BIOINFORMATICS SHEET #III-2

MAIN SEQUENCE FILE FORMATS USED DURING THE PRACTICAL

Sequence information may be stored under different formats. Depending on the software, the format of the input file may be different. Here is a short description of the main formats that will be used during the practical.

FASTA format: extension .fasta or .fa

Used for nucleic or protein sequences
FASTA files will contain at least two lines:
- The first line obligatorily begins with ">" : the header. It contains the name or the identifier of the sequence and comments (optional). Ends with the return command.
- One or several lines of sequences
A multi FASTA file contains several sequences all preceded by its header.

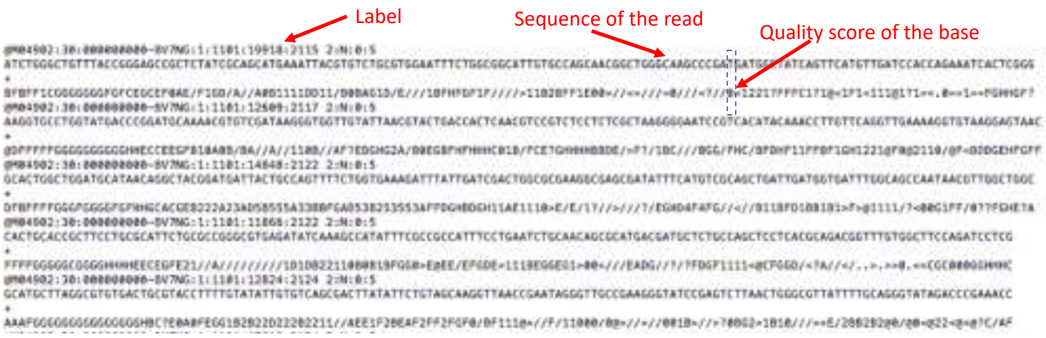
Exemple of a multi FASTA file with two sequences

```
>xyz some other comment
AACCGCTGCGCATCAGCGACCACGTCTGGCAGATCGGCACCGCCAGCATCAGCGCCCTG
CTGGTGAAAACCGACGCCGGCGCGGTGCTCATCGACGGCGGCATGCCCCAGGTGGCCGAC
CACCTGCTGGCCAATATGAAGAAGCTCGGCGTGCAGCCCCAGGACCTGCGCCTGATCCTC
CACAGCCACGCCCACATCGACCACGTCGGCCCGCTGGCGGCGATCAAGCGCGCCACCG
>uvw some other comments
AAGTTTCAGCAAGTTGAACAAGACGTTAAGGCAATTGAAGTTTCTCTTTCTGCTCGTATA
GGTGTTTCCGTTCTTGATACTCAAAATGGAGAATATTGGGATTACAATGGCAATCAGCGC
TTCCCGTTAACAAGTACTTTTAAAACAATAGCTTGCGCTAAATTACTATATGATGCTGAG
CAAGGAAAAGTTAATCCCAATAGTACAGTCGAGATTAAGAAAGCAGATCTTGTGACCTAT
```

FASTQ format: extension .fastq

The FASTQ format is the format of DNA sequences produced by sequencers. FASTQ files contain generally a very large number of individual sequences (several millions for bacterial genomes)
They contain both the information on the sequence of the reads and the information on the quality scores of each nucleotide of the read.
There are always four lines per read. The first line starts with "@" followed by the label.
The second line corresponds to the sequence of the read. The third line starts with "+" and a second copy of the label or only "+". The fourth line corresponds to the Q (quality) scores of each base of the read encoded in ASCII characters. A score of 20 means an error probability of 1/100, a score of 30, an error probability of 1/1000, etc. The way the Q score is encoded depends on the sequencing platform.

Ex: platform Illumina



```
@M04502:30:000000000-BV7NG:1:1101:19918:2115 2:N:0:5
+
@M04502:30:000000000-BV7NG:1:1101:12509:2117 2:N:0:5
+
@M04502:30:000000000-BV7NG:1:1101:14848:2122 2:N:0:5
+
@M04502:30:000000000-BV7NG:1:1101:11868:2122 2:N:0:5
+
@M04502:30:000000000-BV7NG:1:1101:12824:2124 2:N:0:5
+
```

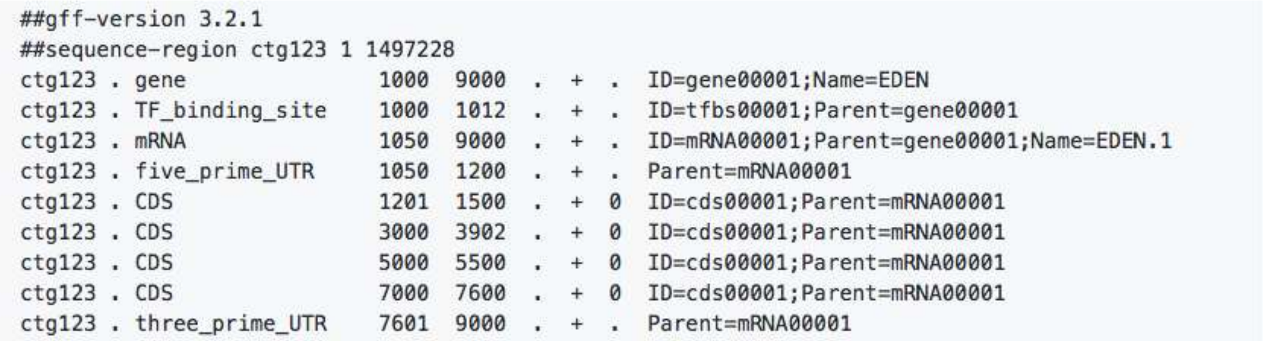
Genbank format: extension .gbk

File created in the GenBank format, a file format used for storing genome information; saves DNA sequences in a plain text format; also contains metadata such as the sample source, a description, and author information. It also often contains annotation of genomic features (genes, promoters, ribosome binding sites, transcription terminators, non-coding RNAs, etc. The GenBank format was developed by the U.S. National Center for Biotechnology Information (NCBI).

```
LOCUS       SCU49845     3240 bp    DNA             PLN       21-JUN-1999
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      Saccharomyces cerevisiae (baker's yeast)
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
            Saccharomycetales; Saccharomycetaceae; Saccharomyces.
REFERENCE   1  (bases 1 to 3240)
  AUTHORS   Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
  TITLE     Cloning and sequence of REV7, a gene whose function is required for
            DNA damage-induced mutagenesis in Saccharomyces cerevisiae
  JOURNAL   Yeast 10 (11), 1503-1509 (1994)
  PUBMED    7871890
FEATURES             Location/Qualifiers
     source          1..3240
                     /organism="Saccharomyces cerevisiae"
                     /db_xref="taxon:4932"
                     /chromosome="IX"
                     /map="9"
     CDS             <1..206
                     /codon_start=3
                     /product="TCP1-beta"
                     /protein_id="AAA98665.1"
                     /db_xref="GI:1293614"
                     /translation="SSIYNGISTSGLDLNNGTIADMRQLGIVESYKLKRAVVSSASEA
                     AEVLLRVDNIIRARPRTANRQHM"
     gene            687..3158
                     /gene="AXL2"
     CDS             687..3158
                     /gene="AXL2"
                     /note="plasma membrane glycoprotein"
                     /codon_start=1
                     /function="required for axial budding pattern of S.
                     cerevisiae"
                     /product="Axl2p"
                     /protein_id="AAA98666.1"
                     /db_xref="GI:1293615"
                     /translation="MTQLQISLLLTATISLLHLVVATPYEAYPIGKQYPPVARVNESF
                     TFQISNDTYKSSVDKTAQITYNCFDLPSWLSFDSSSRTFSGEPSSDLLSDANTTLYFN
                     VILEGTDSADSTSLNNTYQFVVTNRPSISLSSDFNLLALLKNYGYTNGKNALKLDPNE
                     VFNVTFDRSMFTNEESIVSYYGRSQLYNAPLPNWLFFDSGELKFTGTAPVINSAIAPE
                     TSYSFVIIATDIEGFSAVEVEFELVIGAHQLTTSIQNSLIINVTDTGNVSYDLPLNYV
                     YLDDDPISSDKLGSINLLDAPDWVALDNATISGSVPDELLGKNSNPANFSVSIYDTYG
                     DVIYFNFEVVSTTDLFAISSLPNINATRGEWFSYYFLPSQFTDYVNTNVSLEFTNSSQ
                     DHDWVKFQSSNLTLAGEVPKNFDKLSLGLKANQGSQSQELYFNIIGMDSKITHSNHSA
                     NATSTRSSHHSTSTSSYTSSTYTAKISSTSAAATSSAPAALPAANKTSSHNKKAVAIA
                     CGVAIPLGVILVALICFLIFWRRRRENPDDENLPHAISGPDLNNPANKPNQENATPLN
                     NPFDDDASSYDDTSIARRLAALNTLKLDNHSATESDISSVDEKRDSLSGMNTYNDQFQ
                     SQSKEELLAKPPVQPPESPFFDPQNRSSSVYMDSEPAVNKSWRYTGNLSPVSDIVRDS
                     YGSQKTVDTEKLFDLEAPEKEKRTSRDVTMSSLDPWNSNISPSPVRKSVTPSPYNVTK
                     HRNRHLQNIQDSQSGKNGITPTTMSTSSSDDFVPVKDGENFCWVHSMEPDRRPSKKRL
                     VDFSNKSNVNVGQVKDIHGRIPEML"
ORIGIN
        1 gatcctccat atacaacggt atctccacct caggtttaga tctcaacaac ggaaccattg
       61 ccgacatgag acagttaggt atcgtcgaga gttacaagct aaaacgagca gtagtcagct
      121 ctgcatctga agccgctgaa gttctactaa gggtggataa catcatccgt gcaagaccaa
      181 gaaccgccaa tagacaacat atgtaacata tttaggatat acctcgaaaa taataaaccg
      241 ccacactgtc attattataa ttagaaacag aacgcaaaaa ttatccacta tataattcaa
      301 agacgcgaaa aaaaaagaac aacgcgtcat agaacttttg gcaattcgcg tcacaaataa
      361 attttggcaa cttatgtttc ctcttcgagc agtactcgag ccctgtctca agaatgtaat
      421 aatacccatc gtaggtatgg ttaaagatag catctccaca acctcaaagc tccttgccga
      481 gagtcgccct cctttgtcga gtaattttca cttttcatat gagaacttat tttcttattc
      541 tttactctca catcctgtag tgattgacac tgcaacagcc accatcacta gaagaacaga
      601 acaattactt aatagaaaaa ttatatcttc ctcgaaacga tttcctgctt ccaacatcta
      661 cgtatatcaa gaagcattca cttaccatga cacagcttca gatttcatta ttgctgacag
      721 ctactatatc actactccat ctagtagtgg ccacgcccta tgaggcatat cctatcggaa
      781 aacaataccc cccagtggca agagtcaatg aatcgtttac atttcaaatt tccaatgata
     ……..
//
```

General Feature Format (GFF) file extension .gff

A General Feature Format (GFF) file is a simple tab-delimited text file for describing genomic features, such as genes, coding sequences (CDS), RNAs. There are several slightly but significantly different GFF file formats (.gff2, .gff3,..).

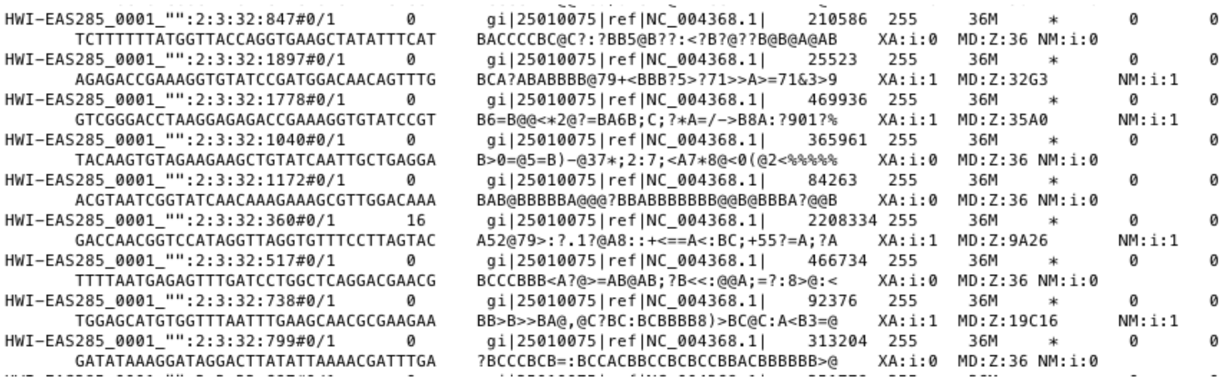


```
##gff-version 3.2.1
##sequence-region ctg123 1 1497228
ctg123 . gene             1000  9000  .  +  .  ID=gene00001;Name=EDEN
ctg123 . TF_binding_site  1000  1012  .  +  .  ID=tfbs00001;Parent=gene00001
ctg123 . mRNA             1050  9000  .  +  .  ID=mRNA00001;Parent=gene00001;Name=EDEN.1
ctg123 . five_prime_UTR   1050  1200  .  +  .  Parent=mRNA00001
ctg123 . CDS              1201  1500  .  +  0  ID=cds00001;Parent=mRNA00001
ctg123 . CDS              3000  3902  .  +  0  ID=cds00001;Parent=mRNA00001
ctg123 . CDS              5000  5500  .  +  0  ID=cds00001;Parent=mRNA00001
ctg123 . CDS              7000  7600  .  +  0  ID=cds00001;Parent=mRNA00001
ctg123 . three_prime_UTR  7601  9000  .  +  .  Parent=mRNA00001
```

Column 1: "seqid"
Column 2: "source"
Column 3: "type"
Columns 4 & 5: "start" and "end"
Column 6: "score"
Column 7: "strand"
Column 8: "phase"
Column 9: "attributes"

Sequence Alignment Map format: extension .sam and .bam

Sequence Alignment Map (SAM) is a text-based format originally for storing biological sequences aligned to a reference sequence. It is widely used for storing data, such as nucleotide sequences, generated by next generation sequencing technologies, and the standard has been broadened to include unmapped sequences. It may contain base-call and alignment qualities and other data.



```
HWI-EAS285_0001_"":2:3:32:847#0/1       0       gi|25010075|ref|NC_004368.1|    210586  255     36M     *       0       0
        TCTTTTTTATGGTTACCAGGTGAAGCTATATTTCAT    BACCCCBC@C?:?BB5@B??:<?B?@??B@B@A@AB    XA:i:0  MD:Z:36 NM:i:0
HWI-EAS285_0001_"":2:3:32:1897#0/1      0       gi|25010075|ref|NC_004368.1|    25523   255     36M     *       0       0
        AGAGACCGAAAGGTGTATCCGATGGACAACAGTTTG    BCA?ABABBBB@79+<BBB?5>?71>>A>=71&3>9    XA:i:1  MD:Z:32G3       NM:i:1
HWI-EAS285_0001_"":2:3:32:1778#0/1      0       gi|25010075|ref|NC_004368.1|    469936  255     36M     *       0       0
        GTCGGGACCTAAGGAGAGACCGAAAGGTGTATCCGT    B6=B@@<*2@?=BA6B;C;?*A=/->B8A:?901?%    XA:i:1  MD:Z:35A0       NM:i:1
HWI-EAS285_0001_"":2:3:32:1040#0/1      0       gi|25010075|ref|NC_004368.1|    365961  255     36M     *       0       0
        TACAAGTGTAGAAGAAGCTGTATCAATTGCTGAGGA    B>0=@5=B)-@37*;2:7;<A7*8@<0(@2<%%%%%    XA:i:0  MD:Z:36 NM:i:0
HWI-EAS285_0001_"":2:3:32:1172#0/1      0       gi|25010075|ref|NC_004368.1|    84263   255     36M     *       0       0
        ACGTAATCGGTATCAACAAAGAAAGCGTTGGACAAA    BAB@BBBBBA@@@?BBABBBBBBB@@B@BBBA?@@B    XA:i:0  MD:Z:36 NM:i:0
HWI-EAS285_0001_"":2:3:32:360#0/1       16      gi|25010075|ref|NC_004368.1|    2208334 255     36M     *       0       0
        GACCAACGGTCCATAGGTTAGGTGTTTCCTTAGTAC    A52@79>:?.1?@A8::+<==A<:BC;+55?=A;?A    XA:i:1  MD:Z:9A26       NM:i:1
HWI-EAS285_0001_"":2:3:32:517#0/1       0       gi|25010075|ref|NC_004368.1|    466734  255     36M     *       0       0
        TTTTAATGAGAGTTTGATCCTGGCTCAGGACGAACG    BCCCBBB<A?@>=AB@AB;?B<<:@@A;=?:8>@:<    XA:i:0  MD:Z:36 NM:i:0
HWI-EAS285_0001_"":2:3:32:738#0/1       0       gi|25010075|ref|NC_004368.1|    92376   255     36M     *       0       0
        TGGAGCATGTGGTTTAATTTGAAGCAACGCGAAGAA    BB>B>>BA@,@C?BC:BCBBBB8)>BC@C:A<B3=@    XA:i:1  MD:Z:19C16      NM:i:1
HWI-EAS285_0001_"":2:3:32:799#0/1       0       gi|25010075|ref|NC_004368.1|    313204  255     36M     *       0       0
        GATATAAAGGATAGGACTTATATTAAAACGATTTGA    ?BCCCBCB=:BCCACBBCCBCBCCBBACBBBBBB>@    XA:i:0  MD:Z:36 NM:i:0
```

The BAM Format is a binary format of SAM files. The SAM format is more human readable, and easier to process by conventional text based processing programs. The BAM format, only machine readable, provides a binary version designed to compress reasonably well.