DNA sequencing alignment and QC

Alignment (mapping)

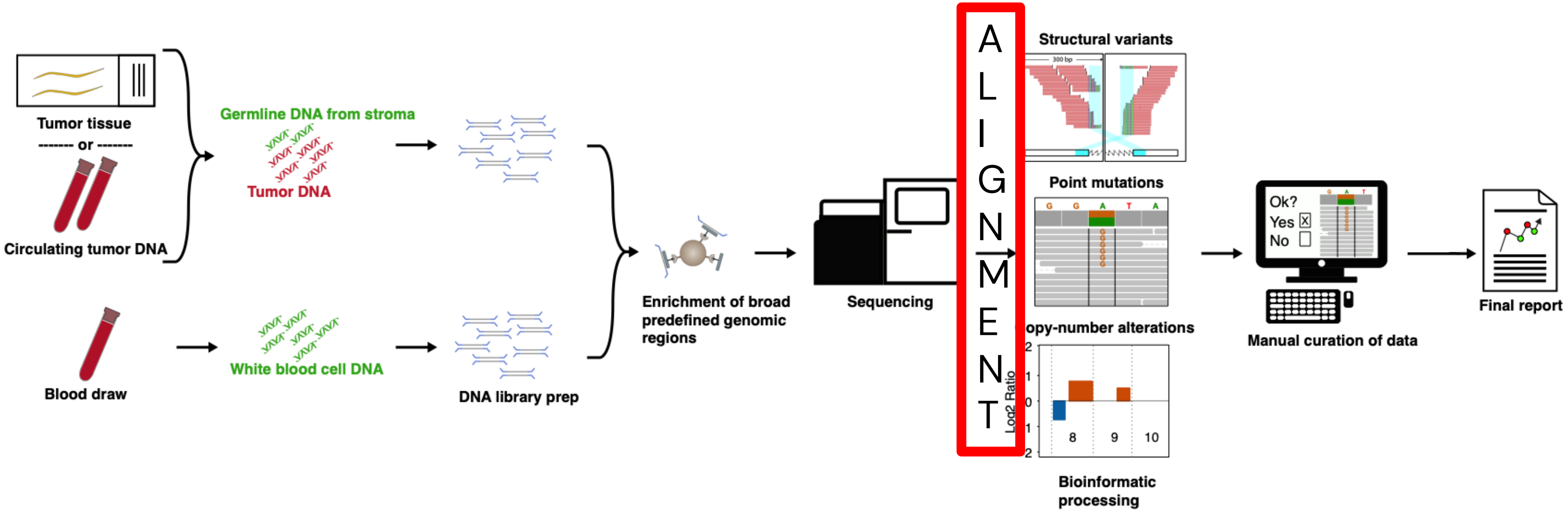
Outline

- Sequence alignment algorithms
- BAM files
- Quiz

Learning outcomes and course content

- Learning outcomes:
  - Understand how to apply technology to obtain relevant information from the cancer genome.
  - Understand the file formats used in high throughput sequencing.
  - Use the command line and running bioinformatic tools. (in exercise)
- Course content:
  - Processing of DNA and RNA sequencing data.
- Focus on DNA here, RNA is covered in on Monday

The DNA analysis process



Tumor tissue

------- or -------

Circulating tumor DNA

Blood draw

Germline DNA from stroma

Tumor DNA

White blood cell DNA

DNA library prep

Enrichment of broad predefined genomic regions

Sequencing

ALIGNMENT

Structural variants

Point mutations

Copy-number alterations

Bioinformatic processing

Manual curation of data

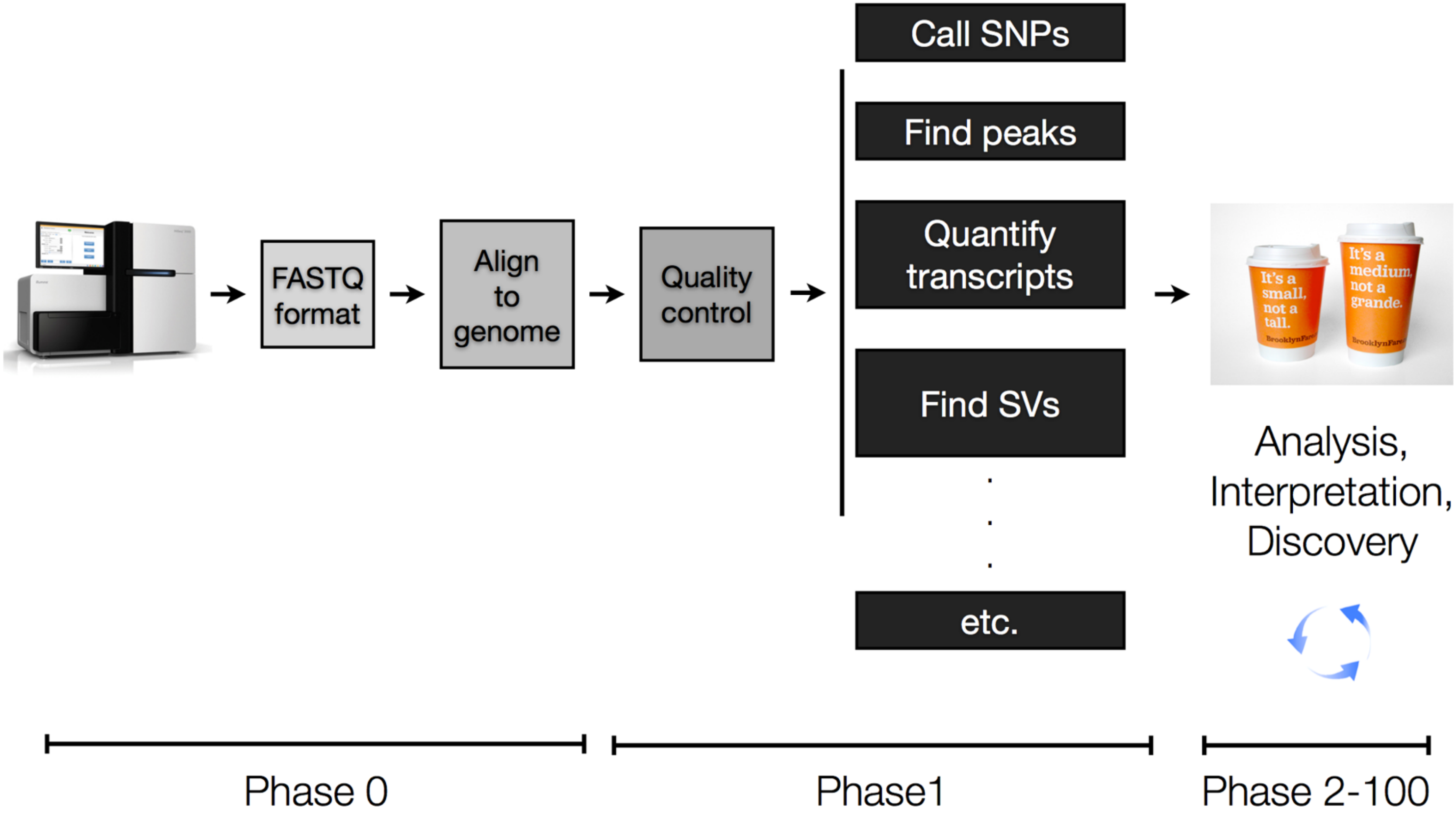
Final report

~14 days

------- Currently using research infrastructure -------

Clinic/pathology → KI biobank → SciLifeLab → Medical Epidemiology and Biostatistics → Oncology

The bioinformatic processing steps



FASTQ format → Align to genome → Quality control →

Call SNPs

Find peaks

Quantify transcripts

Find SVs

.
.
.

etc.

→ Analysis, Interpretation, Discovery

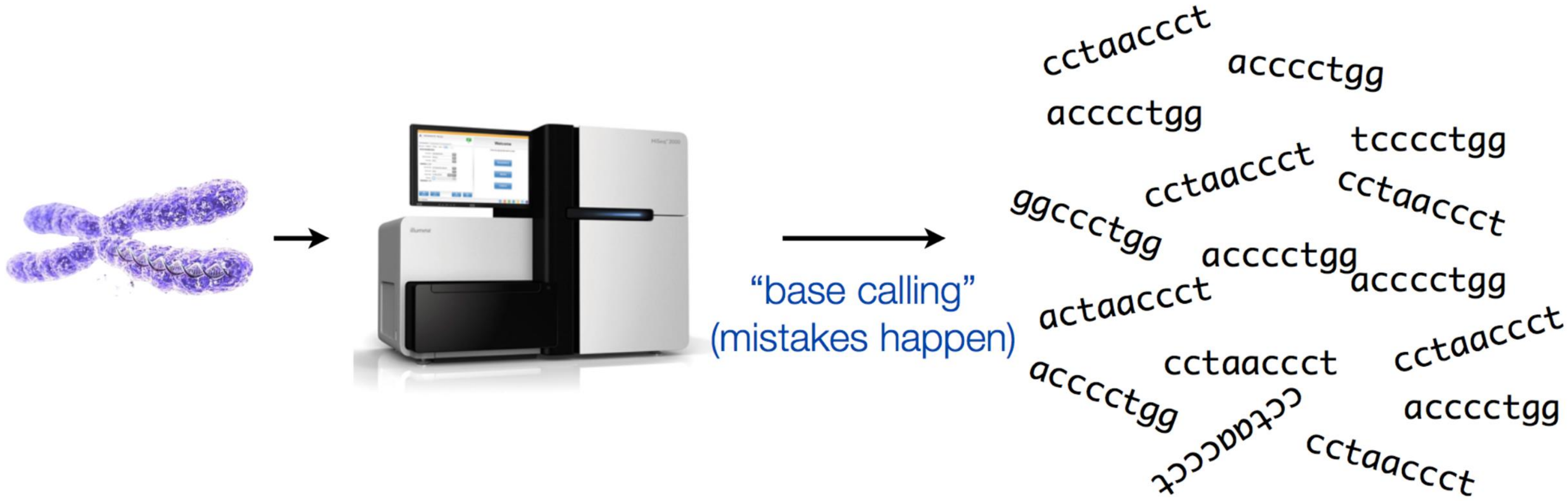
Phase 0

Phase1

Phase 2-100

Slide courtesy of Andrew Farrell (University of Utah)

DNA sequencing

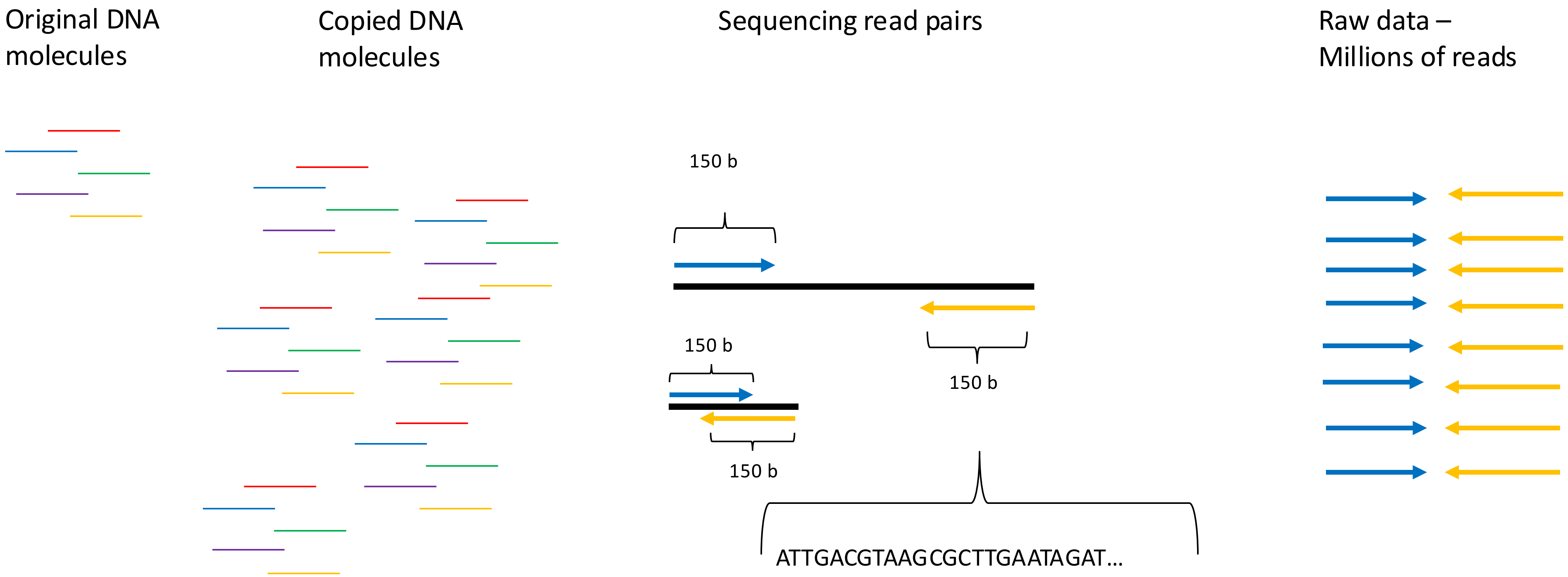


“base calling”
(mistakes happen)

cctaaccct
acccctgg
acccctgg
tcccctgg
cctaaccct
cctaaccct
ggccctgg
acccctgg
acccctgg
actaaccct
cctaaccct
cctaaccct
acccctgg
acccctgg
cctaaccct
cctaaccct

Slide courtesy of Andrew Farrell (University of Utah)

Sequencing reads



Original DNA molecules

Copied DNA molecules

Sequencing read pairs

Raw data – Millions of reads

150 b

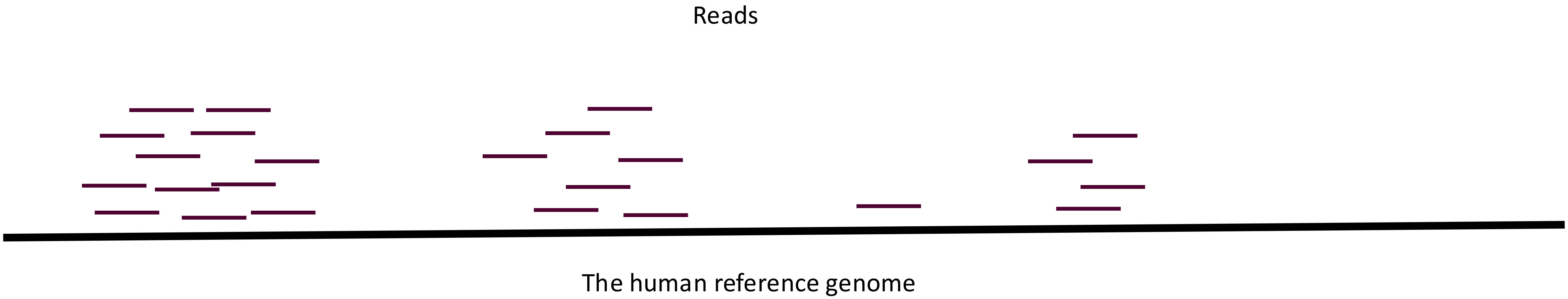
150 b

150 b

150 b

ATTGACGTAAGCGCTTGAATAGAT…

Alignment – putting the puzzle pieces in the right place

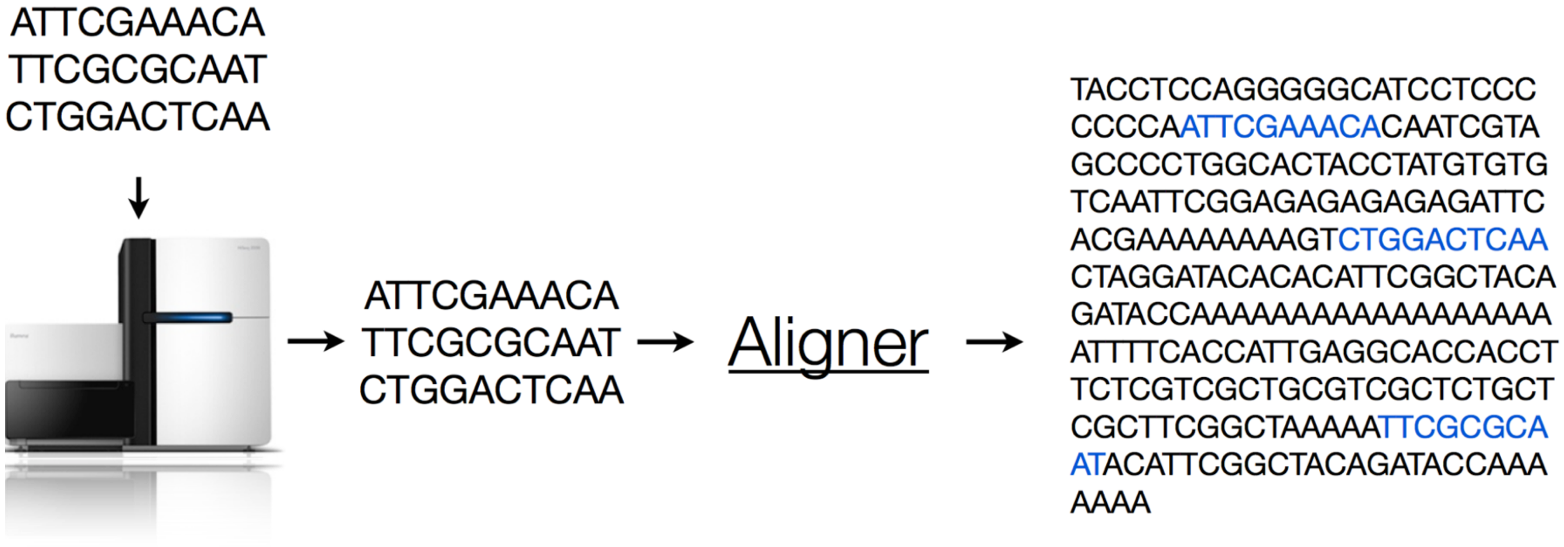


Some parts are easier than others…

Slide courtesy of Andrew Farrell (University of Utah)

Best case scenario

An error-free sequencing technology



ATTCGAAACA
TTCGCGCAAT
CTGGACTCAA

ATTCGAAACA
TTCGCGCAAT
CTGGACTCAA

Aligner

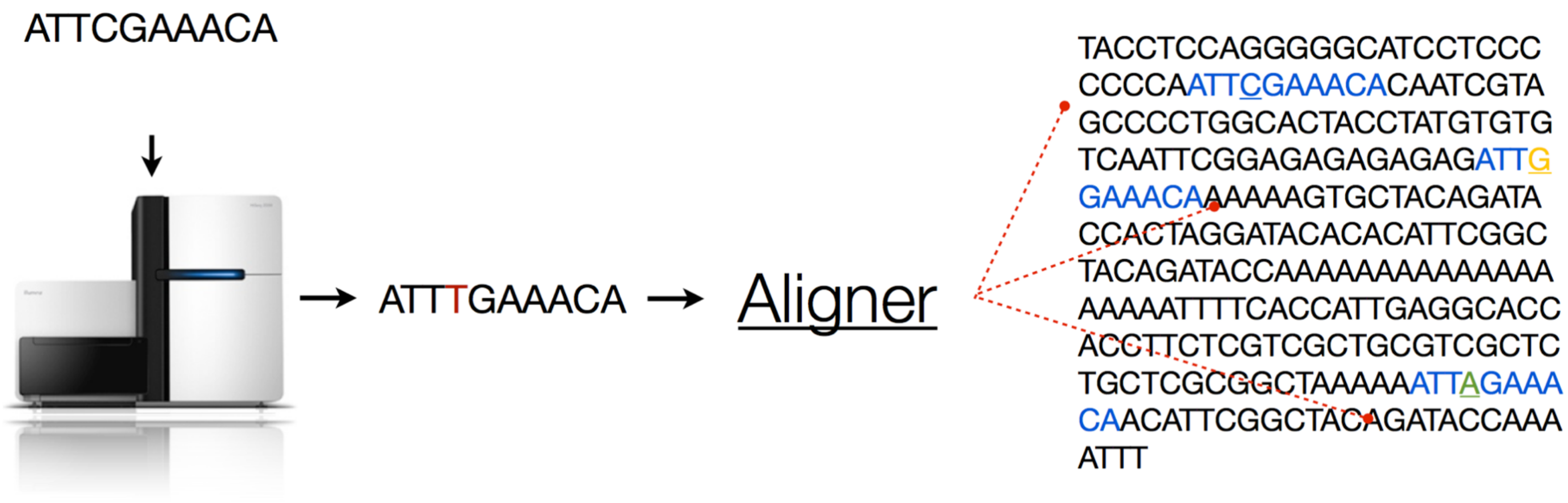
TACCTCCAGGGGGCATCCTCCC
CCCCAATTCGAAACACAATCGTA
GCCCCTGGCACTACCTATGTGTG
TCAATTCGGAGAGAGAGAGATTC
ACGAAAAAAAAGTCTGGACTCAA
CTAGGATACACACATTCGGCTACA
GATACCAAAAAAAAAAAAAAAAAA
ATTTTCACCATTGAGGCACCACCT
TCTCGTCGCTGCGTCGCTCTGCT
CGCTTCGGCTAAAAATTCGCGCA
ATACATTCGGCTACAGATACCAAA
AAAA

Computers are rather good at finding ***exact*** matches.

Slide courtesy of Andrew Farrell (University of Utah)

Reality

Errors happen - frequently; work harder.



ATTCGAAACA

ATTTGAAACA

Aligner

TACCTCCAGGGGGCATCCTCCC
CCCCAATTCGAAACACAATCGTA
GCCCCTGGCACTACCTATGTGTG
TCAATTCGGAGAGAGAGAGATTG
GAAACAAAAAAGTGCTACAGATA
CCACTAGGATACACACATTCGGC
TACAGATACCAAAAAAAAAAAAAA
AAAAATTTTCACCATTGAGGCACC
ACCTTCTCGTCGCTGCGTCGCTC
TGCTCGCGGCTAAAAAATTAGAAA
CAACATTCGGCTACAGATACCAAA
ATTT

“Fuzzy” matching is much more computationally expensive.

Not only errors, but also true variants will differ from the reference

Slide courtesy of Andrew Farrell (University of Utah)

Read alignment to reference

Reference

Read

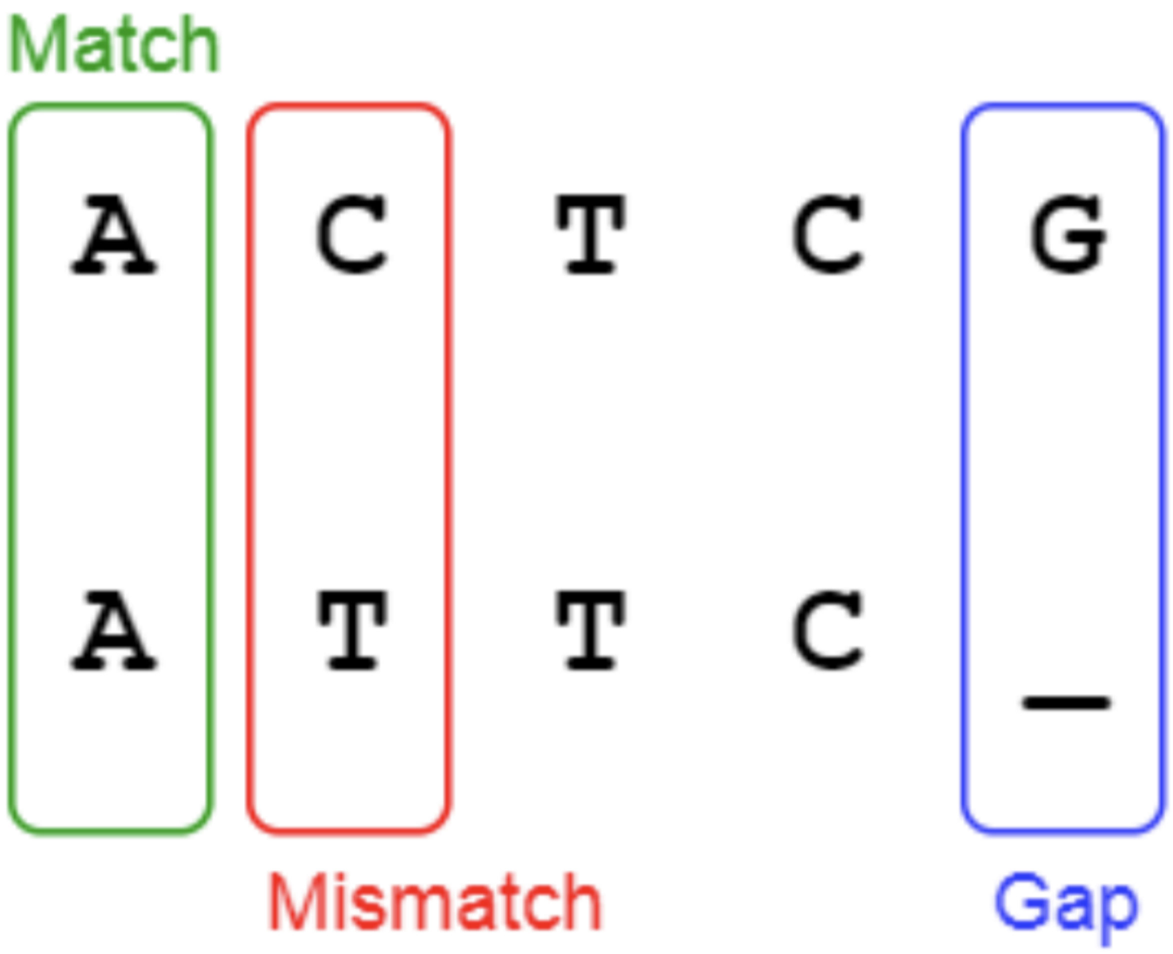


Image: https://bds-au.vlabs.ac.in/exp/pairwise-sequence-alignment/theory.html

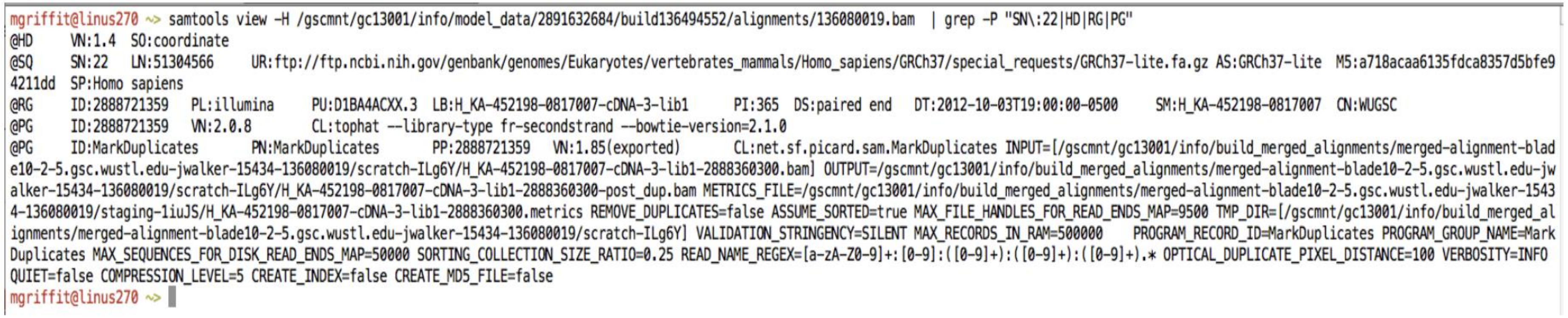
Read alignment algorithms attempt to solve this problem

- There are optimal solutions
  - → Smith-Waterman, Needleman-Wunsch
  - → Computationally expensive (i.e. slow)
- Faster solutions that make some compromises
  - → Hash based solutions
  - → Burrows-Wheeler transform
- Bwa: tool for doing Burrows-Wheeler alignment
  - → bwa-mem: Maximal Exact Match with Burrows-Wheeler, Smith-Waterman for extension and refinement, e.g. around mismatches, insertions, deletions
- Extensive algorithmic details are outside of the scope of this course
- Bottomline: aligners take raw read, determine alignment to reference genome and output a SAM/BAM file

SAM/BAM/CRAM files represent sequence alignments

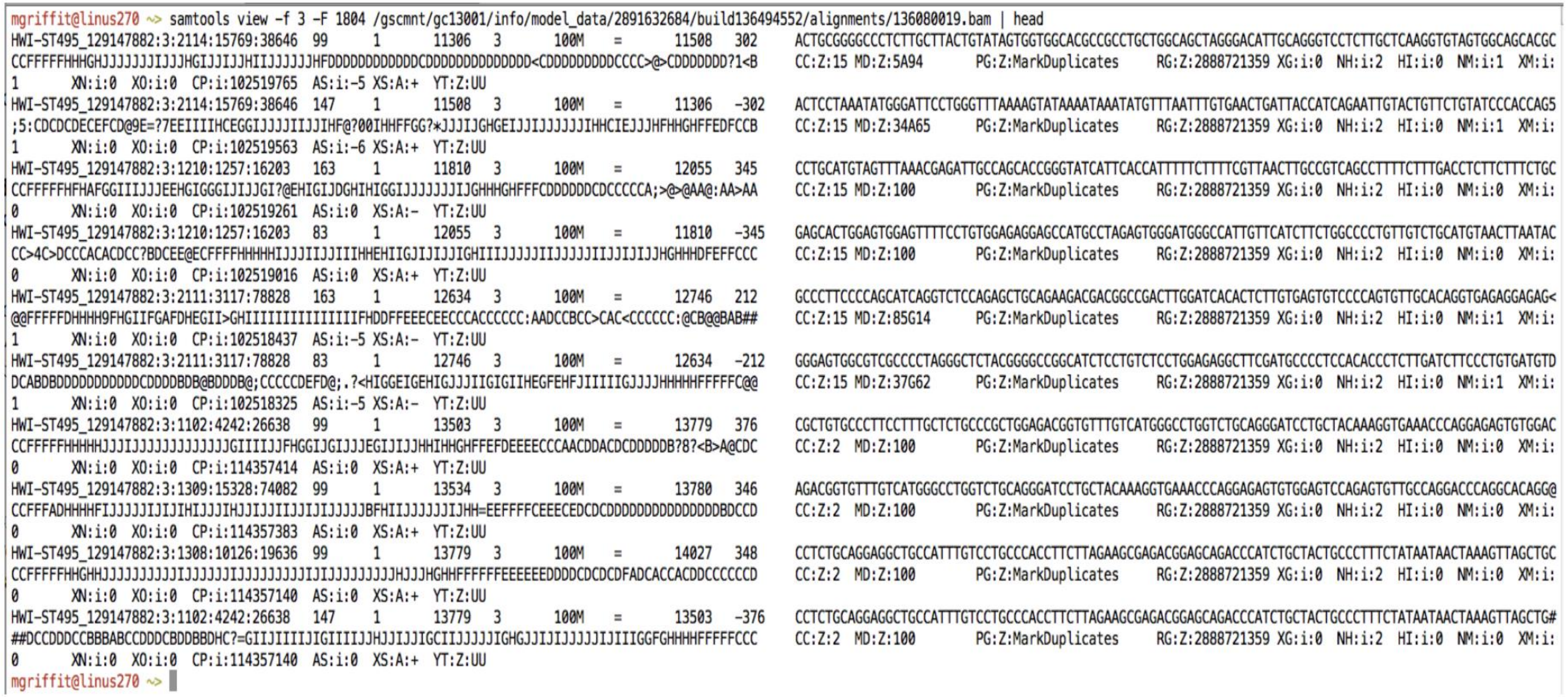
- The specification: http://samtools.sourceforge.net/SAM1.pdf
- The SAM format consists of two sections:
  - → Header section
    - Used to describe source of data, reference sequence, method of alignment, etc.
  - → Alignment section
    - Used to describe the read, quality of the read, and alignment of the read to a region of the genome
- BAM/CRAM are compressed versions of SAM.
  - → BAM compressed using lossless BGZF format
  - → CRAM compressed further using knowledge of reference. May or may not be lossless
- BAM/CRAM files are usually ‘indexed’
  - → A ‘.bai’ file will be found beside the ‘.bam’ file
- Indexing aims to achieve fast retrieval of alignments

Example SAM/BAM header section (abbreviated)



```
mgriffit@linus270 ~> samtools view -H /gscmnt/gc13001/info/model_data/2891632684/build136494552/alignments/136080019.bam  | grep -P "SN\:22|HD|RG|PG"
@HD     VN:1.4  SO:coordinate
@SQ     SN:22   LN:51304566     UR:ftp://ftp.ncbi.nih.gov/genbank/genomes/Eukaryotes/vertebrates_mammals/Homo_sapiens/GRCh37/special_requests/GRCh37-lite.fa.gz AS:GRCh37-lite  M5:a718acaa6135fdca8357d5bfe9
4211dd  SP:Homo sapiens
@RG     ID:2888721359   PL:illumina     PU:D1BA4ACXX.3  LB:H_KA-452198-0817007-cDNA-3-lib1      PI:365  DS:paired end   DT:2012-10-03T19:00:00-0500     SM:H_KA-452198-0817007  CN:WUGSC
@PG     ID:2888721359   VN:2.0.8        CL:tophat --library-type fr-secondstrand --bowtie-version=2.1.0
@PG     ID:MarkDuplicates       PN:MarkDuplicates       PP:2888721359   VN:1.85(exported)       CL:net.sf.picard.sam.MarkDuplicates INPUT=[/gscmnt/gc13001/info/build_merged_alignments/merged-alignment-blad
e10-2-5.gsc.wustl.edu-jwalker-15434-136080019/scratch-ILg6Y/H_KA-452198-0817007-cDNA-3-lib1-2888360300.bam] OUTPUT=/gscmnt/gc13001/info/build_merged_alignments/merged-alignment-blade10-2-5.gsc.wustl.edu-jw
alker-15434-136080019/scratch-ILg6Y/H_KA-452198-0817007-cDNA-3-lib1-2888360300-post_dup.bam METRICS_FILE=/gscmnt/gc13001/info/build_merged_alignments/merged-alignment-blade10-2-5.gsc.wustl.edu-jwalker-1543
4-136080019/staging-1iuJS/H_KA-452198-0817007-cDNA-3-lib1-2888360300.metrics REMOVE_DUPLICATES=false ASSUME_SORTED=true MAX_FILE_HANDLES_FOR_READ_ENDS_MAP=9500 TMP_DIR=[/gscmnt/gc13001/info/build_merged_al
ignments/merged-alignment-blade10-2-5.gsc.wustl.edu-jwalker-15434-136080019/scratch-ILg6Y] VALIDATION_STRINGENCY=SILENT MAX_RECORDS_IN_RAM=500000    PROGRAM_RECORD_ID=MarkDuplicates PROGRAM_GROUP_NAME=Mark
Duplicates MAX_SEQUENCES_FOR_DISK_READ_ENDS_MAP=50000 SORTING_COLLECTION_SIZE_RATIO=0.25 READ_NAME_REGEX=[a-zA-Z0-9]+:[0-9]:([0-9]+):([0-9]+):([0-9]+).* OPTICAL_DUPLICATE_PIXEL_DISTANCE=100 VERBOSITY=INFO
QUIET=false COMPRESSION_LEVEL=5 CREATE_INDEX=false CREATE_MD5_FILE=false
mgriffit@linus270 ~>
```

Example SAM/BAM alignment section (only 10 alignments shown)

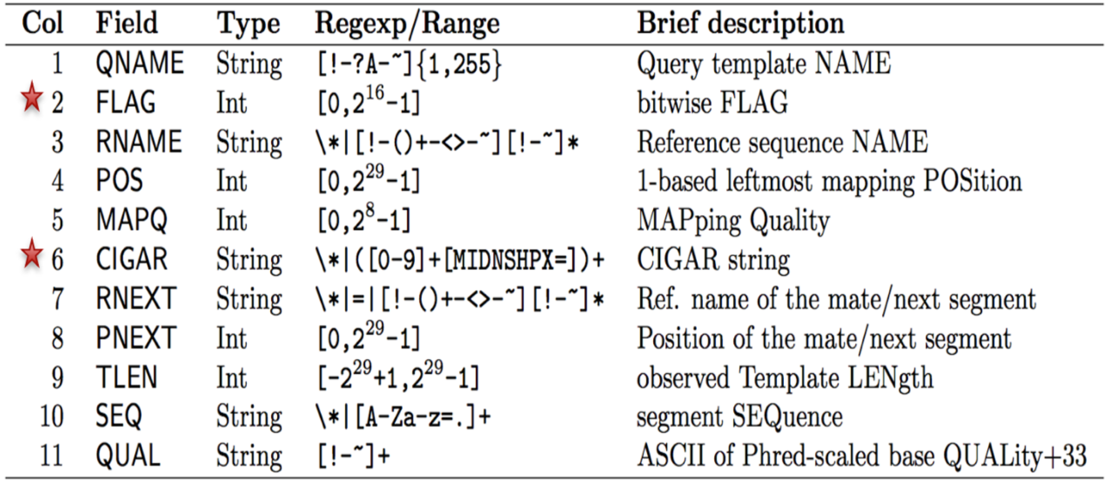


```
mgriffit@linus270 ~> samtools view -f 3 -F 1804 /gscmnt/gc13001/info/model_data/2891632684/build136494552/alignments/136080019.bam | head
HWI-ST495_129147882:3:2114:15769:38646  99      1       11306   3       100M    =       11508   302     ACTGCGGGGCCCTCTTGCTTACTGTATAGTGGTGGCACGCCGCCTGCTGGCAGCTAGGGACATTGCAGGGTCCTCTTGCTCAAGGTGTAGTGGCAGCACGC
CCFFFFFHHHGHJJJJJJJIJJJHGIJJIJJHIIJJJJJJHFDDDDDDDDDDDDCDDDDDDDDDDDDDD<CDDDDDDDDCCCC>@>CDDDDDDD?1<B     CC:Z:15 MD:Z:5A94       PG:Z:MarkDuplicates     RG:Z:2888721359 XG:i:0  NH:i:2  HI:i:0  NM:i:1  XM:i:
1       XN:i:0  XO:i:0  CP:i:102519765  AS:i:-5 XS:A:+  YT:Z:UU
HWI-ST495_129147882:3:2114:15769:38646  147     1       11508   3       100M    =       11306   -302    ACTCCTAAATATGGGATTCCTGGGTTTAAAAGTATAAAATAAATATGTTTAATTTGTGAACTGATTACCATCAGAATTGTACTGTTCTGTATCCCACCAG5
;5:CDCDCDECEFCD@9E=?7EEIIIIHCEGGIJJJJIIJJIHF@?00IHHFFGG?*JJJIJGHGEIJJIJJJJJJIHHCIEJJJHFHHGHFFEDFCCB     CC:Z:15 MD:Z:34A65      PG:Z:MarkDuplicates     RG:Z:2888721359 XG:i:0  NH:i:2  HI:i:0  NM:i:1  XM:i:
1       XN:i:0  XO:i:0  CP:i:102519563  AS:i:-6 XS:A:+  YT:Z:UU
HWI-ST495_129147882:3:1210:1257:16203   163     1       11810   3       100M    =       12055   345     CCTGCATGTAGTTTAAACGAGATTGCCAGCACCGGGTATCATTCACCATTTTTCTTTTCGTTAACTTGCCGTCAGCCTTTTCTTTGACCTCTTCTTTCTGC
CCFFFFFHFHAFGGIIIJJJEEHGIGGGIJIJJGI?@EHIGIJDGHIHIGGIJJJJJJJIJGHHHGHFFFCDDDDDDCDCCCCCA;>@>@AA@:AA>AA     CC:Z:15 MD:Z:100        PG:Z:MarkDuplicates     RG:Z:2888721359 XG:i:0  NH:i:2  HI:i:0  NM:i:0  XM:i:
0       XN:i:0  XO:i:0  CP:i:102519261  AS:i:0  XS:A:-  YT:Z:UU
HWI-ST495_129147882:3:1210:1257:16203   83      1       12055   3       100M    =       11810   -345    GAGCACTGGAGTGGAGTTTTCCTGTGGAGAGGAGCCATGCCTAGAGTGGGATGGGCCATTGTTCATCTTCTGGCCCCTGTTGTCTGCATGTAACTTAATAC
CC>4C>DCCCACACDCC?BDCEE@ECFFFFHHHHHIJJJIIJJIIIHHEHIIGJIJIJJIGHIIIJJJJJIIJJJJJIIJJIJIJJHGHHHDFEFFCCC     CC:Z:15 MD:Z:100        PG:Z:MarkDuplicates     RG:Z:2888721359 XG:i:0  NH:i:2  HI:i:0  NM:i:0  XM:i:
0       XN:i:0  XO:i:0  CP:i:102519016  AS:i:0  XS:A:+  YT:Z:UU
HWI-ST495_129147882:3:2111:3117:78828   163     1       12634   3       100M    =       12746   212     GCCCTTCCCCAGCATCAGGTCTCCAGAGCTGCAGAAGACGACGGCCGACTTGGATCACACTCTTGTGAGTGTCCCCAGTGTTGCACAGGTGAGAGGAGAG<
@@FFFFFDHHHH9FHGIIFGAFDHEGII>GHIIIIIIIIIIIIIIIFHDDFFEEECEECCCACCCCCC:AADCCBCC>CAC<CCCCCC:@CB@@BAB##     CC:Z:15 MD:Z:85G14      PG:Z:MarkDuplicates     RG:Z:2888721359 XG:i:0  NH:i:2  HI:i:0  NM:i:1  XM:i:
1       XN:i:0  XO:i:0  CP:i:102518437  AS:i:-5 XS:A:-  YT:Z:UU
HWI-ST495_129147882:3:2111:3117:78828   83      1       12746   3       100M    =       12634   -212    GGGAGTGGCGTCGCCCCTAGGGCTCTACGGGGCCGGCATCTCCTGTCTCCTGGAGAGGCTTCGATGCCCCTCCACACCCTCTTGATCTTCCCTGTGATGTD
DCABDBDDDDDDDDDDDDCDDDDBDB@BDDDB@;CCCCCDEFD@;.?<HIGGEIGEHIGJJJIIGIGIIHEGFEHFJIIIIIGJJJJHHHHHFFFFFC@@     CC:Z:15 MD:Z:37G62      PG:Z:MarkDuplicates     RG:Z:2888721359 XG:i:0  NH:i:2  HI:i:0  NM:i:1  XM:i:
1       XN:i:0  XO:i:0  CP:i:102518325  AS:i:-5 XS:A:-  YT:Z:UU
HWI-ST495_129147882:3:1102:4242:26638   99      1       13503   3       100M    =       13779   376     CGCTGTGCCCTTCCTTTGCTCTGCCCGCTGGAGACGGTGTTTGTCATGGGCCTGGTCTGCAGGGATCCTGCTACAAAGGTGAAACCCAGGAGAGTGTGGAC
CCFFFFFHHHHHJJJIJJJJJJJJJJJJJGIIIIJJFHGGIJGIJJJEGIJIJJHHIHHGHFFEFDEEEECCCAACDDACDCDDDDDB?8?<B>A@CDC     CC:Z:2  MD:Z:100        PG:Z:MarkDuplicates     RG:Z:2888721359 XG:i:0  NH:i:2  HI:i:0  NM:i:0  XM:i:
0       XN:i:0  XO:i:0  CP:i:114357414  AS:i:0  XS:A:+  YT:Z:UU
HWI-ST495_129147882:3:1309:15328:74082  99      1       13534   3       100M    =       13780   346     AGACGGTGTTTGTCATGGGCCTGGTCTGCAGGGATCCTGCTACAAAGGTGAAACCCAGGAGAGTGTGGAGTCCAGAGTGTTGCCAGGACCCAGGCACAGG@
CCFFFADHHHHFIJJJJJIJIJIHIJJJIHJJIJJIIJJIJIJJJJJBFHIIJJJJJJIJHH=EEFFFFCEEECEDCDCDDDDDDDDDDDDDDDBDCCD     CC:Z:2  MD:Z:100        PG:Z:MarkDuplicates     RG:Z:2888721359 XG:i:0  NH:i:2  HI:i:0  NM:i:0  XM:i:
0       XN:i:0  XO:i:0  CP:i:114357383  AS:i:0  XS:A:+  YT:Z:UU
HWI-ST495_129147882:3:1308:10126:19636  99      1       13779   3       100M    =       14027   348     CCTCTGCAGGAGGCTGCCATTTGTCCTGCCCACCTTCTTAGAAGCGAGACGGAGCAGACCCATCTGCTACTGCCCTTTCTATAATAACTAAAGTTAGCTGC
CCFFFFFHHGHHJJJJJJJJJJIJJJJJJIJJJJJJJJJIJIJJJJJJJJJHJJJHGHHFFFFFFEEEEEEDDDDCDCDCDFADCACCACDDCCCCCCD     CC:Z:2  MD:Z:100        PG:Z:MarkDuplicates     RG:Z:2888721359 XG:i:0  NH:i:2  HI:i:0  NM:i:0  XM:i:
0       XN:i:0  XO:i:0  CP:i:114357140  AS:i:0  XS:A:+  YT:Z:UU
HWI-ST495_129147882:3:1102:4242:26638   147     1       13779   3       100M    =       13503   -376    CCTCTGCAGGAGGCTGCCATTTGTCCTGCCCACCTTCTTAGAAGCGAGACGGAGCAGACCCATCTGCTACTGCCCTTTCTATAATAACTAAAGTTAGCTG#
##DCCDDDCCBBBABCCDDDCBDDBBDHC?=GIIJIIIIJIGIIIIJJHJJIJJIGCIIJJJJJIGHGJJIJIJJJJJIJIIIGGFGHHHHFFFFFCCC     CC:Z:2  MD:Z:100        PG:Z:MarkDuplicates     RG:Z:2888721359 XG:i:0  NH:i:2  HI:i:0  NM:i:0  XM:i:
0       XN:i:0  XO:i:0  CP:i:114357140  AS:i:0  XS:A:+  YT:Z:UU
mgriffit@linus270 ~>
```

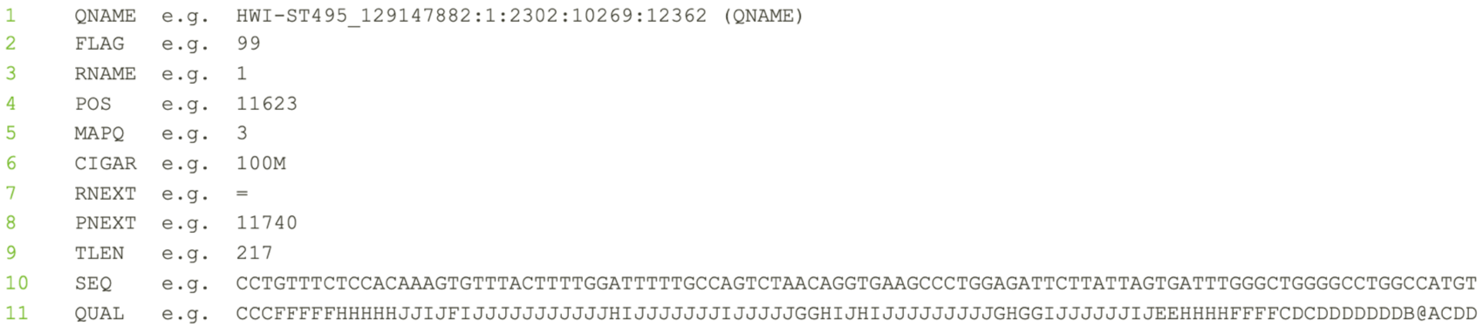
BAM header section provides general information about alignment strategy

- Used to describe source of data, reference sequence, method of alignment, etc.
- Each section begins with character '@' followed by a two-letter record type code. These are followed by two-letter tags and values
- @HD The header line
  - → VN: format version
  - → SO: Sorting order of alignments
- @SQ Reference sequence dictionary
  - → SN: reference sequence name
  - → LN: reference sequence length SP: species
- @RG Read group
  - → ID: read group identifier
  - → CN: name of sequencing center
  - → SM: sample name
- @PG Program
  - → PN: program name
  - → VN: program version

BAM alignment section provides details for each read alignment



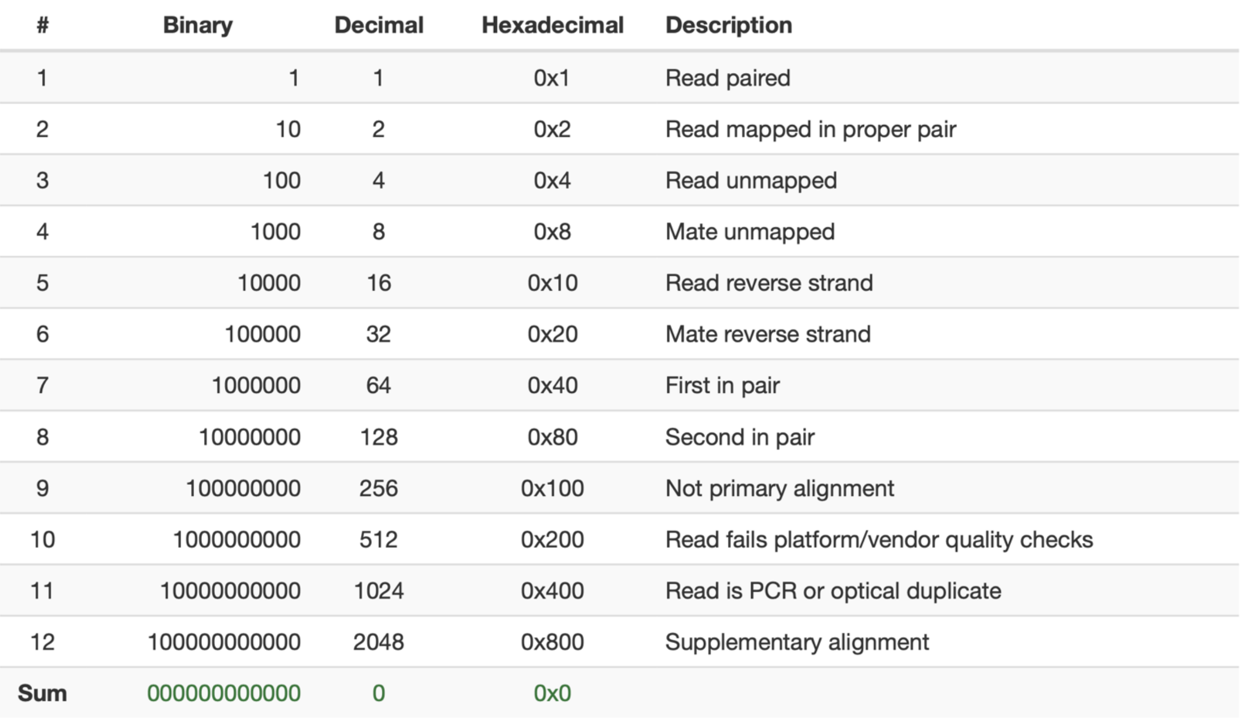
Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,255}`	Query template NAME
★ 2	FLAG	Int	$[0,2^{16}-1]$	bitwise FLAG
3	RNAME	String	`\*|[!-()+-<>-~][!-~]*`	Reference sequence NAME
4	POS	Int	$[0,2^{29}-1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0,2^{8}-1]$	MAPping Quality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



```
1      QNAME   e.g.  HWI-ST495_129147882:1:2302:10269:12362 (QNAME)
2      FLAG    e.g.  99
3      RNAME   e.g.  1
4      POS     e.g.  11623
5      MAPQ    e.g.  3
6      CIGAR   e.g.  100M
7      RNEXT   e.g.  =
8      PNEXT   e.g.  11740
9      TLEN    e.g.  217
10     SEQ     e.g.  CCTGTTTCTCCACAAAGTGTTTACTTTTGGATTTTTGCCAGTCTAACAGGTGAAGCCCTGGAGATTCTTATTAGTGATTTGGGCTGGGGCCTGGCCATGT
11     QUAL    e.g.  CCCFFFFFHHHHHJJIJFIJJJJJJJJJJJHIJJJJJJJIJJJJJGGHIJHIJJJJJJJJJGHGGIJJJJJJIJEEHHHHFFFFCDCDDDDDDDB@ACDD
```

SAM flags describe several alignment properties in a single number

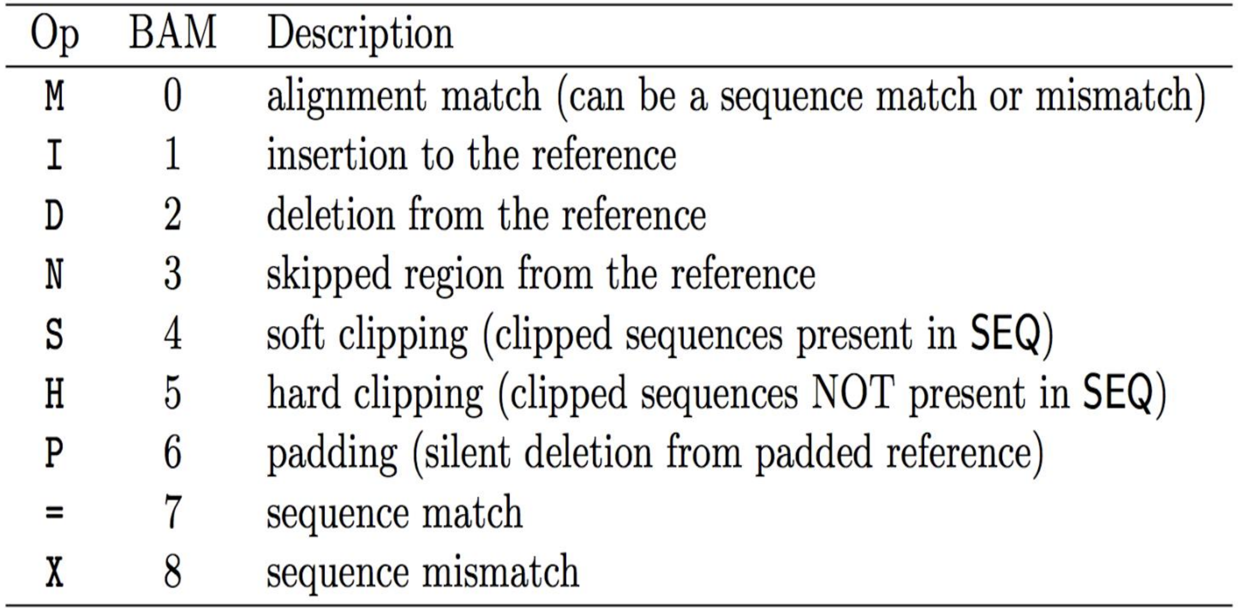
- http://broadinstitute.github.io/picard/explain-flags.html
- 12 bitwise flags describing the alignment
- These flags are stored as a binary string of length 11 instead of 11 columns of data
- Value of '1' indicates the flag is set. e.g. 00100000000
- All combinations can be represented as a number from 1 to 2048 (i.e. 2^{11}–1). This number is used in the BAM/SAM file. You can specify 'required' or 'filter' flags in samtools view using the '-f' and '-F' options respectively



#	Binary	Decimal	Hexadecimal	Description
1	1	1	0x1	Read paired
2	10	2	0x2	Read mapped in proper pair
3	100	4	0x4	Read unmapped
4	1000	8	0x8	Mate unmapped
5	10000	16	0x10	Read reverse strand
6	100000	32	0x20	Mate reverse strand
7	1000000	64	0x40	First in pair
8	10000000	128	0x80	Second in pair
9	100000000	256	0x100	Not primary alignment
10	1000000000	512	0x200	Read fails platform/vendor quality checks
11	10000000000	1024	0x400	Read is PCR or optical duplicate
12	100000000000	2048	0x800	Supplementary alignment
Sum	000000000000	0	0x0	

CIGAR strings similarly describe the entire alignment in as few characters as possible

- The CIGAR string is a sequence of base lengths and associated ‘operations’ that are used to indicate which bases align to the reference (either a match or mismatch), are deleted, are inserted, represent introns, etc.
- e.g. 81M3D19M
  → A 100 bp read consists of: 81 bases of alignment to reference (**m**atch), 3 bases of the reference **d**eleted, 19 bases of alignment (**m**atch)



Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

Deduplication

- Each DNA molecule copied many times during library preparation
- Many reads are duplicates, representing the same original DNA molecule
- We want unique information for downstream processing
- Therefore: deduplication
- Picard MarkDuplicates:
  - → Group reads by start positions of read 1 and read 2
  - → Take the read with highest summed base quality score as the unique read
- Simplest form of deduplication

Alignment quiz

- Menti.com

QC metrics for DNA sequencing

Outline

- Metrics
- Tools
- Correlations
    - Plots
- Quiz

Learning outcomes and course content

- Learning outcomes:
  - Understand how to apply technology to obtain relevant information from the cancer genome.
  - Perform quality control on DNA (and RNA) sequencing data for cancer sequencing purposes.

- Course content:
  - QC of both DNA (and RNA) sequencing data

- Focus on DNA here, RNA QC is covered in RNA lecture

Quality control metrics

- Was the sequencing successful or not?
- Many steps can go wrong
  - → Storage and transportation
  - → Extraction of DNA, DNA input amount
  - → Library prep, e.g. PCR amplification
  - → Capture
  - → Sequencing
  - → Demultiplexing
- Important with quality control metrics
- Is data quality “good enough”?
  - → Requirements can vary a lot depending on the experiment

Quality control metrics

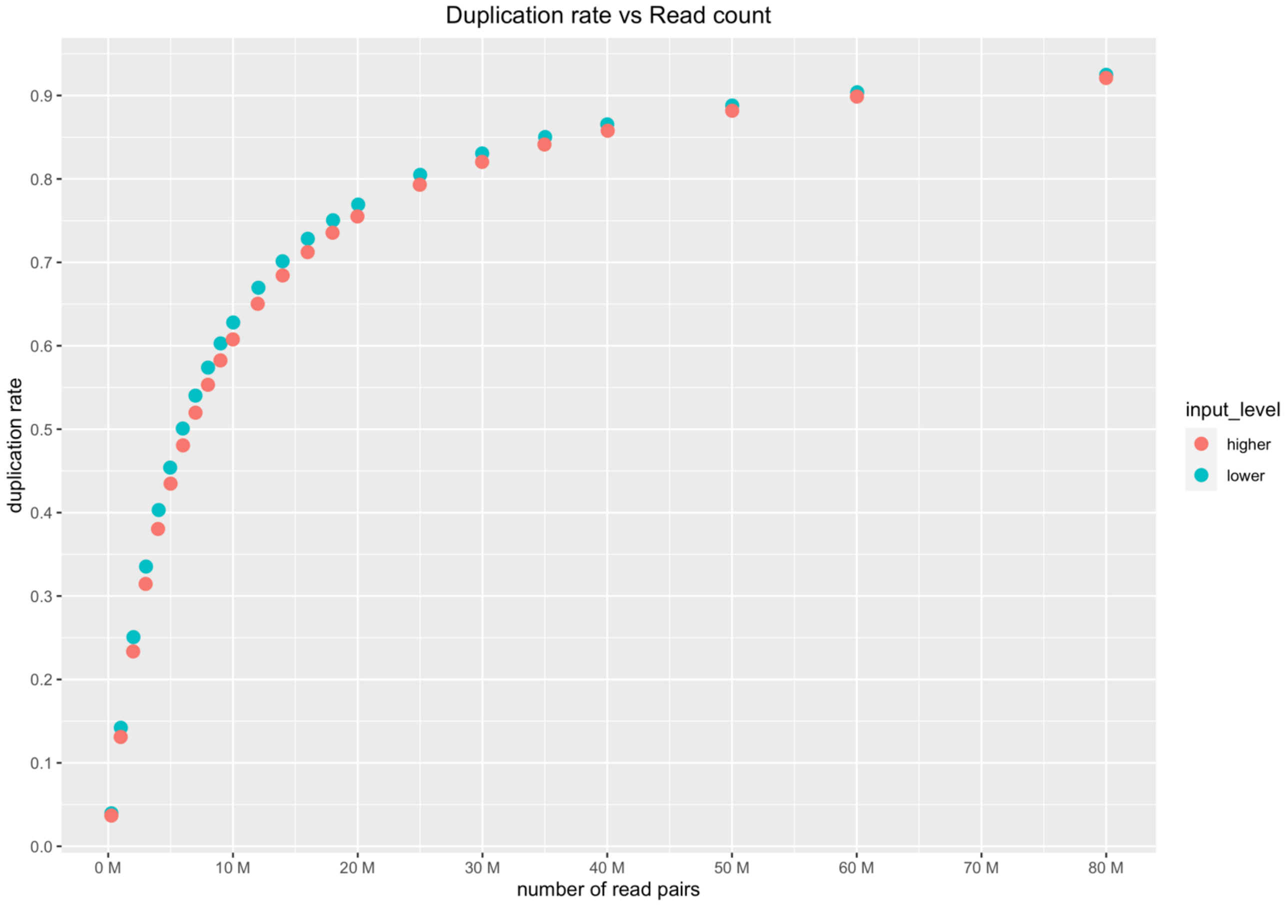
- The most important metrics
  - → **Coverage** (after deduplication) – the average number of (unique) reads covering the targeted regions, also known as depth
  - → **Read count**– total number of reads for a sample
  - → **Duplication rate** – what fraction of all reads where duplicates (not unique)
  - → **Fold enrichment** – how much more the targeted regions are amplified compared to non-targeted regions, x-fold
  - → **Contamination** – DNA from another source

Quality control metrics

- Example of tools for QC
  - → Picard CollectHsMetrics
    - Coverage, fold enrichment
  - → Picard MarkDuplicates
    - Read count, duplication rate
  - → GATK (v ≤3) ContEst
    - Contamination
- Input: bam files
- Output: txt tables,
    - can be parsed in e.g. R for plotting, summary tables etc.

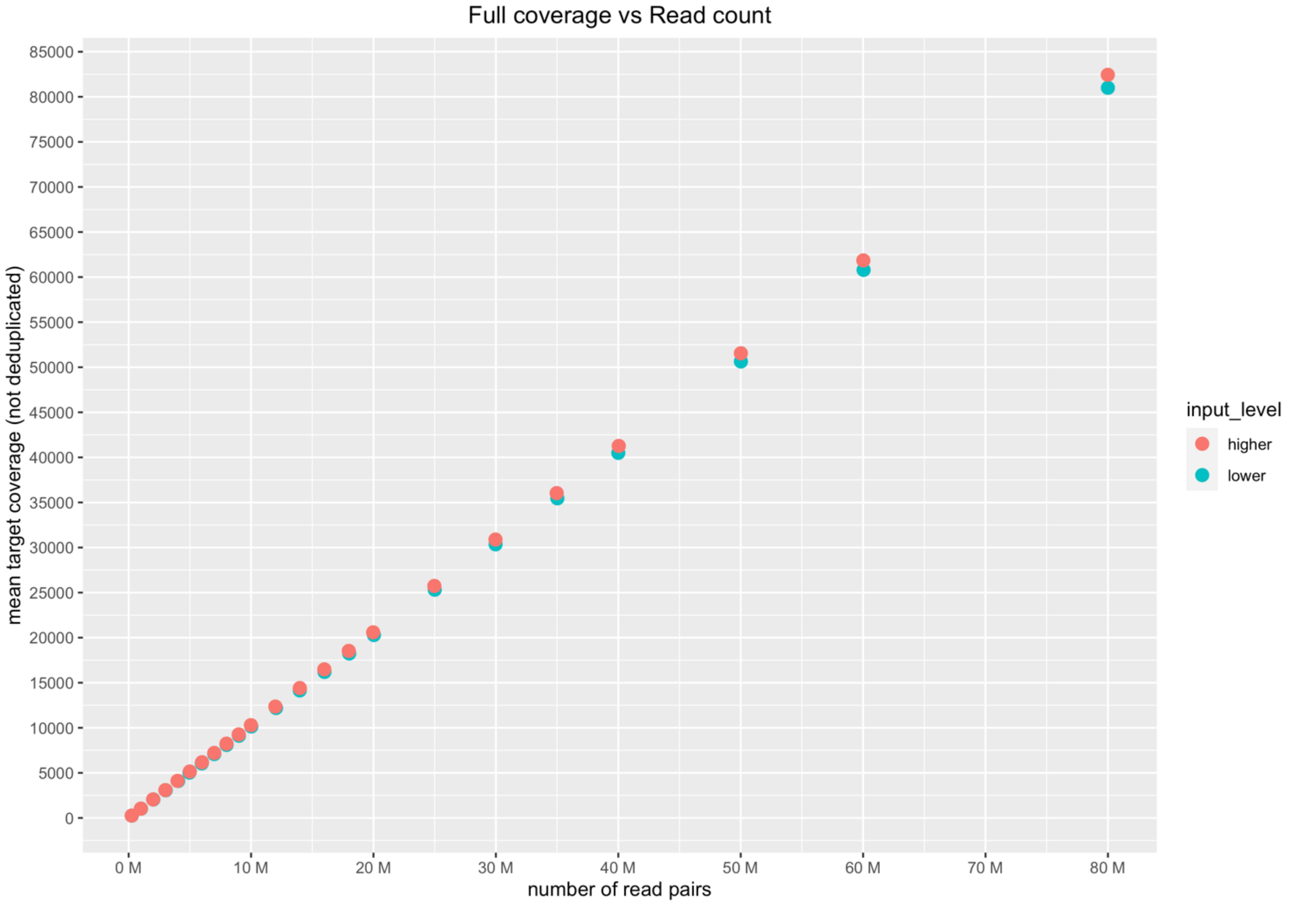
Duplication rate vs read count

- Increasing number of reads give increasing duplication rate
- Small difference in DNA input amount (few ng)
- Higher input amount gives lower duplication rate



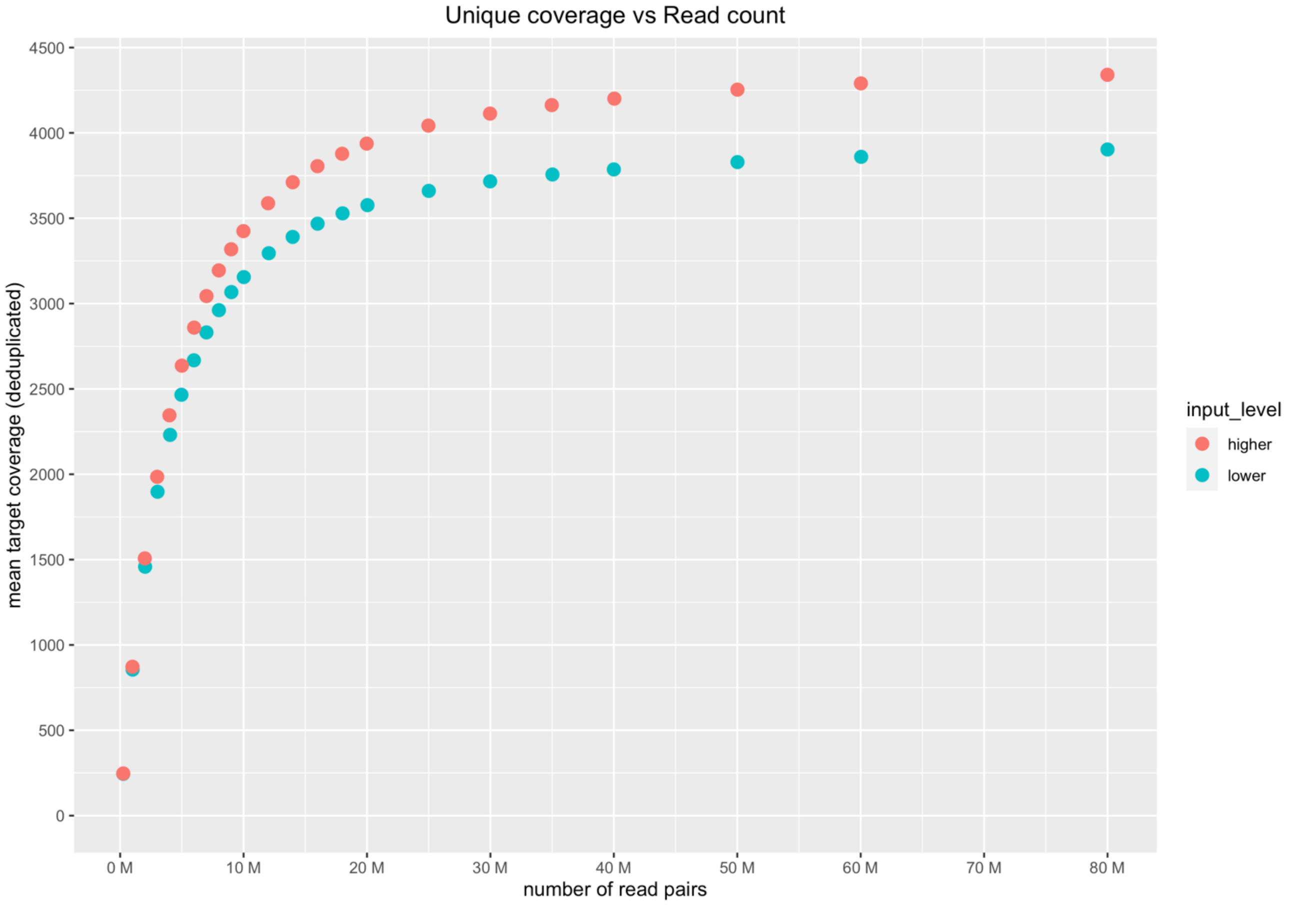
Coverage vs read count

- Including duplicates
  - Not de-duplicated
- Increasing number of reads give higher coverage
- No difference by input amount



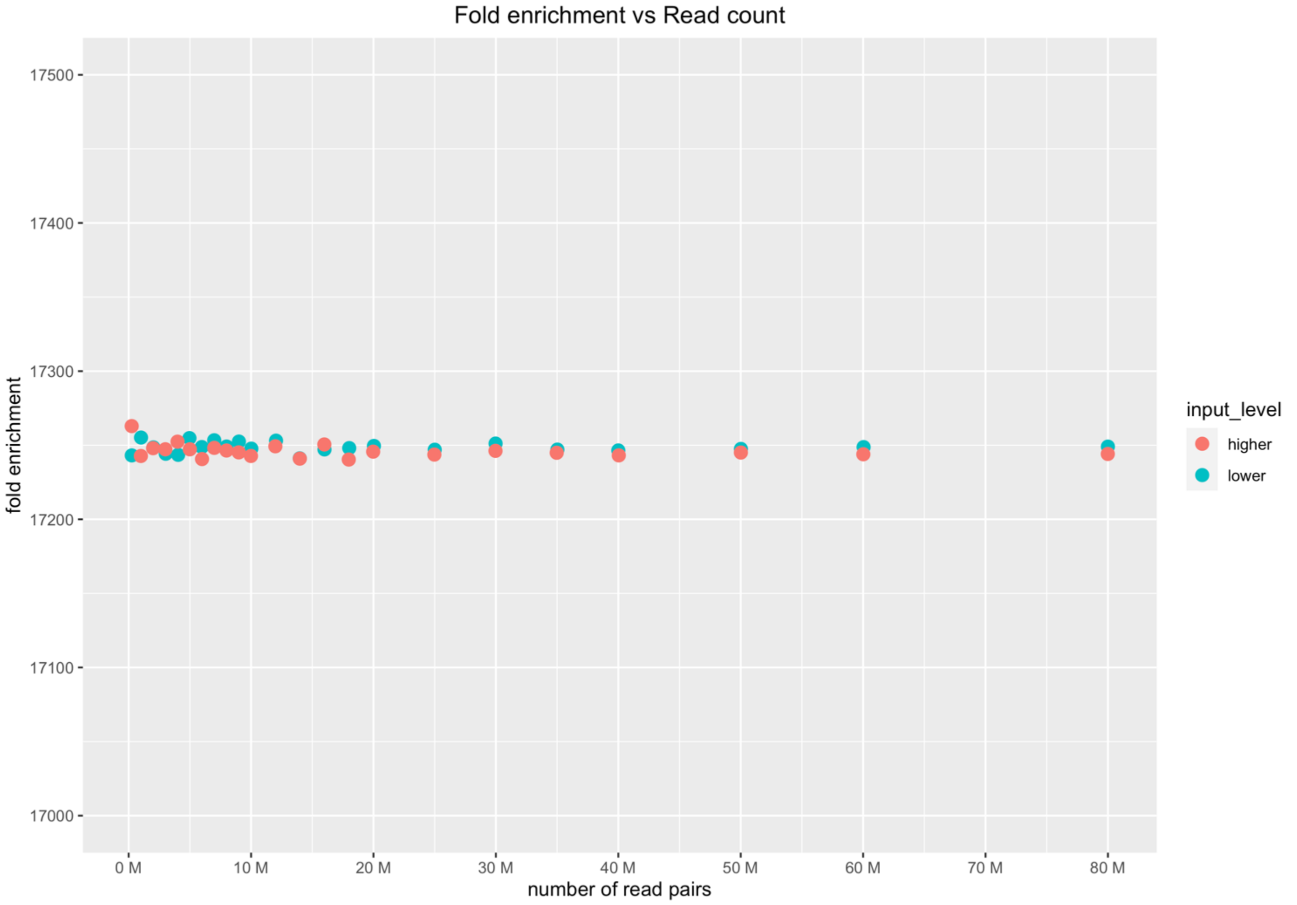
Coverage vs read count

- Unique
  - De-duplicated
- Increasing number of reads give higher coverage
  - Up until a certain level
- Higher input amount gives higher unique coverage
- The max coverage possible depends on input amount
- When max coverage is reached, all unique DNA molecules have been sequenced and all additional reads will be duplicates



Fold enrichment vs read count

- Before de-duplication
- Independent of read count and input amount
- Related to the size of targeted regions
- What could be the cause of a deviating number?



Quality control quiz

- Menti.com

Credits

- Malachi Griffith, Obi Griffith, Zachary Skidmore, Huiming Xia
  - Lecture notes from the course “Introduction to bioinformatics for DNA and RNA sequence analysis (IBDR01)”, 29 October – 2 November, 2018
  - McDonell Genome Institute, Washington University of St Louis School of Medicine