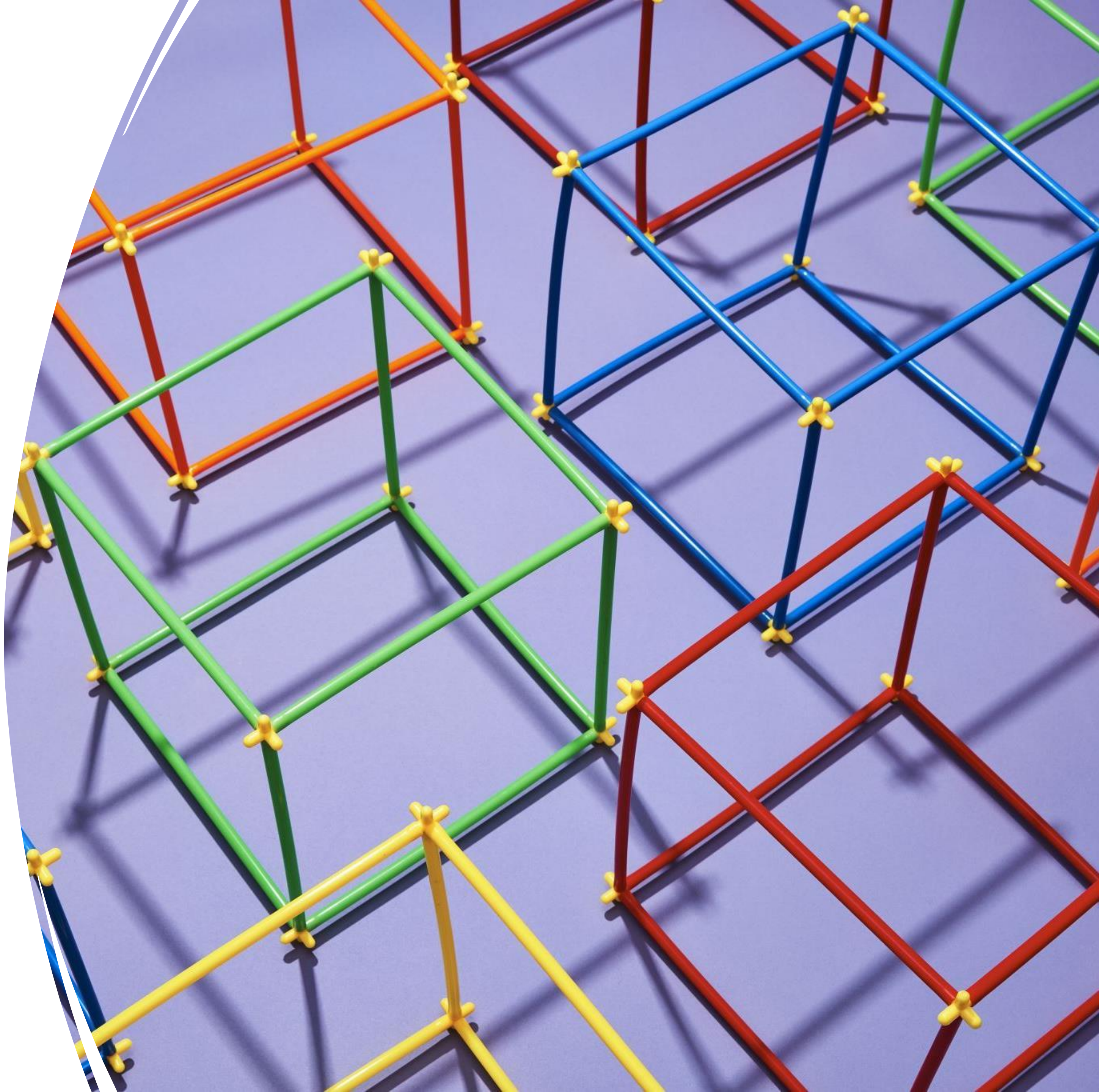
Data Formats in Bioinformatics
&
Where to Find Them

Data format

- definition of the structure of data within a database or file system that gives the information its meaning



Type of data formats

- Sequence formats
- Alignment formats
- Generic feature formats
- Annotation formats
- Protein structure formats
- Index formats
- ...

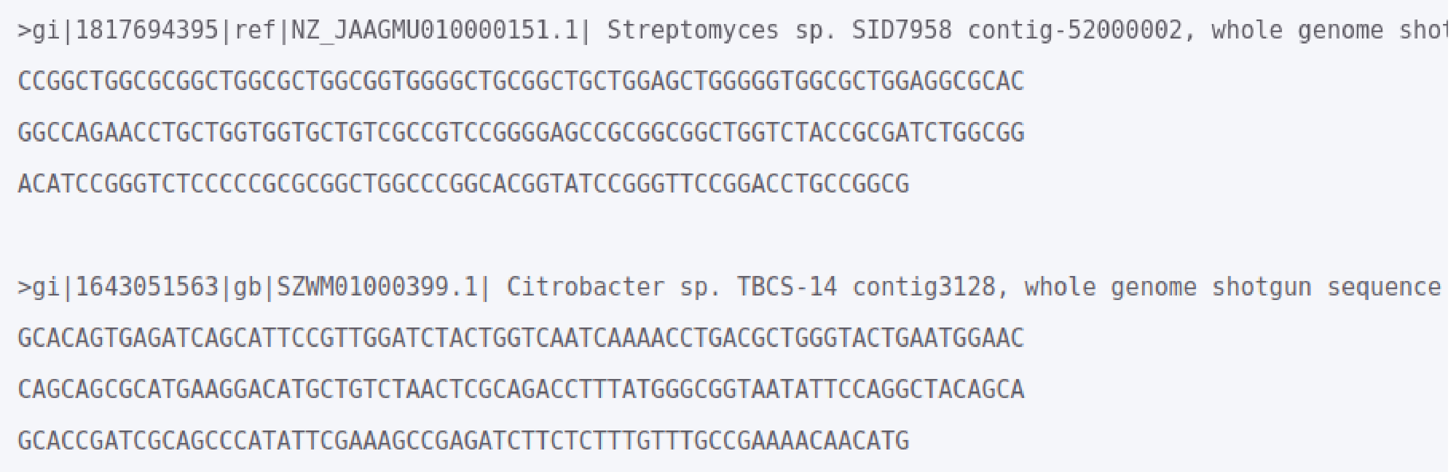
- Text formats (Human readable)
- Binary formats

- Flat formats

- Open formats
- Vendor-lock formats

Sequence formats

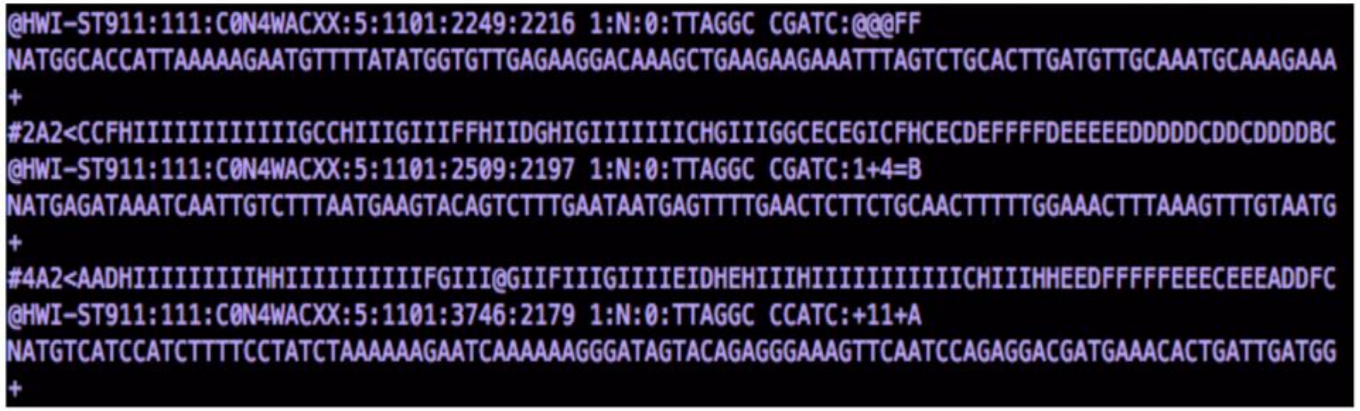
- Fasta
  - Plain sequence
  - Contains header
  - .fasta, .fa, .fna, .faa, .frn



```
>gi|1817694395|ref|NZ_JAAGMU010000151.1| Streptomyces sp. SID7958 contig-52000002, whole genome shot
CCGGCTGGCGCGGCTGGCGCTGGCGGTGGGGCTGCGGCTGCTGGAGCTGGGGGTGGCGCTGGAGGCGCAC
GGCCAGAACCTGCTGGTGGTGCTGTCGCCGTCCGGGGAGCCGCGGCGGCTGGTCTACCGCGATCTGGCGG
ACATCCGGGTCTCCCCCGCGCGGCTGGCCCGGCACGGTATCCGGGTTCCGGACCTGCCGGCG

>gi|1643051563|gb|SZWM01000399.1| Citrobacter sp. TBCS-14 contig3128, whole genome shotgun sequence
GCACAGTGAGATCAGCATTCCGTTGGATCTACTGGTCAATCAAAACCTGACGCTGGGTACTGAATGGAAC
CAGCAGCGCATGAAGGACATGCTGTCTAACTCGCAGACCTTTATGGGCGGTAATATTCCAGGCTACAGCA
GCACCGATCGCAGCCCATATTCGAAAGCCGAGATCTTCTCTTTGTTTGCCGAAAACAACATG
```

Sequence formats

- FastQ
  - Sequencing format
  - Contains quality string (Phred Score)
  - .fastq, .fq

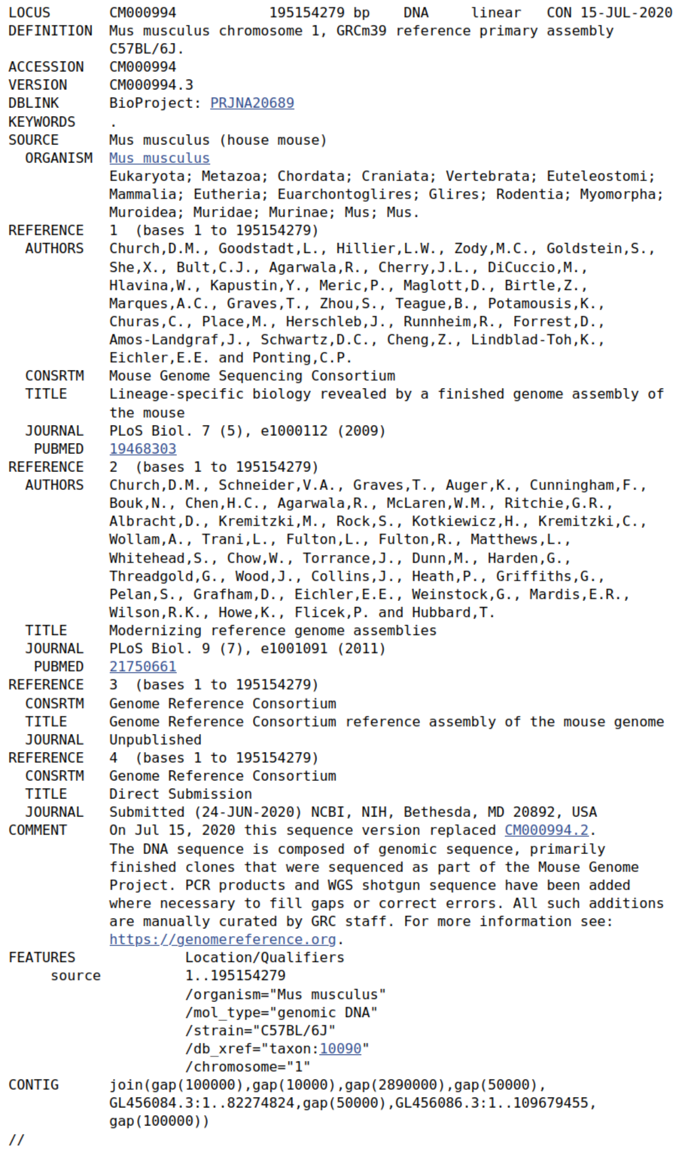


```
@HWI-ST911:111:C0N4WACXX:5:1101:2249:2216 1:N:0:TTAGGC CGATC:@@@FF
NATGGCACCATTAAAAAGAATGTTTTATATGGTGTTGAGAAGGACAAAGCTGAAGAAGAAATTTAGTCTGCACTTGATGTTGCAAATGCAAAGAAA
+
#2A2<CCFHIIIIIIIIIIIIIGCCHIIIGIIIFFHIIDGHIGIIIIIIICHGIIIGGCECEGICFHCECDEFFFFDEEEEEDDDDDCDDCDDDDBC
@HWI-ST911:111:C0N4WACXX:5:1101:2509:2197 1:N:0:TTAGGC CGATC:1+4=B
NATGAGATAAATCAATTGTCTTTAATGAAGTACAGTCTTTGAATAATGAGTTTTGAACTCTTCTGCAACTTTTTGGAAACTTTAAAGTTTGTAATG
+
#4A2<AADHIIIIIIIIIIIHHIIIIIIIIIIIIFGIII@GIIFIIIGIIIIEIDHEHIIIHIIIIIIIIIIIICHIIIHHEEDFFFFFEEECEEEADDFC
@HWI-ST911:111:C0N4WACXX:5:1101:3746:2179 1:N:0:TTAGGC CCATC:+11+A
NATGTCATCCATCTTTTCCTATCTAAAAAAGAATCAAAAAAGGGATAGTACAGAGGGAAAGTTCAATCCAGAGGACGATGAAACACTGATTGATGG
+
```

HW-ST911	the unique instrument name
111	the run id
C0N4WACXX	the flowcell id
5	flowcell lane
1101	tile number within the flowcell lane
2249	'x'-coordinate of the cluster within the tile
2216	'y'-coordinate of the cluster within the tile
1	the member of a pair, 1 or 2 (paired-end or mate-pair reads only)
Y	Y if the read is filtered, N otherwise
18	0 when none of the control bits are on
TTAGGC, CGATC	index sequence

Sequence formats

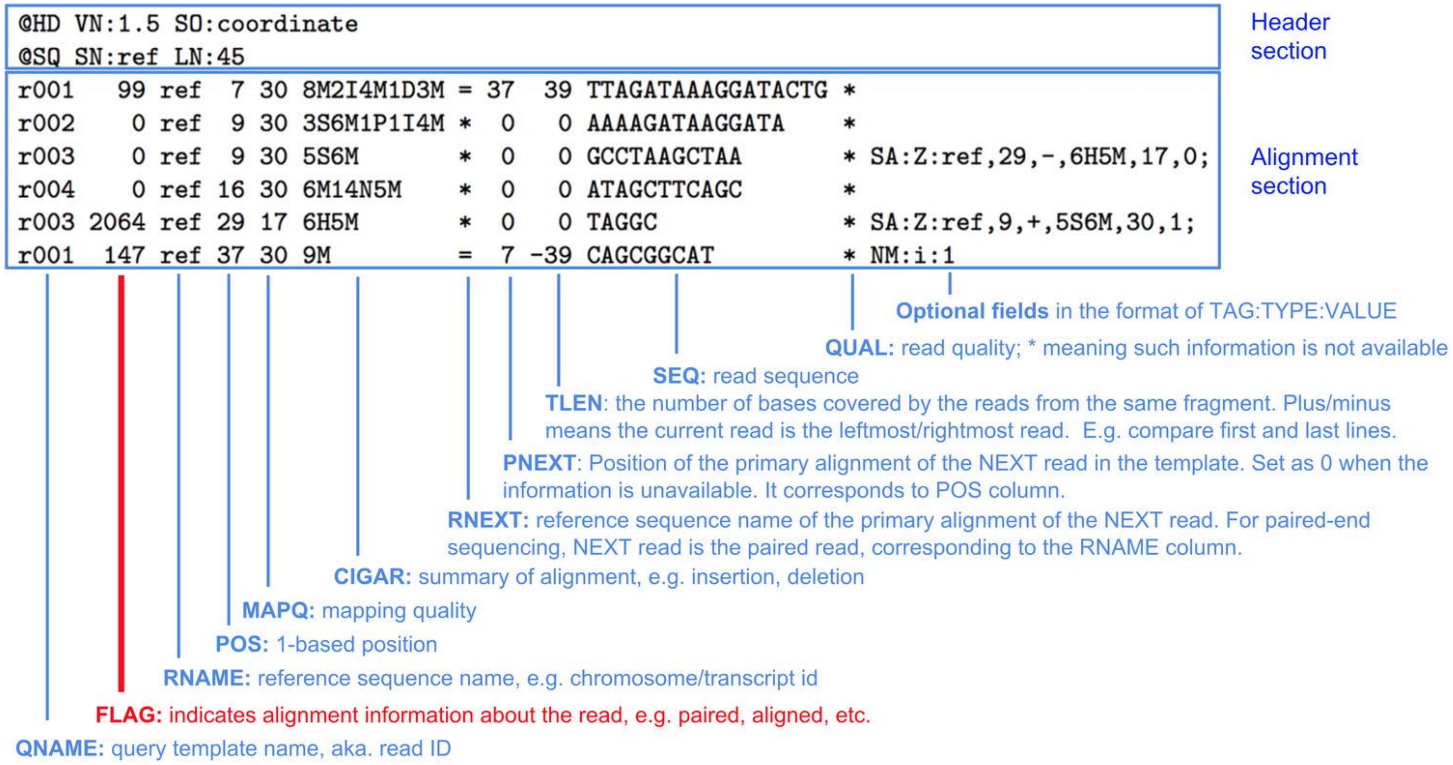
- GenBank
  - Contains addition information about the sequence
  - .genbank, .gb



```
LOCUS       CM000994           195154279 bp    DNA     linear   CON 15-JUL-2020
DEFINITION  Mus musculus chromosome 1, GRCm39 reference primary assembly
            C57BL/6J.
ACCESSION   CM000994
VERSION     CM000994.3
DBLINK      BioProject: PRJNA20689
KEYWORDS    .
SOURCE      Mus musculus (house mouse)
  ORGANISM  Mus musculus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia; Myomorpha;
            Muroidea; Muridae; Murinae; Mus; Mus.
REFERENCE   1  (bases 1 to 195154279)
  AUTHORS   Church,D.M., Goodstadt,L., Hillier,L.W., Zody,M.C., Goldstein,S.,
            She,X., Bult,C.J., Agarwala,R., Cherry,J.L., DiCuccio,M.,
            Hlavina,W., Kapustin,Y., Meric,P., Maglott,D., Birtle,Z.,
            Marques,A.C., Graves,T., Zhou,S., Teague,B., Potamousis,K.,
            Churas,C., Place,M., Herschleb,J., Runnheim,R., Forrest,D.,
            Amos-Landgraf,J., Schwartz,D.C., Cheng,Z., Lindblad-Toh,K.,
            Eichler,E.E. and Ponting,C.P.
  CONSRTM   Mouse Genome Sequencing Consortium
  TITLE     Lineage-specific biology revealed by a finished genome assembly of
            the mouse
  JOURNAL   PLoS Biol. 7 (5), e1000112 (2009)
   PUBMED   19468303
REFERENCE   2  (bases 1 to 195154279)
  AUTHORS   Church,D.M., Schneider,V.A., Graves,T., Auger,K., Cunningham,F.,
            Bouk,N., Chen,H.C., Agarwala,R., McLaren,W.M., Ritchie,G.R.,
            Albracht,D., Kremitzki,M., Rock,S., Kotkiewicz,H., Kremitzki,C.,
            Wollam,A., Trani,L., Fulton,L., Fulton,R., Matthews,L.,
            Whitehead,S., Chow,W., Torrance,J., Dunn,M., Harden,G.,
            Threadgold,G., Wood,J., Collins,J., Heath,P., Griffiths,G.,
            Pelan,S., Grafham,D., Eichler,E.E., Weinstock,G., Mardis,E.R.,
            Wilson,R.K., Howe,K., Flicek,P. and Hubbard,T.
  TITLE     Modernizing reference genome assemblies
  JOURNAL   PLoS Biol. 9 (7), e1001091 (2011)
   PUBMED   21750661
REFERENCE   3  (bases 1 to 195154279)
  CONSRTM   Genome Reference Consortium
  TITLE     Genome Reference Consortium reference assembly of the mouse genome
  JOURNAL   Unpublished
REFERENCE   4  (bases 1 to 195154279)
  CONSRTM   Genome Reference Consortium
  TITLE     Direct Submission
  JOURNAL   Submitted (24-JUN-2020) NCBI, NIH, Bethesda, MD 20892, USA
COMMENT     On Jul 15, 2020 this sequence version replaced CM000994.2.
            The DNA sequence is composed of genomic sequence, primarily
            finished clones that were sequenced as part of the Mouse Genome
            Project. PCR products and WGS shotgun sequence have been added
            where necessary to fill gaps or correct errors. All such additions
            are manually curated by GRC staff. For more information see:
            https://genomereference.org.
FEATURES             Location/Qualifiers
     source          1..195154279
                     /organism="Mus musculus"
                     /mol_type="genomic DNA"
                     /strain="C57BL/6J"
                     /db_xref="taxon:10090"
                     /chromosome="1"
CONTIG      join(gap(100000),gap(10000),gap(2890000),gap(50000),
            GL456084.3:1..82274824,gap(50000),GL456086.3:1..109679455,
            gap(100000))
//
```

Alignment formats

- SAM
  - Sequence Alignment map
  - Contains additional information
- BAM
  - Binary version of SAM
- CRAM
  - BAM with loseless compression
  - .sam, .bam, .cram



```
@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
```

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

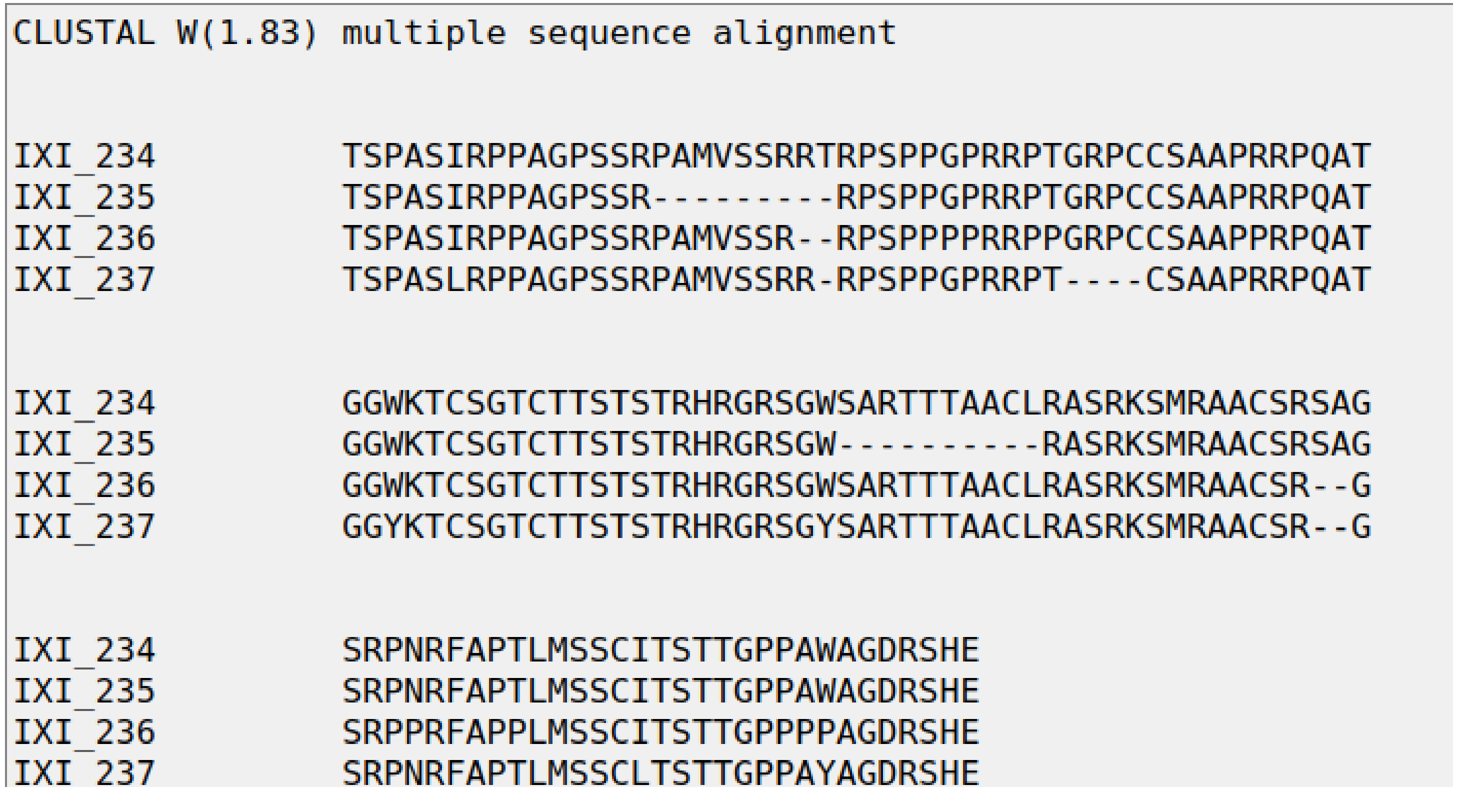
RNAME: reference sequence name, e.g. chromosome/transcript id

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

QNAME: query template name, aka. read ID

Alignment formats

- Clustal
  - Clustal omega
  - Multiple sequence alignment
  - .clusta, .aln



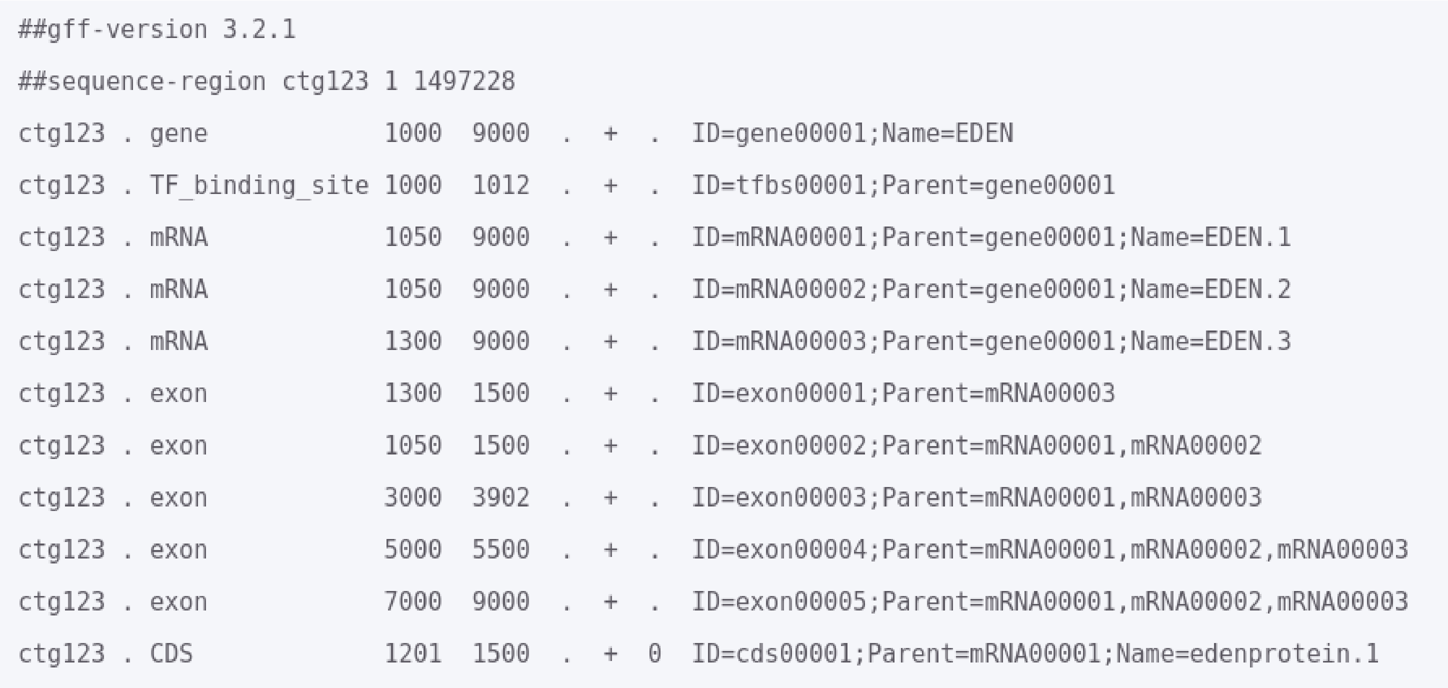
```
CLUSTAL W(1.83) multiple sequence alignment


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


IXI_234         GGWKTCSGTCTTSTSTRHGRSGWSARTTTAACLRASRKSMRAACSRSAG
IXI_235         GGWKTCSGTCTTSTSTRHGRSGW----------RASRKSMRAACSRSAG
IXI_236         GGWKTCSGTCTTSTSTRHGRSGWSARTTTAACLRASRKSMRAACSR--G
IXI_237         GGYKTCSGTCTTSTSTRHGRSGYSARTTTAACLRASRKSMRAACSR--G


IXI_234         SRPNRFAPTLMSSCITSTTGPPAWAGDRSHE
IXI_235         SRPNRFAPTLMSSCITSTTGPPAWAGDRSHE
IXI_236         SRPPRFAPPLMSSCITSTTGPPPPAGDRSHE
IXI_237         SRPNRFAPTLMSSCLTSTTGPPAYAGDRSHE
```

Generic feature formats

- GTF
- GFF
- GFF3

- Describing genes and other features
- Beware of version!

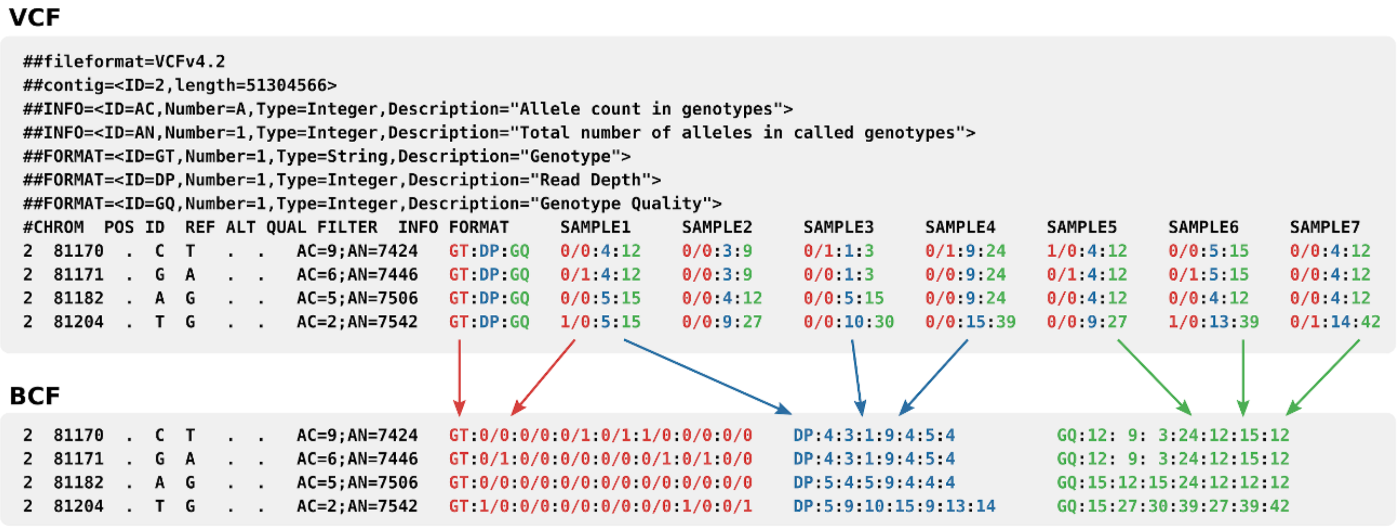
  - .gtf, .gff, .gff3



```
##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 . mRNA            1050  9000  .  +  .  ID=mRNA00002;Parent=gene00001;Name=EDEN.2
ctg123 . mRNA            1300  9000  .  +  .  ID=mRNA00003;Parent=gene00001;Name=EDEN.3
ctg123 . exon            1300  1500  .  +  .  ID=exon00001;Parent=mRNA00003
ctg123 . exon            1050  1500  .  +  .  ID=exon00002;Parent=mRNA00001,mRNA00002
ctg123 . exon            3000  3902  .  +  .  ID=exon00003;Parent=mRNA00001,mRNA00003
ctg123 . exon            5000  5500  .  +  .  ID=exon00004;Parent=mRNA00001,mRNA00002,mRNA00003
ctg123 . exon            7000  9000  .  +  .  ID=exon00005;Parent=mRNA00001,mRNA00002,mRNA00003
ctg123 . CDS             1201  1500  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
```

Annotation formats

- VCF
  - Variant calling format
  - Information about SNP
  - Genotyping projects
  - .vcf
- BCF
  - Binary version of vcf



VCF

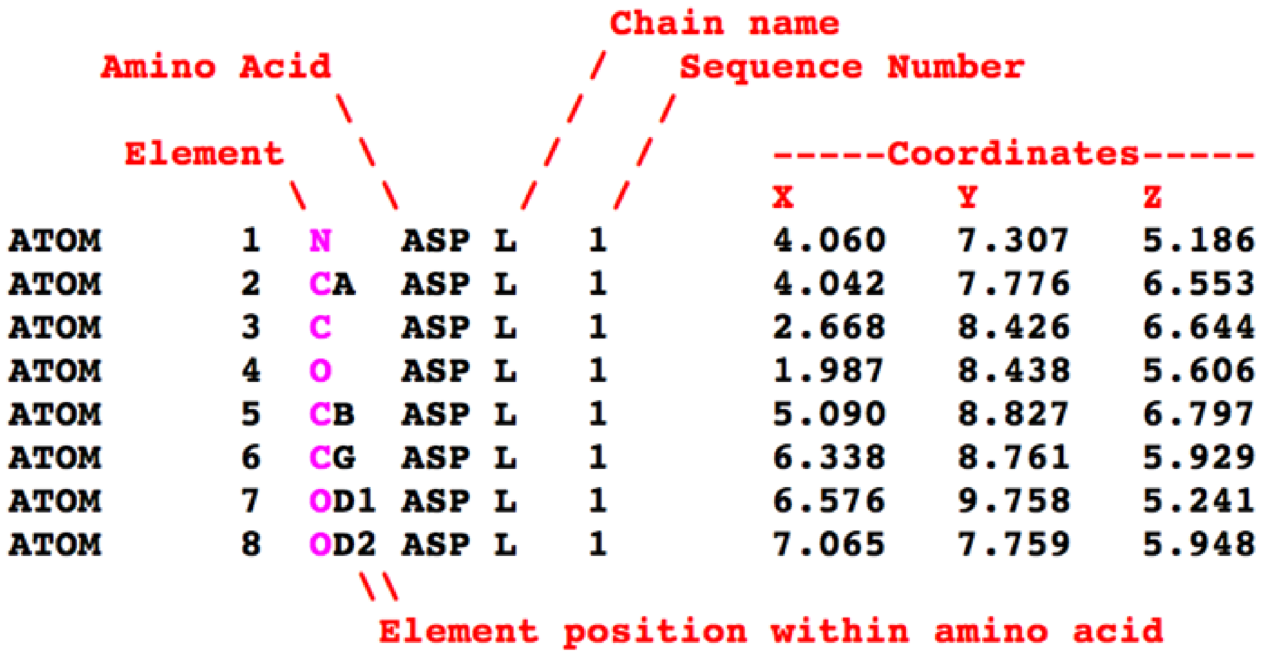
```
##fileformat=VCFv4.2
##contig=<ID=2,length=51304566>
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
#CHROM  POS ID  REF ALT QUAL FILTER  INFO FORMAT     SAMPLE1     SAMPLE2     SAMPLE3     SAMPLE4     SAMPLE5     SAMPLE6     SAMPLE7
2  81170  .  C  T   .  .   AC=9;AN=7424  GT:DP:GQ  0/0:4:12    0/0:3:9     0/1:1:3     0/1:9:24    1/0:4:12    0/0:5:15    0/0:4:12
2  81171  .  G  A   .  .   AC=6;AN=7446  GT:DP:GQ  0/1:4:12    0/0:3:9     0/0:1:3     0/0:9:24    0/1:4:12    0/1:5:15    0/0:4:12
2  81182  .  A  G   .  .   AC=5;AN=7506  GT:DP:GQ  0/0:5:15    0/0:4:12    0/0:5:15    0/0:9:24    0/0:4:12    0/0:4:12    0/0:4:12
2  81204  .  T  G   .  .   AC=2;AN=7542  GT:DP:GQ  1/0:5:15    0/0:9:27    0/0:10:30   0/0:15:39   0/0:9:27    1/0:13:39   0/1:14:42
```

BCF

```
2  81170  .  C  T   .  .   AC=9;AN=7424  GT:0/0:0/0:0/1:0/1:1/0:0/0:0/0   DP:4:3:1:9:4:5:4        GQ:12: 9: 3:24:12:15:12
2  81171  .  G  A   .  .   AC=6;AN=7446  GT:0/1:0/0:0/0:0/0:0/1:0/1:0/0   DP:4:3:1:9:4:5:4        GQ:12: 9: 3:24:12:15:12
2  81182  .  A  G   .  .   AC=5;AN=7506  GT:0/0:0/0:0/0:0/0:0/0:0/0:0/0   DP:5:4:5:9:4:4:4        GQ:15:12:15:24:12:12:12
2  81204  .  T  G   .  .   AC=2;AN=7542  GT:1/0:0/0:0/0:0/0:0/0:1/0:0/1   DP:5:9:10:15:9:13:14    GQ:15:27:30:39:27:39:42
```

Protein structure format

- PDB
  - Information about protein atoms
  - .pdb



```
ATOM      1  N   ASP L   1       4.060   7.307   5.186
ATOM      2  CA  ASP L   1       4.042   7.776   6.553
ATOM      3  C   ASP L   1       2.668   8.426   6.644
ATOM      4  O   ASP L   1       1.987   8.438   5.606
ATOM      5  CB  ASP L   1       5.090   8.827   6.797
ATOM      6  CG  ASP L   1       6.338   8.761   5.929
ATOM      7  OD1 ASP L   1       6.576   9.758   5.241
ATOM      8  OD2 ASP L   1       7.065   7.759   5.948
```

Chain name
Amino Acid
Sequence Number
Element
-----Coordinates-----
X
Y
Z
Element position within amino acid

Index formats

- For quicker searching in bioinformatics formats
- Software dependent
- Binary structure as hash table, suffix tree, k-mer composition, …
- .fai, .bai, .crai, .index, …

Data Compression

- Text formats are commonly compressed
  - .gz
  - .bz2
  - .tar
  - .zip
  - …

Text format X binary format

- Pros:

- Cons:

Open format X Vendor-lock format

- Pros:

- Cons:

Why so many formats?

Why so many formats?

- Compatibility
- Speed
- Readability
- Storage efficiency
- Structuring needs

- Important metadata
    - Transformers
    - Versioning
    - Source!

Example Human reference genome

- GRCh38.p14
- GRCh37
- Hg19
- GCA_000001405.29

- Which one to use?
- What is the difference?
- What are the implications?

Data sources

- Your own laboratory
- Publicly available databases

- Accessible via internet browser
- Dedicated API

NCBI

- https://www.ncbi.nlm.nih.gov/
- National Center for Biotechnology Information
- Aggregation of several data sources into one project

- Genbank
- Refseq
- SRA
- PubMed
- …

EMBL-EBI

- https://www.ebi.ac.uk/
- European Molecular Biology Laboratory - European Bioinformatics Institute
    - https://www.ebi.ac.uk/services/data-resources-and-tools
- ENA – European Nucleotide Archive
    - https://www.ebi.ac.uk/ena/browser/home

UCSC

- https://genome.ucsc.edu/
- University of California, Santa Cruz, Genomic Institute

UniProt

- https://www.uniprot.org/
- Several aggregated data sources about proteins