

Structural variation

(A story of surprise → frustration → hope)



Alexander Sang-Jae Suh

서상재 徐商在

January 23, 2025

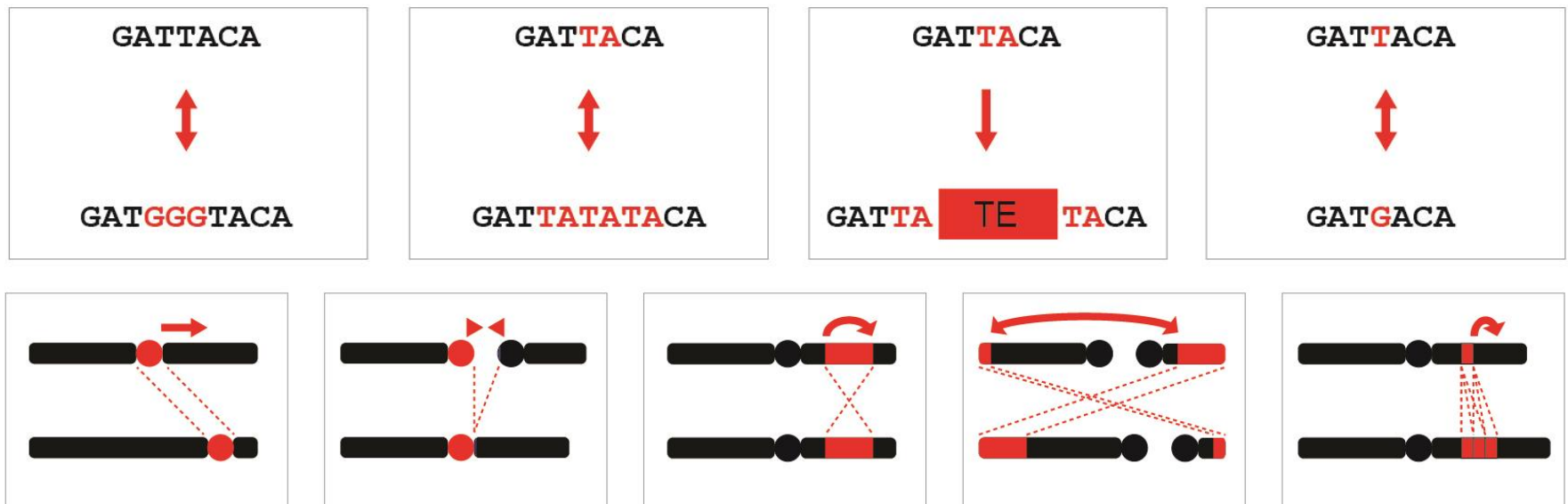


What is structural variation?

Structural variant (SV): genomic variation between individuals affecting the presence, abundance, position, and/or direction of a nucleotide sequence

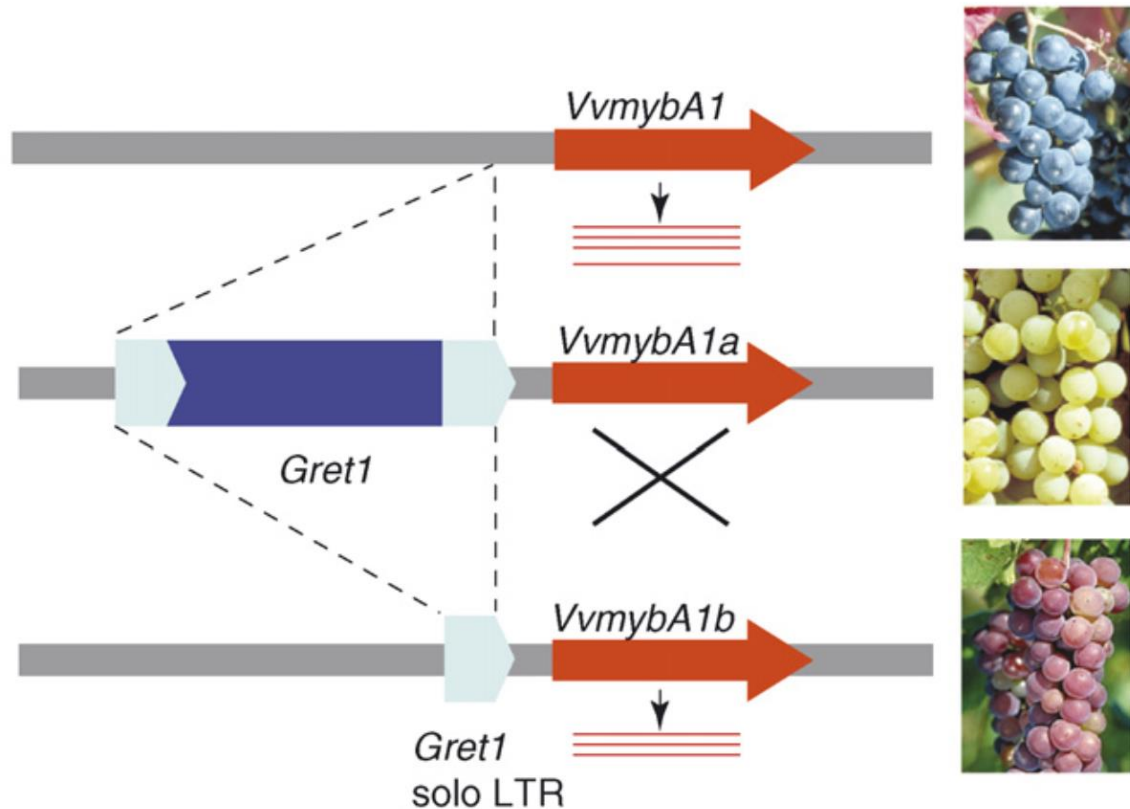
[Mérot et al. 2020, Trends Genet.](#)

Some key mutation types



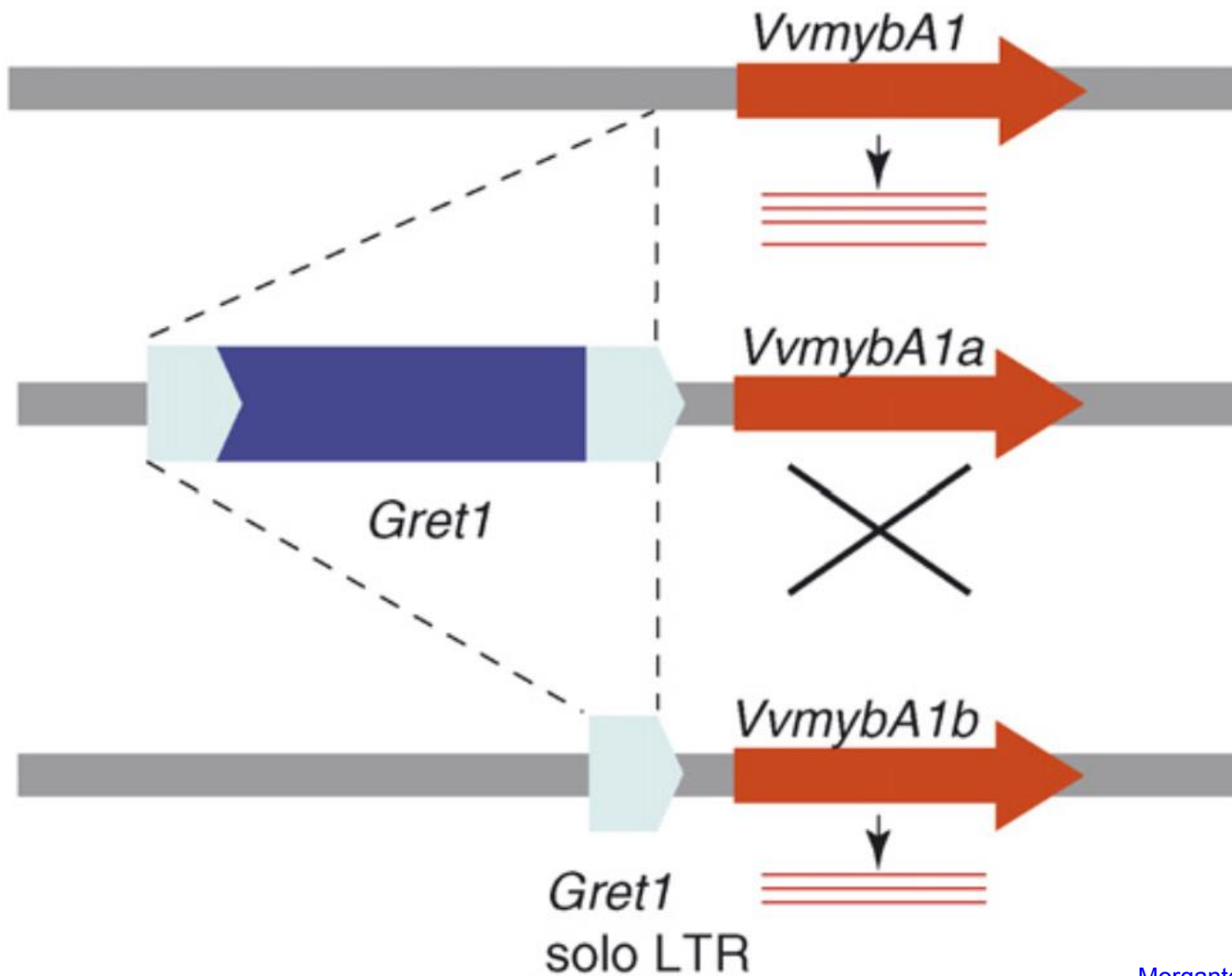
[Berdan et al. 2021, Mol. Ecol.](#)

Part 1: Surprise



A) It's not a SNP!

Delicious effects of SVs



Discovery of gene regulation in 1940s

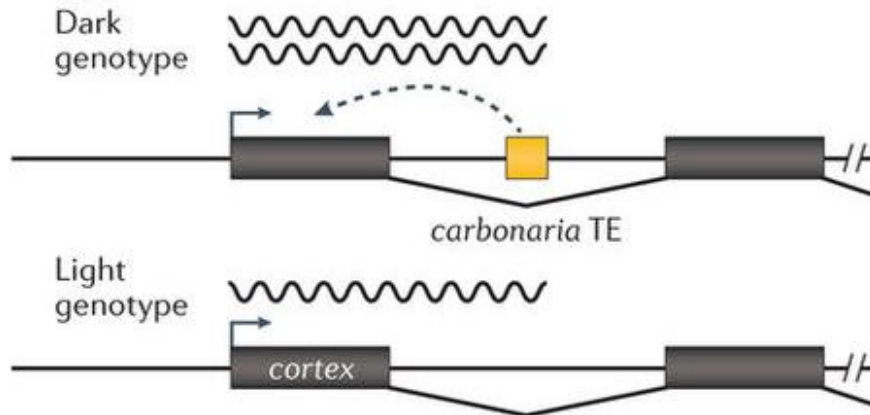


**Barbara
McClintock**
(Nobel Prize in
Physiology or
Medicine 1983)

TE-induced rapid adaptation

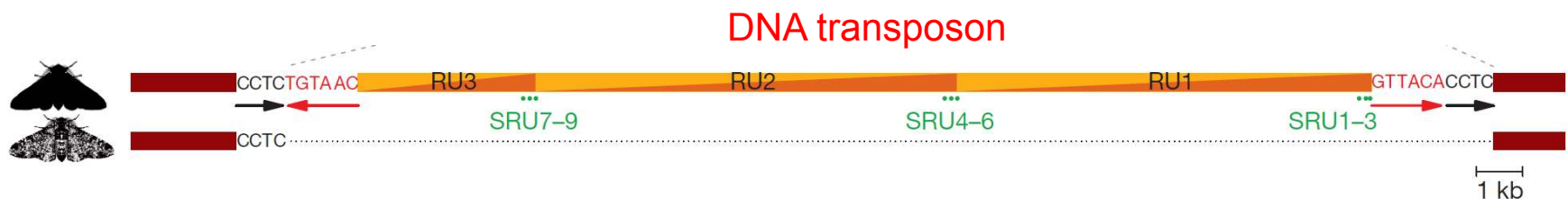
The industrial melanism of the peppered moth is probably the most famous textbook example for adaptation (in only a few decades)!

c Peppered moth



Upregulates cortex, resulting in increased darker coloration

[Chuong et al. 2017 Nat. Rev. Genet.](#)



[Van't Hof et al. 2016, Nature](#)

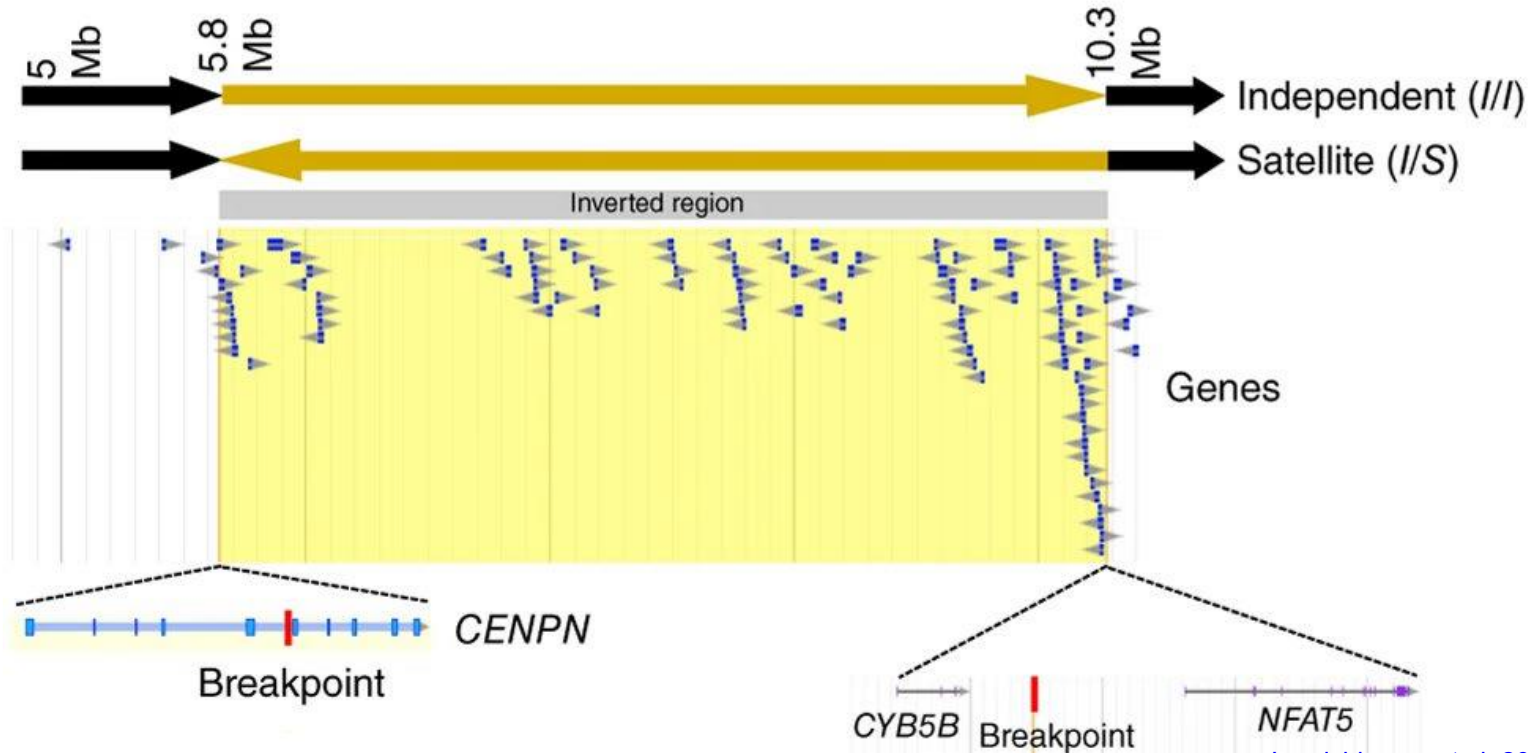
Inversions in ruff reproductive strategies

Independent

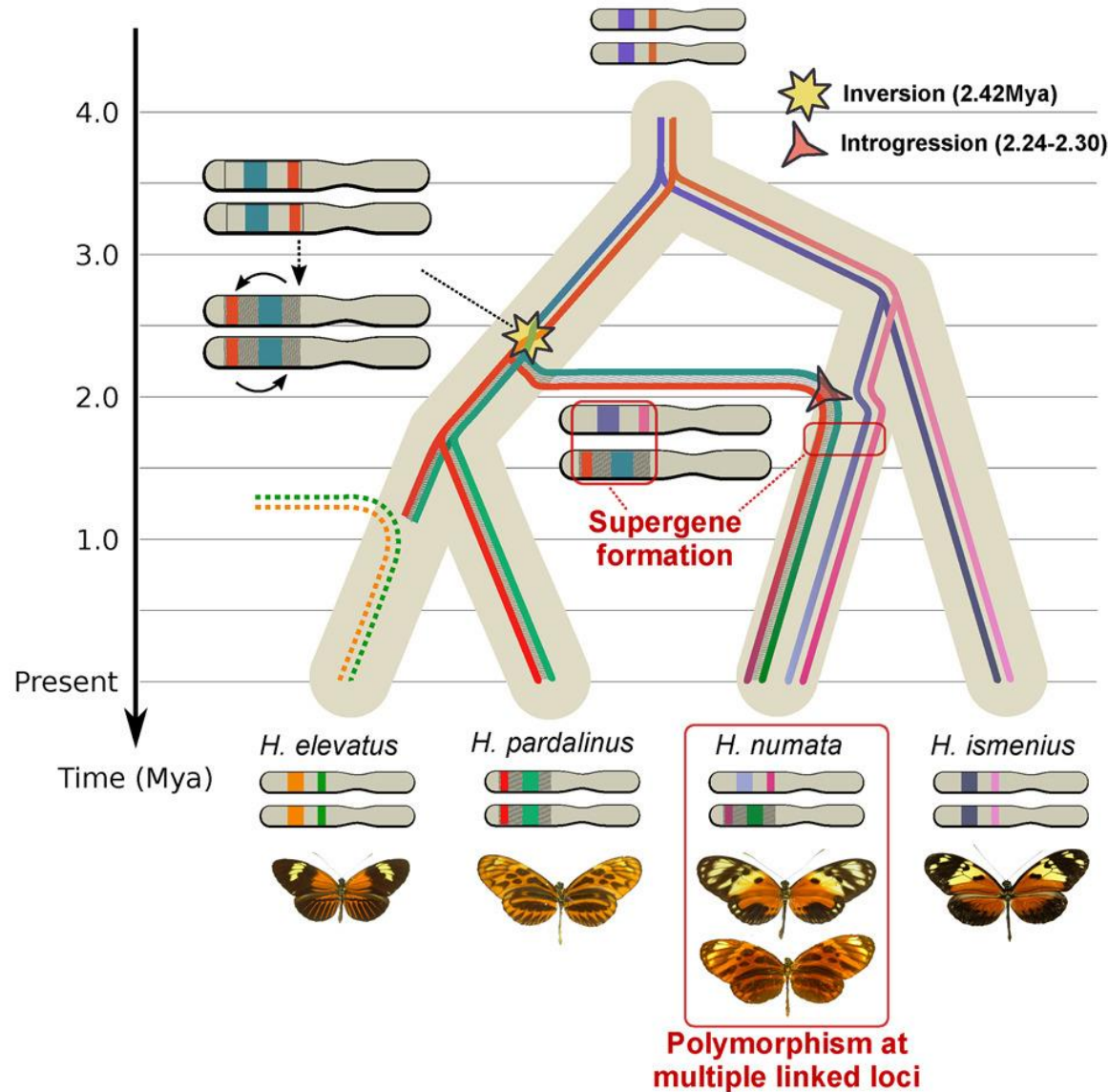
Satellite

Independents

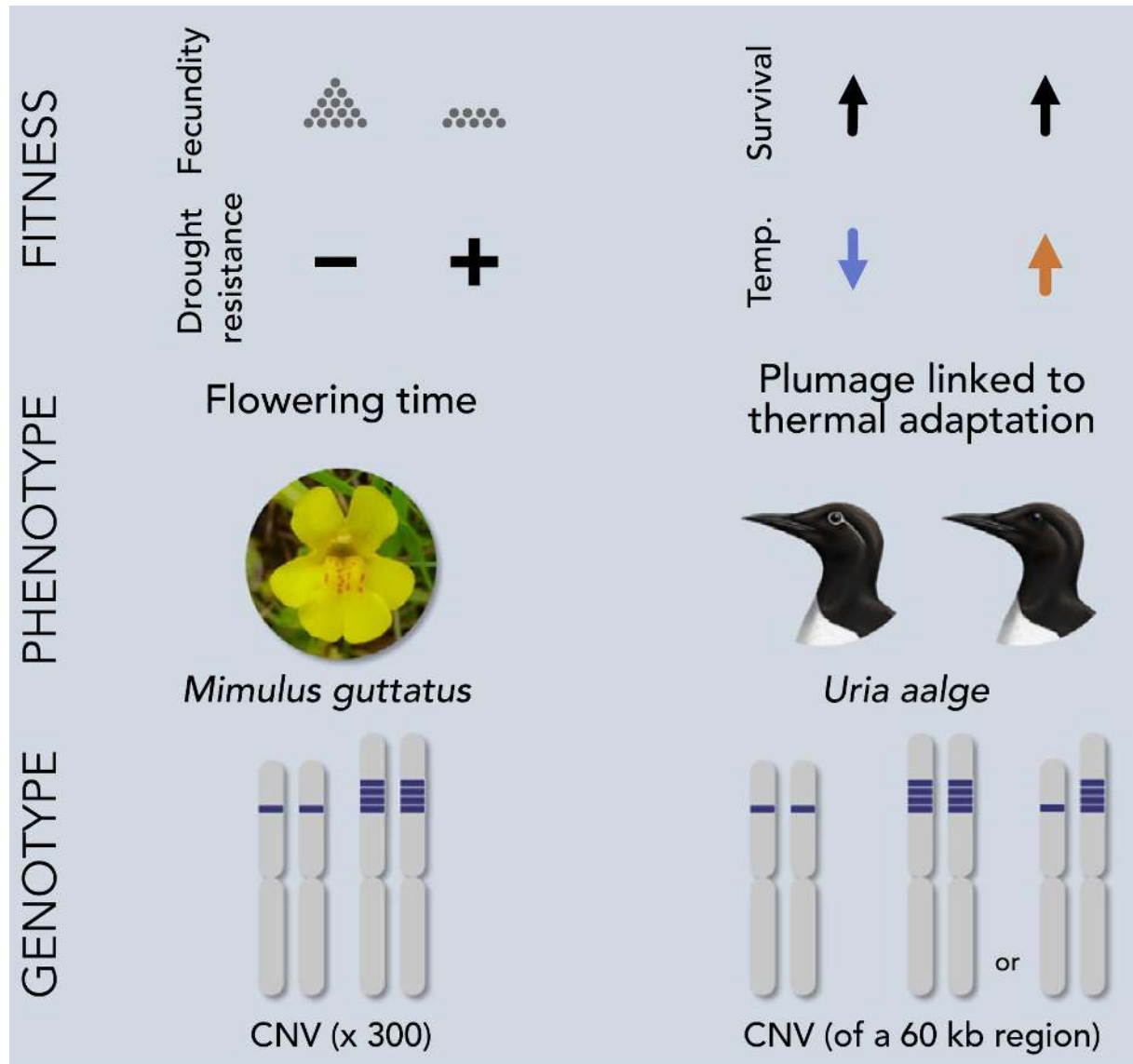
Faeder



Inversion introgression and supergenes

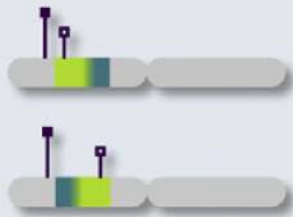


Duplications



High diversity of possible effects

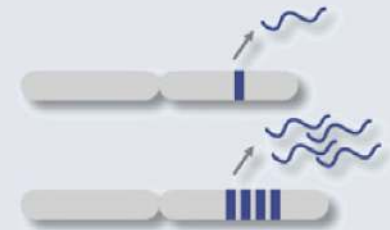
Gene-regulation decoupling



Position-effect variegation



Gene dosage



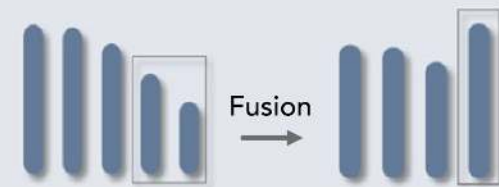
Linear sequence



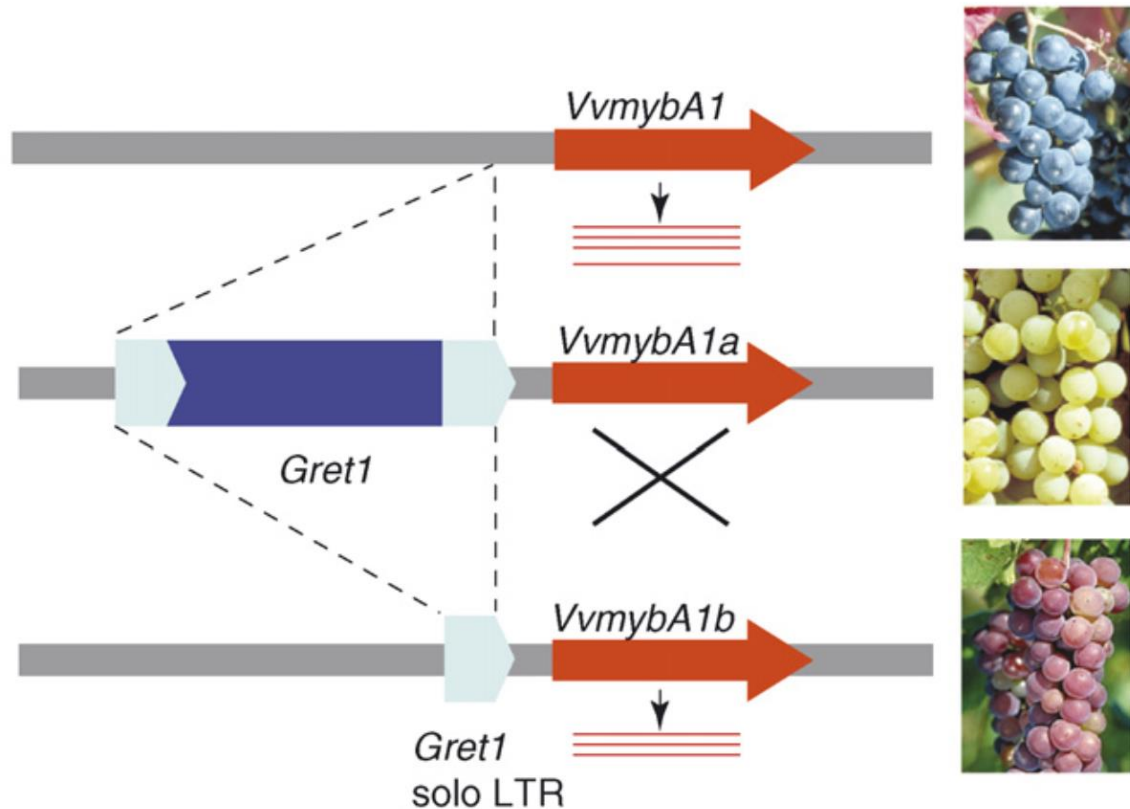
Recombination



Karyotype divergence



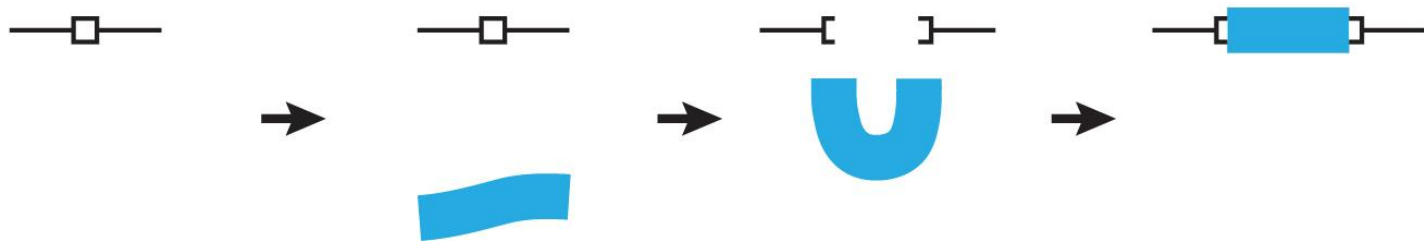
Part 1: Surprise



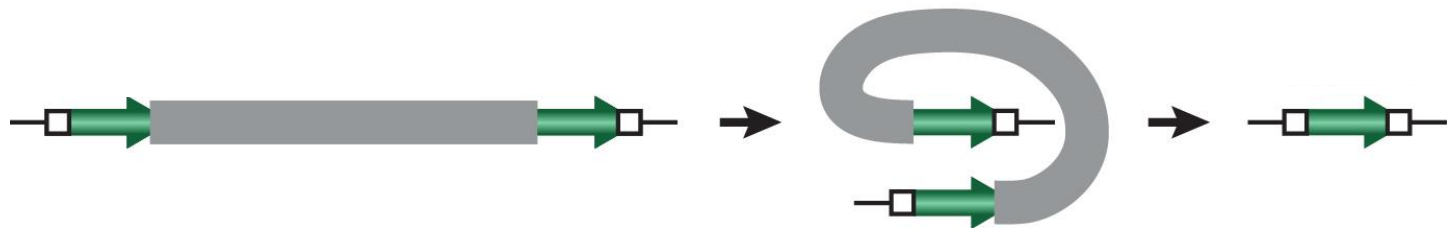
B) Covariation

Two key mechanisms of structural change

Non-homologous end joining (**NHEJ**)
(requires double-strand DNA breaks)

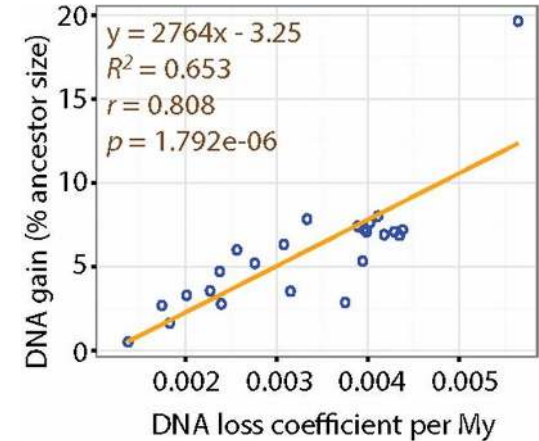
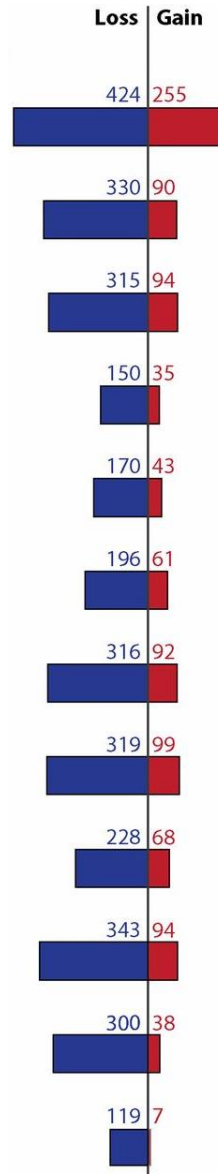
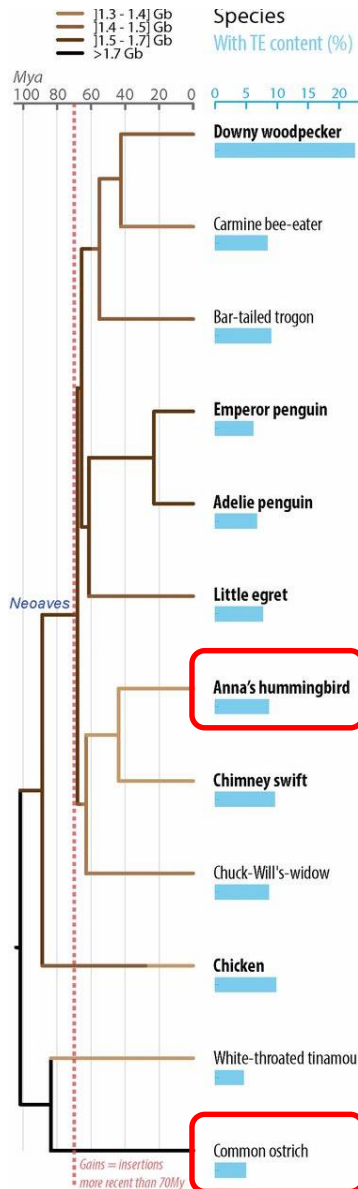


Non-allelic homologous recombination (**NAHR**)
(requires sequence homology)



NHEJ correlates with frequency of DNA damage, NAHR correlates with frequency of (identical, large) repeats

Genome shrinking despite more TEs



Accordian model



**Consider not only
host popgen, but
also TE popgen!**

Genome size and life history traits



Dynamic genome
(more TEs, fast shrinking)



Static genome
(fewer TEs, slow shrinking)

Adaptive processes are often invoked but remain difficult to prove
(few high-quality genome assemblies and lack of popgen data)!



20 Gb



32 Gb

3.2 Gb

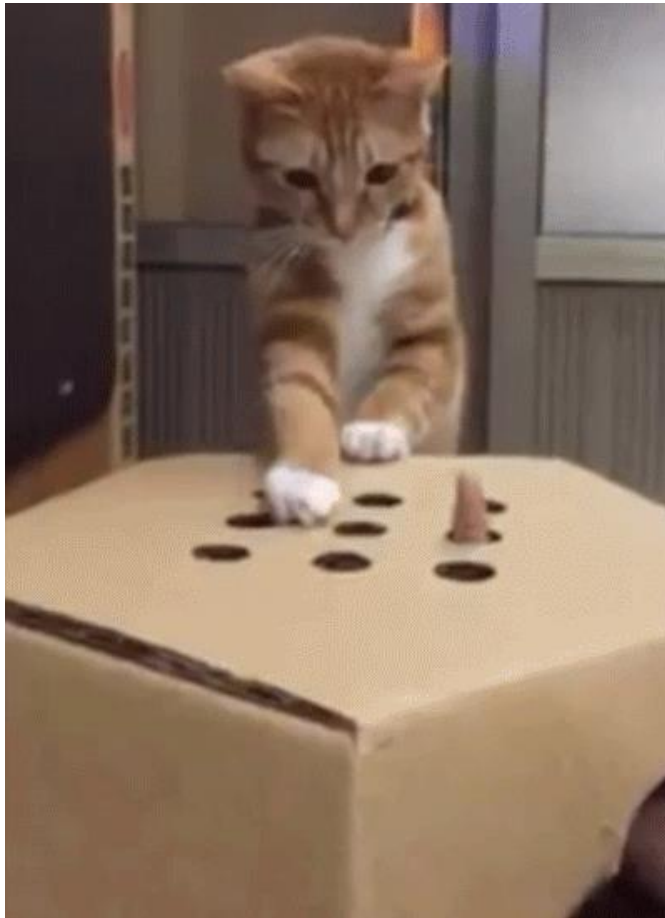
133 Gb



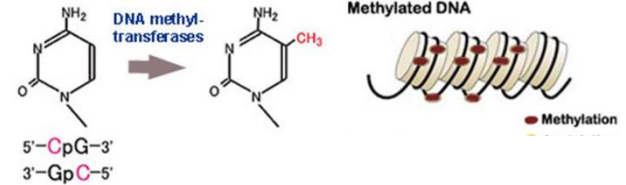
More
context in
[Suh 2021](#)
[TE](#)
[lecture 5](#)

Non-adaptive processes likely contribute to a large or very large degree!

Genomes: whack-a-transposon



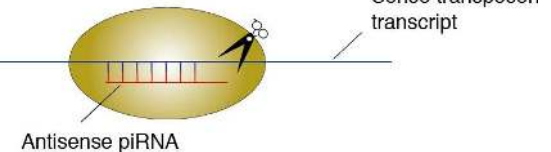
DNA methylation



<http://helicase.pbworks.com/w/page/17605615/DNA%20Methylation>

piRNA pathway

PIWI and/or AUB



<http://ruo.mbl.co.jp/bio/g/product/epigenetics/RNAworld.html>

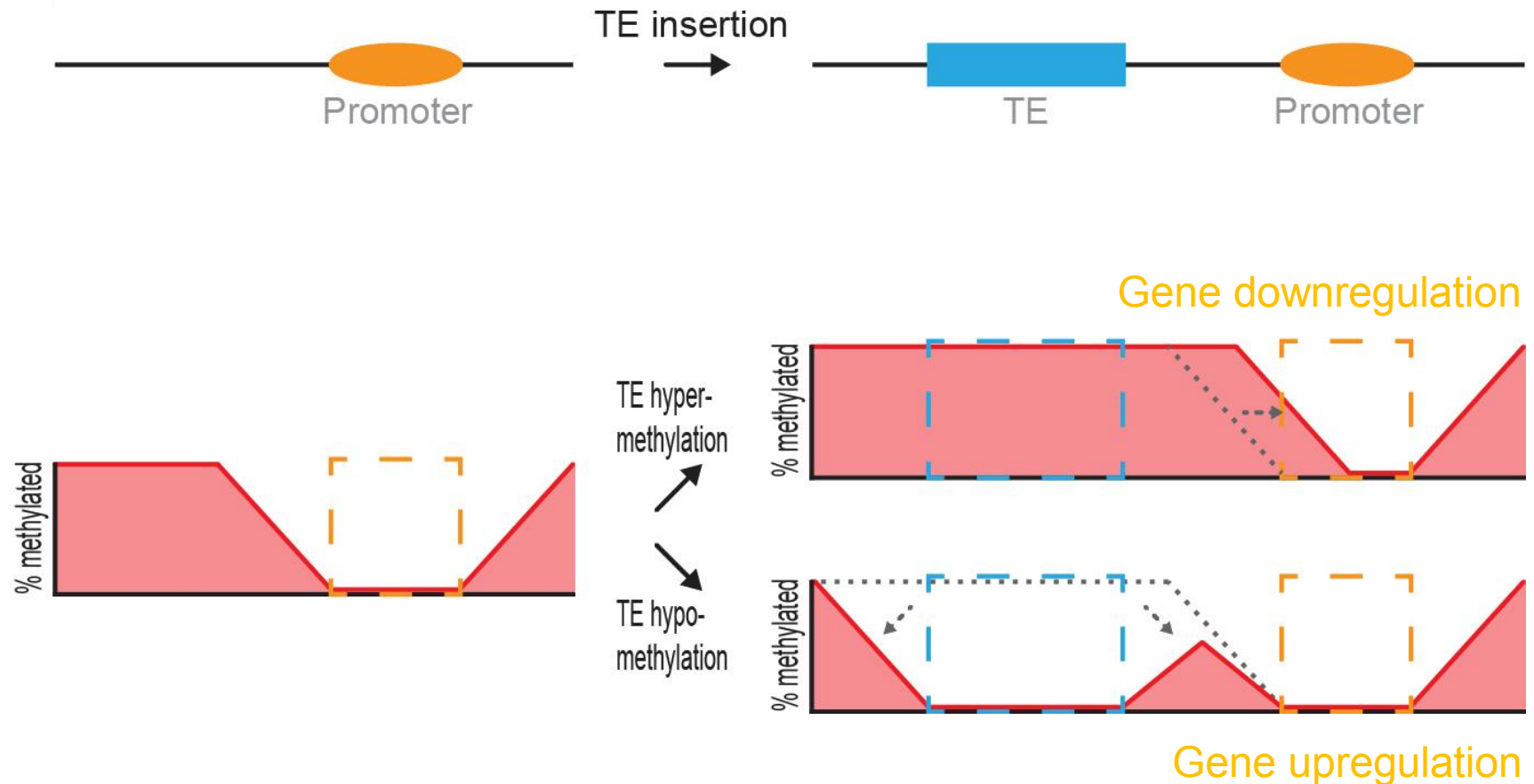
KRAB zinc-finger genes



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

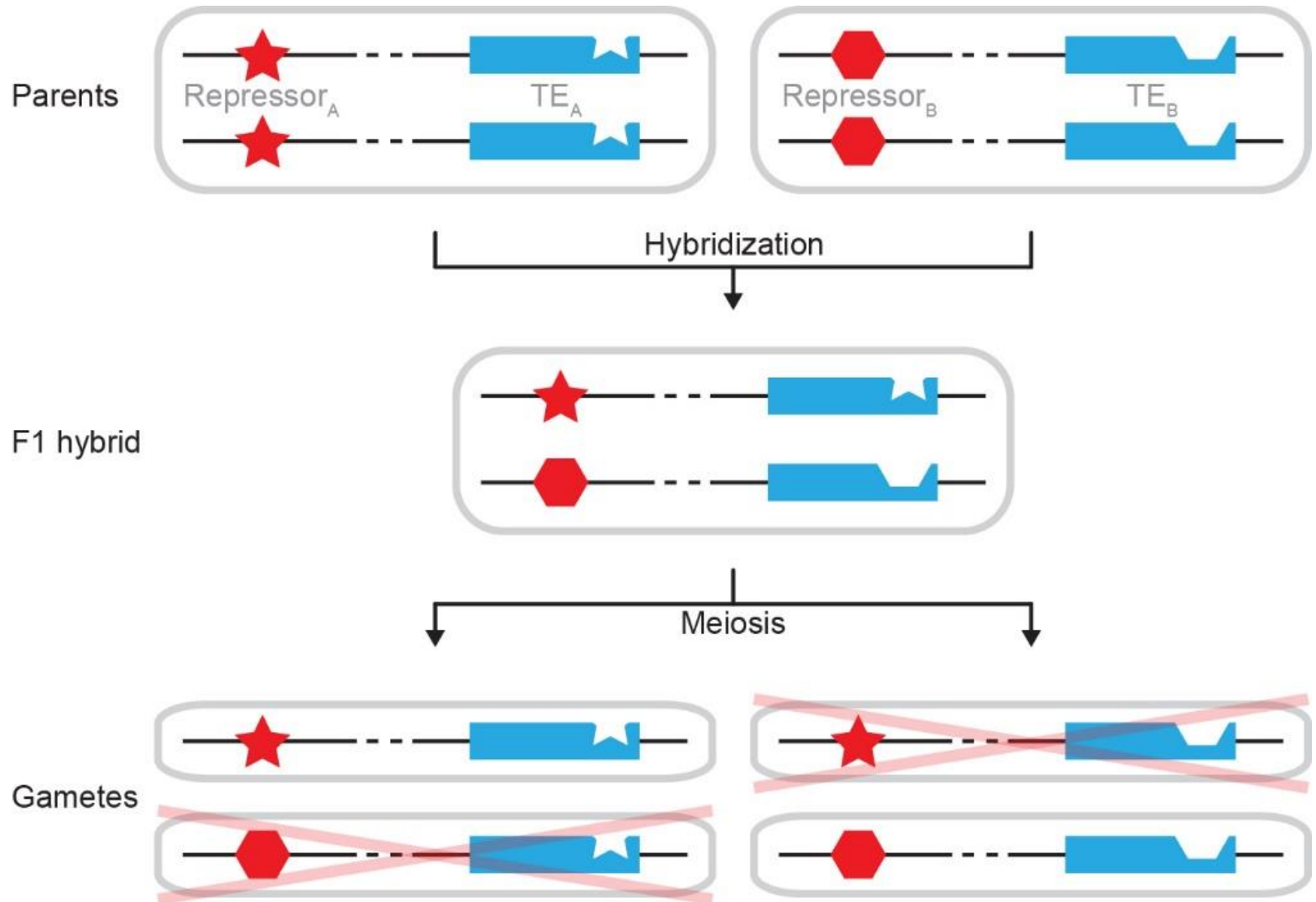
More context in [Suh 2021 TE lecture 6](#)

Covariation between (epi)mutation types

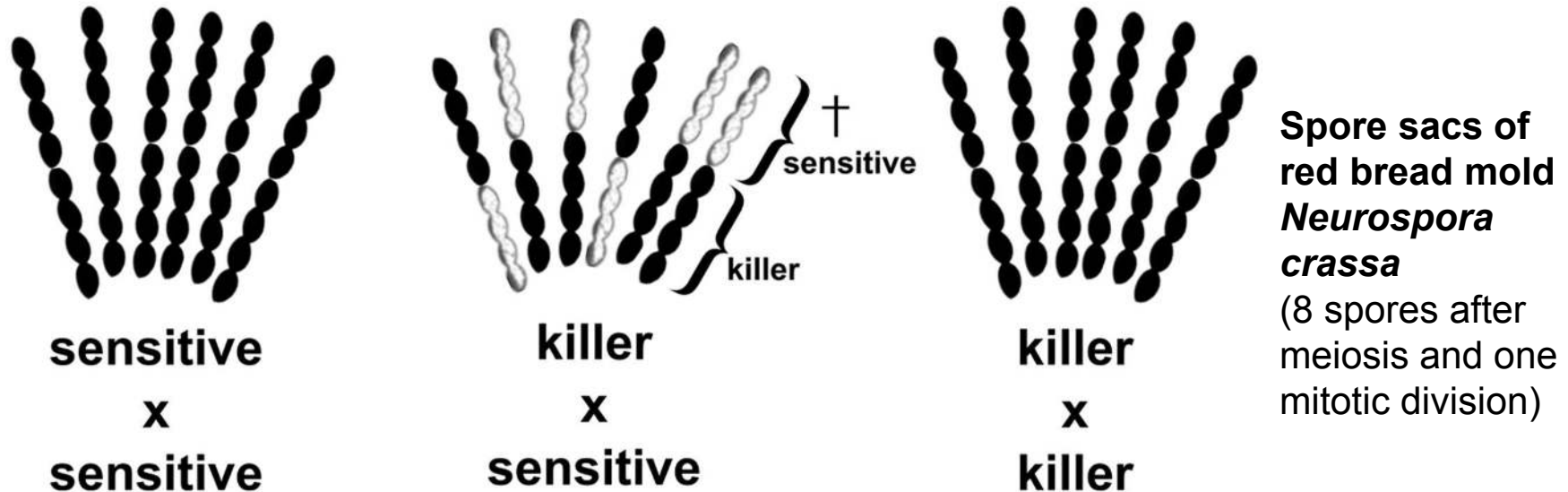


Spillover of DNA methylation and/or histone modifications from new TE insertions to nearby genes!

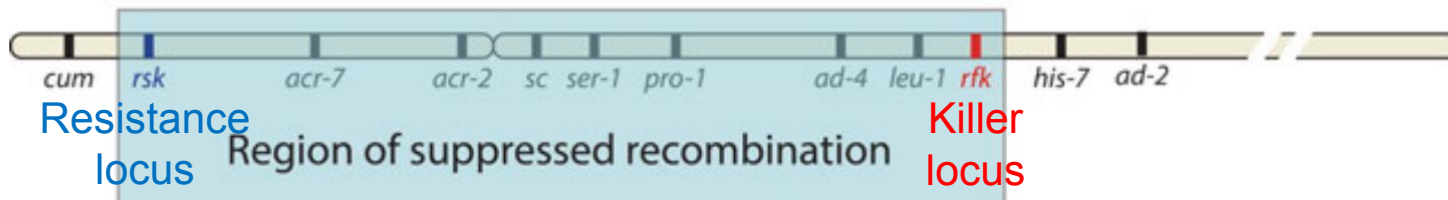
Host-TE conflict and reproductive isolation



Spore/sperm killing of some SVs

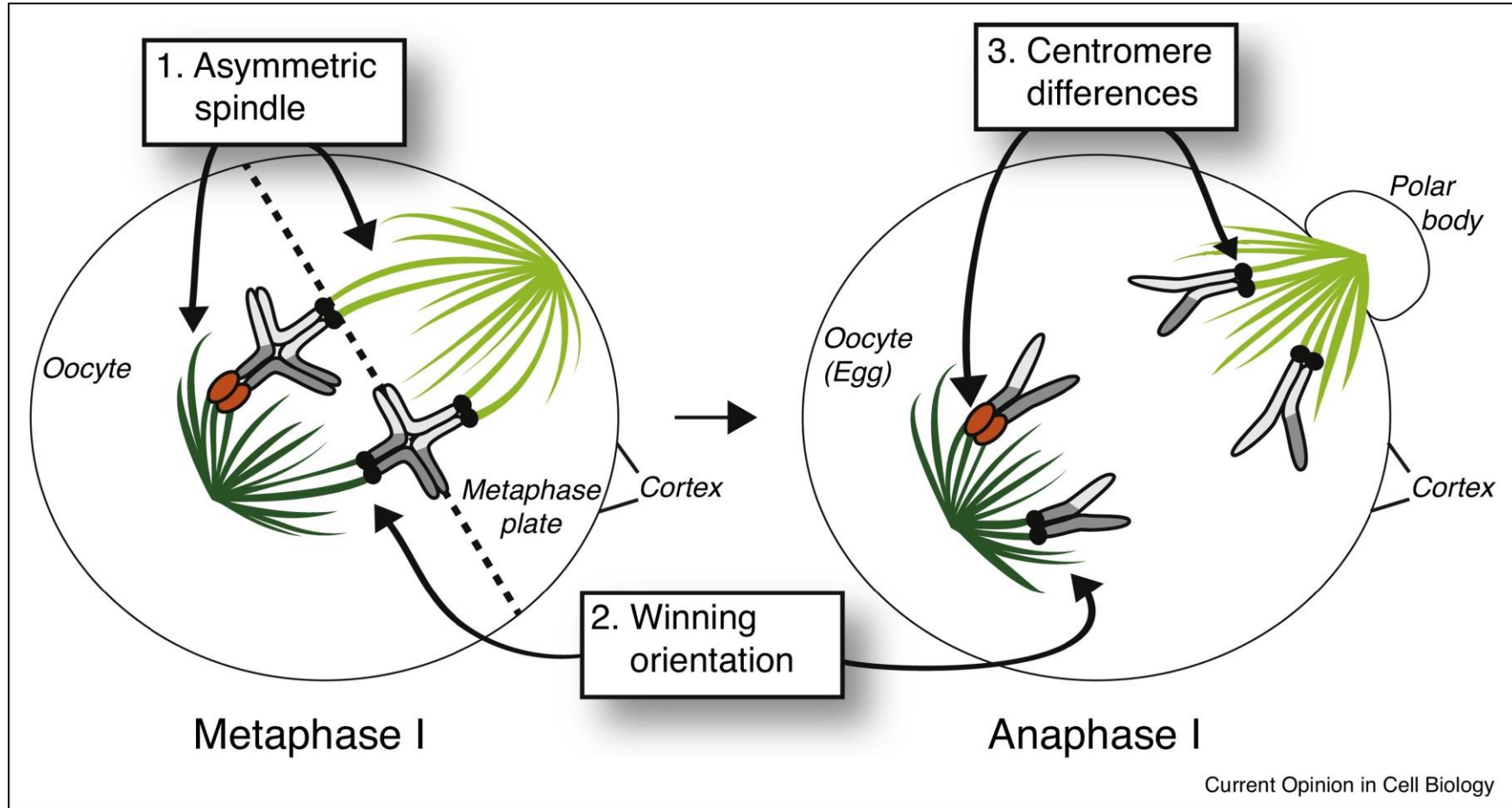


Chromosome 3



If an inversion or duplication leads to gene truncations, a toxin/antitoxin system can evolve to distort its transmission!

Centromere drive of some SVs



If a pericentric inversion or a centromere shift leads to a stronger centromere, it can distort its own transmission!

Questions?



Part 2: Frustration



A) Concepts and methods

What this lecture will not cover

1. Genome assembly: What is (not) assembled?

Primers: [Peona et al. 2018](#), [Peona et al. 2021](#), [Rhie et al. 2021](#), [Nurk et al. 2022](#)

2. Gene and repeat annotation: What is (not) annotated?

Primers: [Yandell & Ence 2012](#), [Suh 2021 TE lecture 4](#), [Goubert et al. 2022](#)

3. Within-individual or germline/soma genome differences

Primers: [Smith et al. 2021](#), [Suh & Dion-Côté 2021](#), [Borodin et al. 2022](#)

4. All SVs, all processes, all effects, all methods, all limitations. Talk to Valentina, Alexander Leonard, and me!



Valentina Peona

9a – 12p	Alex Suh	Structural Variation
2p – 5p	Valentina Peona	Structural Variation Activity
7p – 10p	Alexander Leonard	Pangenomics



Alexander Leonard

Awareness of biology and technology



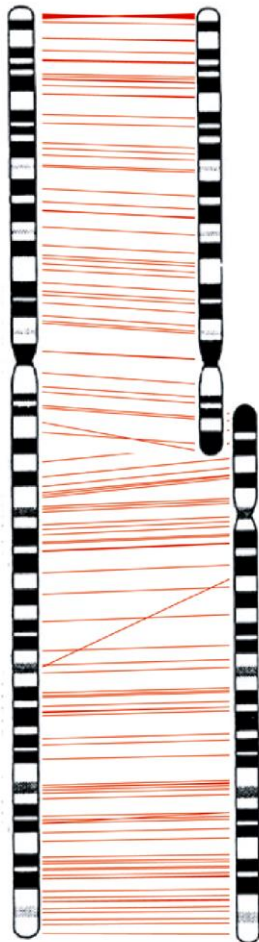
How can we make sure that what we see in our data is what we think it is?

Did we account for biological patterns/processes and technological limitations?

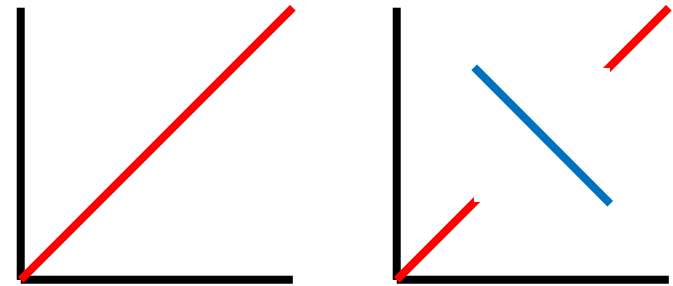
Terminology

Synteny vs. collinearity

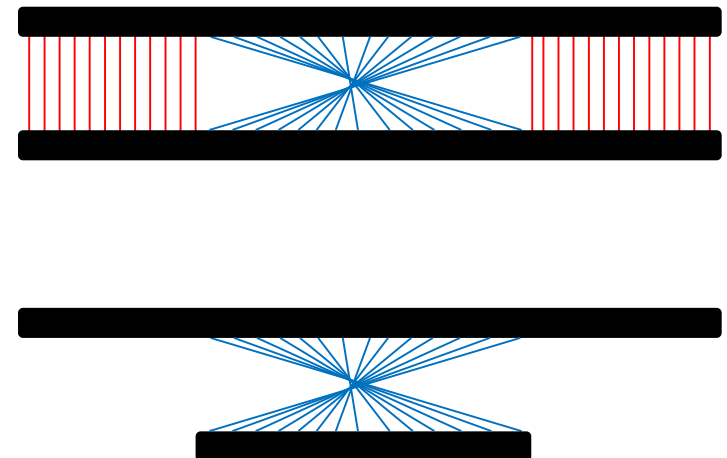
Hs2 Pt12/13



Dot plot



Pattern vs. process



Beware of waves

My SNP
explains
everything!

My inversion
explains
everything!

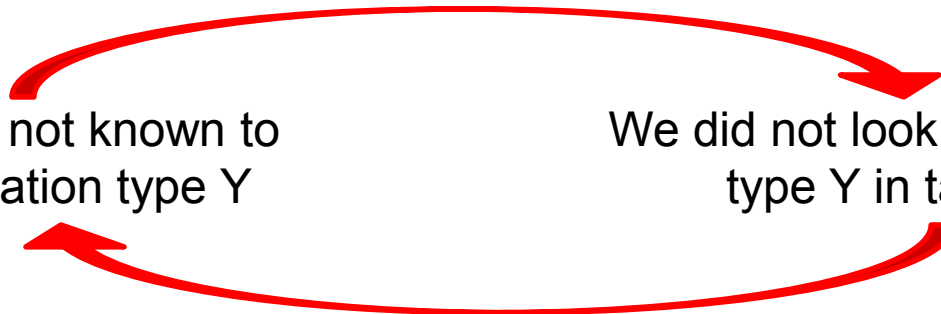
My TE
explains
everything!



Each of these statements can be true, but what if there is covariation with other mutation types?

Taxon X is not known to
have mutation type Y

We did not look for mutation
type Y in taxon X

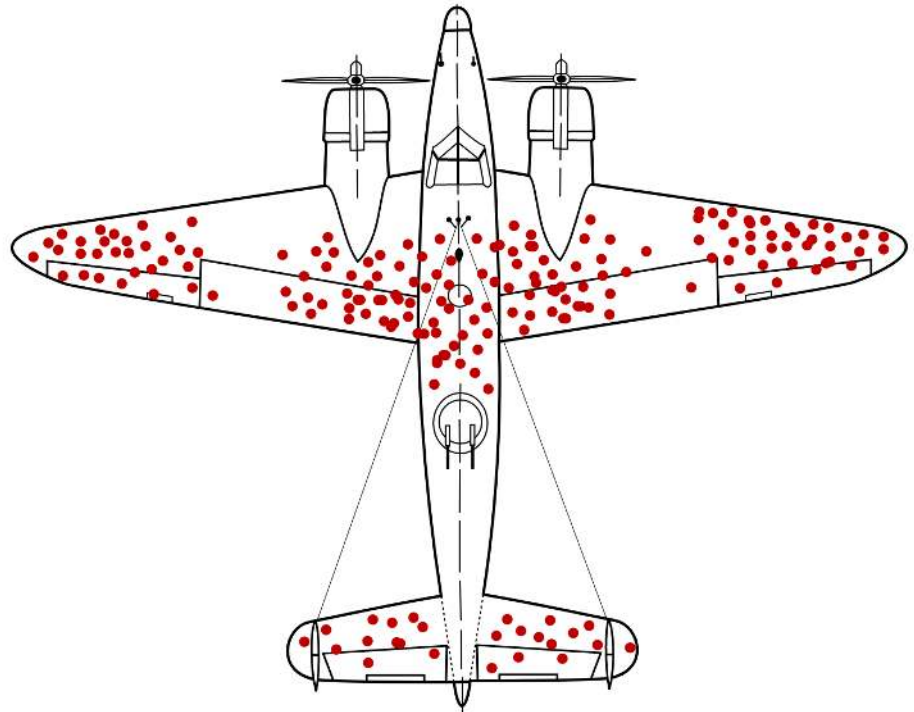


Reflection on biases

Confirmation
bias



Survivorship
bias



My own biases: I like transposable elements, centromere shifts, and simple (but unexpected) answers to complicated questions!

Ultimate vs. proximate causes

Proximate: This TE is beneficial for the host

Ultimate: ~~TEs jump to be beneficial for the host~~

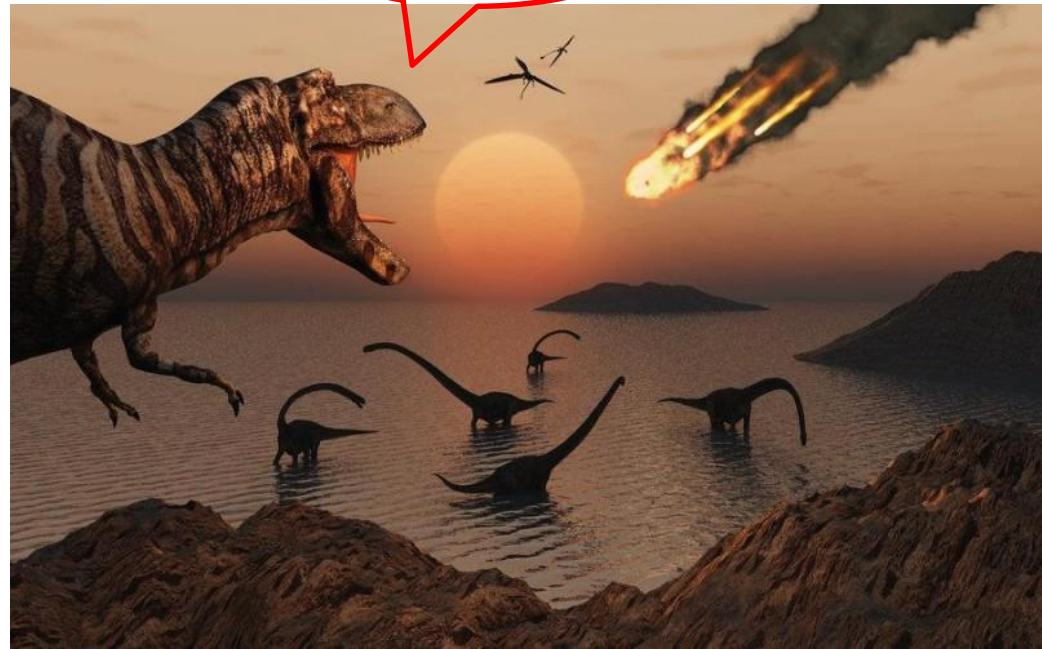
TEs jump because they can

ROAR

Proximate: This asteroid caused diversification

Ultimate: ~~Asteroids land to cause diversification~~

Asteroids land eventually



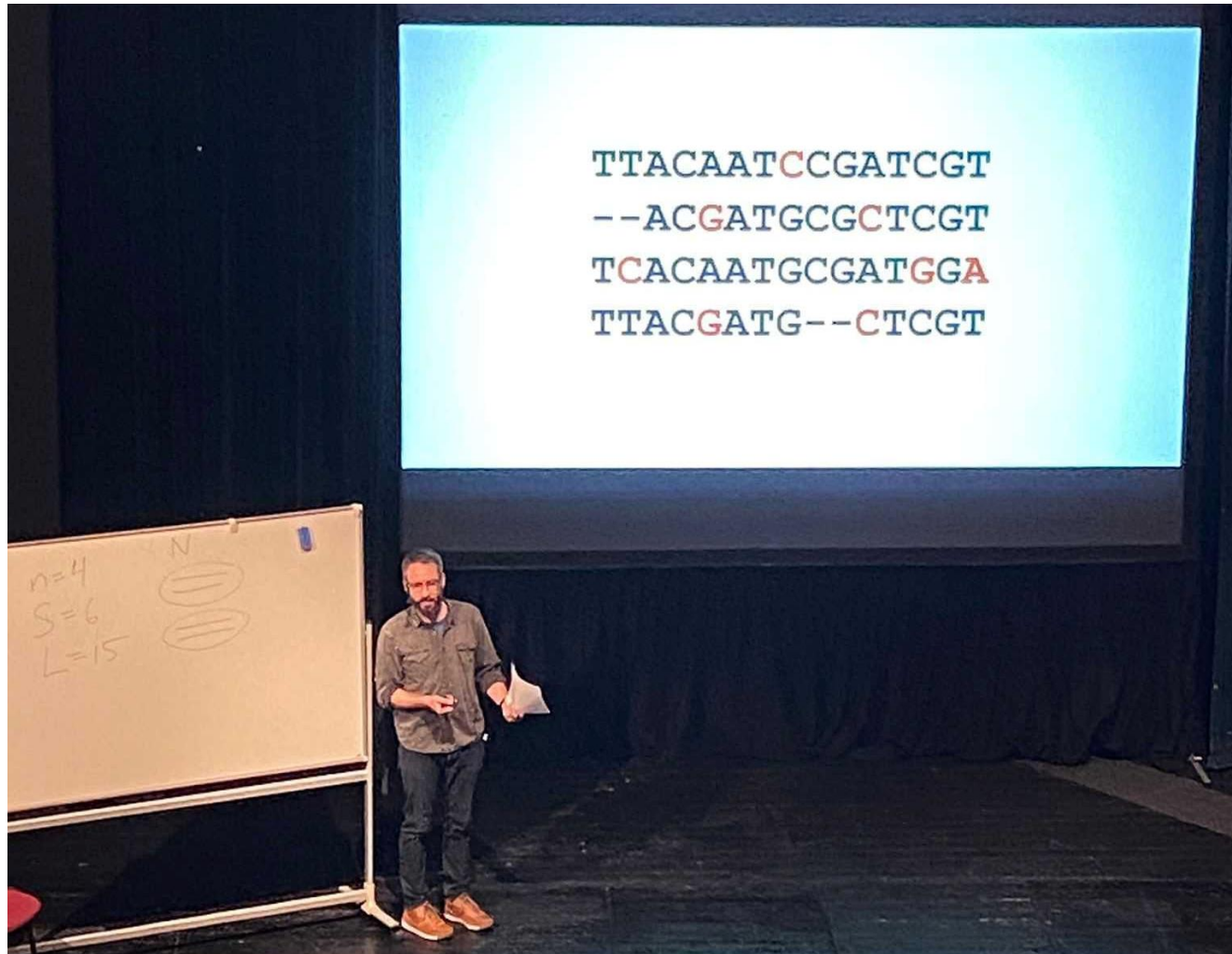
What is the null hypothesis?

~~Guilty until proven innocent~~
Innocent until proven guilty

~~Absence of evidence~~
Evidence of absence

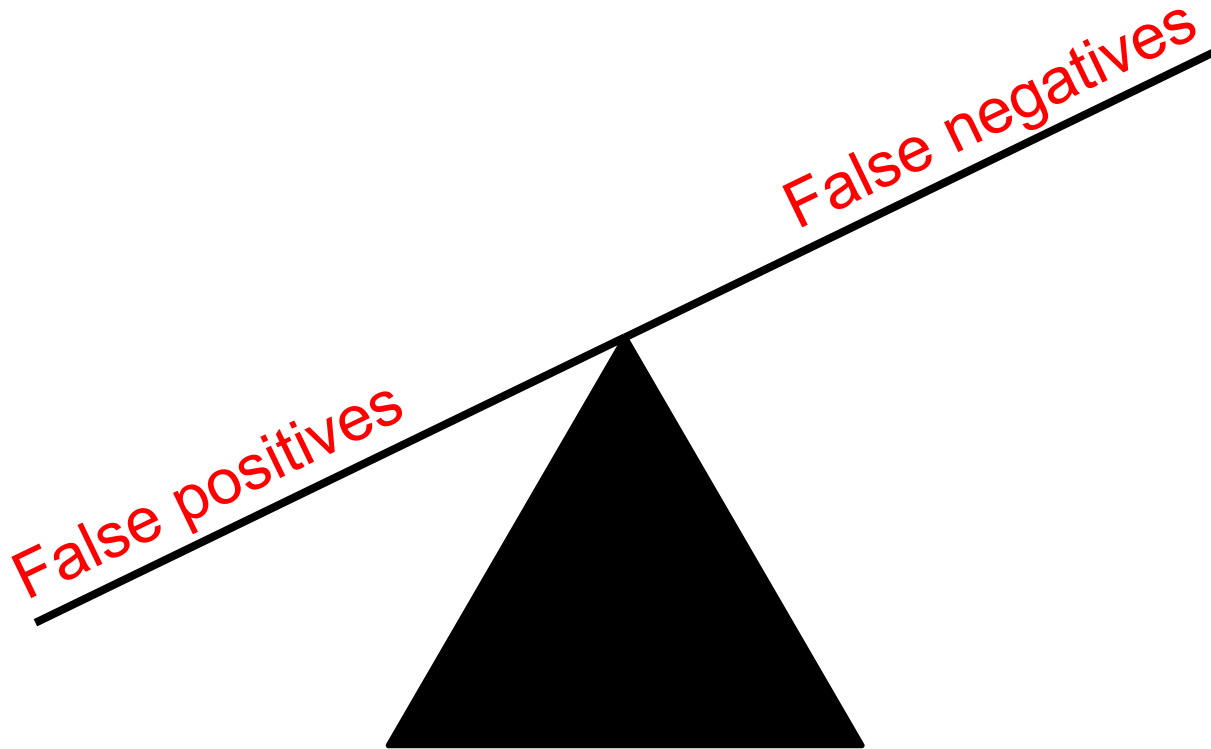


Theory applies to SNPs and to SVs



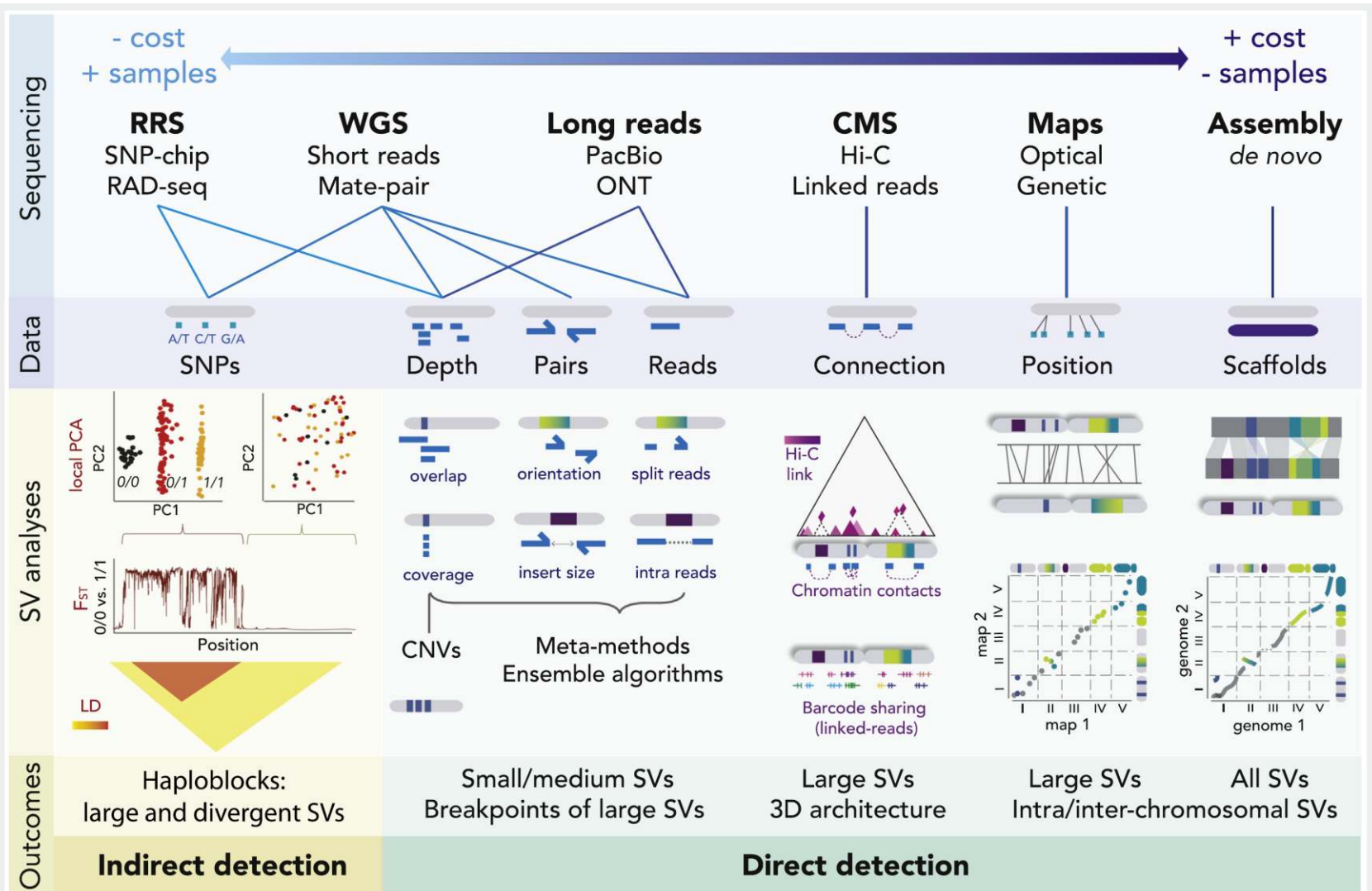
Selection vs. background variation: What SNPs and SVs are there?

SVs are nowhere as established as SNPs



Problem: Reliable SV genotyping (cf. SNP activities in this workshop) + accounting for covariation with other SVs (cf. this lecture) is essential but the SV field is not there yet.

One approach to find them all?



How to pick a tool for finding SVs?

Repeat tools

Description

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

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 are welcome. Please feel free to add missing tools into the table, that would help a lot!

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](#).



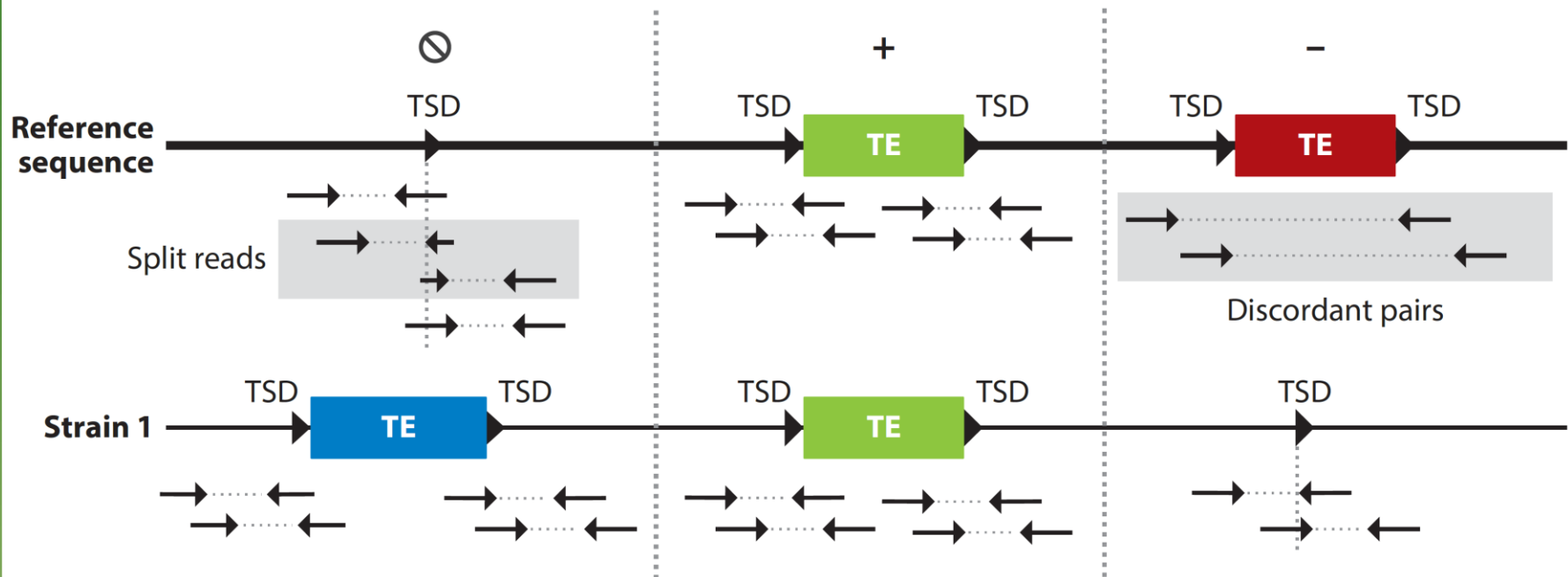
Overview of tools for repeat analysis

Tool↓Find...	DOI↓Find...	Alternate URL↓Find...	Keywords↓Polymorphism
AluMine	https://doi.org/10.1101/588434		Alu, SINE, Genotype, Polymorphism, NGS/HTS
alu-detect	https://doi.org/10.1093/nar/gkt612		Alu, SINE, Genotype, Polymorphism, NGS/HTS, Paired-End

100 tools listed for TE insertion polymorphism analysis!

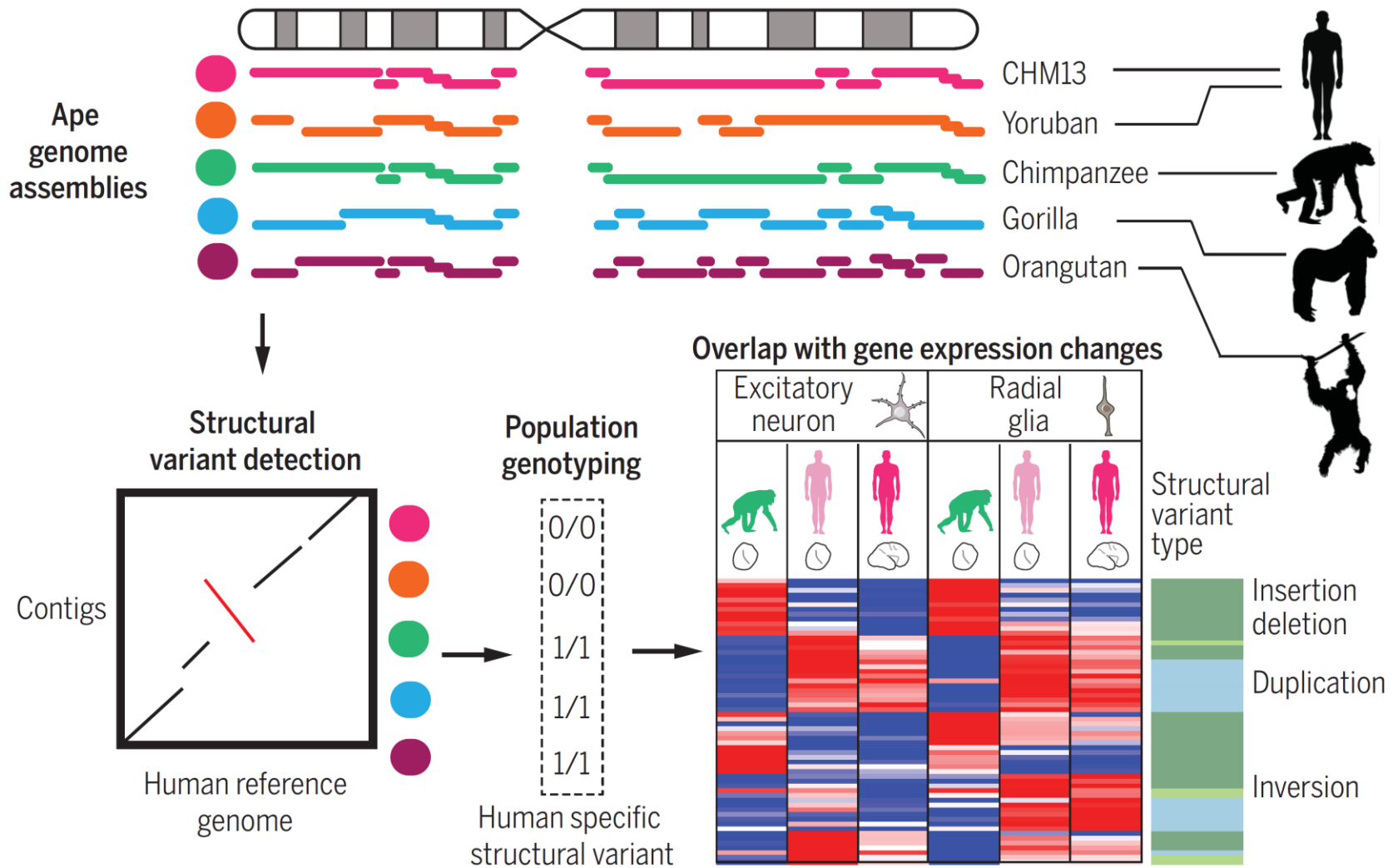
https://tehub.org/en/resources/repeat_tools; [The TE Hub Consortium 2021, Mobile DNA](#)

Read-based SV detection



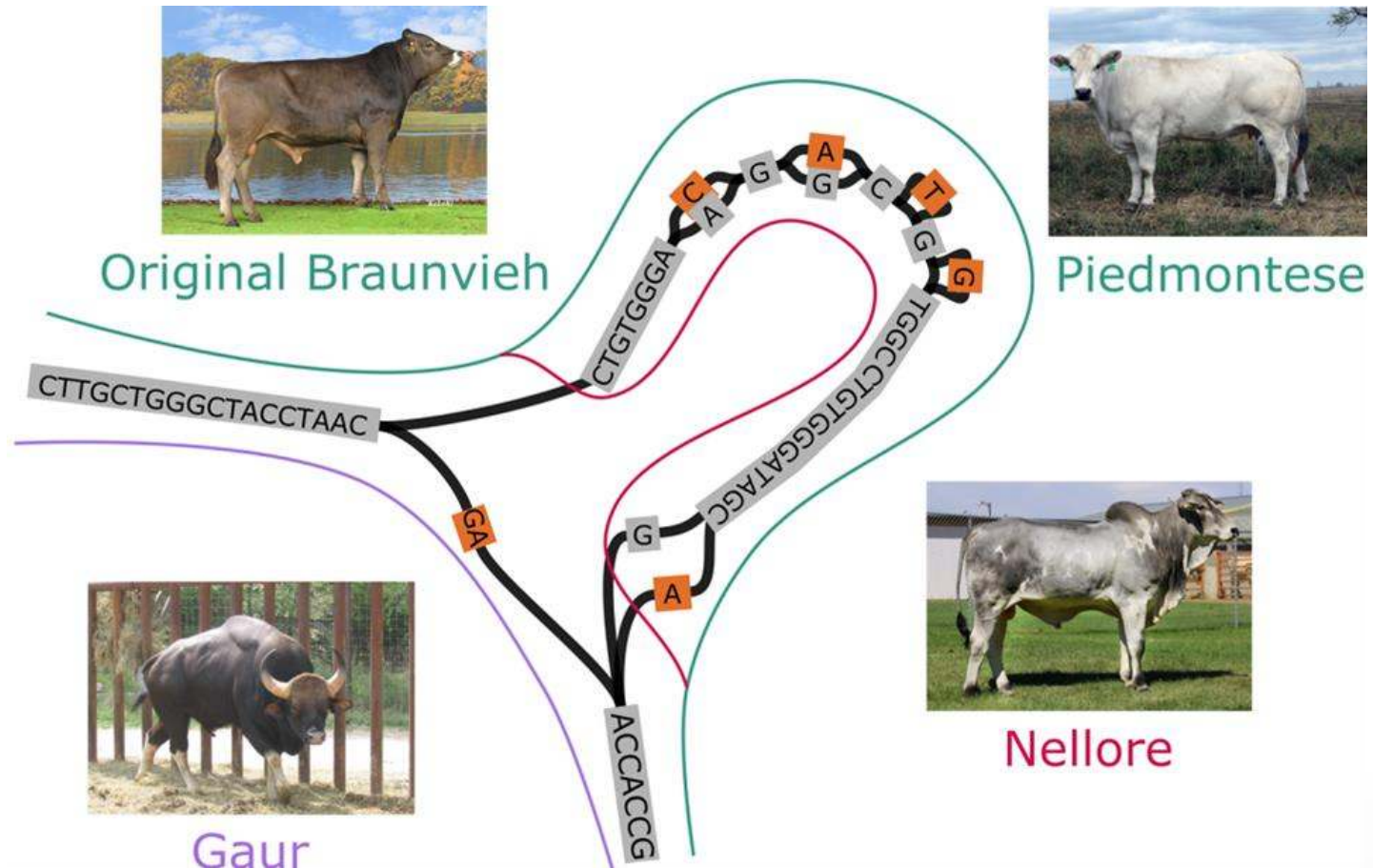
Reliable read mapping and SV scoring is difficult near (other) repeats, near gaps, at misassemblies ...

Assembly-based SV detection



Reliable genome alignment and SV scoring is difficult in highly repetitive regions (if assembled ...)

Graph-based SV detection (pangenomics)



7p – 10p

Alexander Leonard

Pangenomics

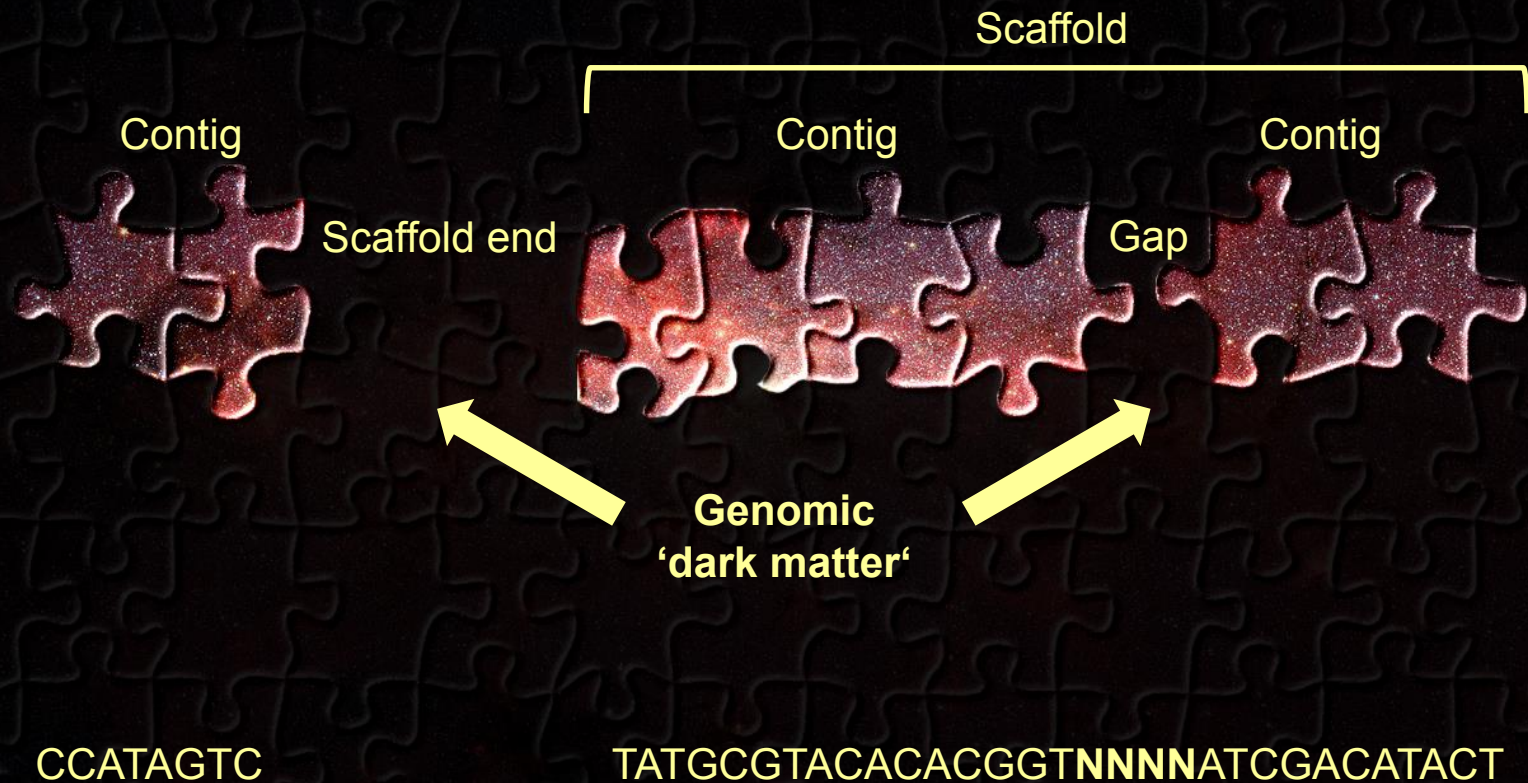


Alexander Leonard

It could all be so easy
(if it wasn't for technological limitations)



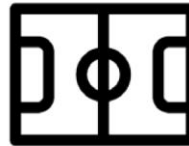
Genomics: a big and messy puzzle



Various sequencing technologies



Distance Rome-Paris
(avian genome)
1,100,000,000 bp



Football field
(OM, LRC, Hi-C)
150,000 bp



Autobus
(long reads)
15,000 bp



Smartphone
(short reads)
150 bp

Input DNA



Short reads



Long reads



Linked reads



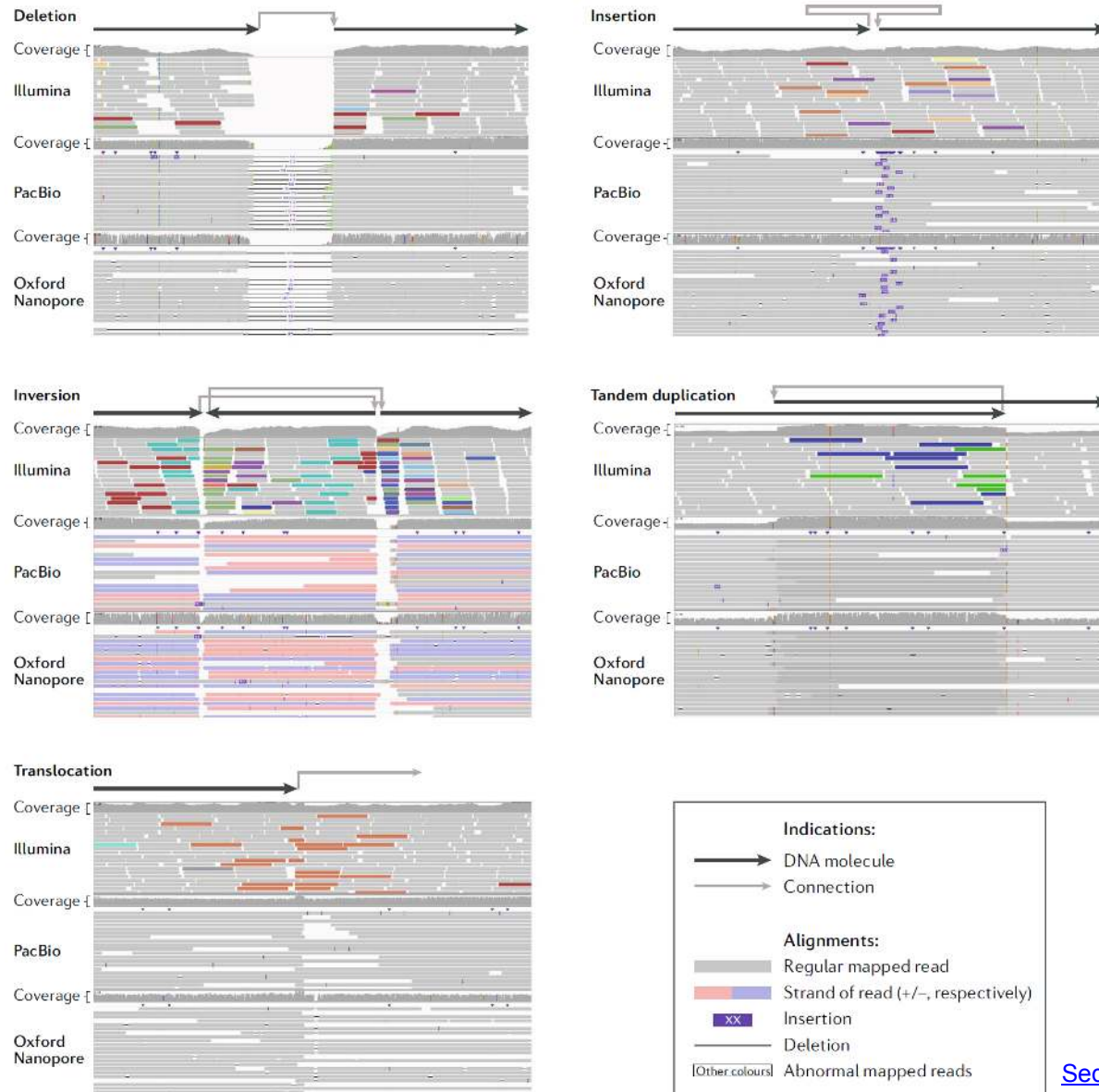
Optical maps



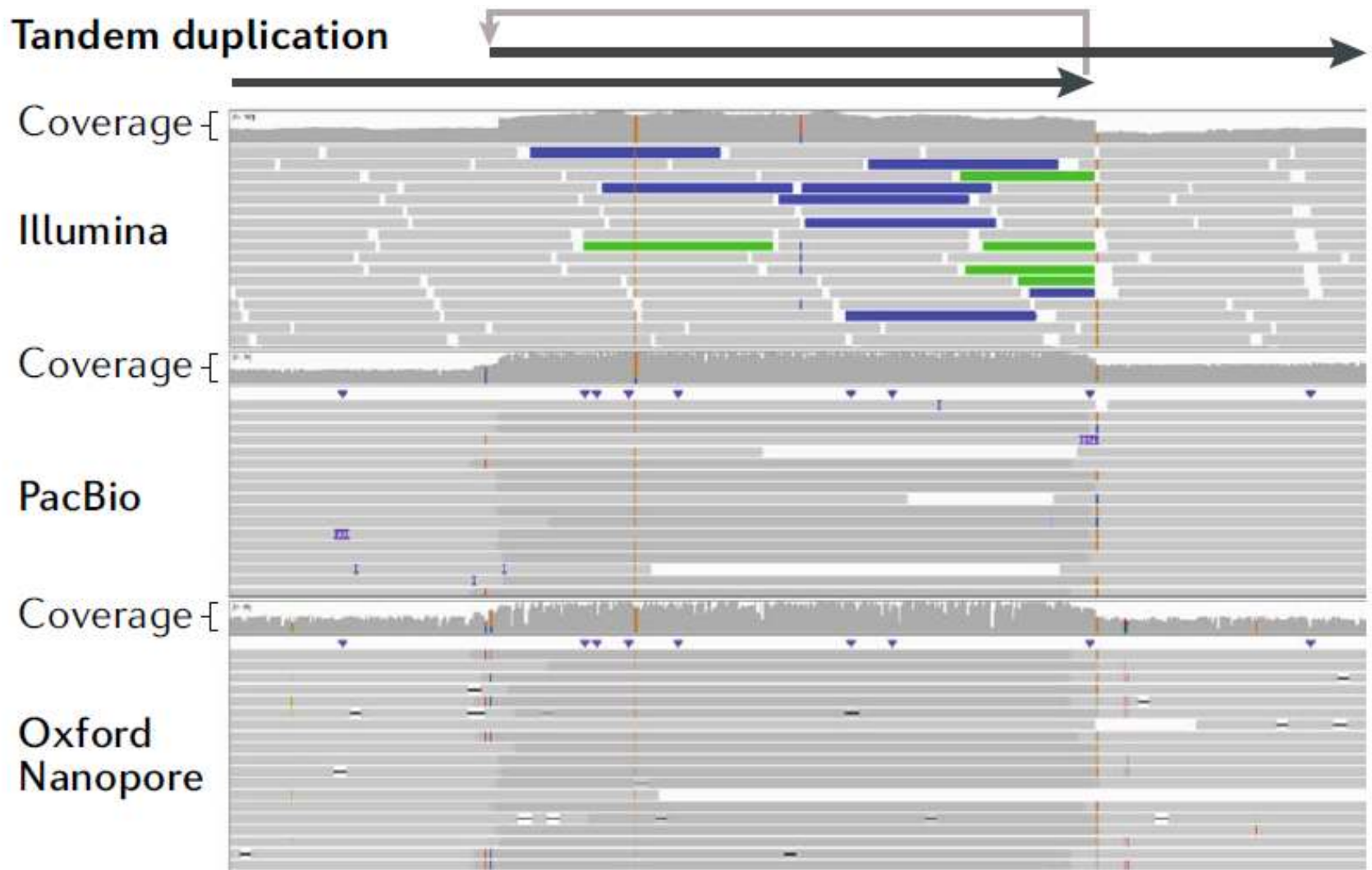
Hi-C maps



SV mapping with longer and longer reads



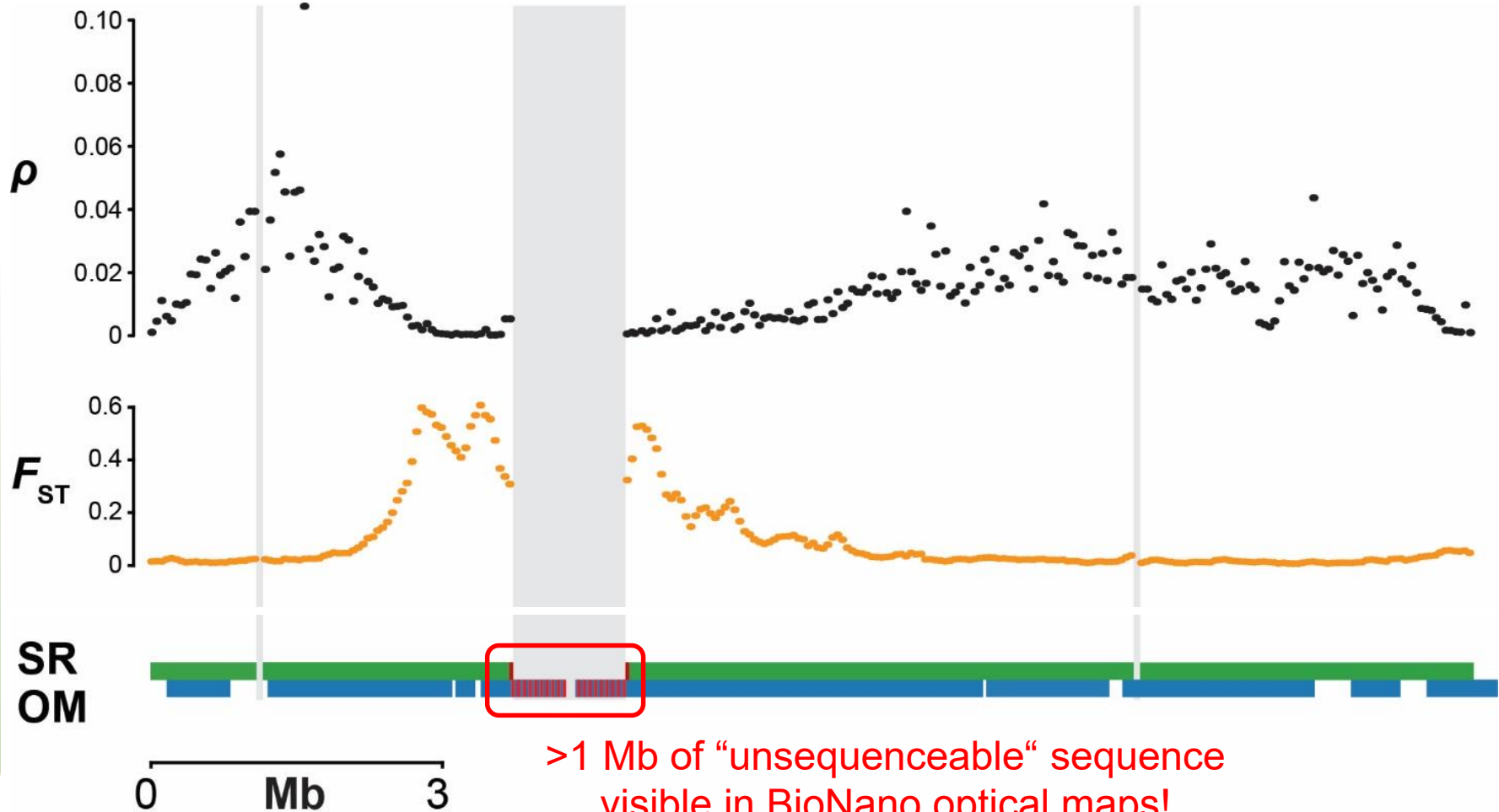
What does coverage variation tell us?



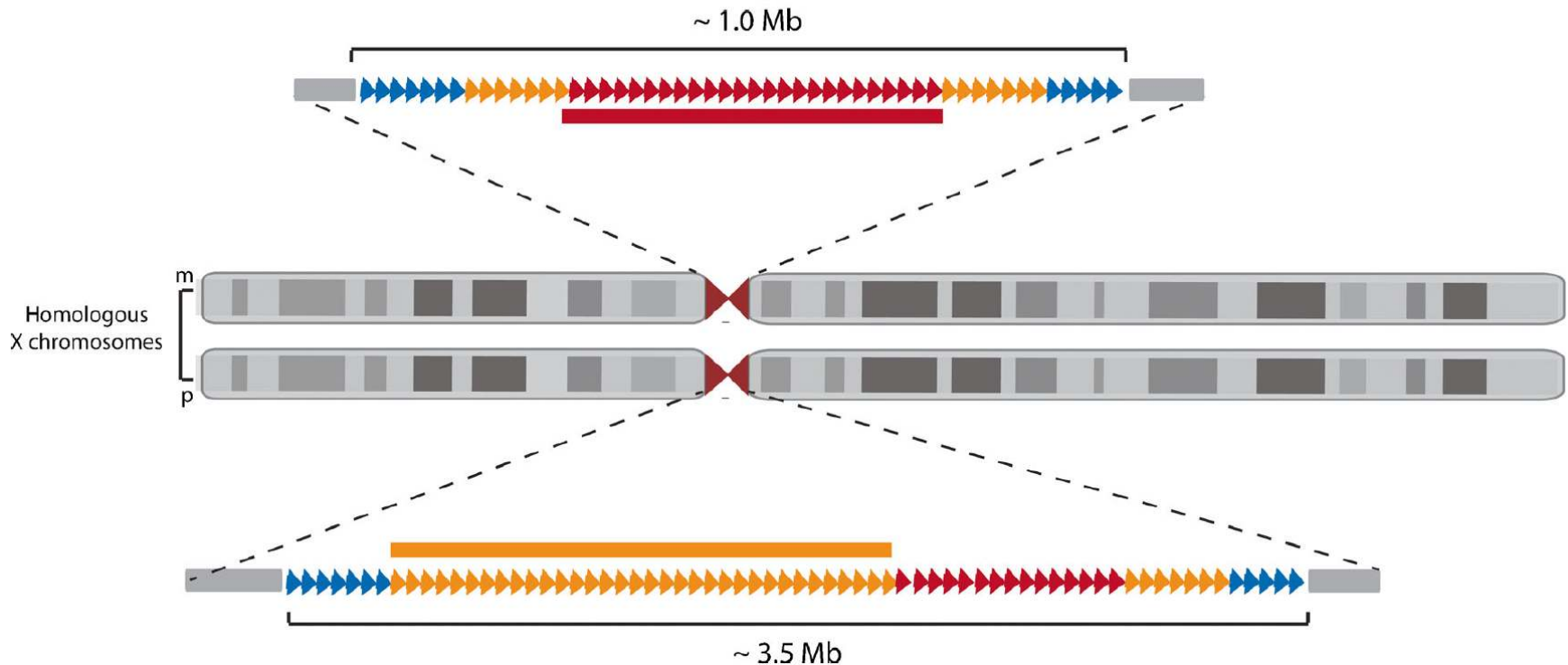
Tandem duplications are (usually) collapsed in assemblies!

Not all gaps are equal

Chromosome 18 of hooded/carrion crow

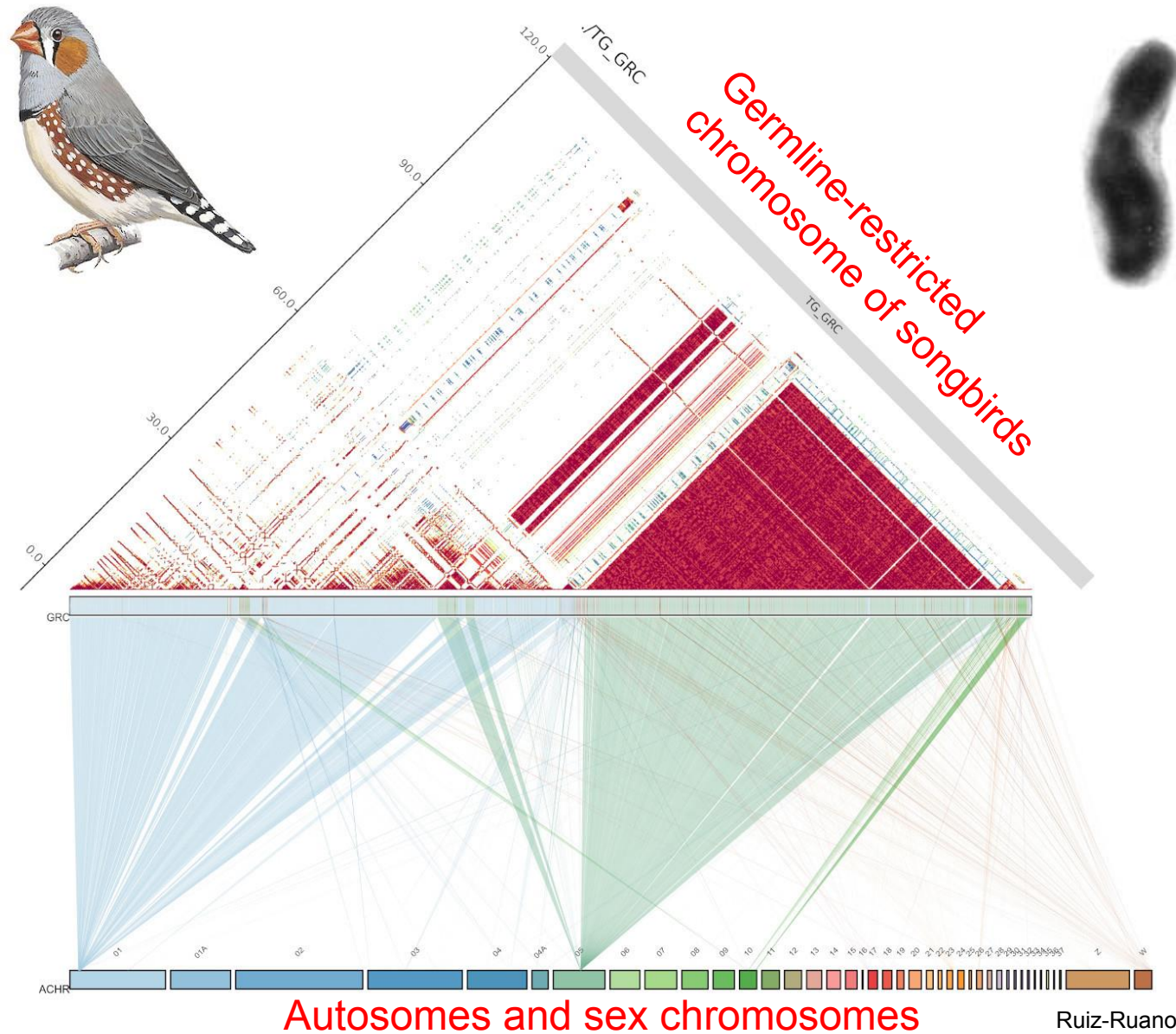


Centromeres are very, very repetitive ...



Rule of thumb: centromeres are not *in* assemblies
but in gaps within or between scaffolds!

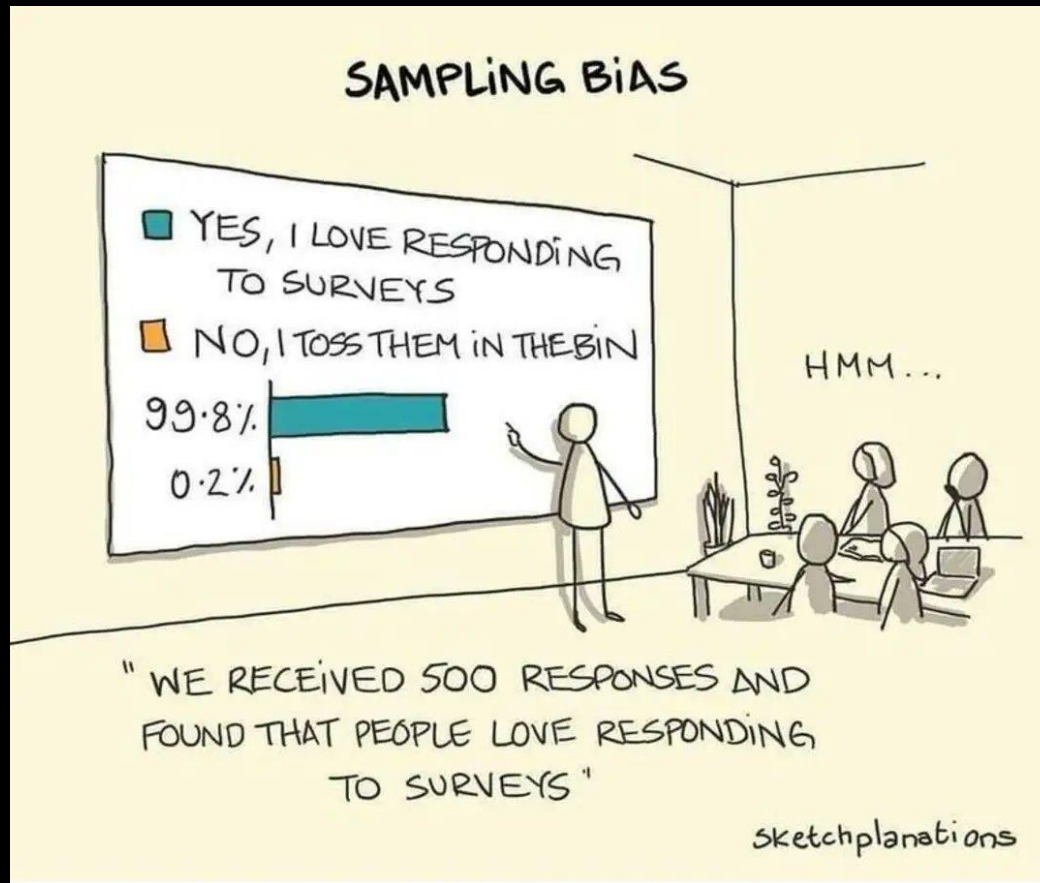
... and so are some chromosomes



Questions?



Coffee break (20 minutes)



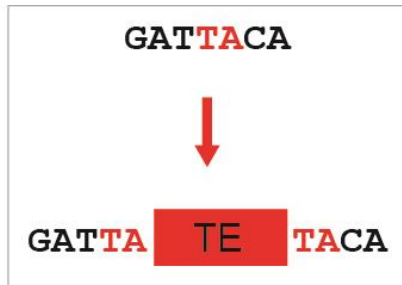
Task: Form random groups of 3 and discuss 1) what SVs you want to study, 2) what SVs you can study, and 3) what data you need to be less frustrated.

Part 2: Frustration



B) Biology and more concepts

Transposable elements are very diverse



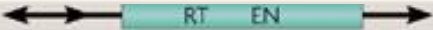
Classification	Structure	TSD	Code	Occurrence	
Order	Superfamily				
Class I (retrotransposons)					
LTR	Copia		4-6	RLC	P, M, F, O
	Gypsy		4-6	RLG	P, M, F, O
	Bel-Pao		4-6	RLB	M
	Retrovirus		4-6	RLR	M
	ERV		4-6	RLE	M
DIRS	DIRS		0	RYD	P, M, F, O
	Ngaro		0	RYN	M, F
	VIPER		0	RYV	O
PLE	Penelope		Variable	RPP	P, M, F, O
LINE	R2		Variable	RIR	M
	RTE		Variable	RIT	M
	Jockey		Variable	RIJ	M
	L1		Variable	RIL	P, M, F, O
	I		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		8	DTA	P, M, F, O
	Mutator		8-9	DTE	M, O
	Transib		8	DTP	P, M
	PiggyBac		3	DTH	P, M, F, O
	CACTA		2-3	DTC	P, M, F
Crypton	Crypton		0	DYC	F
Class II (DNA transposons) - Subclass 2					
Helitron	Helitron		0	DHH	P, M, F
Maverick	Maverick		6	DMM	M, F, O

Structural features			
→	Long terminal repeats	←	Terminal inverted repeats
—	Diagnostic feature in non-coding region	—	Non-coding region
—	Region that can contain one or more additional ORFs	—	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
ENV, Envelope protein	GAG, Capsid protein	HEL, Helicase	INT, Integrase
POL B, DNA polymerase B	RH, RNase H	RPA, Replication protein A (found only in plants)	CYP, Cysteine protease
Tase, Transposase (* with DDE motif)		YR, Tyrosine recombinase	EN, Endonuclease
Species groups			
P, Plants	M, Metazoans	F, Fungi	O, Others

Today's
focus:
LINE,
SINE,
LTR,
TIR

Weirder
TEs in
Suh 2021
TE
lecture 1

Class I: LINE retrotransposons

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

Dear RNA polymerase II,
if you read this,
transcribe me
into RNA

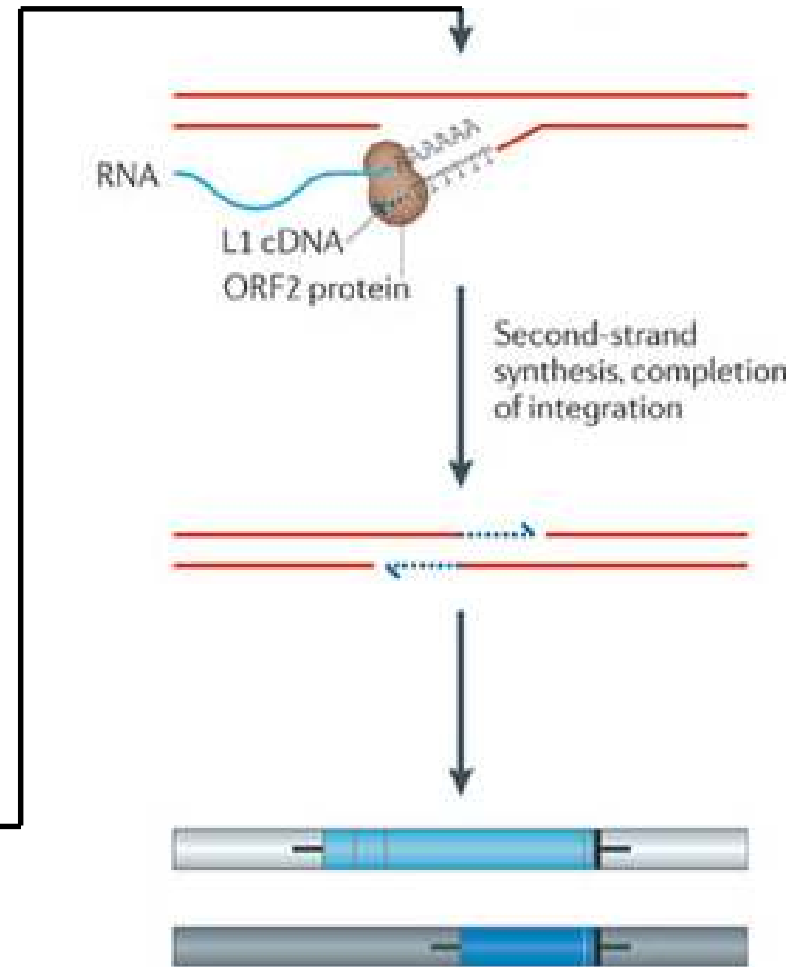
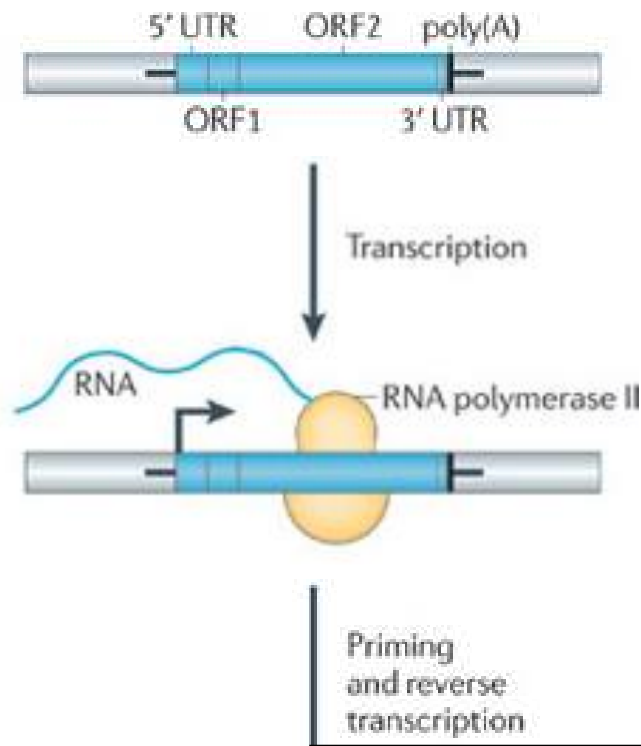
Dear ribosome,
if you read this,
translate me into a
reverse transcriptase

Dear reverse transcriptase,
if you read this,
retropose me somewhere
in the genome



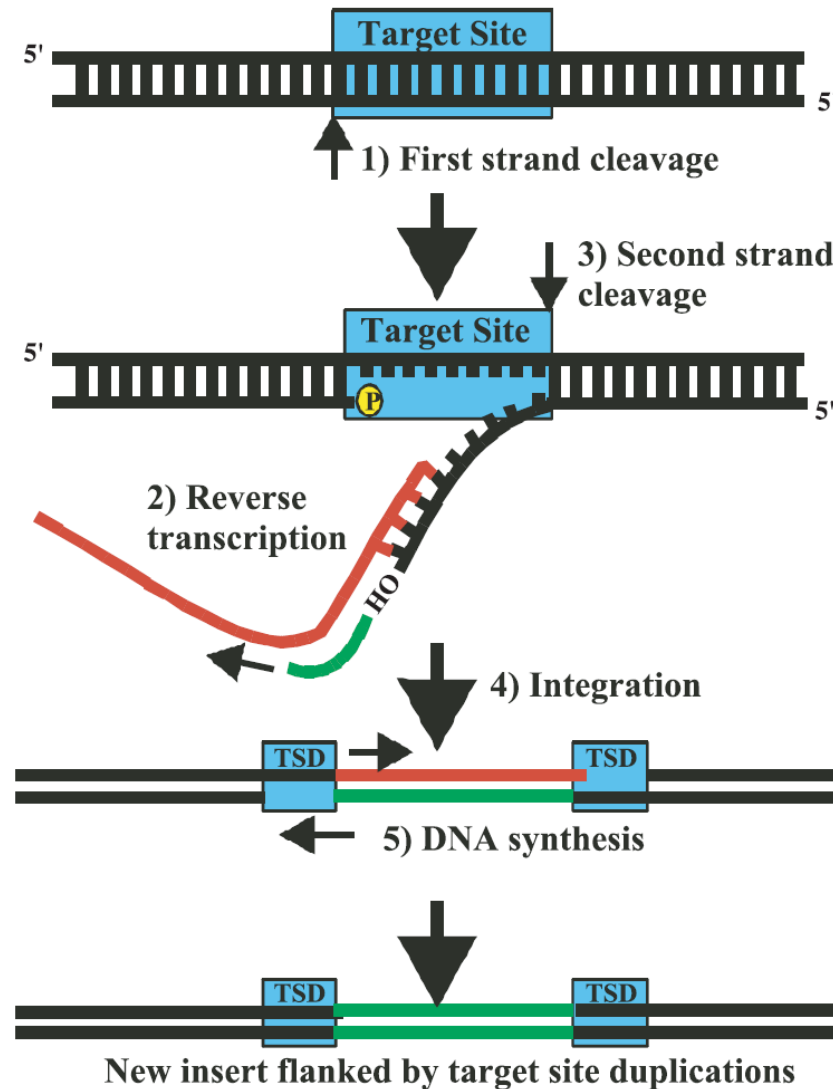
Target-primed reverse transcription (TPRT)

c Non-LTR retrotransposon
Target-site primed reverse transcription



TPRT frequently undergoes premature termination (5' truncation)

Target site duplication (TSD)

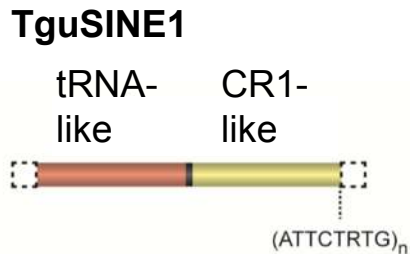
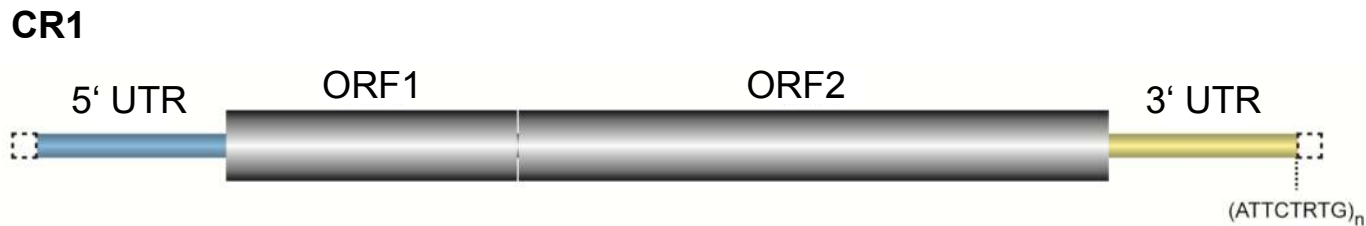


TSDs are a hallmark of nearly all (retro)transposition mechanisms!

Class I: SINE retrotransposons

Classification		Structure	TSD	Code	Occurrence
Order	Superfamily				
<i>Class I (retrotransposons)</i>					
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.



SINEs contain RNA polymerase III promoters, i.e., technically they are selfish small RNAs!

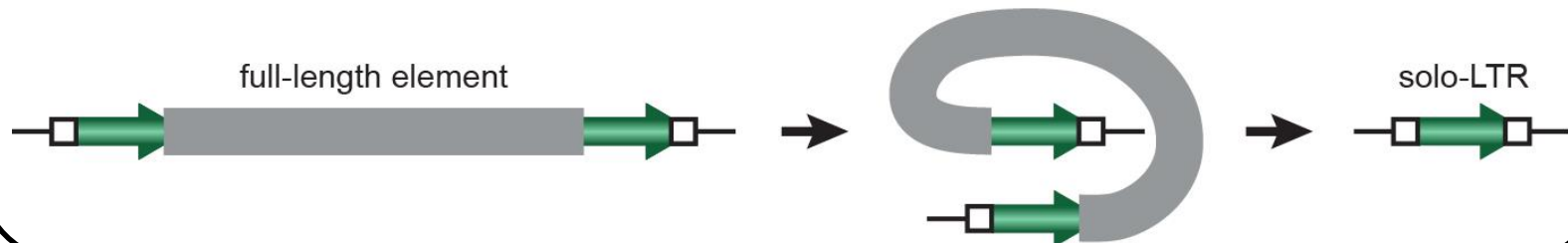
Note: In theory, any small RNA gene (pol III) can become a SINE!

Class I: LTR retrotransposons

Classification		Structure	TSD	Code	Occurrence
Order	Superfamily				
<i>Class I (retrotransposons)</i>					
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

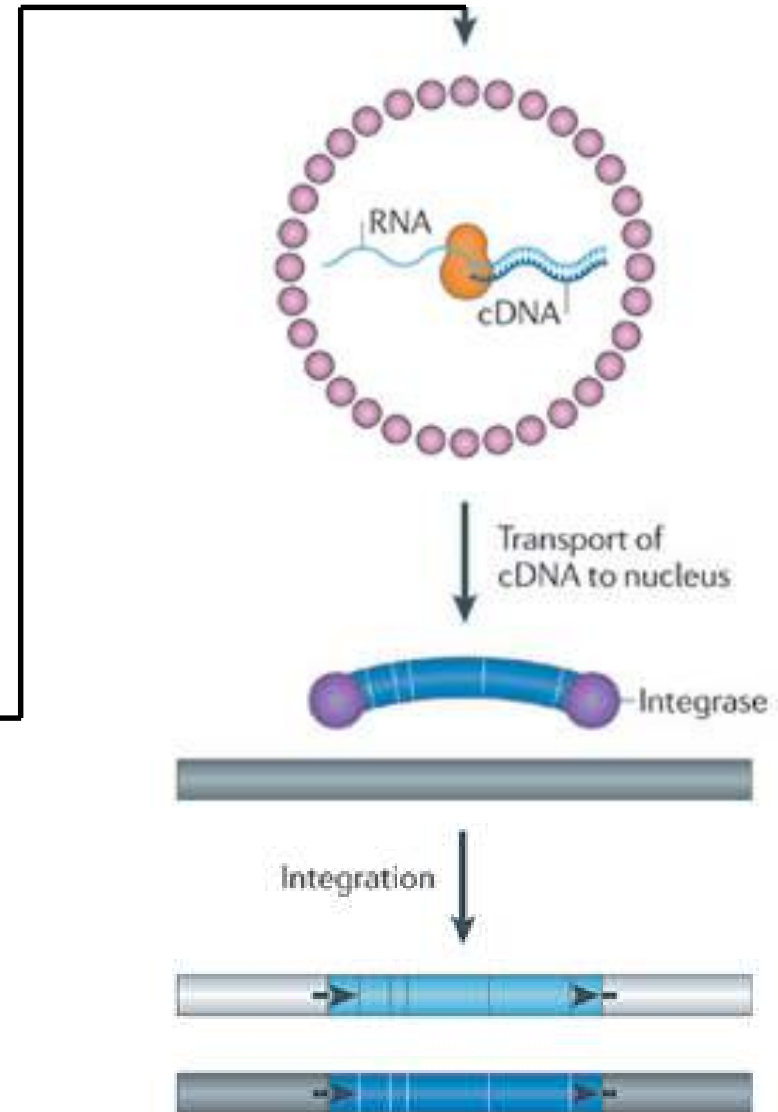
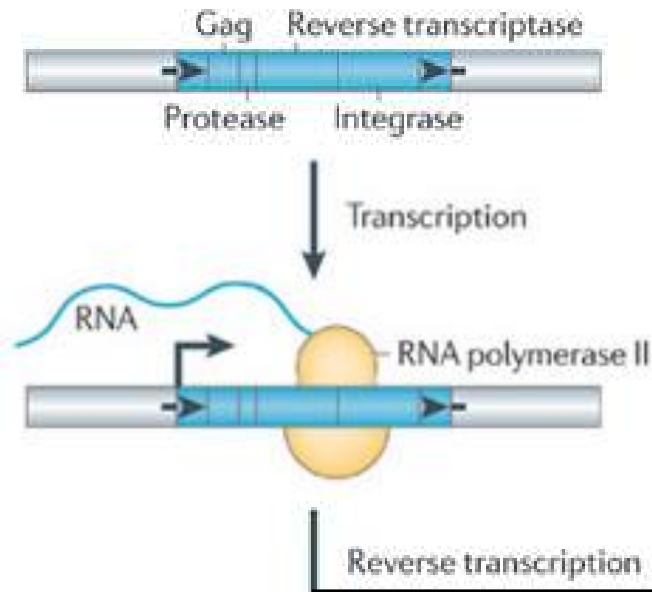


Non-allelic homologous recombination (NAHR):

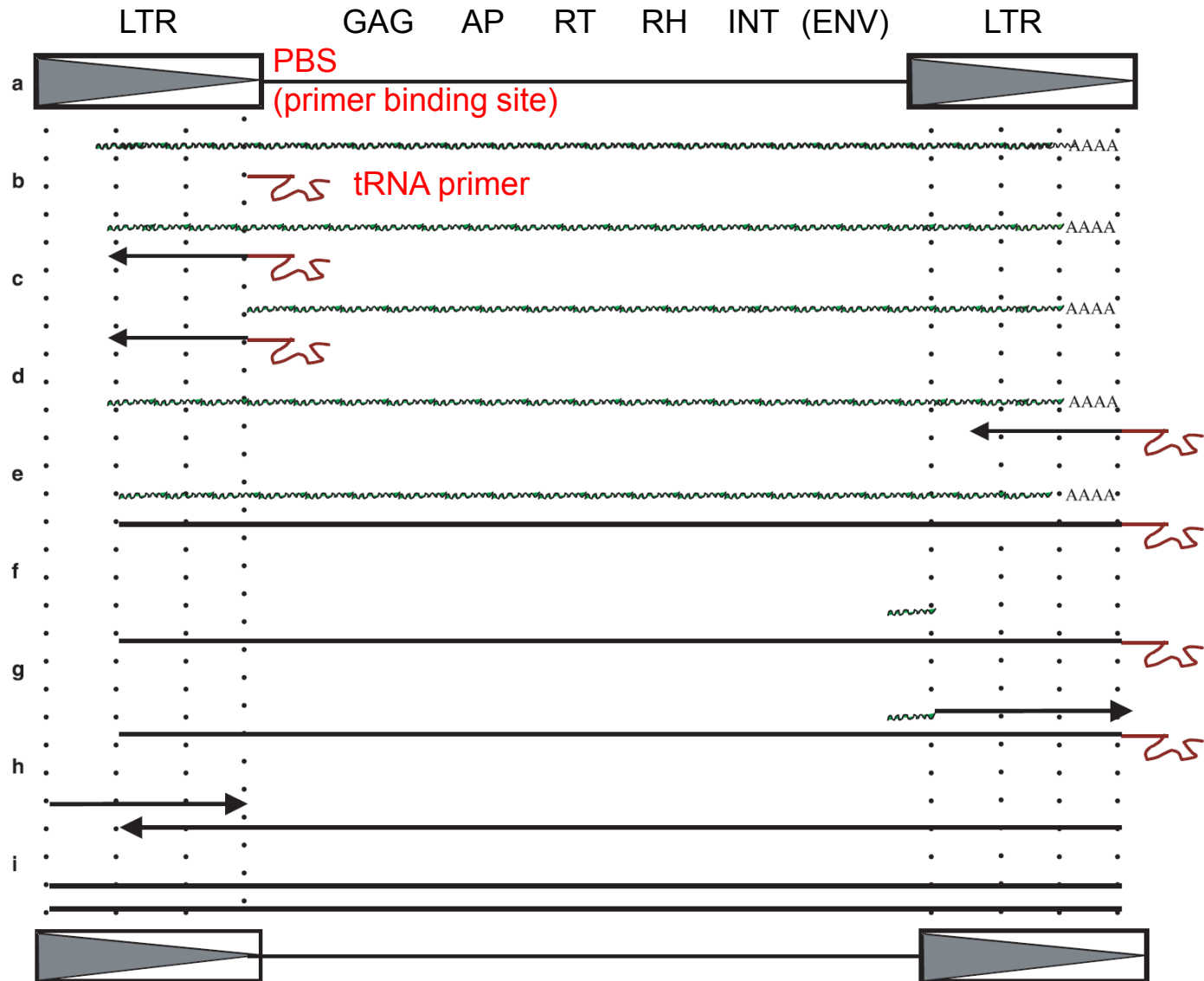


Replicative retrotransposition

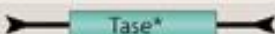
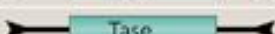
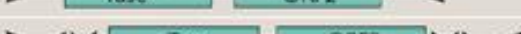
b LTR retrotransposon Replicative retrotransposition



Why LTR retrotransposons have LTRs



Class II: DNA transposons

Classification		Structure	TSD	Code	Occurrence
Order	Superfamily				
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
Crypton	Crypton		0	DYC	F



Cut-and-paste transposition (TIR)

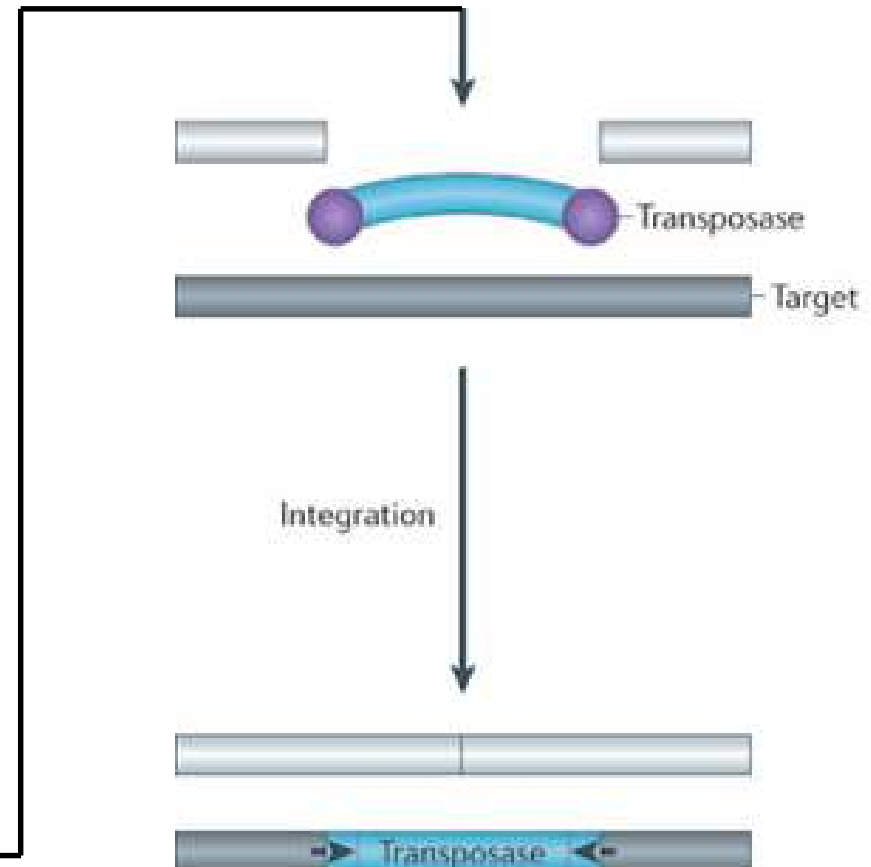
a DNA transposon
'Cut and paste' TE



Transposase
binding

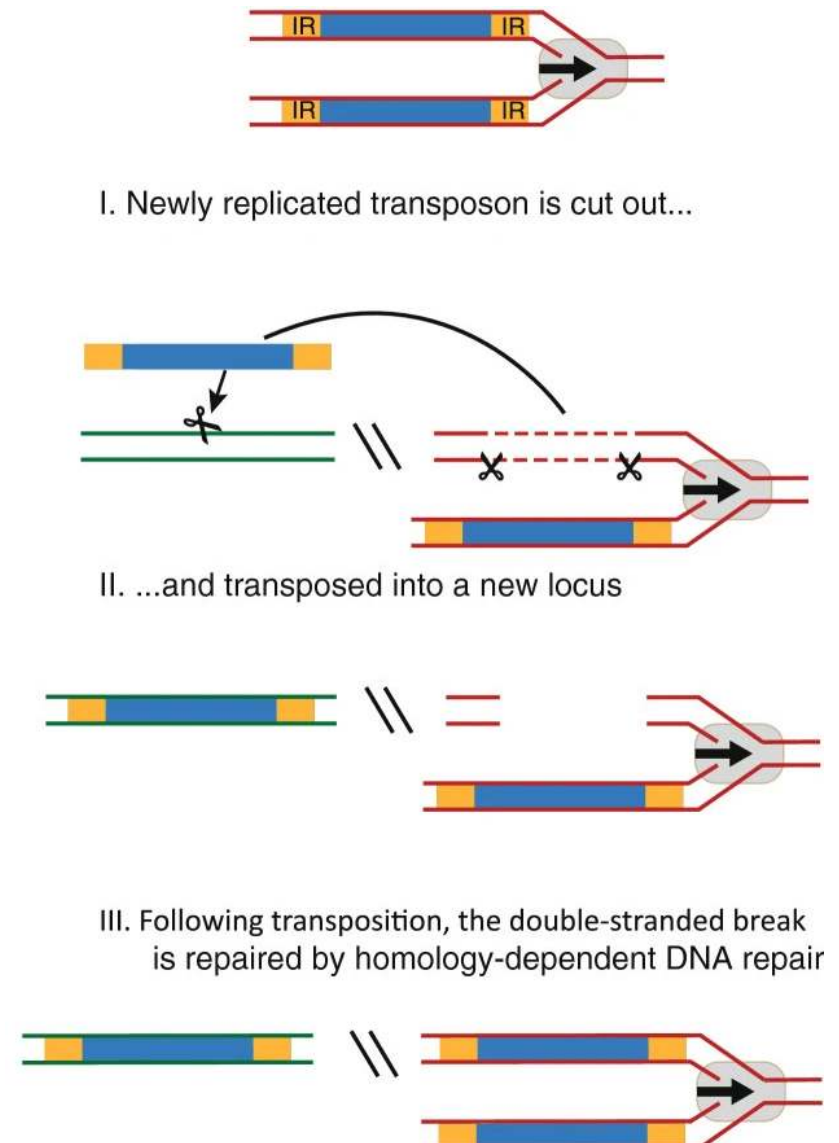
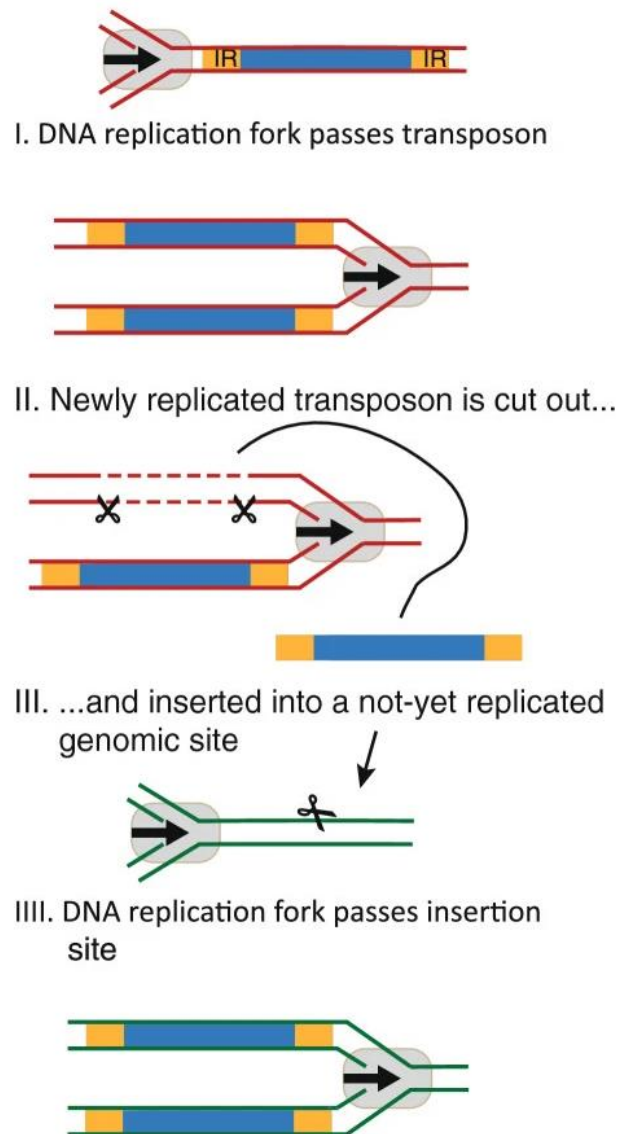


Excision



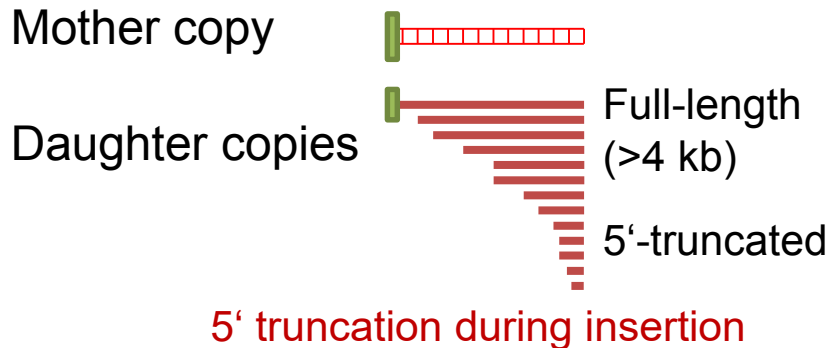
Mobile DNA

How to increase in copy number?

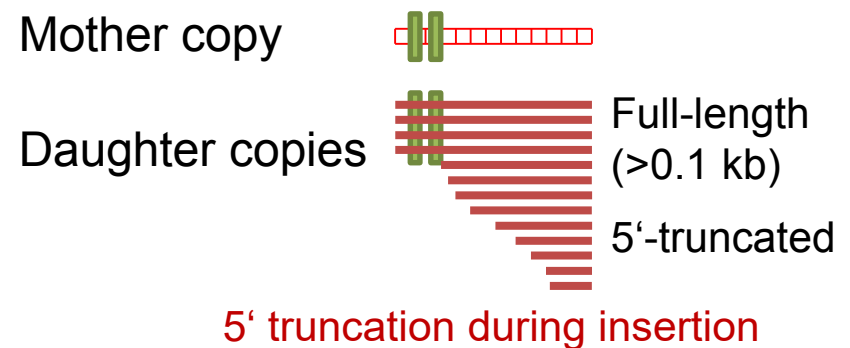


TE ≠ TE

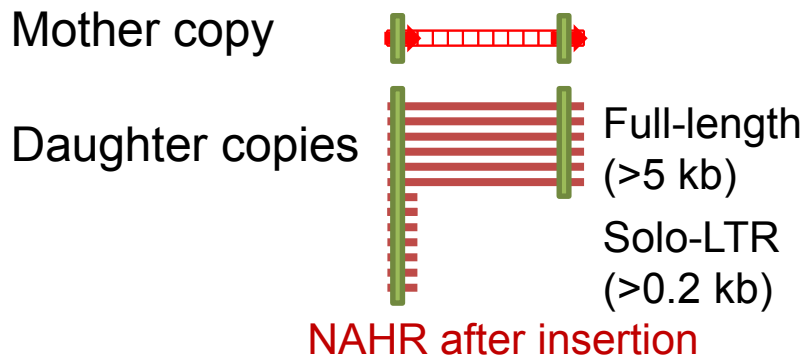
LINE



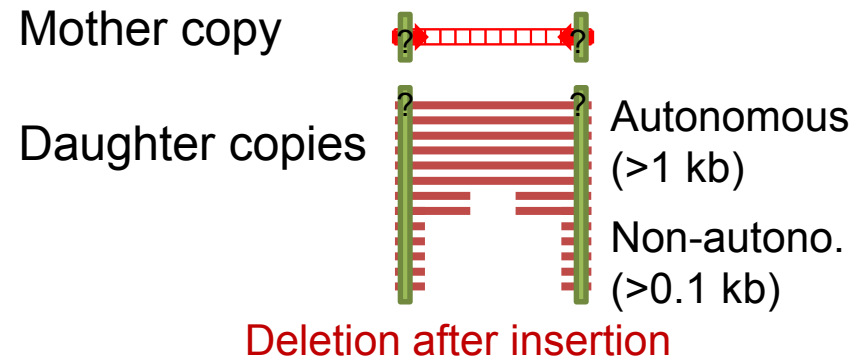
SINE



LTR



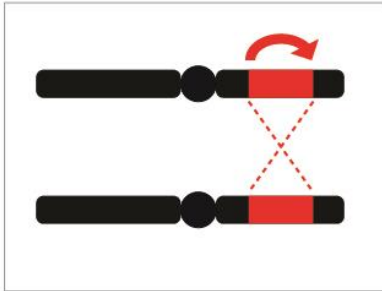
TIR



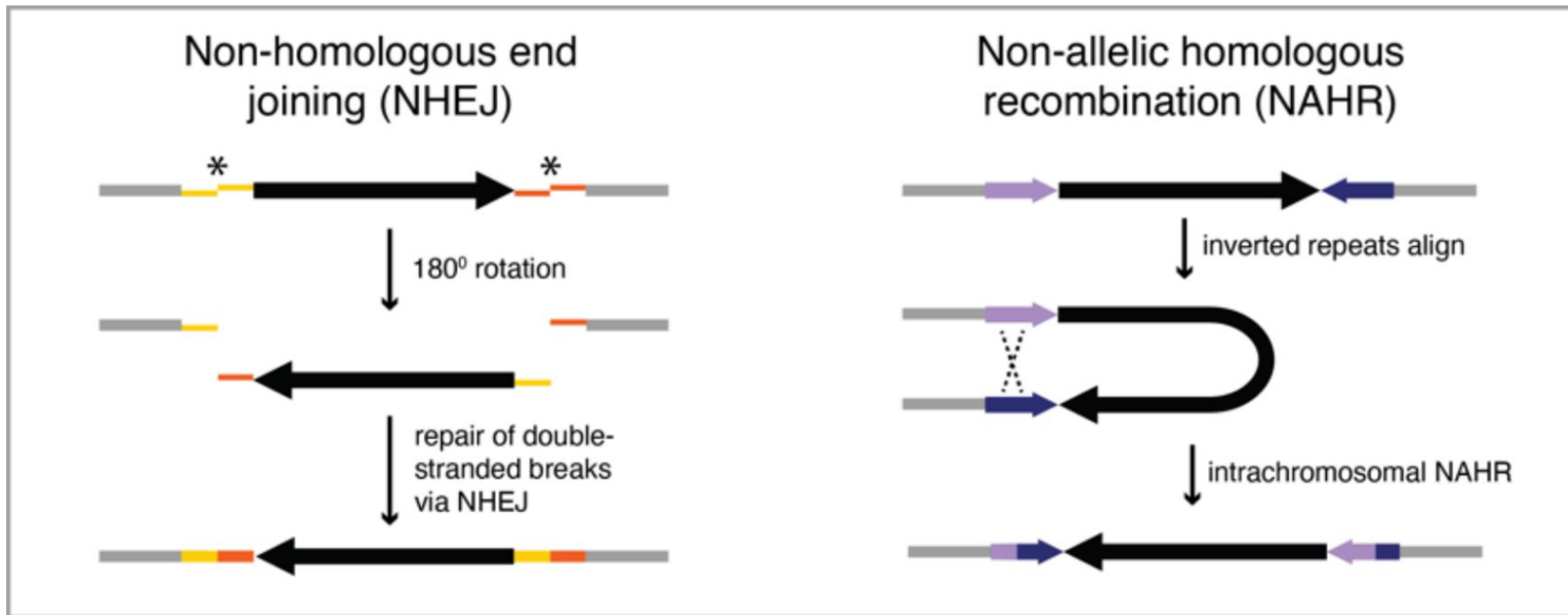
Some TE copies contain regulatory elements, some don't.

More context in [Suh 2021 TE lecture 2](#)

Inversion formation

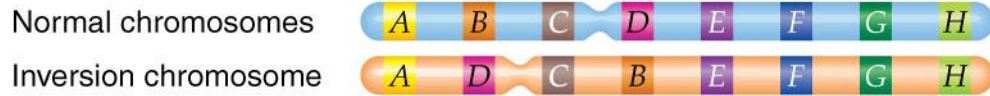


“We found that inversion breakpoints frequently occur in centromeric and telomeric regions and are often flanked by long inverted repeats (0.5-50 kb)”



Assembling or mapping inversion breakpoints is difficult!

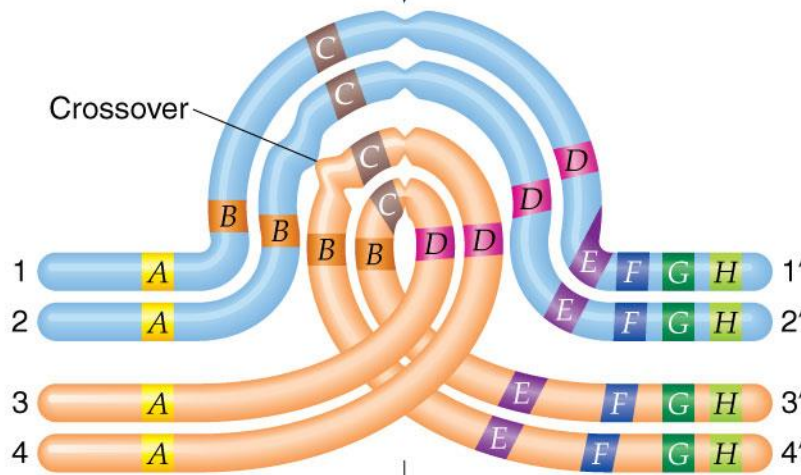
Inversions "reduce" recombination



Meiosis to prophase I

Crossover

Homologous chromosomes



Pericentric
inversion
(heterozygous)

Normal product
(all genes present)



Deletion/duplication
product (*EFGH* deleted;
A duplicated)



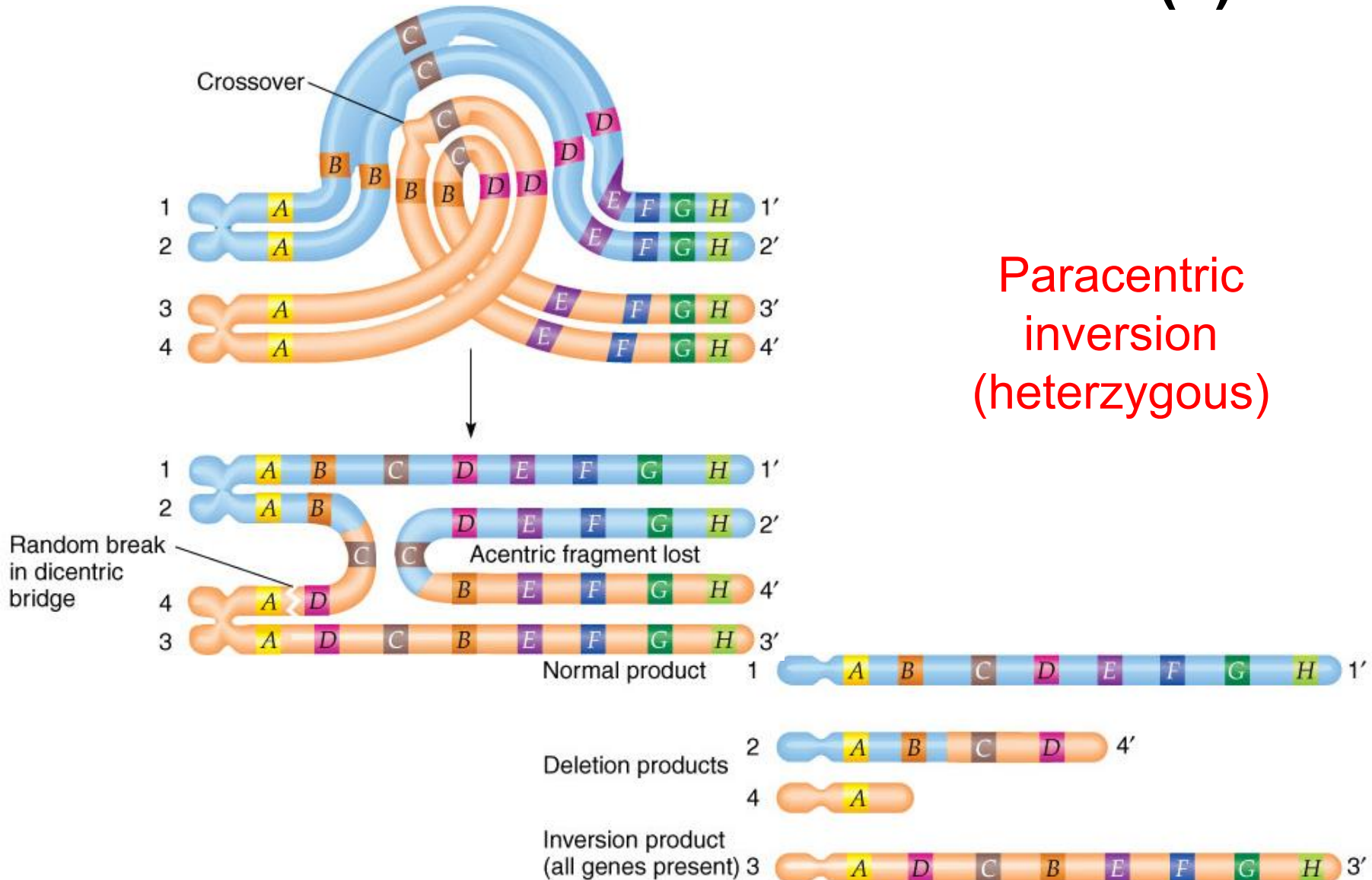
Inversion product
(all genes present)



Deletion/duplication
product (*A* deleted;
EFGH duplicated)

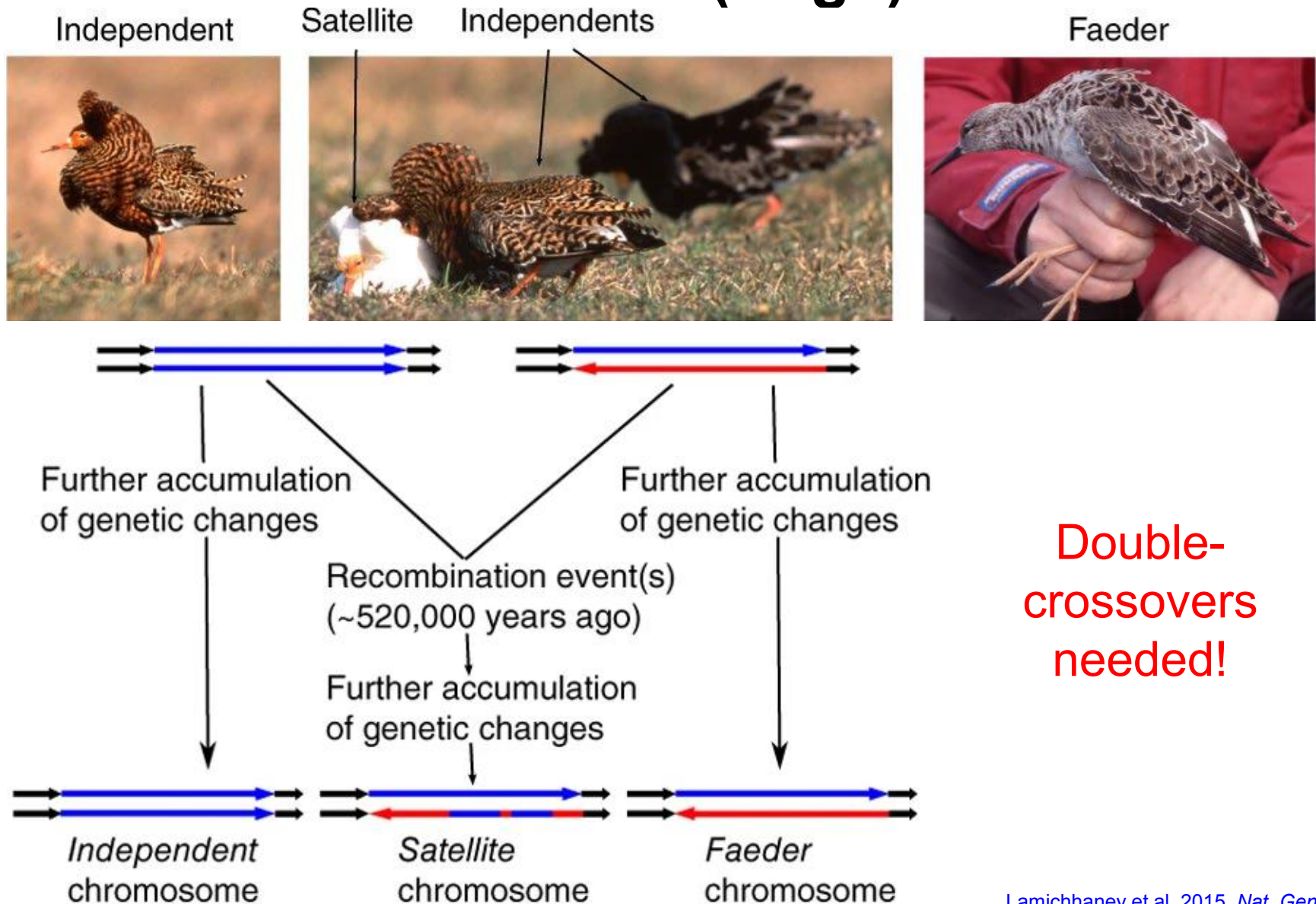


Inversions "reduce" recombination (2)



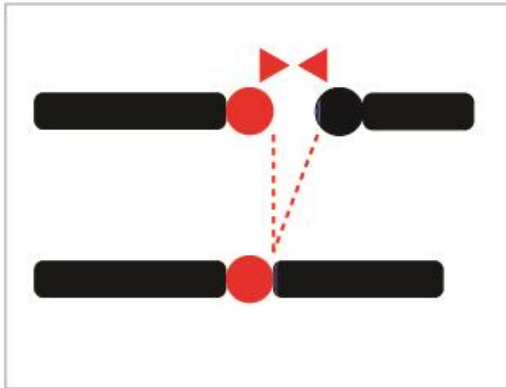
Paracentric
inversion
(heterozygous)

Rare recombination in (large) inversions

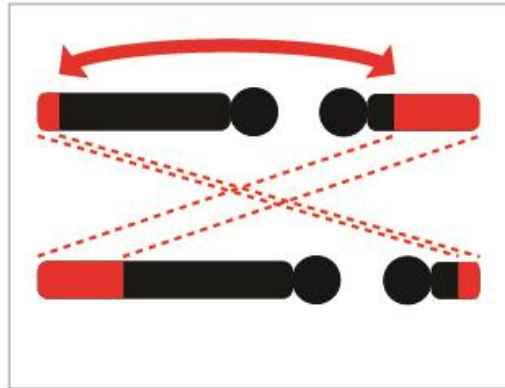


More cases of NAHR

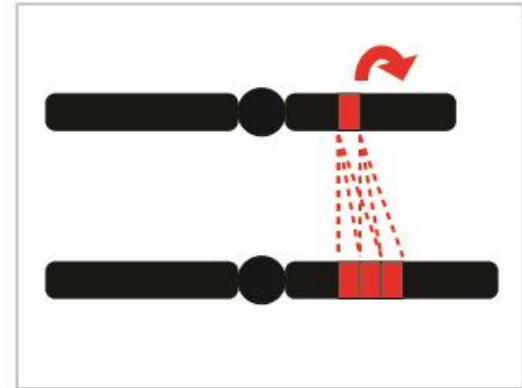
Fusion/fission



Translocation



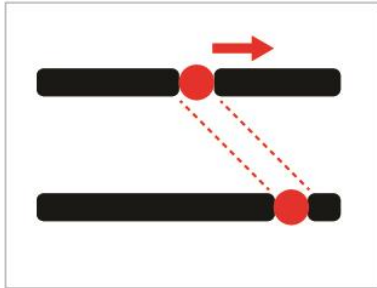
Duplication



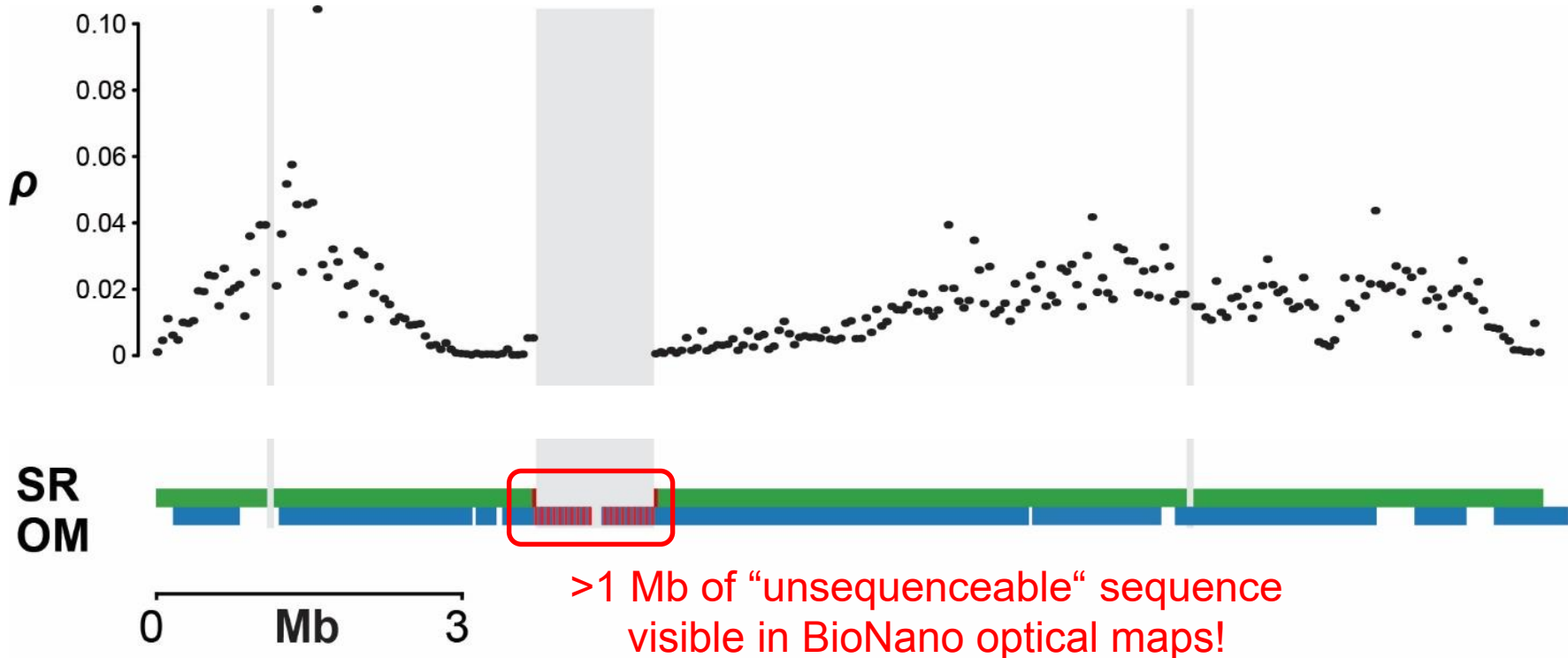
Fusions/fissions/translocations can decrease (new proximity to centromere) or increase (new proximity to telomere) recombination rates

Duplications can increase the chance of further non-allelic homologous recombination (NAHR)

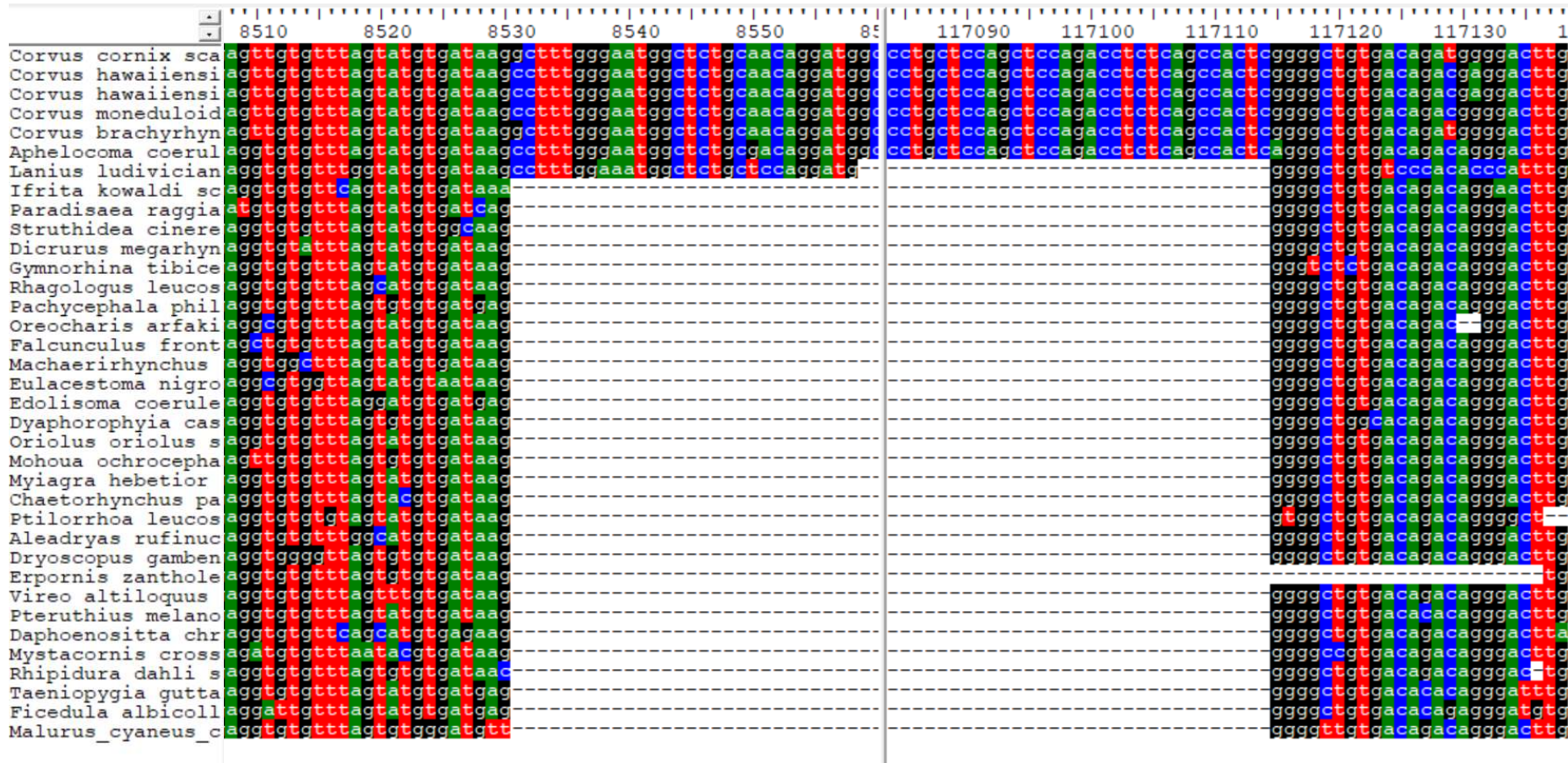
Centromere shifts



Chromosome 18 of hooded/carrion crow

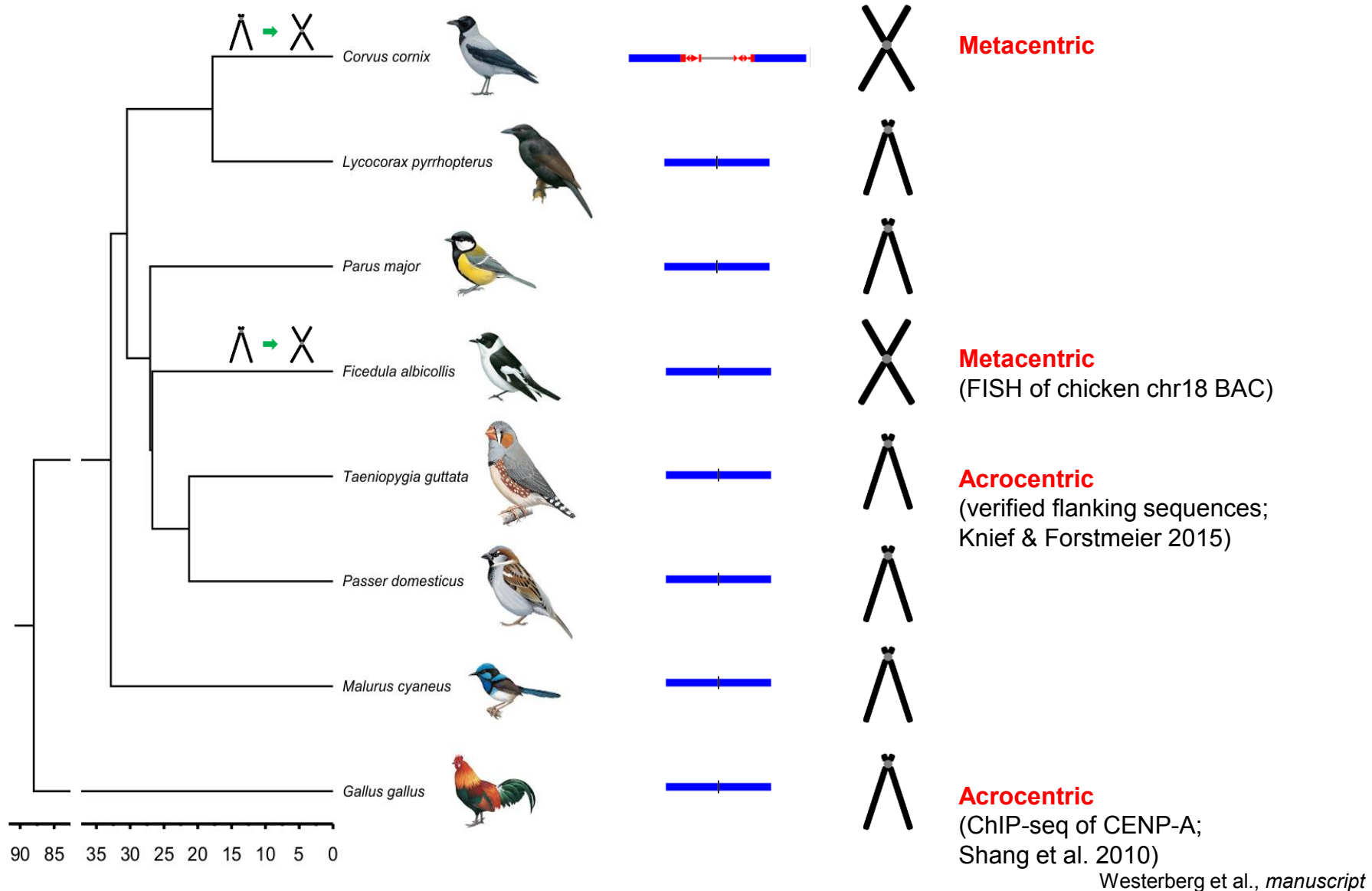


Centromere shifts across songbirds

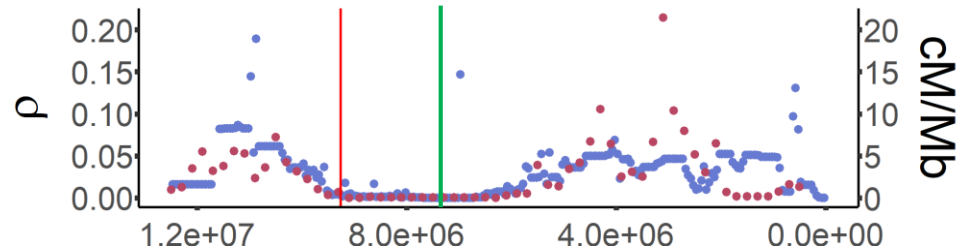
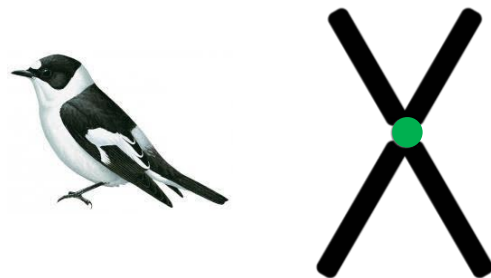
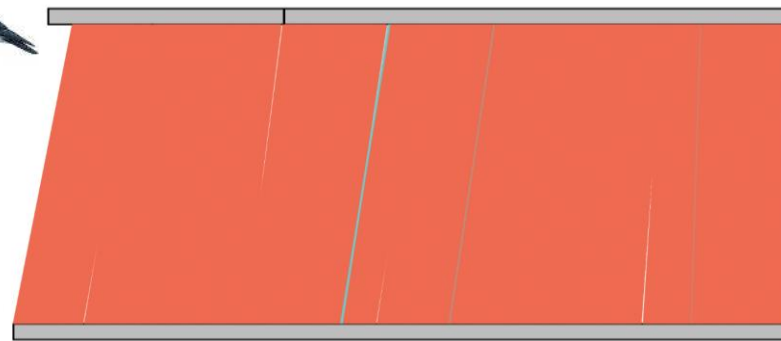
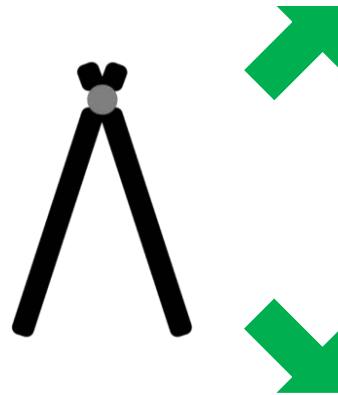
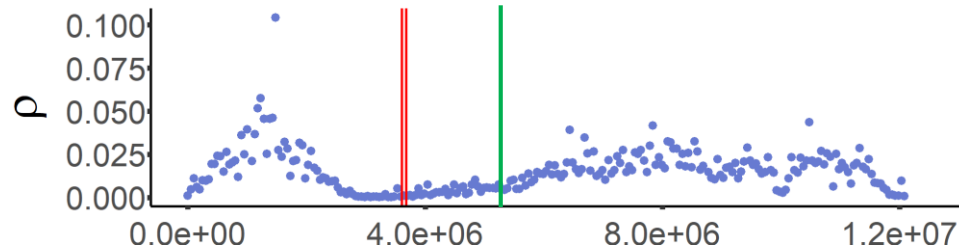
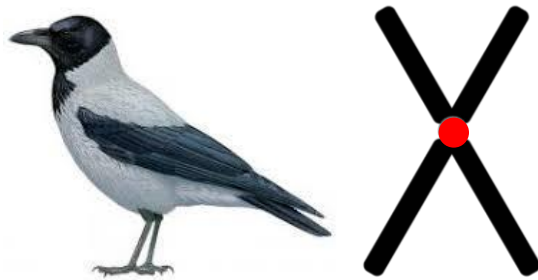


>1 Mb satellite DNA array
inserted in a formerly 5-kb
intergenic region!

Centromere shifts across songbirds



Not so stable chromosomes after all?

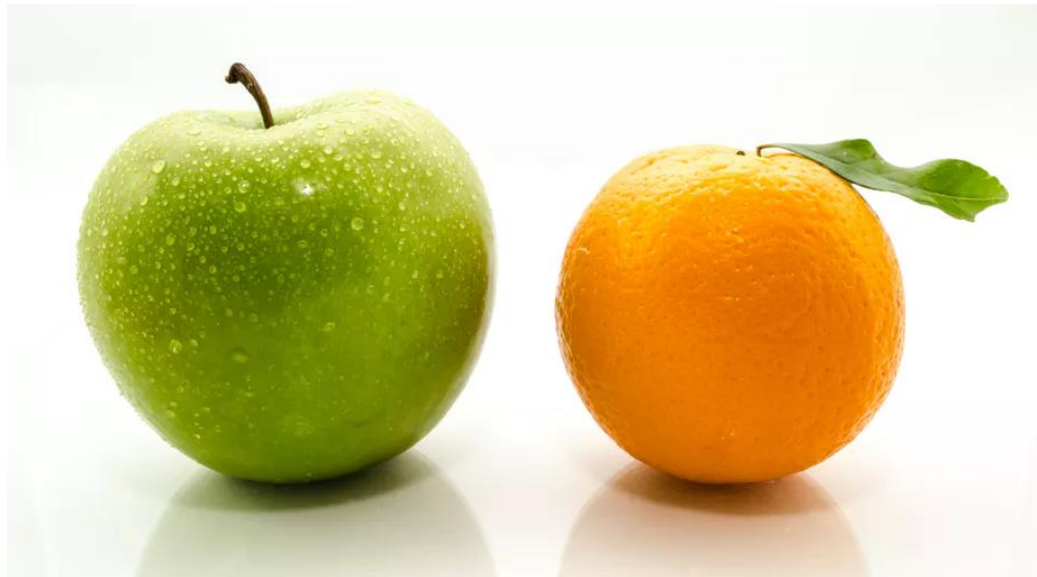


Short break (5 minutes)



Task: Gather in the same groups of 3 and discuss what resources (assembly/read data, gene/repeat annotation) there are for your respective study system.

Part 3: Hope

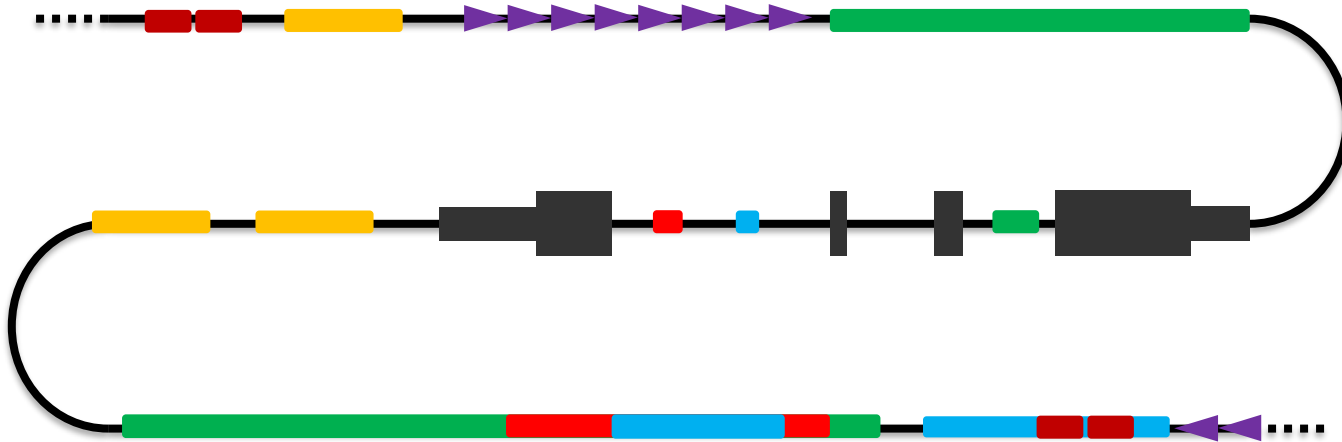


How frustrated are you?

- What types of SVs do you want to study?
- What types of SVs can you study?
- What data do you need to be less frustrated?



Genomes: ecosystems of selfish genes



Interspersed repeats

- Retrotransposons
- DNA transposons
- Endogenous viruses



Tandem repeats

- Satellites
- Minisatellites
- Microsatellites

Biodiversity inside each genome!

Cellular organisms

Phylum

Class

Order

Family

Genus

Species

Individual



Transposable elements

Class

Subclass

Order

Superfamily

Family

Subfamily

Copy



More
context in
[Suh 2021](#)
[TE](#)
[lecture 3](#)

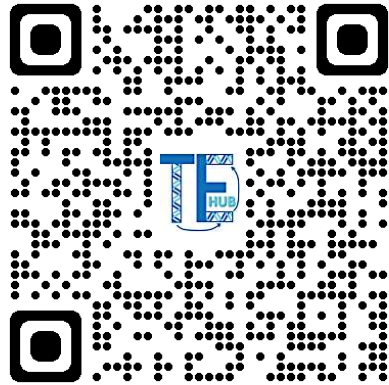
Too much TE data, too few TEologists

Analyses of 600+ insect genomes reveal repetitive element dynamics and highlight biodiversity-scale repeat annotation challenges

John S. Sproul^{1,2,3,12}, Scott Hotaling^{4,5,12}, Jacqueline Heckenhauer^{6,7,12},
Ashlyn Powell⁸, Dez Marshall², Amanda M. Larracuente³, Joanna L. Kelley^{4,9},
Steffen U. Pauls^{6,7,10} and Paul B. Frandsen^{6,8,11}

In most insect lineages, 25%–85% of repetitive sequences were “unclassified” following automated annotation, compared with only ~13% in *Drosophila* species. Although the diversity of available insect genomes has rapidly expanded, we show the rate of community contributions to RE databases has not kept pace, preventing efficient annotation and high-resolution study of REs in most groups. We highlight the tremendous opportunity and need for the biodiversity genomics field to embrace REs and suggest collective steps for making progress toward this goal.

More community initiatives needed



TE Hub website



TE Worldwide Slack
[#te-hub](#) channel

Teaching transposon classification as a means to crowd source the curation of repeat annotation – a tardigrade perspective



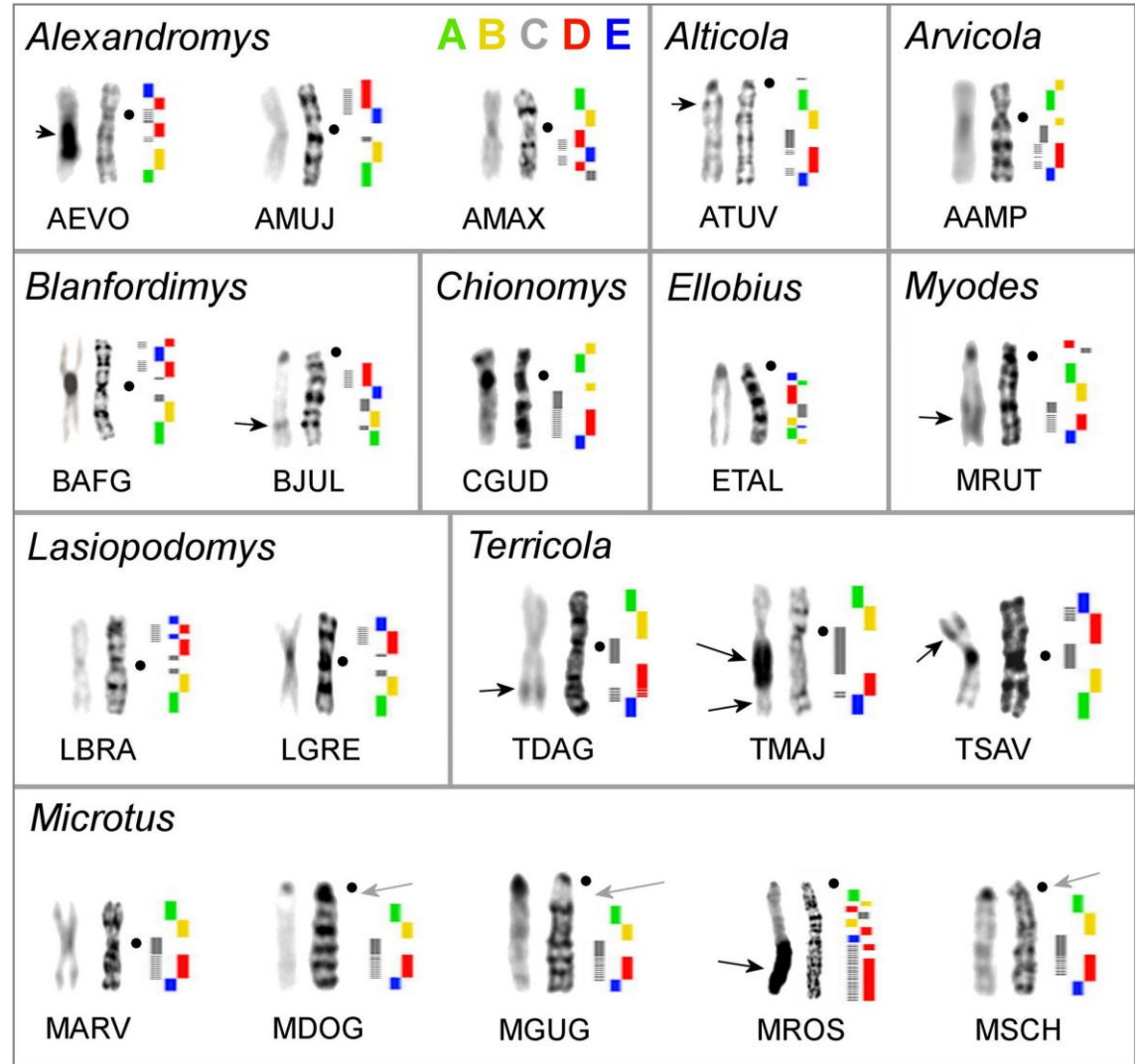
Valentina Peona^{1,2,3*}, Jacopo Martelossi^{4*}, Dureen Almojil⁵, Julia Bocharkina⁶, Ioana Brännström^{7,8}, Max Brown⁹, Alice Cang¹⁰, Tomàs Carrasco-Valenzuela^{11,12}, Jon DeVries¹³, Meredith Doellman^{14,15}, Daniel Elsner¹⁶, Pamela Espíndola-Hernández¹⁷, Guillermo Friis Montoya¹⁸, Bence Gaspar¹⁹, Danijela Zagorski²⁰, Paweł Hałakuc²¹, Beti Ivanovska²², Christopher Laumer²³, Robert Lehmann²⁴, Ljudevit Luka Boštjančič²⁵, Rahia Mashoodh²⁶, Sofia Mazzoleni²⁷, Alice Mouton²⁸, Maria Anna Nilsson²⁵, Yifan Pei^{1,29}, Giacomo Potente³⁰, Panagiotis Provataris³¹, José Ramón Pardos-Blas³², Ravindra Raut³³, Tomasa Sbaffi³⁴, Florian Schwarz³⁵, Jessica Stapley³⁶, Lewis Stevens³⁷, Nusrat Sultana³⁸, Radka Symonova³⁹, Mohadeseh S. Tahami⁴⁰, Alice Urzi⁴¹, Heidi Yang⁴², Abdullah Yusuf⁴³, Carlo Pecoraro⁴⁴ and Alexander Suh^{1,45,46*}



Genomics + cytogenetics = cytogenomics

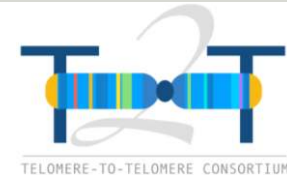
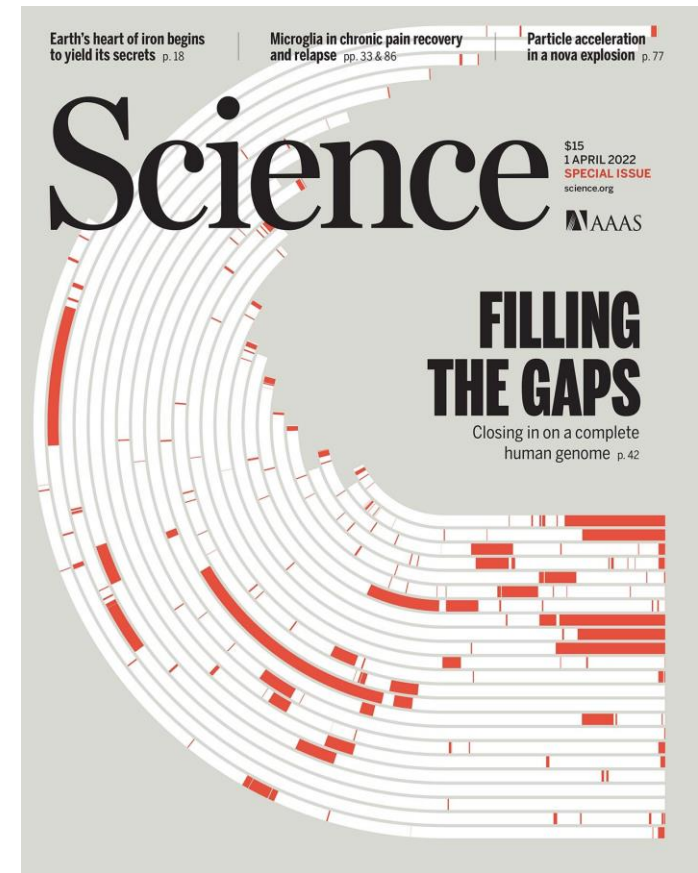


Vole X
chromosomes
(C banding vs.
G banding vs.
in-situ hybridization
of region-specific
DNA probes)



What's next: Telomere-to-telomere omics?

- Nearly 200 million bp more than the previous human reference (GRCh38) with 1956 new genes (99 protein-coding) and 0 assembly gaps!
- Homozygous cell line sequenced with: 120x coverage of Oxford Nanopore ultra-long reads, 70x PacBio CLR long reads, 30x PacBio HiFi long reads, 50x 10X Genomics linked reads, BioNano DLS optical maps, Arima Genomics Hi-C maps.



Money is less of a limitation now than sample amount + quality + repetitiveness!

What's next: Machine learning?

DeepTE: a computational method for de novo classification of transposons with convolutional neural network

[Yan et al. 2020, *Bioinformatics*](#)

TERL: classification of transposable elements by convolutional neural networks FREE

[Pereira da Cruz et al. 2020 *Brief. Bioinform.*](#)

TransposonUltimate: software for transposon classification, annotation and detection

[Riehl et al. 2022, *Nucl. Acids Res.*](#)

Genomic object detection: An improved approach for transposable elements detection and classification using convolutional neural networks

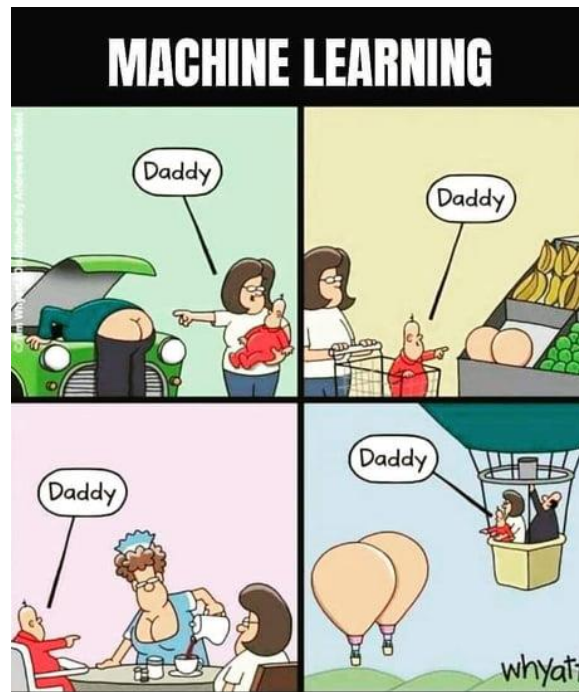
[Orozco-Arias et al. 2023, *PLoS ONE*](#)

TEclass2: Classification of transposable elements using Transformers

[Bickmann et al. 2023 *bioRxiv*](#)

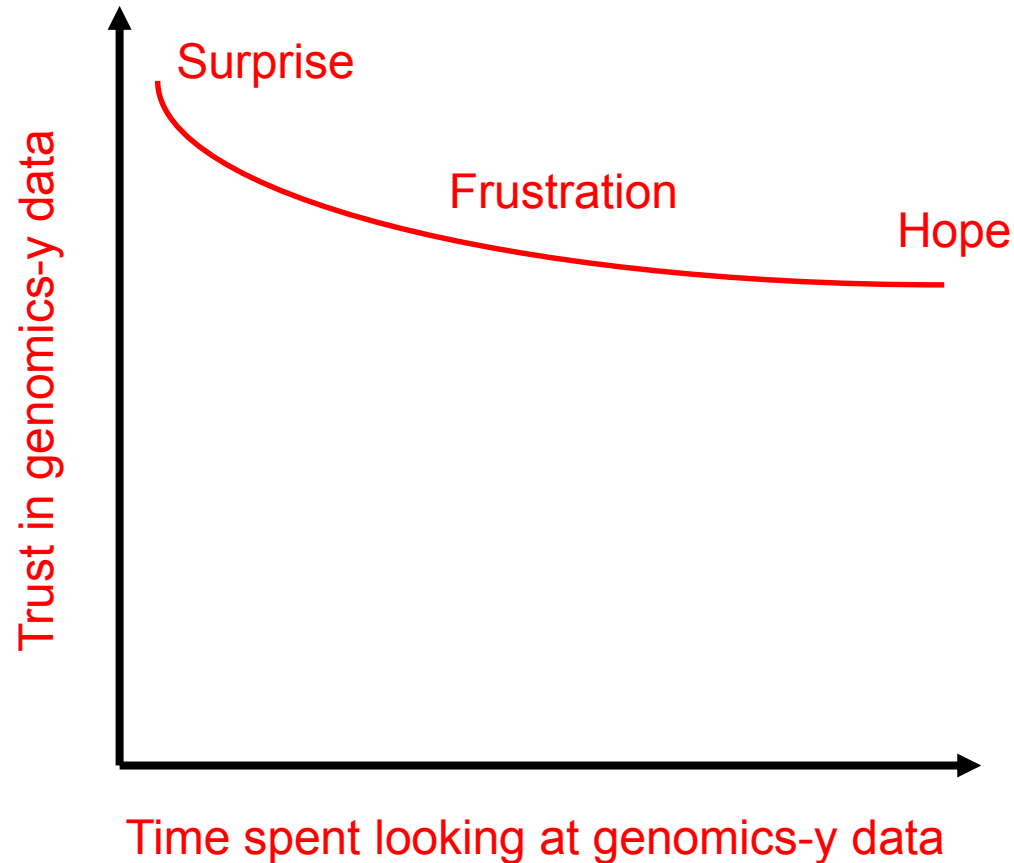
Comprehensive Hierarchical Classification of Transposable Elements based on Deep Learning

[Qi et al. 2024, *bioRxiv*](#)



Prediction: AI training (cf. SV biology and curation) will be a key bottleneck for evaluating machine learning results!

Conclusion: Genomics is no silver bullet



What to take with a grain of salt

1. How can we declare something as absent in a genome (evidence of absence vs. absence of evidence)?
2. How can we study unassembled or underassembled regions (multicopy genes, GC-rich genes, TEs)?
3. How can we compare species with different assembly qualities, data types, or annotation efforts?
4. How can we account for unknown peculiarities (sex chromosomes, B chromosomes, germline/soma genome differences ...)?



Questions?

