Module 3: File formats, QC and Data Processing

Presented by:
Carla Daniela Robles Espinoza

Laboratorio Internacional de Investigación sobre el Genoma Humano
Universidad Nacional Autónoma de México

@daniela_oaks
drobles@liigh.unam.mx

Based on slides by:
Petr Danecek

Next Generation Sequencing Bioinformatics Course
22-27 January 2023 - Santiago - Chile

FACULTAD DE
CIENCIAS BIOLOGICAS
PONTIFICIA
UNIVERSIDAD
CATÓLICA DE CHILE

WELLCOME GENOME CAMPUS
CONNECTING SCIENCE
ADVANCED COURSES+ SCIENTIFIC CONFERENCES

Data Formats

FASTQ

- Unaligned read sequences with base qualities

SAM/BAM

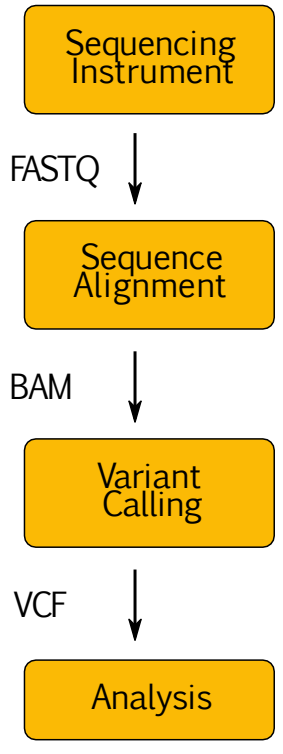
- Unaligned or aligned reads
- Text and binary formats

CRAM

- Better compression than BAM

VCF/BCF

- 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
```

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
```

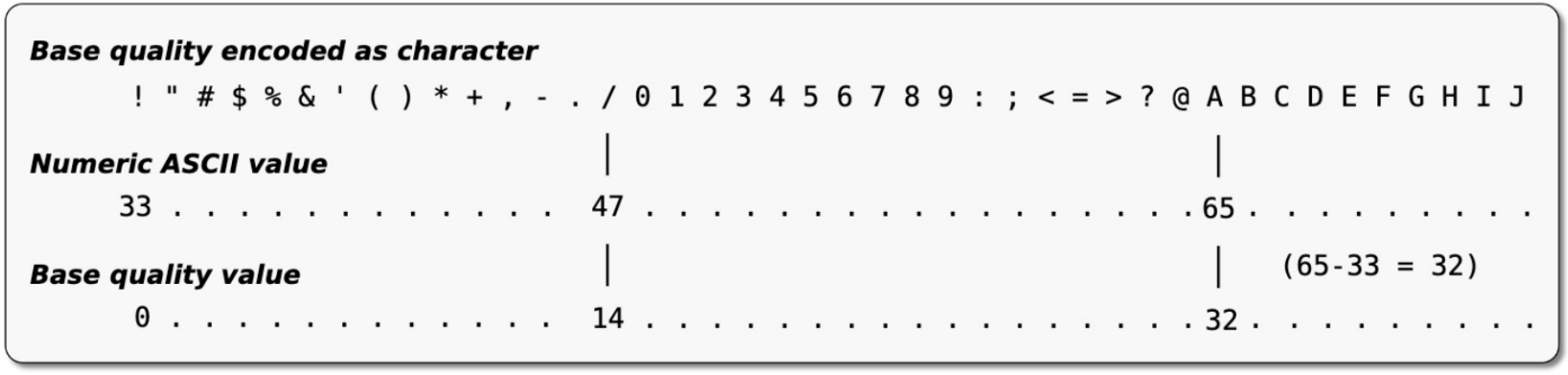
2003	NCBI Build 34	hg16
2004	NCBI Build 35	hg17
2006	NCBI Build 36.1	hg18
2009	GRCh37	hg19
2013	GRCh38	hg38

FASTQ

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



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
- Paired-end sequencing: two FASTQ files or one interleaved file
- 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: $P = 10^{-Q/10}$

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

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

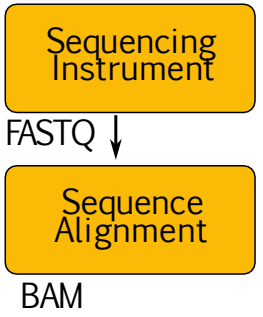
```
@ERR007731.739 IL16_2979:6:1:9:1684/1
CTTGACGACTTGAAAAATGACGAAATCACTAAAAAACGTGAAAAATGAGAAATG
+
BBCBCBBBBBBBABBABBBBBBBABBBBBBBBBBBBBBABAAAABBBBB=@>B
@ERR007731.740 IL16_2979:6:1:9:1419/1
AAAAAAAAAGATGTCATCAGCACATCAGAAAAGAAGGCAACTTTAAAACTTTTC
+
BBABBBABABAABABABBABBBAAA>@B@BBAA@4AAA>.>BAA@779:AAA@A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

- 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;'
```

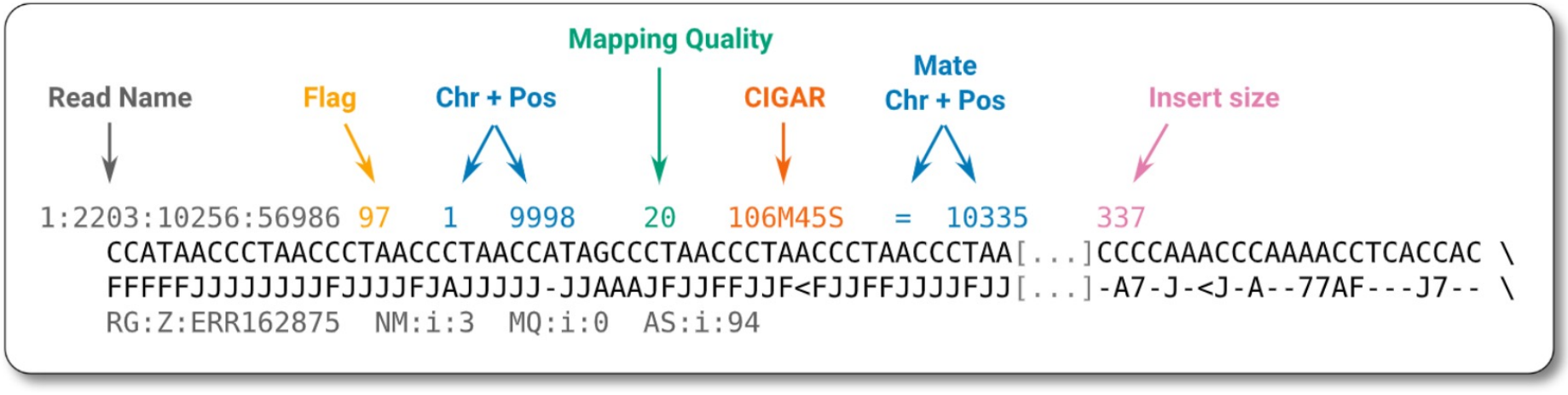
- Beware: multiple quality scores were in use!
  - Sanger, Solexa, Illumina 1.3+

SAM / BAM

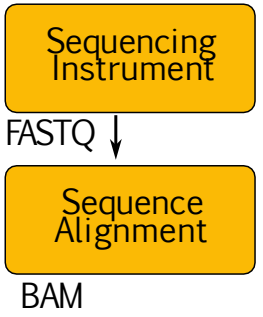


SAM (Sequence Alignment/Map) format

- 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

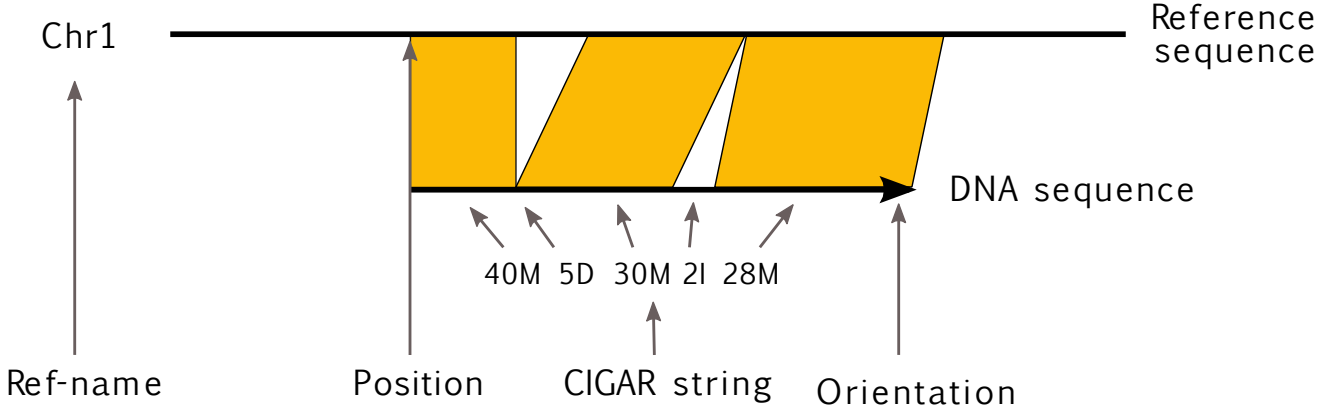


SAM / BAM



SAM (Sequence Alignment/Map) format

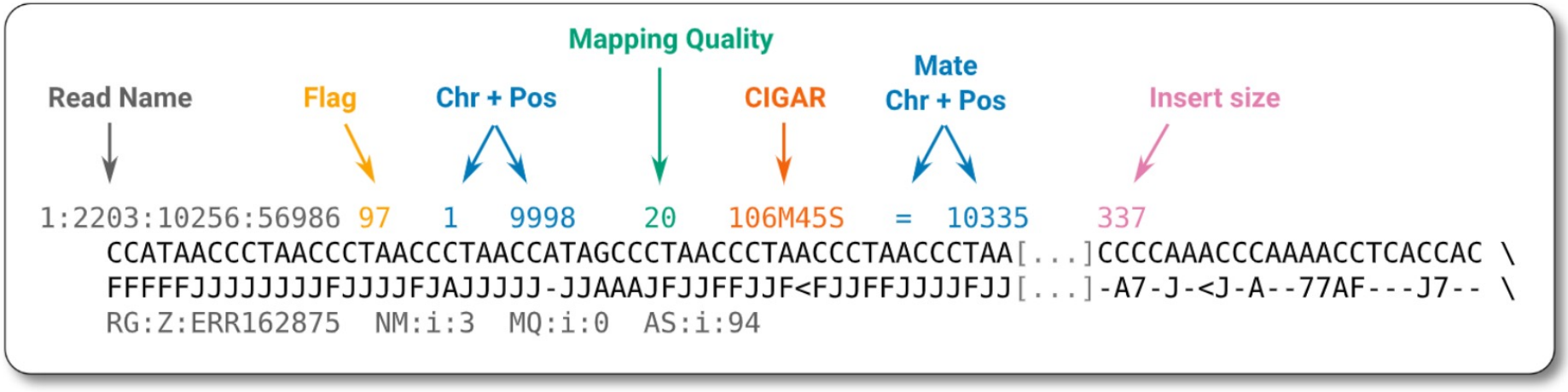
- Unified format for storing read alignments to a reference genome
- Developed by the 1000 Genomes Project group (2009)
- One record (a single DNA fragment alignment) per line describing alignment between fragment and reference
- 11 fixed columns + optional key:type:value tuples



Note that BAM can contain

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

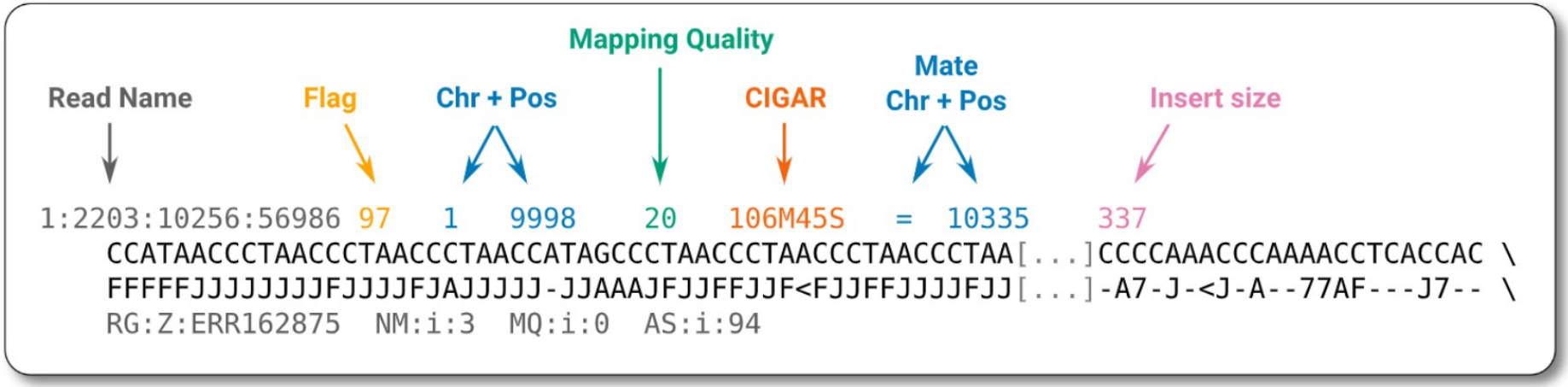
SAM



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

Flags

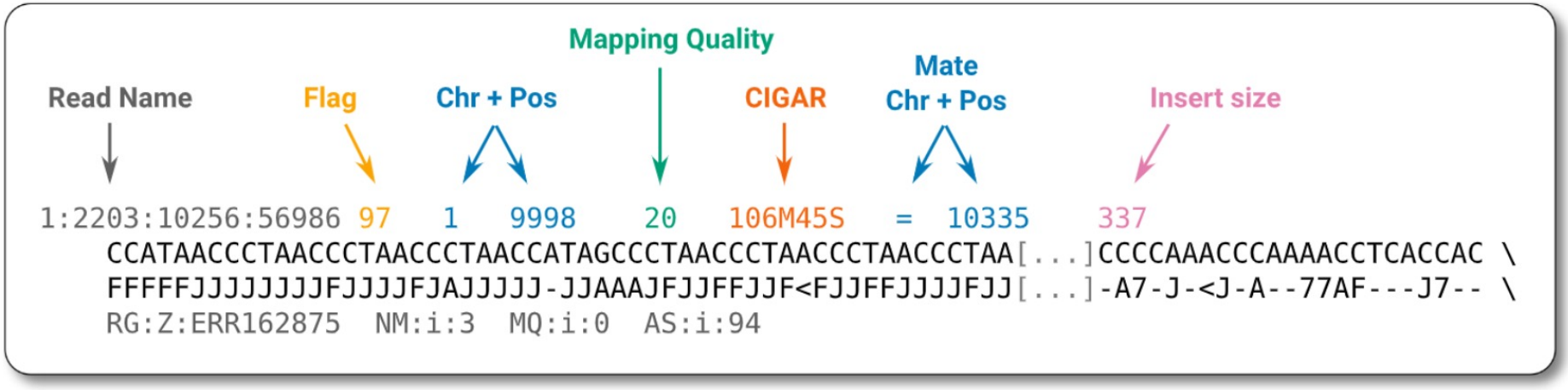


Hex	Dec	Flag	Description
0x1	1	PAIRED	paired-end (or multiple-segment) sequencing technology
0x2	2	PROPER_PAIR	each segment properly aligned according to the aligner
0x4	4	UNMAP	segment unmapped
0x8	8	MUNMAP	next segment in the template unmapped
0x10	16	REVERSE	SEQ is reverse complemented
0x20	32	MREVERSE	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

CIGAR string



Compact representation of sequence alignment

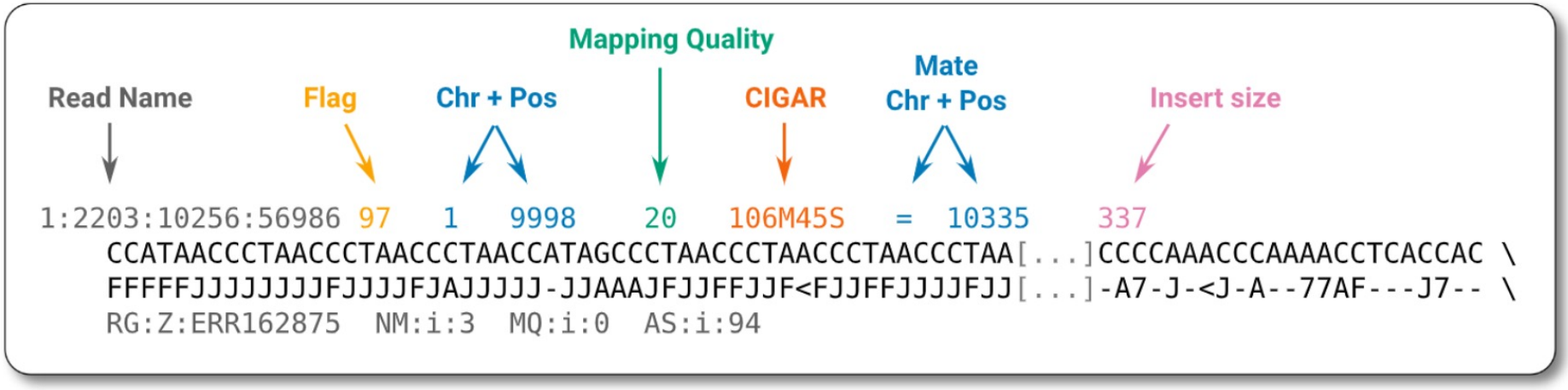
M	alignment match or mismatch
=	sequence match
X	sequence mismatch
I	insertion to the reference
D	deletion from the reference
S	soft clipping (clipped sequences present in SEQ)
H	hard clipping (clipped sequences NOT present in SEQ)
N	skipped region from the reference
P	padding (silent deletion from padded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

Insert size



Insert size

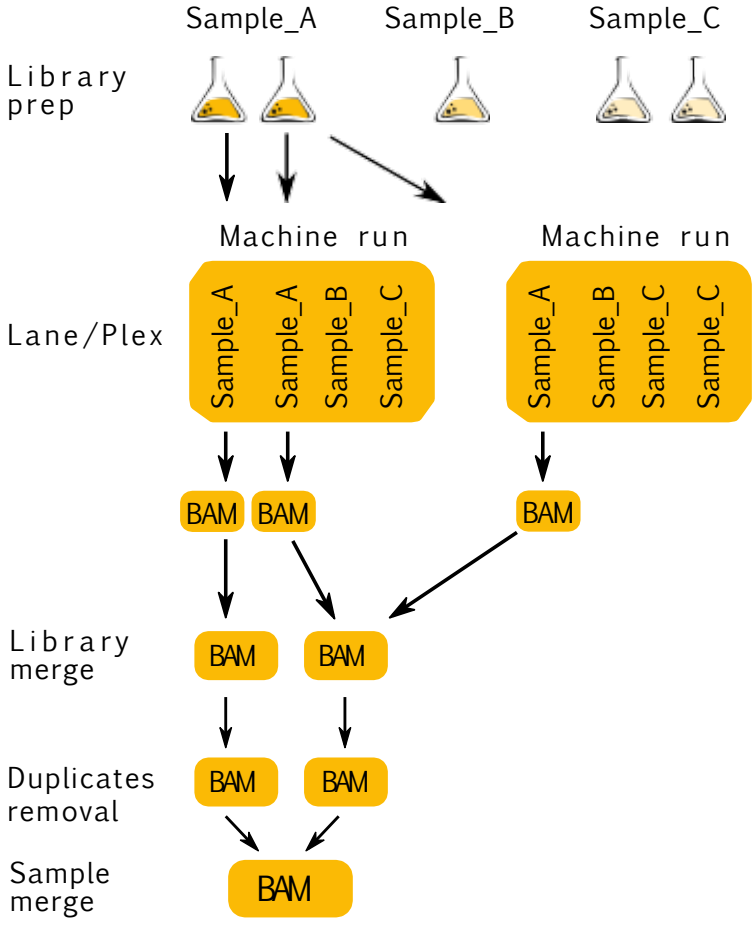
length of the DNA fragment sequenced from both ends by paired-end sequencing:

Optional tags

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

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

- 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:

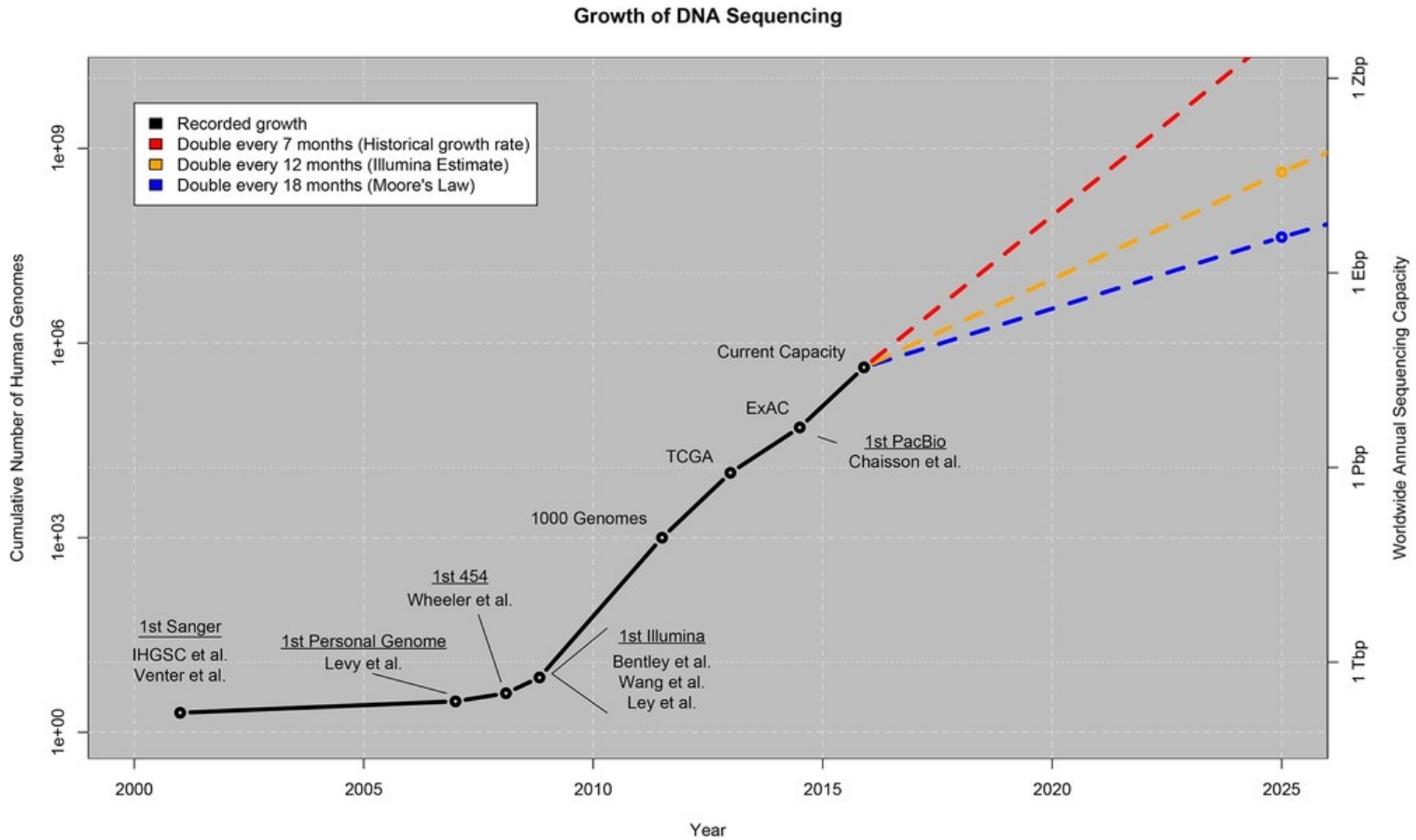
- Samtools (Wellcome Sanger Institute)
- Picard tools (Broad Institute)
- Visualisation: IGV, Ensembl, UCSC

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

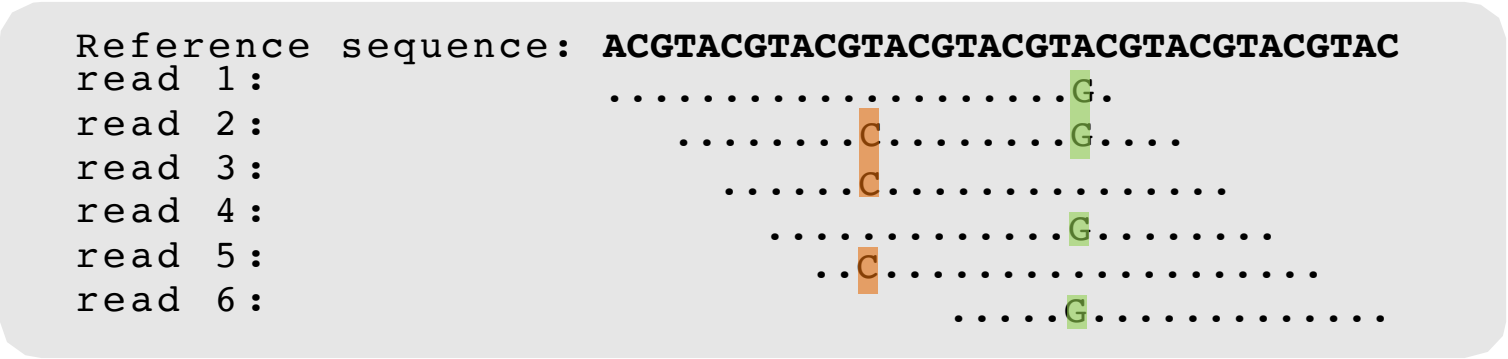
CRAM

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

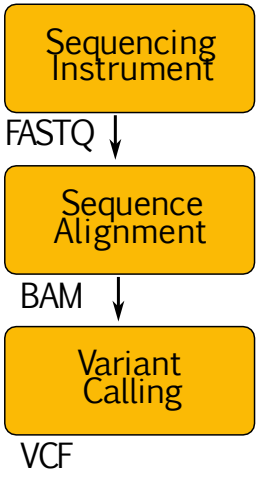
- Support for CRAM added to Samtools/HTSlib in 2014
- Added in Picard/GATK in 2015



VCF: Variant Call Format

File format for storing generic variation data

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



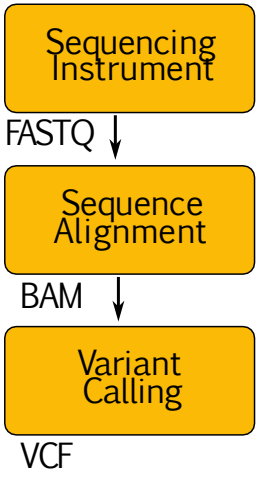

```
...
##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele frequency in population">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths (ref,alt,..)">
...
#CHROM 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
```

- 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



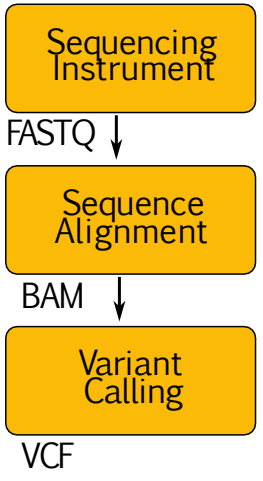

```
...
##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele frequency in population">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths (ref,alt,..)">
...
#CHROM 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
```

- Chromosome and position
- Variant ID
- Reference and alternative alleles
- 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)

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)

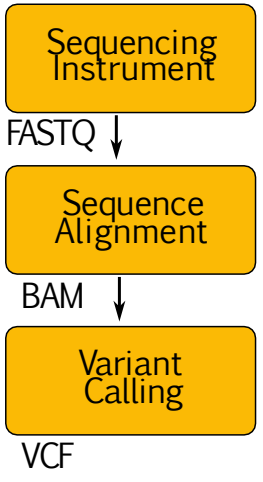
- Homozygous reference (*e.g.*, A/A if the reference allele is A)
- Homozygous alternative (*e.g.*, G/G if the reference allele is A)
- Heterozygous (*e.g.*, C/T)

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

```
...
##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 REF: ACGTACGT ALT: ACTAACGT	POS 3	REF GT	ALT TA
Deletion	ACGTACGT AC--ACGT	2	CGT	C
Insertion	AC--ACGT ACGTACGT	2	C	CGT
Structural variation		2 2	C C	<DEL> <DUP>

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*, **

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
```

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

Global Alliance for Genomics & Health

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

Petr Danecek recommends running:

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

Questions we are interested in:

- Do I have enough coverage with my mapped reads?
- Was the library creation process efficient and problem-free?
- Did the sequencing process créate artifacts?

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

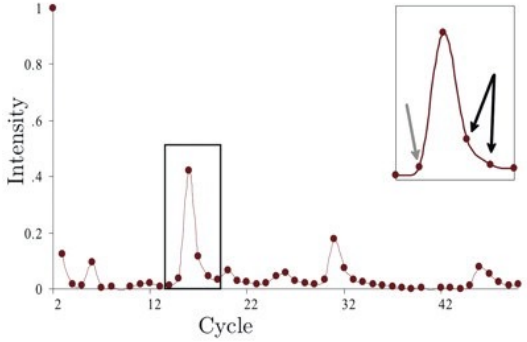
- Be careful to distinguish between the total sequencing yield and on-target bases

Useful coverage:

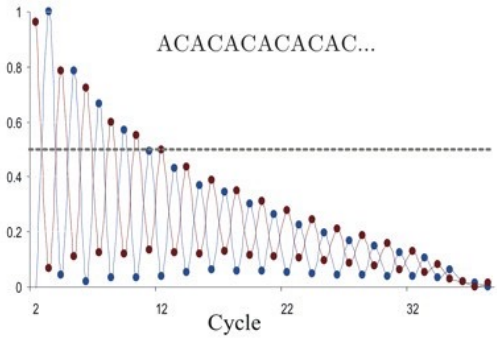
- 15x OK for common germline variants
- 30x OK for most things
- 100-200x for low VAF variants in tumours

Base calling errors

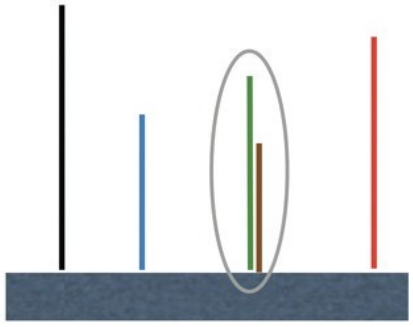
Phasing noise ϕ



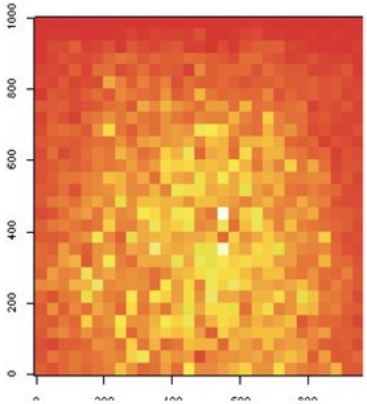
Signal Decay δ



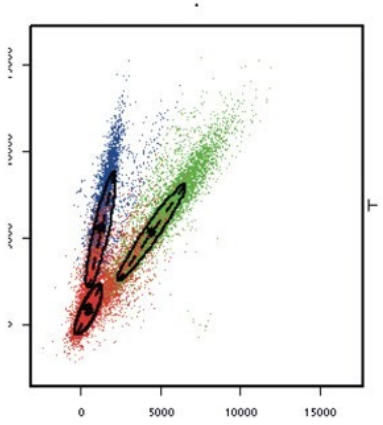
Mixed Cluster μ



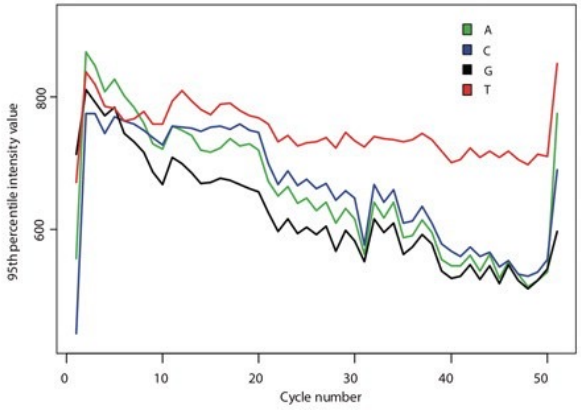
Boundary effects ω



Cross-talk Σ



T fluophore accumulation τ



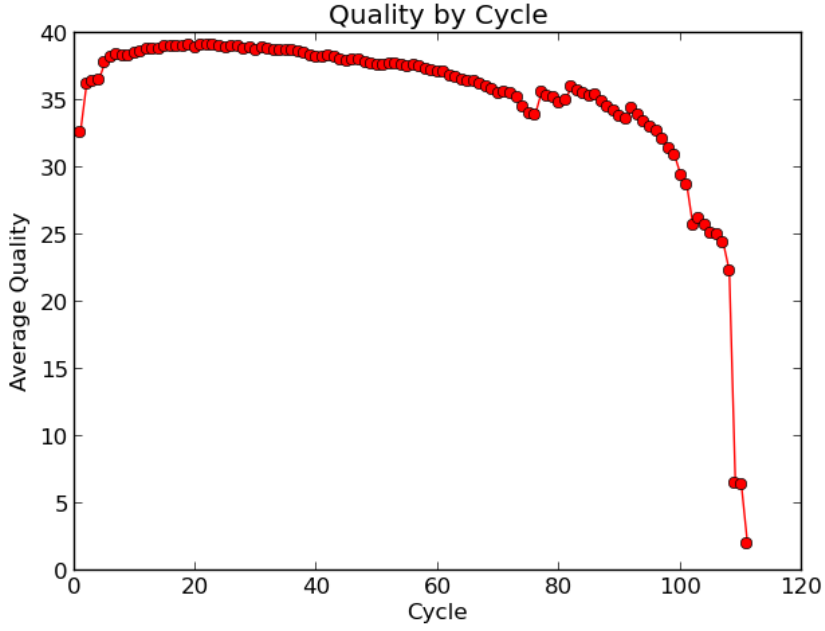
Base-calling for next-generation sequencing platforms, doi: 10.1093/bib/bbq077

Base quality

Sequencing by synthesis: dephasing

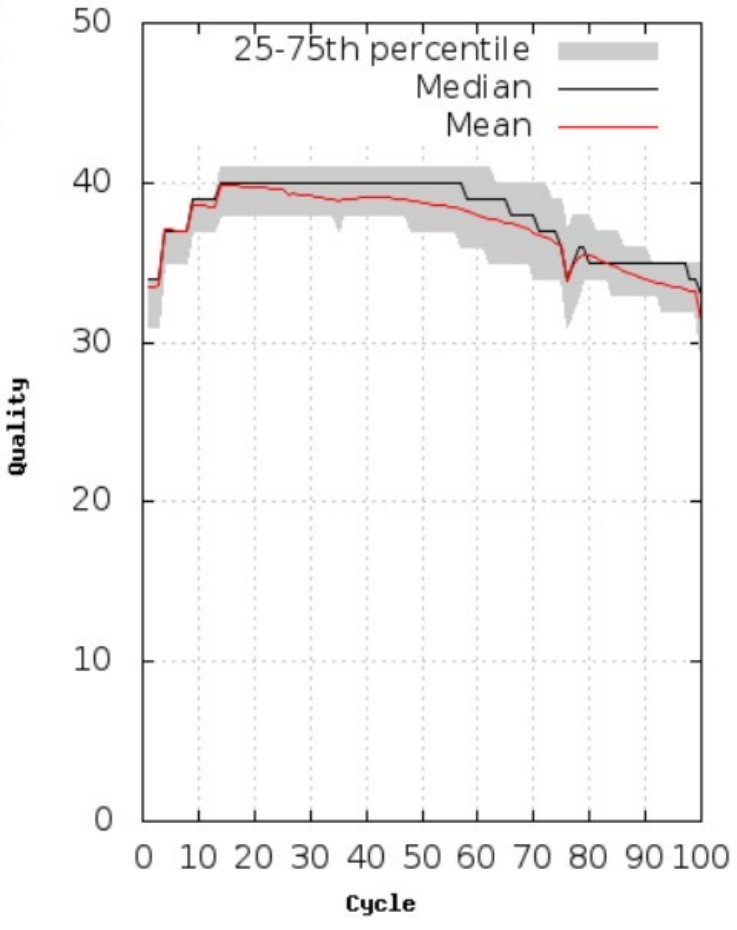
- growing sequences in a cluster gradually desynchronize
- error rate increases with read length

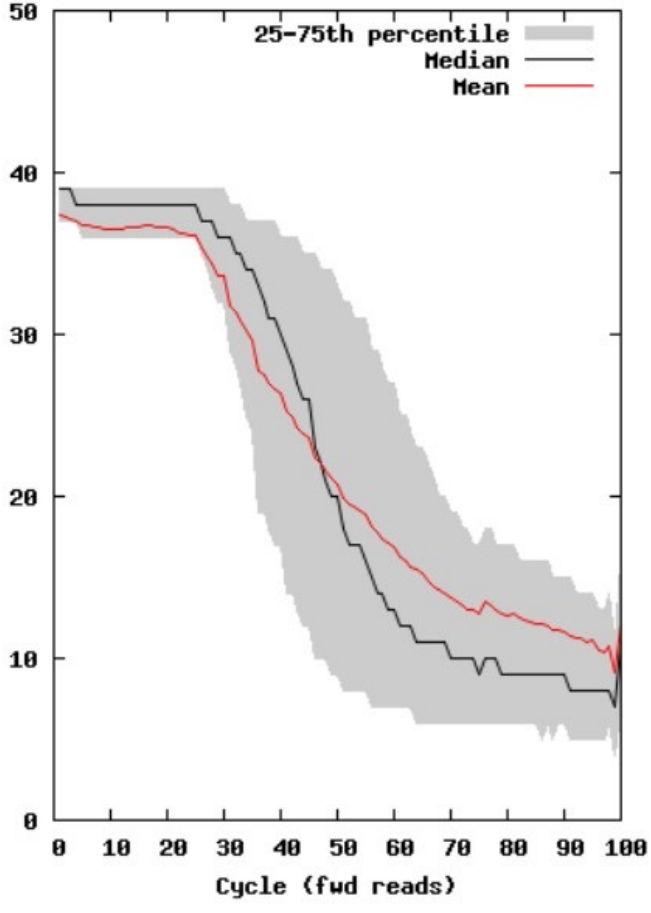
Calculate the average quality at each position across all reads



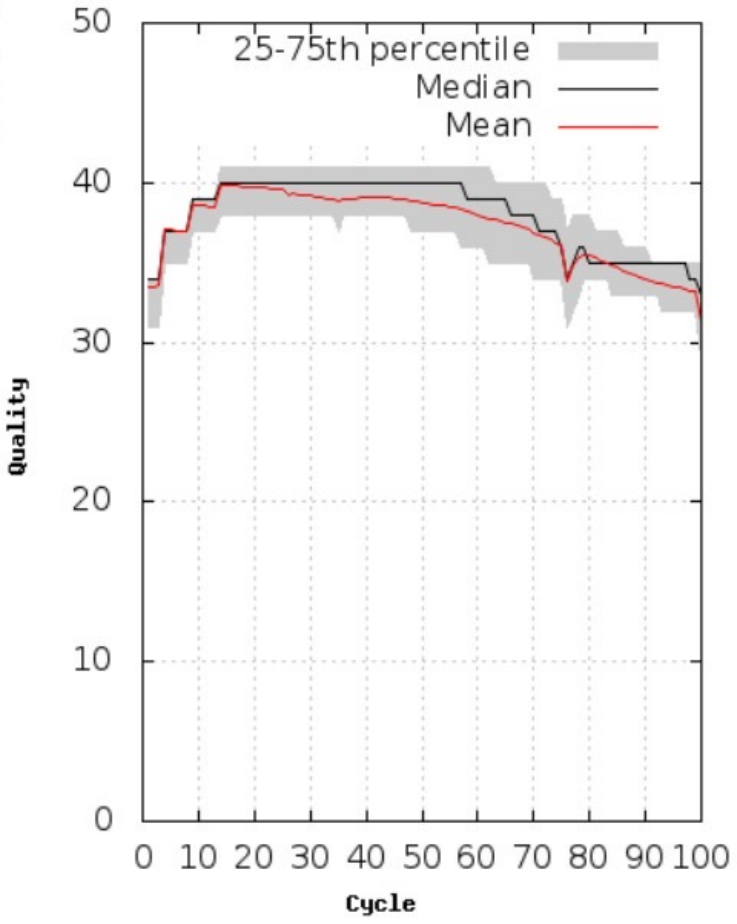
Quality	Probability of error	Call Accuracy
10 (Q10)	1 in 10	90%
20 (Q20)	1 in 100	99%
30 (Q30)	1 in 1000	99.9%
40 (Q40)	1 in 10000	99.99%

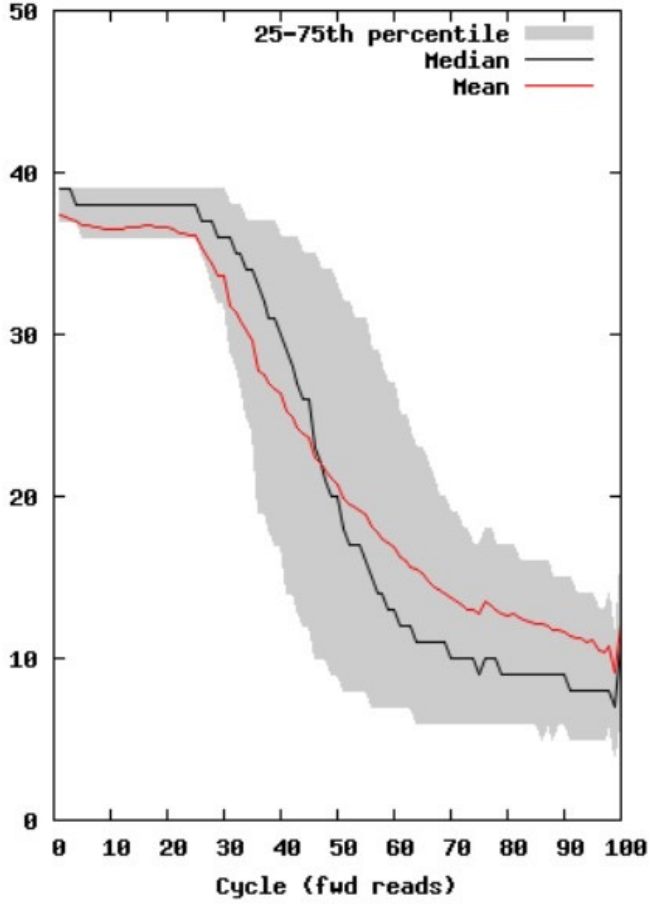
Base quality

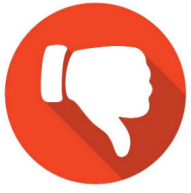




Base quality







Library prep biases: PCR duplicates

Experiments start with small amounts of DNA

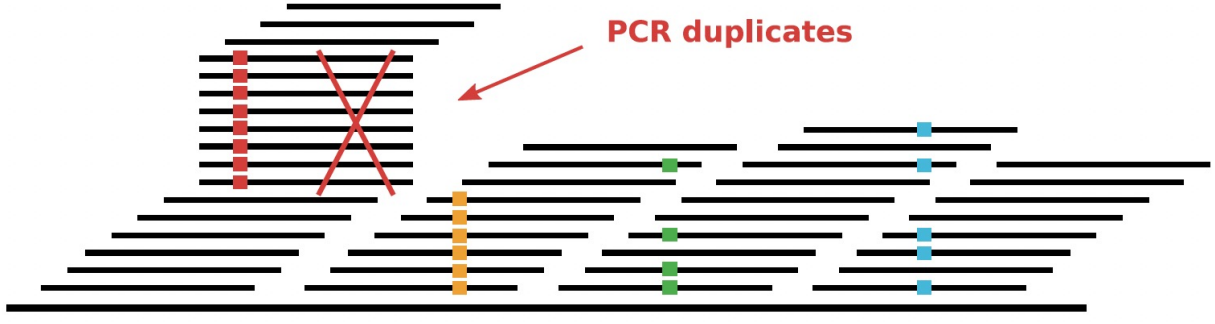
- A PCR amplification step is necessary for Illumina sequencing: one molecule -> many identical molecules

Problem:

- Additional PCR copy molecules are not informative

Solution:

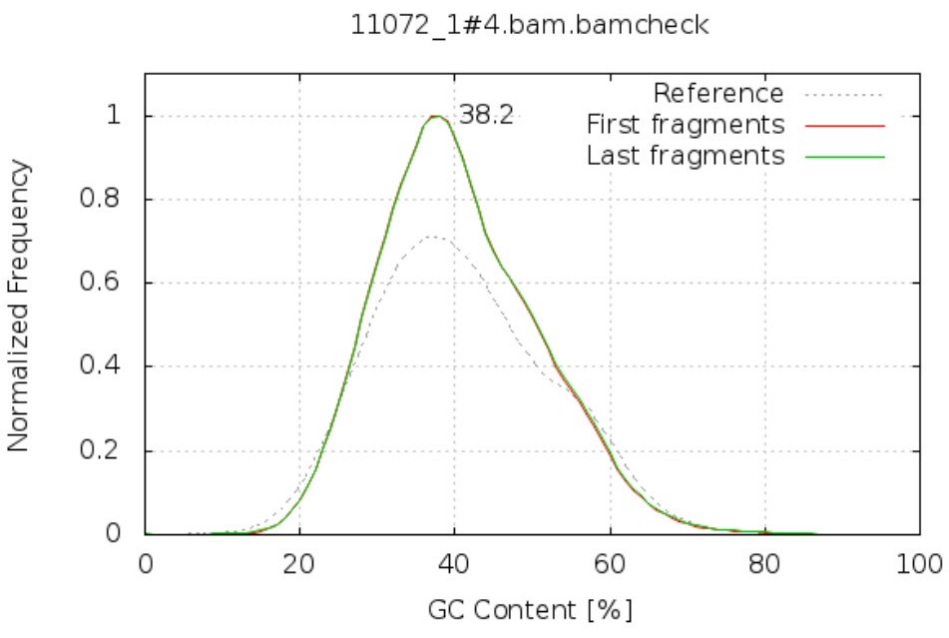
- Infer and mark PCR duplicates, discount in later analysis
    - Mark if reads and their mates start at the same position
- Use Picard MarkDuplicates or samtools markdup
- Typical duplication rates: Exomes 15-20%, Genomes < 5%

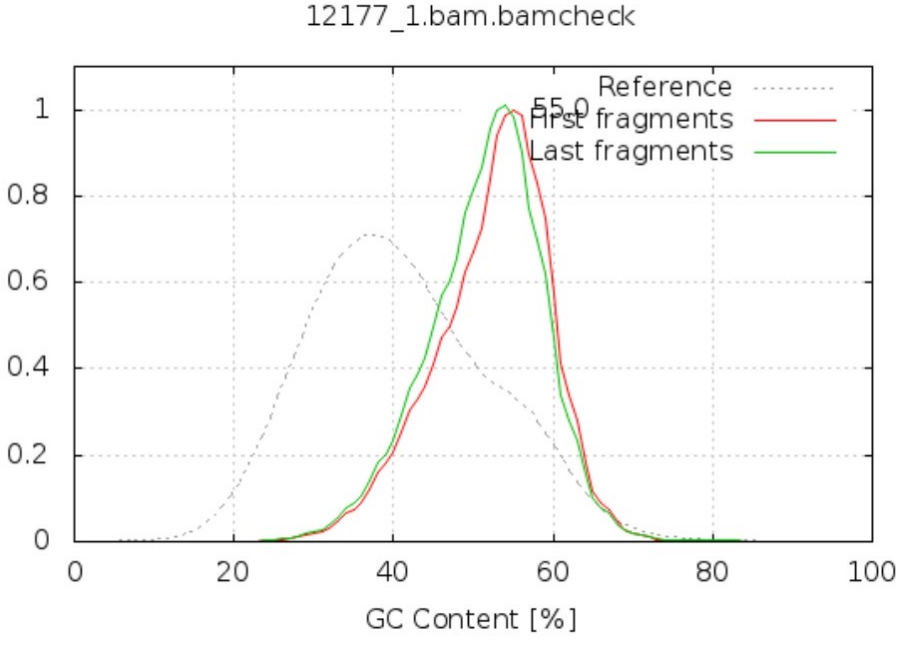


GC bias

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

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

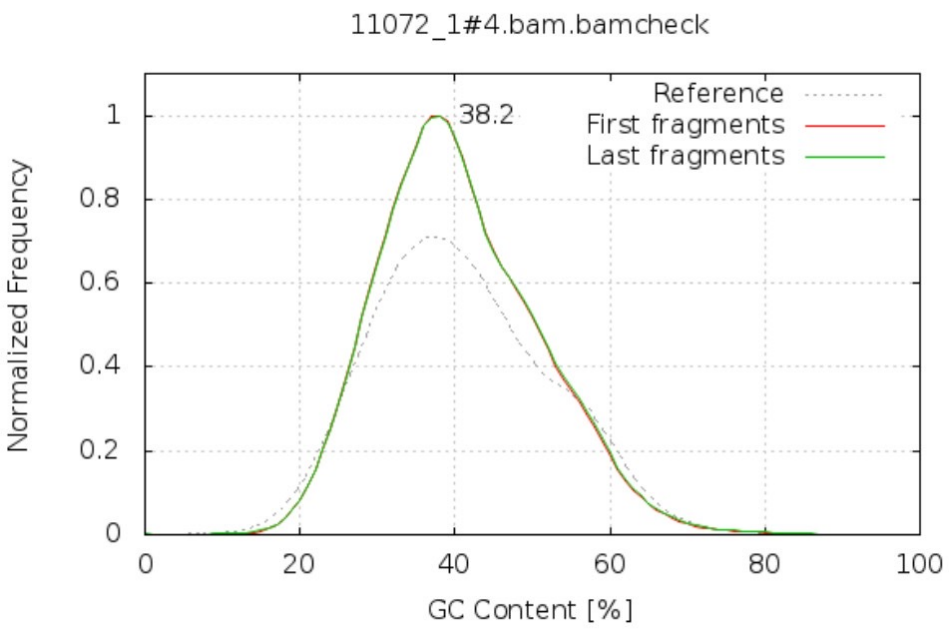


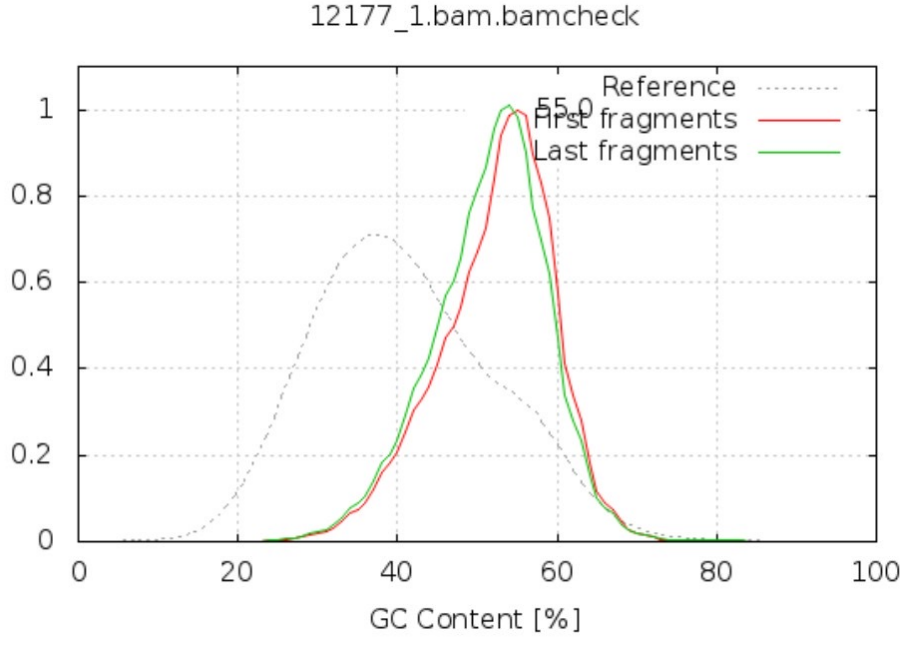


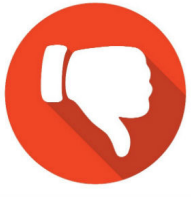
GC bias

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

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

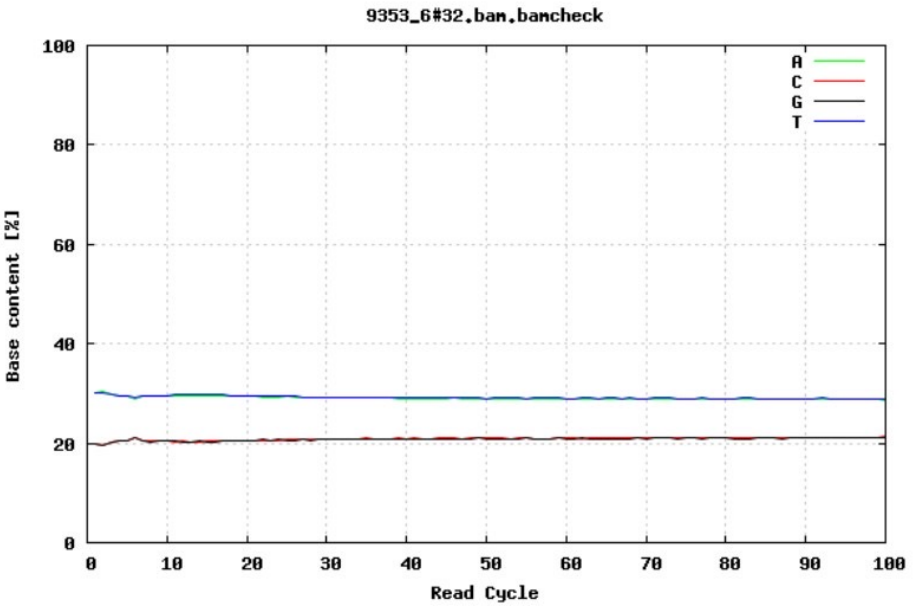


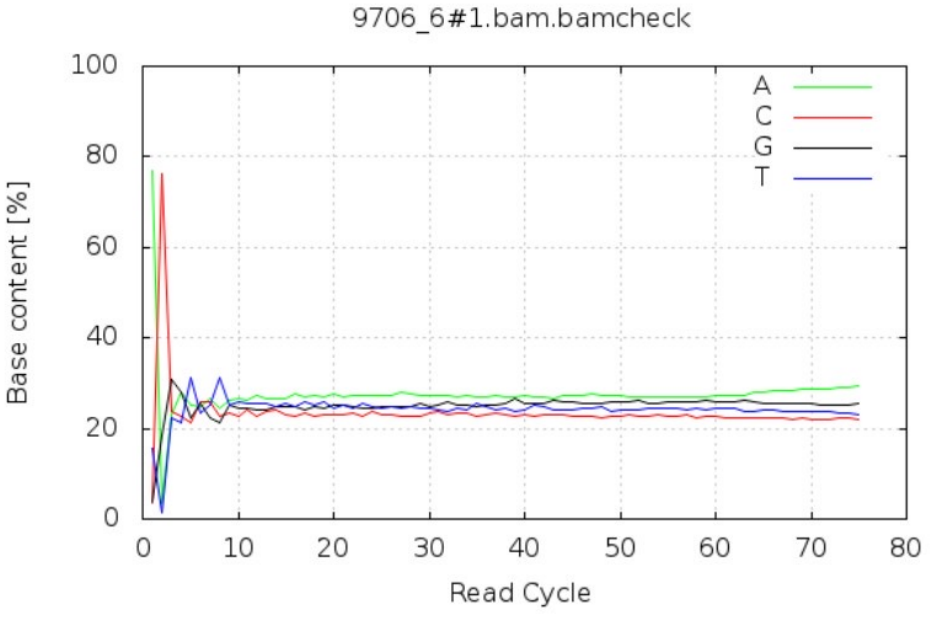




GC content by cycle

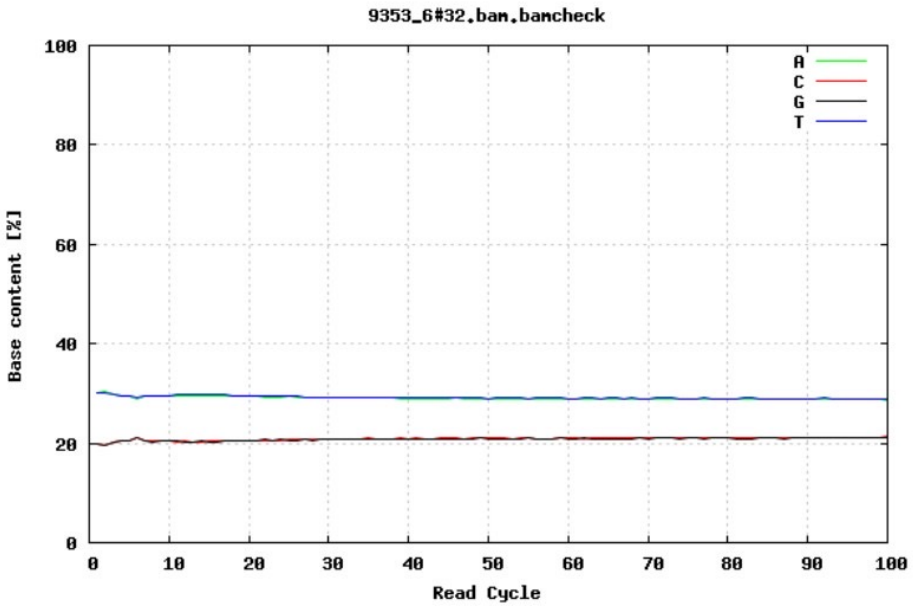
Was the adapter sequence trimmed?

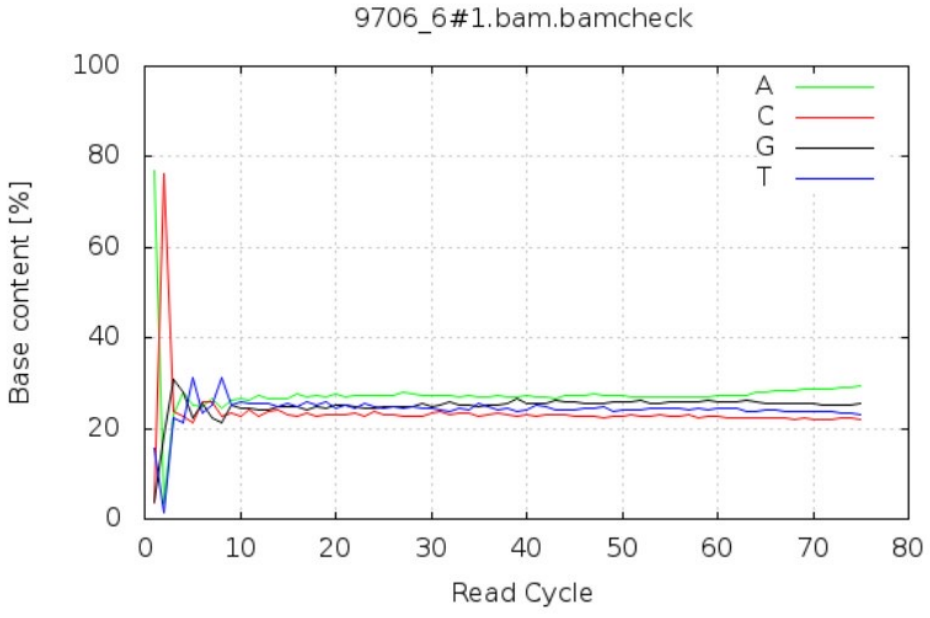


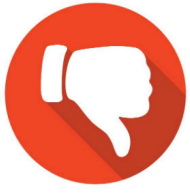


GC content by cycle

Was the adapter sequence trimmed?

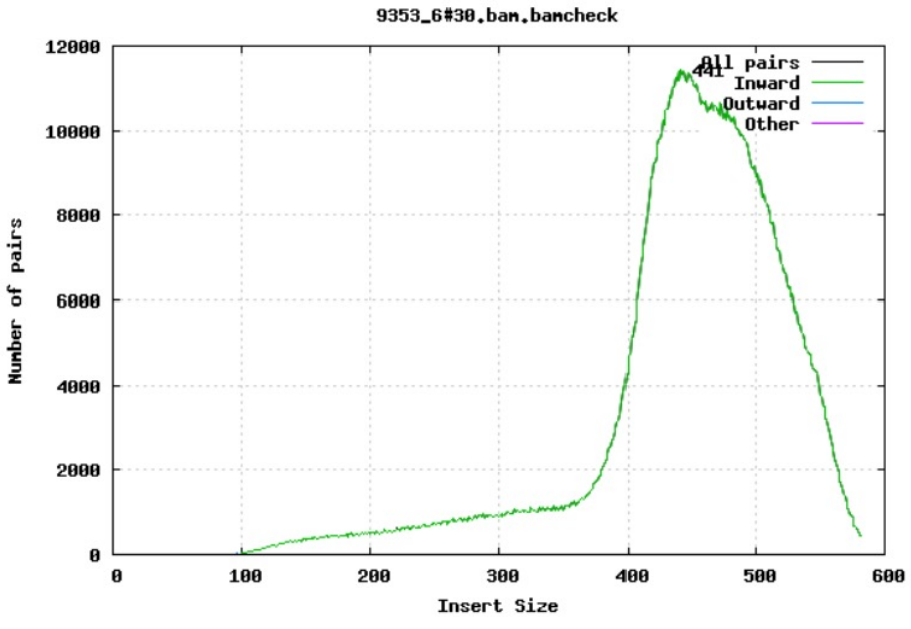


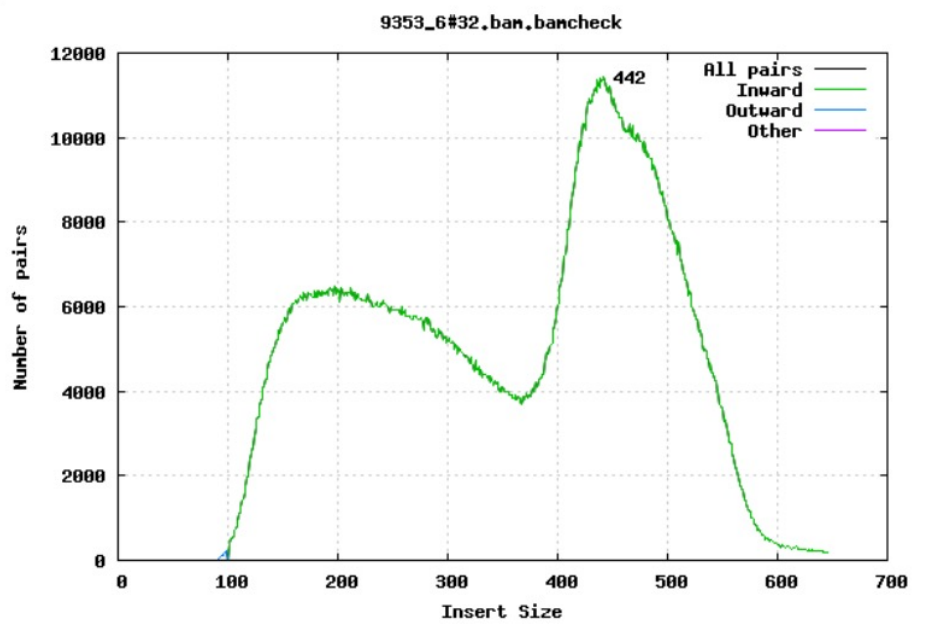




Fragment size

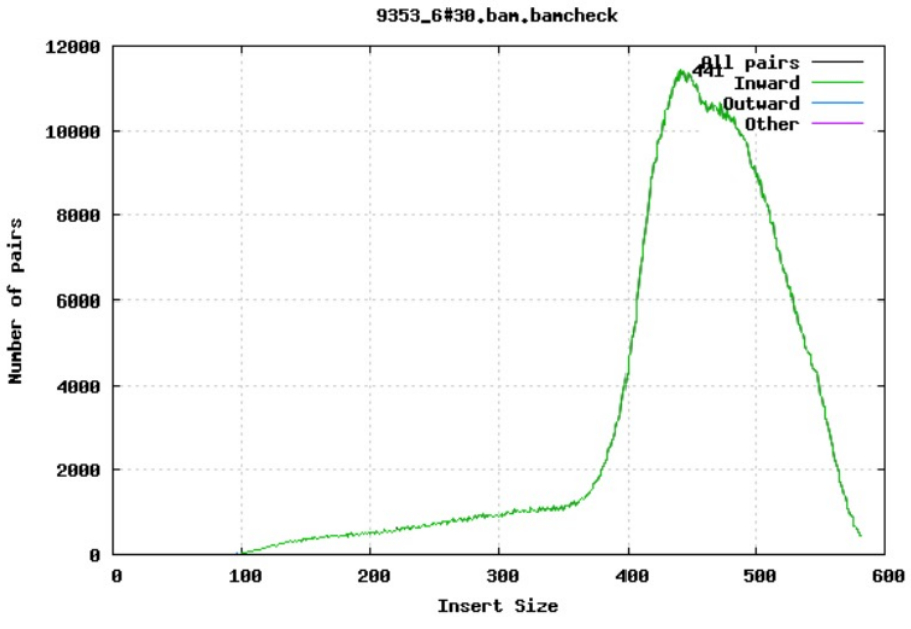
Paired-end sequencing: the size of DNA fragments matters

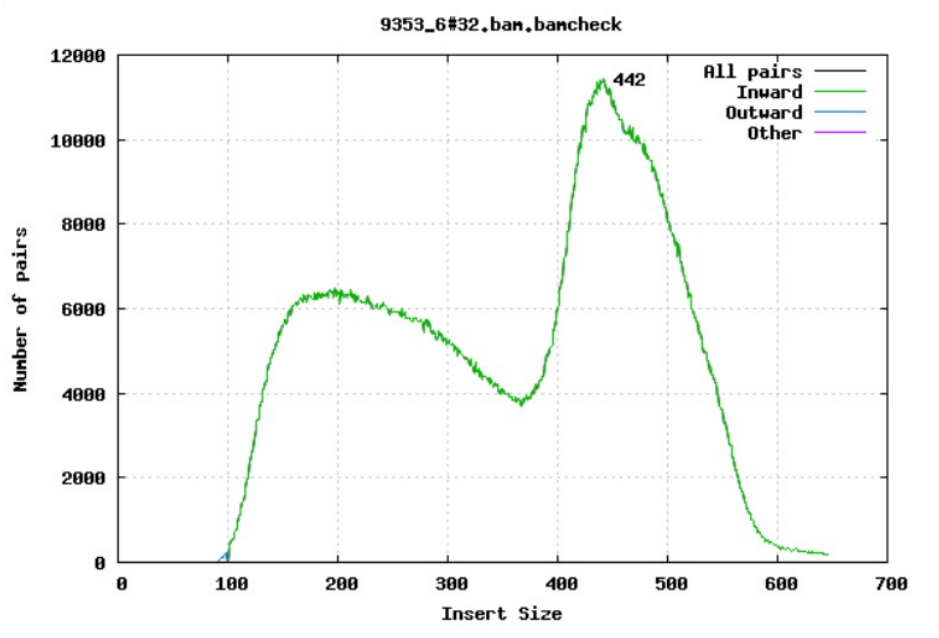


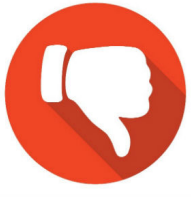


Fragment size

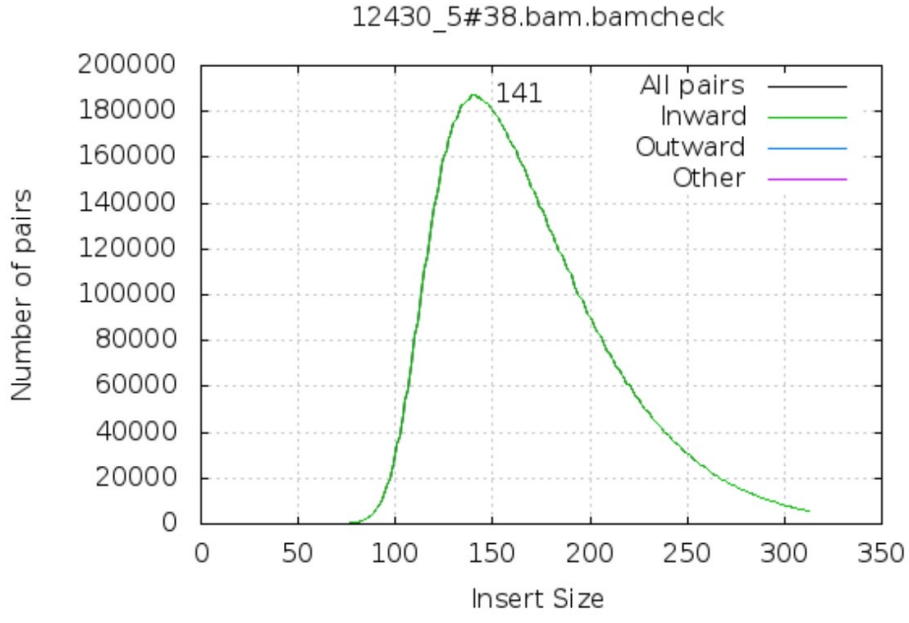
Paired-end sequencing: the size of DNA fragments matters





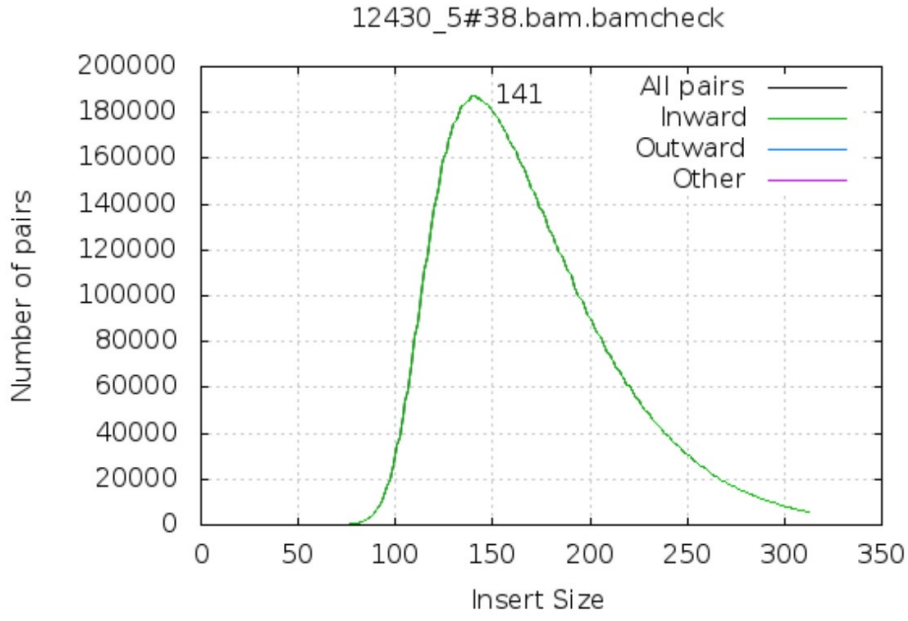


Quiz



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

Quiz



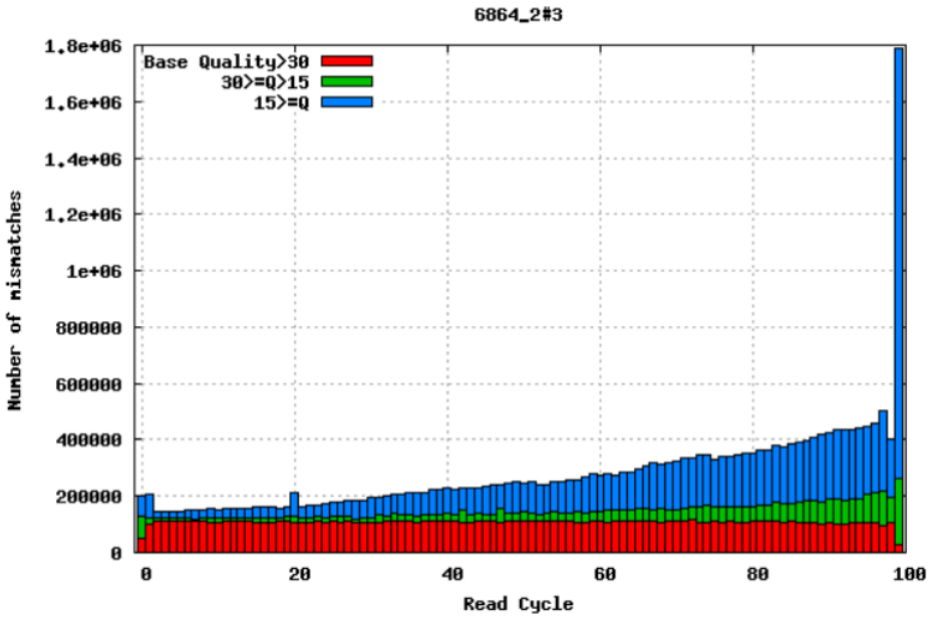
This is 100bp paired-end sequencing. Can you spot any problems??

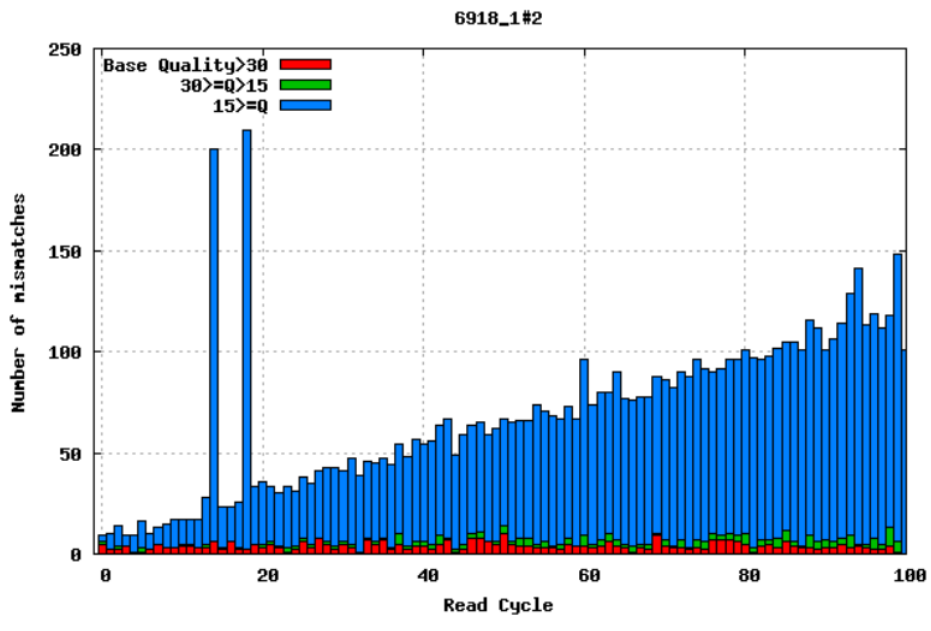
The insert size should be at least 200bp for the mates not to overlap.

Mismatches per cycle

Mismatches in aligned reads (requires reference sequence)

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

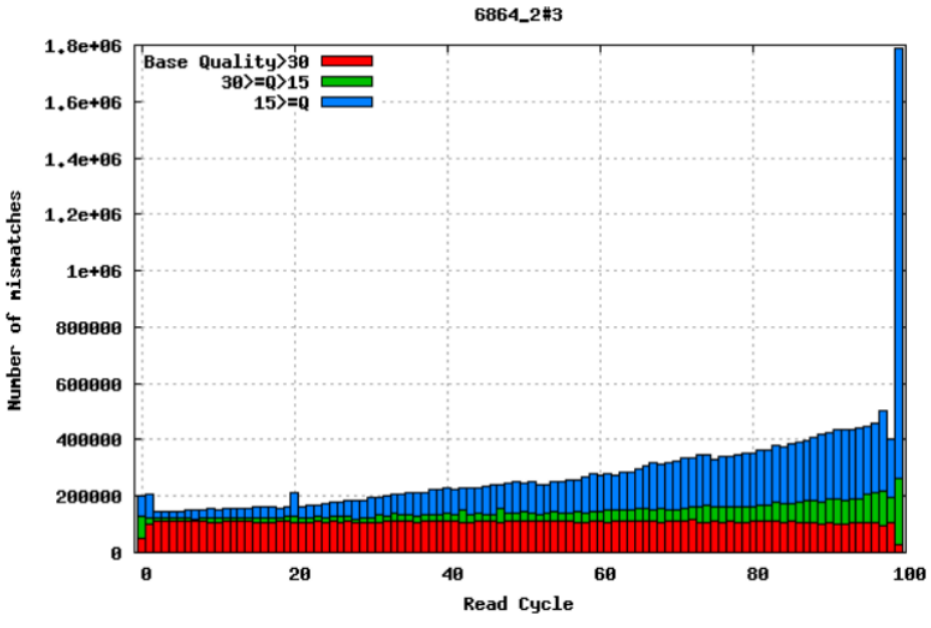


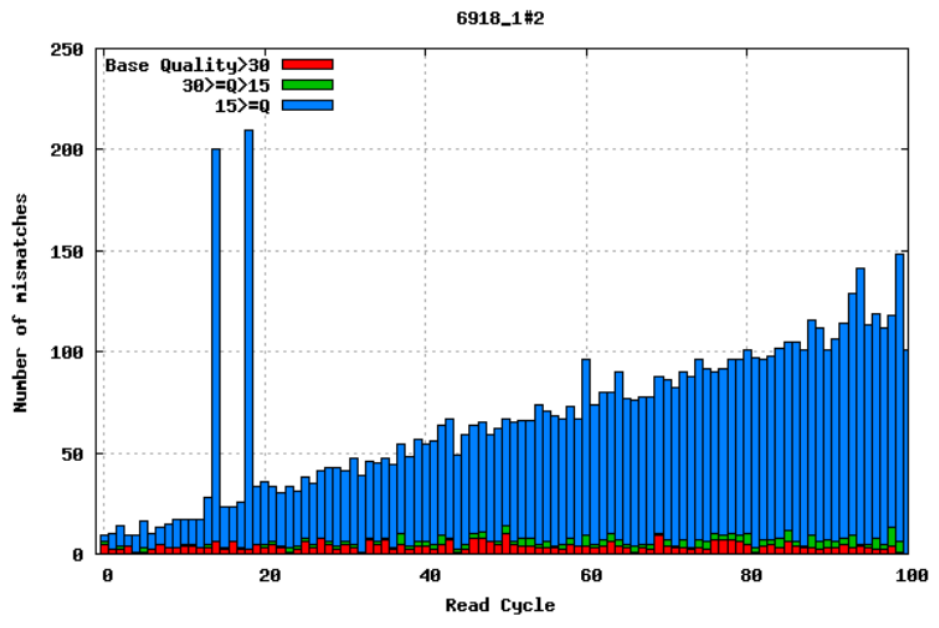


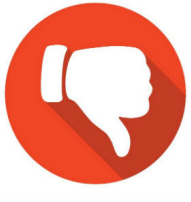
Mismatches per cycle

Mismatches in aligned reads (requires reference sequence)

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



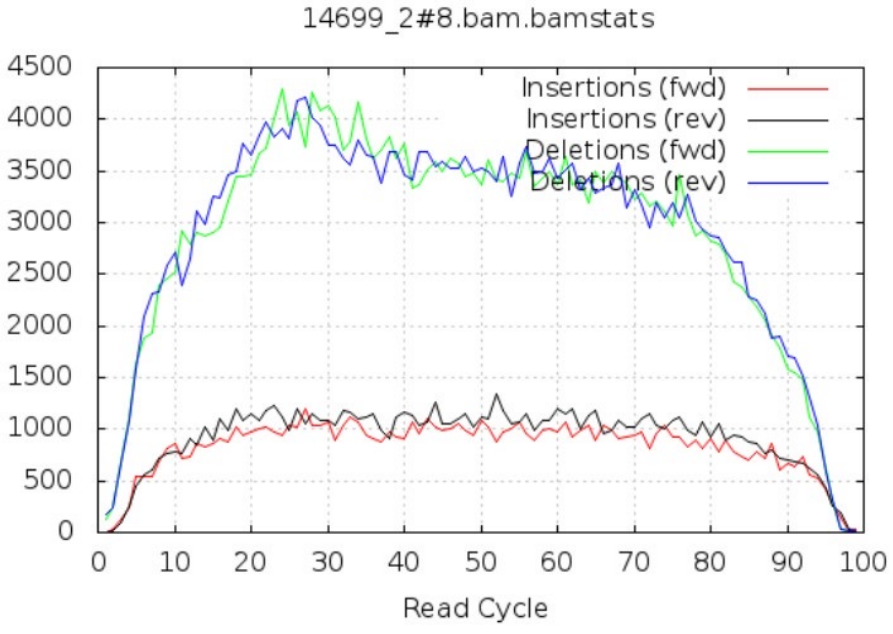


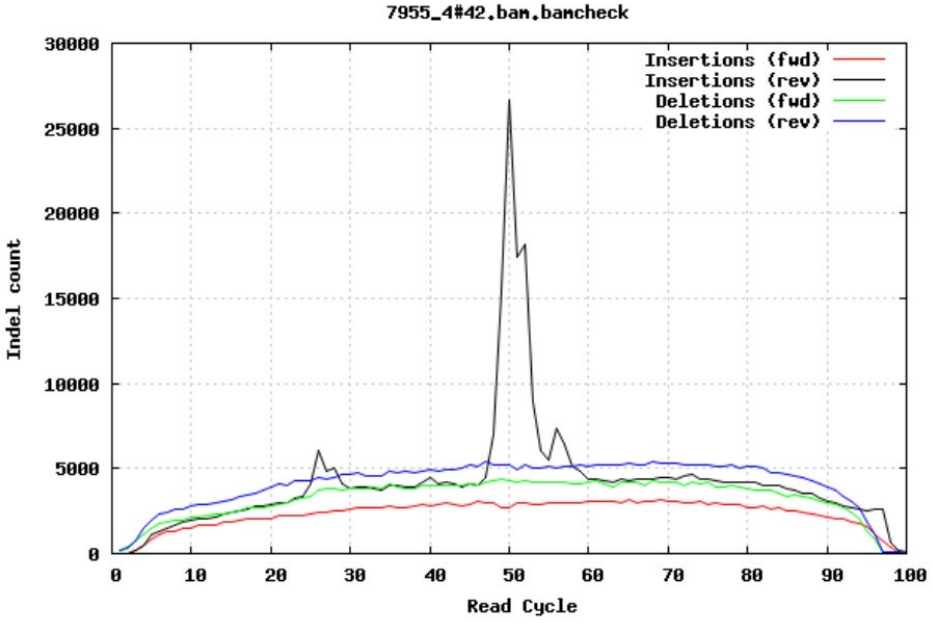


Insertions / Deletions per cycle

False indels

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

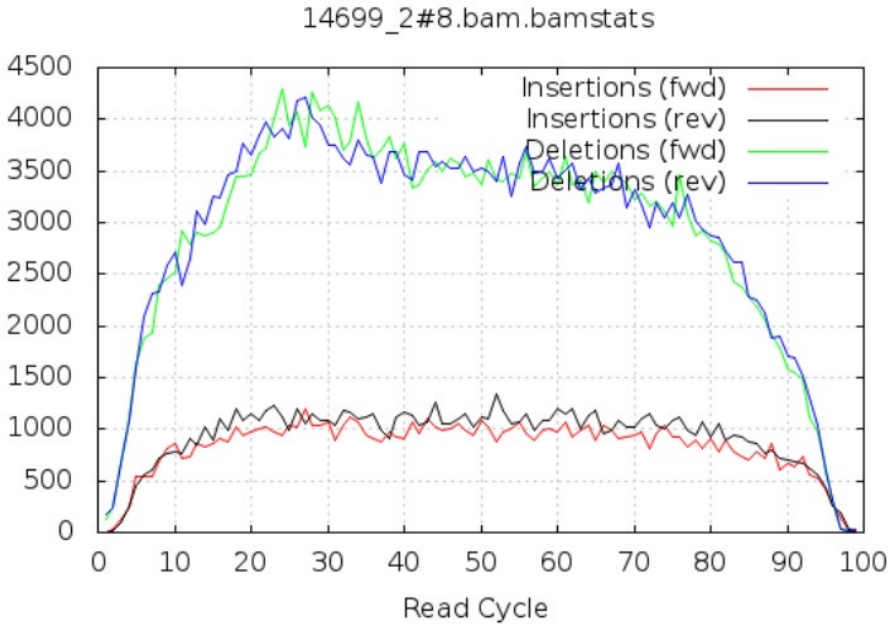


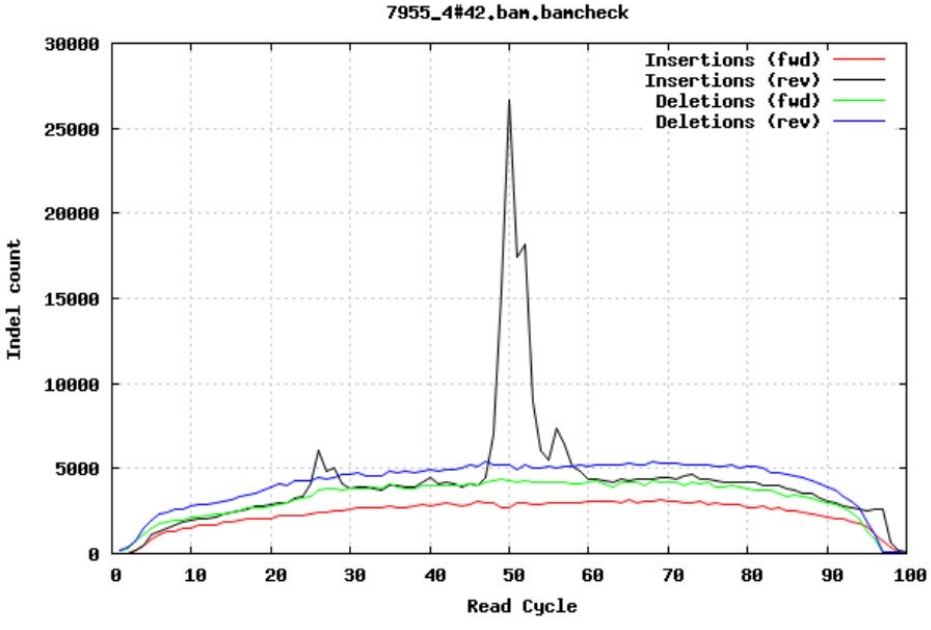


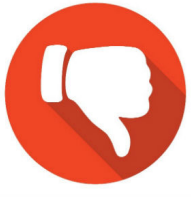
Insertions / Deletions per cycle

False indels

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



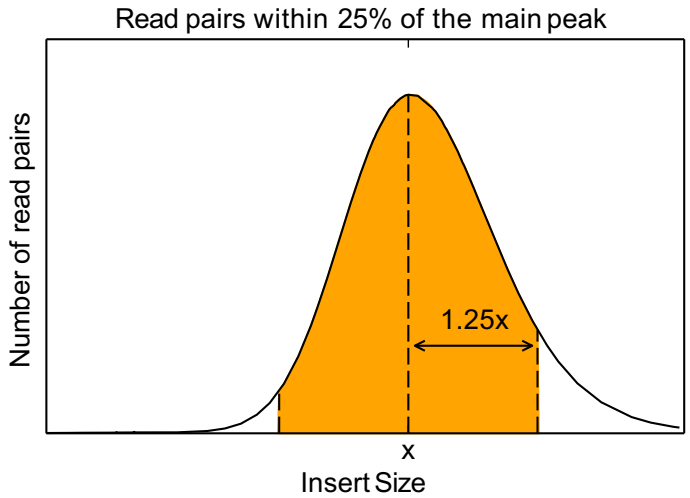




Auto QC tests

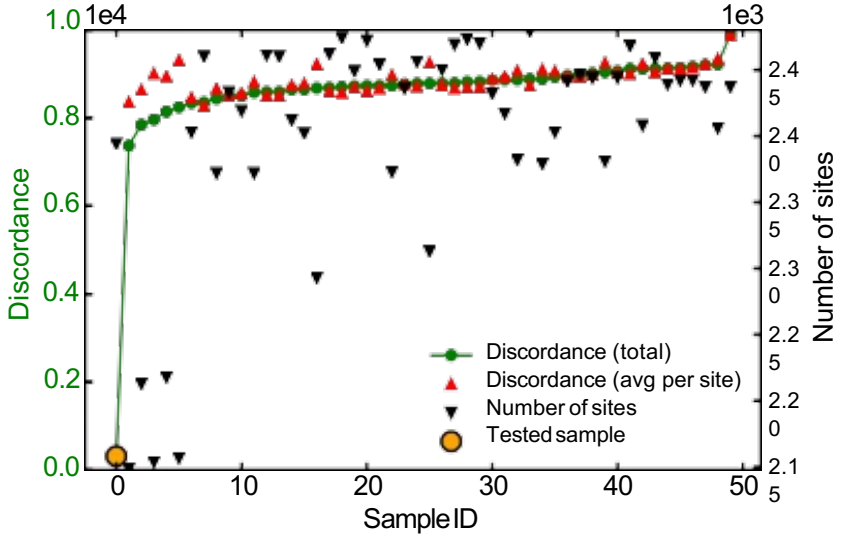
A suggestion for human data:

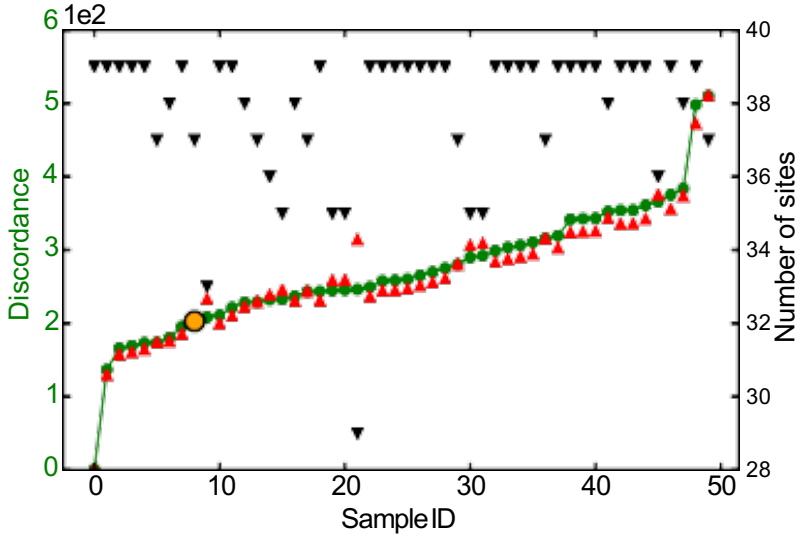
Minimum number of mapped bases	90%
Maximum error rate	0.02%
Maximum number of duplicate reads	5%
Minimum number of mapped reads which are properly paired	80%
Maximum number of duplicated bases due to overlapping read pairs	4%
Maximum in/del ratio	0.82
Minimum in/del ratio	0.68
Maximum indels per cycle,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

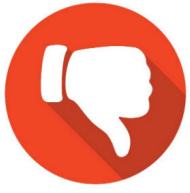


Detecting sample swaps

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 course_data/data_formats

```
cd course_data/data_formats/
```

- Open the exercises, which are in Github or in:

```
/home/manager/course_data/data_formats/practical/data_formats.pdf
```

- Follow the instructions!

Exercise time!

- Solutions (inside course_data/data_formats/practical):

```
course_data/data_formats/practical/ \
.data_formats_solutions.pdf
```