Reference genomes & alignment

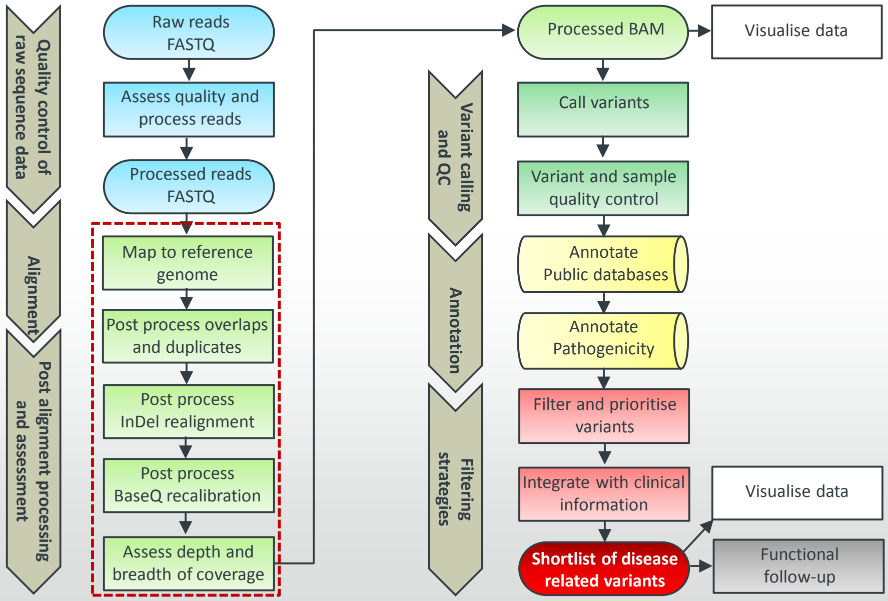
Dr. Reuben J. Pengelly
6th February, 2017

Learning outcomes

At the end of this lecture, you should be able to:

- Describe the purpose of a reference genome and evaluate the current resources
- List and describe the stages of data processing in the context of alignment
- Criticize the current approaches compared to an idealised situation

Analysis workflow



Quality control of raw sequence data

Alignment

Post alignment processing and assessment

Raw reads FASTQ

Assess quality and process reads

Processed reads FASTQ

Map to reference genome

Post process overlaps and duplicates

Post process InDel realignment

Post process BaseQ recalibration

Assess depth and breadth of coverage

Variant calling and QC

Annotation

Filtering strategies

Processed BAM

Visualise data

Call variants

Variant and sample quality control

Annotate Public databases

Annotate Pathogenicity

Filter and prioritise variants

Integrate with clinical information

Shortlist of disease related variants

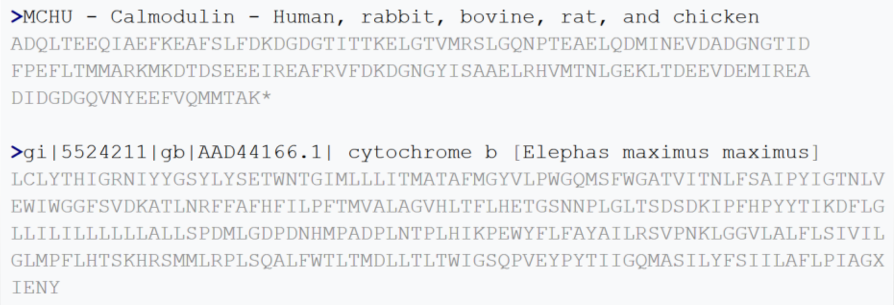
Visualise data

Functional follow-up

Reference genomes

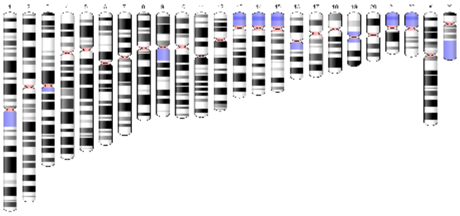
- Theoretical genome to which a sample is compared
  - ~3.2 billion bases
- Not based on any one person
  - Initially European-centric
  - Progress toward global consensus
- GRCh38 (hg38) is latest version

FASTA



```
>MCHU - Calmodulin - Human, rabbit, bovine, rat, and chicken
ADQLTEEQIAEFKEAFSLFDKDGDGTITTKELGTVMRSLGQNPTEAELQDMINEVDADGNGTID
FPEFLTMMARKMKDTDSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREA
DIDGDGQVNYEEFVQMMTAK*

>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLV
EWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLG
LLILILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFLPIAGX
IENY
```

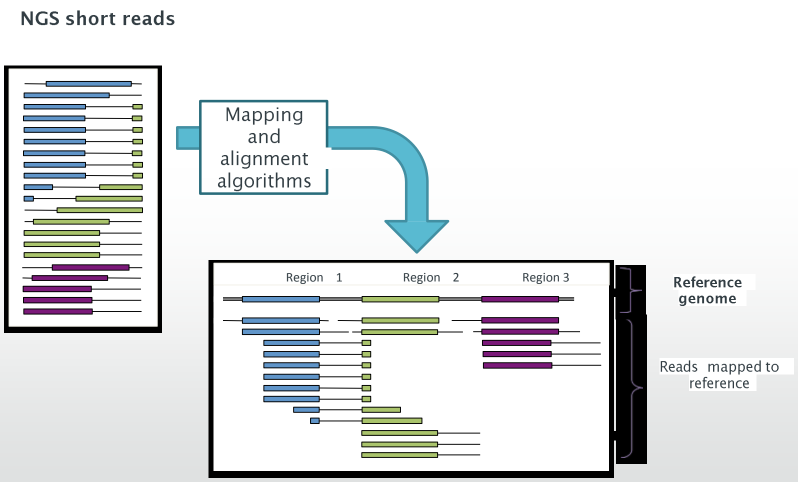
GRCh38



- Most complete reference:
  - GRCh38 (2013) – 3.05 Gb
  - GRCh37 (2009) – 2.90 Gb
  - Reflects global genetic diversity data

- Highly complex regions (i.e. HLA) pose a special challenge
  - If sample's HLA is very different, will fail to align efficiently
  - Alternative HLA references included in GRCh38 to reflect diversity

Alignment/Mapping – a simple idea



Alignment *vs.* assembly

Alignment

- Align short reads to reference genome
- Requires reference genome
- Can have discontinuous sequencing data

Assembly

- Stitching together of short reads
- No reference required
- Must have continuous data

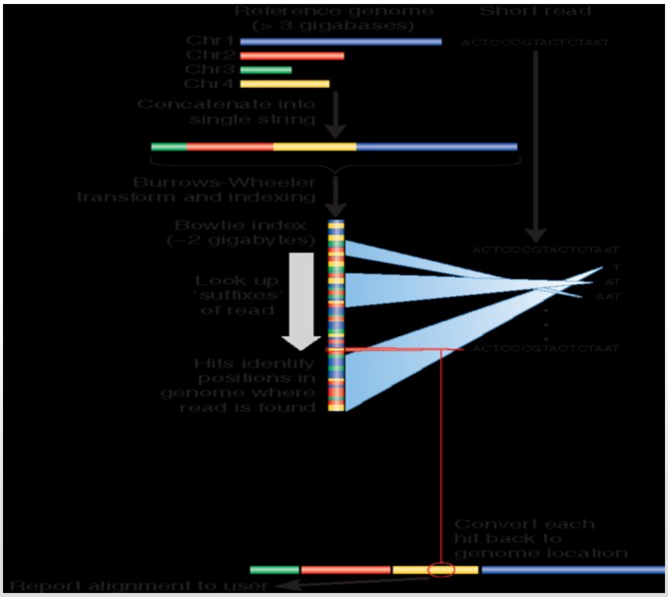
Alignment challenges

- Problems in reference genome
  - Undefined regions
  - Errors
- Diversity from reference genome
  - SNPs, indels & structural rearrangements
- Sequencing errors
- Simple regions (e.g. microsatelites and CAG repeats)

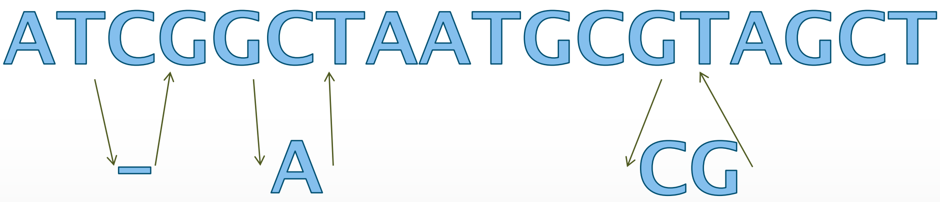
Alignment software

- Tool used must be appropriate for experiment
  - BWA common for human DNA sequencing
    - Many alternatives are available
  - Other aligners (e.g. Bowtie) use for RNA sequencing
- Most intensive step of NGS analysis
- Critical for accuracy of analysis

Burrow-Wheeler transform based algorithm


Reference genome
(> 3 gigabases)
Short read
Chr1
Chr2
Chr3
Chr4
ACTCCCGTACTCTAAT
Concatenate into single string
Burrows-Wheeler transform and indexing
Bowtie index (~2 gigabytes)
ACTCCCGTACTCTAAT
T
AT
AAT
Look up 'suffixes' of read
ACTCCCGTACTCTAAT
Hits identify positions in genome where read is found
Convert each hit back to genome location
Report alignment to user

Graph-based alignment

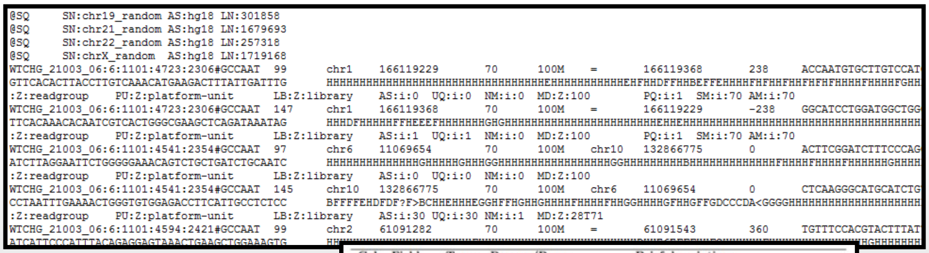


ATCGGCTAATGCGTAGCT

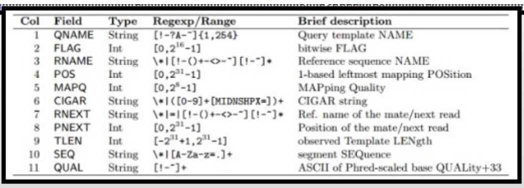
– A CG

SAM files

- Contains aligned reads with position and quality
- Large file (~ 12 GB for exome)



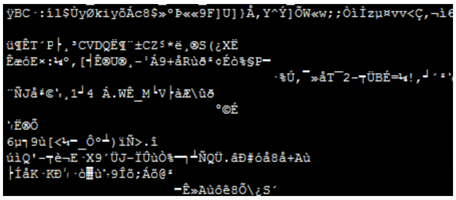
```
@SQ     SN:chr19_random AS:hg18 LN:301858
@SQ     SN:chr21_random AS:hg18 LN:1679693
@SQ     SN:chr22_random AS:hg18 LN:257318
@SQ     SN:chrX_random  AS:hg18 LN:1719168
WTCHG_21003_06:6:1101:4723:2306#GCCAAT  99      chr1    166119229       70      100M    =       166119368       238     ACCAATGTGCTTGTCCAT
GTTCACACTTACCTTGTCAAACATGAAGACTTTATTGATTTG      HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHEHHHHHHHHHHHHHEHFHHDFFHHBEFFEHHHHFHFHHFHHFHFHFHHHHFHHHHFGHH
:Z:readgroup    PU:Z:platform-unit      LB:Z:library    AS:i:0  UQ:i:0  NM:i:0  MD:Z:100        PQ:i:1  SM:i:70 AM:i:70
WTCHG_21003_06:6:1101:4723:2306#GCCAAT  147     chr1    166119368       70      100M    =       166119229       -238    GGCATCCTGGATGGCTGG
TTCACAAACACAATCGTCACTGGGCGAAGCTCAGATAAATAG      HHHDFHHHHHFFHEEEFHHHHHHHGHGHHHHHHHHHHHHHHHHHHHHHHHHEHHEHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH
:Z:readgroup    PU:Z:platform-unit      LB:Z:library    AS:i:1  UQ:i:1  NM:i:0  MD:Z:100        PQ:i:1  SM:i:70 AM:i:70
WTCHG_21003_06:6:1101:4541:2354#GCCAAT  97      chr6    11069654        70      100M    chr10   132866775       0       ACTTCGGATCTTTCCCAG
ATCTTAGGAATTCTGGGGGAAACAGTCTGCTGATCTGCAATC      HHHHHHHHHHHHHHGHHHHHGHHHGGHHHHHHHHHHHHHHHHHHGGHHHHHHHHHBHHHHHHHHHHHHHHFHHHHFHHHHFHHHHHHGHHHH
:Z:readgroup    PU:Z:platform-unit      LB:Z:library    AS:i:0  UQ:i:0  NM:i:0  MD:Z:100
WTCHG_21003_06:6:1101:4541:2354#GCCAAT  145     chr10   132866775       70      100M    chr6    11069654        0       CTCAAGGGCATGCATCTG
CCTAATTTGAAAACTGGGTGTGGAGACCTTCATTGCCTCTCC      BFFFFEHDFDF?F>BCHHEHHHEGGHFFHGHHGHHHHFHHHHFHHGGHHHHGFHHGFFGDCCCDA<GGGGHHHHHHHHHHHHHHHHHHHH
:Z:readgroup    PU:Z:platform-unit      LB:Z:library    AS:i:30 UQ:i:30 NM:i:1  MD:Z:28T71
WTCHG_21003_06:6:1101:4594:2421#GCCAAT  99      chr2    61091282        70      100M    =       61091543        360     TGTTTCCACGTACTTTAT
ATCATTCCCATTTACAGAGGAGTAAACTGAAGCTGGAAAGTG      HH
```



Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,254}`	Query template NAME
2	FLAG	Int	$[0,2^{16}-1]$	bitwise FLAG
3	RNAME	String	`\*|[!-()+-<>-~][!-~]*`	Reference sequence NAME
4	POS	Int	$[0,2^{31}-1]$	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 read
8	PNEXT	Int	$[0,2^{31}-1]$	Position of the mate/next read
9	TLEN	Int	$[-2^{31}+1,2^{31}-1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

BAM format

- Binary alignment/map
  - Lossless compression
  - SAM for exome ~12 Gb
  - BAM for exome ~ 3 Gb
- Machine readable
- Faster to access and process data
- CRAM files are alternative
  - Lossy compression of quality scores



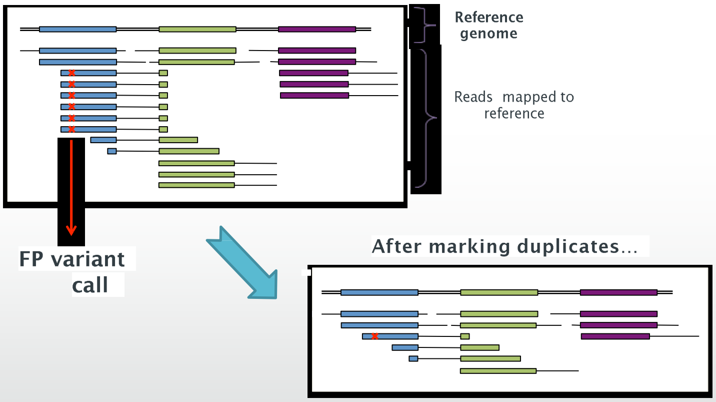
```
ÿBC ·:íl$ÙyØkiyõÁc8$»°Þ««9F]U]}Å,Y^Ý]ÕW«w;;ÒìÌzµ¤vv<Ç,¬ì6

ü¶ÊT´P├,³CVDQË¶¨±CZ*ë,®S(¿XË
ÊæóE×:¼°,[┤Ê®U®,-'Á9+åRùð¹¢Éò%§P━
                                    ·%Ú,¯»åT¯2-┬ÜBÉ=¼!,┘´¹'
¨ÑJå‡©',1┘4 Á.WÊ_M└V├àÆ\ûð
                              °©É
'Ë®Õ
6µ┐9ù[<¼━_Ô°┴)ïÑ>.î
úìQ'-┬è¬E·X9´ÜJ-ÏÛùÒ%━┐┴ÑQÜ.âÐ#óå8å+Aù
├ÍåK·KÐ'·ò▓ù'·9Îõ;Áõ@¹
                         ━Ê»AùôêßÕ\¿S´
```

Medicine

UNIVERSITY OF Southampton

Marking duplicates

✖ = sequencing error propagated in duplicates



Reference genome

Reads mapped to reference

FP variant call

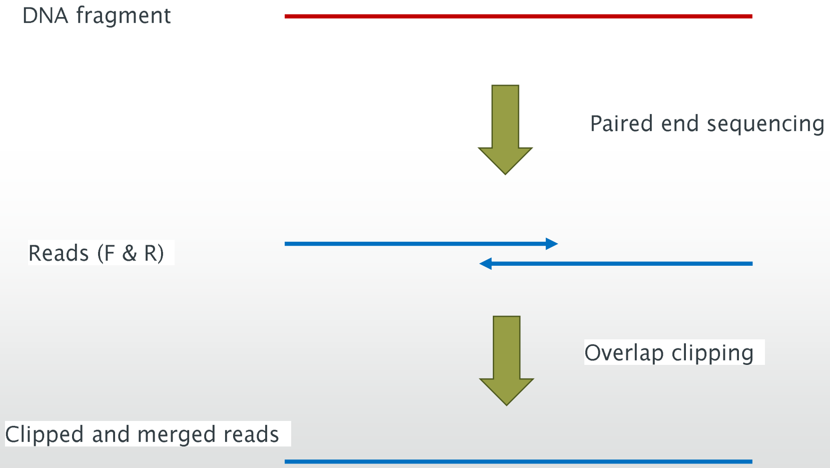
After marking duplicates…

Increase the probability of making the right call

Medicine

UNIVERSITY OF Southampton

Clipping overlaps



DNA fragment

Paired end sequencing

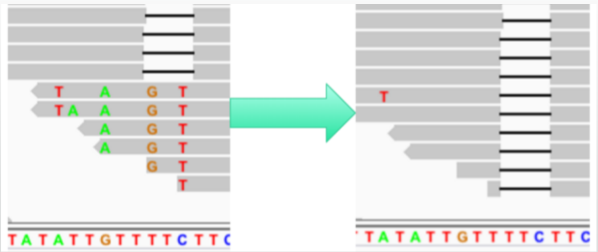
Reads (F & R)

Overlap clipping

Clipped and merged reads

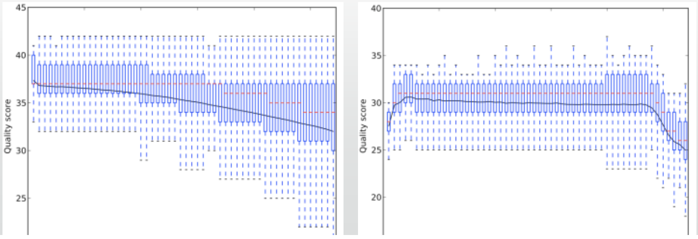
Local realignment

- Indel can shift sequence in reads to make it look like there is a SNP
  - Local realignment can alter this in the BAM file
- Particular issue around homopolymer tracts (e.g. TTTTTT)
- Not required if using Haplotype Caller in GATK



Base quality recalibration

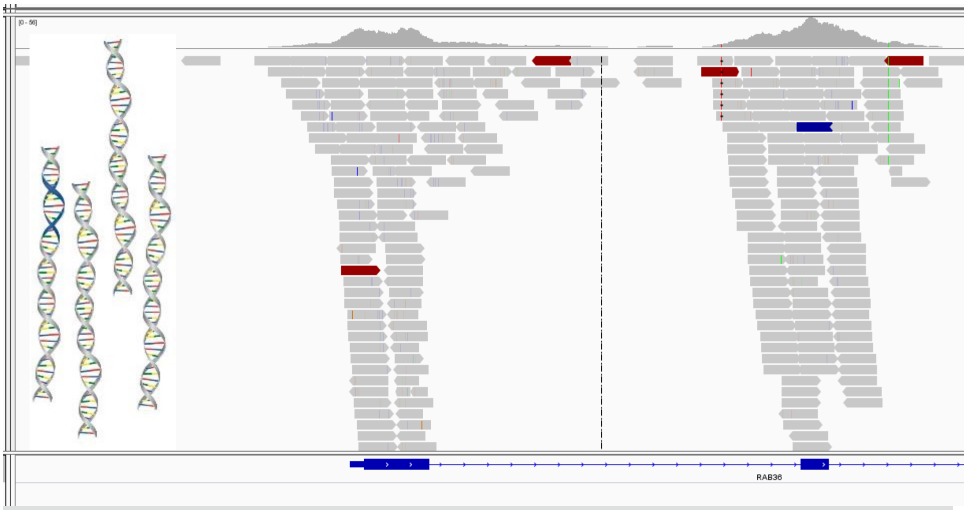
- Sequencer assigns each base call a phred quality score (q), often not accurate to true error rate
- BQR uses known variants to amend base qualities
  - Better reflect true error rate
- Accurate q scores allow for more accurate error calculations during variant calling



Coverage

- Depth of coverage is the number of reads covering a position in the genome
  - Key factor in sensitivity for variant detection
- Coverage of the genome will be variable
  - Sequence biases (GC, paralogous regions)
  - Exome capture leads to biases
  - Targeted capture relies on favourable local sequence to design kit primers or probes

Coverage



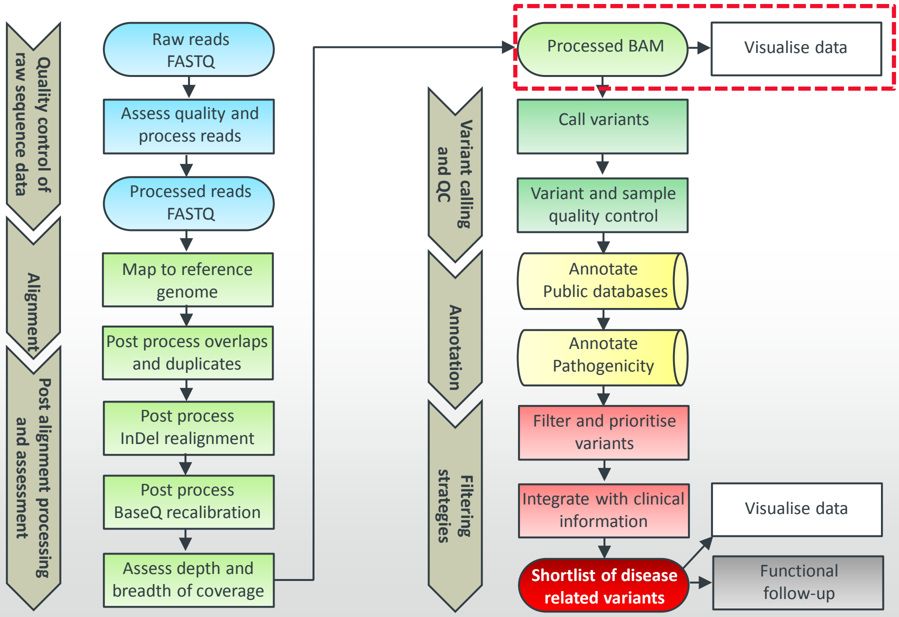
Medicine

Coverage statistics

- May want to assess:
  - Capture coverage to check efficiency in lab
    - Based on target regions
    - Expect ~80% of reads mapped
  - Gene coverage, as our sensitivity relies on this
  - Depth variation for possible copy number variation
- ~20 – 30 X is a good cut-off to assure us of good sensitivity to call a variant

Medicine

Analysis workflow



Quality control of raw sequence data

Alignment

Post alignment processing and assessment

Raw reads FASTQ

Assess quality and process reads

Processed reads FASTQ

Map to reference genome

Post process overlaps and duplicates

Post process InDel realignment

Post process BaseQ recalibration

Assess depth and breadth of coverage

Processed BAM

Visualise data

Variant calling and QC

Annotation

Filtering strategies

Call variants

Variant and sample quality control

Annotate Public databases

Annotate Pathogenicity

Filter and prioritise variants

Integrate with clinical information

Shortlist of disease related variants

Visualise data

Functional follow-up

Medicine

UNIVERSITY OF Southampton

Summary

- Reference genomes underpin genomics, and are continuously improving
- Alignment is the most intensive stage of NGS analysis and underpins all other analysis
- Post-alignment processing let you improve the quality of your data
- IGV is a valuable tool for viewing your raw data to evaluate it visually

Medicine