

Design and Use of RepeatMasker

Jeremy Buhler
jbuhler@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

2

Overview

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

3

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

4

Dfam HOME | SEARCH | BROWSE | CLASSIFICATION | REPOSITORY | DOWNLOAD | PUBLICATIONS | HELP | ABOUT **ISB** Log In

DF0001317.2 [AluYf1]
 AluYf1 subfamily

SUMMARY SEED FEATURES MODEL ANNOTATIONS RELATIONSHIPS DOWNLOAD

Description
 Intermediate between AluY and AluYf, identified both in human and gibbon.

Classification and Taxa
 Classification: Interspersed_Repeat, Transposable_Element, Class_I, Retrotransposon_Line-dependent, Retroposon, SINE, T2A-RNA, Promoter, Non-coding_L1-dependent, L1

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

Citations
 1. AluY elements are passed primarily through paternal germlines. [Jin L, Cheng M, Kapranov I, Zhang J, Harkness D, These Papadimitrakis F. 2002;21:319-320. PubMed](#)

Aliases and External Links
 • Repeat: AluYf1

Example of a repeat family summary page from Dfam

5

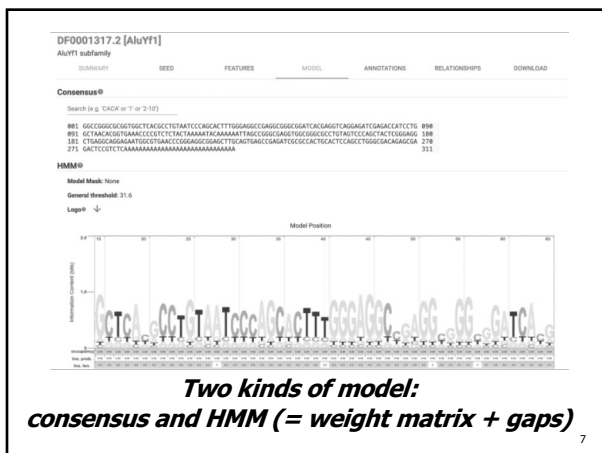
Repeats are DNA Motifs

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

```
accgatagggtatacgtatca-tttacgatac
atcgct-ggtttacgogtcaattcaggatgc
accggt-tgtttacgtagcaattcaggatgc
↓
accgat-ggtttacgtatcaatttaggatac
```

6

6

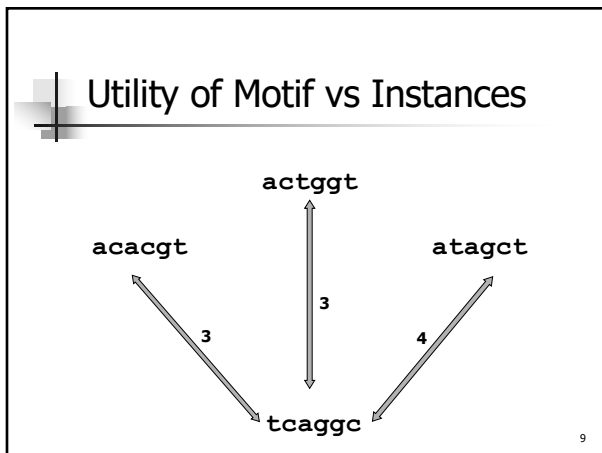


7

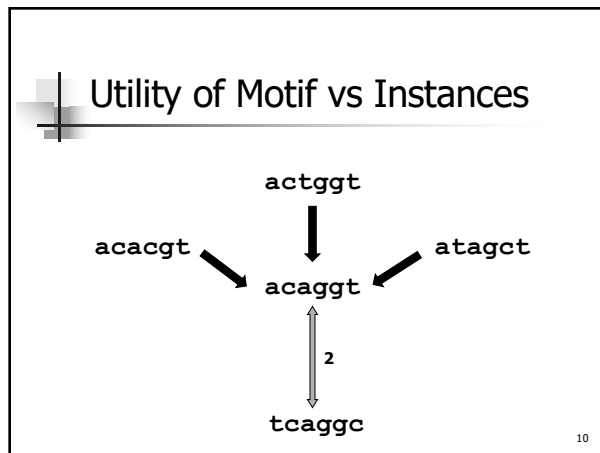
Why Use Motifs for Repeats?

- Faster to compare one sequence/model to genome than many seqs
- Even simple motifs, like a consensus sequence, are better than individual instances for discovering new copies of a repeat.

8



9



10

Types of Repeats Identified

- Interspersed (Alu, LINE, MIR, ...)
- Micro- and mini-satellites
- Noncoding RNAs (tRNA, rRNA, snoRNA, ...)
- Short tandem + low complexity (*agagagag*, *actactactact*, *aaaaataataaaa*, ...)
- Common artifacts (*E. coli*, vectors)

11

Overview

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

12

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

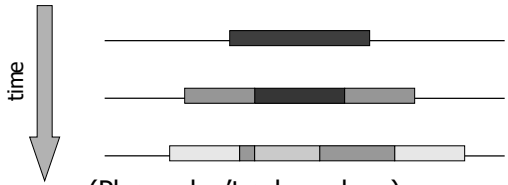
Nested Repeats

- RepeatMasker tries to detect nesting
- (Please don't ask me how)

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

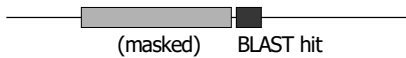
Library Choice

- 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, ...
- Low-copy (chr-specific, inverted, ...)
 - Is it in Dfam? If not, can you BLAST it?

23

In conclusion...

Hey, let's be careful out there!



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

24