

Design and Use of RepeatMasker

Jeremy Buhler
jrbuhler@wustl.edu

1

Parts of RepeatMasker

■ Programs

- Smit AFA, Hubley R, and Green P. "RepeatMasker-Open 4.0." 2013-2015.
<http://www.repeatmasker.org/>
- RMBlast (NCBI variant), HMMER for comparisons

■ Data

- Dfam <https://www.dfam.org>

2

Overview

- Sources of repetitive sequence data
- How RepeatMasker finds repeats
- Issues and limitations

3

Data Source

- Uses a library of known repeat seqs
- Supplied by Dfam ("DNA families DB")
- Repeat families in Dfam are carefully curated using multiple alignment tools.

4

Example of a repeat family summary page from Dfam

DF000001317.2 [AluY1]
 AluY1 subfamily

DESCRIPTION
 Intermediate between AluY1 and AluY14. Identified both in human and gibbon.

Classification and Taxa
 Classification: Interspersed_Repeat; Transposable_Element; Class_I_Retrotransposon; LINE-dependent_Retrotransposon; SINE; SINE_RNA_Promoter; No-core; L1-dependent; Alu
 Taxa: [1] Databases

Curation Details
 Status: Released
 Length: 311
 Target Site Duplication: Unknown

Citations
 1. Active Alu elements are passed primarily through paternal genomes.
 Jurka J, Kapitonov M, Khavronov V, Derjagin J, Kozharyan G.
 These Papal Biol 2002;6:519-529 [PubMed]

Aliases and External Links
 • Repeat: AluY1

Source
 Author: Arndt W, Finn RD, Hubley R, James T, Jurka J, Smit A, Wheeler T

5

Repeats are DNA Motifs

- Repeats occur in multiple instances, so use motif technology to represent them

```
accgatagggtatacgtatca-tttacgatac
atcggt-ggtttacggtcaattcaggatgc
accggt-tgtttacgtagcaatctaggatac
↓
accgat-ggtttacgtatcaatttaggatac
```

6

The Basics

- Uses RMBlast (BLAST-like tool) to compare query to consensus model library
- Uses HMMER (vaguely BLAST-like, but with much fancier math) to compare query to HMM library

13

Partial Repeats

- RepeatMasker will cheerfully report an incomplete match to a repeat.
- Detects best-conserved parts
- Some repeats (retrotransposons) typically incomplete

14

Nested Repeats

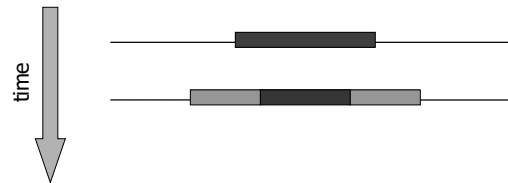
- RepeatMasker tries to detect nesting



15

Nested Repeats

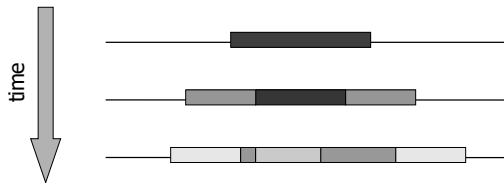
- RepeatMasker tries to detect nesting



16

Nested Repeats

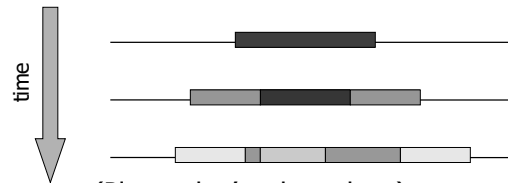
- RepeatMasker tries to detect nesting



17

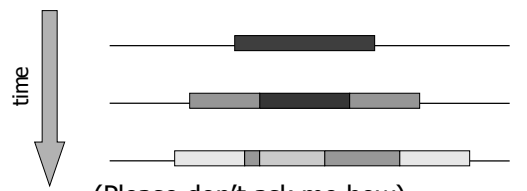
Nested Repeats

- RepeatMasker tries to detect nesting



18

Nested Repeats

- RepeatMasker tries to detect nesting
- 
- (Please don't ask me how)
 - See the [RepeatMasker presentation](#) by Dr. Jessica Storer for details

19

Overview

- Sources of repetitive sequence data
- How *RepeatMasker* finds repeats
- Issues and limitations


20

Library Choice

- Make sure to use correct libraries for your target species
- (Commonly used organisms have preselected library lists)
- Danger: mis-identifications!

21

Incomplete Masking

- Highly diverged repeats can be tough to find
 - Might leave ends of a repeat unmasked
- 
- Is this really a new feature?

22

Use the Right Tool

- Tandem repeats and duplications
 - Dust (short) — Morgulis et al., 2006
 - TRF (long) — Benson et al., 1999
- RNA
 - tRNAscan-SE, Infernal, ...
- Other repeats
 - Search for matches to Dfam (HMMER) and the NCBI nt database (BLAST)
 - Check the "Repeat tools" page on TE Hub

23

In conclusion...

Hey, let's be careful out there!



Neal Wellons; <https://flic.kr/p/ErUFPX>

24