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.

Fungtional 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**

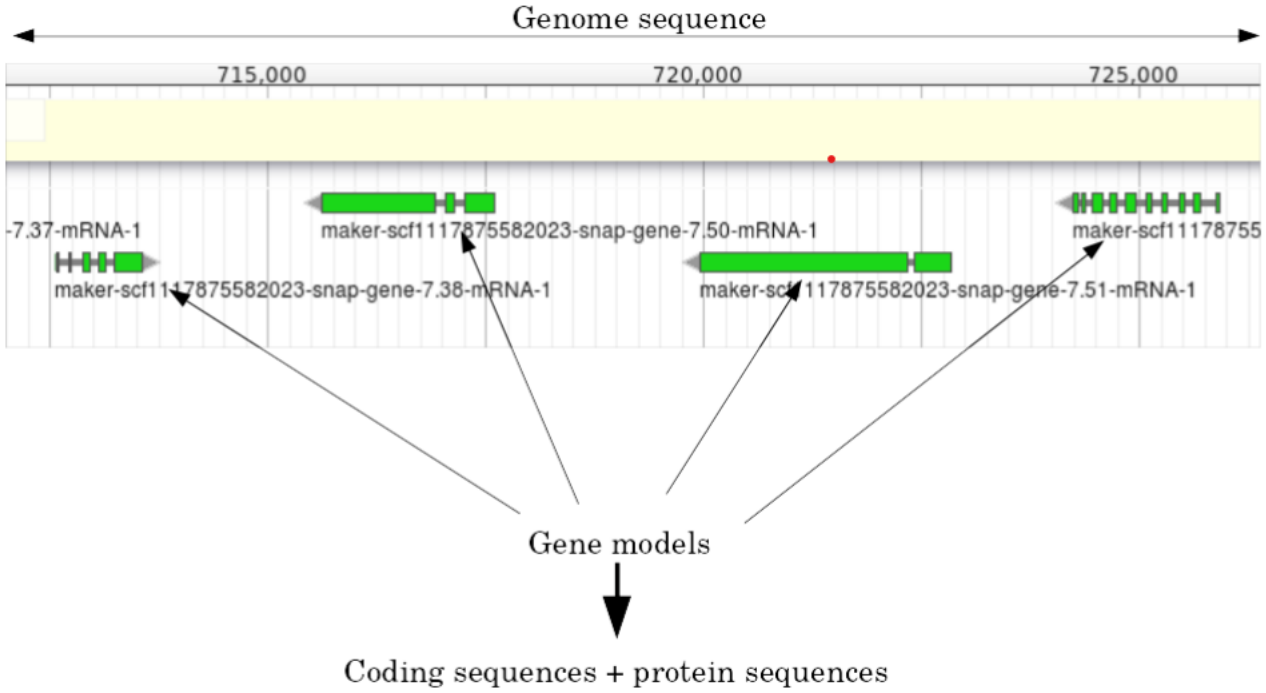
  Use gene prediction tools to identify potential genes in the genome. Popular tools include Augustus, GeneMark, and Glimmer. These tools can be found in the Galaxy Tool Shed.
- 4. **Functional Annotation**

  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.

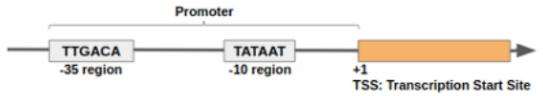
Structural Annotation



Prokaryotic Genes

Promoter:

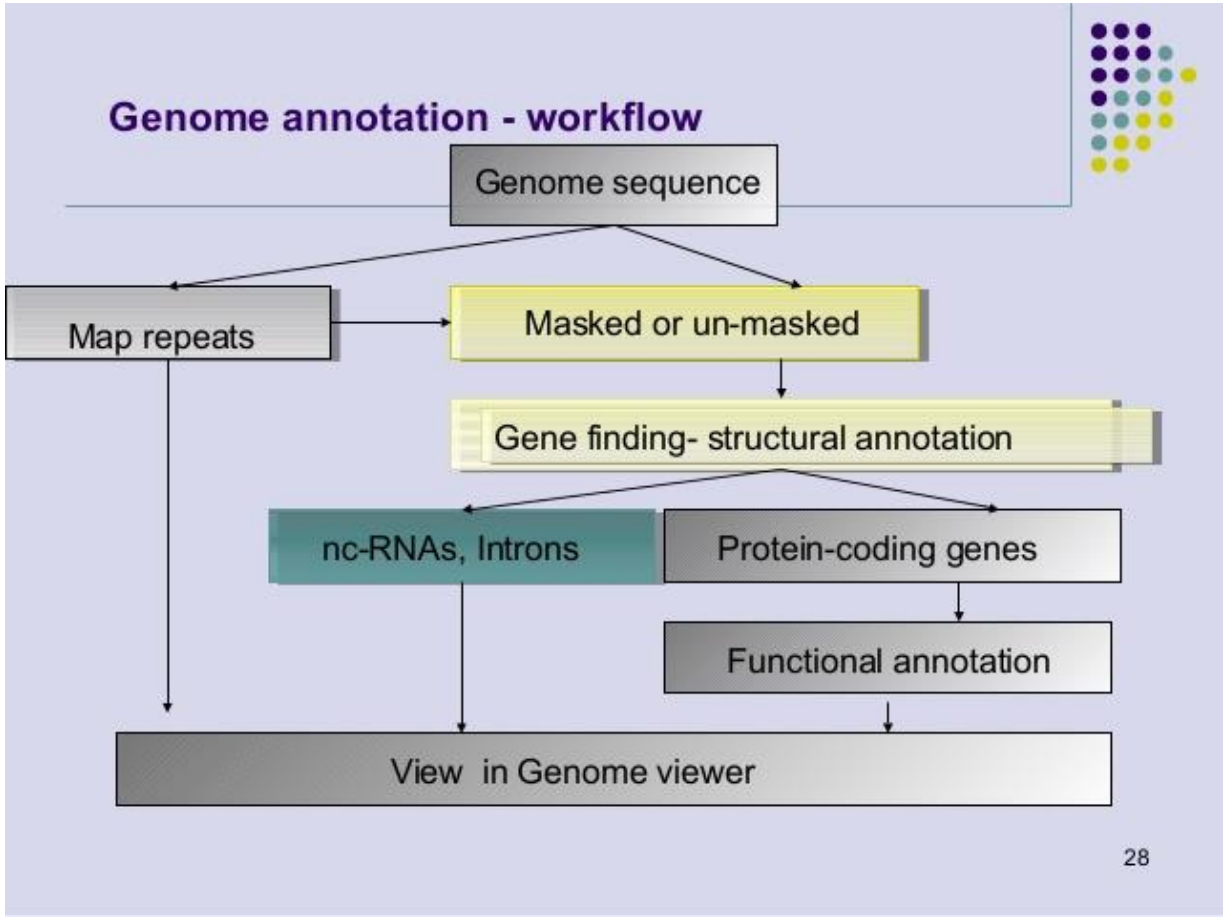
- -35 Region
- TATA Box
- Initiation site (TSS)



Operons:

- Promoter
- Some genes
- A terminator

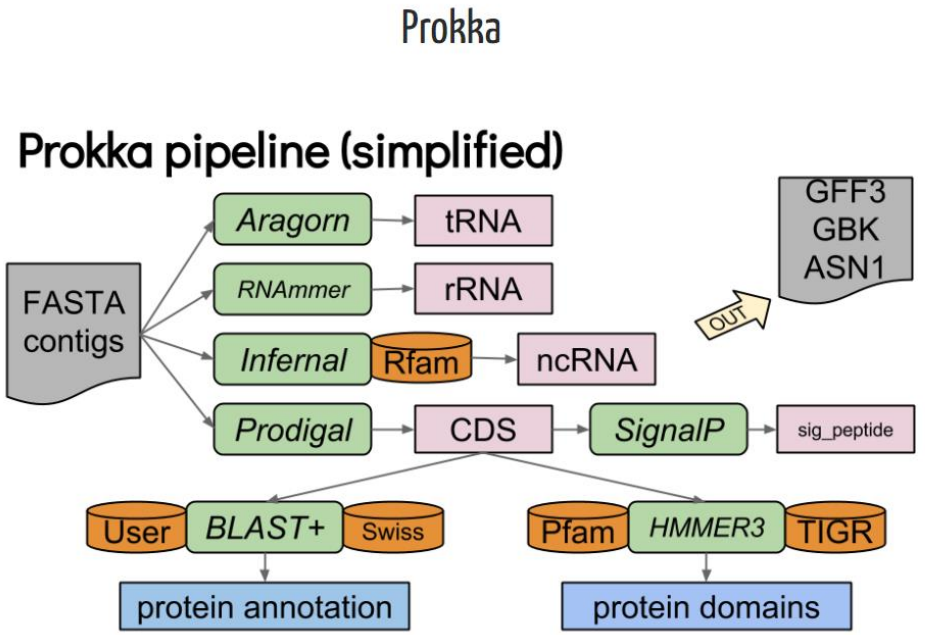
Gennom annotation work flow



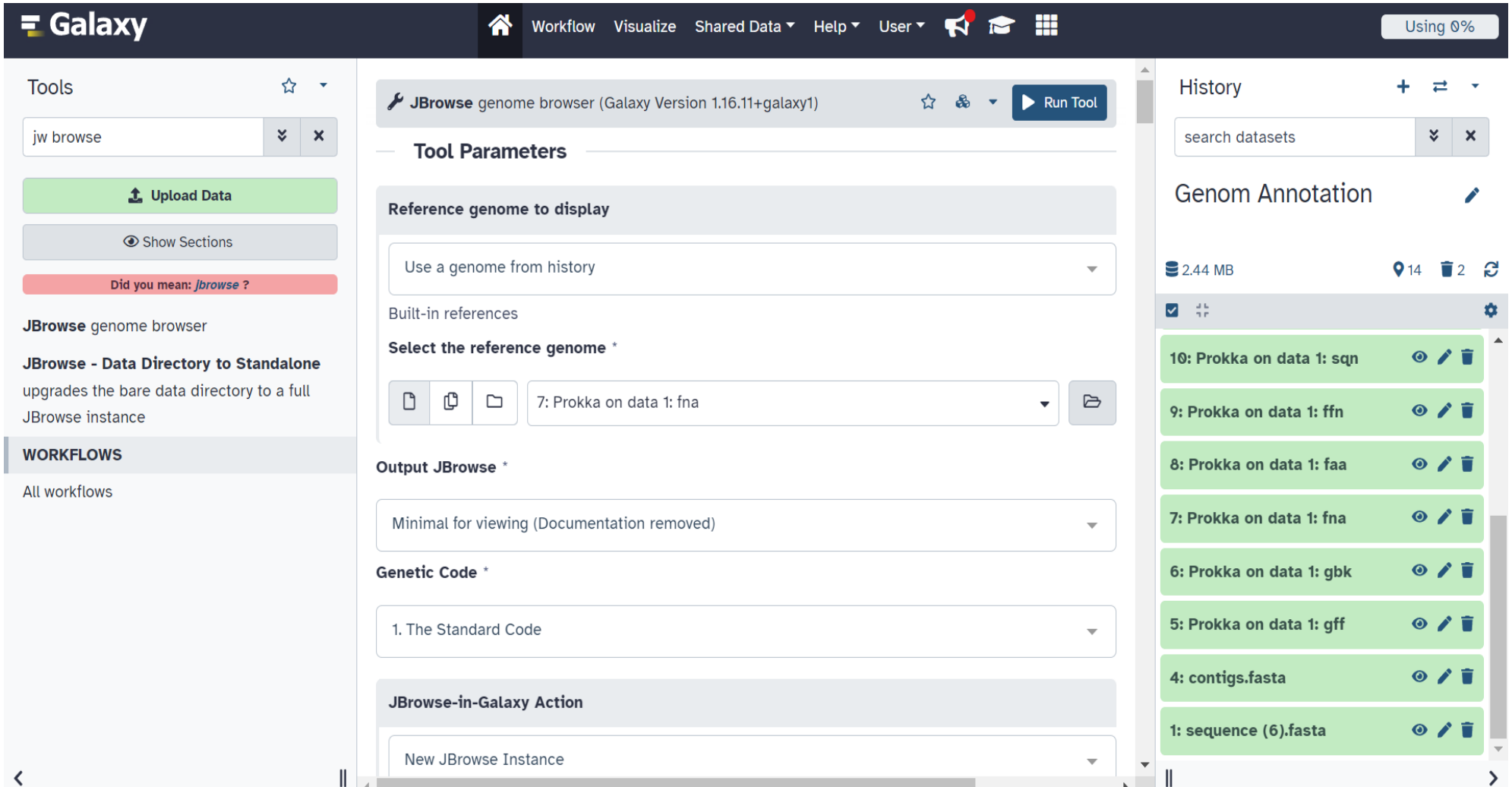
Source : City U Bioinformatics

Genom annotation with Prokka

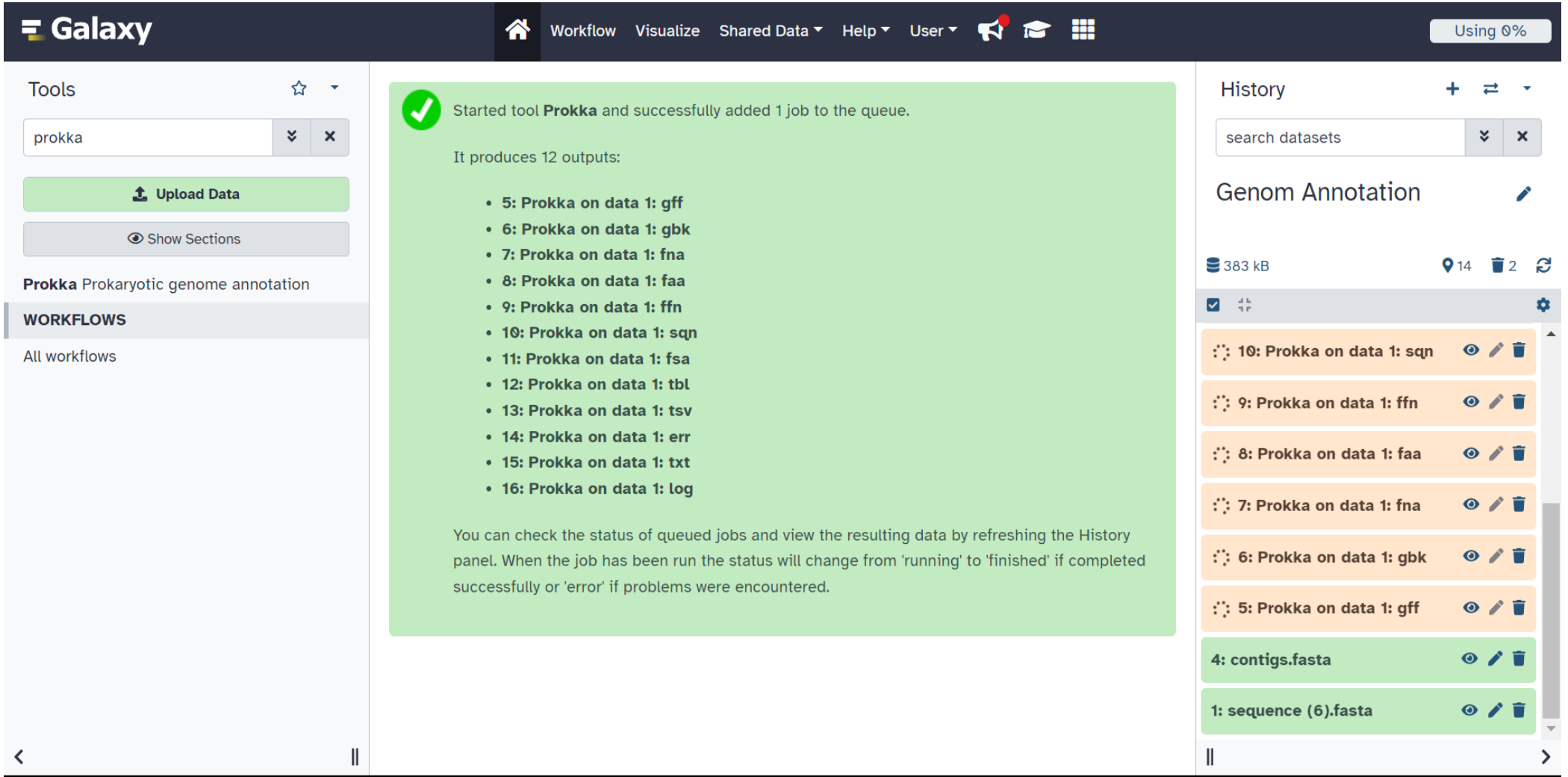
Prokka is a useful tool to annotate a bacterial genome



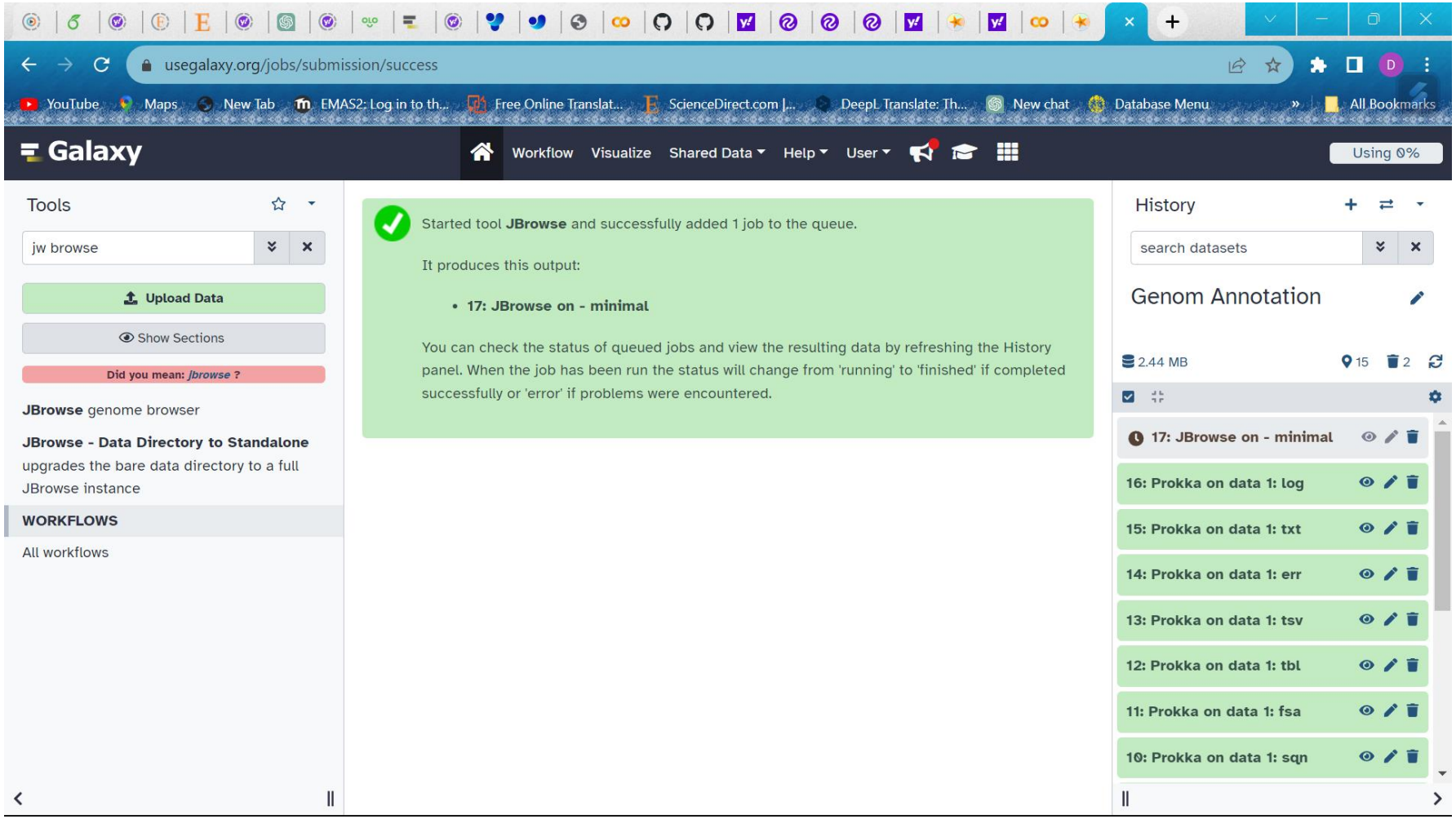
•Step 1 Load Genome into Galaxy



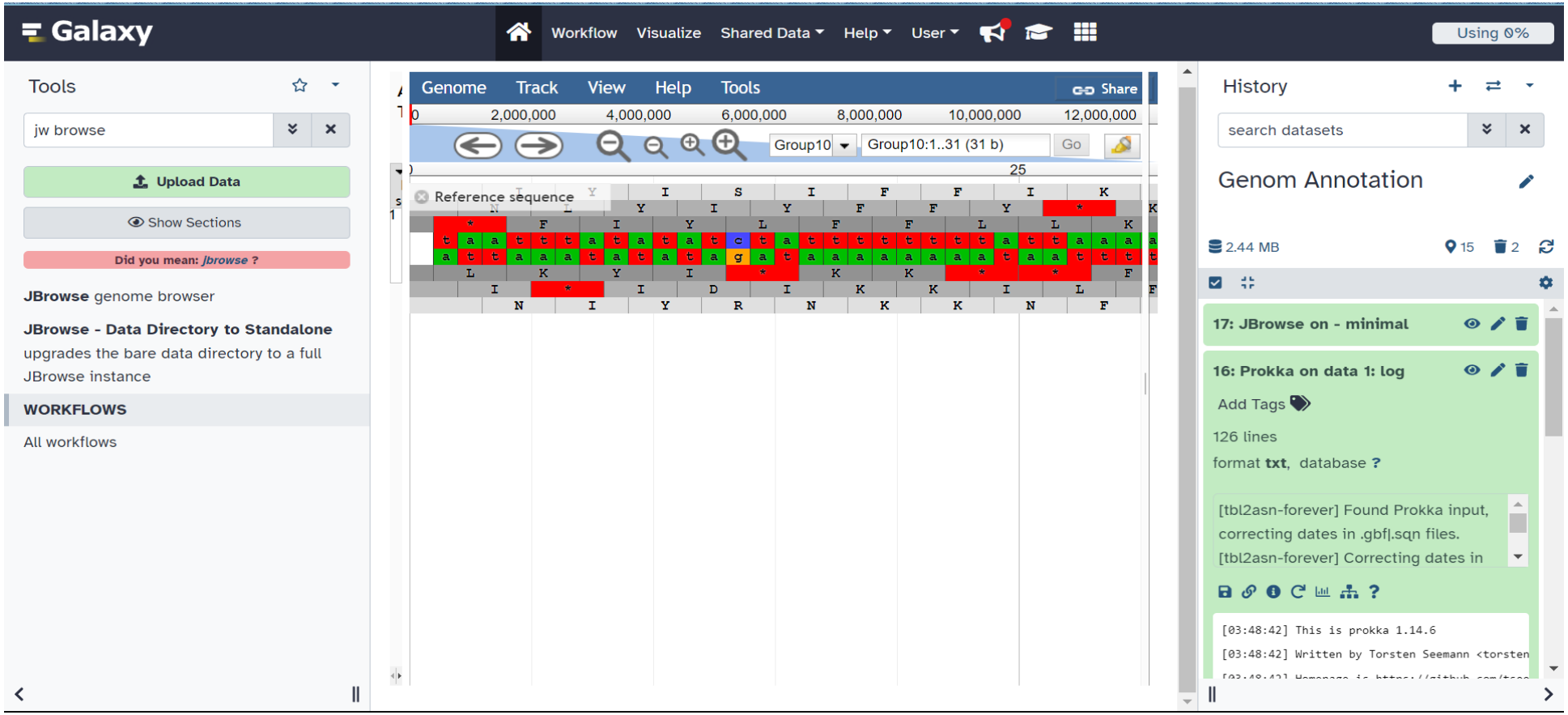
Step 2 Genom Annotation with Prokka



Step 3 View Annotation with Jw Browse



Result


Galaxy
Workflow
Visualize
Shared Data
Help
User
Using 0%
Tools
jw browse
Upload Data
Show Sections
Did you mean: jbrowse ?
JBrowse genome browser
JBrowse - Data Directory to Standalone
upgrades the bare data directory to a full JBrowse instance
WORKFLOWS
All workflows
Genome
Track
View
Help
Tools
Share
Group10
Group10:1..31 (31 b)
Go
Reference sequence
History
search datasets
Genom Annotation
2.44 MB
15
2
17: JBrowse on - minimal
16: Prokka on data 1: log
Add Tags
126 lines
format txt, database ?
[tbl2asn-forever] Found Prokka input, correcting dates in .gbf|.sqn files.
[tbl2asn-forever] Correcting dates in
[03:48:42] This is prokka 1.14.6
[03:48:42] Written by Torsten Seemann <torsten

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