



# **Introduction to Genom Annotation**

Desi Arti

NPM 2206130706

# GENOM ANNOTATION

Genome annotation is the process of identifying and labeling the various functional elements within a DNA sequence, typically referring to the genes and their associated features in a genome. It is a fundamental step in genomics and is crucial for understanding the genetic information encoded in an organism's DNA

# Structural ANNOTATION

- 1. Gene Structure:** Identifying the boundaries and components of protein-coding genes, including exons (coding regions) and introns (non-coding regions), as well as the locations of start codons, stop codons, and splice sites.
- 2. Non-Coding RNA:** Identifying and characterizing non-coding RNA genes, such as transfer RNA (tRNA), ribosomal RNA (rRNA), and small nuclear RNA (snRNA) genes.
- 3. Promoters and Enhancers:** Locating regions in the genome that control gene expression, such as promoters (regions where RNA polymerase binds) and enhancers (regions that enhance transcription).
- 4. Transposable Elements:** Identifying repetitive DNA sequences, transposons, and retrotransposons that can move within the genome and impact gene regulation and structure.
- 5. Splice Sites:** Marking the locations where pre-mRNA molecules are processed, including splice donor and acceptor sites that enable intron removal during gene expression.
- 6. Open Reading Frames (ORFs):** Determining the locations and lengths of potential protein-coding sequences within the genome.
- 7. Protein Domains:** Identifying functional protein domains and motifs within protein-coding regions, which can provide insights into protein function.

# Structural Annotation

## Types of elements:

- Structural annotation is a crucial step in genome analysis, as it provides a detailed understanding of the genetic elements and their organization within an organism's genome.

Here type of elements Structural Annotation:

- genes
- regulatory regions
- ncRNA
- repeat elements
- pseudogenes and paralogs

# Type of element

- 1. Genes:** Genes are segments of DNA that contain the instructions for producing specific proteins or functional RNA molecules. They are the basic units of heredity and play a critical role in determining an organism's traits and functions.
- 2. Regulatory Regions:** Regulatory regions are non-coding segments of DNA that control gene expression. They include promoter regions, enhancers, and silencers, which influence when and to what extent a gene is transcribed and translated into a protein.
- 3. ncRNA (Non-Coding RNA):** Non-coding RNA refers to RNA molecules that do not code for proteins but have important regulatory and functional roles within the cell. Examples include transfer RNA (tRNA), ribosomal RNA (rRNA), small interfering RNA (siRNA), and microRNA (miRNA).
- 4. Repeat Elements:** Repeat elements are sequences of DNA that are repeated within the genome. These include transposable elements, such as retrotransposons and DNA transposons, which can replicate and move within the genome. Repeat elements can impact genome structure and gene regulation.
- 5. Pseudogenes:** Pseudogenes are non-functional copies of genes that have lost their protein-coding or functional RNA capabilities. They may have originated from functional genes but have accumulated mutations that render them non-functional.
- 6. Paralogs:** Paralogs are genes within the same organism that have evolved from a common ancestral gene through gene duplication. They often have similar sequences and may have diverged in function, contributing to genetic diversity and adaptation.

# Functional Annotation

- Functional Annotation is the process of attaching meta-data such as gene ontology terms to structural annotations
  - . It assigns functions to the elements identified in structural annotation
- .

# Gene Annotation In Galaxy

- Gene annotation in Galaxy, a popular bioinformatics platform, involves a series of steps to identify and annotate genes in a genome. Here's a simplified pipeline to perform gene annotation in Galaxy:
- **1. Data Upload:**

Upload the genome sequence you want to annotate in the Galaxy platform. This could be in FASTA format.
- **2. Quality Control (Optional)**

If necessary, perform quality control on the genome data to ensure it's of high quality and free from errors.
- **3. Gene Prediction**

Use gene prediction tools to identify potential genes in the genome. Popular tools include Augustus, GeneMark, and Glimmer. These tools can be found in the Galaxy Tool Shed.
- **4. Functional Annotation**

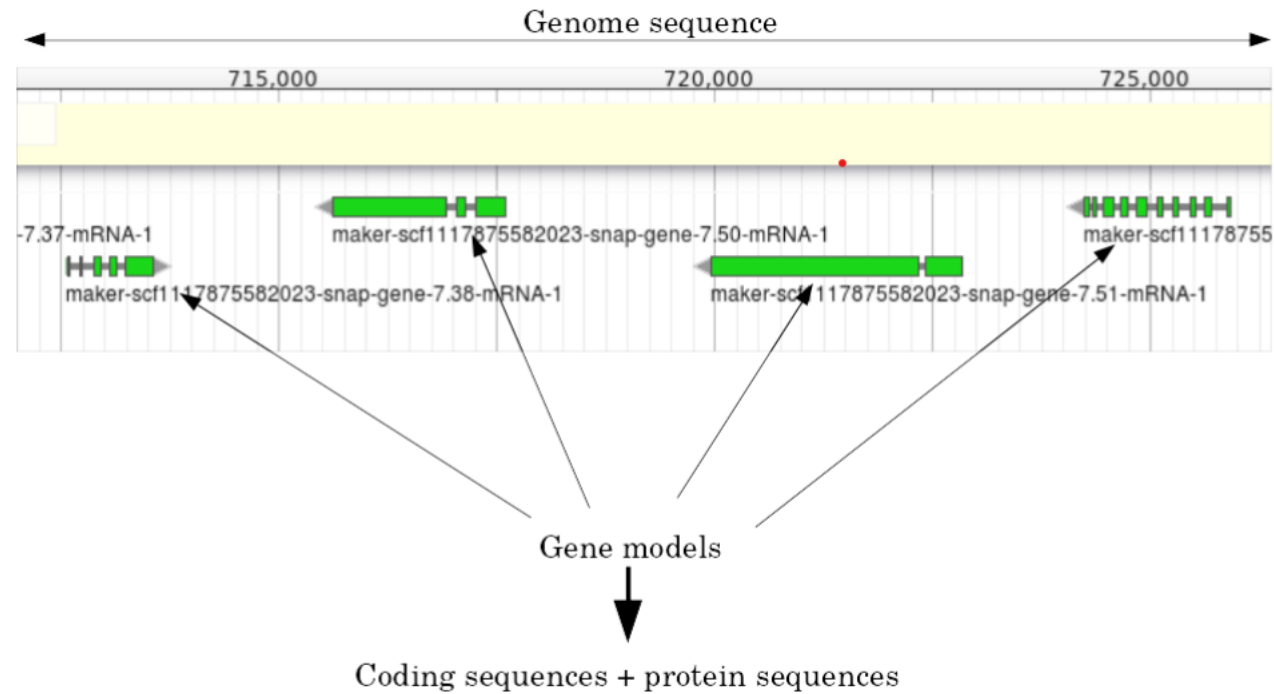
Annotate the predicted genes with functional information. This can be done using tools like InterProScan, BLAST, or HMMER to assign putative functions based on sequence similarity and domain analysis.
- **5. Non-Coding RNA Prediction:**
- Identify non-coding RNA genes, such as tRNAs, rRNAs, and small RNAs, using tools like tRNAscan-SE, Infernal, or other specialized non-coding RNA prediction tools available in Galaxy.

- **6. Repeat Element Identification:**
- Detect repeat elements and transposable elements in the genome using tools like RepeatMasker or RepeatModeler, which can help identify repetitive sequences.
- **7. Structural Annotation:**

Annotate the structural elements of genes, including exons, introns, and splice sites. This can be achieved using tools like Exonerate or Exonerate Transcriptome-to-Genome.
- **8. Combine and Visualize Annotations:**
- Consolidate the various annotations into a comprehensive gene annotation file. Use tools like GFF/GTF merging tools to combine the results.
- **9. Quality Control (Optional):**
- Perform a final quality control check to ensure that the annotations are accurate and coherent.
- **10. Visualization and Reporting:**
- Visualize the gene annotations on a genome browser in Galaxy or export the annotated data for further analysis and interpretation.
- **11. Documentation:**
- Properly document the annotation results, including the parameters and tools used, and save the annotation files for future reference.
- This pipeline provides a general framework for gene annotation in Galaxy. The specific tools and steps to use may vary depending on the organism, the quality of data, and research objectives. You can find and install relevant tools from the Galaxy Tool Shed, and the Galaxy platform provides a user-friendly interface to set up and execute workflows like this.



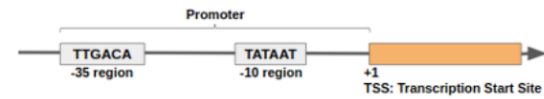
# Structural Annotation



Promoter:

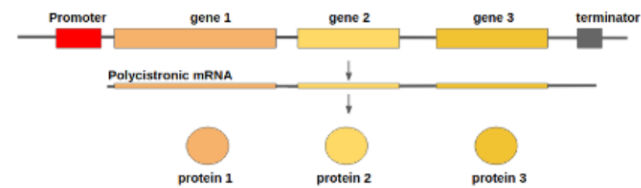
- -35 Region
- TATA Box
- Initiation site (TSS)

## Prokaryotic Genes

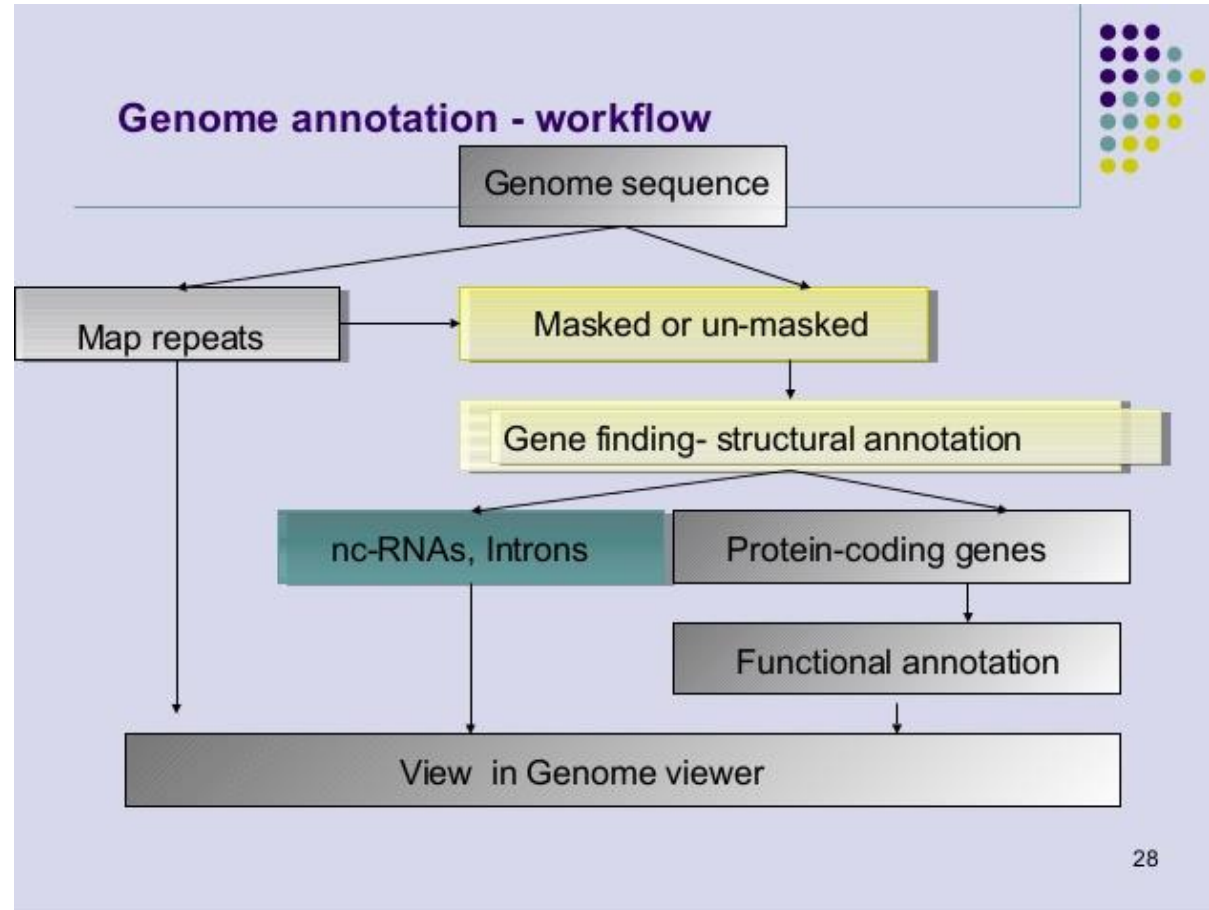


Operons:

- Promoter
- Some genes
- A terminator



# Genom annotation work flow



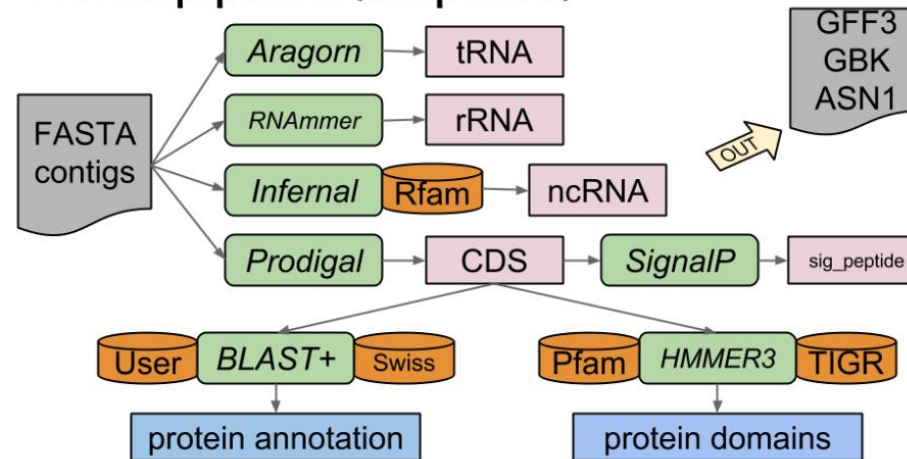
Source : City U Bioinformatics

# Genom annotation with Prokka

Prokka is a useful tool to annotate a bacterial genome

Prokka

## Prokka pipeline (simplified)



# •Step 1 Load Genome into Galaxy

The screenshot displays the Galaxy web interface with the JBrowse genome browser tool selected. The interface is divided into three main sections: Tools, Tool Parameters, and History.

**Tools:** The left sidebar shows the 'JBrowse genome browser' tool. Below the tool name, it states 'JBrowse - Data Directory to Standalone upgrades the bare data directory to a full JBrowse instance'. Under the 'WORKFLOWS' section, it lists 'All workflows'.

**Tool Parameters:** The central panel shows the configuration for the 'JBrowse genome browser (Galaxy Version 1.16.11+galaxy1)'. The 'Reference genome to display' section has a dropdown menu set to 'Use a genome from history'. Below this, the 'Select the reference genome \*' section shows a dropdown menu set to '7: Prokka on data 1: fna'. The 'Output JBrowse \*' section has a dropdown menu set to 'Minimal for viewing (Documentation removed)'. The 'Genetic Code \*' section has a dropdown menu set to '1. The Standard Code'. At the bottom, the 'JBrowse-in-Galaxy Action' section has a dropdown menu set to 'New JBrowse Instance'.

**History:** The right sidebar shows the 'History' section with a search bar. Below the search bar, it lists 'Genom Annotation'. The '2.44 MB' size is displayed. The 'History' section shows a list of datasets, including '10: Prokka on data 1: sqn', '9: Prokka on data 1: ffn', '8: Prokka on data 1: faa', '7: Prokka on data 1: fna', '6: Prokka on data 1: gbk', '5: Prokka on data 1: gff', '4: contigs.fasta', and '1: sequence (6).fasta'.

# Step 2 Genom Annotation with Prokka

The screenshot displays the Galaxy web interface during the execution of the Prokka tool. The top navigation bar includes the Galaxy logo, a home icon, and menu items for Workflow, Visualize, Shared Data, Help, and User. A status indicator on the right shows 'Using 0%'. The left sidebar contains a 'Tools' section with a search bar containing 'prokka', an 'Upload Data' button, and a 'Show Sections' button. Below this, the 'Prokka' tool is listed with the description 'Prokaryotic genome annotation'. The 'WORKFLOWS' section is also visible, showing 'All workflows'. The main content area features a green notification box with a checkmark icon, stating: 'Started tool **Prokka** and successfully added 1 job to the queue. It produces 12 outputs:'. A bulleted list follows, detailing the outputs: 5: Prokka on data 1: gff, 6: Prokka on data 1: gbk, 7: Prokka on data 1: fna, 8: Prokka on data 1: faa, 9: Prokka on data 1: ffn, 10: Prokka on data 1: sqn, 11: Prokka on data 1: fsa, 12: Prokka on data 1: tbl, 13: Prokka on data 1: tsv, 14: Prokka on data 1: err, 15: Prokka on data 1: txt, and 16: Prokka on data 1: log. Below the list, a paragraph explains that the status of queued jobs can be checked in the History panel, which will change from 'running' to 'finished' upon completion or 'error' if problems occurred. The right sidebar shows the 'History' section with a search bar and a list of datasets. The 'Genom Annotation' section is highlighted, showing a list of datasets including '10: Prokka on data 1: sqn', '9: Prokka on data 1: ffn', '8: Prokka on data 1: faa', '7: Prokka on data 1: fna', '6: Prokka on data 1: gbk', '5: Prokka on data 1: gff', '4: contigs.fasta', and '1: sequence (6).fasta'. Each dataset entry includes icons for viewing, editing, and deleting.

**Galaxy**

Workflow Visualize Shared Data Help User

Using 0%

Tools

prokka

Upload Data

Show Sections

**Prokka** Prokaryotic genome annotation

**WORKFLOWS**

All workflows

Started tool **Prokka** and successfully added 1 job to the queue.

It produces 12 outputs:

- 5: Prokka on data 1: gff
- 6: Prokka on data 1: gbk
- 7: Prokka on data 1: fna
- 8: Prokka on data 1: faa
- 9: Prokka on data 1: ffn
- 10: Prokka on data 1: sqn
- 11: Prokka on data 1: fsa
- 12: Prokka on data 1: tbl
- 13: Prokka on data 1: tsv
- 14: Prokka on data 1: err
- 15: Prokka on data 1: txt
- 16: Prokka on data 1: log

You can check the status of queued jobs and view the resulting data by refreshing the History panel. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

History

search datasets

Genom Annotation

383 kB 14 2

10: Prokka on data 1: sqn

9: Prokka on data 1: ffn

8: Prokka on data 1: faa

7: Prokka on data 1: fna

6: Prokka on data 1: gbk

5: Prokka on data 1: gff

4: contigs.fasta

1: sequence (6).fasta

# Step 3 View Annotation with Jw Browse

The screenshot displays the Galaxy web interface at the URL `usegalaxy.org/jobs/submission/success`. The interface is divided into three main sections: Tools, a central message area, and History.

**Tools Panel (Left):** Contains a search bar with "jw browse", an "Upload Data" button, a "Show Sections" button, and a link "Did you mean: jbrowse?". Below this, it lists "JBrowse genome browser" and "JBrowse - Data Directory to Standalone" which upgrades the bare data directory to a full JBrowse instance. A "WORKFLOWS" section is also visible.

**Central Message Area:** A green box with a checkmark icon contains the following text:  
Started tool **JBrowse** and successfully added 1 job to the queue.  
It produces this output:  
• 17: JBrowse on - minimal  
You can check the status of queued jobs and view the resulting data by refreshing the History panel. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

**History Panel (Right):** Shows a list of jobs. The top job is "17: JBrowse on - minimal" with a status of "finished". Below it are several other jobs, including "16: Prokka on data 1: log", "15: Prokka on data 1: txt", "14: Prokka on data 1: err", "13: Prokka on data 1: tsv", "12: Prokka on data 1: tbl", "11: Prokka on data 1: fsa", and "10: Prokka on data 1: sqn". Each job entry includes icons for viewing, editing, and deleting.

# Result

[illegible]



# Conclusion

- Genome annotation with Prokka in Galaxy offers a powerful and user-friendly solution for researchers working with prokaryotic genomes. Prokka simplifies the annotation process by automating various steps, making it accessible to both beginners and experienced bioinformaticians. When executed in the Galaxy platform, it combines the strengths of Prokka with the convenience and versatility of Galaxy's interface and workflow management capabilities.

# Refences

- <https://training.galaxyproject.org/training-material/topics/genome-annotation/tutorials/introduction/slides.html#34>
- <https://training.galaxyproject.org/training-material/topics/genome-annotation/tutorials/annotation-with-prokka/slides.html#11>