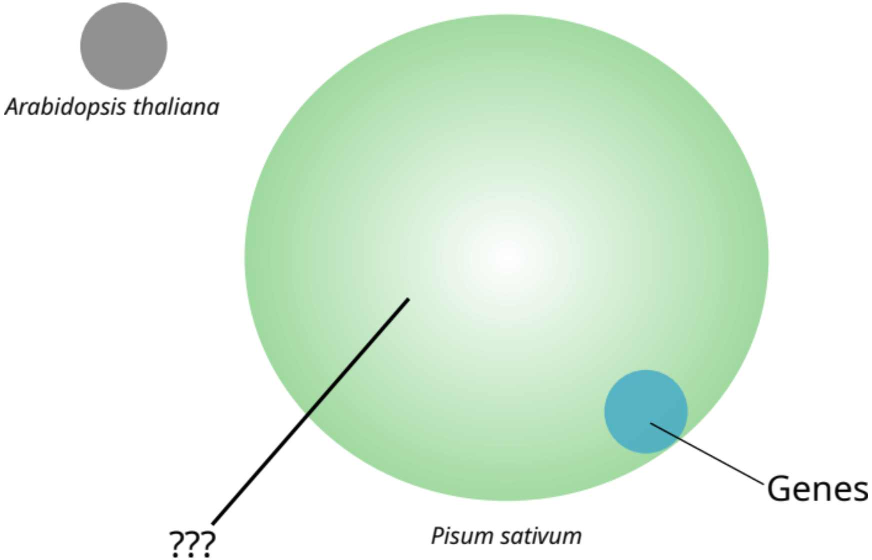
Cruising through an ocean of repeats

J.Kreplak

04/06/2024

Why some genomes are big ?



- Genes content can't always explain genome size

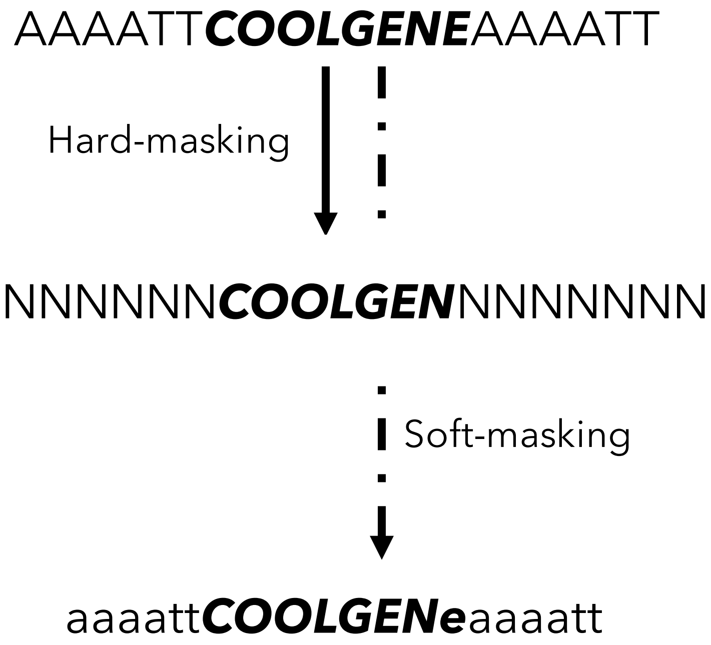
- For some large plants genomes, genes can represent less thant 10 % of the genome total size

I just want to annotate genes !

- Genomes contains a variable amount of repeated elements
- They can be simple repetition of ATGC like satellites…
- Or have the structure of a gene !
- **Transposable elements** (TE), sometimes called, jumping genes, can **be expressed** and are in a larger numbers than genes
- To annotate « useful » genes, it would be preferable to remove them

Hide everything !

- **Hard-masking** done at first ease computation of sequence comparison algorithms.
- Boundaries between a TE and a gene is sometimes difficult to establish.
- Fear of missing and hard-mask interesting genes or part of them.
- Mapper and Gene prediction software begun to authorize **soft-masking** to become able to rescue poorly annotated repeats into genes.



AAAATT***COOLGENE***AAAATT

Hard-masking

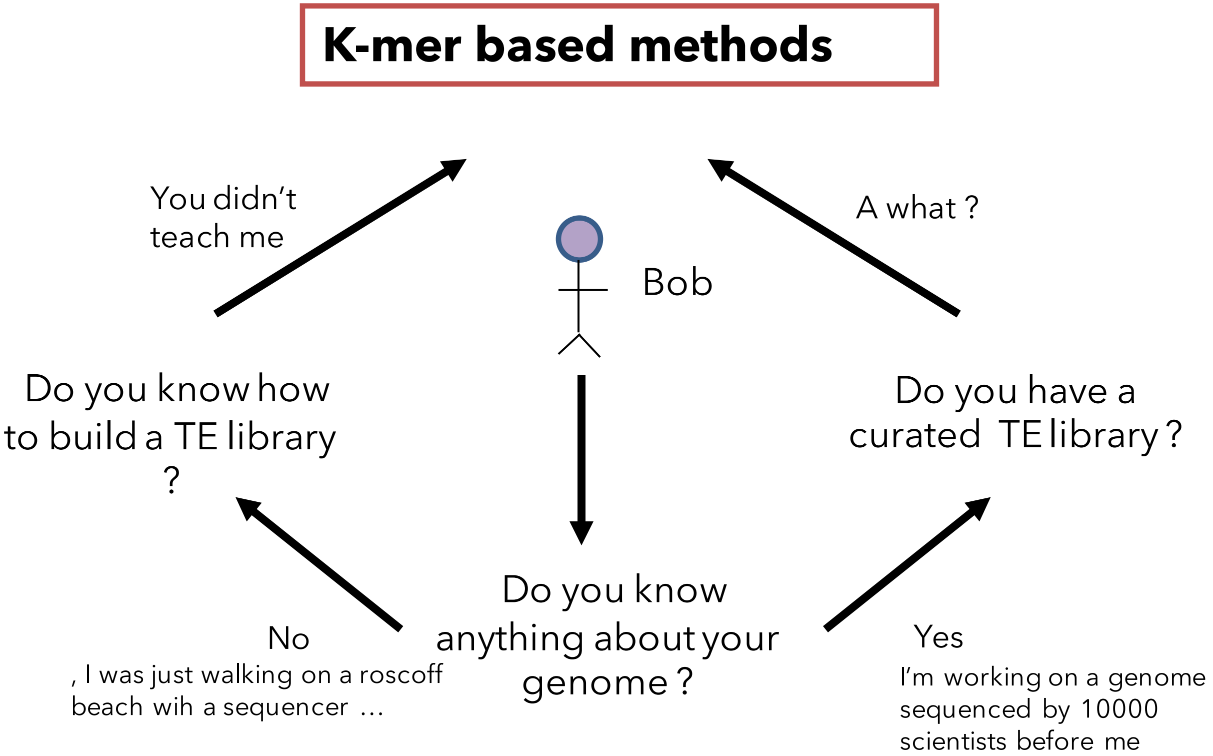
NNNNNN***COOLGEN***NNNNNNN

Soft-masking

aaaatt***COOLGENe***aaaatt

Ultimate masking flowchart

(incomplete)



K-mer based methods

Bob

Do you know anything about your genome ?

No

, I was just walking on a roscoff beach wih a sequencer …

Do you know how to build a TE library ?

You didn't teach me

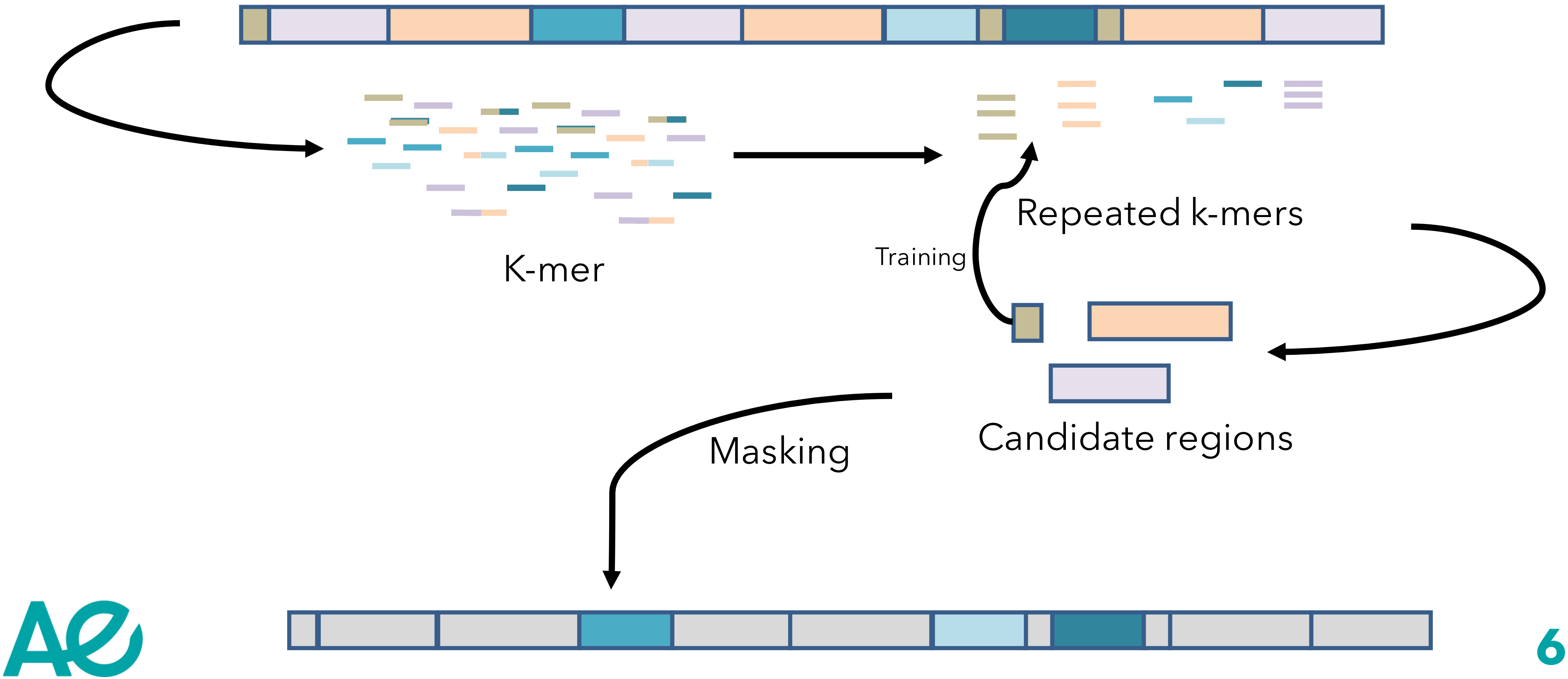
Yes

I'm working on a genome sequenced by 10000 scientists before me

Do you have a curated TE library ?

A what ?

K-mer based methods



You’ll never sleep on this bed

chrom	start	end
contig_1000	3	1703
contig_1001	0	2822
contig_1002	0	598
contig_1004	798	815
contig_1008	126	655
contig_1008	815	945
contig_1008	1111	1163
contig_1008	1369	3015

chrom	start	end	name	Score	Strand
contig_1000	3	1703	repeat-1	.	+
contig_1001	0	2822	repeat-2	.	+
contig_1002	0	598	repeat-3	.	+
contig_1004	798	815	repeat-4	.	+
contig_1008	126	655	repeat-5	.	+
contig_1008	815	945	repeat-6	.	+
contig_1008	1111	1163	repeat-7	.	+
contig_1008	1369	3015	repeat-8	.	+

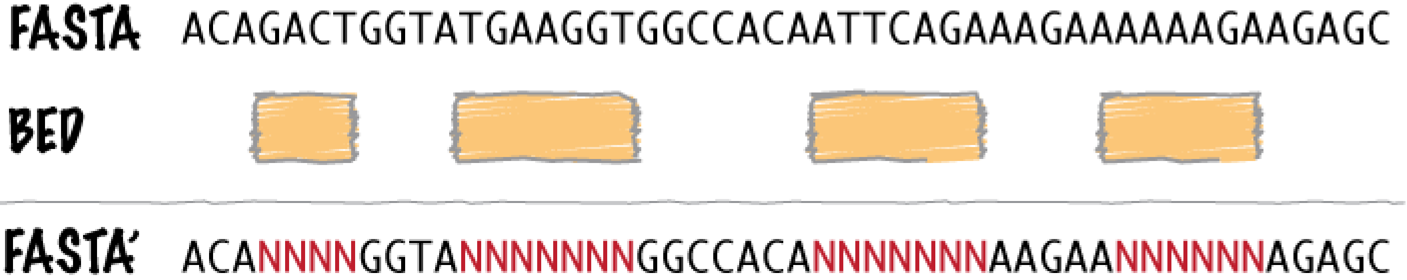
BED3

BED6

0-based coordinates

The arithmetic of coordinates

On galaxy, you can use **Operate on Genomics Intervals** tools and **Bedtools** to play with coordinate.


FASTA
ACAGACTGGTATGAAGGTGGCCACAATTCAGAAAGAAAAAAGAAGAGC
BED
FASTA'
ACANNNNGGTANNNNNNNNGGCCACANNNNNNNNAAGAANNNNNNNAGAGC

maskfasta

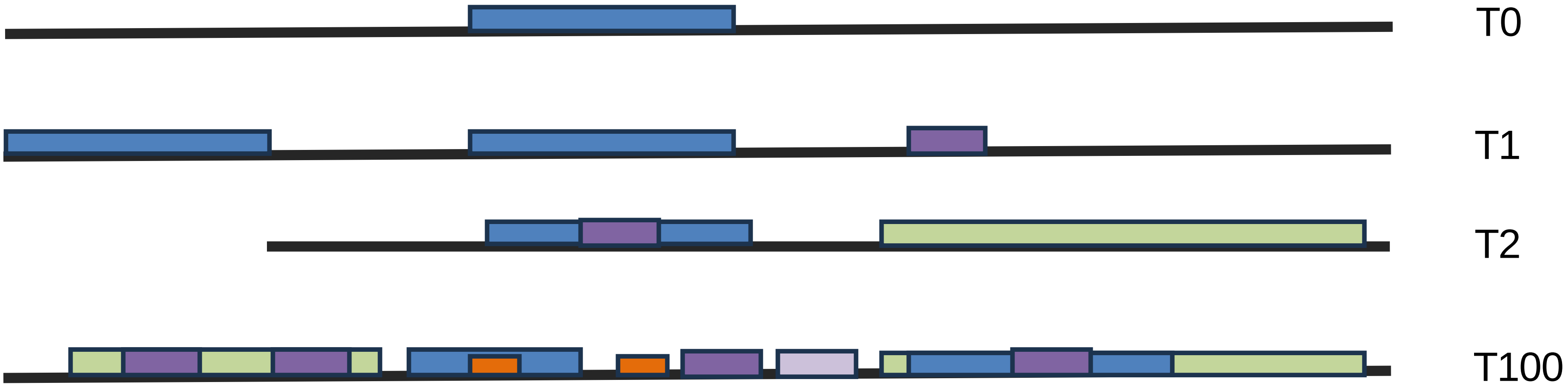
Dark matter ?

- We masked our genome but we don't know anything about what we masked

- HOWEVER repeats aren't dark matter !

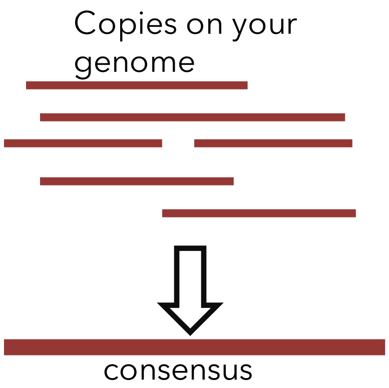
- It could be interesting to have a method that can give you information about what your are masking

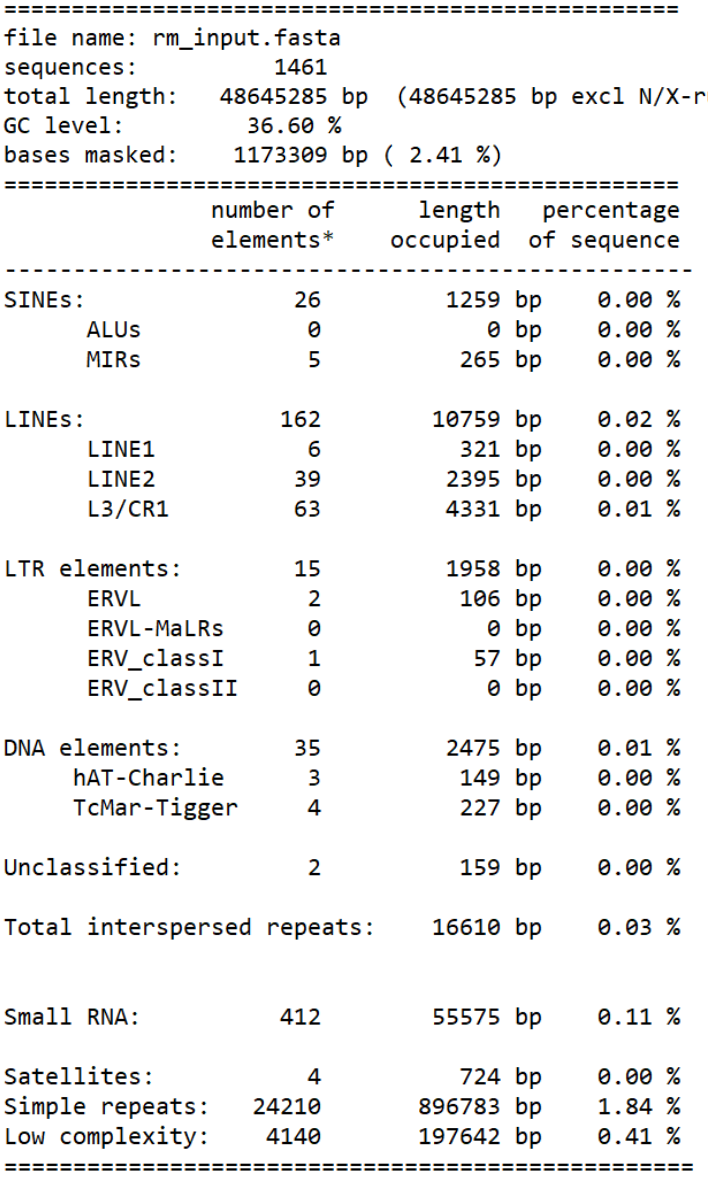
Why TE are hard to annotate ?



RepeatMasker

- RepeatMasker use a library to annotate repeats
- A library contains transposable elements found in one or several organisms
- As they are repeats, if you put inside every repeated sequence that you found, your database will have a lot of redundancy
- To fight that, we use consensus.
- Consensus can be classified



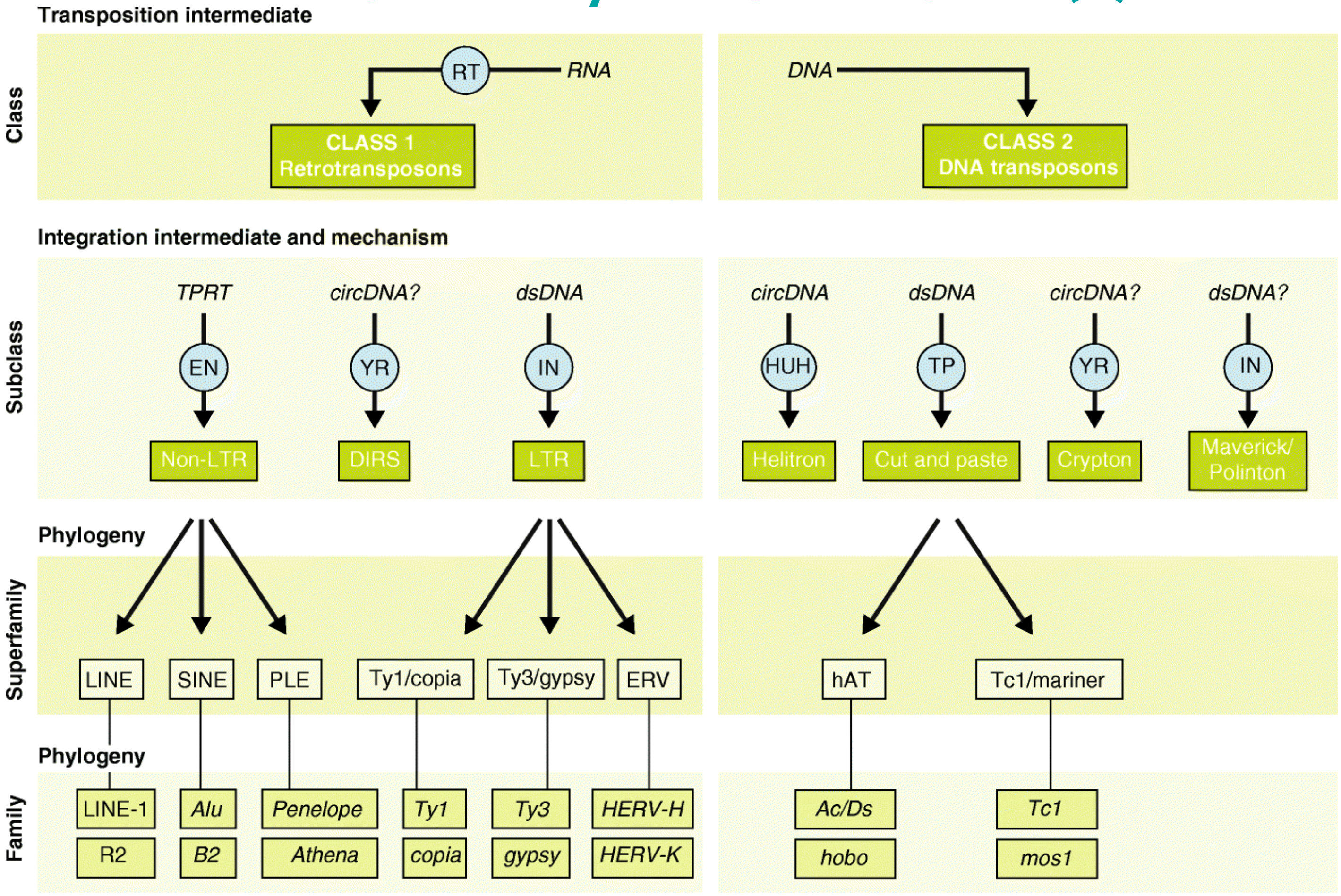


```
==================================================
file name: rm_input.fasta
sequences:          1461
total length:   48645285 bp  (48645285 bp excl N/X-runs)
GC level:         36.60 %
bases masked:    1173309 bp ( 2.41 %)
==================================================
               number of      length   percentage
               elements*    occupied  of sequence
--------------------------------------------------
SINEs:               26         1259 bp    0.00 %
      ALUs            0            0 bp    0.00 %
      MIRs            5          265 bp    0.00 %

LINEs:              162        10759 bp    0.02 %
      LINE1           6          321 bp    0.00 %
      LINE2          39         2395 bp    0.00 %
      L3/CR1         63         4331 bp    0.01 %

LTR elements:        15         1958 bp    0.00 %
      ERVL            2          106 bp    0.00 %
      ERVL-MaLRs      0            0 bp    0.00 %
      ERV_classI      1           57 bp    0.00 %
      ERV_classII     0            0 bp    0.00 %

DNA elements:        35         2475 bp    0.01 %
     hAT-Charlie      3          149 bp    0.00 %
     TcMar-Tigger     4          227 bp    0.00 %

Unclassified:         2          159 bp    0.00 %

Total interspersed repeats:    16610 bp    0.03 %


Small RNA:          412        55575 bp    0.11 %

Satellites:           4          724 bp    0.00 %
Simple repeats:   24210       896783 bp    1.84 %
Low complexity:    4140       197642 bp    0.41 %
==================================================
```

Human Fungi

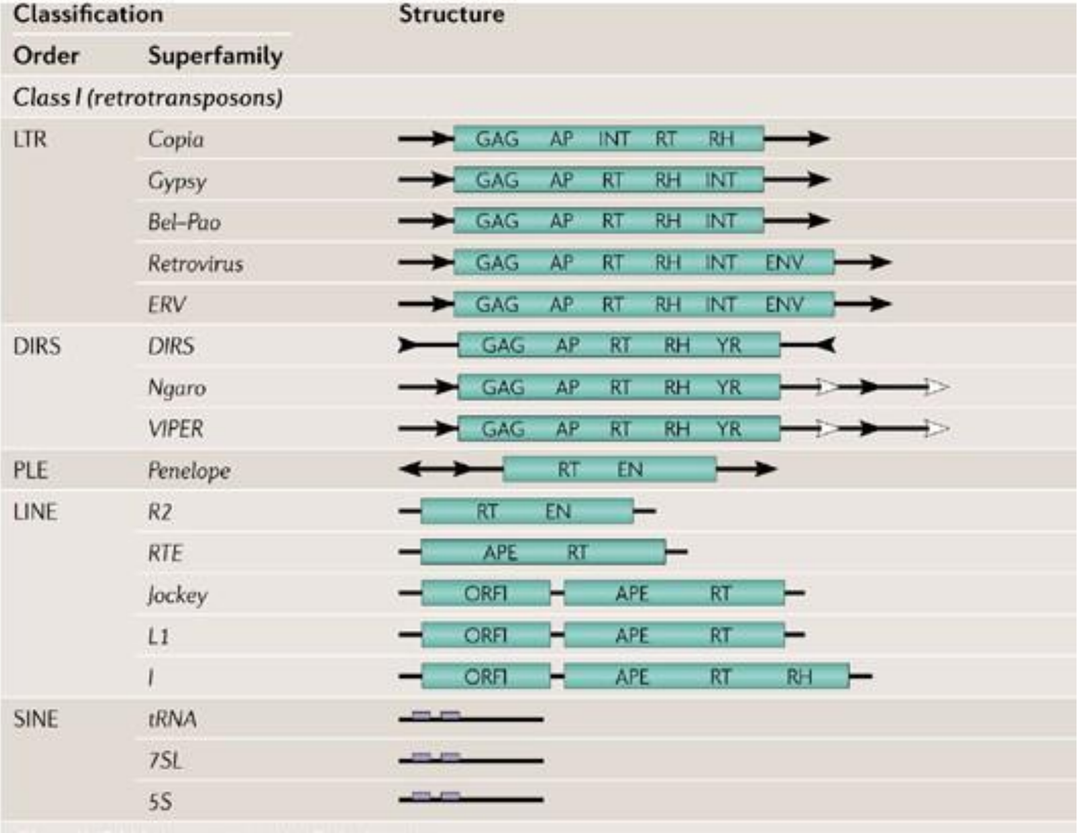
- Do you think that 2.41% is a good result ?
- Most of our results are simple repeats
- A « bad » results is often due to a bad library
- Fungi, Plants and Mammals don’t have the same ratio of TE.
- Plants love LTR
- Mammals prefer DNA elements.

Family Portrait (I)

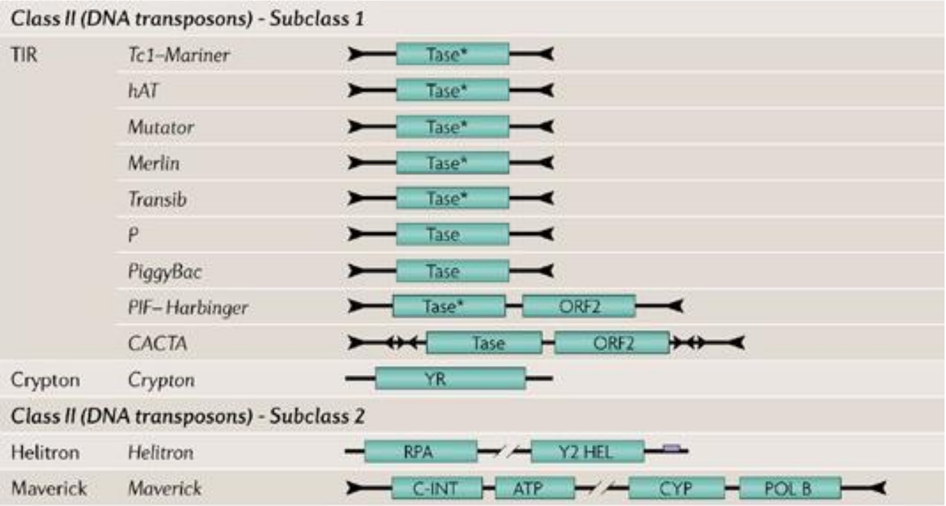


Bourque et al, 2018

Family Portrait (II)



Classification		Structure
Order	**Superfamily**	
Class I (retrotransposons)		
LTR	*Copia*	GAG AP INT RT RH
	Gypsy	GAG AP RT RH INT
	Bel–Pao	GAG AP RT RH INT
	Retrovirus	GAG AP RT RH INT ENV
	ERV	GAG AP RT RH INT ENV
DIRS	*DIRS*	GAG AP RT RH YR
	Ngaro	GAG AP RT RH YR
	VIPER	GAG AP RT RH YR
PLE	*Penelope*	RT EN
LINE	*R2*	RT EN
	RTE	APE RT
	Jockey	ORF1 APE RT
	L1	ORF1 APE RT
	I	ORF1 APE RT RH
SINE	*tRNA*	
	7SL	
	5S	

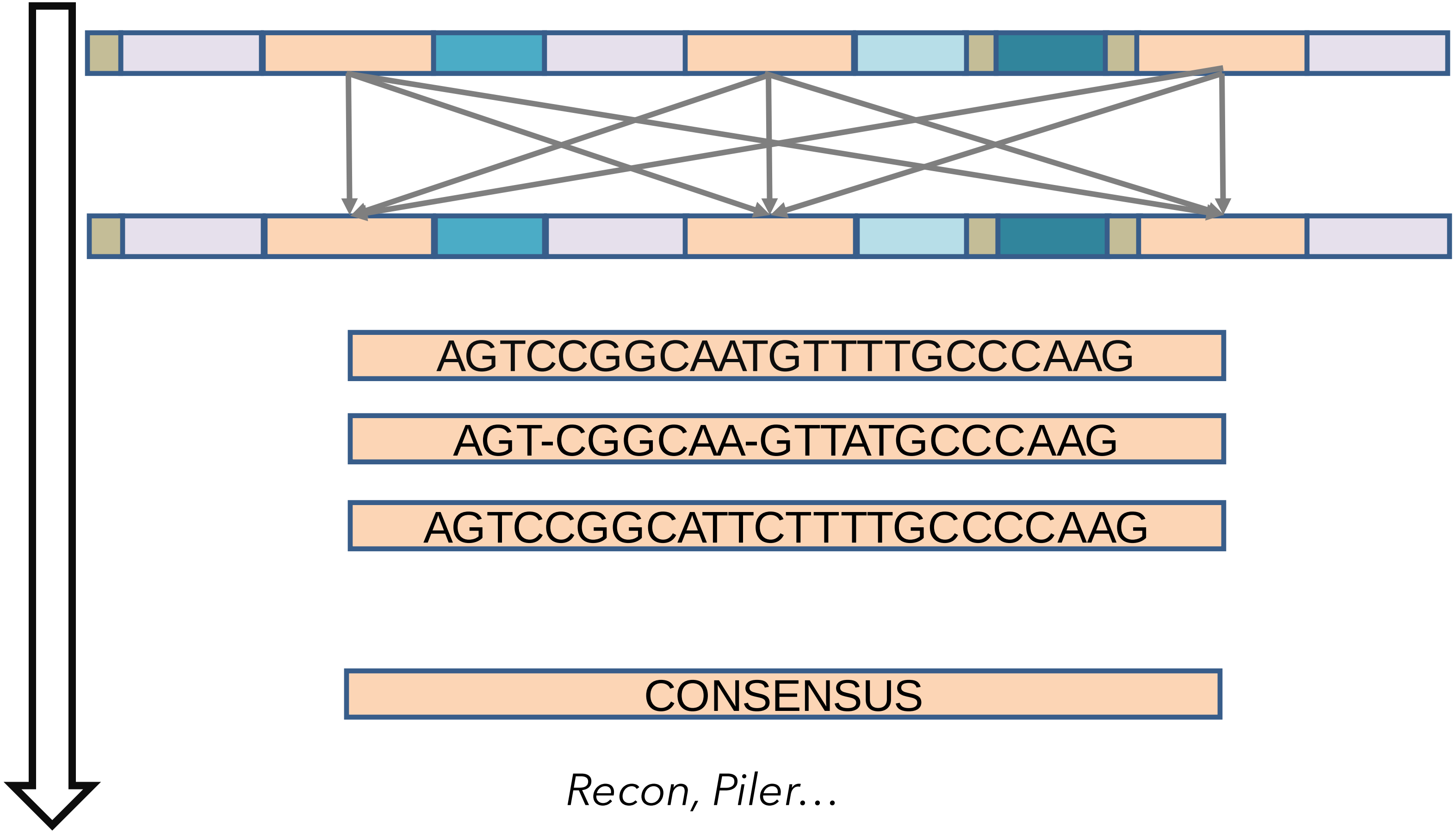


Class II (DNA transposons) - Subclass 1		
TIR	*Tc1–Mariner*	Tase*
	hAT	Tase*
	Mutator	Tase*
	Merlin	Tase*
	Transib	Tase*
	P	Tase
	PiggyBac	Tase
	PIF–Harbinger	Tase* ORF2
	CACTA	Tase ORF2
Crypton	*Crypton*	YR
Class II (DNA transposons) - Subclass 2		
Helitron	*Helitron*	RPA Y2 HEL
Maverick	*Maverick*	C-INT ATP CYP POL B

Wicker et al, 2007

a TE library, pleaaase !

Clustering-based methods



Align against yourself

AGTCCGGCAATGTTTTGCCCAAG

AGT-CGGCAA-GTTATGCCCAAG

AGTCCGGCATTCTTTTGCCCCAAG

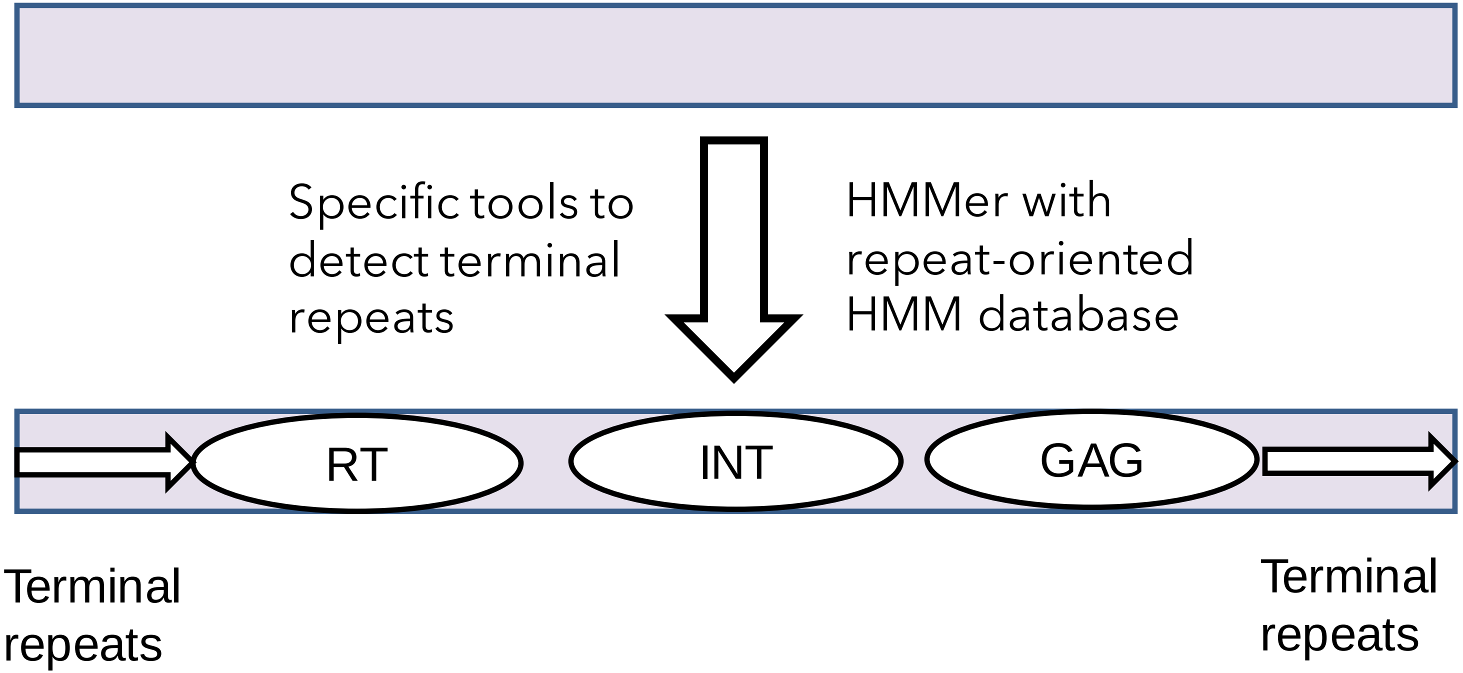
Keep regions with at least 3 copies and multi-align

CONSENSUS

Compute consensus

Recon, Piler…

Structural-based methods



Specific tools to detect terminal repeats

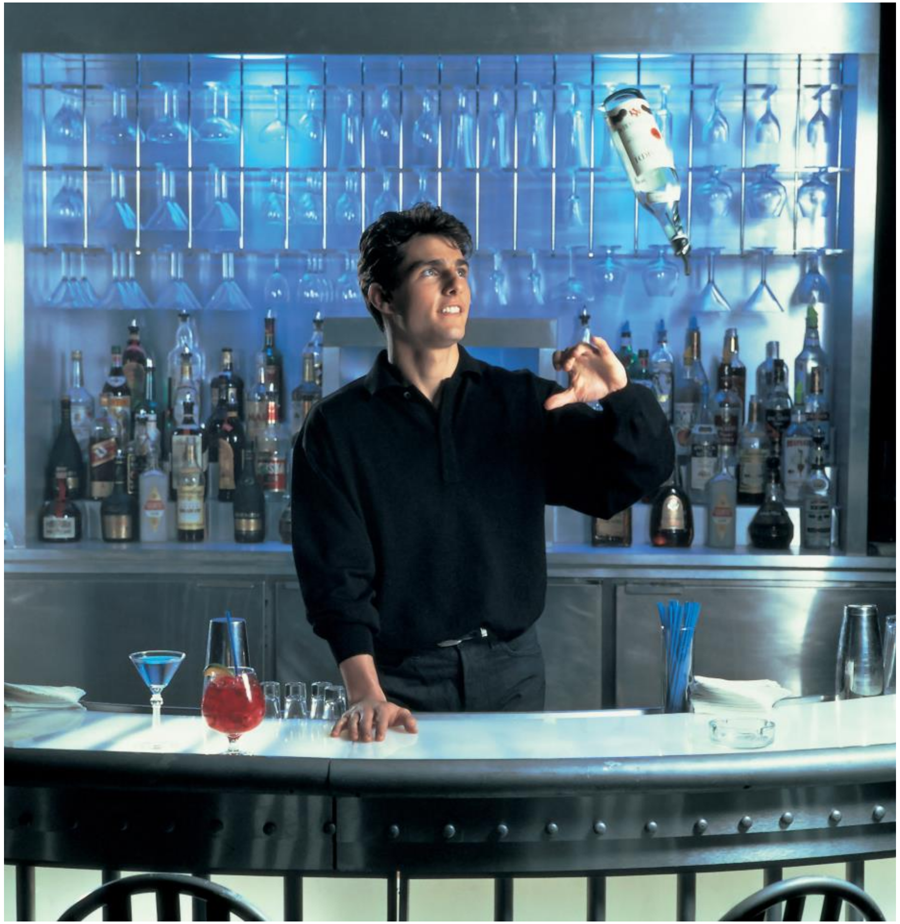
HMMer with repeat-oriented HMM database

Terminal repeats

Terminal repeats

LTRHarvest, MiteFinder, HelitronScanner…

Cocktail !



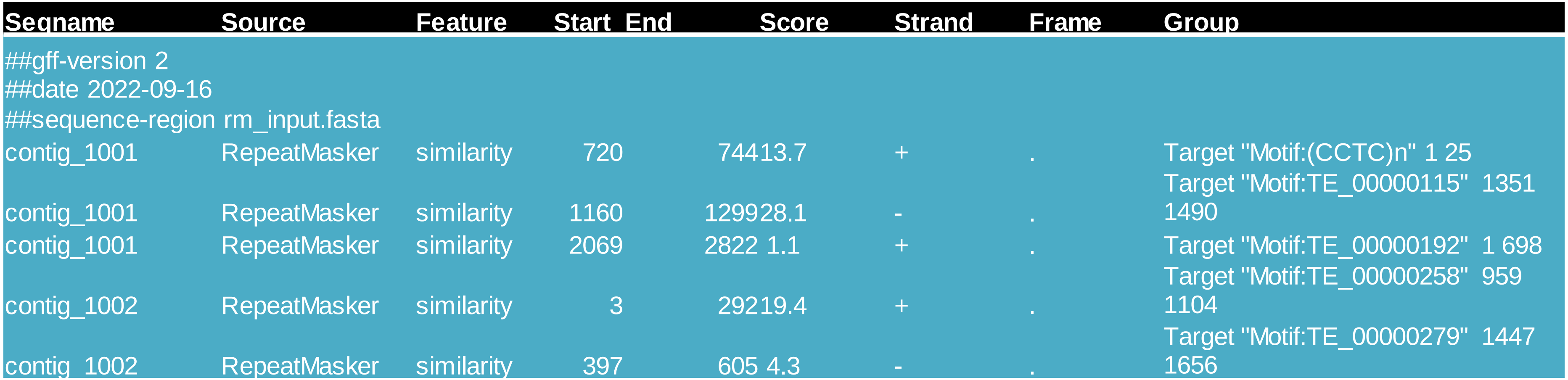
- Recent tools mixed clustering and structural based methods.

- Clustering could fail for ancient elements that have accumulated mutations or with a limited number of copies.

- Structural is really good for families with a clear succession of domains like LTR.

- **RepeatModeler2**, **EDTA** or **Tedenovo** are using both.

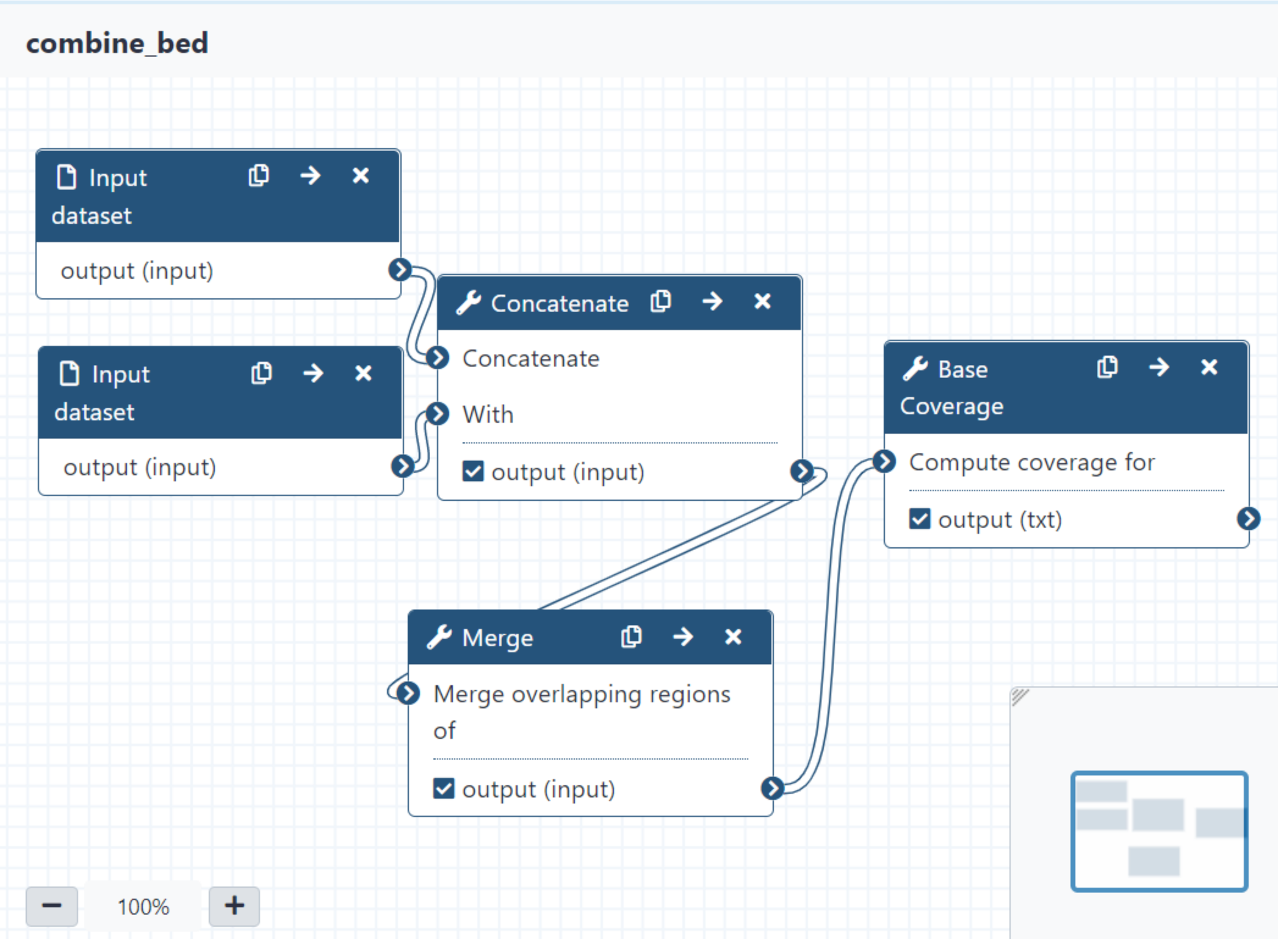
GFF2 DEPRECATED



Seqname	Source	Feature	Start	End	Score	Strand	Frame	Group
##gff-version 2								
##date 2022-09-16								
##sequence-region rm_input.fasta								
contig_1001	RepeatMasker	similarity	720	744	13.7	+	.	Target "Motif:(CCTC)n" 1 25
contig_1001	RepeatMasker	similarity	1160	1299	28.1	-	.	Target "Motif:TE_00000115" 1351 1490
contig_1001	RepeatMasker	similarity	2069	2822	1.1	+	.	Target "Motif:TE_00000192" 1 698
contig_1002	RepeatMasker	similarity	3	292	19.4	+	.	Target "Motif:TE_00000258" 959 1104
contig_1002	RepeatMasker	similarity	397	605	4.3	-	.	Target "Motif:TE_00000279" 1447 1656

Deprecated because unable to work properly with nested feature like genes.
Still used by a lot of tools… :'(:'(

The arithmetic of coordinates II



Expert annotation

- Repeats shouldn’t stay as dark matter
- Genome structure and gene regulations can be heavily impacted by transposable elements
- Annotating properly your repeats can help you to craft meaningful hypothesis on genome structure and gene expression.

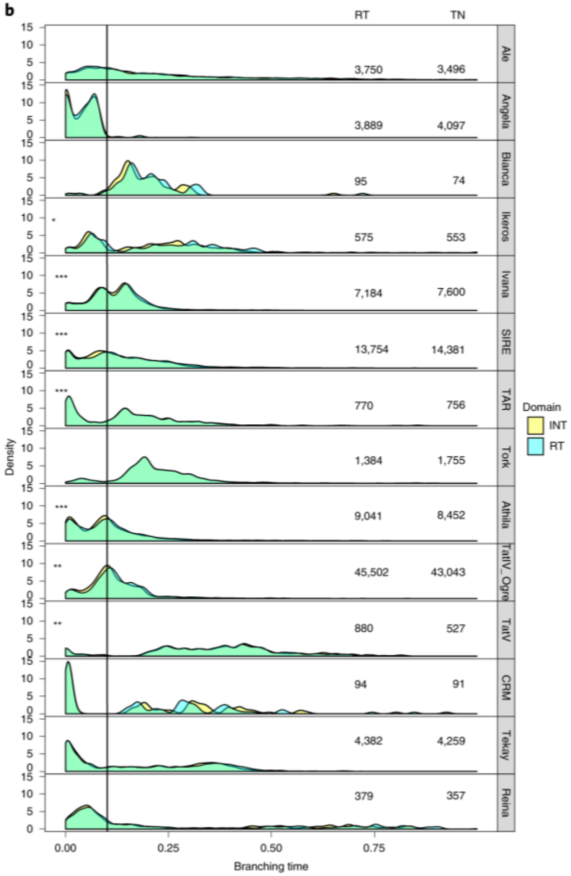
A short and recent bibliography

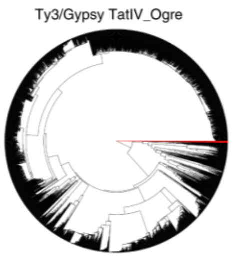
- **EarlGrey** (https://doi.org/10.1093/molbev/msae068)
  - RepeatModeler2-based
  - An iterative step is implemented to obtain larger and better consensus sequences.

- **PanREPET** (unpublished; poster)
  - Basé sur le pipeline REPET (TEdenovo – TEannot)
  - Permet de propager une banque de consensus sur un ensemble de génomes

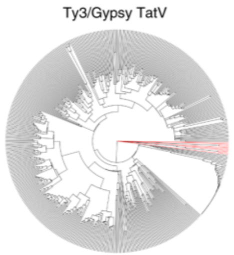
- **DANTE** (https://doi.org/10.1101/2024.04.17.589915)
  - Structural-based approach specialized for plants
  - Really good to identify full-length elements for LTR

Pisum sativum burst

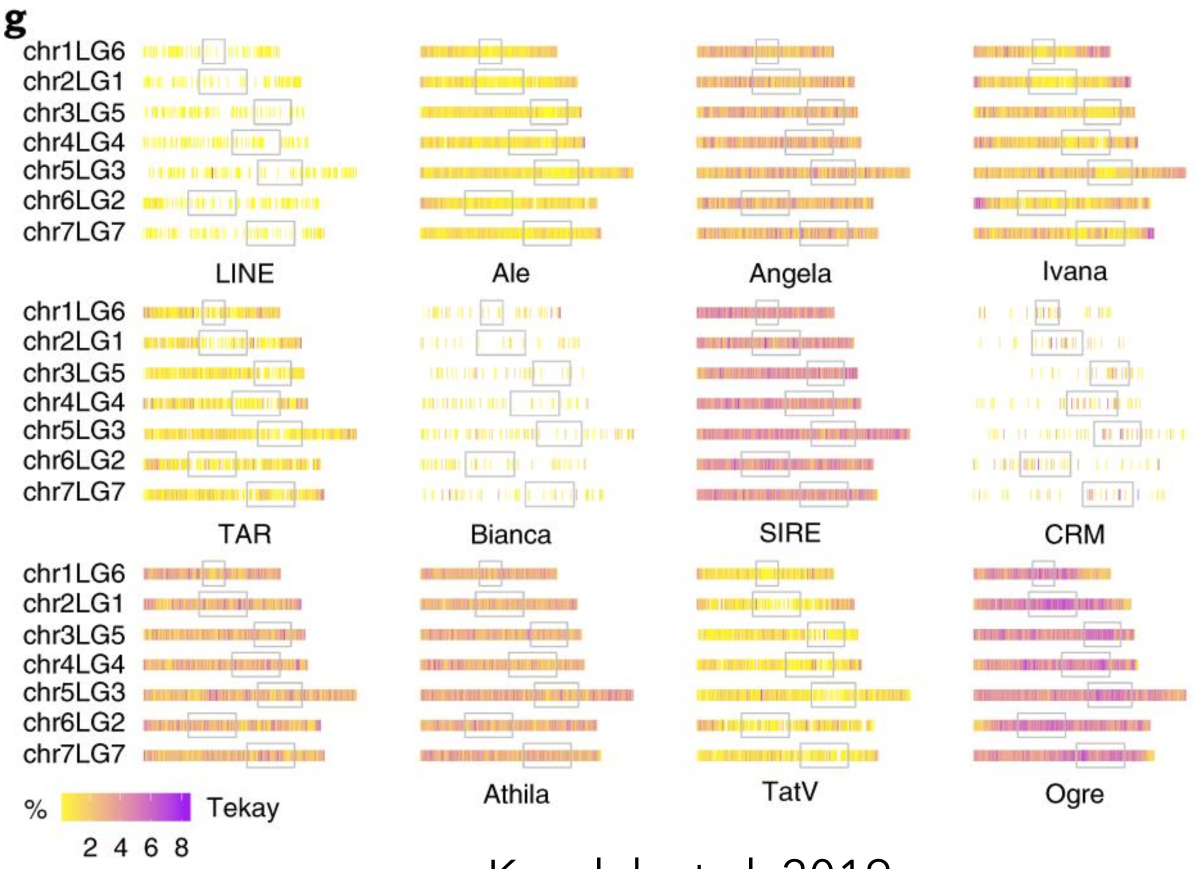




Ty3/Gypsy TatIV_Ogre

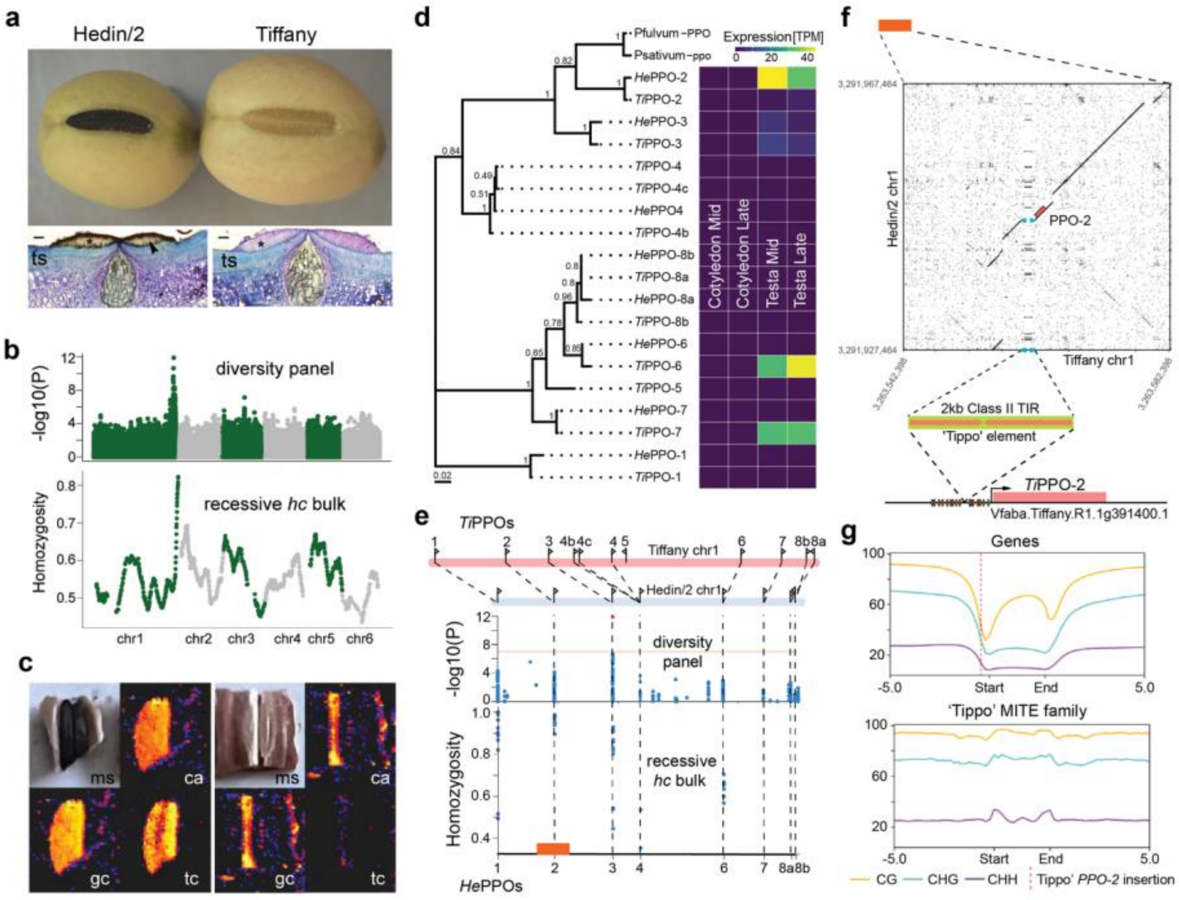


Ty3/Gypsy TatV



Kreplak et al. 2019

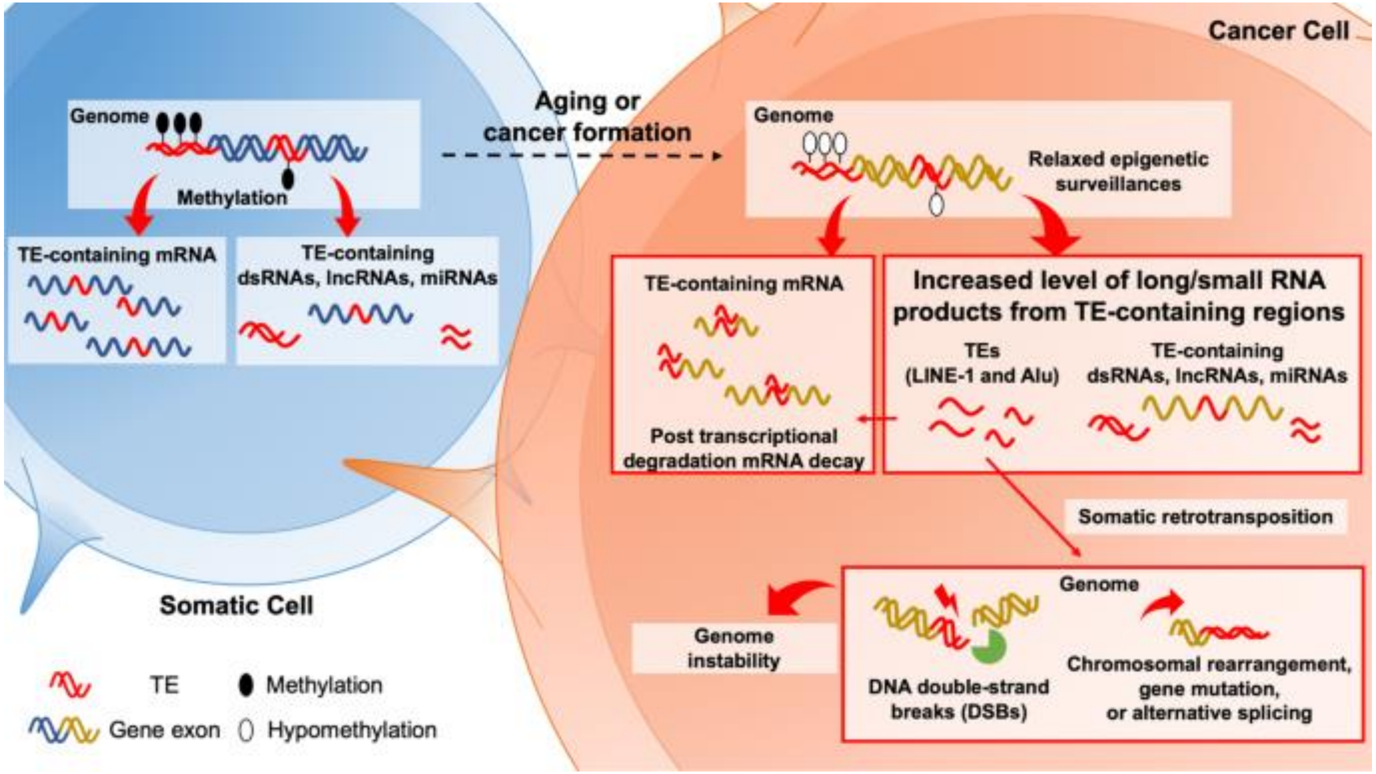
Vicia faba PPO



- Hillum color is a classic phenotype in faba
- Polyphenol oxydase (PPO) were a known potential candidate
- Comparison between two genome assemblies (**Ti**ffany and **He**din) was able to show than a MITE insertion among transcription factor binding site of PPO-2 inhibit its expression.

Jayakodi et al. 2023

Hypomethylation of TE can drive genome instability


Cancer Cell
Aging or cancer formation
Genome
Methylation
TE-containing mRNA
TE-containing dsRNAs, lncRNAs, miRNAs
Somatic Cell
Genome
Relaxed epigenetic surveillances
TE-containing mRNA
Post transcriptional degradation mRNA decay
Increased level of long/small RNA products from TE-containing regions
TEs (LINE-1 and Alu)
TE-containing dsRNAs, lncRNAs, miRNAs
Somatic retrotransposition
Genome
Genome instability
DNA double-strand breaks (DSBs)
Chromosomal rearrangement, gene mutation, or alternative splicing
TE
Methylation
Gene exon
Hypomethylation

Hsu et al. 2021