FastA Format

(modified from: https://learn.gencore.bio.nyu.edu)

File extensions :   file.fa, file.fasta, file.fsa

The official FastA documentation can be found here

FastA format is the most basic format for reporting a sequence and is accepted by almost all sequence analysis program. It only contains a sequence name, a description of the sequence (metadata, sequencer info, annotations, etc.), and the sequence itself – it can be either nucleic acids or amino acids as long as it adheres to the format.
Each sequence consists of at least two lines:

1. The first is the sequence header, which always starts with a '>'
    - Everything from the beginning '>' to the first whitespace is considered the sequence identifier. Everything after that is considered the sequence description (this can be metadata, machine serial number, read orientation, etc.)
2. The sequence itself
    - Note that the sequence can span multiple lines, depending on the length of the sequence.

```
>Chr1 CHROMOSOME dumped from ADB: Jun/20/09 14:53; last updated: 2009-02-02
CCCTAAACCCTAAACCCTAAACCCTAAACCTCTGAATCCTTAATCCCTAAATCCCTAAATCTTTAAATCCTACATCCAT
GAATCCCTAAATACCTAATTCCCTAAACCCGAAACCGGTTTCTCTGGTTGAAAATCATTGTGTATATAATGATAATTTT
ATCGTTTTTATGTAATTGCTTATTGTTGTGTGTAGATTTTTTAAAAATATCATTTGAGGTCAATACAAATCCTATTTCT
TGTGGTTTTCTTTCCTTCACTTAGCTATGGATGGTTTATCTTCATTTGTTATATTGGATACAAGCTTTGCTACGATCTA
CATTTGGGAATGTGAGTCTCTTATTGTAACCTTAGGGTTGGTTTATCTCAAGAATCTTATTAATTGTTTGGACTGTTTA
TGTTTGGACATTTATTGTCATTCTTACTCCTTTGTGGAAATGTTTGTTCTATCAATTTATCTTTTGTGGGAAAATTATT
TAGTTGTAGGGATGAAGTCTTTCTTCGTTGTTGTTACGCTTGTCATCTCATCTCTCAATGATATGGGATGGTCCTTTAG
CATTTATTCTGAAGTTCTTCTGCTTGATGATTTTATCCTTAGCCAAAAGGATTGGTGGTTTGAAGACACATCATATCAA
AAAAGCTATCGCCTCGACGATGCTCTATTTCTATCCTTGTAGCACACATTTTGGCACTCAAAAAAGTATTTTTAGATGT
TTGTTTTGCTTCTTTGAAGTAGTTTCTCTTTGCAAAATTCCTCTTTTTTTTAGAGTGATTTGGATGATTCAAGACTTCTC
GGTACTGCAAAGTTCTTCCGCCTGATTAATTATCCATTTTACCTTTGTCGTAGATATTAGGTAATCTGTAAGTCAACTC
ATATACAACTCATAATTTAAAATAAAATTATGATCGACACACGTTTACACATAAAATCTGTAAATCAACTCATATACCC
GTTATTCCCACAATCATATGCTTTCTAAAAGCAAAAGTATATGTCAACAATTGGTTATAAATTATTAGAAGTTTTCCAC
TTATGACTTAAGAACTTGTGAAGCAGAAAGTGGCAACACCCCCCACCTCCCCCCCCCCCCCCACCCCCCAAATTGAGA
AGTCAATTTTATATAATTTAATCAAATAAATAAGTTTATGGTTAAGAGTTTTTTACTCTCTTTATTTTTCTTTTTCTTT
```

Sequences are expected to be represented in the standard IUB/IUPAC amino acid and nucleic acid codes, with these exceptions:

- lower-case letters are accepted and are mapped into upper-case;

- a single hyphen or dash can be used to represent a gap of indeterminate length;
- in amino acid sequences, U and * are acceptable letters (see below).
- any numerical digits in the query sequence should either be removed or replaced by appropriate letter codes (e.g., N for unknown nucleic acid residue or X for unknown amino acid residue).

The nucleic acid codes are:

```
A --> adenosine           M --> A C (amino)
C --> cytidine            S --> G C (strong)
G --> guanine             W --> A T (weak)
T --> thymidine           B --> G T C
U --> uridine             D --> G A T
R --> G A (purine)        H --> A C T
Y --> T C (pyrimidine)    V --> G C A
K --> G T (keto)          N --> A G C T (any)
                          -  gap of indeterminate length
```

The accepted amino acid codes are:

```
    A ALA alanine                         P PRO proline
    B ASX aspartate or asparagine         Q GLN glutamine
    C CYS cystine                         R ARG arginine
    D ASP aspartate                       S SER serine
    E GLU glutamate                       T THR threonine
    F PHE phenylalanine                   U     selenocysteine
    G GLY glycine                         V VAL valine
    H HIS histidine                       W TRP tryptophan
    I ILE isoleucine                      Y TYR tyrosine
    K LYS lysine                          Z GLX glutamate or
glutamine
    L LEU leucine                         X     any
    M MET methionine                      *     translation
stop
    N ASN asparagine                      -     gap of
indeterminate length
```

Software that use FastA format

FASTA is pretty much a ubiquitous format. Most downstream analysis programs can import data in FASTA format in some way. You will also see FASTA as options for just about all the genome browsers and DB query tools like blast. Just about all multiple-sequence alignment algorithms accept only FastA format. Also, when you download reference genomes they are delivered in this format (usually along with a GFF file).

How are these files generated?

- Some older NGS sequencers either report sequences in this format or the output can be easily convereted to FASTA. For example, Sanger sequencing results in either an abi or scf format file that is easily converted this format.
- Most sequence databases store sequences in FastA format which is available for download.
- FastA can also generated from a FastQ file.

Let's grab one!