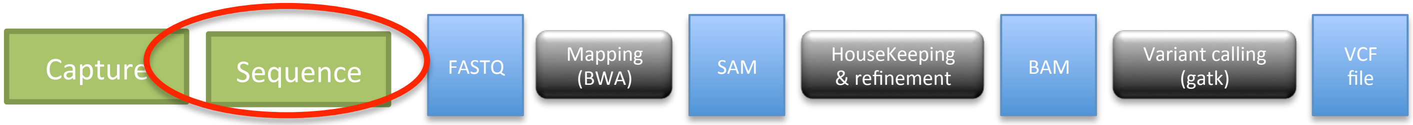
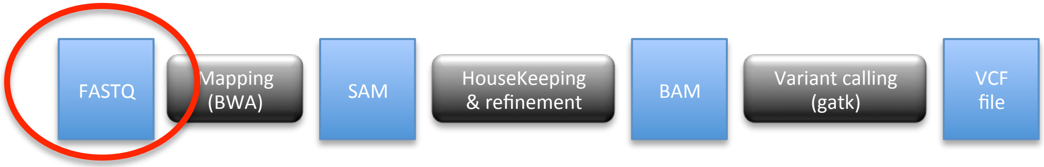


SEQUENCE – ESSENTIALS

Sequencing

- Here we will be assuming that sequencing was covered in a previous module of the course



FASTQ

Mapping (BWA)

SAM

HouseKeeping & refinement

BAM

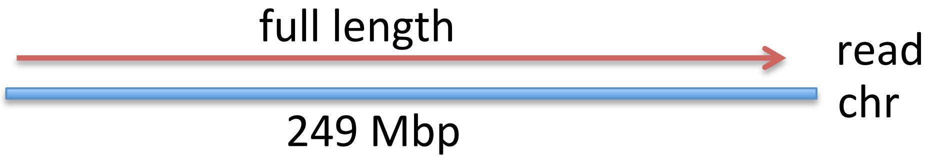
Variant calling (gatk)

VCF file

FASTQ FORMAT – ESSENTIALS

In a perfect world – Perfect sequencing

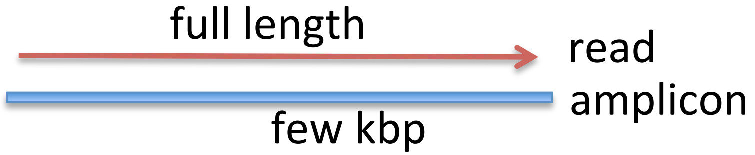
- Perfect sequencing:
  - single molecule (no PCR)
  - **full** length
  - no deterioration of quality
- While we are waiting:
  - Sanger
    - PCR
    - length: some kb
    - limited number of reads
    - high quality
  - HTS (Illumina)
    - PCR
    - 100 bp PE
    - billions of reads
    - high quality, but deteriorating along read


full length

read

chr

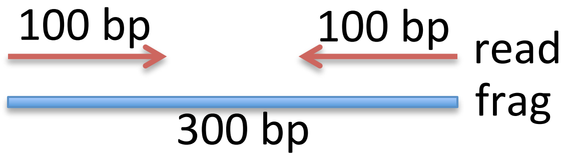
249 Mbp


full length

read

amplicon

few kbp

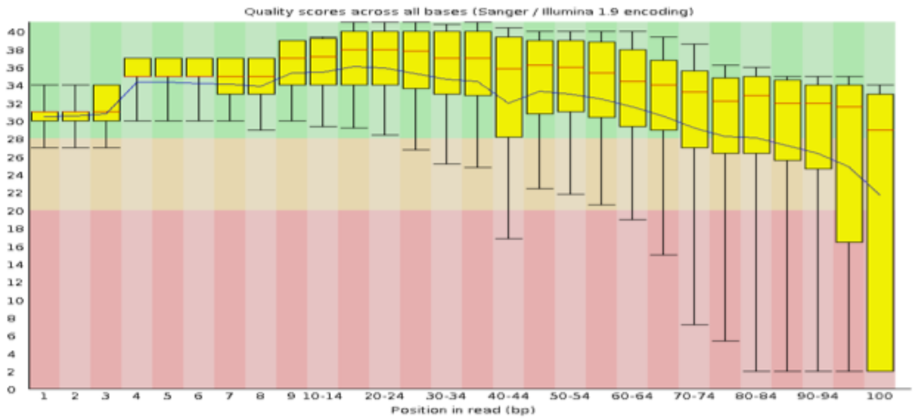

100 bp

100 bp

read

frag

300 bp

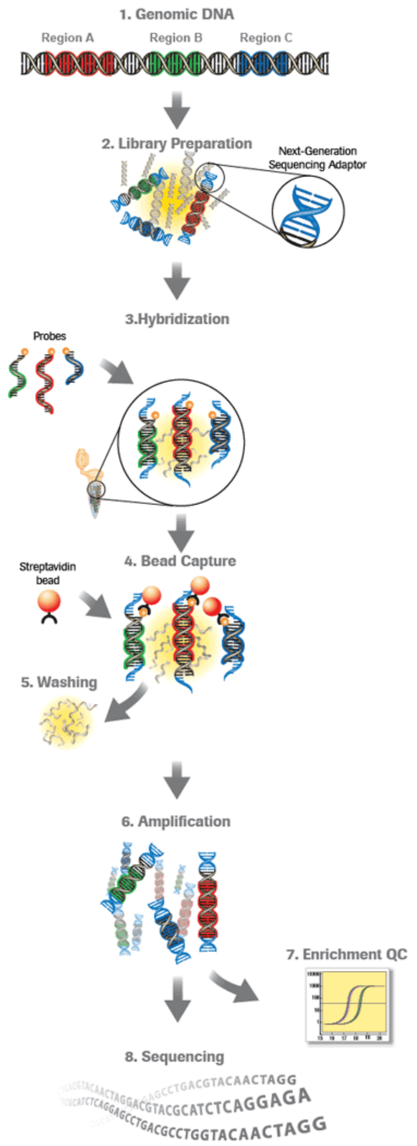


FASTQ format

Described in detail
in the algorithms module

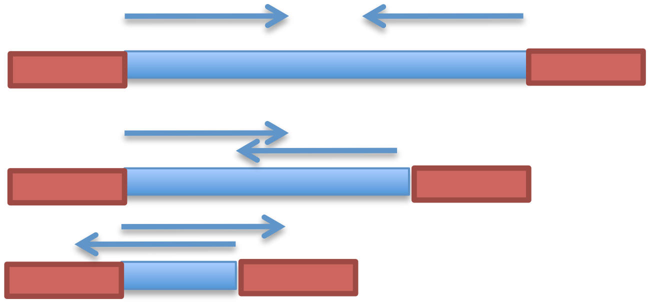
CLEANING UP FASTQ FILES

An overview of exome capture



Sonication

Library prep (sequencing adaptors on)



Problem: adaptor seq (red) in reads

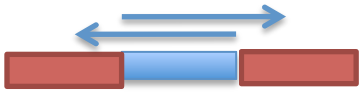
Hybridisation to probes → Possible biases in sequences that hybridise

Bead capture → Possible biases in sequences that elute

Amplification → Possible biases in sequences that amplify

Sequencing → Possible biases in sequences that bridge PCR

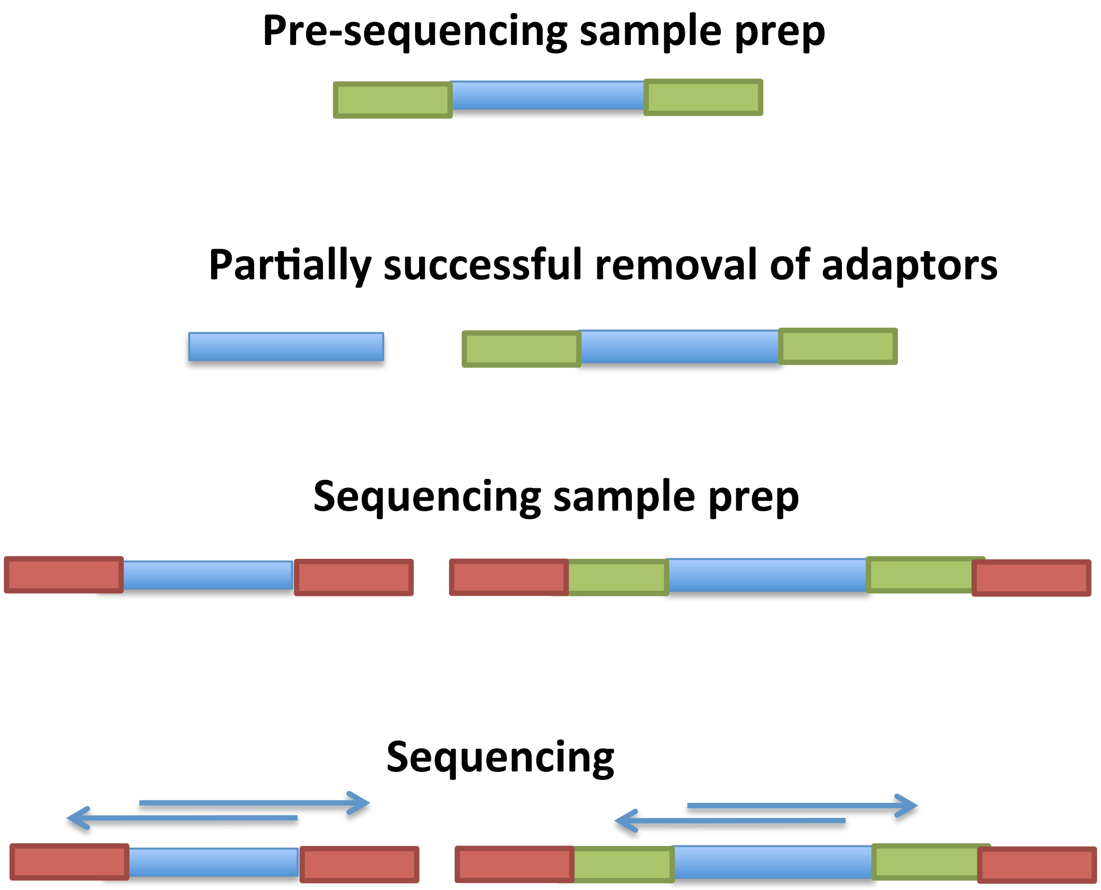
The importance of clean up



Problem: adaptor seq in reads

- There will be non-genomic sequences in reads
  - this will often prevent mapping
- Consequences
  - if short fragments are randomly distributed >> less reads mapping >> reduced coverage >> not a big problem as long as the fraction of reads affected is not too large
  - if short fragments are not randomly distributed (e.g. Halo capture) >> specific areas will suffer from reduced/no coverage >> **big problem**
  - **It all depends on the sample whether something really matters or not**

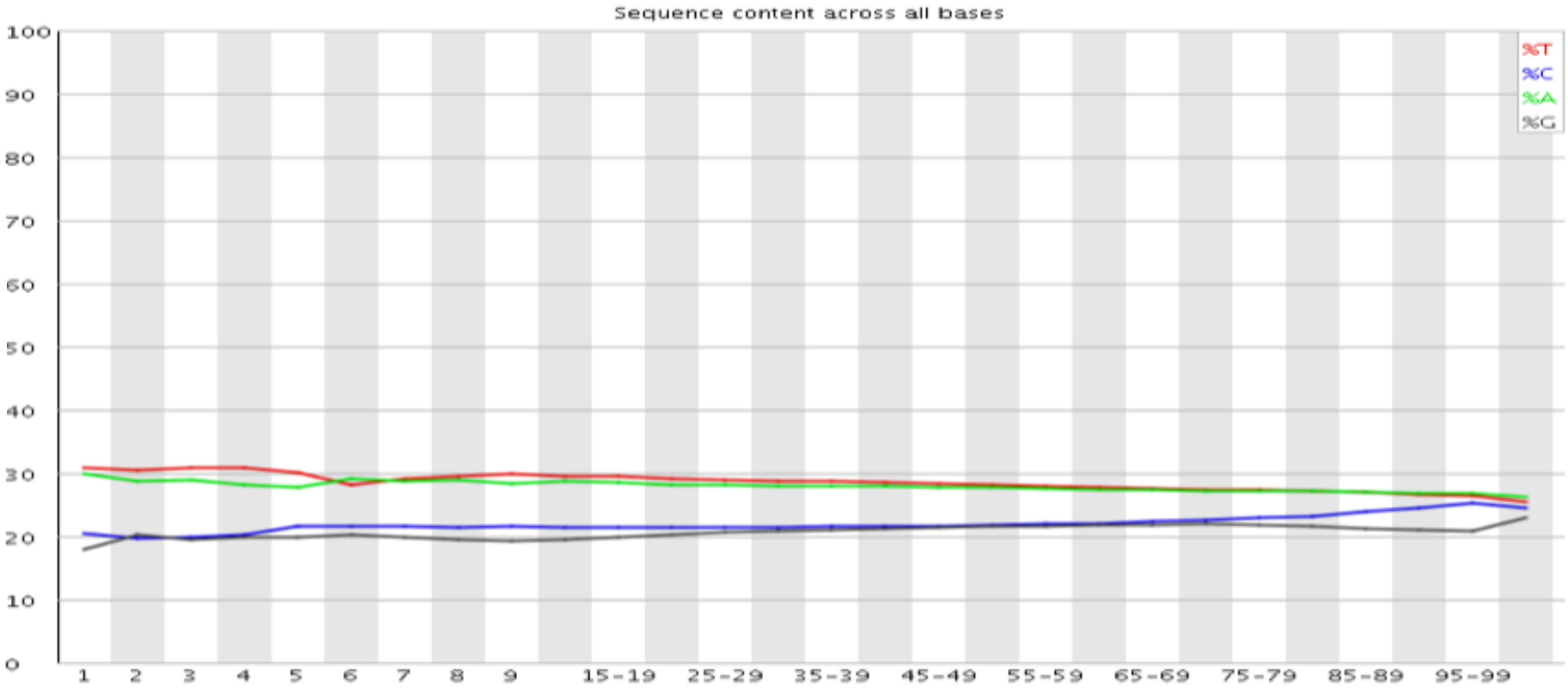
Chip-seq example


Pre-sequencing sample prep
Partially successful removal of adaptors
Sequencing sample prep
Sequencing

- If there are adaptors still attached to the genomic fragments in the material entering sequencing sample prep, the situation can be even more complicated.
- Fastqc report is essential to get a feel for whether clean-up is required

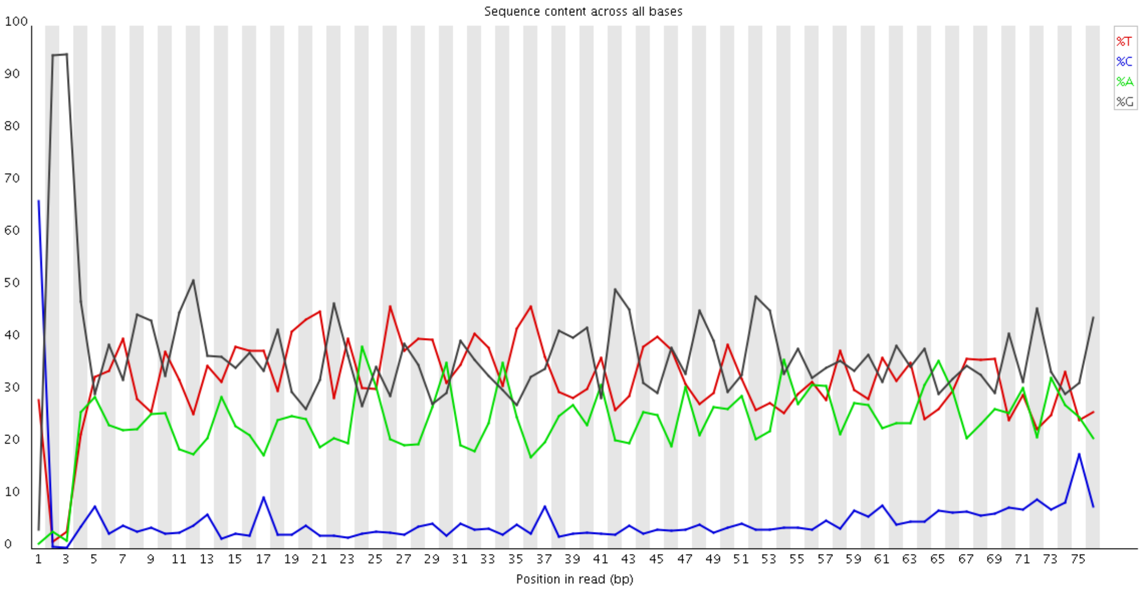
Per cycle sequence content ✓

%

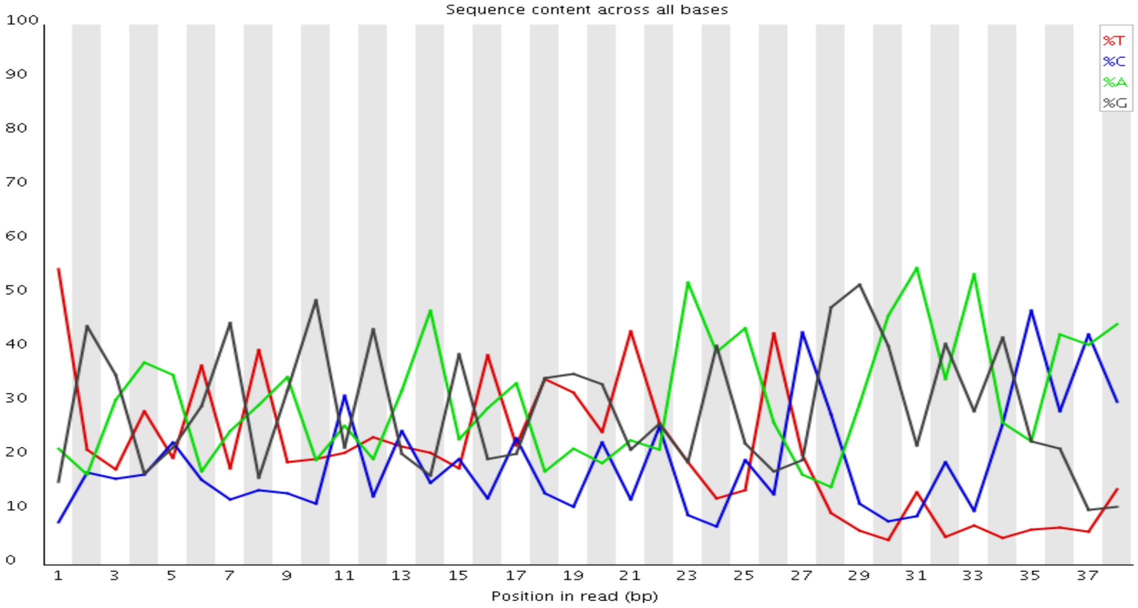


Position in read

Per cycle sequence content



?



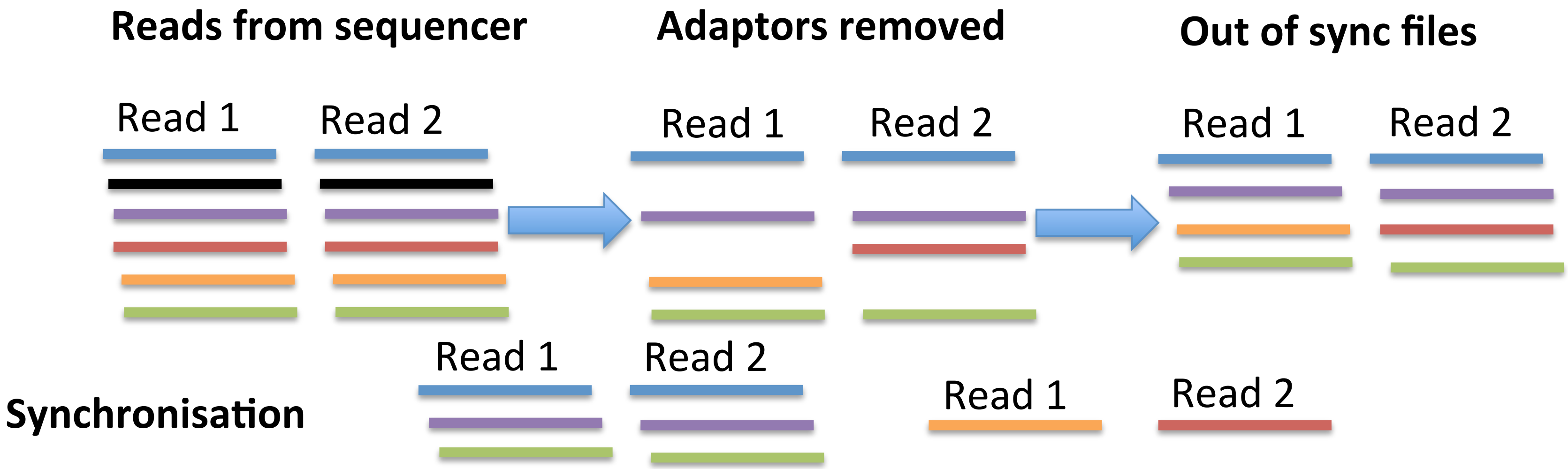
?

Tools for manipulating fasta and fastq files

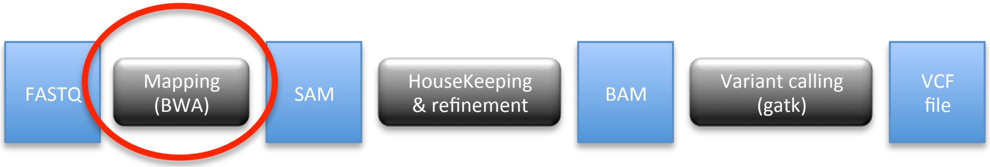
Detailed description
in the sequence pre-processing module

How to perform clean up

- Get an overview of the sample preparation both:
  - pre-sequencing
  - sequencing sample prep
- Run fastqc on the data
  - pay particular attention to the base distribution results
- Clean up using
  - fastx-toolkit
  - cutadapt
  - other tools
- Remember that you need to synchronise the R1 and R2 files (with PE)



NB: Synchronisation is very important even when not clipping off adaptors (files can be unsynchronised for other reasons)



MAPPING WITH BWA

Why mapping?

- The biggest difference with Sanger
  - we did not design and use primers for sequence amplification
  - we sonicated
  - >> we do not know where the reads “originate” from

- For each read
  - we need to determine its likely origin
  - how likely it is that we have correctly identified its origin

BWA mappping algorithm

Described in detail
in the algorithms module

Application to read mapping

- You need to build an index of the genome you wish to search
- Reads are 100 bp long
- But, the longer the string searched for the longer the search time **(longer backtracking)**, thus an advantage to search with a shorter string.
- There are 4 different nucleotides and 3.0E09 bases of genome
  - 4**16 = 4.0E09
  - 4**17 = 17.0E09
- Assuming complex sequence, 16 to 17 bp should be enough for a unique match
- BUT
  - variation in sample relative to reference
  - base call errors
  - **THUS** 17 bp search string is not enough
- BWA searches with a “seed” of 32 and allows 2 differences in the seed
- **Reads with seed matches are aligned using Smith-Waterman**

The Fasta, BWA index and other indexes in inputData

- Log into the machines and run
  cd vc/inputData/human_g1k_v37_chr5/gatkBundle
  ls –l *fasta*
- The biggest file is **human_g1k_v37_chr5.fasta**
- Take a look at this file with **less**
- Next biggest is **human_g1k_v37_chr5.fasta.bwt** What do you think this is?
- Another big one is **human_g1k_v37_chr5.fasta.sa** What do you think this is?
- Files with extension: amb, ann, pac, bwt and sa are generated by indexing of .fasta file
- .dict and .fai are very simple files needed by Picard and other programs

```
bwa index -a bwtsw genome.fa
samtools faidx genome.fa
java -jar CreateSequenceDictionary.jar  R=genome.fa O=genome.dict
samtools dict genome.fa
```

What are desirable characteristics of a read mapper?

- Accurately predict the source of a read
  - in the normal range of base error rates
  - in the normal range of indel frequency and size

- But, not necessary to get the alignment exactly right as this can be done later using multiple sequence alignment (MSA)

Reference	NNNNNCAAGNNNN
Sample	NNNNNCAAAGNNN

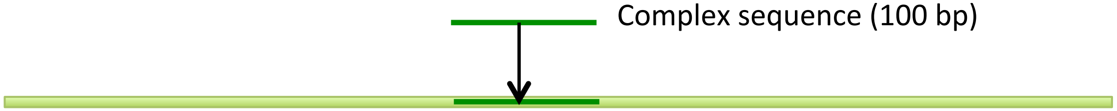
Reference	NNNNNCA_AGGNNN
Correct read align	NNNNNCAAAGNNNN

Reference	NNNNNCAAGGNNN
Alt. align	NNNNNCAAAGNNNN

- Produce an accurate estimate of the reliability of prediction

Factors complicating mapping

The simple case

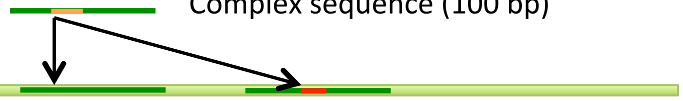


Complex sequence (100 bp)

Millions of reads

Billions of positions
in human genome

Homologous regions



Base call error

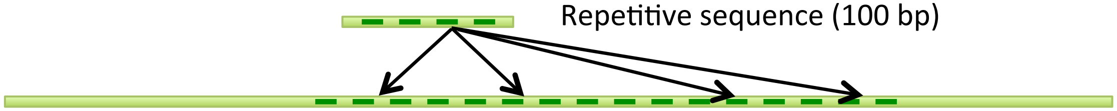
Complex sequence (100 bp)

gene 1A

gene 1B

Risk of mismapping

Repetitive region



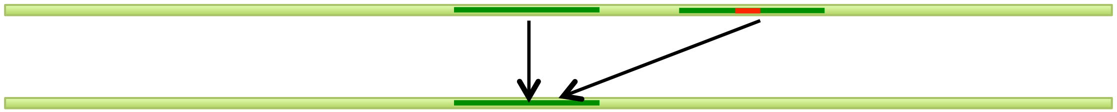
Repetitive sequence (100 bp)

big repetitive region

Impossible to be
map correctly

Structural variation (not in reference)

Duplication in sample which is not present in the reference



Definite
“mismapping”

Different programs

- BWA
- Novoalign
- BOWTIE
- SOAP
- ....
- Most based on BWT: Burrows-Wheeler Transform
  - a very neat computer algorithm for finding the location of substrings within a string
    - can I find atgc in attgcatcgatcga.......
  - requires indexing of string / reference, but enables
    - rapid search, necessary when mapping billions of reads
    - manageable RAM footprint: 2.3 GB for single reads and 3GB for paired-end (for BWA), so runs on an ordinary computer

Mapping quality scores

- The mapping quality score is the Phred-scaled probability of the mapping being incorrect.

$$Q_{\text{sanger}} = -10 \log_{10} p$$

- Probability is computed from the qualities of the mismatched bases between read and reference and quality features of the second best hit (see Li, Ruan, and Durbin 2008)

- BWA provides good mapping qualites with slight overestimation of quality score:
  - empirical error rate 7x10e-06 for Q60 mappings

BWA

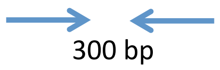
- Fast and accurate short read alignment with Burrows-Wheeler Transform
- But, in practice, makes changes to algorithm to adapt to biological reality and increase speed
- Paired-end mapping
- Like similar programs, randomly places a repetitive read across the multiple equally best positions and mapping quality 0
- Supports multi-threading (as do all BWT aligners)

Some details of paired-end mapping

- Paired-end data contains additional information relative to single read data which can be exploited to identiy abnormal pairs and correctly map both reads in a pair

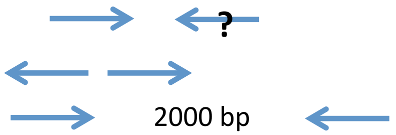
Normal situation

- both reads map
- orientation
- insert size

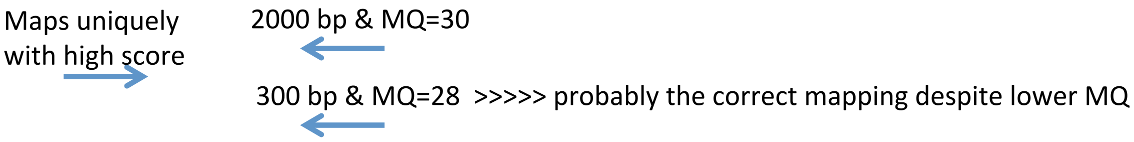


Abnormal situation

- one read does not map
- bad orientation
- big insert size



Exploiting PE information

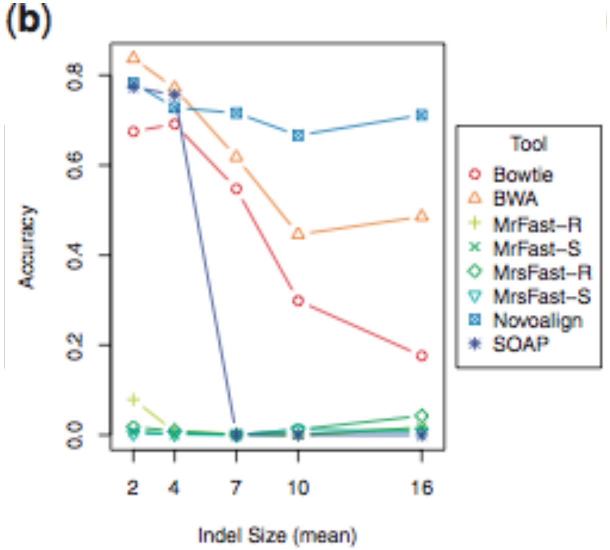


300 bp & MQ=28 >>>>> probably the correct mapping despite lower MQ

- For details of how orientation, insert size and alignment quality are handled by BWA, see "Aligning sequence reads, close sequences and assembly contigs with bwa-mem", Heng Li 2013

Accuracy with varying indel sizes

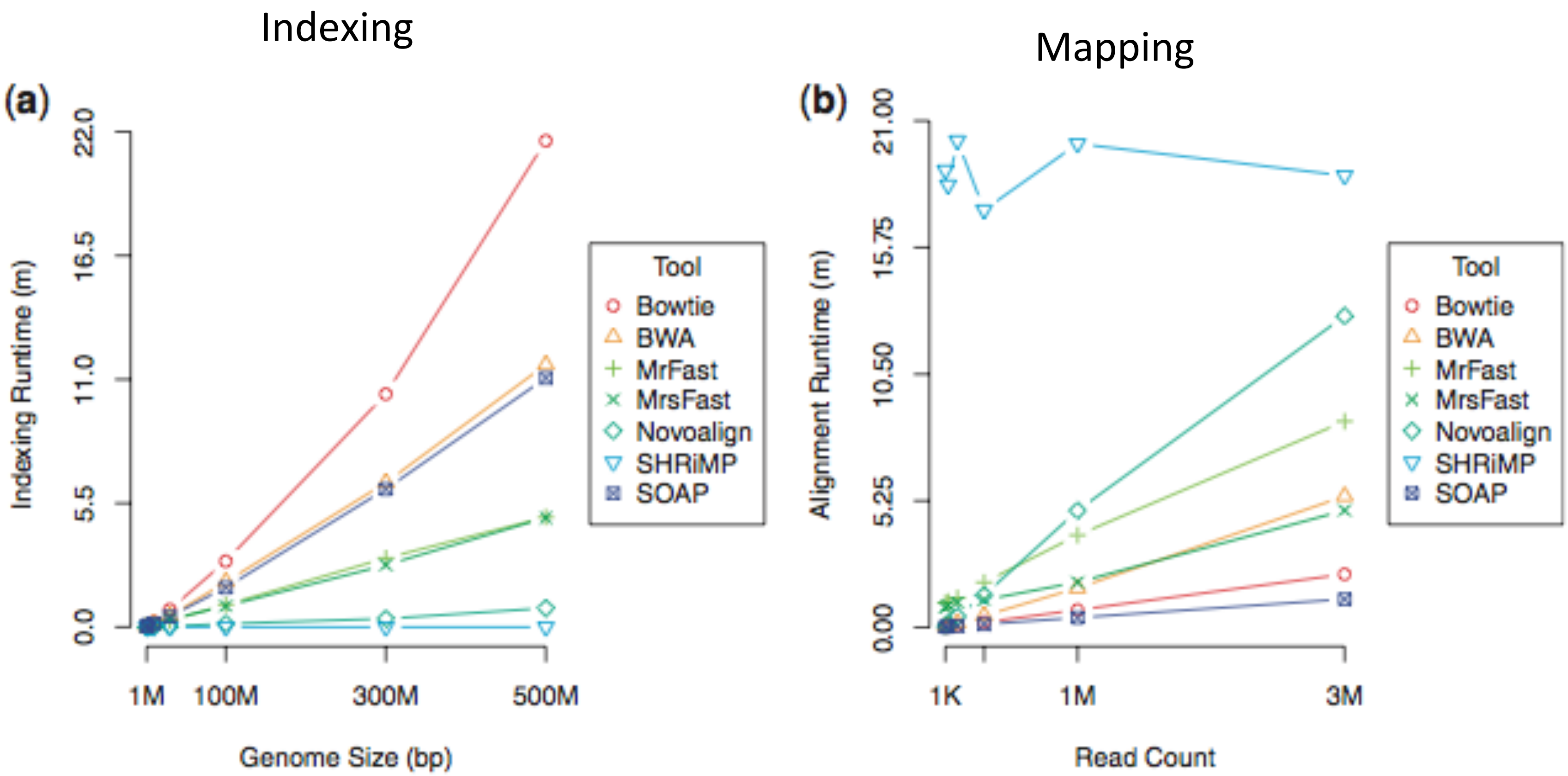
Accuracy



- BWA and Novoalign are best performing

Source: Ruffalo et al. 2011

Runtime


Indexing

Mapping

500 MB genome

Source: Ruffalo et al. 2011

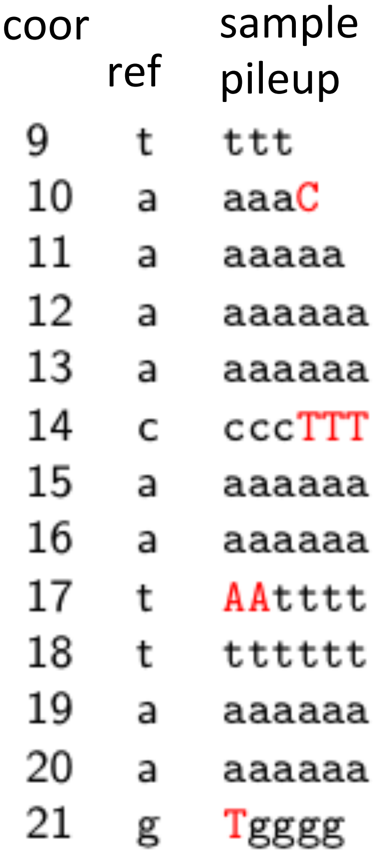
Words of caution about the comparison

- Note this results are for SR 50 bp reads
  - reads are often longer with Illumina
  - can be much longer with other technologies and will require different mapper (see Fast and accurate long-read alignment with Burrows-Wheeler transform)

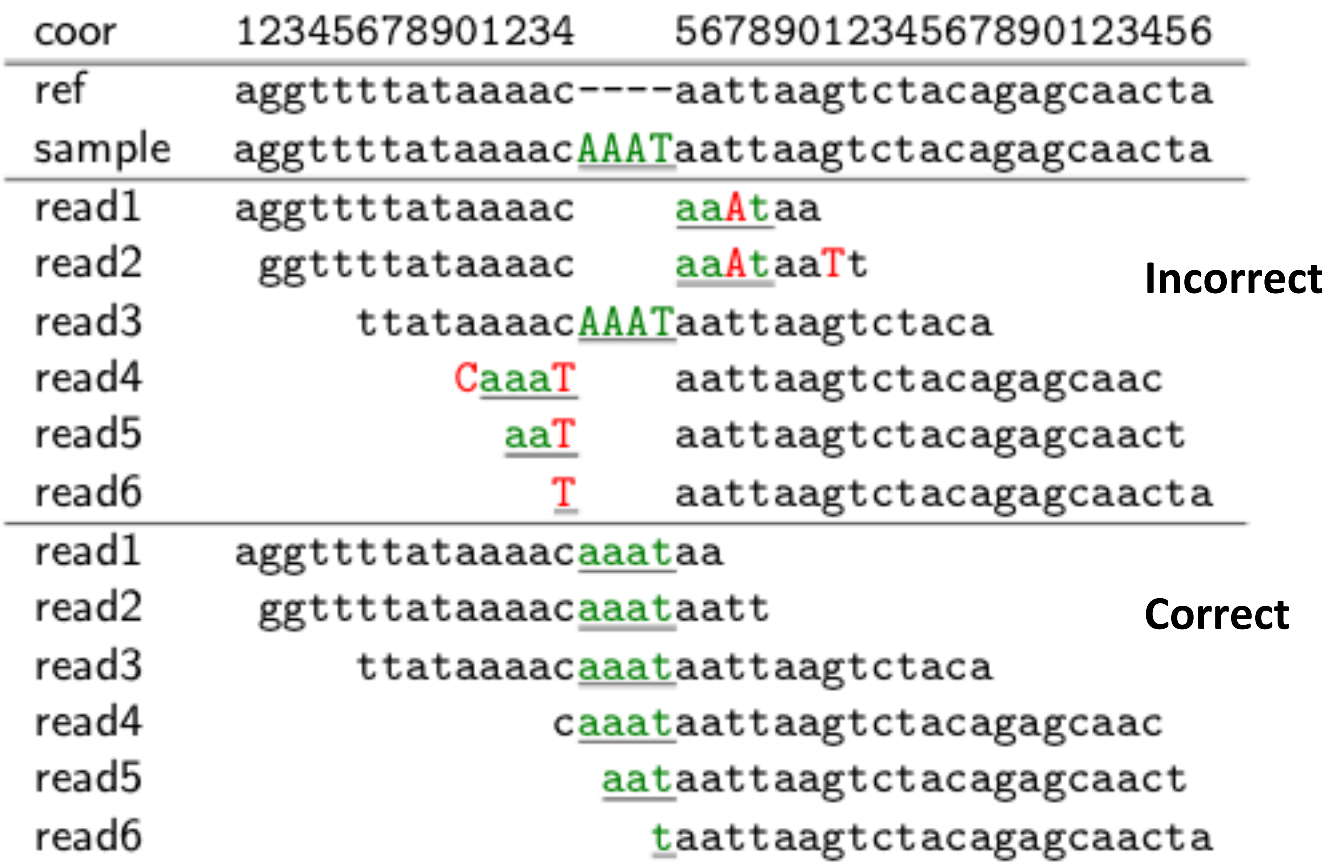
- **But, in summary:**
  - **mappers are not all of the same quality**
  - **bwa and novoalign are amongst the best for Illumina reads**

Alignment errors after mapping

Base stacks



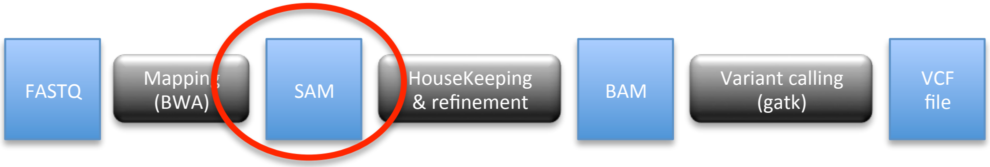
coor	ref	sample pileup
9	t	ttt
10	a	aaaC
11	a	aaaaa
12	a	aaaaaa
13	a	aaaaaa
14	c	cccTTT
15	a	aaaaaa
16	a	aaaaaa
17	t	AAtttt
18	t	tttttt
19	a	aaaaaa
20	a	aaaaaa
21	g	Tgggg



```
coor    12345678901234    5678901234567890123456
ref     aggttttataaaac----aattaagtctacagagcaacta
sample  aggttttataaaacAAATaattaagtctacagagcaacta
read1   aggttttataaaac    aaAtaa
read2    ggttttataaaac    aaAtaaTt                  Incorrect
read3        ttataaaacAAATaattaagtctaca
read4            CaaaT    aattaagtctacagagcaac
read5              aaT    aattaagtctacagagcaact
read6                T    aattaagtctacagagcaacta
read1   aggttttataaaacaaataa
read2    ggttttataaaacaaataatt                      Correct
read3        ttataaaacaaataattaagtctaca
read4                caaataattaagtctacagagcaac
read5                  aataattaagtctacagagcaact
read6                    taattaagtctacagagcaacta
```

>> Can be solved by alignment: considering all mapping reads and reference together (as we shall see later)

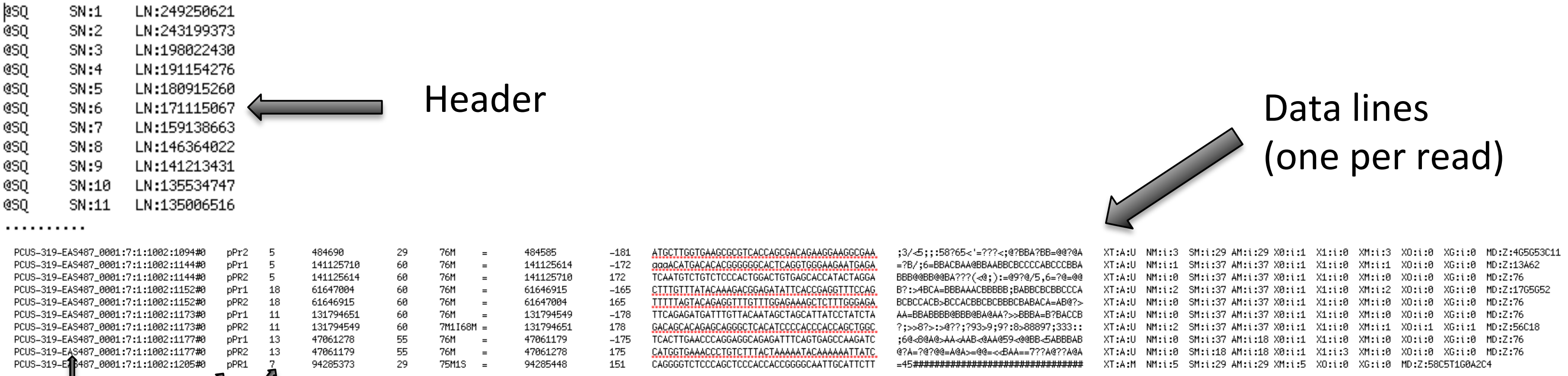
Source: Heng Li, presentation at GSA workshop 2011



FASTQ
Mapping
(BWA)
SAM
HouseKeeping
& refinement
BAM
Variant calling
(gatk)
VCF
file

SAM FORMAT

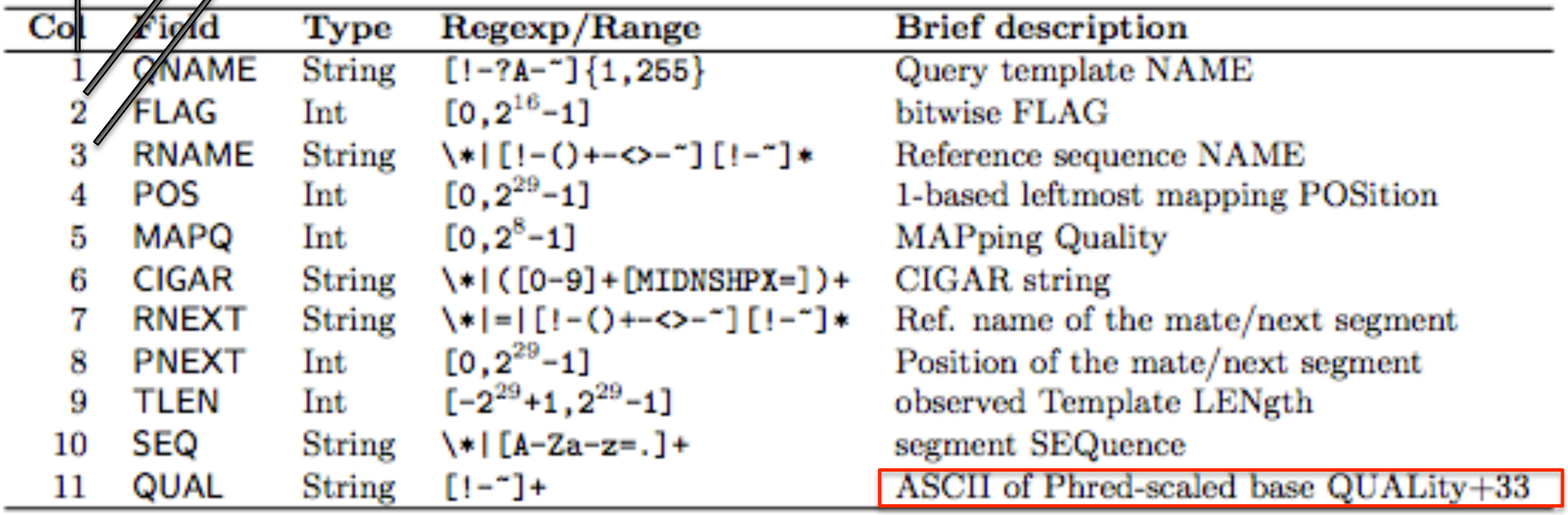
What does the SAM file look like?



```
@SQ     SN:1    LN:249250621
@SQ     SN:2    LN:243199373
@SQ     SN:3    LN:198022430
@SQ     SN:4    LN:191154276
@SQ     SN:5    LN:180915260
@SQ     SN:6    LN:171115067
@SQ     SN:7    LN:159138663
@SQ     SN:8    LN:146364022
@SQ     SN:9    LN:141213431
@SQ     SN:10   LN:135534747
@SQ     SN:11   LN:135006516
..........

PCUS-319-EAS487_0001:7:1:1002:1094#0   pPr2   5    484690      29   76M      =   484585      -181   ATGCTTGGTGAAGCGCGTCACCAGCGACAGAAGGAAGGCGAA   ;3/<5;;:58?65<'=???<;@?BBA?BB=@@?@A    XT:A:U   NM:i:3   SM:i:29 AM:i:29 X0:i:1  X1:i:0  XM:i:3  XO:i:0  XG:i:0  MD:Z:4G5G53C11
PCUS-319-EAS487_0001:7:1:1002:1144#0   pPr1   5    141125710   60   76M      =   141125614   -172   gggACATGACACACGGGGGGCACTCAGGTGGGAAGAATGAGA   =?B/;6=BBACBAA@BBAABBCBCCCCABCCCBBA    XT:A:U   NM:i:1   SM:i:37 AM:i:37 X0:i:1  X1:i:0  XM:i:1  XO:i:0  XG:i:0  MD:Z:13A62
PCUS-319-EAS487_0001:7:1:1002:1144#0   pPR2   5    141125614   60   76M      =   141125710   172    TCAATGTCTGTCTCCCACTGGACTGTGAGCACCATACTAGGA   BBB@@BB@@BA???(<@;):=@9?@/5,6=?@=@@    XT:A:U   NM:i:0   SM:i:37 AM:i:37 X0:i:1  X1:i:0  XM:i:0  XO:i:0  XG:i:0  MD:Z:76
PCUS-319-EAS487_0001:7:1:1002:1152#0   pPr1   18   61647004    60   76M      =   61646915    -165   CTTTGTTTATACAAAGACGGAGATATTCACCGAGGTTTCCAG   B?:>4BCA=BBBAAACBBBBB;BABBCBCBBCCCA    XT:A:U   NM:i:2   SM:i:37 AM:i:37 X0:i:1  X1:i:0  XM:i:2  XO:i:0  XG:i:0  MD:Z:17G5G52
PCUS-319-EAS487_0001:7:1:1002:1152#0   pPR2   18   61646915    60   76M      =   61647004    165    TTTTTAGTACAGAGGTTTGTTTGGAGAAAGCTCTTTGGGAGA   BCBCCACB>BCCACBBCBCBBBCBABACA=AB@?>    XT:A:U   NM:i:0   SM:i:37 AM:i:37 X0:i:1  X1:i:0  XM:i:0  XO:i:0  XG:i:0  MD:Z:76
PCUS-319-EAS487_0001:7:1:1002:1173#0   pPr1   11   131794651   60   76M      =   131794549   -178   TTCAGAGATGATTTGTTACAATAGCTAGCATTATCCTATCTA   AA=BBABBBB@BBB@BA@AA?>>BBBA=B?BACCB    XT:A:U   NM:i:0   SM:i:37 AM:i:37 X0:i:1  X1:i:0  XM:i:0  XO:i:0  XG:i:0  MD:Z:76
PCUS-319-EAS487_0001:7:1:1002:1173#0   pPR2   11   131794549   60   7M1I68M  =   131794651   178    GACAGCACAGAGCAGGGCTCACATCCCCACCCACCAGCTGGC   ?;>>8?>:>@??;?93>9;9?:8>88897;333::    XT:A:U   NM:i:2   SM:i:37 AM:i:37 X0:i:1  X1:i:0  XM:i:1  XO:i:1  XG:i:1  MD:Z:56C18
PCUS-319-EAS487_0001:7:1:1002:1177#0   pPr1   13   47061278    55   76M      =   47061179    -175   TCACTTGAACCCAGGAGGCAGAGATTTCAGTGAGCCAAGATC   ;6@<8@A@>AA<AAB<@AA@59<@@BB<5ABBBAB    XT:A:U   NM:i:0   SM:i:37 AM:i:18 X0:i:1  X1:i:0  XM:i:0  XO:i:0  XG:i:0  MD:Z:76
PCUS-319-EAS487_0001:7:1:1002:1177#0   pPR2   13   47061179    55   76M      =   47061278    175    CATGGTGAAACCCTGTCTTTACTAAAAATACAAAAAATTATC   @?A=?@?@@=A@A>=@@=<<BAA==7??A@??A@A    XT:A:U   NM:i:0   SM:i:18 AM:i:18 X0:i:1  X1:i:3  XM:i:0  XO:i:0  XG:i:0  MD:Z:76
PCUS-319-EAS487_0001:7:1:1002:1205#0   pPR1   7    94285373    29   75M1S    =   94285448    151    CAGGGGTCTCCCAGCTCCCACCACCGGGGCAATTGCATTCTT   =45################################    XT:A:M   NM:i:5   SM:i:29 AM:i:29 XM:i:5  XO:i:0  XG:i:0  MD:Z:58C5T1G0A2C4
```

Header

Data lines
(one per read)

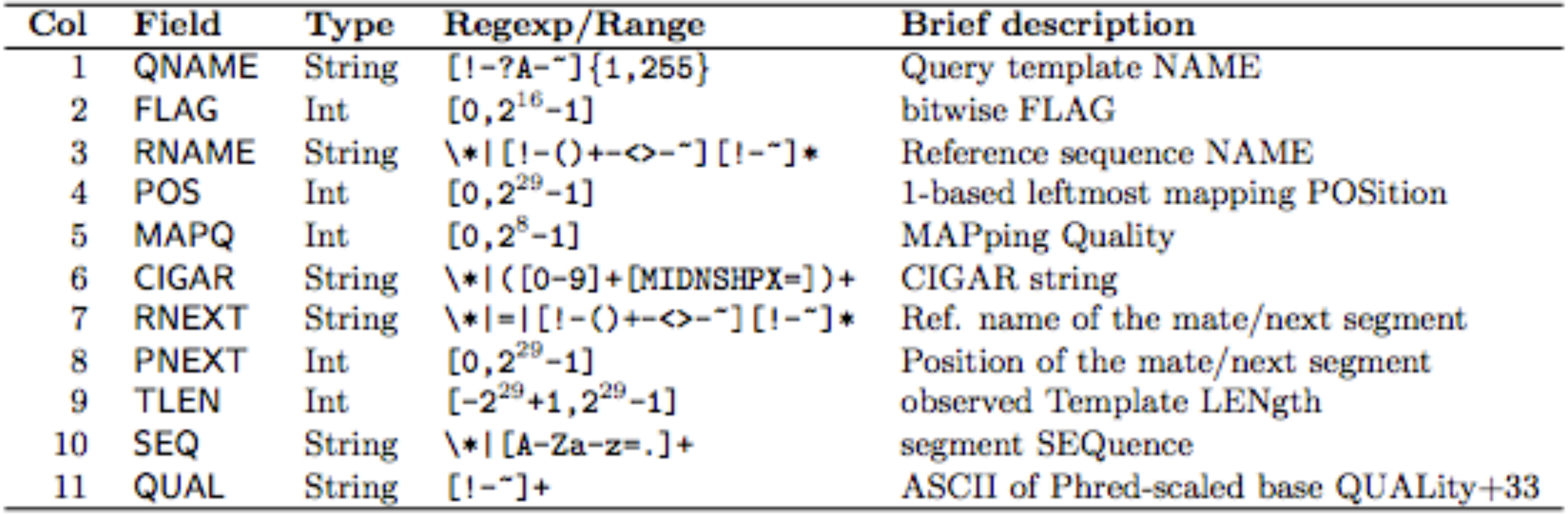


Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	[!-?A-~]{1,255}	Query template NAME
2	FLAG	Int	$[0,2^{16}-1]$	bitwise FLAG
3	RNAME	String	*|[!-()+-<>-~][!-~]*	Reference sequence NAME
4	POS	Int	$[0,2^{29}-1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0,2^{8}-1]$	MAPping Quality
6	CIGAR	String	*|([0-9]+[MIDNSHPX=])+	CIGAR string
7	RNEXT	String	*|=|[!-()+-<>-~][!-~]*	Ref. name of the mate/next segment
8	PNEXT	Int	$[0,2^{29}-1]$	Position of the mate/next segment
9	TLEN	Int	$[-2^{29}+1,2^{29}-1]$	observed Template LENgth
10	SEQ	String	*|[A-Za-z=.]+	segment SEQuence
11	QUAL	String	[!-~]+	ASCII of Phred-scaled base QUALity+33

Source: The SAM format specification

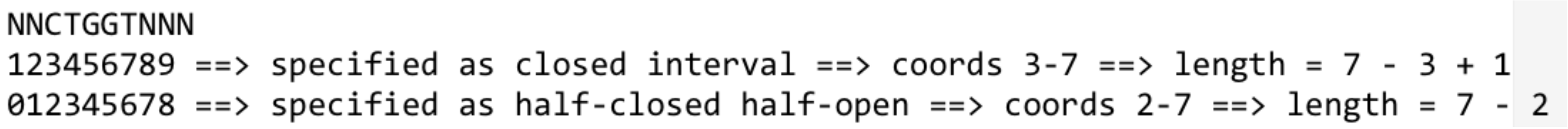
Inspecting one record

```
PCUS-319-EAS487_0001:7:1:1002:1094#0
pPr2
5
484690
29
76M
=
484585
-181
ATGCTTGGTGAAGCGCGTCACCAGCGACAGAAGGAAGGCGAA
;;;;;;;;3/<5;;:58?65<'=???<;@?BBA?BB=@@?@A
```



Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	[!-?A-~]{1,255}	Query template NAME
2	FLAG	Int	[0,2^{16}-1]	bitwise FLAG
3	RNAME	String	*|[!-()+-<>-~][!-~]*	Reference sequence NAME
4	POS	Int	[0,2^{29}-1]	1-based leftmost mapping POSition
5	MAPQ	Int	[0,2^{8}-1]	MAPping Quality
6	CIGAR	String	*|([0-9]+[MIDNSHPX=])+	CIGAR string
7	RNEXT	String	*|=|[!-()+-<>-~][!-~]*	Ref. name of the mate/next segment
8	PNEXT	Int	[0,2^{29}-1]	Position of the mate/next segment
9	TLEN	Int	[-2^{29}+1,2^{29}-1]	observed Template LENgth
10	SEQ	String	*|[A-Za-z=.]+	segment SEQuence
11	QUAL	String	[!-~]+	ASCII of Phred-scaled base QUALity+33

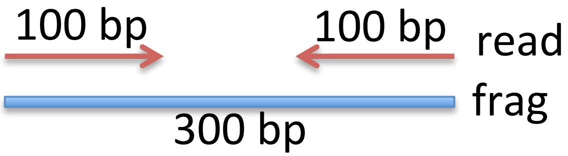
Difference between 1-based and 0-based coordinates

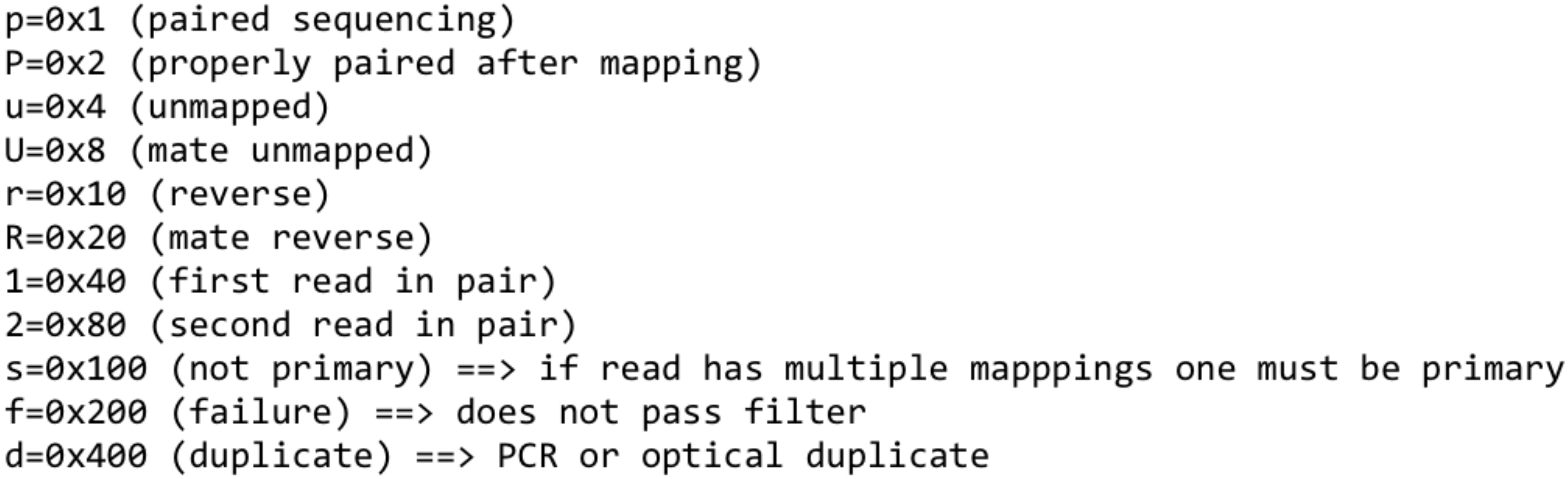


```
NNCTGGTNNN
123456789 ==> specified as closed interval ==> coords 3-7 ==> length = 7 - 3 + 1
012345678 ==> specified as half-closed half-open ==> coords 2-7 ==> length = 7 - 2
```

- SAM (+ VCF and GFF) are 1-based
- BED are 0-based
- Can be very important when manipulating SNP coordinates >> be careful

The FLAG column – a bit wise flag

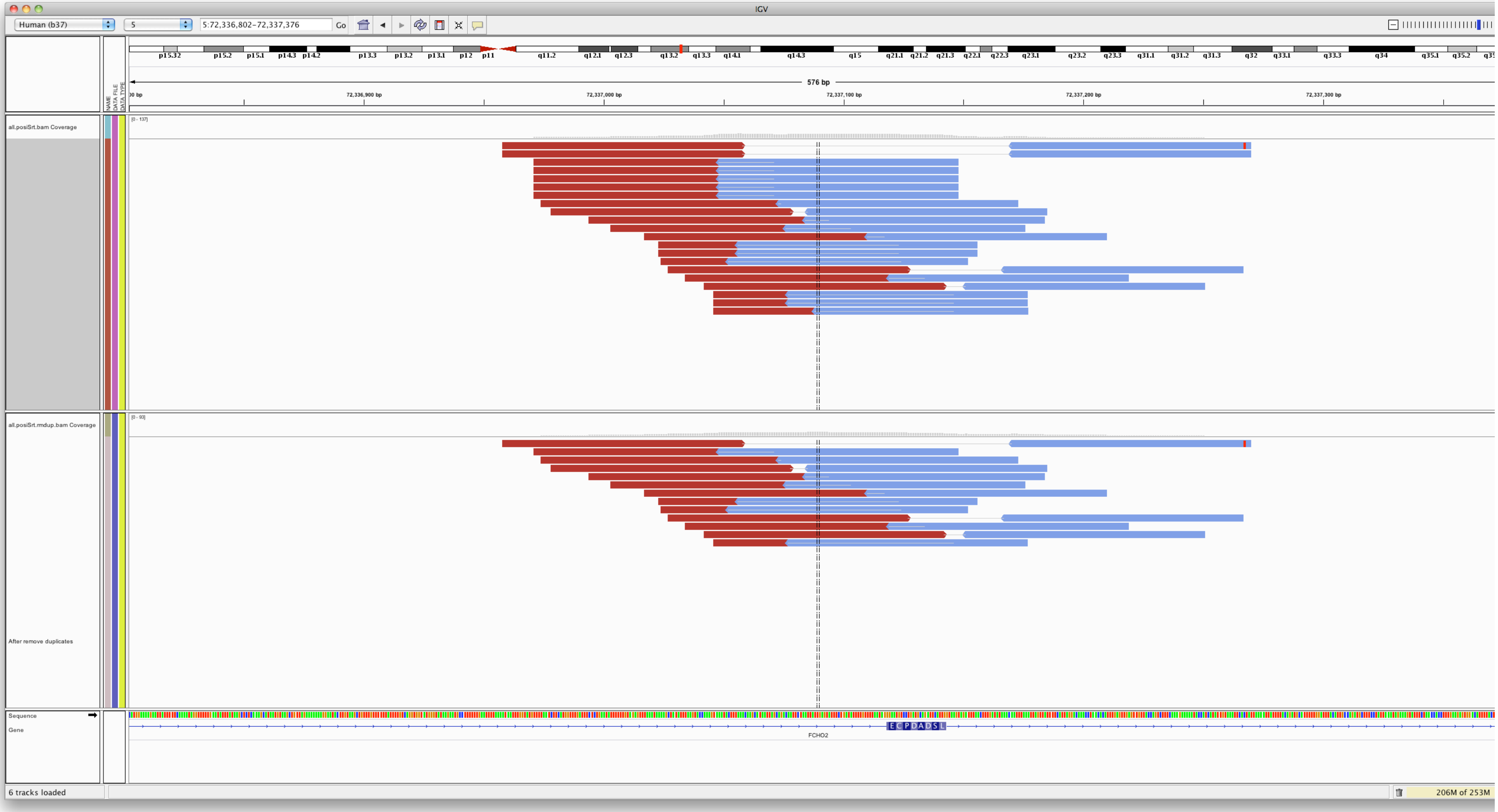




```
p=0x1 (paired sequencing)
P=0x2 (properly paired after mapping)
u=0x4 (unmapped)
U=0x8 (mate unmapped)
r=0x10 (reverse)
R=0x20 (mate reverse)
1=0x40 (first read in pair)
2=0x80 (second read in pair)
s=0x100 (not primary) ==> if read has multiple mapppings one must be primary
f=0x200 (failure) ==> does not pass filter
d=0x400 (duplicate) ==> PCR or optical duplicate
```

- Translate from bit wise flag to readable codes by using **samtools view –X**
- **OR** take a look at **https://broadinstitute.github.io/picard/explain-flags.html**

Introducing flags - What is a duplicate?



About the SAM file produced by BWA

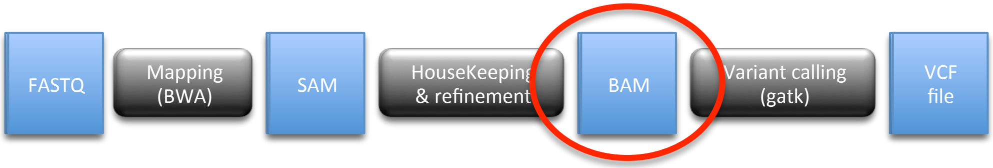
- It contains **all** the reads >> the Picard/GATK paradigm: information is annotated (and not filtered)
  - unique
  - ambiguous
  - unmapped
- It has a number of short comings
  - it takes a lot of space ➔ convert to BAM
  - the mates are not fully updated on each others existence ➔ fixmate
  - it is not sorted ➔ sort
  - it contains PCR duplicates ➔ mark or remove duplicates
  - it does not contain meta-data on the reads (sample, sequencer, etc)

What tools can I use to manipulate the data?

- Two alternative software packages for doing this
  - samtools
    - basic
    - easy to use
  - picard
    - advanced functionality
    - used by GATK team
- For simple tasks like viewing, indexing, sorting ➔ samtools
- For more advanced tasks and preparing for downstream analysis ➔ Picard due to GATKs pickyness

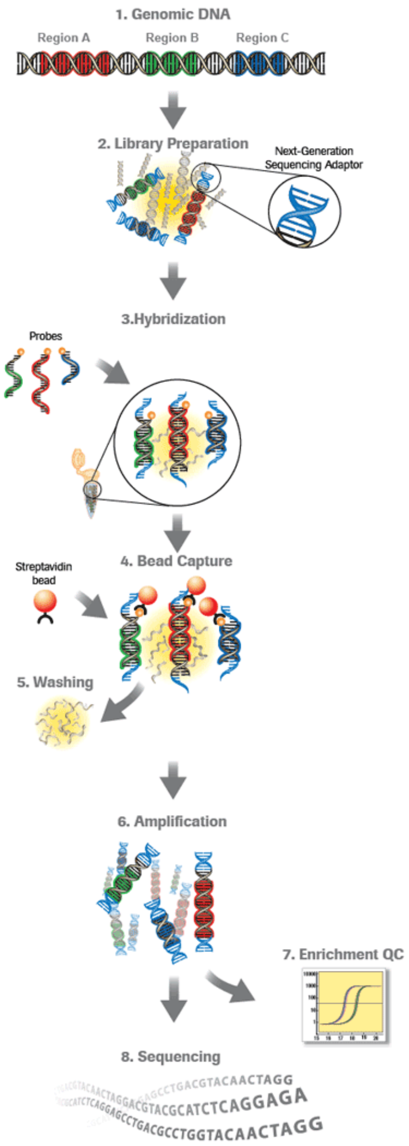
- These tools will be introduced in more detail in the practicals

- **Brief practical: Take a look at your SAM file**



COMPUTING ADVANCED METRICS – PICARD

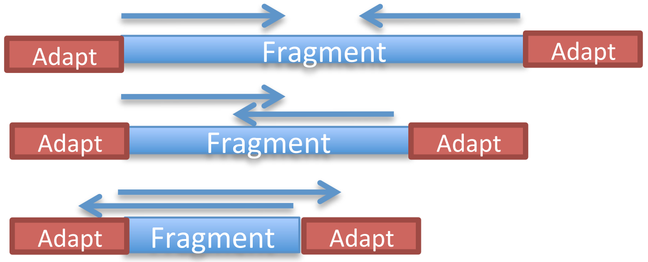
An overview of exome capture – weak points



Sonication

Library prep (sequencing adaptors on)

Problem: **error in sonication** >> adaptor seq in reads >> **unmapped reads**



Hybridisation to probes

Possible biases in sequences that hybridise >> **coverage bias**

Bead capture

Possible biases in sequences that elute >> **coverage bias**

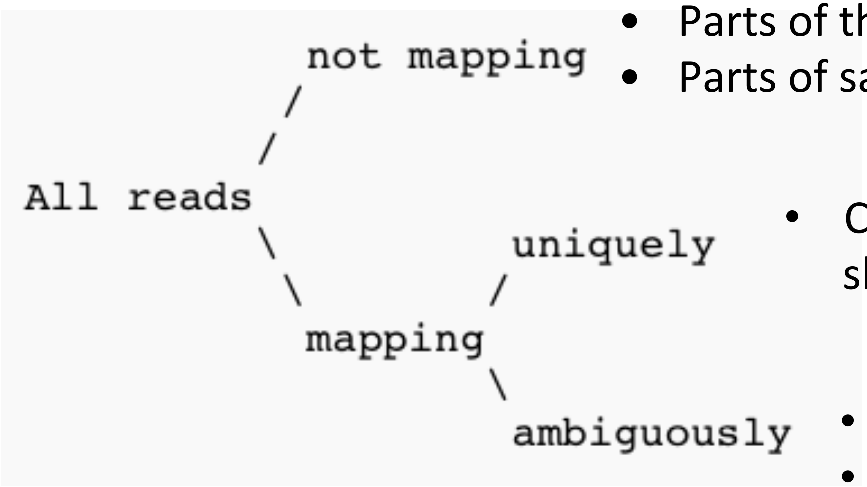
Amplification

Possible biases in sequences that amplify **>> sequence PCR duplicates**

Sequencing

Possible biases in sequences that bridge PCR >> **coverage bias**

Metrics - Basic read classification



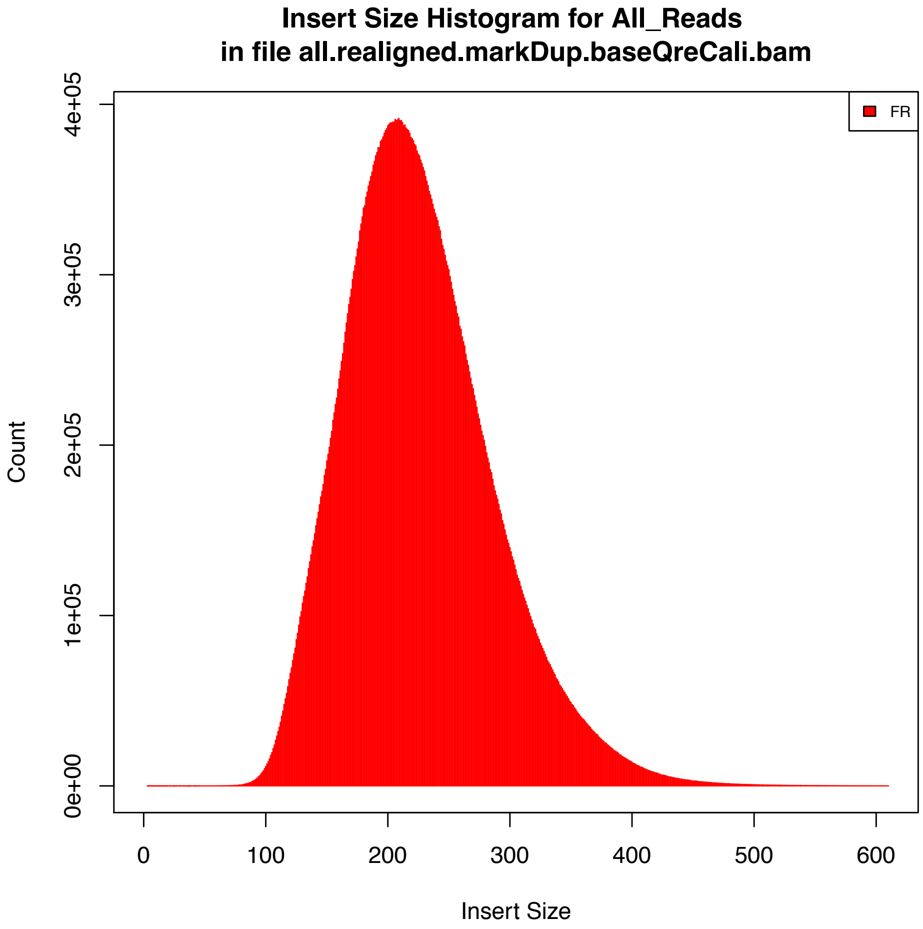
```
               not mapping
              /
             /
All reads
             \
              \           uniquely
                         /
               mapping
                         \
                          ambiguously
```

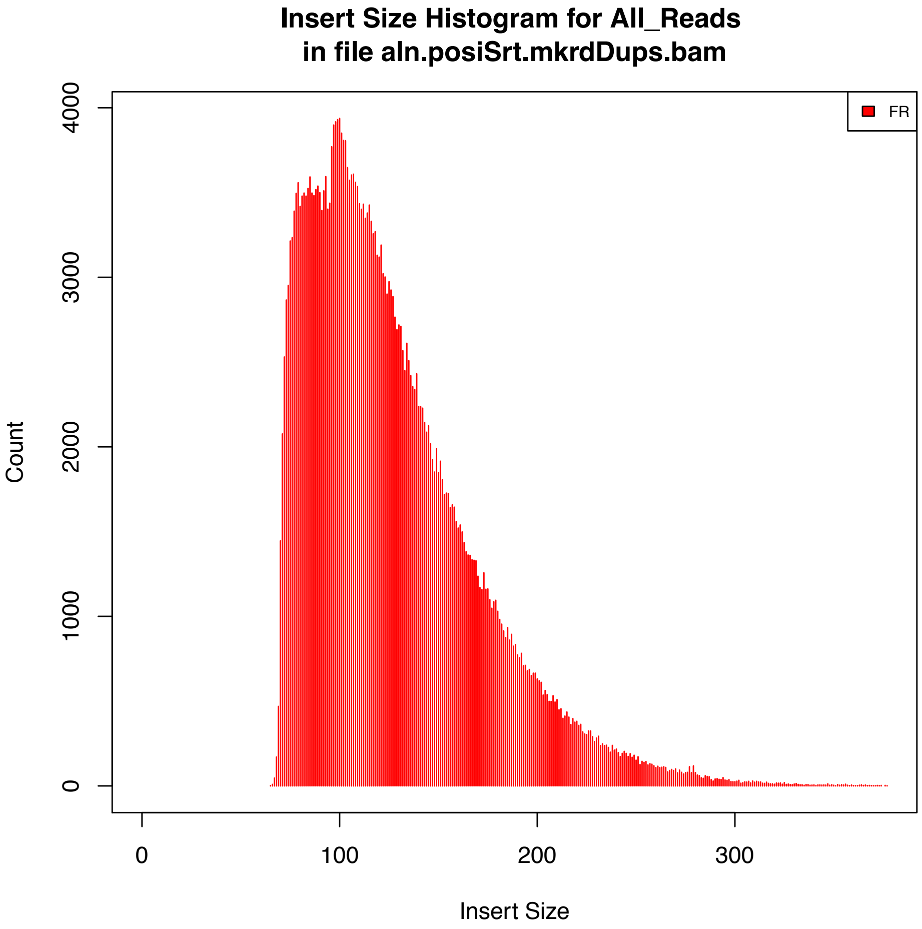
- Contamination from other species
- Reads containing non-genomic DNA e.g. adaptors
- PCR gunk
- Reads with sequencing errors
- Parts of the genome that are not assembled
- Parts of sample affected by structural variation

- Complex sequence with sufficient length should map uniquely

- Typical for repeats
- Also possible for homologous regions

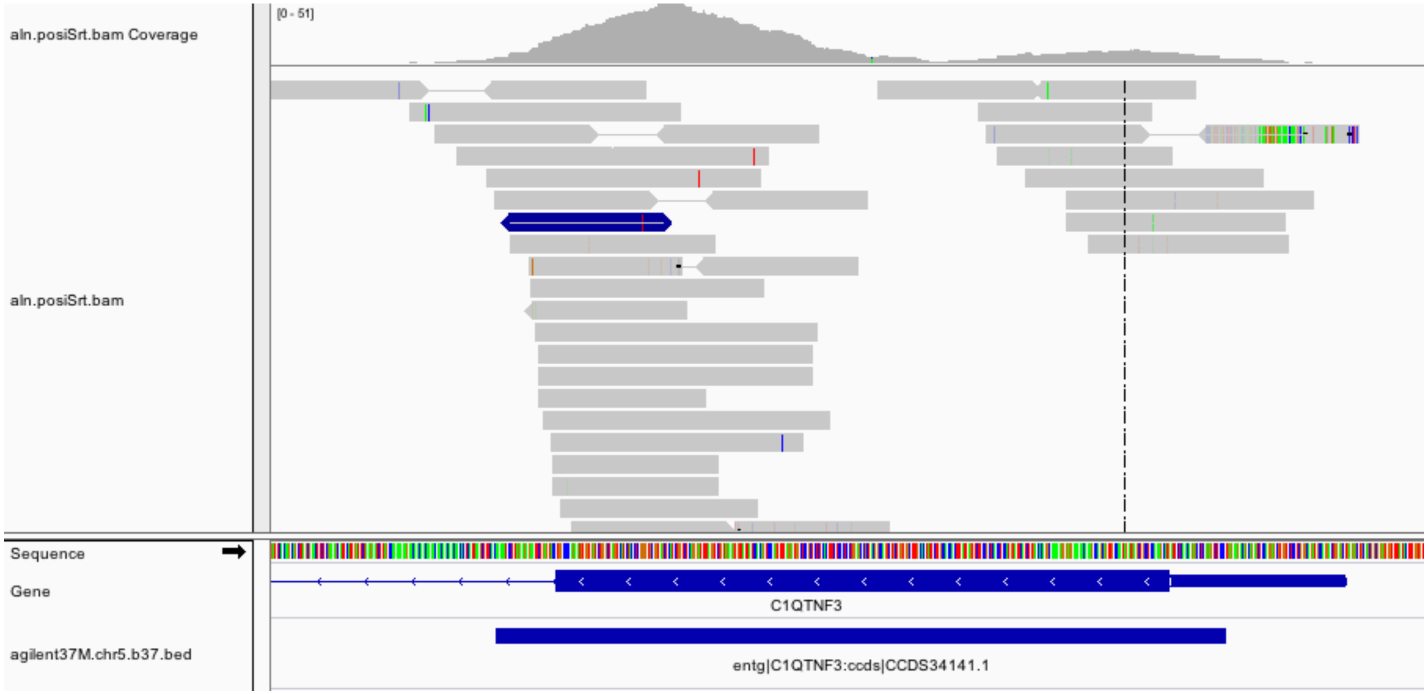
Metrics – Insert sizes





Metrics - Coverage

- Even if doing Whole Genome Sequencing (WGS) >> coverage issues
  - due to repetitive regions
  - due to properties of the DNA e.g. GC content
- Exome sequencing >> Capture by hybridisation



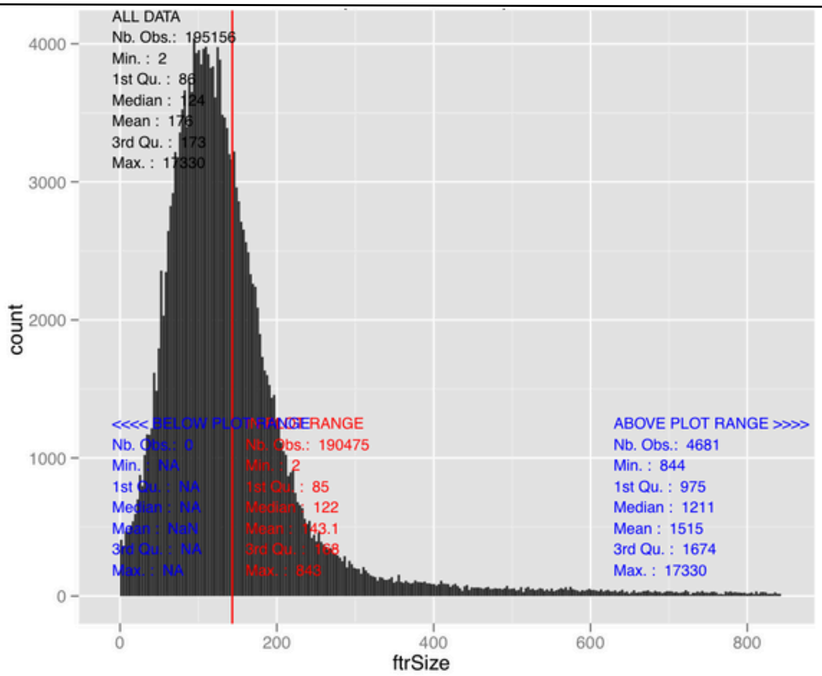
Coverage

Reads

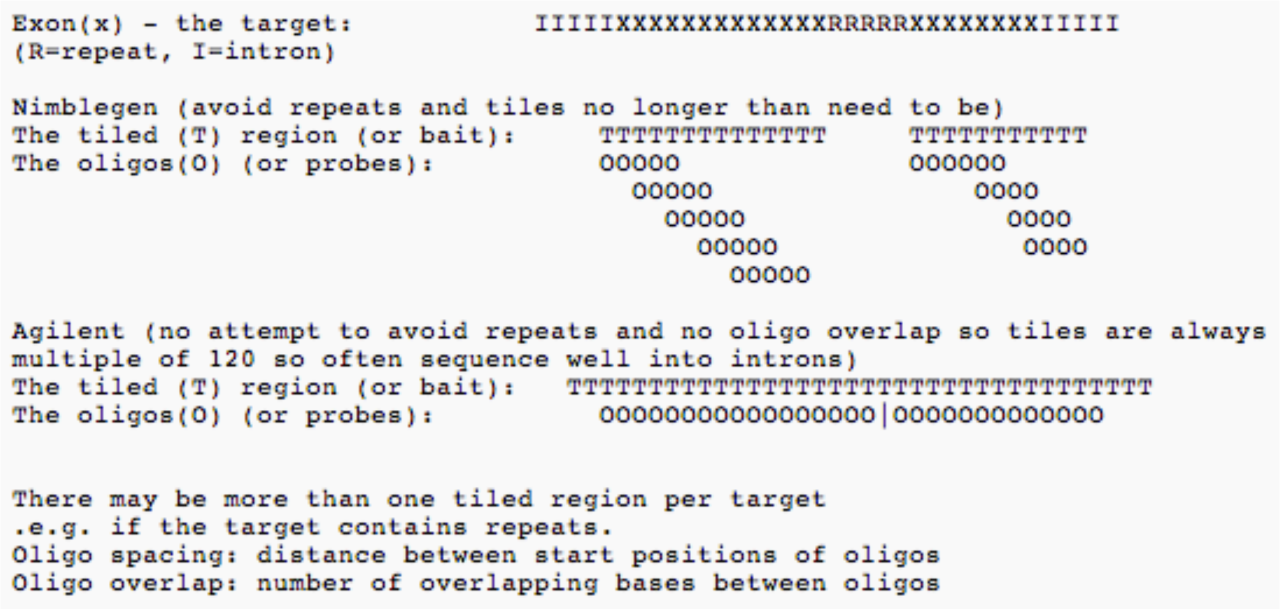
Target

Tile

Exome capture – The nature of the target

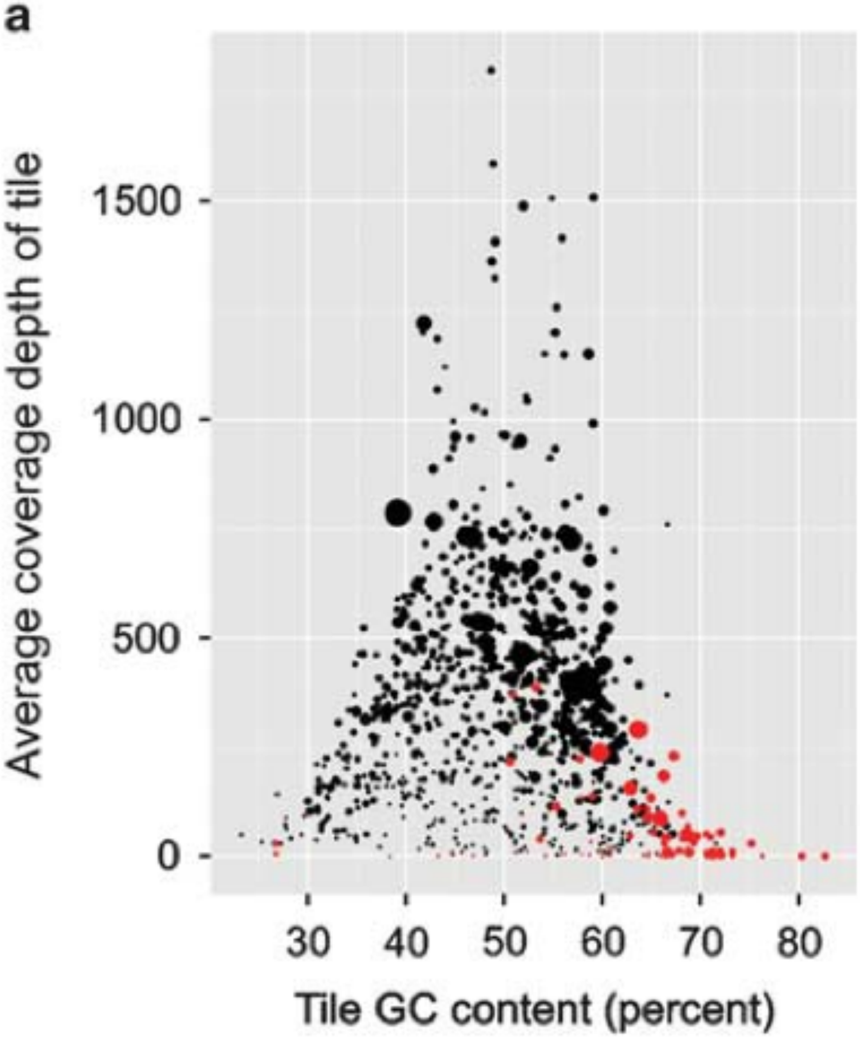


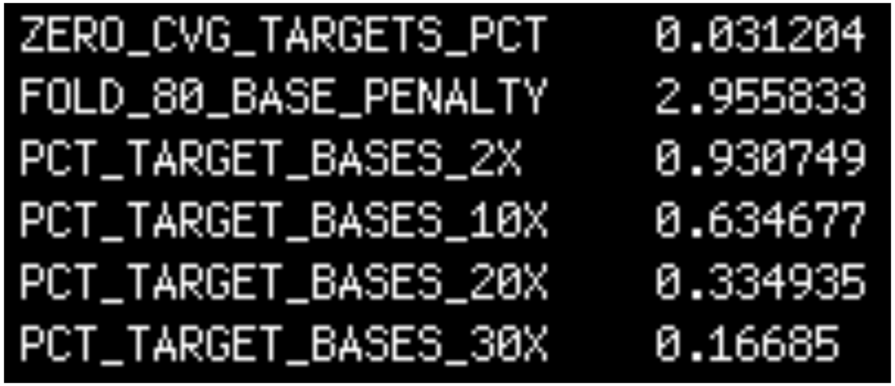
CCDS exon length distribution



```
Exon(x) - the target:           IIIIIXXXXXXXXXXXXXRRRRXXXXXXXXXIIIII
(R=repeat, I=intron)

Nimblegen (avoid repeats and tiles no longer than need to be)
The tiled (T) region (or bait):     TTTTTTTTTTTTTT     TTTTTTTTTTT
The oligos(O) (or probes):          OOOOO              OOOOOO
                                      OOOOO                OOOO
                                        OOOOO                OOOO
                                          OOOOO               OOOO
                                            OOOOO

Agilent (no attempt to avoid repeats and no oligo overlap so tiles are always
multiple of 120 so often sequence well into introns)
The tiled (T) region (or bait):   TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT
The oligos(O) (or probes):          OOOOOOOOOOOOOOOOO|OOOOOOOOOOOOO


There may be more than one tiled region per target
.e.g. if the target contains repeats.
Oligo spacing: distance between start positions of oligos
Oligo overlap: number of overlapping bases between oligos
```

Coverage metrics – Effect of GC content

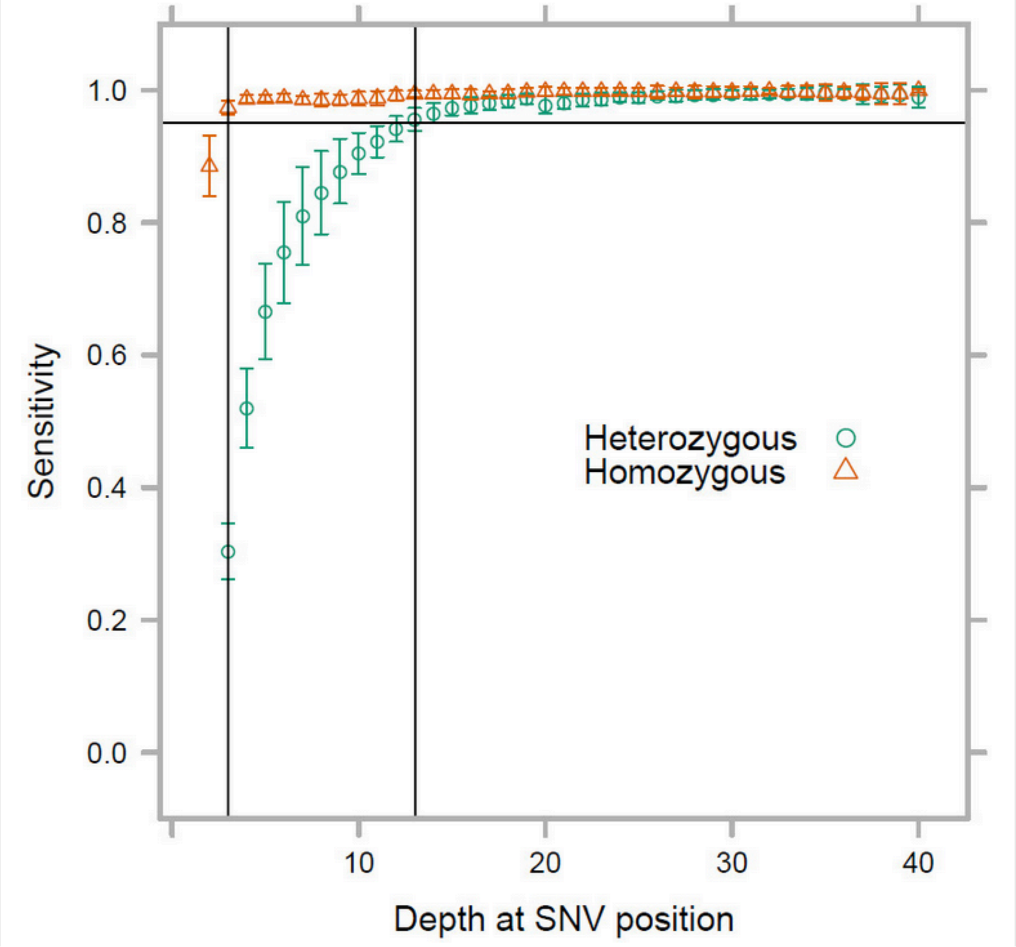




```
ZERO_CVG_TARGETS_PCT     0.031204
FOLD_80_BASE_PENALTY     2.955833
PCT_TARGET_BASES_2X      0.930749
PCT_TARGET_BASES_10X     0.634677
PCT_TARGET_BASES_20X     0.334935
PCT_TARGET_BASES_30X     0.16685
```

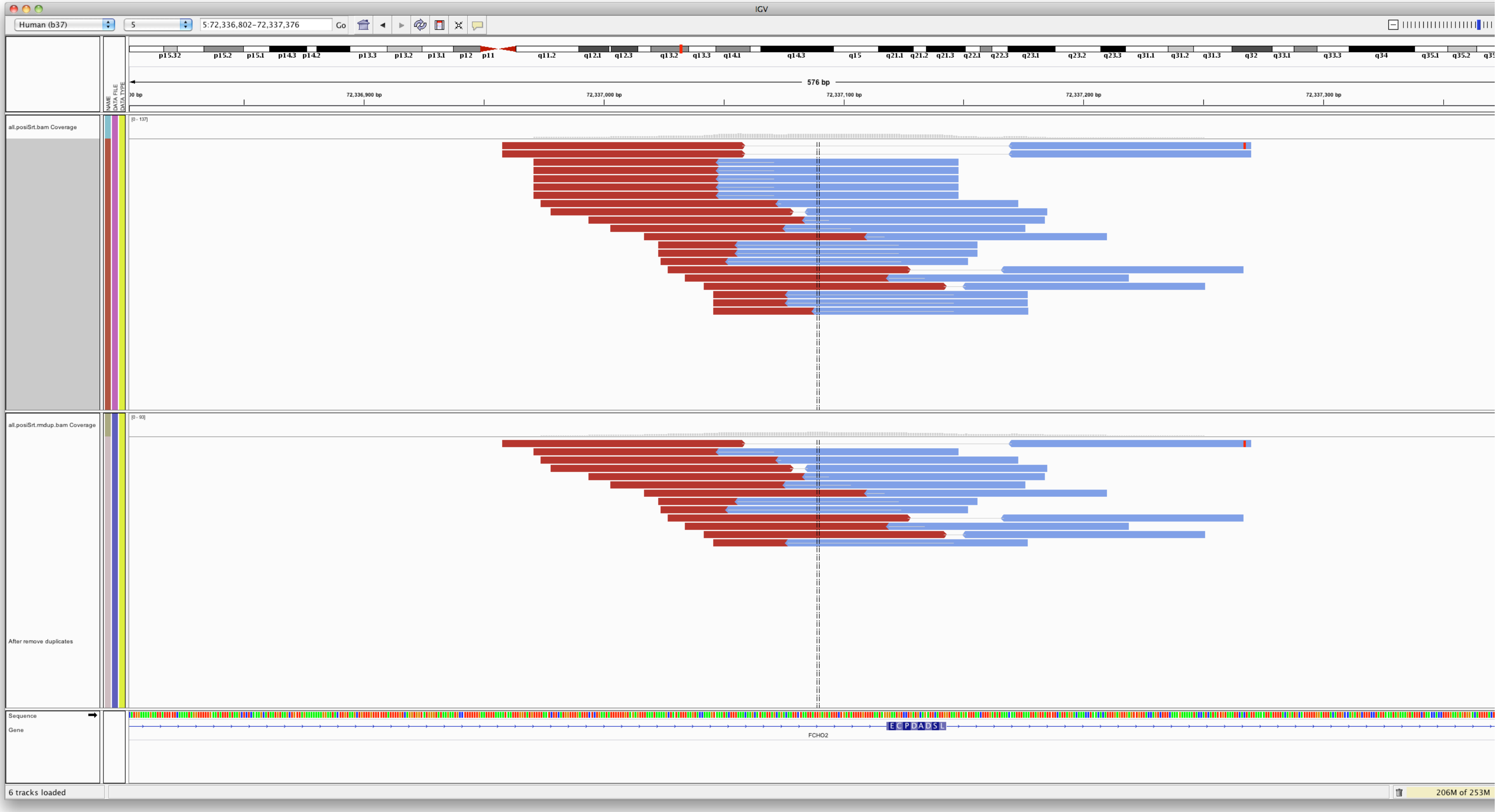
The effect of depth on errors

- Heterozygotes vs homozygote variant sites
- Equal sampling of alleles



From “Quantifying single nucleotide variant detection sensitivity in exome sequencing”
BMC Bioinformatics 2013

Duplicate metrics



Duplicates potentially introduce variant calling errors

021_generatingReports.bash

- Time to see how we can do some of this in practice
- Walk through the fastq and SAM/BAM part of the practical
- Stop when we get to VCF part!
- **Walk through this together with instructor**

APPENDIX

Overview of topics (not in chrono order)

- Software and datasets setup and intro
- Fastq format
- Read mapping (SAM/BAM format)
- IGV
- Variant calling (VCF format)
- Metrics reports (esp coverage – BED format)
- Alignment refinement
- Base quality score recalibration
- Variant annotation and functional filtration

Comparison of different programs (optional)

- Comparison analysis of algorithms for next-generation sequencing read alignment, Ruffalo et al. 2011
  - generate genome in silico with certain indel properties (size and frequency)
  - simulate reads from this genome with certain error rate
  - measure mapper accuracy: the true location of a read is considered the truth and a mapping is considered a prediction
- What to look out for on the next slide
  - Ignore the Fast tools as they are poor mappers (green in legend)
  - Only focus on the parts of the plots that cover realistic range of parameters
- In summary, BWA and Novoalign are both good mapping tools

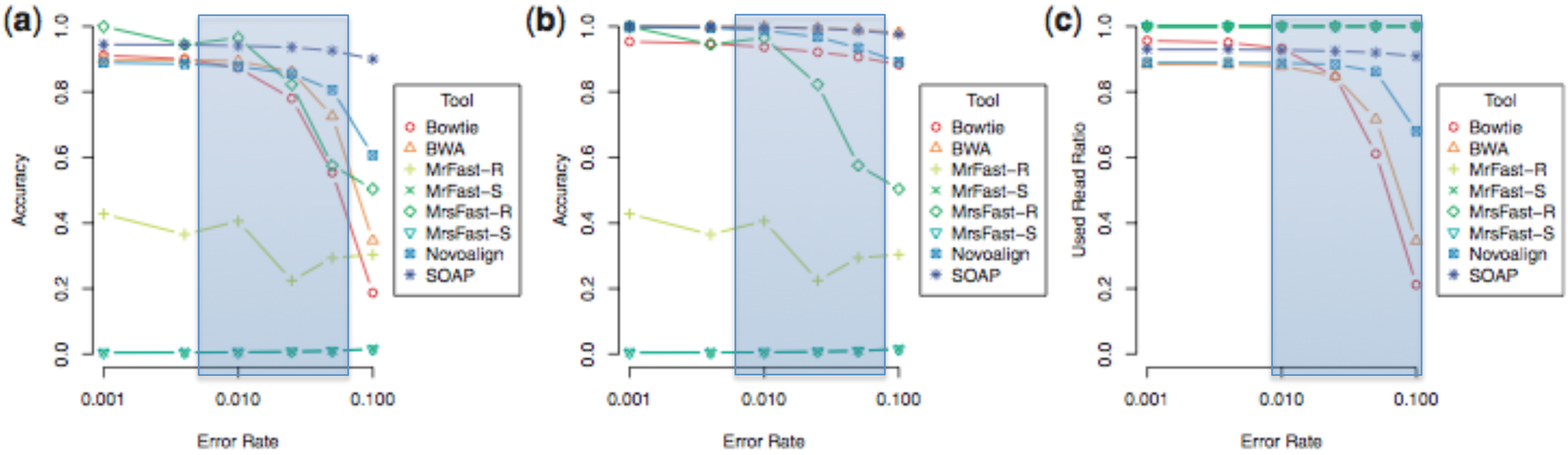
Accuracy with varying error rate

With mapping quality threshold 10

Accuracy

Accuracy

Used read ratio



- No indels
- Higher range of error rate is unrealistic
- Minor differences between good algorithms in range

Source: Ruffalo et al. 2011

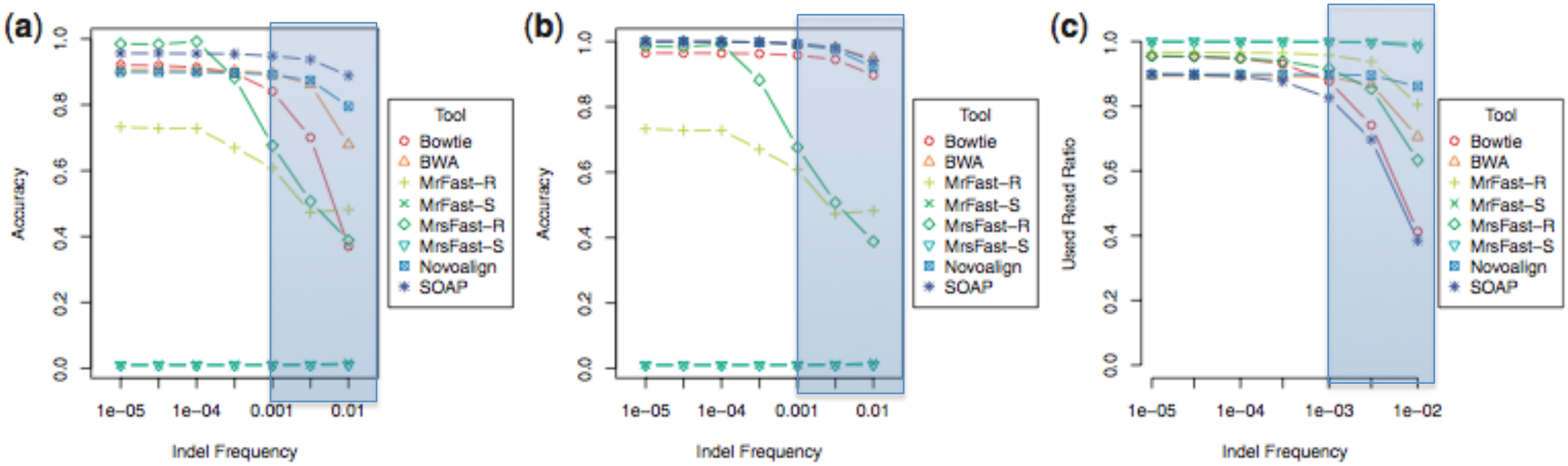
Accuracy with varying indel frequency

With mapping quality threshold 10

Accuracy

Accuracy

Used read ratio



- indel size fixed at 2 ➔ small
- error rate fixed at 0.01 ➔ somewhat high
- only lower indel frequencies are realistic ➔ small differences between programs in relevant range
- soap would have looked a lot worse if indel size had been bigger

Source: Ruffalo et al. 2011

Gigsaw from University of Dundee

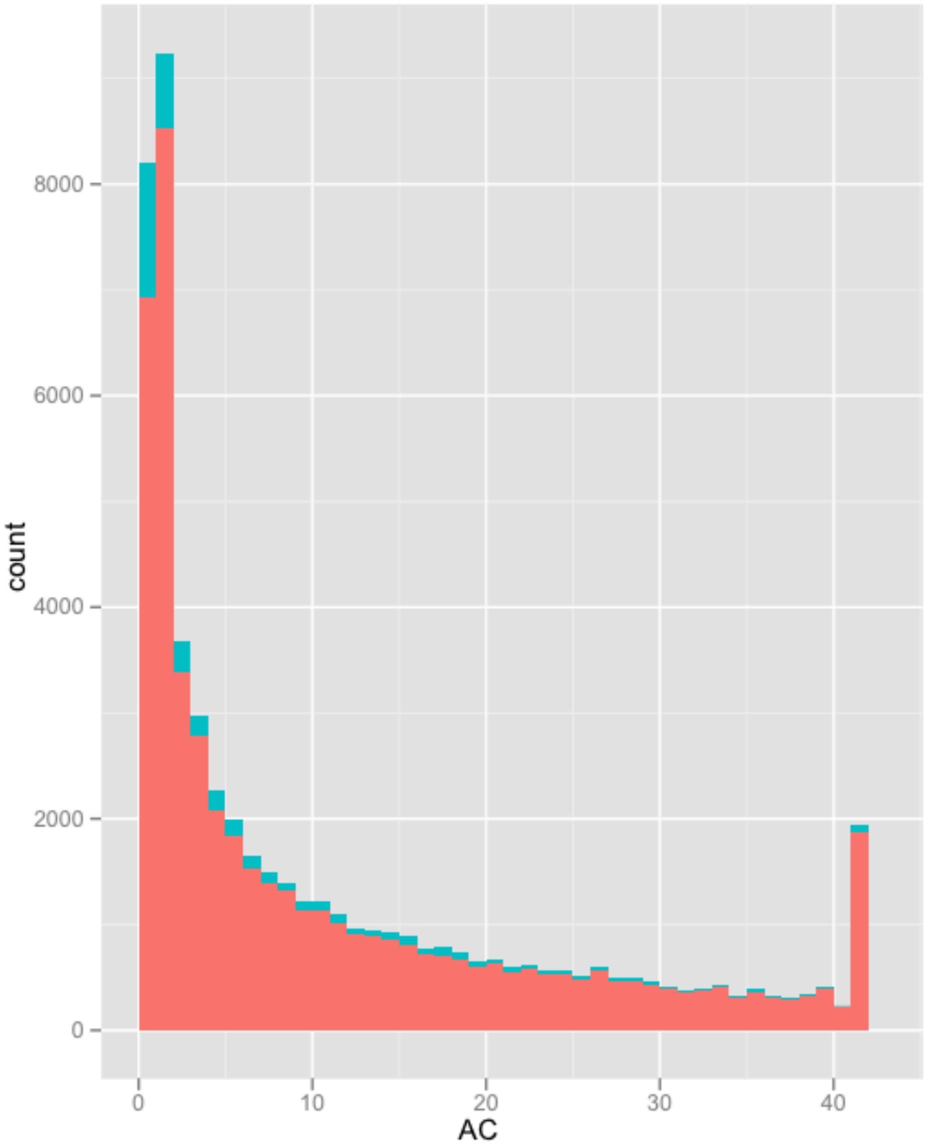
- Four people in each group
- Set up the reference sequence: both forward and reverse (reverse complement)
- Each person takes 10 reads. **Notice that reads are reverse complemented on the back. So use the side that is most convenient.**
- Map individually in parallel: two persons can map to forward and two to reverse.
- Then combine the reads to variant call together.
- In 10 mins, I want to know what you have found out

Gigsaw created by David Martin, University of Dundee, Scotland

Manipulating fasta and fastq files

- Fastx toolkit: http://hannonlab.cshl.edu/fastx_toolkit/
- FASTQ trimmer
- FASTQ quality filter
- FASTQ quality trimmer
- Can do most of the obvious manipulations of fastq/a you may need

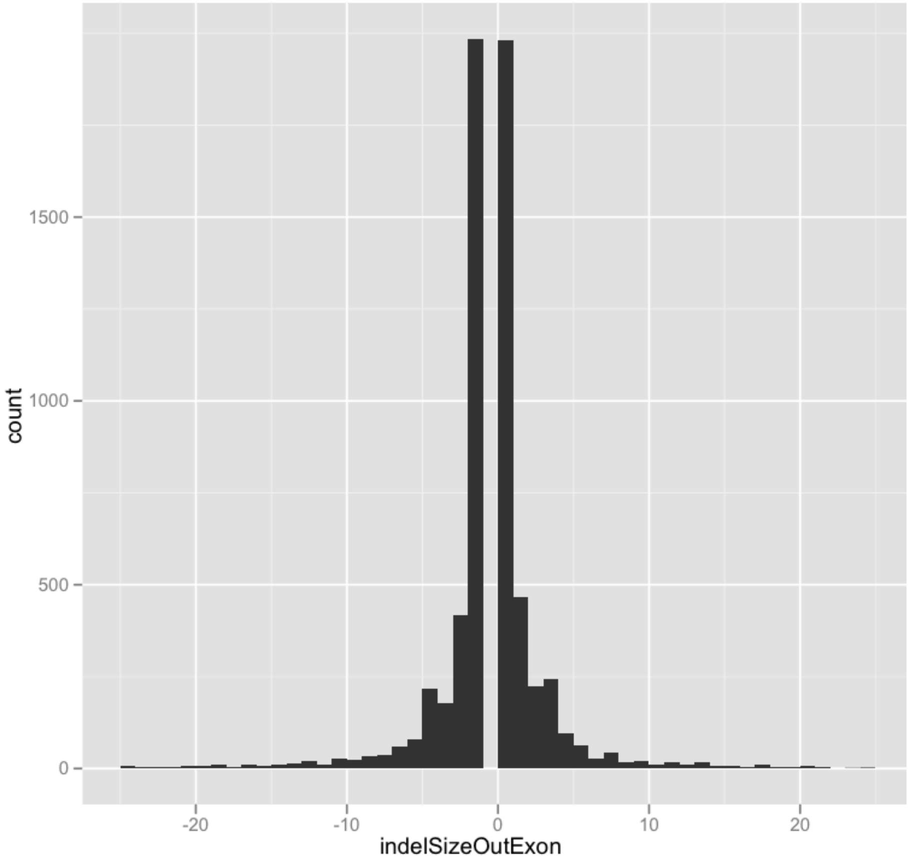
Distribution of Allele Count across 21 exomes

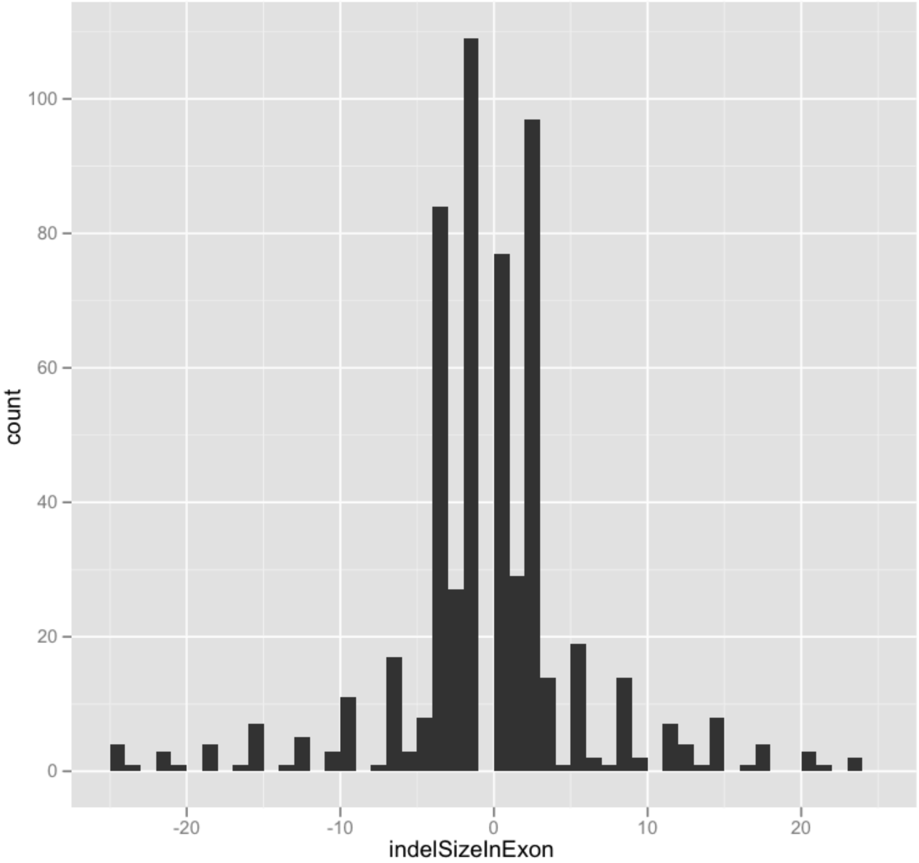


21 individual exomes (of diploid humans) i.e. 42 alleles

SNP numbers and indel size distribution

- Sequencing of human exome
  - 23,602 SNPs in coding exons (approx. 25M bp size)
  - 40,621 SNPs outside coding exons (approx. 25M bp size)





Notice both numbers and pattern in indel figures