Interoperability: File formats

Amelia Harrison

Presented: April 21, 2022

About file formats

What is a file format?

- A standardized layout/structure of information storage

Why is file formatting important?

- File formats are created with specific goals in mind

Do file formats change?

- Yes! File formats have version numbers and evolve over time

File formats and FAIR data

- Interoperability
  - Most software relies on correctly formatted data
  - Many researchers are not able to reformat data themselves
- Reusability
  - Datasets are easier to use the more similar they are
  - Using established, sustainable file formats ensures datasets remain similar

Common file formats in omics research

- Fasta
- Fastq
- BED
- SAM/BAM
- VCF
- GFF
- JSON
- XML

Examples: Fasta unaligned

- Used to store nucleotide and amino acid sequences with headers

```
>tr|A0A679IRD3|A0A679IRD3_9ENTE Multifunctional fusion protein OS=Enterococcus
saigonensis OX=1805431 GN=rphA PE=3 SV=1
MRHDGRQVQQIRPVIIKTNVFKHPEGSVVISFGDTQVVCSATIEERVPHFLRDTGKGWVN
AEYSMLPRATQTRNRRESAKGKLSGRTMEIQRLIARSLRAVVDLEKLGERSIIVDCDVLQ
ADGGTRTASITGAFVALRLAINKLLQANVLTEDPIKEHLAAISVGILSDGTCVTDLDYVE
DFEASVDMNVVMTESGQFVELQGTGEESTFNGEELNEMLVYAKHAINDLIAFQKEALLGN
VALKEVIINENPKTIVIATKNPGKAKEFDALFSAAGYQVKTLLDYPEIPEVEETGHTFEE
NARLKAETIAHLLNRPVLADDSGLSVDALNGMPGVYSARFAGEMKSDAANNAKLLHELTH
VPDEDRTAHFHCTLVFAAPEKNSLSVSADWPGRIGRIPRGDDGFGYDPLFIPQGMEKTAA
ELSRTEKNAISHRGQAMKKLQKEWRTWLEA

> header
NUCLEOTIDEORAMINOACIDSEQUENCE
```

Examples: Fasta aligned

```
>protein1
MRHDGRQVQQIRPVIIKTNVFKHPEGSVVISFGDTQVVCSATIEERVPHFLRDTGKGWVN
AEYSMLPRATQTRNRRESAKGKLSGRTMEIQRLIARSLRAVVDLEKLGERSIIVDCDVL*
>protein2
MRHDGRQVQ--------TNVFKHPEGSVVISFGDTQVVCSATIEERVPHFLRDTGKGWVN
AEYSMLPRATQTRNRRESAKGKLSGRTMEIQRL-ARSLRAVVDLEKLGERSIIVDCDVL*
>protein3
MRHDGRQVQQIRPVIIKTNVFKHPEGSVVISFGDTGIICSATIEERVPHFLRDTGKGWVN
AEYSMLPRATQTRNRRESAKGKLSGRTMEIQRLIARSLRAVVDLEK--ERSIIVDCDVL*
```

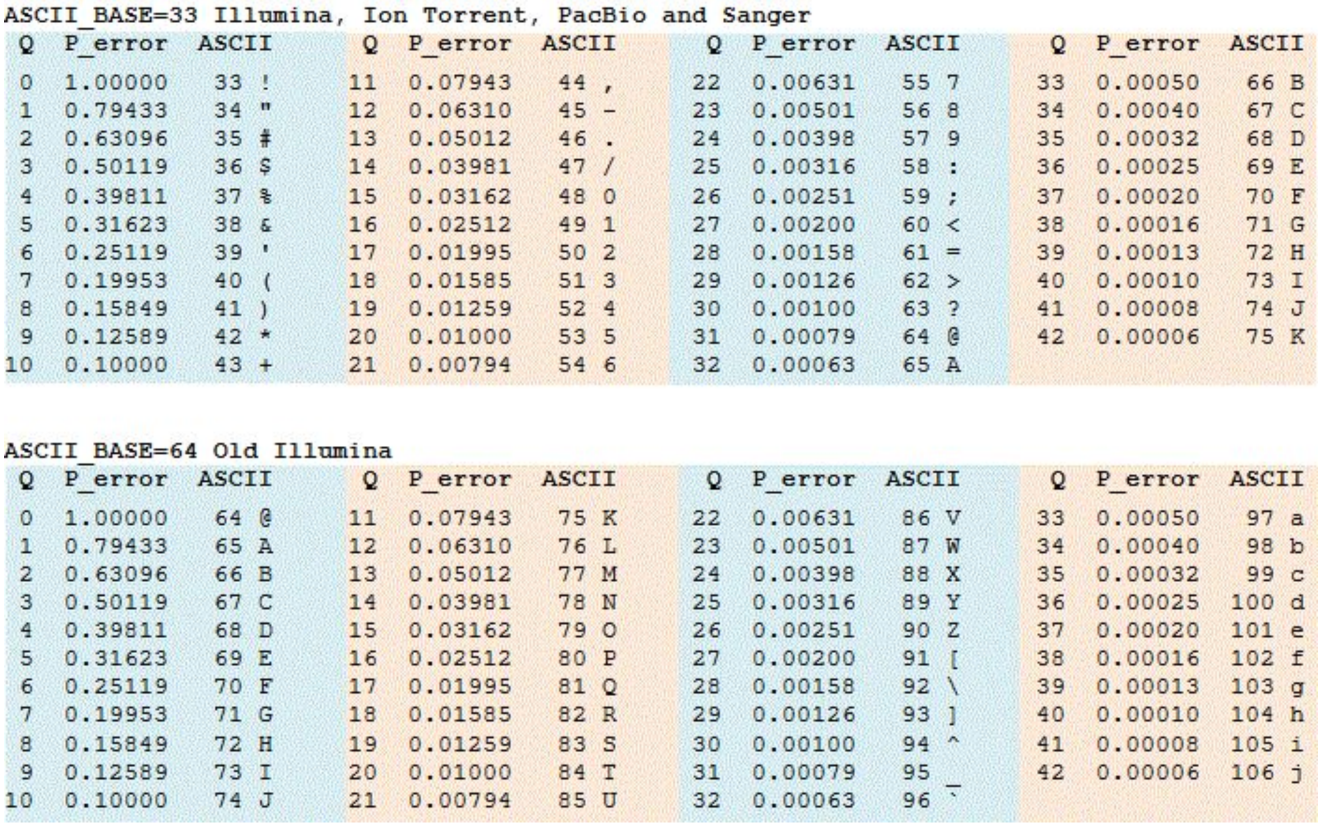
Examples: Fastq

- Used for DNA sequences and their quality scores

```
@K00188:208:HFLNGBBXX:3:1101:1428:1508 2:N:0:CTTGTA
ATAATAGGATCCCTTTTCCTGGAGCTGCCTTTAGGTAATGTAGTATCTNATNGACTGNCNCCANANGGCTAAAGT
+
AAAFFJJJJJJJJJJJJJJJJJFJJFJJJJJFJJJJJJJJJJJJJJJJ#FJ#JJJJF#F#FJJ#F#JJJFJJJJJ
```

```
@information:about:sequencing
DNASEQUENCE
+ (optional comment)
QUALITYSCORE
```

Quality (Phred) 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 `			

https://www.drive5.com/usearch/manual/quality_score.html

Examples: BED (Browser Extensible Data)

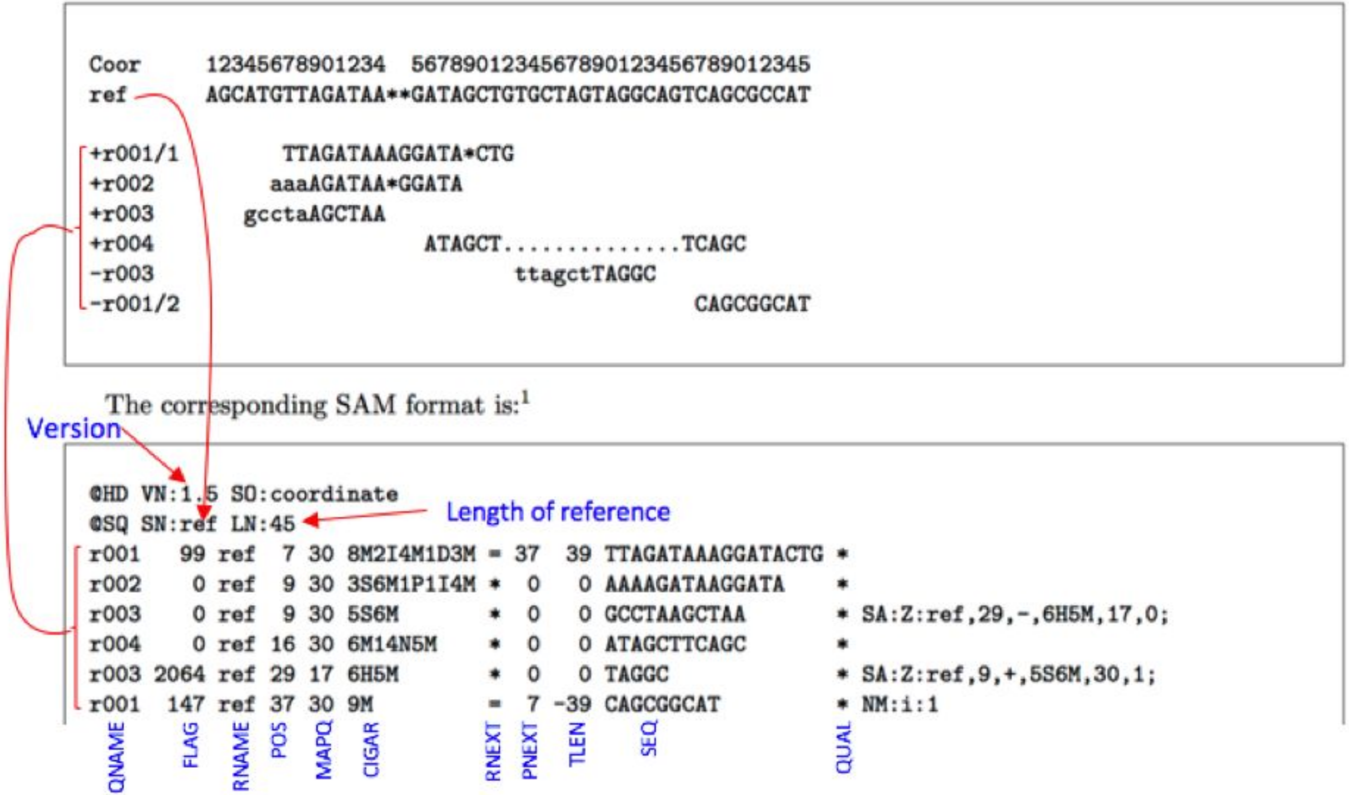
- Tab-separated file used for alignment to a reference
- First three columns are required
- Example: http://useast.ensembl.org/info/website/upload/bed.html

```
Chrom chromStart chromEnd   name score strand thickStart thickEnd otherOptionalCols
chr7  127471196  127472363  Pos1  0   +  127471196  127472363
chr7  127472363  127473530  Pos2  0   +  127472363  127473530
chr7  127473530  127474697  Pos3  0   +  127473530  127474697
chr7  127474697  127475864  Pos4  0   +  127474697  127475864
chr7  127475864  127477031  Neg1  0   -  127475864  127477031
chr7  127477031  127478198  Neg2  0   -  127477031  127478198
chr7  127478198  127479365  Neg3  0   -  127478198  127479365
chr7  127479365  127480532  Pos5  0   +  127479365  127480532
chr7  127480532  127481699  Neg4  0   -  127480532  127481699
```

Examples: BAM/SAM

- SAM: Sequence Alignment Map
- BAM: Binary SAM
- Tab-separated table with a header
- Describes the full alignment of a sequence against a reference
- This tool can help you to decode flags

Annotated SAM example



```
Coor     12345678901234  5678901234567890123456789012345
ref      AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1        TTAGATAAAGGATA*CTG
+r002         aaaAGATAA*GGATA
+r003       gcctaAGCTAA
+r004                     ATAGCT..............TCAGC
-r003                            ttagctTAGGC
-r001/2                                        CAGCGGCAT
```

The corresponding SAM format is:[1]

```
@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
```

Examples: VCF (Variant Call Format)

- Tab-separated table with a header
- After headers lines, contains one line per variant found
- Required columns:
    - #CHROM
    - POS
    - ID
    - REF
    - ALT
    - QUAL
    - FILTER
    - INFO
- A very thorough breakdown of VCF: https://samtools.github.io/hts-specs/VCFv4.2.pdf

VCF

```
##fileformat=VCFv4.3
##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       NA00001         NA00002          NA00003
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  1|0:48:8:51,51   1/1:43:5:.,.
20     17330    .          T     A      3     q10     NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ  0|0:49:3:58,50  0|1:3:5:65,3     0/0:41:3
20     1110696  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  2|1:2:0:18,2     2/2:35:4
20     1230237  .          T     .      47    PASS    NS=3;DP=13;AA=T                   GT:GQ:DP:HQ  0|0:54:7:56,60  0|0:48:4:51,51   0/0:61:2
20     1234567  microsat1  GTC   G,GTCT 50    PASS    NS=3;DP=9;AA=G                    GT:GQ:DP     0/1:35:4        0/2:17:2         1/1:40:3
```

VCF headers

- Metadata is stored in VCF headers
    - Inform what types of data are available in the file
- ##fileformat
    - Version number for the vcf format
- ##FORMAT
    - Tells the program what information about genotype, read depth, etc. to include in the later columns and in what order
- ##FILTER
    - Filtering parameters applied to the SNPs (compare to snp_file.raw.vcf)
- ##INFO
    - Additional information about a SNP

Examples: GFF3 (Generic Feature Format)

- Tab-separated table with 9 columns
- Used to store annotations
  - Often combined with a sequence file (e.g., fasta)
- Official specifications

```
##gff-version 3
ctg123   .   exon   1300   1500   .   +   .   ID=exon00001
ctg123   .   exon   1050   1500   .   +   .   ID=exon00002
ctg123   .   exon   3000   3902   .   +   .   ID=exon00003
ctg123   .   exon   5000   5500   .   +   .   ID=exon00004
ctg123   .   exon   7000   9000   .   +   .   ID=exon00005
```

Examples: JSON (JavaScript Object Notation)

- Not specific to bioinformatics
- Used to store objects and can indicate structure

```
{"menu": {
  "id": "file",
  "value": "File",
  "popup": {
    "menuitem": [
      {"value": "New", "onclick": "CreateNewDoc()"},
      {"value": "Open", "onclick": "OpenDoc()"},
      {"value": "Close", "onclick": "CloseDoc()"}
    ]
  }
}}
```

Examples: XML (Extensible Markup Language)

- Not specific to bioinformatics
- Use case is similar to JSON, but XML is more complex

```
<menu id="file" value="File">
  <popup>
    <menuitem value="New" onclick="CreateNewDoc()" />
    <menuitem value="Open" onclick="OpenDoc()" />
    <menuitem value="Close" onclick="CloseDoc()" />
  </popup>
</menu>
```

Additional resources

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

https://bioinformatics.uconn.edu/resources-and-events/tutorials-2/file-formats-tutorial/

http://useast.ensembl.org/info/website/upload/bed.html

https://www.toptal.com/web/json-vs-xml-part-1#:~:text=JSON%20is%20a%20data%20interchange,of%20any%20XML%20sub%2Dlanguage.