

Structural variation

(A story of surprise → frustration → hope)

June 13, 2022



@alexander_suh



Alexander Sang-Jae Suh

서상재 徐商在

Senior lecturer

University of East Anglia, UK
Uppsala University, Sweden



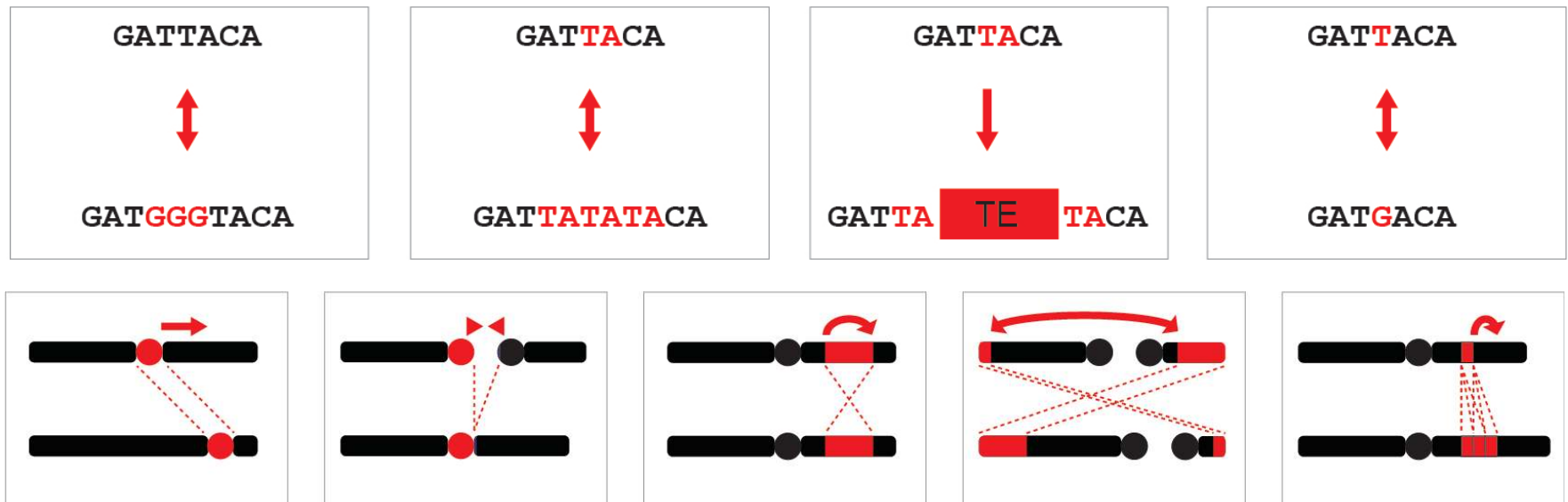
UPPSALA
UNIVERSITET

What is structural variation?

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

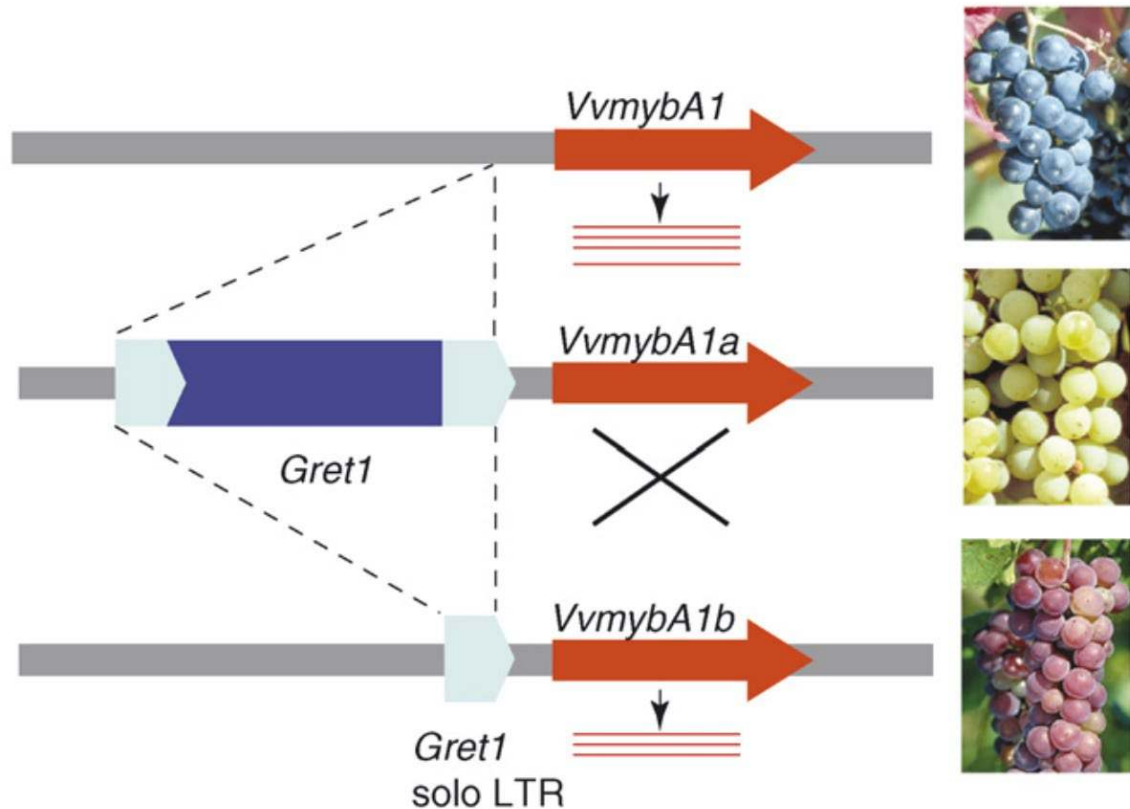
[Mérot et al. 2020, Trends Genet.](#)

Some key mutation types



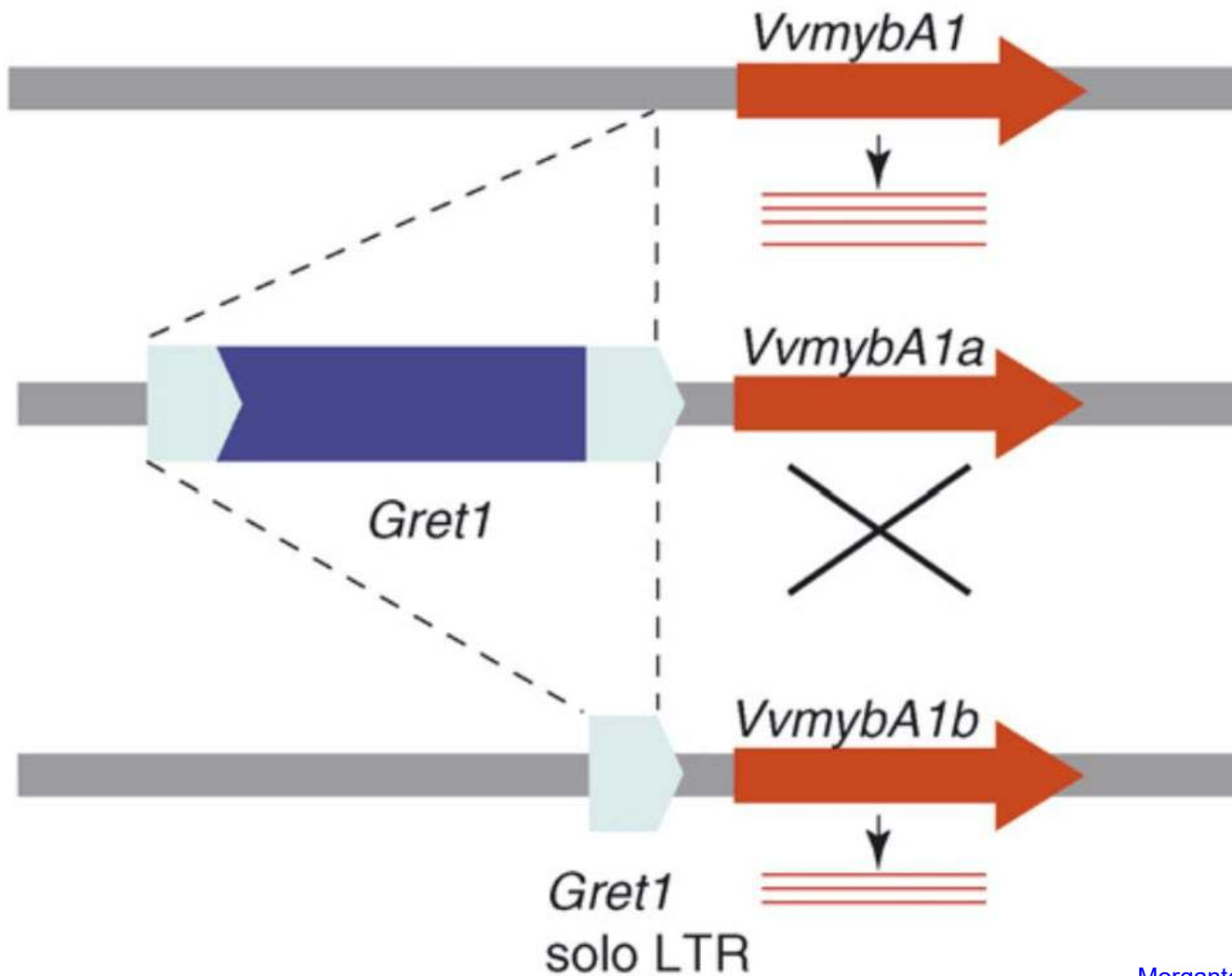
[Berdan et al. 2021, Mol. Ecol.](#)

Part 1: Surprise



A) It's not a SNP!

Delicious effects of SVs



Discovery of gene regulation in 1940s

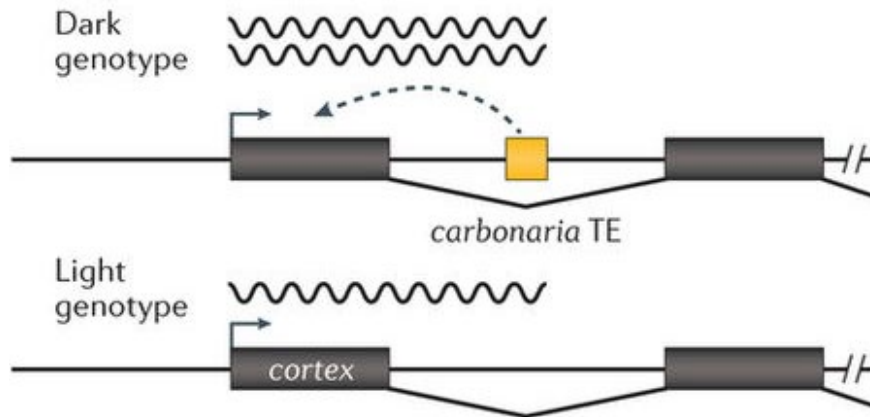


**Barbara
McClintock**
(Nobel Prize in
Physiology or
Medicine 1983)

TE-induced rapid adaptation

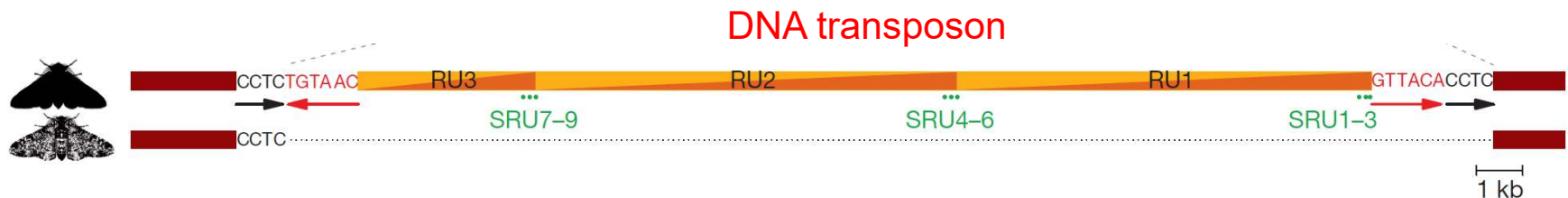
The industrial melanism of the peppered moth is probably the most famous textbook example for adaptation (in only a few decades)!

c Peppered moth



Upregulates cortex, resulting in increased darker coloration

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



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

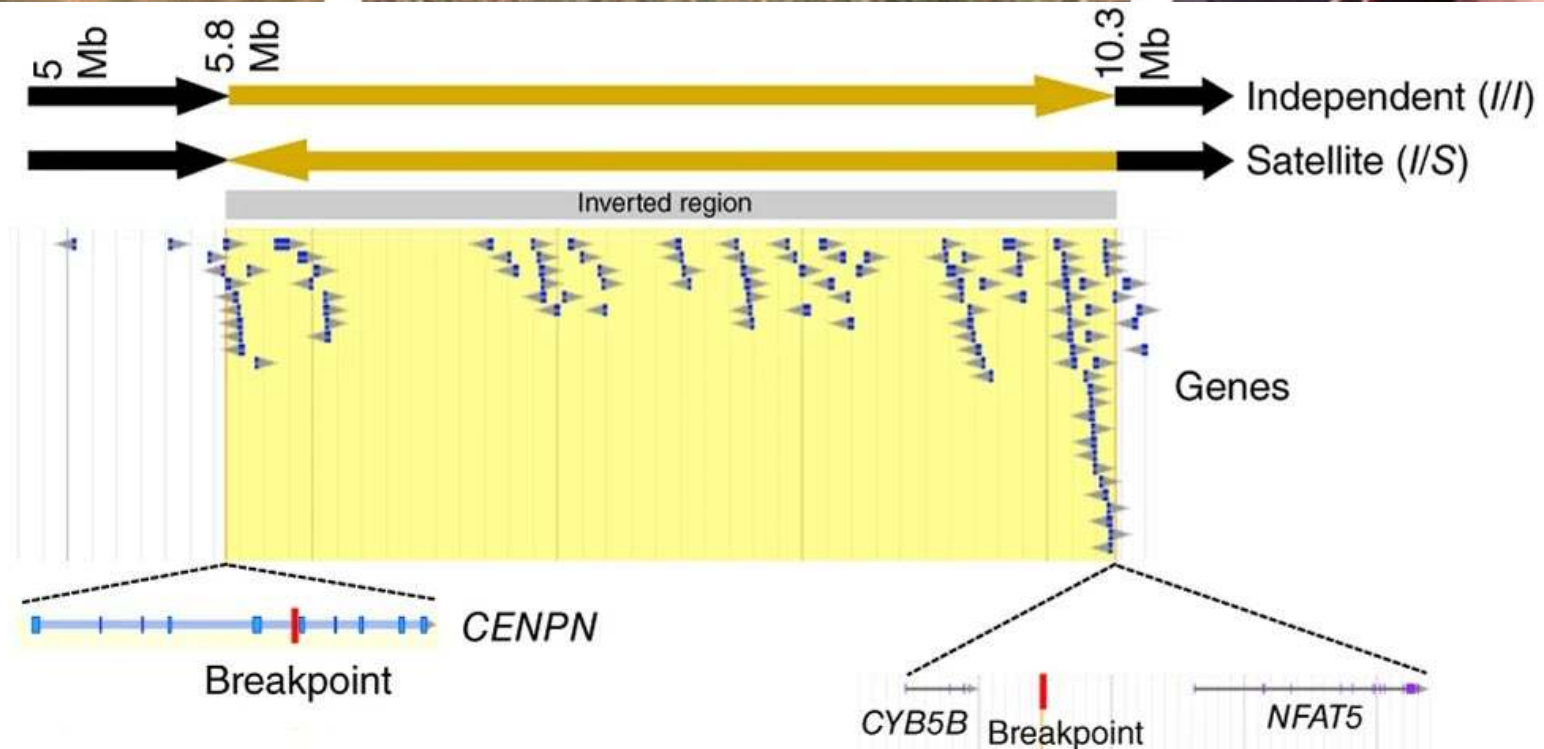
Inversions in ruff reproductive strategies

Independent

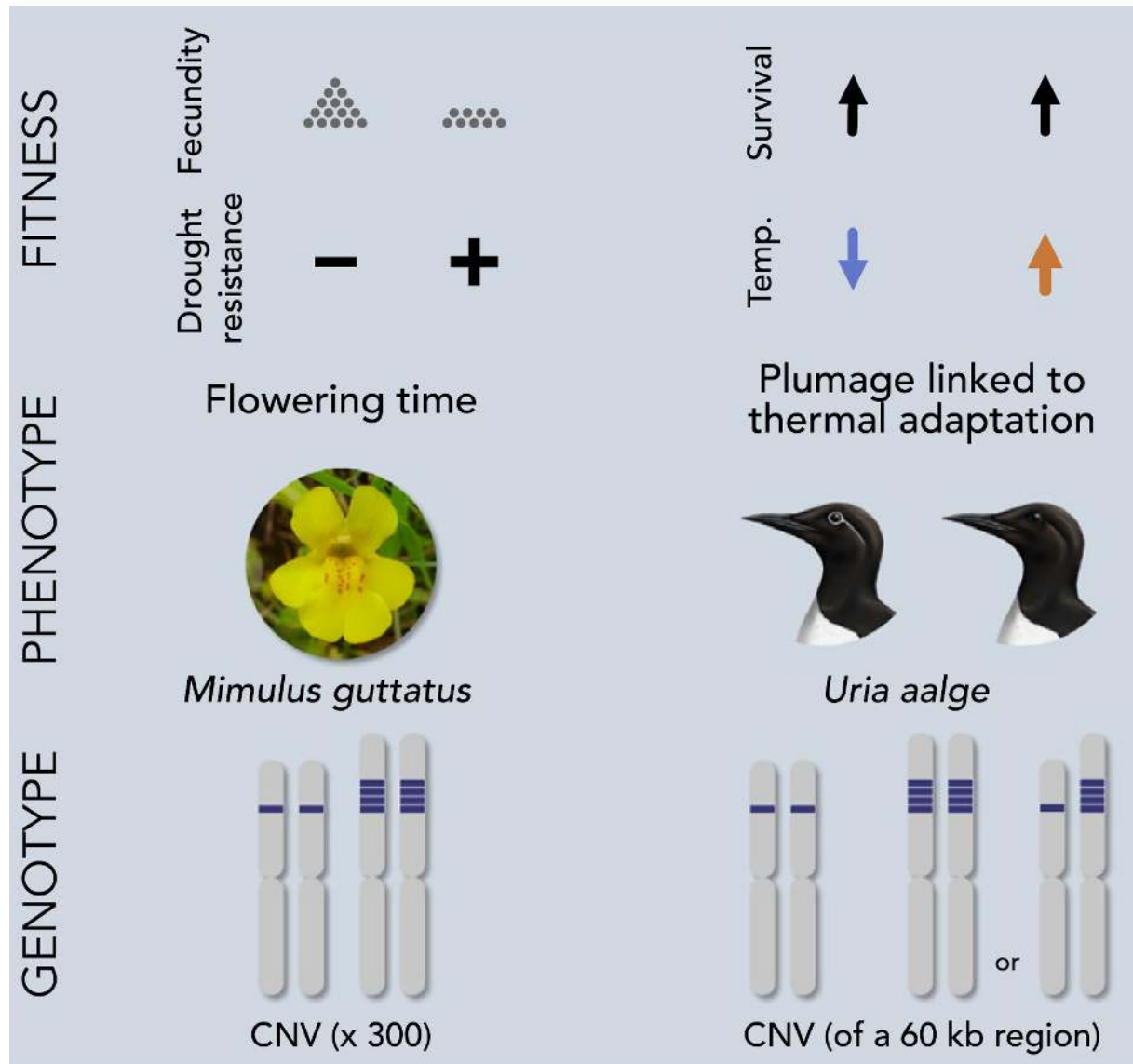
Satellite

Independents

Faeder

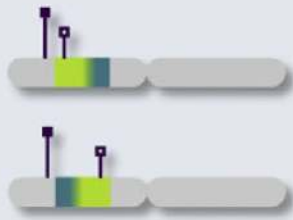


Duplications



High diversity of possible effects

Gene-regulation decoupling



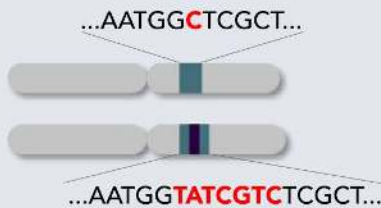
Position-effect variegation



Gene dosage



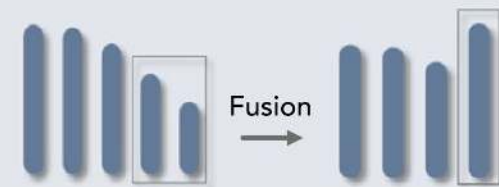
Linear sequence



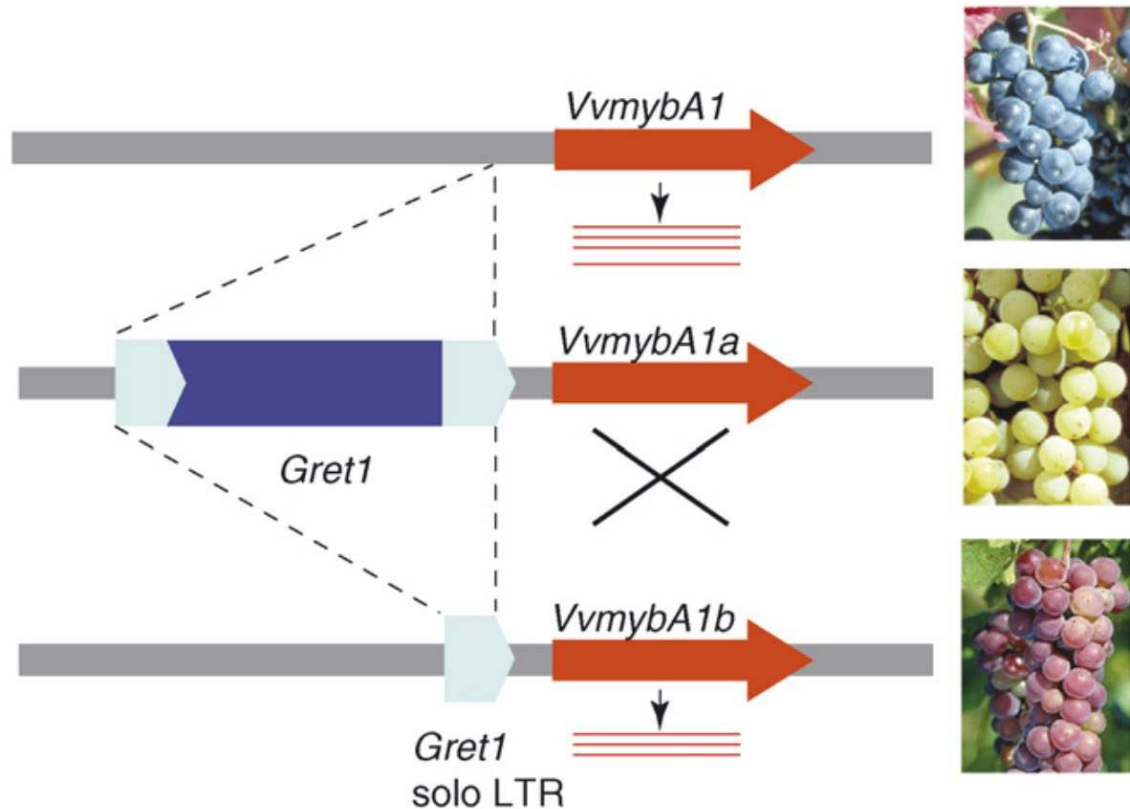
Recombination



Karyotype divergence



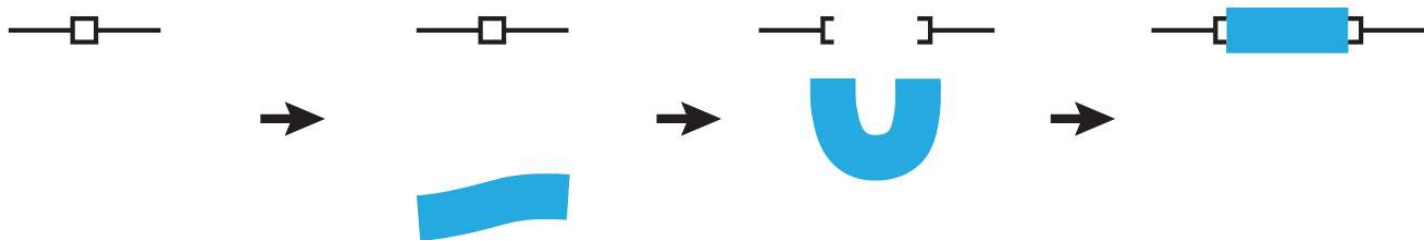
Part 1: Surprise



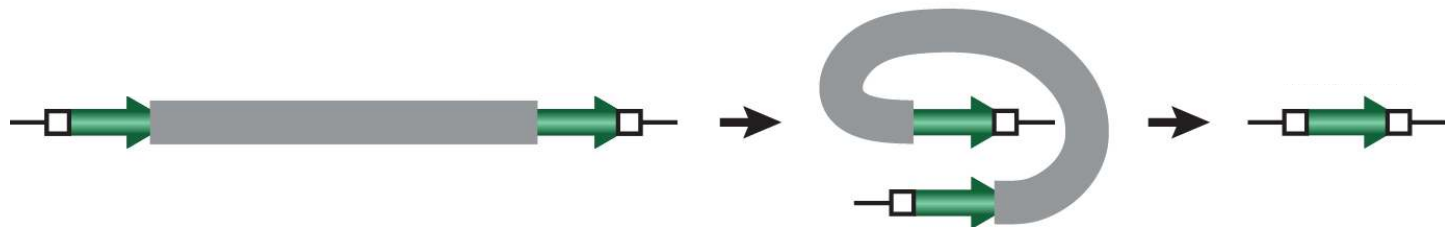
B) Covariation

Two key mechanisms of structural change

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

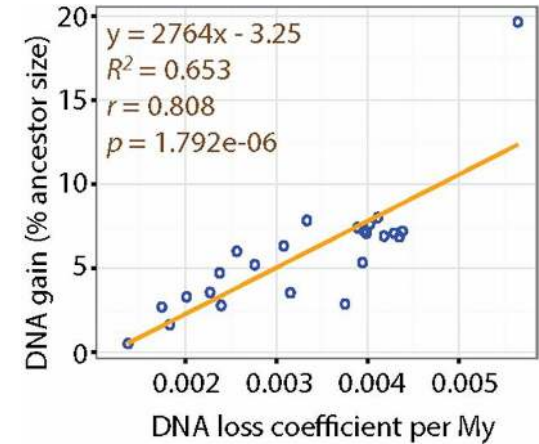
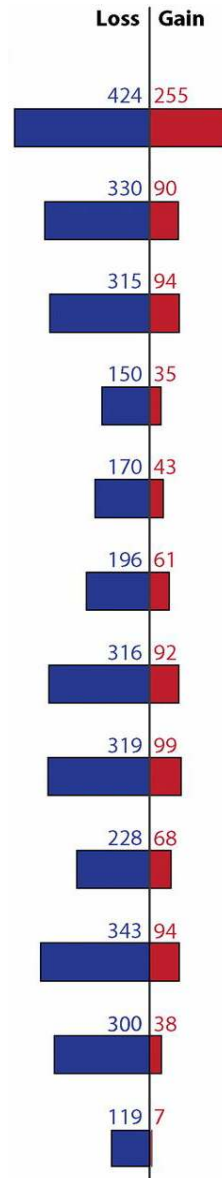
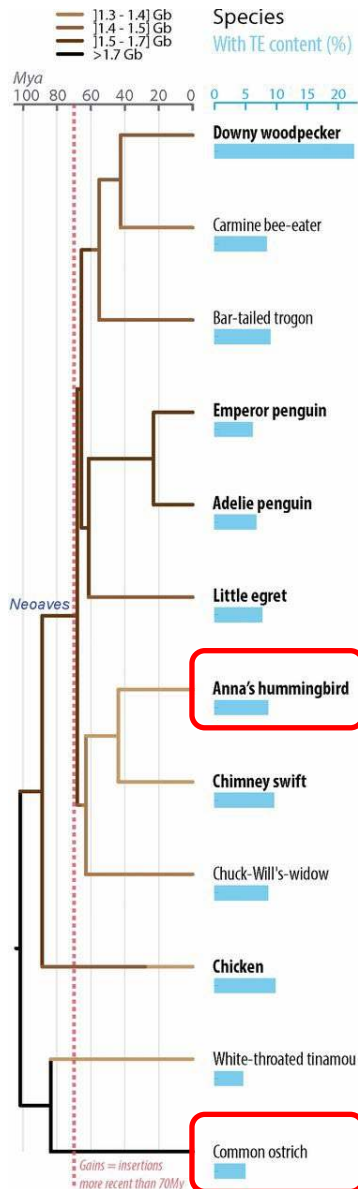


Non-allelic homologous recombination (**NAHR**)
(requires sequence homology)



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

Genome shrinking despite more TEs



Accordion 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



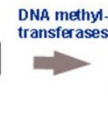
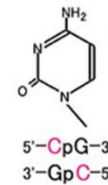
More
context in
[Suh 2021](#)
[TE](#)
[lecture 5](#)

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

Genomes: whack-a-transposon



DNA methylation



Methylated DNA



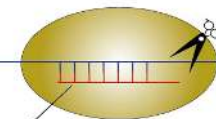
● Methylation

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

piRNA pathway



PIWI and/or AUB



Sense transposon transcript

Antisense piRNA

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

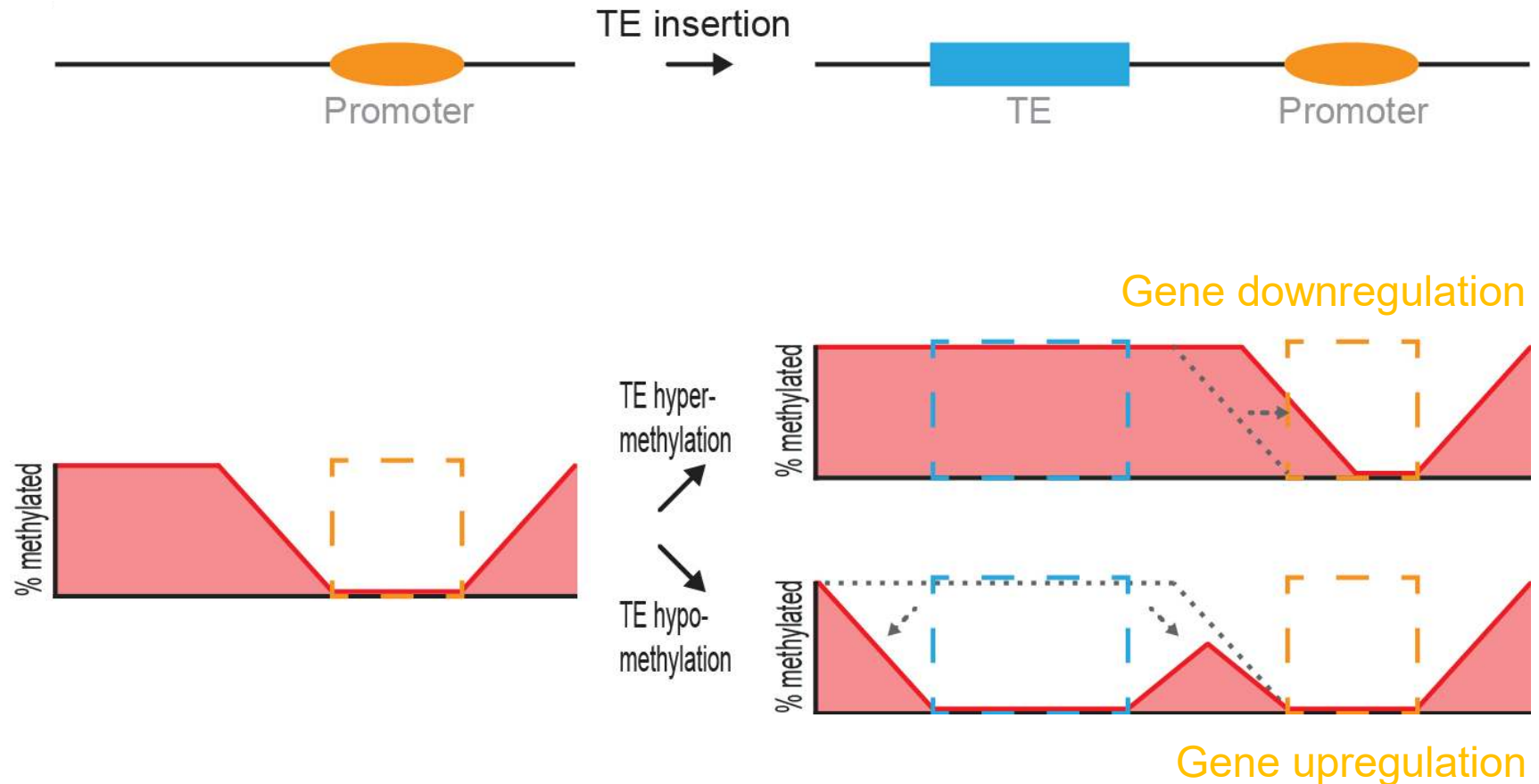
KRAB zinc-finger genes



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

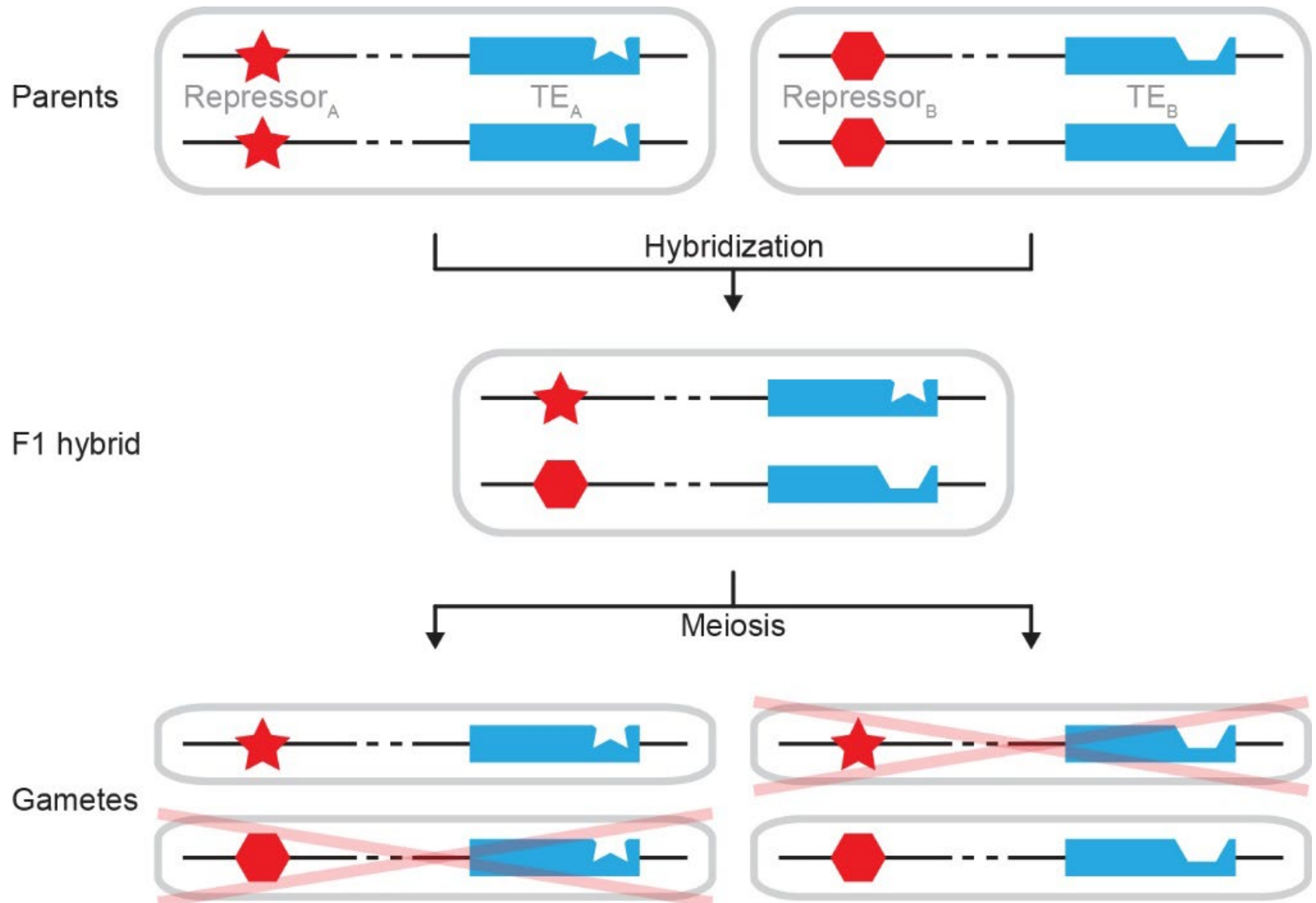
More context in [Suh 2021 TE lecture 6](#)

Covariation between (epi)mutation types

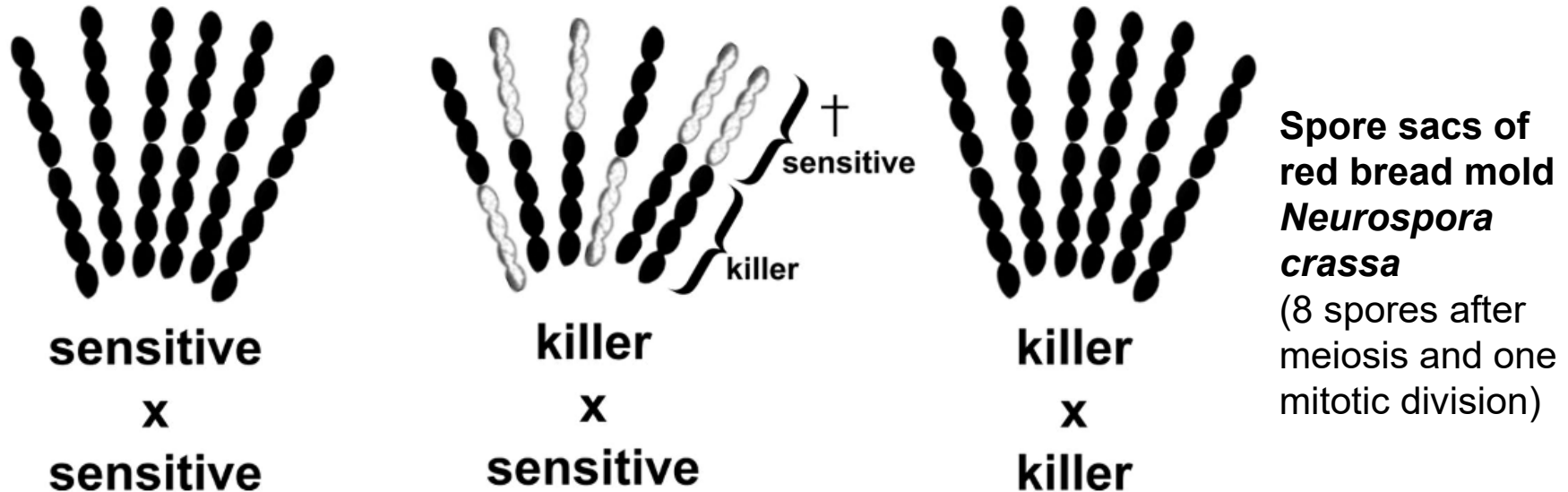


Spillover of DNA methylation and/or histone modifications from new TE insertions to nearby genes!

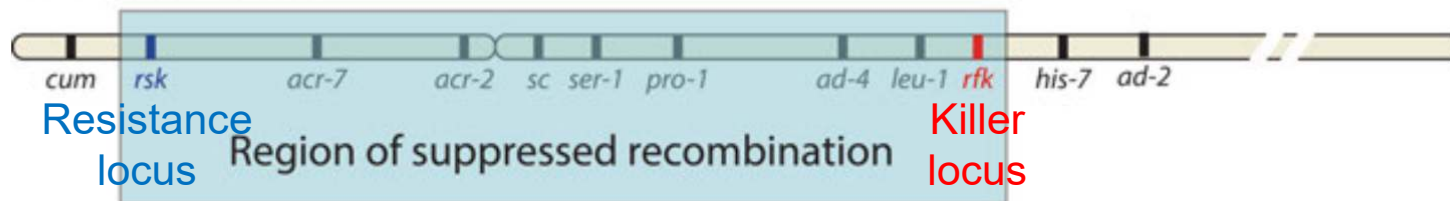
Host-TE conflict and reproductive isolation



Spore/sperm killing of some SVs

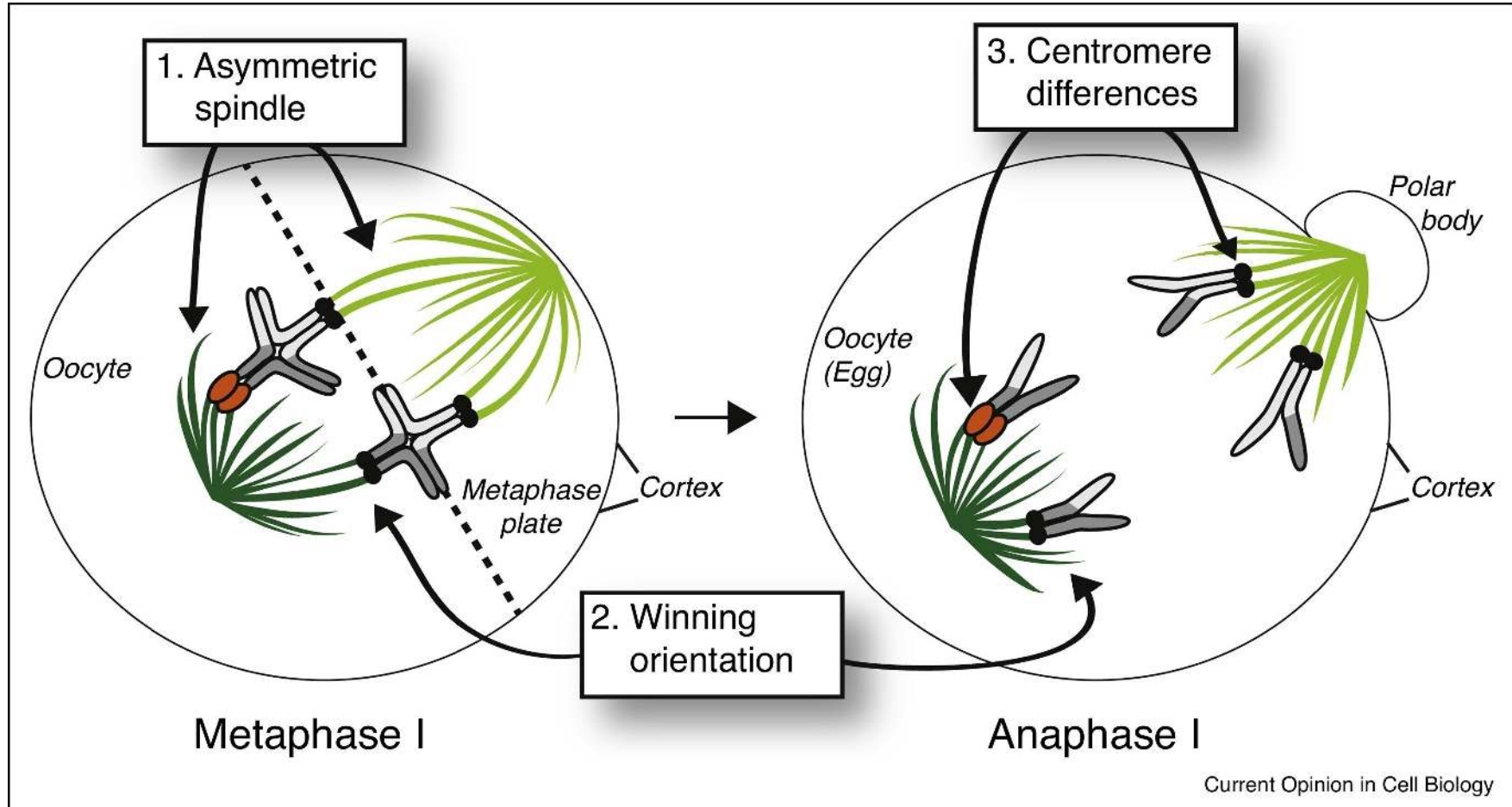


Chromosome 3



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

Centromere drive of some SVs



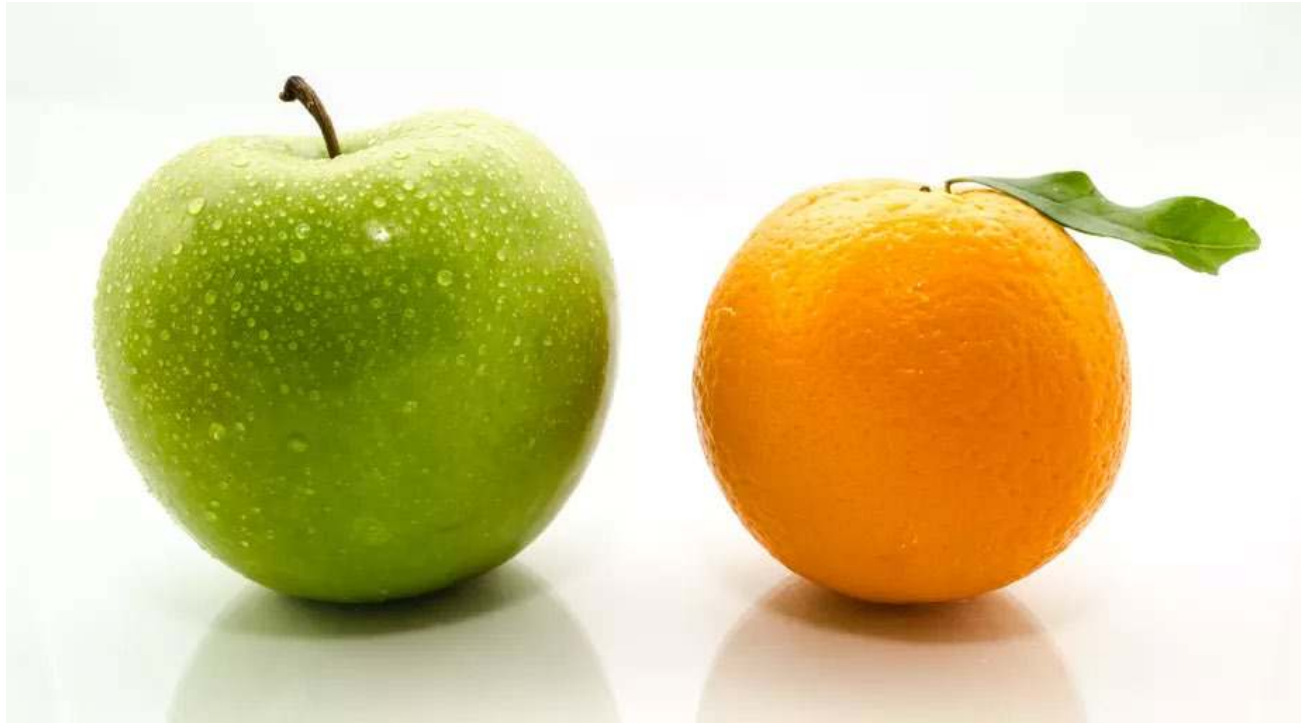
If a pericentric inversion or a centromere shift leads to a stronger centromere, it can distort its own transmission!

Questions?

??!



Part 2: Frustration



A) Concepts and methods

What this lecture will not cover

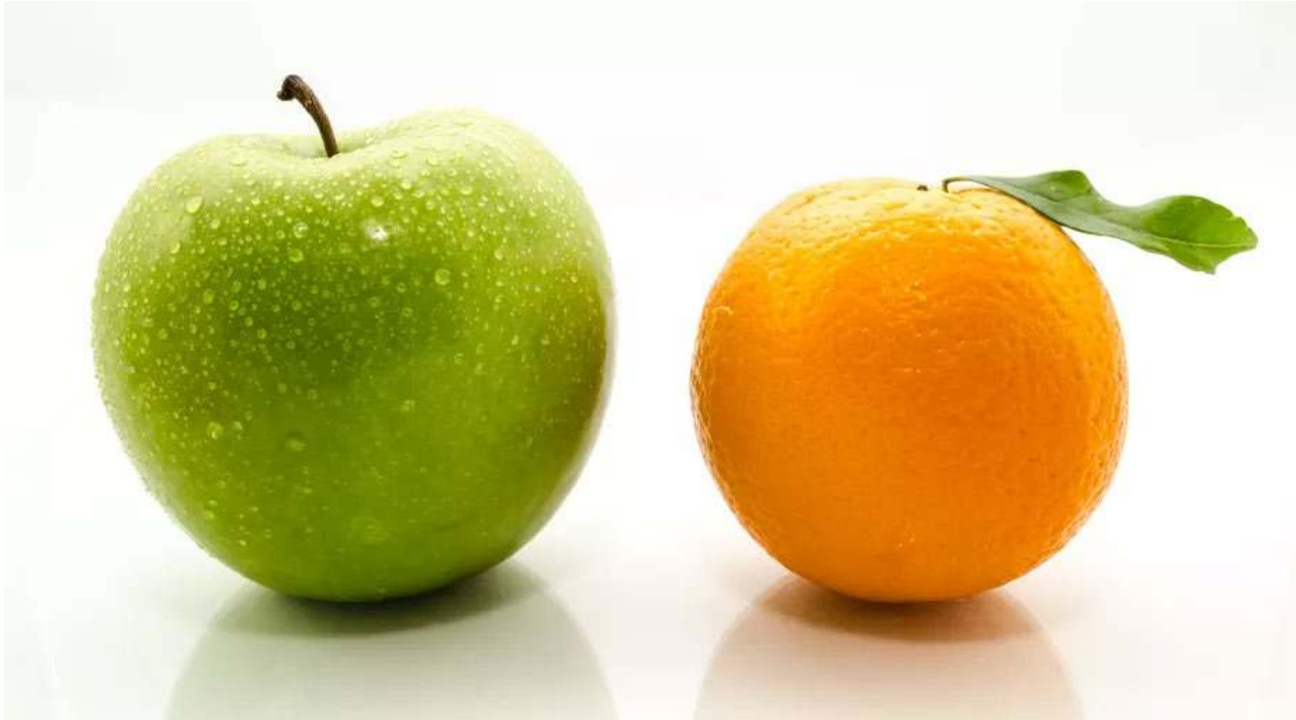
1. Genome assembly: What is (not) assembled?
Primers: [Peona et al. 2018](#), [Peona et al. 2021](#), [Rhie et al. 2021](#), [Nurk et al. 2022](#)
2. Gene and repeat annotation: What is (not) annotated?
Primers: [Yandell & Ence 2012](#), [Suh 2021 TE lecture 4](#), [Goubert et al. 2022](#)
3. Within-individual 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 and me until 10 pm today!

2p – 5p	Alexander Suh	Structural variation	Egon Schiele
7p – 10p	Valentina Peona	Structural variation activity	Prelate



Valentina Peona

Awareness of biology and technology



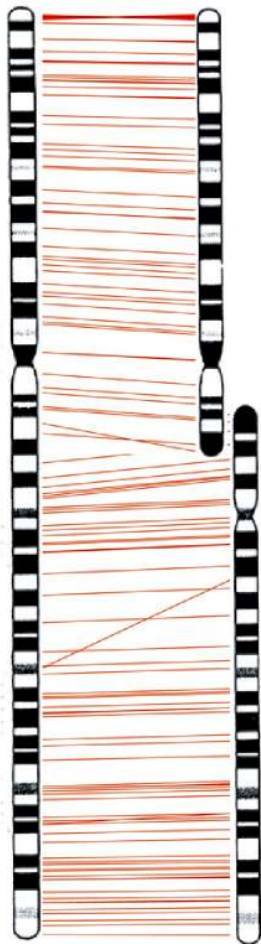
How can we make sure that what we see in our data is what we think it is?

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

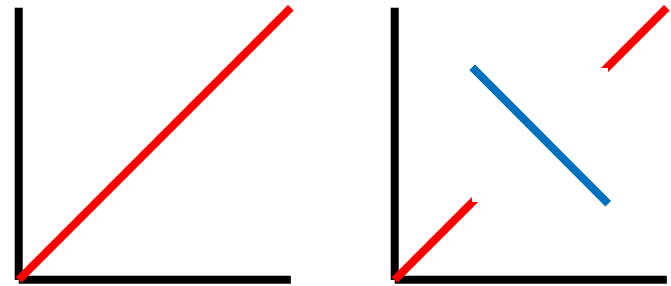
Terminology

Synteny vs. collinearity

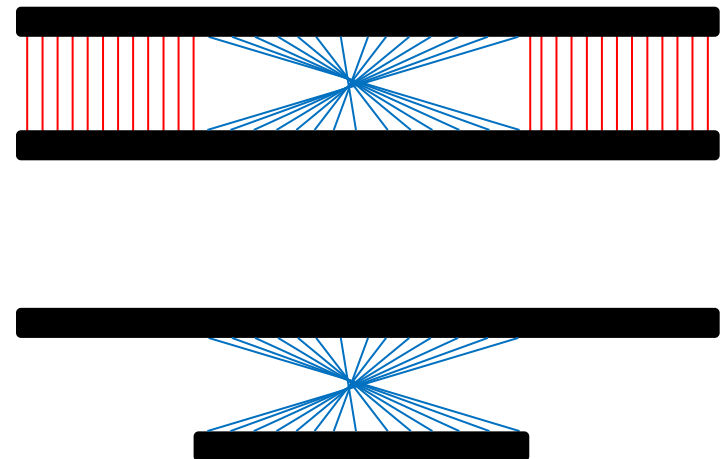
Hs2 Pt12/13



Dot plot



Pattern vs. process



Beware of waves

My SNP
explains
everything!

My inversion
explains
everything!

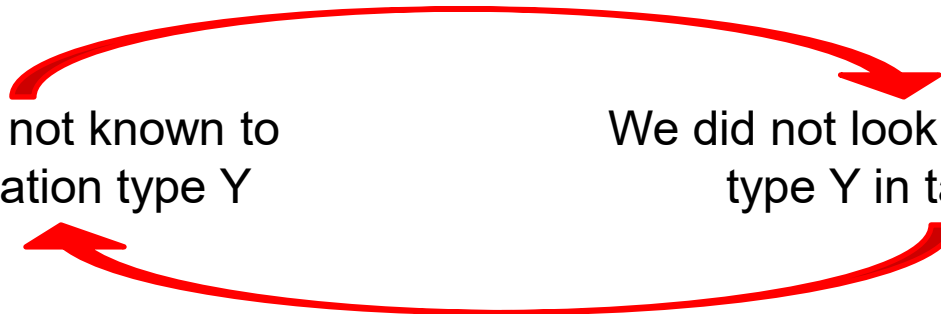
My TE
explains
everything!



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

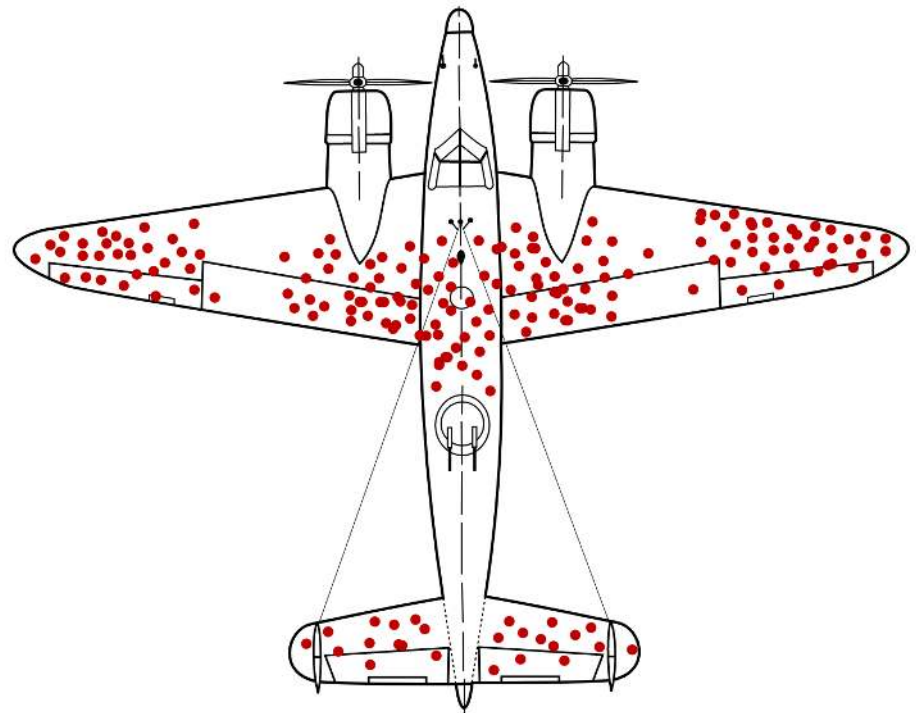


Reflection on biases

Confirmation
bias



Survivorship
bias



My own biases: I like transposable elements, centromere shifts, and simple 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~~

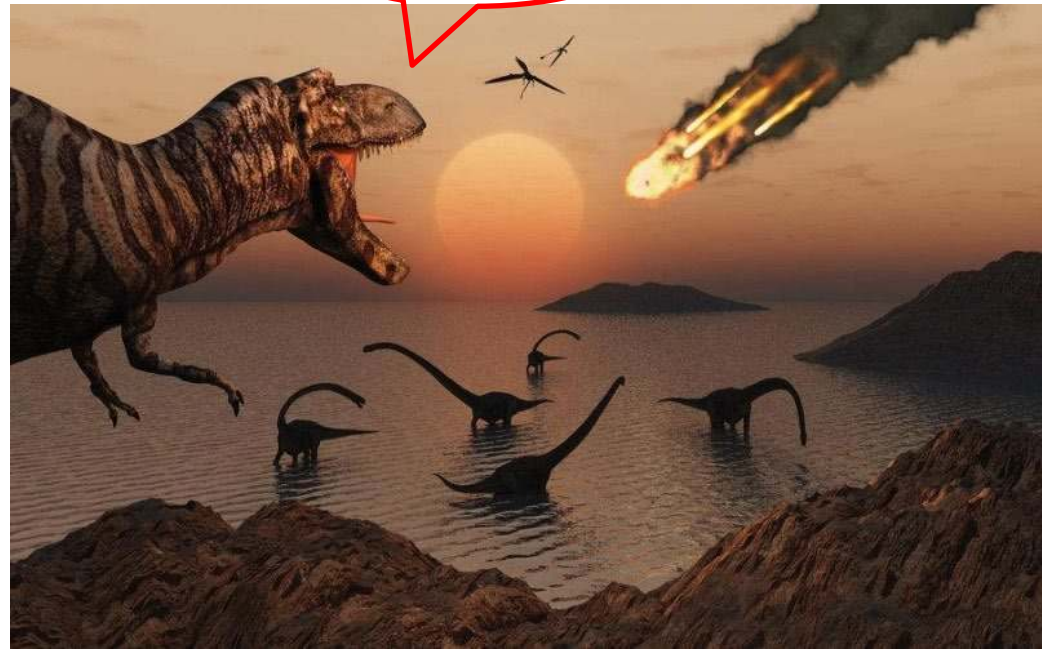
TEs jump because they can

ROAR

Proximate: This asteroid caused diversification

Ultimate: ~~Asteroids land to cause diversification~~

Asteroids land eventually



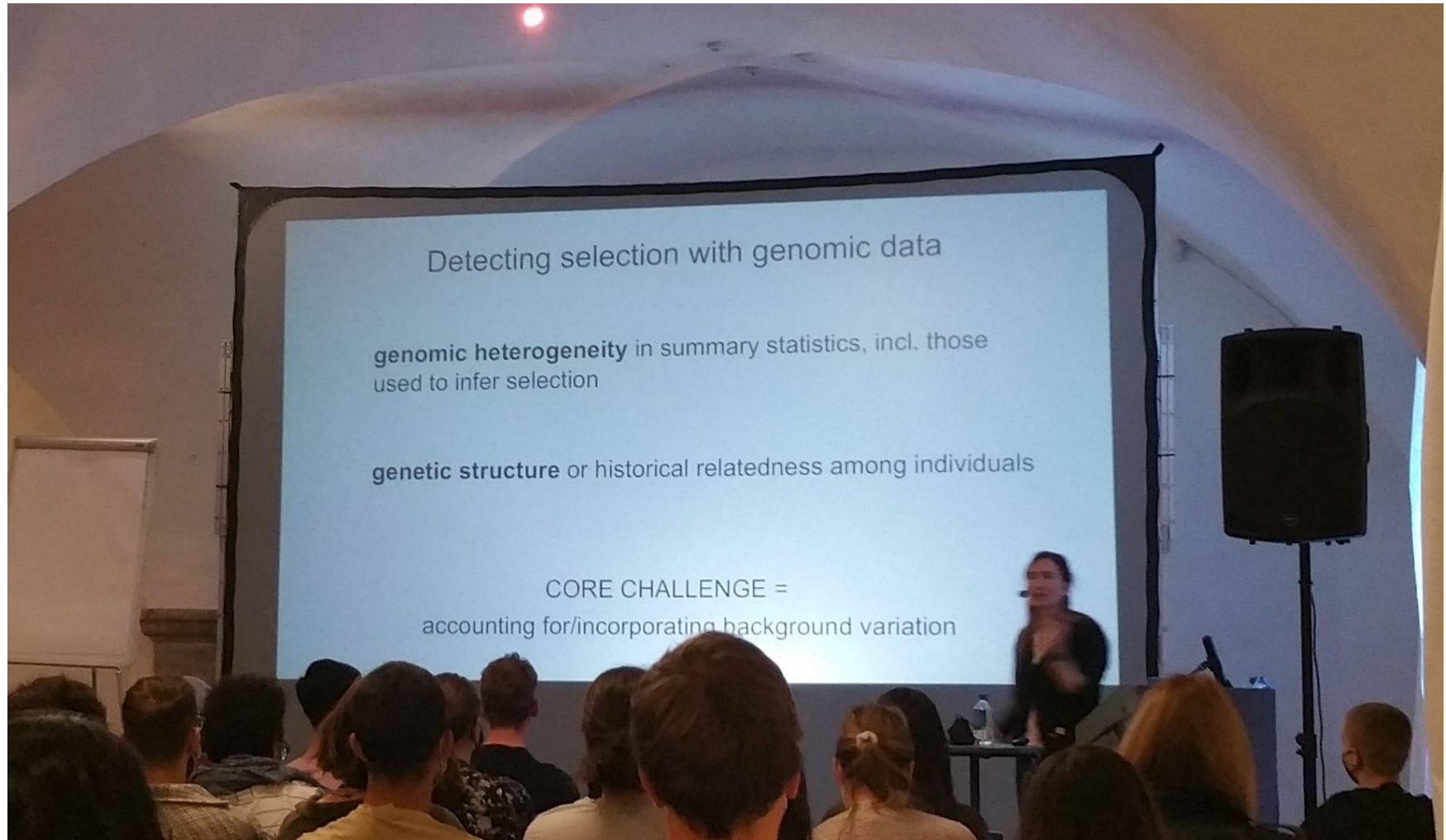
What is the null hypothesis?

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

~~Absence of evidence~~
Evidence of absence

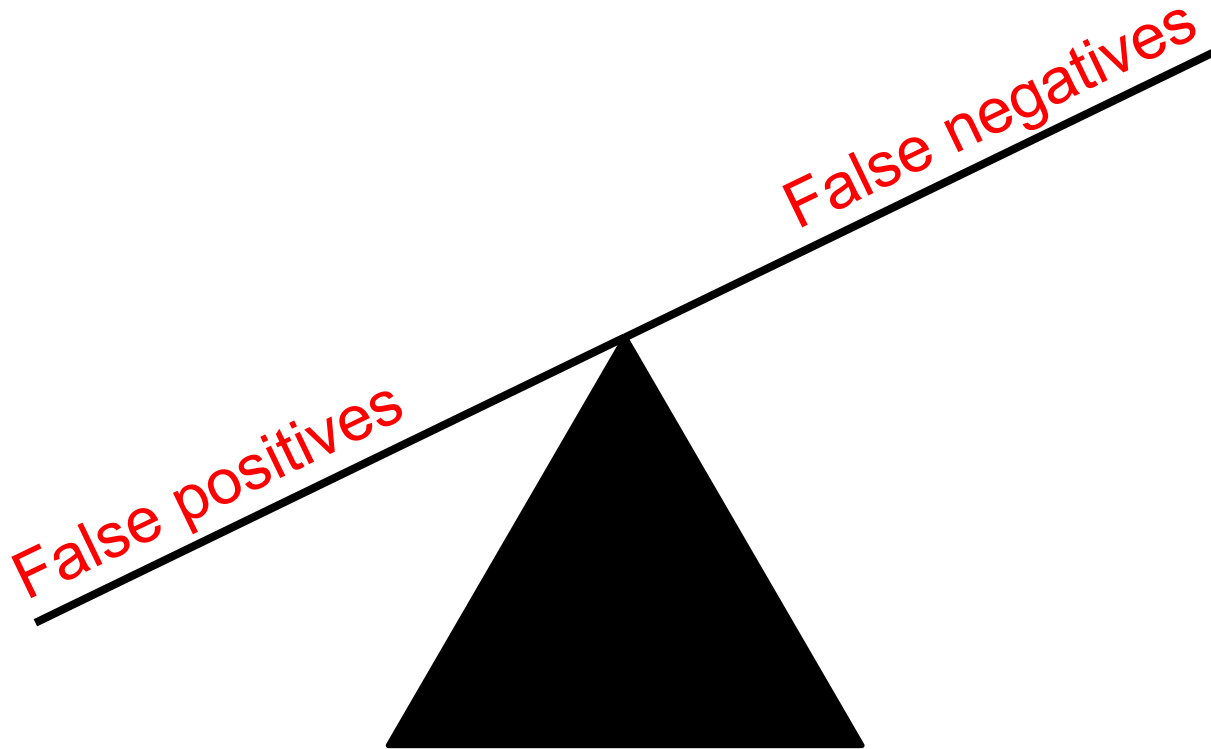


Theory applies to SNPs and to SVs



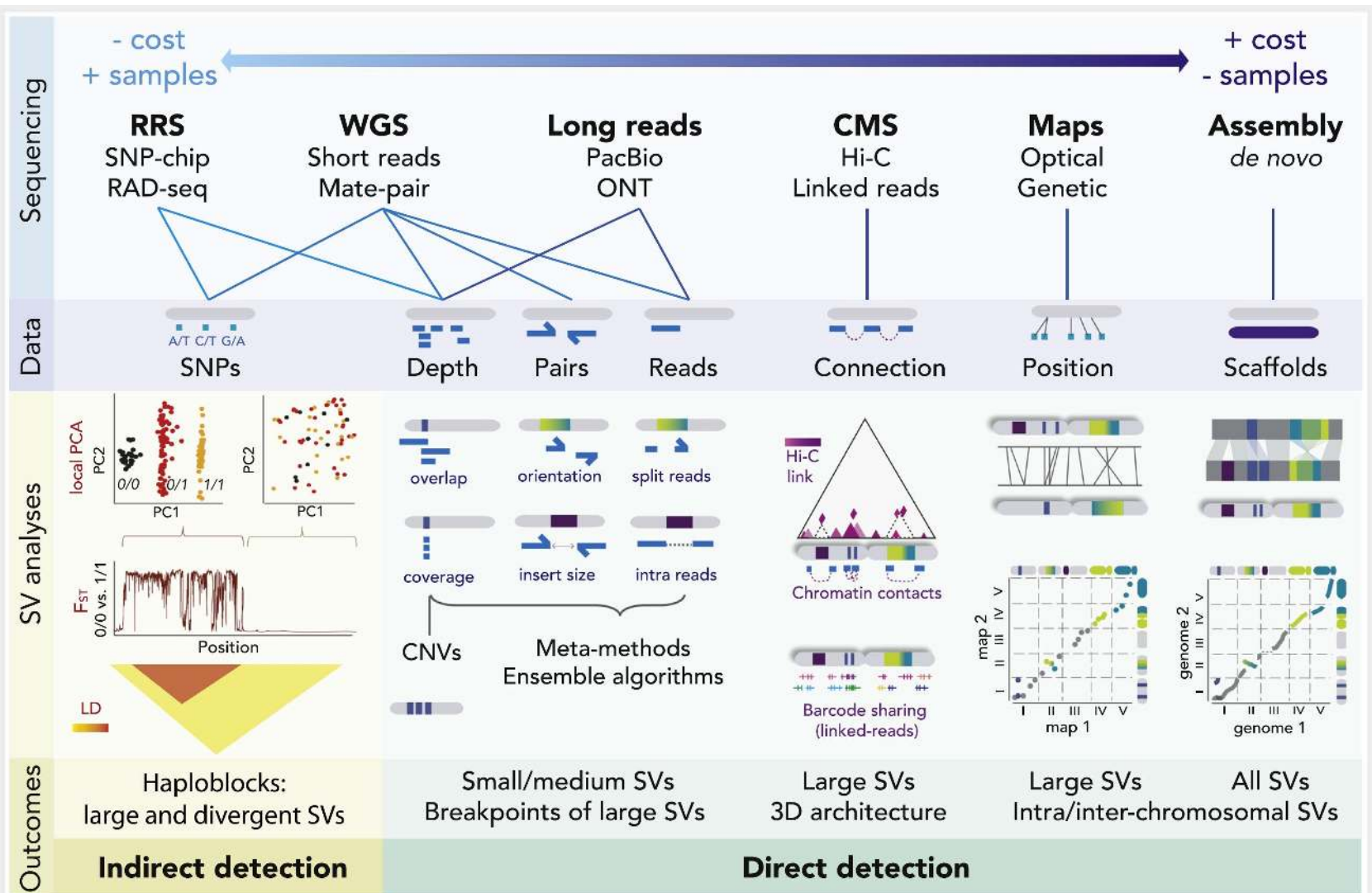
Background variation: What SNPs and SVs are there?

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](#).



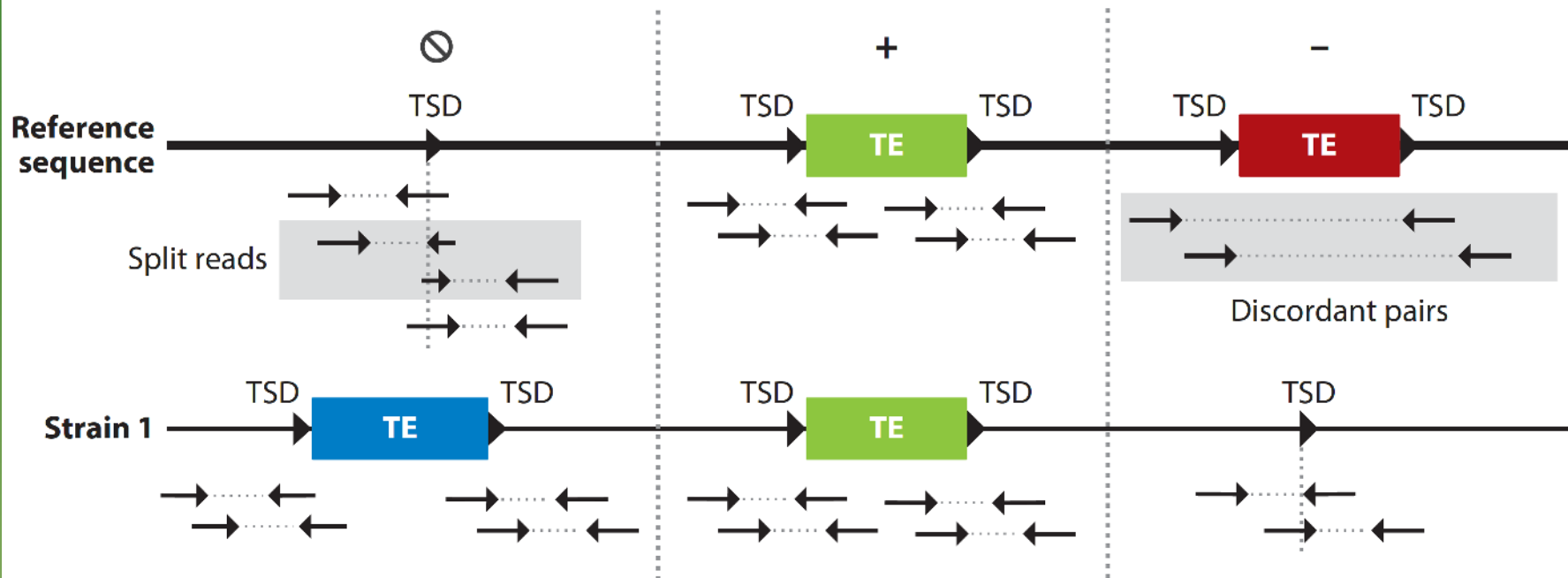
Overview of tools for repeat analysis

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

88 tools listed for TE insertion polymorphism analysis!

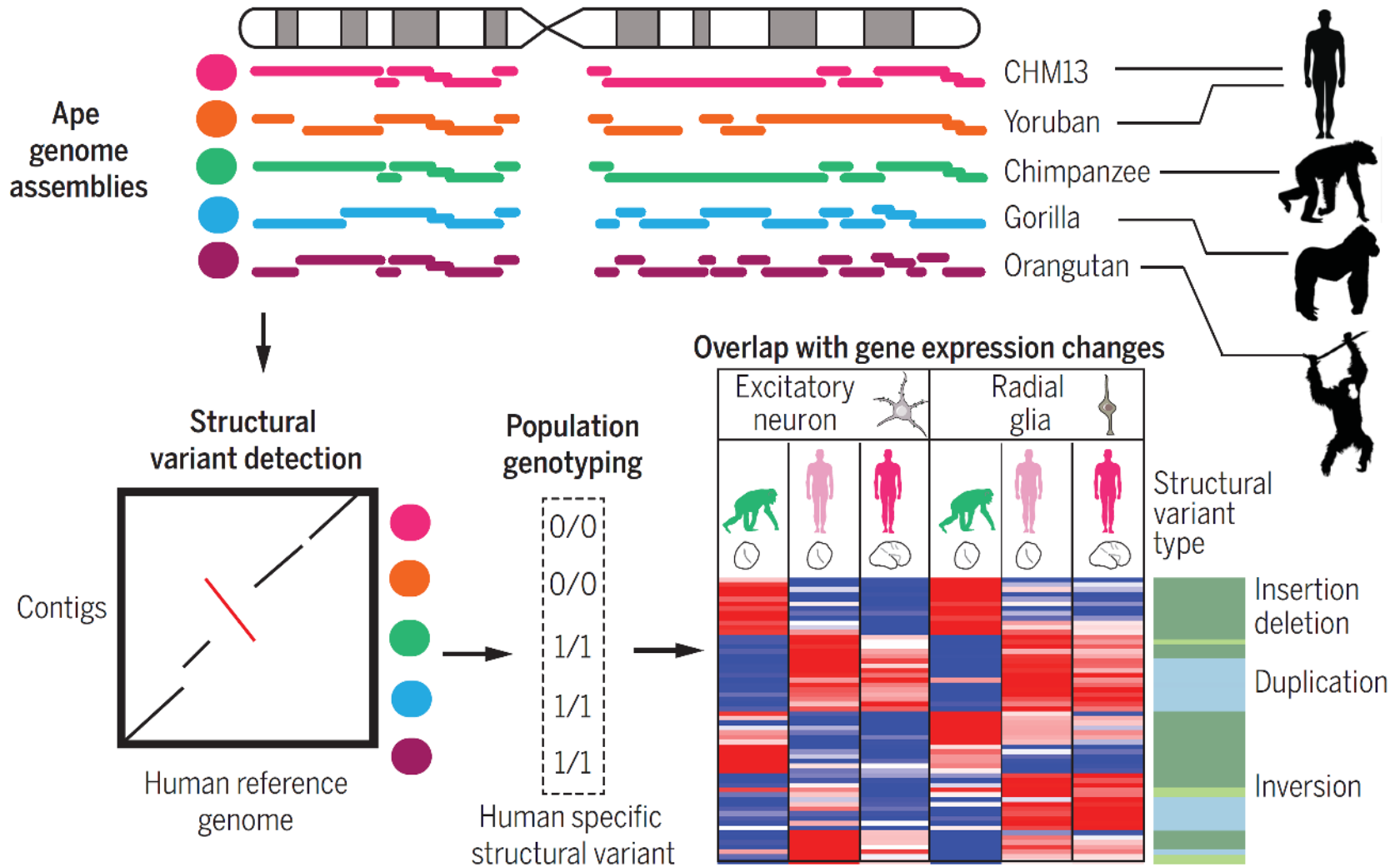
https://tehub.org/en/resources/repeat_tools; The TE Hub Consortium 2021, *Mobile DNA*

Read-based SV detection



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

Assembly-based SV detection

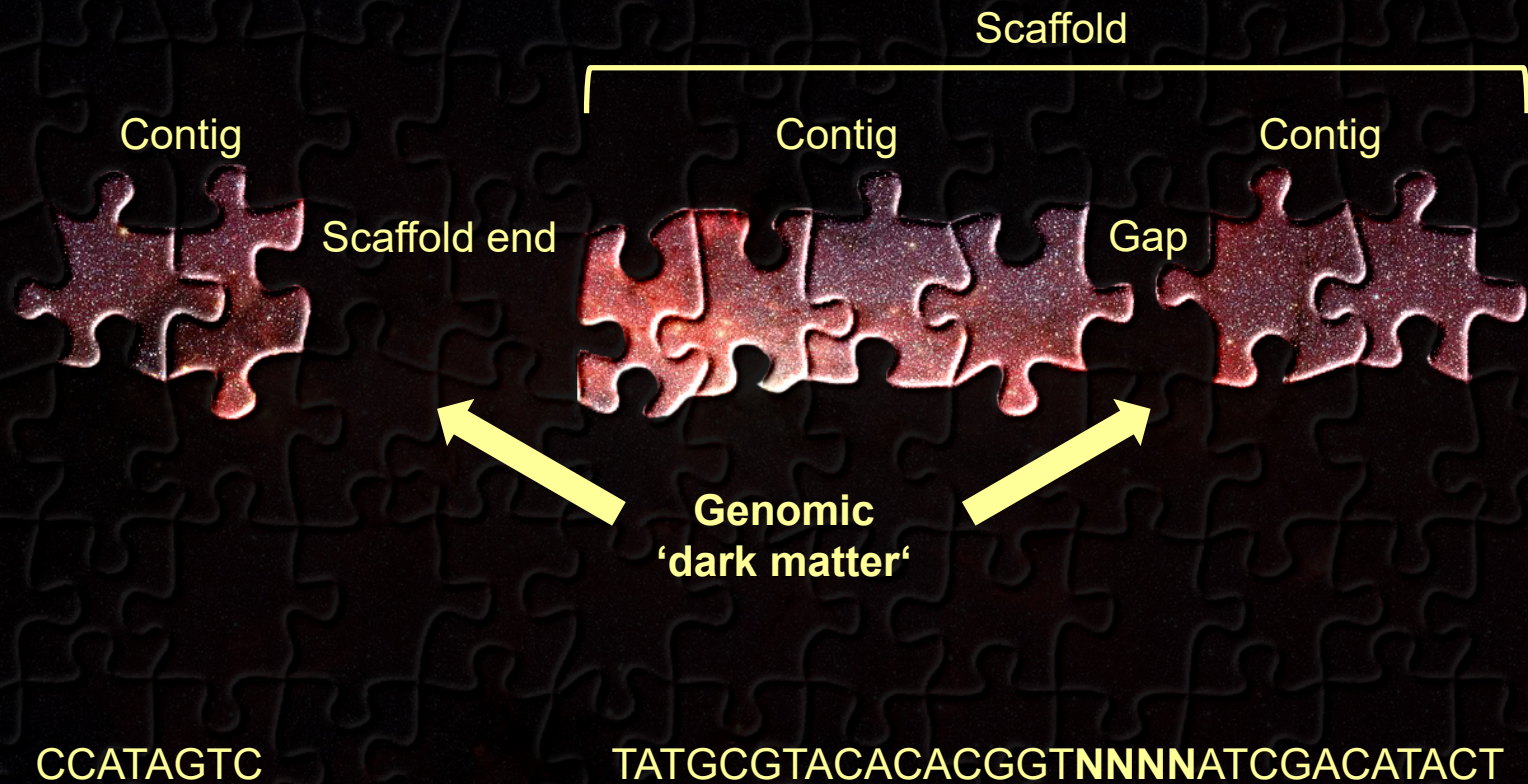


Reliable genome alignment and SV scoring is difficult in highly repetitive regions (if assembled ...)

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



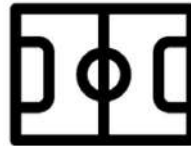
Genomics: a big and messy puzzle



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



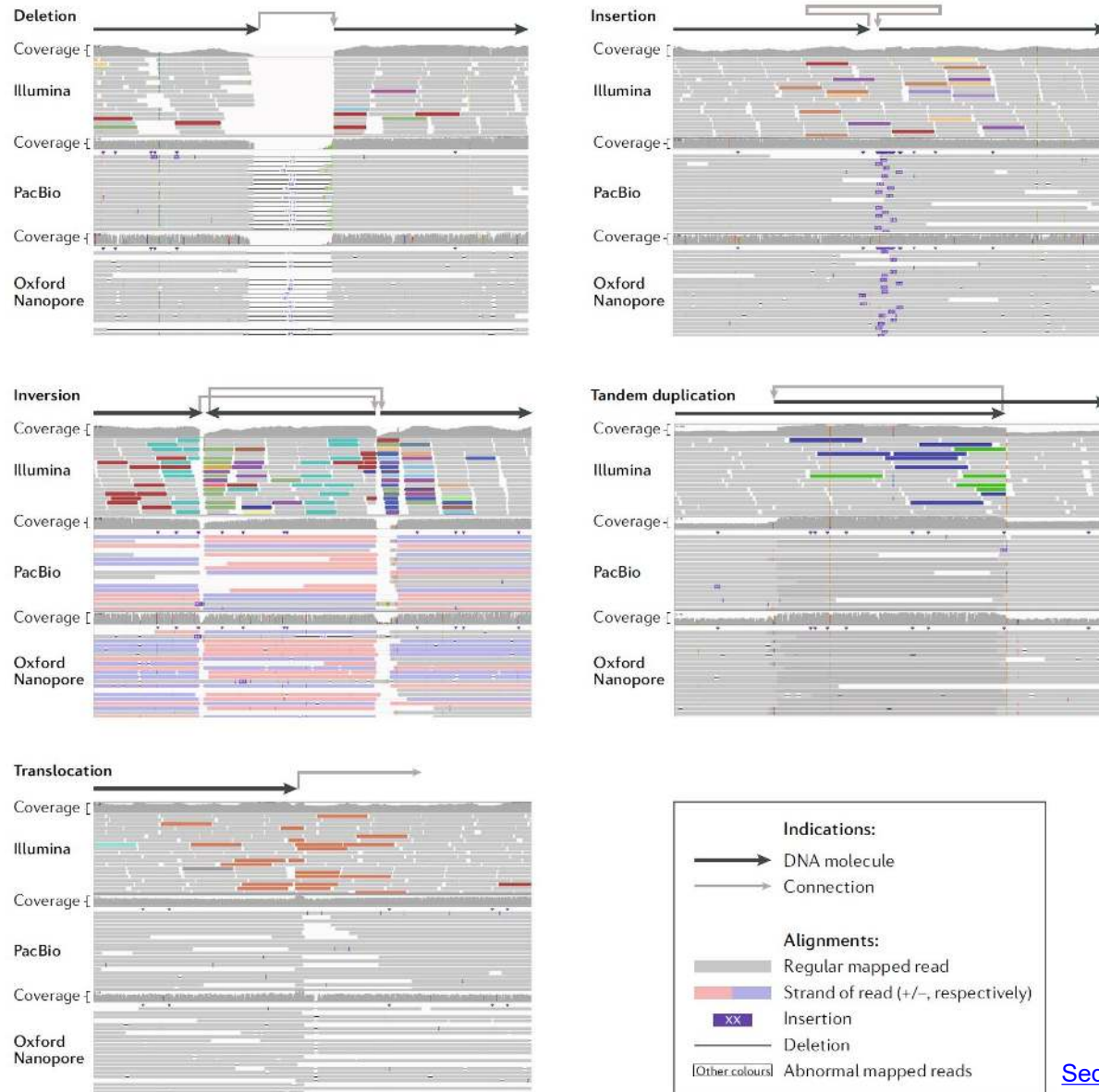
Optical maps



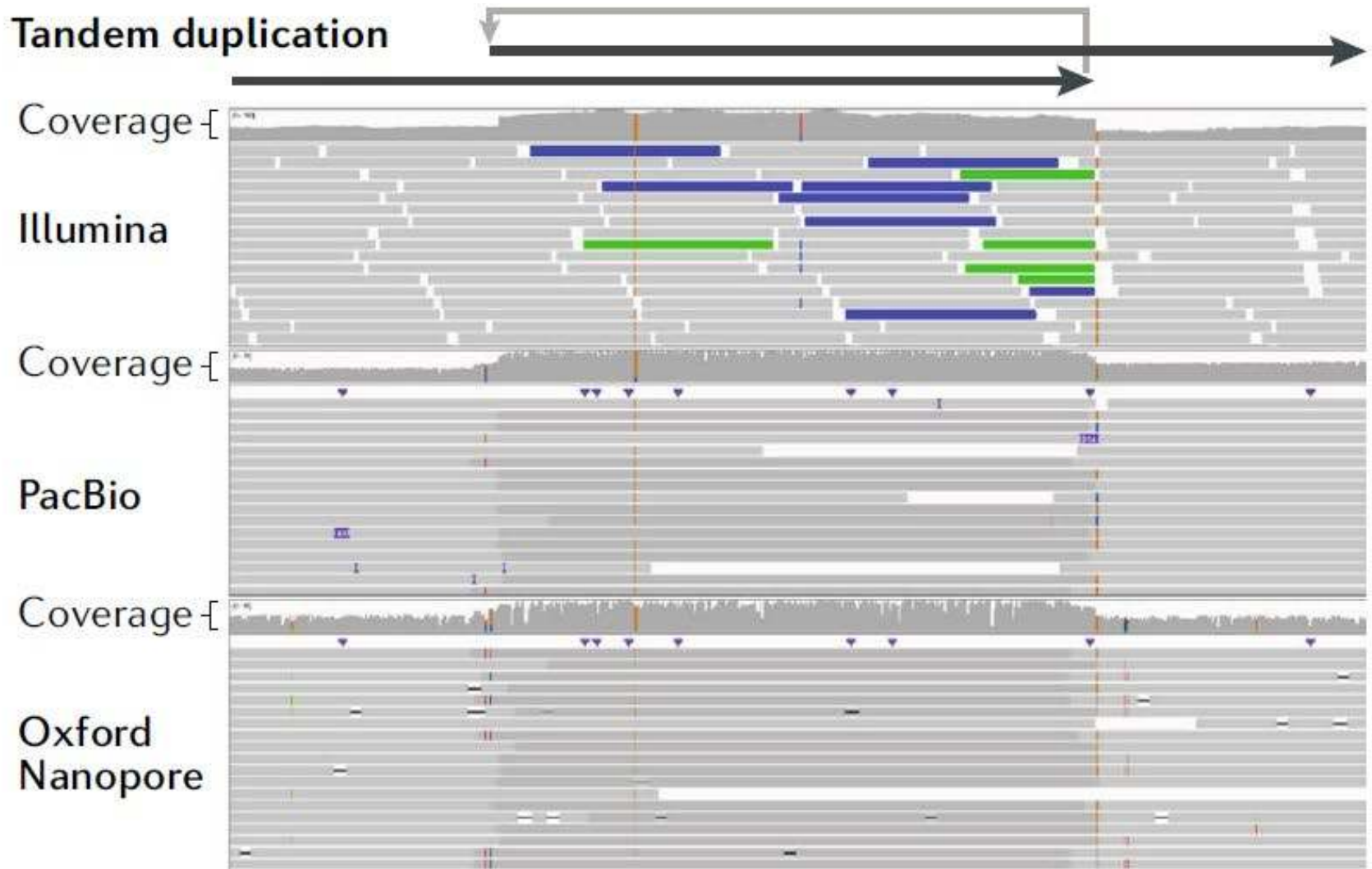
Hi-C maps



SV mapping with longer and longer reads



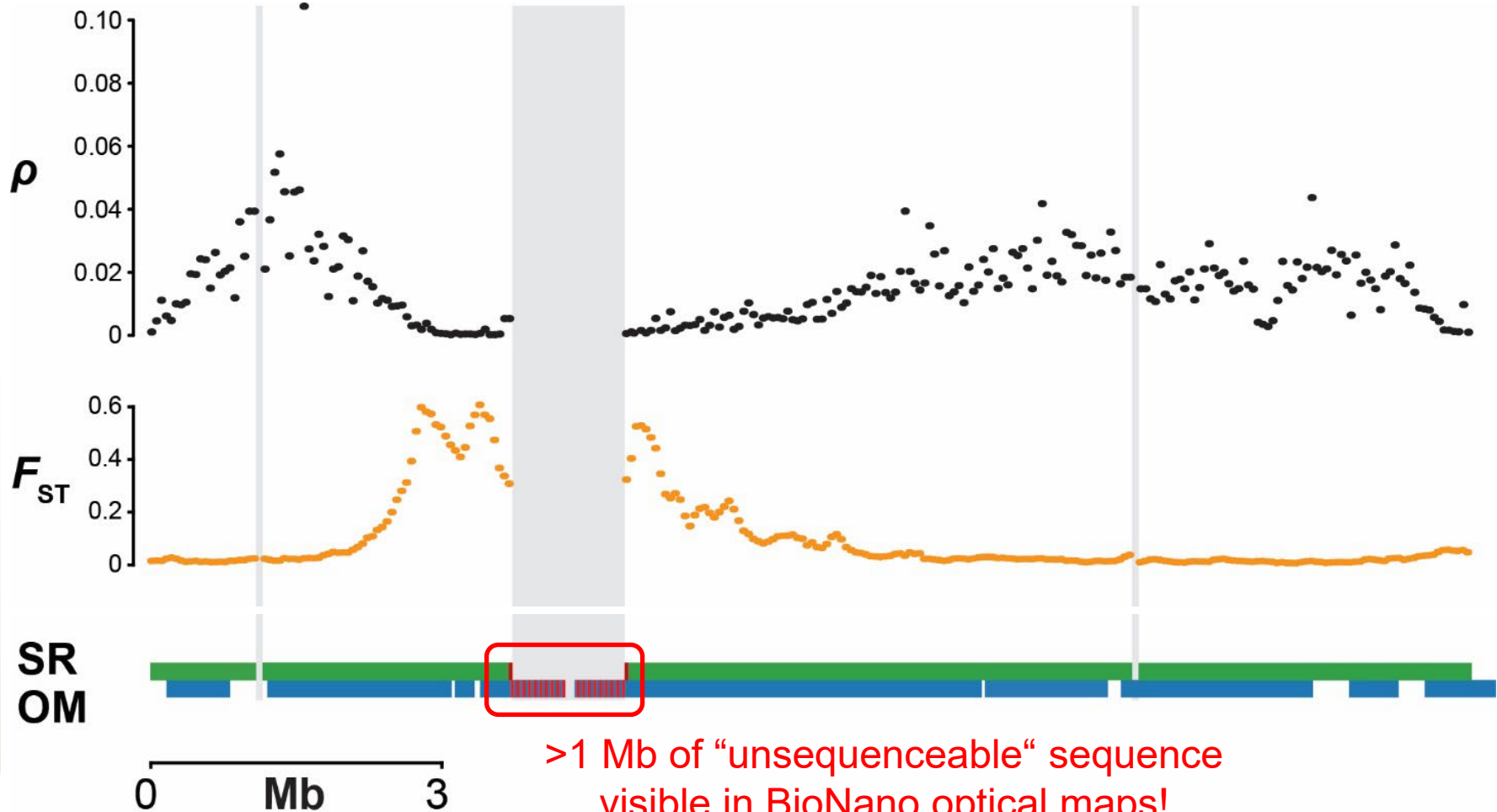
What does coverage variation tell us?



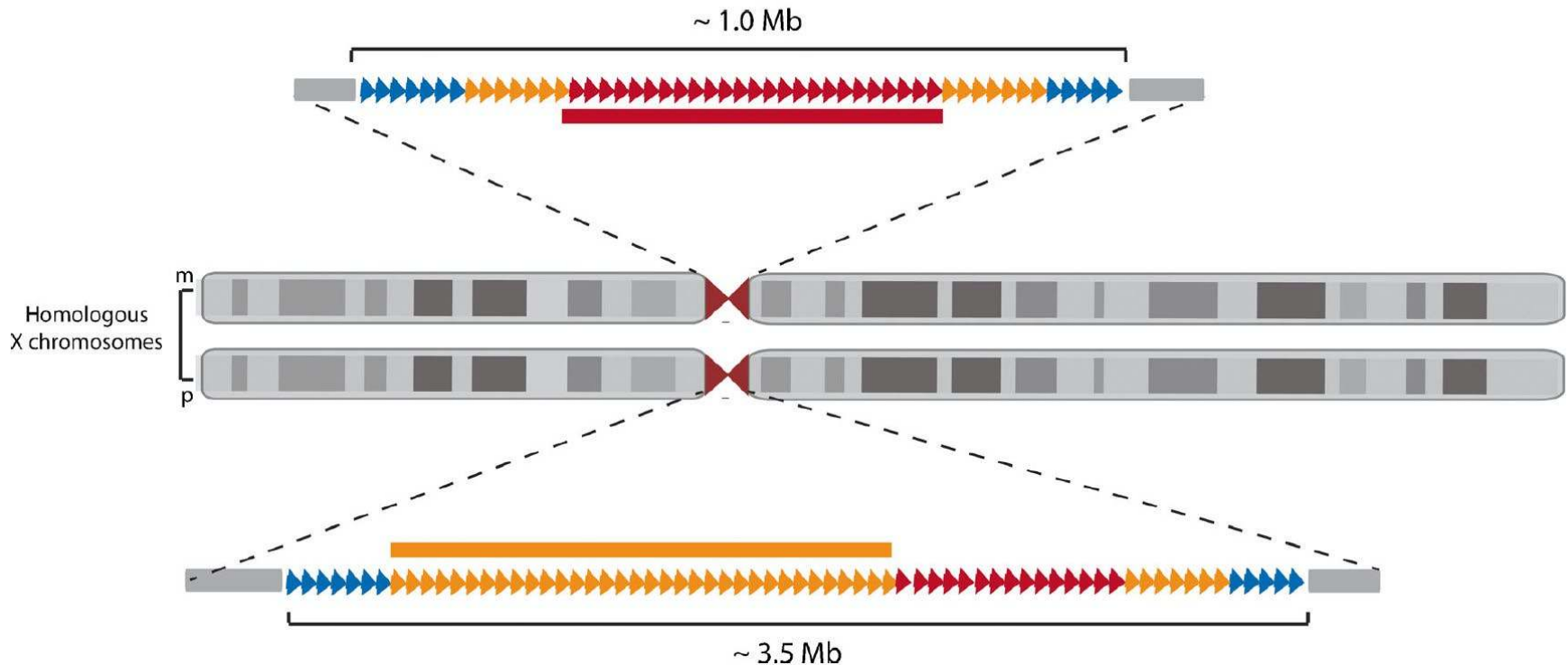
Tandem duplications are (usually) collapsed in assemblies!

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!

Questions?

??!

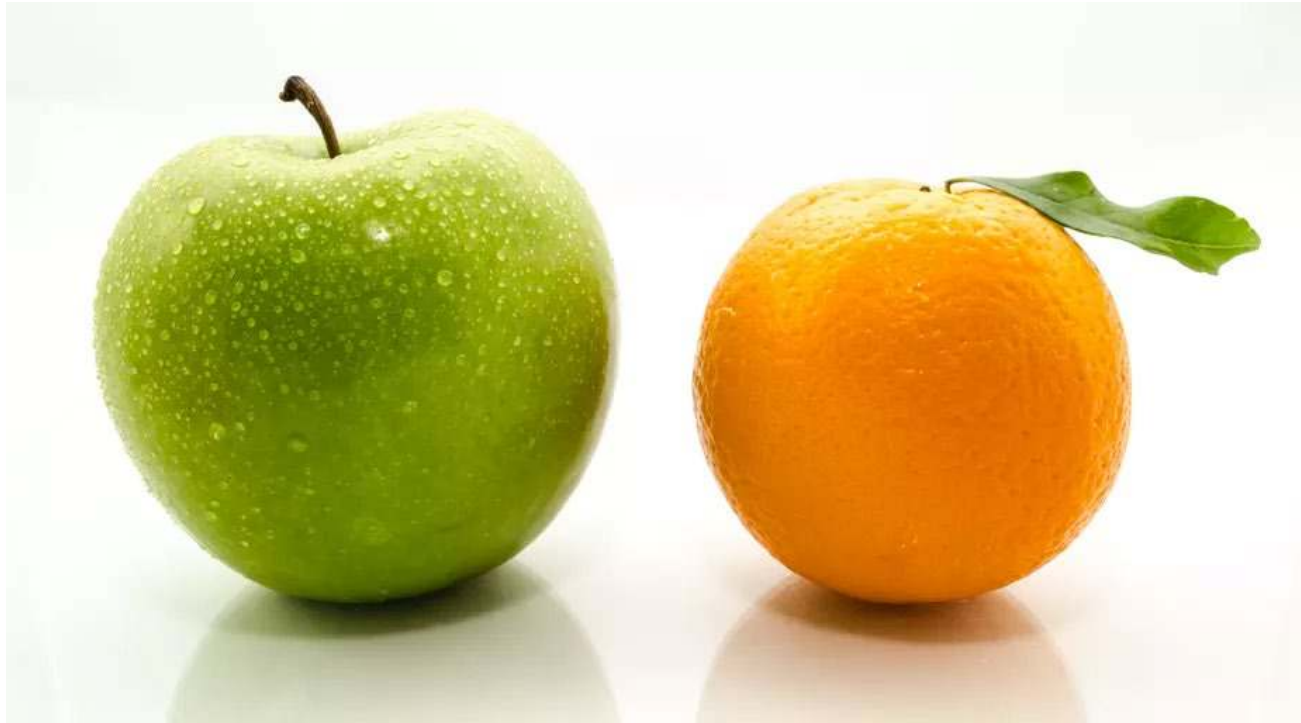


Coffee break (10 minutes)



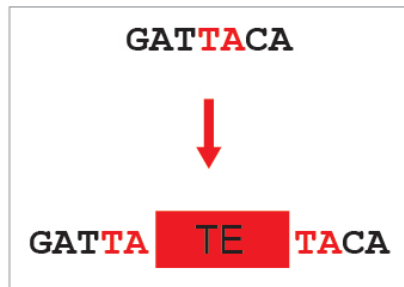
Task: Form random groups of 3 and discuss what resources (assembly quality, gene/repeat annotation) there are for your respective study system.

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		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		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

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

ENV, Envelope protein

GAG, Capsid protein

POL B, DNA polymerase B

RH, RNase H

Tase, Transposase (* with DDE motif)

ATP, Packaging ATPase

HEL, Helicase

RPA, Replication protein A (found only in plants)

YR, Tyrosine recombinase

C-INT, C-integrase

INT, Integrase

ORF, Open reading frame of unknown function

RT, Reverse transcriptase

Y2, YR with YY motif

CYP, Cysteine protease

EN, Endonuclease

Species groups

P, Plants

M, Metazoans

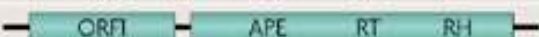
F, Fungi

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	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

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

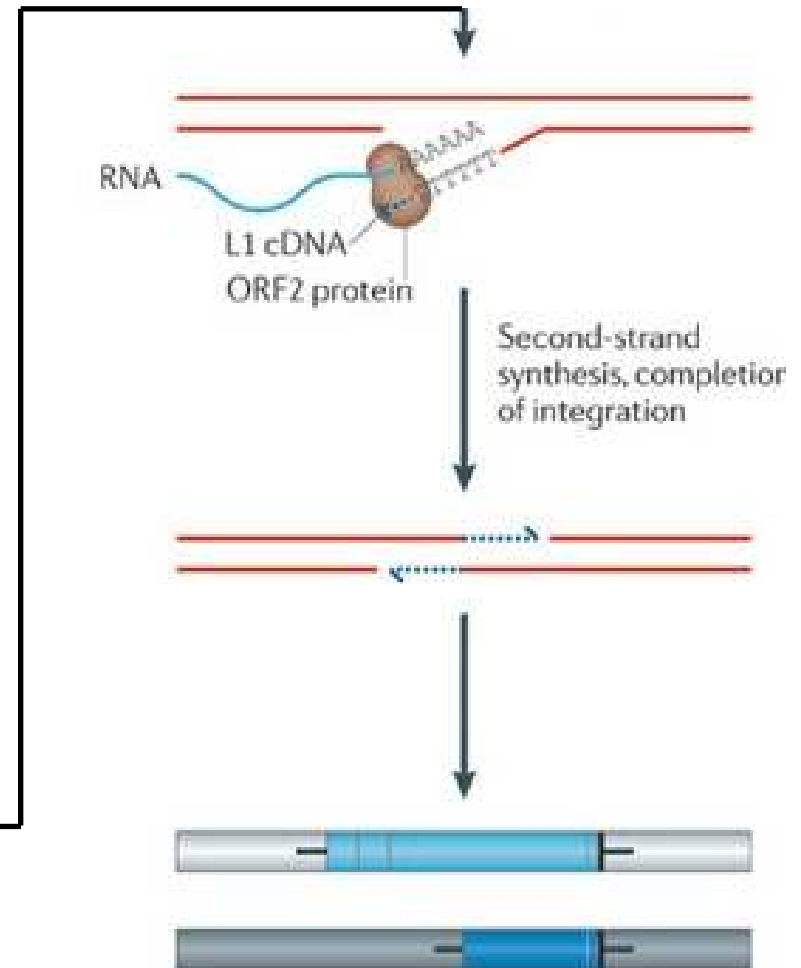
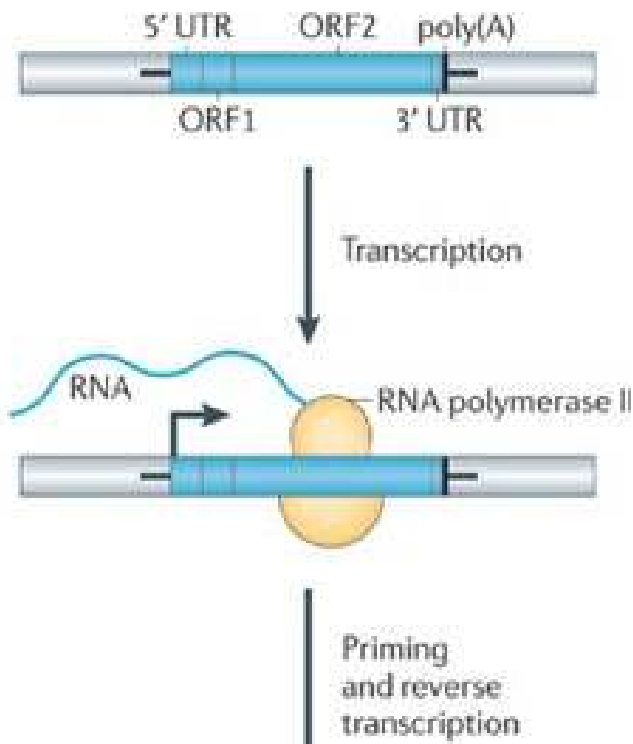
Dear ribosome,
if you read this,
translate me into a
reverse transcriptase

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



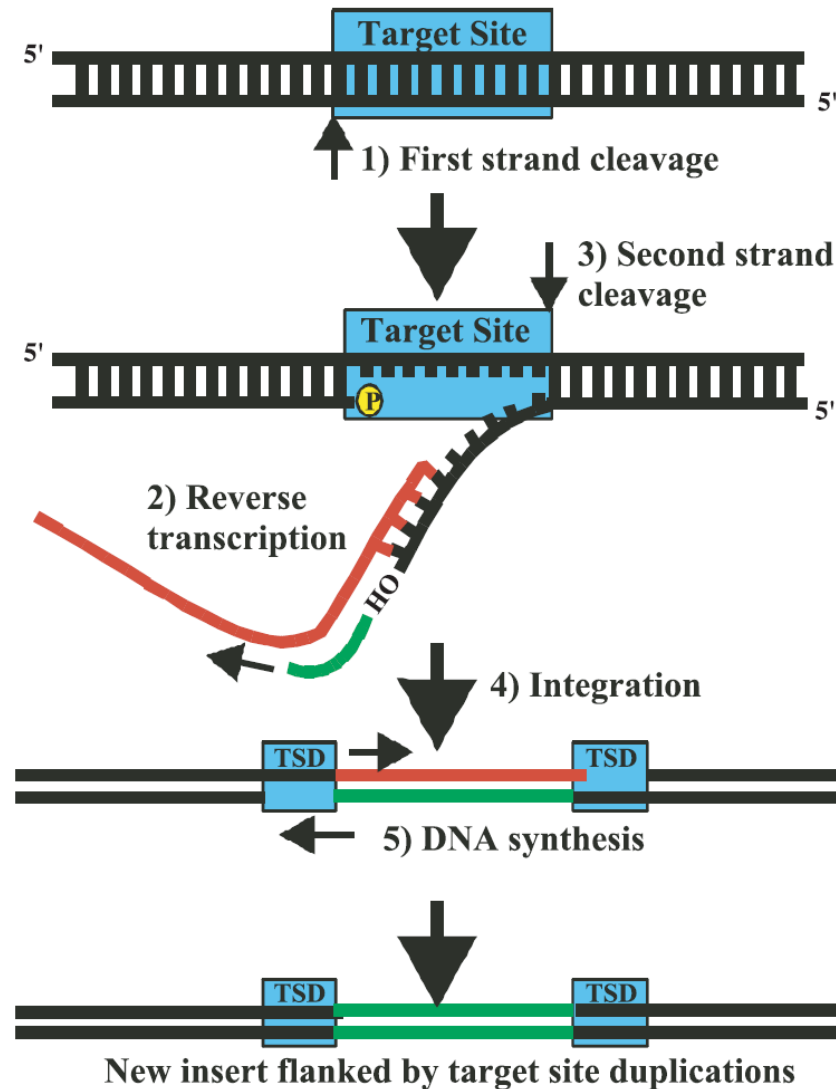
Target-primed reverse transcription (TPRT)

c Non-LTR retrotransposon
Target-site primed reverse transcription



TPRT frequently undergoes premature termination (5' truncation)

Target site duplication (TSD)

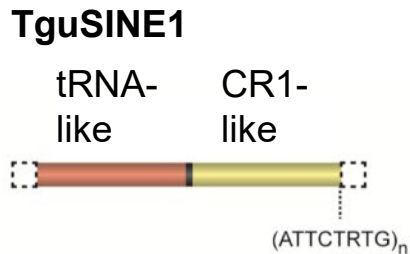
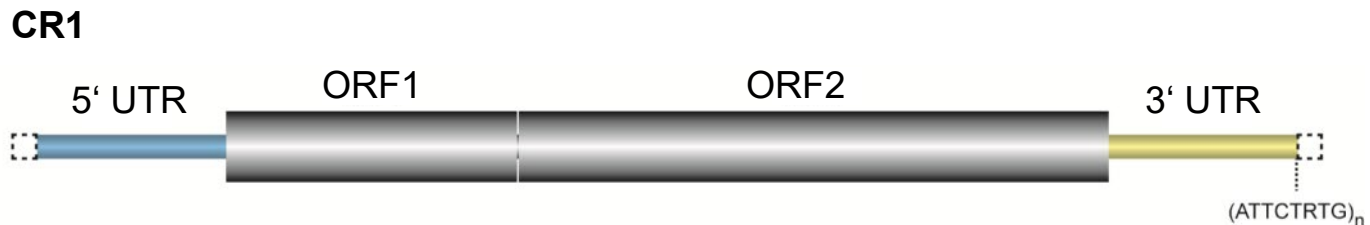


TSDs are a hallmark of nearly all (retro)transposition mechanisms!

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!

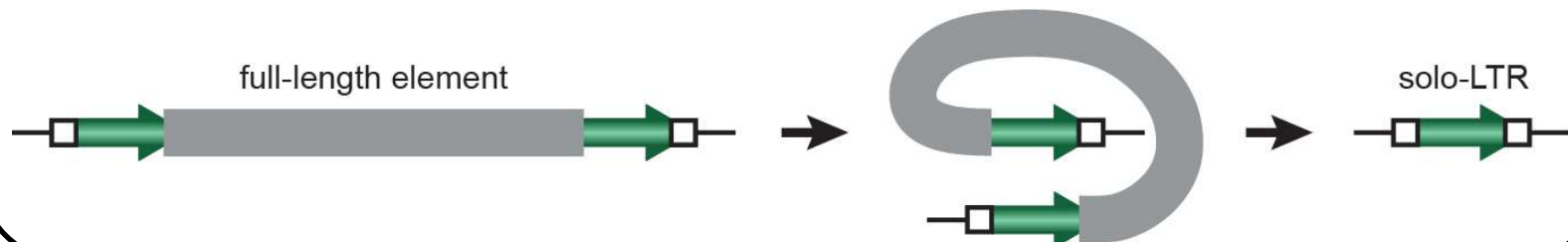
Note: In theory, any small RNA gene (pol III) can become a SINE!

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

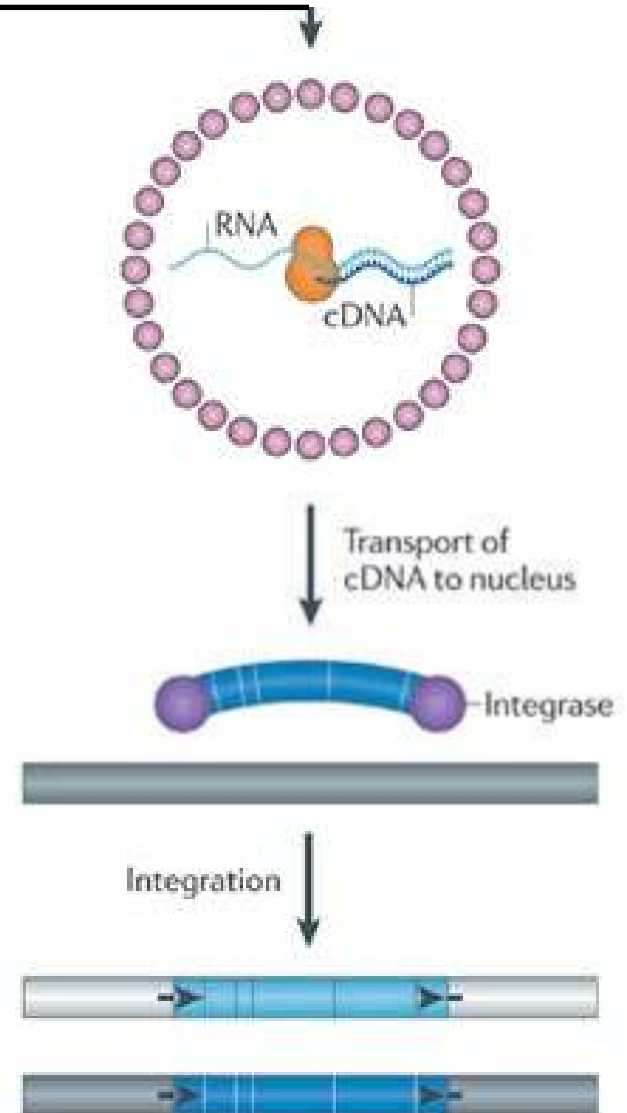
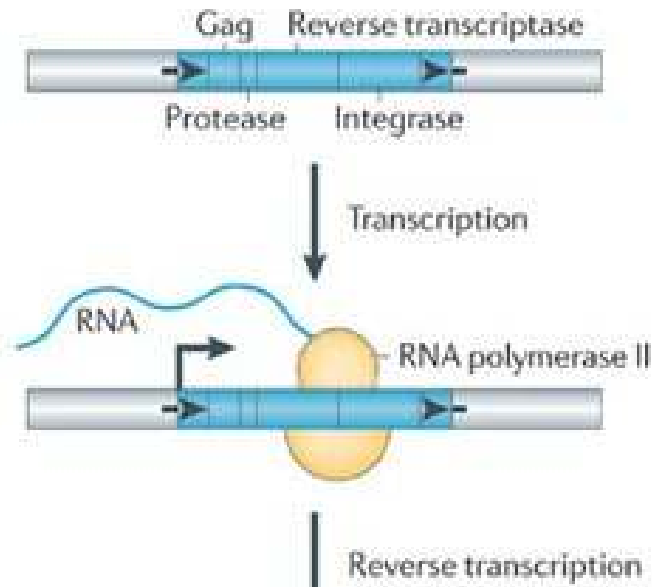


Non-allelic homologous recombination (NAHR):

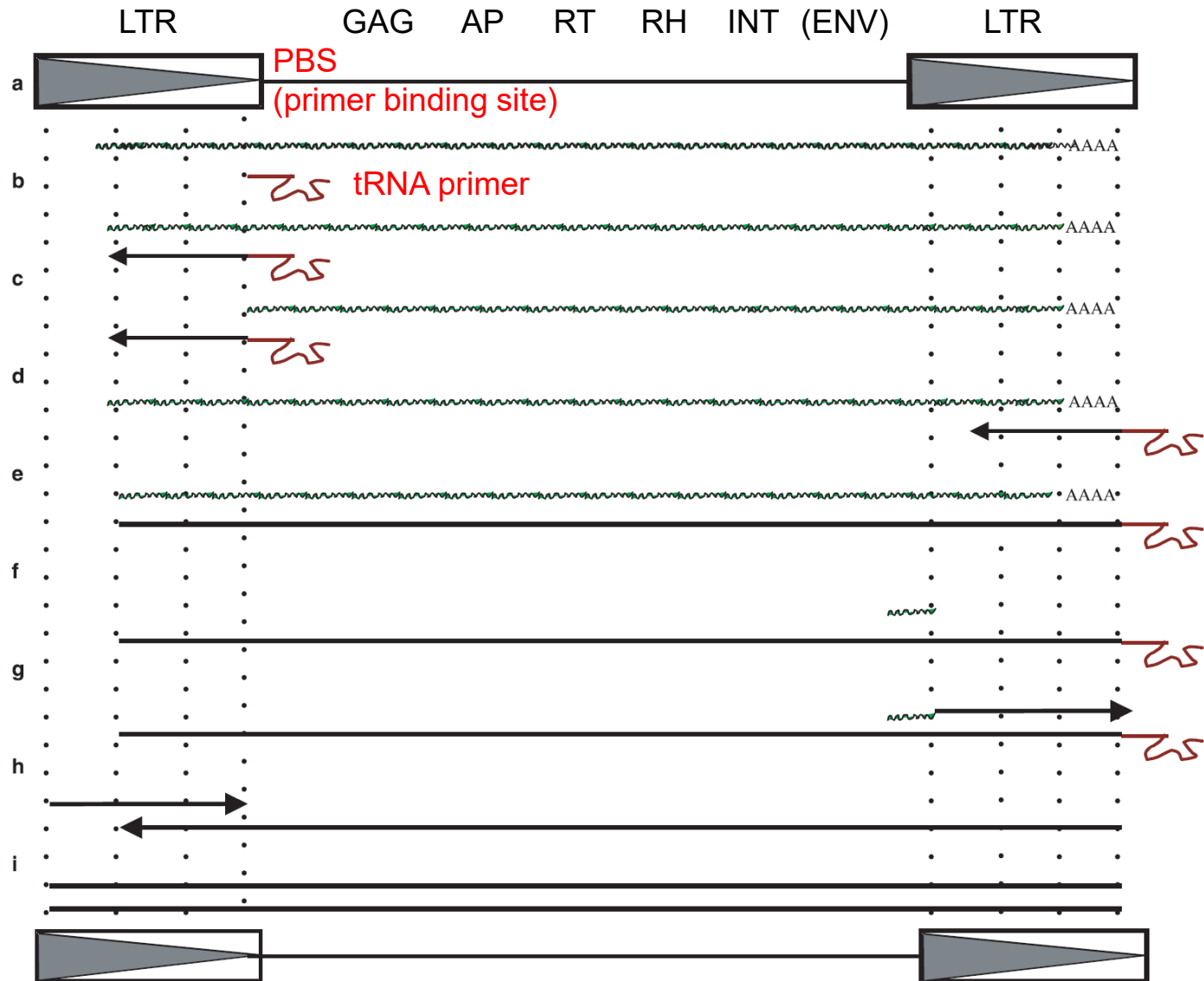


Replicative retrotransposition

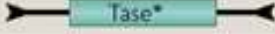
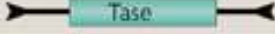
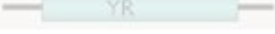
b LTR retrotransposon Replicative retrotransposition



Why LTR retrotransposons have LTRs



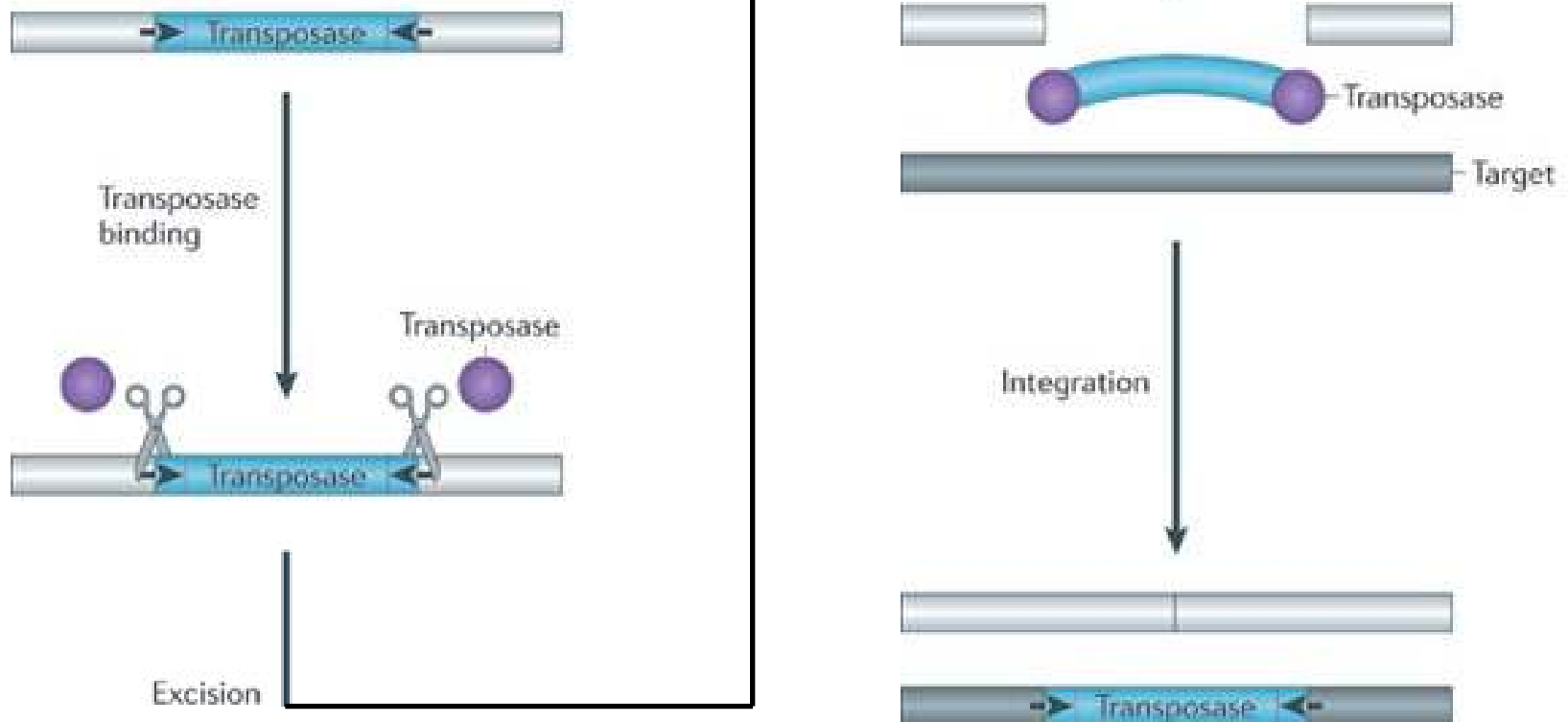
Class II: DNA transposons

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



Cut-and-paste transposition (TIR)

a DNA transposon
'Cut and paste' TE



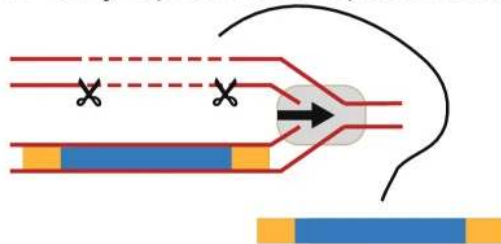
How to increase in copy number?



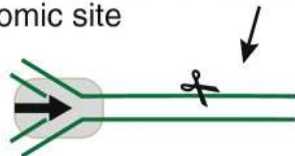
I. DNA replication fork passes transposon



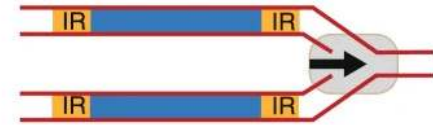
II. Newly replicated transposon is cut out...



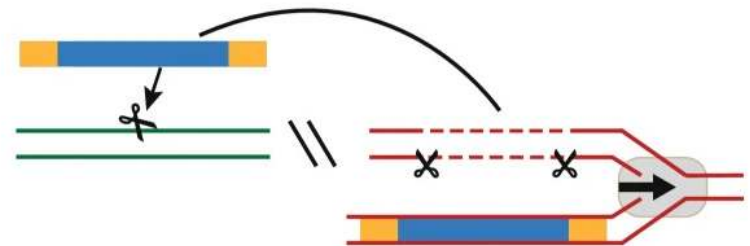
III. ...and inserted into a not-yet replicated genomic site



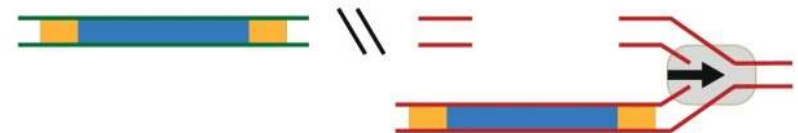
IIII. DNA replication fork passes insertion site



I. Newly replicated transposon is cut out...



II. ...and transposed into a new locus

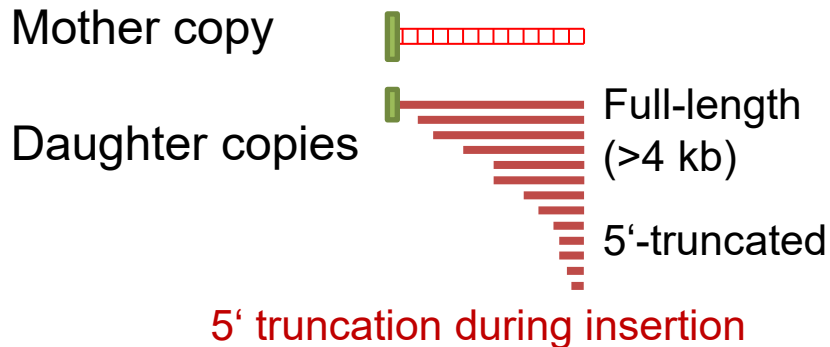


III. Following transposition, the double-strand break is repaired by homology-dependent DNA repair

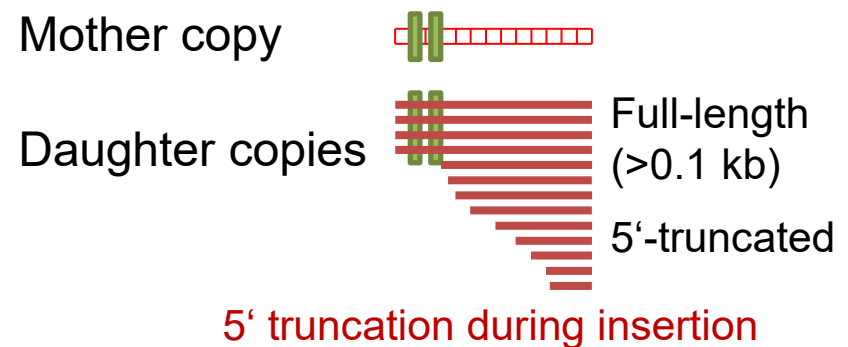


TE ≠ TE

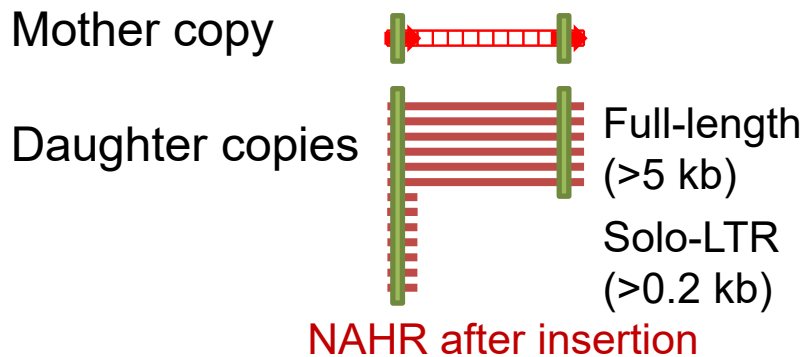
LINE



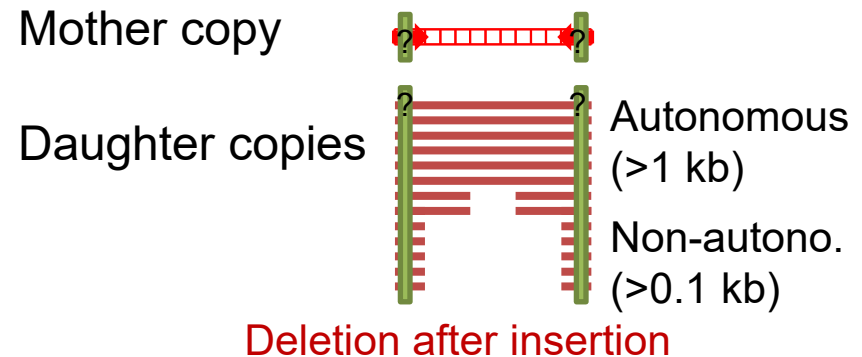
SINE



LTR



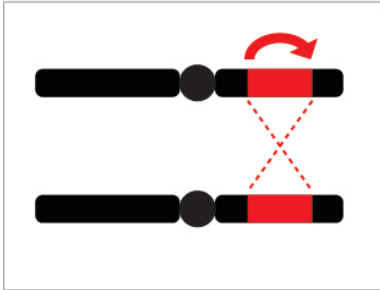
TIR



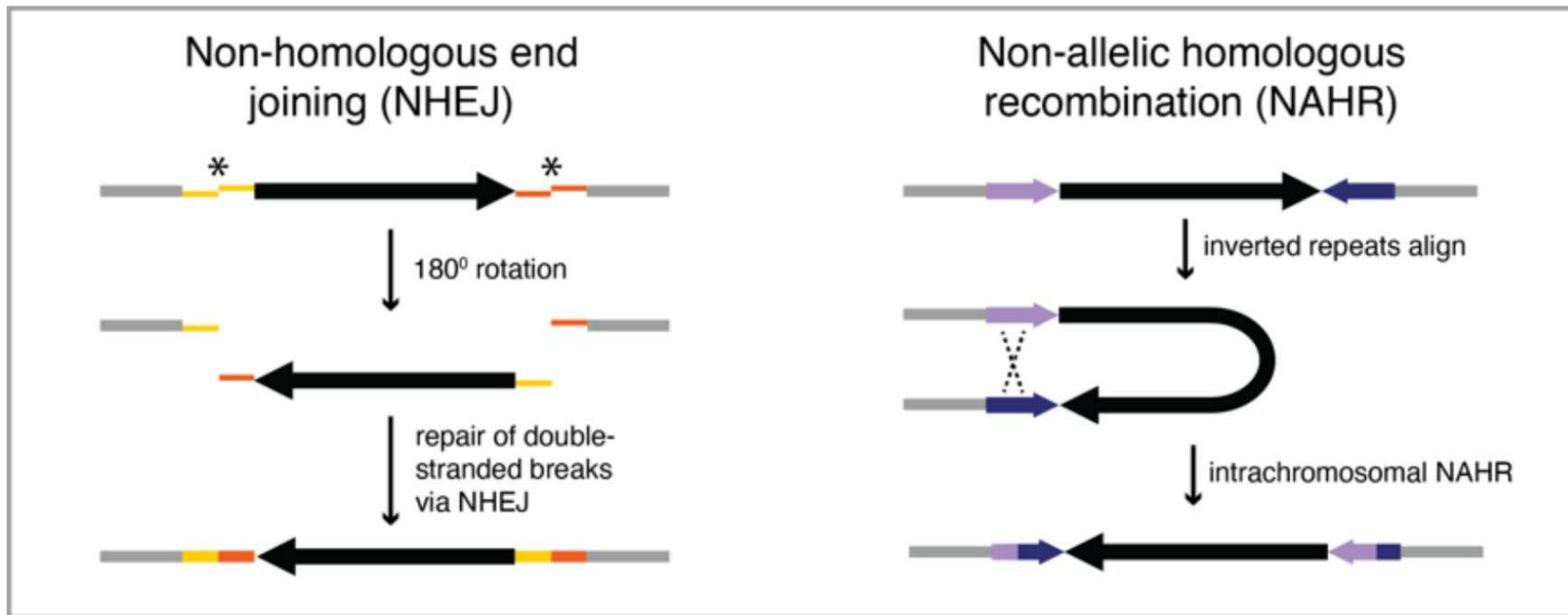
Some TE copies contain regulatory elements, some don't.

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

Inversion formation

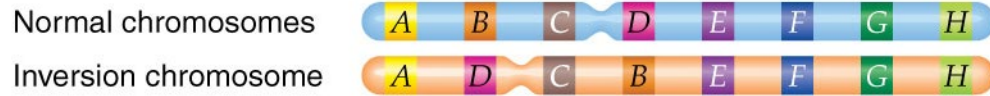


“We found that inversion breakpoints frequently occur in centromeric and telomeric regions and are often flanked by long inverted repeats (0.5-50 kb)”

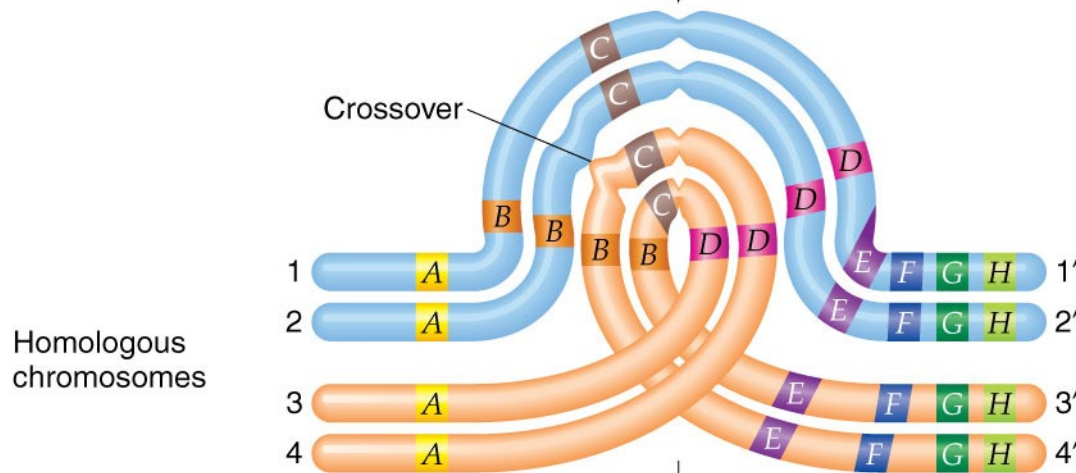


Assembling or mapping inversion breakpoints is difficult!

Inversions reduce recombination



Meiosis to prophase I



Pericentric
inversion
(heterozygous)

Normal product
(all genes present)



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



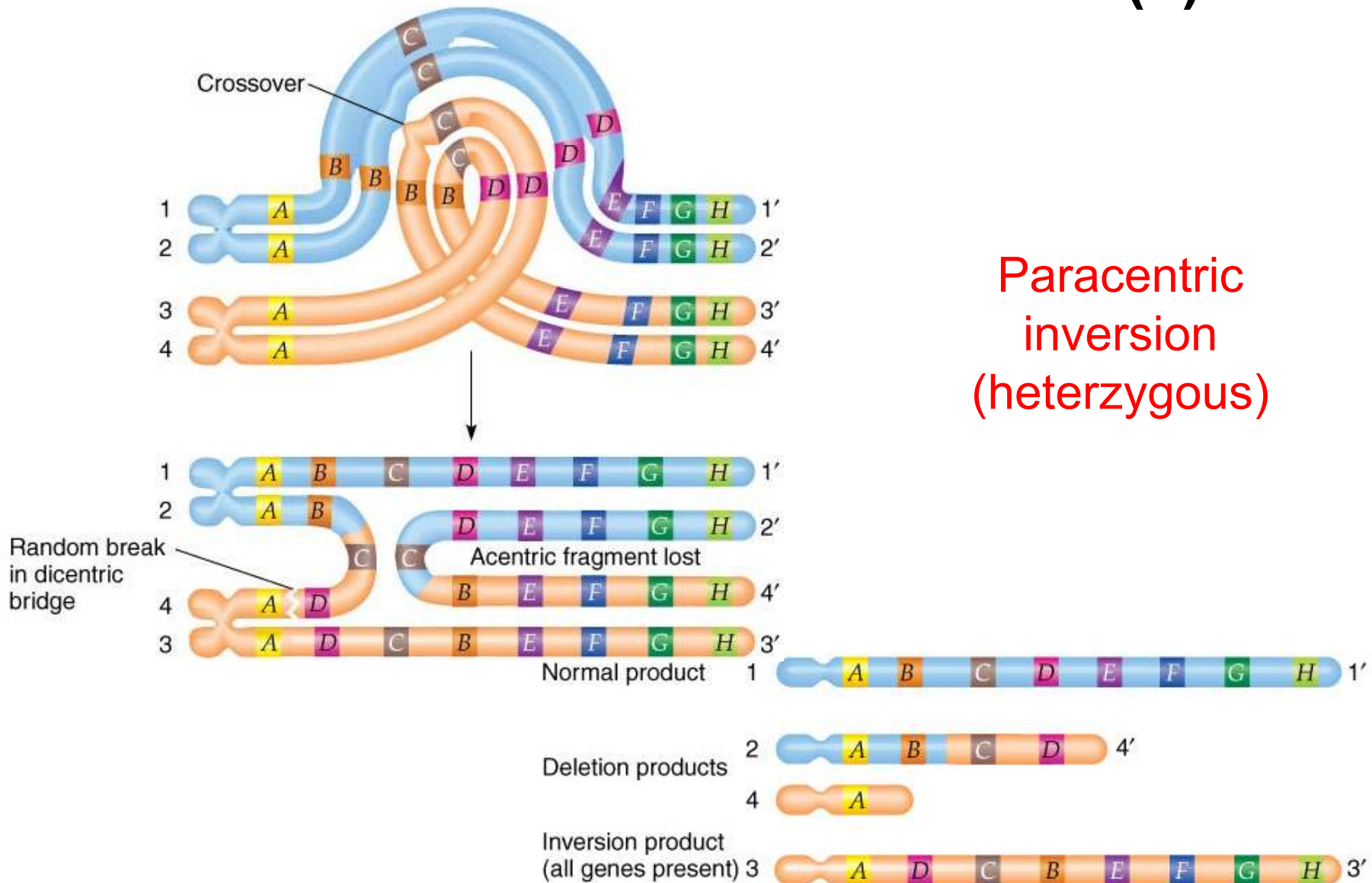
Inversion product
(all genes present)



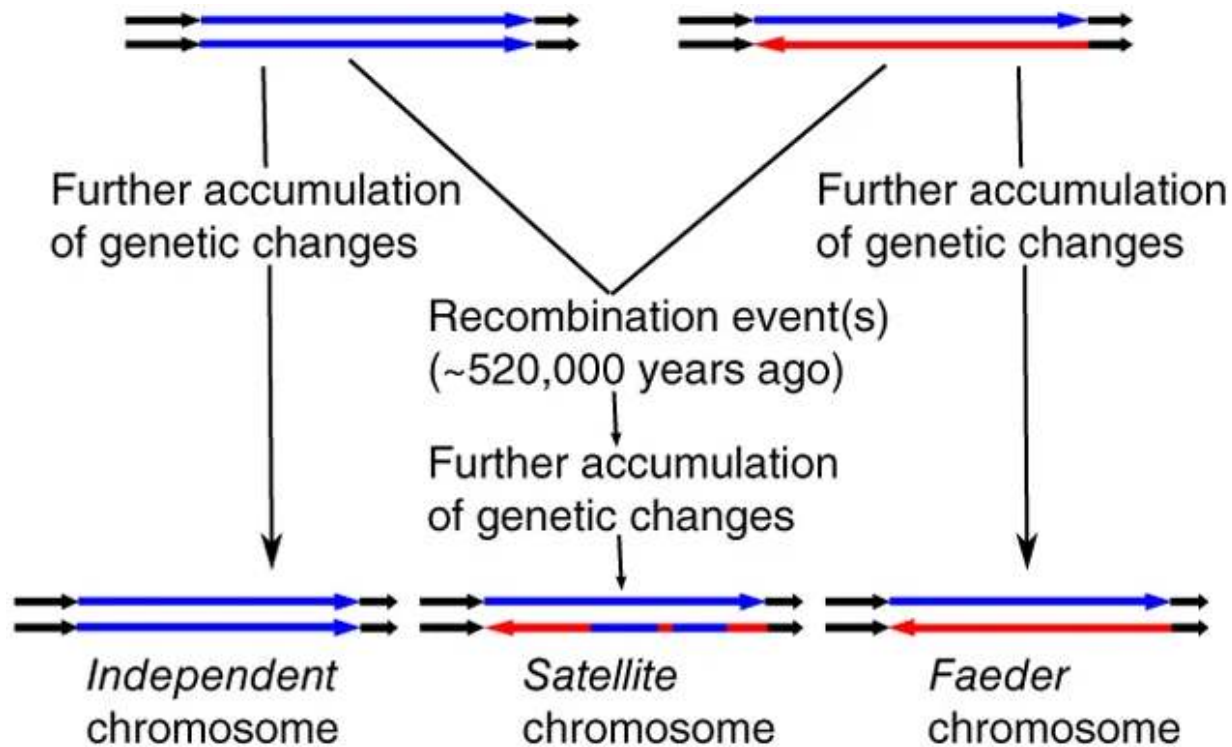
Deletion/duplication
product (*A* deleted;
EFGH duplicated)



Inversions reduce recombination (2)



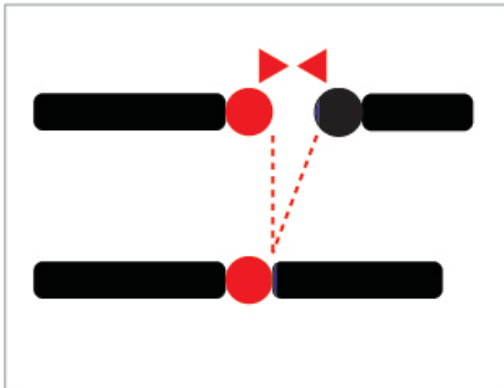
Rare recombination in (large) inversions



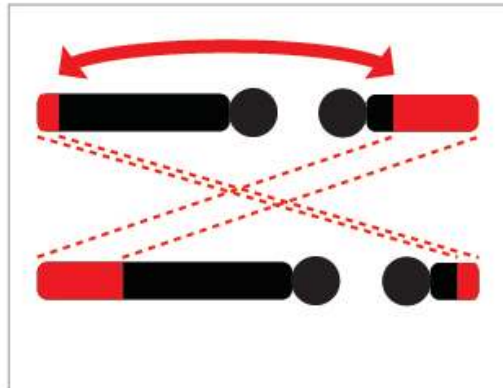
**Double-
crossovers
needed!**

More cases of NAHR

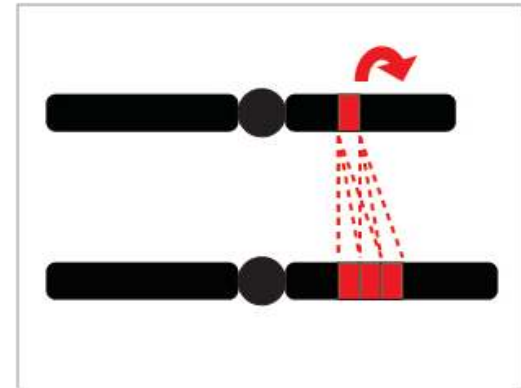
Fusion/fission



Translocation



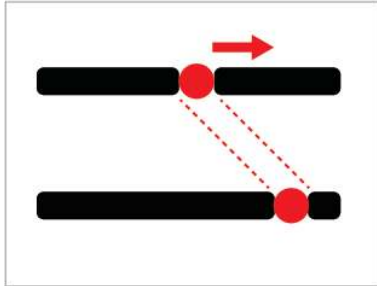
Duplication



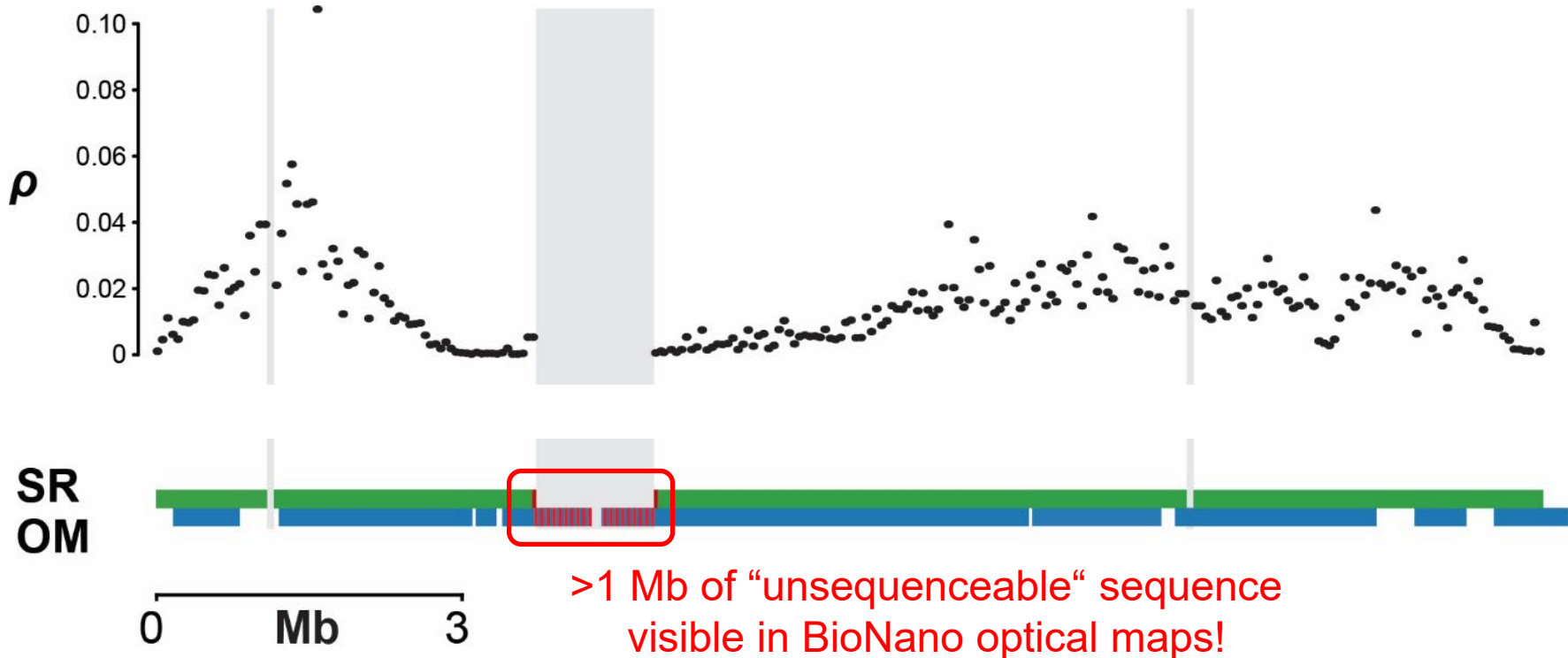
Fusions/fissions/translocations can decrease (new proximity to centromere) or increase (new proximity to telomere) recombination rates

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

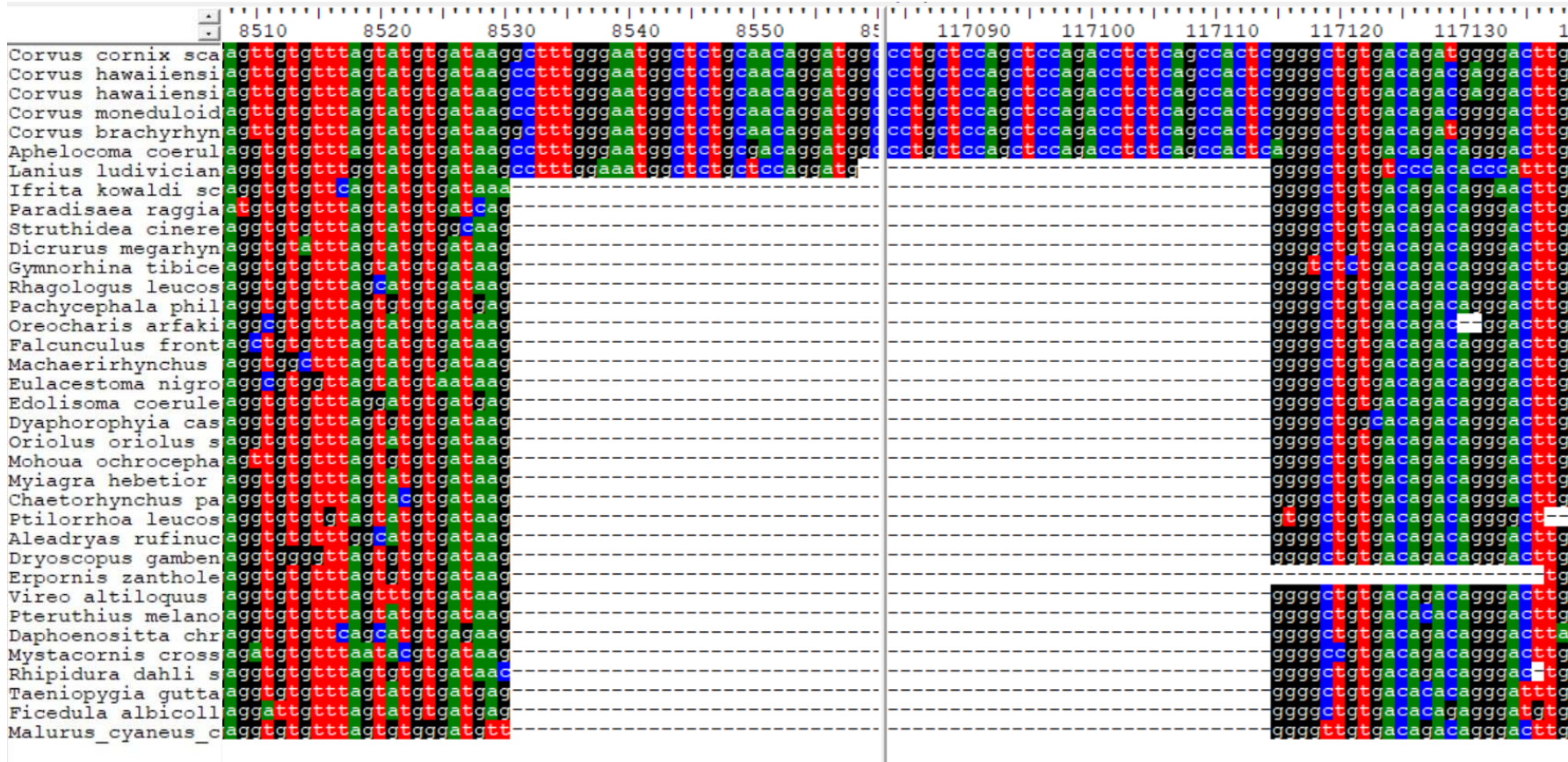
Centromere shifts



Chromosome 18 of hooded/carrion crow

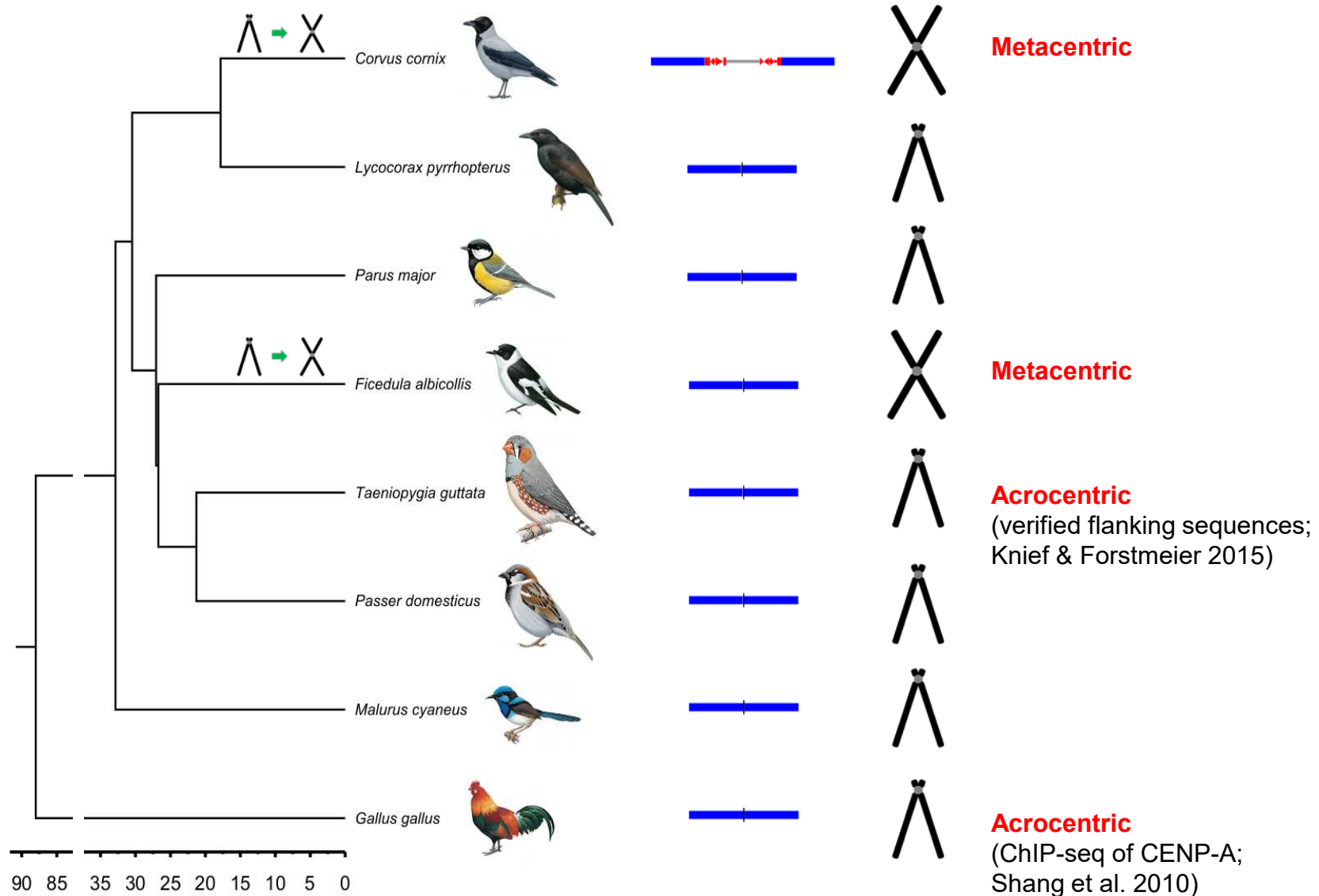


Centromere shifts across songbirds

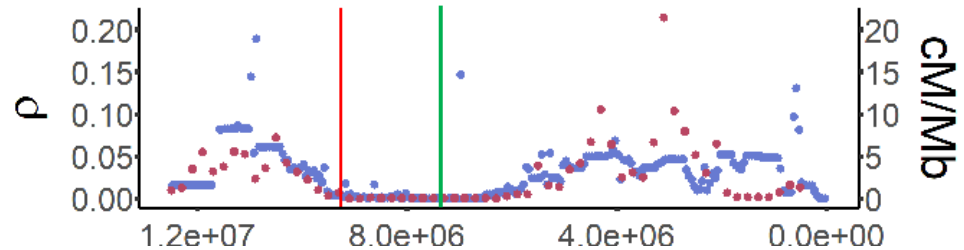
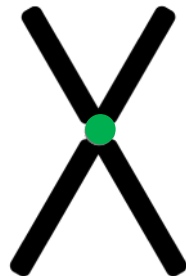
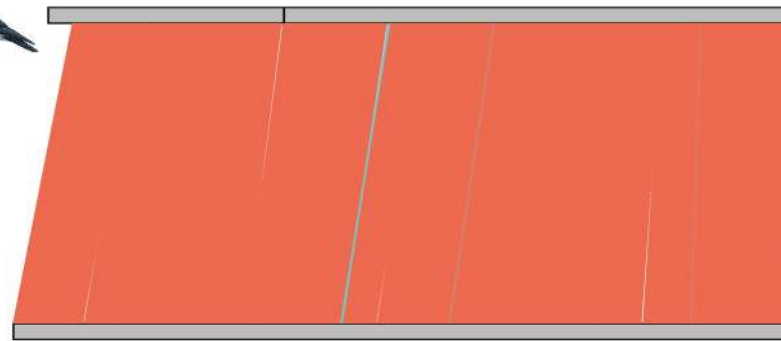
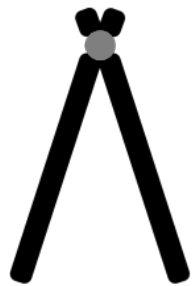
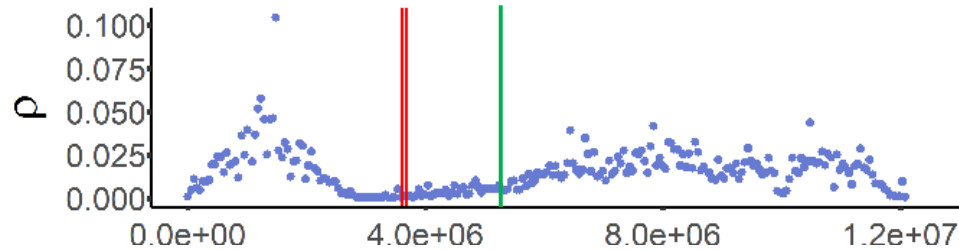
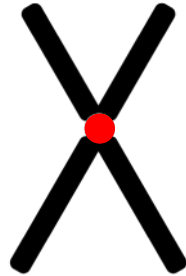


>1 Mb satellite DNA array
inserted in a formerly 5-kb
intergenic region!

Centromere shifts across songbirds



Not so stable chromosomes after all?



Am I stuck in confirmation bias?

Centromere shifts: frequency independent recombination reduction (unlike inversions)

Have centromere shifts been proposed as alternatives to inversions in speciation literature?

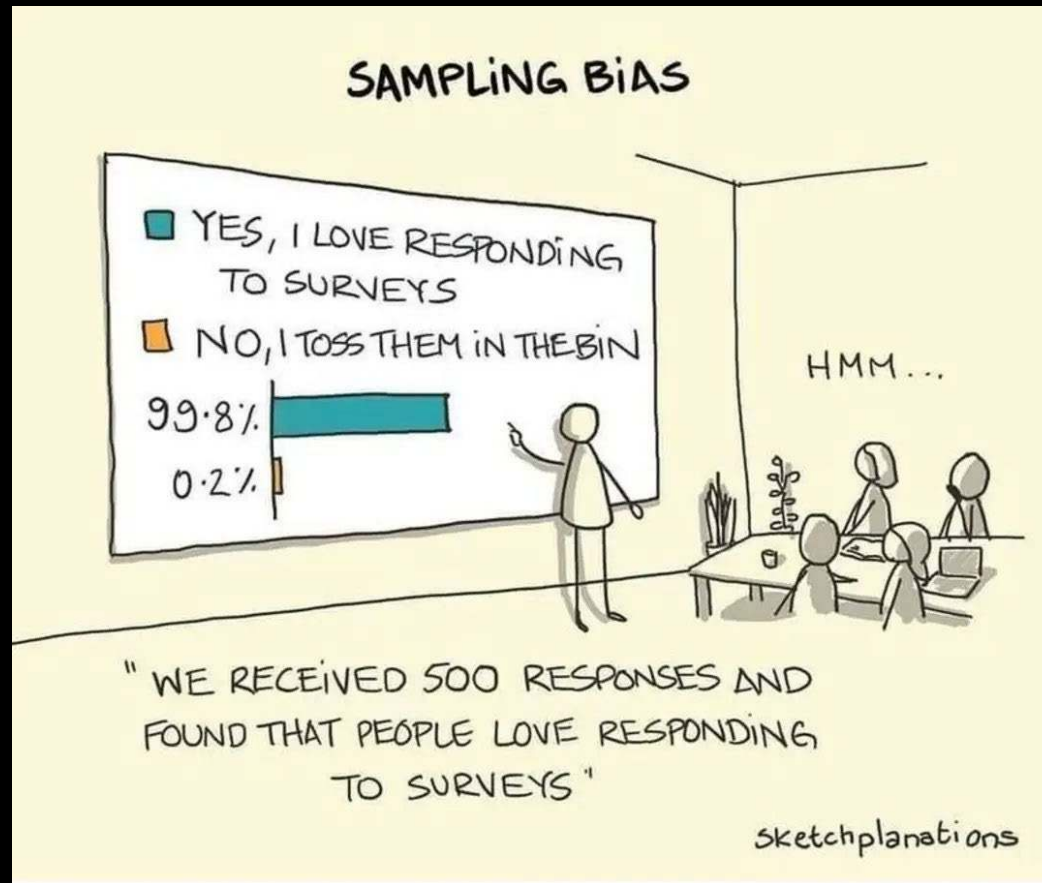
Are some variants that were interpreted as inversions actually centromere shifts?

Can centromere shifts fix more frequently than inversions because of meiotic drive?

Talk to me or email me (alexander.suh@ebc.uu.se) if you have references and/or criticism!

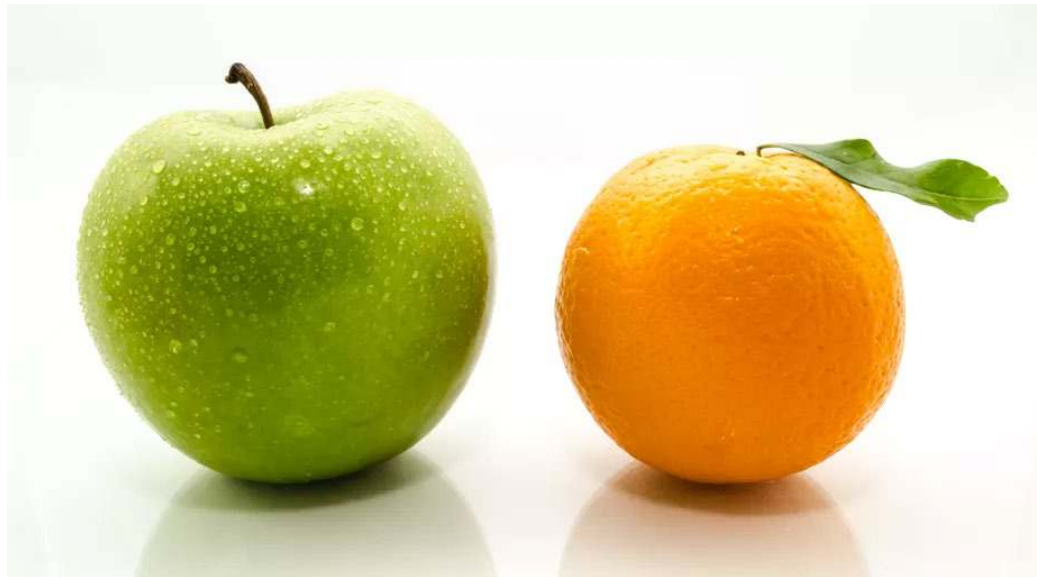
PS: Manuscript in preparation about this model.

Coffee break (20 minutes)



Task: Gather in same 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 3: Hope

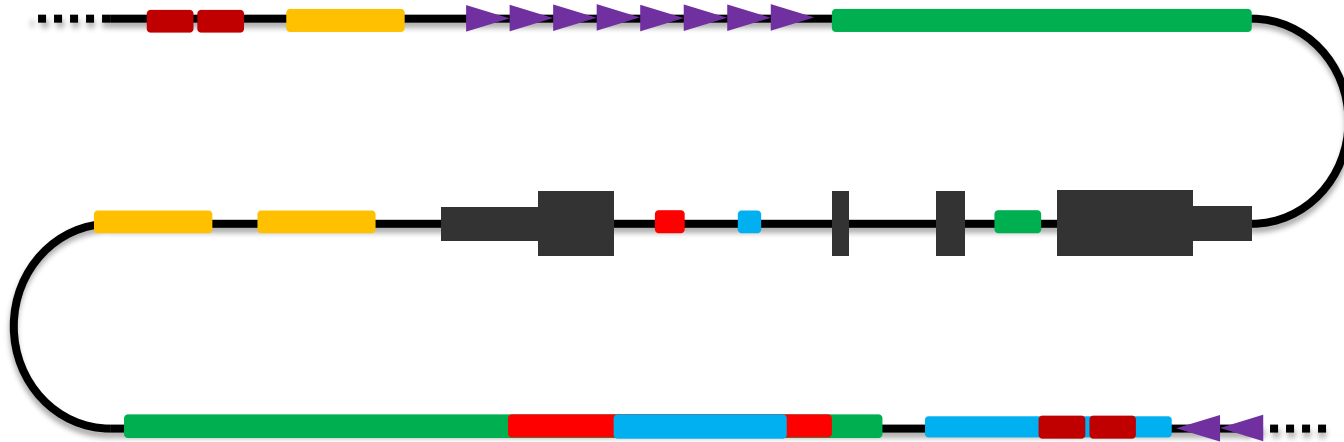


How frustrated are you?

- 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

Repetitive elements in the era of biodiversity genomics: insights from 600+ insect genomes

 John S. Sproul,  Scott Hotaling, Jacqueline Heckenhauer, Ashlyn Powell,  Amanda M. Larracuenta,  Joanna L. Kelley, Steffen U. Pauls,  Paul B. Frandsen

doi: <https://doi.org/10.1101/2022.06.02.494618>

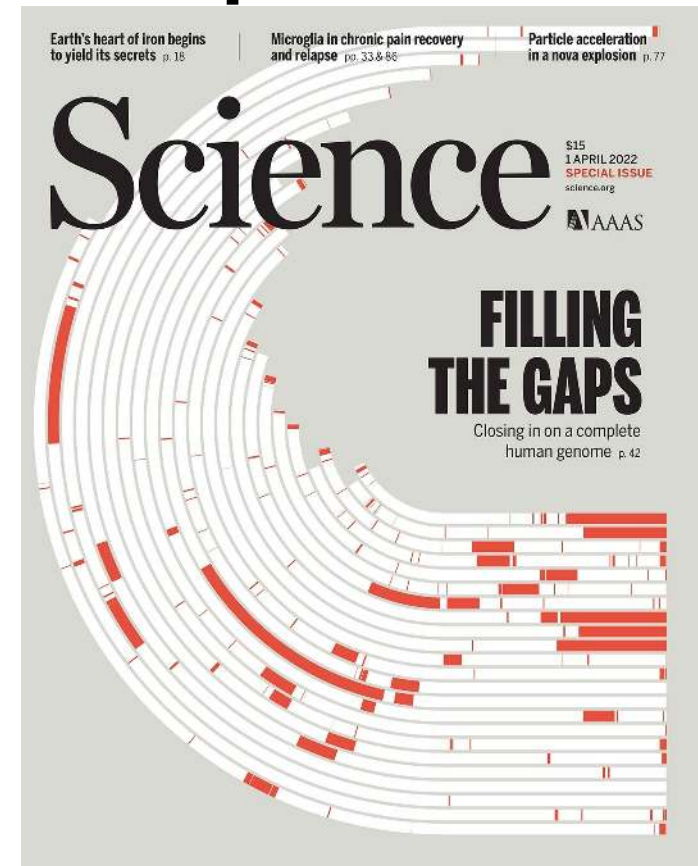
Posted June 03, 2022.

Our

findings suggest this RE-annotation bottleneck, driven largely by uneven taxonomic representation in RE reference databases, is worsening. Although the diversity of available insect genomes has rapidly expanded, the rate of community contributions to RE databases (essential for RE annotation) has not kept pace, preventing high resolution study of REs in most groups. We highlight the tremendous opportunity and need for the field of biodiversity genomics to embrace REs and suggest collective steps for making progress towards this goal.

Complete human genome in April 2022

- 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!

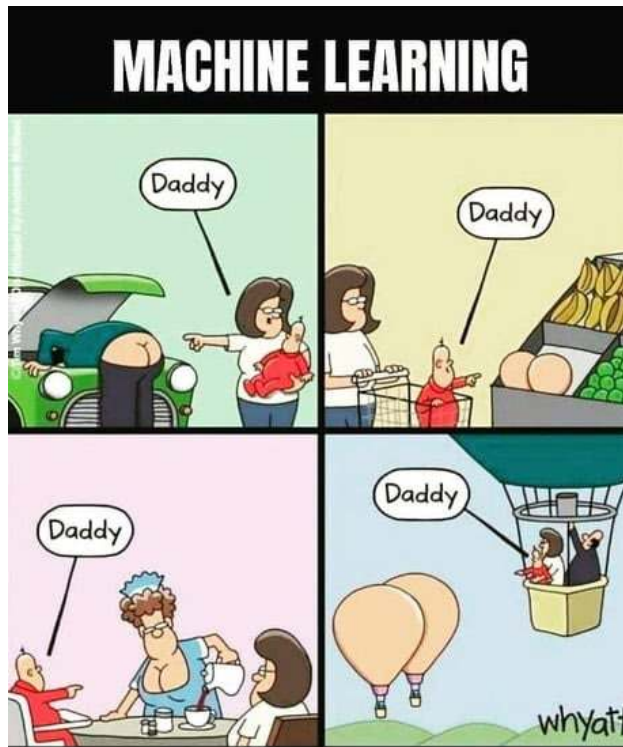
What's next: Machine learning?

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

<https://github.com/LiLabAtVT/DeepTE>, Yan et al. 2020 *Bioinformatics*

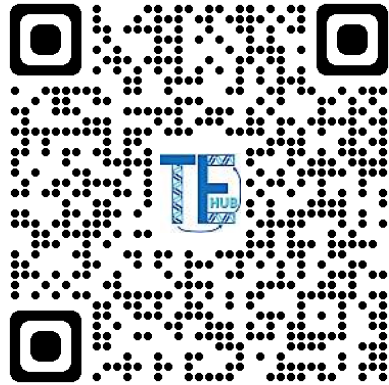
TransposonUltimate: software for transposon classification, annotation and detection

<https://github.com/DerKevinRiehl/TransposonUltimate>, Riehl et al. 2022 *Nucl. Acids Res.*



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

More community initiatives needed



TE Hub website



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

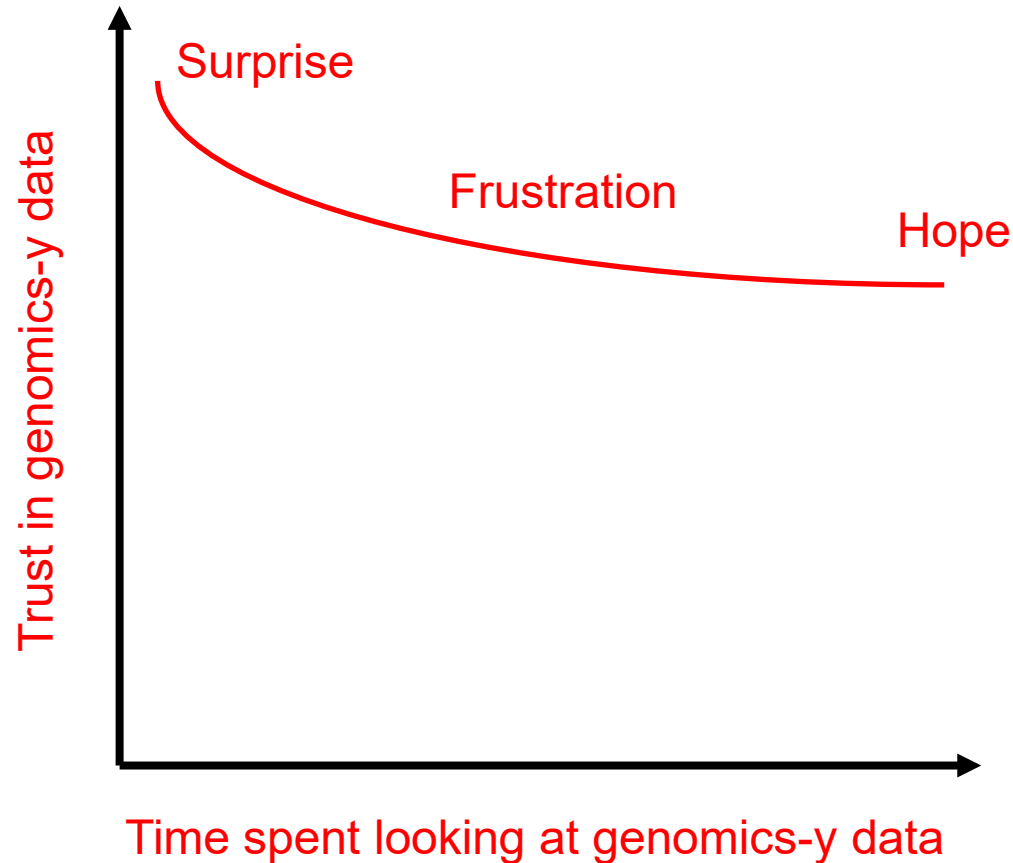
Let's keep in touch in the WPSG 2022
Slack channel [#structural_variation!](#)
(<https://wpsg2022participants.slack.com>)



Alexander Suh 4:38 PM
Please click 👍 for the SV types you are interested in:

- Inversions
👍 1 🗨️
- Transposable elements
👍 1 🗨️
- Translocations
👍 1 🗨️
- Insertions/deletions
👍 1 🗨️
- Fusions/fissions
👍 1 🗨️
- Centromere shifts
👍 1 🗨️
- Duplications
👍 1 🗨️

Conclusion: Genomics is no silver bullet



Genomics + cytogenetics = cytogenomics

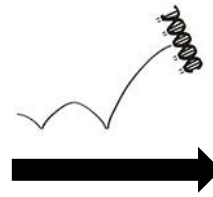
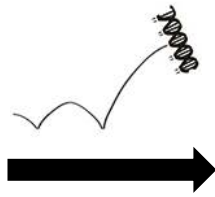


What to take with a grain of salt

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, ...)?



The transposition goes on!



LIB Leibniz Institute for the Analysis
of Biodiversity Change



From 01 April 2023:
Professor at University of Bonn and Head of the
Centre for Molecular Biodiversity Research
in Bonn/Hamburg, Germany

Talk to me or email me (alexander.suh@ebc.uu.se) if
interested in developing cytogenomics and/or genome
annotation/curation for scaling across animals

**PhD/postdoc/researcher
positions available!**



Questions?

??!



Thank you, Český Krumlov!

