MAX PLANCK INSTITUTE
OF IMMUNOBIOLOGY AND EPIGENETICS

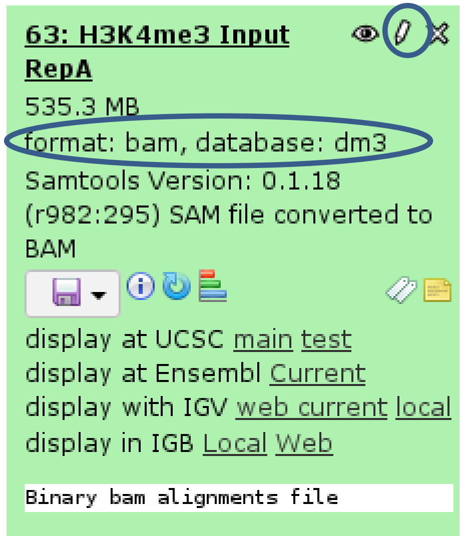
NGS and Galaxy data formats

Sarah Diehl, Friederike Dündar

Bioinformatics Facility, MPI-IE, Freiburg

Data formats in Galaxy

- most tools rely on very specific data format



63: H3K4me3 Input RepA

535.3 MB

format: bam, database: dm3

Samtools Version: 0.1.18 (r982:295) SAM file converted to BAM

display at UCSC main test

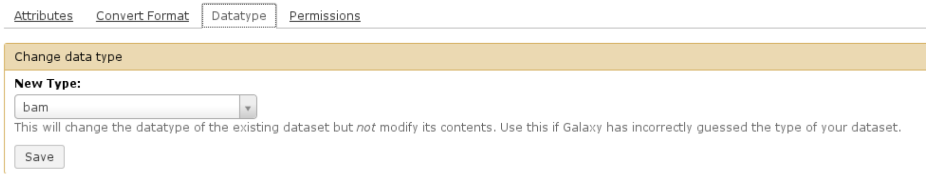
display at Ensembl Current

display with IGV web current local

display in IGB Local Web

```
Binary bam alignments file
```

if you want to change the format, click on "edit" and then choose "data type" within the main window



Change data type

New Type: bam

This will change the datatype of the existing dataset but *not* modify its contents. Use this if Galaxy has incorrectly guessed the type of your dataset.

deepTools data formats

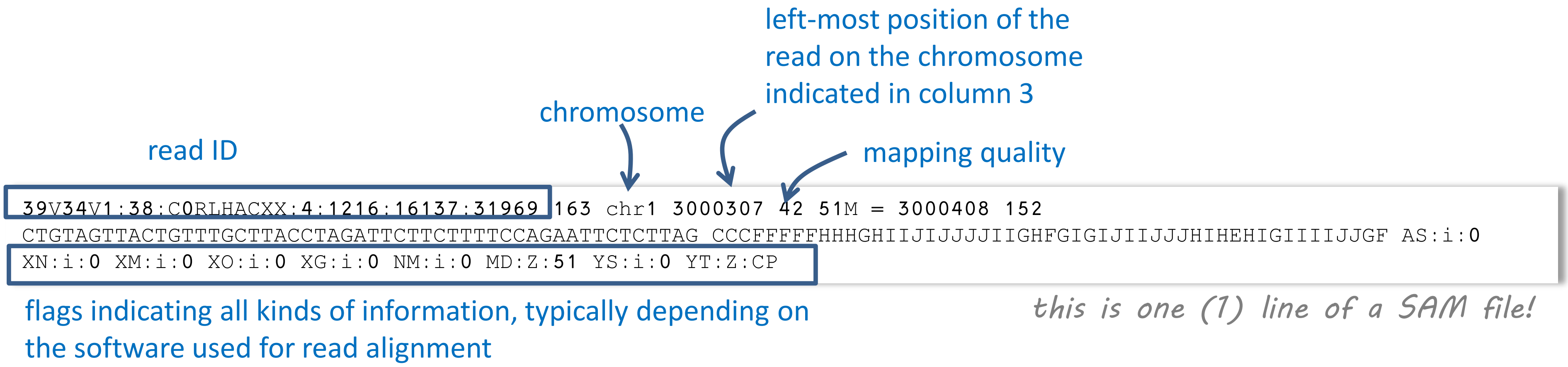
Tool	Input files	Output files
bamCorrelate	2 or more BAM files 1 BED	Image file Table of values
bamFingerprint	2 BAM files	Image file Table of counts
computeGCbias	1 BAM file 1 BED file	Image file Table of frequencies (→ correctGCbias)
correctGCbias	1 BAM file Table of frequencies	BAM or bedGraph or bigWig
bamCoverage	1 BAM file	bedGraph or bigWig
bamCompare	2 BAM files	bedGraph or bigWig
computeMatrix	1 bigWig 1 BED	Zipped matrix of values (→ heatmapper and profiler) Table of values for summary plot Table of values for heatmap BED file of regions used for the computation
heatmapper	Output of computeMatrix	Image file Table of values for summary plot Table of values for heatmap BED file of regions used for the computation
profiler	Output of computeMatrix	Image file Table of values for summary plot Table of values for heatmap BED file of regions used for the computation

black = required, grey = optional

can be viewed in Genome Browsers (might crash)

Formats: SAM/BAM

- preferred format for storage of **aligned** sequencing reads
- SAM = text file, BAM = binary (compressed) version of a SAM, not human-readable (i.e. you will not see anything in the main frame when you click on a BAM file)
- each line contains many information about each single read: where it aligned, how well it aligned, its DNA sequence, whether it has a mate read etc.
- many tools require sorted BAM files and an index file (usually indicated by the file ending .bai) – Galaxy generates those index files automatically as soon as you upload a BAM file
- make sure that the reference genome is always indicated for every data set within Galaxy („database“ entry)



left-most position of the read on the chromosome indicated in column 3

chromosome

read ID

mapping quality

```
39V34V1:38:C0RLHACXX:4:1216:16137:31969 163 chr1 3000307 42 51M = 3000408 152
CTGTAGTTACTGTTTGCTTACCTAGATTCTTCTTTTCCAGAATTCTCTTAG CCCFFFFFHHHGHIIJIJJJJIIGHFGIGIJIIJJJHIHEHIGIIIIJJGF AS:i:0
XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:Z:51 YS:i:0 YT:Z:CP
```

flags indicating all kinds of information, typically depending on the software used for read alignment

this is one (1) line of a SAM file!

for detailed information see http://samtools.sourceforge.net/SAMv1.pdf can be viewed in Genome Browsers

Formats: bedGraph/bigWig

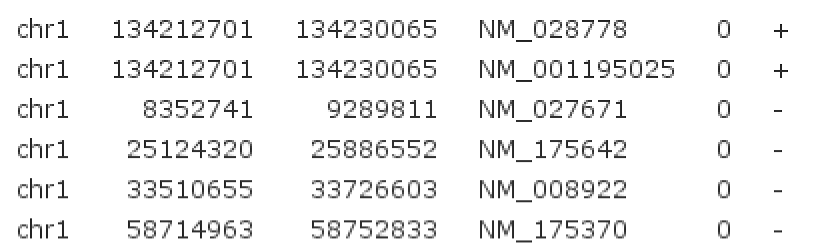
- preferred file formats for storage of genome-wide read coverages
- **bedGraph** = text file, **bigWig** = compressed version of bedGraph (not human-readable)
- no information about individual reads, instead: information about how many reads were mapped to each genomic locus
- much smaller in size than SAM/BAM files

```
chr2   100100   100120   5
chr2   100121   100141   3.2
chr2   100142   100163   13.8
```

these are three lines of a bedGraph file

chromosome, start position, end position, read coverage

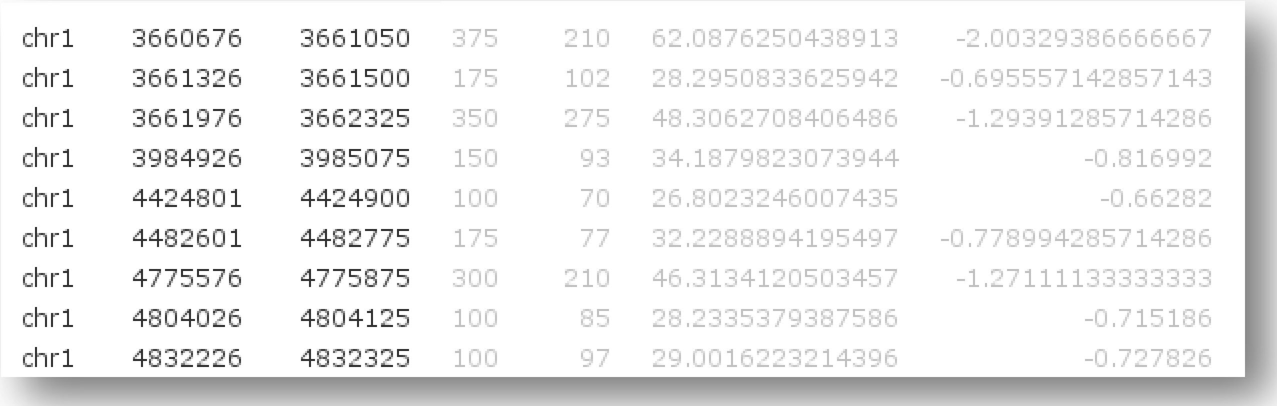
Formats: BED



```
chr1   134212701   134230065   NM_028778        0   +
chr1   134212701   134230065   NM_001195025     0   +
chr1     8352741     9289811   NM_027671        0   -
chr1    25124320    25886552   NM_175642        0   -
chr1    33510655    33726603   NM_008922        0   -
chr1    58714963    58752833   NM_175370        0   -
```

- most common format for **genomic regions** genome.ucsc.edu/FAQ/FAQformat.html#format1
- Column 1-3: same as interval
- Column 4: name
- Column 5: score
- Column 6: strand

Formats: interval

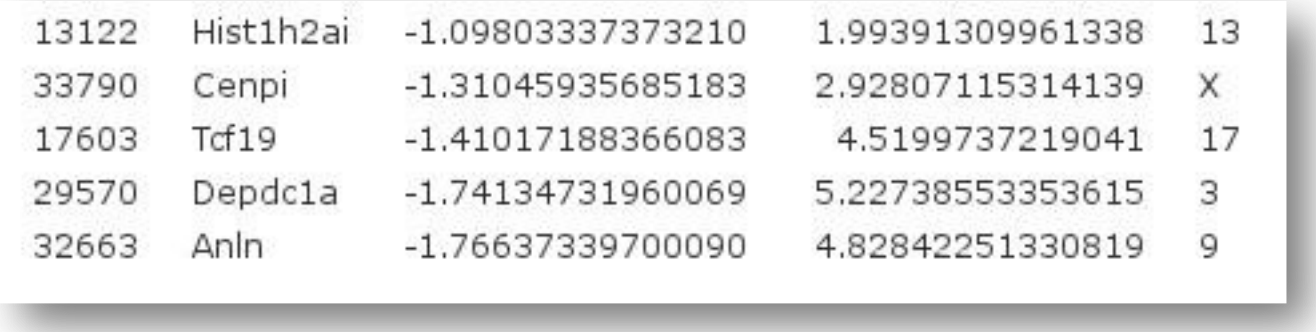


chr1	3660676	3661050	375	210	62.0876250438913	-2.00329386666667
chr1	3661326	3661500	175	102	28.2950833625942	-0.695557142857143
chr1	3661976	3662325	350	275	48.3062708406486	-1.29391285714286
chr1	3984926	3985075	150	93	34.1879823073944	-0.816992
chr1	4424801	4424900	100	70	26.8023246007435	-0.66282
chr1	4482601	4482775	175	77	32.2288894195497	-0.778994285714286
chr1	4775576	4775875	300	210	46.3134120503457	-1.27111133333333
chr1	4804026	4804125	100	85	28.2335379387586	-0.715186
chr1	4832226	4832325	100	97	29.0016223214396	-0.727826

- for **genomic regions**
- Column 1: chromosome
- Column 2: start position
- Column 3: end position
- other columns: anything

*much less stringent than BED format!
(i.e. much more tolerant as only the
first three columns are strictly
defined)*

Formats: tabular



13122	Hist1h2ai	-1.09803337373210	1.99391309961338	13
33790	Cenpi	-1.31045935685183	2.92807115314139	X
17603	Tcf19	-1.41017188366083	4.5199737219041	17
29570	Depdc1a	-1.74134731960069	5.22738553353615	3
32663	Anln	-1.76637339700090	4.82842251330819	9

- most simple format
- column-based
- separated by tabs
- similar to Excel tables

additional format information

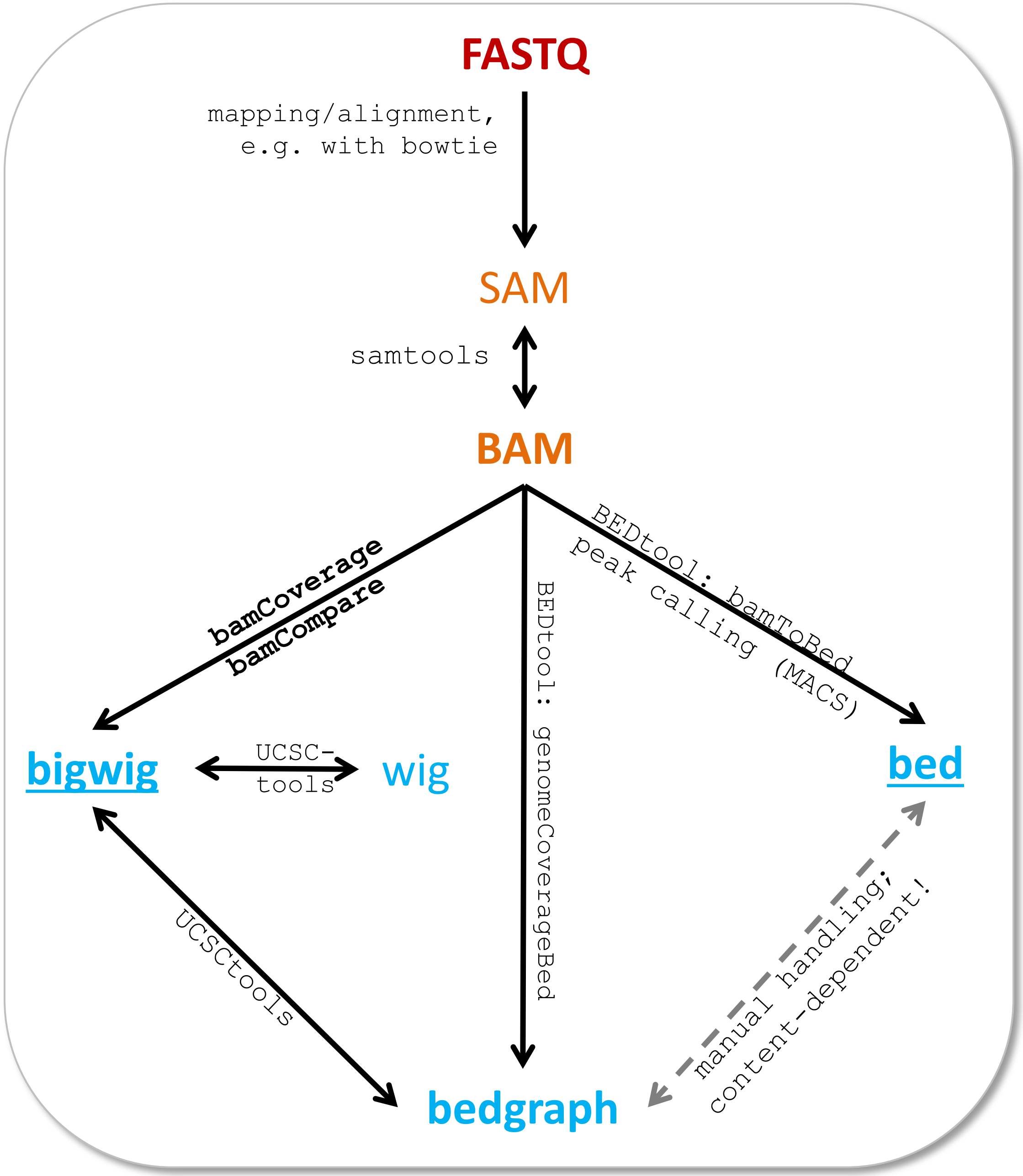
Sequences

- FASTA: wikipedia.org/wiki/FASTA_format
- FASTQ: wikipedia.org/wiki/FASTQ_format

Coverage

- BedGraph
  genome.ucsc.edu/goldenPath/help/bedgraph.html
- Wiggle
  genome.ucsc.edu/goldenPath/help/wiggle.html
- BigWig(gle)
  genome.ucsc.edu/goldenPath/help/bigWig.html

NGS data formats overview



FASTQ

mapping/alignment,
e.g. with bowtie

SAM

samtools

BAM

bamCoverage
bamCompare

BEDtool: genomeCoverageBed

BEDtool: bamToBed
peak calling (MACS)

bigwig

UCSC-
tools

wig

bed

UCSCtools

manual handling;
content-dependent!

bedgraph

reads: **sequence** only

reads:
sequence + genomic localization

coverage files:
read numbers per genomic bin