CEITEC

Central European Institute of Technology

BRNO | CZECH REPUBLIC

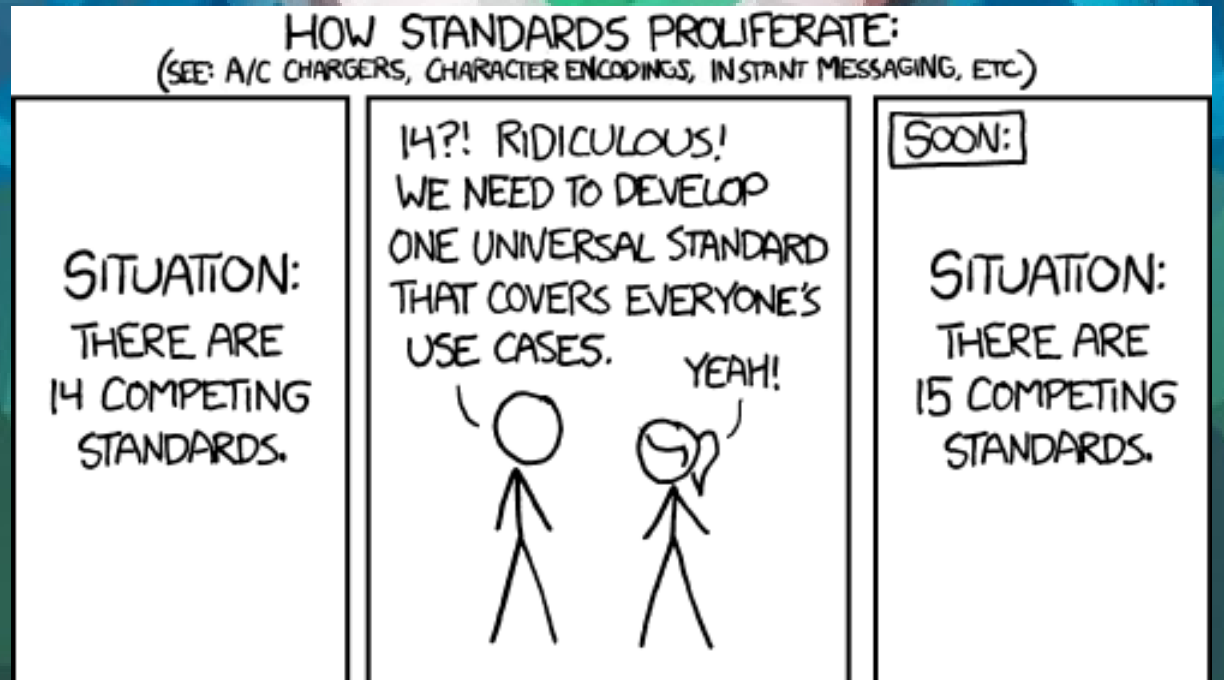
UNIVERSITAS MASARYKIANA BRUNENSIS

Introduction to Bioinformatics (LF:DSIB01)

Week 3 : Filetypes and Browser

Encoding Genomic Information for Bioinformatics Use

- Location Based Formats (.bed)
- Count/Coverage Based Formats (.bedgraph .wig)
- Feature Based Formats (.gtf)
- Sequence Based Formats (.fasta .fastq)
- Multiple Alignment Files
- Alignment Based Formats (.sam)


HOW STANDARDS PROLIFERATE:
(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC)
SITUATION: THERE ARE 14 COMPETING STANDARDS.
14?! RIDICULOUS! WE NEED TO DEVELOP ONE UNIVERSAL STANDARD THAT COVERS EVERYONE'S USE CASES.
YEAH!
SOON:
SITUATION: THERE ARE 15 COMPETING STANDARDS.

BED (Browser Extensible Data) file format

- Tab Delimited Text File
- Number of columns consistent per line
- No Empty fields (some can have “.” as N/A)

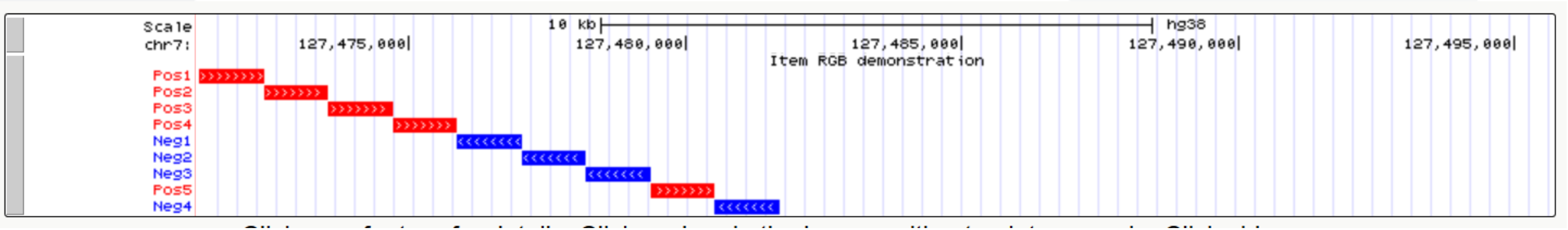
BED 3 columns : chrom, chromStart, chromEnd
BED 6 columns : BED3 + name, score, strand
BED 12 columns : BED6 + thickStart, thickEnd, itemRGB, blockCount, blockSizes, blockStarts

```
chr22 1000 5000 cloneA 960 + 1000 5000 0 2 567,488, 0,3512
chr22 2000 6000 cloneB 900 - 2000 6000 0 2 433,399, 0,3601
```

https://genome.ucsc.edu/FAQ/FAQformat.html#format1

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

BED 3 columns : chrom, chromStart, chromEnd
BED 6 columns : BED3 + name, score, strand
BED 12 columns : BED6 + thickStart, thickEnd, itemRGB, blockCount, blockSizes, blockStarts



BED pros and cons

- Generic
- Human Readable
- Useful for simple genomic loci

- Awkward handling of splice events

- Not useful for variable scores
- Must repeat BED3 info every line

bedGraph

- Used to display continuous data
- Header “browser” + Header “track” + Sorted BED lines (chr, start, stop, score)

```
browser position chr19:49302001-49304701
browser hide all
browser pack refGene encodeRegions
browser full altGraph
#       300 base wide bar graph, autoScale is on by default == graphing
#       limits will dynamically change to always show full range of data
#       in viewing window, priority = 20 positions this as the second graph
#       Note, zero-relative, half-open coordinate system in use for bedGraph format
track type=bedGraph name="BedGraph Format" description="BedGraph format" visibility=full color=200,100,0 altColor=0,100,200 priority=20
chr19 49302000 49302300 -1.0
chr19 49302300 49302600 -0.75
chr19 49302600 49302900 -0.50
chr19 49302900 49303200 -0.25
chr19 49303200 49303500 0.0
chr19 49303500 49303800 0.25
chr19 49303800 49304100 0.50
chr19 49304100 49304400 0.75
chr19 49304400 49304700 1.00
```

wiggle file (.wig)

- Comes in two flavors: variablestep vs fixed step

```
variableStep  chrom=chrN
[span=windowSize]
  chromStartA  dataValueA
  chromStartB  dataValueB
  ... etc ...  ... etc ...
```

```
variableStep chrom=chr2
300701 12.5
300702 12.5
300703 12.5
300704 12.5
300705 12.5
```

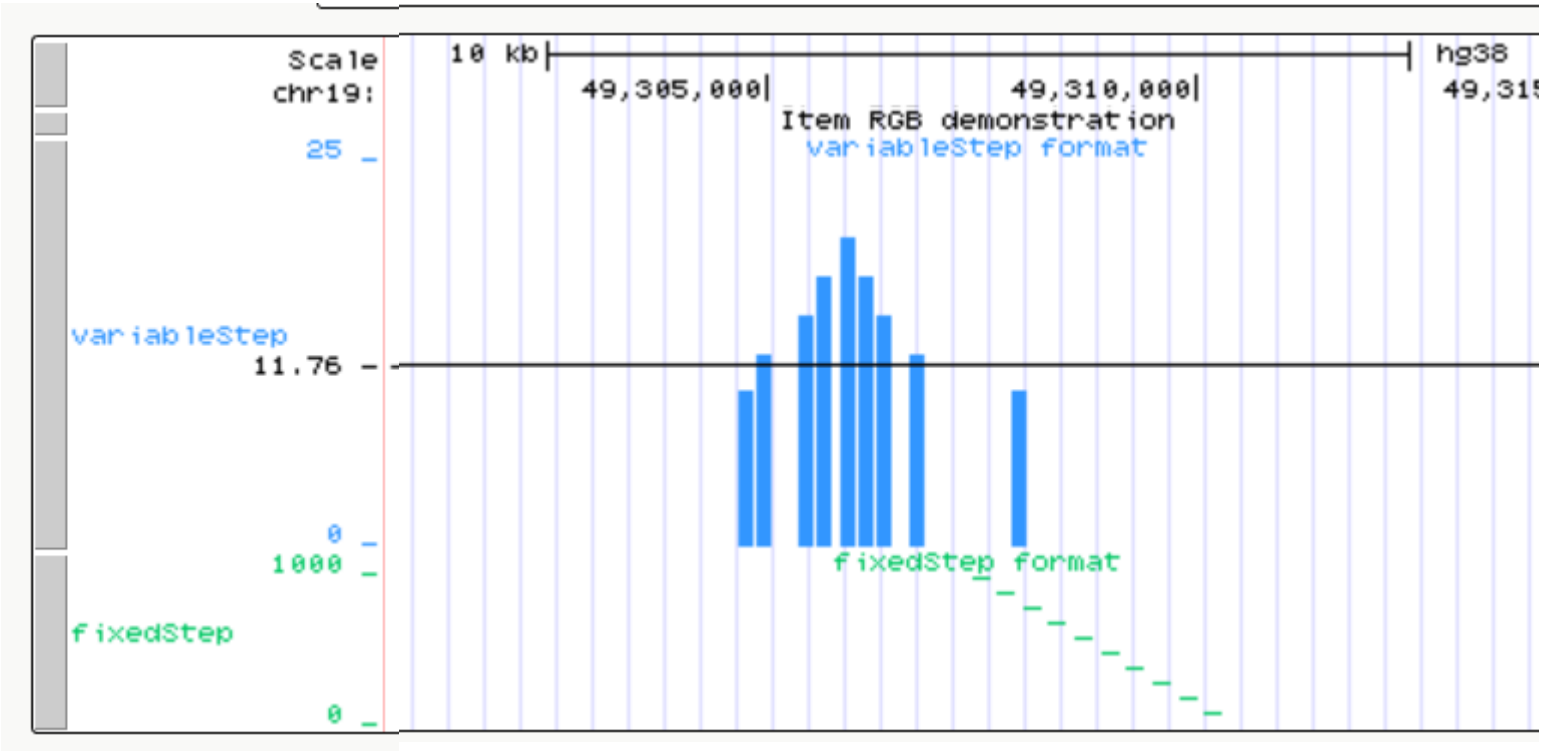
```
variableStep chrom=chr2 span=5
300701 12.5
```

```
fixedStep  chrom=chrN
start=position  step=stepInterval
[span=windowSize]
  dataValue1
  dataValue2
  ... etc ...
```

```
fixedStep chrom=chr3 start=400601 step=100
11
22
33
```

```
fixedStep chrom=chr3 start=400601 step=100 span=5
11
22
33
```

```
browser position chr19:49304200-49310700
browser hide all
#       150 base wide bar graph at arbitrarily spaced positions,
#       threshold line drawn at y=11.76
#       autoScale off viewing range set to [0:25]
#       priority = 10 positions this as the first graph
#       Note, one-relative coordinate system in use for this format
track type=wiggle_0 name="variableStep" description="variableStep format" visibility=full autoScale=off viewLimits=0.0:25.0 color=50,150,255 yLineMark=11.76 yLineOnOff=on priority=10
variableStep chrom=chr19 span=150
49304701 10.0
49304901 12.5
49305401 15.0
49305601 17.5
49305901 20.0
49306081 17.5
49306301 15.0
49306691 12.5
49307871 10.0
#       200 base wide points graph at every 300 bases, 50 pixel high graph
#       autoScale off and viewing range set to [0:1000]
#       priority = 20 positions this as the second graph
#       Note, one-relative coordinate system in use for this format
track type=wiggle_0 name="fixedStep" description="fixedStep format" visibility=full autoScale=off viewLimits=0:1000 color=0,200,100 maxHeightPixels=100:50:20 graphType=points priority=20
fixedStep chrom=chr19 start=49307401 step=300 span=200
1000
 900
 800
 700
 600
 500
 400
 300
 200
 100
```



Wig pros and cons

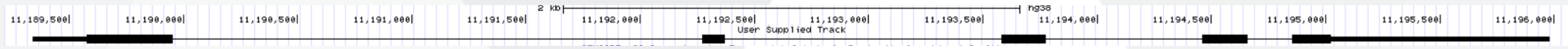
- Compact
- Wide variety of values

- Difficult for human readability
- Difficult to add/subtract lines

General transfer format (.gtf)

- Also commonly known as .gff (general feature format)
- 1. seqname (chr)
- 2. source (program generating seq)
- 3. feature (what is it? “CDS”, “exon”, “enhancer” etc)
- 4. start (position)
- 5. end (inclusive)
- 6. score (can be any float. Ideal: between 1-1000 for UCSC browser)
- 7. strand (“+”, “-”, “.”)
- 8. frame (for exons, frame is between 0-2 representing open reading frame, else “.”)
- 9. group; list of attributes (gene_id, transcript_id etc)

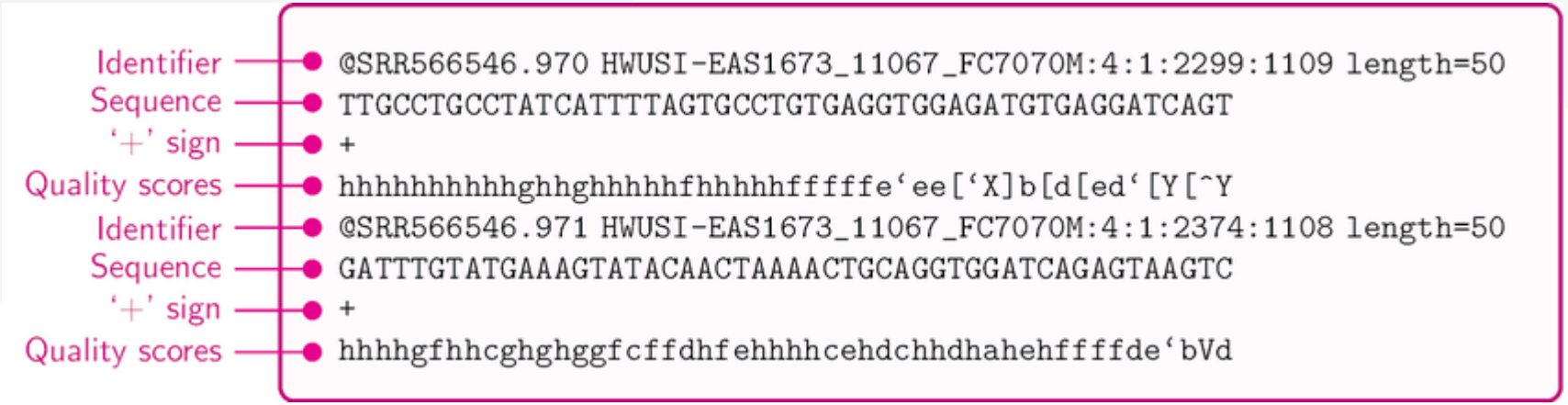
```
chr1    hg38_wgEncodeGencodeBasicV26    start_codon     11189580        11189582        0.000000        +       .       gene_id "ENST00000376819.3"; transcript_id "ENST00000376819.3";
chr1    hg38_wgEncodeGencodeBasicV26    CDS     11189580        11189955        0.000000        +       0       gene_id "ENST00000376819.3"; transcript_id "ENST00000376819.3";
chr1    hg38_wgEncodeGencodeBasicV26    exon    11189341        11189955        0.000000        +       .       gene_id "ENST00000376819.3"; transcript_id "ENST00000376819.3";
chr1    hg38_wgEncodeGencodeBasicV26    CDS     11192270        11192370        0.000000        +       2       gene_id "ENST00000376819.3"; transcript_id "ENST00000376819.3";
chr1    hg38_wgEncodeGencodeBasicV26    exon    11192270        11192370        0.000000        +       .       gene_id "ENST00000376819.3"; transcript_id "ENST00000376819.3";
chr1    hg38_wgEncodeGencodeBasicV26    CDS     11193580        11193774        0.000000        +       0       gene_id "ENST00000376819.3"; transcript_id "ENST00000376819.3";
chr1    hg38_wgEncodeGencodeBasicV26    exon    11193580        11193774        0.000000        +       .       gene_id "ENST00000376819.3"; transcript_id "ENST00000376819.3";
chr1    hg38_wgEncodeGencodeBasicV26    CDS     11194461        11194659        0.000000        +       0       gene_id "ENST00000376819.3"; transcript_id "ENST00000376819.3";
chr1    hg38_wgEncodeGencodeBasicV26    exon    11194461        11194659        0.000000        +       .       gene_id "ENST00000376819.3"; transcript_id "ENST00000376819.3";
chr1    hg38_wgEncodeGencodeBasicV26    CDS     11194854        11195020        0.000000        +       2       gene_id "ENST00000376819.3"; transcript_id "ENST00000376819.3";
chr1    hg38_wgEncodeGencodeBasicV26    stop_codon      11195021        11195023        0.000000        +       .       gene_id "ENST00000376819.3"; transcript_id "ENST00000376819.3";
chr1    hg38_wgEncodeGencodeBasicV26    exon    11194854        11195981        0.000000        +       .       gene_id "ENST00000376819.3"; transcript_id "ENST00000376819.3";
```




```
chr1    hg38_wgEncodeGencodeBasicV26    start_codon     11189580        11189582        0.000000        +       .       gene_id "ENST00000376819.3"; transcript_id "ENST00000376819.3";
chr1    hg38_wgEncodeGencodeBasicV26    CDS     11189580        11189955        0.000000        +       0       gene_id "ENST00000376819.3"; transcript_id "ENST00000376819.3";
chr1    hg38_wgEncodeGencodeBasicV26    exon    11189341        11189955        0.000000        +       .       gene_id "ENST00000376819.3"; transcript_id "ENST00000376819.3";
chr1    hg38_wgEncodeGencodeBasicV26    CDS     11192270        11192370        0.000000        +       2       gene_id "ENST00000376819.3"; transcript_id "ENST00000376819.3";
chr1    hg38_wgEncodeGencodeBasicV26    exon    11192270        11192370        0.000000        +       .       gene_id "ENST00000376819.3"; transcript_id "ENST00000376819.3";
chr1    hg38_wgEncodeGencodeBasicV26    CDS     11193580        11193774        0.000000        +       0       gene_id "ENST00000376819.3"; transcript_id "ENST00000376819.3";
chr1    hg38_wgEncodeGencodeBasicV26    exon    11193580        11193774        0.000000        +       .       gene_id "ENST00000376819.3"; transcript_id "ENST00000376819.3";
chr1    hg38_wgEncodeGencodeBasicV26    CDS     11194461        11194659        0.000000        +       0       gene_id "ENST00000376819.3"; transcript_id "ENST00000376819.3";
chr1    hg38_wgEncodeGencodeBasicV26    exon    11194461        11194659        0.000000        +       .       gene_id "ENST00000376819.3"; transcript_id "ENST00000376819.3";
chr1    hg38_wgEncodeGencodeBasicV26    CDS     11194854        11195020        0.000000        +       2       gene_id "ENST00000376819.3"; transcript_id "ENST00000376819.3";
chr1    hg38_wgEncodeGencodeBasicV26    stop_codon      11195021        11195023        0.000000        +       .       gene_id "ENST00000376819.3"; transcript_id "ENST00000376819.3";
chr1    hg38_wgEncodeGencodeBasicV26    exon    11194854        11195981        0.000000        +       .       gene_id "ENST00000376819.3"; transcript_id "ENST00000376819.3";
```

Fasta & Fastq


Header
Sequence
Header
Sequence
Header
Sequence
>VIT_201s0011g03530.1
AATTAAGCATAAATACTCACTCTTACCCCCTTATTTTCTTATCTCTCATCACTTTTGGTGCGAAG
GACCATGAGAACAAGCTGCAATGGGTGTAGGGTTCTTCGCAAGGCATGCAGCCAAGACTGCATCA
>VIT_201s0011g03540.1
CAGGTAGCGTGAAGTTAAACCCTAGCGCTTTAGACAAACAGCTGTAGTCACCGCCCACAAACACC
AGCCTCTGAGACACCACCTCAAACCTTTCCACTTAAATACACATCCCTCACACCCTTTTCAATTC
>VIT_201s0011g03550.1
CATGCAAAGCTGAACGCGATGCTGTGATTGGTGGTAAGTGGTAGTTGAGTAAATTTGACAGTGAA
GCCGAAATGGTAAAAGACTAAGGCTAGAAGTAGAATACCACTGTTCTTCTCATCACGTGGGCCCA


Identifier
Sequence
'+' sign
Quality scores
Identifier
Sequence
'+' sign
Quality scores
@SRR566546.970 HWUSI-EAS1673_11067_FC7070M:4:1:2299:1109 length=50
TTGCCTGCCTATCATTTTAGTGCCTGTGAGGTGGAGATGTGAGGATCAGT
+
hhhhhhhhhhghhghhhhhfhhhhhfffffe'ee['X]b[d[ed'[Y[^Y
@SRR566546.971 HWUSI-EAS1673_11067_FC7070M:4:1:2374:1108 length=50
GATTTGTATGAAAGTATACAACTAAAACTGCAGGTGGATCAGAGTAAGTC
+
hhhhgfhhcghghggfcffdhfehhhhcehdchhdhahehffffde'bVd

Fastq Quality Scores

ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 $	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

ASCII_BASE=64 Old Illumina

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	64 @	11	0.07943	75 K	22	0.00631	86 V	33	0.00050	97 a
1	0.79433	65 A	12	0.06310	76 L	23	0.00501	87 W	34	0.00040	98 b
2	0.63096	66 B	13	0.05012	77 M	24	0.00398	88 X	35	0.00032	99 c
3	0.50119	67 C	14	0.03981	78 N	25	0.00316	89 Y	36	0.00025	100 d
4	0.39811	68 D	15	0.03162	79 O	26	0.00251	90 Z	37	0.00020	101 e
5	0.31623	69 E	16	0.02512	80 P	27	0.00200	91 [	38	0.00016	102 f
6	0.25119	70 F	17	0.01995	81 Q	28	0.00158	92 \	39	0.00013	103 g
7	0.19953	71 G	18	0.01585	82 R	29	0.00126	93]	40	0.00010	104 h
8	0.15849	72 H	19	0.01259	83 S	30	0.00100	94 ^	41	0.00008	105 i
9	0.12589	73 I	20	0.01000	84 T	31	0.00079	95 _	42	0.00006	106 j
10	0.10000	74 J	21	0.00794	85 U	32	0.00063	96 `			

Identifier	@SRR566546.970 HWUSI-EAS1673_11067_FC7070M:4:1:2299:1109 length=50
Sequence	TTGCCTGCCTATCATTTTAGTGCCTGTGAGGTGGAGATGTGAGGATCAGT
'+' sign	+
Quality scores	hhhhhhhhhhghhghhhhhfhhhhhfffffe‘ee[‘X]b[d[ed‘[Y[^Y
Identifier	@SRR566546.971 HWUSI-EAS1673_11067_FC7070M:4:1:2374:1108 length=50
Sequence	GATTTGTATGAAAGTATACAACTAAAACTGCAGGTGGATCAGAGTAAGTC
'+' sign	+
Quality scores	hhhhgfhhcghghggfcffdhfehhhhcehdchhdhahehffffde‘bVd

Multiple Alignment File

```
CLUSTAL W(1.83) multiple sequence alignment


IXI_234         TSPASIRPPAGPSSRPAMVSSRRTRPSPPGPRRPTGRPCCSAAPRRPQAT
IXI_235         TSPASIRPPAGPSSR---------RPSPPGPRRPTGRPCCSAAPRRPQAT
IXI_236         TSPASIRPPAGPSSRPAMVSSR--RPSPPPPRRPPGRPCCSAAPPRPQAT
IXI_237         TSPASLRPPAGPSSRPAMVSSRR-RPSPPGPRRPT----CSAAPRRPQAT


IXI_234         GGWKTCSGTCTTSTSTRHRGRSGWSARTTTAACLRASRKSMRAACSRSAG
IXI_235         GGWKTCSGTCTTSTSTRHRGRSGW----------RASRKSMRAACSRSAG
IXI_236         GGWKTCSGTCTTSTSTRHRGRSGWSARTTTAACLRASRKSMRAACSR--G
IXI_237         GGYKTCSGTCTTSTSTRHRGRSGYSARTTTAACLRASRKSMRAACSR--G


IXI_234         SRPNRFAPTLMSSCITSTTGPPAWAGDRSHE
IXI_235         SRPNRFAPTLMSSCITSTTGPPAWAGDRSHE
IXI_236         SRPPRFAPPLMSSCITSTTGPPPPAGDRSHE
IXI_237         SRPNRFAPTLMSSCLTSTTGPPAYAGDRSHE
```

Sequence Alignment Map (.sam)

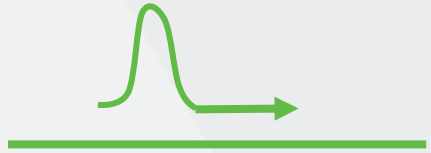
- https://www.samformat.info/

SAM stands for Sequence Alignment/Map format. It is a TAB-delimited text format consisting of a header section, which is optional, and an alignment section. If present, the header must be prior to the alignments. Header lines start with '@', while alignment lines do not. Each alignment line has 11 mandatory fields for essential alignment information such as mapping position, and variable number of optional fields for flexible or aligner specific information.

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

Col	Field	Type	Brief description
1	QNAME	String	Query template NAME
2	FLAG	Int	bitwise FLAG
3	RNAME	String	References sequence NAME
4	POS	Int	1- based leftmost mapping POSition
5	MAPQ	Int	MAPping Quality
6	CIGAR	String	CIGAR String
7	RNEXT	String	Ref. name of the mate/next read
8	PNEXT	Int	Position of the mate/next read
9	TLEN	Int	observed Template LENgth
10	SEQ	String	segment SEQuence
11	QUAL	String	ASCII of Phred-scaled base QUALity+33

Sequence Alignment Map (.sam)



```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

Header section

Alignment section

Optional fields in the format of TAG:TYPE:VALUE

QUAL: read quality; * meaning such information is not available

SEQ: read sequence

TLEN: the number of bases covered by the reads from the same fragment. Plus/minus means the current read is the leftmost/rightmost read. E.g. compare first and last lines.

PNEXT: Position of the primary alignment of the NEXT read in the template. Set as 0 when the information is unavailable. It corresponds to POS column.

RNEXT: reference sequence name of the primary alignment of the NEXT read. For paired-end sequencing, NEXT read is the paired read, corresponding to the RNAME column.

CIGAR: summary of alignment, e.g. insertion, deletion

MAPQ: mapping quality

POS: 1-based position

RNAME: reference sequence name, e.g. chromosome/transcript id

FLAG: indicates alignment information about the read, e.g. paired, aligned, etc.

QNAME: query template name, aka. read ID

.bam format
(Binary SAM)

Encoding Genomic Information for Bioinformatics Use

- Location Based Formats (.bed)
- Count/Coverage Based Formats (.bedgraph .wig)
- Feature Based Formats (.gtf)
- Sequence Based Formats (.fasta .fastq)
- Multiple Alignment Files
- Alignment Based Formats (.sam)