Lecture 8: Functional and Comparative Genomics

Student Handout & In-Class Exercises

Course: BINF301 — Computational Biology
Instructor: Tom Michoel
Date: 9/2/2026
Created with Copilot

1 What is Genomic ”Function”? (Slides 3–5)

There is no single agreed-upon definition of “function” in genomic sequences. Lecture 8 describes three complementary frameworks:

1.1 Genetic Approach

- Function inferred from **loss-of-function phenotypes** after mutation or interference.
- Limited by redundancy and subtle or context-specific phenotypes.

1.2 Evolutionary Approach

- **Comparative genomics**: conserved sequences often under purifying selection.
- Works well for protein-coding regions; regulatory regions evolve rapidly.

1.3 Biochemical Approach (ENCODE)

- Biochemical assays (ChIP-seq, DNase-seq, ATAC-seq, etc.) identify active or bound regions.
- Biochemical signal does not always imply causal function.

Solution. Important teaching point: students should recognize that “function” varies by experimental approach—each approach captures a different biological concept (phenotype, constraint, biochemical activity).

2 The ENCODE Project (Slides 7–12)

ENCODE (Encyclopedia of DNA Elements) aims to catalog:

- Coding genes
- Non-coding RNAs
- Regulatory elements
- Epigenetic marks and chromatin features

Data are available via the ENCODE portal and the UCSC Genome Browser.

Solution. ENCODE provides multi-assay functional data (RNA-seq, ChIP-seq, ATAC-seq, etc.). Students should understand differences in the biological interpretations of these data types.

3 Functional Genomics Methods (Slides 14–22)

3.1 ChIP-seq (Slide 15)

Chromatin Immunoprecipitation + sequencing detects **DNA-protein interactions**.

3.2 ATAC-seq (Slide 16)

Assesses **chromatin accessibility** using transposase accessibility.

3.3 DNase-seq (Slide 17)

Detects DNase I hypersensitive sites → open chromatin.

3.4 FAIRE-seq (Slide 18)

Enriches nucleosome-depleted, regulatory regions.

3.5 RNA-seq (Slide 19)

Measures transcription, gene expression, isoforms, and non-coding RNAs.

3.6 Ribo-seq (Slide 20)

Profiles ribosome-protected RNA fragments → **translation activity**.

3.7 DNA Methylation Assays (Slide 21)

Bisulfite conversion distinguishes methylated vs. unmethylated cytosines.

3.8 Functional Assays (Slide 22)

CRISPR knockouts, RNAi, overexpression → link genomic regions to phenotypes.

Solution. Highlight differences between assays:

- Accessibility (ATAC/FAIRE/DNase) vs binding (ChIP) vs transcription (RNA-seq) vs translation (Ribo-seq).
- Students often conflate RNA-seq with “functional” evidence; clarify it measures expression, not biological effect.

4 Comparative Genomics (Slides 24–41)

4.1 Whole-Genome Alignment (Slides 25–28)

- **Mauve**: multi-genome alignment with rearrangement awareness.
- **MUMmer**: fast pairwise alignment using maximal unique matches.

4.2 Long-Read Genome Alignment (Slides 31–32)

Minimap2 uses a seed–chain–align procedure with minimizers for fast large-scale alignment.

4.3 Variant Detection (Slides 33–34)

Read-to-reference alignment identifies SNPs and indels (e.g. GATK).

4.4 Orthology and Gene Comparisons (Slides 36–41)

- Reciprocal Best Hit (RBH) to detect orthologs.
- **Orthofinder**: finds orthogroups, builds gene trees, infers species trees, identifies duplications/losses.

Solution. Comparative genomics links molecular biology to evolution:

- Conservation → functionality
- Breakpoints → structural evolution
- Orthogroups → shared ancestry vs lineage-specific changes

5 Exercises (with Solutions)

5.1 Exercise 1 — Three Definitions of Function

Give one advantage and one limitation for each approach: genetic, evolutionary, biochemical.

Solution. **Genetic:** Advantage: strong causal inference. Limitation: redundancy hides phenotypes; low throughput.

Evolutionary: Advantage: identifies constrained regions genome-wide. Limitation: fails for rapidly evolving regulatory regions.

Biochemical: Advantage: scalable; maps many element types. Limitation: biochemical signal may reflect correlation, not causation.

5.2 Exercise 2 — Matching Assays to Functions

Match:

1. ChIP-seq
2. ATAC-seq
3. RNA-seq
4. Ribo-seq
5. Bisulfite sequencing

to:

a. Open chromatin

b. Transcription levels

c. DNA-protein binding

d. Translation activity

e. Cytosine methylation

Solution. 1→c, 2→a, 3→b, 4→d, 5→e.

5.3 Exercise 3 — ENCODE Data Interpretation

Choose one assay (e.g. ChIP-seq, ATAC-seq, RNA-seq). State:

- What biological question it answers
- What type of data output it produces

Solution. Examples: **ATAC-seq:** reveals open chromatin; output = peaks representing accessible regions. **ChIP-seq:** detects protein-DNA binding; output = enrichment peaks. **RNA-seq:** measures expression; output = read counts / transcripts.

5.4 Exercise 4 — Comparative Genomics Tools

For each tool, specify its best use case.

Solution. **Mauve:** multiple small genomes, rearrangement-aware alignment. **MUMmer:** pairwise genome alignment, very fast. **Minimap2:** long-read mapping, assembly-to-assembly alignment.

5.5 Exercise 5 — Orthology Reasoning

A1's best hit is B1, and B1's best hit is A1. A2's best hit is B2, but B2's best hit is A1.

Solution. A1–B1: reciprocal best hit → likely orthologs. A2–B2: not reciprocal → suggests paralogy or gene family expansion; B2 may be more similar to duplicated A1.

5.6 Exercise 6 — Variant Detection

Reads at position: A, A, A, G, A, A.

Is this a SNP?

Solution. Yes: one read shows a G while reference = A. Confidence low: only 1/6 supports the alternative allele; could be sequencing error. True SNP calls typically require higher allele frequency and quality metrics.