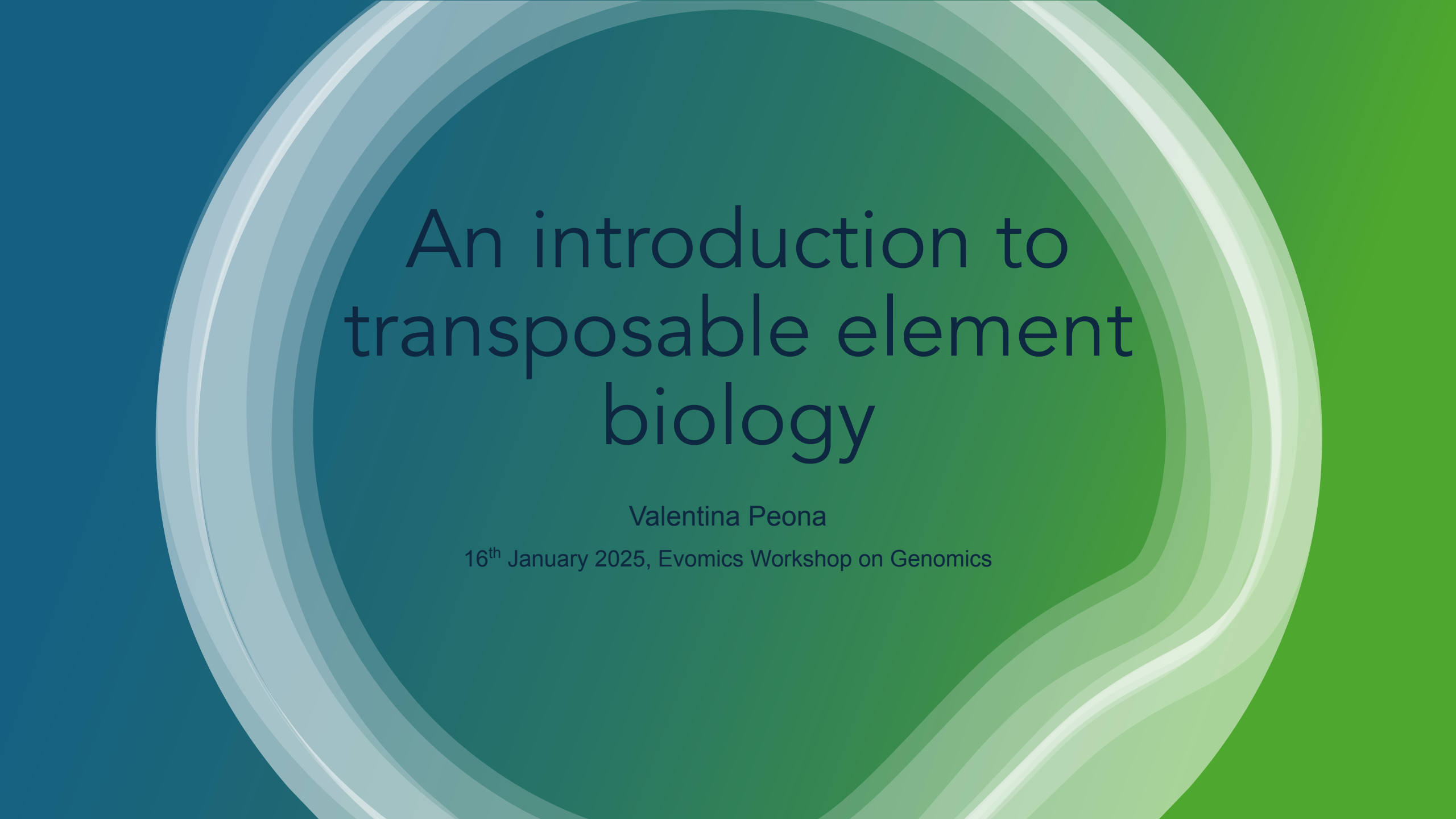




An introduction to transposable element biology

Valentina Peona

16th January 2025, Evomics Workshop on Genomics



Bologna (IT)

Alexander Suh



Uppsala (SE)

Reto Burri
Martin Irestedt



+



Stockholm (SE) Sempach (CH)



2010

2014

2017

2021 2022

2023

Bachelor/Master

Research assistant

PhD

Postdoc



Ecology/Popgen

Comparative genomics

Speciation genomics

Overview

- Part1: intro to TE biology
- Part2: Methods to detect TEs in genomes (+ intro to tutorial)

The diagram illustrates the classification of repetitive DNA sequences into two main categories: interspersed repeats and tandem repeats. A large, stylized blue 'C' shape serves as a background, separating the two categories.

Interspersed repeats: This category is represented by a horizontal line with three colored blocks (red, yellow, red) interspersed along it. Below this, a list of examples is provided:

- Retrotransposons
- DNA transposons
- Endogenous viruses

Tandem repeats: This category is represented by a horizontal line with a series of purple arrowheads pointing to the right, indicating a head-to-tail arrangement. Below this, a list of examples is provided:

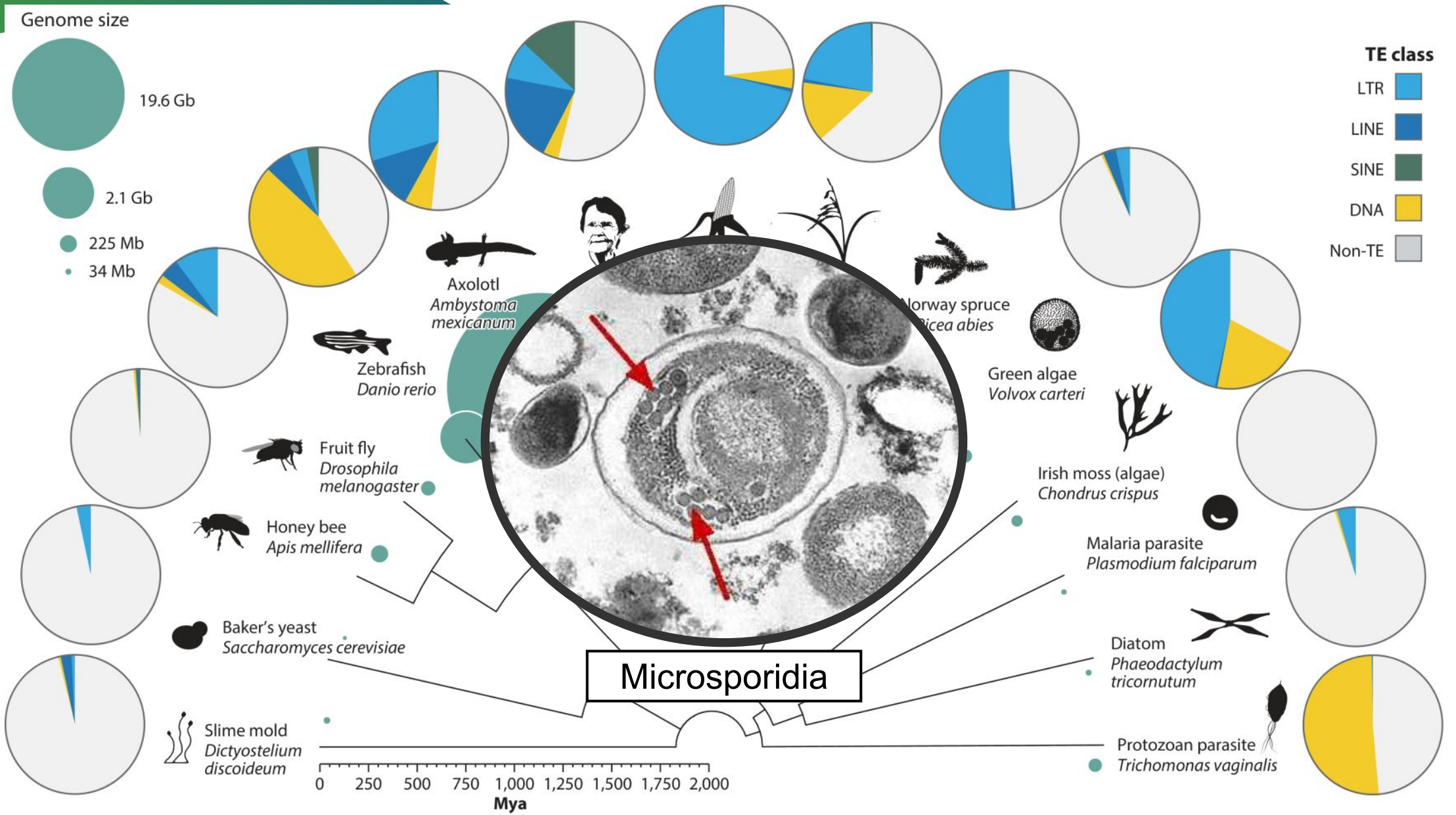
- Satellites
- Minisatellites
- Microsatellites

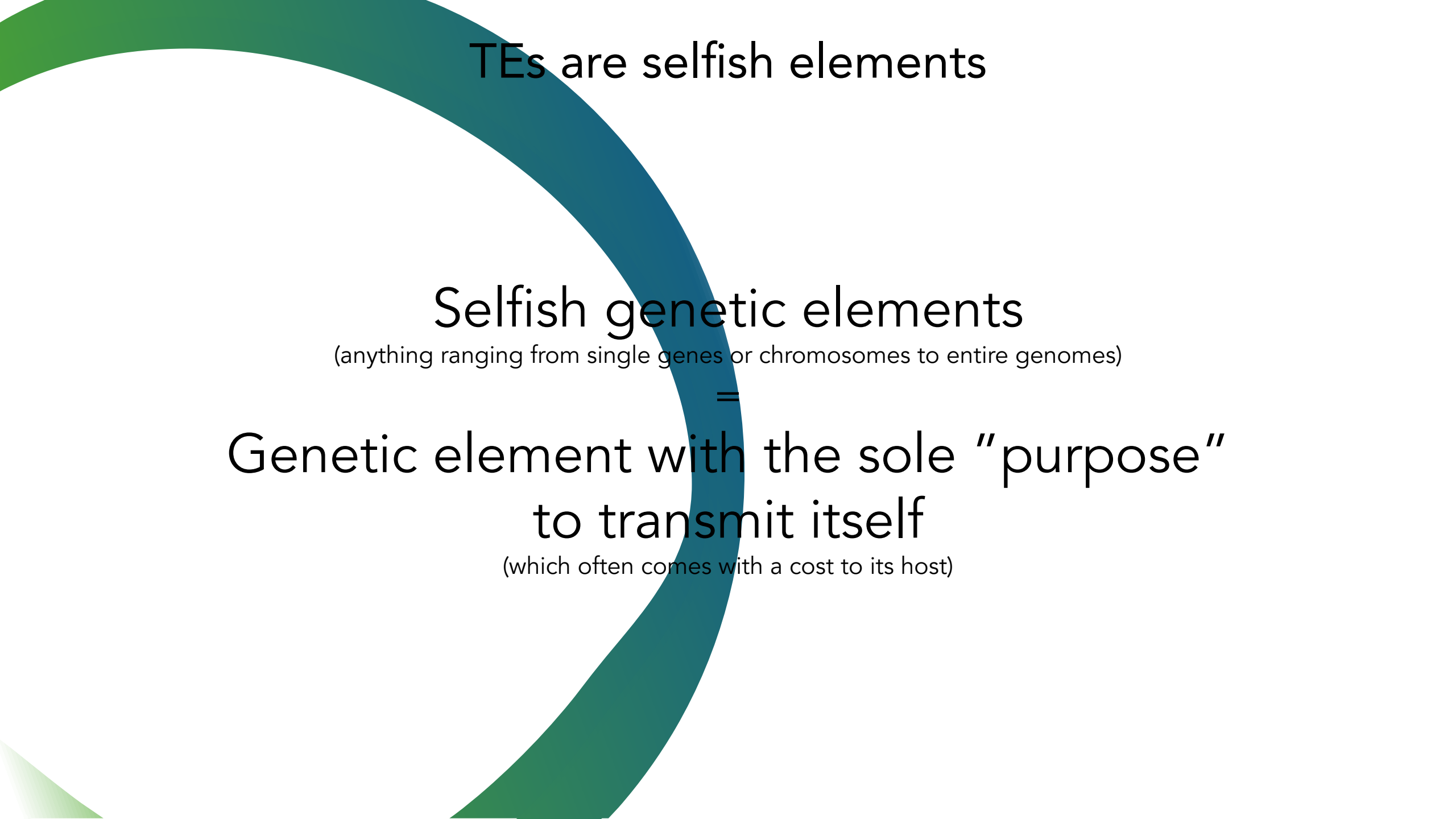
The diagram also shows a complex DNA structure at the top, featuring a loop and various colored blocks (yellow, black, red, blue, green, black) and arrowheads, illustrating the diversity of repetitive DNA sequences.

Barbara McClintock



Nobel Prize
1983





TEs are selfish elements

Selfish genetic elements

(anything ranging from single genes or chromosomes to entire genomes)

=

Genetic element with the sole "purpose"
to transmit itself

(which often comes with a cost to its host)

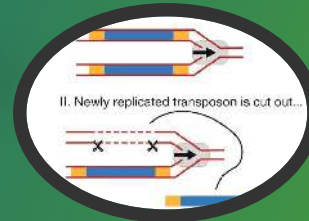
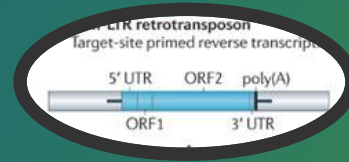


Why are they selfish?

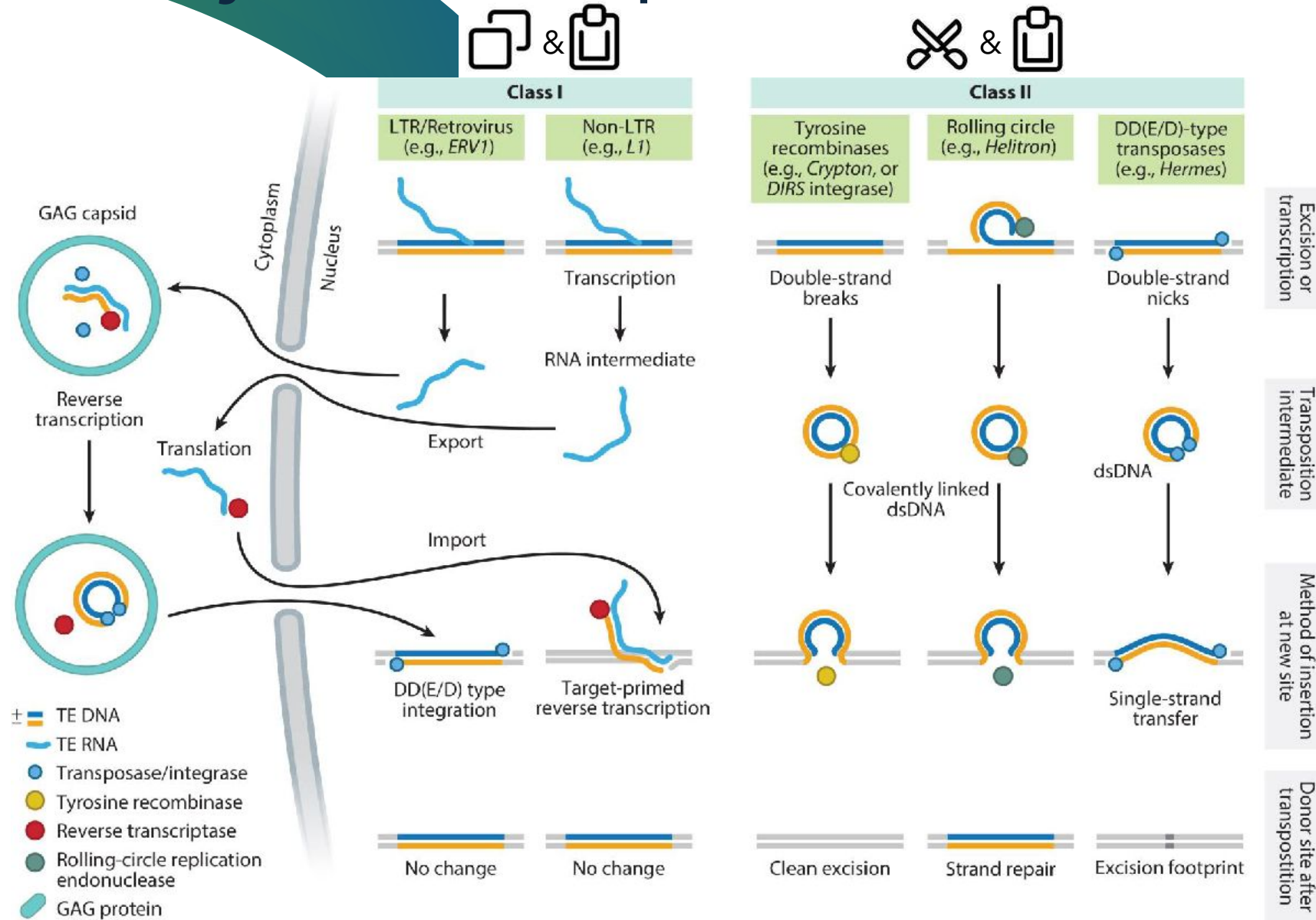
Because
they
can

Main TE categories

- Sequence structure
- Transposition mechanisms
- Effects on genome evolution



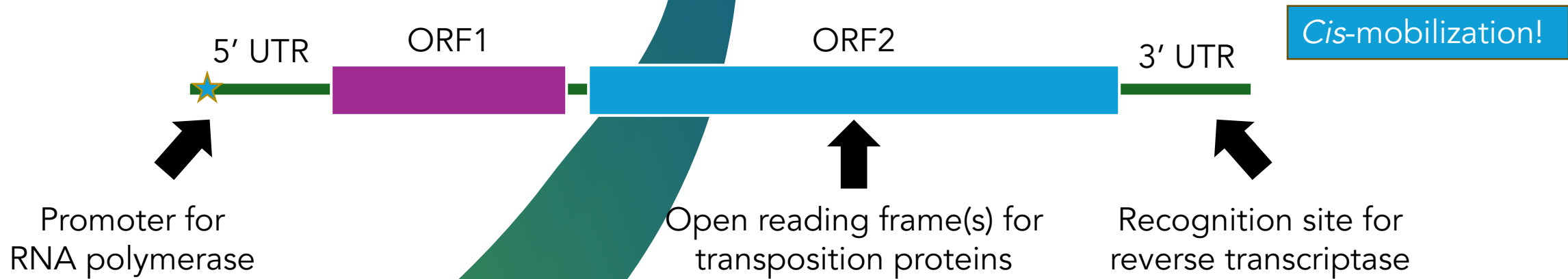
Eukaryotic transposable elements



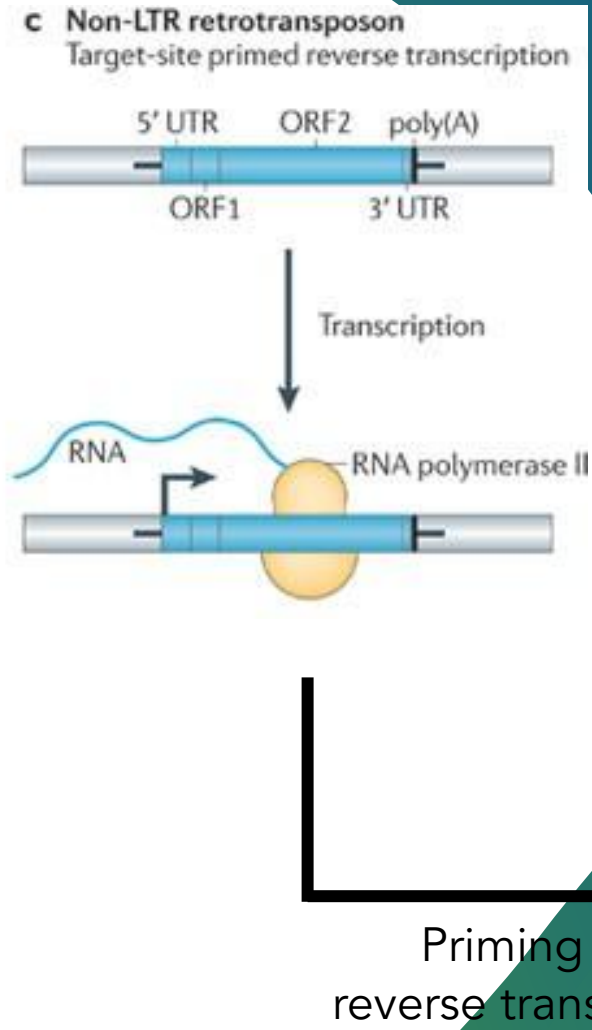
Class I: LINE retrotransposons

Classification		Structure	TSD	Code	Occurrence
Order	Superfamily				
<i>Class I (retrotransposons)</i>					
PLE	<i>Penelope</i>	← RT EN →	Variable	RPP	P, M, F, O
LINE	R2	— RT EN —	Variable	RIR	M
	RTE	— APE RT —	Variable	RIT	M
	<i>Jockey</i>	— ORF1 — APE RT —	Variable	RIJ	M
	L1	— ORF1 — APE RT —	Variable	RIL	P, M, F, O
	I	— ORF1 — APE RT RH —	Variable	RII	P, M, F

RT - retrotranscriptase
 EN - endonuclease
 RH - RNase H
 APE - DNA (apurinic/apyrimidinic site) endonuclease

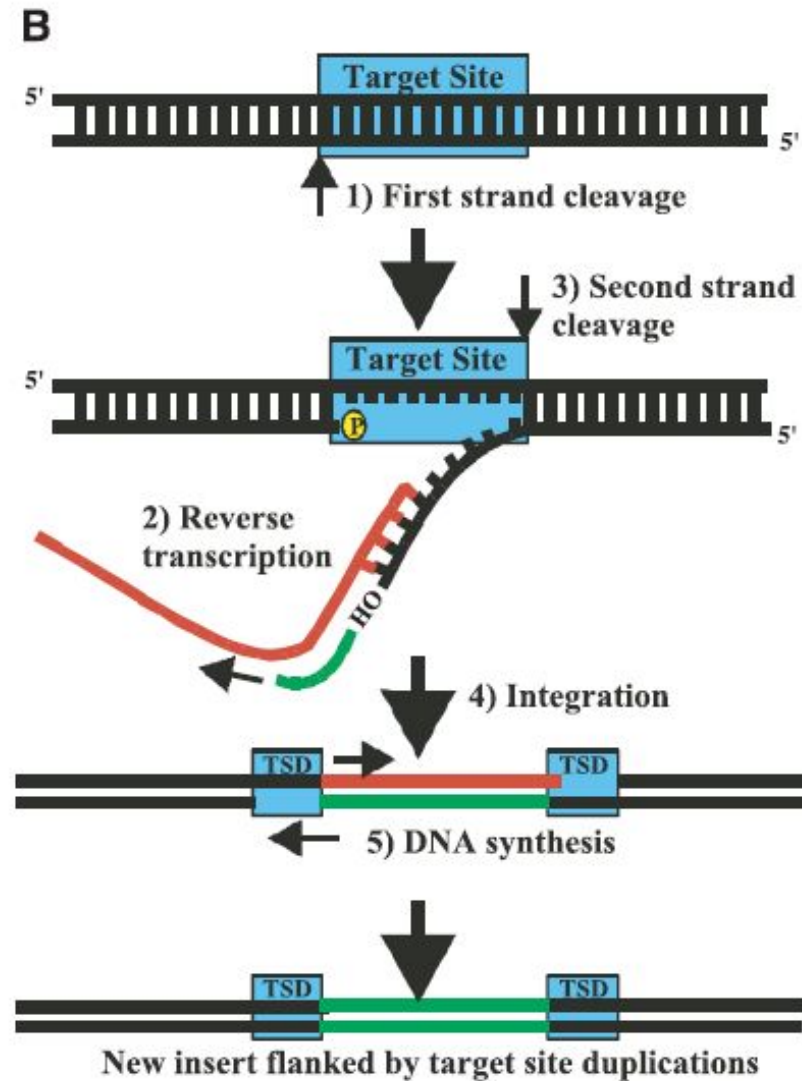


Target-primed reverse transcription (TPRT)



TPRT often undergoes premature 5' truncation and loss of promoters and/or protein

Target site duplications

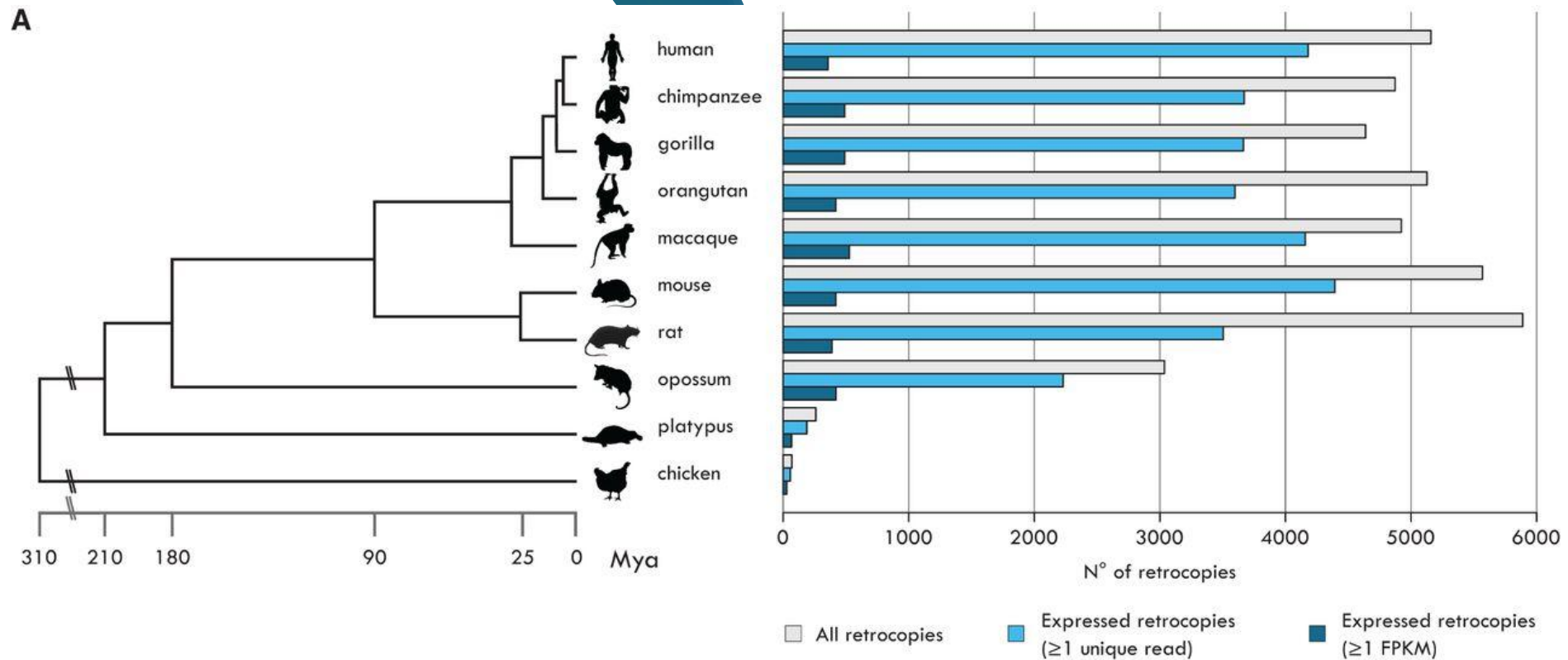
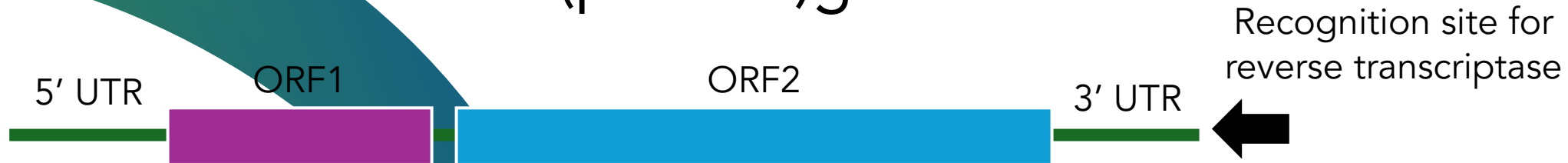


The length of TSDs is important for classification

The length is variable for LINEs but can be of specific lengths for other types of elements

TSDs are the hallmark of most (retro)transposons

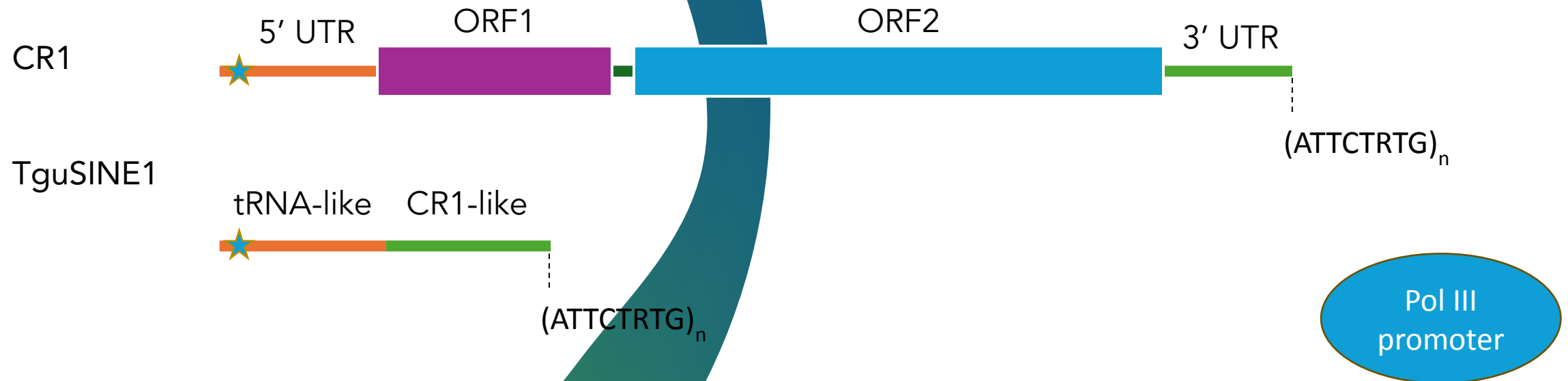
L1 and retro(pseudo)genes



Retrogenes occur when LINE RT recognizes the poly-A

Class I: SINE retrotransposons

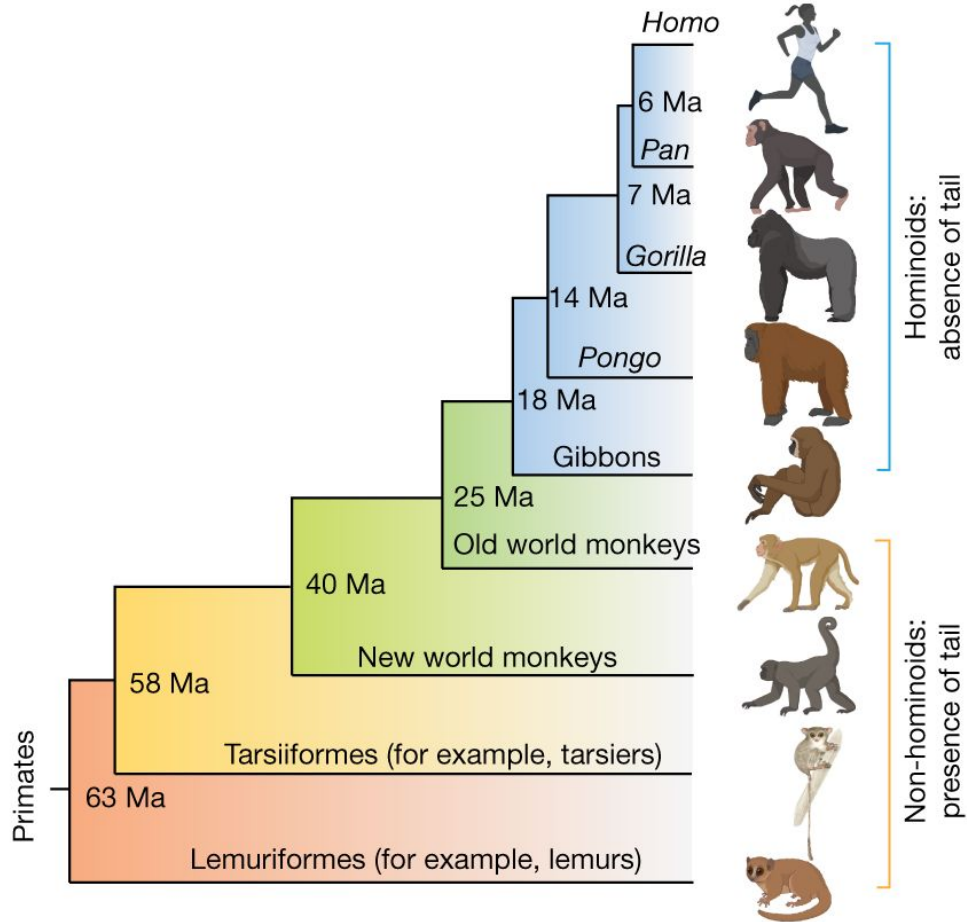
Classification		Structure	TSD	Code	Occurrence
Order	Superfamily				
Class I (retrotransposons)					
SINE	tRNA		Variable	RST	P, M, F
	7SL		Variable	RSL	P, M, F
	5S		Variable	RSS	M, O



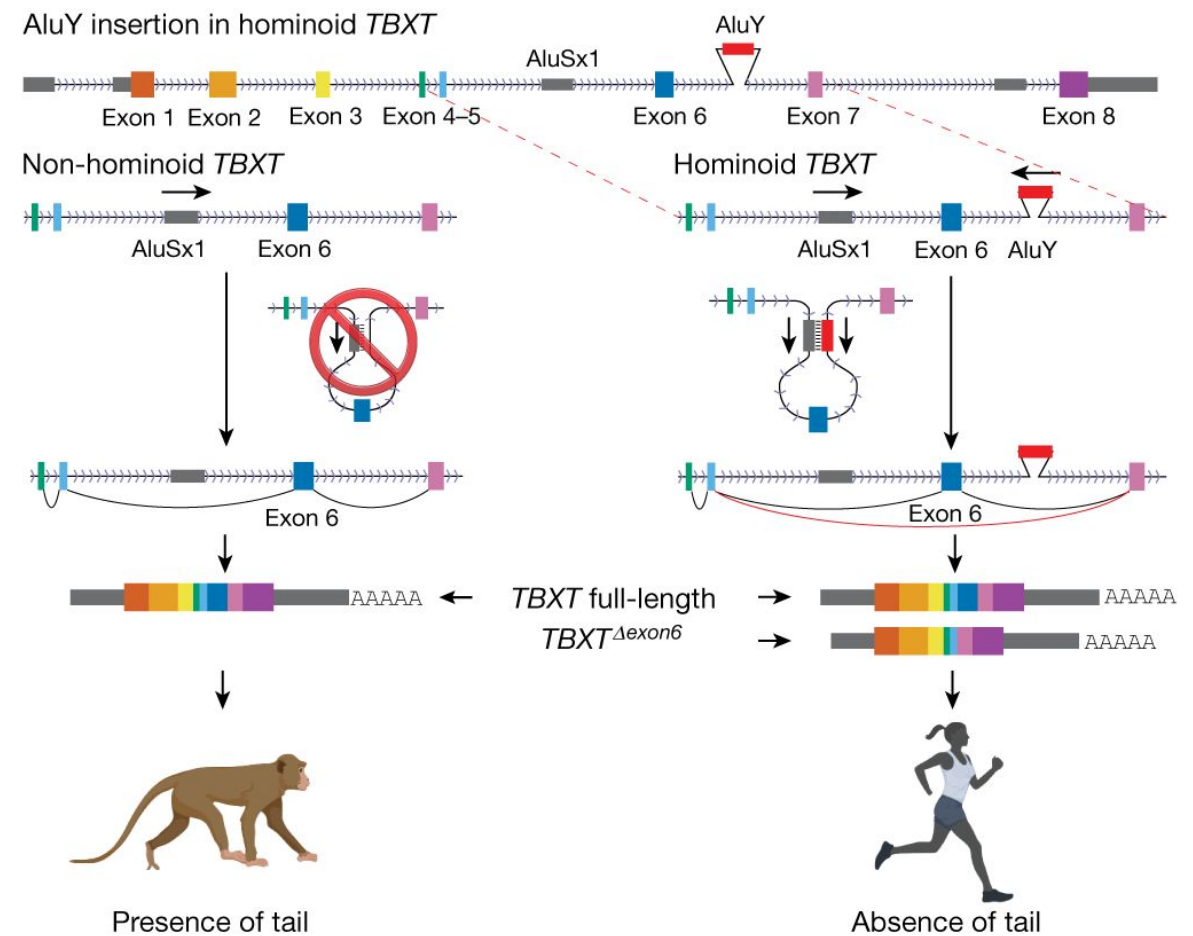
SINEs use the LINE protein machinery to move and replicate – *trans-mobilization*!

SINE *Alu* and alternative splicing

a



c

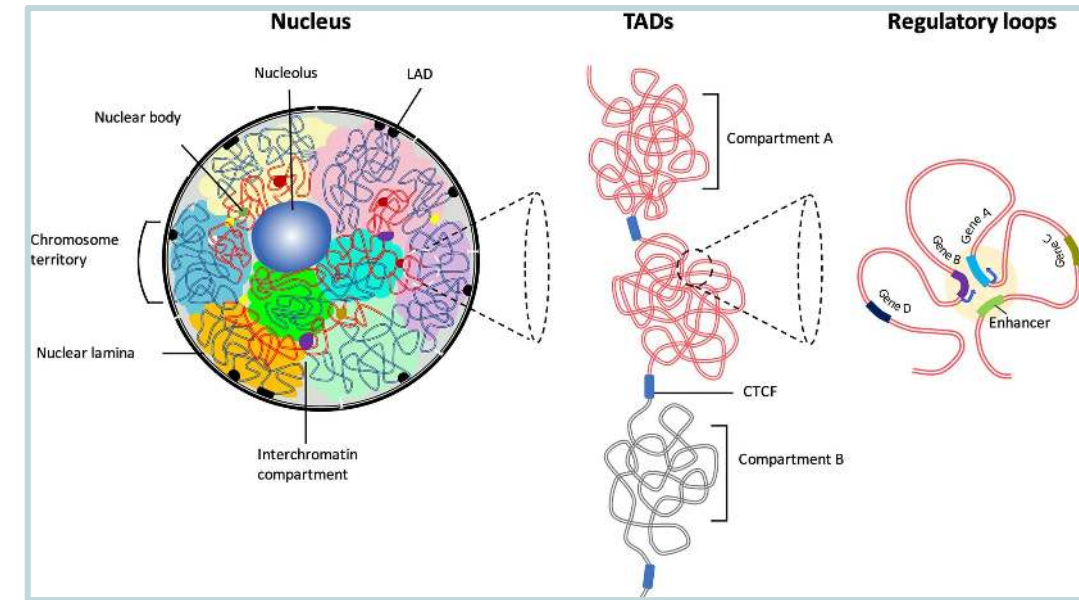
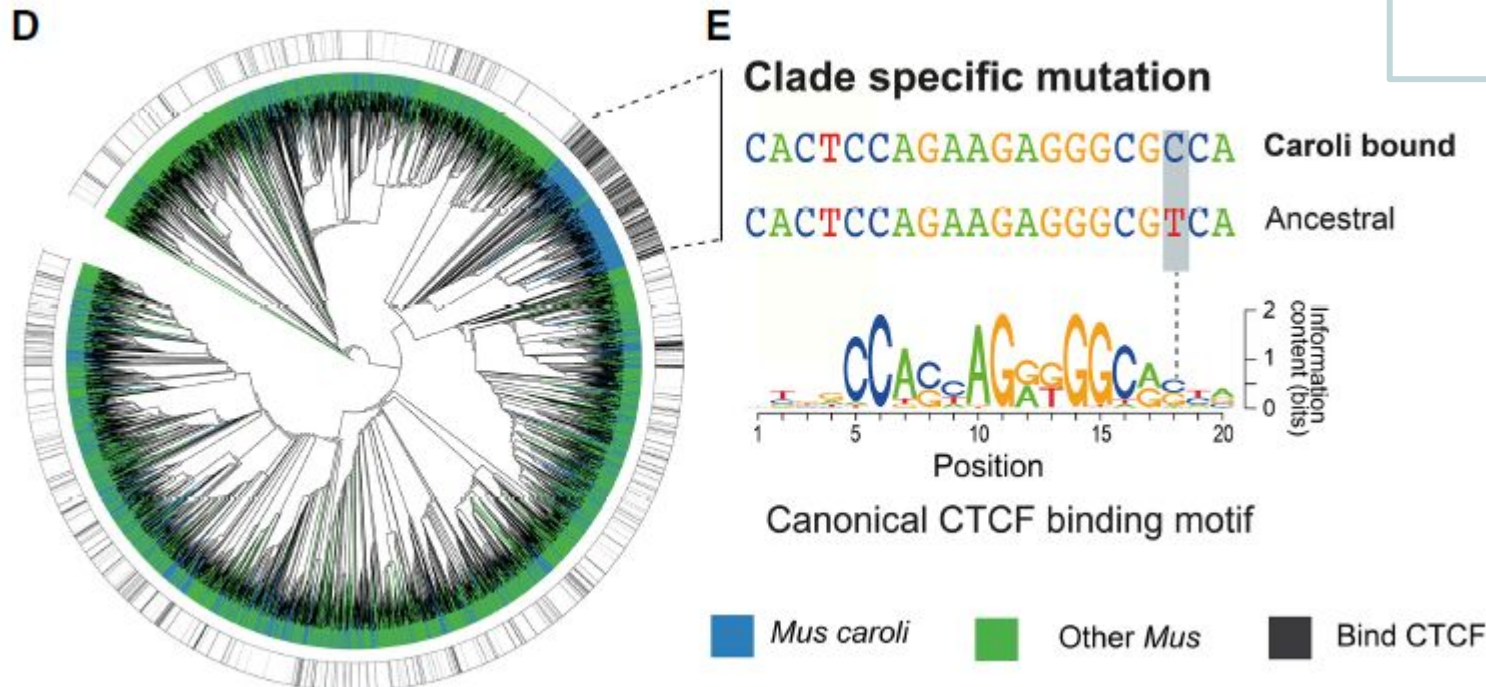


SINEs and 3D genome architecture

Research

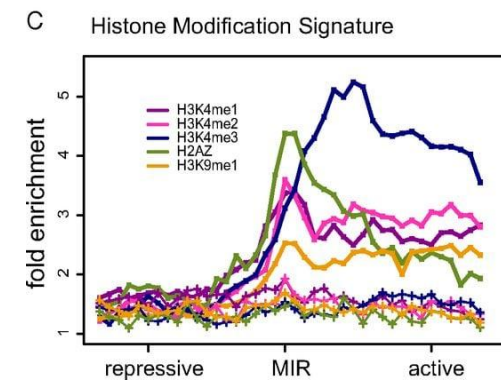
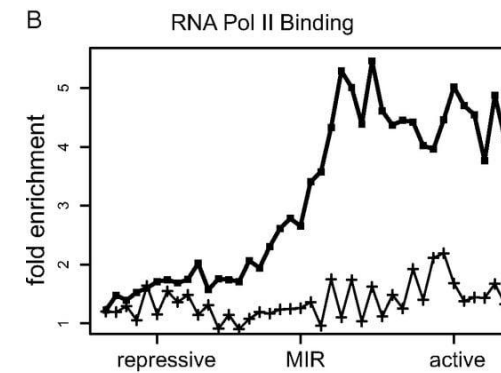
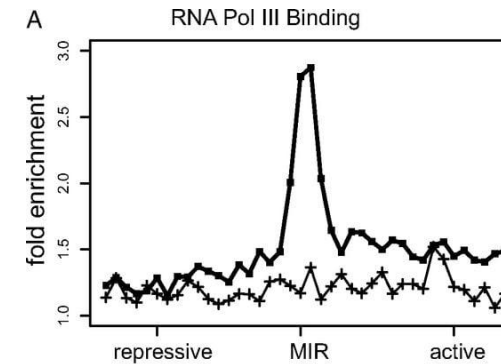
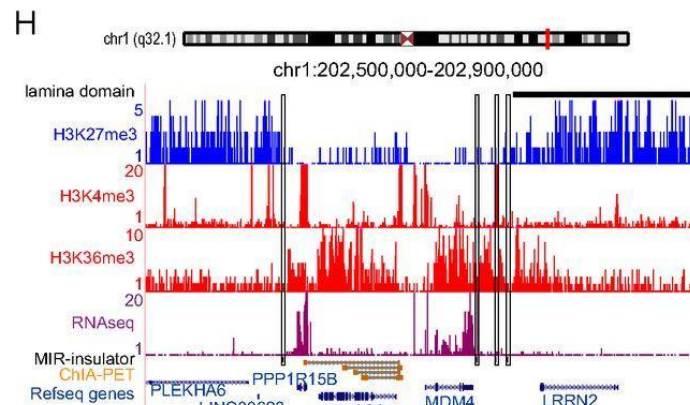
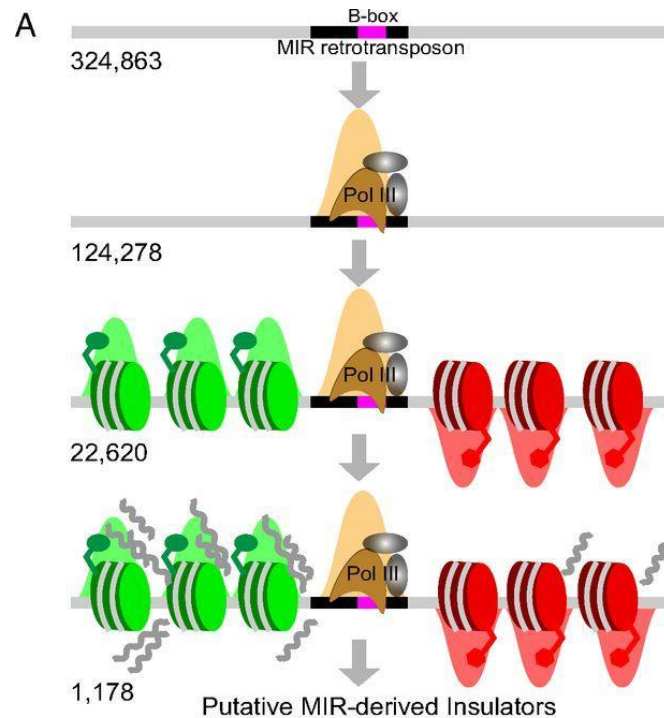
Repeat associated mechanisms of genome evolution and function revealed by the *Mus caroli* and *Mus pahari* genomes

David Thybert,^{1,2} Maša Roller,¹ Fábio C.P. Navarro,³ Ian Fiddes,⁴ Ian Streeter,¹ Christine Feig,⁵ David Martin-Galvez,¹ Mikhail Kolmogorov,⁶ Václav Janoušek,⁷



One SNP in the *Mus caroli* lineage turned SINE B2 into CTCF binding sites

Ancient SINEs became conserved elements and insulators



Class I: LTR retrotransposons

Classification		Structure	TSD	Code	Occurrence
Order	Superfamily				
Class I (retrotransposons)					
LTR	Copia	→ GAG AP INT RT RH →	4-6	RLC	P, M, F, O
	Gypsy	→ GAG AP RT RH INT →	4-6	RLG	P, M, F, O
	Bel-Pao	→ GAG AP RT RH INT →	4-6	RLB	M
	Retrovirus	→ GAG AP RT RH INT ENV →	4-6	RLR	M
	ERV	→ GAG AP RT RH INT ENV →	4-6	RLE	M

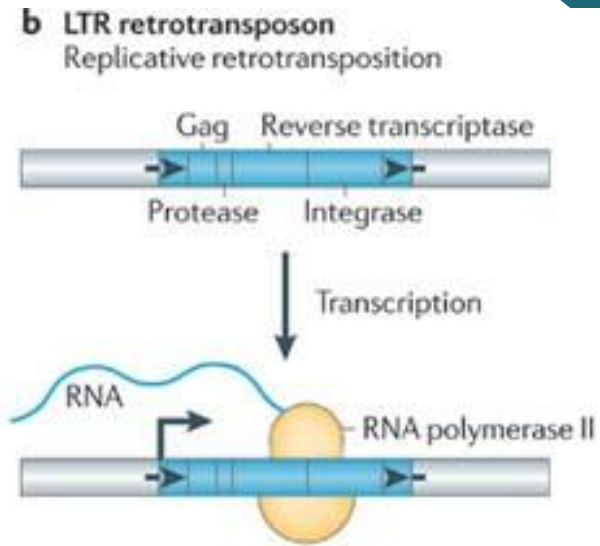
GAG – capsid protein
 AP – aspartic proteinase
 RT - retrotranscriptase
 RH – RNase H
 INT - integrase
 ENV – envelope protein



Non-allelic homologous recombination (NAHR)



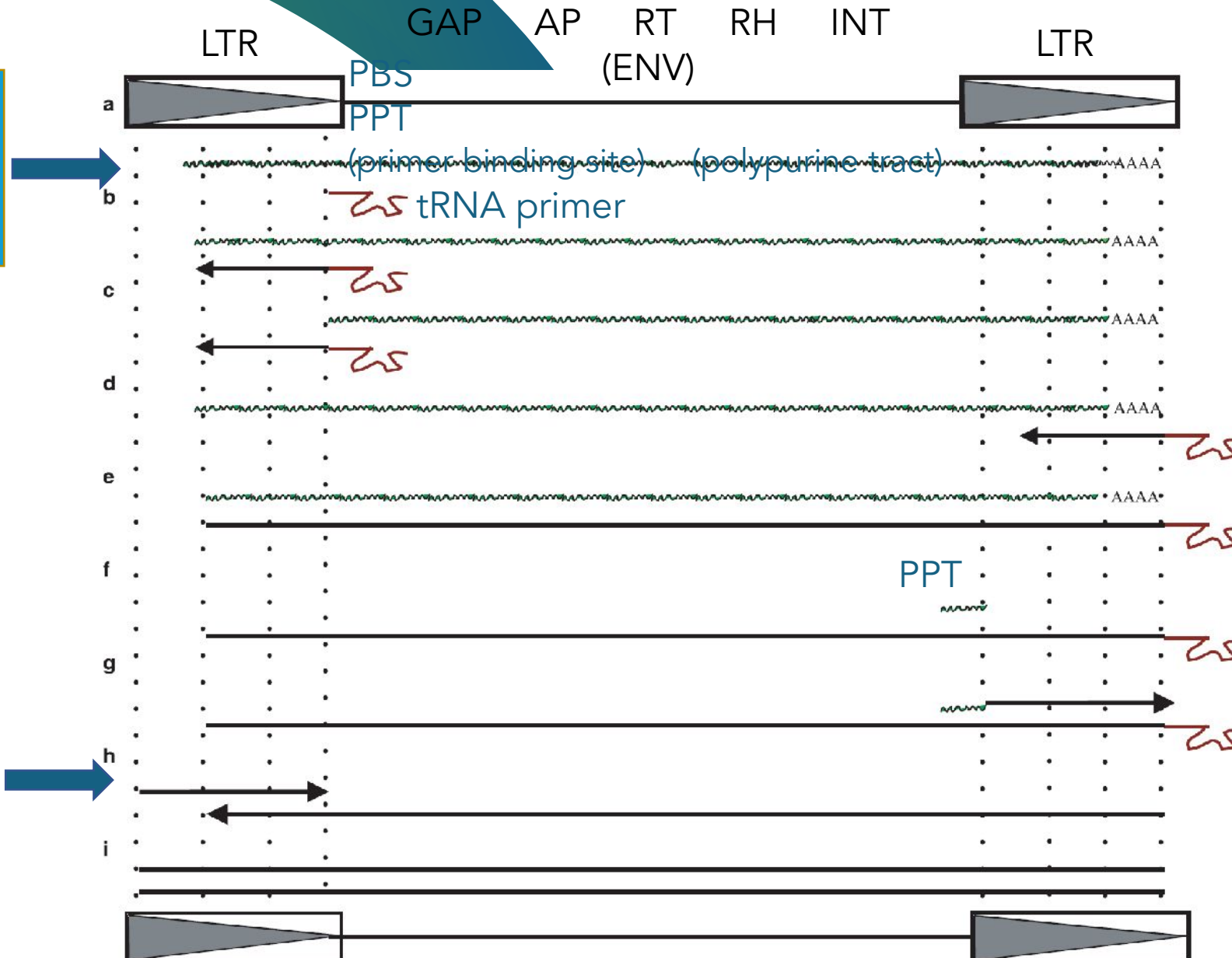
Replicative retrotransposition



Reverse transcription

LTRs are essential for retrotransposition

Transcription starts from an internal promoter!



Insertion in the genome

Transcribed LTR retrotransposon

The synthesis of ds-DNA copy happens within the viral-like particle

Complete new copy ready to be integrated in the genome

The second LTR ensure the restoring of the entire element

LTRs and coloration



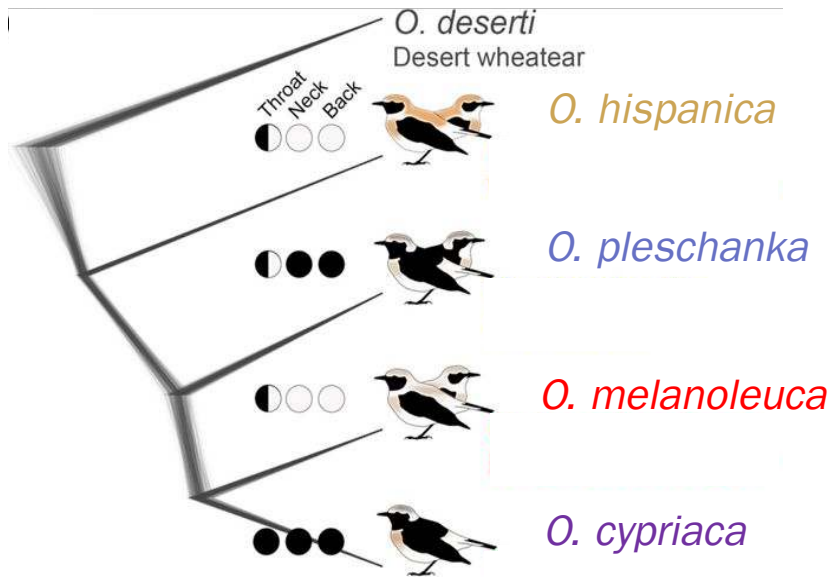
Dave
Lutgen



Madeline
Chase



Fritjof
Lammers



Throat coloration

white
black

white

white

black

black w/

white

black

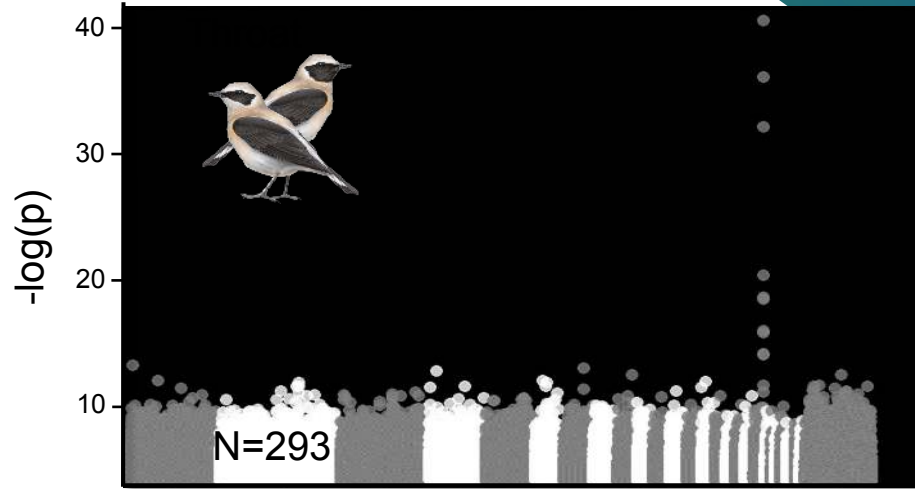
melanoleuca

hybrids

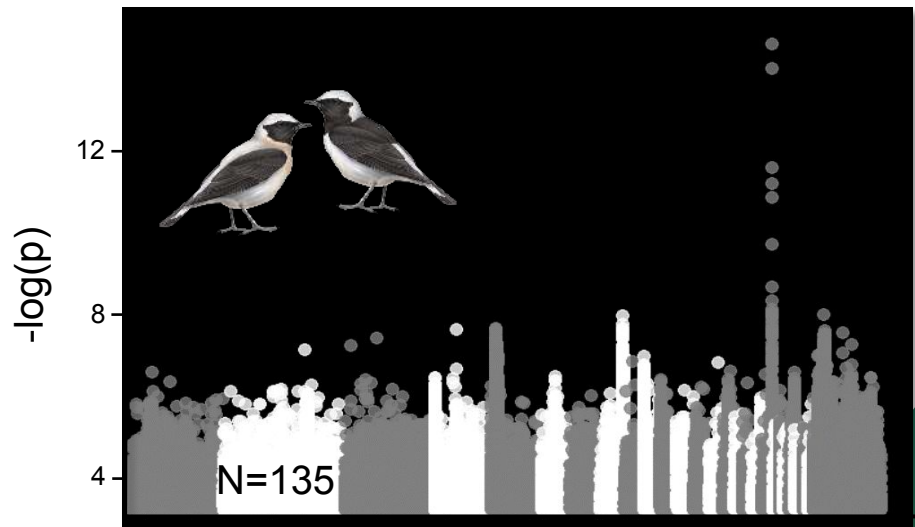
pleschanka



LTRs and coloration



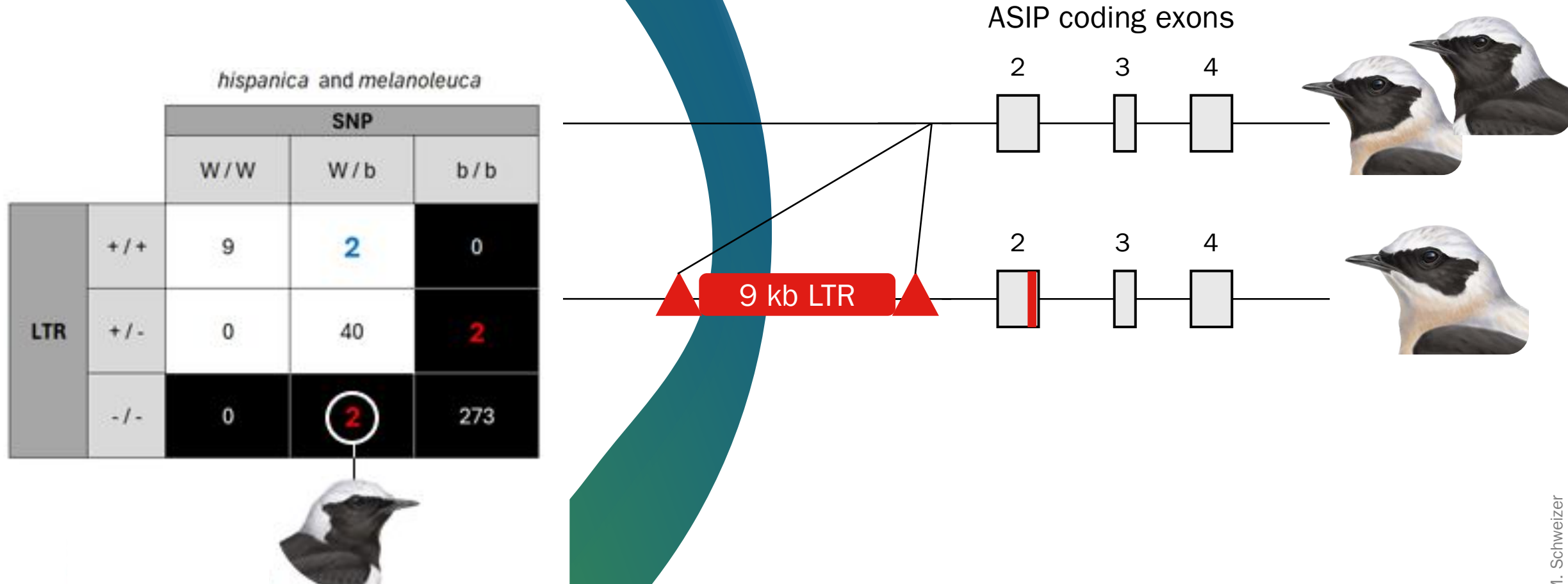
Agouti signalling protein (ASIP): 2 non-synonymous coding SNPs



Agouti signalling protein (ASIP): 17 SNPs up to 45 kb upstream ASIP

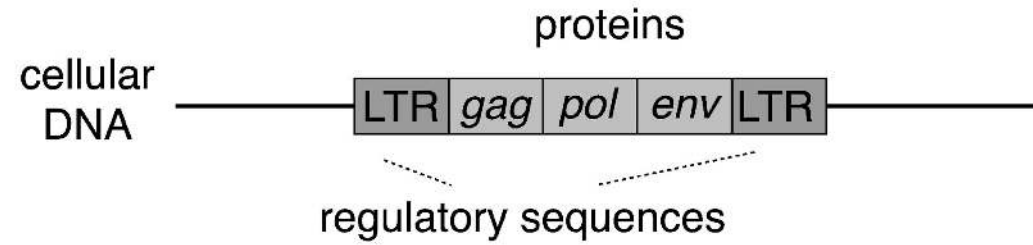
Chromosomes

LTRs and coloration

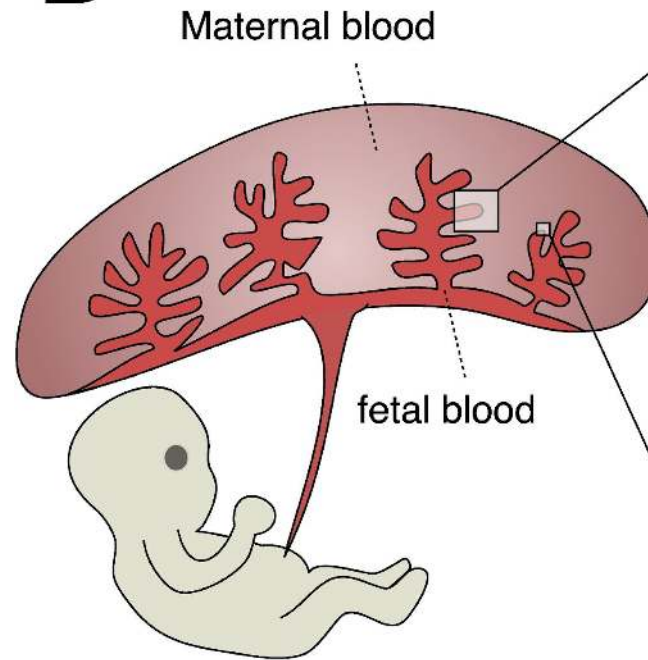


LTRs and placenta

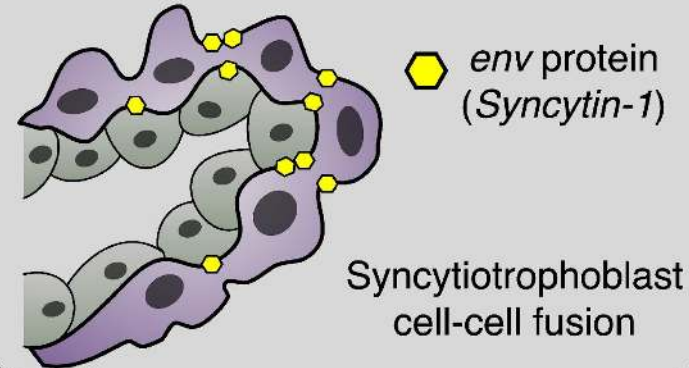
A



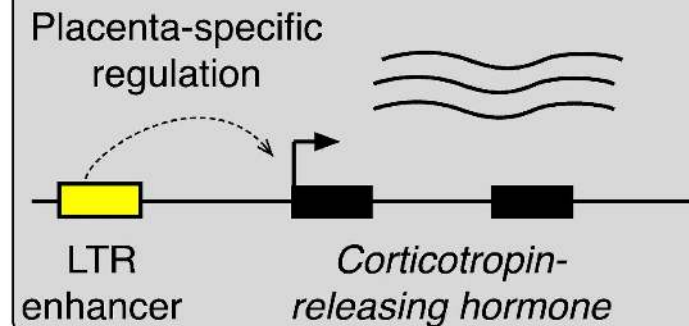
B



Retroviral protein co-option

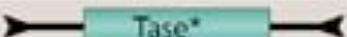
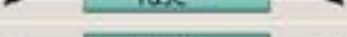
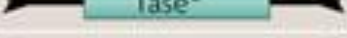


Retroviral LTR co-option



Class II: DNA transposons

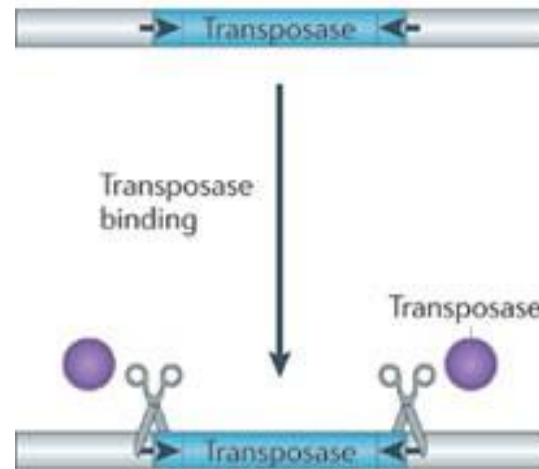
Class II (DNA transposons) - Subclass 1

TIR	Tc1-Mariner		TA	DTT	P, M, F, O
	hAT		8	DTA	P, M, F, O
	Mutator		9-11	DTM	P, M, F, O
	Merlin		8-9	DTE	M, O
	Transib		5	DTR	M, F
	P		8	DTP	P, M
	PiggyBac		TTAA	DTB	M, O
	PIF-Harbinger		3	DTH	P, M, F, O
	CACTA		2-3	DTC	P, M, F



Cut and paste transposition (TIRs)

a DNA transposon
'Cut and paste' TE



Excision

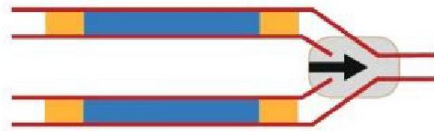


Mobile DNA

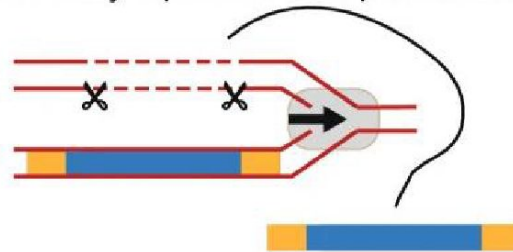
How to increase in copy number?



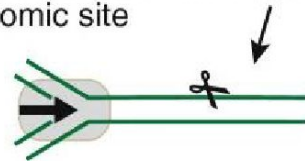
I. DNA replication fork passes transposon



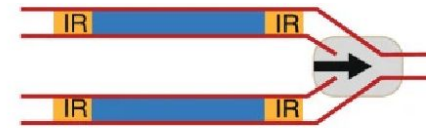
II. Newly replicated transposon is cut out...



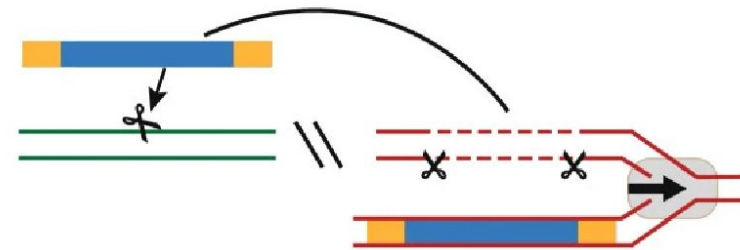
III. ...and inserted into a not-yet replicated genomic site



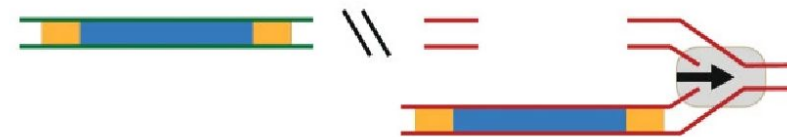
IIII. DNA replication fork passes insertion site



I. Newly replicated transposon is cut out...



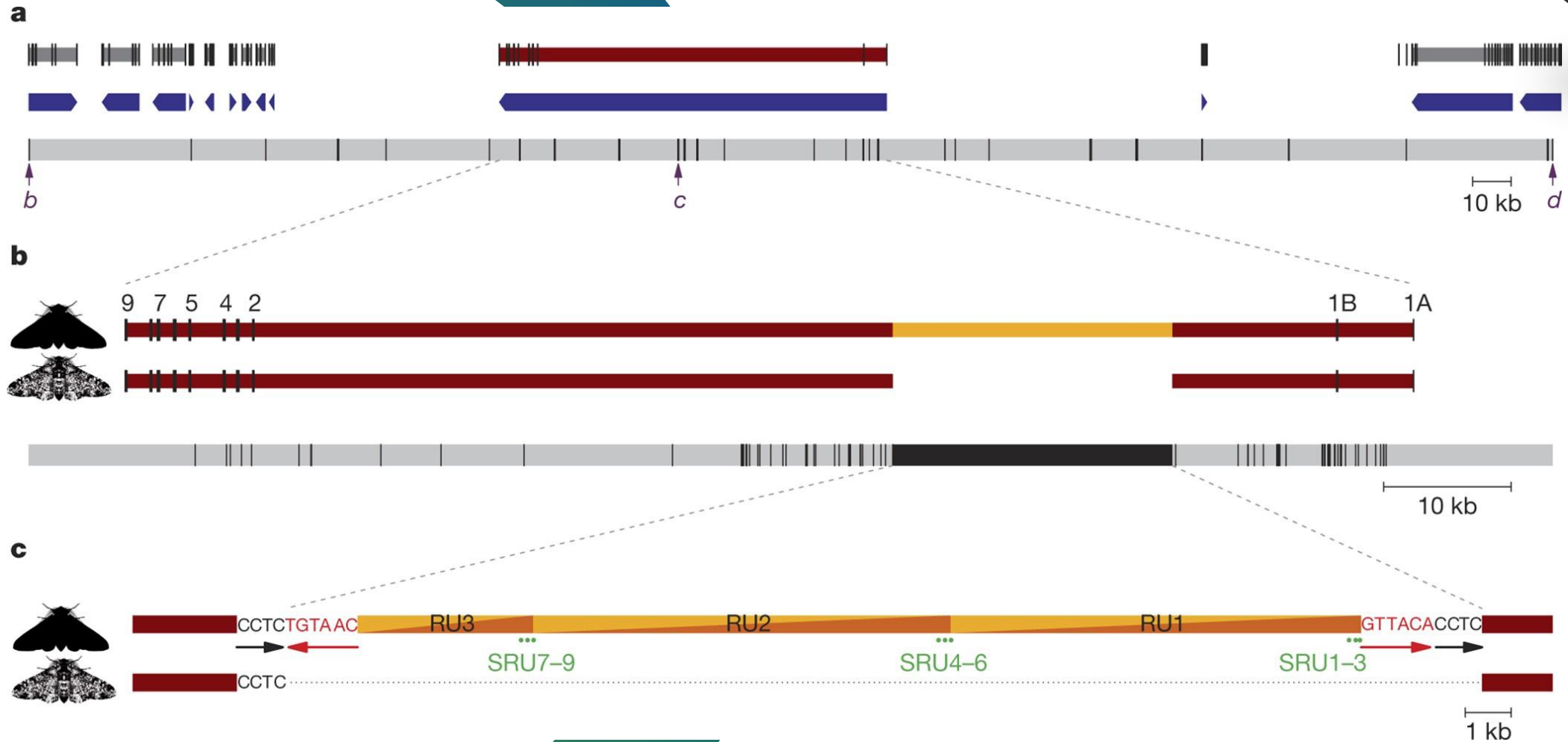
II. ...and transposed into a new locus



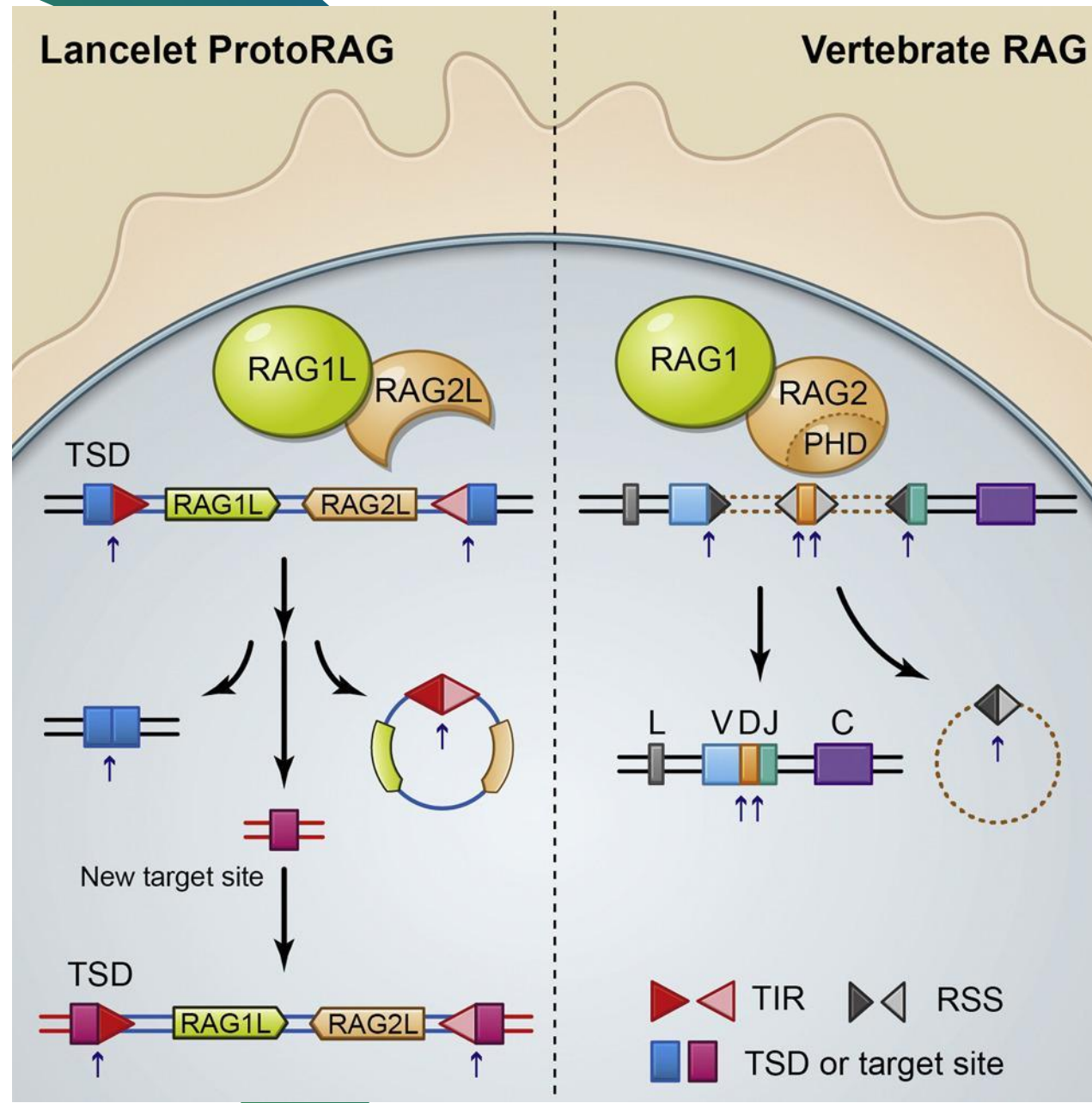
III. Following transposition, the double-stranded break is repaired by homology-dependent DNA repair



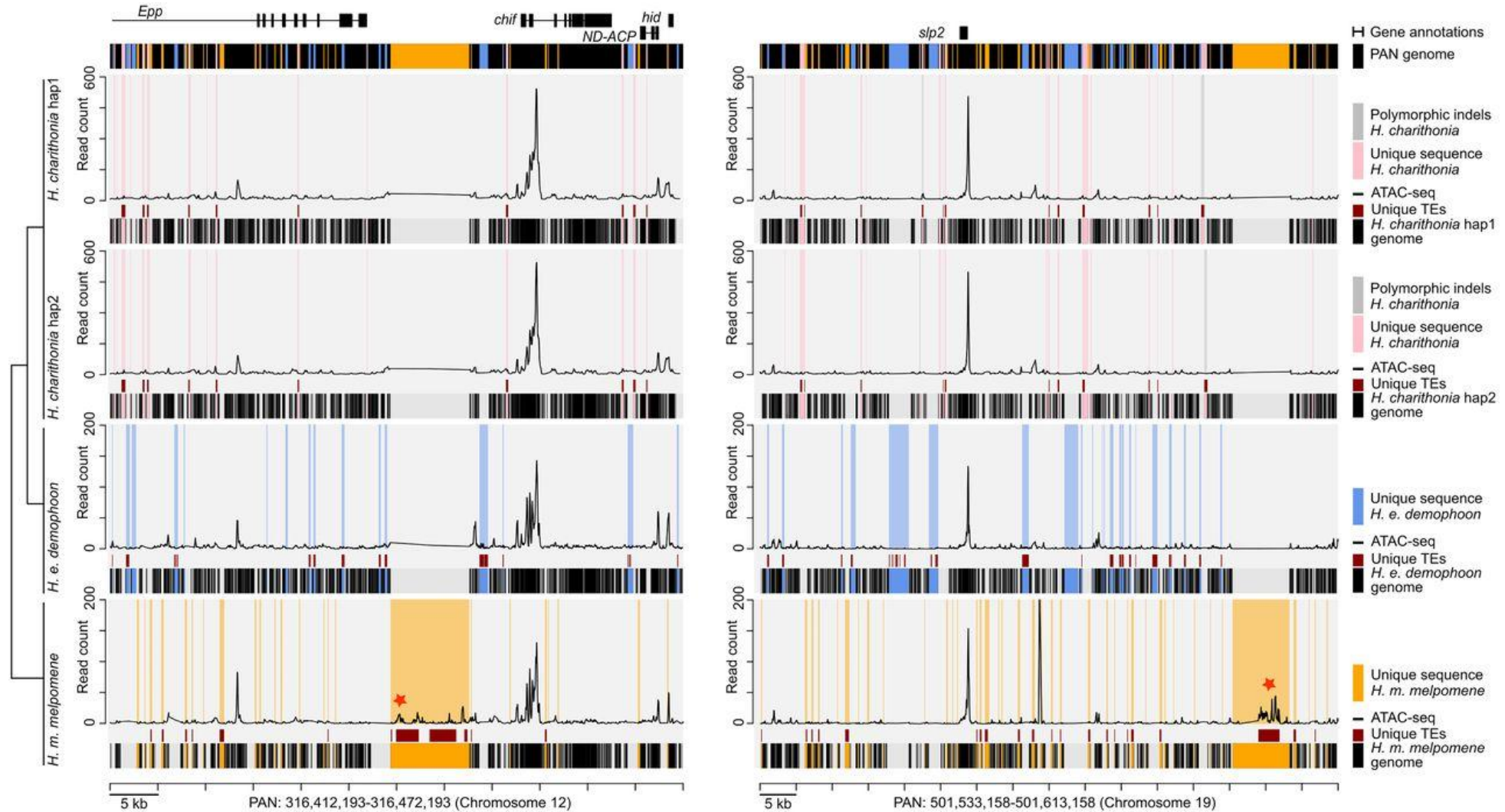
DNA transposons and coloration



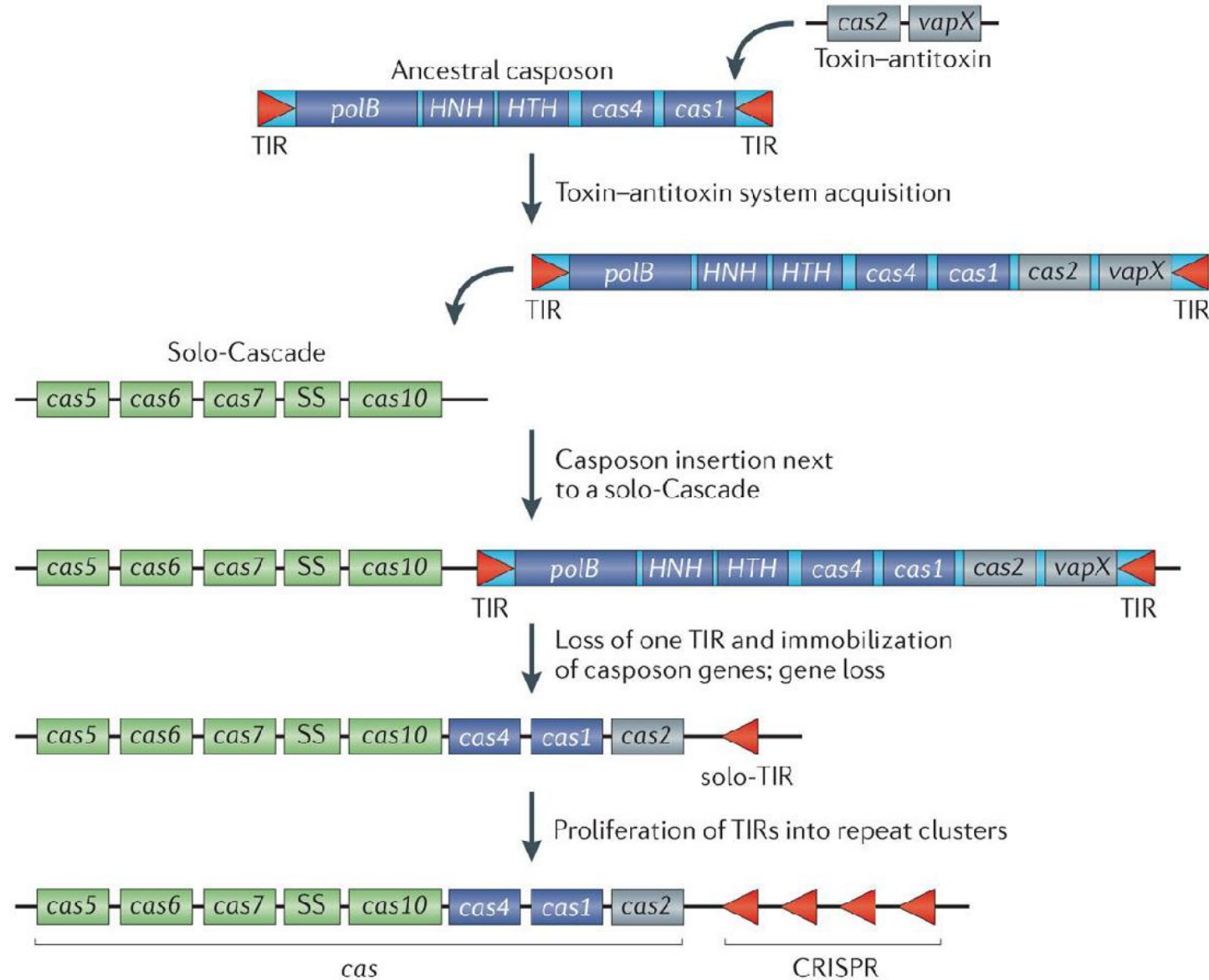
DNA transposons and immune system



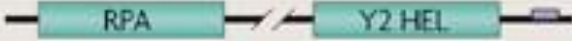
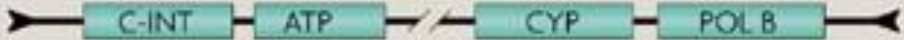
ATAC (Assay for Transposase-Accessible Chromatin)

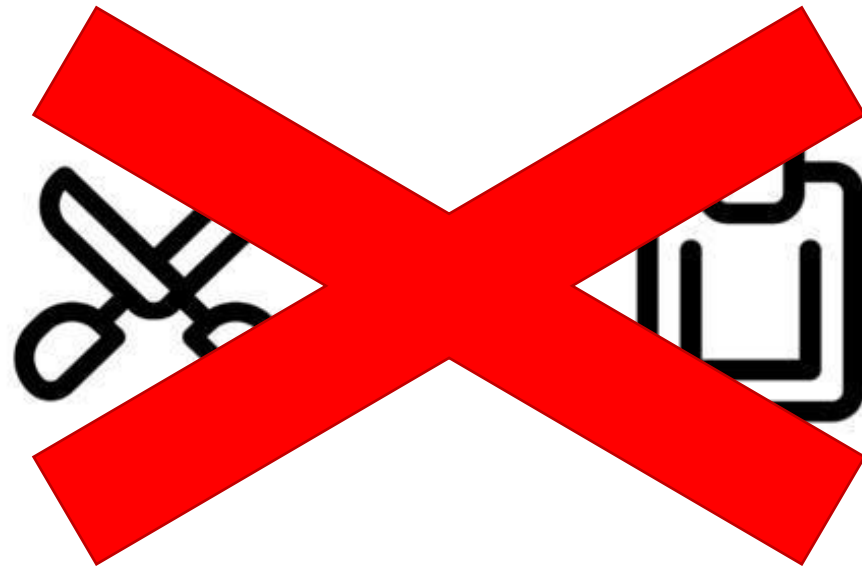


CRISPR-Cas and transposons

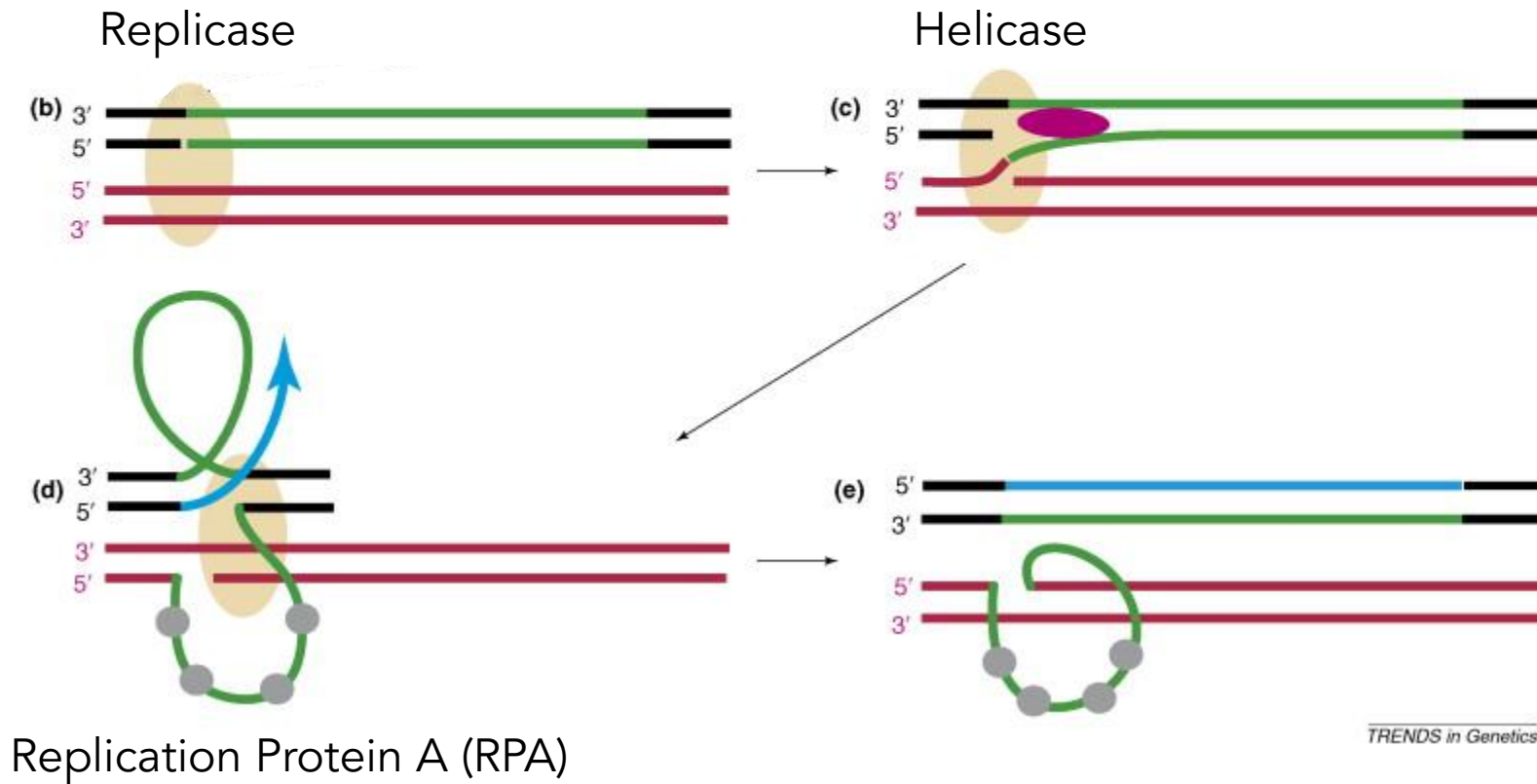


Class II: DNA transposons (subclass 2)

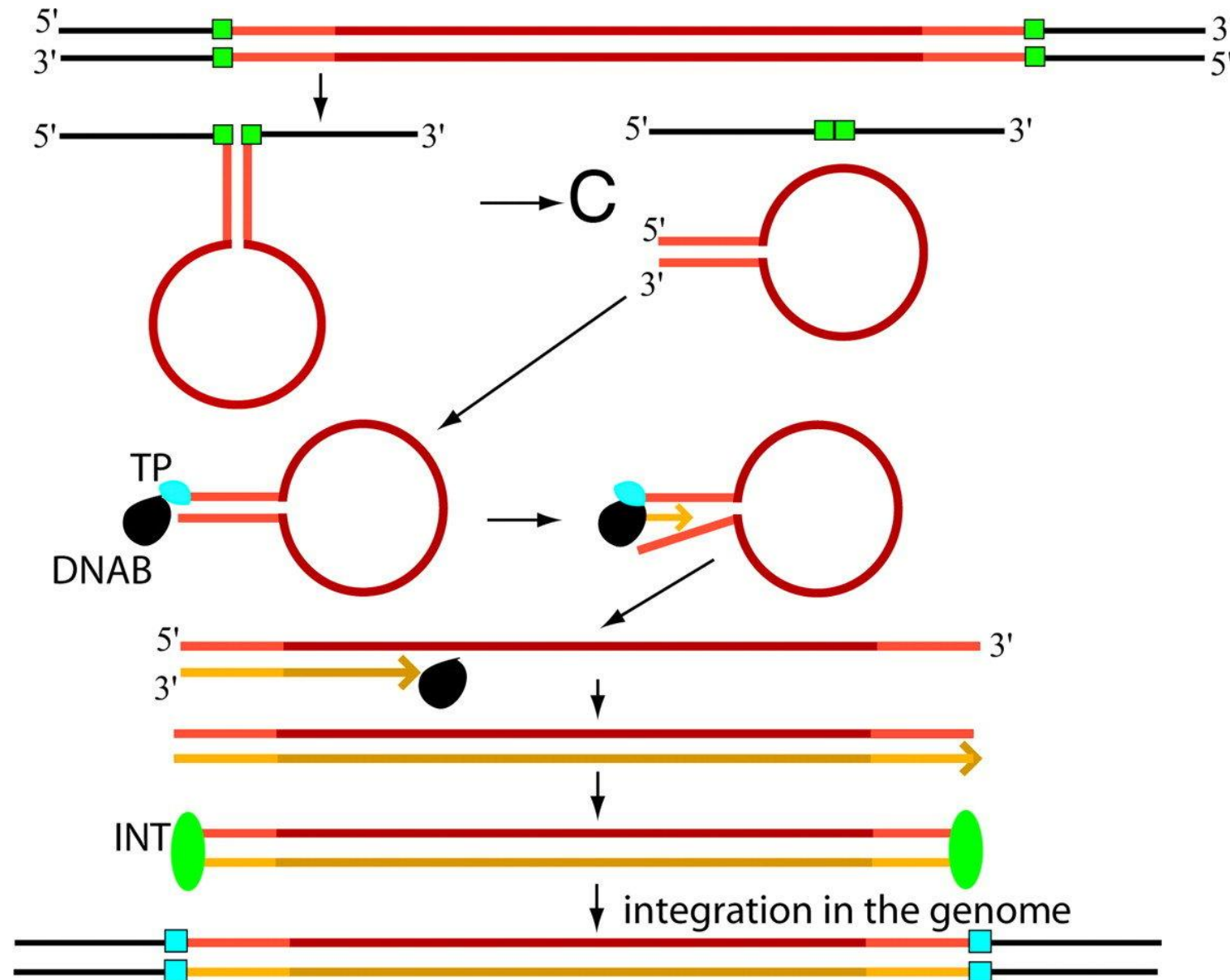
Class II (DNA transposons) - Subclass 2					
Helitron	Helitron		0	DHH	P, M, F
Maverick	Maverick		6	DMM	M, F, O



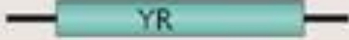
Rolling circle transposition: Helitrons



Self-synthesizing transposition: Mavericks/Polintons



TEs with tyrosine recombinase

Classification		Structure	TSD	Code	Occurrence
Order	Superfamily				
Class I (retrotransposons)					
DIRS	DIRS		0	RYD	P, M, F, O
Class II (DNA transposons) - Subclass 2					
Crypton	Crypton		0	DYC	F

TE integration mechanism occurs via:

- Endonuclease: LINE, SINE, PLE
- DDE-Transposase: TIR
- Integrase: LTR, Maverick/Polinton
- Rep protein: Helitron
- Tyrosine recombinase: DIRS, Crypton

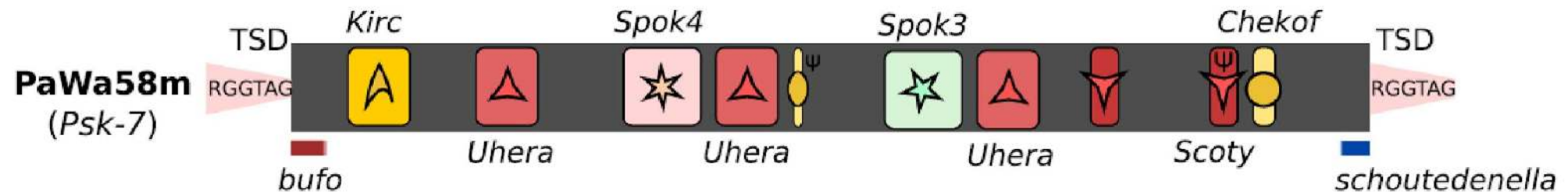
Class I: retrotransposons

Class II: DNA transposons

A hyper-selfish Crypton

Transposable elements carrying meiotic drive genes

YR DNA transposons: *Enterprise* (247 kb, fungus *Podospira anserina*)

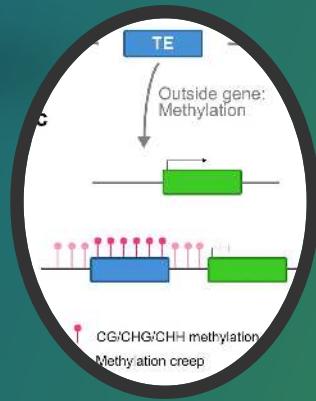


TEs as dynamic mutations

- NAHR



- Methylation



Non-allelic homologous recombination NAHR

Full-length LTR □ solo LTR

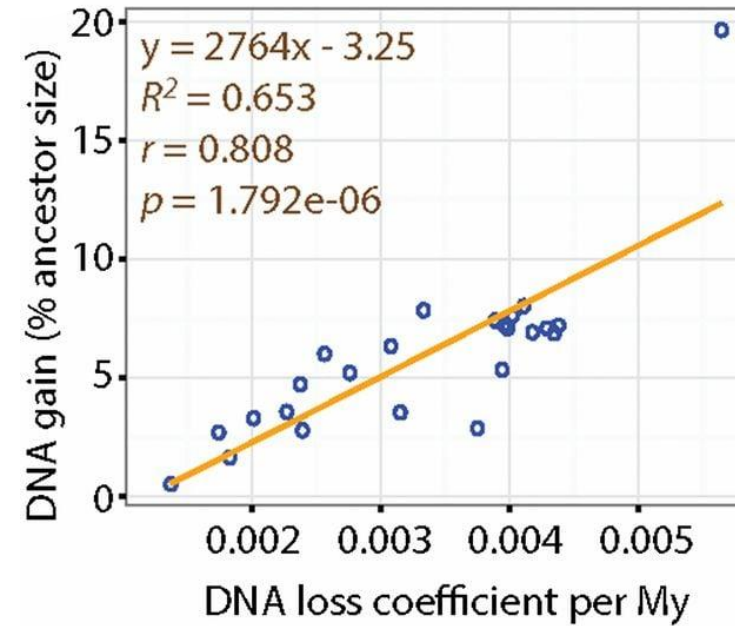
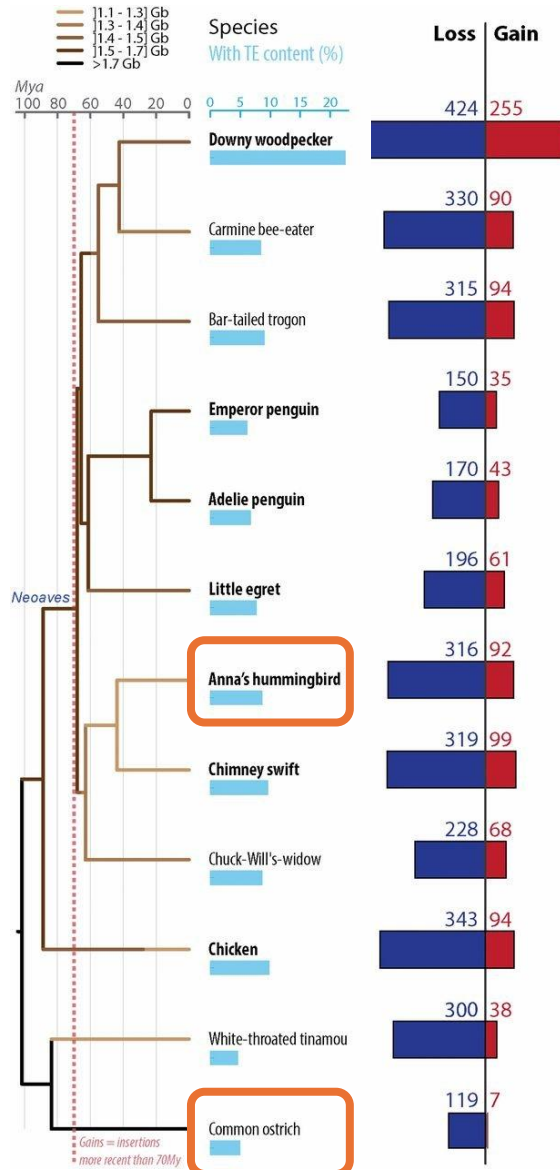


➡ Larger scale deletions and copy number variation

➡ Inversions

Genome size evolution

Accordian model



Consider not only host popgen, but also TE

popgen!

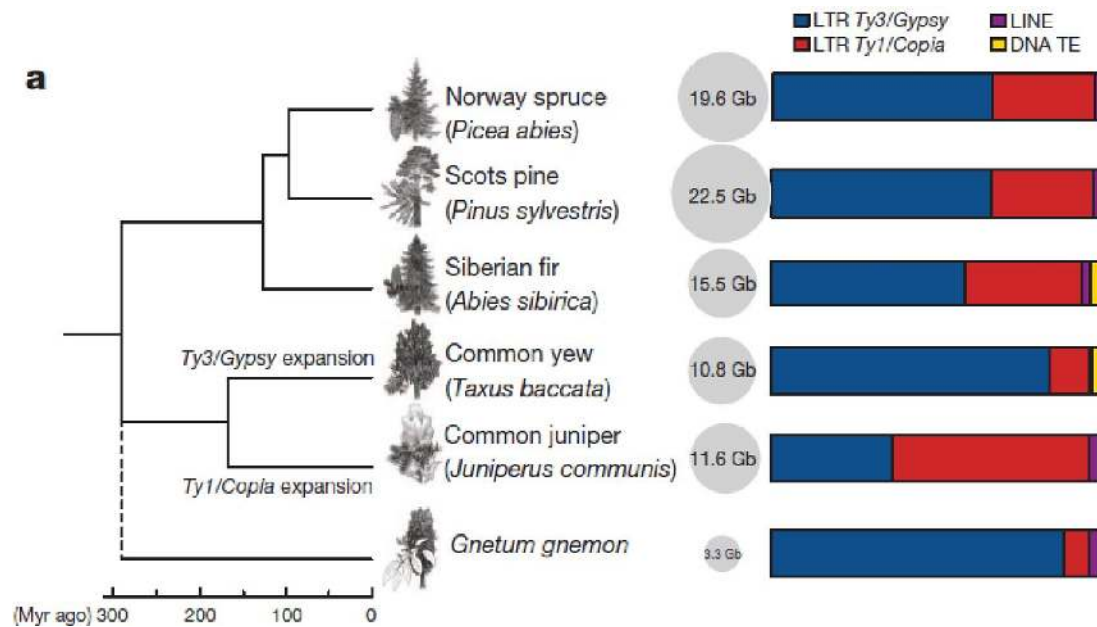
Need to clean up their genomes

VS

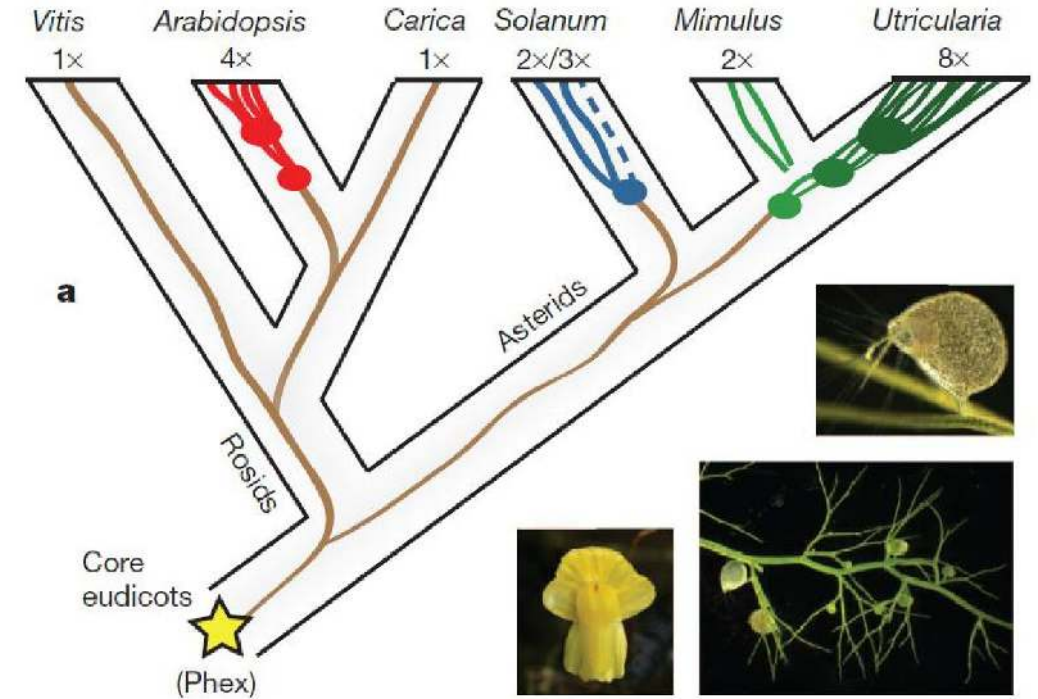
genome structure reflects popgen of the host and TEs

No constraint

Genomic gigantism and dwarfism



Tons of TE activity/accumulation or lack of deletions?

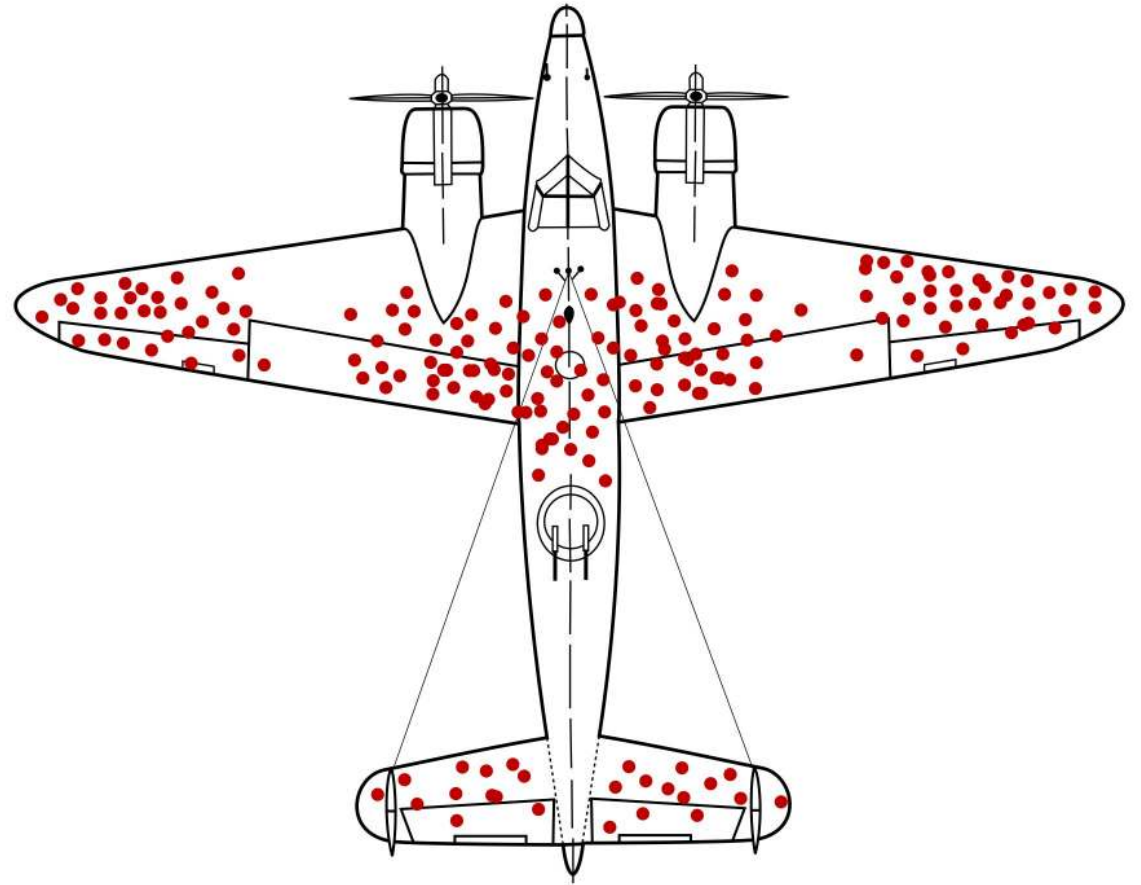


77 Mb genome despite 3 rounds of whole-genome duplications, only 3% repetitive DNA!

Lack of TE activity/accumulation or tons of deletions?

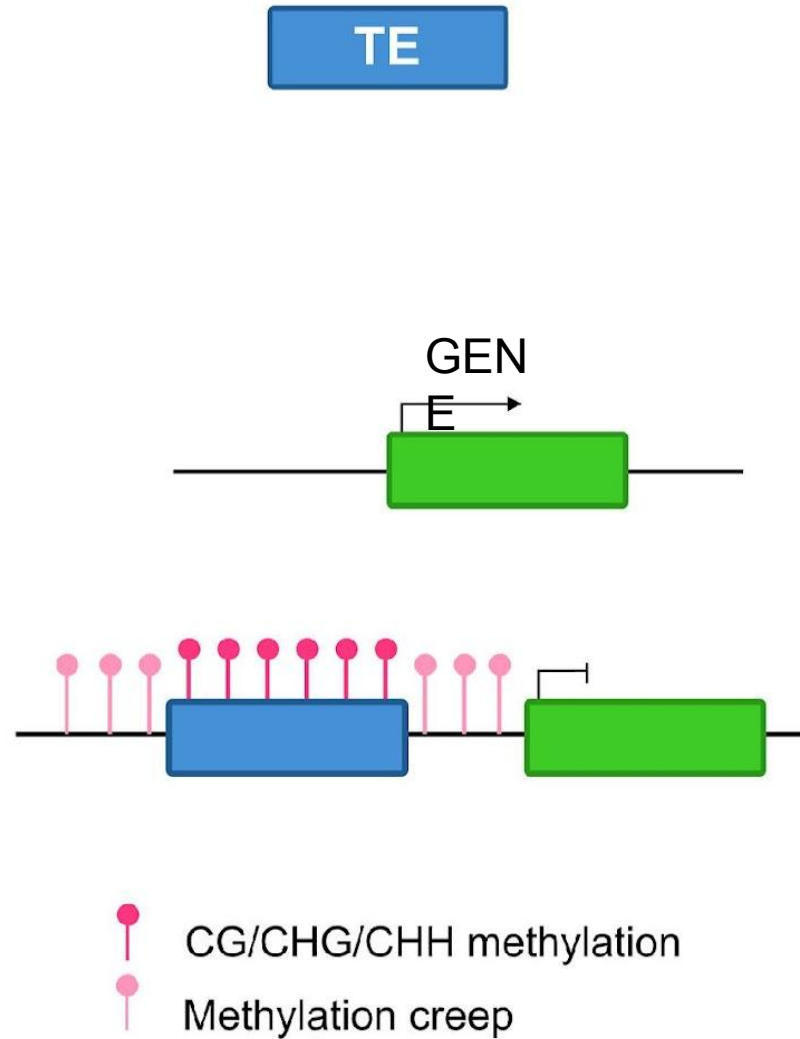
Survivorship bias

- More TEs because of more insertions or more fixations?
- Fewer TEs because of preferential insertion or negative selection?
- Larger genome because of more TE accumulation or fewer deletions?
- Smaller genome because of lower TE accumulation or more deletions?



pattern !=
process

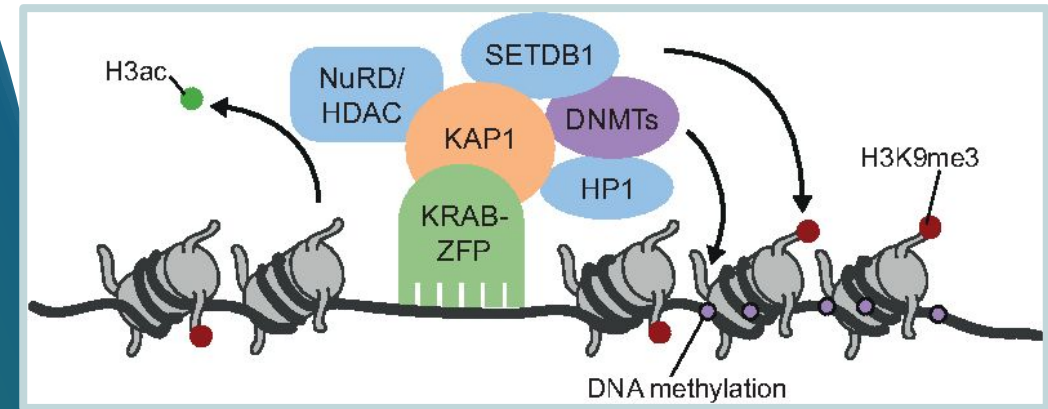
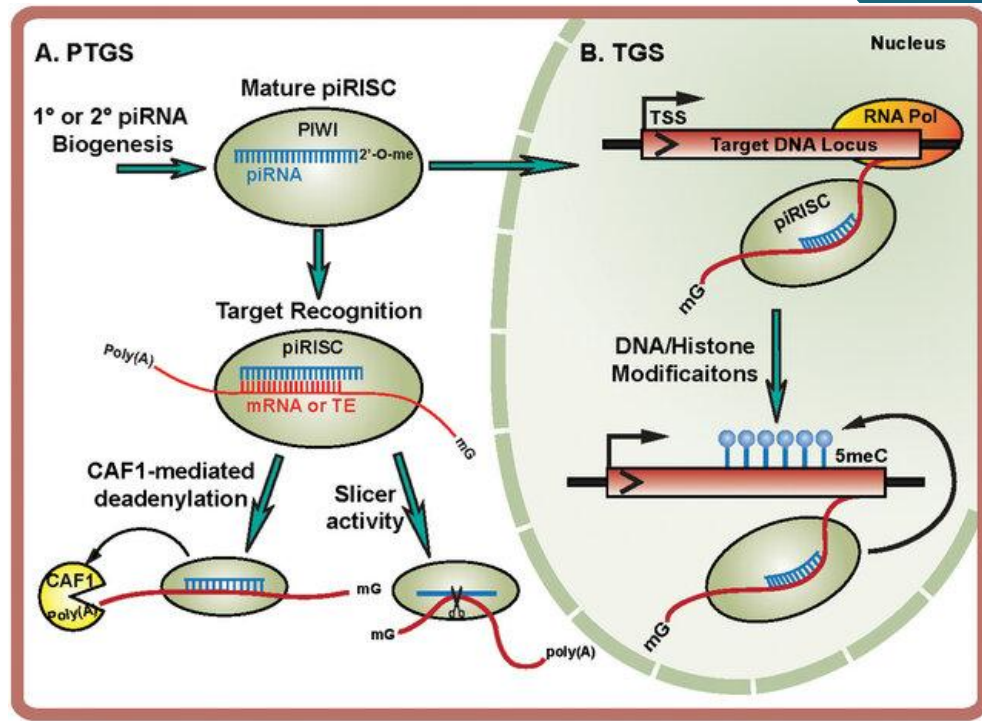
Methylation spillover



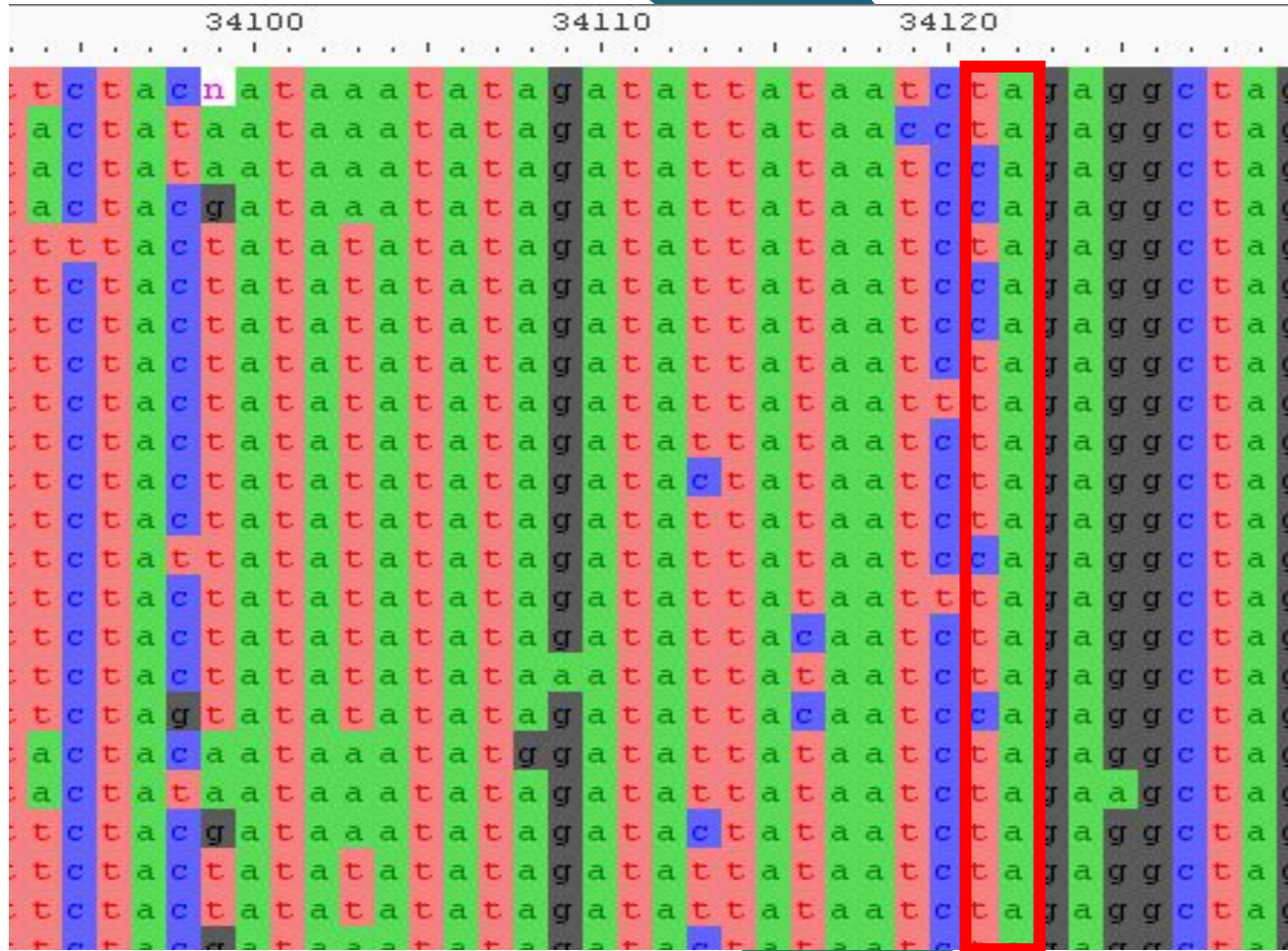
Silencing mechanisms

- DNA level
- Post-transcriptional level

DNA methylation, KRAB and PIWI



Repeat induced point mutation (RIP)



C -> T mutations

In *Neurospora*, mutations that are induced by RIP occur preferentially in CpA dinucleotides

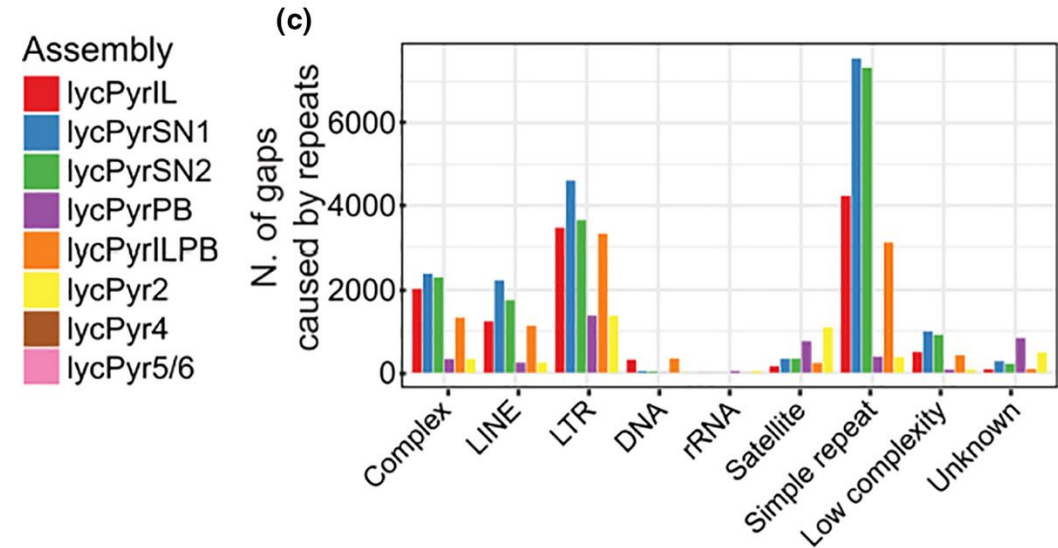
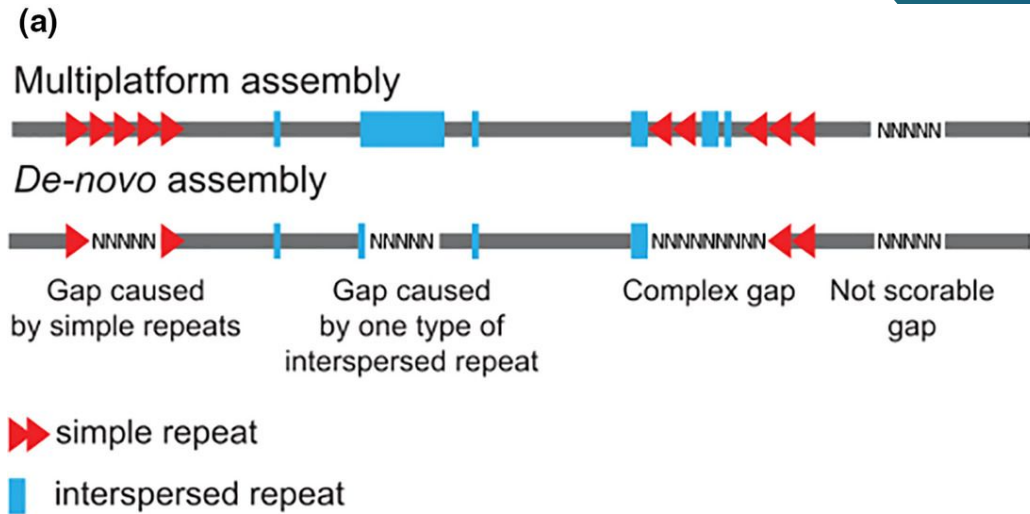
Only a few repeated genes are known to survive RIP

Characterisation and annotation of transposable elements

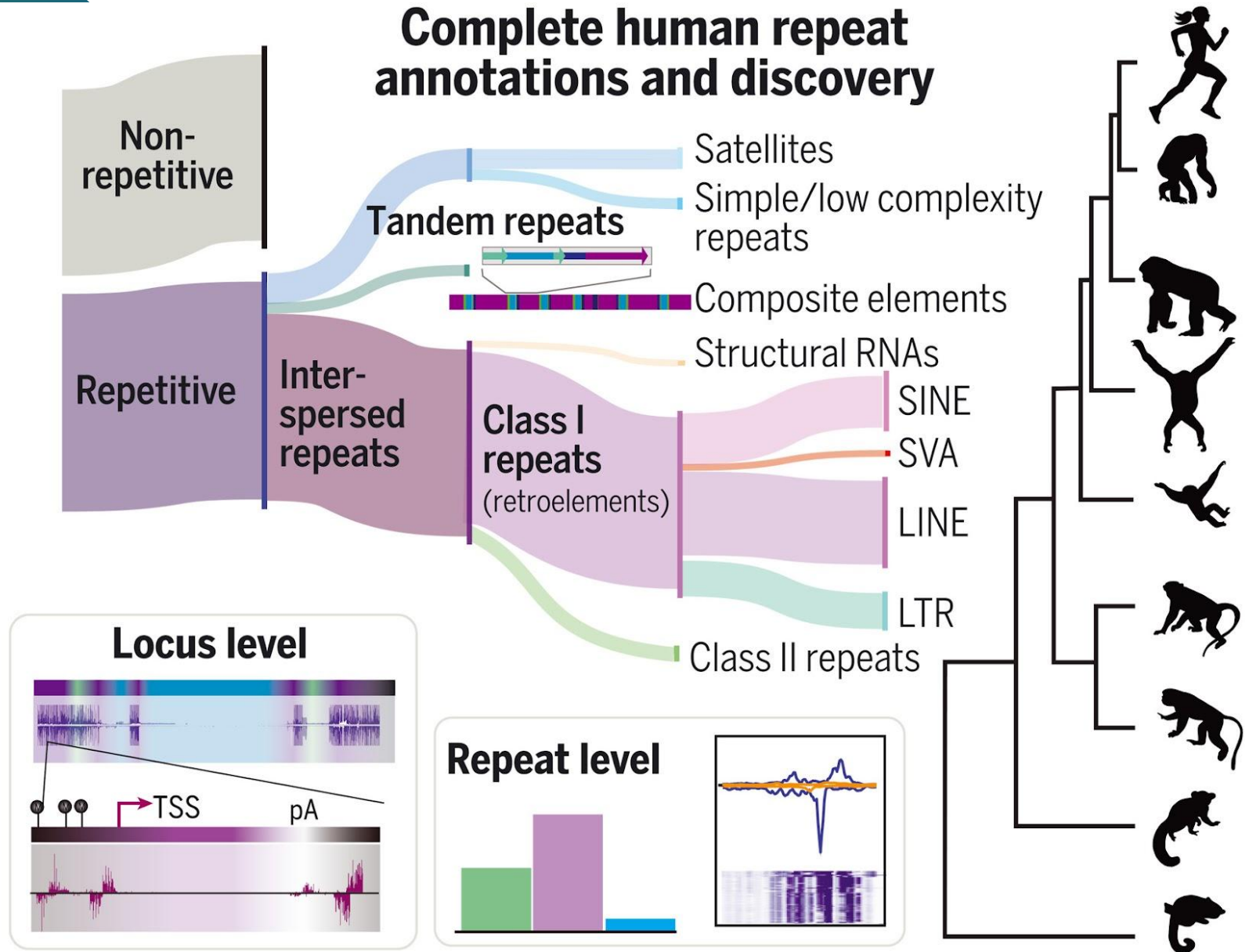
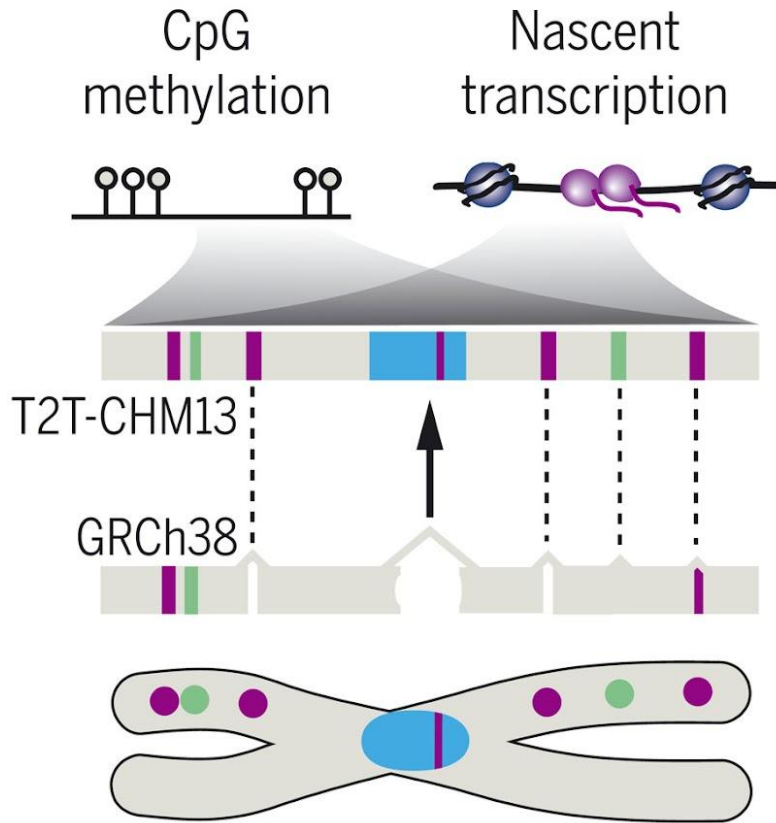
Valentina Peona

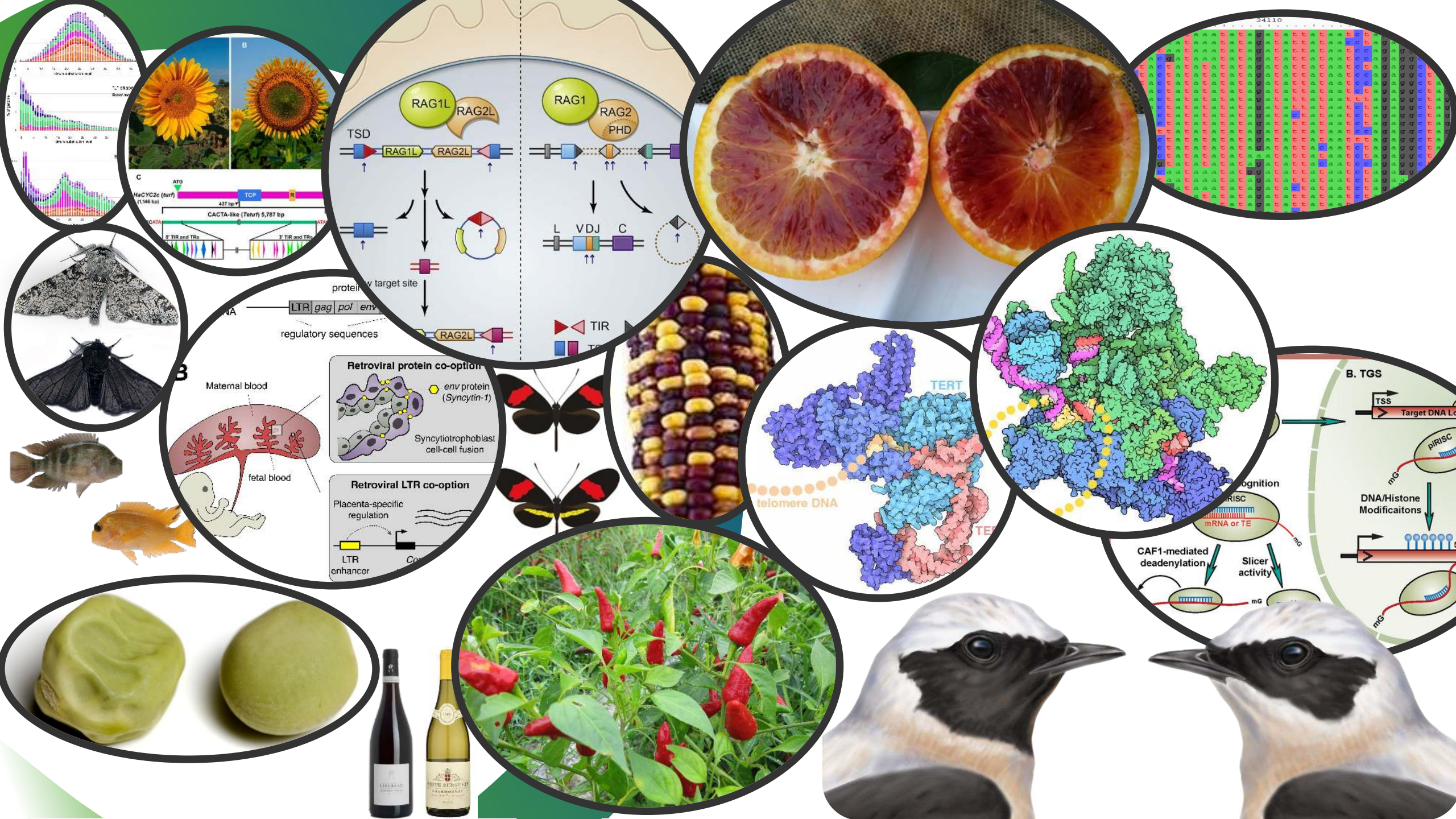
16th January 2025, Evomics Workshop on Genomics

Effects on assemblies



T2T and new human repeats





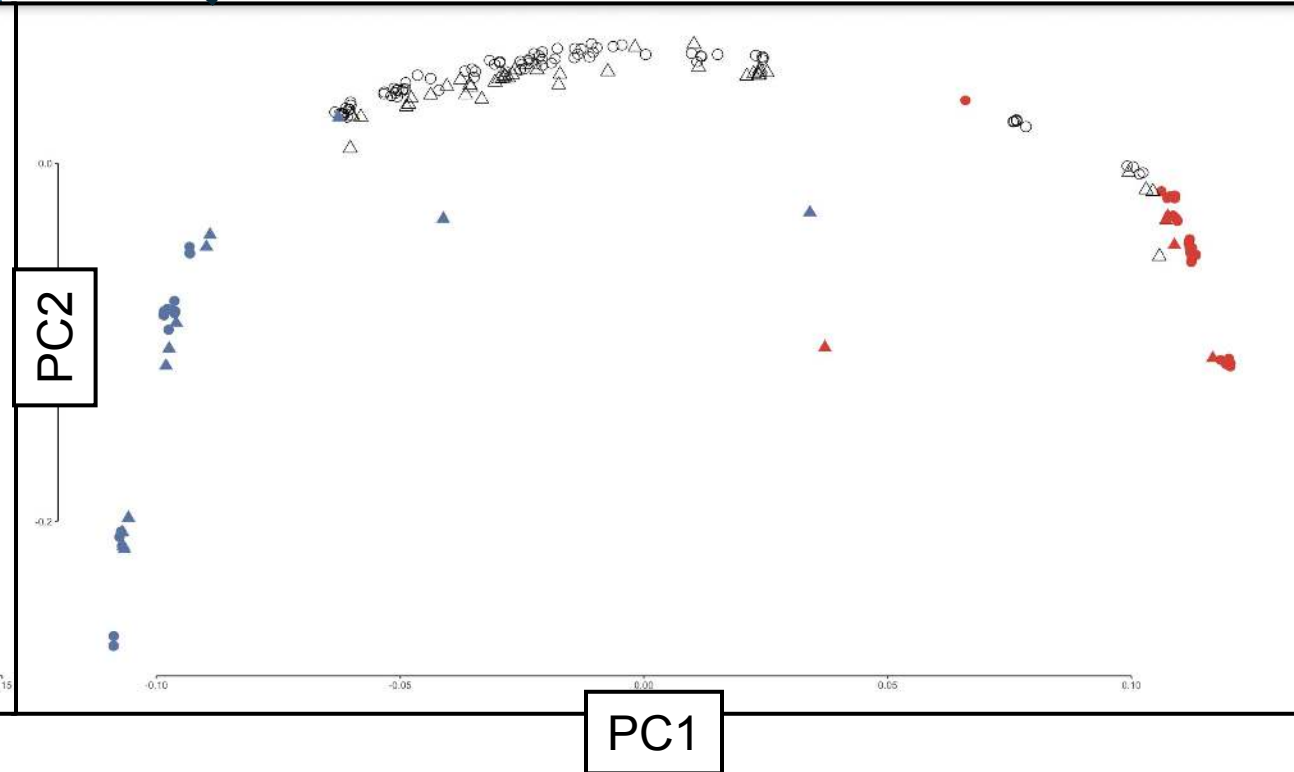
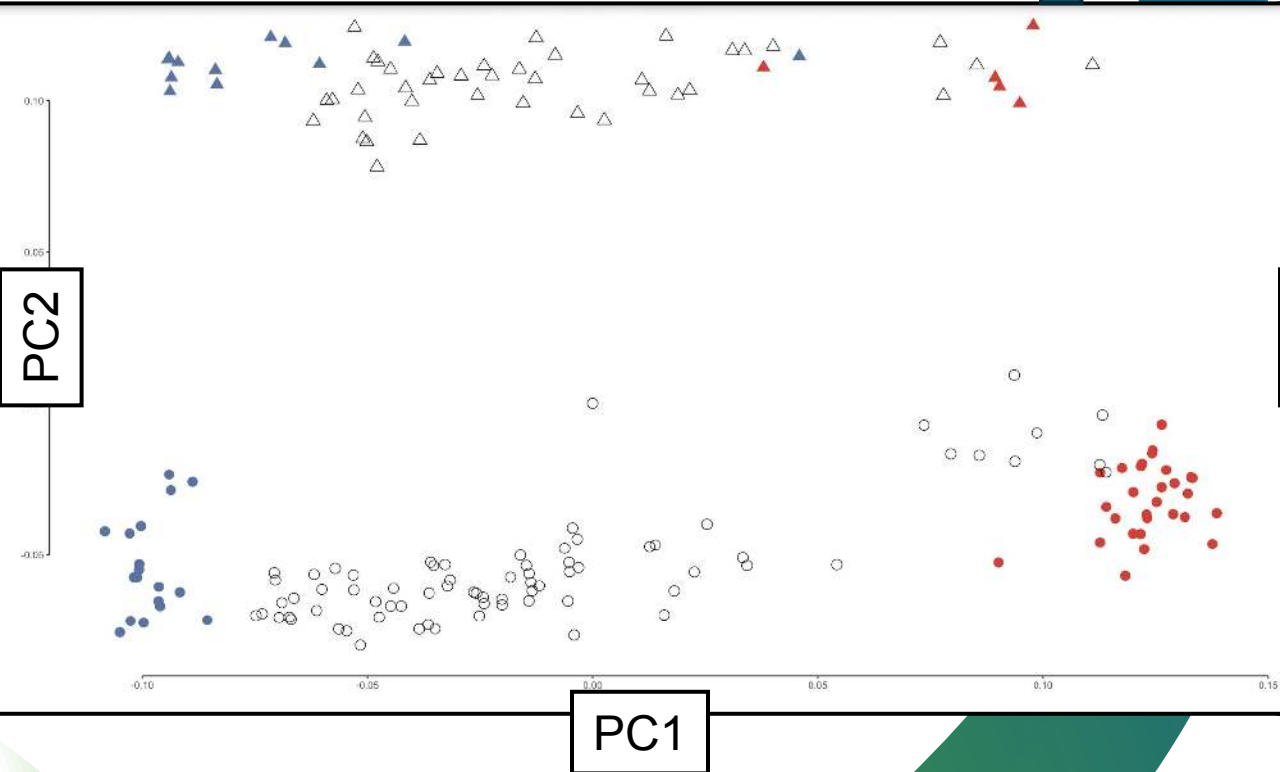
Effects on PCA



Dave
Lutgen

- △ Modern samples
- Historical samples

Removal of SNVs in
repeats
(and INDELS)



TE annotation and lab tutorial

- Characterise the diversity of TEs: RepeatModeler2
- Annotate TEs: RepeatMasker

Characterise the diversity of TEs



We need to know what types of TEs are present to then know where they are



We then need a library, a set of reference/representative sequences (**consensus sequences**) to use



We can use consensus sequences already available in various databases OR create a **de novo library** of consensus sequences



Different approaches for different types of input sequences

A consensus sequence can be a fasta or an HMM model

Genome assembly

All vs all alignment of the genome

- Cluster similar sequences

- Consensus sequences from multi-sequence alignments

Tools like RepeatModeler2, REPET, CARP

Raw reads

- All vs all alignment of the reads (downsampled at 0.1X)

- Cluster similar sequences

- Assembly of the clusters

- Tools like RepeatExplorer2, DNAPipeTE

Classification system

Cellular organisms

Phylum

Class

Order

Family

Genus

Species

Individual



Transposable elements

Class

Subclass

Order

Superfamily

Family

Subfamily

Copy

Consensus
sequence



Properties of a high-quality TE library

- ➡ is **complete** - the entire diversity of repeats is represented
- ➡ contains **nonredundant** consensus sequences - each element is represented only once
- ➡ contains **full-length consensus sequences** - each element is not fragmented/truncated

METHODOLOGY

Open Access



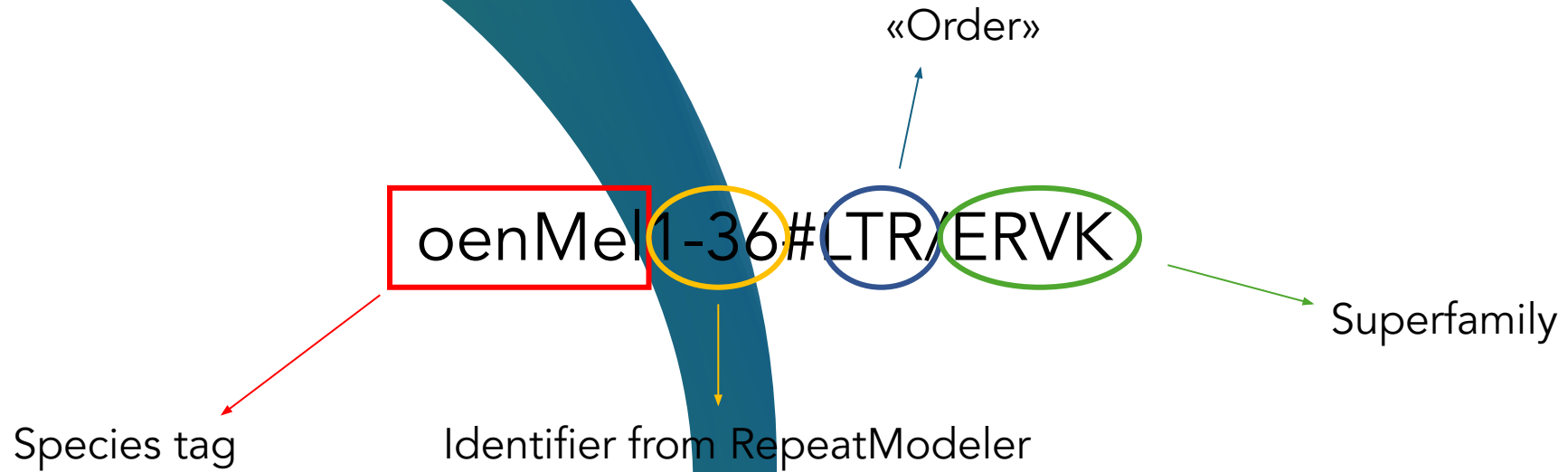
A beginner's guide to manual curation of transposable elements

Clement Goubert^{1,2}, Rory J. Craig³, Agustin F. Bilat⁴, Valentina Peona⁵, Aaron A. Vogan⁵ and Anna V. Protasio^{6,7*}

<https://mobilednajournal.biomedcentral.com/articles/10.1186/s13100-021-00259-7>

Extensive section of supplementary materials with (among the others) video tutorials of how to curate the consensus sequences!

Structure of the consensus sequence name



Other examples

`oenMel1-36_LTR#LTR/ERVK`

`oenMel1-36_int#LTR/ERVK`

`oenMel1-90.inc#LINE/CR1`

TE annotation

Genome.fasta.tbl

Genome.fasta.out (main output)

SW score	perc div.	perc del.	perc ins.	query sequence	position begin	in query end	(left)	matching repeat	repeat class/family	position in repeat begin end (left)	ID
206	6.7	0.0	0.0	NC_030765.1	12662	12691	(4911)	C Unknown-2762_S0	Unknown/Unknown	(1163) 64 35	1
1133	3.5	13.2	0.0	NW_022145616.1	1	174	(75979)	+ MITE_T_6311 NW_022146286.1 1017221 1019177 AT 15 F320	DNA/MITE	1494 1690 (266)	2 *
1863	2.7	0.0	2.2	NW_022145616.1	173	403	(75750)	+ MITE_T_6311 NW_022146286.1 1017221 1019177 AT 15 F320	DNA/MITE	1731 1956 (0)	3
1889	9.7	6.6	12.1	NW_022145616.1	404	811	(75342)	C MITE_T_17069 NW_022145681.1 3085354 3086586 GAATAT 15 F1012	DNA/MITE	(590) 642 255	4
631	13.8	16.6	0.6	NW_022145616.1	692	830	(75323)	C Unknown-1238_S0	Unknown/Unknown	(47) 203 43	5 *
720	11.9	13.8	0.6	NW_022145616.1	861	1012	(75141)	C Unknown-2421_S0	Unknown/Unknown	(15) 241 70	6
212	16.7	2.1	0.0	NW_022145616.1	1079	1126	(75027)	C MITE_T_2060 NW_022146441.1 222914 224876 agct 38 F136	DNA/MITE	(495) 1467 1419	7 *
699	9.3	0.0	0.0	NW_022145616.1	1117	1213	(74940)	+ Unknown-2755_S0	Unknown/Unknown	4 100 (1653)	8
216	16.3	0.0	0.0	NW_022145616.1	1214	1262	(74891)	+ MITE_T_28504 NW_022145617.1 1329355 1331252 ta 38 F1758	DNA/MITE	1026 1074 (823)	9 *
650	17.6	0.9	1.7	NW_022145616.1	1251	1366	(74787)	C Unknown-2726_S0	Unknown/Unknown	(104) 128 14	10
1553	10.6	3.8	2.3	NW_022145616.1	1872	2132	(74021)	+ DNA-2829_S0	DNA/MITE	130 394 (35)	11
846	14.4	2.2	17.7	NW_022145616.1	2107	2334	(73819)	C Unknown-1886_S0	Unknown/Unknown	(514) 562 365	12 *
1192	5.3	3.2	14.7	NW_022145616.1	2133	2352	(73801)	+ MITE3_S0	DNA/MITE	1 198 (1030)	13
1023	7.6	1.4	1.4	NW_022145616.1	2345	2491	(73662)	+ MITE3_S0	DNA/MITE	1082 1228 (0)	14 *
606	19.4	1.4	2.0	NW_022145616.1	2361	2507	(73646)	C DNA-3306_S0	DNA/MITE	(596) 181 36	15 *
239	17.8	6.9	3.3	NW_022145616.1	2819	2876	(73277)	C Unknown-1619_S0	Unknown/Unknown	(271) 452 393	16
996	28.3	5.0	4.4	NW_022145616.1	2879	3827	(72326)	+ LINE-3770_S0	LINE/RTE	424 1282 (33)	17 *

```
101-10630-2060 229 6340 bp 0.75 %
En-Spm 0 0 bp 0.00 %
MuDR-IS905 0 0 bp 0.00 %
PiggyBac 1 129 bp 0.00 %
Tourist/Harbinger 1 465 bp 0.01 %
Other (Mirage,
P-element, Transib) 0 0 bp 0.00 %
Rolling-circles 2 470 bp 0.01 %
Unclassified: 1567 500540 bp 5.89 %
Total interspersed repeats: 4726544 bp 55.60 %

Small RNA: 28 2042 bp 0.02 %

Satellites: 85 43411 bp 0.51 %
Simple repeats: 0 0 bp 0.00 %
Low complexity: 0 0 bp 0.00 %
=====
```

We can find the coordinates of repeats in the genome by aligning our library to the genome

RepeatModeler2

GENOME
FASTA FILE

```
>oenMel  
ATGAGCGCGAGAGGG  
CGAATCCCTAGGCTA  
ACATCGTCCCGCGAT  
GCTTGCTTAGAACCT  
TGCCTAGACCTGAGC  
TCTAGCTTACTGCTA  
GCTTCCGATTACAC  
GATCACCTACATAT  
CTTCACATCCATCTC
```

BuildDatabase -name <name> <genome
file>



RepeatModeler2 -database
<name>



Library of consensus
sequences



RepeatMasker

GENOME
FASTA FILE

```
>oenMel  
ATGAGCGCGAGAGGG  
CGAATCCCTAGGCTA  
ACATCGTCCCGCGAT  
GCTTGCTTAGAACCT  
TGCCTAGACCTGAGC  
TCTAGCTTACTGCTA  
GCTTCCGATTACAC  
GATCACCTACATAT  
CTTCACATCCATCTC
```



```
>consensus1  
ATTGCGCGTTAGGAT  
ATCCCGATCGCCC  
>consensus2  
TGTAGGGAGTCTTGA  
CA  
>consensus3  
ATTTCGGGCTAGGCT  
TGAGGC
```

RepeatMasker -lib <library>
<genome file>



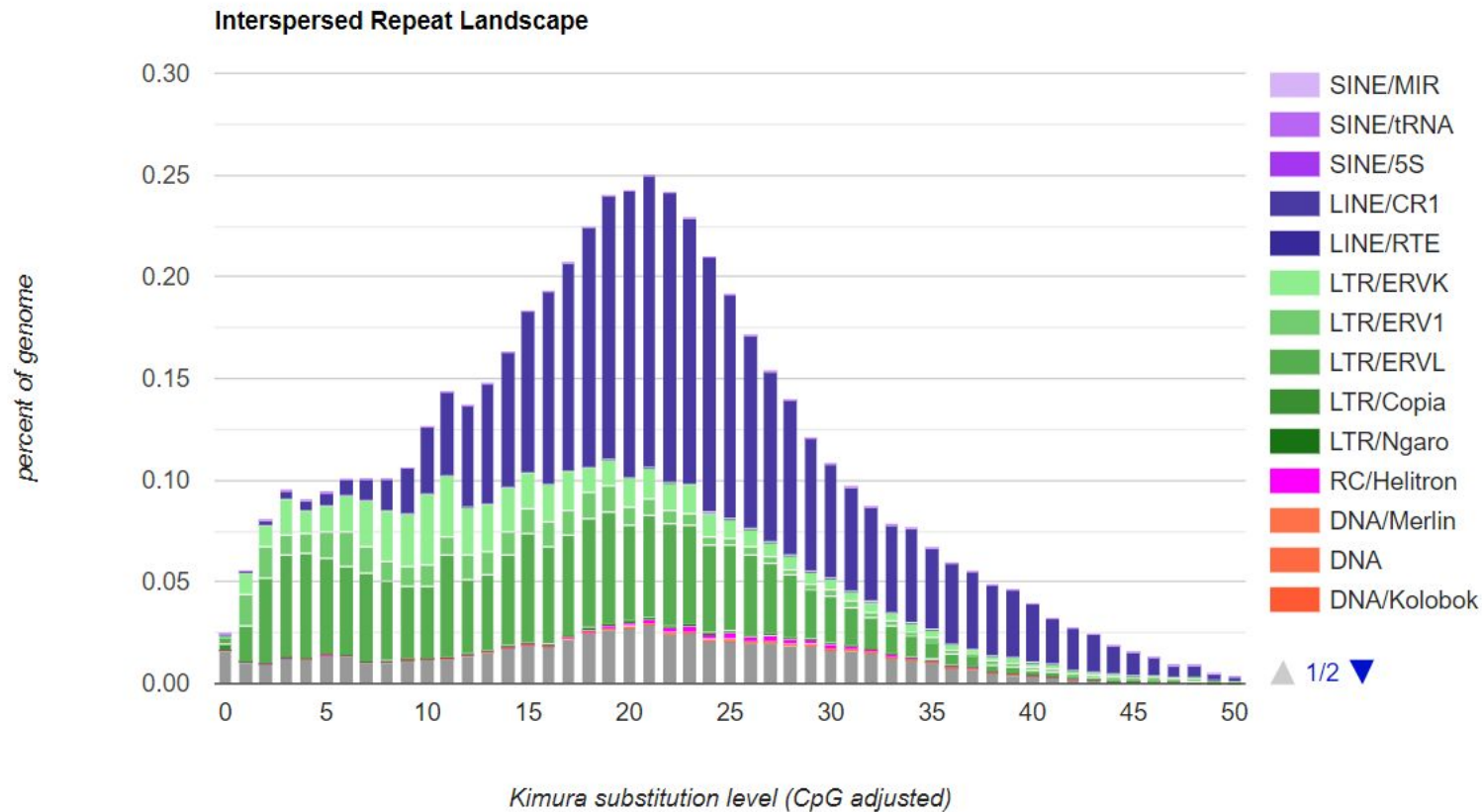
.out + .gff files with
coordinates of repeats



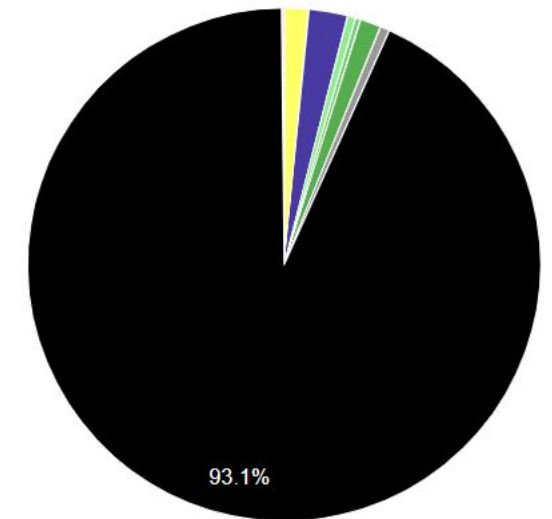
Repeat landscape

Use outputs from RepeaMasker to visualise the repetitive content of the genome

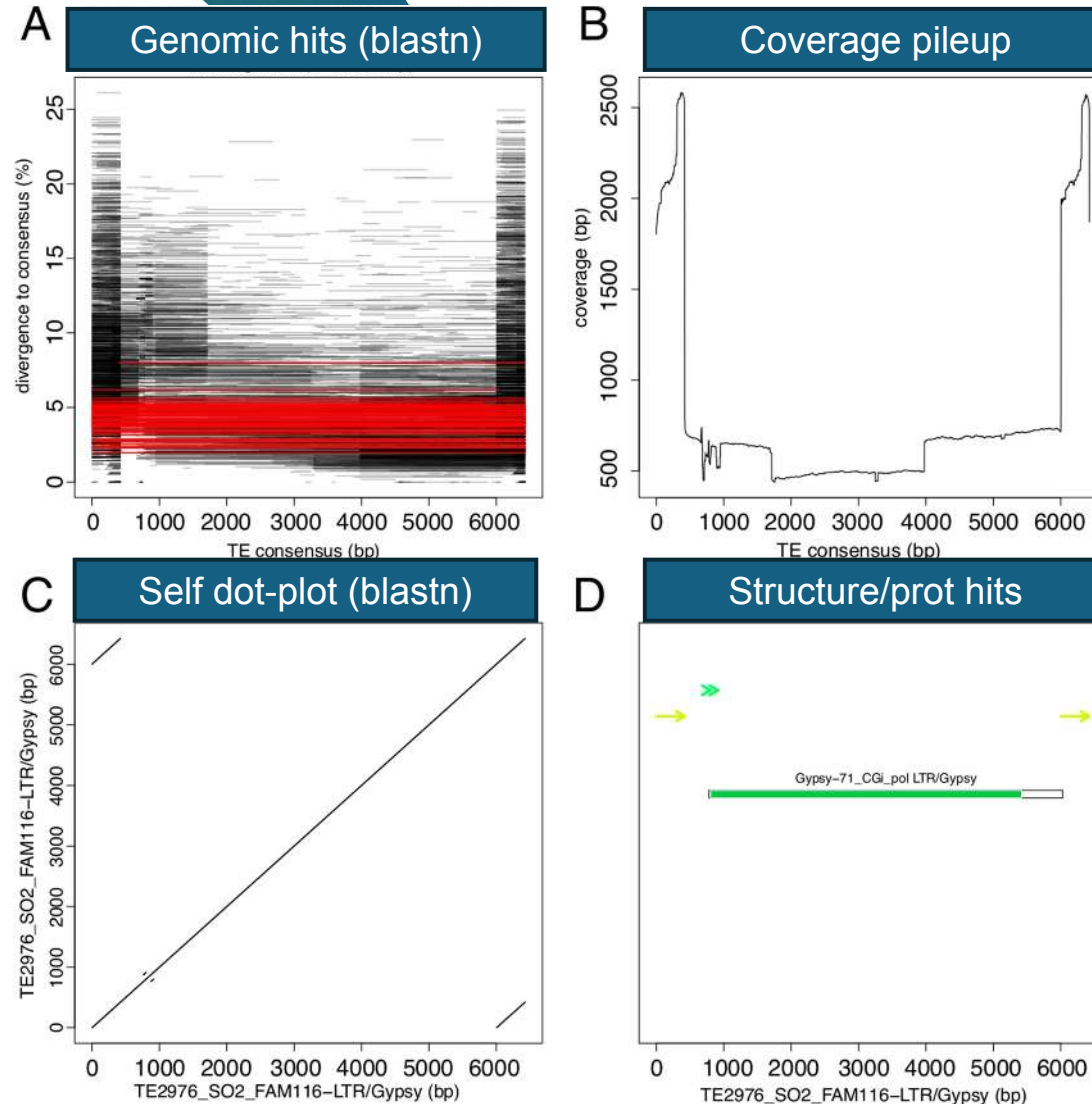
- calcDivergenceFromAlig.pl
- createRepeatLandscape.pl



Genome Fraction

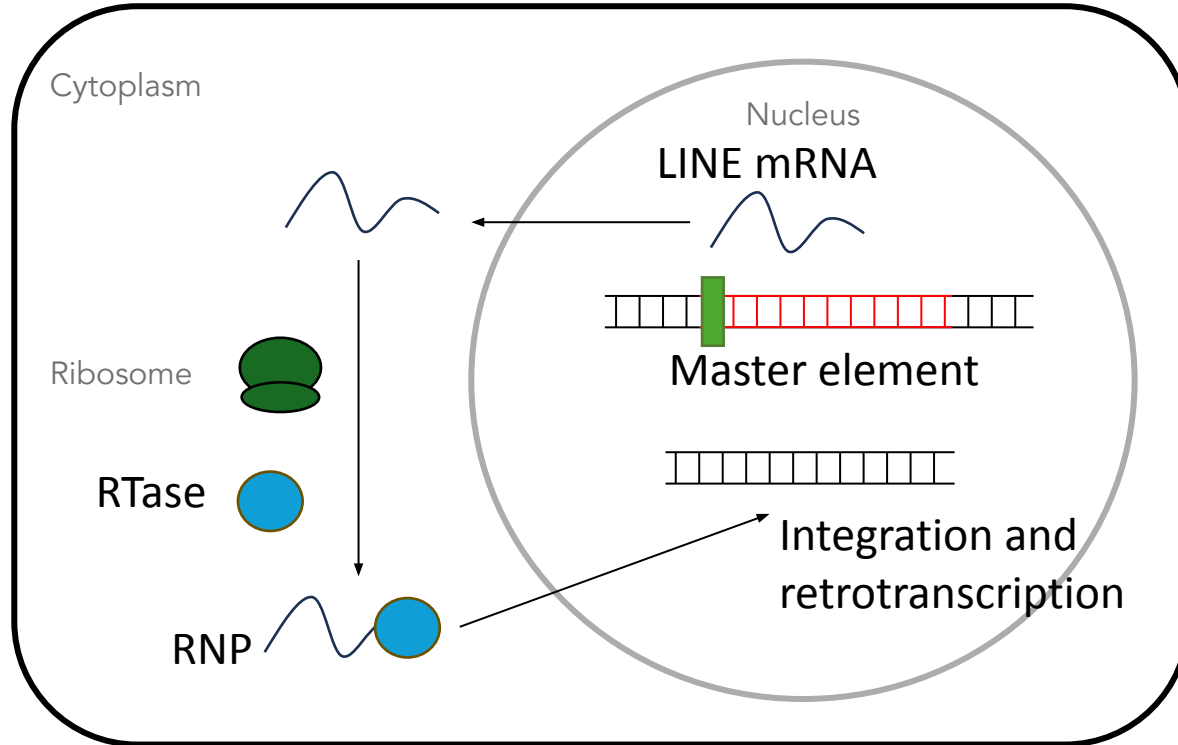


Analyse sequence characteristics of TEs



LINE retrotransposons

Where/when/how in the cell



Target site preference and TSD

Target site preferentiality for sequences similar to the 3' UTR

Target site duplications are of variable length

Requirements for mobility

Transcription: promoter for pol II -> mRNA + polyA

Replication: RTase

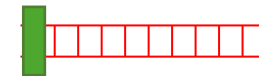
Recognition site for *cis*-mobilisation

Integration: endonuclease

No introns

Content of new copies

Mother copy



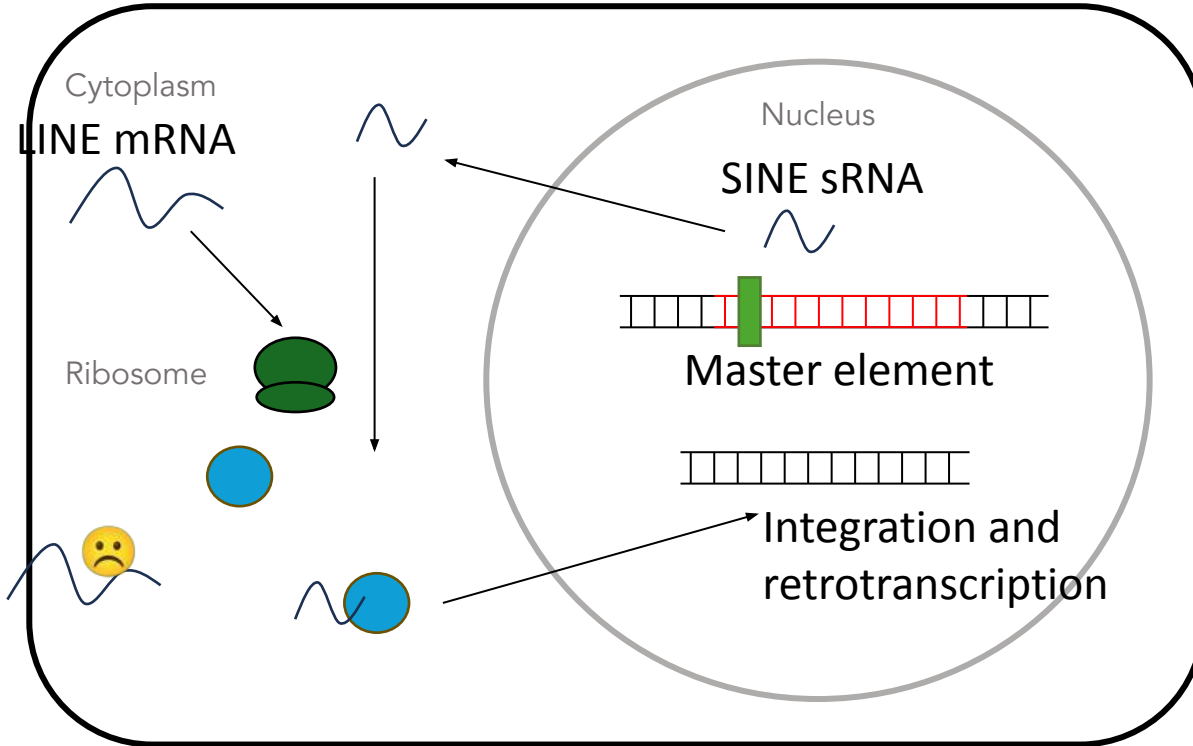
Daughter copies



5' truncation
"dead on arrival"

SINE retrotransposons

Where/when/how in the cell



Target site preference and TSD

Target site preferentiality for sequences similar to the 3' UTR

Target site duplications are of variable length

Requirements for mobility

Transcription: promoter for pol III -> sRNA

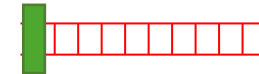
Replication + integration: using LINE derived proteins

Recognition site for *trans*-mobilisation

No introns

Content of new copies

Mother copy



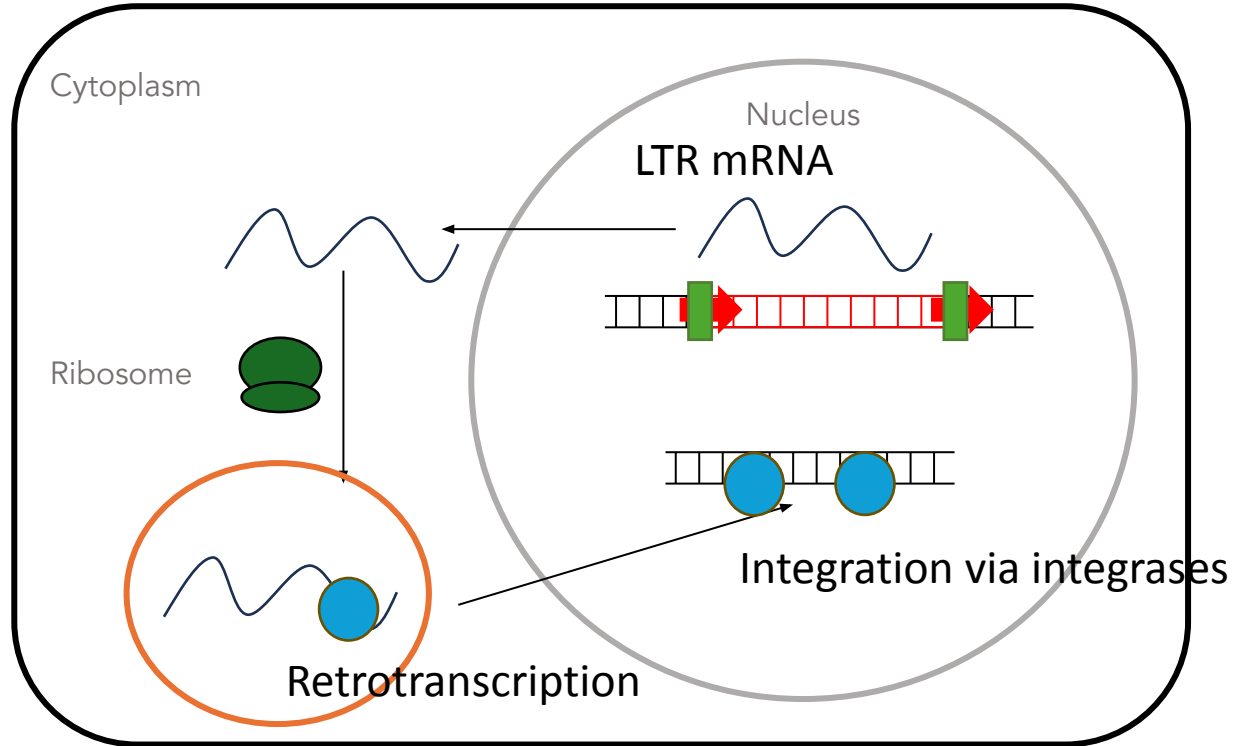
Daughter copies



5' truncation
"dead on arrival"

LTR retrotransposons

Where/when/how in the cell



Target site preference and TSD

No preferentiality for target site

Specific length of target site duplications (4 bp, 5 or 6 bp)

Requirements for mobility

Transcription: promoter for pol II -> mRNA + polyA

Replication: within viral-like particle, gag (capsid), protease, RTase, RNase H

Integration: integrase

Recognition site for *cis*-mobilisation (LTR)

No introns

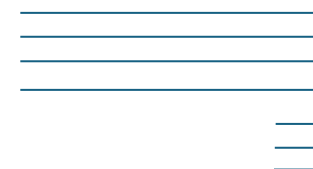
Content of new copies

Mother copy



Daughter

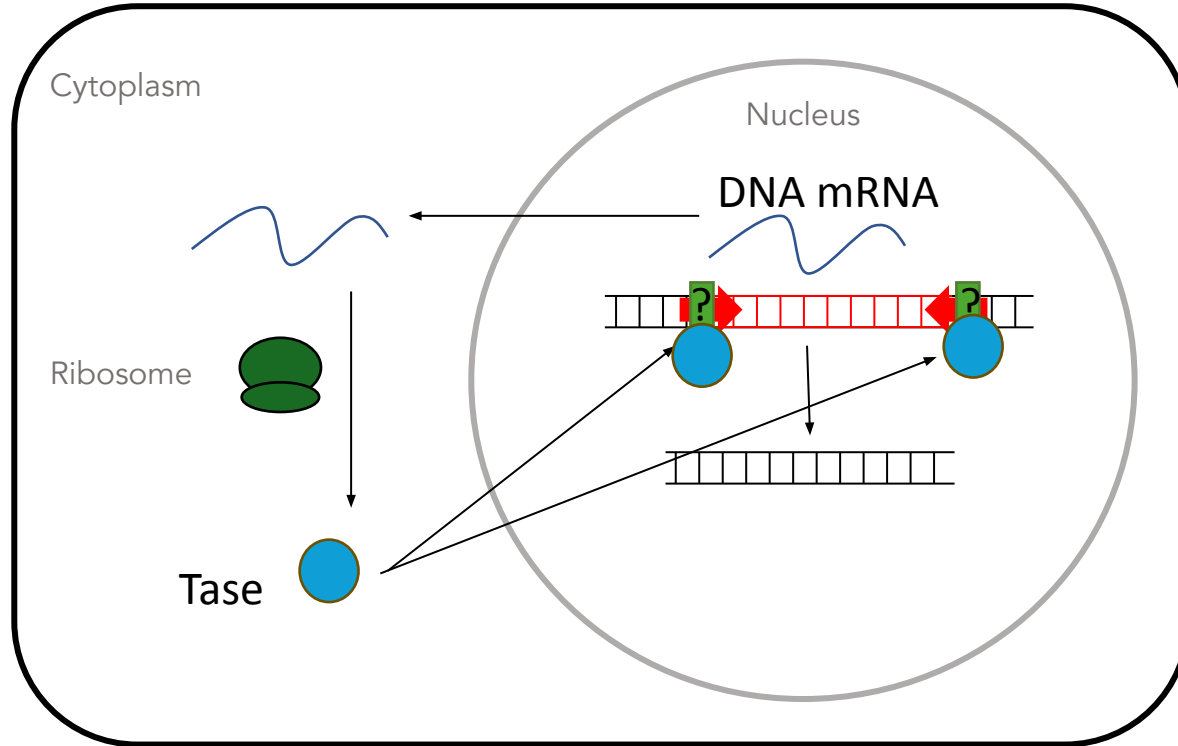
NAHR



Solo-LTR

DNA transposons (TIR)

Where/when/how in the cell



Target site preference and TSD

There can be specificity for target site (e.g., TA, TTAA) and there can be specific target site duplication length (e.g., 8 bp)

Requirements for mobility

Transcription: promoter for pol II -> mRNA + polyA

Mobilisation and integration: transposase (Tase)

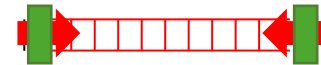
Replication: dependent on host DNA replication

Recognition site in TIRs allows for both *cis*- and trans-mobilisation (with different probabilities)

Might have introns

Content of new copies

Mother copy



Daughter copy



Inclusion of extra DNA
or loss of ORFs