Intro to NGS
Data Processing and Formats

Tyler Linderoth
Copenhagen Popgen Summer course 2022

With some poached materials from Anders Albrechtsen

Some motivations for Next Generation Sequencing, NGS.

- More data!
- Efficient for detecting and characterizing **structural variation**.
- Genome-wide association studies (**GWAS**).
- **Landscape of genetic variability** E.g. How important are average levels of genetic diversity to population viability compared to how it's distributed along the genome.
- **Demography** Ability to infer different genealogical histories along the genome (e.g. PSMC) and summarize variation across the entire genome (e.g. SFS) has greatly enhanced our ability to reconstruct population histories.
- **Role of hybridization in evolution**. NGS has greatly enhanced our view of how species engage one another to influence biodiversity and evolution.
- **Genetic monitoring** through eDNA and metabarcoding.
- Effective for sequencing degraded **historic and ancient DNA**.

What’s your favorite flavor?

- Whole genome sequencing (WGS)
  - Short read (~100-150 bp reads)
  - Long read (>10 kb long reads)

- Reduced representation targets a fraction of the genome.
  - Targeted sequence capture
  - Restriction-associated digest (RAD)
    - single-digest
    - double-digest
    - “Rapture”

Whole genome sequencing (WGS)

Pros:

- Good for discovering both single nucleotide polymorphisms (SNP) and structural variation.
- Better ability to finely identify causal variants in association studies and divergence scans.
- Minimal design and library preparation is easier.
- More contiguity (more LD-based methods on the table).

Cons:

- Require’s a reference genome (for most analyses), which can be difficult to generate depending on the organism.
- More sequencing (i.e. cost) to achieve a desired depth-of-coverage.

Reduced representation

Pros:

- Can increase sample sizes at a fixed budget and depth-of-coverage, which can provide more statistical power and accurate inference.

- *de novo* reference assembly and mapping can be easier since the data has often been pre-screened for “well behaved” regions of the genome.

Cons:

- Gamble on being able to identify particular loci of interest through linkage-disequilibrium (LD).

- Capture kits can be expensive (7,000 USD for a 4 reaction kit), *but* still may be cheaper than additional sequencing required for WGS.

- Can require design.

- RAD is anonymous (you may not be able to find from where in the genome your data are derived).

A typical NGS roadmap

Phase 1. Wetlab data generation.

1) DNA extraction
2) library construction
3) Sequencing

Phase 2. Mapping / Alignment

1) FASTQ quality control (FastQC, cutadapt, trimmomatic, Trim Galore, PEAR, FLASH, super_deduper).
2) Assembly
3) Mapping (bwa, Bowtie, Novoalign).

Phase 3. Variant discovery

1) Site-level quality control
2) Call genotypes and SNPs/INDELS (ANGSD, bcftools, GATK, freeBayes)

Phase 4. Characterize genetic variation and make biological inference

1) Estimate allele frequencies (ANGSD)
2) Population structure, demography, selection.

A typical NGS roadmap

Phase 1. Wetlab data generation.

1) DNA extraction
2) library construction
3) Sequencing

Phase 2. Mapping / Alignment

1) FASTQ quality control (FastQC, cutadapt, trimmomatic, Trim Galore, PEAR, FLASH, super_deduper).
2) Assembly
3) Mapping (bwa, Bowtie, Novoalign).

Phase 3. Variant discovery

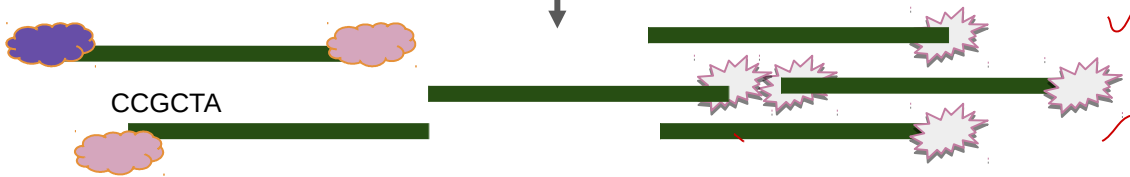
1) Site-level quality control
2) Call genotypes and SNPs/INDELS (ANGSD, bcftools, GATK, freeBayes)

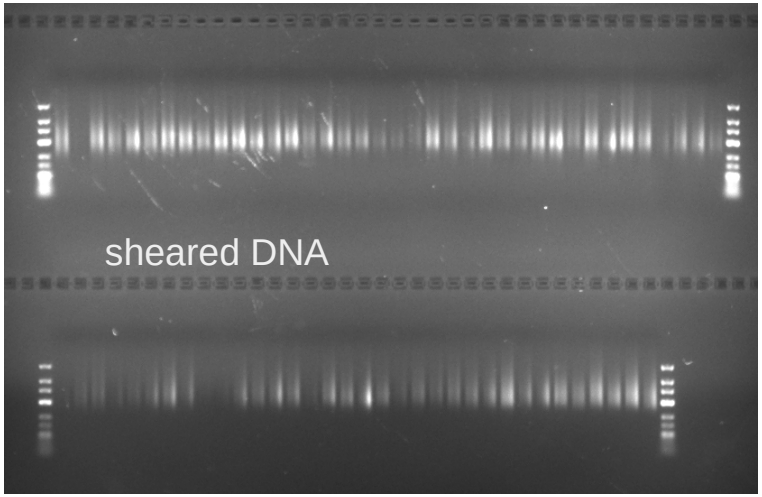
Phase 4. Characterize genetic variation and make biological inference

1) Estimate allele frequencies (ANGSD)
2) Population structure, demography, selection.

Extracted DNA strand

Break up the genomic DNA somehow: sonic shearing, restriction enzymes, mother nature.

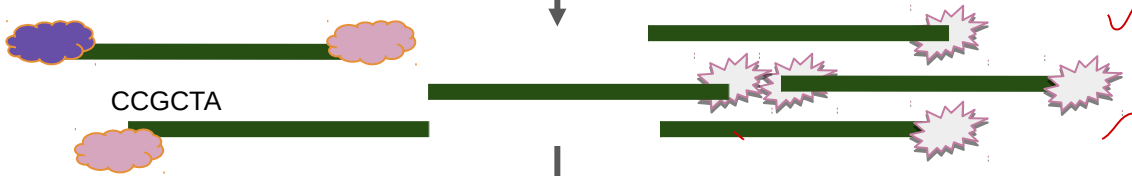




Phase 1: Library preparation

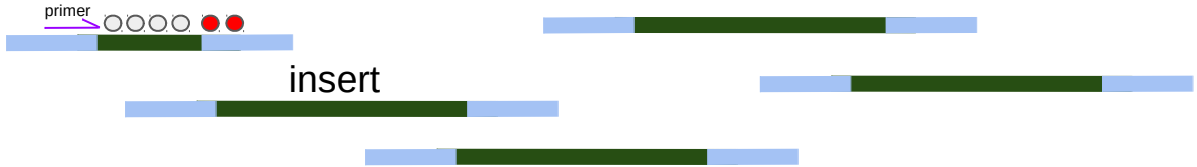
Extracted DNA strand

Break up the genomic DNA somehow: ultrasonic waves, restriction enzymes, mother nature.



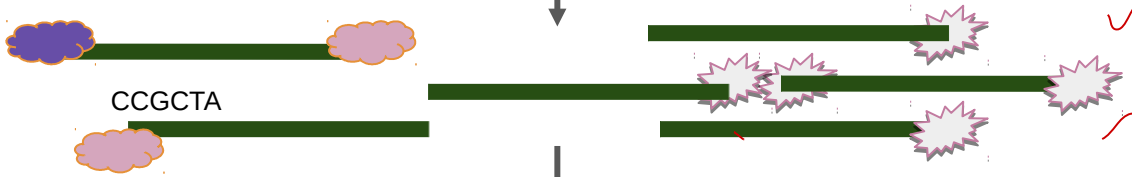
Sequencing into the adapter = adapter contamination!!!

Adapter ligation & indexing



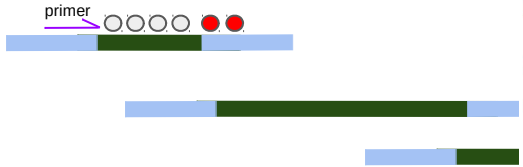
Extracted DNA strand

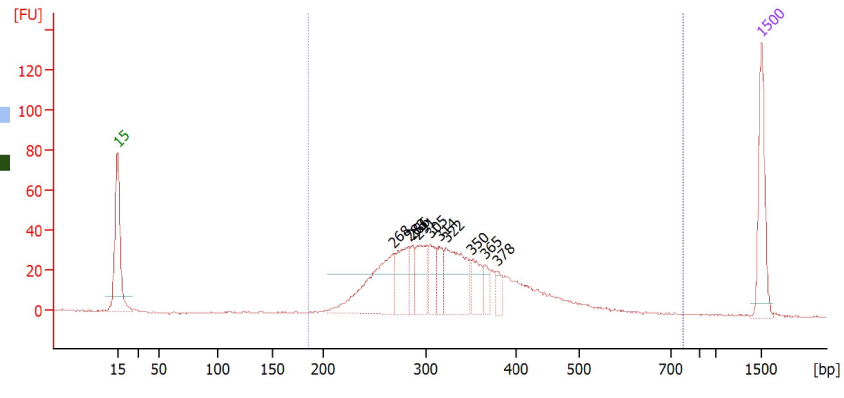
Break up the genomic DNA somehow: ultrasonic waves, restriction enzymes, mother nature.



Adapter ligation & indexing

Sequencing into the adapter = adapter contamination





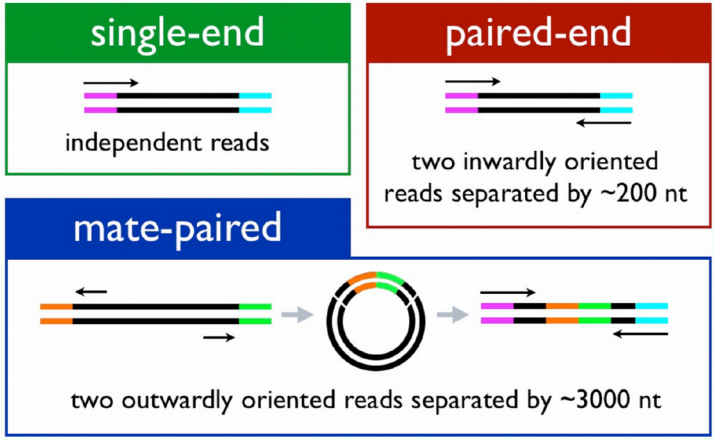
Extracted DNA strand

Break up the genomic DNA somehow: ultrasonic waves, restriction enzymes, mother nature.

Adapter ligation

Sequencing into the adapter = adapter contamination

sequencing



A typical NGS roadmap

Phase 1. Wetlab data generation.

1) DNA extraction
2) library construction
3) Sequencing

Phase 2. Mapping / Alignment

1) FASTQ quality control (FastQC, cutadapt, trimmomatic, Trim Galore, PEAR, FLASH, super_deduper).
2) Assembly
3) Mapping (bwa, Bowtie, Novoalign).

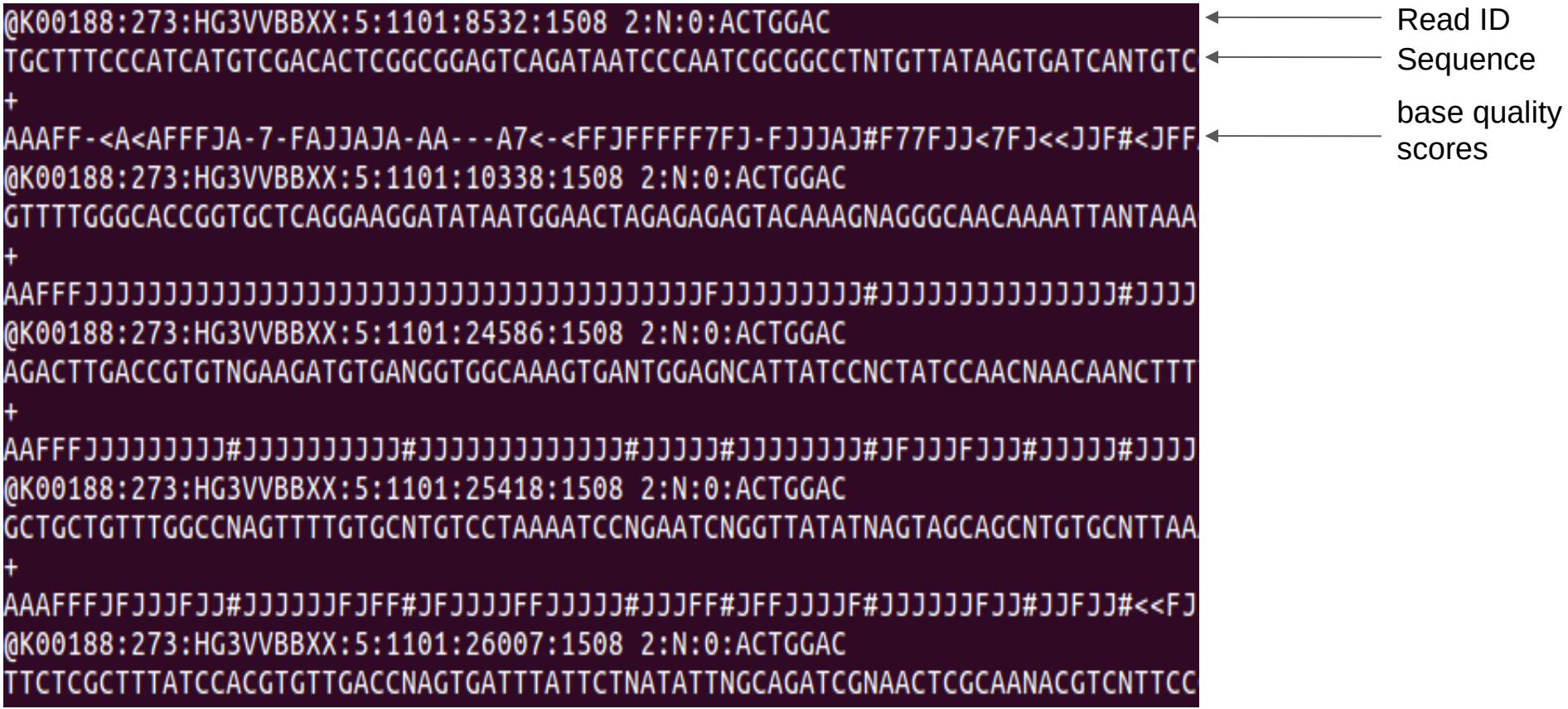
Phase 3. Variant discovery

1) Site-level quality control
2) Call genotypes and SNPs/INDELS (ANGSD, bcftools, GATK, freeBayes)

Phase 4. Characterize genetic variation and make biological inference

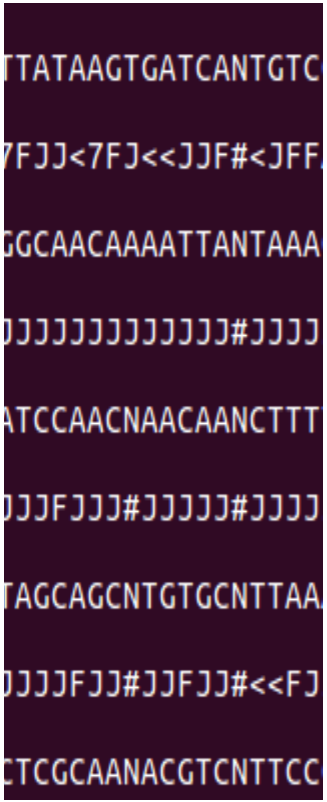
1) Estimate allele frequencies (ANGSD)
2) Population structure, demography, selection.

Raw sequencing reads in FASTQ format



Raw sequencing reads in FASTQ format

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



The quality scores are in ASCII encoding and can be interpreted as the probability of an error, ϵ.

They are in Phred scale: $Qscore = -10*\log_{10}(\epsilon)$

Given the quality score then, $\epsilon = 10^{-Q/10}$

Example:

Raw decimal values are normally offset by 33 (64 on some older platforms). So a Qscore of '5' is 53-33 = 20.

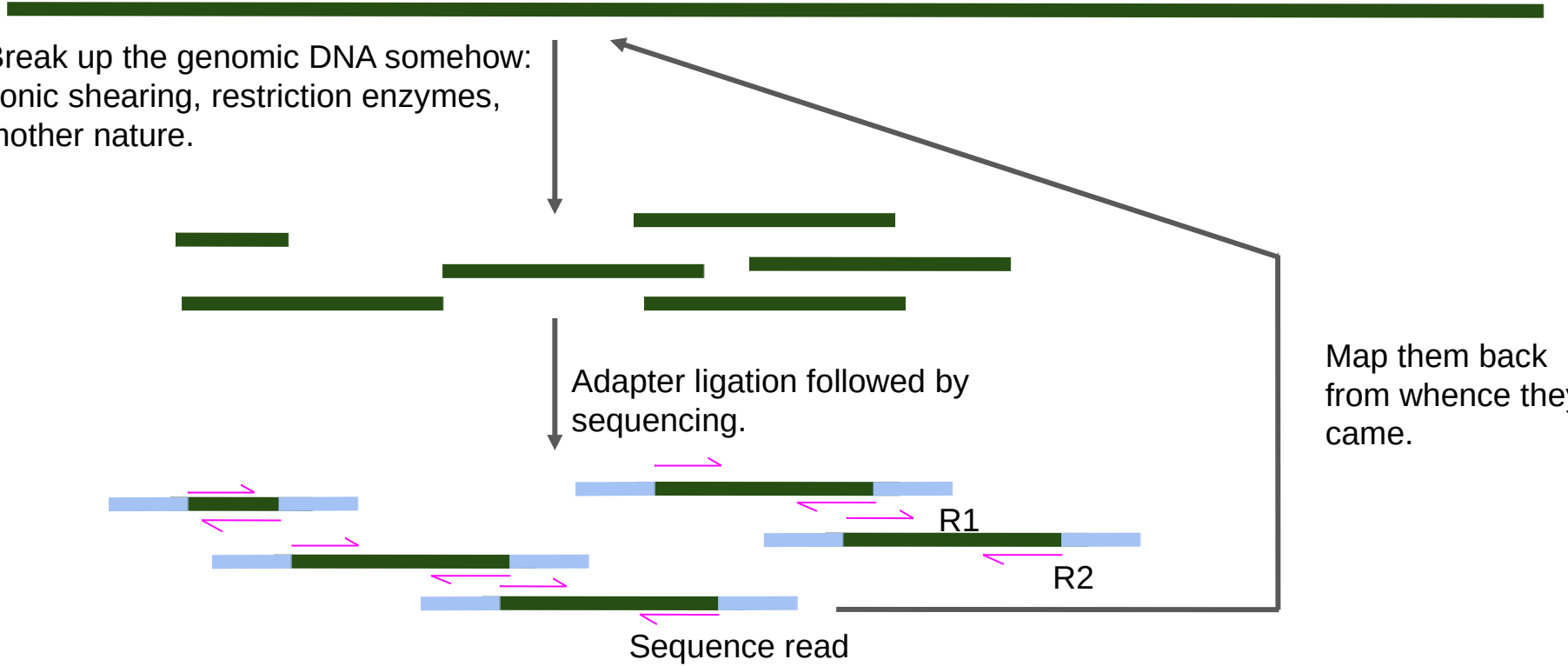
$10^{-20/10} = 1\%$.

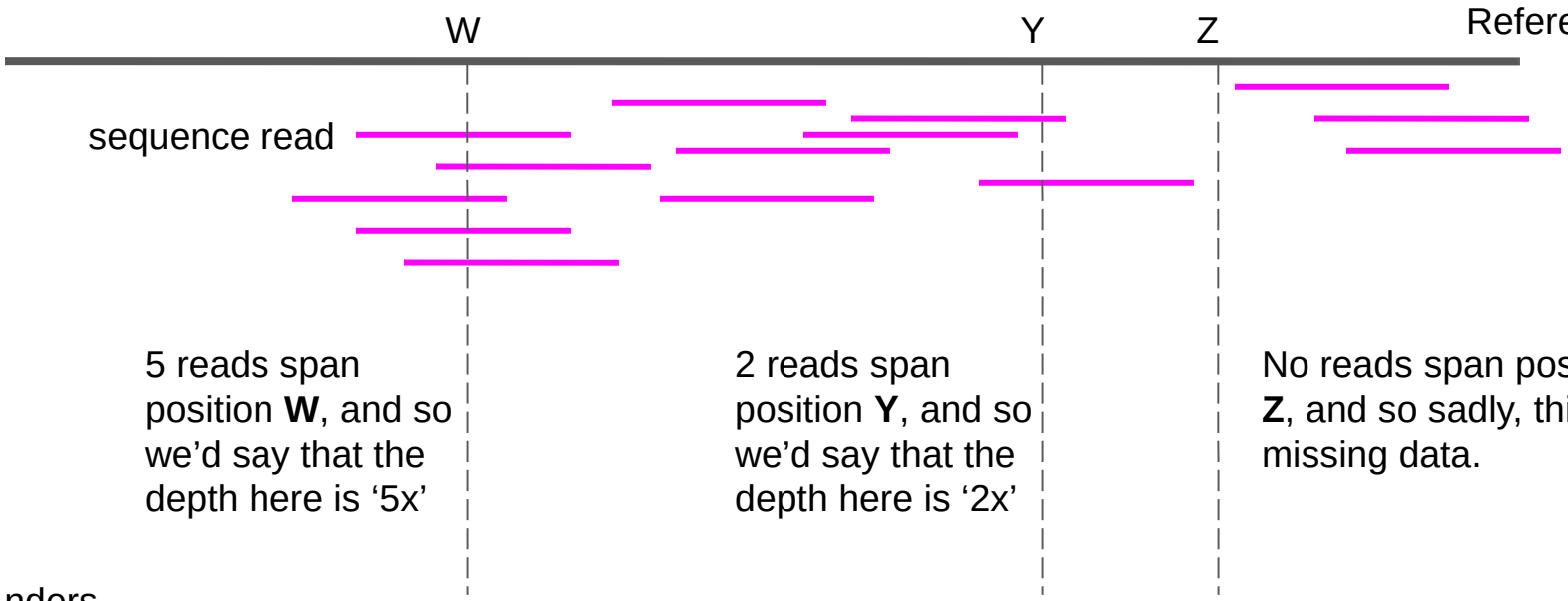
There is a 1% probability that a base with Qscore '5' is an error.

Some things to watch out for in your FASTQ files

- Adapter contamination
- Low quality bases (usually at the ends of reads and can be soft-clipped)
- Low complexity reads
- Excess missing bases (‘N’)
- Overlapping reads

DNA strand (sequence represented in reference assembly)





Anders

My definitions (The literature is not consistent)

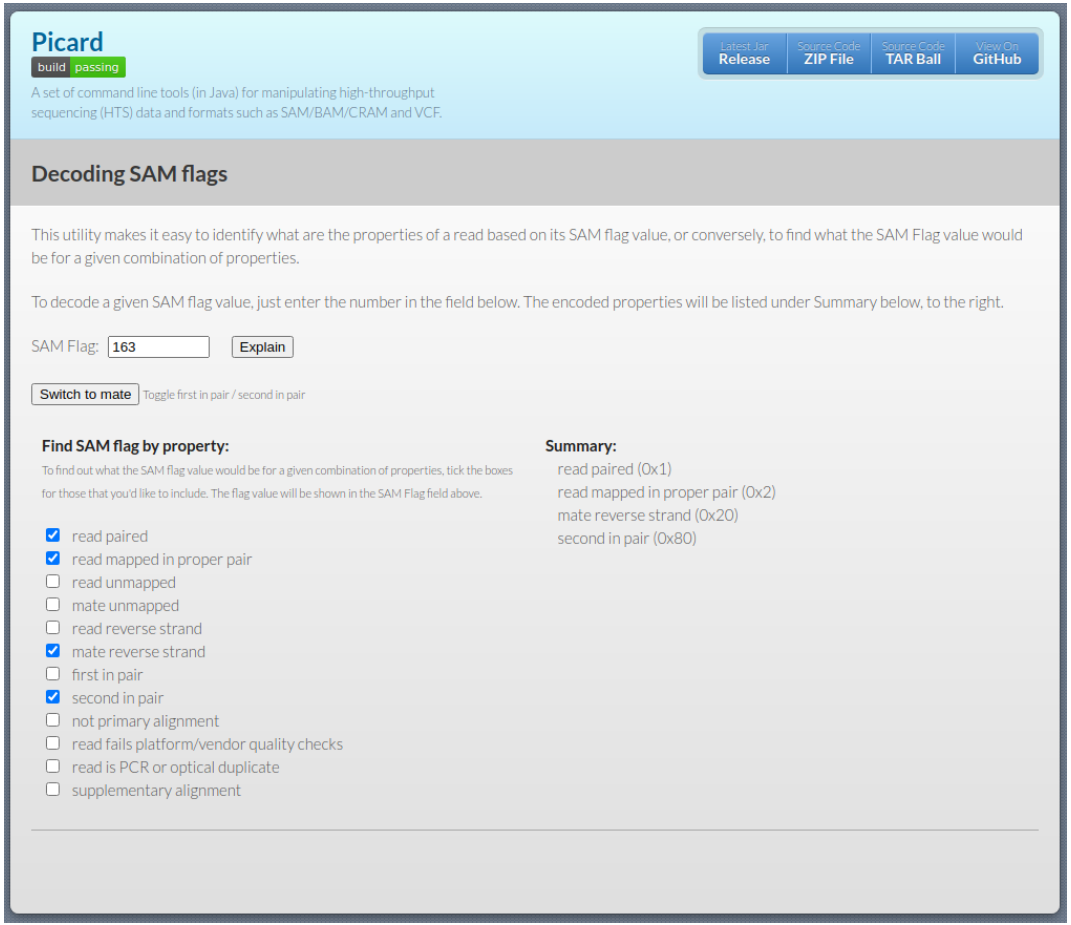
Depth	The number of reads that maps to a position
Coverage	The fraction of the genome (region) with data
Depth-of-coverage	Average depth for sites that are covered

SAM/BAM/CRAM format (for mapped reads)

```
HS36_21523:8:1214:17897:31562   163     chr7    16      23      125M    =       503     612     CGAGAGTGCAGCAGTTTGGTGAAAGGAACAGGTGAAGAAGAGGTGAGCATCCAAGAGACGAGCAAGAGGACAGTTTCAAAGCAGACATTATGGCACAATAAAAGCTGT
AAGAGCAGAGATTCATT   EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE   AS:i:120        XS:i:125        MQ:i:23 MC:Z:125M
ms:i:4587       MD:Z:108G16     NM:i:1  RG:Z:21523_8#1
HS36_21523:8:2114:12416:4261    99      chr7    35      0       125M    =       446     533     TGAAAGGAACAGGTGAAGAAGAGGTGAGCATCCAAGAGACGAGCAAGAGGACAGTTTCAAAGCAGACATTATGGCACAATAAAAGCTGTAAGAGCAGAGATTCATTCA
TATCTTTACTATTTTTC   AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA   AS:i:120        XS:i:125        XA:Z:chr8,-23982542,125M
,0;chr16,-2795556,125M,1;chr18,+31933268,125M,1;chr18,+30035788,125M,2;chr8,-23993448,125M,3;   BC:Z:GAGGATGGTCTTTCCC   QT:Z:BBBBB<FF/BBBBBFF   MQ:i:0  MC:Z:122M3S     ms:i:3656       MD:Z:89G35      NM:i
:1  RG:Z:21523_8#1
HS36_21523:8:1103:15663:2731    99      chr7    45      47      125M    =       593     673     AGGTGAAGAAGAGGTGAGCATCCAAGAGACGAGCAAGAGGACAGTTTCAAAGCAGACATTATGGCACAATAAAAGCTGTGAGAGCAGAGATTCATTCATATCTTTACT
ATTTTTCTTTAGCTTCT   BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB   AS:i:125        XS:i:120        XA:Z:chr8,-23982532,125M
,1;chr16,-2795546,125M,2;chr18,+31933278,125M,2;chr8,-23993438,125M,3;chr18,+30035798,125M,3;   BC:Z:GAGGATGGTCTTTCCC   QT:Z:BBBBBFFFBBBBBFFF   MQ:i:48 MC:Z:125M       ms:i:3343       MD:Z:125        NM:i
:0  RG:Z:21523_8#1
```

<table>
<tr><th>Col</th><th>Field</th><th>Type</th><th>Regexp/Range</th><th>Brief description</th></tr>
<tr><td>1</td><td>QNAME</td><td>String</td><td><code>[!-?A-~]{1,254}</code></td><td>Query template NAME (read ID)</td></tr>
<tr><td>2</td><td>FLAG</td><td>Int</td><td>$[0, 2^{16} - 1]$</td><td>bitwise FLAG</td></tr>
<tr><td>3</td><td>RNAME</td><td>String</td><td><code>*|[:rname:^*=][:rname:]*</code></td><td>Reference sequence NAME[11]</td></tr>
<tr><td>4</td><td>POS</td><td>Int</td><td>$[0, 2^{31} - 1]$</td><td>1-based leftmost mapping POSition</td></tr>
<tr><td>5</td><td>MAPQ</td><td>Int</td><td>$[0, 2^{8} - 1]$</td><td>MAPping Quality</td></tr>
<tr><td>6</td><td>CIGAR</td><td>String</td><td><code>*|([0-9]+[MIDNSHPX=])+</code></td><td>CIGAR string</td></tr>
<tr><td>7</td><td>RNEXT</td><td>String</td><td><code>*|=|[:rname:^*=][:rname:]*</code></td><td>Reference name of the mate/next read</td></tr>
<tr><td>8</td><td>PNEXT</td><td>Int</td><td>$[0, 2^{31} - 1]$</td><td>Position of the mate/next read</td></tr>
<tr><td>9</td><td>TLEN</td><td>Int</td><td>$[-2^{31} + 1, 2^{31} - 1]$</td><td>observed Template LENgth</td></tr>
<tr><td>10</td><td>SEQ</td><td>String</td><td><code>*|[A-Za-z=.]+</code></td><td>segment SEQuence</td></tr>
<tr><td>11</td><td>QUAL</td><td>String</td><td><code>[!-~]+</code></td><td>ASCII of Phred-scaled base QUALity+33</td></tr>
</table>

https://broadinstitute.github.io/picard/explain-flags.html



Can use this tool from the Broad to interpret the bitwise flags.

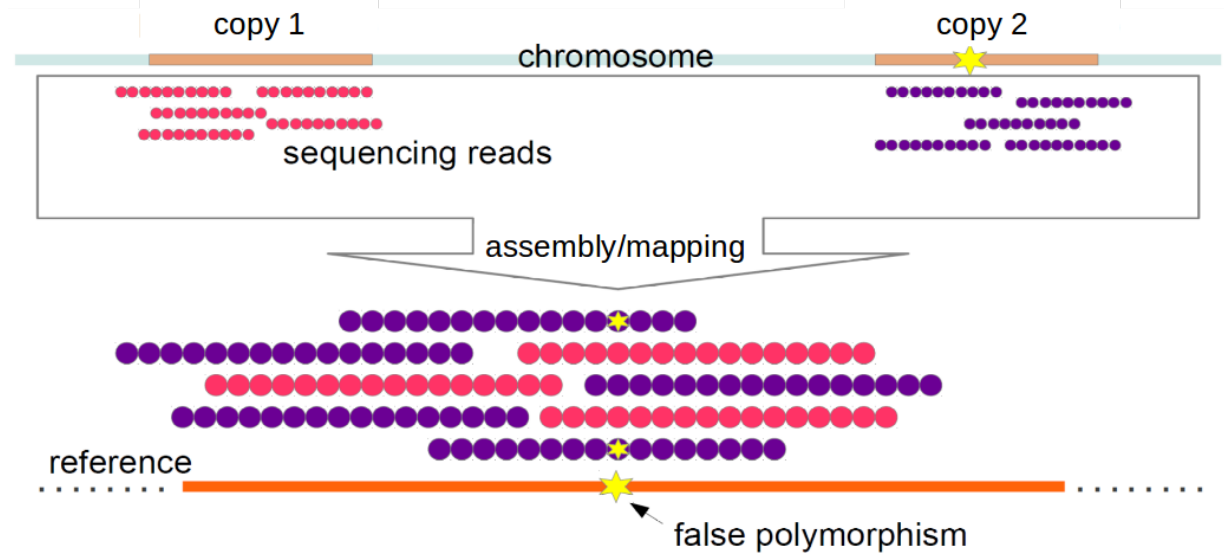
Phase 2: Mapping and alignment

```
HS36_21523:8:1214:17897:31562   163     chr7    16      23      125M    =       503     612     CGAGAGTGCAGCAGTTTGGTGAAAGGAACAGGTGAAGAAGAGGTGAGCATCCAAGAGACGAGCAAGAGGACAGTTTCAAAGCAGACATTATGGCACAATAAAAGCTGT
AAGAGCAGAGATTCATT   EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE   AS:i:120        XS:i:125        MQ:i:23 MC:Z:125M
ms:i:4587       MD:Z:108G16     NM:i:1  RG:Z:21523_8#1
HS36_21523:8:2114:12416:4261    99      chr7    35      0       125M    =       446     533     TGAAAGGAACAGGTGAAGAAGAGGTGAGCATCCAAGAGACGAGCAAGAGGACAGTTTCAAAGCAGACATTATGGCACAATAAAAGCTGTAAGAGCAGAGATTCATTCA
TATCTTTACTATTTTTC   AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA   AS:i:120        XS:i:125        XA:Z:chr8,-23982542,125M
,0;chr16,-2795556,125M,1;chr18,+31933268,125M,1;chr18,+30035788,125M,2;chr8,-23993448,125M,3;   BC:Z:GAGGATGGTCTTTCCC   QT:Z:BBBBB<FF/BBBBBFF   MQ:i:0  MC:Z:122M3S     ms:i:3656       MD:Z:89G35      NM:i
:1  RG:Z:21523_8#1
HS36_21523:8:1103:15663:2731    99      chr7    45      47      125M    =       593     673     AGGTGAAGAAGAGGTGAGCATCCAAGAGACGAGCAAGAGGACAGTTTCAAAGCAGACATTATGGCACAATAAAAGCTGTGAGAGCAGAGATTCATTCATATCTTTACT
ATTTTTCTTTAGCTTCT   BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB   AS:i:125        XS:i:120        XA:Z:chr8,-23982532,125M
,1;chr16,-2795546,125M,2;chr18,+31933278,125M,2;chr8,-23993438,125M,3;chr18,+30035798,125M,3;   BC:Z:GAGGATGGTCTTTCCC   QT:Z:BBBBBFFFBBBBBFFF   MQ:i:48 MC:Z:125M       ms:i:3343       MD:Z:125        NM:i
:0  RG:Z:21523_8#1
```

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,254}`	Query template NAME ID of the read
2	FLAG	Int	$[0, 2^{16}-1]$	bitwise FLAG
3	RNAME	String	`\*|[:rname:^*=][:rname:]*`	Reference sequence NAME[11]
4	POS	Int	$[0, 2^{31}-1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0, 2^{8}-1]$	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[:rname:^*=][:rname:]*`	Reference name of the mate/next read
8	PNEXT	Int	$[0, 2^{31}-1]$	Position of the mate/next read
9	TLEN	Int	$[-2^{31}+1, 2^{31}-1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

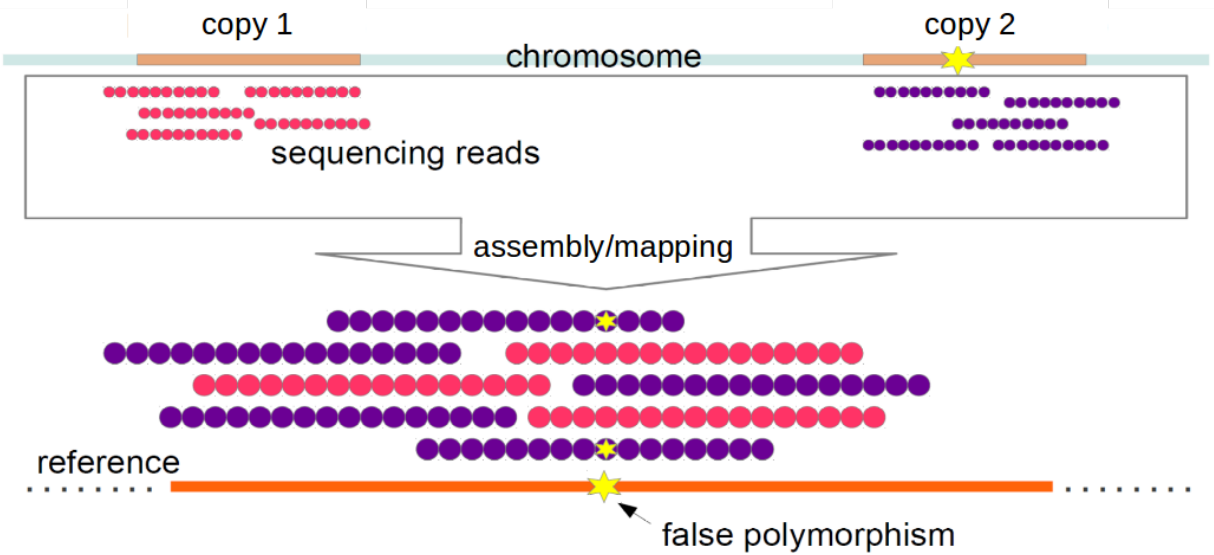
Read HS36_21523:8:1214:17897:31562 is mapped in a properly mapped pair. Its leftmost coordinate is at chr7:16 on the reference and was identical to the reference along all 125 of its bases. The mapping quality is 23 and so could be mismapped with ~0.5% probability.

The problem of paralogs (or any other type of highly similar sequences repeated throughout the genome)



- Sequence depth proportional to the number of copies.
- Tendency to inflate estimates of heterozygosity .

The problem of paralogs (or any other type of repeat sequence)



- Sequence depth proportional to the number of copies.
- Tendency to inflate estimates of heterozygosity.
- Can generate improperly mapped reads

Pileup format (generated with SAMtools)

chr7	10000	T	2	.,	EB	7	.,,,,,^],	>IIBGGE	6	.....,	DEGEGG	9	,.,,.,,.,	DABGIIIII
chr7	10001	C	3	.,^].	EBB	7	.,,,,,,	>IIBGGI	6	.....,	DEGEGG	9	,.,,.,,.,	DABGIIIII
chr7	10002	G	3	.,.	EBB	7	.,,,,,,	>IIBGGI	6	.....,	DEGEGG	9	,.,,.,,.,	DABGIIIII
chr7	10003	G	3	.,.	EBB	7	.,,,,,,	>IIBGGI	6	.....,	DEGEGG	10	,.,,.,,.,^].	DABGIIIIIE
chr7	10004	A	3	.,.	EBB	8	.,,,,,,^].	>IIBGGIE	6	.....,	DEGEGG	10	,.,,.,,.,.	DABGIIIIII
chr7	10005	G	3	.,.	EBB	8	.,,,,,,.	>IIBGGIG	6	.....,	DEGEGG	10	,.,,.,,.,.	DABGIIIIII
chr7	10006	A	3	.,.	EBB	8	.,,,,,,.	>IIBGGIG	6	.....,	DEGEGG	10	,.,,.,,.,.	DABGIIIIII
chr7	10007	G	3	.,.	EBB	8	.,,,,,,.	>IIBGGIG	7	.....,^].	DEGEGGE	10	,.,,.,,.,.	DABGIIIIII
chr7	10008	C	3	.,.	EBB	8	.,,,,,,.	>IIBGGIG	7	.....,.	DEGEGGG	10	,.,,.,,.,.	DABGIIIIII
chr7	10009	A	3	.,.	EBB	8	.,,,,,,.	>IIBGGIG	8	.....,.^],	DEGEGGGB	10	,.,,.,,.,.	DABGIIIIII
chr7	10010	G	3	.,.	EBB	8	.,,,,,,.	>IIBGGIG	8	.....,.,	DEGEGGGB	10	,.,,.,,.,.	DABGIIIIII
chr7	10011	C	3	.,.	EBB	8	.,,,,,,.	>IIBGGIG	8	.....,.,	DEGEGGGB	10	,.,,.,,.,.	DABGIIIIII
chr7	10012	T	3	.,.	EBB	8	.,,,,,,.	>IIBGGIG	8	.....,.,	DEGEGGGB	10	,.,,.,,.,.	DABGIIIIII
chr7	10013	T	3	.,.	EBB	8	.,,,,,,.	>IIBGGIG	9	.....,.,^],	DEGEGGGBE	10	,.,,.,,.,.	DABGIIIIII
chr7	10014	A	4	.,.^].	EBBE	8	.,,,,,,.	>IIBGGIG	9	.G...gG,,	DEGEGGGBG	10	,.,,.,,.,.	D3BGIIIIII
chr7	10015	G	4	.,..	EBBI	8	.,,,,,,.	>IIBGGIG	9	.....,.,,	DEGEGGGBG	10	,.,,.,,.,.	D3BGIIIIII

Left-most fields:

(1) Reference sequence
(2) Position
(3) Reference base

Each individual has 3 fields:

(1) Depth
(2) read bases
(3) quality scores

Base codes:

. = reference match on forward strand
, = reference match on revere strand
[ACGTacgt] = alternate allele (forward strand = uppercase, reverse strand = lowercase)

Pileup format (generated with SAMtools)

chr7	10000	T	2	.,	EB	7	.,,,,,^],	>IIBGGE	6	.....,,	DEGEGG	9	,.,,.,,.,	DABGIIIII
chr7	10001	C	3	.,^].	EBB	7	.,,,,,,	>IIBGGI	6	.....,,	DEGEGG	9	,.,,.,,.,	DABGIIIII
chr7	10002	G	3	.,.	EBB	7	.,,,,,,	>IIBGGI	6	.....,,	DEGEGG	9	,.,,.,,.,	DABGIIIII
chr7	10003	G	3	.,.	EBB	7	.,,,,,,	>IIBGGI	6	.....,,	DEGEGG	10	,.,,.,,.,^].	DABGIIIIIE
chr7	10004	A	3	.,.	EBB	8	.,,,,,,^].	>IIBGGIE	6	.....,,	DEGEGG	10	,.,,.,,.,.	DABGIIIIII
chr7	10005	G	3	.,.	EBB	8	.,,,,,,.	>IIBGGIG	6	.....,,	DEGEGG	10	,.,,.,,.,.	DABGIIIIII
chr7	10006	A	3	.,.	EBB	8	.,,,,,,.	>IIBGGIG	6	.....,,	DEGEGG	10	,.,,.,,.,.	DABGIIIIII
chr7	10007	G	3	.,.	EBB	8	.,,,,,,.	>IIBGGIG	7	.....,^].	DEGEGGE	10	,.,,.,,.,.	DABGIIIIII
chr7	10008	C	3	.,.	EBB	8	.,,,,,,.	>IIBGGIG	7	.....,.	DEGEGGG	10	,.,,.,,.,.	DABGIIIIII
chr7	10009	A	3	.,.	EBB	8	.,,,,,,.	>IIBGGIG	8	.....,.^],	DEGEGGGB	10	,.,,.,,.,.	DABGIIIIII
chr7	10010	G	3	.,.	EBB	8	.,,,,,,.	>IIBGGIG	8	.....,.,	DEGEGGGB	10	,.,,.,,.,.	DABGIIIIII
chr7	10011	C	3	.,.	EBB	8	.,,,,,,.	>IIBGGIG	8	.....,.,	DEGEGGGB	10	,.,,.,,.,.	DABGIIIIII
chr7	10012	T	3	.,.	EBB	8	.,,,,,,.	>IIBGGIG	8	.....,.,	DEGEGGGB	10	,.,,.,,.,.	DABGIIIIII
chr7	10013	T	3	.,.	EBB	8	.,,,,,,.	>IIBGGIG	9	.....,.,^],	DEGEGGGBE	10	,.,,.,,.,.	DABGIIIIII
chr7	10014	A	4	.,.^].	EBBE	8	.,,,,,,.	>IIBGGIG	9	.G...gG,,	DEGEGGGBG	10	,.,,.,,.,.	D3BGIIIIII
chr7	10015	G	4	.,..	EBBI	8	.,,,,,,.	>IIBGGIG	9	.....,.,,	DEGEGGGBG	10	,.,,.,,.,.	D3BGIIIIII

Start of read with ASCII MQ ‘]’

Other characters that show up:
$ = end of read
* = missing base
</> = reference skip

A typical NGS roadmap

Phase 1. Wetlab data generation.

1) DNA extraction
2) library construction
3) Sequencing

Phase 2. Mapping / Alignment

1) FASTQ quality control (FastQC, cutadapt, trimmomatic, Trim Galore, PEAR, FLASH, super_deduper).
2) Assembly
3) Mapping (bwa, Bowtie, Novoalign).

Phase 3. Variant discovery

1) Site-level quality control
2) Call genotypes and SNPs/INDELS (ANGSD, bcftools, GATK, freeBayes)

Phase 4. Characterize genetic variation and make biological inference

1) Estimate allele frequencies (ANGSD)
2) Population structure, demography, selection.

Variant Call Format (VCF)

```
##INFO=<ID=MQ,Number=1,Type=Integer,Description="Average mapping quality">
##INFO=<ID=PV4,Number=4,Type=Float,Description="P-values for strand bias, baseQ bias, mapQ bias and tail distance bias">
##bcftools_callVersion=1.13-3-g89a566b+htslib-1.13-3-gd16bed5
##bcftools_callCommand=call --ploidy 2 -a PV4,GQ,GP -m -P 0.001 -O u; Date=Sun Jul 25 00:26:03 2021
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of samples with data">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele frequency">
##INFO=<ID=ExcHet,Number=A,Type=Float,Description="Test excess heterozygosity; 1=good, 0=bad">
##bcftools_pluginVersion=1.13-3-g89a566b+htslib-1.13-3-gd16bed5
##bcftools_pluginCommand=plugin fill-tags -O b -o /home/tyler/ngs_intro/output/calmas_allsites.bcf.gz -- -tAF,NS,ExcHet; Date=Sun Jul 25 00:26:04 2021
##bcftools_viewVersion=1.3.1-98-ga6a7829+htslib-1.3.1-64-g74bcfd7
##bcftools_viewCommand=view -r chr7:10000-10200 calmas_allsites.bcf.gz; Date=Mon Jul 26 11:44:41 2021
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  CMASS6607982    CMASS6607991    CMASS6389725    CMASS6608065    CMASS6169461    CMASS6169452    CMASS6169443    CMASS63
chr7    10000   .       T       .       4752.11 .       DP=165;AD=165;SCR=21;MQSBZ=0.639526;FS=0;MQ0F=0;AN=78;DP4=89,76,0,0;MQ=56;NS=39 GT:DP:AD:SCR:QS 0/0:2:2:0:69    0/0:7:7:1:258   0/0:6:6
chr7    10001   .       C       .       4795.11 .       DP=166;AD=166;SCR=21;MQSBZ=0.609616;FS=0;MQ0F=0;AN=78;DP4=90,76,0,0;MQ=56;NS=39 GT:DP:AD:SCR:QS 0/0:3:3:0:102   0/0:7:7:1:258   0/0:6:6
chr7    10002   .       G       .       4874.11 .       DP=167;AD=167;SCR=21;MQSBZ=0.580082;FS=0;MQ0F=0;AN=78;DP4=91,76,0,0;MQ=57;NS=39 GT:DP:AD:SCR:QS 0/0:3:3:0:102   0/0:7:7:1:258   0/0:6:6
chr7    10003   .       G       .       4920.11 .       DP=170;AD=170;SCR=21;MQSBZ=0.493647;FS=0;MQ0F=0;AN=78;DP4=94,76,0,0;MQ=57;NS=39 GT:DP:AD:SCR:QS 0/0:3:3:0:102   0/0:7:7:1:258   0/0:6:6
chr7    10004   .       A       .       23.6449 .       DP=172;AD=171;SCR=21;SGB=-0.626698;RPBZ=1.6921;MQBZ=-2.46491;MQSBZ=0.499982;BQBZ=-1.67883;SCBZ=3.32672;FS=0;MQ0F=0;AN=78;DP4=94,77,1,0;
chr7    10005   .       G       .       5018.12 .       DP=173;AD=173;SCR=21;MQSBZ=0.472176;FS=0;MQ0F=0;AN=78;DP4=96,77,0,0;MQ=57;NS=39 GT:DP:AD:SCR:QS 0/0:3:3:0:102   0/0:8:8:1:296   0/0:6:6
chr7    10006   .       A       .       24.4501 .       DP=174;AD=172;SCR=21;VDB=0.14;SGB=-2.84764;RPBZ=2.23109;MQBZ=-3.51912;MQSBZ=0.506159;BQBZ=-2.38185;SCBZ=4.42789;FS=0;MQ0F=0;AN=78;DP4=9
chr7    10007   .       G       .       5087.12 .       DP=176;AD=176;SCR=21;MQSBZ=0.451479;FS=0;MQ0F=0;AN=78;DP4=98,78,0,0;MQ=57;NS=39 GT:DP:AD:SCR:QS 0/0:3:3:0:102   0/0:8:8:1:296   0/0:7:7
chr7    10008   .       C       .       5066.12 .       DP=176;AD=176;SCR=21;MQSBZ=0.451479;FS=0;MQ0F=0;AN=78;DP4=98,78,0,0;MQ=57;NS=39 GT:DP:AD:SCR:QS 0/0:3:3:0:102   0/0:8:8:1:296   0/0:7:7
chr7    10009   .       A       .       5155.12 .       DP=179;AD=179;SCR=21;MQSBZ=0.431514;FS=0;MQ0F=0;AN=78;DP4=100,79,0,0;MQ=57;NS=39       GT:DP:AD:SCR:QS 0/0:3:3:0:102   0/0:8:8:1:296
chr7    10010   .       G       .       5128.12 .       DP=180;AD=180;SCR=21;MQSBZ=0.405226;FS=0;MQ0F=0;AN=78;DP4=101,79,0,0;MQ=57;NS=39       GT:DP:AD:SCR:QS 0/0:3:3:0:102   0/0:8:8:1:296
chr7    10011   .       C       .       5228.12 .       DP=183;AD=183;SCR=21;MQSBZ=0.444839;FS=0;MQ0F=0;AN=78;DP4=102,81,0,0;MQ=57;NS=39       GT:DP:AD:SCR:QS 0/0:3:3:0:102   0/0:8:8:1:296
chr7    10012   .       T       .       5332.12 .       DP=186;AD=186;SCR=21;MQSBZ=0.425788;FS=0;MQ0F=0;AN=78;DP4=104,82,0,0;MQ=57;NS=39       GT:DP:AD:SCR:QS 0/0:3:3:0:102   0/0:8:8:1:296
chr7    10013   .       T       .       5390.12 .       DP=188;AD=188;SCR=21;MQSBZ=0.432328;FS=0;MQ0F=0;AN=78;DP4=105,83,0,0;MQ=57;NS=39       GT:DP:AD:SCR:QS 0/0:3:3:0:102   0/0:8:8:1:296
chr7    10014   .       A       G       259.035 .       DP=190;AD=177,12;SCR=21;VDB=0.311296;SGB=16.1258;RPBZ=-0.0261303;MQBZ=0.689528;MQSBZ=0.382678;BQBZ=0.848991;SCBZ=-1.16673;FS=0;MQ0F=0;A
chr7    10015   .       G       .       5441.12 .       DP=190;AD=190;SCR=21;MQSBZ=0.382678;FS=0;MQ0F=0;AN=78;DP4=107,83,0,0;MQ=57;NS=39       GT:DP:AD:SCR:QS 0/0:4:4:0:142   0/0:8:8:1:296
```

metadata

#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT CMASS6607982

chr7 10014 . A G 259.035 . DP=190;AD=177,12;SCR=21;VDB=0.311296;SGB=16.1258;RPBZ=-0.0261303;MQBZ=0.689528;;BQBZ=0.848991;SCBZ=-1.16673;FS=0;MQ0F=0;AC=3;AN=78;DP4=99,78,8,5;MQ=57;PV4=0.778225,1,1,1;NS=39;AF=0.0384615;ExcHet=0.961039 GT:PL:DP:AD:SCR:QS:GP:GQ 0/0:0,12,125:4:4,0:0:142,0:0.994672,0.00532842,5.66837e-16:22

Variant Call Format (VCF)

#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT CMASS6607982

chr7 10014 . A G 259.035 .
DP=190;AD=177,12;SCR=21;VDB=0.311296;SGB=16.1258;RPBZ=-0.0261303;MQBZ=0.689528;;BQBZ=0.848991;SCBZ=-1.16673;FS=0;MQ0F=0;AC=3;AN=78;DP4=99,78,8,5;MQ=57;PV4=0.778225,1,1,1;NS=39;AF=0.0384615;ExcHet=0.961039 GT:PL:DP:AD:SCR:QS:GP:GQ
0/0:0,12,125:4:4,0:0:142,0:0.994672,0.00532842,5.66837e-16:22

INFO: site-wide info

FORMAT: Format of the information for individuals

Information for the first individual

This is a lot of rich information that we can subset sites with using bcftools.

Site-level quality control

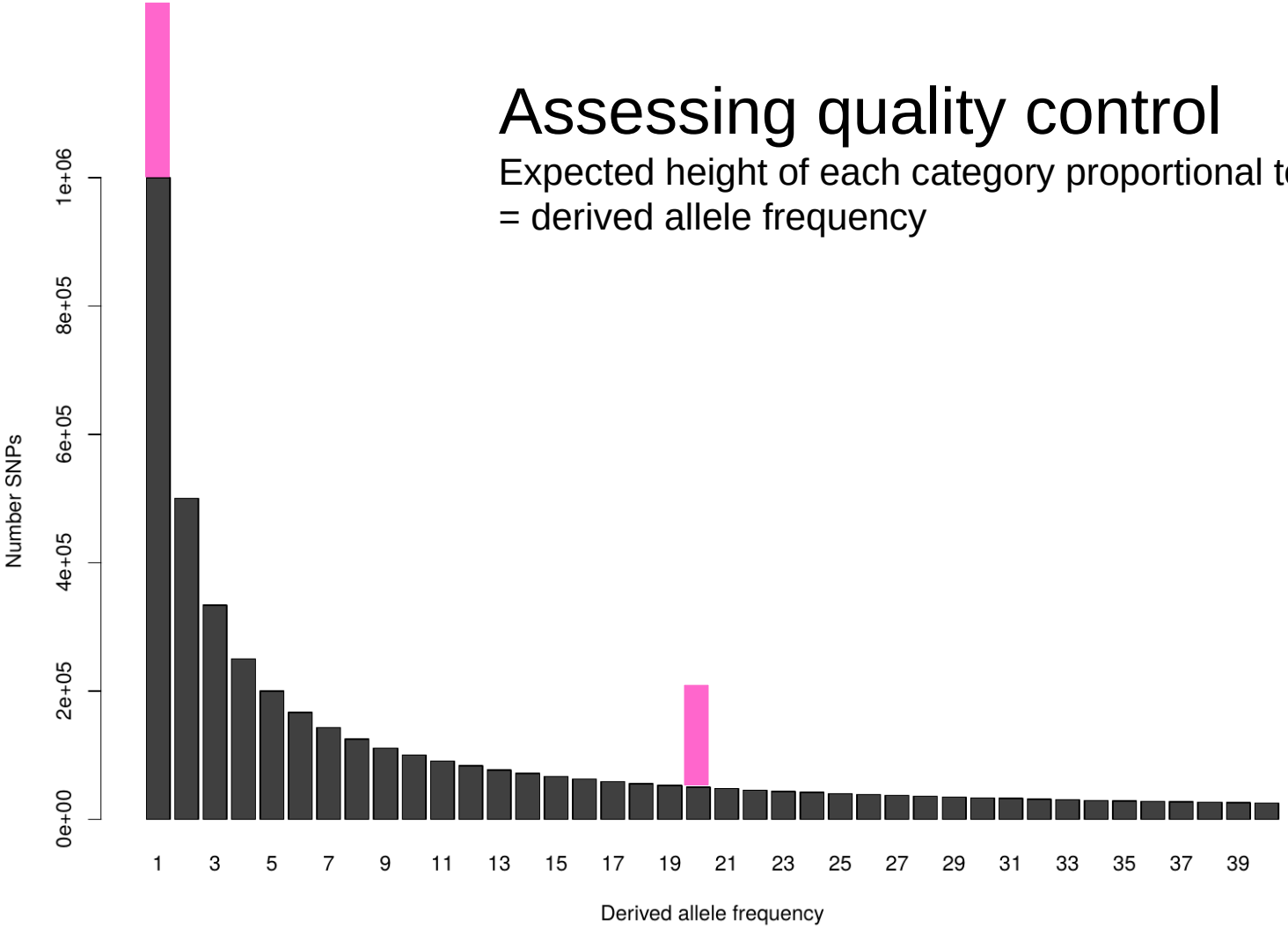
- Excessively low total site sequencing depth
- Excessively high total site sequencing depth
- Evenness of depth across individuals.
- Minimum number of individuals with called genotypes (sufficient high-confidence data)
- Mapping quality and excess mapping quality of zero
- Excess heterozygosity

Filters related to biases between reference and alternate alleles (we expect them to act the same):

- Map quality bias
- Base quality bias
- Read position bias
- Strand bias

Assessing quality control

Expected height of each category proportional to 1/x, x = derived allele frequency



Lets put it all into practice:

https://github.com/tplinderoth/Copenhagen-Popgen-Course/tree/main/ngs_intro_exercises