

Lecture 5: Repeats

Student Handout & In-Class Exercises

Course: BINF301 — Computational Biology

Instructor: Tom Michoel

Date: 28/01/2026

Created with Copilot

1 The C-value Paradox

C-value: the amount of DNA in a haploid genome (pg). Eukaryotic genome size does *not* correlate with organismal “complexity”; closely related species can have vastly different genome sizes. Much of the size variation is explained by non-coding and repetitive DNA.

2 Types of Repetitive DNA

Repeat classification (non-exhaustive):

- **By location:** interspersed vs. tandem; segmental duplications.
- **By structure:** simple/low-complexity vs. composite elements (e.g., LTR).
- **By autonomy:** autonomous vs. non-autonomous (“hitch-hiking”).
- **By replication mode:** copy-and-paste (retrotransposons) vs. cut-and-paste (DNA transposons).
- **By activity:** active vs. inactivated relics.

3 Simple (Tandem) Repeats

- **Satellites:** short motifs (e.g., TATA...) repeated head-to-tail; common in centromeres.
- **Telomeres:** vertebrate consensus TTAGGG.
- **Mechanism:** often formed by *slipped-strand mispairing* during replication.

4 Transposable Elements (TEs)

TEs are mobile DNA first described by McClintock. Two broad classes:

4.1 Retrotransposons (copy-and-paste)

- **LTR retrotransposons:** share features with retroviruses (but lack viral envelope/capsid).
- **Non-LTR: LINEs** (autonomous; ~17% of human genome, mostly inactive); **SINEs** (non-autonomous; up to ~15%).
- **Mechanism:** RNA intermediate, reverse transcription (e.g., TPRT for LINE-1).

4.2 DNA Transposons (cut-and-paste)

- Typically have **TIRs** (terminal inverted repeats) and generate **TSDs** at insertion.
- Families include Type II transposons, Helitrons; some encode their own transposase; Polintons (Type III) encode polymerase/integrase and have TIRs.

5 Repeats and Genome Evolution

- TE proliferation expands genomes; inactive copies accumulate.
- Repeats can generate structural variation, gene duplication/deletion, new regulatory elements,

and occasionally new genes.

- **Horizontal Transposon Transfer (HTT):** transfer across species via vectors (e.g., parasites/viruses); detection remains challenging.

6 Repeat Detection: Core Tools and Ideas

6.1 TRF (Tandem Repeat Finder)

Sliding-window search for periodic k -mers; identifies candidate regions, checks periodic spacing/consistency, and reports tandem repeat units.

6.2 RECON

Alignment-based *de novo* TE discovery:

1. Compute pairwise alignments; define *images* (aligned subsequences).
2. Cluster images via single-linkage (syntopy rules) to infer *elements*.
3. Build inter-element graph; refine edges (primary/secondary) and detect triangles to avoid false family merges; connected components define *families*.

6.3 RepeatScout

Seed-and-extend on frequent k -mers:

- Start from high-frequency k -mers; greedily extend to maximize a consensus scoring function (with penalties for dangling ends).
- Iteratively extract consensus families and remove their instances from the k -mer table.

6.4 RED (REpeat Detector)

Signal-processing + HMM approach:

- Assign adjusted k -mer frequency scores per base; smooth, find local maxima; delineate candidate repeat vs. non-repeat regions.
- Train an HMM on candidates, then scan genome to label repeats.

6.5 RepeatModeler & RepeatMasker

- **RepeatModeler:** orchestrates multiple *de novo* tools to build a species-specific repeat library.
- **RepeatMasker:** screens a genome with known/*de novo* libraries to mask repeats (often very time-consuming).

7 Exercises

7.1 Exercise 1 — C-Value Paradox Reasoning

Prompt. The lungfish genome is $>100\times$ larger than the human genome. Explain why genome size can increase dramatically without increasing organismal complexity.

7.2 Exercise 2 — Classifying Repeat Types

Prompt. Classify each as *tandem* / *interspersed* / *autonomous* / *non-autonomous*:

1. Vertebrate telomere TTAGGG
2. SINE (e.g., Alu)
3. LINE-1
4. Centromeric satellites

7.3 Exercise 3 — TE Replication Mechanisms

Prompt. Match TE type to mechanism:

- | | |
|-------------------------|----------------------------------------|
| A. DNA transposons | 1. Copy-and-paste (RNA intermediate) |
| B. LINEs | 2. Cut-and-paste (DNA intermediate) |
| C. LTR retrotransposons | 3. Target-primed reverse transcription |

7.4 Exercise 4 — Genome Evolution by TEs

Prompt. Provide one beneficial and one harmful evolutionary consequence of TE activity.

7.5 Exercise 5 — Using TRF

Prompt. For sequence ATGATGATGATGCCCCGTA:

1. Identify the tandem repeat motif.
2. How many copies?

7.6 Exercise 6 — RECON vs. RepeatScout

Prompt. Explain how RECON and RepeatScout differ in discovering repeat families.

7.7 Exercise 7 — TE Inactivation

Prompt. Give two reasons why most LINEs are inactive in humans.

7.8 Exercise 8 — RED in Practice (Conceptual)

Prompt. Why does RED adjust k -mer frequencies and then train an HMM?