Briefly: Bioinformatics File Formats

M Britton | March 2019

Overview

- **ASCII Text**
  - **Sequence**
    - **Fasta, Fastq**
  - **~Annotation**
    - **TSV, CSV, BED, GFF, GTF, VCF, SAM**
- **Binary (Data, Compressed, Executable)**
  - **Data**
    - **HDF5**
    - **BAM / CRAM**
    - **2bit**
  - **Compressed**
    - **gzip, bzip2, bgzip**
  - **Executable**

TEXT

Fasta

```
>m54050R1_180210_051102/4194473/0_1421
CCCGGCTGCCCGCCCCGCTCGAAGCGATGACTTGCCGGCGGCCCGACGCGATTAGCTGCCGCGCATGCGATGCGGCCGCGGGCGGCGTGCTGACCTGGCTGGCGGTG
TTGAGCTGCTATACATCCGGCAACAACGCTGCCCAACGACTGACCTGACCGGCCGCCTCGATCCTGGCGGCCGCCGGCCTGGCCTGCGCTTTTCCTTCTTCTCTTTC
CTTC
>m54050R1_180210_051102/4194473/1497_4602
GGCGCGCTGATCGGCAAAACGGCTGGGGCGGCCGGAACACCTTTCAACCGTCGCCAACCGCGATCGCCGCGCGCACCCCGCCTTCCGCGCCGCTGTGGCGTTCCTCG
CCCGTCCTACTCTACTGGCATCCGTCTCATTTCTCCCGCTCTTCCCTCCACCCTTTCCCTGCTCACCGCTTCCGTCTTTTTGTCAACCTCTCCTCTGGGCCGACGAC
GTCGCCGCCTACTGCGACAAAAACCGAGGTCGACAAGGCCCGCCGTTACGACCGTCACCCCGAATTCCATCCGGCTGCTCCGCGGT
>m54050R1_180210_051102/4194551/0_17688
ACCGGACGTACCGCGGGCGGGGGCCTCCCCCCCGGGTGGCTCGGGTGCAGCGCAAATCCTTTCTTTGCTGACCCACCTGCGCAGCGAGTGTGAATCTGTGCGGATCG
AGAAAACAAGAAACCCGGCGGGCCCTGCCTGACGCGCGCCCGTCCCGCCGCGCCCCTTCCGCTTGGCGACGTCGAGTTTTTGACGGGAGGTTTGTCGCTTCGACAGA
CGGGTCCGCCAGCACCCCCTCGTCGCAGTCCCGTTAACTCAGGAAGAACTCCCAGTTGGCCCGGGCATCTGCCAACGCCTCCGGGG
>m54050R1_180210_051102/4194551/17752_17812
AAACATATTATTTTTTATTACTCAAATAATTATTATATTCACCTAATTTTCTTTATTATT
>m54050R1_180210_051102/4194552/0_89
CAGATCGGGGCCCAGCATGGCCACCCGTCCTGCACGTCTACGCGCACTTCGCCGGTGGGGATCGGCAGCGGGAACGGCTCGCGGGCTGG
>m54050R1_180210_051102/4194552/162_490
GCCGCACCCGAGCCGTTCCCGCTGCCGATCCACACCGTCGACGTGCGCGTCGACGTGCAGCCGGCGTCCATGCTTGCCCCGATCTTGGGCTAACAAGCCGCTGCTGA
CACCGACGGACGCCACCGCCCGCGACCAGCTGGCCCGGGCCTCGGTGATGGCGCTGTCCTACCGTCGCGCATTCCCGCGCTCGGCATCTATCAGCCTCGGTGCCGCA
GCGTCATCGACGATGGCGAAACCGTCACTGCACGTTTTCATGACGCGGGGCAGGCAGCGAACCGGGCACATCGGGCATCTACGCCT
```

Fasta

```
>m54050R1_180210_051102/4194473/0_1421
CCCGGCTGCCCGCCCCGCTCGAAGCGATGACTTGCCGGCGGCCCGACGCGATTAGCTGCCGCGCATGCGATGCGGCCGCGGGCGGCGTGCTGACCTGGCTGGCGGTG
TTGAGCTGCTATACATCCGGCAACAACGCTGCCCAACGACTGACCTGACCGGCCGCCTCGATCCTGGCGGCCGCCGGCCTGGCCTGCGCTTTTCCTTCTTCTCTTTC
CTTC
```

Header symbol “>” also redirects stuff into files, so be careful using > in bash commands!

Header text (sequence ID) has formats particular to different organizations and different software, but really has no consistent rules that you can rely on.

Sequence can contain: newline characters (“\n”), ACGT, N, acgt, n, x, . or - (gaps), IUPAC ambiguity codes BDHV etc., alternates like [A/T], amino acid single letter codes (protein fasta; sometimes file name is ‘sequence.fna’ for fasta nucleic acid, or ‘sequence.faa’ for fasta amino acid)

Fastq … “fasta + qualities”

```
@SN638:981:HK7HWBCXX:2:1101:14799:2762 1:N:0:TTAGGC
TGGCGCAACTGCCGATCACCATCGACACCAACGGGTATCTGGTCGCCAAC
+
GGGGGIIIIIIIIIIIIIIIGIIIIIIIIIIIIIIIIIIIIIIIIIIIG
@SN638:981:HK7HWBCXX:2:1101:14784:2782 1:N:0:TTAGGC
CATCATCGAGGACAGCGCCGGTGACCTGGCGGCCCGCATCGGTGCCCCCC
+
GGGGGIIIIIIIIIIIIIIIIIIIIIIIIIGIIIIIIIIIIIIIIIGIIII
@SN638:981:HK7HWBCXX:2:1101:14983:2799 1:N:0:TTAGGC
CGGCGCCGTTGCTGCTGCTGCCGGTGCTGCTTTCGGCGCTGATCGTGCGG
+
GGGGGIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIGIII
@SN638:981:HK7HWBCXX:2:1101:14763:2901 1:N:0:TTAGGC
CCTGACGACGGCACGAAGGACCTCTTCGTCCACTACTCCGAGATCCAGGG
+
GAGGGIGIGGGGGGGGIA.<GGGIGGAGGGGIIGIIGGIIIG<GA.<<GA
```

@Header1
Sequence
+Header2
Qualities

Blocks of four lines for each sequence (sequences shouldn’t occupy more than one line, as they can in fasta). Second header line (starting with “+”) is mandatory, sometimes contains the same header as the first line (that starts with “@”). Why??

The n^{th} quality character applies to the nth nucleotide, and is a number that is *encoded in a single character from the ASCII table*.

Fastq … “fasta + qualities”

```
@SN638:981:HK7HWBCXX:2:1101:14799:2762 1:N:0:TTAGGC
TGGCGCAACTGCCGATCACCATCGACACCAACGGGTATCTGGTCGCCAAC
+
GGGGGIIIIIIIIIIIIIIIGIIIIIIIIIIIIIIIIIIIIIIIIIIG
@SN638:981:HK7HWBCXX:2:1101:14784:2782 1:N:0:TTAGGC
CATCATCGAGGACAGCGCCGGTGACCTGGCGGCCCGCATCGGTGCCCCCC
+
GGGGGIIIIIIIIIIIIIIIIIIIIIIIIIGIIIIIIIIIIIIIIGIIII
@SN638:981:HK7HWBCXX:2:1101:14983:2799 1:N:0:TTAGGC
CGGCGCCGTTGCTGCTGCTGCCGGTGCTGCTTTCGGCGCTGATCGTGCGG
+
GGGGGIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIGIII
@SN638:981:HK7HWBCXX:2:1101:14763:2901 1:N:0:TTAGGC
CCTGACGACGGCACGAAGGACCTCTTCGTCCACTACTCCGAGATCCAGGG
+
GAGGGIGIGGGGGGGGIA.<GGGIGGAGGGGIIGIIGGIIIG<GA.<<GA
```

The “I” for base 16 (“C”) means that that base has a quality of (I’s decimal value: 73) - 33 = 40 (sometimes referred to as “Q40”). Why 33? Because there are 32 non-printable “characters” at the beginning of the ASCII table! (type ‘man ascii’)

Q40 means that the probability of error (that C is actually the wrong basecall) is:

$p_e = 10^{(-40 \,/\, 10)} = 0.0001$, or 1 in 10,000

see also: *https://en.wikipedia.org/wiki/FASTQ_format*

CSV and TSV - comma/tab-separated values

	B01	B02	B03	B04
PDCD1	0	0	0	0
GAL3ST2	0	0	0	0
D2HGDH	55	71	89	101
ING5	1	1	1	1
DTYMK	2	5	7	12
ATG4B	0	0	0	0
THAP4	136	158	85	161
BOK	0	0	0	0
STK25	145	175	195	141

For example, abundances of mRNAs from genes (count data).

(First tab character - “\t” - in column names sometimes omitted for ease of reading by R scripts).

BED - tsv with defined column meanings

```
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  -
```

column	meaning
1	chromosome name
2	feature start coordinate (0-based...?)
3	feature stop coordinate (0-based...?)
4	feature name
5	score (1-1000)
6	strand (‘+’ or ‘-’ or ‘.’ for unknown or not applicable)
…	…

Number of columns used shouldn’t vary within a particular file.

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

GFF / GTF - tsv with defined column meanings

```
chr22  TeleGene enhancer  10000000  10001000  500 +  .  touch1
chr22  TeleGene promoter  10010000  10010100  900 +  .  touch1
chr22  TeleGene promoter  10020000  10025000  800 -  .  touch2
```

column	meaning
1	chromosome / scaffold name
2	source (e.g. software that generated this feature / gene call)
3	feature name (e.g. “exon1”, “enhance”r, “3’-UTR”)
4	feature start coordinate (1-based)
5	feature stop coordinate (1-based)
6	score (1-1000)
7	strand (‘+’ or ‘-’ or ‘.’ for unknown or not applicable)
8	reading frame (0, 1, 2, or “.” if N/A)
9	group (allows grouping features together)

GTF is newer, and shares the first eight (8) columns. Column 9 has additional restrictions in format (gene_id, transcript_id, etc.)

see also: *https://genome.ucsc.edu/FAQ/FAQformat.html#format3*

VCF - tsv with defined column meanings

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL    FILTER  INFO                              FORMAT  sample07 ...
20     14370   rs6054257      G       A       29      PASS    NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ    0|0:48:1:51,51
20     17330   .              T       A       3       q10     NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ    0|0:49:3:58,50
20     111069  rs6040355      A       G,T     67      PASS    NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ    1|2:21:6:23,27
20     123027  .              T       .       47      PASS    NS=3;DP=13;AA=T               GT:GQ:DP:HQ    0|0:54:7:56,60
20     123457  microsat1      GTC     G,GTCT  50      PASS    NS=3;DP=9;AA=G                GT:GQ:DP       0/1:35:4
```

SAM - tsv with defined column meanings

http://www.htslib.org/

See also samtools man page: *http://samtools.sourceforge.net/*

SAM spec grew out of 1000 Genomes Project (see Li et al. 2009 *Bioinformatics* 25:2078)

SAM is plain text; BAM is binary, compressed version of SAM; CRAM is further compressed but not widely used / recognizable by many tools.

SAM - tsv with defined column meanings

```
[...]
@SQ     SN:ctg103993      LN:217
@SQ     SN:ctg103994      LN:222
@SQ     SN:ctg103995      LN:205
@SQ     SN:ctg103996      LN:210
@PG     ID:bwa  PN:bwa  VN:0.7.13-r1126 CL:bwa mem -t 4 -M ../../01_Reference/Transcriptome-Contigs-Build2.fna
../../02-Cleaned/3E/3E_SE.fastq
@PG     ID:bwa-7BC92A6F PN:bwa  VN:0.7.13-r1126 CL:bwa mem -t 4 -M ../../01_Reference/Transcriptome-Contigs-Build2.fna
../../02-Cleaned/3E/3E_R1.fastq ../../02-Cleaned/3E/3E_R2.fastq
K00188:264:HG3WJBBXX:1:1116:14692:35180#0       121     ctg2      128     58      101M    =       128     0
AAGTCTCGACCAAGTGGTTCAGATGGTGACACAGATGTTAGCCCCATCCACCATTCAGTTGCCGTTTTGATAGCTGGAAATCCTGTAAACACAATGCTGAG
FJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFFFAA   NM:i:10
K00188:264:HG3WJBBXX:1:1116:14692:35180#0       181     ctg2      128     0       *            =       128     0
TTTAGTTTTAATTTTTGACTTTGAATAGCGGGAGTCCAGATCGTGTGAACACAGCAGACTGAGCACTCCATTGACAGCCTTCTTCTGTACTTTAGCTATCC
FJFJJFAAJF7F7JJJJAFFFAF<7<AFFJJJFJJJJJJJJJJJJJJJJJJJJJJJJJJFJJJJJJFAJJJJJJJFFFJJJJJJJJJJFFJJJJJJFFFAA   AS:i:0  XS:i:0
K00188:264:HG3WJBBXX:1:1202:11028:9596#0        121     ctg5      45      60      101M    =       45      0
TTCTTTTTTCTACAGTTCATTGTCTGTATAAAGTATGCATCAGGAACAATCTGACTAGGAAGGTAAATAATGTAAAACAGATGATTATTGTATGAAAGTTG
JJJJJJJJJJJJJJJJJJJJJJJJFJJJJJJJJJJJJJJJJJJJJJJJJJJJFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFFFAA   NM:i:8
K00188:264:HG3WJBBXX:1:1202:11028:9596#0        181     ctg5      45      0       *       =       45      0
TCAGCTGTATTAGTAATTTAGTAGAAAAGGTCTTGAGAGAATTATGTTTTTTAAAAATCCACATCACTTCAAACAAAAAGCCCCATTAGAATGGAGGGCCA
FJFJJJJJFJJJJJJFFJJJJFAJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFF-JFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFFFAA   AS:i:0
[...]
```

SAM - tsv with defined column meanings

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

BINARY

HDF5

- **“Hierarchical Data Format” used across many industries**
- **PacBio read data *no longer* comes in bas.h5 / bax.h5 files (instead, you get BAM files) … so let’s forget about HDF5!**

BAM / CRAM - compressed SAM

- *** Don’t dump binary formats to your terminal / shell …**
- **Indexing both BAM and CRAM allow *rapid* random read access to any coordinate range, without uncompressing whole file first**
- **CRAM restricts sequence alphabet, so compression ratio can be greater**
- **CRAM does *lossy* compression of base qualities, also helps compression ratio**

2bit

- **Old format used for sequence in UCSC Genome Browser**
- **Can only store 4 bases per position:**
  - **00 = A**
  - **01 = C**
  - **10 = G**
  - **11 = T**
  - **… N? Lower case acgt for soft masking? Nope ...**

Questions … comments … confusion?