Lecture 6: Genome Annotation
Student Handout & In-Class Exercises

Course: BINF301 — Computational Biology
Instructor: Tom Michoel
Date: 2/2/2026
Created with Copilot

1 What is Genome Annotation? (Slides 2–3)

A genome is a long nucleotide sequence; annotation aims to identify:

- **Structural elements**: gene locations, repeats, ncRNAs, tRNAs, rRNAs, regulatory regions.
- **Functional elements**: functions assigned to predicted genes and RNAs.

Solution. Emphasize: annotation has two parts—finding features and assigning meaning. Make clear to students that much of a genome is *noncoding* and must still be annotated (regulatory sites, ncRNAs, repeats).

2 Annotation Workflow (Slide 3)

A general annotation workflow includes:

1. Repeat masking
2. Gene prediction (ab initio, extrinsic, or combined)
3. Prediction of additional functional elements (ncRNA, tRNA, rRNA)
4. Functional annotation (domains, homology)

Solution. Clarify that these steps occur in both prokaryotic and eukaryotic annotation, but with major differences in complexity.

3 Repeat Masking (Slides 5–8)

Many eukaryotic genomes contain **25–50%** repeats. Masking repeats prevents false gene predictions and reduces the candidate search space.

3.1 Soft vs. Hard Masking (Slide 7)

Soft masking: bases converted to lowercase. **Hard masking:** repeat regions replaced with `N`.

Tools:

- RepeatModeler + RepeatMasker (de novo + masking; slow)
- RED (fast, no classification)

Solution. Make students aware that different gene-finders expect one masking type or the other. Soft masking retains information and is preferred when splice-aware aligners are used.

4 Prokaryotic Genome Annotation (Slides 10–14)

Prokaryotic genomes are simpler due to:

- No introns
- High gene density

4.1 Prokka Pipeline (Slide 11)

Includes: Prodigal (CDS), RNAmmer (rRNA), Aragorn (tRNA), SignalP (signal peptides), Infernal (ncRNAs).

4.2 Prodigal (Slides 12–13)

Uses:

- ORF discovery with GC-bias scoring
- Dynamic programming to select best non-overlapping ORFs
- Hexamer frequencies to refine predictions
- RBS detection to refine start sites

Solution. Focus on: prokaryotic gene finding is essentially a classification of ORFs using statistics; no splice-site modeling needed.

5 Noncoding RNA Detection (Slides 15–18)

Tools:

- RNAmmer (rRNA; HMMs)
- tRNAscan-SE (tRNA; covariance models)
- Infernal (general ncRNAs; covariance models)

Solution. Stress that RNA structure is essential—covariance models capture paired bases and secondary structure not available to simple HMMs.

6 Eukaryotic Gene Prediction (Slides 20–24)

Eukaryotic gene prediction is more complex due to introns, exon variation, UTRs, alternative splicing, and long intergenic regions.

6.1 Ab Initio Prediction (Slides 21–24)

HMM-based models include states for:

- Initial, internal, and terminal exons
- Introns (with splice sites)
- Intergenic regions
- Single-exon genes

Tools: **GeneMark**, **Augustus**.

Solution. Explain training: unsupervised (GeneMark) vs. supervised (Augustus). Important: proper training greatly increases accuracy.

7 Extrinsic Evidence (Slides 25–27)

Sources:

- RNA-seq (expression profiles, intron/exon boundaries)
- Protein homology

Integrated pipelines:

- BRAKER2 / BRAKER3
- TSEBRA

Solution. Correction: RNA-seq alone is insufficient for accurate annotation. Require splice-aware alignment + integration with ab initio predictions.

8 Functional Annotation (Slides 32–39)

Approaches:

- Domain-based: Pfam, CDD
- Homology-based: Reciprocal Best Hit (RBH), BLAST
- Combined systems: InterProScan, eggNOG mapper

Solution. Caution: domain presence does not guarantee exact function; annotations are probabilistic.

9 Annotation Quality Assessment (Slides 41–52)

9.1 BUSCO (Slides 41–43)

Uses lineage-specific universal single-copy orthologs to evaluate:

- Completeness (single-copy, duplicated)
- Missing genes

9.2 OMArk (Slides 45–52)

Evaluates:

- Proteome completeness
- Phylogenetic consistency
- Contamination

Uses OMA gene families and k-mer–based mapping (OMAmer).

Solution. Distinguish roles: BUSCO tests “is most of the conserved gene set present?”, OMArk tests “is the annotation evolutionarily consistent?”.

10 Exercises

10.1 Exercise 1: Why Mask Repeats? (Slides 5–8)

Question: Why must repeats be masked before annotation?

Solution. Repeats cause spurious ORFs and inflate false-positive gene calls; masking reduces errors and computation.

10.2 Exercise 2: Soft vs Hard Masking (Slide 7)

Sequence: `ACGTCGGatatatatatCGATGA`

Question: Which masking type is this? What would the other type look like?

Solution. Lowercase = soft-masked. Hard-masked form: `ACGTCGGNNNNNNNNNNCGATGA`.

10.3 Exercise 3: Prokaryotic Simplicity (Slides 10–11)

Question: List two reasons prokaryotic annotation is easier.

Solution. No introns; high gene density; straightforward ORFs.

10.4 Exercise 4: Ab Initio vs Extrinsic (Slides 21–27)

Question: How do ab initio and extrinsic prediction differ?

Solution. Ab initio uses only genomic sequence/HMMs; extrinsic uses RNA-seq/protein evidence to guide boundaries.

10.5 Exercise 5: Functional Annotation (Slides 32–39)

Question: Name two functional annotation strategies.

Solution. Domain detection (Pfam/CDD) and homology-based inference (BLAST/RBH).

10.6 Exercise 6: BUSCO Interpretation (Slides 41–43)

Question: What does 70% BUSCO completeness imply?

Solution. Many conserved genes missing or fragmented → incomplete assembly/annotation.

10.7 Exercise 7: Tool Matching

Match tools to functions:

a. RNAmmer
b. tRNAscan-SE
c. RepeatMasker
d. InterProScan

Purposes: 1. Protein domain detection 2. rRNA prediction 3. Repeat masking 4. tRNA prediction

Solution. a–2, b–4, c–3, d–1.