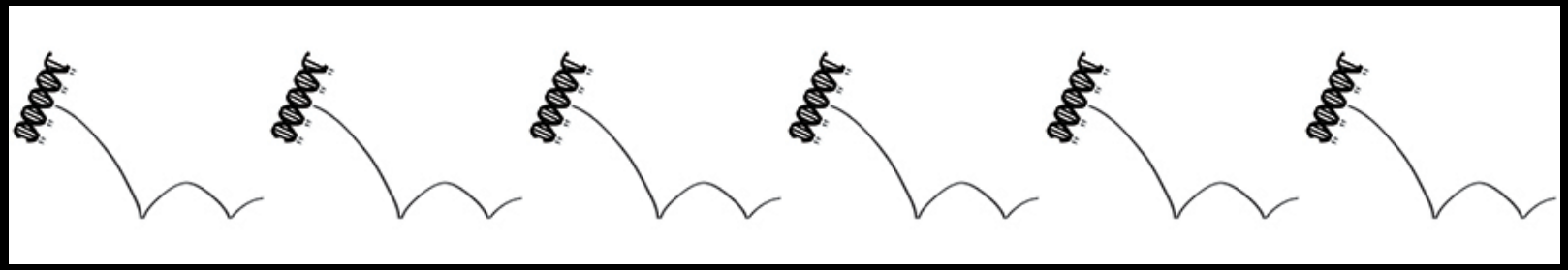
Biology of transposable elements

April 10, 2024



@alexander_suh

Alexander Suh
서상재 徐商在

University of Bonn
LIB Bonn/Hamburg

LIB Leibniz Institute for the Analysis of Biodiversity Change

Schedule for this afternoon

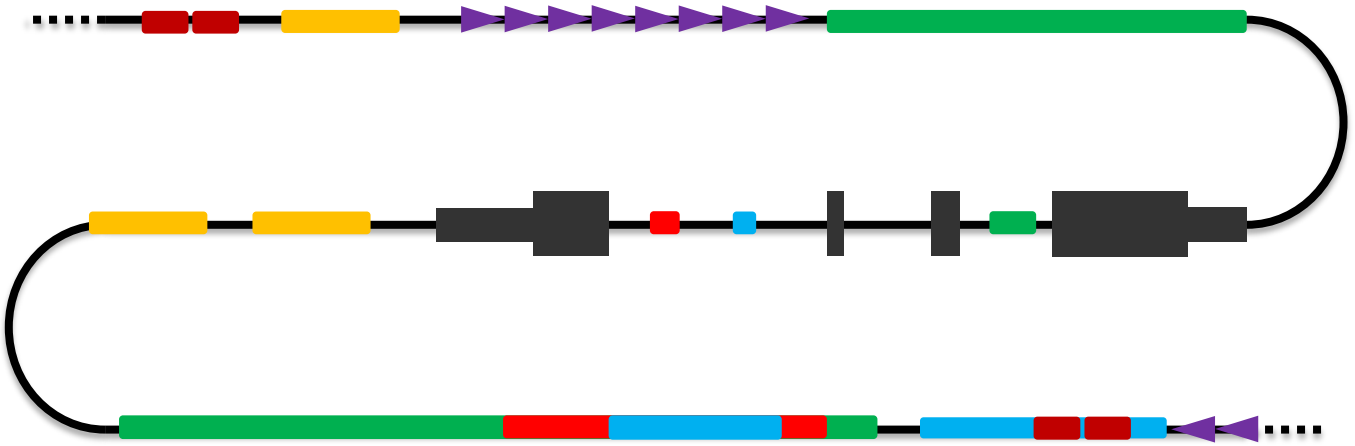
13:00-14:00 Biology of transposable elements (Alexander Suh)

14:00-14:15 Break

14:15-15:45 Visualization and analyses of repeats in R

15:45-16:00 Evaluation and final words

Genomes: microcosms of repeats



Interspersed repeats

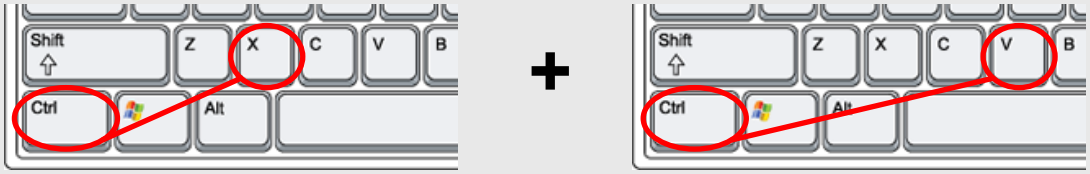
- Retrotransposons
- Retroviruses
- DNA transposons

Tandem repeats

- Some genes (e.g., rRNA genes)
- Satellites (e.g., in centromeres)
- Microsatellites (e.g., in telomeres)

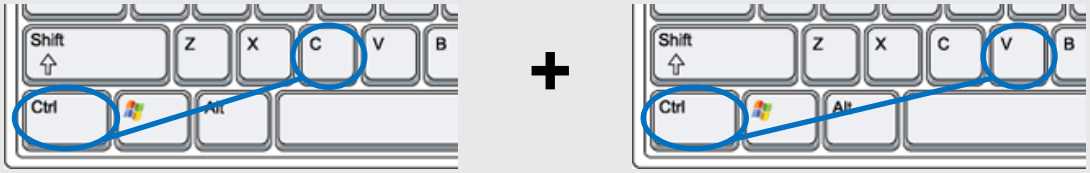
Introduction 1: Repeat diversity

DNA transposons



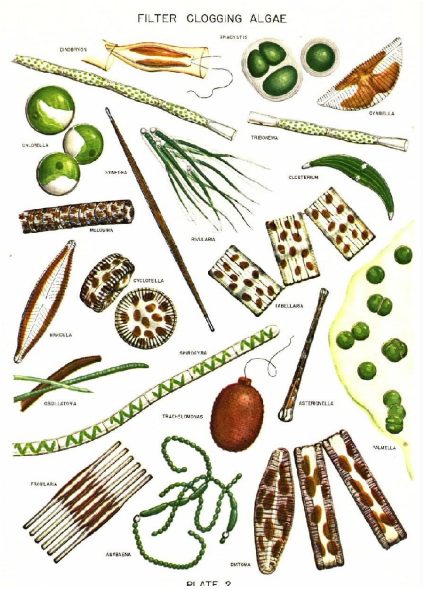
CUT PASTE

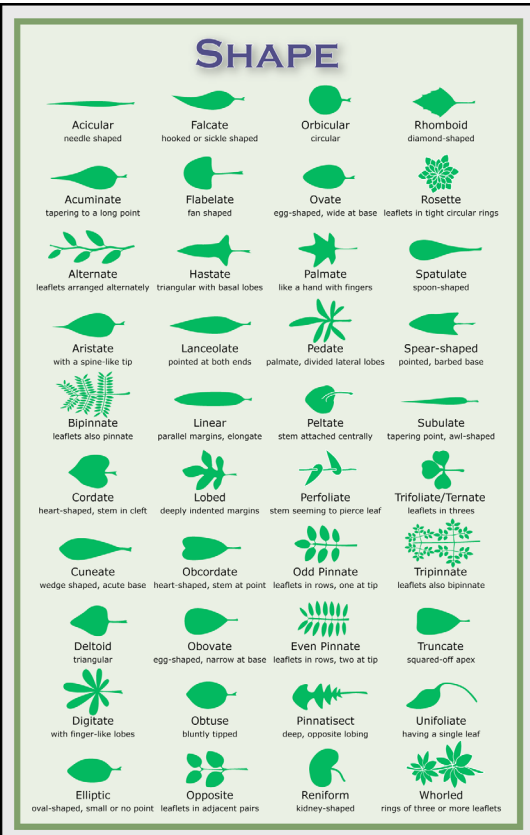
Retro(trans)posons

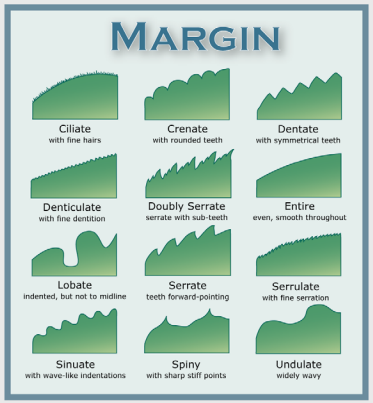


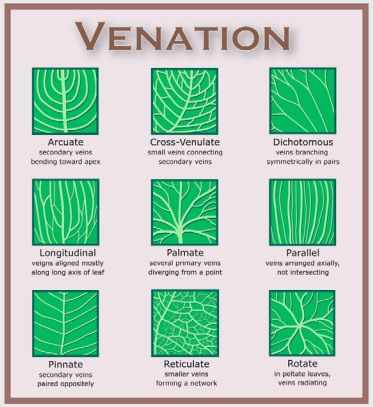
COPY PASTE

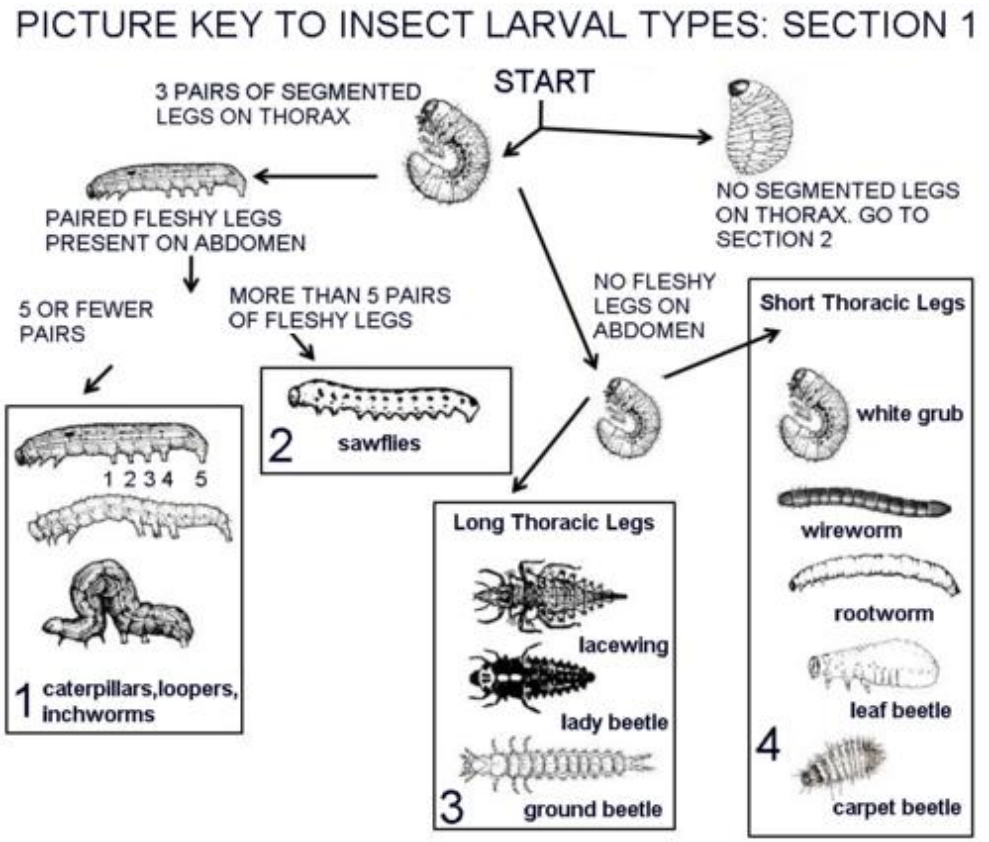

COTTON-TAIL RABBIT
MINK
WEASEL
MUSKRAT
DOG
DEER MOUSE
RAT
CAT
WOODCHUCK
Right hind
fore hind
CHIPMUNK
RACCOON
Right fore
RED SQUIRREL
SKUNK


FILTER CLOGGING ALGAE


SHAPE
Acicular
needle shaped
Falcate
hooked or sickle shaped
Orbicular
circular
Rhomboid
diamond-shaped
Acuminate
tapering to a long point
Flabelate
fan shaped
Ovate
egg-shaped, wide at base
Rosette
leaflets in tight circular rings
Alternate
leaflets arranged alternately
Hastate
triangular with basal lobes
Palmate
like a hand with fingers
Spatulate
spoon-shaped
Aristate
with a spine-like tip
Lanceolate
pointed at both ends
Pedate
palmate, divided lateral lobes
Spear-shaped
pointed, barbed base
Bipinnate
leaflets also pinnate
Linear
parallel margins, elongate
Peltate
stem attached centrally
Subulate
tapering point, awl-shaped
Cordate
heart-shaped, stem in cleft
Lobed
deeply indented margins
Perfoliate
stem seeming to pierce leaf
Trifoliate/Ternate
leaflets in threes
Cuneate
wedge shaped, acute base
Obcordate
heart-shaped, stem at point
Odd Pinnate
leaflets in rows, one at tip
Tripinnate
leaflets also bipinnate
Deltoid
triangular
Obovate
egg-shaped, narrow at base
Even Pinnate
leaflets in rows, two at tip
Truncate
squared-off apex
Digitate
with finger-like lobes
Obtuse
bluntly tipped
Pinnatisect
deep, opposite lobing
Unifoliate
having a single leaf
Elliptic
oval-shaped, small or no point
Opposite
leaflets in adjacent pairs
Reniform
kidney-shaped
Whorled
rings of three or more leaflets


MARGIN
Ciliate
with fine hairs
Crenate
with rounded teeth
Dentate
with symmetrical teeth
Denticulate
with fine dentition
Doubly Serrate
serrate with sub-teeth
Entire
even, smooth throughout
Lobate
indented, but not to midline
Serrate
teeth forward-pointing
Serrulate
with fine serration
Sinuate
with wave-like indentations
Spiny
with sharp stiff points
Undulate
widely wavy


VENATION
Arcuate
secondary veins bending toward apex
Cross-Venulate
small veins connecting secondary veins
Dichotomous
veins branching symmetrically in pairs
Longitudinal
veigns aligned mostly along long axis of leaf
Palmate
several primary veins diverging from a point
Parallel
veins arranged axially, not intersecting
Pinnate
secondary veins paired oppositely
Reticulate
smaller veins forming a network
Rotate
in peltate leaves, veins radiating


PICTURE KEY TO INSECT LARVAL TYPES: SECTION 1
START
3 PAIRS OF SEGMENTED LEGS ON THORAX
NO SEGMENTED LEGS ON THORAX. GO TO SECTION 2
PAIRED FLESHY LEGS PRESENT ON ABDOMEN
5 OR FEWER PAIRS
MORE THAN 5 PAIRS OF FLESHY LEGS
NO FLESHY LEGS ON ABDOMEN
1 2 3 4 5
1 caterpillars,loopers, inchworms
2 sawflies
Long Thoracic Legs
lacewing
lady beetle
3 ground beetle
Short Thoracic Legs
white grub
wireworm
rootworm
leaf beetle
4 carpet beetle

Transposable elements are selfish

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)

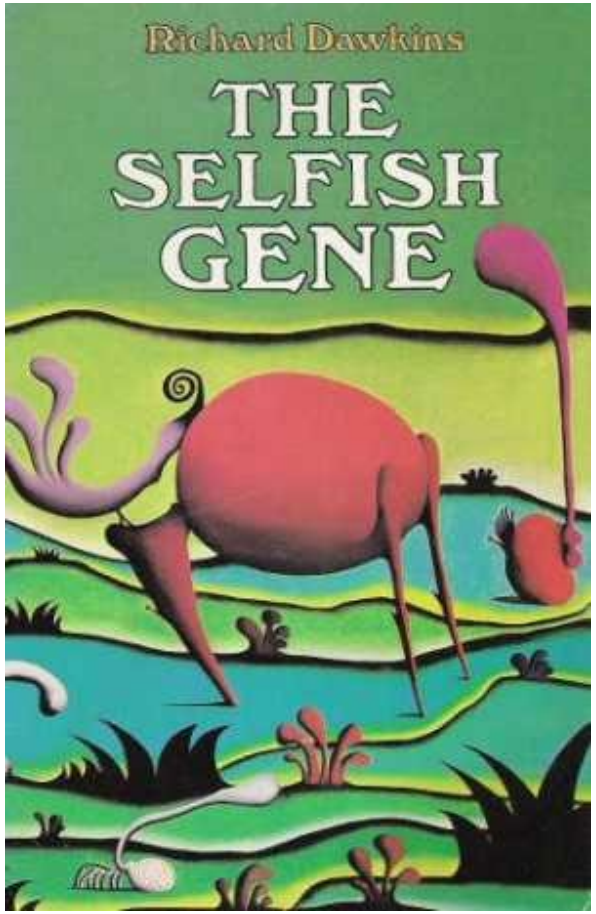
Why do genes behave selfishly?

Because.
They.
Can.

The Selfish Gene

"*We are **survival machines** – robot vehicles blindly programmed to preserve the selfish molecules known as genes.*"

"*What, after all, is so special about genes? The answer is that they are **replicators**.*"



Barbara McClintock
(Nobel Prize in Physiology or Medicine 1983)

Maize: 85% TEs!

Eukaryotic TEs are highly diverse…

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
DIRS	*DIRS*	GAG AP RT RH YR	0	RYD	P, M, F, O
	Ngaro	GAG AP RT RH YR	0	RYN	M, F
	VIPER	GAG AP RT RH YR	0	RYV	O
PLE	*Penelope*	RT EN	Variable	RPP	P, M, F, O
LINE	*R2*	RT EN	Variable	RIR	M
	RTE	APE RT	Variable	RIT	M
	Jockey	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
SINE	*tRNA*		Variable	RST	P, M, F
	7SL		Variable	RSL	P, M, F
	5S		Variable	RSS	M, O
Class II (DNA transposons) - Subclass 1					
TIR	*Tc1–Mariner*	Tase*	TA	DTT	P, M, F, O
	hAT	Tase*	8	DTA	P, M, F, O
	Mutator	Tase*	9–11	DTM	P, M, F, O
	Merlin	Tase*	8–9	DTE	M, O
	Transib	Tase*	5	DTR	M, F
	P	Tase	8	DTP	P, M
	PiggyBac	Tase	TTAA	DTB	M, O
	PIF–Harbinger	Tase* ORF2	3	DTH	P, M, F, O
	CACTA	Tase ORF2	2–3	DTC	P, M, F
Crypton	*Crypton*	YR	0	DYC	F
Class II (DNA transposons) - Subclass 2					
Helitron	*Helitron*	RPA Y2 HEL	0	DHH	P, M, F
Maverick	*Maverick*	C-INT ATP CYP POL B	6	DMM	M, F, O

Structural features

Long terminal repeats; Terminal inverted repeats; Coding region; Non-coding region; Diagnostic feature in non-coding region; Region that can contain one or more additional ORFs

Protein coding domains

AP, Aspartic proteinase; APE, Apurinic endonuclease; ATP, Packaging ATPase; C-INT, C-integrase; CYP, Cysteine protease; EN, Endonuclease; ENV, Envelope protein; GAG, Capsid protein; HEL, Helicase; INT, Integrase; ORF, Open reading frame of unknown function; POL B, DNA polymerase B; RH, RNase H; RPA, Replication protein A (found only in plants); RT, Reverse transcriptase; Tase, Transposase (* with DDE motif); YR, Tyrosine recombinase; Y2, YR with YY motif

Species groups

P, Plants; M, Metazoans; F, Fungi; O, Others

Nature Reviews | Genetics

Wicker et al. 2008, *Nat. Rev. Genet.*

… and so are endogenous viruses!

Group/type	Family or genus	Taxa	Number per haploid genome	Refs
Group I/dsDNA	Baculovirus	Insects	Unknown (hybridization data, no sequencing)	124
Group I/dsDNA	Herpesviridae	Humans	1	125,126
Group I/dsDNA	Nudivirus	Parasitic wasps	Several	127
Group I/dsDNA	Phycodnaviridae	Brown algae	1	128,129
Group II/ssDNA	Circoviridae	Mammals	1 to 2	4,20,130
Group II/ssDNA	Geminiviridae	Tomentosae (tobacco and three other species)	5 to 120	130–132
Group II/ssDNA	Parvoviridae	Mammals; shrimp	1 to 3	4,20,89, 133,134
Group III/dsRNA	Partitiviridae	Plants; arthropods; Protozoa	1 to 4	135
Group III/dsRNA	Reovirus	*Aedes* spp. mosquitoes	1	4
Group III/dsRNA	Totiviridae	Fungi; plants; ticks	1 to 6	16,135,136
Group IV/+ssRNA	Dicistroviridae	Honeybees	1	137
Group IV/+ssRNA	Flaviviridae	Medaka fish; mosquitoes	1 to 4	4,21,138, 139
Group IV/+ssRNA	Potyviridae	Grapes	Several	140
Group V/–ssRNA	Bornaviridae	Vertebrates	1 to 17	4,21,17
Group V/–ssRNA	Bunyaviridae	Ticks	14	4
Group V/–ssRNA	Filoviridae	Mammals	1 to 13	4,21,141
Group V/–ssRNA	Nyavirus	Zebrafish	6	21
Group V/–ssRNA	Orthomyxoviridae	Ticks	1	4
Group V/–ssRNA	Rhabdoviridae	Insects (ticks and mosquitoes)	1 to 28	4
Group VI/ssRNA-RT	Retroviridae	Vertebrates	Several hundreds to several hundreds of thousands	36
Group VII/dsDNA-RT	Hepadnavirus	Passerine birds	15	4,19
Group VII/dsDNA-RT	Pararetrovirus	Plants	A dozen to a thousand	8,11

+, positive sense; –, negative sense; RT, reverse transcriptase.

Feschotte & Gilbert 2012, *Nat. Rev. Genet.*

Class I: LINE retrotransposons

Classification		Structure	TSD	Code	Occurrence
Order	Superfamily				
Class I (retrotransposons)					
PLE	*Penelope*	RT EN	Variable	RPP	P, M, F, O
LINE	*R2*	RT EN	Variable	RIR	M
	RTE	APE RT	Variable	RIT	M
	Jockey	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

Dear RNA polymerase II,

if you read this,
transcribe me
into RNA

Dear ribosome,

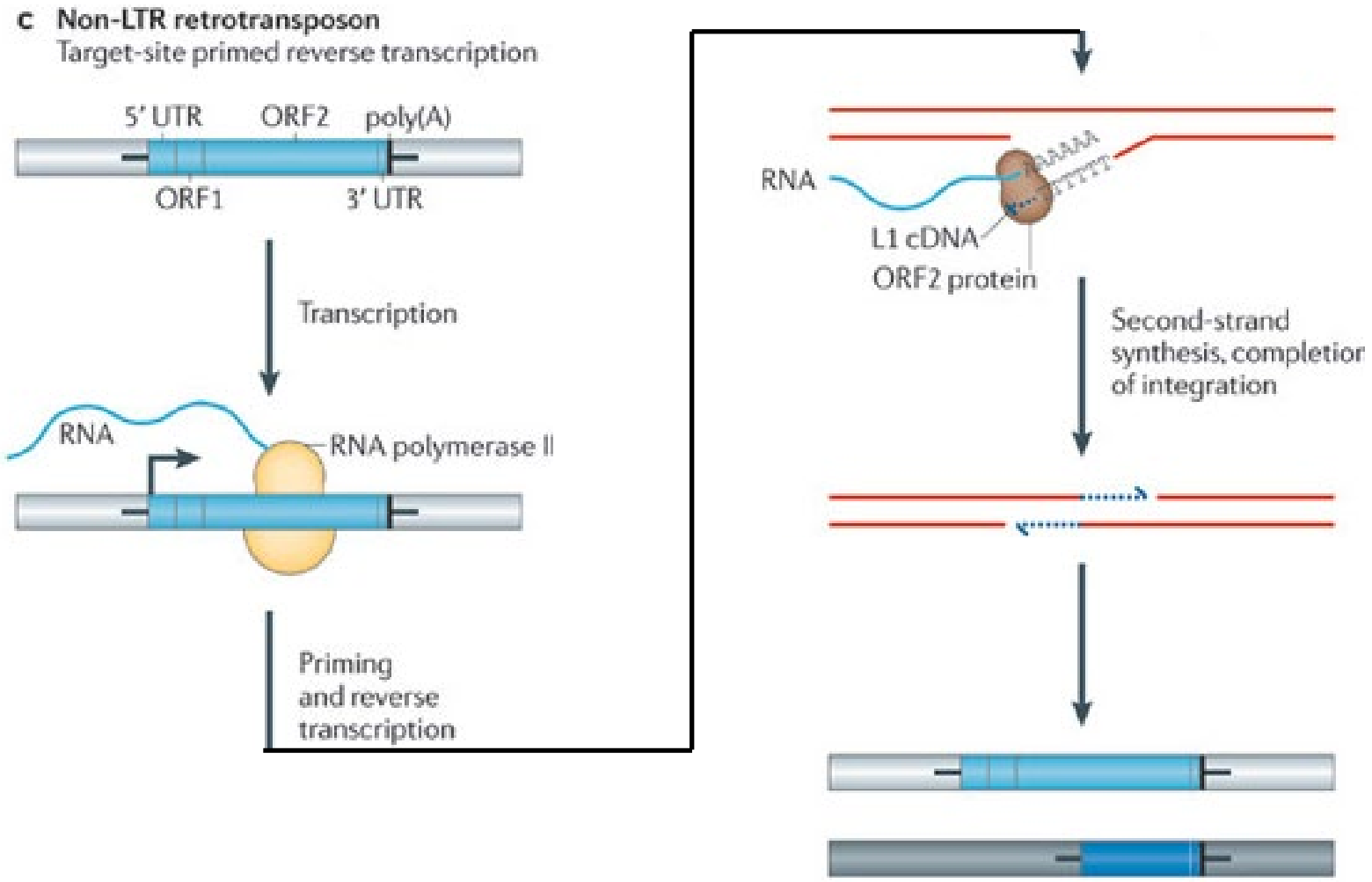
if you read this,
translate me into a
reverse transcriptase

Dear reverse transcriptase,

if you read this,
retropose me somewhere
in the genome

Wicker et al. 2008, *Nat. Rev. Genet.*

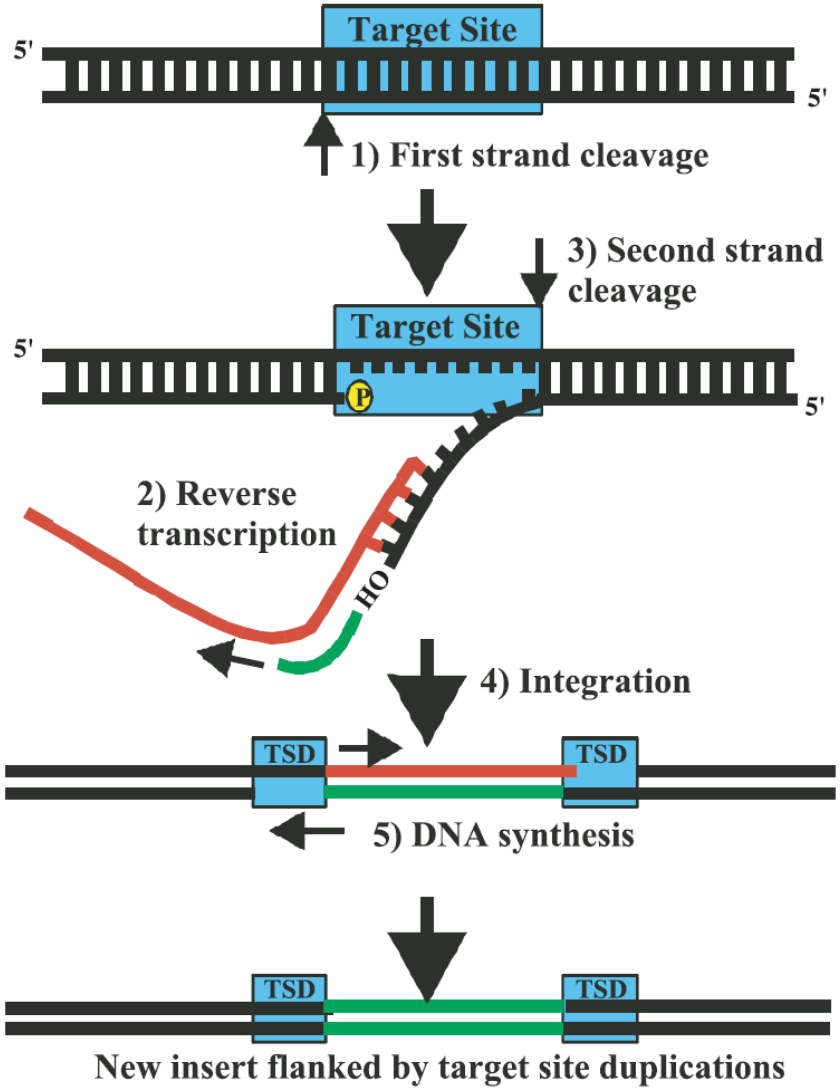
Target-primed reverse transcription (TPRT)



TPRT frequently undergoes premature termination (5‘ truncation)

Levin & Moran 2011, *Nat. Rev. Genet.*

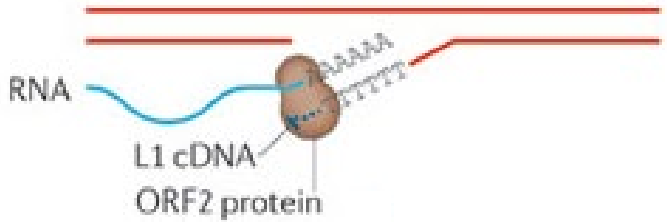
Target site duplication (TSD)

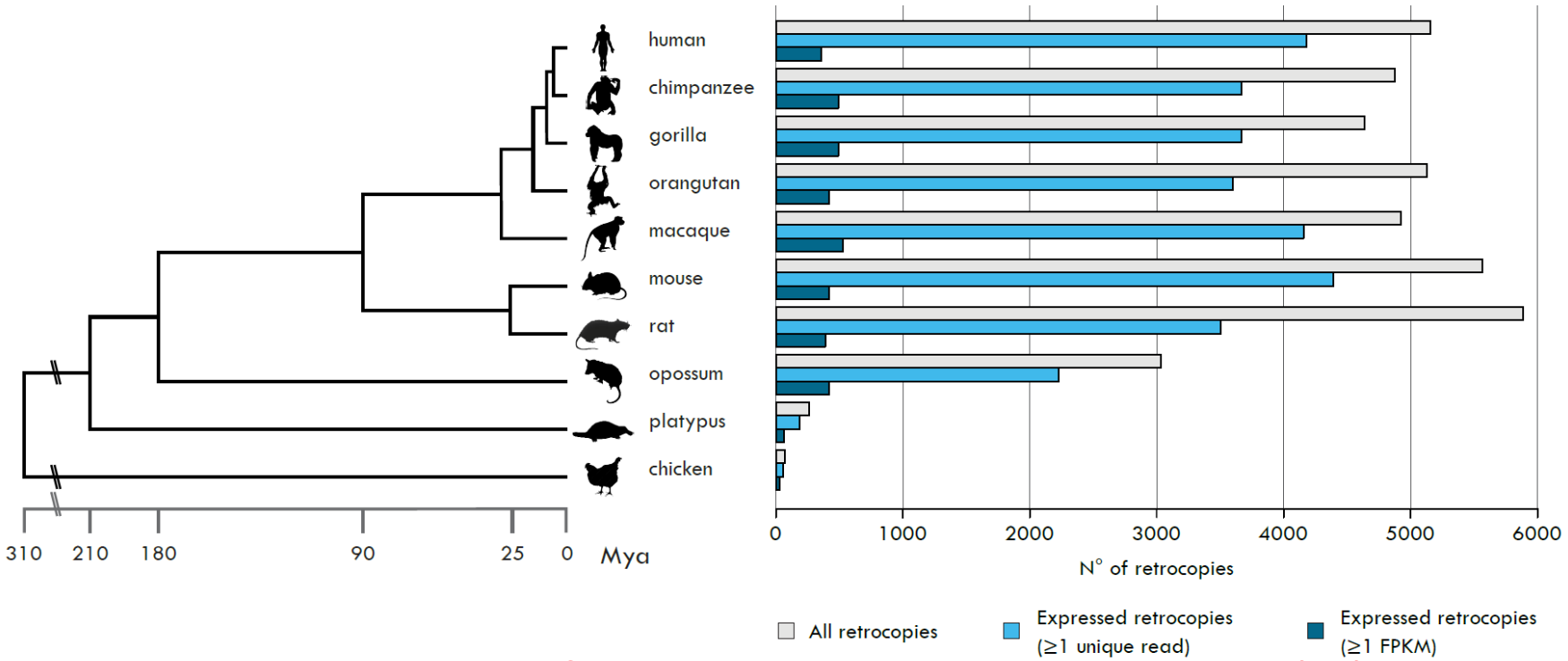


TSDs are a hallmark of nearly all (retro)transposition mechanisms!

Kazazian 2004, *Science*

L1 likes to make retro(pseudo)genes





Intronless retrogenes occur if LINE RT recognizes poly-A tails only (L1)

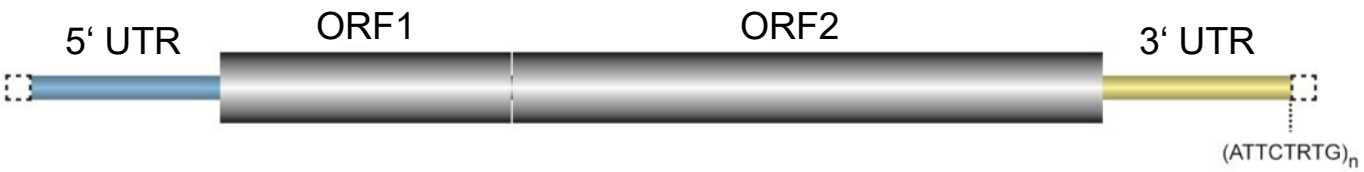
Carelli et al. 2016, *Genome Res.*

Class I: SINE retrotransposons

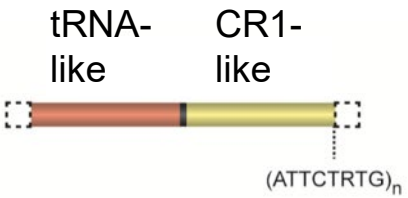
Classification		Structure	TSD	Code	Occurrence
Order	Superfamily				
Class I (retrotransposons)					
SINE	*tRNA*		Variable	RST	P, M, F
	7SL		Variable	RSL	P, M, F
	5S		Variable	RSS	M, O

SINEs are parasites of LINEs! *Trans*-mobilization via LINE enzymes.

CR1



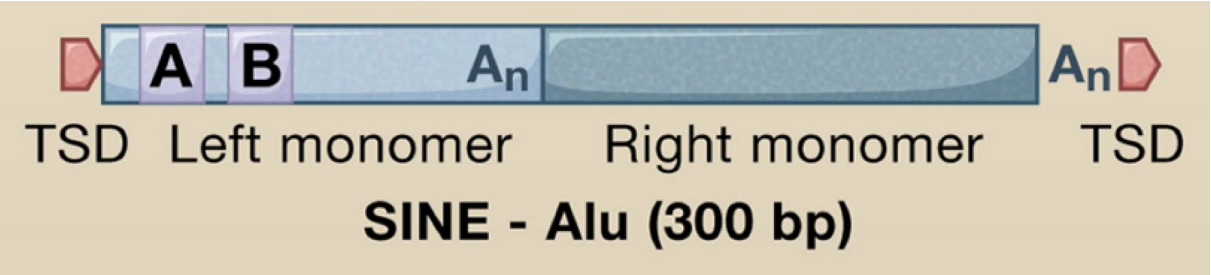
TguSINE1



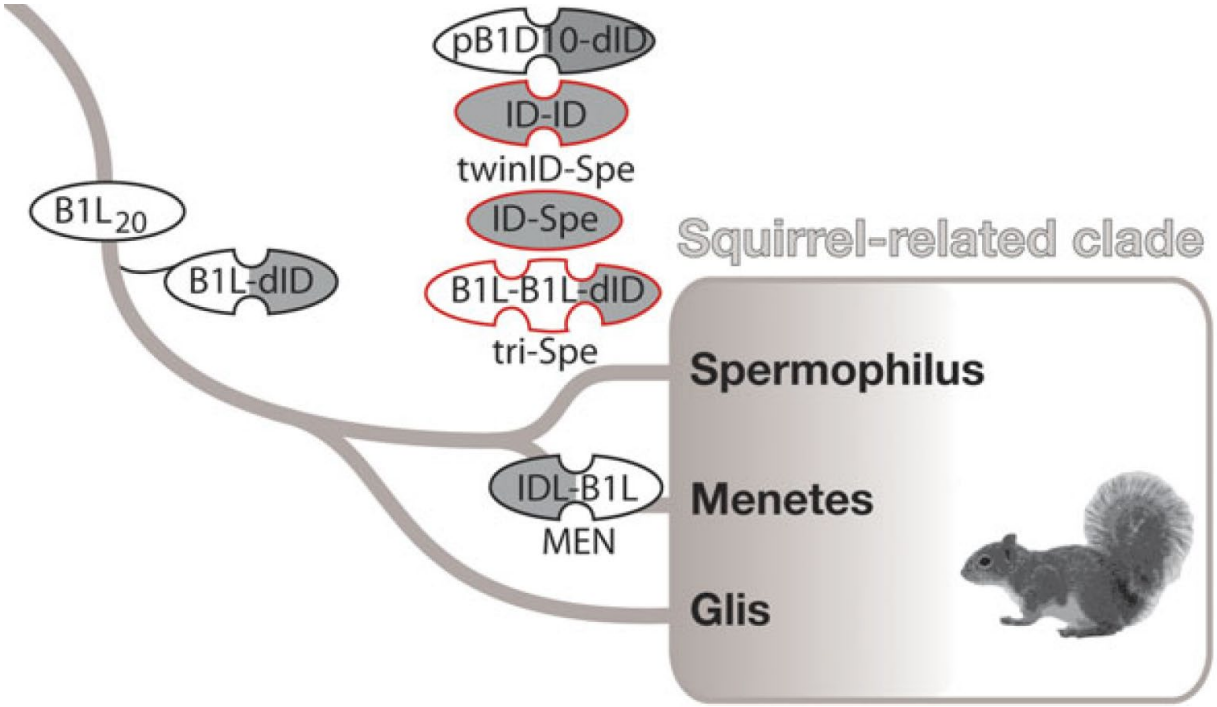
Note: In theory, any small RNA gene (pol III) can become a SINE!

Wicker et al. 2008, *Nat. Rev. Genet.*

SINEs can be monomers, dimers, trimers



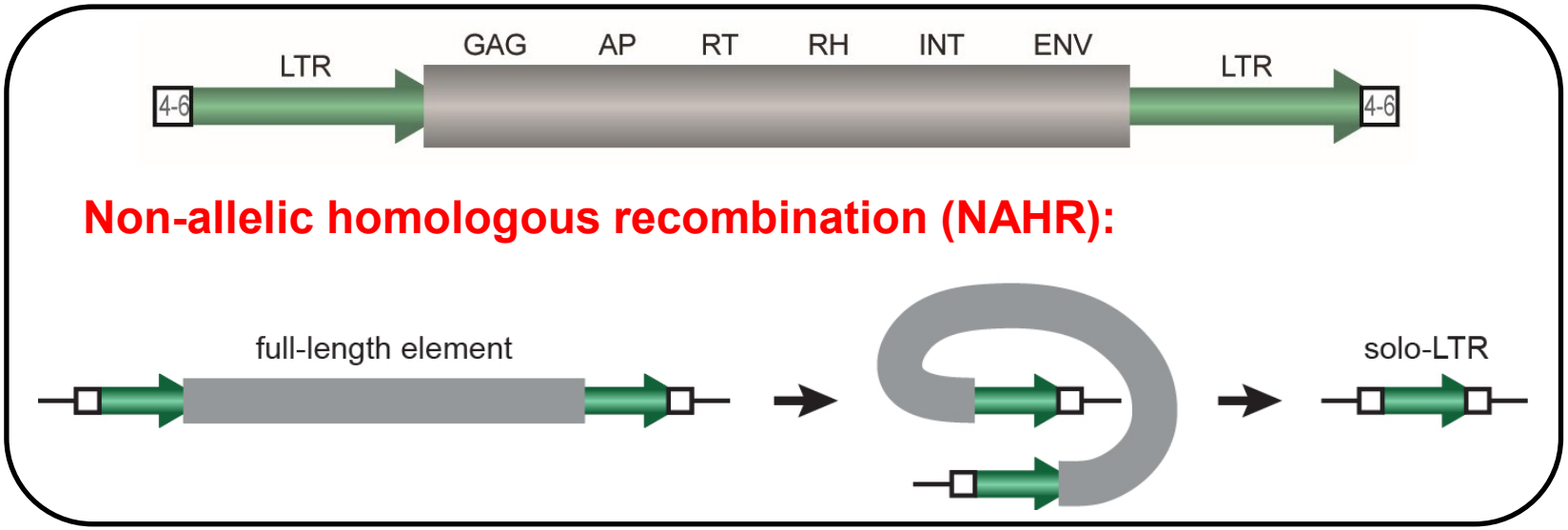
Goodier 2016, *Mob. DNA*



Churakov et al. 2010, *Mol. Biol. Evol.*

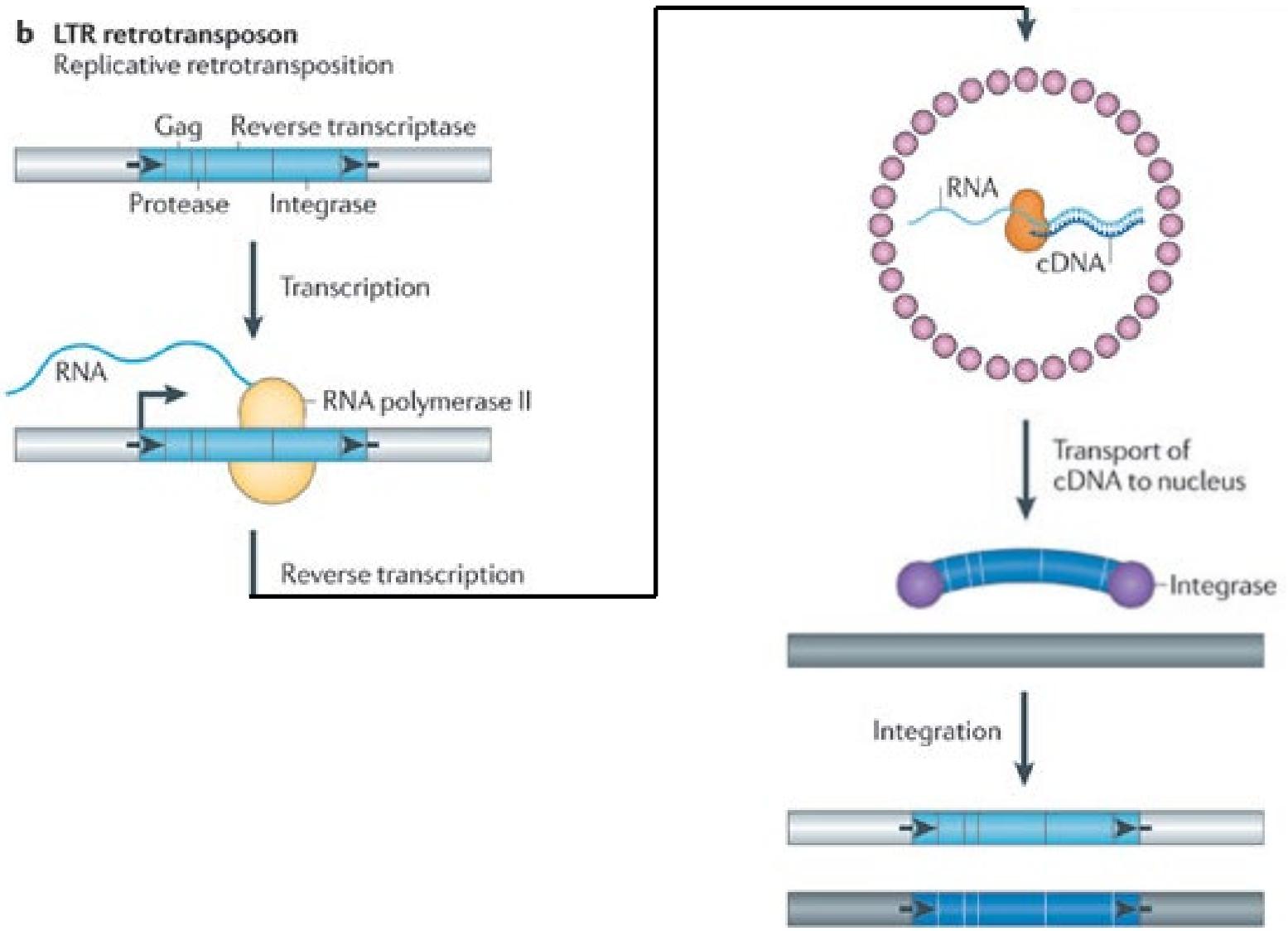
Class I: LTR retrotransposons

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
DIRS	*DIRS*	GAG AP RT RH YR	0	RYD	P, M, F, O
	Ngaro	GAG AP RT RH YR	0	RYN	M, F
	VIPER	GAG AP RT RH YR	0	RYV	O



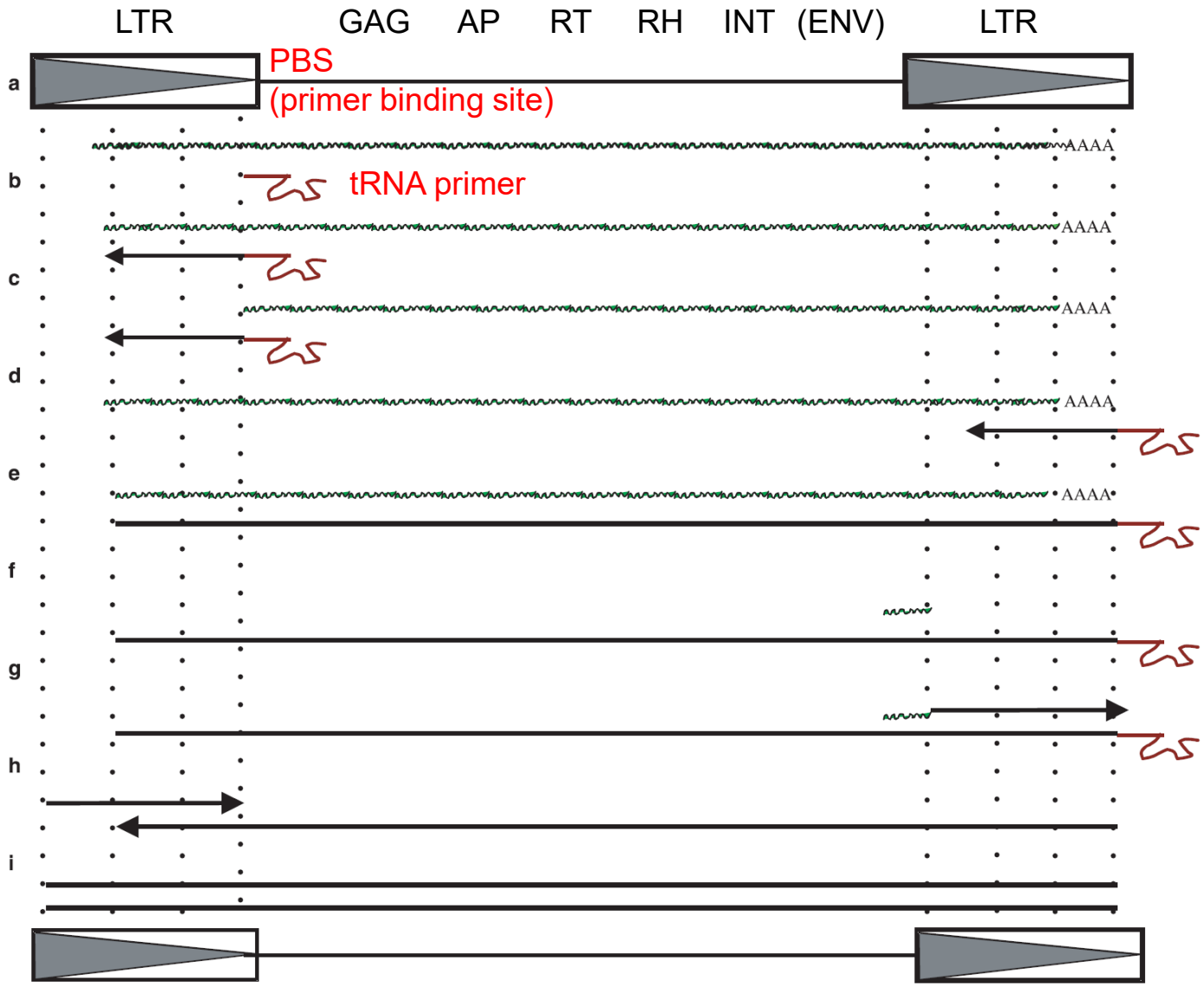
Wicker et al. 2008, *Nat. Rev. Genet.*

Replicative retrotransposition



Levin & Moran 2011, *Nat. Rev. Genet.*

Why LTR retrotransposons have LTRs



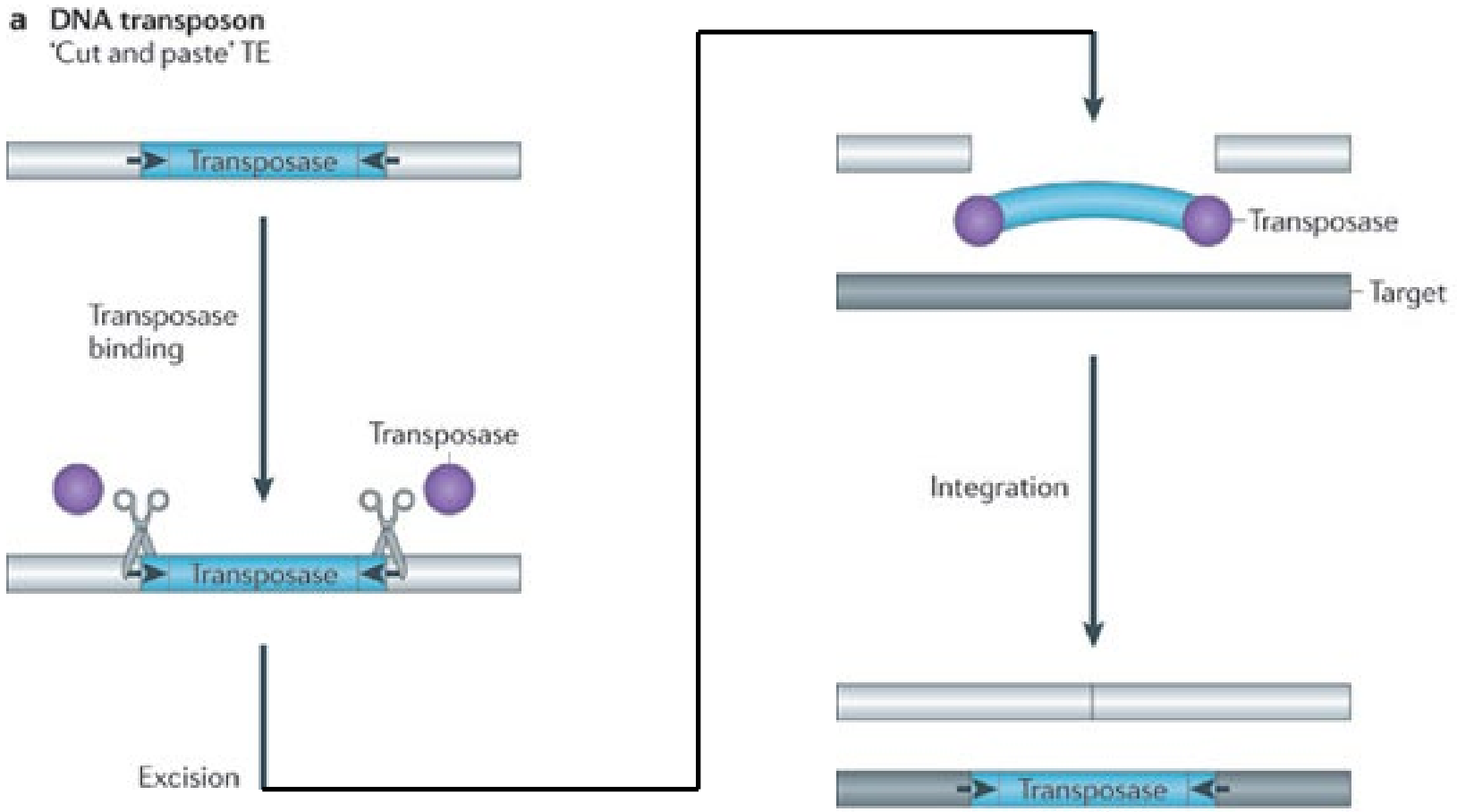
Kazazian 2004, *Science*

Class II: DNA transposons

Classification		Structure	TSD	Code	Occurrence
Order	Superfamily				
Class II (DNA transposons) - Subclass 1					
TIR	*Tc1–Mariner*	Tase*	TA	DTT	P, M, F, O
	hAT	Tase*	8	DTA	P, M, F, O
	Mutator	Tase*	9–11	DTM	P, M, F, O
	Merlin	Tase*	8–9	DTE	M, O
	Transib	Tase*	5	DTR	M, F
	P	Tase	8	DTP	P, M
	PiggyBac	Tase	TTAA	DTB	M, O
	PIF–Harbinger	Tase* ORF2	3	DTH	P, M, F, O
	CACTA	Tase ORF2	2–3	DTC	P, M, F
Crypton	*Crypton*	YR	0	DYC	F

Wicker et al. 2008, *Nat. Rev. Genet.*

Cut-and-paste transposition (TIR)

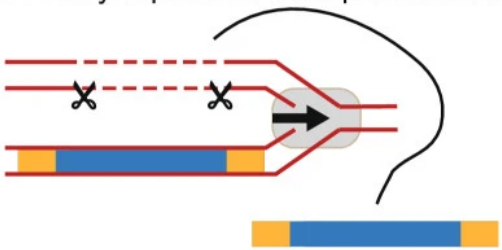


Levin & Moran 2011, *Nat. Rev. Genet.*

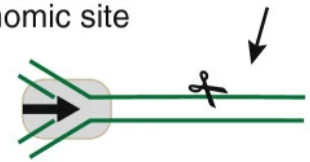
How to increase in copy number?

I. DNA replication fork passes transposon

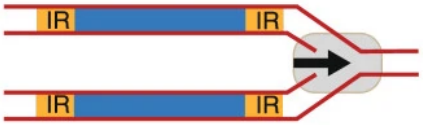
II. Newly replicated transposon is cut out...



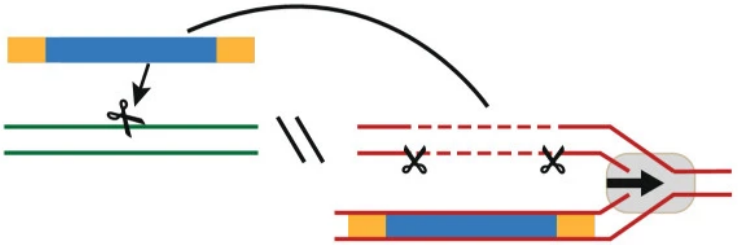
III. ...and inserted into a not-yet replicated genomic site



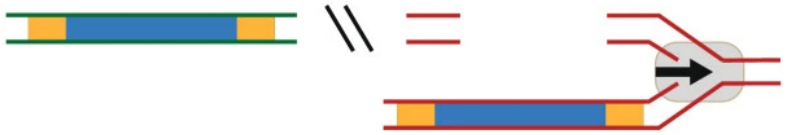
IIII. DNA replication fork passes insertion site



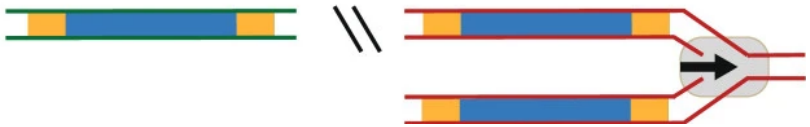
I. Newly replicated transposon is cut out...



II. ...and transposed into a new locus



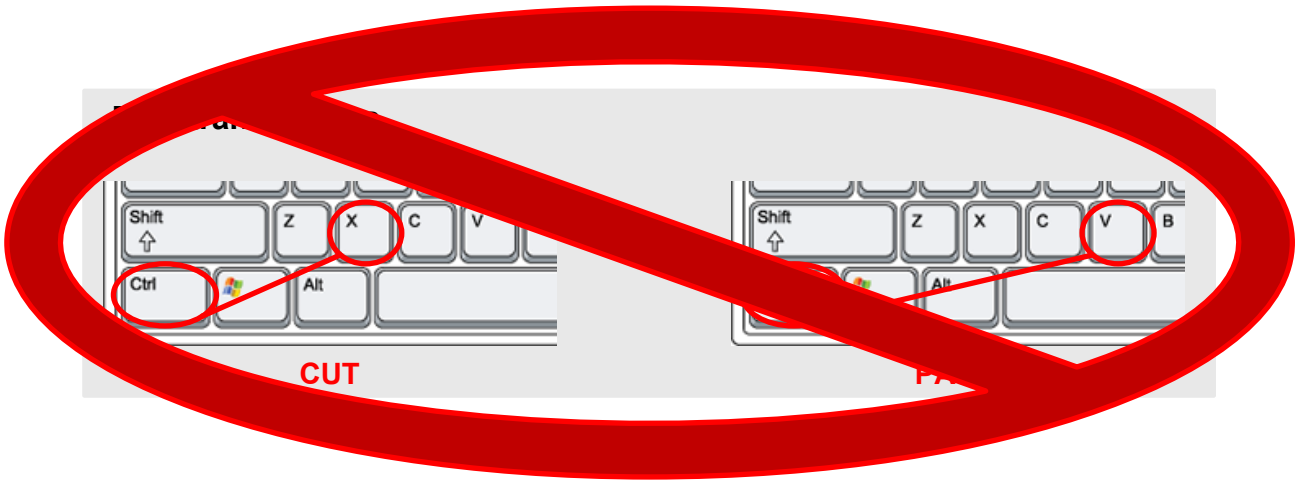
III. Following transposition, the double-stranded break is repaired by homology-dependent DNA repair



Skipper et al. 2013, *J. Biomed. Sci.*

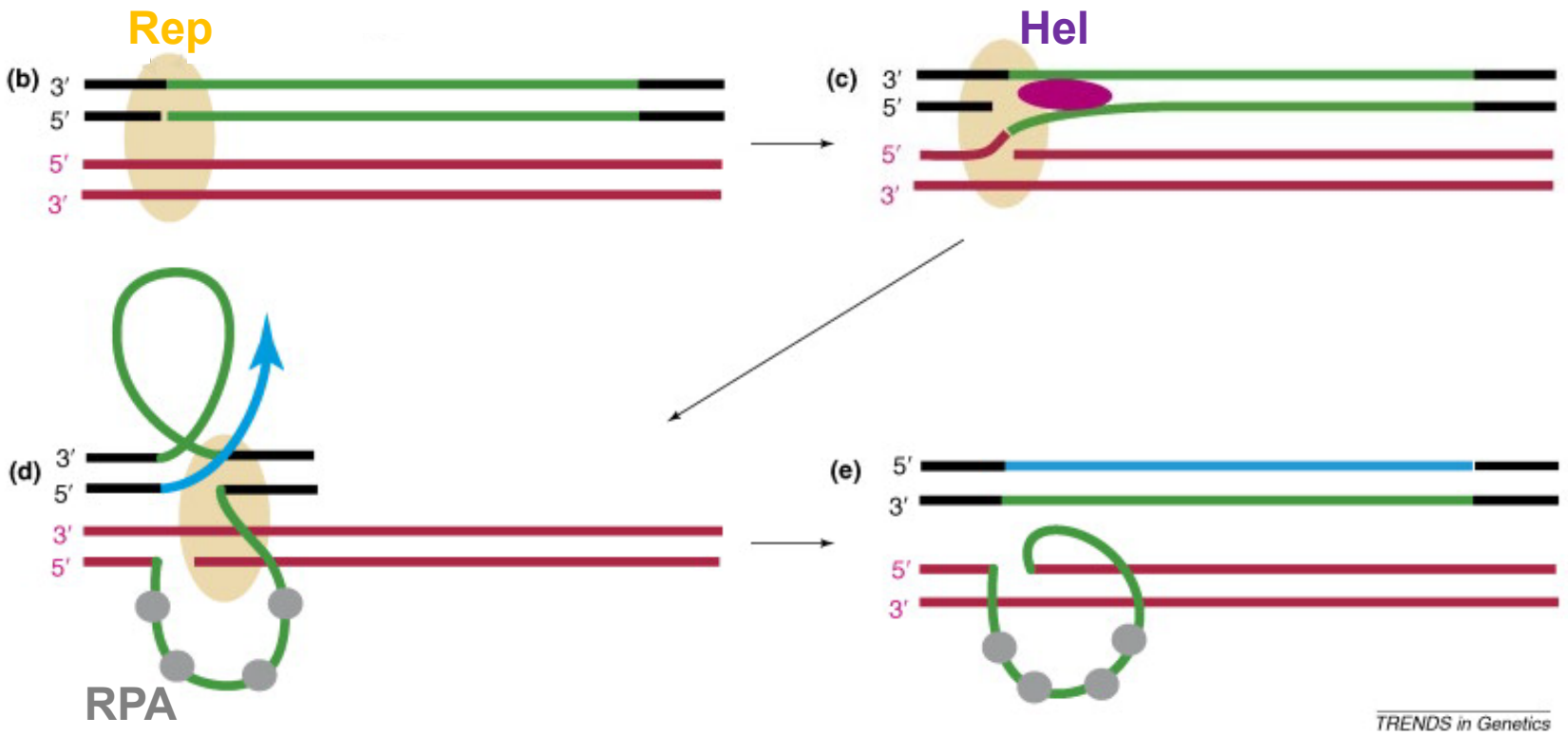
Class II: DNA transposons (subclass 2)

Classification		Structure	TSD	Code	Occurrence
Order	Superfamily				
Class II (DNA transposons) - Subclass 2					
Helitron	*Helitron*	RPA — Y2 HEL	0	DHH	P, M, F
Maverick	*Maverick*	C-INT — ATP — CYP — POL B	6	DMM	M, F, O



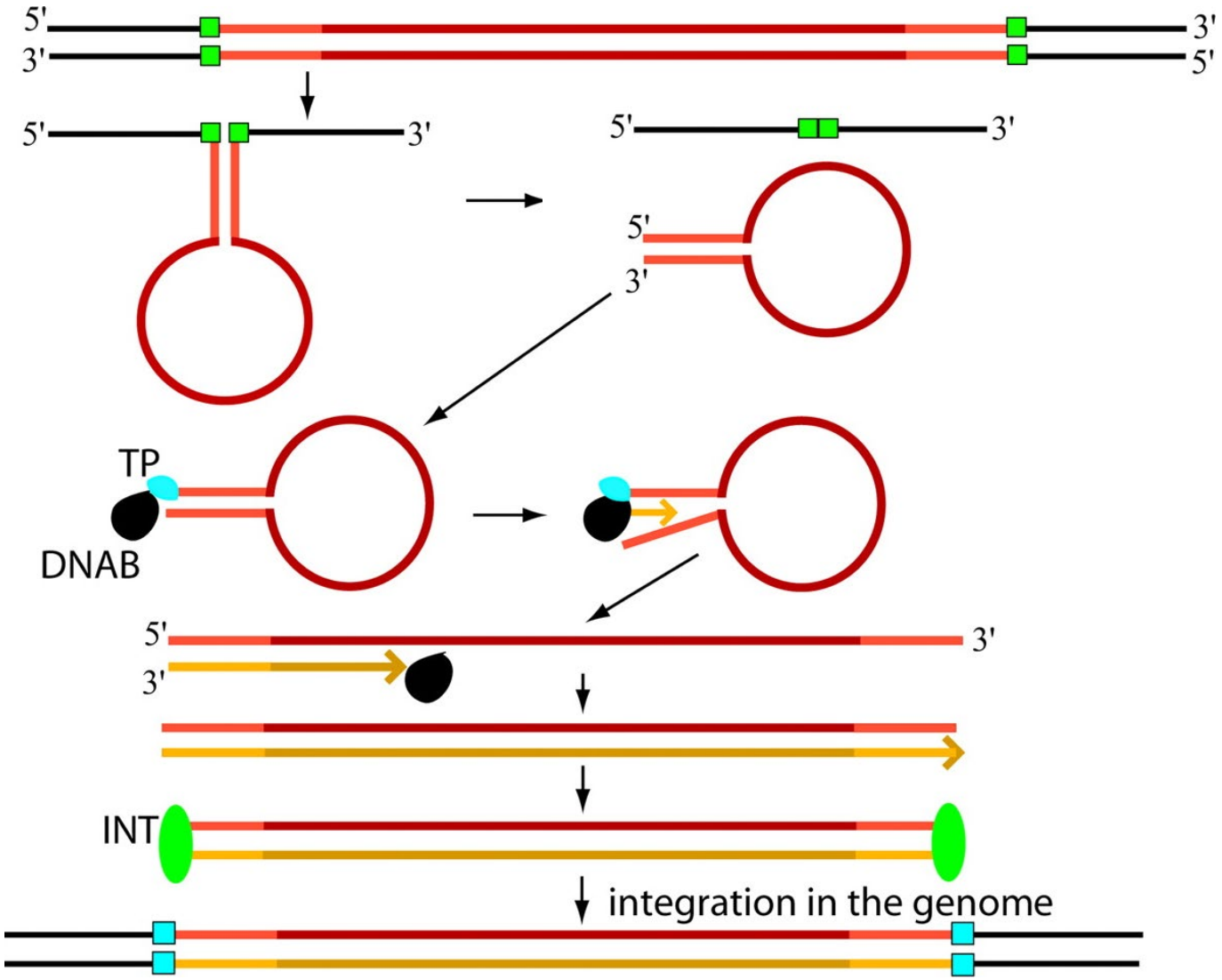
Wicker et al. 2008, *Nat. Rev. Genet.*

Helitrons: rolling-circle transposition



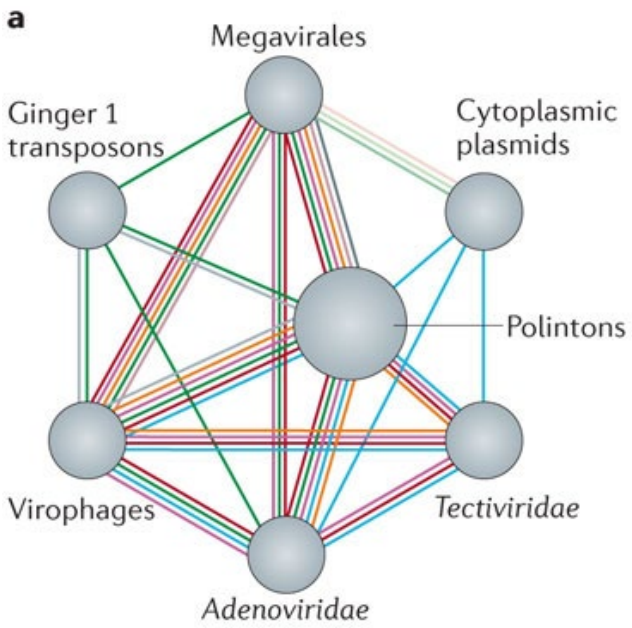
Kapitonov & Jurka 2007 *Trends Genet.*

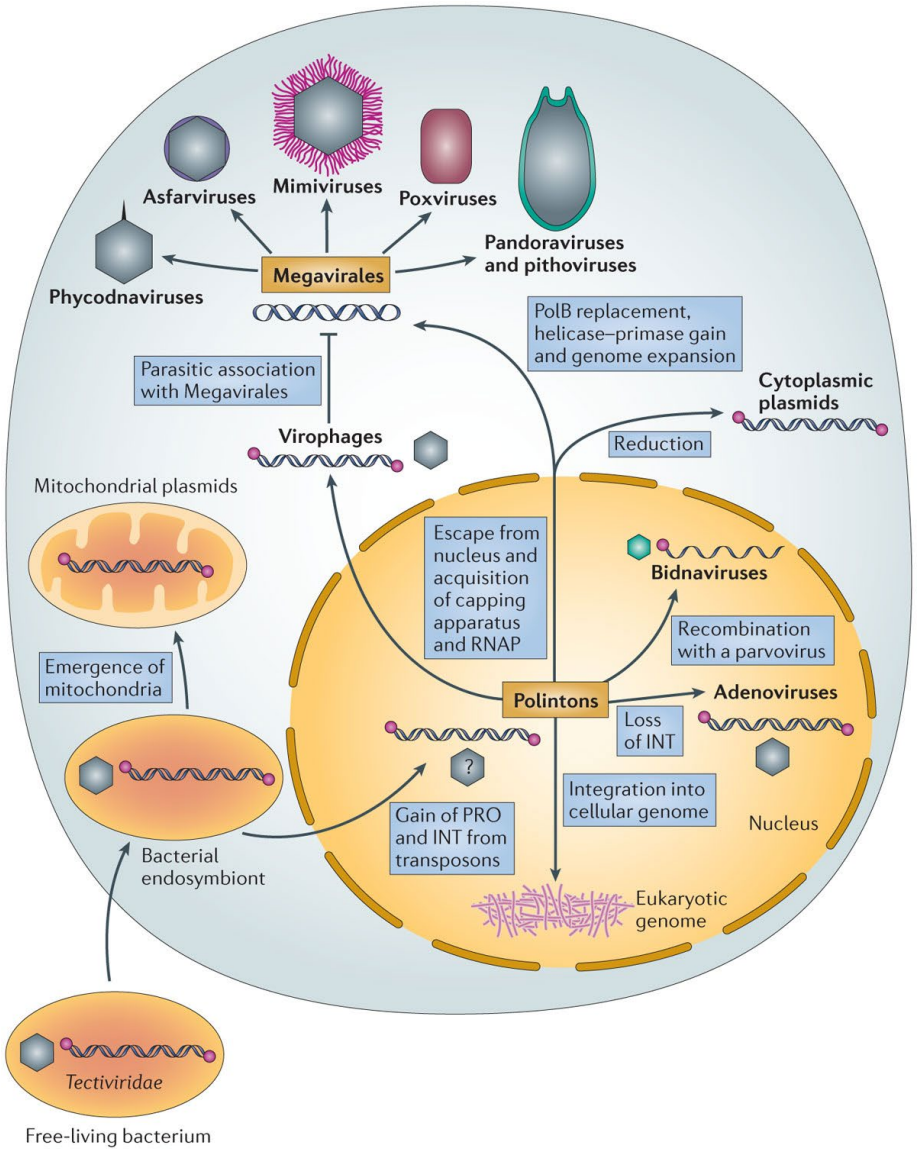
Mavericks: self-synthesizing transposition



Kapitonov & Jurka 2006 *PNAS*

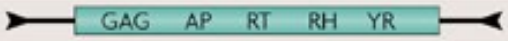
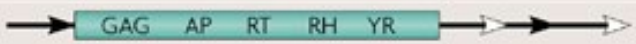
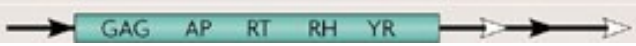
Viral origins of Mavericks/Polintons?





Krupovic & Koonin 2015 *Nat. Rev. Micro.*

Weirdos: TEs with tyrosine recombinase!

Classification		Structure	TSD	Code	Occurrence
Order	Superfamily				
Class I (retrotransposons)					
DIRS	*DIRS*		0	RYD	P, M, F, O
	Ngaro		0	RYN	M, F
	VIPER		0	RYV	O
Class II (DNA transposons) - Subclass 1					
Crypton	*Crypton*		0	DYC	F

TE integration mechanism occurs via:

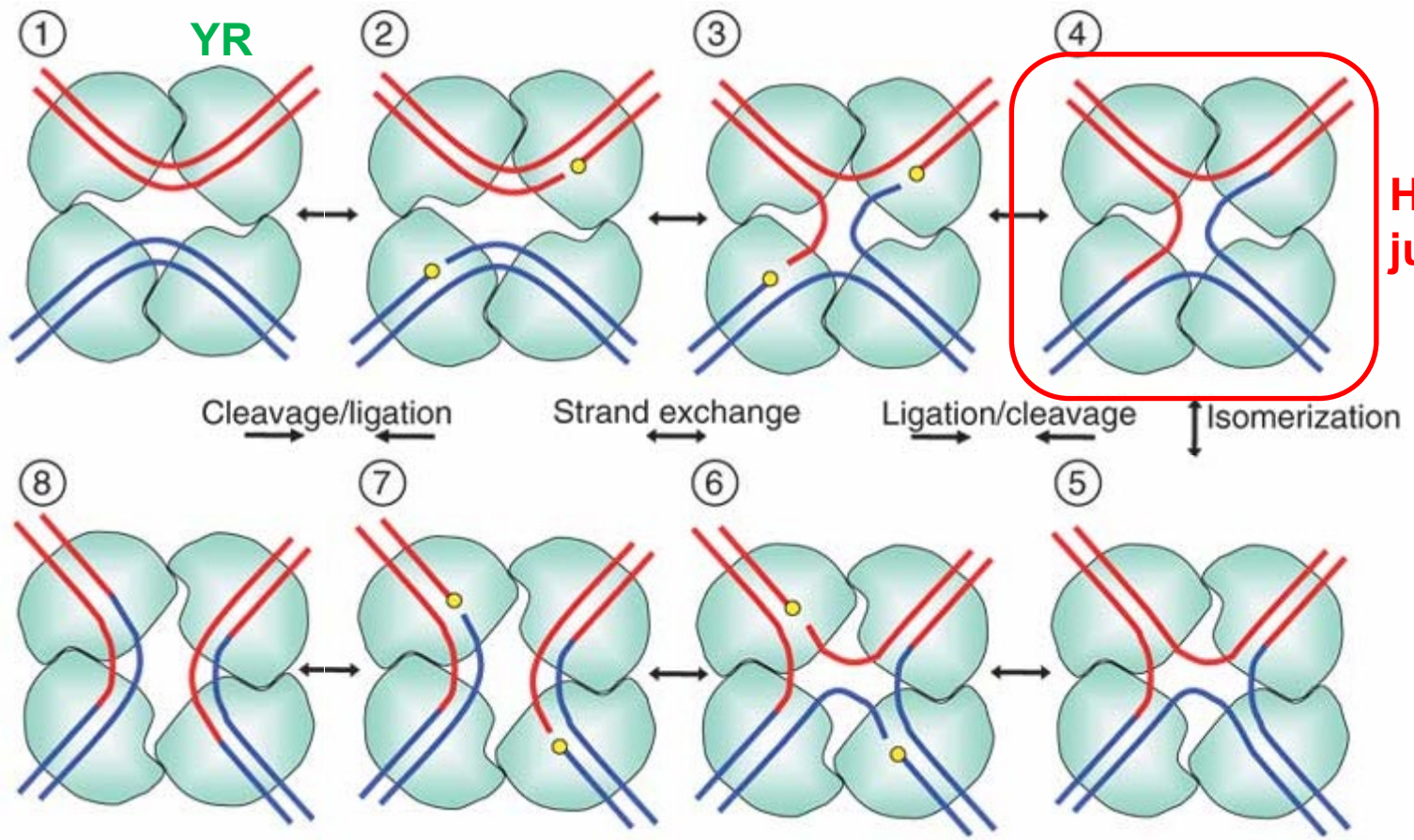
- **Endonuclease**: LINE, SINE, PLE
- **DDE-Transposase**: TIR
- **Integrase**: LTR, Maverick/Polinton
- **Rep protein**: Helitron
- **Tyrosine recombinase**: DIRS, Crypton

Class I (retrotransposons)
Class II (DNA transposons)

Note: TE classification is based on propagation/replication mechanism!

Wicker et al. 2008, *Nat. Rev. Genet.*

Site-specific DNA recombination

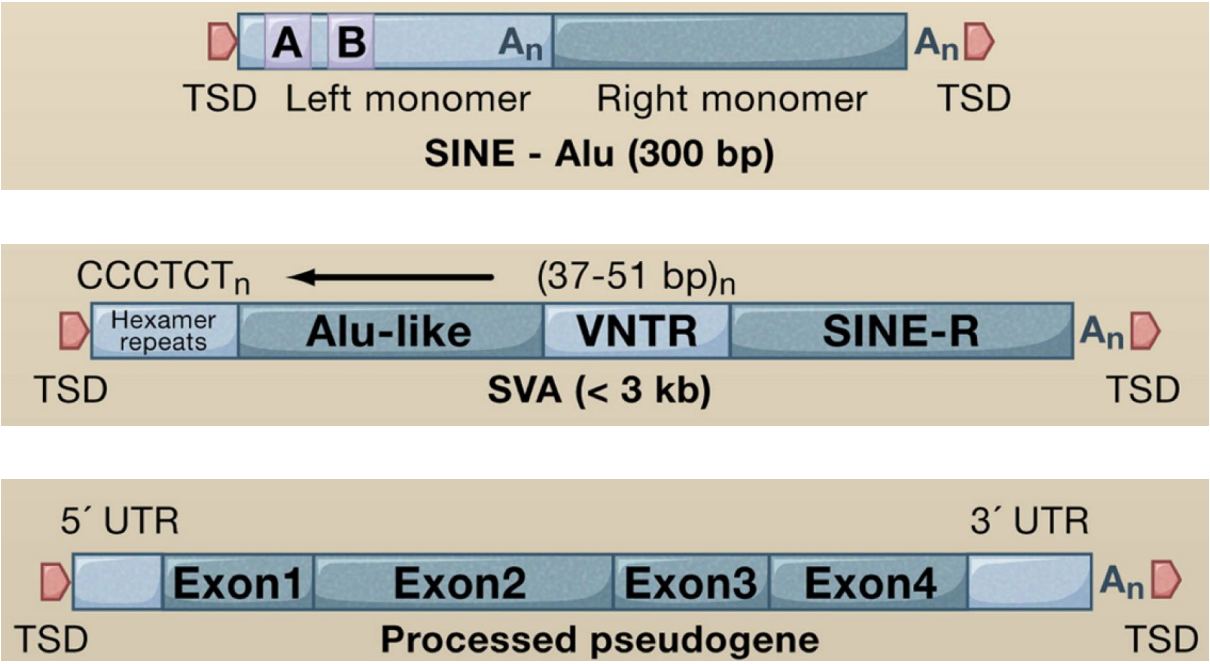


DIRS and Cryptons have no TSDs (but there might be exceptions)

Rice 2005, *Nat. Struct. Mol. Biol.*

Non-autonomous TEs: Not only SINEs

Non-LTR retrotransposons



= retro(pseudo)gene

Goodier 2016, *Mob. DNA*

Everything can be non-autonomous!

LTR retrotransposons

Retrotransposons with non-coding or antisense ORFs

LARDs

(large retrotransposon derivatives)

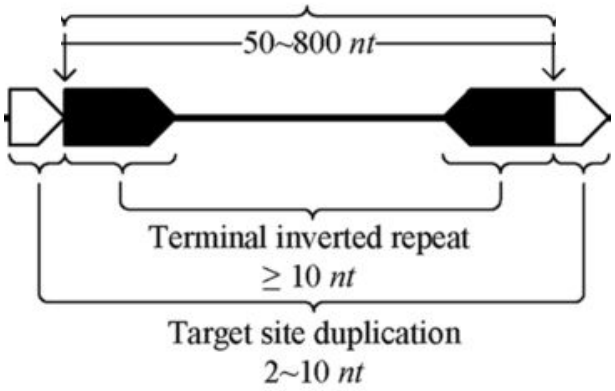
TRIMs

(terminal-repeat retro-transposons in miniature)

Havecker et al. 2004, *Genome Biol.*

DNA transposons

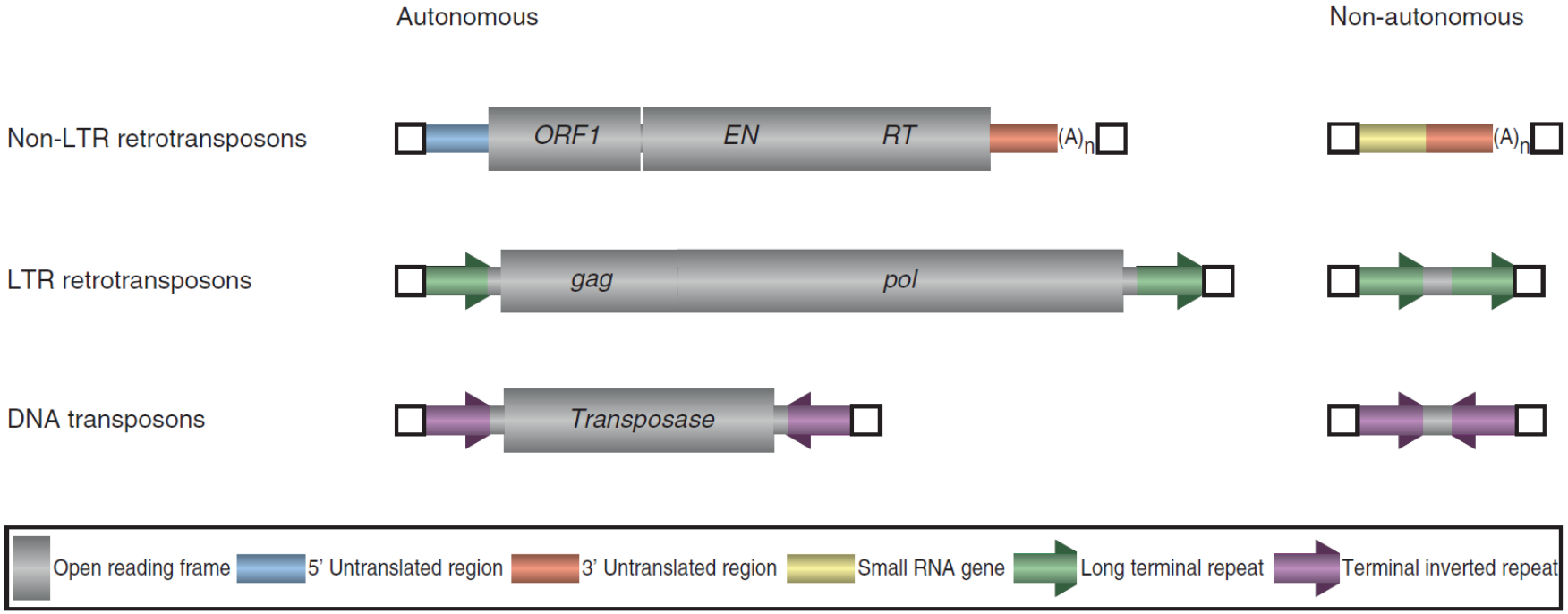
MITE (miniature inverted-repeat transposable element)

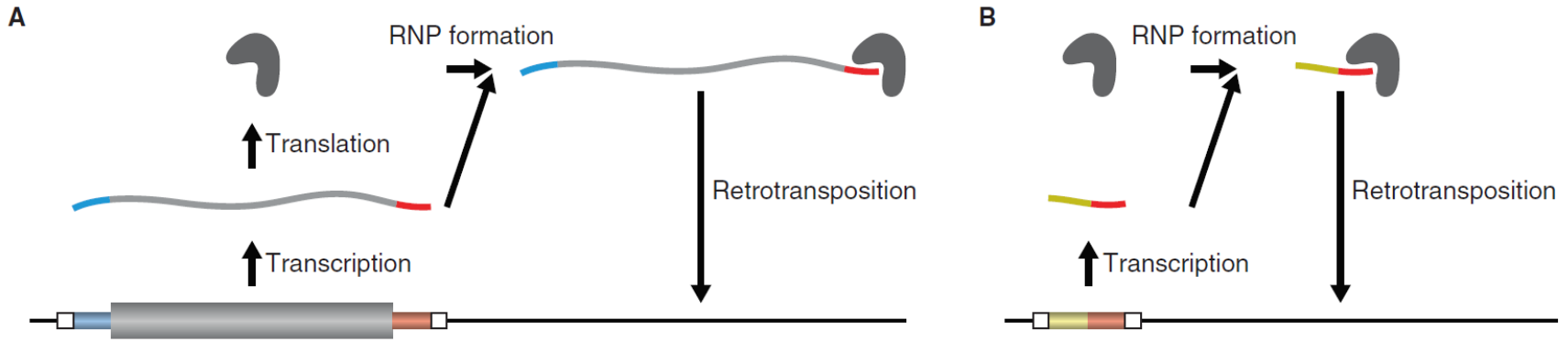


Often inconsistent nomenclature (maybe easier to add suffix "_NA")

Ye et al. 2016, *Sci. Rep.*

Minimum requirements per mechanism

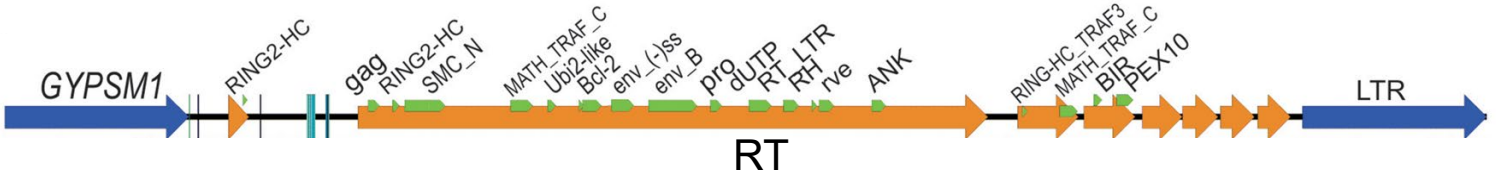




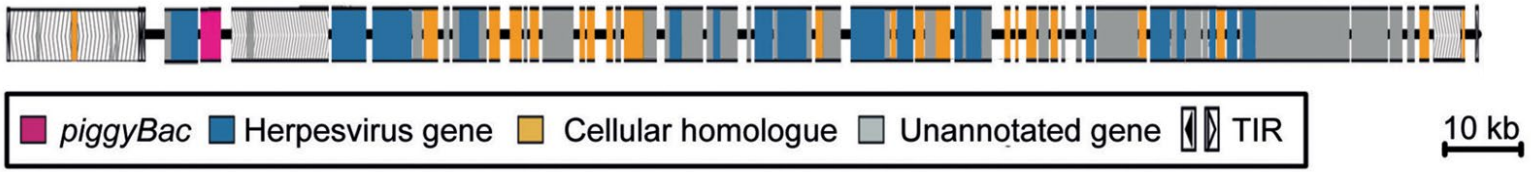
Suh 2019, *Curr. Biol.*

Upper limits for TEs: giant TEs!

LTR retrotransposons: *Burro* (30 kb, planaria *Schmidtea mediterranea*)

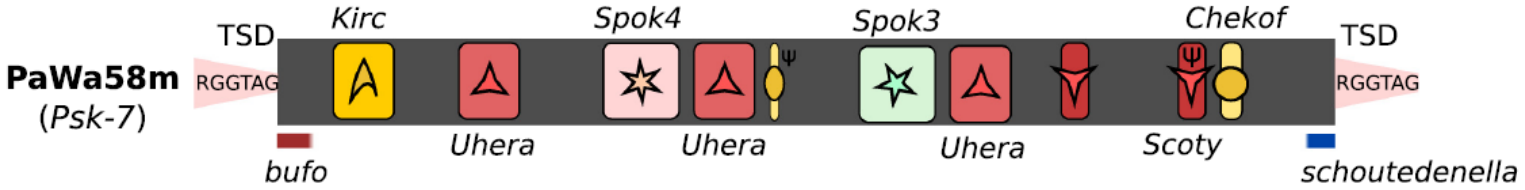


TIR DNA transposons: *Teratorn* (182 kb, medaka *Oryzias latipes*)



Arkhipova 2019, *Genome Biol. Evol.*

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



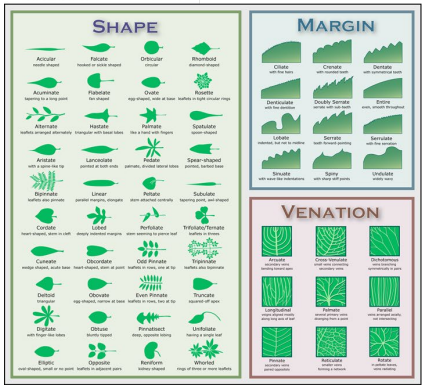
Mobility mechanisms define the upper size limit of TEs

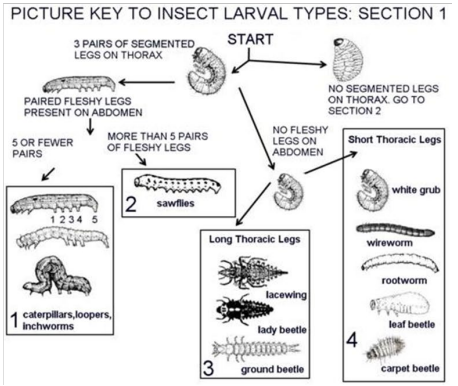
Vogan et al. 2021, *Genome Res.*

Conclusion: Genomes are microcosms!

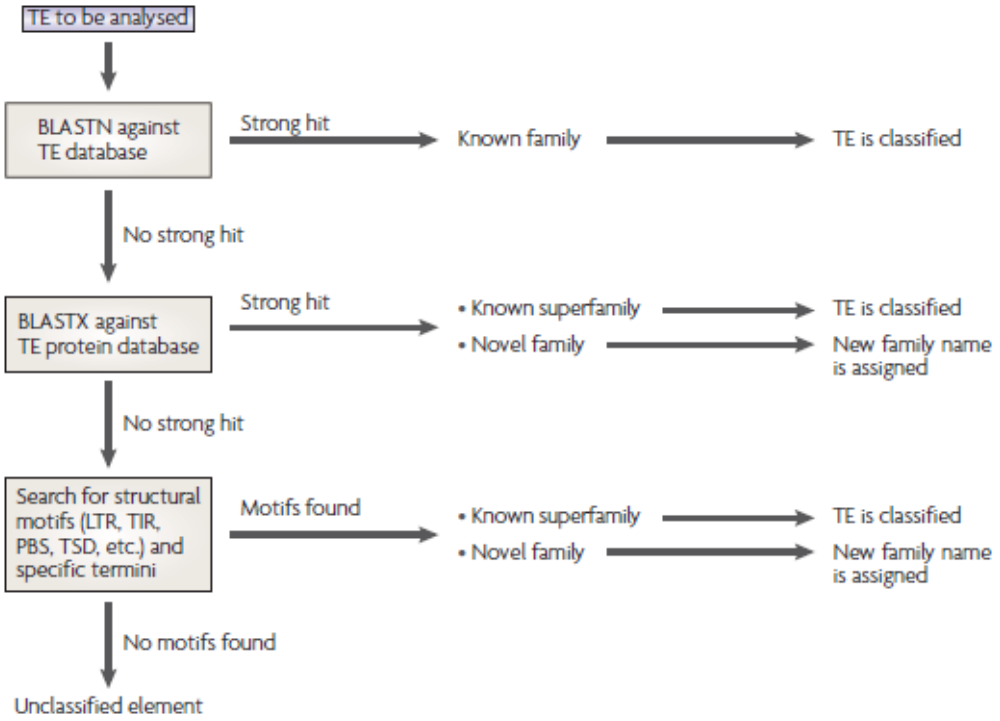
Host genes can become transposons/viruses

Transposons/viruses can become host genes





Introduction 2: Annotation limitations



How to pick a tool for finding (new) TEs?

Repeat tools

Description

This page compiles a list of software for the detection, annotation, analysis, simulation and visualization of repetitive, mobile and selfish DNA.

It is maintained by Tyler A. Elliott and a more metadata rich form of the data can be found here. It was initiated with the help of Elizabeth Smikle and Miduna Rahulan, formerly and currently at the Centre for Biodiversity Genomics at the University of Guelph. Suggestions, updates and error corrections can be directed to Tyler, or by commenting on this page.

We encourage the authors of these tools to create pages for them on TE Hub, so that they can provide more information about their work, and link it back to this table. Please find a template software sheet here.

Overview of tools for repeat analysis

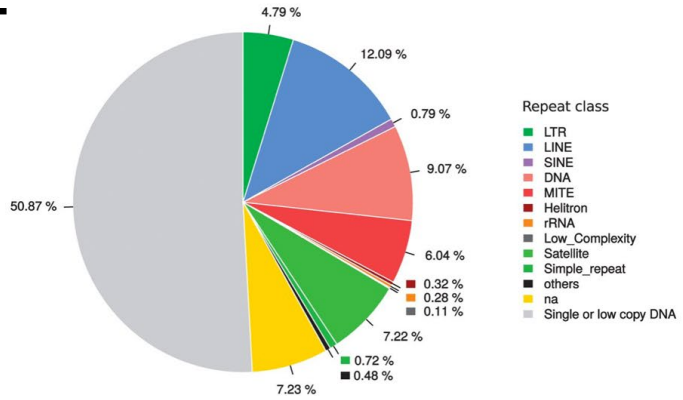
Tool	DOI	Alternate URL <Any>	Keywords De Novo
CARP	https://doi.org/10.1371/journal.pone.0193588		De Novo, Library Generation, Homology, Annotation
de_novo-identification	https://doi.org/10.13140/RG.2.2.27068.69765		De Novo, Eukaryotic Transposon
dnaasm	https://doi.org/10.7490/f1000research.1114626.1		Repeat, NGS/HTS, De Novo
EDTA	https://doi.org/10.1186/s13059-019-1905-y		Library Generation, Filtering, Structure, De Novo

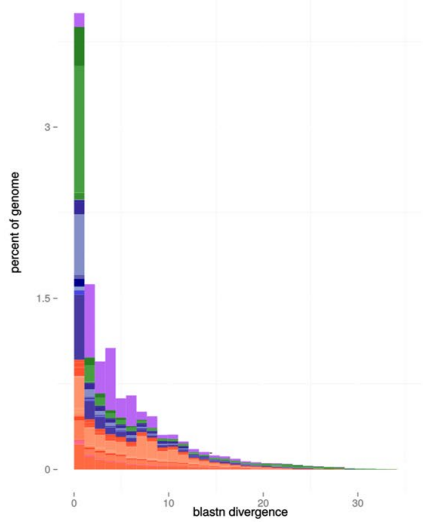
632 tools listed at present (45 alone for de-novo analysis!)!

https://tehub.org/en/resources/repeat_tools; TE Hub Consortium 2021 *Mobile DNA*

TE prediction: (Short-)read-based

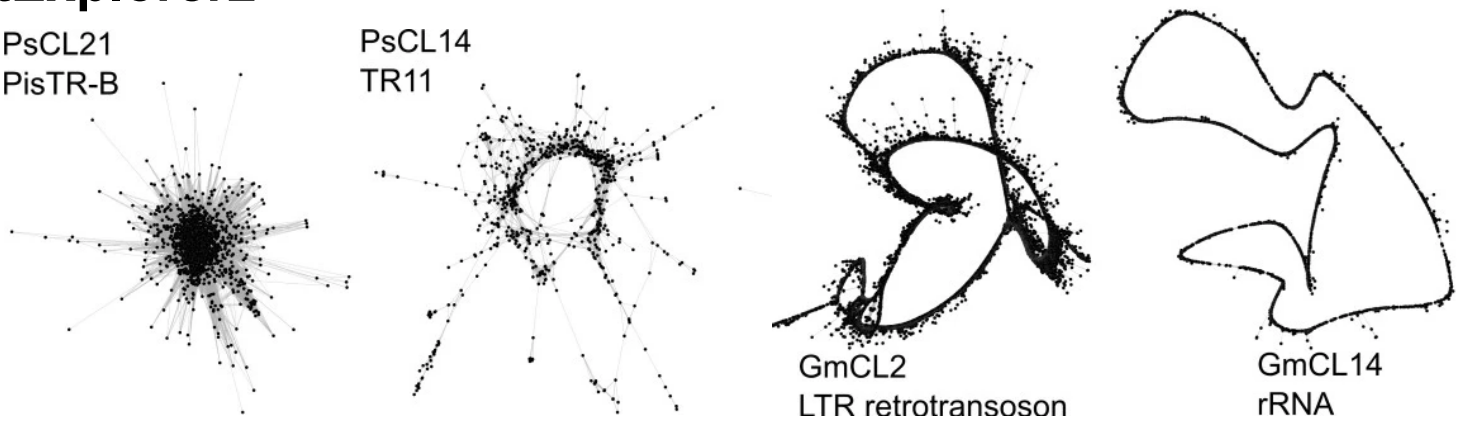
dnaPipeTE





https://github.com/clemgoub/dnaPipeTE; Goubert et al. 2015 *Genome Biol. Evol.*

RepeatExplorer2



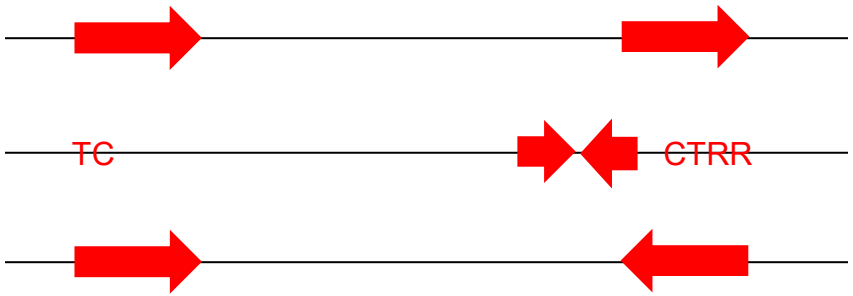
http://repeatexplorer.org/?page_id=818; Novák et al. 2020 *Nat. Protocols*

Pros: read quantification, satellite curation; Cons: hard to curate TEs (no TSDs)

TE prediction: Assembly-based

Structure-based approach:

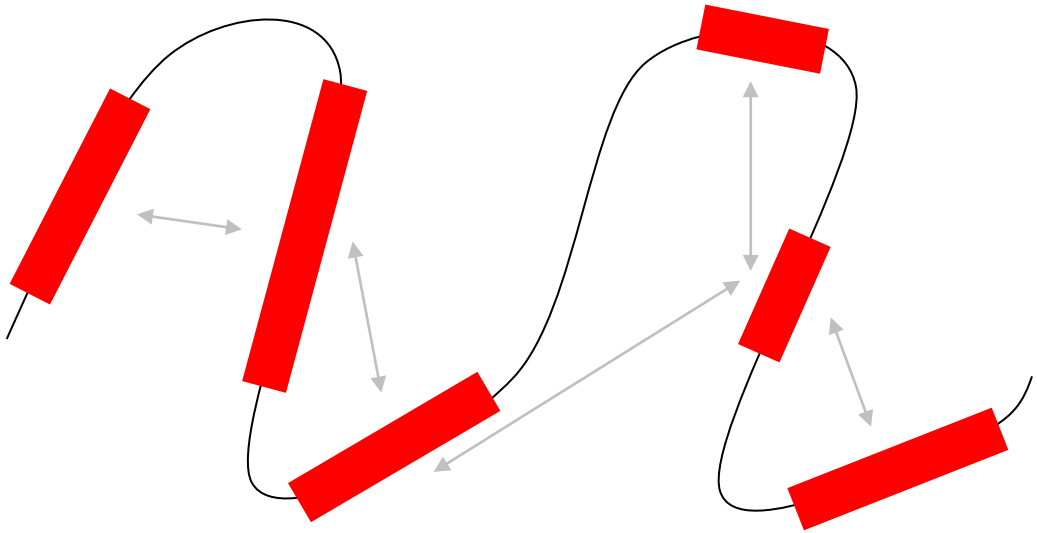
LTRharvest/digest
HelitronScanner
MITE_Hunter_2
…



Pros: good for low-copy TEs; Cons: hard to curate (esp. non-autonomous TEs)

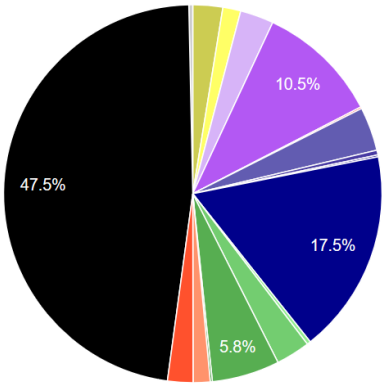
Homology-based approach:

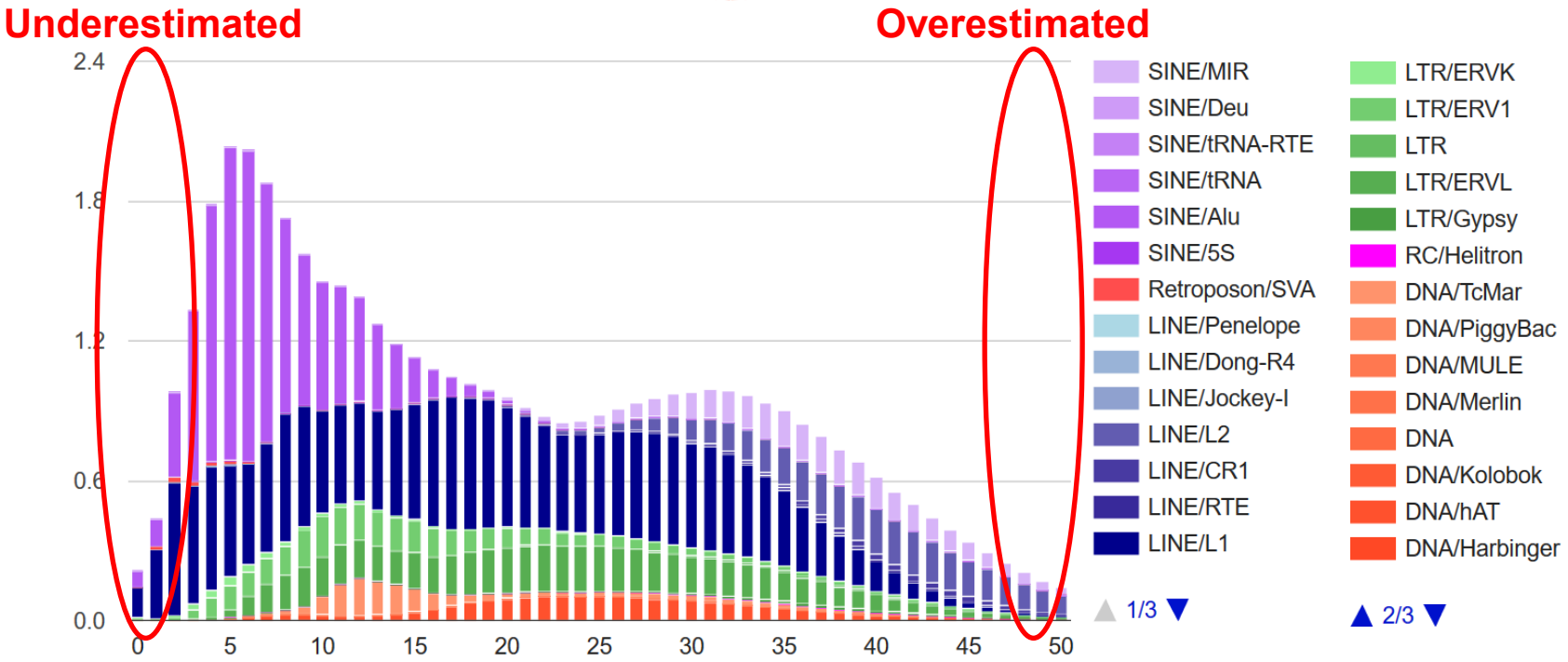
RepeatModeler2
RepeatMasker
Tandem Repeats Finder
REPET
CARP
…



Pros: good for high-copy TEs; Cons: bad for low-copy TEs or bad assemblies

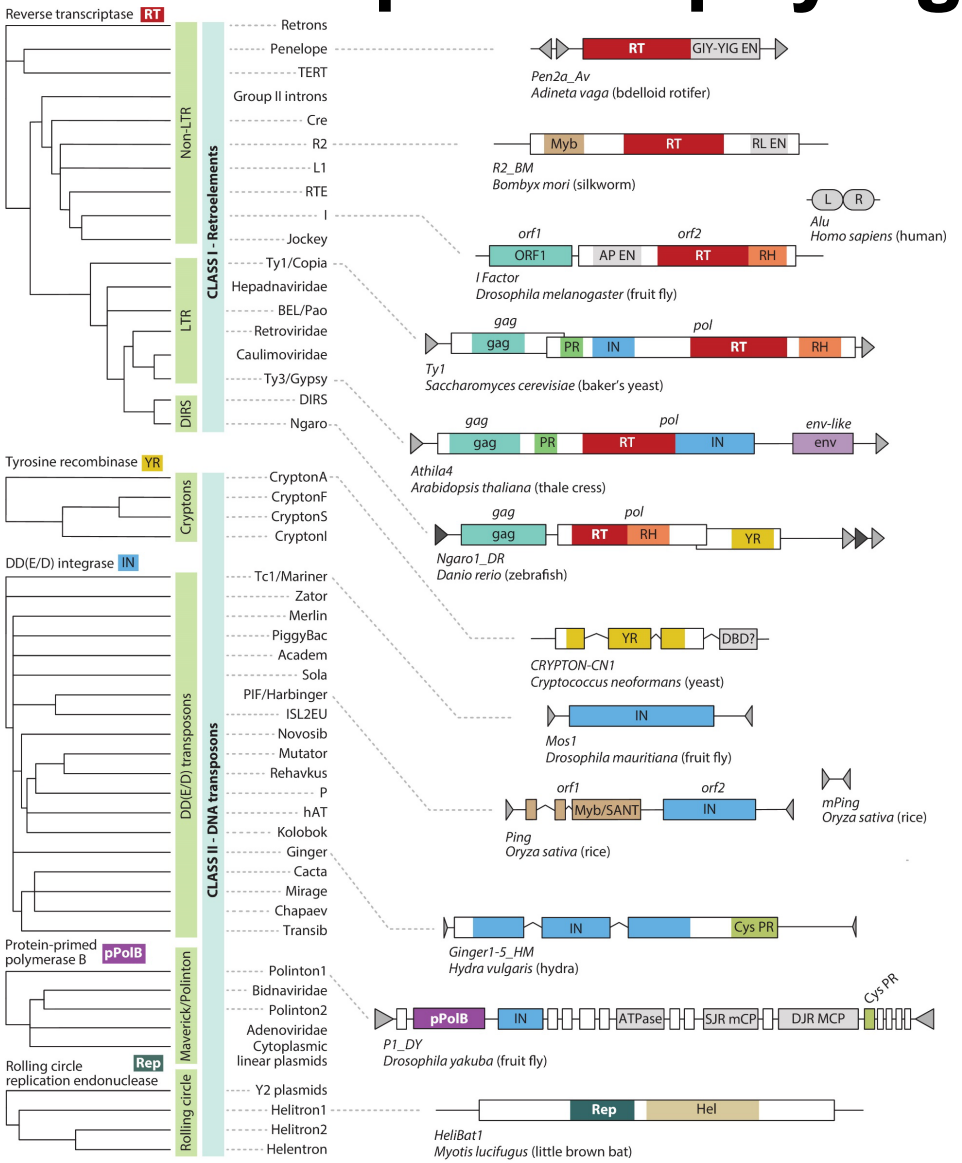
TE landscape of the human genome





http://www.repeatmasker.org/species/hg.html

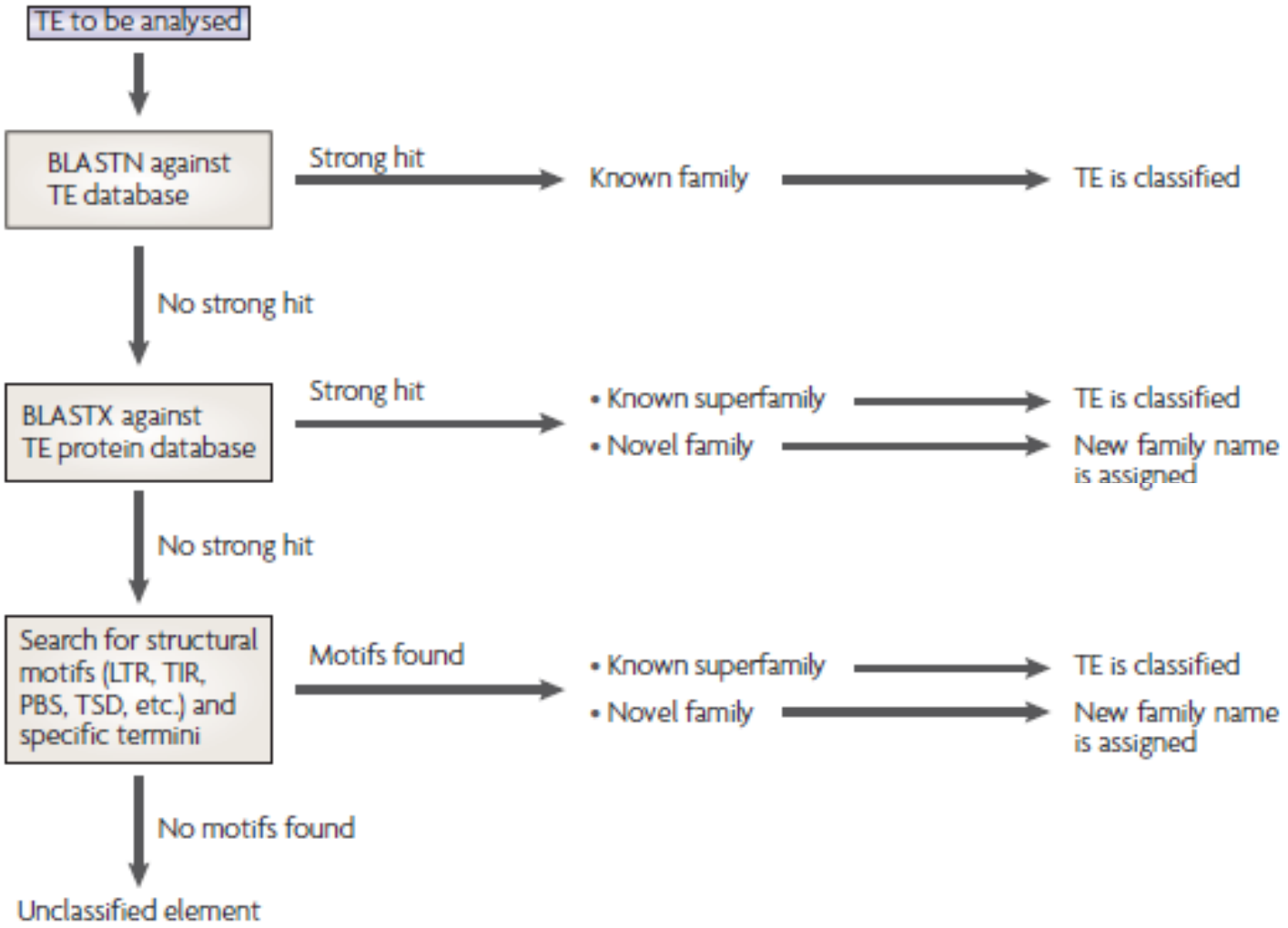
Classification vs. protein phylogeny



Wells JN, Feschotte C. 2020
Annu. Rev. Genet. 54:539–61

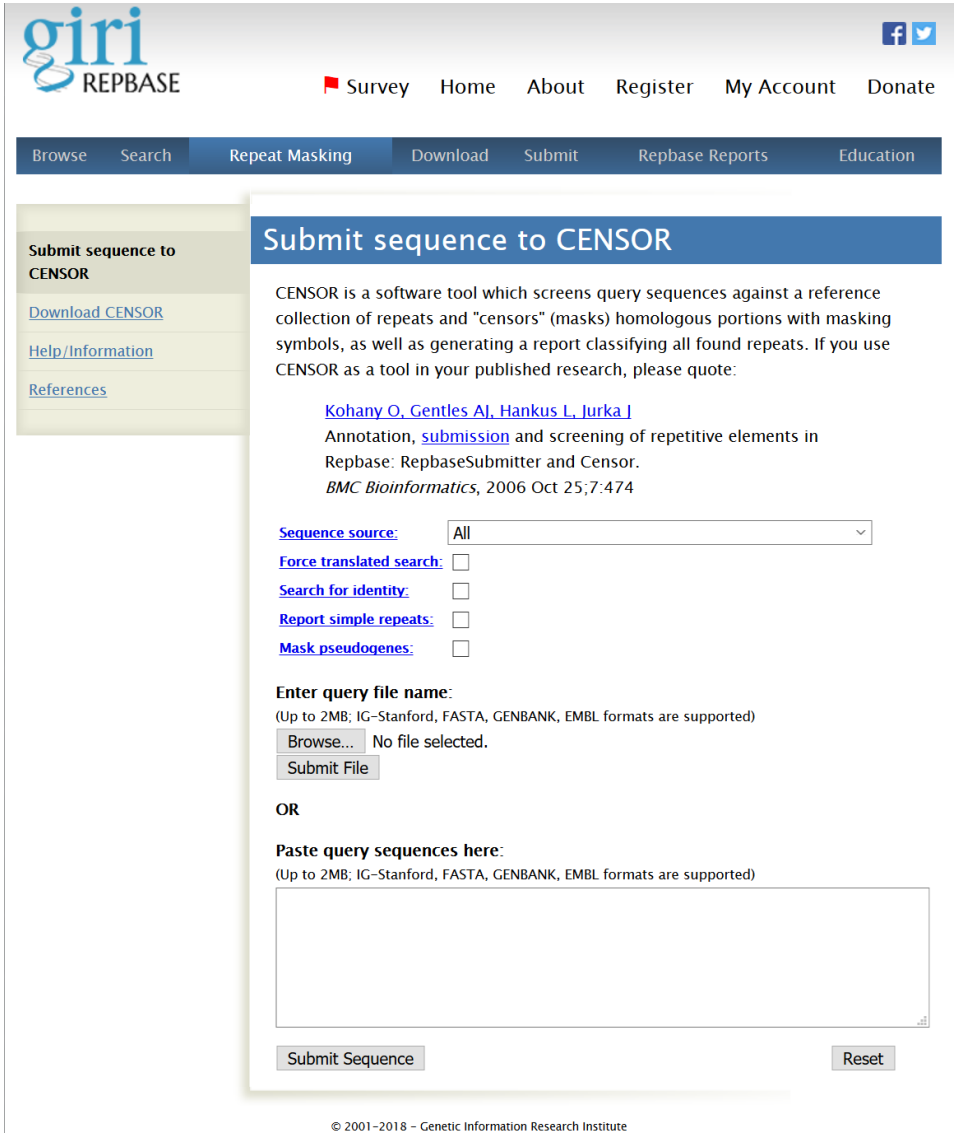
Wells & Feschotte 2020 *Annu. Rev. Genet.*

Stepwise classification of TEs



Wicker et al. 2008, *Nat. Rev. Genet.*

Quick check in Repbase CENSOR



http://www.girinst.org/censor/index.php, Jurka et al. 1996 *Computers Chem.*

Any other database more applicable?

Repeat databases

Description

This page compiles a list of databases for the storage of sequences and metadata associated with repetitive, mobile and selfish DNA.

It is maintained by Tyler A. Elliott and a more metadata-rich form of the data can be found here. It was initiated with the help of Elizabeth Smikle and Miduna Rahulan, formerly at the Centre for Biodiversity Genomics and currently at the University of Guelph. Suggestions, updates and error corrections can be directed to Tyler, or by commenting on this page.

Overview of repeat databases

Resource	DOI	Taxonomic Group <Any>	Repeat Types <Any>
3'UTR-SIRF	https://doi.org/10.1186/1471-2105-8-274	Mammal	SINE
ACLAME	https://doi.org/10.1093/nar/gkp938	Archaea, Bacteria	Plasmid, Virus
alu_ontology	https://doi.org/10.1016/j.jbi.2016.01.010	Homo sapiens	Alu, SINE
ARDB (Antibiotic Resistance Genes Database)	https://doi.org/10.1093/nar/gkn656	Archaea, Bacteria	AMR/Antibiotic Resistance
ArTEDB (Arthropod Transposable Elements Database)	https://doi.org/10.3390/genes10050338	Arthropod	Eukaryotic Transposon

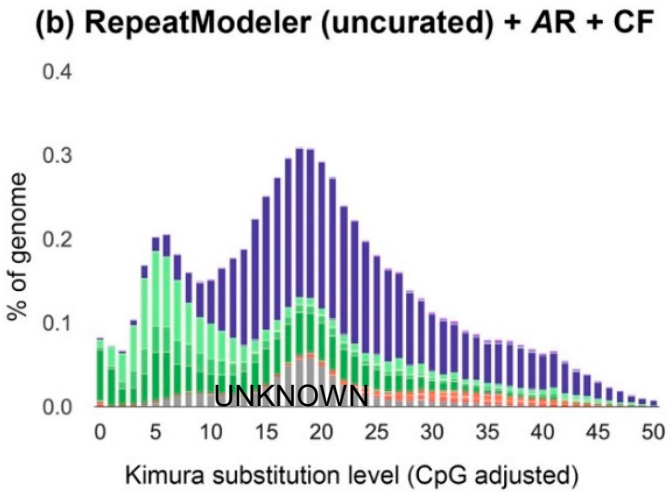
180 databases listed at present!

https://tehub.org/en/resources/repeat_databases; TE Hub Consortium 2021 *Mobile DNA*

Limitations of automatic classification

Species <u>with</u> curated repeat library from close relative (e.g., bird):

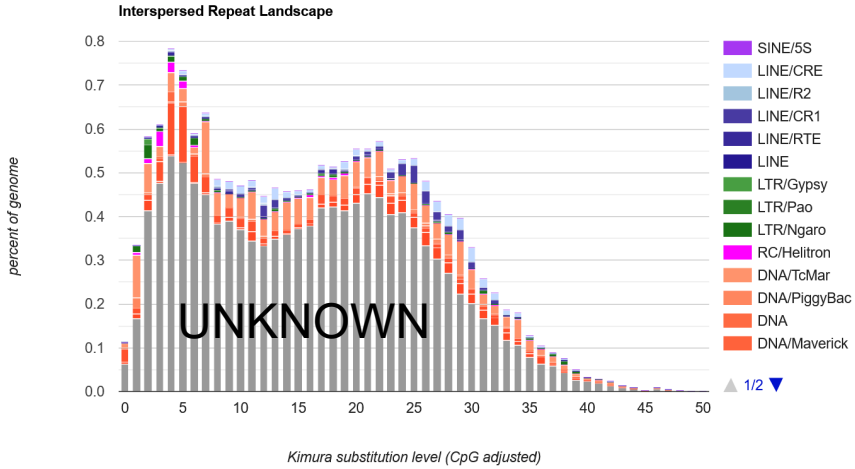
- Multicopy host genes
- Some satellites
- Some non-autonomous TEs
- Some solo-LTRs
- Some very 5'-truncated LINEs/SINEs

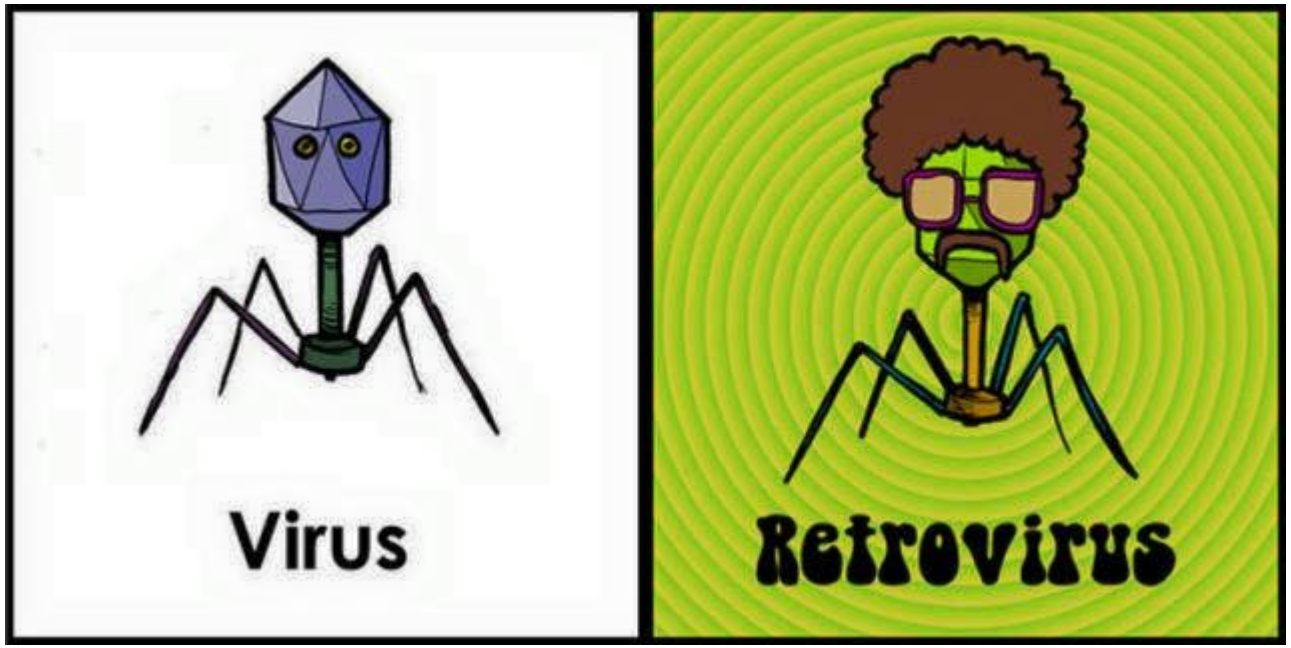


Boman et al. 2019, *Genes*

Species <u>without</u> curated repeat library from close relative (e.g., tardigrade):

- Multicopy host genes
- Satellites
- Non-autonomous TEs
- Solo-LTRs
- Very 5'-truncated LINEs/SINEs
- TEs with protein-coding domains too divergent from repeat databases




Virus
Retrovirus

Schedule for this afternoon

13:00-14:00 Biology of transposable elements (Alexander Suh)

14:00-14:15 Break

14:15-15:45 Visualization and analyses of repeats in R

15:45-16:00 Evaluation and final words

Dropbox folder with a TE classificiation quiz to practice decision making

Preprint on TE annotation curation with a tardigrade case study