Glossary

Annotation file

This is a type of text file that summarizes the different features in a genome such as genes, proteins, and regulatory regions, etc. Annotation files can be generated for single sequences as well as for entire genomes. They have strict formatting rules.

Contig

From the words “contiguous”. A contig is made of a consensus sequence formed by more than one read, which are overlapped to form a longer sequence of DNA. Overlapping reads provide strong evidence for what constitutes the real DNA sequence. When various reads overlap and are in agreement with each other, a consensus is called and the stretch of DNA is named a contig. Note that a contig has no gaps and the nucleotide sequence is known for the whole length of the contig.

Optional: Read more about a contig by following the link to WikiPedia below the Glossary.

EMBL

The European Molecular Biology Laboratory is a scientific institution. EMBL has research laboratories and outposts in Germany, the UK, France, Italy, and Spain.

Enzyme Commission number

A Enzyme Commission number (EC number) is a numerical classification for an enzyme based on the chemical reactions that it catalyses. An EC number consists of the letters “EC”, followed by four numbers, separated by full-stops. For example EC 2.4.1.1 = Glycogen phosphorylase.

GC content

In genomics and genetics, the GC content is the proportion of Guanines (G) and Cytosines (C) present in a given stretch of DNA sequence. The calculation involves counting the number of Gs and Cs in a given stretch of DNA and dividing it by the total number of nucleotides/bases in that DNA stretch. The GC content can be calculated for a whole genome, as well as presented as a value for a given length of DNA. In this way, stretches of DNA with different GC content can be identified.

GenBank

GenBank is a nucleotide sequence database hosted by the National Center for Biotechnology Information (NCBI). Genome sequences can be downloaded easily from this database.

Genome assembly (or sequence assembly)

Genome assembly (or sequence assembly) is the collection of scaffolds or sequenced DNA from a given organism. A genome assembly can be different from a reference genome (see below). For example, the genome assemblies of different bacterial isolates from around the world can be compared to the laboratory strain that was used to build the reference genome.

Optional: Read more about genome assembly by following the WikiPedia link given below the Glossary.

GFF

A file format that usually contains genome annotation information.

Origin

Origin of replication of a bacterial chromosome is a region of specific sequence within the genome, where the replication (copying) of the DNA molecule starts.

Optional: Read more about replication by following the WikiPedia link given below the Glossary.

Pseudogene

A region of the genome that contains a degraded gene. They originate by gene duplication and subsequent loss of function, due to accumulated mutations. Pseudogenes do not code functional proteins.

Optional: Read more about pseudogenes by following the WikiPedia link given below the Glossary.

Read (in sequencing)

In the context of DNA sequencing, a read is a stretch of inferred sequence coming from the sequencing of a DNA fragment. Read length varies depending on the sequencing technology used and could be short (20-30 nucleotides) or very long (several thousand nucleotides).

Optional: Read more about a read (in sequencing) by following the WikiPedia link given below the Glossary.

Reference genome

A reference genome is the representative genome of a given species. Because all individuals in a given species can differ in their exact sequence, a reference genome is used as a standard against which all other sequencing of the same species is compared.

In genome analysis, reference genomes are essential to guarantee reproducibility of results performed by different research groups.

Optional: Read more about a reference genome by by following the WikiPedia link given below the Glossary.

Scaffold

In the context of genomes, a scaffold is a non-contiguous stretch of DNA sequence. A scaffold is formed by linked contigs that are known to be close to each other, based on sequencing information, but which are separated by a gap or stretch of unknown sequence. Although the length of the gap is often known, the sequence is not. The unknown bases are often represented with Ns.