Agenda for today

Genomic data and data formats	14:00-15:00
Break	
Genomic tools on the cluster	15:00-15:50
AWK	15:50-16:30
Q&A session	16:30-17:00

High Performance Computing for genomic applications

Genomic formats

Scientific IT Services
Michal Okoniewski

https://siscourses.ethz.ch/hpc_genomics_2021/

Genomic data formats

- Sequence
  - fasta, fastq
- alignment formats
  - SAM, BAM, CRAM
- Variant description
  - VCF, BCF
- Annotation
  - GTF, GFF
- Genomic ranges
  - BED, WIG, bigWIG

Raw sequence data: fasta

```
>read_no_1
CGGCCTGGAGGCCCTGCAGAACCTGCTGGGCTACAGGTTCGGCGACGAGGG

>read_no_2
GCAGCGTGAGCGCCATCATGGGCAACCCCCAGGTGAAGGCCCACGGCAAGA

>read_no_3
GGGAGACACCCGCACGTGTGGCCCGCATGTATGCTGAGCTCTTCCGCGGAT

>read_no_4
TTTGCCCCGCATCGAGCGGGCTGTGCGGGAAATCCTTCTGGCTGTAGGCGA

>read_no_5
CCTGTGGGGCAAGGTGAACCCCGTGGAGATCGGCGCCGAGAGCCTGGCCAG

>read_no_6
GAGGAGGGCCAGGATCCACCAGAGGAAGGGCCTGCTGTGGTTCATCCCCGC

>read_no_7
CTGCACAGCGACTACAACCTGACCTGGTACAGGAACGGCAGCAACATGCCC

>read_no_8
GTGCTGGGCCTGGCCATCAGCCACTTCCTGCTGGAGCAGTTCCCCGACTAC

>read_no_9
AACCTGGGCGAGTACCTGCTGCTGGGCAAGGGCGAGGAGATGACCGGCGGC

>read_no_10
GTTCCCCGACTACAACGAGGGCGAGCTGAGCAGGCTGAGGAGCGCCATCGT

>read_no_11
CTTCAGCAAGTTCGGCGACCTGAGCAGCGTGAGCGCCATCATGGGCAACCC

>read_no_12
ACCAGAGGAAGGGCCTGCTGTGGTTCATCCCCGCCGCCCTGGAGGACAGCG

>read_no_13
AAGGGCGAGGAGATGACCGGCGGCAGGAGGAAGGCCAGCCTGCTGGCCGAC
```

Raw sequencing data: fastq

```
@700523F:164:CBE5UANXX:5:2302:1106:2069 1:N:0:TCCATTGC
NTCCCCATGCACAGAATATTCAGGCATTTGAAGCCTGCTTTAAGCACTCTAATTTGTTCAAAGTAATAGTACCGGCCCACAATAACACTCGTTTAAGAGCACTAATGCAGGTTTTTAGATCGGAAG
+
#<=BAGBGGGFGGGGGGGGGGGGGGGGGGGGGG>EGG1FGGGGGFGGGGGGGGGGGGGGGGGGEDGGGG::FGEGGEGGGFGGFGFGGG0CGGFGGGGGGGGG000C?FGGGGBGGGGGG/FGBAFG,
@700523F:164:CBE5UANXX:5:2302:1247:2179 1:N:0:TCCATTGC
CACCCAATTAAAATTCTACATGTCATTCCTTGGTCTTCCCTGTCTTTGTAACTTGGAATTCTTATGCAACTTTATAAAGGAAACATGTAATAAGAAAGCGTTTATAATTTAACAGCTCATTCTTGA
+
CCCCCGGGGGGGGGGGGGGGGGGGGGGGGGGFGGGGGGGGCGFGGGEGGGGGEFGGGGGGGGGG:EE>GFFGGGGGGCDFCFGGGFGDGGGFDGFFFG;FGAFGGGGGGGCFG?GGCGGCFGF//
@700523F:164:CBE5UANXX:5:2302:1328:2140 1:N:0:TCCATTGC
CTGCAAGAAACGGATTGTCTGCCGCGATTCTCCAGCGACTGAACATCTCAACACTTGCAAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTCCATTGCTCCGCTCTATCTCGTATGCCGTCTT
+
CCCCCGBG1BEDG/FGGGGGGCG@GGGGGFGG@FGGGDGGGGGGGGFGGG111EGGFGGGGBFGGGGGGGGGGGG:EGGGGGCFGFGGGGCBDFCCFGGGGGGFGFFCGGGGGB:CDGGGGGGGGG
@700523F:164:CBE5UANXX:5:2302:1462:2171 1:N:0:TCCATTGC
TACATTATTCTATCGACTAGAGACTCTTCACCTTGGAGACCAGCTGCGGATATTAGTACGGCCTGTTGAGAAGTTTGCGTATCCCCACCATAAATTTTCAAGGTCCGAGGAGAAAATATCGAGATC
+
CCCCCGGGGGGGGGCGEGGG@FGGGGGGGCGGGGCFGGGGGGGGGGGGGGGGGGFDGGGGEAGGGGGGEFFCFFGGG>GDEF=FEFGGGGGEGGGG;<FFGGC:CGGG@DGGGGGG>.5.67777,7
@700523F:164:CBE5UANXX:5:2302:1499:2246 1:N:0:TCCATTGC
CCTGAATTTAAGCATATTAATTAGGGGAGGAAAAGAAACTAACAAGGATTTTCTTAGTAGCGGCGAGGGAAAAGAAAACAGTTCAGCACTAAGTCACTTTGTCTATATGGCAAAGATCAGAAGAGA
+
CCCCCGDGGFGGGG>1@FCGEFFGGGGGGGBGGGGGFG1>CFG1>CFFGGGGG=FGGGG11E/;EE9/=8;0>80<CF<<0;C0CGGG0<<=CCG0<B:DDG=GGG095C<:GG6..5.7?E.7B-
@700523F:164:CBE5UANXX:5:2302:1628:2124 1:N:0:TCCATTGC
CCTAAGACAGTATTAAAAACAGCCACATTTAGCTGTAAGCAGAGTTAAAGGCTGTGCTTTTATAATTTTGAAATATATATTAAAATATTTATTTTATTAAAAATAAGAGAAGTAAACTAGATTTAA
+
CCCCCGGGGFGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG>EGGFFGGGG=EFGFFGC>GGGGGG>:=EFGFGGGFGGGGGGGGGGGGGGGGGGGG@FGGGG08/BGGGGG//<FEGG:FFC
@700523F:164:CBE5UANXX:5:2302:2129:2178 1:N:0:TCCATTGC
AATTTCCCTTGCAAAACTGCCTAGTATCTGCATTCTTAAGCTTCGTTTCTACTTCTGTGTGCGTTGACTGAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTCCATTGCTCGAAGTGATCTCG
+
CCCCCFFGGGGGGGEGGGGGGGGGGGGGGGG>GGGGGFGGGGGEGGGGGGGCGGGG@GGGEGGGD/0=;EFEFGDDBG=C==FGGGGGGGGFGGG=FGGGGGGEGGFGGGGGGGG@GGGGGGGGG-
```

Fastq record structure

PHRED scores

- Quality as PHRED score

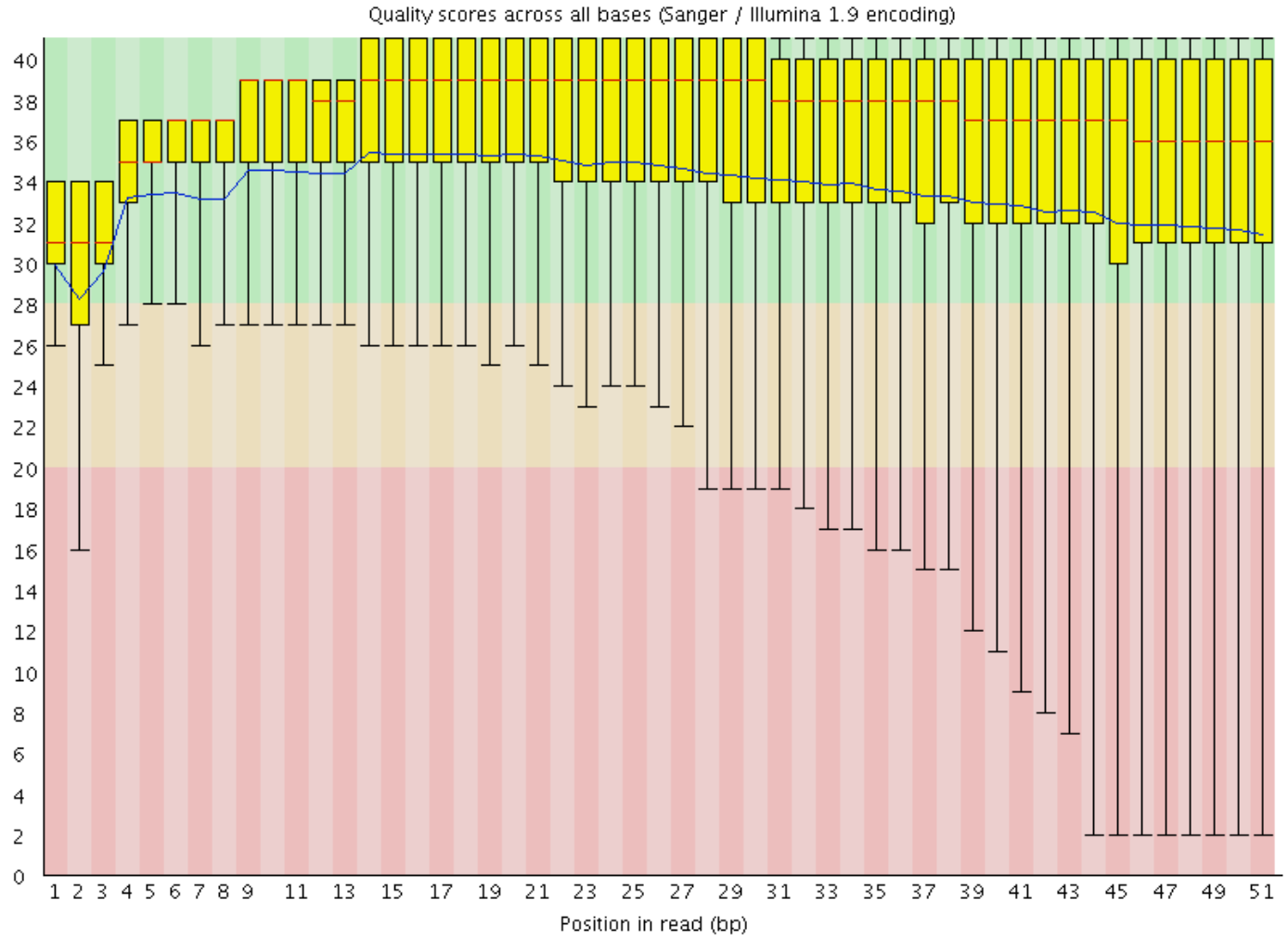
Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.90%
40	1 in 10,000	99.99%
50	1 in 100,000	100.00%
60	1 in 1,000,000	100.00%

- Phred+33 in ASCII:

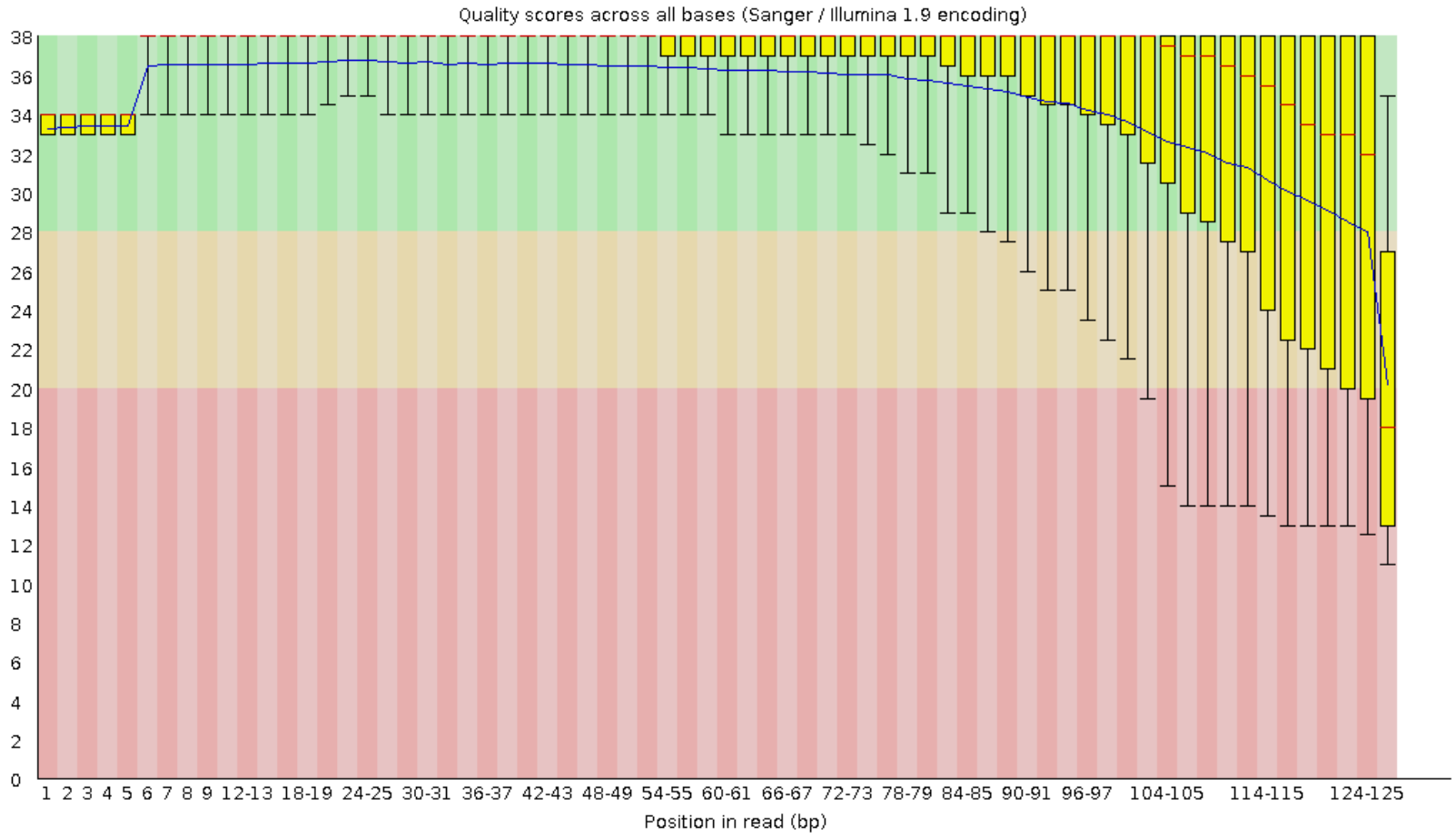
```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
0.........................26...31.......40
```

https://github.com/brentp/bio-playground/blob/master/reads-utils/guess-encoding.py

Fastqc – per base sequence quality check



Fastqc – per base sequence quality check



What can be done with raw sequences

- Quality filtering and summaries (picard/htsjdk,…)
- Trimming (cutadapt, trimmomatic,…)
- Alignment (bowtie, bwa, SHRiMP, STAR,…)
- De-novo assembly (Trinity, velvet, SPADES, SOAPdenovo…)
- …

Alignment formats - SAM

- https://samtools.github.io/hts-specs/SAMv1.pdf

BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 16 2009, pages 2078–2079
doi:10.1093/bioinformatics/btp352

Sequence analysis

The Sequence Alignment/Map format and SAMtools

Heng Li[1,†], Bob Handsaker[2,†], Alec Wysoker[2], Tim Fennell[2], Jue Ruan[3], Nils Homer[4], Gabor Marth[5], Goncalo Abecasis[6], Richard Durbin[1,*] and 1000 Genome Project Data Processing Subgroup[7]

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, UK, [2]Broad Institute of MIT and Harvard, Cambridge, MA 02141, USA, [3]Beijing Institute of Genomics, Chinese Academy of Science, Beijing 100029, China, [4]Department of Computer Science, University of California Los Angeles, Los Angeles, CA 90095, [5]Department of Biology, Boston College, Chestnut Hill, MA 02467, [6]Center for Statistical Genetics, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA and [7]http://1000genomes.org

Received on April 28, 2009; revised on May 28, 2009; accepted on May 30, 2009

Advance Access publication June 8, 2009

Alignment formats - SAM

- Fields of a record in SAM format

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,254}`	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

SAM format – CIGAR strings

- `Example: 26M987N22M1S`

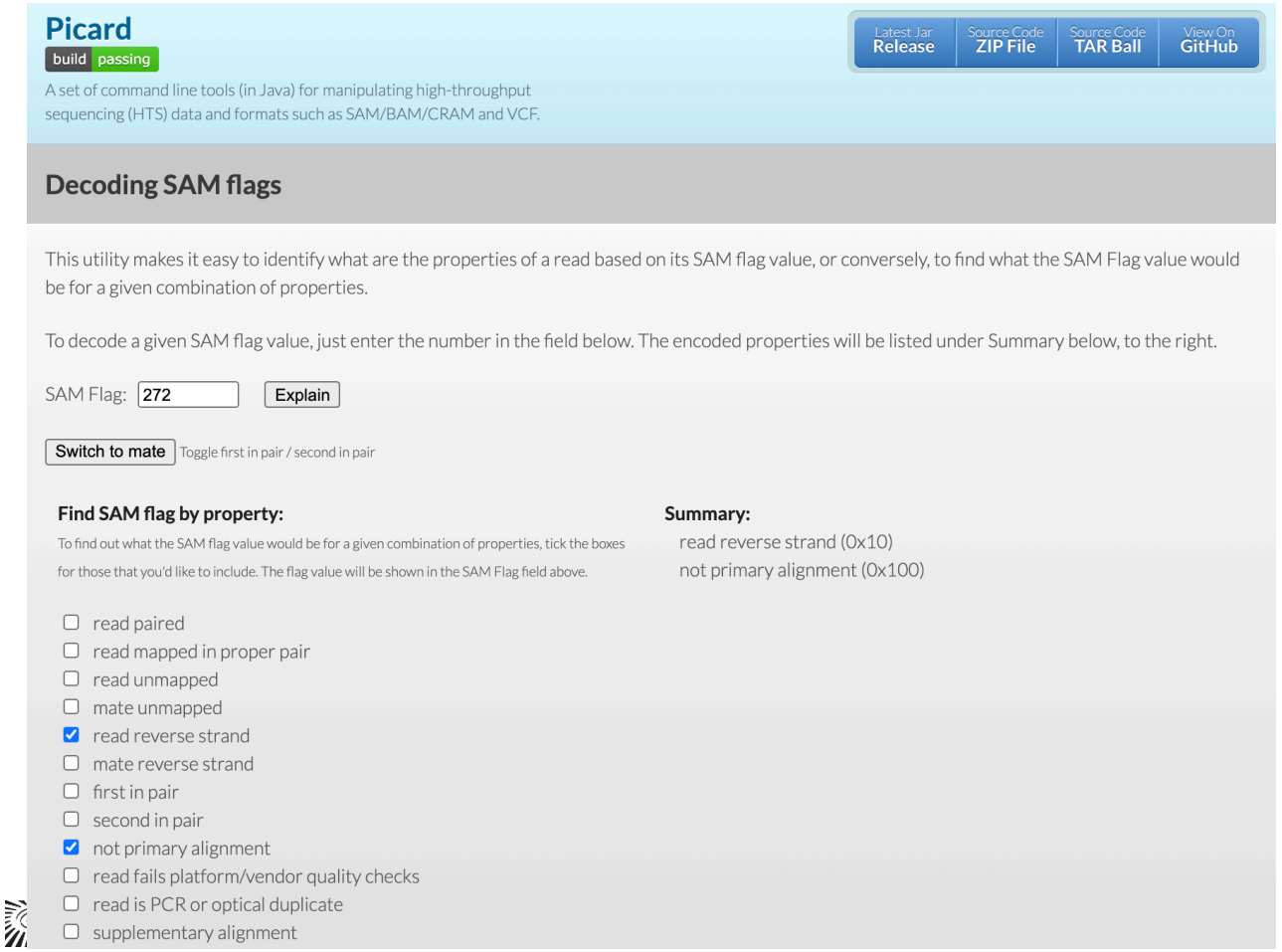
Op	BAM	Description
`M`	0	alignment match (can be a sequence match or mismatch)
`I`	1	insertion to the reference
`D`	2	deletion from the reference
`N`	3	skipped region from the reference
`S`	4	soft clipping (clipped sequences present in SEQ)
`H`	5	hard clipping (clipped sequences NOT present in SEQ)
`P`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

Alignment formats - SAM

```
@SQ     SN:chrM LN:16571
@SQ     SN:chr1 LN:249250621
@SQ     SN:chr2 LN:243199373
@SQ     SN:chr3 LN:198022430
SRR1012931.32   0       chr12   52632509        255     1S48M   *       0       0       NGTGGATGCCCTGAATGATGAGATCAACTTCCTCAGGACCCTCAATGAG
       BP\ccceeeggggiiiiiihiihihiiiiiiiiiiiihiiibghhiiii       NH:i:1  HI:i:1  AS:i:47 nM:i:0
SRR1012931.33   16      chr16   88925155        255     48M1S   *       0       0       GGAGCTGTACCTGGGCCTGCTCTACCCCACGGAGGACTACAAGGTATAC
       iiiihhheciiiiiiiiihiihgeiiiihhiiiiiggggeeeeebab       NH:i:1  HI:i:1  AS:i:45 nM:i:1
SRR1012931.34   16      chr17   60111322        255     24M1478N24M1S   *       0       0       TTAGTCCAAATGGGCATAAGATAACTTGAAATGGGCTATTAGACTGTTN
iiiiiiiiiiiigebihiiiihhhiiiiiiiiiiiiggggeceec\SB       NH:i:1  HI:i:1  AS:i:45 nM:i:0
SRR1012931.35   16      chr20   20031227        255     43M6S   *       0       0       CAGCAGTTATCTGTACCTCAGCCGGGGCTTTGTTTTTCACCTTGGCCAN
       fhfhhghehiiighfghgfhhgfhhgbfhihdihihgc`feccecaYPB       NH:i:1  HI:i:1  AS:i:42 nM:i:0
SRR1012931.36   0       chr3    48680414        255     49M     *       0       0       GCCTTGCCTCCTCAGGCGCTGCCTTCCTGCCCAGACAGGCTGGCATCCA
       bbbeeeeeggggghfgbfgfhiiiihihhihiffcghiiihiiifgiih       NH:i:1  HI:i:1  AS:i:48 nM:i:0
SRR1012931.37   16      chrX    70510616        255     26M987N22M1S    *       0       0       AAATGGGCAACAGGCCAGCAGCCAAAATGAAGGCTTGACTATTGACCTN
ihg`fagfafcgf\gggeigfeifhhhhfhiiiiiigggggcccaa\PB       NH:i:1  HI:i:1  AS:i:45 nM:i:0
SRR1012931.38   0       chr5    139928964       255     1S48M   *       0       0       NTGACTTGTTAGTTCCAGGCCTCCTTTAGTTCTGAGGCAGCTAGACCAG
       BP\ccceeeggggiiiiiihiiiiiiiifhhihiiggfghhhihiiiihg       NH:i:1  HI:i:1  AS:i:47 nM:i:0
SRR1012931.39   16      chr11   6477532 255     49M     *       0       0       CGGAAGACGAGCTCATCCTCAATTGGGTTGTCCTTTCGTTTGCCGCCTG
eeggggiihgifffiihhiiiiiiiiiihghiiiggggeeeeebbb       NH:i:1  HI:i:1  AS:i:48 nM:i:0
SRR1012931.40   0       chr2    3502293 255     49M     *       0       0       GTGGGAGCTCTTCCCCCTACCACTCCCCACCCCAAGGCATCATTTTGGA
bbbeeeeeggggiiiiiiiiiihiiiihiiii_effffhhihhiehc       NH:i:1  HI:i:1  AS:i:48 nM:i:0
SRR1012931.41   0       chr8    38099785        255     49M     *       0       0       GTTTCCTTGAACTTGCTACAGACACATTTTAAGAAAGCCCAAGAAAATC
       babeeeeeggggghiiiiiiiiiiiiiiiiihhiiiiiiiiiiiihgh       NH:i:1  HI:i:1  AS:i:48 nM:i:0
SRR1012931.42   0       chr1    175891126       3       49M     *       0       0       CTGTACTGGAGCCACCCGCGAAAATTCGGCCAGGGTTCTCGCTCTTGTC
       bbbeeeeeggggiihiiiheihhgighiiiihiihiiiiiihfgggcd       NH:i:2  HI:i:1  AS:i:48 nM:i:0
```

Bitwise SAM flags

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




```
> table(tmp2[,1])

      0       4      16     256     272
7522355 5354679 9366254 49581206 74347461
```

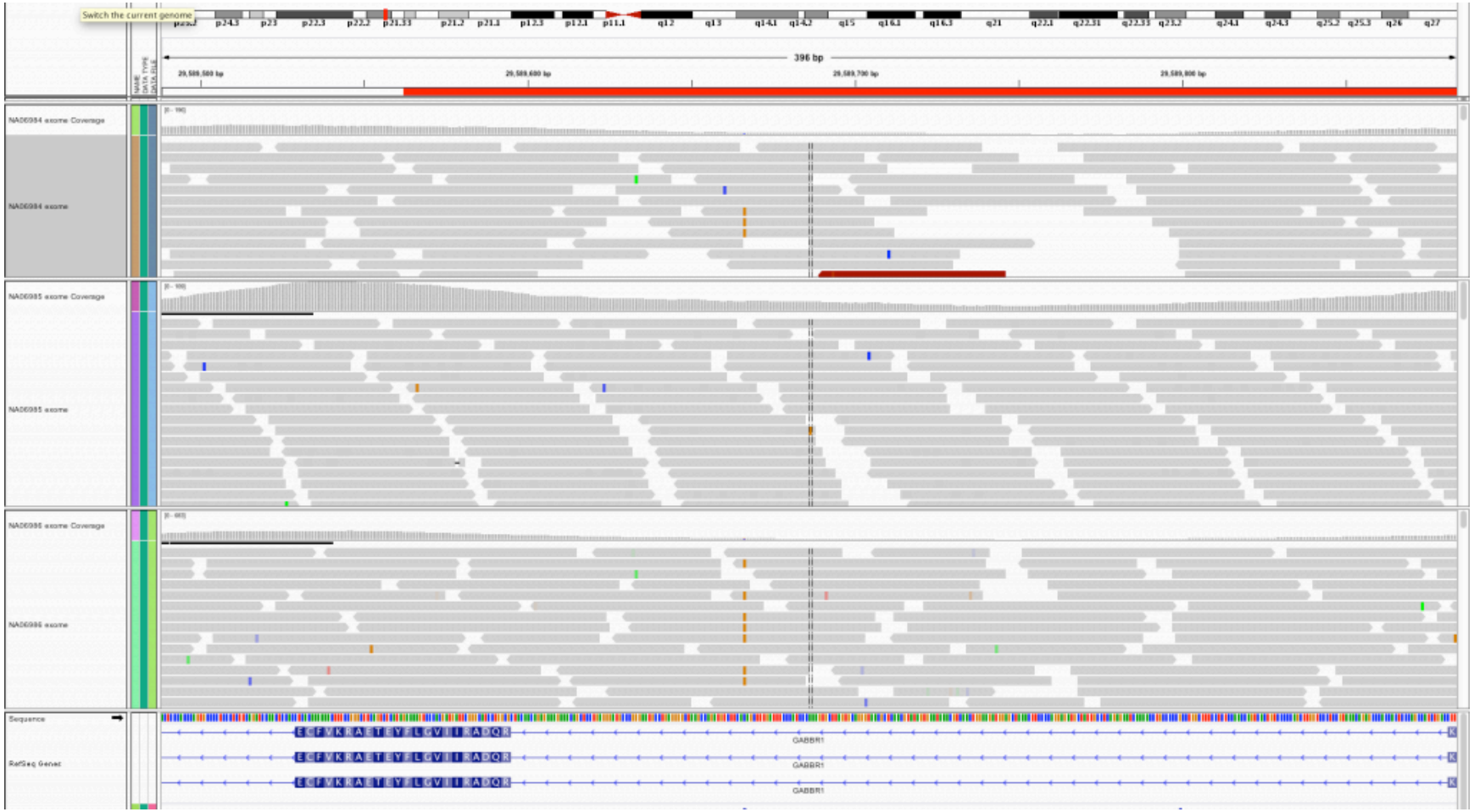
Alignment formats - BAM

- BGZF compressed SAM
- Indexed with R-tree (BAI file)
- BAM and BAI must be placed together
- BAM files may be viewed in the IGV browser
- Other formats available: CRAM, ADAM

What can be done with the alignments

- Processing formats (samtools)
  - http://www.htslib.org/
- Genomic feature extraction
  - Variant calling (samtools pileup, GATK,…)
  - Counting reads according to annotations (HTSeq, featureCount)
  - Junction and isoform discovery (cufflinks, MISO,…)
- Visualization (IGV,…)

IGV browser


Switch the current genome
396 bp
NA06984 exome Coverage
NA06984 exome
NA06985 exome Coverage
NA06985 exome
NA06986 exome Coverage
NA06986 exome
Sequence
RefSeq Genes
GABBR1

Annotation formats – GTF/GFF

- **seqname** - name of the chromosome or scaffold; chromosome names can be given with or without the 'chr' prefix. **Important note**: the seqname must be one used within Ensembl, i.e. a standard chromosome name or an Ensembl identifier such as a scaffold ID, without any additional content such as species or assembly. See the example GFF output below.
- **source** - name of the program that generated this feature, or the data source (database or project name)
- **feature** - feature type name, e.g. Gene, Variation, Similarity
- **start** - Start position of the feature, with sequence numbering starting at 1.
- **end** - End position of the feature, with sequence numbering starting at 1.
- **score** - A floating point value.
- **strand** - defined as + (forward) or - (reverse).
- **frame** - One of '0', '1' or '2'. '0' indicates that the first base of the feature is the first base of a codon, '1' that the second base is the first base of a codon, and so on..
- **attribute** – (column 9) A semicolon-separated list of tag-value pairs, providing additional information about each feature.

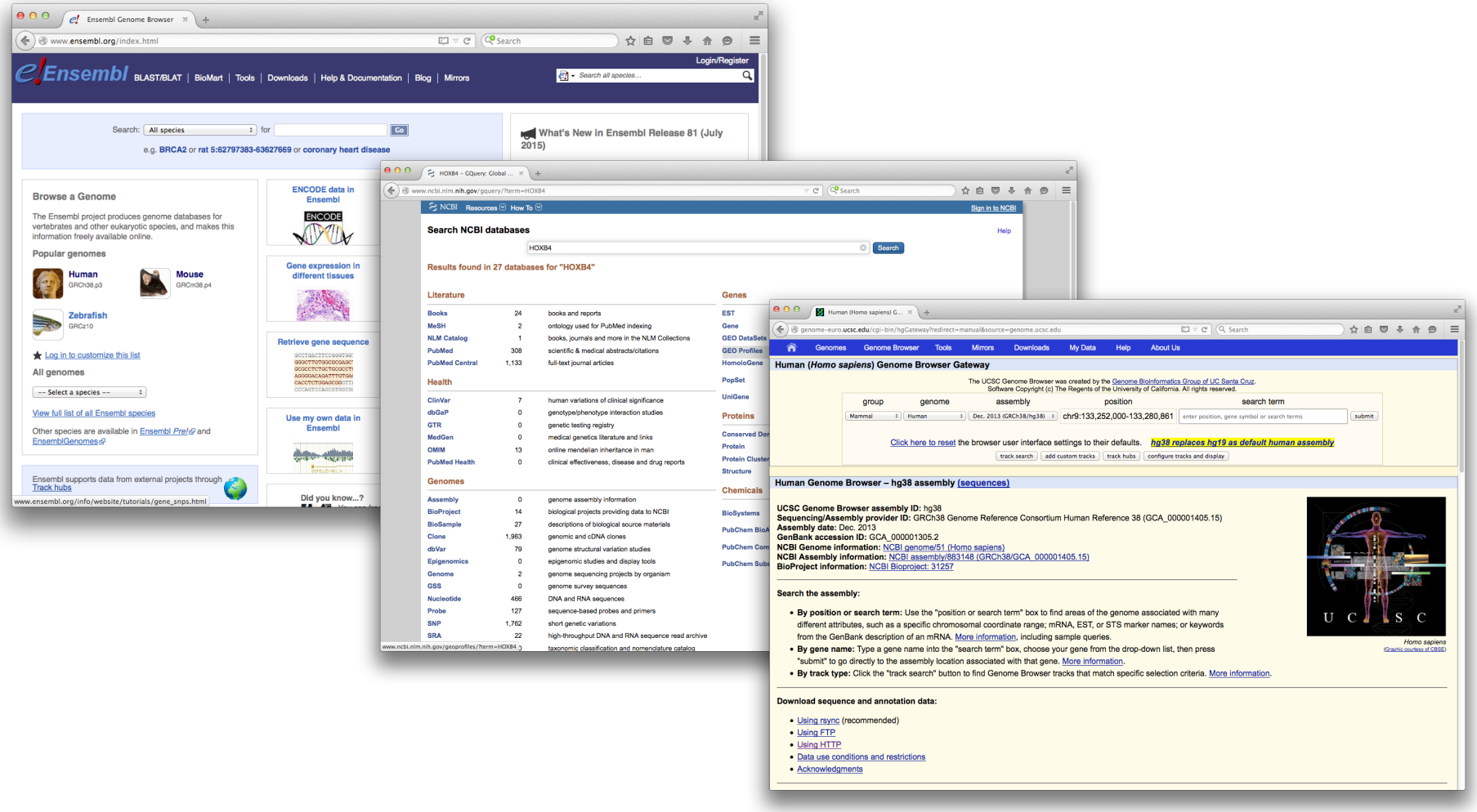
Annotation formats – GTF/GFF

```
11	protein_coding	CDS	96166204	96166448	.	+	2	exon_number "2"; gene_biotype
"protein_coding"; gene_id "ENSMUSG00000038700"; gene_name "Hoxb5"; p_id "P37298"; protein_id "ENSMUSP00000035423";
transcript_id "ENSMUST00000049272"; transcript_name "Hoxb5-001"; tss_id "TSS63862";
11	protein_coding	exon	96166204	96167434	.	+	.	exon_number "2"; gene_biotype
"protein_coding"; gene_id "ENSMUSG00000038700"; gene_name "Hoxb5"; p_id "P37298"; transcript_id "ENSMUST00000049272";
transcript_name "Hoxb5-001"; tss_id "TSS63862";
11	antisense	exon	96166296	96166395	.	-	.	exon_number "1"; gene_biotype
"antisense"; gene_id "ENSMUSG00000085645"; gene_name "0610040B09Rik"; transcript_id "ENSMUST00000140952";
transcript_name "0610040B09Rik-002"; tss_id "TSS16113";
11	protein_coding	stop_codon	96166449	96166451	.	+	0	exon_number "2";
gene_biotype "protein_coding"; gene_id "ENSMUSG00000038700"; gene_name "Hoxb5"; p_id "P37298"; transcript_id
"ENSMUST00000049272"; transcript_name "Hoxb5-001"; tss_id "TSS63862";
11	antisense	exon	96168051	96168224	.	-	.	exon_number "1"; gene_biotype
"antisense"; gene_id "ENSMUSG00000085645"; gene_name "0610040B09Rik"; transcript_id "ENSMUST00000150698";
transcript_name "0610040B09Rik-001"; tss_id "TSS73537";
11	protein_coding	exon	96177998	96180537	.	+	.	exon_number "1"; gene_biotype
"protein_coding"; gene_id "ENSMUSG00000038692"; gene_name "Hoxb4"; p_id "P29329"; transcript_id "ENSMUST00000049241";
transcript_name "Hoxb4-001"; tss_id "TSS19948";
11	non_coding	exon	96178479	96178588	.	+	.	exon_number "1"; gene_biotype
"non_coding"; gene_id "ENSMUSG00000092205"; gene_name "Mir10a"; transcript_id "ENSMUST00000173319"; transcript_name
"Mir10a-001"; tss_id "TSS15362";
11	miRNA	exon	96178479	96178588	.	+	.	exon_number "1"; gene_biotype "miRNA";
gene_id "ENSMUSG00000065519"; gene_name "Mir10a"; transcript_id "ENSMUST00000083585"; transcript_name "Mir10a-201";
tss_id "TSS15362";
11	protein_coding	CDS	96180084	96180537	.	+	0	exon_number "1"; gene_biotype
"protein_coding"; gene_id "ENSMUSG00000038692"; gene_name "Hoxb4"; p_id "P29329"; protein_id "ENSMUSP00000048002";
transcript_id "ENSMUST00000049241"; transcript_name "Hoxb4-001"; tss_id "TSS19948";
11	protein_coding	start_codon	96180084	96180086	.	+	0	exon_number "1";
gene_biotype "protein_coding"; gene_id "ENSMUSG00000038692"; gene_name "Hoxb4"; p_id "P29329"; transcript_id
"ENSMUST00000049241"; transcript_name "Hoxb4-001"; tss_id "TSS19948";
```

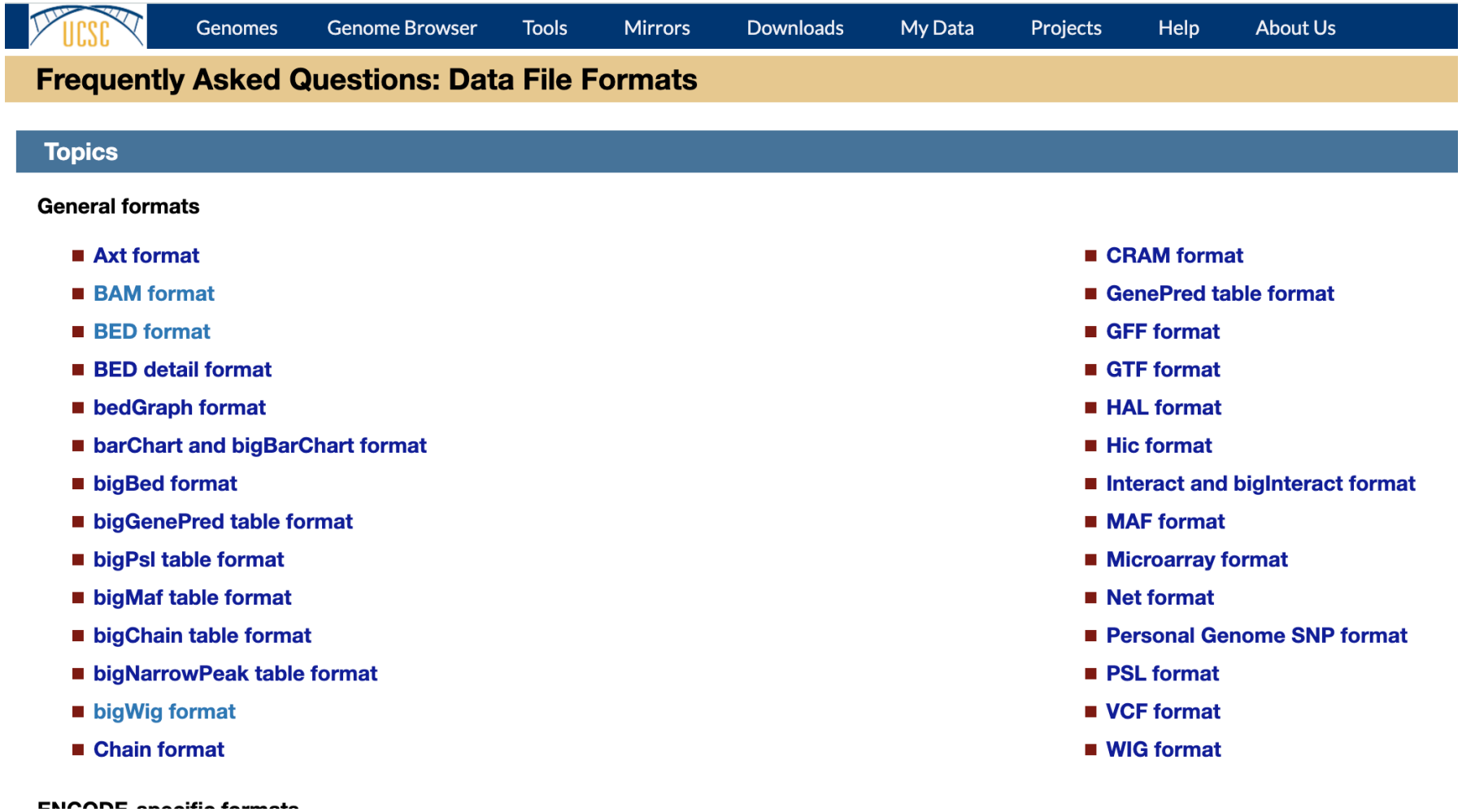
Annotation formats

- There are “dialects” of those formats
  - In particular they differ on the structure of field 9
- Conversion is not always possible
- GFF2 < GTF < GFF3
- Parsing or filtering often needed

Annotation repositories



Genomic ranges formats: BED, WIG, bigWIG



UCSC | Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Projects | Help | About Us

Frequently Asked Questions: Data File Formats

Topics

General formats

- Axt format
- BAM format
- BED format
- BED detail format
- bedGraph format
- barChart and bigBarChart format
- bigBed format
- bigGenePred table format
- bigPsl table format
- bigMaf table format
- bigChain table format
- bigNarrowPeak table format
- bigWig format
- Chain format
- CRAM format
- GenePred table format
- GFF format
- GTF format
- HAL format
- Hic format
- Interact and bigInteract format
- MAF format
- Microarray format
- Net format
- Personal Genome SNP format
- PSL format
- VCF format
- WIG format

ENCODE-specific formats

Genomic ranges formats: BED, WIG, bigWIG

```
browser position chr7:127471196-127495720
browser hide all
track name="ItemRGBDemo" description="Item RGB demonstration" visibility=2 itemRgb="On"
chr7    127471196  127472363  Pos1  0  +  127471196  127472363  255,0,0
chr7    127472363  127473530  Pos2  0  +  127472363  127473530  255,0,0
chr7    127473530  127474697  Pos3  0  +  127473530  127474697  255,0,0
chr7    127474697  127475864  Pos4  0  +  127474697  127475864  255,0,0
chr7    127475864  127477031  Neg1  0  -  127475864  127477031  0,0,255
chr7    127477031  127478198  Neg2  0  -  127477031  127478198  0,0,255
chr7    127478198  127479365  Neg3  0  -  127478198  127479365  0,0,255
chr7    127479365  127480532  Pos5  0  +  127479365  127480532  255,0,0
chr7    127480532  127481699  Neg4  0  -  127480532  127481699  0,0,255
```

Trends to watch GA4GH

- GA4GH is for genomic and health data like W3C standards for the web
  - Work streams: cloud, regulatory, clinical, scaling, security…
  - Driver projects
  - Annual conference
  - Various level of „hard evidence“
- Genomic data toolkit
  - https://www.ga4gh.org/genomic-data-toolkit/
  - a mix of old and new formats

Global Alliance
for Genomics & Health
Collaborate. Innovate. Accelerate.

ETH zürich

Genomic formats…