Structural variation

(A story of surprise → frustration → hope)

January 23, 2025

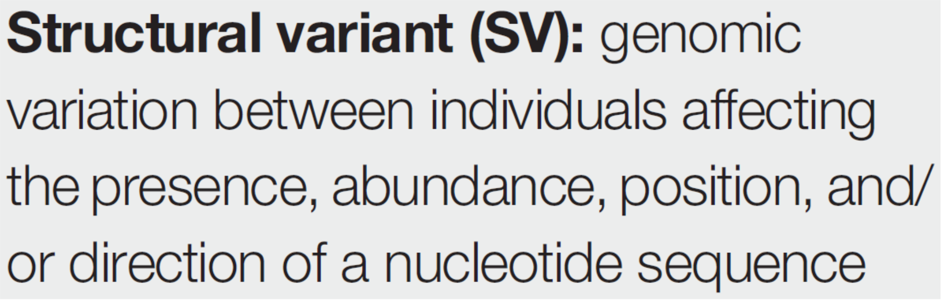
~~@alexander_suh~~

Alexander Sang-Jae Suh
서상재 徐商在

LIB Leibniz Institute for the Analysis of Biodiversity Change

UNIVERSITÄT BONN

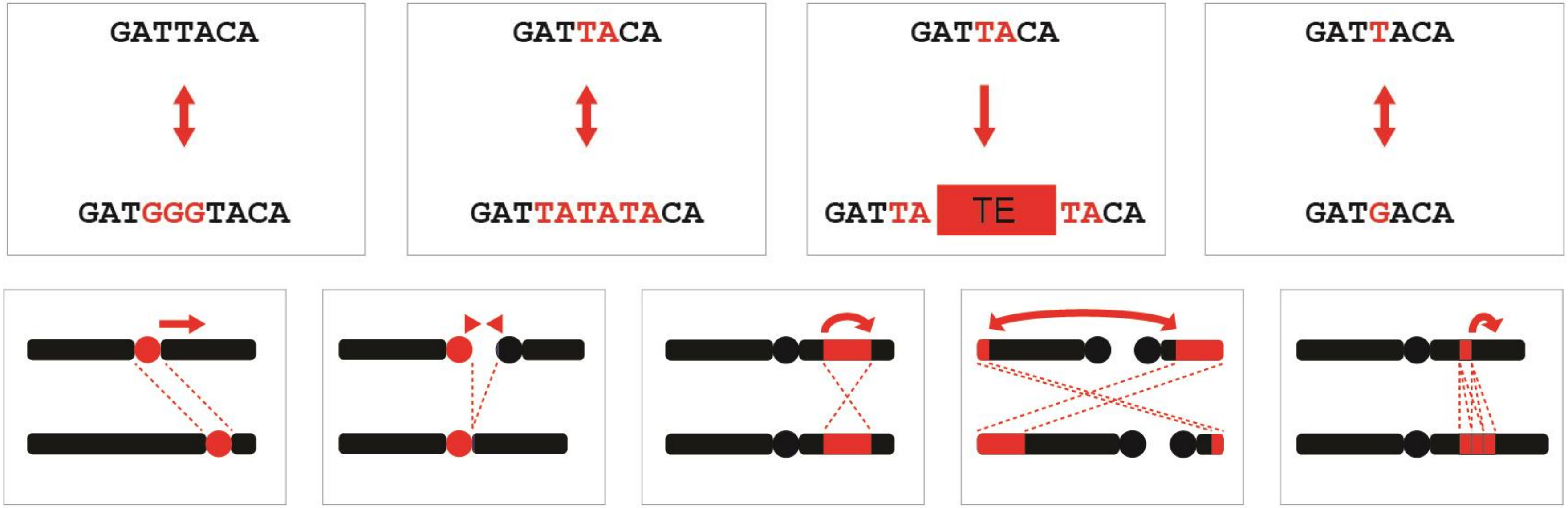
What is structural variation?



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

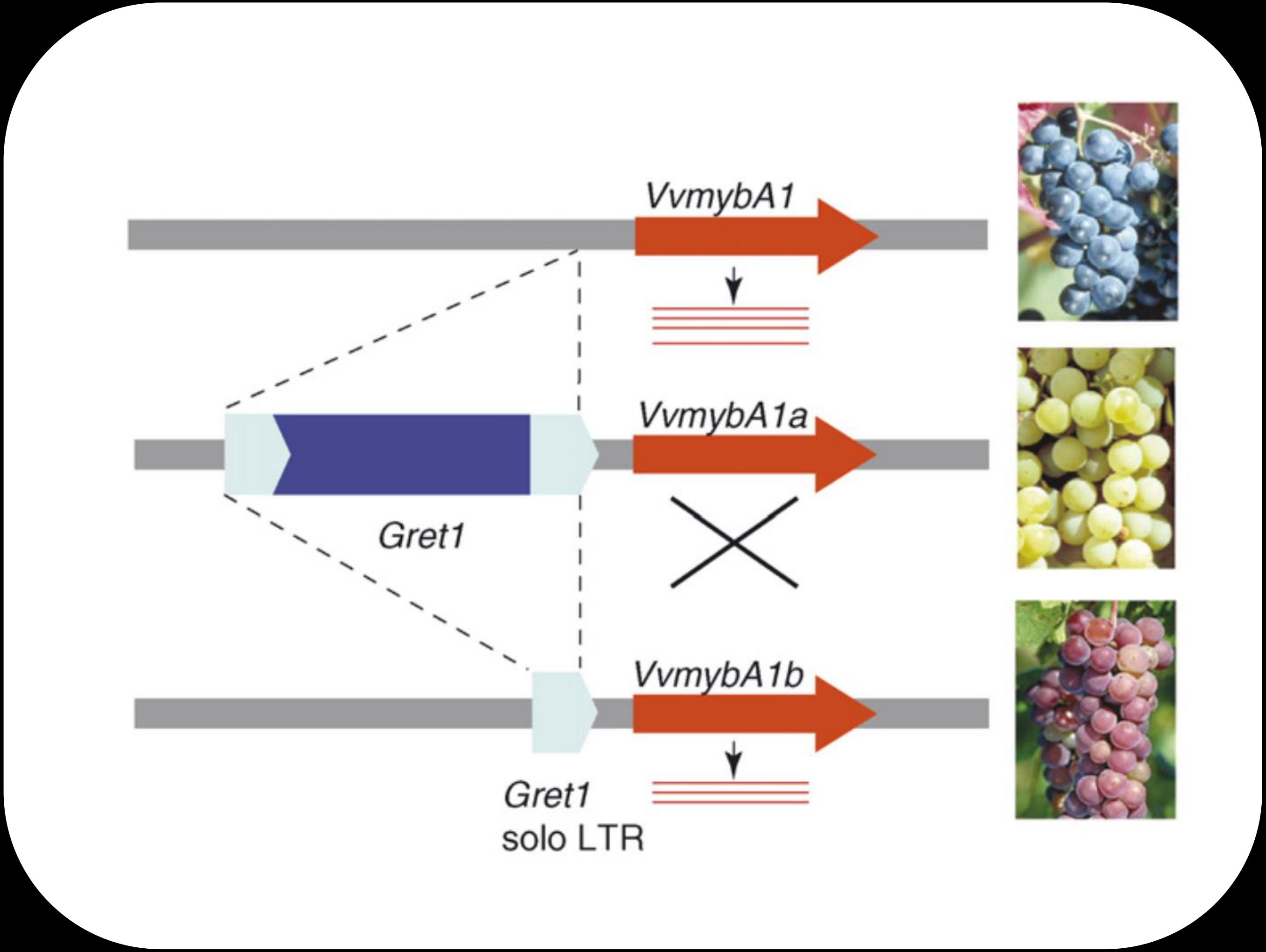
Mérot et al. 2020, *Trends Genet.*

Some key mutation types



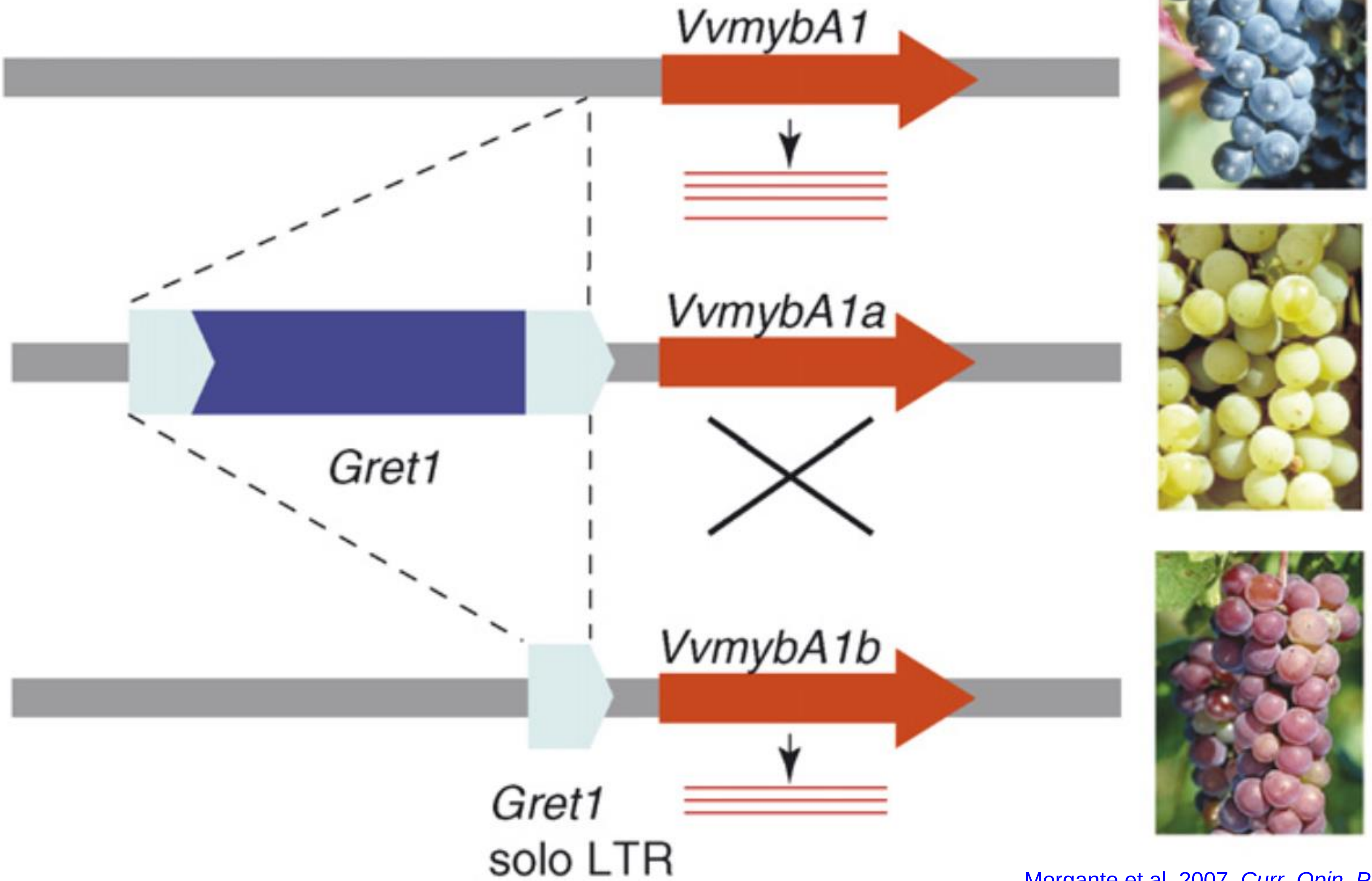
Berdan et al. 2021, *Mol. Ecol.*

Part 1: Surprise



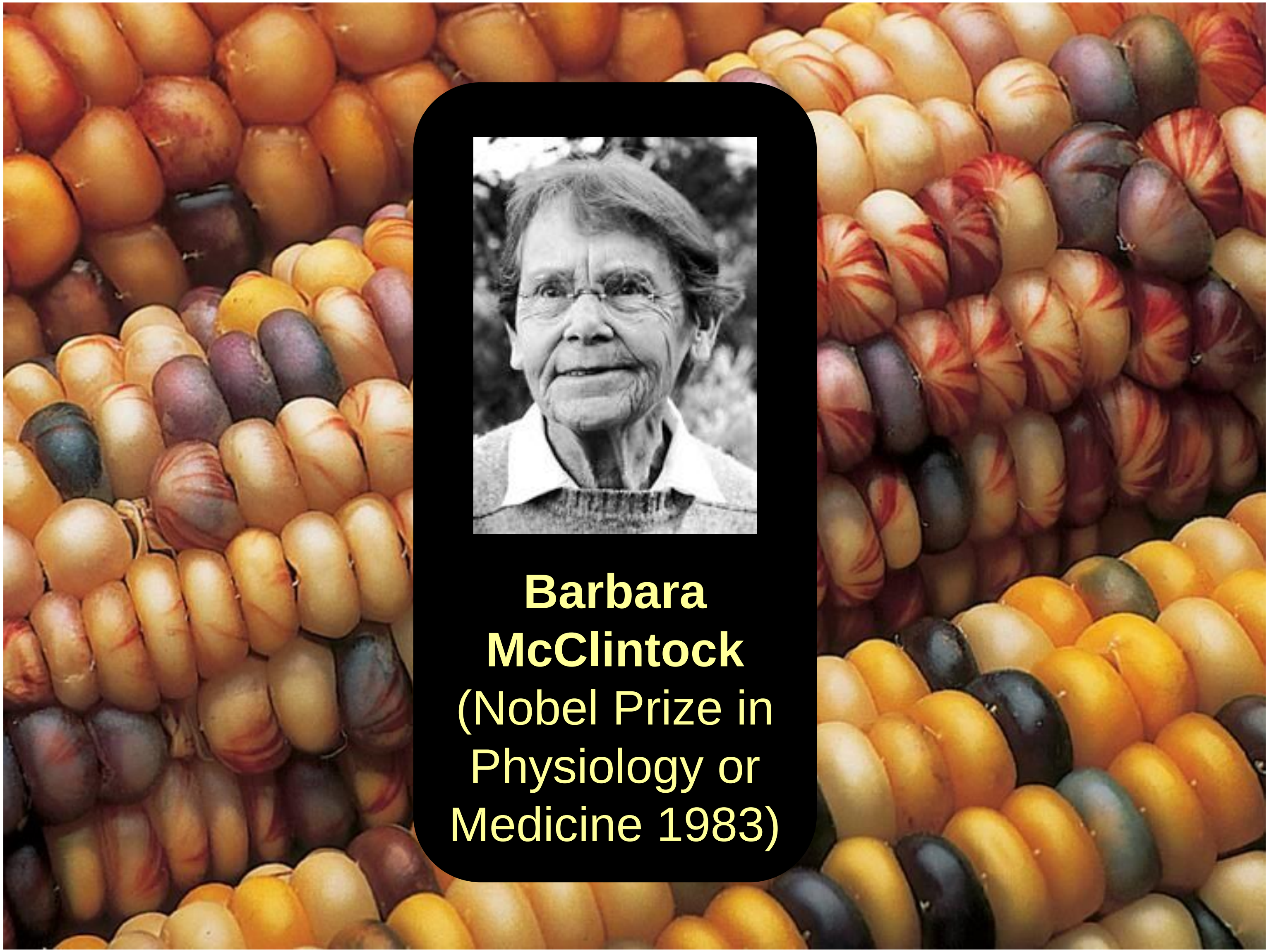
A) It’s not a SNP!

Delicious effects of SVs



Morgante et al. 2007, *Curr. Opin. Plant. Biol.*

Discovery of gene regulation in 1940s

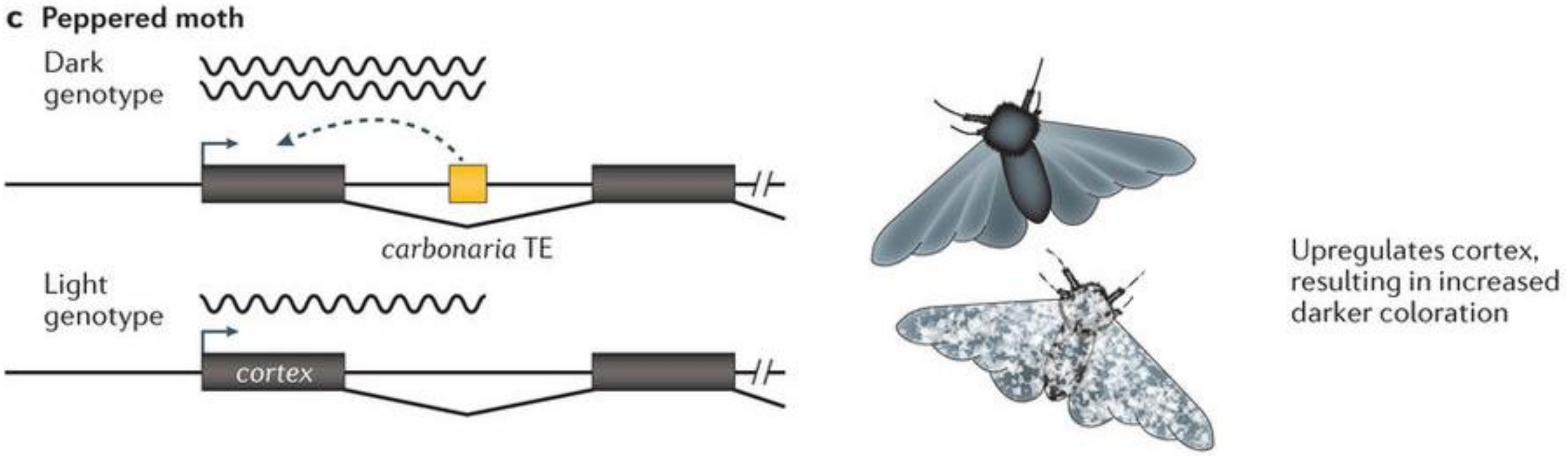


Barbara McClintock

(Nobel Prize in Physiology or Medicine 1983)

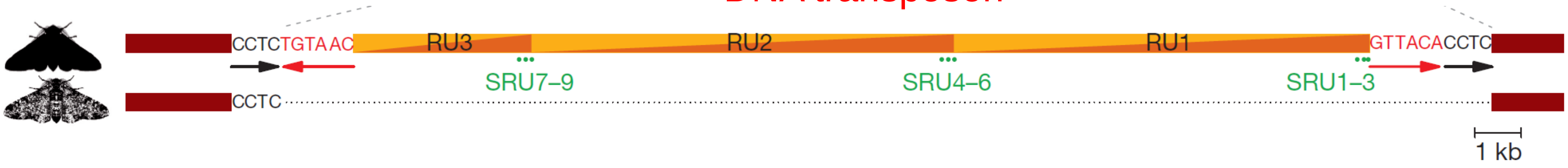
TE-induced rapid adaptation

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



Chuong et al. 2017 *Nat. Rev. Genet.*

DNA transposon

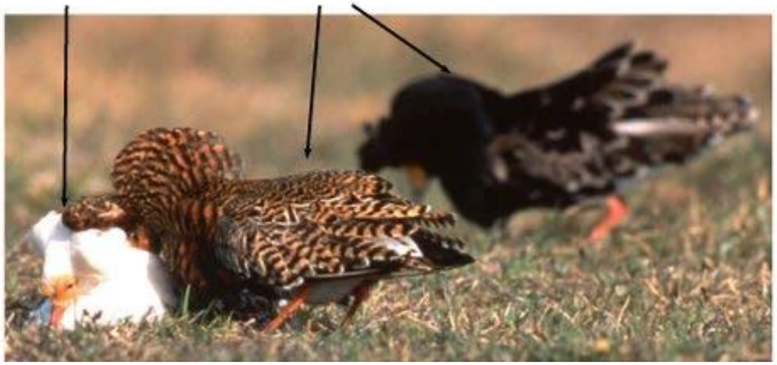


Van‘t Hof et al. 2016, *Nature*

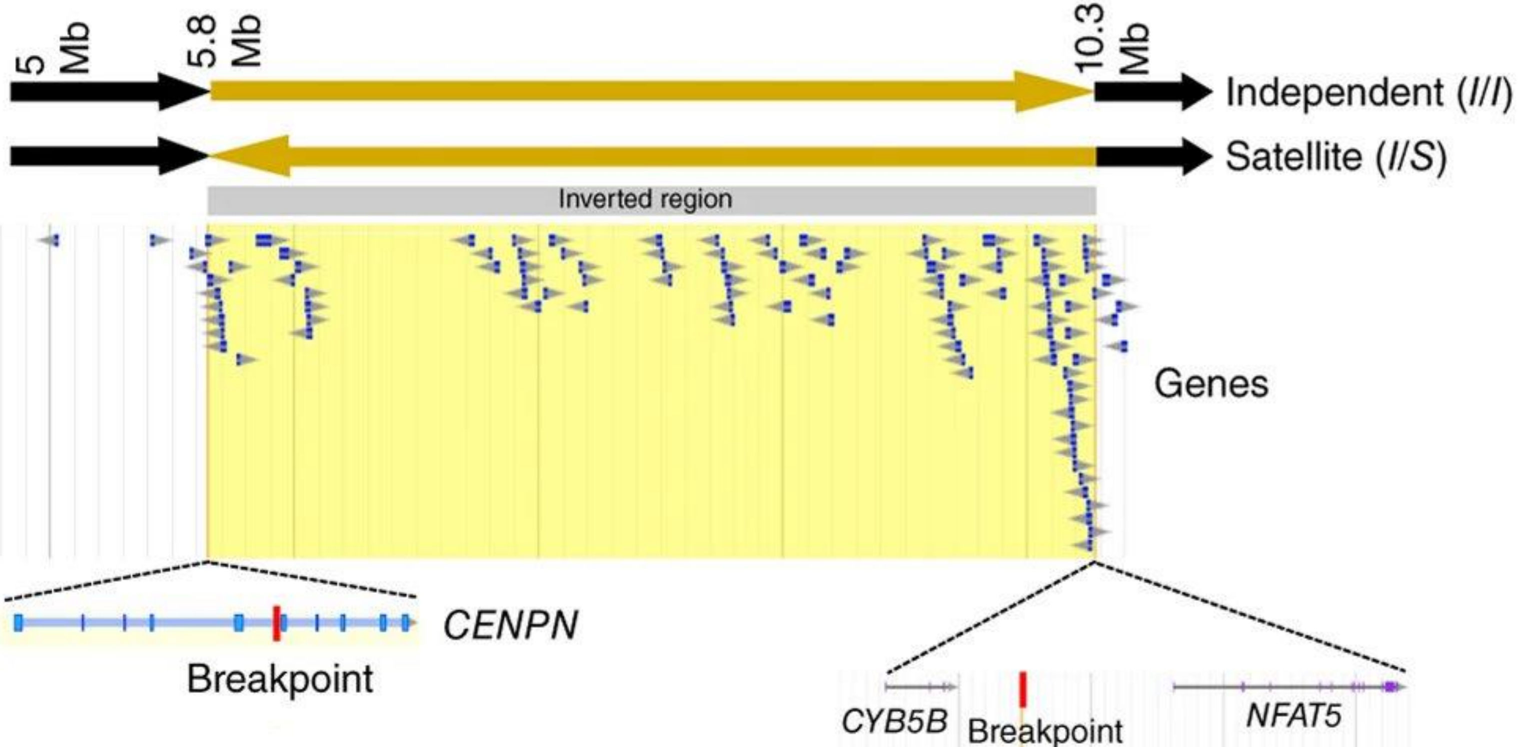
Inversions in ruff reproductive strategies

Independent

Satellite Independents

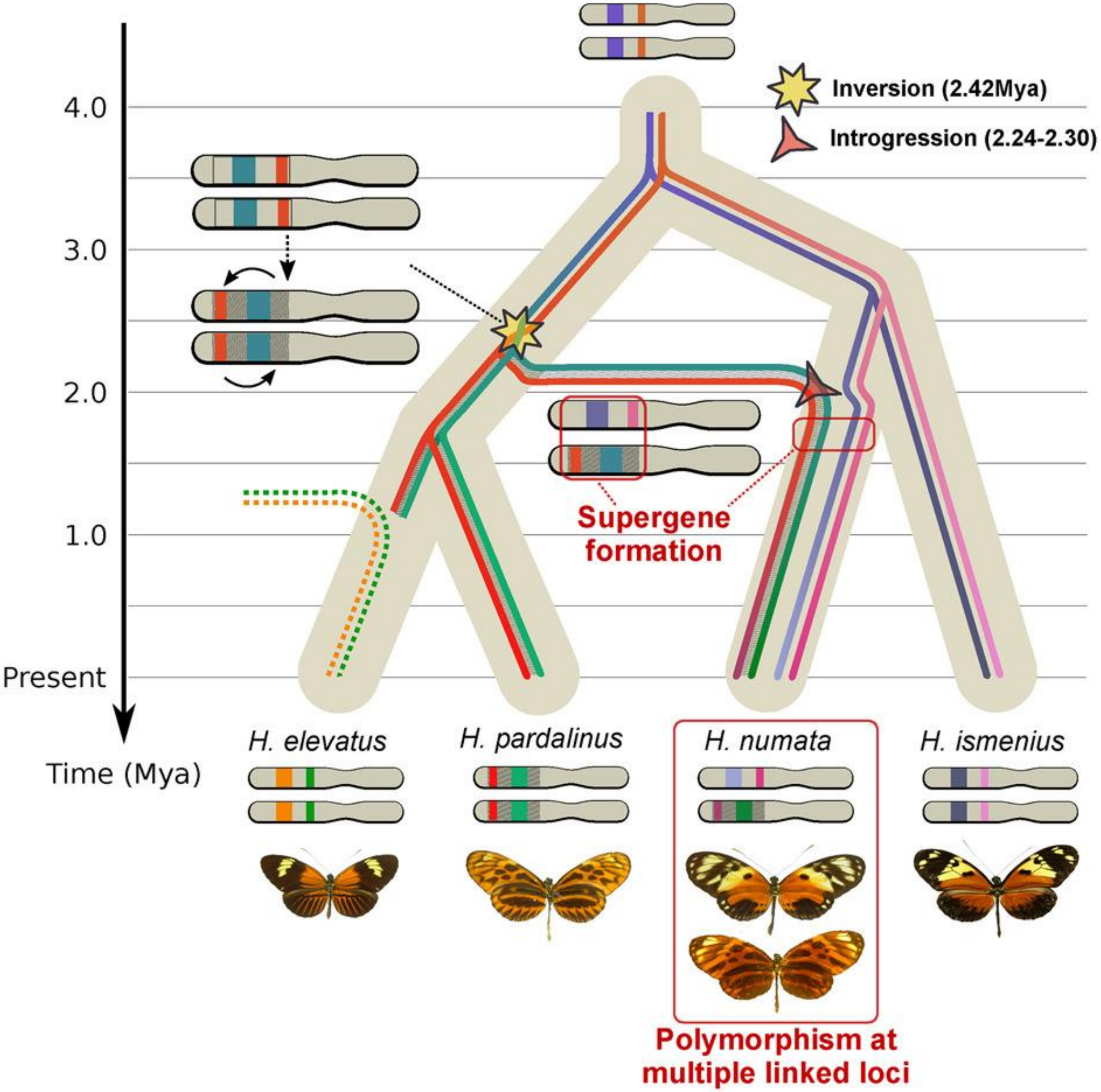


Faeder



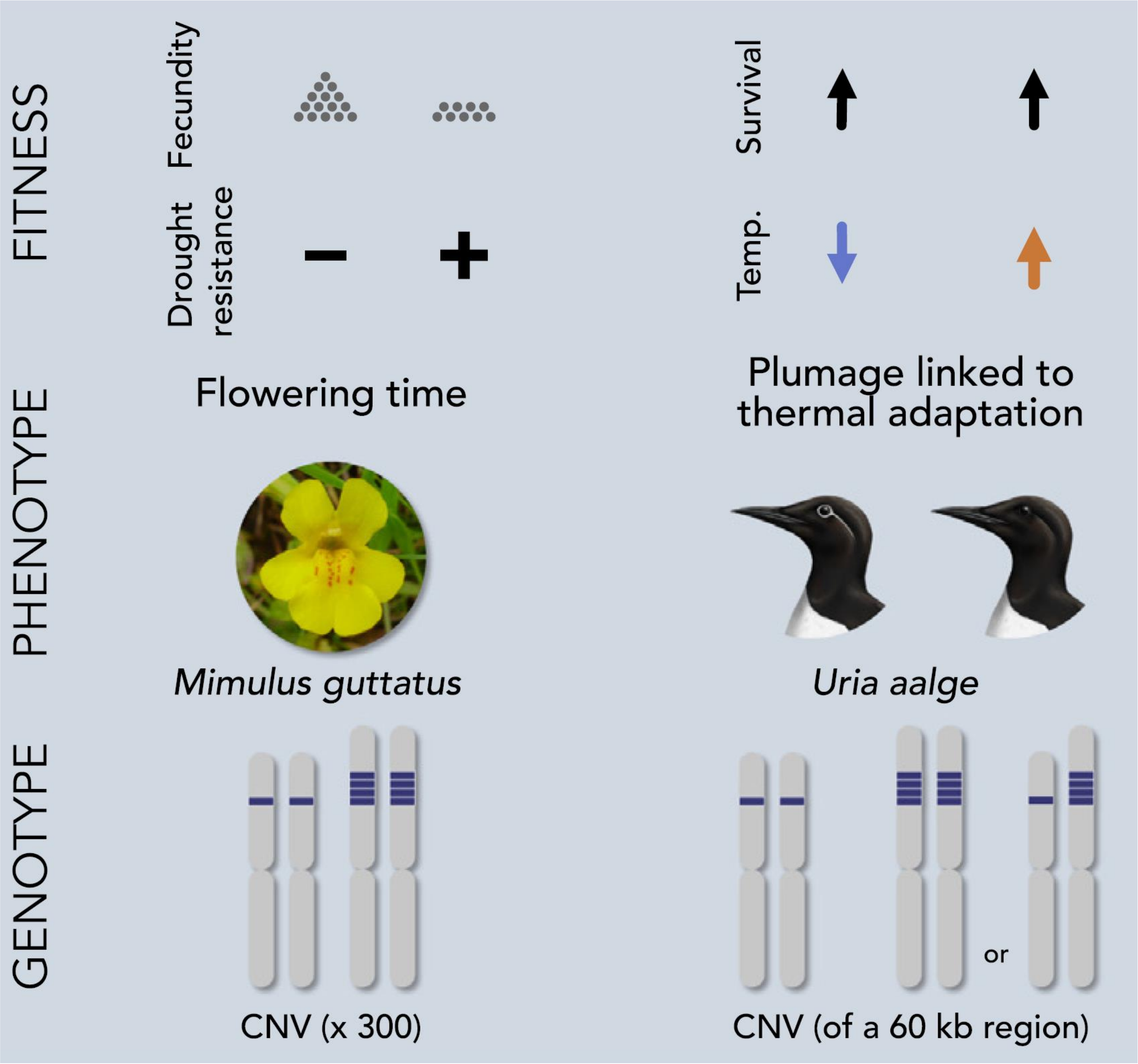
Lamichhaney et al. 2015, *Nat. Genet.*

Inversion introgression and supergenes



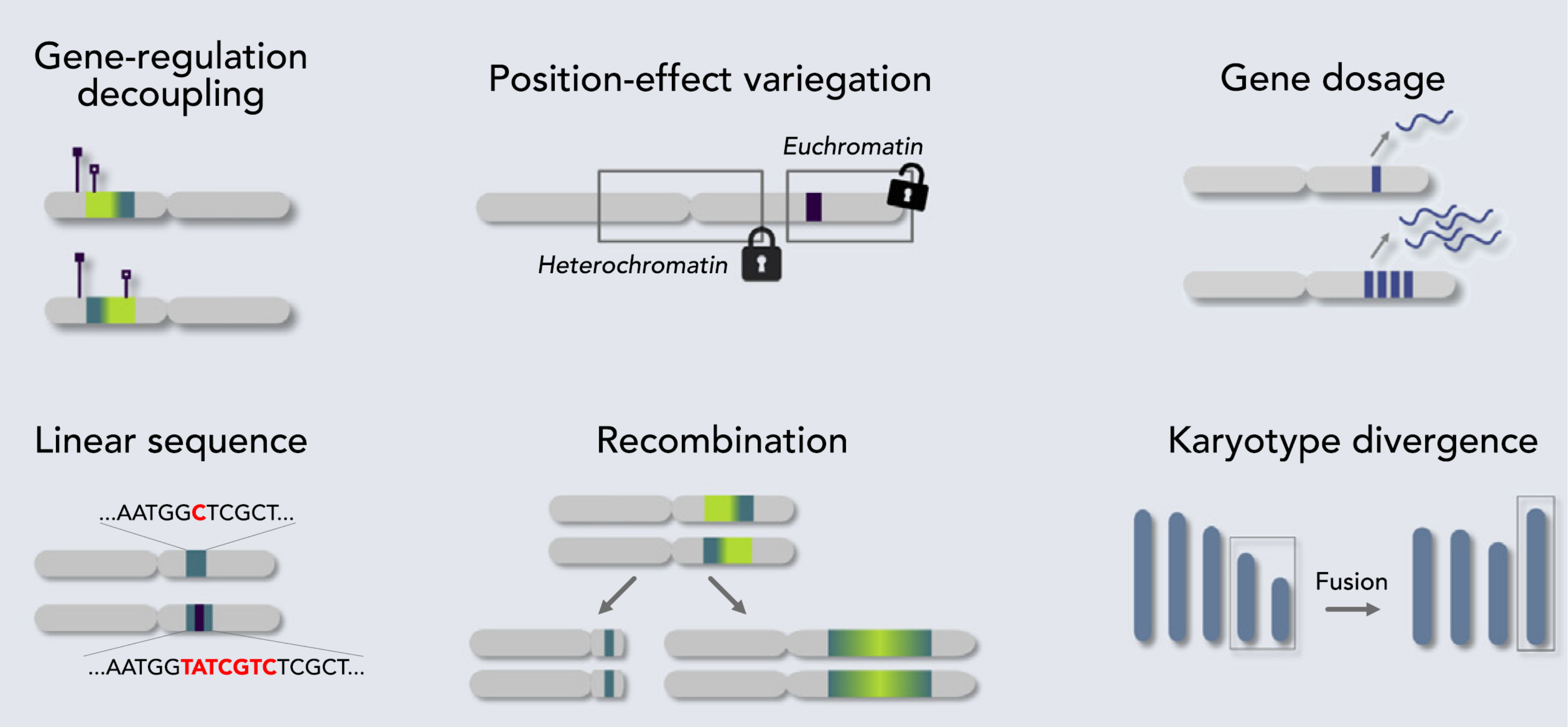
Jay et al. 2018, *Curr. Biol.*

Duplications



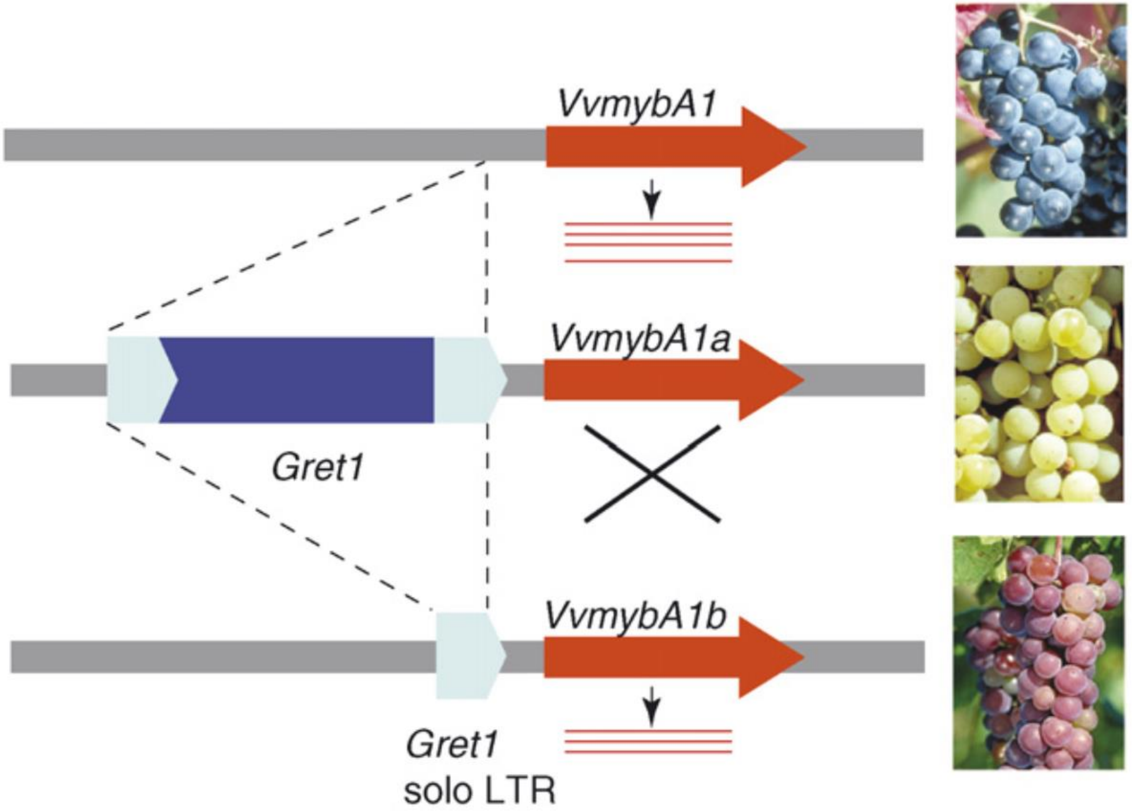
Mérot et al. 2020, *Trends Genet.*

High diversity of possible effects



Mérot et al. 2020, *Trends Genet.*

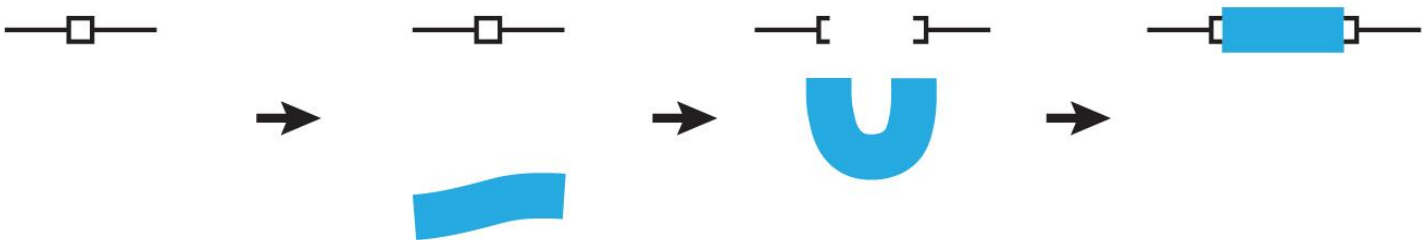
Part 1: Surprise



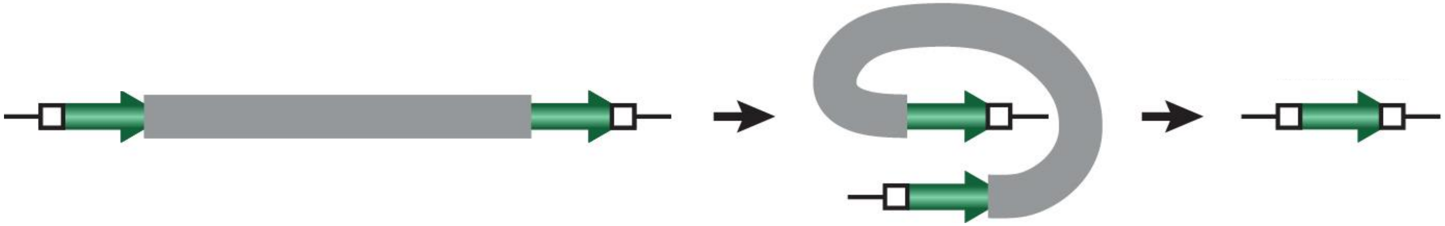
B) Covariation

Two key mechanisms of structural change

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

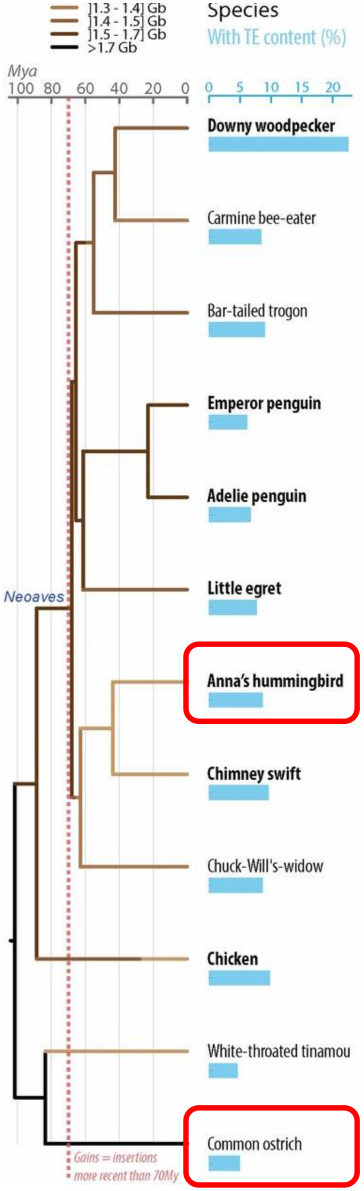


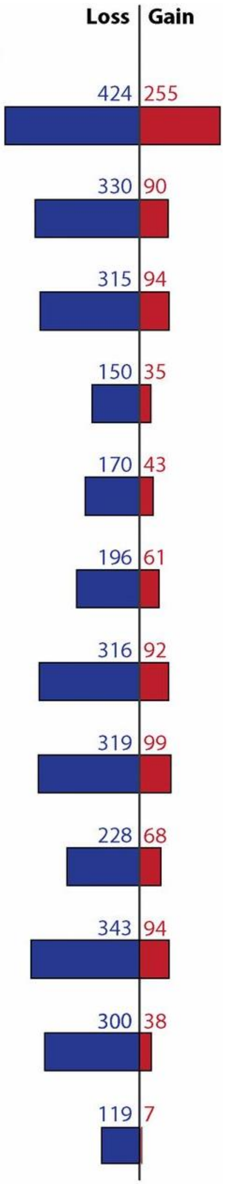
Non-allelic homologous recombination (**NAHR**)
(requires sequence homology)

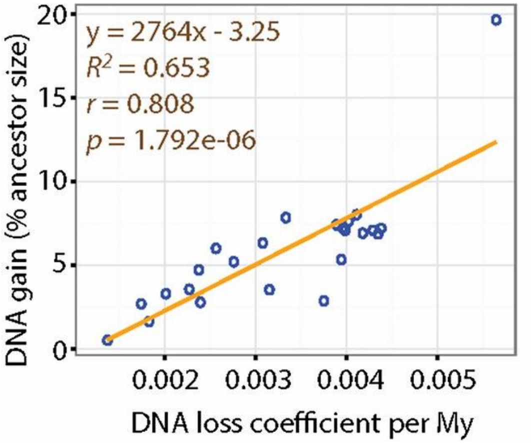


NHEJ correlates with frequency of DNA damage, NAHR correlates with frequency of (identical, large) repeats

Genome shrinking despite more TEs







Accordion model

Consider not only host popgen, but also TE popgen!

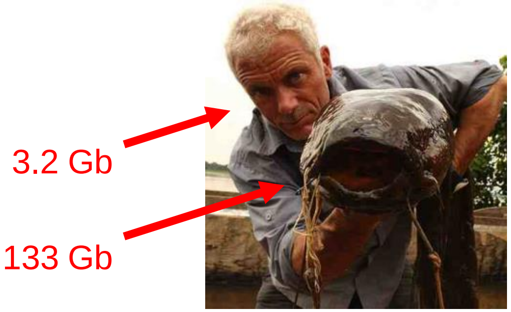
Kapusta et al. 2017, *PNAS*

Genome size and life history traits

Dynamic genome
(more TEs, fast shrinking)

Static genome
(fewer TEs, slow shrinking)

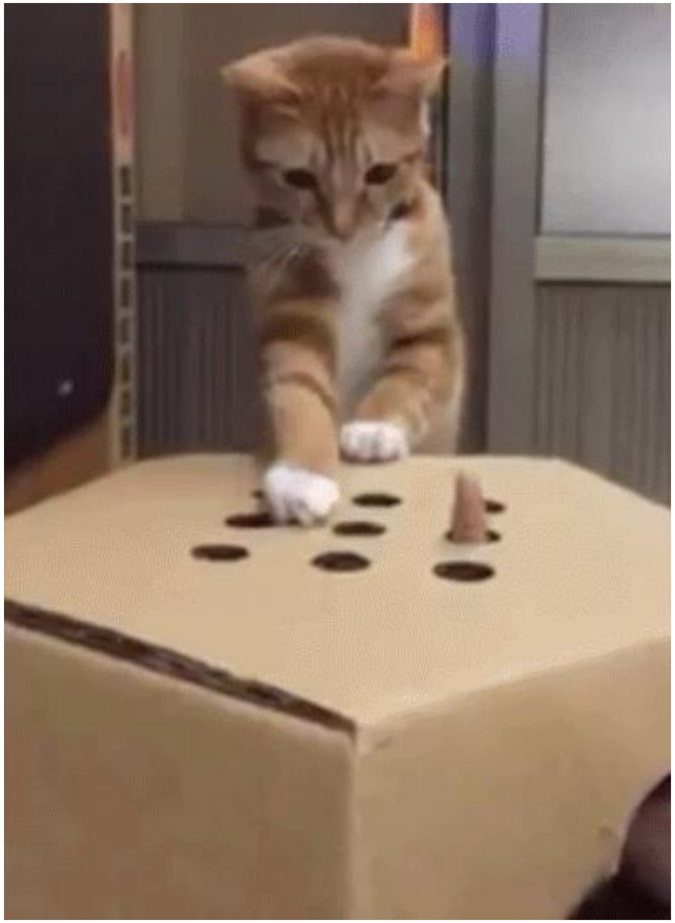
Adaptive processes are often invoked but remain difficult to prove (few high-quality genome assemblies and lack of popgen data)!



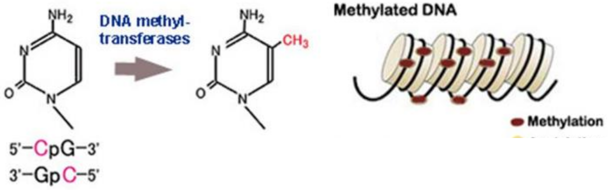
More context in [Suh 2021 TE lecture 5]

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

Genomes: whack-a-transposon

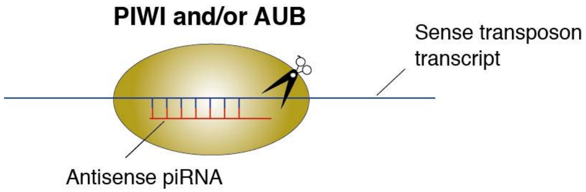


DNA methylation



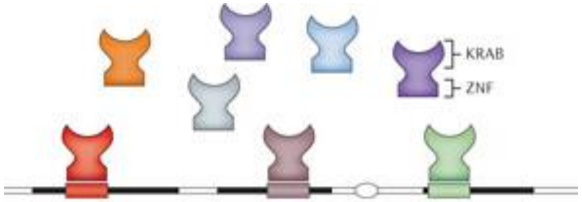
http://helicase.pbworks.com/w/page/17605615/DNA%20Methylation

piRNA pathway



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

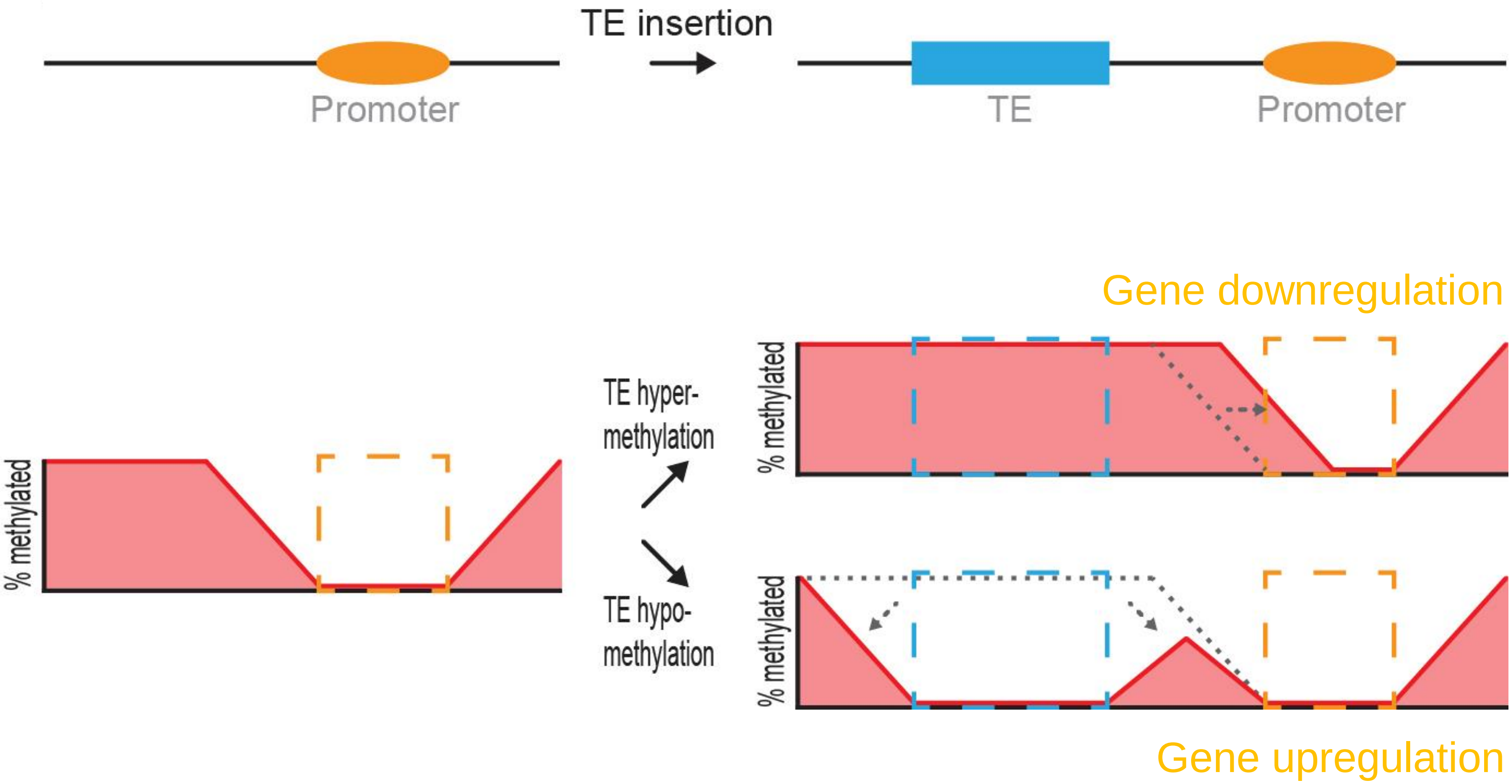
KRAB zinc-finger genes



[Feschotte & Gilbert 2012, *Nat. Rev. Genet.*](#)

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

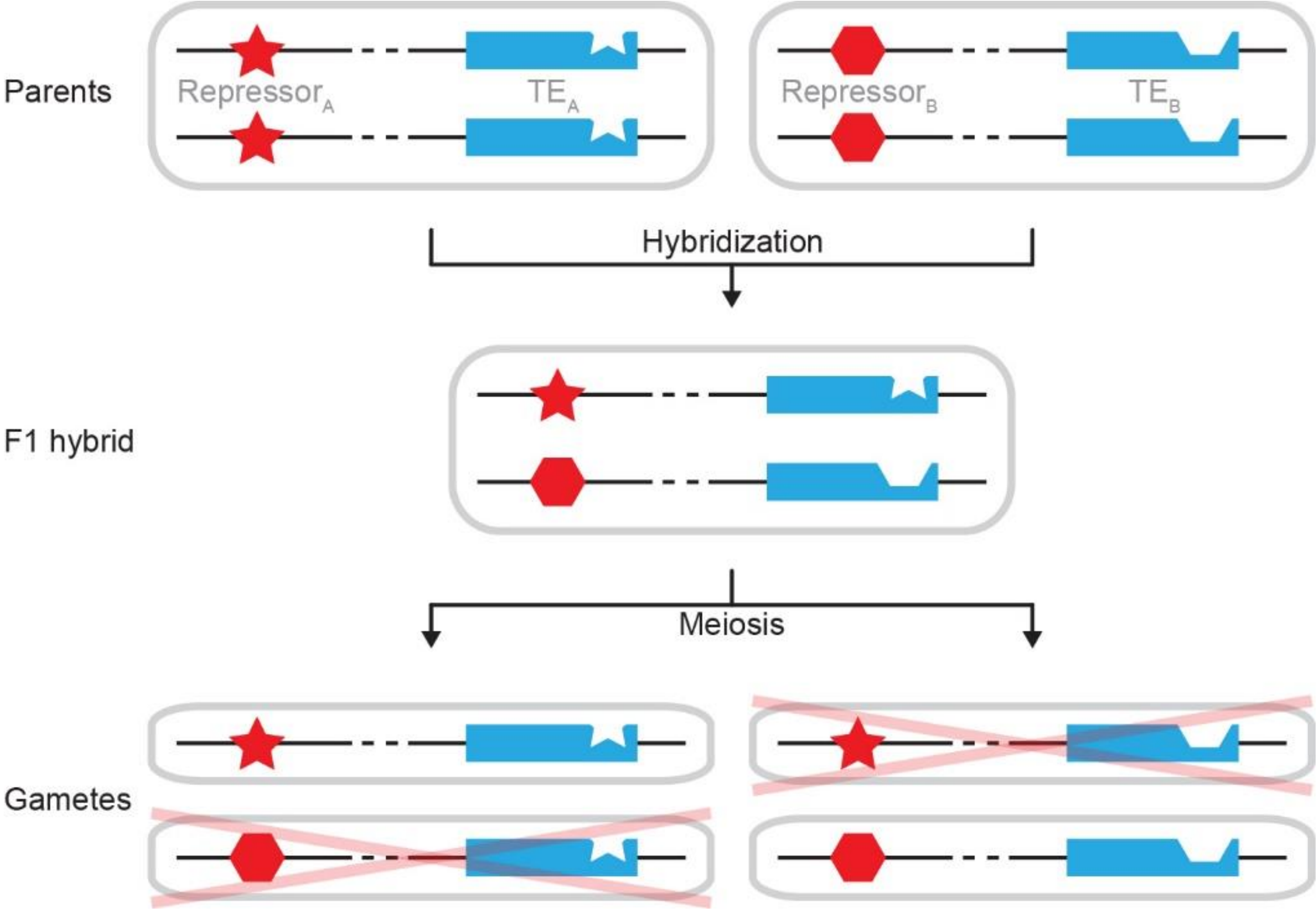
Covariation between (epi)mutation types



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

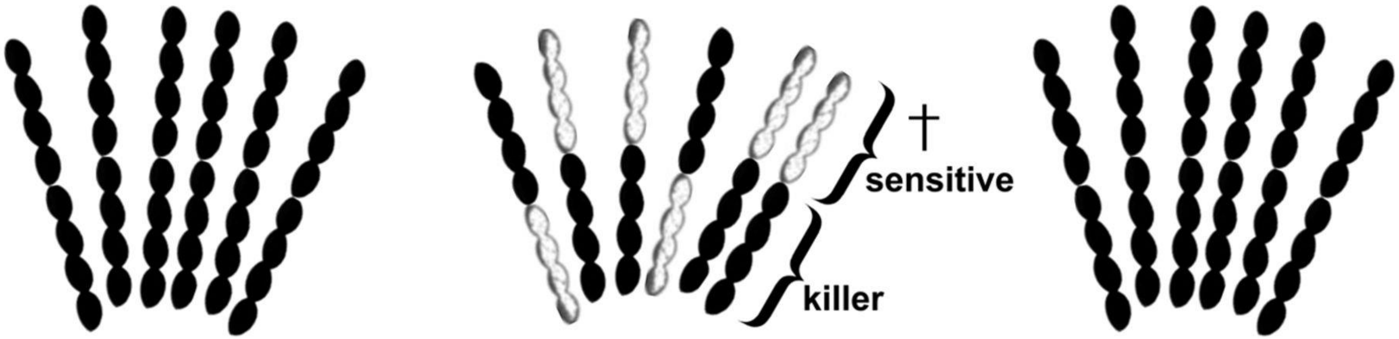
Weissensteiner & Suh 2019 in *Avian Genomics* book

Host–TE conflict and reproductive isolation



Weissensteiner & Suh 2019 in *Avian Genomics* book

Spore/sperm killing of some SVs

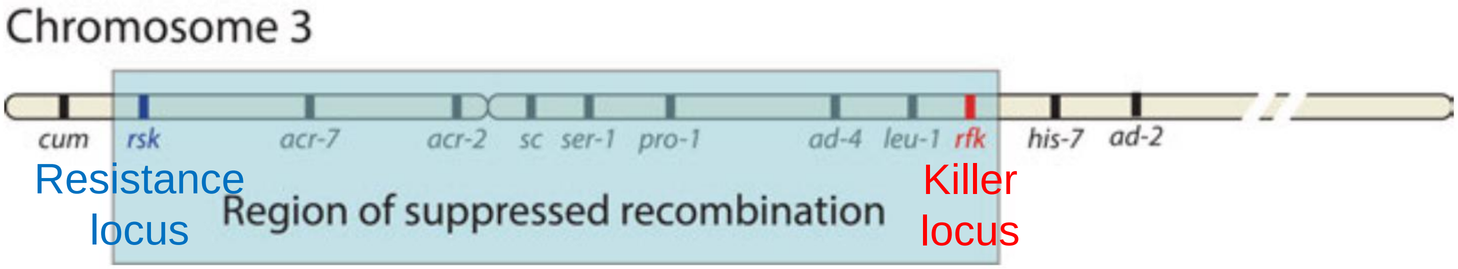


sensitive
x
sensitive

killer
x
sensitive

killer
x
killer

Spore sacs of red bread mold *Neurospora crassa*
(8 spores after meiosis and one mitotic division)



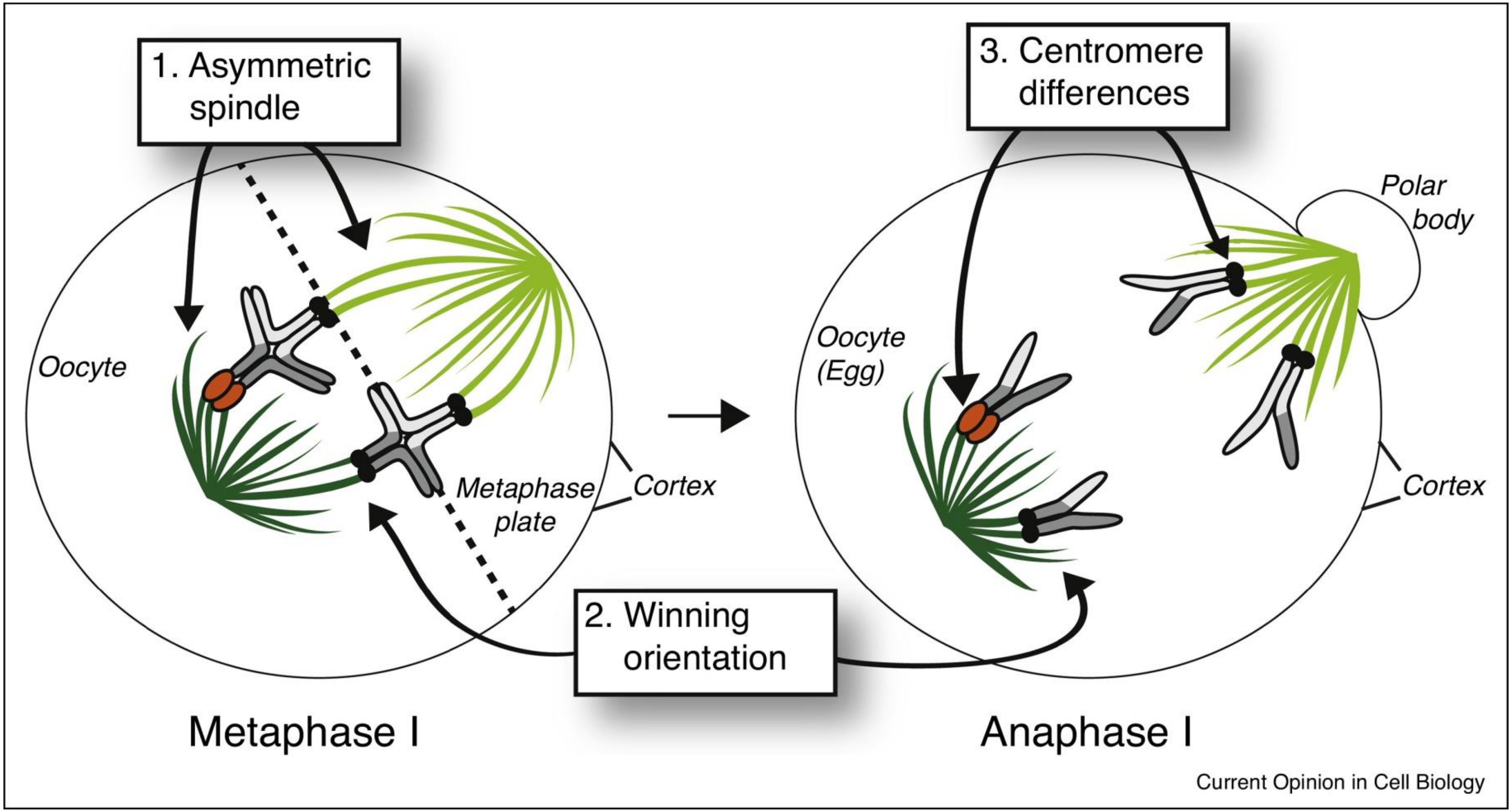
Resistance locus

Killer locus

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

Svedberg 2017 PhD thesis

Centromere drive of some SVs



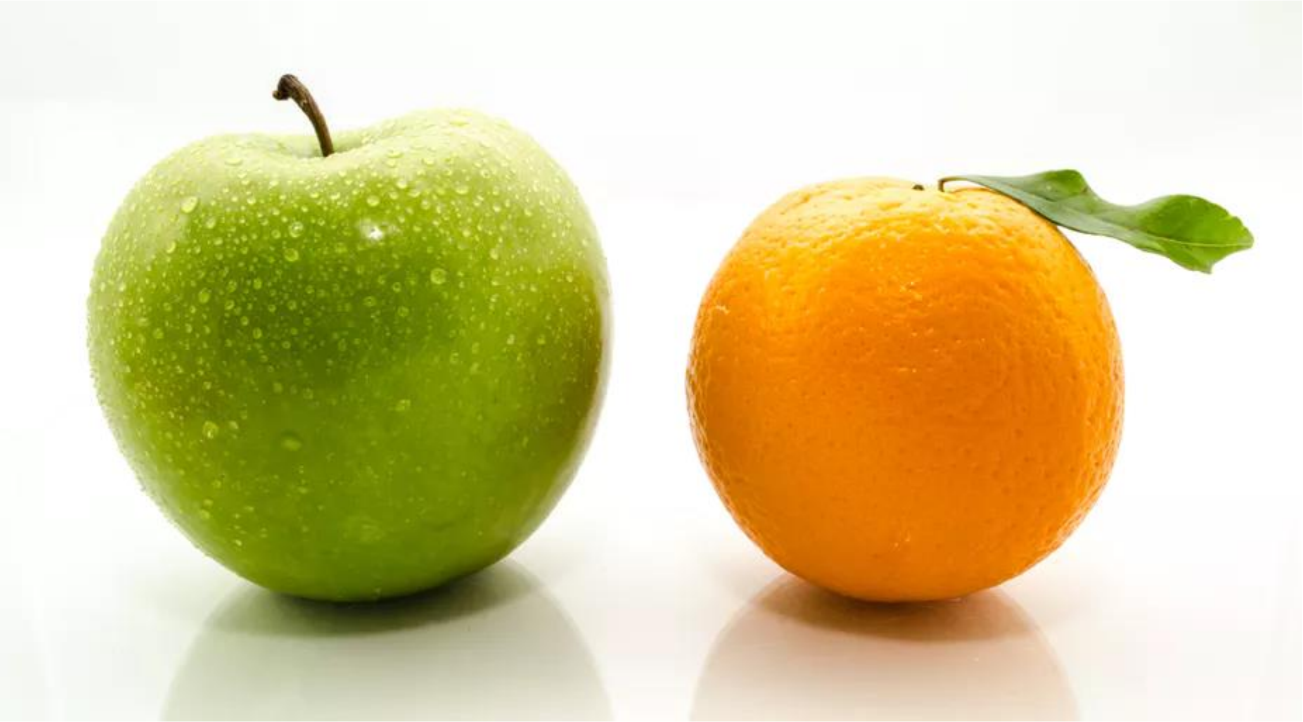
Current Opinion in Cell Biology

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

Kursel & Malik 2018, *Curr. Opin. Cell Biol.*

Questions?

Part 2: Frustration



A) Concepts and methods

What this lecture will not cover

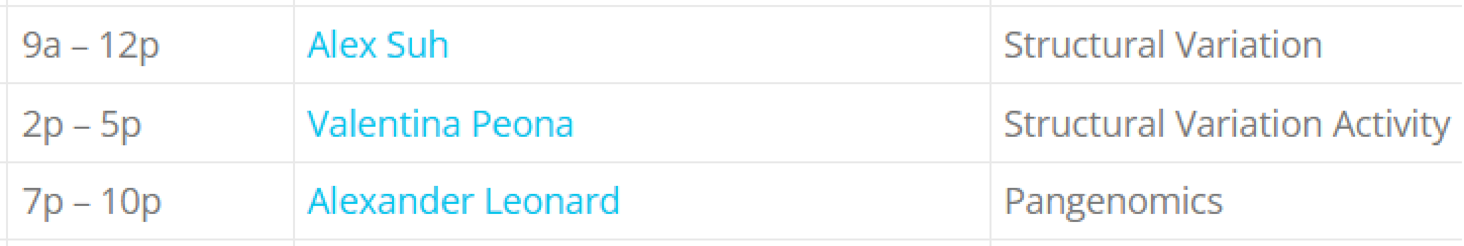
1. Genome assembly: What is (not) assembled?
   Primers: Peona et al. 2018, Peona et al. 2021, Rhie et al. 2021, Nurk et al. 2022

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!

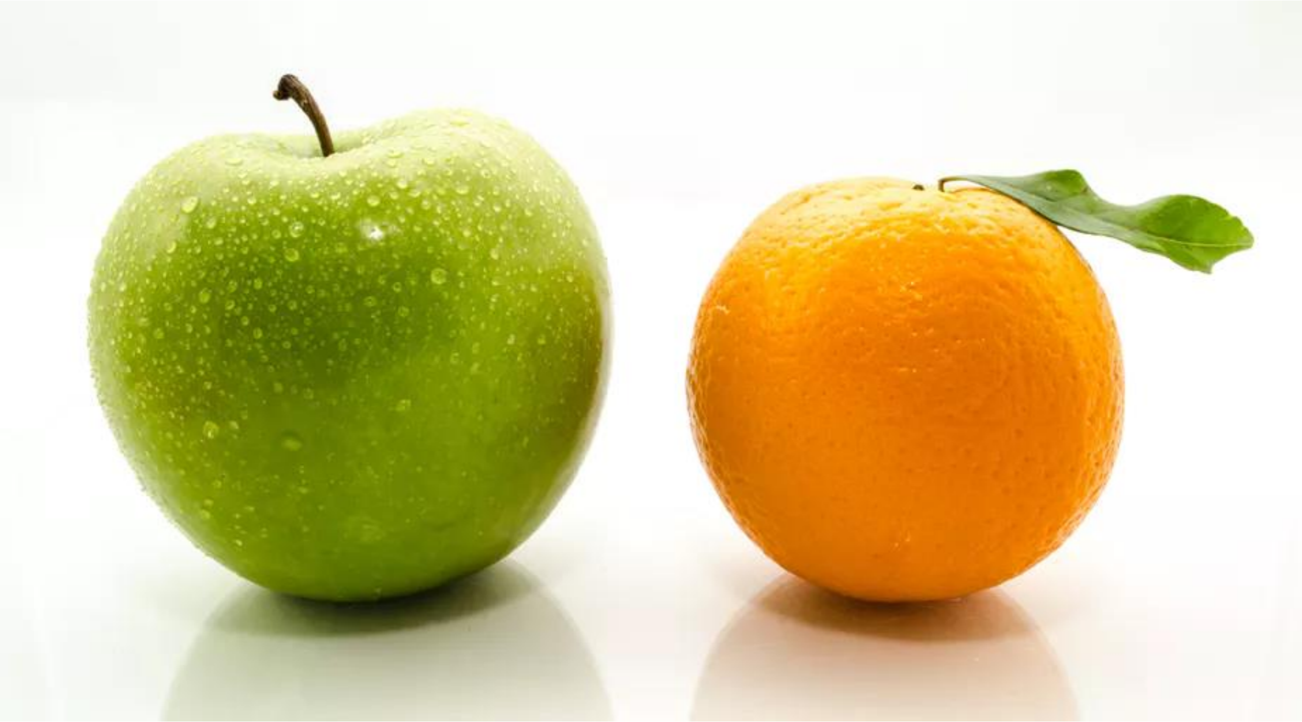
Valentina Peona



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

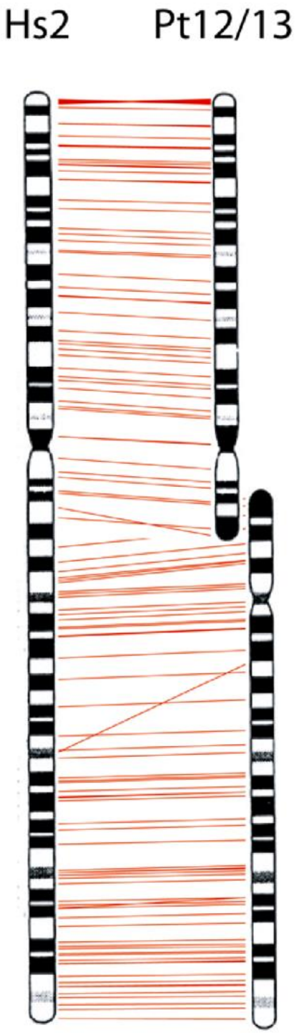


How can we make sure that what we see in our data is what we think it is?

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

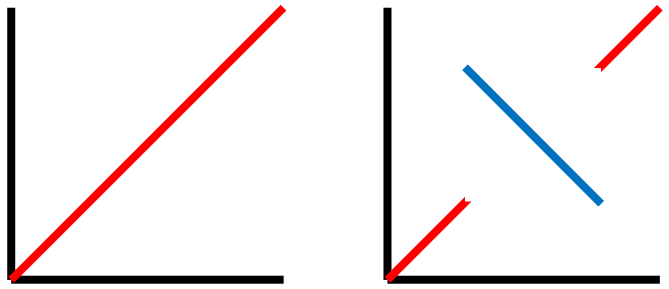
Terminology

Synteny vs. collinearity

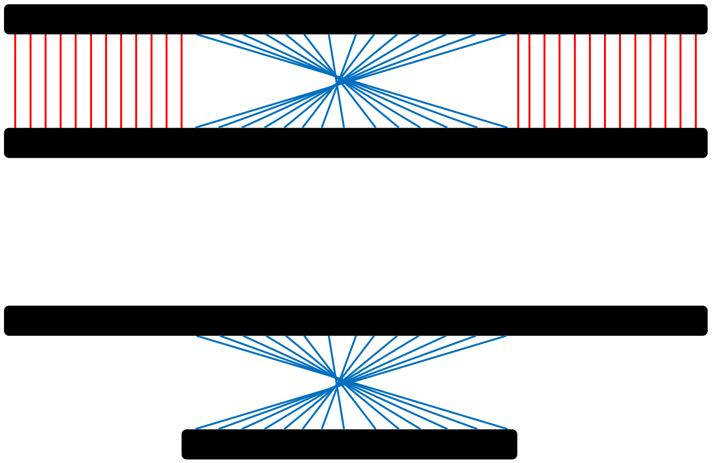


Feuk et al. 2005, *PLoS Genet.*

Dot plot



Pattern vs. process



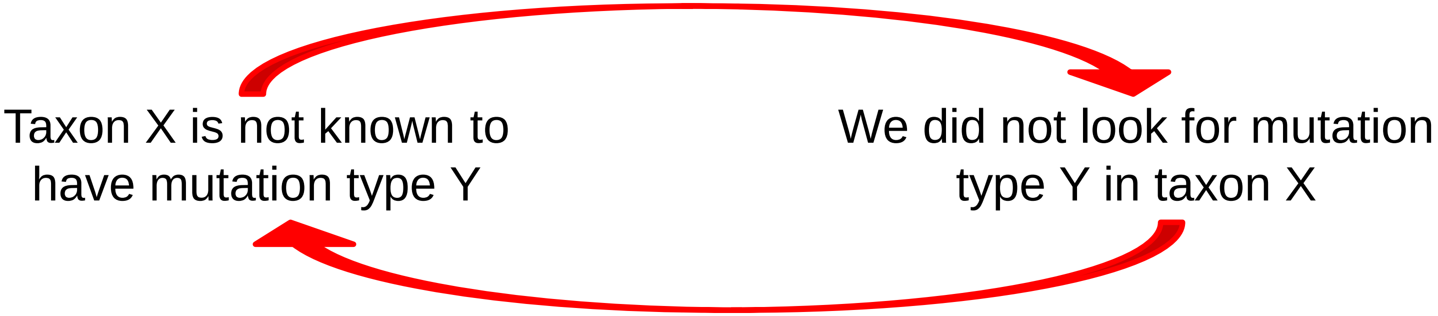
Beware of waves

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?

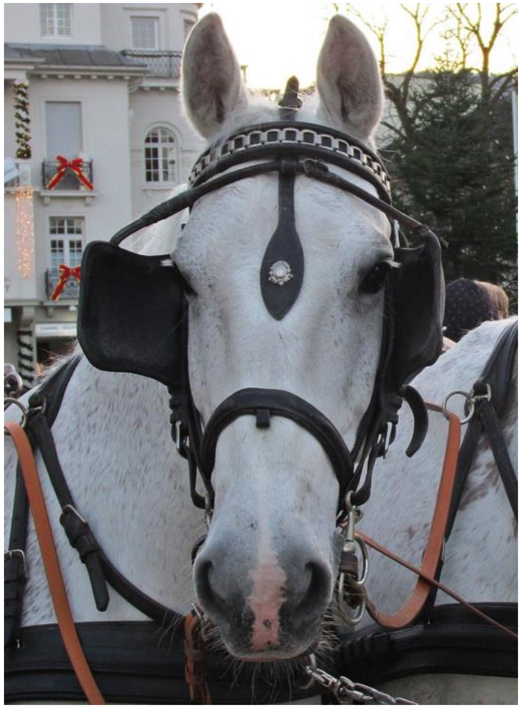


Taxon X is not known to have mutation type Y

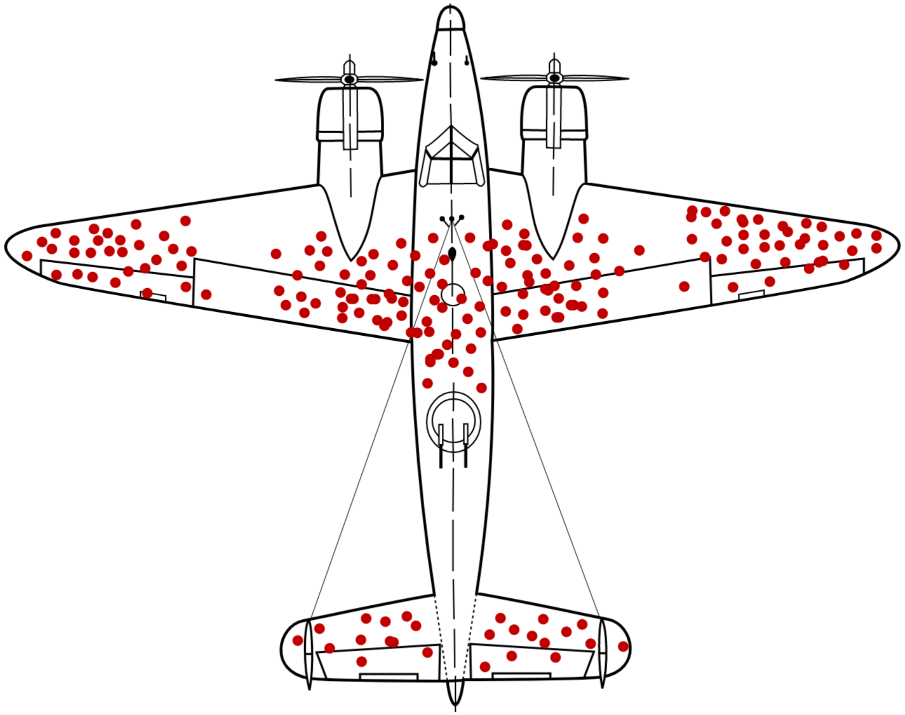
We did not look for mutation type Y in taxon X

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

Proximate: This asteroid caused diversification

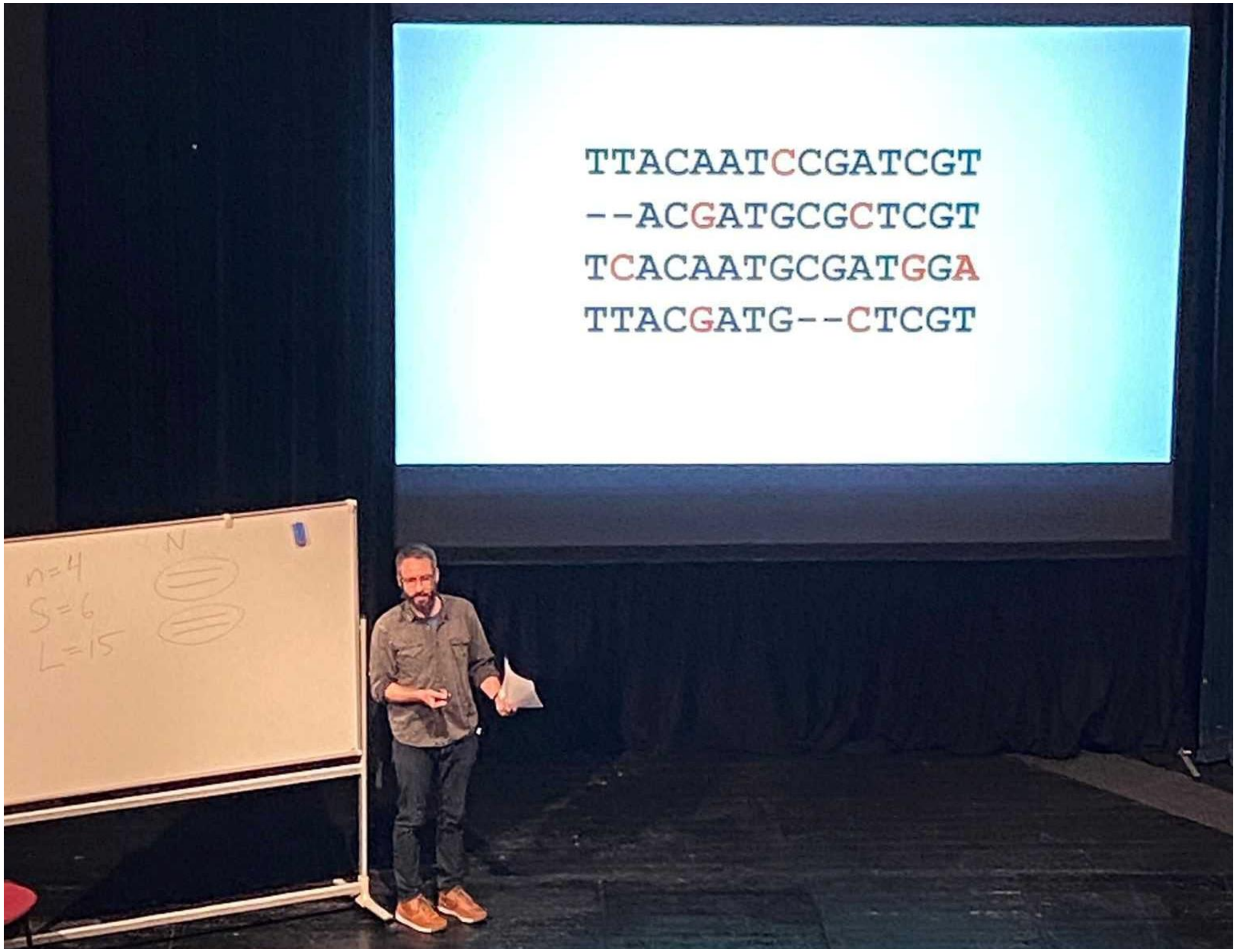
Ultimate: ~~Asteroids land to cause diversification~~
Asteroids land eventually

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

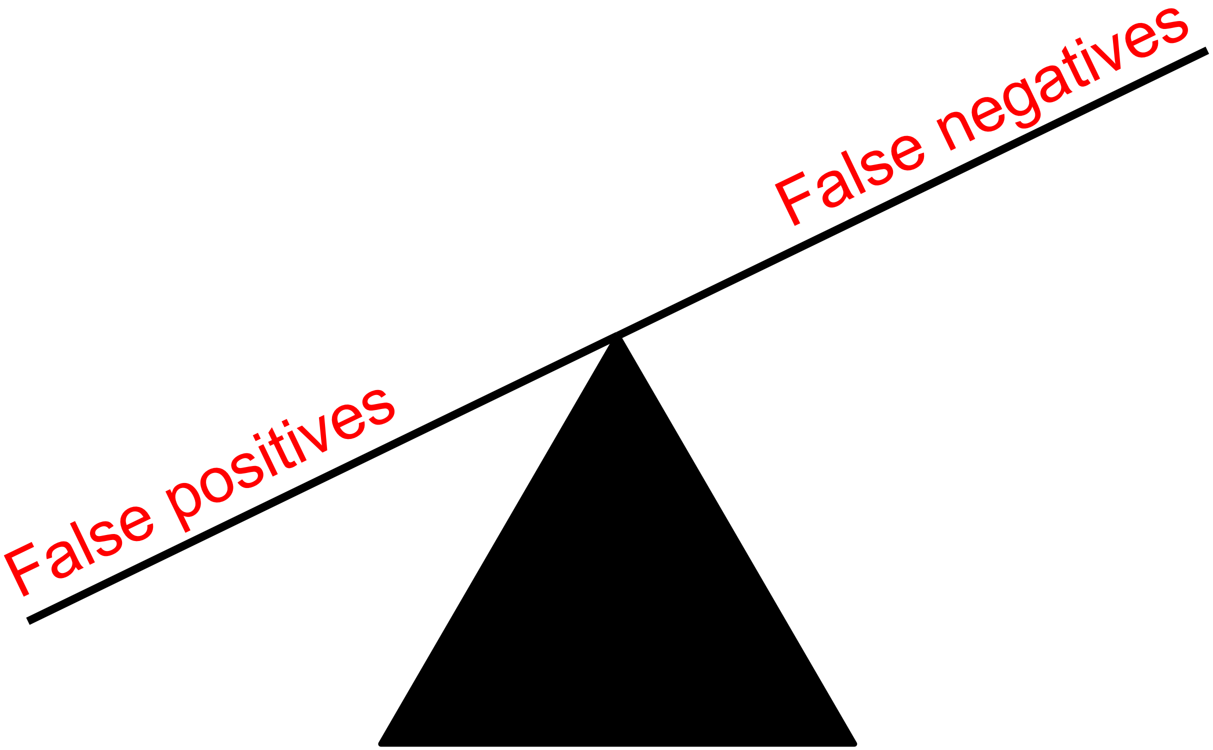


```
TTACAATCCGATCGT
--ACGATGCGCTCGT
TCACAATGCGATGGA
TTACGATG--CTCGT
```

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

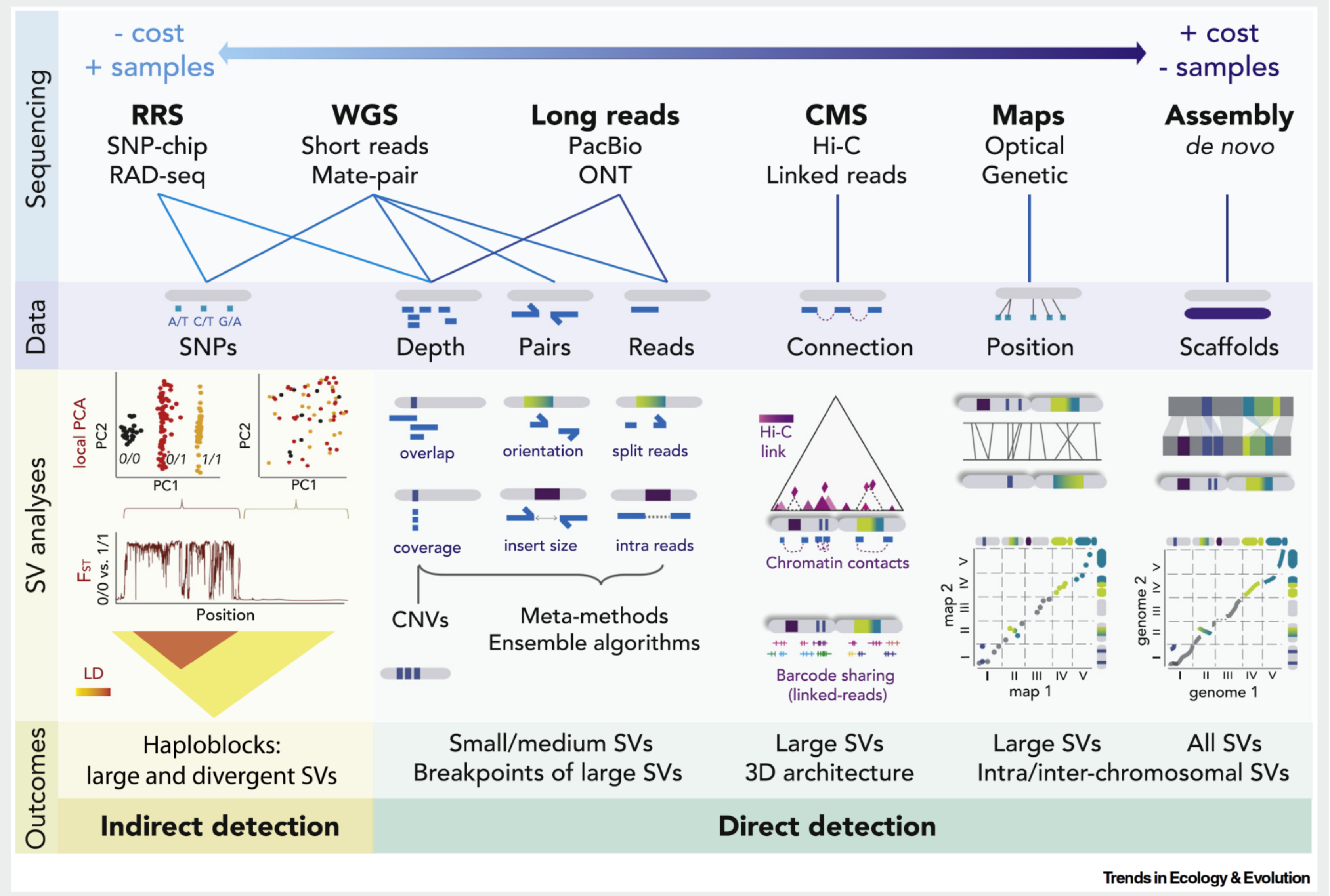
Matthew Hahn‘s January 21 lecture on Molecular Population Genomics

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.

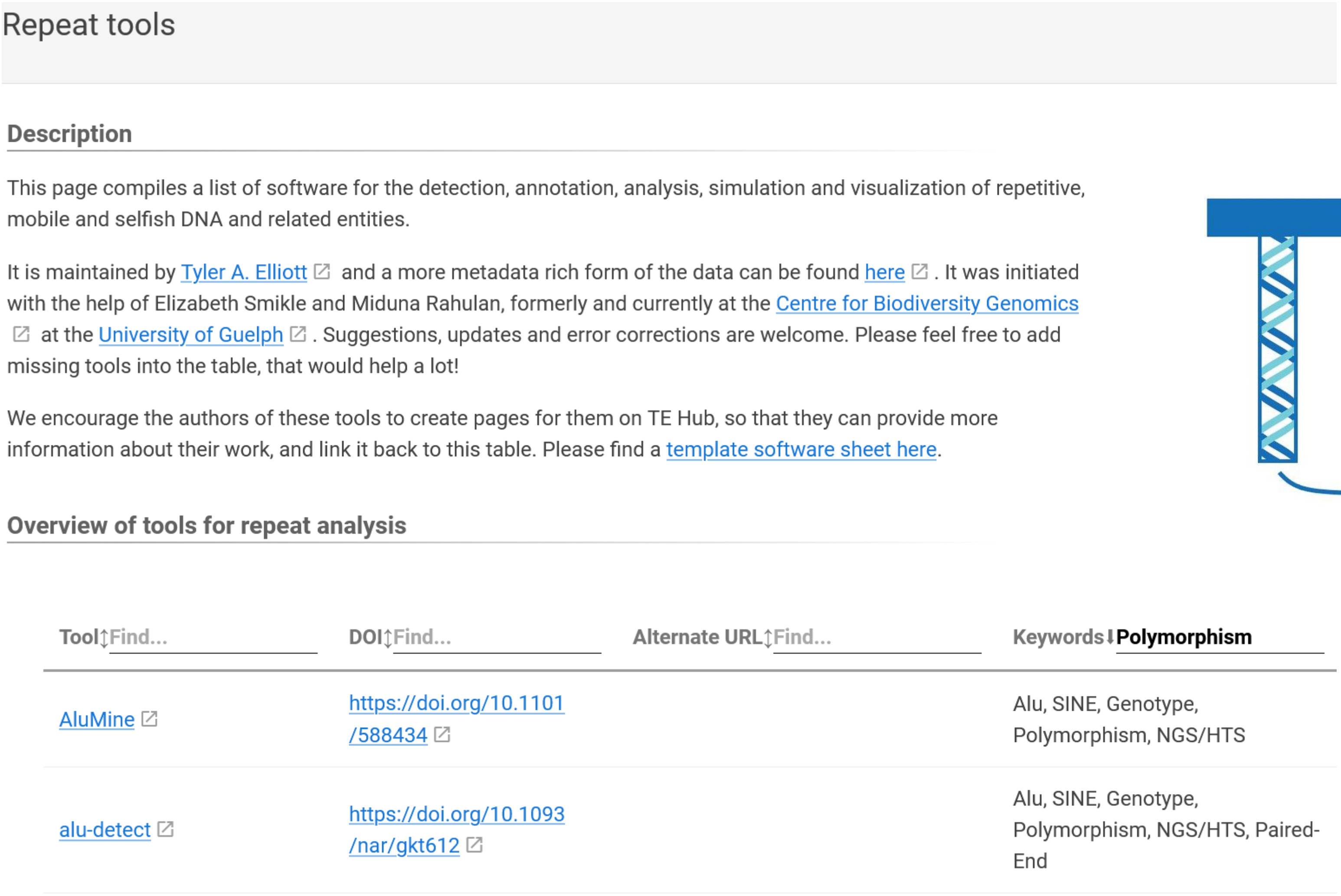
One approach to find them all?



Trends in Ecology & Evolution

Mérot et al. 2020, *Trends Genet.*

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.

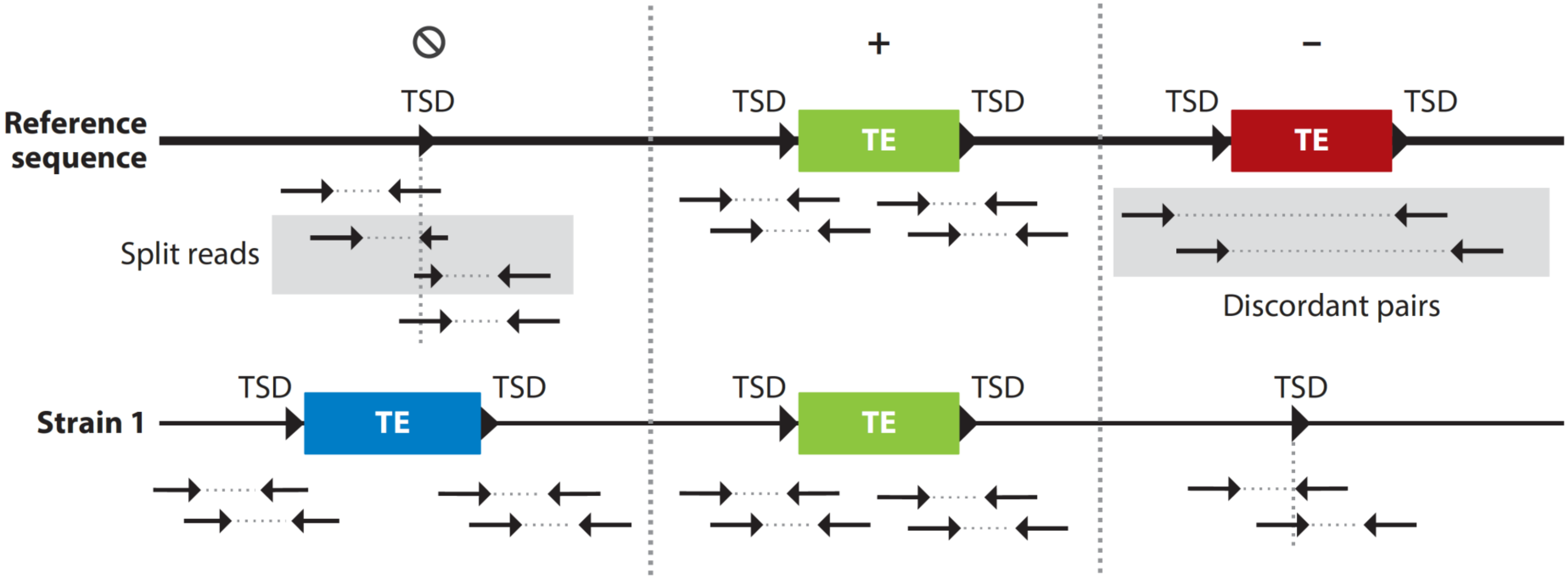
Overview of tools for repeat analysis

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*

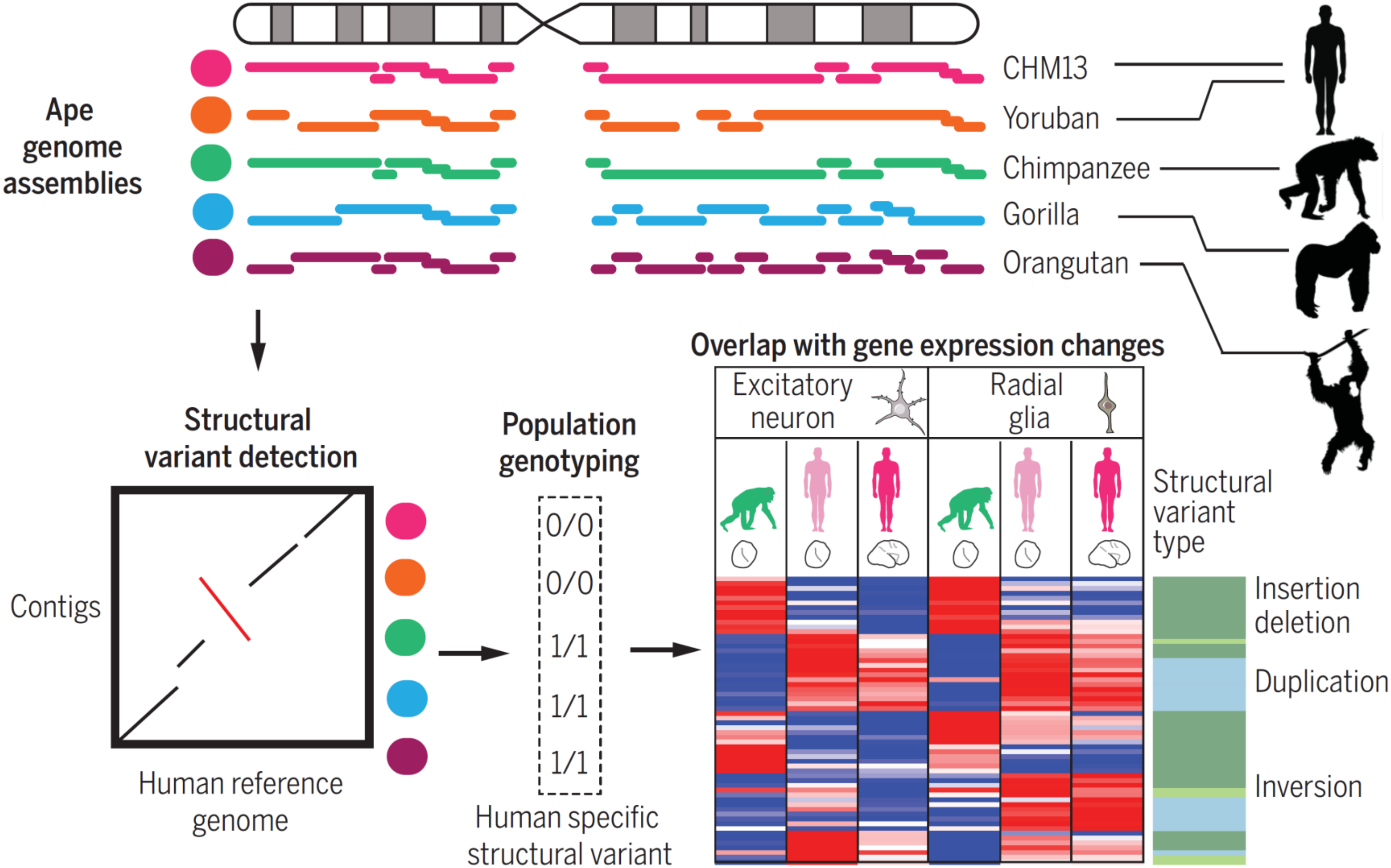
Read-based SV detection



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

Kronenberg et al. 2018, *Science*

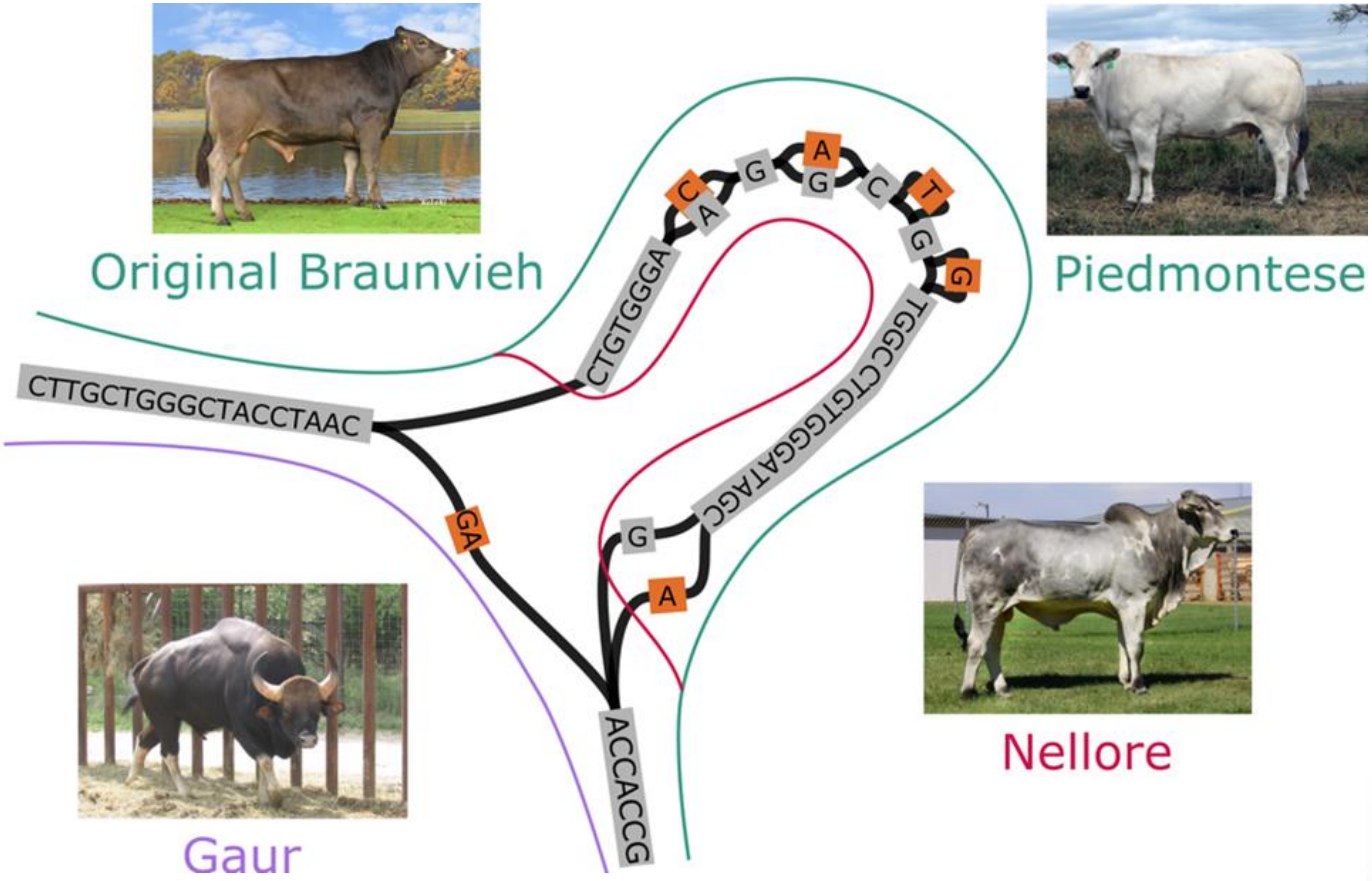
Assembly-based SV detection



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

Barrón et al. 2014, *Annu. Rev. Genet.*

Graph-based SV detection (pangenomics)



7p – 10p	Alexander Leonard	Pangenomics

Alexander Leonard

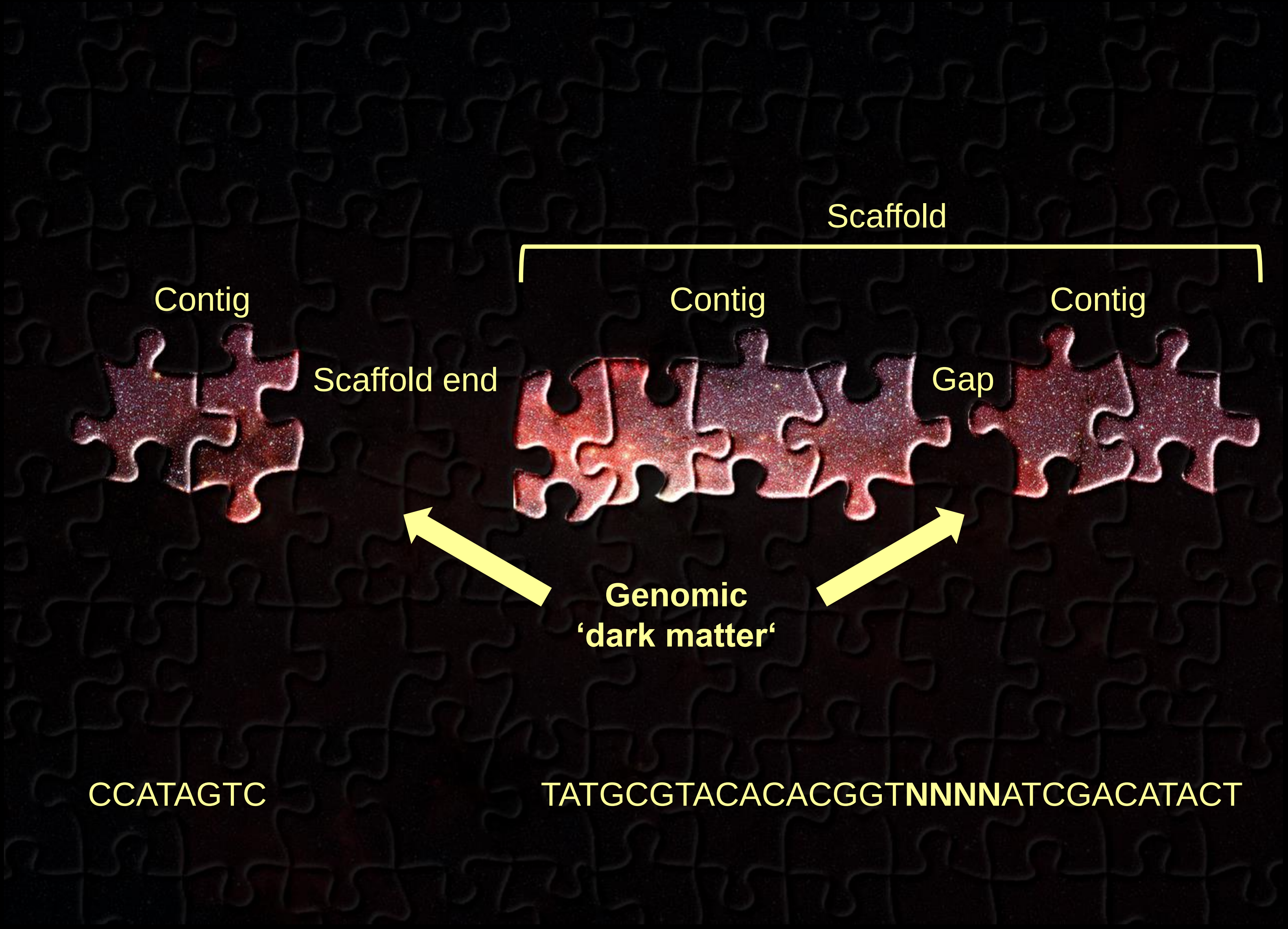
https://usys.ethz.ch/en/news-events/news/archive/2022/06/pangenomes-reveal-differences-between-cattle-and-their-wild-relatives.html

It could all be so easy

(if it wasn‘t for technological limitations)

Karyogram of a human male (Wikipedia)

Genomics: a big and messy puzzle



Scaffold

Contig

Contig

Contig

Scaffold end

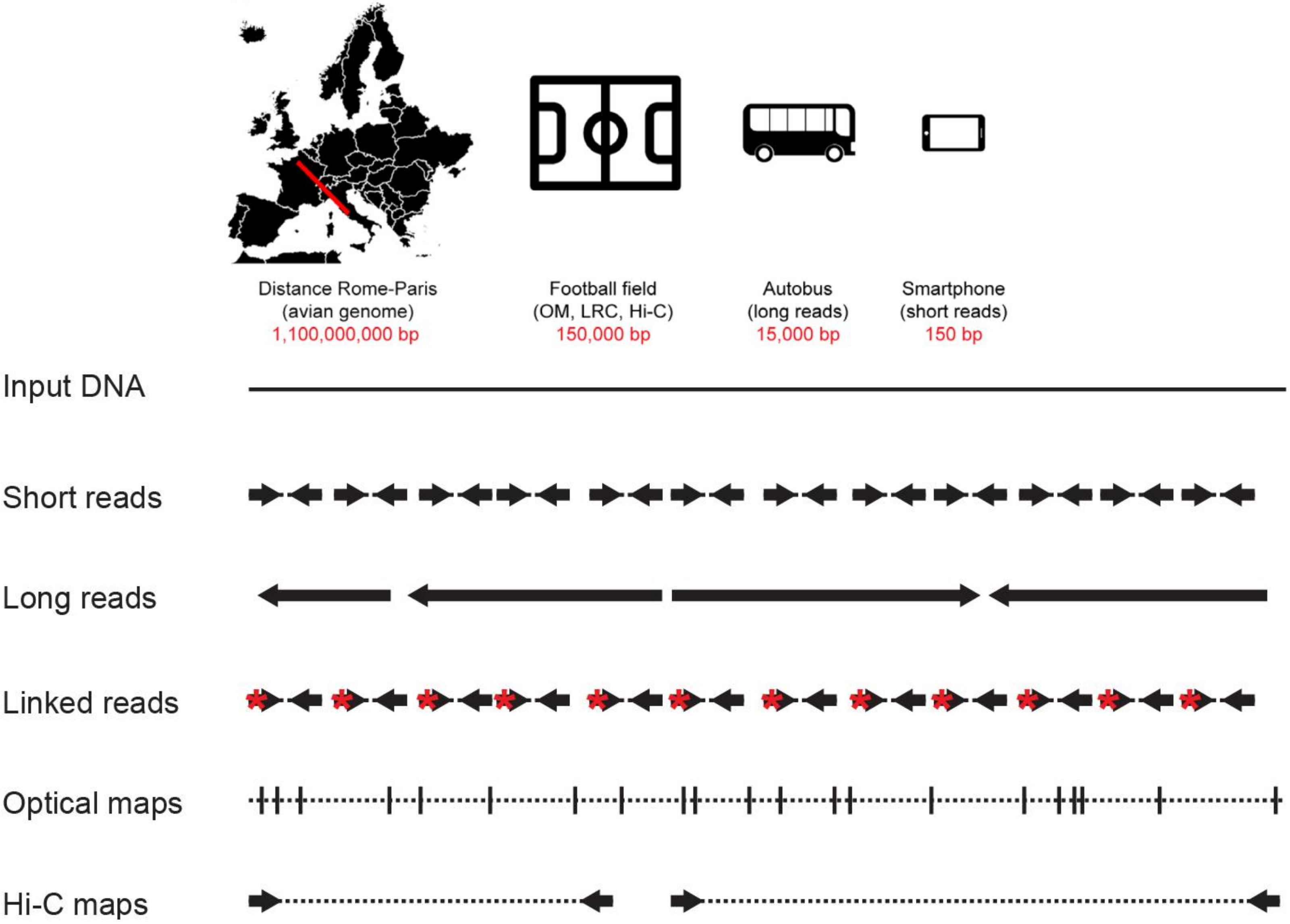
Gap

Genomic
‘dark matter‘

CCATAGTC

TATGCGTACACACGGT**NNNN**ATCGACATACT

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

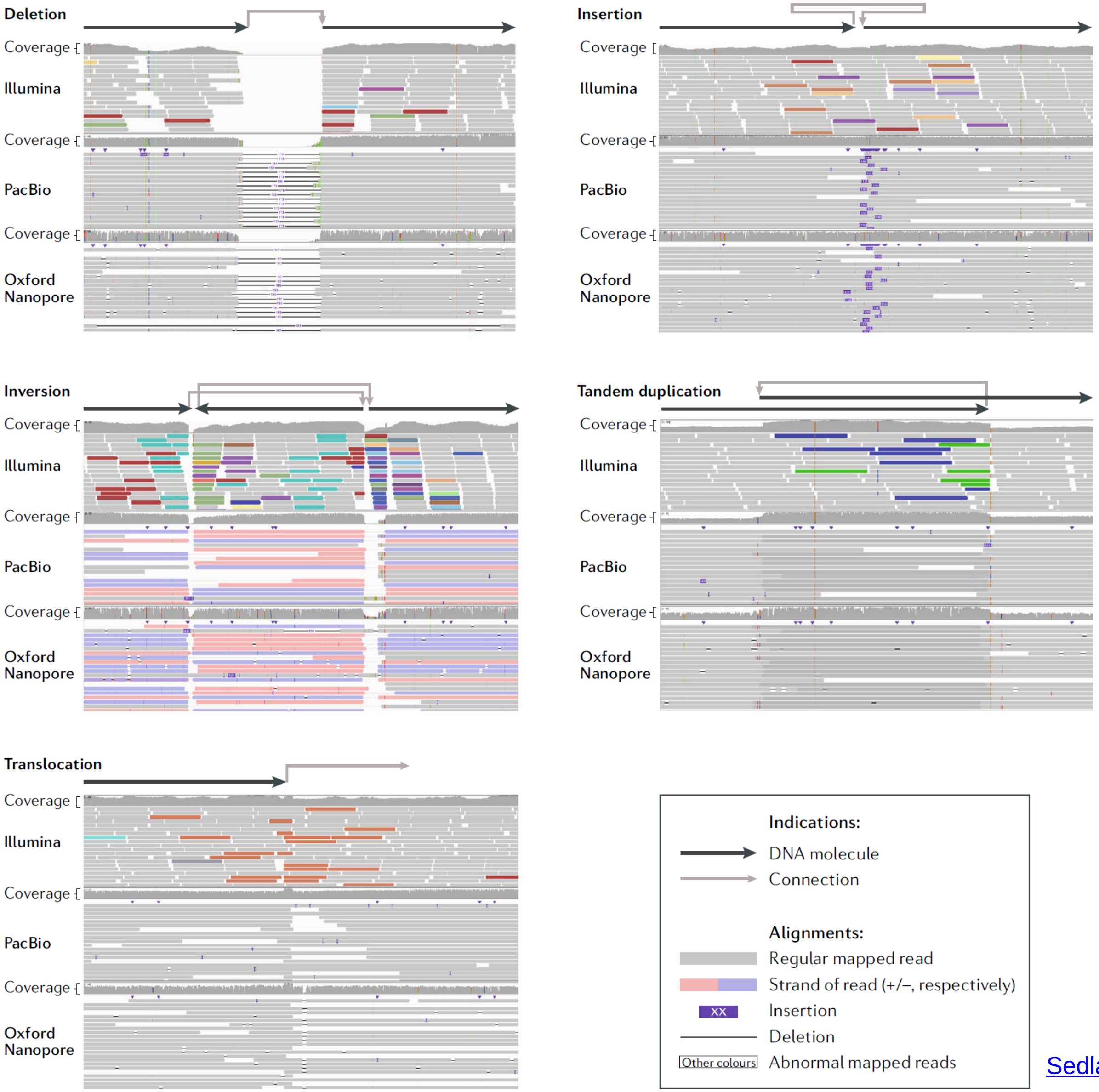
Linked reads

Optical maps

Hi-C maps

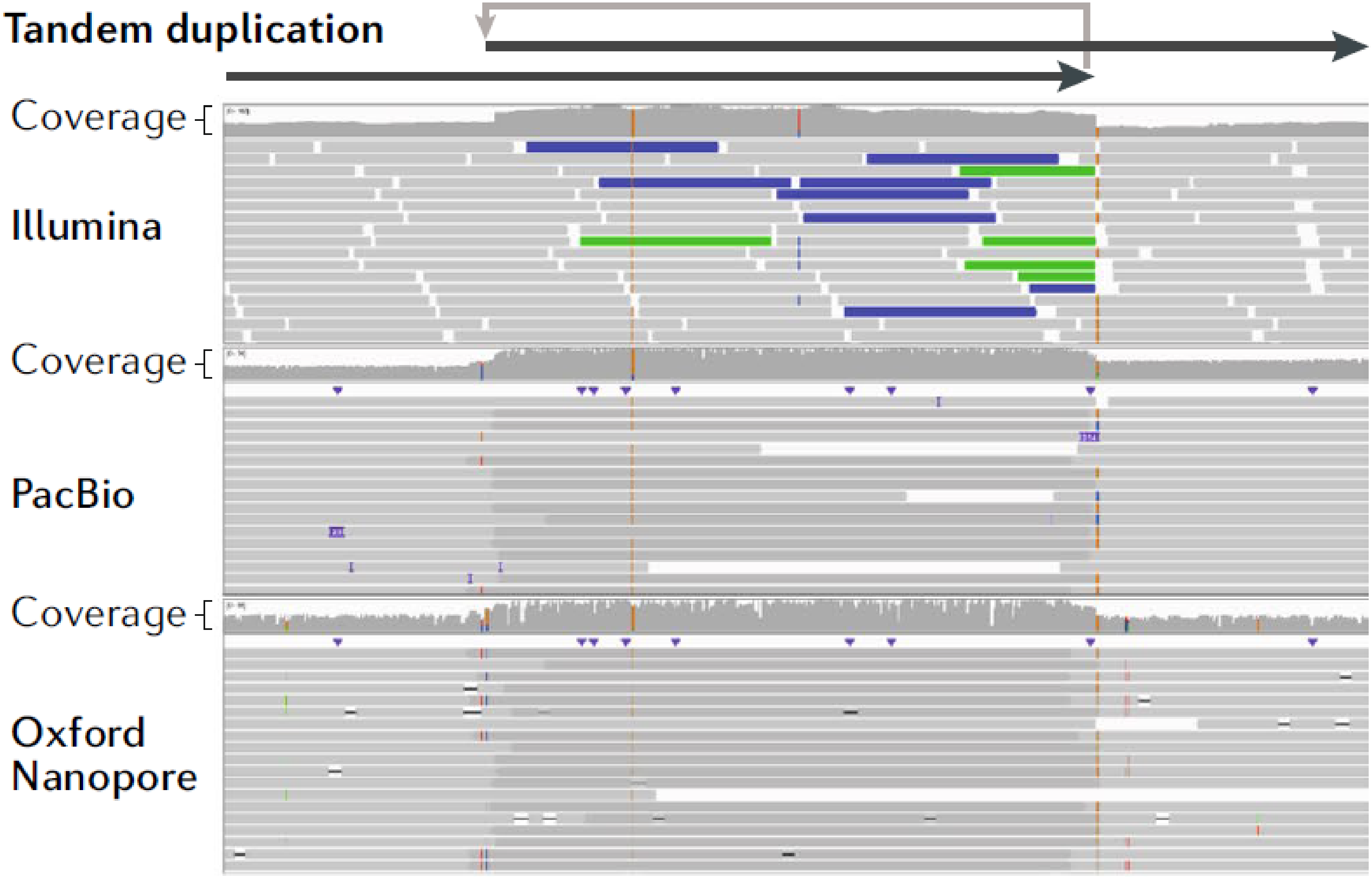
Peona et al. 2018, *Mol. Ecol. Res.*

SV mapping with longer and longer reads



Sedlazeck et al. 2018 *Nat. Rev. Genet.*

What does coverage variation tell us?

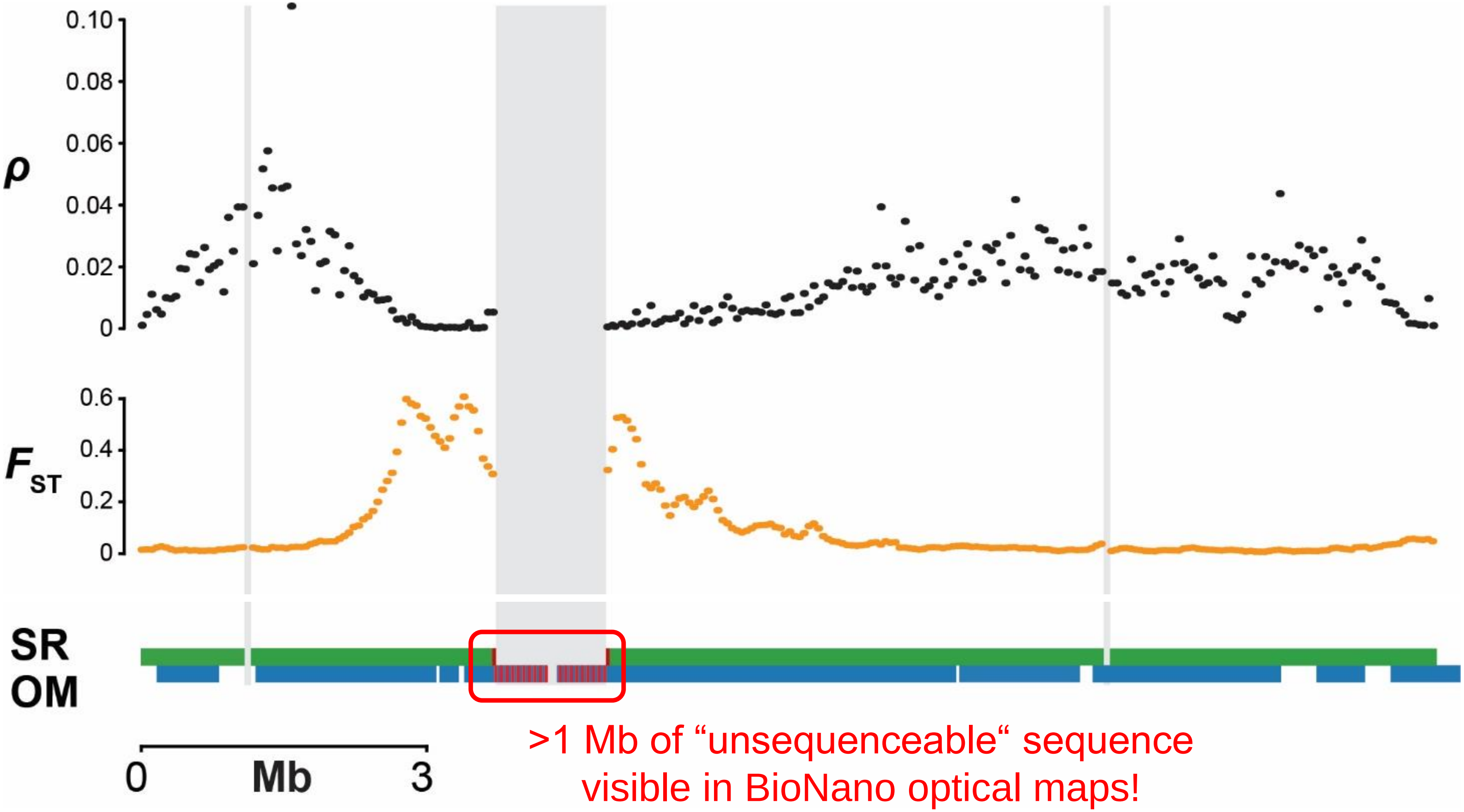


Tandem duplications are (usually) collapsed in assemblies!

Sedlazeck et al. 2018 *Nat. Rev. Genet.*

Not all gaps are equal

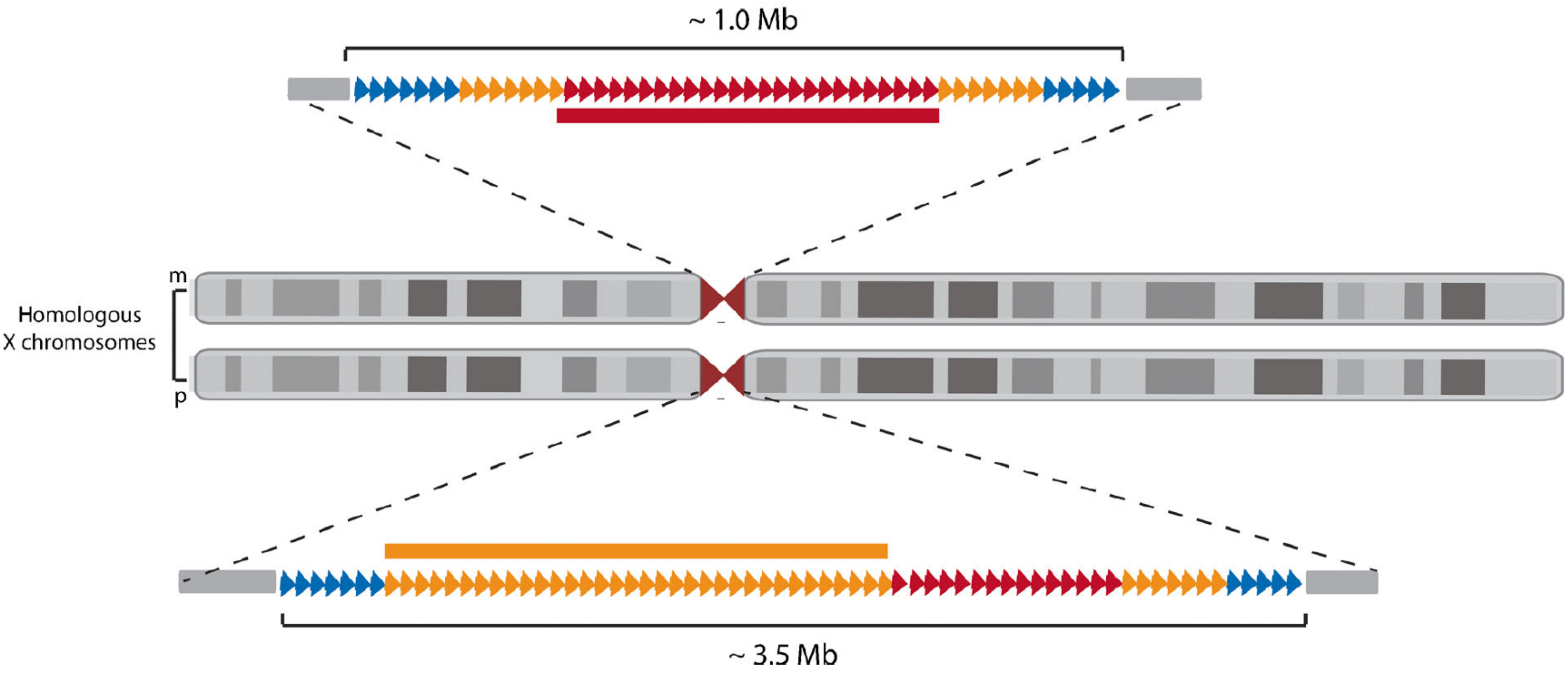
Chromosome 18 of hooded/carrion crow



>1 Mb of “unsequenceable“ sequence visible in BioNano optical maps!

Weissensteiner et al. 2017, *Genome Res.*

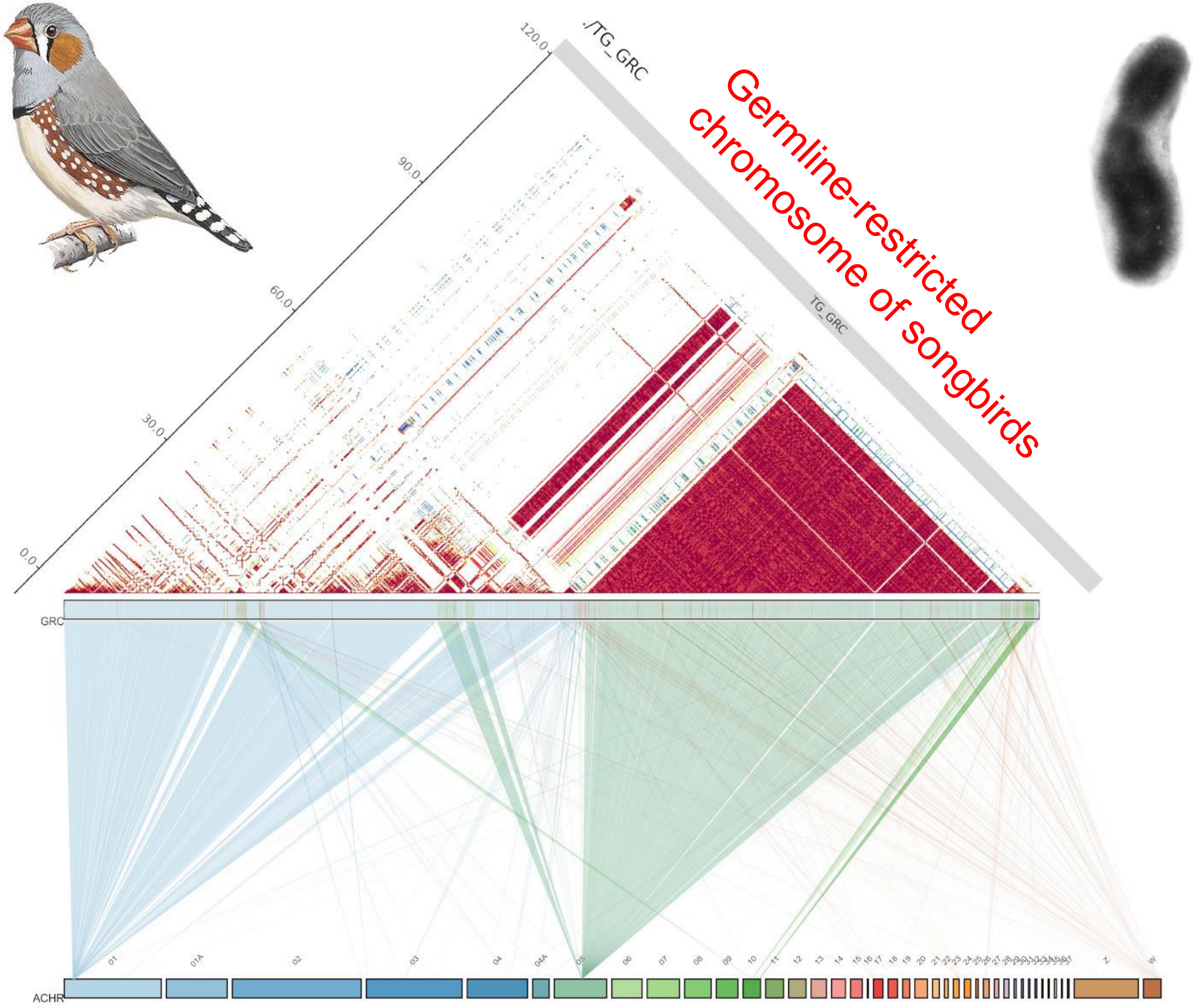
Centromeres are very, very repetitive …



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

Miga 2015, *Chromosome Res.*

… and so are some chromosomes



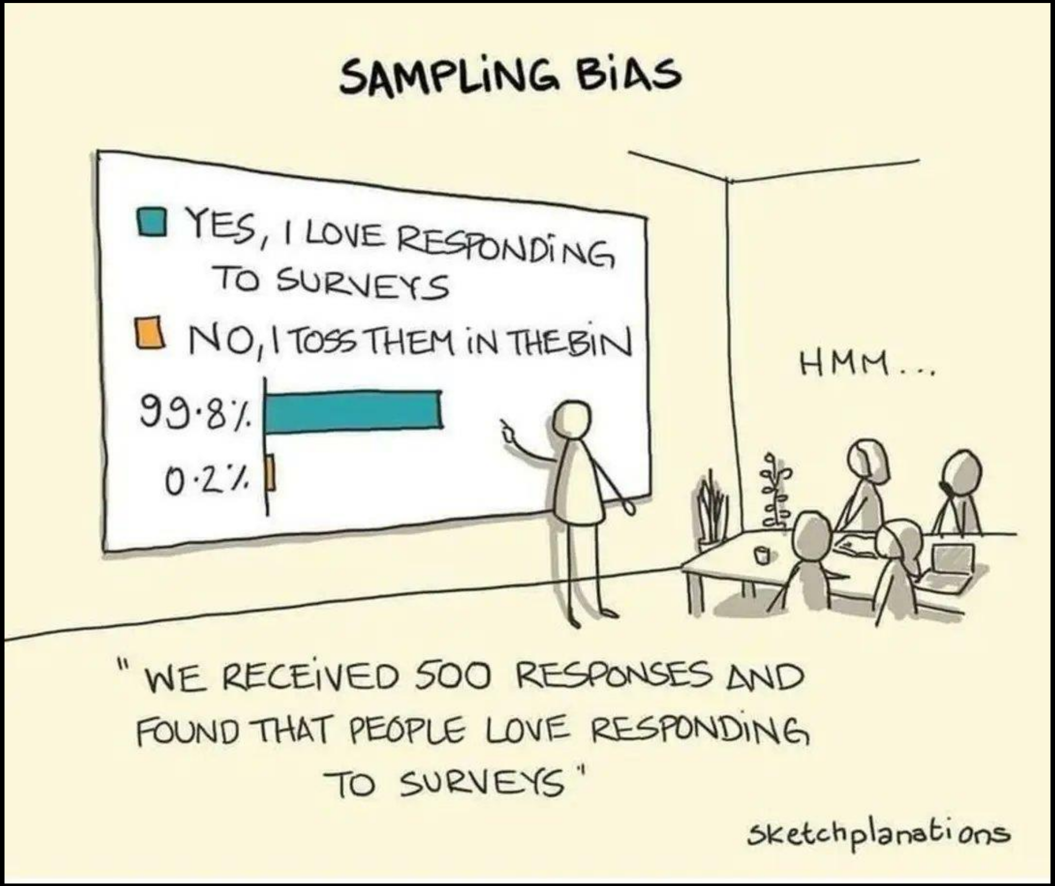
Germline-restricted chromosome of songbirds

Autosomes and sex chromosomes

Ruiz-Ruano et al., *in prep.*

Questions?

Coffee break (20 minutes)

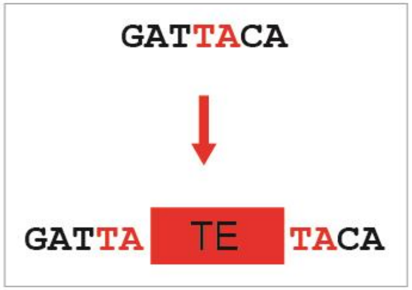

SAMPLING BIAS
🟦 YES, I LOVE RESPONDING TO SURVEYS
🟧 NO, I TOSS THEM IN THE BIN
99.8%
0.2%
HMM...
"WE RECEIVED 500 RESPONSES AND FOUND THAT PEOPLE LOVE RESPONDING TO SURVEYS"
sketchplanations

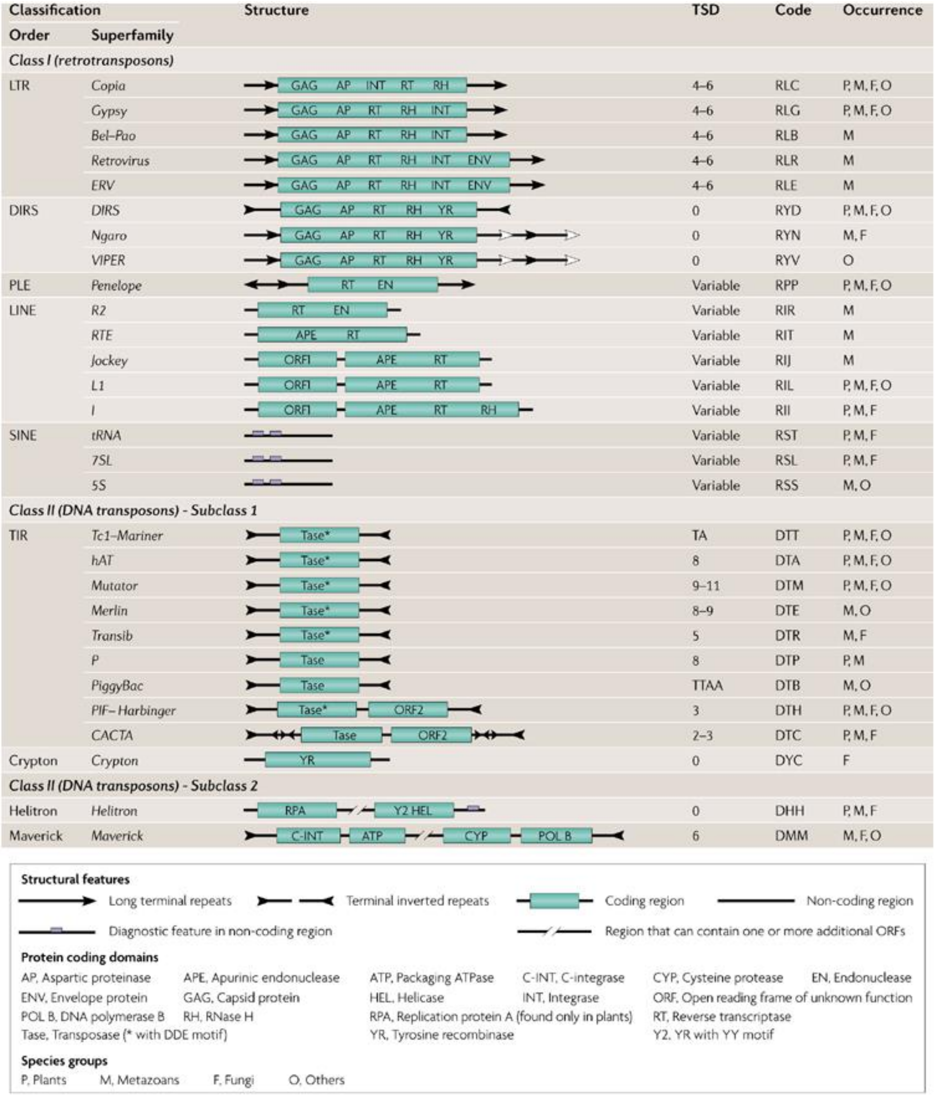
Task: Form random groups of 3 and discuss 1) what SVs you want to study, 2) what SVs you can study, and 3) what data you need to be less frustrated.

Part 2: Frustration

B) Biology and more concepts

Transposable elements are very diverse





Classification Order	Superfamily	Structure	TSD	Code	Occurrence
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	ORFI APE RT	Variable	RIJ	M
	L1	ORFI APE RT	Variable	RIL	P, M, F, O
	I	ORFI 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

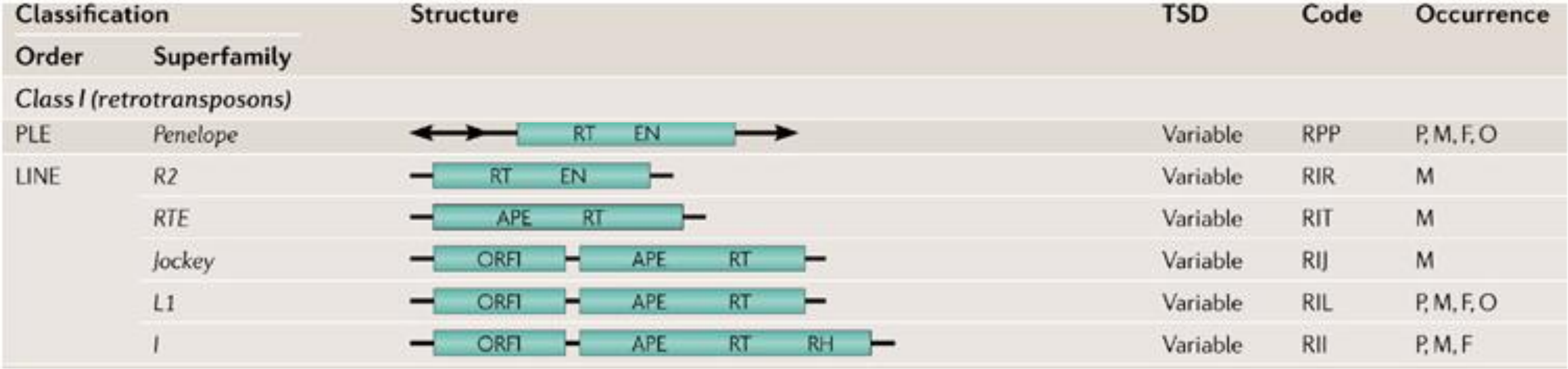
Nature Reviews | Genetics

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

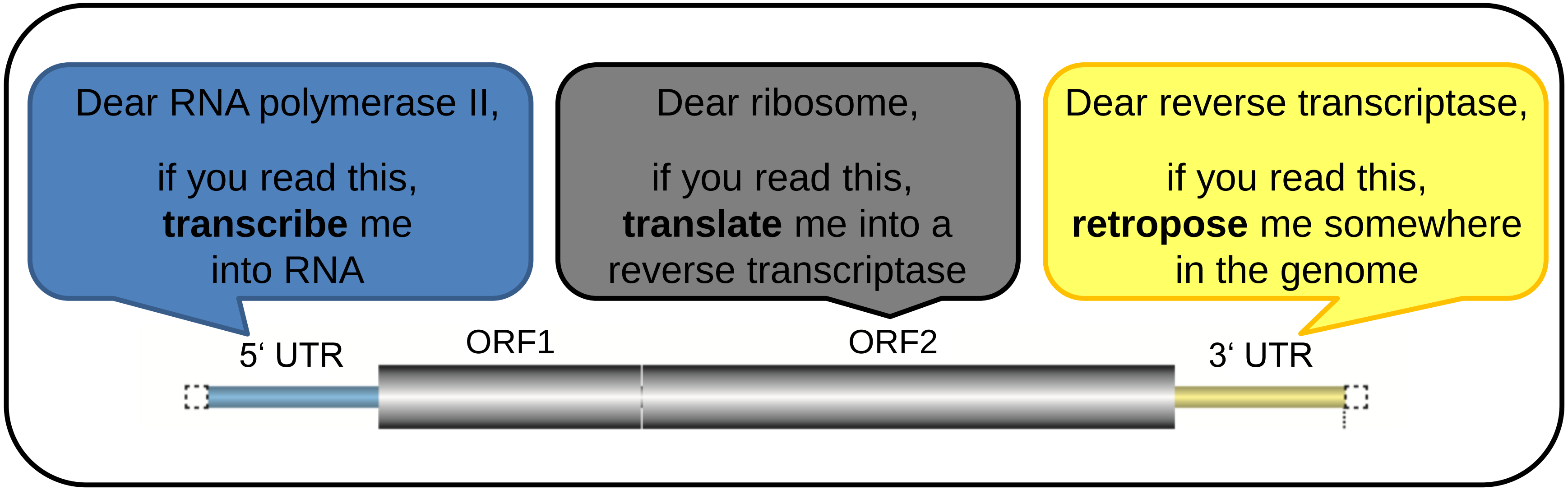
Weirder TEs in Suh 2021 TE lecture 1

Wicker et al. 2008, *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,

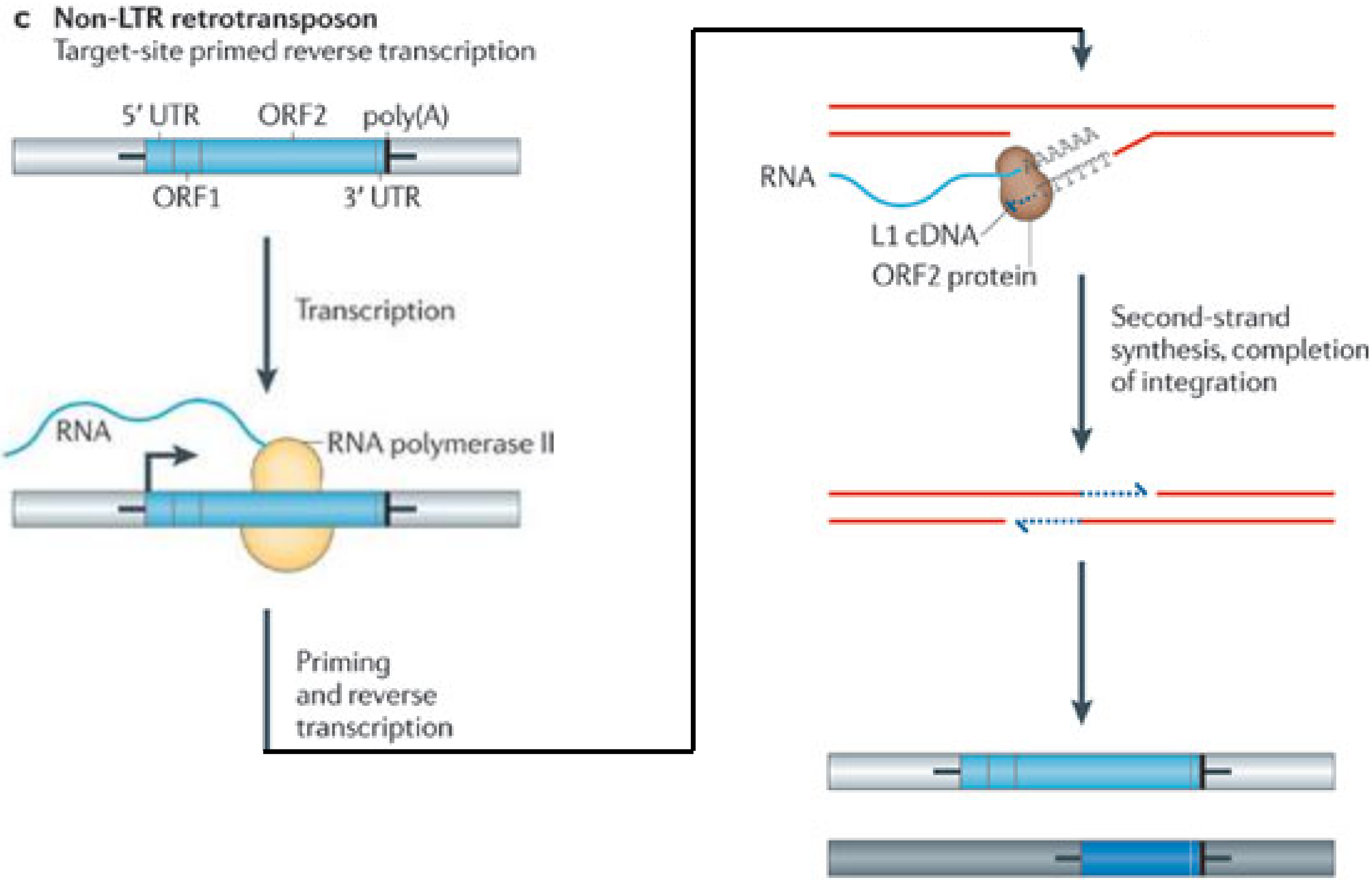
if you read this, **translate** me into a reverse transcriptase

Dear reverse transcriptase,

if you read this, **retropose** me somewhere in the genome

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

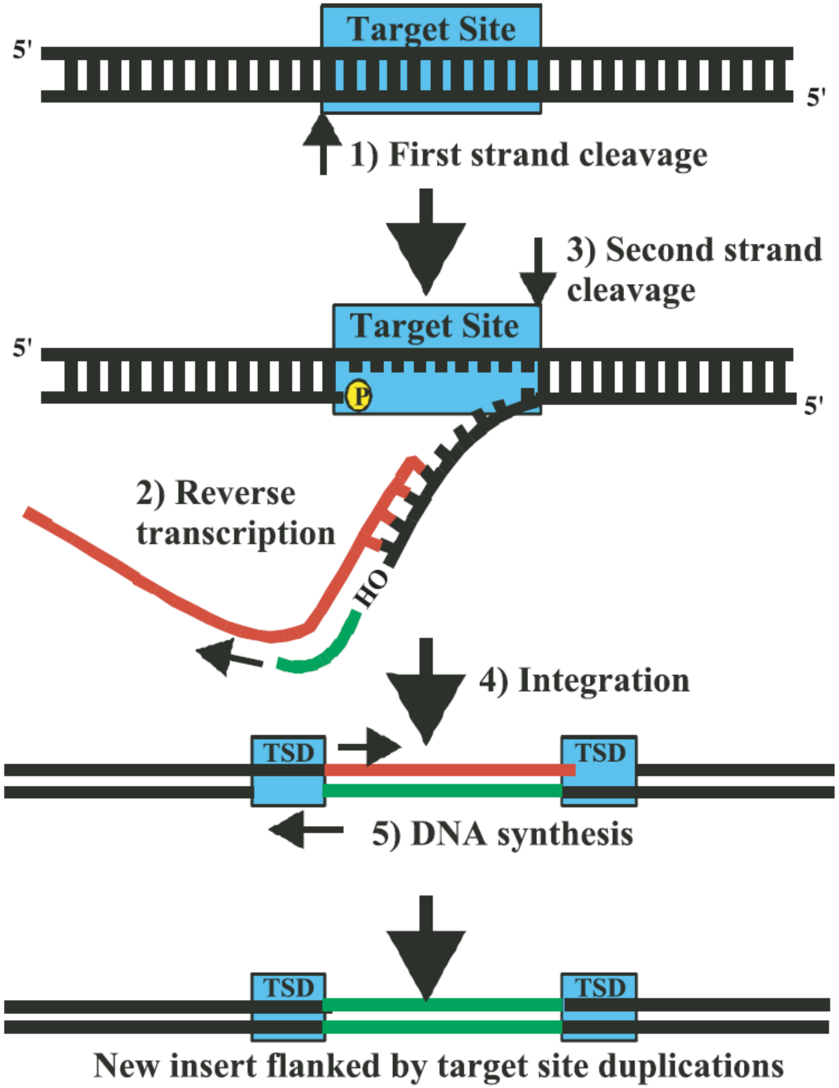
Target-primed reverse transcription (TPRT)


c Non-LTR retrotransposon
Target-site primed reverse transcription
5′ UTR
ORF2
poly(A)
ORF1
3′ UTR
Transcription
RNA
RNA polymerase II
Priming and reverse transcription
RNA
AAAAAA
TTTTT
L1 cDNA
ORF2 protein
Second-strand synthesis, completion of integration

TPRT frequently undergoes premature termination (5‘ truncation)

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

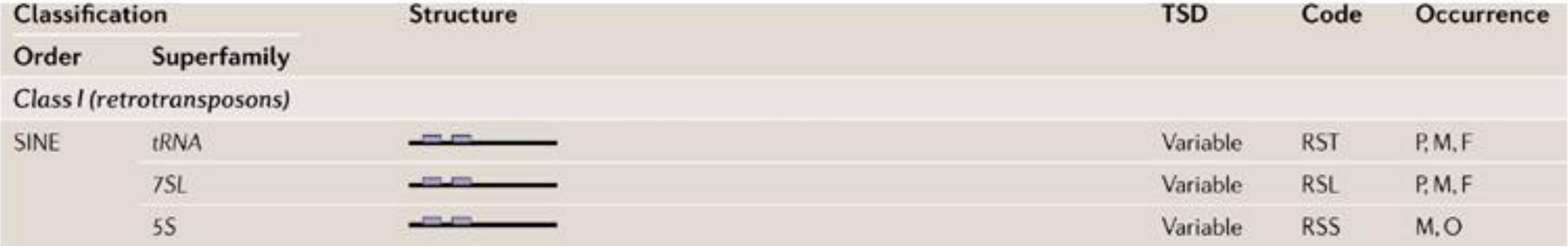
Target site duplication (TSD)



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

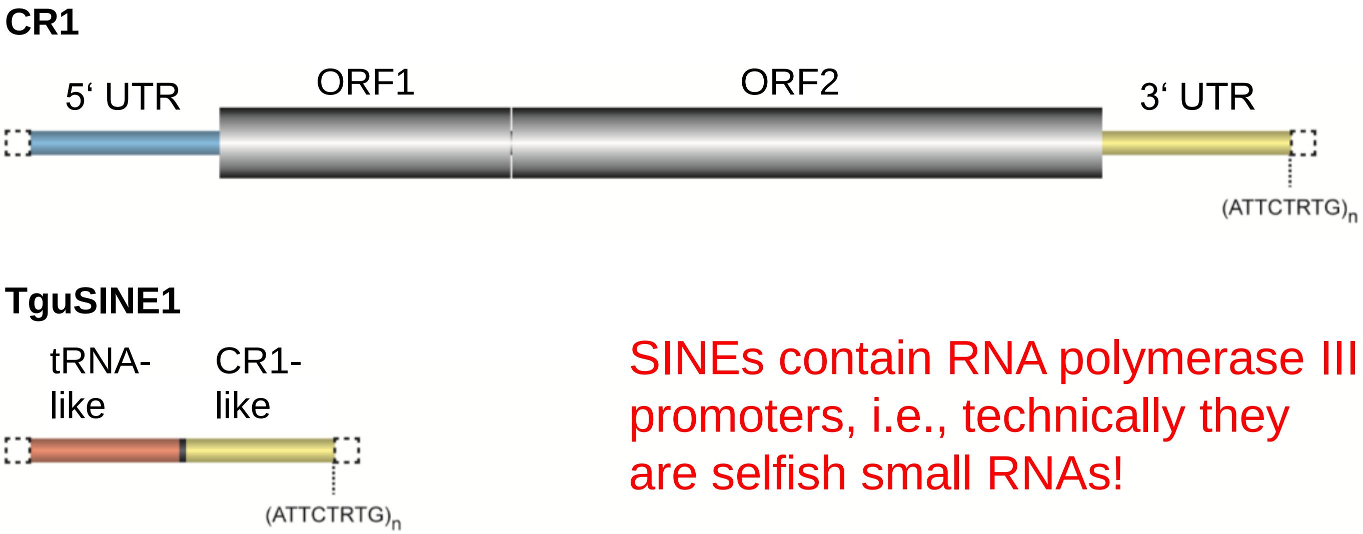
Kazazian 2004, *Science*

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.

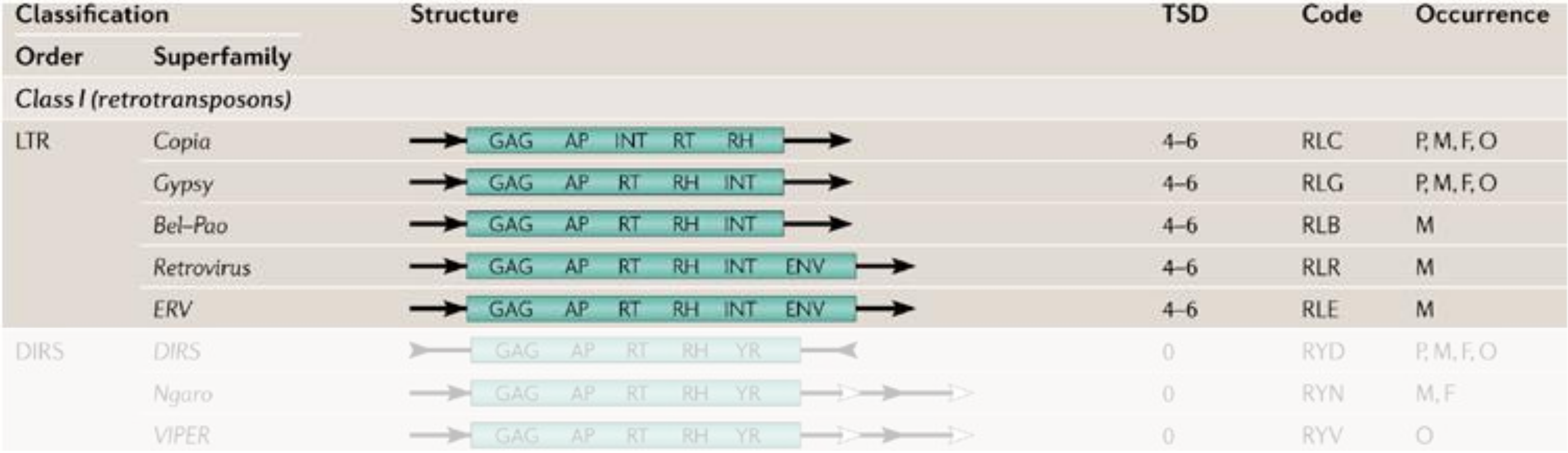


SINEs contain RNA polymerase III promoters, i.e., technically they are selfish small RNAs!

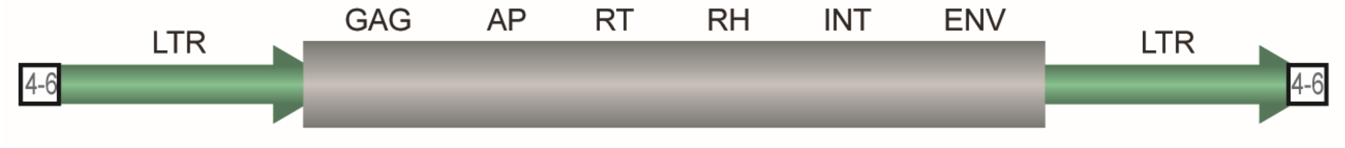
Note: In theory, any small RNA gene (pol III) can become a SINE!

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

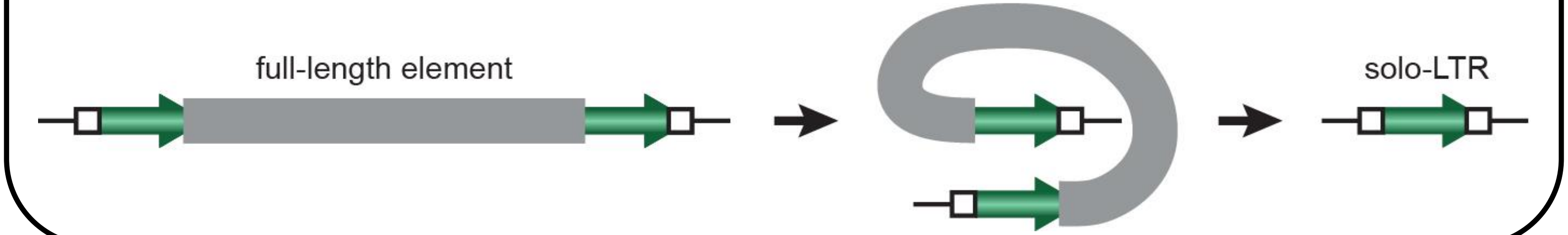
Class I: LTR retrotransposons



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

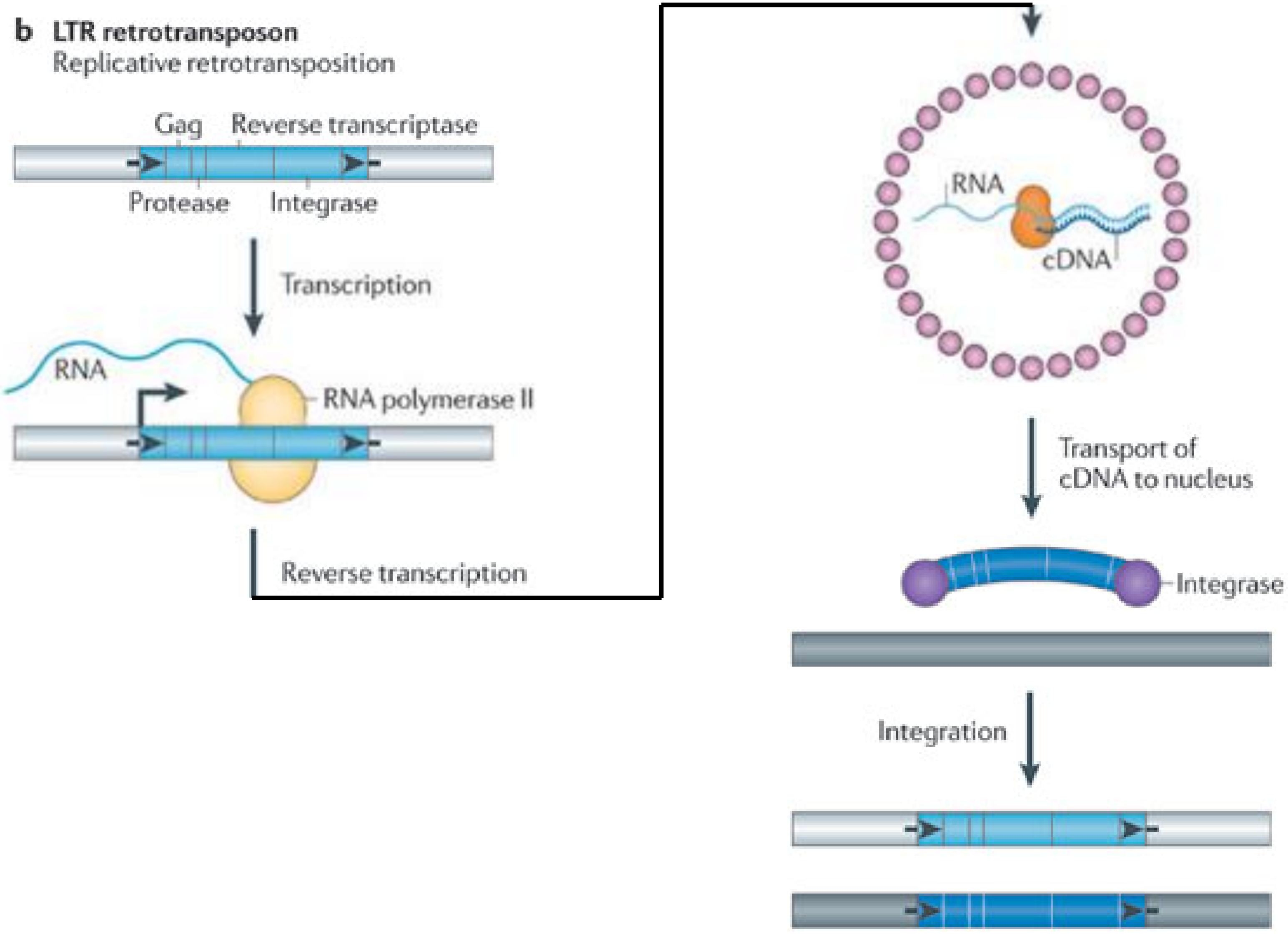


Non-allelic homologous recombination (NAHR):



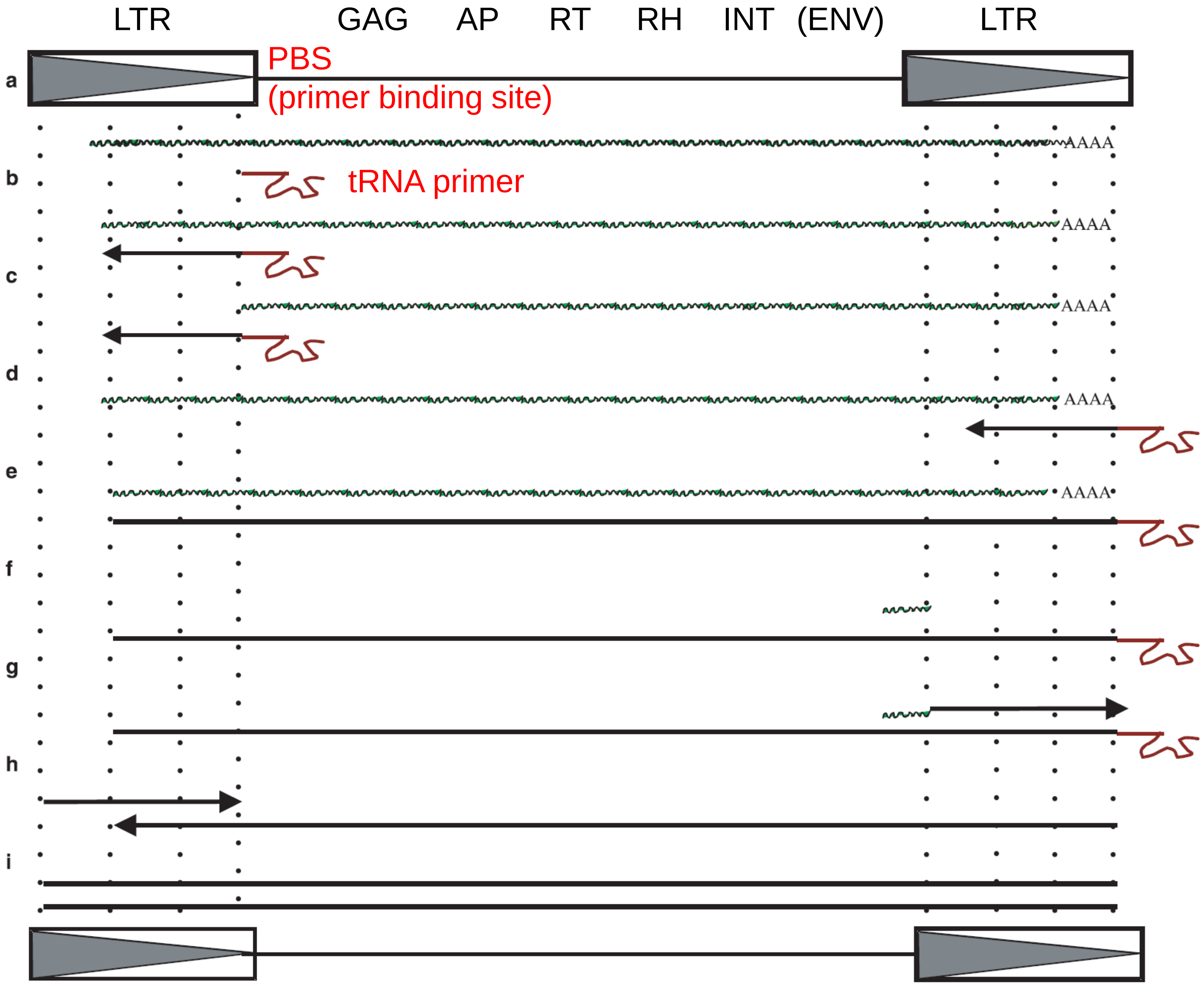
Wicker et al. 2008, *Nat. Rev. Genet.*

Replicative retrotransposition



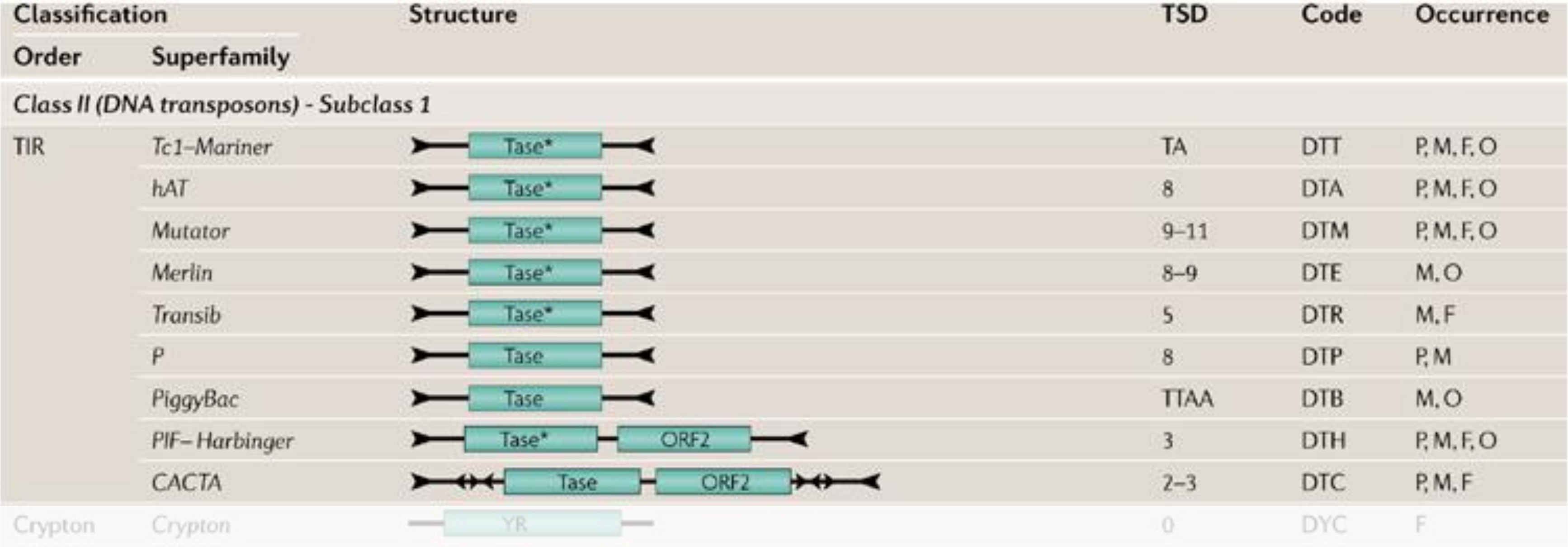
Levin & Moran 2011, *Nat. Rev. Genet.*

Why LTR retrotransposons have LTRs



Kazazian 2004, *Science*

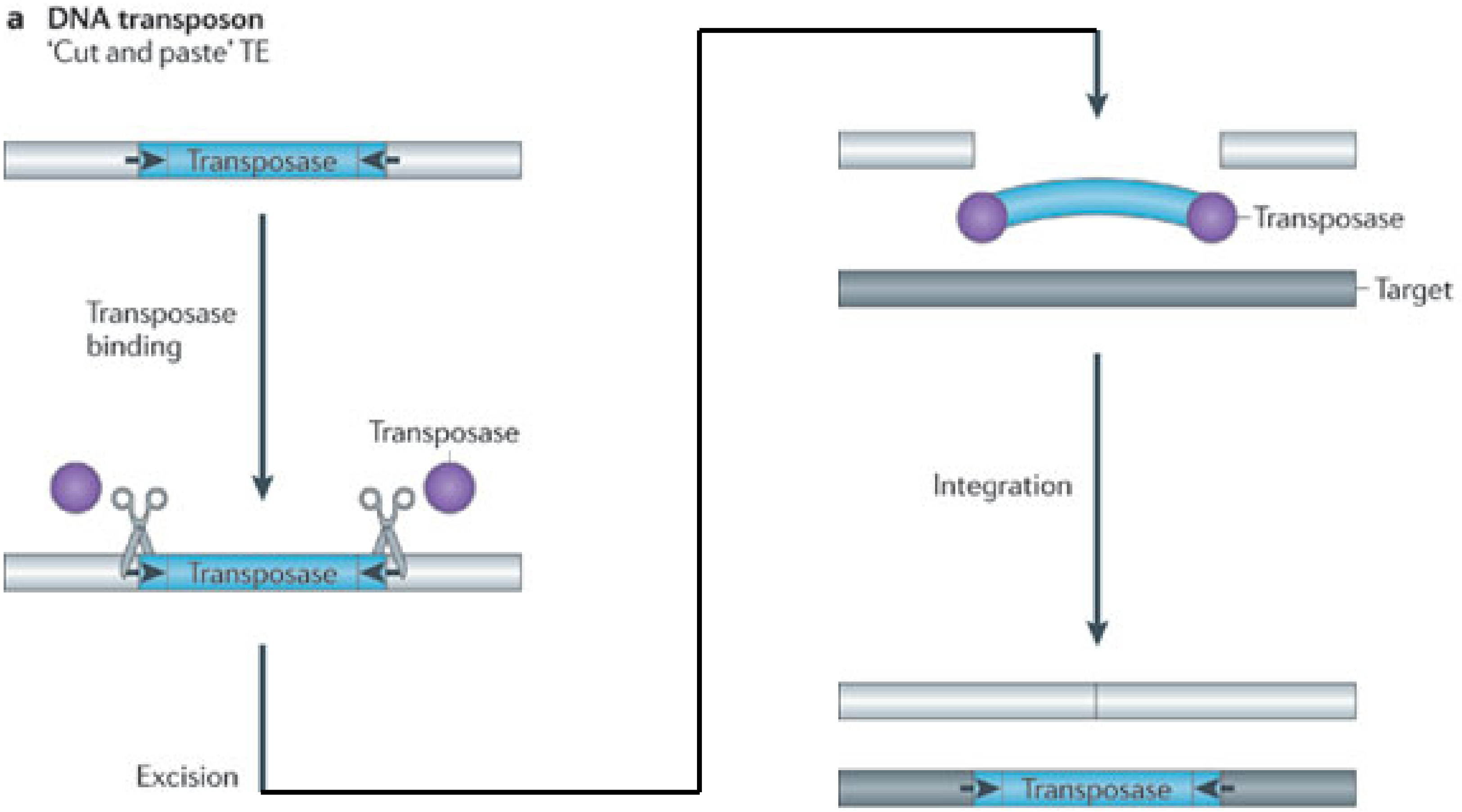
Class II: DNA transposons



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.*

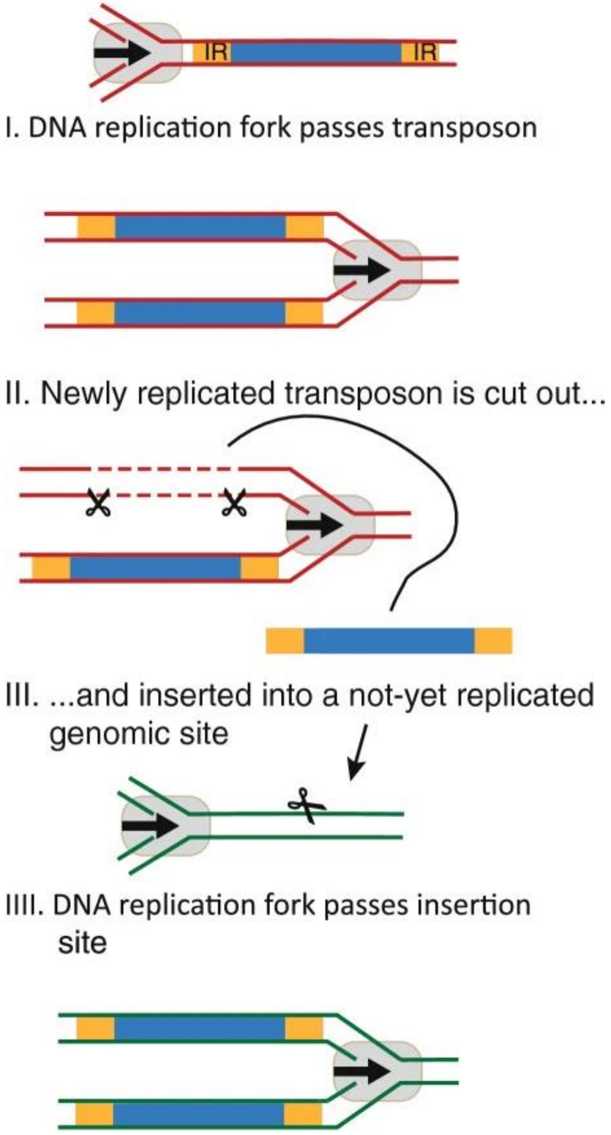
Cut-and-paste transposition (TIR)



Mobile DNA

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

How to increase in copy number?

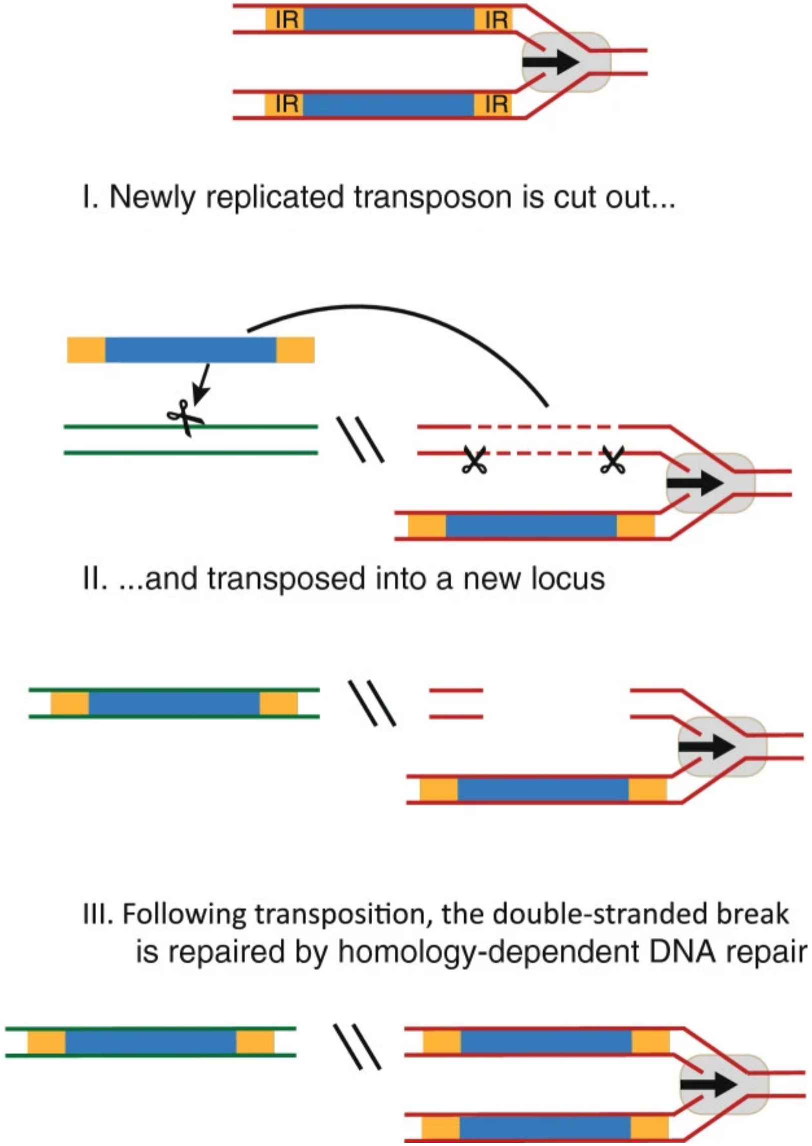


I. DNA replication fork passes transposon

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

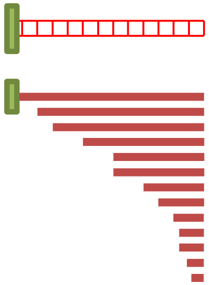
Skipper et al. 2013, *J. Biomed. Sci.*

TE ≠ TE

LINE

Mother copy

Daughter copies



Full-length
(>4 kb)

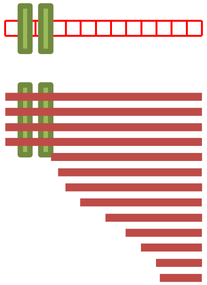
5‘-truncated

5‘ truncation during insertion

SINE

Mother copy

Daughter copies



Full-length
(>0.1 kb)

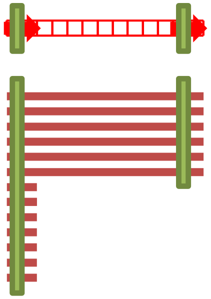
5‘-truncated

5‘ truncation during insertion

LTR

Mother copy

Daughter copies



Full-length
(>5 kb)

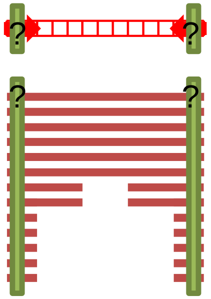
Solo-LTR
(>0.2 kb)

NAHR after insertion

TIR

Mother copy

Daughter copies



Autonomous
(>1 kb)

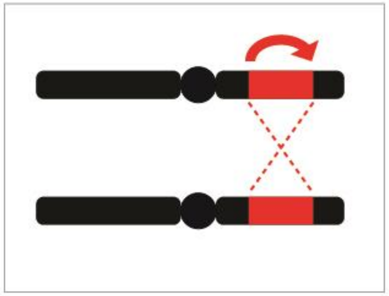
Non-autono.
(>0.1 kb)

Deletion after insertion

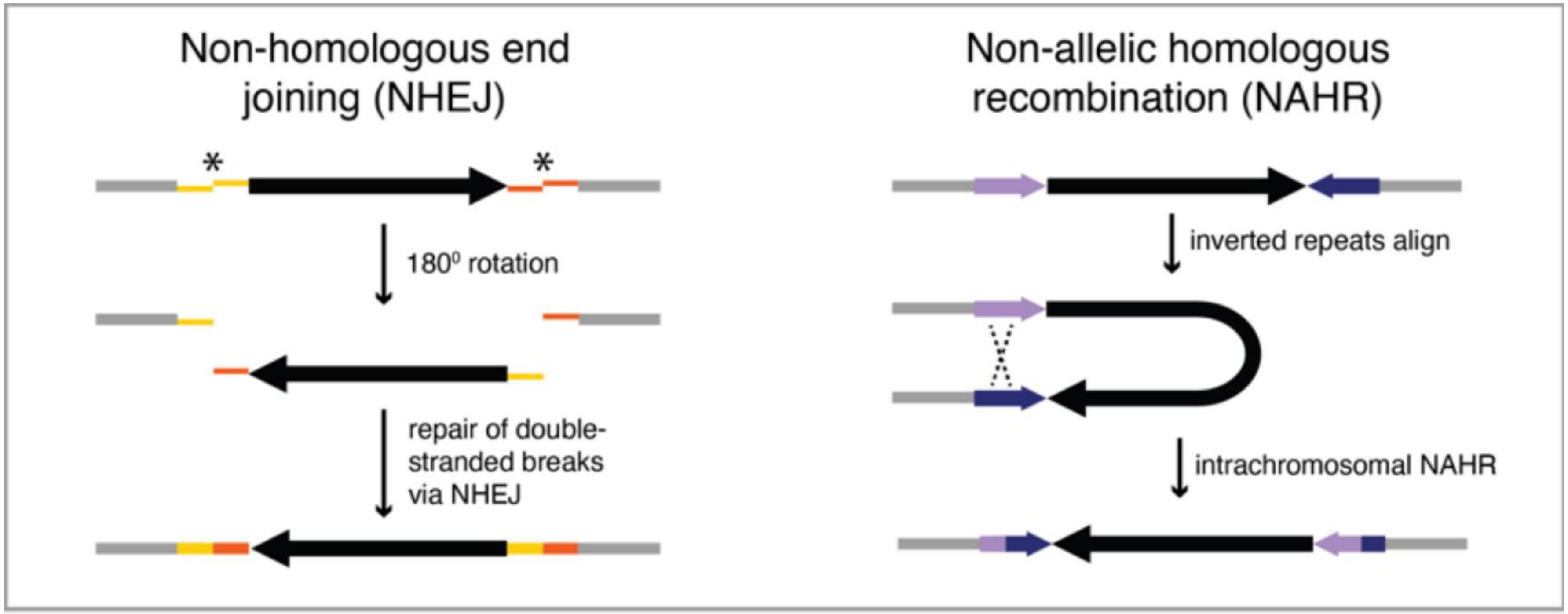
Some TE copies contain regulatory elements, some don‘t.

More context in Suh 2021 TE lecture 2

Inversion formation



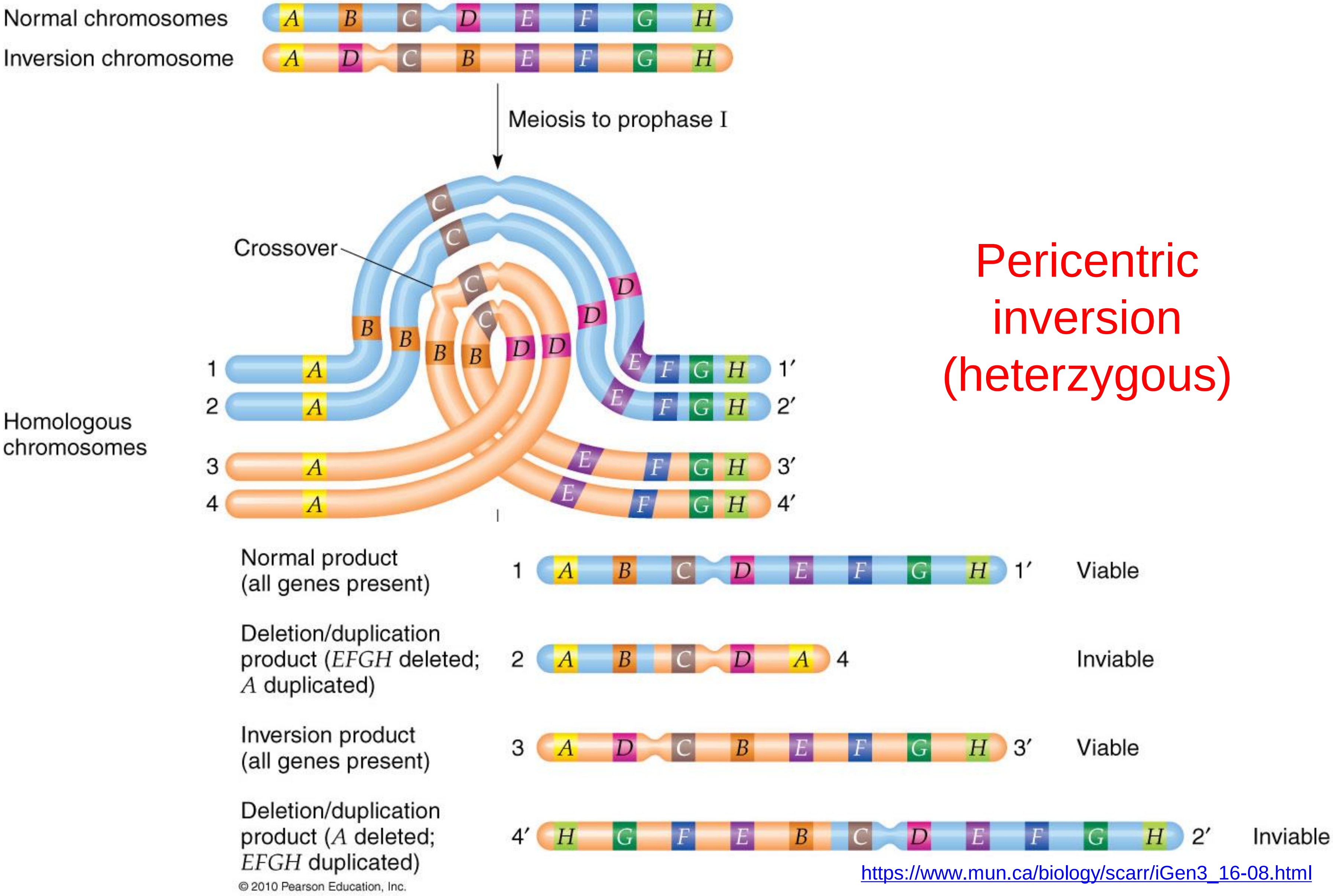
“We found that inversion breakpoints frequently occur in centromeric and telomeric regions and are often flanked by long inverted repeats (0.5-50 kb)”


Non-homologous end joining (NHEJ)
*
*
180⁰ rotation
repair of double-stranded breaks via NHEJ
Non-allelic homologous recombination (NAHR)
inverted repeats align
intrachromosomal NAHR

Assembling or mapping inversion breakpoints is difficult!

Harringmeyer & Hoekstra 2022, *bioRxiv*

Inversions "reduce" recombination

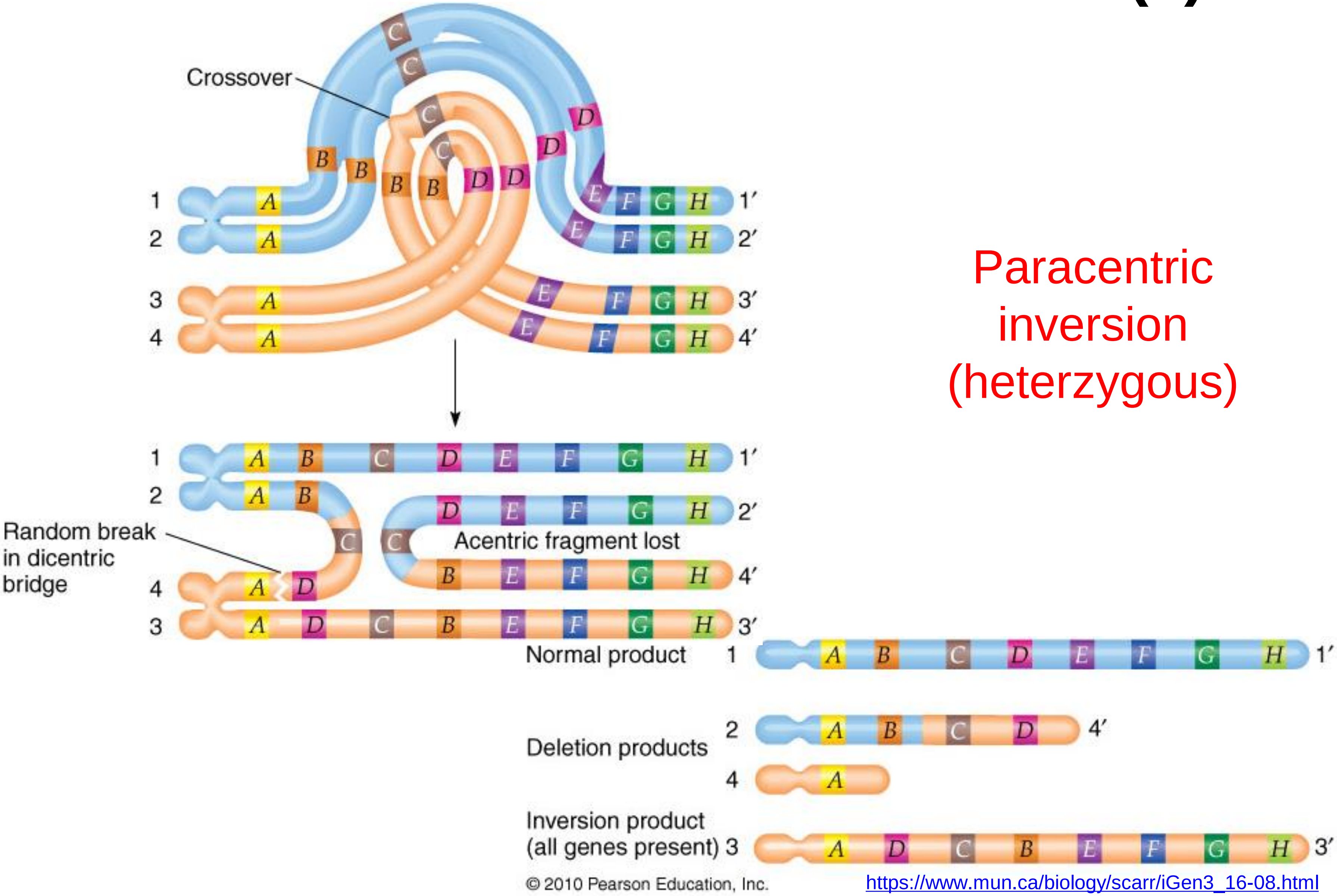


© 2010 Pearson Education, Inc.

Pericentric inversion (heterzygous)

https://www.mun.ca/biology/scarr/iGen3_16-08.html

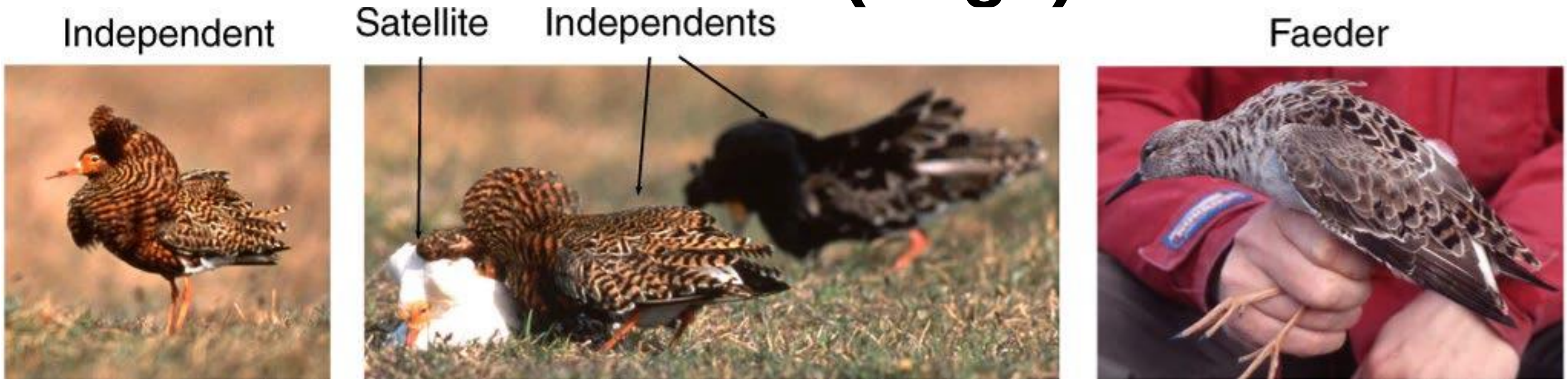
Inversions "reduce" recombination (2)

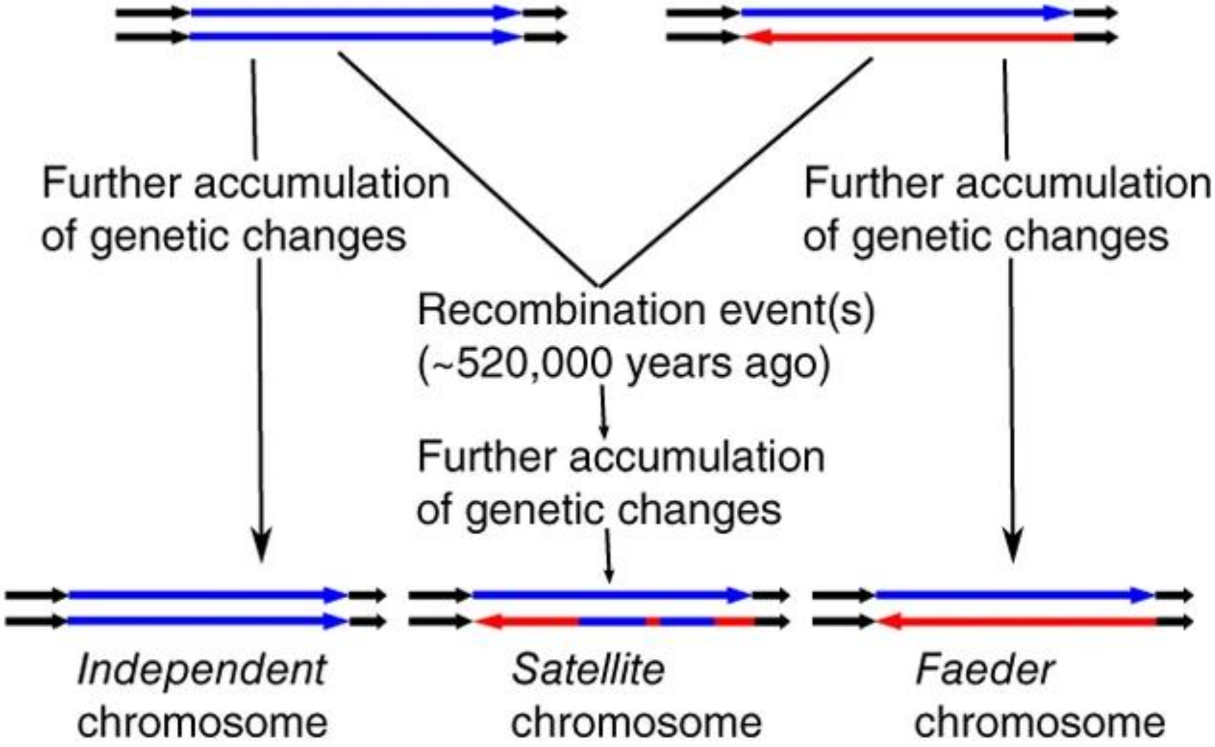

Crossover
1
2
3
4
A
B
C
D
E
F
G
H
1′
2′
3′
4′
Random break in dicentric bridge
Acentric fragment lost
Normal product
Deletion products
Inversion product (all genes present)
© 2010 Pearson Education, Inc.

Paracentric inversion (heterzygous)

https://www.mun.ca/biology/scarr/iGen3_16-08.html

Rare recombination in (large) inversions





Further accumulation of genetic changes

Further accumulation of genetic changes

Recombination event(s) (~520,000 years ago)

Further accumulation of genetic changes

Independent chromosome

Satellite chromosome

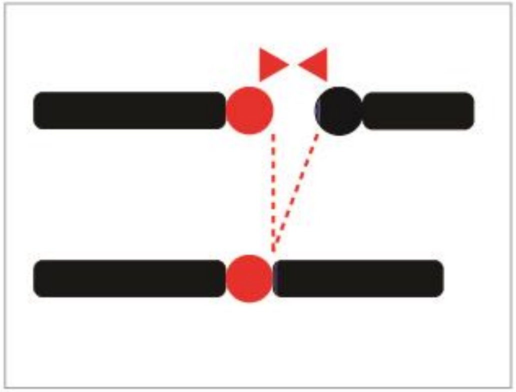
Faeder chromosome

Double-crossovers needed!

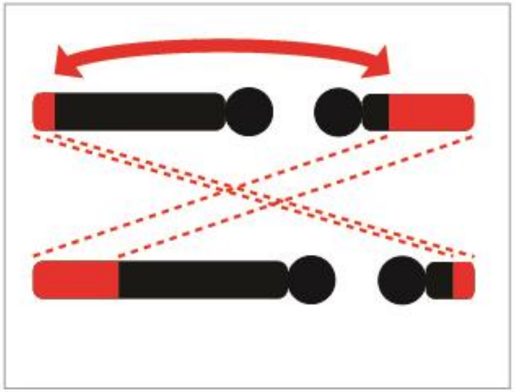
Lamichhaney et al. 2015, *Nat. Genet.*

More cases of NAHR

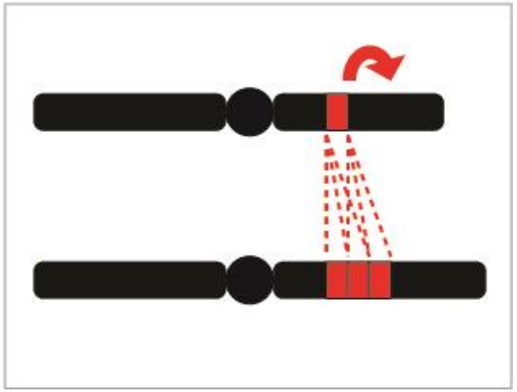
Fusion/fission



Translocation



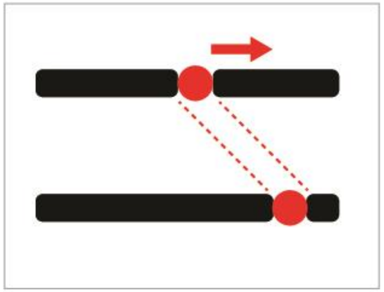
Duplication



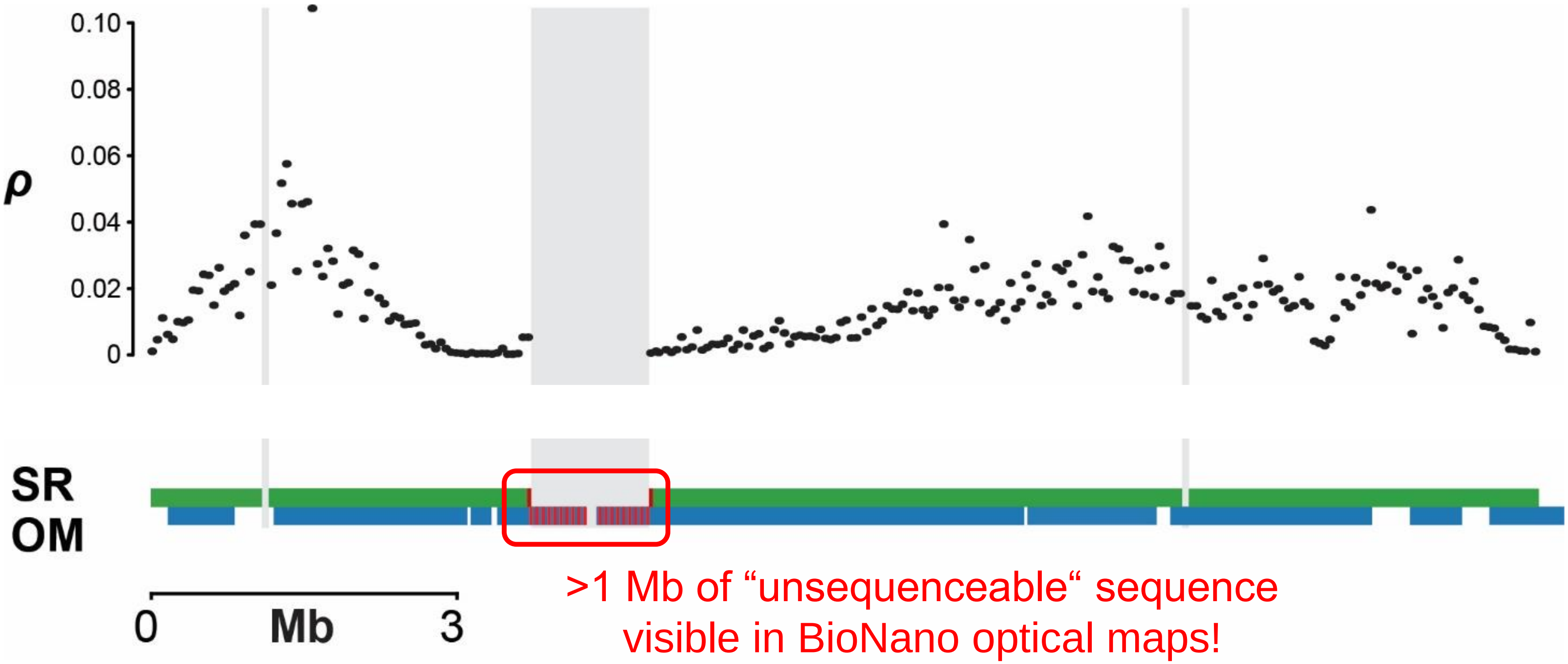
Fusions/fissions/translocations can decrease (new proximity to centromere) or increase (new proximity to telomere) recombination rates

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

Centromere shifts



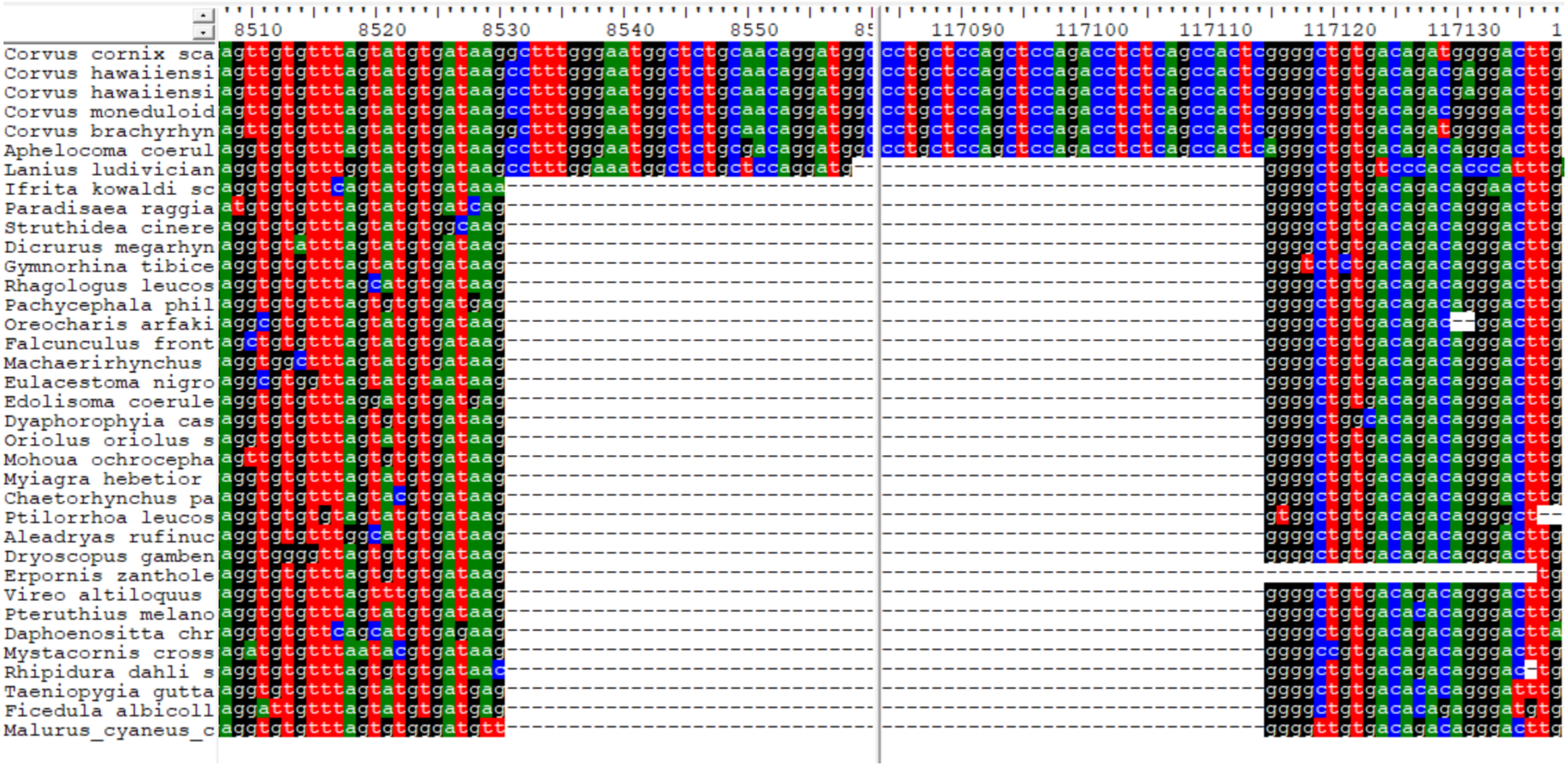
Chromosome 18 of hooded/carrion crow



>1 Mb of “unsequenceable“ sequence visible in BioNano optical maps!

Weissensteiner et al. 2017, *Genome Res.*

Centromere shifts across songbirds

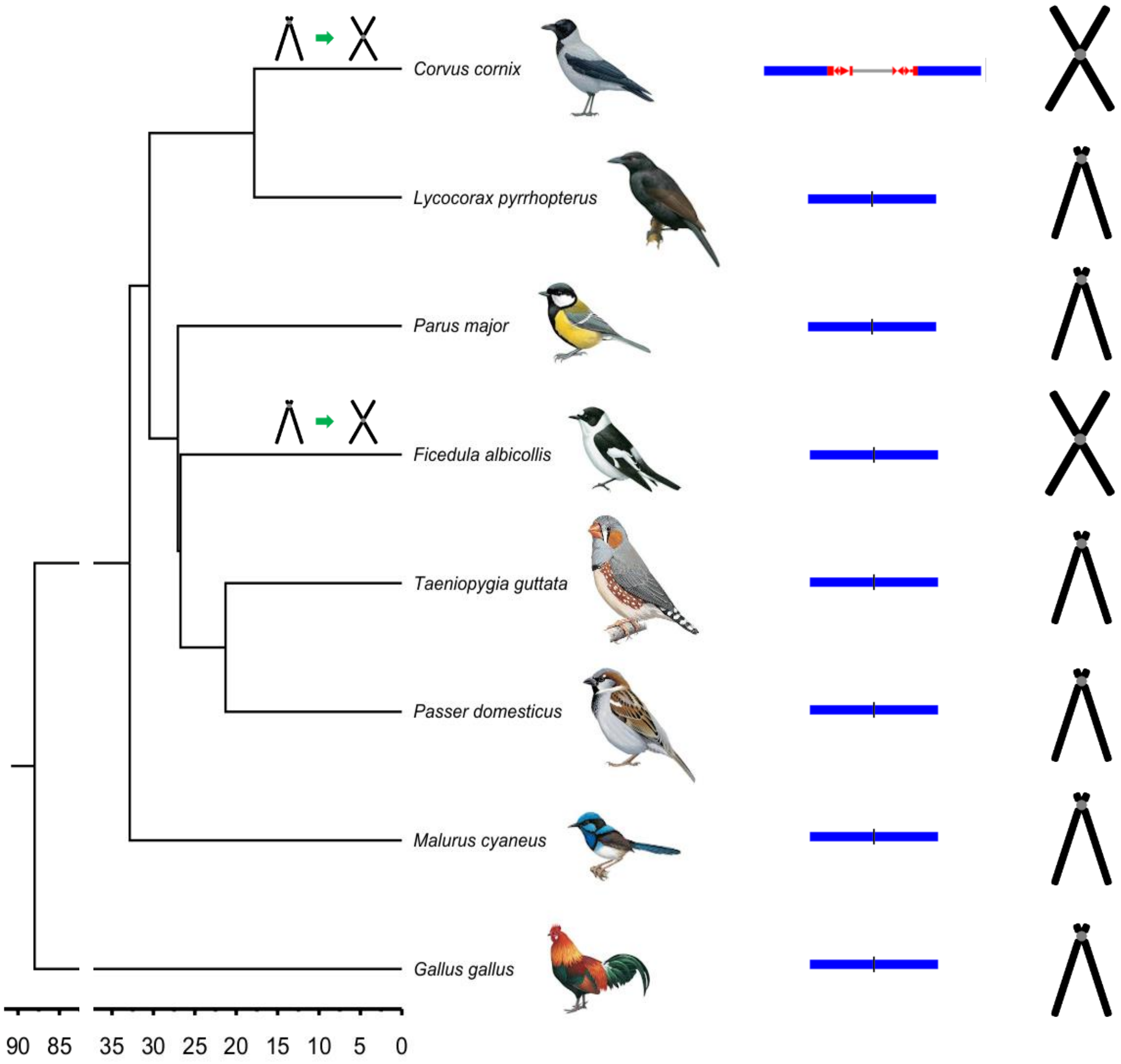


```
                  8510      8520      8530      8540      8550      85     117090    117100    117110    117120    117130    1
Corvus cornix sca agttgtgtttagtatgtgataaggctttgggaatggctctgcaacaggatggc cctgctccagctccagacctctcagccactcggggctgtgacagatggggacttg
Corvus hawaiiensi agttgtgtttagtatgtgataagcctttgggaatggctctgcaacaggatggc cctgctccagctccagacctctcagccactcggggctgtgacagacgaggacttg
Corvus hawaiiensi agttgtgtttagtatgtgataagcctttgggaatggctctgcaacaggatggc cctgctccagctccagacctctcagccactcggggctgtgacagacgaggacttg
Corvus moneduloid agttgtgtttagtatgtgataagcctttgggaatggctctgcaacaggatggc cctgctccagctccagacctctcagccactcggggctgtgacagacggggacttg
Corvus brachyrhyn agttgtgtttagtatgtgataaggctttgggaatggctctgcaacaggatggc cctgctccagctccagacctctcagccactcggggctgtgacagatggggacttg
Aphelocoma coerul aggtgtgtttagtatgtgataagcctttgggaatggctctgcgacaggatggc cctgctccagctccagacctctcagccactcagggctgtgacagacagggacttg
Lanius ludivician aggtgtgtttggtatgtgataagcctttggaaatggctctgctccaggatg-- -------------------------------ggggctgtgtcccacacccatttg
Ifrita kowaldi sc aggtgtgttcagtatgtgataaa------------------------------ -------------------------------ggggctgtgacagacaggaacttg
Paradisaea raggia atgtgtgtttagtatgtgatcag------------------------------ -------------------------------ggggctgtgacagacagggacttg
Struthidea cinere aggtgtgtttagtatgtggcaag------------------------------ -------------------------------ggggctgtgacagacagggacttg
Dicrurus megarhyn aggtgtatttagtatgtgataag------------------------------ -------------------------------ggggctgtgacagacagggacttg
Gymnorhina tibice aggtgtgtttagtatgtgataag------------------------------ -------------------------------gggtctctgacagacagggacttg
Rhagologus leucos aggtgtgtttagcatgtgataag------------------------------ -------------------------------ggggctgtgacagacagggacttg
Pachycephala phil aggtgtgtttagtgtgtgatgag------------------------------ -------------------------------ggggctgtgacagacagggacttg
Oreocharis arfaki aggcgtgtttagtatgtgataag------------------------------ -------------------------------ggggctgtgacagac--ggacttg
Falcunculus front agctgtgtttagtatgtgataag------------------------------ -------------------------------ggggctgtgacagacagggacttg
Machaerirhynchus  aggtggctttagtatgtgataag------------------------------ -------------------------------ggggctgtgacagacagggacttg
Eulacestoma nigro aggcgtggttagtatgtaataag------------------------------ -------------------------------ggggctgtgacagacagggacttg
Edolisoma coerule aggtgtgtttaggatgtgatgag------------------------------ -------------------------------ggggctgtgacagacagggacttg
Dyaphorophyia cas aggtgtgtttagtgtgtgataag------------------------------ -------------------------------ggggctggcacagacagggacttg
Oriolus oriolus s aggtgtgtttagtatgtgataag------------------------------ -------------------------------ggggctgtgacagacagggacttg
Mohoua ochrocepha agttgtgtttagtgtgtgataag------------------------------ -------------------------------ggggctgtgacagacagggacttg
Myiagra hebetior  aggtgtgtttagtatgtgataag------------------------------ -------------------------------ggggctgtgacagacagggacttg
Chaetorhynchus pa aggtgtgtttagtacgtgataag------------------------------ -------------------------------ggggctgtgacagacagggacttg
Ptilorrhoa leucos aggtgtgtgtagtatgtgataag------------------------------ -------------------------------gtggctgtgacagacaggggct--
Aleadryas rufinuc aggtgtgtttggcatgtgataag------------------------------ -------------------------------ggggctgtgacagacagggacttg
Dryoscopus gamben aggtggggttagtgtgtgataag------------------------------ -------------------------------ggggctgtgacagacagggacttg
Erpornis zanthole aggtgtgtttagtgtgtgataag------------------------------ -----------------------------------------------------tg
Vireo altiloquus  aggtgtgtttagtttgtgataag------------------------------ -------------------------------ggggctgtgacagacagggacttg
Pteruthius melano aggtgtgtttagtatgtgataag------------------------------ -------------------------------ggggctgtgacacacagggacttg
Daphoenositta chr aggtgtgttcagcatgtgagaag------------------------------ -------------------------------ggggctgtgacagacagggactta
Mystacornis cross agatgtgtttaatacgtgataag------------------------------ -------------------------------ggggccgtgacagacagggacttg
Rhipidura dahli s aggtgtgtttagtgtgtgataac------------------------------ -------------------------------ggggctgtgacagacagggac-tg
Taeniopygia gutta aggtgtgtttagtatgtgatgag------------------------------ -------------------------------ggggctgtgacacacagggatttg
Ficedula albicoll aggattgtttagtatgtgatgag------------------------------ -------------------------------ggggctgtgacacagagggatgtg
Malurus_cyaneus_c aggtgtgtttagtgtgggatgtt------------------------------ -------------------------------ggggttgtgacagacagggacttg
```

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

Westerberg et al., *manuscript*

Centromere shifts across songbirds



Metacentric

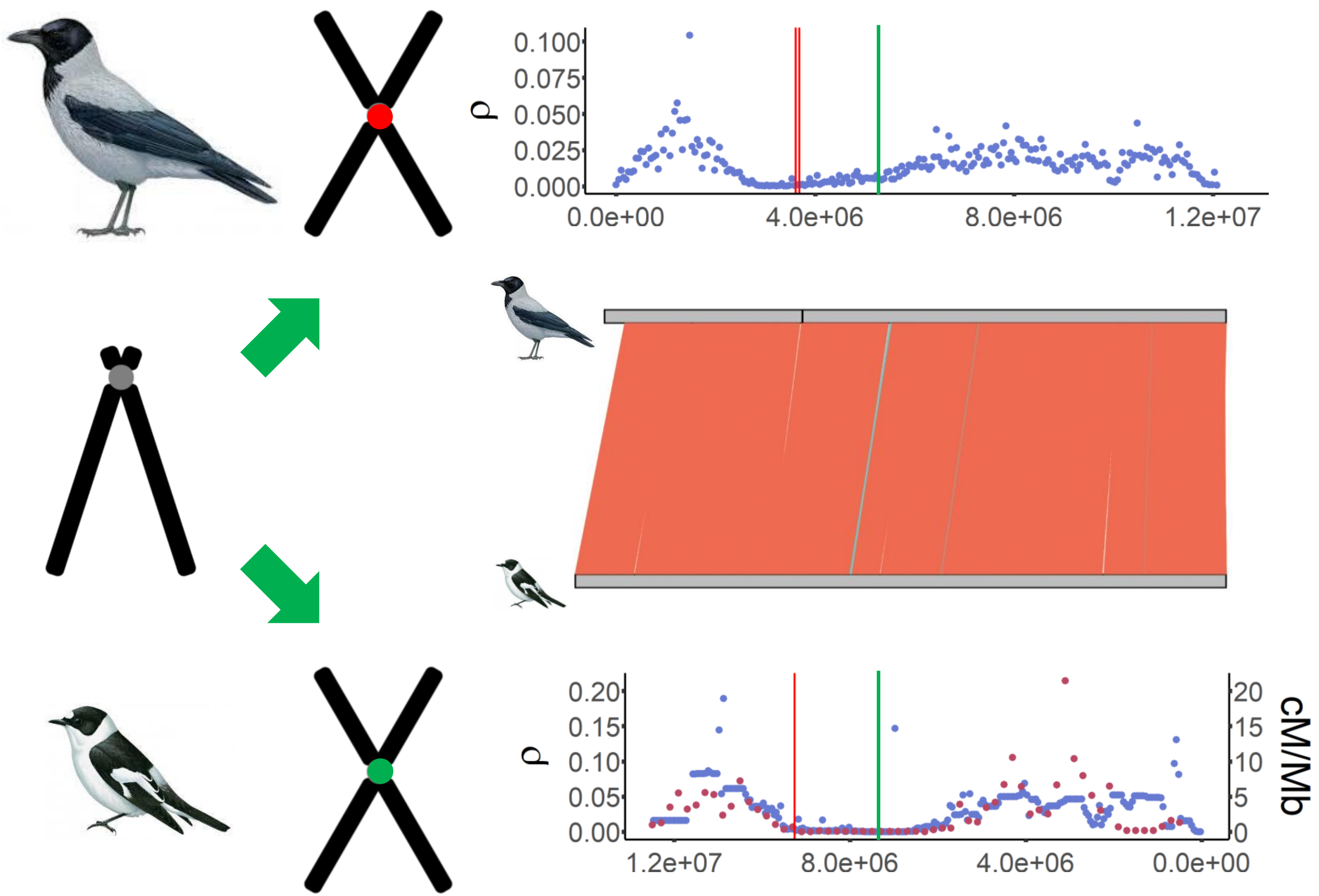
Metacentric
(FISH of chicken chr18 BAC)

Acrocentric
(verified flanking sequences;
Knief & Forstmeier 2015)

Acrocentric
(ChIP-seq of CENP-A;
Shang et al. 2010)

Westerberg et al., *manuscript*

Not so stable chromosomes after all?



Westerberg et al., *manuscript*

Short break (5 minutes)

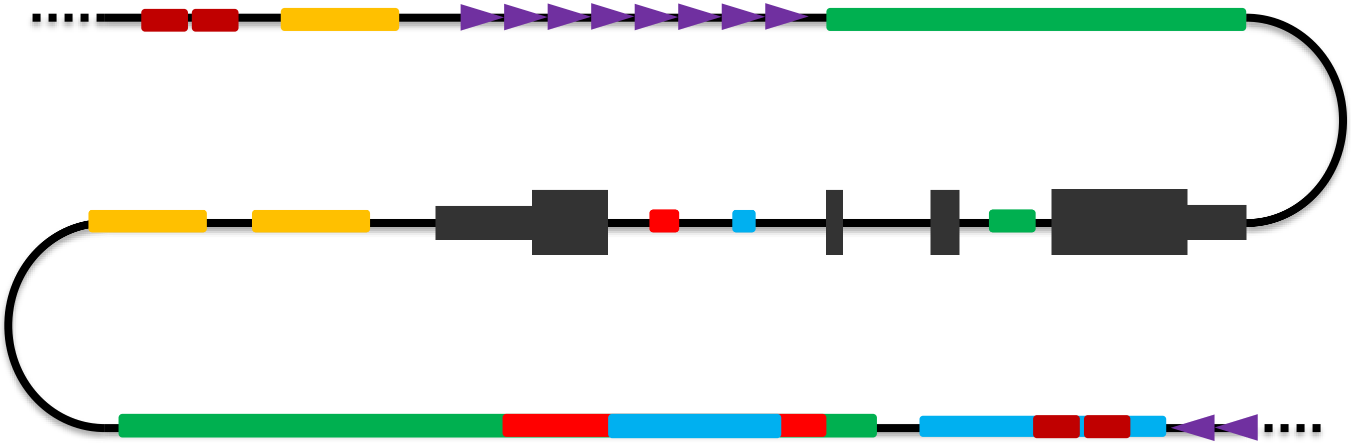
Task: Gather in the same groups of 3 and discuss what resources (assembly/read data, gene/repeat annotation) there are for your respective study system.

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

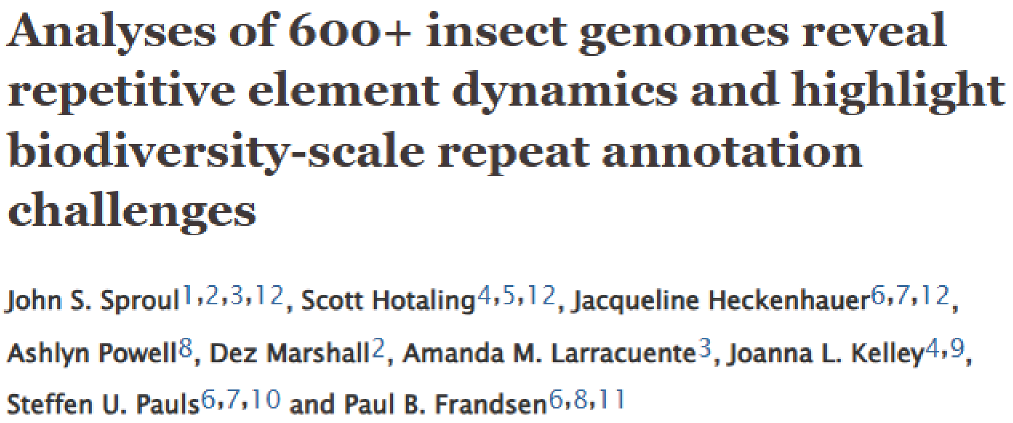
Family

Subfamily

Copy

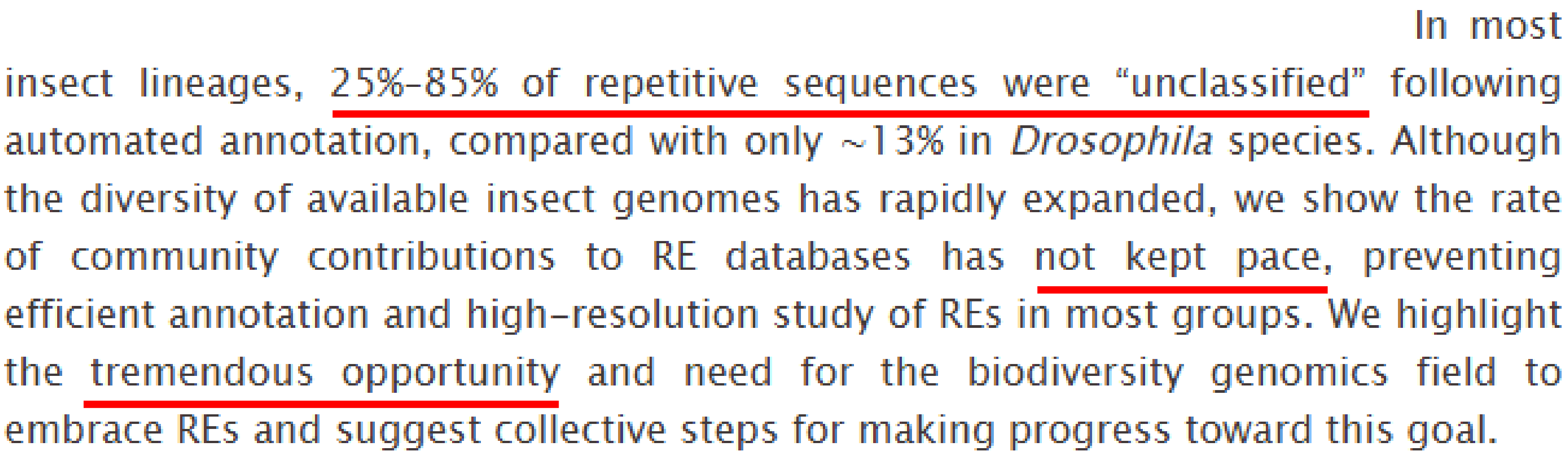
More context in Suh 2021 TE lecture 3

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 Powell[8], Dez Marshall[2], Amanda M. Larracuente[3], Joanna L. Kelley[4,9], Steffen U. Pauls[6,7,10] and Paul B. Frandsen[6,8,11]



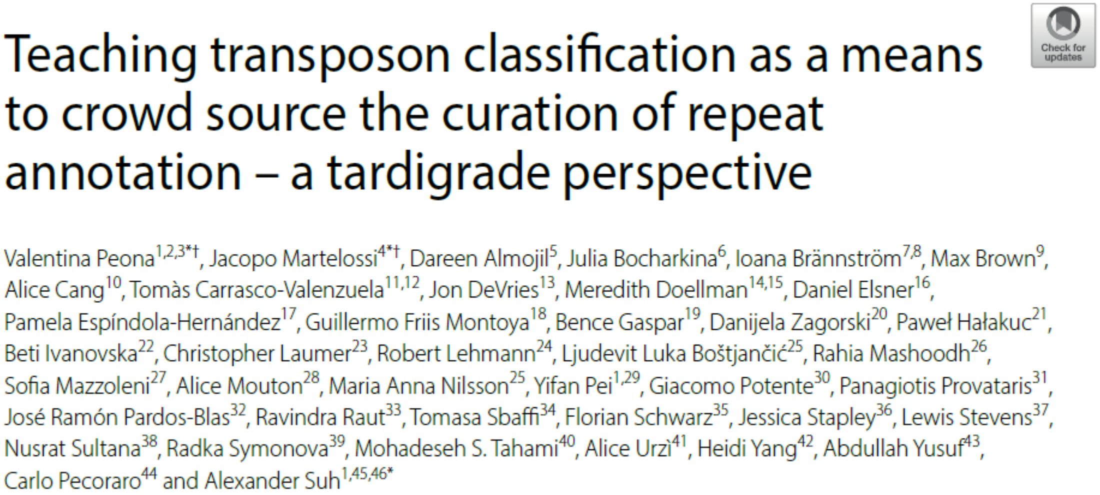
In most insect lineages, 25%–85% of repetitive sequences were “unclassified” following automated annotation, compared with only ~13% in *Drosophila* species. Although the diversity of available insect genomes has rapidly expanded, we show the rate of community contributions to RE databases has not kept 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.

Sproul et al. 2023, *Genome Res*

More community initiatives needed

TE Hub website

TE Worldwide Slack
#te-hub channel



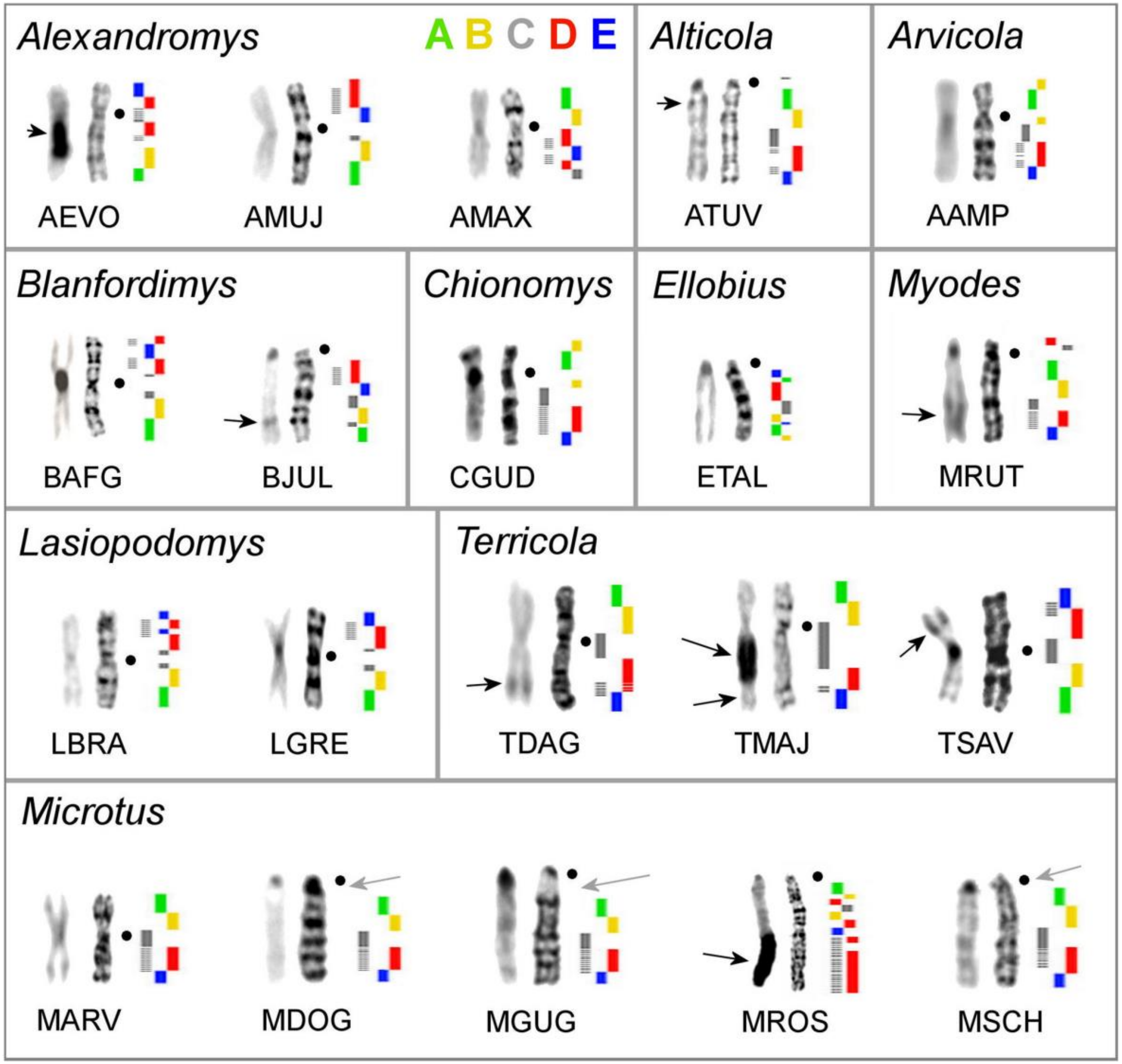
Teaching transposon classification as a means to crowd source the curation of repeat annotation – a tardigrade perspective

Valentina Peona[1,2,3*†], Jacopo Martelossi[4*†], Dareen Almojil[5], Julia Bocharkina[6], Ioana Brännström[7,8], Max Brown[9], Alice Cang[10], Tomàs Carrasco-Valenzuela[11,12], Jon DeVries[13], Meredith Doellman[14,15], Daniel Elsner[16], Pamela Espíndola-Hernández[17], Guillermo Friis Montoya[18], Bence Gaspar[19], Danijela Zagorski[20], Paweł Hałakuc[21], Beti Ivanovska[22], Christopher Laumer[23], Robert Lehmann[24], Ljudevit Luka Boštjančić[25], Rahia Mashoodh[26], Sofia Mazzoleni[27], Alice Mouton[28], Maria Anna Nilsson[25], Yifan Pei[1,29], Giacomo Potente[30], Panagiotis Provataris[31], José Ramón Pardos-Blas[32], Ravindra Raut[33], Tomasa Sbaffi[34], Florian Schwarz[35], Jessica Stapley[36], Lewis Stevens[37], Nusrat Sultana[38], Radka Symonova[39], Mohadeseh S. Tahami[40], Alice Urzì[41], Heidi Yang[42], Abdullah Yusuf[43], Carlo Pecoraro[44] and Alexander Suh[1,45,46*]

Peona & Martelossi et al. 2024, *Mobile DNA*

Genomics + cytogenetics = cytogenomics

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



Romanenko et al. 2020, *Sci. Rep.*

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.

Earth's heart of iron begins to yield its secrets p. 18

Microglia in chronic pain recovery and relapse pp. 33 & 86

Particle acceleration in a nova explosion p. 77

Science

$15
1 APRIL 2022
SPECIAL ISSUE
science.org

AAAS

FILLING THE GAPS

Closing in on a complete human genome p. 42

TELOMERE-TO-TELOMERE CONSORTIUM

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

Nurk et al. 2022, *Science*

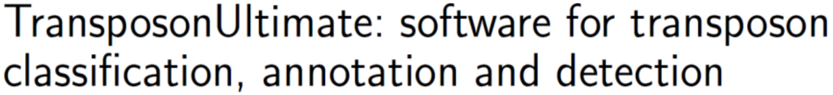
What’s next: Machine learning?

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.*

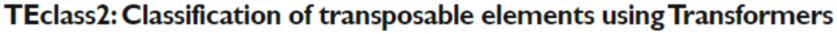


TransposonUltimate: software for transposon classification, annotation and detection

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*

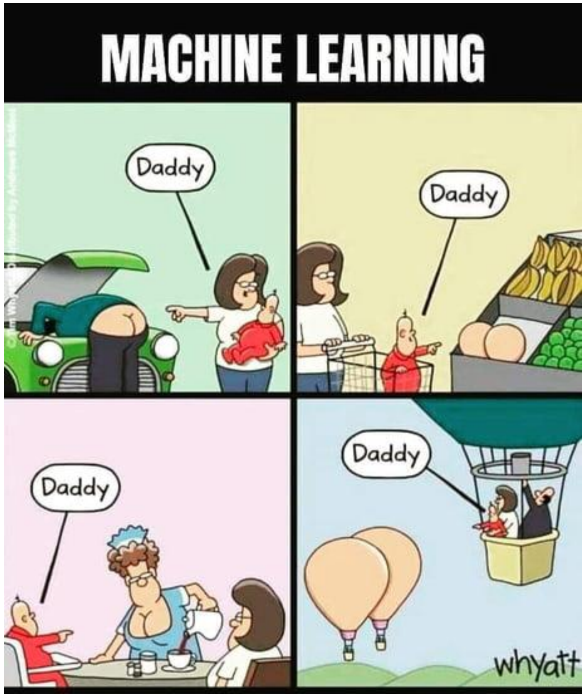


TEclass2: Classification of transposable elements using Transformers

Bickmann et al. 2023 *bioRxiv*

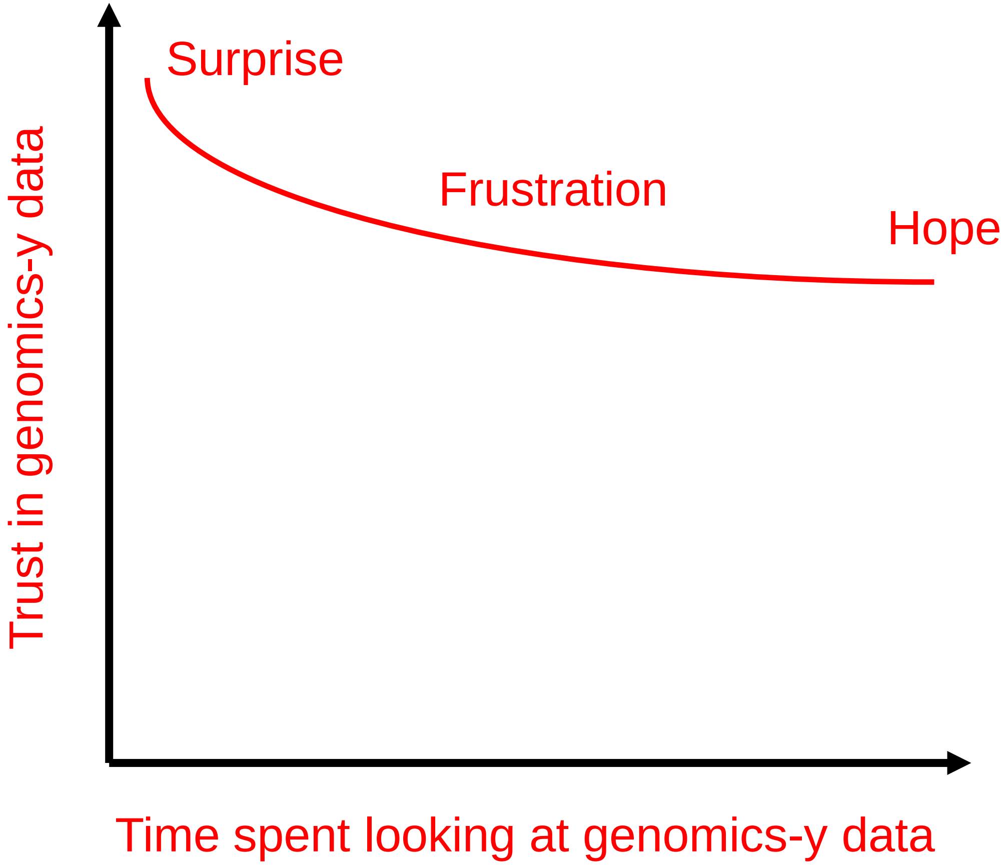
Comprehensive Hierarchical Classification of Transposable Elements based on Deep Learning

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?