Intro to NGS
Data Processing and Formats

Tyler Linderoth
Copenhagen Popgen Summer course 2021

With some poached materials from Anders Albrechtsen

Some motivations for NGS.

- Genome-wide vs. targeted sequencing data is much more conducive to identifying and characterizing the evolutionary roles of **structural variation**.

- Genome-wide association studies (**GWAS**). Much more efficient than candidate gene approaches (though identifying truly novel gene functions is still a challenge).

- Discerning how the **landscape of genetic variability** across the genome influences fitness. E.g. there's recently been debate over how important overall levels of genetic diversity are to population viability compared to how it's distributed in the genome.

- Being able to infer different genealogical histories along the genome has greatly enhanced our ability to reconstruct **population histories** (e.g. PSMC).

- Much more efficient for identifying and characterizing **hybridization** between species. Without NGS, finding regions of adaptive introgression for instance could be a needle-in-a-haystack endeavor. This has greatly enhanced our view of how species engage with one another to influence biodiversity and evolution.

- **Genetic monitoring** through eDNA and metabarcoding.

- Effective for sequencing degraded **historic and ancient DNA**.

So NGS provides "genome-wide" data, but how does it do it and what are the different flavors?

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

- Reduced representation, which target some fraction of the genome.
    - Targeted sequence capture.
    - Restriction-associated digest (RAD).
        - single-digest
        - double-digest
    - RAD-capture.

So you've gathered that NGS provides "genome-wide" data, but how does it do it and what are the different flavors?

Whole genome sequencing (WGS)

Pros:
- Most amiable for discovering both SNP and structural variation.

- Better ability to finely identify causal variants in association studies and divergence scans.

- Library preparation is easier.

Cons:
- Require's a reliable reference genome, which can be difficult to generate depending on the organism.

- More sequencing and cost to achieve a target sequencing depth-of-coverage. May not be able to achieve high enough coverage to reliably phase and so could preclude the use of powerful haplotype-based analytical methods.

So you've gathered that NGS provides "genome-wide" data, but how does it do it and what are the different flavors?

Reduced representation, which target some fraction of the genome.

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 can be easier since the data has often been pre-screened for regions that will be difficult to assemble.

Cons:
- Gamble on being able to identify particular regions of interest through linkage-disequilibrium (LD), and often the extent of LD is unknown.

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

- RAD is anonymous (you may not be able to find from where in the genome your SNPs are actually located unless there are suitable references).

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) Quality control sites to ensure any genetic variation or (lack of genetic variation) at them is deemed reliable.
2) Screen for signatures indicative of SV (samtools, IGV, ngsParalog)
3) Call SNPs and genotypes (ANGSD, bcftools, GATK, freeBayes)

Phase 4. Characterize genetic variation and make biological inference
1) Estimate allele frequencies (ANGSD)
2) Population structure, demography, selection.

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) Quality control sites to ensure any genetic variation or (lack of genetic variation) at them is deemed reliable.
2) Screen for signatures indicative of SV (samtools, IGV, ngsParalog)
3) Call SNPs and genotypes (ANGSD, bcftools, GATK, freeBayes)

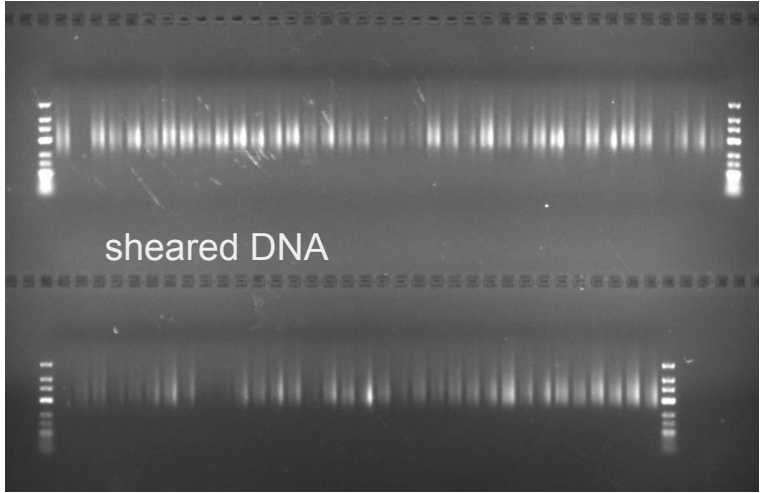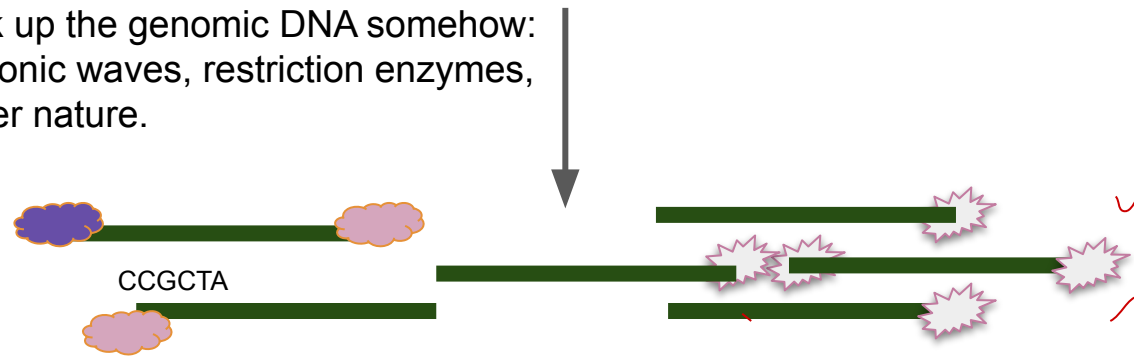**Phase 4. Characterize genetic variation and make biological inference**
1) Estimate allele frequencies (ANGSD)
2) Population structure, demography, selection.

Extracted DNA strand

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

sonicator

CCGCTA

sheared DNA

Extracted DNA strand

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

sonicator

CCGCTA

Adapter ligation

Sequencing into the adapter = adapter contamination

primer

insert

Phase 1: wetlab data generation

Extracted DNA strand

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

sonicator

CCGCTA

Adapter ligation

Sequencing into the adapter = adapter contamination

primer

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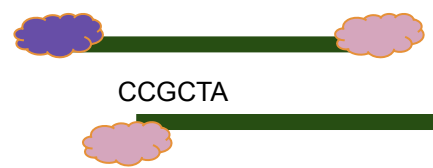
Break up the genomic DNA somehow:
ultrasonic waves, restriction enzymes,
mother nature.

single-end
independent reads

paired-end
two inwardly oriented
reads separated by ~200 nt

mate-paired
two outwardly oriented reads separated by ~3000 nt

Adapter ligation

Sequencing into the adapter
= adapter contamination

primer

sequencing

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) Quality control sites to ensure any genetic variation or (lack of genetic variation) at them is deemed reliable.
2) Screen for signatures indicative of SV (samtools, IGV, ngsParalog)
3) Call SNPs and genotypes (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

```
@K00188:273:HG3VVBBXX:5:1101:8532:1508 2:N:0:ACTGGAC
TGCTTTCCCATCATGTCGACACTCGGCGGAGTCAGATAATCCCAATCGCGGCCTNTGTTATAAGTGATCANTGTC
+
AAAFF-<A<AFFFJA-7-FAJJAJA-AA---A7<-<FFJFFFFF7FJ-FJJJAJ#F77FJJ<7FJ<<JJF#<JFF
@K00188:273:HG3VVBBXX:5:1101:10338:1508 2:N:0:ACTGGAC
GTTTTGGGCACCGGTGCTCAGGAAGGATATAATGGAACTAGAGAGAGTACAAAGNAGGGCAACAAAATTANTAAA
+
AAFFFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFJJJJJJJJJ#JJJJJJJJJJJJJJJ#JJJJ
@K00188:273:HG3VVBBXX:5:1101:24586:1508 2:N:0:ACTGGAC
AGACTTGACCGTGTNGAAGATGTGANGGTGGCAAAGTGANTGGAGNCATTATCCNCTATCCAACNAACAANCTTT
+
AAFFFJJJJJJJJJ#JJJJJJJJJJ#JJJJJJJJJJJJJ#JJJJJ#JJJJJJJJ#JFJJJFJJJ#JJJJJ#JJJJ
@K00188:273:HG3VVBBXX:5:1101:25418:1508 2:N:0:ACTGGAC
GCTGCTGTTTGGCCNAGTTTTGTGCNTGTCCTAAAATCCNGAATCNGGTTATATNAGTAGCAGCNTGTGCNTTAA
+
AAAFFFJFJJJFJJ#JJJJJJFJFF#JFJJJJFFJJJJJ#JJJFF#JFFJJJJF#JJJJJJFJJ#JJFJJ#<<FJ
@K00188:273:HG3VVBBXX:5:1101:26007:1508 2:N:0:ACTGGAC
TTCTCGCTTTATCCACGTGTTGACCNAGTGATTTATTCTNATATTNGCAGATCGNAACTCGCAANACGTCNTTCC
```

Read ID

Sequence

base quality scores

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

TTATAAGTGATCANTGTC

7FJJ<7FJ<<JJF#<JFF

GGCAACAAAATTANTAAA

JJJJJJJJJJJJ#JJJJ

ATCCAACNAACAANCTTT

JJJFJJJ#JJJJJ#JJJJ

TAGCAGCNTGTGCNTTAA

JJJJFJJ#JJFJJ#<<FJ

CTCGCAANACGTCNTTCC

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

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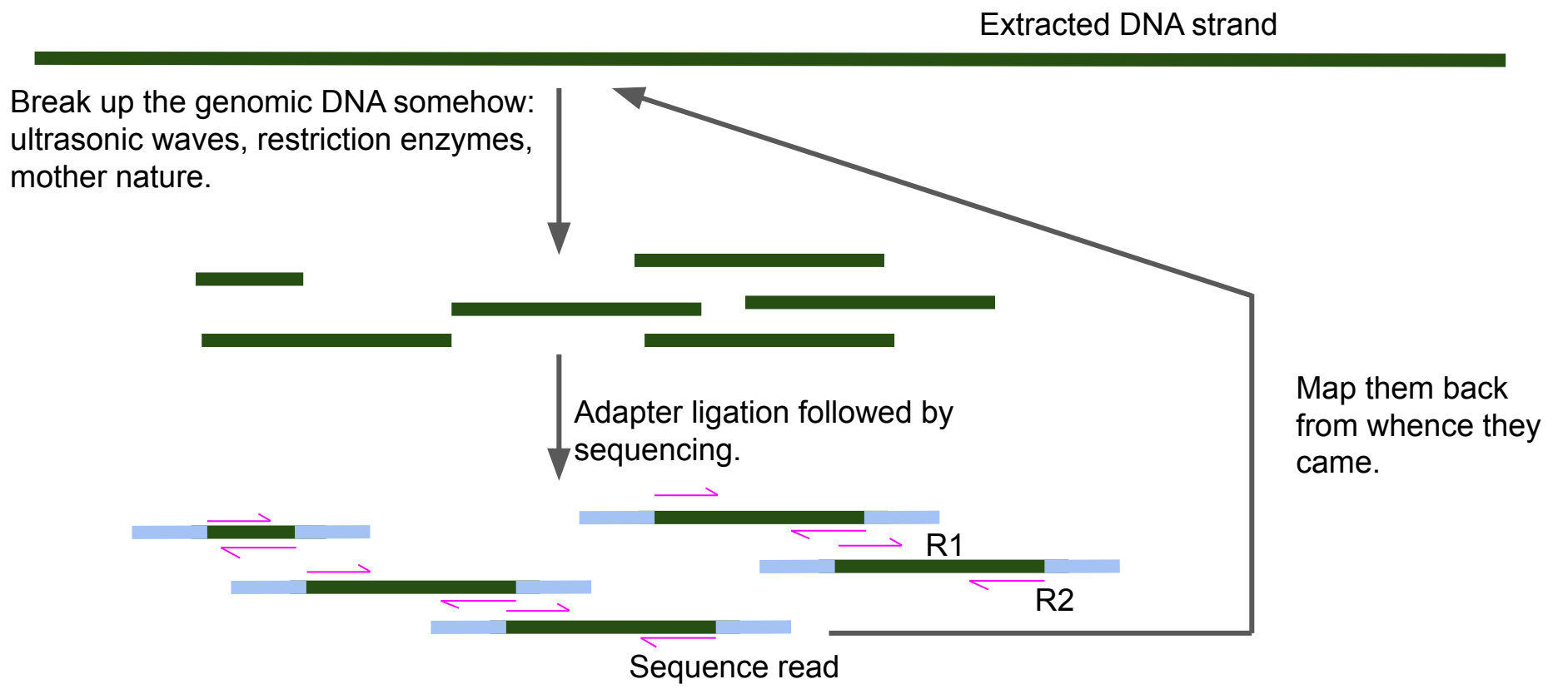$Qscore = -10*\log_{10}(\epsilon)$

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

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.
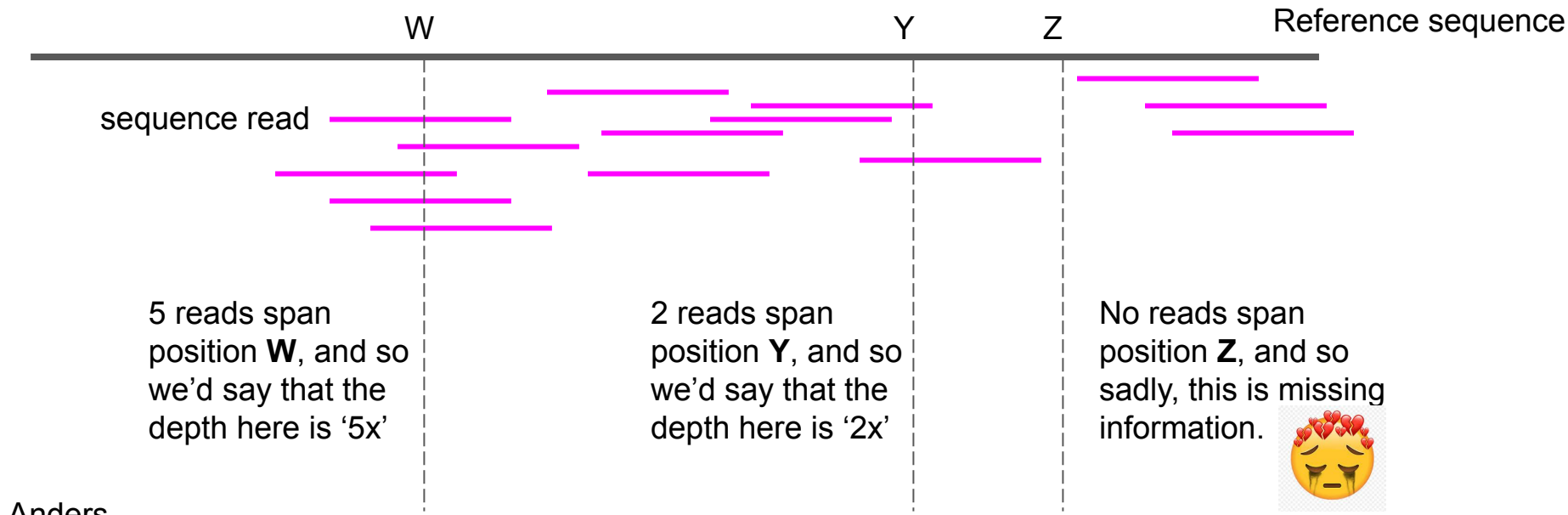
FASTQ quality issues.

- **Adapter contamination (this is the only quality issue I mostly worry about these days)**
- Low quality bases at the ends of reads (best handled by mappers like BWA through soft-clipping)
- Contamination
- Low Complexity reads (this is best handled by the mapper these days)
- Excess missing bases ('N')

Extracted DNA strand

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

Adapter ligation followed by
sequencing.

Map them back
from whence they
came.

R1

R2

Sequence read

Phase 2: Mapping and alignment

W   Y   Z   Reference sequence

sequence read

5 reads span
position **W**, and so
we'd say that the
depth here is '5x'

2 reads span
position **Y**, and so
we'd say that the
depth here is '2x'

No reads span
position **Z**, and so
sadly, this is missing
information.

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    EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE    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    AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA    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    BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB    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:$^\wedge$*=][: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:$^\wedge$*=][: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	

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    EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE    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
TATCTTACTATTTTTC    AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA    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    AGGTGAAGAAGAGGTGAGCATCCAAGAGACGAGCAAGAGGACAGTTTCAAAGCAGACATTATGGCACAATAAAAGCTGTGTGAGCAGAGATTCATTCATATCTTTACT
ATTTTTCTTTAGCTTCT    BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB    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

Based on the first entry, the left-most coordinate of read HS36_21523:8:1214:17897:31562 mapped in a properly mapped pair to chr7:16 in the reference and was identical to the reference along all 125 of its bases. The mapping quality is not particularly high: 23, so it could potentially be mismapped (~0.5% probability).

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 .

Phase 2: Mapping and alignment

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

Phase 2: Mapping and alignment

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

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) Quality control sites to ensure any genetic variation or (lack of genetic variation) at them is deemed reliable.
2) Screen for signatures indicative of SV (samtools, IGV, ngsParalog)
3) Call SNPs and genotypes (ANGSD, bcftools, GATK, freeBayes)

Phase 4. Characterize genetic variation and make biological inference
1) Estimate allele frequencies (ANGSD)
2) Population structure, demography, selection.

Phase 2: Variant discovery

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=0.84
8991;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
0/0:0,12,125:4:4,0:0:0:142,0:0.994672,0.00532842,5.66837e-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=0.84
8991;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
```

Site-wide info

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 (excessive missing data).
- Excessively high total site sequencing depth (indicative of sites with problematic mapping due to collapsed repeats for example).
- 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 (addresses collapsed mapping)

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 the effect of quality control
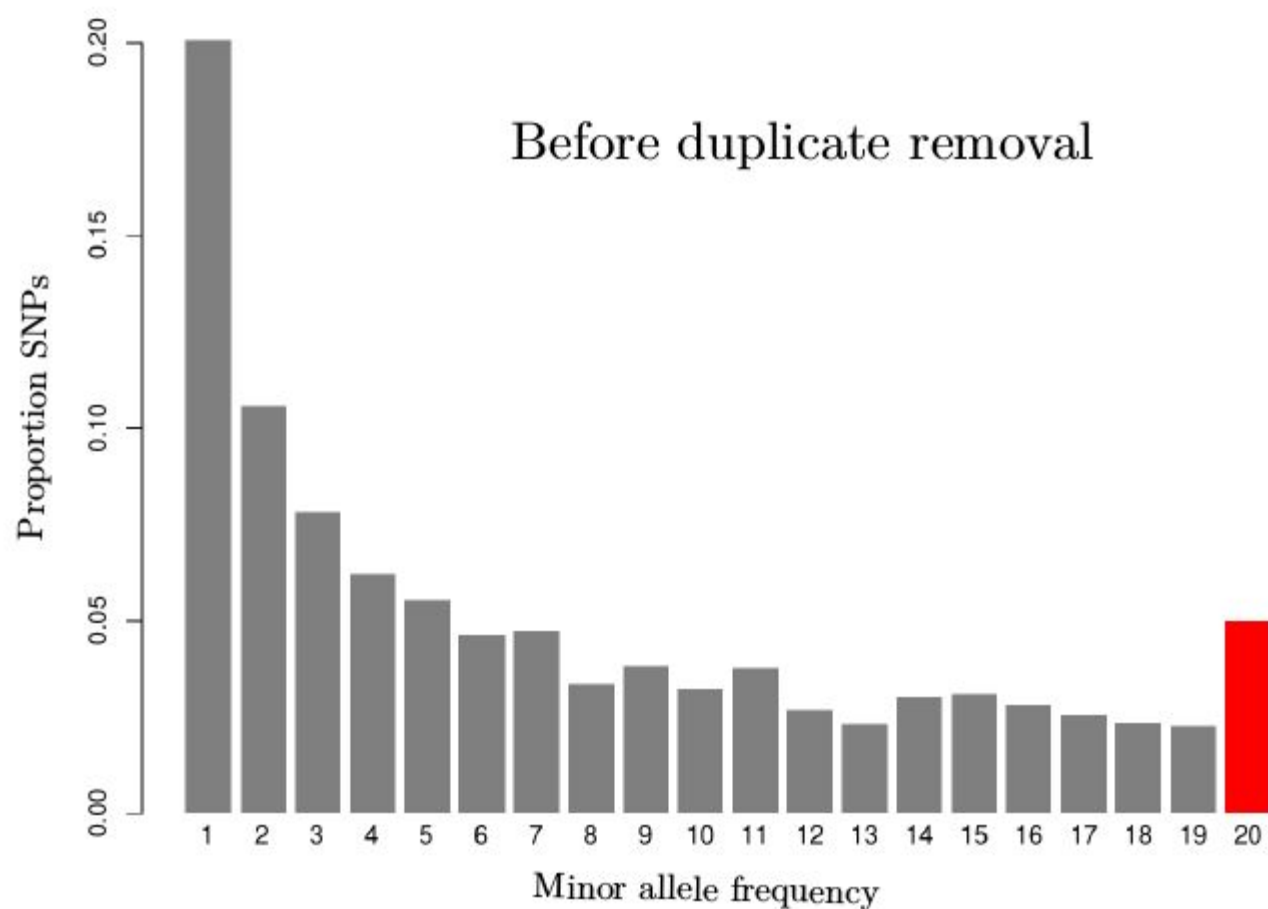
Now put it all into practice:

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