



HTS data formats and Quality Control

petr.danecek@sanger.ac.uk



Data Formats

FASTQ

- ▶ Unaligned read sequences with base qualities

SAM/BAM

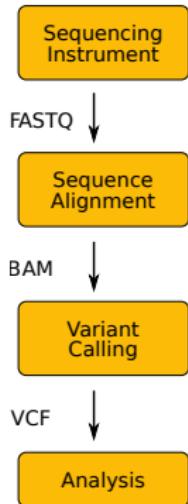
- ▶ Unaligned or aligned reads
- ▶ Text and binary formats

CRAM

- ▶ Better compression than BAM

VCF/BCF

- ▶ Flexible variant call format
- ▶ Arbitrary types of sequence variation
- ▶ SNPs, indels, structural variations



Specifications maintained by the Global Alliance for Genomics and Health

FASTA - reference genome

FASTA - reference genome

2003	NCBI Build 34	hg16
2004	NCBI Build 35	hg17
2006	NCBI Build 36.1	hg18
2009	GRCh37	hg19
2013	GRCh38	hg38

FASTQ

- ▶ Simple format for raw unaligned sequencing reads
- ▶ Extension to the FASTA file format
- ▶ Sequence and an associated per base quality score

Read 1

```
@ERR007731.739 IL16_2979:6:1:9:1684/1      ← Read name  
CTTGACGACTTAAAAATGACGAAATCACTAAAAACGTAAAAATGAGAAATG... ← Sequence  
+  
BBCBCBBBBBBBABBABBBBBBBABBBBBBBBBBABA...=@>BB... ← Base qualities  
@ERR007731.740 IL16_2979:6:1:9:1419/1  
AAAAAAAAAGATGTCATCAGCACATCAGAAAAGAAGGCAACTTTAAACTTT...  
+  
BBABBBBABABAABABABBABBBAAA>@B@BAA@4AAA>.>BAA@779:AAA@A...  
Read 2
```

- ▶ Quality encoded in ASCII characters with decimal codes 33-126
 - ▶ ASCII code of "A" is 65, the corresponding quality is $Q = 65 - 33 = 32$
 - ▶ Phred quality score: $P = 10^{-Q/10}$
`perl -e 'printf "%d\n", ord("A")-33;'`
- ▶ Beware: multiple quality scores were in use!
 - ▶ Sanger, Solexa, Illumina 1.3+
- ▶ Paired-end sequencing produces two FASTQ files

FASTQ

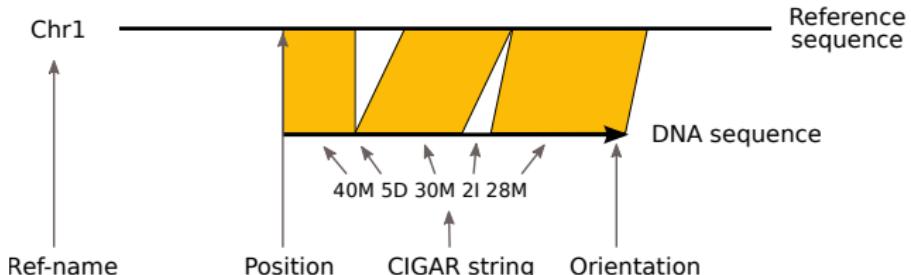
Quality	Probability of error	Accuracy
10 (Q10)	1 in 10	90%
20 (Q20)	1 in 100	99%
30 (Q30)	1 in 1000	99.9%
40 (Q40)	1 in 10000	99.99%

ASCII Table

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

SAM (Sequence Alignment/Map) format

- ▶ Single unified format for storing read alignments to a reference genome
- ▶ Developed by the 1000 Genomes Project group in 2009
- ▶ One record (a single DNA fragment alignment) per line describing alignment between fragment and reference
- ▶ 11 fixed columns + optional key:type:value tuples



Note that BAM can contain

- ▶ unmapped reads
- ▶ multiple alignments of the same read
- ▶ supplementary (chimeric) reads

SAM

```
$ samtools view -h file.bam | less

@HD VN:1.0 GO:none SO:coordinate
@SQ SN:1 LN:249250621 UR:hs37d5.fa.gz AS:NCBI37 M5:1b22b98cdeb4a9304cb5d48026a85128 SP:Human
@SQ SN:2 LN:243199373 UR:hs37d5.fa.gz AS:NCBI37 M5:a0d9851da00400dec1098a9255ac712e SP:Human
@RG ID:1 PL:ILLUMINA PU:13350_1 LB:13350_1 SM:13350_1 CN:SC
@PG ID:bwa PN:bwa VN:0.7.10-r806 CL:bwa mem hs37d5.fa.gz 13350_1_1.fq 13350_1_1.fq
1:2203:10256:56986 97 1 9998 20 106M45S = 10335 0 \
CCATAACCTTAACCTTAACCTAACCATAGCCCTAACCTAACCTAACCTAACCT[...]CAAACCCACCCCCAAACCCAAAACCTCACCAC \
FFFFFJJJJJJFJJFJAJJJJ-JJAAAJFJJFFJJF<JJFFJJJJFJJJJFF[...]<---F----A7-J-<J-A--77AF---J7-- \
MD:Z:1G24C2A76 PG:Z:MarkDuplicates RG:Z:1 NM:i:3 MQ:i:0 AS:i:94 XS:i:94
```

SAM fields

1	QNAME	Query NAME of the read or the read pair
2	FLAG	Bitwise FLAG (pairing, strand, mate strand, etc.)
3	RNAME	Reference sequence NAME
4	POS	1-Based leftmost POSition of clipped alignment
5	MAPQ	MAPping Quality (Phred-scaled)
6	CIGAR	Extended CIGAR string (operations: MIDNSHPX=)
7	MRNM	Mate Reference NaMe ('=' if same as RNAME)
8	MPOS	1-Based leftmost Mate POSition
9	ISIZE	Inferred Insert SIZE
10	SEQ	Query SEQuence on the same strand as the reference
11	QUAL	Query QUALity (ASCII-33=Phred base quality)
12-	OTHER	Optional fields

CIGAR string

Compact representation of sequence alignment

- M alignment match or mismatch
- = sequence match
- X sequence mismatch
- I insertion to the reference
- D deletion from the reference
- S soft clipping (clipped sequences present in SEQ)
- H hard clipping (clipped sequences NOT present in SEQ)
- N skipped region from the reference
- P padding (silent deletion from padded reference)

Examples:

Ref: ACGTACGTACGTACGT

Read: ACGT~~----~~ACGTACGA

Cigar: 4M 4D 8M

Ref: ACGT~~----~~ACGTACGT

Read: ACGTACGTACGTACGT

Cigar: 4M 4I 8M

Ref: ACTCAGTG--GTATCGTTAAC

Read: ACGCA-TGCAGTTAGACGTACGT

Cigar: 5M 1D 2M 2I 2M 7S

CIGAR string

Compact representation of sequence alignment

- M alignment match or mismatch
- = sequence match
- X sequence mismatch
- I insertion to the reference
- D deletion from the reference
- S soft clipping (clipped sequences present in SEQ)
- H hard clipping (clipped sequences NOT present in SEQ)
- N skipped region from the reference
- P padding (silent deletion from padded reference)

Examples:

Ref: ACGTACGTACGTACGT

Read: ACGT~~----~~ACGTACGA

Cigar: 4M 4D 8M

Ref: ACGT~~----~~ACGTACGT

Read: ACGTACGTACGTACGT

Cigar: 4M 4I 8M

Ref: ACTCAGTG--GTATCGTTAAC

Read: ACGCA-TGCAGTTAGACGTACGT

Cigar: 5M 1D 2M 2I 2M 7S

Ref: TGTCGTCACGCATG---CAGTTTTTTAAAA

Read: ACGTACGAAGCATGCGCAGTACGACGTTCG

Cigar: ???

CIGAR string

Compact representation of sequence alignment

- M alignment match or mismatch
- = sequence match
- X sequence mismatch
- I insertion to the reference
- D deletion from the reference
- S soft clipping (clipped sequences present in SEQ)
- H hard clipping (clipped sequences NOT present in SEQ)
- N skipped region from the reference
- P padding (silent deletion from padded reference)

Examples:

Ref: ACGTACGTACGTACGT

Read: ACGT~~----~~ACGTACGA

Cigar: 4M 4D 8M

Ref: ACGT~~----~~ACGTACGT

Read: ACGTACGTACGTACGT

Cigar: 4M 4I 8M

Ref: ACTCAGTG--GTATCGTTAAC

Read: ACGCA-TGCAGTTAGACGTACGT

Cigar: 5M 1D 2M 2I 2M 7S

Ref: TGTCGTCACGCATG---CAGTTTTTTAAAA

Read: ACGTACGAAGCATGCGGCAGTACGACGTTCG

Cigar: 7H 7M 3I 4M 7S

Flags

Hex	Dec	Flag	Description
0x1	1	PAIRED	paired-end (or multiple-segment) sequencing technology
0x2	2	PROPER_PAIR	each segment properly aligned according to the aligner
0x4	4	UNMAP	segment unmapped
0x8	8	MUNMAP	next segment in the template unmapped
0x10	16	REVERSE	SEQ is reverse complemented
0x20	32	MREVERSE	SEQ of the next segment in the template is reversed
0x40	64	READ1	the first segment in the template
0x80	128	READ2	the last segment in the template
0x100	256	SECONDARY	secondary alignment
0x200	512	QCFAIL	not passing quality controls
0x400	1024	DUP	PCR or optical duplicate
0x800	2048	SUPPLEMENTARY	supplementary alignment

Bit operations made easy

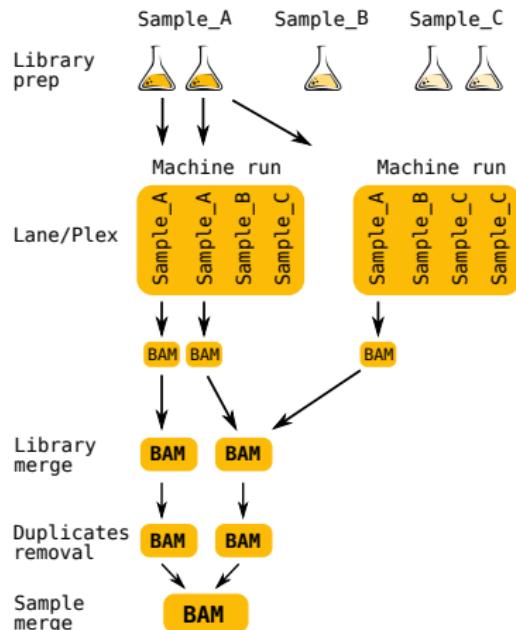
- ▶ **python**
0x1 | 0x2 | 0x20 | 0x80 .. 163
`bin(163) .. 10100011`
- ▶ **samtools flags**
`0xa3 163 PAIRED,PROPER_PAIR,MREVERSE,READ2`

Optional tags

Each lane has a unique RG tag that contains meta-data for the lane

RG tags

- ▶ ID: SRR/ERR number
- ▶ PL: Sequencing platform
- ▶ PU: Run name
- ▶ LB: Library name
- ▶ PI: Insert fragment size
- ▶ SM: Individual
- ▶ CN: Sequencing center



SAM/BAM tools

Several tools and programming APIs for interacting with SAM/BAM files

Samtools - Wellcome Sanger Institute (<http://www.htslib.org>)

- ▶ convert between SAM, BAM, CRAM
- ▶ sort, index
- ▶ flagstat - summary of the mapping flags
- ▶ merge multiple BAM files
- ▶ rmdup - remove PCR duplicates from the library preparation

Picard tools - Broad Institute (<https://www.broadinstitute.org/gatk/>)

- ▶ MarkDuplicates, CollectAlignmentSummaryMetrics, CreateSequenceDictionary, SamToFastq, MeanQualityByCycle, FixMateInformation etc.

Others

- ▶ Bio-SamTool - Perl (<http://search.cpan.org/~lds/Bio-SamTools/>)
- ▶ Pysam - Python (<https://github.com/pysam-developers/pysam>)
- ▶ R - Bioconductor/Rsamtools

BAM Visualisation

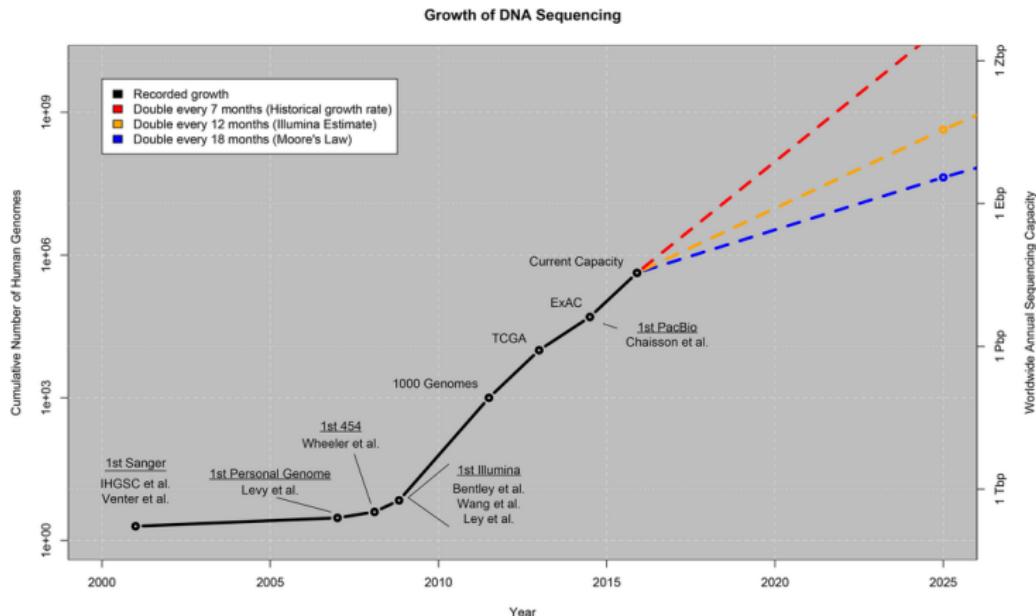
- ▶ IGV: <http://www.broadinstitute.org/igv/>
- ▶ BamView, LookSeq, Gap5, Tablet, Ensembl, UCSC, Bambino, Biodalliance...

Reference based Compression

BAM files are too large

- ~1.5-2 bytes per base pair

Increases in disk capacity are being far outstripped by sequencing technologies



Zachary D. Stephens, et al, Big Data: Astronomical or Genomical? DOI: 10.1371/journal.pbio.1002195

Reference based Compression

BAM files are too large

- ▶ ~1.5-2 bytes per base pair

Increases in disk capacity are being far outstripped by sequencing technologies

BAM stores all of the data

- ▶ Every read base
- ▶ Every base quality
- ▶ Using a single conventional compression technique for all types of data

Reference sequence:	ACGTACGTACGTACGTACGTACGTACGTAC
read 1:	ACGTACGTACGTACGTACGTGC
read 2:	TACGTACGCACTACGTGCGTA
read 3:	CGTACGCCACGTACGTACGTACG
read 4:	TACGTACGTACGTGCGTACGTAC
read 5:	CGCACGTACGTACGTACGTACG
read 6:	TACGTGCGTACGTACGTAC

Reference based Compression

BAM files are too large

- ▶ ~1.5-2 bytes per base pair

Increases in disk capacity are being far outstripped by sequencing technologies

BAM stores all of the data

- ▶ Every read base
- ▶ Every base quality
- ▶ Using a single conventional compression technique for all types of data

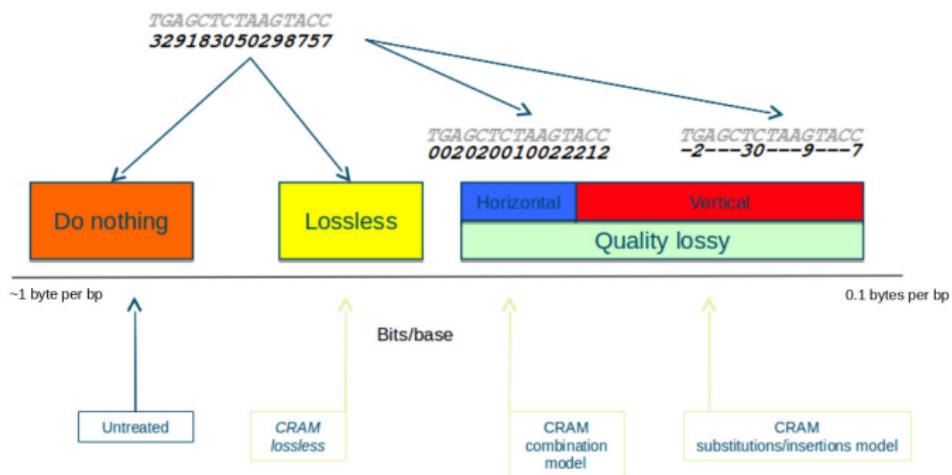
Reference sequence: ACGTACGTACGTACGTACGTACGTACGTAC
read 1: G.
read 2: C.....
read 3: C.....
read 4: G.....
read 5: C.....
read 6: G.....

CRAM

Three important concepts

- ▶ Reference based compression
- ▶ Controlled loss of quality information
- ▶ Different compression methods to suit the type of data, e.g. base qualities vs. metadata vs. extra tags

In lossless mode: 60% of BAM size

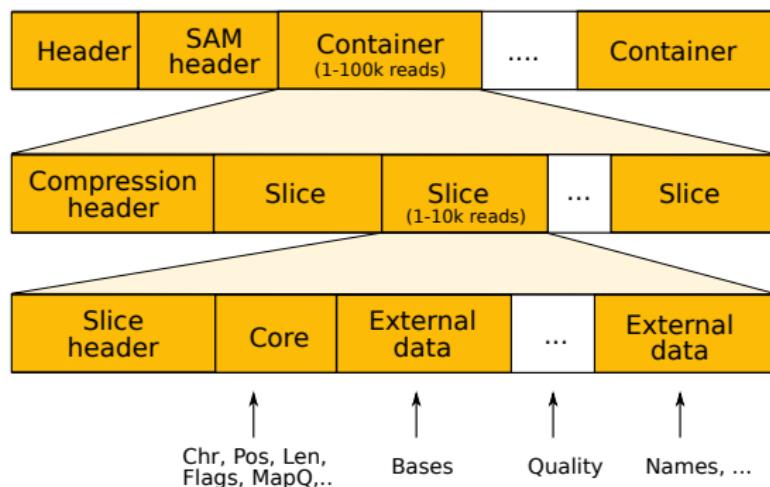


CRAM

Three important concepts

- ▶ Reference based compression
- ▶ Controlled loss of quality information
- ▶ Different compression methods to suit the type of data, e.g. base qualities vs. metadata vs. extra tags

In lossless mode: 60% of BAM size



CRAM

Three important concepts

- ▶ Reference based compression
- ▶ Controlled loss of quality information
- ▶ Different compression methods to suit the type of data, e.g. base qualities vs. metadata vs. extra tags

In lossless mode: 60% of BAM size

Support for CRAM

- ▶ added to Samtools/HTSlip in 2014, to GATK in 2015
- ▶ CRAM is now mature and used in production pipelines
 - ▶ all sequencing data by default in CRAM format
 - ▶ 40% disk space saving immediately

VCF: Variant Call Format



File format for storing genetic variation. Must be able to

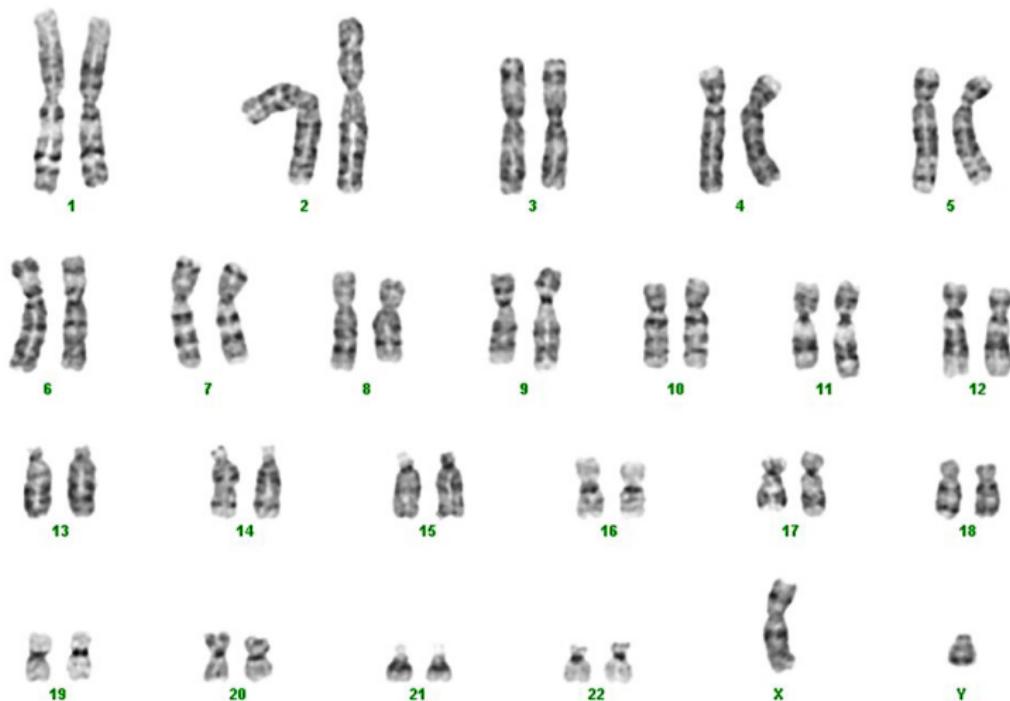
- ▶ accommodate all types of variation
 - ▶ SNPs, short insertions and deletions, large events
- ▶ multiple samples
- ▶ terse, parsable, user-extensible

Some terminology: alleles & genotype



Each individual...

Some terminology: alleles & genotype



...has 22 pairs of autosomal chromosomes...

Some terminology: alleles & genotype

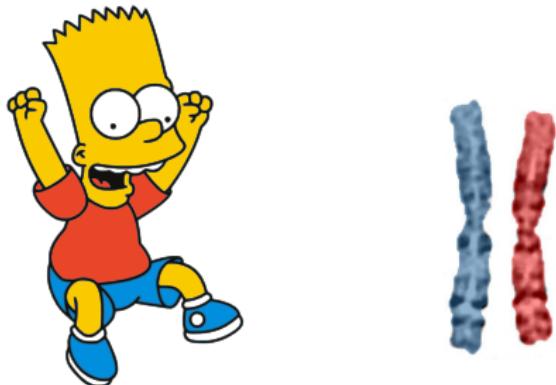


We inherit one copy from each parent

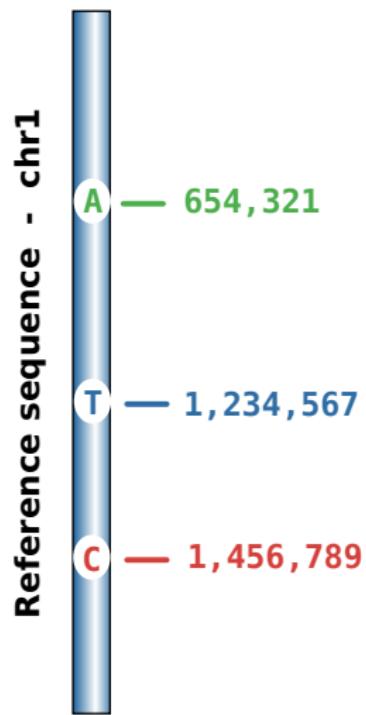
Some terminology: alleles & genotype



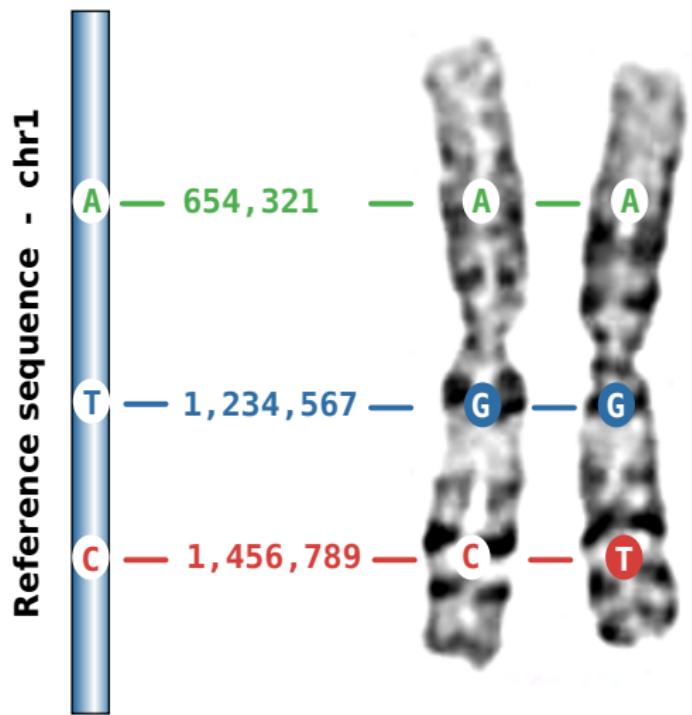
We inherit one copy from each parent



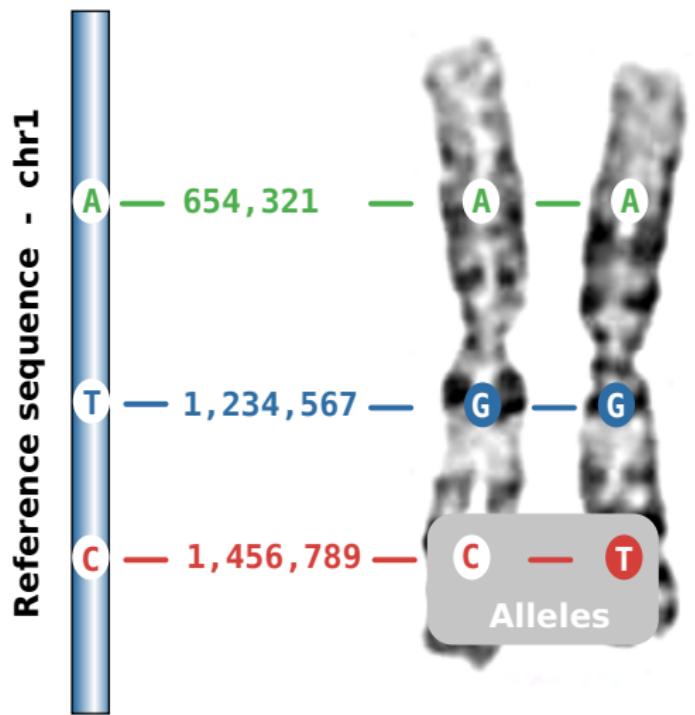
Some terminology: alleles & genotype



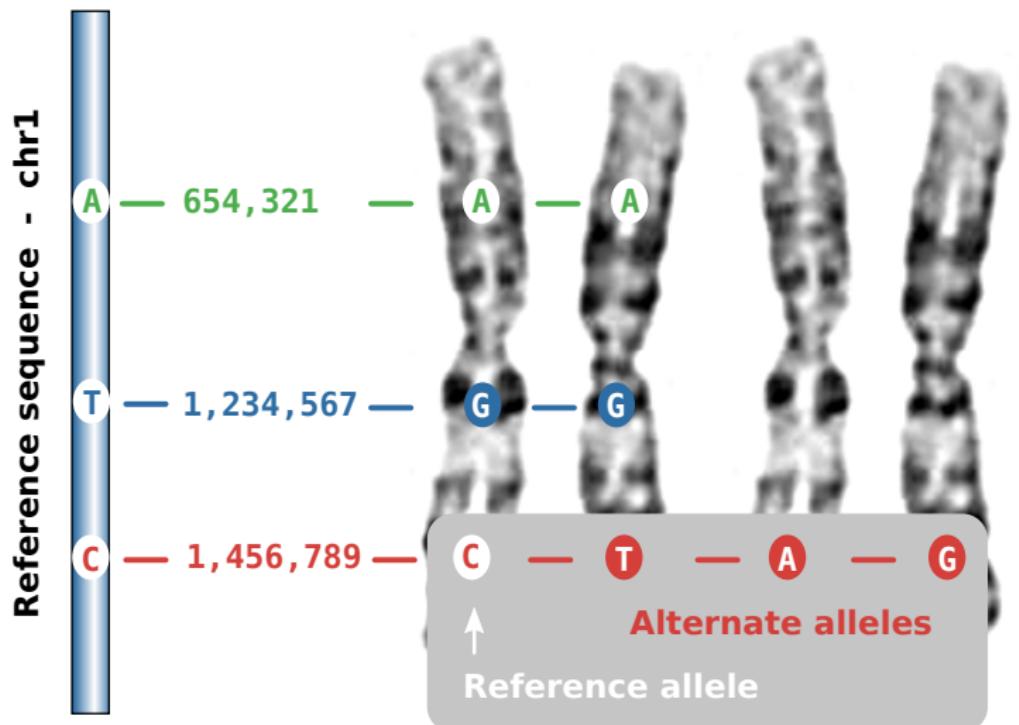
Some terminology: alleles & genotype



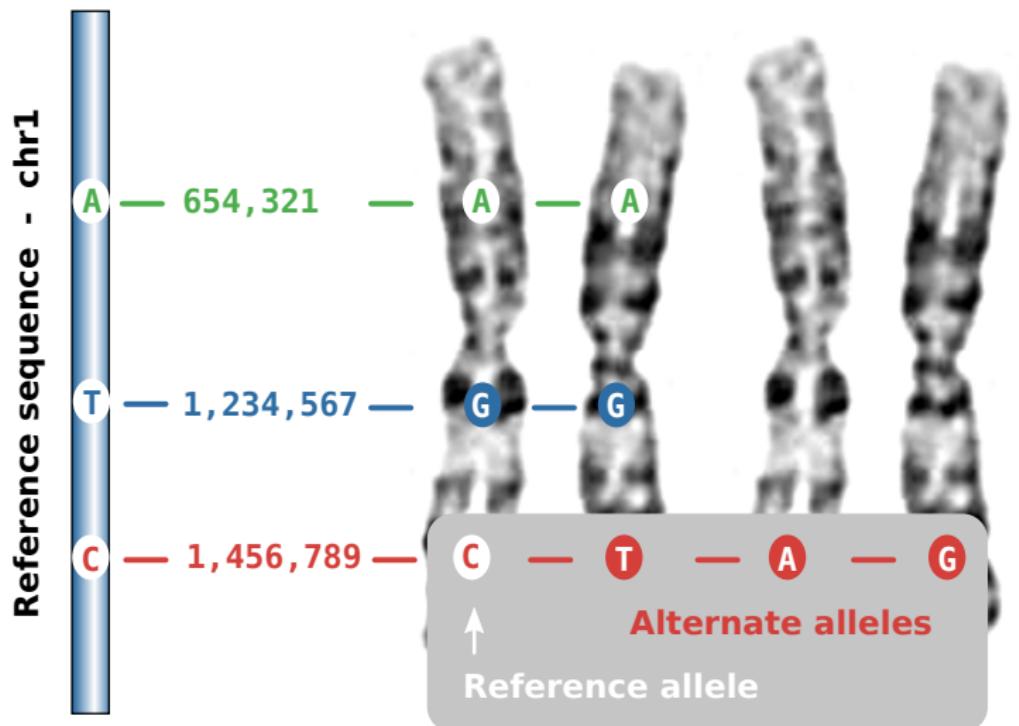
Some terminology: alleles & genotype



Some terminology: alleles & genotype

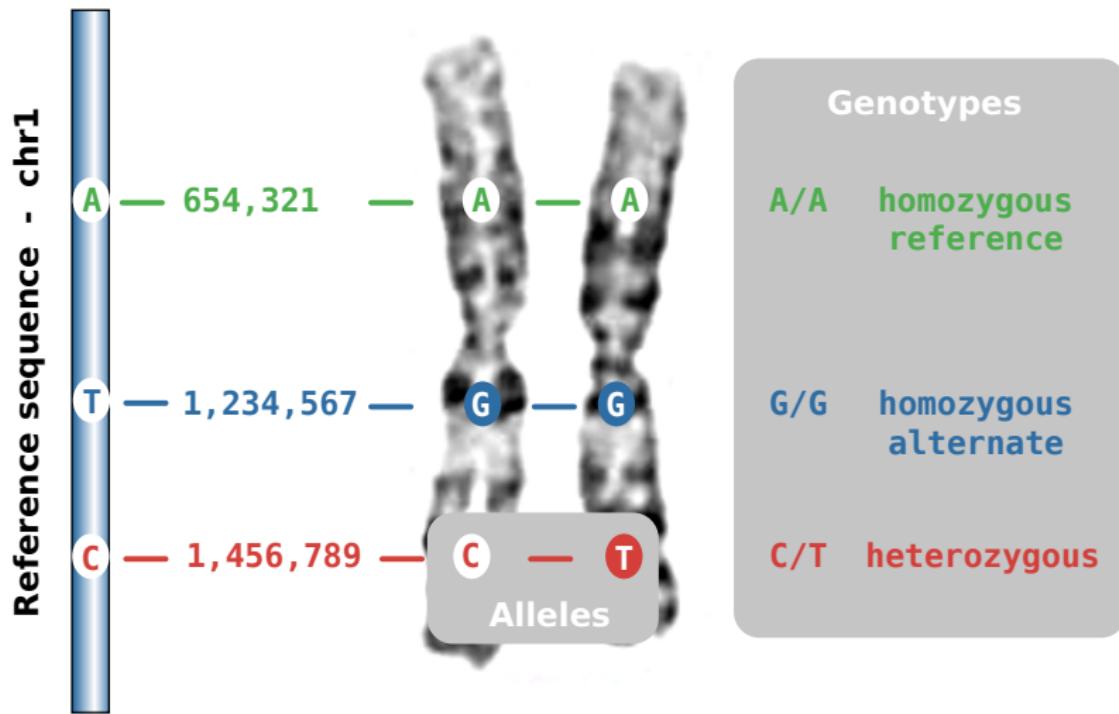


Some terminology: alleles & genotype



Q: What is the maximum number of alleles that can be observed at a site?

Some terminology: alleles & genotype



VCF notation

Allele numbering

CHR:POS	REF	ALT	S1	S2	S3
2:81170	A	C,T,G	A/A	A/C	T/G
	0,	1,2,3	0/0	0/1	2/3

VCF notation

Allele numbering

CHR:POS	REF	ALT	S1	S2	S3
2:81170	A	C,T,G	A/A	A/C	T/G
	0,	1,2,3	0/0	0/1	2/3

Columns

- ▶ chromosome and position
- ▶ variant ID (or “.” if not set)
- ▶ quality of the call (or “.”)
- ▶ soft filter (e.g. is the site low quality, low depth, etc)
- ▶ optional per-site information in the INFO column
- ▶ optional per-sample information in the FORMAT columns (one column per sample)

Optional tags

- ▶ DP .. sequencing depth (INFO/DP vs FORMAT/DP)
- ▶ GQ .. genotype quality (FORMAT/GQ)

VCF: Variant Call Format

File format for storing variation data

- ▶ Tab-delimited text, parsable by standard UNIX commands
- ▶ Flexible and user-extensible
- ▶ Compressed with BGZF (bgzip), indexed with TBI or CSI (tabix)

VCF header	Mandatory header lines
	##fileformat=VCFv4.0 ##fileDate=20100707 ##source=VCFtools ##reference=NCBI36 ##INFO=<ID=AA,Number=1>Type=String,Description="Ancestral Allele"> ##INFO=<ID=H2,Number=0>Type=Flag,Description="HapMap2 membership"> ##FORMAT=<ID=GT,Number=1>Type=String,Description="Genotype"> ##FORMAT=<ID=GQ,Number=1>Type=Integer,Description="Genotype Quality (phred score)"> ##FORMAT=<ID=GL,Number=3>Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)"> ##FORMAT=<ID=DP,Number=1>Type=Integer,Description="Read Depth"> ##ALT=<ID=DEL,Description="Deletion"> ##INFO=<ID=SVTYPE,Number=1>Type=String,Description="Type of structural variant"> ##INFO=<ID=END,Number=1>Type=Integer,Description="End position of the variant">
Body	Optional header lines (meta-data about the annotations in the VCF body)
	#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1 SAMPLE2
	1 1 . ACG A,AT . PASS . GT:DP 1/2:13 0/0:29 1 2 rs1 C T,CT . PASS H2;AA=T GT:GQ 0/1:100 2/2:70 1 5 . A G . PASS GT:GQ 1/0:77 1/1:95 1 100 . T . PASS SVTYPE=DEL;END=300 GT:GQ:DP 1/1:12:3 0/0:20
Deletion	Reference alleles (GT=0)
SNP	Alternate alleles (GT>0 is an index to the ALT column)
Large SV	Phased data (G and C above are on the same chromosome)
Insertion	Other event

VCF / BCF

VCFs can be very big

- ▶ compressed VCF with 3781 samples, human data:
 - ▶ 54 GB for chromosome 1
 - ▶ 680 GB whole genome

VCFs can be slow to parse

- ▶ text conversion is slow
- ▶ main bottleneck: FORMAT fields

```
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1 SAMPLE2 SAMPLE3 SAMPLE4 SAMPLE5
1 3 . A G . PASS AC=67;AN=5400;DP=2809 GT:PL:DP:GQ 1/1:0,9,73:26:22 0/0:0,9,73:13:31 0/0:0,9,73:48:99 1/0:255,0,75:32:15 1/0:255,0,75:32:15
1 4 . A T . PASS AC=15;AN=6800;DP=6056 GT:PL:DP:GQ 0/0:0,9,73:13:31 1/0:255,0,75:32:15 0/0:0,2,80:14:90 1/1:0,9,73:26:22 0/0:0,9,73:13:31
1 5 . C T . PASS AC=20;AN=6701;DP=5234 GT:PL:DP:GQ 1/0:255,0,75:32:15 0/0:0,2,170:14:90 1/1:0,9,73:13:31 0/0:0,6,50:13:80 0/0:0,2,80:14:90
1 6 . A G . PASS AC=67;AN=5400;DP=2809 GT:PL:DP:GQ 1/1:0,9,73:26:22 0/0:0,9,73:13:31 0/0:0,9,73:48:99 1/0:255,0,75:32:15 1/0:255,0,75:32:15
1 7 . A T . PASS AC=15;AN=6800;DP=6056 GT:PL:DP:GQ 0/0:0,9,73:13:31 1/0:255,0,75:32:15 0/0:0,2,80:14:90 1/1:0,9,73:26:22 0/0:0,9,73:13:31
```

BCF

- ▶ binary representation of VCF
- ▶ fields rearranged for fast access

CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE1	SAMPLE2	SAMPLE3	SAMPLE4	SAMPLE5
1	6	.	A	G	.	PASS	AC=67;AN=540	GT:PL:DP:GQ	1/1:0,9,73:26:22	0/0:0,9,73:13:31	0/0:0,9,73:48:99	1/0:255,0,75:32:15	1/0:255,0,75:32:15
1	6	.	A	G	.	PASS	AC=67;AN=540	GT:1/1:0:0:0:0:1:0:1:0	PL:0,9,73:0,9,73:0,9,73:255,0,75:255,0,75	DP:26:13:48:32:32	GQ:22:31:99:15:15		

gVCF

Often it is not enough not know *variant* sites only

- ▶ was a site dropped because of a reference call or because of missing data?
- ▶ we need evidence for both variant and non-variant positions in the genome

gVCF

- ▶ blocks of reference-only sites can be represented in a single record using the INFO/END tag
- ▶ symbolic alleles <*> for incremental calling
 - ▶ raw, "callable" gVCF
 - ▶ calculate genotype likelihoods only once (an expensive step)
 - ▶ then call incrementally as more samples come in

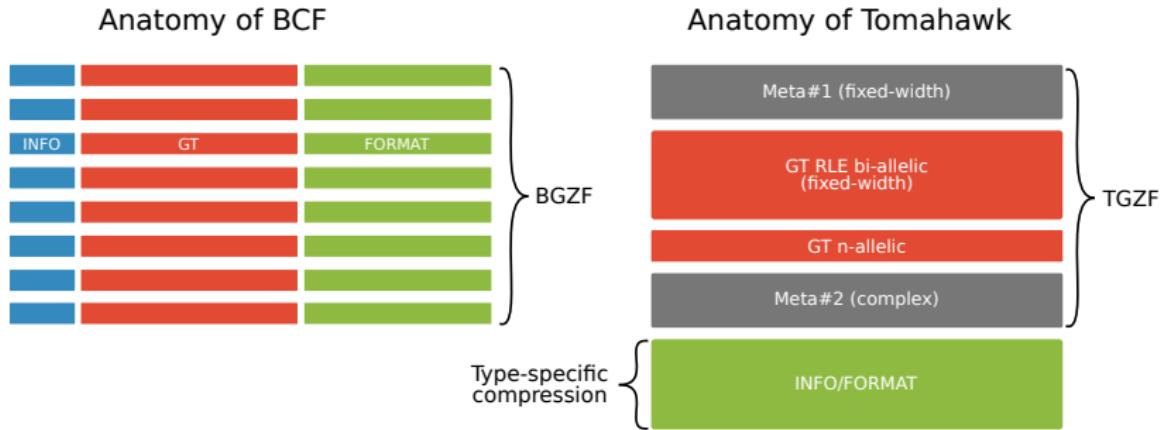
#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	Sample
19	9902	.	G	<*>	.	.	END=9915;MinDP=0	PL:DP	0,0,0:0
19	9916	.	C	<*>	.	.	END=9922;MinDP=5	PL:DP	0,15,137:5
19	9923	.	G	<*>	.	.	END=9948;MinDP=10	PL:DP	0,30,214:10
19	9949	.	G	A,<*>	.	.	DP=28	PL:DP	0,60,255,78,255,255:27
19	9950	.	C	<*>	.	.	END=9958;MinDP=28	PL:DP	0,84,255:28
19	9959	.	G	T,<*>	.	.	DP=34	PL:DP	0,82,255,99,255,255:34
19	9960	.	C	<*>	.	.	END=9969;MinDP=34	PL:DP	0,102,255:34

Symbolic "unobserved" allele
Represents any other possible alternate allele

A block of 10 sites with
at least 34 reference reads

Genotype likelihoods
for CC, C*, **

Optimizing variant calls for speed



New TWK format by Marcus Klarqvist (under development)

- ▶ BCF still too slow for querying hundreds of thousands and millions of samples
- ▶ bigger but 100x faster for certain operations on GTs

The commands I run:

```
 samtools stats file.bam > file.bam.stats  
 plot-bamstats -p plots/ file.bam.stats
```

The questions I want to answer:

- ▶ Do I have enough read coverage with my mapped reads?
- ▶ Was the library creation process efficient and problem-free?
- ▶ Did the sequencing process create artefacts?

Read coverage

Read coverage / depth

- ▶ is every genomic position “covered” to a sufficient depth?
- ▶ average depth: number-of-reads / target-size
 - ▶ the whole human genome .. target-size = 3Gb
 - ▶ the exomes .. target-size = 50Mb

Exomes

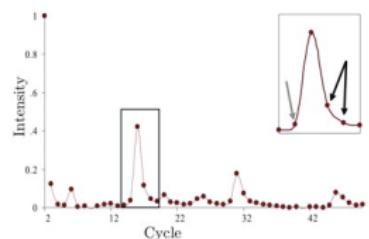
- ▶ be careful to distinguish between the total sequencing yield and on-target bases

Useful coverage

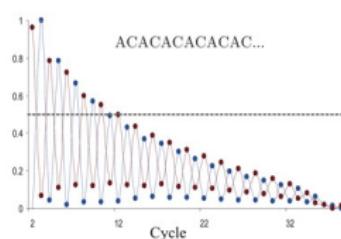
- ▶ 15x ok for common germline variants
- ▶ 30x ok for most things
- ▶ 100-200x for low VAF variants in tumors

Base calling errors

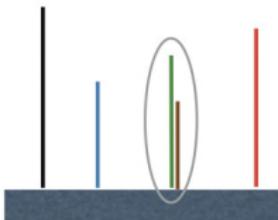
Phasing noise ϕ



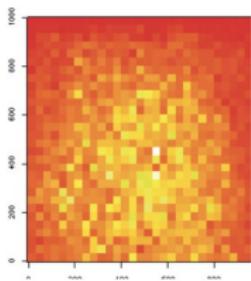
Signal Decay δ



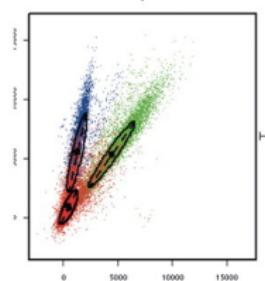
Mixed Cluster μ



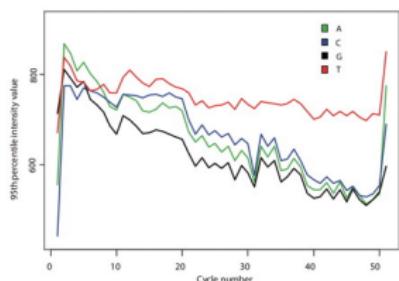
Boundary effects ω



Cross-talk Σ



T fluorophore accumulation τ

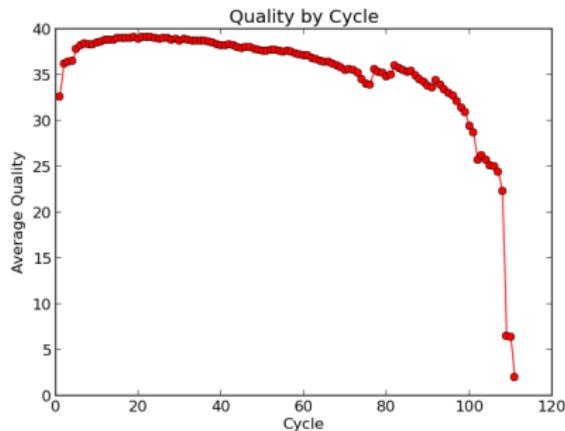


Base quality

Sequencing by synthesis: dephasing

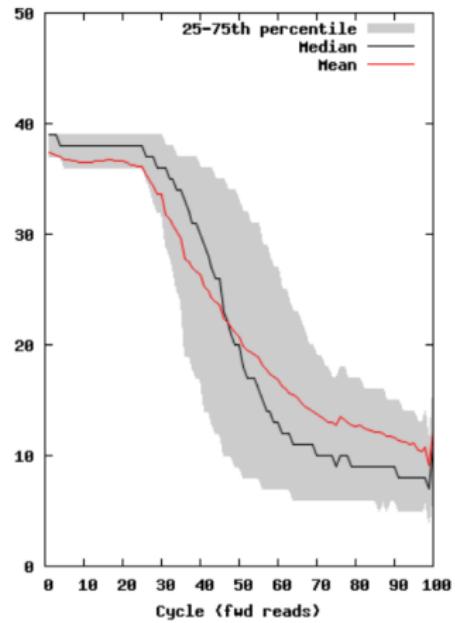
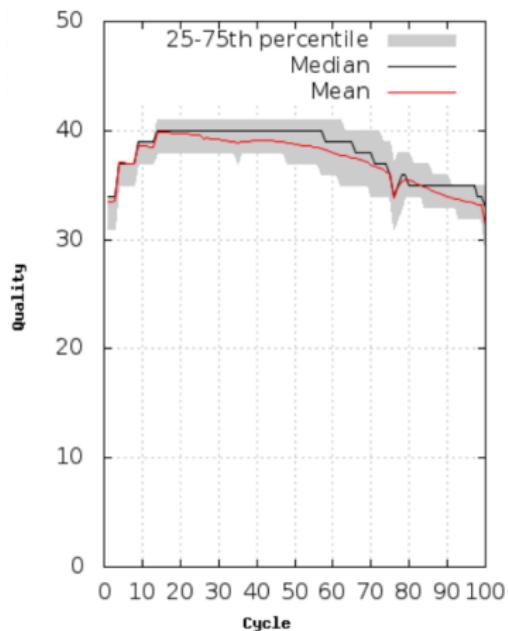
- ▶ growing sequences in a cluster gradually desynchronize
- ▶ error rate increases with read length

Calculate the average quality at each position across all reads

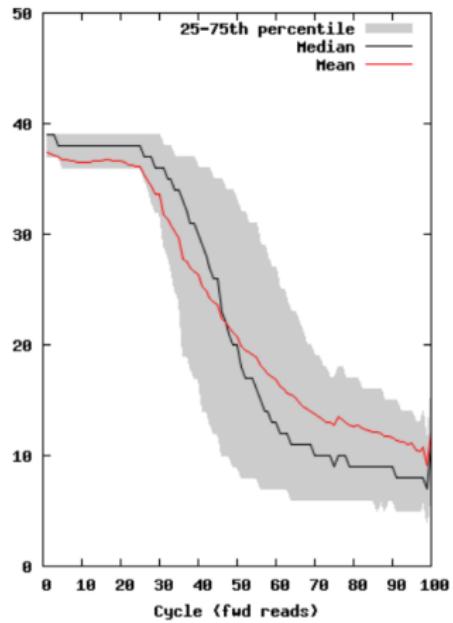
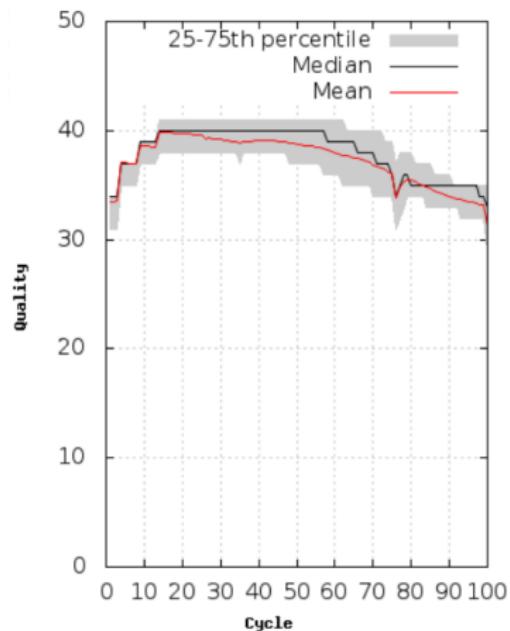


Quality	Probability of error	Call Accuracy
10 (Q10)	1 in 10	90%
20 (Q20)	1 in 100	99%
30 (Q30)	1 in 1000	99.9%
40 (Q40)	1 in 10000	99.99%

Base quality



Base quality



Library prep biases: PCR duplicates

Experiments start with small amounts of DNA

- ▶ a PCR amplification step is necessary for Illumina sequencing: one molecule => many identical molecules

Problem:

- ▶ additional PCR-copy molecules are not informative

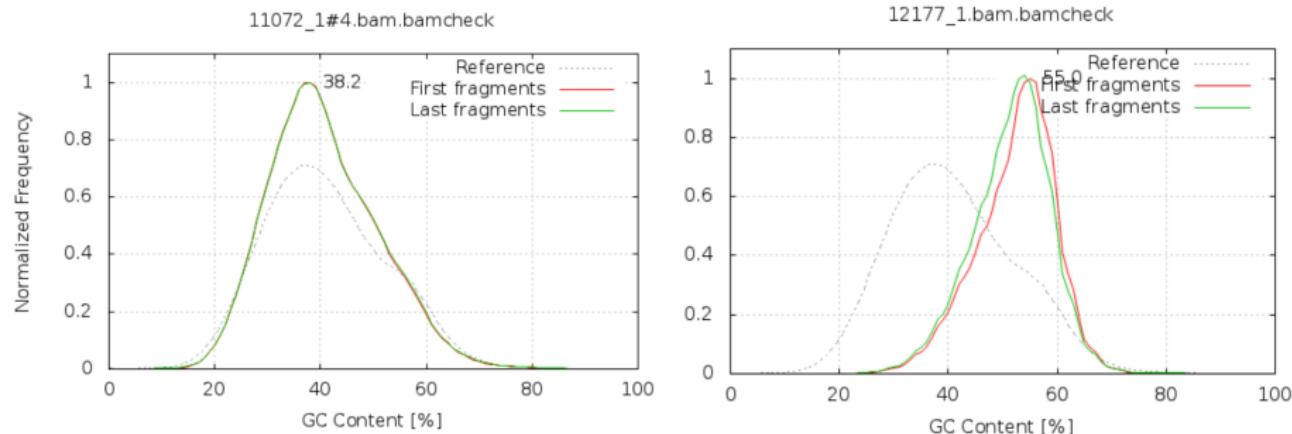
Solution:

- ▶ infer and mark PCR-duplicates, discount in later analysis
 - ▶ mark if reads and their mates start at the same position
- ▶ use `picard MarkDuplicates` or `samtools markdup`
- ▶ typical dup rates: Exomes ~ 15-20%, Genomes < 5%

GC bias

GC- and AT-rich regions are more difficult to amplify

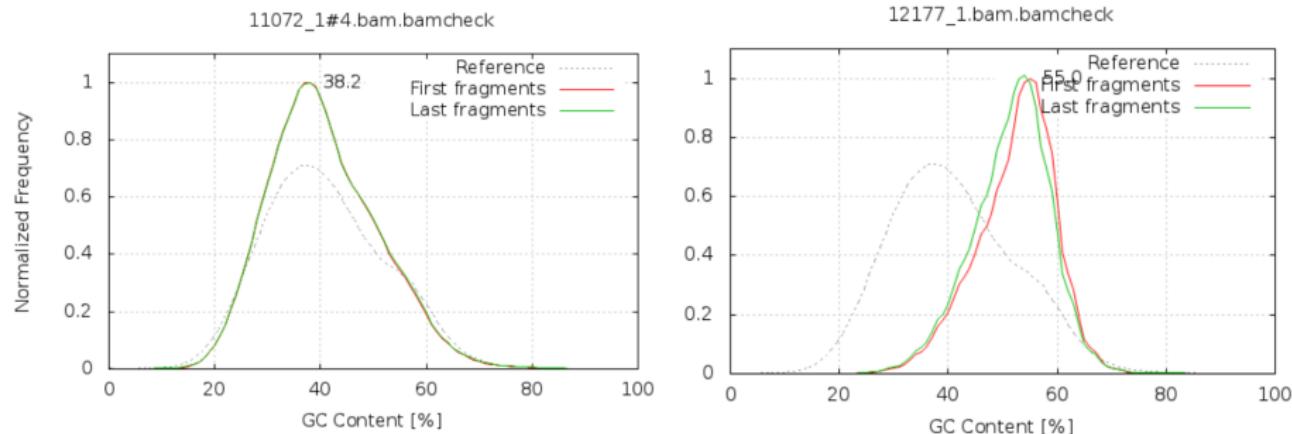
- ▶ compare the GC content against the expected distribution (reference sequence)



GC bias

GC- and AT-rich regions are more difficult to amplify

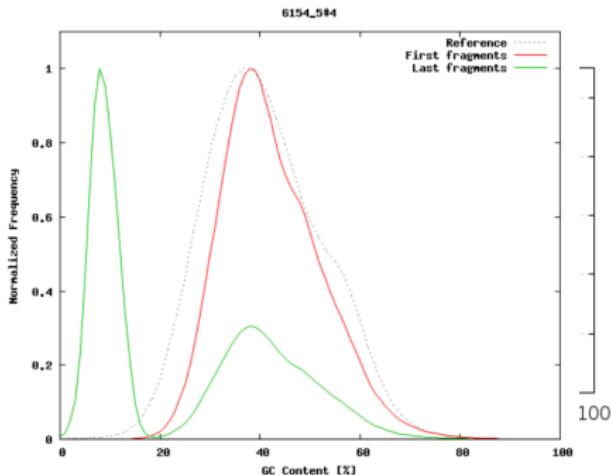
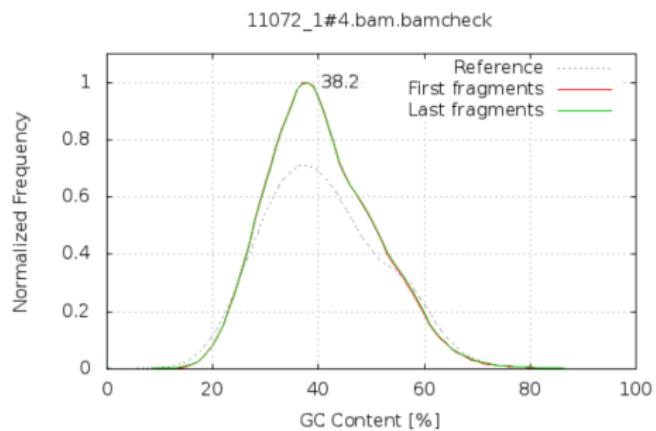
- ▶ compare the GC content against the expected distribution (reference sequence)



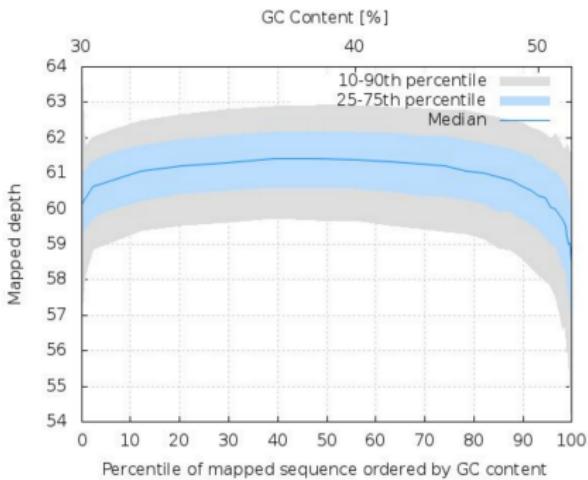
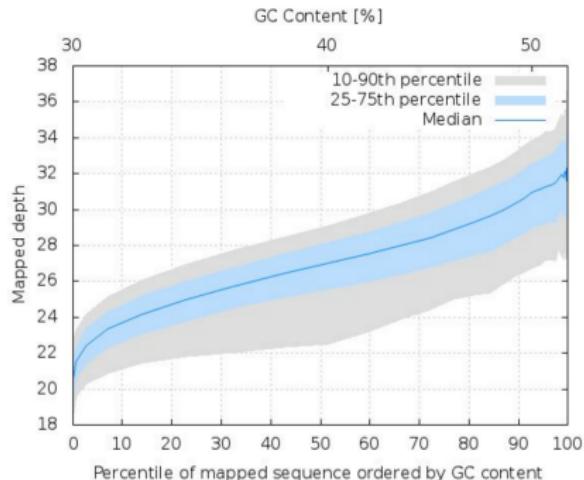
GC bias

GC- and AT-rich regions are more difficult to amplify

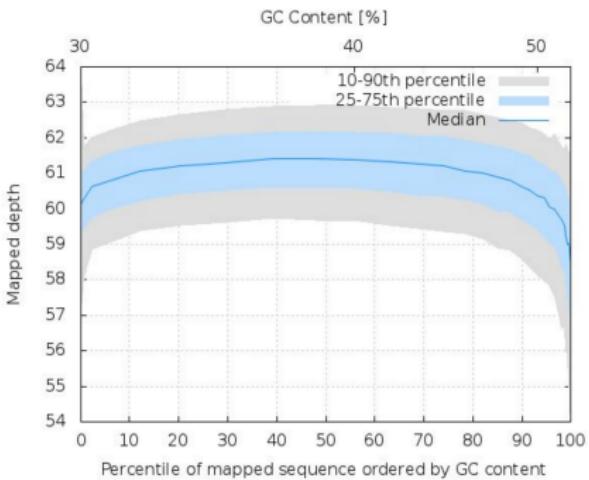
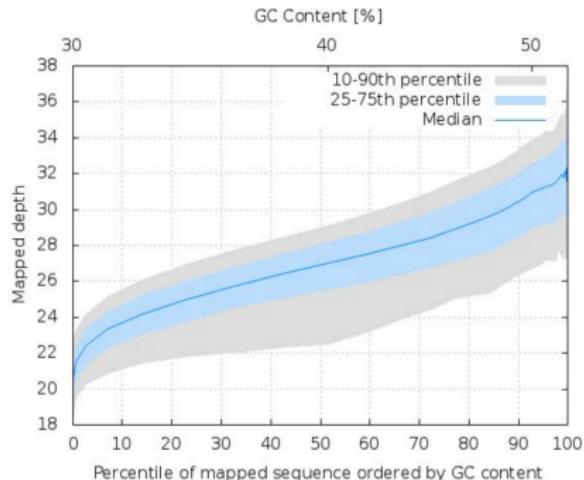
- ▶ compare the GC content against the expected distribution (reference sequence)



GC content vs depth

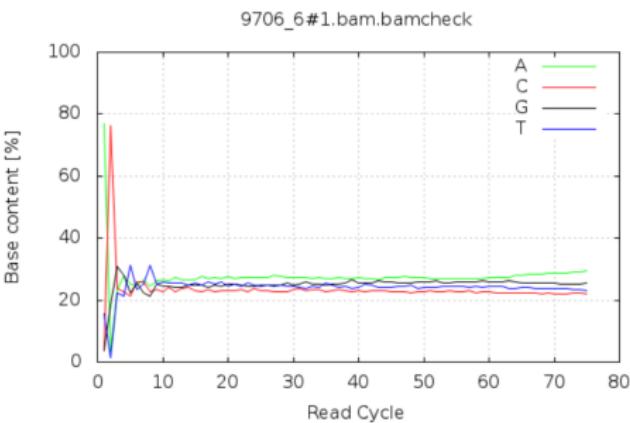
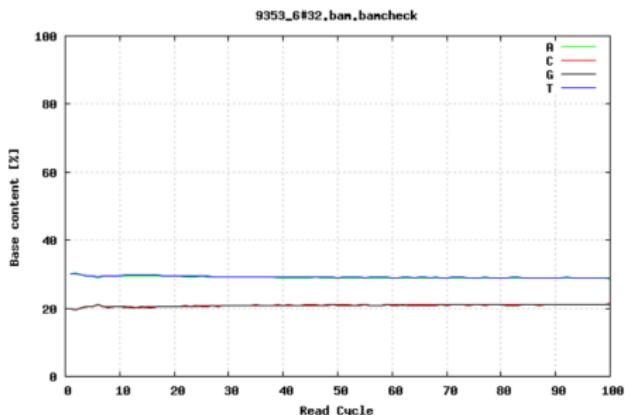


GC content vs depth



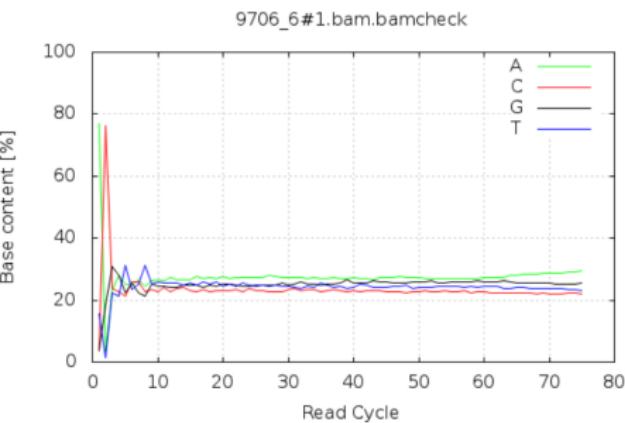
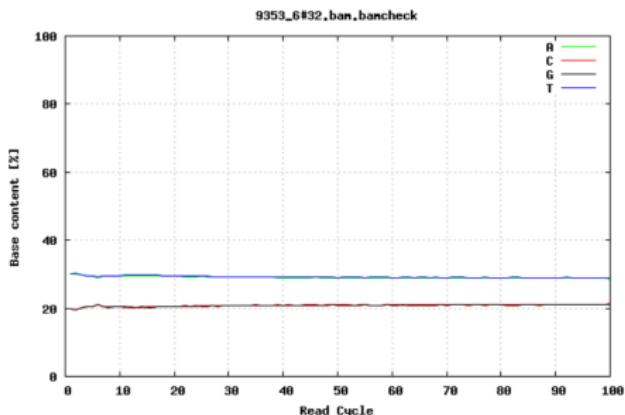
GC content by cycle

Was the adapter sequence trimmed?



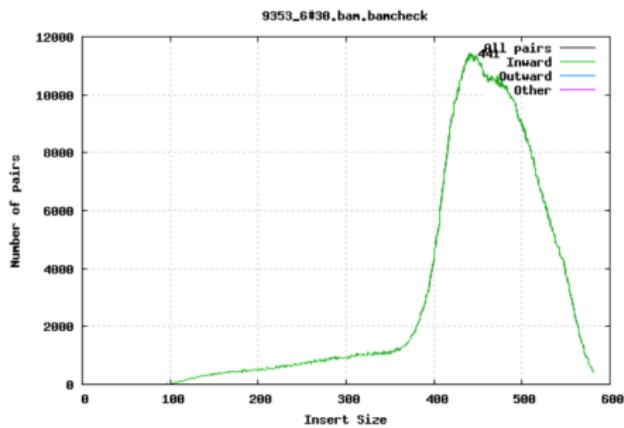
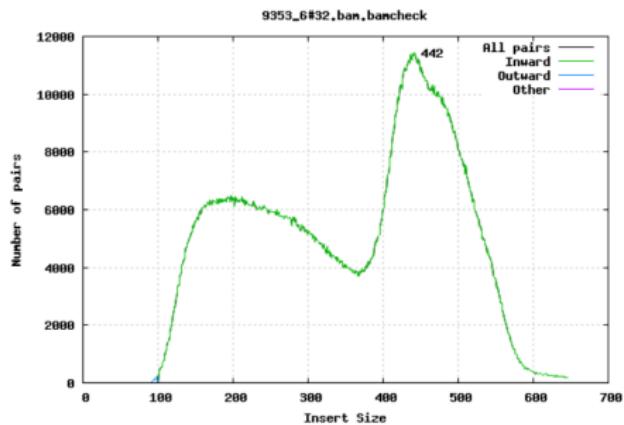
GC content by cycle

Was the adapter sequence trimmed?



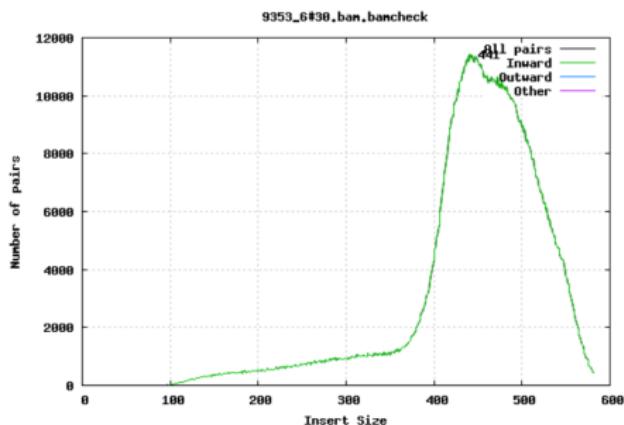
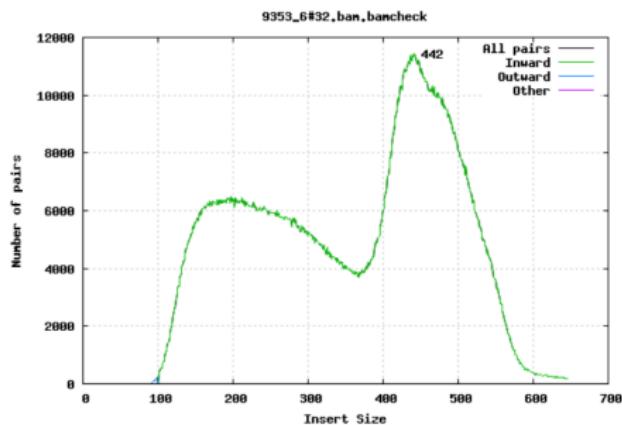
Fragment size

Paired-end sequencing: the size of DNA fragments matters

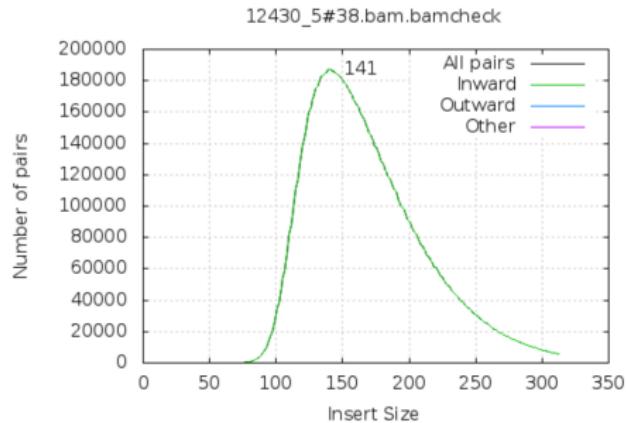


Fragment size

Paired-end sequencing: the size of DNA fragments matters

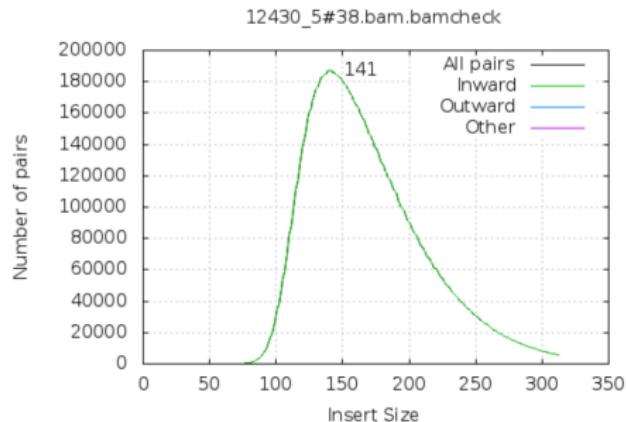


Quiz



This is 100bp paired-end sequencing. Can you spot any problems??

Quiz



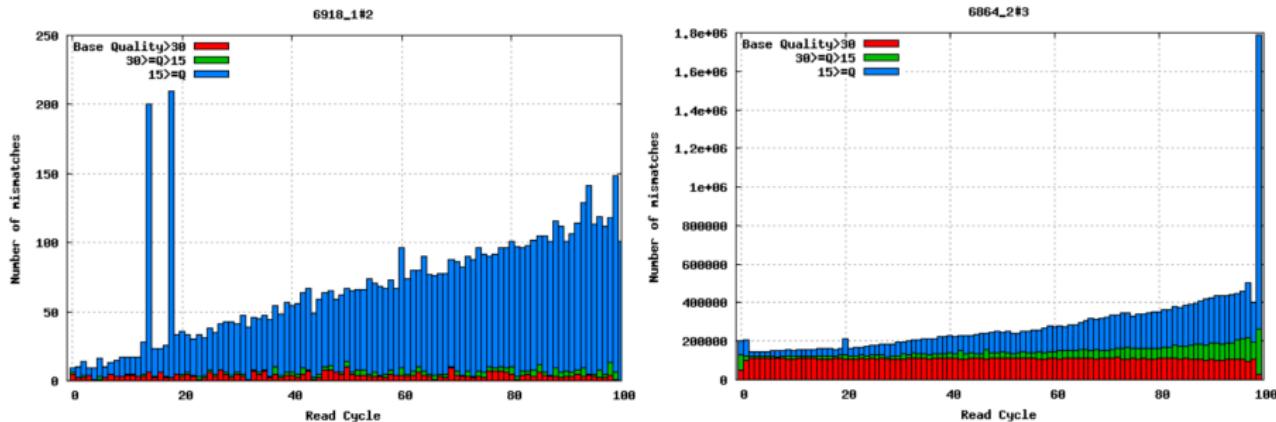
This is 100bp paired-end sequencing. Can you spot any problems??



Mismatches per cycle

Mismatches in aligned reads (requires reference sequence)

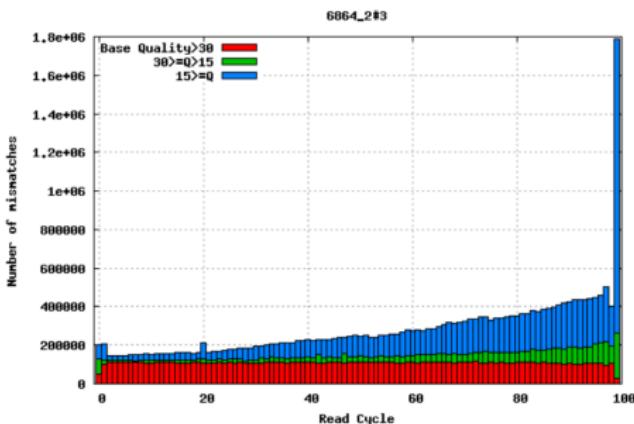
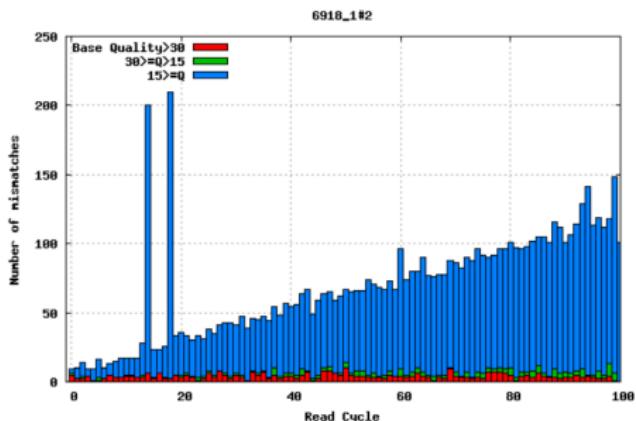
- ▶ detect cycle-specific errors
- ▶ base qualities are informative!



Mismatches per cycle

Mismatches in aligned reads (requires reference sequence)

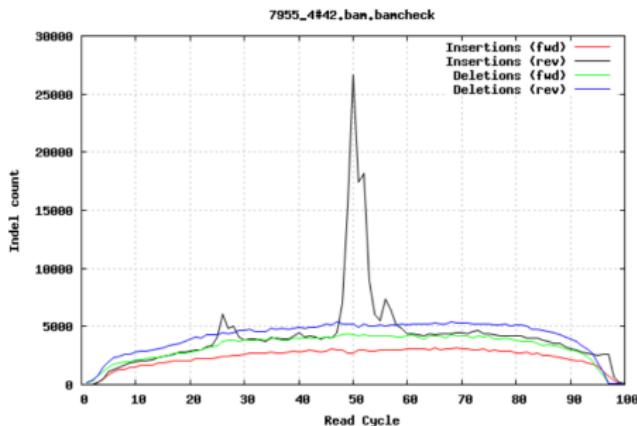
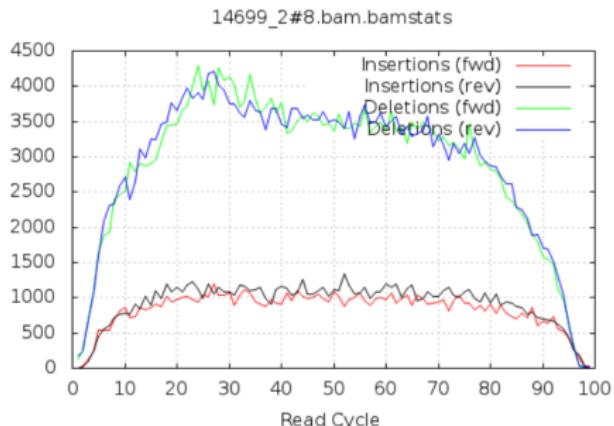
- ▶ detect cycle-specific errors
- ▶ base qualities are informative!



Insertions / Deletions per cycle

False indels

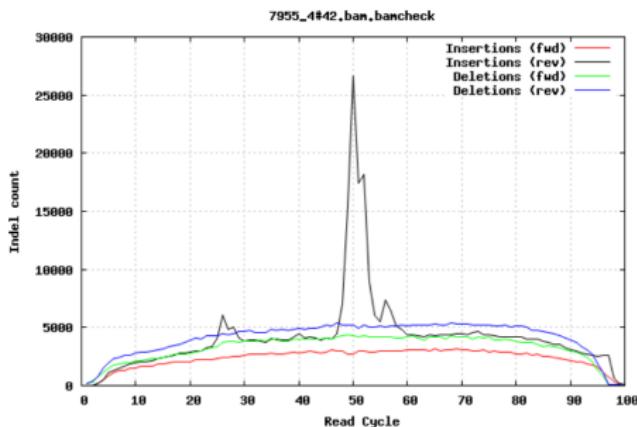
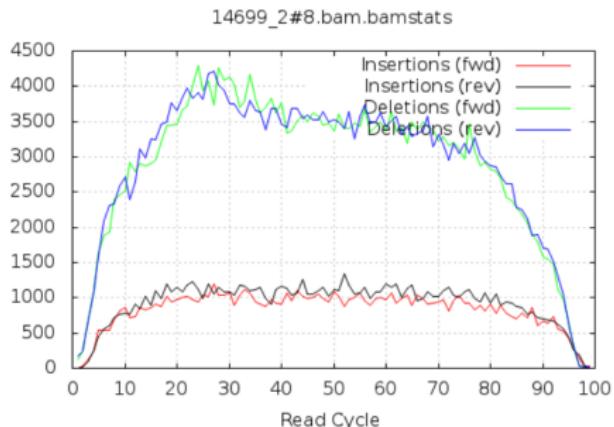
- air bubbles in the flow cell can manifest as false indels



Insertions / Deletions per cycle

False indels

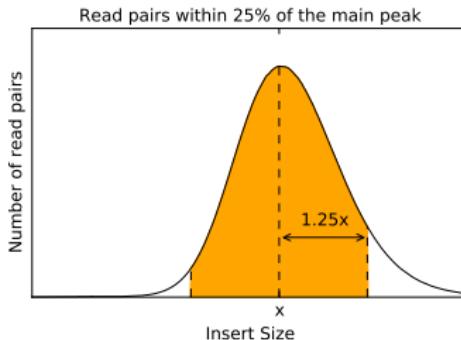
- air bubbles in the flow cell can manifest as false indels



Auto QC tests

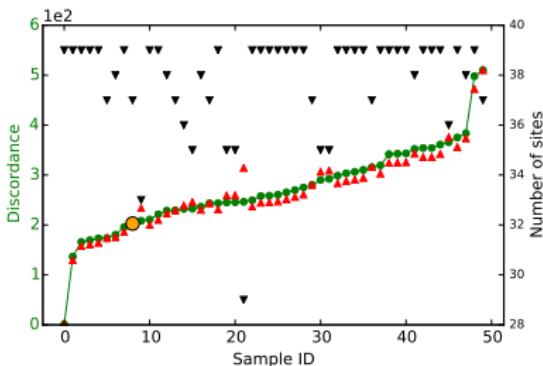
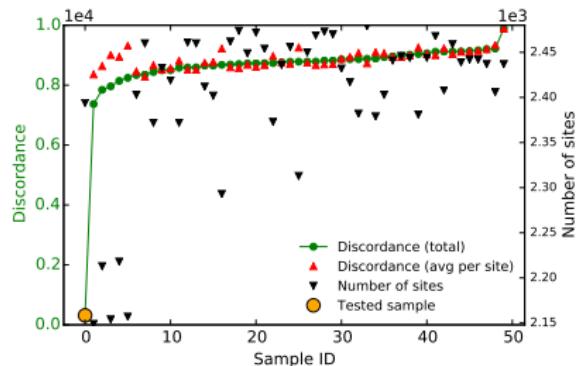
A suggestion for human data:

Minimum number of mapped bases	90%
Maximum error rate	0.02%
Maximum number of duplicate reads	5%
Minimum number of mapped reads which are properly paired	80%
Maximum number of duplicated bases due to overlapping read pairs	4%
Maximum in/del ratio	0.82
Minimum in/del ratio	0.68
Maximum indels per cycle, factor above median	8
Minimum number of reads within 25% of the main peak	80%



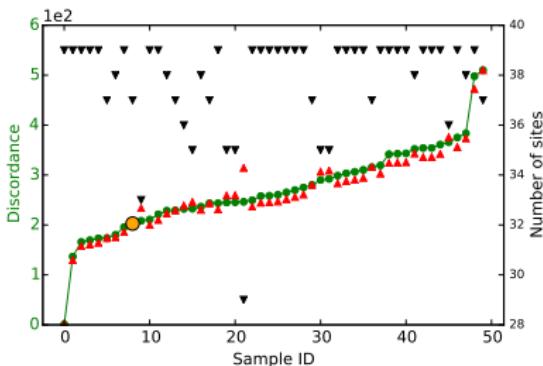
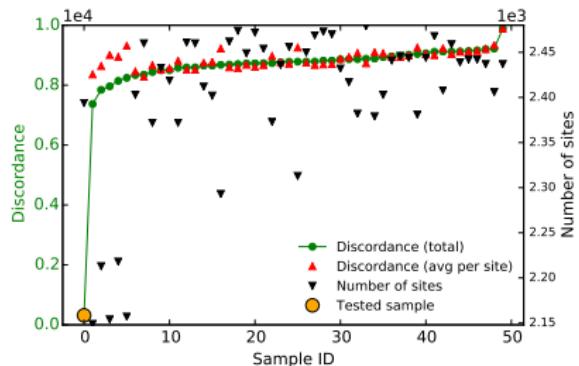
Detecting sample swaps

Check the identity against a known set of variants



Detecting sample swaps

Check the identity against a known set of variants



Software used to produce graphs in these slides

- ▶ samtools stats and plot-bamstats
- ▶ bcftools gtcheck
- ▶ matplotlib