Bioinformatics Files

Bioinformatics analyses are key for understanding biology. However, bioinformatics analyses output a number of unique file formats that can be hard to interpret. These files are often compressed because they can be very large and usually require specialized tools, rather than standard programs, to be opened and analyzed. In this guide, we review some of the common files that can be provided as part of a standard bioinformatics analysis.

FASTA files

FASTA files are plain text files that contain sequence data for RNA, DNA, or proteins. They start with a ">" symbol and an identifier, and the next line(s) contain the nucleotide sequence. FASTA files are used for sequences of many lengths including reference genomes and IVT RNAs.

```
>ID
AUGCCGAAUUAGCA
```

FASTQ files

FASTQ files are plain text files that contain sequence data along with quality scores that indicate how reliable each part of the sequence is. Each entry in a FASTQ file starts with a "@" symbol and an identifier, followed by the sequence, a "+" symbol, and the quality scores. FASTQ files are the first output of next-generation sequencing-based analytics.

```
@SEQ_ID
AUGCCGAAUUAGCA
+
!!'(*(!''!!!(*
```

SAM files

Sequence Alignment and Map (SAM) files are plain text files that describe the alignment of sequences from a FASTQ file compared to a reference sequence. SAM files start with a header containing alignment information in lines beginning with an "@" symbol, including details about the reference sequence(s) and their length(s) (shown as "@SQ" lines) as well as the program used for alignment ("@PG" lines). Following the header, each entry provides information for an individual sequence, including where it aligns on the reference sequence, the alignment location as coordinates, and any gaps or mismatches between the sequence and the reference.

```
@HD    VN:1.6 SO:coordinate
@SQ    SN:chrRNA    LN:1000
@PG    ID:Program

rna_read_001 0 chr1 100 255 14M * 0 0
ATGCCGAATTAGCA !!'(*(!''!!!(*
```

BAM files

Binary Alignment and Map (BAM) files are compressed, binary files that contain the same sequence alignment information as SAM files. Since BAM files are in binary format, they need to be read by special programs and can't be opened in a text editor. Due to their smaller file size, BAM files are used by most bioinformatics tools in place of SAM files.

BAM index files

BAM index files are indexes for BAM files. Similar to a table of contents, they are a guide to where to find specific reads within a BAM file and are required by most bioinformatics tools to enable efficient BAM file access.

bigWig files

bigWig files are indexed, binary files of dense, continuous sequence coverage data. These files are created from wiggle (.wig) or bedGraph files and represent the distribution of sequencing data along a reference. bigWig files can be loaded into genome browsers, such as IGV, to visualize the sequencing data. Since bigWig files are indexed, only the portions needed to make a specific graph are accessed by the genome browser, making them a faster file to use for data visualizations than BAM files.

BED files

Browser Extensible Data (BED) files are plain text files that define features in an RNA or DNA reference sequence. Each feature is defined by its position along the reference and must include at least three required fields:

1. The name of the reference sequence (such as chromosome, contig, or other region)

2. A numerical starting position of the feature on the reference sequence

3. A numerical ending position of the feature on the reference sequence

There are up to nine additional optional fields that can be present and are used to provide more detail about the feature, such as which strand of DNA it is on, its name, or associated numerical values like scores of p-values.

```
chr9    1    145
```

The output of Eclipsebio's analytics platforms typically includes all of these file types. If you need assistance interpreting and visualizing these files in a genome browser, you can check out our eBlog.

Interested in obtaining bioinformatics support for your research or drug development goals? Contact Eclipsebio to learn more.