Practical 4
Next Generation Sequencing

General overview practicals

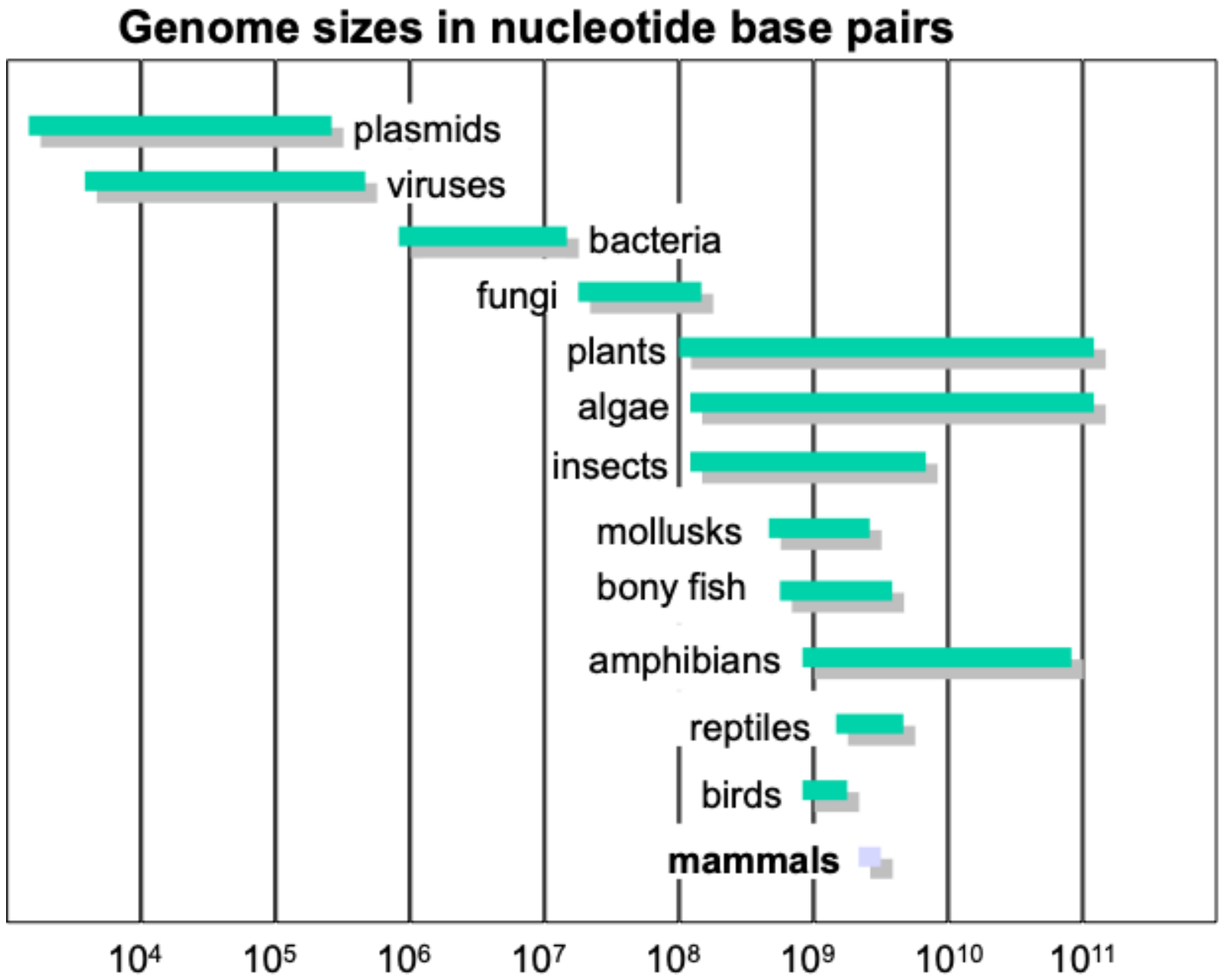
- P1: Databases, pairwise alignment & MSA
- P2: Sequence similarity searching (BLAST) & sequence motifs
- P3: Phylogeny, protein structure & gene ontology
- P4: Next generation sequencing

Next generation sequencing (NGS):

- What is it?
- Why do we do it?
- How do we do it?

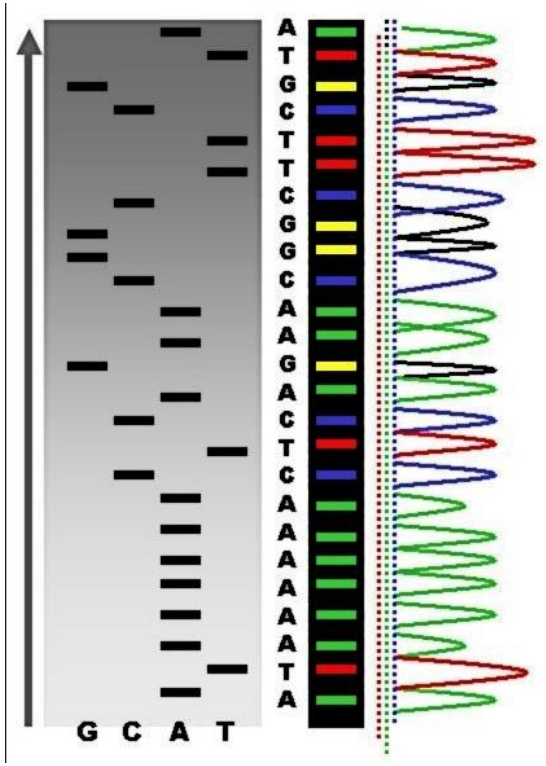
Applications of Genome Sequencing		
Purpose	Template *	Example
De novo sequencing	Genome sequencing	Sequencing >1000 influenza genomes
	Ancient DNA	Extinct Neanderthal genome
	Metagenomics	Human gut
Resequencing	Whole genomes	Individual humans
	Genomic regions	Assessment of genomic rearrangements or disease-associated regions
	Somatic mutations	Sequencing mutations in cancer
Transcriptome	Full-length transcripts	Defining regulated messenger RNA transcripts
	Serial Analysis of Gene Expression (SAGE)	
	Noncoding RNAs	Identifying and quantifying microRNAs in samples
Epigenetics	Methylation changes	Measuring methylation changes in cancer

* Template: starting material; the focus of your experiment



http://www3.kumc.edu/jcalvet/PowerPoint/bioc801b.ppt

Background: Sanger Sequencing



- Oldest technique (°1977)
- 10 euros/sequence
- 800-1000BP (relatively long reads)
- High quality & low throughput
- Commonly used to determine one or several genes
- Reference genomes
  - Viral: 1977
  - Bacterial: 1995
  - Human: 2000
- Mechanism: https://www.youtube.com/watch?v=FvHRio1yyhQ

Next Generation sequencing

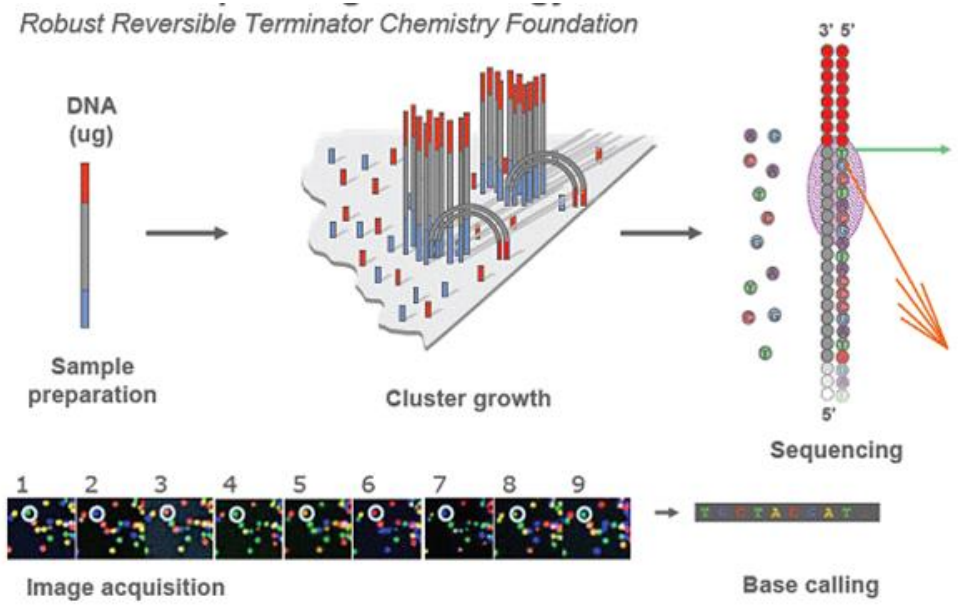
illumina®

- Based on **massive parallel sequencing** (allowing millions of fragments to be sequenced simultaneously.
- Second generation sequencing (90's)

⇒Higher throughput and lower costs (compared to Sanger)

Most popular platform: Illumina (Solexa technology)

Sequencing technique: Illumina



- 50-600Bp – often 150bp paired end
- Most used technique for genome sequencing
- Easy sample prep
- Principle = sequencing by synthesis
- Sequencing quality decreases towards the end of the read (phasing error!)
- https://www.youtube.com/watch?v=womKfikWlxM

Sequencing techniques (2)

- PacBio or SMRT sequencing
- Single Molecule Real Time sequencing
- Popularity quickly increasing
- Very long reads >10kb (up to >20kb)
  - Extremely useful for *de novo* assembly
- Was very low read quality, now up to 99.9% accuracy
- Epigenetic base modifications can be detected
- https://www.youtube.com/watch?v=_lD8JyAbwEo

Sequencing techniques (3)

- Oxford Nanopore Technologies
- Different approach compared to the ‘sequencing by synthesis’ principle
- https://www.youtube.com/watch?v=hs0FdiTHMbc
- Gigantic read lengths
- Portable: MinION

What comes out of the machine?

FastQ file containing all the sequenced reads

1 file for forward reads, 1 for reverse

Each read:

- Line 1: ‘@ + description’ . Descriptions contains information about machine, location, etc
- Line 2: sequence
- Line 3: '+' and can be followed by 1 again
- Line 4: Phred scores converted to ASCII codes

FastQ

Read 1

Read 2

```
@M00984:14:000000000-AA0HF:1:1101:21362:1290 1:N:0:3
GCGGTGTGAGATGTTTGCTTGCTAGTGTTTTCAGGAGTAGACAACATGAGAGAGCGCACAATAA
+
C9CCCGGDGFFGGGGGFDFEFGGGFGGA@EEAGG<FFFEFFGGGGGFGFGGF8@FG@+FFDDGD
@M00984:14:000000000-AA0HF:1:1101:21119:1292 1:N:0:3
GACAGAAAAGGCAGAGAGGTGCACGCCGTATACTGTGTTCACTCCAGAGTGCTACTGGCACAAT
+
CCCCCGDFGFGGGGGGGGGFDEFACFF7FCGGGGGGDGFG<FFFFGEFFGGGGFFDFGFFGGGE
```

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

Phred-Score

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

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%

NGS analysis

1. Generate sequence data
2. Quality control
3. Trimming (optional)
4. Quality control
5. Mapping
6. Remove PCR duplicates
7. Variant calling

NGS Pipeline Overview

Sample

Sequencer

Unmapped Reads
FASTQ

fastp, FASTQC,...

Quality Control

Quality trimming
fastp, trimmomatic

Unmapped Reads
FASTQ

fastp, FASTQC,...

Quality Control

Mapping to reference genome
Bowtie2, STAR, BWA-MEM, HiSat2, ...

Remove PCR duplicates
Picard, Markduplicates, rmdup, ...

Mapped Reads
BAM/SAM

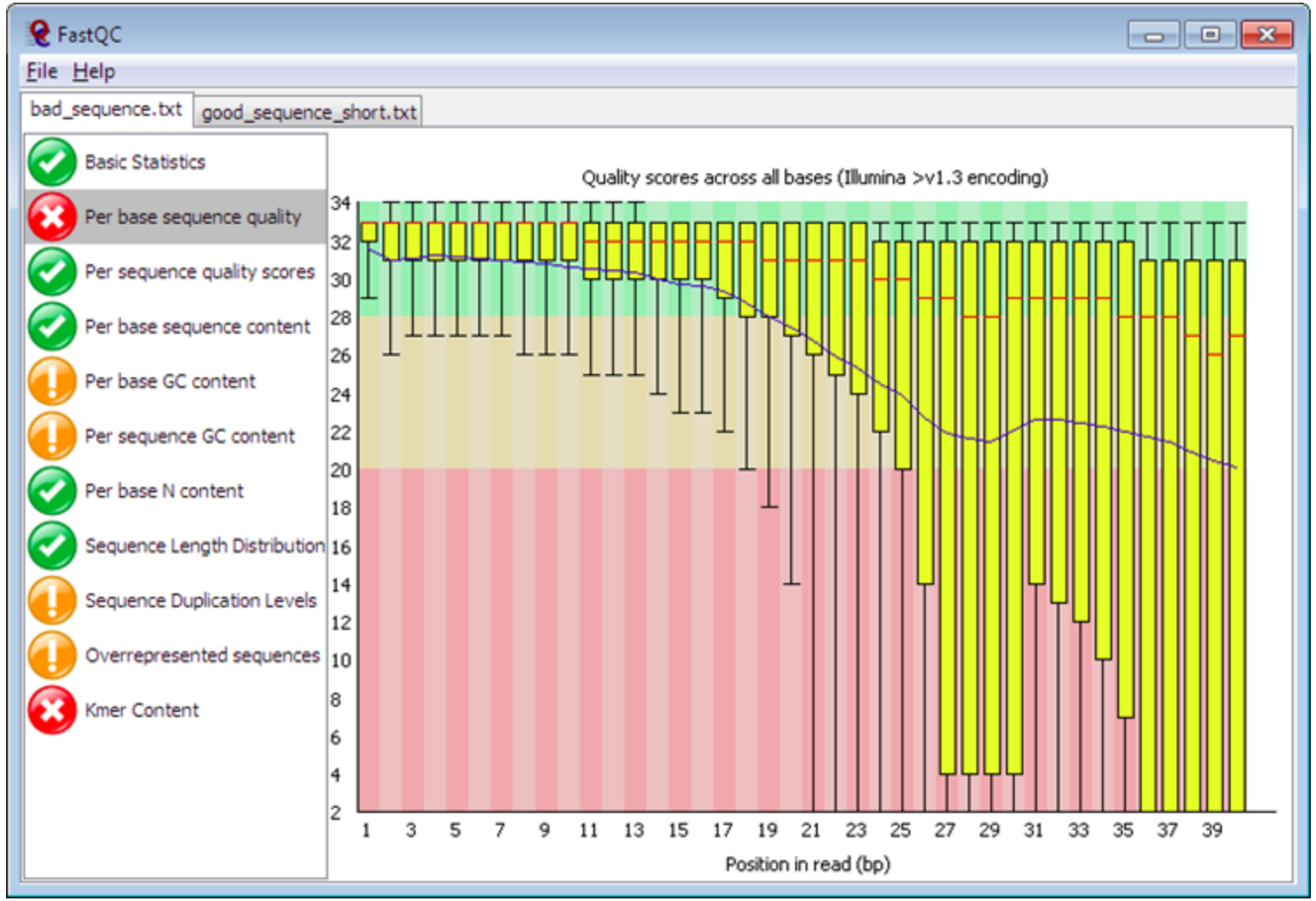
Mapped Reads
BAM/SAM

Variant calling
Lofreq, GATK, samtools

Variants (SNP/Indel)
VCF

VariantDB
Database

Quality control



- **Use a tool** to assess the **quality** of .fastq or .fastq.gz files (your sequencing data)
- **Different criteria** are measured e.g. overrepresented sequences, per base sequence quality, …
- Very nice and **easy** way to **visually** inspect data quality!

Quality control

fastp report

Summary

General

fastp version:	0.23.4 (https://github.com/OpenGene/fastp)
sequencing:	paired end (101 cycles + 101 cycles)
mean length before filtering:	101bp, 101bp
mean length after filtering:	100bp, 100bp
duplication rate:	0.005259%
Insert size peak:	169

Before filtering

total reads:	874.684000 K
total bases:	88.343084 M
Q20 bases:	74.224393 M (84.018340%)
Q30 bases:	65.932495 M (74.632322%)
GC content:	49.210692%

After filtering

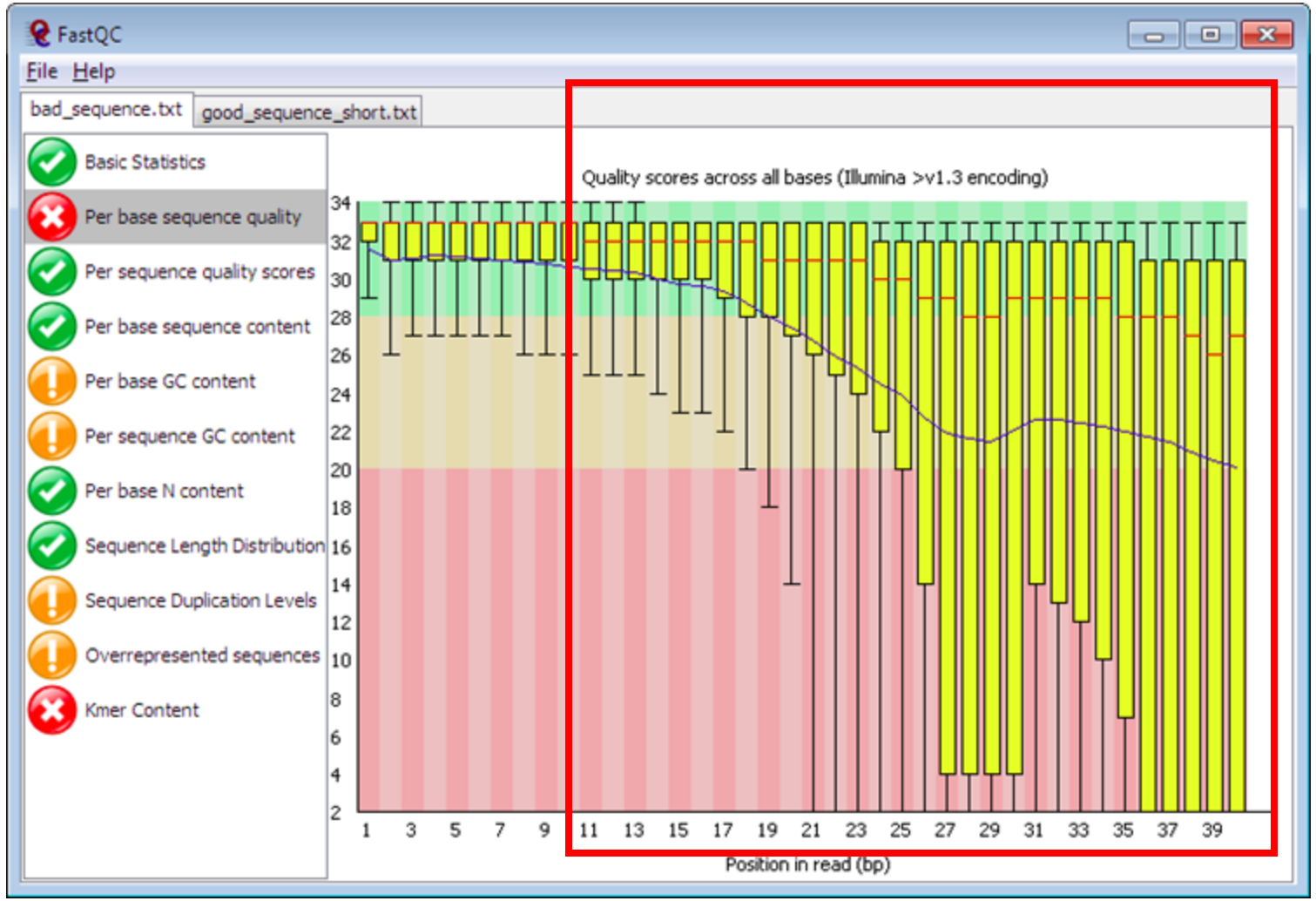
total reads:	594.676000 K
total bases:	59.793370 M
Q20 bases:	54.480386 M (91.114426%)
Q30 bases:	49.429680 M (82.667493%)
GC content:	48.446567%

Filtering result

reads passed filters:	594.676000 K (67.987525%)
reads with low quality:	279.876000 K (31.997384%)
reads with too many N:	132 (0.015091%)
reads too short:	0 (0.000000%)

- **Use a tool** to assess the **quality** of .fastq or .fastq.gz files (your sequencing data)
- **Different criteria** are measured e.g. overrepresented sequences, per base sequence quality, …
- Very nice and **easy** way to **visually** inspect data quality!

Quality control



- This case: bad quality data (low quality scores, especially towards the end)
- Is not always so clear as in this example
- Inspect criteria that fail (red cross)

- How do we solve this?

NGS Pipeline Overview

Sample

Sequencer

Unmapped Reads
FASTQ

fastp, FASTQC,...

Quality Control

Quality trimming
fastp, trimmomatic

Unmapped Reads
FASTQ

fastp, FASTQC,...

Quality Control

Mapping to reference genome
Bowtie2, STAR, BWA-MEM, HiSat2, ...

Remove PCR duplicates
Picard, Markduplicates, rmdup, ...

Mapped Reads
BAM/SAM

Mapped Reads
BAM/SAM

Variant calling
Lofreq, GATK, samtools

Variants (SNP/Indel)
VCF

VariantDB
Database

Trimming

Trimming is the process of removing low-quality bases, adapter sequences, and other unwanted regions from raw sequencing reads (FASTQ files) before downstream analysis. It ensures that only high-quality portions of the reads are used for alignment, variant calling, or other bioinformatics steps.

Types:

- Adapter trimming
- Quality trimming
- Length trimming

NGS Pipeline Overview

Sample

Sequencer

Unmapped Reads
FASTQ

fastp, FASTQC,...

Quality Control

Quality trimming
fastp, trimmomatic

Unmapped Reads
FASTQ

fastp, FASTQC,...

Quality Control

Mapping to reference genome
Bowtie2, STAR, BWA-MEM, HiSat2, ...

Remove PCR duplicates
Picard, Markduplicates, rmdup, ...

Mapped Reads
BAM/SAM

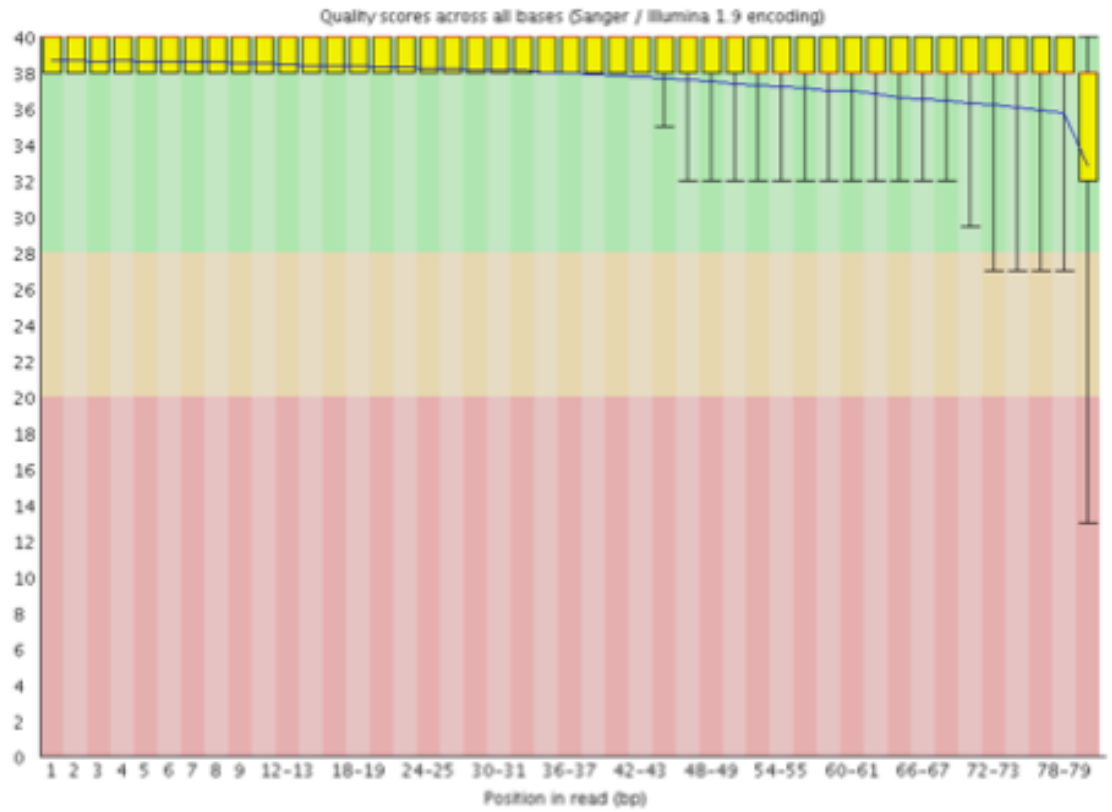
Mapped Reads
BAM/SAM

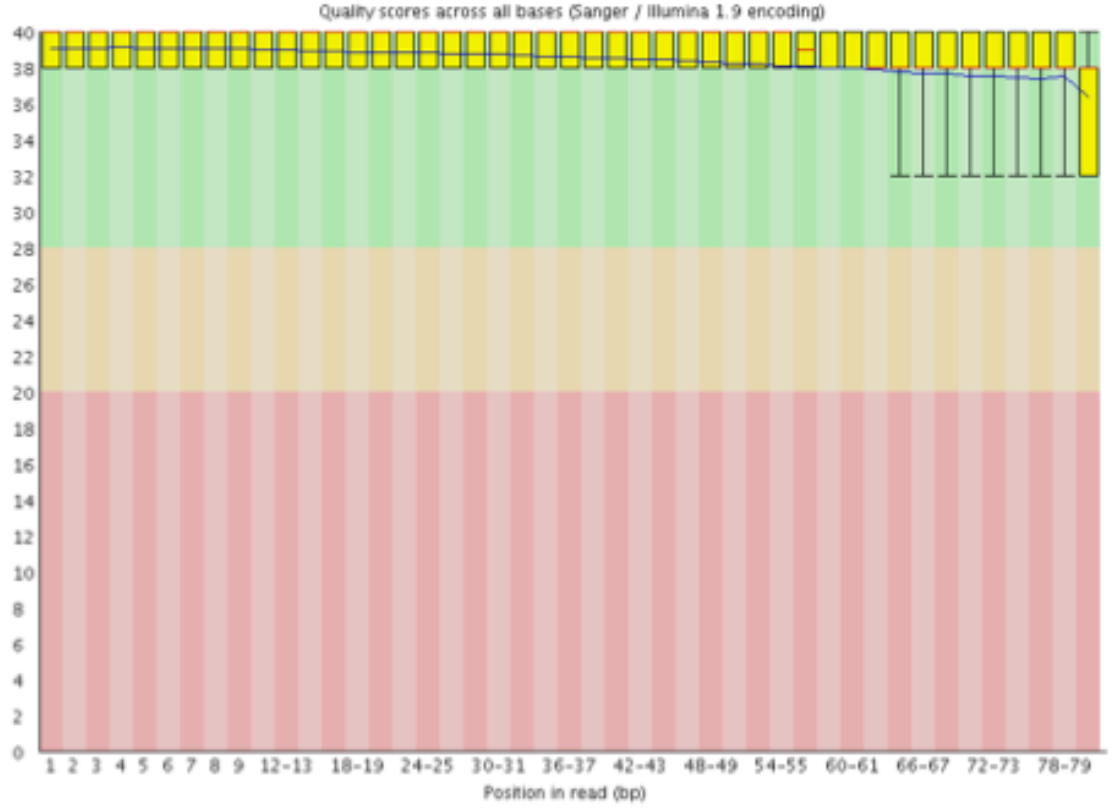
Variant calling
Lofreq, GATK, samtools

Variants (SNP/Indel)
VCF

VariantDB
Database

Quality control after trimming





NGS Pipeline Overview

Sample

Sequencer

Unmapped Reads
FASTQ

fastp, FASTQC,...

Quality Control

Quality trimming
fastp, trimmomatic

Unmapped Reads
FASTQ

fastp, FASTQC,...

Quality Control

Mapping to reference genome
Bowtie2, STAR, BWA-MEM, HiSat2, ...

Remove PCR duplicates
Picard, Markduplicates, rmdup, ...

Mapped Reads
BAM/SAM

Mapped Reads
BAM/SAM

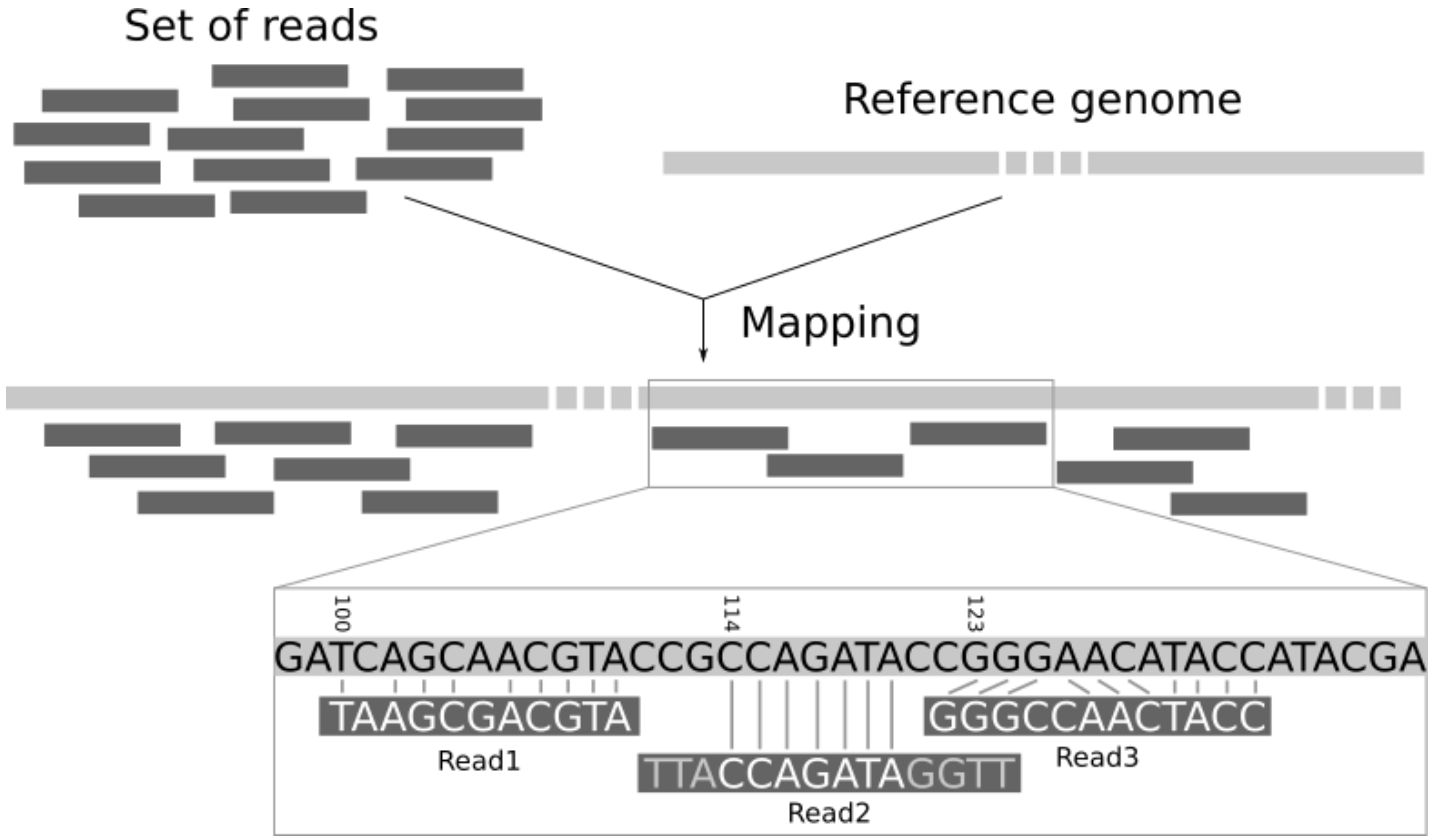
Variant calling
Lofreq, GATK, samtools

Variants (SNP/Indel)
VCF

VariantDB
Database

Mapping

- “Map” your processed data to the reference genome



NGS Pipeline Overview

Sample

Sequencer

Unmapped Reads
FASTQ

fastp, FASTQC,...

Quality Control

Quality trimming
fastp, trimmomatic

Unmapped Reads
FASTQ

fastp, FASTQC,...

Quality Control

Mapping to reference genome
Bowtie2, STAR, BWA-MEM, HiSat2, ...

Remove PCR duplicates
Picard, Markduplicates, rmdup, ...

Mapped Reads
BAM/SAM

Mapped Reads
BAM/SAM

Variant calling
Lofreq, GATK, samtools

Variants (SNP/Indel)
VCF

VariantDB
Database

NGS Pipeline Overview

Sample

Sequencer

Unmapped Reads
FASTQ

fastp, FASTQC,…

Quality Control

Quality trimming
fastp, trimmomatic

Unmapped Reads
FASTQ

fastp, FASTQC,…

Quality Control

Mapping to reference genome
Bowtie2, STAR, BWA-MEM, HiSat2, …

Remove PCR duplicates
Picard, Markduplicates, rmdup, …

Mapped Reads
BAM/SAM

Mapped Reads
BAM/SAM

Variant calling
Lofreq, GATK, samtools

Variants (SNP/Indel)
VCF

VariantDB
Database

This practical:

- Exercise on FASTQ and the ASCII format

- Using Galaxy to perform the NGS pipeline

  Galaxy is a scientific workflow and **data analysis platform** that makes **computational biology** accessible to research scientists that do not have computer programming experience

9.1 FASTQ file format

Determine the phred quality scores of the underlined bases, given these are Illumina reads (offset = +33). What is the probability that these bases are wrong?

```
@M00984:14:000000000-AA0HF:1:1101:23031:1298 1:N:0:1
CGTGCCAACGGCACTCGTACACGAGTTGTACAGAAACTGAT
+
CCCCCGGGGGGGGGGGGGGGGGFDGGGGGGGFFGGGGG9F
```

- **G**: Quality score “C” = 67 in ASCII table

- **A**: Quality score “9” = 57 in ASCII table

9.1 FASTQ file format

Determine the phred quality scores of the underlined bases, given these are Illumina reads (offset = +33). What is the probability that these bases are wrong?

```
@M00984:14:000000000-AA0HF:1:1101:23031:1298 1:N:0:1
CGTGCCAACGGCACTCGTACACGAGTTGTACAGAAACTGAT
+
CCCCCGGGGGGGGGGGGGGGGGFDGGGGGGGFFGGGGG9F
```

- **G**: Quality score “C” = 67 in ASCII table
  - Subtract offset: 67 – 33 = 34

- **A**: Quality score “9” = 57 in ASCII table
  - Subtract offset: 57 – 33 = 24

9.1 FASTQ file format

Determine the phred quality scores of the underlined bases, given these are Illumina reads (offset = +33). What is the probability that these bases are wrong?

```
@M00984:14:000000000-AA0HF:1:1101:23031:1298 1:N:0:1
CGTGCCAACGGCACTCGTACACGAGTTGTACAGAAACTGAT
+
CCCCCGGGGGGGGGGGGGGGGGFDGGGGGGGFFGGGGG9F
```

- **G**: Quality score "C" = 67 in ASCII table
  - Subtract offset: 67 – 33 = 34

- **A**: Quality score "9" = 57 in ASCII table
  - Subtract offset: 57 – 33 = 24

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

$$\downarrow$$

$$P = 10^{\frac{Q}{-10}}$$

9.1 FASTQ file format

Determine the phred quality scores of the underlined bases, given these are Illumina reads (offset = +33). What is the probability that these bases are wrong?

```
@M00984:14:000000000-AA0HF:1:1101:23031:1298 1:N:0:1
CGTGCCAACGGCACTCGTACACGAGTTGTACAGAAACTGAT
+
CCCCCGGGGGGGGGGGGGGGGGFDGGGGGGGFFGGGGG9F
```

- **G**: Quality score "C" = 67 in ASCII table
  - Subtract offset: 67 – 33 = 34
  - Probability this base is wrong is $10^{\frac{34}{-10}} = 0.000398$
- **A**: Quality score "9" = 57 in ASCII table
  - Subtract offset: 57 – 33 = 24
  - Probability this base is wrong is $10^{\frac{24}{-10}} = 0.00398$

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

$$\downarrow$$

$$P = 10^{\frac{Q}{-10}}$$

9.1 FASTQ file format

What ASCII Offset was used below? Important information: The Q score never exceeds 40 or falls below 0.

```
@SRR038845.3 HWI-EAS038:6:1:0:1938 length=36
CAACGAGTTCACACCTTGGCCGACAGGCCCGGGTAA
+SRR038845.3 HWI-EAS038:6:1:0:1938 length=36
BA@7>B=>:>>7@7@>>9=BAA?;>52;>:9=8.=A
```

9.1 FASTQ file format

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

Sanger and newest Illumina machines (>1.8): Offset +33

9.1 FASTQ file format

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

Sanger and newest Illumina machines (>1.8): Offset +33

Solexa/Illumina 1.0: +59

9.1 FASTQ file format

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

Sanger and newest Illumina machines (>1.8): Offset +33

Solexa/Illumina 1.0: +59

Illumina 1.3 -1.8: +64

9.1 FASTQ file format

What ASCII Offset was used below? Important information: The Q score never exceeds 40 or falls below 0.

```
@SRR038845.3 HWI-EAS038:6:1:0:1938 length=36
CAACGAGTTCACACCTTGGCCGACAGGCCCGGGTAA
+SRR038845.3 HWI-EAS038:6:1:0:1938 length=36
BA@7>B=>:>>7@7@>>9=BAA?;>52;>:9=8.=A
```

- We can see digits in the quality score -> what does this tell us?

9.1 FASTQ file format

What ASCII Offset was used below? Important information: The Q score never exceeds 40 or falls below 0.

```
@SRR038845.3 HWI-EAS038:6:1:0:1938 length=36
CAACGAGTTCACACCTTGGCCGACAGGCCCGGGTAA
+SRR038845.3 HWI-EAS038:6:1:0:1938 length=36
BA@7>B=>:>>7@7@>>9=BAA?;>52;>:9=8.=A
```

- We can see digits in the quality score -> what does this tell us?

- Digits range in values between 48-57
  - Offset +33: 15-24 → Sanger and newest Illumina machines (>1.8) encoding
  - Offset +59: below 0!
  - Offset +64: below 0!

9.1 FASTQ file format

What ASCII Offset was used below (and so, which format is it)?

```
@HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
TTAATTGGTAAATAAATCTCCTAATAGCTTAGATNTTACCTTNNNNNNNNNNTAGTTTCTTG
+HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
efcfffffcfeefffcffffffddf`feed]`]_Ba_^__[YBBBBBBBBBBRTT\]][]dd
```

- Similarly, e.g. ‘f’ has ASCII value **102**.

9.1 FASTQ file format

What ASCII Offset was used below (and so, which format is it)?

```
@HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
TTAATTGGTAAATAAATCTCCTAATAGCTTAGATNTTACCTTNNNNNNNNNNTAGTTTCTTG
+HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
efcfffffcfeefffcffffffddf`feed]`]_Ba_^__[YBBBBBBBBBBRTT\]][]dd
```

- Similarly, e.g. ‘f’ has ASCII value **102**.

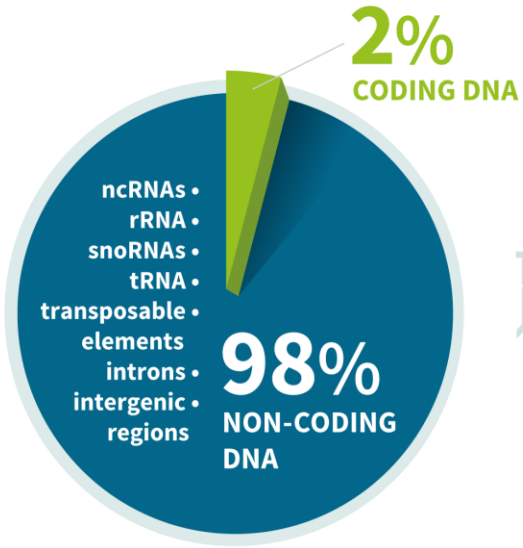
- Offset +33: above 40!
- Offset +59: above 40!
- Offset +64: 38 → Illumina 1.3 -1.8 encoding

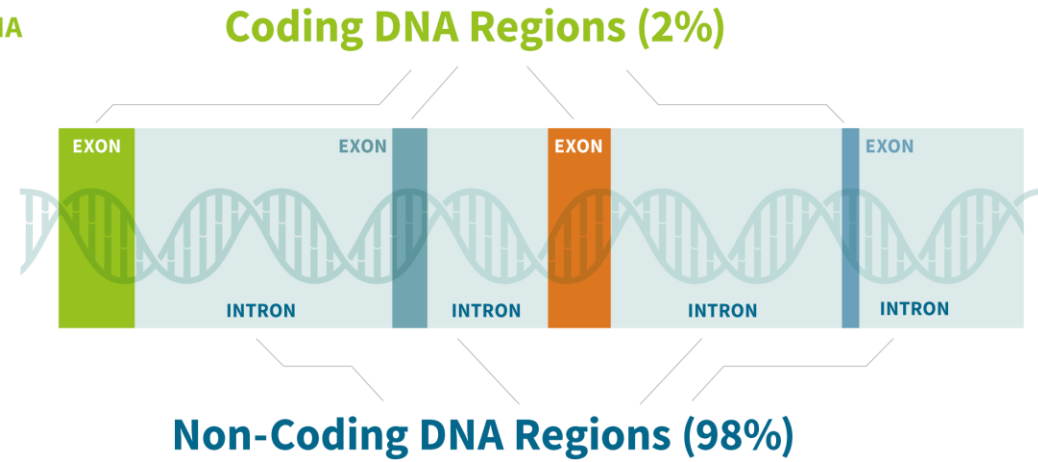
Galaxy NGS analysis

About sequencing data

Why not sequence the complete human genome, but look at exomes only?

- Very large size difference
- More price and time efficient
- Often good enough for studying genetic diseases





Quality control and trimming

1. Which other files get created in addition to the fastq files?

2. What is the percentage of reads that are retained?

3. What is the most frequent 5-mer (sequence of 5 nucleotides) in the R2 dataset before and after fastp quality control? Which 5-mers do you think are biologically relevant?

Quality control and trimming

1. Which other files get created in addition to the fastq files?

   A fastp report both in json and in html format


```
{
    "summary": {
        "fastp_version": "0.23.4",
        "sequencing": "paired end (101 cycles + 101 cycles)",
        "before_filtering": {
            "total_reads":874684,
            "total_bases":88343084,
            "q20_bases":74224393,
            "q30_bases":65932495,
            "q20_rate":0.840183,
            "q30_rate":0.746323,
            "read1_mean_length":101,
            "read2_mean_length":101,
            "gc_content":0.492107
        },
        "after_filtering": {
            "total_reads":594676,
            "total_bases":59793370,
            "q20_bases":54480386,
            "q30_bases":49429680,
            "q20_rate":0.911144,
            "q30_rate":0.826675,
            "read1_mean_length":100,
            "read2_mean_length":100,
            "gc_content":0.484466
        }
    },
    "filtering_result": {
        "passed_filter_reads": 594676,
        "low_quality_reads": 279876,
        "too_many_N_reads": 132,
        "too_short_reads": 0,
        "too_long_reads": 0
    },
    "duplication": {
        "rate": 5.25904e-05
    },
    "insert_size": {
        "peak": 169,
        "unknown": 134544,
        "histogram": [0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,7,10,8,11,10,7,18,
    },
```


fastp report

Summary

General

fastp version:	0.23.4 (https://github.com/OpenGene/fastp)
sequencing:	paired end (101 cycles + 101 cycles)
mean length before filtering:	101bp, 101bp
mean length after filtering:	100bp, 100bp
duplication rate:	0.005259%
Insert size peak:	169

Before filtering

total reads:	874.684000 K
total bases:	88.343084 M
Q20 bases:	74.224393 M (84.018340%)
Q30 bases:	65.932495 M (74.632322%)
GC content:	49.210692%

After filtering

total reads:	594.676000 K
total bases:	59.793370 M
Q20 bases:	54.480386 M (91.114426%)
Q30 bases:	49.429680 M (82.667493%)
GC content:	48.446567%

Filtering result

reads passed filters:	594.676000 K (67.987525%)
reads with low quality:	279.876000 K (31.997384%)
reads with too many N:	132 (0.015091%)
reads too short:	0 (0.000000%)

Quality control and trimming

1. Which other files get created in addition to the fastq files?
2. What is the percentage of reads that are retained?

SRR12733957

Filtering result

reads passed filters:	594.676000 K (67.987525%)
reads with low quality:	279.876000 K (31.997384%)
reads with too many N:	132 (0.015091%)
reads too short:	0 (0.000000%)

SRR11954102

Filtering result

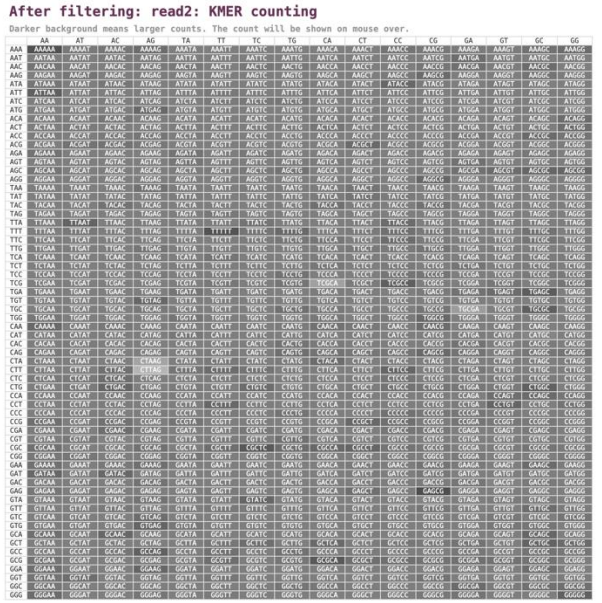
reads passed filters:	2.649948 M (90.657881%)
reads with low quality:	270.890000 K (9.267470%)
reads with too many N:	2.066000 K (0.070680%)
reads too short:	116 (0.003968%)

Quality control and trimming

1. Which other files get created in addition to the fastq files?
2. What is the percentage of reads that are retained?
3. What is the most frequent 5-mer (sequence of 5 nucleotides) in the R2 dataset before and after fastp quality control? Which 5-mers do you think are biologically relevant?

Before: GGGGG

After: AAAAA and TTTTT -> the latter can be part of the polyA tail and thus biologically relevant, while polyG is most likely a sequencing artefact.



Convert SAM to BAM Format

- How does the filesize differ between the .sam and the .bam file?

BAM is significantly smaller than SAM

- What do the first 5 rows of the .sam file look like? Could you do the same for the .bam?

Convert SAM to BAM Format

SAM

```
@SQ     SN:NC_045512.2  LN:29903
@PG     ID:bwa-mem2     PN:bwa-mem2     VN:2.2.1        CL:bwa-mem2 mem GCF_009858895.2_ASM985889v3_ge
SRR12733957.1   77      *       0       0       *       *       0       0
GGATAAGGCGCAGCGGTCGGGCTGAACGGGGGGTTCGTGCACACAGCCCAGCTTGGAGCGAACGACCTACACCGAACTGAGATACCTACAGCGTGAGCTAT
SRR12733957.1   141     *       0       0       *       *       0       0       GCTTCCAGGGGGAAACGCCTGG
FFFFFFFF:FFF:FFFF,::F,FF:F:F:FFFFFF,F,F:FFF:FF::FFFF,FFF:F,FFF,F,,:F,,F:F,F,,F,F,:F,FFFF,,FFFF,F:,FF:
SRR12733957.3   77      *       0       0       *       *       0       0
CAGCAGGGTGAGGTGCTCCTTGAAGGAAGCGCCGATCCCGGCGGCGCGCAATTCGTCGACGAAGTTGAAGAACATGGGCTGCTGCTTAACCCGCGCACCGC
SRR12733957.3   141     *       0       0       *       *       0       0       CGCCTCATCGGATGCCTCGGCA
FFFFFFFFFFFFF:FF:FFFFFFF:F:,:F:,FFFFF,FFF::FFFFF:FF,FFFFF,,F:FFFF,F,FFFFF:FFFFFF,FFFFFFFFF,FFF:F:FFF:
SRR12733957.4   77      *       0       0       *       *       0       0       TTTATAAGGGATTTTGGCGATT
```

BAM

```
6=©E—R
x
©TåQ|I_…à‘ø√.ª3≥À0É¸¨ıflB®ßÊXÂiûi÷Ç<ÕSêí≈Ì^1ΔìaZ^Ld+
∏»flÛ2à(Pé≥È
$°ëq6“å…D$ât_ÕûÀ9÷õjÔÏvE◊{C‘bi°«Rt5◊Ó‰'O◊¶^G∫π~±♣À6ΔjcΩsª˙aÏ=_&9^♣ÿ♣‰º≠.$*â     BÑÕ˘“ÿ_(úμ
Ç¸»©“"ZHπ`AB,!YFC@‰áç=X6äú‰áA, ÏÑ/À$?…3Qu˜ÓfiΩªflÓ~°=wfjÓº›{˜Ó>ı‘™ZμæÎo+•æ≤hVY+ıÆ˙¸ËûRJΩ¸·>;
üı´À~˝Kõo€◊áÒÂ;ÍS/flQ{—{{ë8ìh
©T)%;Æ)ãw:Ñÿ¯Eò–ïå)%O˙Ÿñ^¶Î°ìnfi˘E˘∫ì%∞¸»Ï¢õ‘ı']7ªÏ∫^g¸Aªò}fiFßø—n](ˆìÿd™Å!ë∑ŒSÚmÕ^G#—
f°hG§˝≥ÓdkÎOÜÆOΔ˝~íÏh«“ËByW≠èÏjB8áTS®H‘˙.590ÆZ™)x§8&°jäf_æÌª^˘%≤~9˜˜∫M◊KÇmN…qè•—•íRL≤¬[˘≈&
y∞ûä≥,¥˘ûMás>uN∫n>ùs©o‡¬^∑}‘–±Ã¸flıÛÀÆÎ€#Û∑ÙìÓX]*ƒ$qQJ&)ût♣/!Äñ(&qCËKNÑï[£ì°öu©_–§^“≠¶S?¢ÕÙ
```

Convert SAM to BAM Format

- How does the filesize differ between the .sam and the .bam file?

BAM is significantly smaller than SAM

- What do the first 5 rows of the .sam file look like? Could you do the same for the .bam?

The sam is a fastq-like format, while the bam cannot be opened this way, it is a binary file.

Removing duplicates (MarkDuplicates)

- Why is it important to identify and remove PCR duplicates in the analysis of NGS data?

Cause: The same fragment being sequenced several times = Artificial overrepresentation of a single molecule → Not true sequencing replicates

Consequence: Falsely inflate the abundance of certain sequences, especially when you have low concentrations of starting DNA → Overestimate genome coverage, allele frequency and expression levels.

Variant calling

- What is the role of the reference genome in the variant calling step? How does the choice of reference genome affect the interpretation of results? What if you would have a SARS-CoV-2 reference genome from later on in the pandemic?
- What is the significance of the min-cov (Minimal Coverage) and min-bq (Minimum BaseQ) parameters? How do these parameters affect the sensitivity and specificity of the variant calling?

Variant calling

- What is the role of the reference genome in the variant calling step? How does the choice of reference genome affect the interpretation of results? What if you would have a SARS-CoV-2 reference genome from later on in the pandemic?

- Template for aligning reads and identifying variants
- Different reference genome -- different variant calls
- Later genome -- evolutionary changes of the virus -- no longer matching -- different variant calls -- take into account the current dominant strain

Variant calling

- What is the significance of the min-cov (Minimal Coverage) and min-bq (Minimum BaseQ) parameters? How do these parameters affect the sensitivity and specificity of the variant calling?
  - "Minimal Coverage" : minimum read depth coverage.
    - Higher values **increase specificity** by reducing false positives but may miss low-frequency variants.
  - "Minimum baseQ" : minimum base quality score
    - Higher values **increase specificity** by filtering out low-quality bases but **may reduce sensitivity**.
  - Balance between sensitivity and specificity, sequencing depth, and the expected variant frequencies.

Annotating Variant Effects

- What is the most frequent variant type, and what percentage of the total variants does this constitute?
- Not all mutations in the DNA lead to a difference on the protein level. How many of the mutations will have no functional impact? Is this more or less common than a mutation that changes the amino acid, and by how much?

Annotating Variant Effects

- What is the most frequent variant type, and what percentage of the total variants does this constitute?

snpEff eff/ann output:

Type	Count	Percent
DEL	8	4.571429%
INS	5	2.857143%
SNP	162	92.571429%

Number variants by type

Type	Total
SNP	162
MNP	0
INS	5
DEL	8
MIXED	0
INV	0
DUP	0
BND	0
INTERVAL	0
Total	175

SNP (Single Nucleotide Polymorphism) = the substitution of one nucleotide for another -- errors in DNA replication or due to mutagens.

Annotating Variant Effects

- Not all mutations in the DNA lead to a difference on the protein level. How many of the mutations will have no functional impact? Is this more or less common than a mutation that changes the amino acid, and by how much?

Type	Count	Percent
MISSENSE	281	71.501272%
NONSENSE	31	7.888041%
SILENT	81	20.610687%

Number of effects by functional class

Type (alphabetical order)	Count	Percent
MISSENSE	283	71.646%
NONSENSE	31	7.848%
SILENT	81	20.506%

Silent mutations (20.51%) : no functional impact on the protein level.

Missense mutations (51.14%) : change the amino acid

Nonsense mutations (7.85%) : change the amino acid and create a stop codon

MultiQC

- What is the purpose of a MultiQC report in the context of bioinformatics analysis?

MultiQC

- What is the purpose of a MultiQC report in the context of bioinformatics analysis?

Combine results from multiple samples and analyses all into 1 report.

- Easier comparisons of metrics between samples and tools
- Facilitate interpretation of large datasets
- Overview of data quality, processing statistics and issues

Final results

Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7	Column 8	Column 9	Column 10	Column 11	Column 12
CHROM	POS	REF	ALT	QUAL	DP	AF	SB	DP4	EFF[*].IMPACT	EFF[*].FUNCLASS	EFF[*].EFFECT
NC_045512.2	84	C	T	7114.0	208	0.975962	0	0,1,102,105	MODIFIER	NONE	intergenic_region
NC_045512.2	160	G	T	144.0	254	0.031496	14	166,77,10,0	MODIFIER	NONE	intergenic_region
NC_045512.2	219	G	T	87.0	546	0.010989	10	335,205,6,0	MODIFIER	NONE	intergenic_region
NC_045512.2	241	C	T	23752.0	679	0.967599	0	0,0,416,261	MODIFIER	NONE	intergenic_region
NC_045512.2	443	GT	G	53.0	478	0.006276	5	170,314,2,1	HIGH	NONE	frameshift_variant