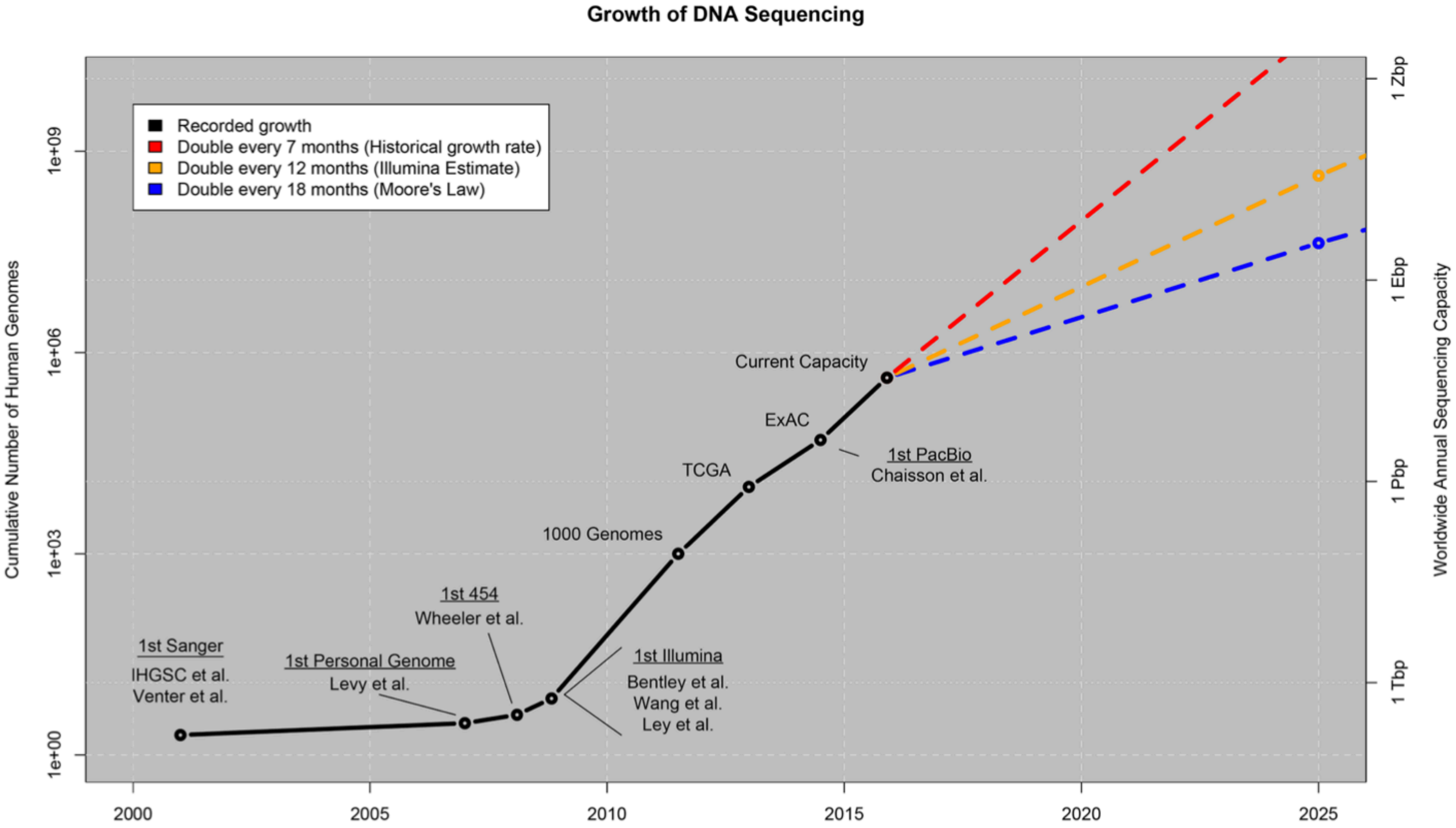
ETH zürich University of Zurich UZH functional genomics center zurich f g c z

Introduction to NGS

Hubert Rehrauer

University of Zurich UZH
ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

NGS Data Increase



- NGS data increases faster than computer speed

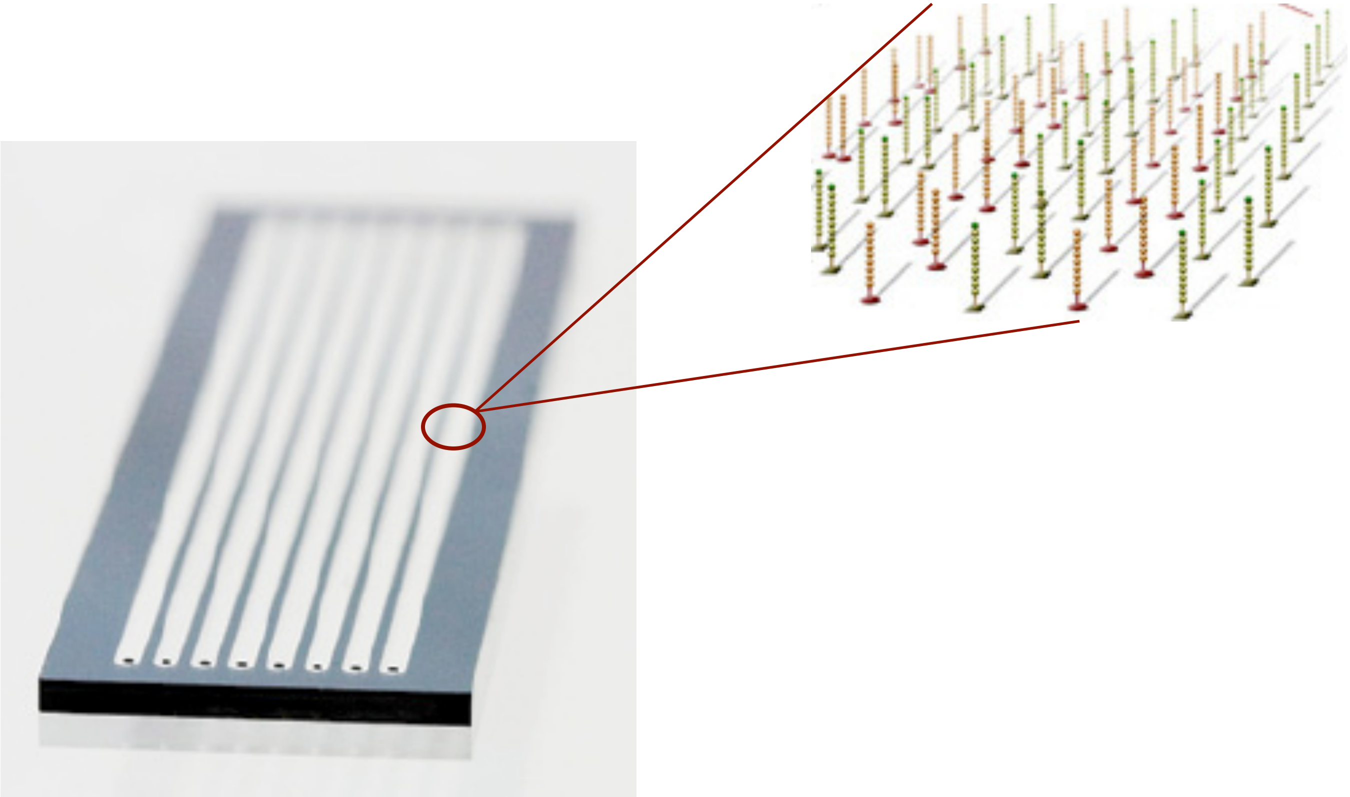
Stephens ZD etal. (2015) Big Data: Astronomical or Genomical? PLoS Biol 13(7): e1002195.

Ingredients for the success

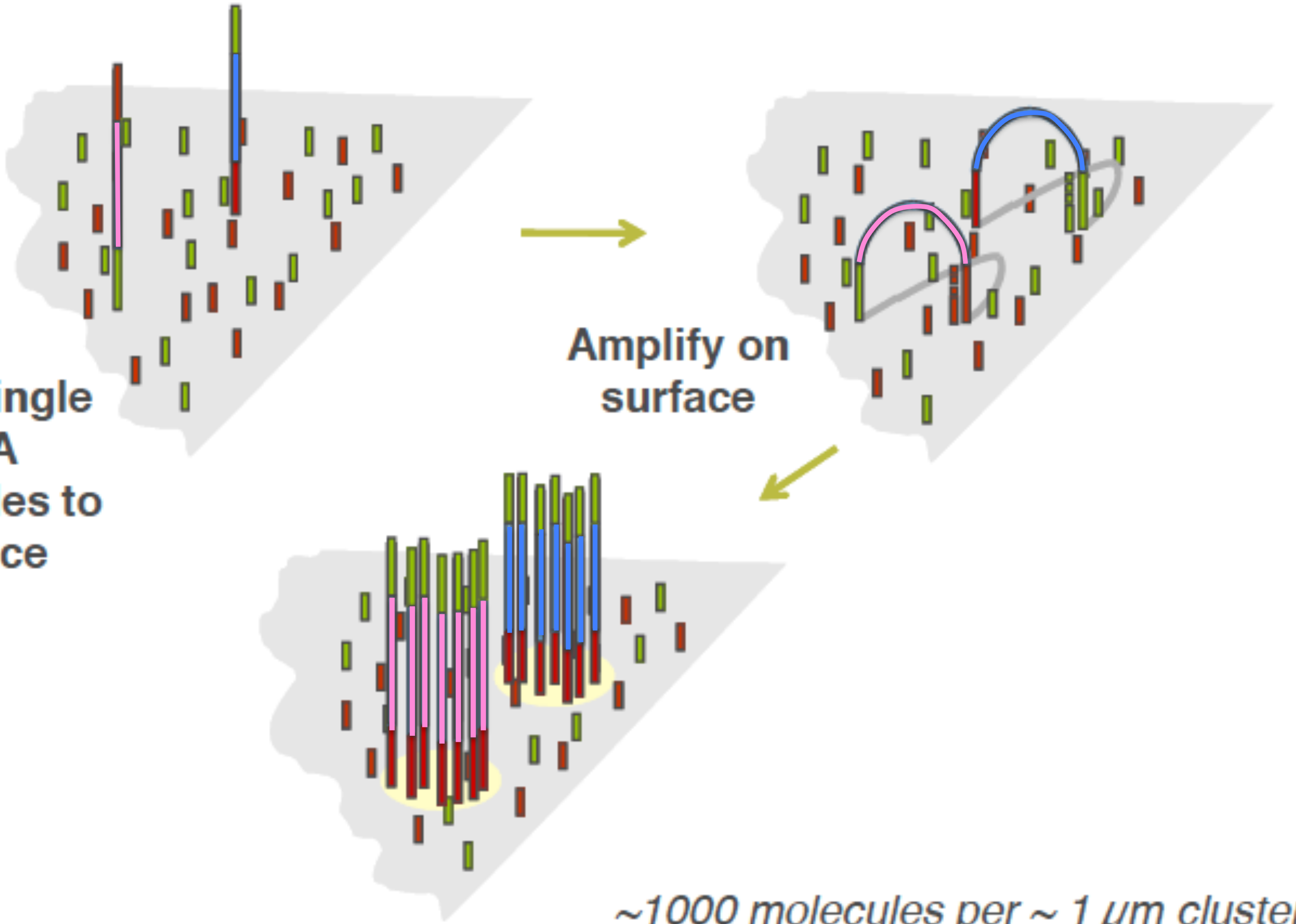
- Evolution has yielded DNA and RNA molecules for information storage and transfer. They have good properties to be read (**measured**)
- NGS technologies rely on
  - **massive parallelization**
  - measurement process is done by individual molecules (**cheap and fast**)

2nd generation sequencers	3rd generation sequencers	
Illumina	Pac Bio	optical
Ion Torrent	Oxford Nanopores	non- optical

Illumina Flow cell



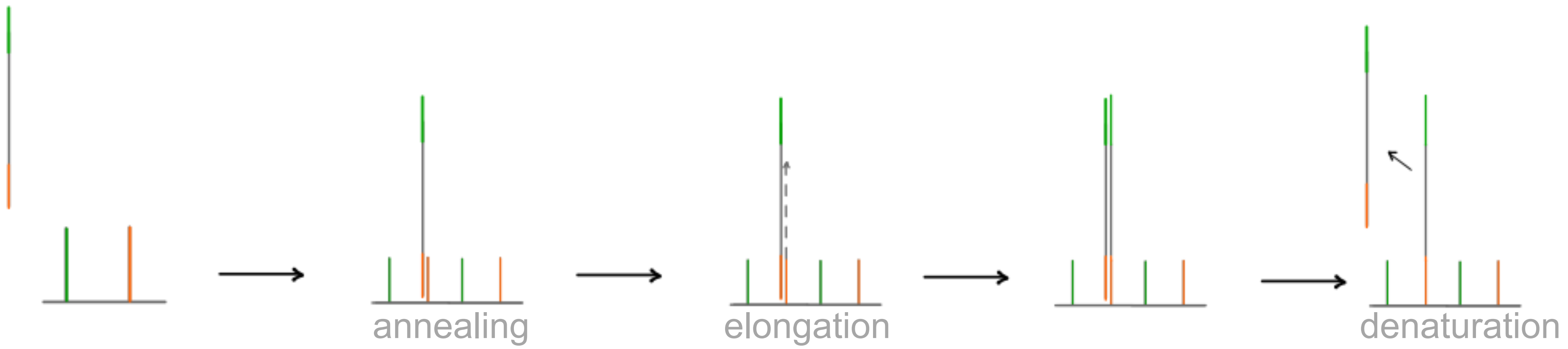
Cluster generation overview



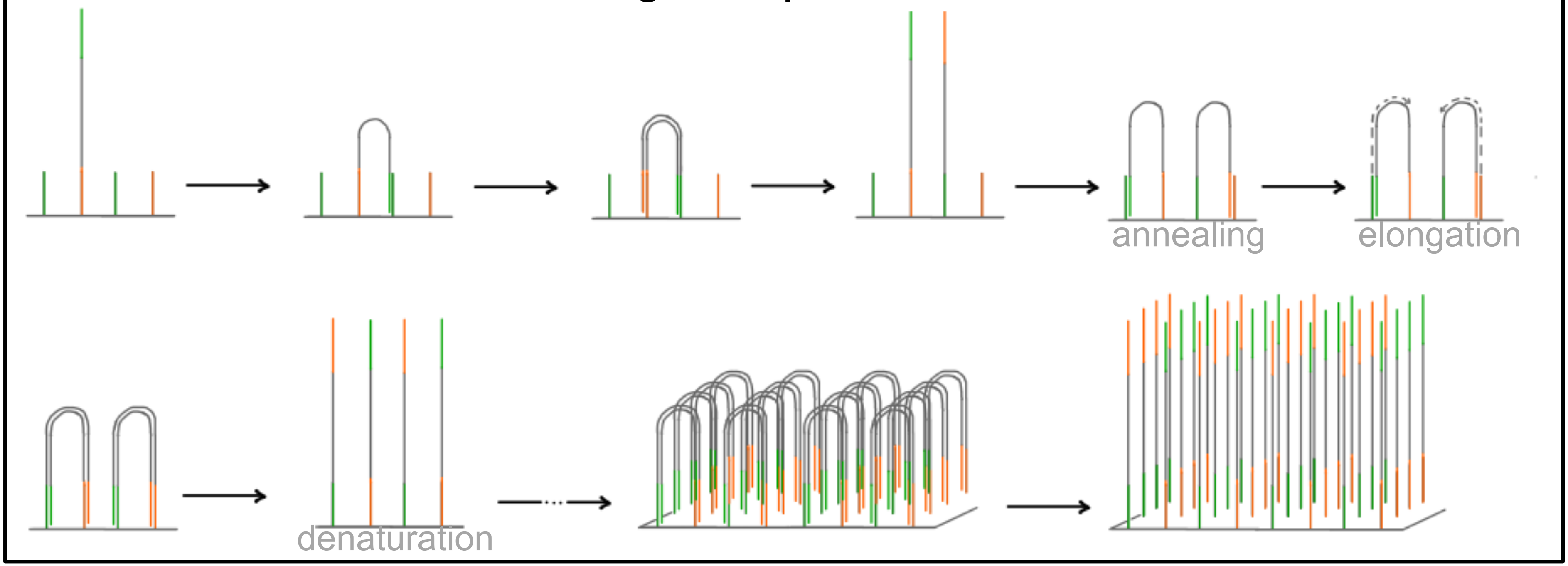
Bind single DNA molecules to surface

Amplify on surface

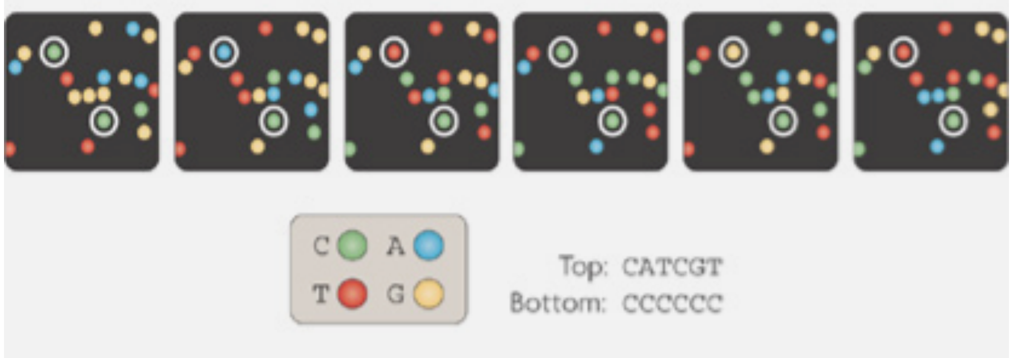
~1000 molecules per ~ 1 µm cluster

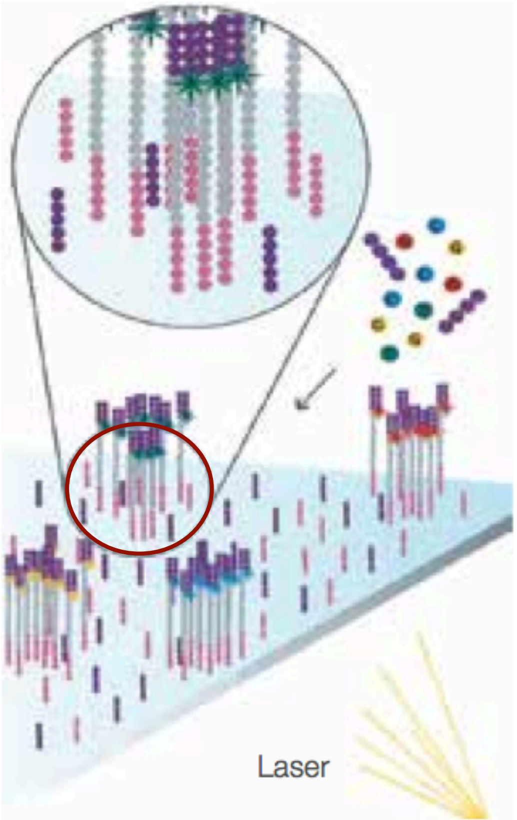


bridge amplification

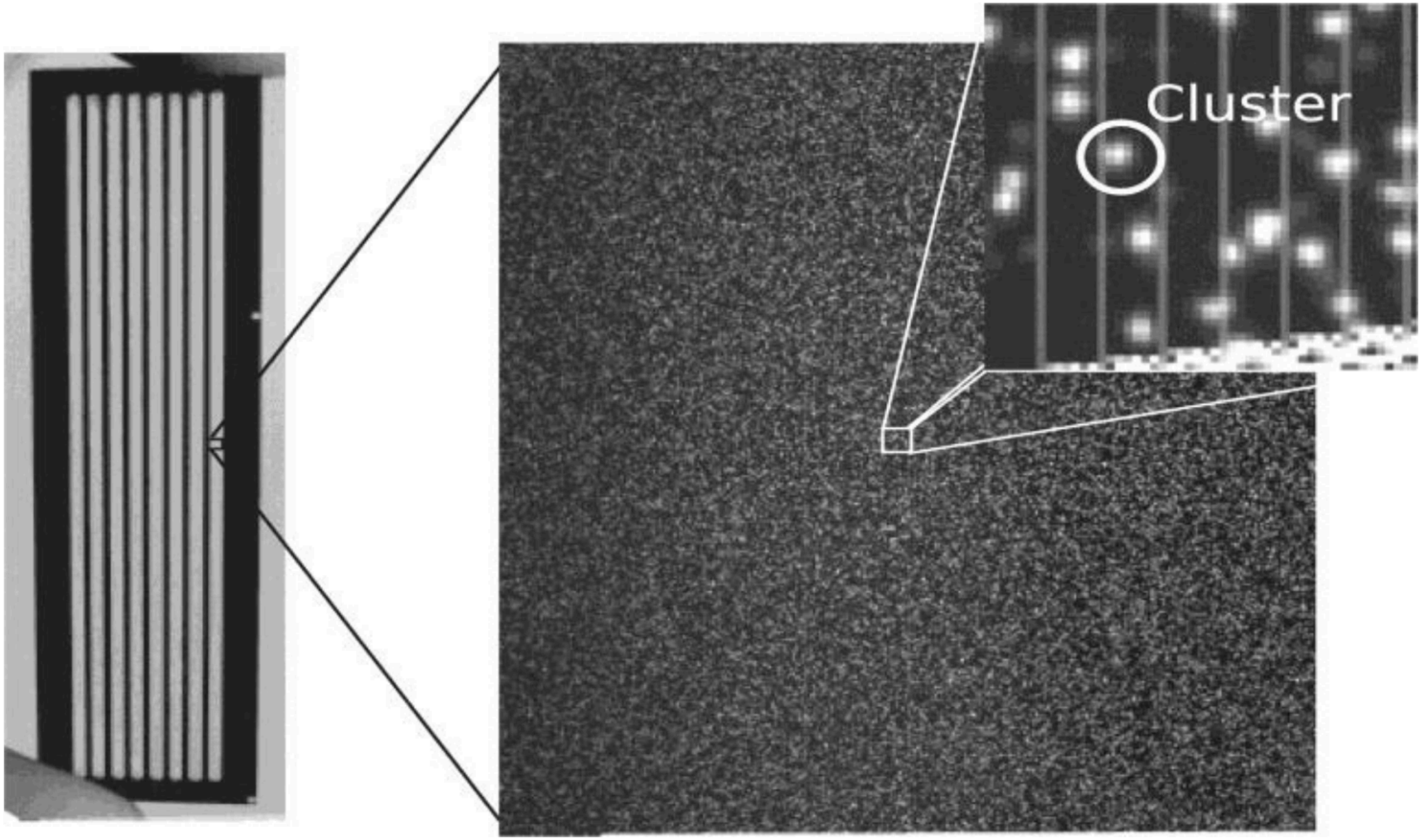


Illumina Sequencing



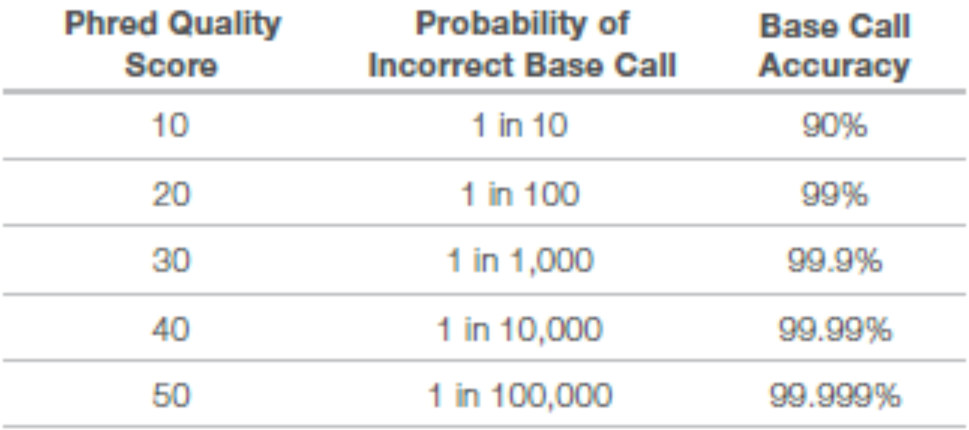


The first sequencing cycle begins by adding four labeled reversible terminators, primers, and DNA polymerase.


Cluster

Phred scores measure base call accuracy

- P
  - error probability of a given base call
- *Q*
  - *$-10\log_{10}P$*
- Assign to each base
- Range from 0-41



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

Ewing B, Green P. 1998. Genome Res. 8(3):186-194.

http://en.wikipedia.org/wiki/Phred_quality_score

Phred scores are stored with sequences

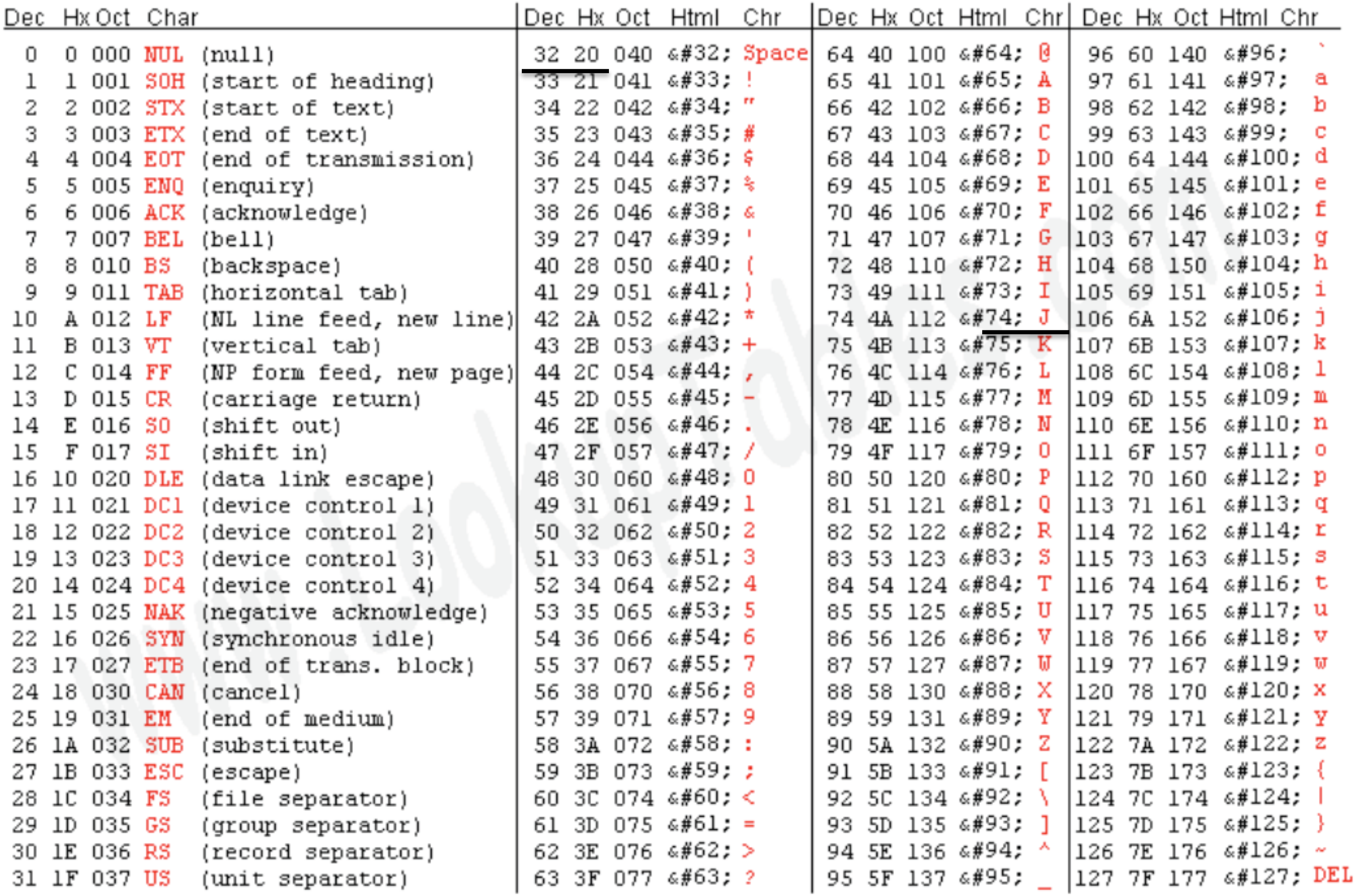
- FASTQ
  - 4 lines:
    1. Header line for Read (starts with “@” and the sequence ID)
    2. Sequence
    3. Header line for Qualities (starts with “+”)
    4. Quality score (represented in ASCII format)

```
@HWI-ST1034:40:C08PJACXX:2:1101:20681:1994 1:N:0:ATCACG

CTCGNAGACTGGCAACTTGTTCTGGTTTACTGCACCTTCTTTTAAAGGCAGAAAGGC

+

CCCF#2ADHHHGHJJJIJJHIIIJJHIJJJJJJJJJBGIIJJJJJJJJJJJJJJJIJ
```

Phred scores can be **ASCII encoded**

- Add an offset and convert the sum to ASCII

- Current format
  - **Illumina 1.9 (i.e. Sanger format)**
  - Phred scoring: 0-41;
  - Offset: 33
  - 41+33=74 (J)
  - All current sequencers



Dec	Hx	Oct	Char	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr
0	0	000	NUL (null)	32	20	040	&#32;	Space	64	40	100	&#64;	@	96	60	140	&#96;	`
1	1	001	SOH (start of heading)	33	21	041	&#33;	!	65	41	101	&#65;	A	97	61	141	&#97;	a
2	2	002	STX (start of text)	34	22	042	&#34;	"	66	42	102	&#66;	B	98	62	142	&#98;	b
3	3	003	ETX (end of text)	35	23	043	&#35;	#	67	43	103	&#67;	C	99	63	143	&#99;	c
4	4	004	EOT (end of transmission)	36	24	044	&#36;	$	68	44	104	&#68;	D	100	64	144	&#100;	d
5	5	005	ENQ (enquiry)	37	25	045	&#37;	%	69	45	105	&#69;	E	101	65	145	&#101;	e
6	6	006	ACK (acknowledge)	38	26	046	&#38;	&	70	46	106	&#70;	F	102	66	146	&#102;	f
7	7	007	BEL (bell)	39	27	047	&#39;	'	71	47	107	&#71;	G	103	67	147	&#103;	g
8	8	010	BS (backspace)	40	28	050	&#40;	(	72	48	110	&#72;	H	104	68	150	&#104;	h
9	9	011	TAB (horizontal tab)	41	29	051	&#41;	)	73	49	111	&#73;	I	105	69	151	&#105;	i
10	A	012	LF (NL line feed, new line)	42	2A	052	&#42;	*	74	4A	112	&#74;	J	106	6A	152	&#106;	j
11	B	013	VT (vertical tab)	43	2B	053	&#43;	+	75	4B	113	&#75;	K	107	6B	153	&#107;	k
12	C	014	FF (NP form feed, new page)	44	2C	054	&#44;	,	76	4C	114	&#76;	L	108	6C	154	&#108;	l
13	D	015	CR (carriage return)	45	2D	055	&#45;	-	77	4D	115	&#77;	M	109	6D	155	&#109;	m
14	E	016	SO (shift out)	46	2E	056	&#46;	.	78	4E	116	&#78;	N	110	6E	156	&#110;	n
15	F	017	SI (shift in)	47	2F	057	&#47;	/	79	4F	117	&#79;	O	111	6F	157	&#111;	o
16	10	020	DLE (data link escape)	48	30	060	&#48;	0	80	50	120	&#80;	P	112	70	160	&#112;	p
17	11	021	DC1 (device control 1)	49	31	061	&#49;	1	81	51	121	&#81;	Q	113	71	161	&#113;	q
18	12	022	DC2 (device control 2)	50	32	062	&#50;	2	82	52	122	&#82;	R	114	72	162	&#114;	r
19	13	023	DC3 (device control 3)	51	33	063	&#51;	3	83	53	123	&#83;	S	115	73	163	&#115;	s
20	14	024	DC4 (device control 4)	52	34	064	&#52;	4	84	54	124	&#84;	T	116	74	164	&#116;	t
21	15	025	NAK (negative acknowledge)	53	35	065	&#53;	5	85	55	125	&#85;	U	117	75	165	&#117;	u
22	16	026	SYN (synchronous idle)	54	36	066	&#54;	6	86	56	126	&#86;	V	118	76	166	&#118;	v
23	17	027	ETB (end of trans. block)	55	37	067	&#55;	7	87	57	127	&#87;	W	119	77	167	&#119;	w
24	18	030	CAN (cancel)	56	38	070	&#56;	8	88	58	130	&#88;	X	120	78	170	&#120;	x
25	19	031	EM (end of medium)	57	39	071	&#57;	9	89	59	131	&#89;	Y	121	79	171	&#121;	y
26	1A	032	SUB (substitute)	58	3A	072	&#58;	:	90	5A	132	&#90;	Z	122	7A	172	&#122;	z
27	1B	033	ESC (escape)	59	3B	073	&#59;	;	91	5B	133	&#91;	[	123	7B	173	&#123;	{
28	1C	034	FS (file separator)	60	3C	074	&#60;	<	92	5C	134	&#92;	\	124	7C	174	&#124;	|
29	1D	035	GS (group separator)	61	3D	075	&#61;	=	93	5D	135	&#93;	]	125	7D	175	&#125;	}
30	1E	036	RS (record separator)	62	3E	076	&#62;	>	94	5E	136	&#94;	^	126	7E	176	&#126;	~
31	1F	037	US (unit separator)	63	3F	077	&#63;	?	95	5F	137	&#95;	_	127	7F	177	&#127;	DEL

Source: www.LookupTables.com

Quality control is to know your reads

- Library construction could introduce bias
  - Fragmentation, ligation, amplification
  - GC bias
  - Over-amplification
  - Contamination

- Sequencing errors
  - Chemical, optical, computational

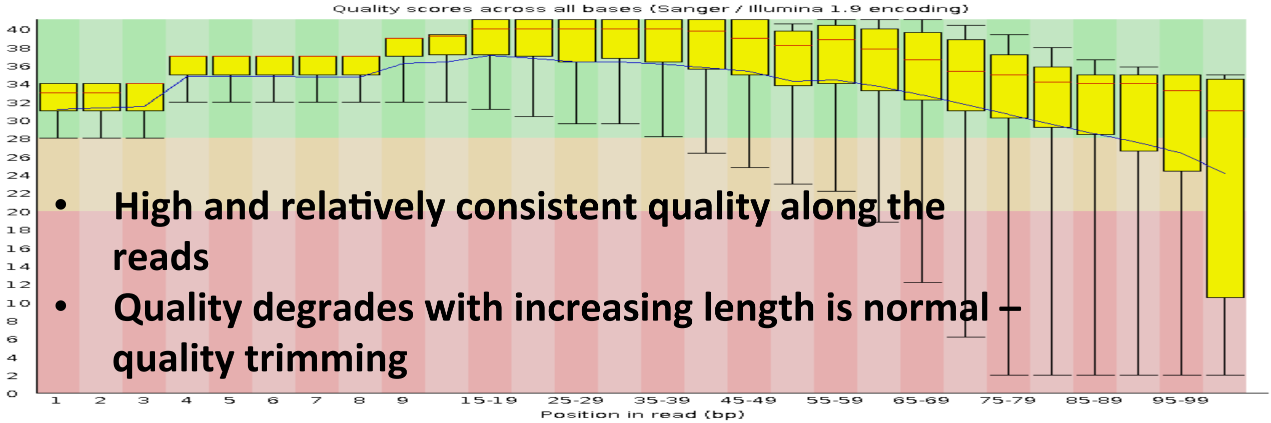
Platform	Primary error	Error rate (%)
Illumina	Substitution	0.1
PacBio	Indel	12
PGM	Indel	1
454	Indel	1
Oxford Nanopore	Indel	20-40

Glenn, TC. 2011. Molecular Ecology Resources.
Carneiro , MO. 2012. BMC Genomics

Per base sequence quality - FastQC

- Range of quality values across all bases at each position

☑



- **High and relatively consistent quality along the reads**
- **Quality degrades with increasing length is normal – quality trimming**

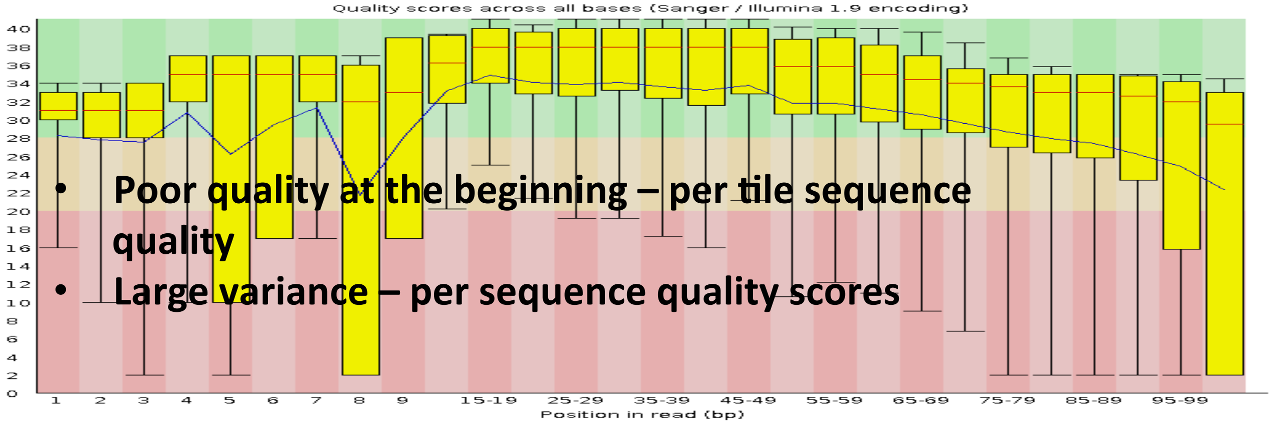
Green: >Q28, good

Orange: >Q20, reasonable

Red:<Q20, poor

Median > Q25

☒



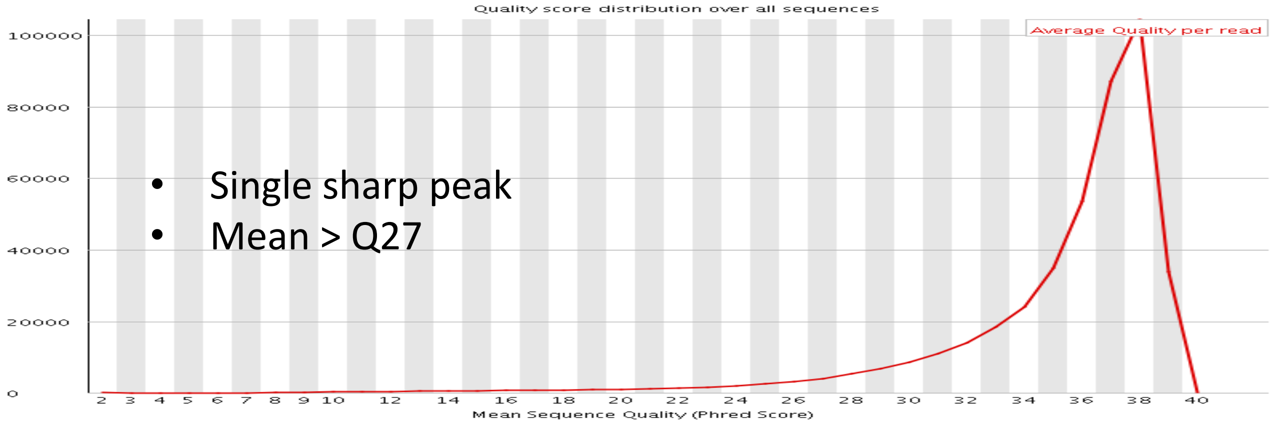
- **Poor quality at the beginning – per tile sequence quality**
- **Large variance – per sequence quality scores**

Median < Q20

Per sequence quality scores - FastQC

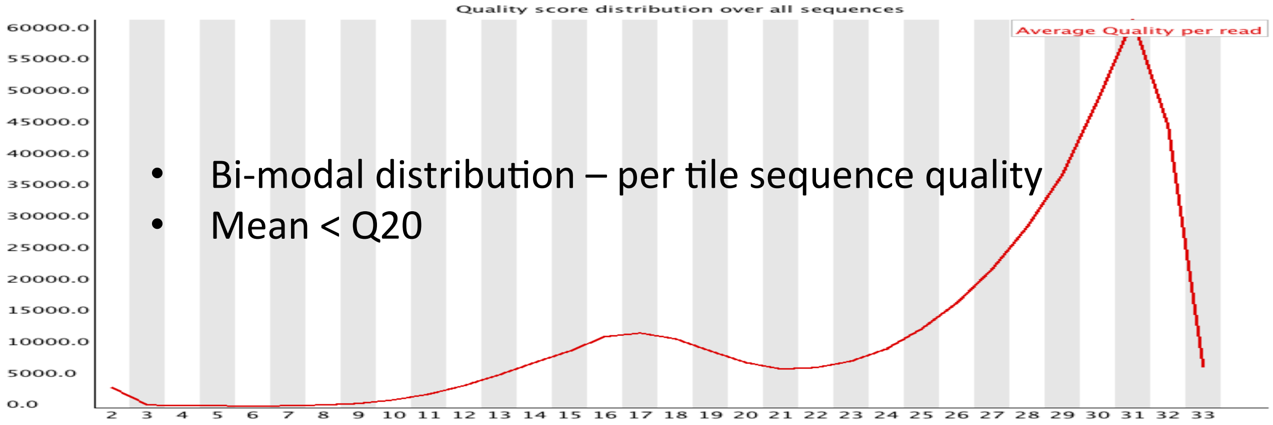
- Subset of sequences with universally low quality values

☑



- Single sharp peak
- Mean > Q27

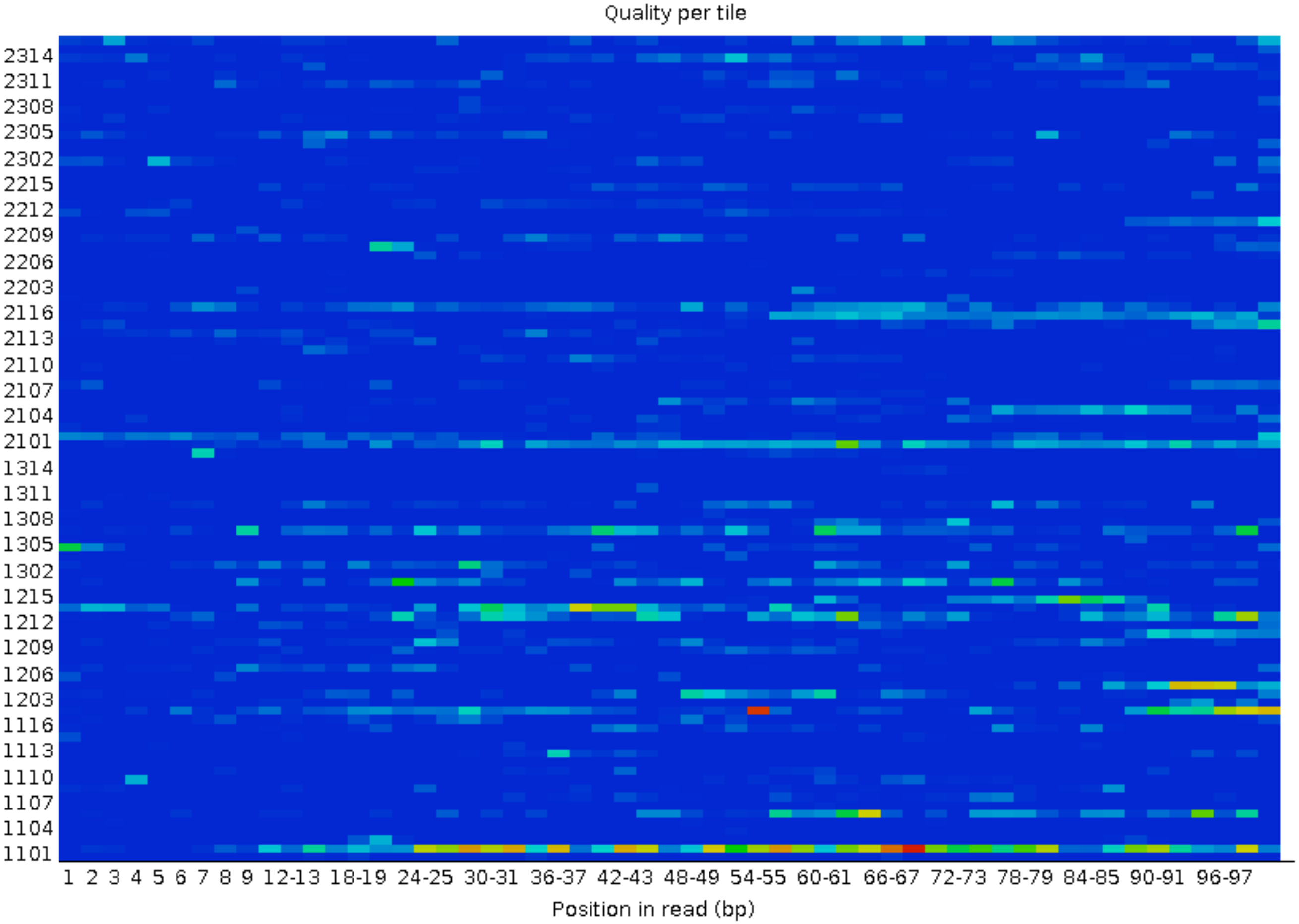
☒



- Bi-modal distribution – per tile sequence quality
- Mean < Q20

Per tile sequence quality - FastQC

- Quality scores from each tile across all bases - loss in quality associated with only one part of the flowcell



Deviation from average quality

Cold colors: ≥ average

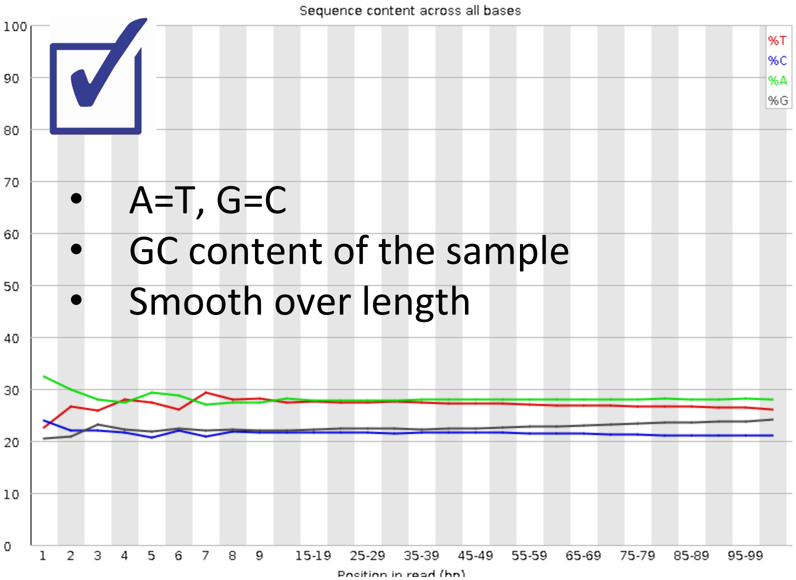
Hotter color: worse quality

Good: universal blue

Failure: < average - 5

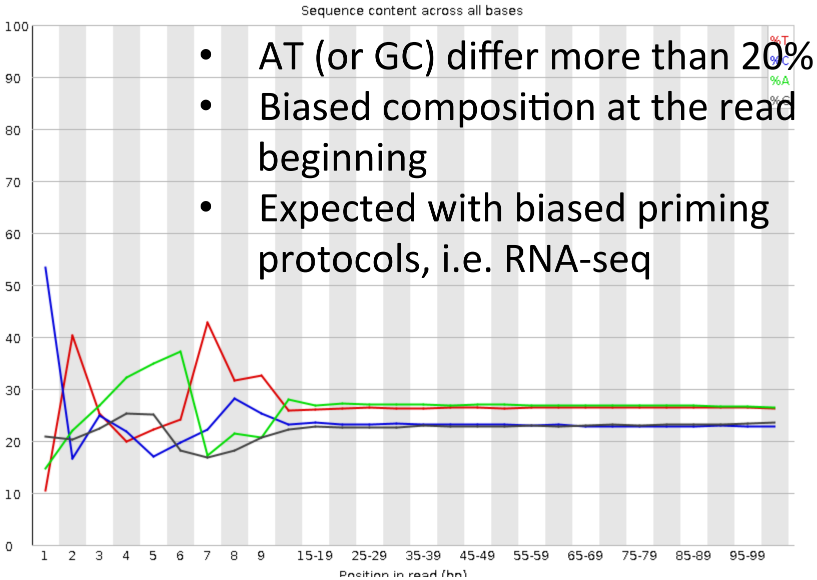
Per base sequence content - FastQC

- The portion of A, T, G, and C at each position

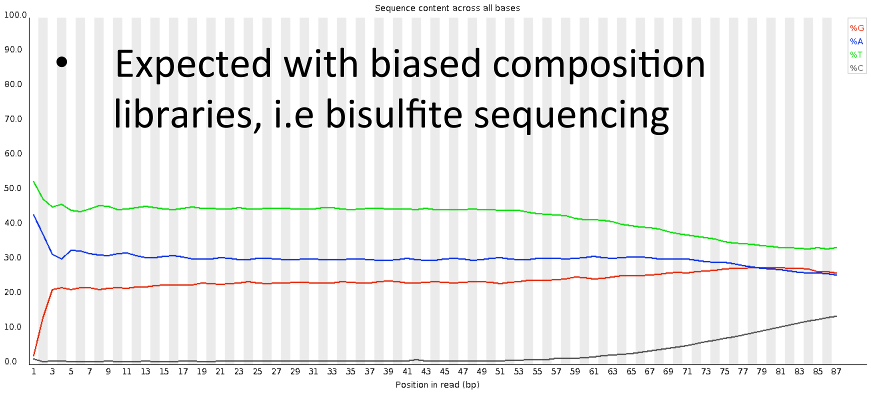


✓

- A=T, G=C
- GC content of the sample
- Smooth over length



- AT (or GC) differ more than 20%
- Biased composition at the read beginning
- Expected with biased priming protocols, i.e. RNA-seq



- Expected with biased composition libraries, i.e bisulfite sequencing

Biases in Illumina transcriptome sequencing caused by random hexamer priming

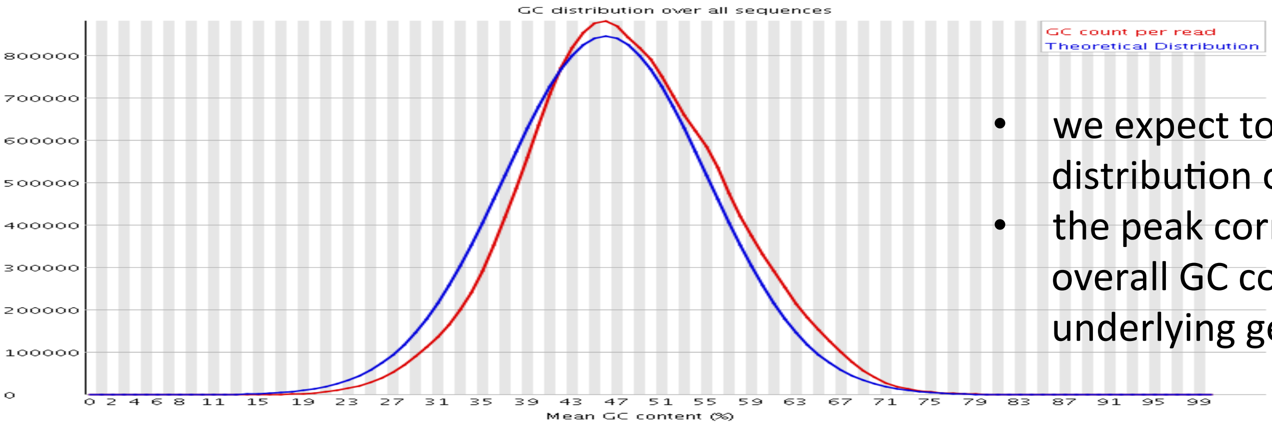
Kasper D. Hansen[1,*], Steven E. Brenner[2] and Sandrine Dudoit[1,3]

Treatment of DNA with bisulfite converts cytosine to uracil, but leaves methylated cytosine unaffected. Therefore, DNA that has been treated with bisulfite retains only methylated cytosines.

Per sequence GC content - FastQC

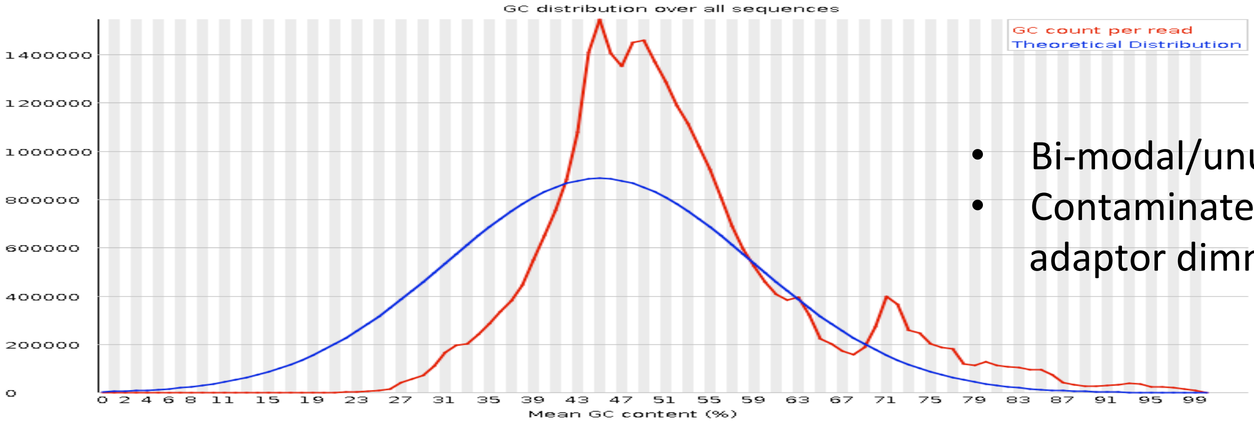
- Distribution of average GC in all reads

☑



- we expect to see a roughly normal distribution of GC content
- the peak corresponds to the overall GC content of the underlying genome

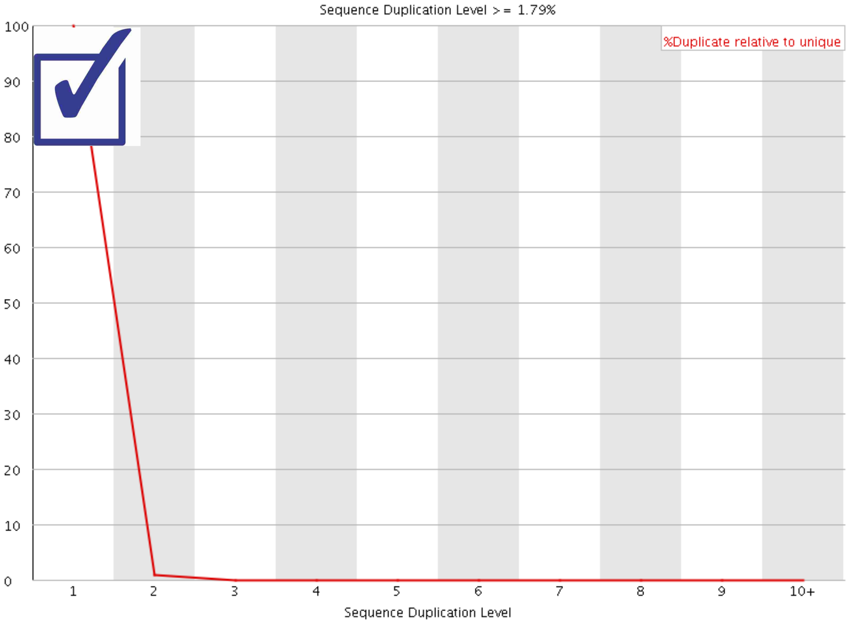
☒

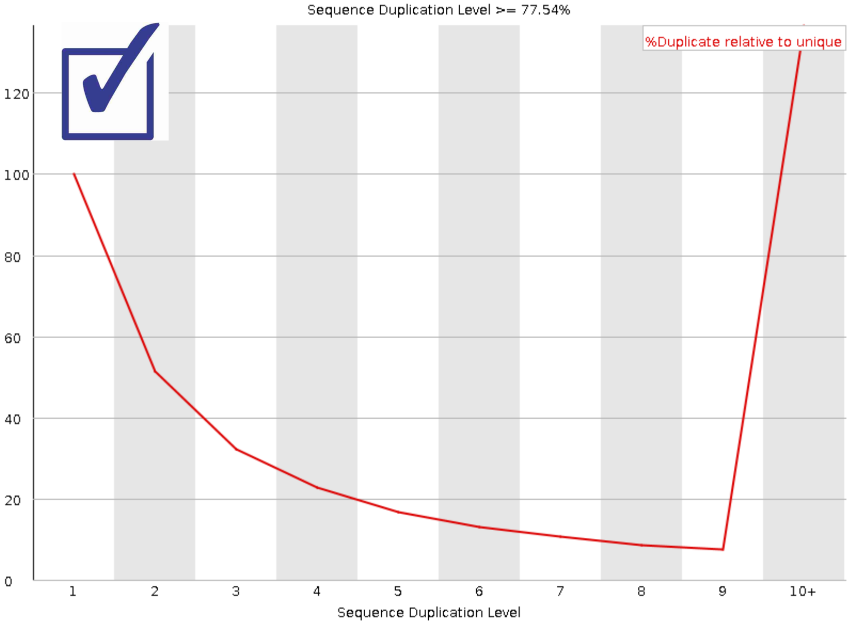


- Bi-modal/unusual distribution
- Contaminated/biased subset, i.e. adaptor dimmers, rRNA etc

Sequence duplication - FastQC

- Relative number of sequences with different degrees of duplication





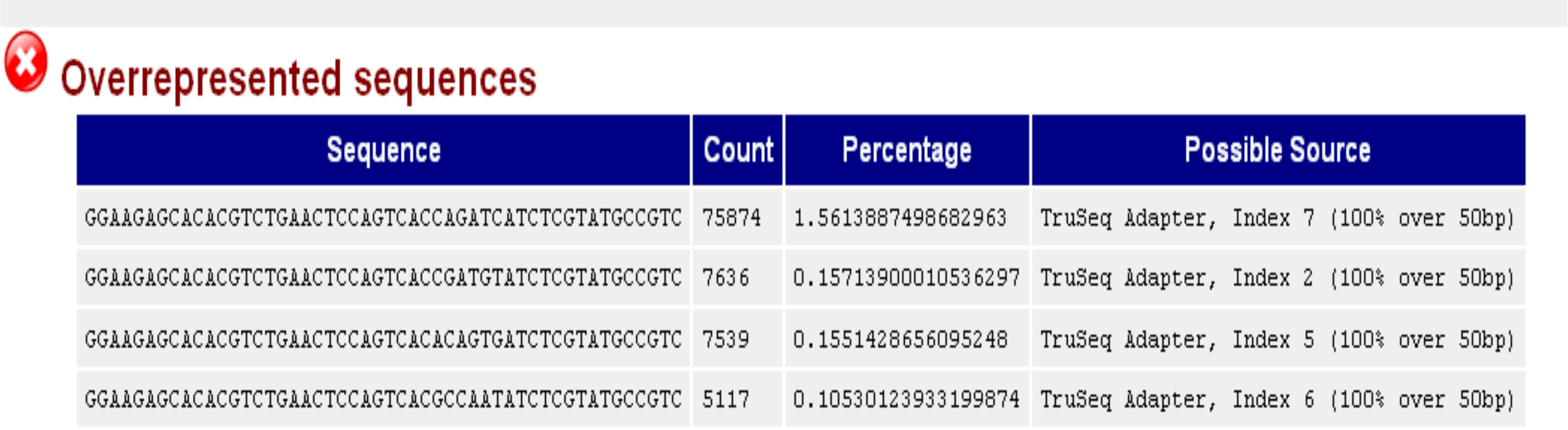
- Essentially no duplication

High duplication levels:

- DNA-seq: PCR over amplification, too little input material
- Normal in RNA-seq: high expression

Overrepresented sequences - FastQC

- Sequences make up >0.1 % of the total
- Compare those with a contamination database for finding contamination (i.e. adaptor dimmers)

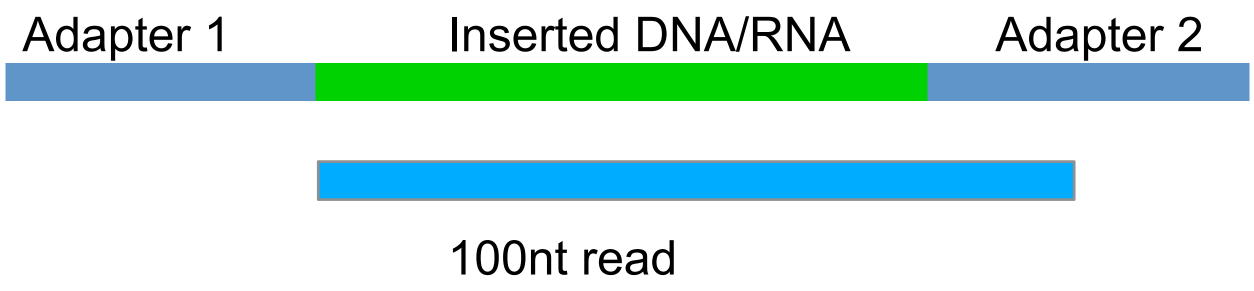


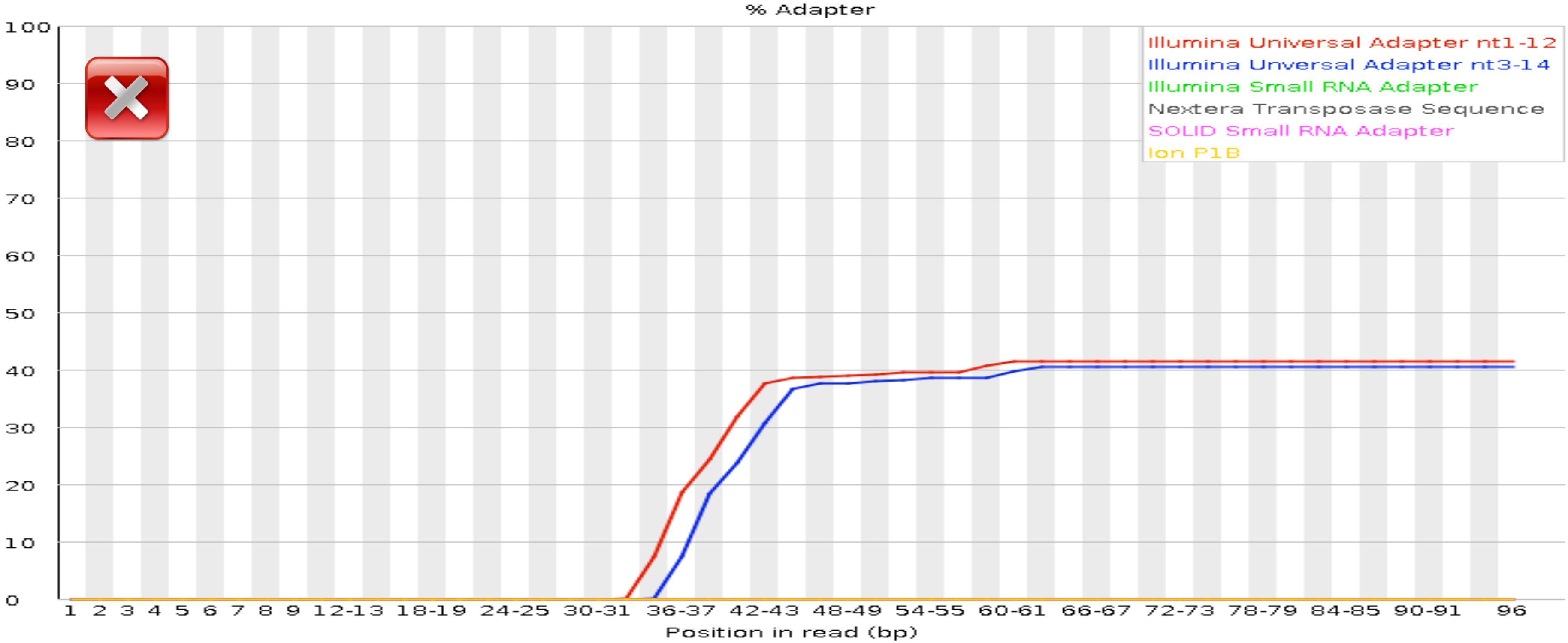
Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GGAAGAGCACACGTCTGAACTCCAGTCACCAGATCATCTCGTATGCCGTC	75874	1.5613887498682963	TruSeq Adapter, Index 7 (100% over 50bp)
GGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCTCGTATGCCGTC	7636	0.15713900010536297	TruSeq Adapter, Index 2 (100% over 50bp)
GGAAGAGCACACGTCTGAACTCCAGTCACACAGTGATCTCGTATGCCGTC	7539	0.1551428656095248	TruSeq Adapter, Index 5 (100% over 50bp)
GGAAGAGCACACGTCTGAACTCCAGTCACGCCAATATCTCGTATGCCGTC	5117	0.10530123933199874	TruSeq Adapter, Index 6 (100% over 50bp)

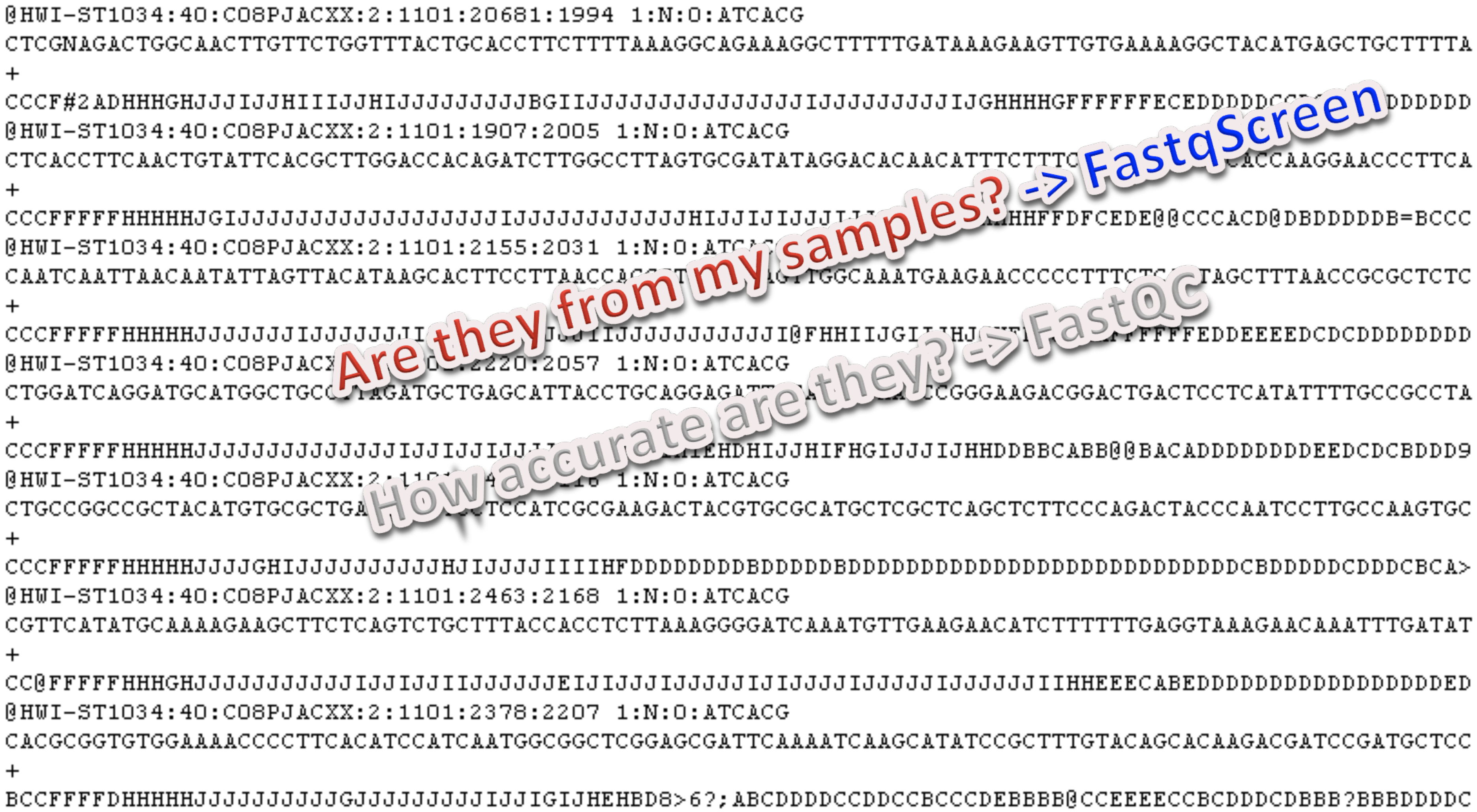
- Can be normal and biologically meaningful
  - highly expressed transcripts
  - high copy number repeats
  - Less diverse library (amplicons)

Adapter Content - FastQC





Millions of reads with base resolution

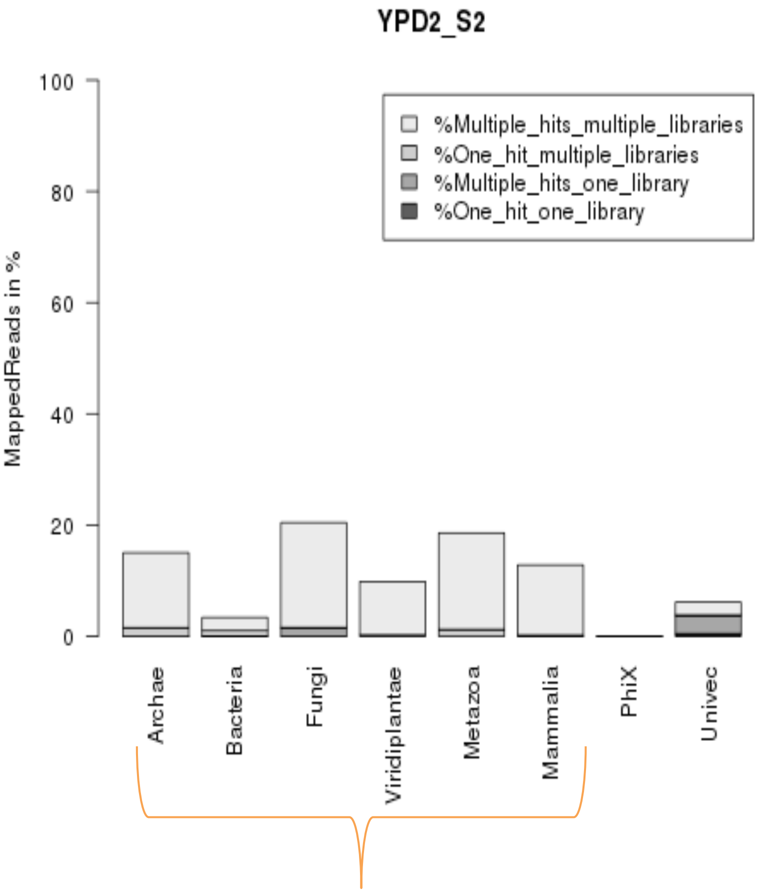


```
@HWI-ST1034:40:C08PJACXX:2:1101:20681:1994 1:N:0:ATCACG
CTCGNAGACTGGCAACTTGTTCTGGTTTACTGCACCTTCTTTTAAAGGCAGAAAGGCTTTTTGATAAAGAAGTTGTGAAAAGGCTACATGAGCTGCTTTTA
+
CCCF#2ADHHHGHJJJIJJHIIIJJHIJJJJJJJJJBGIIJJJJJJJJJJJJJJJJIJJJJJJJJJIJGHHHHGFFFFFECEDDDDDCC[illegible]DDDDDD
@HWI-ST1034:40:C08PJACXX:2:1101:1907:2005 1:N:0:ATCACG
CTCACCTTCAACTGTATTCACGCTTGGACCACAGATCTTGGCCTTAGTGCGATATAGGACACAACATTTCTTTC[illegible]CACCAAGGAACCCTTCA
+
CCCFFFFFHHHHHJGIJJJJJJJJJJJJJJJJJJJJJIJJJJJJJJJJJJJHIJJIJIJJJ[illegible]HHFFDFCEDE@@CCCACD@DBDDDDDB=BCCC
@HWI-ST1034:40:C08PJACXX:2:1101:2155:2031 1:N:0:ATC[illegible]
CAATCAATTAACAATATTAGTTACATAAGCACTTCCTTAACCA[illegible]AGTTGGCAAATGAAGAACCCCCTTTCT[illegible]TAGCTTTAACCGCGCTCTC
+
CCCFFFFFHHHHHJJJJJJJIJJJJJJJI[illegible]JJJIIJJJJJJJJJJJI@FHHIIJGI[illegible]HJ[illegible]FFFFFEDDEEEEDCDCDDDDDDDD
@HWI-ST1034:40:C08PJACX[illegible]:2220:2057 1:N:0:ATCACG
CTGGATCAGGATGCATGGCTGCC[illegible]AGATGCTGAGCATTACCTGCAGGAGA[illegible]CCGGGAAGACGGACTGACTCCTCATATTTTGCCGCCTA
+
CCCFFFFFHHHHHJJJJJJJJJJJJJJIJJIJJIJJ[illegible]EHDHIJJHIFHGIJJJIJHHDDBBCABB@@BACADDDDDDDDEEDCDCBDDD9
@HWI-ST1034:40:C08PJACXX:2:11[illegible]16 1:N:0:ATCACG
CTGCCGGCCGCTACATGTGCGCTGA[illegible]TCCATCGCGAAGACTACGTGCGCATGCTCGCTCAGCTCTTCCCAGACTACCCAATCCTTGCCAAGTGC
+
CCCFFFFFHHHHHJJJJGHIJJJJJJJJJJJHJIJJJJIIIIHFDDDDDDDDBDDDDDBDDDDDDDDDDDDDDDDDDDDDDDDDDCBDDDDDCDDDCBCA>
@HWI-ST1034:40:C08PJACXX:2:1101:2463:2168 1:N:0:ATCACG
CGTTCATATGCAAAAGAAGCTTCTCAGTCTGCTTTACCACCTCTTAAAGGGGATCAAATGTTGAAGAACATCTTTTTTGAGGTAAAGAACAAATTTGATAT
+
CC@FFFFFHHHGHJJJJJJJJJJJIJJIJJIIJJJJJJEIJIJJJIJJJJJIJIJJJJIJJJJJIJJJJJJIIHHEEECABEDDDDDDDDDDDDDDDDDED
@HWI-ST1034:40:C08PJACXX:2:1101:2378:2207 1:N:0:ATCACG
CACGCGGTGTGGAAAACCCCTTCACATCCATCAATGGCGGCTCGGAGCGATTCAAAATCAAGCATATCCGCTTTGTACAGCACAAGACGATCCGATGCTCC
+
BCCFFFFDHHHHHJJJJJJJJJJGJJJJJJJJJIJJIGIJHEHBD8>6?;ABCDDDDCCDDCCBCCCDEBBBB@CCEEEECCBCDDDCDBBB?BBBDDDDC
```

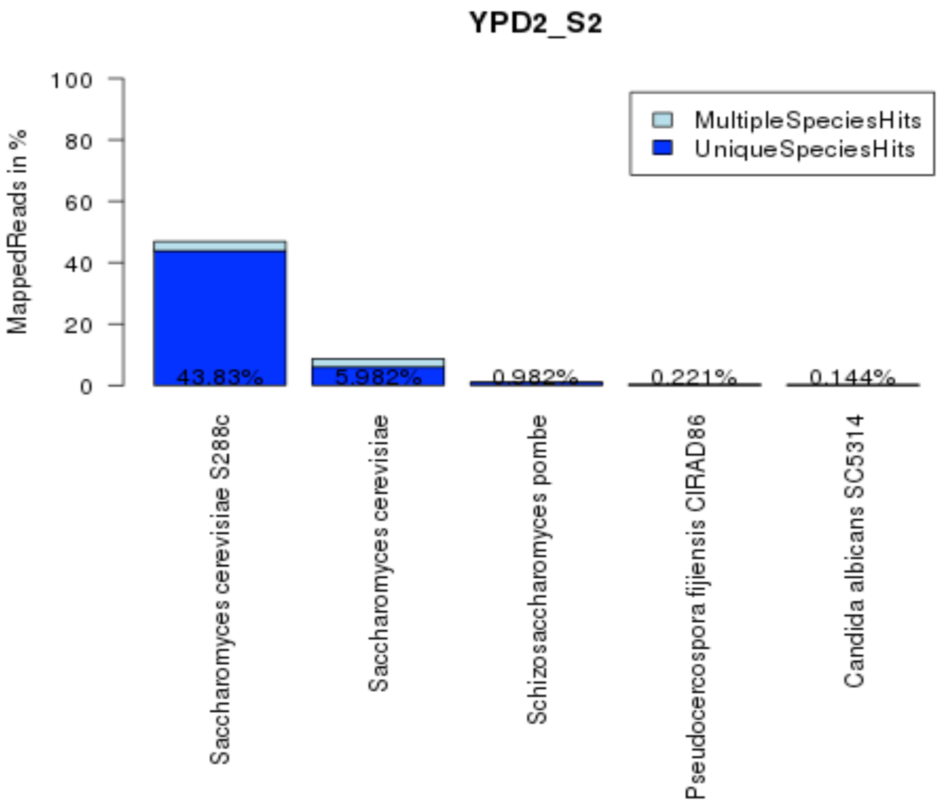
Are they from my samples? -> FastqScreen

How accurate are they? -> FastQC

Contamination Check - FastqScreen



rRNA-Content (Silva)



Mapping to RefSeq mRNAs
(all species)

Data preprocessing common tasks

1. Trimming: remove bad bases from (end(s) of) reads
   - Adaptor sequence
   - Low quality bases

2. Filtering: remove bad reads
   - Low quality reads
   - Contaminating sequences
   - Low complexity reads (repeats)
   - Short (<20bp) reads – they slow down mapping software

Data preprocessing software

- PRINSEQ
  - http://prinseq.sourceforge.net/
  - Quality/hard trimming, quality filtering, reformat, ...
- Trimmomatic
  - http://www.usadellab.org/cms/?page=trimmomatic
  - Adaptor trimming, quality trimming &filtering, ...
- FlexBar (FAR)
  - http://sourceforge.net/projects/theflexibleadap/
  - Flexible barcode detection and adapter removal
- FASTX
  - http://hannonlab.cshl.edu/fastx_toolkit/
  - Reformat, stats, collapse duplicated reads, trim, filter, reverse compliment
- TagCleaner
  - http://tagcleaner.sourceforge.net
  - Trim MIDs or adaptors, demultiplexing
- DeconSeq
  - http://deconseq.sourceforge.net
  - Remove potential contaminants

Recommendations

- Always generate quality plots for all libraries
- Trim and/or filter data if needed
  - always trim and filter for de-novo transcriptome assembly