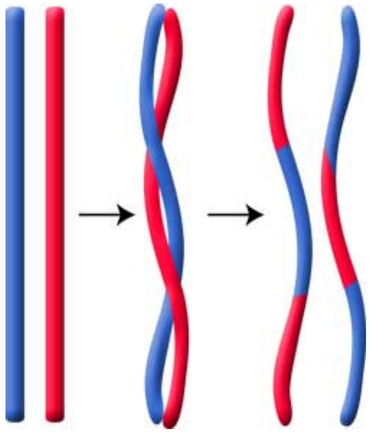
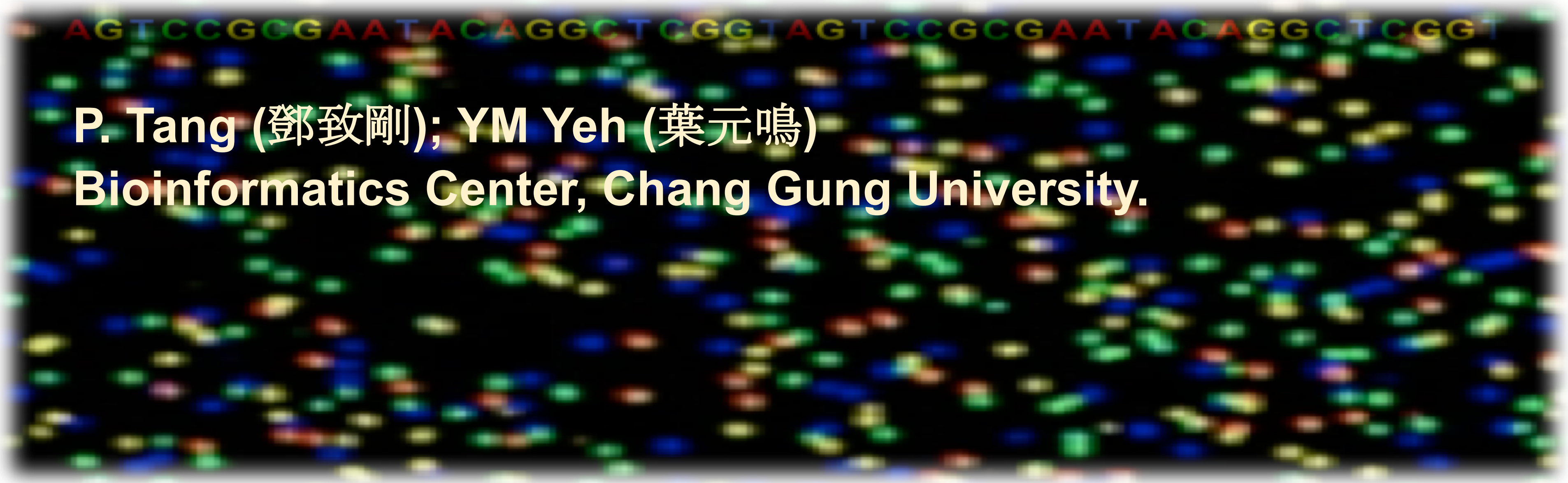
03 Mar 2017

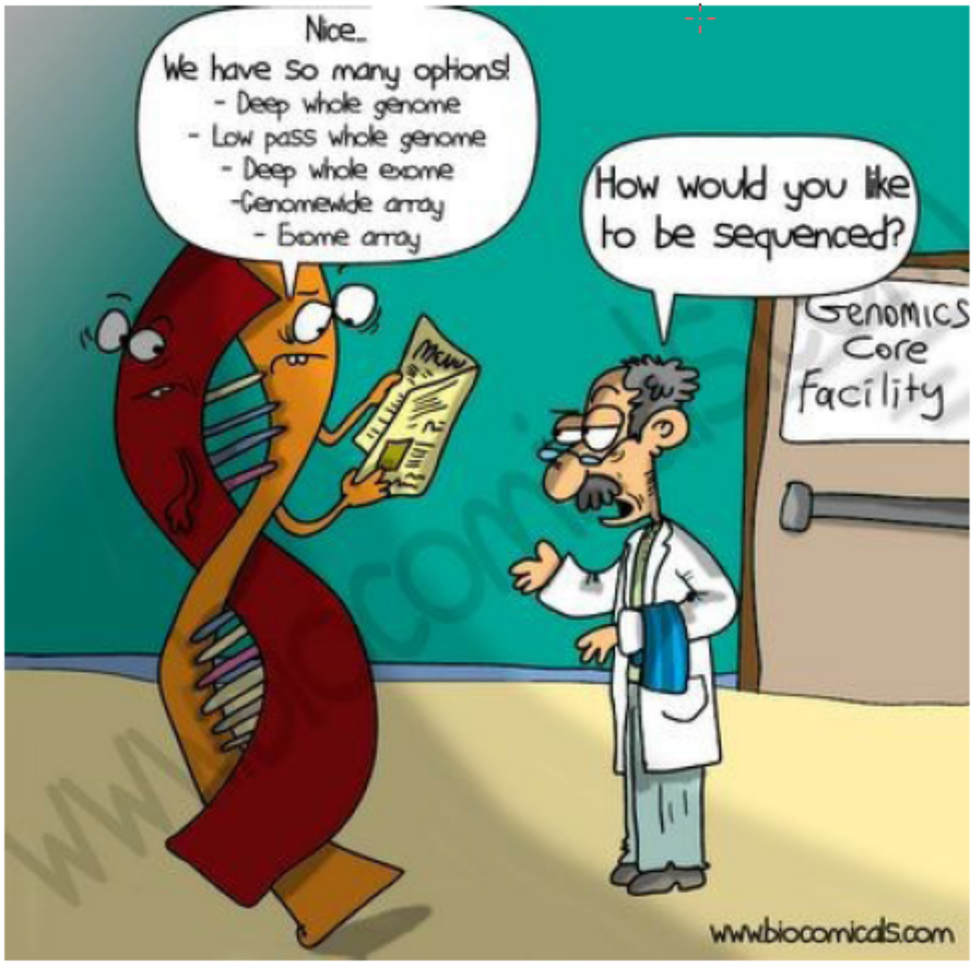


Databases and Tools for High Throughput Sequencing Analysis



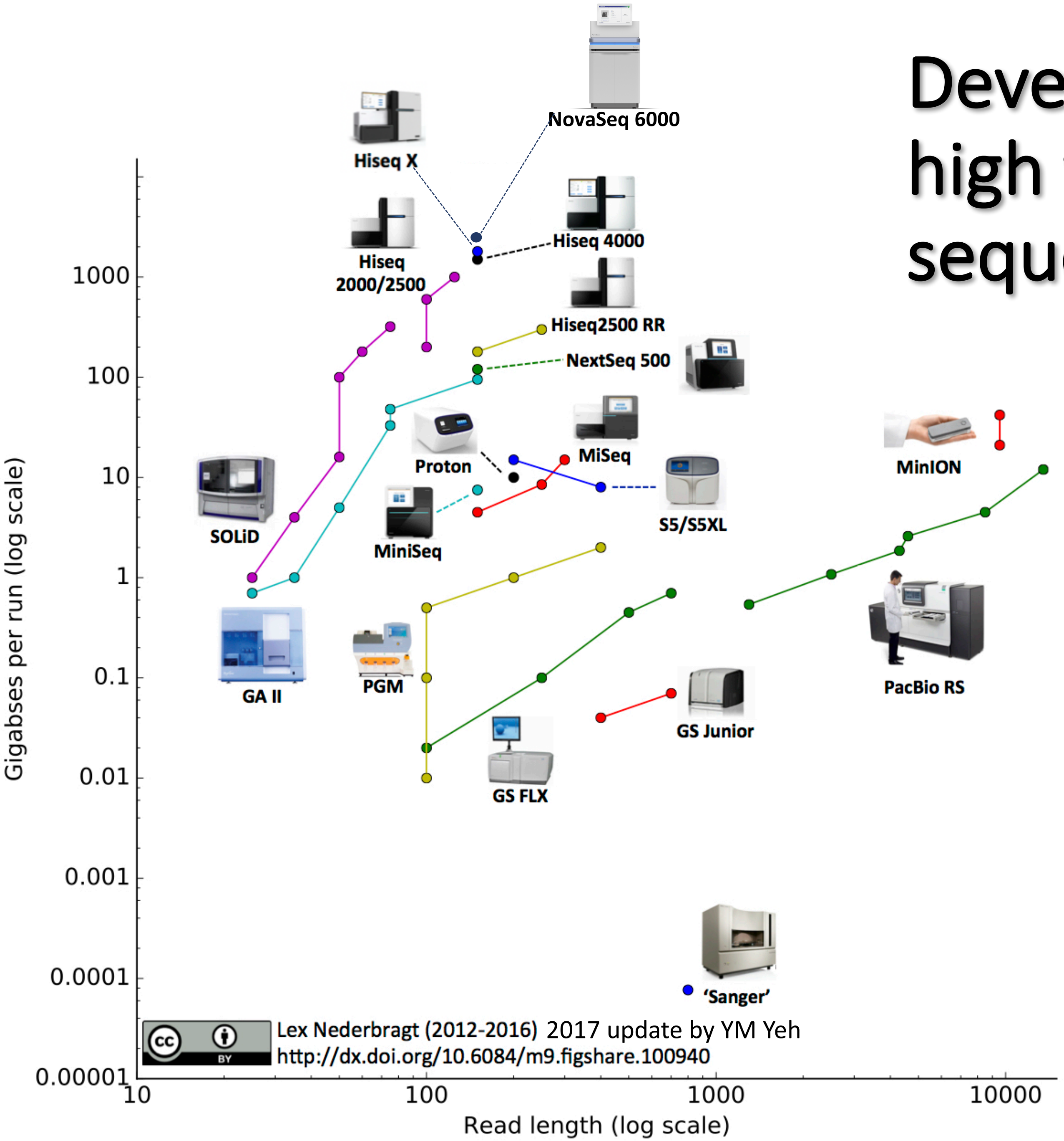
P. Tang (鄧致剛); YM Yeh (葉元鳴)
Bioinformatics Center, Chang Gung University.

High-throughput Sequencing (HTSeq) Platforms


Nice... We have so many options! - Deep whole genome - Low pass whole genome - Deep whole exome -Genomewide array - Exome array
How would you like to be sequenced?
Genomics Core Facility
www.biocomicals.com

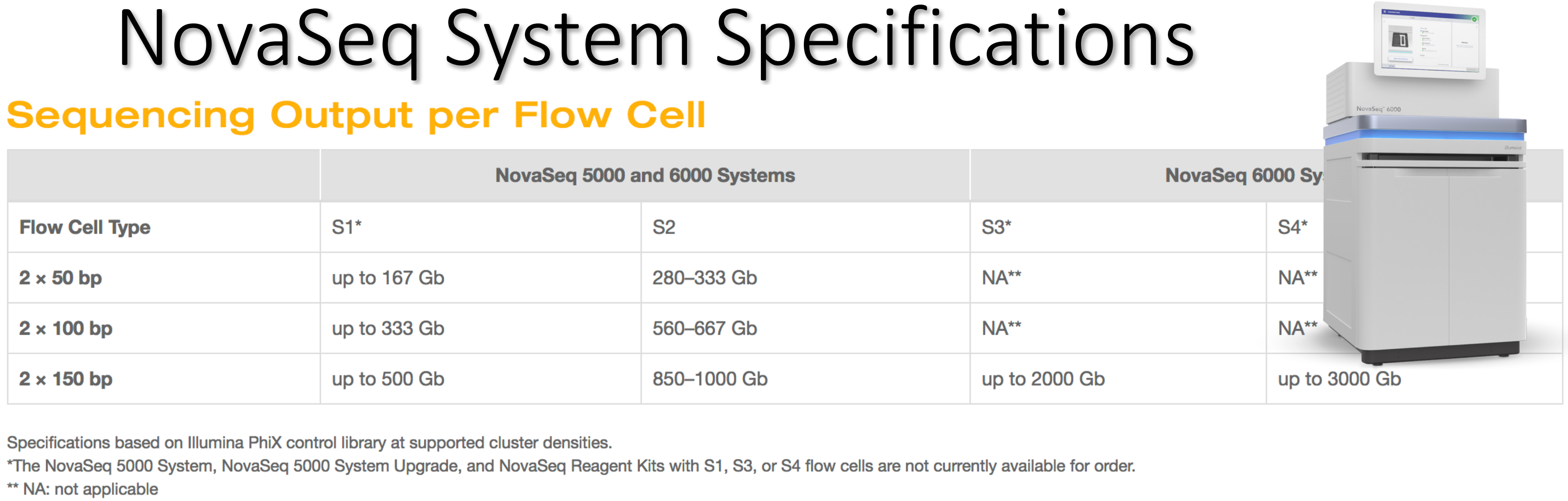
- 454 Sequencing / Roche
    - ~~GS Junior System~~
    - ~~GS FLX+ System~~
- Illumina
    - NovaSeq System
    - HiSeq System
    - NextSeq
    - MiniSeq/MiSeq
- Ion Torrent / Thermo
    - Personal Genome Machine
    - Proton
    - S5/S5XL
- Pacific Biosciences
    - PacBio RS
- Oxford Nanopore Teechnologies
    - MinION

Developments in high throughput sequencing



Lex Nederbragt (2012-2016) 2017 update by YM Yeh
http://dx.doi.org/10.6084/m9.figshare.100940

NovaSeq System Specifications



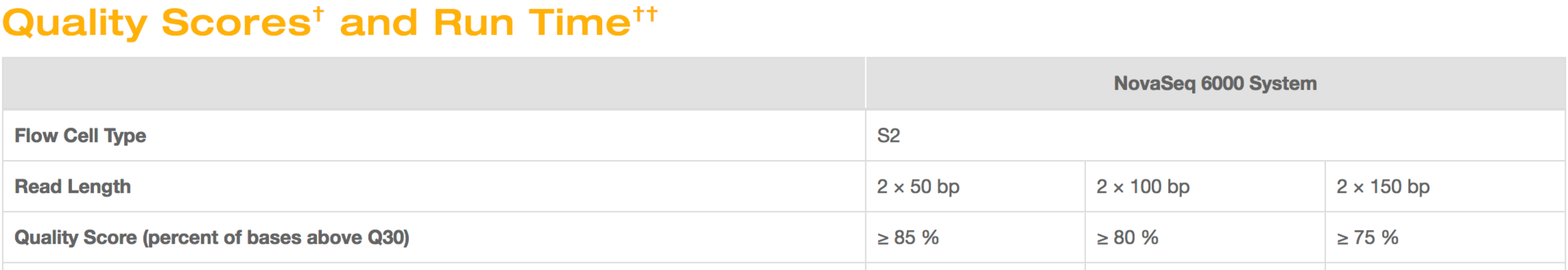
Sequencing Output per Flow Cell

	NovaSeq 5000 and 6000 Systems		NovaSeq 6000 Sy	
Flow Cell Type	S1*	S2	S3*	S4*
2 × 50 bp	up to 167 Gb	280–333 Gb	NA**	NA**
2 × 100 bp	up to 333 Gb	560–667 Gb	NA**	NA**
2 × 150 bp	up to 500 Gb	850–1000 Gb	up to 2000 Gb	up to 3000 Gb

Specifications based on Illumina PhiX control library at supported cluster densities.

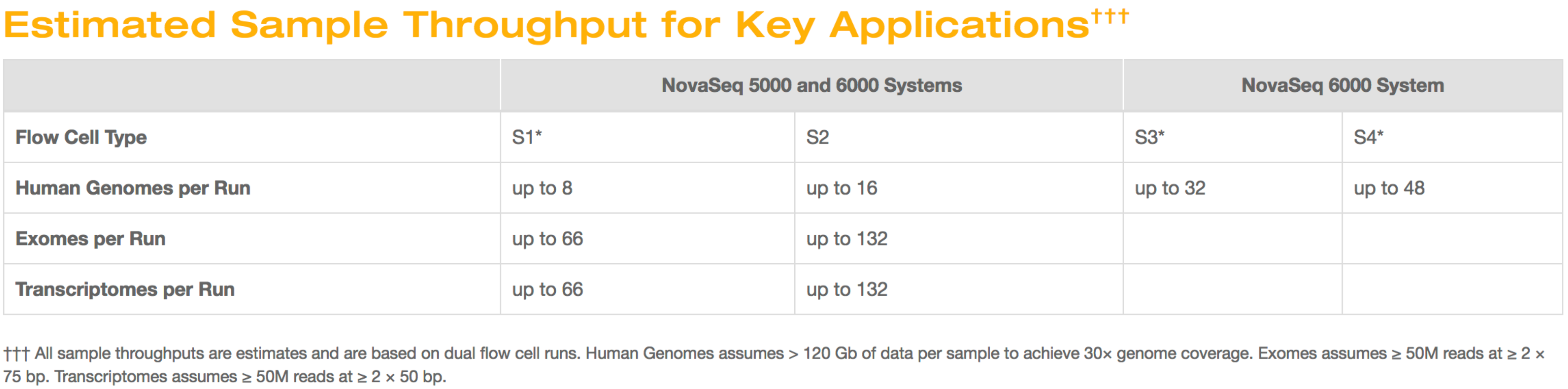
*The NovaSeq 5000 System, NovaSeq 5000 System Upgrade, and NovaSeq Reagent Kits with S1, S3, or S4 flow cells are not currently available for order.

** NA: not applicable



Quality Scores† and Run Time††

	NovaSeq 6000 System		
Flow Cell Type	S2		
Read Length	2 × 50 bp	2 × 100 bp	2 × 150 bp
Quality Score (percent of bases above Q30)	≥ 85 %	≥ 80 %	≥ 75 %



Estimated Sample Throughput for Key Applications†††

	NovaSeq 5000 and 6000 Systems		NovaSeq 6000 System	
Flow Cell Type	S1*	S2	S3*	S4*
Human Genomes per Run	up to 8	up to 16	up to 32	up to 48
Exomes per Run	up to 66	up to 132		
Transcriptomes per Run	up to 66	up to 132		

††† All sample throughputs are estimates and are based on dual flow cell runs. Human Genomes assumes > 120 Gb of data per sample to achieve 30× genome coverage. Exomes assumes ≥ 50M reads at ≥ 2 × 75 bp. Transcriptomes assumes ≥ 50M reads at ≥ 2 × 50 bp.

Compare the Illumina High-throughput machine

	S1	S2	S3	S4	2500 HO	4000	X
Reads per flowcell (billion)	1.6	3.3	6.6	10	2	2.8	3.44
Lanes per flowcell	2	2	4	4	8	8	8
Reads per lane (million)	800	1650	1650	2500	250	350	430
Throughput per lane (Gb)	240	495	495	750	62.5	105	129
Throughput per flowcell (Gb)	480	990	1980	3000	500	840	1032
Total Lanes	4	4	8	8	16	16	16
Total Flowcells	2	2	2	2	2	2	2
Run Throughput (Gb)	960	1980	3960	6000	1000	1680	2064
Run Time (days)	2-2.5	2-2.5	2-2.5	2-2.5	6	3.5	3

X 10

Interpreting raw data

Illumina

Capillary (e.g. AB 3730)

Roche 454

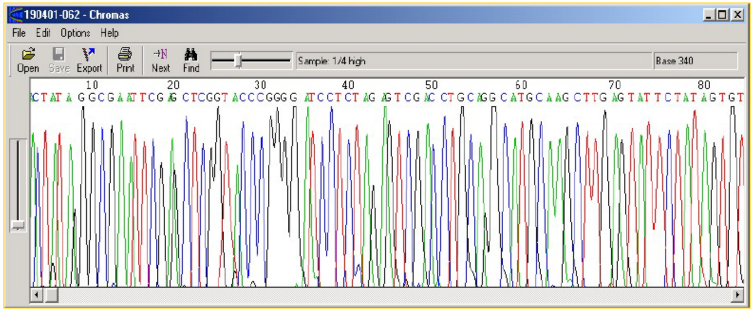
AB SOLiD

Helicos

PACIFIC
BIOSCIENCES™

Raw Data Format: fasta

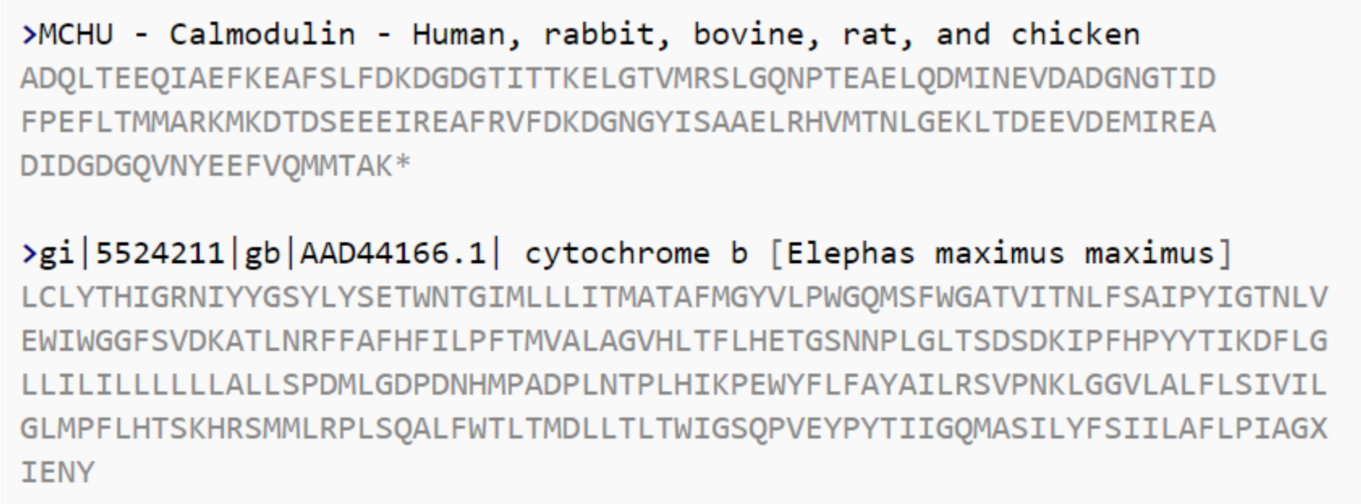
- fasta (Sanger)



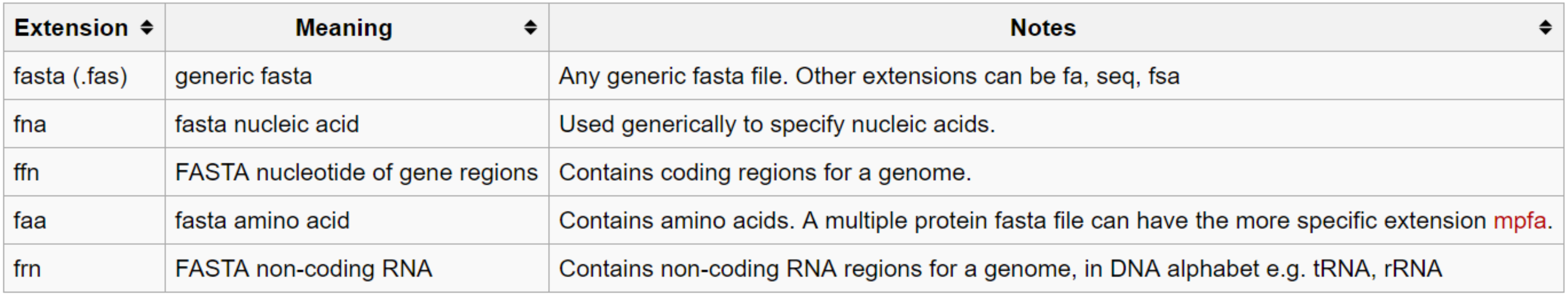
FASTA

Header line “>”

Sequence



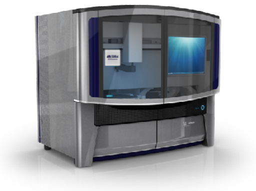
```
>MCHU - Calmodulin - Human, rabbit, bovine, rat, and chicken
ADQLTEEQIAEFKEAFSLFDKDGDGTITTKELGTVMRSLGQNPTEAELQDMINEVDADGNGTID
FPEFLTMMARKMKDTDSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREA
DIDGDGQVNYEEFVQMMTAK*

>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLV
EWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLG
LLILILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFLPIAGX
IENY
```



Extension ◆	Meaning ◆	Notes ◆
fasta (.fas)	generic fasta	Any generic fasta file. Other extensions can be fa, seq, fsa
fna	fasta nucleic acid	Used generically to specify nucleic acids.
ffn	FASTA nucleotide of gene regions	Contains coding regions for a genome.
faa	fasta amino acid	Contains amino acids. A multiple protein fasta file can have the more specific extension mpfa.
frn	FASTA non-coding RNA	Contains non-coding RNA regions for a genome, in DNA alphabet e.g. tRNA, rRNA

All Platforms have Errors

Illumina


SoLiD

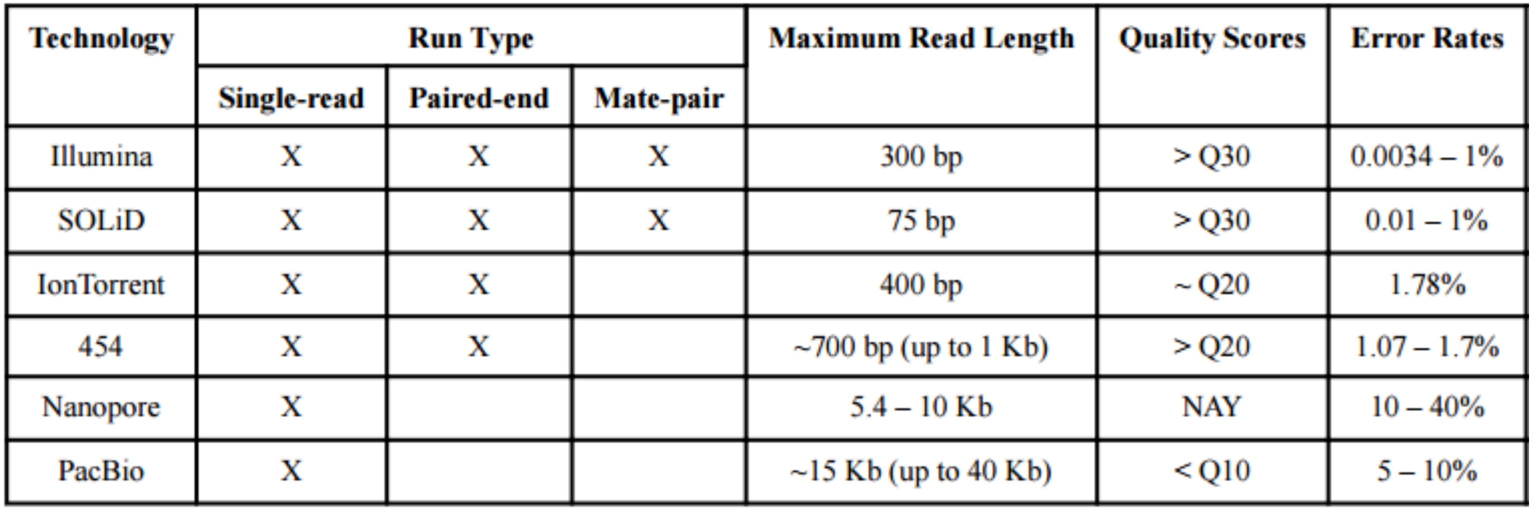
Ion Torrent

Roche 454

PacBio

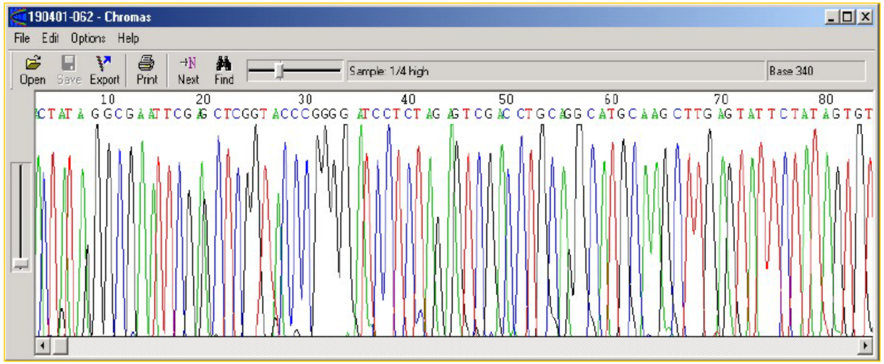
Nanopore

1. Removal of low quality bases/ Low complexity regions
2. Removal of adaptor sequences
3. Homopolymer-associated base call errors (3 or more identical DNA bases) causes higher number of (artificial) frameshifts

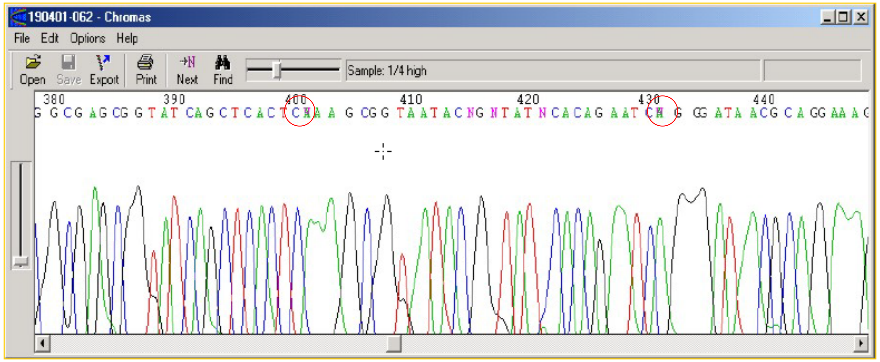


Technology	Run Type			Maximum Read Length	Quality Scores	Error Rates
	Single-read	Paired-end	Mate-pair			
Illumina	X	X	X	300 bp	> Q30	0.0034 – 1%
SOLiD	X	X	X	75 bp	> Q30	0.01 – 1%
IonTorrent	X	X		400 bp	~ Q20	1.78%
454	X	X		~700 bp (up to 1 Kb)	> Q20	1.07 – 1.7%
Nanopore	X			5.4 – 10 Kb	NAY	10 – 40%
PacBio	X			~15 Kb (up to 40 Kb)	< Q10	5 – 10%

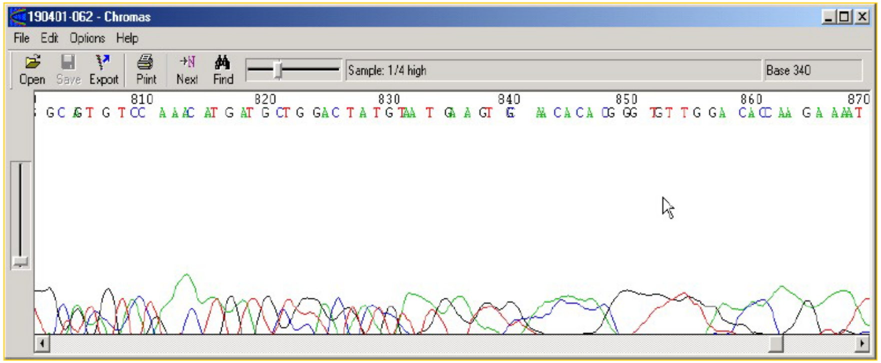
Trace File


190401-062 - Chromas
File Edit Options Help
Open Save Export Print Next Find
Sample: 1/4 high
Base 340
10 20 30 40 50 60 70 80
ACTATAGGCGAATTCGAGCTCGGTACCCGGGGATCCTCTAGAGTCGACCTGCAGGCATGCAAGCTTGAGTATTCTATAGTGT

High quality region - NO ambiguities (Ns)

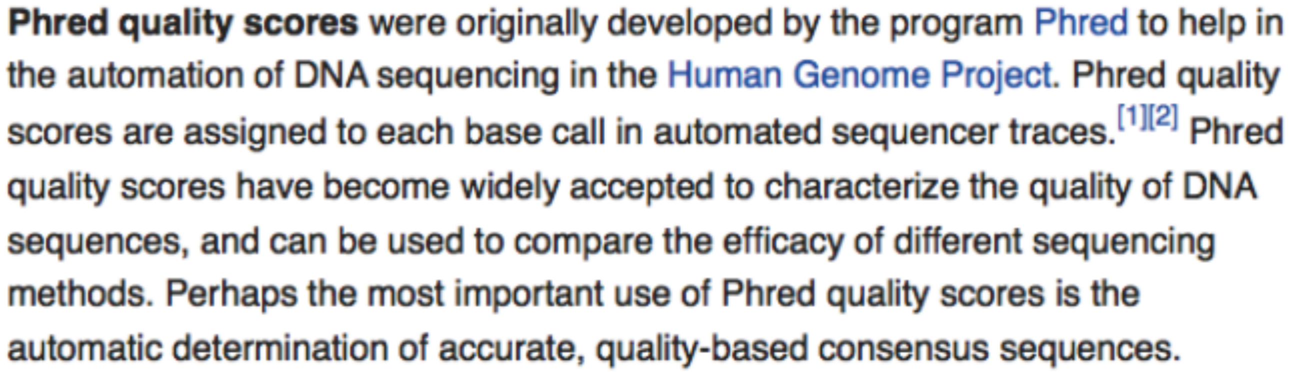

190401-062 - Chromas
File Edit Options Help
Open Save Export Print Next Find
Sample: 1/4 high
380 390 400 410 420 430 440
GGCGAGCGGTATCAGCTCACTCNAAGCGGTAATACNGNTATNCACAGAATCNGGGATAACGCAGGAAAG

Medium quality region - SOME ambiguities (Ns)


190401-062 - Chromas
File Edit Options Help
Open Save Export Print Next Find
Sample: 1/4 high
Base 340
810 820 830 840 850 860 870
GCAGTGTCCAAACATGATGCTGGACTATGTAATGAAGTGAACACACGGGTGTTGGACACCAAGAAAAT

Poor quality region - LOW confidence

Accessing Quality: phred scores



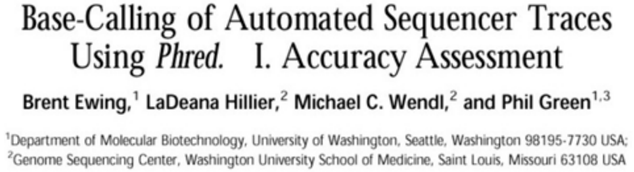
Phred quality scores were originally developed by the program Phred to help in the automation of DNA sequencing in the Human Genome Project. Phred quality scores are assigned to each base call in automated sequencer traces.[1][2] Phred quality scores have become widely accepted to characterize the quality of DNA sequences, and can be used to compare the efficacy of different sequencing methods. Perhaps the most important use of Phred quality scores is the automatic determination of accurate, quality-based consensus sequences.

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

$$Q = -10\log_{10} P$$

P=error probability of a given base call

Genome Research **8**: 175-185, 1998

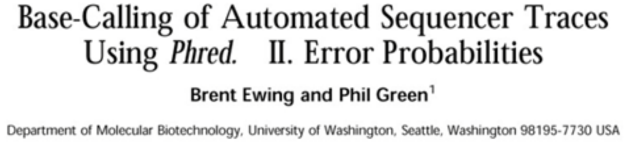


Base-Calling of Automated Sequencer Traces Using *Phred.* I. Accuracy Assessment

Brent Ewing,[1] LaDeana Hillier,[2] Michael C. Wendl,[2] and Phil Green[1,3]

[1]Department of Molecular Biotechnology, University of Washington, Seattle, Washington 98195-7730 USA;
[2]Genome Sequencing Center, Washington University School of Medicine, Saint Louis, Missouri 63108 USA

Genome Research **8**: 186-194, 1998

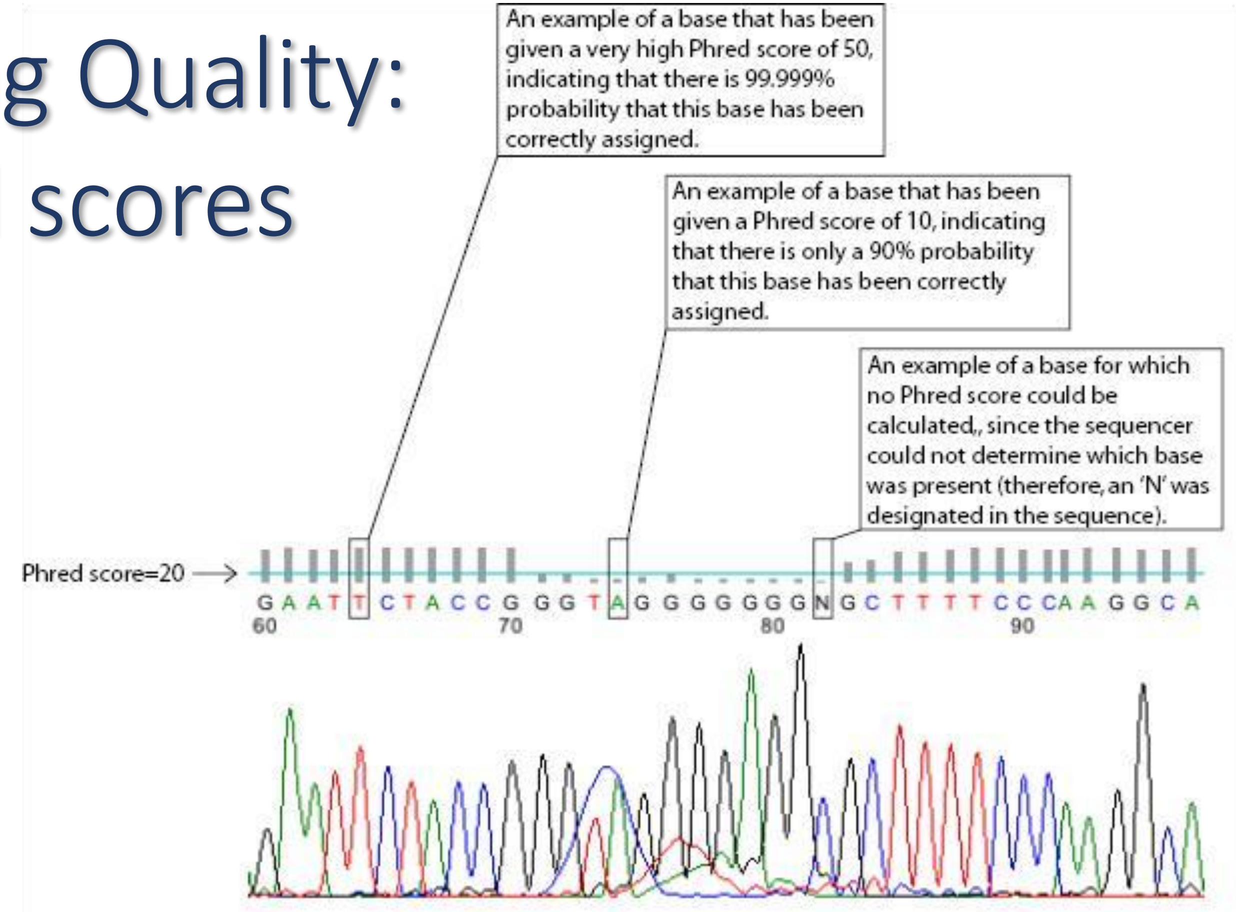


Base-Calling of Automated Sequencer Traces Using *Phred.* II. Error Probabilities

Brent Ewing and Phil Green[1]

Department of Molecular Biotechnology, University of Washington, Seattle, Washington 98195-7730 USA

Accessing Quality: phred scores

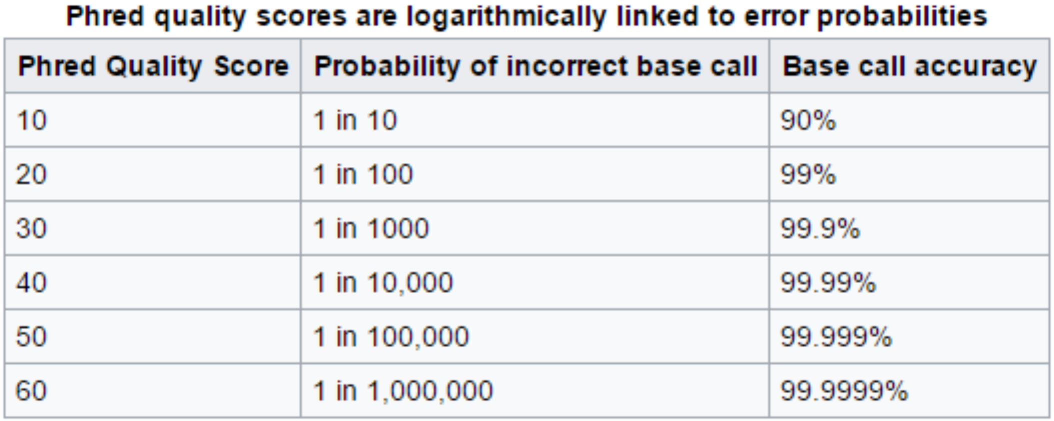


An example of a base that has been given a very high Phred score of 50, indicating that there is 99.999% probability that this base has been correctly assigned.

An example of a base that has been given a Phred score of 10, indicating that there is only a 90% probability that this base has been correctly assigned.

An example of a base for which no Phred score could be calculated,, since the sequencer could not determine which base was present (therefore, an 'N' was designated in the sequence).

Phred score=20
G A A T T C T A C C G G G T A G G G G G G N G C T T T T C C C A A G G C A
60
70
80
90

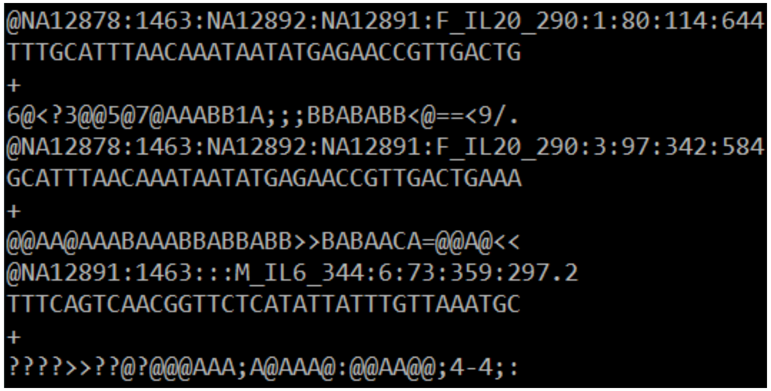


Phred quality scores are logarithmically linked to error probabilities

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

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

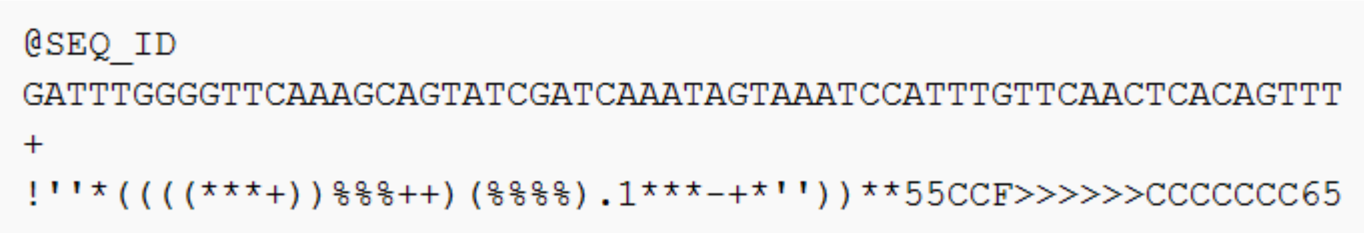
Raw Data Format: fastq



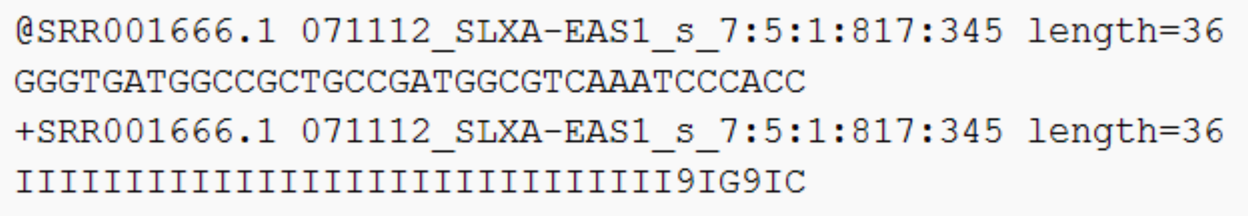
```
@NA12878:1463:NA12892:NA12891:F_IL20_290:1:80:114:644
TTTGCATTTAACAAATAATATGAGAACCGTTGACTG
+
6@<?3@@5@7@AAABB1A;;;BBABABB<@==<9/.
@NA12878:1463:NA12892:NA12891:F_IL20_290:3:97:342:584
GCATTTAACAAATAATATGAGAACCGTTGACTGAAA
+
@@AA@AAABAAABBABBABB>>BABAACA=@@A@<<
@NA12891:1463:::M_IL6_344:6:73:359:297.2
TTTCAGTCAACGGTTCTCATATTATTTGTTAAATGC
+
????>>??@?@@@AAA;A@AAA@:@@AA@@;4-4;:
```

- **FASTA**
  - Header line “>”
  - Sequence
- **FASTQ**
  - Add QVs encoded as single byte ASCII codes
- Most aligners accept FASTA/Q as input
- Issue: data is volumous (2 bytes per base for FASTQ)
- Do PHRED scaled values provide the most information?

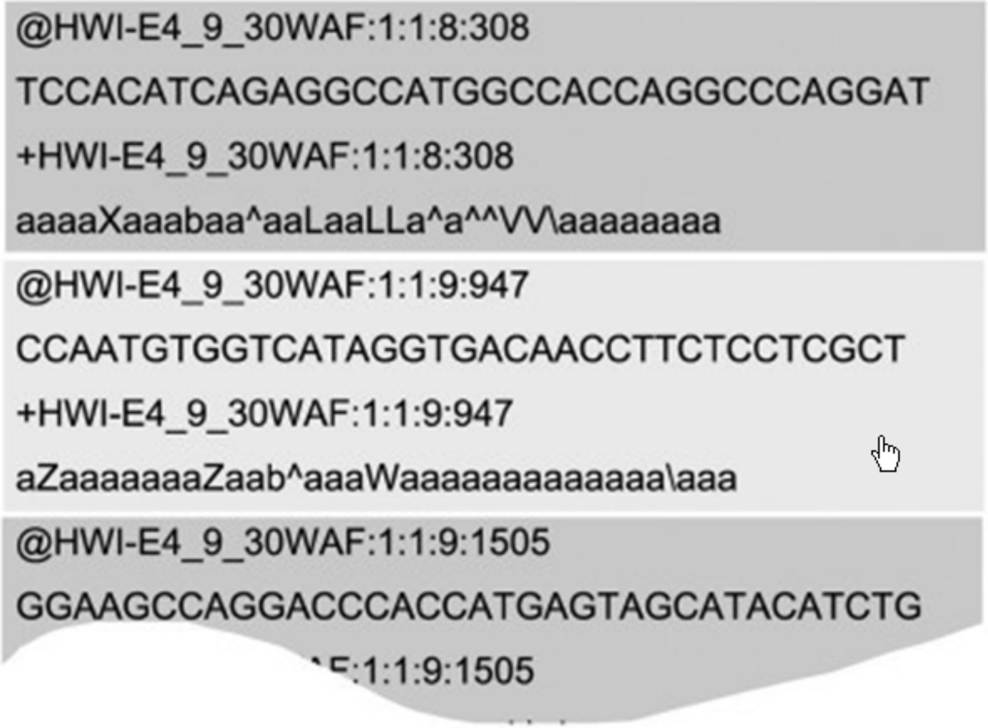
Raw Data Format: fastq



```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

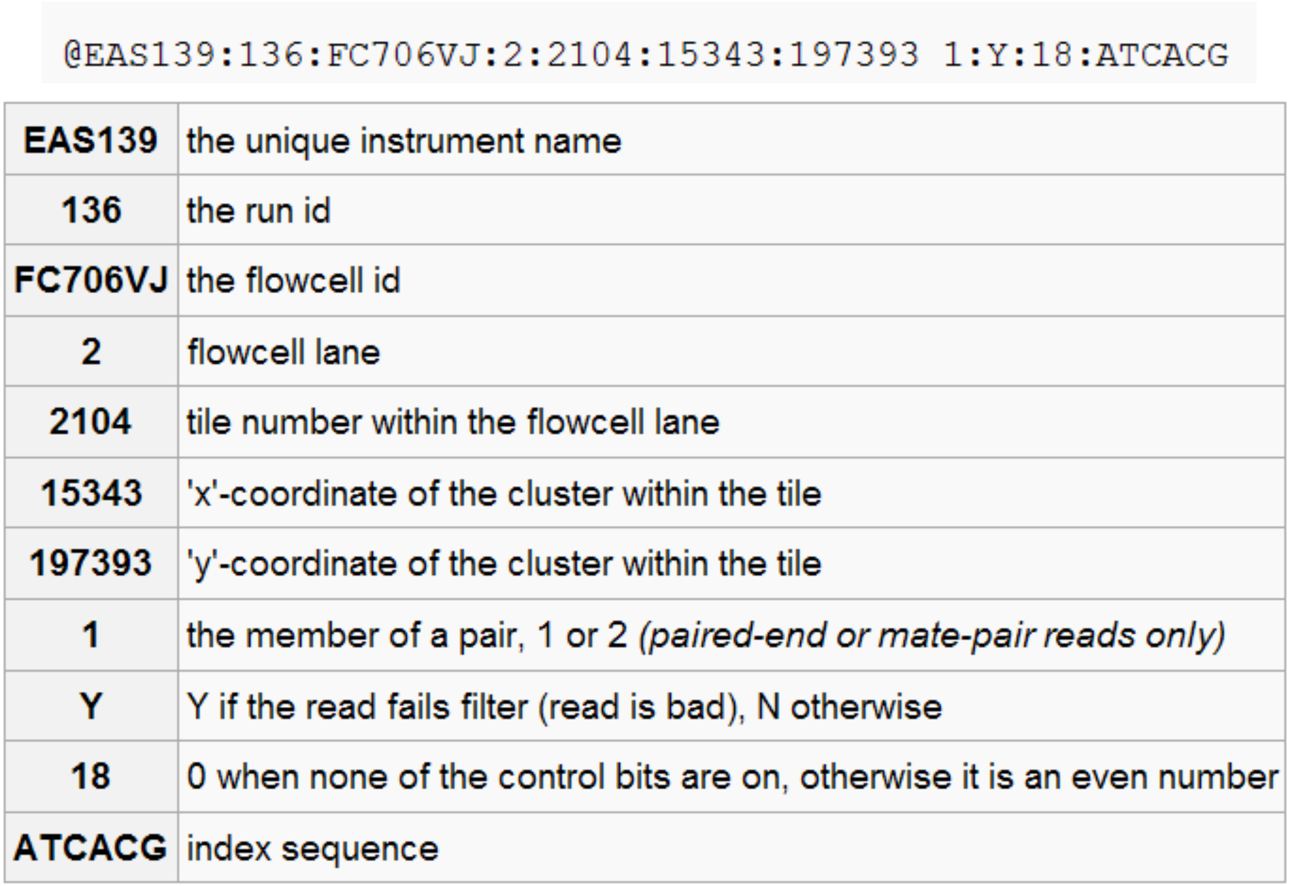


```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```



```
@HWI-E4_9_30WAF:1:1:8:308
TCCACATCAGAGGCCATGGCCACCAGGCCCAGGAT
+HWI-E4_9_30WAF:1:1:8:308
aaaaXaaabaa^aaLaaLLa^a^^VV\aaaaaaaa
@HWI-E4_9_30WAF:1:1:9:947
CCAATGTGGTCATAGGTGACAACCTTCTCCTCGCT
+HWI-E4_9_30WAF:1:1:9:947
aZaaaaaaaZaab^aaaWaaaaaaaaaaaaa\aaa
@HWI-E4_9_30WAF:1:1:9:1505
GGAAGCCAGGACCCACCATGAGTAGCATACATCTG
F:1:1:9:1505
```

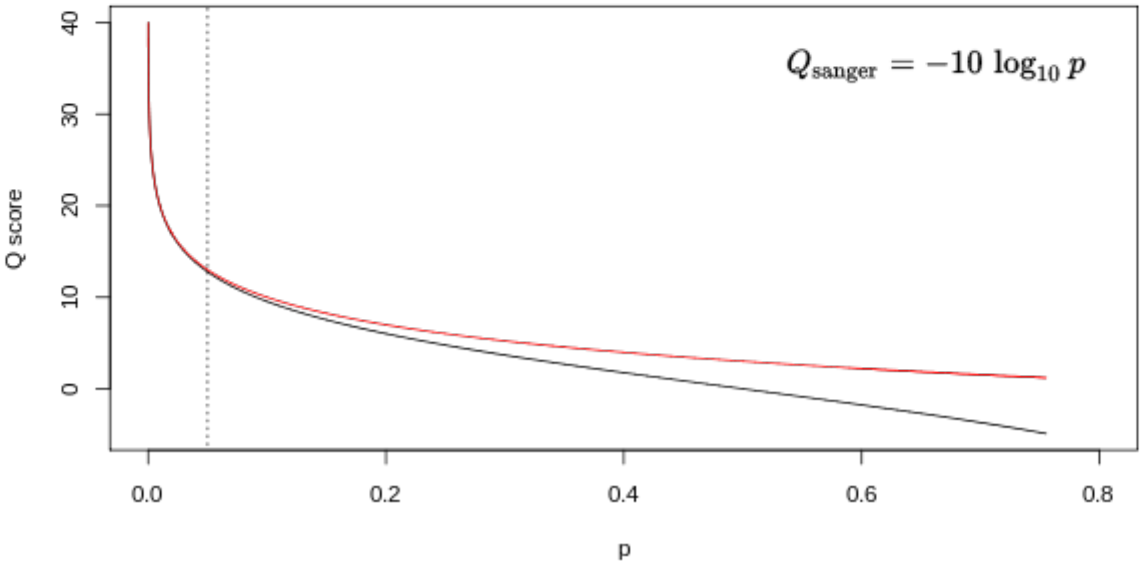
Raw Data Format: fastq



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

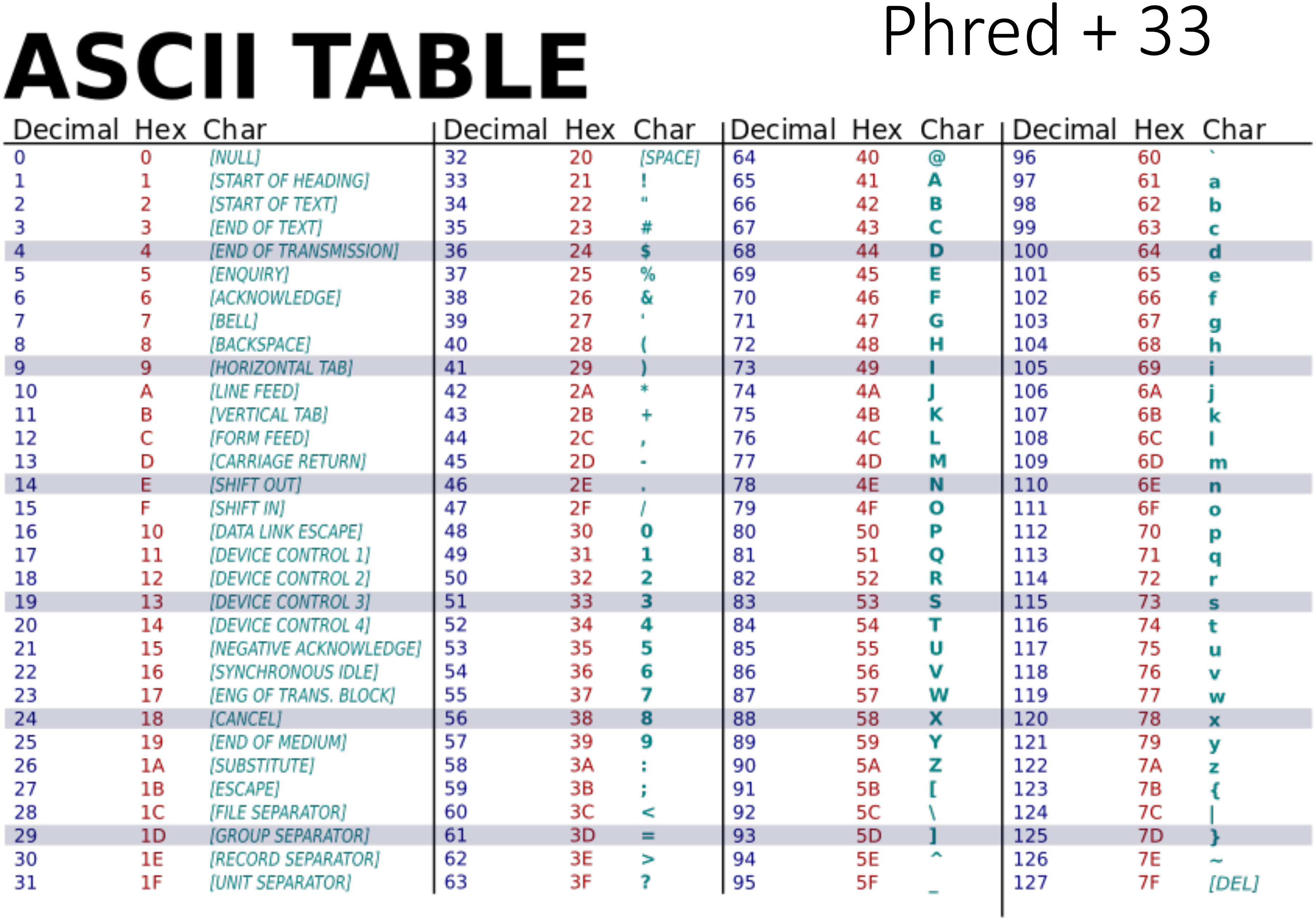
EAS139	the unique instrument name
136	the run id
FC706VJ	the flowcell id
2	flowcell lane
2104	tile number within the flowcell lane
15343	'x'-coordinate of the cluster within the tile
197393	'y'-coordinate of the cluster within the tile
1	the member of a pair, 1 or 2 *(paired-end or mate-pair reads only)*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

Fastq Quality



Relationship between *Q* and *p* using the Sanger (red) and Solexa (black) equations (described above). The vertical dotted line indicates *p* = 0.05, or equivalently, *Q* ≈ 13.

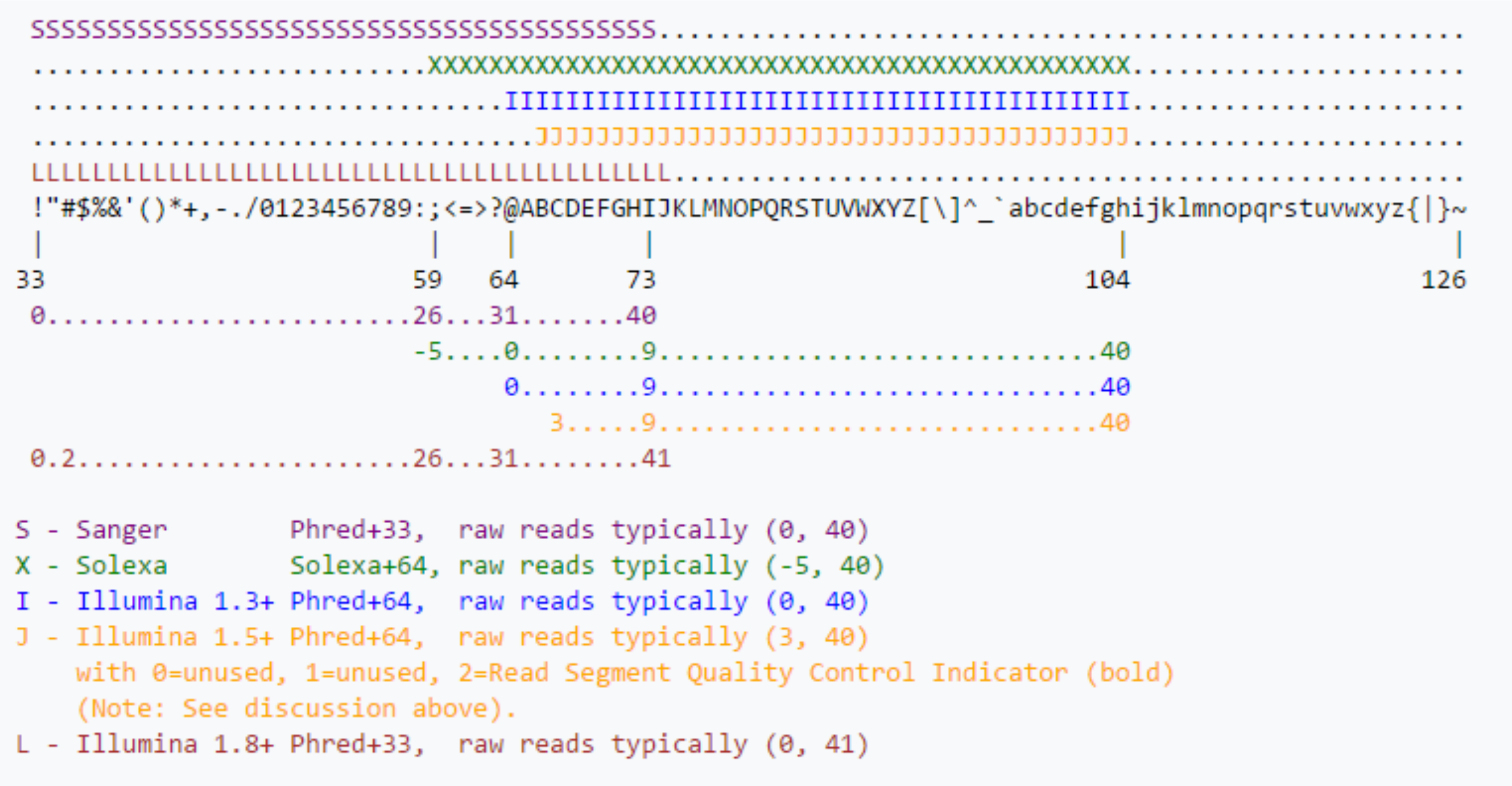
Phred + 33



ASCII TABLE

Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char
0	0	[NULL]	32	20	[SPACE]	64	40	@	96	60	`
1	1	[START OF HEADING]	33	21	!	65	41	A	97	61	a
2	2	[START OF TEXT]	34	22	"	66	42	B	98	62	b
3	3	[END OF TEXT]	35	23	#	67	43	C	99	63	c
4	4	[END OF TRANSMISSION]	36	24	$	68	44	D	100	64	d
5	5	[ENQUIRY]	37	25	%	69	45	E	101	65	e
6	6	[ACKNOWLEDGE]	38	26	&	70	46	F	102	66	f
7	7	[BELL]	39	27	'	71	47	G	103	67	g
8	8	[BACKSPACE]	40	28	(	72	48	H	104	68	h
9	9	[HORIZONTAL TAB]	41	29	)	73	49	I	105	69	i
10	A	[LINE FEED]	42	2A	*	74	4A	J	106	6A	j
11	B	[VERTICAL TAB]	43	2B	+	75	4B	K	107	6B	k
12	C	[FORM FEED]	44	2C	,	76	4C	L	108	6C	l
13	D	[CARRIAGE RETURN]	45	2D	-	77	4D	M	109	6D	m
14	E	[SHIFT OUT]	46	2E	.	78	4E	N	110	6E	n
15	F	[SHIFT IN]	47	2F	/	79	4F	O	111	6F	o
16	10	[DATA LINK ESCAPE]	48	30	0	80	50	P	112	70	p
17	11	[DEVICE CONTROL 1]	49	31	1	81	51	Q	113	71	q
18	12	[DEVICE CONTROL 2]	50	32	2	82	52	R	114	72	r
19	13	[DEVICE CONTROL 3]	51	33	3	83	53	S	115	73	s
20	14	[DEVICE CONTROL 4]	52	34	4	84	54	T	116	74	t
21	15	[NEGATIVE ACKNOWLEDGE]	53	35	5	85	55	U	117	75	u
22	16	[SYNCHRONOUS IDLE]	54	36	6	86	56	V	118	76	v
23	17	[ENG OF TRANS. BLOCK]	55	37	7	87	57	W	119	77	w
24	18	[CANCEL]	56	38	8	88	58	X	120	78	x
25	19	[END OF MEDIUM]	57	39	9	89	59	Y	121	79	y
26	1A	[SUBSTITUTE]	58	3A	:	90	5A	Z	122	7A	z
27	1B	[ESCAPE]	59	3B	;	91	5B	[	123	7B	{
28	1C	[FILE SEPARATOR]	60	3C	<	92	5C	\	124	7C	|
29	1D	[GROUP SEPARATOR]	61	3D	=	93	5D	]	125	7D	}
30	1E	[RECORD SEPARATOR]	62	3E	>	94	5E	^	126	7E	~
31	1F	[UNIT SEPARATOR]	63	3F	?	95	5F	_	127	7F	[DEL]

Fastq Quality Encoding



```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
 ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
 .................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                        59   64       73                            104                   126
 0........................26...31.......40
                          -5....0........9.............................40
                                0........9.............................40
                                   3.....9.............................40
 0.2......................26...31........41

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

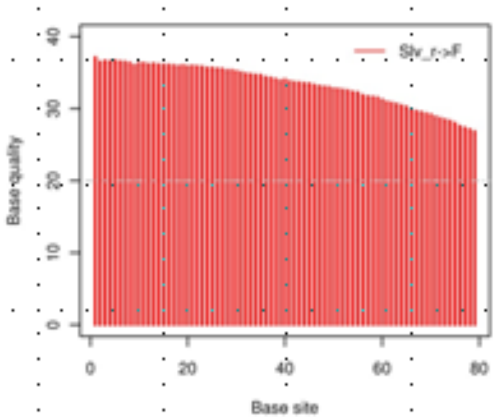
Quality Control

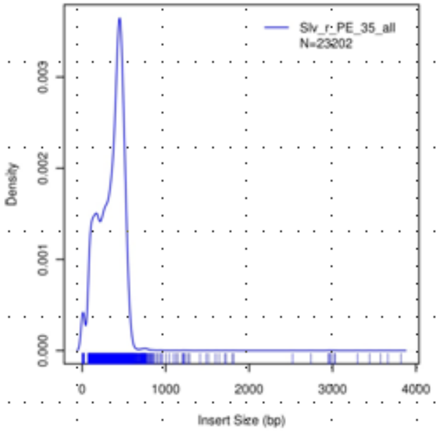
Read quality distribution

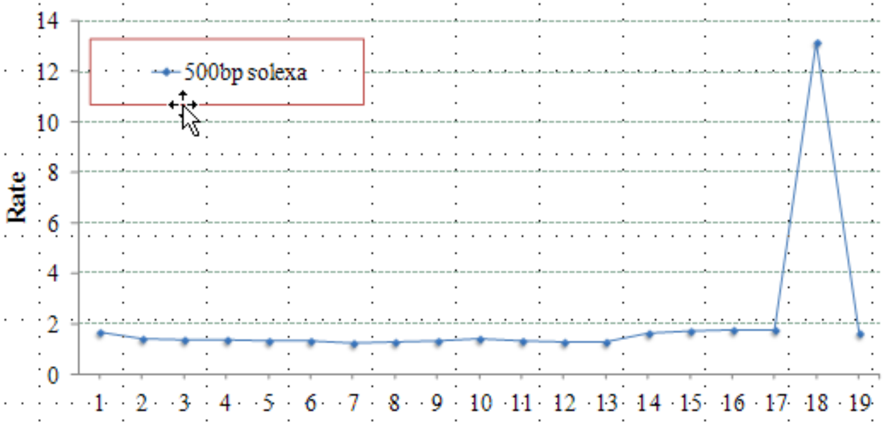
Library insert size

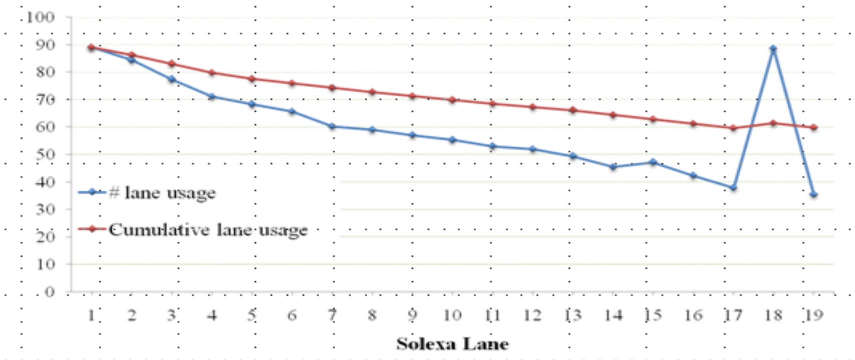
Mapping Rate

Duplication assessment

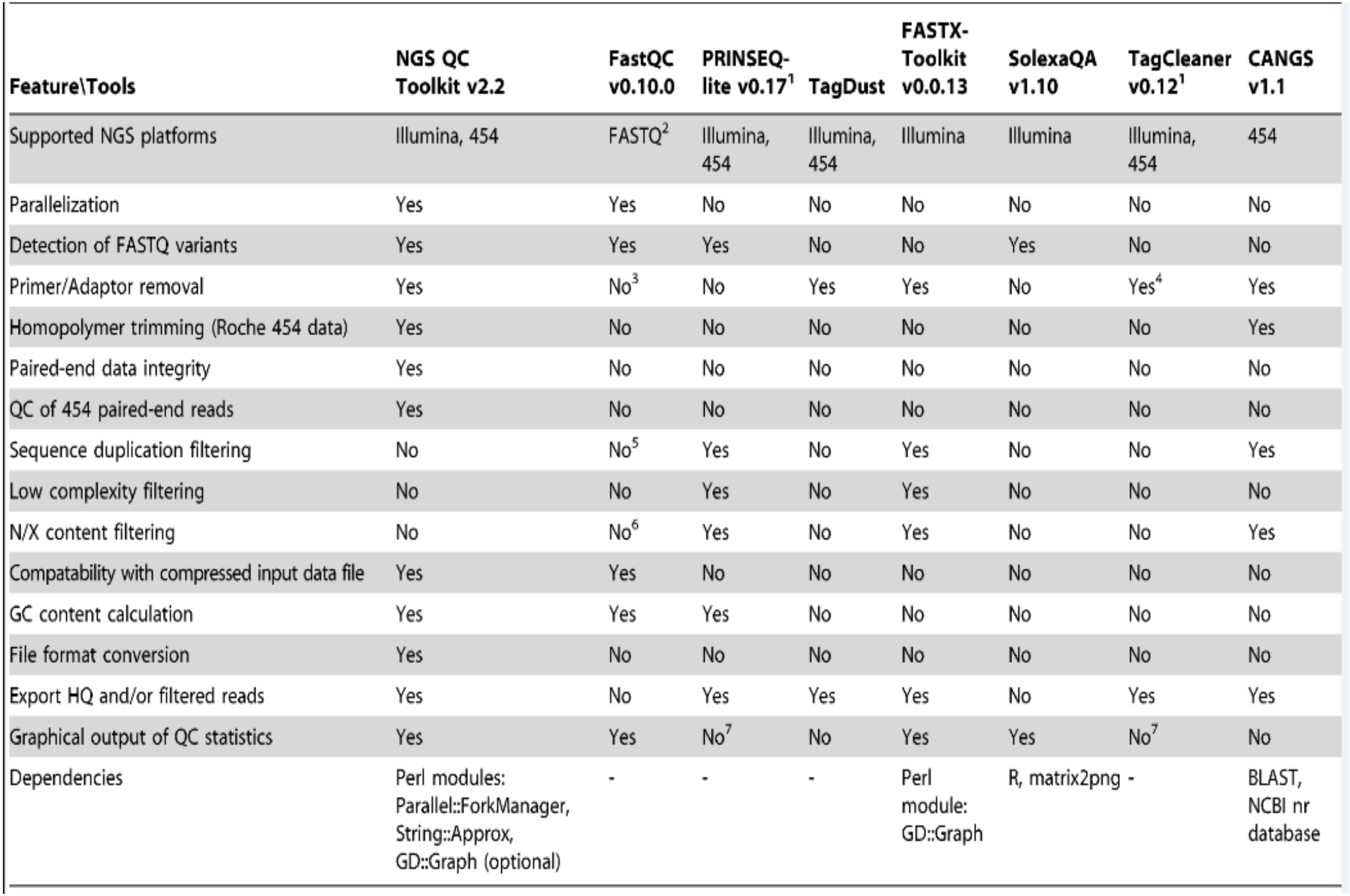






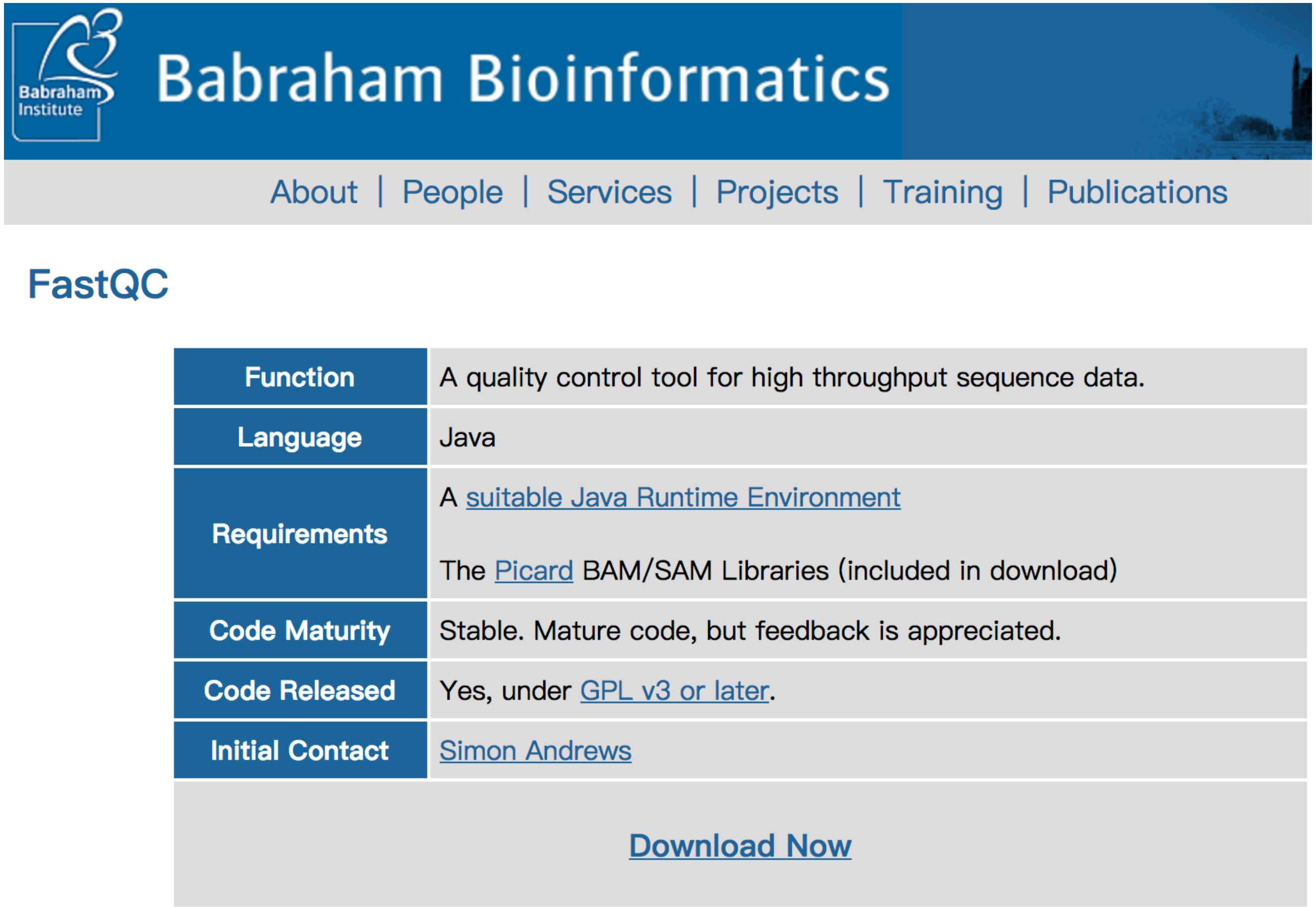


Quality Control Tools



Feature\Tools	NGS QC Toolkit v2.2	FastQC v0.10.0	PRINSEQ-lite v0.17[1]	TagDust	FASTX-Toolkit v0.0.13	SolexaQA v1.10	TagCleaner v0.12[1]	CANGS v1.1
Supported NGS platforms	Illumina, 454	FASTQ[2]	Illumina, 454	Illumina, 454	Illumina	Illumina	Illumina, 454	454
Parallelization	Yes	Yes	No	No	No	No	No	No
Detection of FASTQ variants	Yes	Yes	Yes	No	No	Yes	No	No
Primer/Adaptor removal	Yes	No[3]	No	Yes	Yes	No	Yes[4]	Yes
Homopolymer trimming (Roche 454 data)	Yes	No	No	No	No	No	No	Yes
Paired-end data integrity	Yes	No	No	No	No	No	No	No
QC of 454 paired-end reads	Yes	No	No	No	No	No	No	No
Sequence duplication filtering	No	No[5]	Yes	No	Yes	No	No	Yes
Low complexity filtering	No	No	Yes	No	Yes	No	No	No
N/X content filtering	No	No[6]	Yes	No	Yes	No	No	Yes
Compatability with compressed input data file	Yes	Yes	No	No	No	No	No	No
GC content calculation	Yes	Yes	Yes	No	No	No	No	No
File format conversion	Yes	No	No	No	No	No	No	No
Export HQ and/or filtered reads	Yes	No	Yes	Yes	Yes	No	Yes	Yes
Graphical output of QC statistics	Yes	Yes	No[7]	No	Yes	Yes	No[7]	No
Dependencies	Perl modules: Parallel::ForkManager, String::Approx, GD::Graph (optional)	-	-	-	Perl module: GD::Graph	R, matrix2png	-	BLAST, NCBI nr database

FastQC

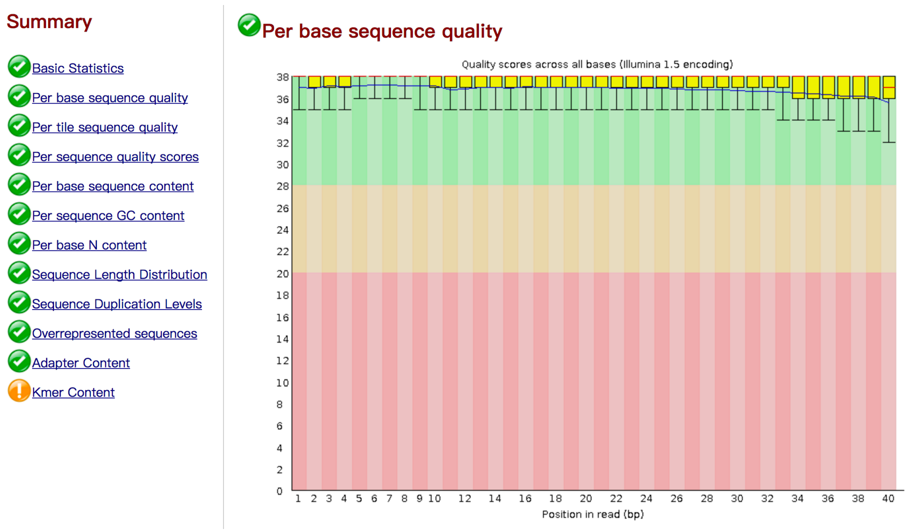

Babraham Institute
Babraham Bioinformatics
About | People | Services | Projects | Training | Publications

FastQC

Function	A quality control tool for high throughput sequence data.
Language	Java
Requirements	A suitable Java Runtime Environment The Picard BAM/SAM Libraries (included in download)
Code Maturity	Stable. Mature code, but feedback is appreciated.
Code Released	Yes, under GPL v3 or later.
Initial Contact	Simon Andrews

Download Now

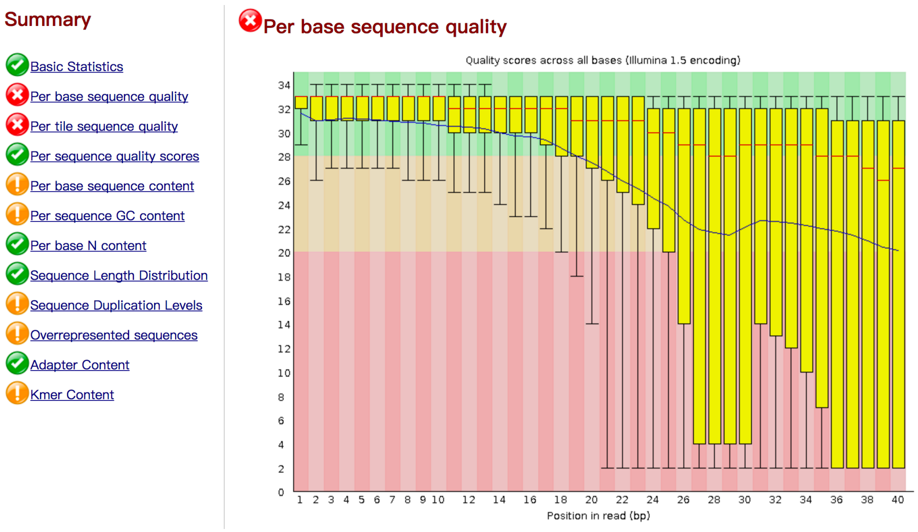
Example Reports



Summary

- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per tile sequence quality
- ✓ 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
- ! Kmer Content

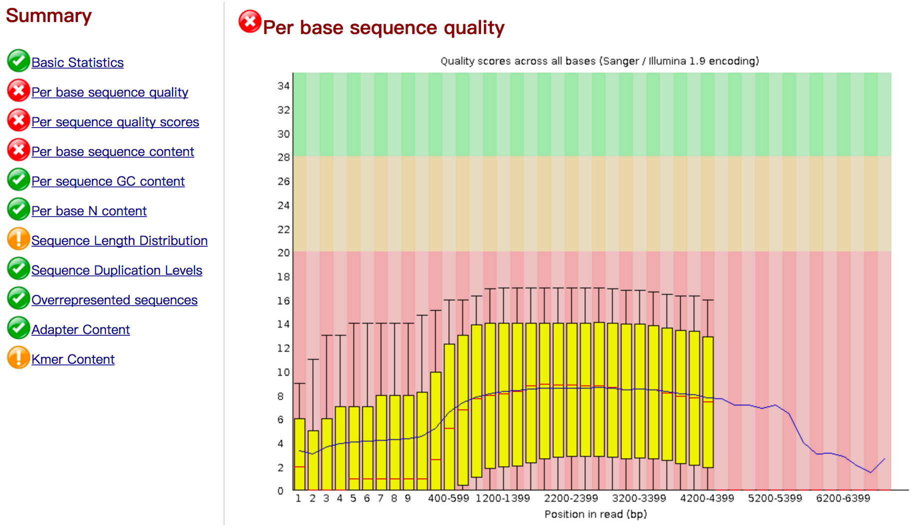
✓ **Per base sequence quality**



Summary

- ✓ Basic Statistics
- ✗ Per base sequence quality
- ✗ Per tile sequence quality
- ✓ 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
- ! Kmer Content

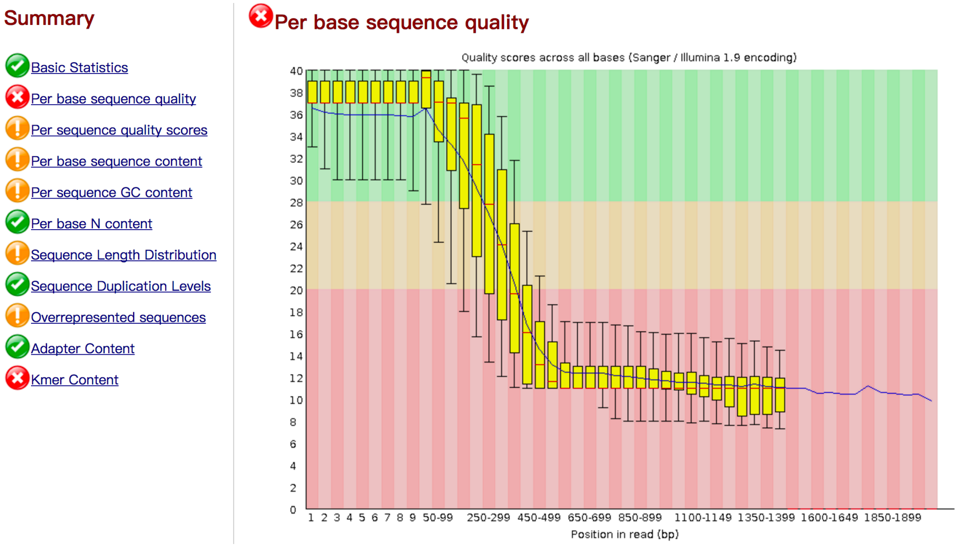
✗ **Per base sequence quality**



Summary

- ✓ Basic Statistics
- ✗ Per base sequence quality
- ✗ 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
- ! Kmer Content

✗ **Per base sequence quality**

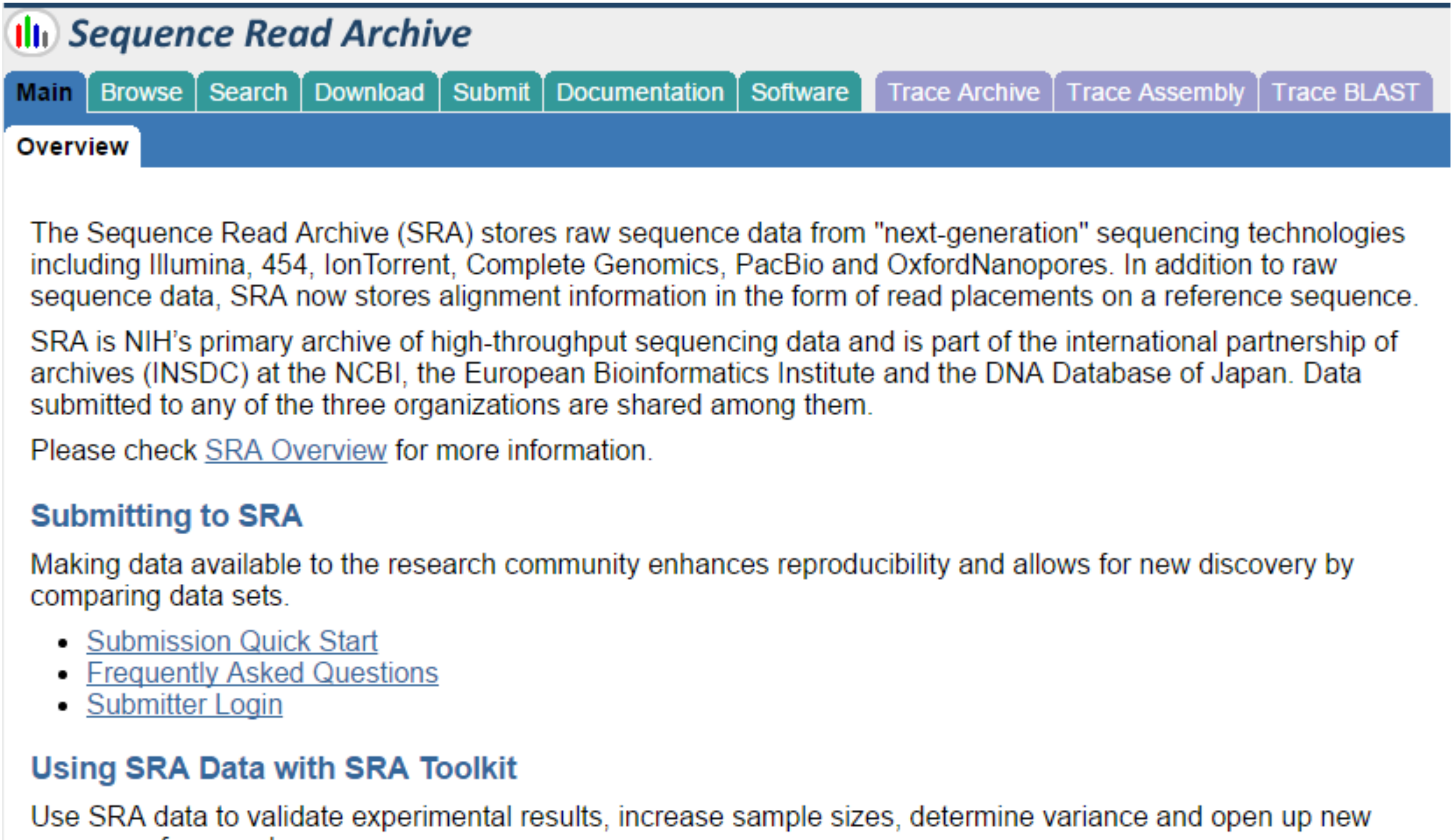


Summary

- ✓ Basic Statistics
- ✗ Per base sequence quality
- ! 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
- ✗ Kmer Content

✗ **Per base sequence quality**

SRA



The Sequence Read Archive (SRA) stores raw sequence data from "next-generation" sequencing technologies including Illumina, 454, IonTorrent, Complete Genomics, PacBio and OxfordNanopores. In addition to raw sequence data, SRA now stores alignment information in the form of read placements on a reference sequence.

SRA is NIH's primary archive of high-throughput sequencing data and is part of the international partnership of archives (INSDC) at the NCBI, the European Bioinformatics Institute and the DNA Database of Japan. Data submitted to any of the three organizations are shared among them.

Please check SRA Overview for more information.

Submitting to SRA

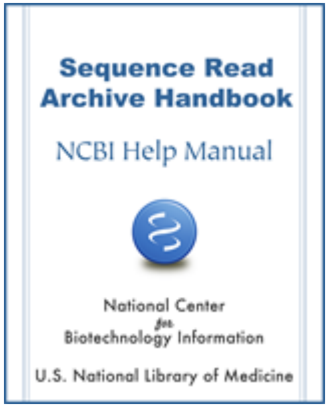
Making data available to the research community enhances reproducibility and allows for new discovery by comparing data sets.

- Submission Quick Start
- Frequently Asked Questions
- Submitter Login

Using SRA Data with SRA Toolkit

Use SRA data to validate experimental results, increase sample sizes, determine variance and open up new avenues of research.

- Documentation
- Usage Guide
- Download
- Get sources code on GitHub (for developers using SRA)

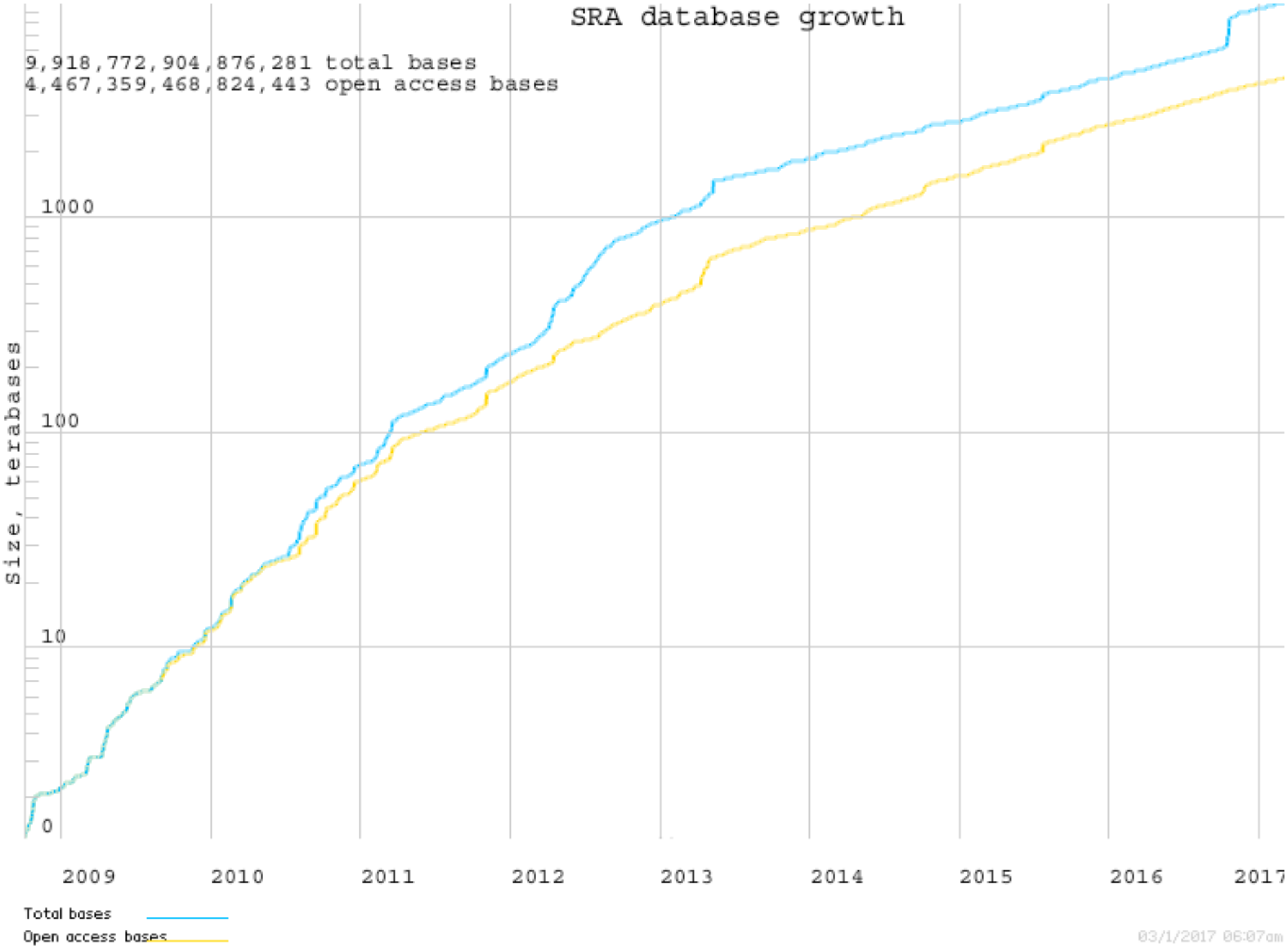


Sequence Read Archive Handbook

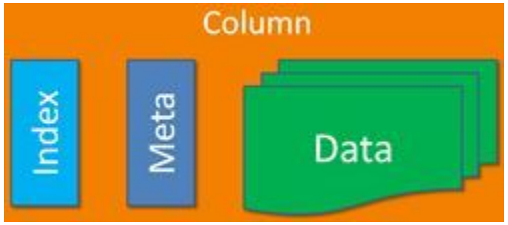
NCBI Help Manual

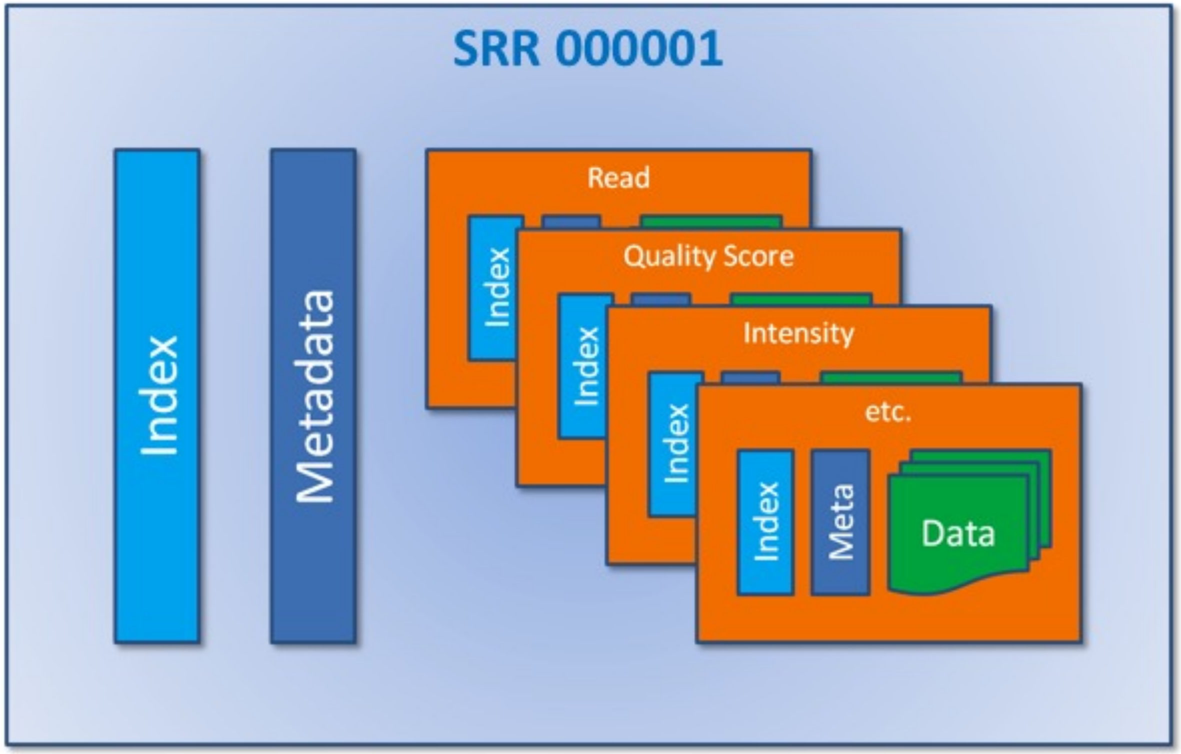
National Center for Biotechnology Information

U.S. National Library of Medicine


SRA database growth
9,918,772,904,876,281 total bases
4,467,359,468,824,443 open access bases
Size, terabases
1000
100
10
0
2009
2010
2011
2012
2013
2014
2015
2016
2017
Total bases
Open access bases
03/1/2017 06:07am

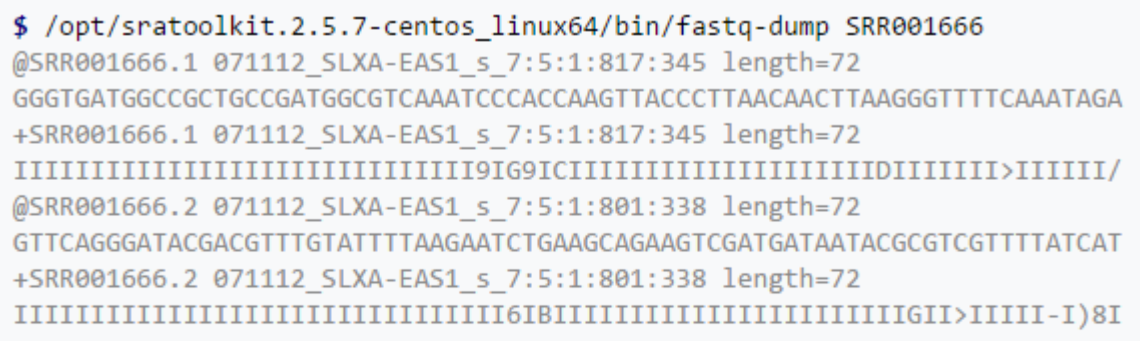
SRA Data Structure



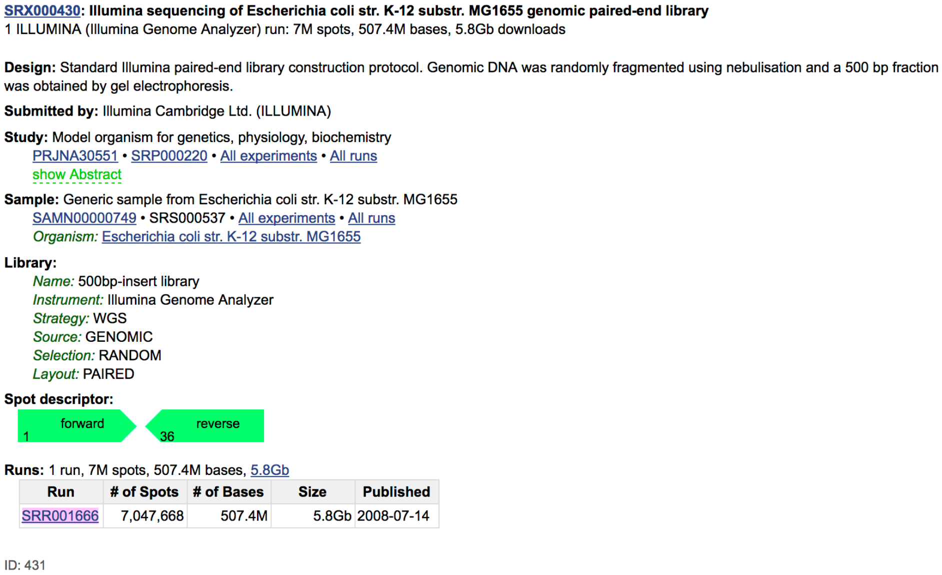


NCBI Sequence Read Archive (fastq)

an NCBI-assigned identifier, and the description holds the original identifier from Solexa/Illumina (as described above) plus the read length. Sequencing was performed in paired-end mode (~500bp insert size), see SRR001666. Notably in the above output the paired-end information was lost when the data was extracted from the NCBI SRA using fastq-dump with default settings.



```
$ /opt/sratoolkit.2.5.7-centos_linux64/bin/fastq-dump SRR001666
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=72
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACCAAGTTACCCTTAACAACTTAAGGGTTTTCAAATAGA
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=72
IIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9ICIIIIIIIIIIIIIIIIIIIIDIIIIIII>IIIIII/
@SRR001666.2 071112_SLXA-EAS1_s_7:5:1:801:338 length=72
GTTCAGGGATACGACGTTTGTATTTTAAGAATCTGAAGCAGAAGTCGATGATAATACGCGTCGTTTTATCAT
+SRR001666.2 071112_SLXA-EAS1_s_7:5:1:801:338 length=72
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII6IBIIIIIIIIIIIIIIIIIIIIIIIGII>IIIII-I)8I
```



SRX000430: Illumina sequencing of Escherichia coli str. K-12 substr. MG1655 genomic paired-end library
1 ILLUMINA (Illumina Genome Analyzer) run: 7M spots, 507.4M bases, 5.8Gb downloads

Design: Standard Illumina paired-end library construction protocol. Genomic DNA was randomly fragmented using nebulisation and a 500 bp fraction was obtained by gel electrophoresis.

Submitted by: Illumina Cambridge Ltd. (ILLUMINA)

Study: Model organism for genetics, physiology, biochemistry
PRJNA30551 • SRP000220 • All experiments • All runs
show Abstract

Sample: Generic sample from Escherichia coli str. K-12 substr. MG1655
SAMN00000749 • SRS000537 • All experiments • All runs
Organism: Escherichia coli str. K-12 substr. MG1655

Library:
Name: 500bp-insert library
Instrument: Illumina Genome Analyzer
Strategy: WGS
Source: GENOMIC
Selection: RANDOM
Layout: PAIRED

Spot descriptor:

Runs: 1 run, 7M spots, 507.4M bases, 5.8Gb

Run	# of Spots	# of Bases	Size	Published
SRR001666	7,047,668	507.4M	5.8Gb	2008-07-14

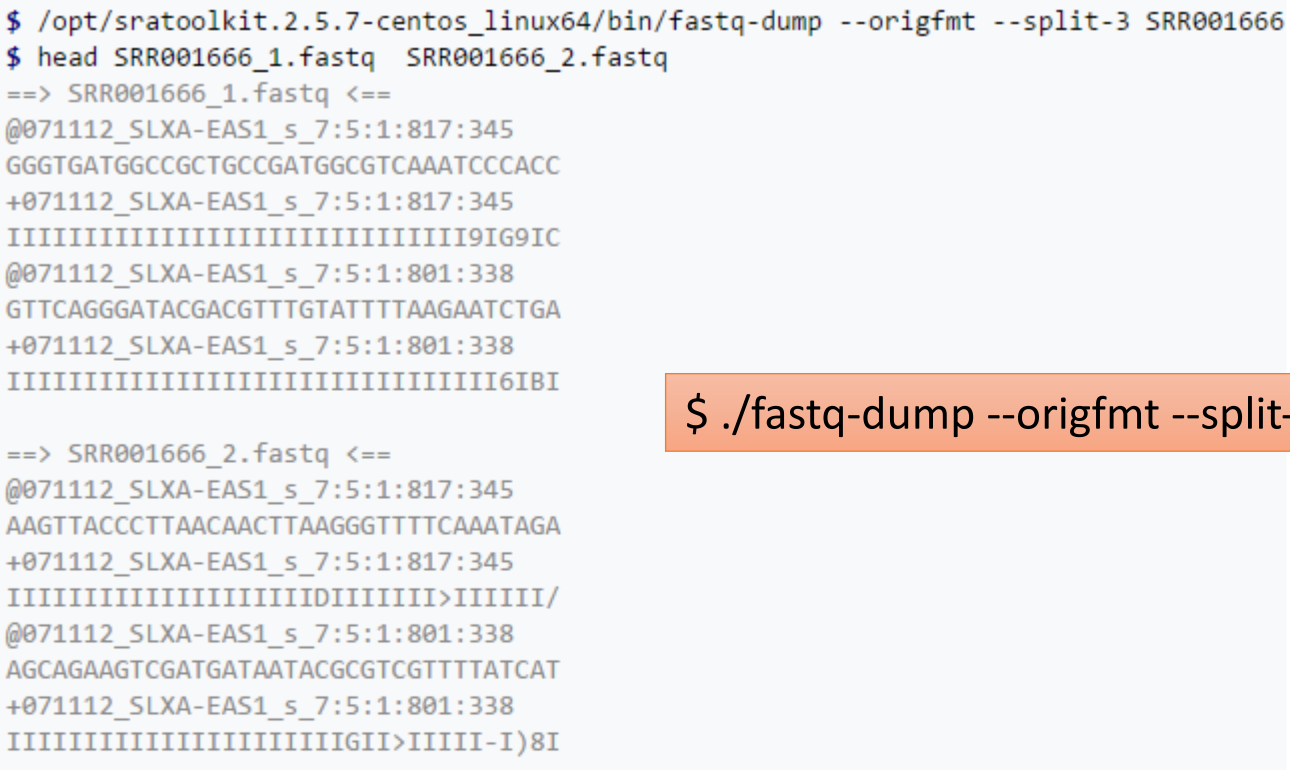
ID: 431

```
$ ./prefetch SRR001666
$ ./fastq-dump SRR001666
```

NCBI Sequence Read Archive (fastq)

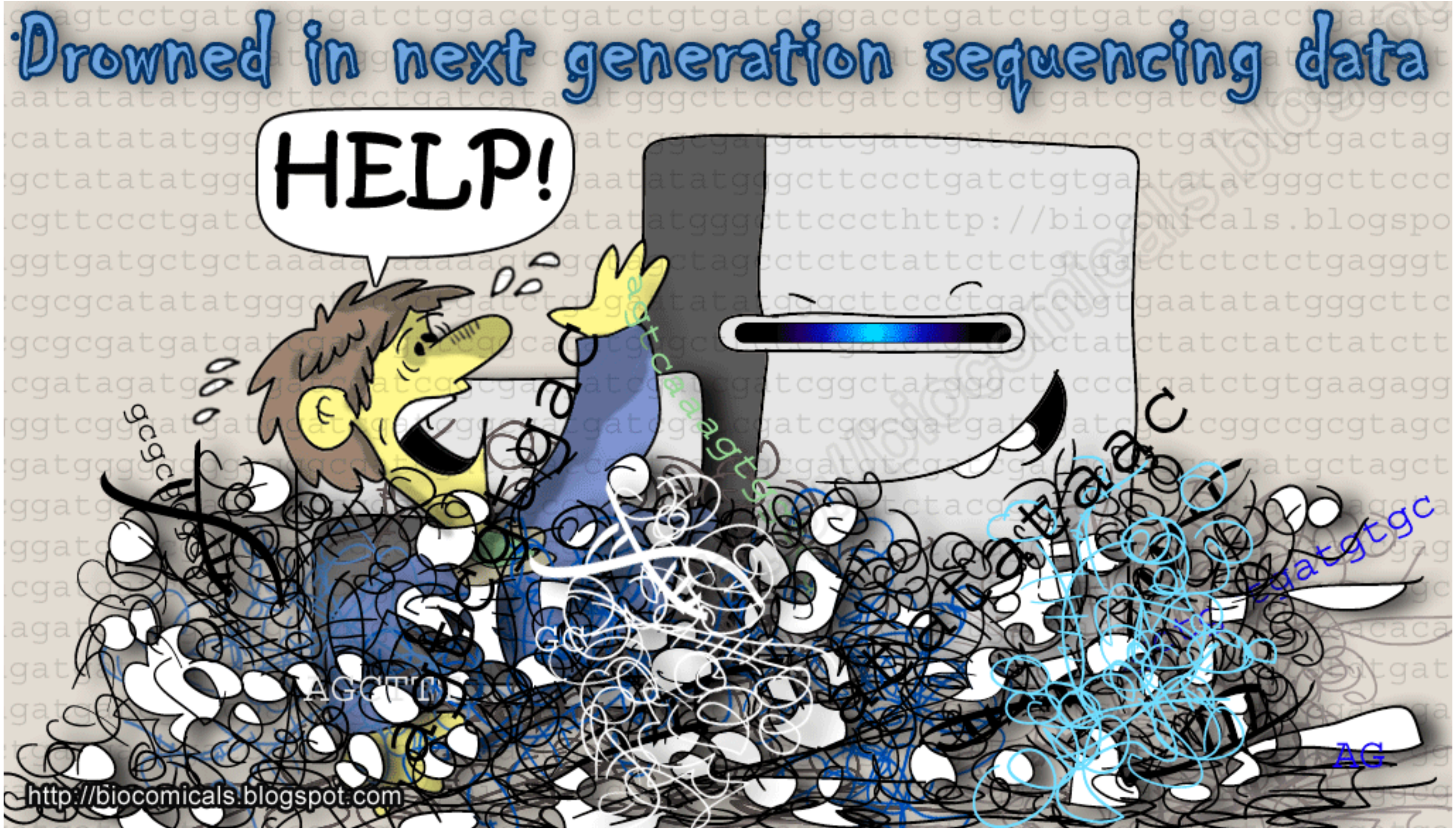
Further to note, with newer fastq-dump the extracted sequences have double-length and it turns out fastq-dump concatenates sequence of the forward and reverse reads together into a non-sense:

Better approach is to preserve original accessions and split into two or three files (forward, reverse, singletons)

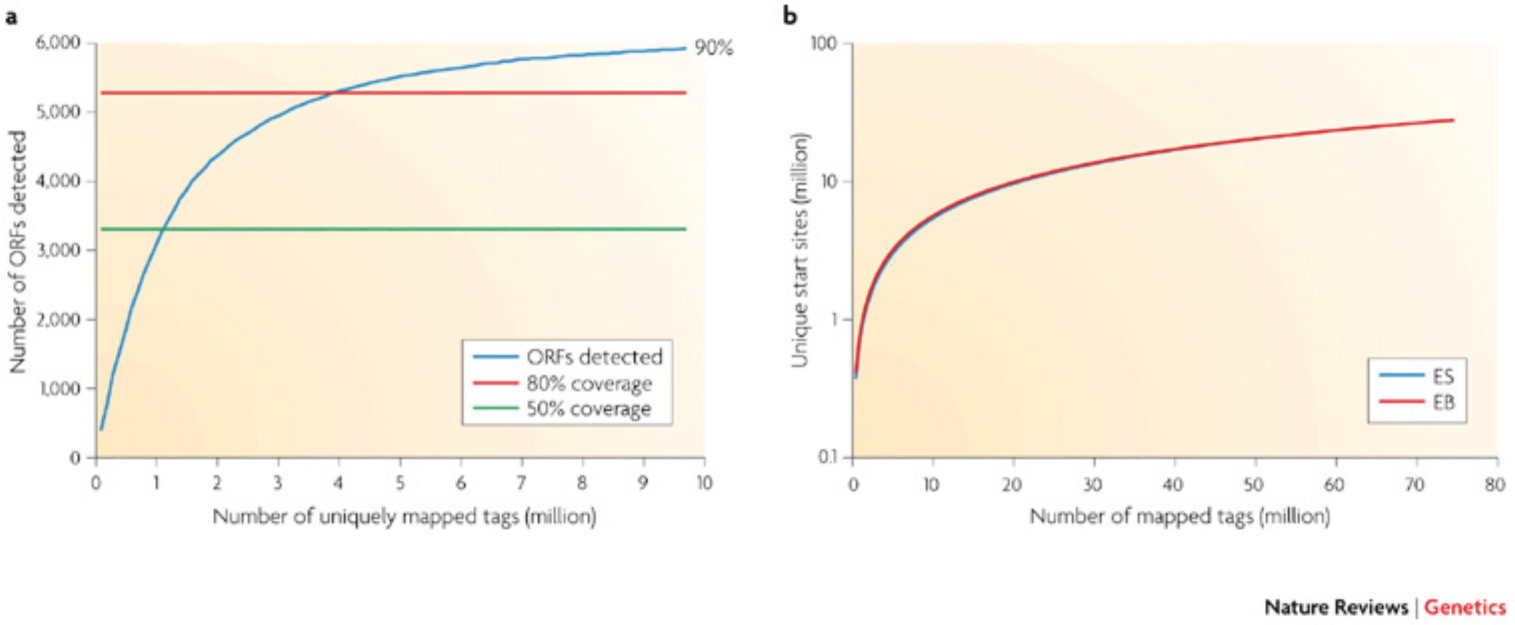


```
$ /opt/sratoolkit.2.5.7-centos_linux64/bin/fastq-dump --origfmt --split-3 SRR001666
$ head SRR001666_1.fastq  SRR001666_2.fastq
==> SRR001666_1.fastq <==
@071112_SLXA-EAS1_s_7:5:1:817:345
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+071112_SLXA-EAS1_s_7:5:1:817:345
IIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
@071112_SLXA-EAS1_s_7:5:1:801:338
GTTCAGGGATACGACGTTTGTATTTTAAGAATCTGA
+071112_SLXA-EAS1_s_7:5:1:801:338
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII6IBI

==> SRR001666_2.fastq <==
@071112_SLXA-EAS1_s_7:5:1:817:345
AAGTTACCCTTAACAACTTAAGGGTTTTCAAATAGA
+071112_SLXA-EAS1_s_7:5:1:817:345
IIIIIIIIIIIIIIIIIIIIDIIIIIII>IIIIII/
@071112_SLXA-EAS1_s_7:5:1:801:338
AGCAGAAGTCGATGATAATACGCGTCGTTTTATCAT
+071112_SLXA-EAS1_s_7:5:1:801:338
IIIIIIIIIIIIIIIIIIIIIIGII>IIIII-I)8I
```

```
$ ./fastq-dump --origfmt --split-3 SRR001666
```


Drowned in next generation sequencing data
HELP!
http://biocomicals.blogspot.com

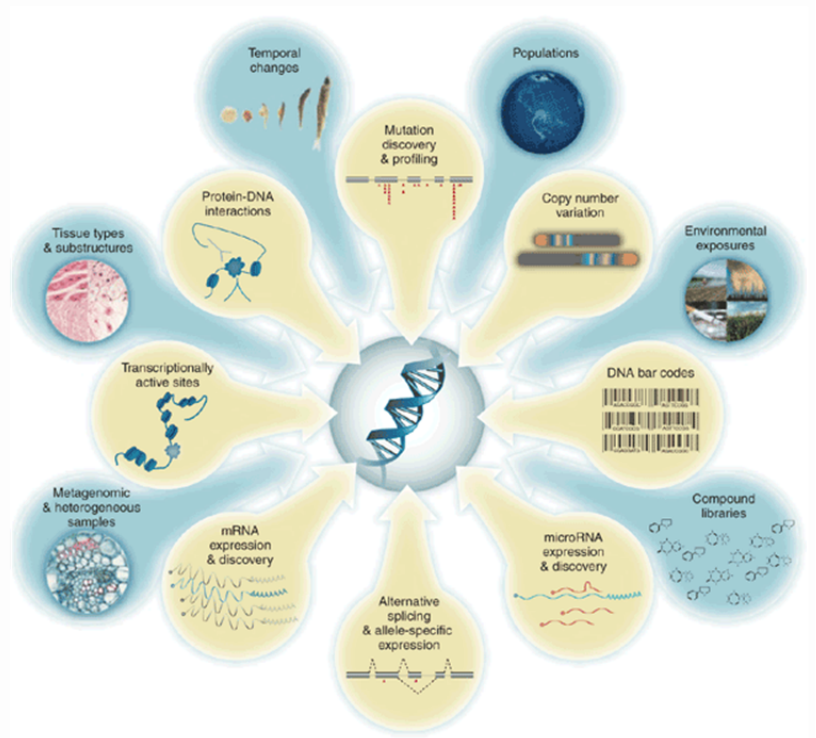
How deep should we go? coverage



- **a** | 80% of yeast genes (genome size ~120 Mb) were detected at 4 million uniquely mapped RNA-Seq reads, and coverage reaches a plateau afterwards despite the increasing sequencing depth. Expressed genes are defined as having at least four independent reads from a 50-bp window at the 3' end.
- **b** | The number of unique start sites detected starts to reach a plateau when the depth of sequencing reaches 80 million in two mouse transcriptomes. ES, embryonic stem cells; EB, embryonic body.

doi:10.1038/nrg2484

Applications on Biomedical Sciences


Temporal changes
Populations
Mutation discovery & profiling
Protein-DNA interactions
Copy number variation
Tissue types & substructures
Environmental exposures
Transcriptionally active sites
DNA bar codes
Metagenomic & heterogeneous samples
mRNA expression & discovery
Alternative splicing & allele-specific expression
microRNA expression & discovery
Compound libraries

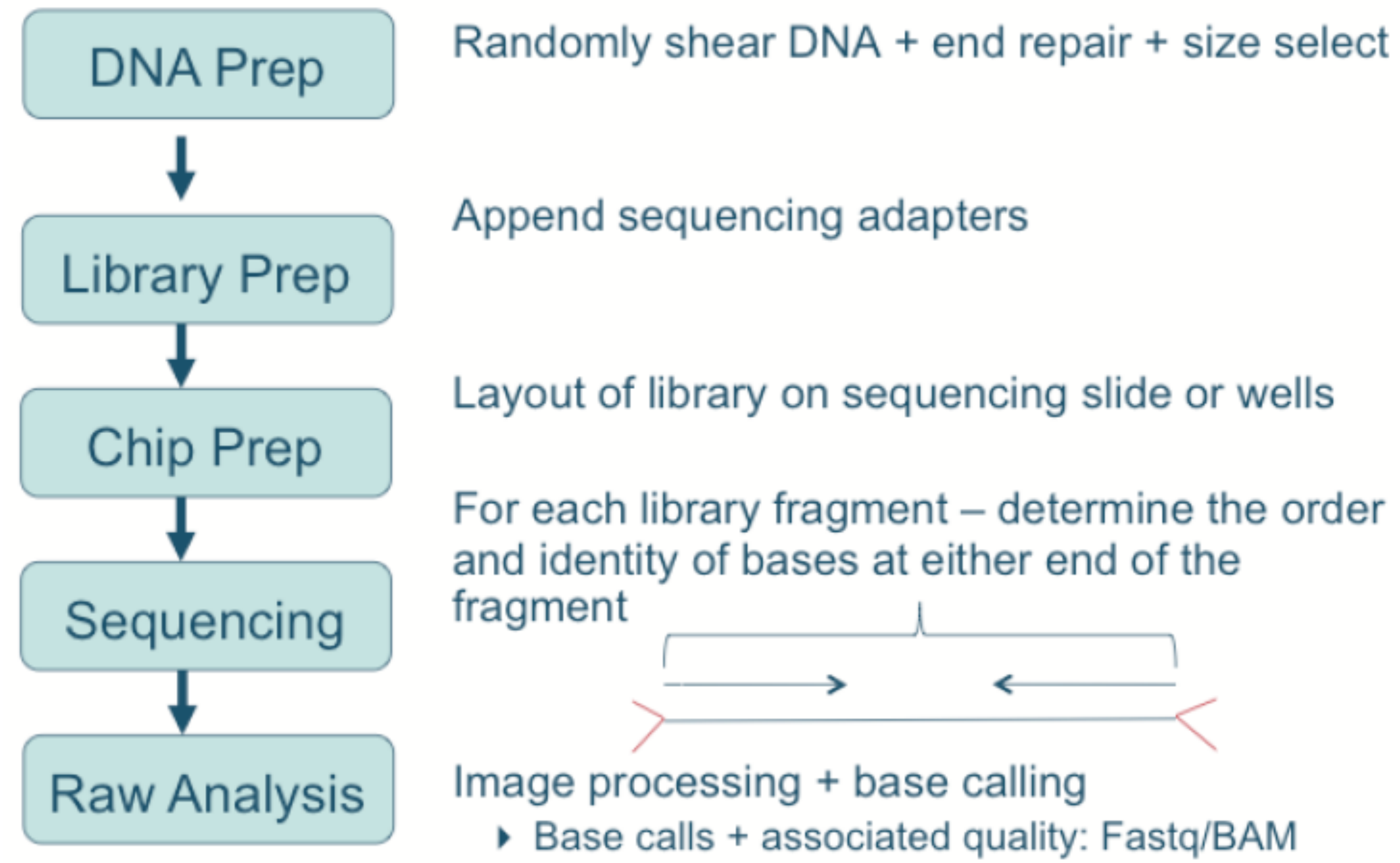
DNA

- Whole Genome Sequencing
- Exome Sequencing
- De novo Genome Sequencing
- Metagenome Sequencing
- ChIP Sequencing

RNA

- Small RNA Sequencing
- Transcriptome Sequencing
- De novo Transcriptome Sequencing
- Metatranscriptome Sequencing

HTseq Experiment



DNA Prep

Randomly shear DNA + end repair + size select

Library Prep

Append sequencing adapters

Chip Prep

Layout of library on sequencing slide or wells

Sequencing

For each library fragment – determine the order and identity of bases at either end of the fragment

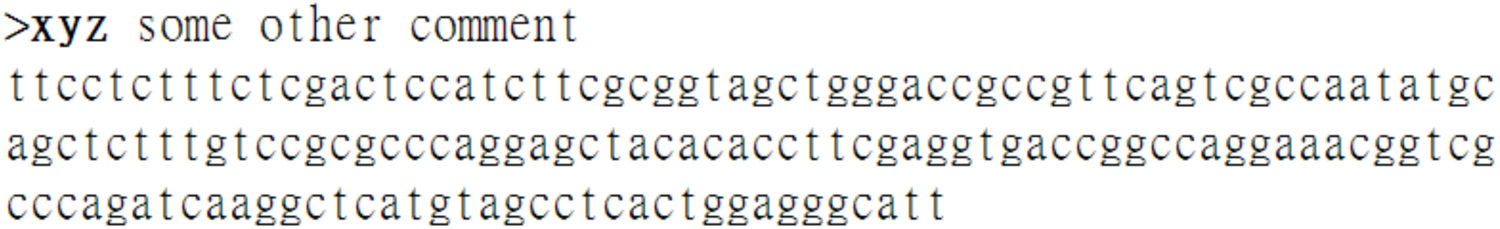
Raw Analysis

Image processing + base calling

- Base calls + associated quality: Fastq/BAM

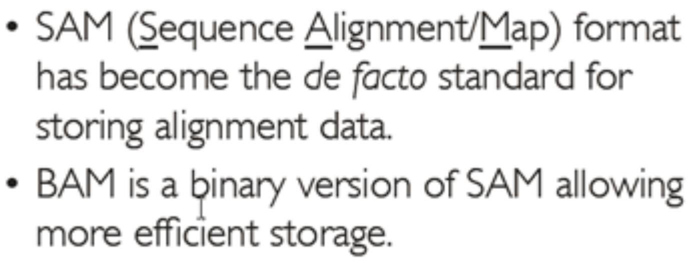
Data Format Types

- Raw Sequence Data e.g. fasta/fastq

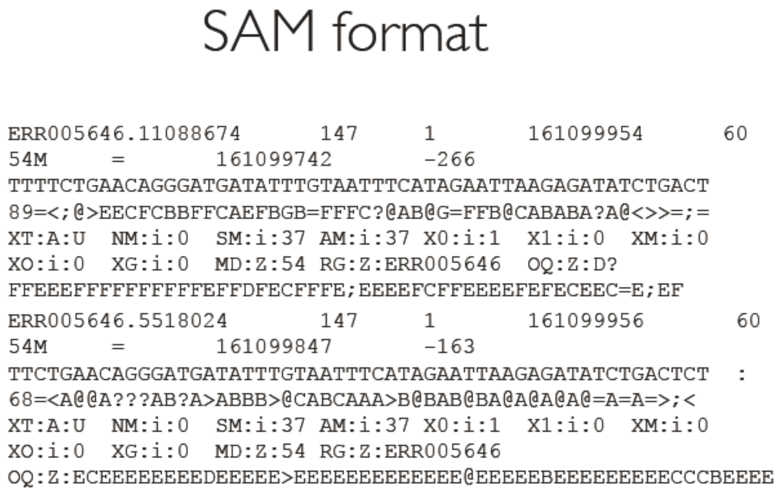


```
>xyz some other comment
ttcctctttctcgactccatcttcgcggtagctgggaccgccgttcagtcgccaatatgc
agctctttgtccgcgcccaggagctacacaccttcgaggtgaccggccaggaaacggtcg
cccagatcaaggctcatgtagcctcactggagggcatt
```

- Aligned data e.g. SAM/BAM



- SAM (Sequence Alignment/Map) format has become the *de facto* standard for storing alignment data.
- BAM is a binary version of SAM allowing more efficient storage.



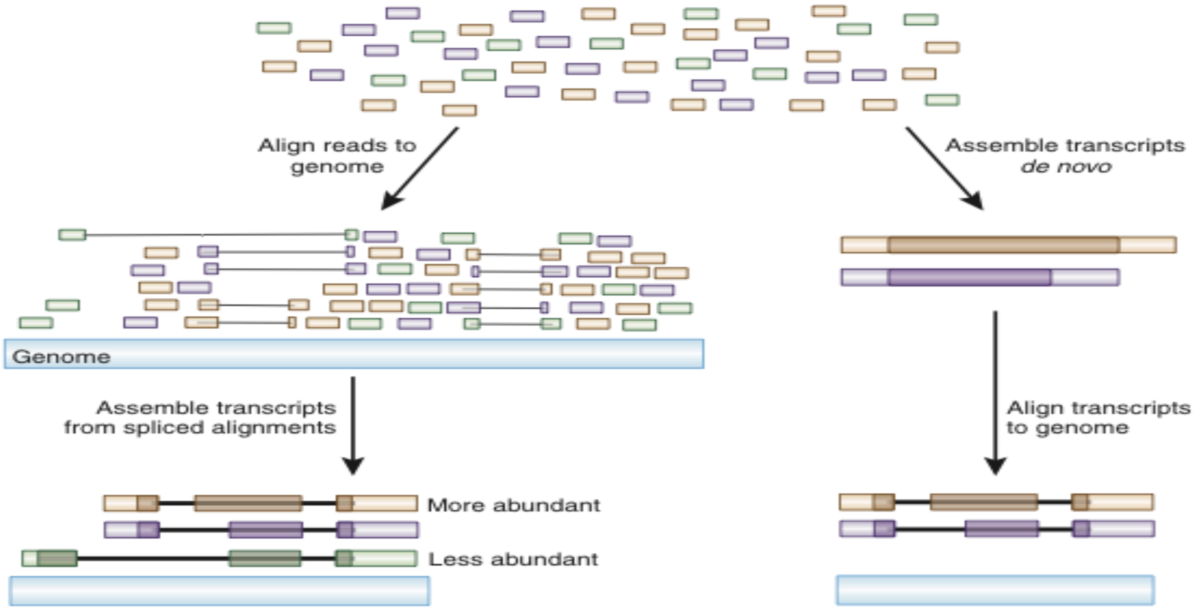
SAM format

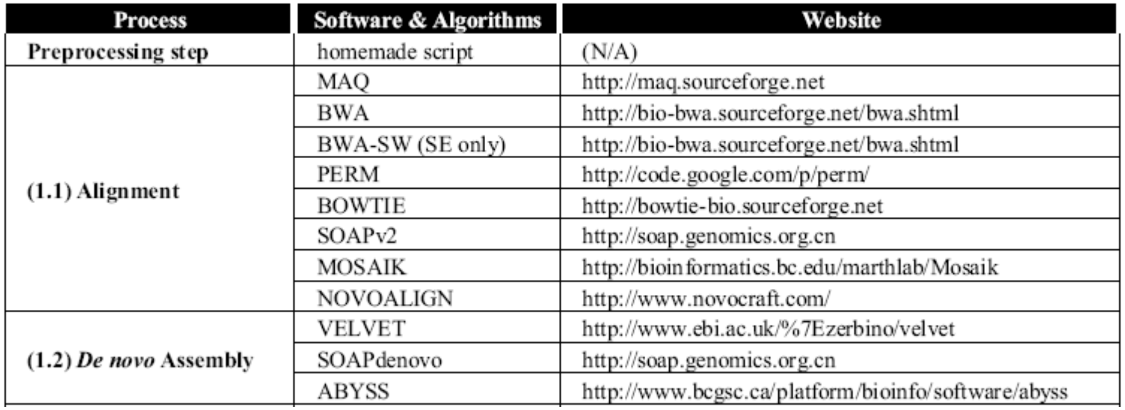
```
ERR005646.11088674      147     1       161099954       60
54M     =       161099742       -266
TTTTCTGAACAGGGATGATATTTGTAATTTCATAGAATTAAGAGATATCTGACT
89=<;@>EECFCBBFFCAEFBGB=FFFC?@AB@G=FFB@CABABA?A@<>>=;=
XT:A:U  NM:i:0  SM:i:37 AM:i:37 X0:i:1  X1:i:0  XM:i:0
XO:i:0  XG:i:0  MD:Z:54 RG:Z:ERR005646  OQ:Z:D?
FFEEEFFFFFFFFFFEFFDFECFFFE;EEEEFCFFEEEEFEFECEEC=E;EF
ERR005646.5518024       147     1       161099956       60
54M     =       161099847       -163
TTCTGAACAGGGATGATATTTGTAATTTCATAGAATTAAGAGATATCTGACTCT  :
68=<A@@A???AB?A>ABBB>@CABCAAA>B@BAB@BA@A@A@A@=A=A=>;<
XT:A:U  NM:i:0  SM:i:37 AM:i:37 X0:i:1  X1:i:0  XM:i:0
XO:i:0  XG:i:0  MD:Z:54 RG:Z:ERR005646
OQ:Z:ECEEEEEEEEDEEEEE>EEEEEEEEEEEEE@EEEEEBEEEEEEEEEECCCBEEEE
```

- Processed data e.g. BED

Analysis Strategies:

Reference Sequence Alignment (**Mapping**) vs ***de novo* Assembly**

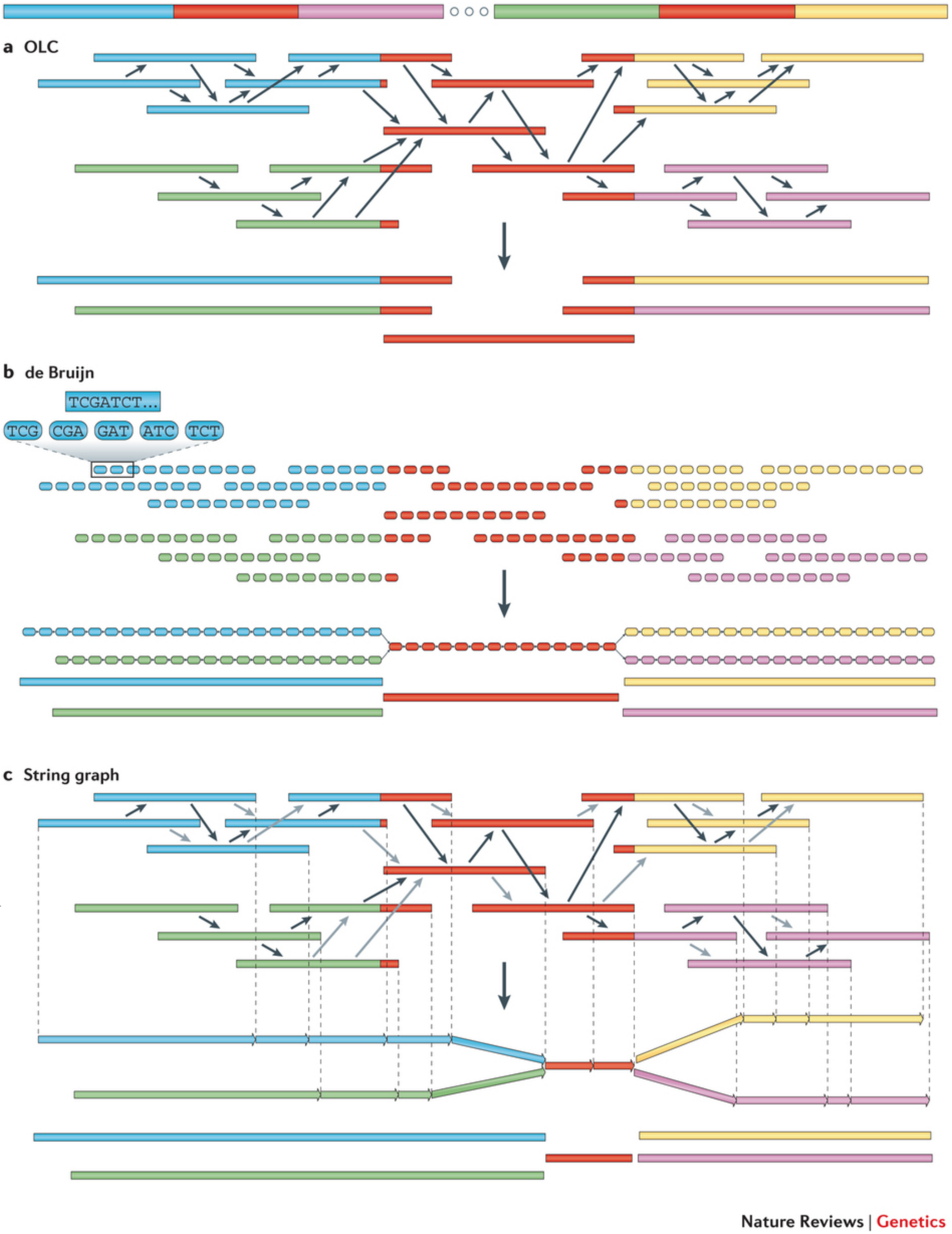

Align reads to genome
Assemble transcripts *de novo*
Genome
Assemble transcripts from spliced alignments
More abundant
Less abundant
Align transcripts to genome



Process	Software & Algorithms	Website
Preprocessing step	homemade script	(N/A)
(1.1) Alignment	MAQ	http://maq.sourceforge.net
	BWA	http://bio-bwa.sourceforge.net/bwa.shtml
	BWA-SW (SE only)	http://bio-bwa.sourceforge.net/bwa.shtml
	PERM	http://code.google.com/p/perm/
	BOWTIE	http://bowtie-bio.sourceforge.net
	SOAPv2	http://soap.genomics.org.cn
	MOSAIK	http://bioinformatics.bc.edu/marthlab/Mosaik
	NOVOALIGN	http://www.novocraft.com/
(1.2) *De novo* Assembly	VELVET	http://www.ebi.ac.uk/%7Ezerbino/velvet
	SOAPdenovo	http://soap.genomics.org.cn
	ABYSS	http://www.bcgsc.ca/platform/bioinfo/software/abyss

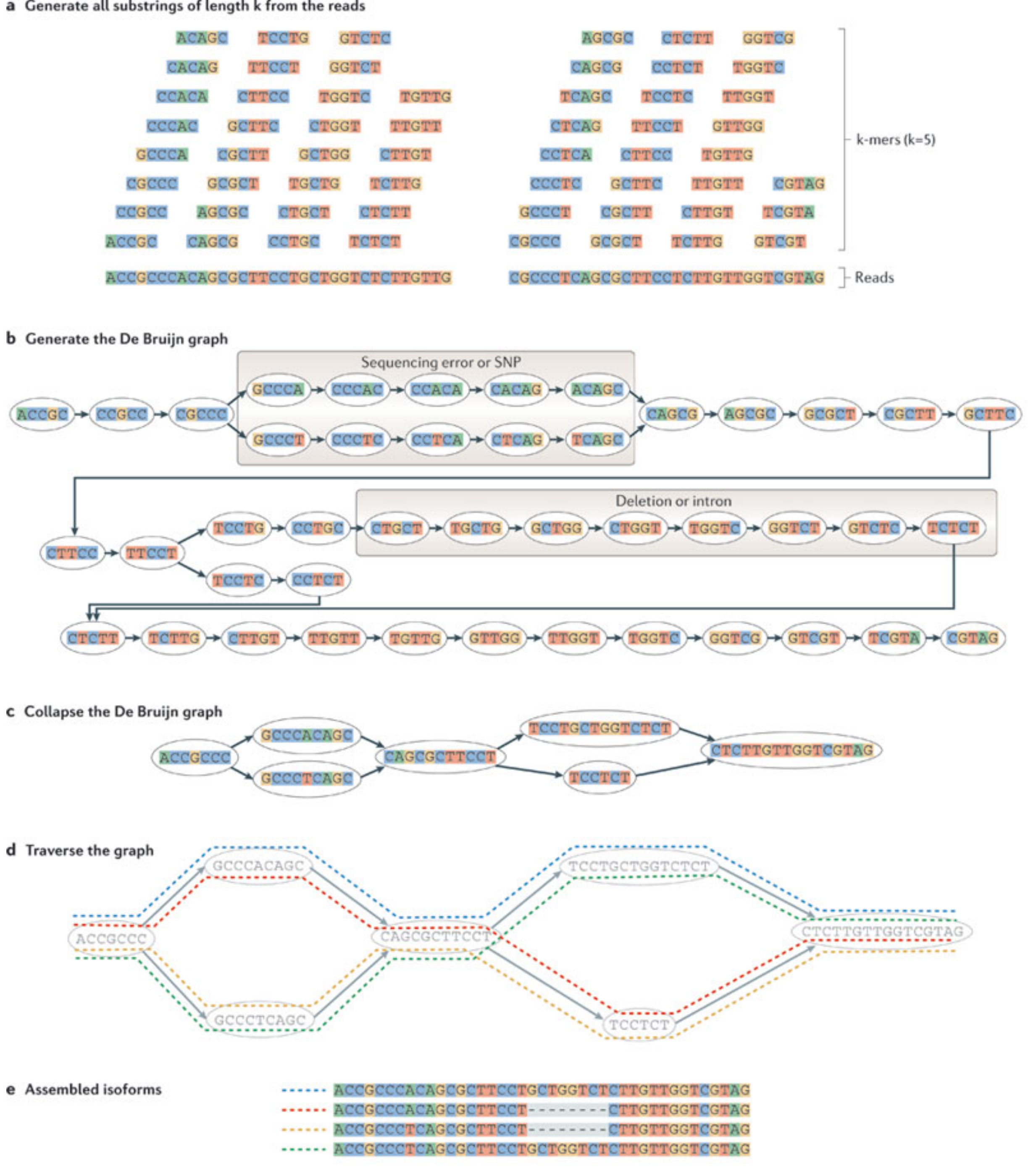
de novo Assembly

- Genomics assembly:
    - Velvet,
    - SOAPdenovo
- Transcript assembly:
    - Trinity



http://player.slideplayer.com/27/9065734

K-mers

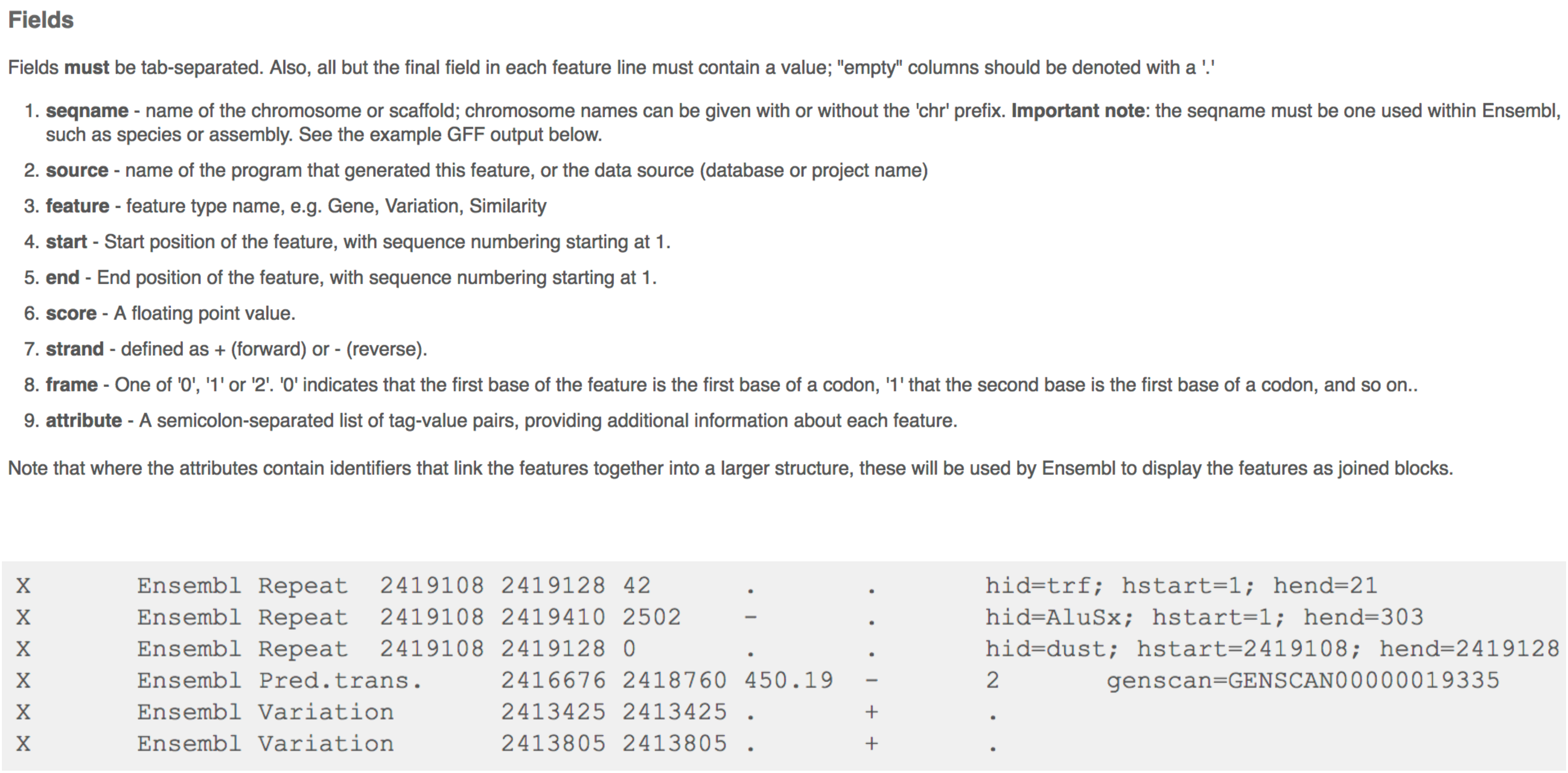


Nature Reviews | Genetics

Reference Genome Example:

assembly	Genome	years
GRCh38/hg38	Human	Dec. 2013
GRCh37/hg19	Human	Feb. 2009
NCBI36/hg18	Human	Mar. 2006
NCBI35/hg17	Human	May 2004
NCBI34/hg16	Human	July 2003
GRCm38/mm10	Mouse	Dec. 2011
NCBI37/mm9	Mouse	July 2007
NCBI36/mm8	Mouse	Feb. 2006
NCBI35/mm7	Mouse	Aug. 2005
RGSC 6.0/rn6	Rat	Jul. 2014
RGSC 5.0/rn5	Rat	Mar. 2012
Baylor 3.4/rn4	Rat	Nov. 2004
BDGP R6+ISO1 MT/dm6	D. melanogaster	Aug. 2014
BDGP R5/dm3	D. melanogaster	Apr. 2006

GFF/GTF File Format



Fields

Fields **must** be tab-separated. Also, all but the final field in each feature line must contain a value; "empty" columns should be denoted with a '.'

1. **seqname** - name of the chromosome or scaffold; chromosome names can be given with or without the 'chr' prefix. **Important note**: the seqname must be one used within Ensembl, such as species or assembly. See the example GFF output below.
2. **source** - name of the program that generated this feature, or the data source (database or project name)
3. **feature** - feature type name, e.g. Gene, Variation, Similarity
4. **start** - Start position of the feature, with sequence numbering starting at 1.
5. **end** - End position of the feature, with sequence numbering starting at 1.
6. **score** - A floating point value.
7. **strand** - defined as + (forward) or - (reverse).
8. **frame** - One of '0', '1' or '2'. '0' indicates that the first base of the feature is the first base of a codon, '1' that the second base is the first base of a codon, and so on..
9. **attribute** - A semicolon-separated list of tag-value pairs, providing additional information about each feature.

Note that where the attributes contain identifiers that link the features together into a larger structure, these will be used by Ensembl to display the features as joined blocks.

```
X       Ensembl Repeat  2419108 2419128 42      .       .       hid=trf; hstart=1; hend=21
X       Ensembl Repeat  2419108 2419410 2502    -       .       hid=AluSx; hstart=1; hend=303
X       Ensembl Repeat  2419108 2419128 0       .       .       hid=dust; hstart=2419108; hend=2419128
X       Ensembl Pred.trans.     2416676 2418760 450.19  -       2       genscan=GENSCAN00000019335
X       Ensembl Variation       2413425 2413425 .       +       .
X       Ensembl Variation       2413805 2413805 .       +       .
```

UCSC Table Browser

Our tools

- **Genome Browser**
  interactively visualize genomic data
- **BLAT**
  rapidly align sequences to the genome
- **Table Browser**
  download data from the Genome Browser database

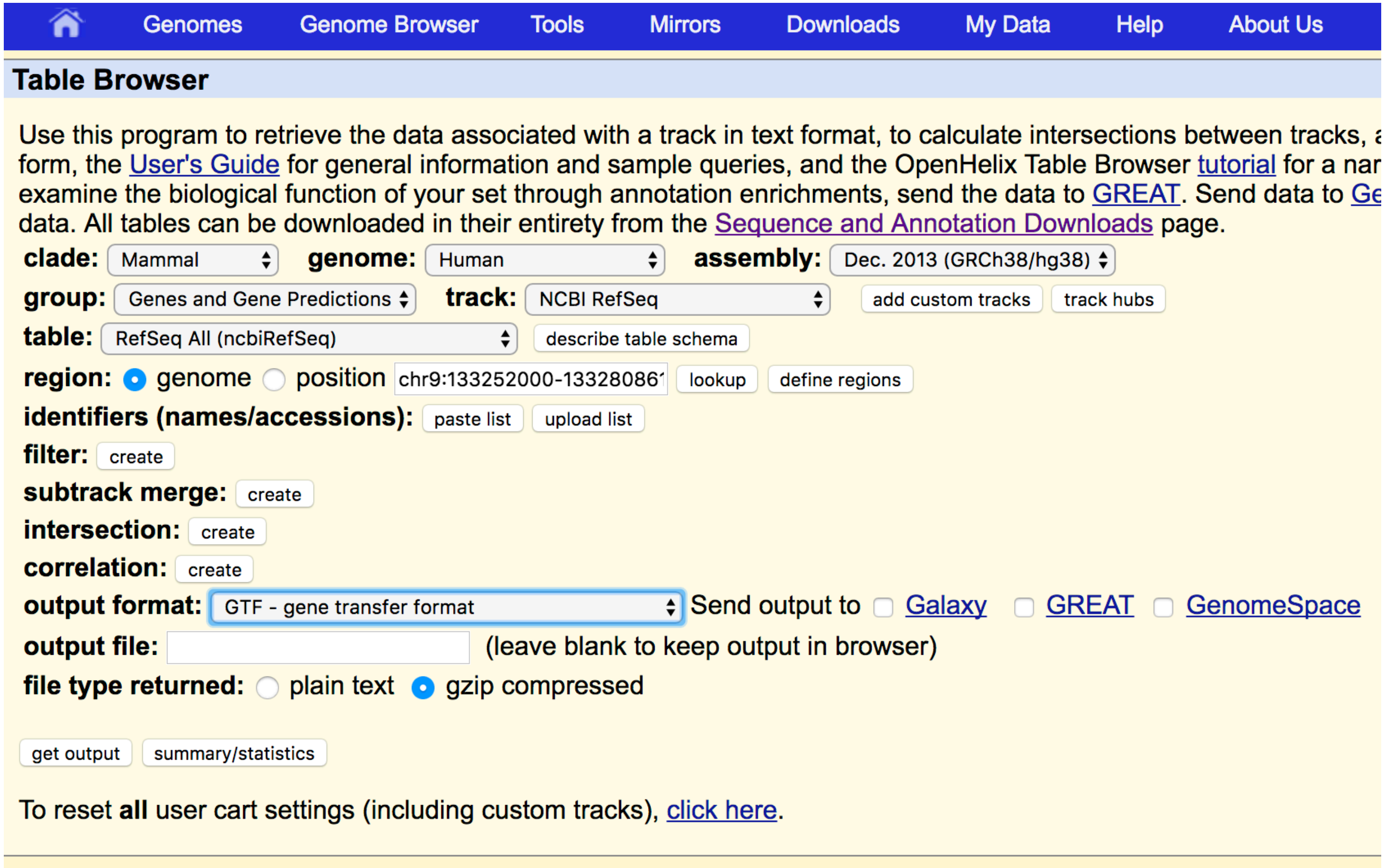


Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, a
form, the User's Guide for general information and sample queries, and the OpenHelix Table Browser tutorial for a nar
examine the biological function of your set through annotation enrichments, send the data to GREAT. Send data to Ge
data. All tables can be downloaded in their entirety from the Sequence and Annotation Downloads page.

clade: Mammal **genome:** Human **assembly:** Dec. 2013 (GRCh38/hg38)

group: Genes and Gene Predictions **track:** NCBI RefSeq add custom tracks track hubs

table: RefSeq All (ncbiRefSeq) describe table schema

region: ● genome ○ position chr9:133252000-13328086' lookup define regions

identifiers (names/accessions): paste list upload list

filter: create

subtrack merge: create

intersection: create

correlation: create

output format: GTF - gene transfer format Send output to □ Galaxy □ GREAT □ GenomeSpace

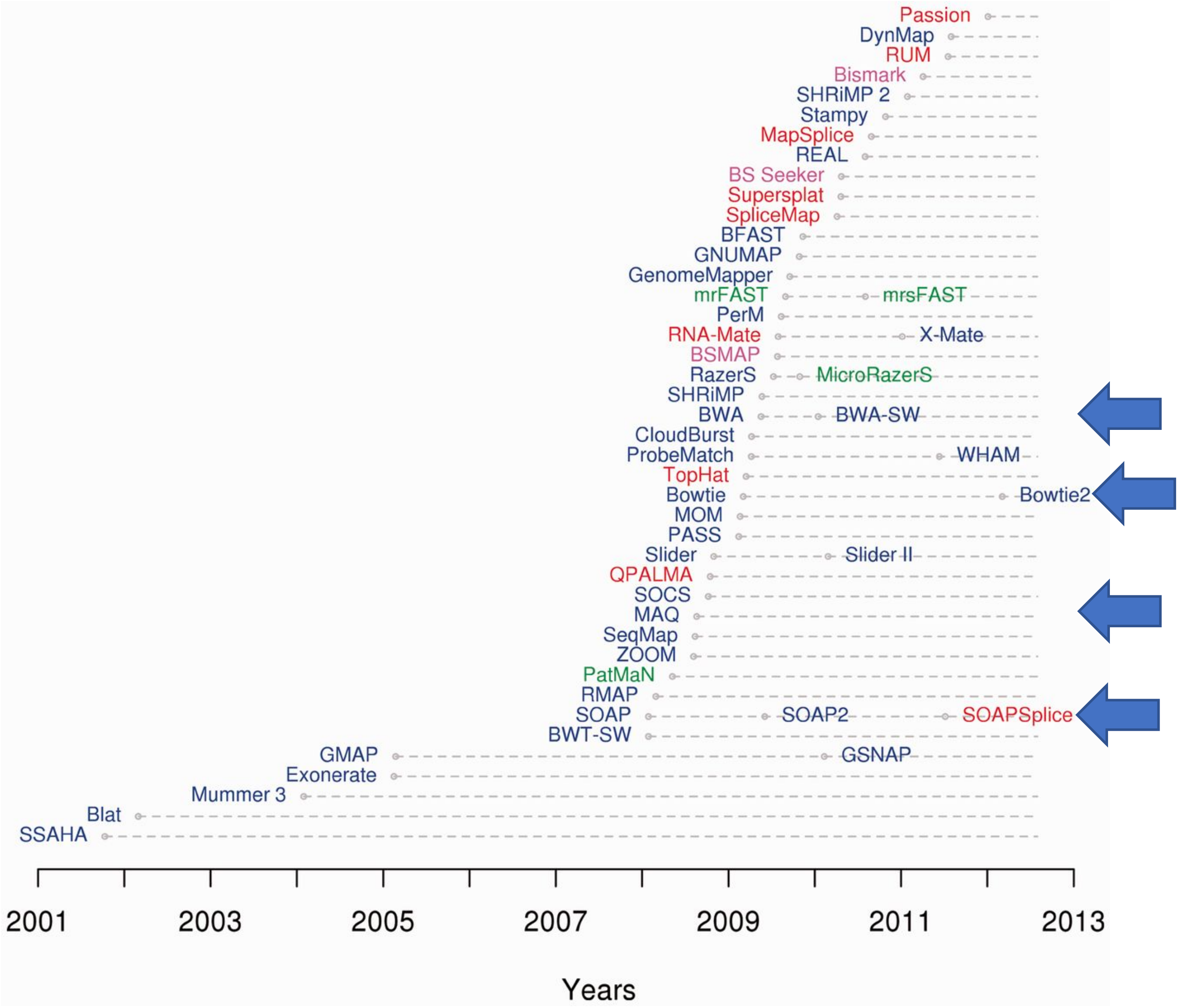
output file: (leave blank to keep output in browser)

file type returned: ○ plain text ● gzip compressed

get output summary/statistics

To reset **all** user cart settings (including custom tracks), click here.

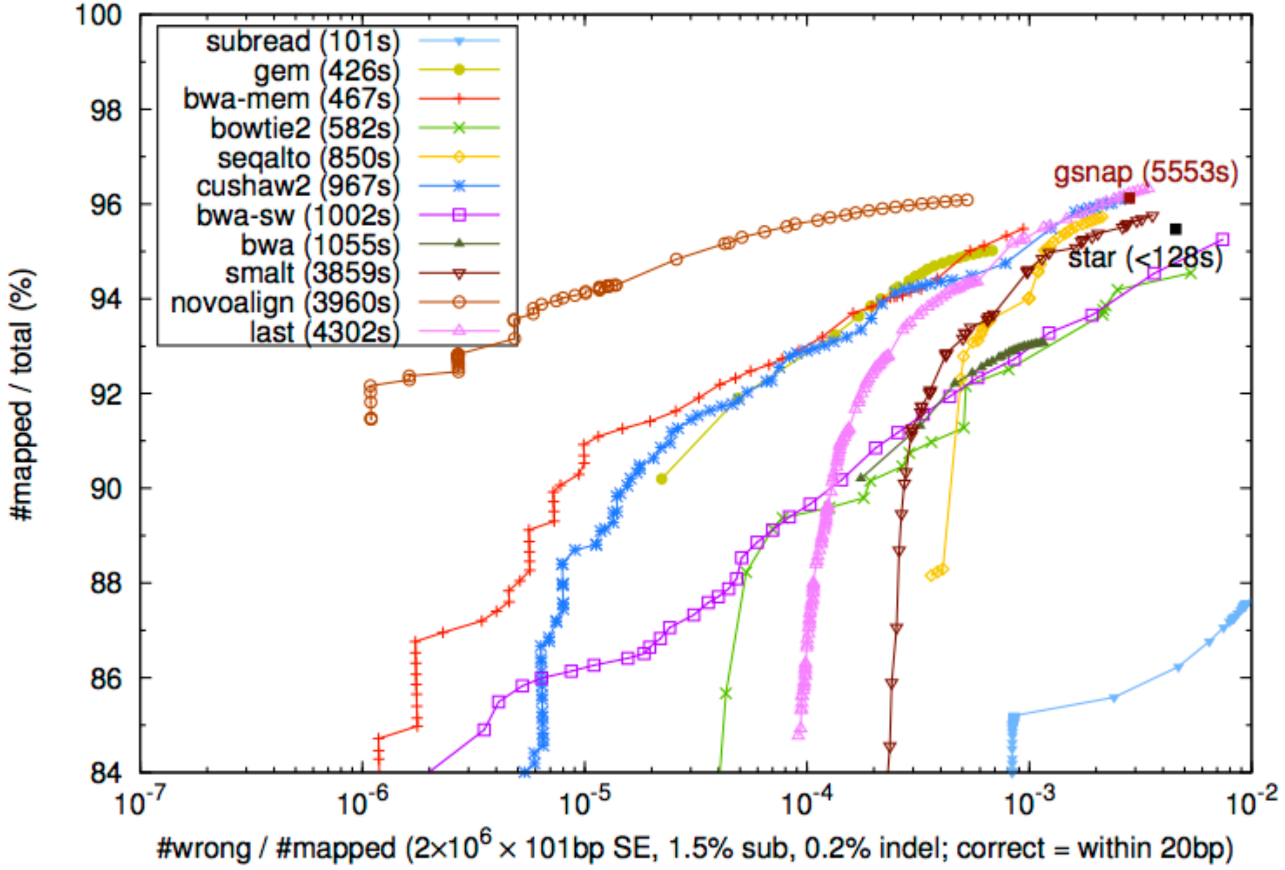
Reference Sequence Alignment (Mapping)



DNA mappers are plotted in blue, RNA mappers in red, miRNA mappers in green and bisulphite mappers in purple.

PMID: 23060614

Comparison of Mapping tools (ROC curve)

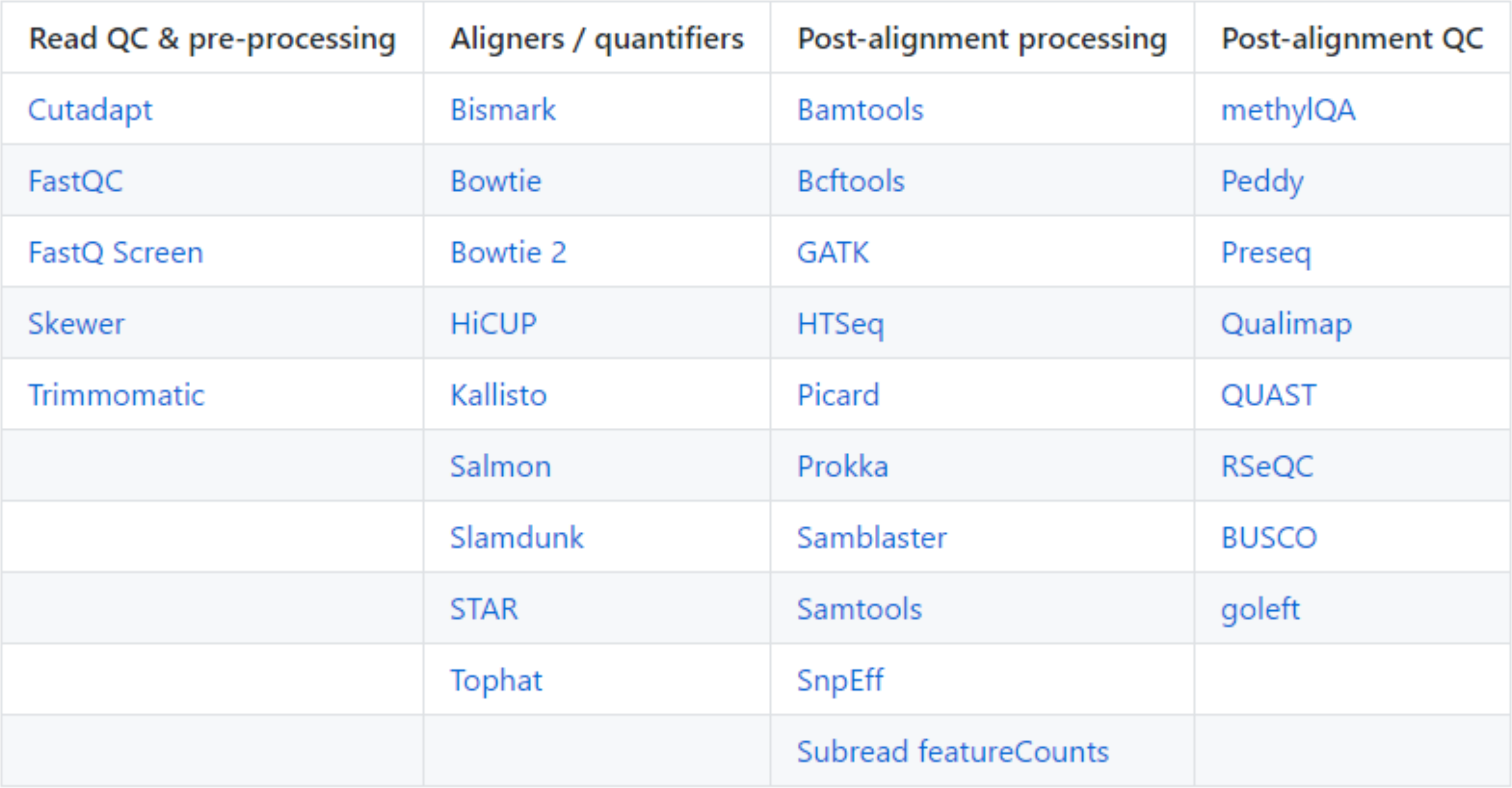


- ChIP, RNA-seq ➔ bowtie2 → cufflinks
- SNP, Indels, methylation ➔ BWA → GATK

http://seqanswers.com/forums/showthread.php?t=15200&page=5

MultiQC

Aggregate bioinformatics results across many samples into a single report.



Read QC & pre-processing	Aligners / quantifiers	Post-alignment processing	Post-alignment QC
Cutadapt	Bismark	Bamtools	methylQA
FastQC	Bowtie	Bcftools	Peddy
FastQ Screen	Bowtie 2	GATK	Preseq
Skewer	HiCUP	HTSeq	Qualimap
Trimmomatic	Kallisto	Picard	QUAST
	Salmon	Prokka	RSeQC
	Slamdunk	Samblaster	BUSCO
	STAR	Samtools	goleft
	Tophat	SnpEff	
		Subread featureCounts	

MultiQC: Summarize analysis results for multiple tools and samples in a single report
Philip Ewels, Måns Magnusson, Sverker Lundin and Max Käller
Bioinformatics (2016)
doi: 10.1093/bioinformatics/btw354
PMID: 27312411

SciLifeLab

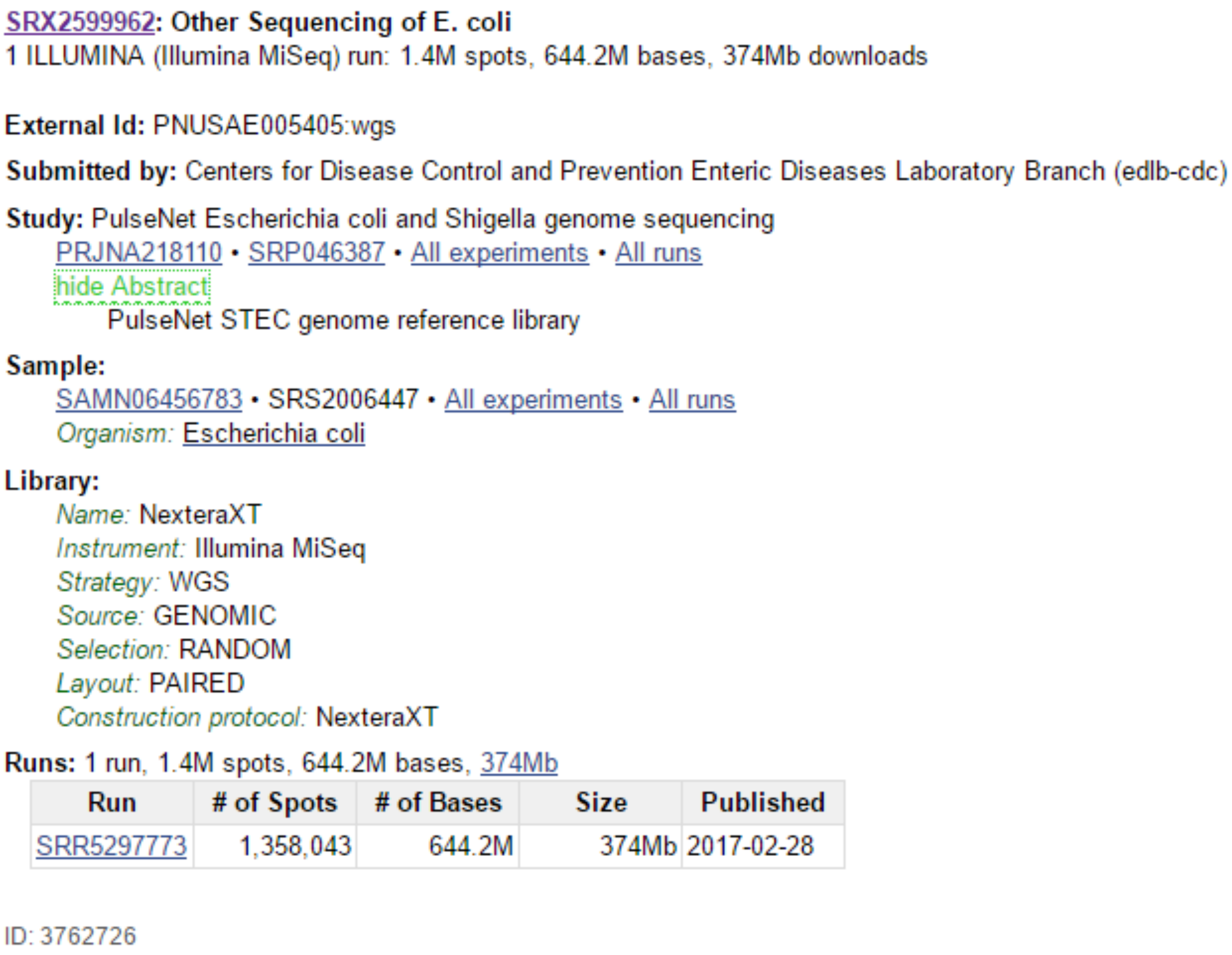
RNA-Seq

- This report was generated using logs from an analysis accidentally run on ChIP-Seq data from the *BI Human Reference Epigenome Mapping Project: ChIP-Seq in human subject* dataset (SRP001534).
- Initial QC was done using FastQC, followed by trimming with TrimGalore! (a wrapper around cutadapt). Reads were aligned using STAR and overlaps counted with featureCounts.

Whole-Genome Sequencing

- The data from this report comes from an analysis of HapMap trio samples, run by the National Genomics Infrastructre(NGI) at SciLifeLab, Sweden. Initial quality control was done using FastQC and FastQ Screen. Reads were processed with GATK and the aligned reads analysed using Picard. Downstream QC was done using Qualimap BamQC andSnpEff.

SRA & FastQC Exercise



SRX2599962: Other Sequencing of E. coli
1 ILLUMINA (Illumina MiSeq) run: 1.4M spots, 644.2M bases, 374Mb downloads

External Id: PNUSAE005405:wgs

Submitted by: Centers for Disease Control and Prevention Enteric Diseases Laboratory Branch (edlb-cdc)

Study: PulseNet Escherichia coli and Shigella genome sequencing
PRJNA218110 • SRP046387 • All experiments • All runs
hide Abstract
PulseNet STEC genome reference library

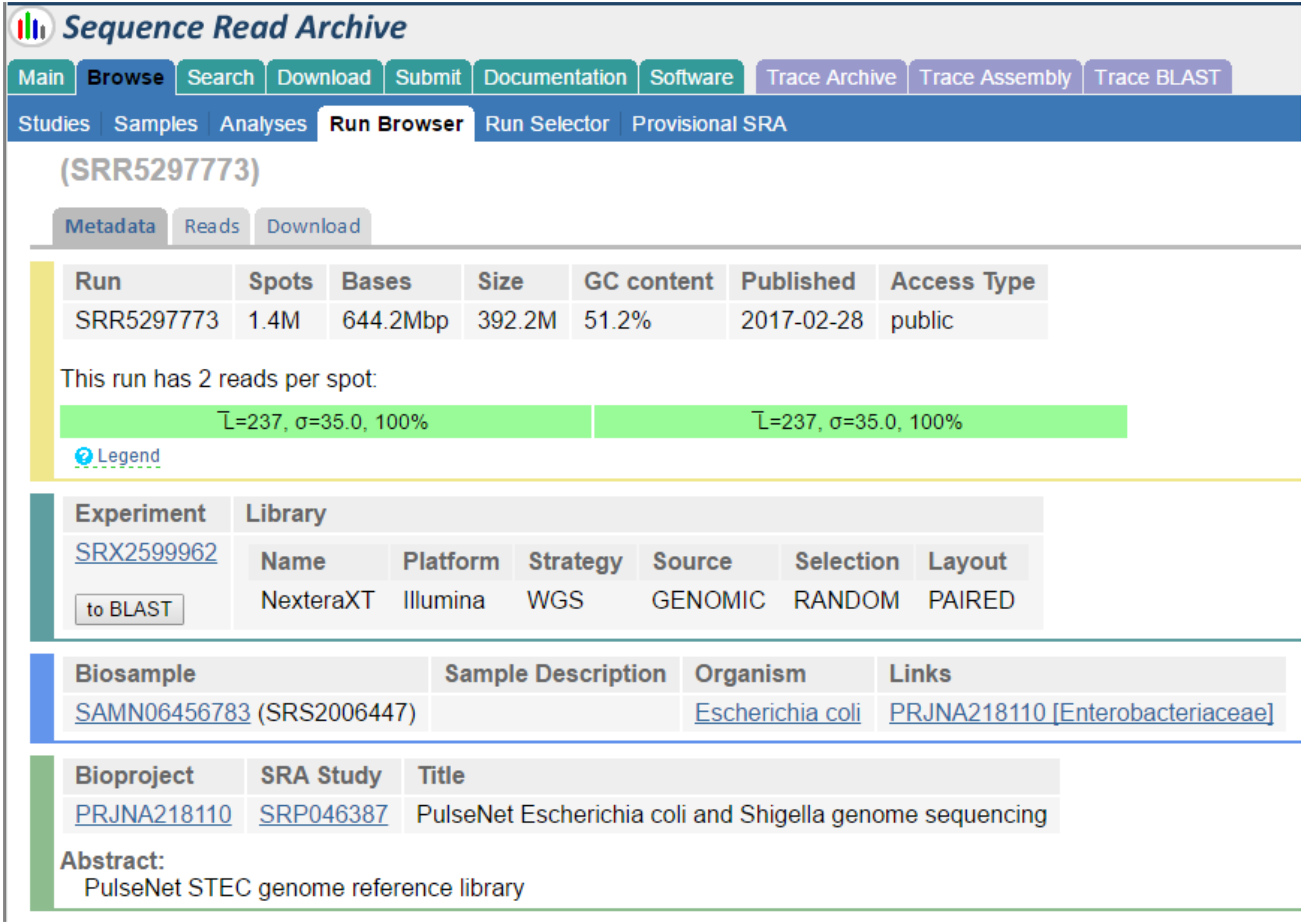
Sample:
SAMN06456783 • SRS2006447 • All experiments • All runs
Organism: Escherichia coli

Library:
Name: NexteraXT
Instrument: Illumina MiSeq
Strategy: WGS
Source: GENOMIC
Selection: RANDOM
Layout: PAIRED
Construction protocol: NexteraXT

Runs: 1 run, 1.4M spots, 644.2M bases, 374Mb

Run	# of Spots	# of Bases	Size	Published
SRR5297773	1,358,043	644.2M	374Mb	2017-02-28

ID: 3762726



(SRR5297773)

Run	Spots	Bases	Size	GC content	Published	Access Type
SRR5297773	1.4M	644.2Mbp	392.2M	51.2%	2017-02-28	public

This run has 2 reads per spot:

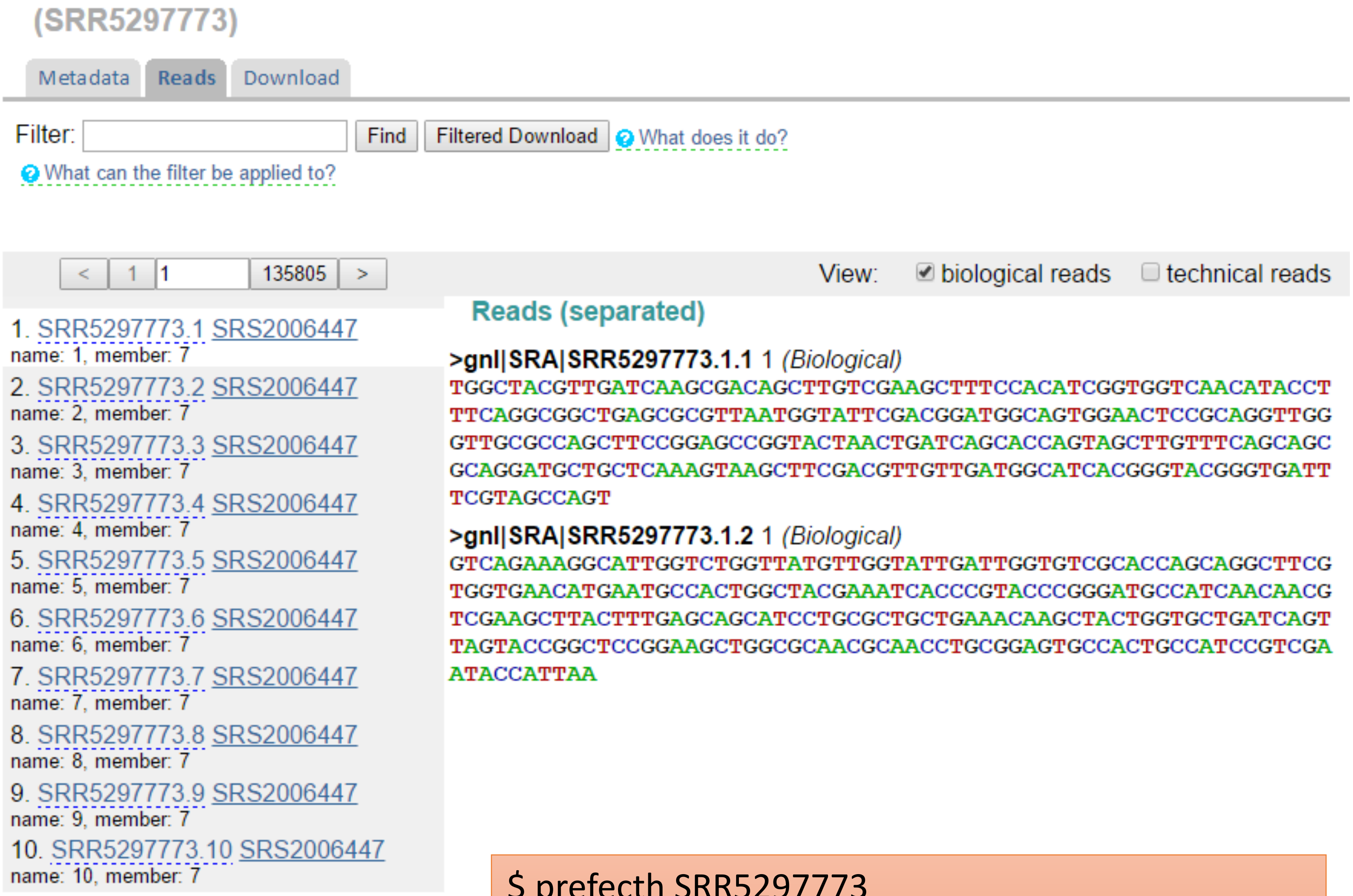
L̄=237, σ=35.0, 100%	L̄=237, σ=35.0, 100%

Experiment	Library					
SRX2599962	Name	Platform	Strategy	Source	Selection	Layout
	NexteraXT	Illumina	WGS	GENOMIC	RANDOM	PAIRED

Biosample	Sample Description	Organism	Links
SAMN06456783 (SRS2006447)		Escherichia coli	PRJNA218110 [Enterobacteriaceae]

Bioproject	SRA Study	Title
PRJNA218110	SRP046387	PulseNet Escherichia coli and Shigella genome sequencing

Abstract:
PulseNet STEC genome reference library



(SRR5297773)

1. SRR5297773.1 SRS2006447
name: 1, member: 7
2. SRR5297773.2 SRS2006447
name: 2, member: 7
3. SRR5297773.3 SRS2006447
name: 3, member: 7
4. SRR5297773.4 SRS2006447
name: 4, member: 7
5. SRR5297773.5 SRS2006447
name: 5, member: 7
6. SRR5297773.6 SRS2006447
name: 6, member: 7
7. SRR5297773.7 SRS2006447
name: 7, member: 7
8. SRR5297773.8 SRS2006447
name: 8, member: 7
9. SRR5297773.9 SRS2006447
name: 9, member: 7
10. SRR5297773.10 SRS2006447
name: 10, member: 7

Reads (separated)

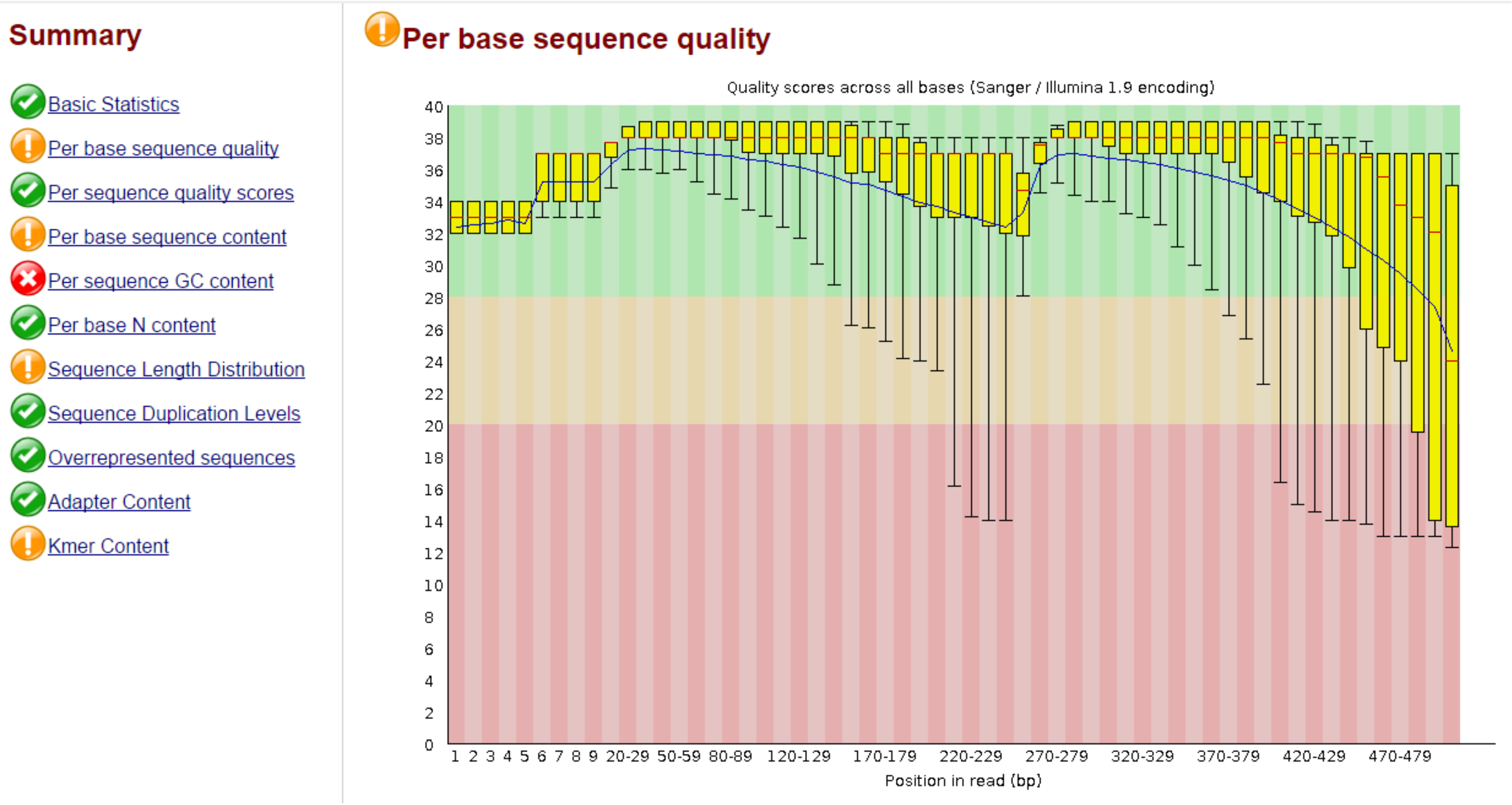
```
>gnl|SRA|SRR5297773.1.1 1 (Biological)
TGGCTACGTTGATCAAGCGACAGCTTGTCGAAGCTTTCCACATCGGTGGTCAACATACCT
TTCAGGCGGCTGAGCGCGTTAATGGTATTCGACGGATGGCAGTGGAACTCCGCAGGTTGG
GTTGCGCCAGCTTCCGGAGCCGGTACTAACTGATCAGCACCAGTAGCTTGTTTCAGCAGC
GCAGGATGCTGCTCAAAGTAAGCTTCGACGTTGTTGATGGCATCACGGGTACGGGTGATT
TCGTAGCCAGT
>gnl|SRA|SRR5297773.1.2 1 (Biological)
GTCAGAAAGGCATTGGTCTGGTTATGTTGGTATTGATTGGTGTCGCACCAGCAGGCTTCG
TGGTGAACATGAATGCCACTGGCTACGAAATCACCCGTACCCGGGATGCCATCAACAACG
TCGAAGCTTACTTTGAGCAGCATCCTGCGCTGCTGAAACAAGCTACTGGTGCTGATCAGT
TAGTACCGGCTCCGGAAGCTGGCGCAACGCAACCTGCGGAGTGCCACTGCCATCCGTCGA
ATACCATTAA
```

```
$ prefecth SRR5297773
$ fastq-dump SRR5297773
$ fastq-dump --split-files SRR5297773
```

- Install
  - “Putty” http://www.putty.org/
  - “filezilla” https://filezilla-project.org/

IP:120.126.1.41

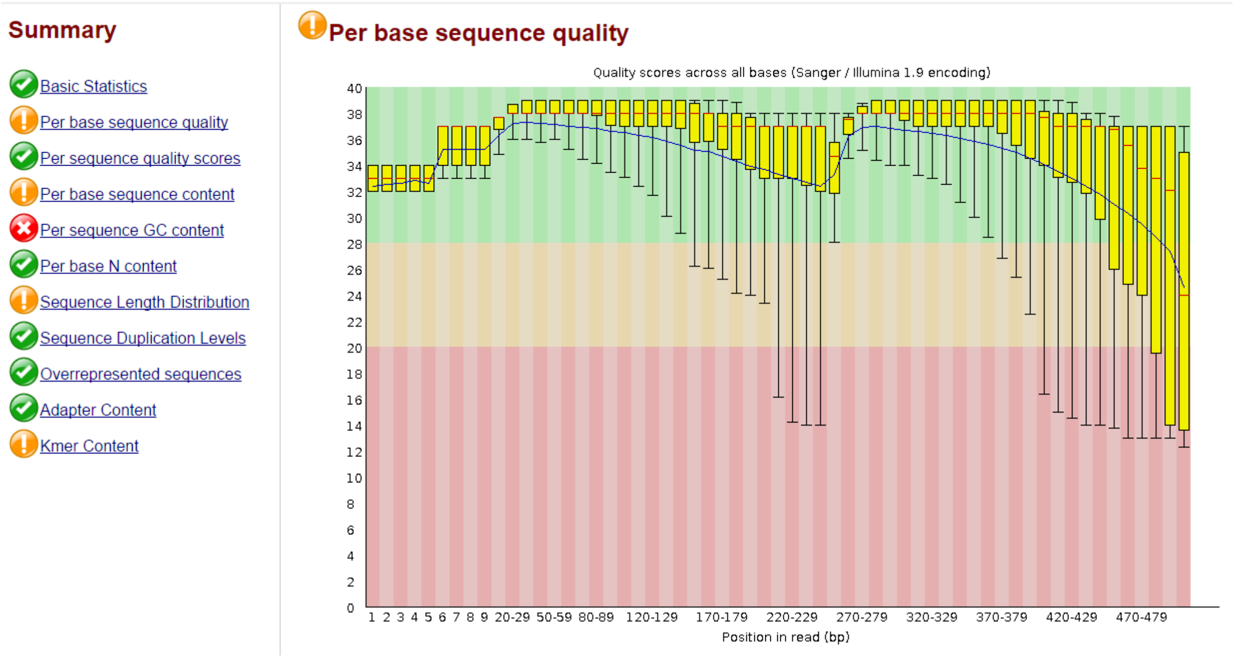
ID: std01 ~std12



Summary

- ✓ Basic Statistics
- ! Per base sequence quality
- ✓ 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
- ! Kmer Content

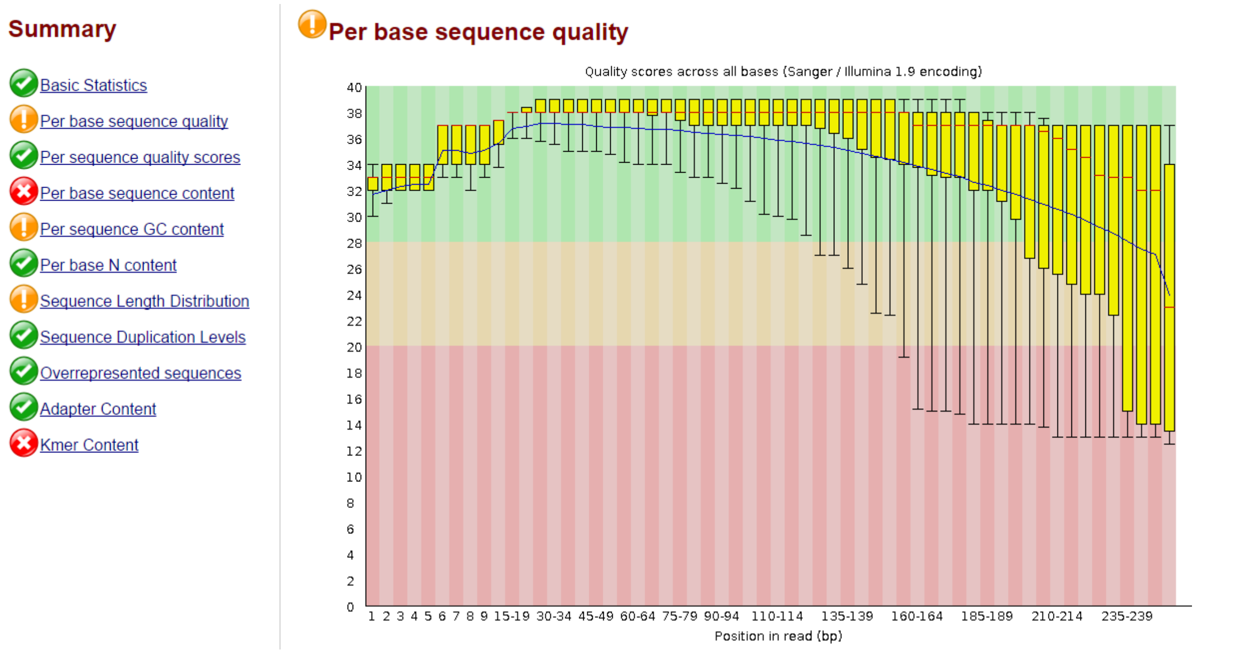
! Per base sequence quality



Summary

- ✓ Basic Statistics
- ! Per base sequence quality
- ✓ 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
- ! Kmer Content

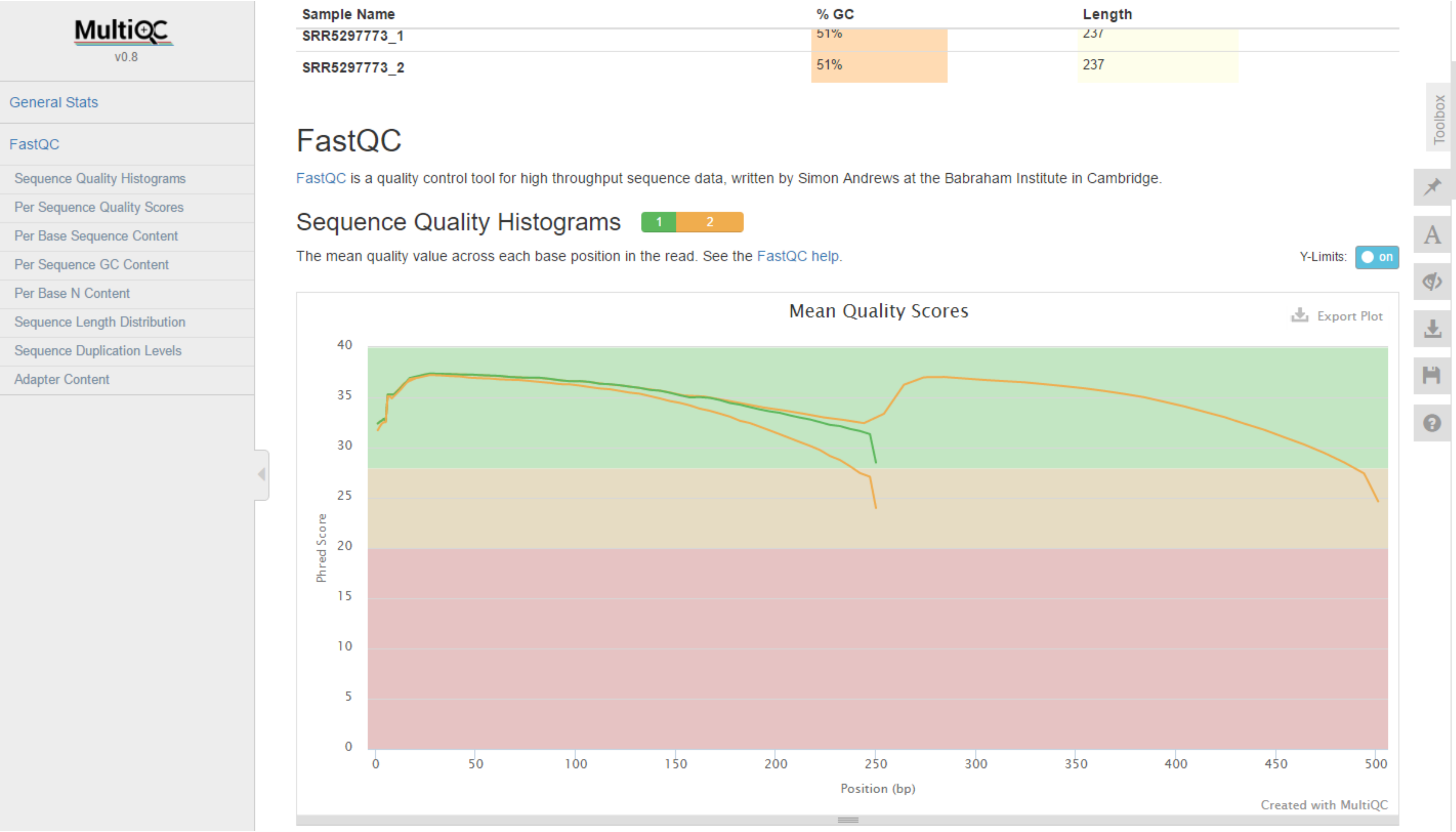
! Per base sequence quality



Summary

- ✓ Basic Statistics
- ! Per base sequence quality
- ✓ 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
- ✗ Kmer Content

! Per base sequence quality



MultiQC
v0.8
General Stats
FastQC
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
Adapter Content
Toolbox

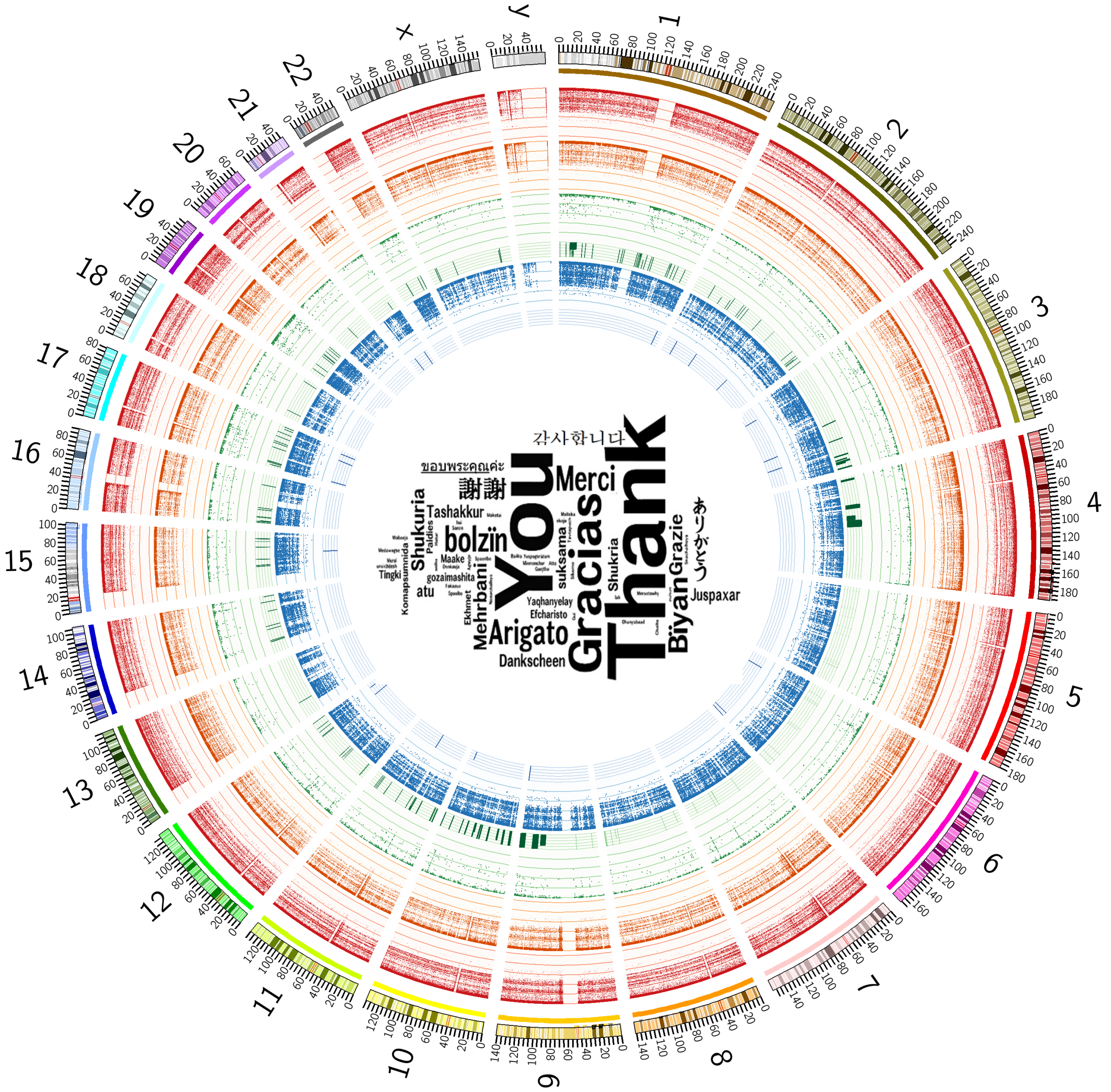
Sample Name	% GC	Length
SRR5297773_1	51%	237
SRR5297773_2	51%	237

FastQC

FastQC is a quality control tool for high throughput sequence data, written by Simon Andrews at the Babraham Institute in Cambridge.

Sequence Quality Histograms

The mean quality value across each base position in the read. See the FastQC help.


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
x
y
Thank
You
Gracias
Merci
감사합니다
ขอบพระคุณค่ะ
謝謝
ありがとう
Grazie
Biyan
Juspaxar
Shukria
suksama
Arigato
Dankscheen
Efcharisto
Yaqhanyelay
Mehrbani
Ekhmet
bolzïn
Tashakkur
Shukuria
Paldies
Komapsumnida
Tingki
atu
gozaimashita
Maake
Spasibo
Dhanyabaad
Merastawhy