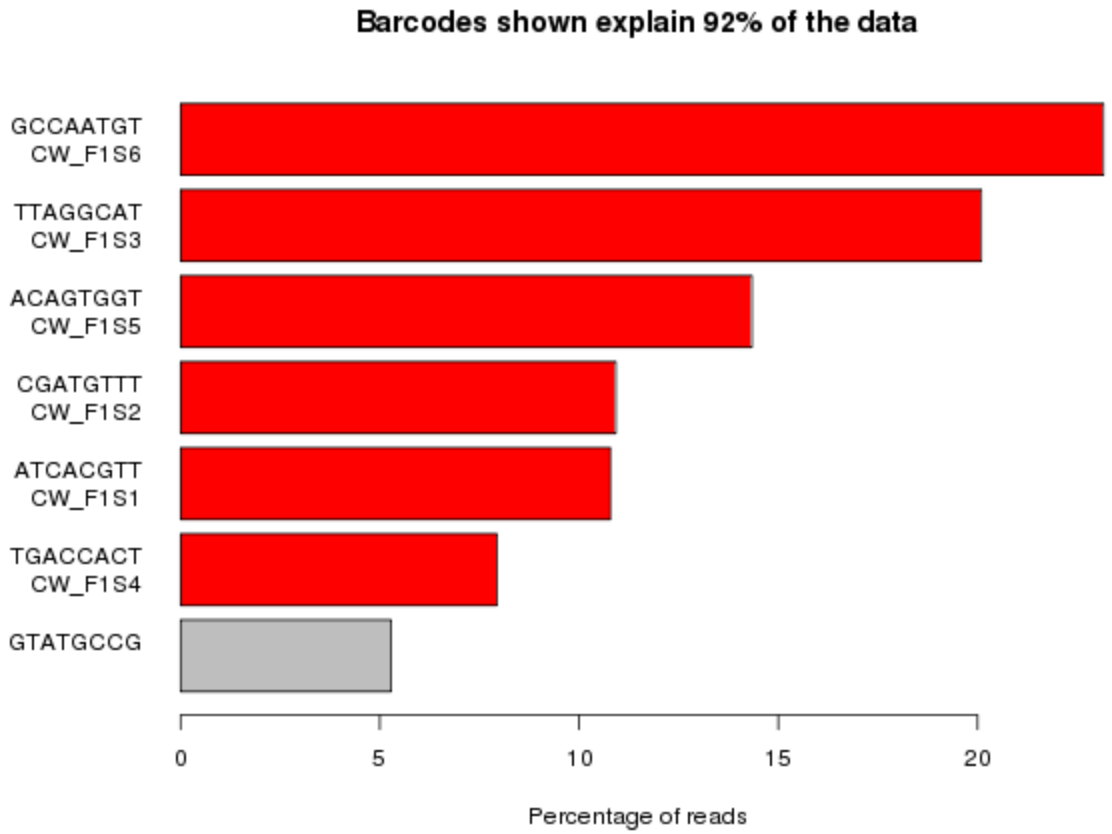
- Adapter Content

- Mapping Quality

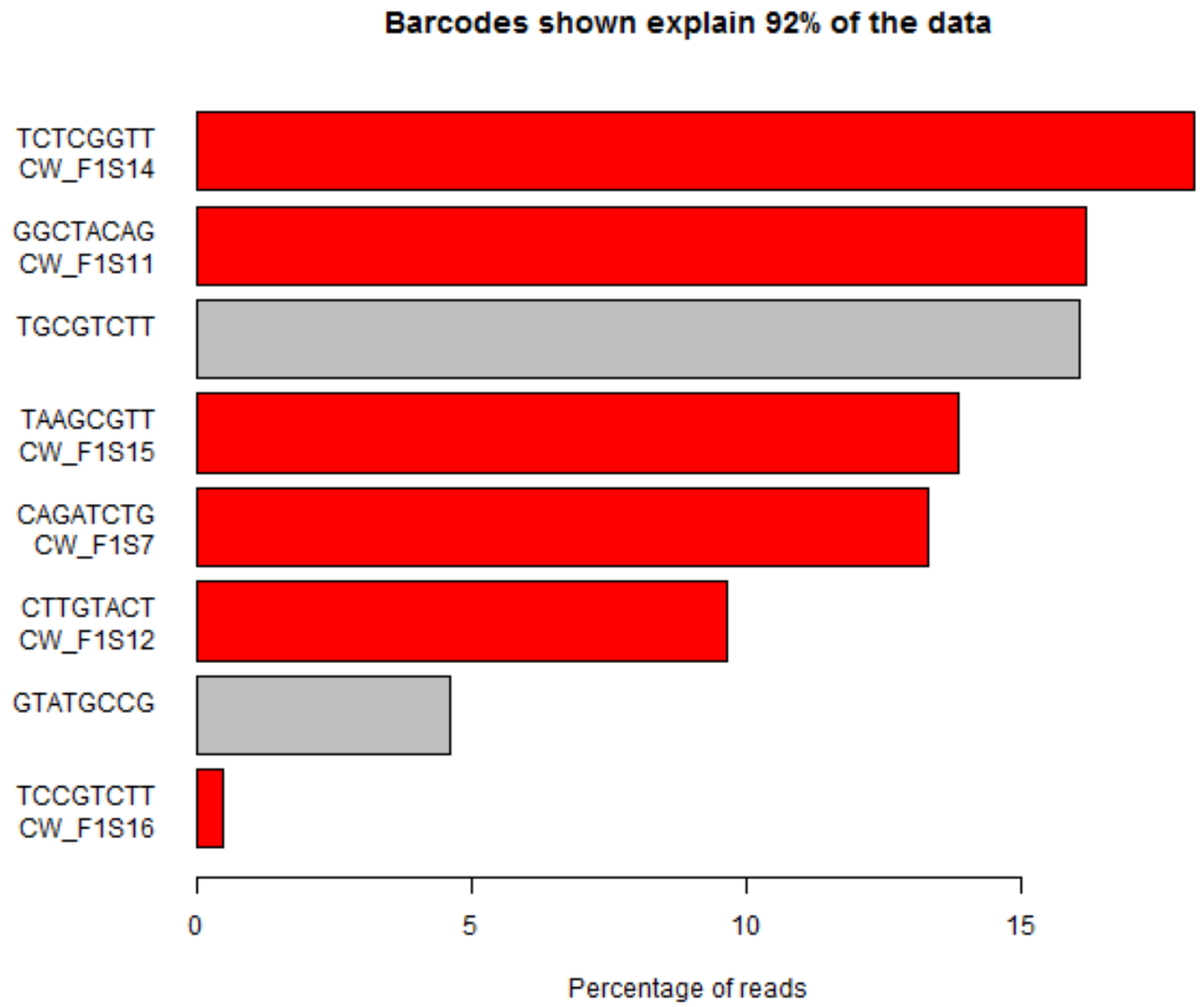


Demultiplexing: Barcode Sequences

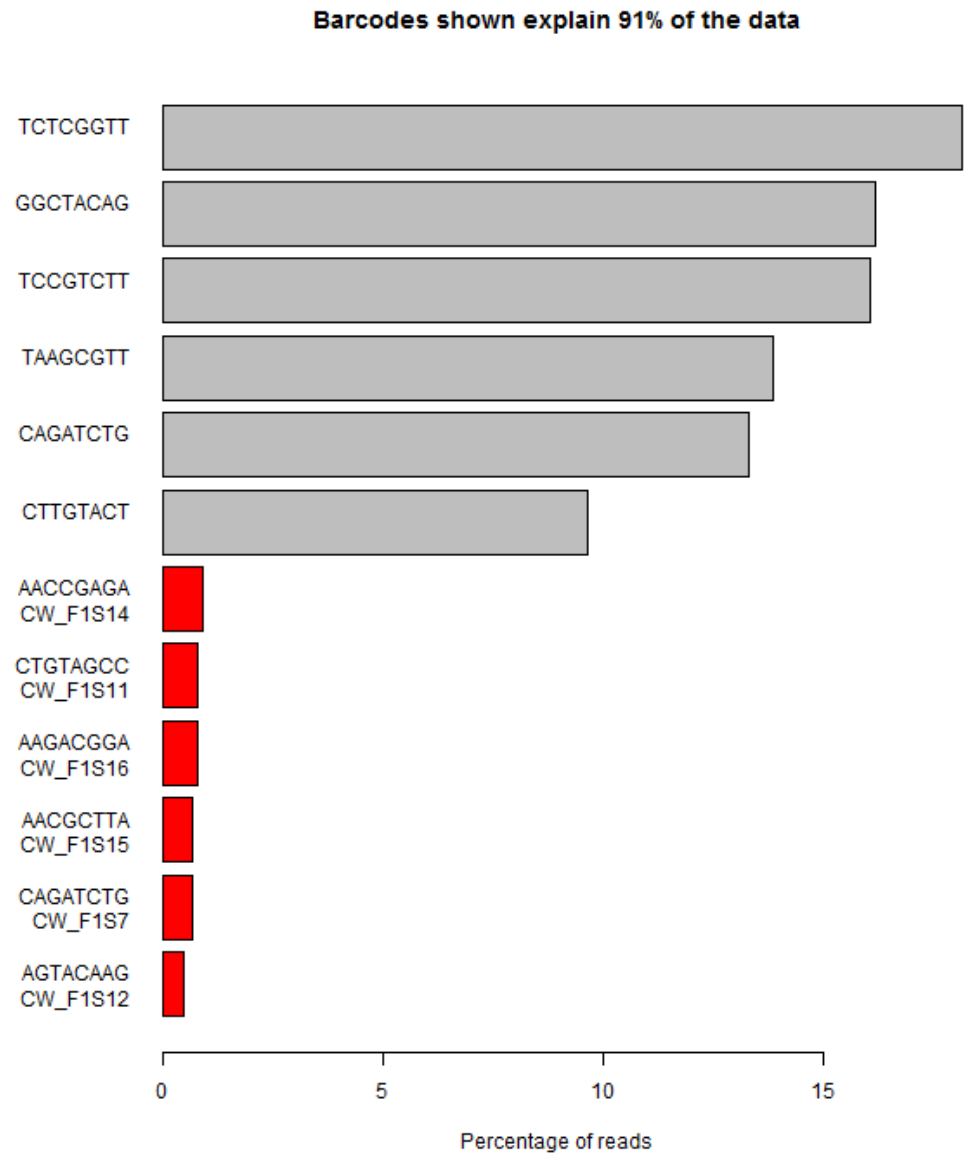




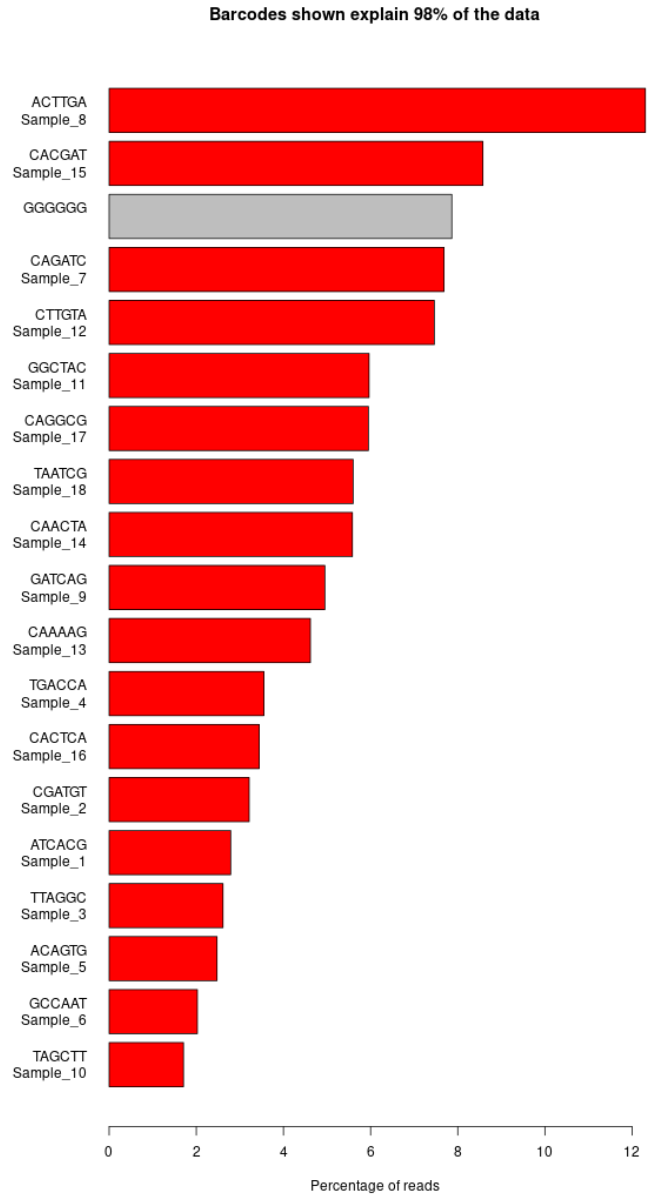
Demultiplexing: Barcode Sequences



Demultiplexing: Barcode Sequences



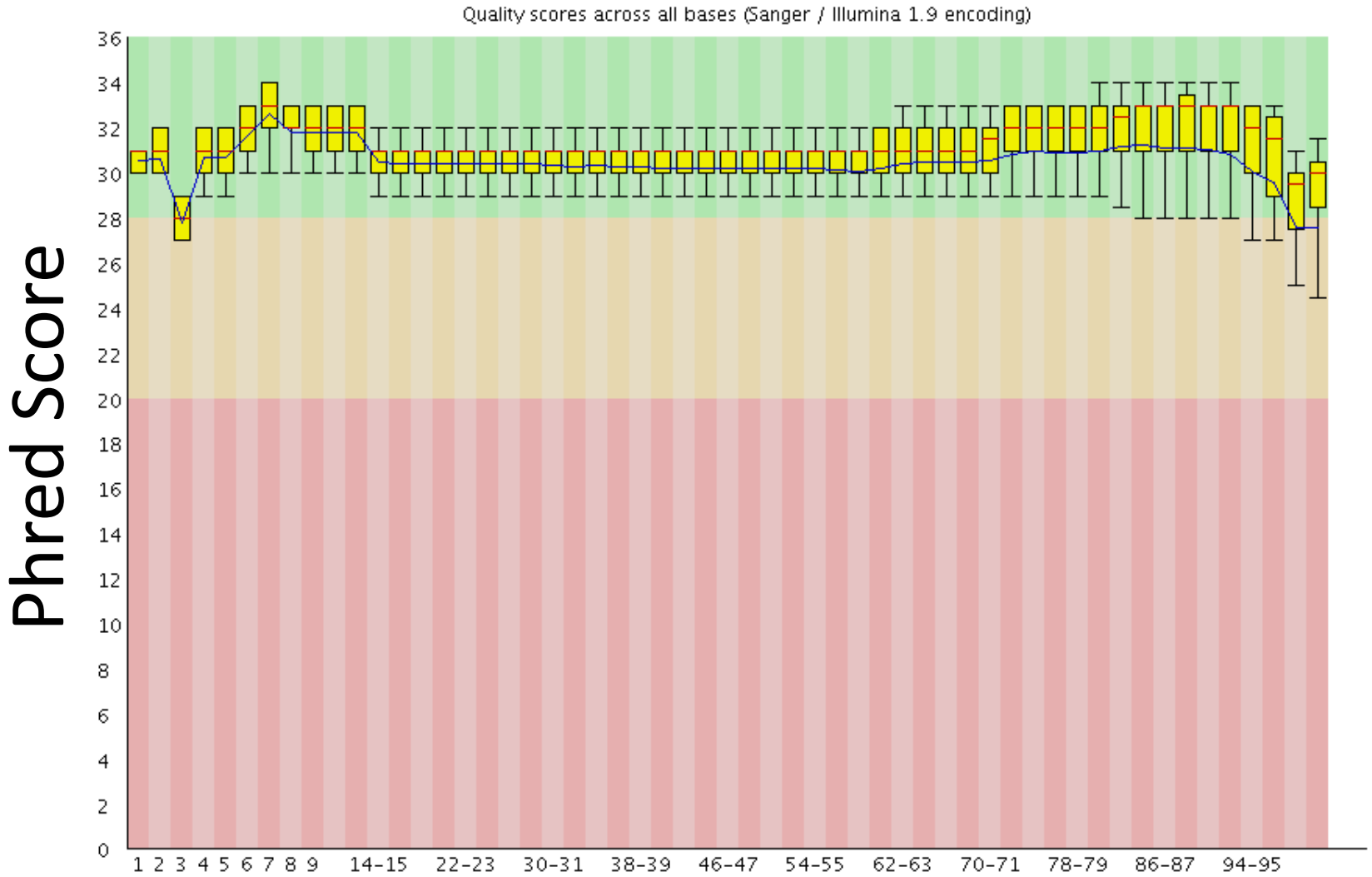
Demultiplexing: Barcode Sequences



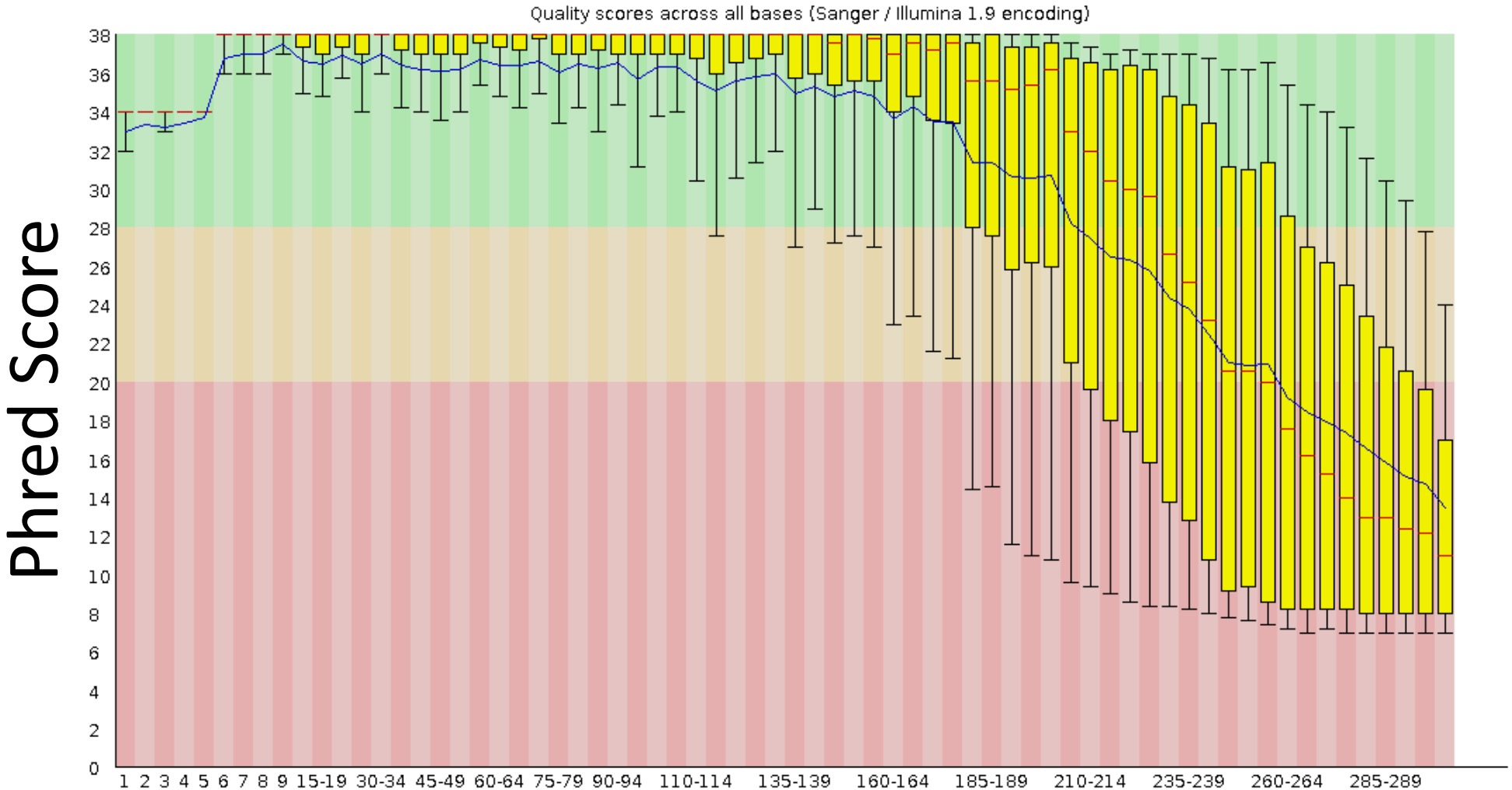
FastQC Report

Per base sequence quality

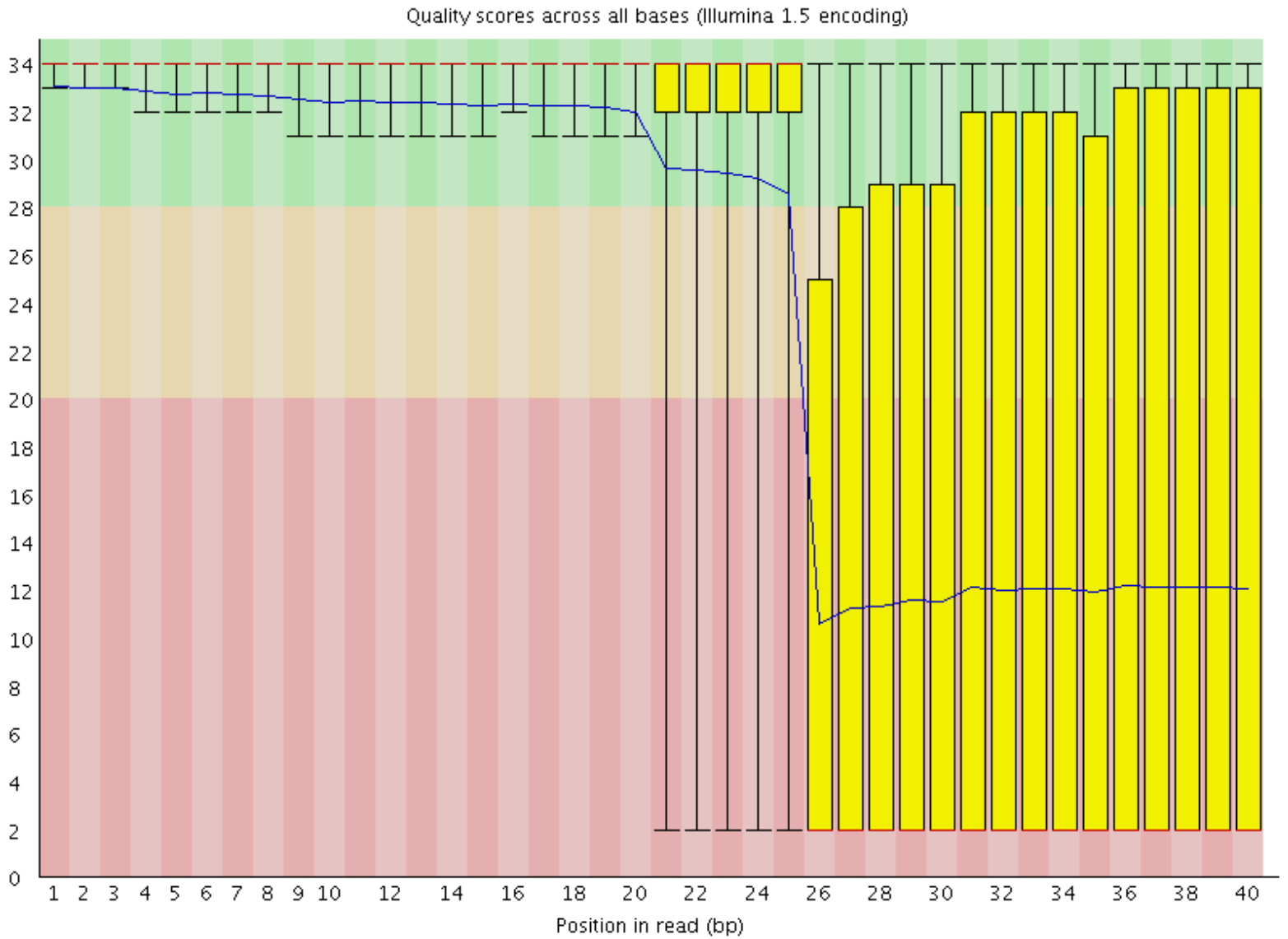
Base Call Qualities – Per Cycle



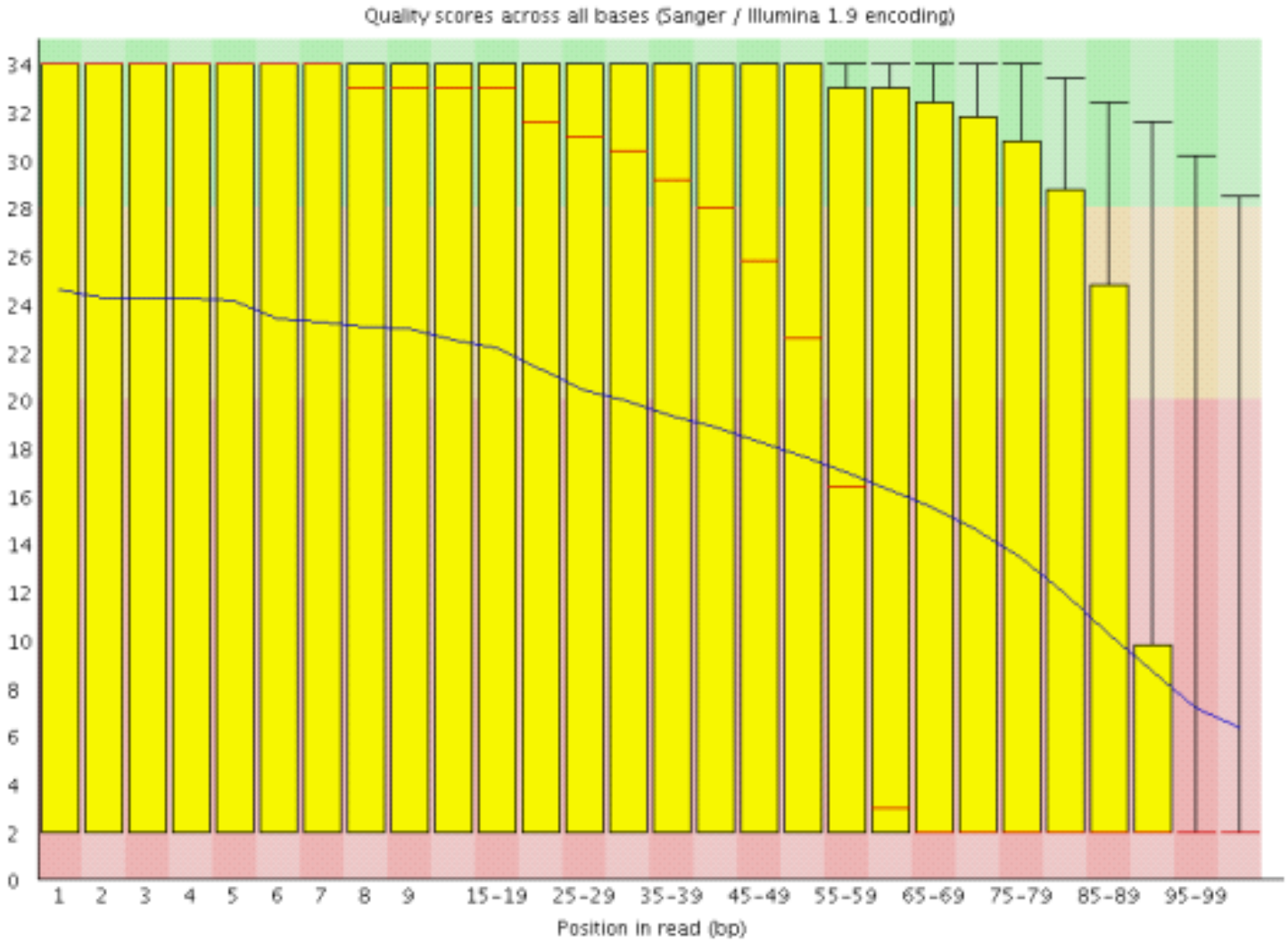
Base Call Qualities – Per Cycle



Base Call Qualities – Per Cycle



Diagnosing Poor Base Call Qualities

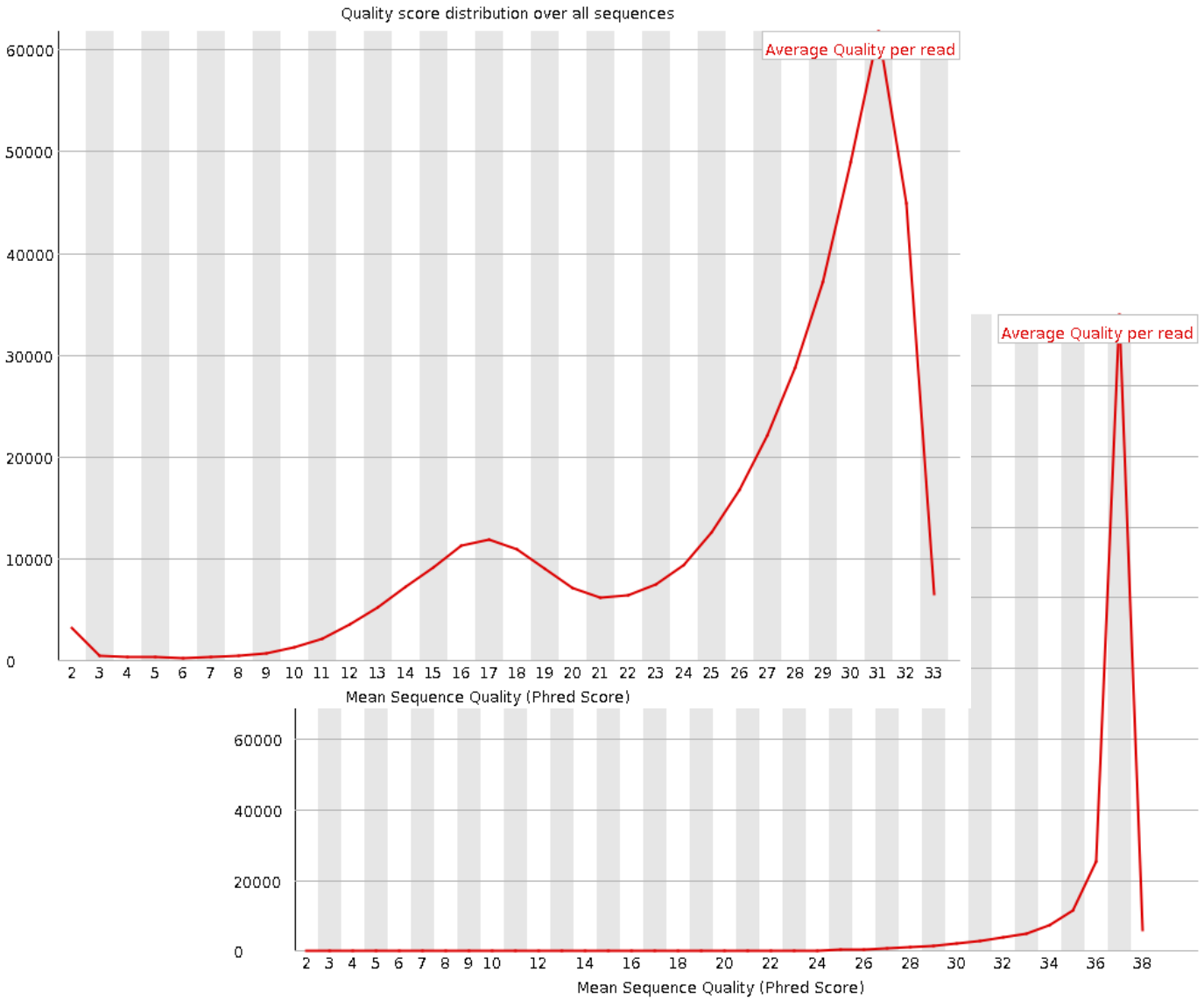


- Not everything is bad
- Can see identify why some parts are bad and others aren't?
- May help to fix future runs

Per-Read Quality

FastQC Report

Per sequence quality scores

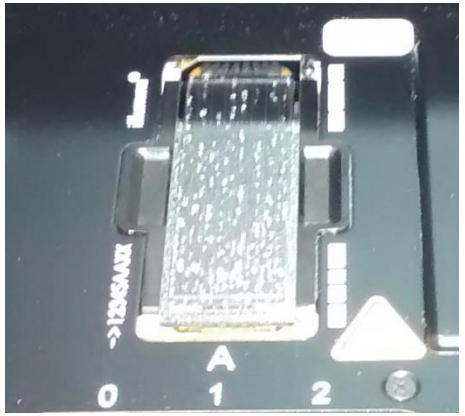


- Are all reads equally affected?
- Is there a subset of reads which are always poor whilst others are good?

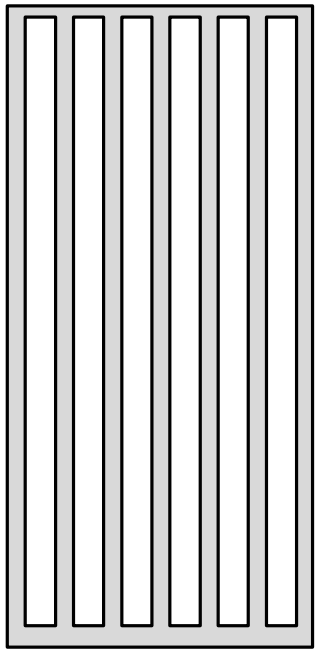
FastQC Report

Per tile sequence quality

Positional Quality

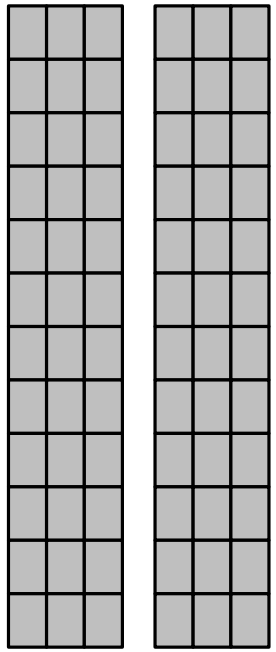


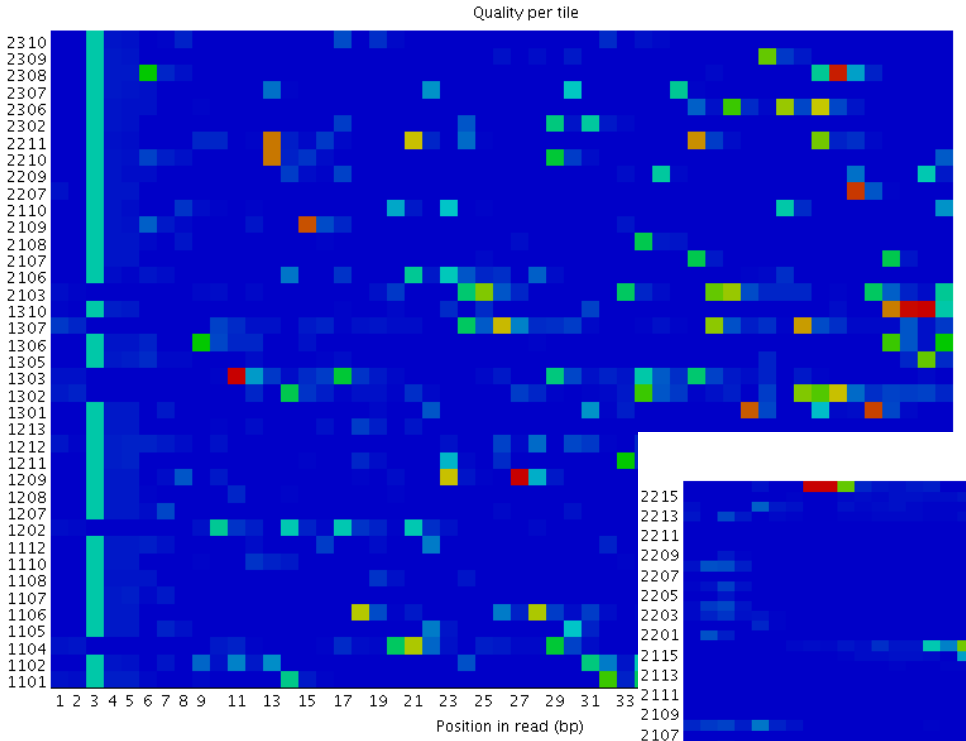
Lanes



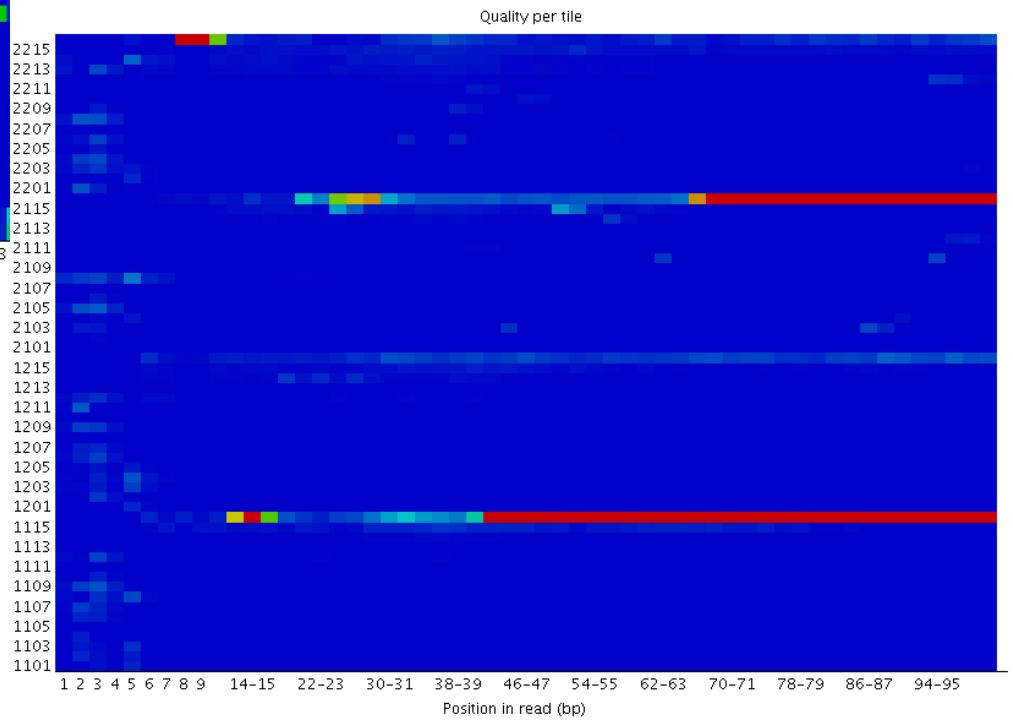
Tiles

Top Bottom



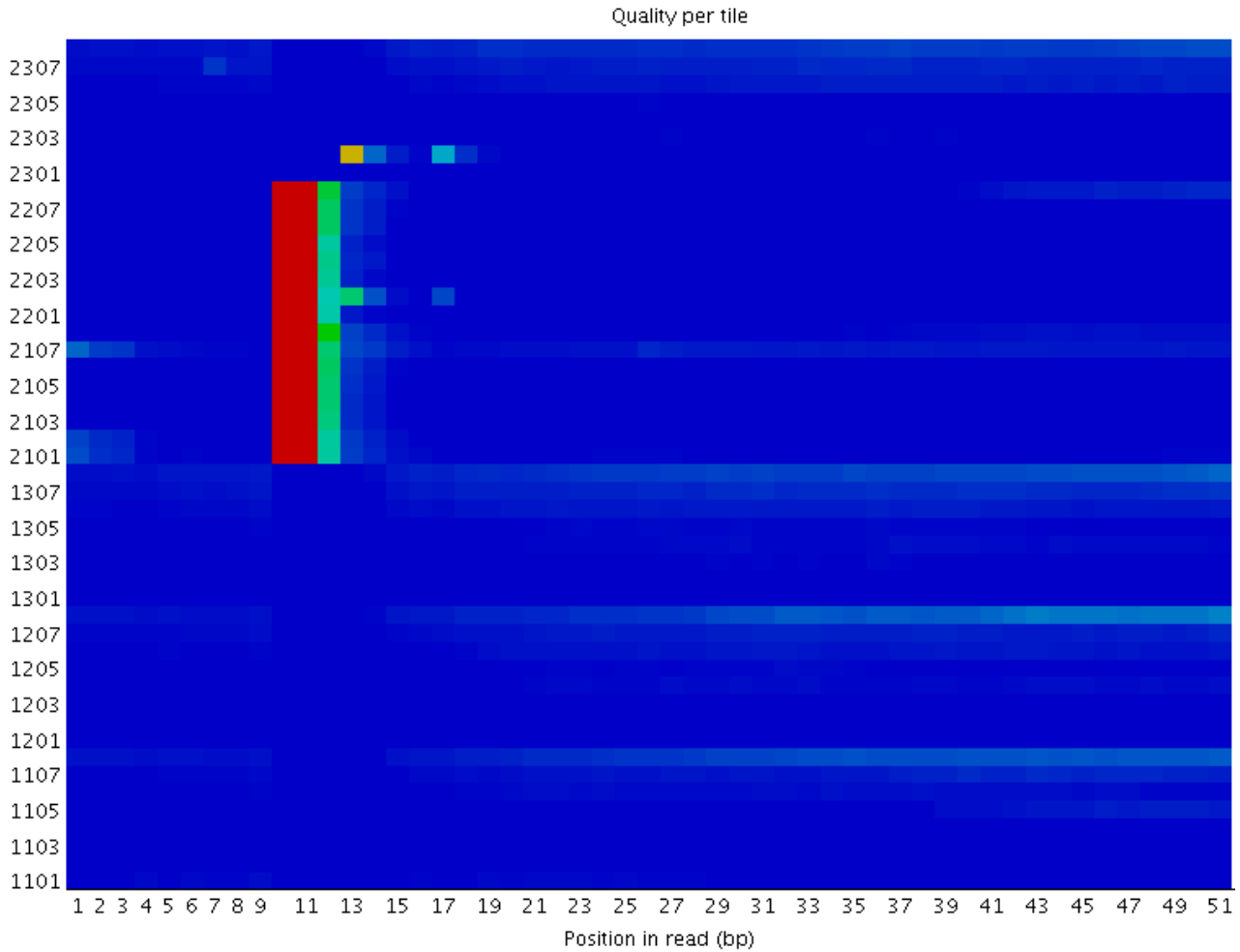


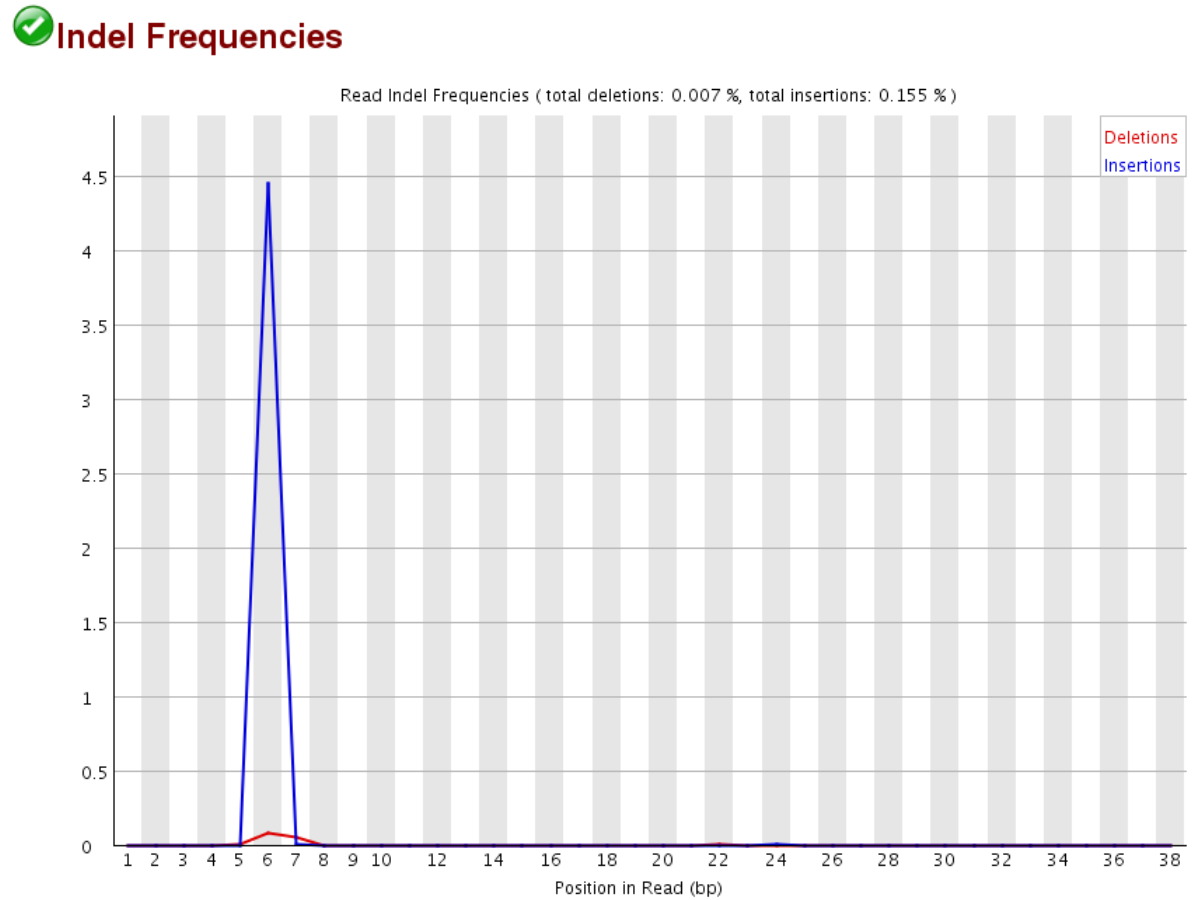
Random Patterning



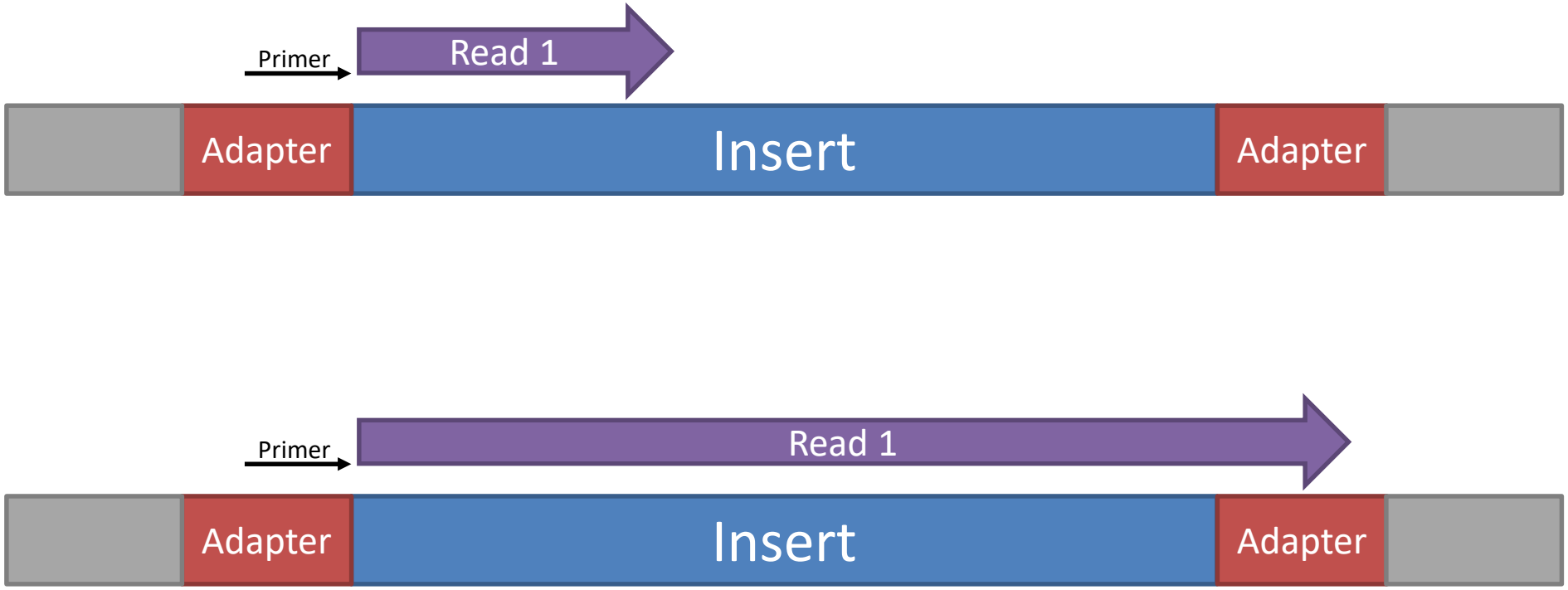
Position Specific Patterning

Positional Quality





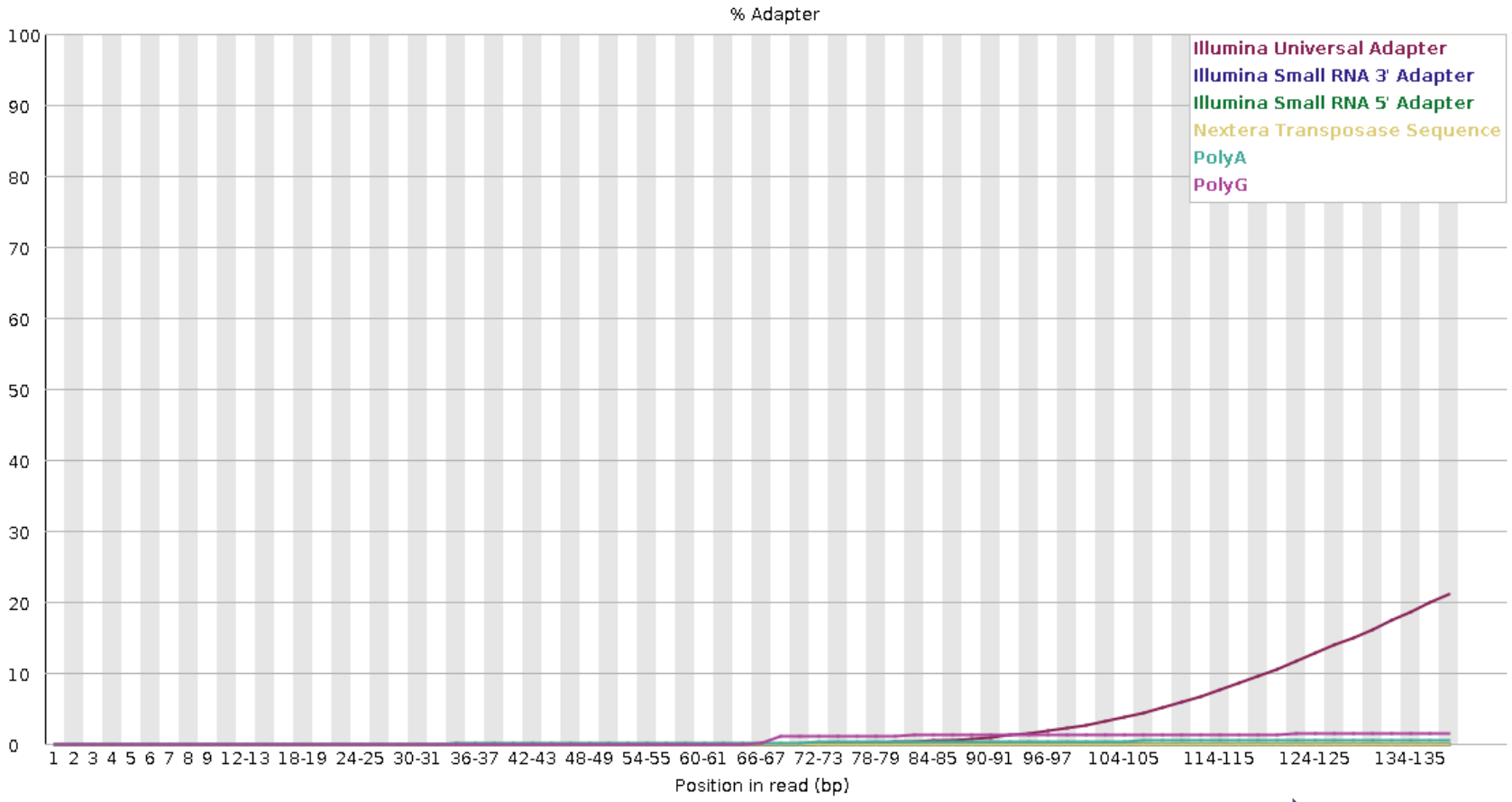
Read-through Adapters

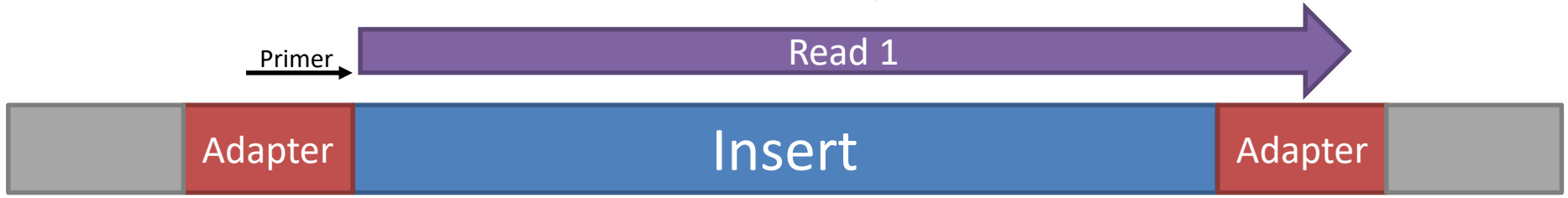


FastQC Report

Adapter Content

Measuring Read-though Adapters

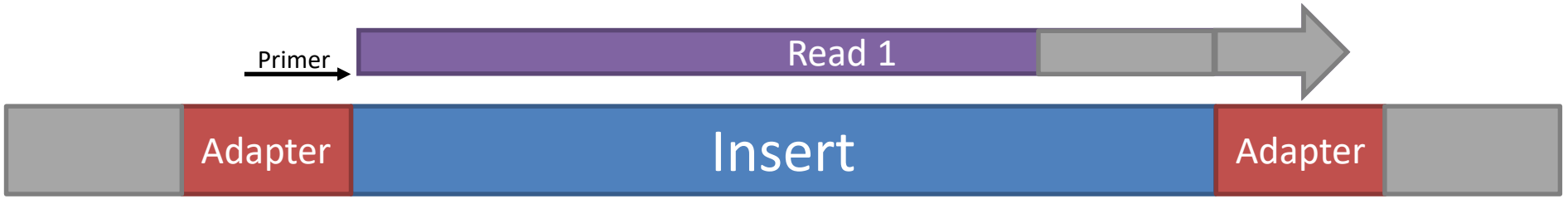




Clean-Up Options

Trimming 3’ end:

- Remove adapter read through
- Remove poor quality bases



Some quality issues may need to also remove specific reads

Despite issues may still be good enough for what is needed e.g. mapping

Mapping Statistics

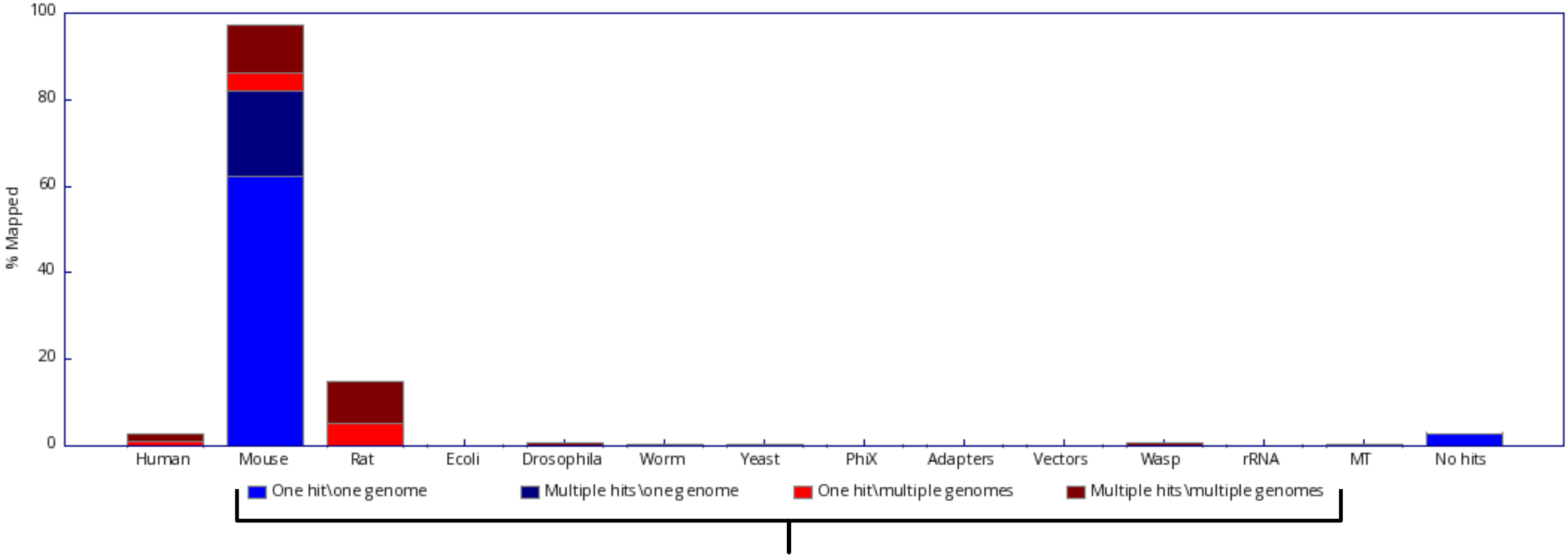
```
Time loading forward index: 00:01:10
Time loading reference: 00:00:05
Multiseed full-index search: 00:20:47
24548251 reads; of these:
  24548251 (100.00%) were paired; of these:
    1472534 (6.00%) aligned concordantly 0 times
    21491188 (87.55%) aligned concordantly exactly 1 time
    1584529 (6.45%) aligned concordantly >1 times
94.00% overall alignment rate
Time searching: 00:20:52
Overall time: 00:22:02
```

If many reads do not map to the expected genome…

…Where do they come from?

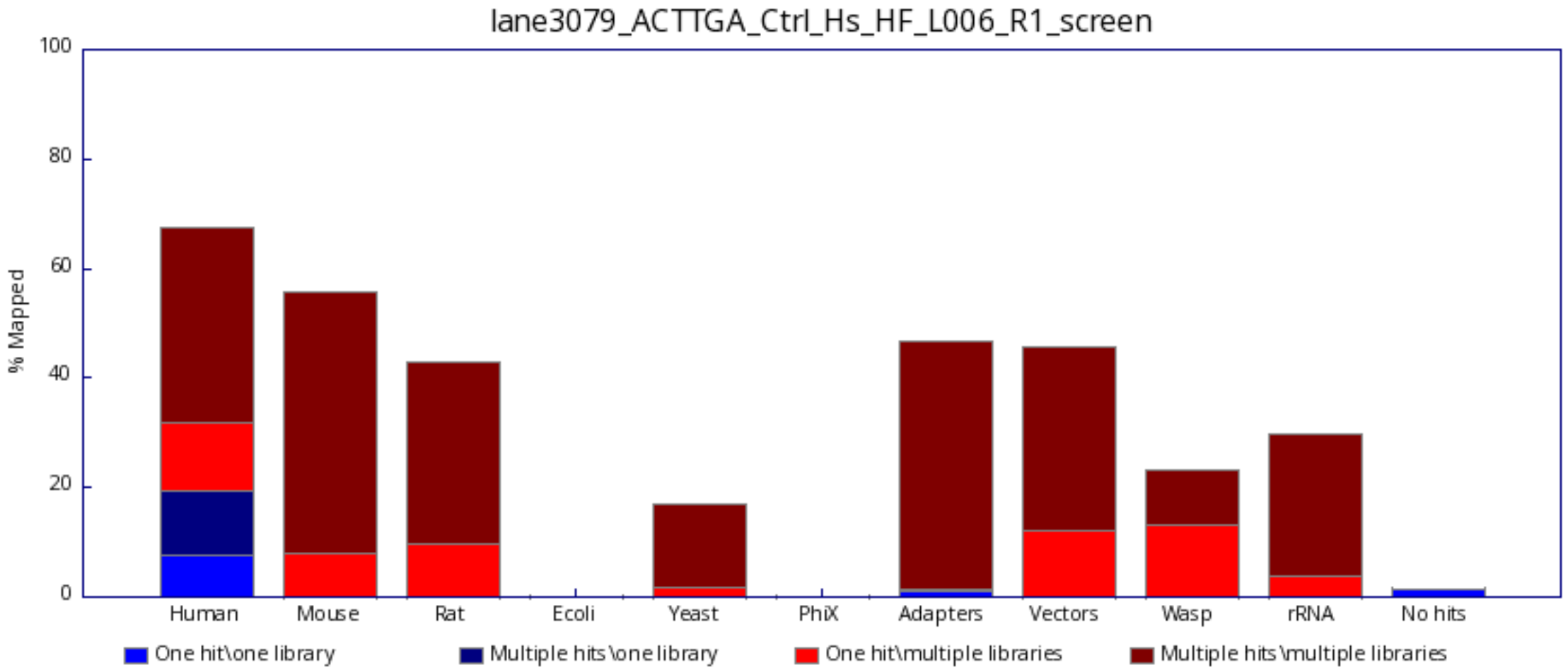
Library Screening

Map your reads against a range of reference genomes

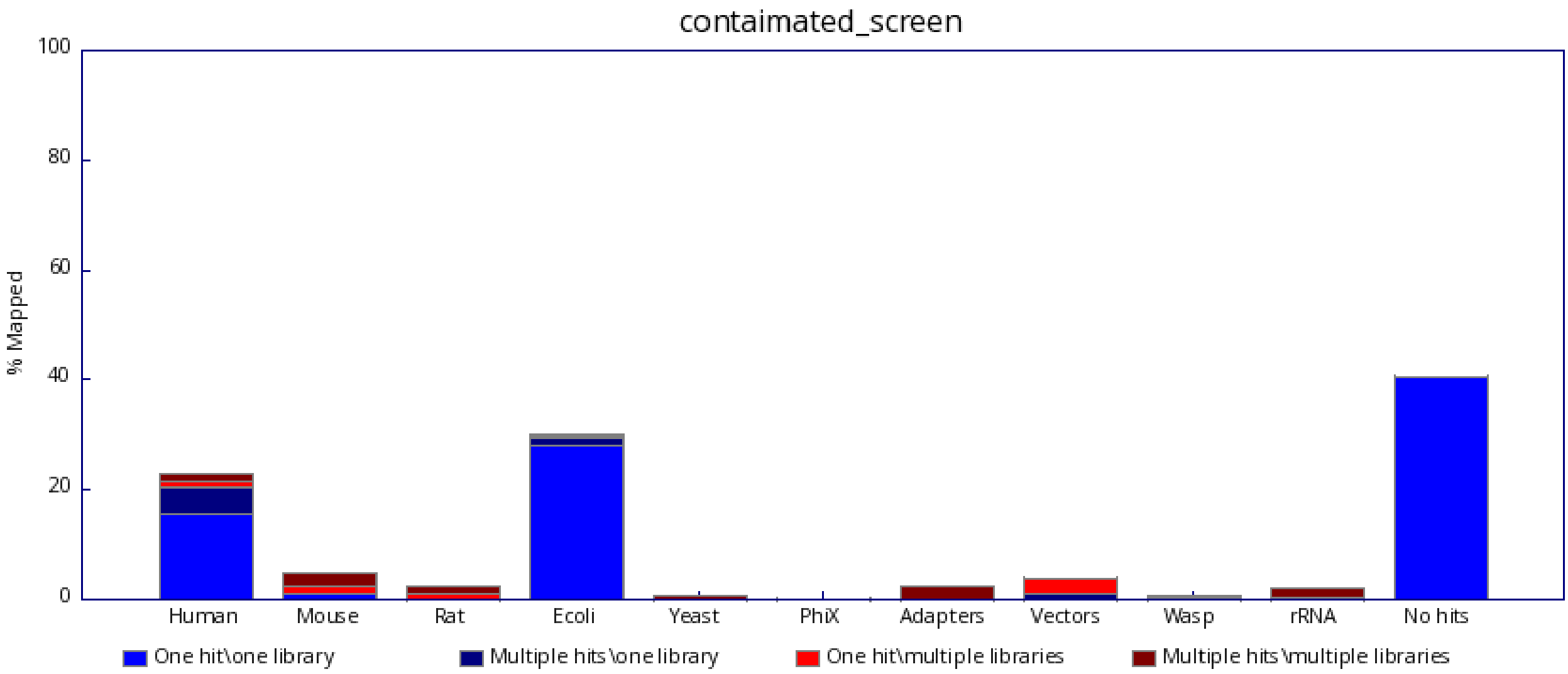


Classify matches as: unique to one species & single or multiple mapping

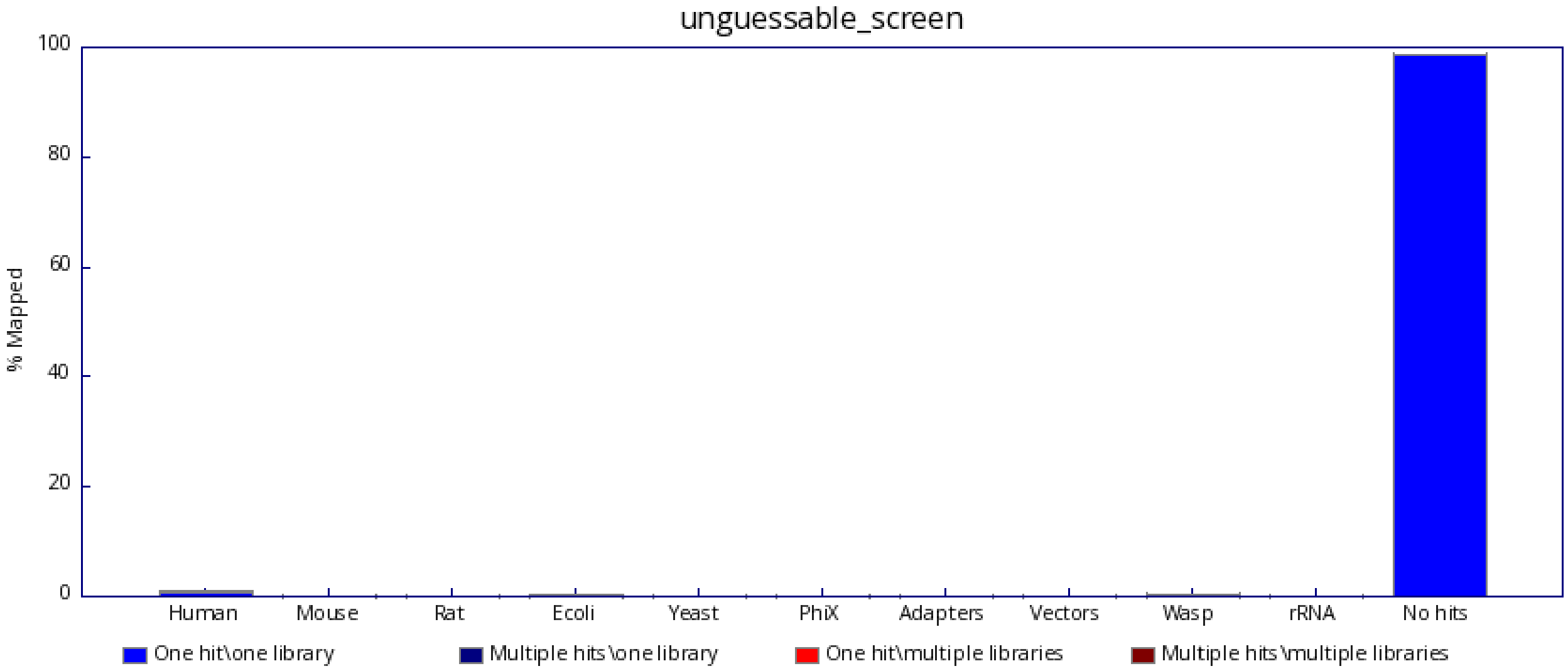
Library Screening



Library Screening



Library Screening



Assessing Library Dependent Metrics

Babraham Bioinformatics

Library Dependent QC Metrics

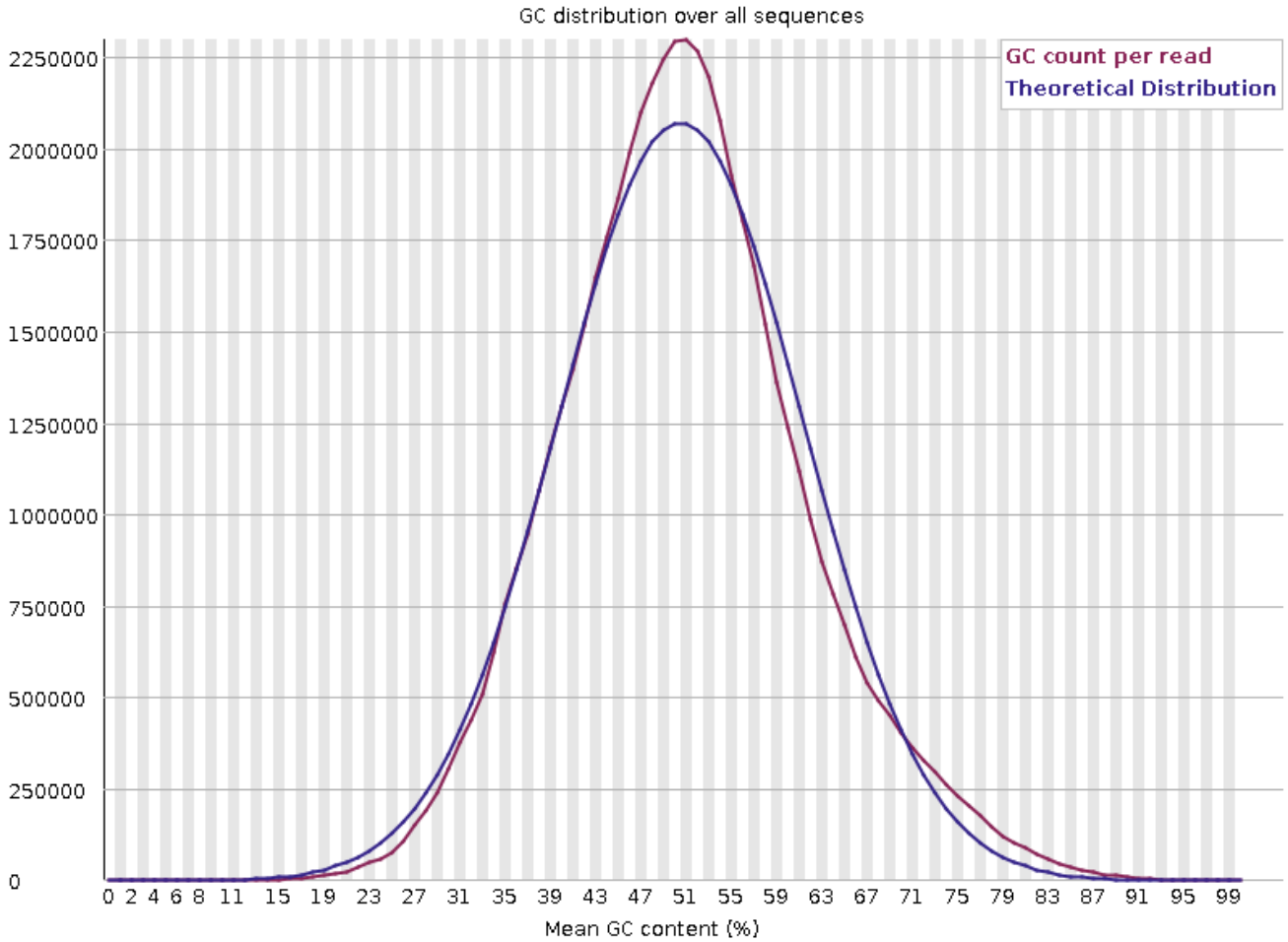
FastQC expects a Genomic Library: Do you?

Some QC metrics will be influenced by what you are sequencing

 Concern or Expected?

- GC Content
- Base Composition
- Duplication

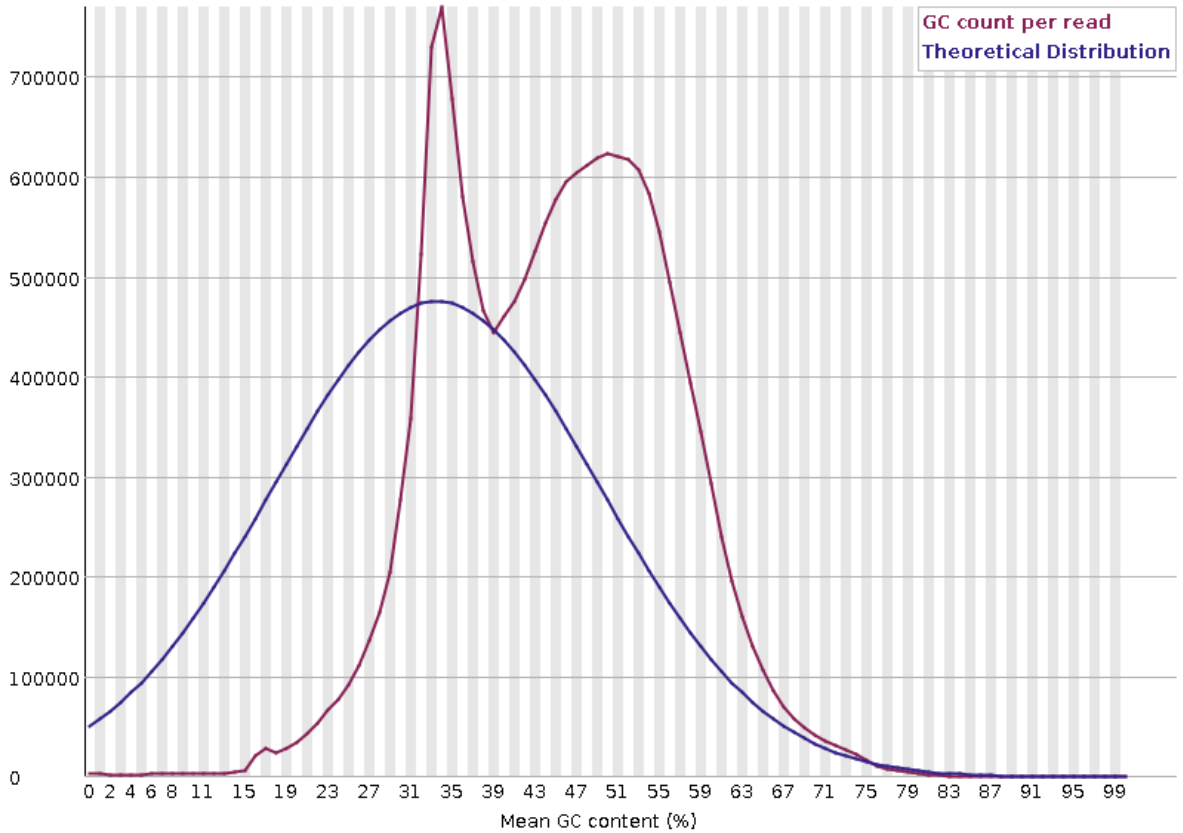
Library GC Content

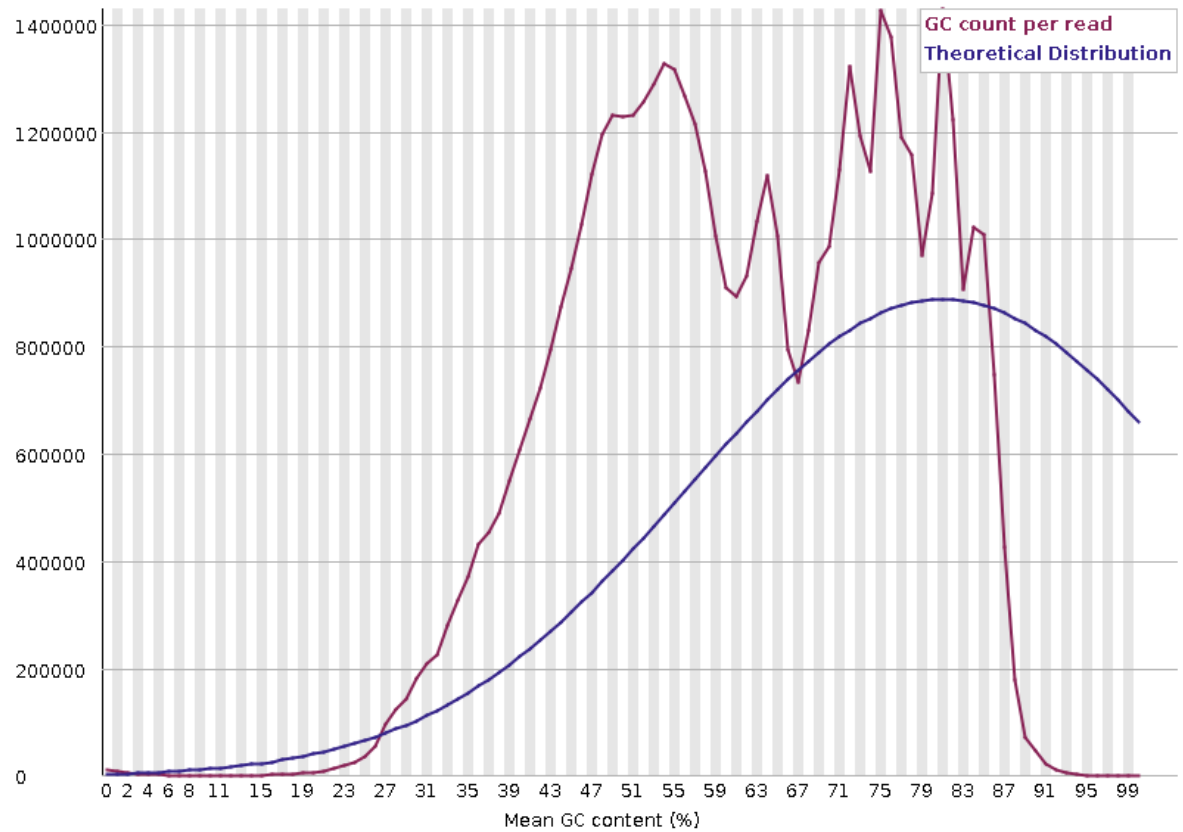


- Generic summary of library composition at a read level
- Expect a normally distributed set of values centred on the overall GC content

Sharp Peaks in GC

Concern or Expected





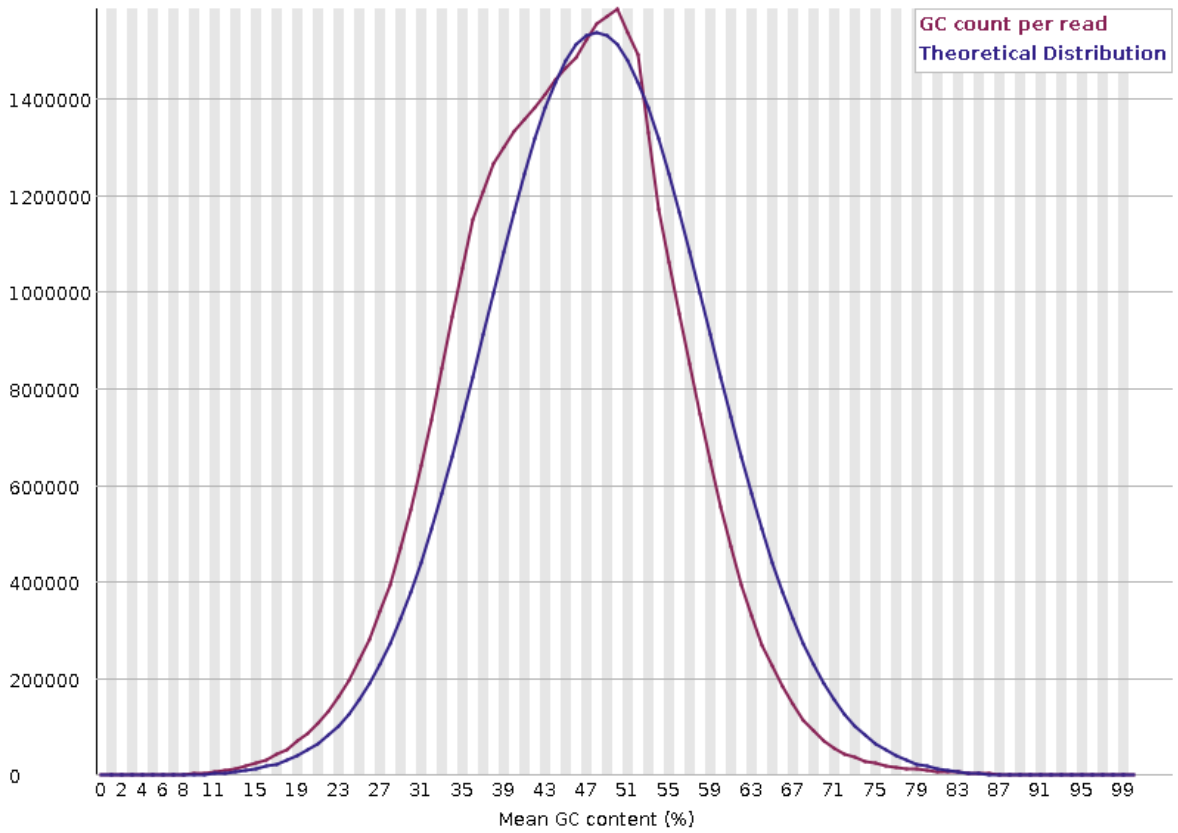
Specific Contamination with single sequence or closely related sequences

Artificial sequences, ribosomal RNA, contaminants

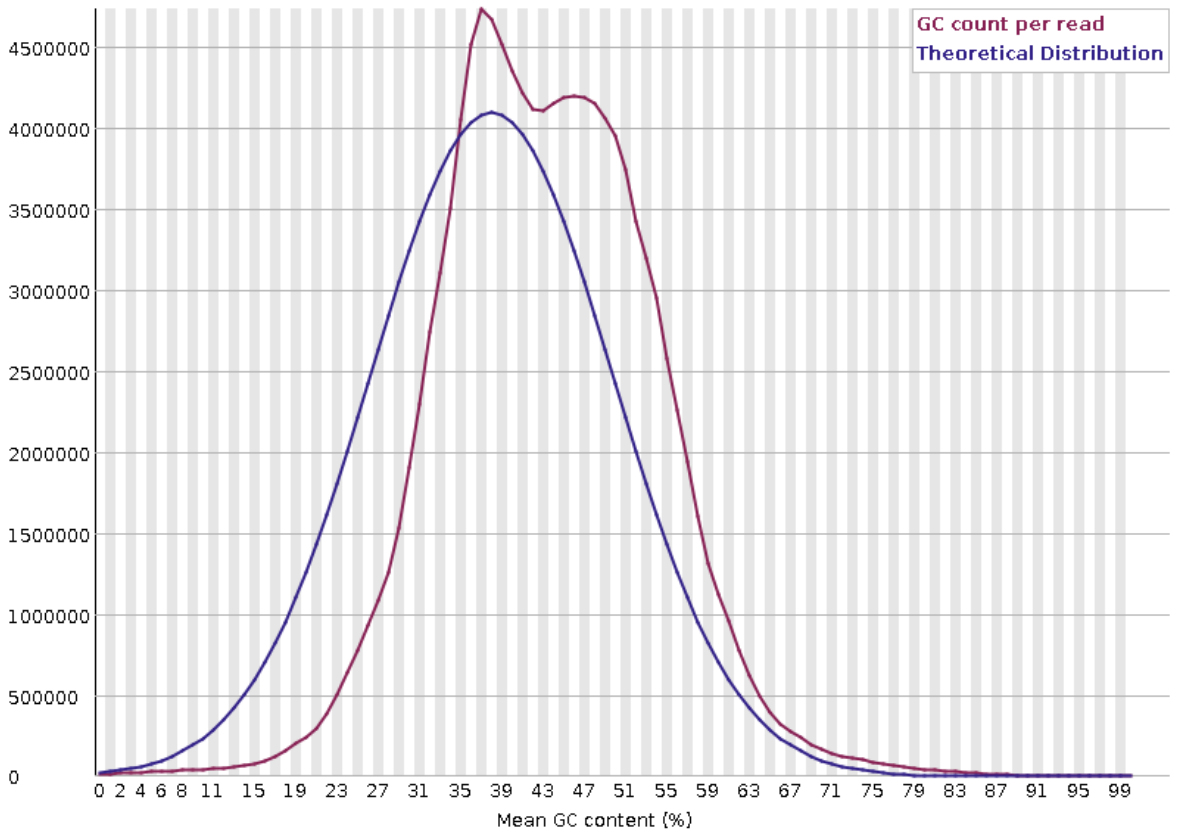
Broader Peaks in GC

Concern or Expected

Mouse **Drosophila**



H3K4me1 ChIP: enrich for regions rich in GC

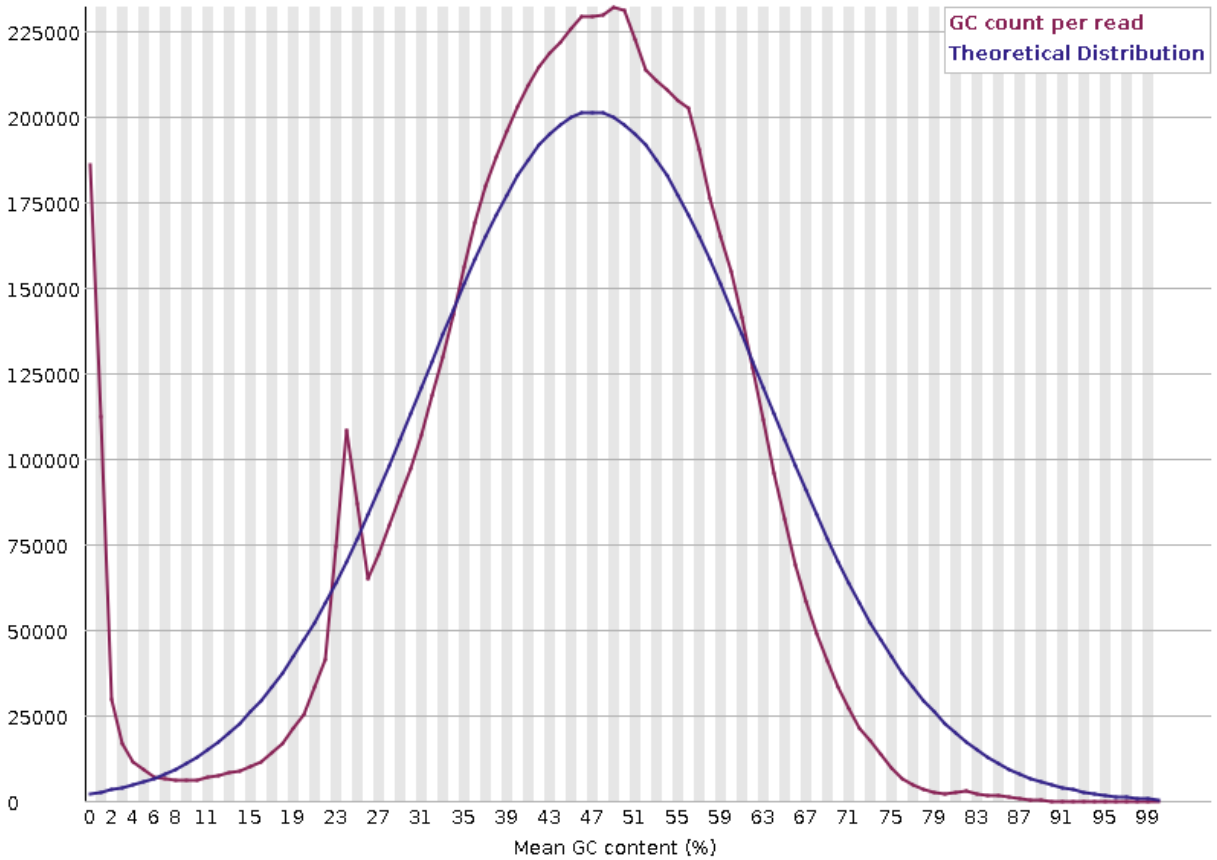


More extensive mixture of reads with different GC content

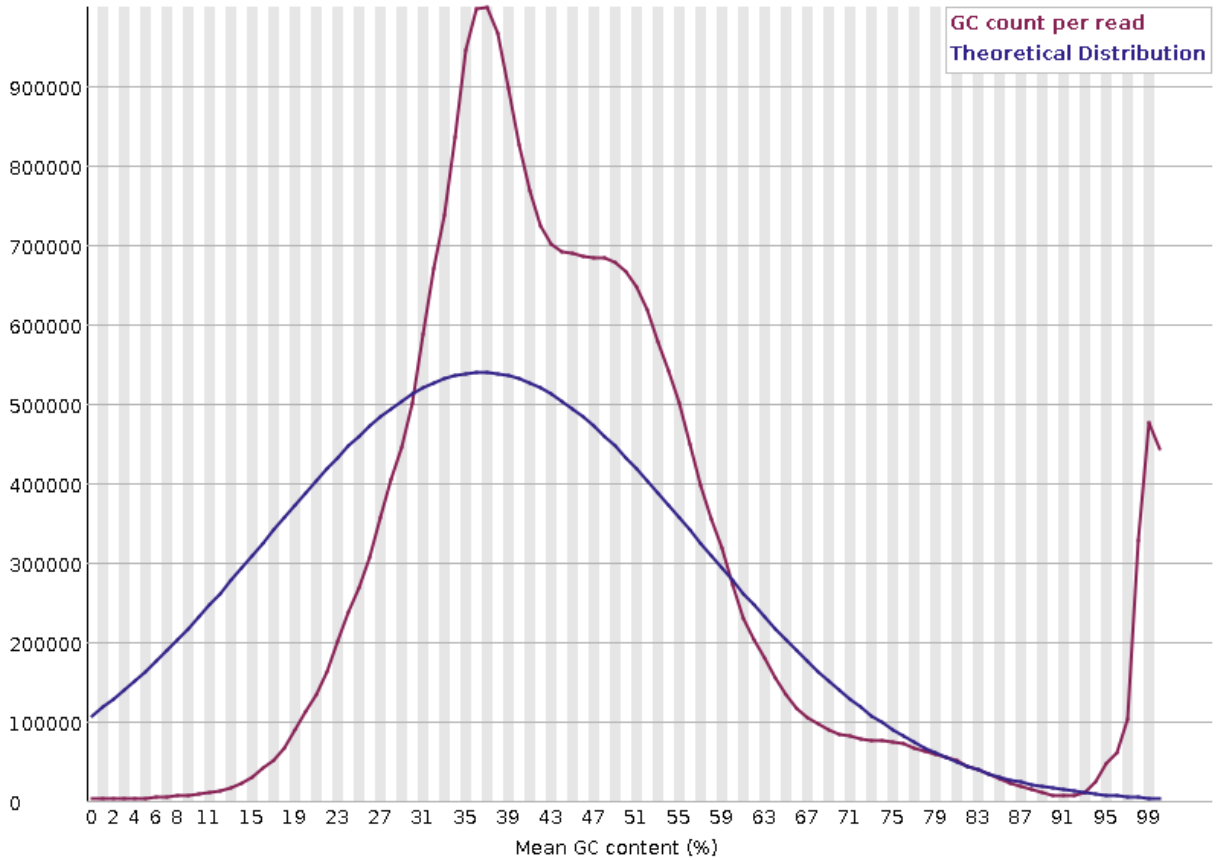
GC Skew

Concern or Expected

PolyA's

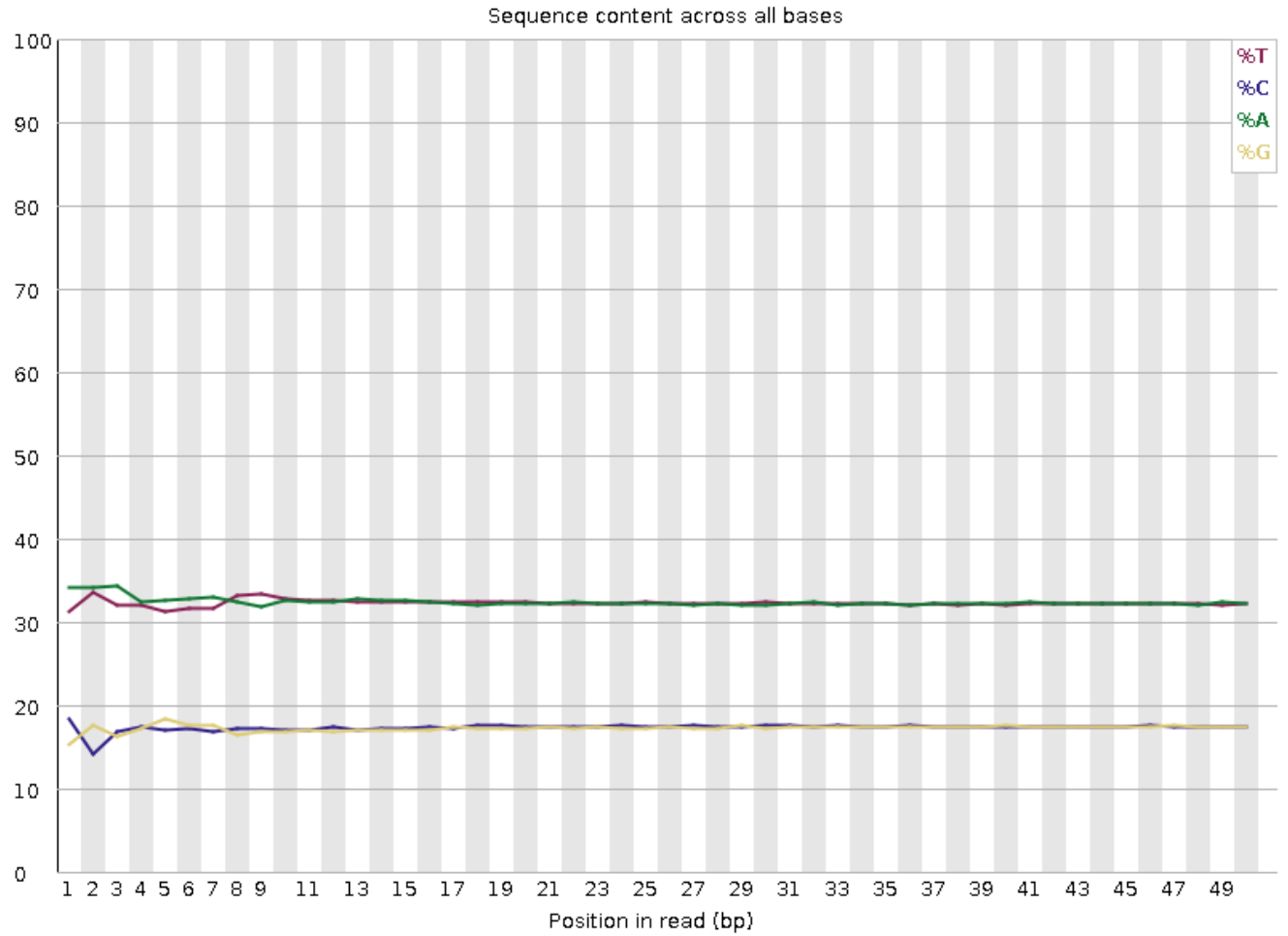


PolyG's



More extensive subset of reads with extreme differences in GC

Library Base Composition

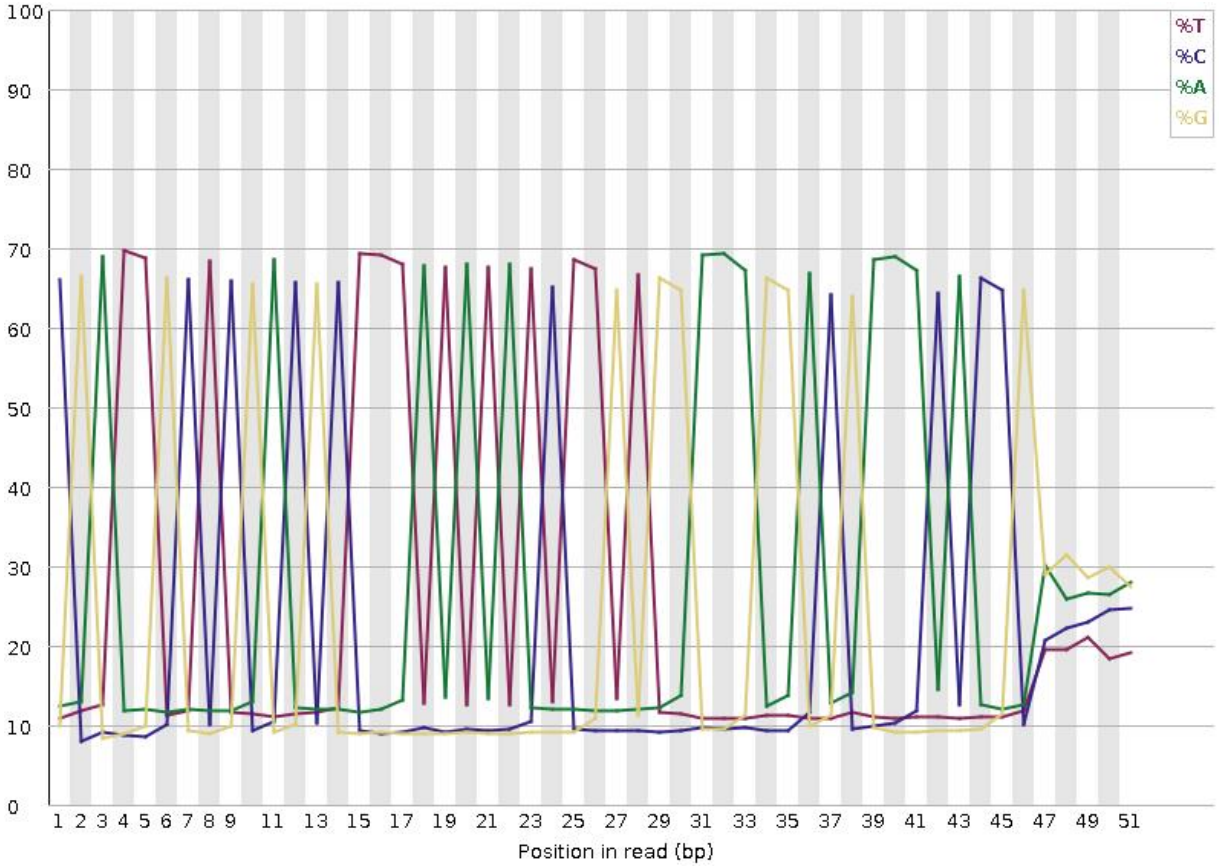


- For every chemistry cycle we can look at the number of ATGC we call
- For Libraries with random start positions the composition should be the same for all cycles

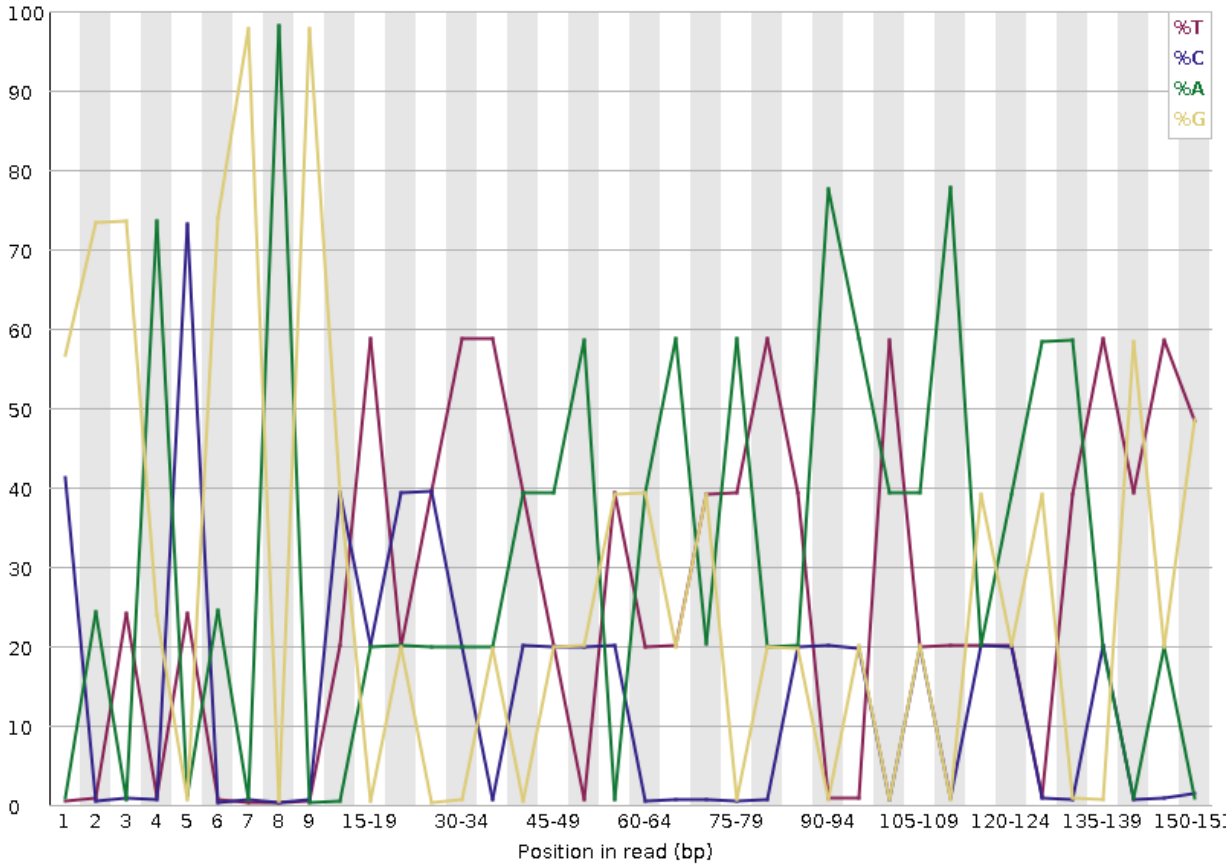
Bias Composition Throughout

Concern or Expected

Wrong Sequence



Amplicon

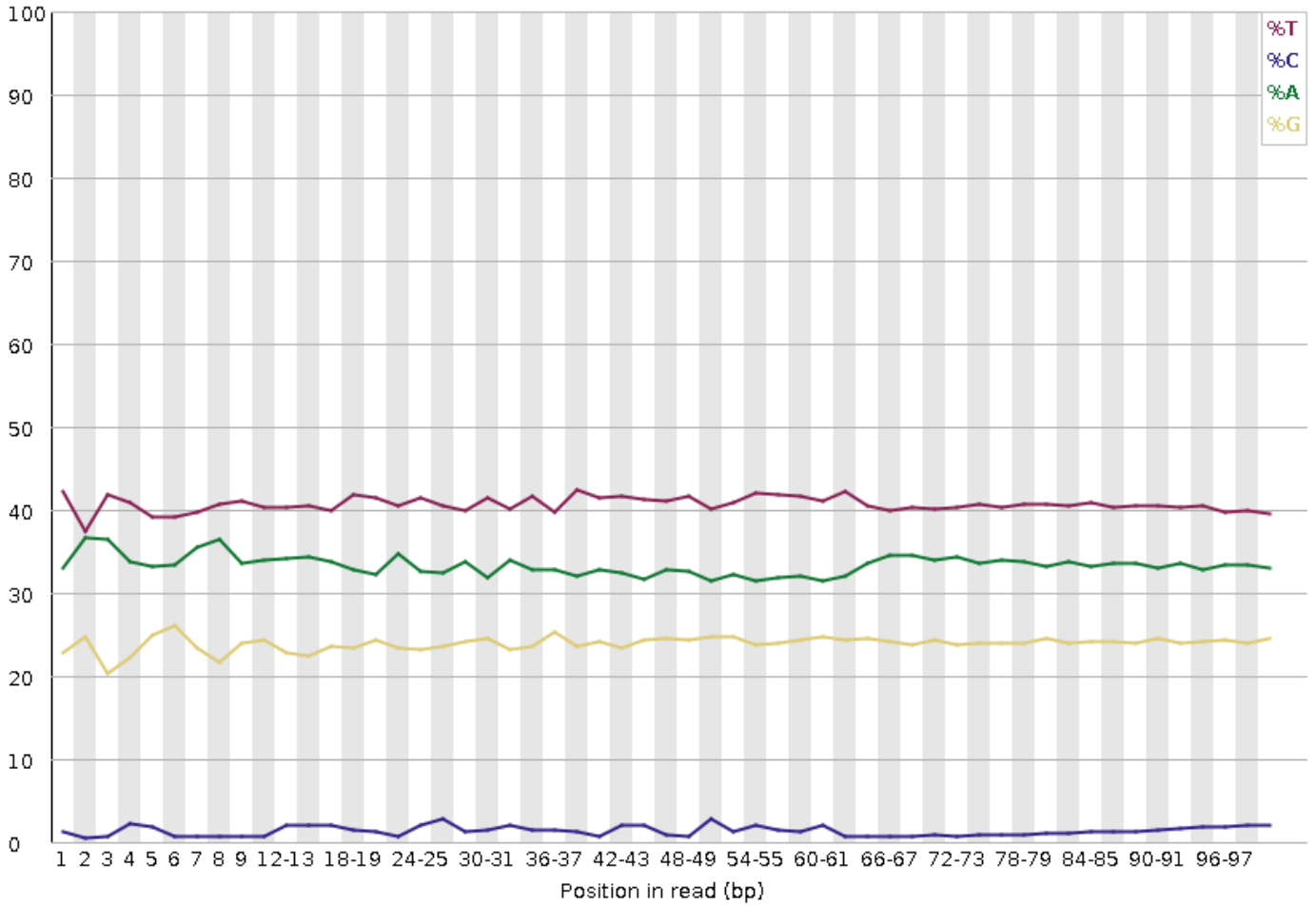


Proportional biases of bases at specific positions: Very low diversity

Bias Composition Throughout

Concern or Expected

WGBS

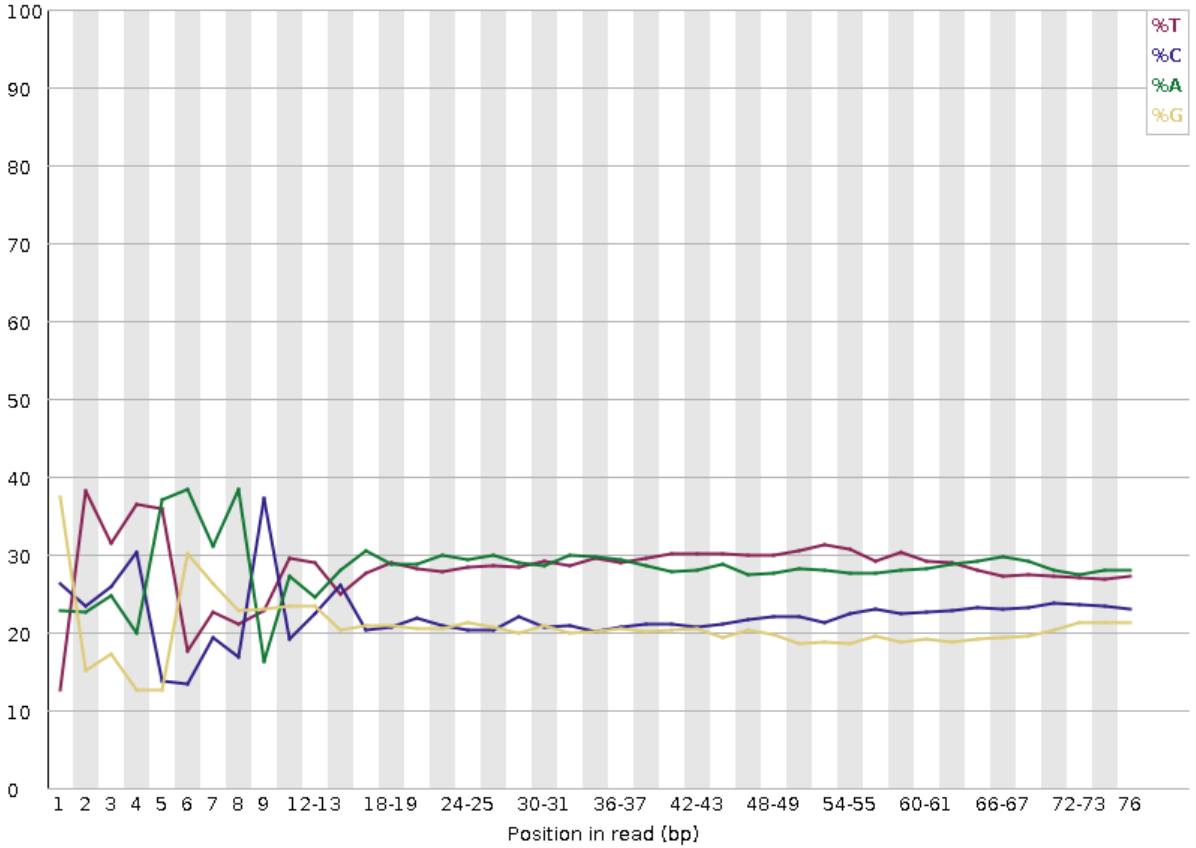


Consistent disproportional expression of bases

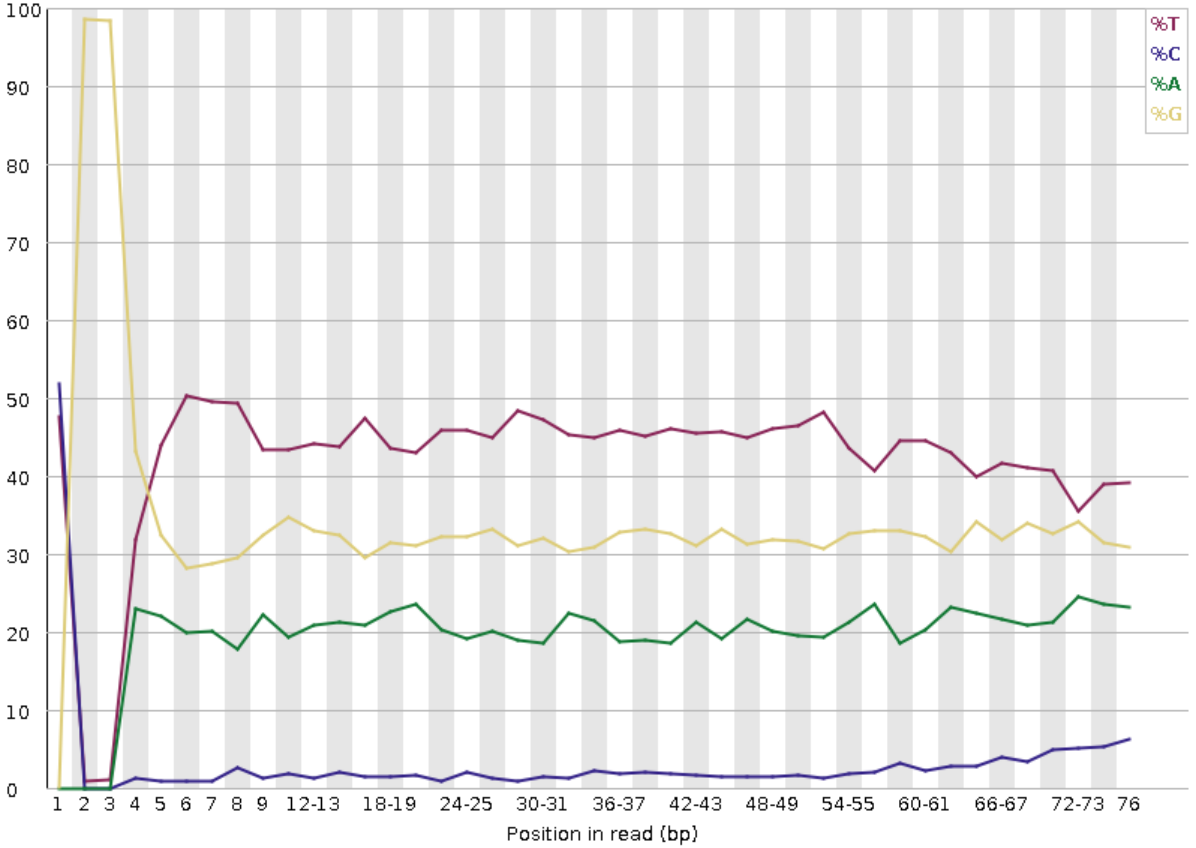
Bias Composition at 5’ end

Concern or Expected

ATAC – Transposases



RRBS – Restriction Start Site

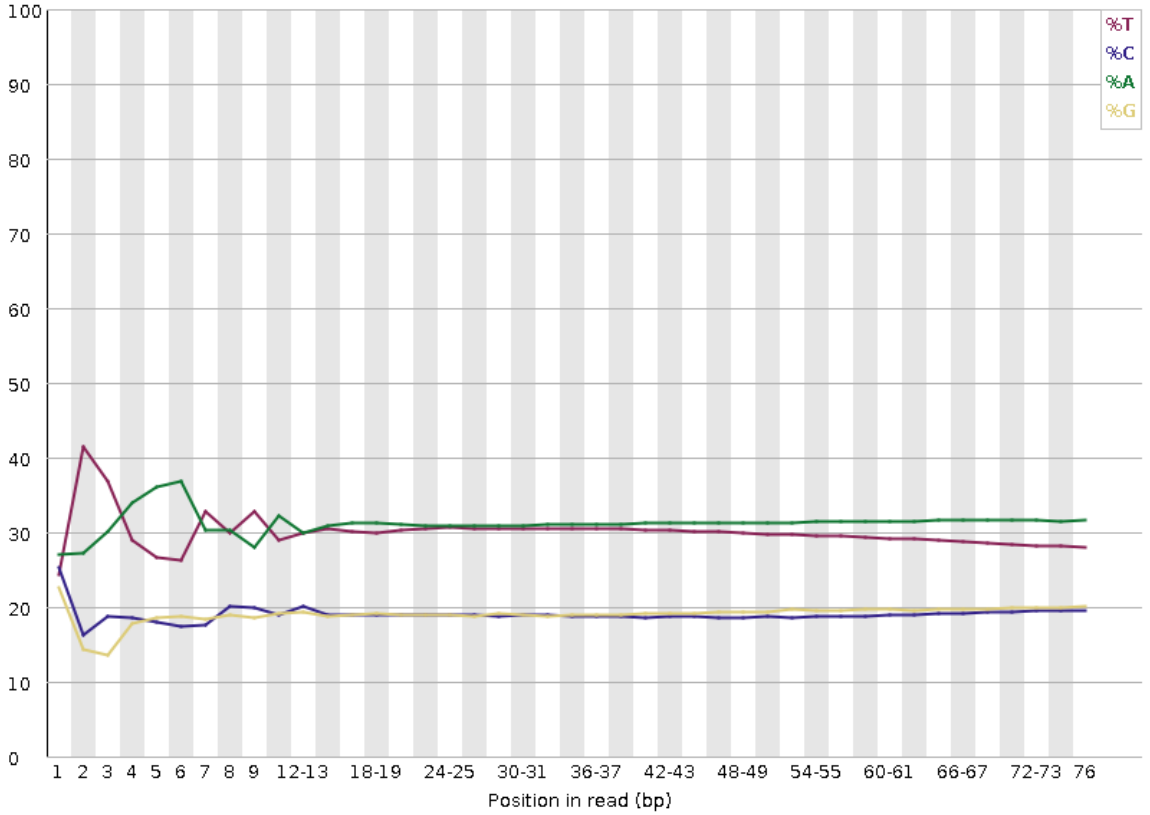


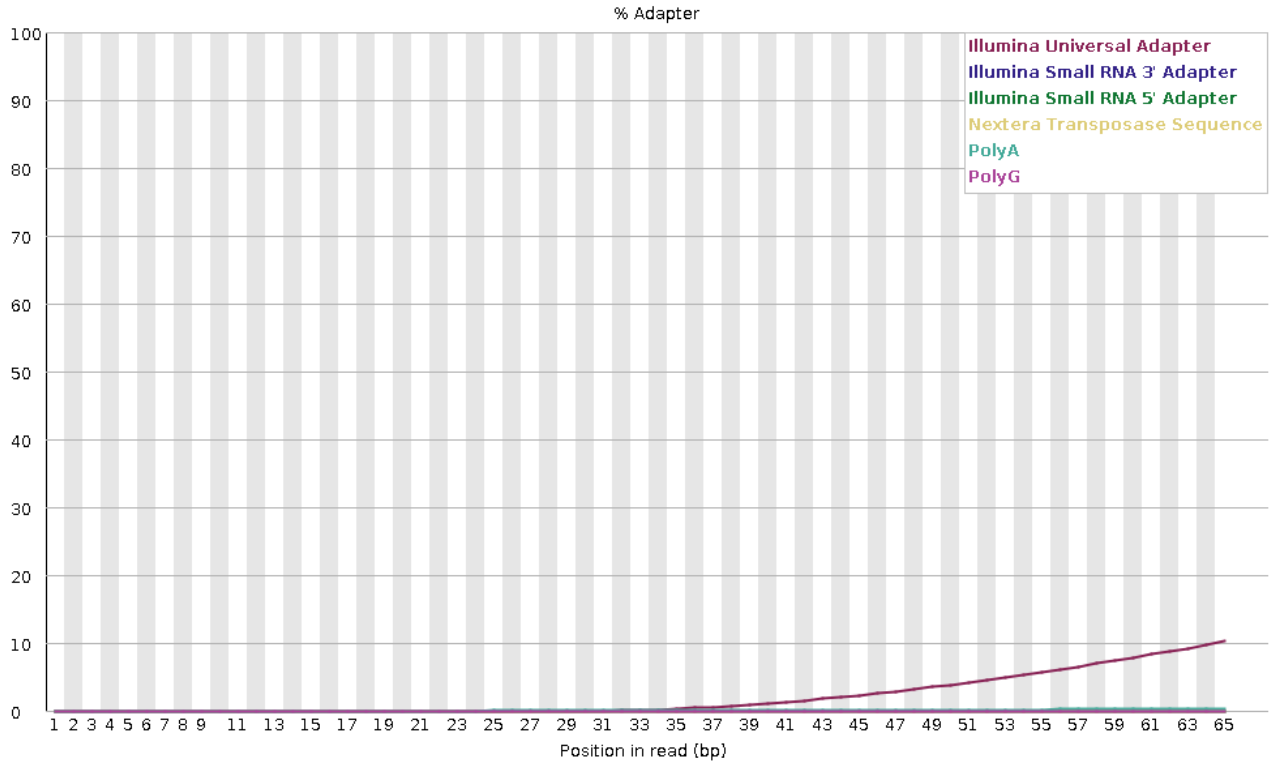
Proportional biases of bases at the start of a read: A preferred start site

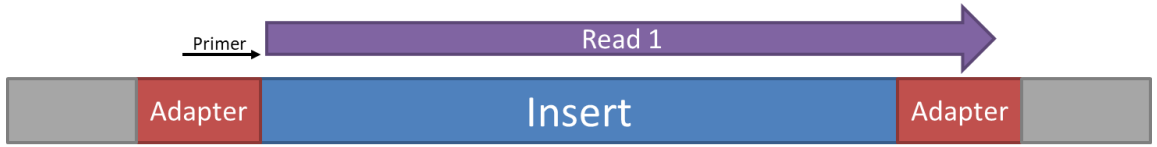
Bias Composition at 3’ end

Concern or Expected

Proportional biases of bases at the end of a read: consistent closing sequence



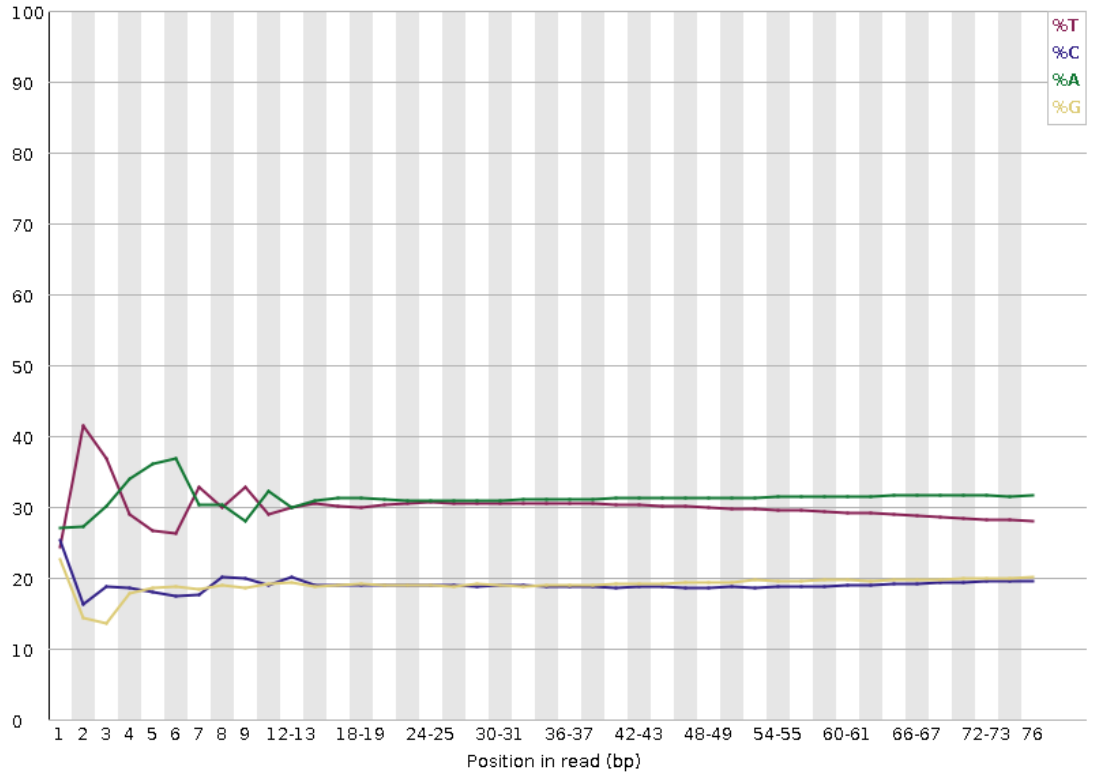


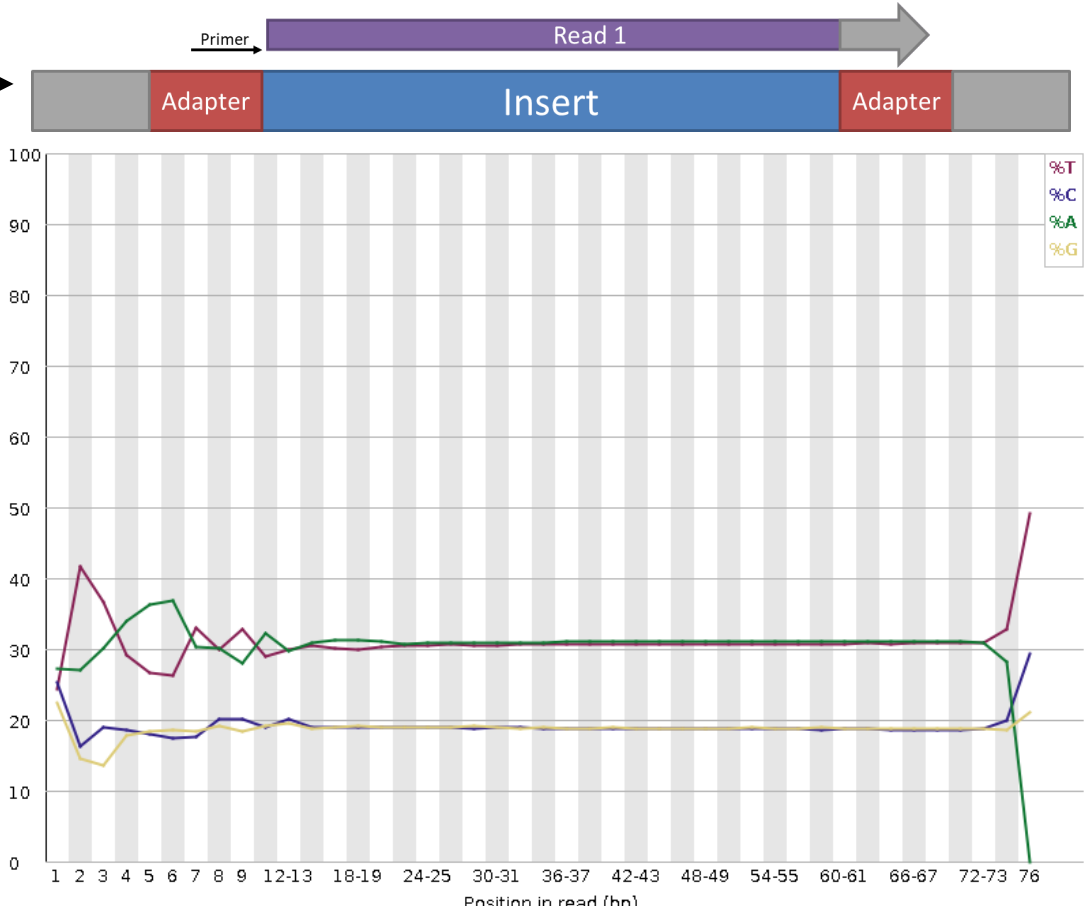


Bias Composition at 3’ end

Concern or Expected

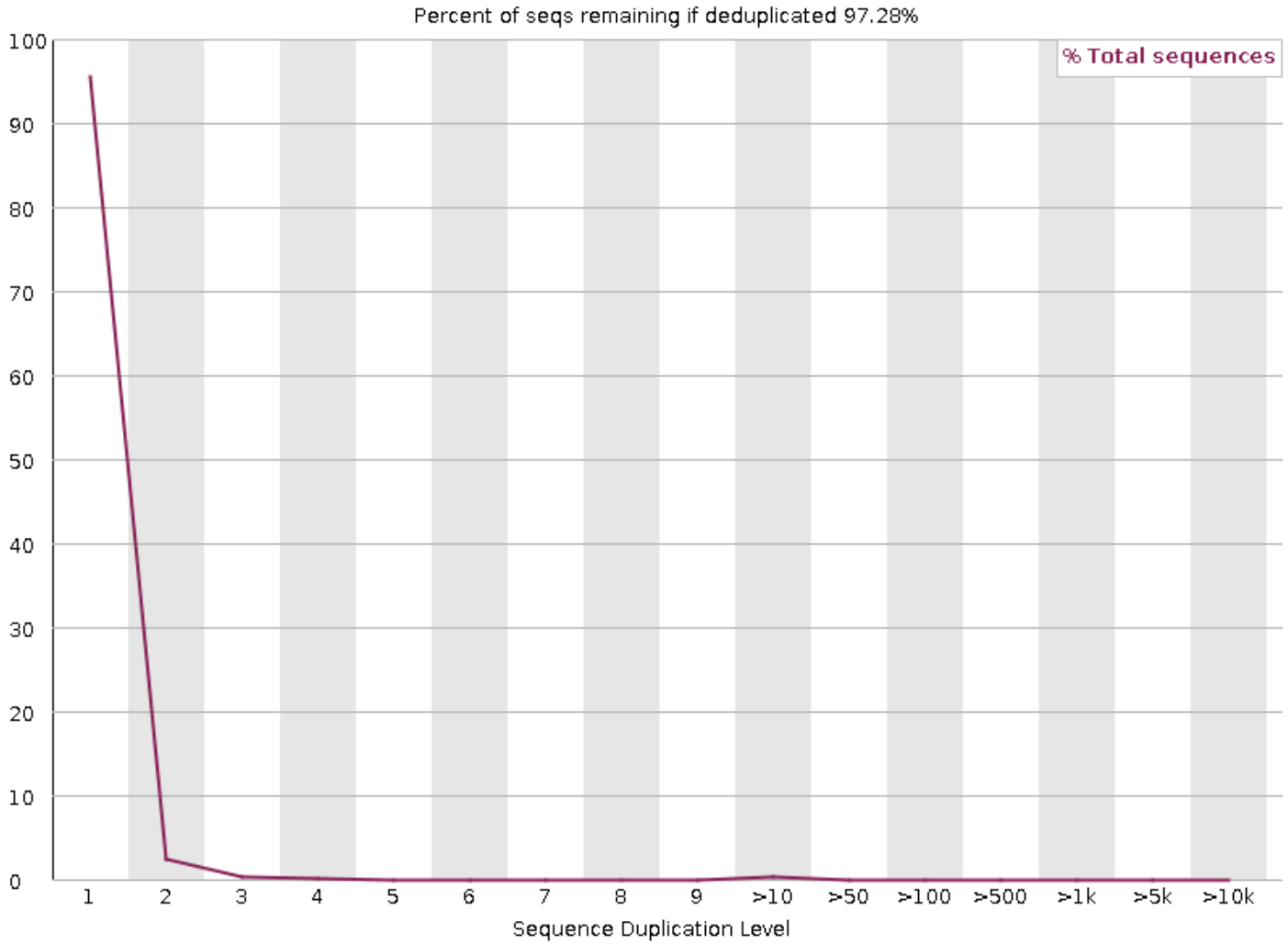
Proportional biases of bases at the end of a read: consistent closing sequence





Bioinformatics processing can also influence QC metrics!

Duplication

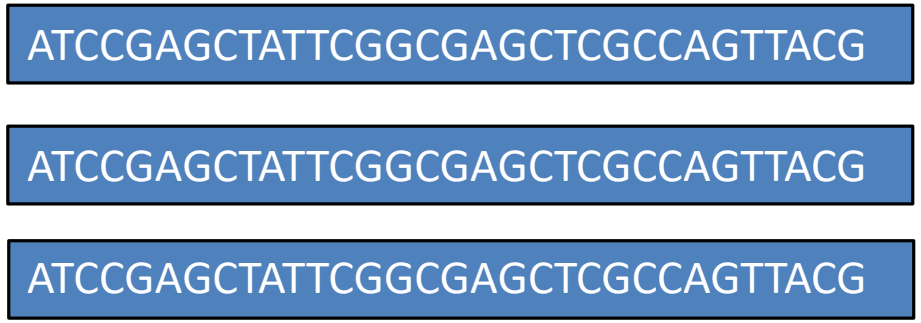


- How frequently the exact same sequence appears in your library
- For WGS expect most sequences to be unique

Duplication

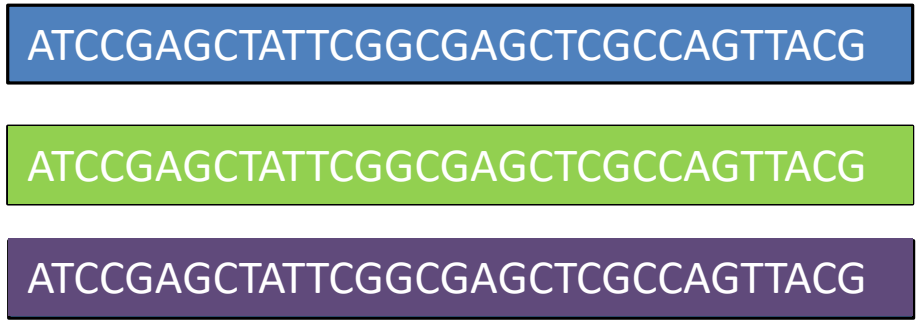
If the exact same sequence appears more than once it could be…

Technical:



- PCR duplicates

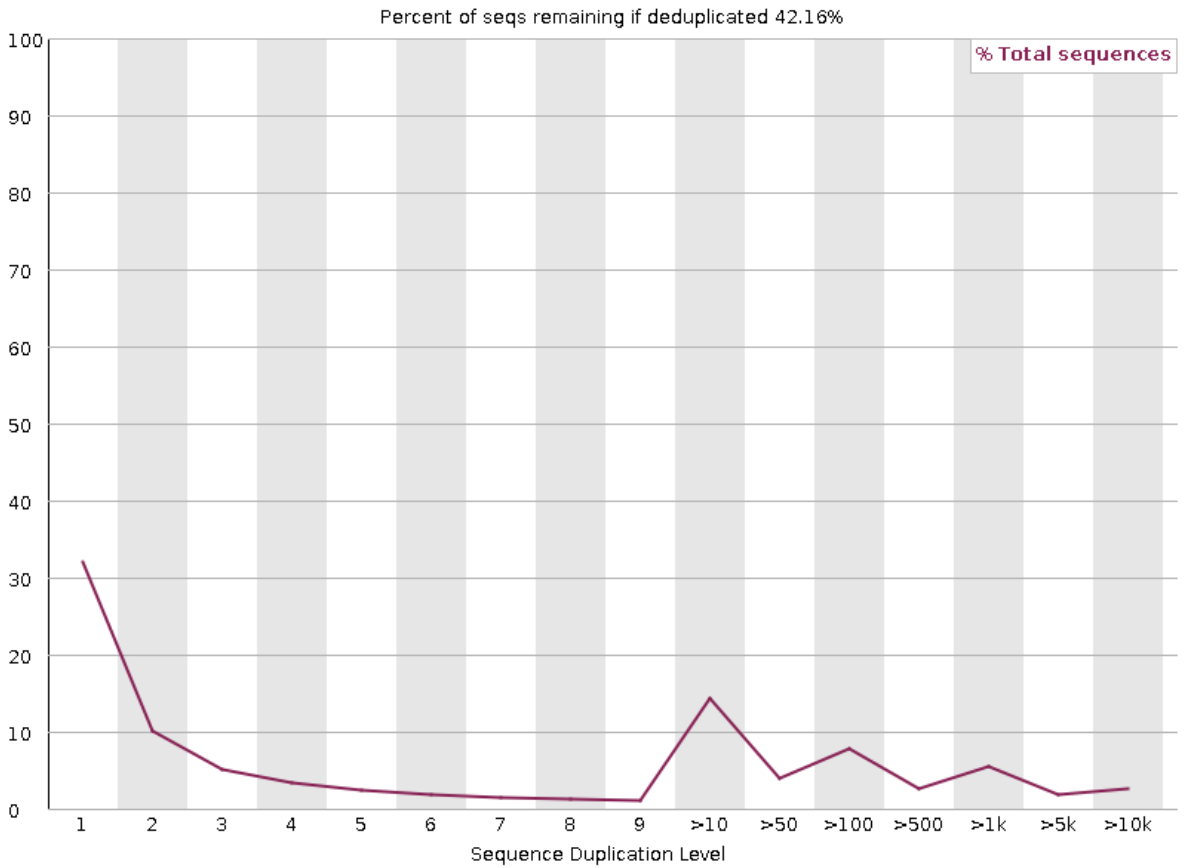
Coincidental:



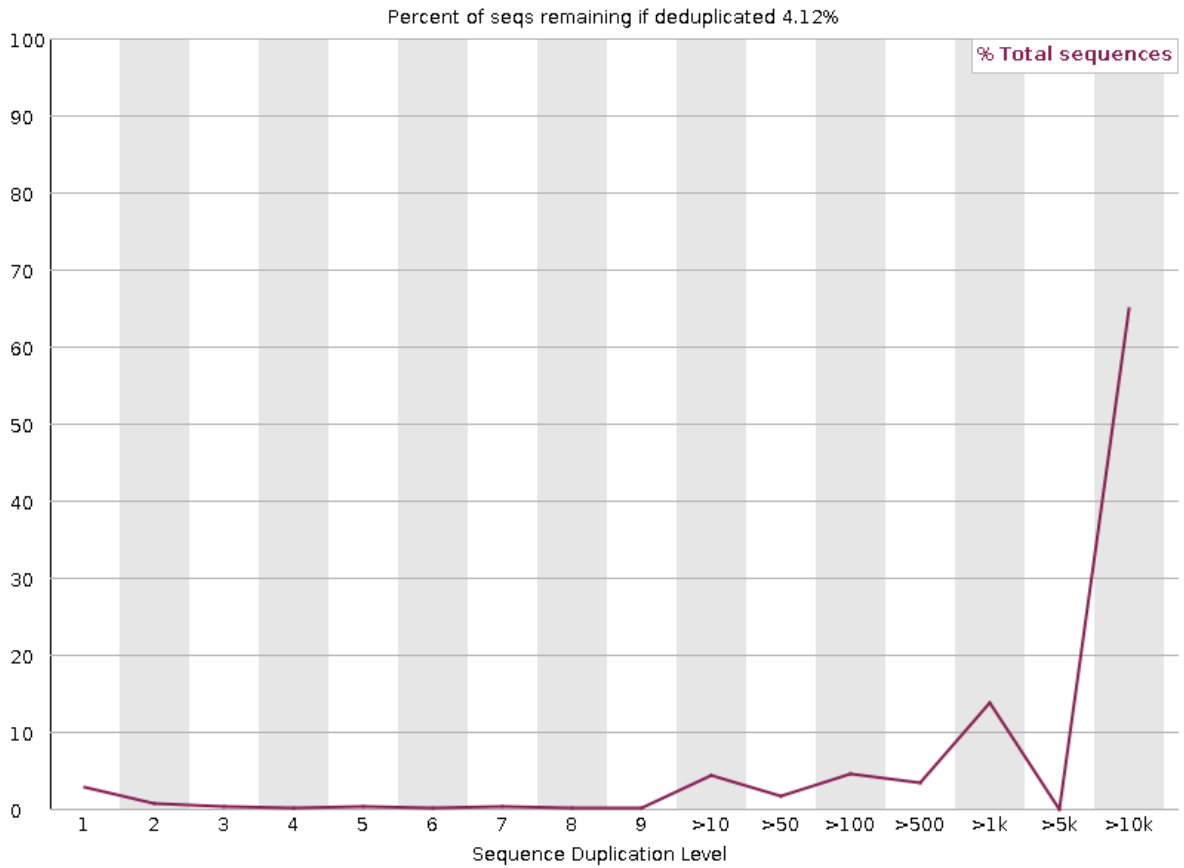
- Deep sequencing
- Highly present sequences
- Restricted diversity libraries

Duplication
Concern or Expected

RNA-Seq



Amplicon



BUT could have technical duplication with expected coincidental duplication!

Overrepresented Sequences

- Extreme duplication
- The exact same sequence is a significant proportion of the whole library (which might not be duplicated overall)

- Poly Sequences
- Specific Sequences

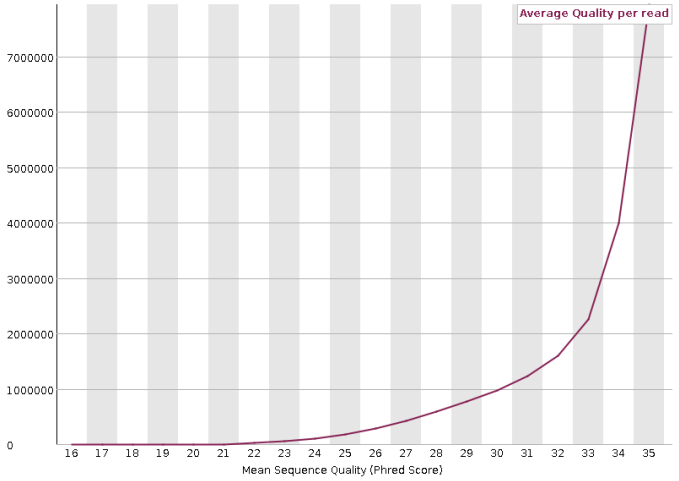
Sources of Poly Sequences

PolyN – Quality too poor to make any calls

Sequence	Count	Percentage
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN	462344	1.070097045533307
GNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN	232540	0.5382147642627897
ANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN	127291	0.29461553090984244
CNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN	87792	0.20319493671694688
TNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN	85181	0.19715176672688003
GANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN	48918	0.11322090753507845

PolyG – Empty space in 2 colour chemistry, can be technically “high quality” calls

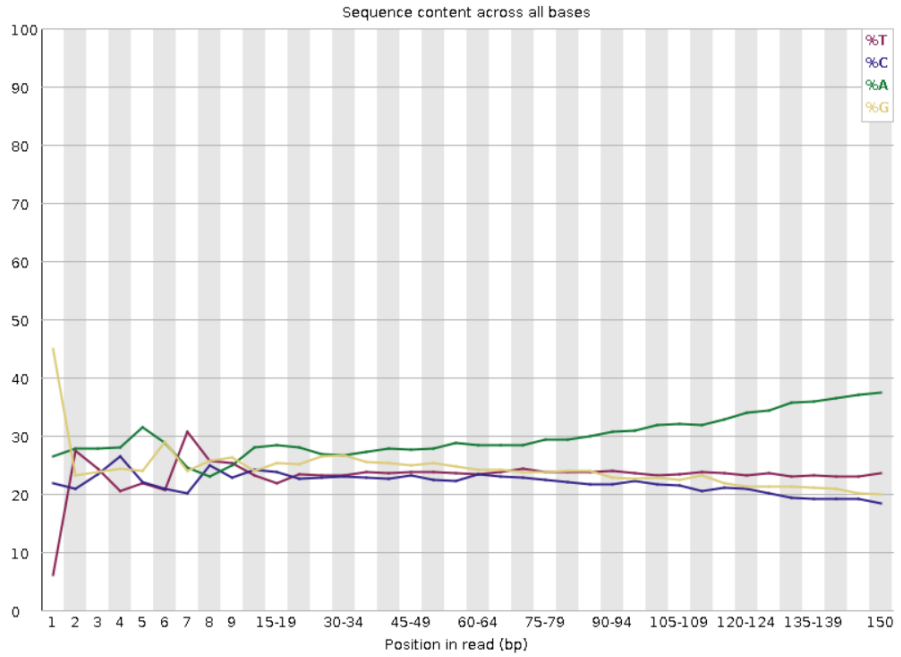
Sequence	Count	Percentage
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG	406221	1.9711963990403654
TGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG	315458	1.530766931420275
AGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG	161606	0.7841966940737117
TTGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG	46937	0.2277628320095653
CGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG	32291	0.15669279264590566
ATGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG	24237	0.1176105792746838

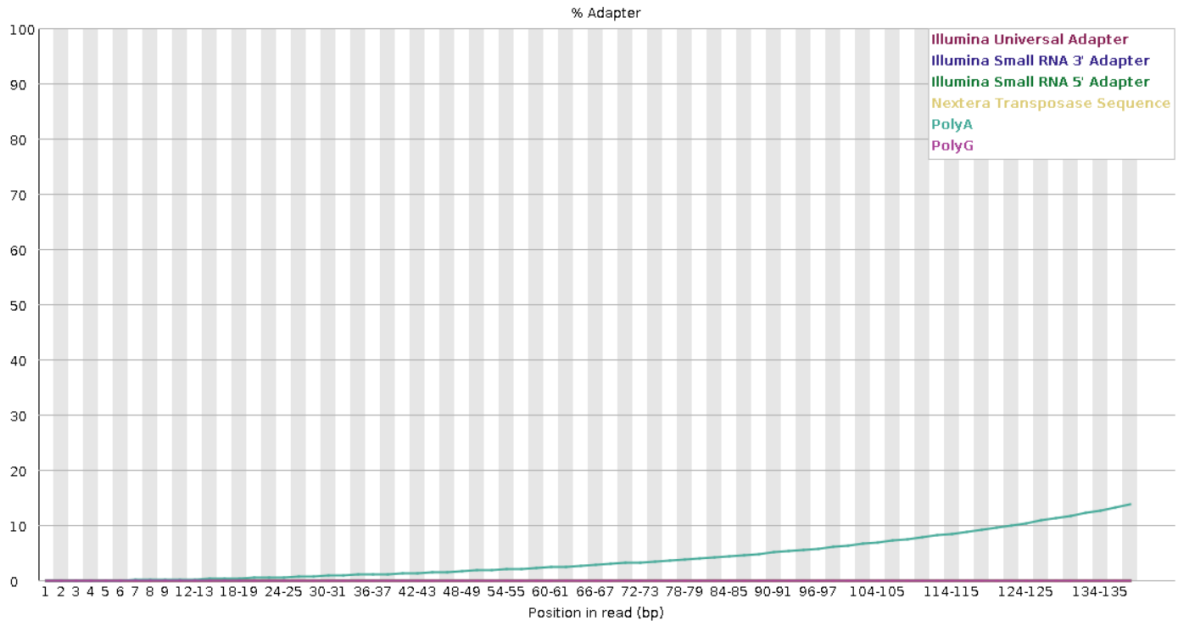


Sources of Poly Sequences

PolyA (or PolyT) – Common in RNA-Seq

Sequence	Count	Percentage
TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT	68355	1.7344041279604823
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	67792	1.7201188595230343



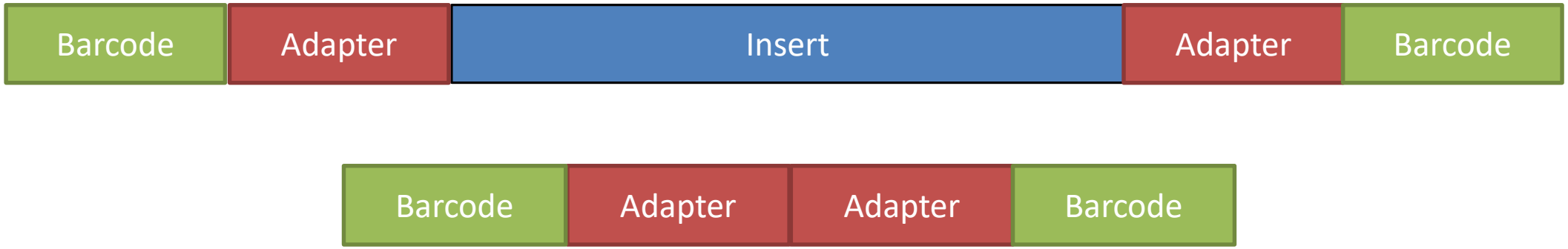


Overrepresented Specific Sequences

- Normally artificial sequences (primers, adapters, vectors etc)
- Can search a database of known sequences to find matches

Sequence	Count	Percentage
GATCGGAAGAGCACACGTCTGAACTCCAGTCACCTTGTAATCTCGTATGC	17957	0.14359551756800035

Example of an Adapter dimer:



Overrepresented Specific Sequences

- Other potential sources…

Sequence	Count	Percentage	Possible Source
GGCTTCCTCGGCCCCGGGATTCGGCGAAAGCTGCGGCCGGAGGGCTGTAA	746766	1.360148419566899	No Hit

Basic Local Alignment Search Tool

Sequence	Count	Percentage	Possible Source
CTTATACACATCTCCGAGCCCACGAGACTAAGGCGAATCTCGTATGCCGT	2767629	5.149013792611521	No Hit

Ribosomal

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Mus musculus large subunit ribosomal RNA gene, partial sequence	Mus musculus	93.5	93.5	100%	1e-15	100.00%	4731	MN537140.1
Mus musculus clone contig 15 chromocenter region genomic sequence	Mus musculus	93.5	93.5	100%	1e-15	100.00%	884	KX121621.1
Mus musculus genome assembly, chromosome: 18	Mus musculus	93.5	186	100%	1e-15	100.00%	89877872	OX439032.1
Mus musculus genome assembly, chromosome: 16	Mus musculus	93.5	93.5	100%	1e-15	100.00%	96079412	OX439031.1
Mus musculus genome assembly, chromosome: 18	Mus musculus	93.5	93.5	100%	1e-15	100.00%	90037828	OX390161.1
Mus musculus genome assembly, chromosome: 16	Mus musculus	93.5	93.5	100%	1e-15	100.00%	97401718	OX390159.1
	Mus musculus	93.5	93.5	100%	1e-15	100.00%	62939505	OX389813.1
	Mus musculus	93.5	93.5	100%	1e-15	100.00%	89861325	OX389812.1
e	Mus musculus	93.5	93.5	100%	1e-15	100.00%	15928	GU372691.1
th enriched library, clone:F730219H1...	Mus musculus	93.5	93.5	100%	1e-15	100.00%	910	AK155774.1
ength enriched library, clone:F63021...	Mus musculus	93.5	93.5	100%	1e-15	100.00%	1045	AK155253.1
Mus musculus CNR gene for cadherin-related neuronal receptor, complete cds	Mus musculus	93.5	93.5	100%	1e-15	100.00%	10521	AB114630.1
Mus musculus putative membrane-associated guanylate kinase 1 (Magi-1) mRNA, alternatively spliced b form,...	Mus musculus	93.5	93.5	100%	1e-15	100.00%	5371	AF027503.1
Chain L5, Mus musculus 28S ribosomal RNA	Mus musculus	86.1	86.1	100%	2e-13	98.00%	4731	7CPU_L5
Mus musculus 45S pre-ribosomal RNA (Rn45s), ribosomal RNA	Mus musculus	86.1	86.1	100%	2e-13	98.00%	13400	NR_046233.2
TPA: Mus musculus ribosomal DNA, complete repeating unit	Mus musculus	86.1	86.1	100%	2e-13	98.00%	45306	BK000964.3
Mus musculus 28S ribosomal RNA (Rn28s1), ribosomal RNA	Mus musculus	86.1	86.1	100%	2e-13	98.00%	4730	NR_003279.1
M.musculus 45S pre rRNA gene	Mus musculus	86.1	86.1	100%	2e-13	98.00%	22118	X82564.1
Mouse 28S ribosomal RNA	Mus musculus	86.1	86.1	100%	2e-13	98.00%	4712	X00525.1

Contaminant

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Escherichia phage Lambda_ev058 genome assembly, chromosome: 1	Escherichia pha...	93.5	186	100%	1e-15	100.00%	47678	LR597651.1
Escherichia phage Lambda_ev017 genome assembly, chromosome: 1	Escherichia pha...	93.5	93.5	100%	1e-15	100.00%	50126	NC_049948.1
	Bacillus phage...	93.5	93.5	100%	1e-15	100.00%	45023	MH538296.1
	Escherichia coli	93.5	93.5	100%	1e-15	100.00%	2780	MH213709.1
	Polynucleobacte...	93.5	93.5	100%	1e-15	100.00%	1655757	LT606948.1
	Bacillales bacter...	93.5	93.5	100%	1e-15	100.00%	4793	LC663156.1
	bacterium	93.5	93.5	100%	1e-15	100.00%	4287	LC663089.1
Bacterium ARSSAG-00000681 DNA, putative prophage region, clone: 00000681_pp1	bacterium	93.5	93.5	100%	1e-15	100.00%	6385	LC663013.1
Bacterium ARSSAG-00000681 DNA, putative prophage region, clone: 00000681_pp3	bacterium	93.5	93.5	100%	1e-15	100.00%	24280	LC662963.1
Rhodobacteraceae bacterium ARSSAG-00000591 DNA, putative prophage region, clone: 00000591_pp1	Paracoccaceae...	93.5	93.5	100%	1e-15	100.00%	5070	LC662864.1
Escherichia coli strain O111 chromosome, complete genome	Escherichia coli	89.8	391	100%	2e-14	100.00%	5288508	CP101307.1
Deinococcus grandis ATCC 43672 DNA, complete genome	Deinococcus gra...	87.9	258	100%	7e-14	98.00%	3241502	AP021849.1

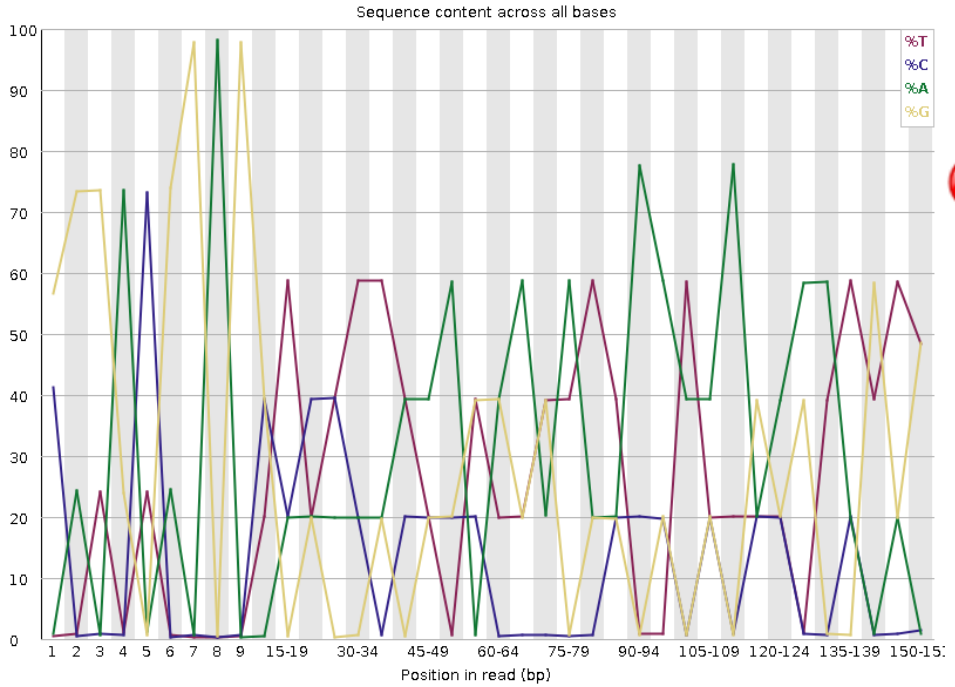
Overrepresented Specific Sequences

Which of these libraries would you expect to flag with overrepresented specific sequences?

A) Whole Genome Bisulfite Library
B) Amplicon Library
C) RNAseq Library

Overrepresented Specific Sequences

Which of these libraries would you expect to flag with overrepresented sequences?



Overrepresented sequences

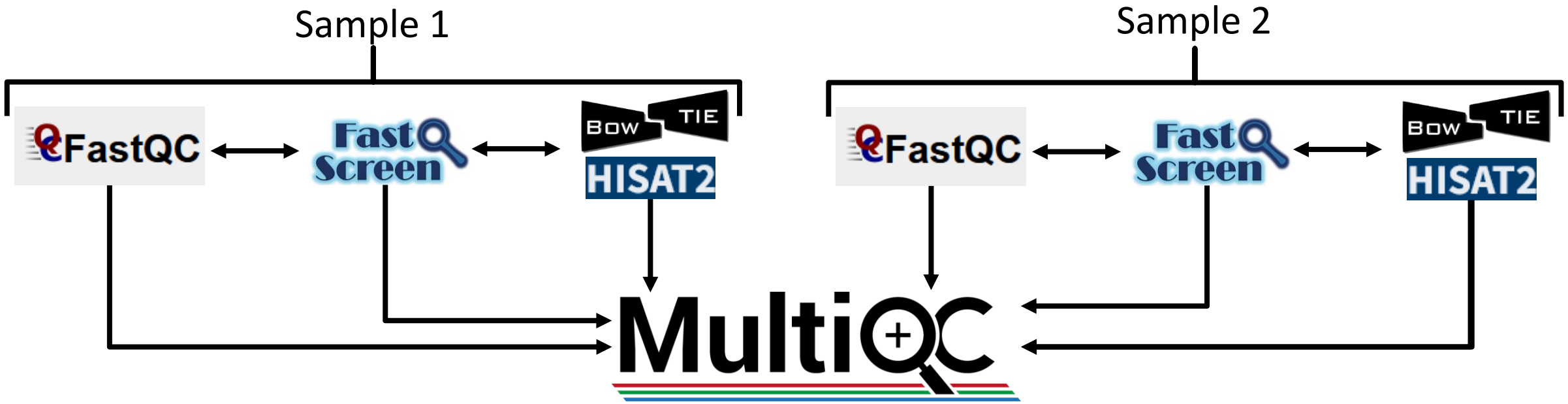
Sequence	Count	Percentage	Possible Source
CGGACGGAGGGCCTATTTCCCATGATTCCTTCATATTTGCATATACGATA	70194	37.84369541308145	No Hit
GGGACGGAGGGCCTATTTCCCATGATTCCTTCATATTTGCATATACGATA	55496	29.919561795087446	No Hit
GATGTAGAGGGCCTATTTCCCATGATTCCTTCATATTTGCATATACGATA	40221	21.684350132625994	No Hit
TGGACGGAGGGCCTATTTCCCATGATTCCTTCATATTTGCATATACGATA	281	0.1514955467857066	No Hit

Assessing Consistency

Babraham Bioinformatics

Aggregated Statistics

Individual QC reports are useful but helpful to have a wider picture

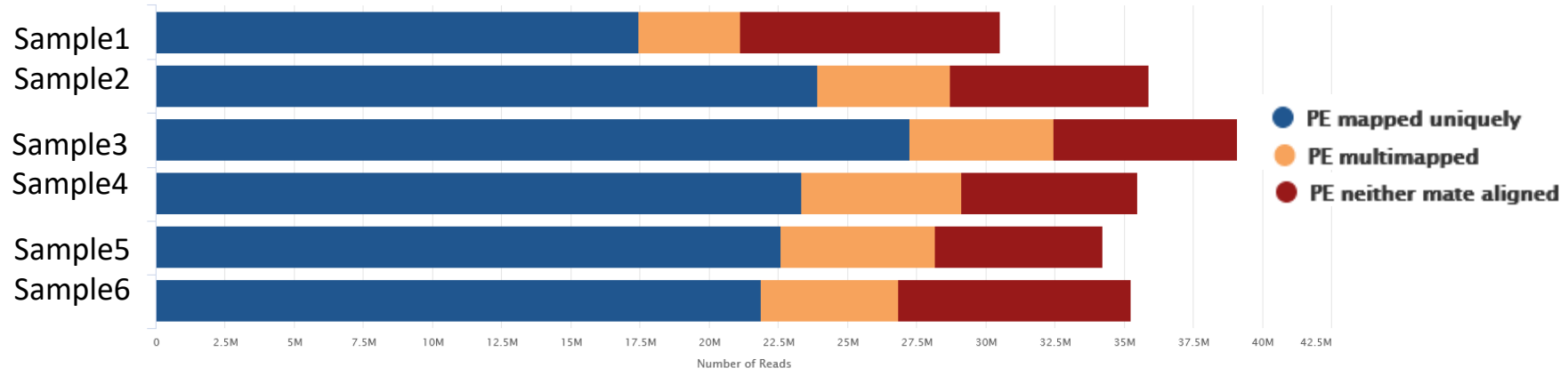


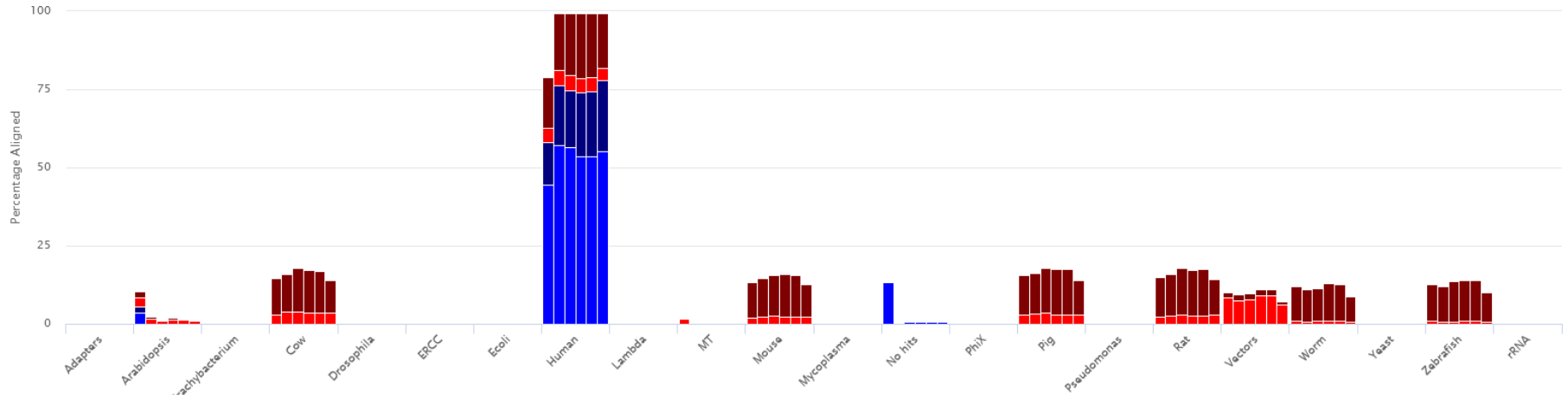
Aggregate and plot range of QC stats together

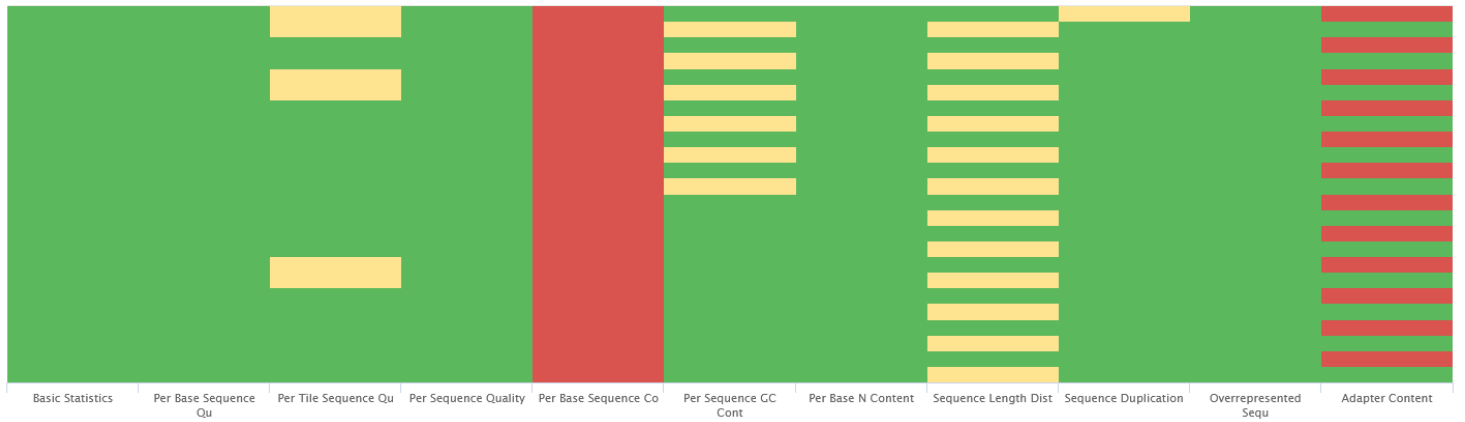
MultiQC currently has modules to support 128 different bioinformatics tools, https://multiqc.info/modules/

MultiQC

- General Stats
- Bowtie 2 / HiSAT2
- Cutadapt
  - Filtered Reads
  - Trimmed Sequence Lengths (3')
- FastQ Screen
- FastQC
  - Sequence Counts
  - Sequence Quality Histograms
  - 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
  - Status Checks





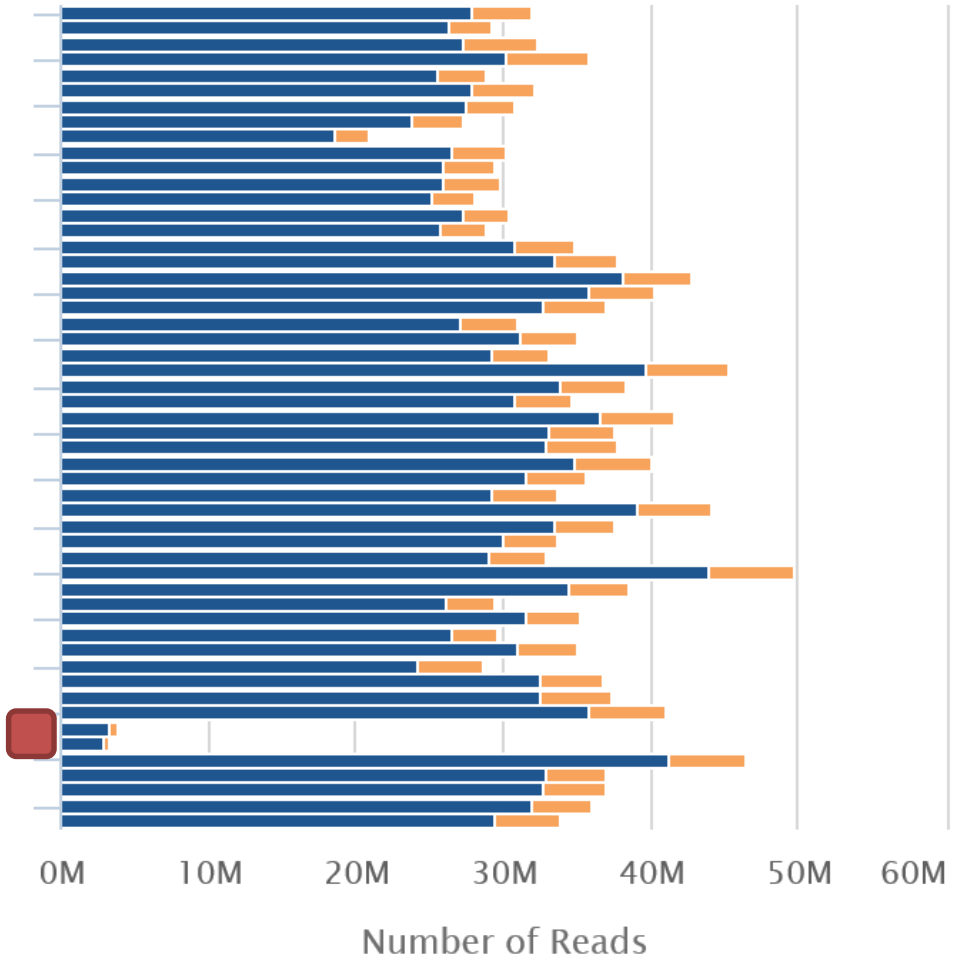


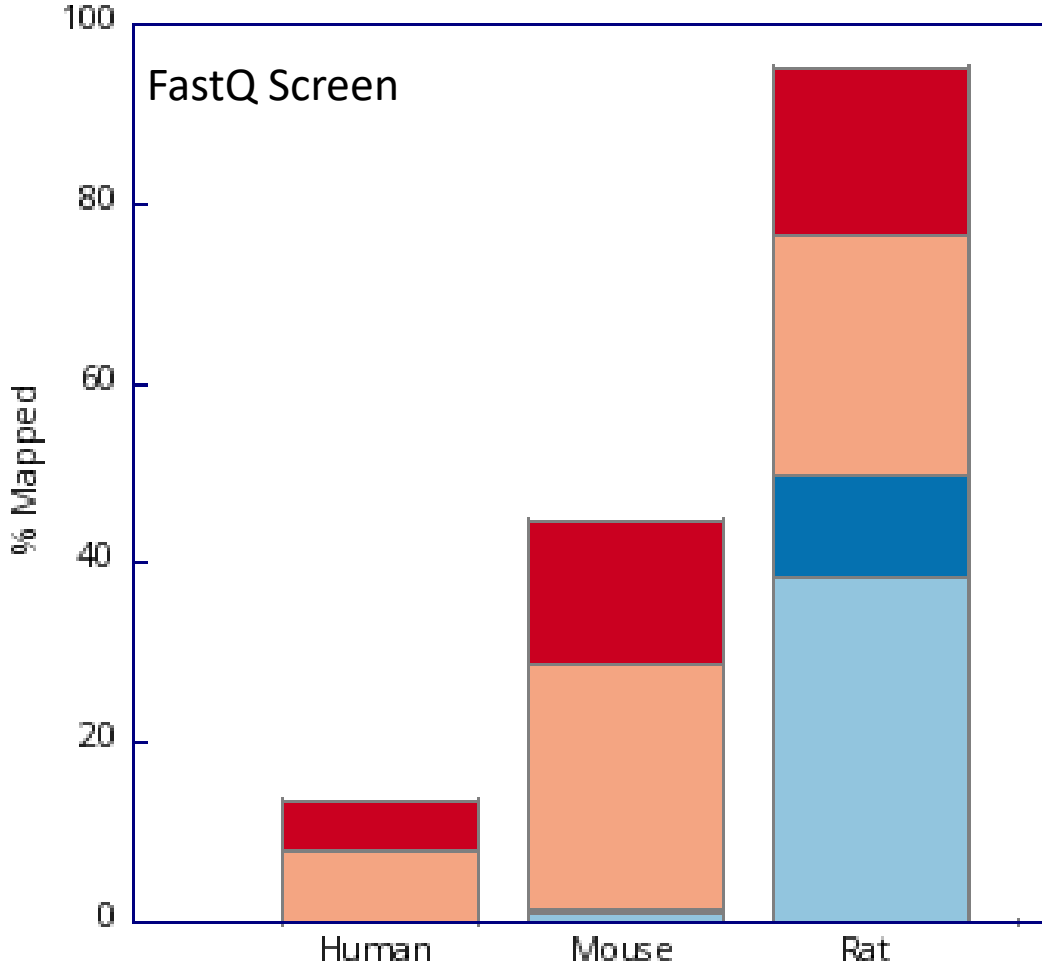
6 Samples, 24 files:

Paired end
Pre/post trimming

Aggregated Mapping Stats

Identify local QC problems by spotting samples that behave differently

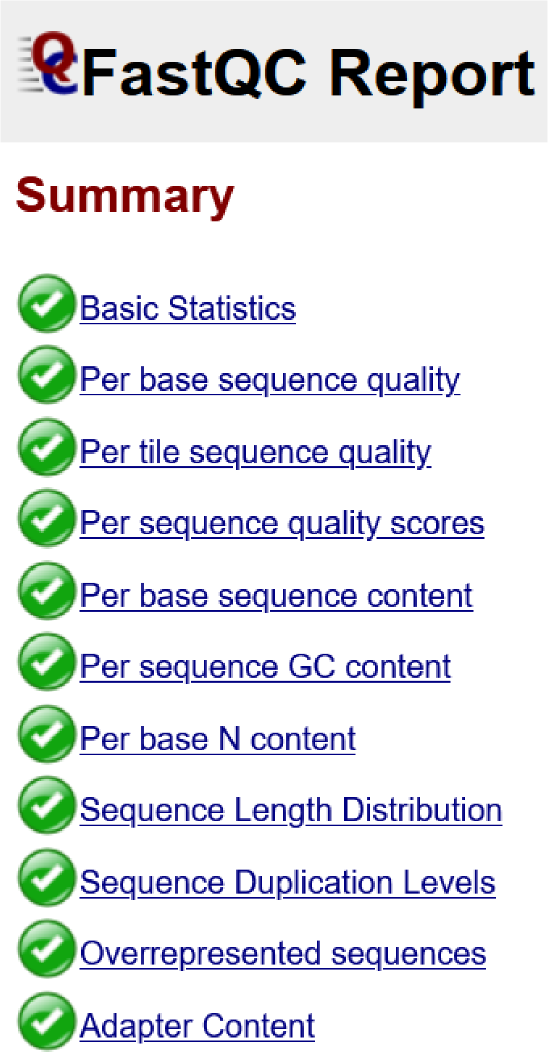


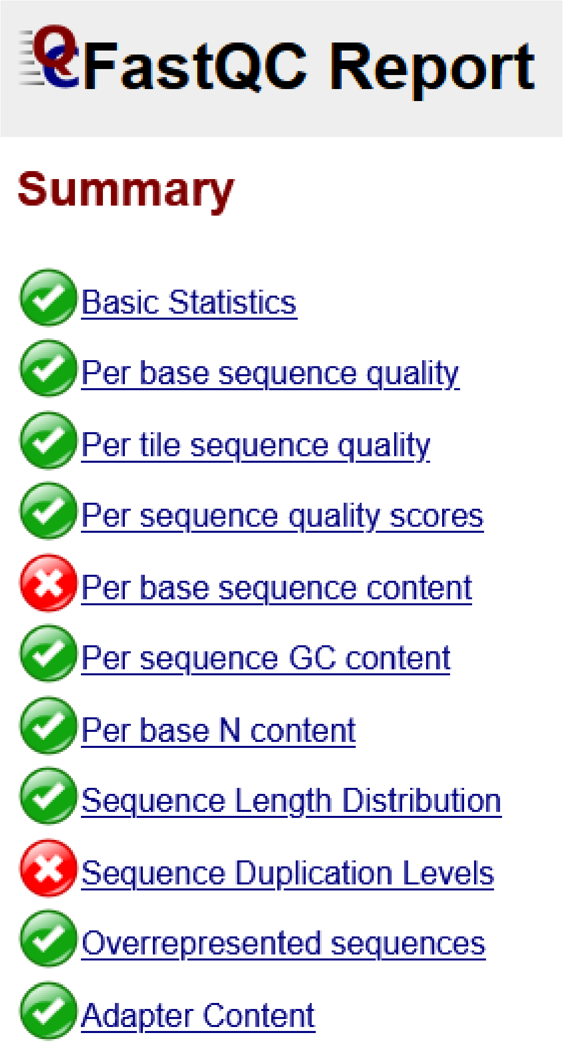


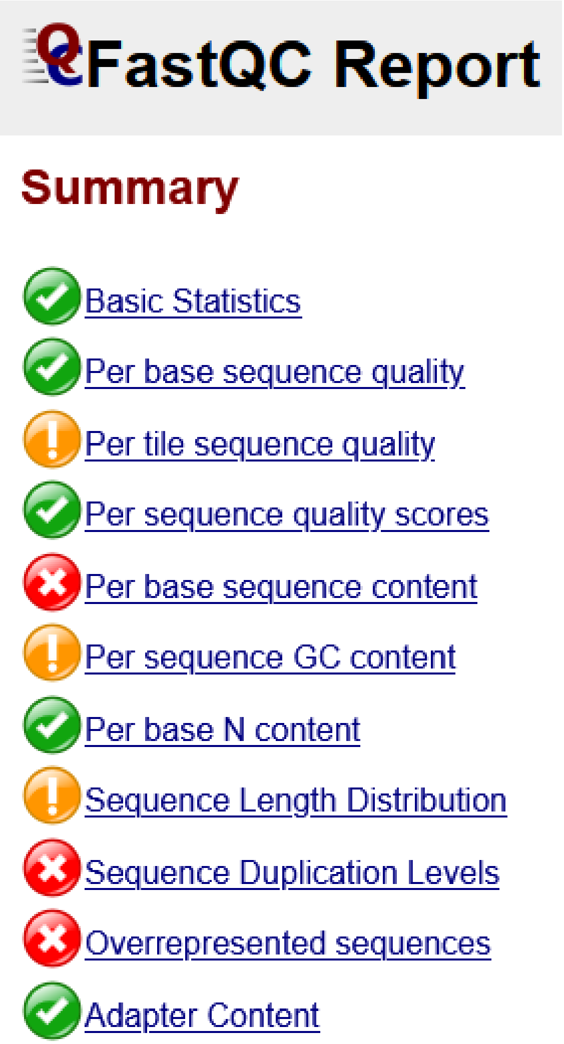
Putting QC into Practice

Babraham Bioinformatics

Expectations and Observations are Key



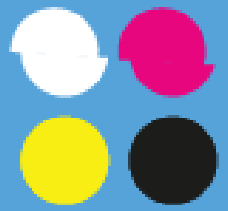




Can you tell if these libraries are any good?

QC In A Nut-Shell

Good Science ☺



QCFAIL.com

Articles about common next-generation sequencing problems

Exercise: Assess the Quality of Public Sequencing Datasets