09:00-12:00

- Interpretation and visualization of CAFE5 results
- Introduction to GO enrichment analyses
- GO enrichment analyses
- Wrap-up discussion

12:00-13:00 Lunch

13:00-14:00

- Talk by Alex Suh

14:00-14:15 Break

14:15-16:00

- Analyses of repeats in R

Genome assembly, annotation and comparative genomics

Day 3, afternoon

Teachers: Lars Grønvold, Thu-Hien To, Bram Danneels, Helle Tessand Baalsrud, Ole K. Tørresen

Norwegian Biodiversity & Genomics Conference 2024
10th April

Repeat detection

in a nutshell

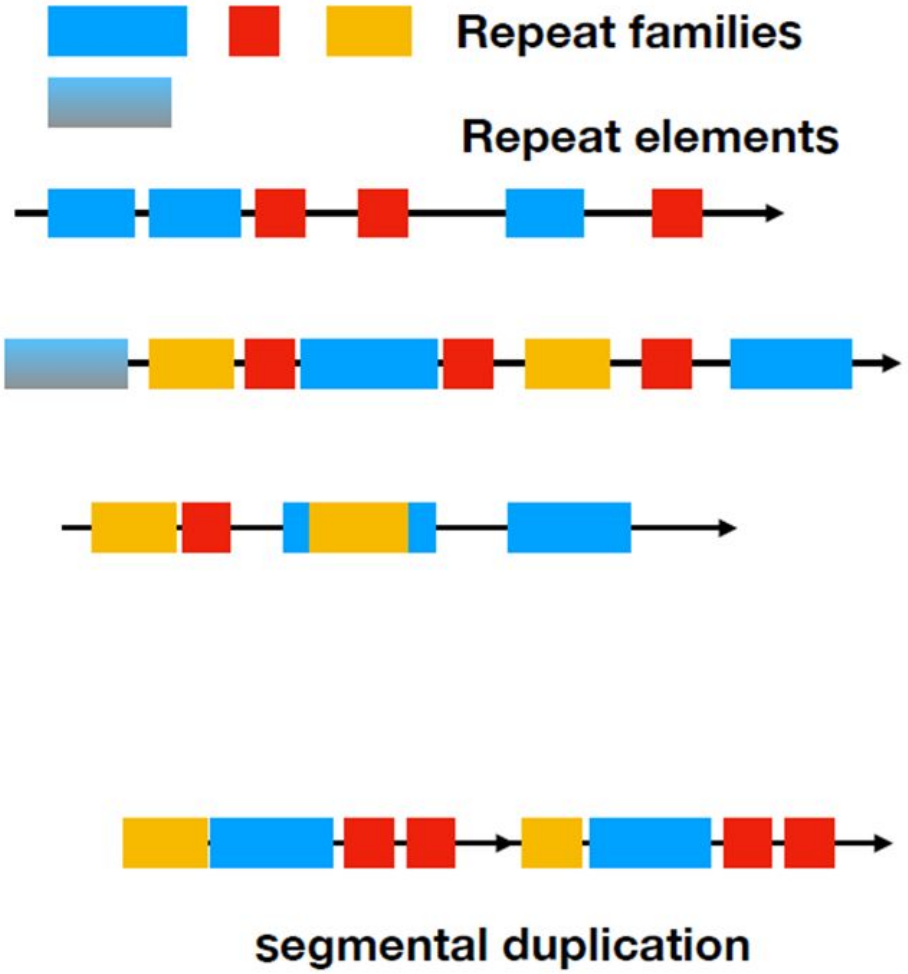
Repeat Detection

Repeats are abundant and repetitive, but can degenerate quickly

Repeat analysis:

- Repeat detection
- Repeat classification
- Repeat masking

Relies mainly on “older” software (except: RED)



De novo repeat detection

Tandem Repeat Finder

Recon: genomic alignments

RepeatScout: frequent k-mer seeding

JOURNAL ARTICLE

Tandem repeats finder: a program to analyze DNA sequences

Gary Benson

Nucleic Acids Research, Volume 27, Issue 2, 1 January 1999, Pages 573–580,
https://doi.org/10.1093/nar/27.2.573
Published: 01 January 1999 **Article history**

Genome Res. 2002 Aug; 12(8): 1269–1276.
doi: 10.1101/gr.88502
PMCID: PMC186642
PMID: 12176934

Automated De Novo Identification of Repeat Sequence Families in Sequenced Genomes

Zhirong Bao and Sean R. Eddy[1]

De novo identification of repeat families in large genomes

Alkes L. Price, Neil C. Jones, Pavel A. Pevzner

Bioinformatics, Volume 21, Issue suppl_1, , Pages i351–i358,
https://doi.org/10.1093/bioinformatics/bti1018
Published: 01 June 2005 **Article history**

General *de novo* repeat detection

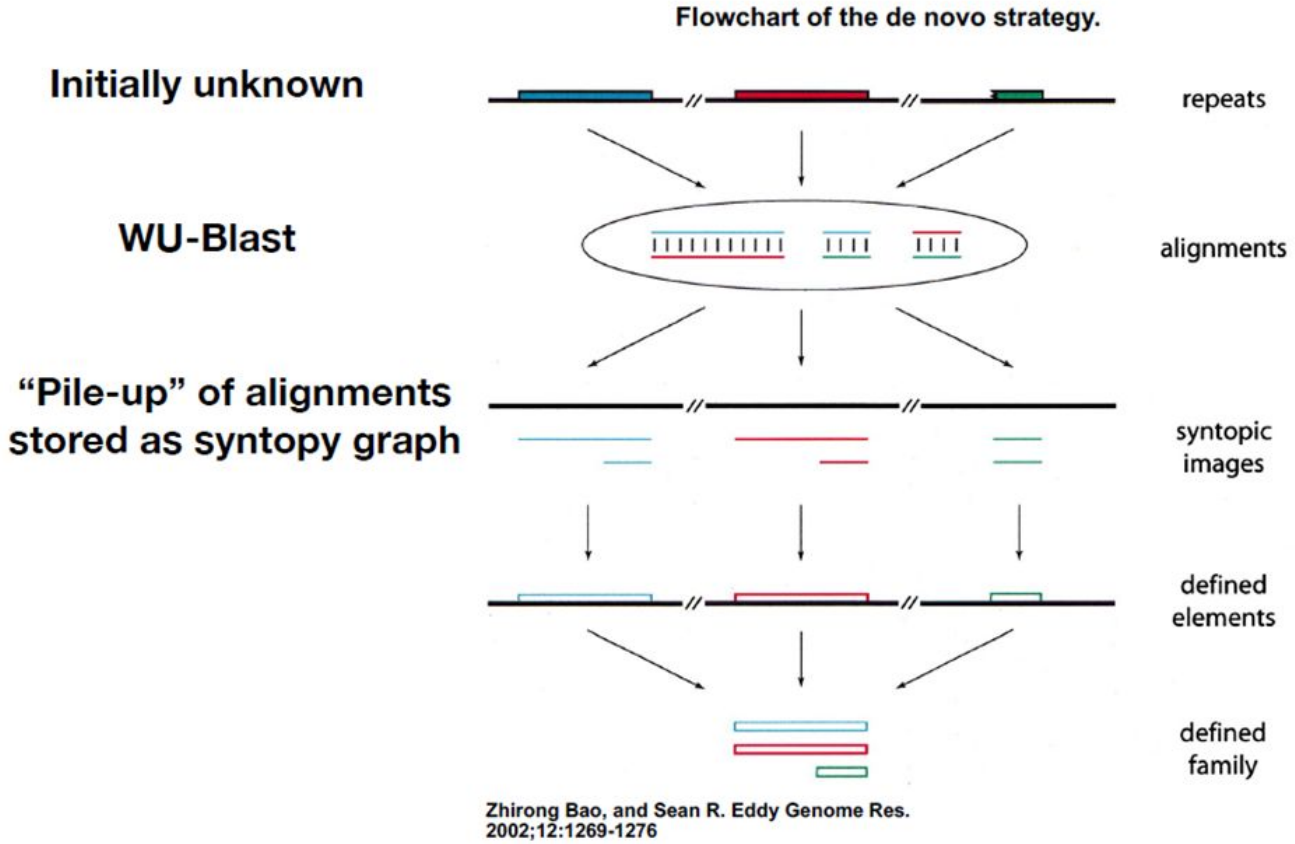
Detecting repeated sequences is not that difficult in itself

- Self-alignment
- High-frequency k-mers

Problem is defining meaningful repeat families is difficult:

- Sequence degradation, indels, divergence of copies, ...

RECON



Bao & Eddy (2002); Genome Research

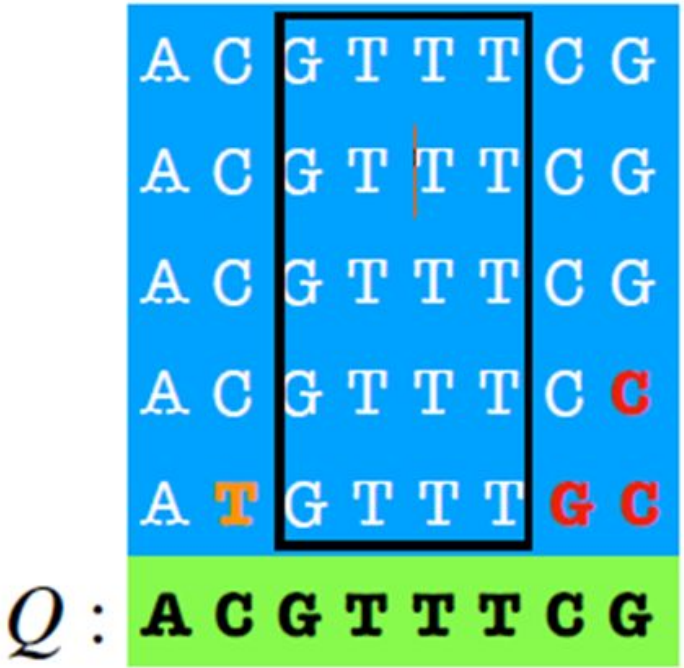
RepeatScout

Idea:

- Repeat sequences generate many identical *k*-mers
- Possibly, repeats extend to left and right of these repeated *k*-mer

Approach.

- Start from high-frequency *k*-mers
- Greedy extension of both *k*-mer ends
- Calculate consensus sequence score Q
- Continue until score doesn't improve



Price *et al*. (2005); Bioinformatics

RED - REpeat Detection

Detects candidate repeat regions based on:

- Adjusted counts of *k*-mers
- Signal processing technique
  - Separate signal from noise (repeats from non-repeats)
- Second derivative test
  - Identify local maxima (find distinct elements)

Candidate regions are used to train Hidden Markov Models

Girgis (2015); BMC Bioinformatics

RED - (Dis)Advantages

RED is:

- Self-learning repeat detection
- Very fast (faster than other tools)
- Sensitive to both tandem repeats and transposable elements
  - Recon/RepeatScout only to transposable elements
- Works on all types of genomes (draft, complete, pro-/eukaryote, ...)

But, RED does not classify repeats into families, and doesn’t tell anything about what type of repeats it finds.

Girgis (2015); BMC Bioinformatics

Repeat Masking

2 ways of masking

Softmasking:

ACGTCGGatatatatatCGATGATGGACTCCTACggtggtggtggtCTA

Hardmasking:

ACGTCGGNNNNNNNNNNCGATGATGGACTCCTACNNNNNNNNNNNNCTA

Important to check what kind of masking software requires!

RepeatModeler & RepeatMasker

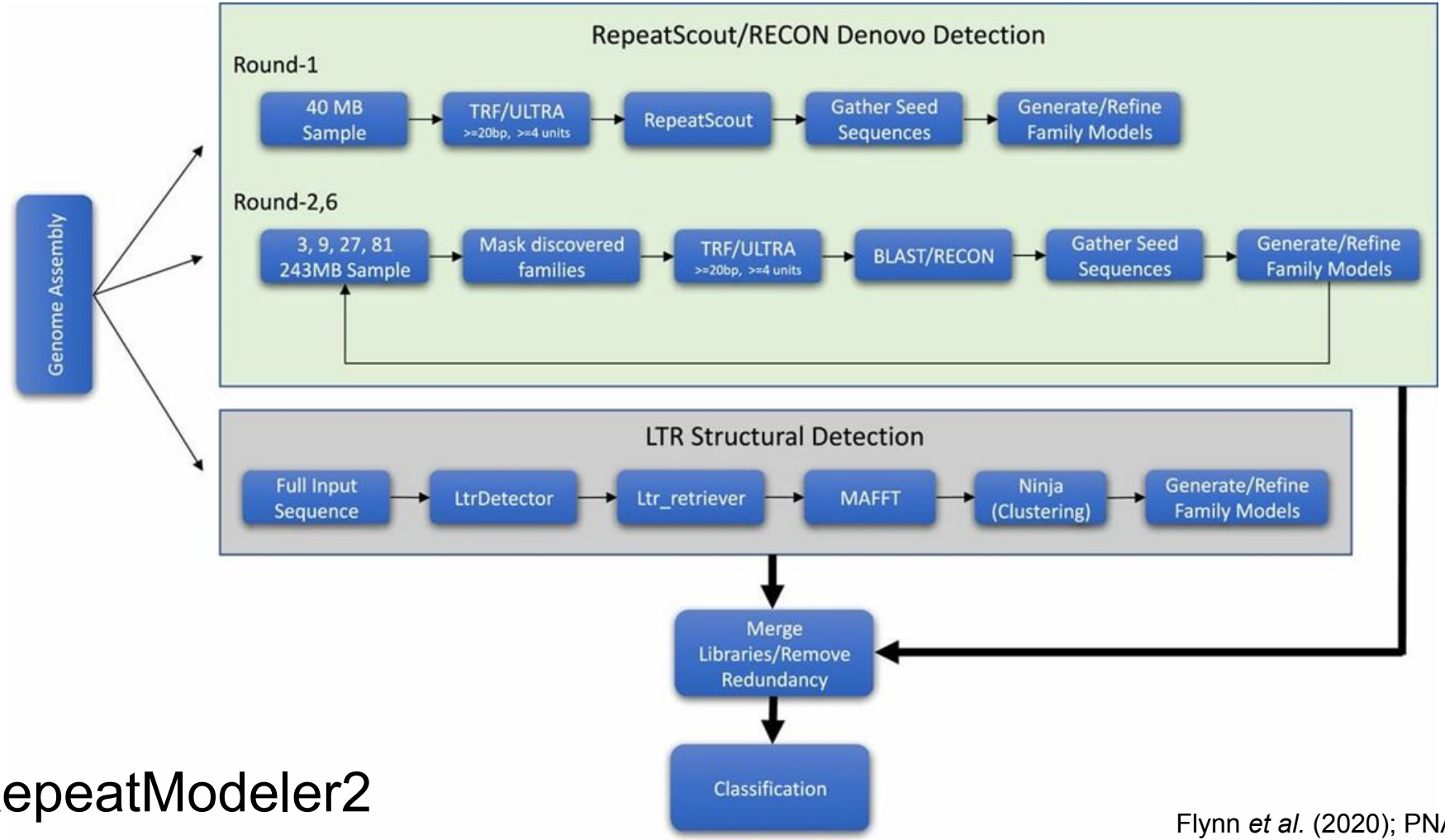
RepeatModeler and RepeatMasker are commonly used tools:

RepeatModeler is combination of repeat detection tools to create a species-specific library of repeats

RepeatMasker can use public or RepeatModeler libraries to perform repeat identification and masking in a genome

These tools are very slow to run

(multiple days/weeks on an average eukaryotic genome)

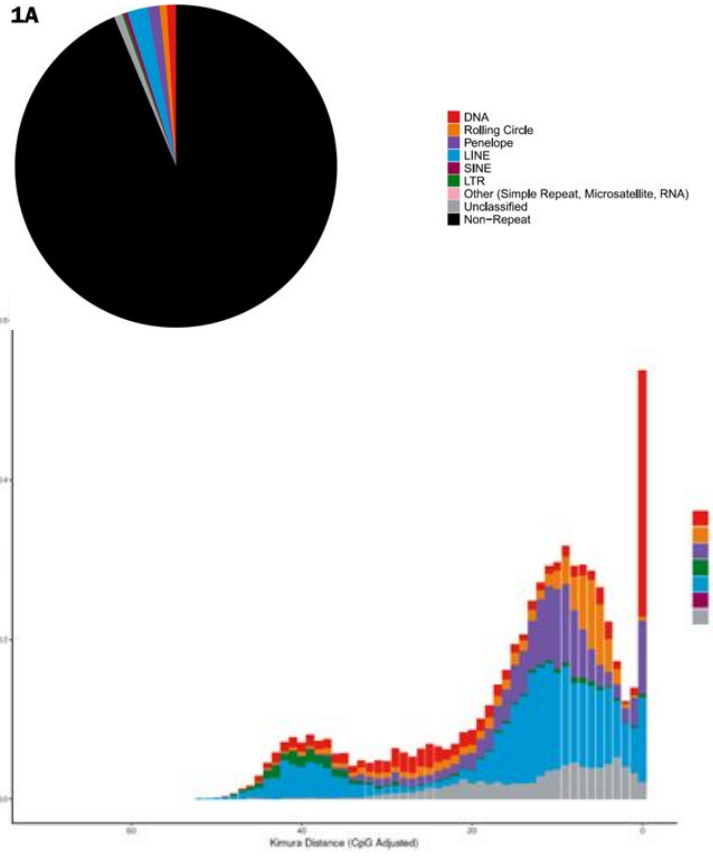


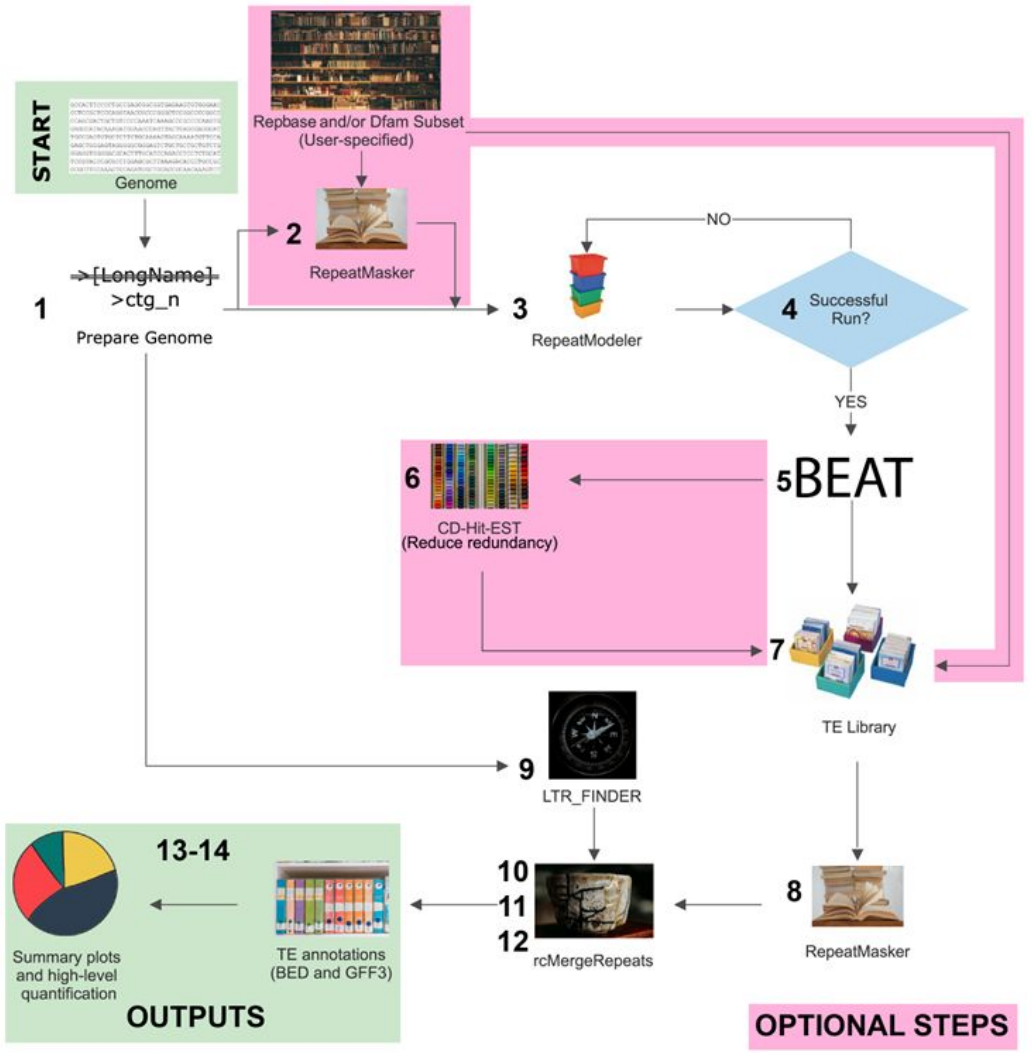
RepeatModeler2

Flynn *et al.* (2020); PNAS

Baril *et al.* (2024); Molecular Biology and Evolution
https://github.com/TobyBaril/EarlGrey

Earl Grey (TE)





09:00-12:00

- Interpretation and visualization of CAFE5 results
- Introduction to GO enrichment analyses
- GO enrichment analyses
- Wrap-up discussion

12:00-13:00 Lunch

13:00-14:00

- Talk by Alex Suh

14:00-14:15 Break

14:15-16:00

- Analyses of repeats in R

Evaluation form