Annotation Field Description

- **Contig_id:**

  Contig Identifier

- **Gen_id:**

  Identifier given in this project to the predicted gene

- **Gene/RNA:**

  It shows whether it is a protein coding gene or an RNA

  start_is_canonical

  If the field displays "true" this gene has a canonical START codon usually found in bacteria

- **start_position**

  Gene start position

- **end_is_canonical**

- If the field displays "true" this gene has a canonical STOP codon usually found in bacteria

- **end_position**

  Gene end position (STOP codon not included)

- **Strand**

  Contig chain where the gene is located

- **Hit_def**

  Name and data of the contig where the gene is located and to which start and end positions make reference

- **Similar_to**

  If we are dealing with CDSs this field contains the Uniprot reference protein Identifier on which this gene annotation is based. If it is RNA this field includes the identifier and data of similar RNA.

- **Protein names**

  Name of the reference protein in Uniprot

- **Organism**

  Organism to which that Uniprot reference protein belongs

- **Comment (FUNCTION)**

  Data related to the function of that Uniprot reference protein

  EC numbers

  Enzyme code corresponding to the Uniprot reference protein, if it has a typified enzymatic function

- **InterPro**

  InterPro motifs present in the Uniprot reference protein

- **Gene Ontology**

  Annotation using Gene Ontology terms corresponding to the Uniprot reference protein

- **Pathway**

  Data concerning metabolic or signalling pathways the Uniprot reference protein is involved with

- **Protein family**

  Protein families to which the Uniprot reference protein belongs

- **Keywords**

  Functional keywords annotating the Uniprot reference protein

Length

Length of the Uniprot reference protein

Subcellular locations

Subcellular location of the Uniprot reference protein

- **PubMed ID**

  Identifiers corresponding to Pubmed references connected with the Uniprot reference protein

Intragenic_stops

Stop codons detected inside the gene. These stops may be really there, if we are dealing with a pseudogene, or may be due to errors inherent to sequencing technology

- **Frameshifts**

  Reading-frame changes detected inside the gene. Frameshifts may be really there, if we are dealing with a pseudogene or genes like transposases or another type of proteins that use phase variation mechanisms to regulate their expression, or may be due to errors inherent to sequencing technology

- **Gene status**

  The content of this field comes from analyzing overlapping and selecting those genes more similar to already known proteins when overlapping is more than 102 bases between predicted genes. The possible cases are the following:

  - **Selected**: Genes with no overlapping or genes chosen as the best in case of overlapping stretch greater than 102 bases.
  - **Selected_minor_threshold**: Genes with an overlapping with other genes lesser than 102 bases
  - **Dismissed**: Genes dismissed due to overlapping more than 102 bases with other predicted genes that get a better score in similarity to known proteins.

- **gene_dismissed_by**

"Dismissed" genes include the id of the gene responsible for its dismission

- **Evalue**

  Value related to the degree of similarity.The lesser the E the more similarity with the reference protein or RNA.

- **Nucleotide sequence**

  Nucleotide sequence of the predicted gene

- **Aminoacid sequence**

  Aminoacid sequence of the protein encoded by that gene. The sequence is only included in the genes:

  - With canonical start
  - With canonical end
  - With no intragenic stop codons
  - With no intragenic frame-shitft

- **Canonical CDS**

  This field will display "canonical CDS" when the protein coding gene has:

  - Canonical start
  - Canonical end
  - No intragenic stop codones
  - No intragenic frameshifts

In case of a gene that do not fulfill these requirements it will display "non-canonical CDS".

In case of RNA the field will be empty.