

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)

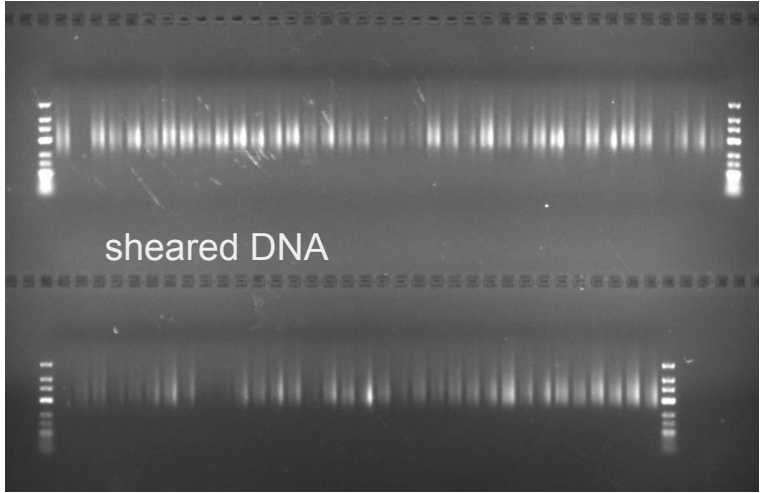
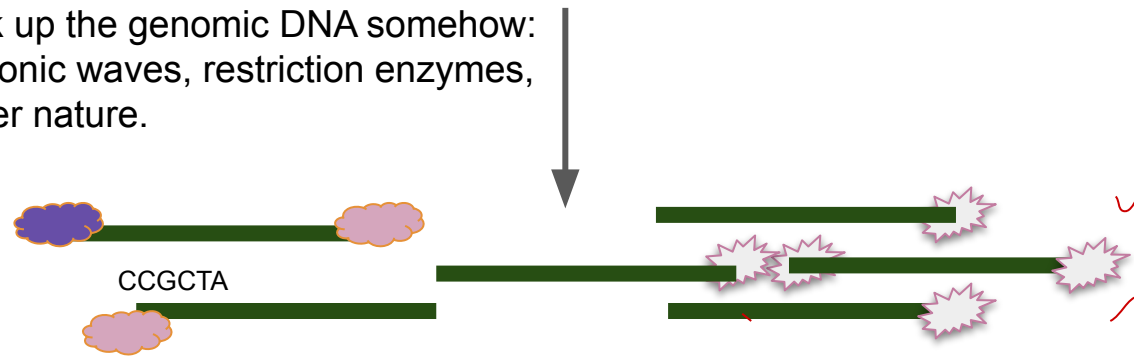
Phase 4. Characterize genetic variation and make biological inference

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

Phase 1: wetlab data generation

Extracted DNA strand

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



Phase 1: wetlab data generation

Extracted DNA strand

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



Sequencing into the adapter
= adapter contamination



insert



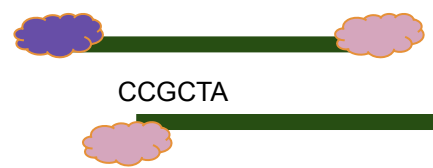
Adapter ligation



Phase 1: wetlab data generation

Extracted DNA strand

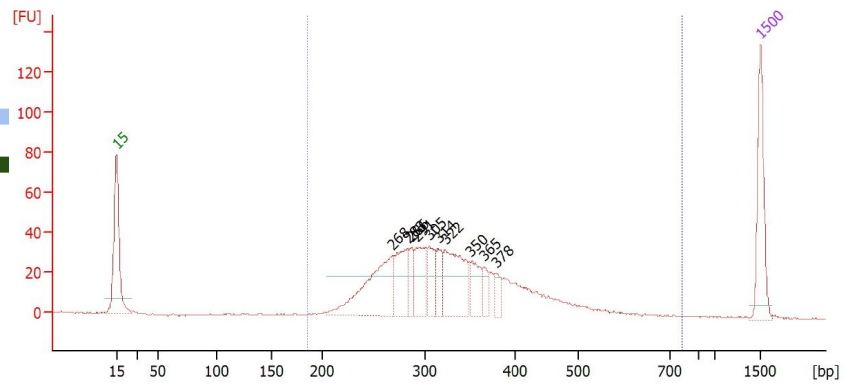
Break up the genomic DNA somehow:
ultrasonic waves, restriction enzymes,
mother nature.



Sequencing into the adapter
= adapter contamination



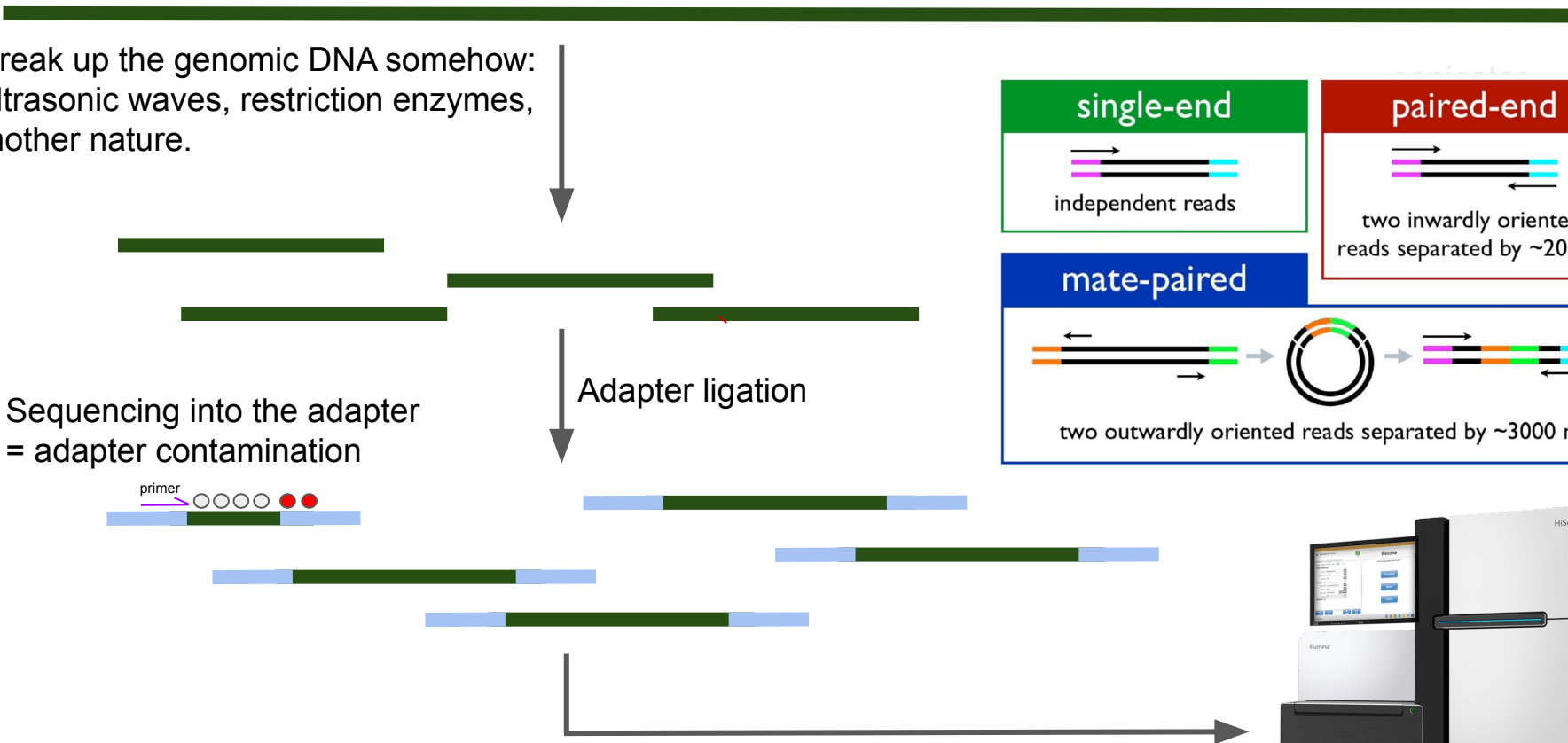
Adapter ligation



Phase 1: wetlab data generation

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

Extracted DNA strand



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

[illegible]

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

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

They are in Phred scale:

$$Qscore = -10 \cdot \log_{10}(\epsilon)$$

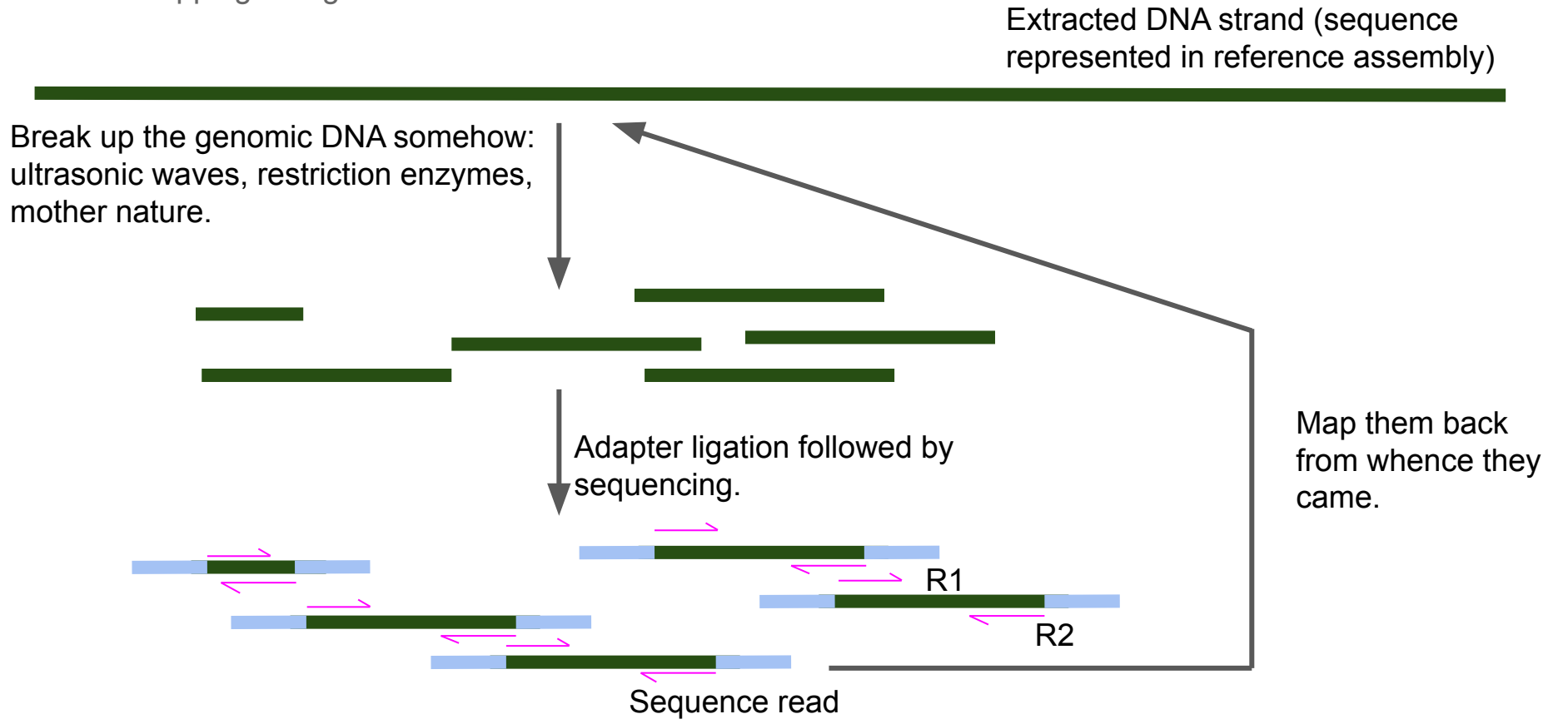
Given the quality score
then, $\epsilon = 10^{-Q/10}$

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.

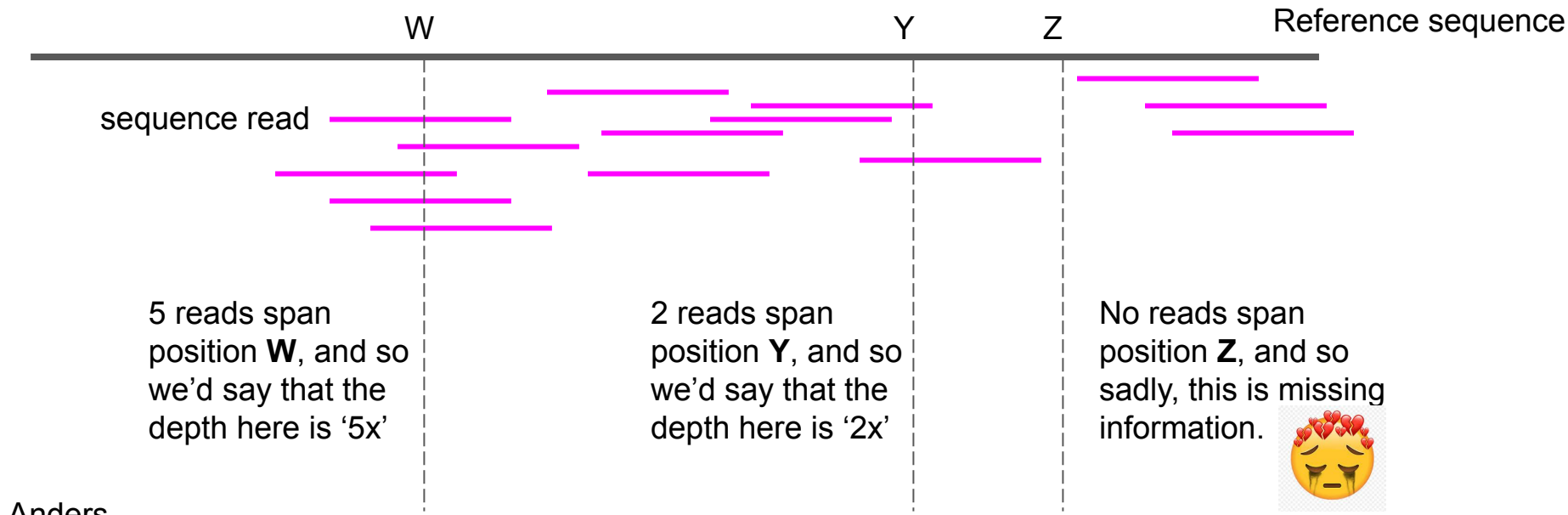
FASTQ quality issues.

- **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')

Phase 2: Mapping & Alignment



Phase 2: Mapping and alignment



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)

[illegible]

Col	Field	Type	Regex/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

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

Picard

build passing

Latest Jar
Release

Source Code
ZIP File

Source Code
TAR Ball

View On
GitHub

A set of command line tools (in Java) for manipulating high-throughput sequencing (HTS) data and formats such as SAM/BAM/CRAM and VCF.

Decoding SAM flags

This utility makes it easy to identify what are the properties of a read based on its SAM flag value, or conversely, to find what the SAM Flag value would be for a given combination of properties.

To decode a given SAM flag value, just enter the number in the field below. The encoded properties will be listed under Summary below, to the right.

SAM Flag:

Toggle first in pair / second in pair

Find SAM flag by property:

To find out what the SAM flag value would be for a given combination of properties, tick the boxes for those that you'd like to include. The flag value will be shown in the SAM Flag field above.

- ☒ read paired
- ☒ read mapped in proper pair
- ☐ read unmapped
- ☐ mate unmapped
- ☐ read reverse strand
- ☒ mate reverse strand
- ☐ first in pair
- ☒ second in pair
- ☐ not primary alignment
- ☐ read fails platform/vendor quality checks
- ☐ read is PCR or optical duplicate
- ☐ supplementary alignment

Summary:

read paired (0x1)
read mapped in proper pair (0x2)
mate reverse strand (0x20)
second in pair (0x80)

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

Phase 2: Mapping and alignment

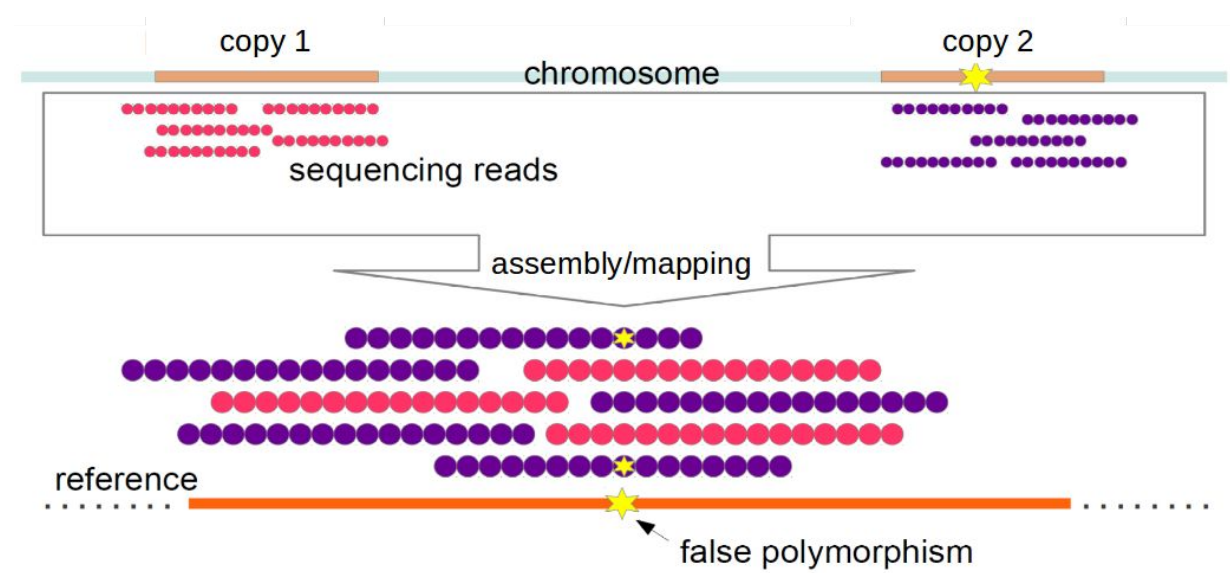
[illegible]

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).

Phase 2: Mapping and alignment

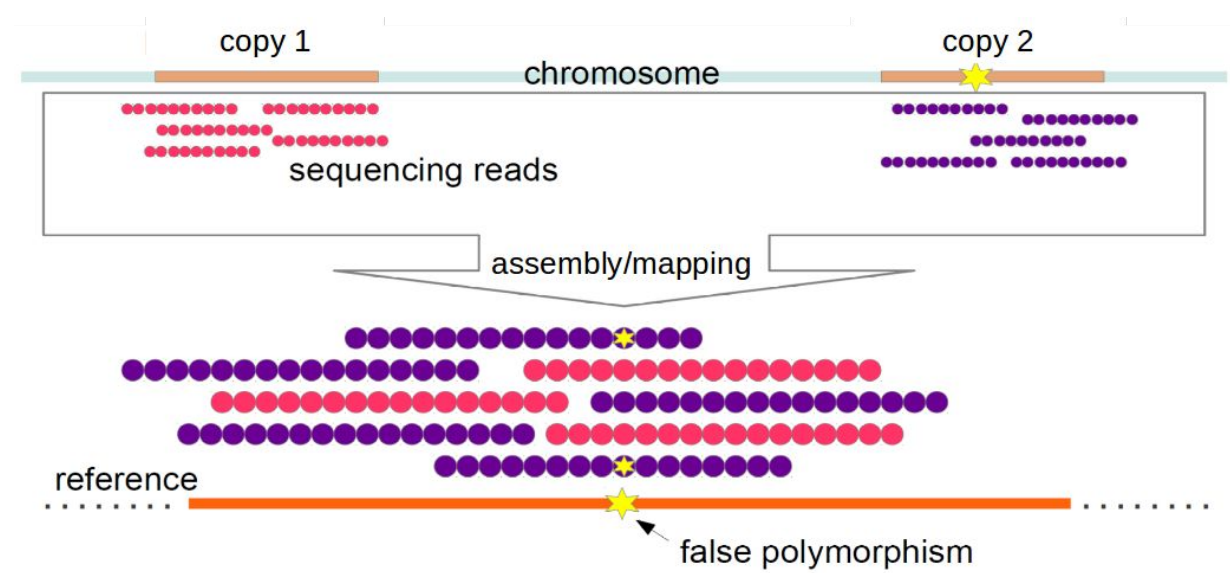
The problem of paralogs (or any other type of repeat sequence)



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

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	.,.,.,.,,	DABGIIIIII
chr7	10001	C	3	.,^].	EBB	7	.,.,.,.,	>IIBGGI	6	.,.,.,,	DEGEGG	9	.,.,.,.,,	DABGIIIIII
chr7	10002	G	3	.,.	EBB	7	.,.,.,.,	>IIBGGI	6	.,.,.,,	DEGEGG	9	.,.,.,.,,	DABGIIIIII
chr7	10003	G	3	.,.	EBB	7	.,.,.,.,	>IIBGGI	6	.,.,.,,	DEGEGG	10	.,.,.,.,.^].	DABGIIIIIE
chr7	10004	A	3	.,.	EBB	8	.,.,.,.,^].	>IIBGGIE	6	.,.,.,,	DEGEGG	10	.,.,.,.,.,	DABGIIIIIII
chr7	10005	G	3	.,.	EBB	8	.,.,.,.,.	>IIBGGIG	6	.,.,.,,	DEGEGG	10	.,.,.,.,.,	DABGIIIIIII
chr7	10006	A	3	.,.	EBB	8	.,.,.,.,.	>IIBGGIG	6	.,.,.,,	DEGEGG	10	.,.,.,.,.,	DABGIIIIIII
chr7	10007	G	3	.,.	EBB	8	.,.,.,.,.	>IIBGGIG	7	.,.,.,.^].	DEGEGGE	10	.,.,.,.,.,	DABGIIIIIII
chr7	10008	C	3	.,.	EBB	8	.,.,.,.,.	>IIBGGIG	7	.,.,.,.	DEGEGGG	10	.,.,.,.,.,	DABGIIIIIII
chr7	10009	A	3	.,.	EBB	8	.,.,.,.,.	>IIBGGIG	8	.,.,.,.^].	DEGEGGGB	10	.,.,.,.,.,	DABGIIIIIII
chr7	10010	G	3	.,.	EBB	8	.,.,.,.,.	>IIBGGIG	8	.,.,.,.,,	DEGEGGGB	10	.,.,.,.,.,	DABGIIIIIII
chr7	10011	C	3	.,.	EBB	8	.,.,.,.,.	>IIBGGIG	8	.,.,.,.,,	DEGEGGGB	10	.,.,.,.,.,	DABGIIIIIII
chr7	10012	T	3	.,.	EBB	8	.,.,.,.,.	>IIBGGIG	8	.,.,.,.,,	DEGEGGGB	10	.,.,.,.,.,	DABGIIIIIII
chr7	10013	T	3	.,.	EBB	8	.,.,.,.,.	>IIBGGIG	9	.,.,.,.,.^].	DEGEGGGBE	10	.,.,.,.,.,	DABGIIIIIII
chr7	10014	A	4	.,.^].	EBBE	8	.,.,.,.,.	>IIBGGIG	9	.G...gG,,	DEGEGGGBG	10	.,.,.,.,.,	D3BGIIIIIII
chr7	10015	G	4	.,.,.	EBBI	8	.,.,.,.,.	>IIBGGIG	9	.,.,.,.,,	DEGEGGGBG	10	.,.,.,.,.,	D3BGIIIIIII

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 reverse strand
- [ACGTacgt] = alternate allele (forward strand = uppercase, reverse strand = lowercase)

Phase 2: Mapping and alignment

Pileup format (generated with SAMtools)

chr7	10000	T	2	.,	EB	7	.,.,.,.^],	>IIBGGE	6	.,.,.,,	DEGEGG	9	.,.,.,.,,	DABGIIIIII
chr7	10001	C	3	.,^].	EBB	7	.,.,.,.,	>IIBGGI	6	.,.,.,,	DEGEGG	9	.,.,.,.,,	DABGIIIIII
chr7	10002	G	3	.,.	EBB	7	.,.,.,.,	>IIBGGI	6	.,.,.,,	DEGEGG	9	.,.,.,.,,	DABGIIIIII
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)

```
##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;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

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=0.848991;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

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.

Phase 2: Variant discovery

Site-level quality control

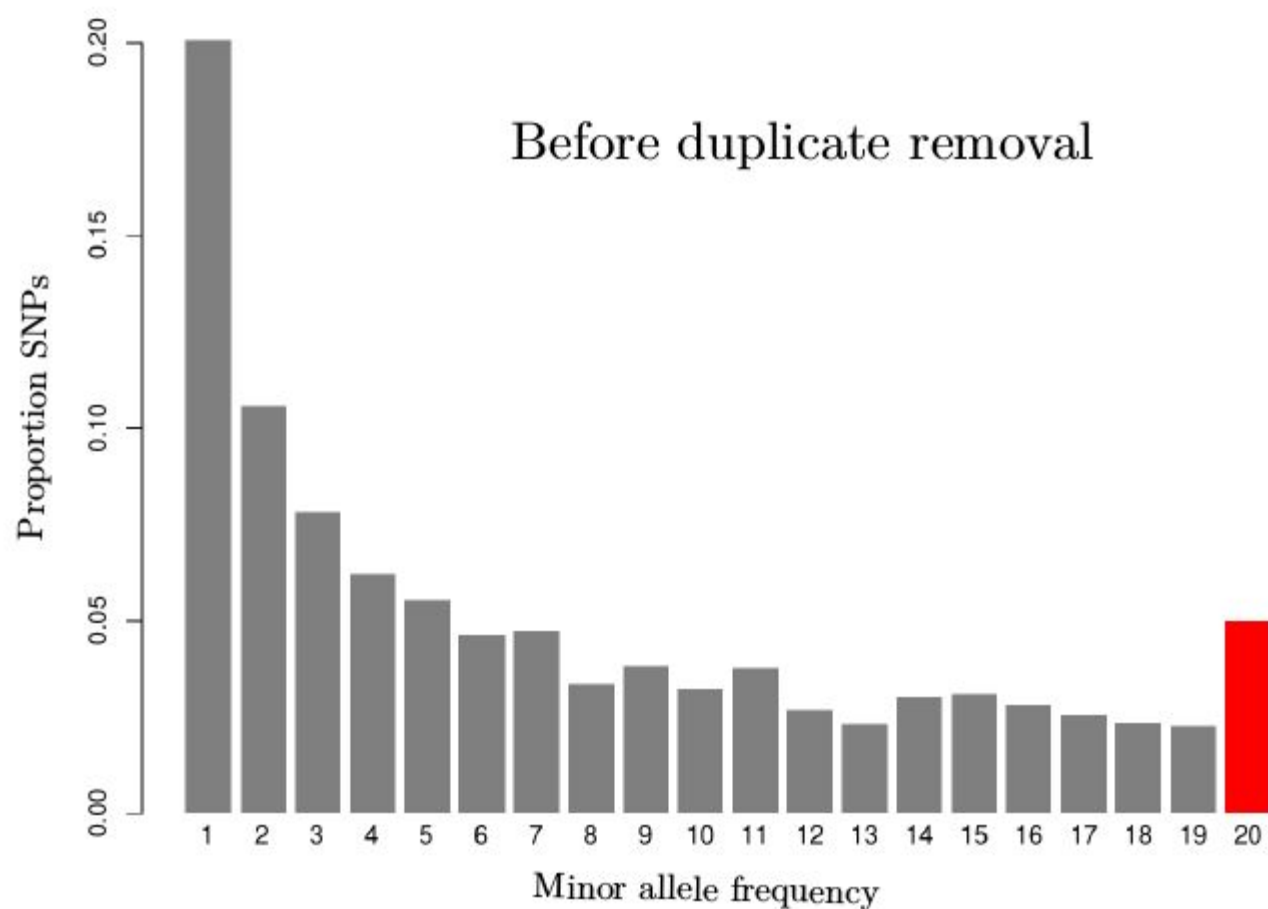
- Excessively low total site sequencing depth (excessive missing data).
- Excessively high total site sequencing depth (indicative of sites with problematic mapping due to collapsed repeats for example).
- Evenness of depth across individuals.
- Minimum number of individuals with called genotypes (sufficient 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

Phase 2: Variant discovery

Assessing the effect of quality control



Now put it all into practice:

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