Module 1: File formats, QC and Data Processing

Presented by:
Eric Dawson[1] and Carla Daniela Robles Espinoza[2]

[1]Bioinformatics Scientist (Genomics / AI) - Nvidia
[2]Laboratorio Internacional de Investigación sobre el Genoma Humano
Universidad Nacional Autónoma de México

Based on slides by:
Petr Danecek

@daniela_oaks
drobles@liigh.unam.mx

@erictdawson
eric.t.dawson@gmail.com

Cancer Genome Analysis

26 November – 1 December 2023 – Montevideo, Uruguay

LIIGH-UNAM

WELLCOME GENOME CAMPUS
CONNECTING SCIENCE
ADVANCED COURSES+ SCIENTIFIC CONFERENCES

Data Formats

FASTQ

- Unaligned read sequences with base qualities

SAM/BAM

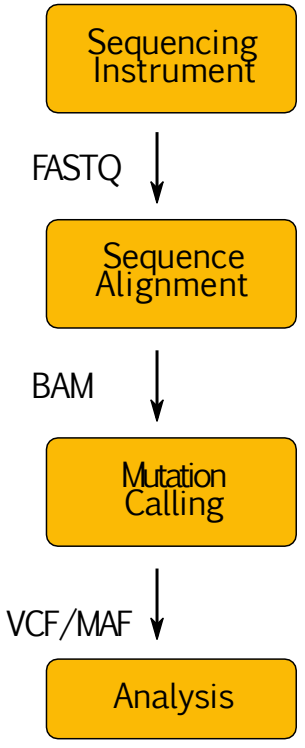
- Unaligned or aligned reads
- Text and binary formats

CRAM

- Better compression than BAM

VCF/BCF or MAF

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



BAM/VCF specifications maintained by the Global Alliance for Genomics and Health
MAF documentation can be found in NCI's Genomic Data Commons

FASTA – reference genome

```
>1 dna:chromosome chromosome:GRCh37:1:1:249250621:1
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
TATTCAAAAAATTGAGAATTTCTGACCACTTAACAAACCCACAGAAAATCCACCCGAGTG
CACTGAGCACGCCAGAAATCAGGTGGCCTCAAAGAGCTGCTCCCACCTGAAGGAGACGCG
CTGCTGCTGCTGTCGTCCTGCCTGGCGCCTTGGCCTACAGGGGCCGCGGTTGAGGGTGGG
AGTGGGGGTGCACTGGCCAGCACCTCAGGAGCTGGGGGTGGTGGTGGGGGCGGTGGGGGT
GGTGTTAGTACCCCATCTTGTAGGTCTGAAACACAAAGTGTGGGGTGTCTAGGGAAGAAG
>2
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
AAAAGCATTTATGCTACAAATTACTATGGTAATTATGCTACAAATTTATGGTACCATAAA
TTACCATAGTAATTTGTAGCATAAATTTGTACTATGGTACAAATTACATGGGAGAGTGAA
GGTGGGTTAAAACATTCATATTAAAGAACTTCCACTCAGATTGCAAGAAAAGAGAGAGGA
ATGGAGATGGTAGCACAAGTCCCTACAATAAAAGTAGATGTTTTGAGATCAGTTCTATTT
```

FASTA – reference genome

```
>1 dna:chromosome chromosome:GRCh37:1:1:249250621:1
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
TATTCAAAAAATTGAGAATTTCTGACCACTTAACAAACCCACAGAAAATCCACCCGAGTG
CACTGAGCACGCCAGAAATCAGGTGGCCTCAAAGAGCTGCTCCCACCTGAAGGAGACGCG
CTGCTGCTGCTGTCGTCCTGCCTGGCGCCTTGGCCTACAGGGGCCGCGGTTGAGGGTGGG
AGTGGGGGTGCACTGGCCAGCACCTCAGGAGCTGGGGGTGGTGGTGGGGGCGGTGGGGGT
GGTGTTAGTACCCCATCTTGTAGGTCTGAAACACAAAGTGTGGGGTGTCTAGGGAAGAAG
>2
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
AAAAGCATTTATGCTACAAATTACTATGGTAATTATGCTACAAATTTATGGTACCATAAA
TTACCATAGTAATTTGTAGCATAAATTTGTACTATGGTACAAATTACATGGGAGAGTGAA
GGTGGGTTAAAACATTCATATTAAAGAACTTCCACTCAGATTGCAAGAAAAGAGAGAGGA
ATGGAGATGGTAGCACAAGTCCCTACAATAAAAGTAGATGTTTTGAGATCAGTTCTATTT
```

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

```
@ERR007731.739 IL16_2979:6:1:9:1684/1
CTTGACGACTTGAAAAATGACGAAATCACTAAAAAACGTGAAAAATGAGAAATG
+
BBCBCBBBBBBBABBABBBBBBBABBBBBBBBBBBBBBABAAAABBBBB=@>B
@ERR007731.740 IL16_2979:6:1:9:1419/1
AAAAAAAAAGATGTCATCAGCACATCAGAAAAGAAGGCAACTTTAAAACTTTTC
+
BBABBBABABAABABABBABBBAAA>@B@BBAA@4AAA>.>BAA@779:AAA@A
```

- 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;'
```

FASTQ

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

```
@ERR007731.739 IL16_2979:6:1:9:1684/1
CTTGACGACTTGAAAAATGACGAAATCACTAAAAAACGTGAAAAATGAGAAATG
+
BBCBCBBBBBBBABBABBBBBBBABBBBBBBBBBBBBBABAAAABBBBB=@>B
@ERR007731.740 IL16_2979:6:1:9:1419/1
AAAAAAAAAGATGTCATCAGCACATCAGAAAAGAAGGCAACTTTAAAACTTTTC
+
BBABBBABABAABABABBABBBAAA>@B@BBAA@4AAA>.>BAA@779:AAA@A
```

- 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;'
```

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%

FASTQ

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	~
28	FS (file separator)	61	=	94	^	127	DEL
29	GS (group separator)	62	>	95	_		
30	RS (record separator)	63	?	96	`		
31	US (unit separator)	64	@	97	a		
32	SPACE	65	A	98	b		

FASTQ

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

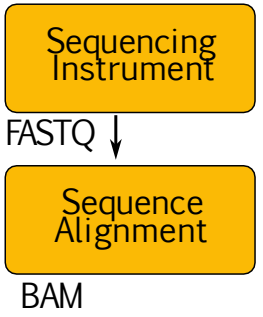
```
@ERR007731.739 IL16_2979:6:1:9:1684/1
CTTGACGACTTGAAAAATGACGAAATCACTAAAAAACGTGAAAAATGAGAAATG
+
BBCBCBBBBBBBABBABBBBBBBABBBBBBBBBBBBBBABAAAABBBBB=@>B
@ERR007731.740 IL16_2979:6:1:9:1419/1
AAAAAAAAAGATGTCATCAGCACATCAGAAAAGAAGGCAACTTTAAAACTTTTC
+
BBABBBABABAABABABBABBBAAA>@B@BBAA@4AAA>.>BAA@779:AAA@A
```

- 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;'
```

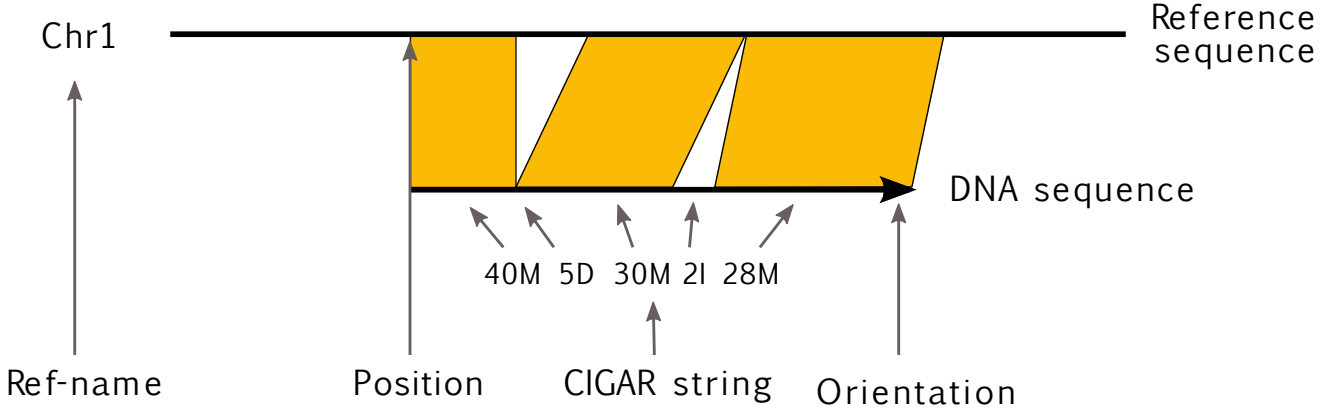
- Beware: multiple quality scores were in use!
  - Sanger, Solexa, Illumina 1.3+
- Paired-end sequencing produces two FASTQ files

SAM / BAM



SAM (Sequence Alignment/Map) format

- Unified format for storing read alignments to a reference genome
- Developed by the 1000 Genomes Project group (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           0   106M45S =  10335       0 \
     CCATAACCCTAACCCTAACCCTAACCATAGCCCTAACCCTAACCCTAACCCTAACCCT[...]CAAACCCACCCCCAAACCCAAAACCTCACCAC\
     FFFFFJJJJJJJJFJJJJFJAJJJJJ-JJAAAJFJJFFJJF<FJJFFJJJJFJJJJFF[...]<---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--GT
Read:           ACGCA-TGCAGTtagacgt
Cigar:          5M 1D 2M 2I 2M 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

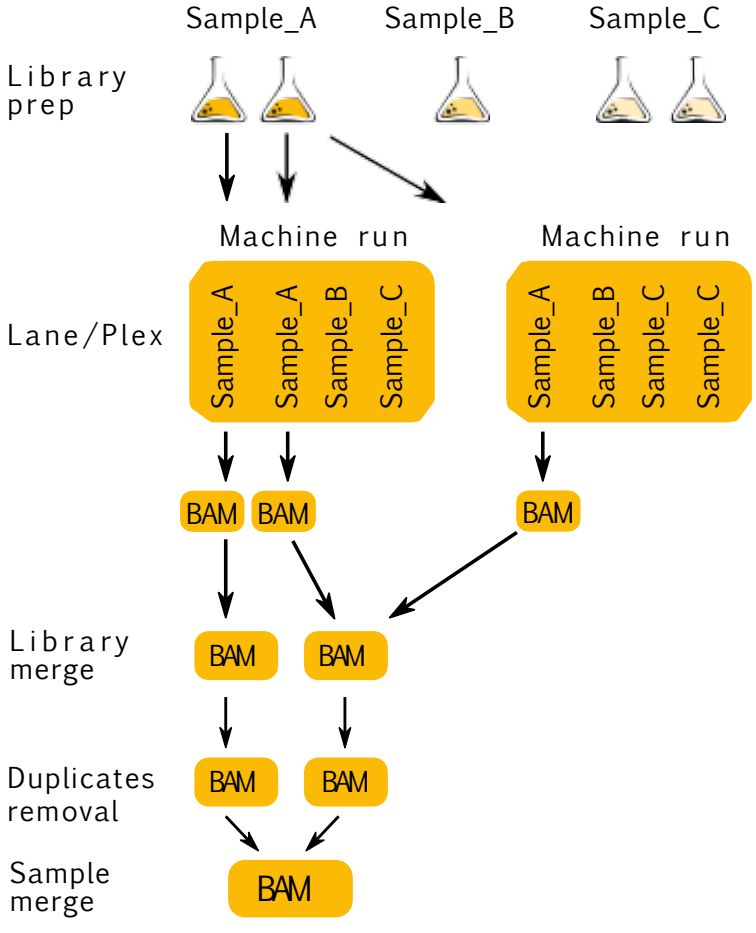
- samtools flags

  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



BAM

BAM (Binary Alignment/Map) format

- Binary version of SAM
- Developed for fast processing and random access
    - BGZF (Block GZIP) compression for indexing

Key features

- Can store alignments from most mappers
- Supports multiple sequencing technologies
- Supports indexing for quick retrieval/viewing
- Compact size (e.g. 112Gbp Illumina = 116GB disk space)
- Reads can be grouped into logical groups e.g. lanes, libraries, samples
- Widely supported by variant calling packages and viewers`

SAM/BAM tools

Several tools and programs for interacting with SAM/BAM files:

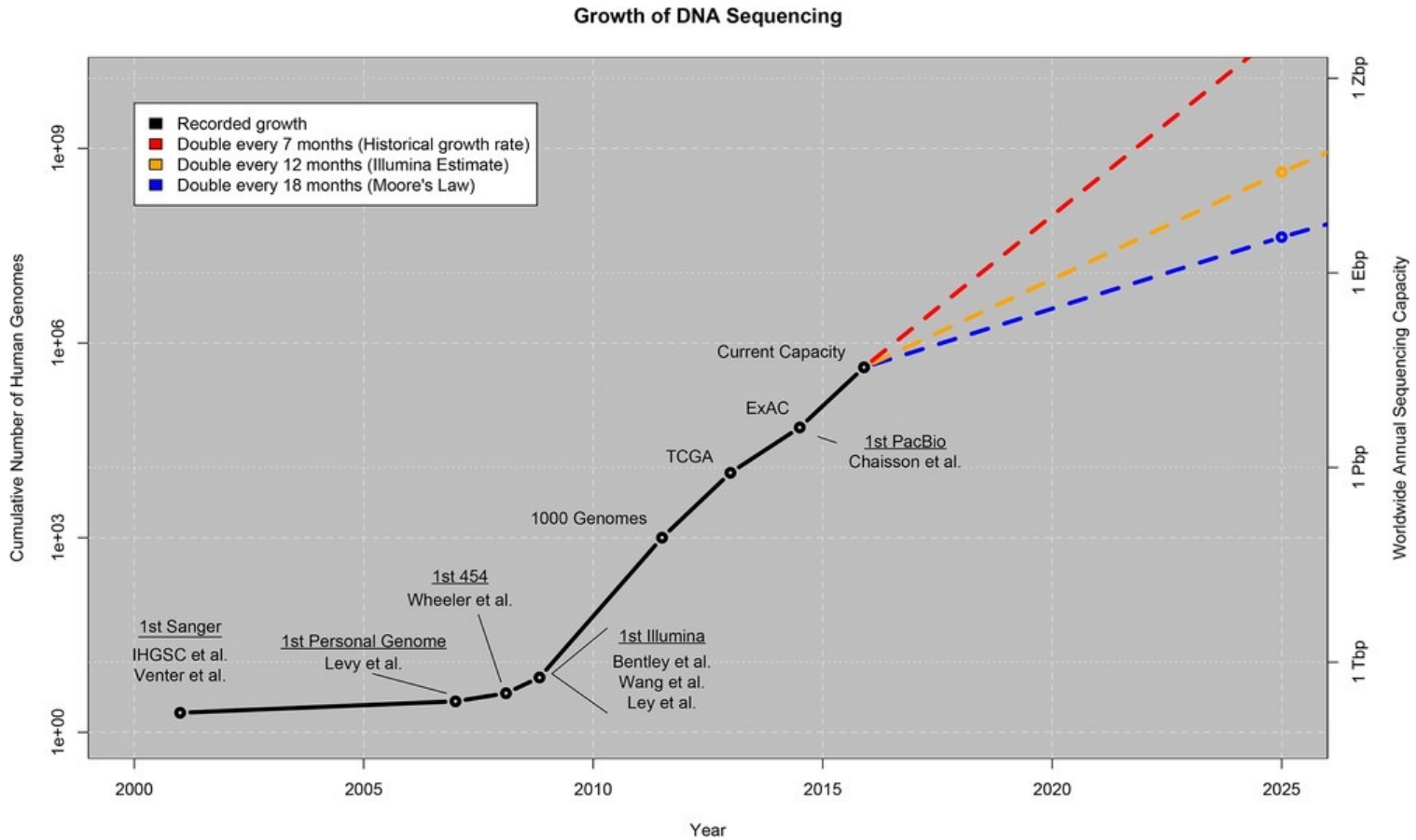
- Samtools (Wellcome Sanger Institute)
- Picard tools (Broad Institute)
- Visualisation: IGV, Ensembl, UCSC

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-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

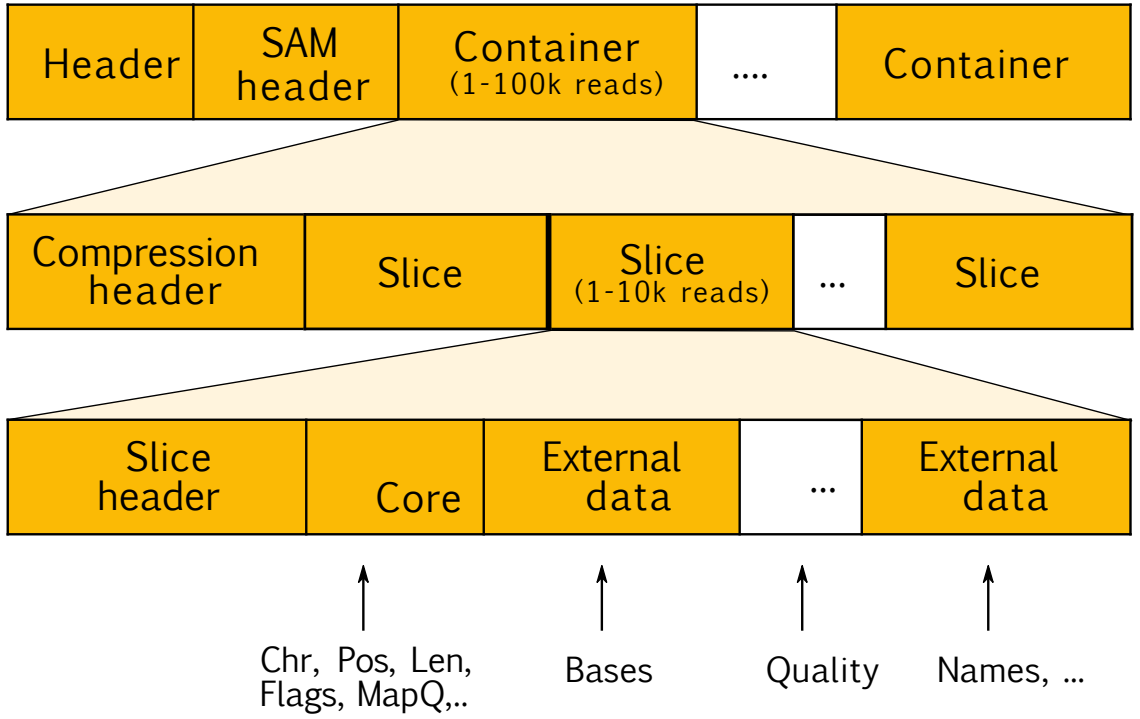
CRAM

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 is now mature and used in production pipelines

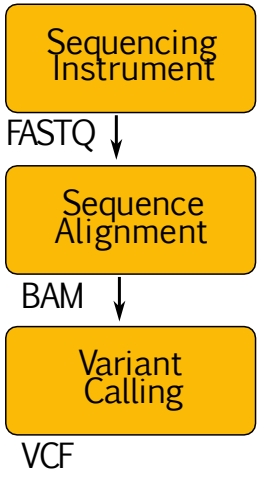
- Support for CRAM added to Samtools/HTSlib in 2014
- Added in Picard/GATK in 2015



VCF: Variant Call Format

File format for storing generic variation data

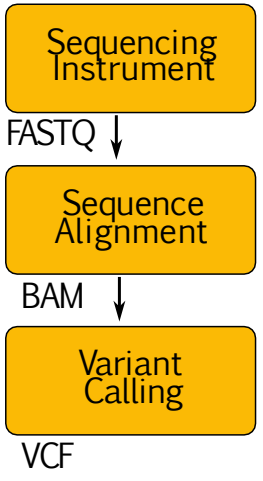
- Accommodate all types of variation: SNPs, short indels, large events
- Multiple samples



VCF: Variant Call Format

File format for storing generic variation data

- Accommodate all types of variation: SNPs, short indels, large events
- Multiple samples



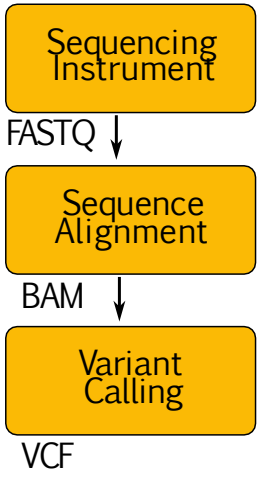
Columns

- Chromosome and position
- Variant ID
- Quality of the call
- Soft filtering (*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), GQ (genotype quality)

VCF: Variant Call Format

File format for storing generic variation data

- Accommodate all types of variation: SNPs, short indels, large events
- Multiple samples



Genotypes (for diploid individuals)

- Homozygous reference (*e.g.*, A/A if the reference allele is A)
- Homozygous alternative (*e.g.*, G/G if the reference allele is A)
- Heterozygous (*e.g.*, C/T)

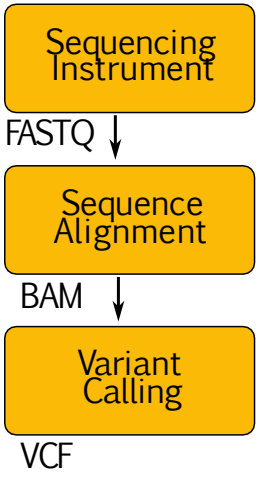
Allele numbering (for VCF notation):

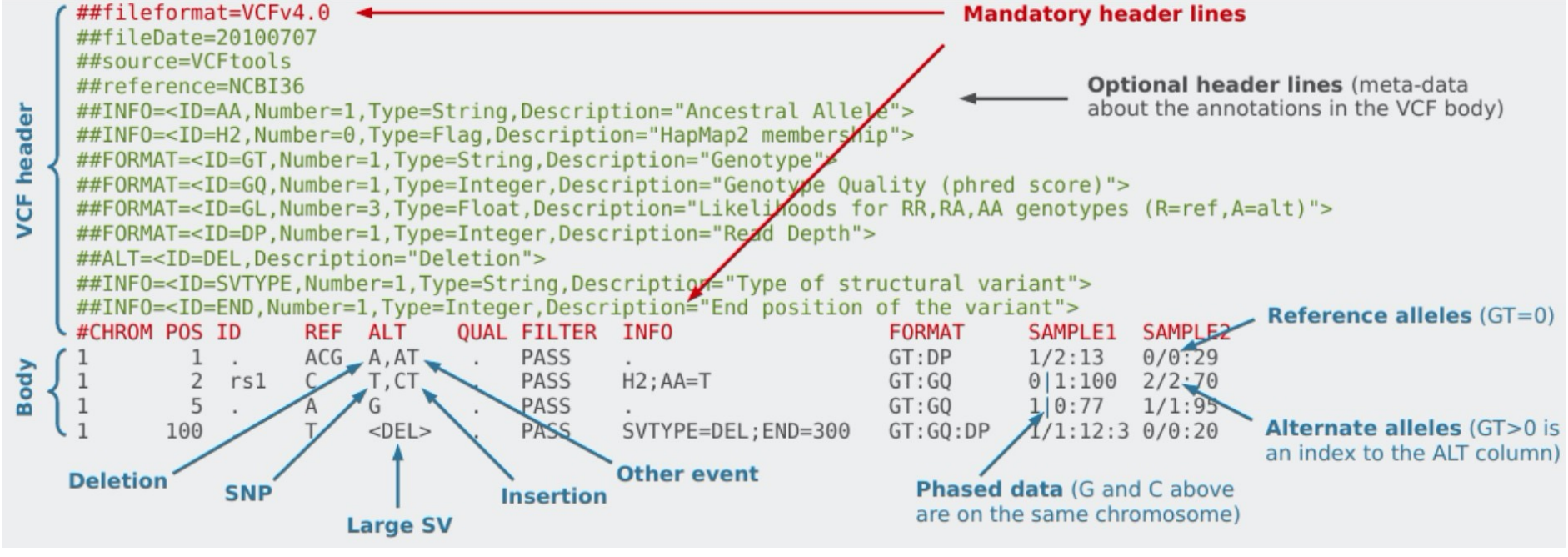
- Reference allele is 0, first alternative allele is 1, second is 2, etc
- Homozygous reference (0/0)
- Homozygous alternative (1/1, 2/2, etc.)
- Heterozygous (0/1, 1/2, etc)

VCF: Variant Call Format

File format for storing generic variation data

- Accommodate all types of variation: SNPs, short indels, large events
- Multiple samples
- Tab-delimited text, parsable by standard UNIX commands
- Flexible and user-extensible
- Compressed with BGZF (bgzip), indexed with TBI or CSI (tabix)

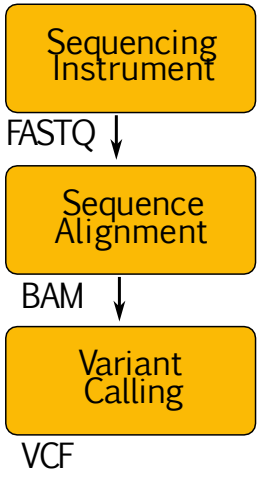




VCF: Variant Call Format

File format for storing generic variation data

- Accommodate all types of variation: SNPs, short indels, large events
- Multiple samples
- Tab-delimited text, parsable by standard UNIX commands
- Flexible and user-extensible
- Compressed with BGZF (bgzip), indexed with TBI or CSI (tabix)



Some annotations important in tumour VCFs:

- AF: Allelic fraction of alternate alleles in tumour sample
- GERMQ: Phred-scaled quality that ALT alleles are not germline
- PON: Site found in panel of normals

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
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

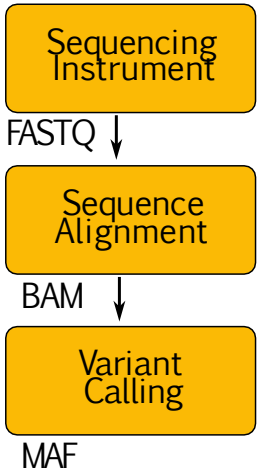
```
#CHROM POS ID REF ALT QUAL FILTER INFO          FORMAT      SAMPLE1          SAMPLE2          SAMPLE3          SAMPLE4             SAMPLE5
1      6   .  A   G   .    PASS   AC=67;AN=540  GT:PL:DP:GQ  1/1:0,9,73:26:22  0/0:0,9,73:13:31  0/0:0,9,73:48:99  1/0:255,0,75:32:15  1/0:255,0,75:32:15

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
```

MAF

File format for storing aggregated variation data

- Stands for "Mutation-Annotation Format" and is a tab-delimited text file containing aggregated information from VCF files (NCI-GDC)
- It aggregates lots of information – 120+ fields per mutation! You can review all these at:



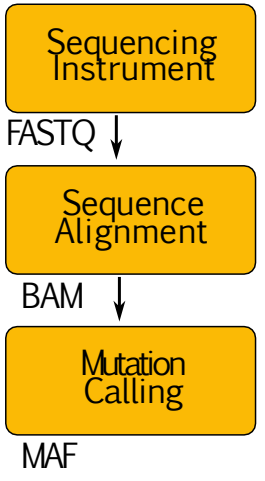
https://docs.gdc.cancer.gov/Data/File_Formats/MAF_Format/

Column
1 - Hugo_Symbol
2 - Entrez_Gene_Id
3 - Center
4 - NCBI_Build
5 - Chromosome
6 - Start_Position
7 - End_Position
8 - Strand
9 - Variant_Classification
10 - Variant_Type
11 - Reference_Allele
12 - Tumor_Seq_Allele1
13 - Tumor_Seq_Allele2
14 - dbSNP_RS
15 - dbSNP_Val_Status
16 - Tumor_Sample_Barcode
17 - Matched_Norm_Sample_Barcode
18 - Match_Norm_Seq_Allele1
19 - Match_Norm_Seq_Allele2
20 - Tumor_Validation_Allele1
21 - Tumor_Validation_Allele2
22 - Match_Norm_Validation_Allele1
23 - Match_Norm_Validation_Allele2
24 - Verification_Status
25 - Validation_Status
26 - Mutation_Status
27 - Sequencing_Phase
28 - Sequence_Source
29 - Validation_Method
30 - Score
31 - BAM_File
32 - Sequencer
33 - Tumor_Sample_UUID
34 - Matched_Norm_Sample_UUID
35 - HGVSc
36 - HGVSp
37 - HGVSp_Short
38 - Transcript_ID
39 - Exon_Number
40 - t_depth
41 - t_ref_count
42 - t_alt_count
43 - n_depth
44 - n_ref_count
45 - n_alt_count
46 - all_effects
47 - Allele
48 - Gene
49 - Feature
50 - Feature_type
51 - One_Consequence
52 - Consequence
53 - cDNA_position
54 - CDS_position
55 - Protein_position
56 - Amino_acids
57 - Codons
58 - Existing_variation
59 - ALLELE_NUM
60 - DISTANCE
61 - TRANSCRIPT_STRAND
62 - SYMBOL
63 - SYMBOL_SOURCE
64 - HGNC_ID
65 - BIOTYPE
66 - CANONICAL
67 - CCDS
68 - ENSP
69 - SWISSPROT
70 - TREMBL
71 - UNIPARC
72 - RefSeq
73 - SIFT
74 - PolyPhen
75 - EXON
76 - INTRON
77 - DOMAINS
78 - GMAF

MAF

You can convert between VCF and MAF via vcf2maf tools!

```
perl vcf2maf.pl --input-vcf [vcf_file] --output-maf
[maf_file] --vep-path /cm/shared/apps/vep/ensembl-vep-
release-106.1/ --vep-data /mnt/Archives/vep/106/38/ --ref-
fasta [fasta_file] --tumor-id [tumor] --normal-id [normal]
```



Global Alliance for Genomics and Health

International coalition dedicated to improving human health
Mission

- establish a common framework to enable sharing of genomic and clinical data

Working groups

- clinical
- regulatory and ethics
- security
- data

Data working group

- beacon project .. test the willingness of international sites to share genetic data
- BRCA challenge .. advance understanding of the genetic basis of breast and other cancers
- matchmaker exchange .. locate data on rare phenotypes or genotypes
- reference variation .. describe how genomes differ so researchers can assemble and interpret them
- benchmarking .. develop variant calling benchmark toolkits for germline, cancer, and transcripts
- file formats .. CRAM, SAM/BAM, VCF/BCF

File formats

- http://samtools.github.io/hts-specs/

Quality Control

Biases in sequencing

- Base calling accuracy
- Read cycle vs. base content
- GC vs. depth
- Indel ratio

Biases in mapping

Genotype checking

- Sample swaps
- Contaminations

Read coverage

Read coverage / depth

- Is every genomic position covered to a sufficient depth?
- Average depth: number of reads / target size
    - Whole human genome: 3Gb
    - Human exome: 50Mb

Exomes

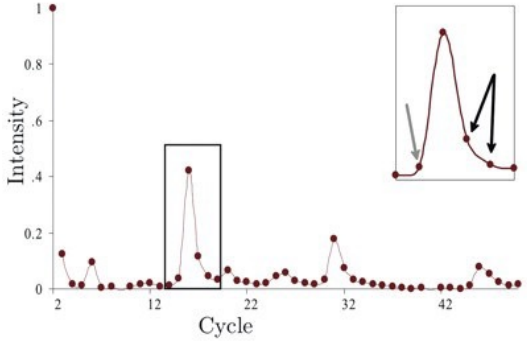
- Be careful to distinguish between the total sequencing yield and on-target bases

Useful coverage:

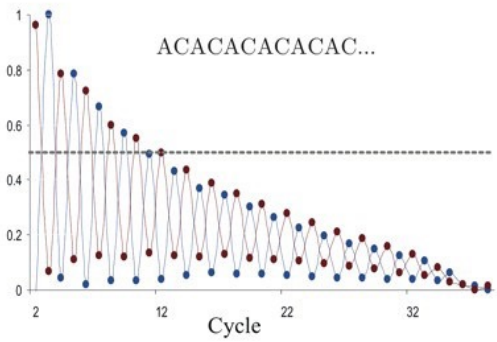
- 15x OK for common germline variants
- 30x OK for most things
- 100-200x for low VAF variants in tumours

Base calling errors

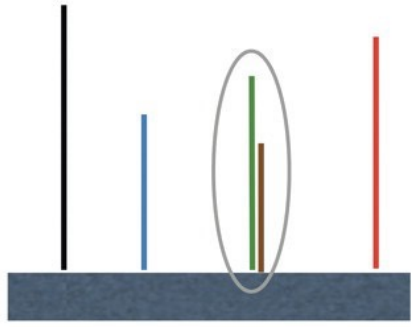
Phasing noise ϕ



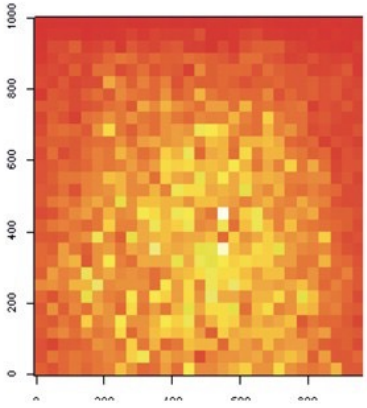
Signal Decay δ



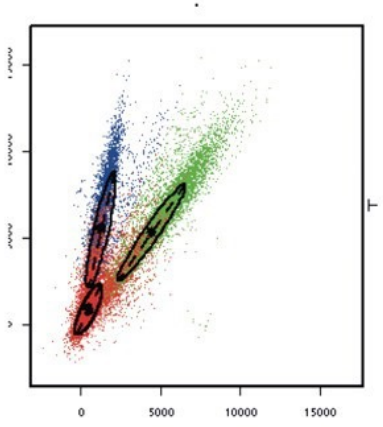
Mixed Cluster μ



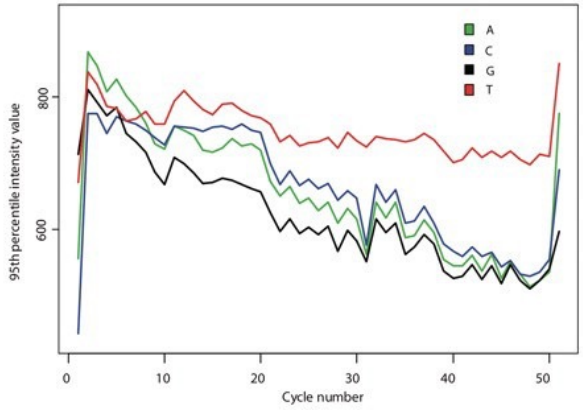
Boundary effects ω



Cross-talk Σ



T fluophore accumulation τ



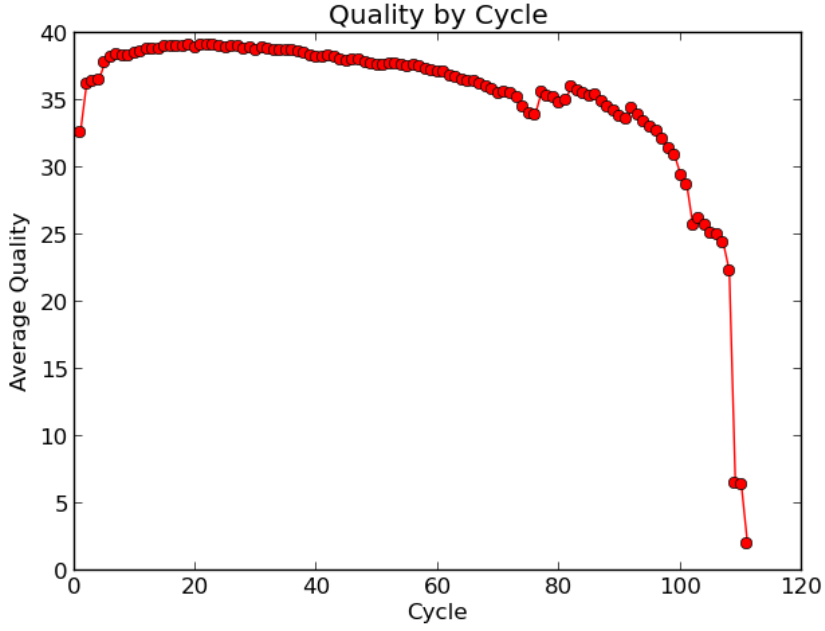
Base-calling for next-generation sequencing platforms, doi: 10.1093/bib/bbq077

Base quality

Sequencing by synthesis: dephasing

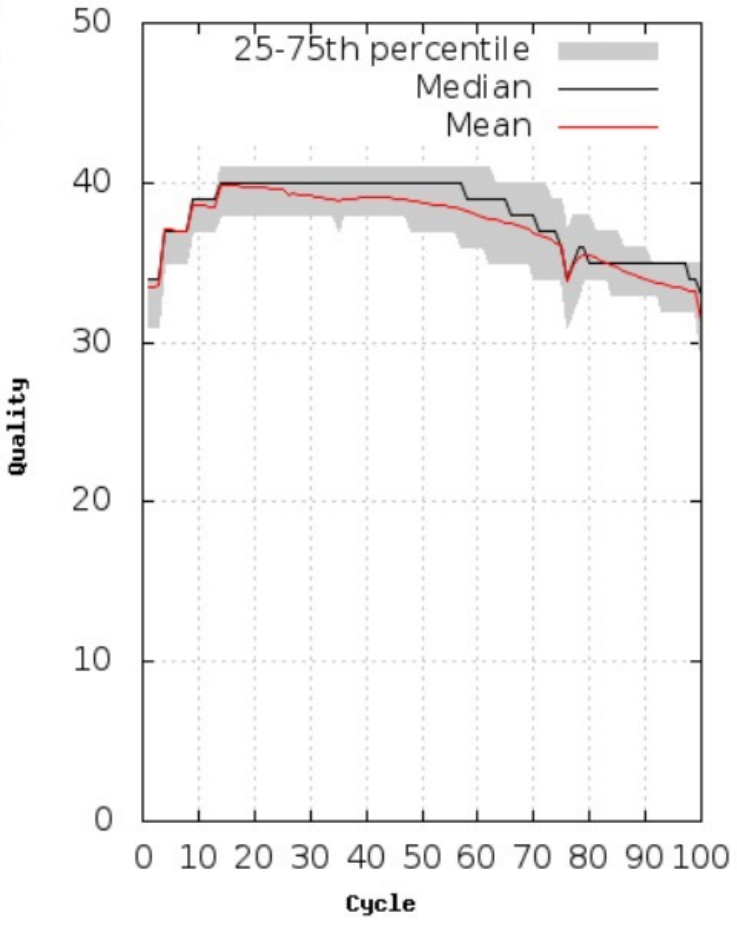
- growing sequences in a cluster gradually desynchronize
- error rate increases with read length

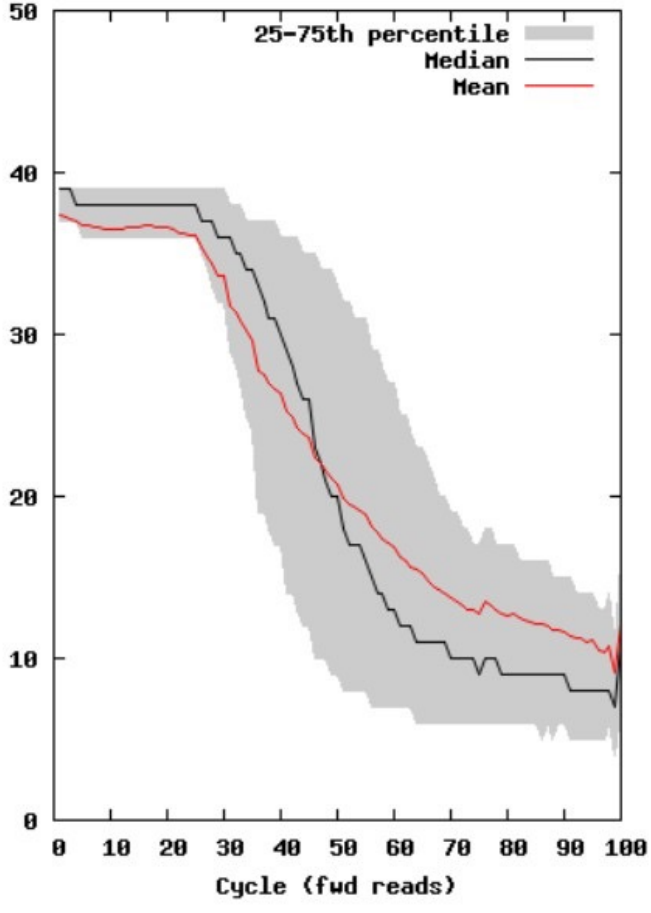
Calculate the average quality at each position across all reads



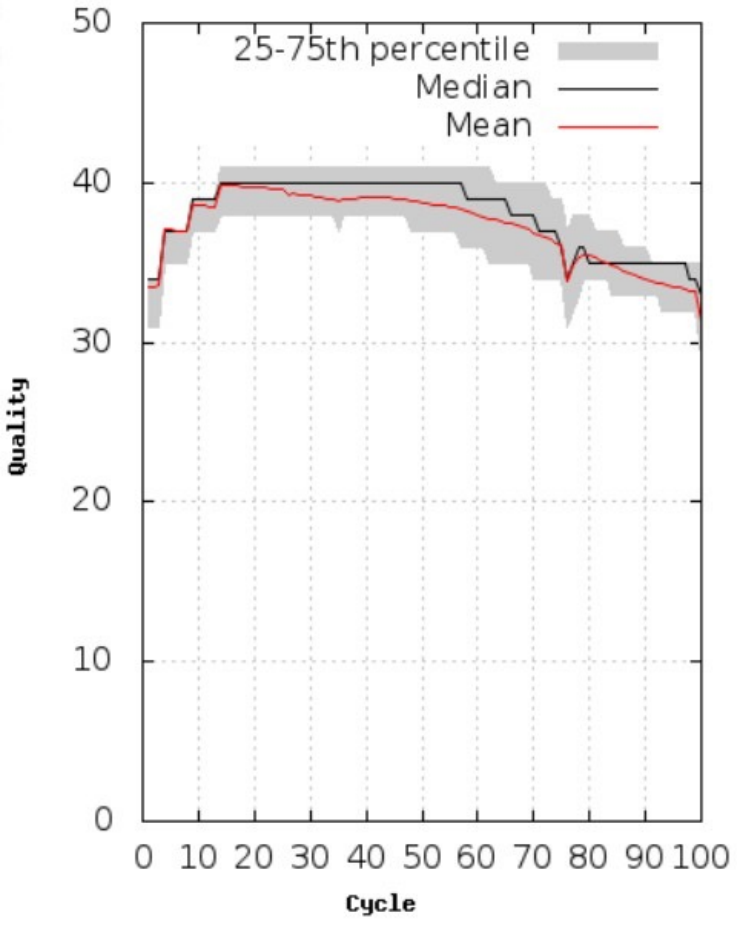
Quality	Probability of error	Call Accuracy
10 (Q10)	1 in 10	90%
20 (Q20)	1 in 100	99%
30 (Q30)	1 in 1000	99.9%
40 (Q40)	1 in 10000	99.99%

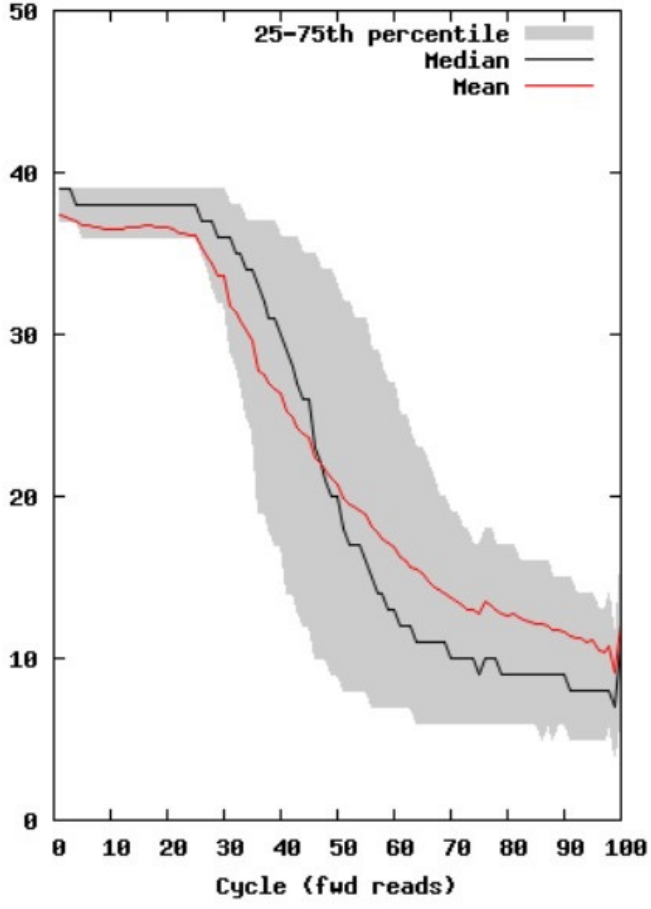
Base quality

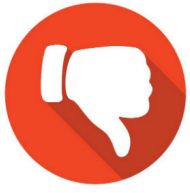




Base quality







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

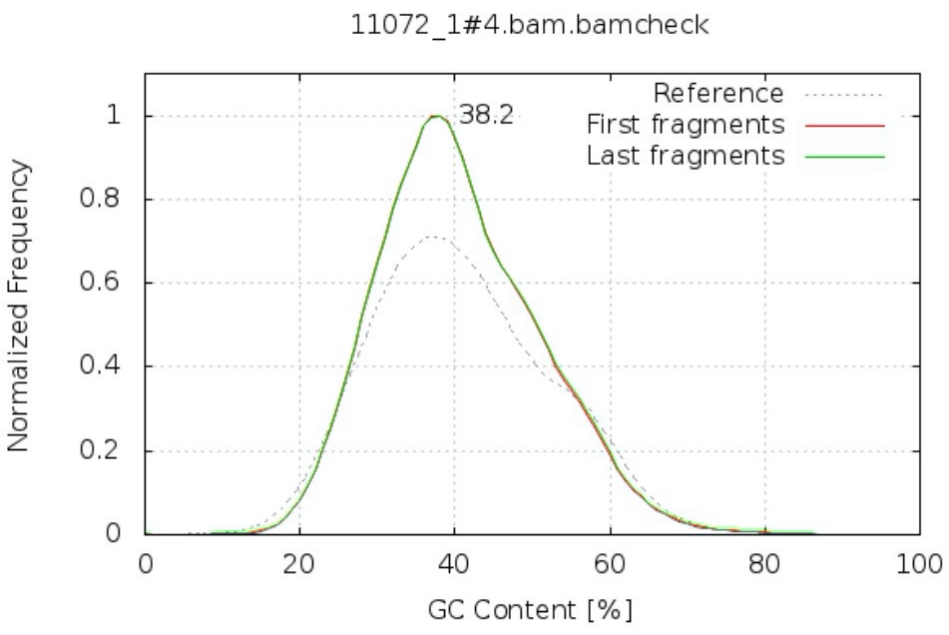
Solution:

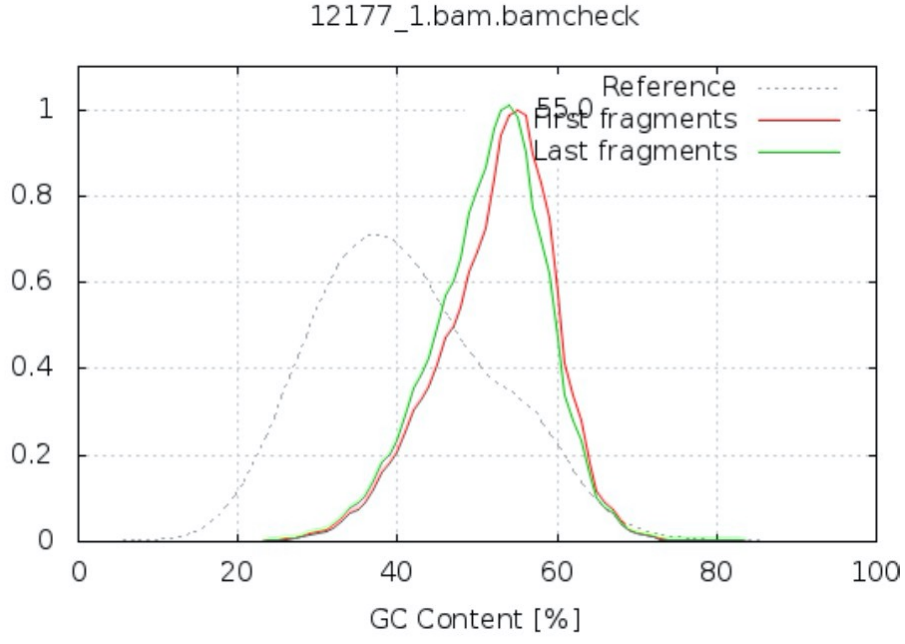
- Infer and mark PCR duplicates, discount in later analysis
    - Mark if reads and their mates start at the same position
- Use Picard MarkDuplicates or samtools markdup
- Typical duplication rates: Exomes 15-20%, Genomes < 5%

GC bias

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

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

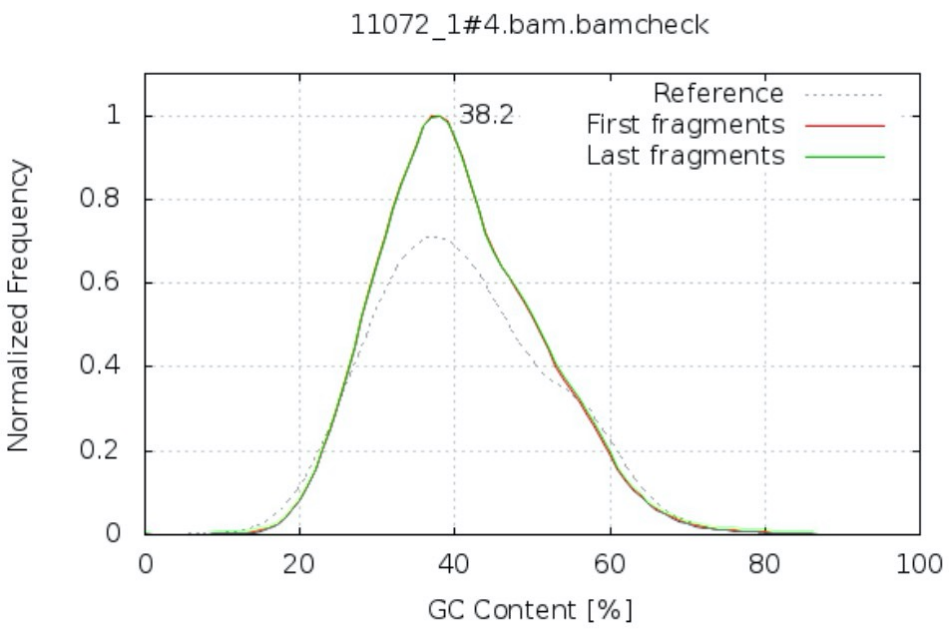


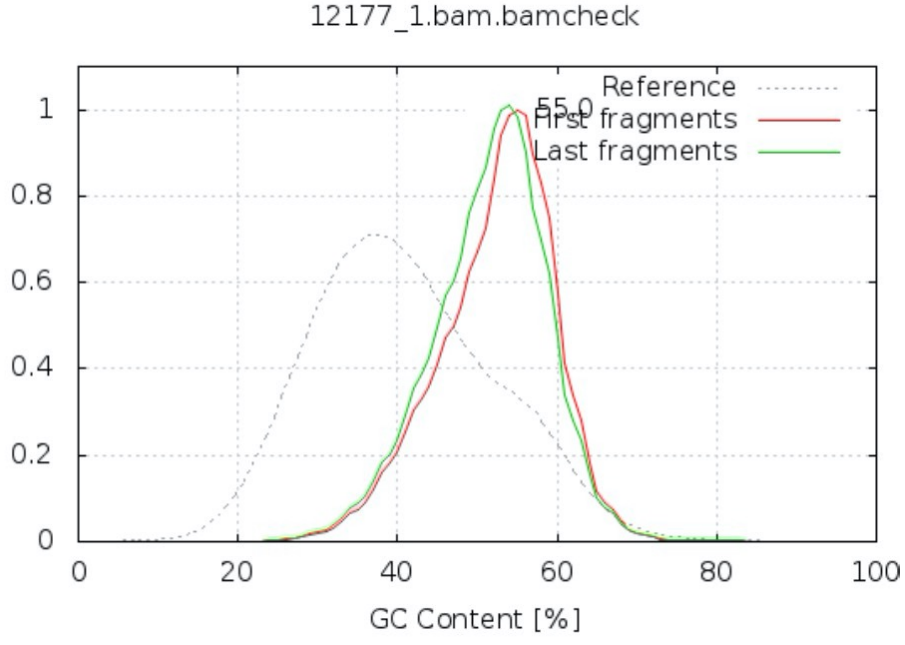


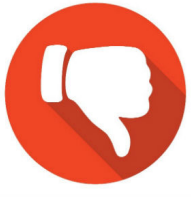
GC bias

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

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

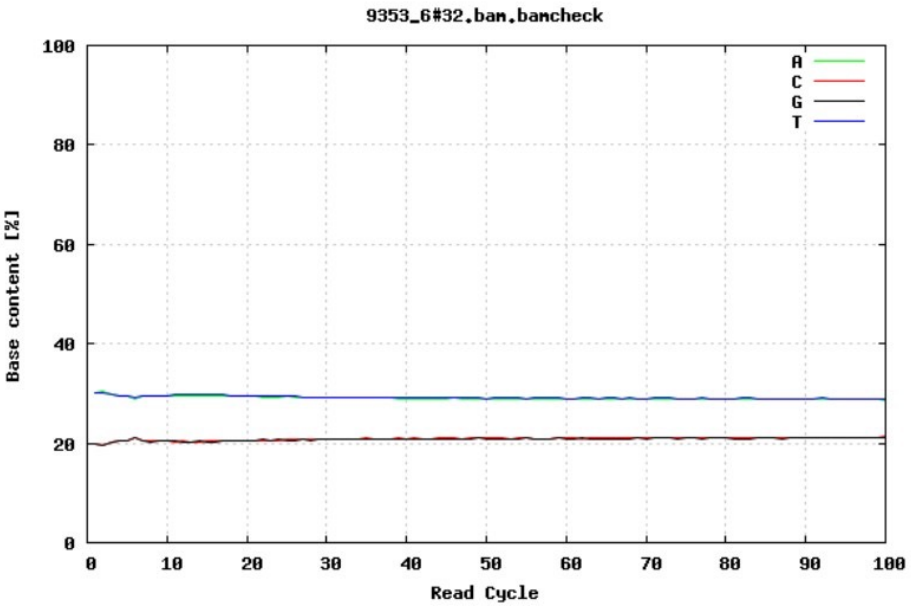


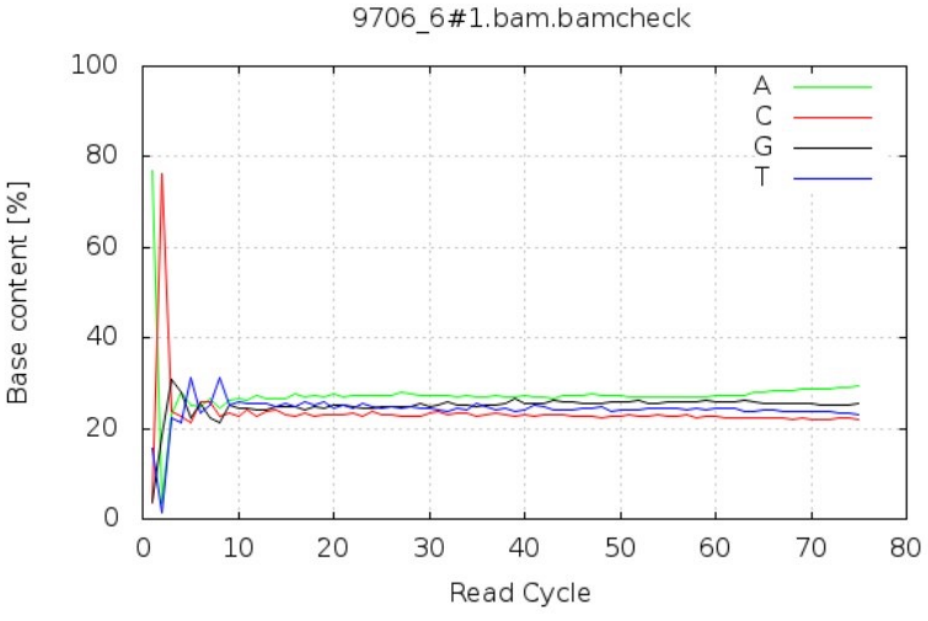




GC content by cycle

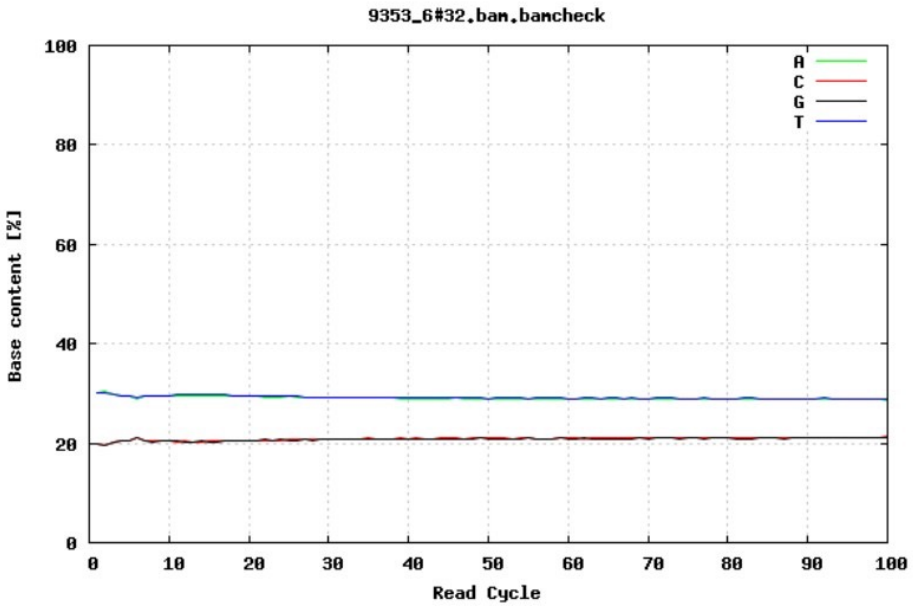
Was the adapter sequence trimmed?

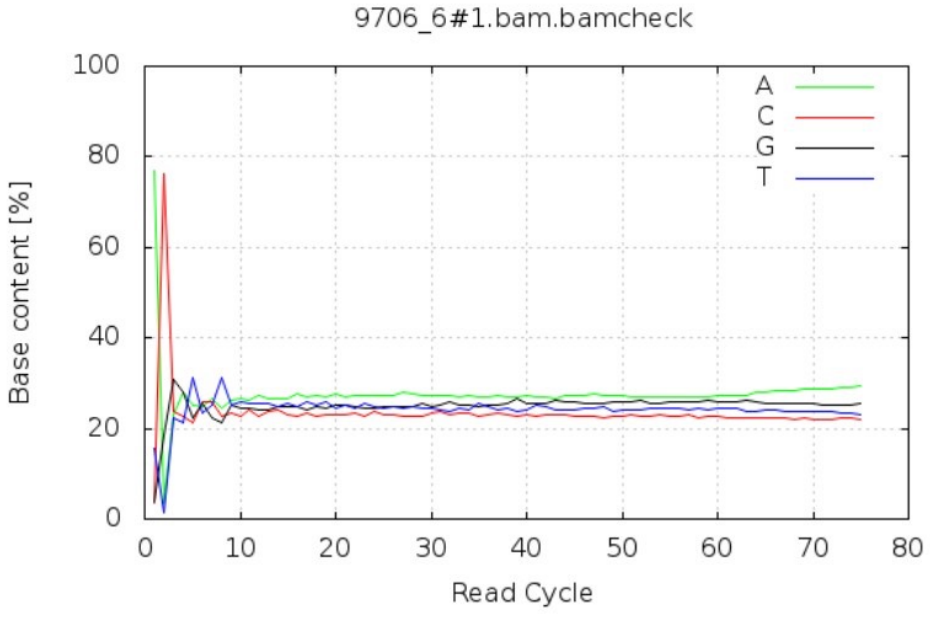


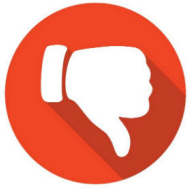


GC content by cycle

Was the adapter sequence trimmed?

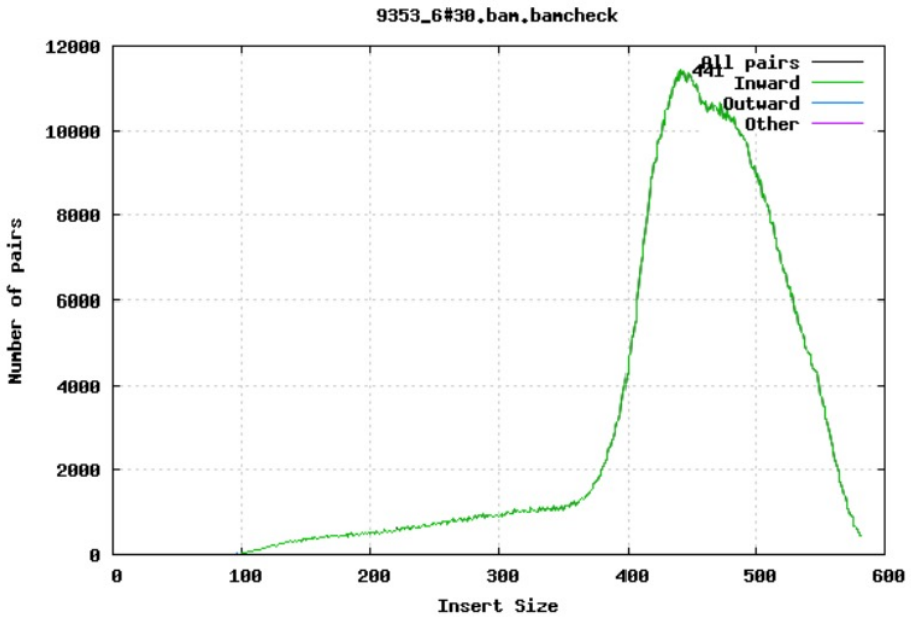


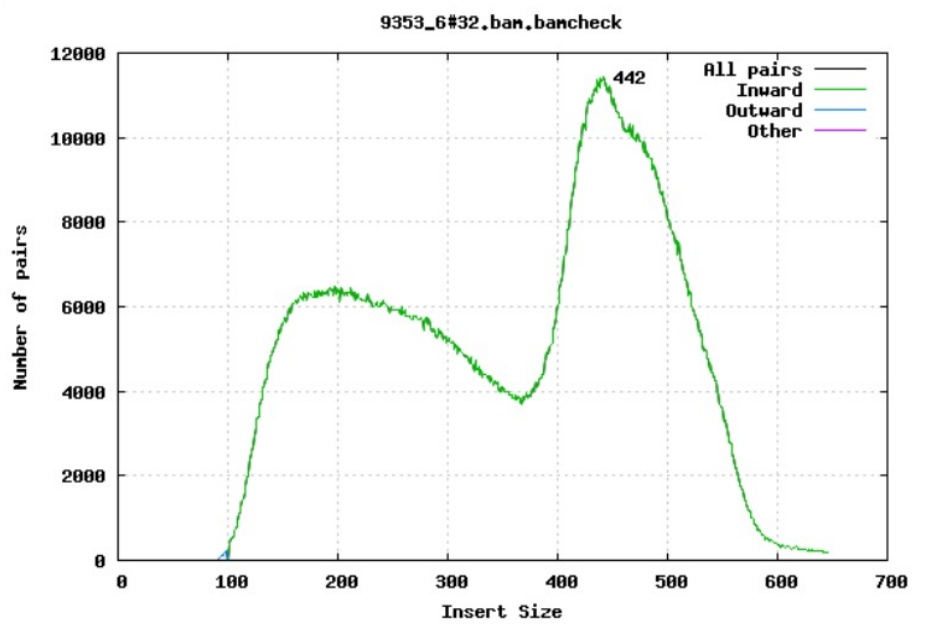




Fragment size

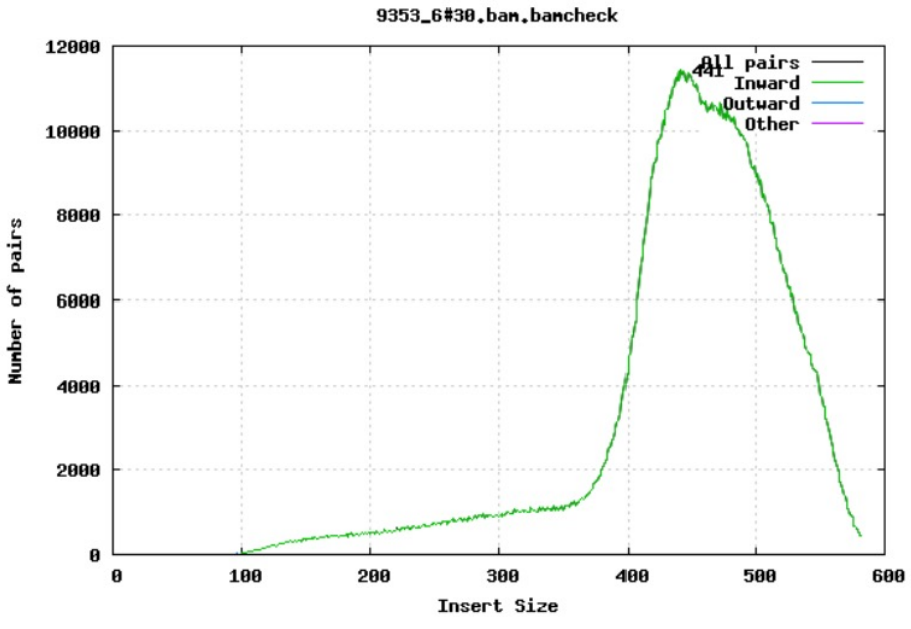
Paired-end sequencing: the size of DNA fragments matters

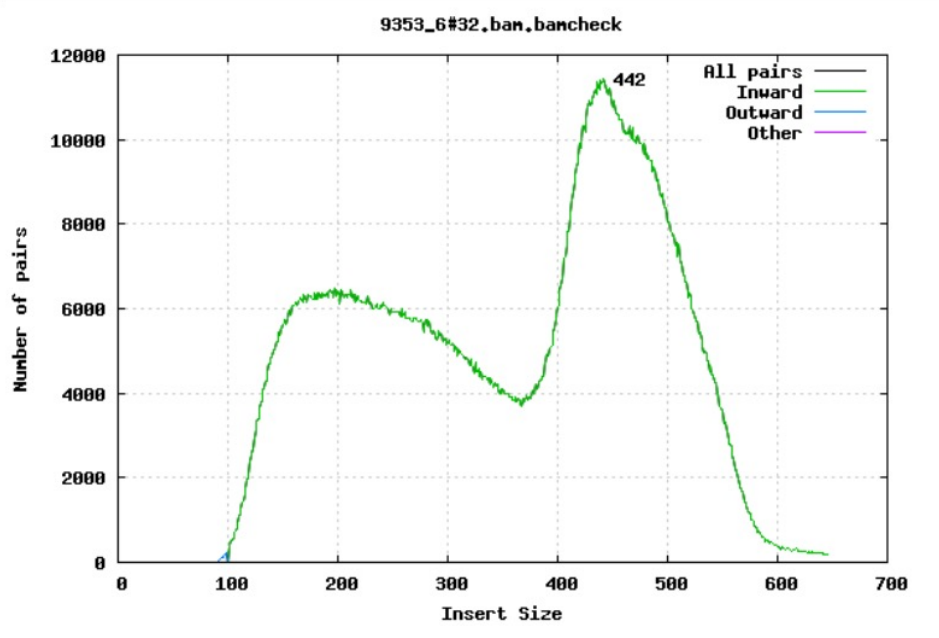


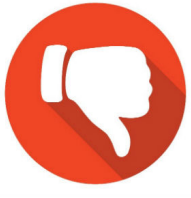


Fragment size

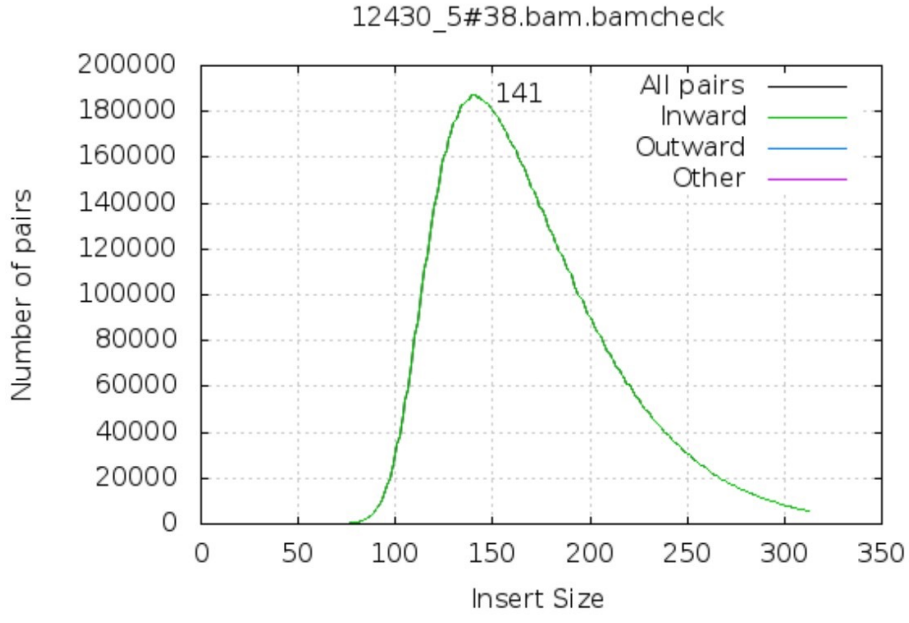
Paired-end sequencing: the size of DNA fragments matters





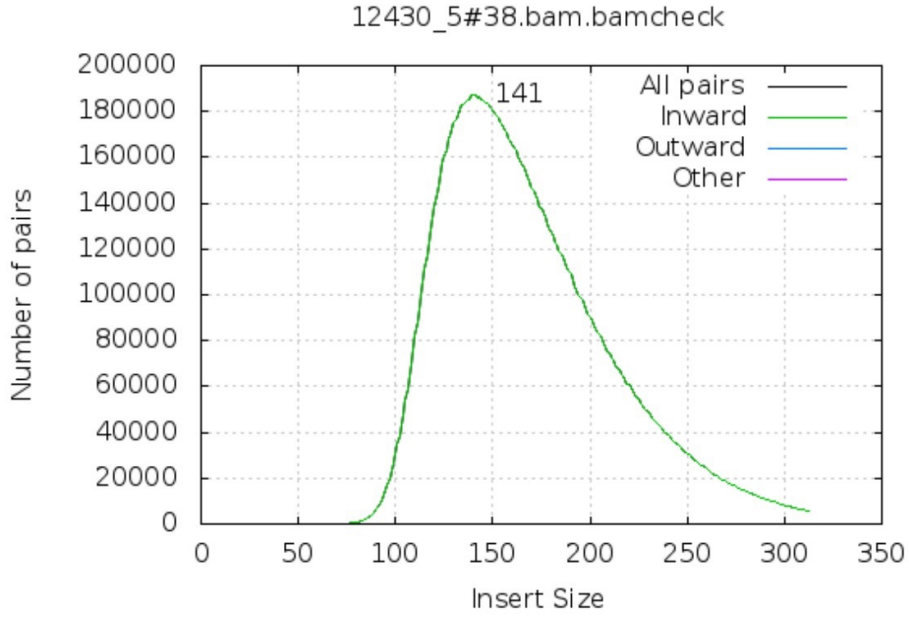


Quiz



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

Quiz



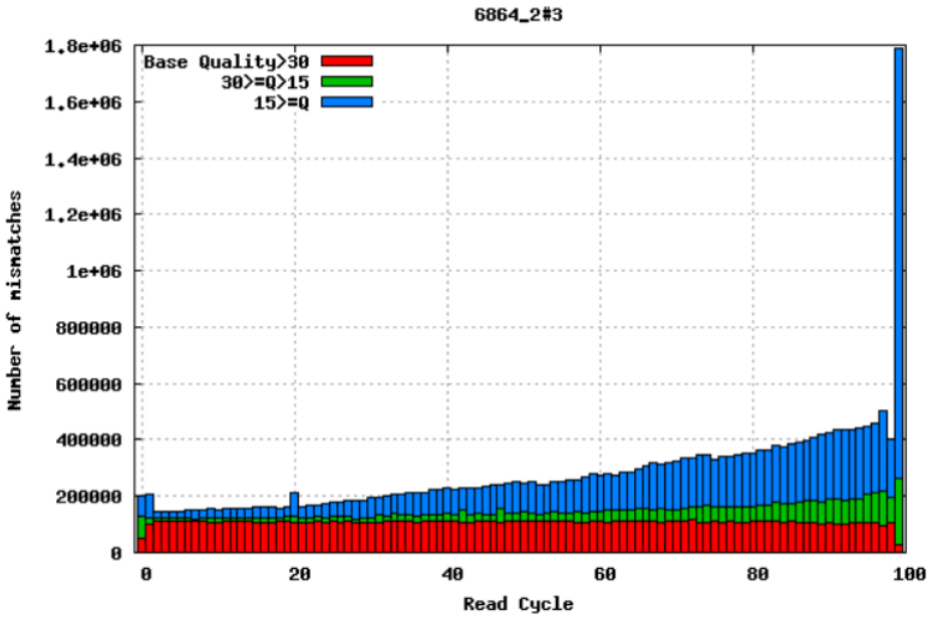
This is 100bp paired-end sequencing. Can you spot any problems??

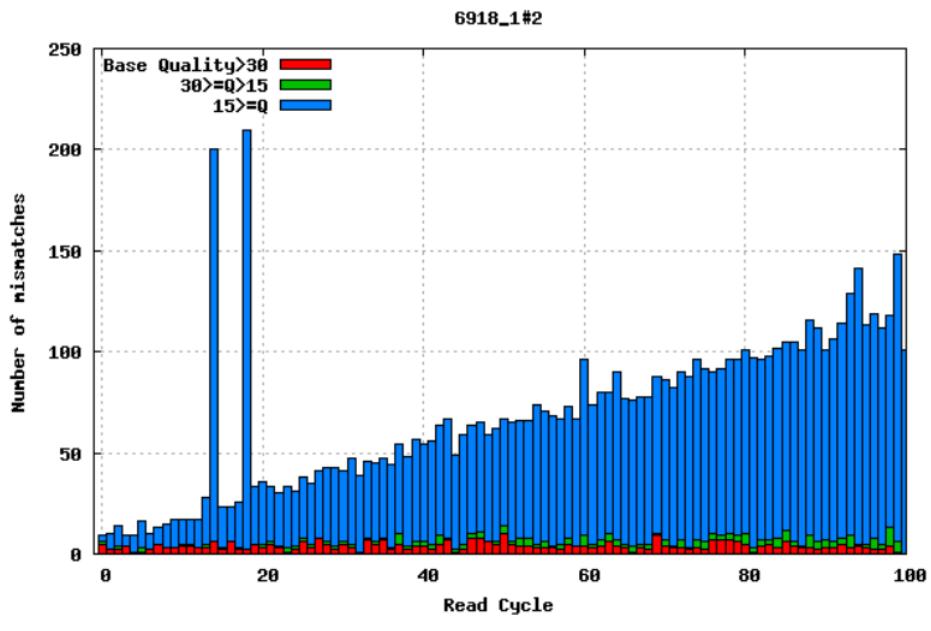
The insert size should be at least 200bp for the mates not to overlap.

Mismatches per cycle

Mismatches in aligned reads (requires reference sequence)

- detect cycle-specific errors
- Base qualities are informative!

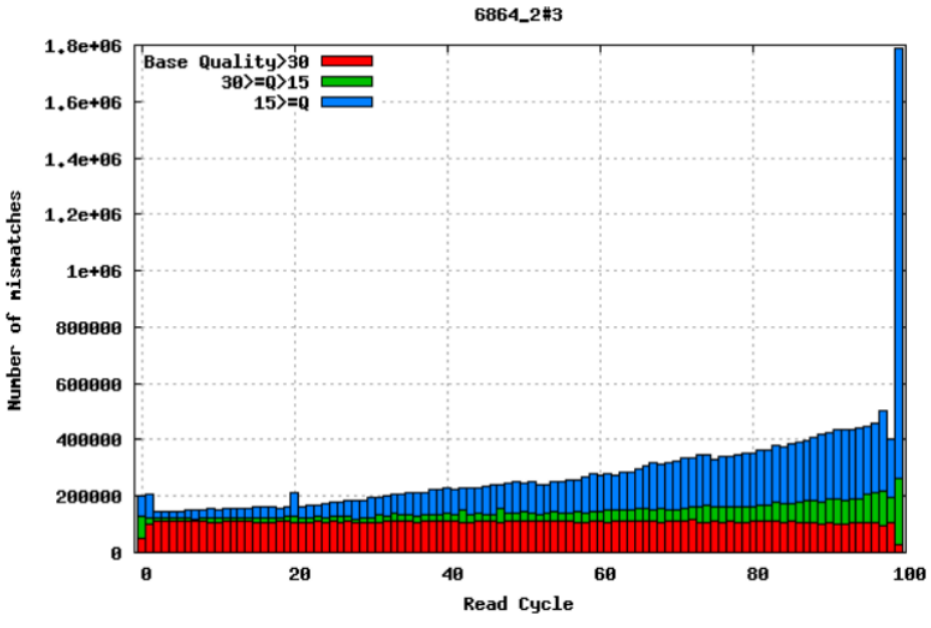


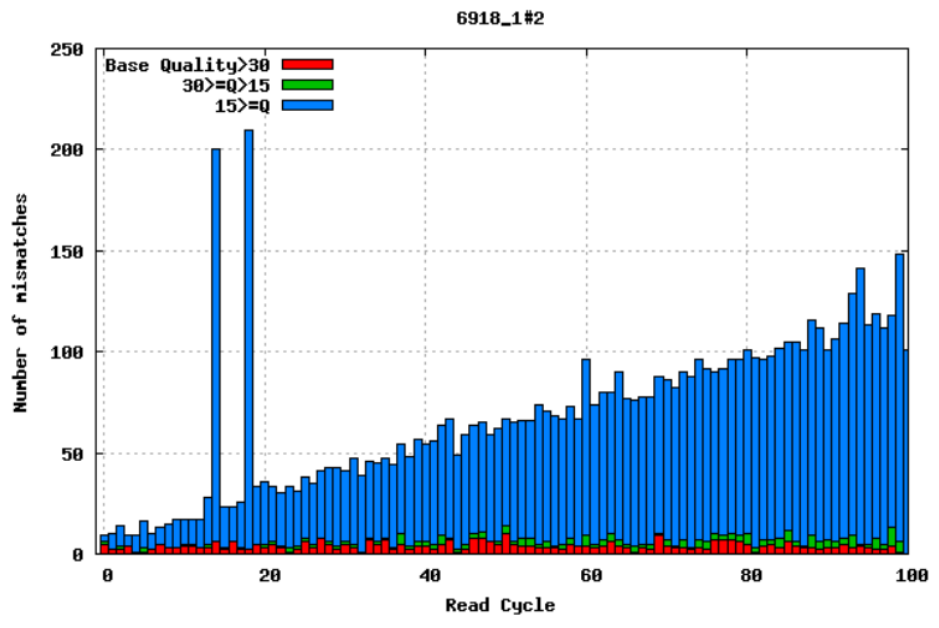


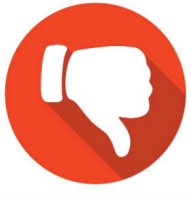
Mismatches per cycle

Mismatches in aligned reads (requires reference sequence)

- detect cycle-specific errors
- Base qualities are informative!



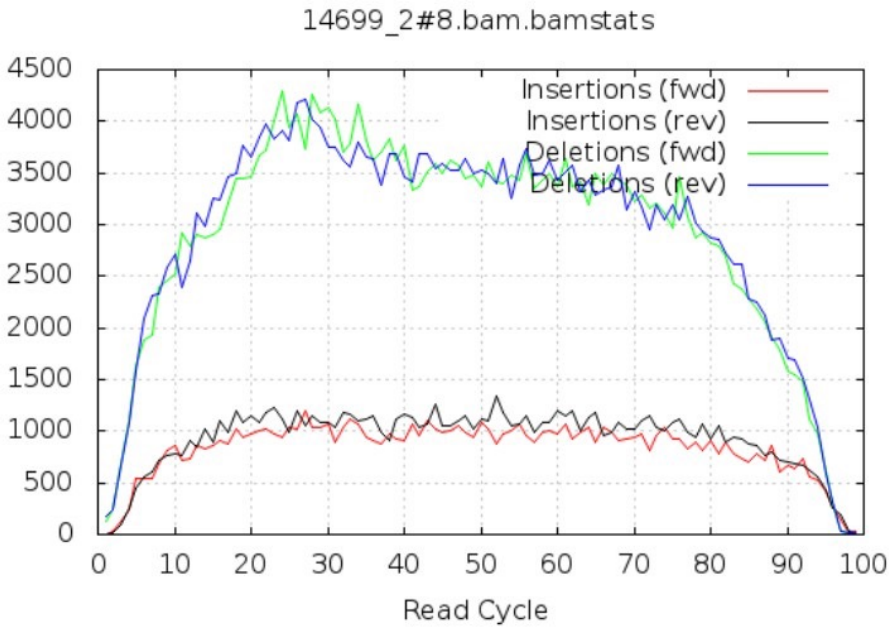


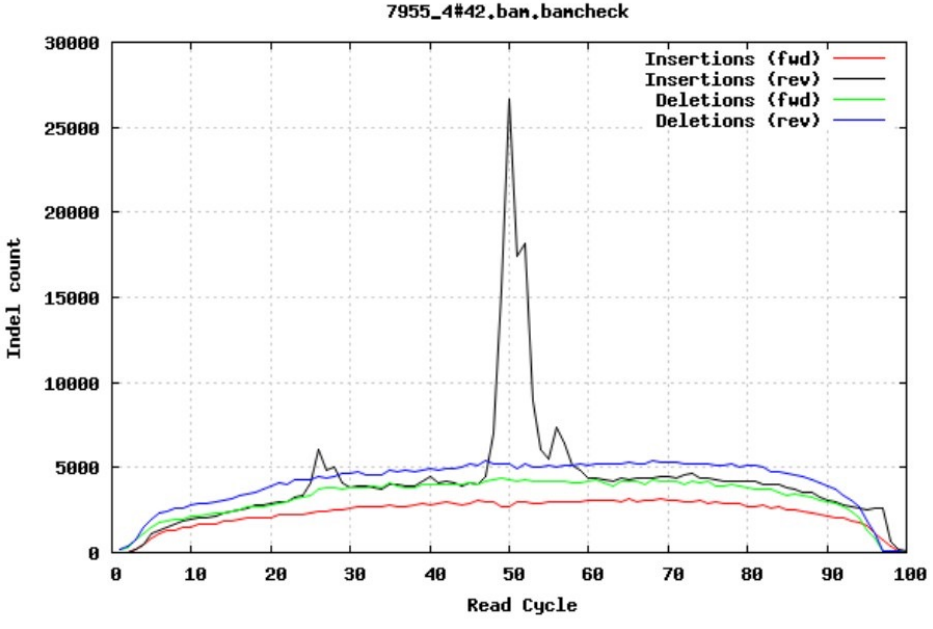


Insertions / Deletions per cycle

False indels

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

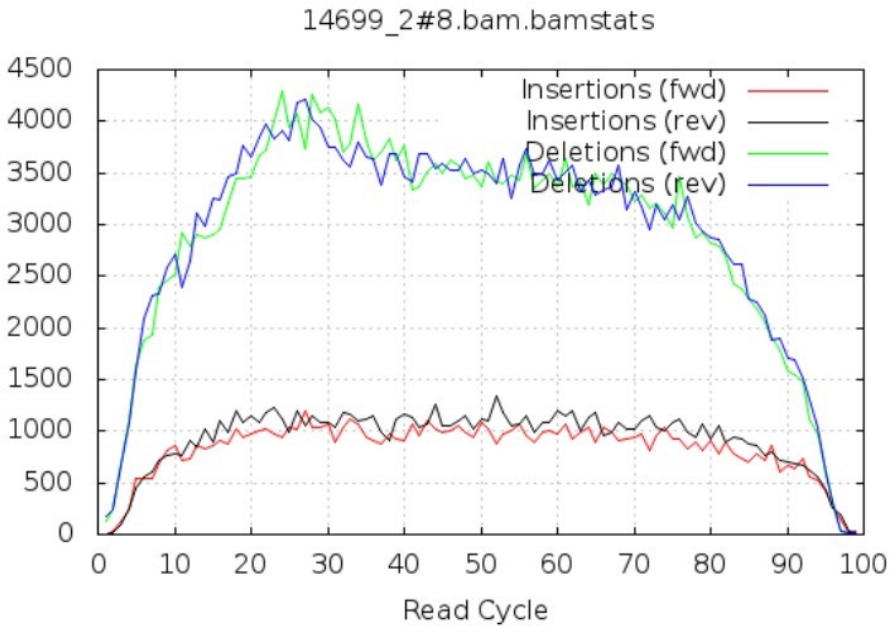


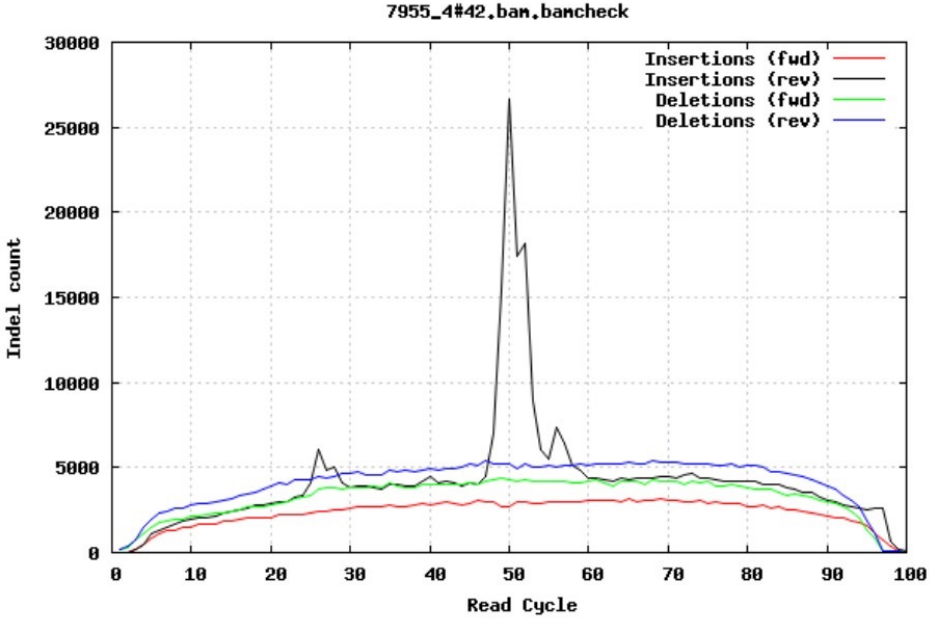


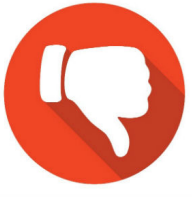
Insertions / Deletions per cycle

False indels

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



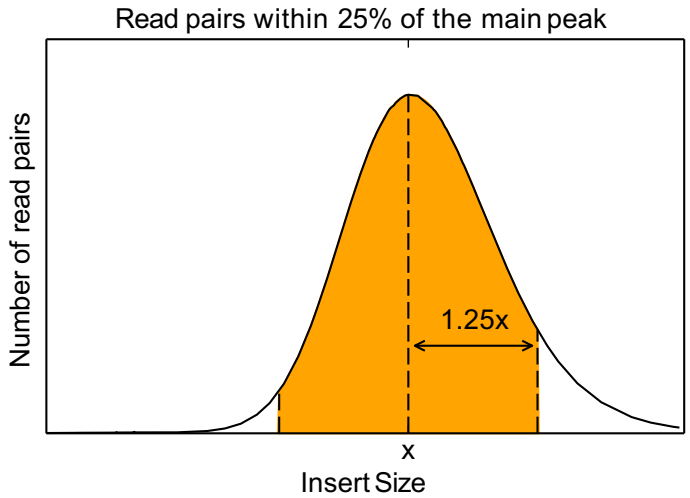




Auto QC tests

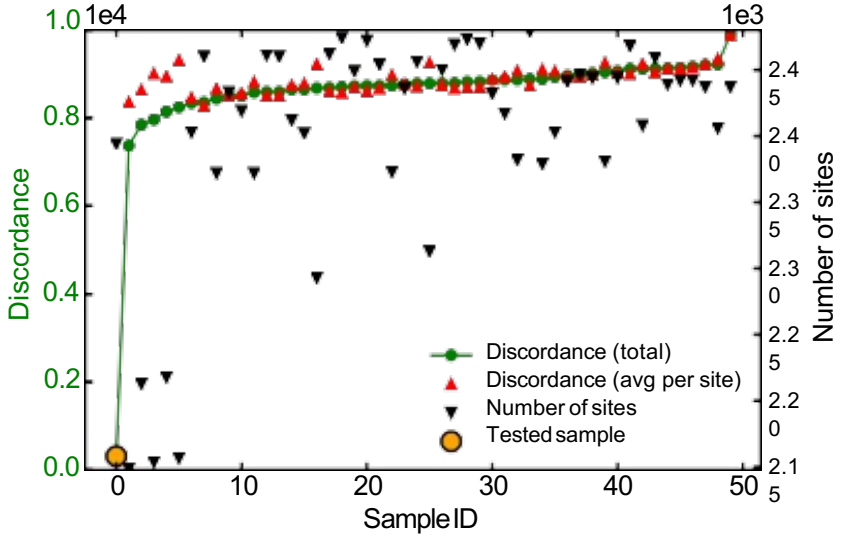
A suggestion for human data:

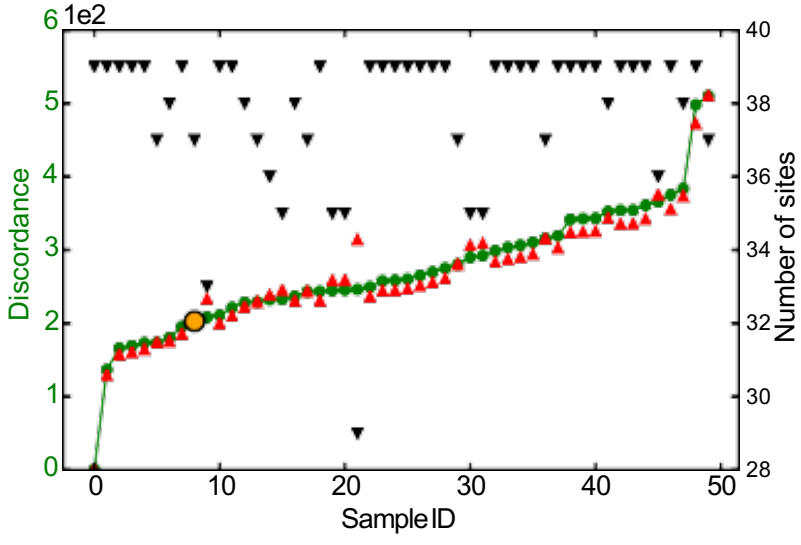
Minimum number of mapped bases	90%
Maximum error rate	0.02%
Maximum number of duplicate reads	5%
Minimum number of mapped reads which are properly paired	80%
Maximum number of duplicated bases due to overlapping read pairs	4%
Maximum in/del ratio	0.82
Minimum in/del ratio	0.68
Maximum indels per cycle,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

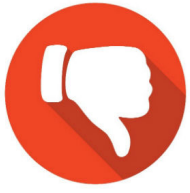


Detecting sample swaps

Check the identity against a known set of variants







Software

Software used to produce graphs in these slides

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

Exercise time!

- Open your VM
- Open a terminal window.
- Go to the Module 1 folder:

```
cd /home/manager/Module_1
```

- Open the exercises, which are in the GitHub

  Data formats and QC

- Follow the instructions!

Exercise time!

- Solutions will be shared through Slack!