HTS data formats and Quality Control

petr.danecek@sanger.ac.uk

wellcome trust sanger institute

Data Formats

FASTQ

- Unaligned read sequences with base qualities

SAM/BAM

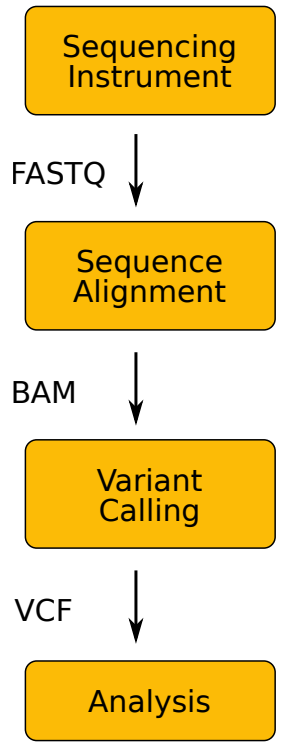
- Unaligned or aligned reads
- Text and binary formats

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

```
>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

```
Read 1  @ERR007731.739 IL16_2979:6:1:9:1684/1      <-- Read name
        CTTGACGACTTGAAAAATGACGAAATCACTAAAAAACGTGAAAAATGAGAAATG...   <-- Sequence
        +
        BBCBCBBBBBBBABBABBBBBBBABBBBBBBBBBBBBBABAAAABBBBB=@>BB...   <-- Base qualities
Read 2  @ERR007731.740 IL16_2979:6:1:9:1419/1
        AAAAAAAAAGATGTCATCAGCACATCAGAAAAGAAGGCAACTTTAAAACTTTTC...
        +
        BBABB/ABABAABABABBABBBAAA>@B@BBAA@4AAA>.>BAA@779:AAA@A...
```

- Simple format for raw unaligned sequencing reads
- Paired-end sequencing: two FASTQ files or one interleaved file

- Quality encoded in ASCII characters with decimal codes 33-126
  - ASCII code of "A" is 65, the corresponding quality is Q$= 65 - 33 = 32$

```
Base quality encoded as character
      ! " # $ % & ' ( ) * + , - . / 0 1 2 3 4 5 6 7 8 9 : ; < = > ? @ A B C D E F G H I J
Numeric ASCII value
     33 . . . . . . . . . . . . . . 47 . . . . . . . . . . . . . . . . . . . . 65 . . . . . . . . .
Base quality value
      0 . . . . . . . . . . . . . . 14 . . . . . . . . . . . . . . . . . . . . 32 . . . . . . . . .
                                                                              (65-33 = 32)
```

  - Beware: multiple quality scores were in use!
    - Sanger, Solexa, Illumina 1.3+
    - See https://en.wikipedia.org/wiki/FASTQ_format for details

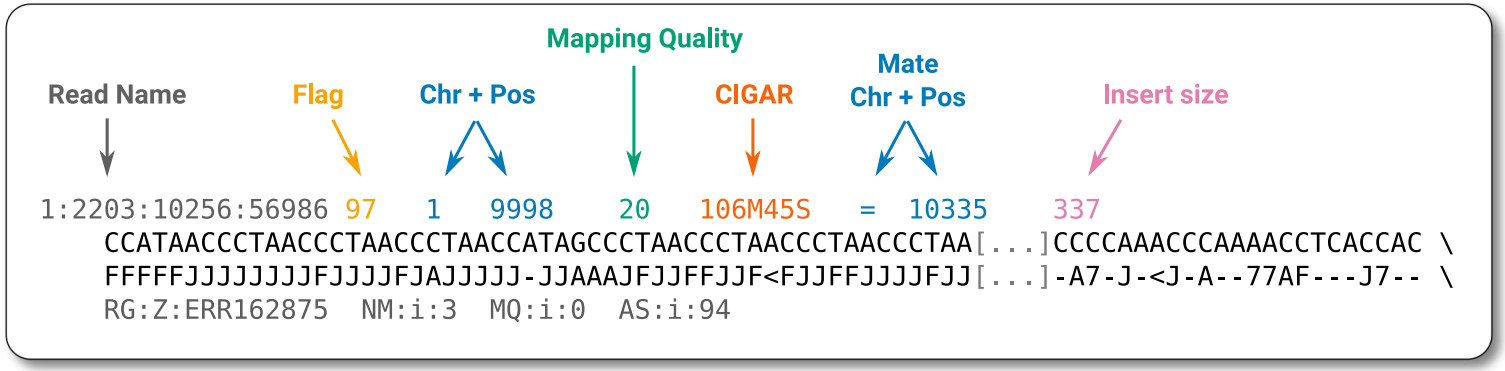
  - `perl -e 'printf "%d\n",ord("A")-33;'`

Quality = Phred-scaled probability of an error

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%

$$Q = -10 \log_{10} P \qquad \ldots \qquad P = 10^{-Q/10}$$

SAM / BAM: Sequence Alignment/Map format



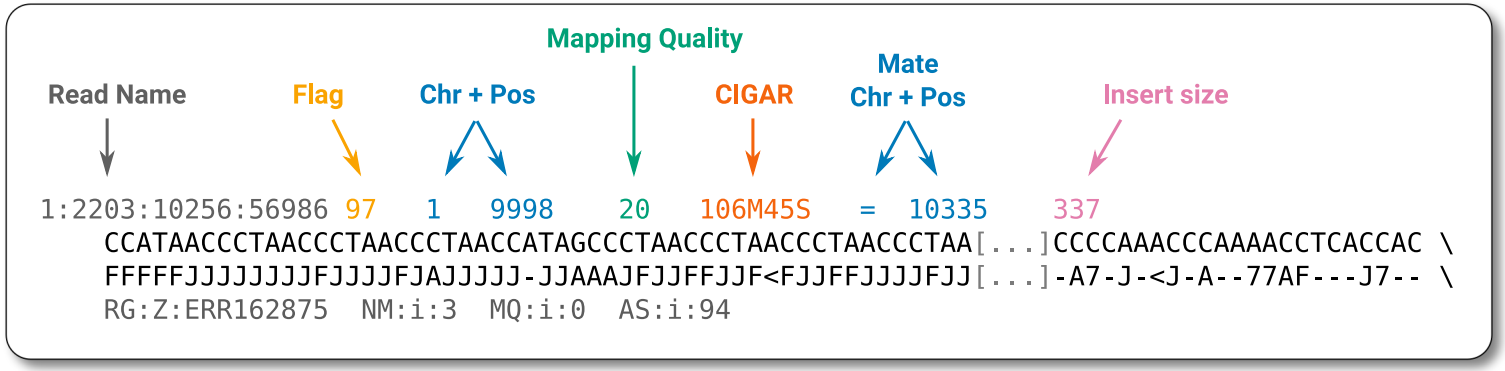
Flag

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

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

SAM / BAM: Sequence Alignment/Map format



CIGAR string

compact representation of sequence alignment:

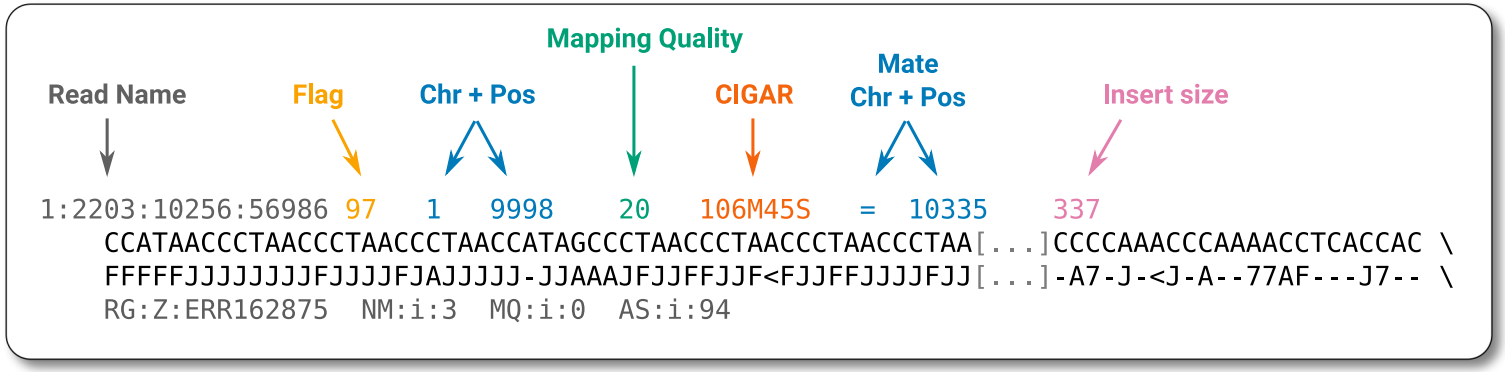
M	alignment match or mismatch
=	sequence match
X	sequence mismatch
I	insertion to the reference
D	deletion from the reference
S	soft clipping (clipped sequences present in SEQ)
H	hard clipping (clipped sequences NOT present in SEQ)
N	skipped region from the reference
P	padding (silent deletion from padded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

```
Ref:   ACGTACGTACTGT
Read:  ACGT----ACTGA
Cigar: 4M 4D 5M
```

```
Ref:   ACGT----ACGTA
Read:  ACGTACGTACGTA
Cigar: 4M 4I 5M
```

```
Ref:   CTCAGTG-GTCATCGTT
Read:  CGCA-TGAGTCTAGACG
Cigar: 4M 1D 2M 1I 3M 6S
```

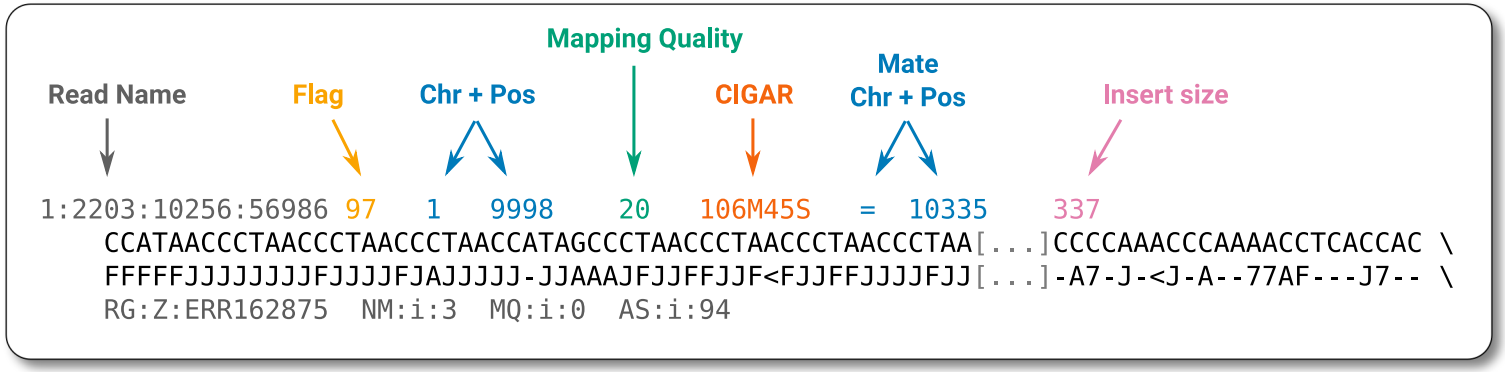
SAM / BAM: Sequence Alignment/Map format



Insert size

length of the DNA fragment sequenced from both ends by paired-end sequencing:

SAM / BAM: Sequence Alignment/Map format



Optional tags

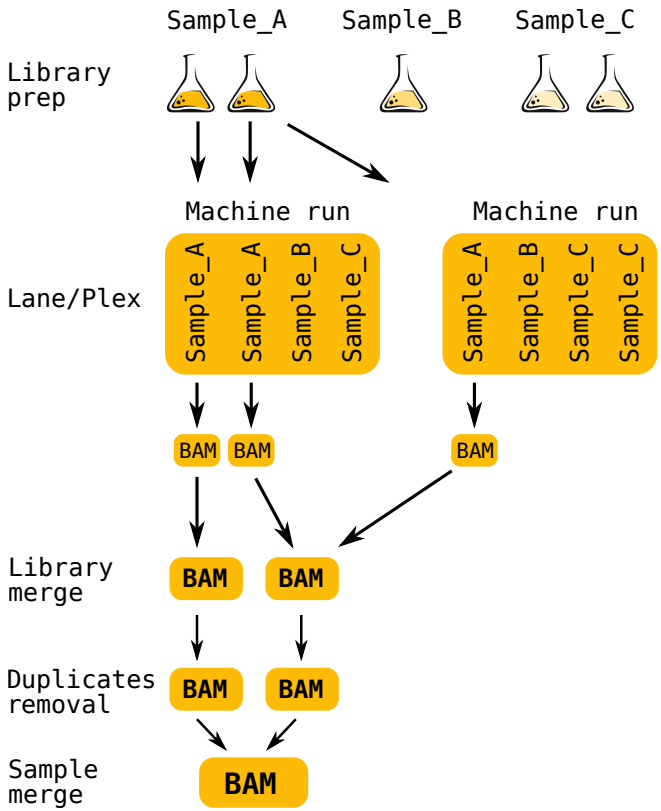
AS	Alignment score by the aligner
NM	Edit distance to the reference
MQ	Mapping quality of the mate
RG	Read group

Read Group

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

BAM specification

http://samtools.github.io/hts-specs/SAMv1.pdf
http://samtools.github.io/hts-specs/SAMtags.pdf



SAM / BAM tools

```
$ 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 \
    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
```

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

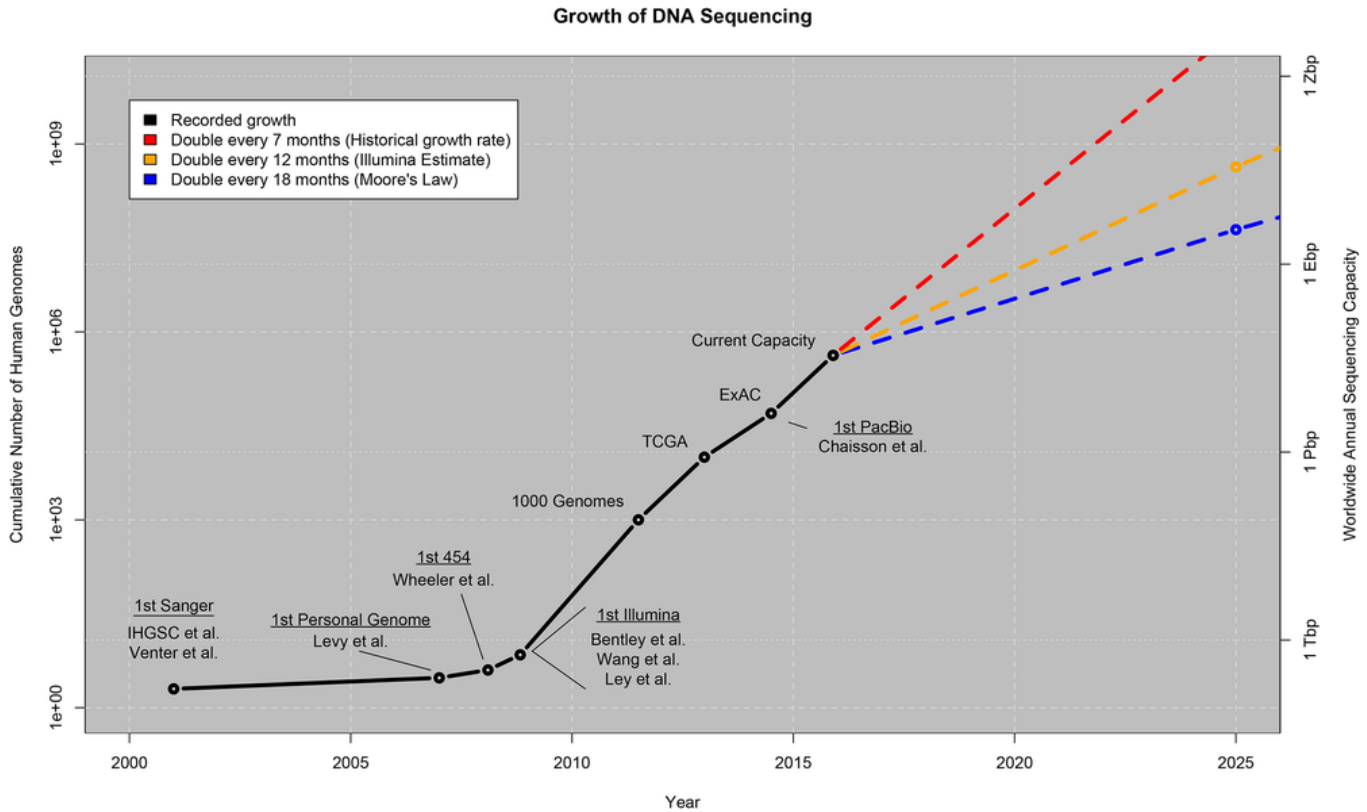
- IGV: http://www.broadinstitute.org/igv/
- BamView, LookSeq, Gap5, Tablet, Ensembl, UCSC, Bambino, Biodalliance...

CRAM: 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

CRAM: 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.............
```

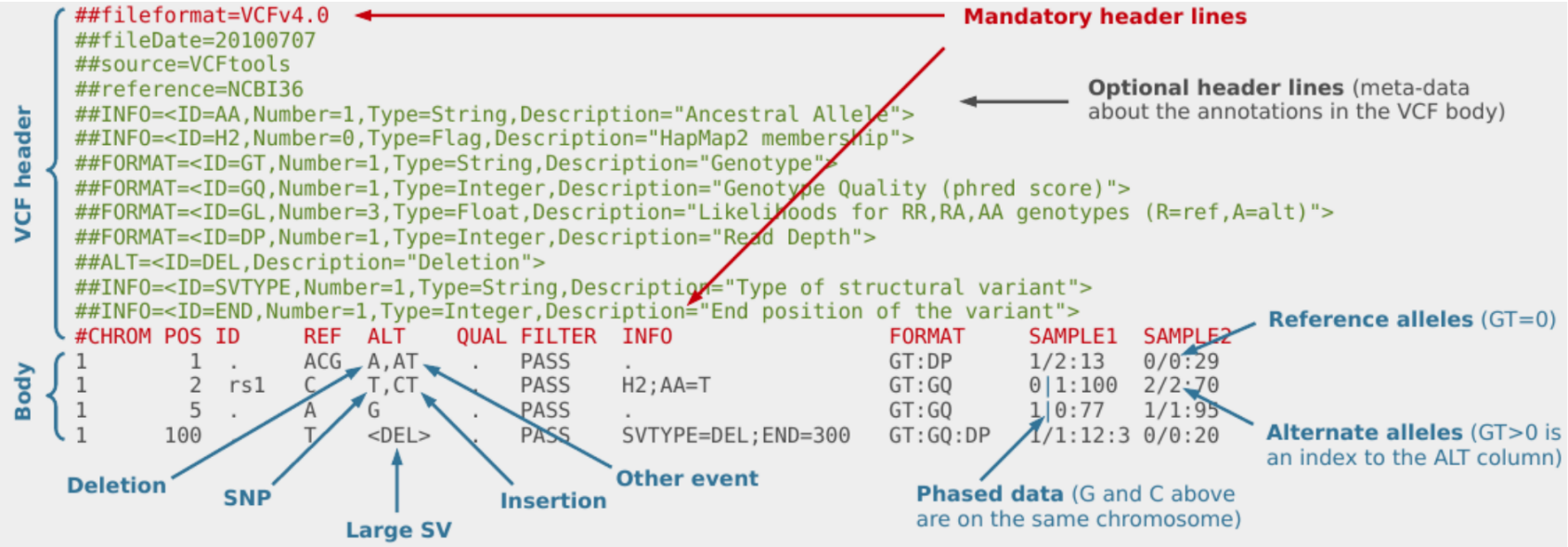
CRAM: in lossless mode 60% of BAM size

- Reference based compression
- Controlled loss of quality information
- Different compression methods for different type of data

Support for CRAM

- added to Samtools/HTSlib 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 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 anatomy

```
...
##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele frequency in population">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths (ref,alt,..)">
...
#CHROM POS  ID  REF ALT  QUAL FILTER  INFO          FORMAT  SAMPLE1    SAMPLE2   SAMPLE3
11   24535  .   G   A    243  PASS    DP=221;AF=0.5 GT:AD   0/1:73,15  0/0:48,0  0/1:71,14
```

Row-oriented, tab-delimited file with eight mandatory columns (CHROM-INFO)

VCF anatomy

```
...
##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele frequency in population">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths (ref,alt,..)">
...
#CHROM POS  ID  REF ALT  QUAL FILTER  INFO          FORMAT  SAMPLE1    SAMPLE2    SAMPLE3
11    24535 .   G   A    243  PASS    DP=221;AF=0.5 GT:AD   0/1:73,15  0/0:48,0   0/1:71,14
```

Genomic coordinates

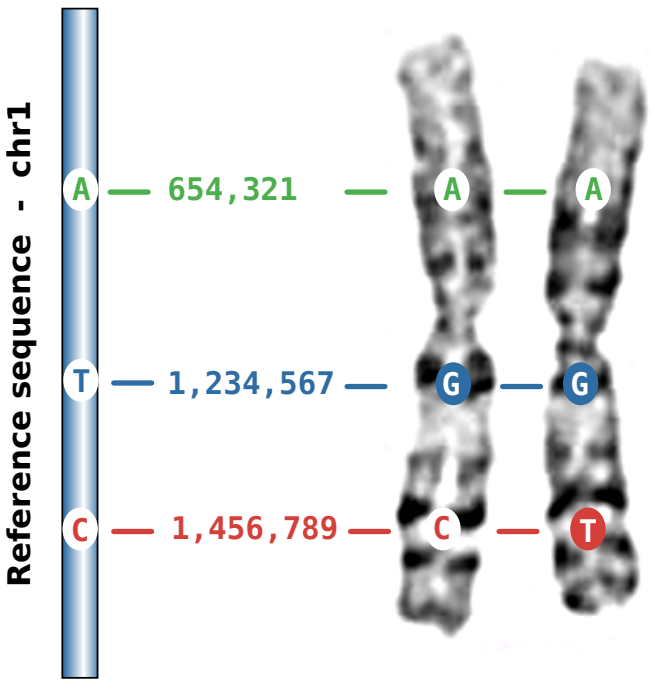
VCF anatomy

```
...
##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele frequency in population">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths (ref,alt,..)">
...
#CHROM POS   ID  REF ALT   QUAL FILTER  INFO          FORMAT  SAMPLE1    SAMPLE2   SAMPLE3
11    24535  .   G   A     243   PASS    DP=221;AF=0.5 GT:AD   0/1:73,15  0/0:48,0  0/1:71,14
```

Arbitrary string, typically a dbSNP RefSNP id. Dot for missing value.

VCF anatomy

```
...
##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele frequency in population">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths (ref,alt,..)">
...
#CHROM POS   ID  REF ALT  QUAL FILTER  INFO          FORMAT  SAMPLE1    SAMPLE2    SAMPLE3
11    24535  .   G   A    243  PASS    DP=221;AF=0.5 GT:AD   0/1:73,15  0/0:48,0   0/1:71,14
```



VCF anatomy

```
...
##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele frequency in population">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths (ref,alt,..)">
...
#CHROM POS   ID  REF ALT  QUAL FILTER  INFO          FORMAT  SAMPLE1    SAMPLE2   SAMPLE3
11    24535  .   G   A    243  PASS    DP=221;AF=0.5 GT:AD   0/1:73,15  0/0:48,0  0/1:71,14
```

Although in theory phred-scaled probability, don't expect truly probabilistic interpretation in practice.

VCF anatomy

```
...
##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele frequency in population">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths (ref,alt,..)">
...
#CHROM POS   ID  REF ALT   QUAL FILTER INFO           FORMAT  SAMPLE1    SAMPLE2   SAMPLE3
11    24535  .   G   A     243  PASS   DP=221;AF=0.5  GT:AD   0/1:73,15  0/0:48,0  0/1:71,14
```

Soft-filter variants with e.g. low quality, low depth, etc.

VCF anatomy

```
...
##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele frequency in population">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths (ref,alt,..)">
...
#CHROM POS   ID  REF ALT   QUAL FILTER INFO        FORMAT  SAMPLE1    SAMPLE2    SAMPLE3
11    24535  .   G   A     243  PASS   DP=221;AF=0.5 GT:AD   0/1:73,15  0/0:48,0   0/1:71,14
```

Per-site annotations. Here **DP** is the cumulative read depth across all samples and **AF** allele frequency of the allele in general population.

VCF anatomy

```
...
##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele frequency in population">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths (ref,alt,..)">
...
#CHROM POS   ID  REF ALT   QUAL FILTER  INFO          FORMAT  SAMPLE1    SAMPLE2   SAMPLE3
11    24535  .   G   A     243  PASS    DP=221;AF=0.5 GT:AD   0/1:73,15  0/0:48,0  0/1:71,14
```

Per-sample annotations. Here **GT** (genotype) and **AD** (allelic depth) will be present for each sample.

VCF anatomy

```
...
##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele frequency in population">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths (ref,alt,..)">
...
#CHROM POS  ID  REF ALT  QUAL FILTER  INFO          FORMAT  SAMPLE1    SAMPLE2   SAMPLE3
11   24535  .   G   A    243  PASS    DP=221;AF=0.5 GT:AD   0/1:73,15  0/0:48,0  0/1:71,14
```

Per-sample values listed in the same order as specified in the FORMAT column, separated by a colon.

VCF anatomy

```
...
##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele frequency in population">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths (ref,alt,..)">
...
#CHROM POS   ID  REF ALT   QUAL FILTER  INFO            FORMAT  SAMPLE1    SAMPLE2    SAMPLE3
11    24535  .   G   A     243  PASS    DP=221;AF=0.5   GT:AD   0/1:73,15  0/0:48,0   0/1:71,14
12   153927  .   C   CA,T  15   LowQ    AF=0,0.1        GT      2/2        1/2        0/1
```

Multiple alternate alleles can be present in one row.

VCF anatomy

```
...
##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele frequency in population">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths (ref,alt,..)">
...
#CHROM POS   ID  REF ALT   QUAL FILTER  INFO          FORMAT  SAMPLE1   SAMPLE2   SAMPLE3
11    24535  .   G   A     243  PASS    DP=221;AF=0.5 GT:AD   0/1:73,15 0/0:48,0  0/1:71,14
12   153927  .   C   CA,T  15   LowQ    AF=0,0.1      GT      2/2       1/2       0/1
```

All variation types can be represented:

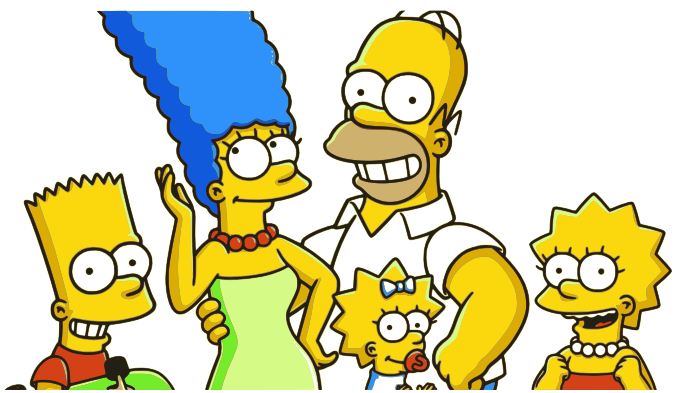
MNP	POS:	12345678	POS	REF	ALT
	REF:	**ACGTACGT**	**3**	**GT**	**TA**
	ALT:	**ACTAACGT**			
Deletion		**ACGTACGT**	**2**	**CGT**	**C**
		AC--ACGT			
Insertion		**AC--ACGT**	**2**	**C**	**CGT**
		ACGTACGT			
Structural variation			**2**	**C**	**<DEL>**
			2	**C**	**<DUP>**

VCF anatomy

```
...
##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele frequency in population">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths (ref,alt,..)">
...
#CHROM POS   ID  REF ALT   QUAL FILTER  INFO            FORMAT  SAMPLE1    SAMPLE2    SAMPLE3
11    24535  .   G   A     243   PASS    DP=221;AF=0.5   GT:AD   0/1:73,15  0/0:48,0   0/1:71,14
12   153927  .   C   CA,T   15   LowQ    AF=0,0.1        GT      2/2        1/2        0/1
                 0   1  2                                       T/T        CA/T       C/CA
```

Genotype (GT) is represented as a 0-based index into the array of REF and ALT alleles

One file can contain zero, one or many samples



Genome VCF (gVCF)

VCF

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
19	9902	.	G	.	.	.	DP=0
19	9903	.	T	.	.	.	DP=0
19	9904	.	A	.	.	.	DP=0
19	9905	.	C	.	.	.	DP=0
19	9906	.	G	.	.	.	DP=5
19	9907	.	T	.	.	.	DP=7
19	9908	.	A	.	.	.	DP=10
19	9909	.	C	.	.	.	DP=13
19	9910	.	G	A	.	.	DP=15
19	9911	.	T	.	.	.	DP=14
19	9912	.	A	.	.	.	DP=19
19	9913	.	C	.	.	.	DP=23
19	9914	.	G	.	.	.	DP=22
19	9915	.	T	.	.	.	DP=17
19	9916	.	G	T	.	.	DP=18
19	9917	.	A	.	.	.	DP=19
19	9918	.	C	.	.	.	DP=16
19	9919	.	G	.	.	.	DP=25
19	9920	.	T	.	.	.	DP=23

→

gVCF

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
19	9902	.	G	.	.	.	MinDP=0;END=9905
19	9906	.	G	.	.	.	MinDP=5;END=9909
19	9910	.	G	A	.	.	DP=15
19	9911	.	T	.	.	.	MinDP=14;END=9915
19	9916	.	G	T	.	.	DP=18
19	9917	.	A	.	.	.	MinDP=16;END=9920

Often it is not sufficient to keep only *variant* sites:

- is there **no alternate allele** or is there **no coverage**???
- need evidence for both variant and non-variant positions in the genome

VCF vs 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
```

Quality Control

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

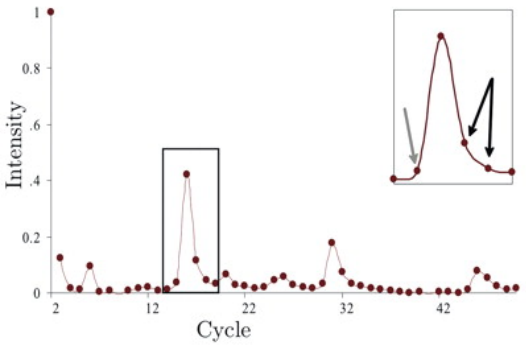
- be careful to distinguish between the total sequencing yield and on-target bases

Useful coverage

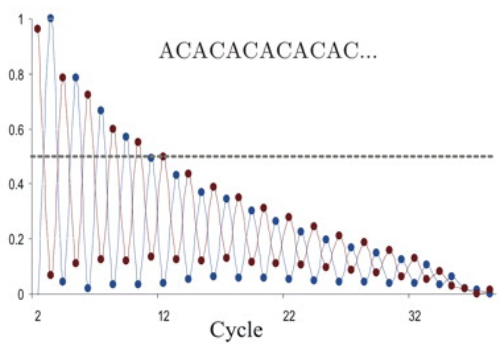
- 15x ok for common germline variants
- 30x ok for most things
- 100-200x for low VAF variants in tumors

Base calling errors

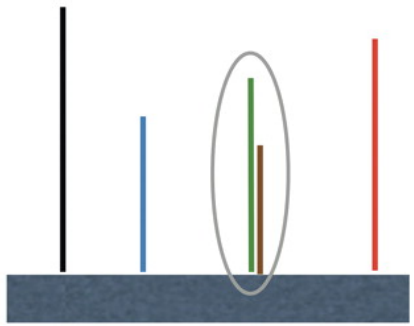
Phasing noise ϕ



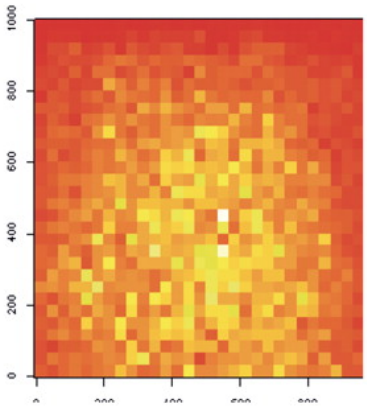
Signal Decay δ



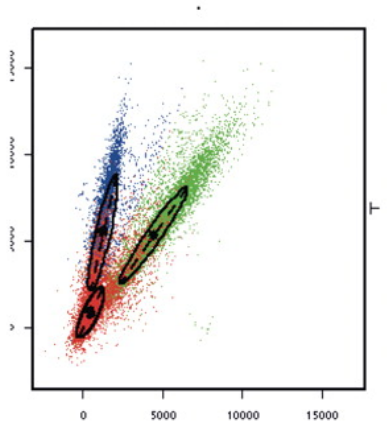
Mixed Cluster μ



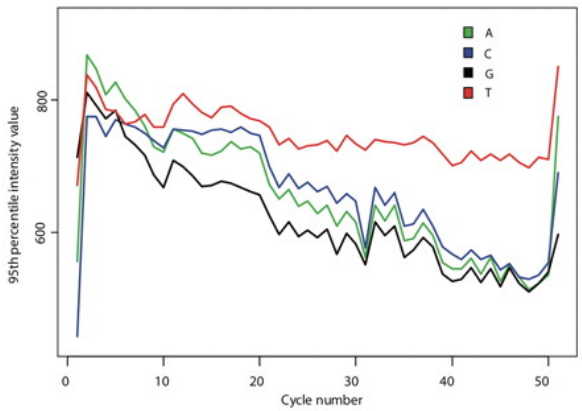
Boundary effects ω



Cross-talk Σ



T fluophore accumulation τ



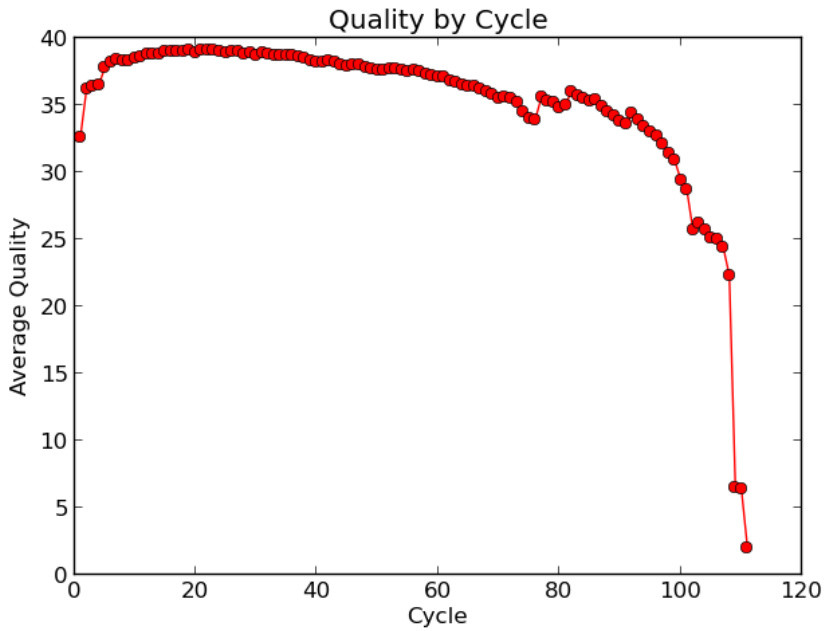
Base-calling for next-generation sequencing platforms, doi: 10.1093/bib/bbq077

Base quality

Sequencing by synthesis: dephasing

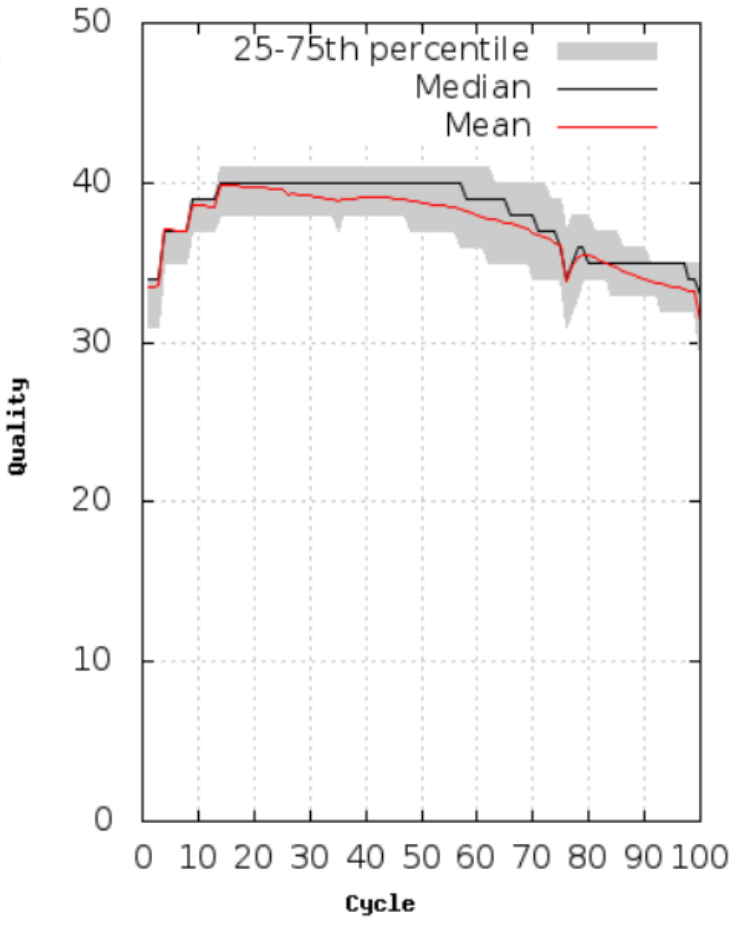
- growing sequences in a cluster gradually desynchronize
- error rate increases with read length

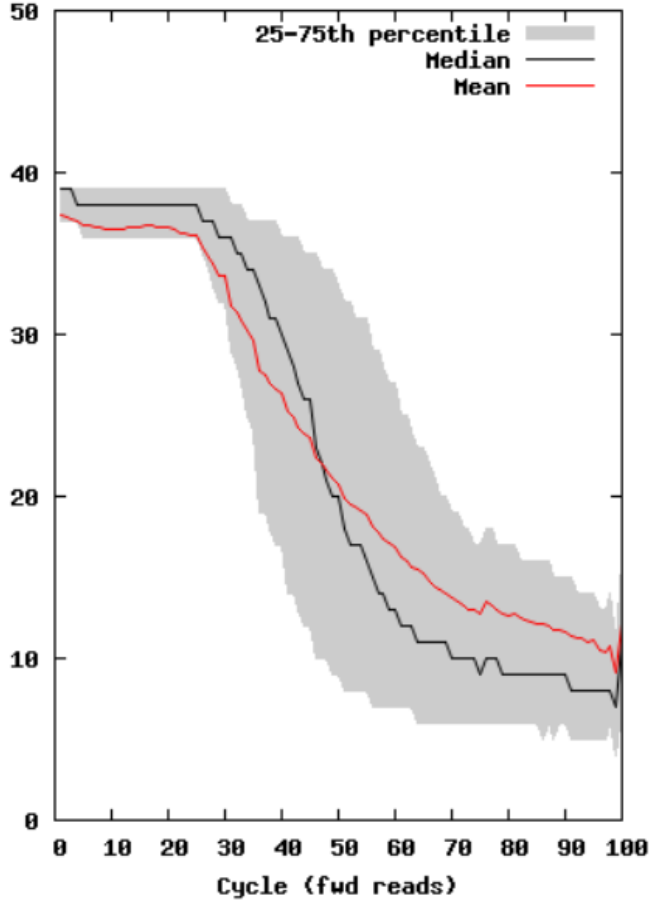
Calculate the average quality at each position across all reads



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





Library prep biases: PCR duplicates

Experiments start with small amounts of DNA

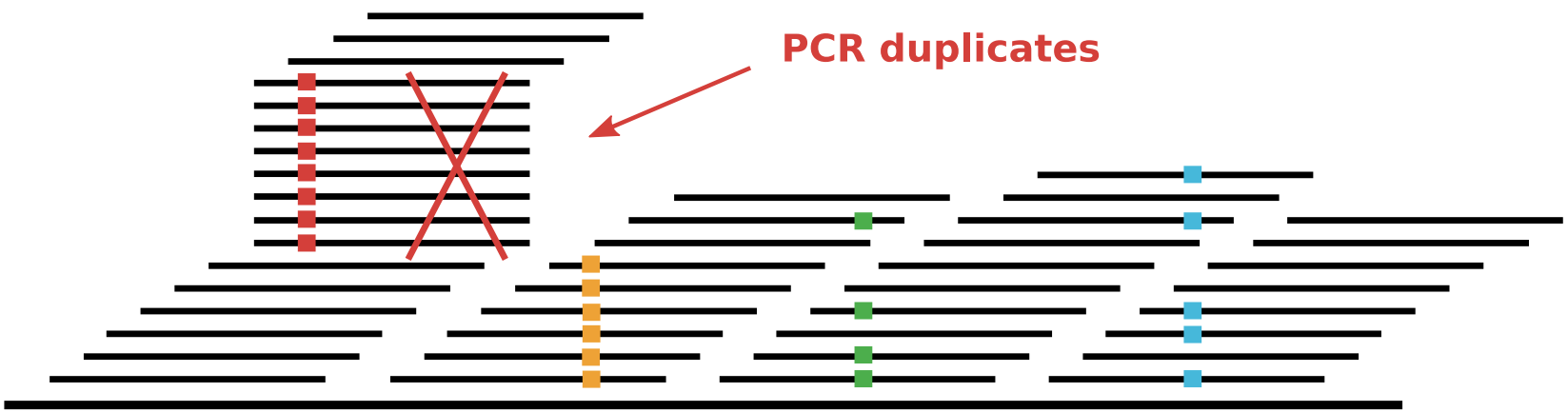
- a PCR amplification step is necessary for Illumina sequencing: one molecule => many identical molecules

Problem:

- additional PCR-copy molecules are not informative

Solution:

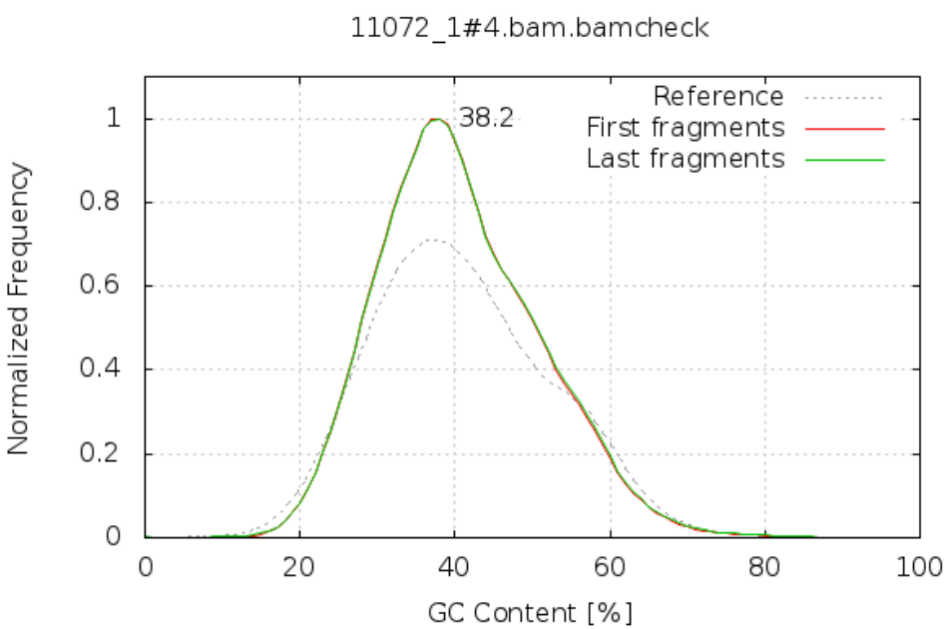
- infer and mark PCR-dupliates, discount in later analysis
  - mark if reads and their mates start at the same position
- use `picard MarkDuplicates` or `samtools markdup`
- typical dup rates: Exomes ~ 15-20%, Genomes < 5%

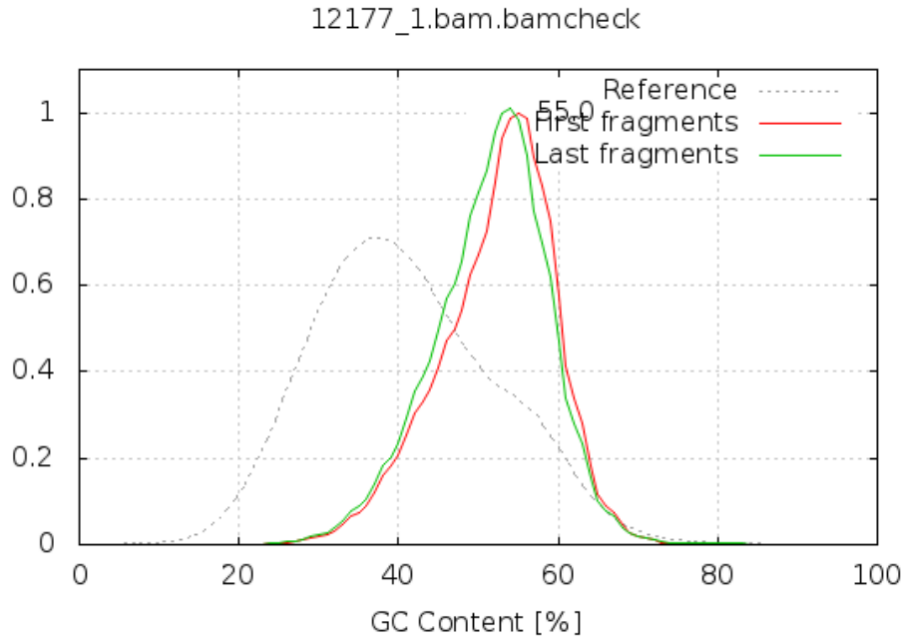


GC bias

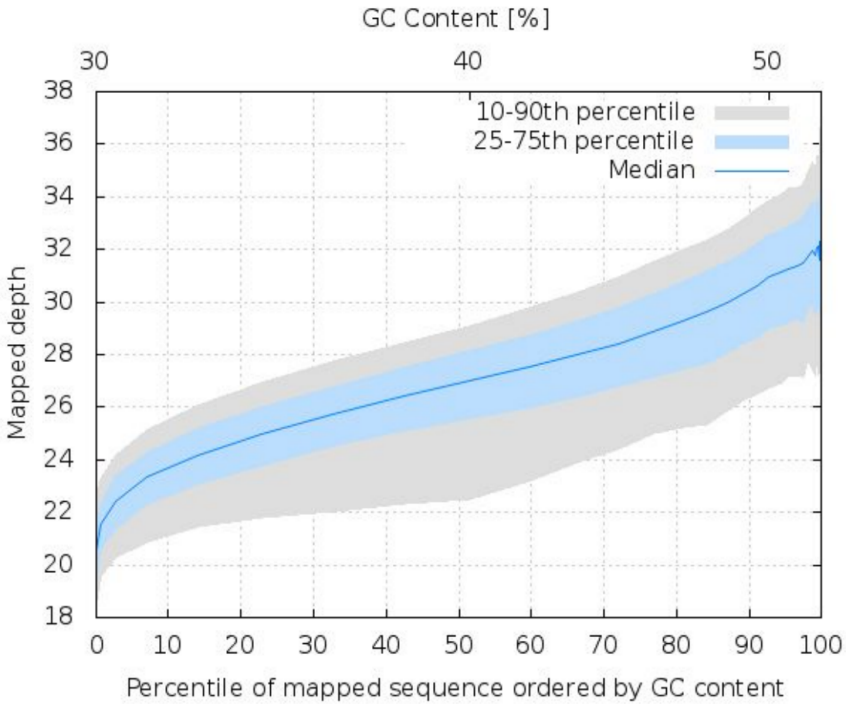
GC- and AT-rich regions are more difficult to amplify

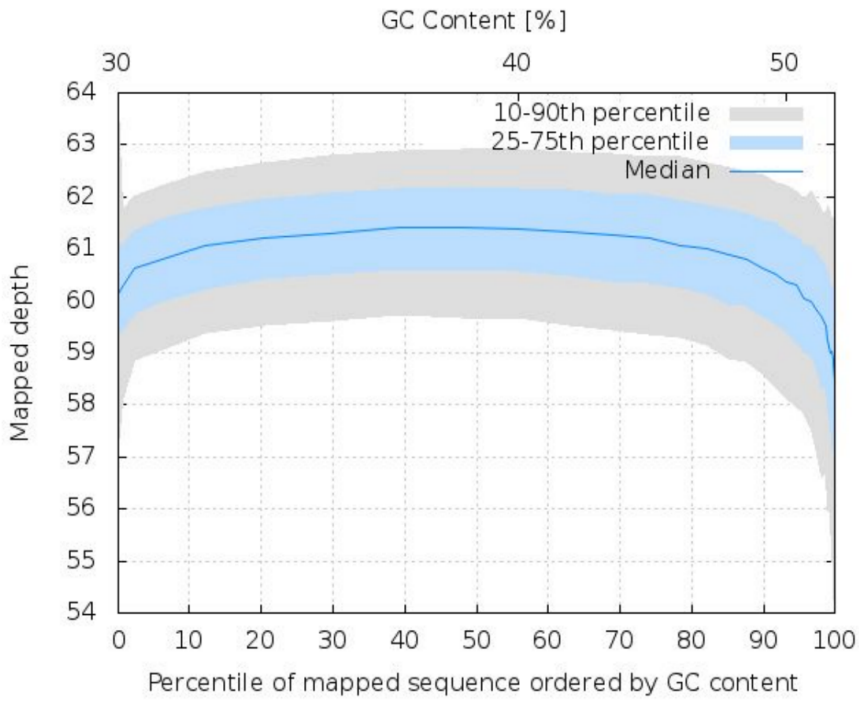
- compare the GC content against the expected distribution (reference sequence)





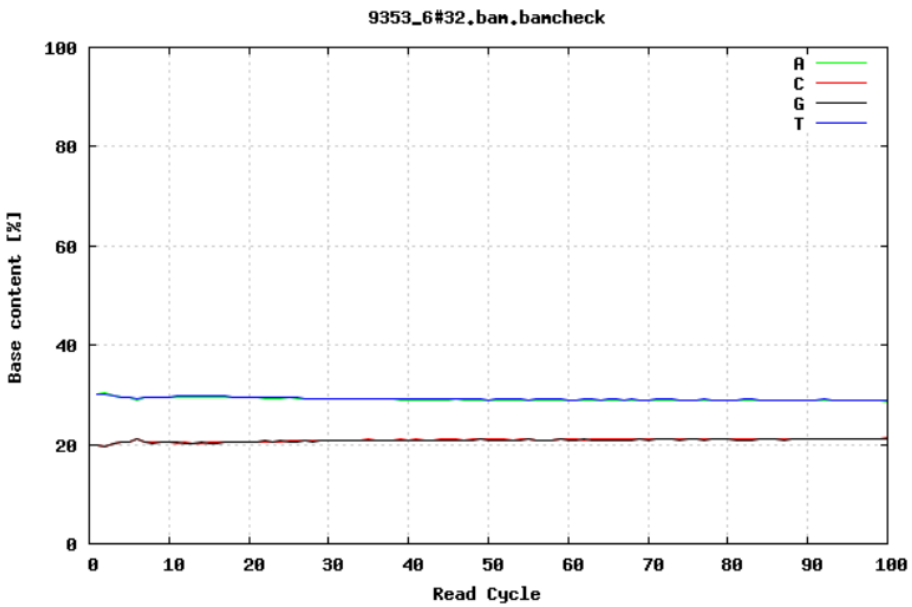
GC content vs depth

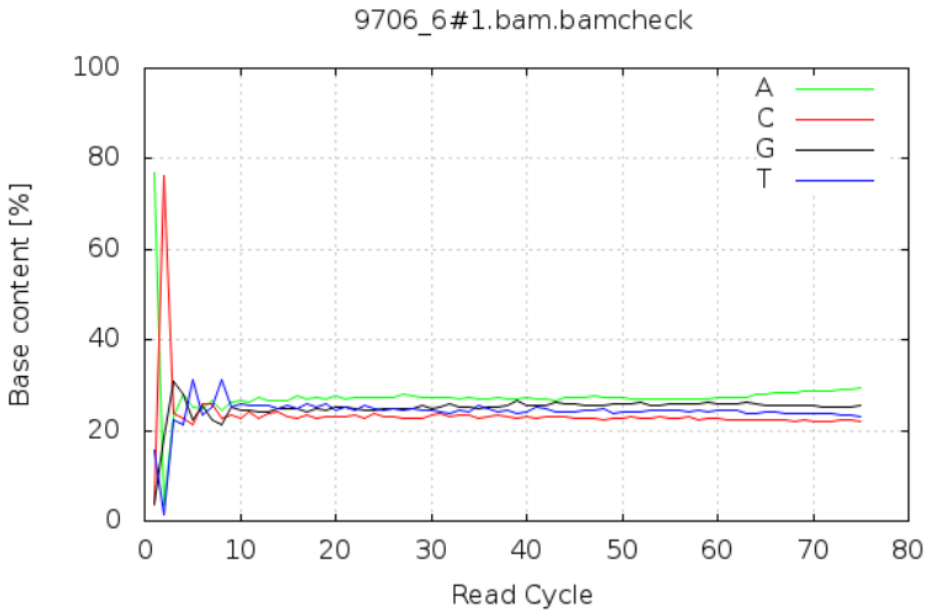




GC content by cycle

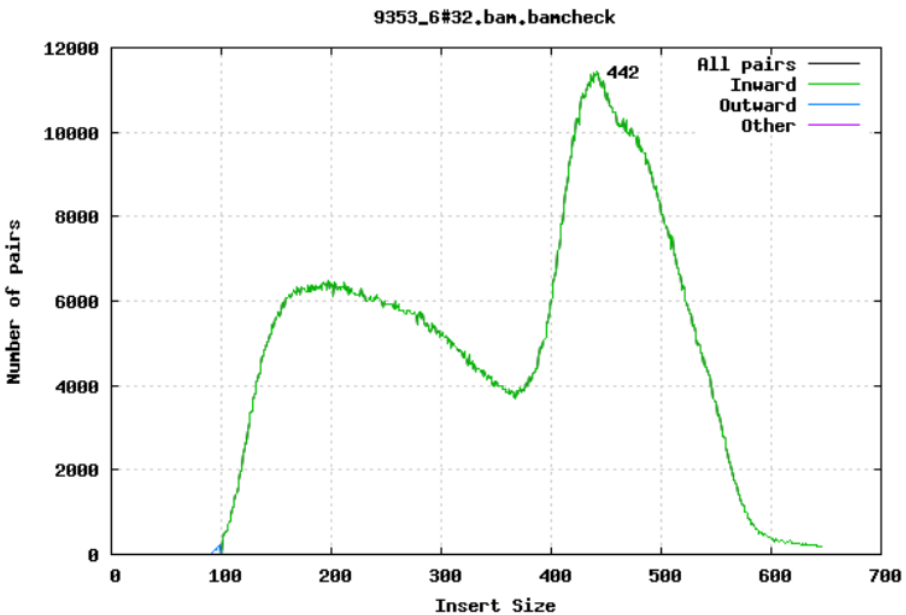
Was the adapter sequence trimmed?

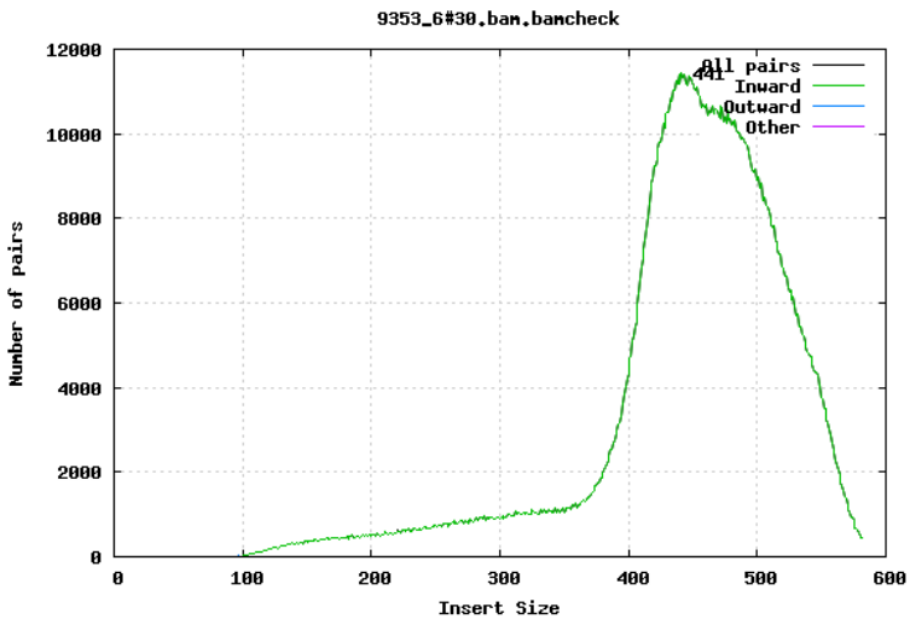




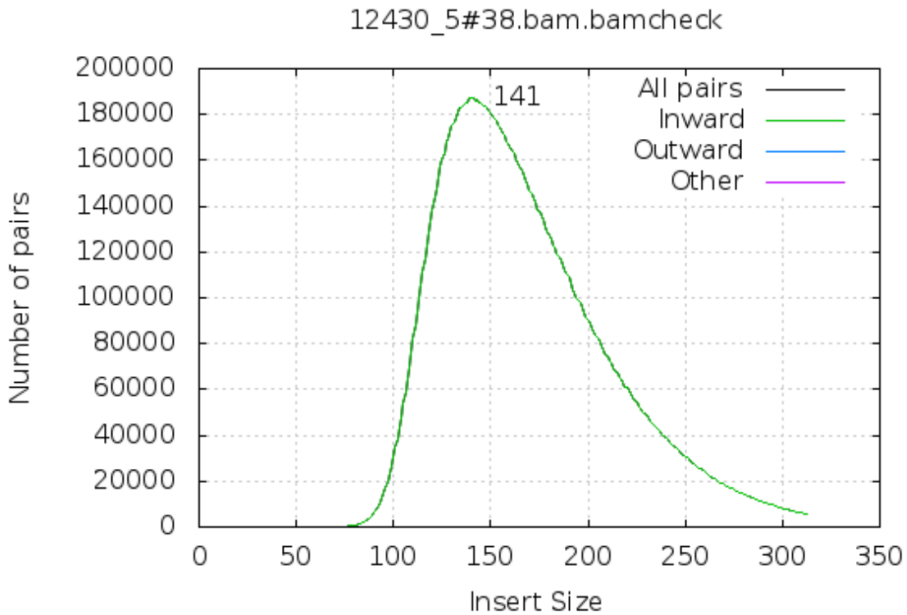
Fragment size

Paired-end sequencing: the size of DNA fragments matters





Quiz

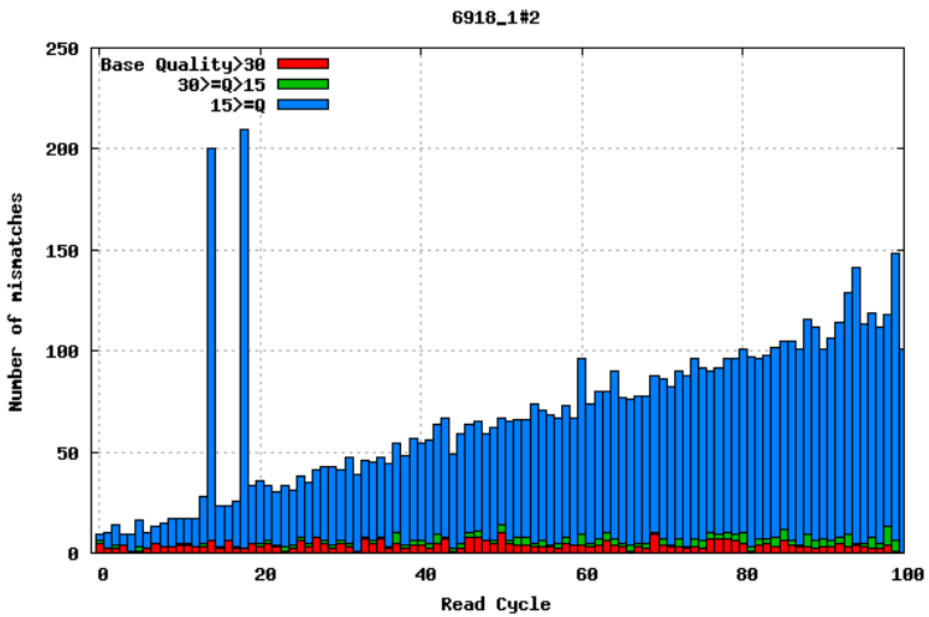


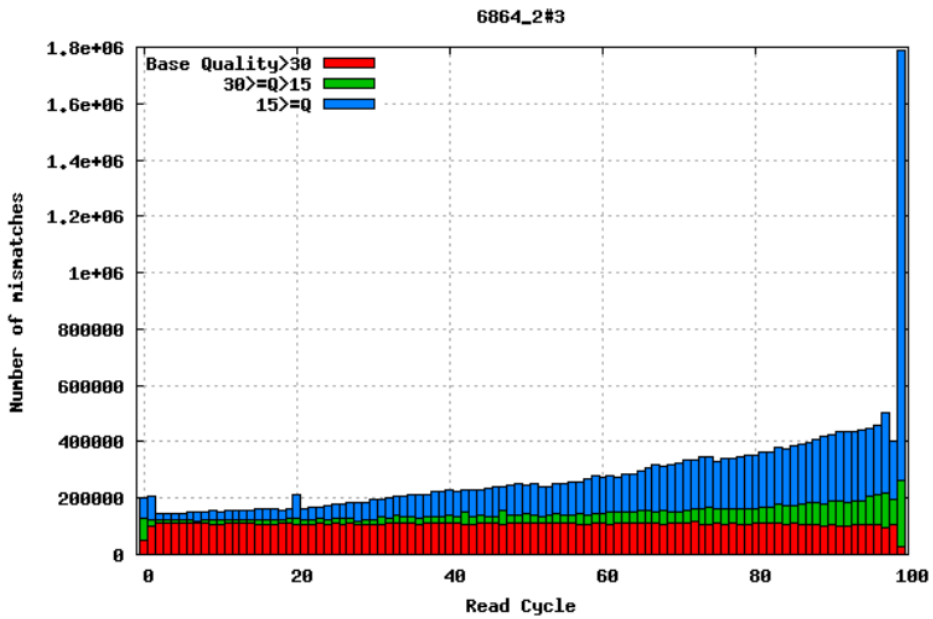
This is 100bp paired-end sequencing. Can you spot any problems??

Mismatches per cycle

Mismatches in aligned reads (requires reference sequence)

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

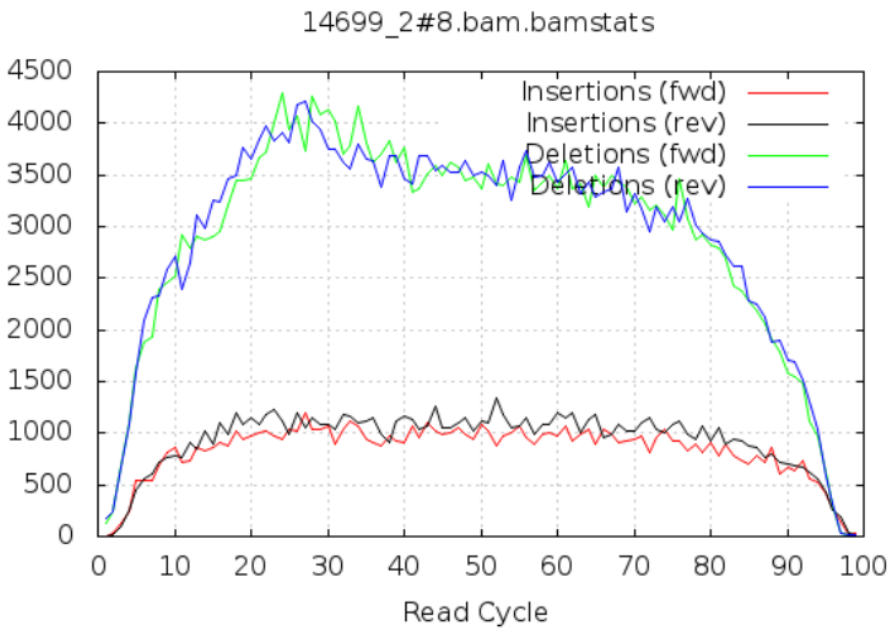


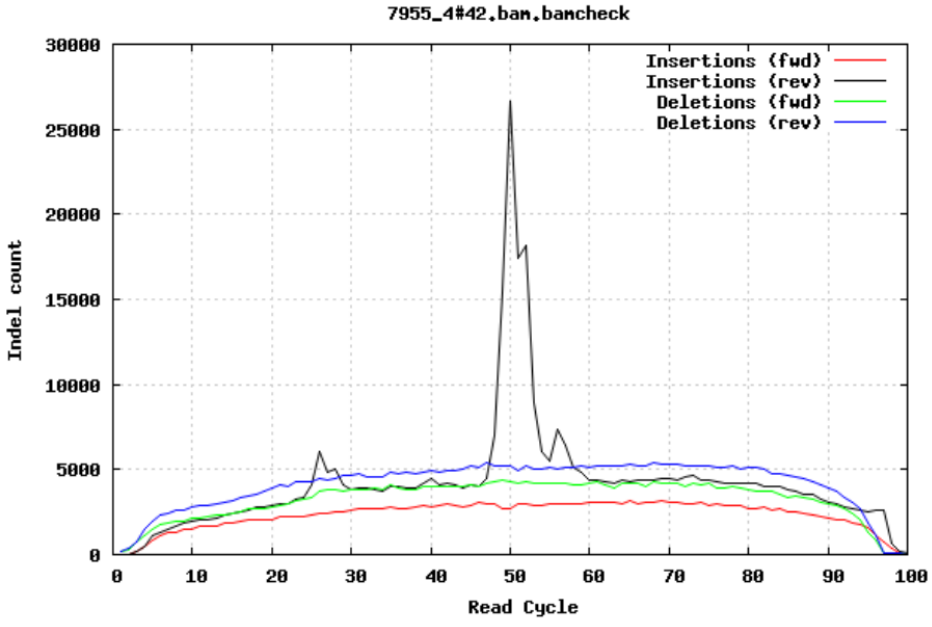


Insertions / Deletions per cycle

False indels

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

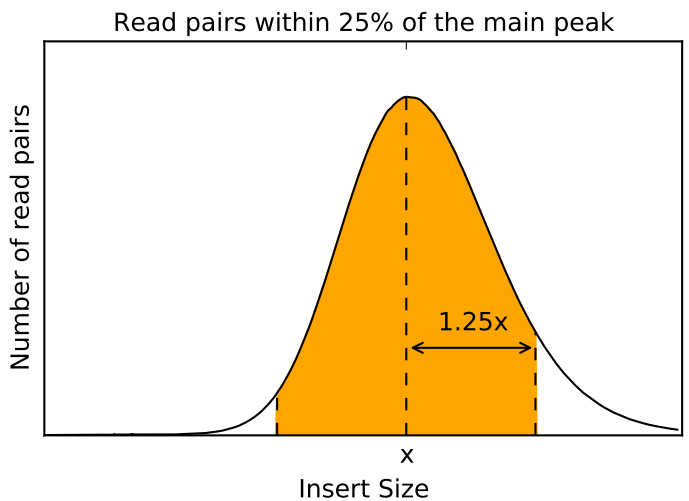




Auto QC tests

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
Maximimum indels per cycle, factor above median	8
Minimum number of reads within 25% of the main peak	80%



Detecting contamination and sample swaps

Detect sample mixture from population allele frequency

https://genome.sph.umich.edu/wiki/VerifyBamID

Check sample identity against a known set of variants

Cheat Sheet

File formats specifications

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

Index FASTA file

```
samtools faidx ref.fa
```

View a SAM/BAM/CRAM or a slice of it

```
samtools view file.bam | less
samtools view file.bam chr1:300000-310000 | less
```

Generate and plot stats

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

Index VCF/BCF

```
bcftools index file.vcf
```

View VCF/BCF or a slice of it

```
bcftools view file.vcf | less
bcftools view -r chr1:300000-310000 file.vcf | less
```

Generate and plot stats

```
bcftools stats -s - file.vcf > file.txt
plot-vcfstats -p plots/ file.txt
```