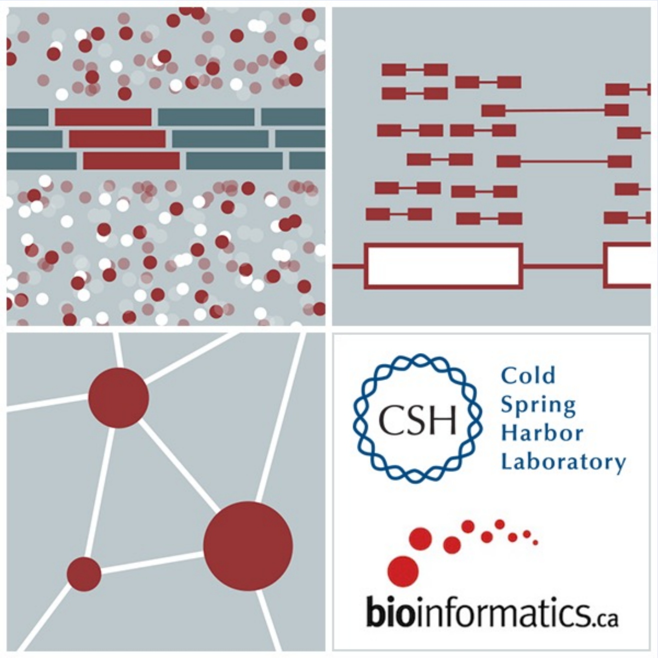
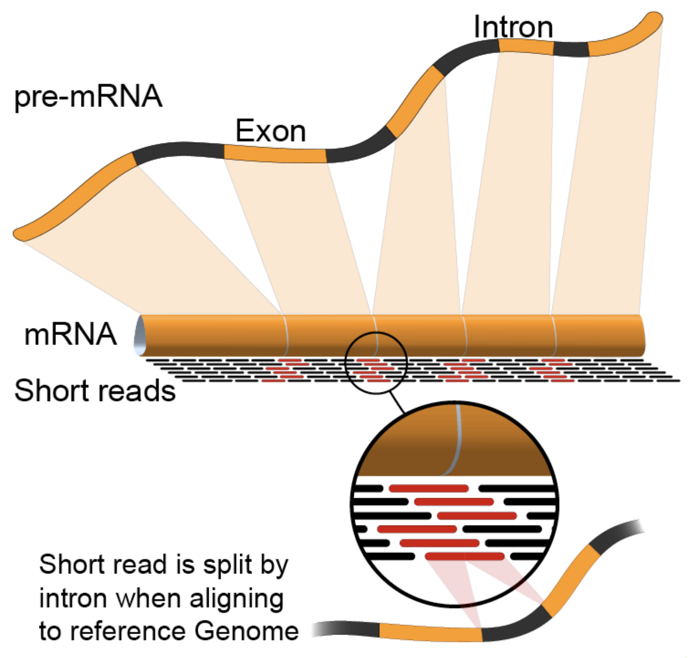
Cold Spring Harbor Laboratory

RNA-Seq Module 2: SAM/BAM/BED file formats

Arpad Danos, Felicia Gomez, Obi Griffith, Malachi Griffith,
My Hoang, Mariam Khanfar, Chris Miller, Kartik Singhal

Advanced Sequencing Technologies & Bioinformatics Analysis November 5-19, 2023




Intron
pre-mRNA
Exon
mRNA
Short reads
Short read is split by intron when aligning to reference Genome

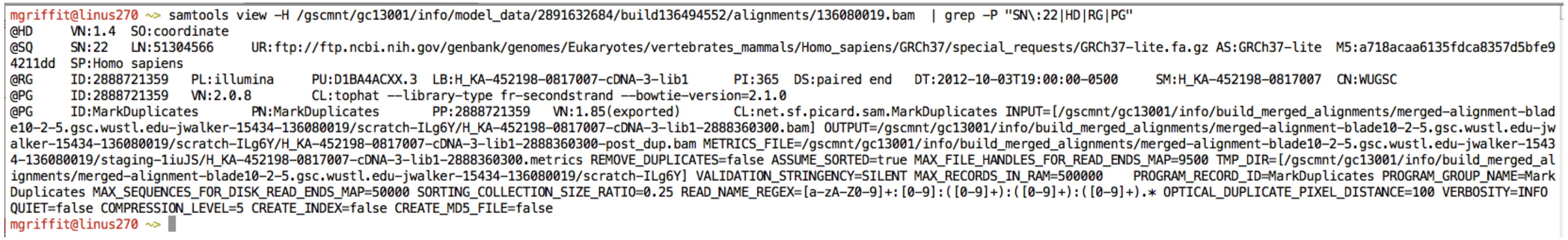
Washington University in St. Louis
School of Medicine

Introduction to the SAM/BAM format

- The specification
  - http://samtools.sourceforge.net/SAM1.pdf
- SAM is uncompressed text data
- BAM is a compressed version of SAM
  - lossless BGZF format
- BAM files are usually ‘indexed’
  - A ‘.bai’ file will be found beside the ‘.bam’ file
  - Indexing provides fast retrieval of alignments overlapping a specified region without going through all alignments.
  - BAM must be sorted by the reference ID and then the leftmost coordinate before indexing

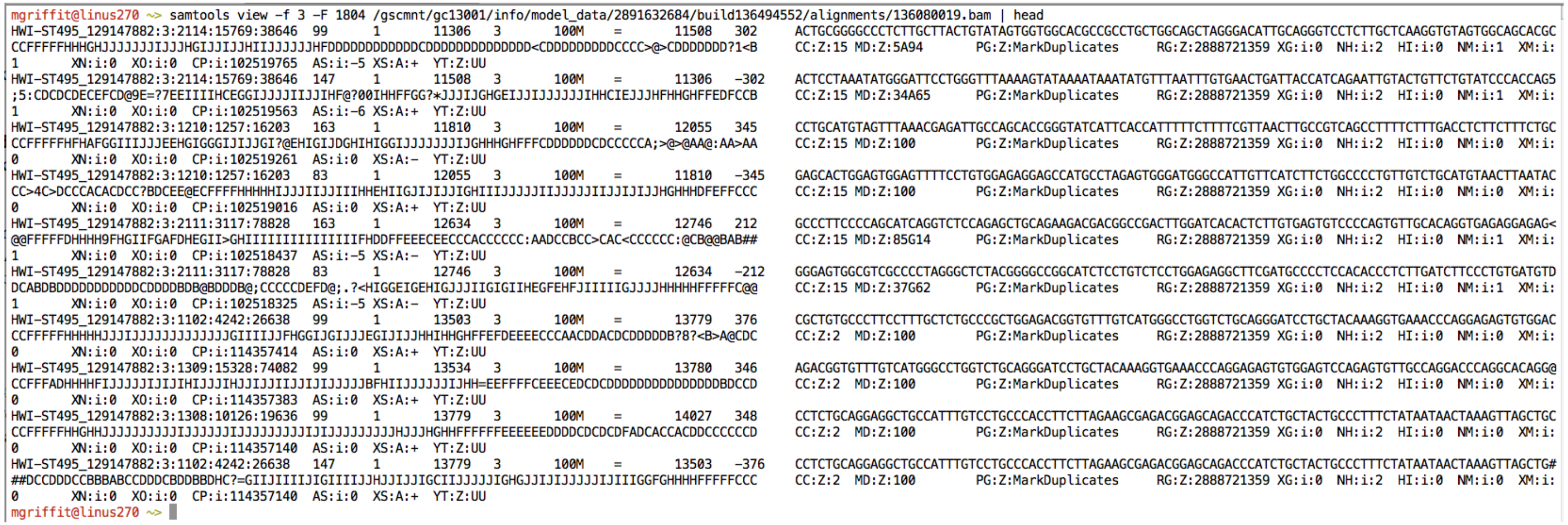
Example of SAM/BAM file format

Example SAM/BAM/CRAM 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/CRAM 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
DCABDBDDDDDDDDDDDCDDDDBDB@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
CCFFFFFHHGHHJJJJJJJJJJIJJJJJJIJJJJJJJJJIJIJJJJJJJJJHJJJHGHHFFFFFEEEEEEDDDDCDCDCDFADCACCACDDCCCCCCD     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 ~> 
```

SAM/BAM header section

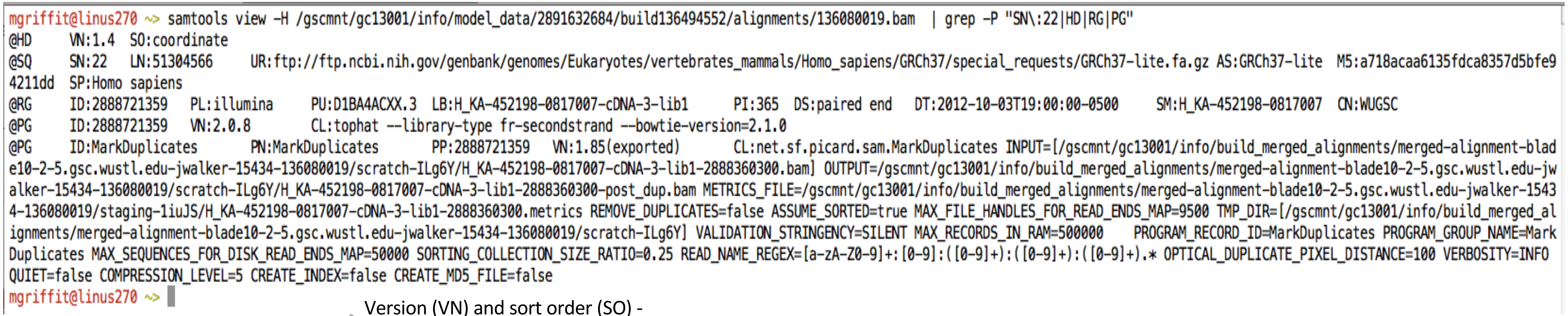
- Used to describe source of data, reference sequence, method of alignment, etc.

- 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

A BAM file is divided in header and alignment sections

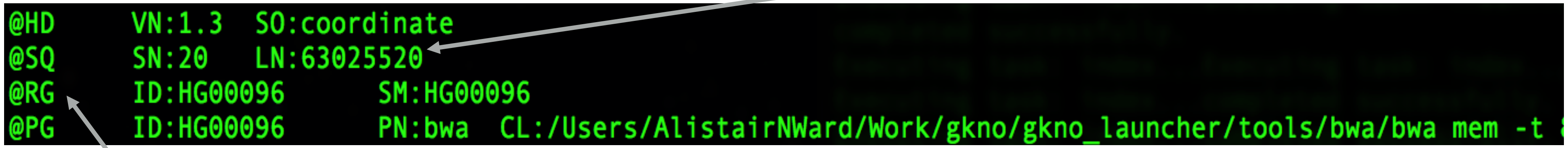
Example SAM/BAM header section (abbreviated)



```
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 ~>
```

Version (VN) and sort order (SO) - Important!

Reference sequence (SQ) and sequence length (LN)



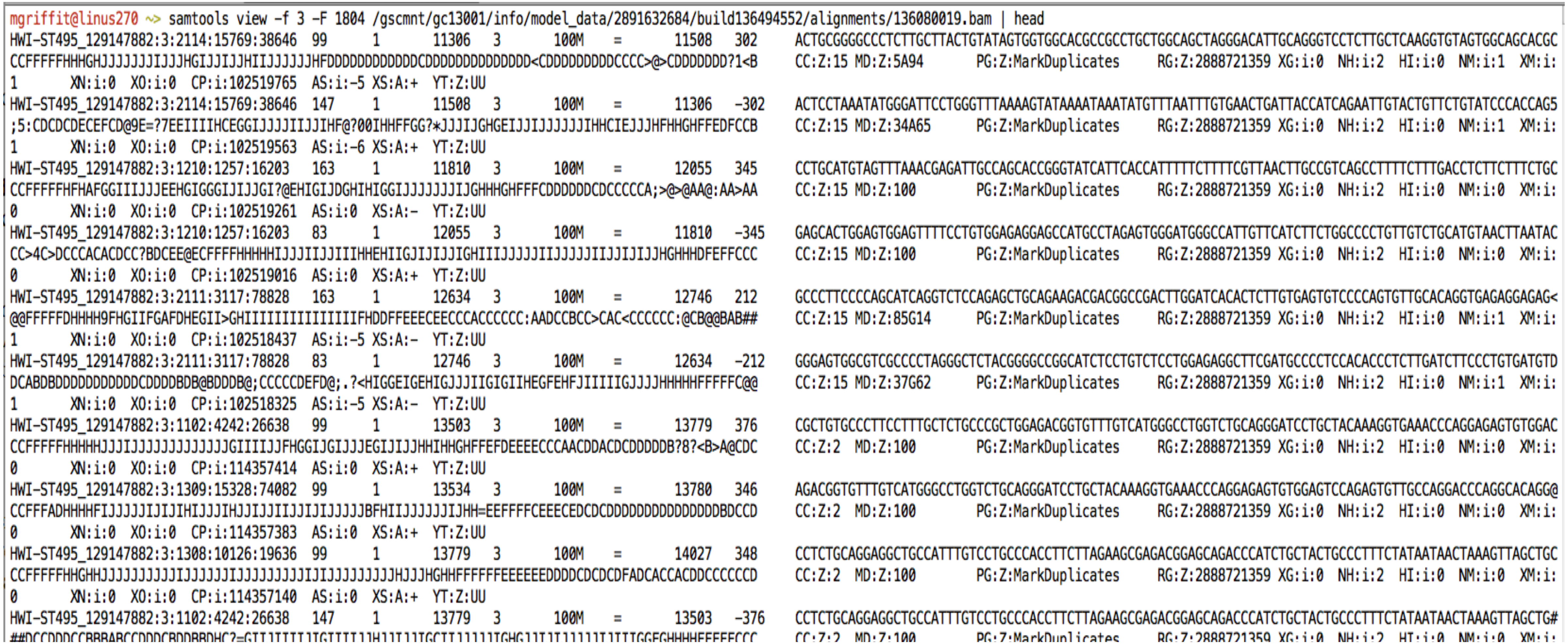
```
@HD     VN:1.3  SO:coordinate
@SQ     SN:20   LN:63025520
@RG     ID:HG00096      SM:HG00096
@PG     ID:HG00096      PN:bwa  CL:/Users/AlistairNWard/Work/gkno/gkno_launcher/tools/bwa/bwa mem -t
```

Read group (RG) and sample (SM)

Programs (PG) that have been run on the data

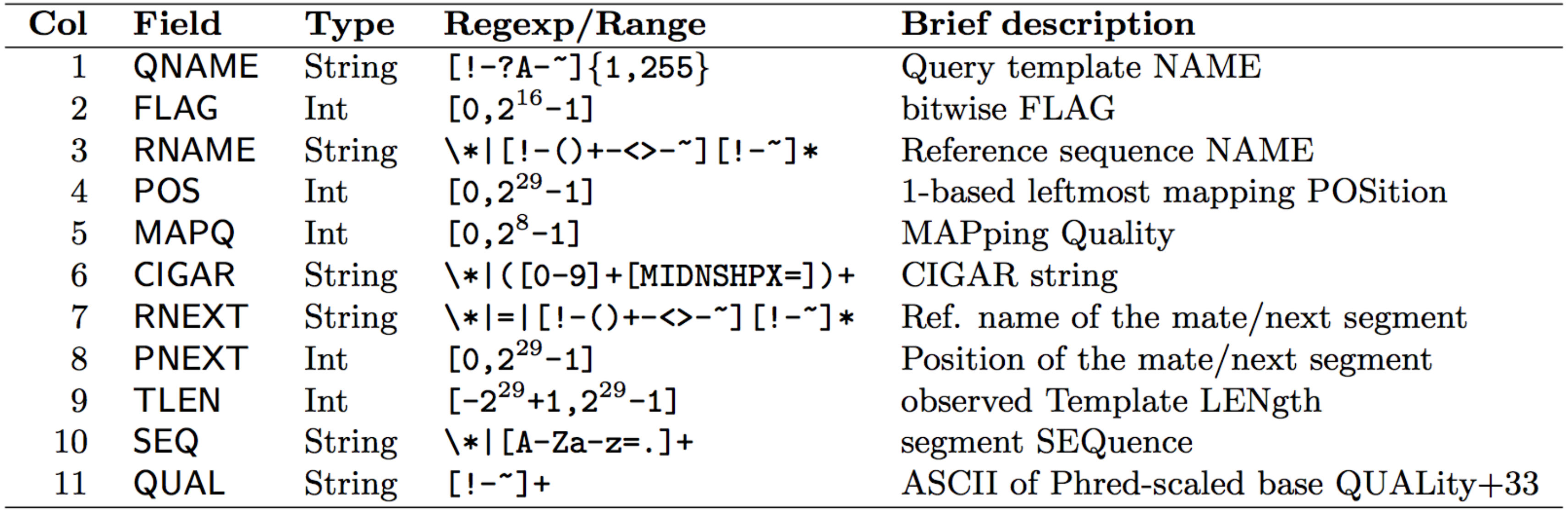
A BAM file is divided in header and alignment sections

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<CDDDDDDDDDCCCC>@>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    ACTCCTAAATATGGATTCCTGGGTTTAAAAGTATAAAATAAATATGTTTAATTTGTGAACTGATTACCATCAGAATTGTACTGTTCTGTATCCCACCAG5
;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
DCABDBDDDDDDDDDDDCDDDDBDB@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@
CCFFFADHHHHFIJJJJIJIJIHIJJJIHJJIJJIIJJIJIJJJJJBFHIIJJJJJJIJHH=EEFFFFCEEECEDCDCDDDDDDDDDDDDDDDDBDCCD    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?=GIIJIIIIJJGIIIIJJHJJIJJIGCIIJJJJJIGHGJJIJIJJJJIJIIIGGFGHHHHFFFFFCCC    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:
```

SAM/BAM alignment section

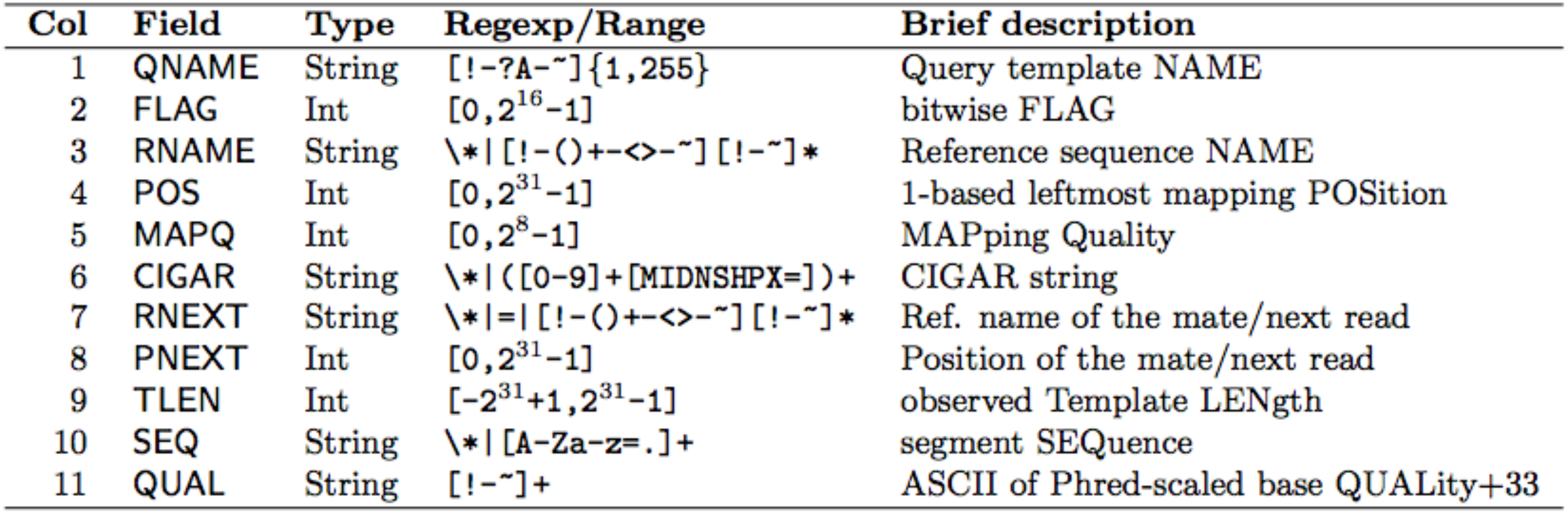


<table>
<tr><th>Col</th><th>Field</th><th>Type</th><th>Regexp/Range</th><th>Brief description</th></tr>
<tr><td>1</td><td>QNAME</td><td>String</td><td>[!-?A-~]{1,255}</td><td>Query template NAME</td></tr>
<tr><td>2</td><td>FLAG</td><td>Int</td><td>[0,2^{16}-1]</td><td>bitwise FLAG</td></tr>
<tr><td>3</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^{29}-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 segment</td></tr>
<tr><td>8</td><td>PNEXT</td><td>Int</td><td>[0,2^{29}-1]</td><td>Position of the mate/next segment</td></tr>
<tr><td>9</td><td>TLEN</td><td>Int</td><td>[-2^{29}+1,2^{29}-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+33</td></tr>
</table>

Example values

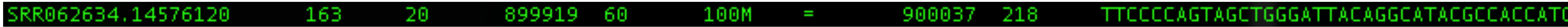
1	QNAME	e.g.	HWI-ST495_129147882:1:2302:10269:12362
2	FLAG	e.g.	99
3	RNAME	e.g.	1
4	POS	e.g.	11623
5	MAPQ	e.g.	3
6	CIGAR	e.g.	100M
7	RNEXT	e.g.	=
8	PNEXT	e.g.	11740
9	TLEN	e.g.	217
10	SEQ	e.g.	CCTGTTTCTCCACAAAGTGTTTACTTTTGGATTTTTGCCAGTCTAACAGGTGAAGCCCTGGAGATTCTTATTAGTGATTTGGGCTGGGGCCTGGCCATGT
11	QUAL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

SAM Format – Information Fields



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^{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

```
1                    2        3          4        5          6         7            8         9           10
```

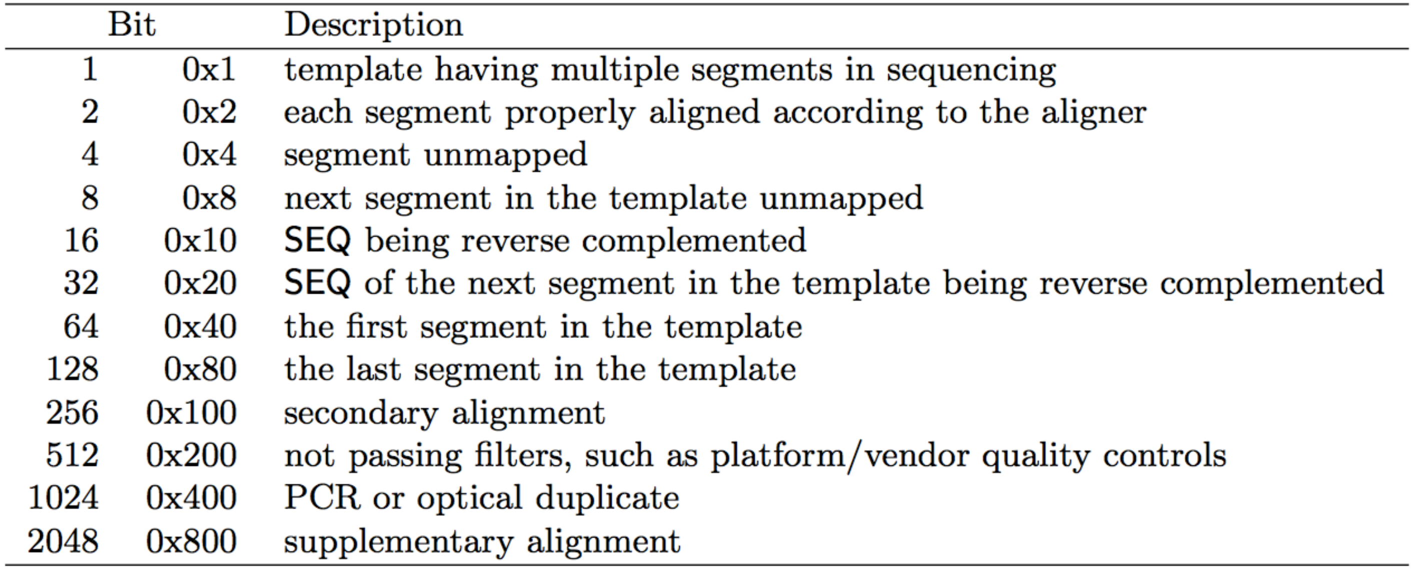


```
SRR062634.14576120      163     20      899919  60      100M    =       900037  218     TTCCCCAGTAGCTGGGATTACAGGCATACGCCACCAT
```

?   ?

SAM/BAM flags explained

- 12 bitwise flags describing the alignment
- Stored as a binary string of length 12 instead of 12 columns of data
- Value of ‘1’ indicates the flag is set. e.g. 001000000000
- All combinations can be represented as a number from 0 to 4095 (i.e. 2^{12}-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

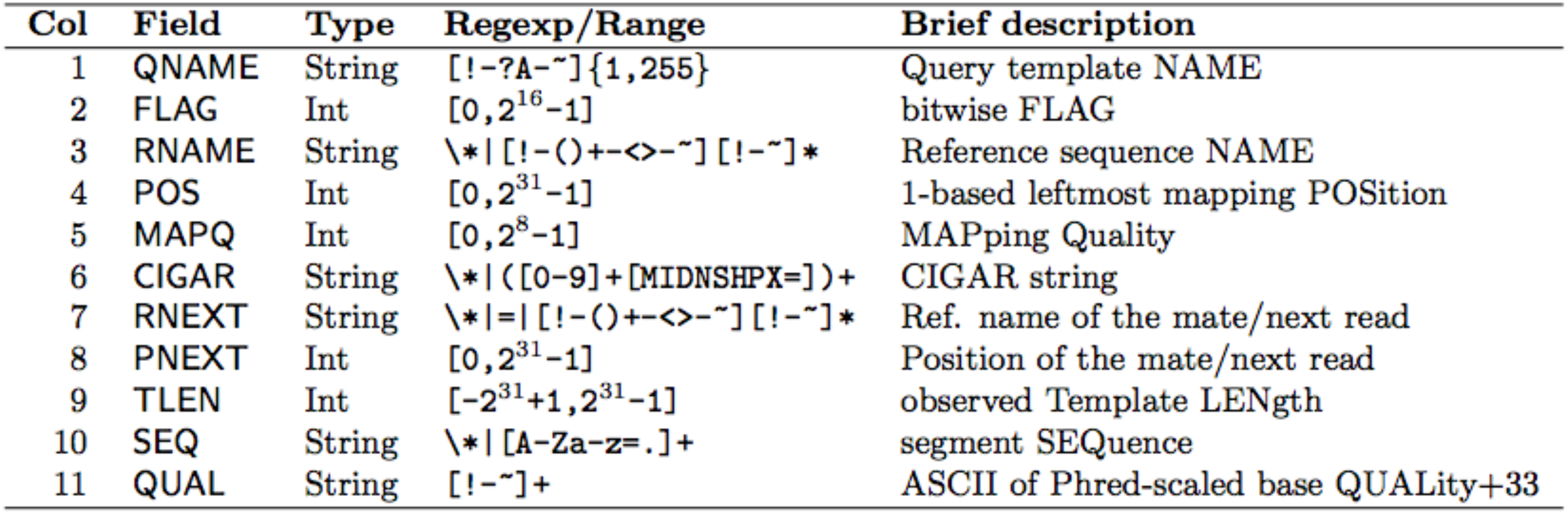


Bit		Description
1	0x1	template having multiple segments in sequencing
2	0x2	each segment properly aligned according to the aligner
4	0x4	segment unmapped
8	0x8	next segment in the template unmapped
16	0x10	SEQ being reverse complemented
32	0x20	SEQ of the next segment in the template being reverse complemented
64	0x40	the first segment in the template
128	0x80	the last segment in the template
256	0x100	secondary alignment
512	0x200	not passing filters, such as platform/vendor quality controls
1024	0x400	PCR or optical duplicate
2048	0x800	supplementary alignment

Note that to maximize confusion, each bit is described in the SAM specification using its hexadecimal representation (i.e., '0x10' = 16 and '0x40' = 64).

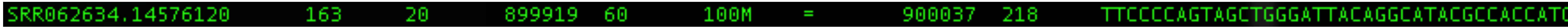
http://broadinstitute.github.io/picard/explain-flags.html

SAM Format – Information Fields



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^{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

1 2 3 4 5 6 7 8 9 10



```
SRR062634.14576120   163   20   899919  60   100M   =   900037  218   TTCCCCAGTAGCTGGGATTACAGGCATACGCCACCAT
```

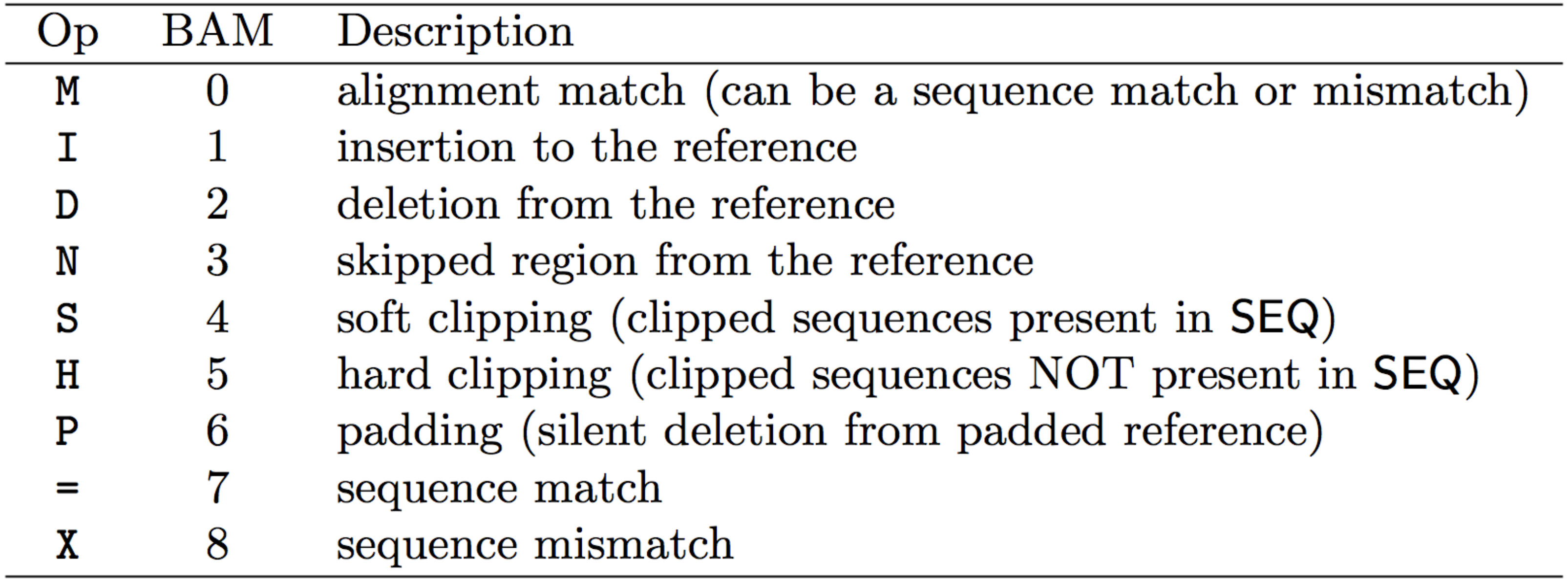
? ?

Slide courtesy of Andrew Farrell

CIGAR strings explained

- The ‘CIGAR’ (**C**ompact **I**diosyncratic **G**apped **A**lignment **R**eport)

- The CIGAR string is a sequence of base lengths and associated ‘operations’ indicating which bases align to the reference (either a match or mismatch), are deleted, are inserted, represent introns, etc.



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

- e.g. 81M859N19M
    - A 100 bp read consists of: 81 bases of alignment to reference, 859 bases skipped (an intron), 19 bases of alignment

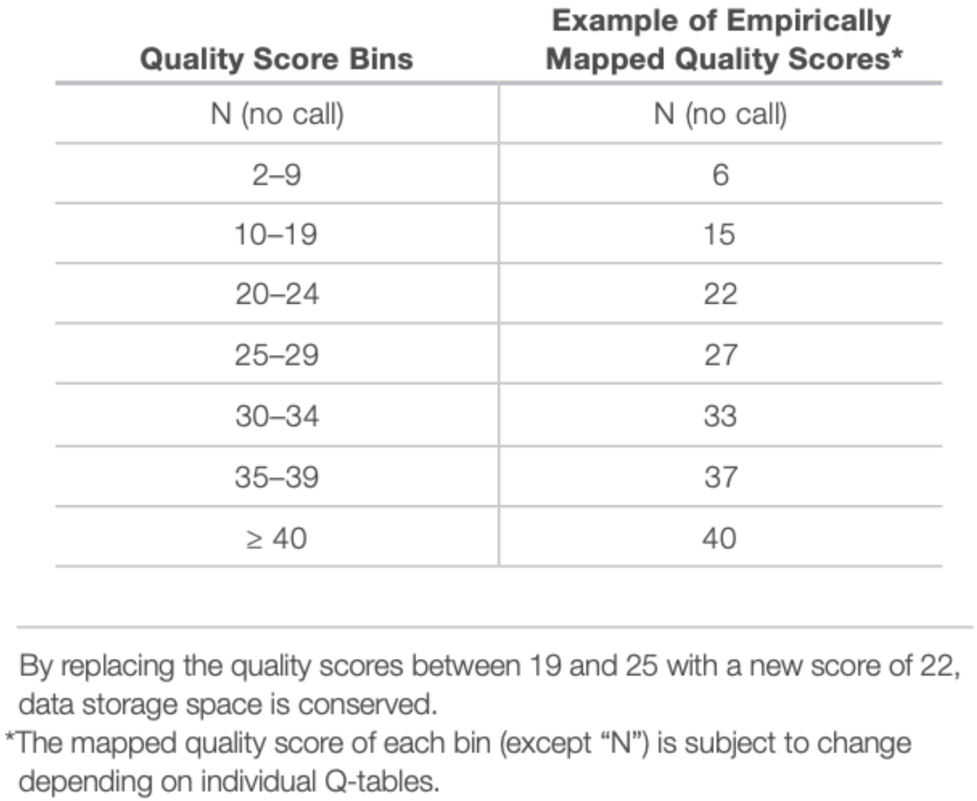
CRAM files

- CRAM is an ultra-compressed version of a BAM file
  - Usually between 30-60% smaller than the corresponding BAM

- Stores “diffs” from the reference genome
  - requires the matching reference genome to restore original data!

- Base quality binning may be used as well

- Some tools still require conversion back to bam



Quality Score Bins	Example of Empirically Mapped Quality Scores*
N (no call)	N (no call)
2–9	6
10–19	15
20–24	22
25–29	27
30–34	33
35–39	37
≥ 40	40

By replacing the quality scores between 19 and 25 with a new score of 22, data storage space is conserved.

*The mapped quality score of each bin (except “N”) is subject to change depending on individual Q-tables.

Introduction to the BED format

- When working with BAM files, it is very common to want to examine a focused subset of the reference genome
    - e.g. the exons of a gene

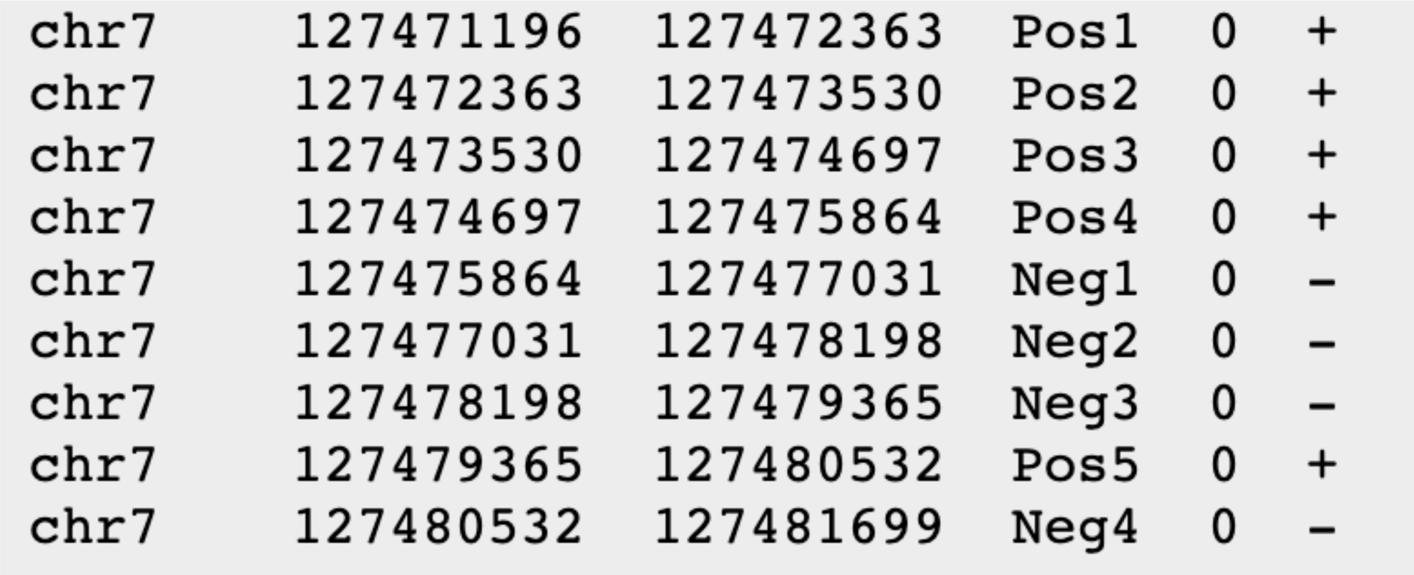
- These subsets are commonly specified in ‘BED’ files
    - https://genome.ucsc.edu/FAQ/FAQformat.html#format1

- Many BAM manipulation tools accept regions of interest in BED format

- Basic BED format (tab separated):
    - Chromosome name, start position, end position (BED3)
    - Coordinates in BED format are 0 based

Introduction to the BED format

- There are several flavors of BED format: BED3, BED4, BED6, BED8, etc
- First 3 fields always required: chr, start, stop
- Followed by up to 9 additional optional fields: name, score, strand, thickStart, thickEnd, itemRGB, blockCount, blockSizes, blockStarts



```
chr7    127471196  127472363  Pos1  0  +
chr7    127472363  127473530  Pos2  0  +
chr7    127473530  127474697  Pos3  0  +
chr7    127474697  127475864  Pos4  0  +
chr7    127475864  127477031  Neg1  0  -
chr7    127477031  127478198  Neg2  0  -
chr7    127478198  127479365  Neg3  0  -
chr7    127479365  127480532  Pos5  0  +
chr7    127480532  127481699  Neg4  0  -
```

Manipulation of SAM/BAM and BED files

- Several tools are used ubiquitously in sequence analysis to manipulate these files

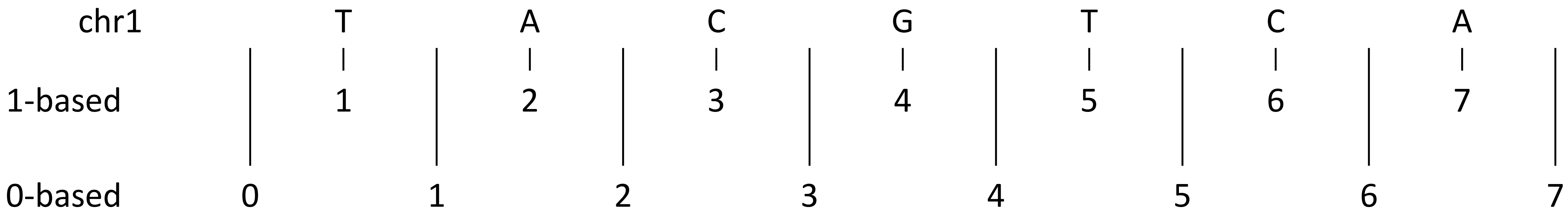
- SAM/BAM files
  - samtools
  - bamtools
  - Picard

- BED files
  - bedtools
  - bedops

Common sources of confusion

- Genomic coordinate systems
- Genome builds
- Variant representation

Genomic coordinates – 1 vs 0 based



chr1		T		A		C		G		T		C		A	
1-based		1		2		3		4		5		6		7	
0-based	0		1		2		3		4		5		6		7

	1-based	0-based
Indicate a single nucleotide	chr1:4-4 G	chr1:3-4 G
Indicate a range of nucleotides	chr1:2-4 ACG	chr1:1-4 ACG
Indicate a single nucleotide variant	chr1:5-5 T/A	chr1:4-5 T/A

- 1-based : Single nucleotides, variant positions, or ranges are specified directly by their corresponding nucleotide numbers
  - GFF, SAM, VCF, Ensembl browser, …
- 0-based: Single nucleotides, variant positions, or ranges are specified by the coordinates that flank them
  - BED, BAM, UCSC browser, …

Genome builds

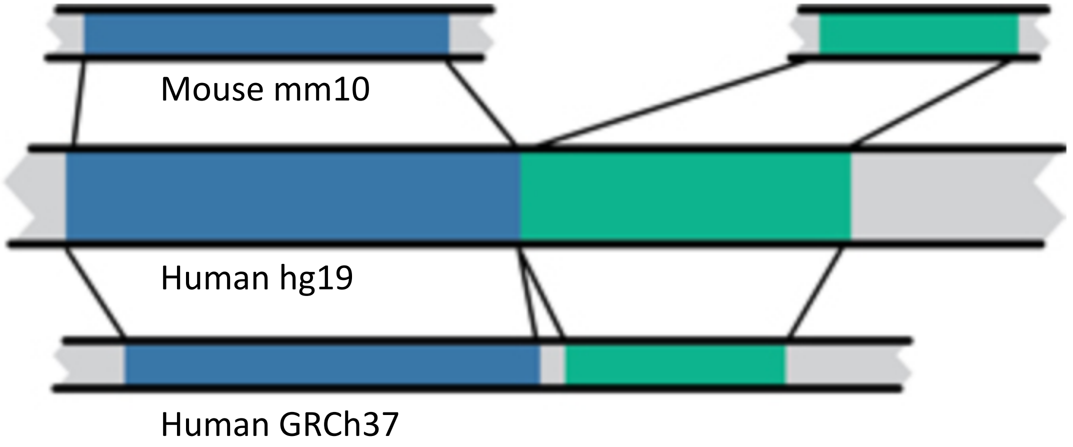
Reference Genome builds

Current human: GRCh38, hg38, b38
alternates: GRCh38v2_ccdg,
GRCh38_full_analysis_set_plus_decoy_hla

Previous human: GRCh37, hg19, b37

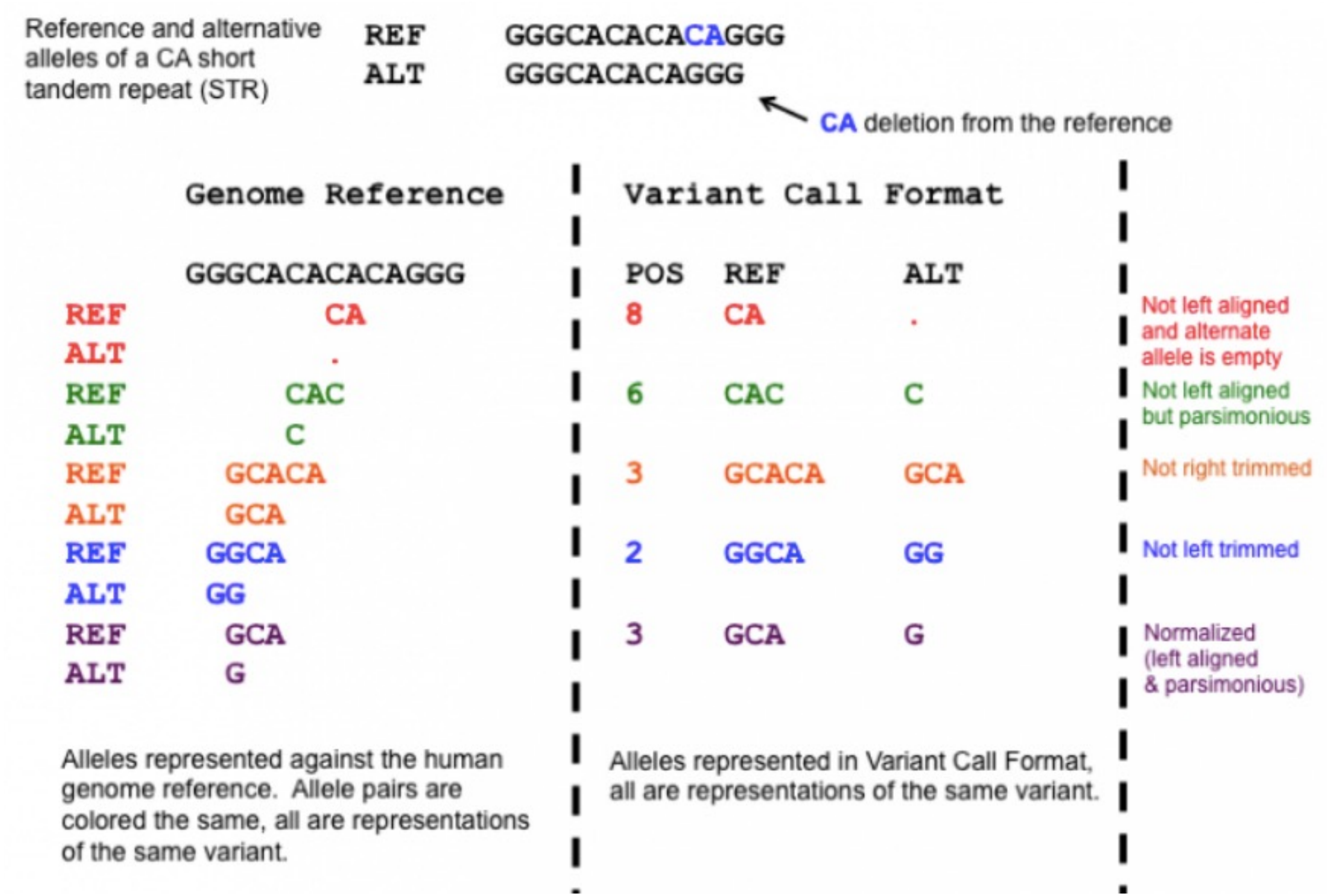
Current mouse: GRCm38, mm10

Lift-over



For a detailed discussion of various human reference genome flavors refer here:
https://pmbio.org/module-02-inputs/0002/02/01/Reference_Genome/

Variant shifting (alignment) and parsimony/trimming



Reference and alternative alleles of a CA short tandem repeat (STR)

REF GGGCACACACAGGG
ALT GGGCACACAGGG

CA deletion from the reference

	Genome Reference	Variant Call Format			
	GGGCACACACAGGG	POS	REF	ALT	
REF	CA	8	CA	.	Not left aligned and alternate allele is empty
ALT	.				
REF	CAC	6	CAC	C	Not left aligned but parsimonious
ALT	C				
REF	GCACA	3	GCACA	GCA	Not right trimmed
ALT	GCA				
REF	GGCA	2	GGCA	GG	Not left trimmed
ALT	GG				
REF	GCA	3	GCA	G	Normalized (left aligned & parsimonious)
ALT	G				

Alleles represented against the human genome reference. Allele pairs are colored the same, all are representations of the same variant.

Alleles represented in Variant Call Format, all are representations of the same variant.

Parsimony: representing variant in as few nucleotides as possible without reducing the length of any allele to 0

Left (right) aligning = shifting the start position of a variant as far to the left (right) as possible

How should I sort my SAM/BAM file?

- Generally BAM files are sorted by position
  - This is for performance reasons
    - When sorted and indexed, arbitrary positions in a massive BAM file can be accessed rapidly

- Certain tools require a BAM sorted by read name
  - Usually this is when we need to easily identify both reads of a pair
    - The insert size between two reads may be large
    - In fusion detection we are interested in read pairs that map to different chromosomes

We are on a Coffee Break & Networking Session