CS273A

Lecture 11: Neutral evolution: repetitive elements

MW 1:30-2:50pm in Clark **S361*** (behind Peet's)

Profs: Serafim Batzoglou & Gill Bejerano

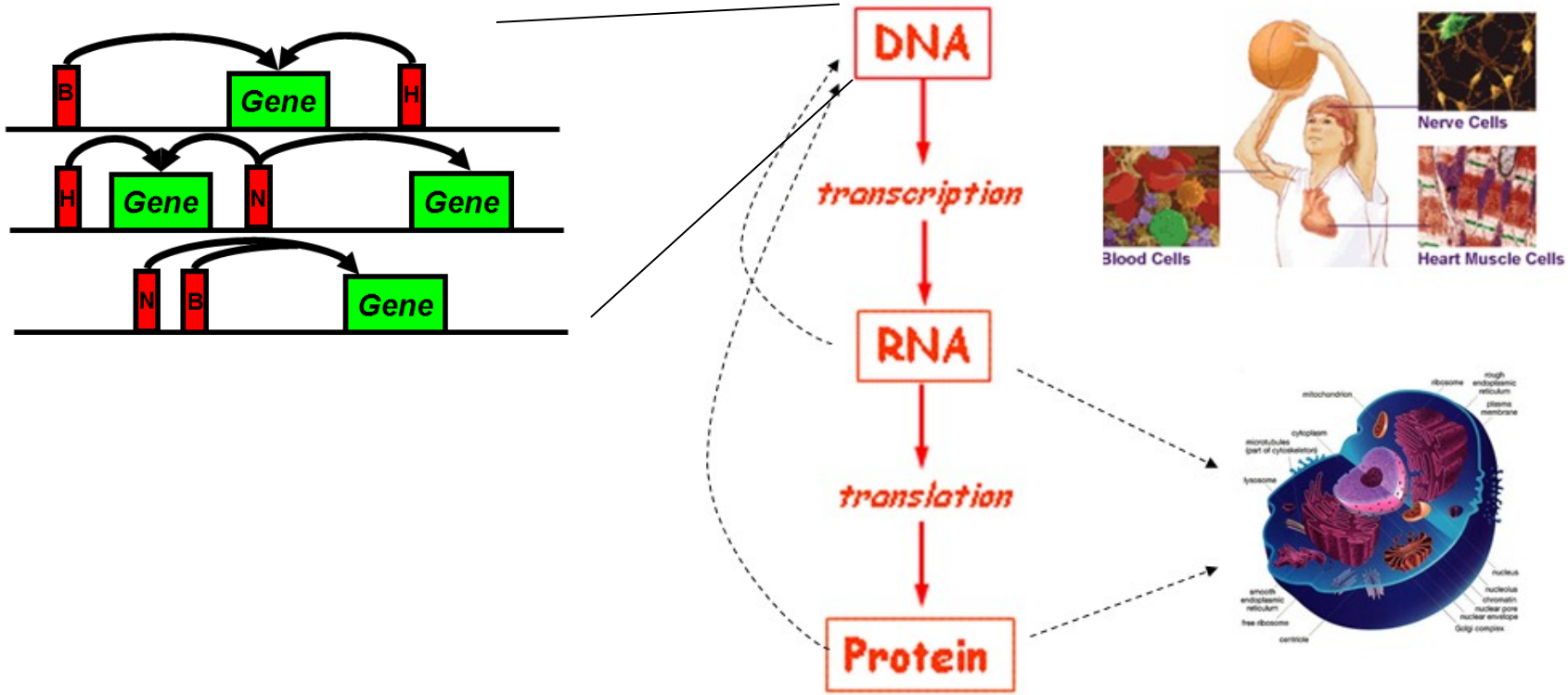
CAs: Karthik Jagadeesh & Johannes Birgmeier

* Mostly: track on website/piazza

Announcements

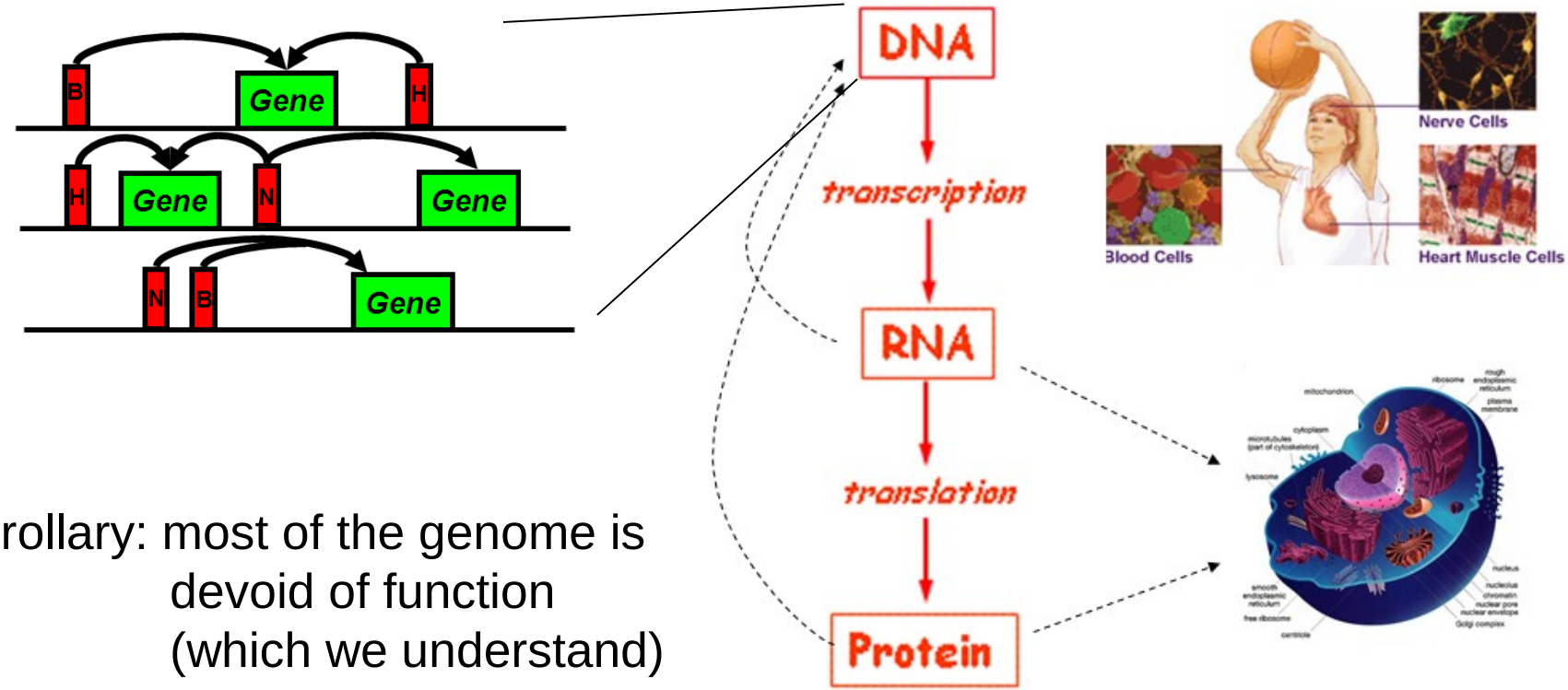
- PS1 is in. PS2 is out…

The Functional Genome



Type	# in genome
genes	20,000
ncRNA	20,000
cis elements	1,000,000

The Functional Genome



Corollary: most of the genome is devoid of function (which we understand)

Type	# in genome	% of genome
genes	20,000	2-3%
ncRNA	20,000	2%
cis elements	1,000,000	10-15%

Genome Evolution

“Nothing in Biology Makes Sense
Except in the Light of Evolution”

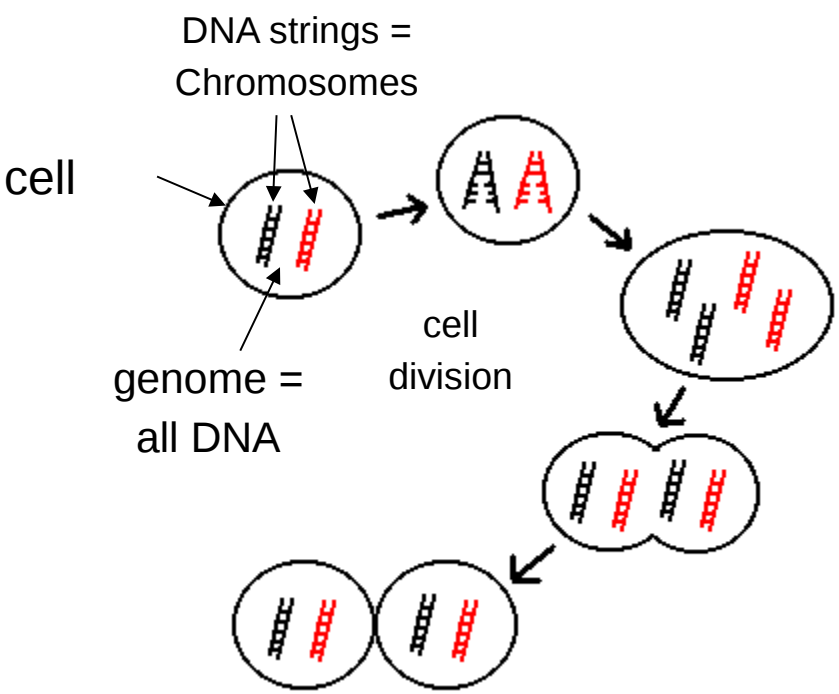
Theodosius Dobzhansky

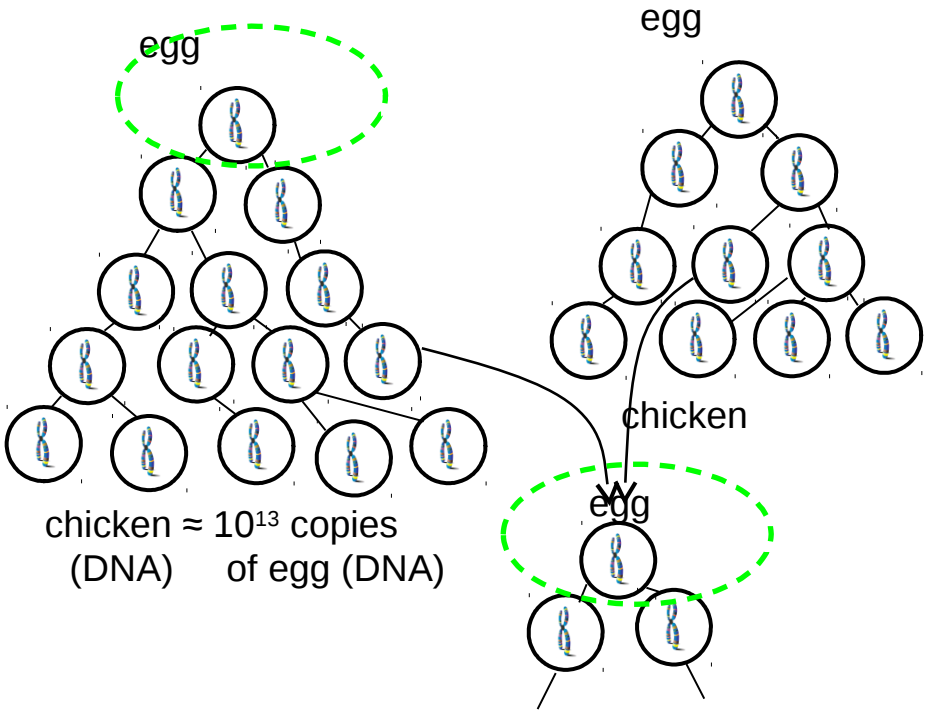
One Cell, One Genome, One Replication

Every cell holds a copy of all its DNA = its genome.

The human body is made of ~10^{13} cells.

All originate from a *single* cell through *repeated* cell divisions.





Every Genome is Different

DNA Replication is imperfect – between individuals of the same species, even between the cells of an individual.

junk *functional*

...ACGTACGACTGACTAGCATCGACTACGA...

chicken

egg

TT CAT

...ACGTACGACTGACTAGCATCGACTGACGA...

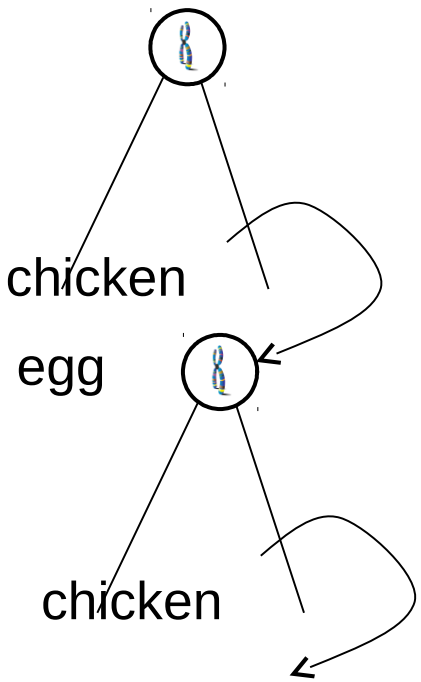
"anything goes"

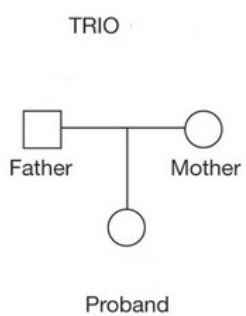
many changes are **not** tolerated

chicken

This has bad implications – disease,
and good implications – adaptation.

Human Mutation Rate

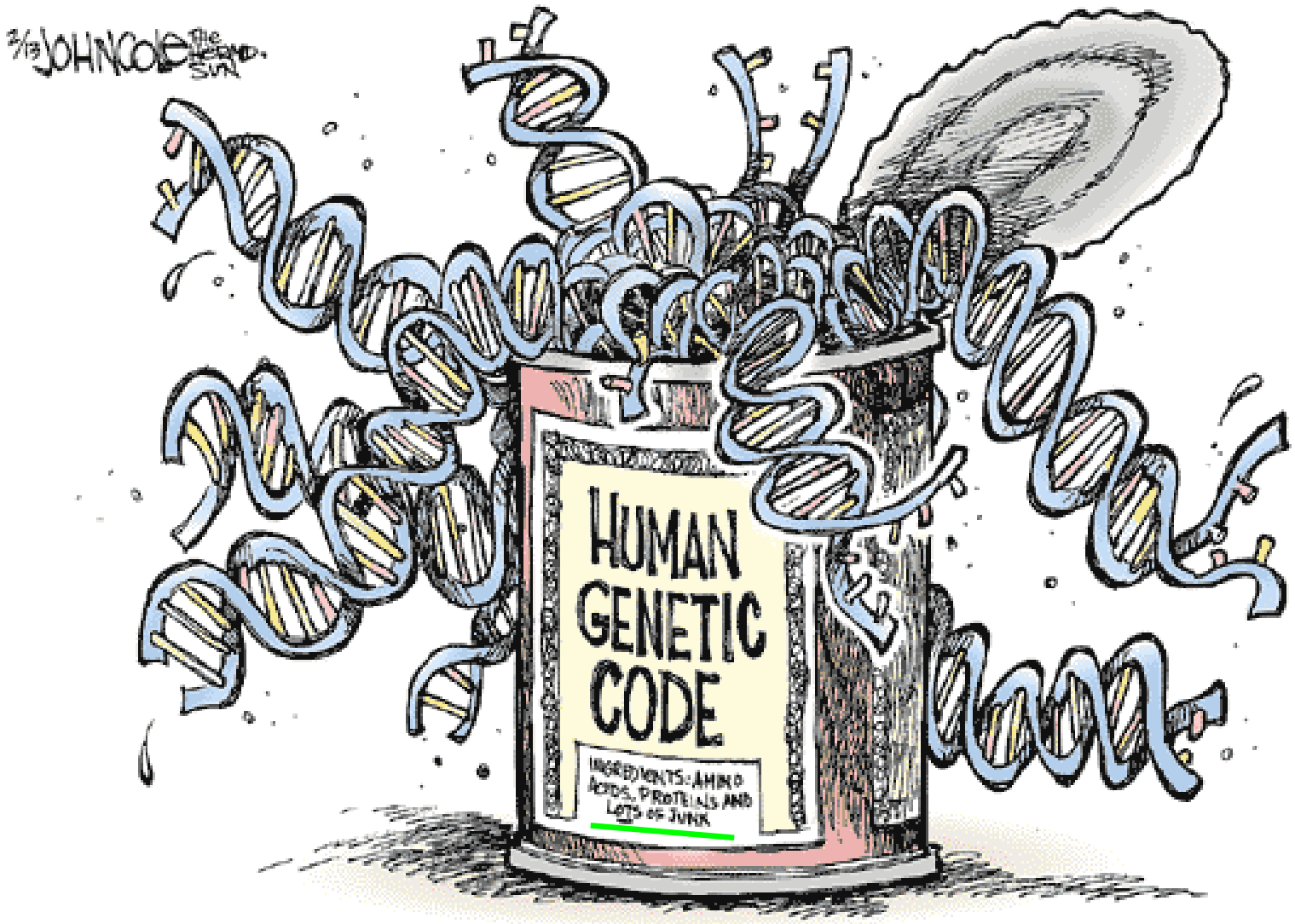




- Recent sequencing analysis suggests ~40 new mutations in a child that were not present in either parent.
- Mutations range from the smallest possible (single base pair change) to the largest – whole genome duplication (to be discussed).
- Selection does not tolerate all of these mutation, but it sure does tolerate some.

Genome Content

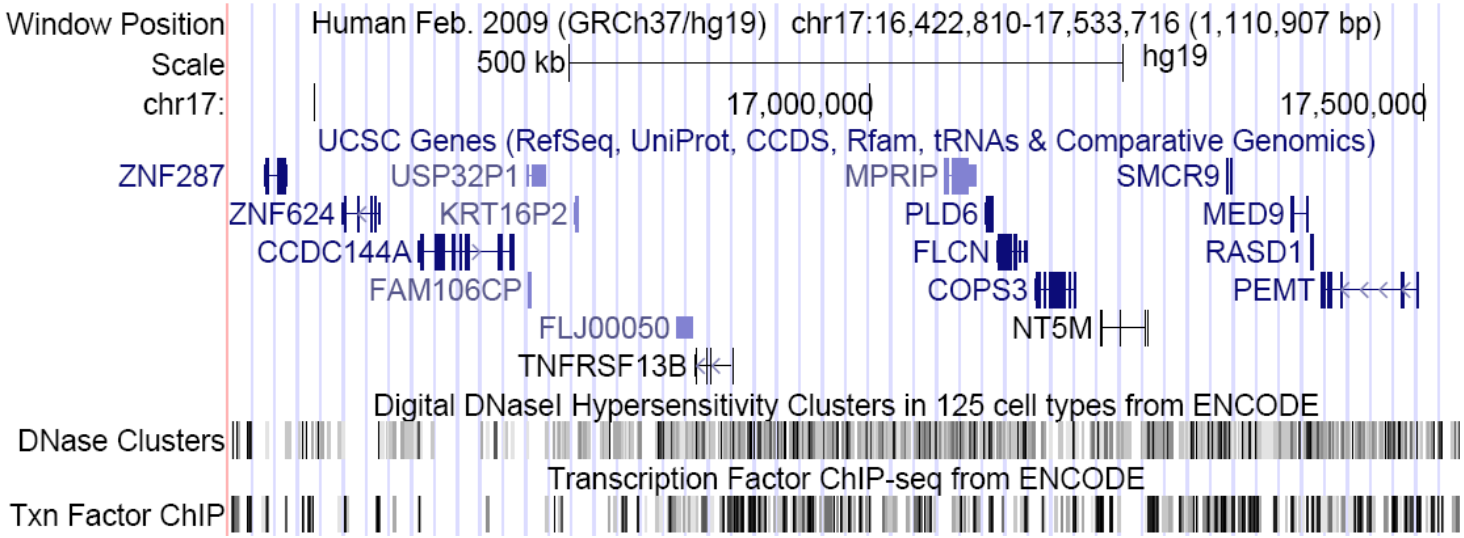
Why this cartoon?



Genome Composition

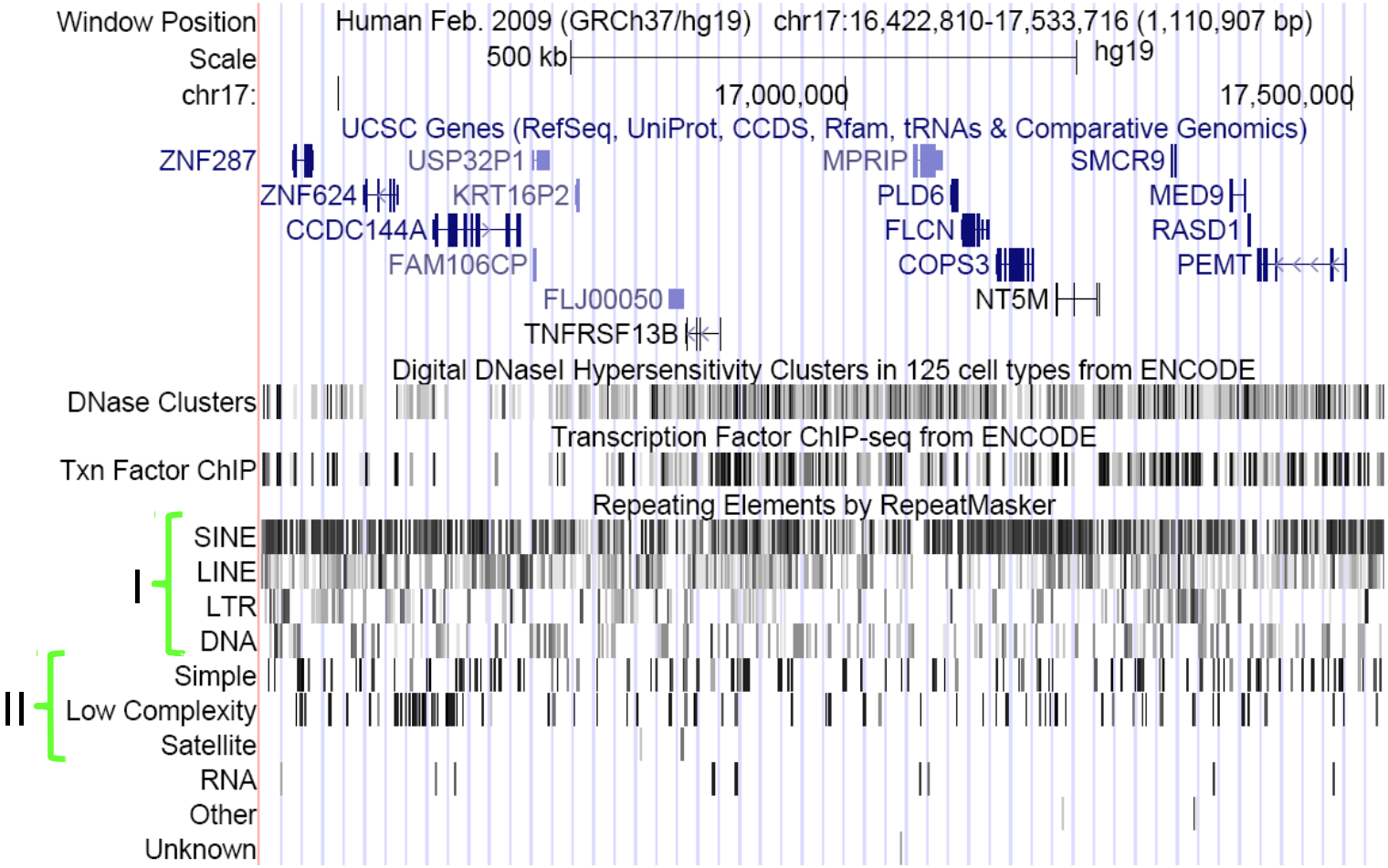
The functional genome takes about 20% of the genome.

The remaining 80% is far from homogeneous…



Sequences that repeat many times in the genome

- Take up cumulatively a whooping *half* of the genome
- Come in two major, very different, flavors



I. Interspersed Repeats / TEs

Transposable elements are pieces of genetic information that somehow manage to multiply themselves and move around in the genome.

[Adapted from Lunter]

I. Interspersed Repeats / TEs

Transposable elements are pieces of genetic information that somehow manage to multiply themselves and move around in the genome.

History: First suspected in 1940 from work by Barbare McClintock on genomic instability in maize. Existence of transposable elements was proven experimentally in 1970s. She received Nobel prize in 1983.

[Adapted from Lunter]

I. Interspersed Repeats / TEs

Transposable elements are pieces of genetic information that somehow manage to multiply themselves and move around in the genome.

History: First suspected in 1940 from work by Barbare McClintock on genomic instability in maize. Existence of transposable elements was proven experimentally in 1970s. She received Nobel prize in 1983.

Four classes of transposable elements live in our genome:

- DNA transposons
- LINEs (long interspersed nuclear elements), retroposons
- SINEs (short interspersed nuclear elements), non-autonomous retroposons
- Retroviruses and retrovirus-like LTR (long terminal repeat) retrotransposons

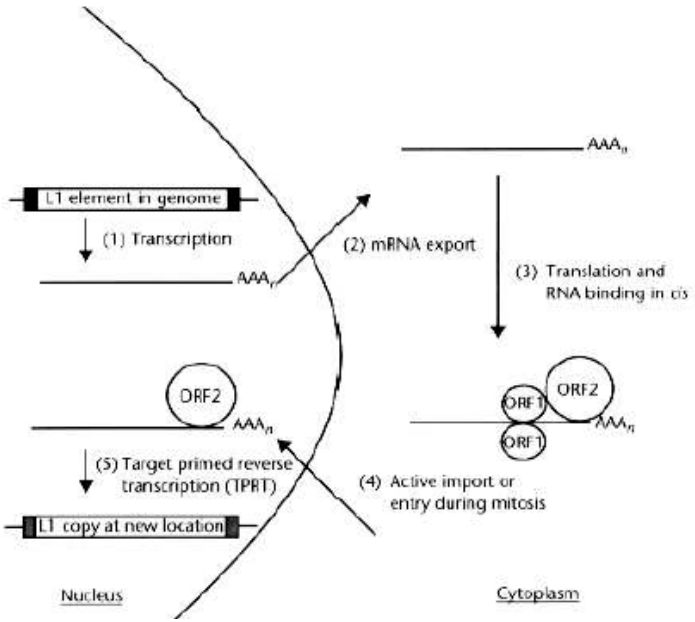
[Adapted from Lunter]

LINE & SINE Elements

LINEs have

- their own (pol II) promotors,
- two ORFs coding for protein,
- 3' binding site for ORF2 protein,
- poly-A tail

Act in cis, i.e. proteins coded by LINE bind to own mRNA.

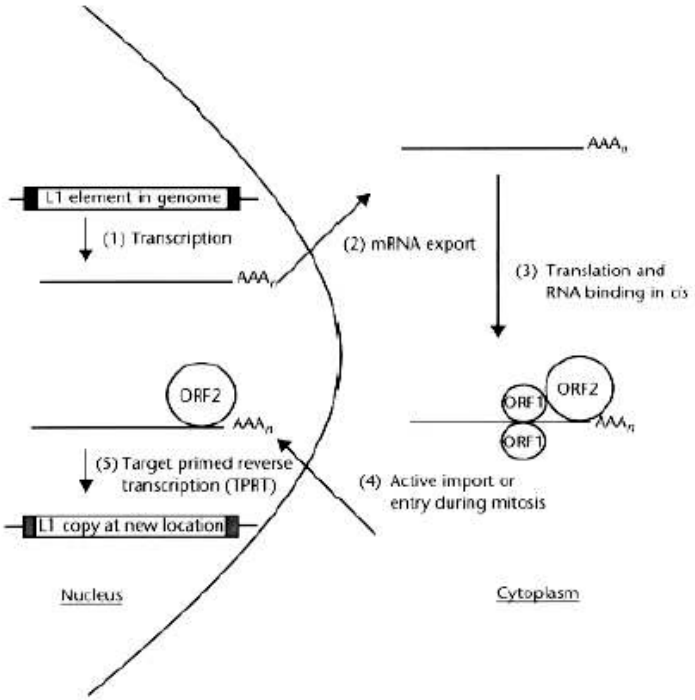


LINE & SINE Elements

LINEs have

- their own (pol II) promotors,
- two ORFs coding for protein,
- 3' binding site for ORF2 protein,
- poly-A tail

Act in cis, i.e. proteins coded by LINE bind to own mRNA.



From: EHG R 53

After translation and binding to own mRNA, the LINE element:

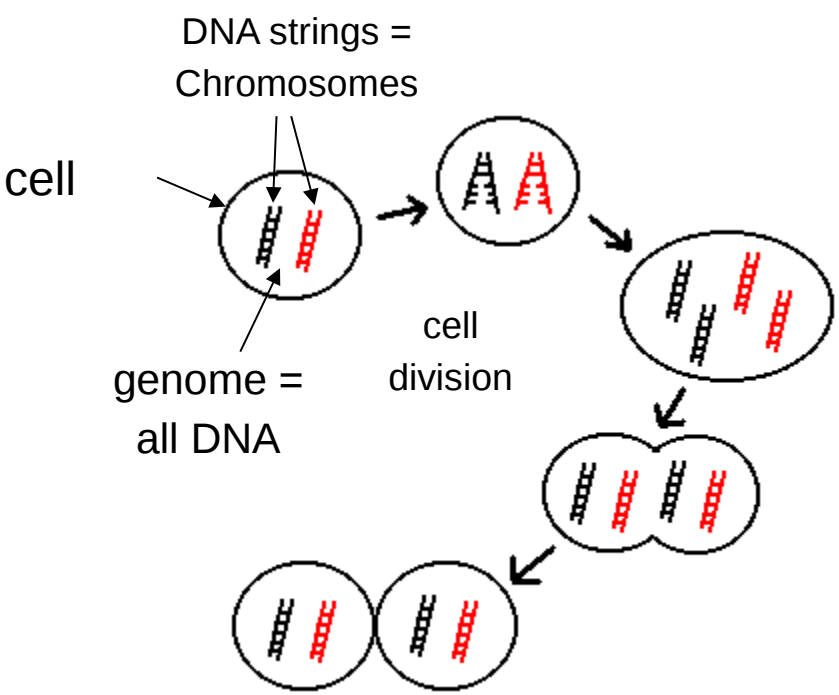
- Gets transported back into nucleus;
- Cleaves host DNA, preferentially at `TT|AAAA`;
- Transcribes a DNA copy from RNA directly into genome. New copy is flanked by a 7–20 bp target site duplication from cleaved-and-repaired host DNA.

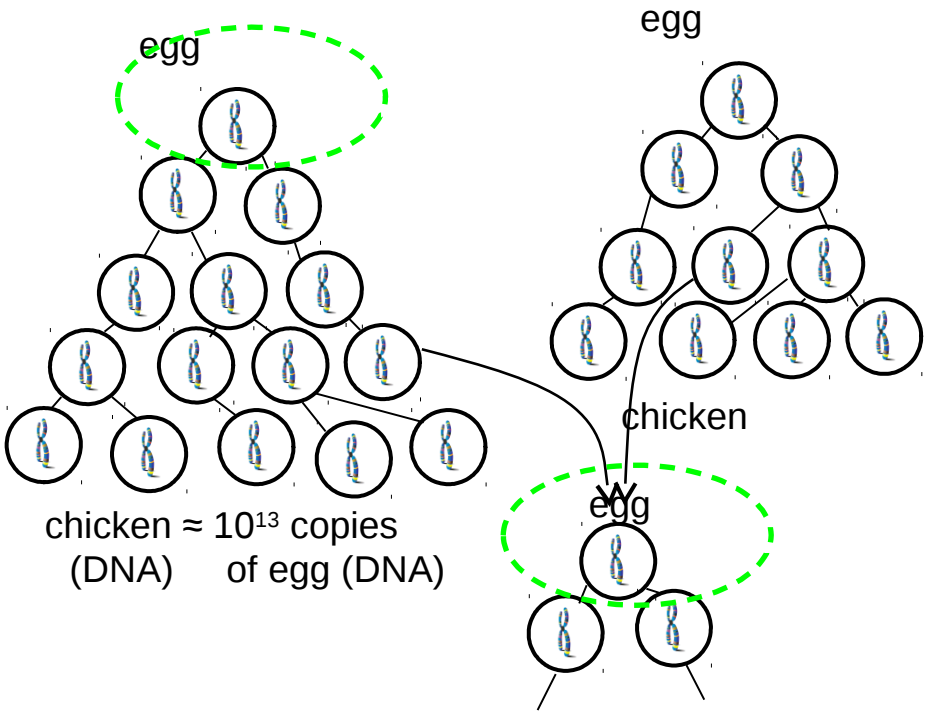
LINEs	Autonomous	ORF1 ORF2 *(pol)* AAA	1–5 kb	20,000–40,000	21%
SINEs	Nonautonomous	A B AAA	100–300 bp	1,500,000	13%

Genomic Transmission

For repeat copies to accumulate through human generations they must make it into the <u>germline</u> cells (eggs & sperms).

Equally true for any genomic mutation.





Classes of Interspersed Repeats

Classes of interspersed repeat in the human genome

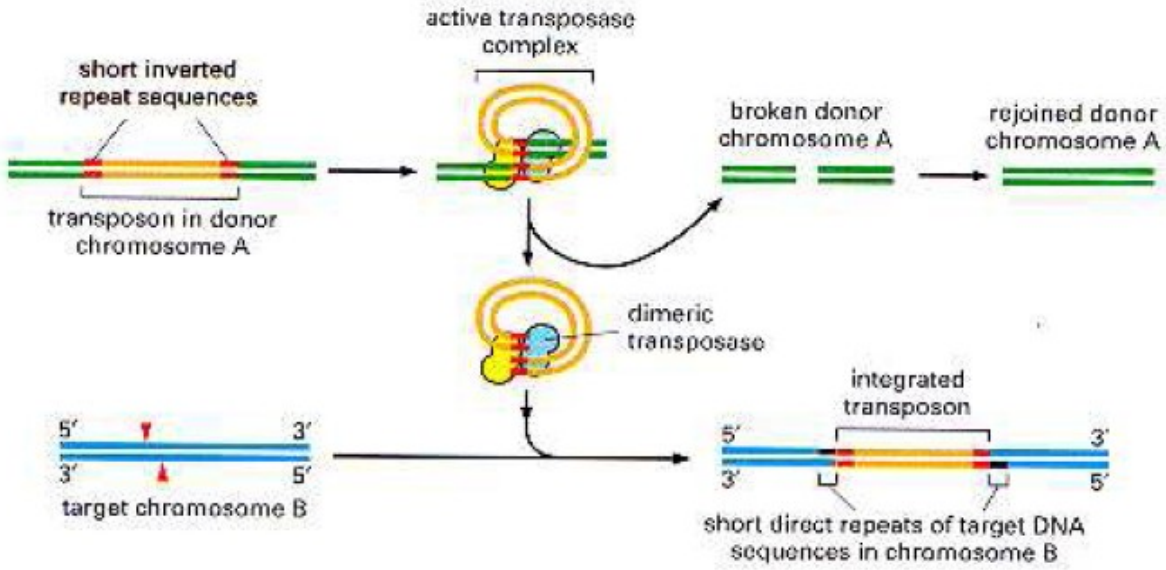
Element	Transposition	Structure	Length	Copy number	Fraction of genome
LINEs	Autonomous	ORF1 ORF2 *(pol)* AAA	1–5 kb	20,000–40,000	21%
SINEs	Nonautonomous	A B AAA	100–300 bp	1,500,000	13%
Retrovirus-like elements	Autonomous	*gag* *pol* *(env)*	6–11 kb	450,000	8%
	Nonautonomous	*gag*	1.5–3 kb		
DNA transposons	Autonomous	*transposase*	2–3 kb	300,000	3%
	Nonautonomous		80–3000 bp		

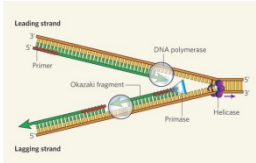
- LINEs and SINEs were first distiguished by their length. Turned out to have different 'lifestyle' and are now distinguished by that.
- DNA transposons and retrotransposons code for *transposase* (or related *integrase*). Insert double-stranded DNA into host genome.
- LINE retroposons and retrovirus-like retrotransposons code for *reverse transcriptase*. Go through intermediate RNA phase.

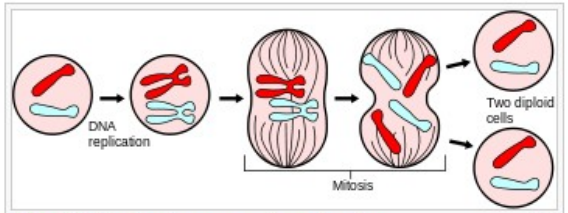
From: Nature, Feb. 2001

DNA Transposons

Transposons move by a cut-and-paste mechanism.



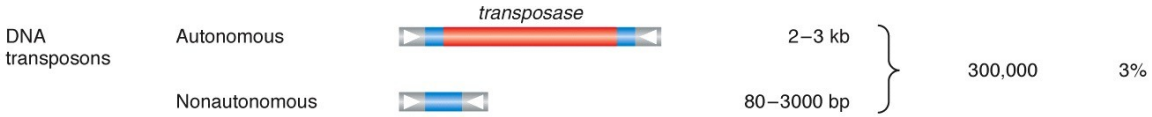




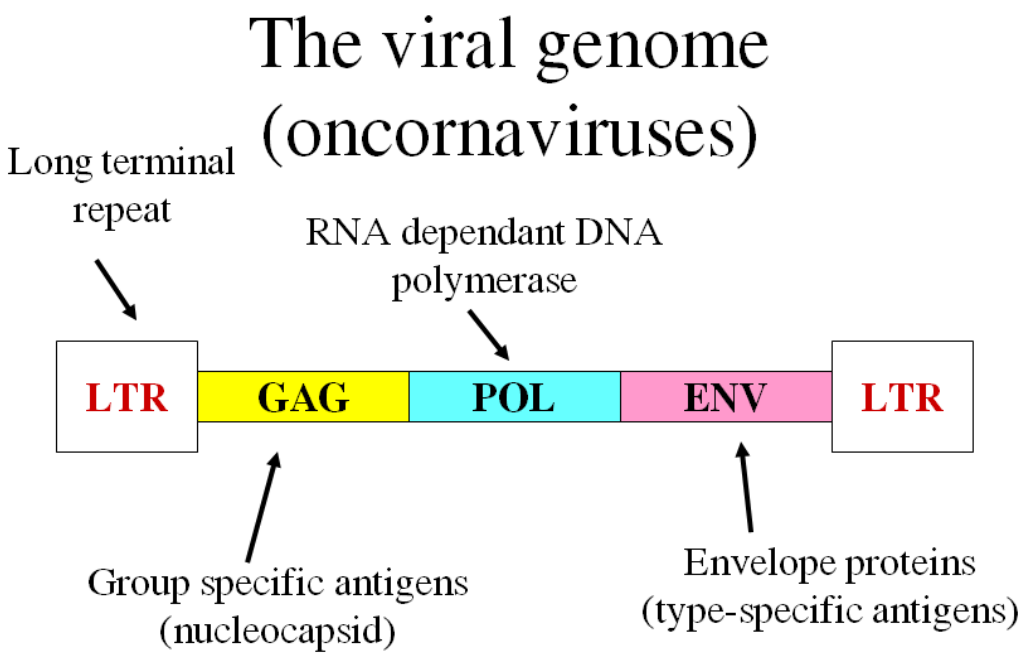
Mitosis divides the chromosomes in a cell nucleus.

Multiply when excising themselves during mitosis, when DNA repair mechanisms can recover removed portion from newly duplicated strand.

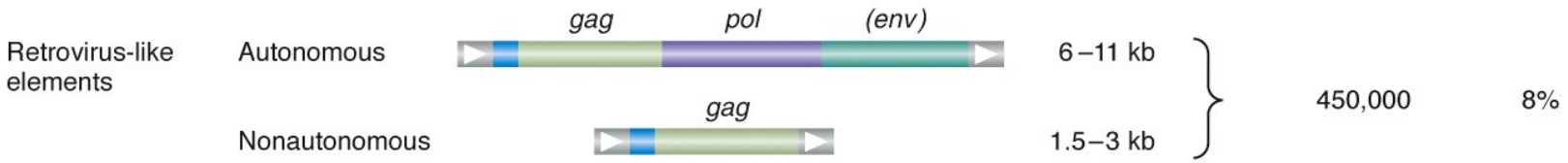
Work in trans, i.e. gene gets translated, then transposase looks for "itself" in genome. Recognises itself by 10 – 30 bp stretch, so often binds to inactive transposon. Result: mutations accumulate, copying becomes less efficient.



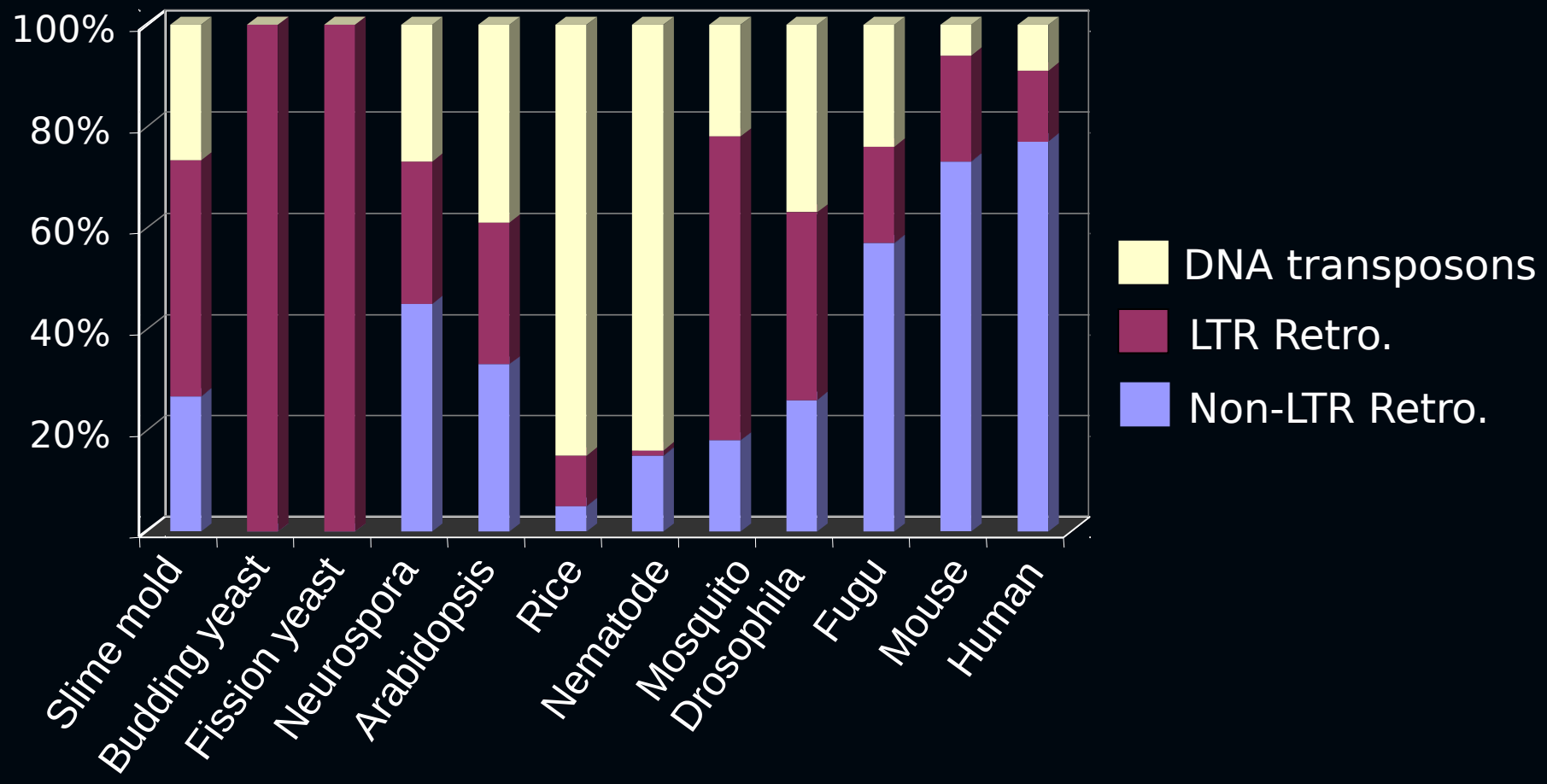
Retrovirus-like Elements



All three genes - GAG, POL, ENV - required for replication



TE composition and assortment vary among eukaryotic genomes



Feschotte & Pritham 2006

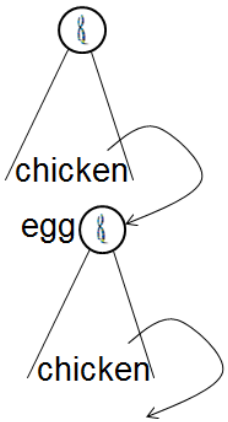
Repeats: mostly neutral

Most repeat events/instances are neutral.

Ie, a repeat instance is dropped in a new place, and joins the rest of the neutral DNA, gradually decaying over time.

Many repeat copies are “dead as a duck” on arrival at their new location (eg 5’ truncation).

Some instances may be active (spawn new instances) for a while, but when an active copy is hit by a mutation – the host is not affected, the instance is inactivated and decays away.



junk

...ACGTACGAC

TT

...ACGTACGAC

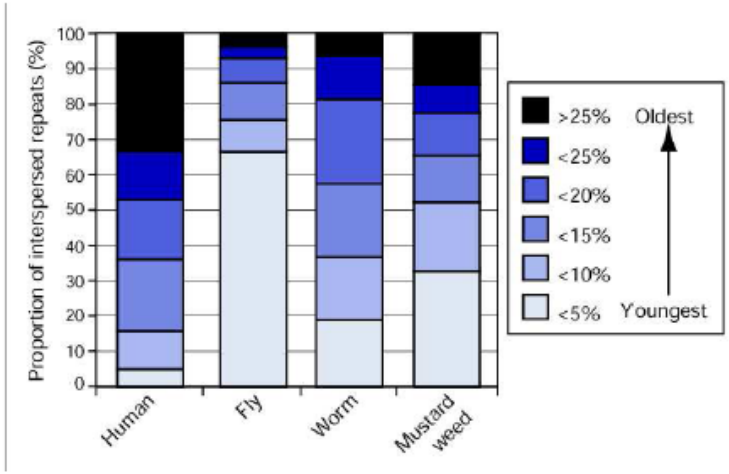
“anything goes”

Repeat Ages

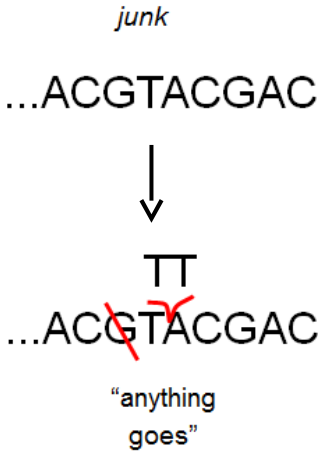
Activity of transposable elements

Activity varies greatly per organism:

- Humans: Rather quiet, ≈ 50 active LINEs, no or very few active DNA transposons, no LTRs through to be active.
- Mice: ≈ 3000 active LINEs, many active DNA transposons, many active LTRs.
- Maize: Genome size doubled in last ≈ 3 Myr because of transposon insertions.



From: Nature vol 420, 5 Dec. 2002



In fruitfly, most TEs have few mutations (relative to concensus = ancestor): young.

In human DNA, there are relatively few young transposable elements.

INTERSPECIES VARIATION IN GENOME SIZE WITHIN VARIOUS GROUPS OF ORGANISMS

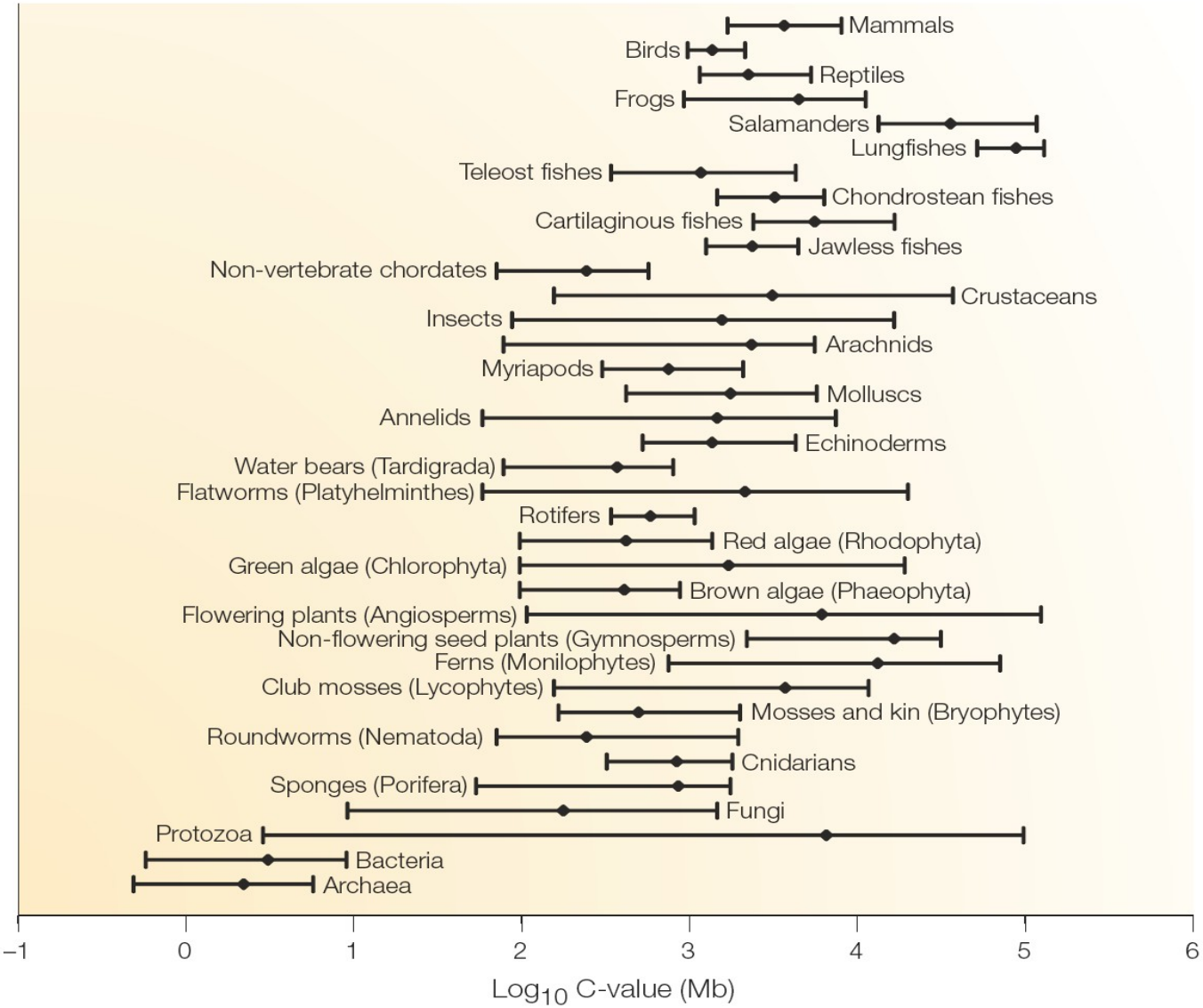
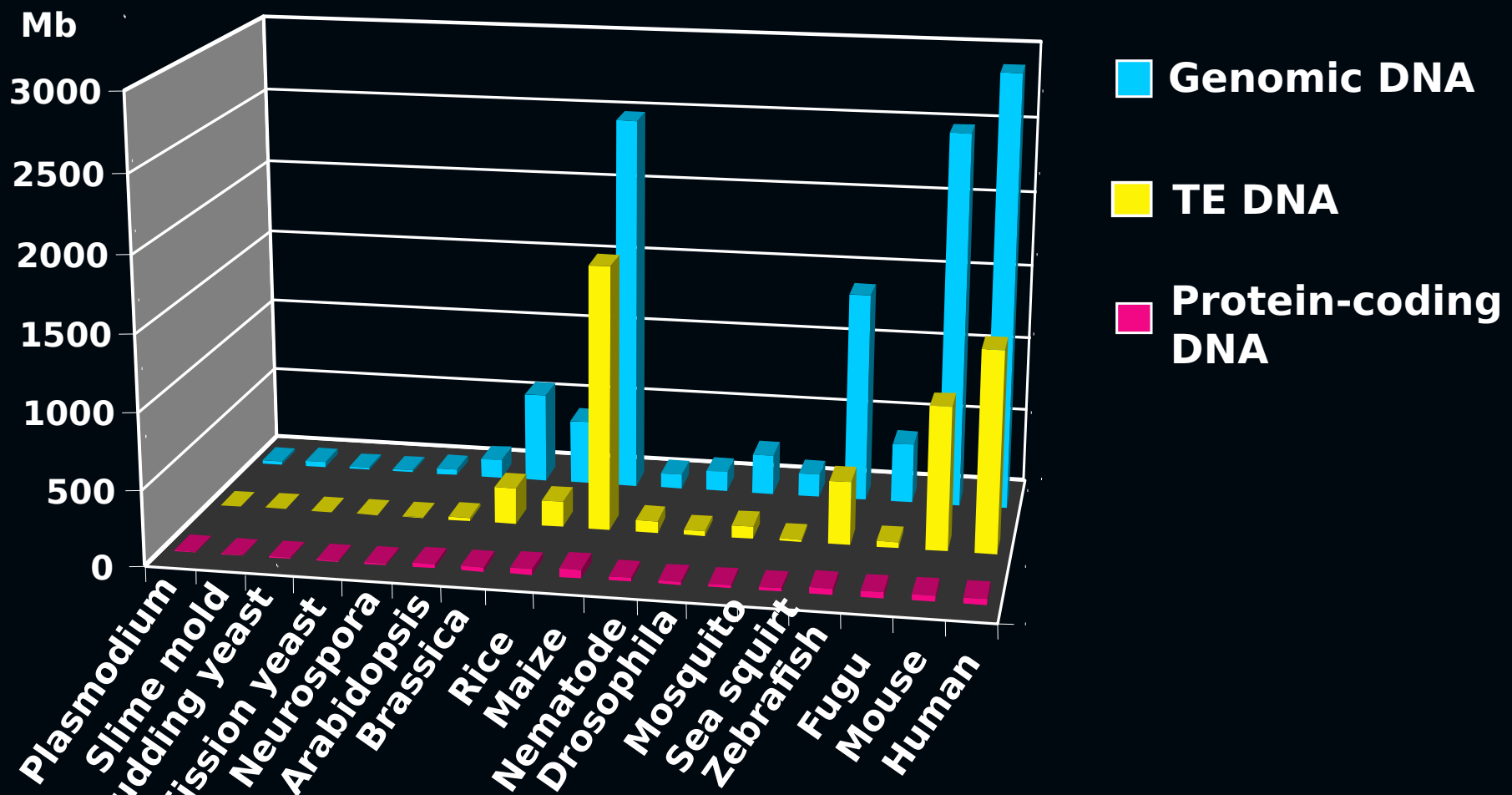


Figure from Ryan Gregory (2005)

The amount of TE correlate positively with genome size



Feschotte & Pritham 2006

The proportion of protein-coding genes decreases with genome size, while the proportion of TEs increases with genome size

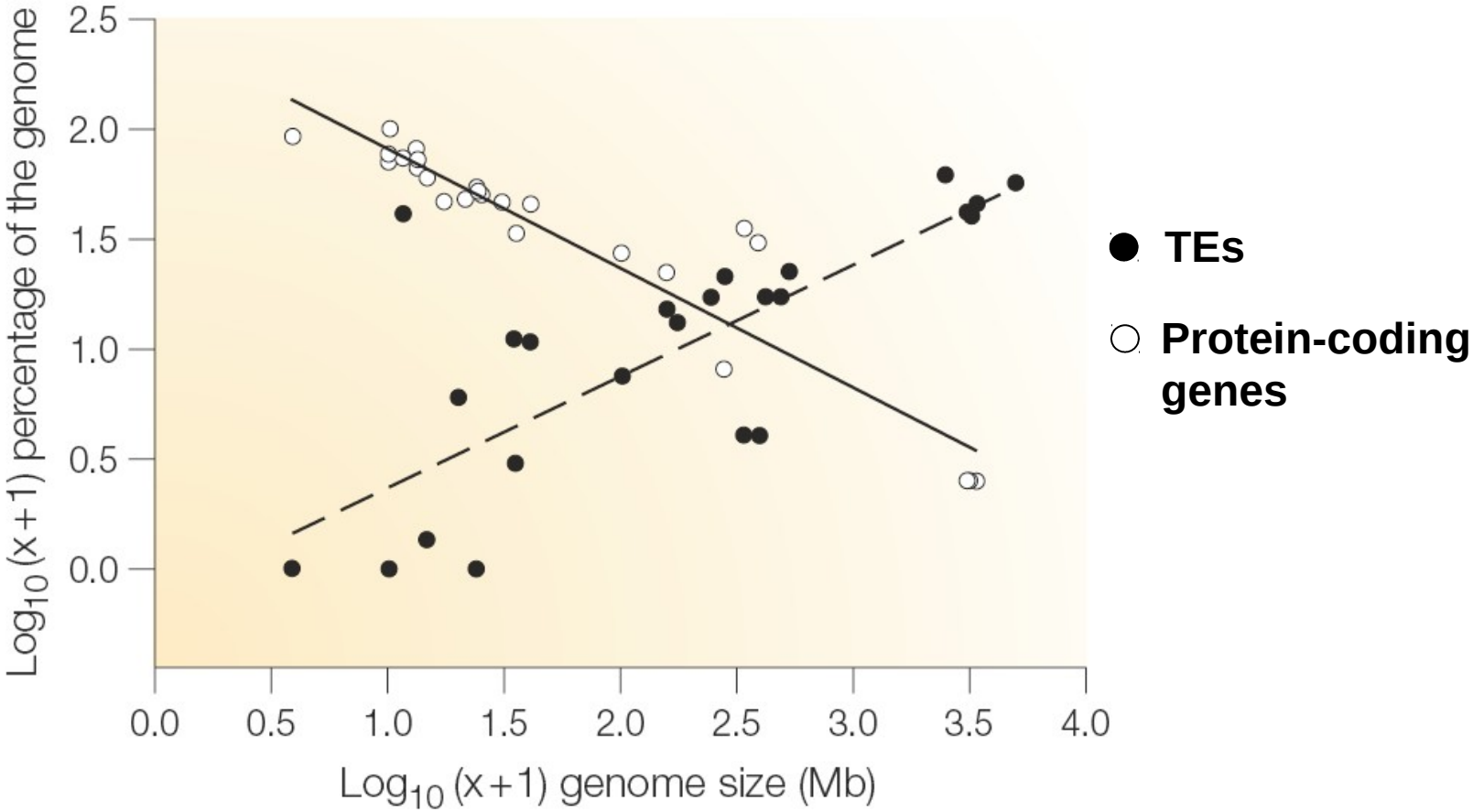


Figure 1 | **The relative contributions of two key components of eukaryotic genomes.** The relationships between haploid genome size and the percentage of the genome that consists of protein-coding genes (white circles) and transposable elements (black circles) are shown. The data are based on species that have been the subject of large-scale sequencing studies. Larger genomes contain proportionately fewer genes and more transposable elements than small genomes. A $\log_{10}(x+1)$ transformation was used because some tiny genomes contain no recognizable transposable elements.

Gregory, Nat Rev Genet 2005

Repeats: not just neutral

So far we treated all repeat proliferation events as neutral.

While the majority of them appear to be neutral, this is certainly not the case for all repeat instances.

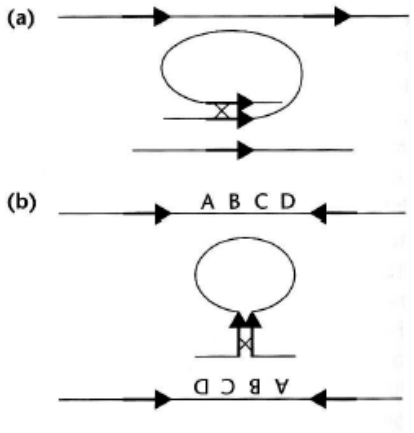
And because there are so many repeat instances even a small fraction of all repeats can be a big set compared to other types of elements in the genome.

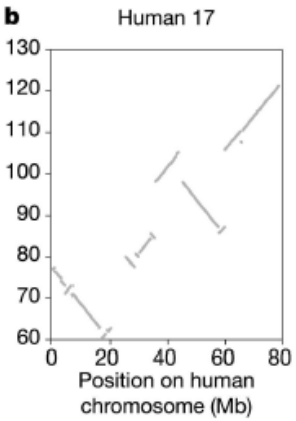
(Eg, 1% of ½ the genome is still a lot)

Transposable elements: Effect on genome

High copy number of transposable elements provide many opportunities for unequal homologous recombination.

When this happens within a chromosome, leads to deletions or inversions.





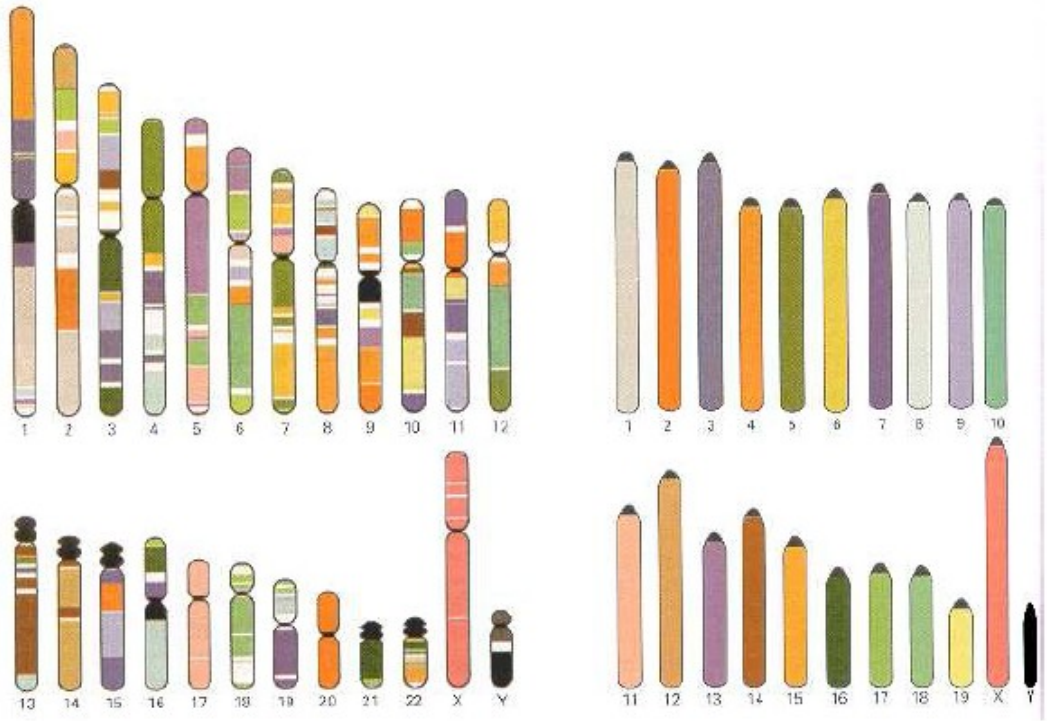
EHG T 622; Nature Feb. 2001

Direct evidence:

- Existence of solo-LTRs, result of recombination between two LTRs flanking one (or two) LTR-retrotransposon(s). EHG T 622
- 20% of Alus have no flanking target-site repeats. CW Schmid, Nuc Acids Res 1998 26(20) 4541

Transposable elements: Effect on genome

When unequal homologous recombination occurs *between* chromosomes, chromosome rearrangements occur.



From: Alberts et al., The Cell, after Nature vol 420, 5 Dec. 2002

(Right: mouse chromosomes. Left: human chromosomes, colored according to which mouse chromosome region correspond to)

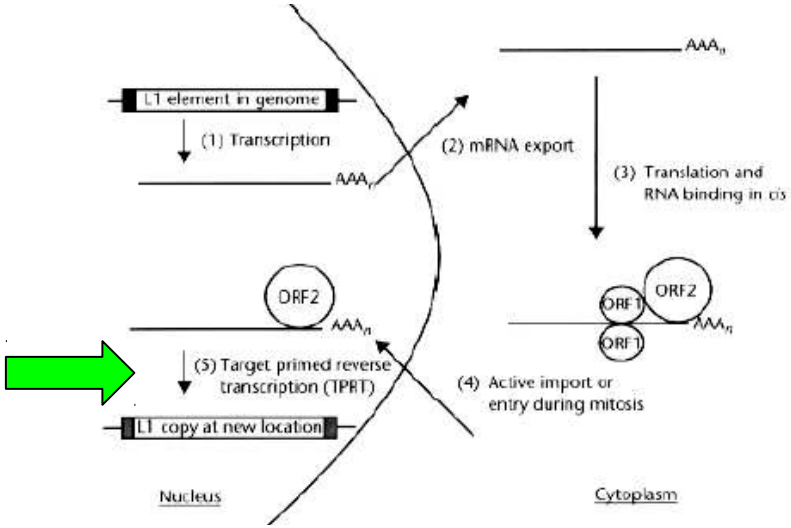
Repeats & Retroposed Genes

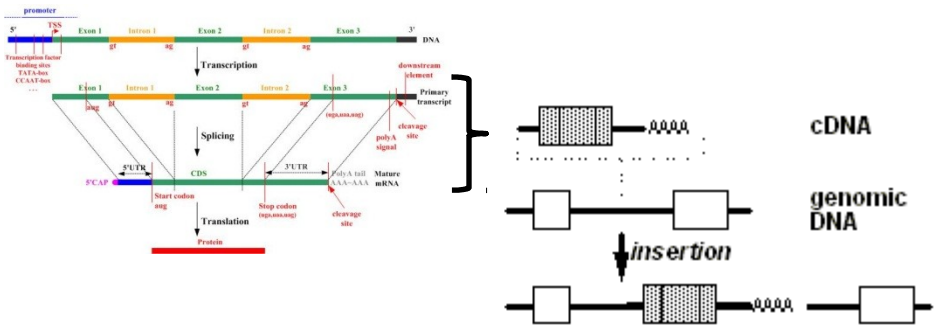
Remember how LINEs reverse transcribe copies of themselves back into the genome? How they sometimes reverse transcribe SINEs "by mistake"? Well, they also grab m/ncRNAs and reverse transcribe them into the genome!

<u>Retrogenes ("retrotranscribed")</u>:

Protein coding RNA that was reverse transcribed and inserted back into the genome.

The RNA can be grabbed at any stage (partial/full transcript, before/during/after all introns are spliced).

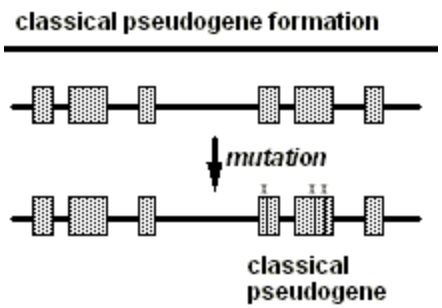


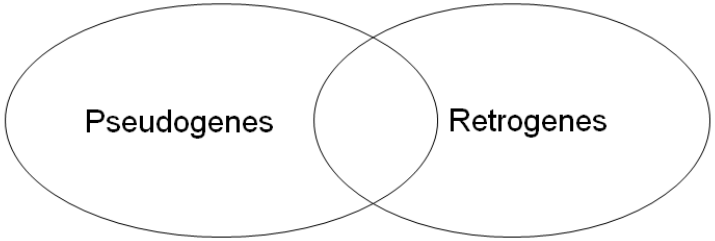


Retroposed Genes & Pseudogenes

<u>Pseudogenes (“dead genes”)</u>:

Genomic sequences that resemble (originated from) genes that no longer make proteins.

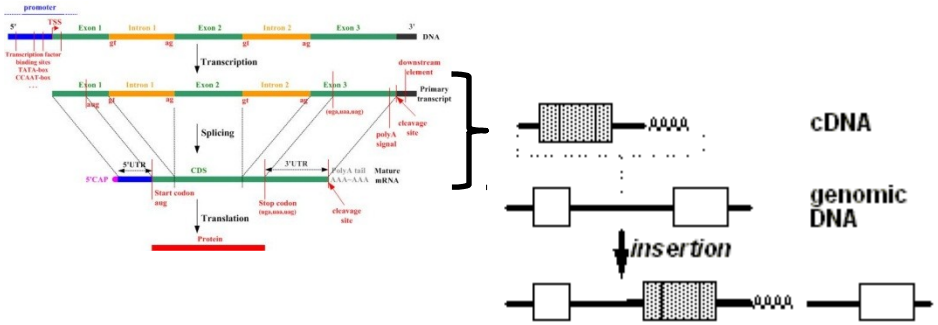




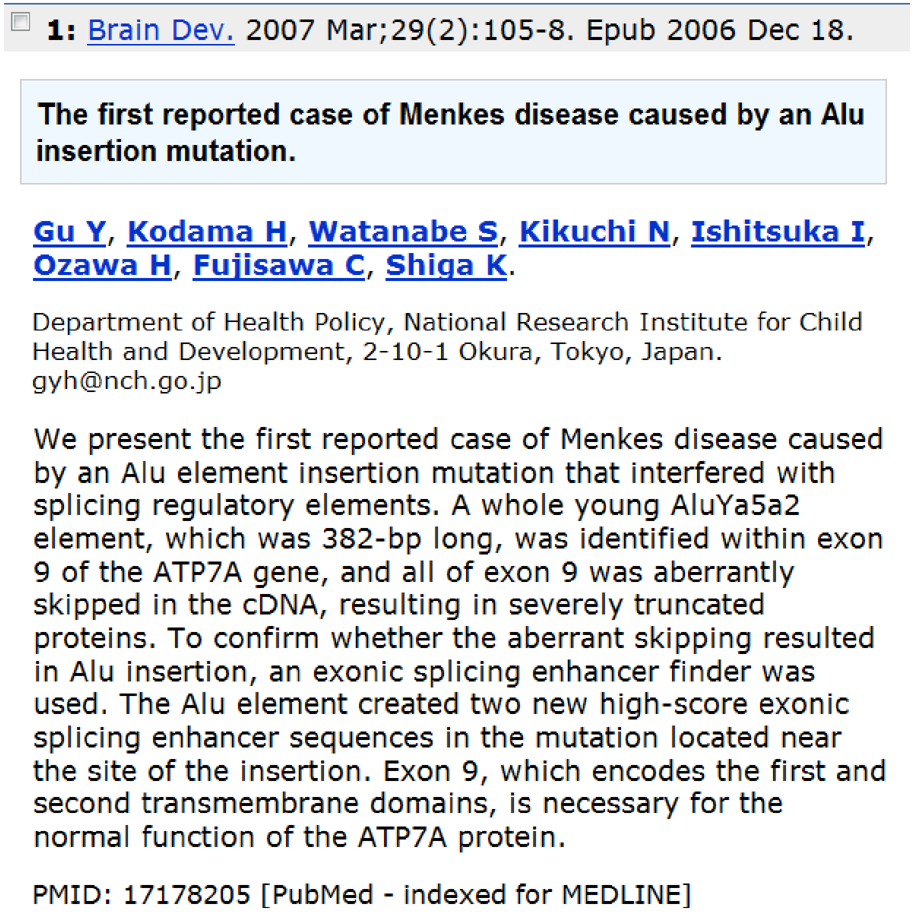
<u>Retrogenes (“retrotranscribed”)</u>:

Protein coding RNA that was reverse transcribed and inserted back into the genome.

The RNA can be grabbed at any stage (partial/full transcript, before/during/after all introns are spliced).



Repeat Insertions Can “Break Things”



☐ **1:** Brain Dev. 2007 Mar;29(2):105-8. Epub 2006 Dec 18.

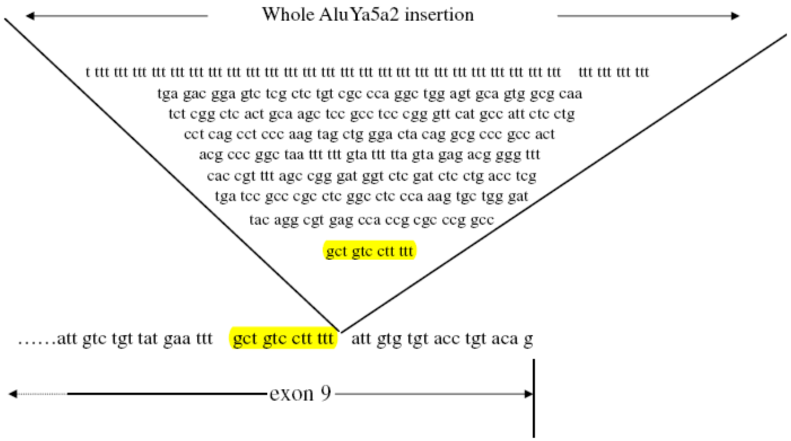
The first reported case of Menkes disease caused by an Alu insertion mutation.

Gu Y, **Kodama H**, **Watanabe S**, **Kikuchi N**, **Ishitsuka I**, **Ozawa H**, **Fujisawa C**, **Shiga K**.

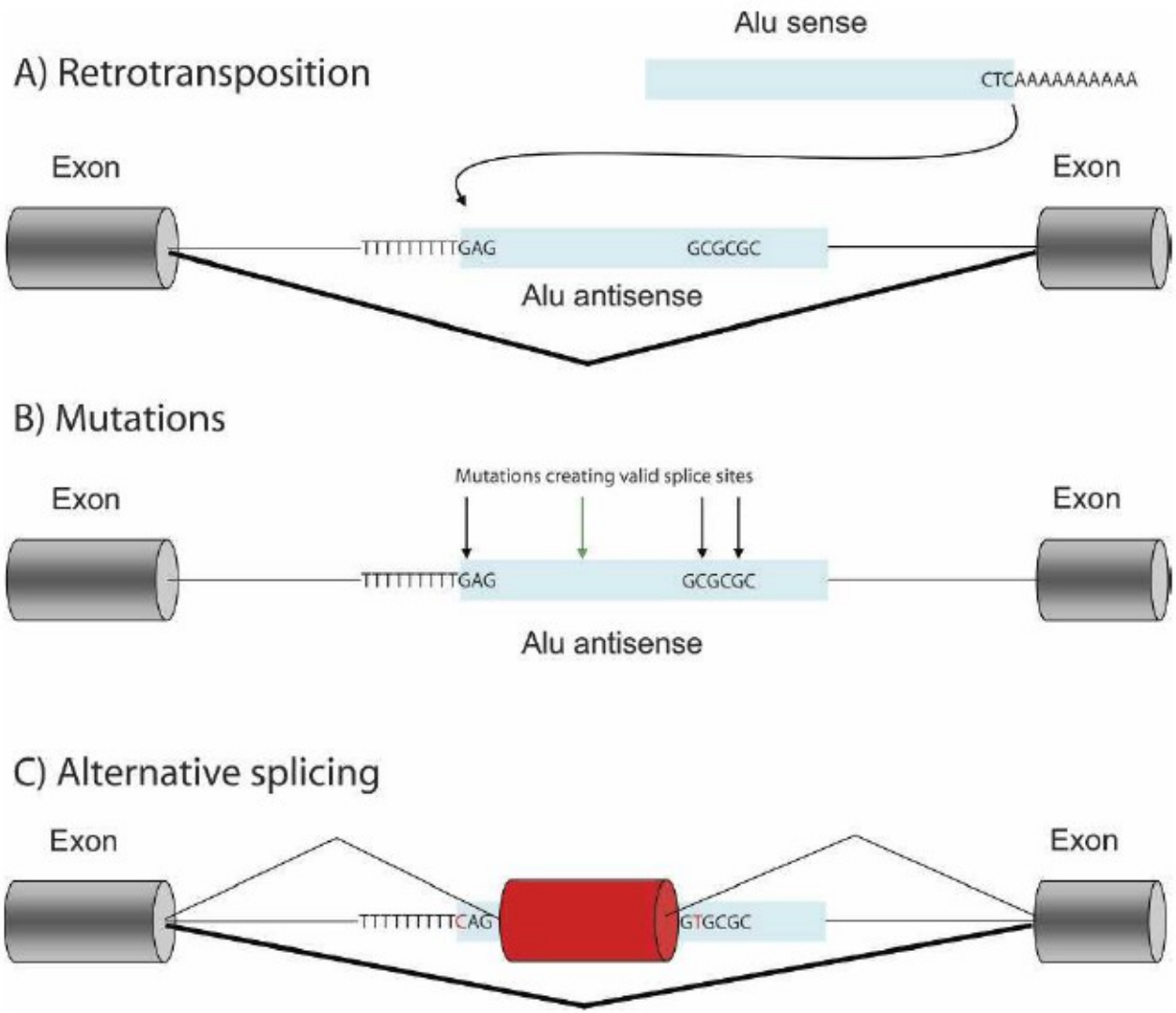
Department of Health Policy, National Research Institute for Child Health and Development, 2-10-1 Okura, Tokyo, Japan.
gyh@nch.go.jp

We present the first reported case of Menkes disease caused by an Alu element insertion mutation that interfered with splicing regulatory elements. A whole young AluYa5a2 element, which was 382-bp long, was identified within exon 9 of the ATP7A gene, and all of exon 9 was aberrantly skipped in the cDNA, resulting in severely truncated proteins. To confirm whether the aberrant skipping resulted in Alu insertion, an exonic splicing enhancer finder was used. The Alu element created two new high-score exonic splicing enhancer sequences in the mutation located near the site of the insertion. Exon 9, which encodes the first and second transmembrane domains, is necessary for the normal function of the ATP7A protein.

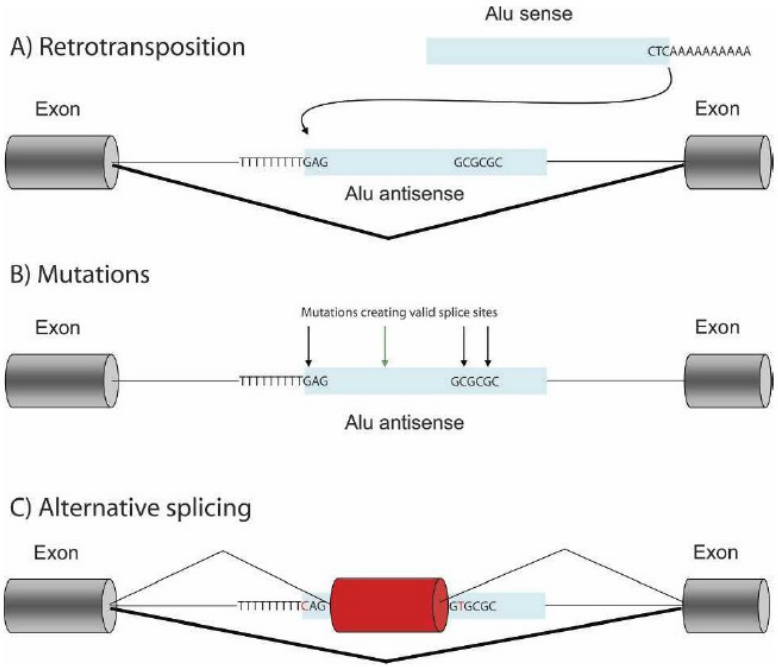
PMID: 17178205 [PubMed - indexed for MEDLINE]



Repeat Insertions Can “Make Things”



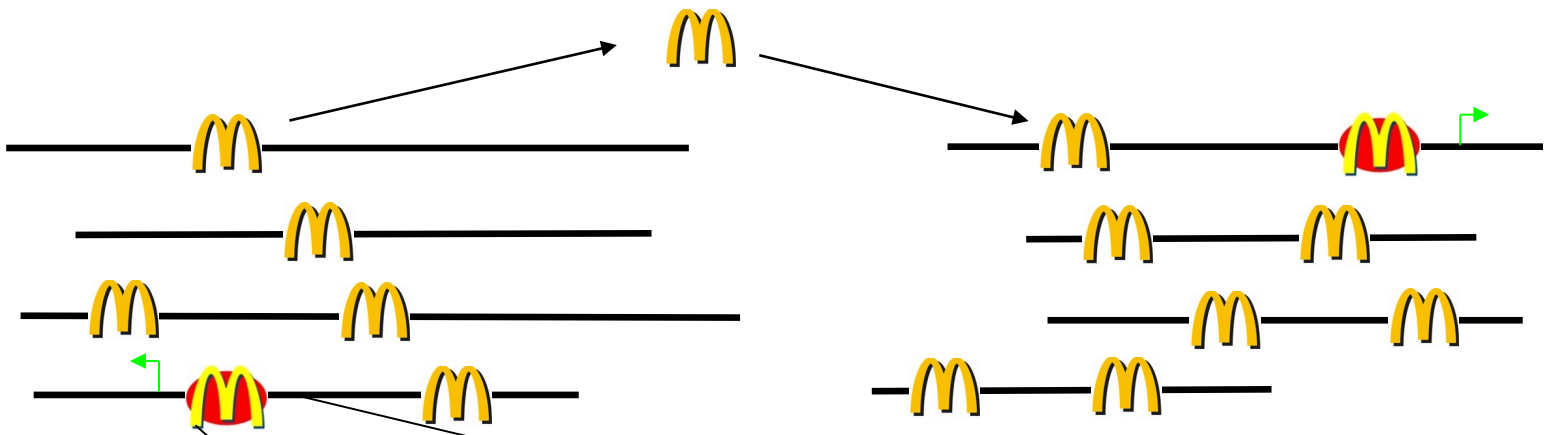
Any Sequence Can Become Functional



Random mutation (especially in a large place like our genome) can create functional DNA elements out of neutrally evolving sequences.

So is there anything special about a piece of DNA from a repetitive origin that takes on a new function?

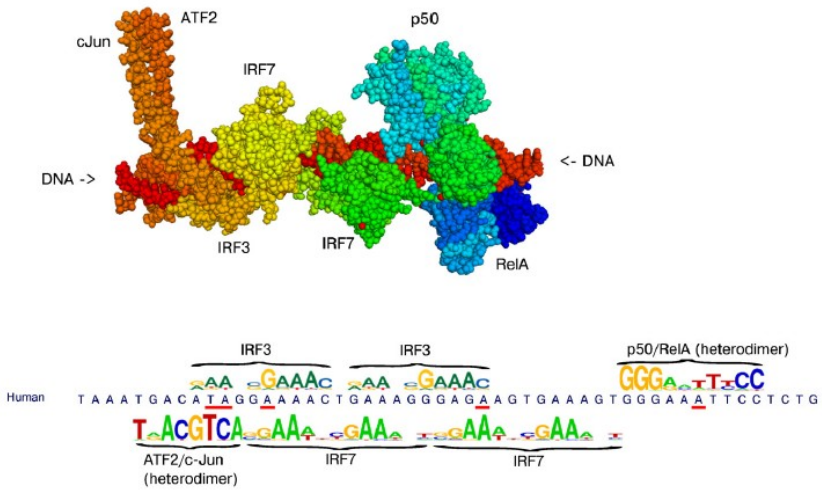
Regulatory elements from Mobile Elements



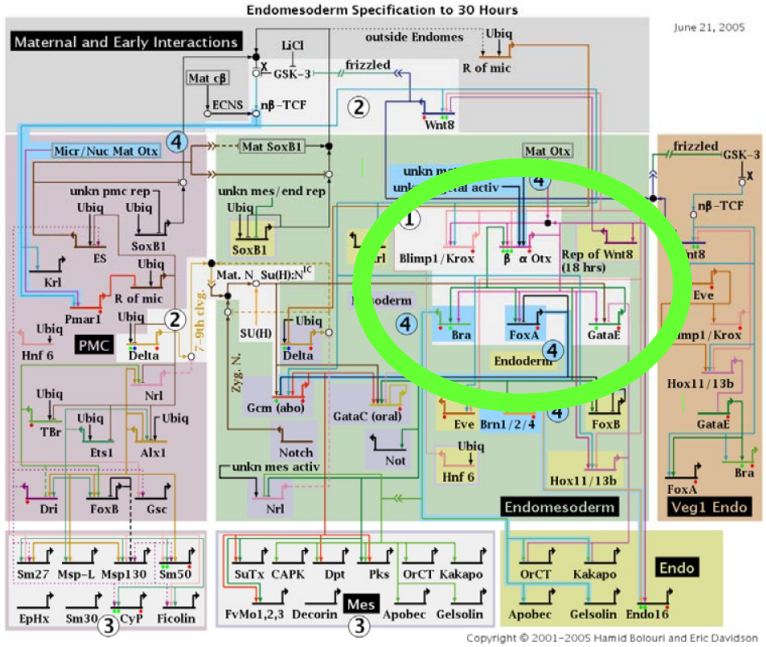
Co-option event, probably due to favorable genomic *context*

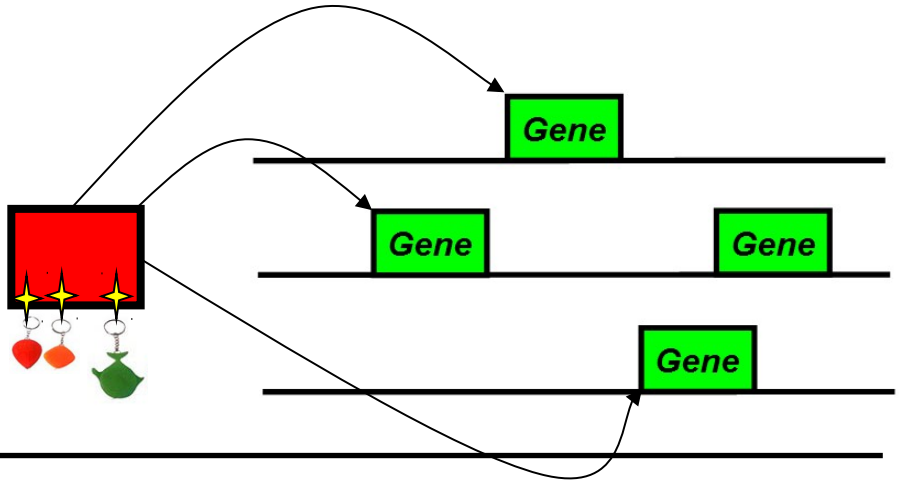
[Yass is a small town in New South Wales, Australia.]

Britten & Davidson Hypothesis: Repeat to Rewire!

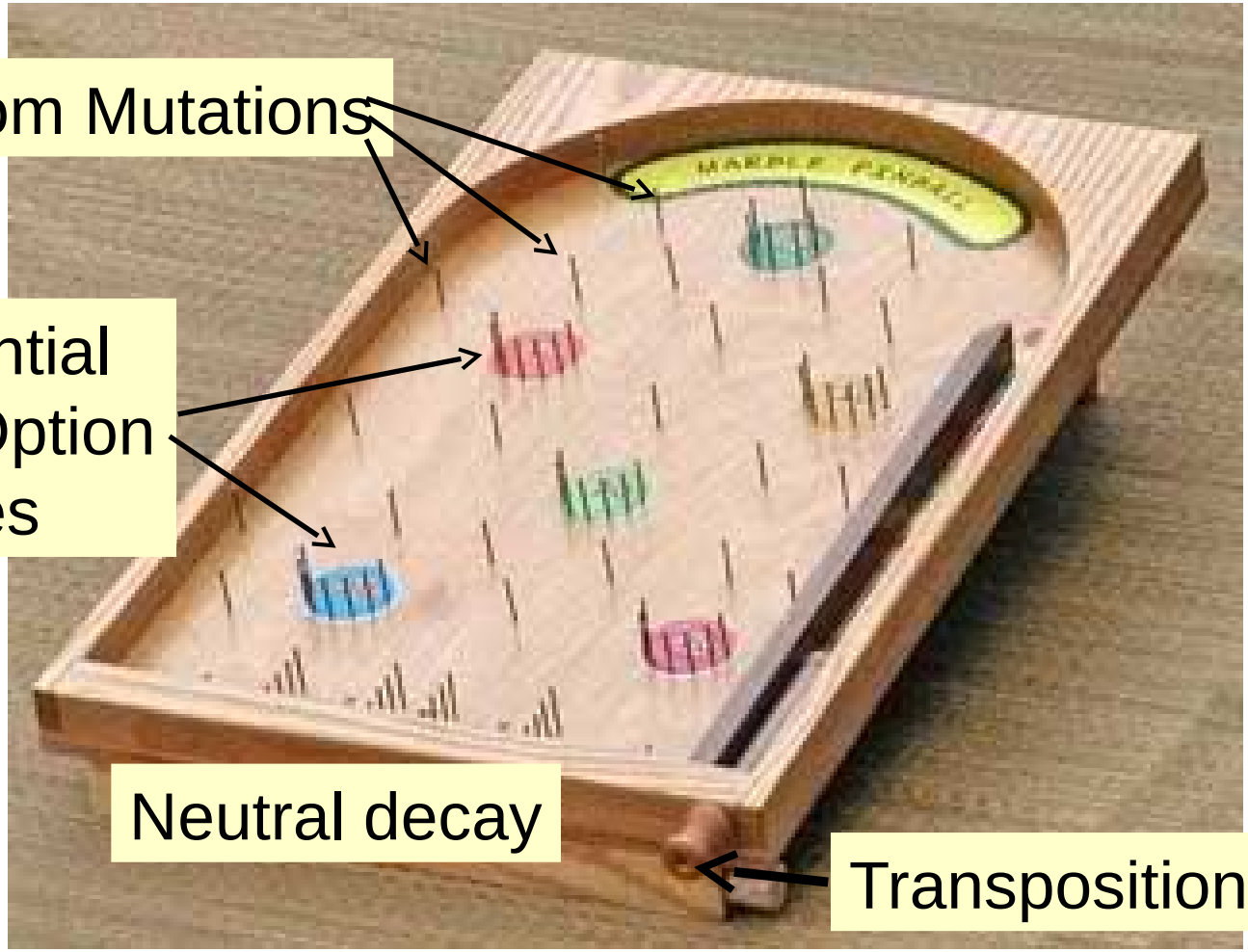


Enhancer structure reminder

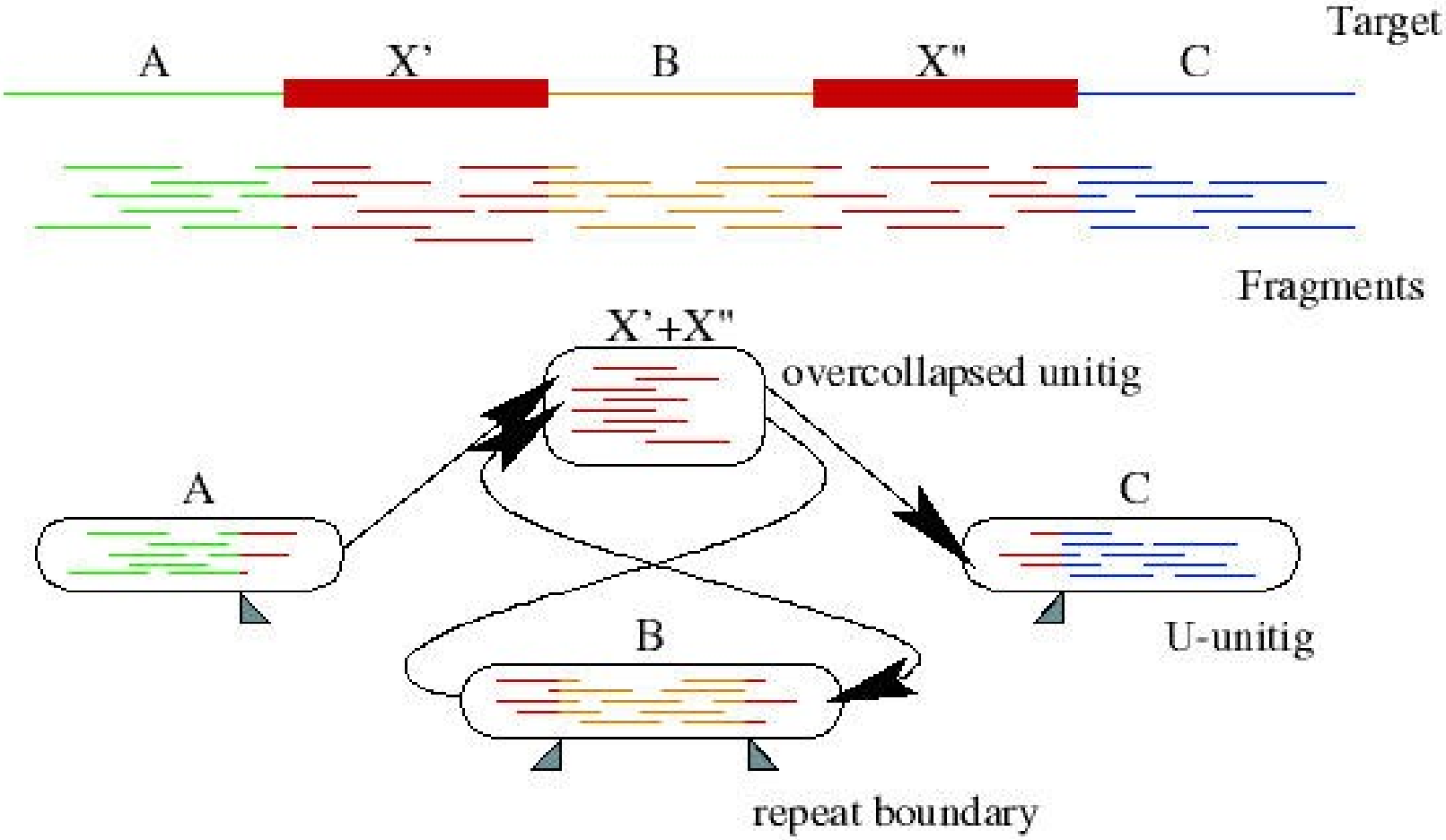




The Road to Co-Option



Assemby Challenges



Inferring Phylogeny Using Repeats

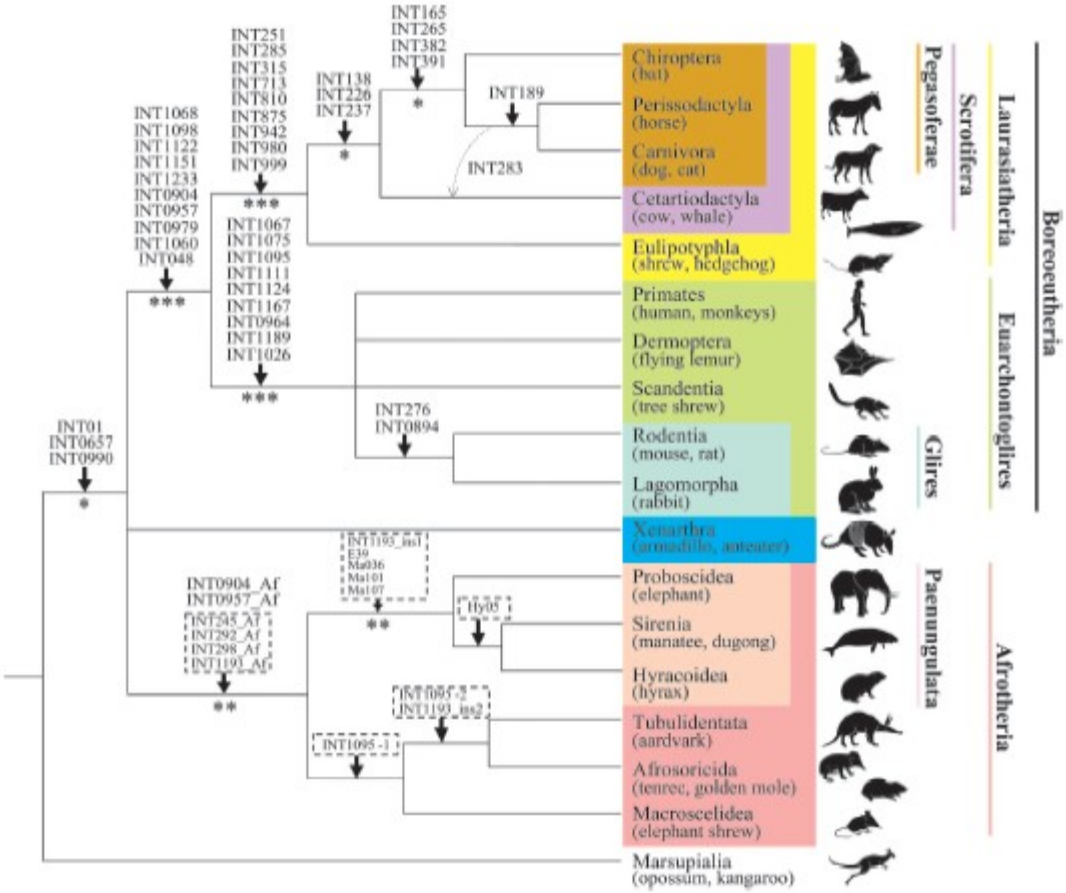
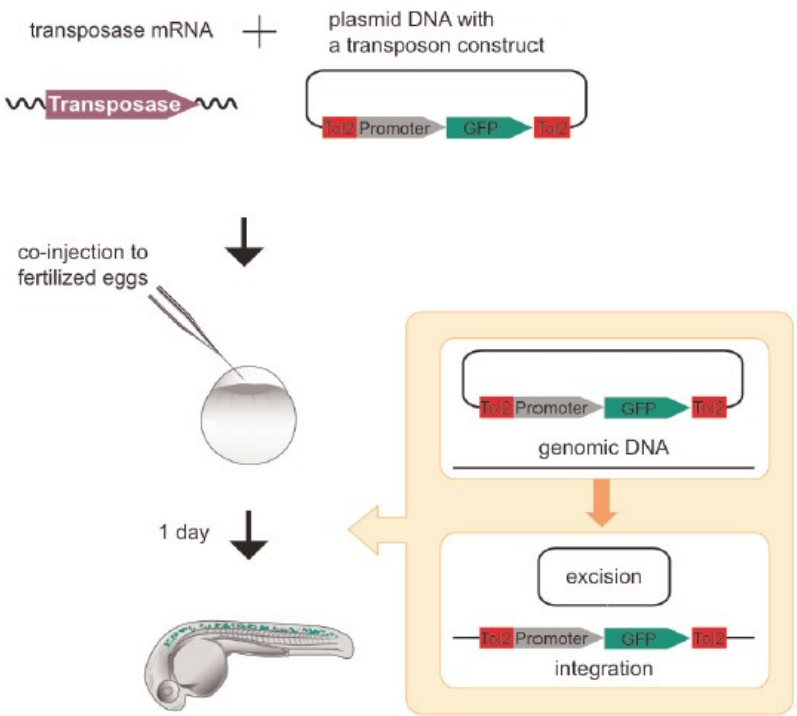


Fig. 2. An interordinal mammalian phylogeny reconstructed by our retroposon insertion analysis. Downward arrows denote insertions of retroposons into each lineage. Locus INT283, denoted by a dashed arrow, supports the monophyly of Cetartiodactyla, Perissodactyla, and Carnivora. The loci surrounded by dashed lines in Afrotheria were identified in our previous study (22). Asterisks below the branches denote that the monophylies are statistically significant (*, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$).

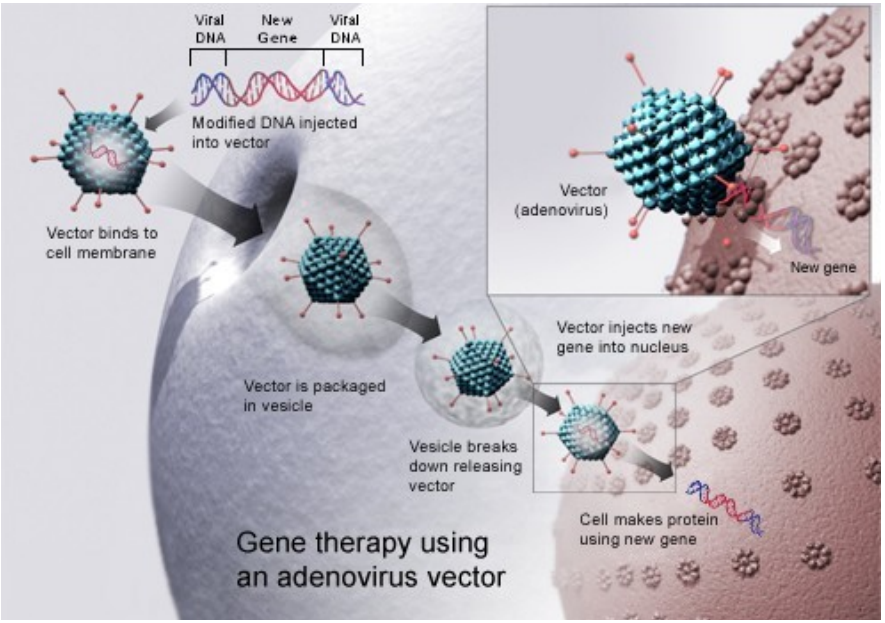
[Nishihara et al, 2006]

Transposons as Genetics Engineering Tools

The medaka fish *Tol2* element is an autonomous transposon that encodes a fully functional transposase. The transposase protein can catalyze transposition of a transposon construct that has 200 and 150 base pairs of DNA from the left and right ends of the *Tol2* sequence, respectively. These sequences contain essential terminal inverted repeats and subterminal sequences. DNA inserts of fairly large sizes (as large as 11 kilobases) can be cloned between these sequences without reducing transpositional activity. The *Tol2* transposon system has been shown to be active in all vertebrate cells tested thus far, including zebrafish, *Xenopus*, chicken, mouse, and human. In this review I describe and discuss how the *Tol2* transposon is being applied to transgenic studies in these vertebrates, and possible future applications.



Human Gene Therapy



Repeats: fun conspiracy theories

1. Repeats wreck so much havoc in the genome, by inserting themselves, deleting segments between instances and more – they make the genome feel like a “rolling sea”. Maybe it is because of them that enhancers “learned” to work irrespective of distance and orientation?

2. When the last active copy of a repeat dies, all instances of the repeat are now decaying. Wait long enough and they lose resemblance to each other. Look in 200My and you never know they belonged to the same repeat family. So… if half the genome is recognizable as repetitive now, how much of the genome originated from repeats? Most of it?

Repeats: fun conspiracy theories

3. If repeats do significantly accelerate the rate of creation of novel functional (gene/regulation) elements – how many functional elements today came from repeats (including old ones we no longer can recognize as such)? Most?

4. Is that why our genome “tolerates” these elements?

5. You make a conspiracy theory…

6. You think of ways* to solve one!

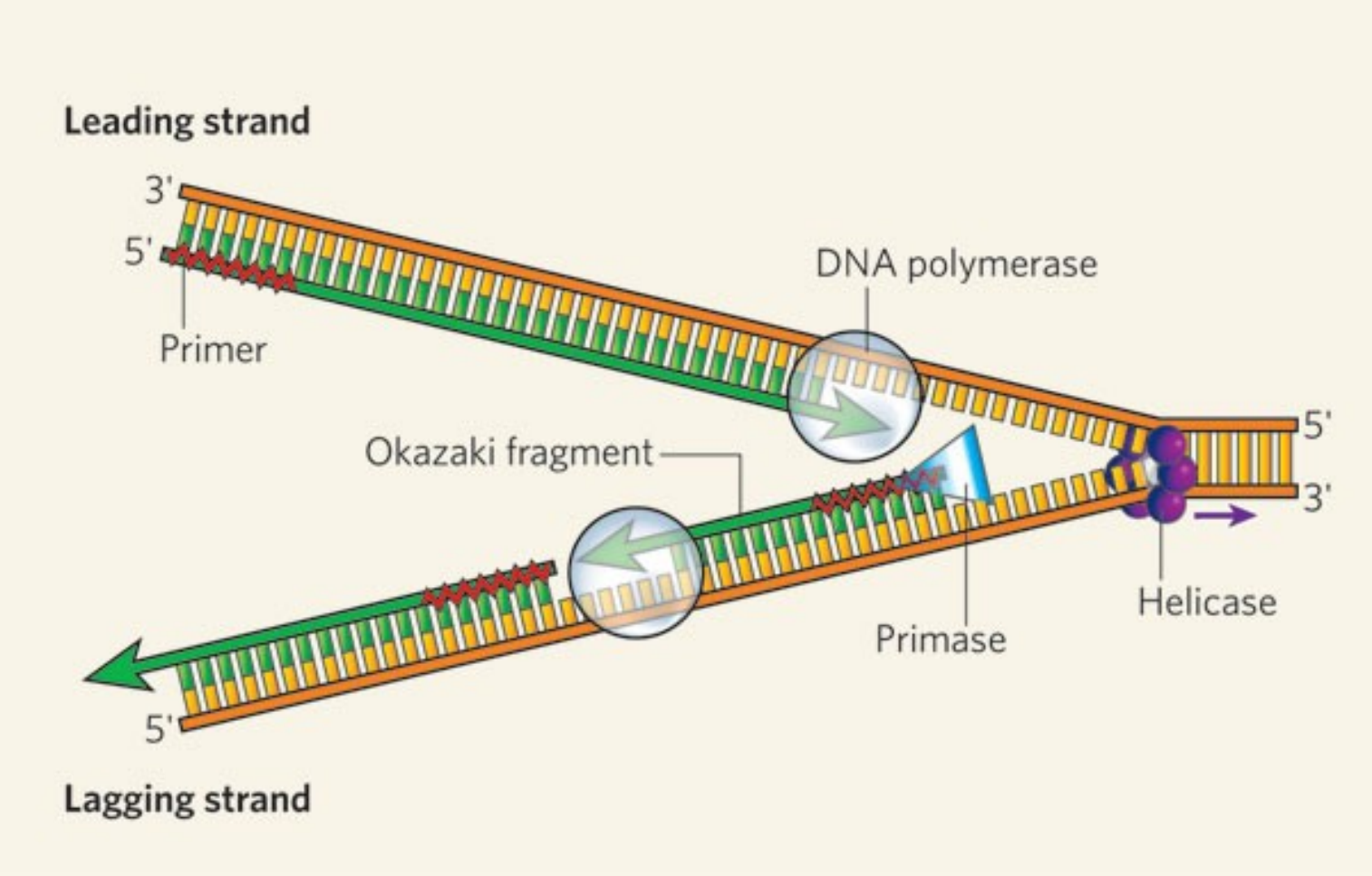
* Computationally. Evolution is mostly computational business.

II. Simple Repeats

- Every possible motif of mono-, di, tri- and tetranucleotide repeats is vastly overrepresented in the human genome.
- These are called microsatellites,
  Longer repeating units are called minisatellites,
  The real long ones are called satellites.

  AAAAAAAAA
  CACACACAC
  CAACAACAA
- Highly polymorphic in the human population.
- Highly heterozygous in a single individual.
- As a result microsatellites are used in paternity testing, forensics, and the inference of demographic processes.
- There is no clear definition of how many repetitions make a simple repeat, nor how imperfect the different copies can be.
- Highly variable between species: e.g., using the same search criteria the mouse & rat genomes have 2-3 times more microsatellites than the human genome. They're also longer in mouse & rat.

DNA Replication



Simple Repeats Create Funky DNA structures

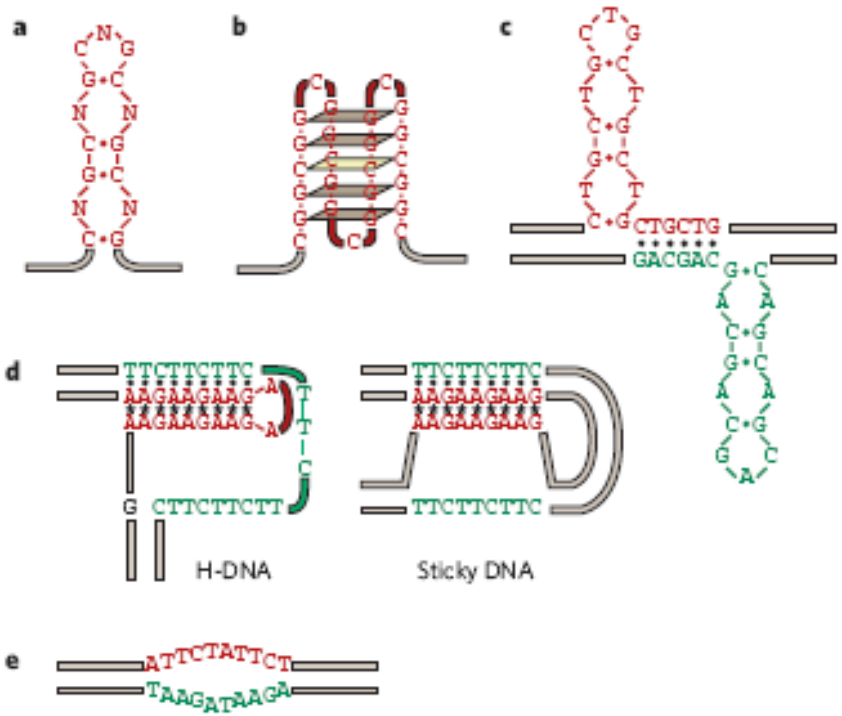


Figure 2 | Unusual DNA structures formed by expandable repeats. Repetitive DNA can form several unusual structures, examples of which are shown. The structure-prone strand of the repetitive run is shown in red, its complementary strand in green, and flanking DNA in beige. **a**, An imperfect hairpin formed by $(CNG)_n$ repeats. **b**, A quadruplex-like structure formed by the $(CGG)_n$ repeat. Brown rectangles indicate G quartets, and the yellow rectangle indicates an i motif. **c**, A slipped-stranded structure formed by the $(CTG)_n \bullet (CAG)_n$ repeat. **d**, H-DNA and sticky DNA formed by the $(GAA)_n \bullet (TTC)_n$ repeat. Only one possible isoform, in which the homopurine strand is donated to the triplex, is shown for both structures. Reverse Hoogsteen pairing is indicated by asterisks. **e**, A DNA-unwinding element formed by the $(ATTCT)_n \bullet (AGAAT)_n$ repeat.

These Bumps Give The DNA Polymerase Hiccups

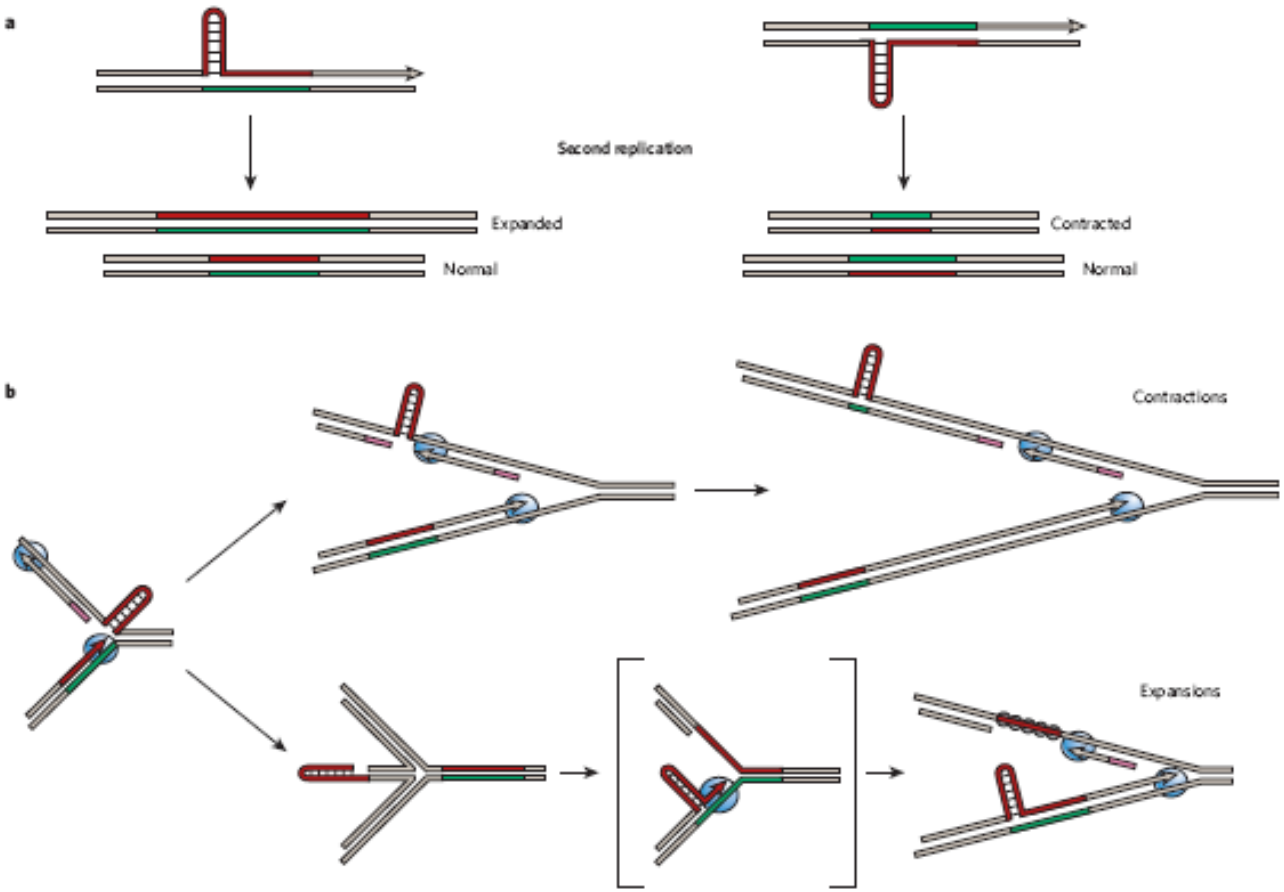


Figure 3 | Replication mechanisms for repeat expansion. a, After two rounds of replication, formation of a repetitive hairpin on the nascent strand results in repeat expansions (left panel), whereas the presence of the same structure on the template strand results in repeat contractions (right panel). **b**, A model for repeat instability based on replication fork stalling and restarting within the repetitive run is shown. Repeat contractions (upper pathway) occur when the machinery for the lagging-strand synthesis skips the repetitive hairpin on the lagging-strand template. Repeat expansions (lower pathway) can occur during replication fork reversal and restart, leading to the formation of a repetitive hairpin on the nascent leading strand. The structure-prone strand of the repetitive run is shown in red, its complementary strand in green, and flanking DNA in beige. DNA polymerases are shown in blue, primers for Okazaki fragments in pink, and single-stranded-DNA-binding proteins as grey circles. The bracketed intermediate contains a hairpin on the nascent strand, which can also be stabilized by MSH2–MSH3.

Expandable Repeats and Disease

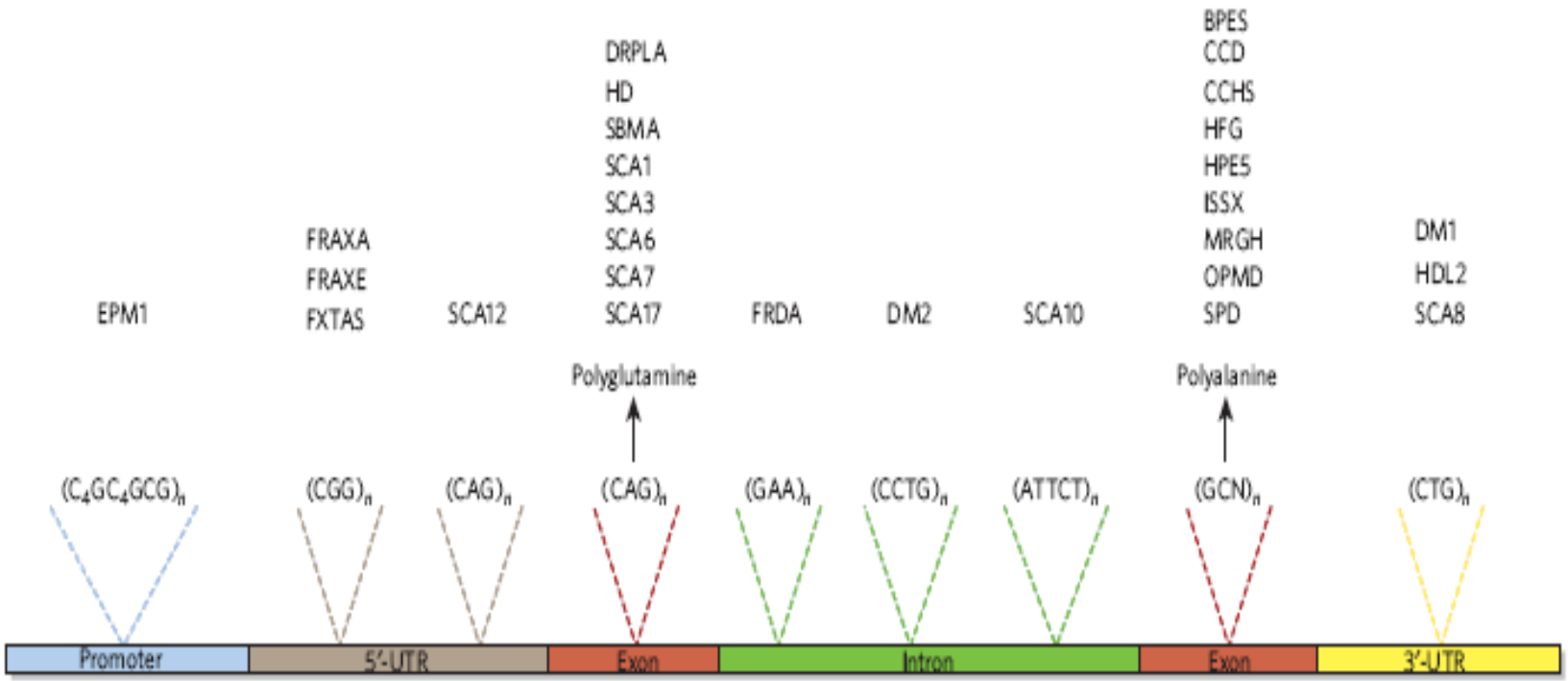


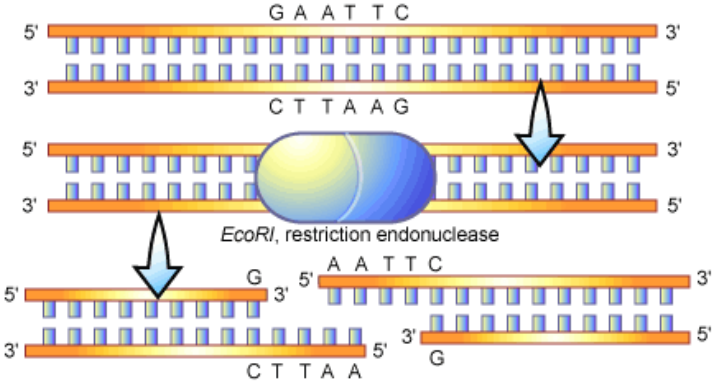
Figure 1 | Location of expandable repeats responsible for human diseases. The sequence and location within a generic gene of expandable repeats that cause human diseases are shown, and the associated diseases are listed. BPES, blepharophimosis, ptosis and epicanthus inversus; CCD, cleidocranial dysplasia; CCHS, congenital central hypoventilation syndrome; DM, myotonic dystrophy; DRPLA, dentatorubral–pallidoluysian atrophy; EPM1, progressive myoclonic epilepsy 1; FRAXA, fragile X syndrome; FRAXE, fragile X mental retardation associated with *FRAXE* site; FRDA, Friedreich's ataxia; FXTAS, fragile X tremor and ataxia syndrome; HD, Huntington's disease; HDL2, Huntington's-disease-like 2; HFG, hand–foot–genital syndrome; HPE5, holoprosencephaly 5; ISSX, X-linked infantile spasm syndrome; MRGH, mental retardation with isolated growth hormone deficiency; OPMD, oculopharyngeal muscular dystrophy; SBMA, spinal and bulbar muscular atrophy; SCA, spinocerebellar ataxia; SPD, synpolydactyly.

Restriction Enzymes

- Restriction enzymes recognize and make a cut within specific DNA sequences, known as ***restriction sites***.
- This is usually a 4-6 base pair palindromic sequence.
- Naturally found in different types of bacteria
- Bacteria use restriction enzymes to protect themselves from foreign DNA
- Many have been isolated and sold for use in lab work

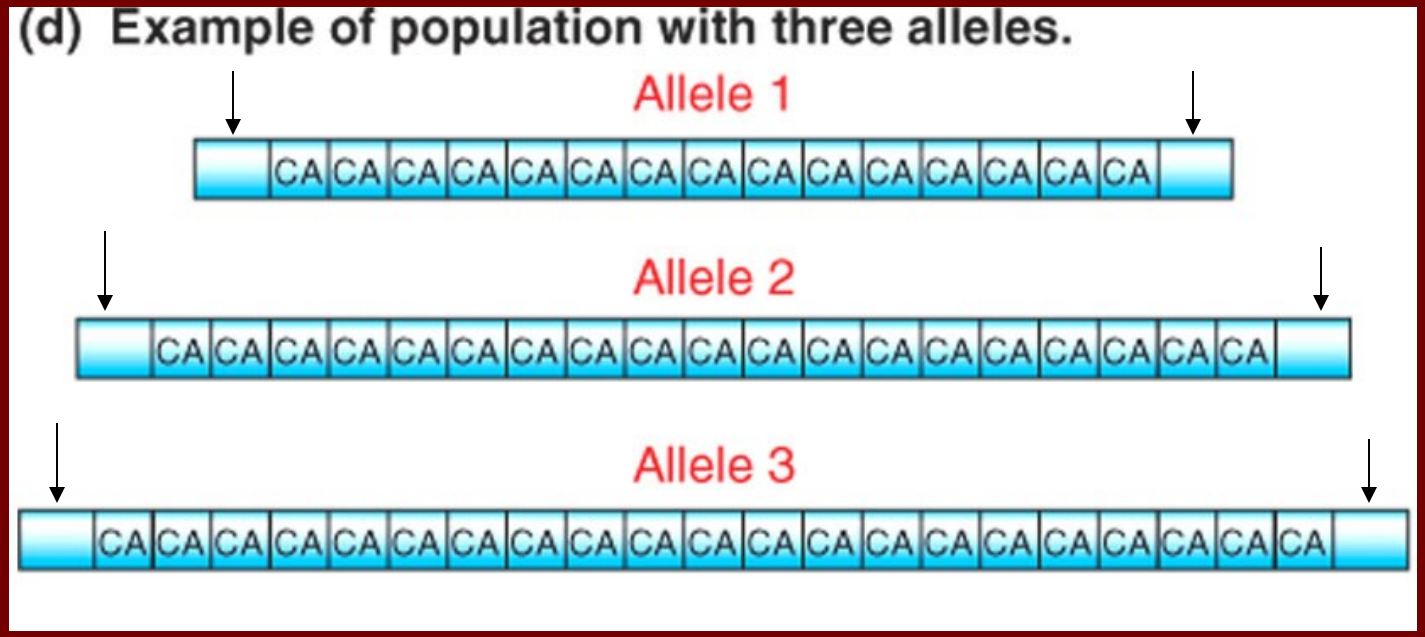
blunt end

sticky end

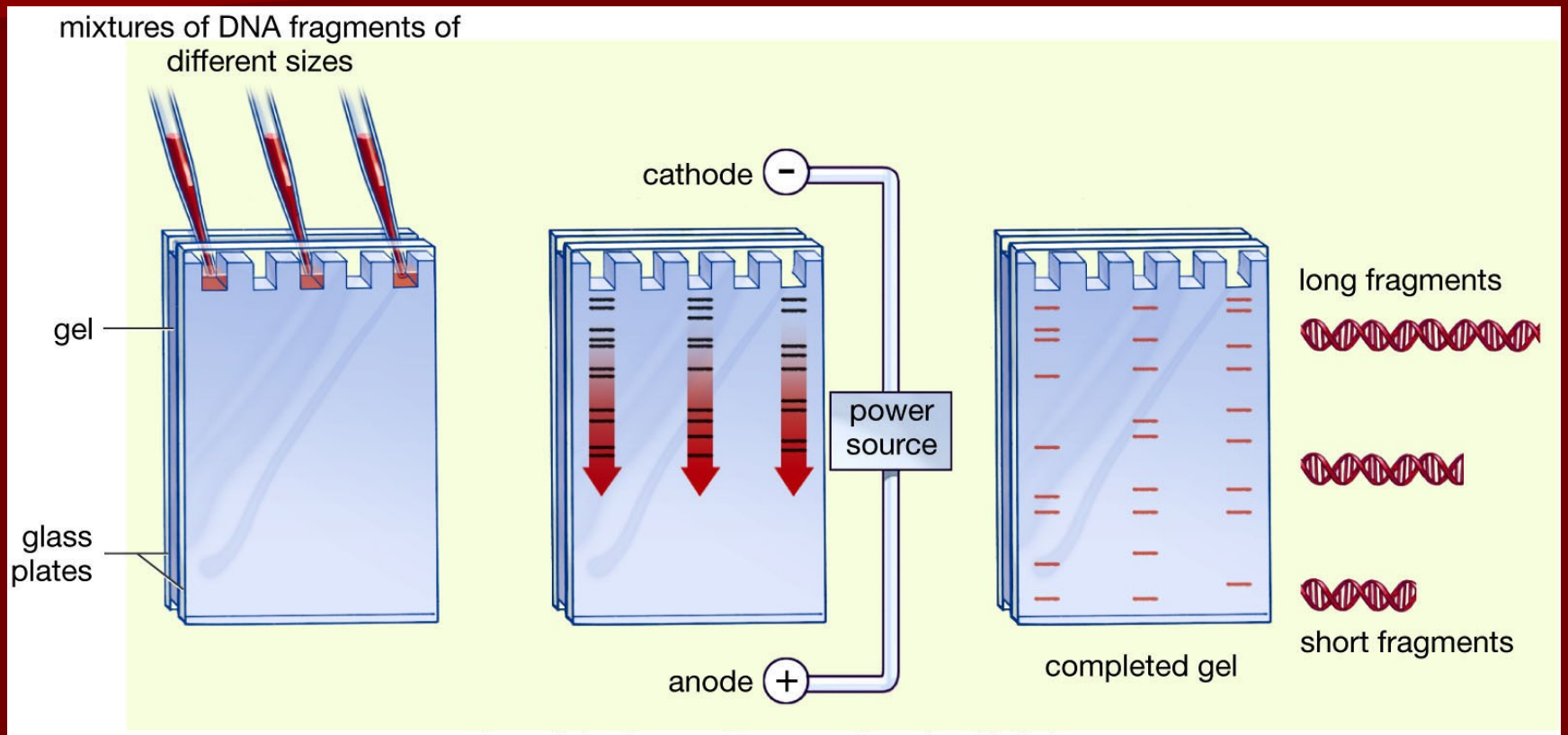


DNA Fingerprint Basics

DNA fragments of different size will be produced by a restriction enzyme that cuts at the points shown by the arrows.

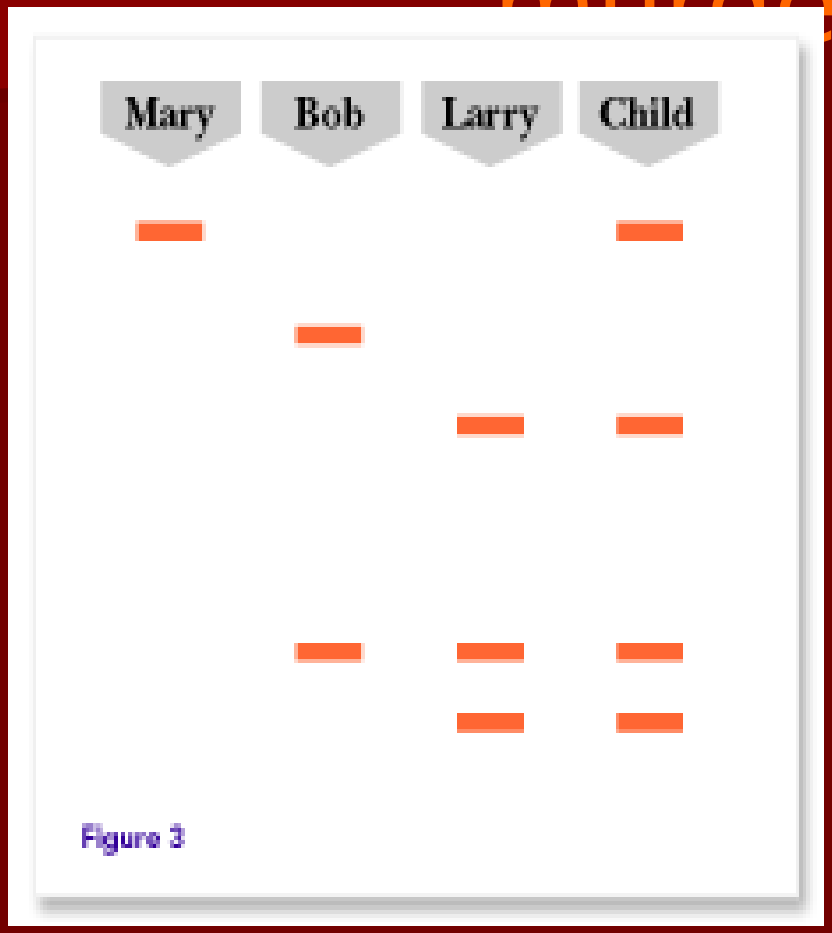


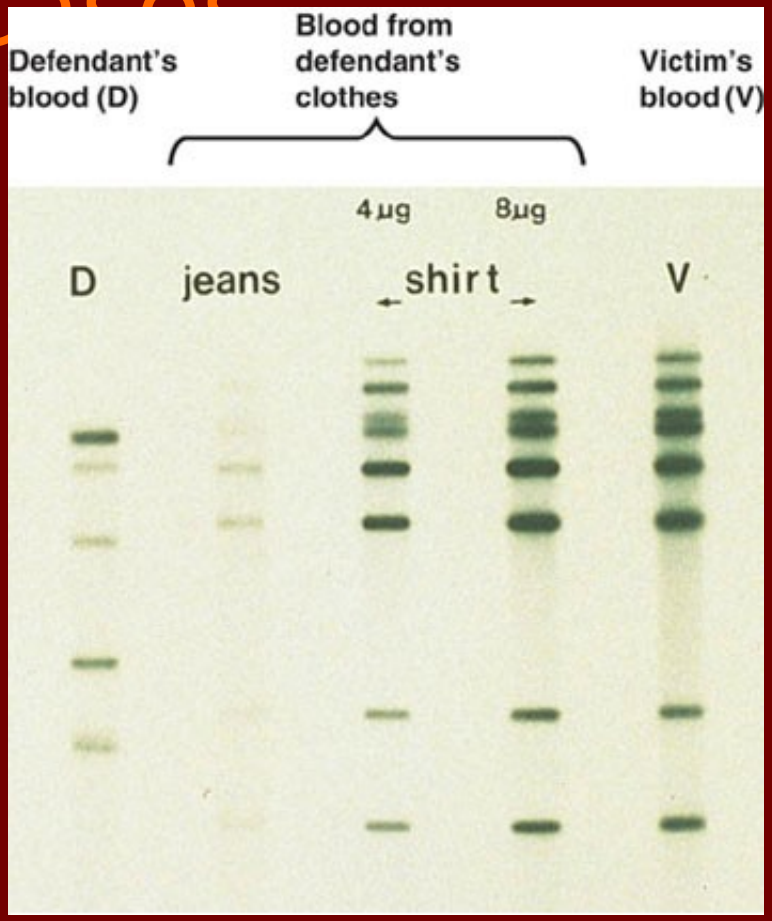
DNA fragments are then separated based on size using gel electrophoresis.



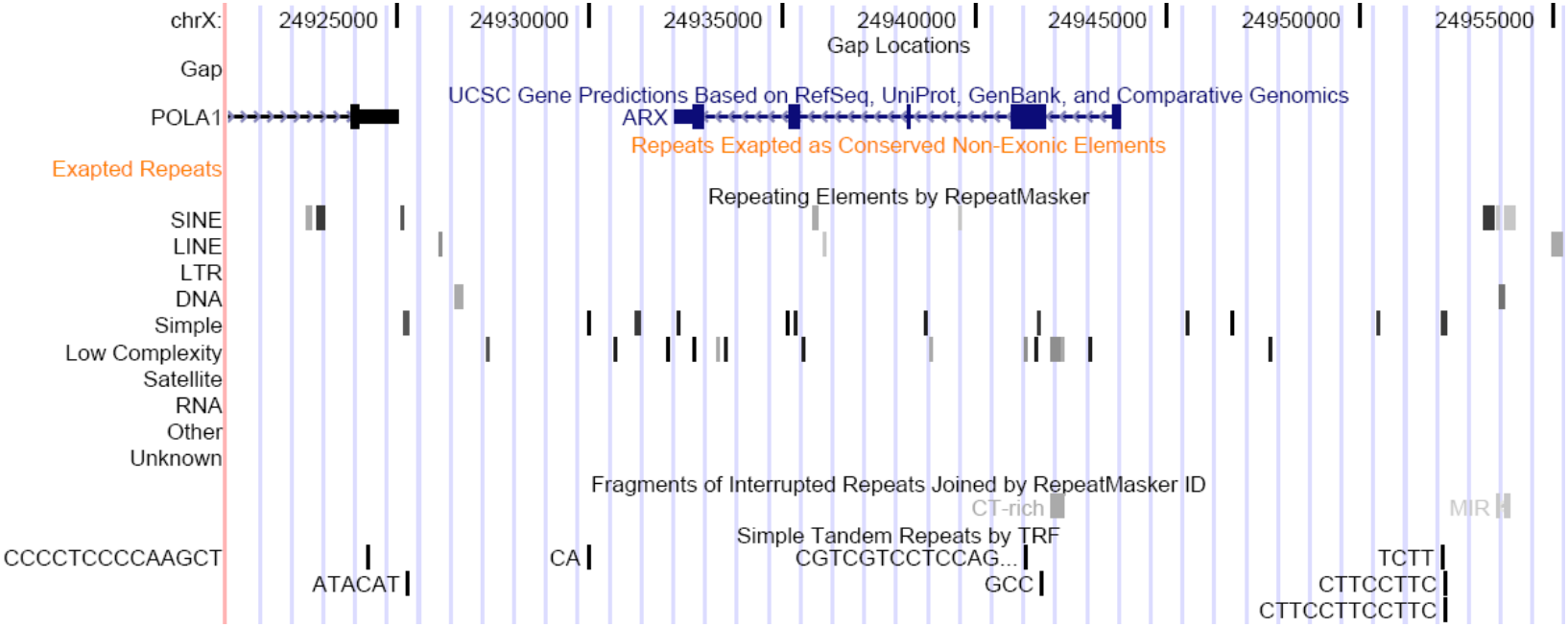
Copyright © 2005 Pearson Prentice Hall, Inc.

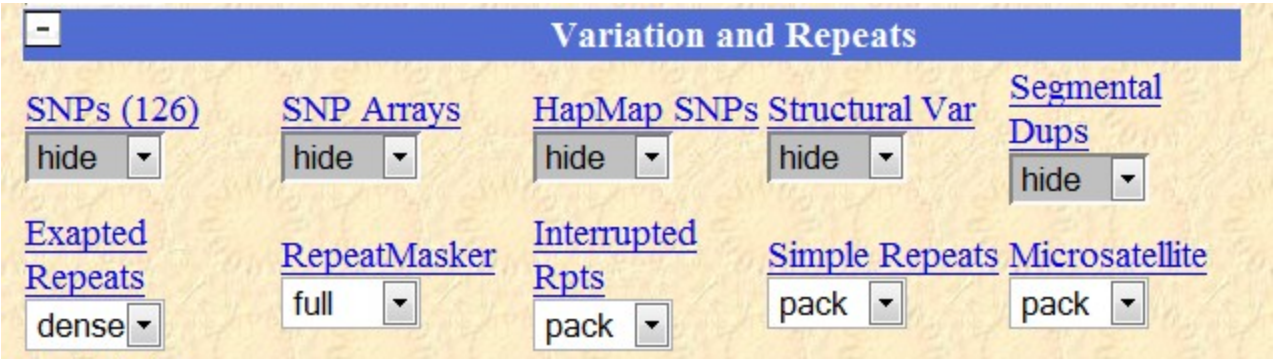
DNA Fingerprinting can be used in paternity testing or murder cases





There are Tracks for it



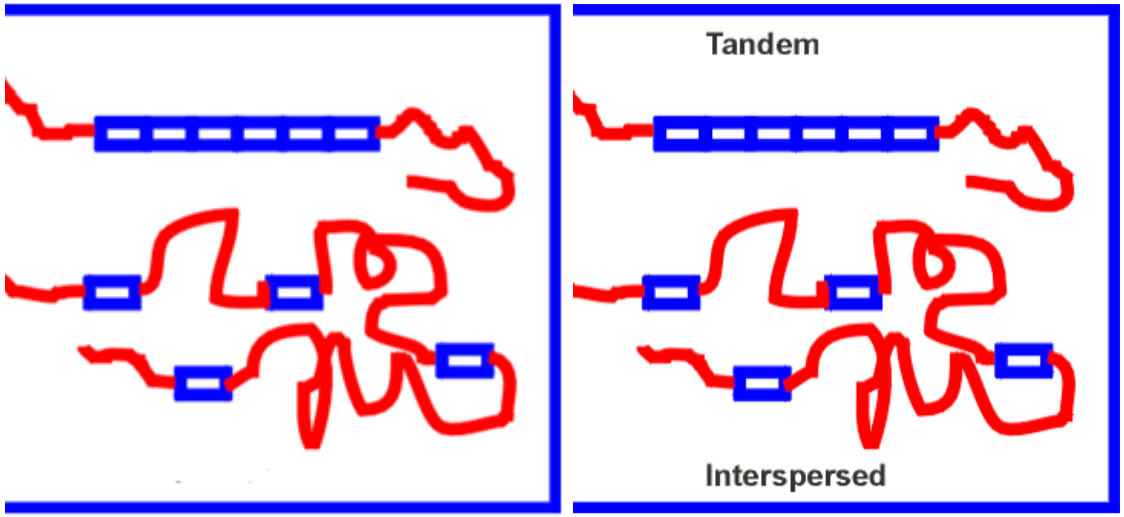


Interspersed vs. Simple Repeats

From an evolutionary point of view transposons and simple repeats are very different.

Different instances of the same transposon share common ancestry (but not necessarily a direct common progenitor).

Different instances of the same simple repeat most often do not.



Now you really know most everything

In the Genome:

Genes (up to 5% of genome)
  coding and non coding (exons, introns)

Gene regulation (15% of genome)
  proteins: transcription factors, chromatin remodelers, ...
  RNA genes: microRNAs, antisense, guide RNAs…
  DNA elements: TF binding sites, promoters, enhancers, ...

Repetitive sequences (50% of genome)
  Interspersed repeats (transposons that hop around)
  Simple repeats (local replication “sore spots”)

Categories are not mutually exclusive.

Function comes & goes with evolution = mutation + selection