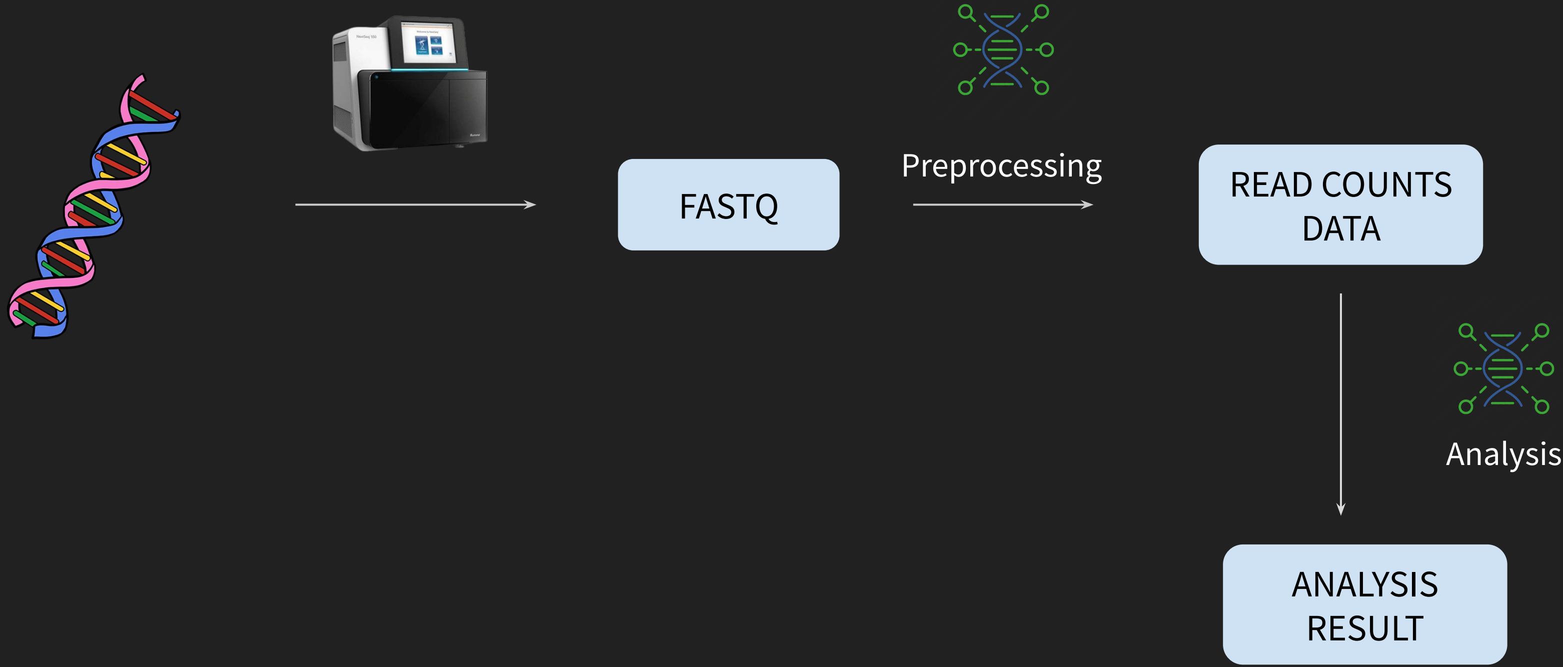
Introduction to Bioinformatics

More on UNIX and Bioinformatics Data Formats

Credit to https://mdozmorov.github.io/ for this session material

A Quick Glimpse

NGS (Bisulfite Sequencing)



FASTQ

Preprocessing

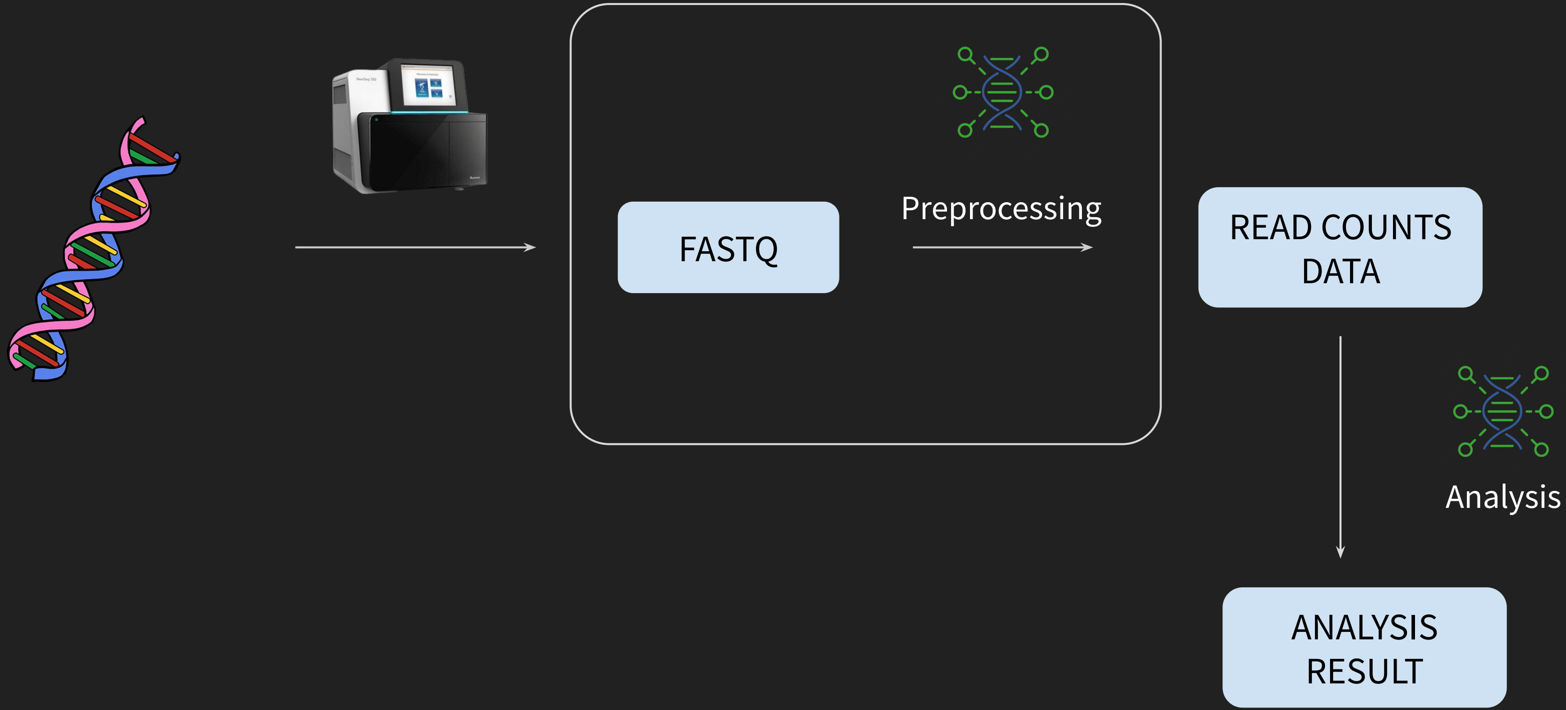
READ COUNTS DATA

Analysis

ANALYSIS RESULT

A Quick Glimpse

NGS (Bisulfite Sequencing)



FASTQ

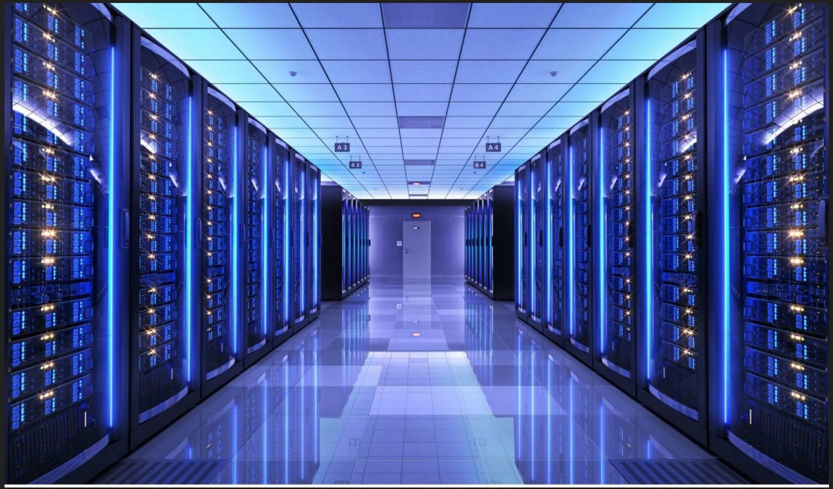
Preprocessing

READ COUNTS DATA

Analysis

ANALYSIS RESULT

Why do we need to learn UNIX?

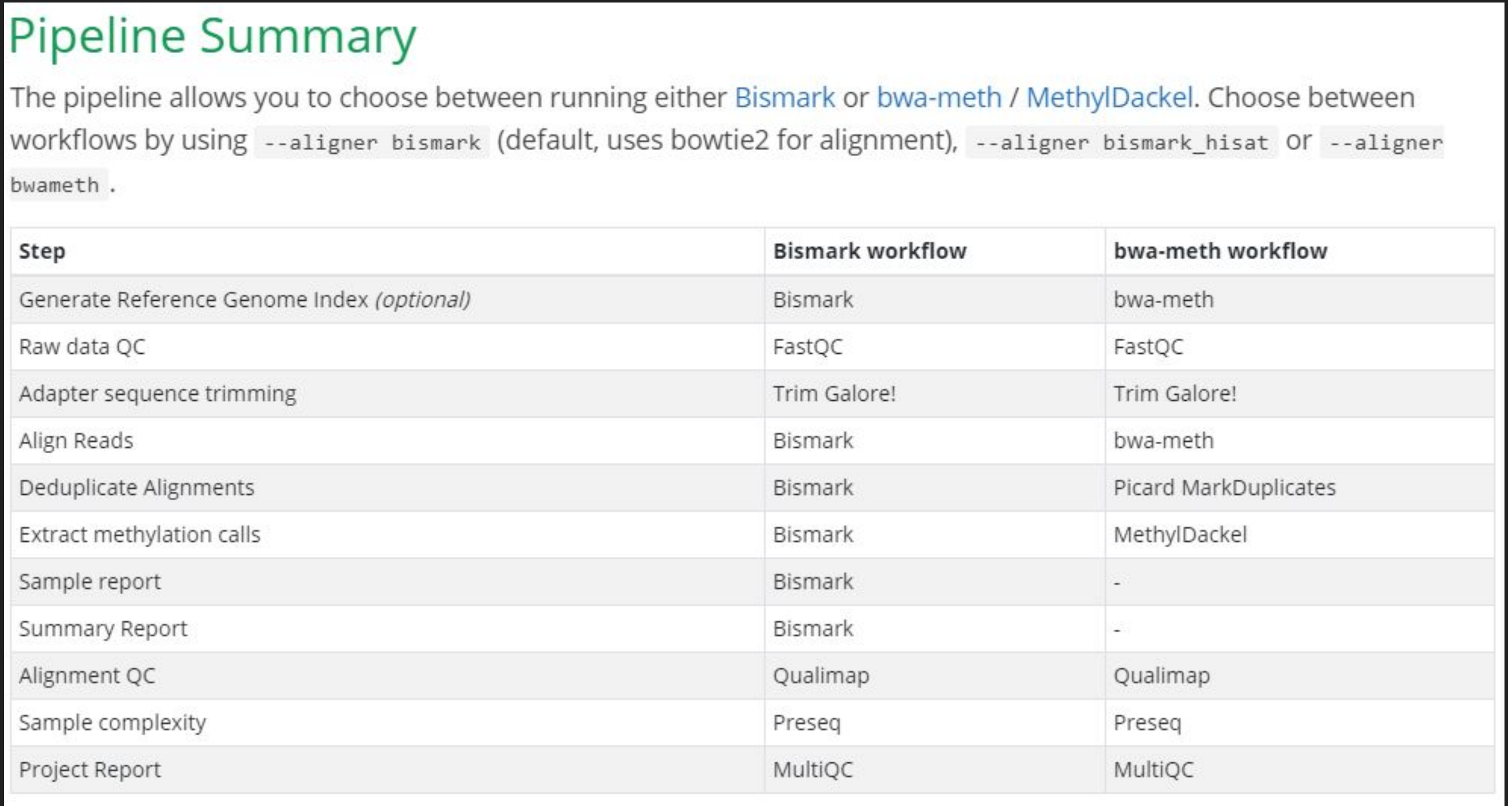


- Most of the software in the preprocessing part for NGS analysis is only available and can be run in UNIX system only
- Most high performance computing is using UNIX

What are we doing and what are we processing?

- During the preprocessing part we will preprocess the output from the sequencing machine
- The output is mostly looking like DNA sequence since we’re sequencing DNA
- Our job here is to convert this data so it’d be usable for analysis

Preprocessing Pipeline using MethylSeq



Pipeline Summary

The pipeline allows you to choose between running either Bismark or bwa-meth / MethylDackel. Choose between workflows by using `--aligner bismark` (default, uses bowtie2 for alignment), `--aligner bismark_hisat` or `--aligner bwameth`.

Step	Bismark workflow	bwa-meth workflow
Generate Reference Genome Index *(optional)*	Bismark	bwa-meth
Raw data QC	FastQC	FastQC
Adapter sequence trimming	Trim Galore!	Trim Galore!
Align Reads	Bismark	bwa-meth
Deduplicate Alignments	Bismark	Picard MarkDuplicates
Extract methylation calls	Bismark	MethylDackel
Sample report	Bismark	-
Summary Report	Bismark	-
Alignment QC	Qualimap	Qualimap
Sample complexity	Preseq	Preseq
Project Report	MultiQC	MultiQC

File Types

- Plain text file formats
  - Information often structured into lines and columns
  - Human-readable
  - Easy to process

- Binary file formats
  - Not human-readable
  - Require special software for processing
  - Efficient storage
  - (significant) reduction to file size when compared to a plain text counterpart (e.g. 75 % space saved)

Common File Formats that You Will Encounter

- **FASTA** - Simple collections of named DNA/protein sequences (text)
- **FASTQ** - Extension of FASTA format, contains additional quality information. Widely used for storing unaligned sequencing reads (text)
- **SAM/BAM** - Alignments of sequencing reads to a reference genome (text/binary)
- **BED** - Region-based genome annotation information (e.g. a list of genes and their genomic locations).
- **GFF/GTF** - gene-centric annotations (text)
- **VCF** - variant call format, to store information about genomic variants (text)
- **CSV/TSV** - Usually stores read counts/expression information per sample

FASTA format

The nucleic acid codes that can be found in FASTA file:

A --> adenosine
T --> thymidine
C --> cytidine
S --> G C (strong)
G --> guanine

W --> A T (weak)
B --> G T C
U --> uridine
N --> A G C T
R --> G A (purine)
Y --> T C (pyrimidine)

Example of fasta format http://hgdownload.cse.ucsc.edu/goldenPath/hg38/chromosomes/

Quick UNIX Check

- **How long is chrY?**

  -
    ```
    $ grep -v ">" hg38.chrY.fa | grep -o "[ATCGatcg]" | wc -l 26415043
    ```
- **How many adenosines are there? $**

  -
    ```
    $ grep -v ">" hg38.chrY.fa | grep -o -i "A" | wc -l 7886192
    ```

FASTQ format

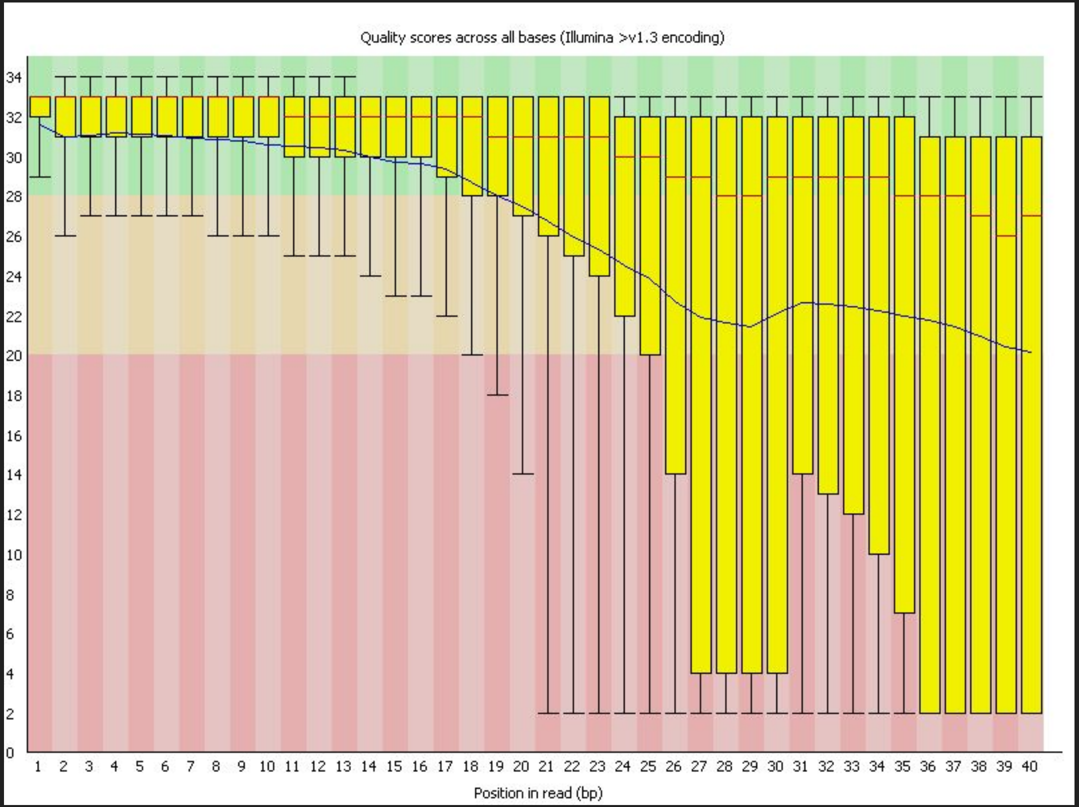
- Nearly all sequencing technologies produce sequencing reads in FASTQ format
  - **Sequence ID** @SEQ_ID
  - **Sequence** GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
  - **Separator** +
  - **Quality scores** !''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>

FASTQ Quality Scores (Phred Scores)

- PHRED Base quality (Q) – integer value derived from the estimated probability (P) of the corresponding base being determined wrong
  - $Q = -10 * \log10(Perr)$ (rounded to nearest integer)

- PHRED Base quality (Q) – integer value derived from the estimated probability (P) of the corresponding base being determined wrong A higher quality score is better (>=20 is considered “good”)
  - Score of 10 means 10% of probability of it’s being error
  - Score of 20 means 1%
  - Score of 30 means 0.1% etc

FastQC Helps Quality Control



More information on interpreting:

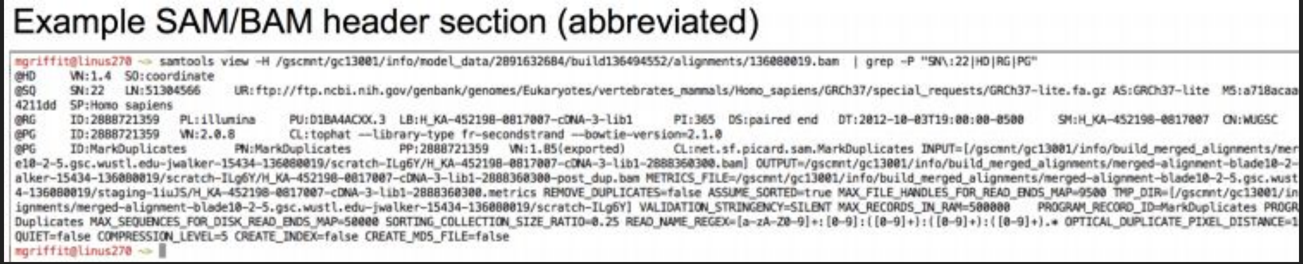
https://hbctraining.github.io/Intro-to-rnaseq-hpc-salmon/lessons/qc_fastqc_assessment.html

Sequence Alignment Map (SAM)

- Intended for storing read alignments against reference sequences
- Has a binary version with good software support (BAM format)

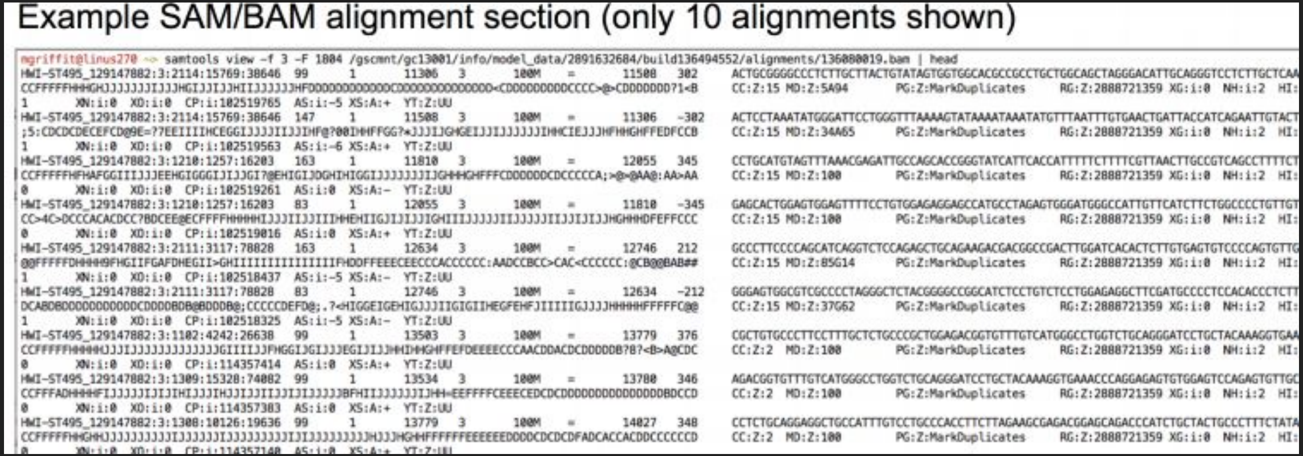
- The SAM format consists of two sections:
  - Header section Used to describe source of data, reference sequence, method of alignment, etc.
  - Alignment section Used to describe the read, quality of the read, and nature alignment of the read to a region of the genome

Sequence Alignment Map (SAM)



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:a718acaa
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/mer
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-
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.wust
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/in
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 PROGR
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=1
QUIET=false COMPRESSION_LEVEL=5 CREATE_INDEX=false CREATE_MD5_FILE=false
mgriffit@linus270 ~> 
```



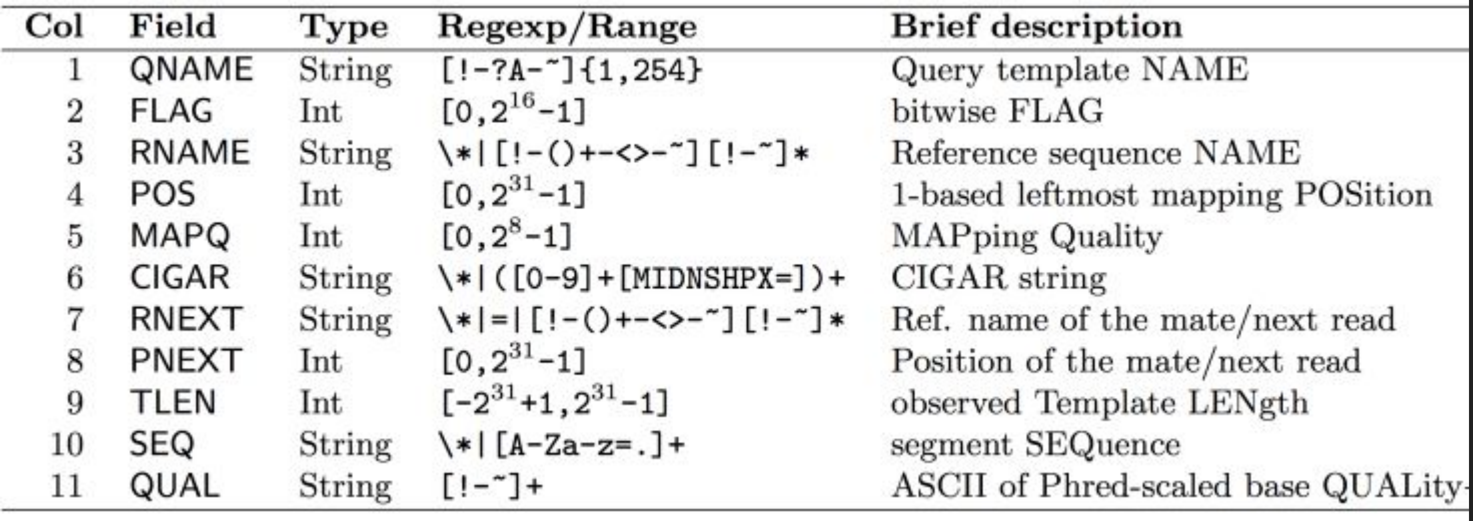
Example SAM/BAM alignment section (only 10 alignments shown)

```
mgriffit@linus270 ~> samtools view -f 3 -F 1004 /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     ACTGCGGGGCCCTCTTGCTTACTGTATAGTGGTGGCACGCCGCCTGCTGGCAGCTAGGGACATTGCAGGGTCCTCTTGCTCAA
CCFFFFFHHHGHJJJJJJJJIJJJHGIJJIJJHIIJJJJJJHFDDDDDDDDDDDDCDDDDDDDDDDDDDD<CDDDDDDDDDCCCC>@>CDDDDDDD?1<B     CC:Z:15 MD:Z:5A94       PG:Z:MarkDuplicates     RG:Z:2888721359 XG:i:0  NH:i:2  HI:
1       XN:i:0  XD: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    ACTCCTAAATATGGGATTCCTGGGTTTAAAAGTATAAAATAAATATGTTTAATTTGTGAACTGATTACCATCAGAATTGTACT
;5:CDCDCDECEFCD@9E=??EEIIIIHCEGGIJJJJIIJJIHF@?00IHHFFGG?*JJJIJGHGEIJJIJJJJJJIHHCIEJJJHFHHGHFFEDFCCB     CC:Z:15 MD:Z:34A65      PG:Z:MarkDuplicates     RG:Z:2888721359 XG:i:0  NH:i:2  HI:
1       XN:i:0  XD: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     CCTGCATGTAGTTTAAACGAGATTGCCAGCACCGGGTATCATTCACCATTTTTCTTTTCGTTAACTTGCCGTCAGCCTTTTCT
CCFFFFFHFHAFGGIIIJJJEEHGIGGGIJIJJGI?@EHIGIJDGHIHIGGIJJJJJJJJIJGHHHGHFFFCDDDDDDCDCCCCCA;>@>@AA@:AA>AA     CC:Z:15 MD:Z:100        PG:Z:MarkDuplicates     RG:Z:2888721359 XG:i:0  NH:i:2  HI:
0       XN:i:0  XD: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    GAGCACTGGAGTGGAGTTTTCCTGTGGAGAGGAGCCATGCCTAGAGTGGGATGGGCCATTGTTCATCTTCTGGCCCCTGTTGT
CC>4C>DCCCACACDCC?BDCEE@ECFFFFHHHHHIJJJIIJJIIIHHEHIIGJIJIJJIGHIIIJJJJJIIJJJJJIIJJIJIJJHGHHHDFEFFCCC     CC:Z:15 MD:Z:100        PG:Z:MarkDuplicates     RG:Z:2888721359 XG:i:0  NH:i:2  HI:
0       XN:i:0  XD: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     GCCCTTCCCCAGCATCAGGTCTCCAGAGCTGCAGAAGACGACGGCCGACTTGGATCACACTCTTGTGAGTGTCCCCAGTGTTG
@@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:
1       XN:i:0  XD: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    GGGAGTGGCGTCGCCCCTAGGGCTCTACGGGGCCGGCATCTCCTGTCTCCTGGAGAGGCTTCGATGCCCCTCCACACCCTCTT
DCABDBDDDDDDDDDDDCDDDDBDB@BDDDB@;CCCCCDEFD@;.?<HIGGEIGEHIGJJJIIGIGIIHEGFEHFJIIIIIGJJJHHHHHFFFFFC@@     CC:Z:15 MD:Z:37G62      PG:Z:MarkDuplicates     RG:Z:2888721359 XG:i:0  NH:i:2  HI:
1       XN:i:0  XD: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     CGCTGTGCCCTTCCTTTGCTCTGCCCGCTGGAGACGGTGTTTGTCATGGGCCTGGTCTGCAGGGATCCTGCTACAAAGGTGAA
CCFFFFFHHHHHJJJIJJJJJJJJJJJJJJGIIIIJJFHGGIJGIJJJEGIJIJJHHIHHGHFFEFDEEEECCCAACDDACDCDDDDDB?B?<B>A@CDC     CC:Z:2  MD:Z:100        PG:Z:MarkDuplicates     RG:Z:2888721359 XG:i:0  NH:i:2  HI:
0       XN:i:0  XD: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     AGACGGTGTTTGTCATGGGCCTGGTCTGCAGGGATCCTGCTACAAAGGTGAAACCCAGGAGAGTGTGGAGTCCAGAGTGTTGC
CCFFFADHHHHFIJJJJJIJIJIHIJJJIHJJIJJIIJJIJIJJJJJBFHIIJJJJJJIJHH=EEFFFFCEEECEDCDCDDDDDDDDDDDDDDDBDCCD     CC:Z:2  MD:Z:100        PG:Z:MarkDuplicates     RG:Z:2888721359 XG:i:0  NH:i:2  HI:
0       XN:i:0  XD: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     CCTCTGCAGGAGGCTGCCATTTGTCCTGCCCACCTTCTTAGAAGCGAGACGGAGCAGACCCATCTGCTACTGCCCTTTCTATA
CCFFFFFHHGHHJJJJJJJJJJIJJJJJJIJJJJJJJJJIJIJJJJJJJJJHJJJHGHHFFFFFFEEEEEEDDDDCDCDCDFADCACCACDDCCCCCCD     CC:Z:2  MD:Z:100        PG:Z:MarkDuplicates     RG:Z:2888721359 XG:i:0  NH:i:2  HI:
0       XN:i:0  XD:i:0  CP:i:114357140  AS:i:0  XS:A:+  YT:Z:UU
```

SAM/BAM Header Section

- Used to describe source of data, reference sequence, method of alignment, etc.
- Each section begins with ‘@’ followed by a two-letter record type code. These are followed by two-letter tags and values, example:
    - @HD The header line
    - VN: format version
    - SN: reference sequence name
    - LN: reference sequence length
    - SP: species

SAM/BAM Alignment Section



<table>
<tr><th>Col</th><th>Field</th><th>Type</th><th>Regexp/Range</th><th>Brief description</th></tr>
<tr><td>1</td><td>QNAME</td><td>String</td><td>[!-?A-~]{1,254}</td><td>Query template NAME</td></tr>
<tr><td>2</td><td>FLAG</td><td>Int</td><td>$[0,2^{16}-1]$</td><td>bitwise FLAG</td></tr>
<tr><td>3</td><td>RNAME</td><td>String</td><td>*|[!-()+-<>-~][!-~]*</td><td>Reference sequence NAME</td></tr>
<tr><td>4</td><td>POS</td><td>Int</td><td>$[0,2^{31}-1]$</td><td>1-based leftmost mapping POSition</td></tr>
<tr><td>5</td><td>MAPQ</td><td>Int</td><td>$[0,2^{8}-1]$</td><td>MAPping Quality</td></tr>
<tr><td>6</td><td>CIGAR</td><td>String</td><td>*|([0-9]+[MIDNSHPX=])+</td><td>CIGAR string</td></tr>
<tr><td>7</td><td>RNEXT</td><td>String</td><td>*|=|[!-()+-<>-~][!-~]*</td><td>Ref. name of the mate/next read</td></tr>
<tr><td>8</td><td>PNEXT</td><td>Int</td><td>$[0,2^{31}-1]$</td><td>Position of the mate/next read</td></tr>
<tr><td>9</td><td>TLEN</td><td>Int</td><td>$[-2^{31}+1,2^{31}-1]$</td><td>observed Template LENgth</td></tr>
<tr><td>10</td><td>SEQ</td><td>String</td><td>*|[A-Za-z=.]+</td><td>segment SEQuence</td></tr>
<tr><td>11</td><td>QUAL</td><td>String</td><td>[!-~]+</td><td>ASCII of Phred-scaled base QUALity</td></tr>
</table>

Tools to work with BAM/SAM

- **samtools** - view, sort, index, QC, stats on SAM/BAM files, and more
- **sambamba** - view, sort, index, merge, stats, mark duplicates. fast laternative to samtools
- **picard** - QC, validation, duplicates removal and many more utility tools

BED File Formats

- Text-based, tab-separated list of genomic regions
- Each region is specified by a reference sequence and the start and end positions on it
- Optionally, each region can have additional properties defined – E.g. strand, name, score, color
- Intended for visualizing genomic annotations in IGV, UCSC Genome Browser (context of expression, regulation, variation, conservation, . . .)

BED File Formats

- 3 mandatory columns (must be in correct order)
    - “chrom” – chromosome
    - “chromStart” – the first base of the region with respect to the chromosome (counting starts from 0)
    - “chromEnd” – the first base after the region with respect to the chromosome [chromStart, chromEnd) allows easy region-length calculation
    - Optional fields: “name”, “score”, “strand”, other annotation columns

Example of BED File Formats

```
chr1 115263684 115263685 rs10489525 0 +
chr12 97434219 97434220 rs6538761 0 +
chr14 102360744 102360745 rs7142002 0 +
 chr16 84213683 84213684 rs4150167 0 -
chr2 206086170 206086171 rs4675502 0 +
chr20 14747470 14747471 rs4141463 0 +
```

BED File Formats

- 9 additional optional fields, their order is binding (unlike with SAM format).
- All regions must have the same optional fields
- Most important optional fields:
  - “name” – name of the region
  - “score” – score value between 0 and 1000 (read-count, transformed p-value, “quality”, . . .) Can be interpreted as shades of grey during visualization
  - “strand” – either “+” or “-” (not “1”/“-1”) BED12 format specification available

Tools to work with BED File Formats

- **bedtools** - universal tools for manipulating genomic regions
- **bedops** - complementary to bedtools, providing additional functionality and speedup

Genomic Data Resources

- **GEO**: Gene Expression Omnibus.
  - Host array- and sequencing-based data.
- **ArrayExpress**: European version of GEO.
  - Better curated than GEO but has less data.
- **SRA**: Sequence Read Archive. Designed for hosting large scale high-throughput sequencing data, e.g., high speed file transfer. Data are required to be deposited in one of the databases when paper is accepted

Sequence Read Archive

- The NCBI database which stores sequence data obtained from next generation sequence (NGS) technology
- Archives raw NGS data for various organisms from several platforms (FASTQ files) Serves as a starting point for “secondary analyses”
- Provides access to data from human clinical samples to authorized users who agree to the datasets’ privacy and usage mandates
- Search metadata to locate the sequence reads for download and further downstream analyses

Getting data from SRA

- The NCBI sratoolkit provides two command line tools to allow local BLAST searches against specific sra files directly
  - fastq-dump: Convert SRA data into fastq format
  - prefetch: Allows command-line downloading of SRA, dbGaP, and ADSP data
  - sam-dump: Convert SRA data to sam format
- .sra files are NOT FASTQ files - need to further convert them using sratoolkit