Variant Calling
(using High-throughput Sequencing Data)

November 2012

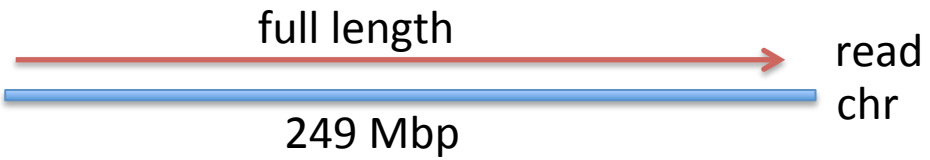
Tim Hughes

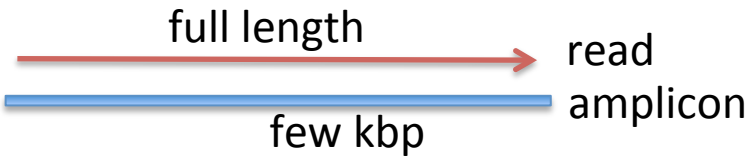
NORWEGIAN SEQUENCING CENTRE

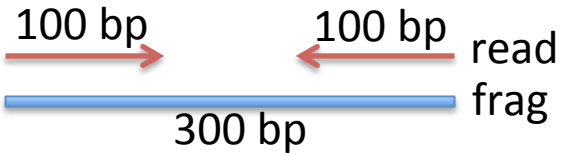
INTRODUCTION

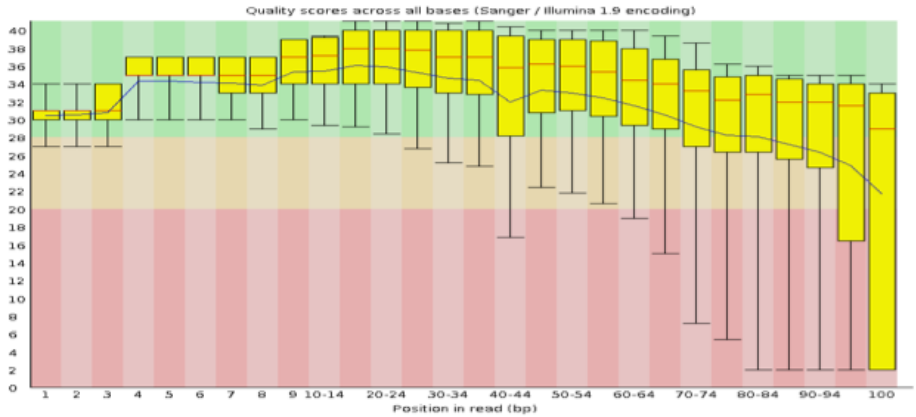
In a perfect world – Perfect sequencing

- **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







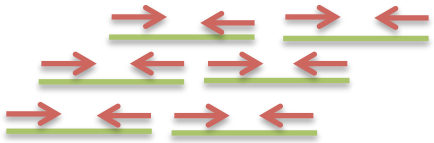


A quick overview of the HTS workflow

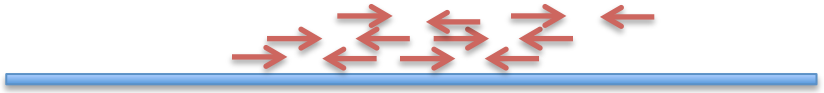
Fragment sample

Capture

Sequence

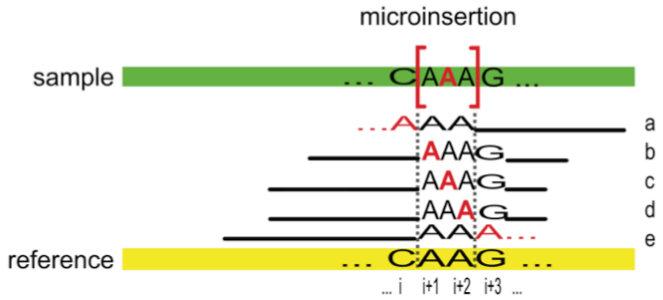


Map



Align

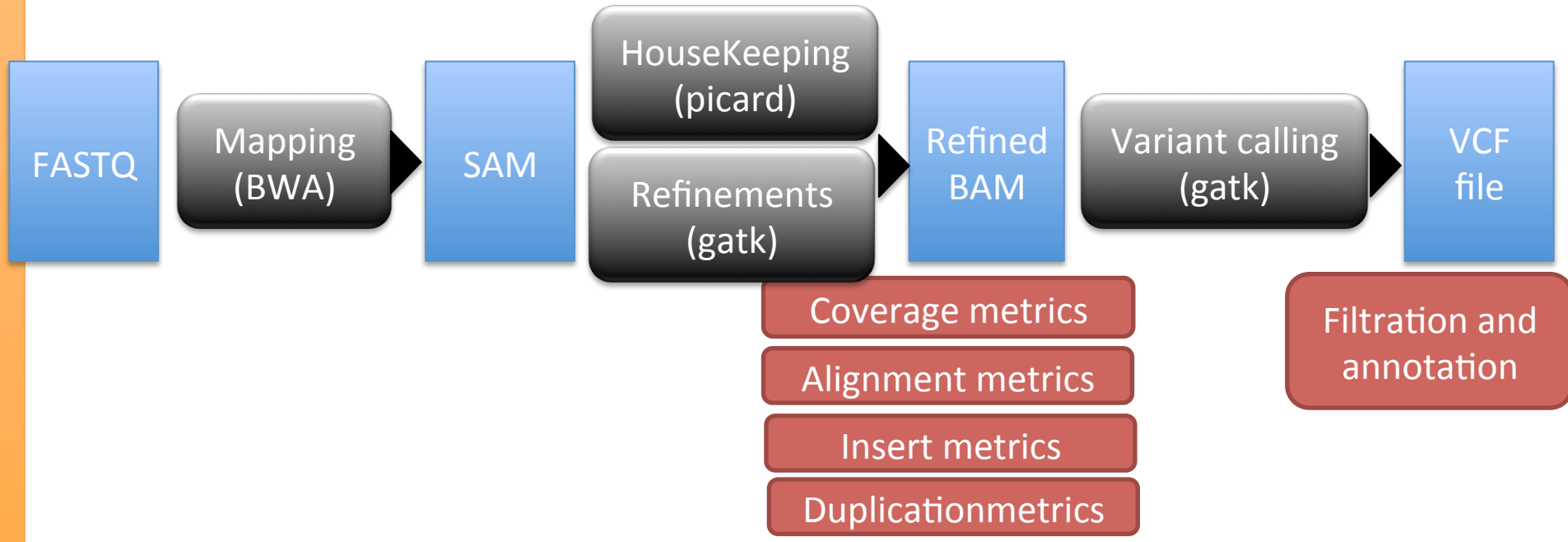
Variant call



Sample mutation

Poor alignment >> FN micro indel + FP SNP

In a bit more detail



GSA team at the Broad Institute

- A large fraction of the materials and software in this course are produced by the **Genome Sequencing and Analysis Group** team at the Broad Institute
- Information sources
  - http://www.broadinstitute.org/gsa/wiki
  - http://www.getsatisfaction.com/gsa
- People
  - Mark A. DePristo, Manager of Medical and Population Genetics Analysis
  - Eric Banks, Team Lead
  - Guillermo del Angel
  - Ryan Poplin
  - Kiran Garimella, Team Lead
  - Mauricio Carneiro
  - Chris Hartl
  - Khalid Shakir, Team Lead
  - Matthew Hanna
  - David Roazen
- Others at the Broad
  - Heng Li: samtools and bwa
  - Tim Fennell: picard
  - Alec Wysoker: picard
- And others outside the Broad
  - sources at bottom of slides

Overview of topics (not in chrono order)

- Software and datasets 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 filtration

DATASETS

Introduction of dataset

- reads_exomeCapt_chr5 in fastq format (**reads_agilentV1_chr5**)
  - real reads from exome capture (**real**)
  - simulated: known mutations and simulated reads (**simul**) – same regions as real dataset

- reference data (**human_g1k_v37_chr5**)
  - **agilentV1** >> definition of capture tiles in different formats
  - **gatkBundle** >> reference data in fasta format and vcf files of known variants (dbSNP, 1000 genomes, hapmap)

- Formats >> we will return to these later

Naming and ordering of chromosome/contigs

	Hg18 (UCSC)	B36 (NCBI)
Contig prefix	chr	none
Mitochondrial contig	chrM	MT
Contig order	chrM, chr1, chr2, ..., chrX, chrY	1, 2, ..., X, Y, MT

- Genome references
  - Fasta file: must have .fasta extension + respect naming and order
  - Fai file (created by samtools faidx): contig, size, location, basesPerLine ➔ for efficient random access
  - Dict file (created by Picard CreateSequenceDictionary): SAM style header describing the contents of the fasta file ➔ for names and length of original file
- ROD (reference ordered data)
  - GATK supports several common file formats for reading ROD data: VCF, UCSC formatted dbSNP, BED
- dbSNP files
  - Must also be ROD
  - Generated by GSA from the dbSNP db using a bit of bash, awk and a perl script: sortByRef.pl. Full details: http://www.broadinstitute.org/gsa/wiki/index.php/The_DBSNP_rod
- All of the above delivered for human as part of the GATK resource bundle
  - Other species may also be available
  - Help on generating for another species see GATK wiki or getsatisfaction.com/gsa

GENETICS 101

Any questions?

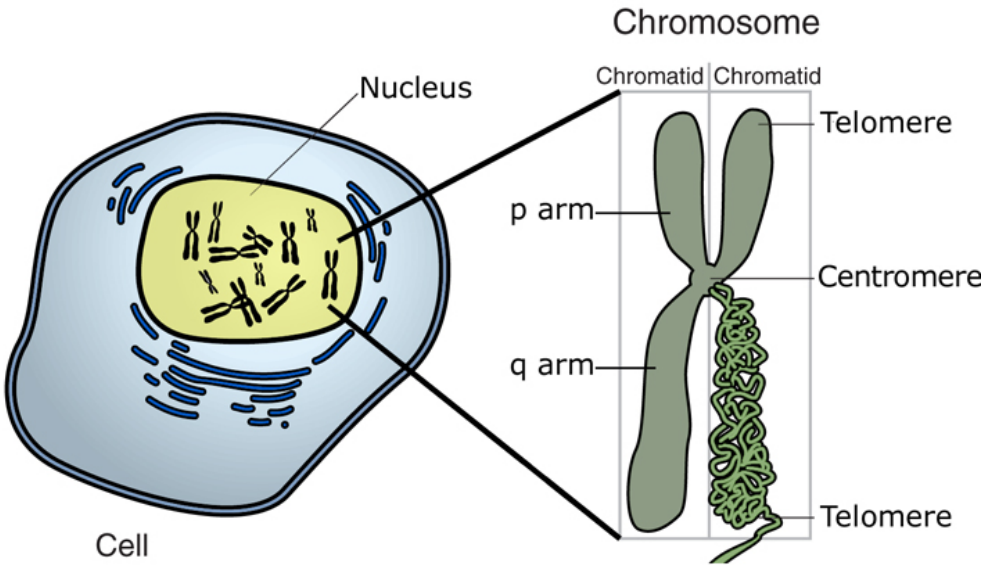
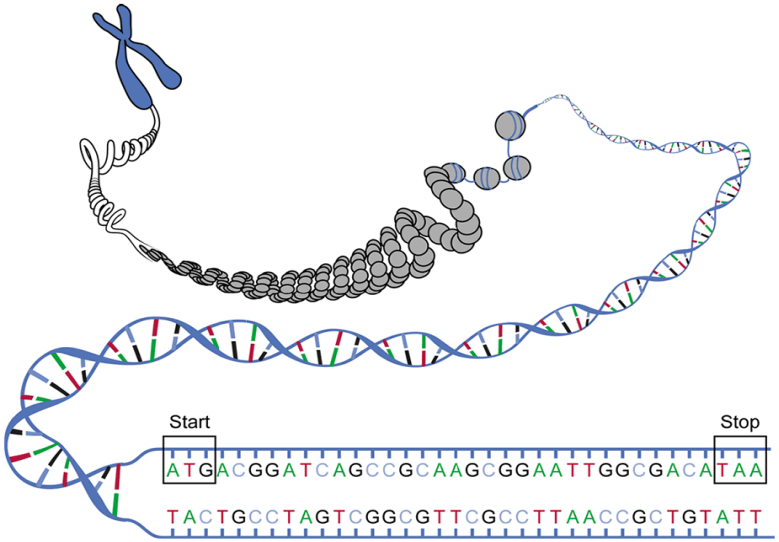
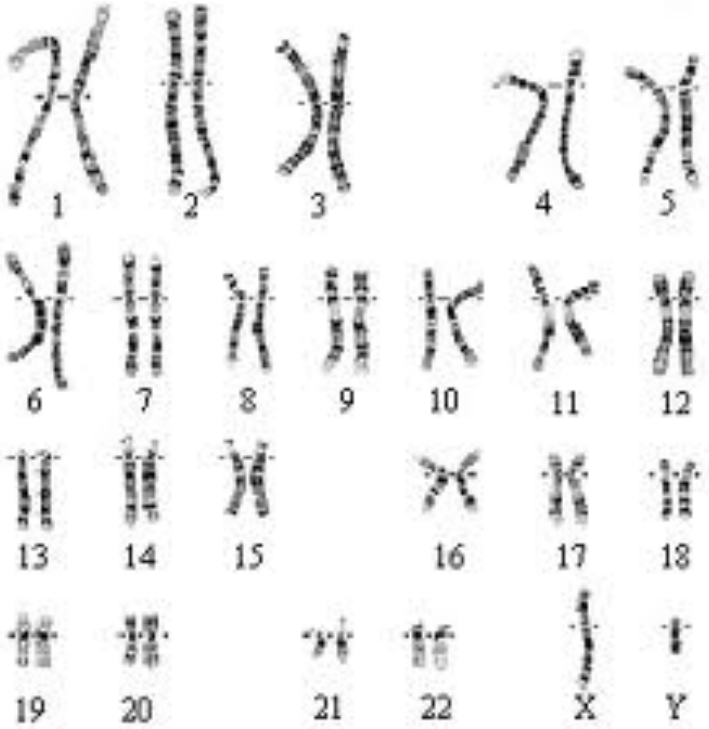


Image adapted from: National Human Genome Research Institute.



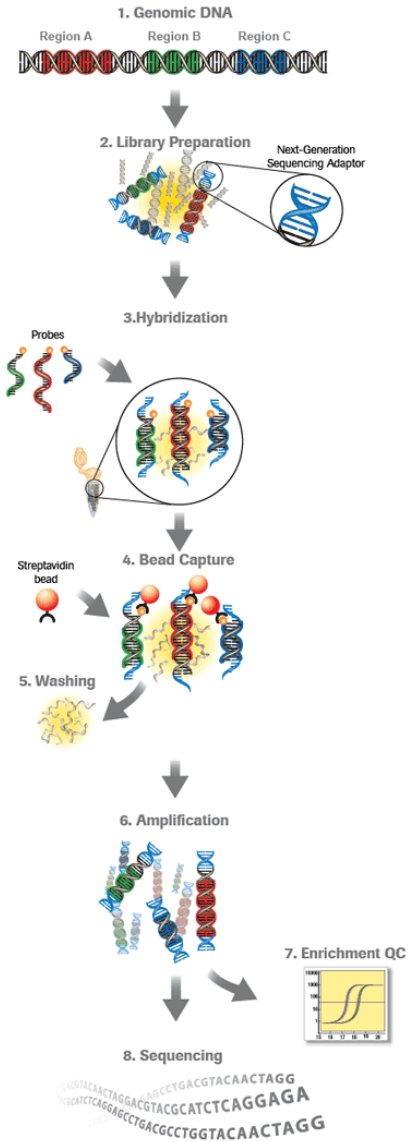


- Cells
- chromosomes
- homo, hetero

NORWEGIAN SEQUENCING CENTRE

EXOME CAPTURE – ESSENTIALS

An overview of exome capture



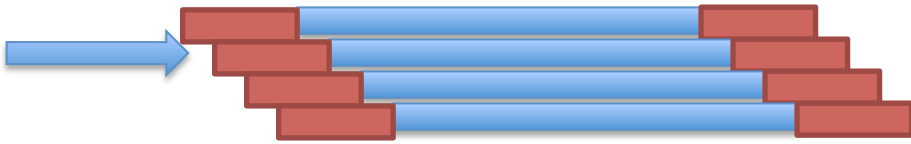
Sonication

Library prep (sequencing adaptors on)

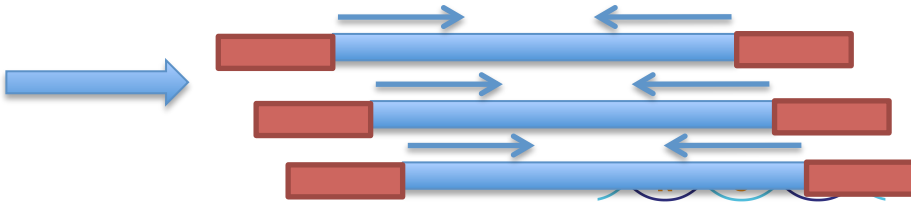
Hybridisation to probes

Bead capture

Amplification



Sequencing



SEQUENCING – ESSENTIALS

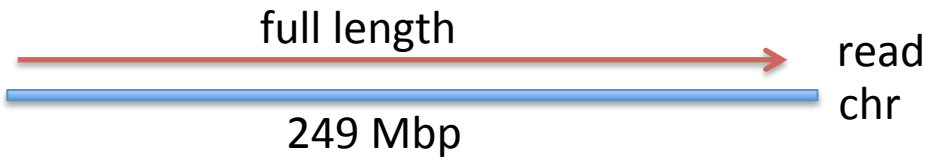
Sequencing

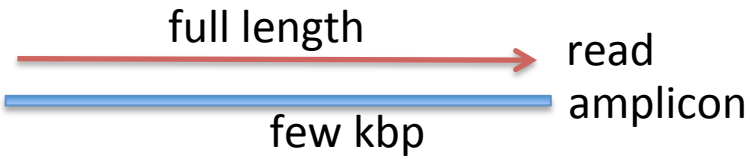
Covered by Robert

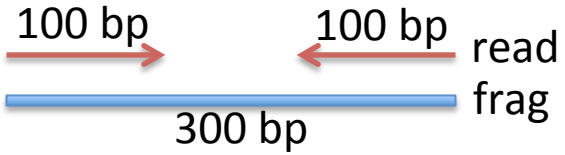
FASTQ FORMAT – ESSENTIALS

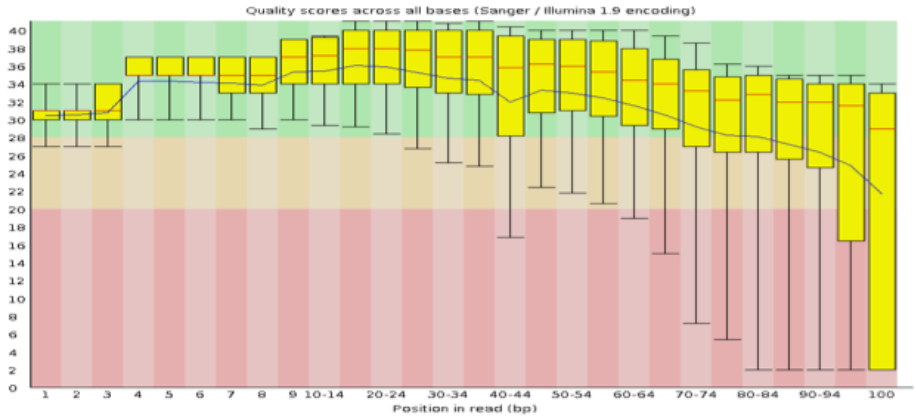
In a perfect world – Perfect sequencing

- 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









Fastq format – fasta with qualities

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((****+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

- p = the probability that the corresponding base call is wrong
- Qualities $Q_{\text{sanger}} = -10 \log_{10} p$
  - p = 0.1 ➔ Q = 10
  - p = 0.01 ➔ Q = 20
  - P = 0.001 ➔ Q = 30
- Encoding: Sanger/Phred format can encode a quality score from 0 to 93 using ASCII 33 to 126: Q + 33 ➔ ASCII code

Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr
32	20	040		Space	64	40	100	@	@
33	21	041	!	!	65	41	101	A	A
34	22	042	"	"	66	42	102	B	B
35	23	043	#	#	67	43	103	C	C
36	24	044	$	$	68	44	104	D	D
37	25	045	%	%	69	45	105	E	E
38	26	046	&	&	70	46	106	F	F
39	27	047	'	'	71	47	107	G	G
40	28	050	(	(	72	48	110	H	H
41	29	051	)	)	73	49	111	I	I
42	2A	052	*	*	74	4A	112	J	J
43	2B	053	+	+	75	4B	113	K	K
44	2C	054	,	,	76	4C	114	L	L
45	2D	055	-	-	77	4D	115	M	M
46	2E	056	.	.	78	4E	116	N	N
47	2F	057	/	/	79	4F	117	O	O
48	30	060	0	0	80	50	120	P	P
49	31	061	1	1	81	51	121	Q	Q
50	32	062	2	2	82	52	122	R	R
51	33	063	3	3	83	53	123	S	S
52	34	064	4	4	84	54	124	T	T
53	35	065	5	5	85	55	125	U	U
54	36	066	6	6	86	56	126	V	V
55	37	067	7	7	87	57	127	W	W
56	38	070	8	8	88	58	130	X	X
57	39	071	9	9	89	59	131	Y	Y
58	3A	072	:	:	90	5A	132	Z	Z
59	3B	073	;	;	91	5B	133	[	[
60	3C	074	<	<	92	5C	134	\	\
61	3D	075	=	=	93	5D	135	]	]
62	3E	076	>	>	94	5E	136	^	^
63	3F	077	?	?	95	5F	137	_	_

Source: http://en.wikipedia.org/wiki/FASTQ_format

Different quality formulas and encodings

- Beware of different versions

```
  SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
  ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
  ..............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII..............................
  ...............................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ..............................
  !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
  |                          |    |        |                              |                     |
 33                        59   64       73                            104                   126

 S - Sanger, Illumina 1.8+     Phred+33,  raw reads typically (0, 41)
 X - Solexa                    Solexa+64, raw reads typically (-5, 40)
 I - Illumina 1.3+             Phred+64,  raw reads typically (0, 40)
 J - Illumina 1.5+             Phred+64,  raw reads typically (3, 40)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
```

- With sequence data recently produced, you do not need to worry
- For older data, be careful
- Tools such as BWA have an option for older formats

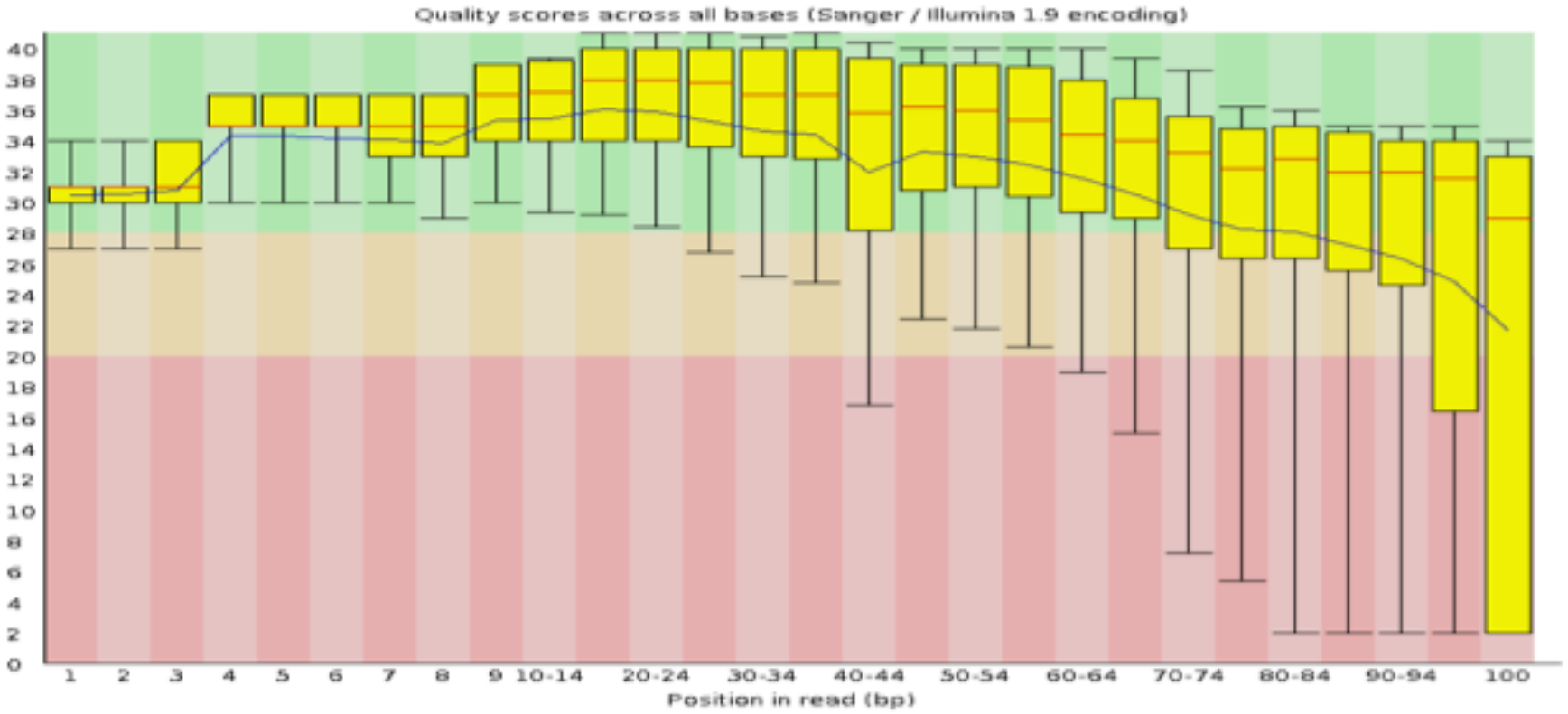
Illumina sequence identifiers

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
```

EAS139	the unique instrument name
136	the run id
FC706VJ	the flowcell id
2	flowcell lane
2104	tile number within the flowcell lane
15343	'x'-coordinate of the cluster within the tile
197393	'y'-coordinate of the cluster within the tile
1	the member of a pair, 1 or 2 *(paired-end or mate-pair reads only)*
Y	Y if the read is filtered, N otherwise
18	0 when none of the control bits are on, otherwise it is an even number
ATCACG	index sequence

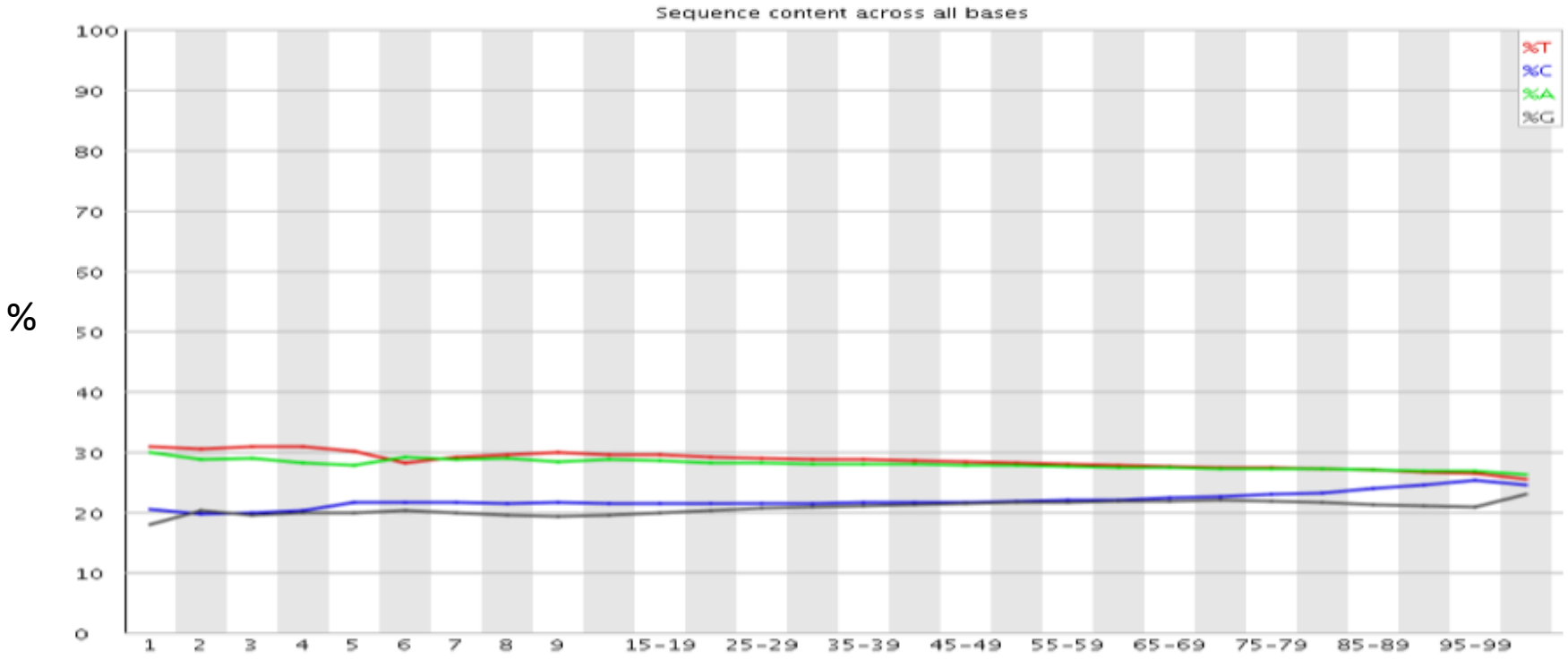
FastQC - Per cycle quality distribution



Position in read

FastQC - Per cycle sequence content

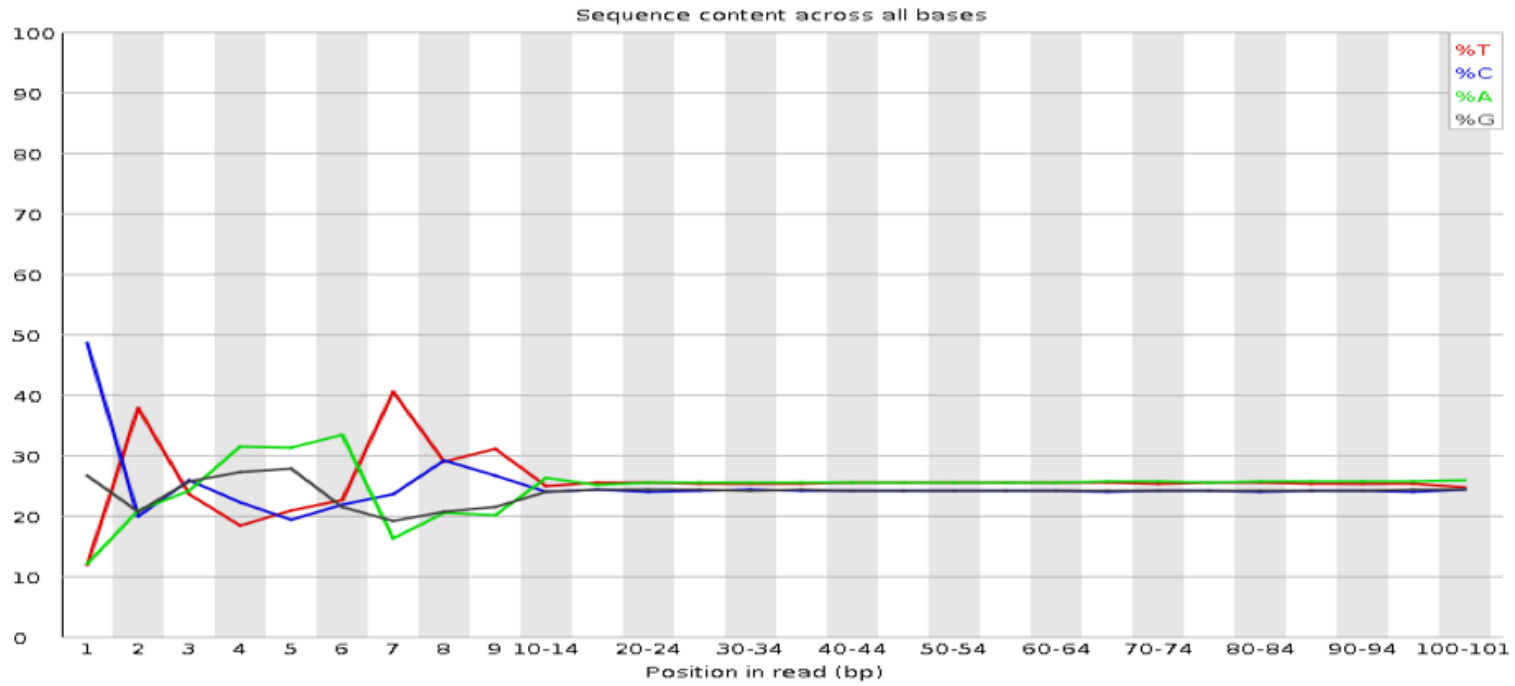
Exome sequencing



NORWEGIAN SEQUENCING CENTRE

FastQC - Per cycle sequence content

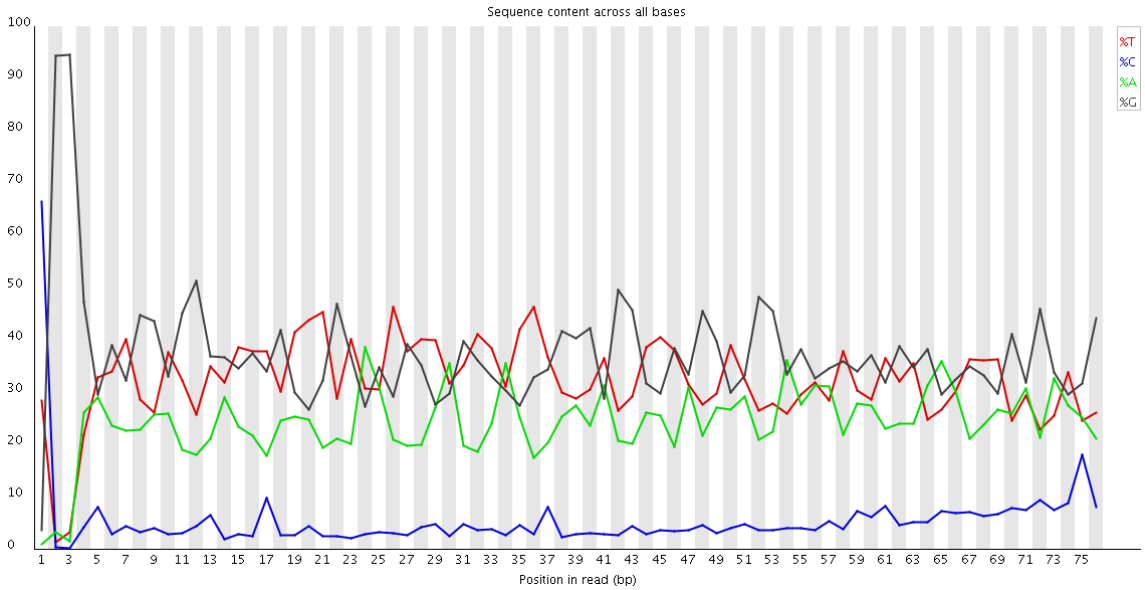
mRNA sequencing



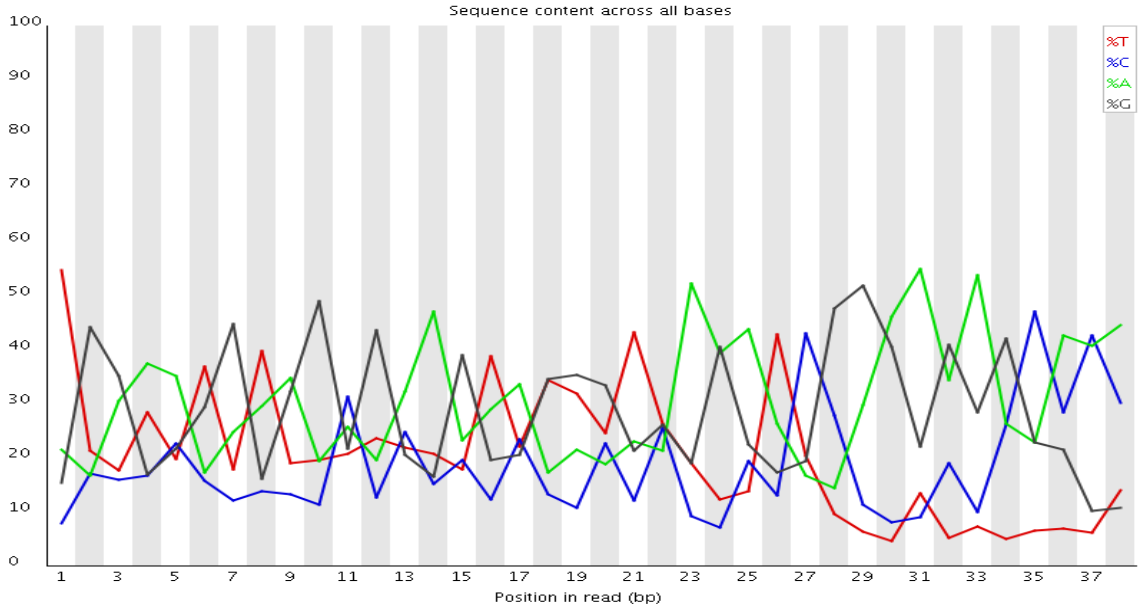
Position in read

NORWEGIAN SEQUENCING CENTRE

FastQC - Per cycle sequence content



?



?

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

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

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

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)

- All programs do not necessarily produce good estimates of mapping quality

- 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
- Theoretically allows for differences and guarantees to find all intervals with x differences
- But, in practice, makes changes to algorithm to adapt to biological reality and increase speed
  - different penalties for mismatches, gap open and extension
  - uses a seed approach: no more than x differences in the first 32bp of a read to increase speed
  - the alignment error rate (fraction of wrong alignments out of confident mappings in simulation) only marginally increases, but substantially improves speed
  - implementation modifications to speed up computation time
- 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)

Mapping errors

Base stacks

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
read3          ttataaaacAAATaattaagtctaca
read4              CaaaT    aattaagtctacagagcaac
read5                aaT    aattaagtctacagagcaact
read6                 T     aattaagtctacagagcaacta
read1     aggttttataaaacaaataa
read2      ggttttataaaacaaataatt
read3          ttataaaacaaataattaagtctaca
read4                caaataattaagtctacagagcaac
read5                  aataattaagtctacagagcaact
read6                   taattaagtctacagagcaacta
```

Incorrect

Correct

>> Can be solved by alignment: considering all mapping reads and reference together

Source: Heng Li, presentation at GSA workshop 2011

NORWEGIAN SEQUENCING CENTRE


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
.........
```

Header

Data lines
(one per read)

```
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:465653C11
PCUS-319-EAS487_0001:7:1:1002:1144#0  pPr1  5   141125710  60  76M   =  141125614  -172  GGGACATGACACACGGGGGCCACTCAGGTGGGAAGAATGAGA  =?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:1765G52
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  7M168M =  131794651  178   GACAGCACAGAGCAGGGCTCACATCCCCACCCACCAGCTGGC  ?;>>87>:>@??;?93>9;97:8>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<@AA859<@@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==???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:58C5T16OA2C4
```

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

NORWEGIAN SEQUENCING CENTRE

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
```

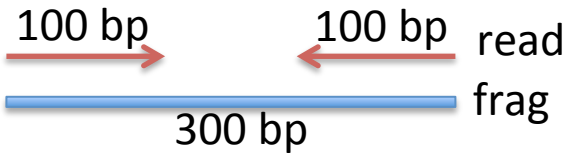
Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,255}`	Query template NAME
2	FLAG	Int	$[0,2^{16}-1]$	bitwise FLAG
3	RNAME	String	`\*|[!-()+-<>-~][!-~]*`	Reference sequence NAME
4	POS	Int	$[0,2^{29}-1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0,2^{8}-1]$	MAPping Quality
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

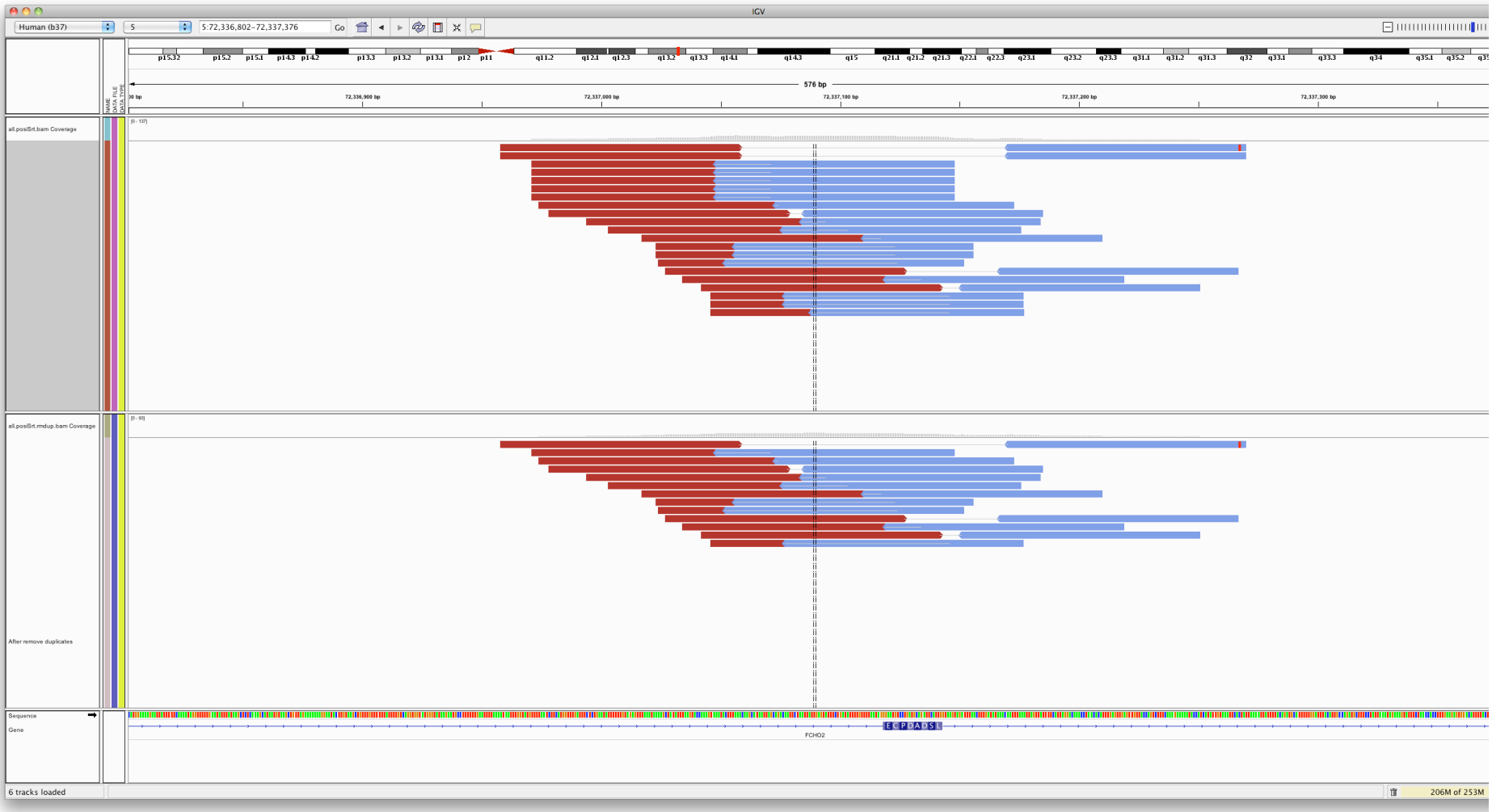
The FLAG column – a bit wise flag




```
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**

What is a duplicate?



NORWEGIAN SEQUENCING CENTRE

About the SAM file produced by BWA

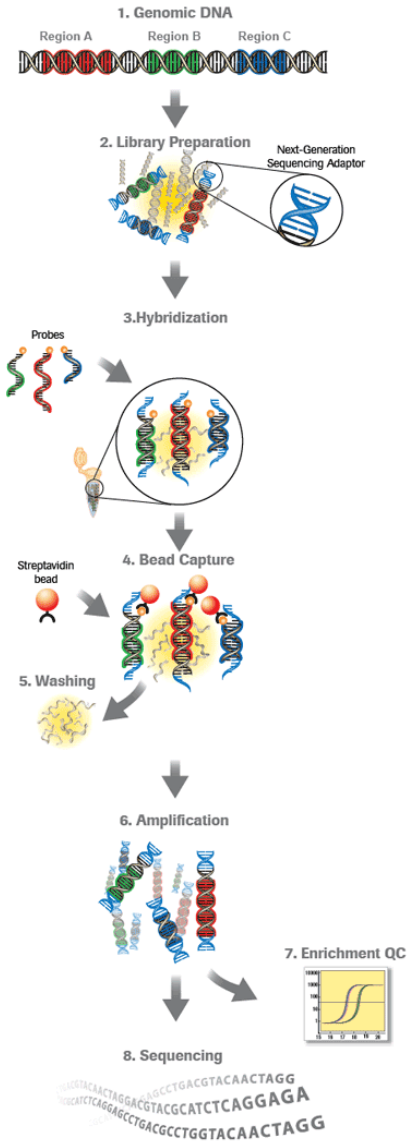
- It contains **all** the reads >> the Picard/GATK paradigm: information is annotated (and not filtered)
  - unique
  - ambiguous
  - unmapped
- It has a number of short comings
  - it takes a lot of space ➔ convert to BAM
  - the mates are not fully updated on each others existence ➔ fixmate
  - it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

IGV practical on a basic BAM file

PRACTICAL

COMPUTING ADVANCED METRICS – PICARD

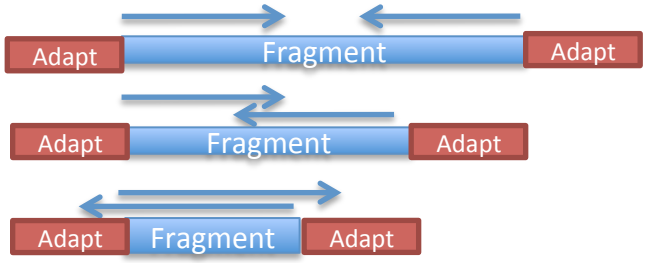
An overview of exome capture – weak points



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**

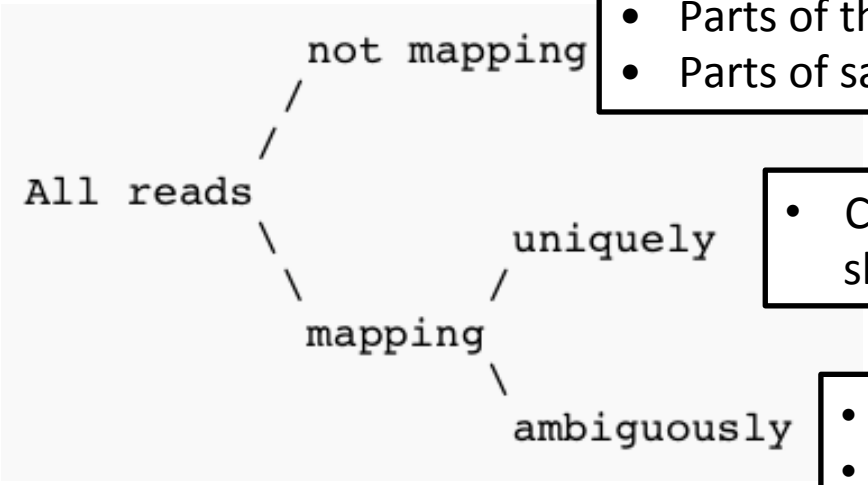
Bead capture → Possible biases in sequences that elute >> **coverage bias**

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

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

NORWEGIAN SEQUENCING CENTRE

Metrics - Basic read classification

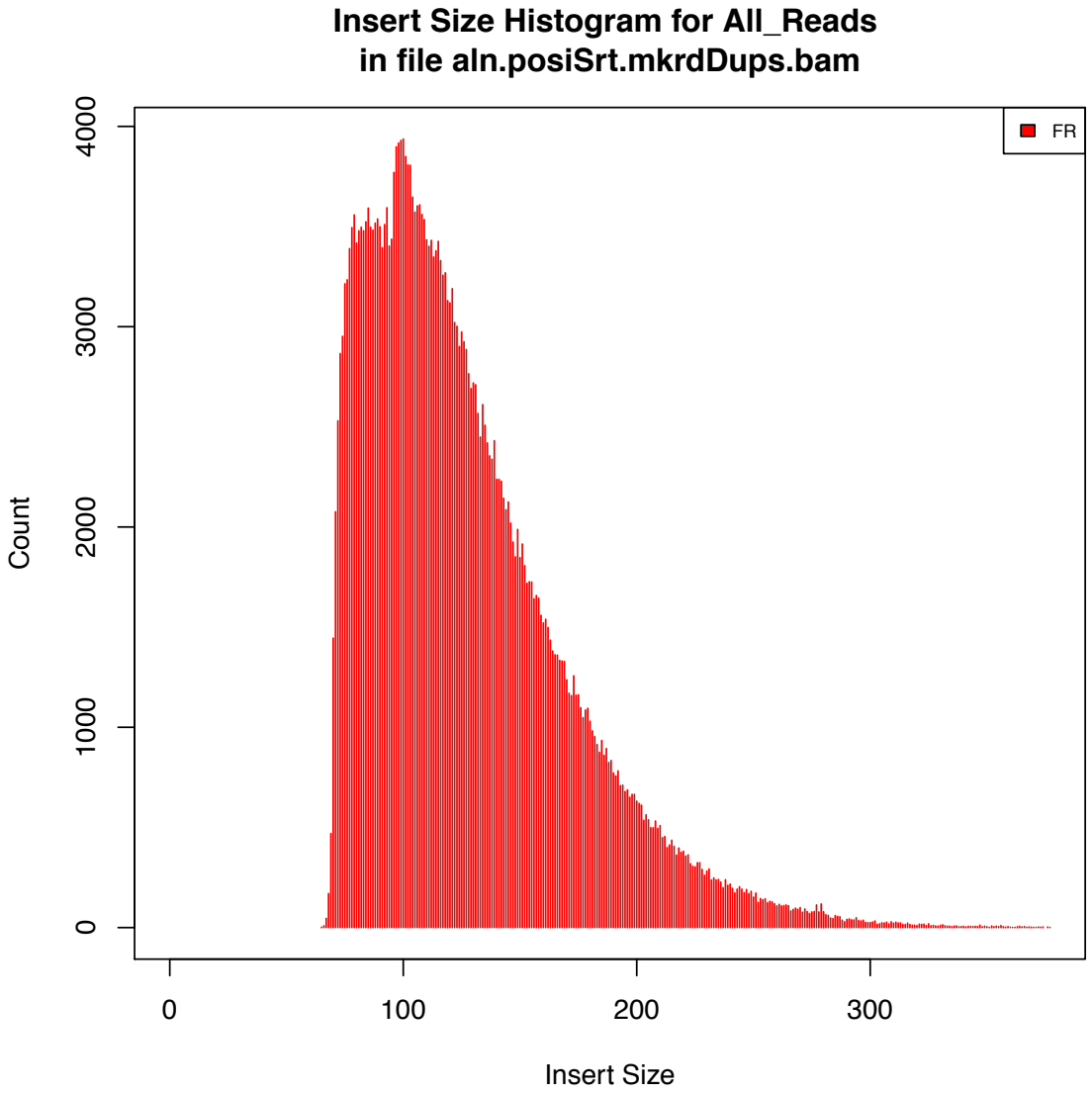


- Contamination from other species
- Reads containing non-genomic DNA e.g. adaptors
- PCR gunk
- Reads with sequencing errors
- Parts of the genome that are not assembled
- Parts of sample affected by structural variation

- Complex sequence with sufficient length should map uniquely

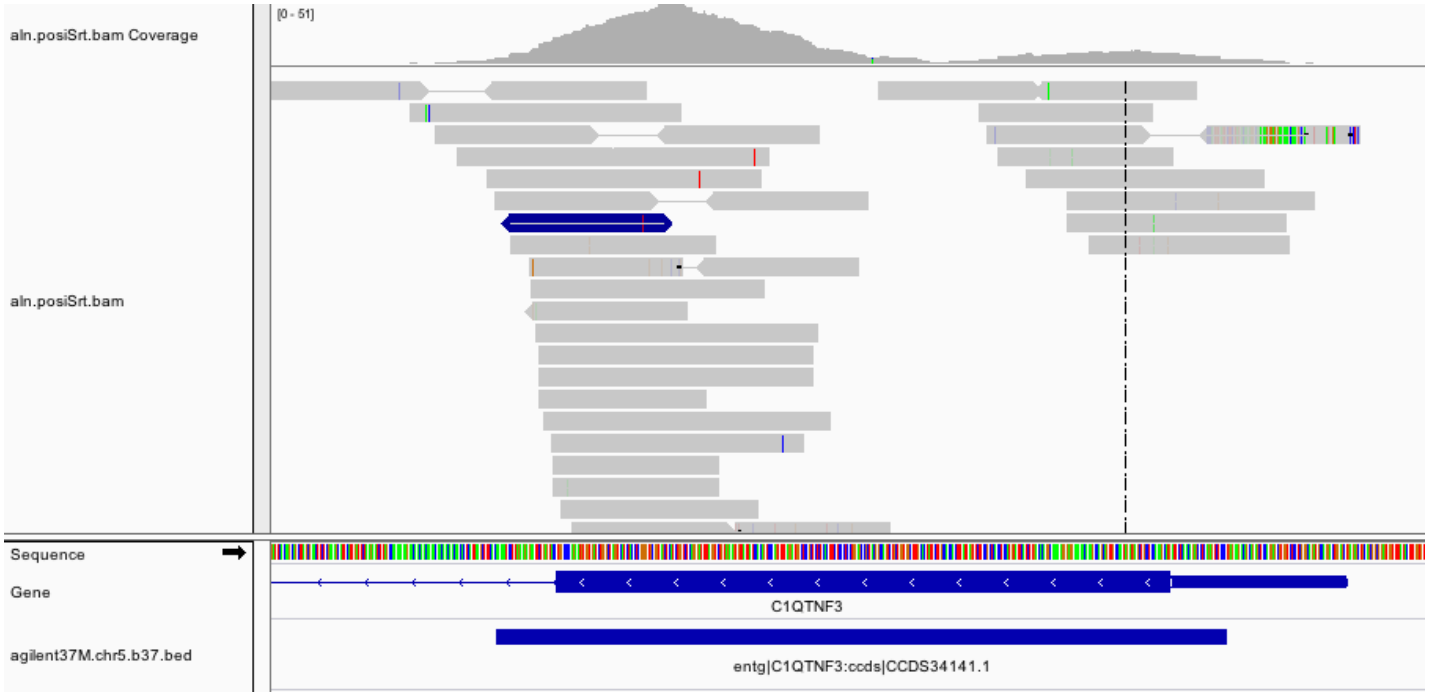
- Typical for repeats
- Also possible for homologous regions

Metrics – Insert sizes

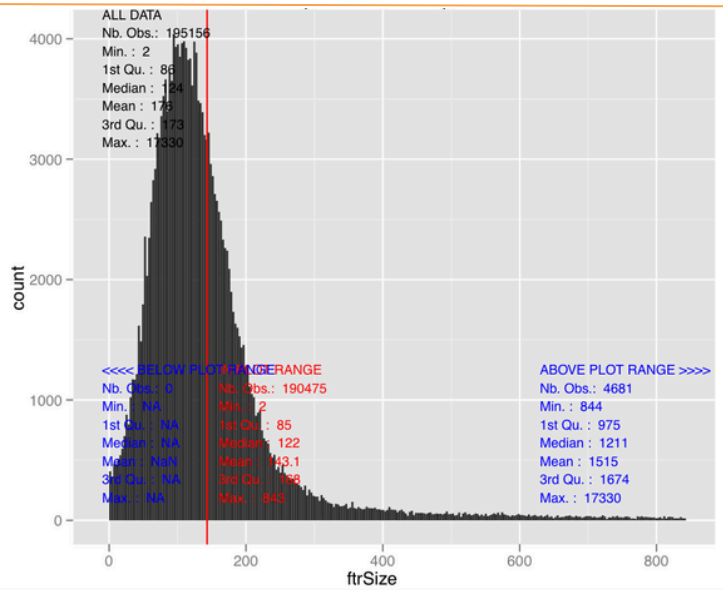


Metrics - Coverage

- 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



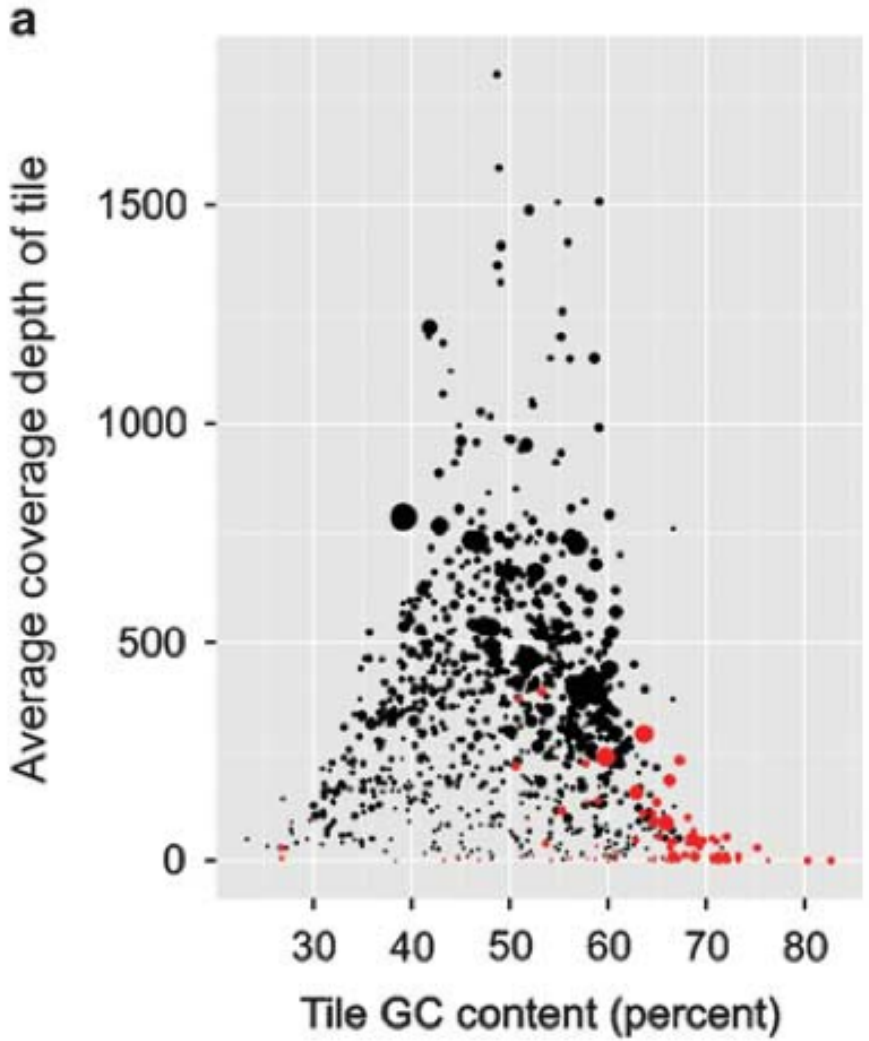
Exome capture – The nature of the target



CCDS exon length distribution

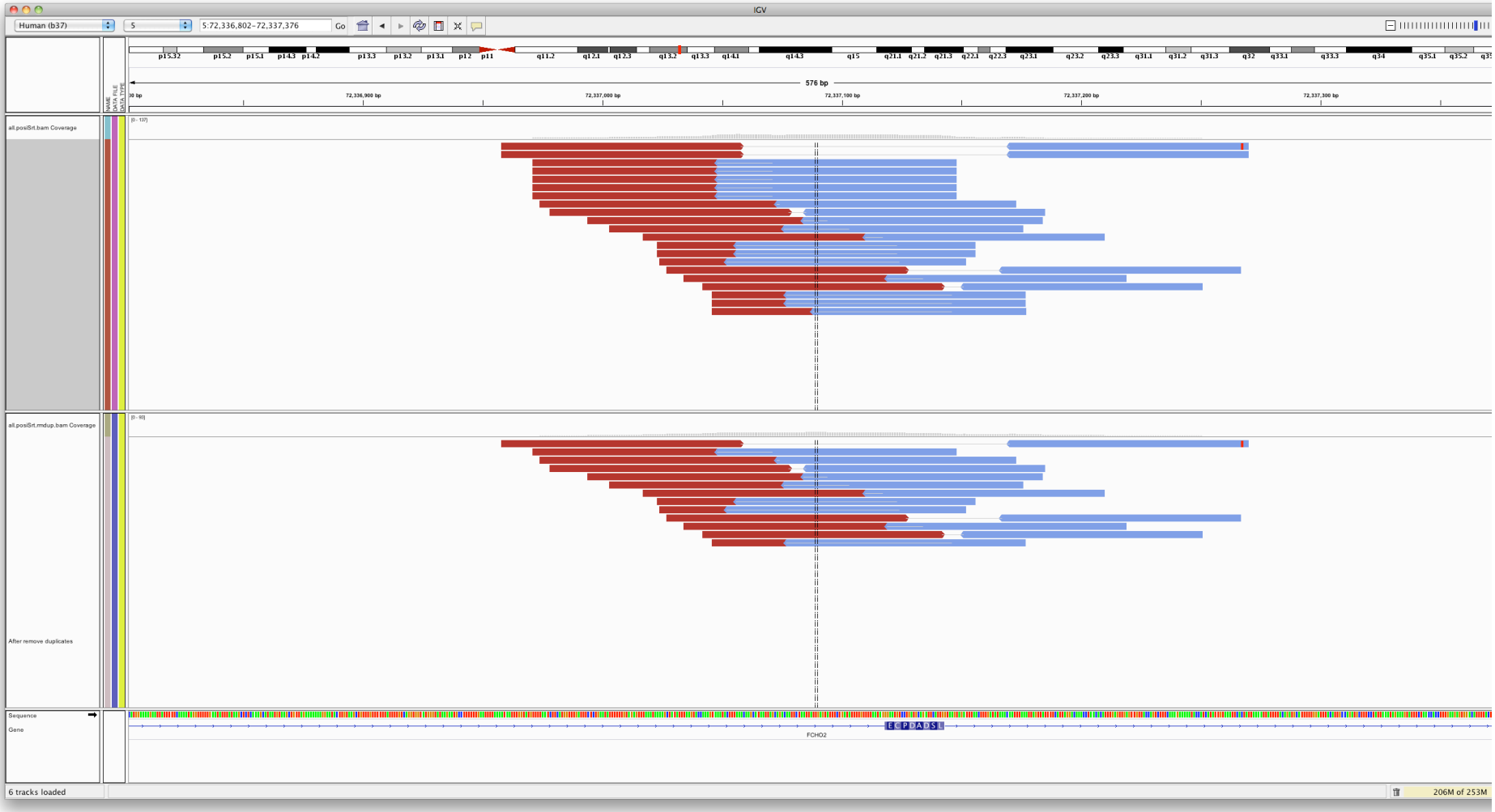
```
Exon(x) - the target:              IIIIIXXXXXXXXXXXXXRRRRRXXXXXXXXIIIII
(R=repeat, I=intron)

Nimblegen (avoid repeats and tiles no longer than need to be)
The tiled (T) region (or bait):      TTTTTTTTTTTTT     TTTTTTTTTT
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):   TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT
The oligos(O) (or probes):          OOOOOOOOOOOOOOOO|OOOOOOOOOOOO


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
```

Metrics - Coverage



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

What is a duplicate?



Duplicates potentially introduce variant calling errors

NB: it does not always make sense to remove duplicates e.g. Halo capture >> **another example of having to think of what we are doing**

NORWEGIAN SEQUENCING CENTRE

RE-ALIGNMENT

Mapping errors require re-alignment

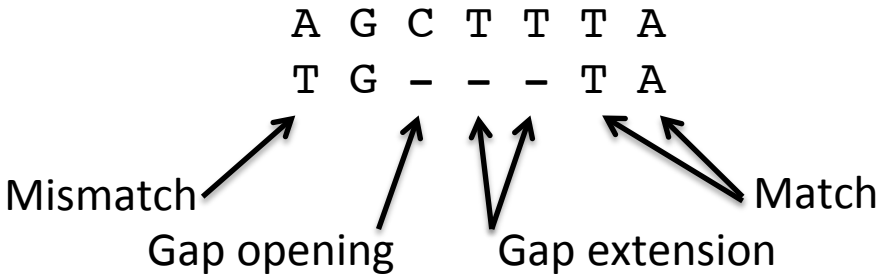
			coor	12345678901234	567890123456789012345 6
9	t	ttt	ref	aggttttataaaac	----aattaagtctacagagcaacta
10	a	aaaC	sample	aggttttataaaac	AAATaattaagtctacagagcaacta
11	a	aaaaa	read1	aggttttataaaac	aaAtaa
12	a	aaaaaa	read2	ggttttataaaac	aaAtaaTt
13	a	aaaaaa	read3	ttataaaac	AAATaattaagtctaca
14	c	cccTTT	read4	CaaaT	aattaagtctacagagcaac
15	a	aaaaaa	read5	aaT	aattaagtctacagagcaact
16	a	aaaaaa	read6	T	aattaagtctacagagcaacta
17	t	AAtttt	read1	aggttttataaaac	aaataa
18	t	tttttt	read2	ggttttataaaac	aaataatt
19	a	aaaaaa	read3	ttataaaac	aaataattaagtctaca
20	a	aaaaaa	read4		caaataattaagtctacagagcaac
21	g	Tgggg	read5		aataattaagtctacagagcaact
			read6		taattaagtctacagagcaacta

Source: Heng Li, presentation at GSA workshop 2011

NORWEGIAN SEQUENCING CENTRE

Alignment

- Key component of alignment algorithm is the scoring
  - negative contribution to score
    - opening a gap
    - extending a gap
    - mismatches
  - positive contribution to score
    - matches



- When aligning two sequences there **is only one set of differences** to consider

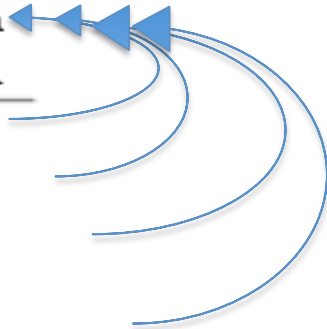
- In a multiple sequence alignment, **one has to consider all pairs of differences** in the scoring algorithm

Few mismatches when considering one-to-one

Base stacks

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
read3          ttataaaacAAATaattaagtctaca
read4              CaaaT    aattaagtctacagagcaac
read5                aaT    aattaagtctacagagcaact
read6                 T     aattaagtctacagagcaacta
read1     aggttttataaaacaaataa
read2      ggttttataaaacaaataatt
read3          ttataaaacaaataattaagtctaca
read4                  caaataattaagtctacagagcaac
read5                    aataattaagtctacagagcaact
read6                     taattaagtctacagagcaacta
```



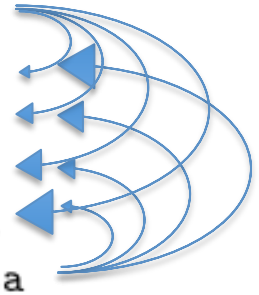
NORWEGIAN SEQUENCING CENTRE

Lots of mismatch in all-to-all if reads mismapped

Base stacks

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     567890123456789012345678
ref       aggttttataaaac----aattaagtctacagagcaacta
sample    aggttttataaaacAAATaattaagtctacagagcaacta
read1     aggttttataaaac    aaAtaa
read2      ggttttataaaac    aaAtaaTt
read3          ttataaaacAAATaattaagtctaca
read4              CaaaT    aattaagtctacagagcaac
read5                aaT    aattaagtctacagagcaact
read6                 T     aattaagtctacagagcaacta
read1     aggttttataaaacaaataa
read2      ggttttataaaacaaataatt
read3          ttataaaacaaataattaagtctaca
read4                caaataattaagtctacagagcaac
read5                  aataattaagtctacagagcaact
read6                   taattaagtctacagagcaacta
```



No mismatches between reads

NORWEGIAN SEQUENCING CENTRE

Mapping vs. alignment

Mapping vs. alignment

Mapping

- A mapping is the region where a read sequence is placed.
- A mapping is regarded to be correct if it overlaps the true region.

Alignment

- An alignment is the detailed placement of each base in a read.
- An alignment is regarded to be correct only if each base is placed correctly.

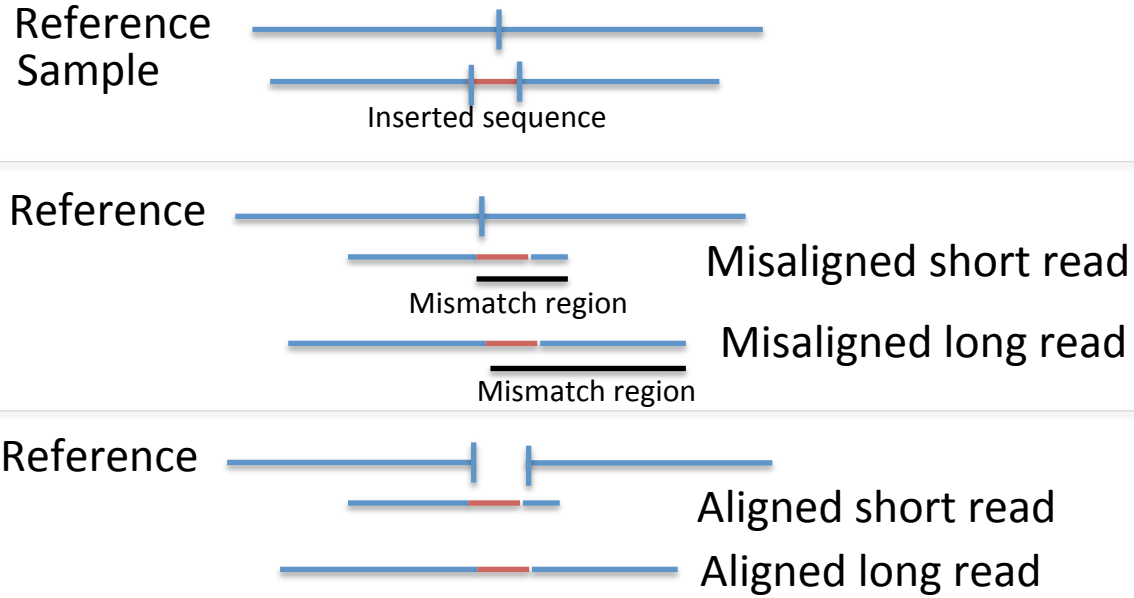
The problem

- A read mapper is fairly good at mapping, may not be good at alignment.
- This is because the true alignment minimizes differences between reads, but the read mapper only sees the reference.

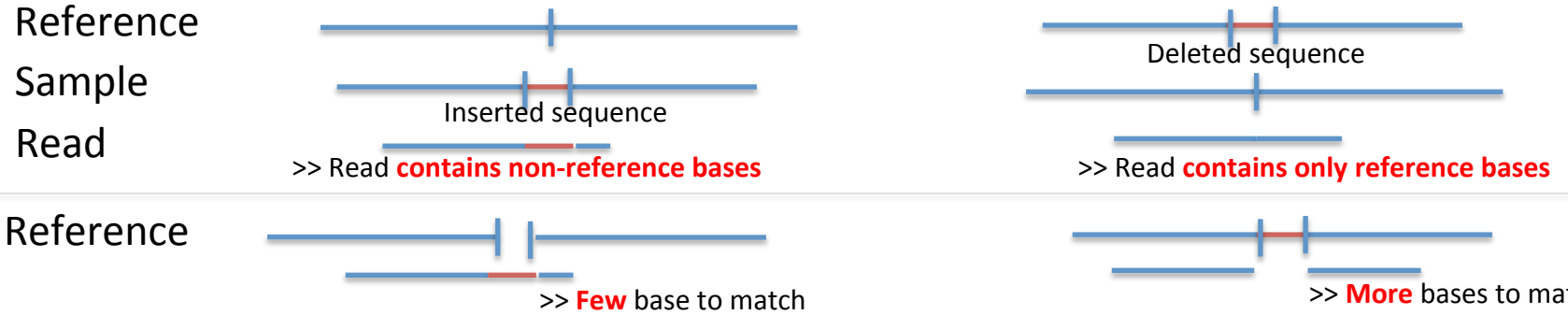
Source: Heng Li, presentation at GSA workshop 2011

Detection of indels

Effect of read length



Asymetry between insertions and deletions

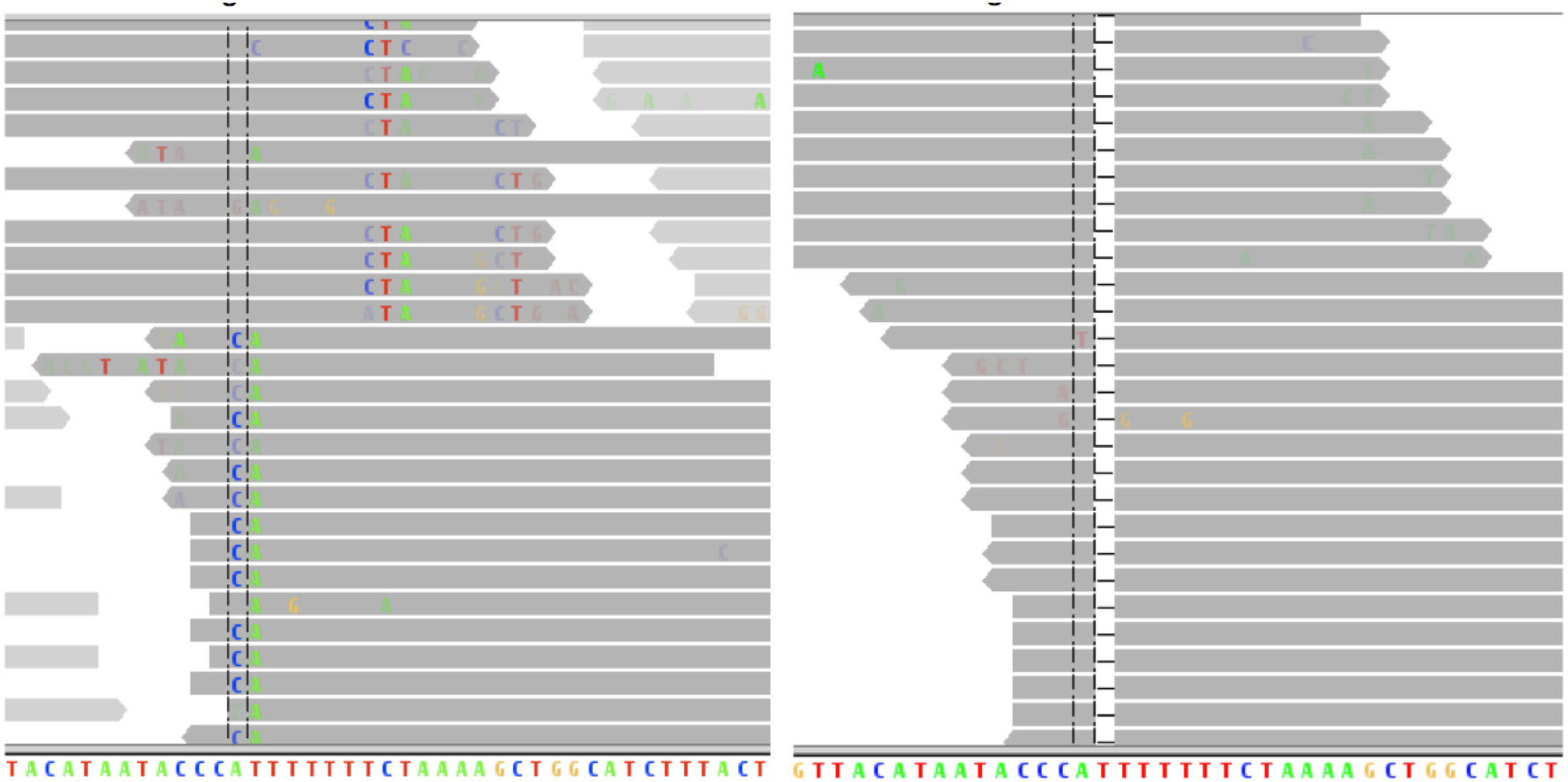


>> insertion and deletion of same size, but more likely to detect the deletion

Local realignment around indels

Before

After



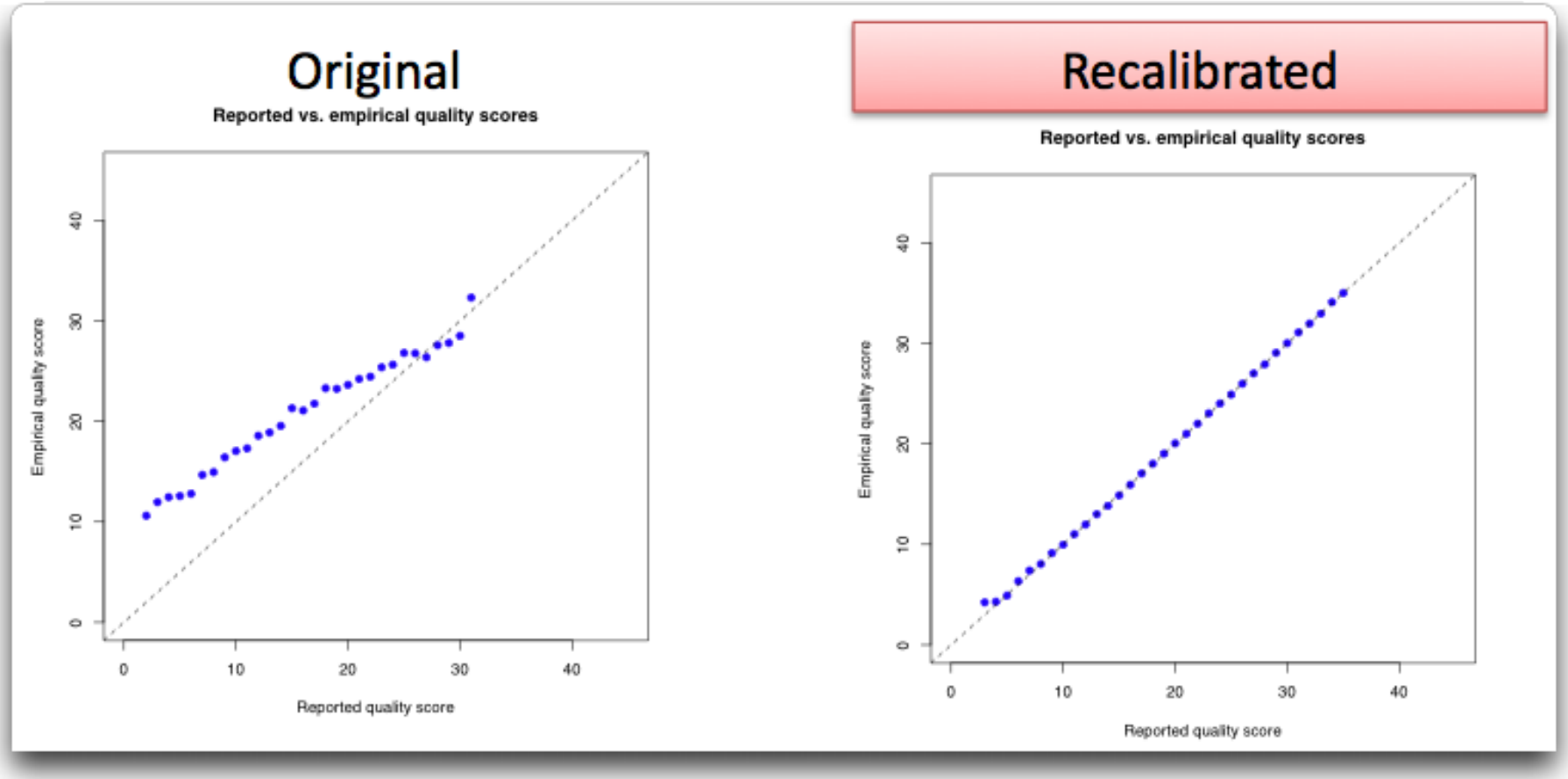
http://www.broadinstitute.org/files/shared/mpg/nextgen2010/nextgen_sivachenko.pdf

BASE QUALITY SCORE RECALIB.

Theoretical vs Empirical error rates / qualities

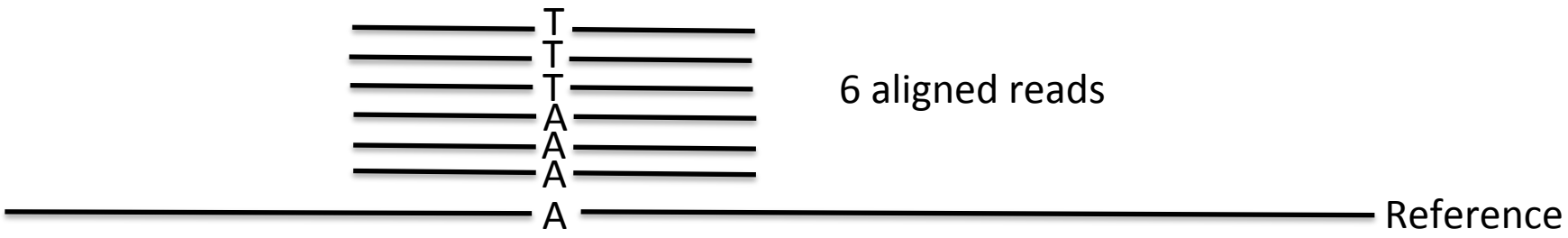
- The qualities in the fastq file are computed using a model
- This model is not perfect >> there are discrepancies between the model and the empirical error rate
- We can compute a good approximation of the empirical error rate by identifying all sites where there are mismatches between the read and the reference (being careful to ignore sites with known SNPs)
- We can analyse whether there are parameters of the bases that covary with the discrepancy
  - e.g. cycle
- We can use these quantified covariances to recalibrate the base qualities >> more accurate qualities

Result of recalibration



VCF FORMAT – MORE DETAILS

A way of thinking of the variant calling process



- Compute:
  - the probability that the site is variant
  - the likelihood of the different genotypes

- Difference between variant site and genotype:
  - ref is A, aligned bases are TTTTTTTAA
  - highly likely that the site is variant
  - less clear what the genotype is: T/A or T/T?

- Complex mathematical models involved in both allele frequency calculations and genotype likelihoods >> wise to use the recommended option settings in the tool documentation (as we have done in the practicals)

Bayesian variant caller

Input

Reference is C, observing 4C and 2T, all with base quality 30.

Likelihood of data

- $P(D|CC) = \Pr\{\text{two Q30 errors}\} = 10^{-(30+30)/10} = 10^{-6}$
- $P(D|TT) = \Pr\{\text{four Q30 errors}\} = 10^{-(30*4)/10} = 10^{-12}$
- $P(D|CT) = \Pr\{\text{sample 6 reads from 2 chr}\} = 1/2^6 = 1.56 \times 10^{-2}$

Posterior

- Prior: $P(CC) = 0.9985$, $P(CT) = 0.001$ and $P(TT) = 0.0005$

$$P(CC|D) = \frac{P(D|CC)P(CC)}{P(D|CC)P(CC) + P(D|CT)P(CT) + P(D|TT)P(TT)}$$

- Get: $P(CC|D) = 0.06$, $P(CT|D) = 0.94$ and $P(TT|D) = 3 \times 10^{-11}$

Source: Heng Li, presentation at GSA workshop 2011

VCF format

```
##fileformat=VCFv4.0
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=1000GenomesPilot-NCBI36
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF ALT    QUAL FILTER INFO                              FORMAT      NA00001        NA00
20     14370   rs6054257 G   A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51
20     17330   .         T   A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50
20     1110696 rs6040355 A   G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27
20     1230237 .         T   .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60
20     1234567 microsat1 GTCT G,GTACT 50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4
```

Meta data: definitions of tags used elsewhere in data lines

Header line

Data lines

Variant columns

Genotype columns

NORWEGIAN SEQUENCING CENTRE

Columns of data lines

- **CHROMO**
- **POS**: the reference position with the 1st base having position 1
- **ID**: an id; rs number if dbSNP variant
- **REF**: reference base.
  - The value in POS refers to the position of the first base in the string
  - for indels, the reference string must include the base before the event (and this must be reflected in POS)
- **ALT**: comma sepearated list of alternate non-ref alleles called on at least one of the samples
  - if no alternate alleles then the missing value should be used "."
- **QUAL**: phred-scaled quality score of the assertion made in ALT (whether variant or non-variant)
- **FILTER**: PASS if the position has passed all filters (defined in meta-data).
- **INFO**: additional information

REF and ALT

Reference a t C g a >> C is reference base

				REF	ALT
Variant	a t **G** g a >> C is a G	→	20 3 .	C	G
Variant	a t **–** g a >> C is deleted	→	20 **2** .	TC	T
Variant	a t C**a**g a >> A is inserted	→	20 3 .	C	CA

Source: The SAM format specification

REF and ALT

Reference a t C g a >> C is reference base

			REF	ALT
Variant	a t **G** g a >> C is a G	→	20 3 . C G	
Variant	a t **–** g a >> C is deleted	→	**20 2** . TC T	

		REF	ALT
To represent both in the same record	→	**20 2** . TC	**T,TG**

INFO, FORMAT, and genotypes

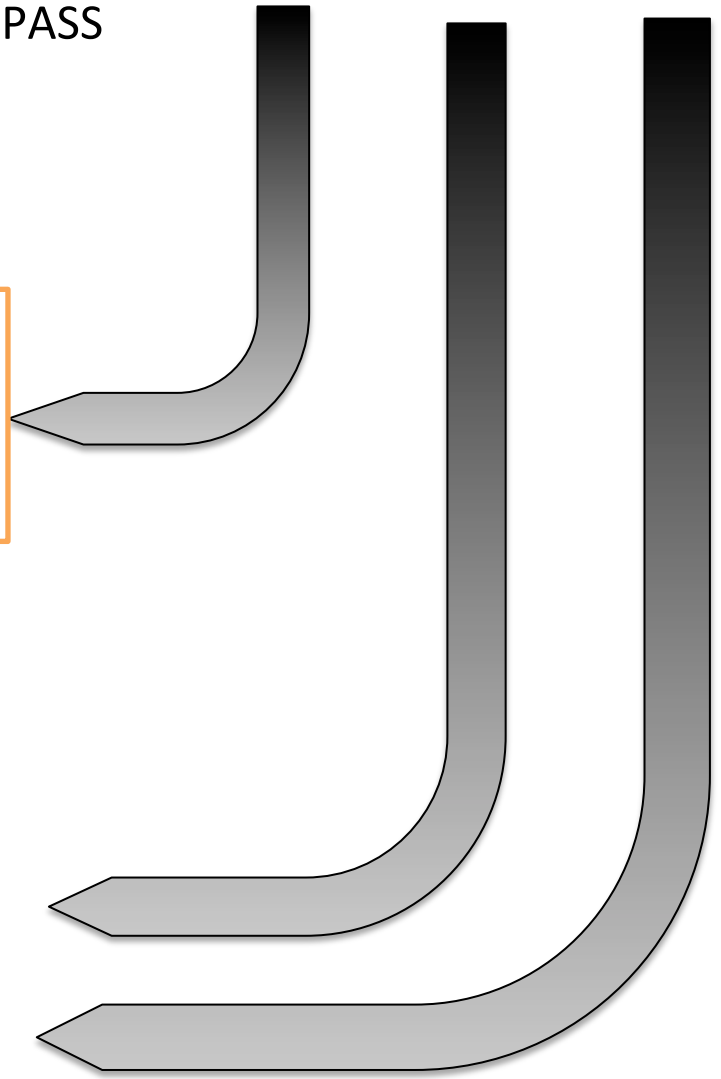
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT sample1
1 801943 rs7516866 C T 9787.34 PASS

AC=2;
AF=1.00;
AN=2;
BaseQRankSum=1.009;
DB;
DP=556;
FS=18.302;
MQ=44.04;
MQ0=38;
MQRankSum=5.122;
QD=17.60;
ReadPosRankSum=3.375

GT:AD:DP:GQ:PL

1/1:37,518:556:99:9787,685,0

We will explore these fields when we discuss filtering



NORWEGIAN SEQUENCING CENTRE

Genotype fields

- Format field specifies type of data present for each genotype
  - GT:AD:DP:GQ:PL
  - fields defined in metadata header

- GT: genotype, encoded as alleles separated by either | or /
  - 0 for the ref, 1 for the 1st allele listed in ALT, 2 for the second, etc
  - REF=A and ALT=T
    - genotype 0/1 means hetero A/T
    - genotype 1/1 means homo T/T
  - /: genotype unphased and | genotype phased
- DP: read depth at position for sample
- GQ: genotype quality encoded as a phred quality
- etc.....

Homozygous SNP

#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT

1 801943 rs7516866 **C** **T** 9787.34 PASS

AC=2;AF=1.00;AN=2;BaseQRankSum=1.009;DB;DP=556;DS;Dels=0.00;FS=18.302;HRun=1;HaplotypeScore=4.6410;MQ=44.04;MQ0=38;MQRankSum=5.122;QD=17.60;ReadPosRankSum=3.375

GT:AD:DP:GQ:PL **1/1**:37,518:556:99:9787,685,0

Heterozygous SNP

#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT

1 1918488 rs4350140 **A** **G** 233.10 PASS

AC=1;AF=0.50;AN=2;BaseQRankSum=1.349;DB;DP=33;DS;Dels=0.00;FS=0.000;HRun=0;HaplotypeScore=0.0000;MQ=68.18;MQ0=1;MQRankSum=0.436;QD=7.06;ReadPosRankSum=1.547

GT:AD:DP:GQ:PL **0/1**:21,12:33:99:263,0,620

Homozygous deletion

#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT

1 1289367 rs35062587 **CTG** **C** 3139.27 PASS

AC=2;AF=1.00;AN=2;DB;DP=66;DS;FS=0.000;HRun=0;HaplotypeScore=223.1329;MQ=68.34;MQ0=1;QD=47.56

GT:AD:DP:GQ:PL **1/1**:0,66:65:99:3181,196,0

Heterozygous insertion

#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT

1 17948305 . **G** **GGGCCACAGCAG** 3581.32 PASS

AC=1;AF=0.50;AN=2;BaseQRankSum=-2.638;DP=54;DS;FS=0.000;HRun=0;HaplotypeScore=552.8152;MQ=70.65;MQ0=2;MQRankSum=3.258;QD=66.32;ReadPosRankSum=0.320

GT:AD:DP:GQ:PL **0/1**:44,10:52:99:3581,0,3730

FILTERING

The rationale for filtering

- To eliminate False Positive variants from variant list

- What causes errors in variant calling?
  - sequencing errors >> should be accounted for by base quality + recalibration + marking of duplicates
  - Incorrect alignment >> Re-alignment step should have reduced this problem **but not eliminated it**

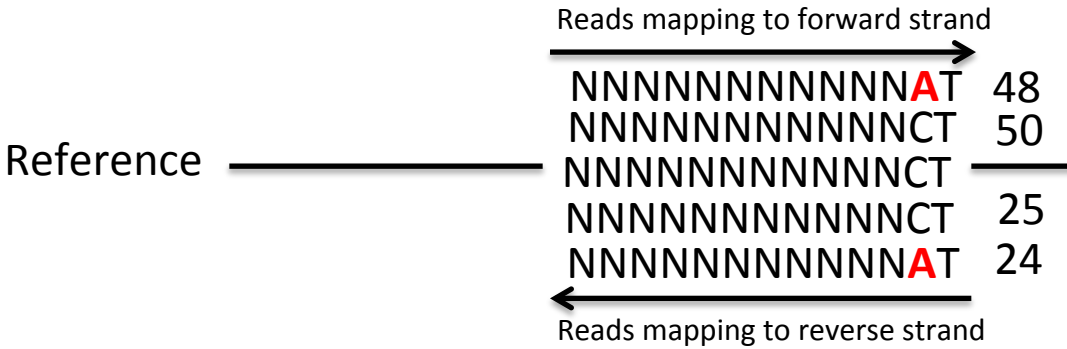
- Tell tale signs of suspicious variants
  - poorly mapped reads (ambiguity)
    - MQ: Root Mean Square of MAPQ of all reads at locus
    - MQ0: Number of MAPQ 0 reads at locus
  - biased support for the **REF** and **ALT** alleles
    - MQRankSum: Mapping quality rank sum test
    - ReadPosRankSum: Read position rank sum test
    - Strand bias and FS:

INFO fields – important for filtering

- **QD:** variant quality score over depth
  - Confidence in the site being variant should increase with increasing depth
- **MQ:** RMS MAPQ of all reads at locus
  - Regions of excessively low mapping quality are ambiguously mapped and variants called within are suspicious
- **MQ0:** number of MAPQ 0 reads at locus
- **MQRankSum:** Mapping quality rank sum test
  - If the alternate bases are more likely to be found on reads with lower MQ than reference bases then the site is likely mismapped
- **Haplotype score:** Probability that the reads in a window around the variant can be explained by at most two haplotypes
- **FS:** fisher exact test of read strand
  - If the reference-carrying reads are balanced between forward and reverse strands then the alternate-carrying reads should be as well
- **ReadPosRankSum:** Read position rank sum test
  - If the alternate bases are biased towards the beginning or end of the reads then the site is likely a mapping artifact

Strand bias (assume heterozygote)

This is **NOT** strand bias: strand bias is **NOT** about more reads mapping to one of the strands than the other

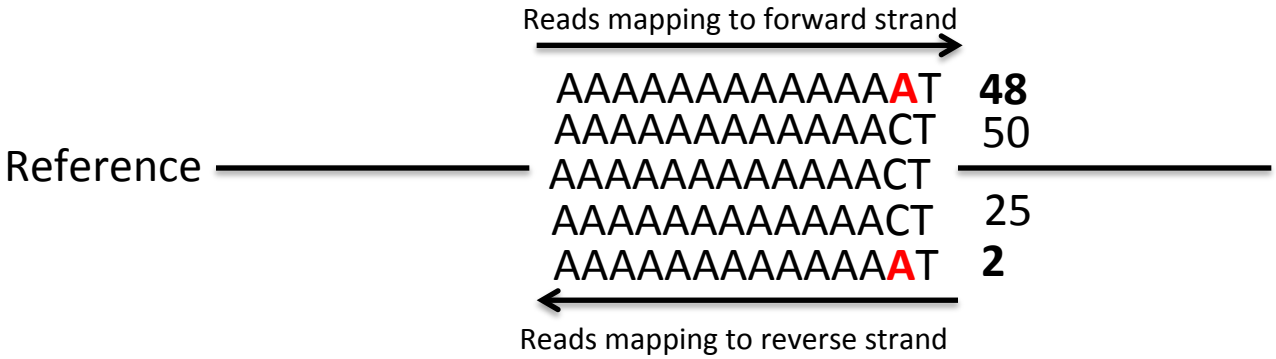


		Fw	Rev
Ref	C	50	25
Alt	A	48	24

Clearly more reads mapping to FW than Rev

But, Fw/Rev ratio is same for Ref allele and Alt: 50/25 = 48/24

This **IS** strand bias



		Fw	Rev
Ref	C	50	25
Alt	A	48	2

50/25 != 48/2

NORWEGIAN SEQUENCING CENTRE

Hard vs. soft filtering

- Can set thresholds for these INFO fields and request that all thresholds are passed for a variant to be considered valid

- Which fields to you use and where do you set the thresholds?
  - use datasets of known SNPs and compare their INFO fields to those likely FP variants
  - fields that provide a good separation can be used as filters

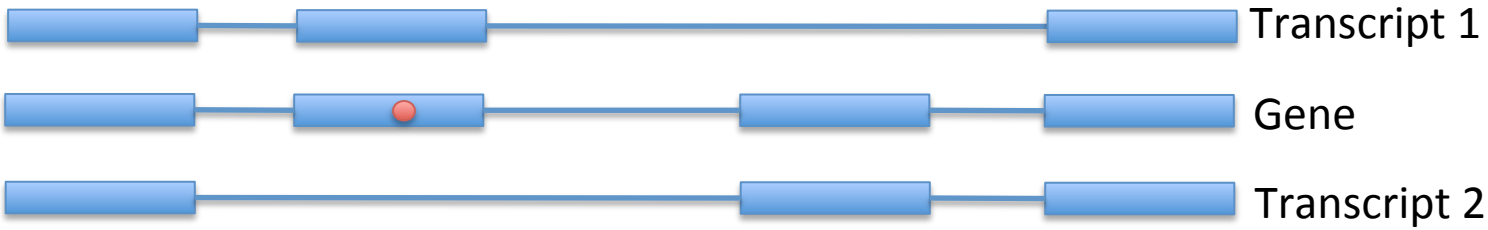
- Disadvantage of **hard filtering**
  - works with hard cut-offs

- Variant Quality Score Recalibration (GATK) or **soft filtering**

ANNOTATION

What is annotation?

- Adding information about the variants
- Two broad categories of annotations
  - annotation that **depend on gene models**
    - coding/non-coding
    - if coding: synonymous / non-synonymous
    - if non-synonymous >> what is the impact on protein structure (Polyphen, SIFT, etc)
  - annotations that **do not depend on gene models**
    - variant frequency in different database / different populations
    - degree of conservation across species
- Considerable complications caused by different gene models



- Two approaches to problem
  - decide **ex-ante** what which transcript to use for each gene
  - annotate with all transcript for a given gene and pick the **highest impact effect**

Annotation software

- Two sets of software
  - Annovar
    - provides a wide range of annotations that can be applied with one tool
    - we have experienced some inconsistencies in the results e.g. non-synonymous SNPs without polyphen score
  - SNPEff and dbNSFP (non-synoymous functional prediction)

- Both tested by GATK team
  - recommended snpEff, but with strict requirements
  - **snpEff version 2.0.5** (not 2.0.5d)
  - db should be **GRCh37.64** (which is the ensembl database version 64)
  - should use the option **-onlyCoding true** (using false can cause erroneous annotation)

- GATKs VariantAnnotator to pick the highest impact.

- Finally, also annotate with **dbNSFP, which contains:**
  - variant frequencies
  - conservation scores
  - protein function effect

snpEff annotation

31942920 . G T 683.93 PASS

AC=1;AF=0.50;AN=2;BaseQRankSum=4.358;DP=73;DS;Dels=0.00;FS=0.000;HRun=0;
HaplotypeScore=1.7876;MQ=69.76;MQ0=0;MQRankSum=0.977;QD=9.37;ReadPosR
ankSum=0.508; VQSLOD=1.6292;culprit=QD

SNPEFF_TRANSCRIPT_ID=ENST00000421060;
SNPEFF_GENE_NAME=SFI1;
SNPEFF_EXON_ID=exon_22_31942847_31942957;
SNPEFF_CODON_CHANGE=Gag/Tag;
SNPEFF_AMINO_ACID_CHANGE=E114*;
SNPEFF_EFFECT=STOP_GAINED;
SNPEFF_FUNCTIONAL_CLASS=NONSENSE;
SNPEFF_GENE_BIOTYPE=processed_transcript;
SNPEFF_IMPACT=HIGH;

GT:AD:DP:GQ:PL 0/1:42,31:73:99:714,0,981

Example of annotation with dbNSFP

766910 rs1809933 C T 556.42 PASS

AC=1;AF=0.50;AN=2;BaseQRankSum=1.366;DB;DP=30;Dels=0.00;FS=0.000;HRun=0;HaplotypeScore=1.8675;MQ=47.46;MQ0=0;MQRankSum=-0.651;QD=18.55;ReadPosRankSum=-1.757;SB=-109.24;

SNPEFF_AMINO_ACID_CHANGE=R42Q;SNPEFF_CODON_CHANGE=cGg/cAg;SNPEFF_EFFECT=NON_SYNONYMOUS_CODING;SNPEFF_EXON_ID=exon_5_766813_767034;SNPEFF_FUNCTIONAL_CLASS=MISSENSE;SNPEFF_GENE_BIOTYPE=processed_transcript;SNPEFF_GENE_NAME=ZDHHC11B;SNPEFF_IMPACT=MODERATE;SNPEFF_TRANSCRIPT_ID=ENST00000382776;

dbnsfp1000Gp1_ASN_AF=0.8199300699300699;
dbnsfpEnsembl_transcriptid=ENST00000508859,ENST00000382776;
dbnsfp1000Gp1_AFR_AF=0.75;
dbnsfp1000Gp1_EUR_AF=0.71;
dbnsfp1000Gp1_AF=0.76;
dbnsfpGERP++_RS=1.43;
dbnsfpGERP++_NR=2.68;
dbnsfp29way_logOdds=3.0289;
dbnsfpSIFT_score=1.000000

GT:AD:DP:GQ:PL 0/1:5,25:30:98:586,0,98

Filtering a functionally annotated VCF file

PRACTICAL

CONCLUDING REMARKS