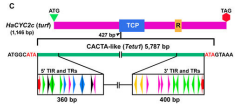
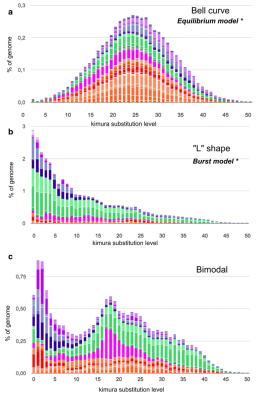



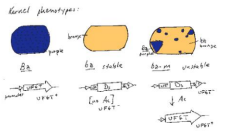

# BIOINFORMATIC ANALYSIS OF TRANSPOSABLE ELEMENTS

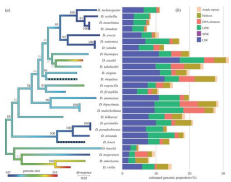
3rd-7th November 2025

## Lecture 4
## TEs and genome evolution

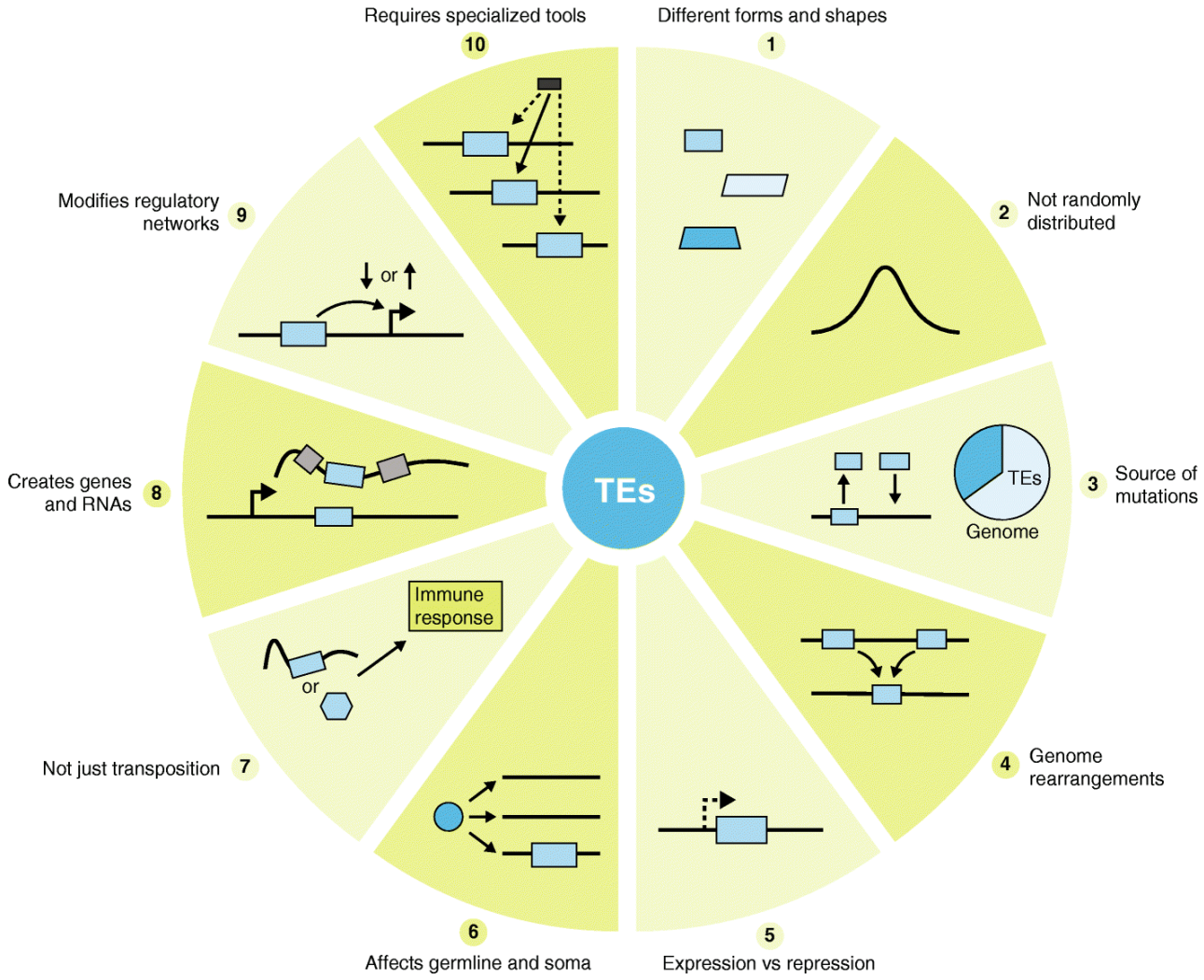

Valentina Peona

NATURHISTORISKA RIKSMUSEET



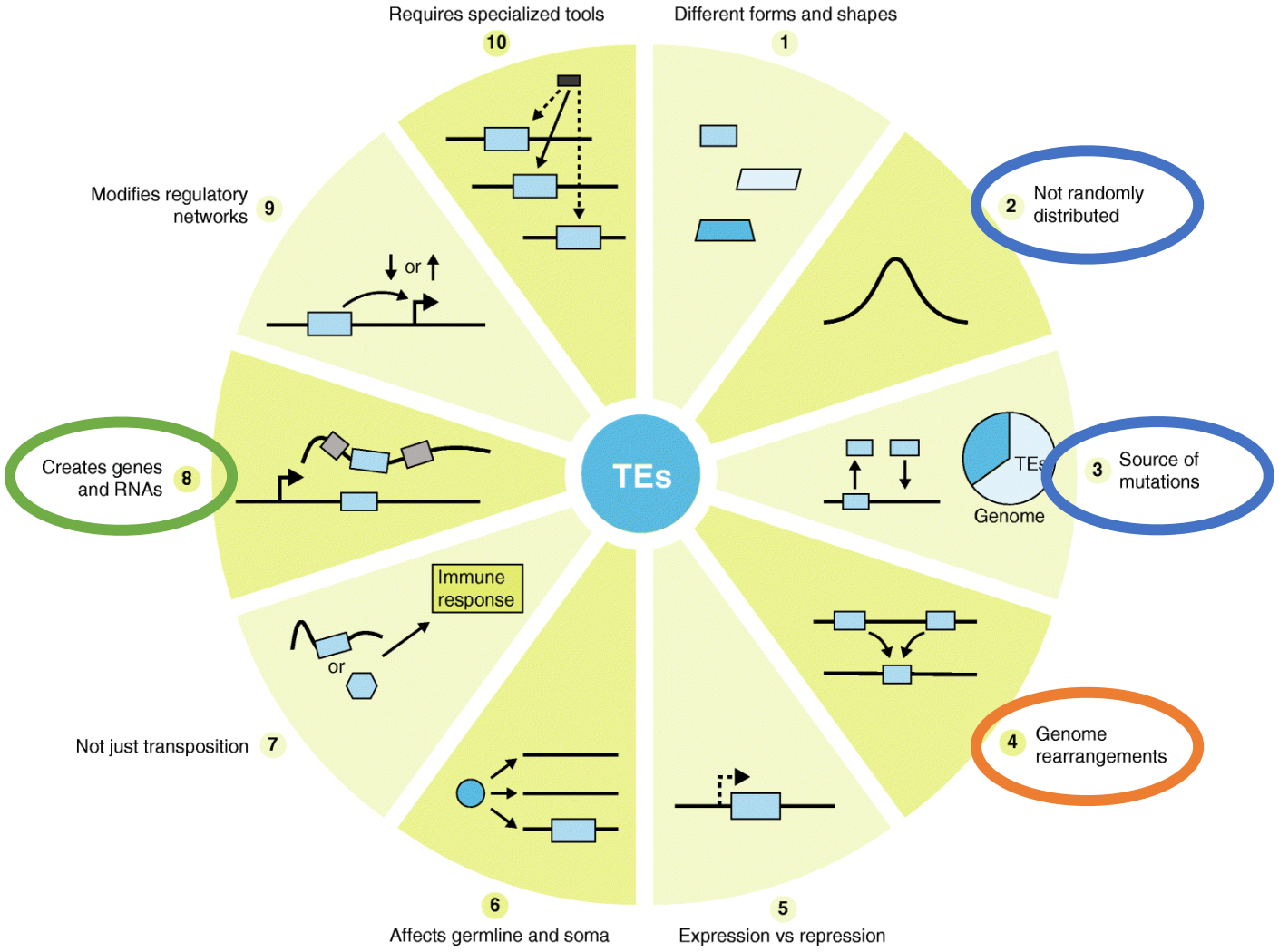


# «Ten things you should know about TEs»




Bourque et al 2018, Mobile DNA

# «Ten things you should know about TEs»




Bourque et al 2018, Mobile DNA

# Drivers of genome evolution

1. Structure: from size to 3D organization

2. Ecology: waves of transposition and differential accumulation

3. Gene evolution: TEs as innovators and constraint

# Structural changes

## Mechanisms

1. Non-homologous end joining DNA repair - NHEJ

2. Non-allelic homologous recombination - NAHR

3. Copy paste transposition

4. Cut and paste transposition

# Structural changes

## Mechanisms

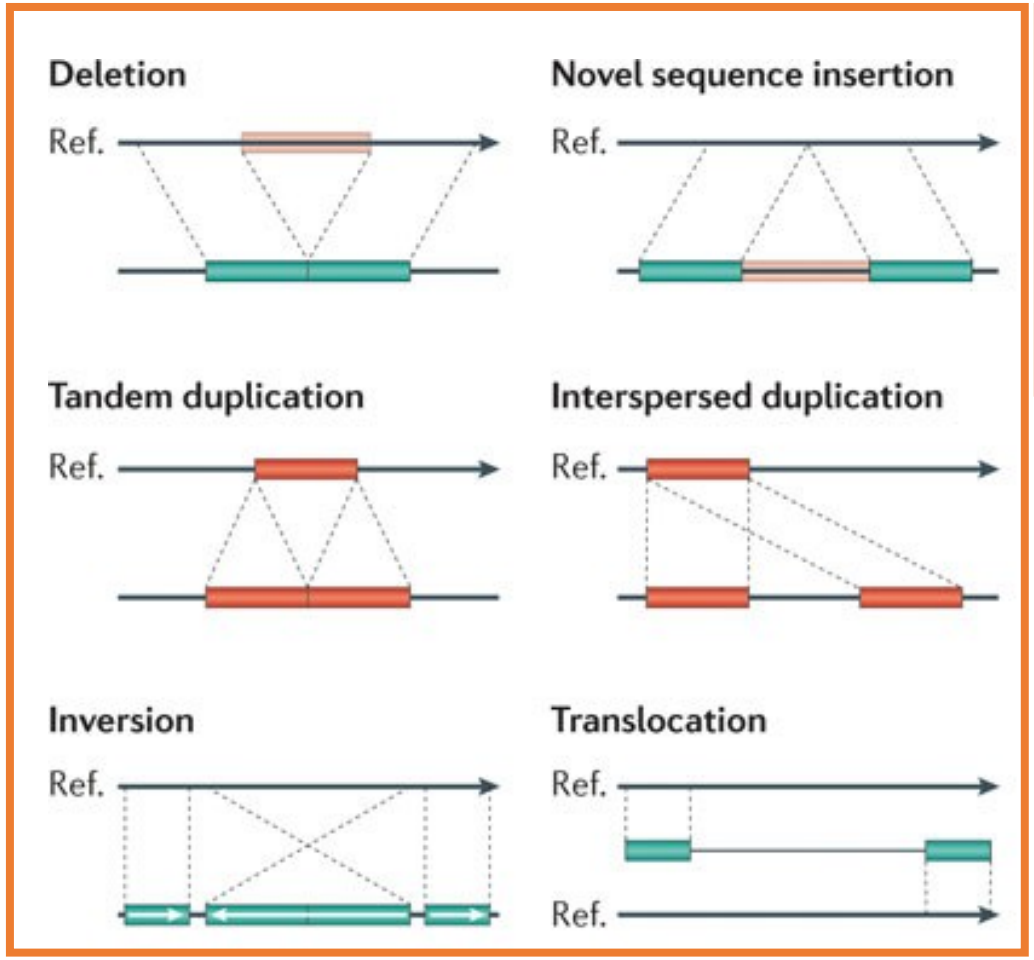
1. Non allelic homologous recombination - NAHR

2. Copy paste transposition

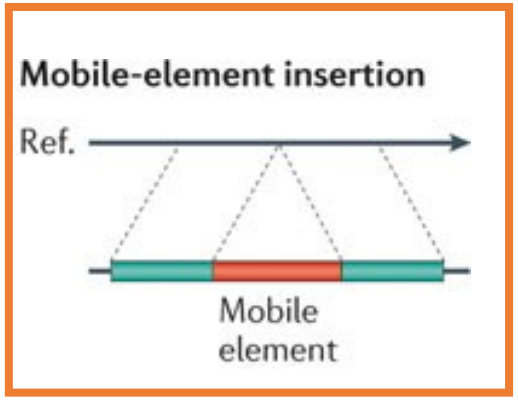
3. Cut and paste transposition

# Genome rearrangements

## Types




NAHR
NHEJ

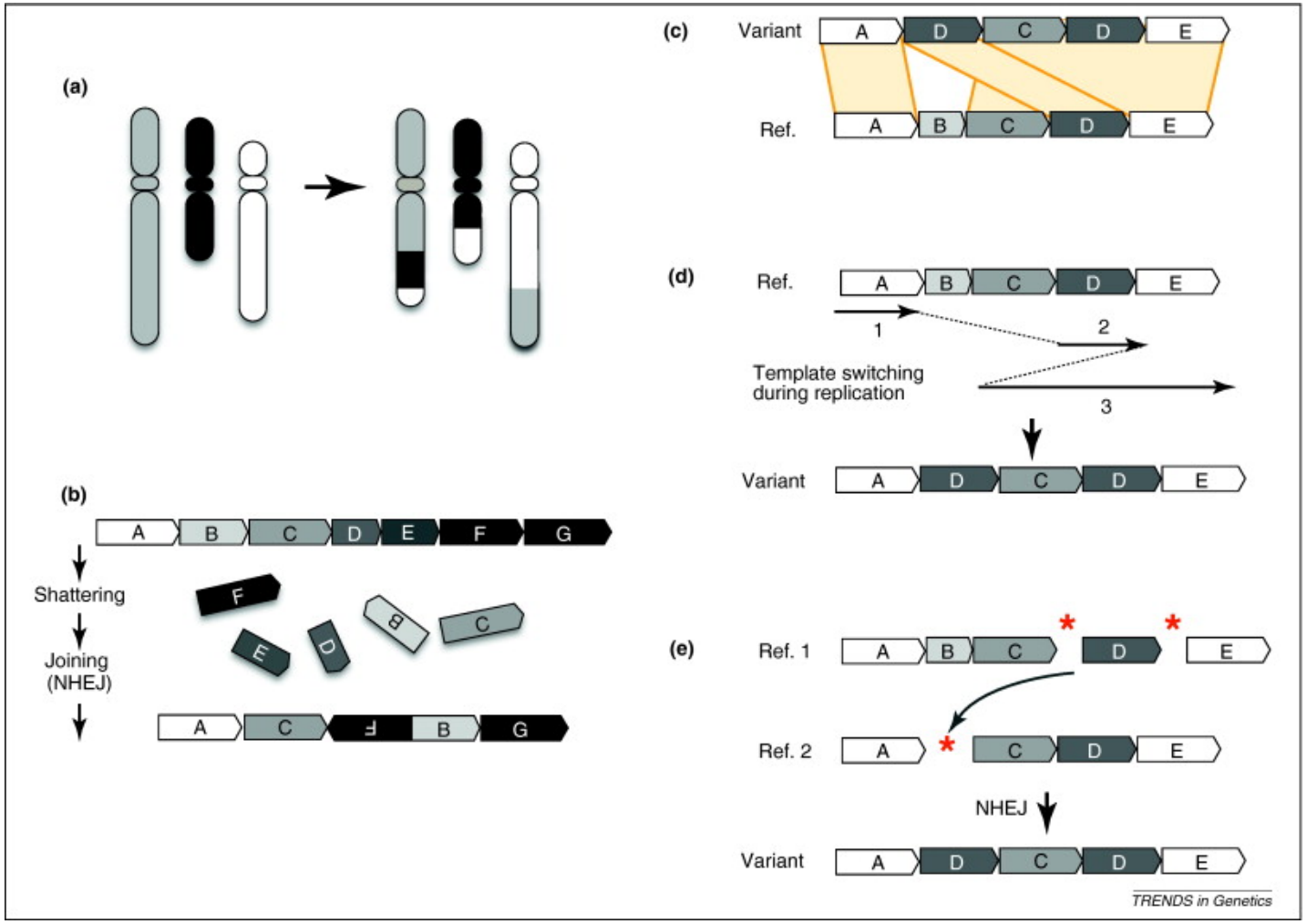



Transposition and retrotransposition

Nature Reviews | Genetics

Alkan et al 2011, Nat Rev Gen

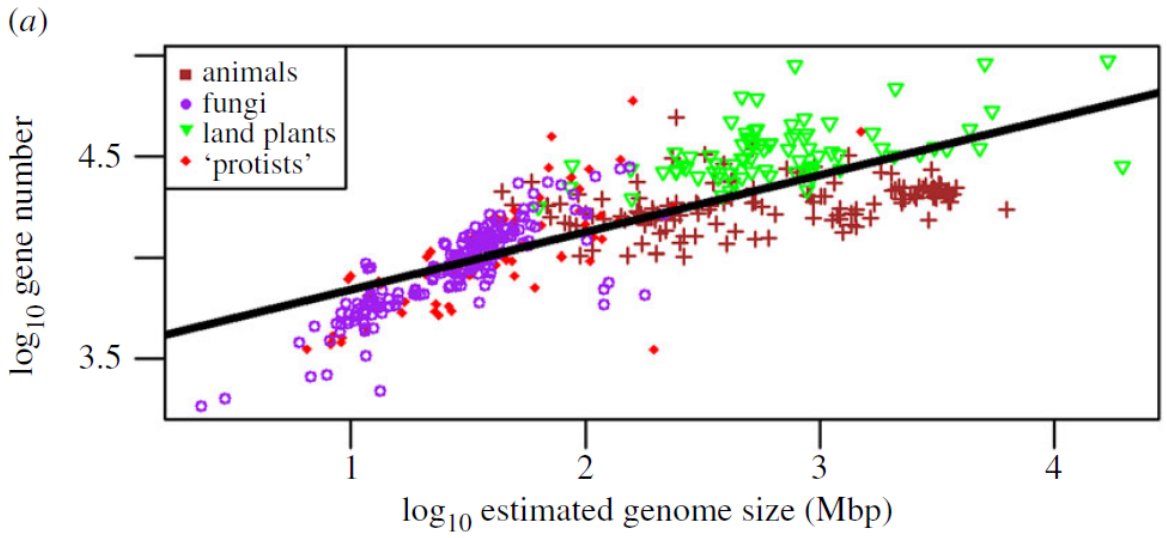
# Complex rearrangements




Quilan and Hall, 2012 Trends in Genetics

# Genome size evolution

## C-value paradox

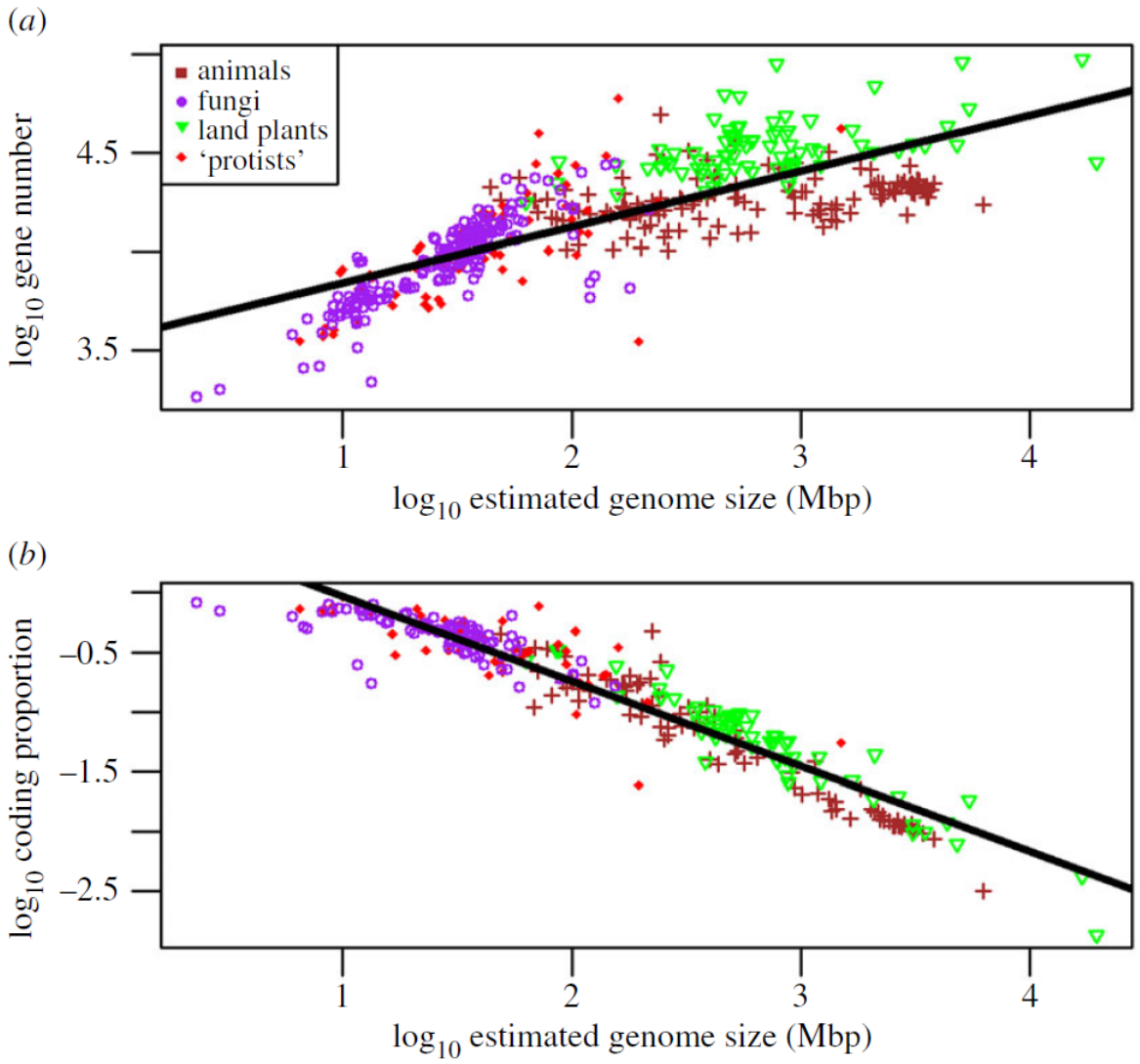



Positive relationship with number of genes

Elliot and Gregory, 2015 Phil Trans B

# Genome size evolution

## C-value paradox




Positive relationship with number of genes

Negative relationship with proportion

Bigger genome = more non-functional (for the host) DNA

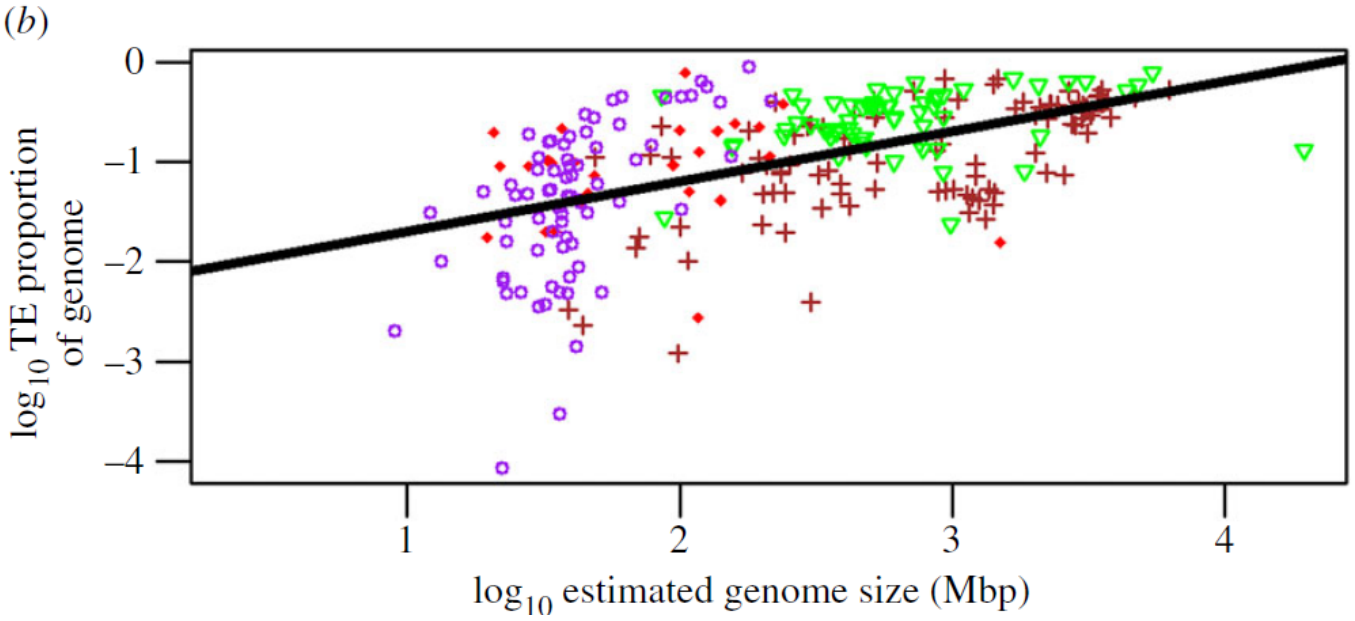
Elliot and Gregory, 2015 Phil Trans B

# Genome size evolution

| Species | C-value (kb) |
|---|---|
| Navicola pelliculosa | 35,000 |
| Drosophila melanogaster | 180,000 |
| Gallus domesticus | 1,200,000 |
| Lampreta planrei | 1,900,000 |
| Xenopus laevis | 3,100,000 |
| Homo sapiens | 3,400,000 |
| Nicotiana tabaccum | 3,800,000 |
| Paramecium caudaatum | 8,600,000 |
| Allium cepa | 18,000,000 |
| Lilium formosanum | 36,000,000 |
| Amphiuma means | 84,000,000 |
| Protopterus aethiopicus | 140,000,000 |
| Amboeba proteus | 290,000,000 |
| Amoeba dubia | 670,000,000 |

Molecular Evolution, Li, W.-H. 1997

C-value paradox



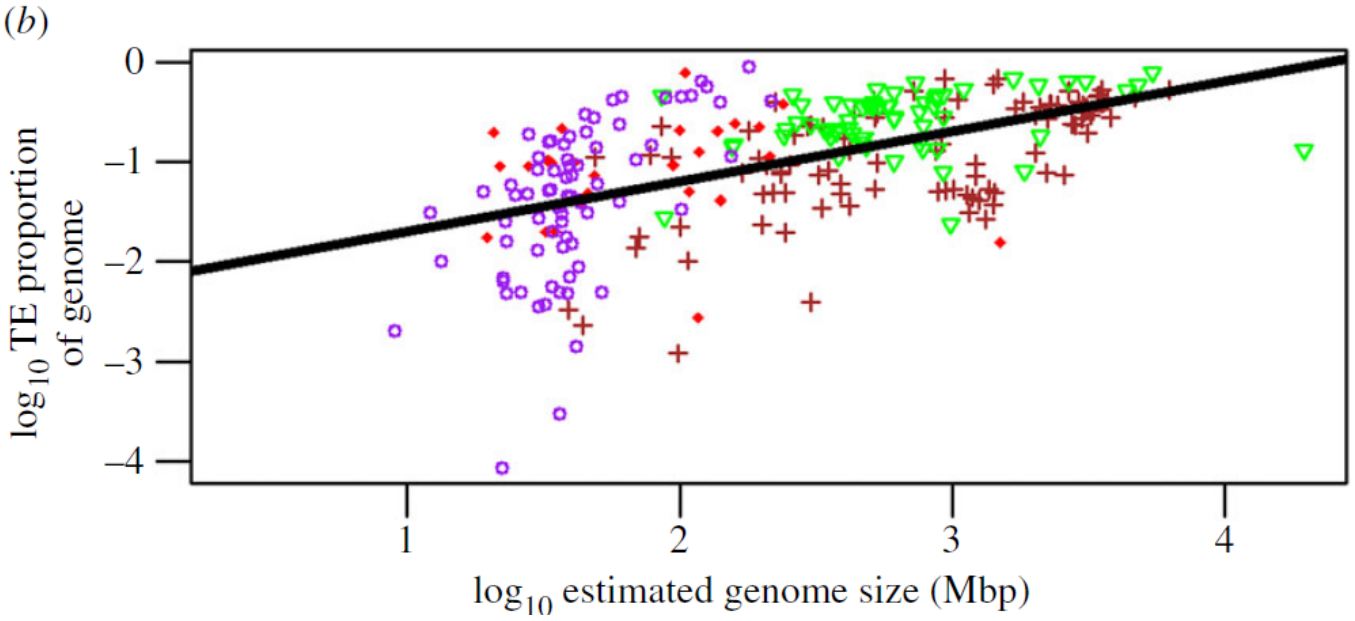

Elliot and Gregory, 2015 Phil Trans B

# Genome size evolution

| Species | C-value (kb) |
|---|---|
| Navicola pelliculosa | 35,000 |
| Drosophila melanogaster | 180,000 |
| Gallus domesticus | 1,200,000 |
| Lampreta planrei | 1,900,000 |
| Xenopus laevis | 3,100,000 |
| Homo sapiens | 3,400,000 |
| Nicotiana tabaccum | 3,800,000 |
| Paramecium caudaatum | 8,600,000 |
| Allium cepa | 18,000,000 |
| Lilium formosanum | 36,000,000 |
| Amphiuma means | 84,000,000 |
| Protopterus aethiopicus | 140,000,000 |
| Amboeba proteus | 290,000,000 |
| Amoeba dubia | 670,000,000 |

Molecular Evolution, Li, W.-H. 1997

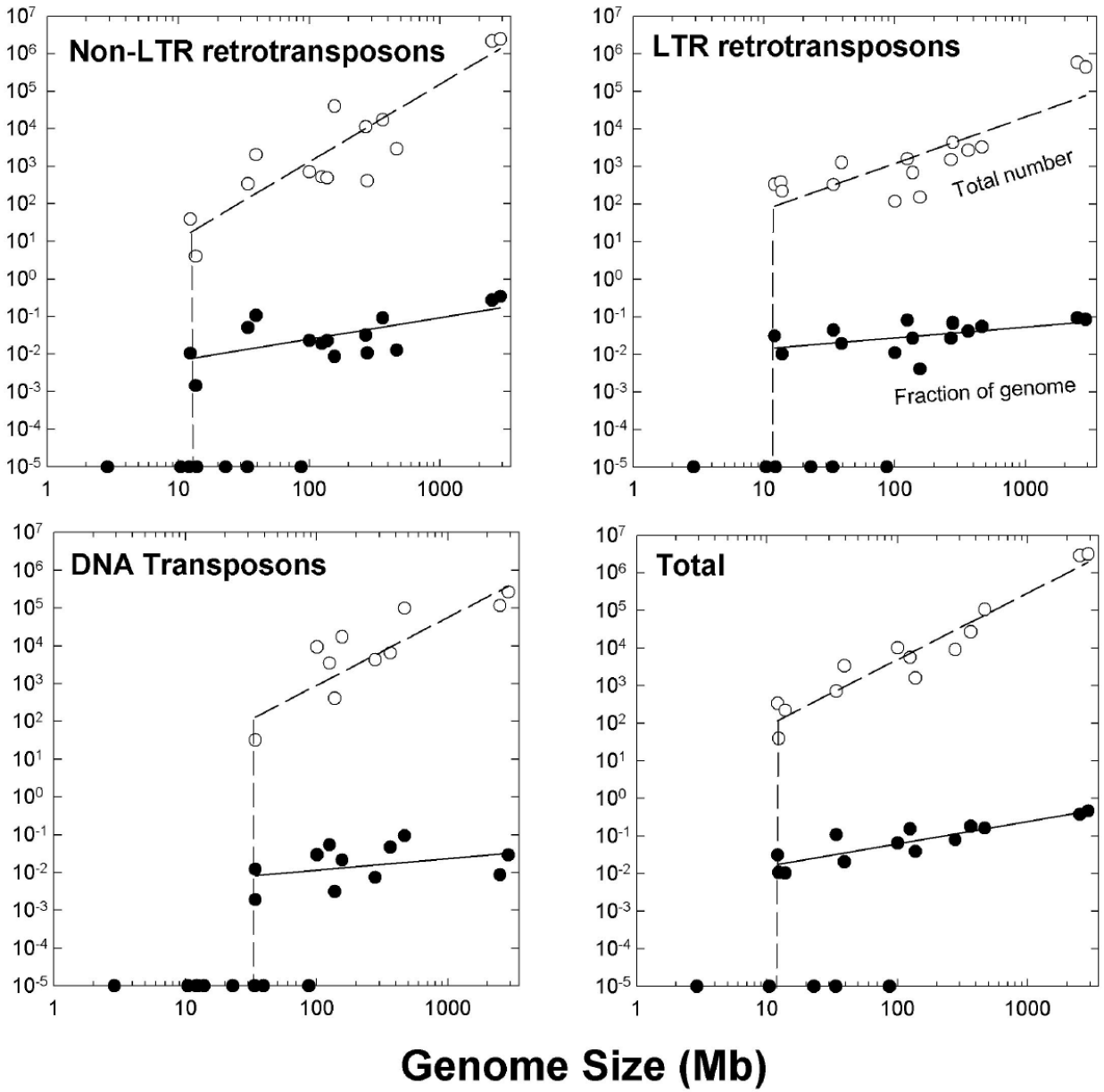
C-value paradox




- Just because we find a lot of TEs does not mean that the genome «needs» them, maybe genomes are just stuck with TEs and can deal with them
- Ultimate vs proximate cause for TE presence
- Important distinction for when we design hypotheses

Elliot and Gregory, 2015 Phil Trans B

# Nothing in evolution makes sense…




**Why do we have more/less TEs?**

"Consistent with theoretical expectations, all three classes of mobile elements appear to have a threshold genome size below which mobile elements are unable to establish, an intermediate range in which only a fraction of species harbor them, and an upper threshold (~100 Mb) above which all species are infected (1) (Fig. 4). By extrapolation from the mutation rate cited above, the critical effective population size above which a unicellular eukaryote population appears to be immune to retrotransposon proliferation is ~$7 \times 10^7$, whereas for DNA-based transposons it is ~$2 \times 10^7$."

- Need space for jumping and for accumulating
- Genomes that are too small may not tolerate well the presence of TEs
- Effective population size, recombination rate, …

Lynch and Conery, 2003 Science

# … except in the light of population genetics

## The frailty of adaptive hypotheses for the origins of organismal complexity

**Michael Lynch***

Department of Biology, Indiana University, Bloomington, IN 47405

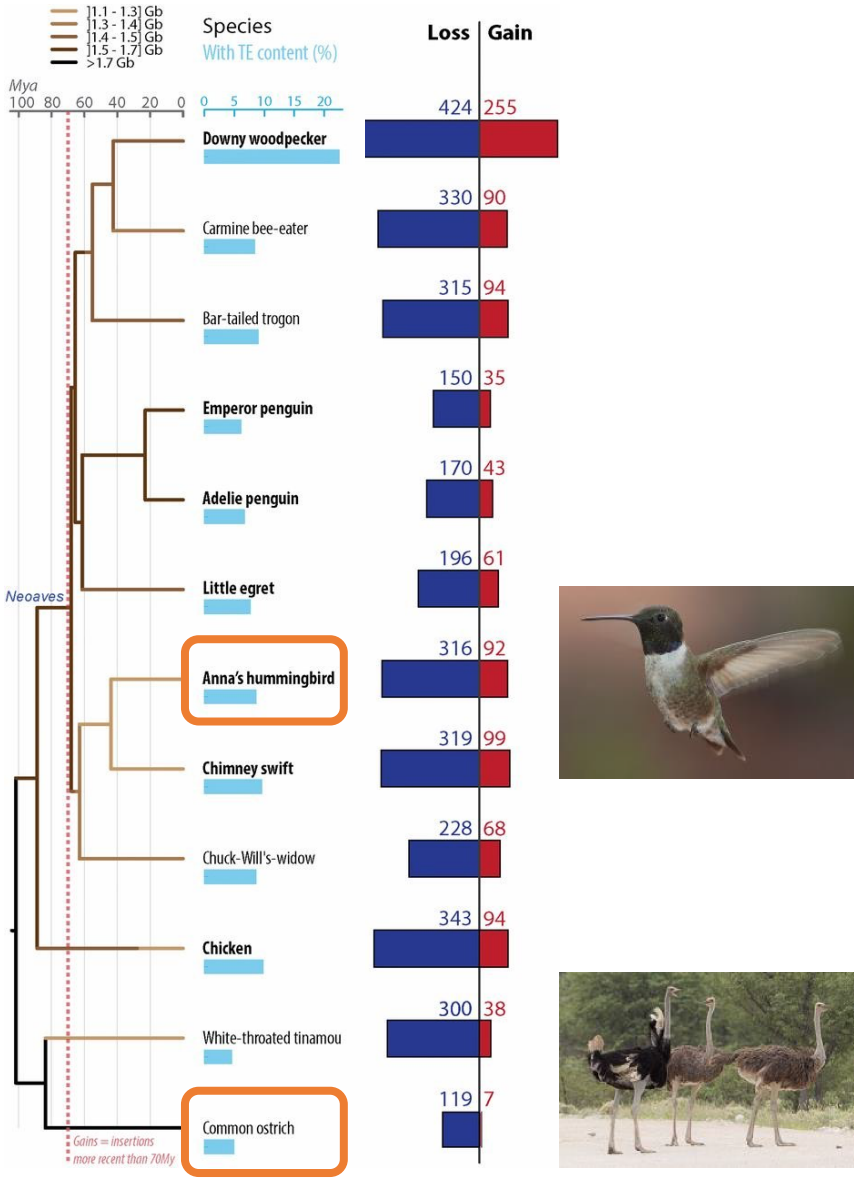
The vast majority of biologists engaged in evolutionary studies interpret virtually every aspect of biodiversity in adaptive terms. This narrow view of evolution has become untenable in light of recent observations from genomic sequencing and population-genetic theory. Numerous aspects of genomic architecture, gene structure, and developmental pathways are difficult to explain without invoking the nonadaptive forces of genetic drift and mutation. In addition, emergent biological features such as complexity, modularity, and evolvability, all of which are current targets of considerable speculation, may be nothing more than indirect by-products of processes operating at lower levels of organization. These issues are examined in the context of the view that the origins of many aspects of biological diversity, from gene-structural embellishments to novelties at the phenotypic level, have roots in nonadaptive processes, with the population-genetic environment imposing strong directionality on the paths that are open to evolutionary exploitation.

Lynch 2007, PNAS

1

# Genome size evolution

## Accordion model





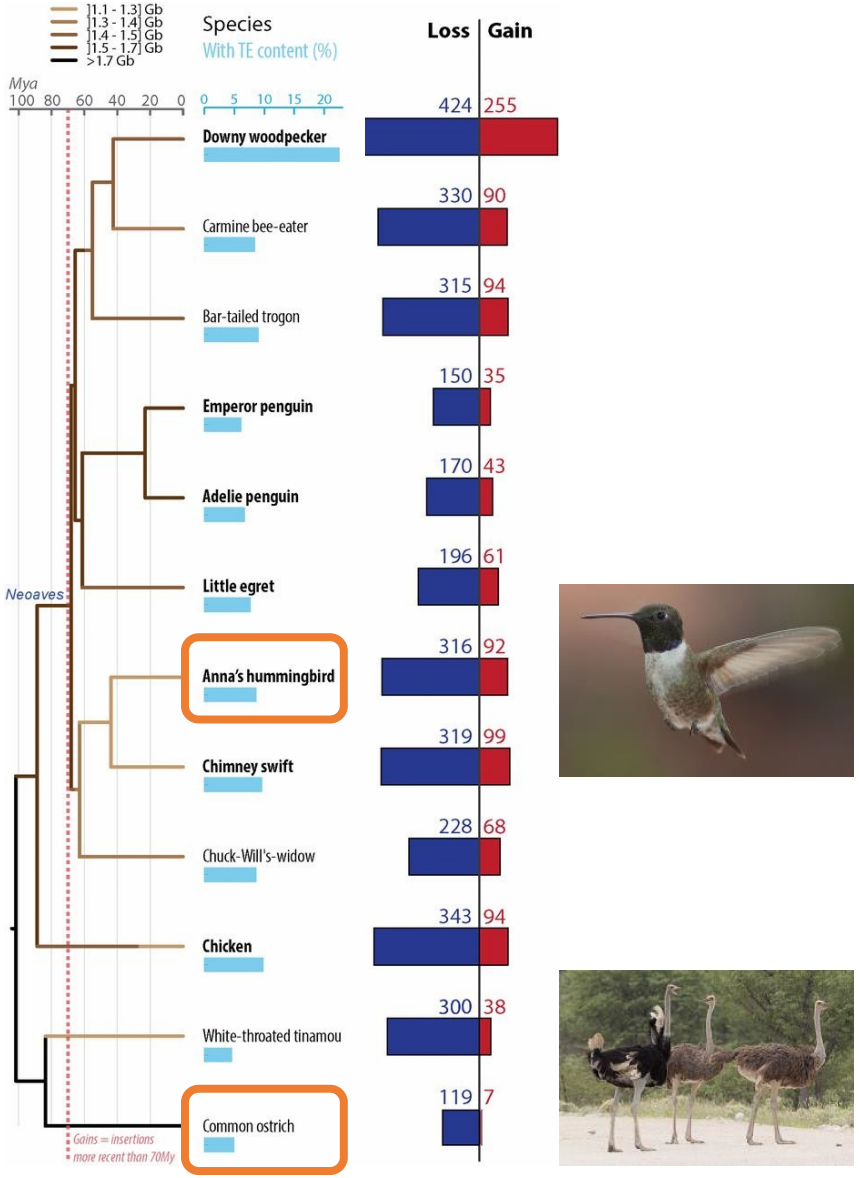



Kapusta et al 2017, PNAS

1

# Genome size evolution

## Accordion model







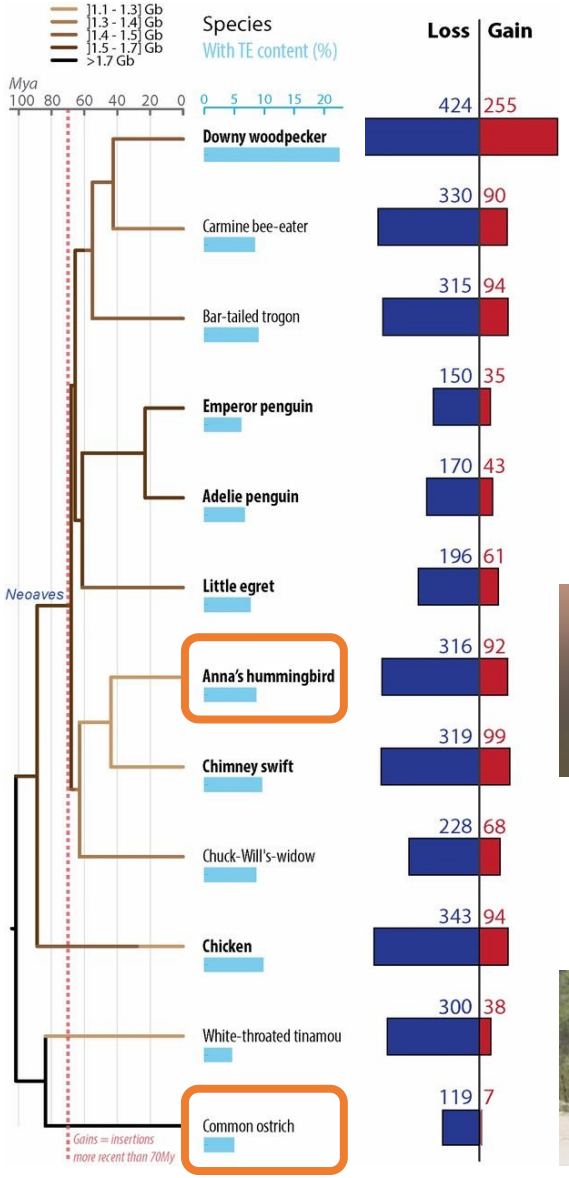

No constraint


Kapusta et al 2017, PNAS

1

# Genome size evolution

## Accordion model




No constraint

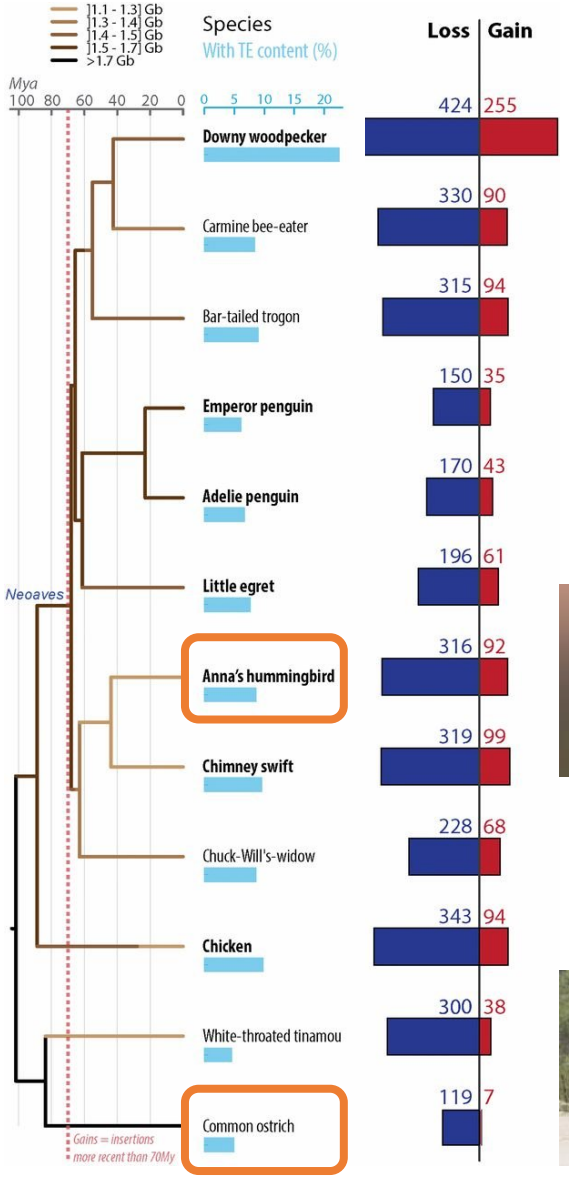

Consider not only host popgen, but also TE popgen!

Kapusta et al 2017, PNAS

1

# Genome size evolution

## Accordion model



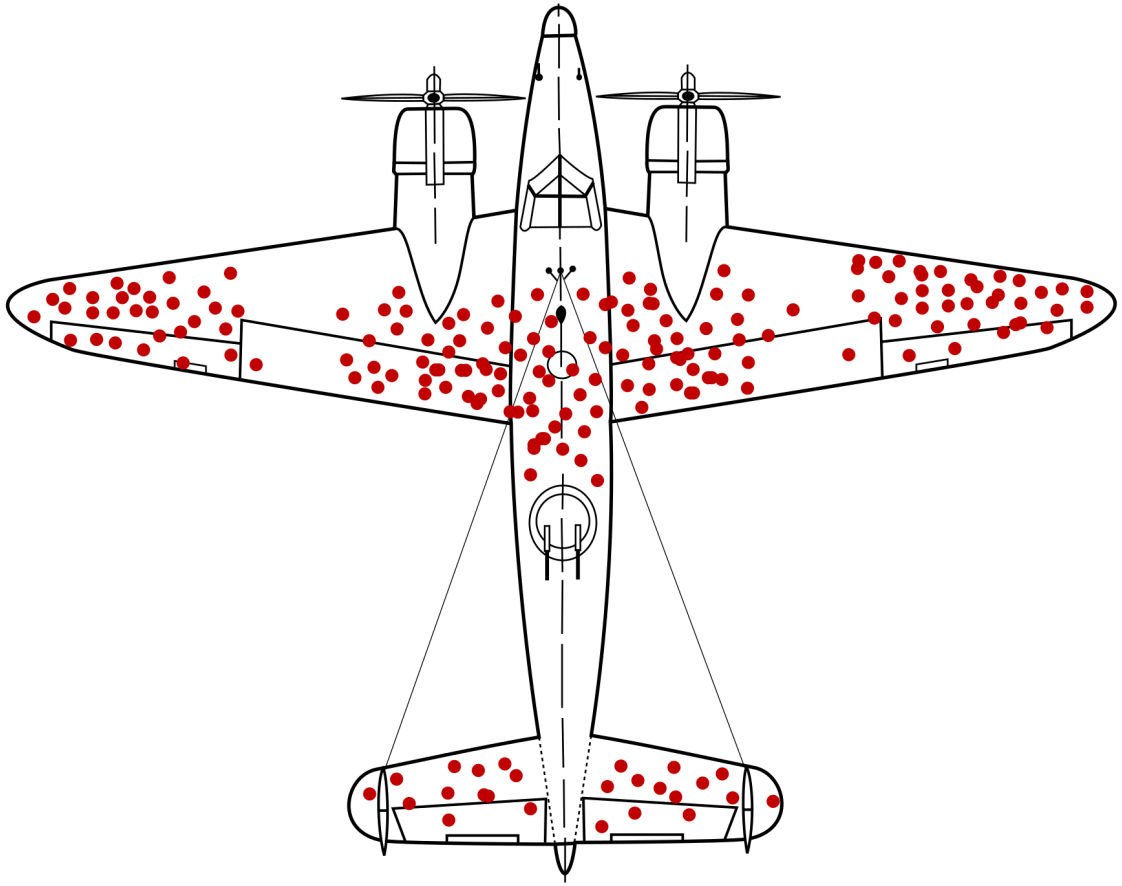

No constraint


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

Kapusta et al 2017, PNAS

# Survivorship bias



pattern != process

1

# Survivorship bias

- More TEs because of more insertions or more fixations?

  TE popgen vs host popgen

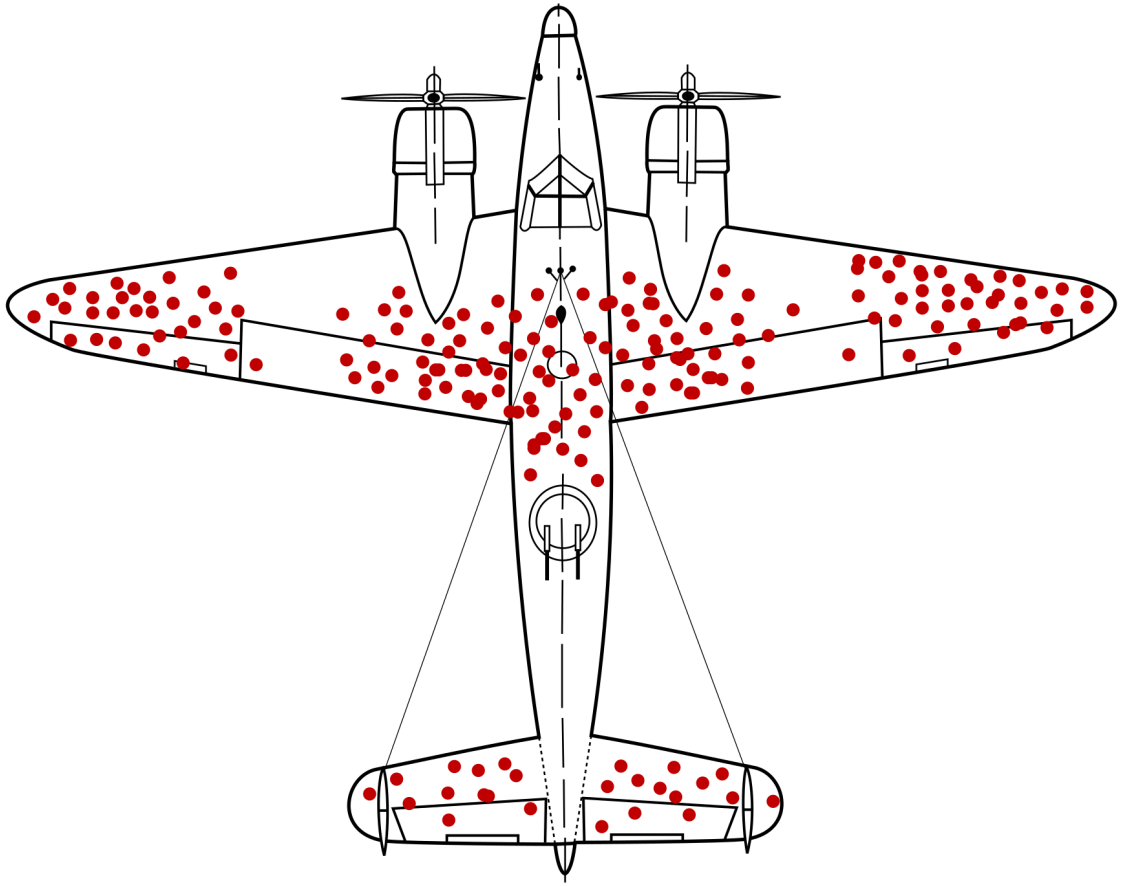
- Fewer TEs because of preferential insertion or negative selection?

  you cannot see detrimental insertions

- Larger genome because of more TE accumulation or fewer deletions?

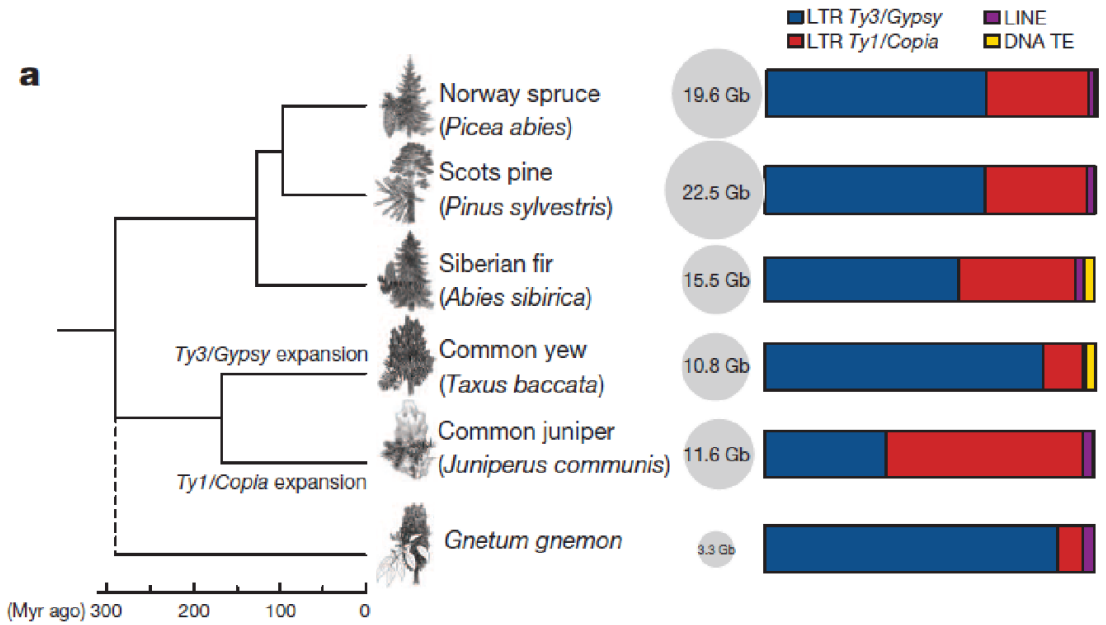
  lower deletion rate

- Smaller genome because of lower TE accumulation or more deletions?

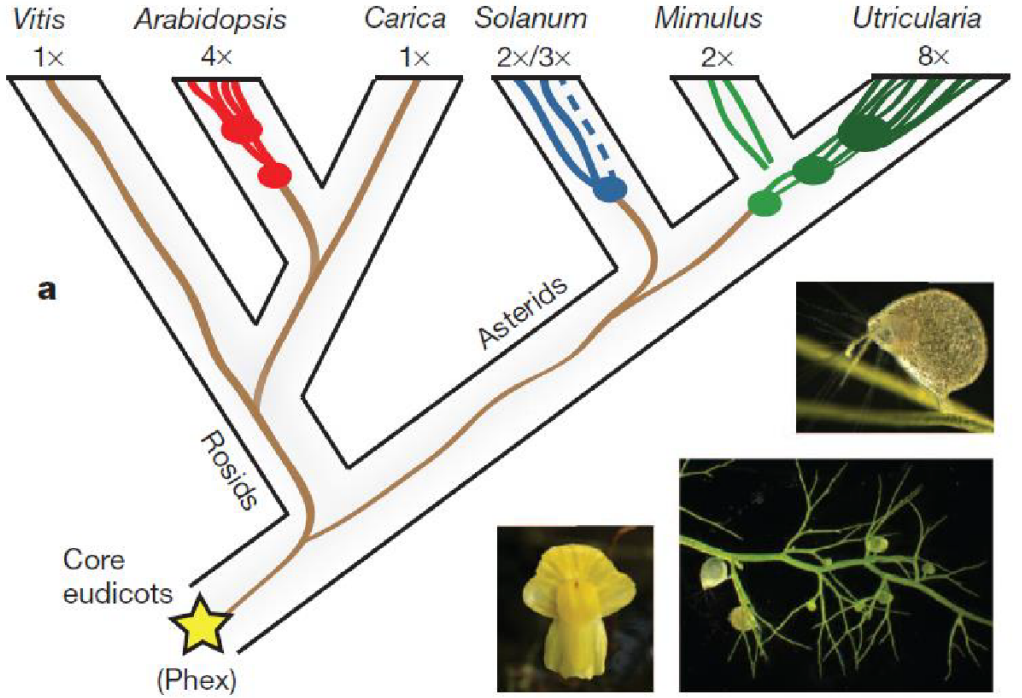


pattern != process

# Genomic gigantism and dwarfism




Tons of TE activity/accumulation or lack of deletions?



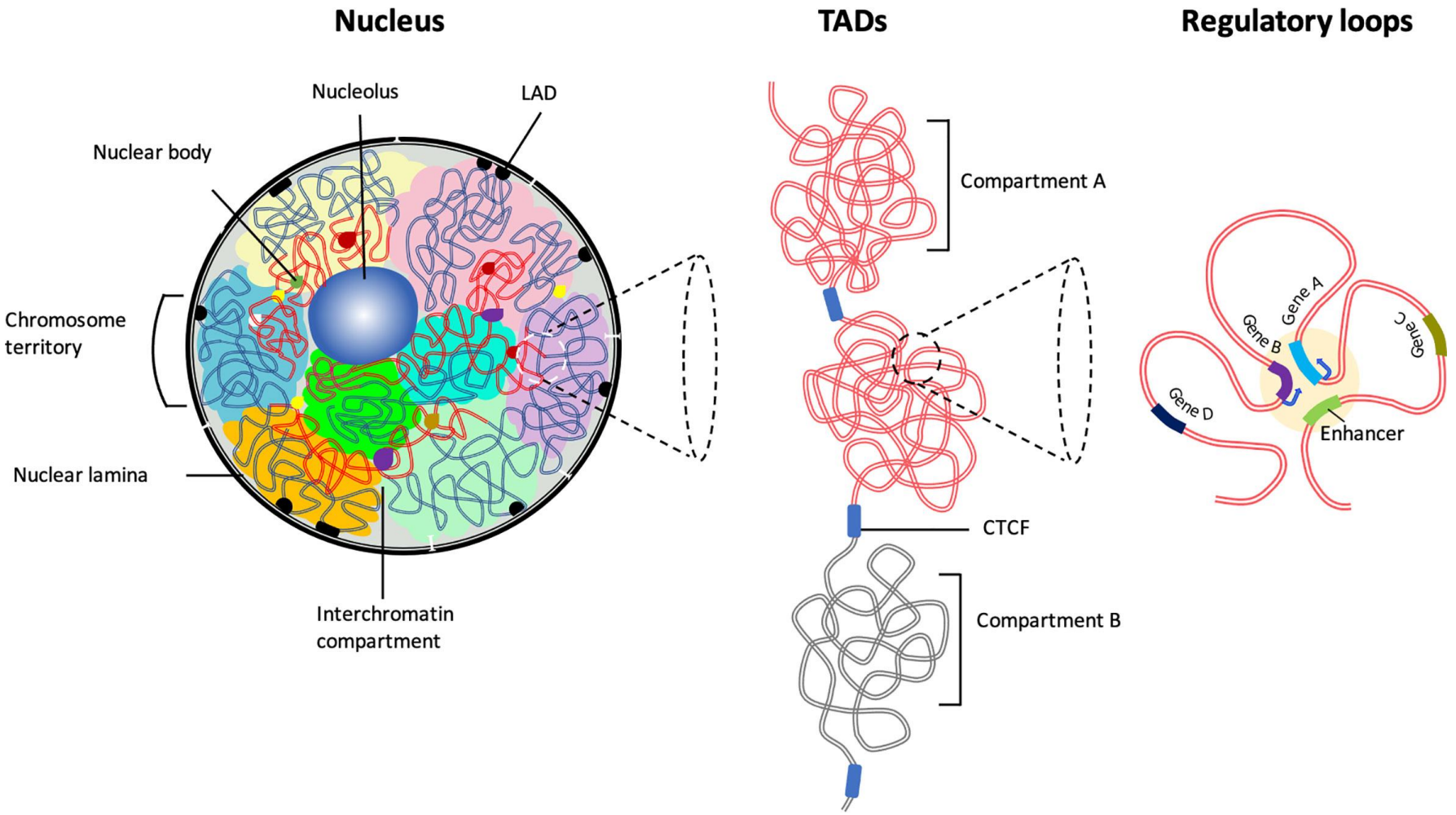

**77 Mb genome** despite 3 rounds of whole-genome duplications, only 3% repetitive DNA!

Lack of TE activity/accumulation or tons of deletions?

Nysted et al 2013, Nature

1

# 3D genome structure



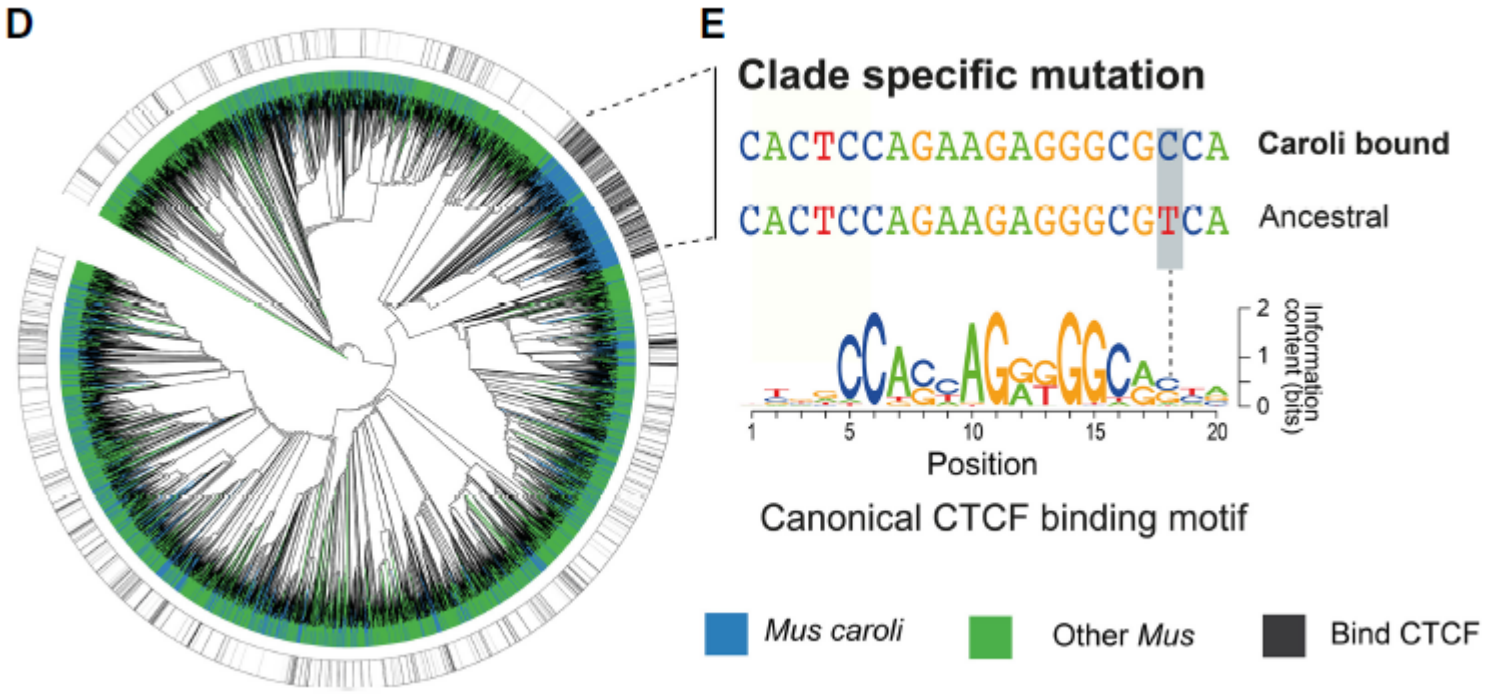

Pongubala and Murre 2021, Front Immunol

# 3D genome structure

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

# 3D genome structure

## MIR retrotransposon sequences provide insulators to the human genome

Jianrong Wang[a], Cristina Vicente-García[b,c], Davide Seruggia[b,c], Eduardo Moltó[b,c], Ana Fernandez-Miñán[d], Ana Neto[d], Elbert Lee[e], José Luis Gómez-Skarmeta[d], Lluís Montoliu[b,c], Victoria V. Lunyak[e], and I. King Jordan[a,f,1]

**MIR insulators appear to be CCCTC-binding factor (CTCF) independent** and show a distinct local chromatin environment with marked peaks for RNA Pol III and a number of histone modifications, suggesting that MIR insulators recruit transcriptional complexes and chromatin modifying enzymes in situ to help establish chromatin and regulatory domains in the human genome. **The provisioning of insulators by MIRs across the human genome suggests a specific mechanism by which TE sequences can be used to modulate gene regulatory networks.**

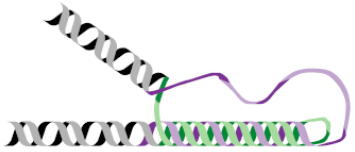
# Non-B DNA and TEs

**Slipped strand**

Direct repeat (28 bp)

Repeat Spacer Repeat

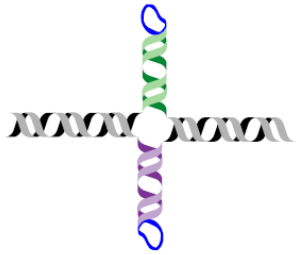
**Triple helix (H-DNA)**



Mirror repeat (58 bp)

Repeat Repeat

**Cruciform**



Inverted repeat (21 bp)

Repeat Spacer Repeat

**Bent DNA**

A-phased repeat (26 bp)

AAAA $N_{4-9}$ AAAA $N_{4-9}$ AAAA

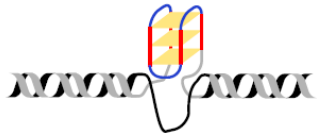
A-tract Spacer A-tract Spacer A-tract

**Z-DNA**

Purine-pyrimidine motif (12 bp)

$(YR-YR)_n$

**G-quadruplex (G4)**



G4 motif (35 bp)

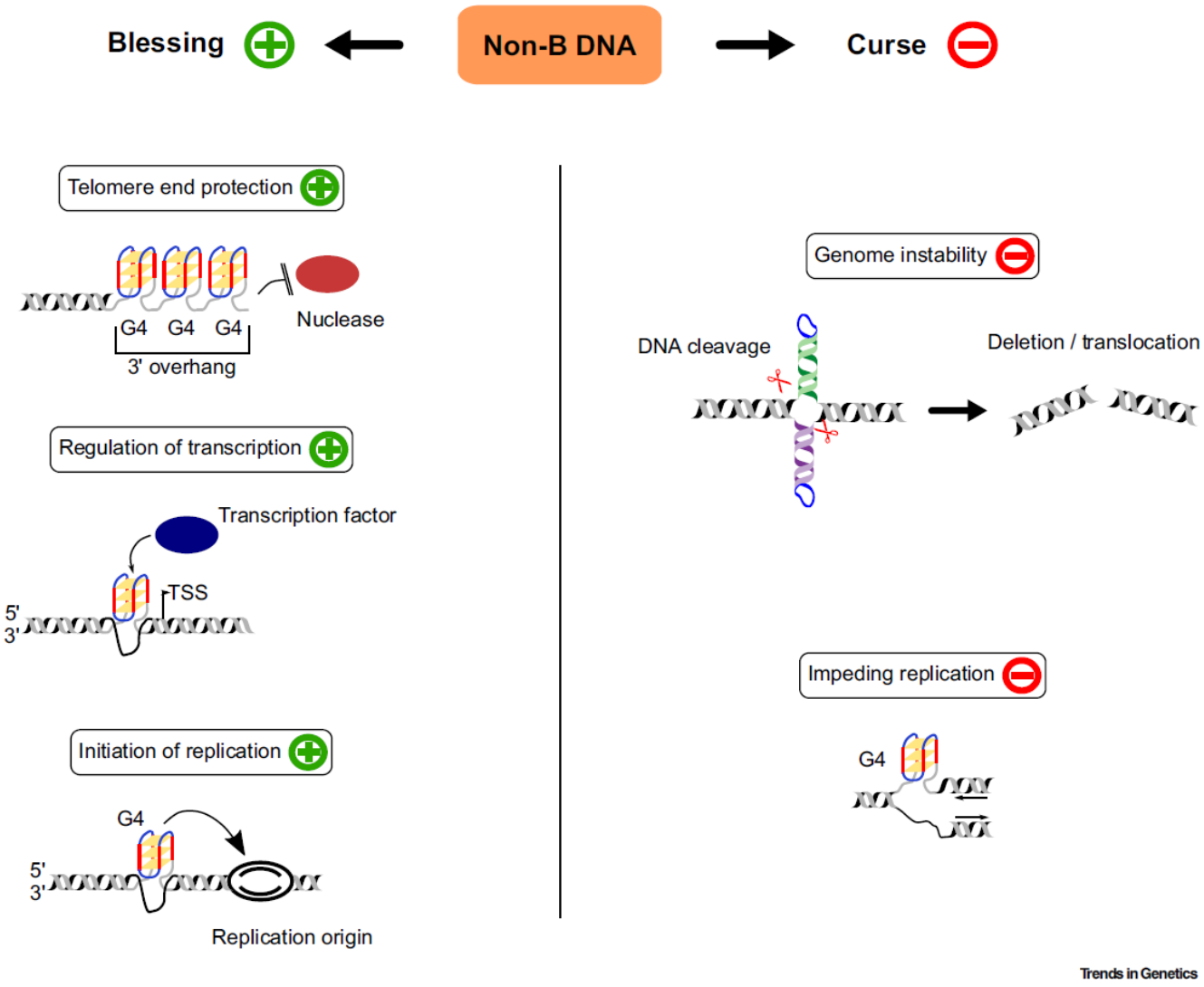
GGGNNNGGGNNNGGGNNNGGG

Stem-Loop Stem-Loop Stem-Loop Stem

Trends in Genetics

Makova and Weissensteiner, 2022 Trends in Genetics

# Non-B DNA and TEs



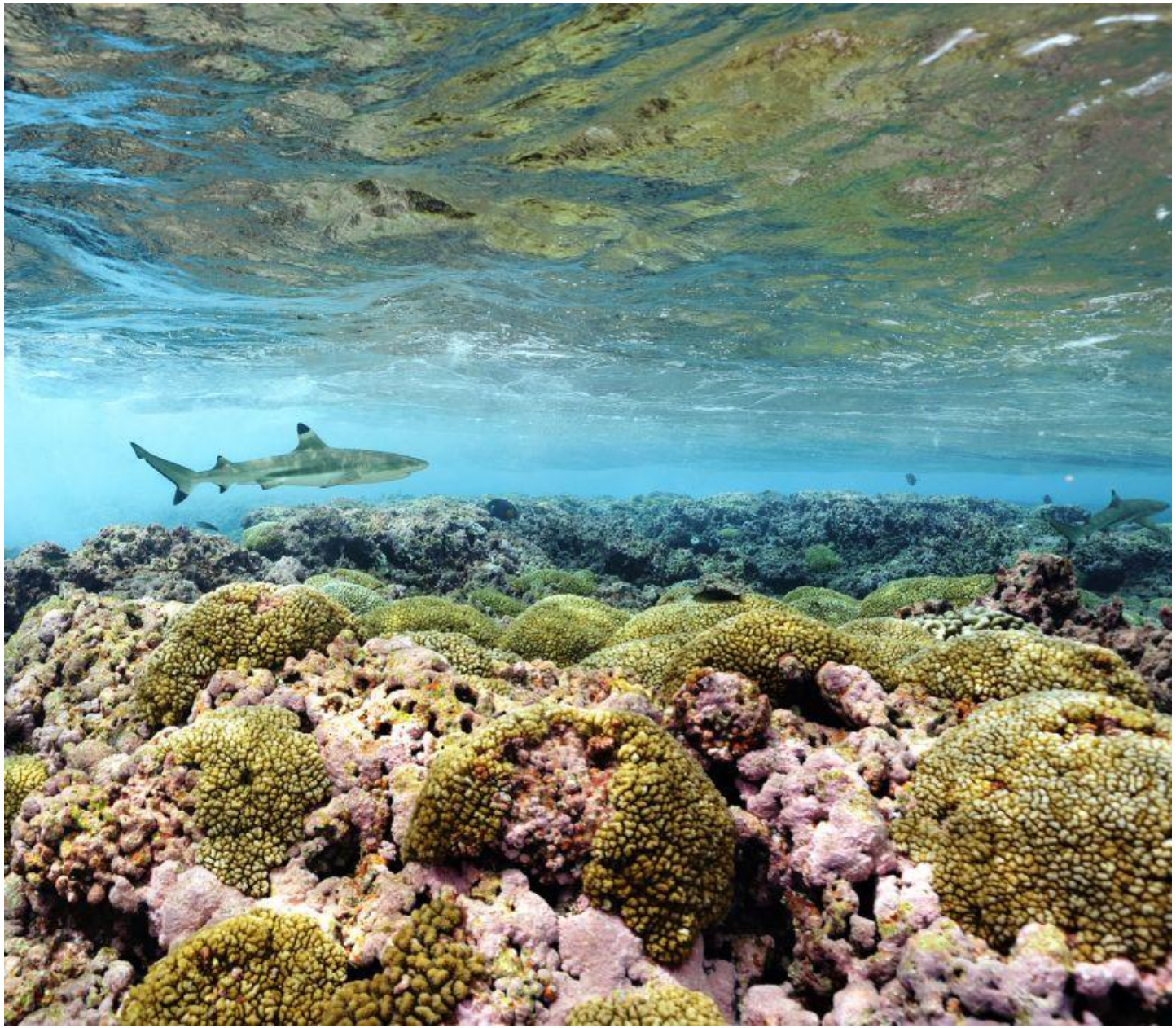

**TEs were proposed to serve as vehicles of spreading non-B DNA across the genome**

Non-B DNA located in TEs may play a role in:

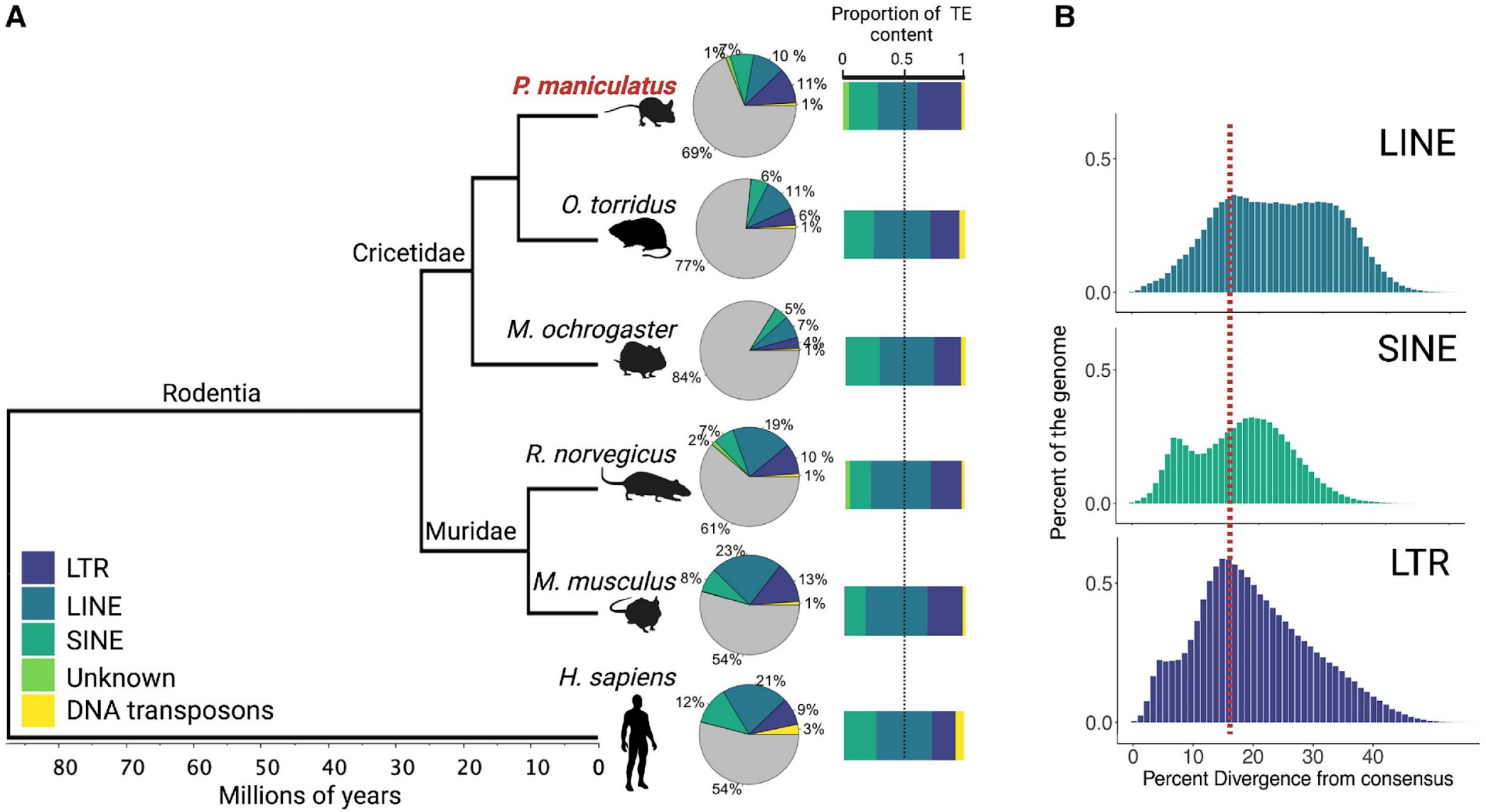
- regulating activity of neighboring genes by affecting the activity of promoters
- regulating epigenetic state of discrete genomic regions and facilitating genome silencing, including TE silencing, via spreading heterochromatin
- reshuffling of genomic DNA via recombination at TEs
- G4 motifs within hominid-specific SVA retrotransposons are enriched in cancer genome breakpoints

Makova and Weissensteiner, 2022 Trends in Genetics

# Genomes as ecosystems

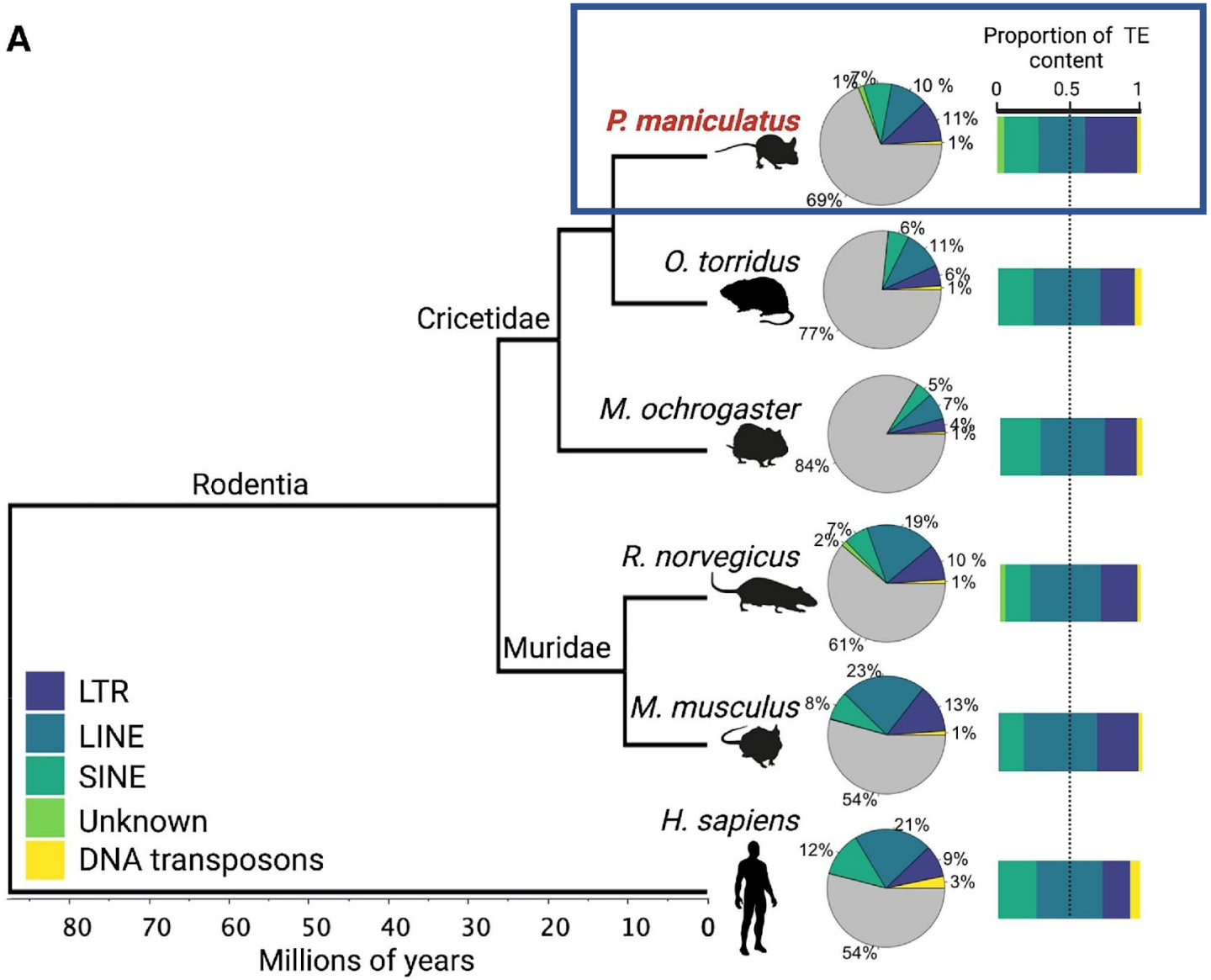


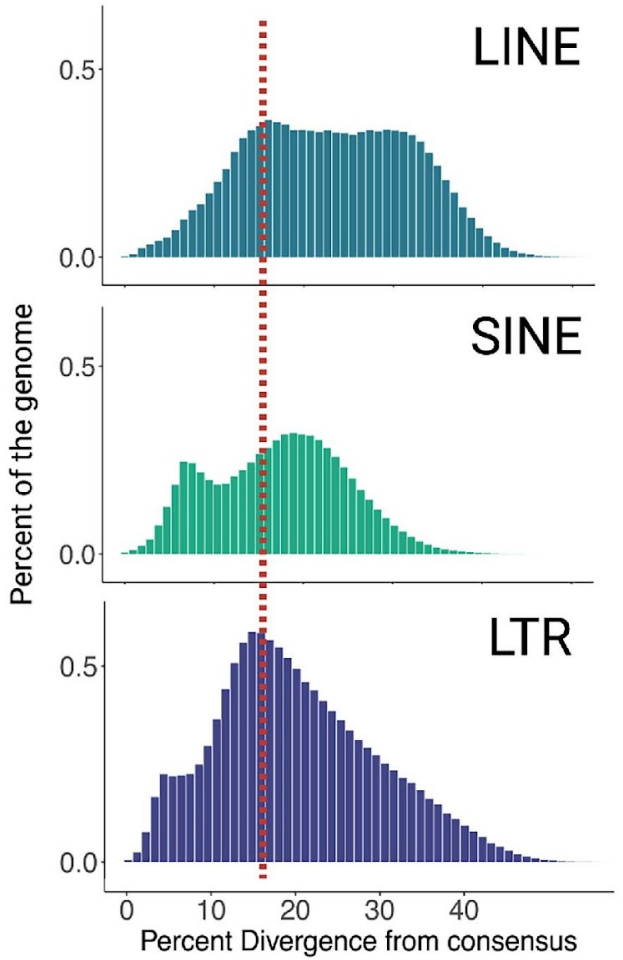
# Genomes as ecosystems




Gozashti et al, 2023 MBE

# Genomes as ecosystems



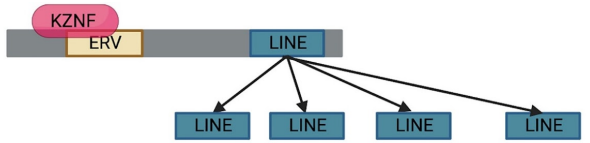




Gozashti et al, 2023 MBE

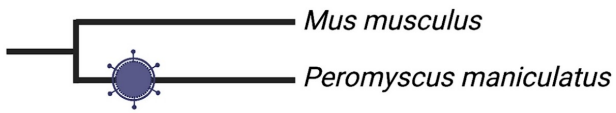
# Genomes as ecosystems

Time

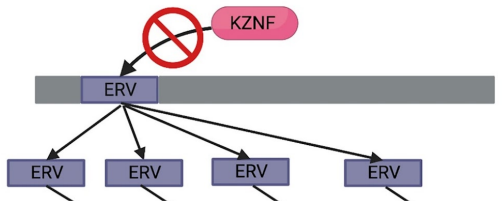
LINEs outcompete ERVs in ancestral lineage of *P. maniculatus* and *M. musculus* partly due to biased host defense targeting of ERVs




MysTR endogenized




Host defenses initially ineffective against new foreign ERVs



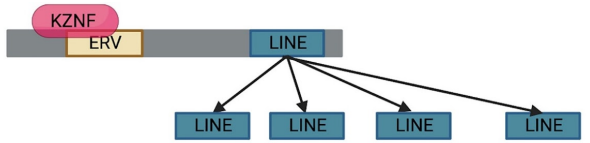

Massive expansion of ERVs (followed by expansion of KZNFs)

Gozashti et al, 2023 MBE

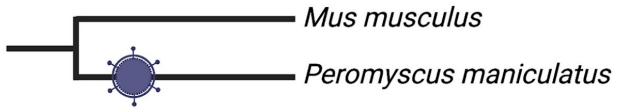
# Genomes as ecosystems

Time

LINEs outcompete ERVs in ancestral lineage of *P. maniculatus* and *M. musculus* partly due to biased host defense targeting of ERVs

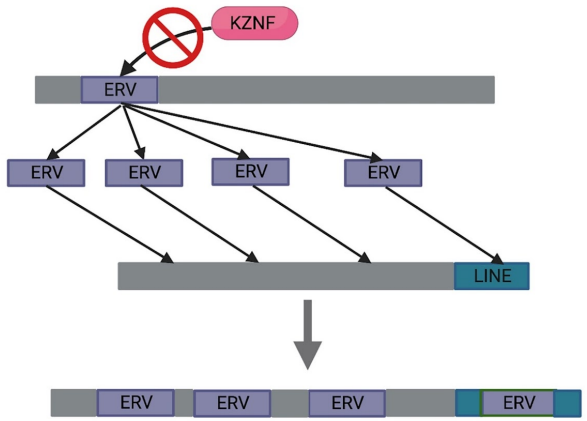



MysTR endogenized




Host defenses initially ineffective against new foreign ERVs

Massive expansion of ERVs (followed by expansion of KZNFs)




ERVs outcompete LINEs for genomic sites

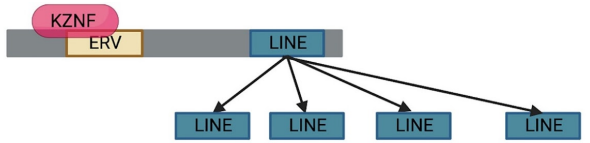
Fewer transposition-competent LINEs

Gozashti et al, 2023 MBE

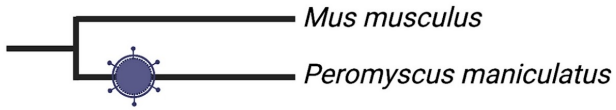
# Genomes as ecosystems

Time

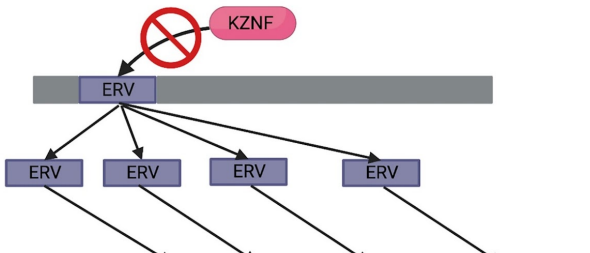
LINEs outcompete ERVs in ancestral lineage of *P. maniculatus* and *M. musculus* partly due to biased host defense targeting of ERVs




MysTR endogenized




Host defenses initially ineffective against new foreign ERVs

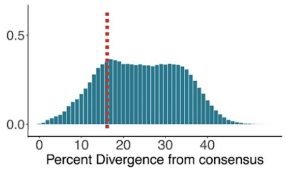



Massive expansion of ERVs (followed by expansion of KZNFs)

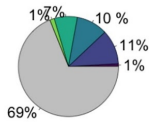
ERVs outcompete LINEs for genomic sites

Fewer transposition-competent LINEs


Reduced LINE activity in deer mouse lineage

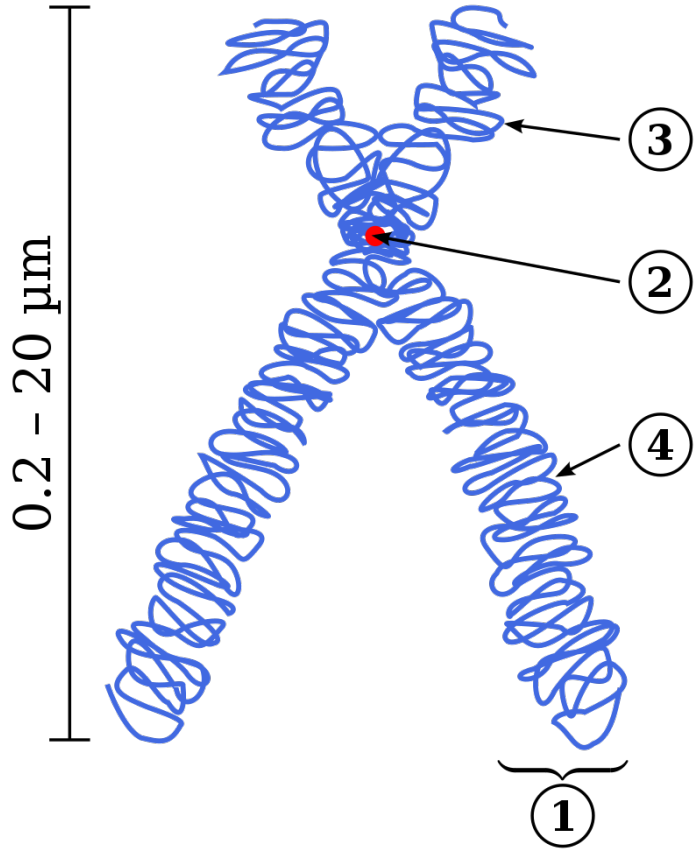


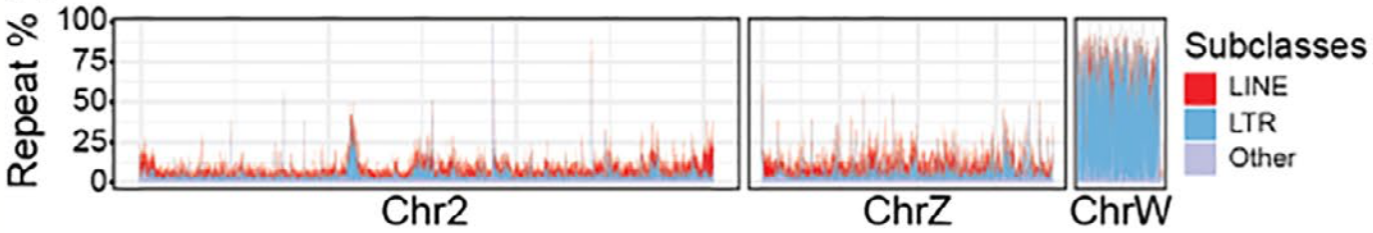

ERVs occupy more of the genome than LINEs




Gozashti et al, 2023 MBE

# TEs are not uniformily distributed




Peona et al. 2021a, Mol Ecol Res

2

# Target specificity

R2

RESEARCH ARTICLE | STRUCTURAL BIOLOGY

## Structure of the R2 non-LTR retrotransposon initiating target-primed reverse transcription

MAX E. WILKINSON, CHRIS J. FRANGIEH, RHIANNON K. MACRAE, AND FENG ZHANG Authors Info & Affiliations

SCIENCE • 6 Apr 2023 • Vol 380, Issue 6642 • pp. 301-308 • DOI: 10.1126/science.adg7883

RESEARCH ARTICLE | GENETICS

## The retrotransposon R2 maintains *Drosophila* ribosomal DNA repeats

Jonathan O. Nelson, Alyssa Slicko, and Yukiko M. Yamashita Authors Info & Affiliations

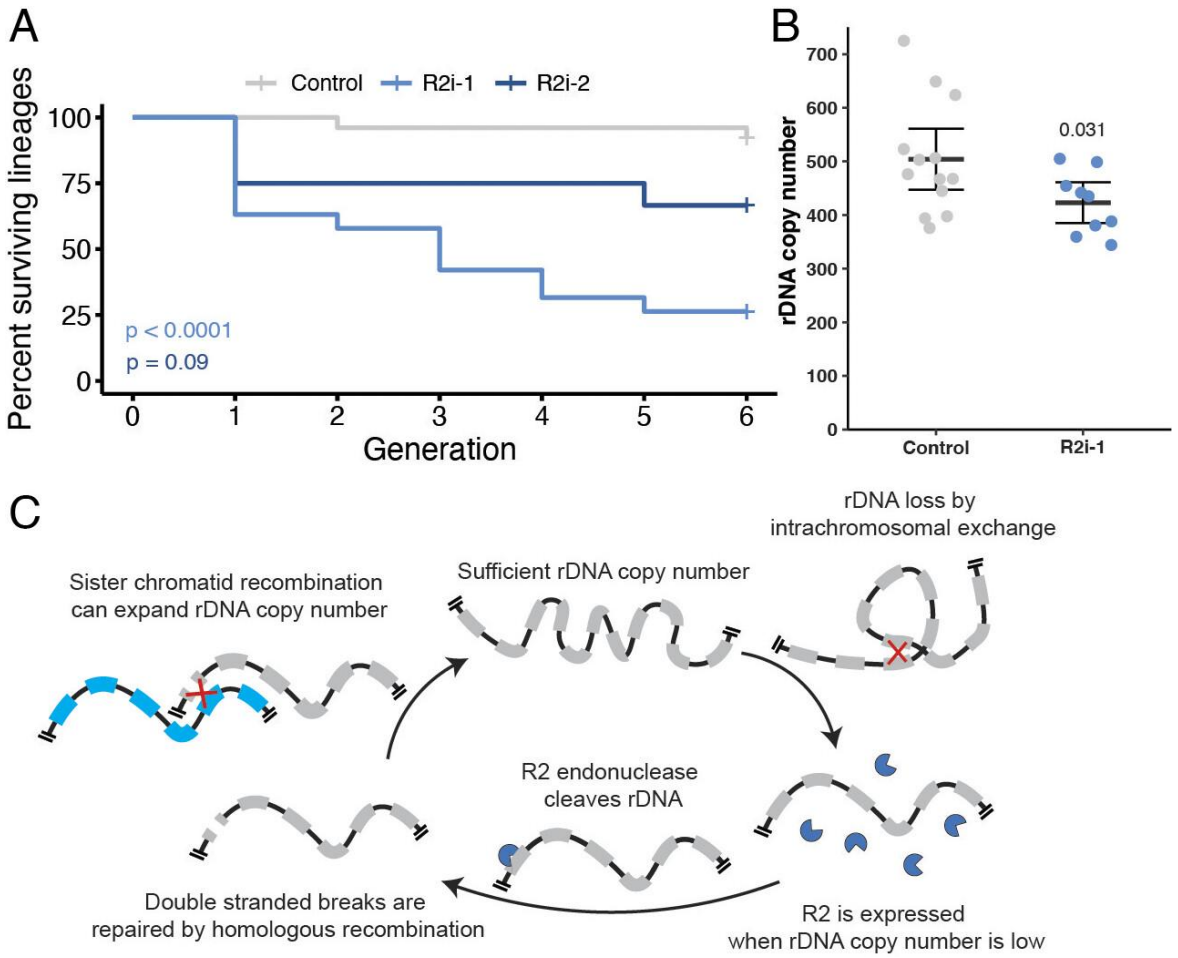
Edited by R. Scott Hawley, Stowers Institute for Medical Research, Kansas City, MO; received December 20, 2022; accepted May 3, 2023

**May 30, 2023** | 120 (23) e2221613120 | https://doi.org/10.1073/pnas.2221613120

https://twitter.com/maxewilkinson/status/1644043579067887617

# Target specificity

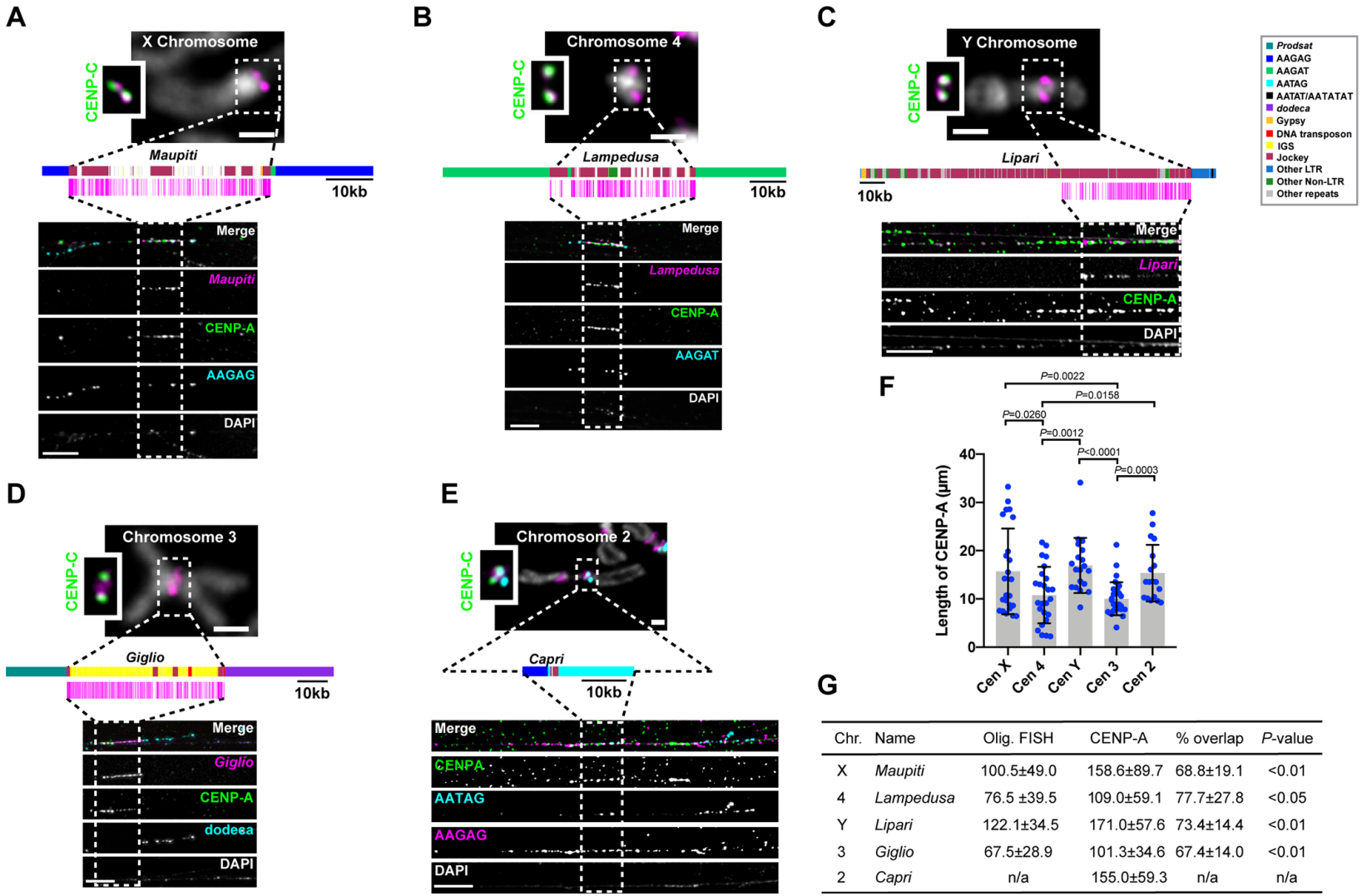
## R2




Nelson et al 2023, PNAS

# Non-recombining regions

## Centromeres



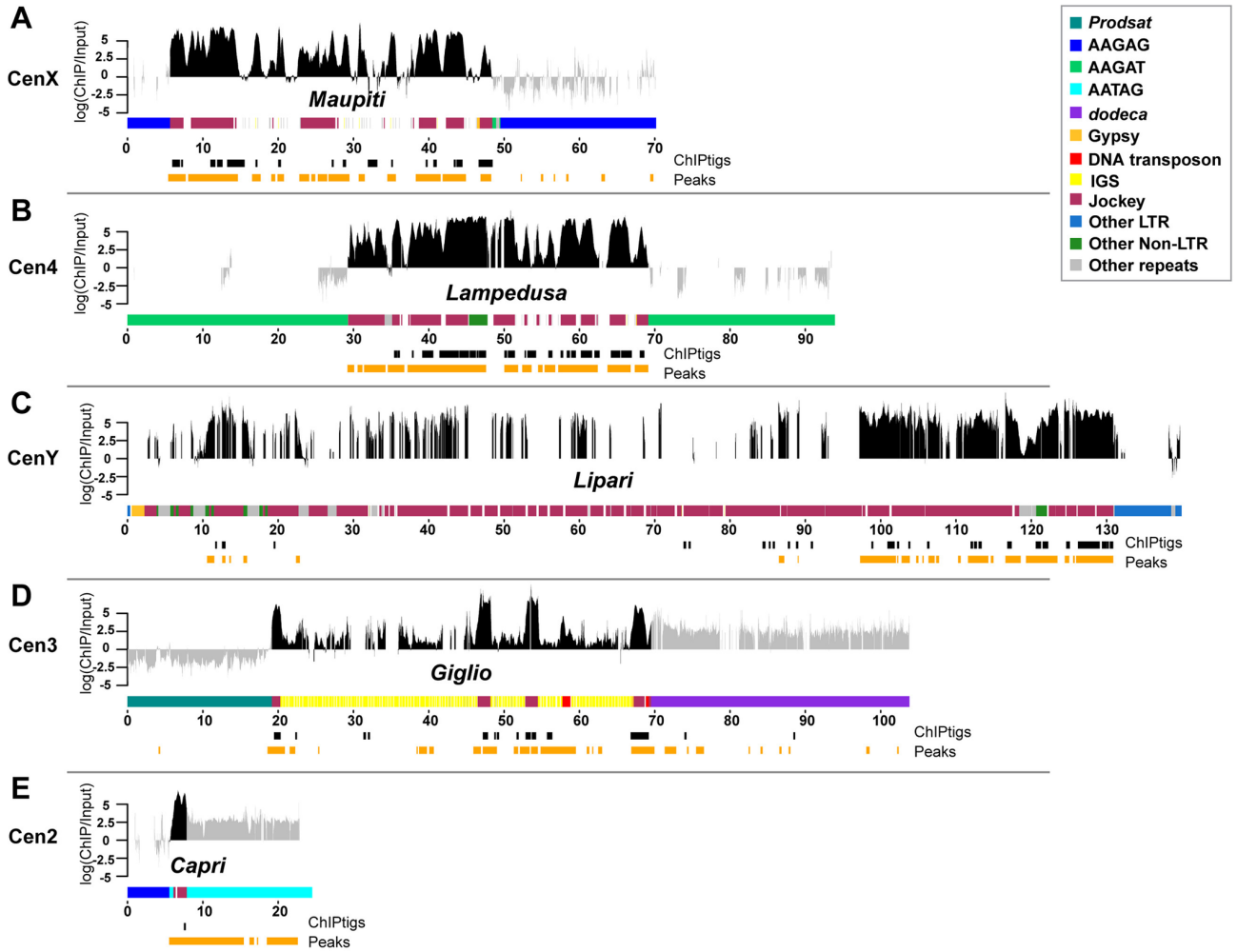

G

| Chr. | Name | Olig. FISH | CENP-A | % overlap | *P*-value |
|---|---|---|---|---|---|
| X | *Maupiti* | 100.5±49.0 | 158.6±89.7 | 68.8±19.1 | <0.01 |
| 4 | *Lampedusa* | 76.5 ±39.5 | 109.0±59.1 | 77.7±27.8 | <0.05 |
| Y | *Lipari* | 122.1±34.5 | 171.0±57.6 | 73.4±14.4 | <0.01 |
| 3 | *Giglio* | 67.5±28.9 | 101.3±34.6 | 67.4±14.0 | <0.01 |
| 2 | *Capri* | n/a | 155.0±59.3 | n/a | n/a |

Chang et al 2019

# Non-recombining regions

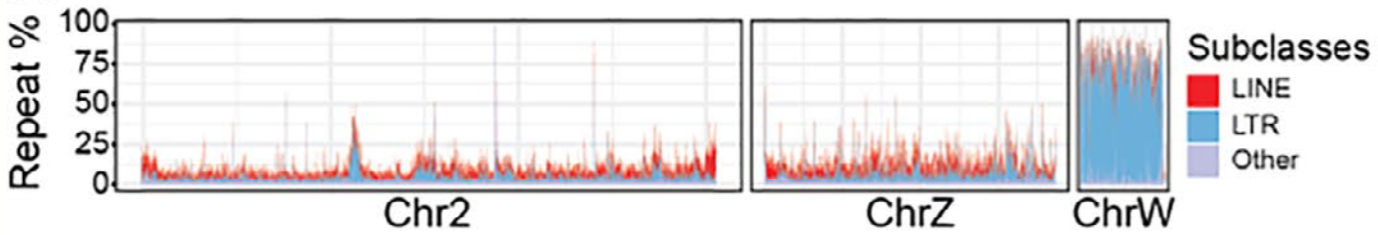
## Centromeres




Chang et al 2019

# Non-recombining regions

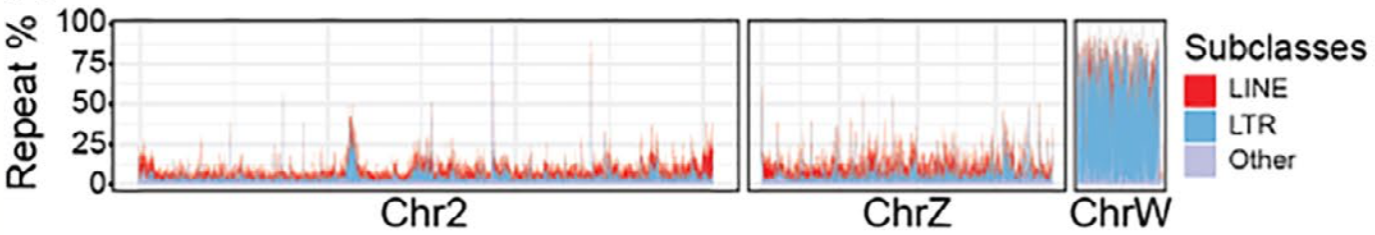
Sex chromosomes



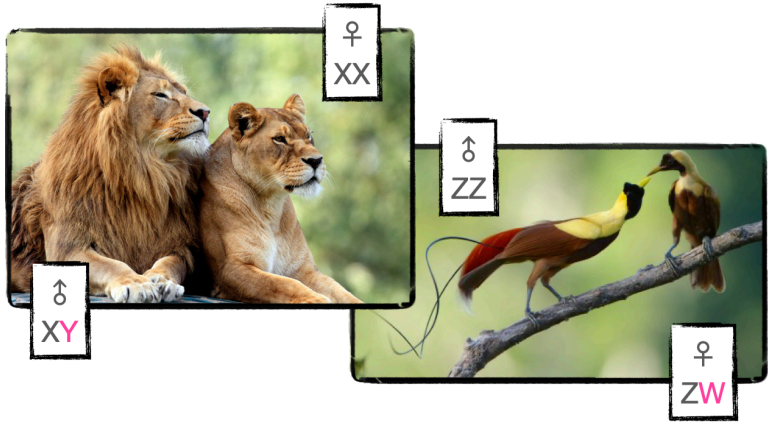

Peona et al. 2021a, Mol Ecol Res

# Non-recombining regions

## Sex chromosomes




Genetic sex determination

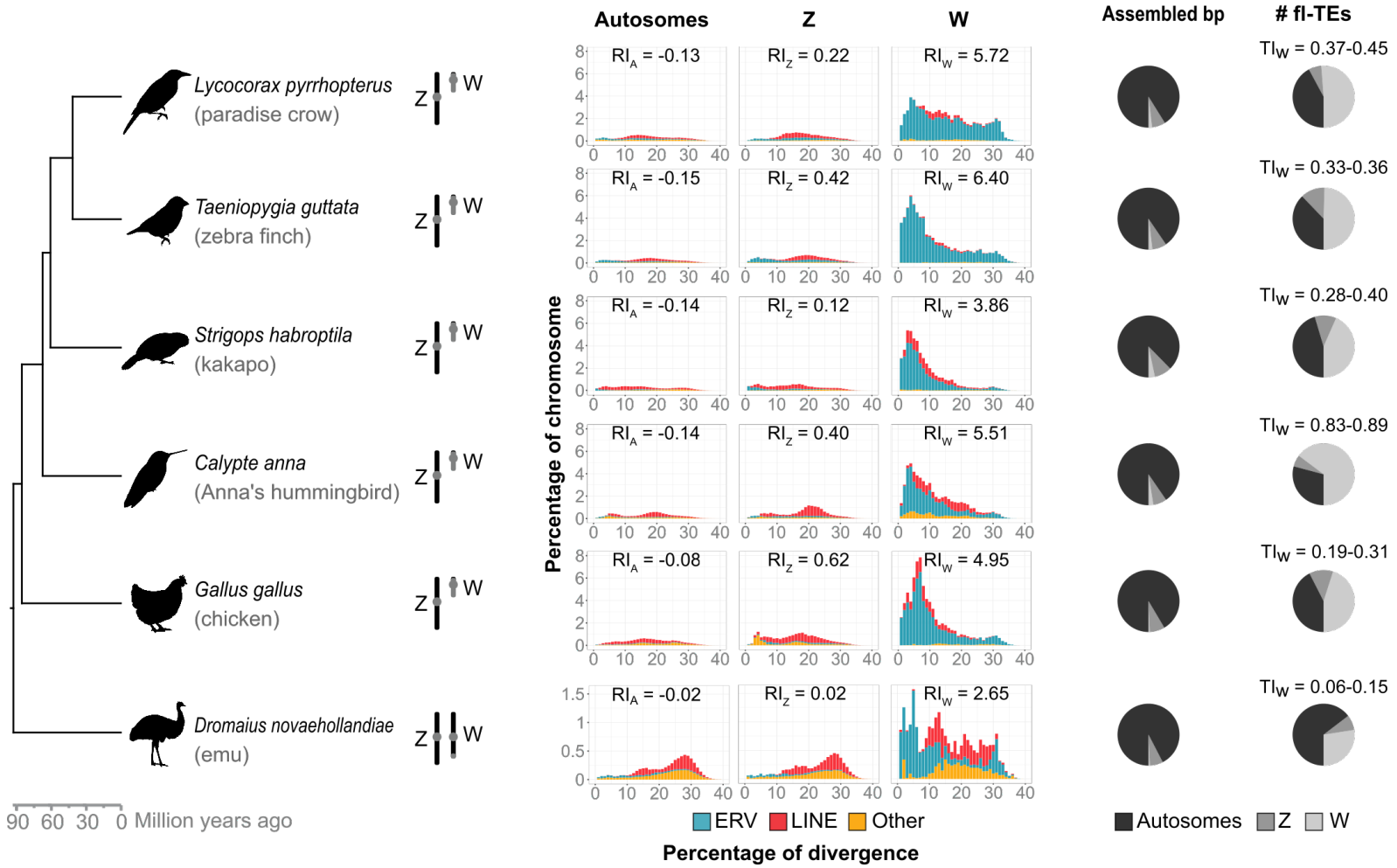



Peona et al. 2021a, Mol Ecol Res

2

# Non-recombining regions

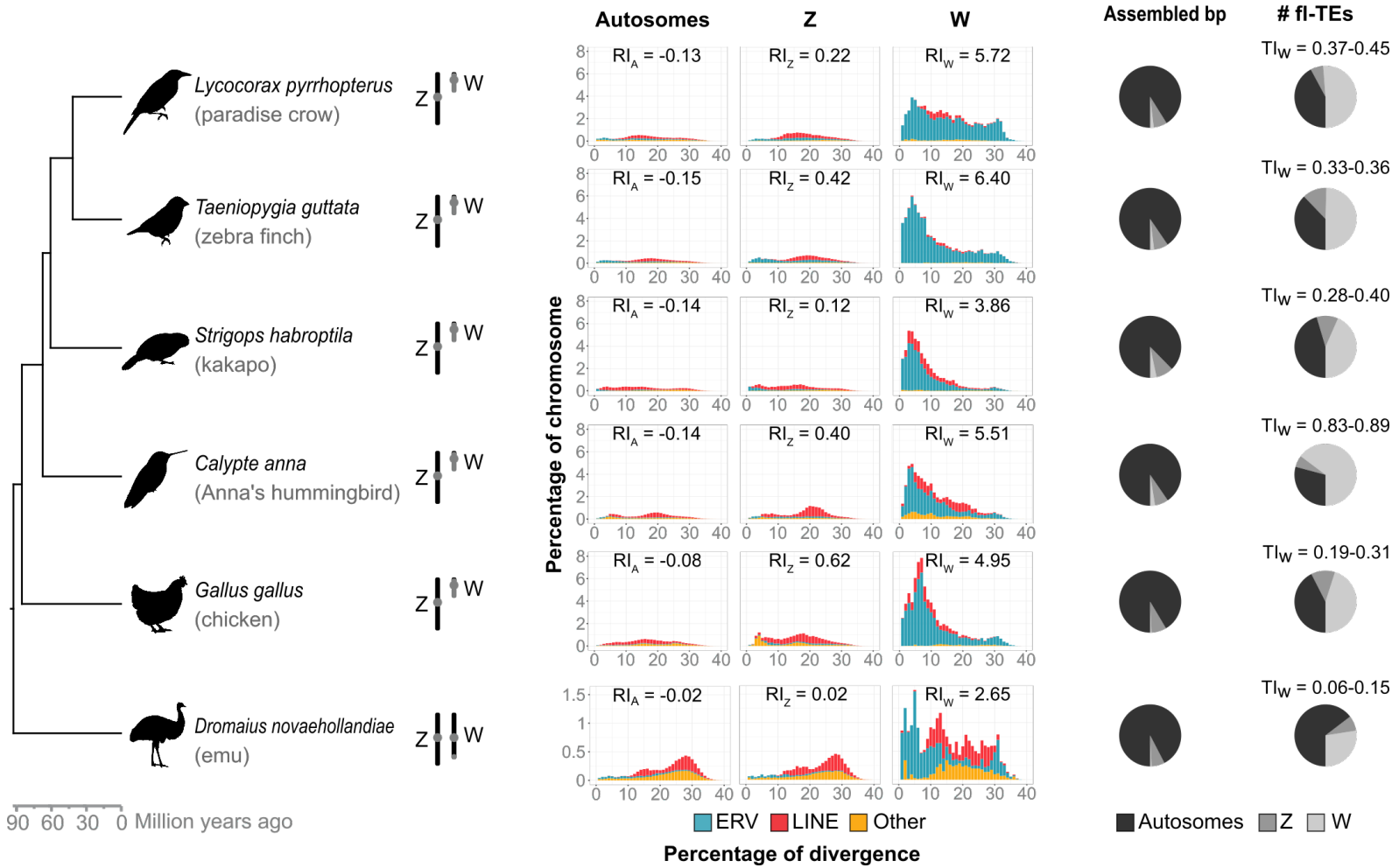
## Sex chromosomes




Peona et al. 2021b, Phil Trans B

# Non-recombining regions

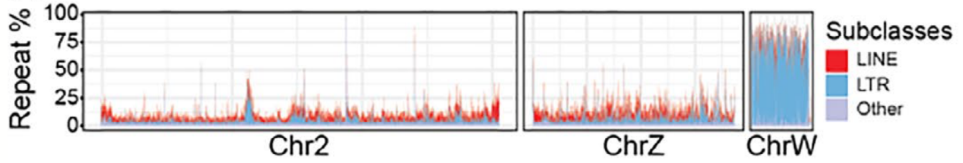
## Sex chromosomes



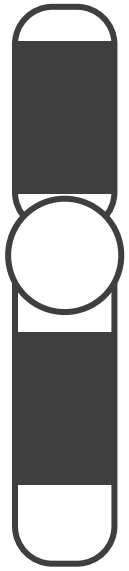

Half of all intact TEs is on W alone

Peona et al. 2021b, Phil Trans B

# Active transposable elements

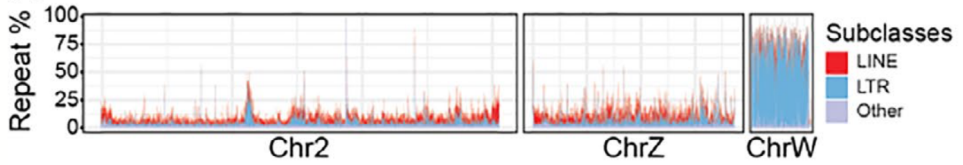



W

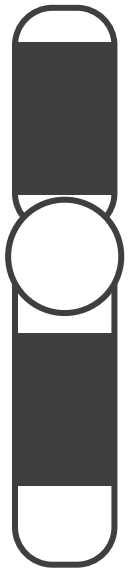


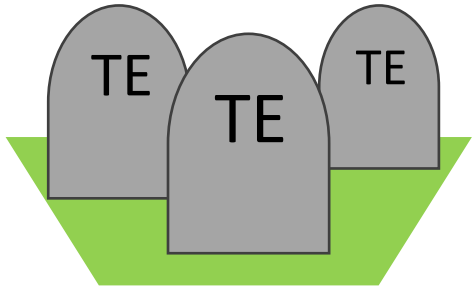
TE = transposable element

# Active transposable elements



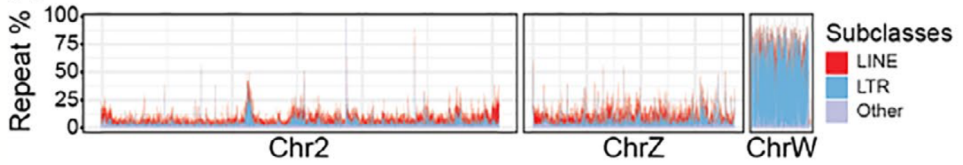

W



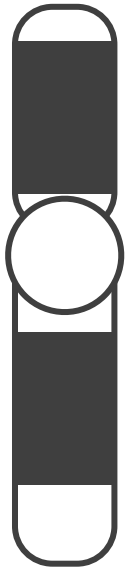



TE = transposable element

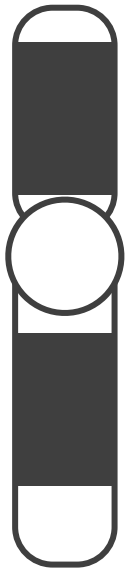
# Active transposable elements

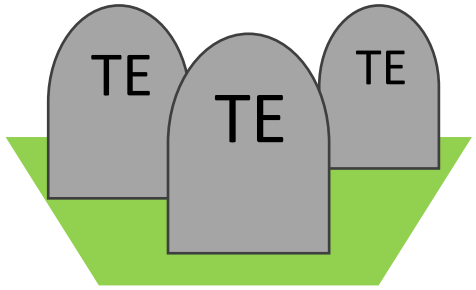



W



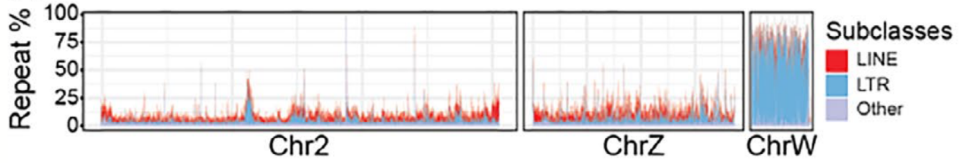
W

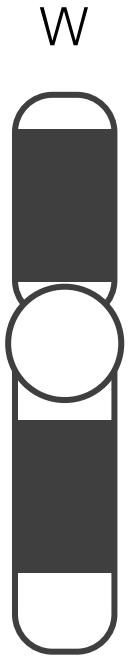


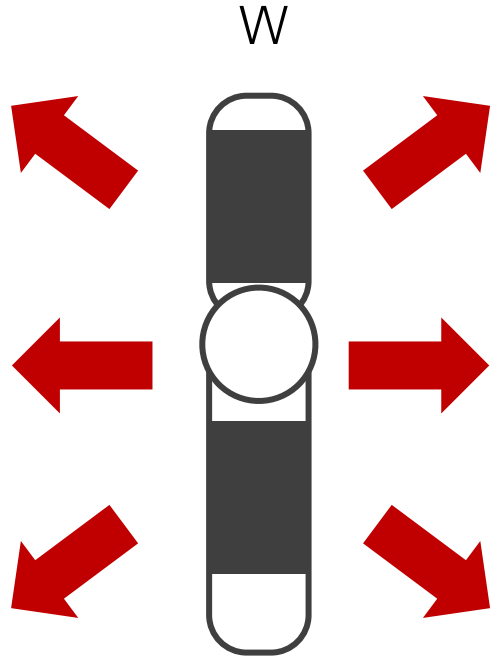


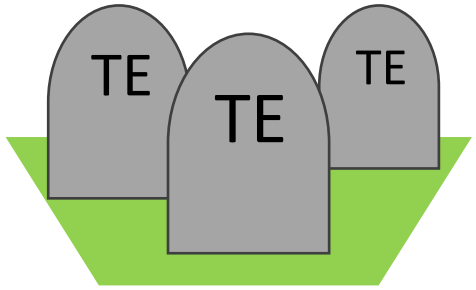

TE = transposable element

# Active transposable elements



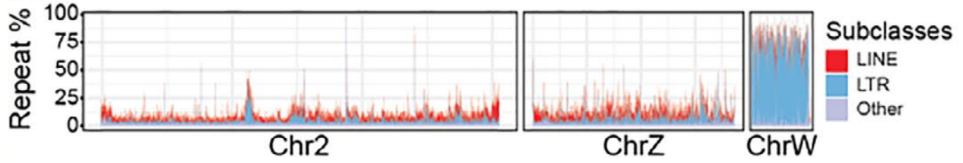



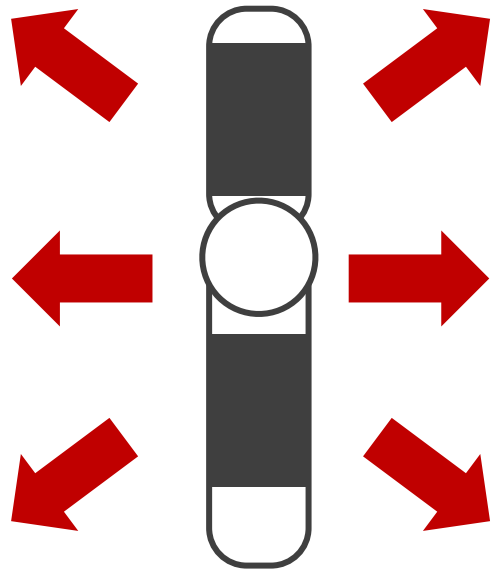

TE = transposable element

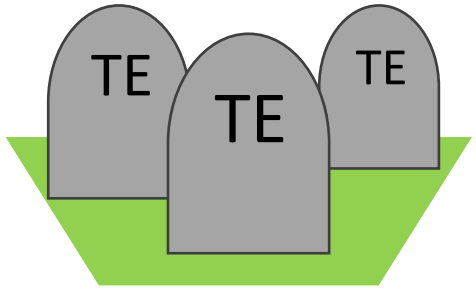
# Active transposable elements




W

W






TE = transposable element

2

# Find W-linked TEs

Kinsella et al.,

2

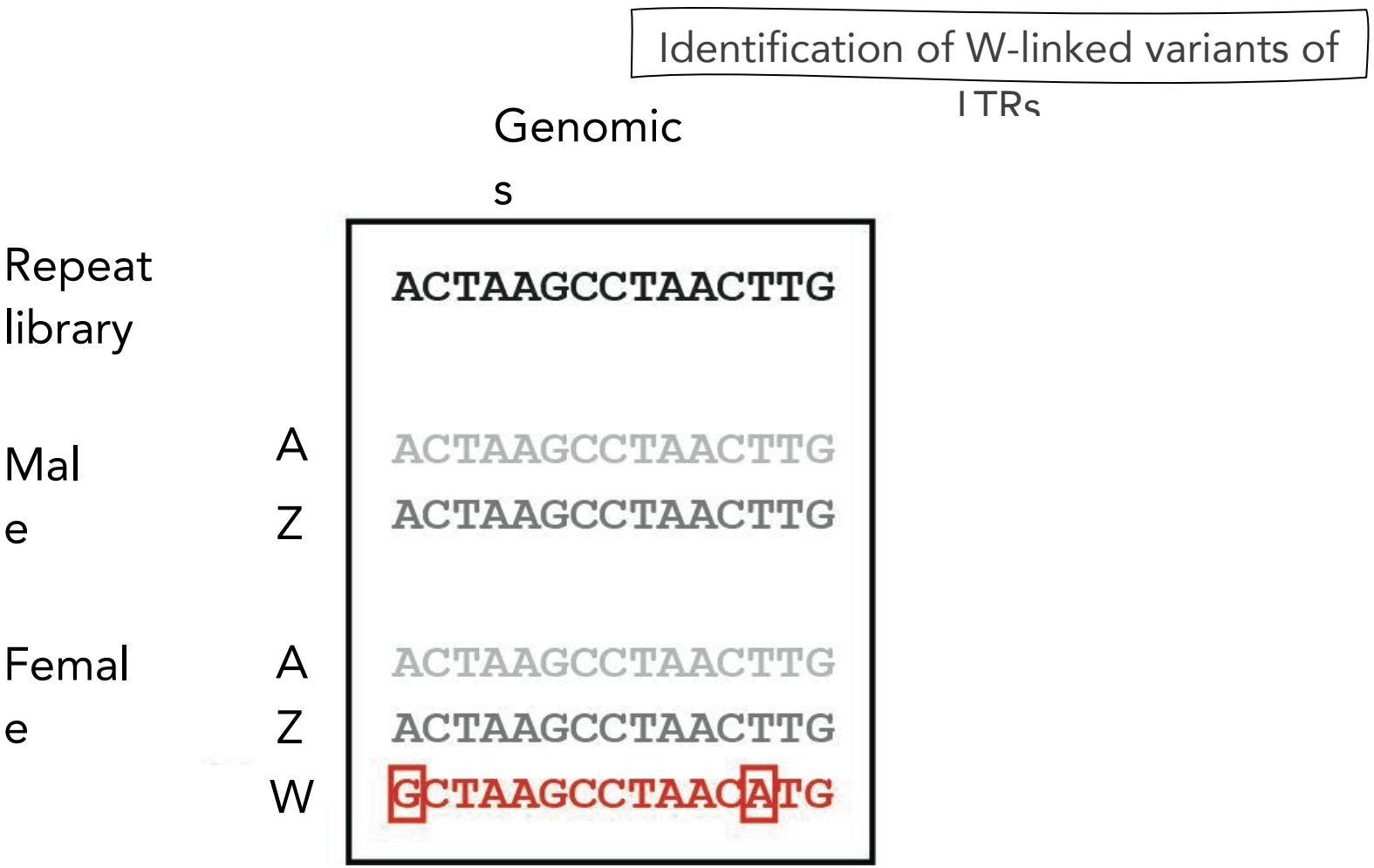
# Find W-linked TEs

Identification of W-linked variants of ITRs

Kinsella et al.,

# Find W-linked TEs

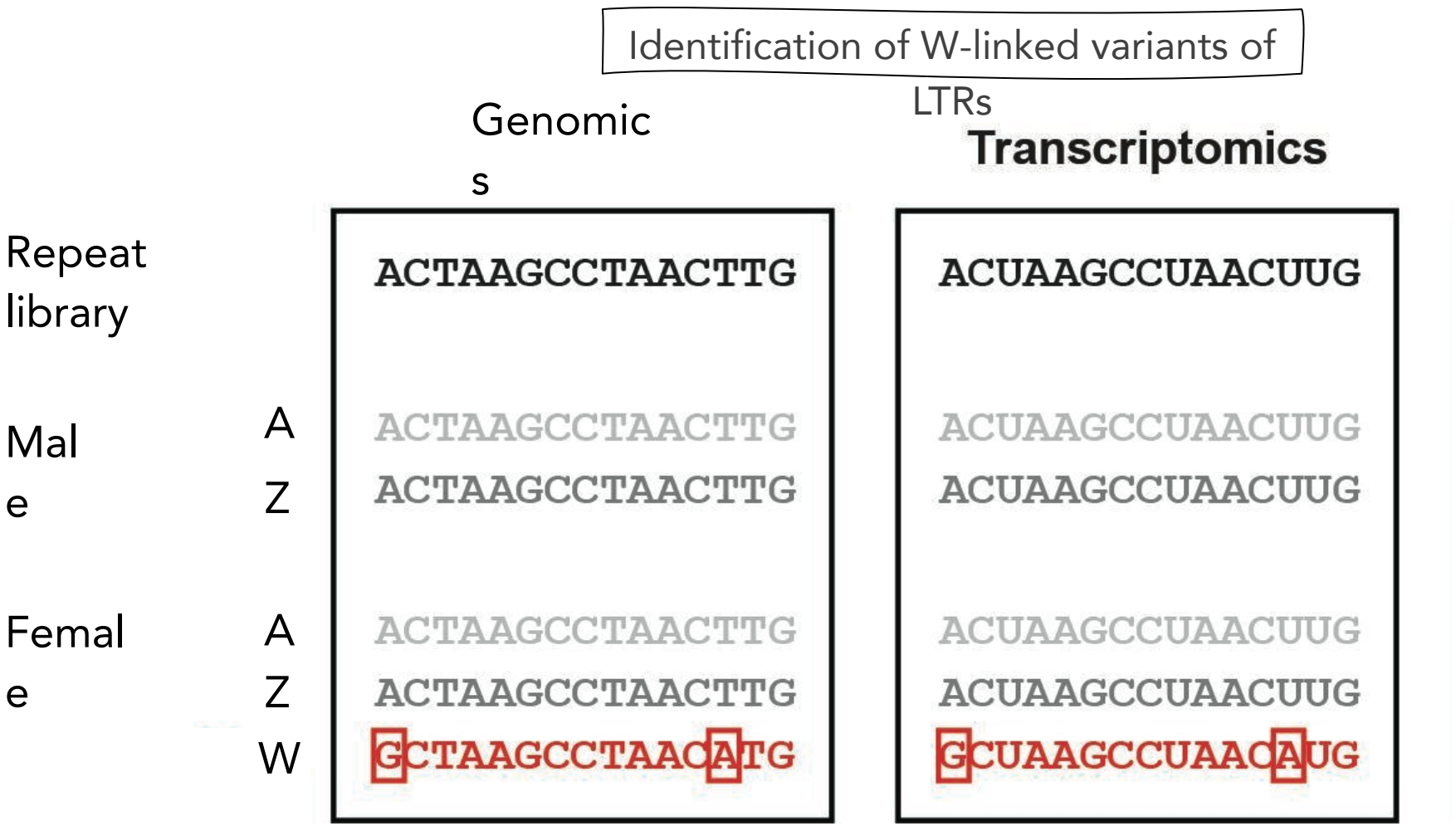
Identification of W-linked variants of ITRs




Kinsella et al.,

# Find W-linked TEs

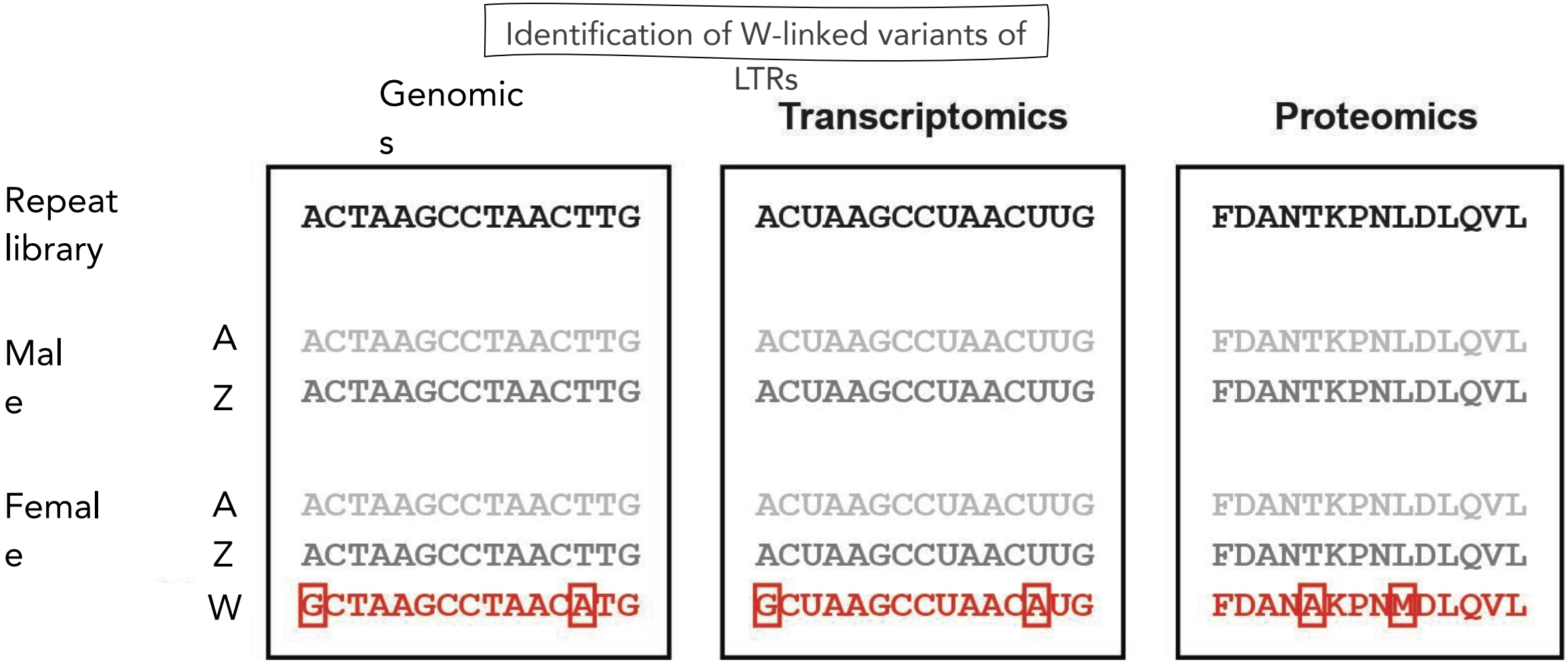
Identification of W-linked variants of LTRs




Kinsella et al.,

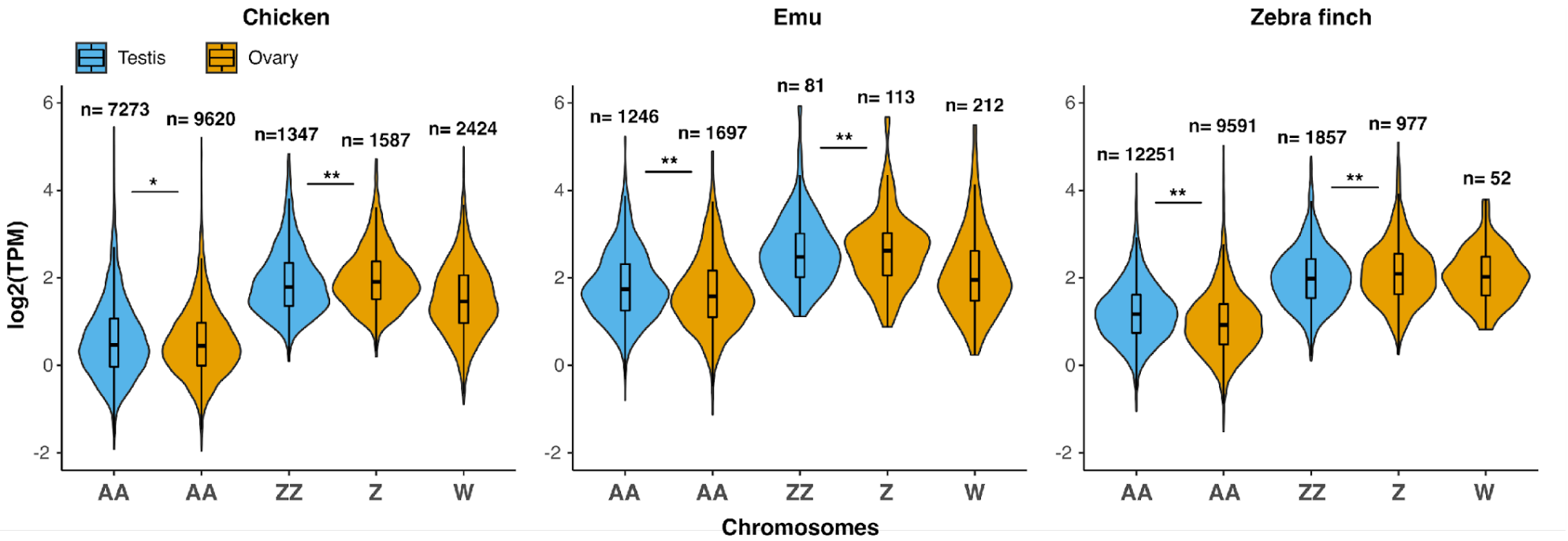
# Find W-linked TEs

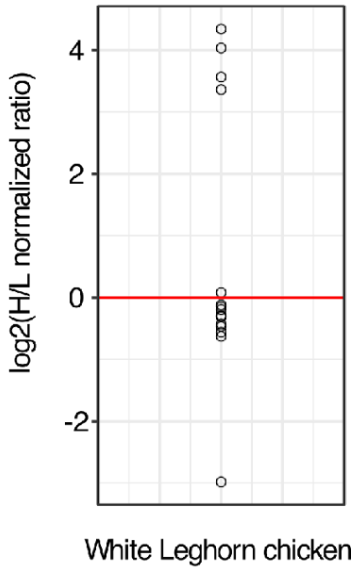
Identification of W-linked variants of LTRs




Kinsella et al.,

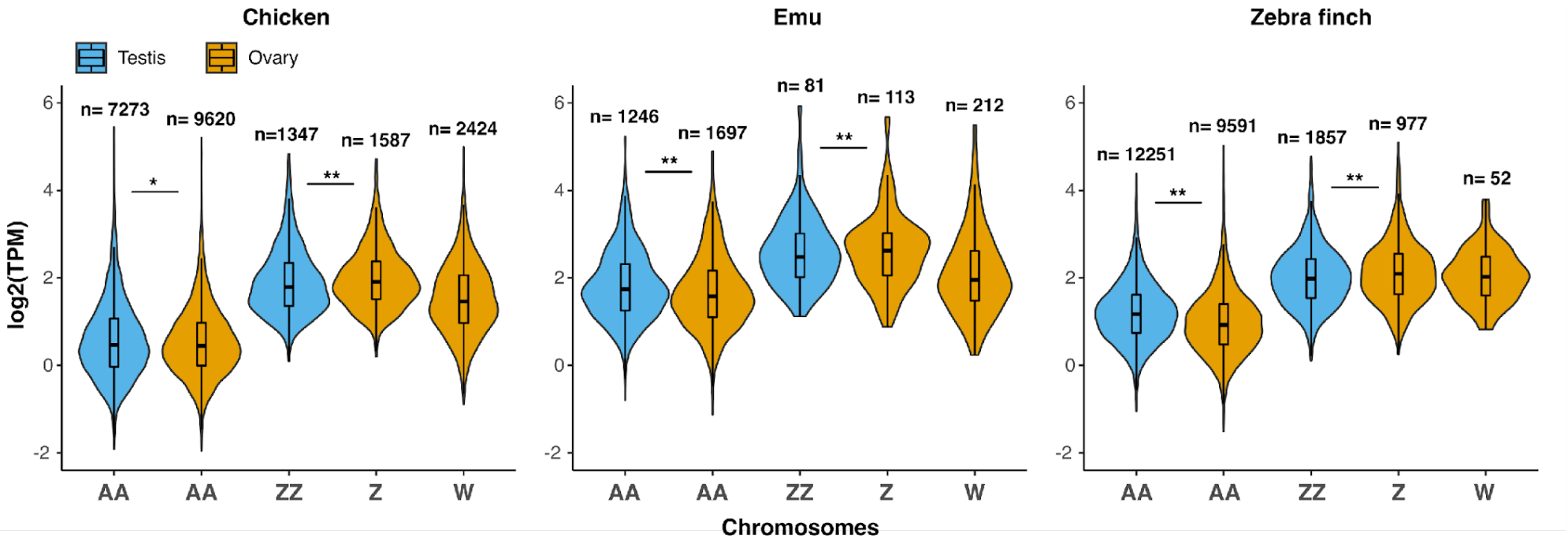
# W-linked TEs are more expressed

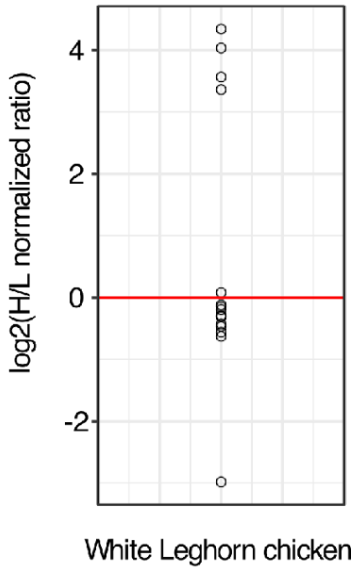



Peona et al. 2021b, Phil Trans B

# W-linked TEs are more expressed

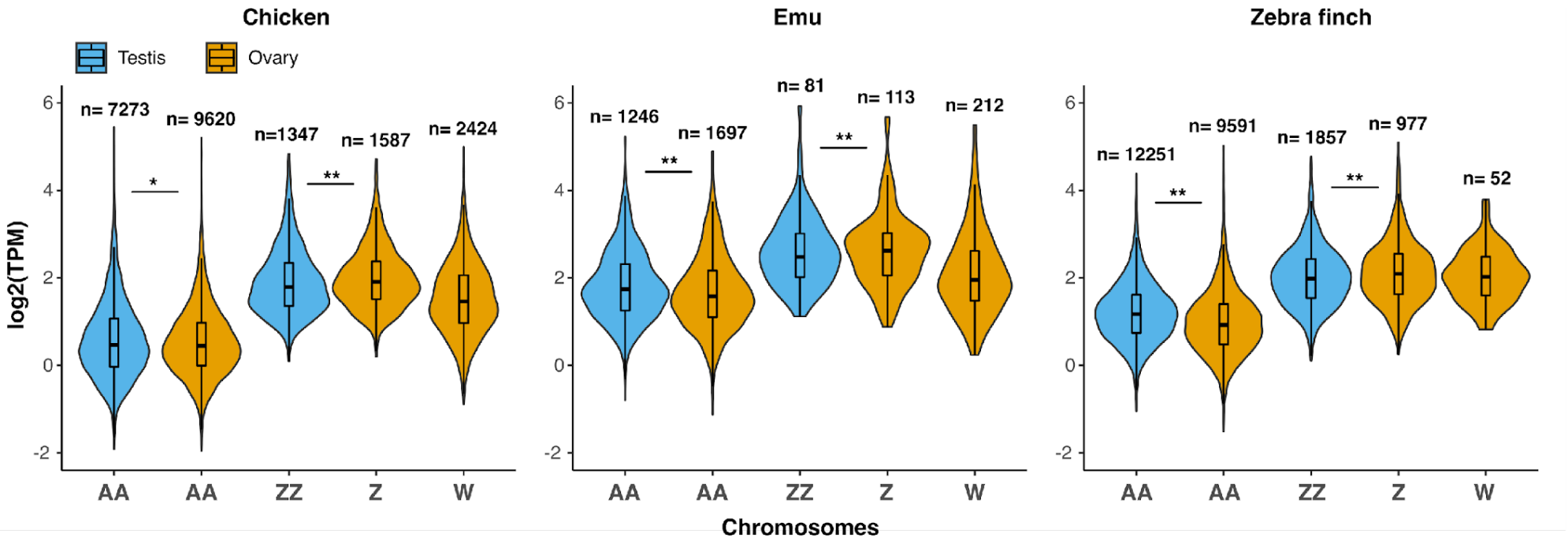


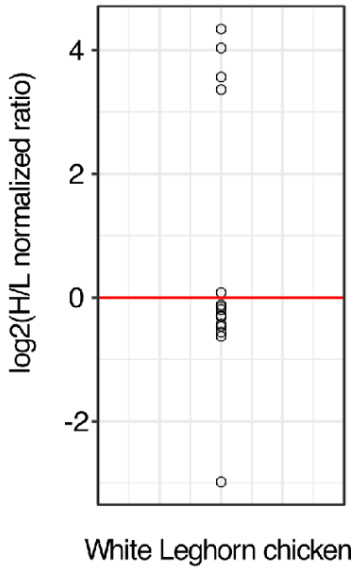




A reservoir (refugium) for active TEs that can have fitness effects on females

Peona et al. 2021b, Phil Trans B

# W-linked TEs are more expressed

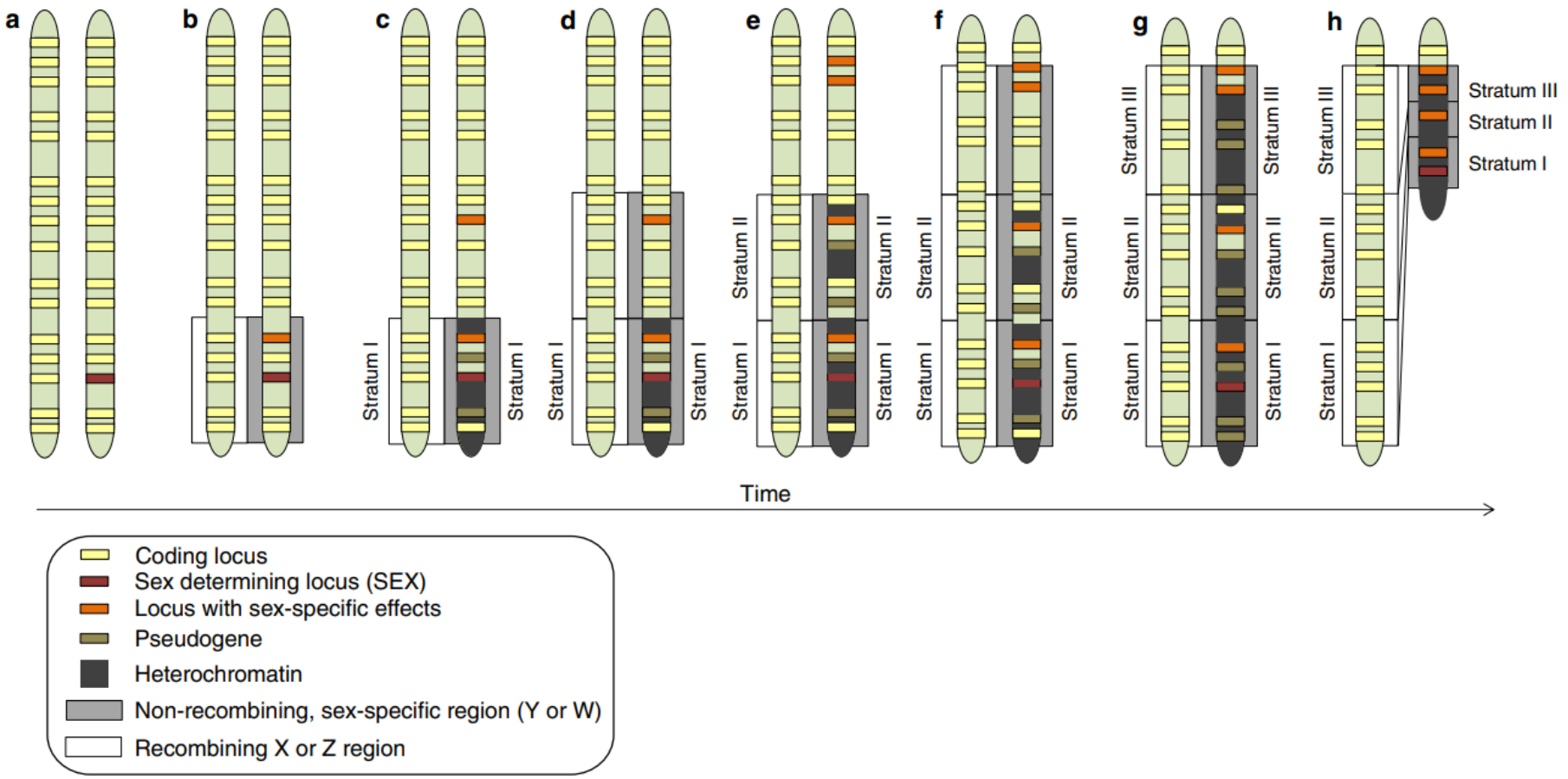



- A reservoir (refugium) for active TEs that can have fitness effects on females
- According to short reads no full-length elements were present on the W

Peona et al. 2021b, Phil Trans B

# Sex chromosome evolution



# Gene evolution

Innovators VS
constraint

# Gene structure

Introns

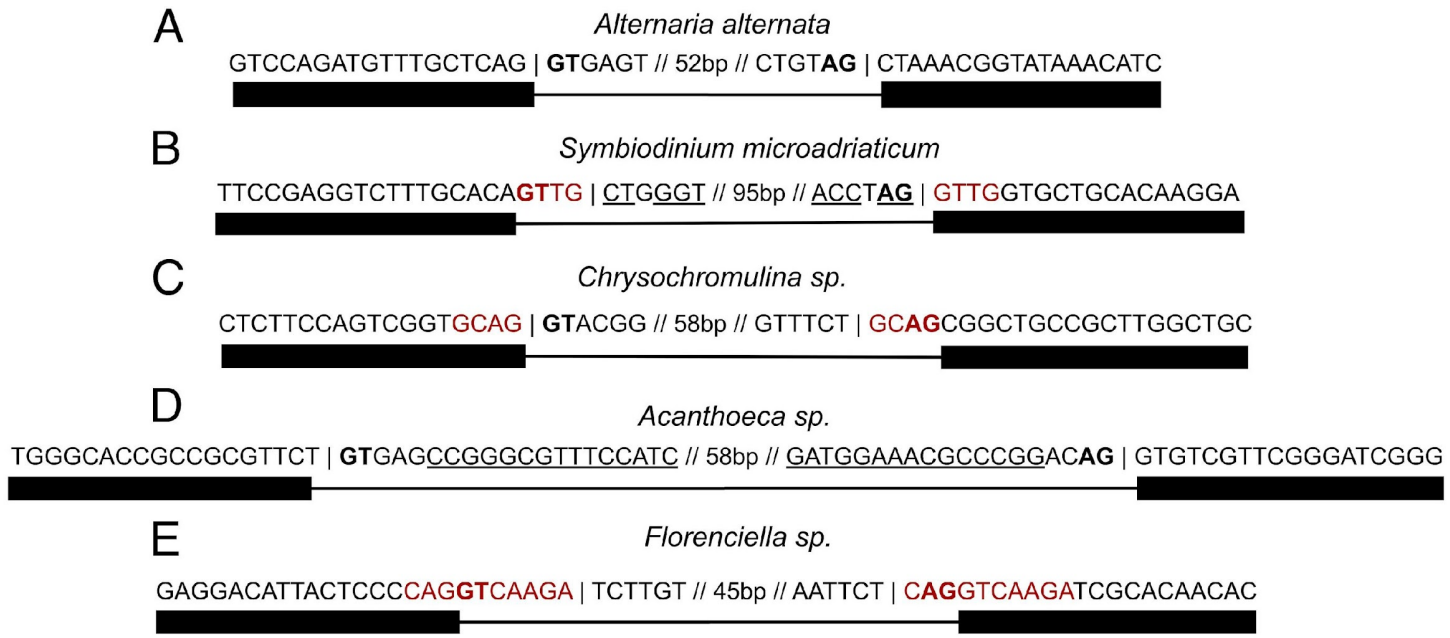
## Prp8, the pivotal protein of the spliceosomal catalytic center, evolved from a retroelement-encoded reverse transcriptase

MENSUR DLAKIĆ[1] and ARCADY MUSHEGIAN[2,3]

[1]Department of Microbiology, Montana State University, Bozeman, Montana 59717, USA
[2]Stowers Institute for Medical Research, Kansas City, Missouri 64110, USA
[3]Department of Microbiology, Kansas University Medical Center, Kansas City, Kansas 66160, USA

# Gene structure

## Introners

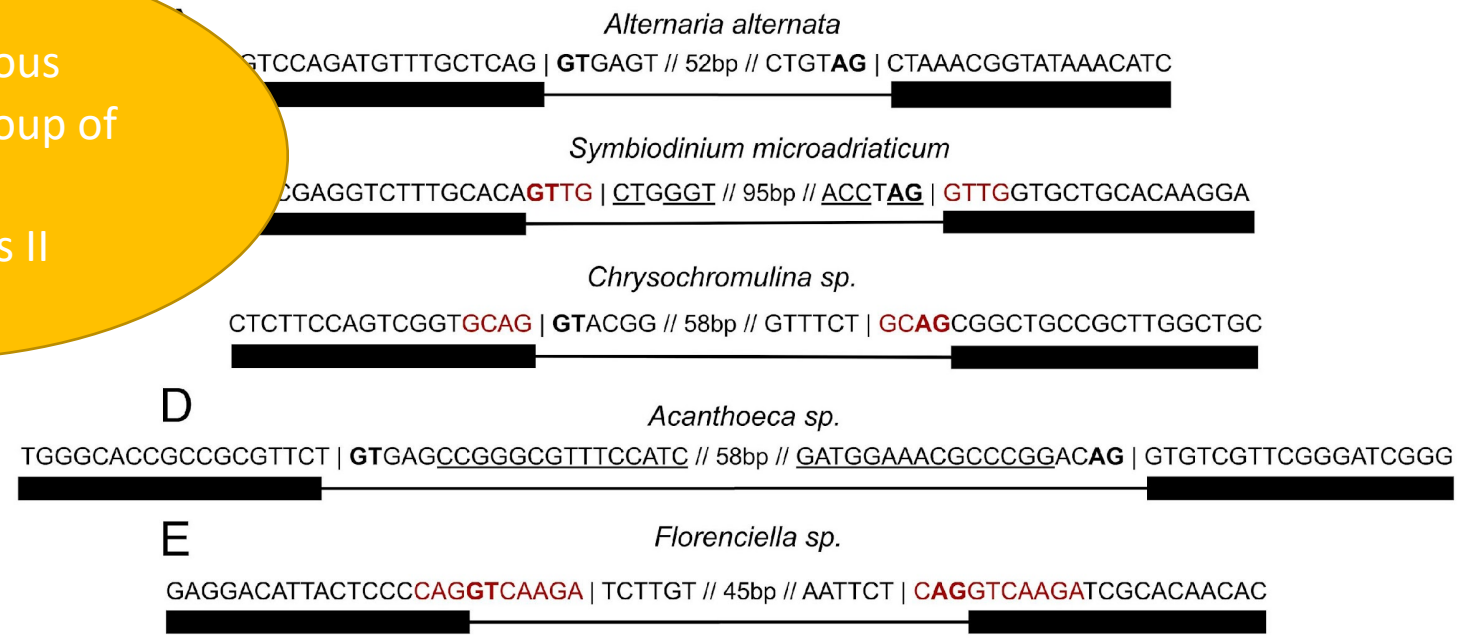



Gozashti et al 2022, PNAS

# Gene structure

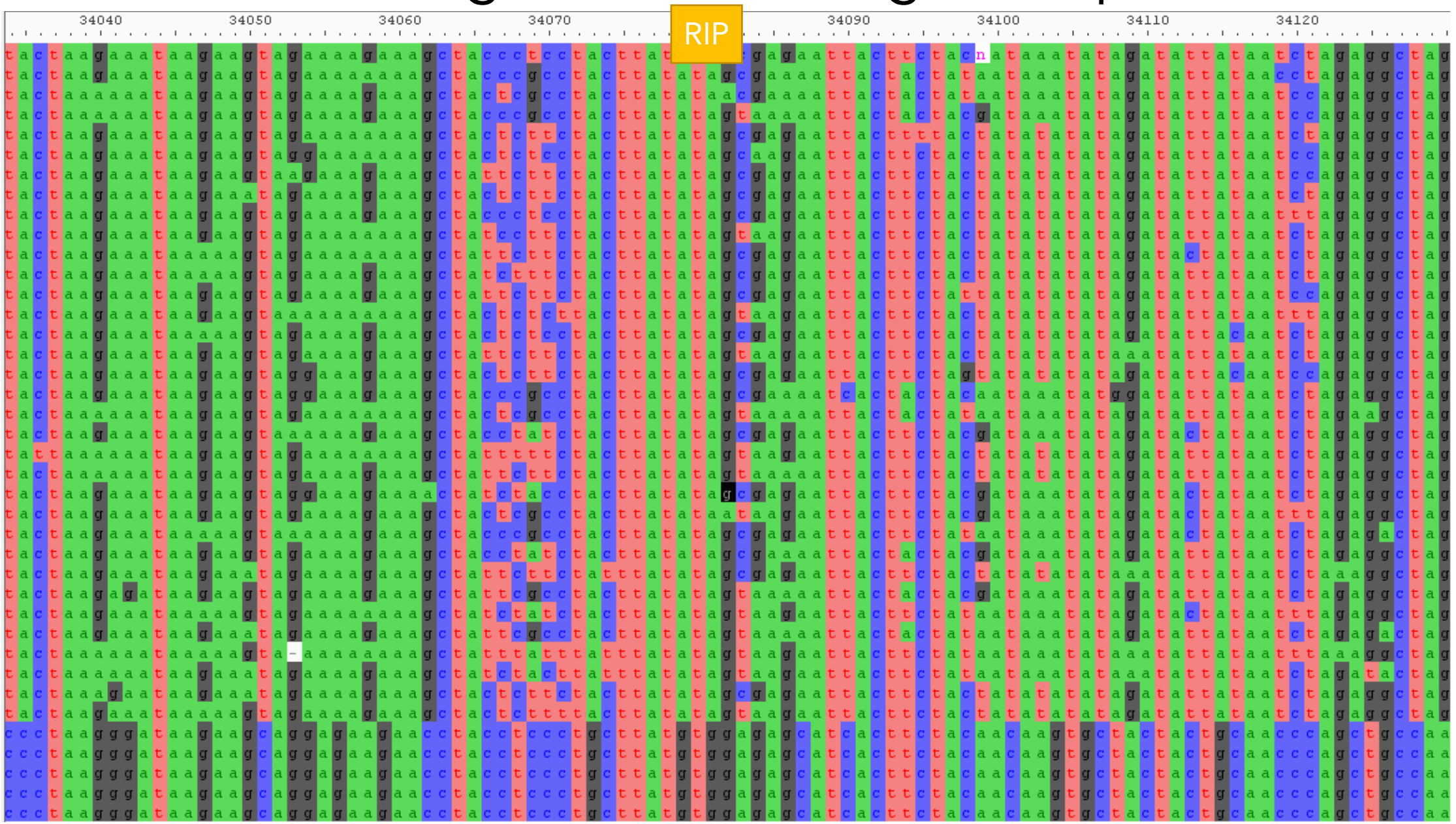
## Introners

Heterogeneous polyphyletic group of TEs.
Mostly Class II

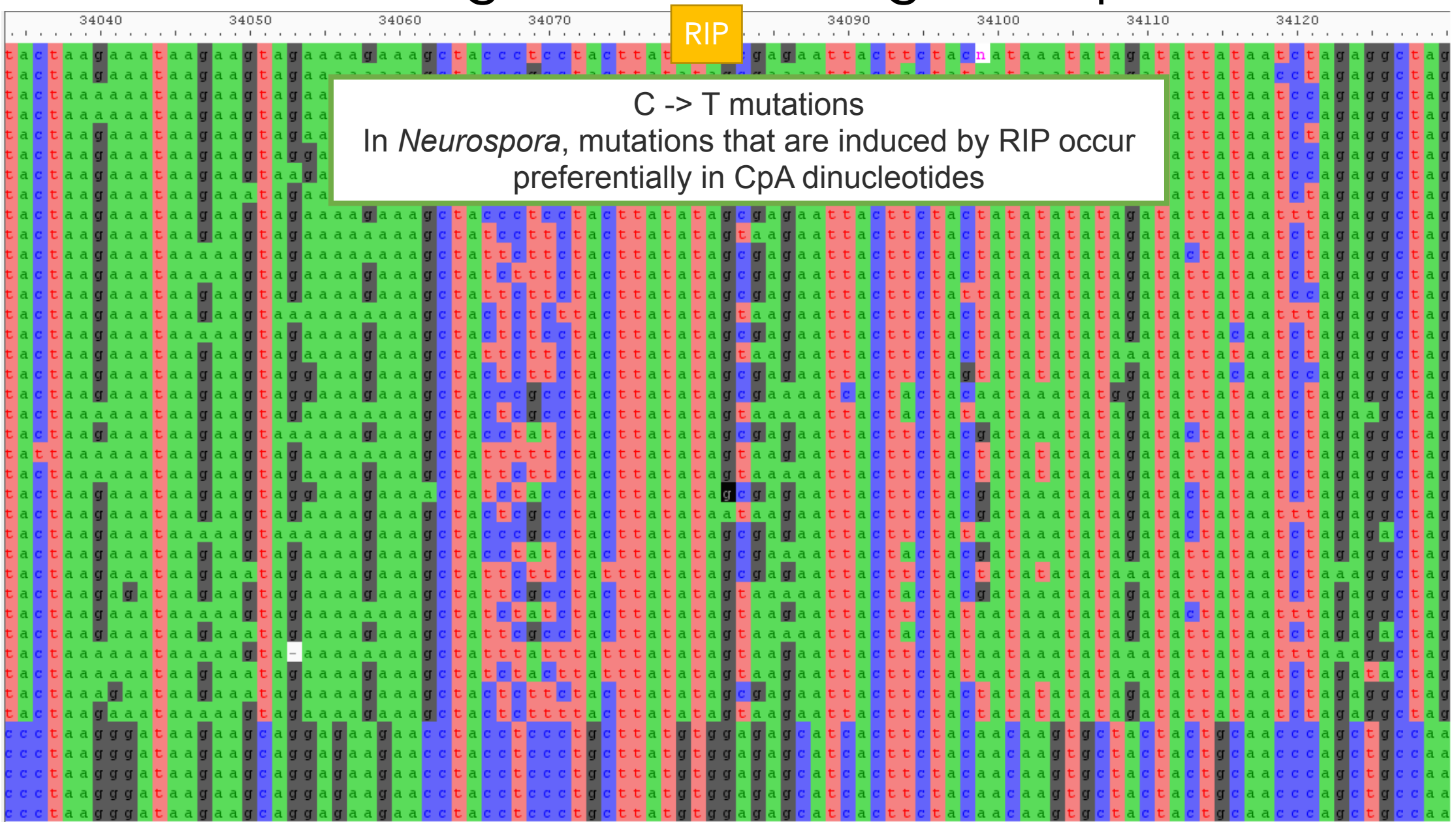



Gozashti et al 2022, PNAS

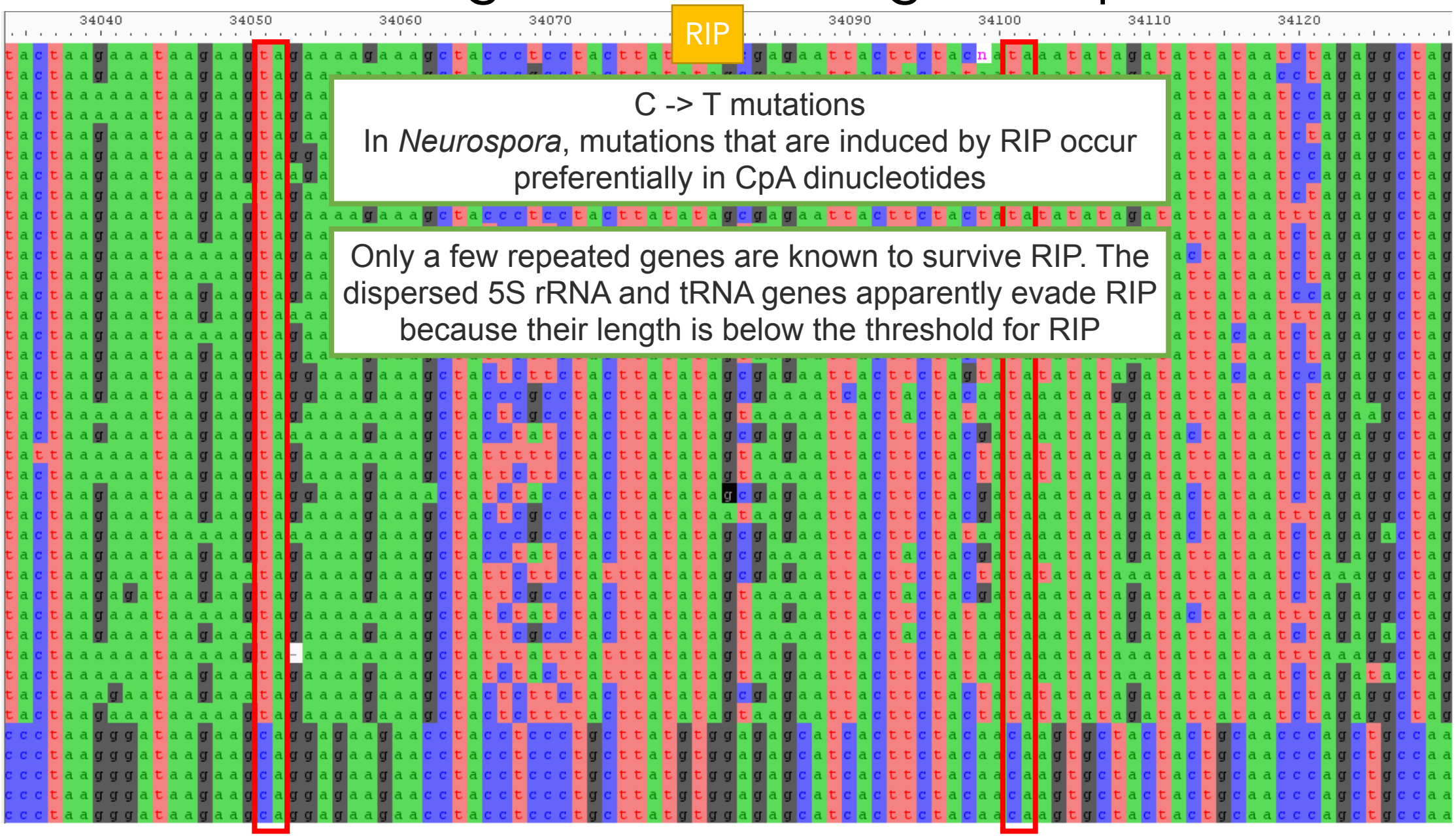
# Defences against TEs block gene duplication



# Defences against TEs block gene duplication



# Defences against TEs block gene duplication




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. The dispersed 5S rRNA and tRNA genes apparently evade RIP because their length is below the threshold for RIP