Intro to NGS data

Anders Albrechtsen

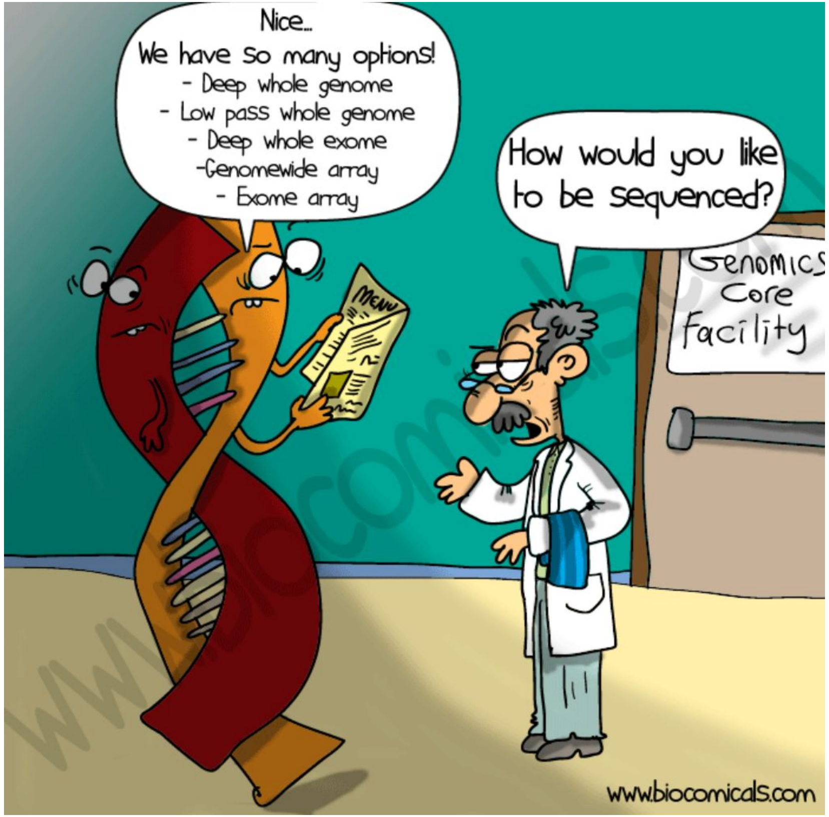
KØBENHAVNS UNIVERSITET

slido

Join at slido.com
#9464757

ⓘ Click **Present with Slido** or install our Chrome extension to display joining instructions for participants while presenting.

Many type of sequencing

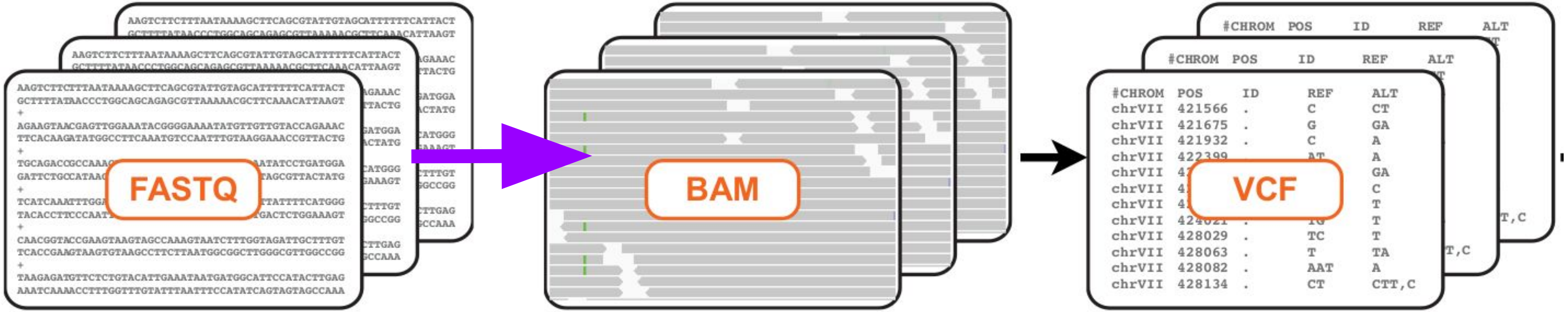


slido

Which kind of sequencing data are you working with?

ⓘ Click **Present with Slido** or install our Chrome extension to activate this poll while presenting.

This session

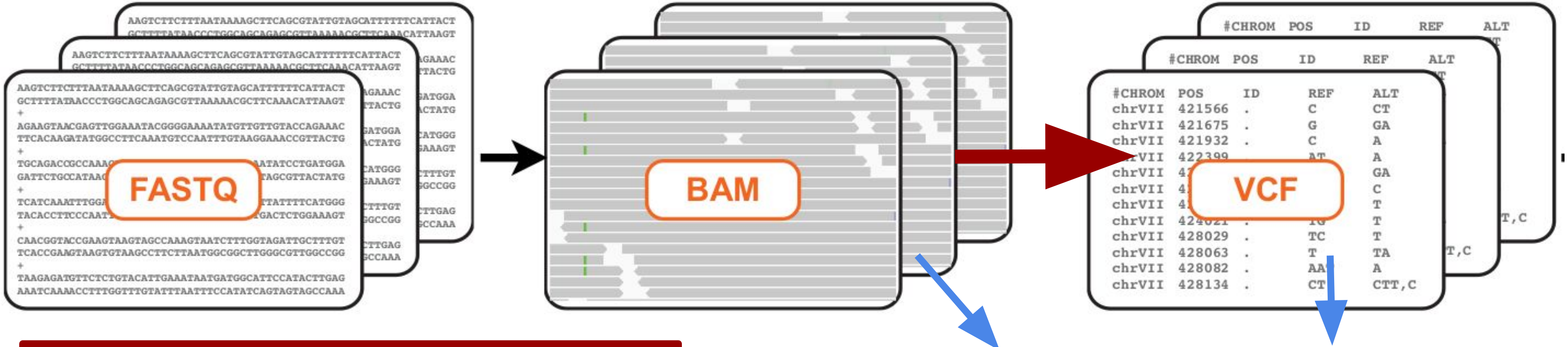


This afternoon

- Sequencing data types
- QC
- Alignment and mapping
- Exploring bam files

KØBENHAVNS UNIVERSITET

This session

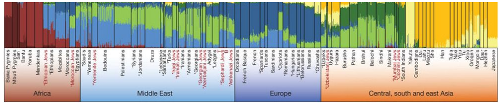


Tomorrow

- Genotype likelihoods (GL)
- Estimate allele frequencies
- Calling variable sites
- Calling genotype

Later

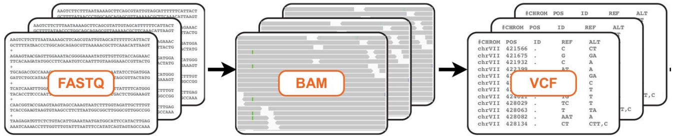
- Perform analysis from Genotypes or GLs

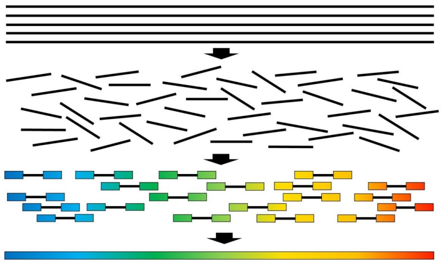


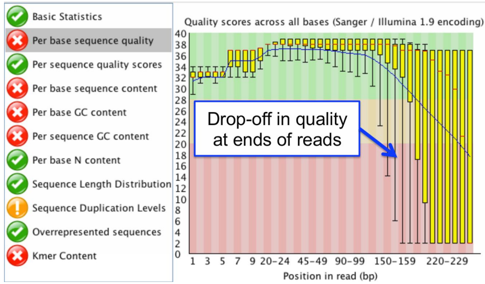
Objectives this afternoon

To understand

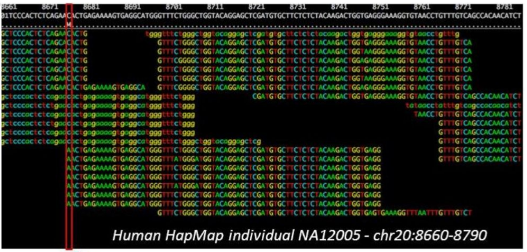
- Types of sequencing
  - Single/pair end, mate pair
- FastQ files
  - Quality
  - reads
- QC
  - Adapter contamination
  - Duplicated read
  - Sequenicng errors
  -
- Bam files
  - Mapping quality
  - Exploring variants



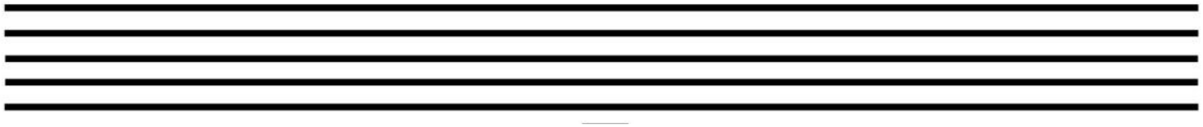



✓ Basic Statistics
✗ Per base sequence quality
✓ Per sequence quality scores
✗ Per base sequence content
✗ Per base GC content
✗ Per sequence GC content
✓ Per base N content
✓ Sequence Length Distribution
! Sequence Duplication Levels
✓ Overrepresented sequences
✗ Kmer Content

Drop-off in quality at ends of reads

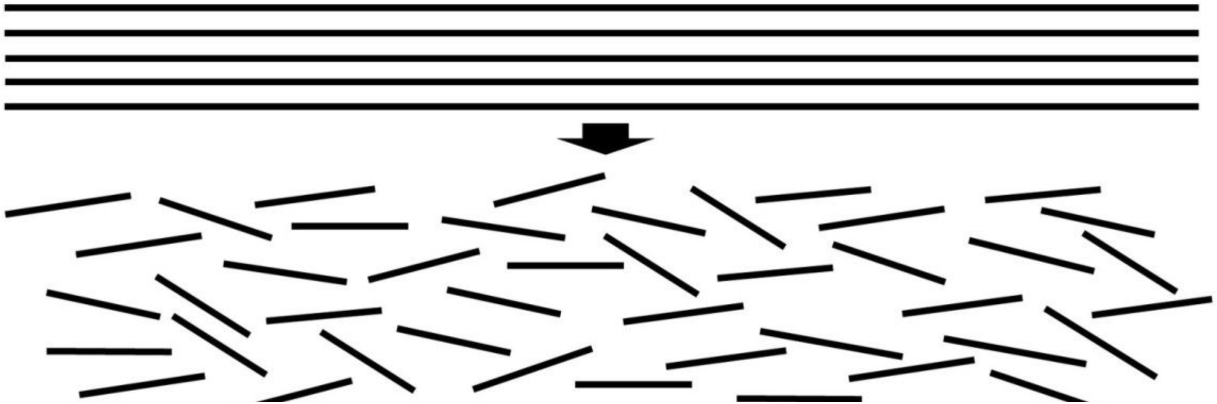

Human HapMap individual NA12005 - chr20:8660-8790

High throughput sequencing



Multiple versions of the same genome (DNA from many cells)

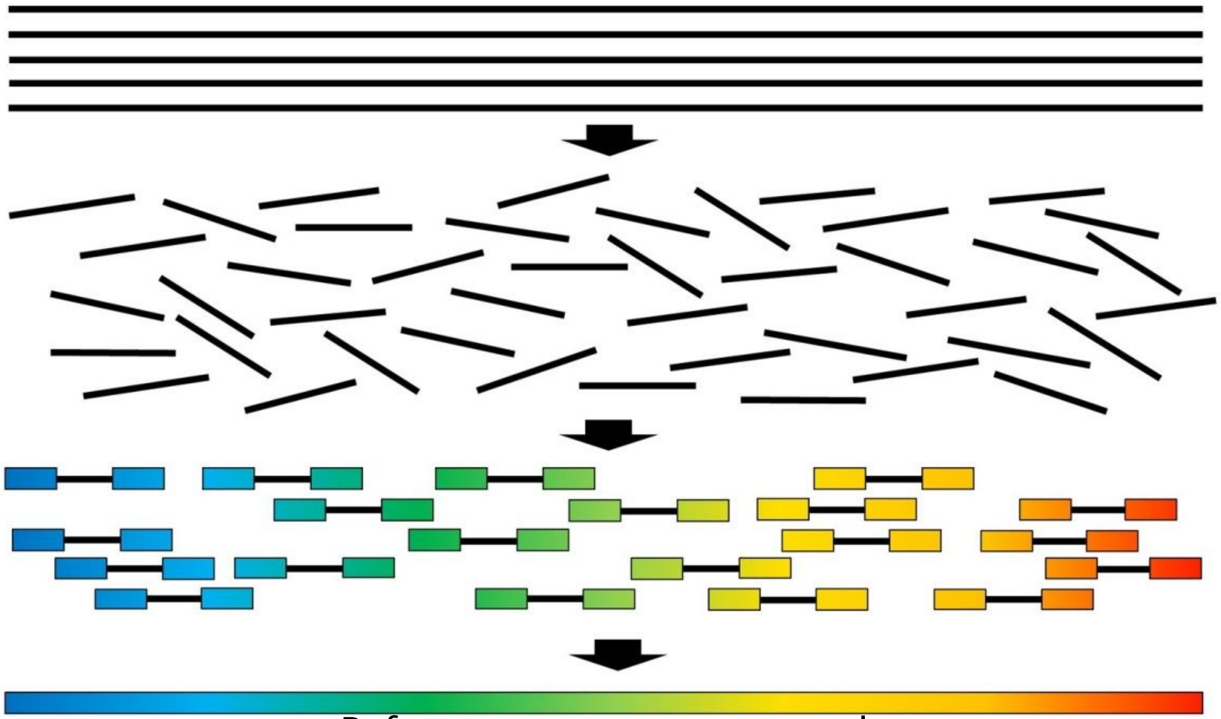
Next generation sequencing



Multiple versions of the same genome (DNA from many cells)

Genome fractured into small fragments

Next generation sequencing



Multiple versions of the same genome (DNA from many cells)

Genome fractured into small fragments

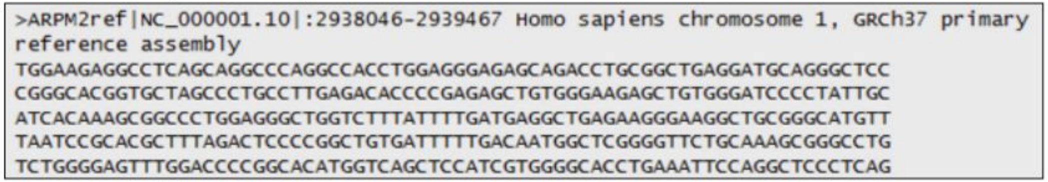
Fragments sequenced and ordered according to position on the genome

Reference genome or assemply

KØBENHAVNS UNIVERSITET

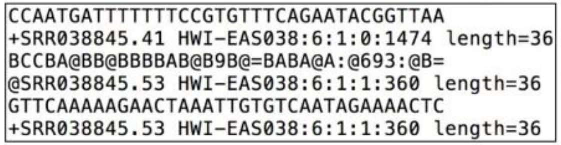
Data formats

Genome (FASTA)



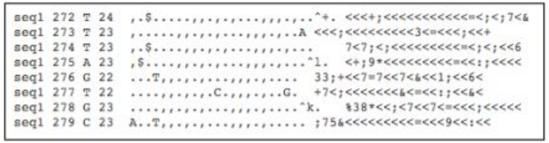
```
>ARPM2ref|NC_000001.10|:2938046-2939467 Homo sapiens chromosome 1, GRCh37 primary
reference assembly
TGGAAGAGGCCTCAGCAGGCCCAGGCCACCTGGAGGGAGAGCAGACCTGCGGCTGAGGATGCAGGGCTCC
CGGGCACGGTGCTAGCCCTGCCTTGAGACACCCCGAGAGCTGTGGGAAGAGCTGTGGGATCCCCTATTGC
ATCACAAAGCGGCCCTGGAGGGCTGGTCTTTATTTTGATGAGGCTGAGAAGGGAAGGCTGCGGGCATGTT
TAATCCGCACGCTTTAGACTCCCCGGCTGTGATTTTTGACAATGGCTCGGGGTTCTGCAAAGCGGGCCTG
TCTGGGGAGTTTGGACCCCGGCACATGGTCAGCTCCATCGTGGGGCACCTGAAATTCCAGGCTCCCTCAG
```

Reads (FASTQ)



```
CCAATGATTTTTTTCCGTGTTTCAGAATACGGTTAA
+SRR038845.41 HWI-EAS038:6:1:0:1474 length=36
BCCBA@BB@BBBBAB@B9B@=BABA@A:@693:@B=
@SRR038845.53 HWI-EAS038:6:1:1:360 length=36
GTTCAAAAAGAACTAAATTGTGTCAATAGAAAACTC
+SRR038845.53 HWI-EAS038:6:1:1:360 length=36
```

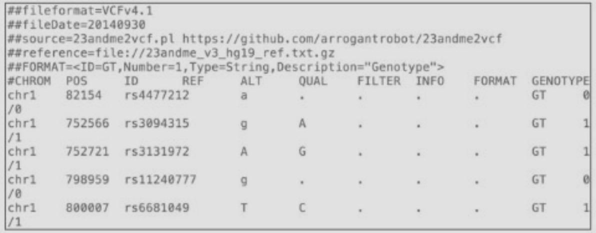
Mapped Reads (mpileup, BAM)



```
seq1 272 T 24  ,.$.....,,.,.,...,,,.,..^+. <<<+;<<<<<<<<<<<=<;<;7<&
seq1 273 T 23  ,.....,,.,.,...,,,.,..A <<<;<<<<<<<<<3<=<<<;<<+
seq1 274 T 23  ,.$....,,.,.,...,,,.,...    7<7;<;<<<<<<<<<=<;<;<<6
seq1 275 A 23  ,$....,,.,.,...,,,.,...^l.  <+;9*<<<<<<<<<=<<:;<<<<
seq1 276 G 22  ...T,,.,.,...,,,.,....  33;+<<7=7<<7<&<<1;<<6<
seq1 277 T 22  ....,,.,.,.C.,,,.,..G.  +7<;<<<<<<<&<=<<:;<<&<
seq1 278 G 23  ....,,.,.,...,,,.,....^k.   %38*<<;<7<<7<=<<<;<<<<<
seq1 279 C 23  A..T,,.,.,...,,,.,..... ;75&<<<<<<<<<=<<<9<<:<<
```

tomorrow

Variants (VCF)

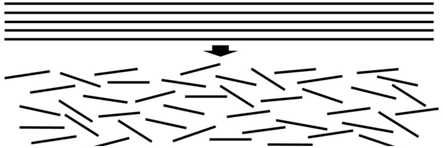


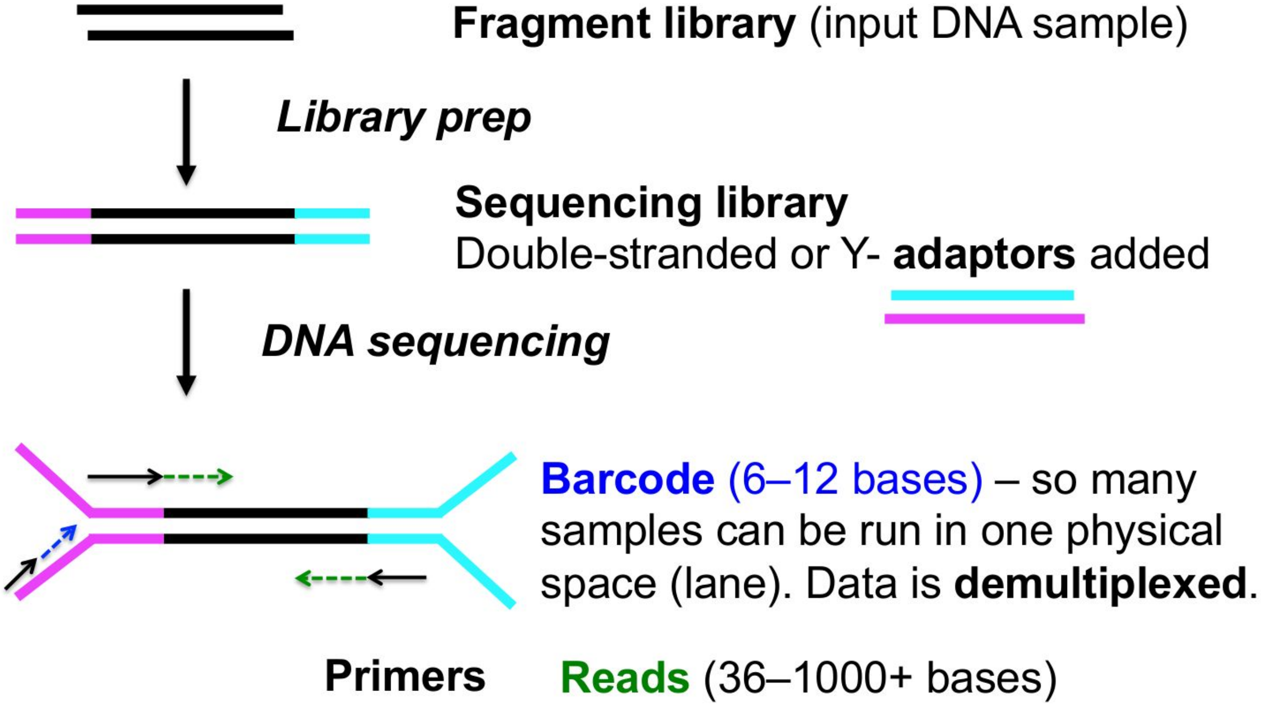
```
##fileformat=VCFv4.1
##fileDate=20140930
##source=23andme2vcf.pl https://github.com/arrogantrobot/23andme2vcf
##reference=file://23andme_v3_hg19_ref.txt.gz
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
#CHROM  POS     ID        REF       ALT     QUAL    FILTER  INFO    FORMAT  GENOTYPE
chr1    82154   rs4477212           a       .       .       .       .       GT      0
/0
chr1    752566  rs3094315           g       A       .       .       .       GT      1
/1
chr1    752721  rs3131972           A       G       .       .       .       GT      1
/1
chr1    798959  rs11240777          g       .       .       .       .       GT      0
/0
chr1    800007  rs6681049           T       C       .       .       .       GT      1
/1
```

KØBENHAVNS UNIVERSITET

Fragment library

Fragment DNA





Fragment library (input DNA sample)

Library prep

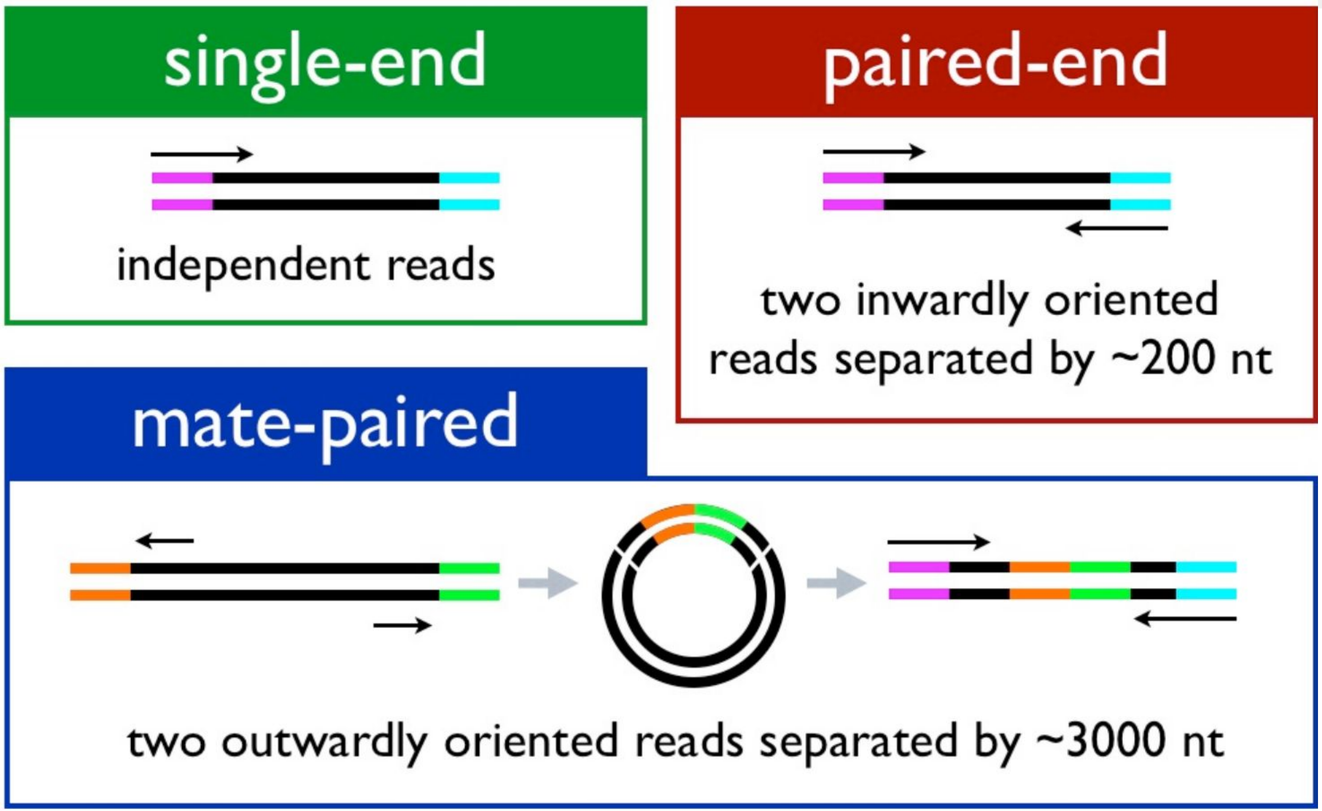
Sequencing library
Double-stranded or Y- **adaptors** added

DNA sequencing

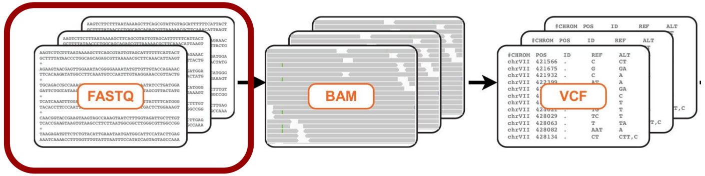
Barcode (6–12 bases) – so many samples can be run in one physical space (lane). Data is **demultiplexed**.

Primers **Reads** (36–1000+ bases)

Single or pair of fq files.


single-end
independent reads
paired-end
two inwardly oriented reads separated by ~200 nt
mate-paired
two outwardly oriented reads separated by ~3000 nt

fastQ (.fq.gz)



```
a`X_\Va\J`KaYJHG^]b\a^BBBBBBBBBBBBB   <-- quality score
@FC42BF1AAXX:6:1:5:732#0/1            <-- read ID
TGATTCTCTCGATATCCAGTCCTTAGTGNCATAGN   <-- read (bases)
+
a^_aaaa`aa`_aaa_aaa`__``_`VBBBBBBBB
@FC42BF1AAXX:6:1:5:492#0/1
AACAGTGGGAGGCTGCAGCAGGAGGATTNCTGAAN
+
ababb_abbbZbabaab^`aaTaabbaBBBBBBBB
@FC42BF1AAXX:6:1:5:480#0/1
ACCTCCTCAGAGTTCTCGAGCTCGAGAANTCTGGN
```

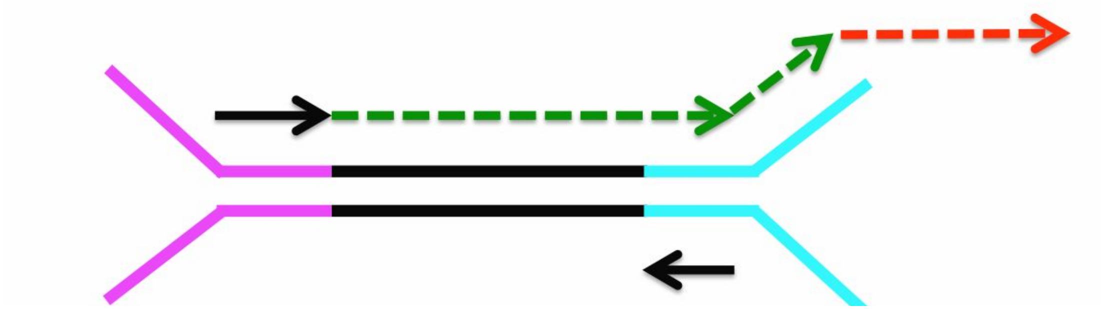
KØBENHAVNS UNIVERSITET

Selected issues with sequencing data

- **Adapter contamination of data**
  - If the DNA is too short we will sequence the adapters

Adapter contamination

If the DNA fragment is too small you will sequencing into the adapter+junk



AACAGTGGAGGCTGCAGCAGGAGGAAAAAAAAAA

Solution: Identify the problem using fastQC and trim the 3’ end of the read to remove the **adapter** + **junk (AAA…)** if needed

KØBENHAVNS UNIVERSITET

Selected issues with sequencing data

- Adapter contamination of data
  - If the DNA is too short we will sequence the adapters
- **Sequencing errors**
  - The reads by have errors

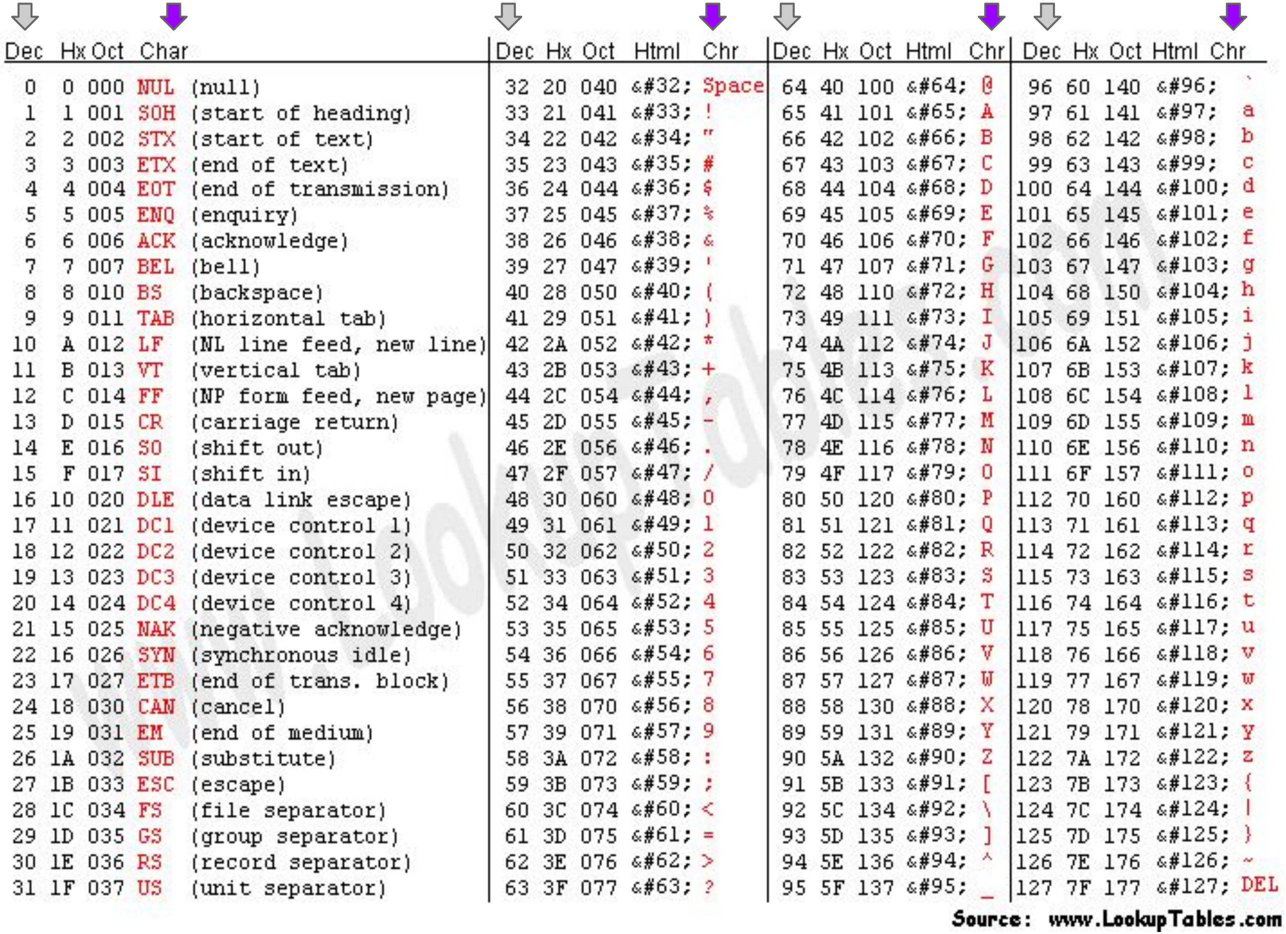
Sequencing error

```
FastQ file
a`X_\Va\J`KaYJHG^]b\a^BBBBBBBBBBBBB    <-- quality score
@FC42BF1AAXX:6:1:5:732#0/1             <-- read ID
TGATTCTCTCGATATCCAGTCCTTAGTGNCATAGN    <-- read (bases)
TGACTCTCTCGATATCAAGTCCTTAGTGNCATAGN    <-- sequenced DNA fragment
```

Solution: Translate the quality score to error rates

Identify the scale of the problem using fastQC

Use the error rates when calling genotypes



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

Source: www.LookupTables.com

www.asciitable.com/

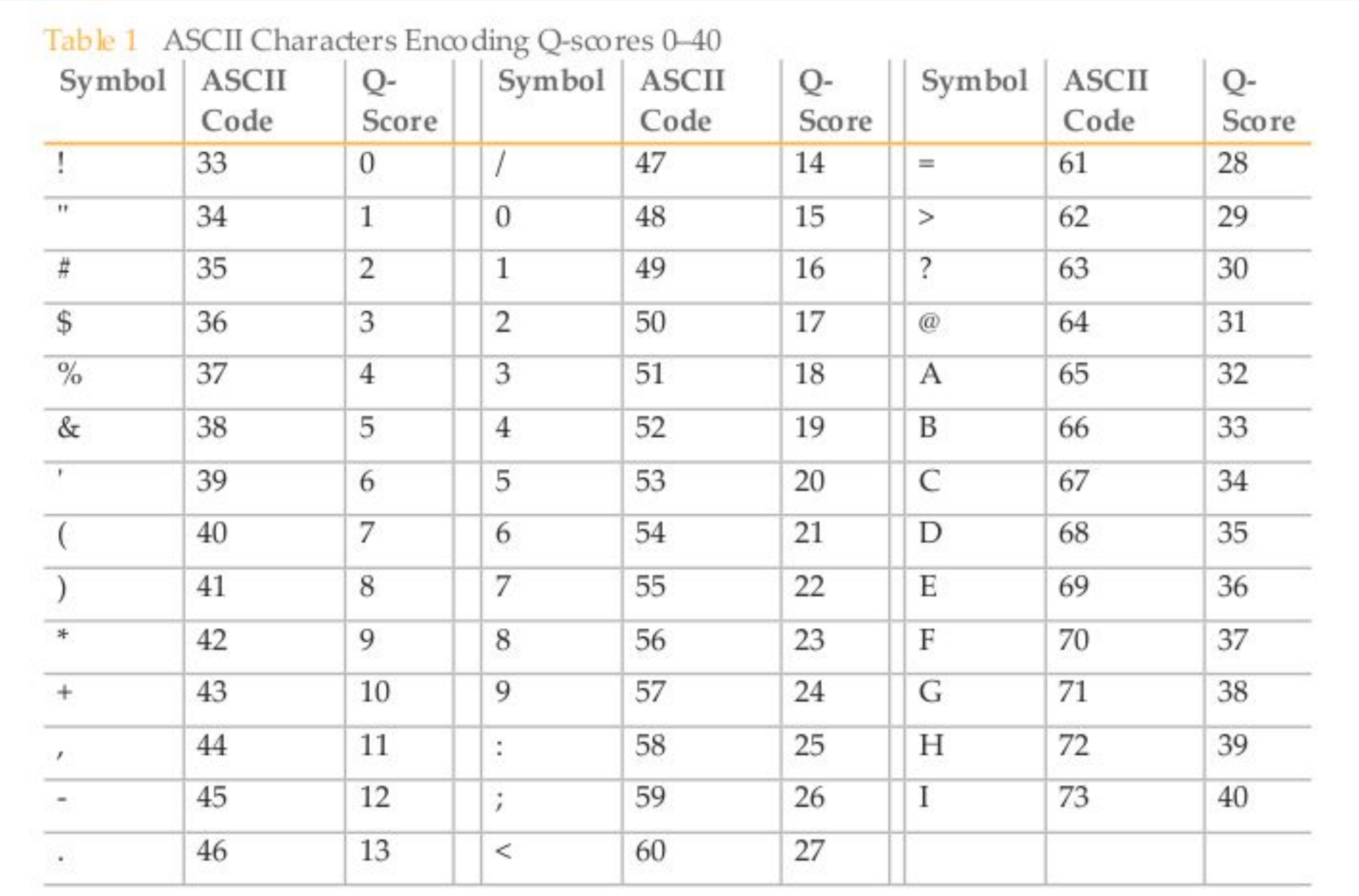


Table 1 ASCII Characters Encoding Q-scores 0–40

Symbol	ASCII Code	Q-Score	Symbol	ASCII Code	Q-Score	Symbol	ASCII Code	Q-Score
!	33	0	/	47	14	=	61	28
"	34	1	0	48	15	>	62	29
#	35	2	1	49	16	?	63	30
$	36	3	2	50	17	@	64	31
%	37	4	3	51	18	A	65	32
&	38	5	4	52	19	B	66	33
'	39	6	5	53	20	C	67	34
(	40	7	6	54	21	D	68	35
)	41	8	7	55	22	E	69	36
*	42	9	8	56	23	F	70	37
+	43	10	9	57	24	G	71	38
,	44	11	:	58	25	H	72	39
-	45	12	;	59	26	I	73	40
.	46	13	<	60	27			

quality scores/Phred scores

```
a`X_\Va\J`KaYJHG^]b\a^BBBBBBBBBBBBB    <-- quality score
```

Ascii	Dec	Qscore (Dec -33)	Error (ϵ)
+	43	10	10%
5	53	20	1%
?	63	30	0.1%
I	73	40	0.01%

Convert Qscores to seuquencing error rates

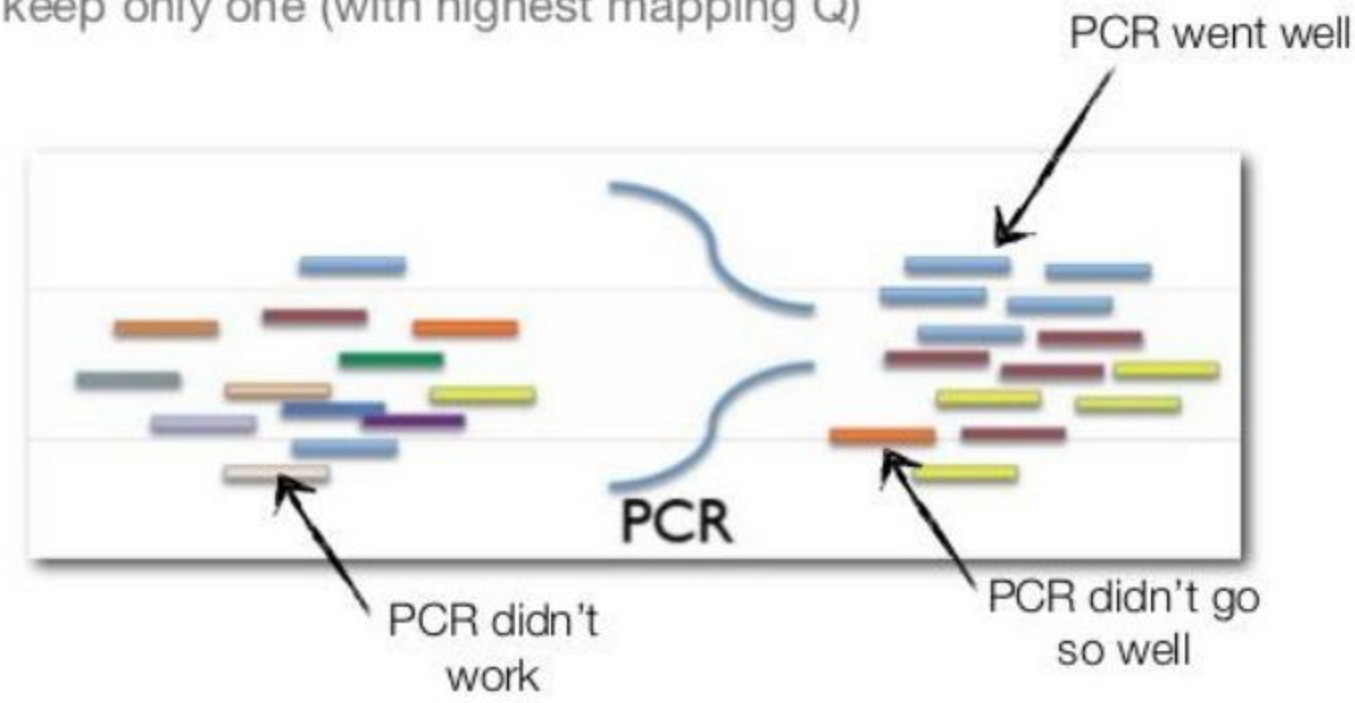
$$\text{Qscore} = -10\log_{10}(\epsilon) \quad \Leftrightarrow \quad \epsilon = 10^{-Q/10}$$

Selected issues with sequencing data

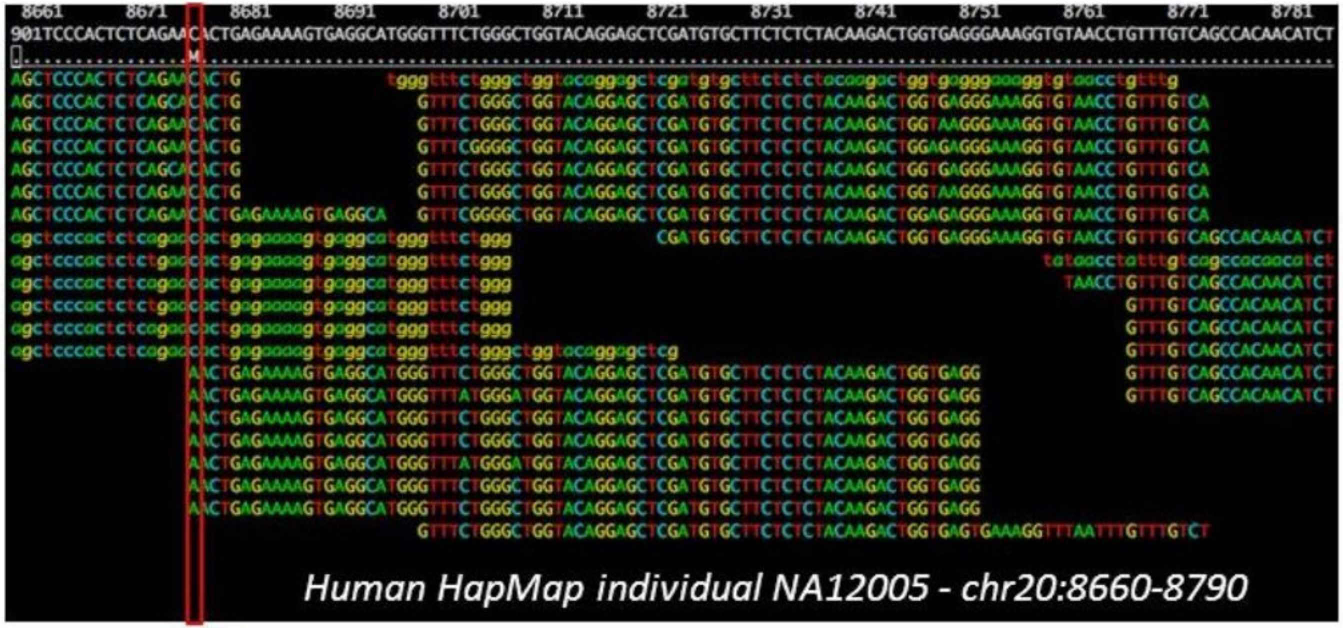
- Adapter contamination of data
  - If the DNA is too short we will sequence the adapters
- Sequencing errors
  - The reads by have errors
- **PCR or optical duplicates**
  - Reads can be duplicated ether from PCR or from the chip

Duplicated reads

> keep only one (with highest mapping Q)



Duplicated reads can cause


8661
8671
8681
8691
8701
8711
8721
8731
8741
8751
8761
8771
8781
901TCCCACTCTCAGAACACTGAGAAAAGTGAGGCATGGGTTTCTGGGCTGGTACAGGAGCTCGATGTGCTTCTCTCTACAAGACTGGTGAGGGAAAGGTGTAACCTGTTTGTCAGCCACAACATCT
M

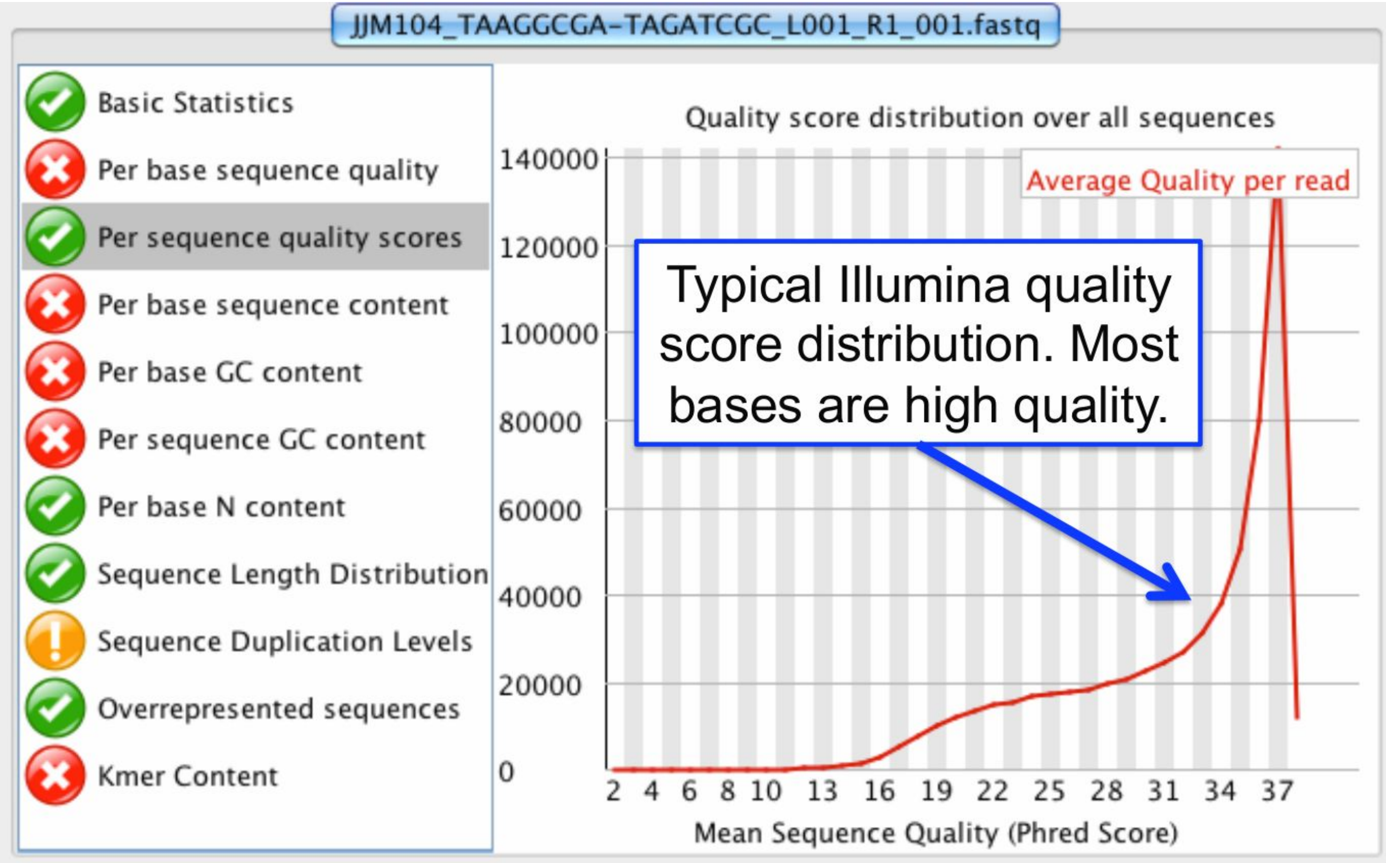
Human HapMap individual NA12005 - chr20:8660-8790

Solution: Identify the problem using fastQC

Identify the duplicated reads and remove or mark them

FAST QC

Easy to use tool for evaluating the quality of your data (fastQ or Bam files)

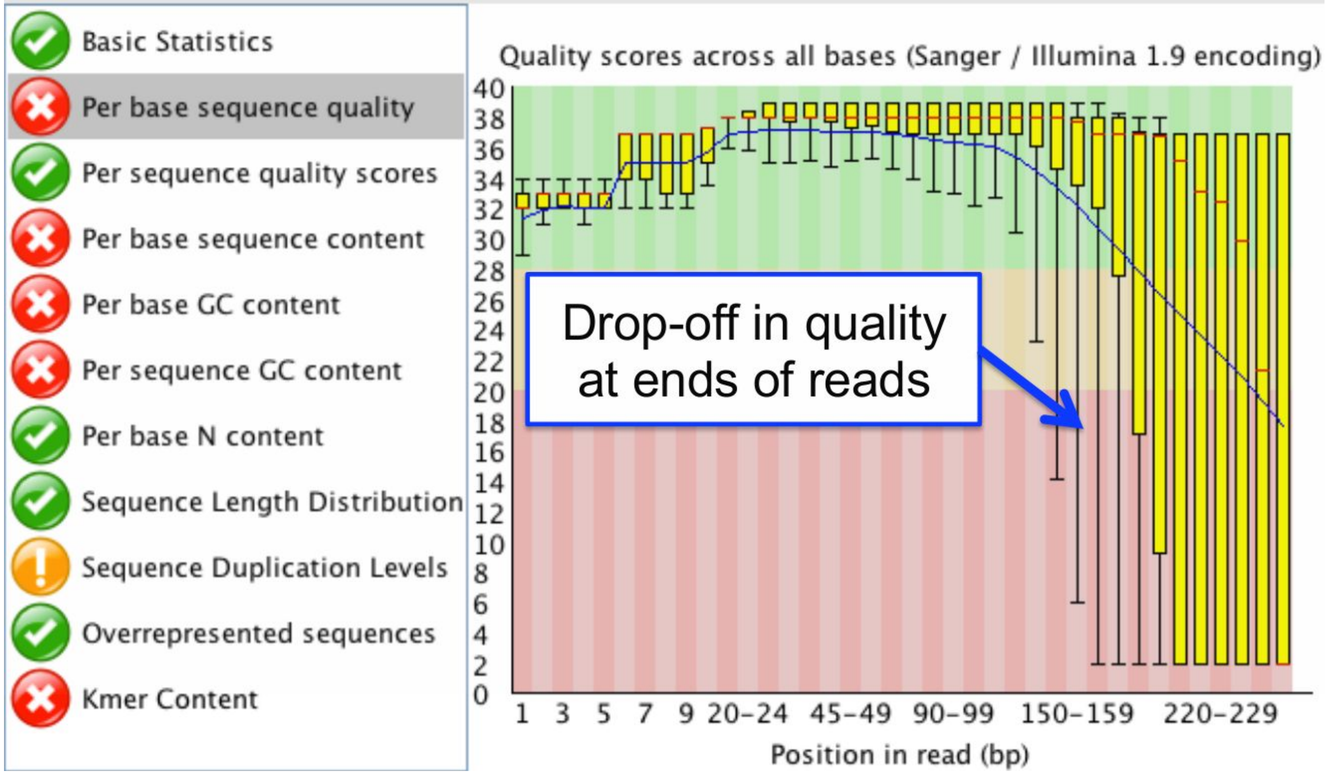


JJM104_TAAGGCGA-TAGATCGC_L001_R1_001.fastq

- ✓ Basic Statistics
- ✗ Per base sequence quality
- ✓ Per sequence quality scores
- ✗ Per base sequence content
- ✗ Per base GC content
- ✗ Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ! Sequence Duplication Levels
- ✓ Overrepresented sequences
- ✗ Kmer Content

Typical Illumina quality score distribution. Most bases are high quality.

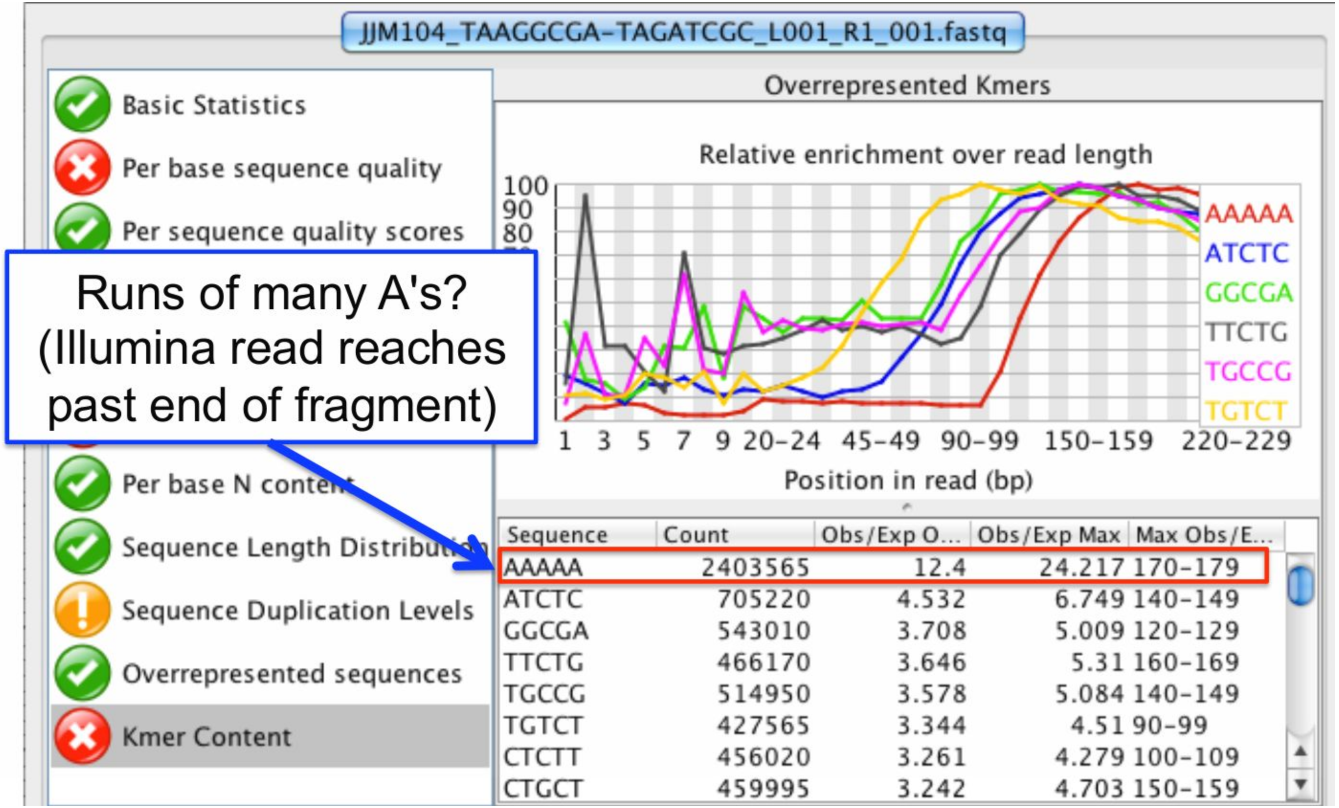
Quality for each cycle



- ✓ Basic Statistics
- ✗ Per base sequence quality
- ✓ Per sequence quality scores
- ✗ Per base sequence content
- ✗ Per base GC content
- ✗ Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ! Sequence Duplication Levels
- ✓ Overrepresented sequences
- ✗ Kmer Content

Drop-off in quality at ends of reads

Kmer/ adapter

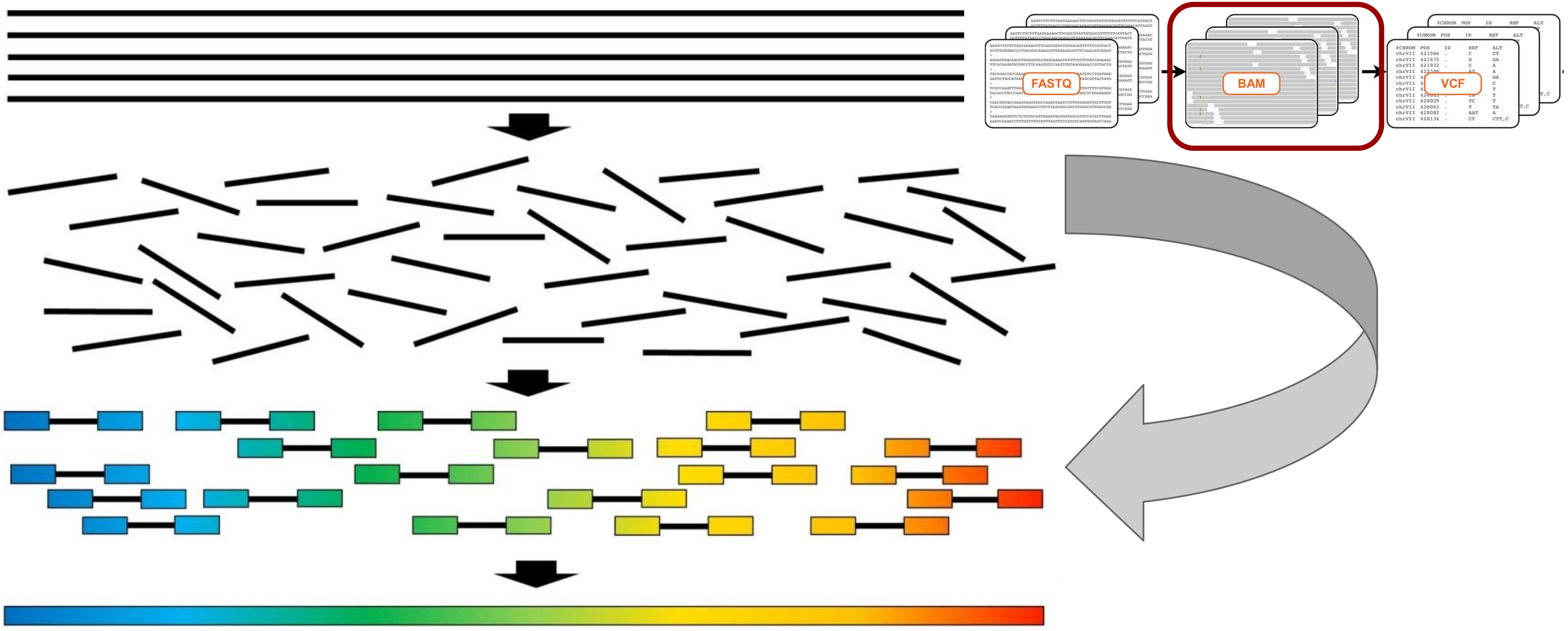


JJM104_TAAGGCGA-TAGATCGC_L001_R1_001.fastq
Basic Statistics
Per base sequence quality
Per sequence quality scores
Per base N conten
Sequence Length Distributio
Sequence Duplication Levels
Overrepresented sequences
Kmer Content
Overrepresented Kmers
Relative enrichment over read length
100
90
80
AAAAA
ATCTC
GGCGA
TTCTG
TGCCG
TGTCT
1 3 5 7 9 20-24 45-49 90-99 150-159 220-229
Position in read (bp)

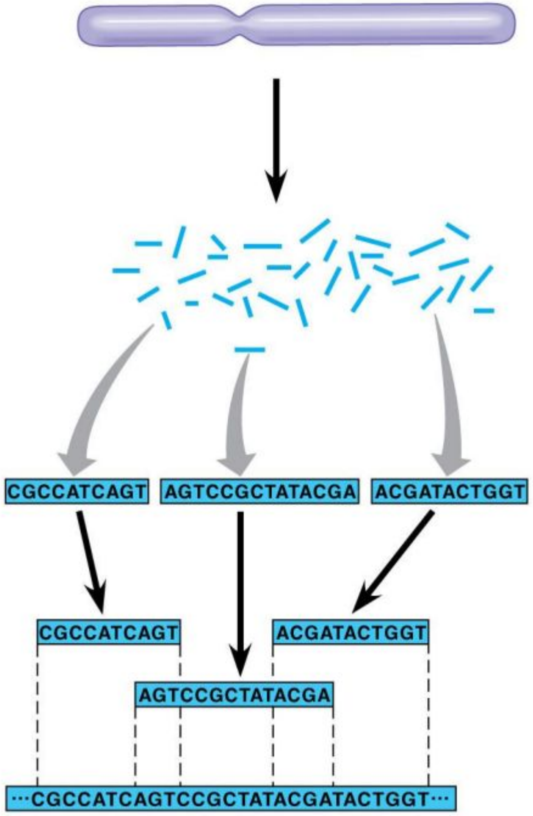
Runs of many A's?
(Illumina read reaches past end of fragment)

Sequence	Count	Obs/Exp O...	Obs/Exp Max	Max Obs/E...
AAAAA	2403565	12.4	24.217	170-179
ATCTC	705220	4.532	6.749	140-149
GGCGA	543010	3.708	5.009	120-129
TTCTG	466170	3.646	5.31	160-169
TGCCG	514950	3.578	5.084	140-149
TGTCT	427565	3.344	4.51	90-99
CTCTT	456020	3.261	4.279	100-109
CTGCT	459995	3.242	4.703	150-159

Mapping - alignment of reads



Alignment and mapping

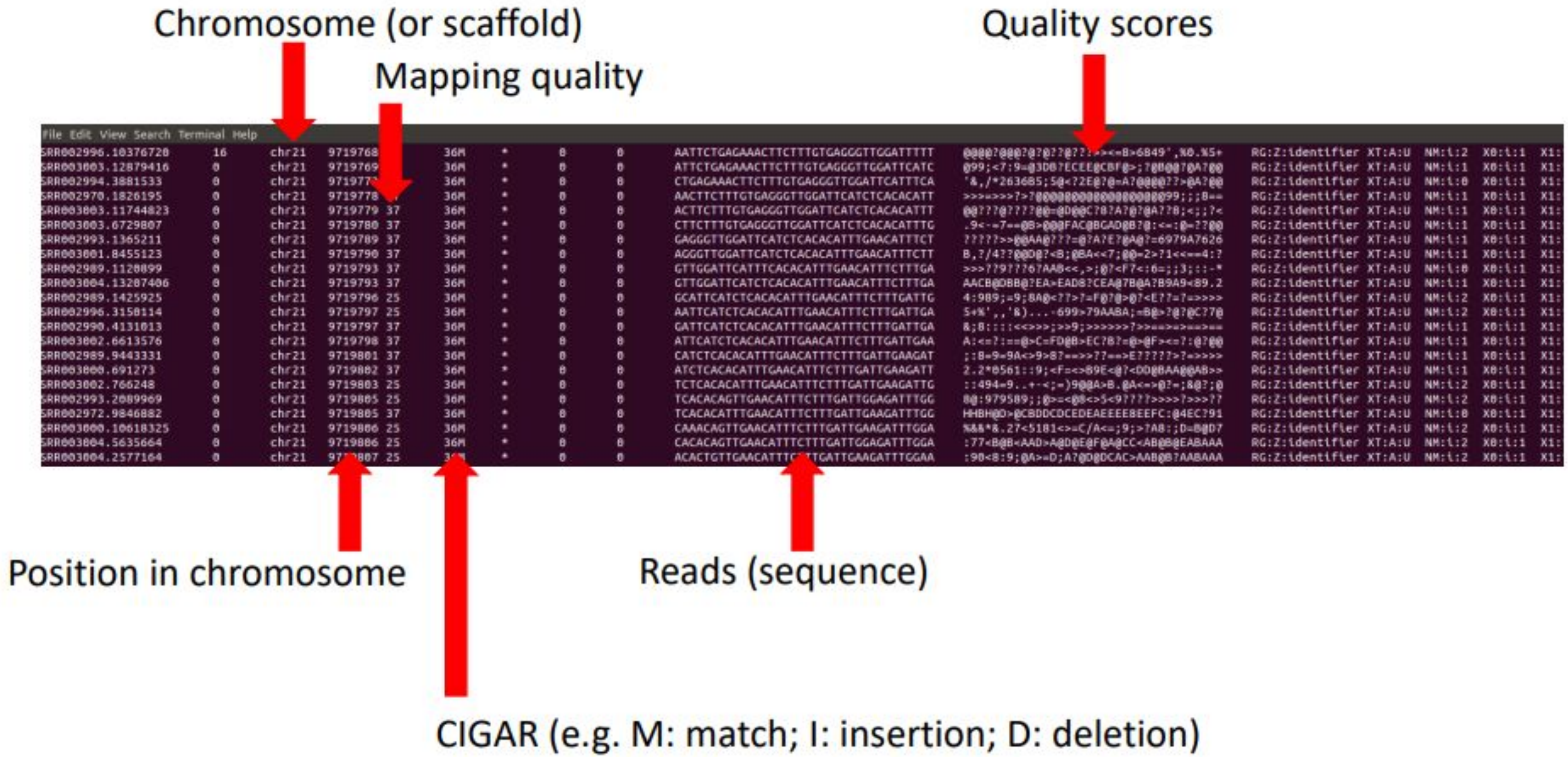


.bam/.sam file

```
reads    TTTGTTCTTTCTTTCTCTCTAGTCTTCTT ...
Qscore   NVFVN]^]`^_]^^U]]`][_VS[_^Z]_ ...
Position chr4 53351385
Mismatch 2 (in cigar string)
strand   +
mapQ     30
Mate     mapped chr4 53351145
Alt map  chr2 15331145 with 2 mismatch
```

KØBENHAVNS UNIVERSITET

bam/sam file



```
SRR002996.10376720      16      chr21   9719768          36M     *       0       0       AATTCTGAGAAACTTCTTTGTGAGGGTTGGATTTTT    @@@@?@@@?@?@??@??  ><=B>6849',%0.%5+    RG:Z:identifier XT:A:U  NM:i:2  X0:i:1  X1:
SRR003003.12879416      0       chr21   9719769          36M     *       0       0       ATTCTGAGAAACTTCTTTGTGAGGGTTGGATTCATC    @99;<7:9=@3DB?ECEE@CBF@>;?@B@@?@A?@@    RG:Z:identifier XT:A:U  NM:i:1  X0:i:1  X1:
SRR002994.3881533       0       chr21   971977           36M     *       0       0       CTGAGAAACTTCTTTGTGAGGGTTGGATTCATTTCA    '&,/*263685;5@<?2E@?@=A?@@@@??>@A?@@    RG:Z:identifier XT:A:U  NM:i:0  X0:i:1  X1:
SRR002970.1826195       0       chr21   9719778          36M     *       0       0       AACTTCTTTGTGAGGGTTGGATTCATCTCACACATT    >>>=>>>?>?@@@@@@@@@@@@@@@@@@99;;;8==    RG:Z:identifier XT:A:U  NM:i:1  X0:i:1  X1:
SRR003003.11744823      0       chr21   9719779 37       36M     *       0       0       ACTTCTTTGTGAGGGTTGGATTCATCTCACACATTT    @@???@????@@=@D@@C?8?A?@?@A??8;<;;?<    RG:Z:identifier XT:A:U  NM:i:1  X0:i:1  X1:
SRR003003.6729807       0       chr21   9719780 37       36M     *       0       0       CTTCTTTGTGAGGGTTGGATTCATCTCACACATTTG    .9<-=7==@B>@@@FAC@BGAD@B?@:<=:@=??@@    RG:Z:identifier XT:A:U  NM:i:1  X0:i:1  X1:
SRR002993.1365211       0       chr21   9719789 37       36M     *       0       0       GAGGGTTGGATTCATCTCACACATTTGAACATTTCT    ?????>>@@AA@???=@?A?E?@A@?=6979A7626    RG:Z:identifier XT:A:U  NM:i:1  X0:i:1  X1:
SRR003001.8455123       0       chr21   9719790 37       36M     *       0       0       AGGGTTGGATTCATCTCACACATTTGAACATTTCTT    B,?/4??@@D@?<B;@BA<<7;@@=2>?1<<==4:?    RG:Z:identifier XT:A:U  NM:i:1  X0:i:1  X1:
SRR002989.1128899       0       chr21   9719793 37       36M     *       0       0       GTTGGATTCATTTCACACATTTGAACATTTCTTTGA    >>>??9???6?AA8<<,>;@?<F7<:6=;;3;::-*    RG:Z:identifier XT:A:U  NM:i:0  X0:i:1  X1:
SRR003004.13287406      0       chr21   9719793 37       36M     *       0       0       GTTGGATTCATCTCACACATTTGAACATTTCTTTGA    AACB@DBB@?EA>EAD8?CEA@7B@A?B9A9<89.2    RG:Z:identifier XT:A:U  NM:i:1  X0:i:1  X1:
SRR002989.1425925       0       chr21   9719796 25       36M     *       0       0       GCATTCATCTCACACATTTGAACATTTCTTTGATTG    4:989;=9;8A@<??>?=F@?@>@?<E??=?=>>>>    RG:Z:identifier XT:A:U  NM:i:2  X0:i:1  X1:
SRR002996.3158114       0       chr21   9719797 25       36M     *       0       0       AATTCATCTCACACATTTGAACATTTCTTTGATTGA    5+%',,'&)...-699>79AABA;=B@>?@?@C?7@    RG:Z:identifier XT:A:U  NM:i:2  X0:i:1  X1:
SRR002990.4131013       0       chr21   9719797 37       36M     *       0       0       GATTCATCTCACACATTTGAACATTTCTTTGATTGA    &;8::::<<>>>;>>9;>>>>>>?>>==>=>==>==    RG:Z:identifier XT:A:U  NM:i:1  X0:i:1  X1:
SRR003002.6613576       0       chr21   9719798 37       36M     *       0       0       ATTCATCTCACACATTTGAACATTTCTTTGATTGAA    A:<=?:==@>C=FD@B>EC?B?=@>@F><=?:@?@@    RG:Z:identifier XT:A:U  NM:i:1  X0:i:1  X1:
SRR002989.9443331       0       chr21   9719801 37       36M     *       0       0       CATCTCACACATTTGAACATTTCTTTGATTGAAGAT    ;:8=9=9A<>9>8?==>>??==>E?????>?=>>>>    RG:Z:identifier XT:A:U  NM:i:1  X0:i:1  X1:
SRR003000.691273        0       chr21   9719802 37       36M     *       0       0       ATCTCACACATTTGAACATTTCTTTGATTGAAGATT    2.2*0561::9;<F=<>89E<@?<DD@BAA@@AB>>    RG:Z:identifier XT:A:U  NM:i:1  X0:i:1  X1:
SRR003002.766248        0       chr21   9719803 25       36M     *       0       0       TCTCACACATTTGAACATTTCTTTGATTGAAGATTG    ::494=9..+-<;=)9@@A>B.@A<=>@?=;8@?;@    RG:Z:identifier XT:A:U  NM:i:2  X0:i:1  X1:
SRR002993.2089969       0       chr21   9719805 25       36M     *       0       0       TCACACAGTTGAACATTTCTTTGATTGGAGATTTGG    8@:979589;;@>=<@8<>5<9????>>>>?>>>??    RG:Z:identifier XT:A:U  NM:i:2  X0:i:1  X1:
SRR002972.9846882       0       chr21   9719805 37       36M     *       0       0       TCACACATTTGAACATTTCTTTGATTGAAGATTTGG    HHBH@D>@CBDDCDCEDEAEEEE8EEFC:@4EC?91    RG:Z:identifier XT:A:U  NM:i:0  X0:i:1  X1:
SRR003000.10618325      0       chr21   9719806 25       36M     *       0       0       CAAACAGTTGAACATTTCTTTGATTGAAGATTTGGA    %&&*&.27<5181<>=C/A<=;9;>?A8:;D=8@D7    RG:Z:identifier XT:A:U  NM:i:2  X0:i:1  X1:
SRR003004.5635664       0       chr21   9719806 25       36M     *       0       0       CACACAGTTGAACATTTCTTTGATTGGAGATTTGGA    :77<B@B<AAD>A@D@E@F@A@CC<AB@B@EABAAA    RG:Z:identifier XT:A:U  NM:i:2  X0:i:1  X1:
SRR003004.2577164       0       chr21   97  807 25       3       *       0       0       ACACTGTTGAACATTTC  TTGATTGAAGATTTGGAA    :98<8:9;@A>=D;A7@D@DCAC>AAB@B?AABAAA    RG:Z:identifier XT:A:U  NM:i:2  X0:i:1  X1:
```

Chromosome (or scaffold)

Mapping quality

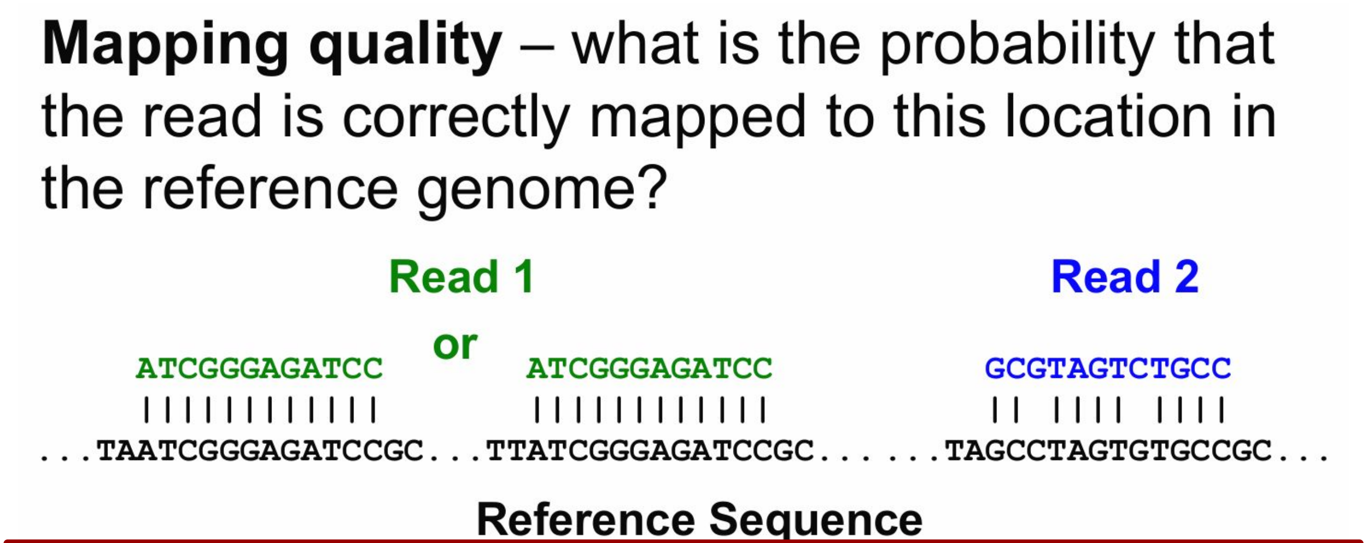
Quality scores

Position in chromosome

Reads (sequence)

CIGAR (e.g. M: match; I: insertion; D: deletion)

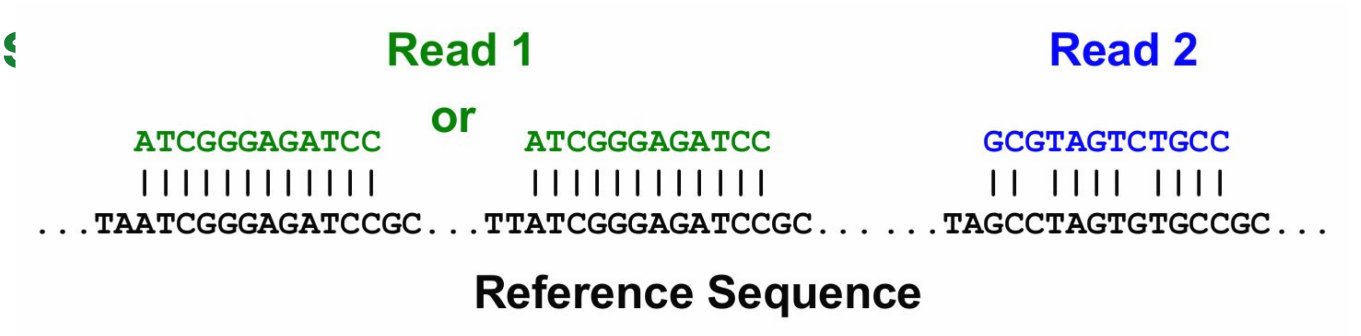
Mapping quality



Mapping quality – what is the probability that the read is correctly mapped to this location in the reference genome?

Read 1 can be mapped two places on the genome while **Read 2** only maps to one

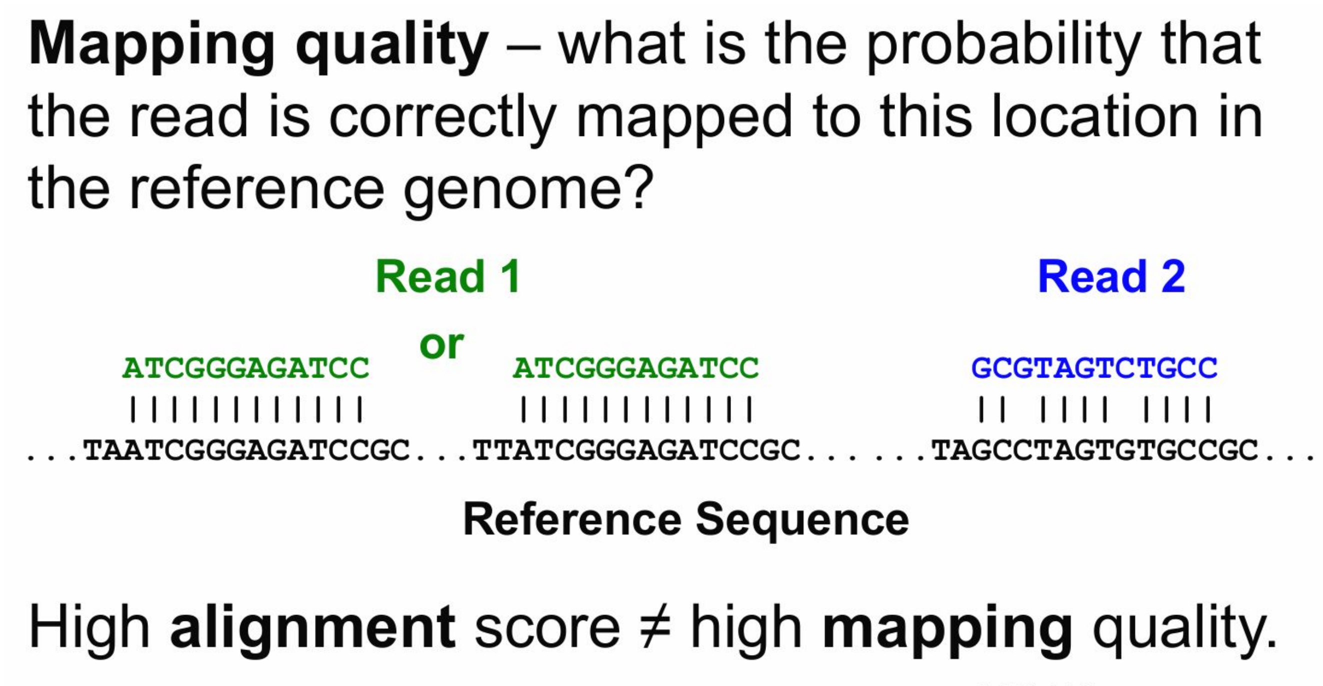
- **Which of the two reads has the highest mapping quality?**



Which read will have the highest mapping quality

ⓘ Click **Present with Slido** or install our Chrome extension to activate this poll while presenting.

Mapping quality



Mapping quality – what is the probability that the read is correctly mapped to this location in the reference genome?

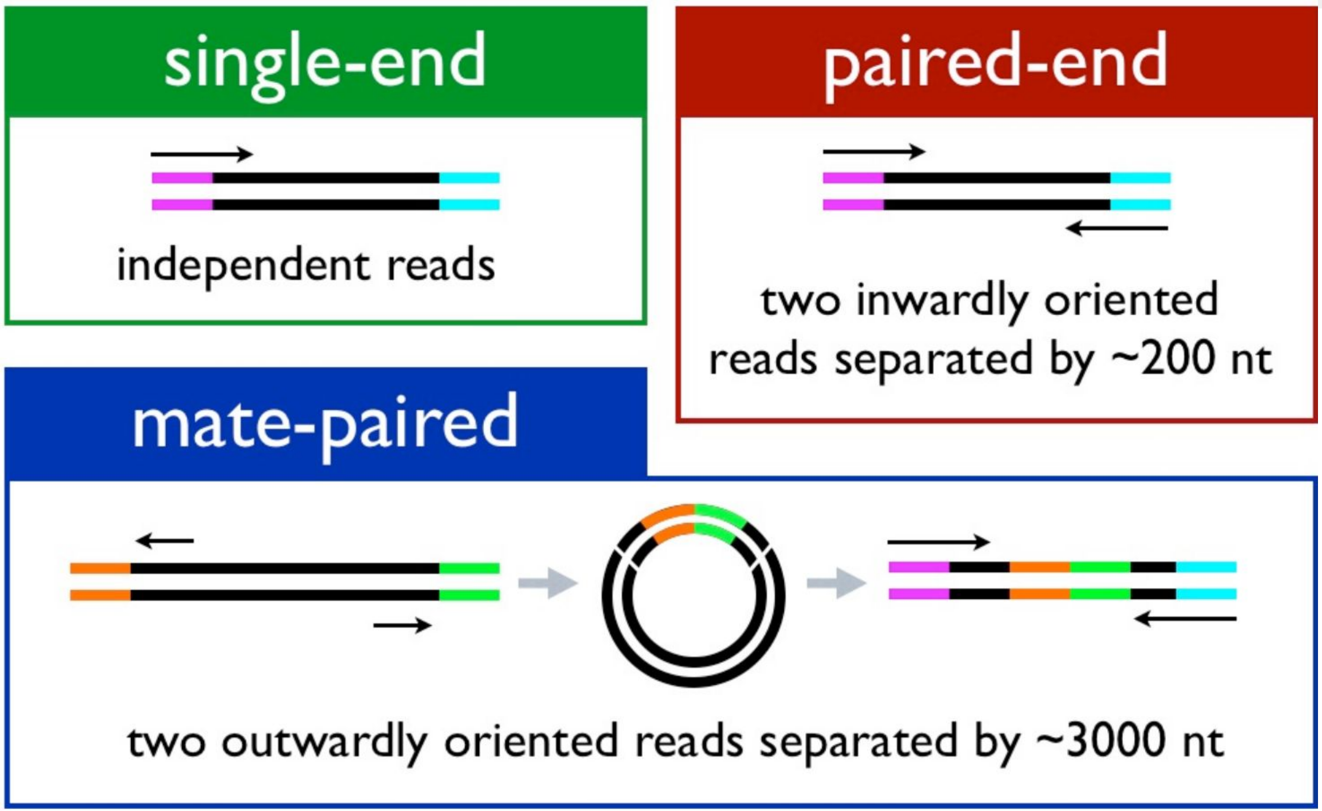
```
                         Read 1                                  Read 2

                           or
     ATCGGGAGATCC             ATCGGGAGATCC                  GCGTAGTCTGCC
     ||||||||||||             ||||||||||||                  ||  ||||  ||||
...TAATCGGGAGATCCGC...TTATCGGGAGATCCGC... ...TAGCCTAGTGTGCCGC...
```

Reference Sequence

High **alignment** score ≠ high **mapping** quality.

Why use paired end sequencing?

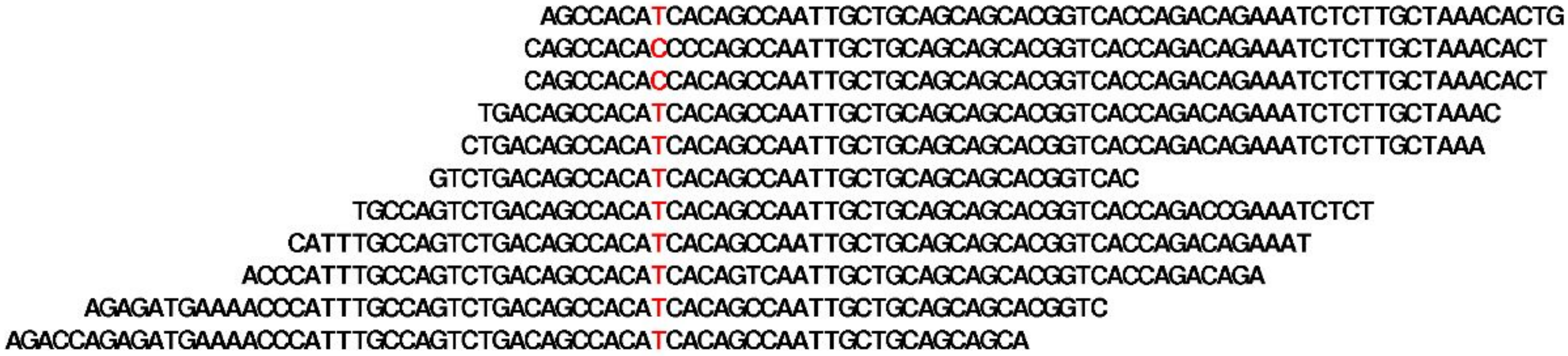

single-end
independent reads
paired-end
two inwardly oriented reads separated by ~200 nt
mate-paired
two outwardly oriented reads separated by ~3000 nt

Sequencing Depth

Sequencing depth is the number of reads covering a position

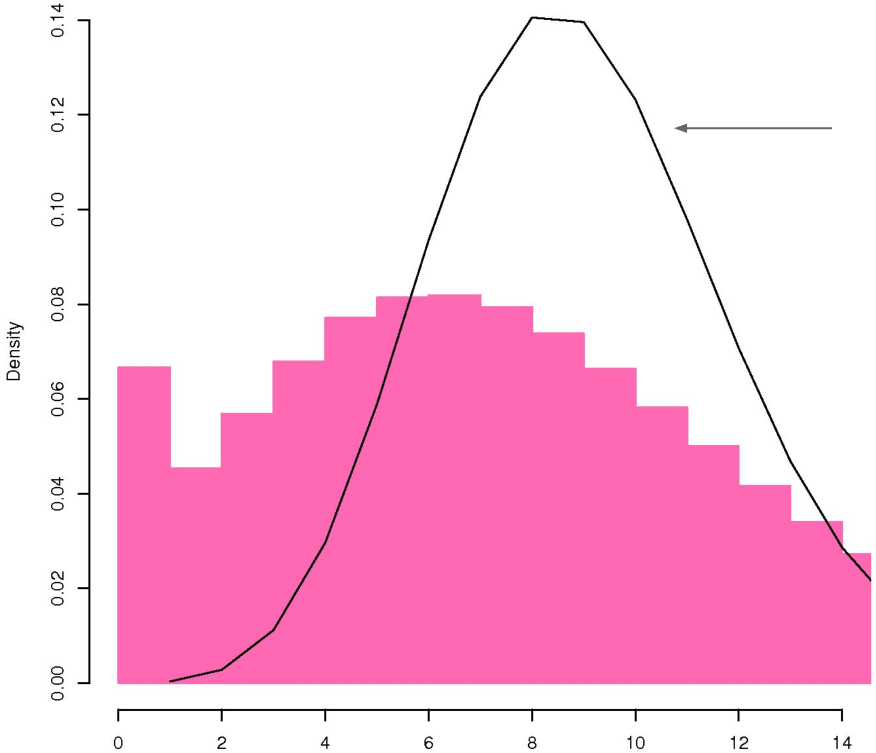
Average depth is often written as X e.g. **15X** sequencing

Coverage = depth or the fraction of genome with data



```
                                  AGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCTCTTGCTAAACACTG
                                 CAGCCACACCCCAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCTCTTGCTAAACACT
                                 CAGCCACACCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCTCTTGCTAAACACT
                              TGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCTCTTGCTAAAC
                             CTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCTCTTGCTAAA
                           GTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCAC
                      TGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACCGAAATCTCT
                  CATTTGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAAT
               ACCCATTTGCCAGTCTGACAGCCACATCACAGTCAATTGCTGCAGCAGCACGGTCACCAGACAGA
     AGAGATGAAAACCCATTTGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTC
AGACCAGAGATGAAAACCCATTTGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCA
```

Depth distribution

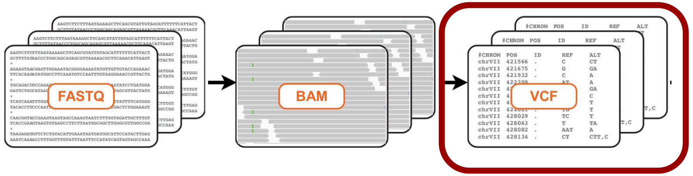


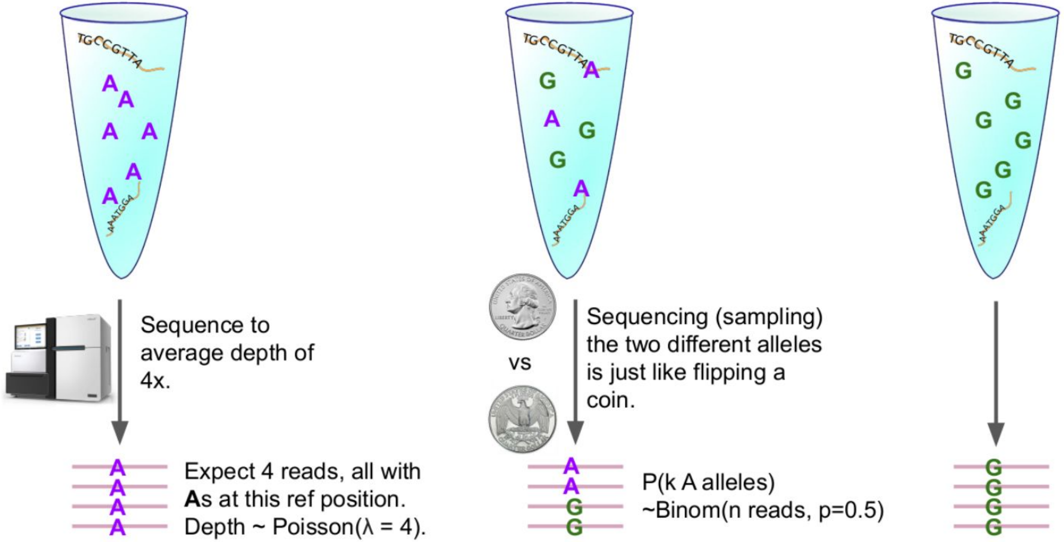
Sequencing depth

Theoretical distribution (Poisson) for 8X avg. depth if reads mapped perfectly and there was no bias

Why don’t we observe genotype

Each allele is sequenced separately and alleles are sampled with replacement





Sequence to average depth of 4x.

Expect 4 reads, all with As at this ref position. Depth ~ Poisson(λ = 4).

Sequencing (sampling) the two different alleles is just like flipping a coin.

P(k A alleles) ~Binom(n reads, p=0.5)

Why don’t we observe genotype

Question: Assuming an error rate of 1%
Is the individual heterozygous C/T?

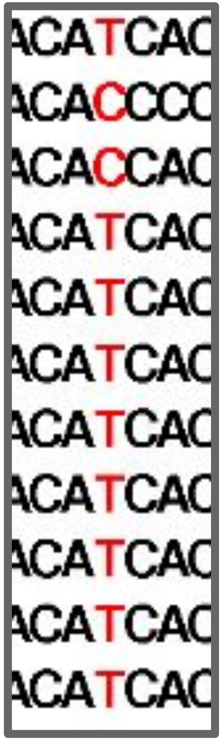
```
                                  AGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCTCTTGCTAAACACTG
                                 CAGCCACACCCCAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCTCTTGCTAAACACT
                                 CAGCCACACCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCTCTTGCTAAACACT
                              TGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCTCTTGCTAAAC
                             CTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCTCTTGCTAAA
                           GTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCAC
                      TGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACCGAAATCTCT
                  CATTTGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAAT
               ACCCATTTGCCAGTCTGACAGCCACATCACAGTCAATTGCTGCAGCAGCACGGTCACCAGACAGA
     AGAGATGAAAACCCATTTGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTC
AGACCAGAGATGAAAACCCATTTGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCA
```

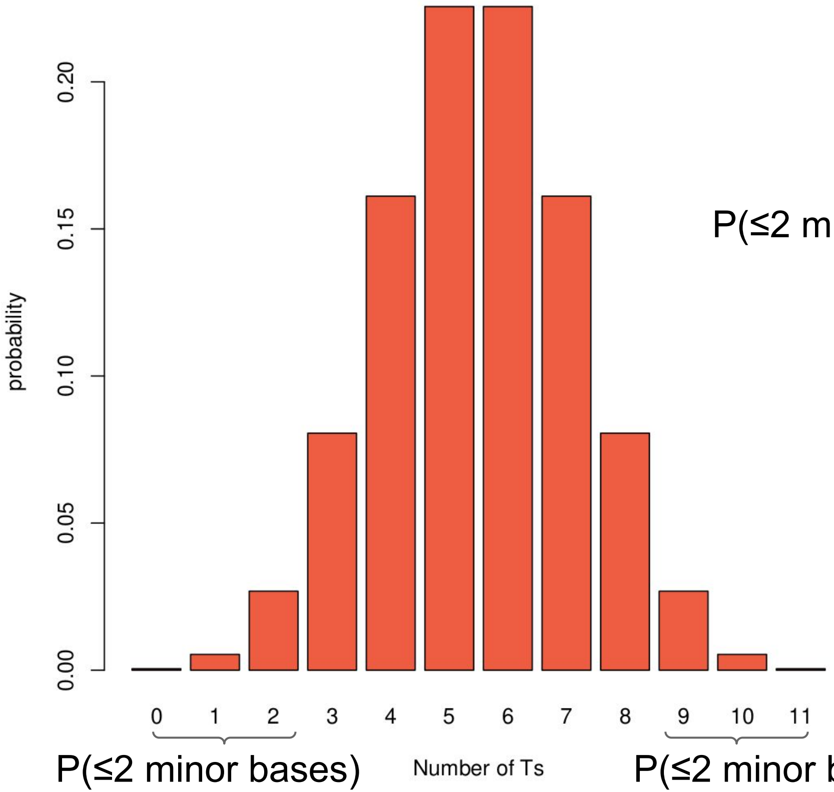
slido

Genotype is more likely

ⓘ Click **Present with Slido** or install our Chrome extension to activate this poll while presenting.

Assuming heterozygous (C/T)



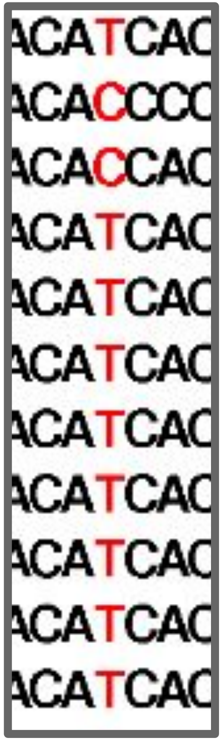


P(≤2 minor bases | T/A) = 0.065

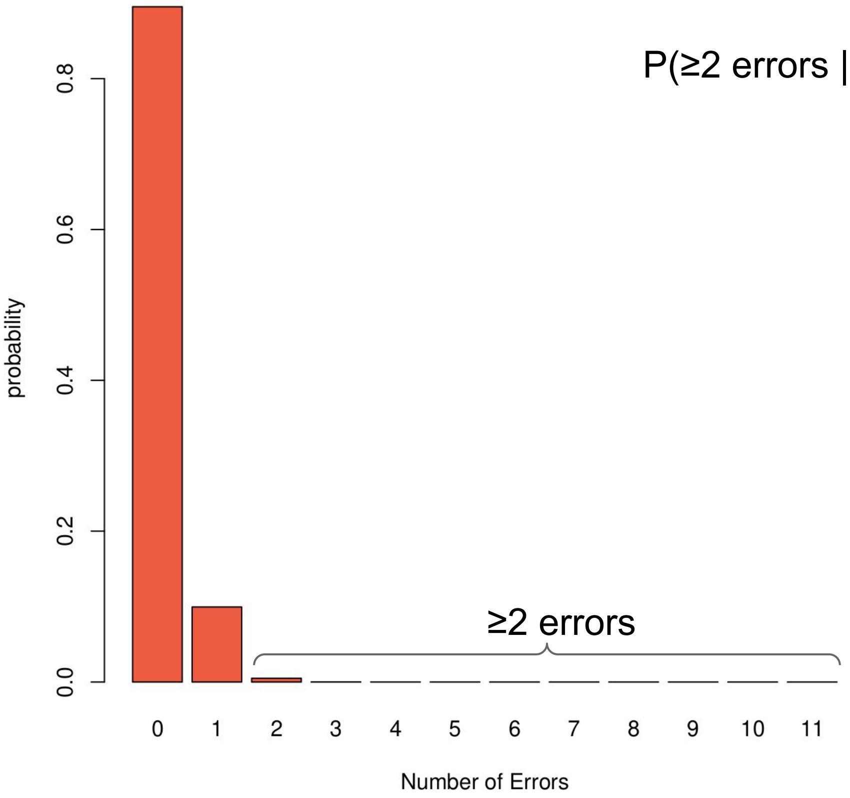
P(≤2 minor bases)

P(≤2 minor bases)

Assuming homozygous (T/T)



ACATCAC
ACACCCC
ACACCAC
ACATCAC
ACATCAC
ACATCAC
ACATCAC
ACATCAC
ACATCAC
ACATCAC
ACATCAC

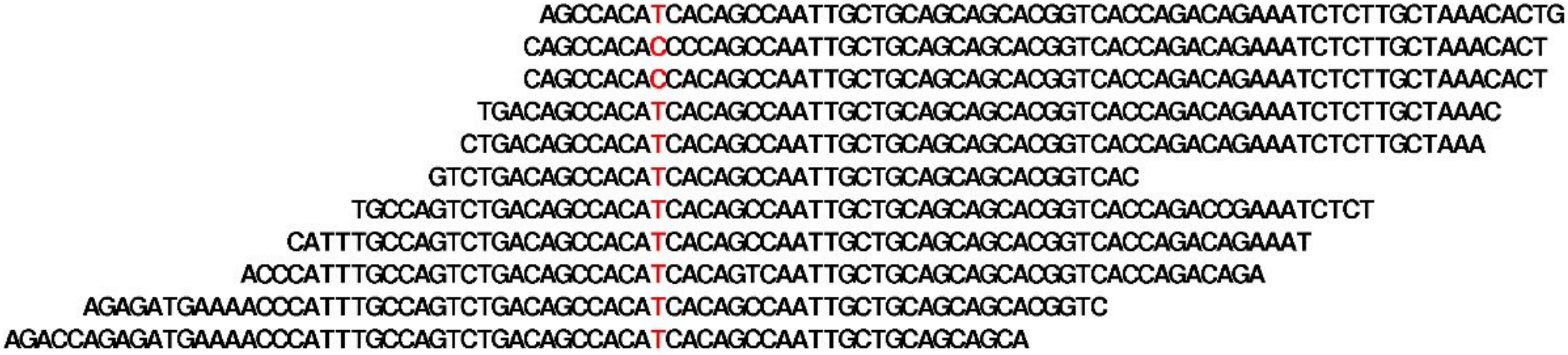


P(≥2 errors | T/T) = 0.0052

Why don’t we observe genotype

P(≥2 errors | T/T) = 0.0052　　P(≤2 minor bases | T/C) = 0.065

Question: Assuming an error rate of 1%
Is the individual heterozygous C/T?

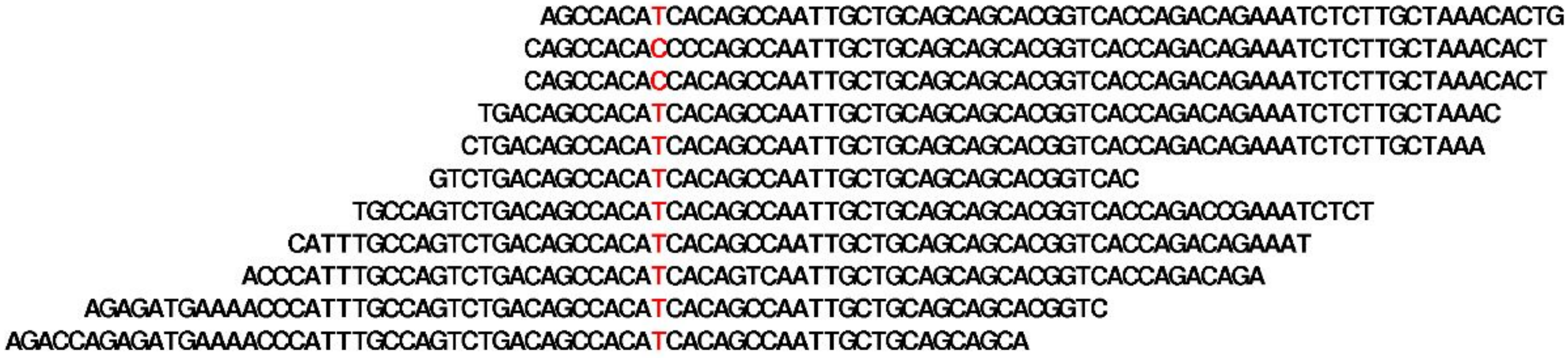


```
                                  AGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCTCTTGCTAAACACTG
                                 CAGCCACACCCCAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCTCTTGCTAAACACT
                                 CAGCCACACCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCTCTTGCTAAACACT
                              TGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCTCTTGCTAAAC
                             CTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCTCTTGCTAAA
                           GTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCAC
                      TGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACCGAAATCTCT
                  CATTTGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAAT
               ACCCATTTGCCAGTCTGACAGCCACATCACAGTCAATTGCTGCAGCAGCACGGTCACCAGACAGA
     AGAGATGAAAACCCATTTGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTC
AGACCAGAGATGAAAACCCATTTGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCA
```

Why don’t we observe genotype

P(≥2 errors | T/T) = 0052   P(≤2 minor bases | T/C) = 0.065   Heterozygosity is 0.1%

Question: Assuming an error rate of 1%
Is the individual heterozygous C/T?



```
                                  AGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCTCTTGCTAAACACTG
                                 CAGCCACACCCCAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCTCTTGCTAAACACT
                                 CAGCCACACCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCTCTTGCTAAACACT
                              TGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCTCTTGCTAAAC
                             CTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAATCTCTTGCTAAA
                           GTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCAC
                      TGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACCGAAATCTCT
                  CATTTGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTCACCAGACAGAAAT
               ACCCATTTGCCAGTCTGACAGCCACATCACAGTCAATTGCTGCAGCAGCACGGTCACCAGACAGA
     AGAGATGAAAACCCATTTGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCACGGTC
AGACCAGAGATGAAAACCCATTTGCCAGTCTGACAGCCACATCACAGCCAATTGCTGCAGCAGCA
```

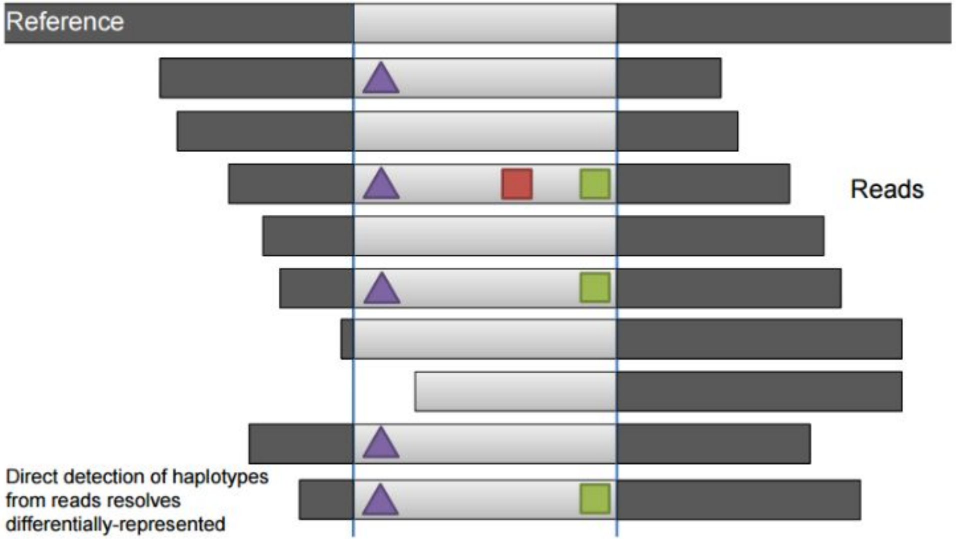
Multiple variants on the same reads

How many variants?

- Assembly-based caller (as in GATK)

Local re-alignment around putative variants; better resolution for INDELs detection.

- Haplotype-based caller (as in freebayes)



Direct detection of haplotypes from reads resolves differentially-represented

Figure from Erik Garrison

Time for exercises

Go to

popgen.dk/popgen24github

KØBENHAVNS UNIVERSITET