Tertiary Analysis

- Working with BAM/SAM files

SAM Specification

- Latest vesion SAMv1 (http://samtools.github.io/hts-specs/)

Highlights

- Header
- Alignment Section
  - FLAG
  - POS convention (always the 5' end)
  - MAPQ (read MAQ paper PHRED score for Prob(mismapped))
  - CIGAR (Does not give mismatches; on IN/DEL; 50M!=perfect match necessarily)
  - MATE INFO (RNEXT, PNEXT, TLEN)
  - SEQ
  - QUAL
- TAGS
  - Standarized
  - Custom

SAM/BAM Format

Proliferation of alignment formats over the years: Cigar, psl, gff, xml etc.

SAM (Sequence Alignment/Map) format

- Single unified format for storing read alignments to a reference genome

BAM (Binary Alignment/Map) format

- Binary equivalent of SAM
- Developed for fast processing/indexing

Advantages

- Can store alignments from most aligners
- Supports multiple sequencing technologies
- Supports indexing for quick retrieval/viewing
- Compact size (e.g. 112Gbp Illumina = 116Gbytes disk space)
- Reads can be grouped into logical groups e.g. lanes, libraries, individuals/genotypes
- Supports second best base call/quality for hard to call bases

Possibility of storing raw sequencing data in BAM as replacement to SRF & fastq

Read Entries in SAM

No.	Name	Description
1	QNAME	Query NAME of the read or the read pair
2	FLAG	Bitwise FLAG (pairing, strand, mate strand, etc.)
3	RNAME	Reference sequence NAME
4	POS	1-Based leftmost POSition of clipped alignment
5	MAPQ	MAPping Quality (Phred-scaled)
6	CIGAR	Extended CIGAR string (operations: MIDNSHP)
7	MRNM	Mate Reference NaMe ('=' if same as RNAME)
8	MPOS	1-Based leftmost Mate POSition
9	ISIZE	Inferred Insert SIZE
10	SEQ	Query SEQuence on the same strand as the reference
11	QUAL	Query QUALity (ASCII-33=Phred base quality)

Heng Li , Bob Handsaker , Alec Wysoker , Tim Fennell , Jue Ruan , Nils Homer , Gabor Marth , Goncalo Abecasis , Richard Durbin , and 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools, *Bioinformatics,* 25:2078-2079

Extended Cigar Format

Cigar has been traditionally used as a compact way to represent a sequence alignment

Operations include

- M - match or mismatch
- I - insertion
- D - deletion

SAM extends these to include

- S - soft clip
- H - hard clip
- N - skipped bases
- P – padding

E.g.

```
Read:   ACGCA-TGCAGTtagacgt
Ref:    ACTCAGTG—-GT
Cigar:  5M1D2M2I2M7S
```

What is the cigar line?

```
E.g.    Read:   ACGCA-TGCAGTtagacgt
        Ref:    ACTCAGTG—-GT
        Cigar:  5M1D2M2I2M7S

E.g.    Read:   tgtcgtcACGCATG---CAGTtagacgt
        Ref:           ACGCATGCGGCAGT
        Cigar:
```

Read Group Tag

Each lane (or equivalent unit) has a unique read group (RG) tag

1000 Genomes

- Meta information derived from DCC

RG tags

- ID: SRR/ERR number
- PL: Sequencing platform
- PU: Run name
- LB: Library name
- PI: Insert fragment size
- SM: Individual
- CN: Sequencing center

Activity 2: Interpreting SAM/BAM files

From reading page 4 of the SAM specification, look at the following line from the header of the BAM file:

```
@RG     ID:ERR001711    PL:ILLUMINA     LB:g1k-sc-NA12878-CEU-1 PI:200  DS:SRP000032
SM:NA12878      CN:SC
```

What does RG stand for?

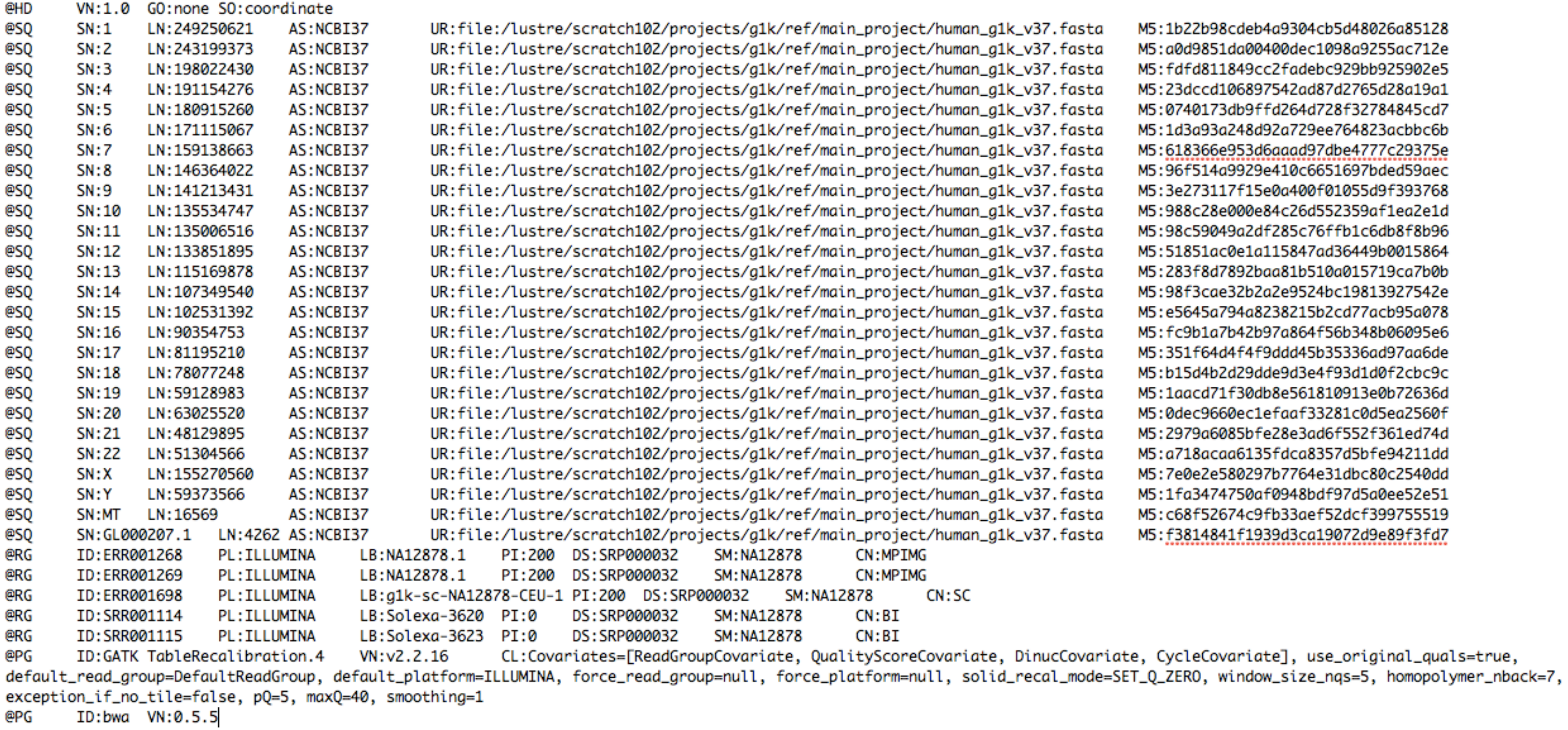
What is the sequencing platform?

What is the sequencing centre?

What is the lane accession number?

What is the expected fragment insert size?

1000 Genomes BAM File



```
@HD     VN:1.0  GO:none SO:coordinate
@SQ     SN:1    LN:249250621    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:1b22b98cdeb4a9304cb5d48026a85128
@SQ     SN:2    LN:243199373    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:a0d9851da00400dec1098a9255ac712e
@SQ     SN:3    LN:198022430    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:fdfd811849cc2fadebc929bb925902e5
@SQ     SN:4    LN:191154276    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:23dccd106897542ad87d2765d28a19a1
@SQ     SN:5    LN:180915260    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:0740173db9ffd264d728f32784845cd7
@SQ     SN:6    LN:171115067    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:1d3a93a248d92a729ee764823acbbc6b
@SQ     SN:7    LN:159138663    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:618366e953d6aaad97dbe4777c29375e
@SQ     SN:8    LN:146364022    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:96f514a9929e410c6651697bded59aec
@SQ     SN:9    LN:141213431    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:3e273117f15e0a400f01055d9f393768
@SQ     SN:10   LN:135534747    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:988c28e000e84c26d552359af1ea2e1d
@SQ     SN:11   LN:135006516    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:98c59049a2df285c76ffb1c6db8f8b96
@SQ     SN:12   LN:133851895    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:51851ac0e1a115847ad36449b0015864
@SQ     SN:13   LN:115169878    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:283f8d7892baa81b510a015719ca7b0b
@SQ     SN:14   LN:107349540    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:98f3cae32b2a2e9524bc19813927542e
@SQ     SN:15   LN:102531392    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:e5645a794a8238215b2cd77acb95a078
@SQ     SN:16   LN:90354753     AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:fc9b1a7b42b97a864f56b348b06095e6
@SQ     SN:17   LN:81195210     AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:351f64d4f4f9ddd45b35336ad97aa6de
@SQ     SN:18   LN:78077248     AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:b15d4b2d29dde9d3e4f93d1d0f2cbc9c
@SQ     SN:19   LN:59128983     AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:1aacd71f30db8e561810913e0b72636d
@SQ     SN:20   LN:63025520     AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:0dec9660ec1efaaf33281c0d5ea2560f
@SQ     SN:21   LN:48129895     AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:2979a6085bfe28e3ad6f552f361ed74d
@SQ     SN:22   LN:51304566     AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:a718acaa6135fdca8357d5bfe94211dd
@SQ     SN:X    LN:155270560    AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:7e0e2e580297b7764e31dbc80c2540dd
@SQ     SN:Y    LN:59373566     AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:1fa3474750af0948bdf97d5a0ee52e51
@SQ     SN:MT   LN:16569        AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:c68f52674c9fb33aef52dcf399755519
@SQ     SN:GL000207.1   LN:4262 AS:NCBI37       UR:file:/lustre/scratch102/projects/g1k/ref/main_project/human_g1k_v37.fasta    M5:f3814841f1939d3ca19072d9e89f3fd7
@RG     ID:ERR001268    PL:ILLUMINA     LB:NA12878.1    PI:200  DS:SRP000032    SM:NA12878      CN:MPIMG
@RG     ID:ERR001269    PL:ILLUMINA     LB:NA12878.1    PI:200  DS:SRP000032    SM:NA12878      CN:MPIMG
@RG     ID:ERR001698    PL:ILLUMINA     LB:g1k-sc-NA12878-CEU-1 PI:200  DS:SRP000032    SM:NA12878      CN:SC
@RG     ID:SRR001114    PL:ILLUMINA     LB:Solexa-3620  PI:0    DS:SRP000032    SM:NA12878      CN:BI
@RG     ID:SRR001115    PL:ILLUMINA     LB:Solexa-3623  PI:0    DS:SRP000032    SM:NA12878      CN:BI
@PG     ID:GATK TableRecalibration.4    VN:v2.2.16      CL:Covariates=[ReadGroupCovariate, QualityScoreCovariate, DinucCovariate, CycleCovariate], use_original_quals=true,
default_read_group=DefaultReadGroup, default_platform=ILLUMINA, force_read_group=null, force_platform=null, solid_recal_mode=SET_Q_ZERO, window_size_nqs=5, homopolymer_nback=7,
exception_if_no_tile=false, pQ=5, maxQ=40, smoothing=1
@PG     ID:bwa  VN:0.5.5
```

samtools view –H my.bam

How is the BAM file sorted?
How many different sequencing centres contributed lanes to this BAM file?
What is the alignment tool used to create this BAM file?

How many different sequencing libraries are there in this BAM? Hint: RG tag

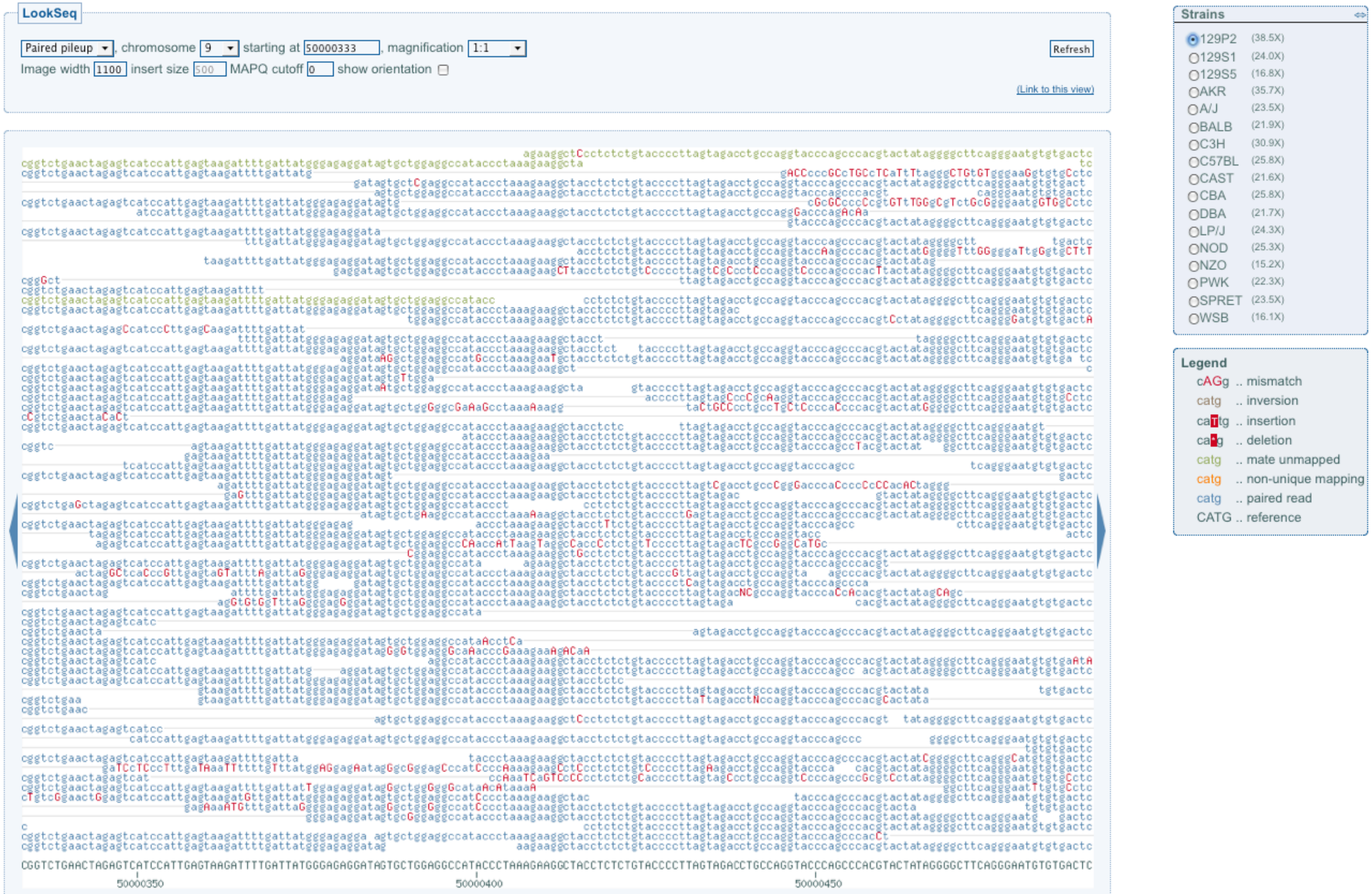
SAM/BAM Tools

Well defined specification for SAM/BAM

Several tools and programming APIs for interacting with SAM/BAM files

- Samtools - Sanger/C (http://samtools.sourceforge.net)
  - Convert SAM <-> BAM
  - Sort, index, BAM files
  - Flagstat – summary of the mapping flags
  - Merge multiple BAM files
  - Rmdup – remove PCR duplicates from the library preparation
- Picard - Broad Institute/Java (http://picard.sourceforge.net)
  - MarkDuplicates, CollectAlignmentSummaryMetrics, CreateSequenceDictionary, SamToFastq, MeanQualityByCycle, FixMateInformation…….
- Bio-SamTool – Perl (http://search.cpan.org/~lds/Bio-SamTools/)
- Pysam – Python (http://code.google.com/p/pysam/)

BAM Visualisation



LookSeq

Paired pileup ▾, chromosome 9 ▾ starting at 50000333, magnification 1:1 ▾ Refresh

Image width 1100 insert size 500 MAPQ cutoff 0 show orientation □

(Link to this view)

```
CGGTCTGAACTAGAGTCATCCATTGAGTAAGATTTTGATTATGGGAGAGGATAGTGCTGGAGGCCATACCCTAAAGAAGGCTACCTCTCTGTACCCCTTAGTAGACCTGCCAGGTACCCAGCCCACGTACTATAGGGGCTTCAGGGAATGTGTGACTC
            50000350                                               50000400                                                 50000450
```

Strains

● 129P2	(38.5X)
○ 129S1	(24.0X)
○ 129S5	(16.8X)
○ AKR	(35.7X)
○ A/J	(23.5X)
○ BALB	(21.9X)
○ C3H	(30.9X)
○ C57BL	(25.8X)
○ CAST	(21.6X)
○ CBA	(25.8X)
○ DBA	(21.7X)
○ LP/J	(24.3X)
○ NOD	(25.3X)
○ NZO	(15.2X)
○ PWK	(22.3X)
○ SPRET	(23.5X)
○ WSB	(16.1X)

Legend

cAGg .. mismatch
catg .. inversion
caTtg .. insertion
ca■g .. deletion
catg .. mate unmapped
catg .. non-unique mapping
catg .. paired read
CATG .. reference

http://www.sanger.ac.uk/mousegenomes

Flags

- major headache for humans but the right thing to do.
  - But why on earth is strand bit 4 and not bit 1; the thing you want most should be in the first bit: even == positive, odd == negative
  - old samtools had -X option but really not that much better

Dec	Hex	Flags	Dec	Hex	Flags	Dec	Hex	Flags
65	0x41	p1	69	0x45	pu1	73	0x49	pU1
81	0x51	pr1	97	0x61	pR1	113	0x71	prR1
117	0x75	purR1	121	0x79	pUrR1	129	0x81	p2
133	0x85	pu2	137	0x89	pU2	145	0x91	pr2
161	0xa1	pR2	177	0xb1	prR2	181	0xb5	purR2
185	0xb9	pUrR2	321	0x141	p1s	329	0x149	pU1s
337	0x151	pr1s	353	0x161	pR1s	369	0x171	prR1s
377	0x179	pUrR1s	385	0x181	p2s	401	0x191	pr2s
417	0x1a1	pR2s	433	0x1b1	prR2s	1089	0x441	p1d
1097	0x449	pU1d	1105	0x451	pr1d	1121	0x461	pR1d
1137	0x471	prR1d	1145	0x479	pUrR1d	1153	0x481	p2d

Flags; better solution

- PICARD page is a life saver; bookmarkit or download it https://broadinstitute.github.io/picard/explain-flags.html

Samtools / Picard

- When there is overlap, my honest advice, use Picard
- Unless you are doing pipes/streams
  - But probably should not be doing those anyway
- However samtools view is prehaps the most used samfile command ever (really)
  - go over options

PICARD

- Two main uses
  - manipulating SAM/BAMs
    - AddRG, Sort, Index & MarkDup in almost every pipeline
    - Mark Duplicates a key step in many cases
  - BAM stats
    - Alignment Stats
    - Insert Size
    - Duplicates Stats
    - and a bunch of misc other stuff
- Wins award for friendliest bioinformatics tool

Mark/Remove Duplicates

- PCR amplification is present in almost in all library preps
- Depending on number of cycles (amount of amplifiction) you can get PCR run aways
  - a single molecule is copied 100-1,000 of times
- Severe problem in variant (mutation) detection
  - if that molecule had an error the error gets amplified
- Mark Duplicates is a critial part of most pipelines
  - And the duplication statistics are a measure of library quality

Multi-mapper issue

- Many pipeline simple filter these reads out.
- BWA MEM problem
  - No longer sets simple flag
  - if using filter on MAPQ
- If using multi-mappers in uniq-mode need to really make sure:
  - how the algorithm deals with high multiplicity
  - random choice?
- Bowtie/SHRiMP for exhaustive multi-mappers
- CSEM (http://deweylab.biostat.wisc.edu/csem/)
  - impute likely position of multi-mappers by looking at surronding unique mappers.

Other bioinformatics file formats

Other range formats

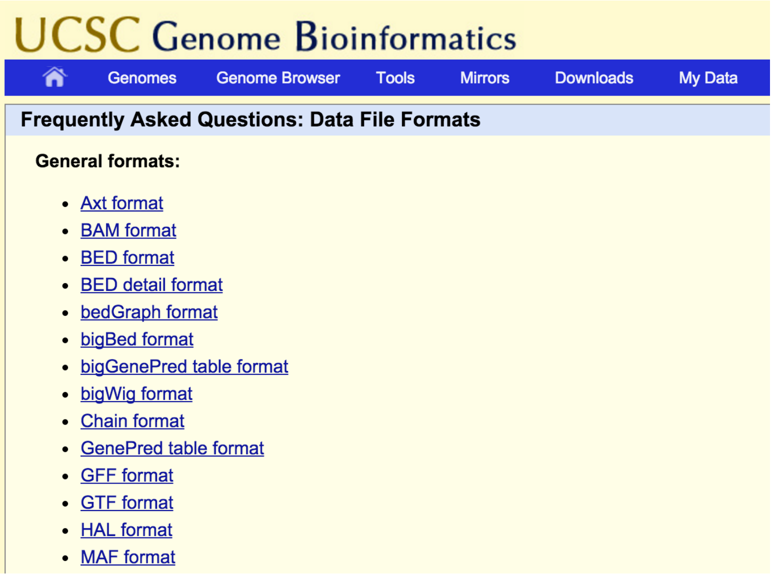
- BED (0-offset)
  - stand 3 column format:
    - chromsome
    - start (first base is 0)
    - end
  - various extended version
- Interval List (1-offest)
  - Used by Picard:
  - Genome Header so you know what the reference is
  - Standard 5 column format
    - Chromsome
    - Start (first base is 1)
    - End
    - Strand (REQUIRED)
    - Feature Name (REQUIRED)

Other range formats, continued

- GFF/GTF: General Feature Format (1-offset)
    - 9 Columns (see http://www.ensembl.org/info/website/upload/gff.html)
        - but 9th column is a COMMENT field that can pretty much hold anything arbtrary key/value pairs
- GTF: General Transfer Format == GFF v2
    - GFF with “rules” (kind of) about what goes in column 9

Other range formats; UCSC

General Formats:



Frequently Asked Questions: Data File Formats

General formats:

- Axt format
- BAM format
- BED format
- BED detail format
- bedGraph format
- bigBed format
- bigGenePred table format
- bigWig format
- Chain format
- GenePred table format
- GFF format
- GTF format
- HAL format
- MAF format

Swiss Army knife of range formats

BEDTOOLS

- Genome Arithmetic
- Handles:
  - BED
  - BAM
  - GFF/GTF
  - VCF
- Another package that is also very useful: GenomicRanges in R