Quality control

https://jobsite.procore.com/this-friday-s-funny-the-history-of-quality-control

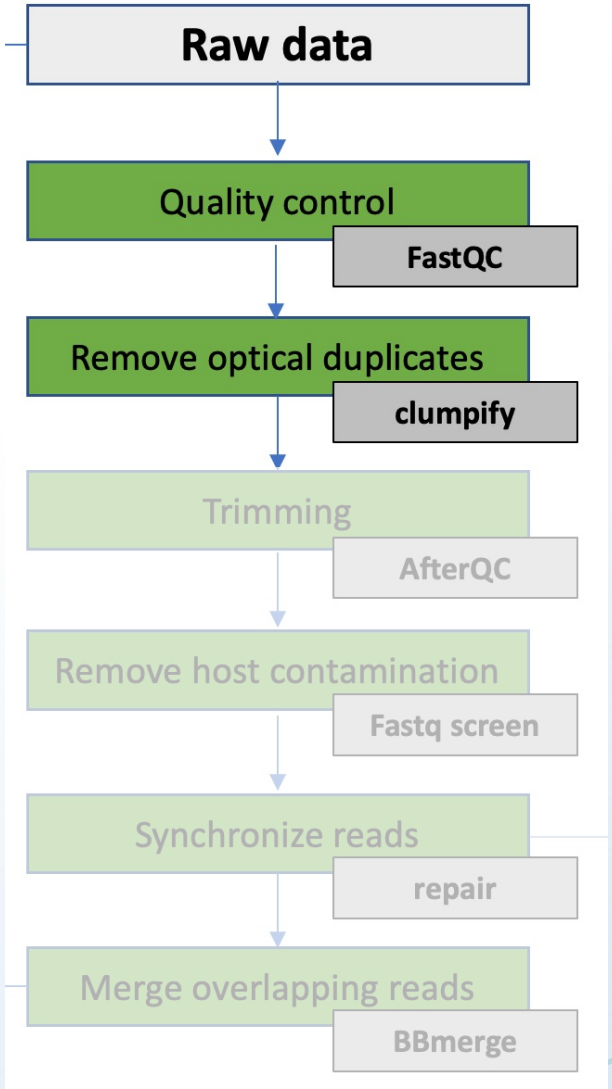
Overview of this talk

Sequencing recap

Data integrity

Quality control of sequence data

Optical duplicates in data



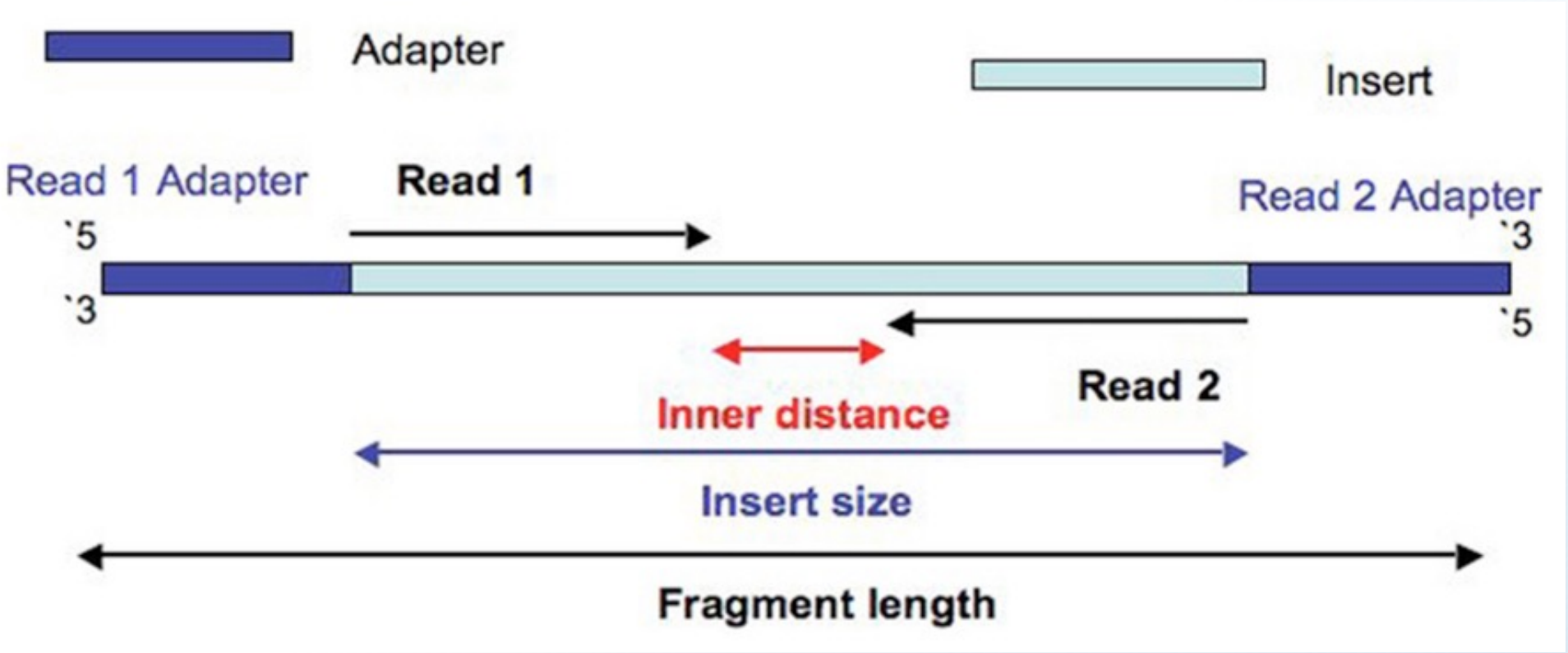
Wet lab QC - DNA quality is always checked prior to library preparation and sequencing

Methods for quantification and evaluation of NGS libraries:

	Agarose Gel Electrophoresis	Spectrophotometry	qPCR	Digital PCR (dPCR)
Quantification	Relative	Absolute (*)	Relative	Absolute (**)
Sensitivity	+	++	++++	+++++
Quality Assessment	+	+	+++	+++++
Cost	+	++	+++	+++

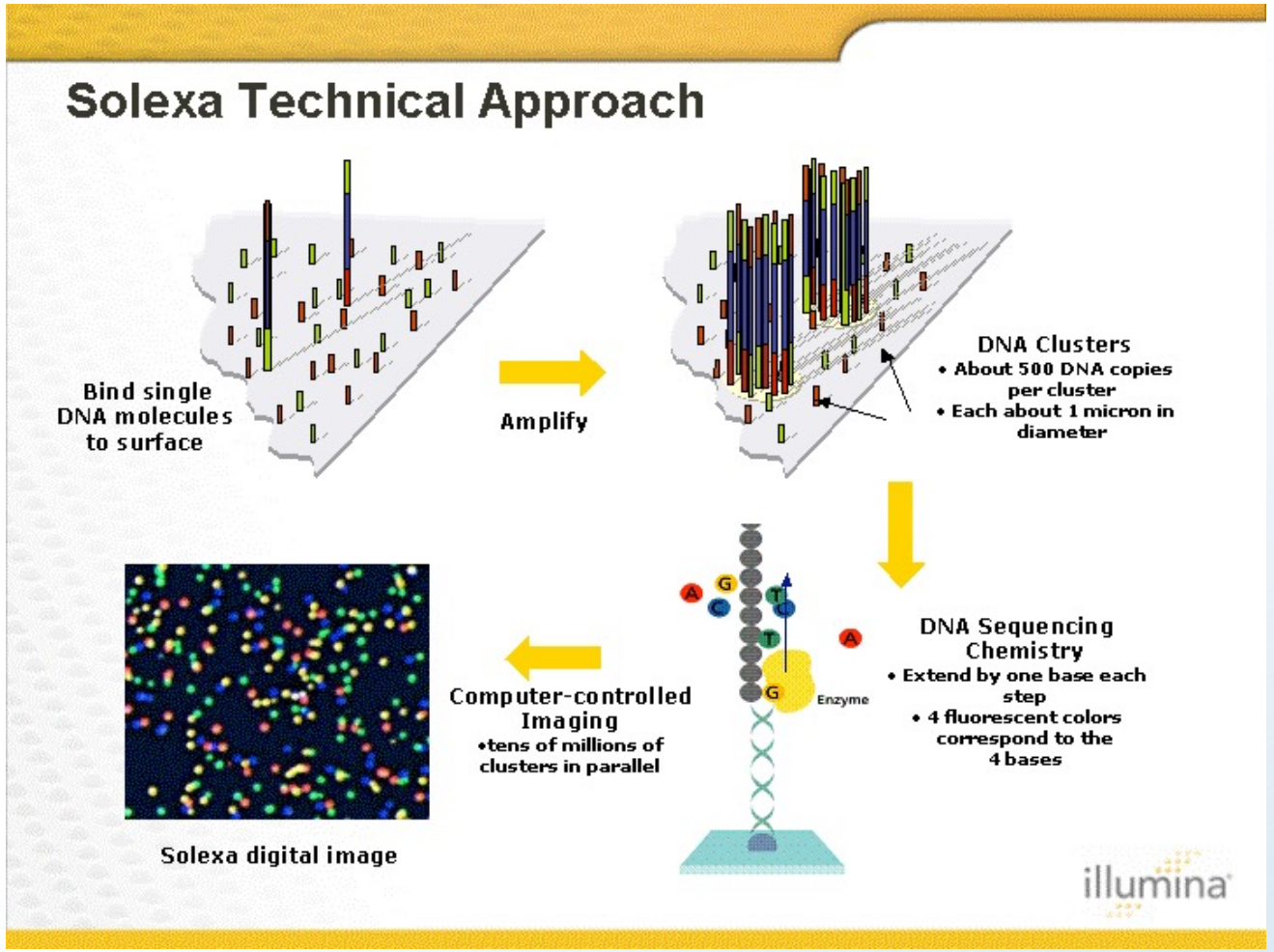
DNA sequencing recap – Illumina technology

Paired-end (PE) Illumina data



Sequencing quality scores measure the probability that a base is called incorrectly

Base calls are made directly from signal intensity measurements



Sequencing technologies are not perfect and do produce errors

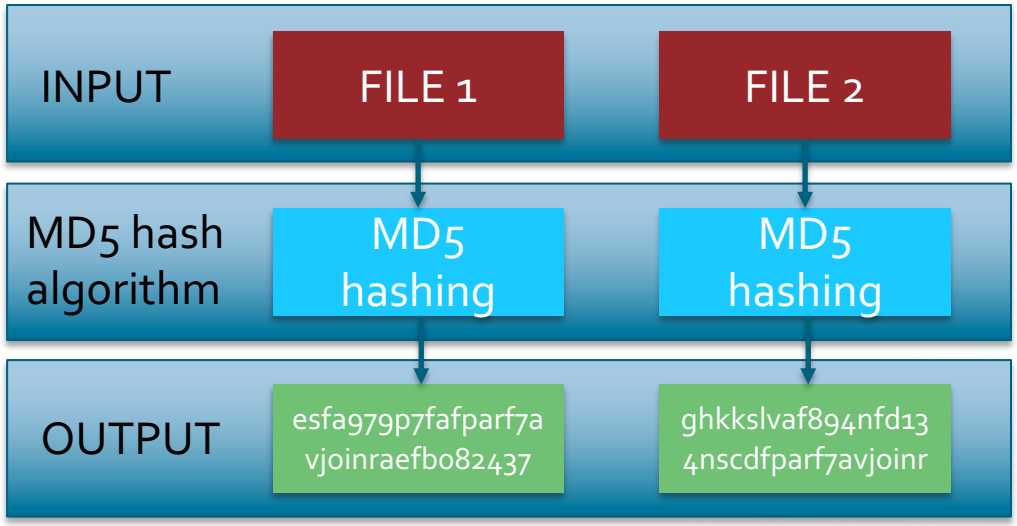
QC ensures that the data used for downstream analysis does not contain (too many) errors and poor quality sequences

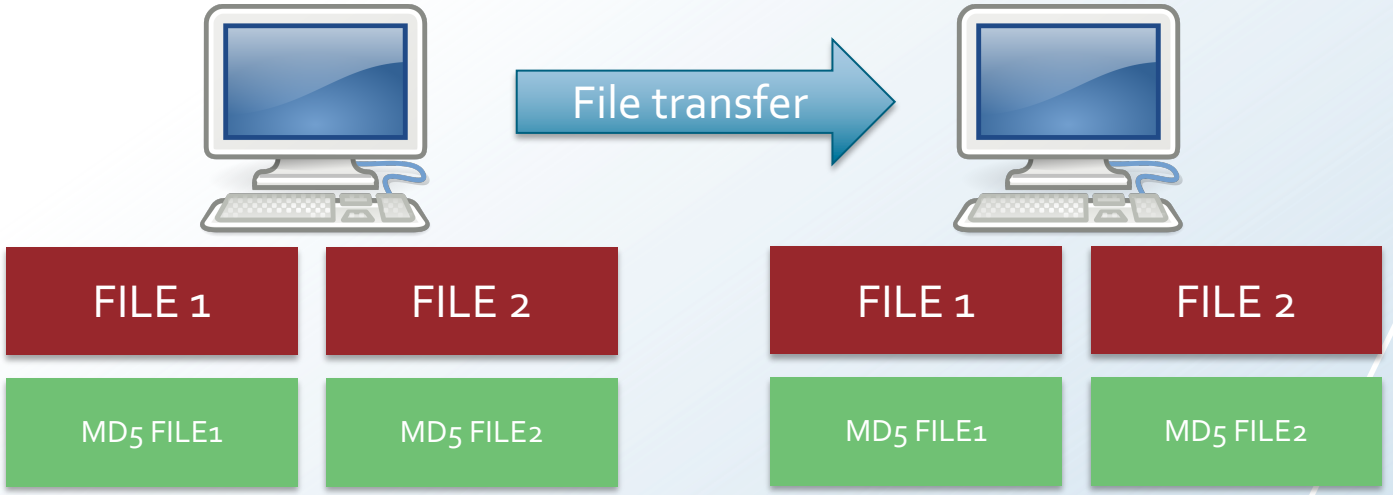
Check that the data is complete

File checksums ensure data integrity, eg. MD5

Calculate file checksums before transfer

After the transfer, checksum the transferred file





Check the data quantity

Generate simple statistics of FASTA/Q files using eg. SeqKit

```
$ seqkit stat *.fg
file            format type   num_seqs  sum_len      min_len  avg_len  max_len
dataset_A.fq    FASTQ DNA     2,675,748 244,807,643  50       99       100
dataset_B.fq    FASTQ DNA     3,789,194 388,099,750  100      100      100
dataset_C.fq    FASTQ DNA     9,186,045 918,604,500  100      100      100
```

Check the format - The Illumina output are FASTQ files

Sequence read name

1 and 2 refer to the forward and reverse (PE) read

Read sequence (base calls)

Sequence quality. One character per base

Sequence/quality line separator

Check the format - The Illumina output are FASTQ files

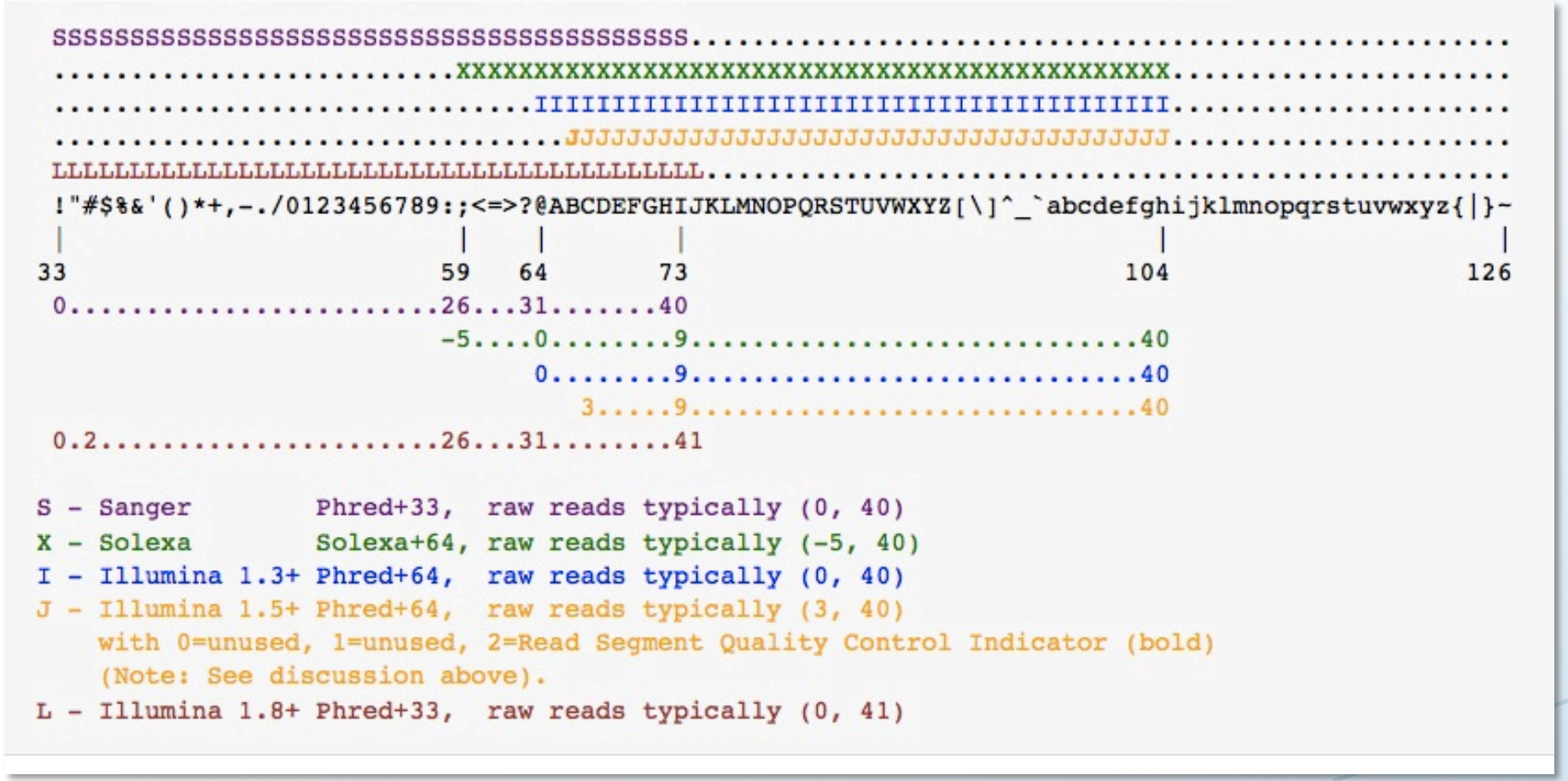
The FASTQ header

```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
```

EAS139	the unique instrument name
136	the run id
FC706VJ	the flowcell id
2	flowcell lane
2104	tile number within the flowcell lane
15343	'x'-coordinate of the cluster within the tile
197393	'y'-coordinate of the cluster within the tile
1	the member of a pair, 1 or 2 *(paired-end or mate-pair reads only)*
Y	Y if the read is filtered, N otherwise
18	0 when none of the control bits are on, otherwise it is an even number
ATCACG	index sequence

A range of sequence quality scores depending on technology and the base caller

Most modern sequencing machine, such as MiSeq use Illumina 1.8+



The FASTQ quality scores derives from using ASCII encoding

ASCII codes represent text in computers, originally based on the English alphabet

Dec	Hex	Name	Char	Ctrl-char	Dec	Hex	Char	Dec	Hex	Char	Dec	Hex	Char
0	0	Null	NUL	CTRL-@	32	20	Space	64	40	@	96	60	`
1	1	Start of heading	SOH	CTRL-A	33	21	!	65	41	A	97	61	a
2	2	Start of text	STX	CTRL-B	34	22	"	66	42	B	98	62	b
3	3	End of text	ETX	CTRL-C	35	23	#	67	43	C	99	63	c
4	4	End of xmit	EOT	CTRL-D	36	24	$	68	44	D	100	64	d
5	5	Enquiry	ENQ	CTRL-E	37	25	%	69	45	E	101	65	e
6	6	Acknowledge	ACK	CTRL-F	38	26	&	70	46	F	102	66	f
7	7	Bell	BEL	CTRL-G	39	27	'	71	47	G	103	67	g
8	8	Backspace	BS	CTRL-H	40	28	(	72	48	H	104	68	h
9	9	Horizontal tab	HT	CTRL-I	41	29	)	73	49	I	105	69	i
10	0A	Line feed	LF	CTRL-J	42	2A	*	74	4A	J	106	6A	j
11	0B	Vertical tab	VT	CTRL-K	43	2B	+	75	4B	K	107	6B	k
12	0C	Form feed	FF	CTRL-L	44	2C	,	76	4C	L	108	6C	l
13	0D	Carriage feed	CR	CTRL-M	45	2D	-	77	4D	M	109	6D	m
14	0E	Shift out	SO	CTRL-N	46	2E	.	78	4E	N	110	6E	n
15	0F	Shift in	SI	CTRL-O	47	2F	/	79	4F	O	111	6F	o
16	10	Data line escape	DLE	CTRL-P	48	30	0	80	50	P	112	70	p
17	11	Device control 1	DC1	CTRL-Q	49	31	1	81	51	Q	113	71	q
18	12	Device control 2	DC2	CTRL-R	50	32	2	82	52	R	114	72	r
19	13	Device control 3	DC3	CTRL-S	51	33	3	83	53	S	115	73	s
20	14	Device control 4	DC4	CTRL-T	52	34	4	84	54	T	116	74	t
21	15	Neg acknowledge	NAK	CTRL-U	53	35	5	85	55	U	117	75	u
22	16	Synchronous idle	SYN	CTRL-V	54	36	6	86	56	V	118	76	v
23	17	End of xmit block	ETB	CTRL-W	55	37	7	87	57	W	119	77	w
24	18	Cancel	CAN	CTRL-X	56	38	8	88	58	X	120	78	x
25	19	End of medium	EM	CTRL-Y	57	39	9	89	59	Y	121	79	y
26	1A	Substitute	SUB	CTRL-Z	58	3A	:	90	5A	Z	122	7A	z
27	1B	Escape	ESC	CTRL-[	59	3B	;	91	5B	[	123	7B	{
28	1C	File separator	FS	CTRL-\	60	3C	<	92	5C	\	124	7C	|
29	1D	Group separator	GS	CTRL-]	61	3D	=	93	5D	]	125	7D	}
30	1E	Record separator	RS	CTRL-^	62	3E	>	94	5E	^	126	7E	~
31	1F	Unit separator	US	CTRL-_	63	3F	?	95	5F	_	127	7F	DEL

Q30 is considered a benchmark for sequence quality in next-generation sequencing

- A quality score of 99% (Q30) will have an incorrect base call probability of 1 in 1000

Relationship Between Sequencing Quality Score and Base Call Accuracy:

Quality Score	Probability of Incorrect Base Call	Inferred Base Call Accuracy
10 (Q10)	1 in 10	90%
20 (Q20)	1 in 100	99%
30 (Q30)	1 in 1000	99.9%

Simple FASTQ/FASTA manipulations using SeqKit

SeqKit is a cross-platform ultrafast comprehensive toolkit that can perform common manipulations of FASTA/Q files

Eg converting, searching, filtering, deduplication, splitting

Works on all major operating systems, including Windows, Linux, and Mac OS X, and can be directly used without any dependencies

Easy to use on command line:

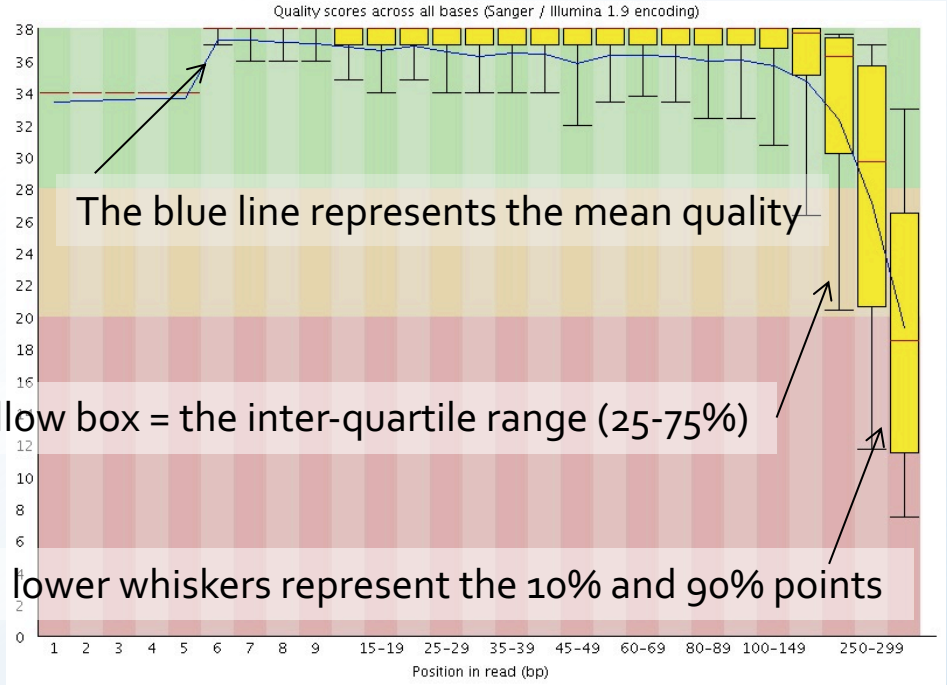
$ seqkit <subcommand>

List of subcommands can be found here: https://github.com/shenwei356/seqkit

When performing Quality Control (QC) you generate a general summary of the input data

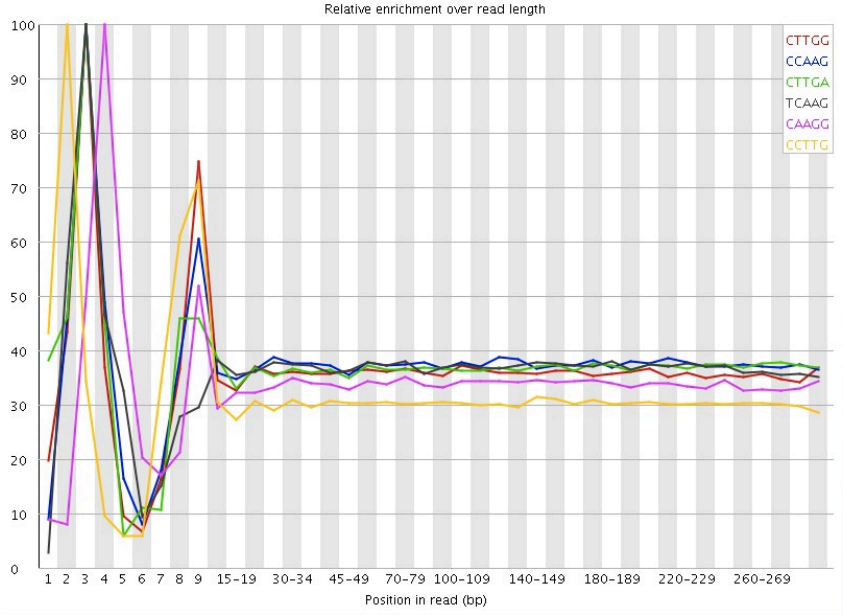
The range of quality values across all bases at each position in the FASTQ file

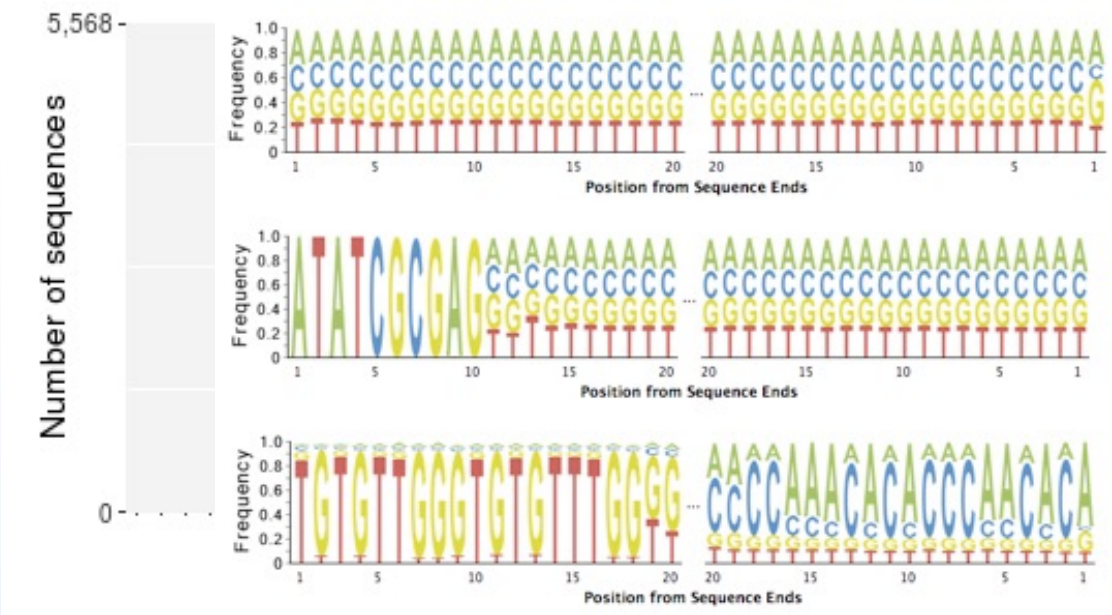
Measure	Value
Filename	CAV3_S1_L001_R1_001.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	2172538
Filtered Sequences	0
Sequence length	35-301
%GC	41



Low quality sequences can cause problems during downstream analysis

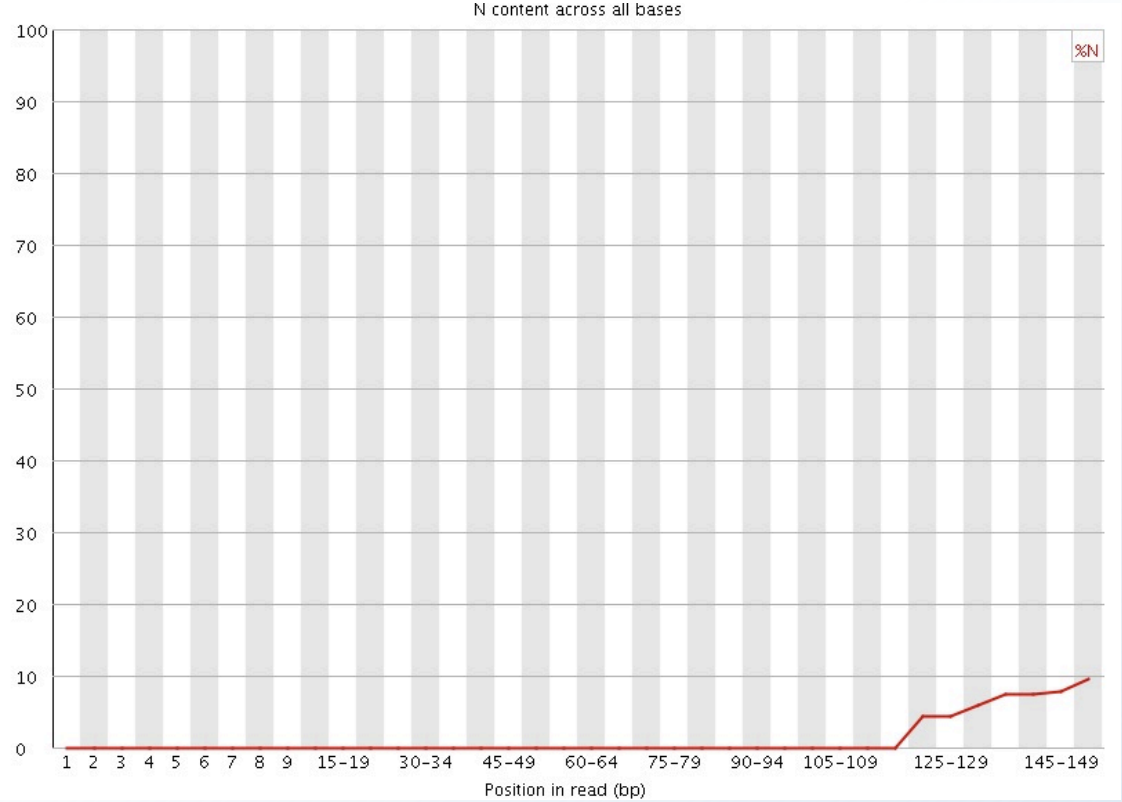
Analyzing K-mer content may identify artifacts in sequence reads, eg. multiplex identifiers, adapters, and primer sequences





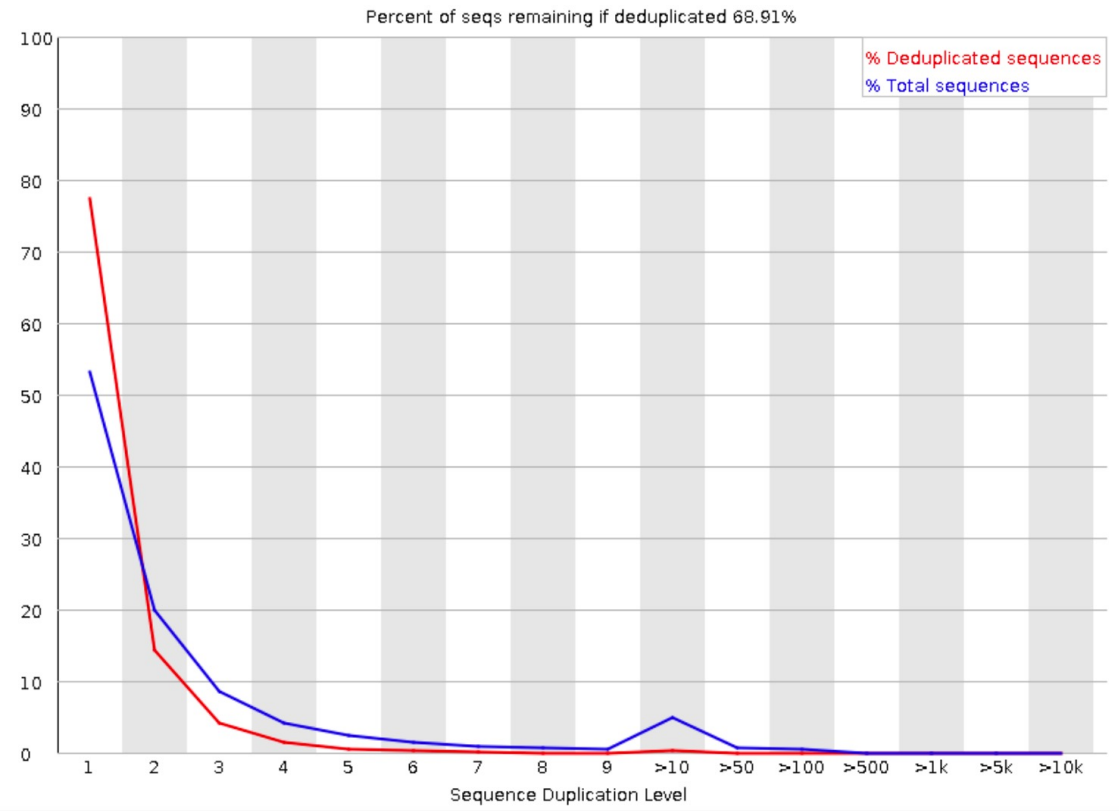
Ambiguous bases (Ns) arises when a sequencer is unable to make a base call with sufficient confidence

A high number of Ns can be a sign for a low quality sequence and can cause problems during downstream analysis



Identification of sequence duplication since ideally, no reads should not start at the same position and have the same errors

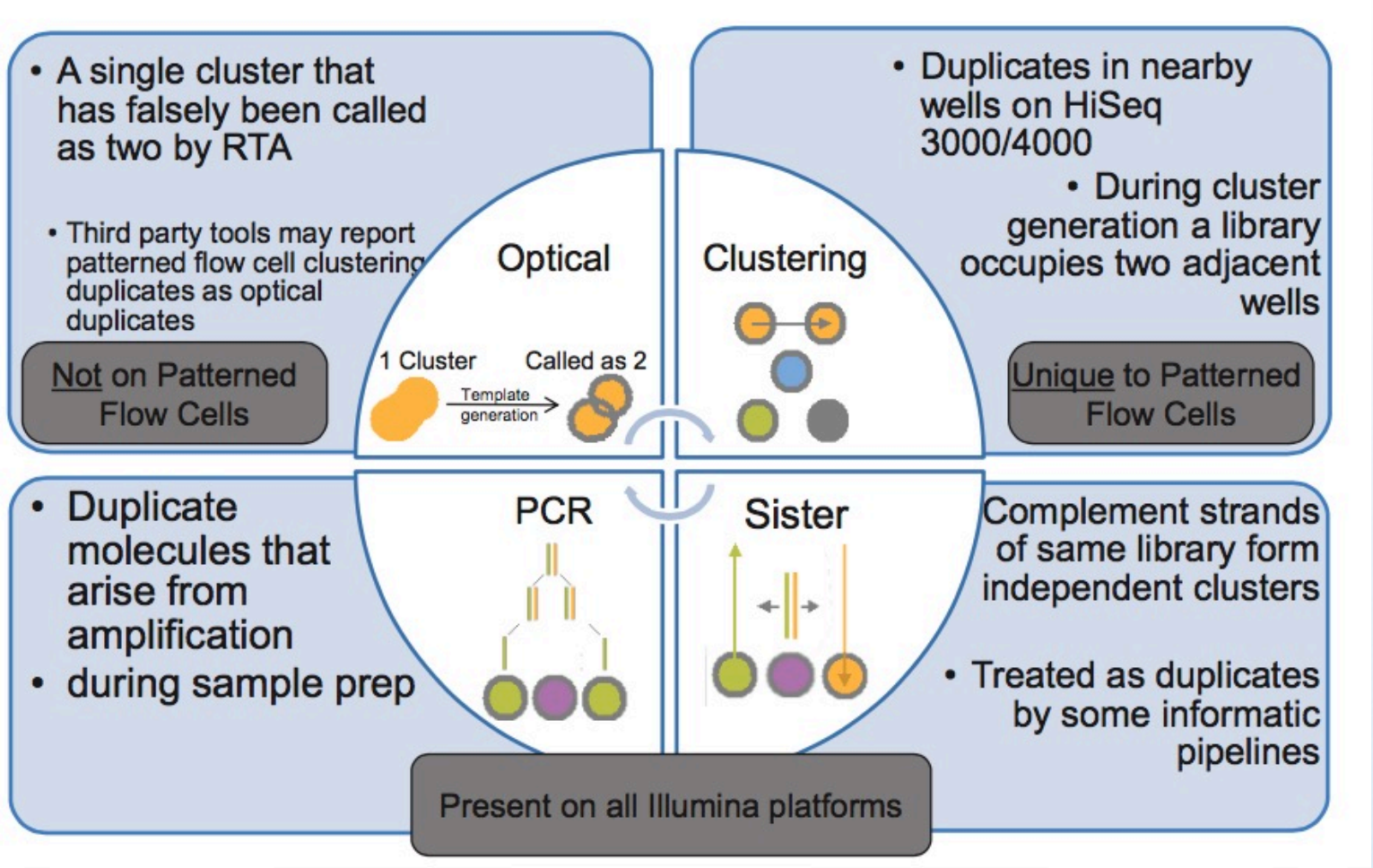
Most likely to indicate some kind of enrichment bias



Duplicates in NGS data

Optical duplicates

PCR and enrichment duplicates



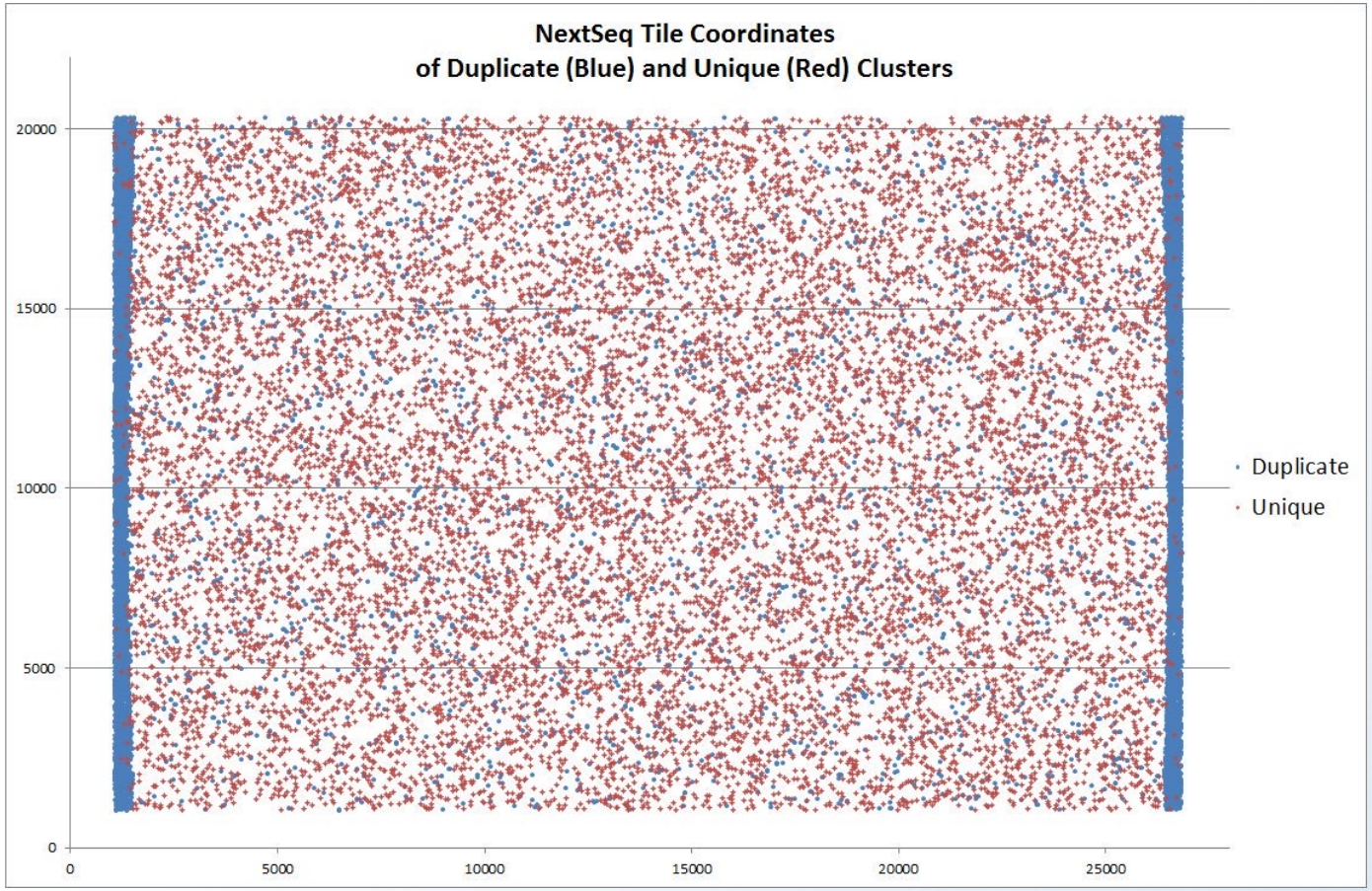
Local re-clustering

Exclusion Amplification duplicates

Illumina

Duplicates in NGS data

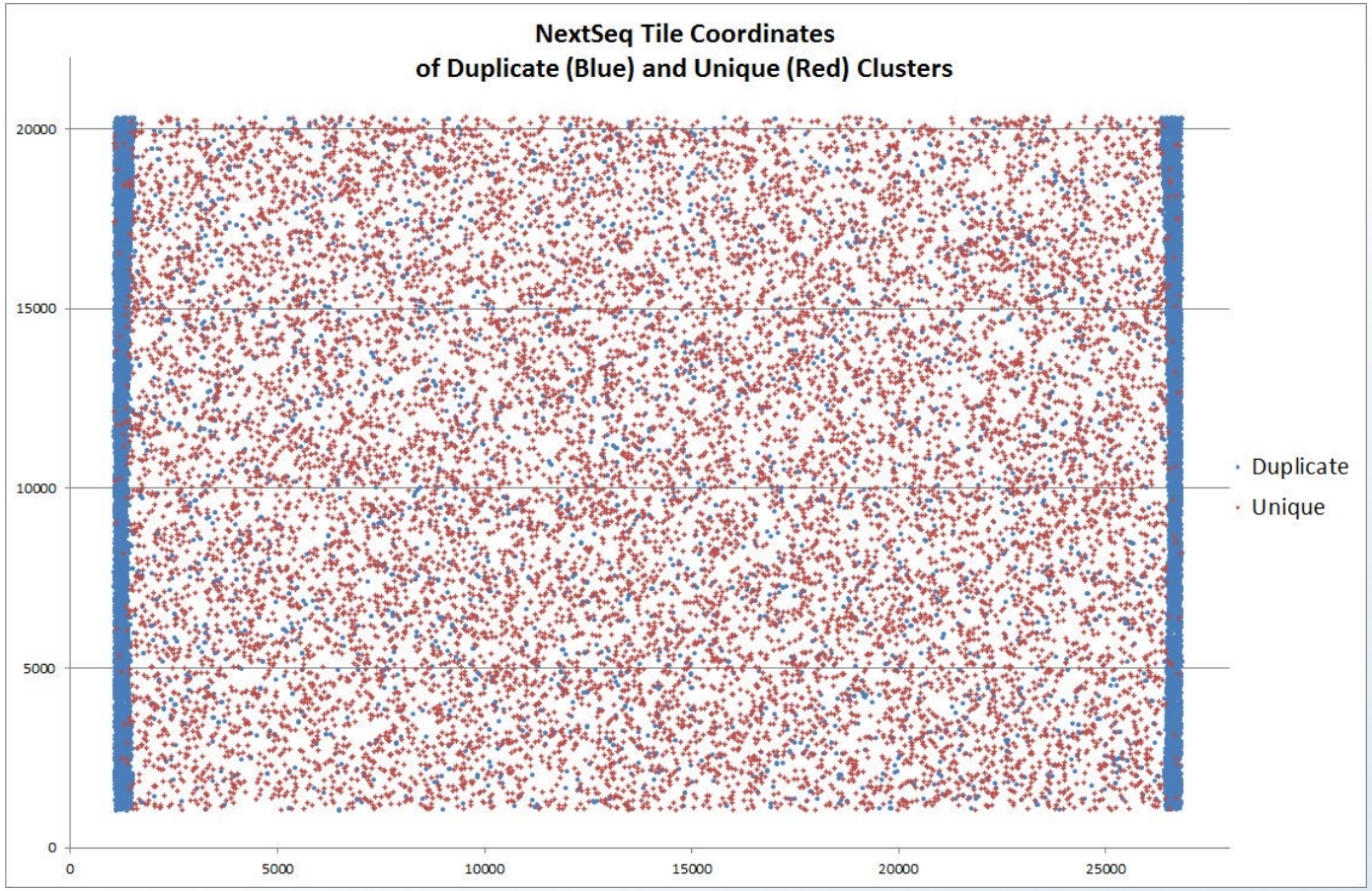
Tile-edge duplicates may account for >80% of the duplicates



Brian Bushell

Remove duplicates from NGS data

Clumpify removes duplicates and reorders reads to maximize gzip compression



Brian Bushell

Popular QC tools

	FastQC	Prinseq	Kraken
Standalone	+	+	+
Web tool	-	+	-
GUI	+	+	-
Summary statistics	+	+	+
Adaptor trimming	-	+	+
Quality trimming	-	+	+
Format conversion	-	+	
Sequencing technologies	Illumina, PacBio	Illumina, 454	Illumina