Quality Control for Sequencing Experiments

v2024-11

Simon Andrews, Sarah Inglesfield

simon.andrews@babraham.ac.uk

sarah.inglesfield@babraham.ac.uk

Babraham Bioinformatics

Interests in QC

QC packages

FastQC

Librarian

Fast Screen

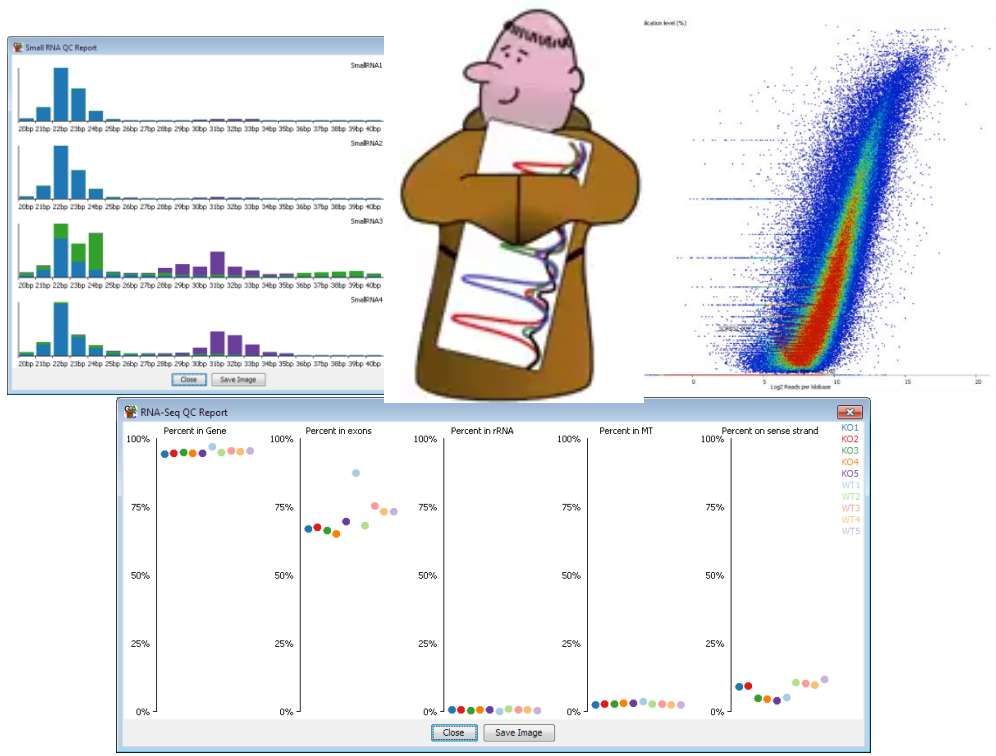
Application specific QC

Bismark

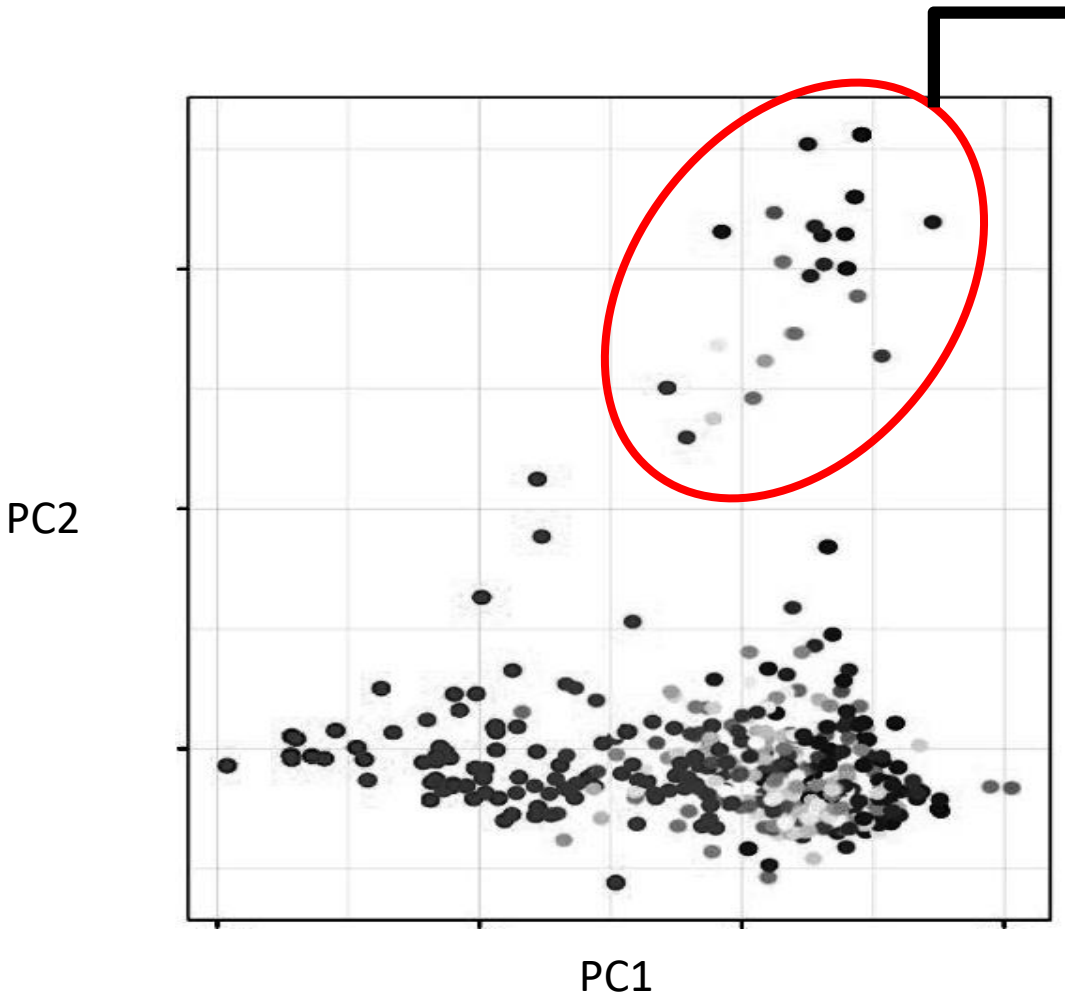
HiCUP

Data visualisation QC

SeqMonk



What is the Point of QC? An Example…



PC2 Genes (85 total)

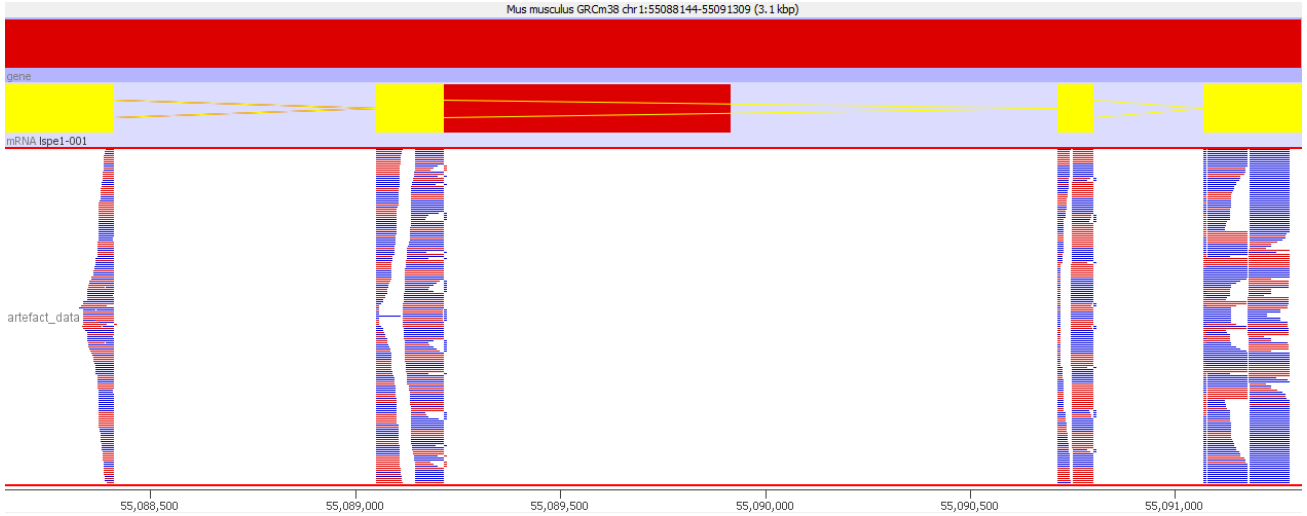
- No clear biological theme
- No clear connection to system

What is going on?

What is the Point of QC? An Example…

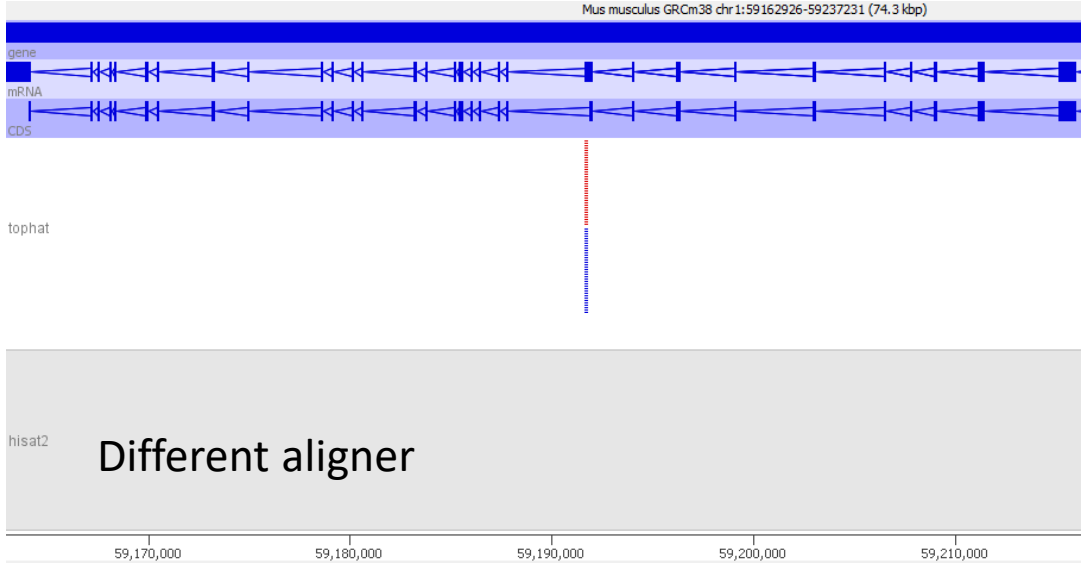
Expectations

Normal RNA-seq Data for a Gene



Observations

RNA-seq Data for a PCA Gene



Conclusion: The original separation was a technical artefact of no biological interest

What is the Point of QC?

Technical Problems …

- Don’t always cause pipelines to fail
- Don’t prevent hits being generated
- These hits can look biologically real

Real biology…

- Can cause unexpected, interesting behaviour of data

Set Pipelines can miss things….

QC Saves Time, Effort and Money!!

- Better to know asap what you’re dealing with
- Want to be sure any follow-up work will be worth it

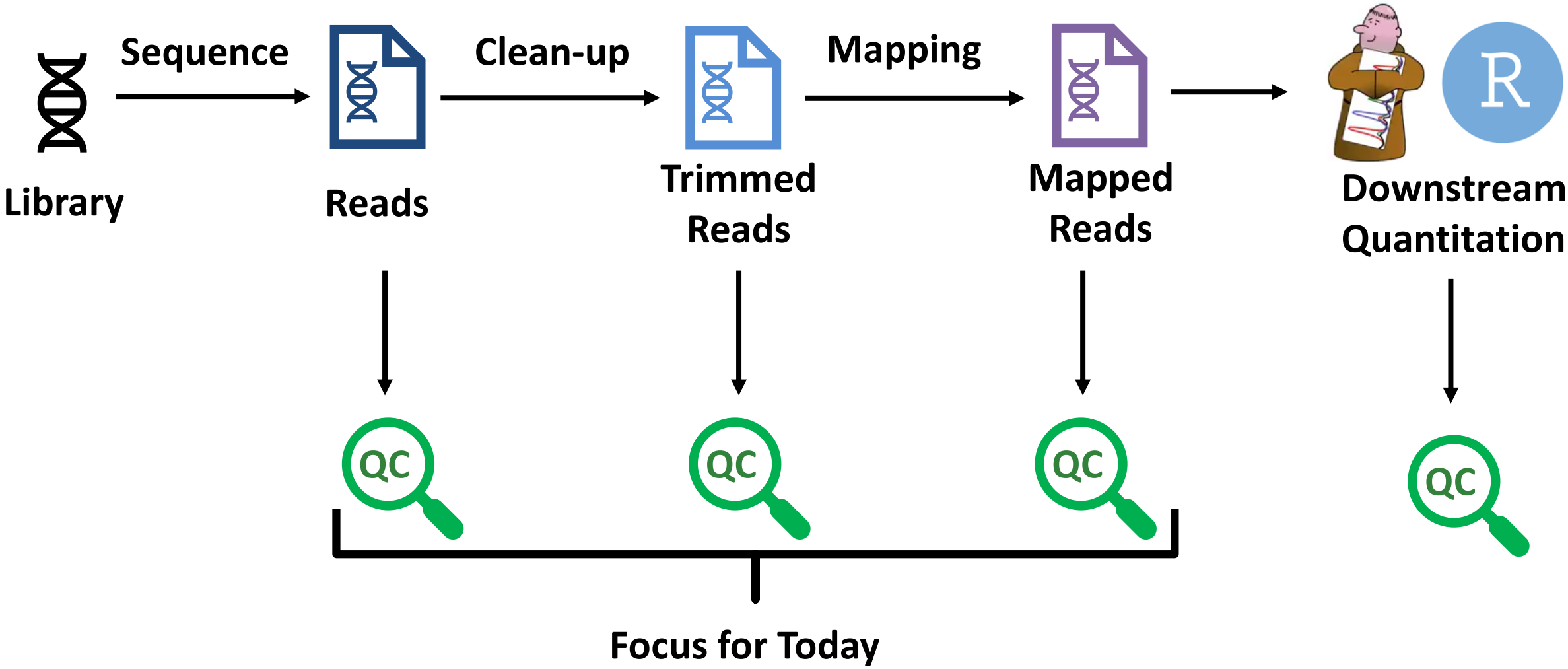
Course Structure

How Does Sequencing Work?	What can QC tell us?
• Fundamentals of Illumina Sequencing • The Format of Sequencing Data • How QC Programmes Fit In	• Universal metrics • Library Dependent metrics • Consistency

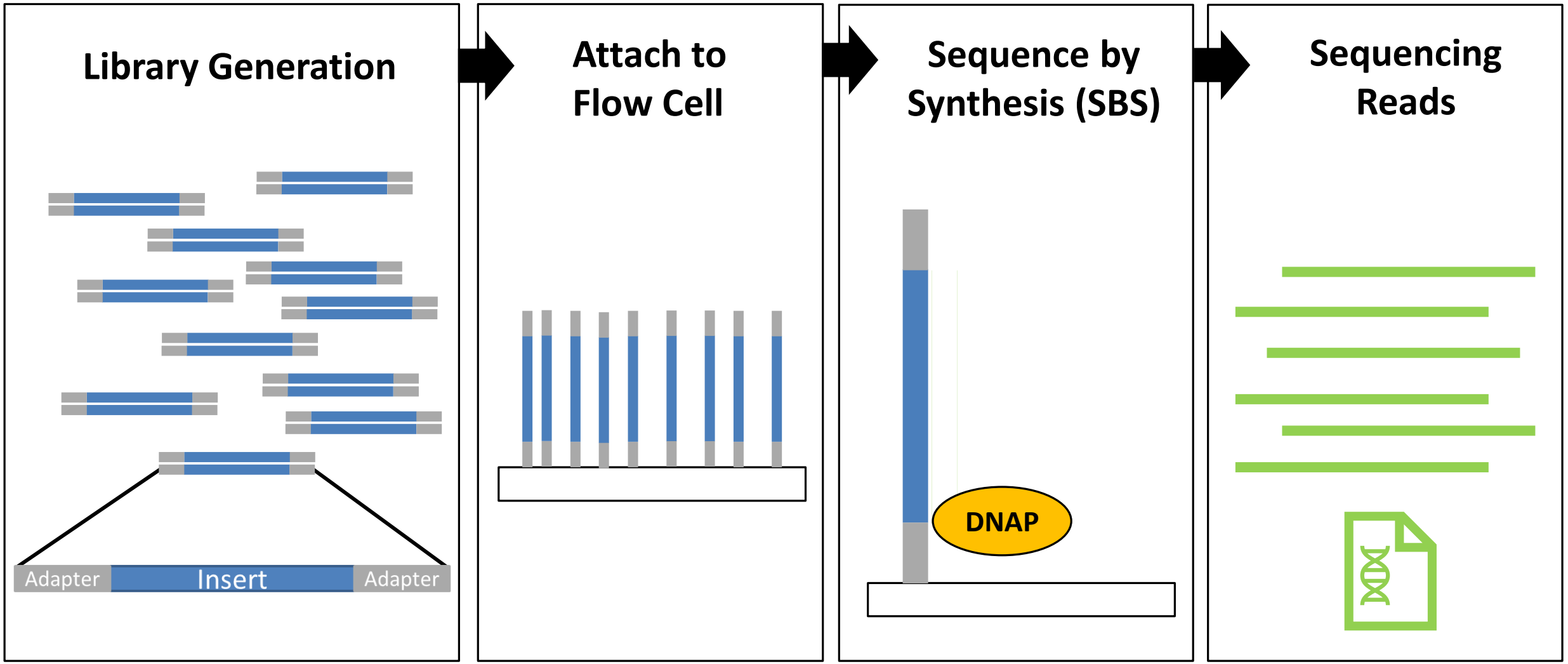
How Does (Illumina) Sequencing Work?

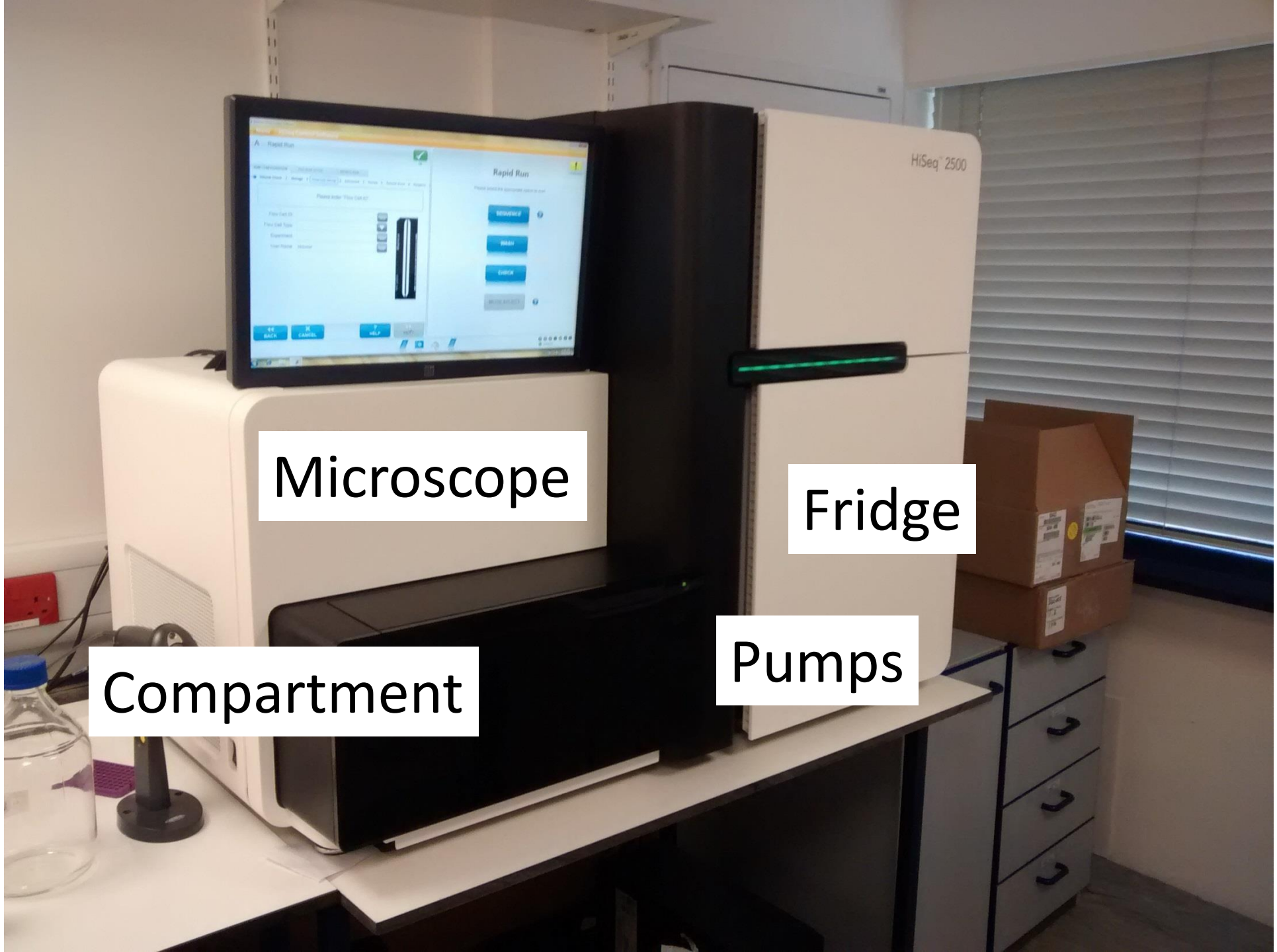
Babraham Bioinformatics

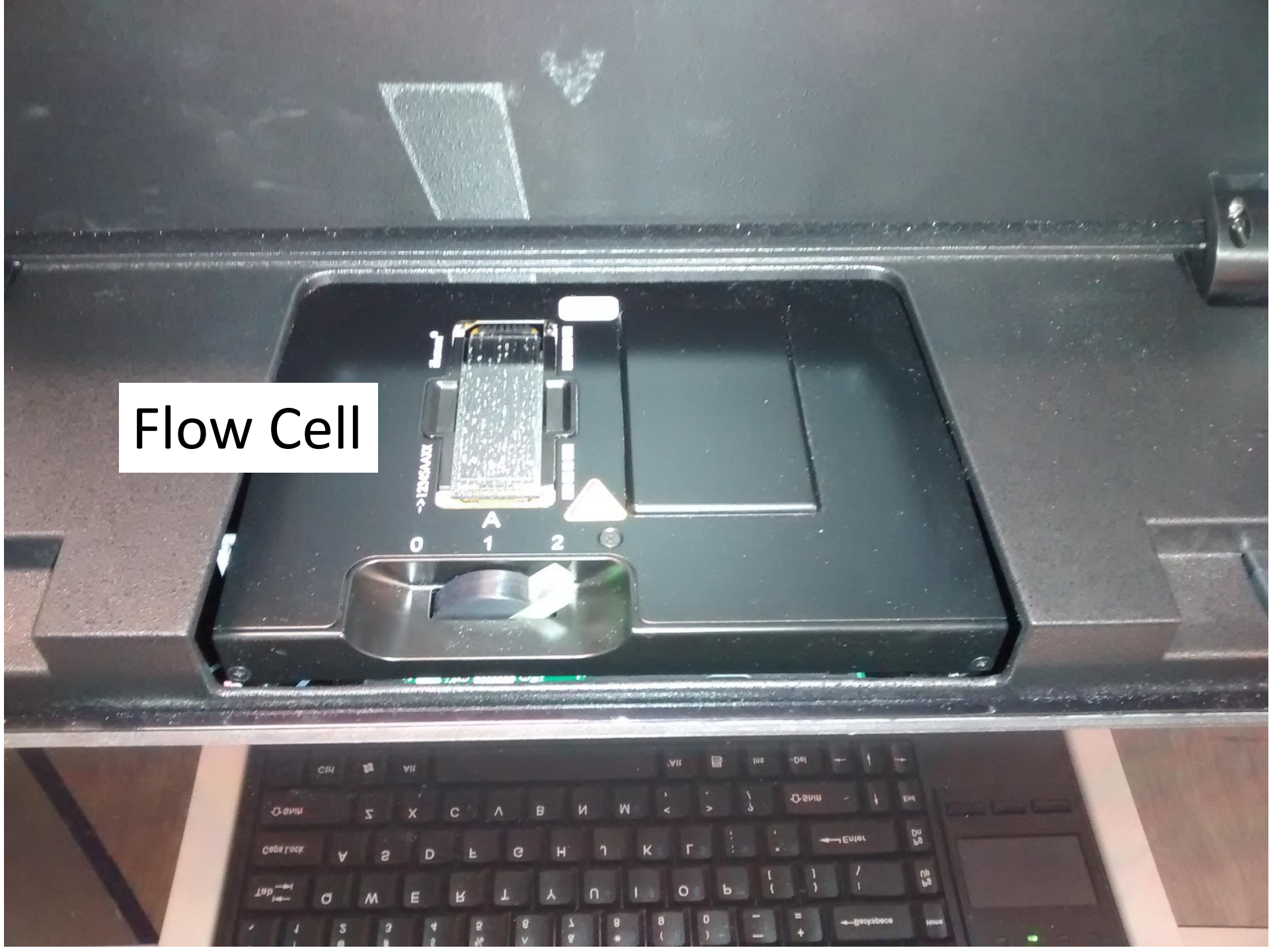
Processing Sequencing Data



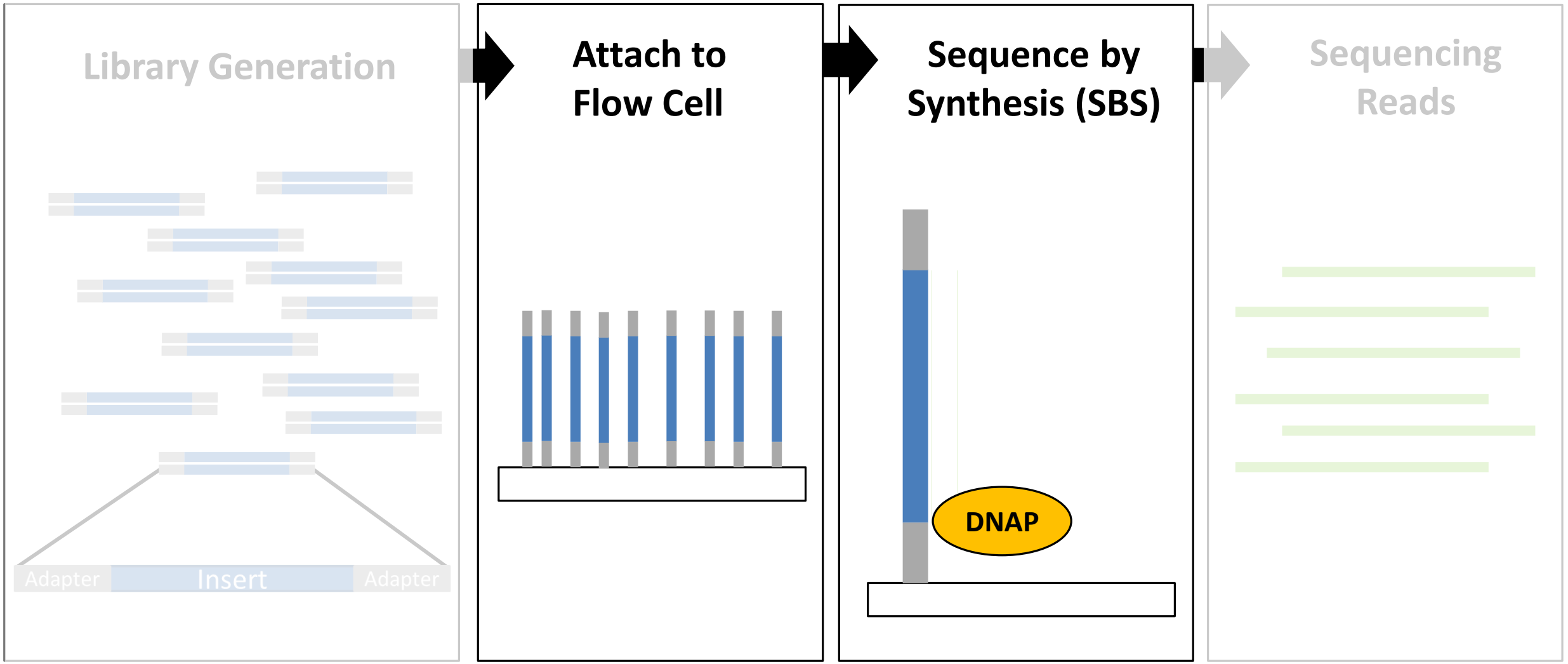
Illumina Sequencing: An Overview



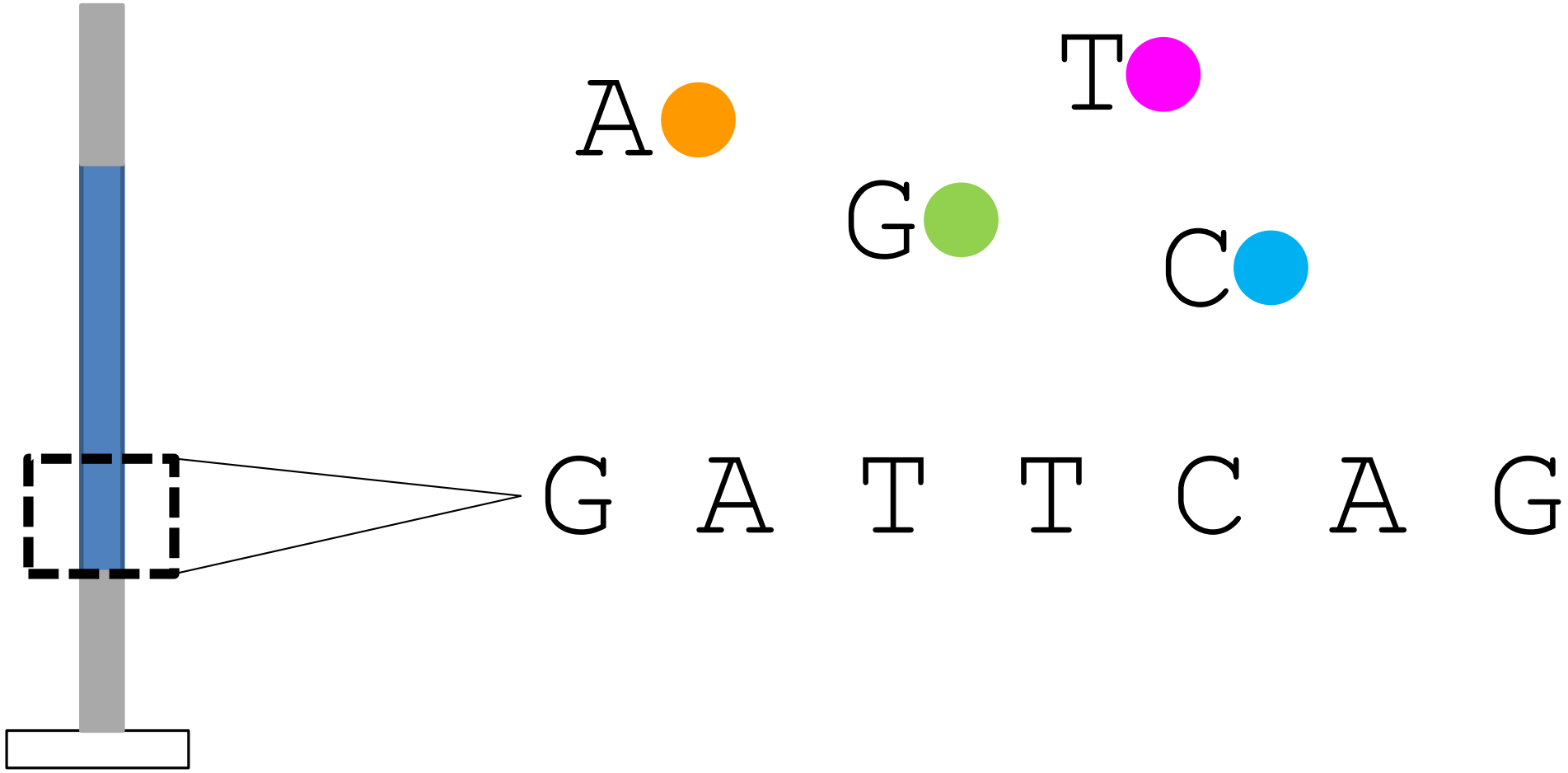

Microscope
Fridge
Compartment
Pumps
HiSeq 2500
Rapid Run


Flow Cell
A
0
1
2

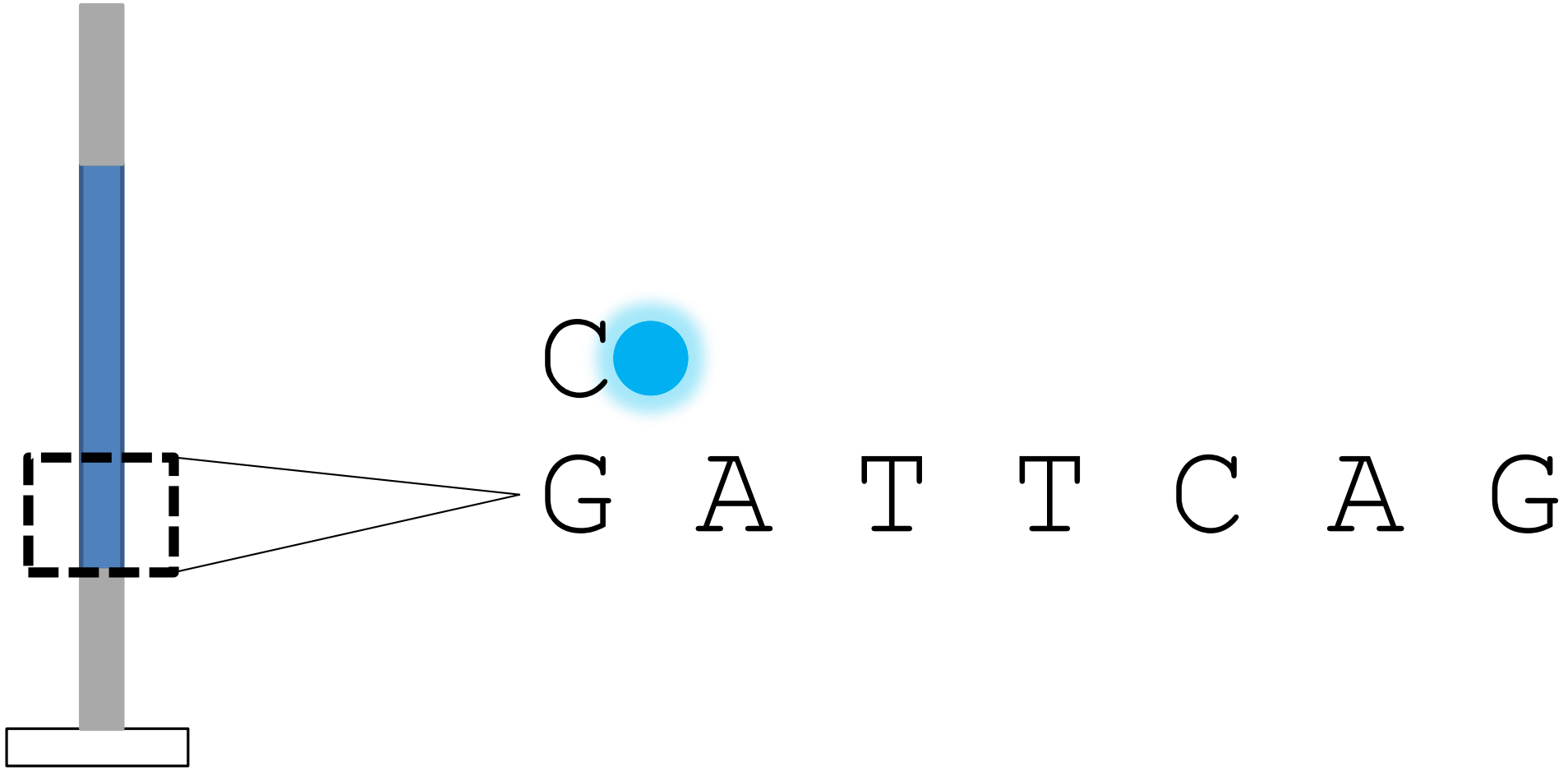
Illumina Sequencing: An Overview



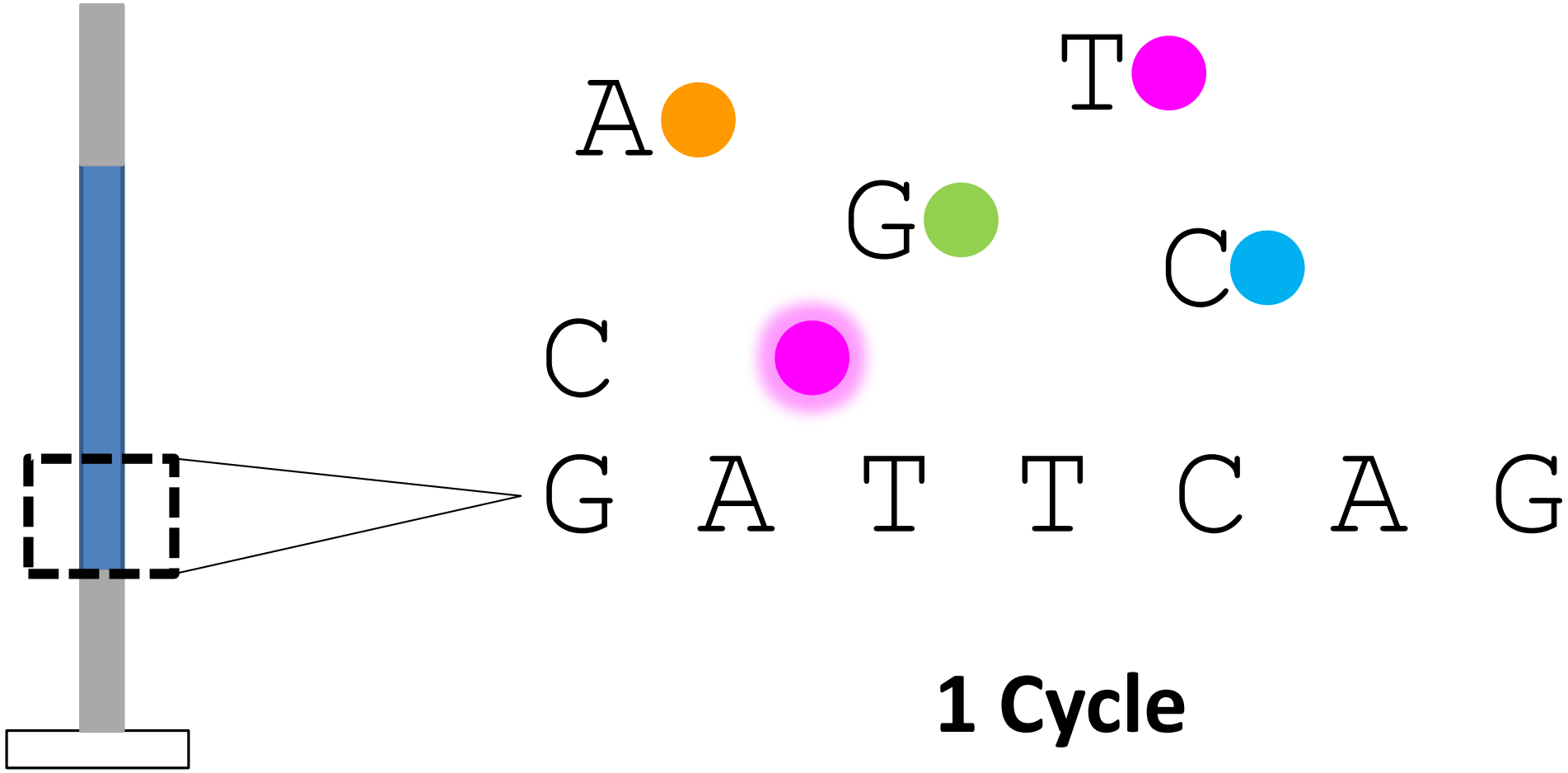
SBS For a Single Molecule



SBS For a Single Molecule

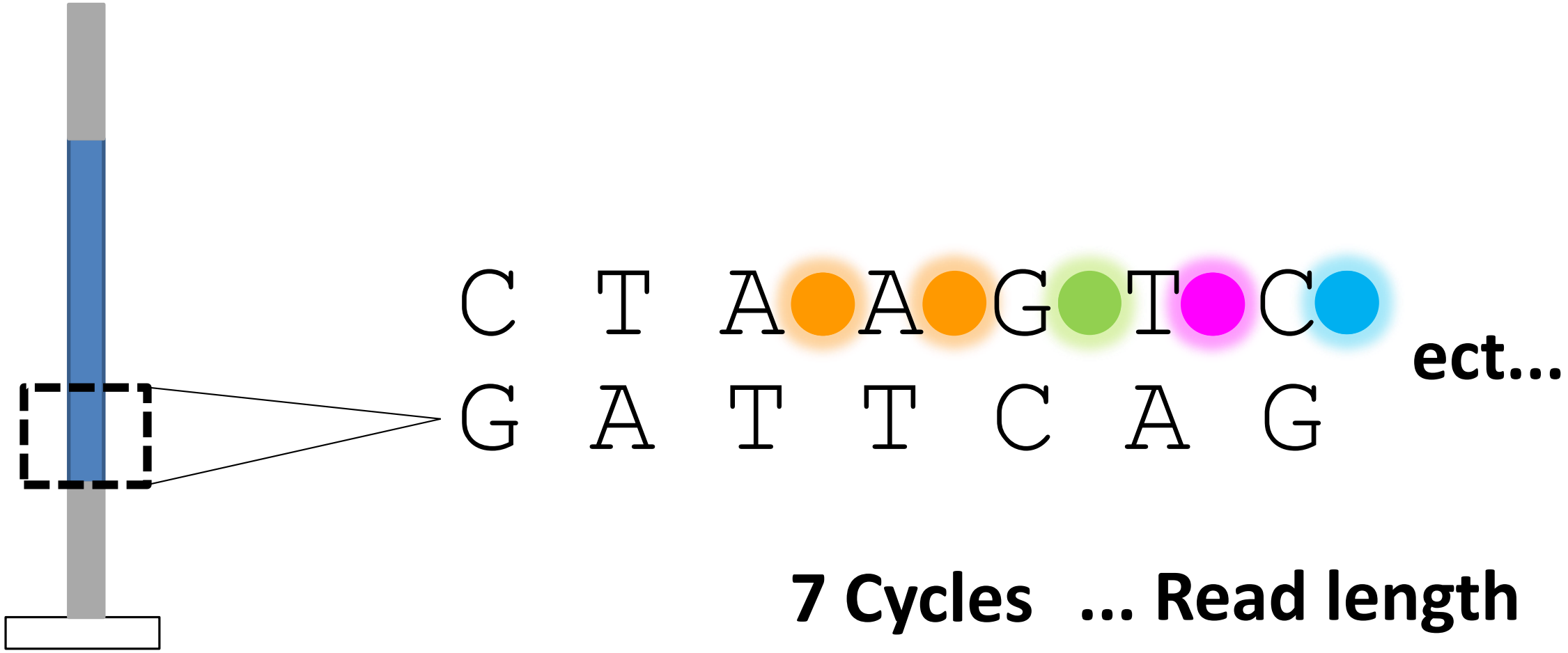


SBS For a Single Molecule

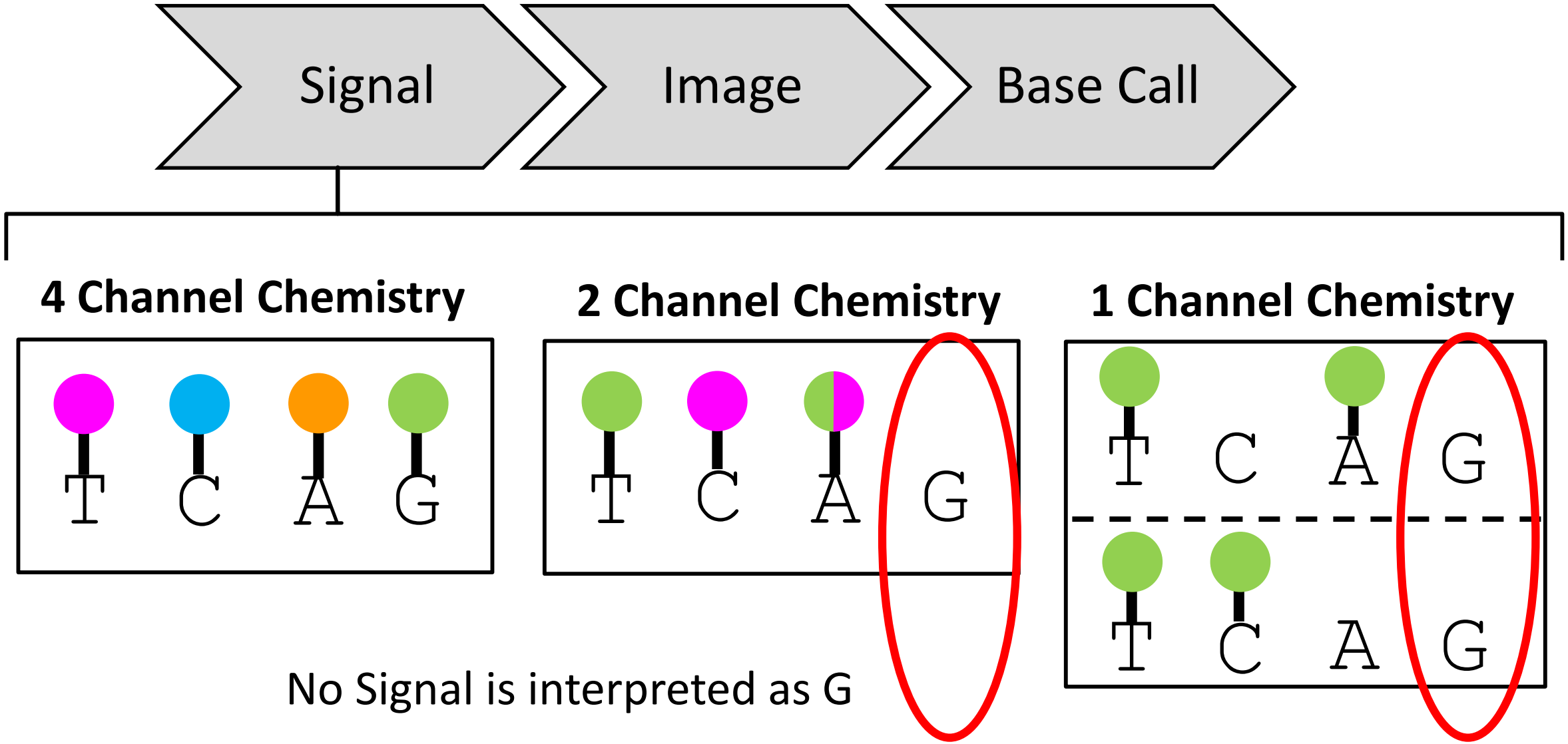


1 Cycle

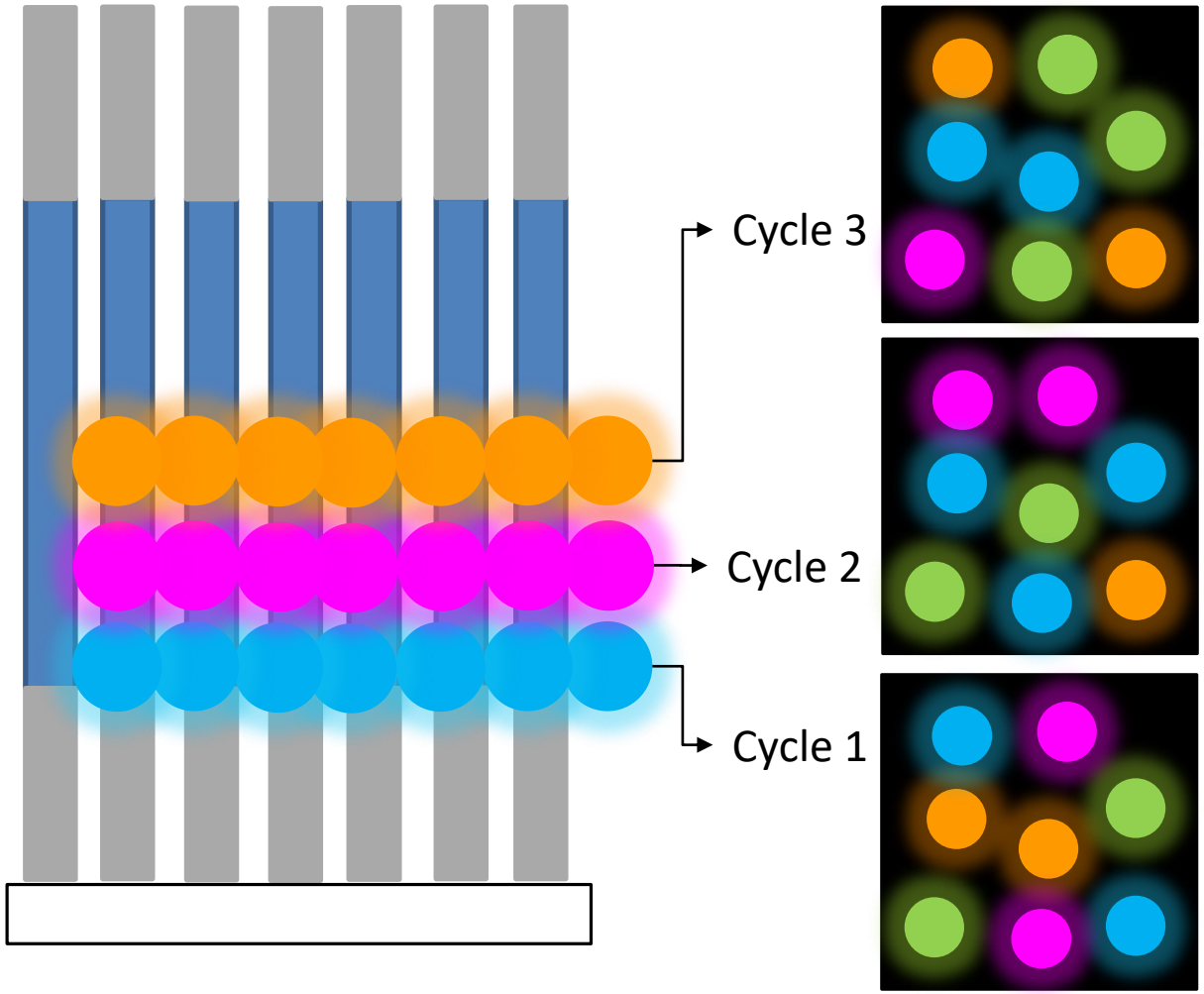
SBS For a Single Molecule



Comparing Chemistry

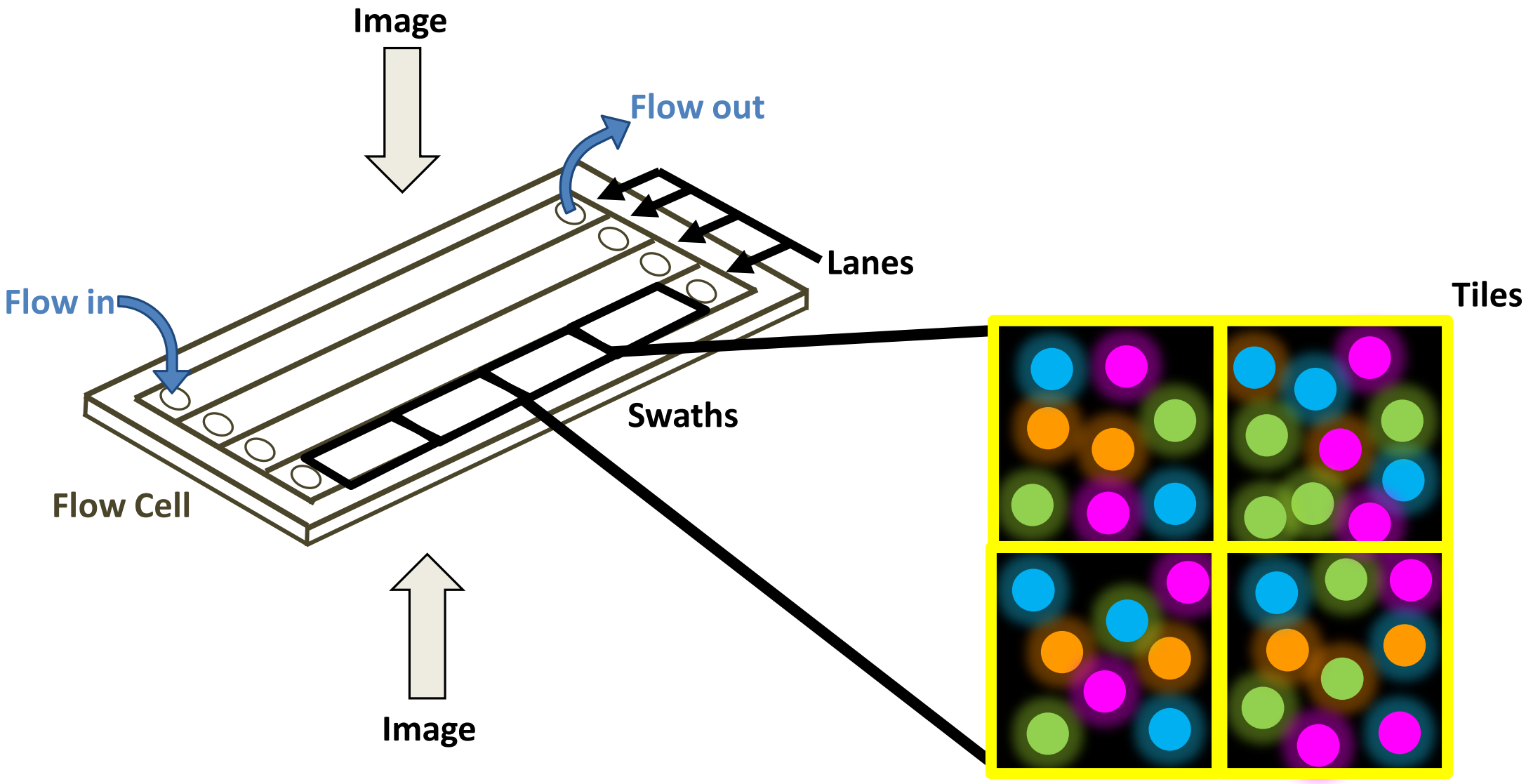


Detecting a Signal

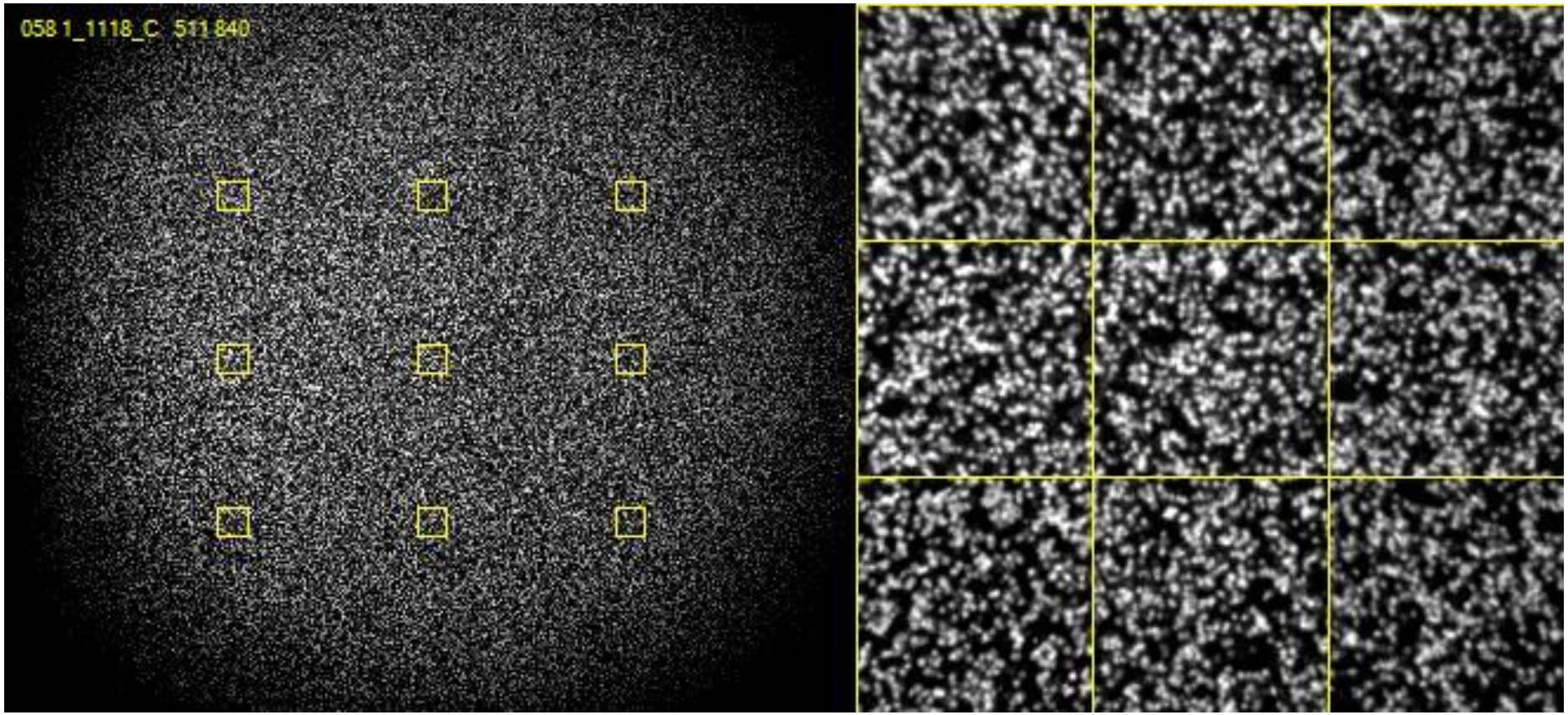


- One Molecule isn’t Enough
- Amplify to generate Cluster
- Cluster Molecules sequenced
- Multiple Clusters on a Flow Cell

Flow Cell Imaging

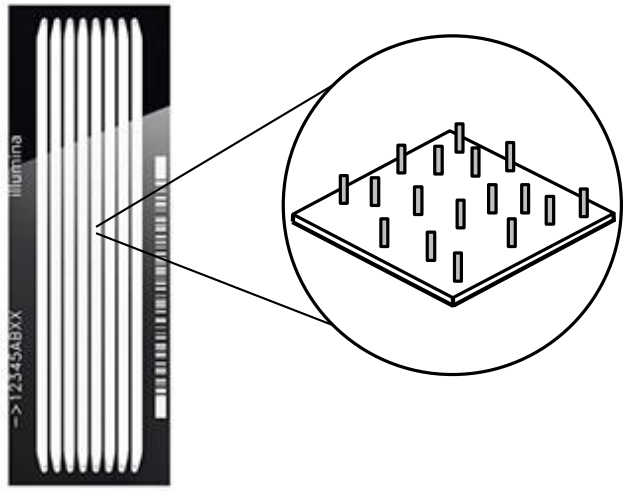


Real Illumina Sequence Data

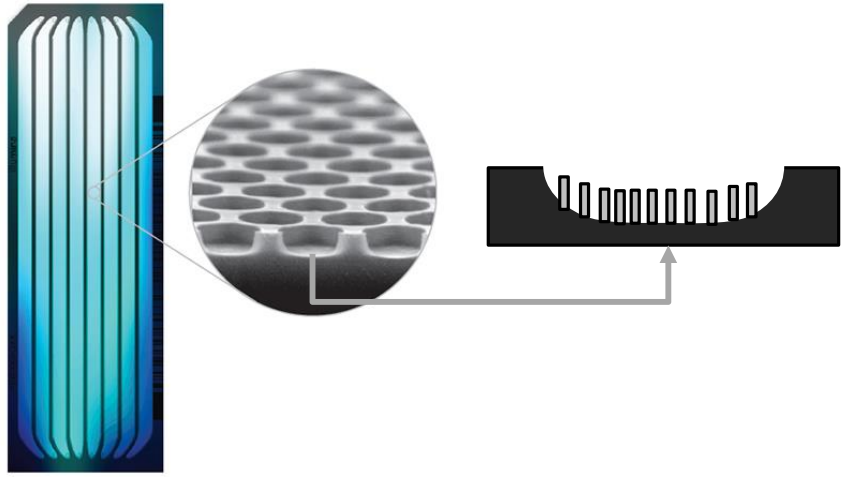


Creating Clusters

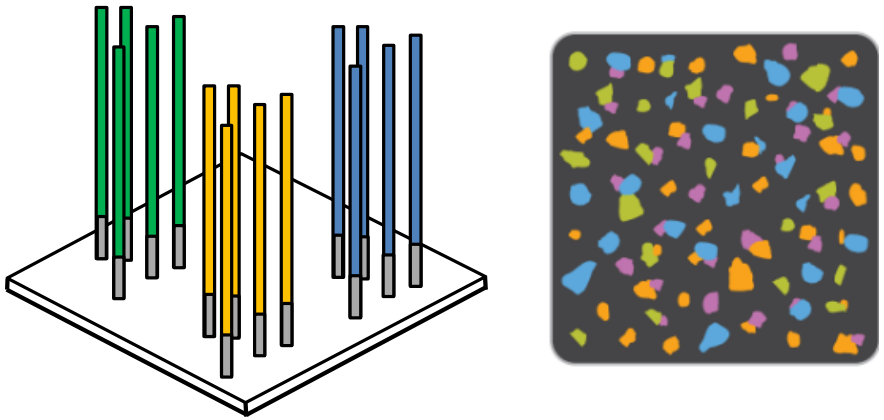
Non Patterned Flow Cell



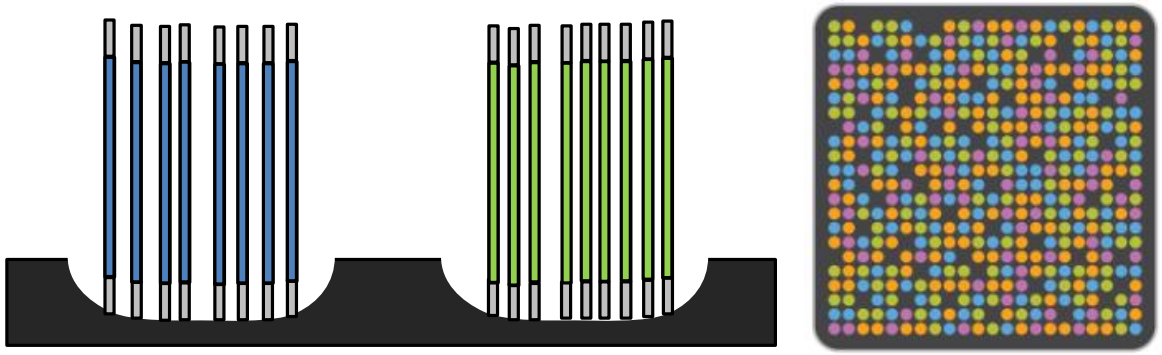
Patterned Flow Cell



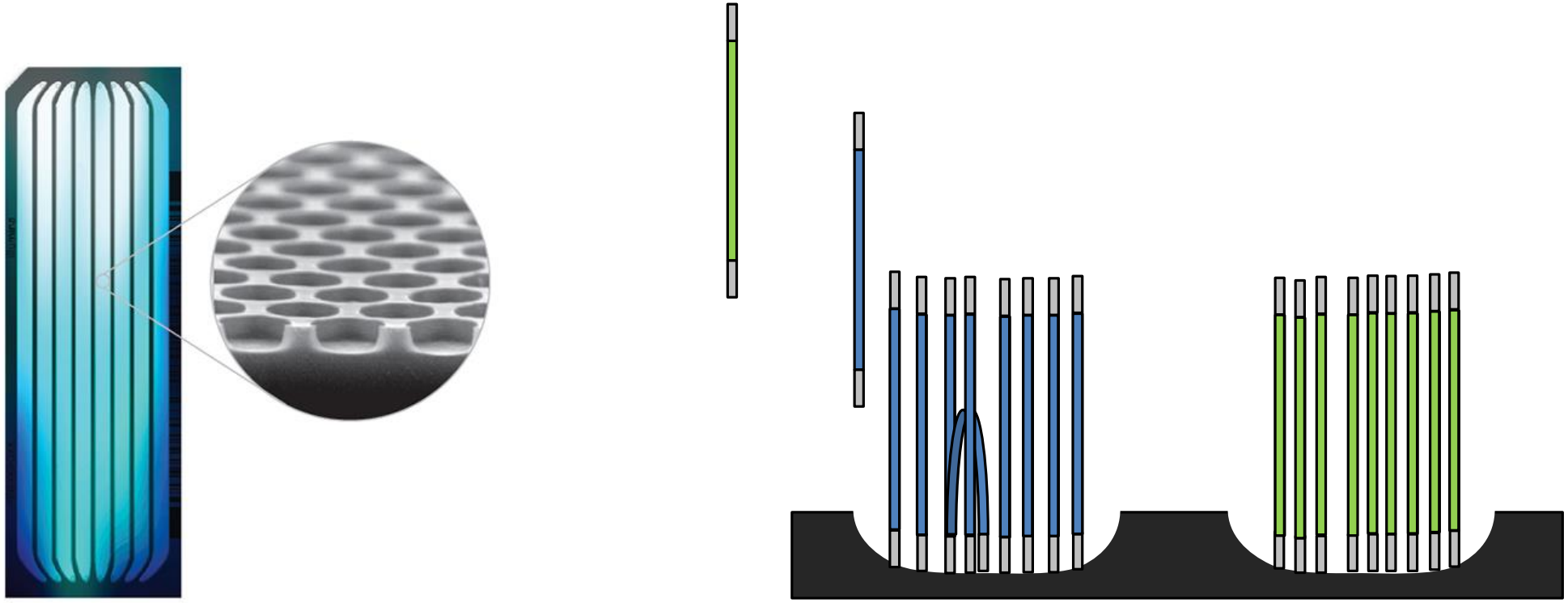
Random Clusters



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
- Small amount of starting material

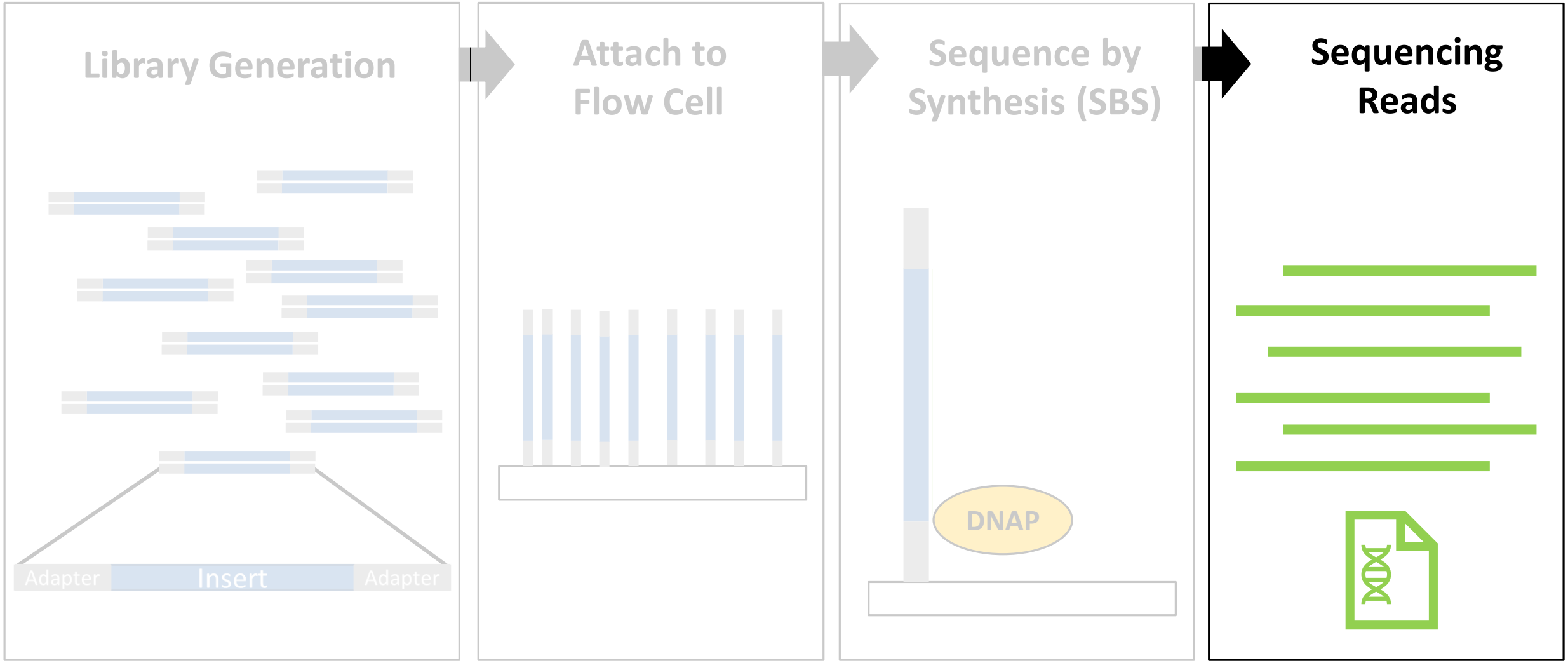
Bad

- 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 or Paired?

Insert R1

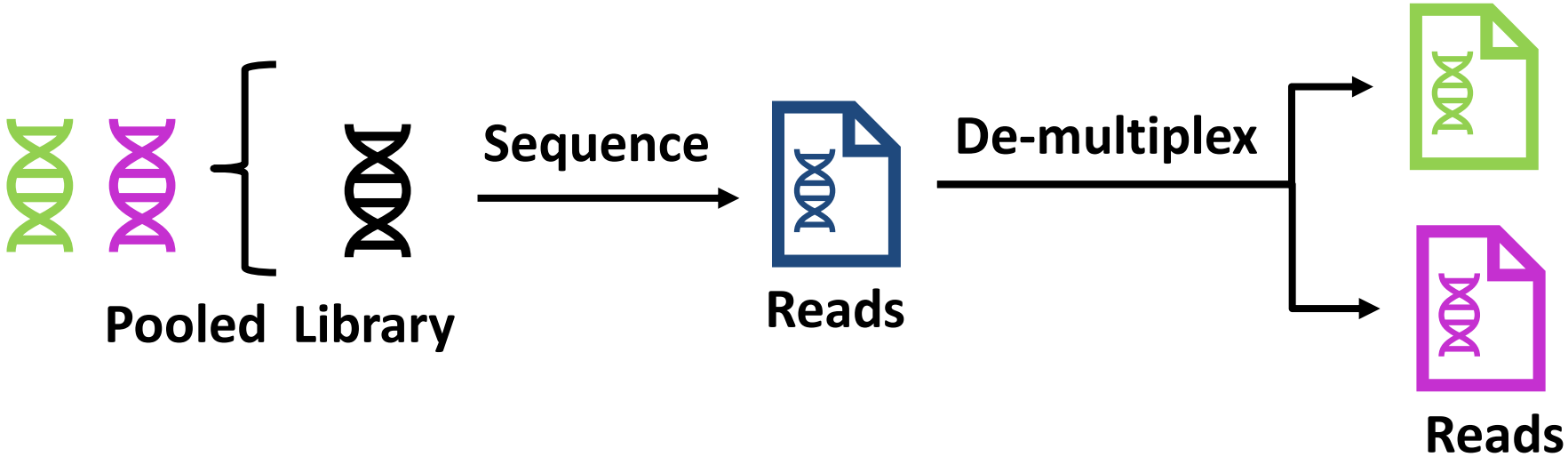
Single End
(1 file)

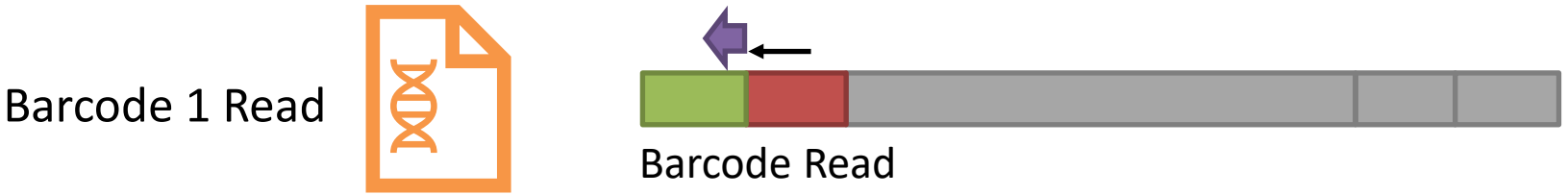
+

Paired End
(2 files)

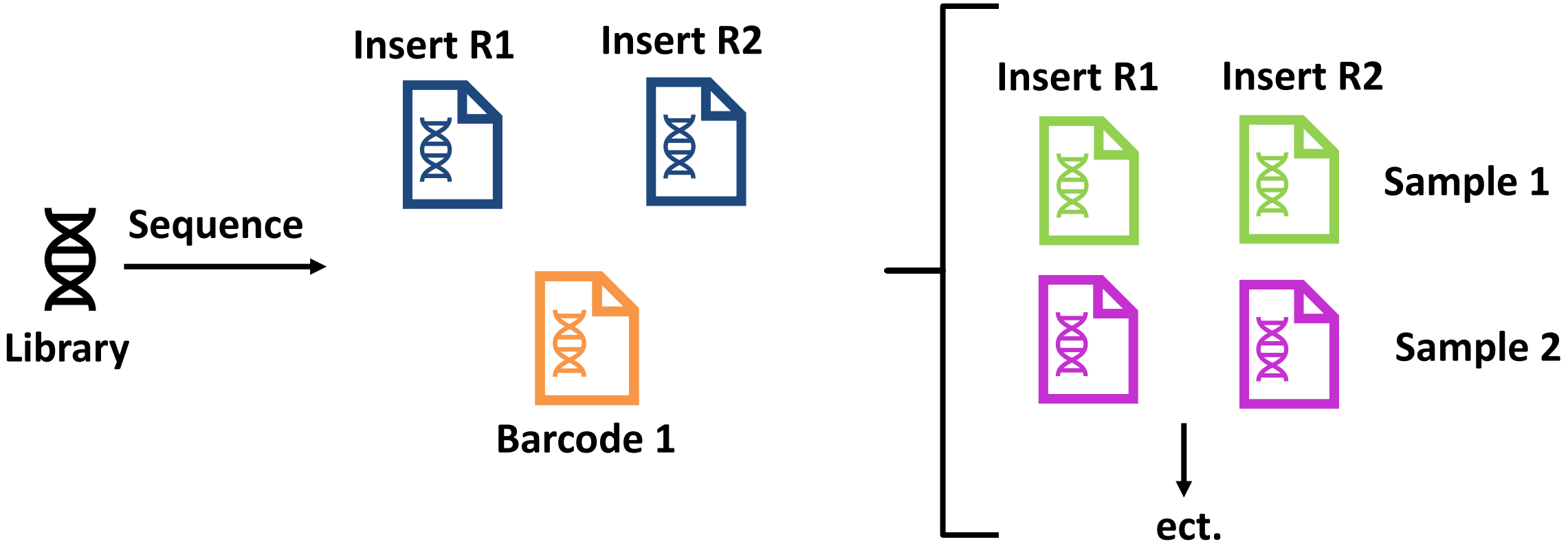
Insert R2

What Reads Do You Get: Multiplexed?





So For a Paired End, Single Index Multiplexed Library….



So what is actually in these files?

FastQ Format Data

Sequence 1

```
@HWUSI-EAS611:34:6669YAAXX:1:1:5069:1159 1:N:0:
TCGATAATACCGTTTTTTTCCGTTTGATGTTGATACCATT
+
IIHIIHIIIIIIIIIIIIIIIIIIIIIIIIHIIIIHIIIII
```

Sequence 2

```
@HWUSI-EAS611:34:6669YAAXX:1:1:5243:1158 1:N:0:
TATCTGTAGATTTCACAGACTCAAATGTAAATATGCAGAG
+
DF=DBD<BBFGGGGGGGBD@GGGD4@CA3CGG>DDD:D,B
```

Sequence 3

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

A Single FastQ Entry

1. `@HWUSI-EAS611:34:6669YAAXX:1:1:5266:1162 1:N:0:`
2. `GGAGGAAGTATCACTTCCTTGCCTGCCTCCTCTGGGGCCT`
3. `+`
4. `:GBGGGGGGGGGGDGGDEDGGDGGGGDHHDHGHHGBGG:GG`

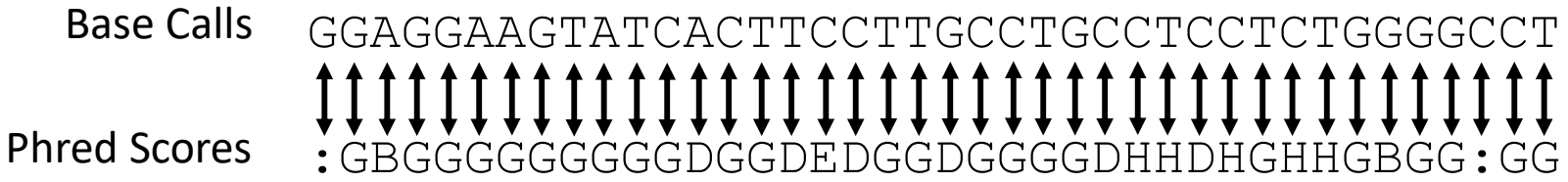
1. Header - starts with @
2. Base calls (can include N or IUPAC codes)
3. Mid-line - starts with + usually empty
4. Quality scores (= Phred Scores)

Illumina Header Sections (Line 1)

```
@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 (Line 4)



How it's calculated:

- Start from (p) - the probability that the reported call is incorrect
- 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

Higher Phred Score
Higher Confidence

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

Interpreting Phred Scores

: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

Interpreting Phred Scores

: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
```

Interpreting Phred Scores

: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

Interpreting Phred Scores

: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
```

Interpreting Phred Scores

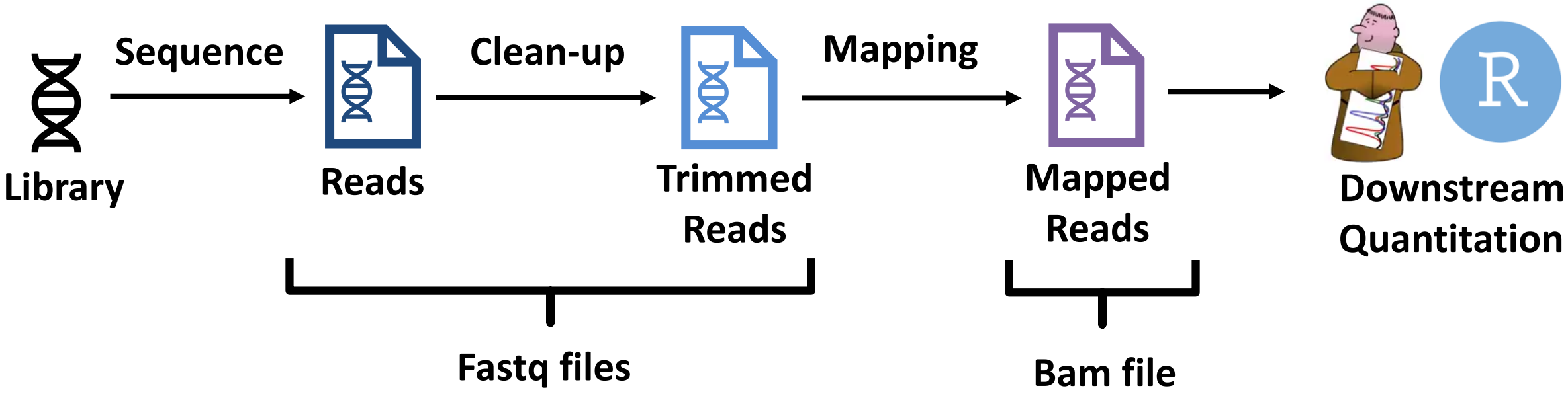
:GBGGGGGGGGGDGGDEDGGDGGGGDHHDHGHHGBGG:GG

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
```

Further Processing: Beyond Raw FastQ Files



Aligned Data – BAM Files

Expanded file containing alignment data (+ FastQ file details), in 2 Sections:

Header Section

`samtools view -H **filename.bam**`

General Metadata

- Information on:
  - Technical details of File
  - Reference
  - Programmes Involved
- Each line begins with @ + 2 letter code

Alignments Section

`samtools view **filename.bam** | less`

Details of sequence alignments:

- One line per read
- 11 required columns (tab separated)
- Information on:
  - Sequence
  - Where alignments and how it aligns

Need special programs to read, normally 'samtools'

BAM Header Section: Example

```
@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"
```

Header line Format Version Sorting Order

Reference Sequence Name Length

Program Information ID Name Version Command

BAM Alignments Section: Example

```
HWI-D00436:394:CBGLBANXX:1:1101:1222:1861       16      chr18      57944851
60      50M     *       0       0
AAAAGATCTCTTGATTTAGAATTTTCTCTCAAATGTGAGGGACTTTTATN
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGBA<=#      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
```

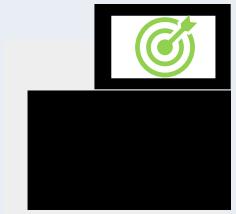
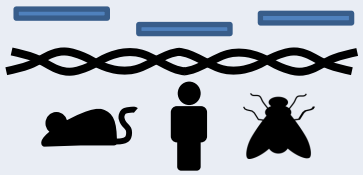
Sections

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) — `GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGBA<=#`
12. Other Tags

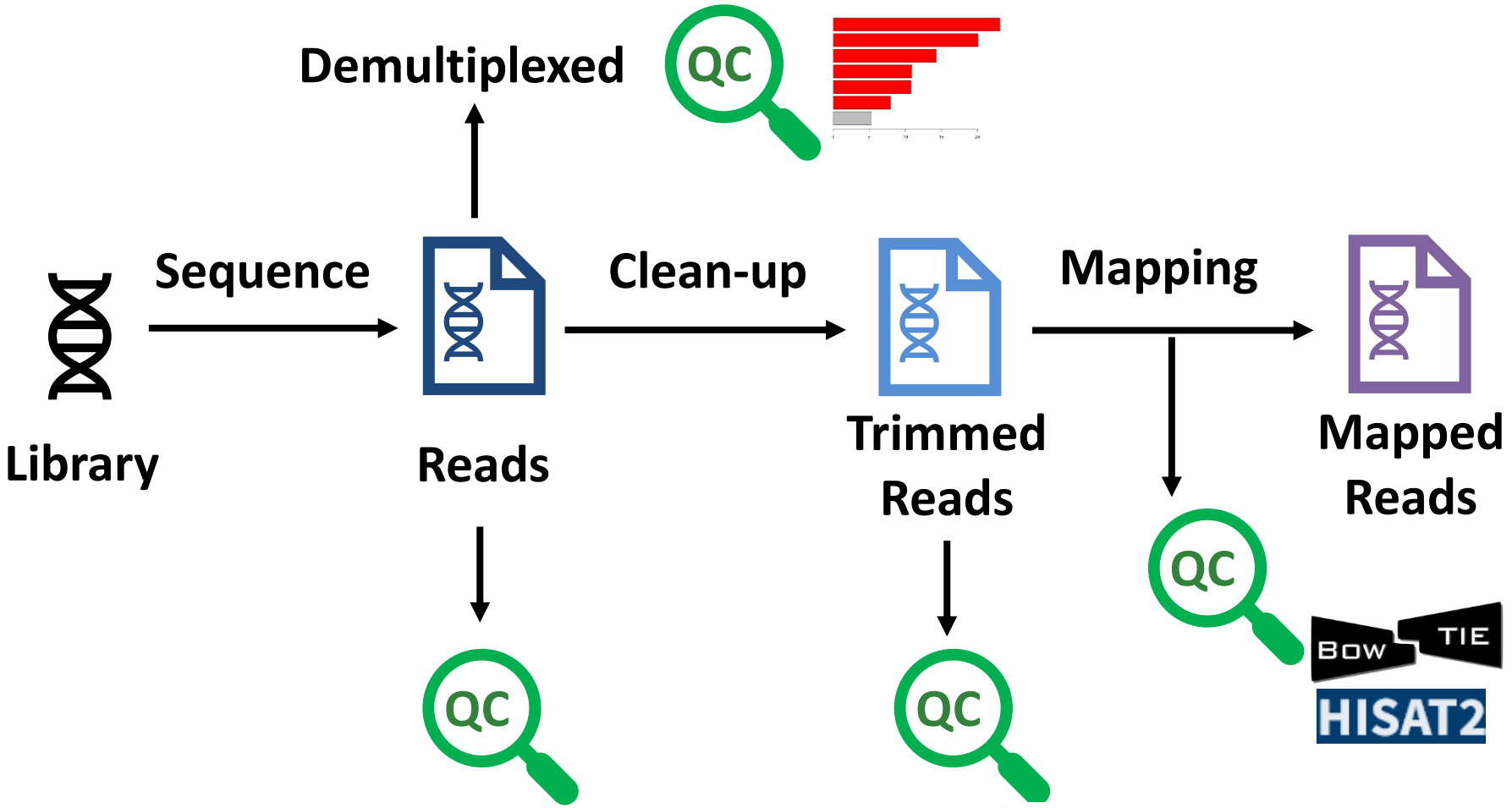
How QC Programmes Fit Into Processing Pipelines

Babraham Bioinformatics

QC metrics can we work with

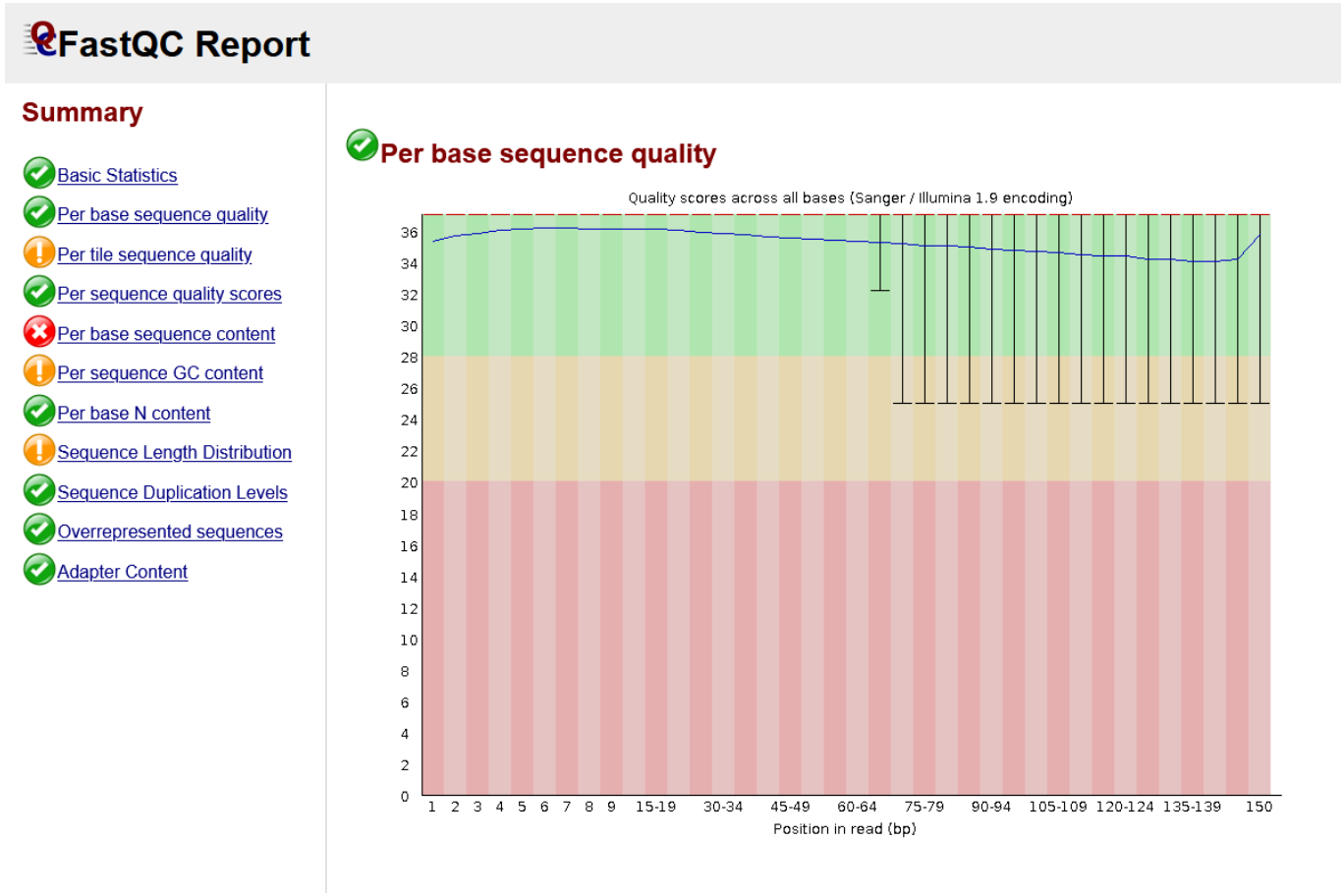
Phred Scores 	**How the Sequencer Performed** • At different cycles • Across different locations • For different reads
Library Composition 	**The Nature of our Sequenced Reads** • Biases • Contaminants • Duplication
Mapping 	**Where our Sequencing Reads Come From** • Species (plural or singular!) • Region (repetitive or unique)

QC Programmes in Data Processing



MultiQC

FastQC




```
fastqc seqfile1 seqfileN
```

```
fastqc *.fastq.gz
```

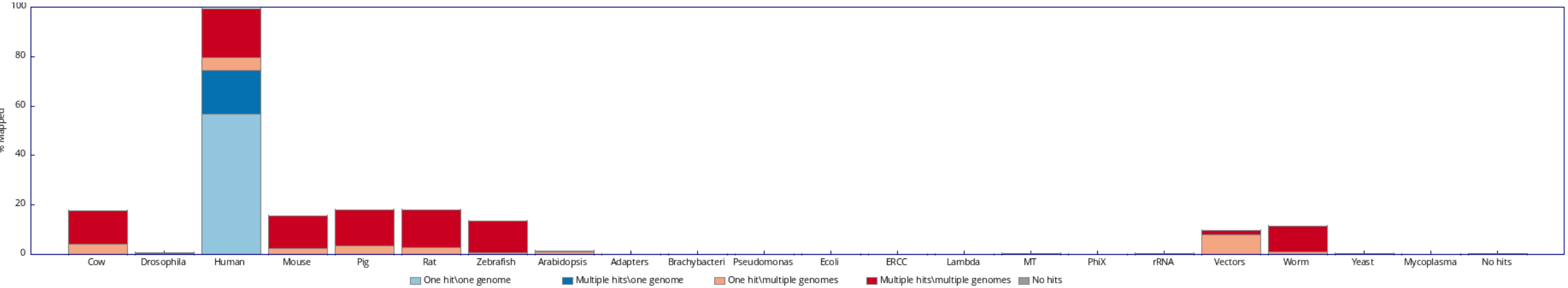
- Reads raw fastq file(s)
- Performs multiple checks
  - Pass/warn/fail
  - Compares to genomic library
- Generates a HTML Report

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

FastQ Screen

```
fastq_screen seqfile1 seqfileN
```

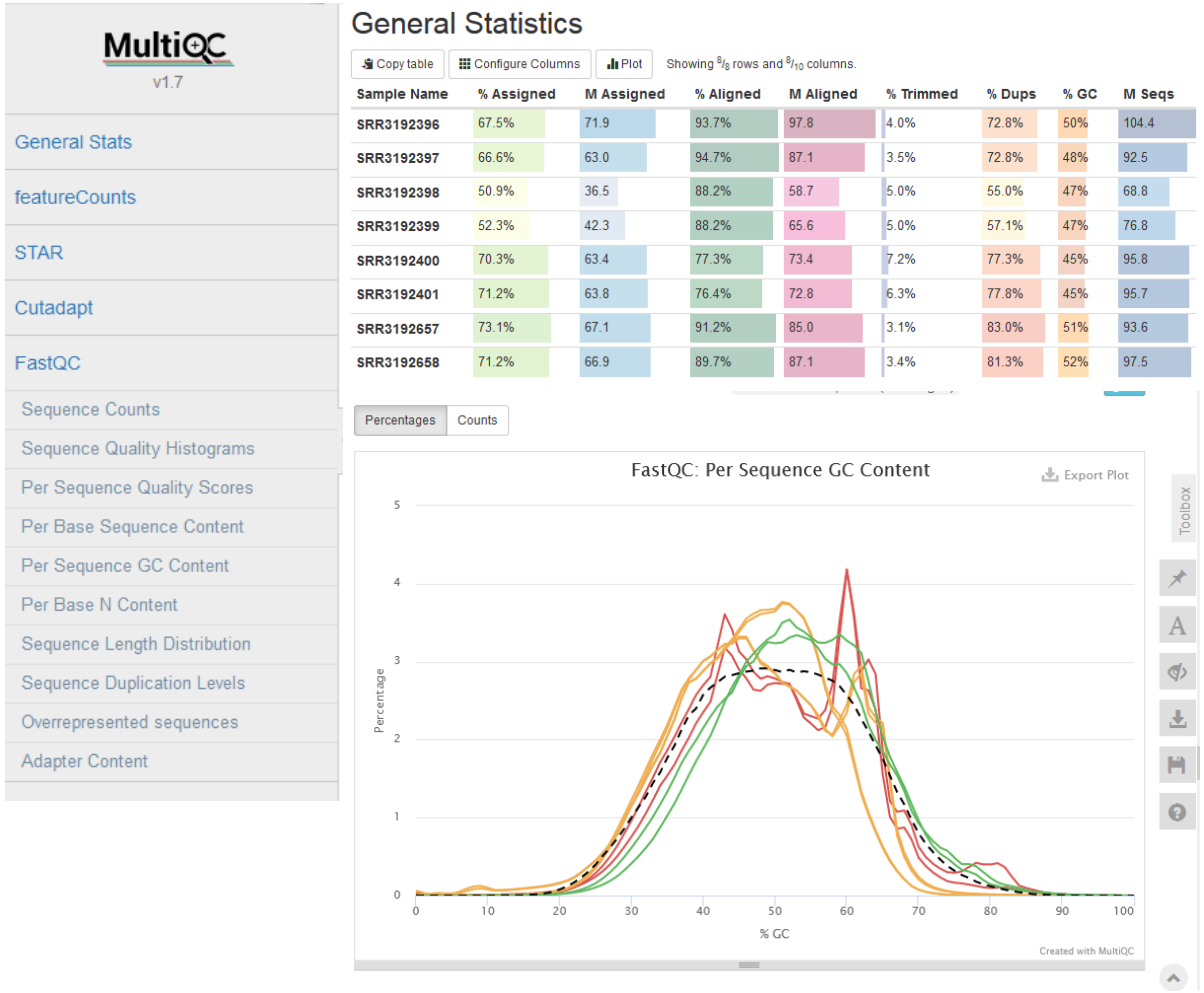
```
fastq_screen *.fastq.gz
```



- Reads fastq file(s)
- Maps against a range of species / contaminants
- Identifies unexpected sequences in your library
- Generates a HTML Report

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

MultiQC




```
multiqc directory_with_reports
```

```
multiqc .
```

- Reads in all QC files in a directory
- Aggregates QC information from multiple samples
- Large number of programs supported
- Generates a combined HTML report

https://multiqc.info/

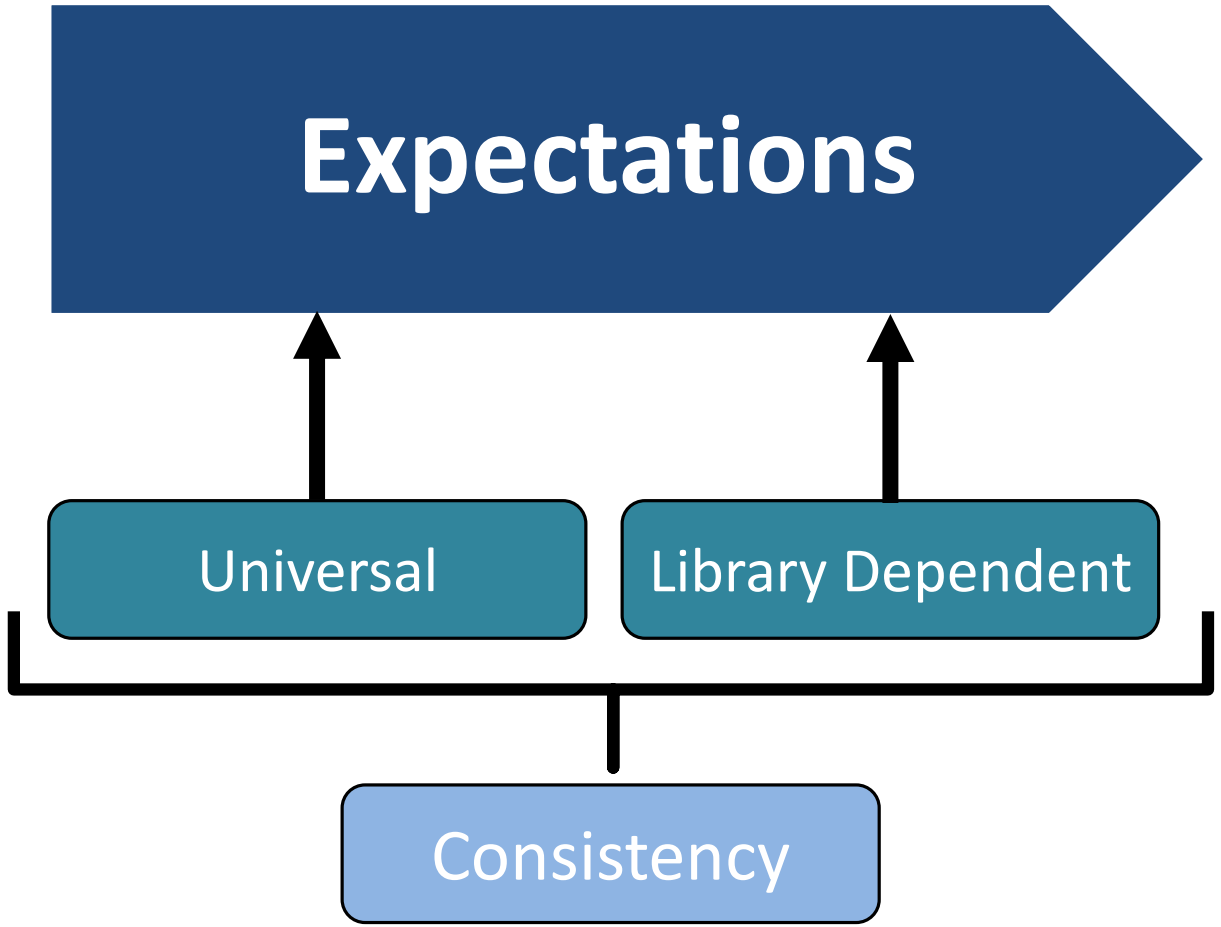
What can QC tell us?

Babraham Bioinformatics

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:

Assessing Universal Metrics

Babraham Bioinformatics

Context is Key for QC

QC metrics that behave the same for all libraries

What Does it Mean?

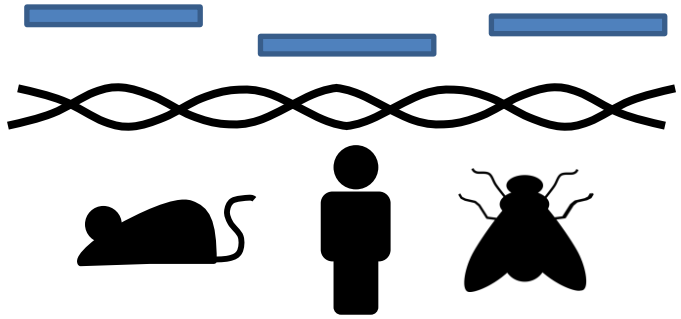
Universal QC Metrics

- Demultiplexing

- Base Call Quality

- Adapter Content

- Mapping Quality



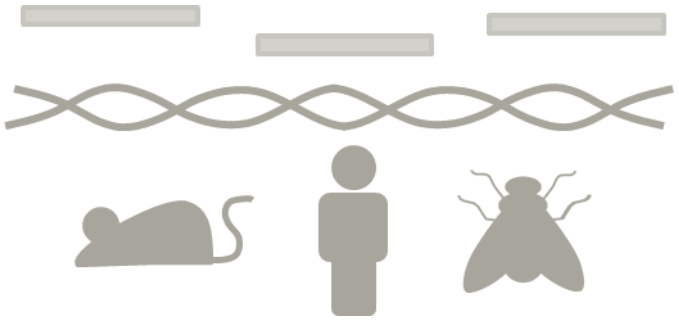
Universal QC Metrics

- Demultiplexing

- Base Call Quality

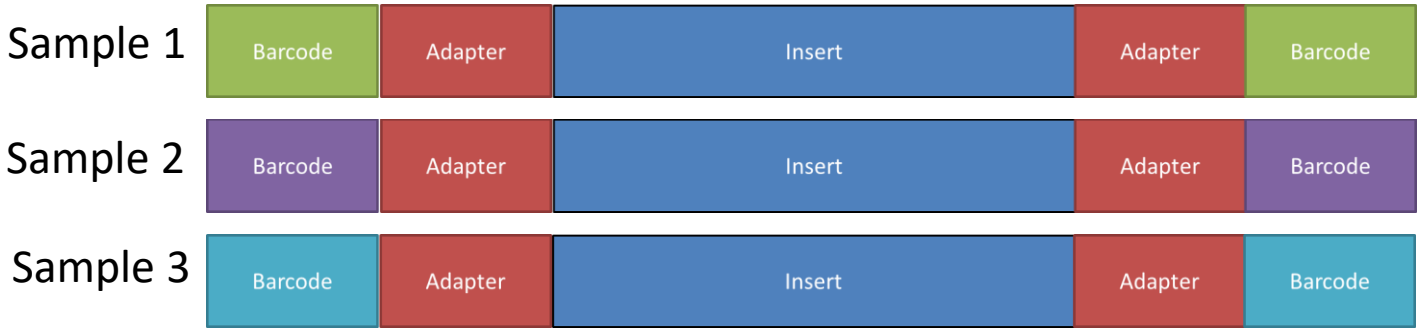
- Adapter Content

- Mapping Quality



Demultiplexing: Expectations

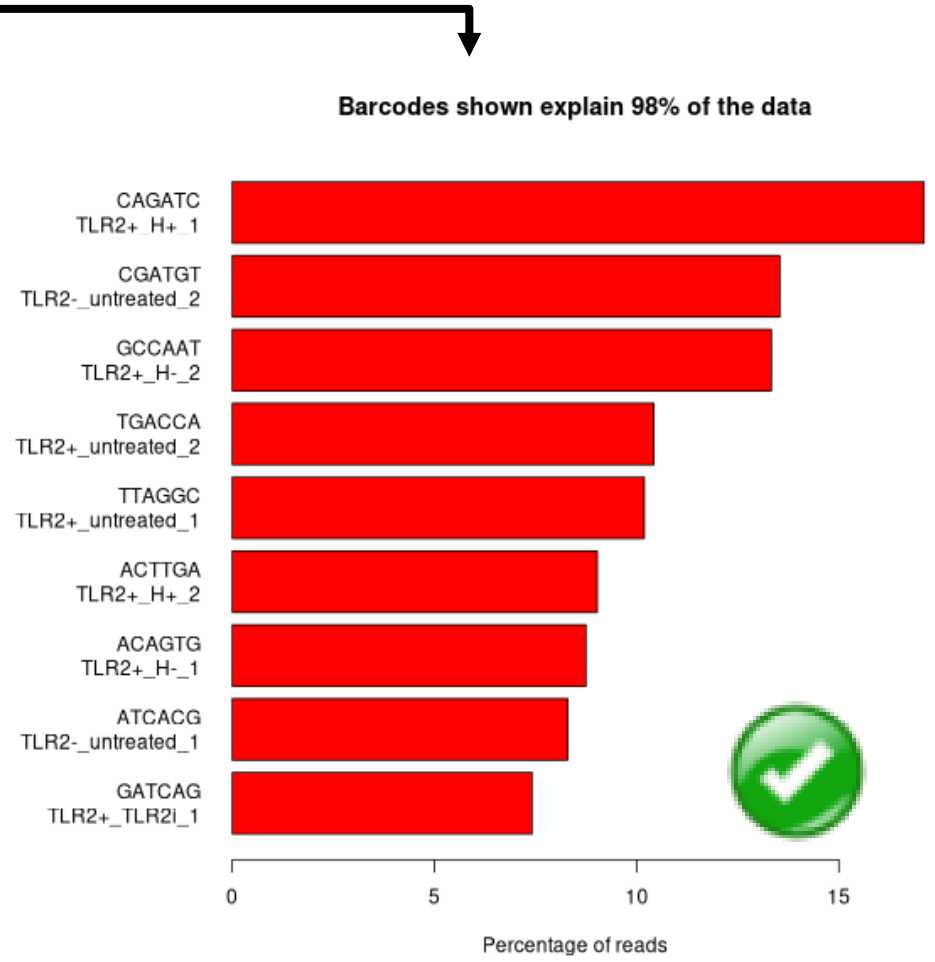
Only the barcodes we assigned to samples should be present—**no others**



ect…

What barcode sequences have we found?

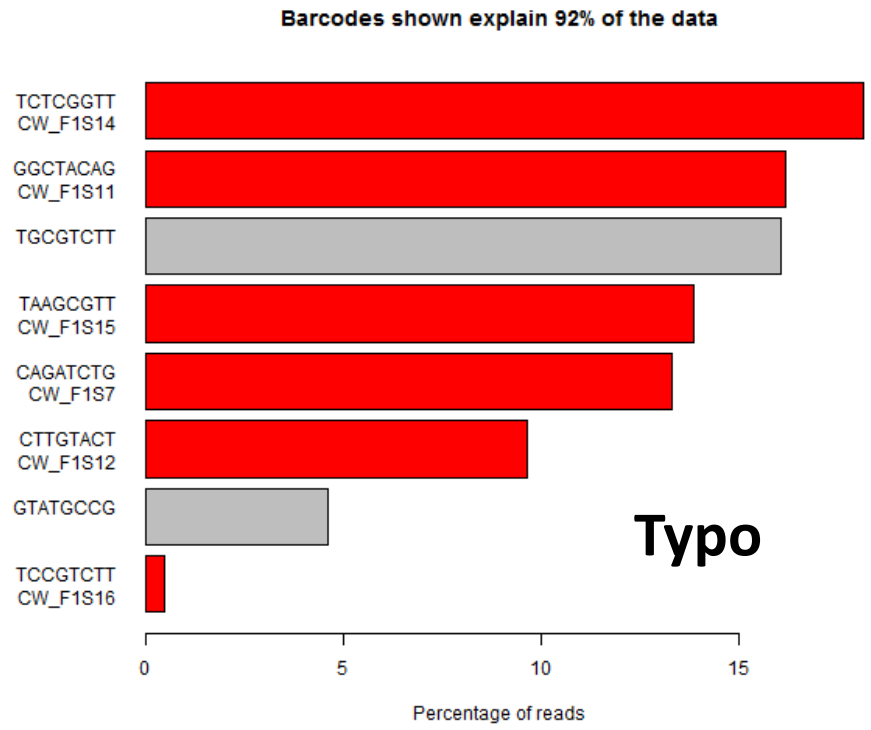
Expected Barcode Unknown Sequence



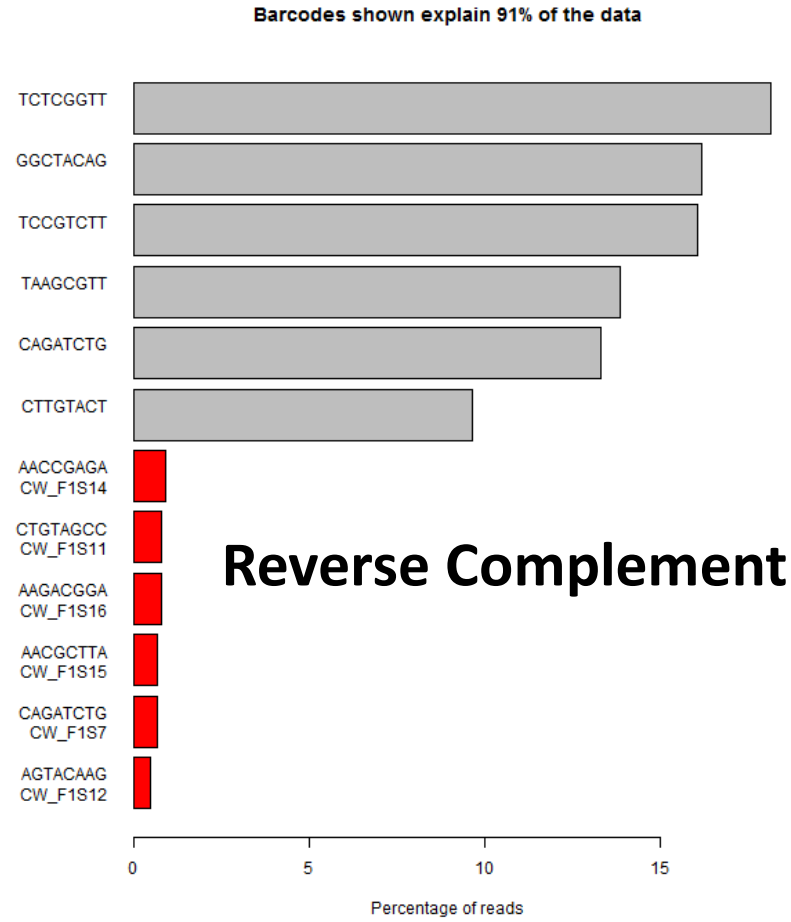
What could unknown barcode sequences mean?

Demultiplexing: Unknown Barcode Sequences

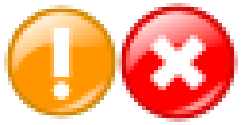
Expected Barcode Unknown Sequence



Typo



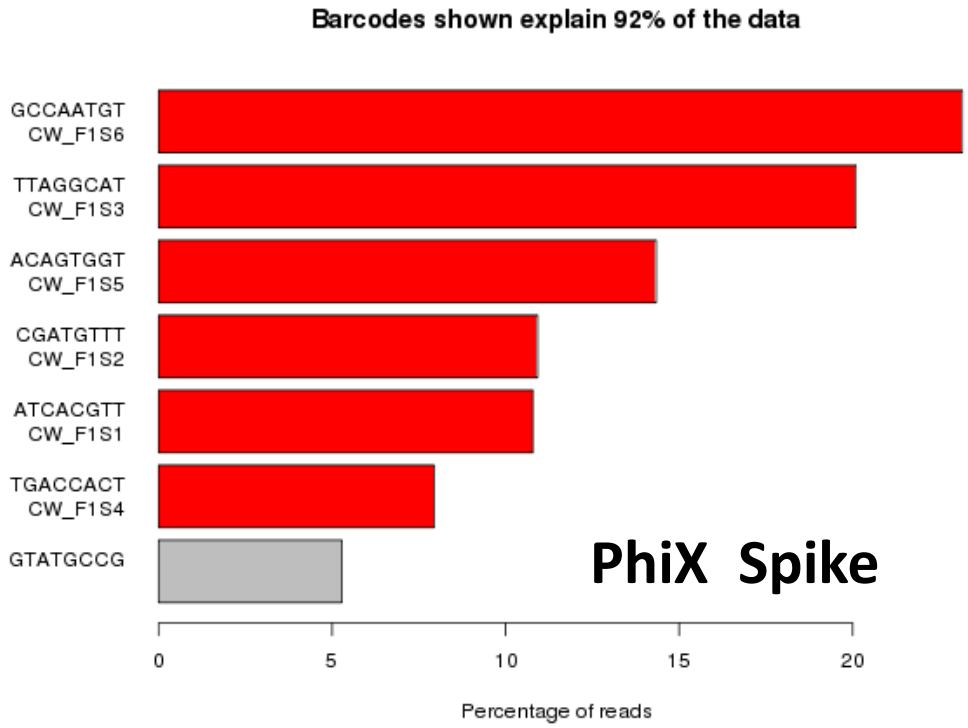
Reverse Complement



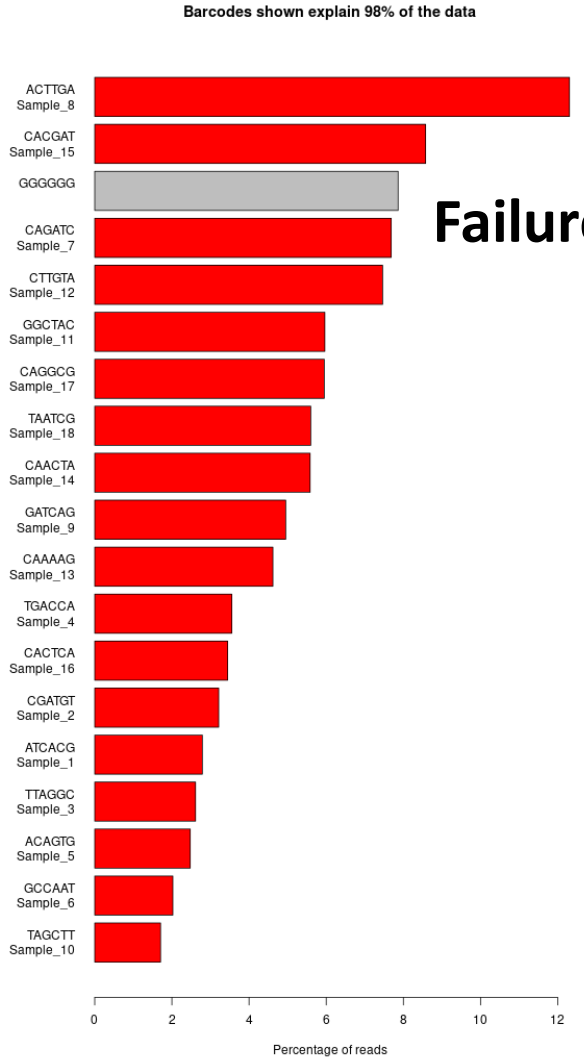
Human Error is a really common source of barcode issues

Demultiplexing: Unknown Barcode Sequences

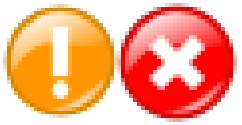
Expected Barcode Unknown Sequence



PhiX Spike



Failure in Priming site



Sometimes more technical in nature…

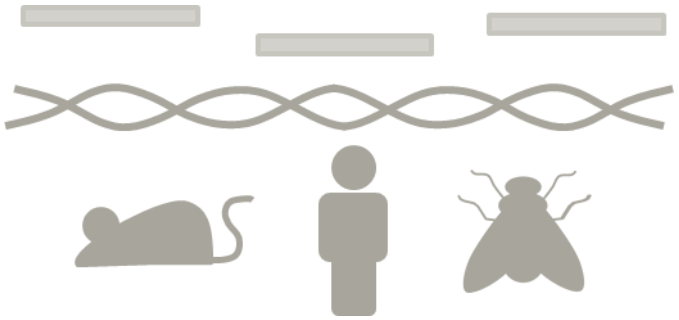
Universal QC Metrics

- Demultiplexing

- Base Call Quality

- Adapter Content

- Mapping Quality

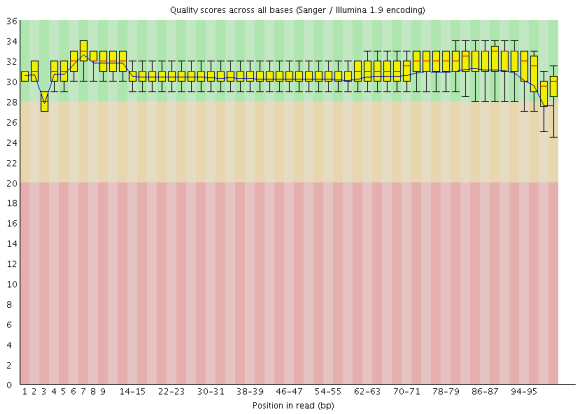


Base Call Quality: Expectations

Illumina Sequencers are technically reliable, so **we expect confident calls**

Phred scores

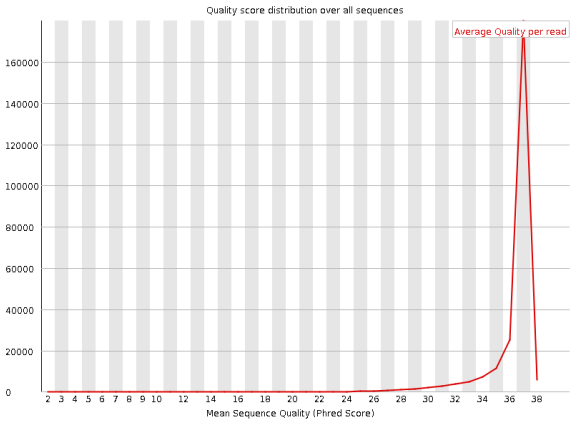
Per base/cycle



FastQC Report

Per base sequence quality

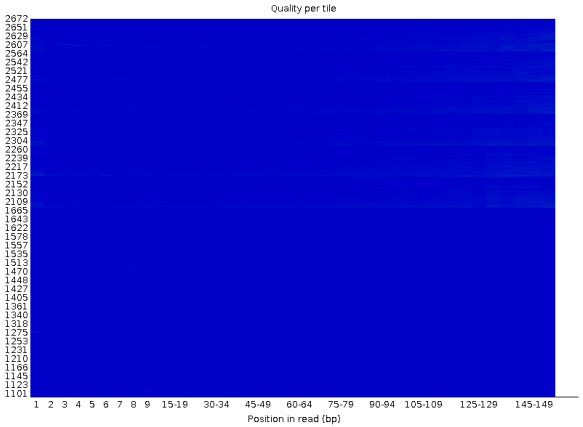
Per Read



FastQC Report

Per sequence quality scores

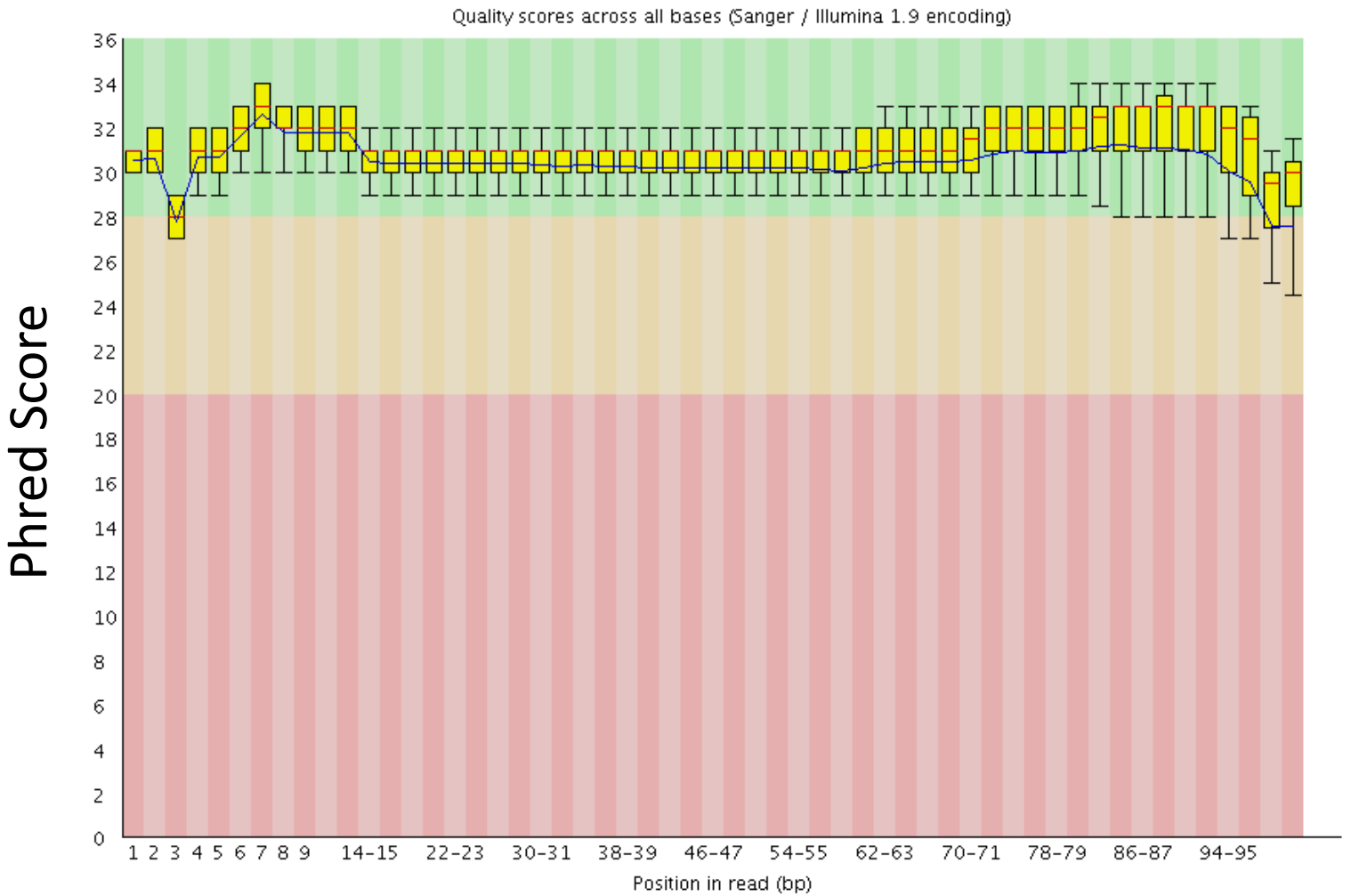
Location



FastQC Report

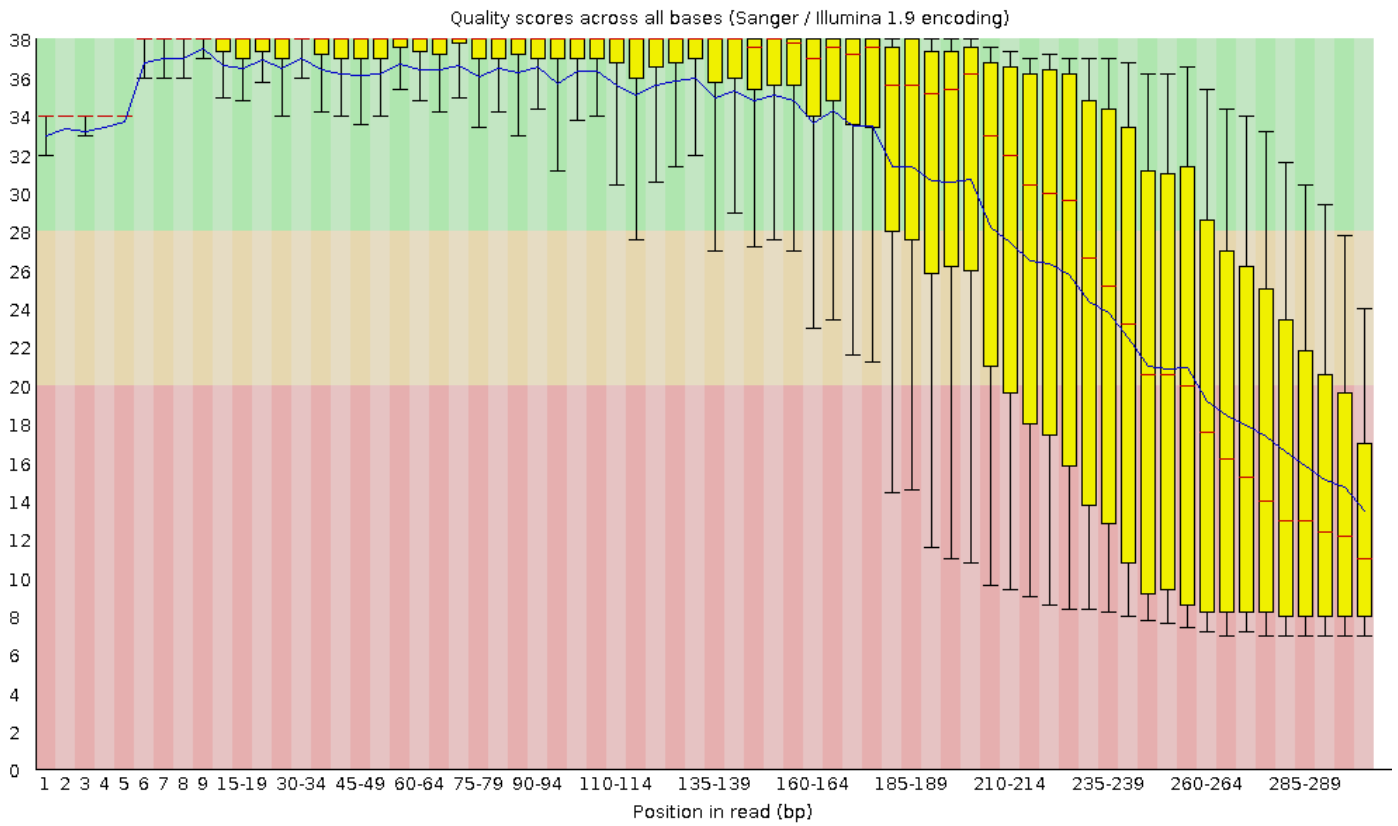
Per tile sequence quality

Per Base Sequence Quality Plot

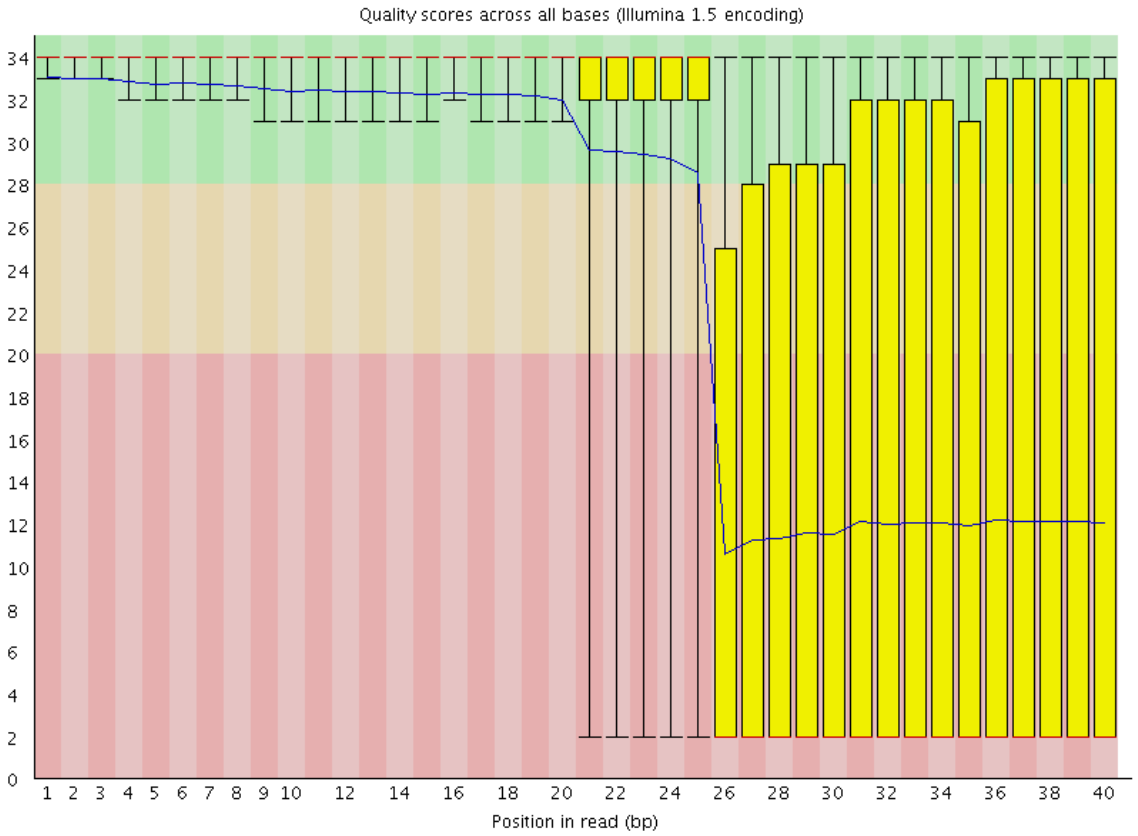


Read Position / Cycles of Chemistry

Base Call Qualities – Problems At Given Cycles

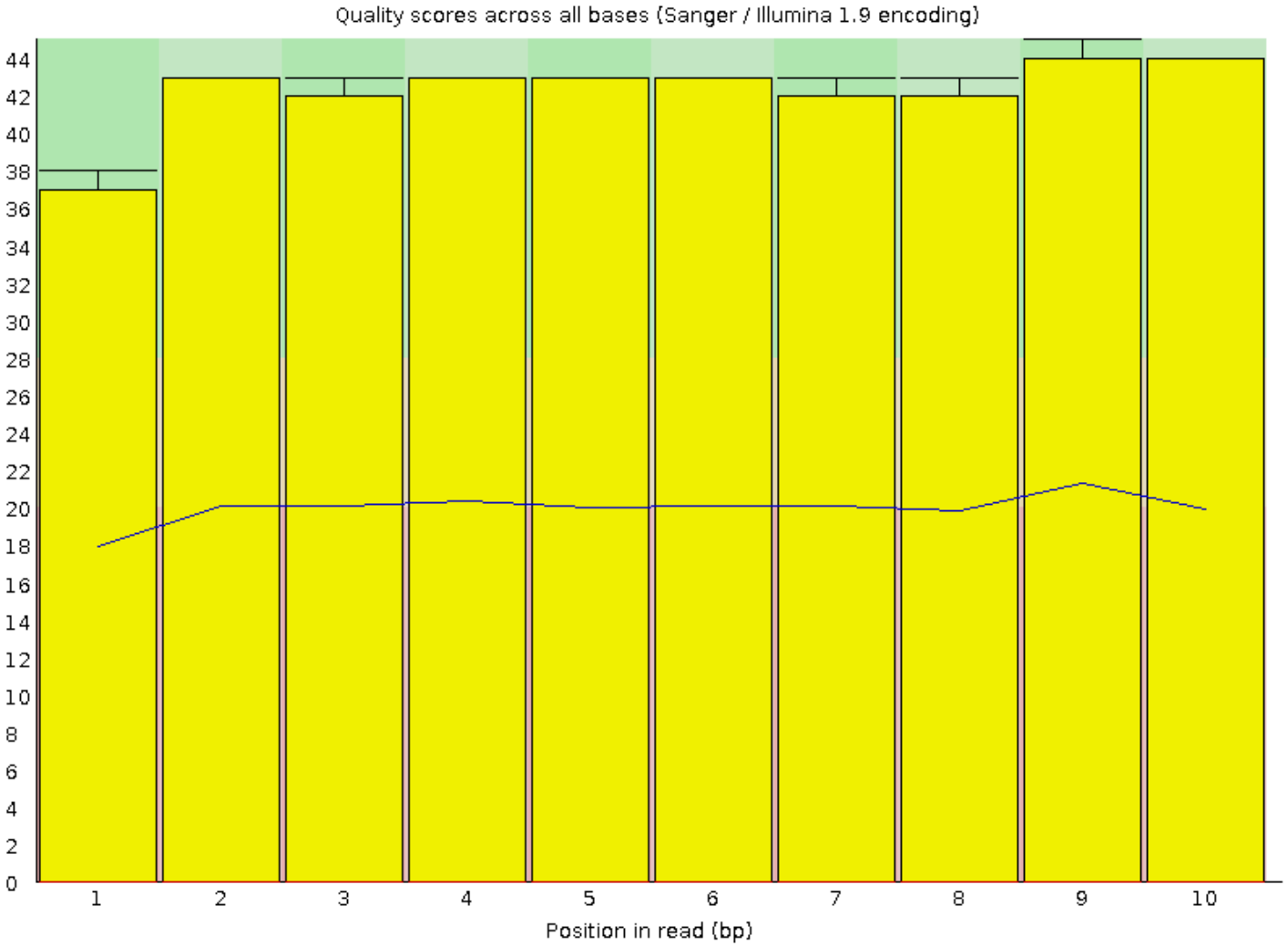


Clusters get out of sync over long read length



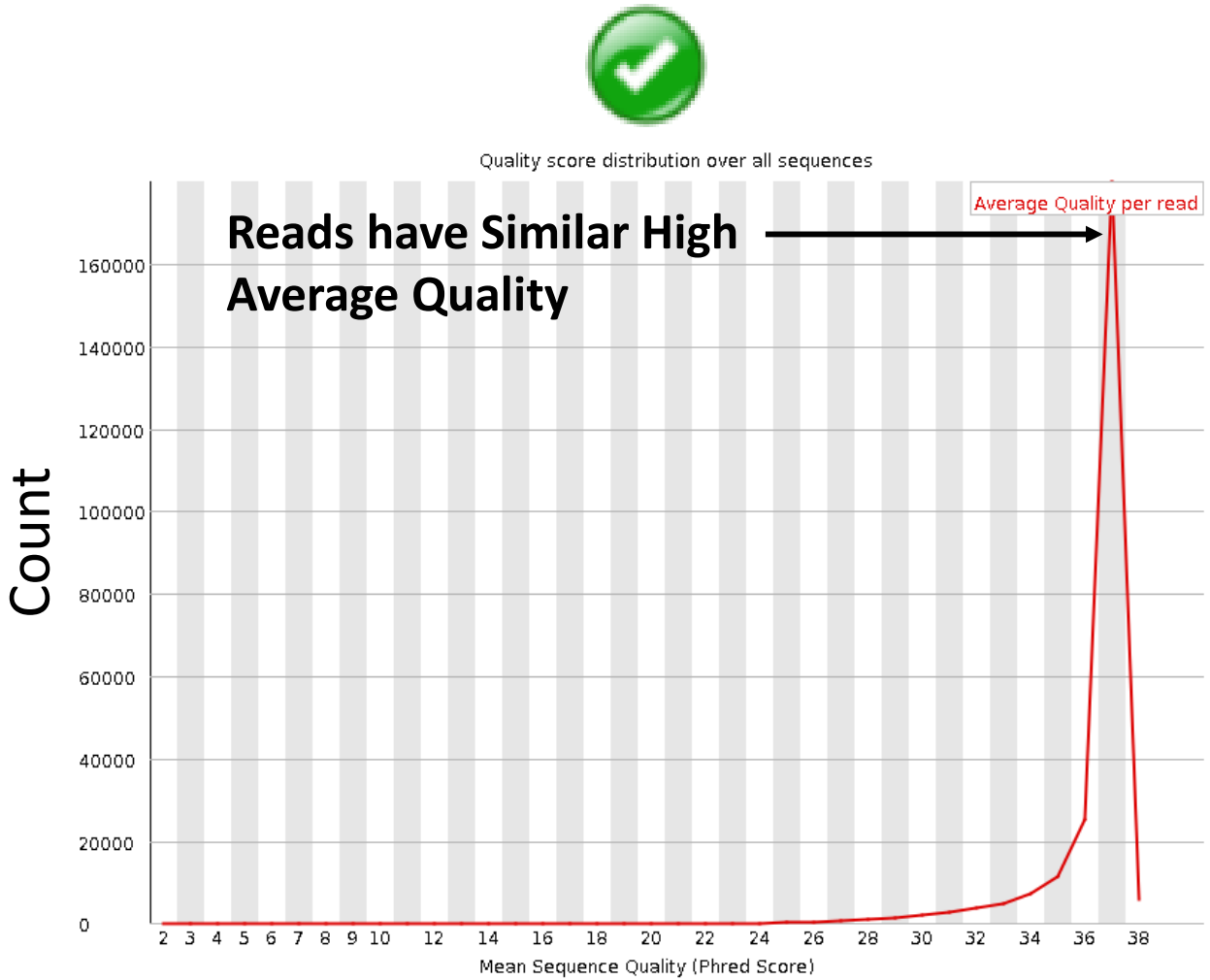
Technical problem with the sequencer

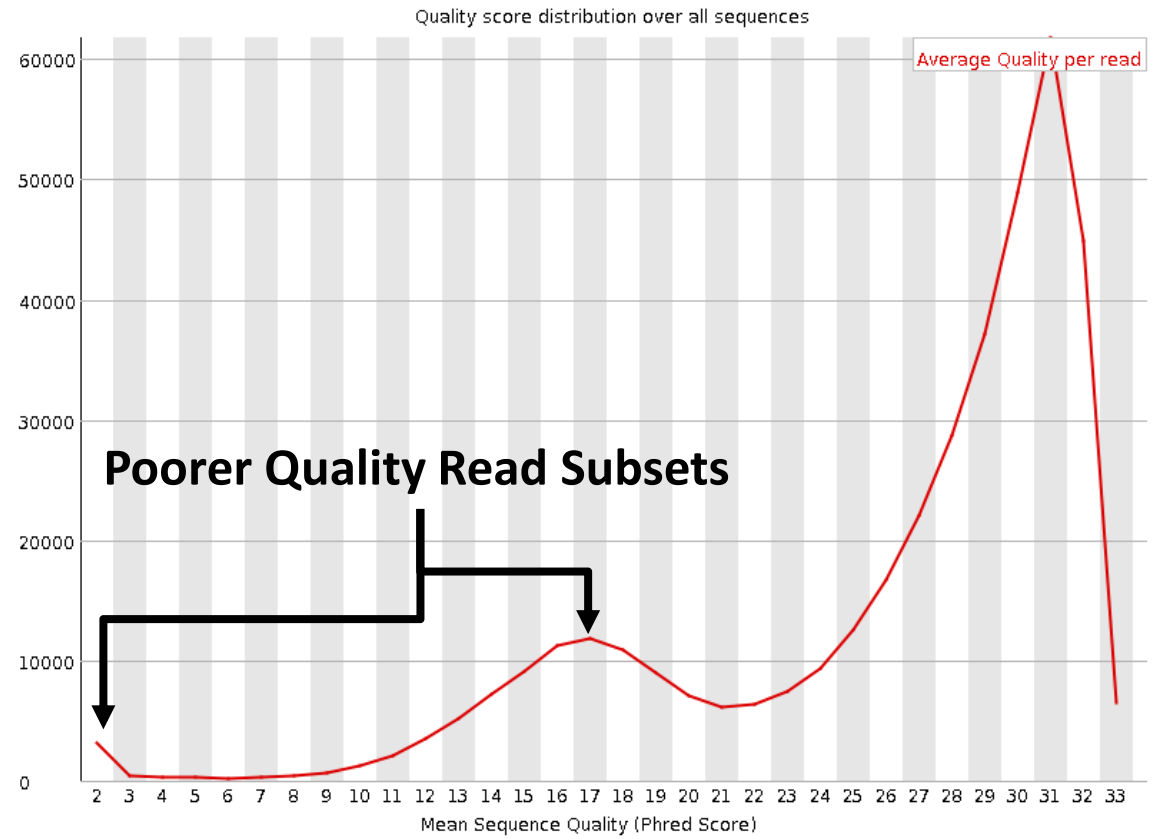
Diagnosing Less Clear Cut Base Call Problems



- Not everything is bad
- Why are some calls bad and not others?
- What can we learn
  - Salvage this run
  - Fix future runs

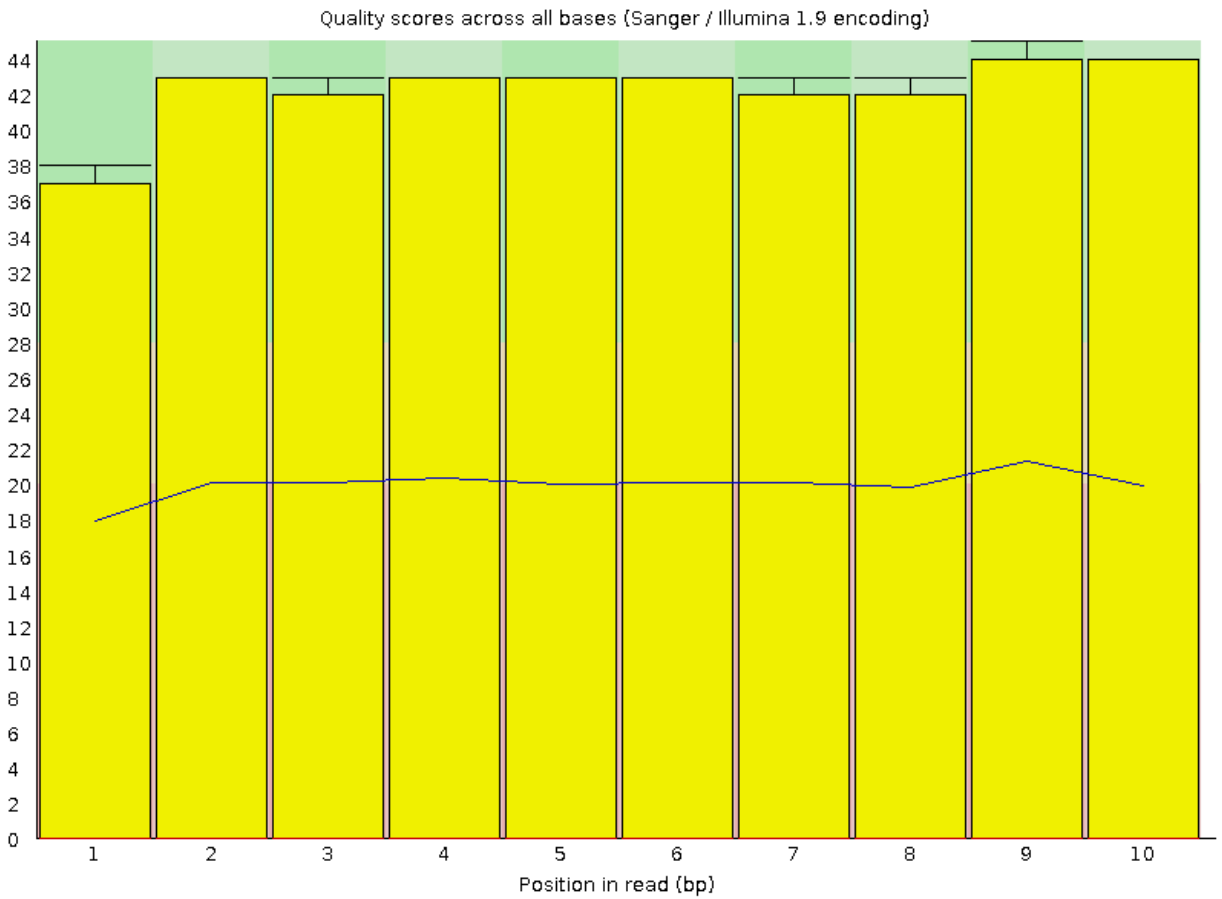
Per Sequence Quality Score Plot

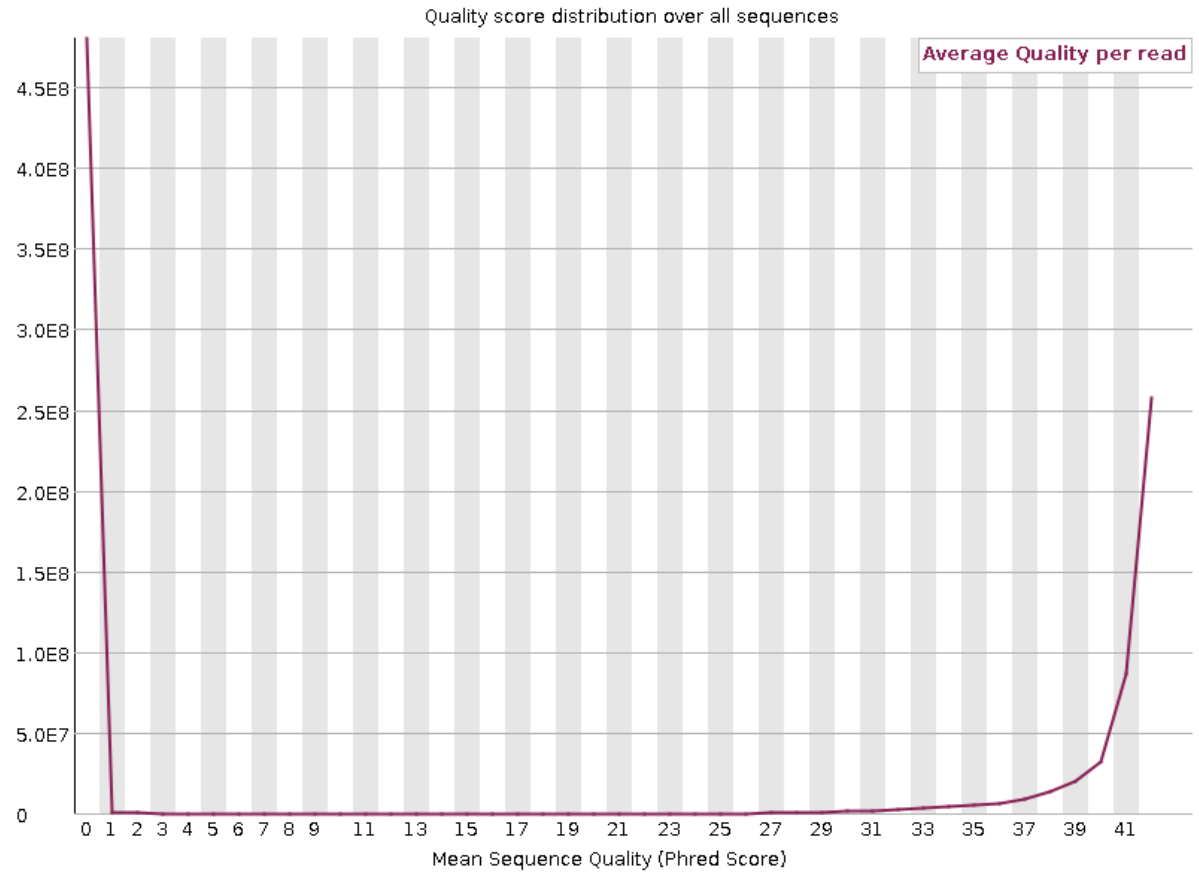




Mean Sequence Phred Score

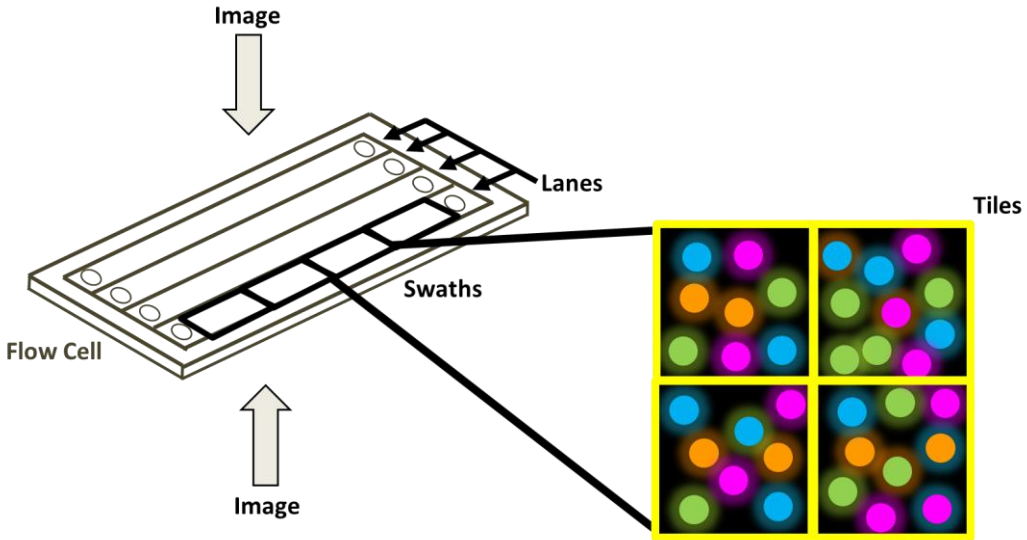
Diagnosing Less Clear Cut Base Call Problems



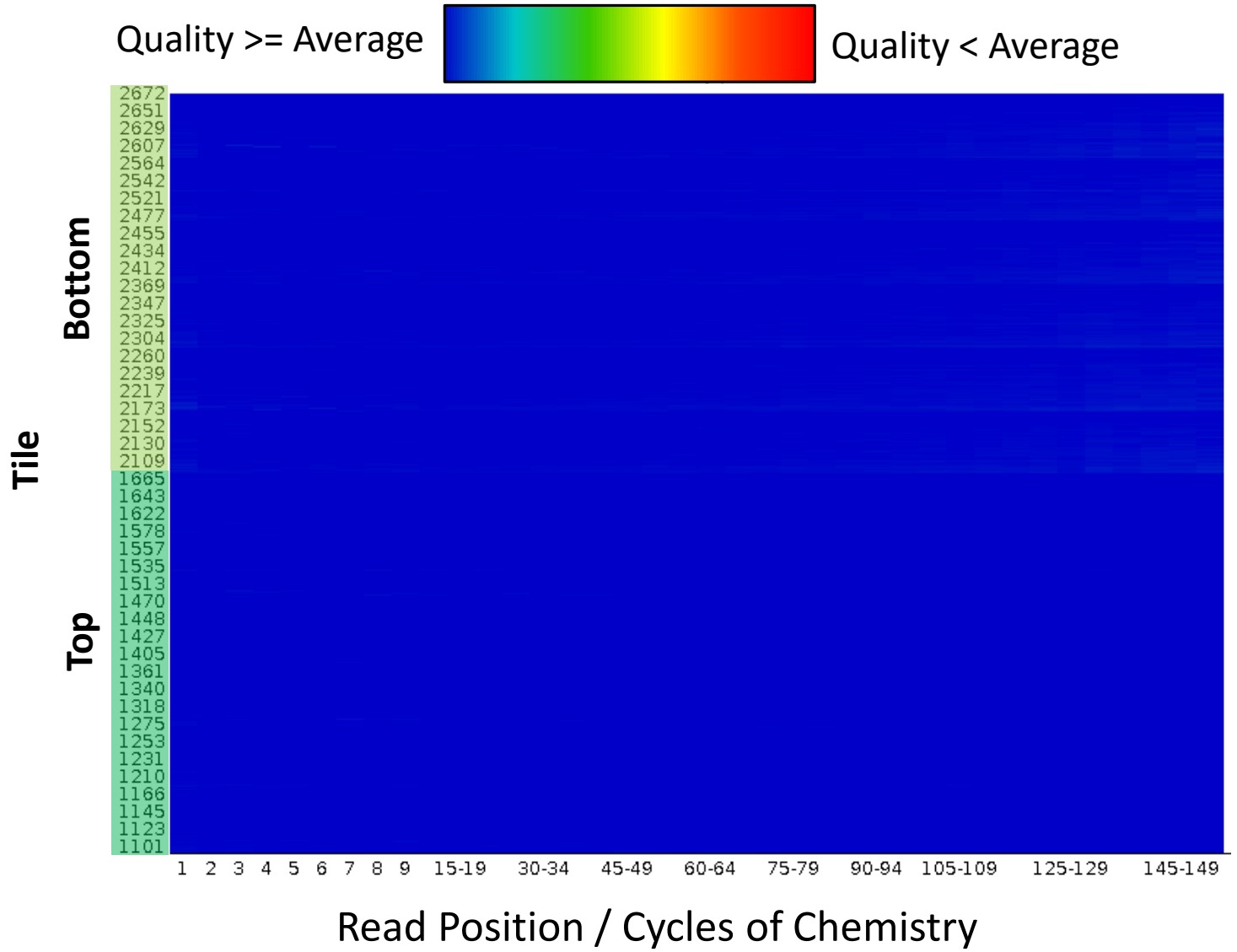


Lot of Reads with really low Scores

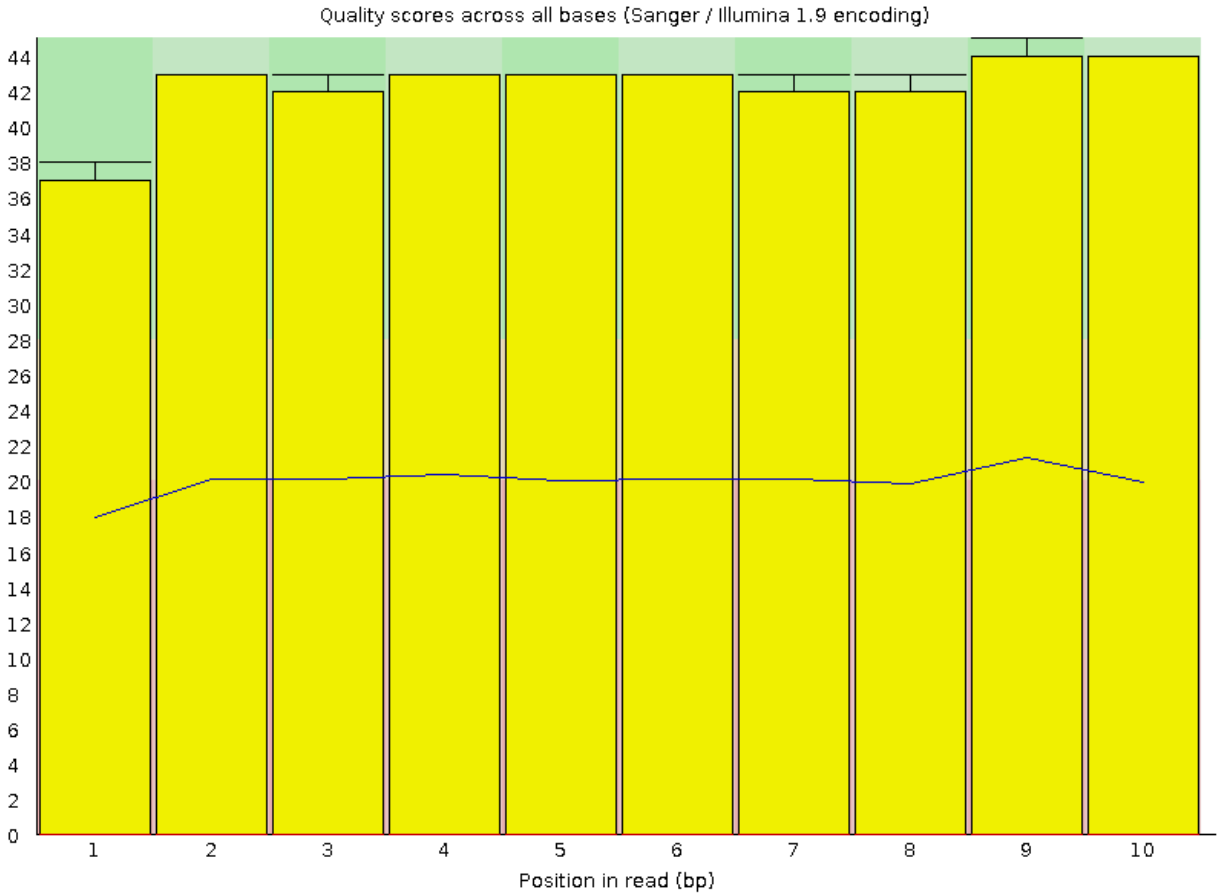
Positional Quality

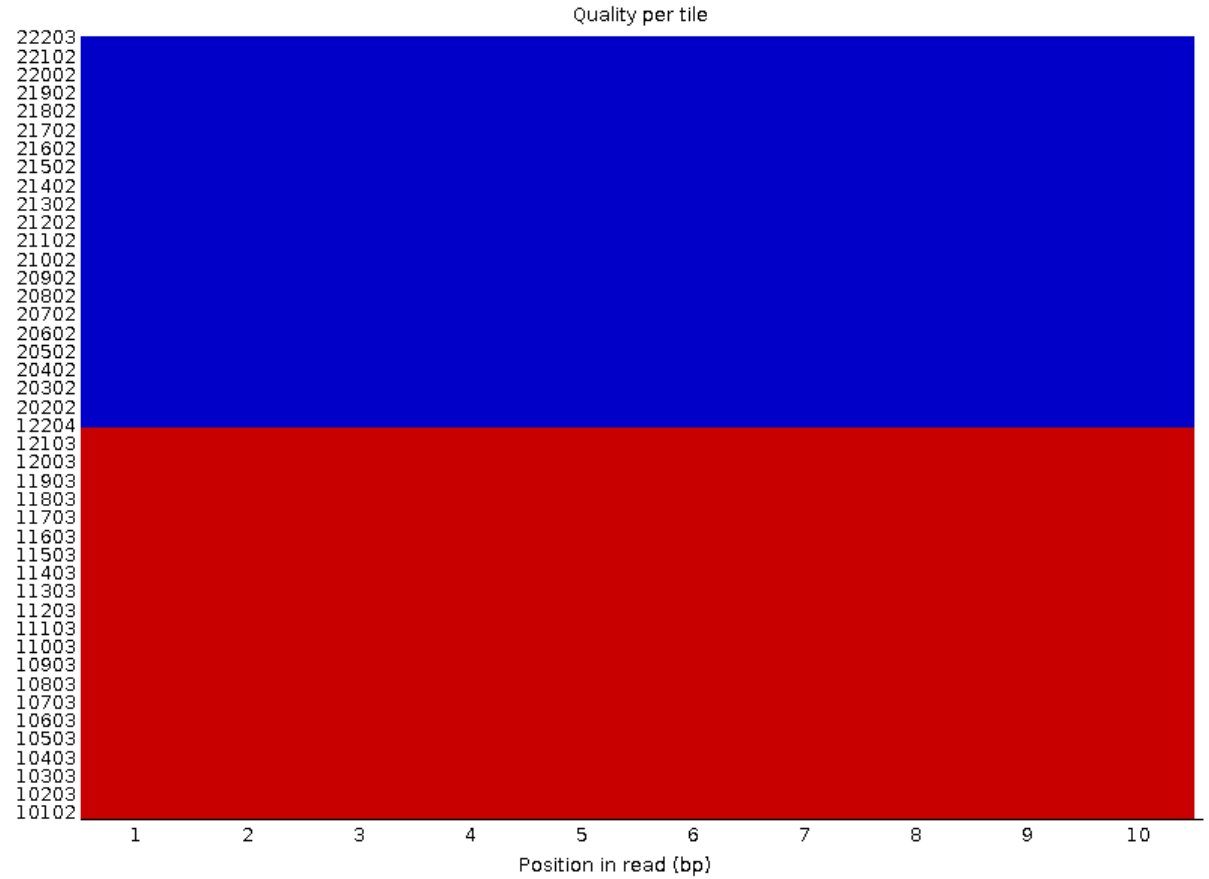


A good tile plot is a blue square



Diagnosing Less Clear Cut Base Call Problems



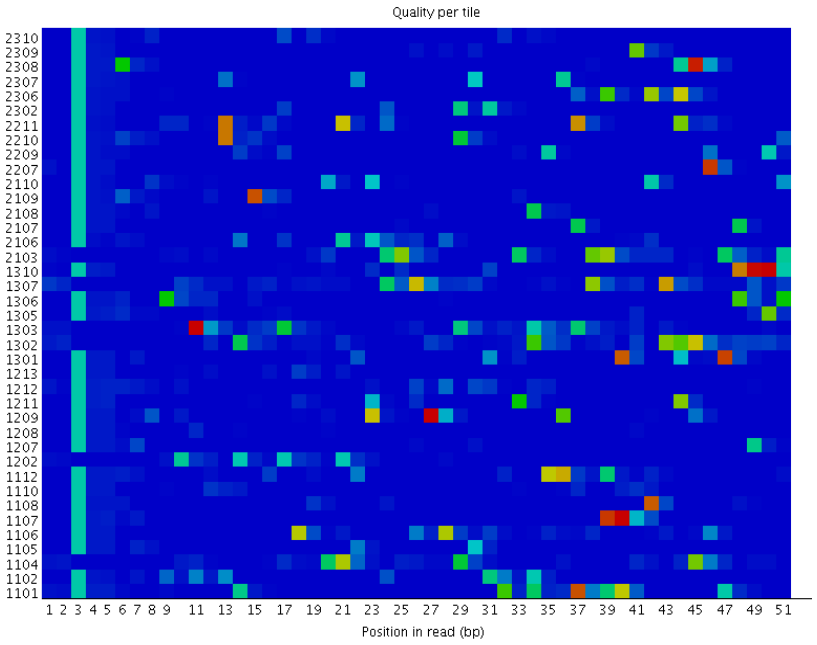


Focusing Fail!

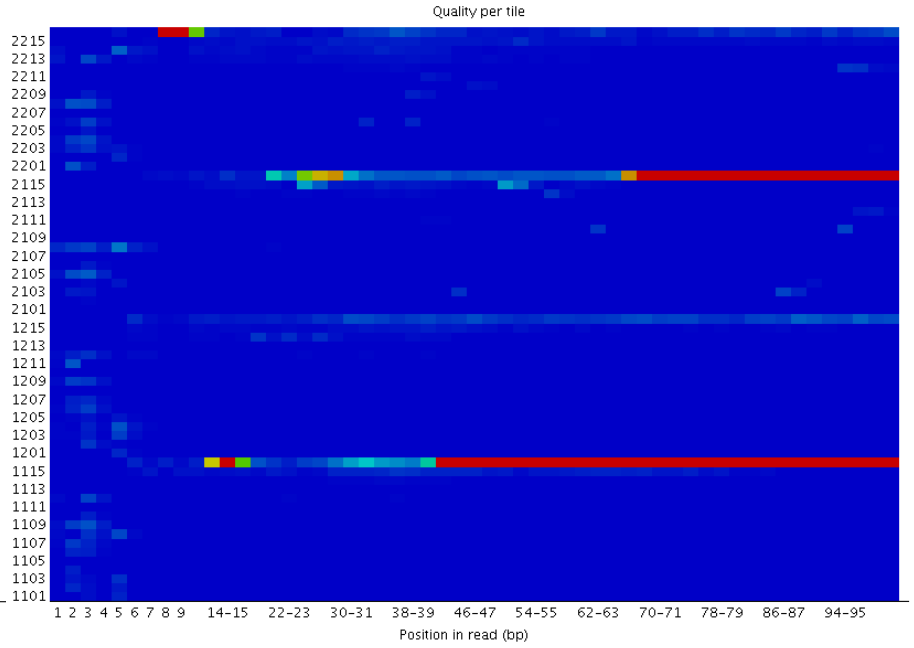
More Examples of Positional Fails

Position Specific Patterning

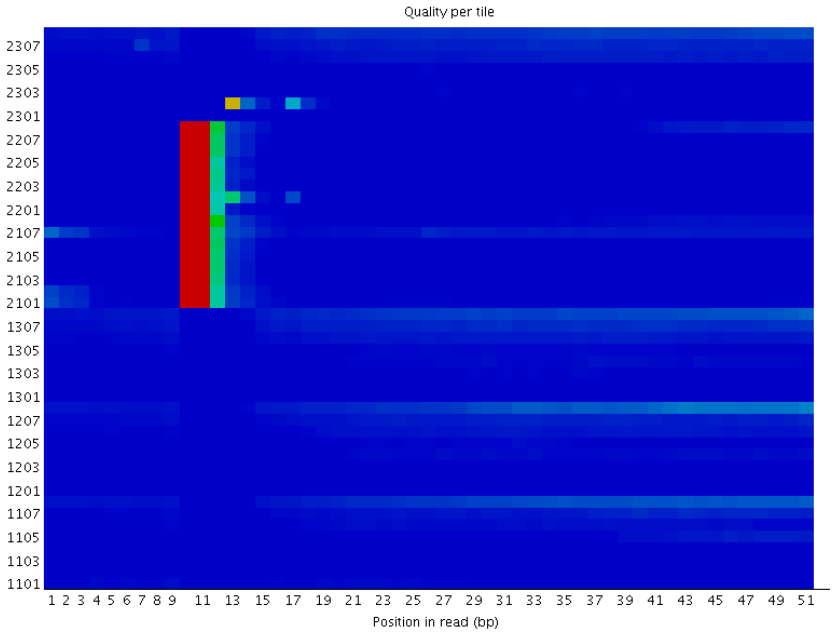
Random Pattern



Permanent



Transient



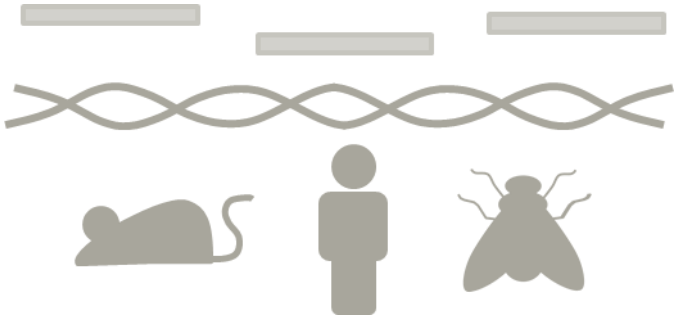
Overloading of Flow Cell

Obstruction

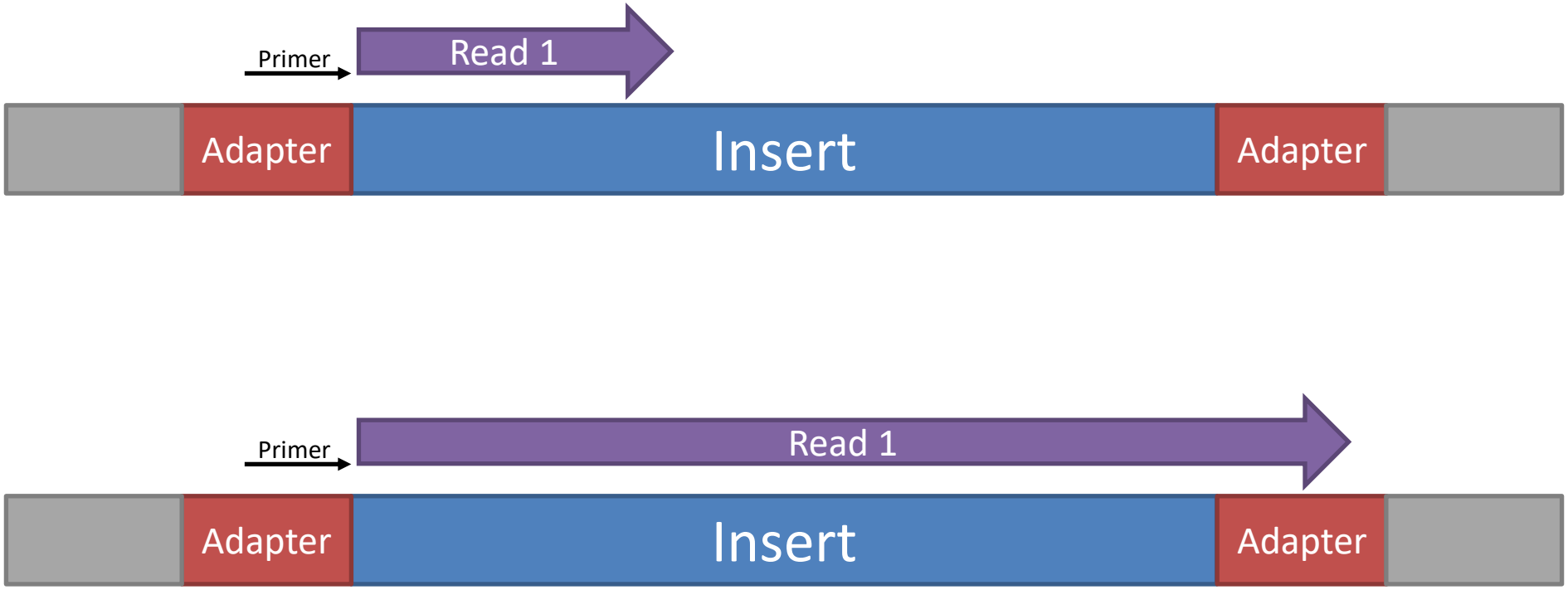
Bubble

Universal QC Metrics

- Demultiplexing
- Base Call Quality
- Adapter Content
- Mapping Quality



Adapters Content: Expectations

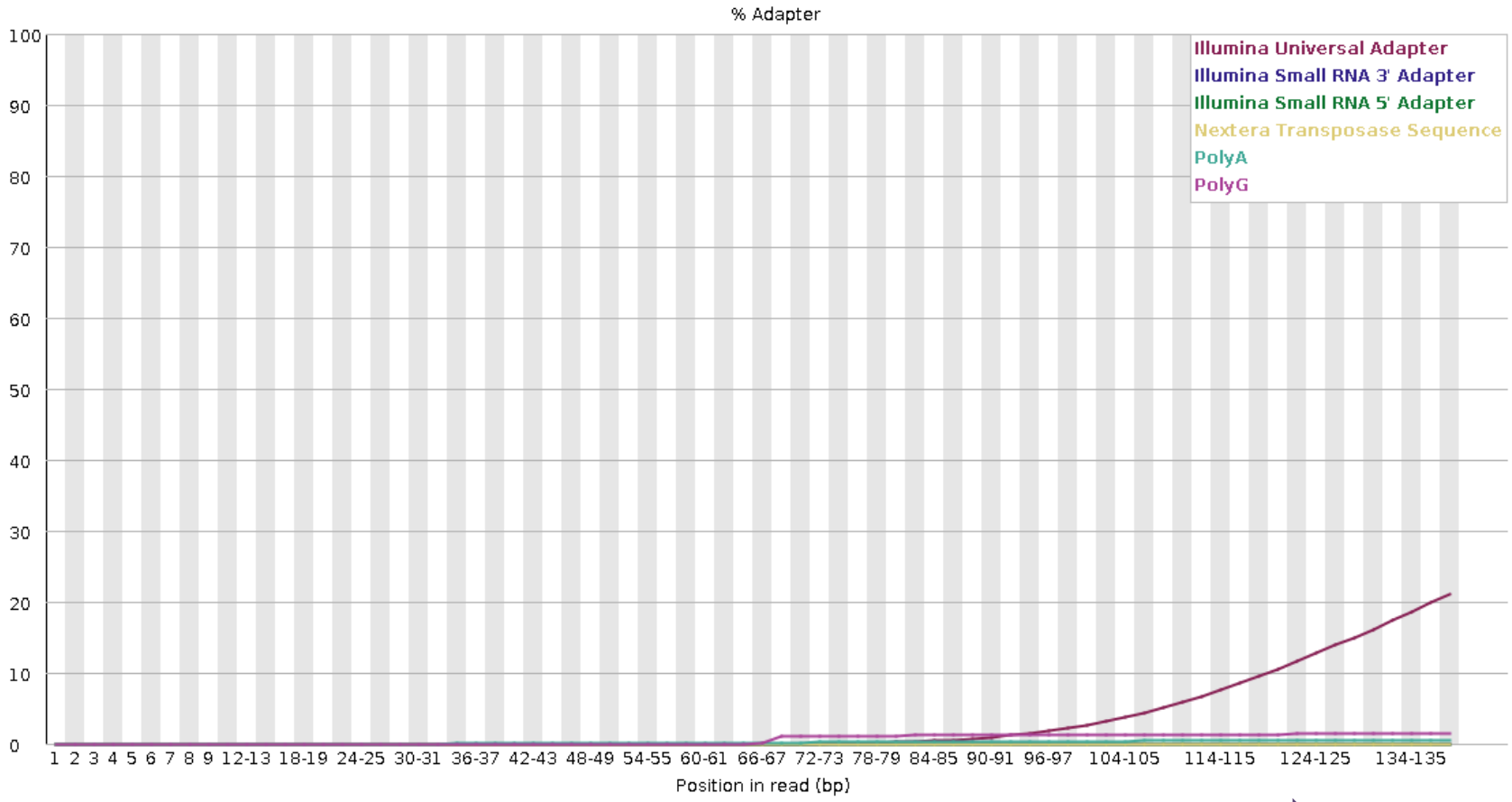


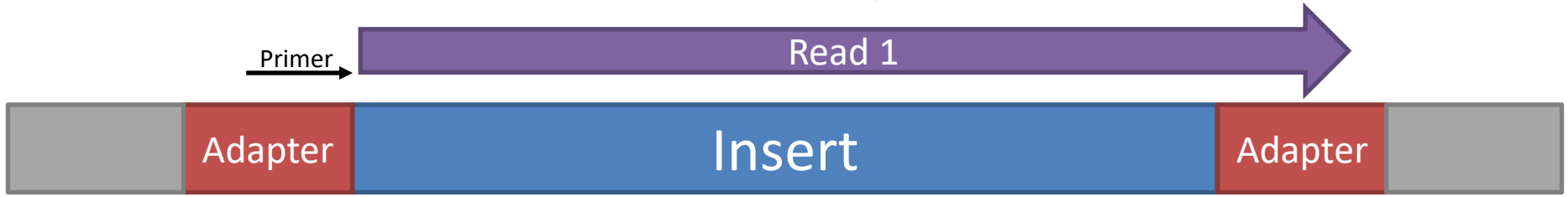
Due to variable insert and read length, **we may sequence adapter at the end of our reads**

FastQC Report

Adapter Content

Measuring Read-though Adapters

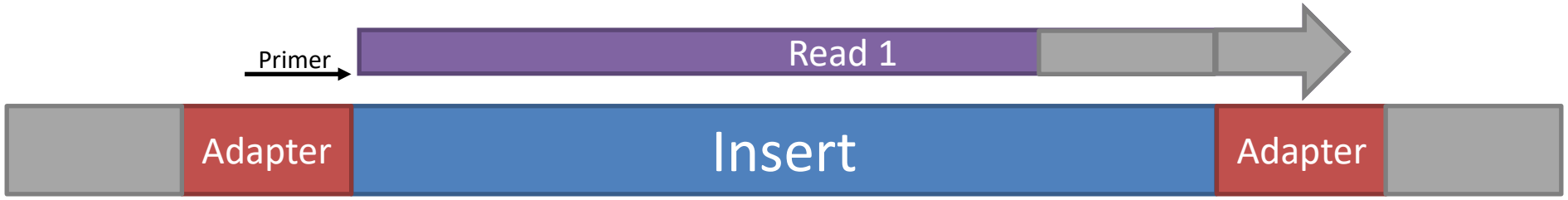




Clean-Up Options (Adapters & Poor Calls)

Trimming 3’ end:

- Remove adapter read through
- Remove poor quality bases



Remove Specific Reads

- Average Quality
- Location on Flow Cell

Live with it:

- Sometimes it’s good enough e.g. mapping

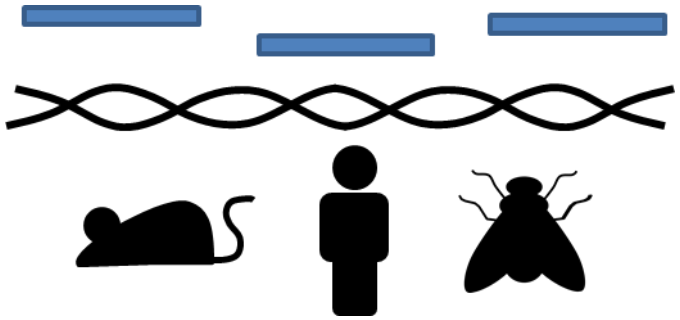
Universal QC Metrics

- Demultiplexing

- Base Call Quality

- Adapter Content

- Mapping Quality

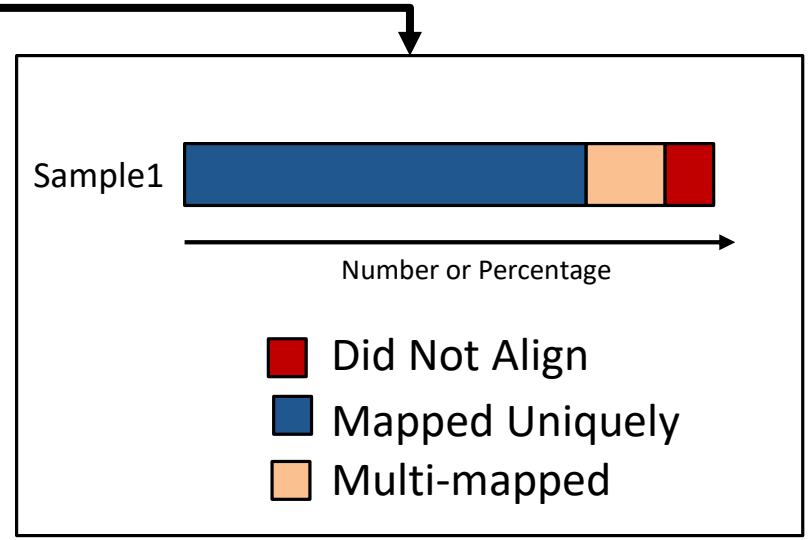


Mapping: Expectations

We Expect Reads to Align to the Species Sequenced

HISAT2

```
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
```

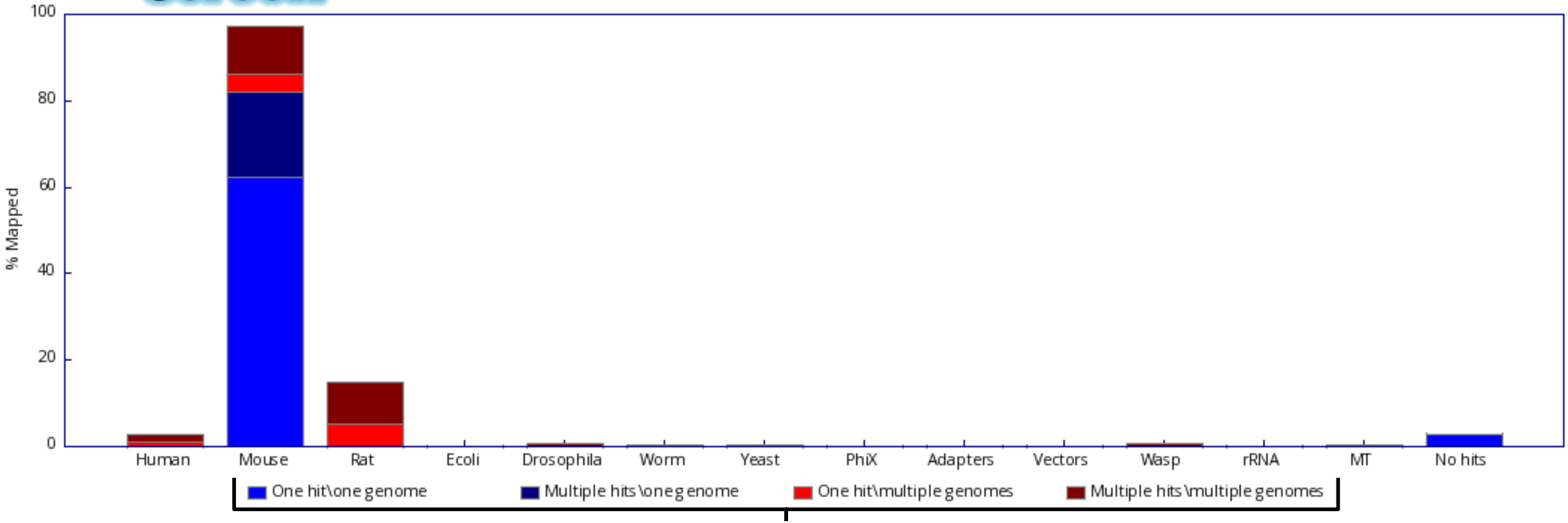


Alignment Will Vary!

But if many reads do not align as expected… Where are they from?

Checking Mapping issues: Library Screening

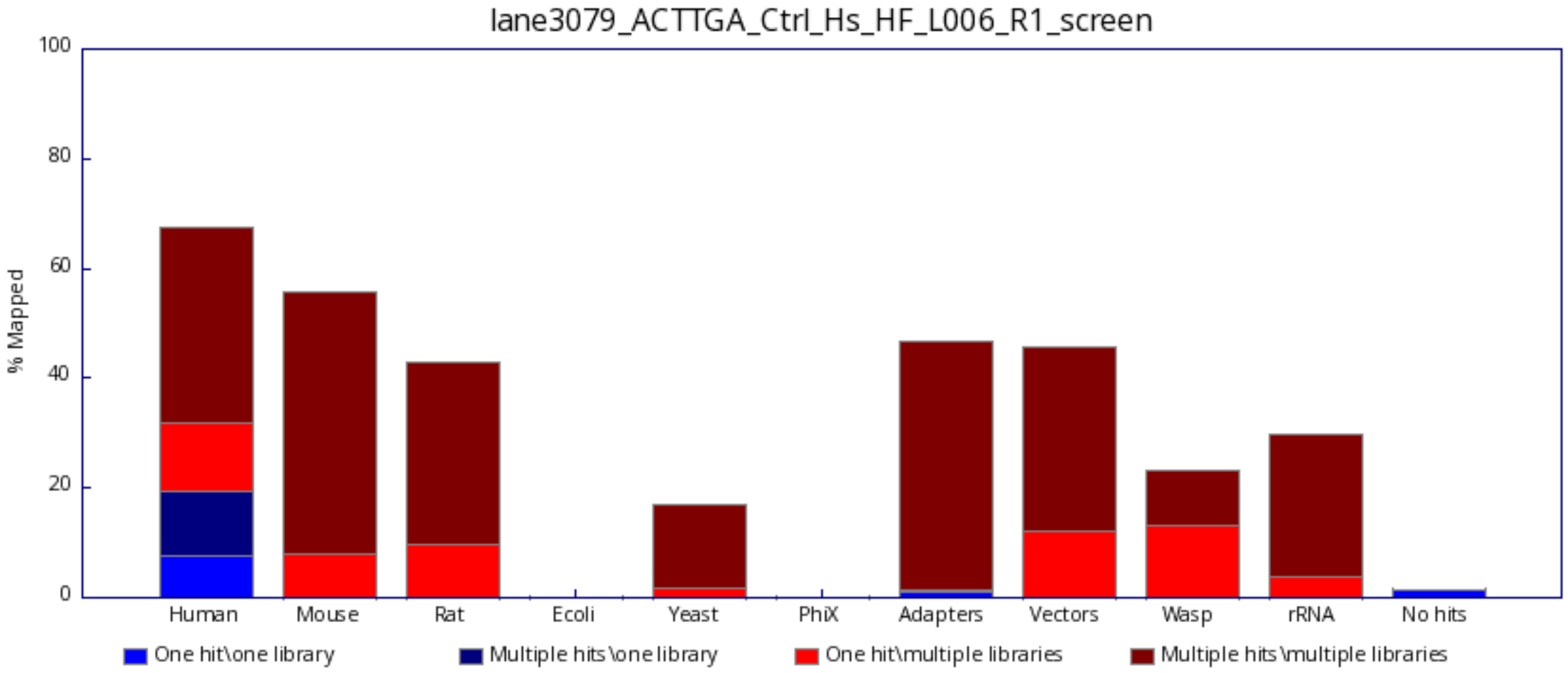
Map reads against a range of reference genomes



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

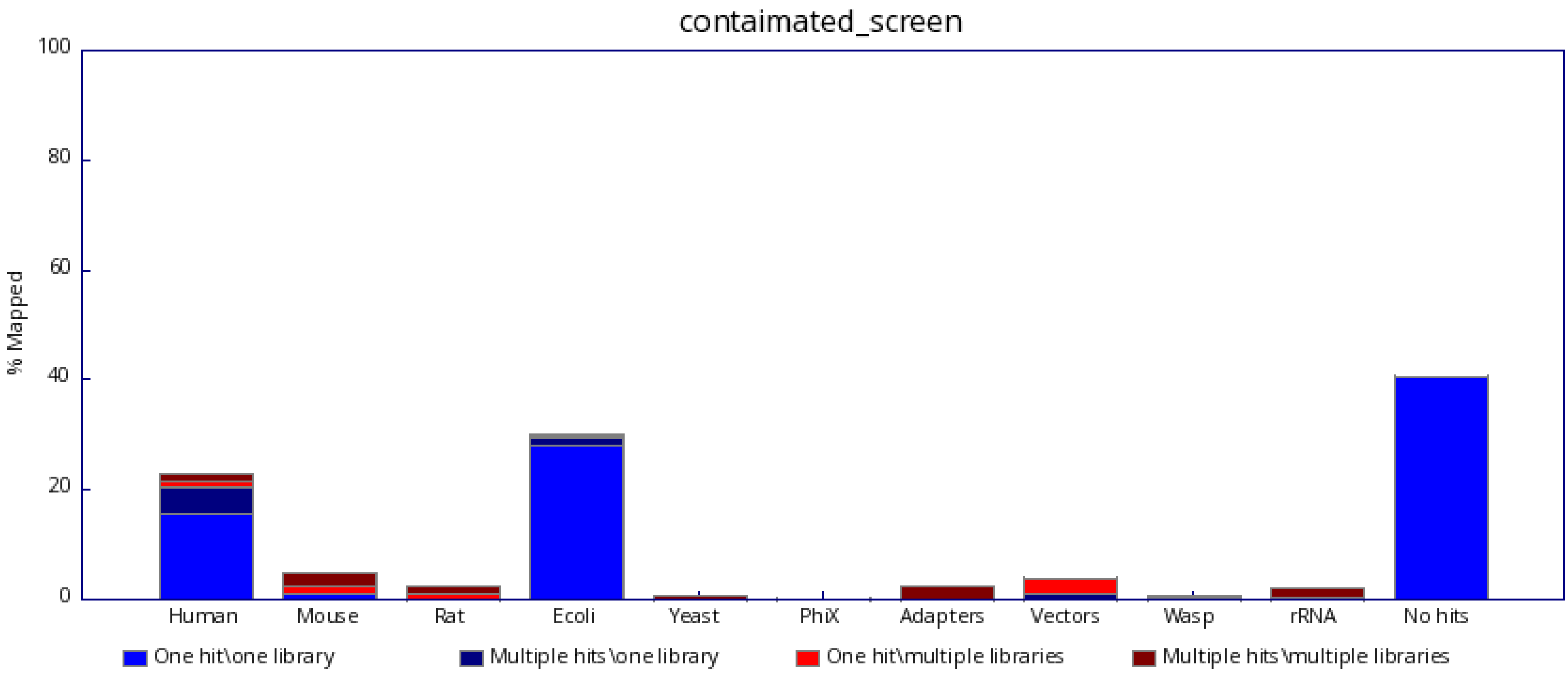
One Genome Hits Should Match The Species Sequenced

Library Screening

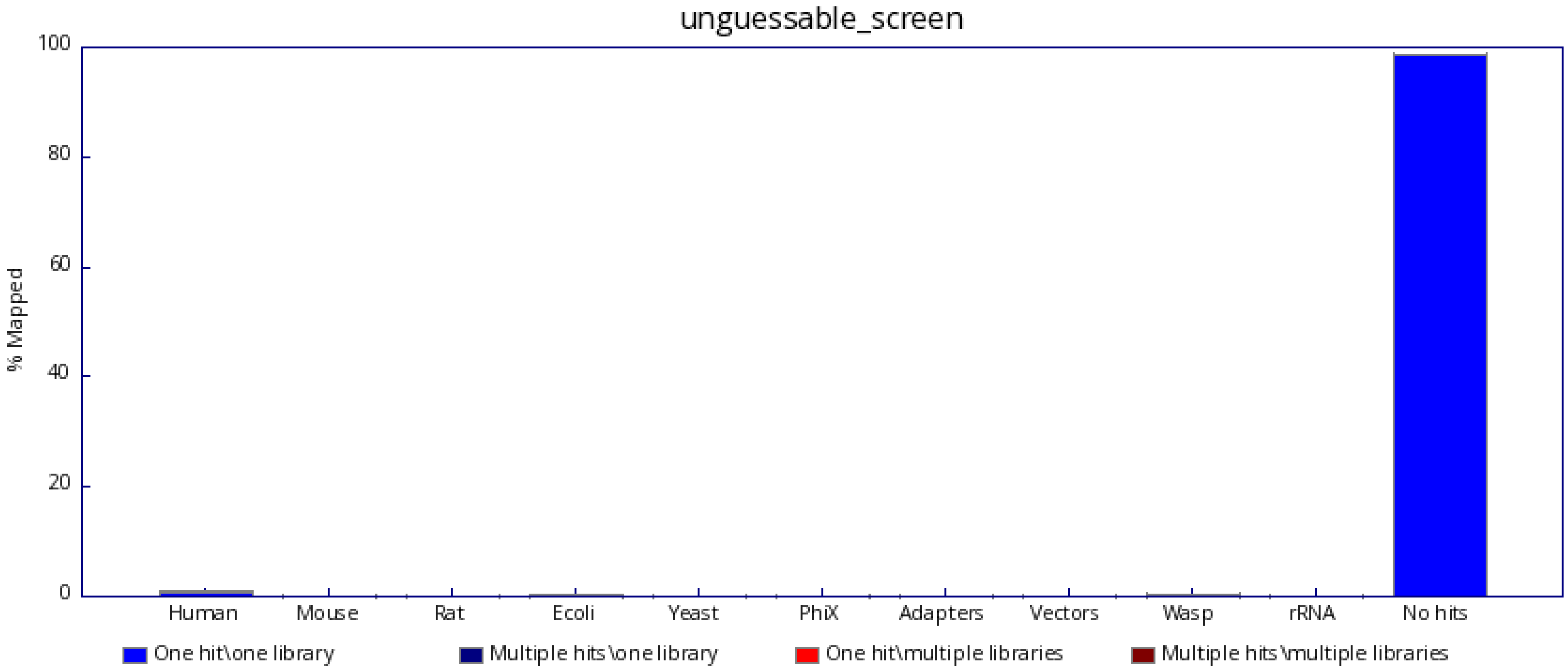


One Genome Hits Should Match The Species Sequenced

Library Screening



Library Screening



Exercise Part 1: Assessing Universal Metrics

Dataset 1

Dataset 2

Dataset 3

FastQC

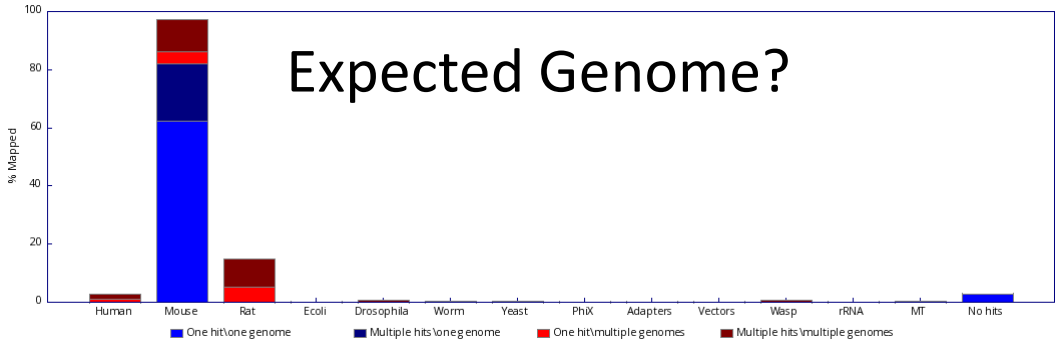
one for each read

Concern?

- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Adapter Content

FastScreen

Expected Genome?



There are QC problems with all of these libraries, can you spot them?

Assessing Library Dependent Metrics

Babraham Bioinformatics

Context is Key for QC

Expectations → **Observations**

Library Dependent

Metrics which vary with type of Library

Concern or Expected?

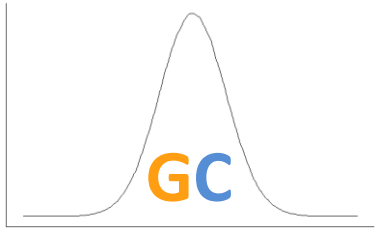
What Does it Mean?

FastQC expects a Genomic Library: Do you?

Library Dependent QC Metrics

From the Base Sequence:

- GC Content



- Base Composition

GATC

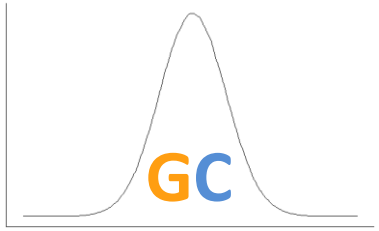
- Duplication

GATCTACGAGTTACGATCAGT
GATCTACGAGTTACGATCAGT
GATCTACGAGTTACGATCAGT
GATCTACGAGTTACGATCAGT

Library Dependent QC Metrics

From the Base Sequence:

- GC Content



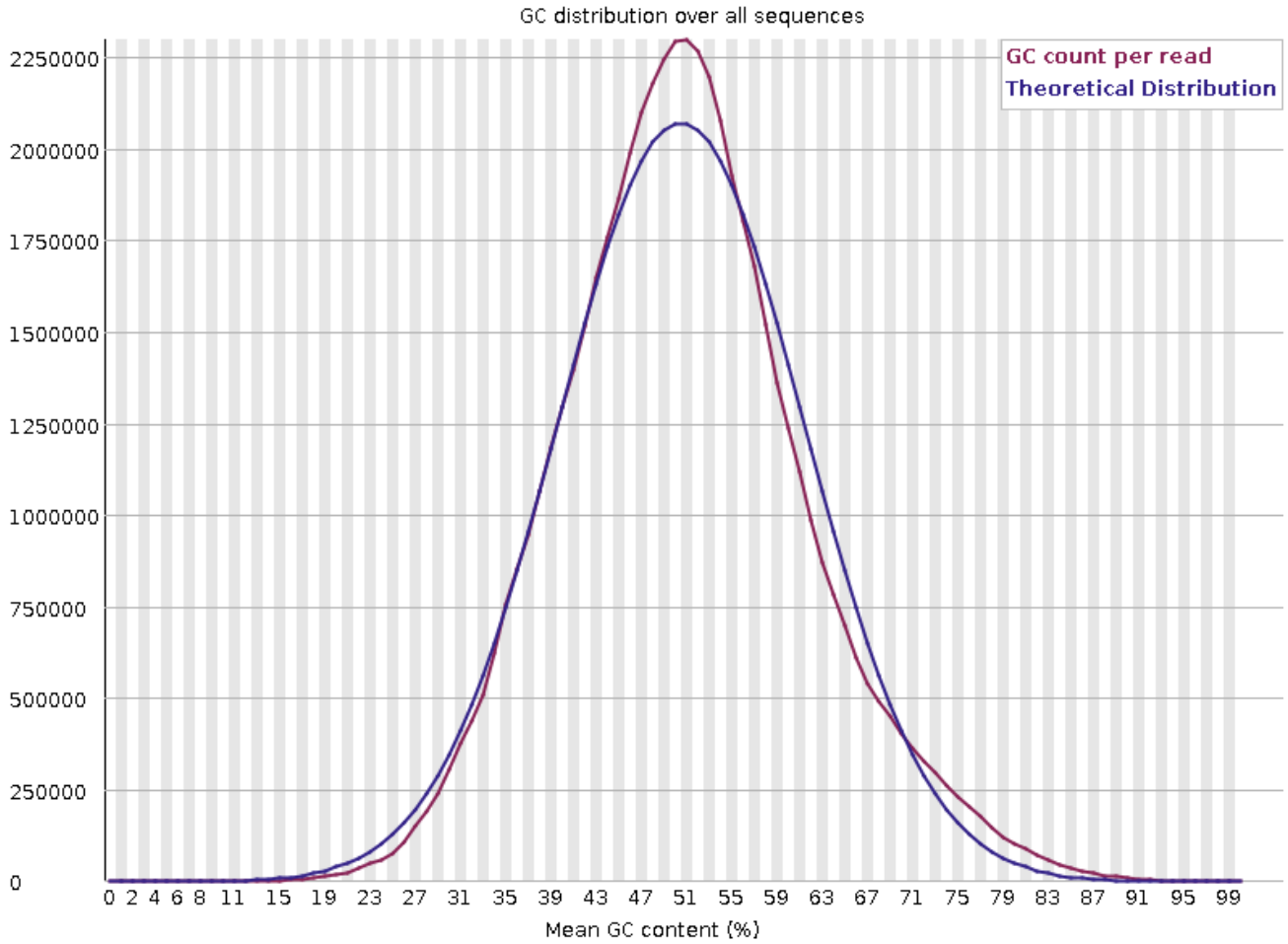
- Base Composition

GATC

- Duplication

GATCTACGAGTTACGATCAGT
GATCTACGAGTTACGATCAGT
GATCTACGAGTTACGATCAGT
GATCTACGAGTTACGATCAGT

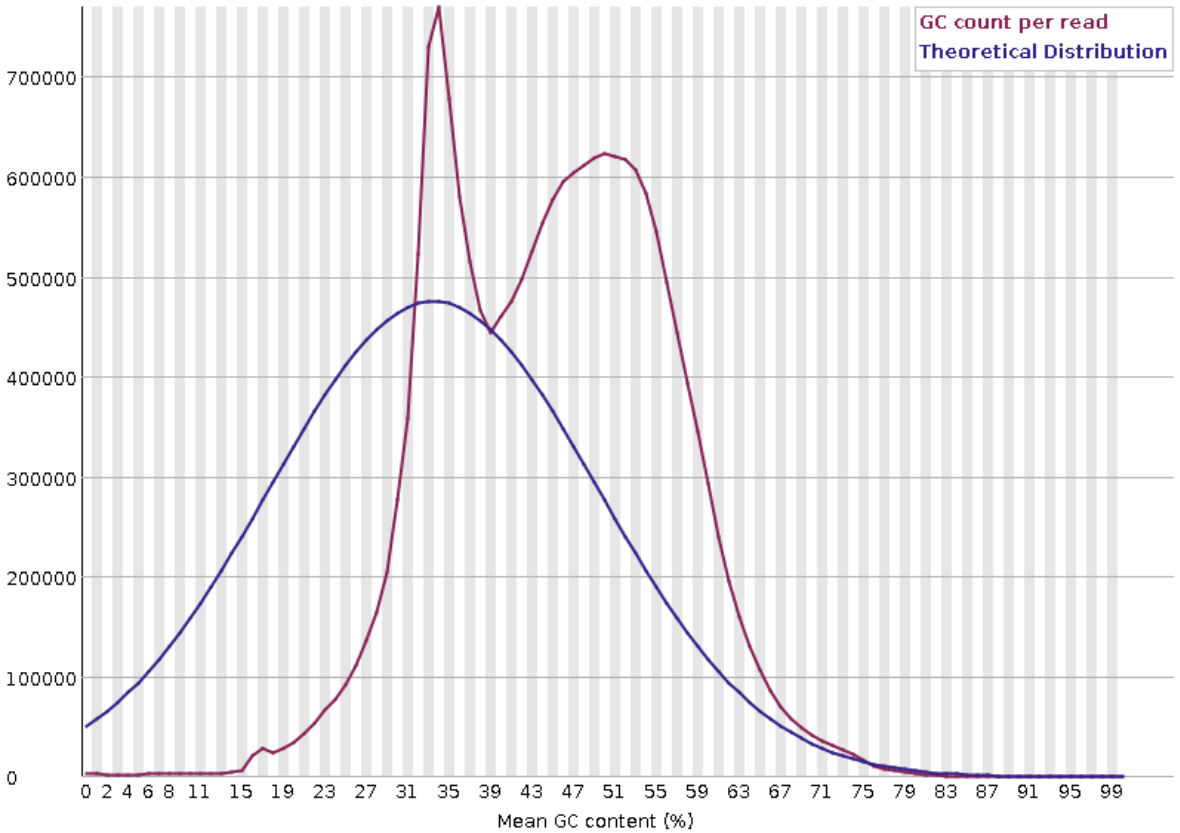
Library GC Content

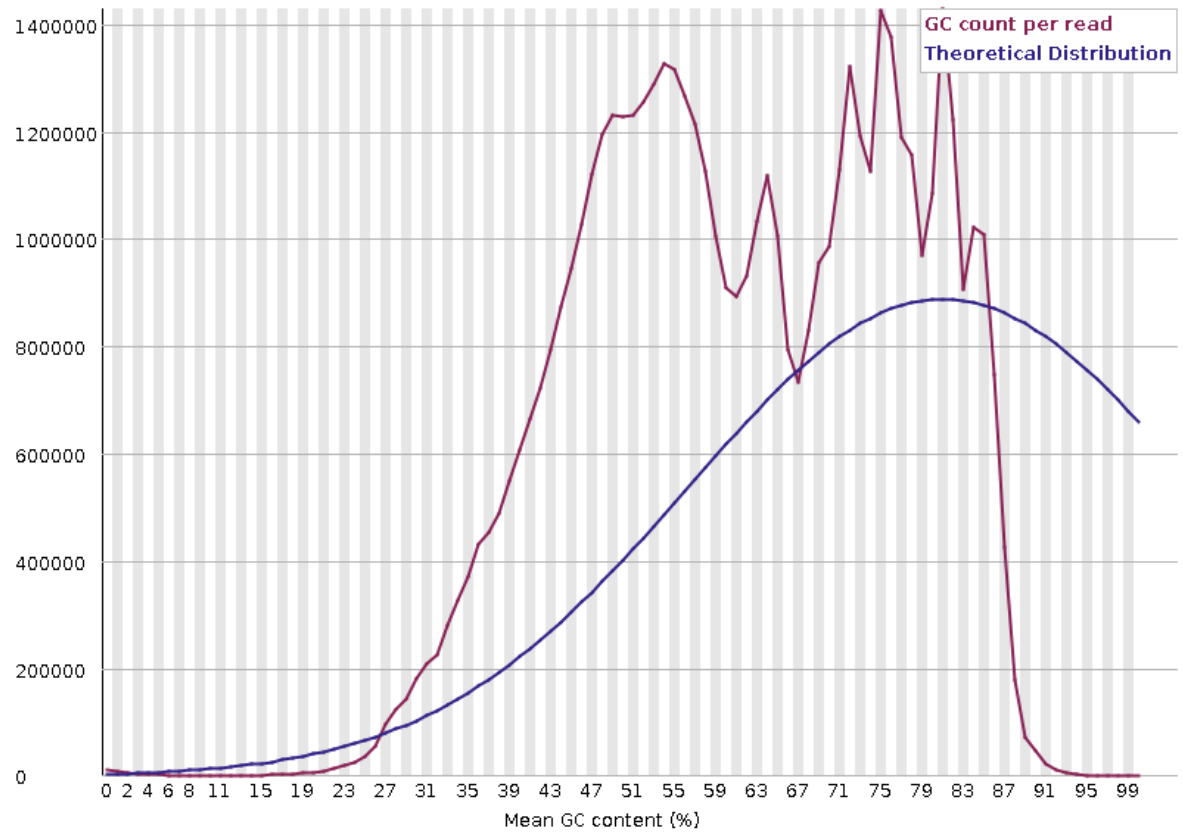


- Generic summary of library composition at a read level
- Expect a normally distributed set of values centred on the overall GC content

Sharp Peaks in GC

Concern or Expected



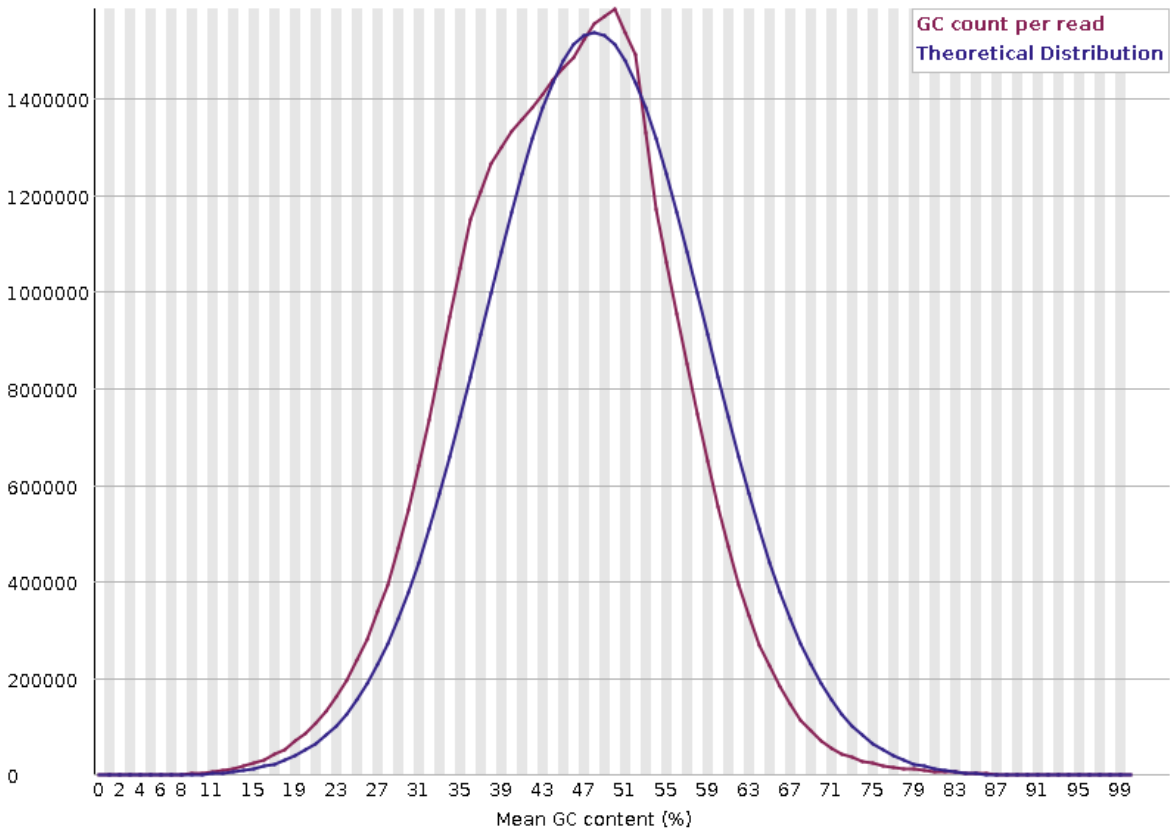


Specific Contamination with single sequence or closely related sequences

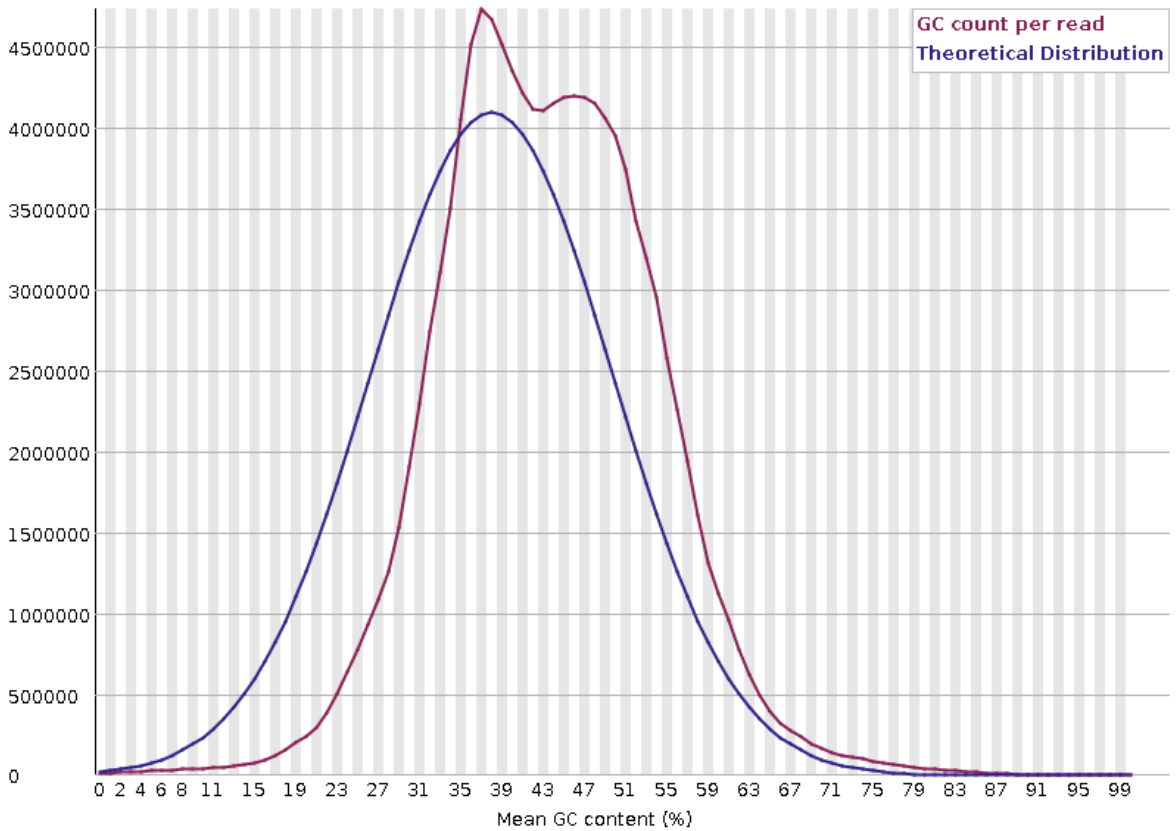
Artificial sequences, ribosomal RNA, contaminants

Broader Peaks in GC
Concern or Expected

Mouse **Drosophila**



H3K4me1 ChIP: enrich for regions rich in GC

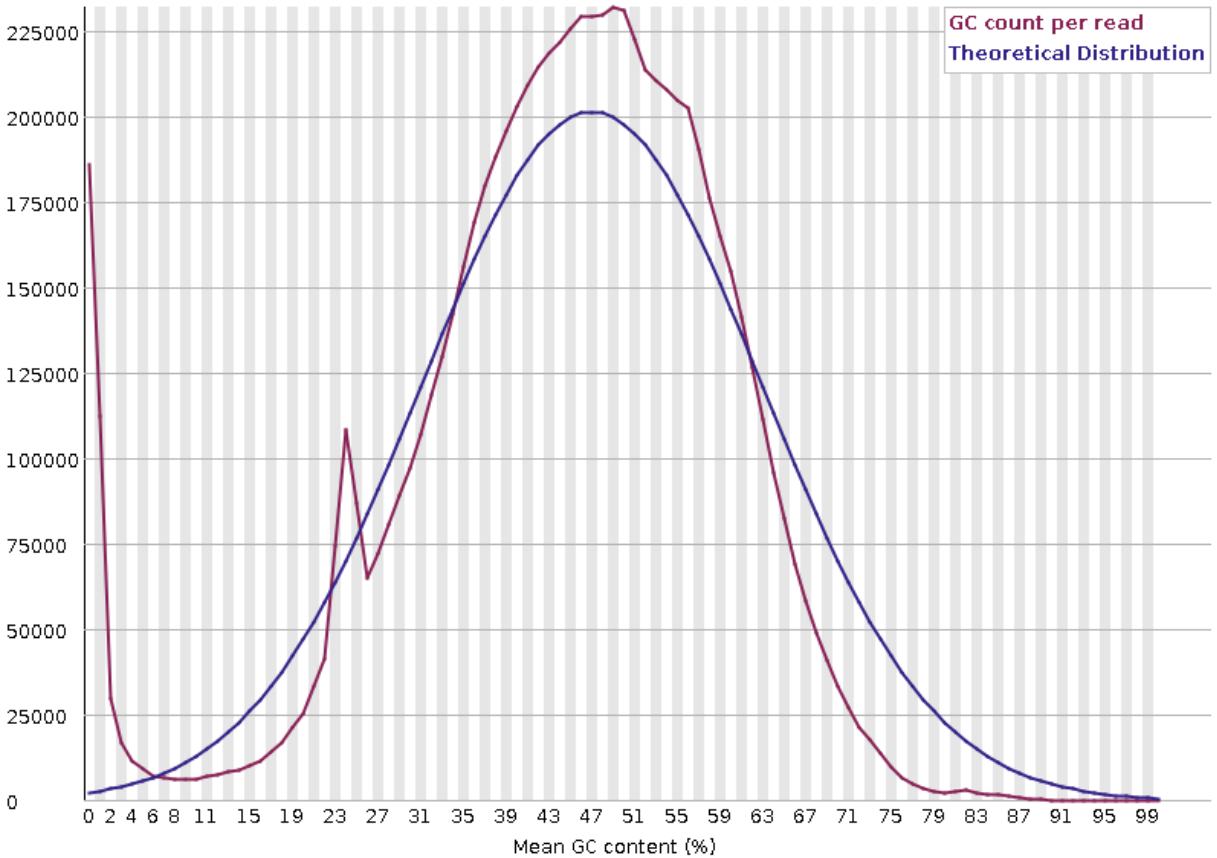


More extensive mixture of reads with different GC content

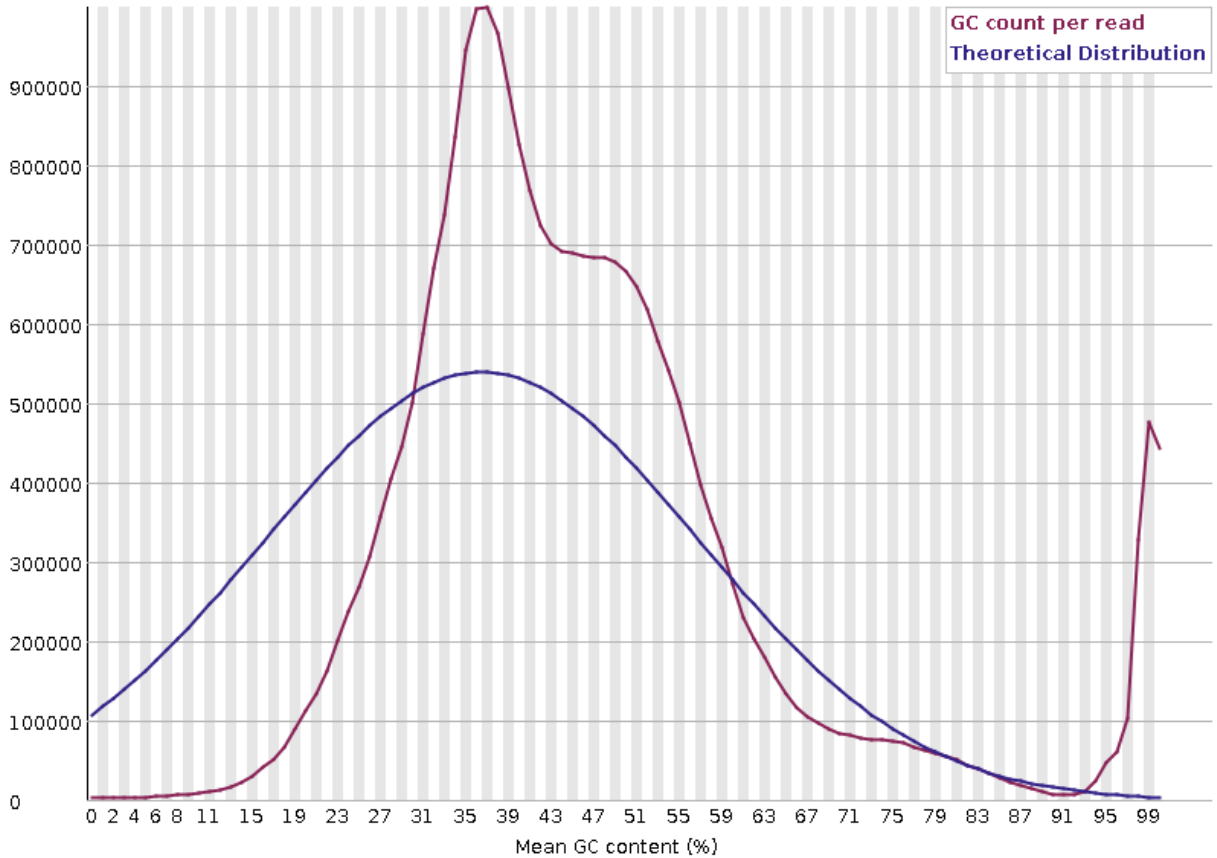
GC Skew

Concern or Expected

PolyA's



PolyG's

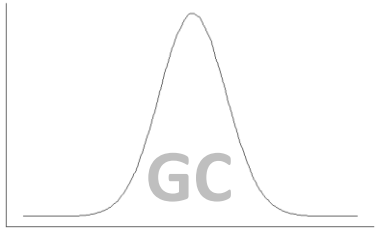


More extensive subset of reads with extreme differences in GC

Library Dependent QC Metrics

From the Base Sequence:

- GC Content



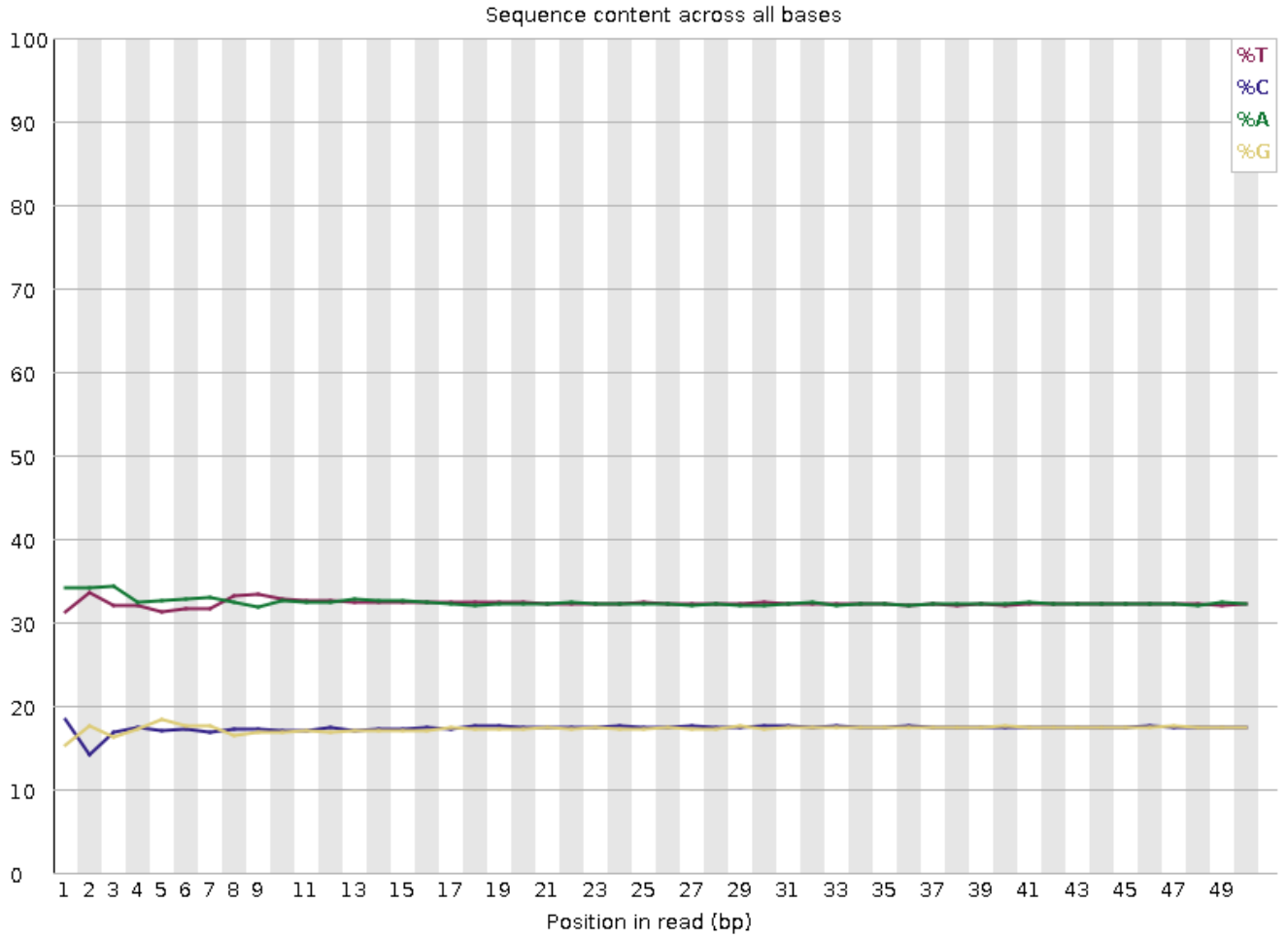
- Base Composition

GATC

- Duplication

GATCTACGAGTTACGATCAGT
GATCTACGAGTTACGATCAGT
GATCTACGAGTTACGATCAGT
GATCTACGAGTTACGATCAGT

Library Base Composition

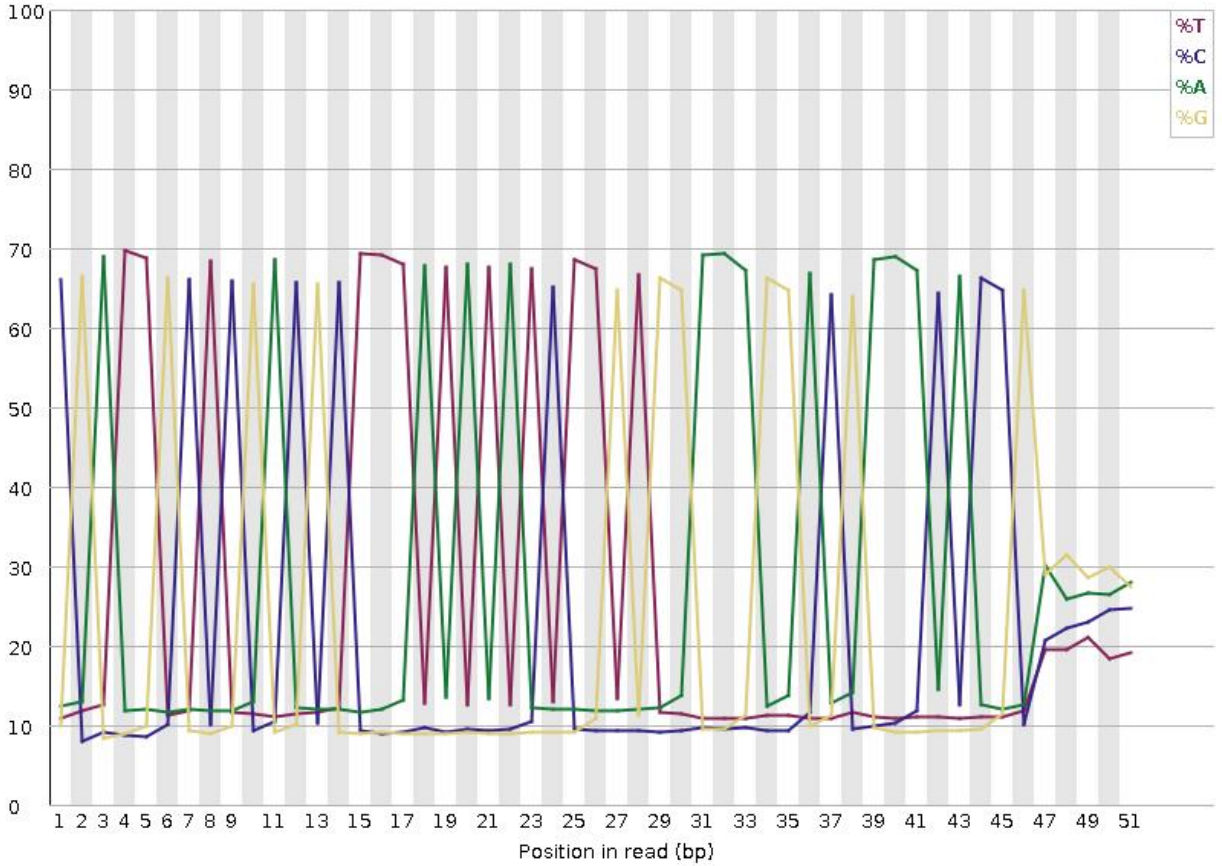


- For every chemistry cycle we can look at the number of ATGC we call
- For Libraries with random start positions the composition should be the same for all cycles

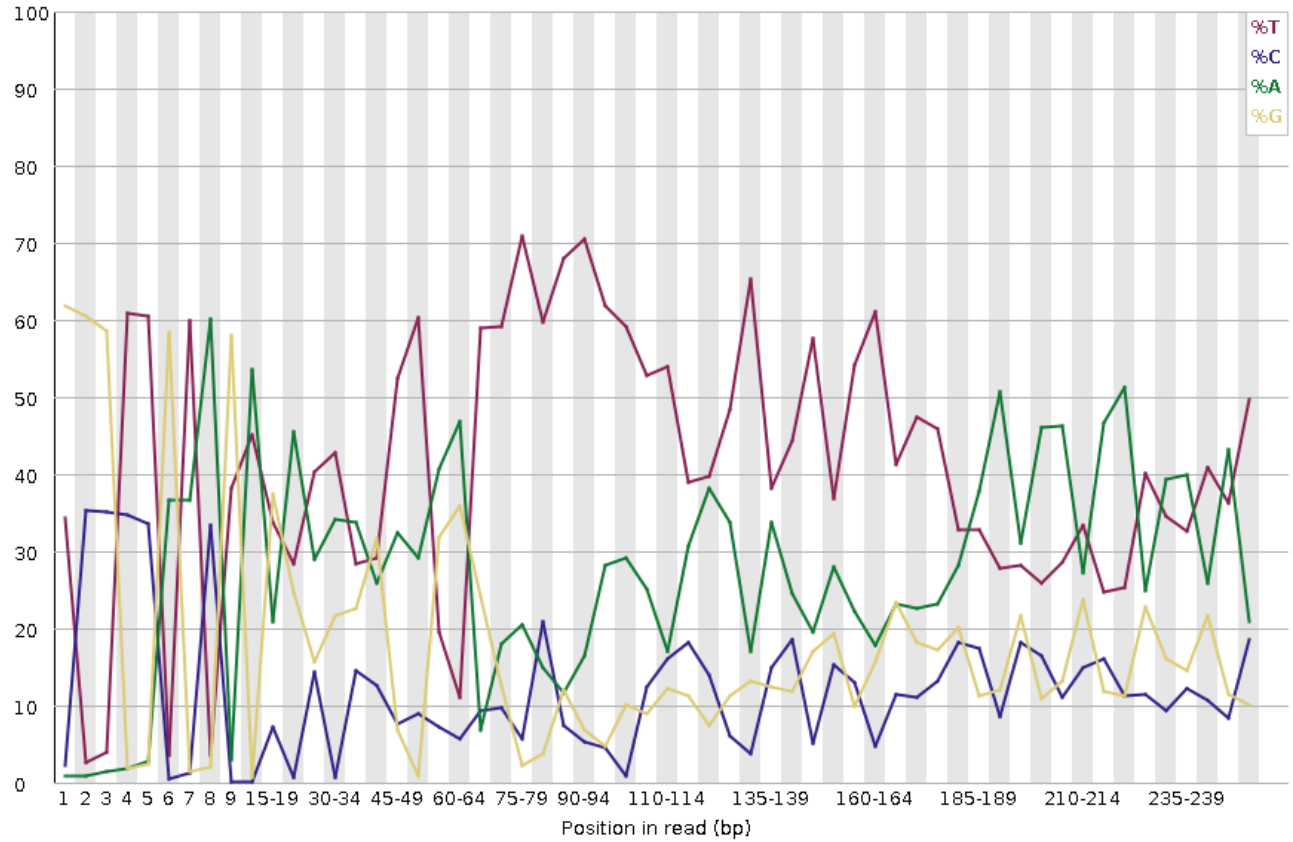
Bias Composition Throughout

Concern or Expected

Wrong Sequence



Amplicon

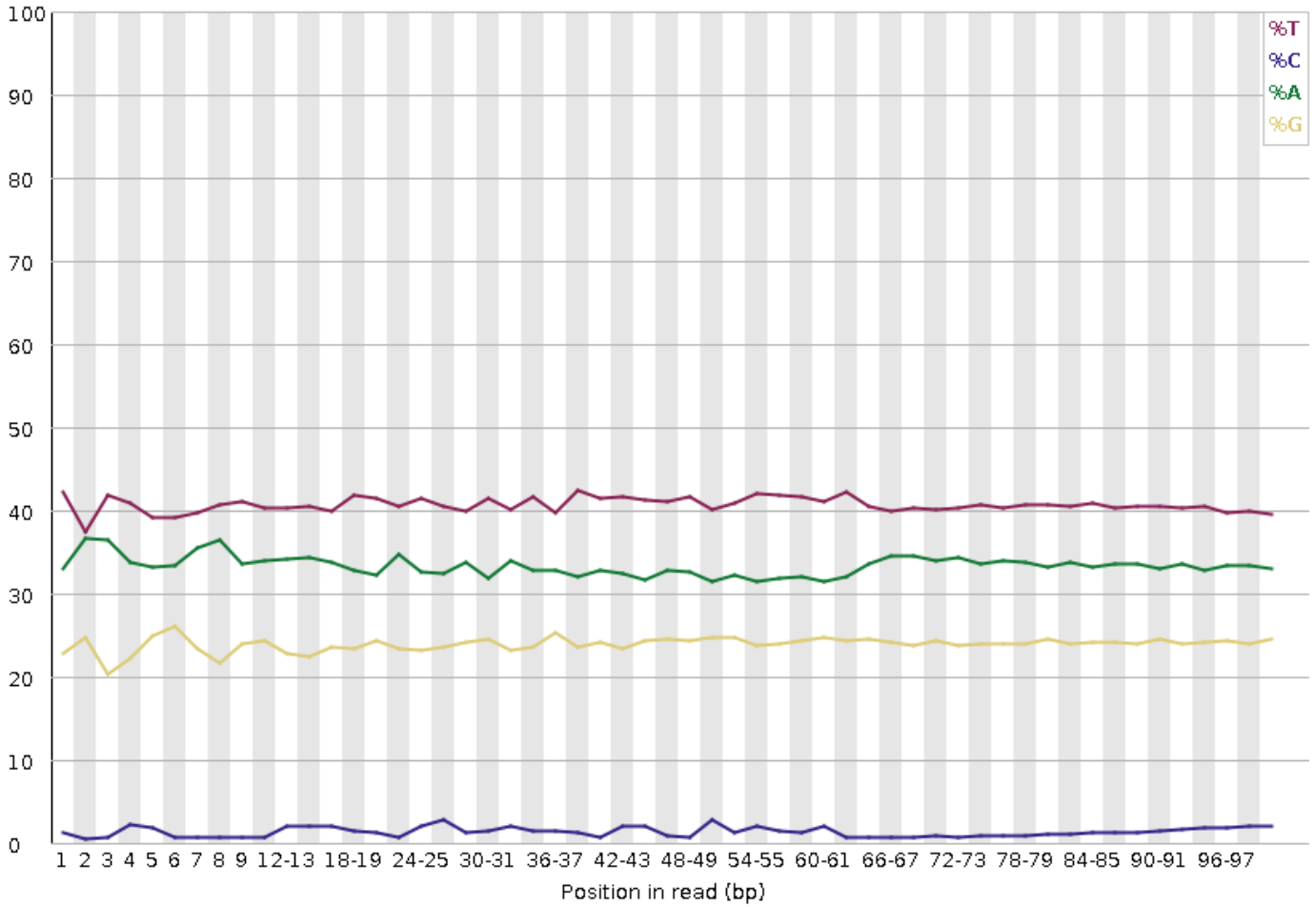


Proportional biases of bases at specific positions: Very low diversity

Bias Composition Throughout

Concern or Expected

WGBS

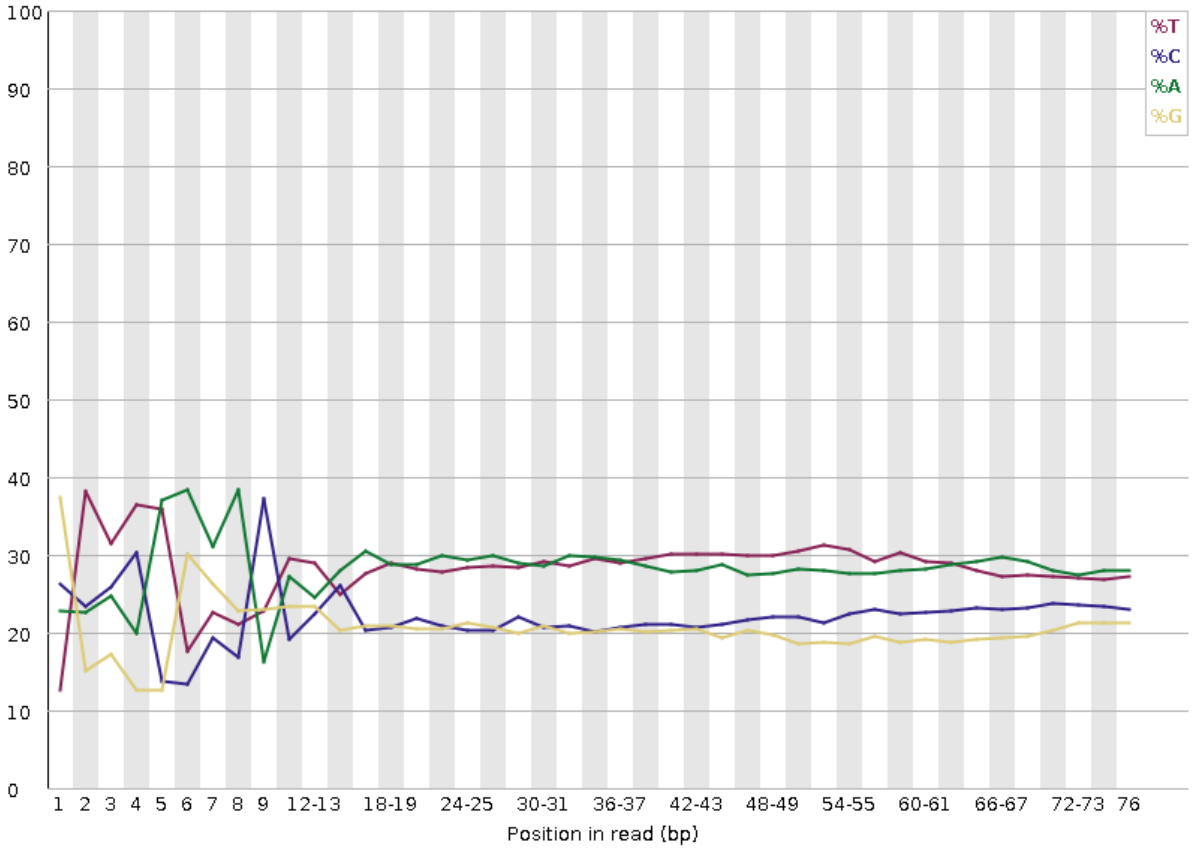


Consistent disproportional expression of bases

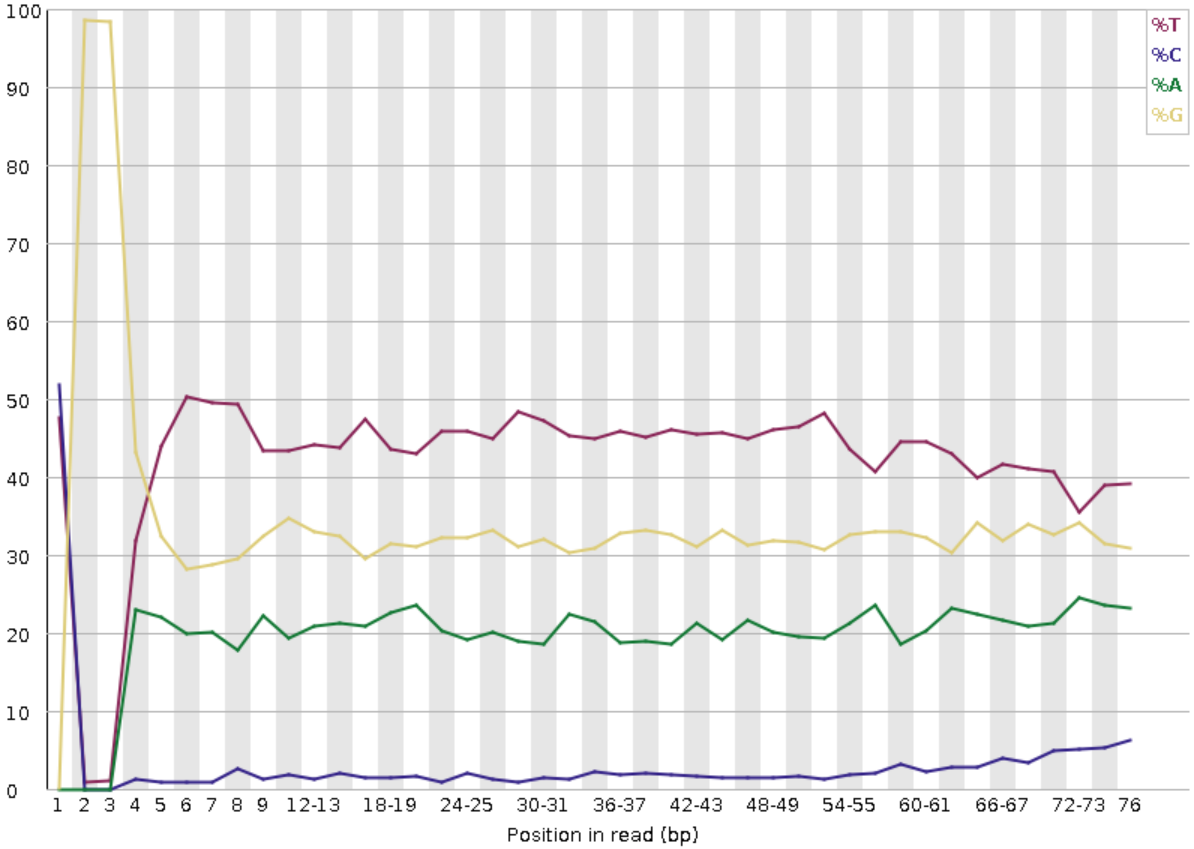
Bias Composition at 5’ end

Concern or Expected

ATAC – Transposases



RRBS – Restriction Start Site

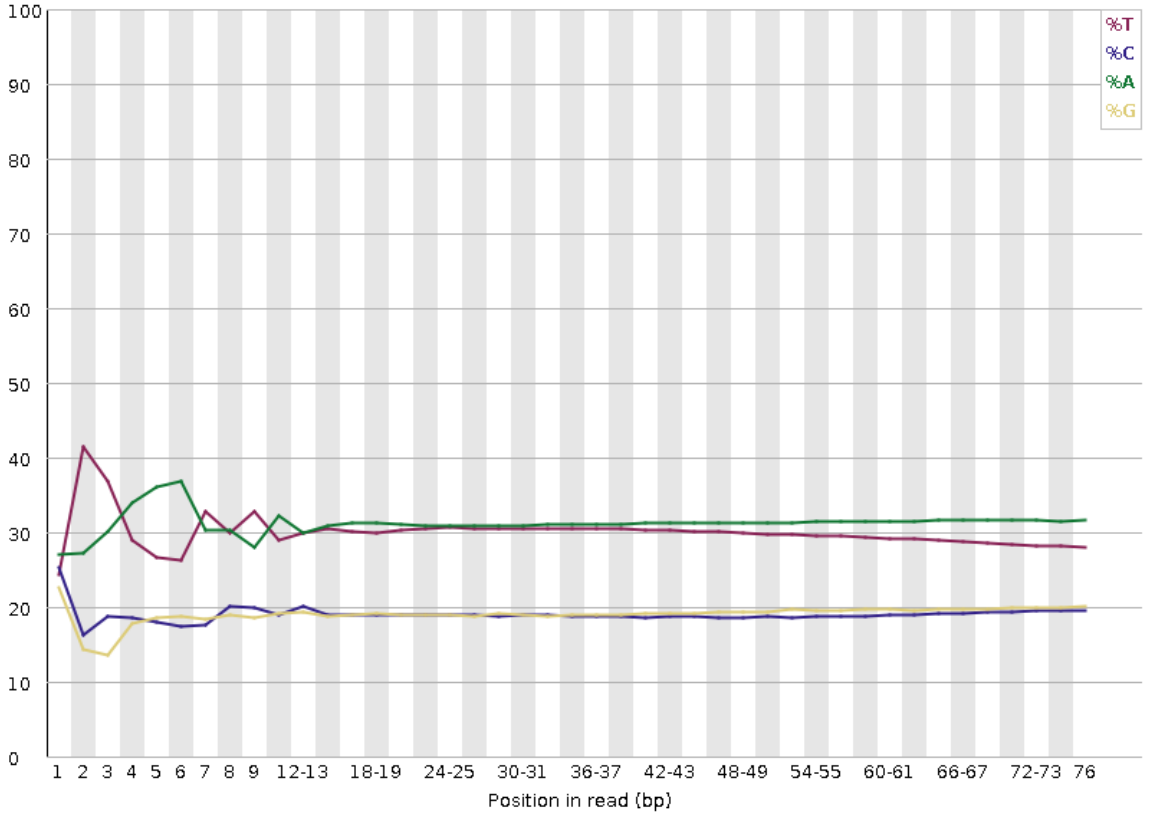


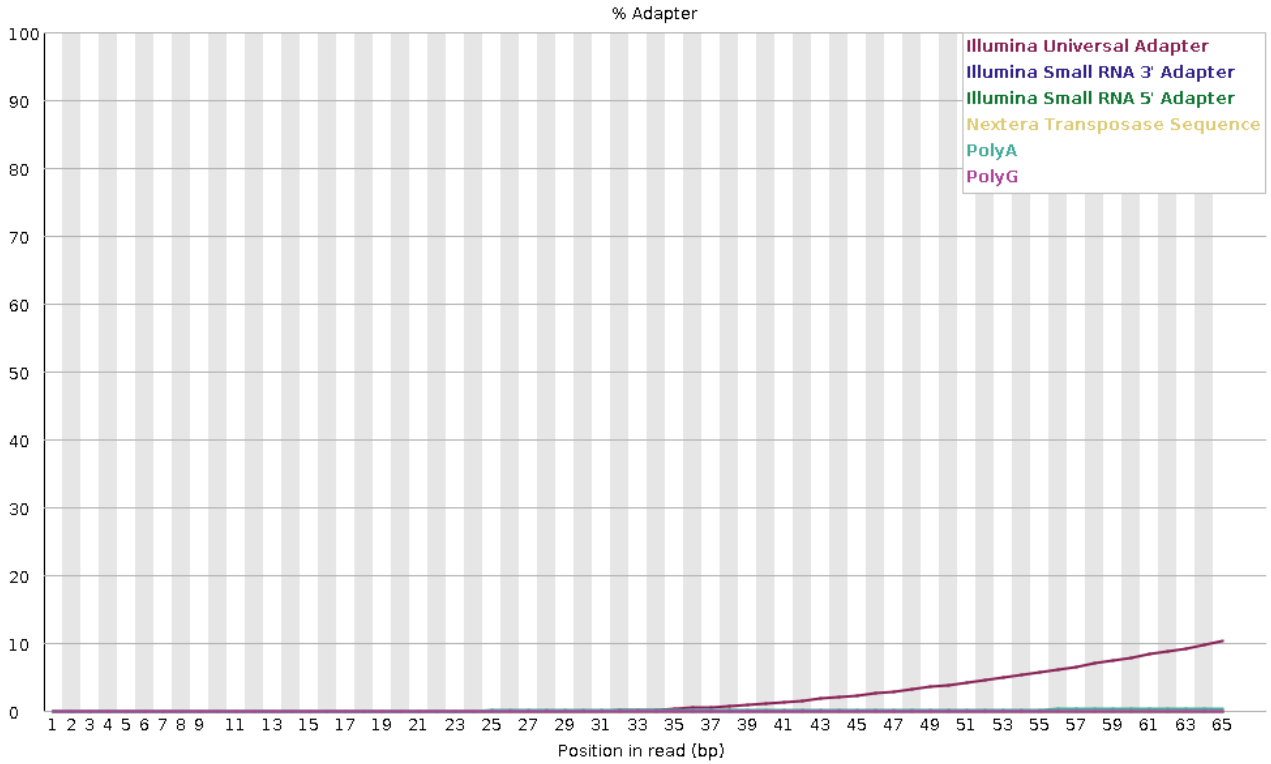
Proportional biases of bases at the start of a read: A preferred start site

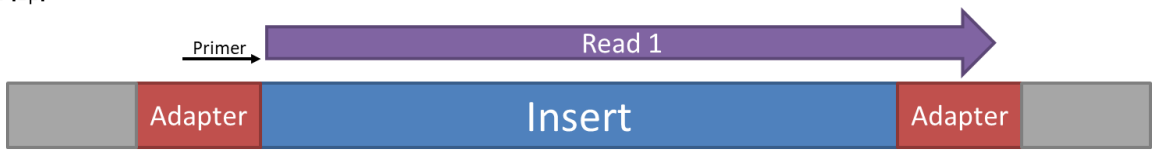
Bias Composition at 3’ end

Concern or Expected

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



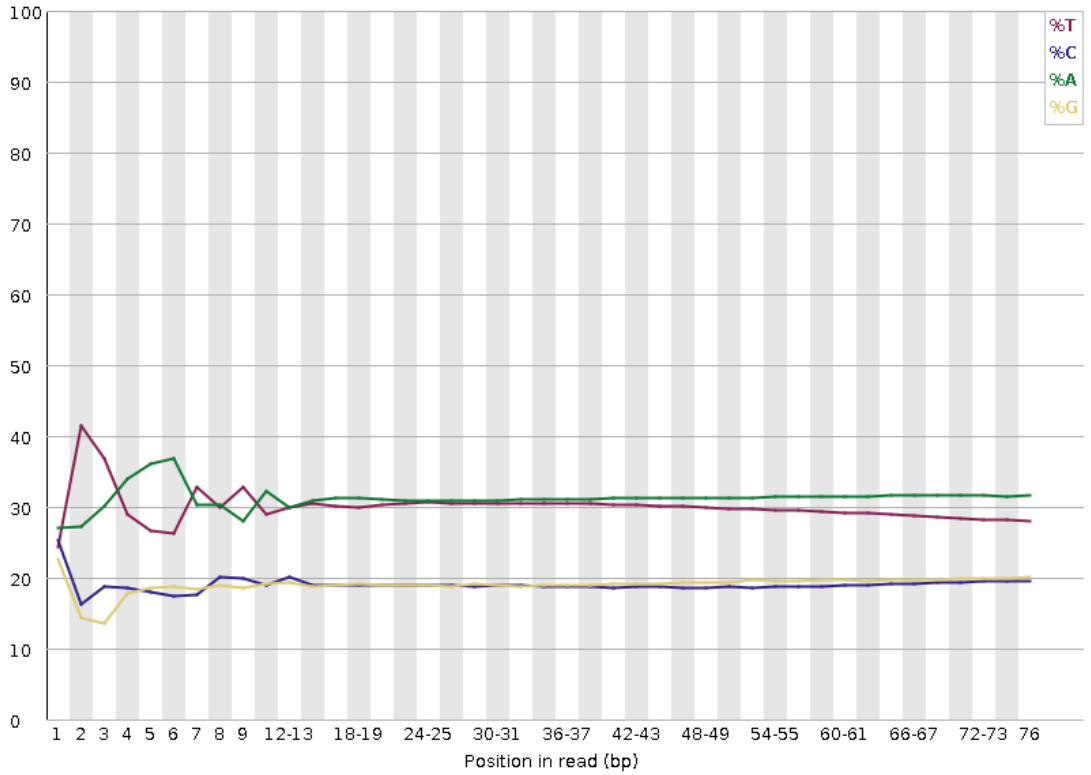


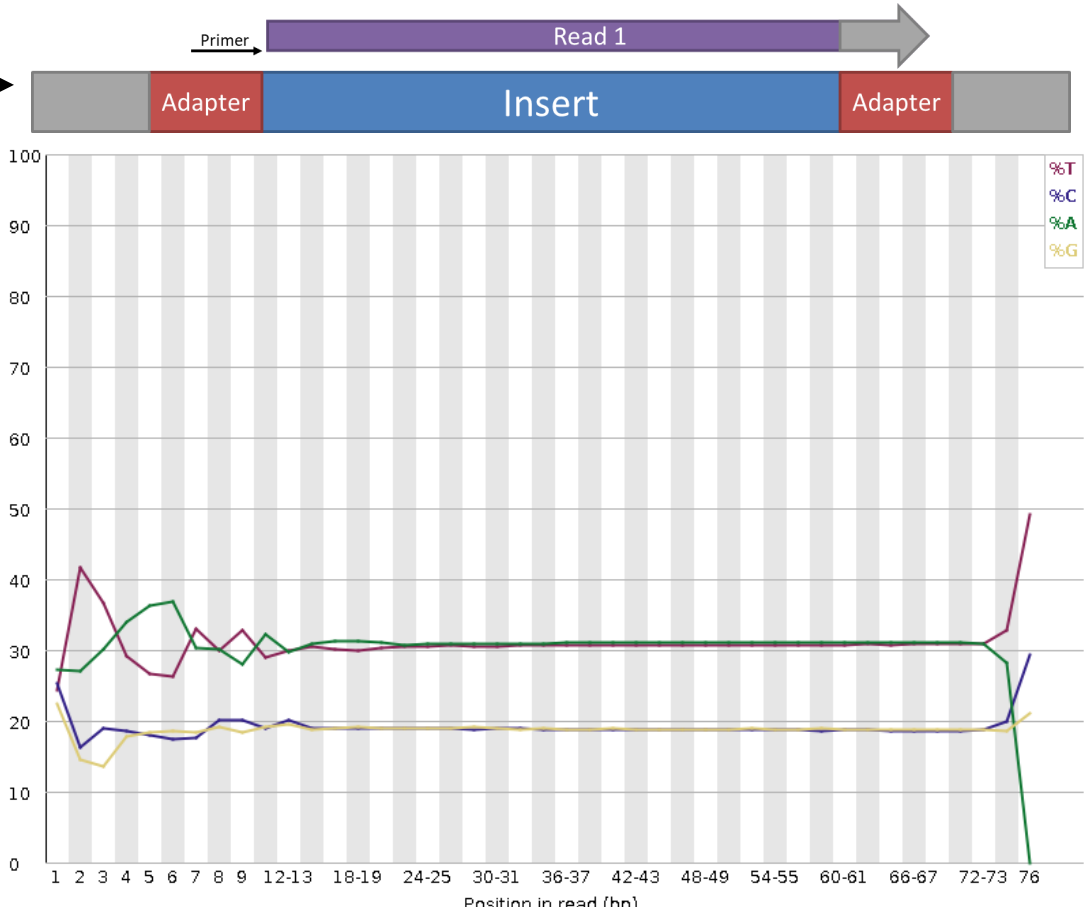


Bias Composition at 3’ end

Concern or Expected

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



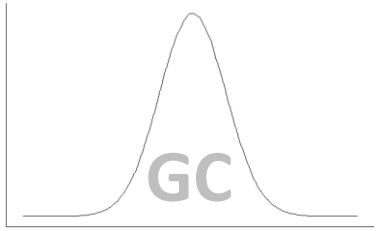


Bioinformatics processing can also influence QC metrics!

Library Dependent QC Metrics

From the Base Sequence:

- GC Content



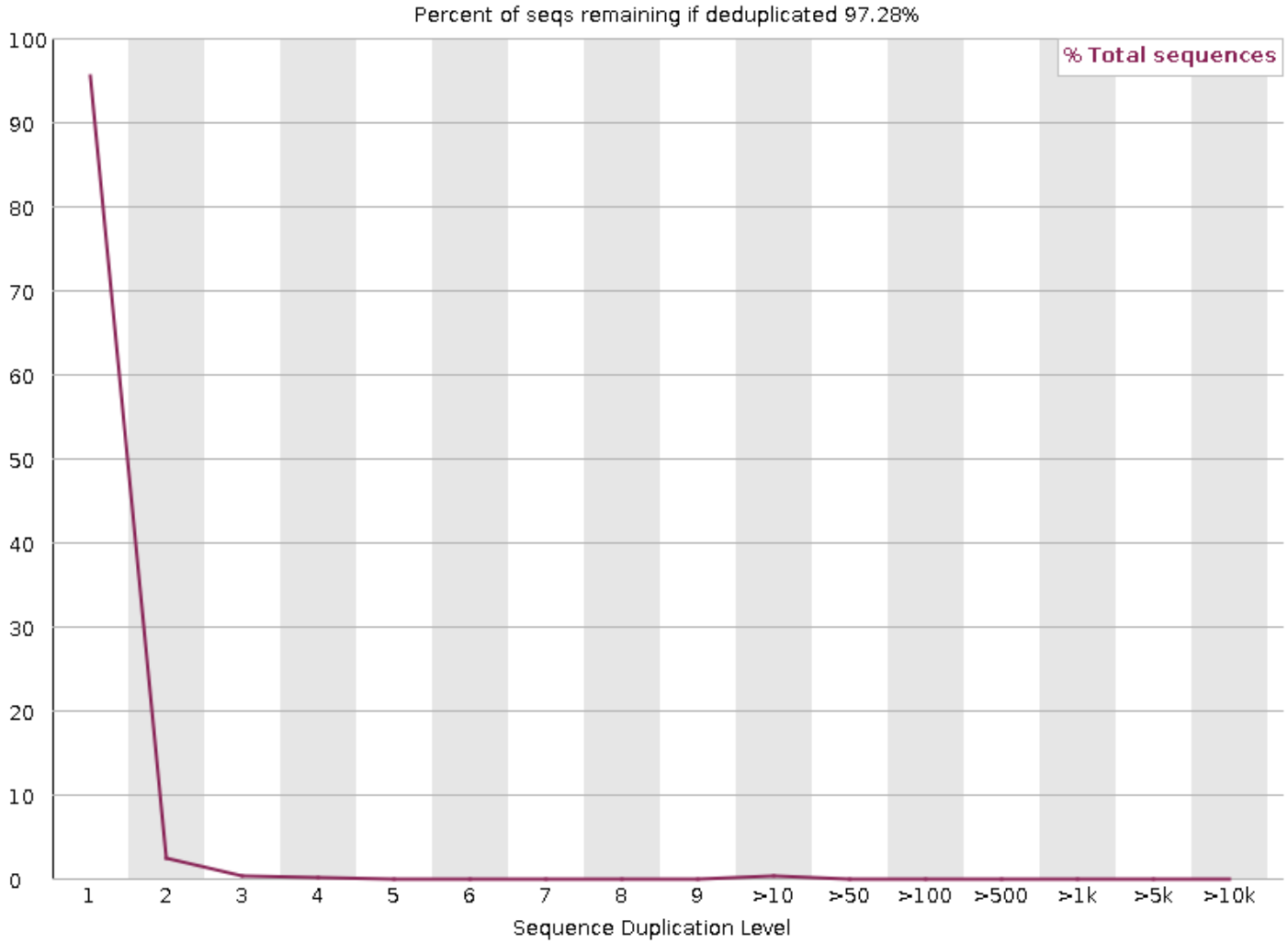
- Base Composition

GATC

- Duplication

GATCTACGAGTTACGATCAGT
GATCTACGAGTTACGATCAGT
GATCTACGAGTTACGATCAGT
GATCTACGAGTTACGATCAGT

Duplication

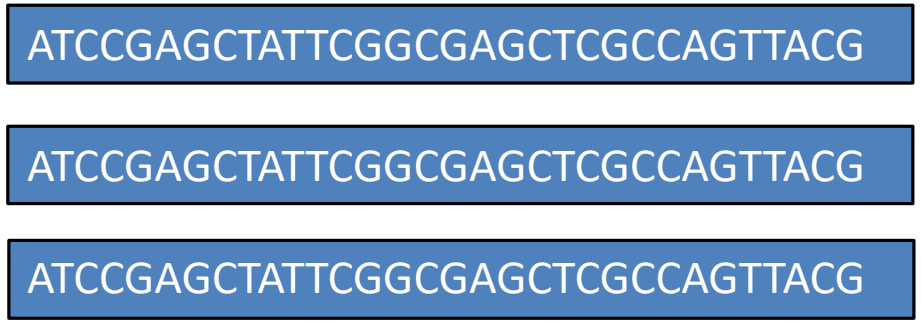


- How frequently the exact same sequence appears in your library
- For WGS expect most sequences to be unique

Duplication

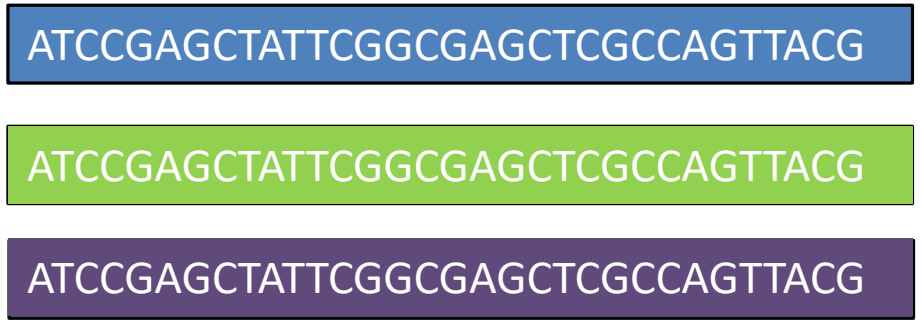
If the exact same sequence appears more than once it could be…

Technical:



- PCR duplicates

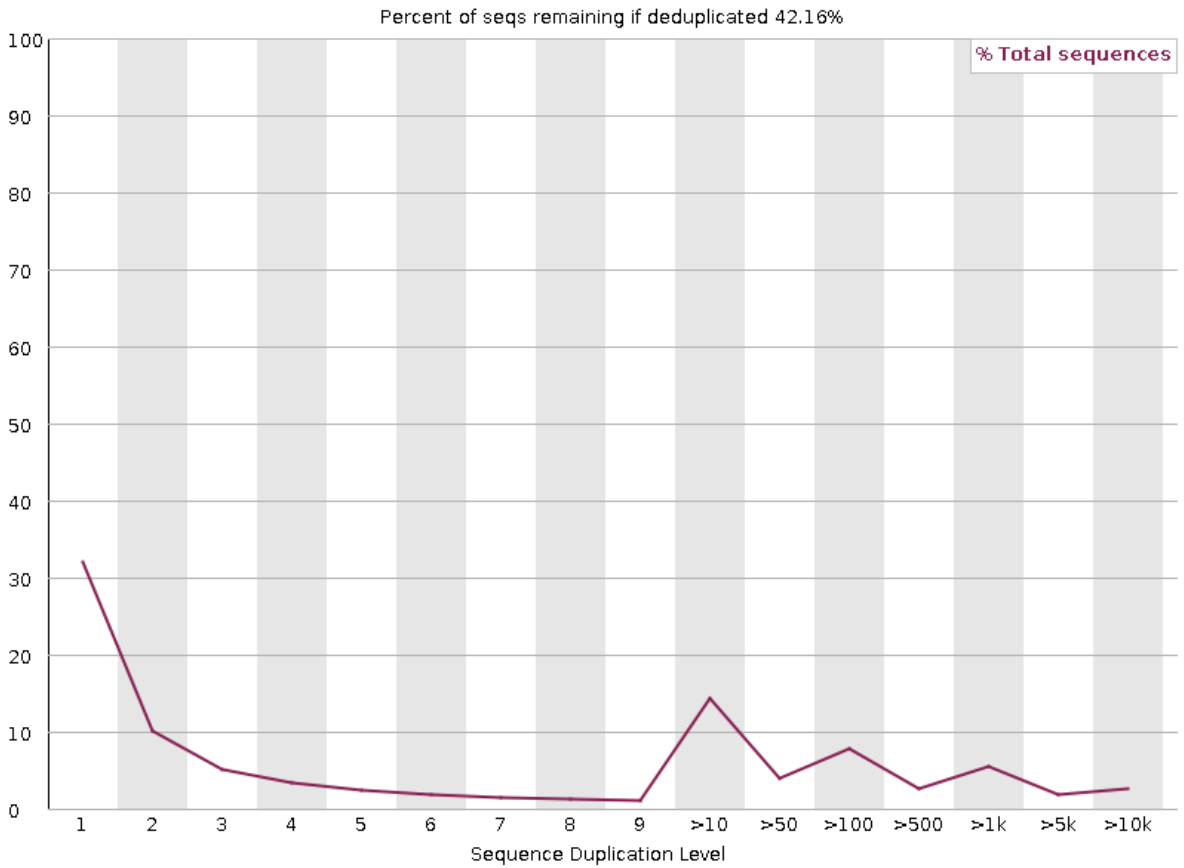
Coincidental:



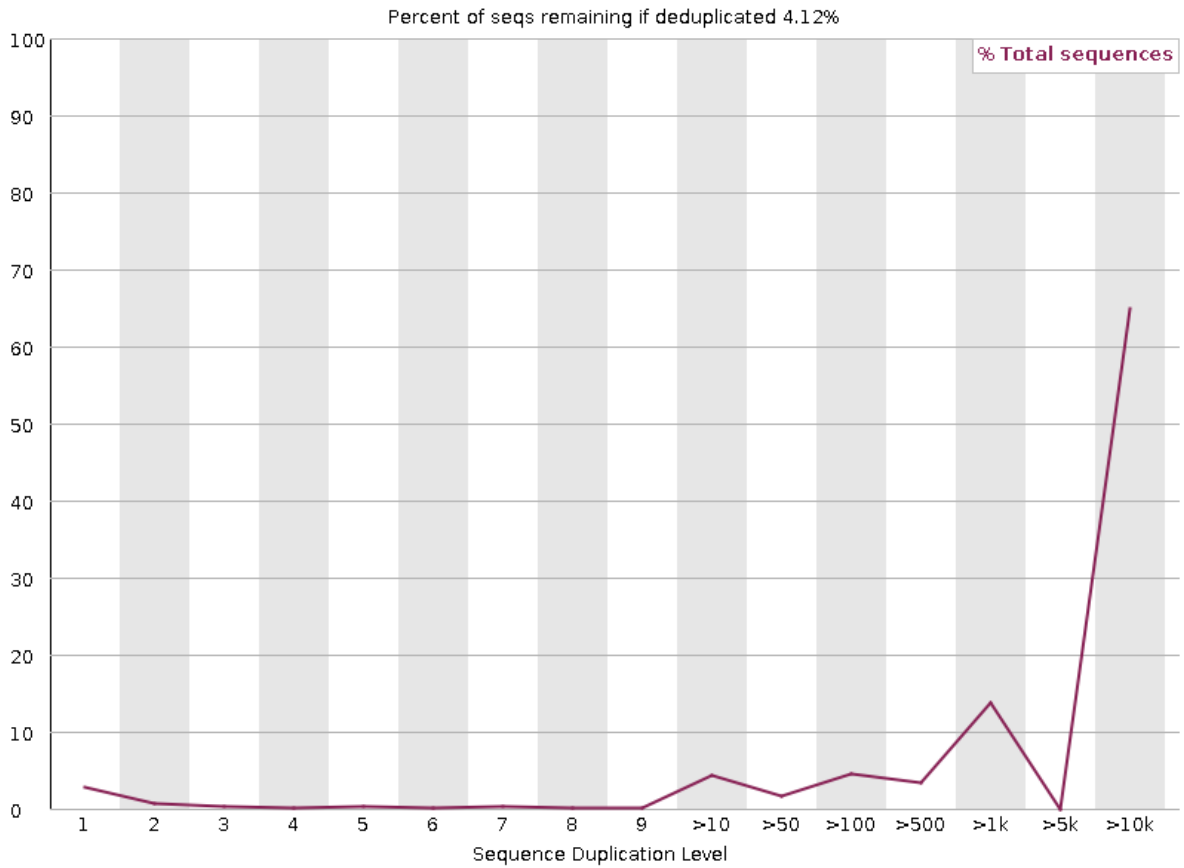
- Deep sequencing
- Highly present sequences
- Restricted diversity libraries

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

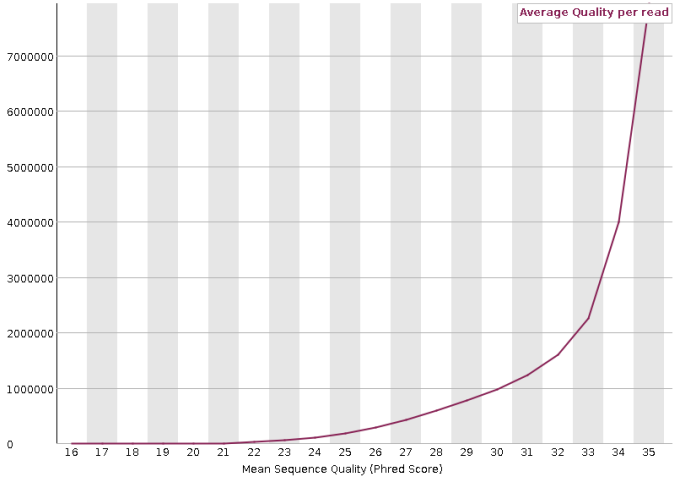
Sources of Poly Sequences

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

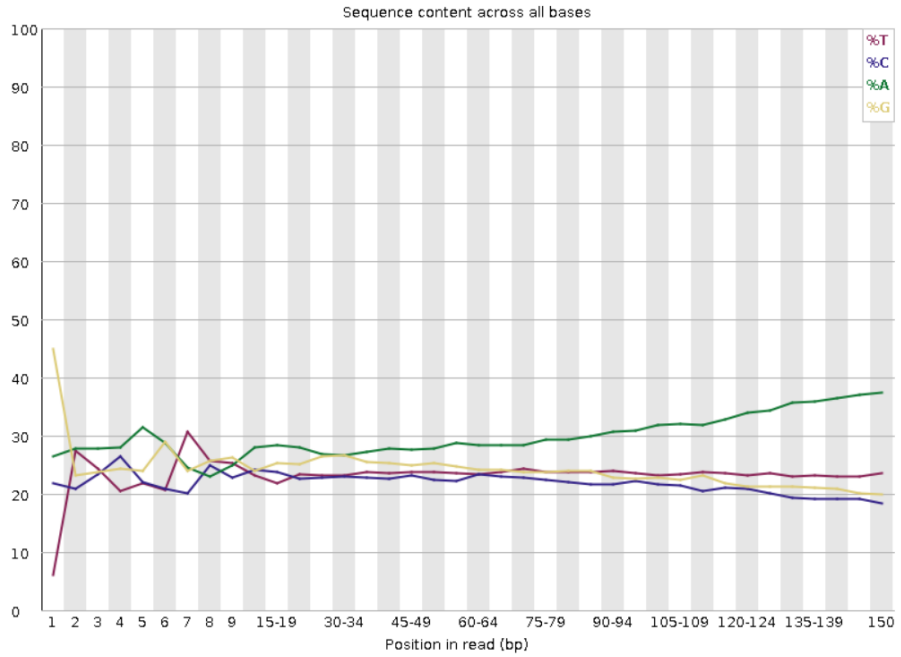
Sequence	Count	Percentage
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG	406221	1.9711963990403654
TGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG	315458	1.530766931420275
AGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG	161606	0.7841966940737117
TTGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG	46937	0.2277628320095653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

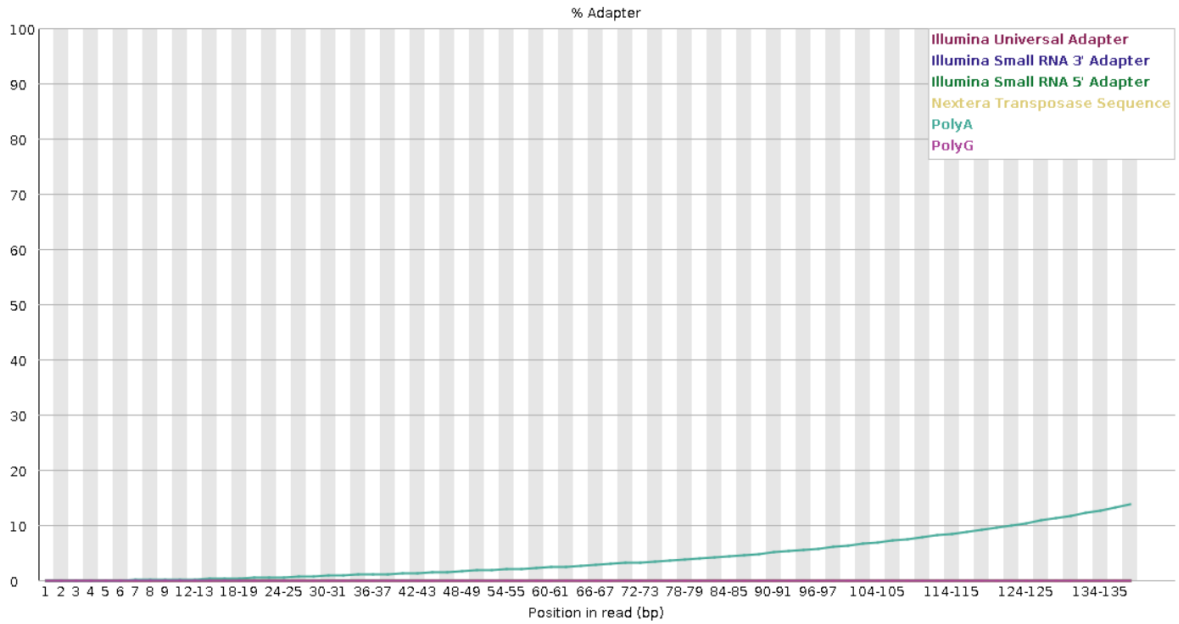


Sources of Poly Sequences

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

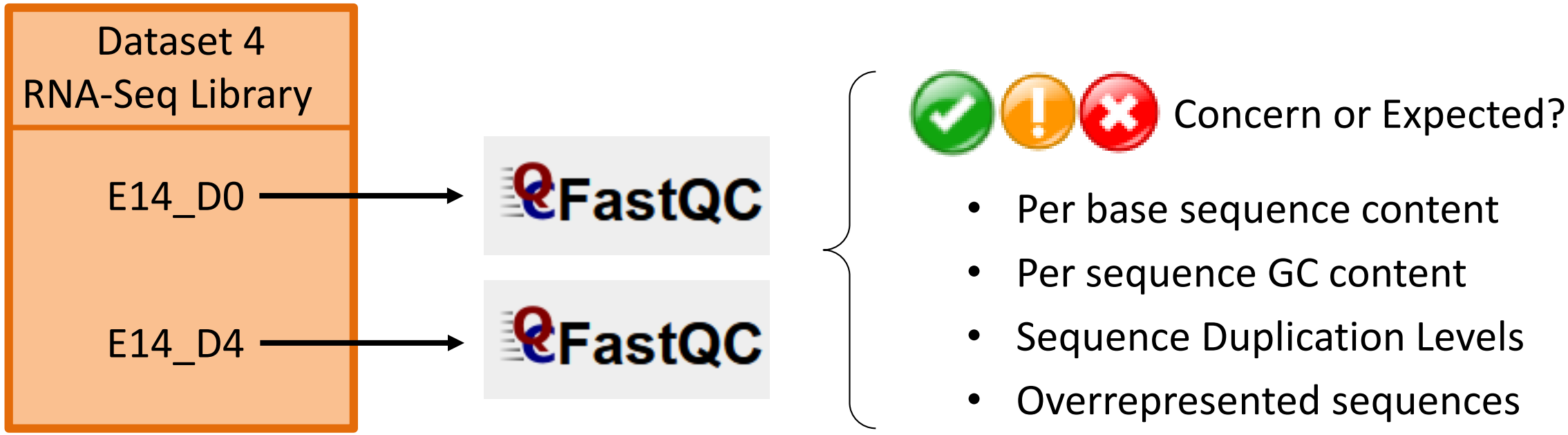
Ribosomal

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

Exercise Part 2: Assessing Library Dependent Metrics

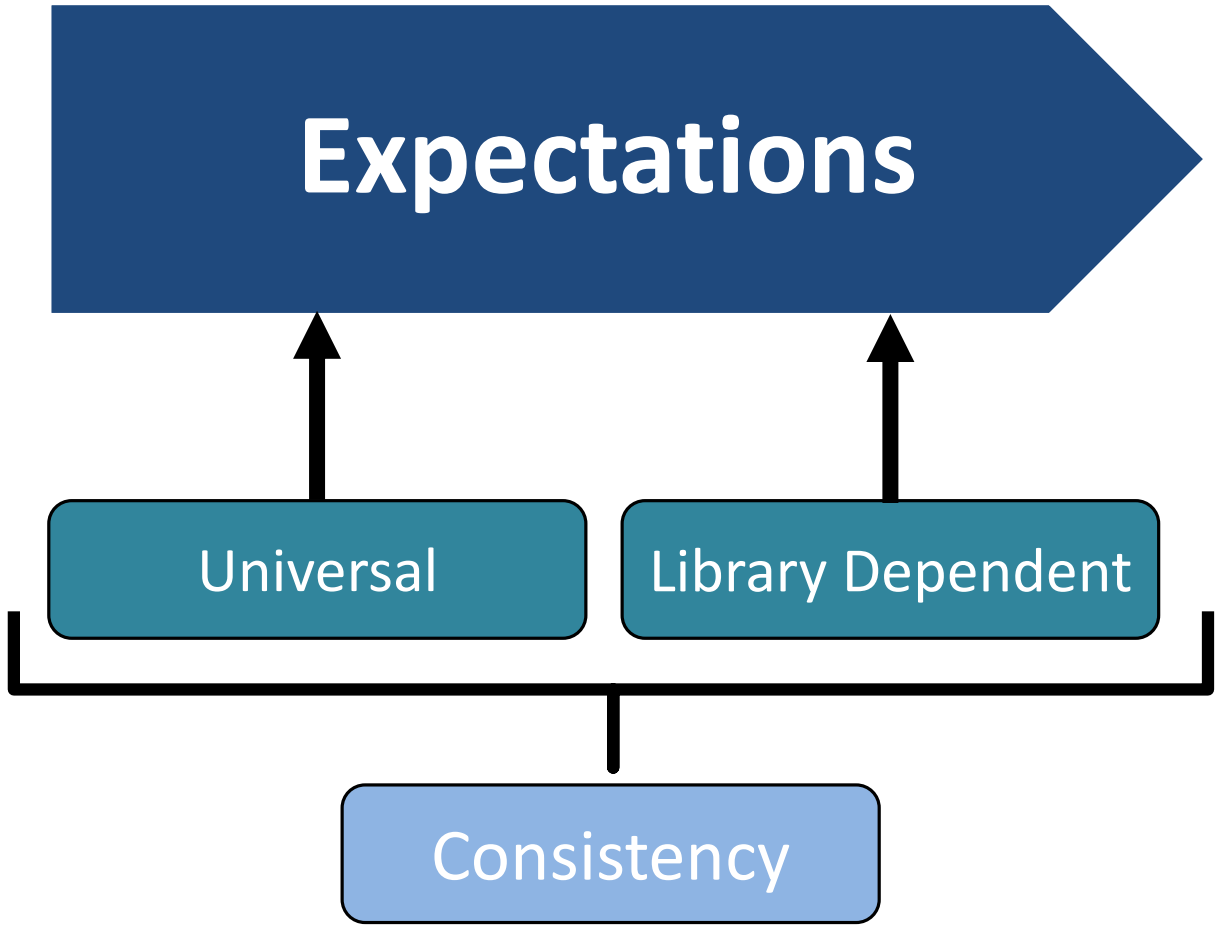


One of these samples is normal for an RNA-Seq library, the other is not. Which is which?

Assessing Consistency

Babraham Bioinformatics

Context is Key for QC

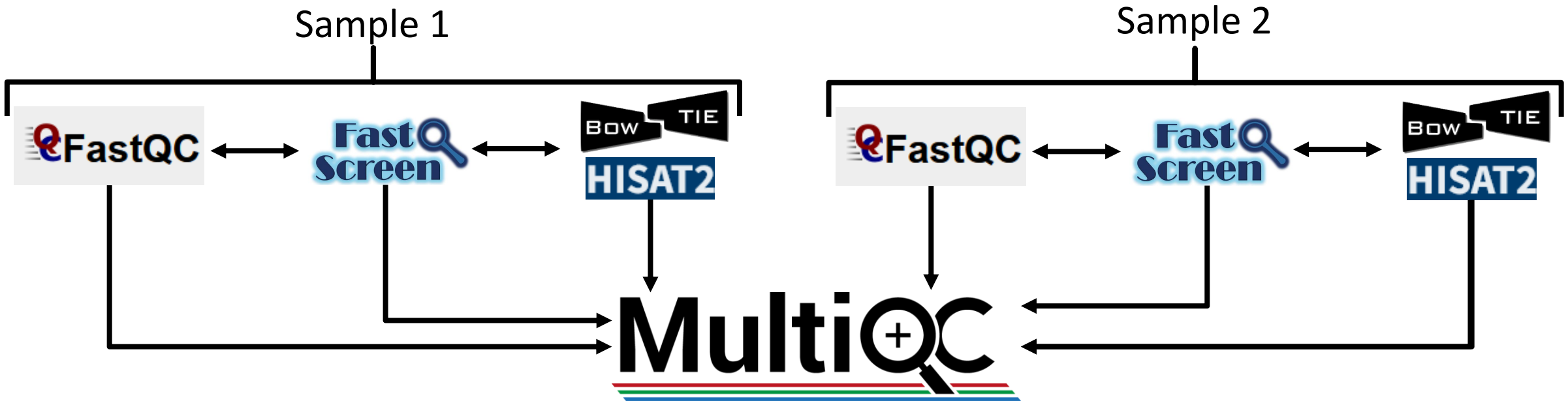


Individual Library:

Replicate Libraries:

Aggregated Statistics

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

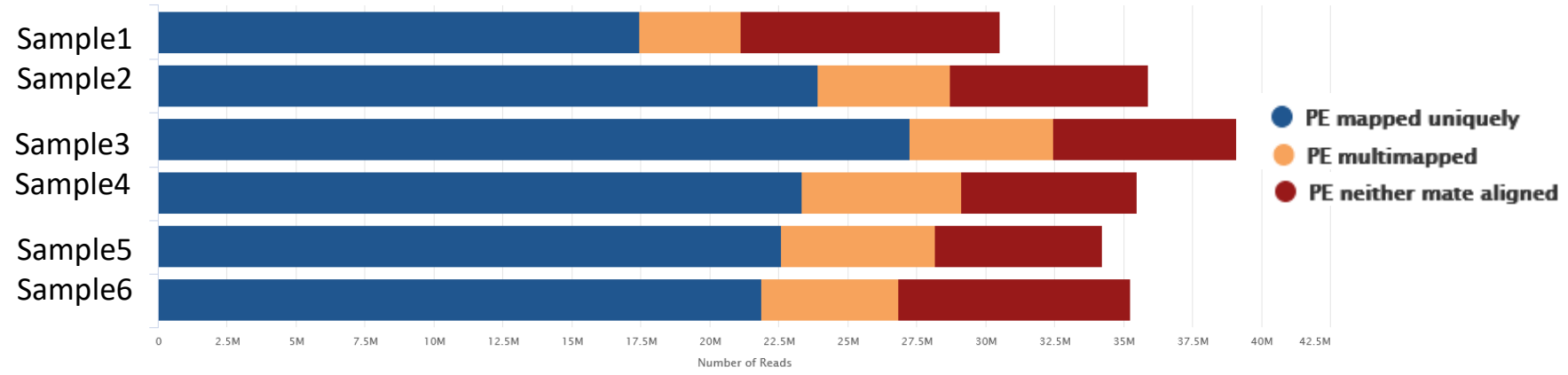


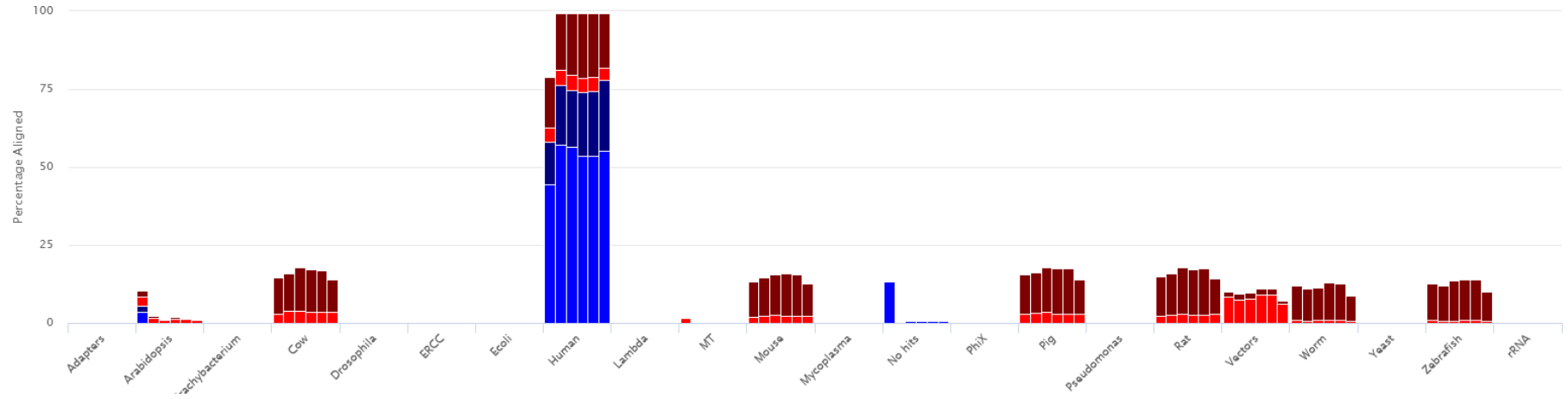
Aggregate and plot range of QC stats together

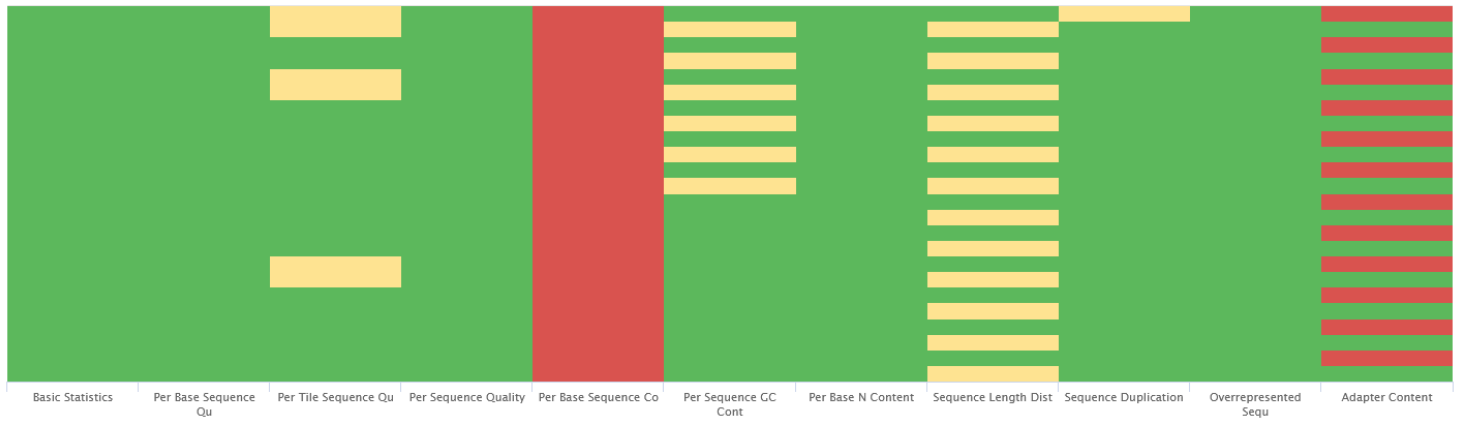
MultiQC currently has modules to support 128 different bioinformatics tools, https://multiqc.info/modules/

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





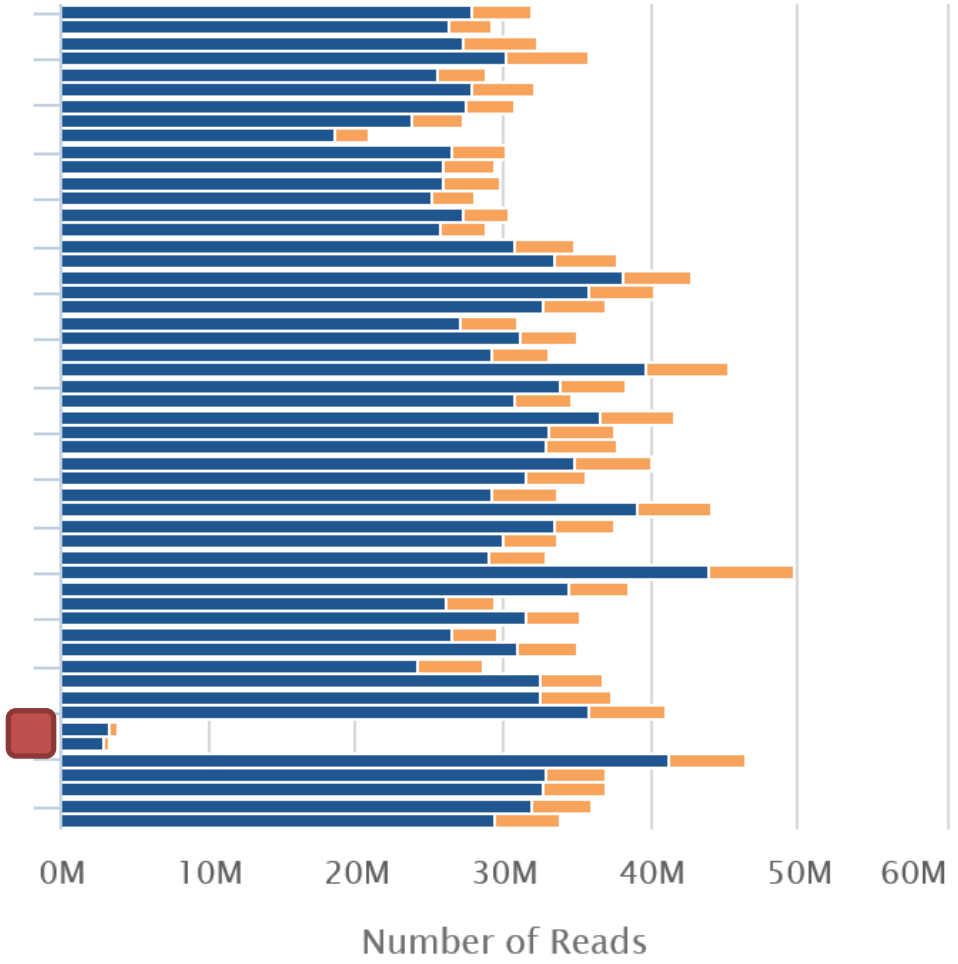


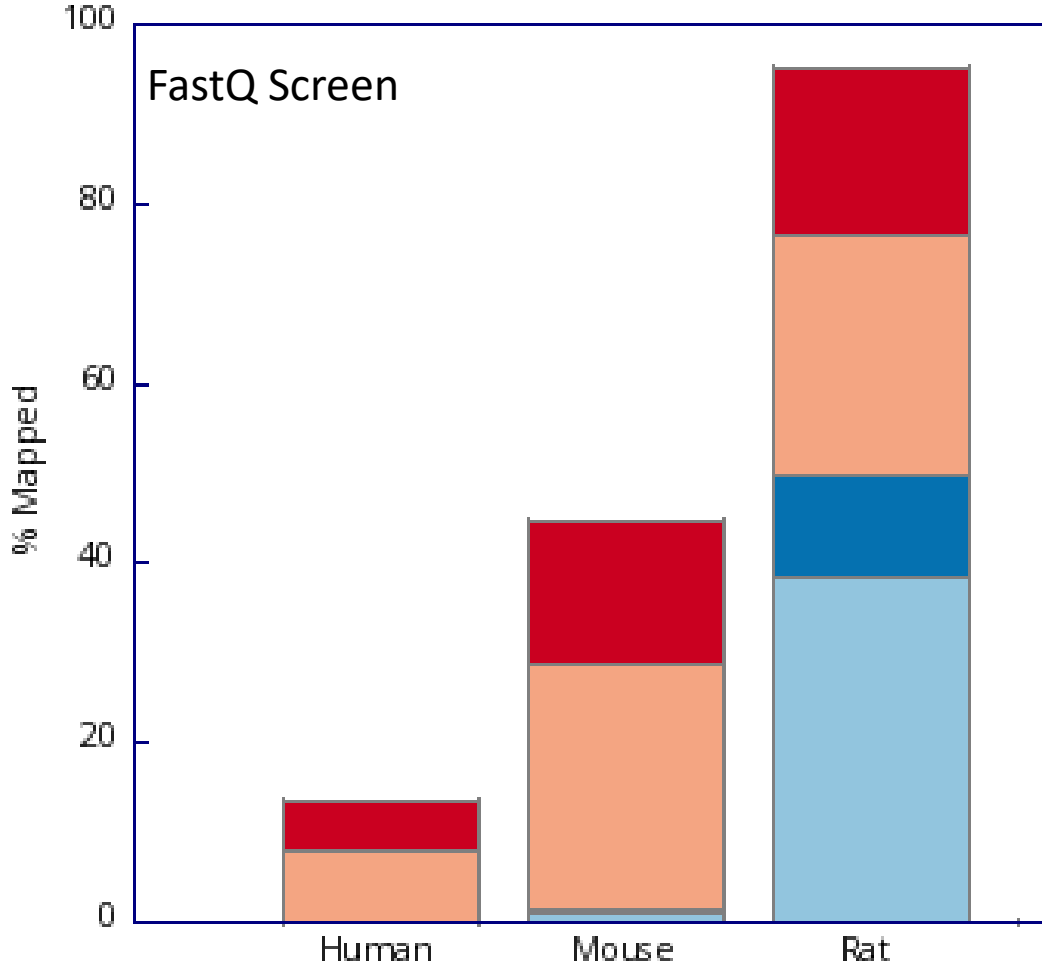
6 Samples, 24 files:

Paired end
Pre/post trimming

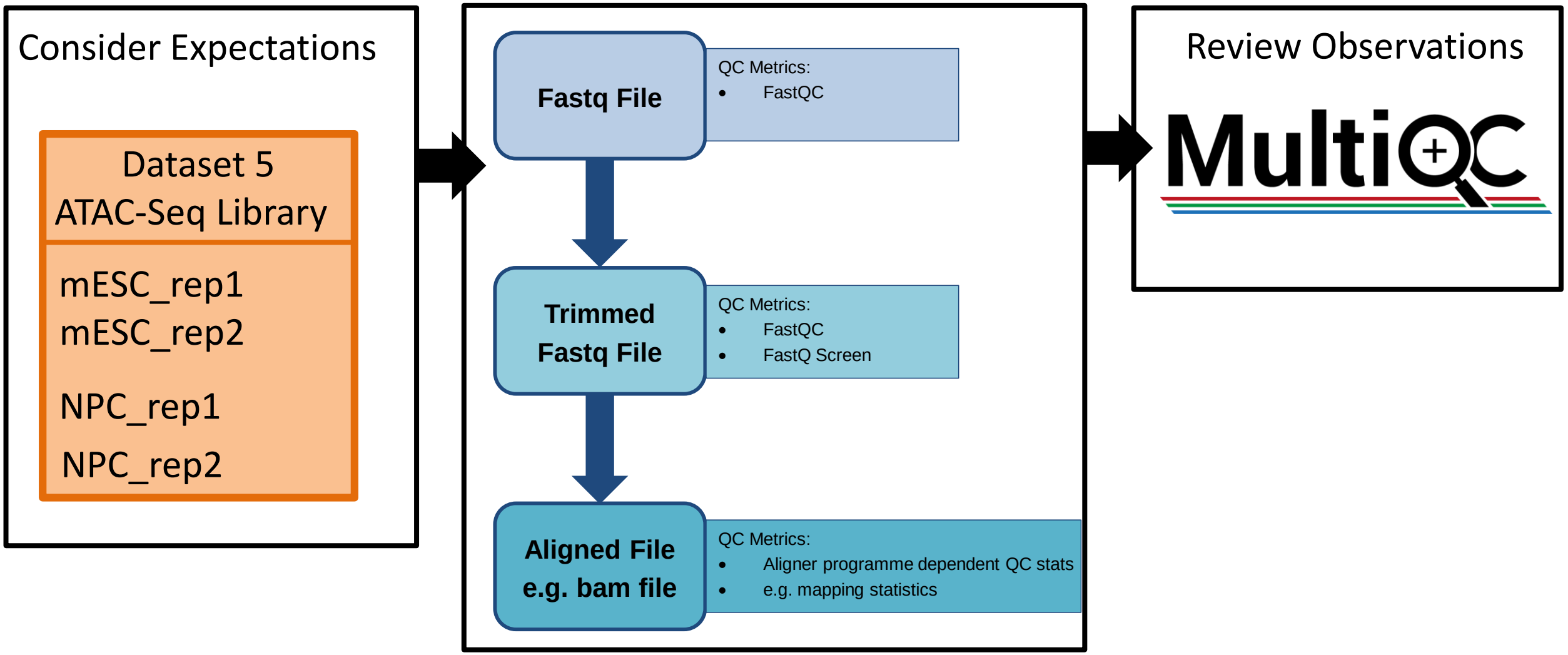
Aggregated Mapping Stats

Identify local QC problems by spotting samples that behave differently





Exercise Part 3: Putting it All Together with MultiQC

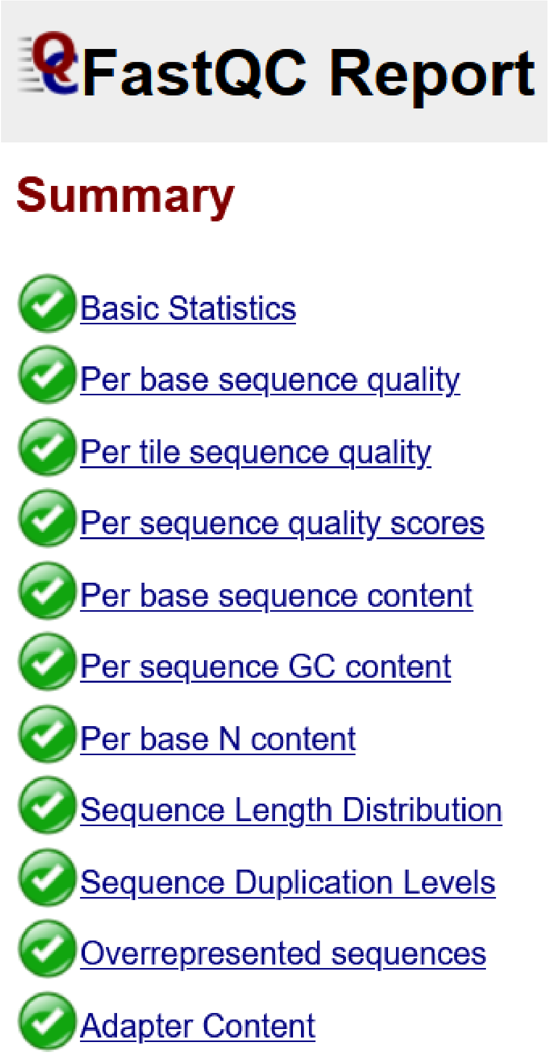


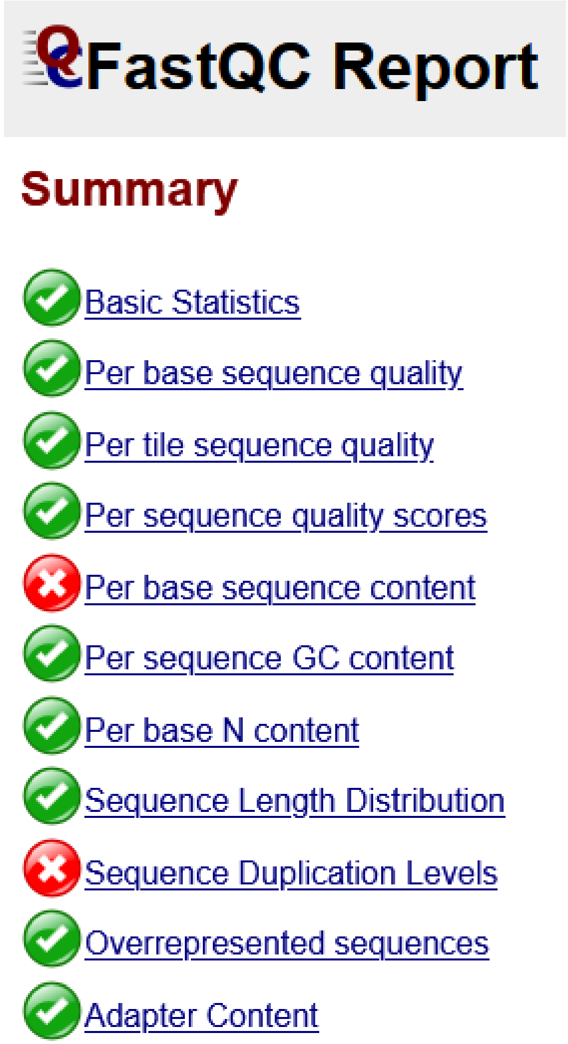
There is a QC problem, can you tell what it is and whether the data is usable?

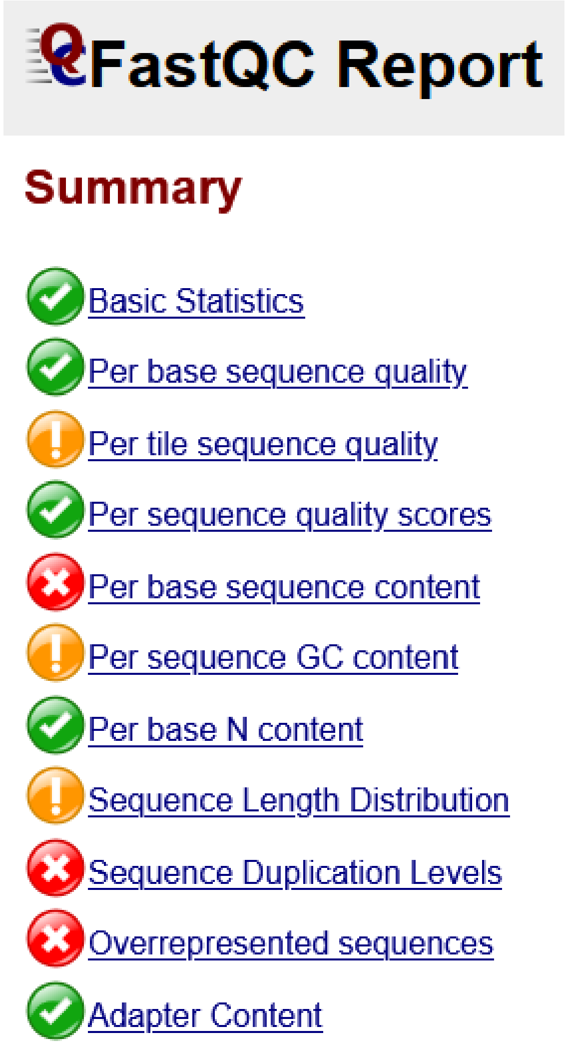
Final Thoughts

Babraham Bioinformatics

Expectations and Observations are Key







Can you tell if these libraries are any good?

QC In A Nut-Shell

Good Science ☺

Useful Links

https://sequencing.qcfail.com/

FastQC http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
FastQ Screen https://www.bioinformatics.babraham.ac.uk/projects/fastq_screen/
MultiQC https://multiqc.info/