25/09/2019

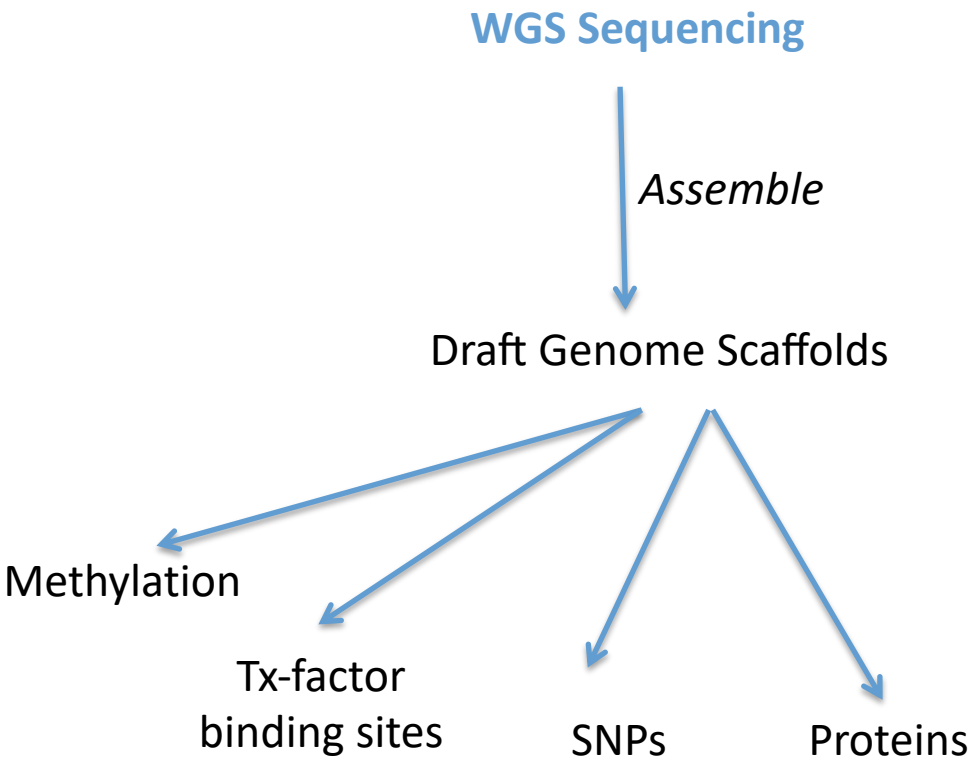
RNA Seq analysis

Cleaning

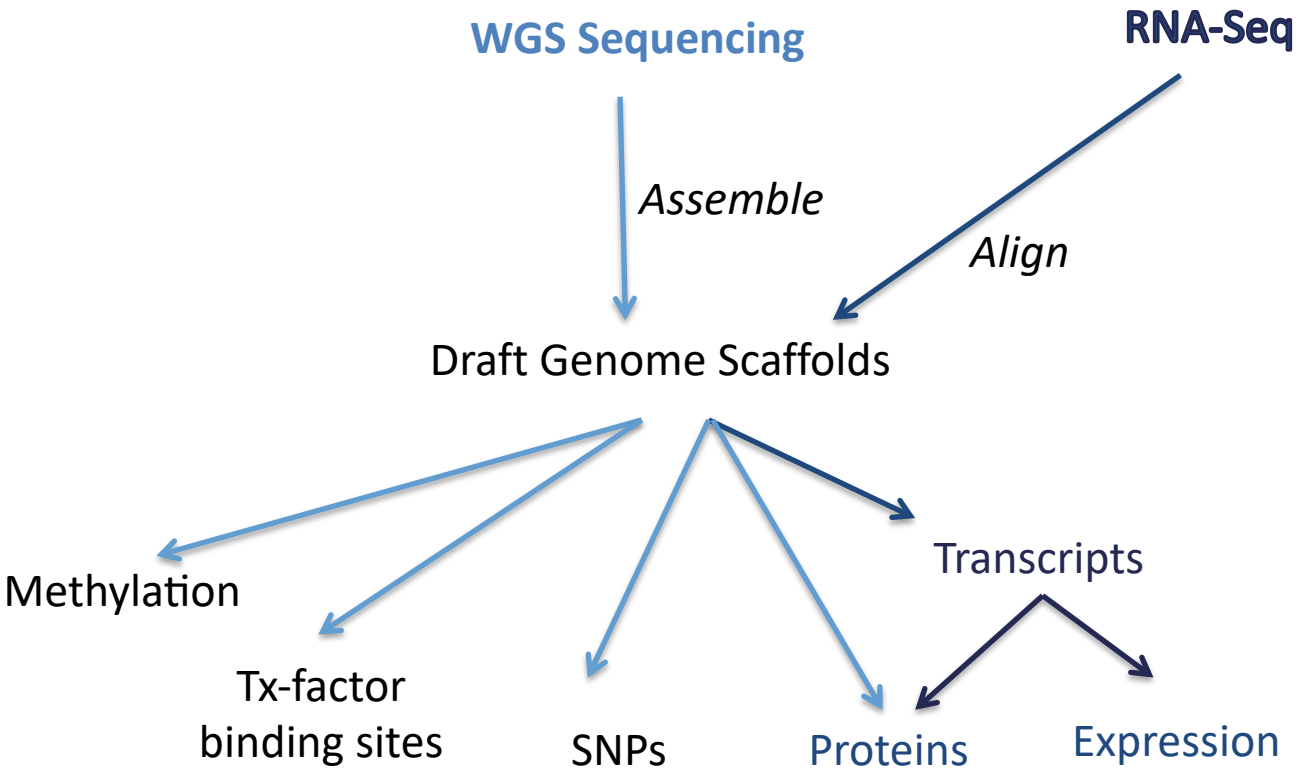
Platform ABiMS
SouthGreen

INTRODUCTION

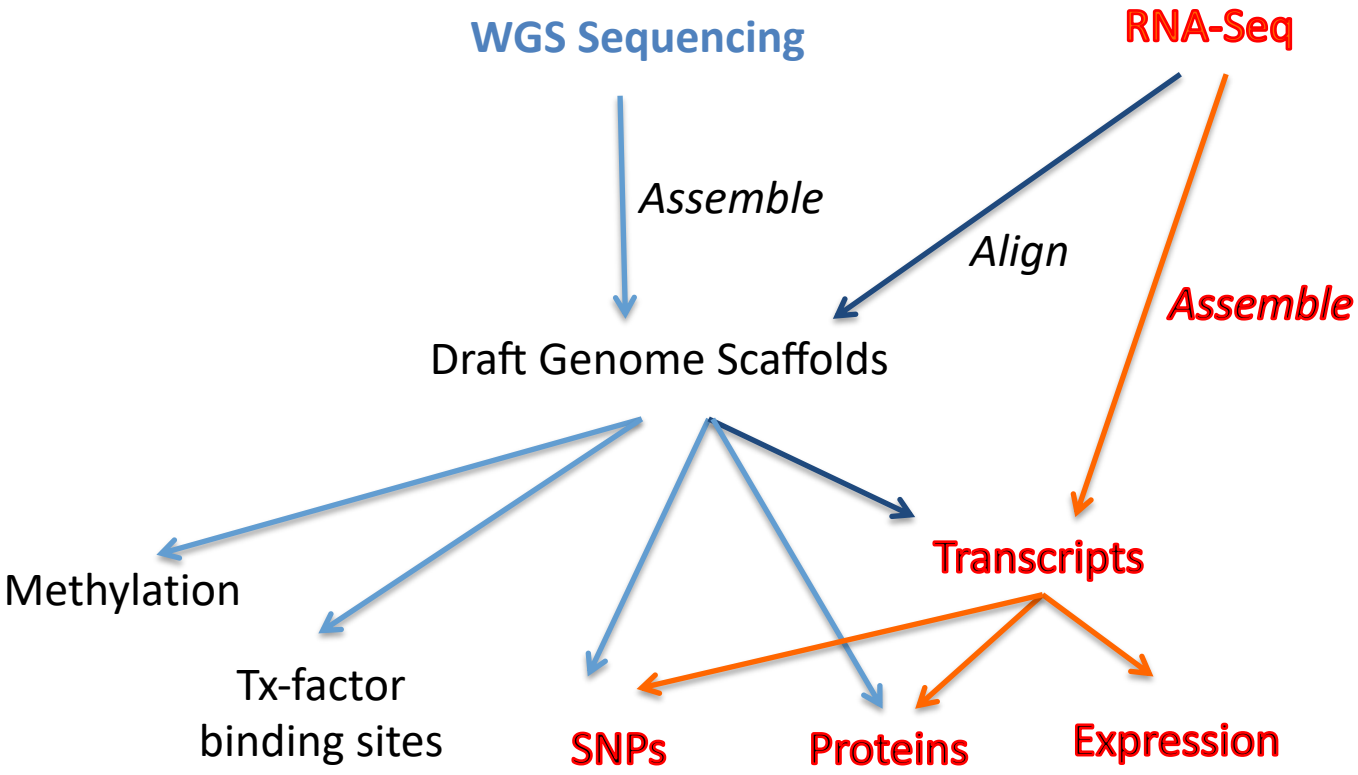
A Paradigm for Genomic Research



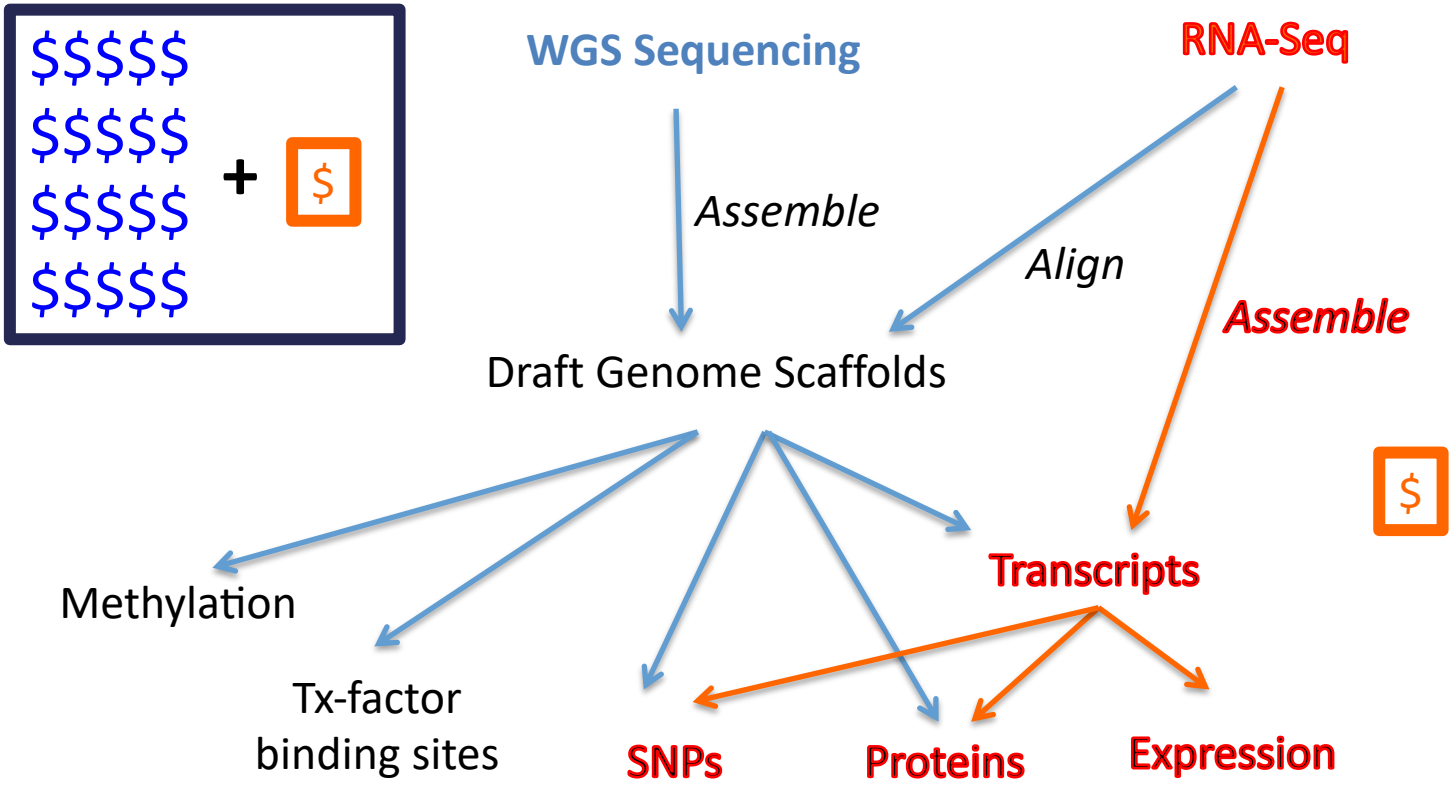
A Paradigm for Genomic Research



A *Maturing* Paradigm for Transcriptome Research



A ***Maturing*** Paradigm for Transcriptome Research



A **Maturing** Paradigm for Transcriptome Research

WGS Sequencing

RNA-Seq

Assemble

Align

Assemble

Draft Genome Scaffolds

Methylation

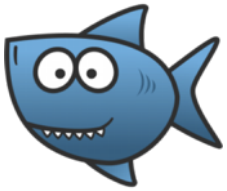
Tx-factor binding sites

Transcripts

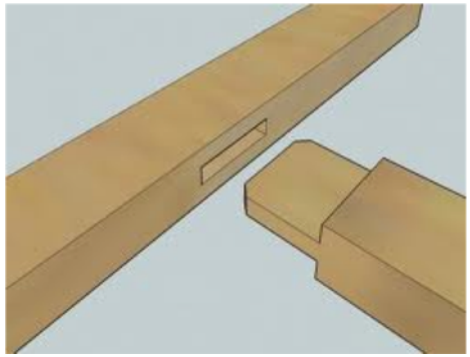
SNPs

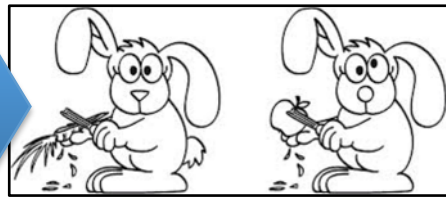
Proteins

Expression



RNA Seq de novo analysis workflow





Data Cleaning

Why do we care about cleaning ?

Why do we care about cleaning ?

RAW SEQUENCES

```
@D16GHACXX:8:2308:19491:200306 2:N:0:CGATGT
GACCCTATGAAGCTTTACTGTAACTTGAAATTGGTTTCGGGTTTTATTTG
+
7@?DB;BDD?FDHIIGIBHGFHF@FJHHB<FHHE48CGGGBBGCGGHIIG
@D16GHACXX:8:2308:19471:200307 2:N:0:CGATGT
AATCTGTTTTCCCTTGAATAGCCGCTCCTGTTAAAACCCTTGTAGTTTCT
+
@CCFFFFFHHGHHJIIHEIIIIJJIJGIJJJICHEHHJJIJJJJJIJJJJ
@D16GHACXX:8:2308:19410:200308 2:N:0:CGATGT
TATATATATATTAGTTCAGTAGTTTCATGTCTATTGCCAGCTTCGTGTTA
+
DGGIGIIJGHGGIJBHHHJJIIFCEADBEDCDDBDDD-9<A:AAD#####
@D16GHACXX:8:2308:19363:200321 2:N:0:CGATGT
CGTGCCAAGTTTGATTTCGTATTTATGTACCACATATTTCTATTTGAACA
+
BCBFFFFFHHHHHGJJJJJFHIJIIJJIJJJJJJJJJJJJJJJJJJJJJ
@D16GHACXX:8:2308:19258:200323 2:N:0:CGATGT
TGATTCGGGATAGGTGTTGGAAATGCGTGCATATTTTGGTTGGCGTAGCG
+
BCCFFFFFHDHFHJAEGGGGJJHIIJHEGIIIFGIJJJIDFHIJIGHFIG
@D16GHACXX:8:2308:19335:200326 2:N:0:CGATGT
GCCGCGAGGTTAAGGTTTTCACCGTCGGACGCGTTGCATGCCCGCTCAAC
+
```

Why do we care about cleaning ?

RAW SEQUENCES

```
@D16GHACXX:8:2308:19491:200306 2:N:0:CGATGT
GACCCTATGAAGCTTTACTGTAACTTGAAATTGGTTTCGGGTTTTATTTG
+
7@?DB;BDD?FDHIIGIBHGFHF@FJHHB<FHHE48CGGGBBGCGGHIIG
@D16GHACXX:8:2308:19471:200307 2:N:0:CGATGT
AATCTGTTTTCCCTTGAATAGCCGCTCCTGTTAAAACCCTTGTAGTTTCT
+
@CCFFFFFHHGHHJIIHEIIIIJJIJGIJJJICHEHHJJIJJJJJIJJJJ
@D16GHACXX:8:2308:19410:200308 2:N:0:CGATGT
TATATATATATTAGTTCAGTAGTTTCATGTCTATTGCCAGCTTCGTGTTA
+
DGGIGIIJGHGGIJBHHHJJIIFCEADBEDCDDBDDD-9<A:AAD#####
@D16GHACXX:8:2308:19363:200321 2:N:0:CGATGT
CGTGCCAAGTTTGATTTCGTATTTATGTACCACATATTTCTATTTGAACA
+
BCBFFFFFHHHHHGJJJJJFHIJIIJJIJJJJJJJJJJJJJJJJJJJJJ
@D16GHACXX:8:2308:19258:200323 2:N:0:CGATGT
TGATTCGGGATAGGTGTTGGAAATGCGTGCATATTTTGGTTGGCGTAGCG
+
BCCFFFFFHDHFHJAEGGGGJJHIIJHEGIIIFGIJJJIDFHIJIGHFIG
@D16GHACXX:8:2308:19335:200326 2:N:0:CGATGT
GCCGCGAGGTTAAGGTTTTCACCGTCGGACGCGTTGCATGCCCGCTCAAC
+
```

AMAZING
TRANSCRIPTOME !!!

Why do we care about cleaning ?

RAW SEQUENCES

```
@D16GHACXX:8:2308:19491:200306 2:N:0:CGATGT
GACCCTATGAAGCTTTACTGTAACTTGAAATTGGTTTCGGGTTTTATTTG
+
7@?DB;BDD?FDHIIGIBHGFHF@FJHHB<FHHE4BCGGGBBGCGGHIIG
@D16GHACXX:8:2308:19471:200307 2:N:0:CGATGT
AATCTGTTTTCCCTTGAATAGCCGCTCCTGTTAAAACCCTTGTAGTTTCT
+
@CCFFFFFHHGHHJIIHEIIIIJJIJGIJJJICHEHHJJIJJJJJIJJJJ
@D16GHACXX:8:2308:19410:200308 2:N:0:CGATGT
TATATATATATTAGTTCAGTAGTTTCATGTCTATTGCCAGCTTCGTGTTA
+
DGGIGIIJGHGGIJBHHHJJIIFCEADBEDCDDBDDD~9<A:AAD#####
@D16GHACXX:8:2308:19363:200321 2:N:0:CGATGT
CGTGCCAAGTTTGATTTCGTATTTATGTACCACATATTTCTATTTGAACA
+
BCBFFFFFHHHHHGJJJJJFHIJIIJJIJJJJJJJJJJJJJJJJJJJJJ
@D16GHACXX:8:2308:19258:200323 2:N:0:CGATGT
TGATTCGGGATAGGTGTTGGAAATGCGTGCATATTTTGGTTGGCGTAGCG
+
BCCFFFFFHDHFHJAEGGGGJJHIIJHEGIIIFGIJJJIDFHIJIGHFIG
@D16GHACXX:8:2308:19335:200326 2:N:0:CGATGT
GCCGCGAGGTTAAGGTTTTCACCGTCGGACGCGTTGCATGCCCGCTCAAC
+
```

NO !!

AMAZING
T[illegible]ISCRIP[illegible]ME !!!

Data Cleaning

- Unknown nucleotides
- Bad quality nucleotides
- Adaptors and primers sub-sequences
- Poly A/T tails
- Low complexity sequences
- rRNA sequences
- Contaminant sequences
- Short length sequences

But also:

- Removing singletons
- In-silico normalization
- Sequencing errors correction
- …

Bias should be corrected in reverse order of their generation

1. Sequencing biases (bad quality, unknowns)
2. Library preparation
   - Adaptors and primers sequences
   - Poly A/T tails
3. Biological sample (low complexity, rRNA, contaminants)

CNRS UPMC
Station Biologique
Roscoff

But first… What kind of data do we have ?

NGS sequences

- **Illumina**, 454 (Roche), Ion Torrent, Solid, PacBio, MinION, …
- Single, **Paired-end**,
- Sequences length: 25, 35, 50, **75, 100, 150**, 250, 500, 700, 800, … base pairs
- File format: **Fastq**, 2 files (.fasta + .qual), Colorspace, Fast5

1. Sequencing biases

- Unknown nucleotides (Ns)
- Bad quality nucleotides
- Hexamers biases (random priming) ? (Illumina. Now corrected ?)

- Why do we need to correct those ?
  - To remove a lot of sequencing errors (detrimental to the vast majority of assemblers)
  - Because most de-bruijn graph based assemblers can’t handle unknown nucleotides

Tools

- PRINSEQ2, FASTX Toolkit, Trimmomatic…
- http://prinseq.sourceforge.net/index.html
  - Perl software for PReprocessing and INformation of SEQuence data
  - Not the fastest, but very exhaustive
  - 2 versions. We use the command-line version: prinseq_lite.pl
- Now Trimmomatic

2. Adaptors & primers sequences

- Can be found in 3’ end if insert size is too short

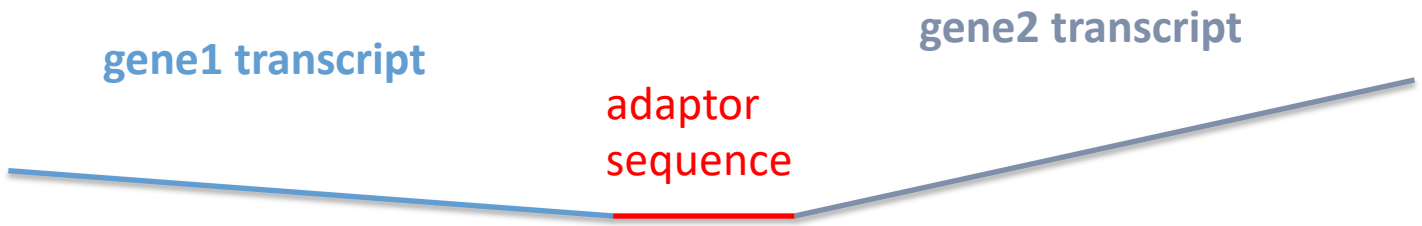
Normal case:
insert size > sequencing length

Abnormal case:
insert size < sequencing length

2. Adaptors & primers sequences

- Can be found in 3’ end if insert size is too short

- Why do we need to remove those ?
  - Because they can lead to “bridges” (links) between unrelated sequences (eg. 2 genes) and generate chimeras



Tools

- Trimmomatic, cutadapt, far, btrim, SeqTrim, TagCleaner, solexaQA, ...

- http://code.google.com/p/cutadapt/

- Trimming of adaptors sequences from NGS data

3. Poly A/T tails, low complexity reads

- Some poly A/T tails can be left during library preparation

- Poly A/T or low complexity sequences can also lead to “bridges” between unrelated sequences and generate chimeras

```
>
ACGTAGCTACTAGCTGACGATTCCCGTAGATCATCGGATAAAAAAAAAAAAAAAAAAAAAA
>
TTTTTTTTTTTTTTTTTTTTTTTTTTTTACTGCGTAGCACATGGCTATTATTTCGGCCATCAA
>
CGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCG
>
ATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGAT
```

Tools

PRINSEQ 2

- Trimming poly A/T tails
  - From 5'-end and 3'-end
  - w/ nucleotide nb >= 5
- Filtering low complexity sequences
  - Entropy < 70 (out of 100)
- Filtering short reads (< 50 nu)

4. Contaminations

- Most RNA-seq libraries comprise ribosomal RNA that you may want to remove

- Contaminations can also occur with foreign RNA/DNA (PhiX, Bacteria, …)

Tools

- SortMeRNA, riboPicker, DeconSeq
- Easy identification and removal of rRNA-like sequences
- For RNAseq and DNAseq

NGS Data basics : FASTQ format, SE data

FASTA format:

```
>61DFRAAXX100204:1:100:10494:3070
AAACAACAGGGCACATTGTCACTCTTGTATTTGAAAAACACTTTCCGGCCAT
```

FASTQ format:

```
@61DFRAAXX100204:1:100:10494:3070
AAACAACAGGGCACATTGTCACTCTTGTATTTGAAAAACACTTTCCGGCCAT
+
ACCCCCCCCCCCCCCCCCCCCCCCCCCCCCCBC?CCCCCCCCC@@CACCCCCA
```

Read

Quality values

Quality Scores

Sequencers can assign a “confidence” value per call based on how ambiguous the base call is

The sequencer will estimate the probability that a given base call is NOT correct (Erwing 1998)

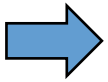
Ewing B, Green P (1998). "Base-calling of automated sequencer traces using phred. II. Error probabilities". Genome Res. 8 (3): 186–194. doi:10.1101/gr.8.3.186. PMID 9521922.

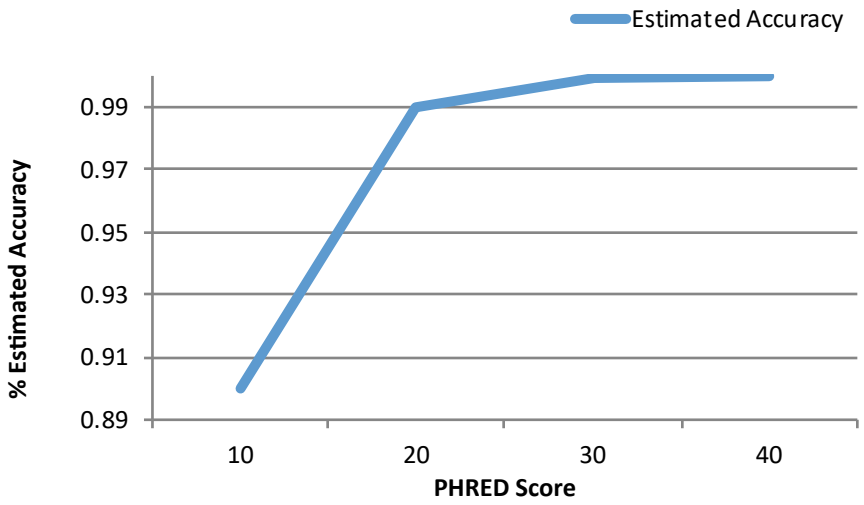
Quality Scores

PHRED Score is defined as
$q = -10 \times \log_{10}(p)$
(Erwing 1998)

P = probability call is not correct

P	-10*log10(p)	Est. Accuracy = 1-P
0.1	10	0.9
0.01	20	0.99
0.001	30	0.999
0.0001	40	0.9999





Ewing B, Green P (1998). "Base-calling of automated sequencer traces using phred. II. Error probabilities". Genome Res. 8 (3): 186–194. doi:10.1101/gr.8.3.186. PMID 9521922.

NGS Data basics : FASTQ format, SE data

```
@C060CACXX:1:2108:04435:81967
AGAGAATGGTAC
+
?@@DDDFFHFFH
@C060CACXX:1
GTGCATTCTTAT
+
CCCFFFFFHHH
@C060CACXX:1
CTCCTTTCCCAT
+
==>AA@?;?++@<=>AC>BB4,A7,,3?A>4+???A<@BBBA7):*111*?0?3:=?A>A
@C060CACXX:1:1305:16126:134486
ATCTATTCCTGAACAGGTCAATTTTAATGACTGATTCTTCAATCCGTGGTGGTCGAGATG
+
;>=AAAAABB+@=@C3+?++<,,33<=C<+?77+*:=7*1?A?=3?0:0=A<A3(<AA##
@C060CACXX:1:1308:04529:41884
ATTTGCCATCCCTGCATTGTGCGTGGTTTTCAGCAGCTTTTTAACAGGTGTTGTTTTTAT
+
@@<DDDEAFHHFDIGEEGGE9FGHHIA@FGIIGIIGIIJJJJIIIIEHDDBFFBCGHGII
@C060CACXX:1:2202:06955:98871
CTGAGATCTTCTTTAATTTCTTTCTTCAGGGACTTGAAGTTTTTATCATACAGATCTTTC
+
BCCDFFFFHHHHHJJJJJJJJJJIJJJIJJJIIJJJGIIFIJJJJJJJJJJIJJJJJJIJ
@C060CACXX:1:1105:15276:91210
TAGGAATCAGCGTGAGCTGTATTCTGACGGAGAATCTCTTCTGGTACCAGAAGGTTTGGA
+
?7?>BDD:C3:02@+AE2<3AEEDF++<))?D?DD4BDB9DDIIDBDD49DB;8.48@5@
@C060CACXX:1:1301:16367:35650
CGCTCTCCAAGCTCCTCCTCCTGGCCCTCAGCTTCTGTGGCTTTCTGGTCTTCACCAACC
+
==<;A8A7+?A7?CB9AAACA++++2<?)5@3*1????*0:?=>**00/*9AA43))==A
@C060CACXX:1:1205:17708:111304
CTGGTAGTAAAGTAGCTGCATGGAGTTCACCTGCAGTTCGTGCTGCTTGGCGCCGACCCA
+
?@@DABB=CC<,C:ACG4CFE4@E;+<?+<C3CDCFF?91::)0:?<93BG(7;;''58(
@C060CACXX:1:1208:13509:106734
GCTTTGTGGTCTTCACCAACCTTTCTCTGCAGAACAACACCATAGGCACCTATCAGCTGG
+
@CCFFFDFHFHHHJIJIJJJJJJJJJIJIIJJJJIIJJJEHIIJIGIIJJJJJJJIHJG
@C060CACXX:1:1101:03034:113094
ATTCTCCGTCAGAATACAGCTCACGCTGATTCCTATTACTGTAGGTGTAATCCTAAATTC
+
@CCFFFFFHHHFHIIIJIHIIIJJIIHIJEIJJGJBHGIGGDDFCDHEFFCIBGICHIIG
.
.
.
.
```


```
@C060CACXX:1:1305:16126:134486
ATCTATTCCTGAACAGGTCAATTTTAATGACTGATTCTTCAATCCGTGGTGGTCGAGATG
+
;>=AAAAABB+@=@C3+?++<,,33<=C<+?77+*:=7*1?A?=3?0:0=A<A3(<AA##
```


Standard format is 4 lines per read:

1. Unique read identifier.
2. Read sequence.
3. Either read identifier again or a place holder like “+”.
4. Phred-like base quality scores [Q:0-40].

$Q = -10 \log_{10}(e)$, where e is the estimated probability of a wrong base. So the probability that a base call is an error is:

* 0.01% if Q=40
* 0.1% if Q=30
* 1% if Q=20
* 10% if Q=10

NGS Data basics : FASTQ format, PE data

Quality Score Encodings

–Why not just have numbers?

```
@CCRI0219:135:D243EACXX:1:1101:1682:1955 1:N:0:ACAGTG
CGTTCAGT…
+
3131303537373739…
```

Quality symbols to the rescue

From National Center for Genome Analysis Support: http://ncgas.org

Quality Score Encodings

- Letters are represented deep down in the computer as numbers

- The quality score + a constant number (33 or 64, usually) is the number, which is converted to the quality symbol using ASCII

From National Center for Genome Analysis Support: http://ncgas.org

ASCII Table

Dec	Hx	Oct	Char
0	0	000	NUL (null)
1	1	001	SOH (start of heading)
2	2	002	STX (start of text)
3	3	003	ETX (end of text)
4	4	004	EOT (end of transmission)
5	5	005	ENQ (enquiry)
6	6	006	ACK (acknowledge)
7	7	007	BEL (bell)
8	8	010	BS (backspace)
9	9	011	TAB (horizontal tab)
10	A	012	LF (NL line feed, new line)
11	B	013	VT (vertical tab)
12	C	014	FF (NP form feed, new page)
13	D	015	CR (carriage return)
14	E	016	SO (shift out)
15	F	017	SI (shift in)
16	10	020	DLE (data link escape)
17	11	021	DC1 (device control 1)
18	12	022	DC2 (device control 2)
19	13	023	DC3 (device control 3)
20	14	024	DC4 (device control 4)
21	15	025	NAK (negative acknowledge)
22	16	026	SYN (synchronous idle)
23	17	027	ETB (end of trans. block)
24	18	030	CAN (cancel)
25	19	031	EM (end of medium)
26	1A	032	SUB (substitute)
27	1B	033	ESC (escape)
28	1C	034	FS (file separator)
29	1D	035	GS (group separator)
30	1E	036	RS (record separator)
31	1F	037	US (unit separator)

Dec	Hx	Oct	Html	Chr
32	20	040		Space
33	21	041	!	!
34	22	042	"	"
35	23	043	#	#
36	24	044	$	$
37	25	045	%	%
38	26	046	&	&
39	27	047	'	'
40	28	050	(	(
41	29	051	)	)
42	2A	052	*	*
43	2B	053	+	+
44	2C	054	,	,
45	2D	055	-	-
46	2E	056	.	.
47	2F	057	/	/
48	30	060	0	0
49	31	061	1	1
50	32	062	2	2
51	33	063	3	3
52	34	064	4	4
53	35	065	5	5
54	36	066	6	6
55	37	067	7	7
56	38	070	8	8
57	39	071	9	9
58	3A	072	:	:
59	3B	073	;	;
60	3C	074	<	<
61	3D	075	=	=
62	3E	076	>	>
63	3F	077	?	?

Dec	Hx	Oct	Html	Chr
64	40	100	@	@
65	41	101	A	A
66	42	102	B	B
67	43	103	C	C
68	44	104	D	D
69	45	105	E	E
70	46	106	F	F
71	47	107	G	G
72	48	110	H	H
73	49	111	I	I
74	4A	112	J	J
75	4B	113	K	K
76	4C	114	L	L
77	4D	115	M	M
78	4E	116	N	N
79	4F	117	O	O
80	50	120	P	P
81	51	121	Q	Q
82	52	122	R	R
83	53	123	S	S
84	54	124	T	T
85	55	125	U	U
86	56	126	V	V
87	57	127	W	W
88	58	130	X	X
89	59	131	Y	Y
90	5A	132	Z	Z
91	5B	133	[	[
92	5C	134	\	\
93	5D	135	]	]
94	5E	136	^	^
95	5F	137	_	_

Dec	Hx	Oct	Html	Chr
96	60	140	`	`
97	61	141	a	a
98	62	142	b	b
99	63	143	c	c
100	64	144	d	d
101	65	145	e	e
102	66	146	f	f
103	67	147	g	g
104	68	150	h	h
105	69	151	i	i
106	6A	152	j	j
107	6B	153	k	k
108	6C	154	l	l
109	6D	155	m	m
110	6E	156	n	n
111	6F	157	o	o
112	70	160	p	p
113	71	161	q	q
114	72	162	r	r
115	73	163	s	s
116	74	164	t	t
117	75	165	u	u
118	76	166	v	v
119	77	167	w	w
120	78	170	x	x
121	79	171	y	y
122	7A	172	z	z
123	7B	173	{	{
124	7C	174			|
125	7D	175	}	}
126	7E	176	~	~
127	7F	177	&#127;	DEL

Source: www.LookupTables.com

Phred score 20

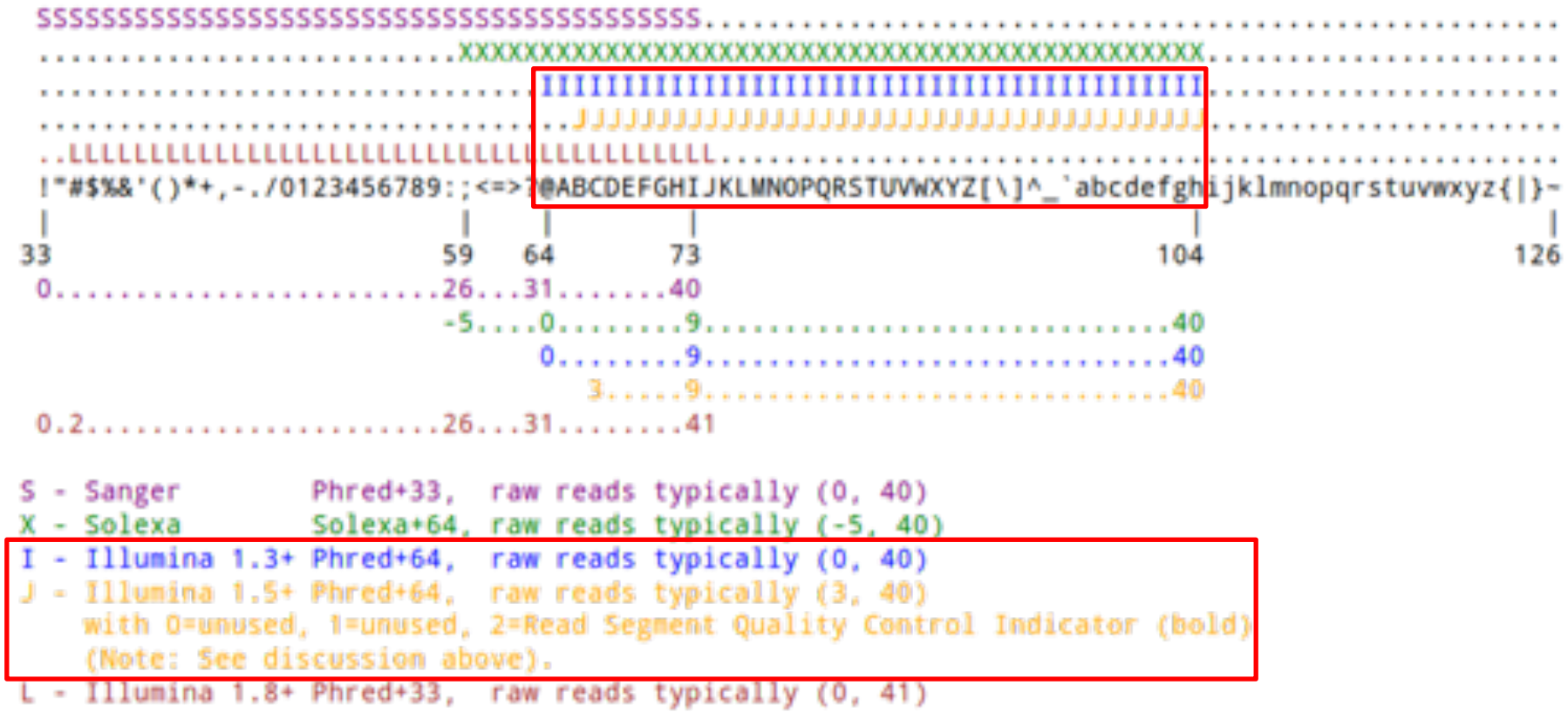
20+33 = 53 = 5

20+64 = 84= T

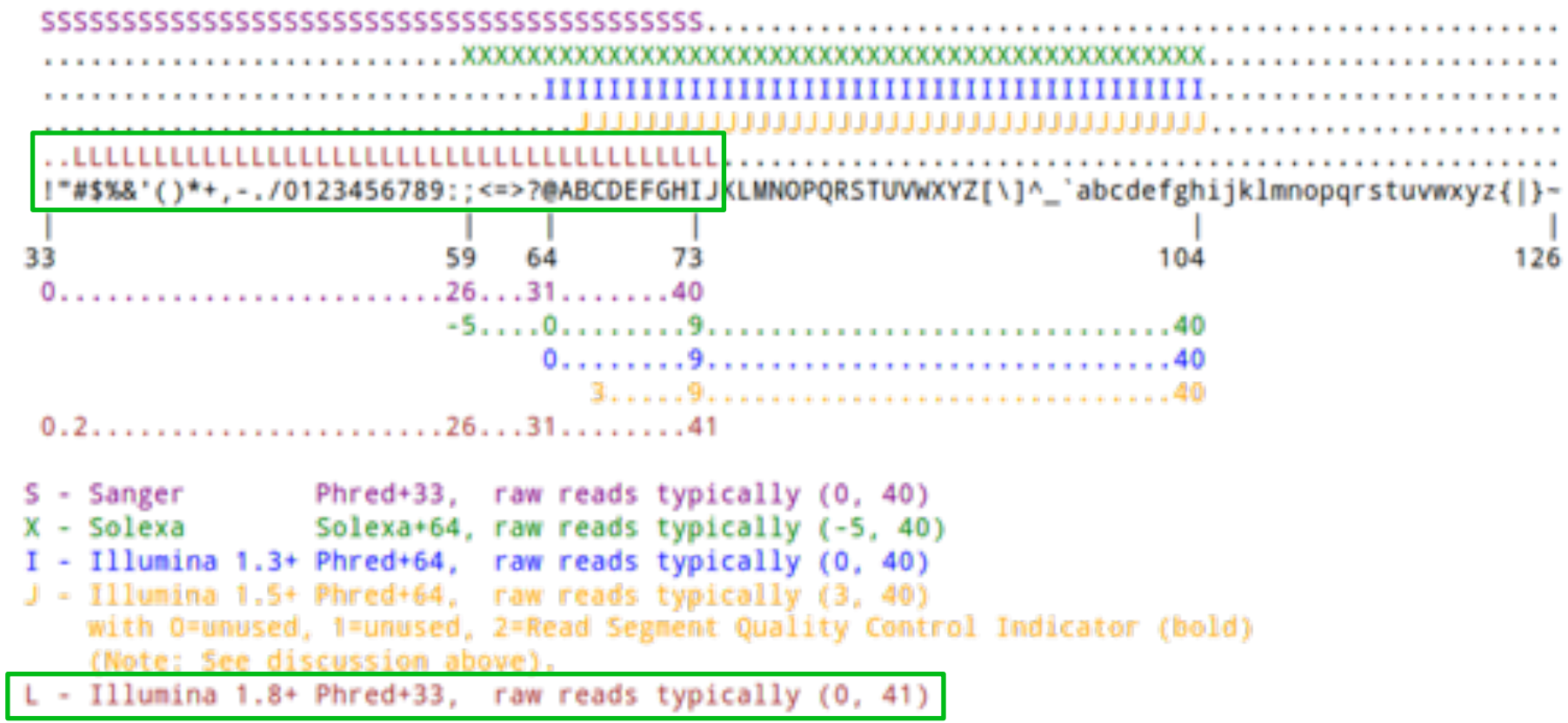
From National Center for Genome Analysis Support: http://ncgas.org

FASTQ quality encoding

FASTQ quality encoding



FASTQ quality encoding



FASTQ quality encoding

```
@MERCURE_0127:7:1101:1162:2110#CTTGTA/1
TAATAACCCATTAAATACCAATCCAGAAAGCAGCGTGGGTTCAATTCCCAAGATCGGAAG
+MERCURE_0127:7:1101:1162:2110#CTTGTA/1
bbbeeeeeggggiiiihfgfffgihhiihfhfcab``aKZ^]b]]_]`b^^_b``[a__
@MERCURE_0127:7:1101:1182:2111#CTTGTA/1
ACTTACCTCCTGACCCCCCAAAGCCTACTCTCCACTTGCCTGGATGAGCGCAGCTCCAAC
+MERCURE_0127:7:1101:1182:2111#CTTGTA/1
bbbeeeeeeggggghiihhihiiiiiigaaabb`b`b]`b`b^`T]T]bc_aOEETR___BB
```

```
@HWI-ST227:191:D16GHACXX:8:2308:20216:200677 1:N:0:CGATGT
GCCATTGATGGTGGTGTGTGTTTGGTTGGTTGTTGGATGGGGGTGGGGGGTGTGGTGCG
+
++1BD2222==2A+2+2<3CFFIIA<E)1?C:)0?)*0*0?D@###############
@HWI-ST227:191:D16GHACXX:8:2308:20300:200513 1:N:0:CGATGT
CGTTGTTCCTCGCGACGAGAAAAGTGCAGACGGTTTAGGGATCATCGGTATTTCGTGCG
+
?@?ADDDDDBCF@HIEIAGDHB;DDBHGIIEBG:FBDGHBD@CA+9:>098595?CCC<
```

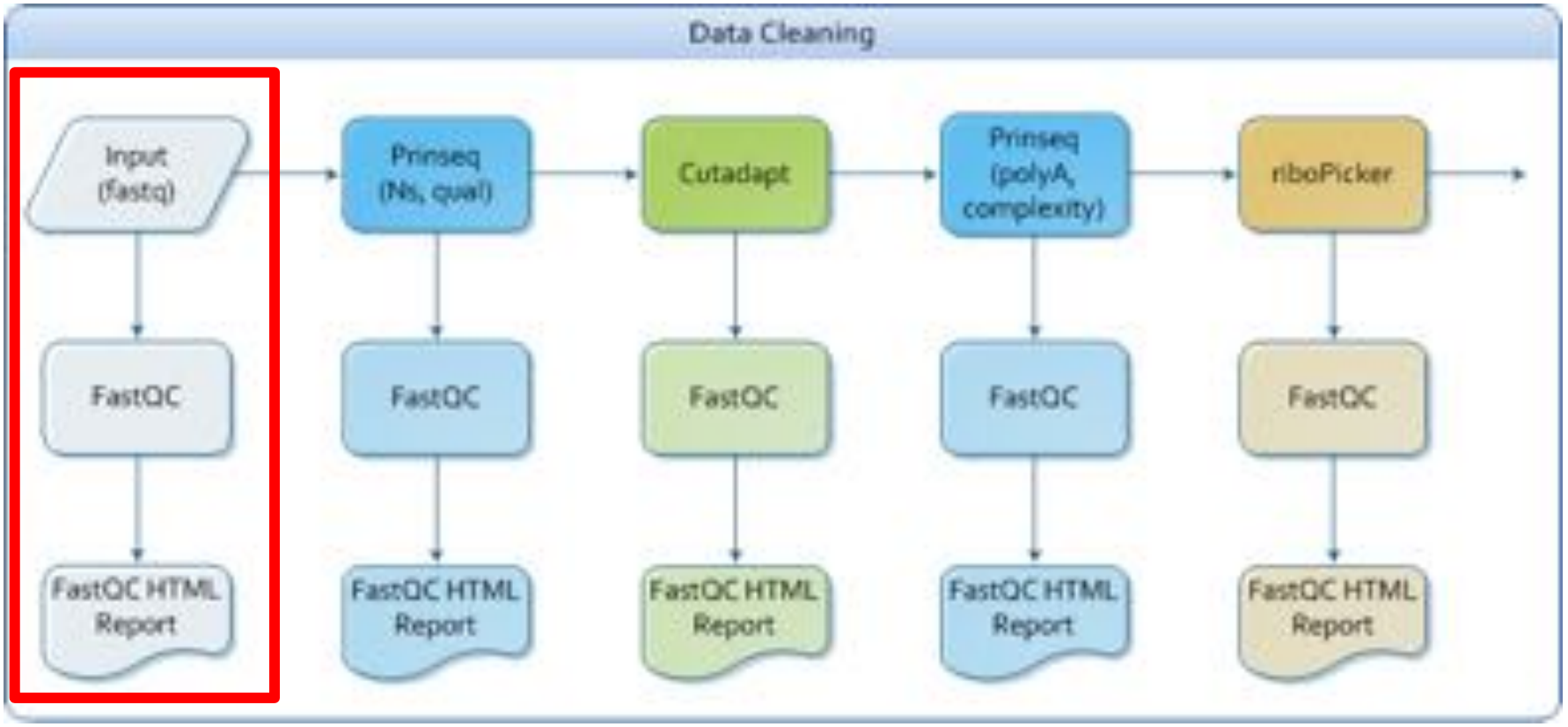
FASTQ quality encoding

```
@MERCURE_0127:7:1101:1162:2110#CTTGTA/1
TAATAACCCATTAAATACCAATCCAGAAAGCAGCGTGGGTTCAATTCCCAAGATCGGAAG
+MERCURE_0127:7:1101:1162:2110#CTTGTA/1
bbbeeeeeggggiiiihfgfffgihhiihfhfcab``aKZ^]b]]_]`b^^_b``[a__
@MERCURE_0127:7:1101:1182:2111#CTTGTA/1
ACTTACCTCCTGACCCCCCAAAGCCTACTCTCCACTTGCCTGGATGAGCGCAGCTCCAAC
+MERCURE_0127:7:1101:1182:2111#CTTGTA/1
bbbeeeeeeggggghiihhihiiiiiigaaabb`b`b]`b`b^`T]T]bc_aOEETR___BB
```

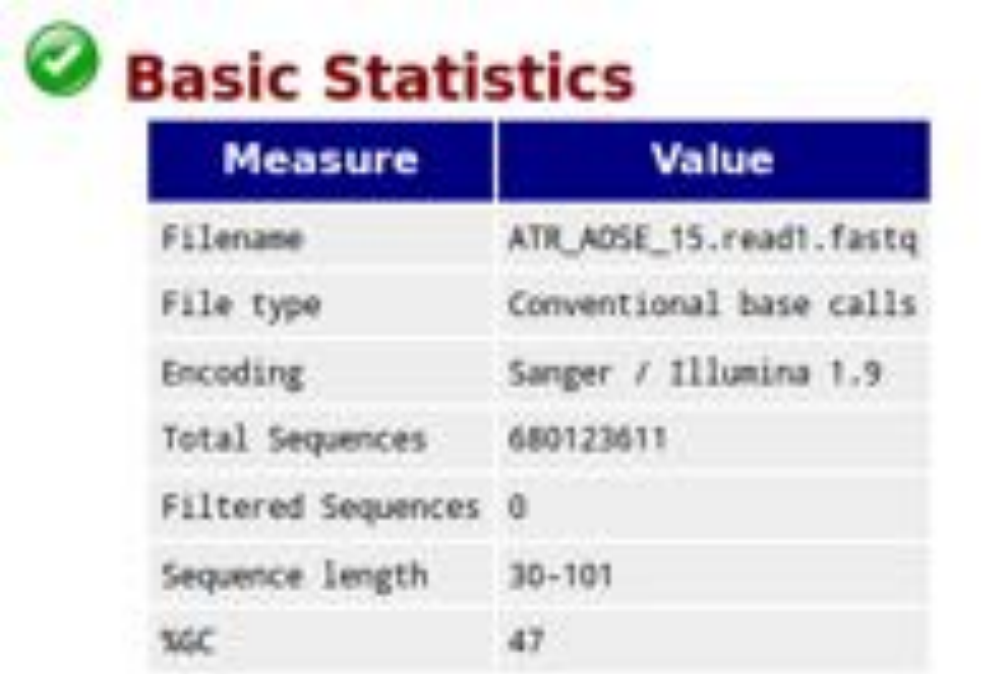
Phred+64

```
@HWI-ST227:191:D16GHACXX:8:2308:20216:200677 1:N:0:CGATGT
GCCATTGATGGTGGTGTGTGTTTGGTTGGTTGTTGGATGGGGGTGGGGGGTGTGGTGCG
+
++1BD2222==2A+2+2<3CFFIIA<E)1?C:)0?)*0*0?D@###############
@HWI-ST227:191:D16GHACXX:8:2308:20300:200513 1:N:0:CGATGT
CGTTGTTCCTCGCGACGAGAAAAGTGCAGACGGTTTAGGGATCATCGGTATTTCGTGCG
+
?@?ADDDDDBCF@HIEIAGDHB;DDBHGIIEBG:FBDGHBD@CA+9:>098595?CCC<
```

Phred+33


Data Cleaning
Input (fastq)
Prinseq (Ns, qual)
Cutadapt
Prinseq (polyA, complexity)
riboPicker
FastQC
FastQC
FastQC
FastQC
FastQC
FastQC HTML Report
FastQC HTML Report
FastQC HTML Report
FastQC HTML Report
FastQC HTML Report

FastQC : Basic Statistics



Basic Statistics

Measure	Value
Filename	ATR_AOSE_15.read1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	680123611
Filtered Sequences	0
Sequence length	30-101
%GC	47

FastQC : Basic Statistics

```
@MERCURE_0127:7:1101:1162:2110#CTTGTA/1
TAATAACCCATTAAATACCAATCCAGAAAGCAGCGTGGGTTCAATTCCCAAGATCGGAAG
+MERCURE_0127:7:1101:1162:2110#CTTGTA/1
bbbeeeeeggggggiiiihfgfffgihhiihfhfcab``aKZ^]b]]_]`b^^_b``[a__
@MERCURE_0127:7:1101:1182:2111#CTTGTA/1
ACTTACCTCCTGACCCCCCAAAGCCTACTCTCCACTTGCCTGGATGAGCGCAGCTCCAAC
+MERCURE_0127:7:1101:1182:2111#CTTGTA/1
bbbeeeeeggggghiihhihiiiiiigaaabb`b`b]`b`b^``T]T]bc_aOEETR___BB
```

Basic Statistics

Measure	Value
Filename	AWA_COSW_7_1_D0BF9ACXX.IND12.fastq
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	120620512
Filtered Sequences	0
Sequence length	101
%GC	45

Phred+64

```
@HWI-ST227:191:D16GHACXX:8:2308:20216:200677 1:N:0:CGATGT
GCCATTGATGGTGGTGTGTGTTTGGTTGGTTGTTGGATGGGGGTGGGGGGTGTGGTGCG
+
++1BD2222==2A+2+2<3CFFIIA<E)1?C:)0?)*0*0?D@###############
@HWI-ST227:191:D16GHACXX:8:2308:20300:200513 1:N:0:CGATGT
CGTTGTTCCTCGCGACGAGAAAAGTGCAGACGGTTTAGGGATCATCGGTATTTCGTGCG
+
?@?ADDDDDBCF@HIEIAGDHB;DDBHGIIEBG:FBDGHBD@CA+9:>098595?CCC<
```

Basic Statistics

Measure	Value
Filename	ATR_AOSE_15.read1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	680123611
Filtered Sequences	0
Sequence length	30-101
%GC	47

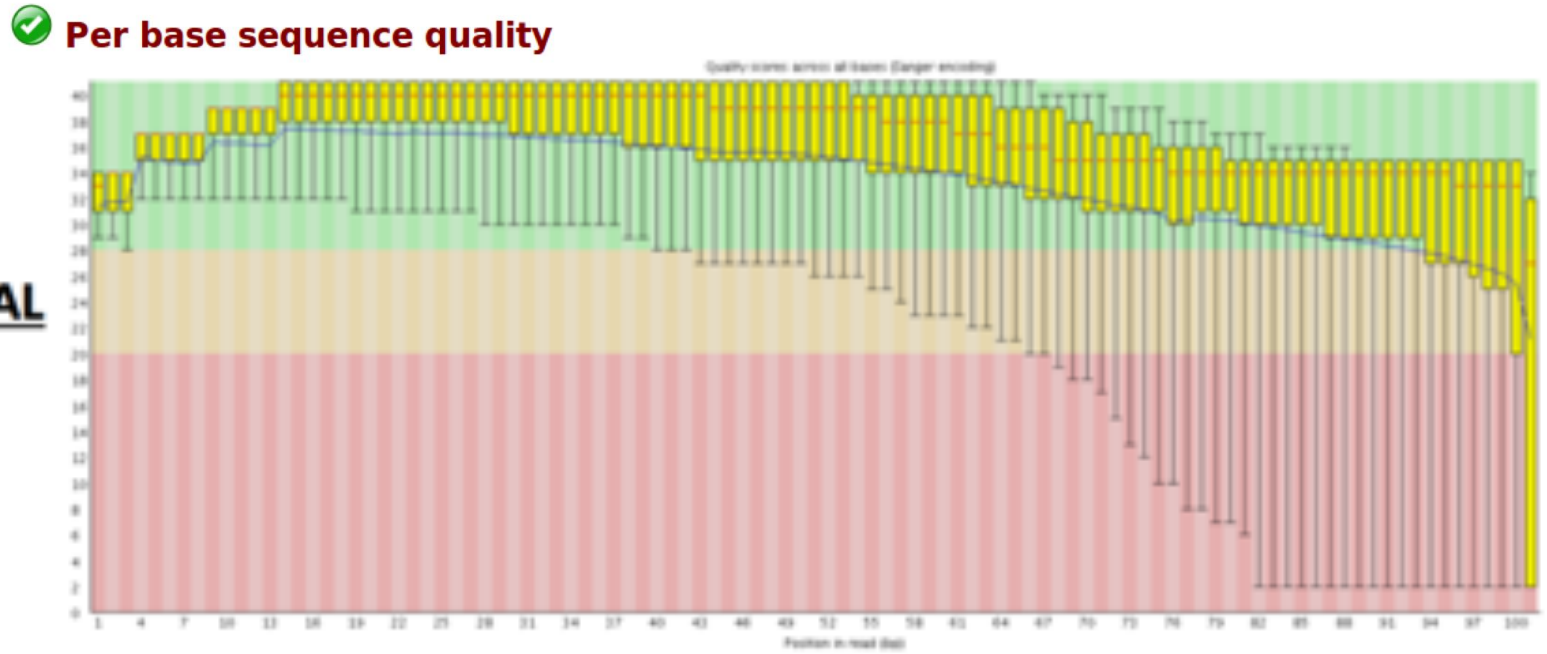
Phred+33

FastQC : Per base sequence quality

This plot shows the base quality score distribution for all reads in a lane, with each read position considered independently.

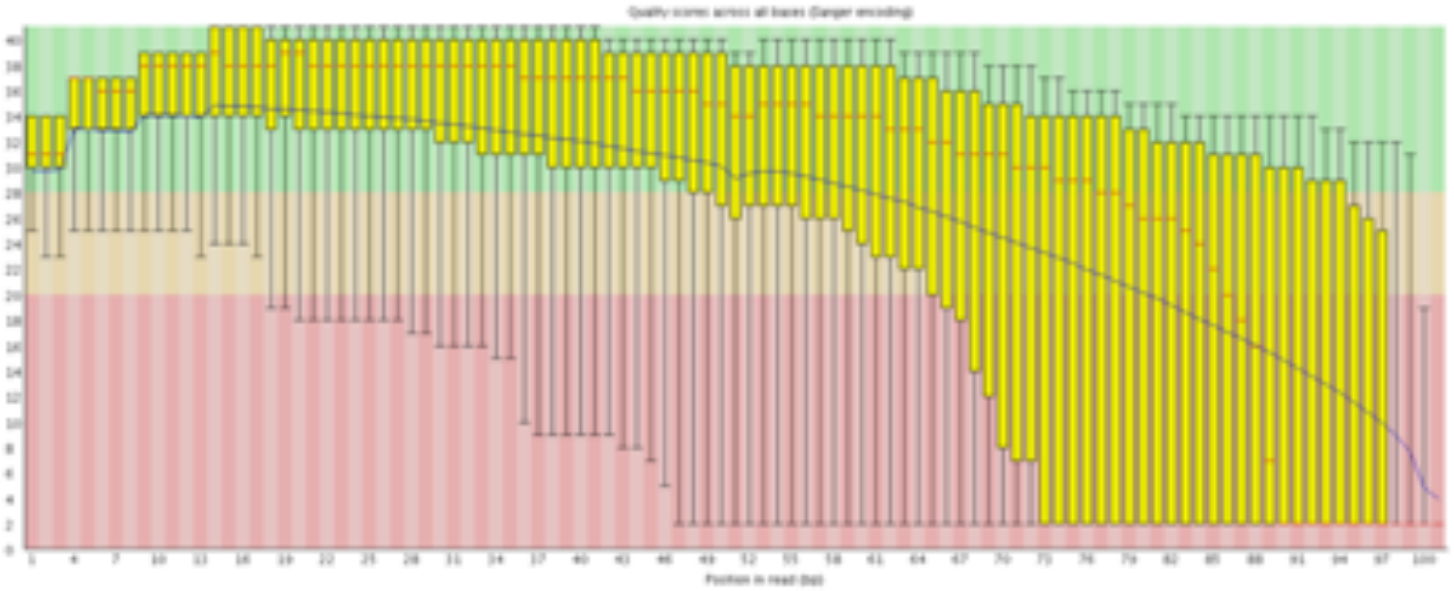
- x-axis = position in read (bp)
- y-axis = Phred-like base quality score [pink=0-20, tan=20-30, green=30-40]
- red bar = median score, blue line = mean score
- yellow box = 25th to 75th percentile, black whiskers = 10th to 90th percentile

GOOD/NORMAL LANE



FastQC : Per base sequence quality

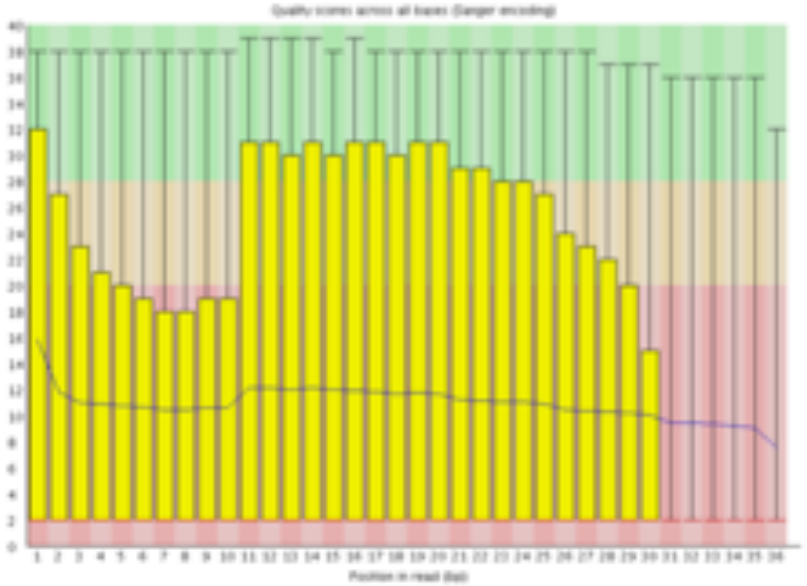
SALVAGEABLE LANE



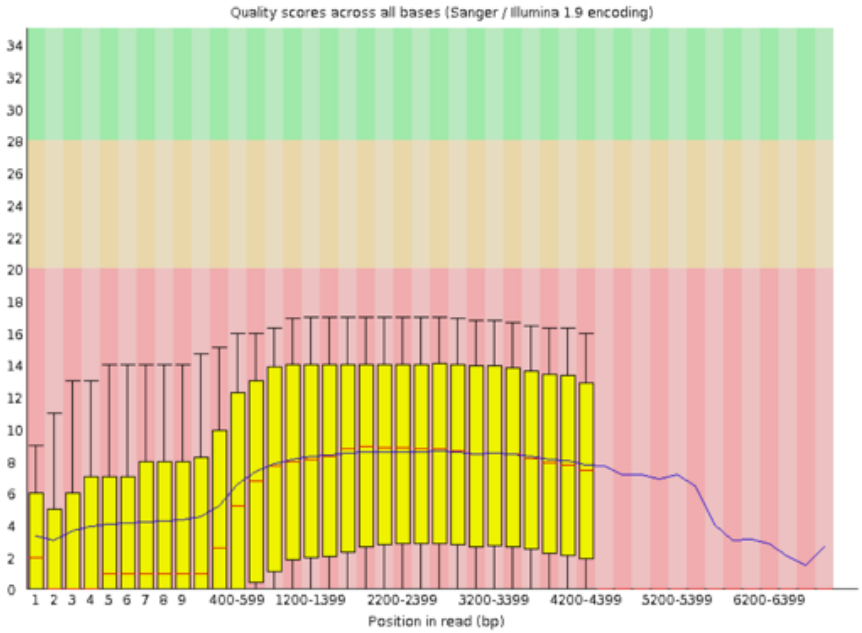
FAILED LANE

FastQC : Per base sequence quality

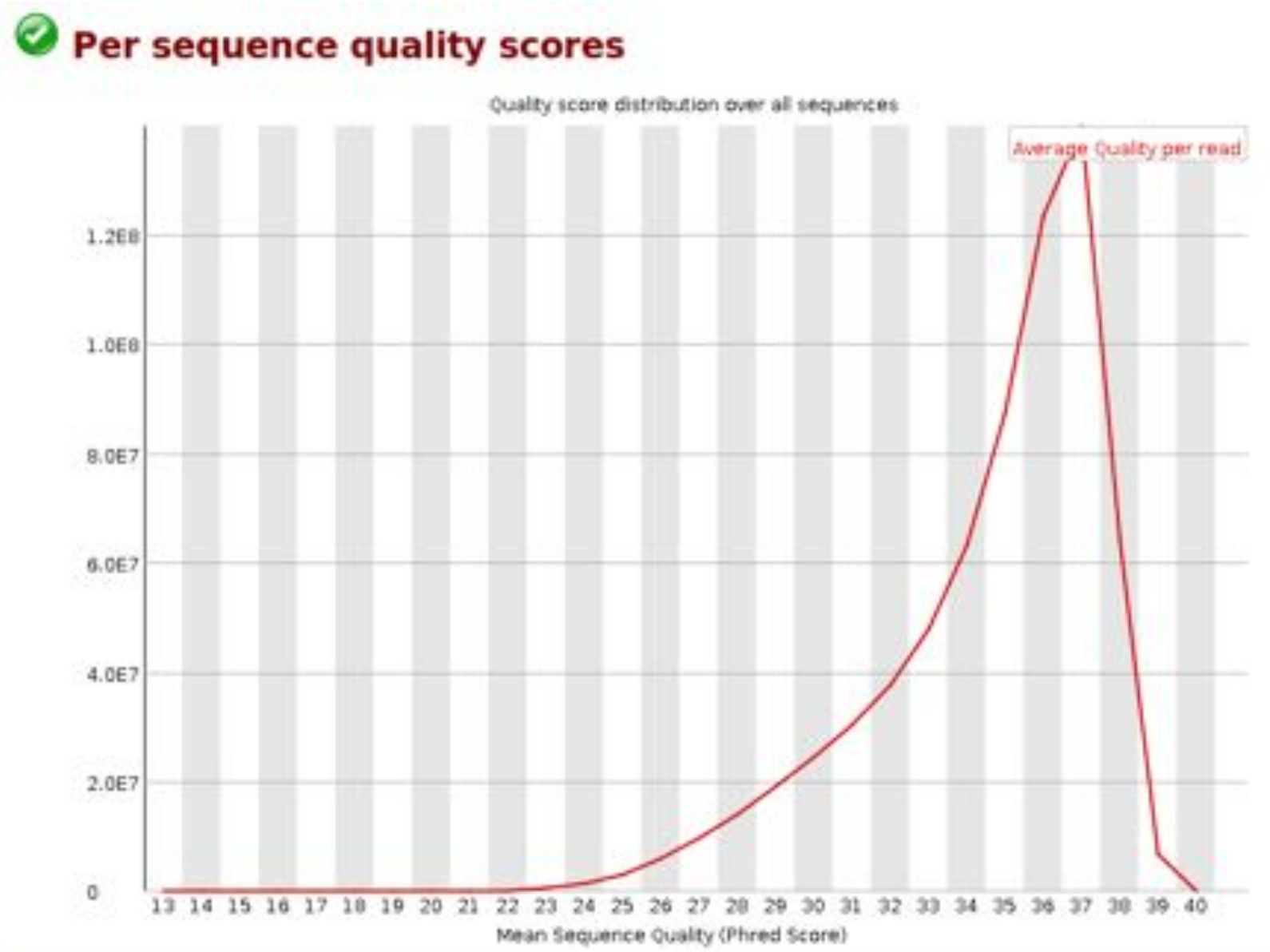
FAILED LANE



PACBIO



FastQC: Per sequence quality scores

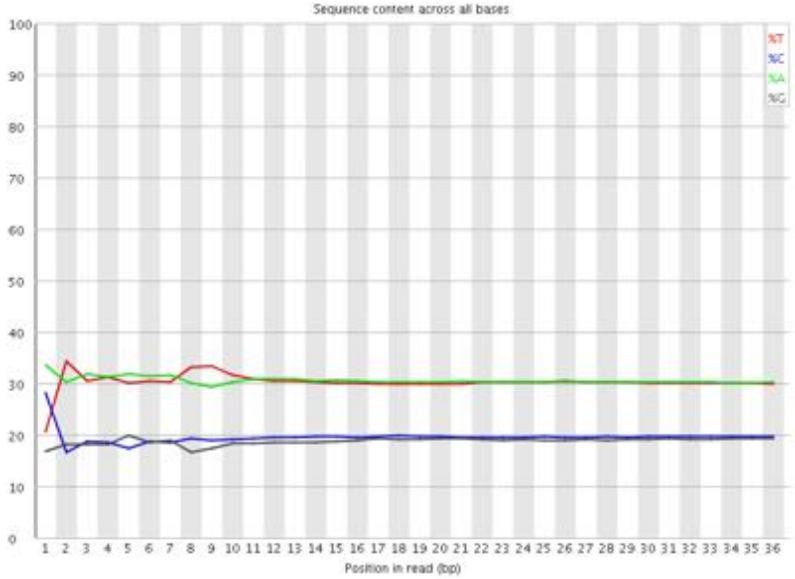


FastQC: Per base sequence content

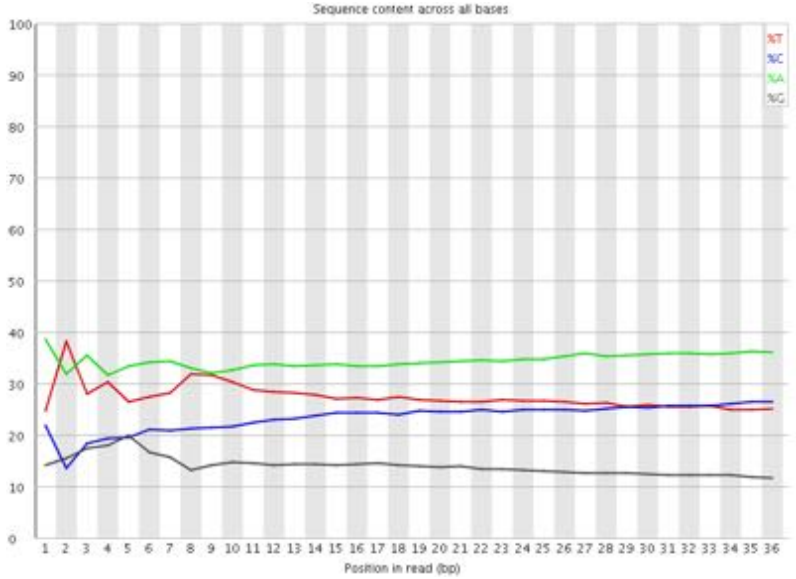
This plot shows the nucleotide distribution per read position for all reads in a lane.

- x-axis = position in read (bp)
- y-axis = % of all reads in the lane
- colors refer to individual nucleotides: **A**, **C**, **G**, **T**

GOOD LANE



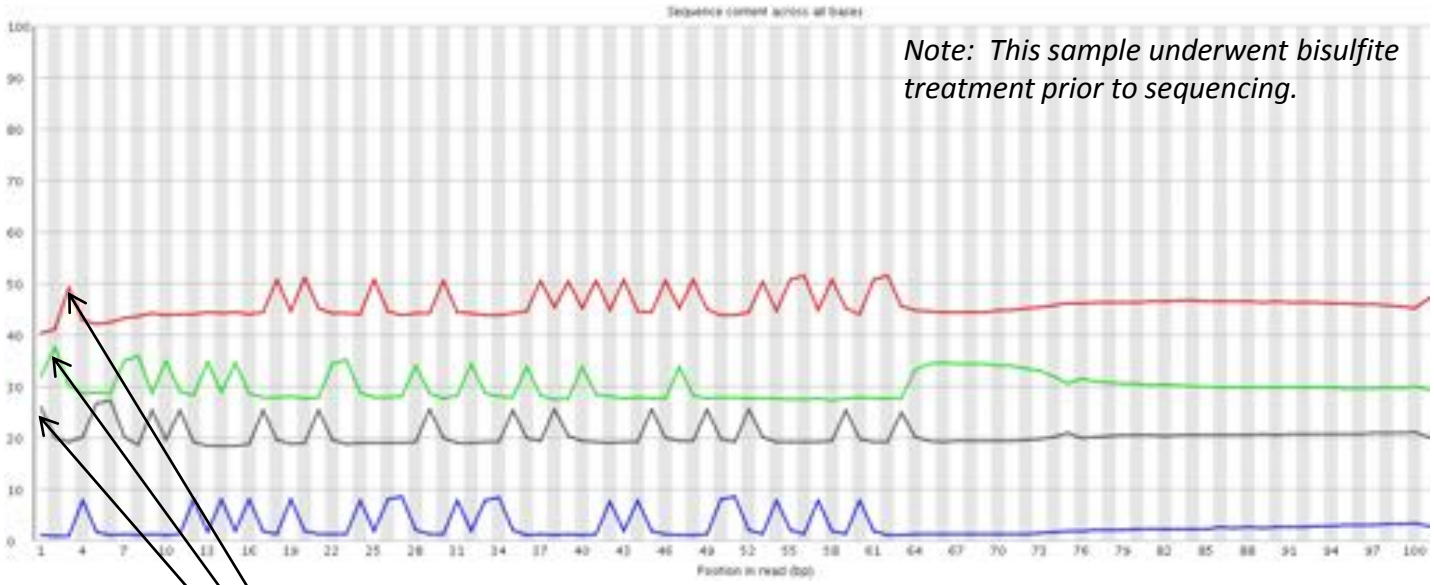
BAD LANE



Can this be fixed? **No.**

FastQC: Per base sequence content

This lane has a different problem – one sequence motif is highly over-represented.



Note: This sample underwent bisulfite treatment prior to sequencing.

primer/adapter sequence: **GATCGGAAGAGCACACGTCTGAACTCCAGTCACACAGTGATCTCGTATGCCGTCTTCTGCTTG**

In this lane, ~10% of reads have the adapter sequence & the rest are normal.

<u>Can this be fixed?</u> Yes. Simply remove the reads w/ adapter contamination, and everything that's left should be fine. (Talk to a bioinformatics analyst for help.)

FastQC: Per sequence GC content

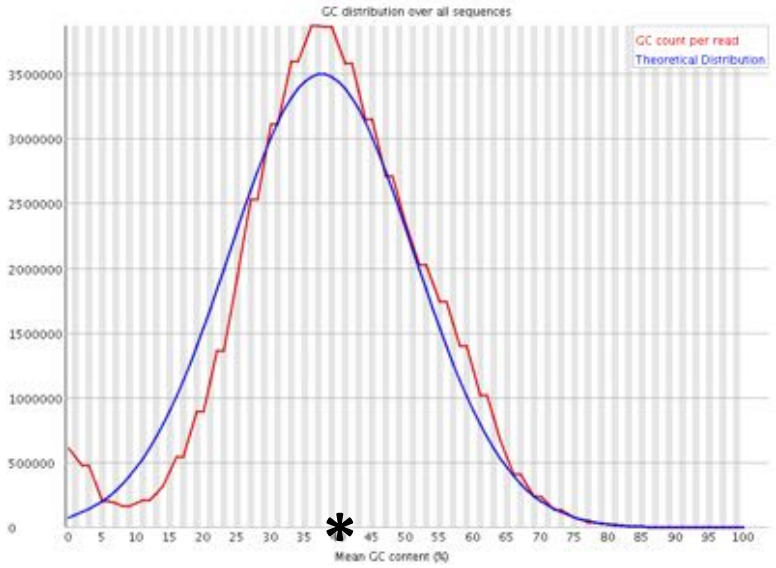
This plot shows the distribution of GC content per read for all reads in a lane.

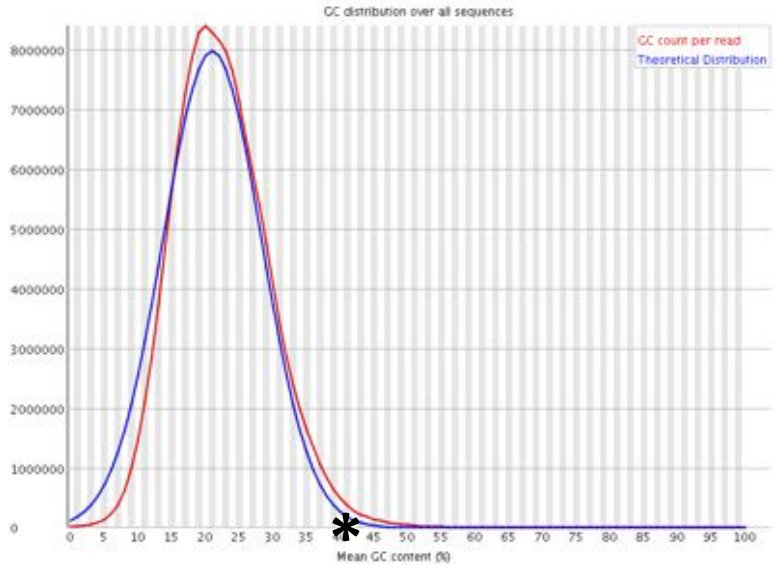
- x-axis = mean GC content (%)
- y-axis = # of reads
- red: observed read count, blue: theoretical distribution (given observed)

GOOD LANE

mouse genome ≈ 40% GC

BAD LANE

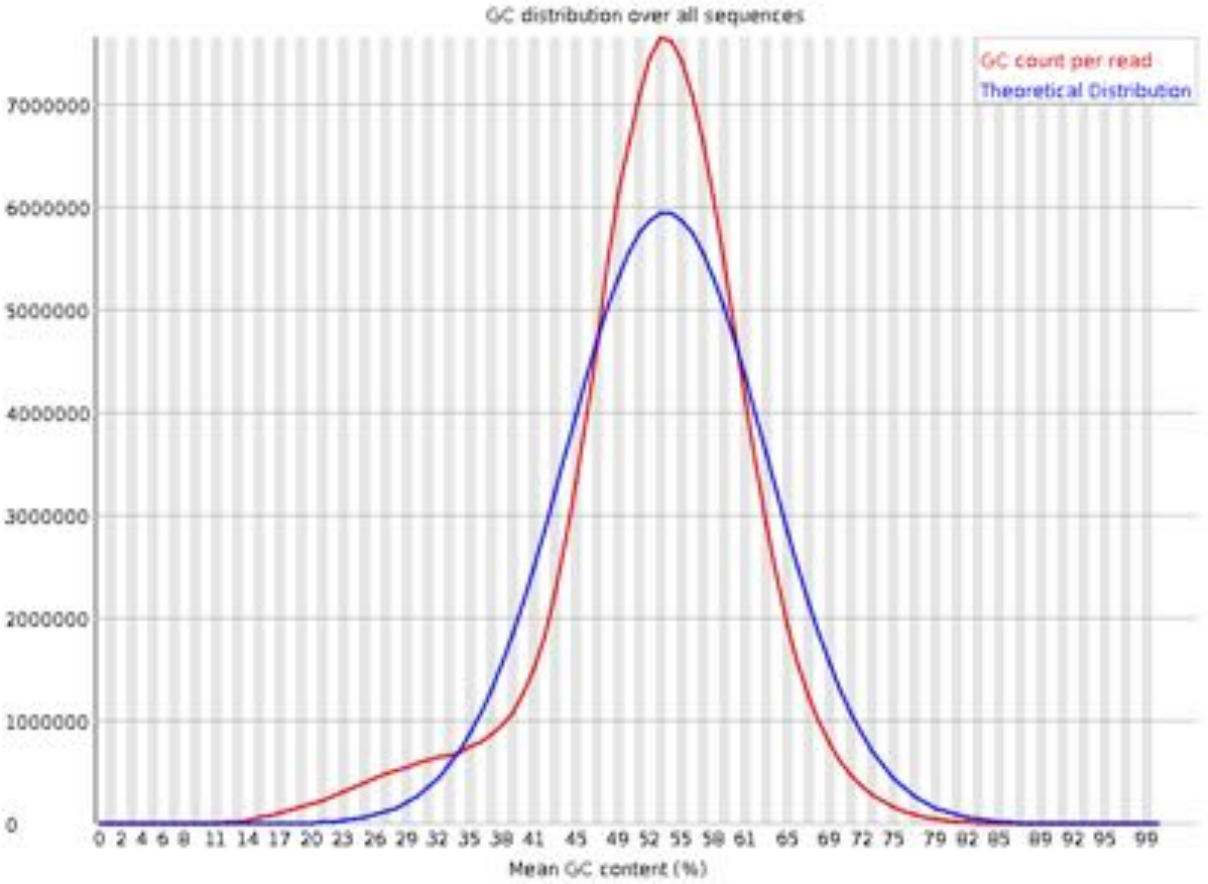




Can this be fixed? **No.**

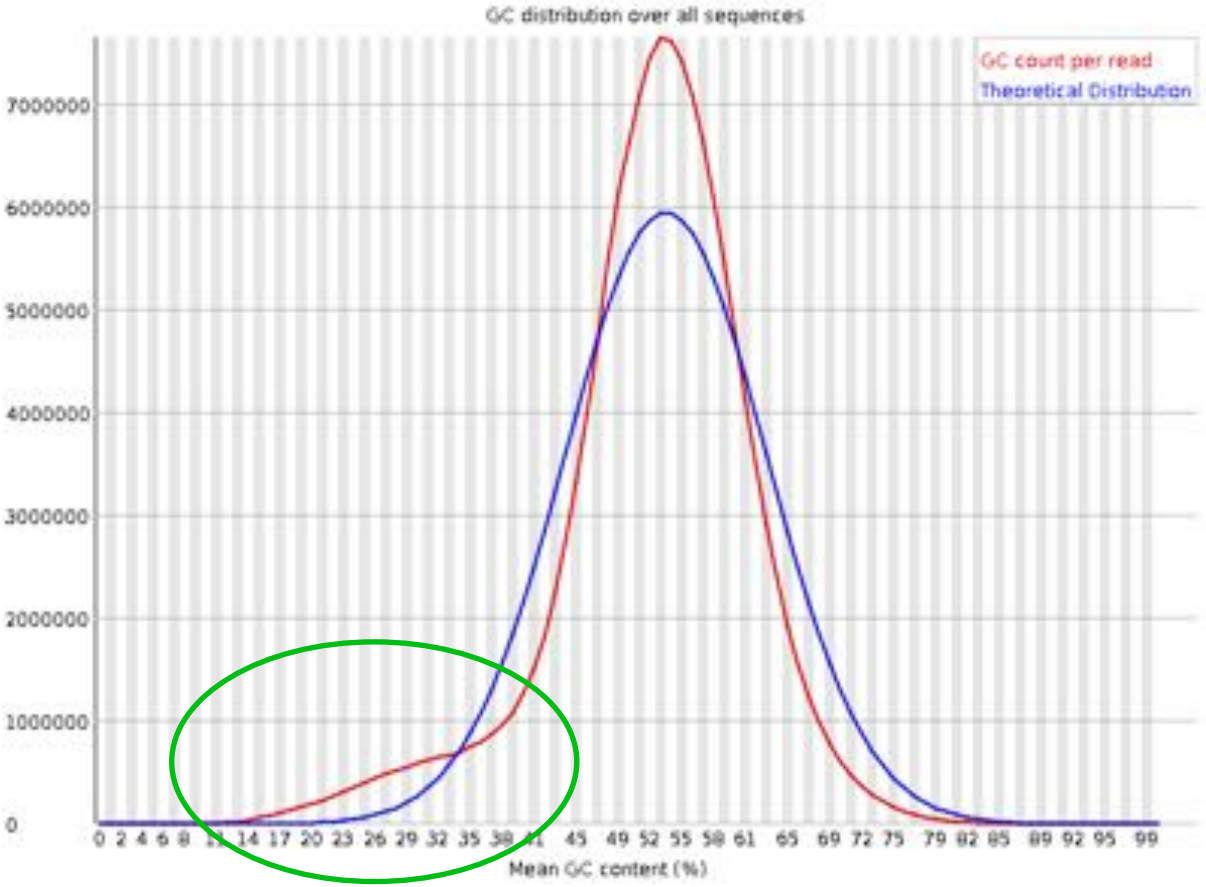
FastQC: Per sequence GC content

- A contamination ?



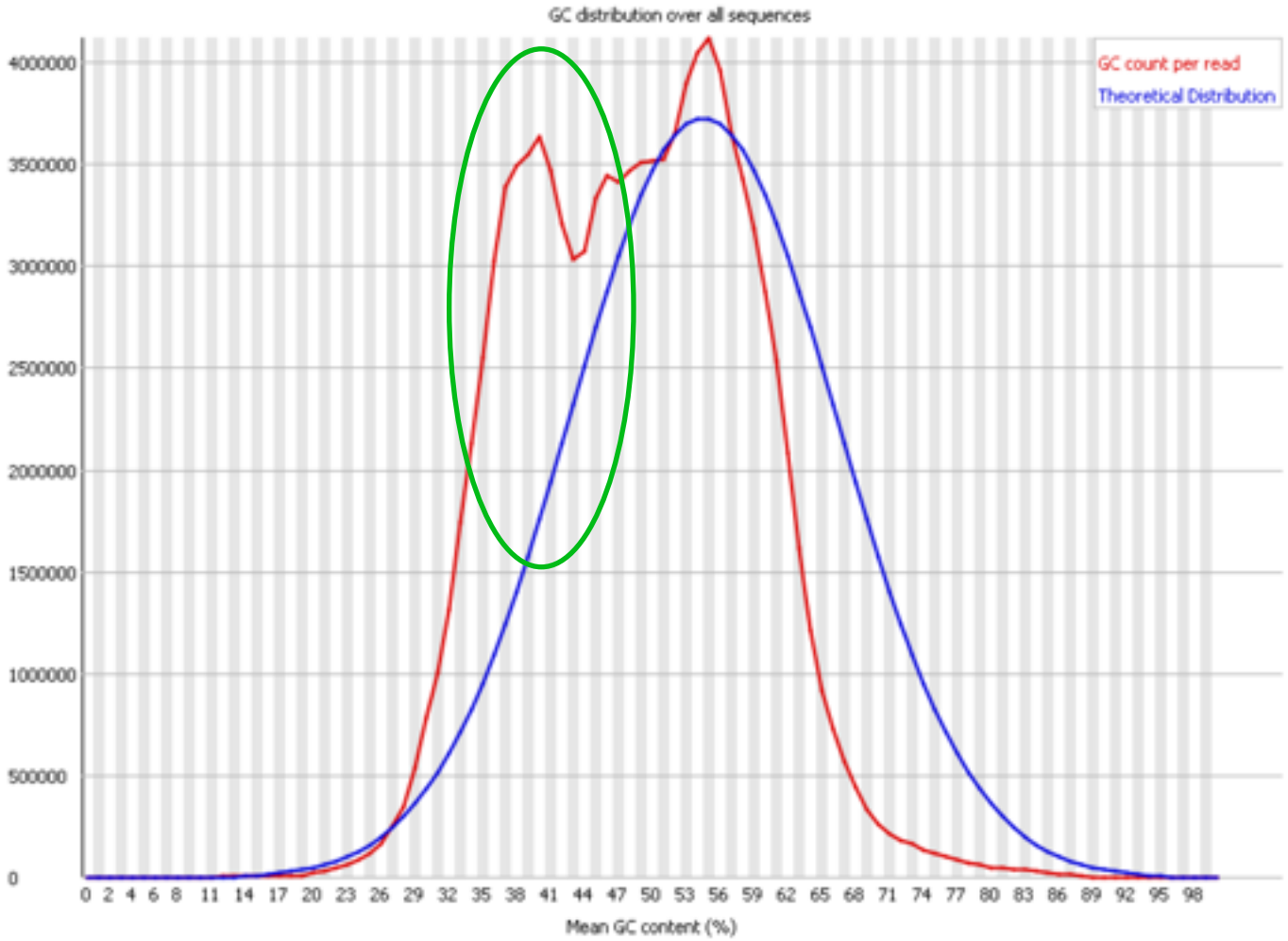
FastQC: Per sequence GC content

- A contamination ?

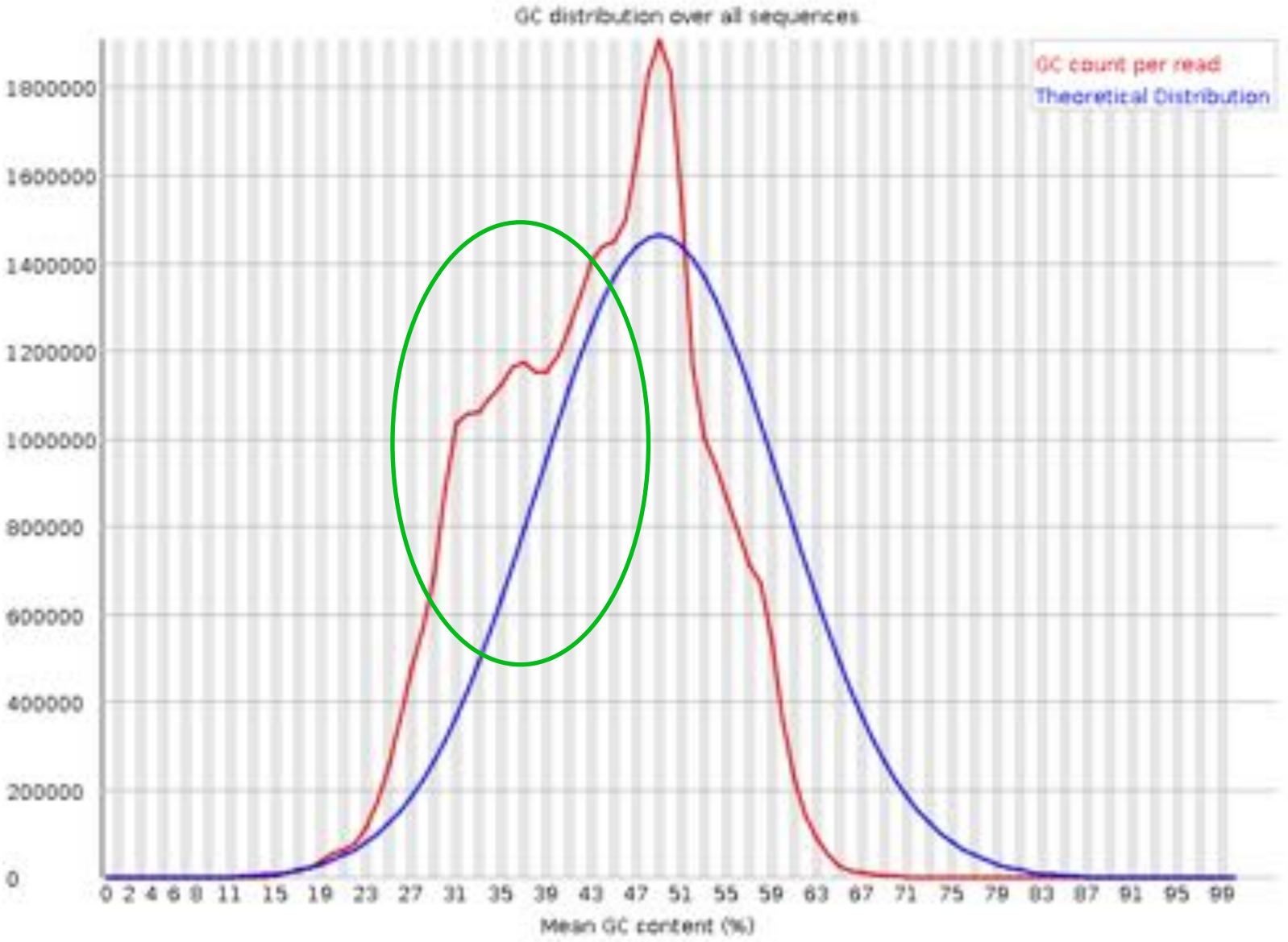


<u>Can this be fixed ?</u> Maybe…

FastQC: Per sequence GC content

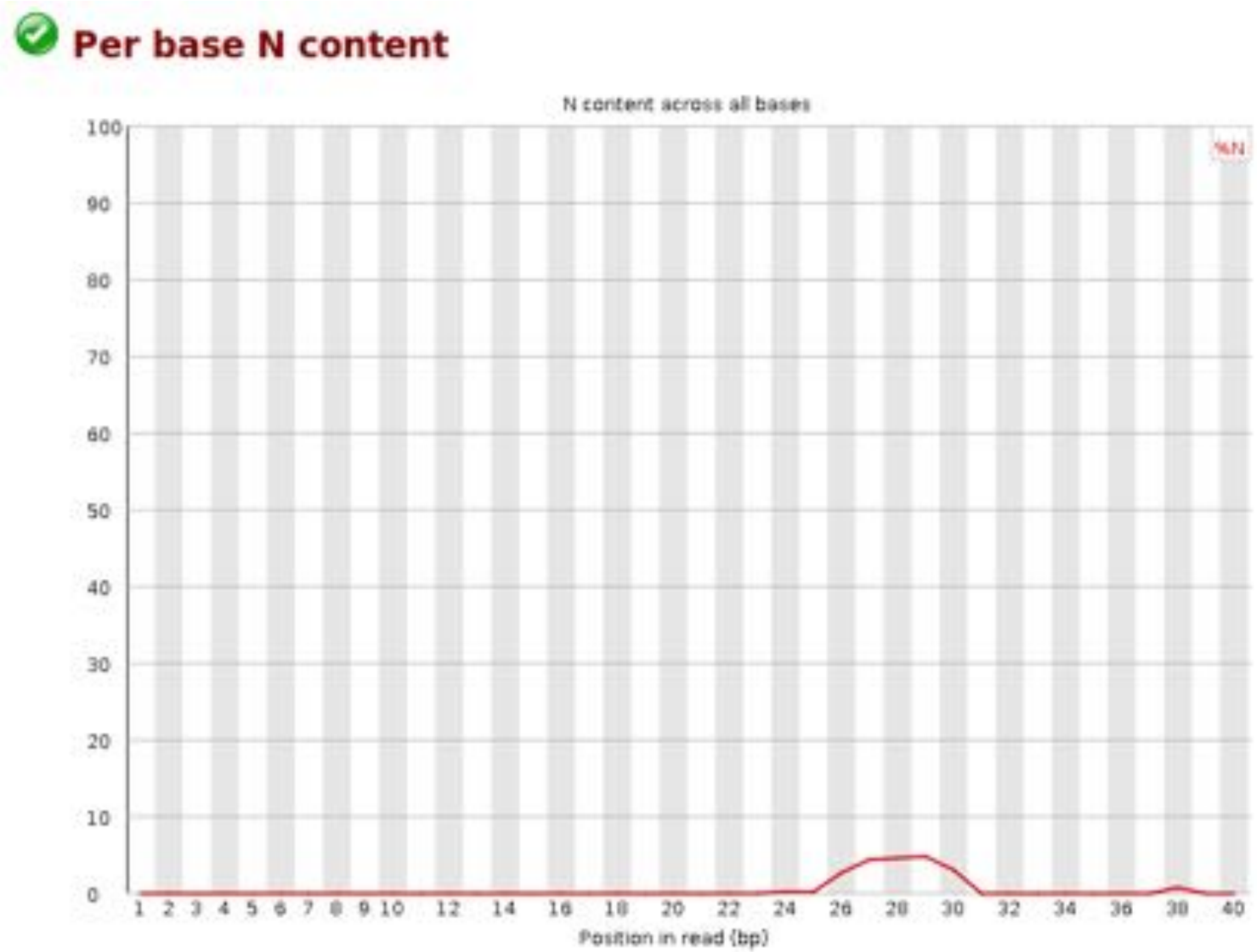


Third-party contamination : detection

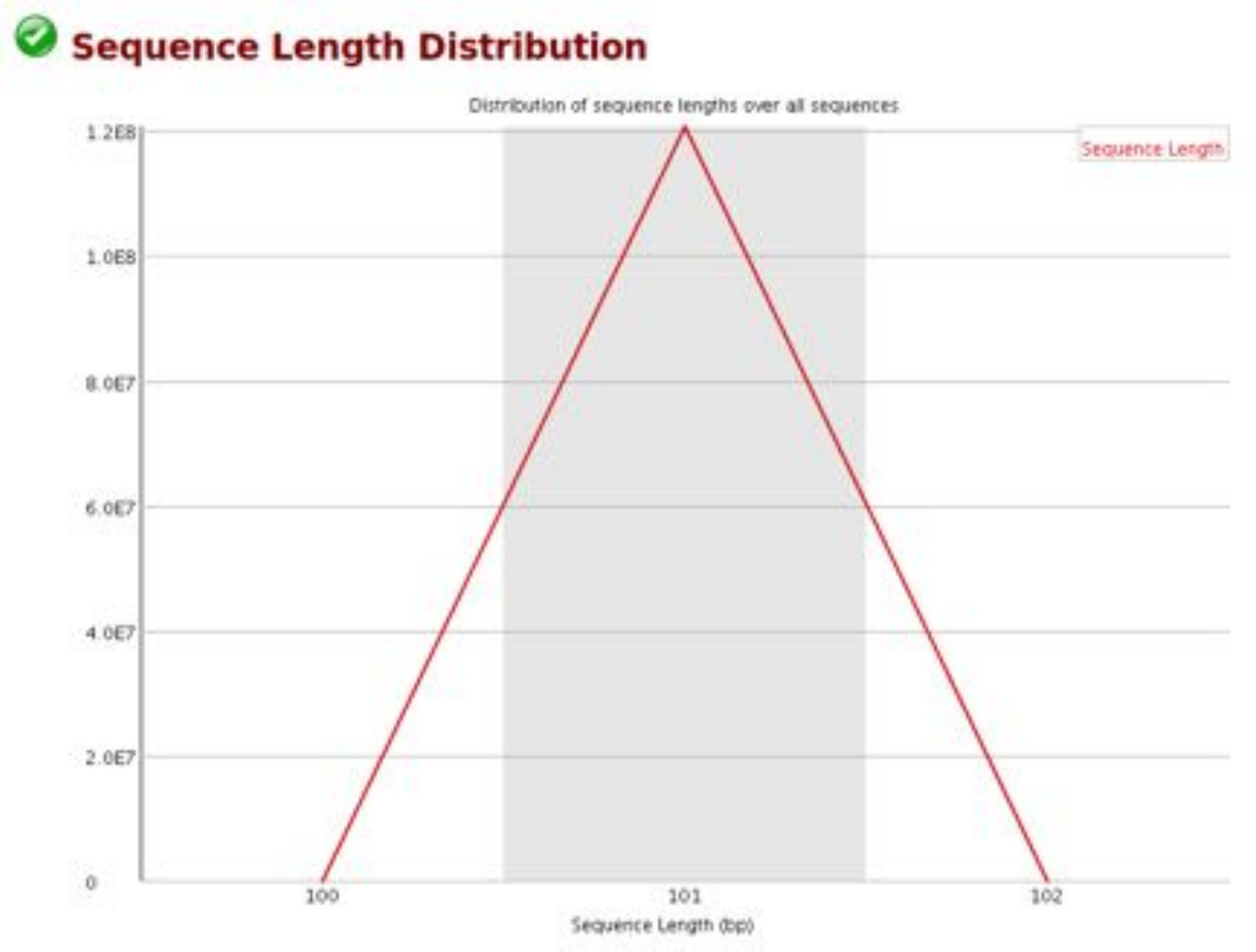


Sabellaria alveolata : mantle transcriptome

FastQC: Per base N content



FastQC: Sequence Length Distribution

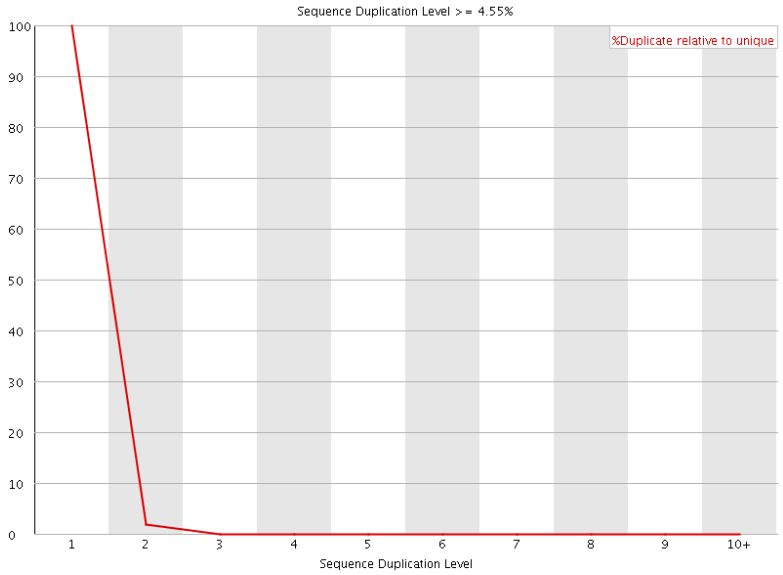


FastQC: Sequence Duplication Levels

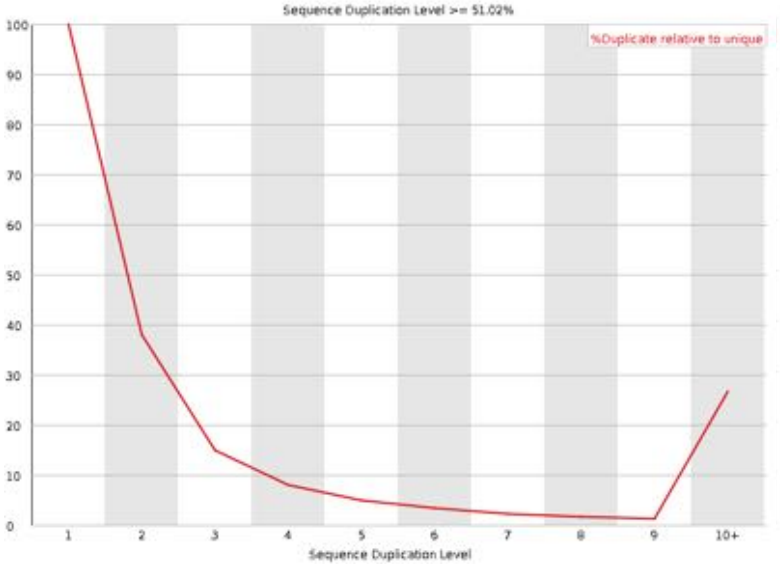
This plot shows the degree of duplication for a subset of reads in a lane.

- x-axis = sequence duplication level
- y-axis = % duplicates relative to unique reads

GOOD LANE

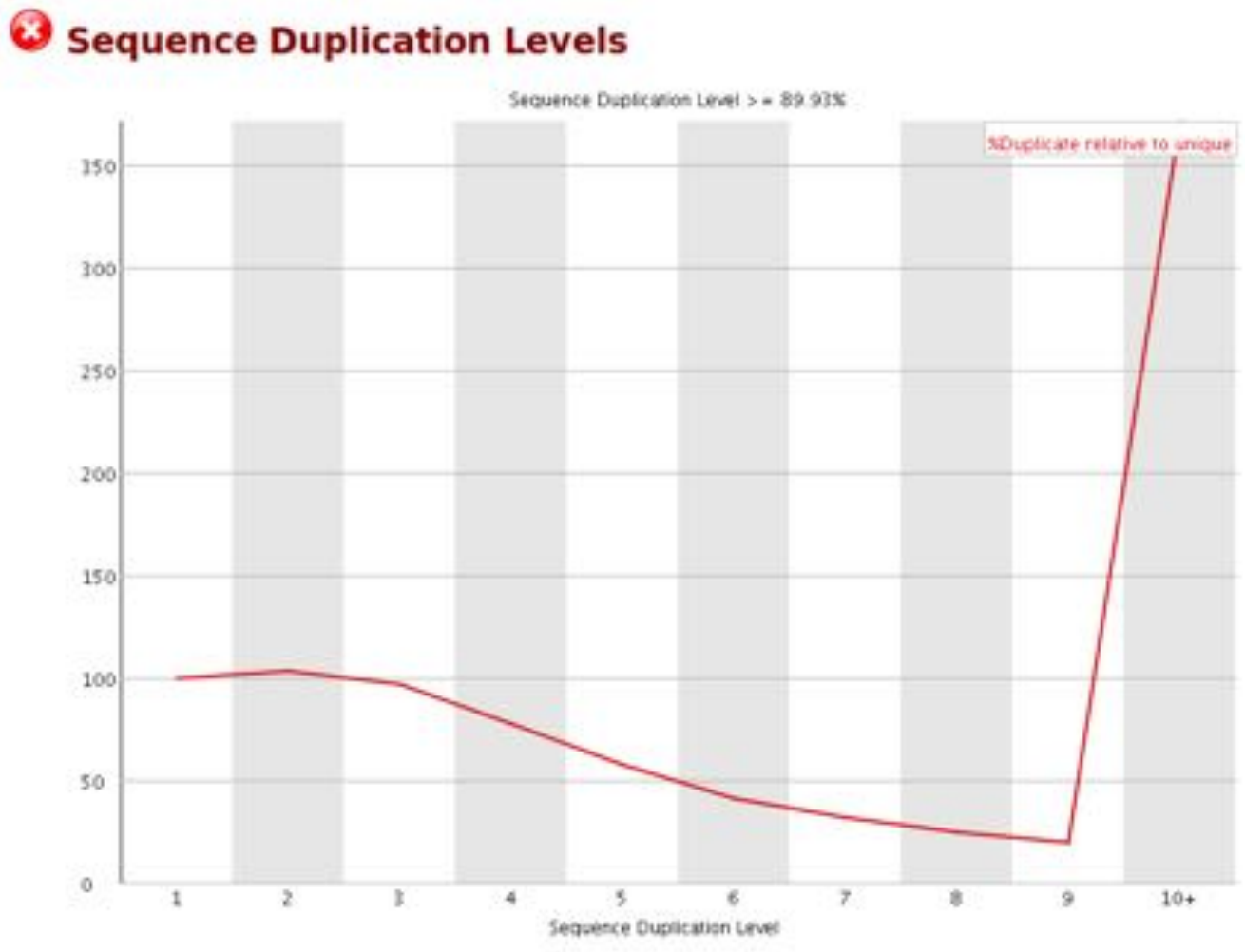


BAD LANE



Can this be fixed? Maybe.

FastQC: Sequence Duplication Levels



<u>Can this be fixed?</u> Hem…

FastQC: Overrepresented sequences

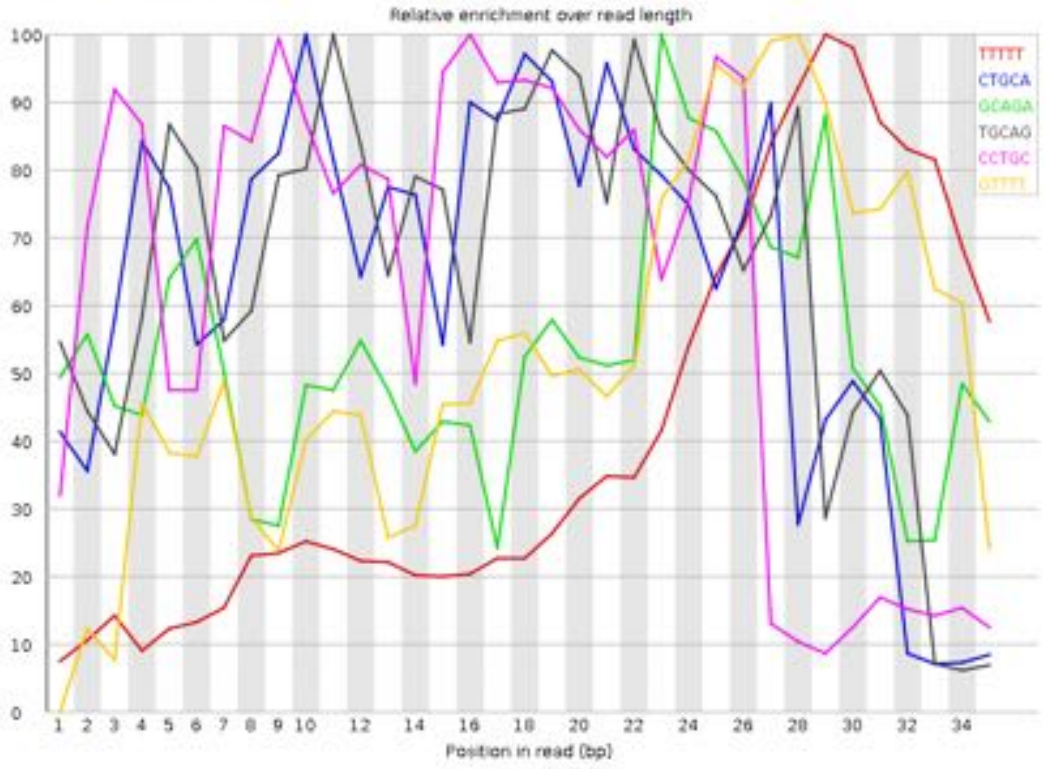
Overrepresented sequences

Sequence	Count	Percentage	Possible Source
AGAGTTTTATCGCTTCCATGACGCAGAAGTTAACACTTTC	2065	0.5224039181558763	No Hit
GATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATG	2047	0.5178502762542754	No Hit
ATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATGA	2014	0.5095019327680071	No Hit
CGATAAAAATGATTGGCGTATCCAACCTGCAGAGTTTTAT	1913	0.4839509420979134	No Hit
GTATCCAACCTGCAGAGTTTTATCGCTTCCATGACGCAGA	1879	0.47534961850600066	No Hit
AAAAATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCT	1846	0.4670012750197325	No Hit
TCATGGAAGCGATAAAACTCTGCAGGTTGGATACGCCAAT	665	0.16823177025358726	No Hit
TCTGCGTCATGGAAGCGATAAAACTCTGCAGGTTGGATAC	627	0.15861852623909656	No Hit
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCT	624	0.1578595859221631	Illumina Paired End PCR Primer 2 (100% over 40bp)
CCTGCAGAGTTTTATCGCTTCCATGACGCAGAAGTTAACA	613	0.15507680476007366	No Hit
CGGTTCAGCAGGAATGCCGAGATCGGAAGAGCGGTTCAGC	599	0.15153508328105078	Illumina Paired End PCR Primer 2 (96% over 25bp)
TCTGCAGGTTGGATACGCCAATCATTTTTATCGAAGCGCG	585	0.1479933618020279	No Hit
CGCTTAAAGCTACCAGTTATATGGCTGGGGGGTTTTTTTT	552	0.13964501831575965	No Hit
CTCTGCAGGTTGGATACGCCAATCATTTTTATCGAAGCGC	532	0.1345854162028698	No Hit
CTGCGTCATGGAAGCGATAAAACTCTGCAGGTTGGATACG	515	0.13028475440691342	No Hit
CTGCAGGTTGGATACGCCAATCATTTTTATCGAAGCGCGC	505	0.12775495335046852	No Hit
GCTTAAAGCTACCAGTTATATGGCTGGGGGGTTTTTTTTG	411	0.10397482341988626	No Hit

Adapter dimers
rRNA
Satellite sequences

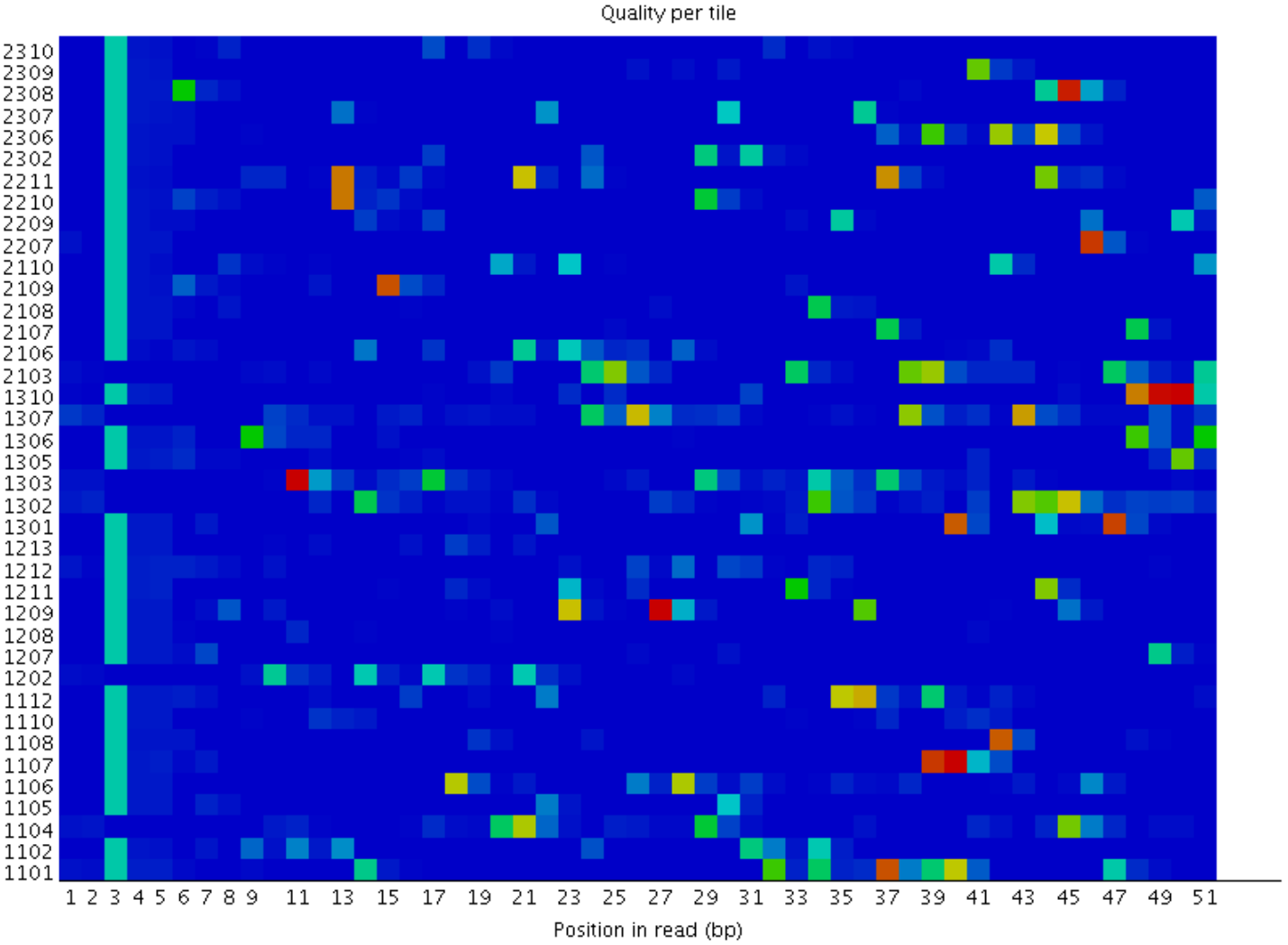
FastQC: Kmer Content

Kmer Content



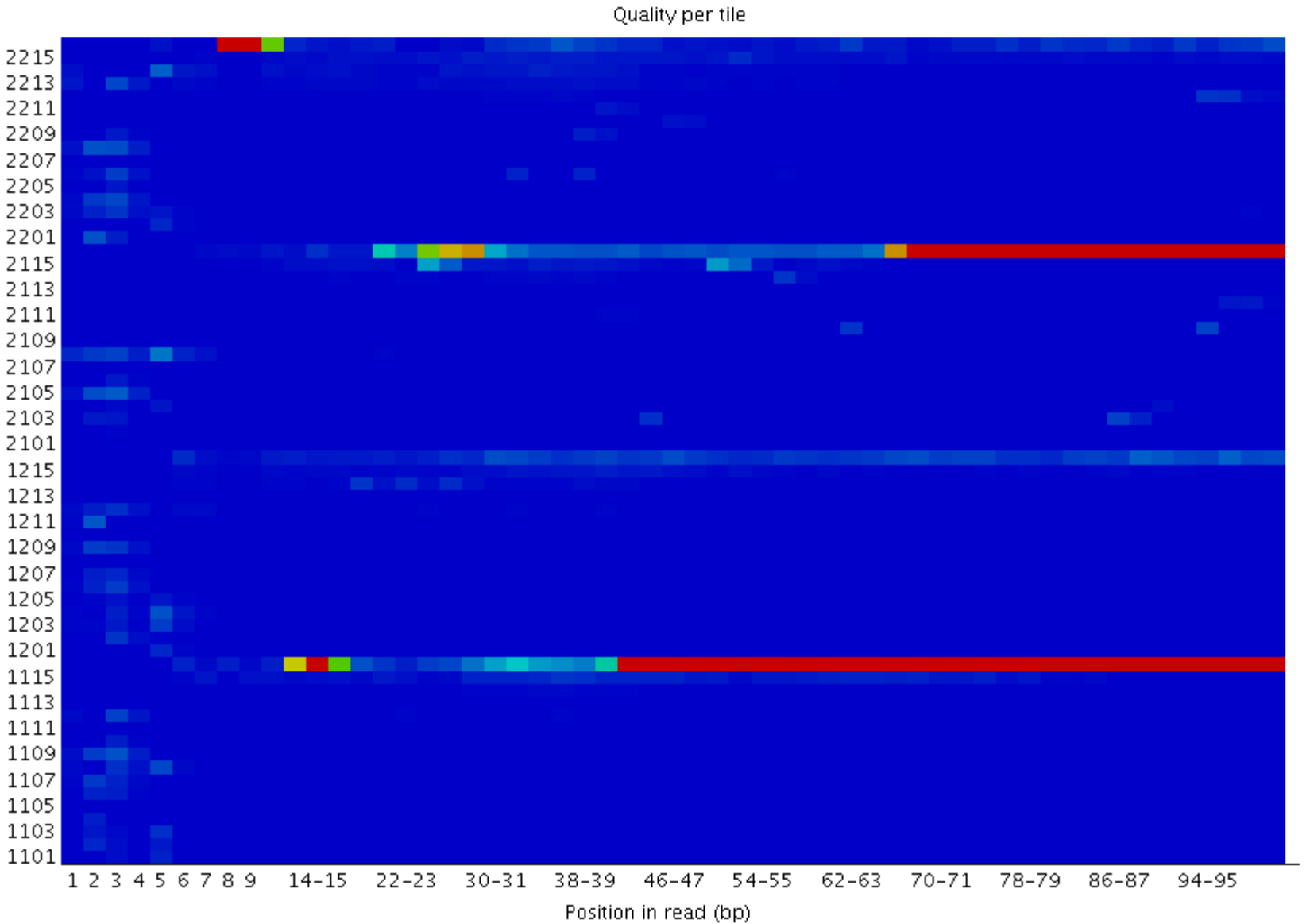
Sequence	Count	Obs/Exp Overall	Obs/Exp Max	Max Obs/Exp Posi
TTTTT	192940	8.590186	21.06293	29
CTGCA	90975	7.7906475	12.251836	10
GCAGA	84910	7.163295	13.539302	23
TGCAG	92470	7.002405	10.671717	11
CCTGC	57235	5.4987235	8.729035	16
GTTTT	108205	5.324498	10.243909	28
CAACC	49005	5.2869425	9.85526	13
ATCGC	58320	4.9942355	8.029807	29
CCAAC	46220	4.9864807	9.408141	12
AAAAA	62285	4.7588468	8.0126295	5
CAGAG	56370	4.7555633	7.148592	20
ACCTG	55315	4.736902	7.919266	15
CGCCA	44035	4.7130895	8.830201	35
GGGGG	63675	4.67525	16.94222	27
GCAGG	55380	4.6350074	17.521912	19
AAAAC	51945	4.452569	8.159592	24
TATCG	64615	4.4271946	8.394971	34
GCTGG	58505	4.3952427	10.37436	18
AACCT	50775	4.382863	7.691214	14
TTATC	70080	4.3444843	7.810299	33
TTTTA	87340	4.332125	7.8541703	28
TTTAT	86645	4.297653	7.9511886	35
CGCTT	54695	4.2042785	6.9374876	31

Tile Problems - Overclustering

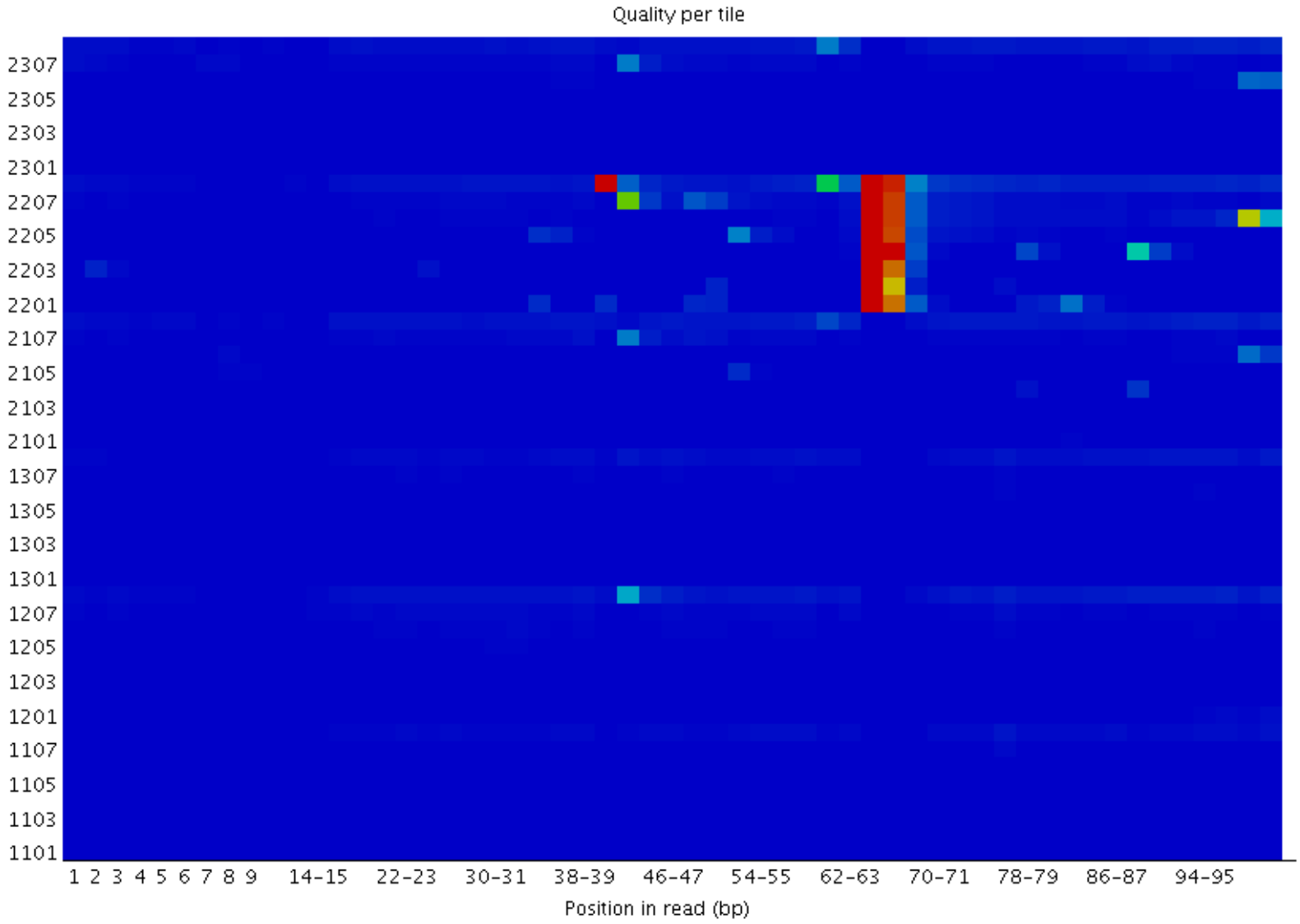


Tile Problems – Consistent tile fail

Repet sequences



Tile problems – transient tile fail



Bubble

Reasons for seeing warnings or errors on this plot could be

- transient problems such as : bubbles going through the flowcell,
- or they could be more permanent problems such as smudges on the flowcell or debris inside the flowcell lane.

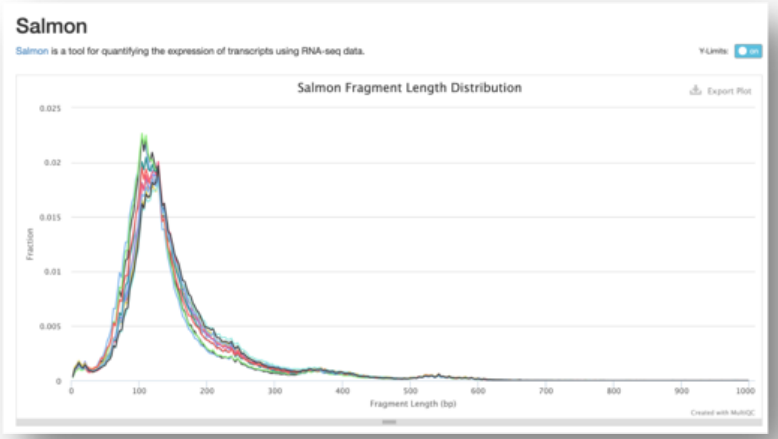
MultiQC

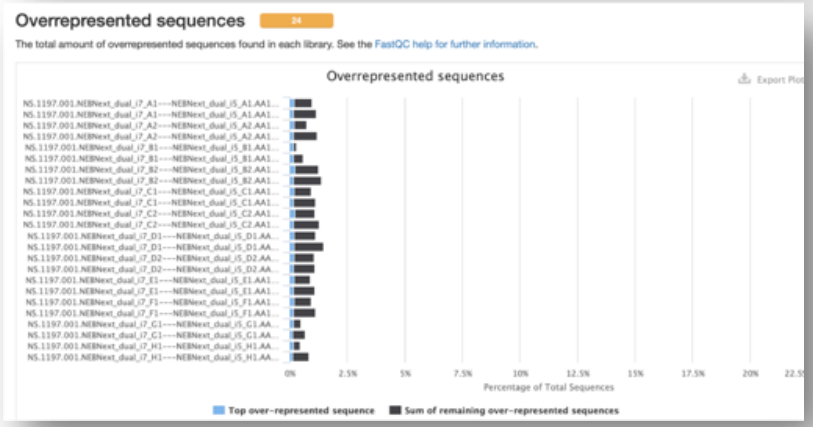
MultiQC

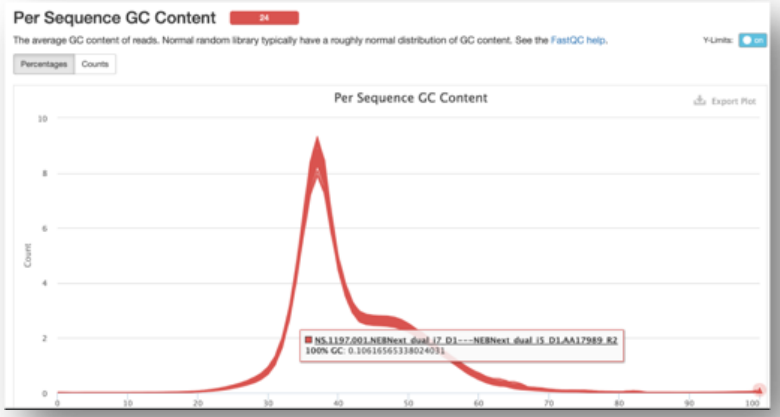
Phil Ewels
phil.ewels@scilifelab.se

More than 50 Modules
- Pre-alignment
- Aligners
- Post-alignment

Sample Name	% Aligned	M Aligned	% Dups	% GC	M Seqs
0_DS-1	26.3%	15.9			
0_DS-2	27.1%	19.7			
0_DS-3	28.3%	36.6			
0_PBS-1	26.4%	16.6			
0_PBS-2	26.8%	17.0			
0_PBS-3	28.1%	21.4			
7_6_DS-1	26.6%	19.7			
7_6_DS-2	27.3%	16.0			
7_6_DS-3	26.0%	18.4			
7_6_PBS-1	27.9%	19.0			
7_6_PBS-2	27.5%	16.5			
7_6_PBS-3	27.9%	25.4			
NS.1197.001.NEBNext_dual_i7_A1---NEBNext_dual_i5_A1.AA16701_R1			51.5%	41%	68.0
NS.1197.001.NEBNext_dual_i7_A1---NEBNext_dual_i5_A1.AA16701_R2			48.1%	42%	68.0
NS.1197.001.NEBNext_dual_i7_A2---NEBNext_dual_i5_A2.AA17680_R1			49.3%	41%	70.7
NS.1197.001.NEBNext_dual_i7_A2---NEBNext_dual_i5_A2.AA17680_R2			46.6%	41%	70.7
NS.1197.001.NEBNext_dual_i7_B1---NEBNext_dual_i5_B1.AA17673_R1			47.1%	41%	60.1
NS.1197.001.NEBNext_dual_i7_B1---NEBNext_dual_i5_B1.AA17673_R2			43.7%	42%	60.1
NS.1197.001.NEBNext_dual_i7_B2---NEBNext_dual_i5_B2.AA17664_R1			50.0%	41%	60.3





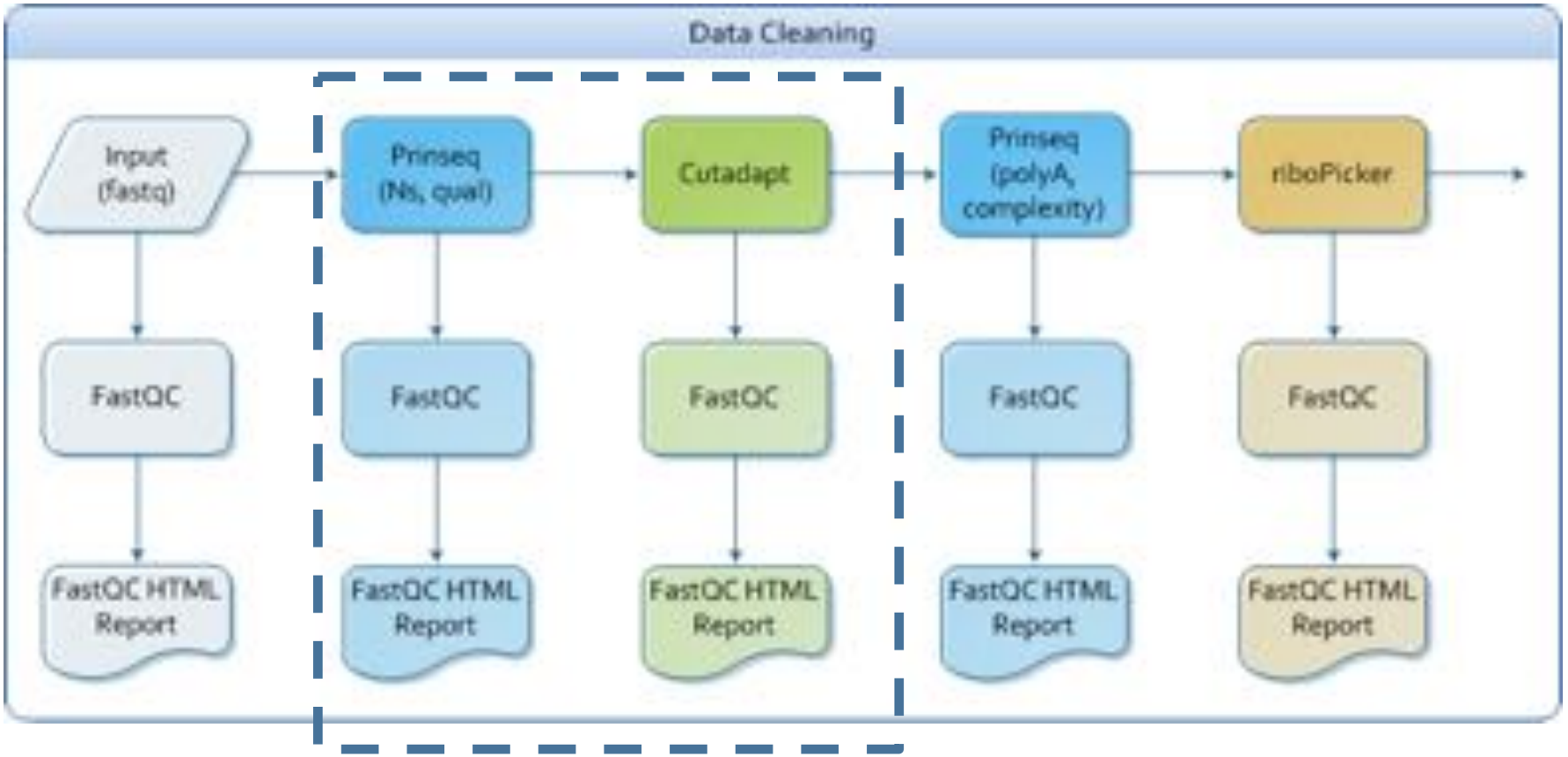


Practice

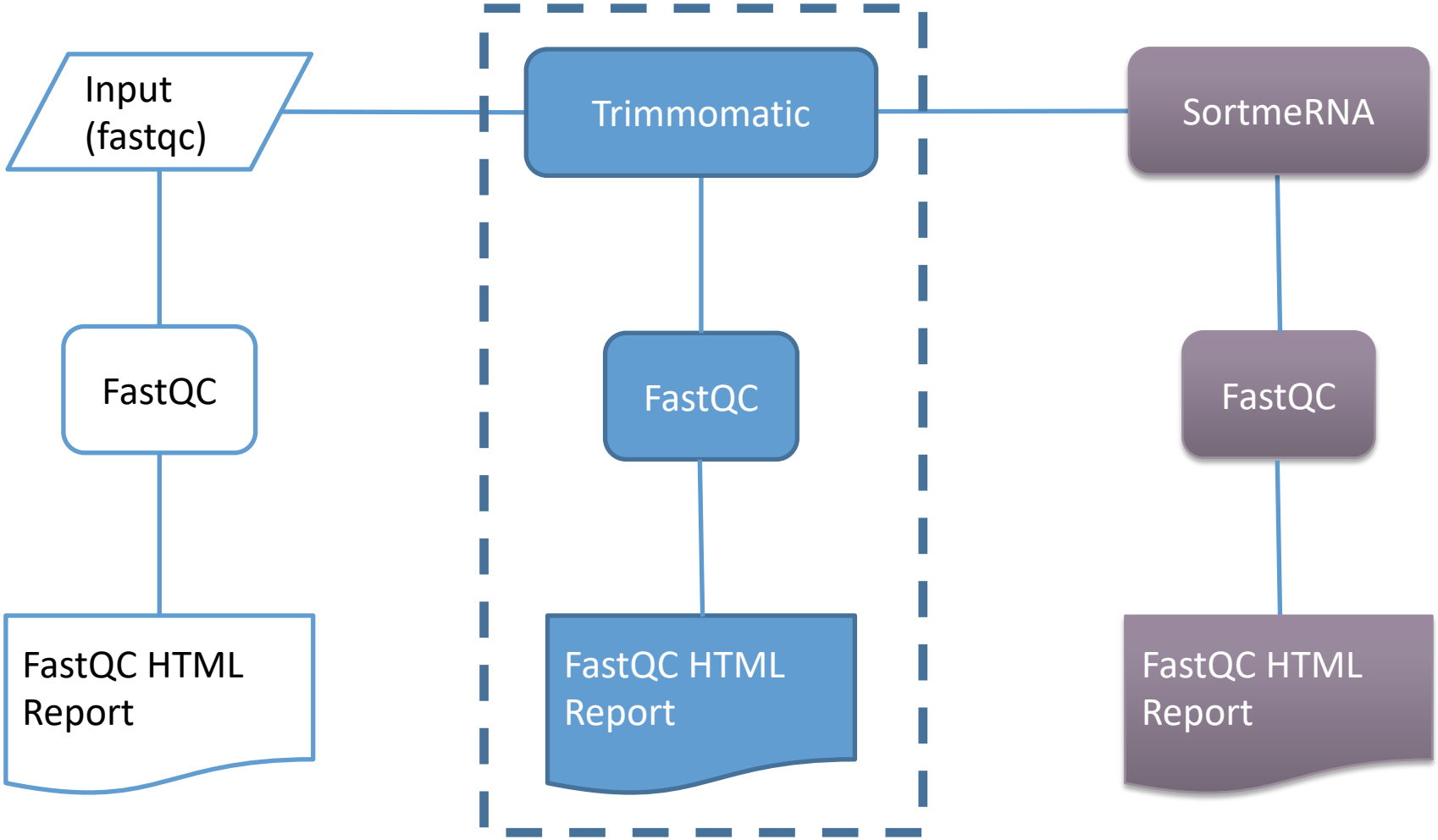
1

Aller sur la practice 1 Checking Reads Quality *du github*

Quality cleanning



Data cleaning



Trimmomatic

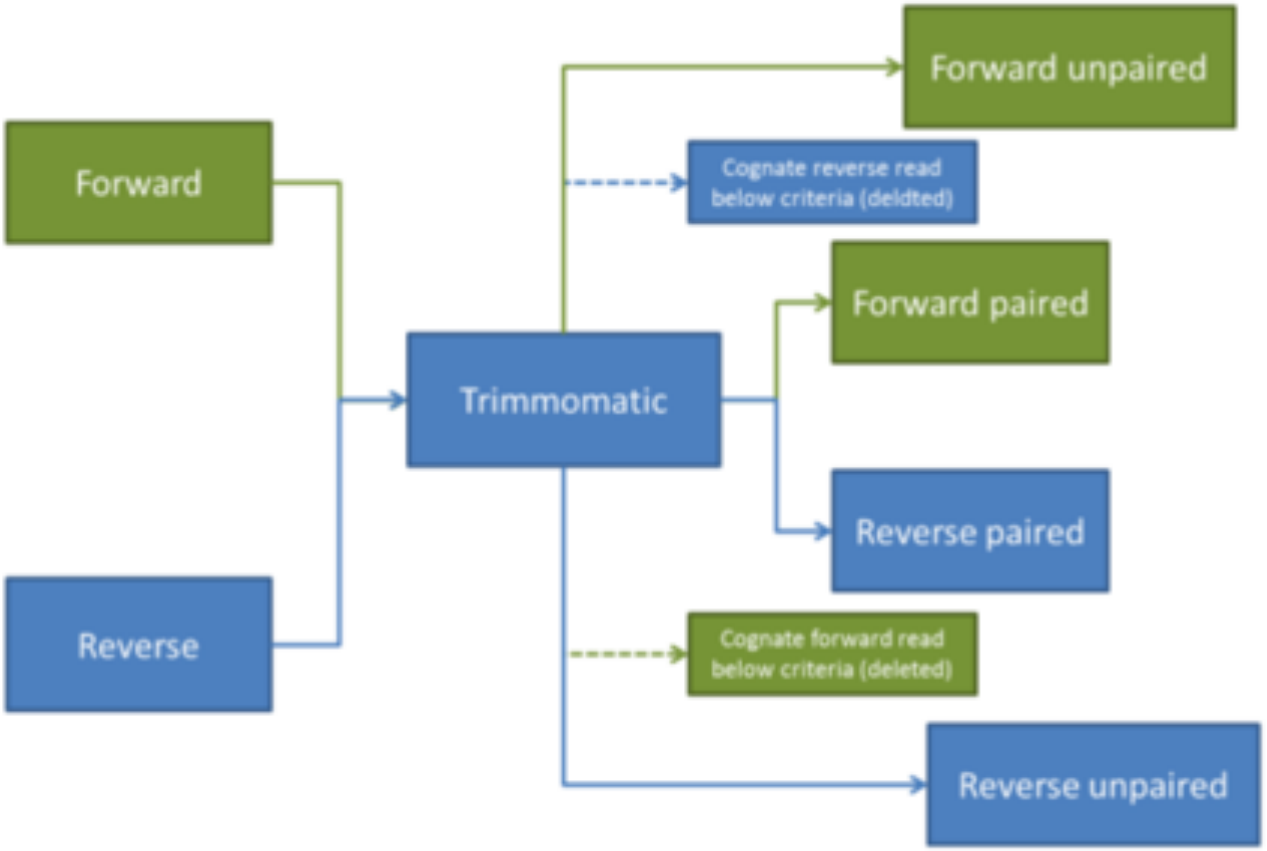


Figure 1: Flow of reads in Trimmomatic Paired End mode

Trimmomatic

- The different processing steps occur in the order in which the steps are specified on the command line.

- It is recommended in most cases that adapter clipping, if required, is done as early as possible, since correctly identifying adapters using partial matches is more difficult.

Trimmomatic

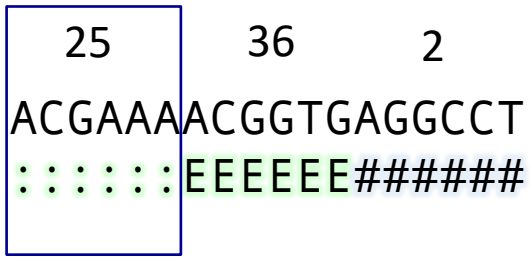
1. Cut adapters and other illumina- specific sequences from the reads
2. Cut bases off the start of a read, if below a threshold quality (3)
3. Cut bases off the end of a read, if below a threshold quality (3)

Trimmomatic

4. Perform a sliding window trimming, cutting once the average quality within the window falls below a threshold (windows = 4; mean = 20)
5. Drop reads with average quality below a threshold (25)
6. Cut the read to a specified length (depending of the reads initial lenght : (50-100bp)

Trimming Based on Quality

Sliding windows and minimum vs. average quality scores



Average:	25
Min:	25
Max:	25

Target:
Average below 20

From National Center for Genome Analysis Support: http://ncgas.org

Trimming Based on Quality

Sliding windows and minimum vs. average quality scores

Step Size = 5
Window Size = 6

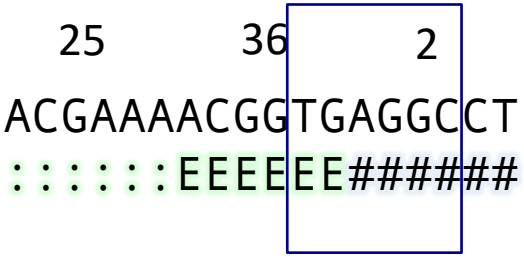
Average:	34.2
Min:	25
Max:	36

Target:
Average below 20

From National Center for Genome Analysis Support: http://ncgas.org

Trimming Based on Quality

Sliding windows and minimum vs. average quality scores



Step Size = 5
Window Size = 6

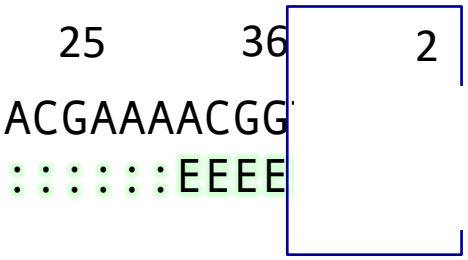
Average:	13.3
Min:	2
Max:	36

Target:
Average below 20

From National Center for Genome Analysis Support: http://ncgas.org

Trimming Based on Quality

Sliding windows and minimum vs. average quality scores



Step Size = 5
Window Size = 6

Average:	13.3
Min:	2
Max:	36

Target:
Average below 20

From National Center for Genome Analysis Support: http://ncgas.org

Trimmomatic command

```
java -jar trimmomatic.jar  PE -phred33
\ lib1_1.fastq lib1_2.fastq                 Raw reads
\ lib1_1.P.qtrim lib1_1.U.qtrim             Paired and unpaired reads1
\ lib1_2.P.qtrim lib1_2.U.qtrim             Paired and unpaired reads2
\ ILLUMINACLIP:illumina.fa:2:30:10          Adapters
\ SLIDINGWINDOW:4:5 LEADING:5 TRAILING:5 MINLEN:25
```

```
Input Read Pairs: 2 000 000
Both Surviving: 1 879 345 (93.97%)
Forward Only Surviving: 94 153 (4.71%)
Reverse Only Surviving: 18 098 (0.90%)
Dropped: 8 404 (0.42%)

TrimmomaticPE: Completed successfully
```

Trimming effect

Recent publications have identified contradictory results of the effects of trimming raw reads on the quality of the assembly

-> How de novo assemblers manage the variable reads size?

-> Should we prefer a complete removal of the read to the deletion of the only poor quality part?

-> Add later additional cleaning step

- Del Fabbro, C., Scalabrin, S., Morgante, M., & Giorgi, F. M. (2013). An Extensive Evaluation of Read Trimming Effects on Illumina NGS Data Analysis. PLoS ONE, 8(12), e85024. doi:10.1371/journal.pone.0085024

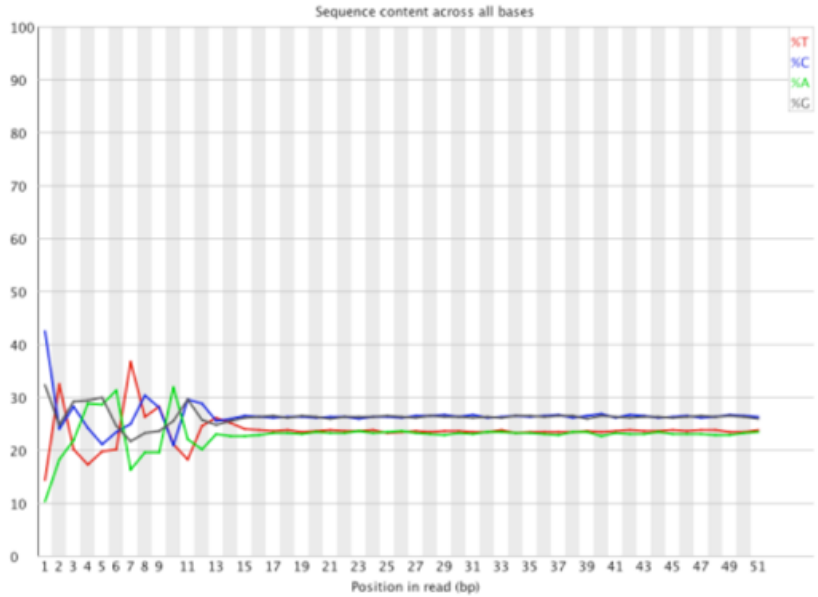
-> *"trimming is beneficial in RNA-Seq, SNP identification and genome assembly procedures, with the best effects evident for intermediate quality thresholds (Q between 20 and 30)"*

- MacManes, M. D. (2014, November). On the optimal trimming of high-throughput mRNAseq data. Biorxiv. doi:10.1101/000422

-> *"Although very aggressive quality trimming is common, this study suggests that a more gentle trimming, specifically of those nucleotides whose Phred score < 2 or < 5, is optimal for most studies across a wide variety of metrics."*

- Sleep, J. A., Schreiber, A. W., & Baumann, U. (2013). Sequencing error correction without a reference genome. BMC Bioinformatics, 14(1), 367. doi:10.1186/gb-2011-12-11-r112

Beginnings of reads



Bias in sequence composition is often (always?) seen in the first 12-15 bp in Illumina RNA-seq data sets

Thought to be due to issues with “random” hexamer priming

Hansen et al. (2010) **Biases in Illumina transcriptome sequencing caused by random hexamer priming** Nucleic Acids Res. 2010 July; 38(12): e131. doi: 10.1093/nar/gkq224

Not clear if trimming the 5’ helps here.

Duplicate sequences

Observing identical sequences in a sequencing run could result from

- Genuine, multiple observations of the same sequence from different source molecules
- Amplification from PCR steps in library preparation or sequencing
- Optical duplicates
- Exhausting the library; sequencing the same molecule several times

Note:

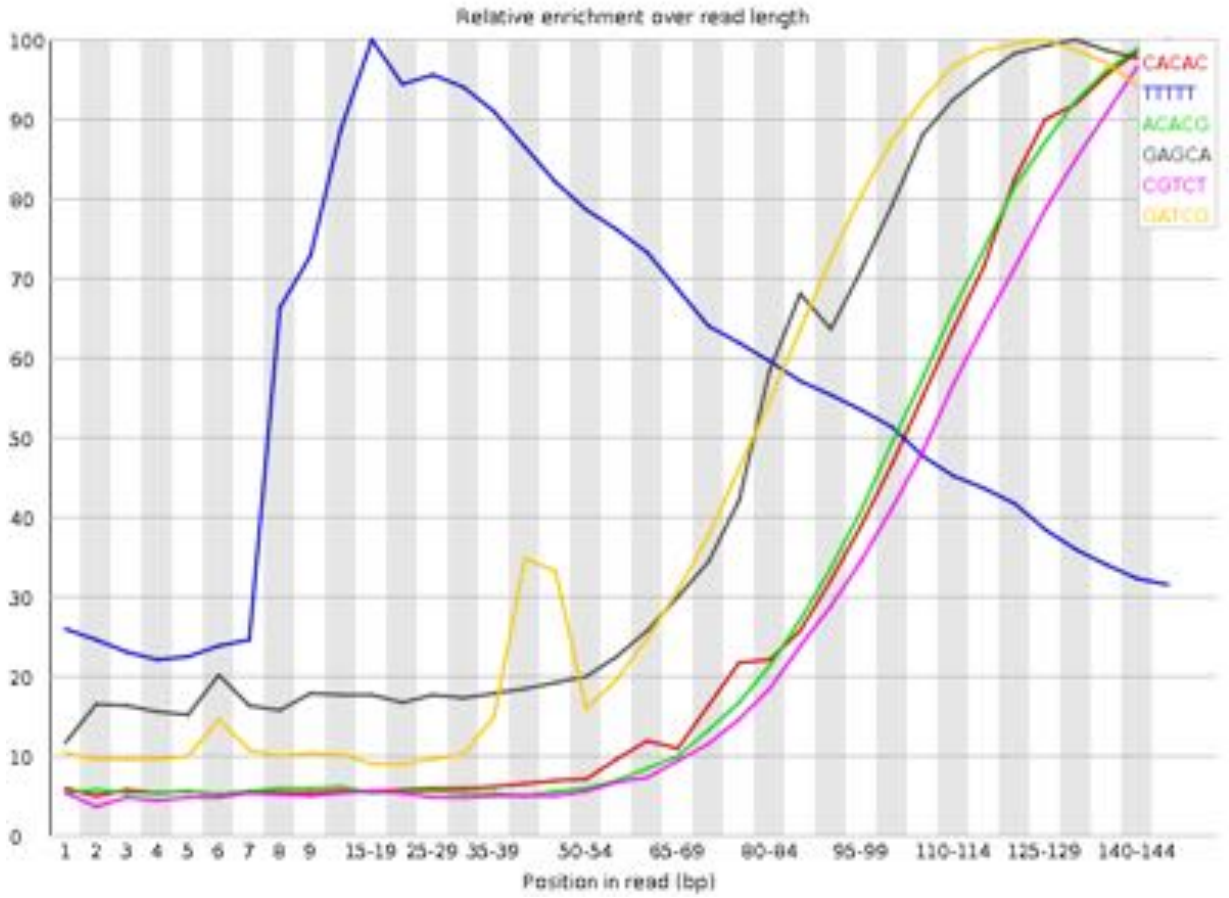
For resequencing applications (whole-genome, exome sequencing) it is standard practice to remove duplicate sequences. For RNA-seq, things are more complicated.

Duplicates are usually removed after mapping because it is simple. E g look for paired-end reads where both mates map to the same coordinates.

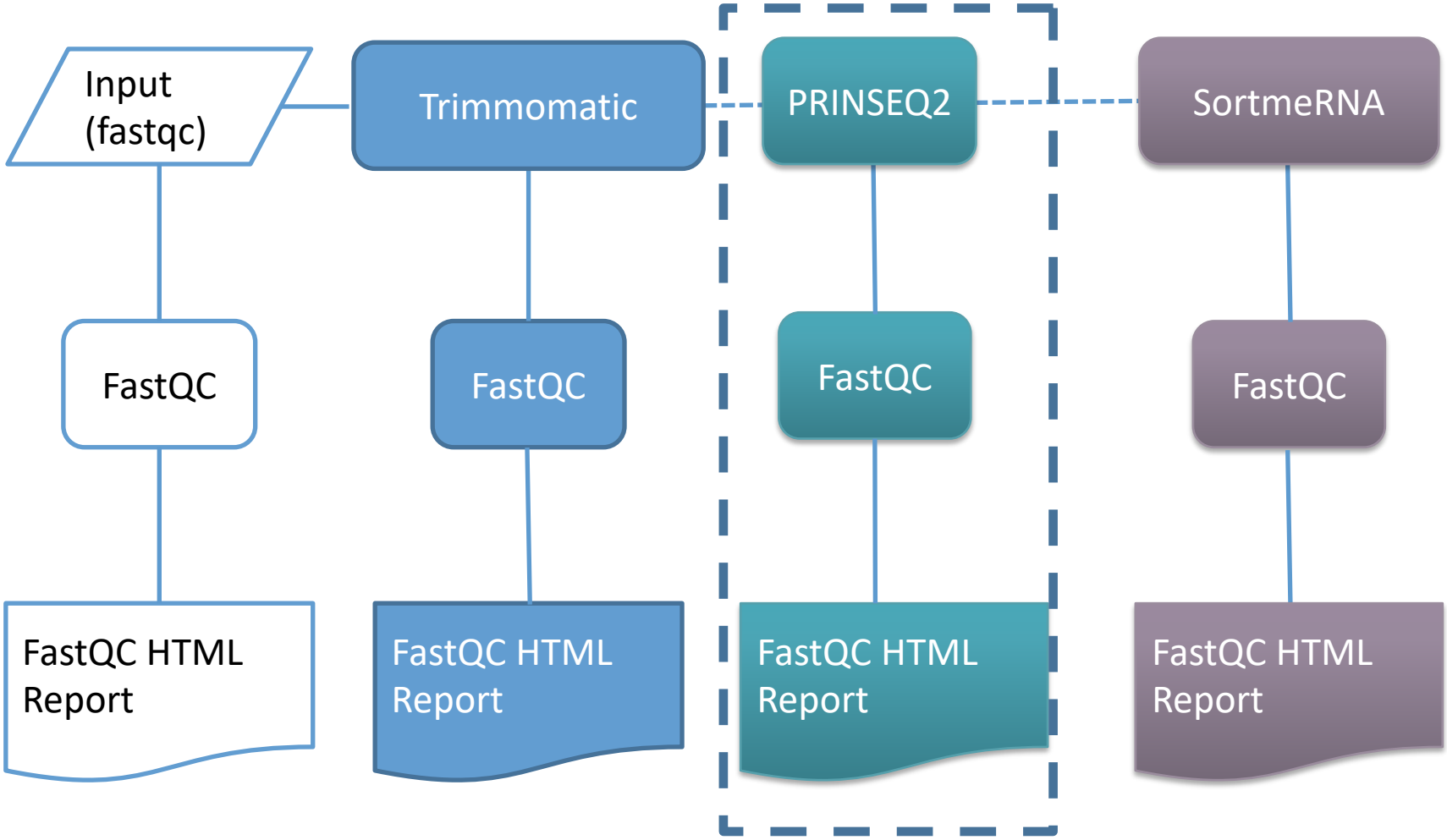
Poly-A Tails and Other Artifacts

–Library Prep – retained and sequenced poly-As/poly-Ts

–When to suspect this:



Data cleaning



Poly-A Tails and Other Artifacts

–PRINSEQ (Schmieder 2011) for trimming poly-Ts – takes a % of the read that contains T's and sorts them out

Conservatively, 60% of a read is T?
Kick it out.

Filter on % base, sequence complexity, duplicates

Schmieder R and Edwards R: Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 2011, 27:863-864. [PMID: 21278185]

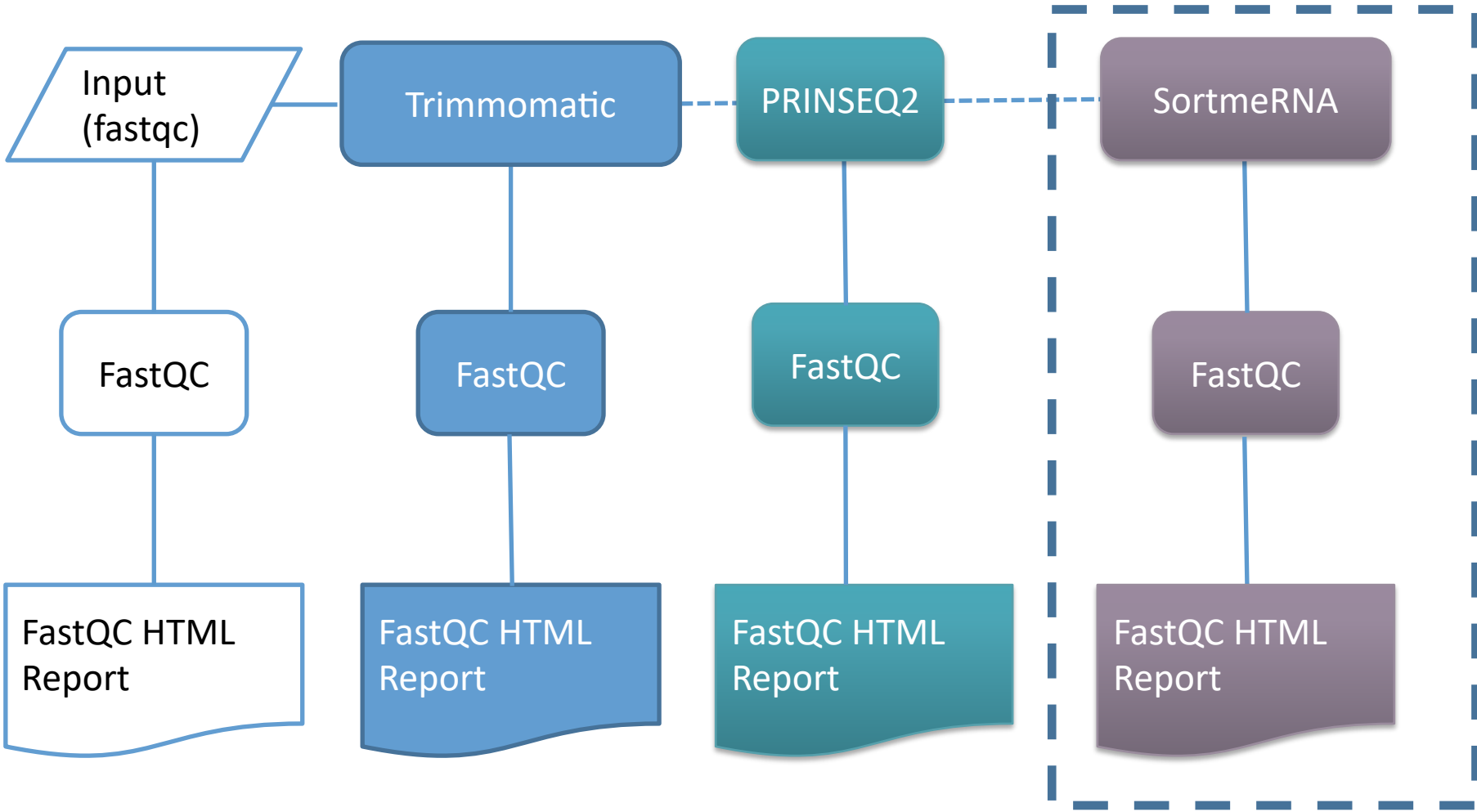
PRINSEQ step 2

- Trimming poly A/T tails
  - From 5'-end and 3'-end
- Filtering low complexity sequences
  - Entropy < 70 (out of 100)

Entropy < 50

- Filtering short reads (< 50 nu)

Data cleaning



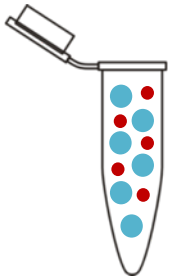
Contaminations

Euphausia superba (Uwe Kils. 2011)

Contaminations

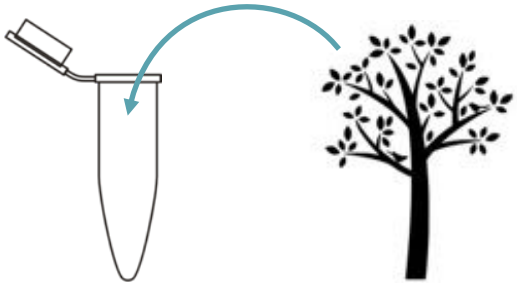
Euphausia superba (Uwe Kils. 2011)

Contaminations



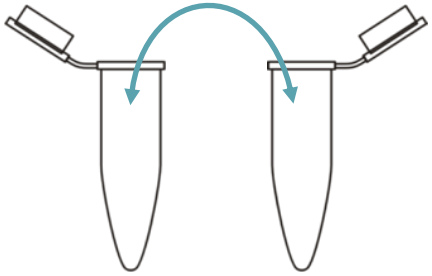
in-contamination

for ex. rRNA



third-party contamination

for ex. food - parasite



cross-contamination

for ex. experiment

- Most of (all) Illumina sequencing dataset are somewhat contaminated
- Illumina sequencing is especially susceptible to contamination due to the coverage depth
- It seems inherent to the method
- “Index misassignment between multiplexed libraries is a known issue” (Illumina, Inc., 2018); it potentially can produce contaminations in the sequenced datasets

Thanks to Serafim Nenarokov

rRNA contamination

One of the most common contamination

90-95% of total RNA correspond to rRNA

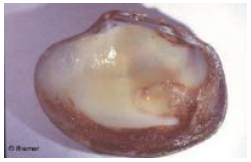
Hopefully it belongs to the sequenced organism but can also belongs to symbiont parasite or …. Aliens

rRNA contamination

One of the most common contamination

90-95% of total RNA correspond to rRNA

Hopefully it belongs to the sequenced organism but can also belongs to symbiont parasite or …. Aliens



Ruditapes philippinarui

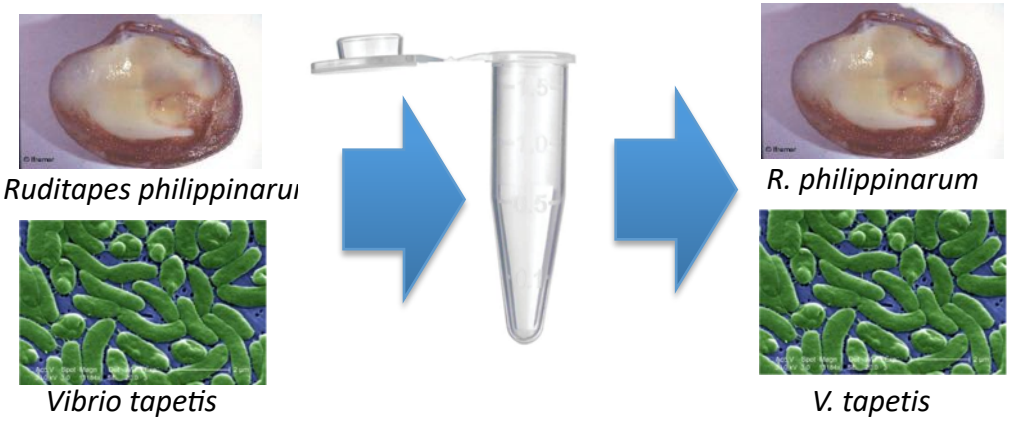
Vibrio tapetis

rRNA contamination

One of the most common contamination

90-95% of total RNA correspond to rRNA

Hopefully it belongs to the sequenced organism but can also belongs to symbiont parasite or …. Aliens

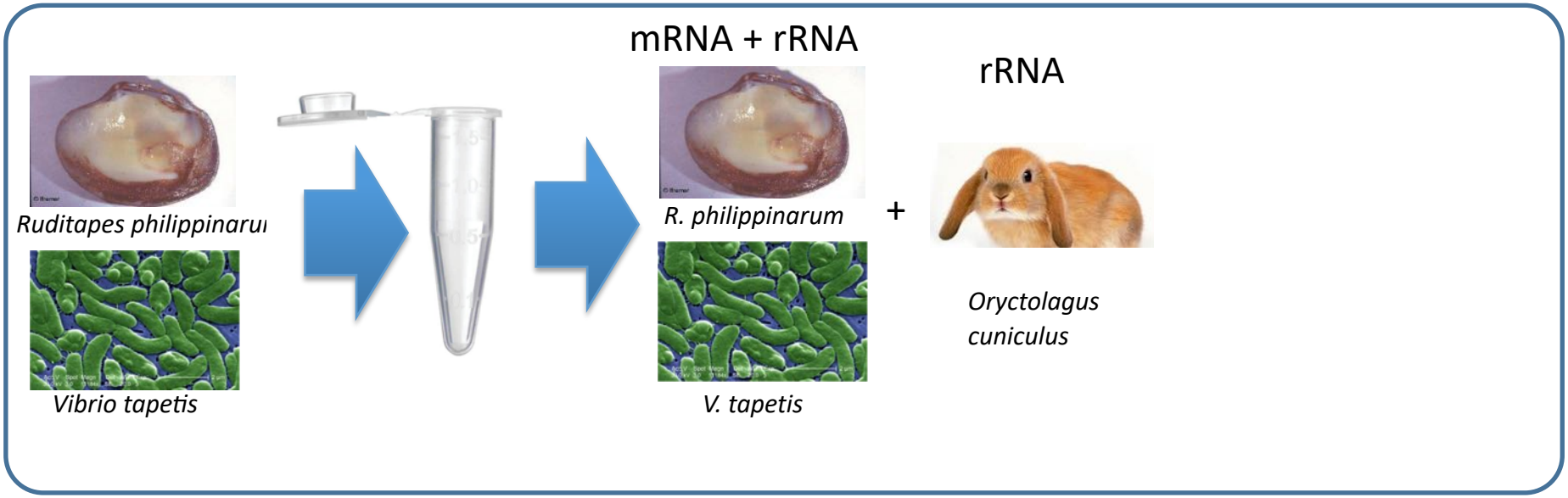


rRNA contamination

One of the most common contamination

90-95% of total RNA correspond to rRNA

Hopefully it belongs to the sequenced organism but can also belongs to symbiont parasite or …. Aliens

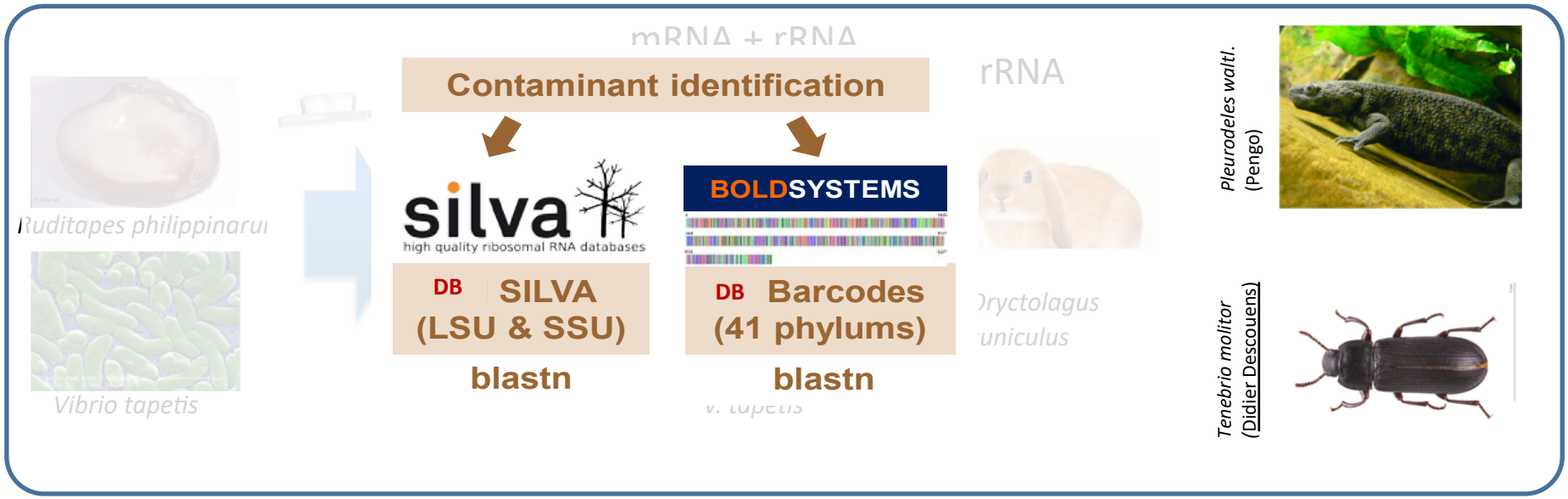


rRNA contamination

One of the most common contamination

90-95% of total RNA correspond to rRNA

Hopefully it belongs to the sequenced organism but can also belongs to symbiont parasite or …. Aliens

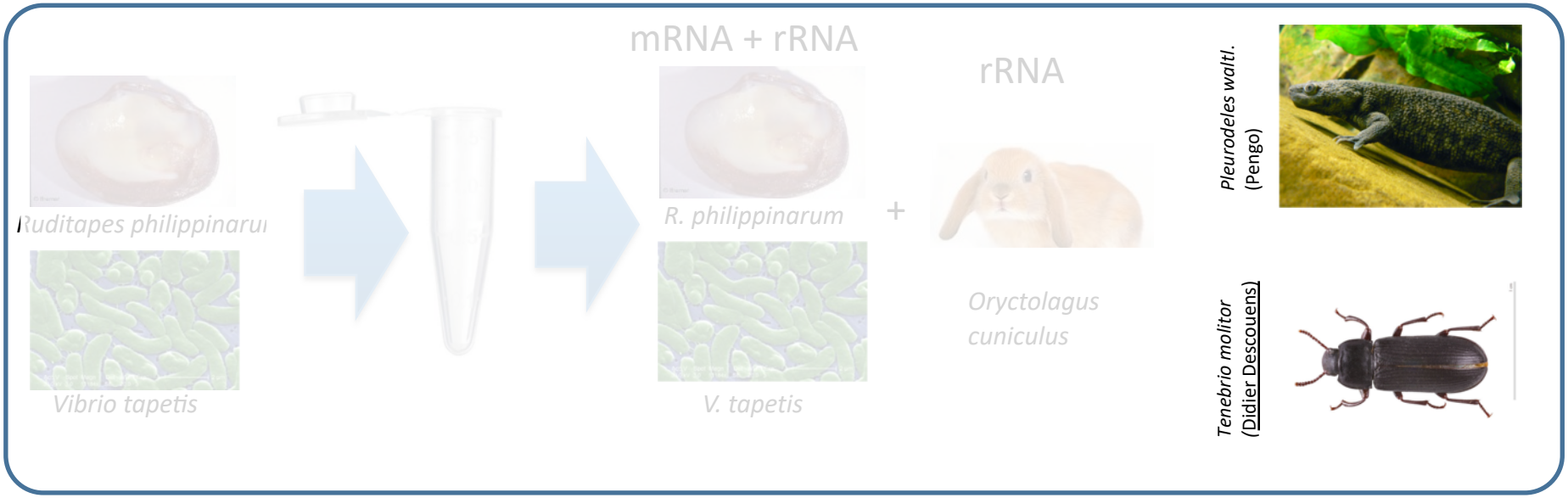


rRNA contamination

One of the most common contamination

90-95% of total RNA correspond to rRNA

Hopefully it belongs to the sequenced organism but can also belongs to symbiont parasite or …. Aliens



Solutions:

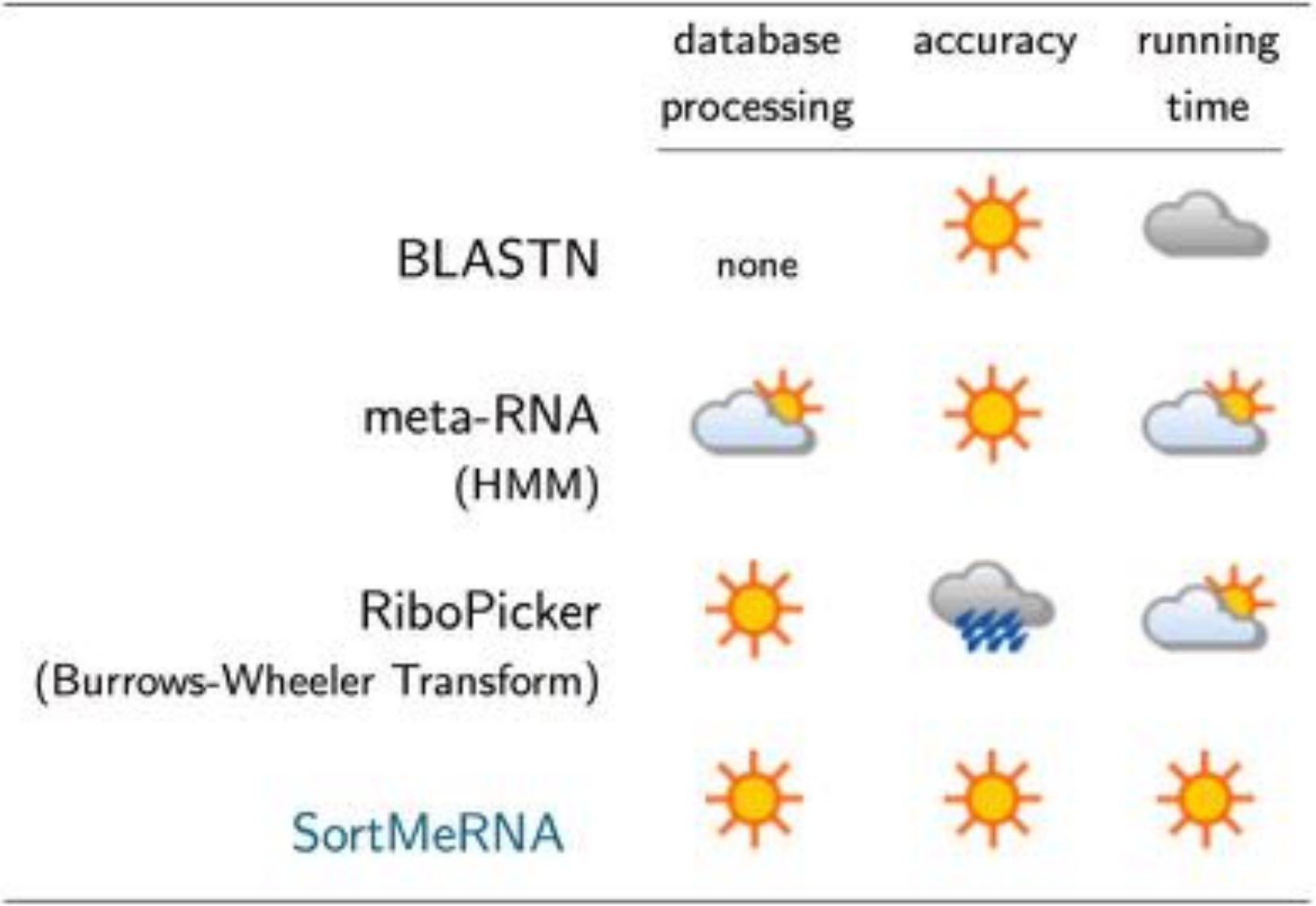
Prior to sequencing :

- Ribodepletion kits
- Selection polyA

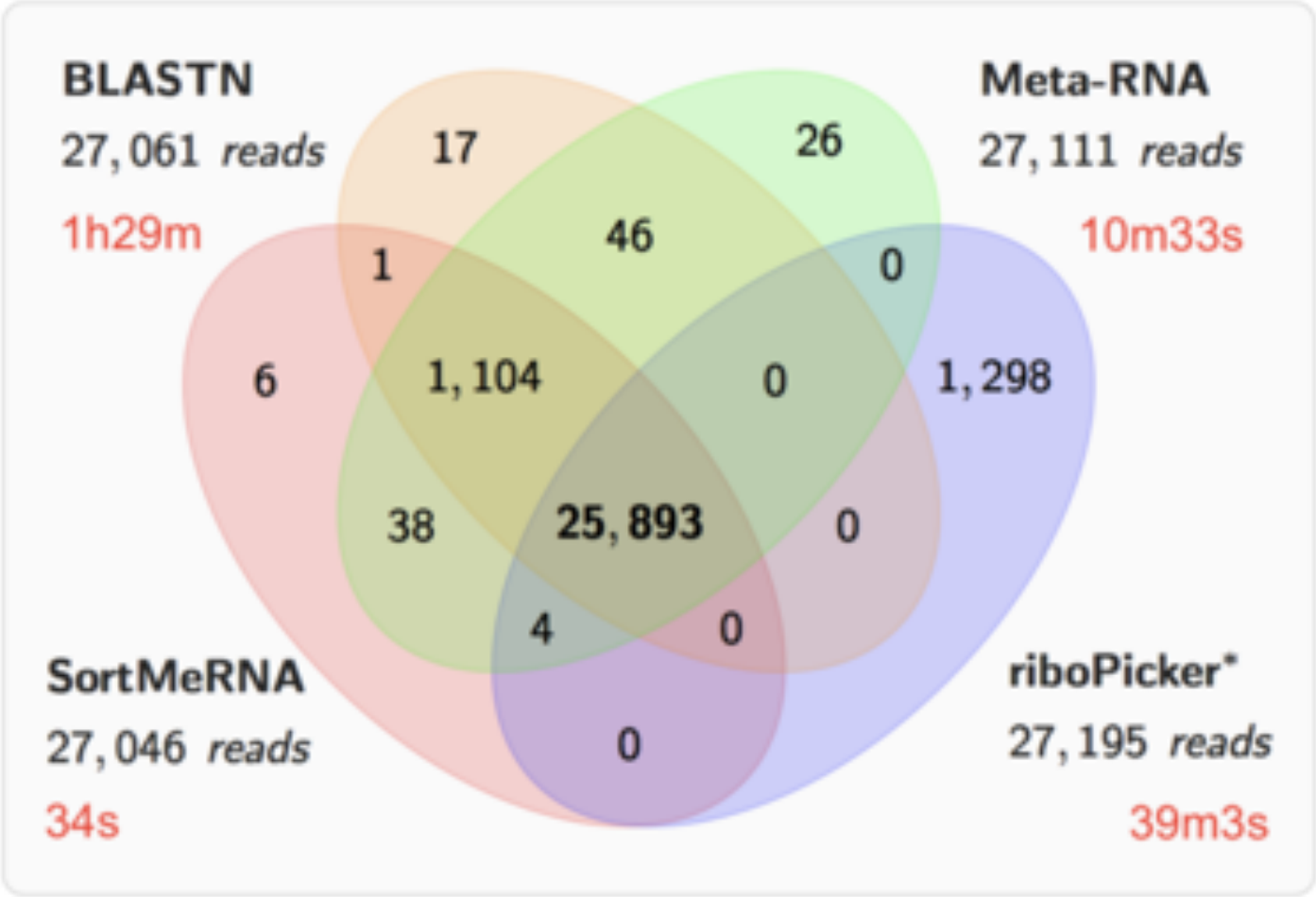
After sequencing :

- Remove rRNA reads from raw reads
- Detect rRNA transcripts

SortMeRNA/ribopicker/…

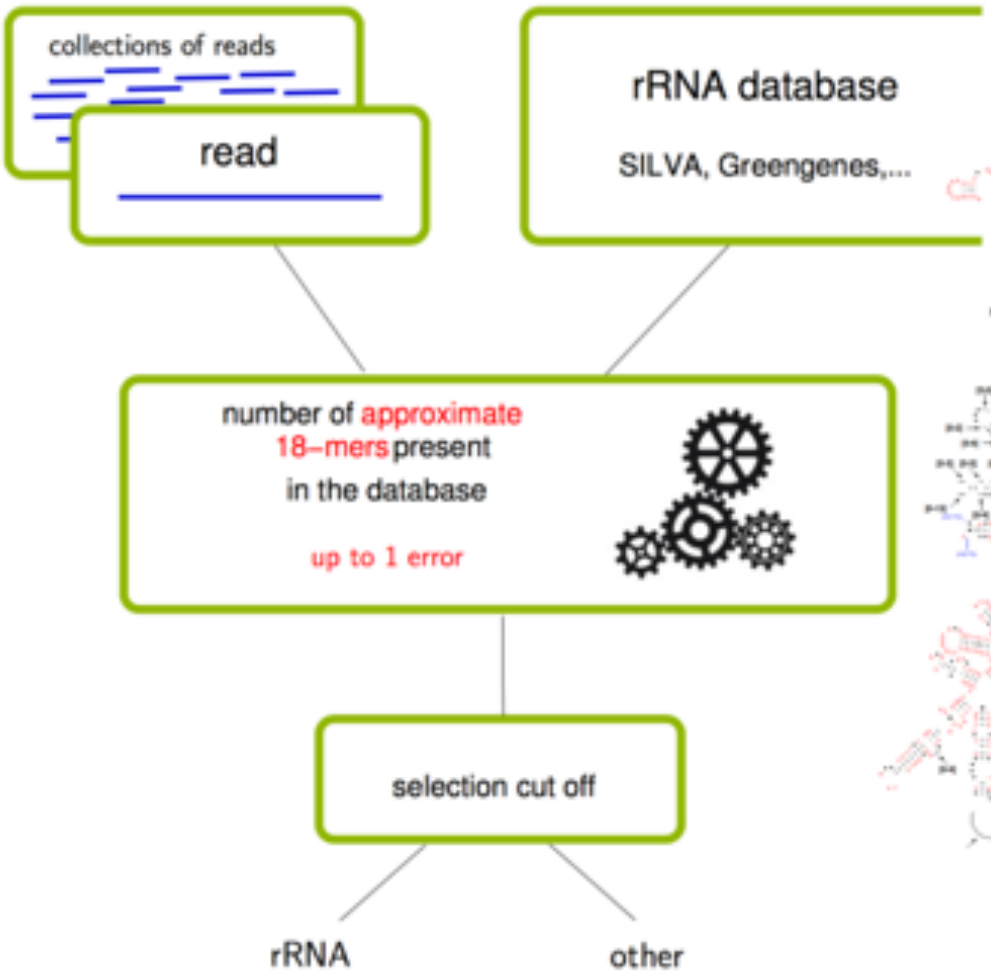


SortMeRNA/ribopicker/...

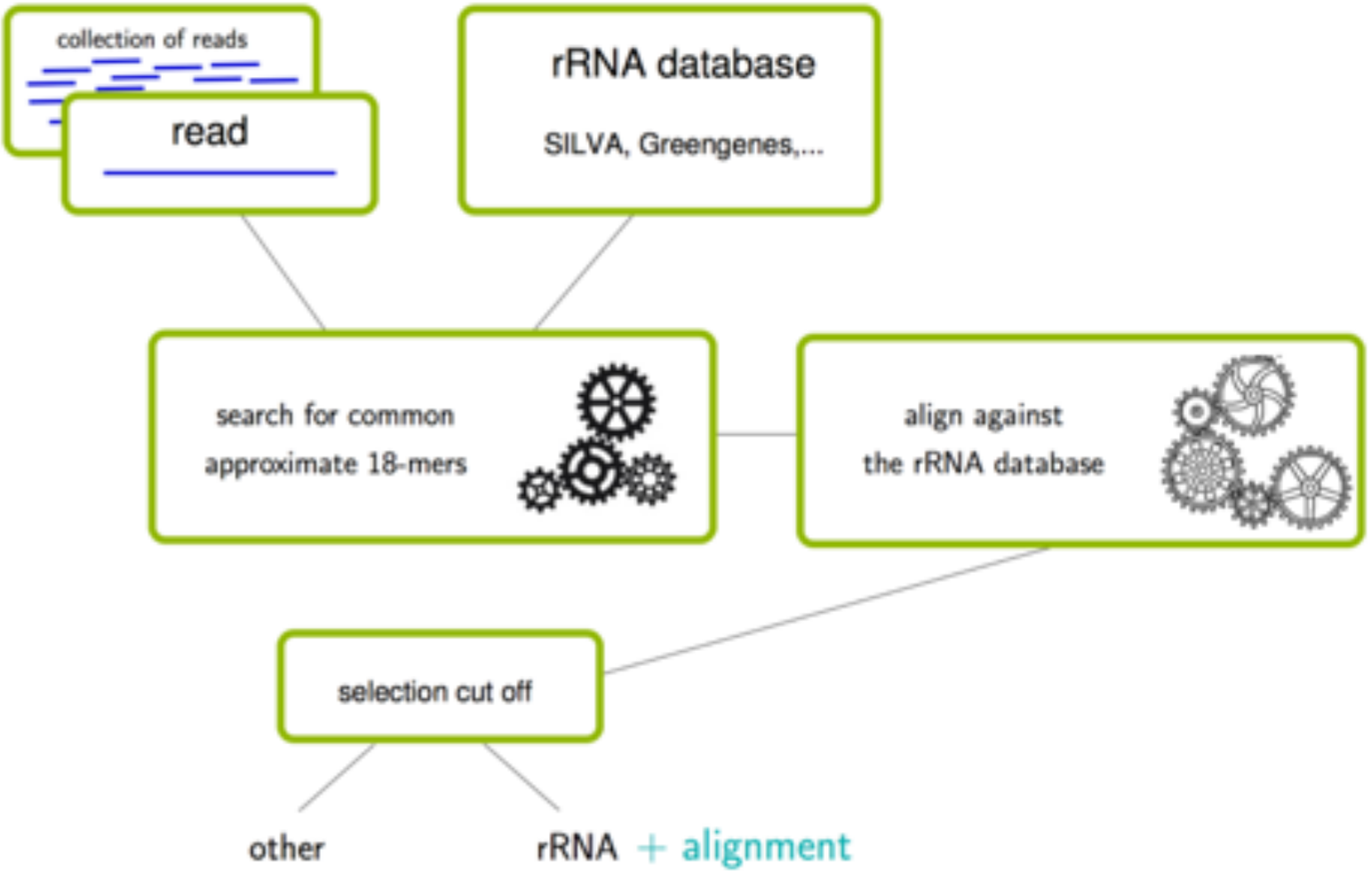


105,873 454 reads from photosynthetic metatranscriptome (SRR106861)

SortMeRNA


collections of reads
read
rRNA database
SILVA, Greengenes,...
number of approximate
18–mers present
in the database
up to 1 error
selection cut off
rRNA
other

SortMeRNA


collection of reads
read
rRNA database
SILVA, Greengenes,...
search for common
approximate 18-mers
align against
the rRNA database
selection cut off
other
rRNA + alignment

SortMeRNA commands

```
> merge-paired-reads.sh read_1.fq read_2.fq read-
interleaved.fq
```

```
>sortmerna --fastx -a 4 --log --paired_out -e 0.1 --id
0.97 --coverage 0.97
\--ref silva-bac-16s-id90.fasta,silva-bac-16s-id90:
\silva-bac-23s-id98.fasta,silva-bac-23s-id98:
\silva-euk-18s-id95.fasta,silva-euk-18s-id95:
\silva-euk-28s-id98.fasta,silva-euk-28s-id98:
\rfam-5s-database-id98.fasta,rfam-5s-database-id98:
\rfam-5.8s-database-id98.fasta,rfam-5.8s-database-id98
--reads read-interleaved.fq --other output_mRNA.fastq
fastq --aligned output_aligned.fastq
```

Reference DB

```
>unmerge-paired-reads.sh output_mRNA.fastq read-
sortmerna_1.fq read-sortmerna_2.fq
```

SortMeRNA results

```
Results:
Total reads = 34 196 864
Total reads for de novo clustering = 4 084 914
Total reads passing E-value threshold = 30 122 173 (88.08%)
Total reads failing E-value threshold = 4 074 691 (11.92%)
Minimum read length = 150
Maximum read length = 150
Mean read length = 150
By database:
silva-bac-16s-id90.fasta          6.95%
silva-bac-23s-id98.fasta          18.75%
silva-euk-18s-id95.fasta          9.97%
silva-euk-28s-id98.fasta          52.42%
rfam-5s-database-id98.fasta       0.00%
rfam-5.8s-database-id98.fasta     0.00%

Total reads passing %id and %coverage thresholds = 26 037 259
```

rRNA detection and assignation

- Assemble rRNA reads : Trinity, etc …
- Similarity search against : nr, Greengene, SILVA

- Detect rRNA in *denovo* assembly
  - Blast
  - RNAMMER

Detect rRNA transcripts : RNAMMER

The program uses hidden Markov models trained on data from the 5S ribosomal RNA database and the European ribosomal RNA database project

```
# ---------------------------------------------------------------------------------------------------------
##gff-version2##source-version RNAmmer-1.2##date 2009-11-16
##Type DNA
# seqname     source      feature  start       end          score        +/-  frame  attribute
# ---------------------------------------------------------------------------------------------------------
AE000511      RNAmmer-1.2  rRNA   448462       448577       49.2         +      .      5s_rRNA
AE000511      RNAmmer-1.2  rRNA   1473564      1473679      49.2         -      .      5s_rRNA
AE000511      RNAmmer-1.2  rRNA   1045067      1045183      40.3         +      .      5s_rRNA
AE000511      RNAmmer-1.2  rRNA   445339       448223       3056.5       +      .      23s_rRNA
AE000511      RNAmmer-1.2  rRNA   1473918      1476803      3032.8       -      .      23s_rRNA
AE000511      RNAmmer-1.2  rRNA   1207586      1209074      1801.4       -      .      16s_rRNA
AE000511      RNAmmer-1.2  rRNA   1511140      1512627      1803.6       -      .      16s_rRNA
```

Lagesen K, Hallin PF, Rødland E, Stærfeldt HH, Rognes T Ussery DW **RNammer: consistent annotation of rRNA genes in genomic sequences**
Nucleic Acids Res. 2007 Apr 22.

Alternative Barnap :
https://github.com/tseemann/barrnap

RNAmmer - Barnap

```
> Trinotate-3.0.1/util/rnammer_support/RnammerTranscriptome.pl
--transcriptome Assembly.fasta  –org_type (arc|bac|euk) --
path_to_rnammer /usr/local/genome2/rnammer/rnammer

>bedtools getfasta -fi Assembly.fasta -bed
rnammer_predictions.gff > trancripts_rrna.fasta
```

```
> /usr/local/genome2/barrnap-master2/bin/barrnap --kingdom bac
--threads 10 --outfasta rrna_bact.fasta Assembly.fasta
```

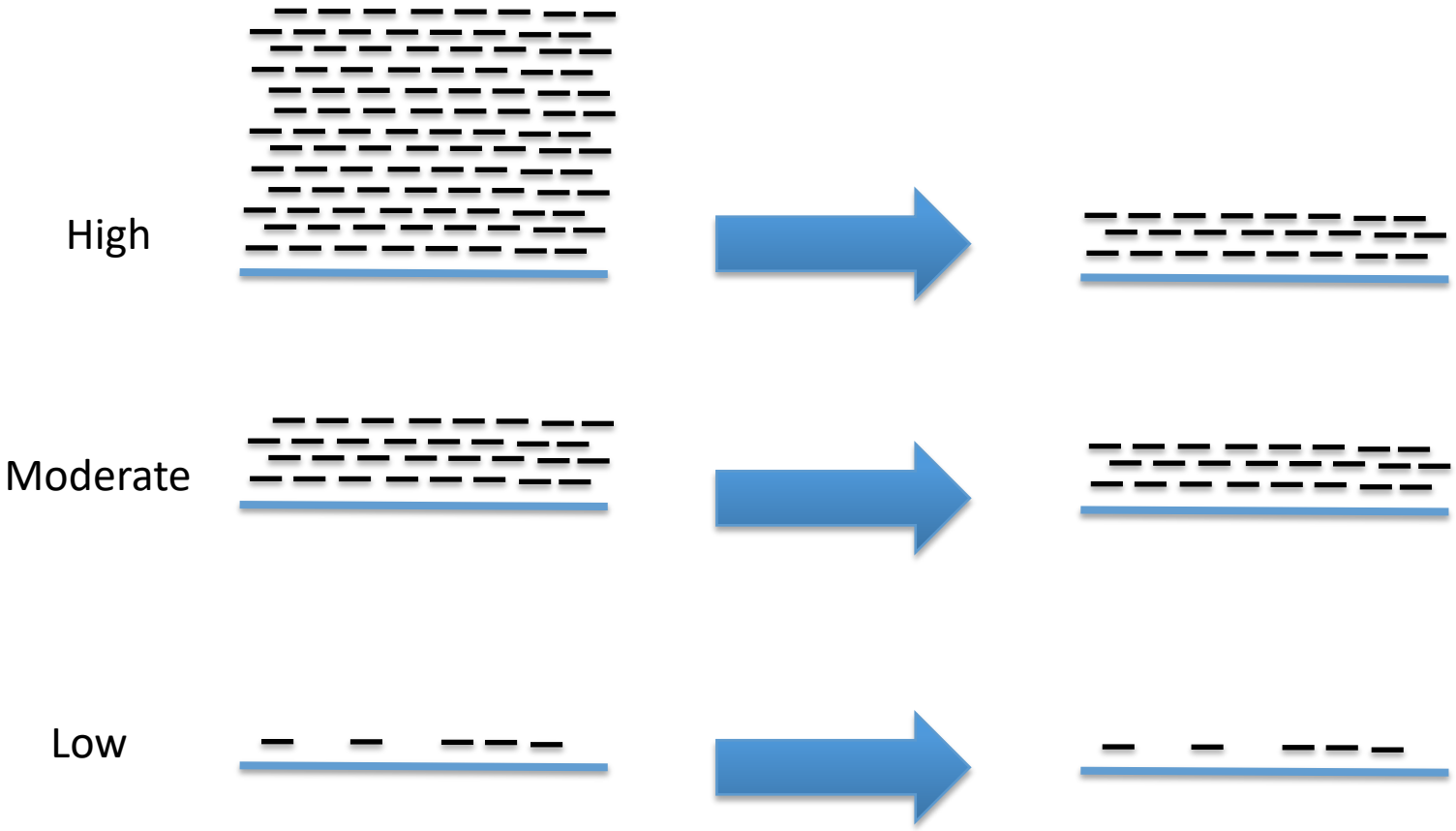
Digital Normalization

NGS reads normalization (by Trinity)

- Context:
  - By definition RNAseq display a wide range of expressions Very low expressed → Very highly expressed transcripts
  - The information given by reads from high expression transcripts is redundant, and very high coverage also brings more sequencing errors
  - De-novo assemblers do not benefit from coverage increase beyond a certain point, and fewer data means quicker assemblies

➔How to decrease coverage of highly expressed transcripts without decreasing that of low expressed transcripts ?

In silico normalization of reads



NGS reads normalization (by Trinity)

1. Count kmers in all the data (Jellyfish):

e.g. for k = 5

>
CAGTCGATCA

>
CGATCAGTCG

NGS reads normalization (by Trinity)

1. Count kmers in all the data (Jellyfish):

e.g. for k = 5

>
CAGTCGATCA

>
CGATCAGTCG

CAGTC	1

NGS reads normalization (by Trinity)

1. Count kmers in all the data (Jellyfish):

e.g. for k = 5

>
CAGTCGATCA

>
CGATCAGTCG

CAGTC	1
AGTCG	1

NGS reads normalization (by Trinity)

1. Count kmers in all the data (Jellyfish):

e.g. for k = 5

>
CAGTCGATCA

>
CGATCAGTCG

CAGTC	1
AGTCG	1
GTCGA	1

NGS reads normalization (by Trinity)

1. Count kmers in all the data (Jellyfish):

e.g. for k = 5

>
CAGTCGATCA

>
CGATCAGTCG

CAGTC	1
AGTCG	1
GTCGA	1
TCGAT	1

NGS reads normalization (by Trinity)

1. Count kmers in all the data (Jellyfish):

e.g. for k = 5

>
CAGTCGATCA

>
CGATCAGTCG

CAGTC	1
AGTCG	1
GTCGA	1
TCGAT	1
CGATC	1

NGS reads normalization (by Trinity)

1. Count kmers in all the data (Jellyfish):

e.g. for k = 5

>
CAGTCGATCA

>
CGATCAGTCG

CAGTC	1
AGTCG	1
GTCGA	1
TCGAT	1
CGATC	1
GATCA	1

NGS reads normalization (by Trinity)

1. Count kmers in all the data (Jellyfish):

e.g. for k = 5

>
CAGTCGATCA

>
CGATCAGTCG

CAGTC	1
AGTCG	1
GTCGA	1
TCGAT	1
CGATC	2
GATCA	1

NGS reads normalization (by Trinity)

1. Count kmers in all the data (Jellyfish):

e.g. for k = 5

>
CAGTCGATCA

>
CGATCAGTCG

CAGTC	1
AGTCG	1
GTCGA	1
TCGAT	1
CGATC	2
GATCA	2

NGS reads normalization (by Trinity)

1. Count kmers in all the data (Jellyfish):

e.g. for k = 5

>
CAGTCGATCA

>
CGATCAGTCG

CAGTC	1
AGTCG	1
GTCGA	1
TCGAT	1
CGATC	2
GATCA	2
ATCAG	1

NGS reads normalization (by Trinity)

1. Count kmers in all the data (Jellyfish):

e.g. for k = 5

>
CAGTCGATCA

>
CGATCAGTCG

CAGTC	1
AGTCG	1
GTCGA	1
TCGAT	1
CGATC	2
GATCA	2
ATCAG	1
TCAGT	1

NGS reads normalization (by Trinity)

1. Count kmers in all the data (Jellyfish):

e.g. for k = 5

>
CAGTCGATCA

>
CGATCAGTCG

CAGTC	2
AGTCG	1
GTCGA	1
TCGAT	1
CGATC	2
GATCA	2
ATCAG	1
TCAGT	1

NGS reads normalization (by Trinity)

1. Count kmers in all the data (Jellyfish):

e.g. for k = 5

>
CAGTCGATCA

>
CGATCAGTCG

CAGTC	2
AGTCG	2
GTCGA	1
TCGAT	1
CGATC	2
GATCA	2
ATCAG	1
TCAGT	1

NGS reads normalization (by Trinity)

1. Count kmers in all the data (Jellyfish):

e.g. for k = 5

>
CAGTCGATCA

>
CGATCAGTCG

CAGTC	2
AGTCG	2
GTCGA	1
TCGAT	1
CGATC	2
GATCA	2
ATCAG	1
TCAGT	1
...	

NGS reads normalization (by Trinity)

1. Count kmers in all the data (Jellyfish):
   - with k = 25

2. For each read, compute the median, average and stdev kmers coverage

NGS reads normalization (by Trinity)

1. Count kmers in all the data (Jellyfish):
   - with k = 25

2. For each read, compute the median, average and stdev kmers coverage

3. Accept a read with a probability of:

NGS reads normalization (by Trinity)

3. Accept a read with a probability of:

e.g. with $max\ coverage = 30$

Read_A: $median\ coverage = 60$ → $\frac{max\_coverage}{median} = 0.5$

→ Read_A has a 50% chance of being kept

Read_B: $median\ coverage = 10$ → $\frac{max\_coverage}{median} = 3$

→ Read_B has a 300% chance of being kept ;-)

→ Read_B will be kept

NGS reads normalization (by Trinity)

3. Accept a read with a probability of:

Read_A comes from a highly expressed transcript and is 2 times more covered than the threshold. We know its information is also contained by other reads.

➔ So it has less chance to be kept.

Read_B comes from a low expressed transcript, way below the threshold. Its information is not very redondant, we will need it for the assembly.

➔ So it will absolutly be kept

NGS reads normalization (by Trinity)

1. Count kmers in all the data (Jellyfish):
   - with k = 25

2. For each read, compute the `median`, `average` and `stdev` kmers coverage

3. Accept a read with a probability of: *`maxcov`*/*`median`*

4. Remove a read if: *`standartdev`*/*`average`* (CV)> 1 (100%)

A high variability in a read kmer coverage means there is probably a lot of sequencing errors in this read

NGS reads normalization (by Trinity)

- Pros:
  - Reduce the data to be assembled
    → faster assemblies
    → RAM requirement highly reduced
  - Remove reads with potentially lots of sequencing errors
    → better assemblies ?

- Cons:
  - Small loss of information → slightly worse assemblies ?
  - Stringent filter on kmer coverage variability
    → loss of low expressed alternative transcripts (splice junctions) ?

  Trinity normalisation procedure quite greedy

  → Use khmer instead (https://github.com/dib-lab/khmer)

Stand alone normalisation

```
$TRINITY_HOME/util/insilico_read_normalization.pl
\ --seqType fq --JM 1G --max_cov 50
\ --left lib1_1.P.qtrim --right lib2_2.P.qtrim
\ --pairs_together --output insil_norm_ex
```

```
1189570 / 1879312 = 63.30% reads selected during normalization.
1094 / 1879312 = 0.06% reads discarded as likely aberrant based on
coverage profiles.

Normalization complete. See outputs:
insil_norm_ex/lib1_1.P.qtrim.normalized_K25_C50_pctSD200.fq
insil_norm_ex/lib1_2.P.qtrim.normalized_K25_C50_pctSD200.fq
```

Trinity normalisation