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.

Solution. Key messages to stress: (1) Genome size varies over orders of magnitude among eukaryotes. (2) Differences are largely driven by repeat/TE content rather than gene number. (3) Avoid equating genome size with organismal complexity.

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.

Solution. Helpful framing: *Interspersed* repeats are dispersed copies (often TEs), while *tandem* repeats are adjacent copies (e.g., satellites). Autonomy indicates whether the element encodes the machinery needed for its own mobilization.

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.

Solution. Emphasize the distinction between tandem repeats (periodic, adjacent) and interspersed elements (TEs). Simple repeats are important for chromosome structure (centromeres/telomeres), but can also complicate assembly.

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.

Solution. Talking points: (1) **Autonomous vs. non-autonomous:** LINEs carry ORF1/ORF2; SINEs *lack* these and parasitize LINE machinery. (2) **Mechanistic fingerprints:** TIRs/TSDs for DNA transposons; LTRs for LTR retrotransposons; target-primed reverse transcription for LINE-1. (3) **Activity:** In humans, most copies are ancient and inactive, but remnants dominate genome composition.

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.

Solution. Balance the narrative: although many TEs are neutral or deleterious at insertion, over long timescales they contribute to innovation (cis-regulatory rewiring, exon shuffling) and plasticity. HTT is rare in eukaryotes but increasingly documented.

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.

Solution. Students should know that TRF is specialized for tandem repeats (satellites, microsatellites) and not for interspersed TEs.

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*.

Solution. Contrast with TRF: RECON is for interspersed repeats/TEs; handles partial copies/degeneration via graph refinement.

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.

Solution. RepeatScout tends to be faster than all-vs-all alignment (RECON) and is effective when families still share abundant k-mers.

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.

Solution. Key selling points: very fast, self-learning, detects both tandem and interspersed repeats; but does not classify families by type.

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).

Solution. A typical pipeline: *RepeatModeler* to build library → *RepeatMasker* to annotate/mask. Useful for downstream gene prediction and variant calling.

7 Exercises

7.1 Exercise 1 — C-Value Paradox Reasoning

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

Solution. Large genomes typically reflect accumulation of repeats/TEs, polyploidy events, and reduced DNA removal mechanisms, not expanded protein-coding gene sets. Many TE insertions become inactive but persist, inflating genome size while contributing little to gene count or "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

Solution. 1) Telomere TTAGGG: *tandem* (simple repeat). 2) SINE/Alu: *interspersed*, *non-autonomous* (uses LINE machinery). 3) LINE-1: *interspersed*, *autonomous* (encodes ORF1/ORF2). 4) Centromeric satellites: *tandem* repeats.

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

Solution. A→2 (DNA transposons cut-and-paste); B→3 (LINE-1 uses TPRT); C→1 (LTR retrotransposons copy-and-paste via RNA).

7.4 Exercise 4 — Genome Evolution by TEs

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

Solution. **Beneficial:** Creation of novel regulatory elements (TE-derived enhancers/promoters), exon shuffling, or substrate for gene duplication.
Harmful: Disruption of coding/regulatory regions upon insertion; ectopic recombination between repeats causing deletions/rearrangements.

7.5 Exercise 5 — Using TRF

Prompt. For sequence `ATGATGATGATGCCCGTA`:

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

Solution. (a) Motif = `ATG`. (b) There are 4 adjacent copies: `ATG ATG ATG ATG` followed by non-repetitive sequence.

7.6 Exercise 6 — RECON vs. RepeatScout

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

Solution. **RECON:** all-vs-all alignments to cluster *images* → infer *elements* → inter-element graph refinement → families; robust to partial/degenerate copies but computationally heavier.
RepeatScout: frequent k-mer seeding with greedy consensus extension; faster and relies on abundance of short exact words; may struggle if repeats are too diverged to yield frequent k-mers.

7.7 Exercise 7 — TE Inactivation

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

Solution. (1) Accumulation of disabling mutations in ORF1/ORF2 or promoter regions over evolutionary time;
(2) Host defense mechanisms (DNA methylation, small RNAs) suppress TE expression;
(3) Truncation at insertion (common for LINE-1) yields defective copies.

7.8 Exercise 8 — RED in Practice (Conceptual)

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

Solution. Adjusted k-mer scores highlight sequence segments with unexpected repetitiveness beyond base composition; smoothing and local maxima detection define candidate "repeat" vs. "non-repeat" regions. The HMM then generalizes these patterns to label the whole genome efficiently and consistently.