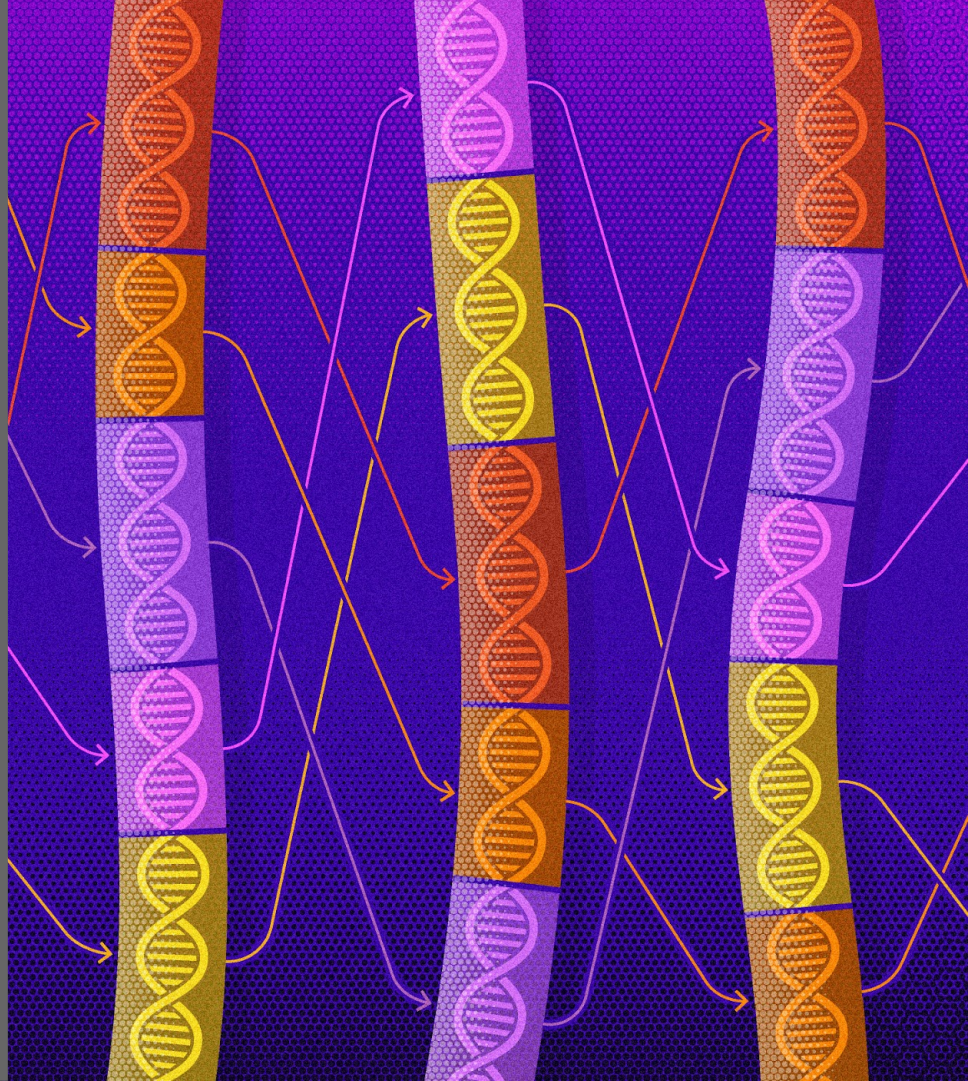
Molecular parasites and host genomes: how do transposable elements drive genome evolution?

Alba Marino - University of Montpellier (France)

alba.marino@umontpellier.fr

General Systematics - 23/07/2022

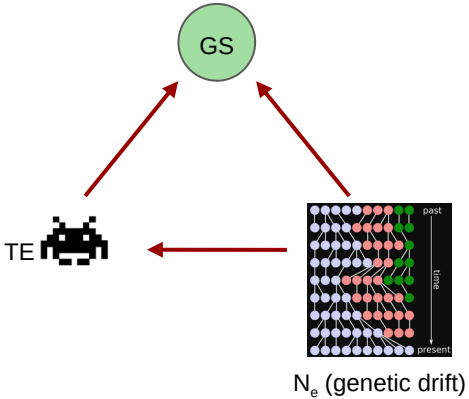


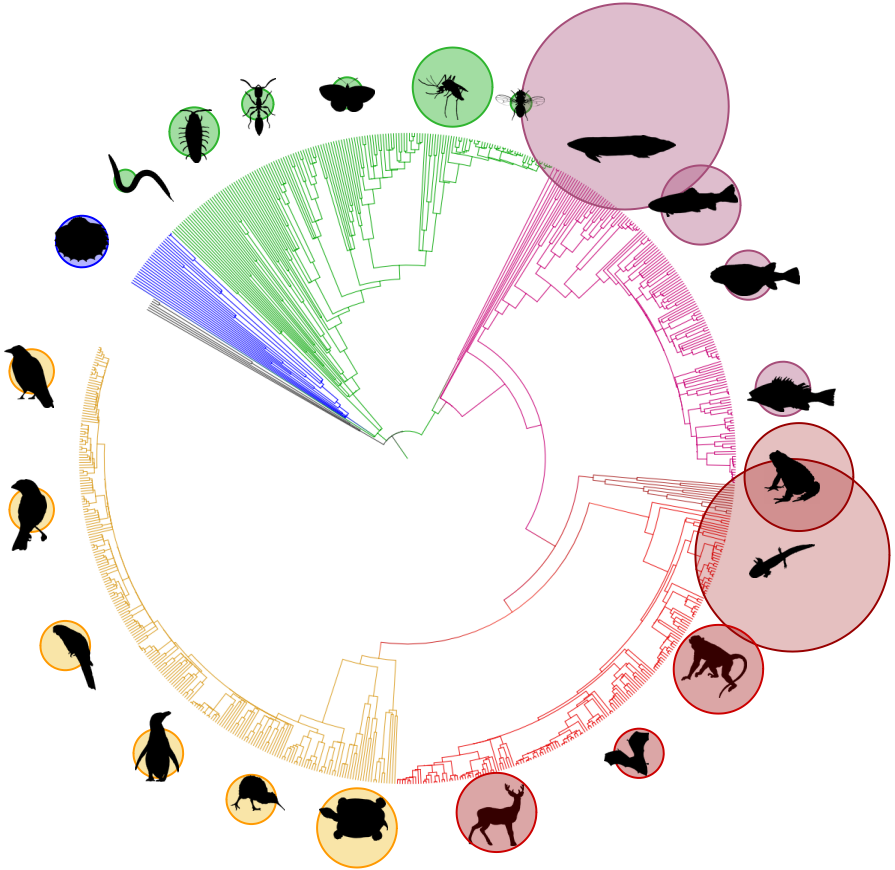
Genome size variation in animals: impact of effective population size and transposable elements

Animals show a remarkable variation in their genome sizes (GS) and one of the major contributors to this variation are transposable elements (TEs). Genetic drift has been posited to be the non-adaptive force driving the patterns of TEs expansion and ultimately those of GS variation.

Objective

Test the general validity of the hypothesis of neutral variation of GS at broader and smaller phylogenetic distance in multiple animal groups



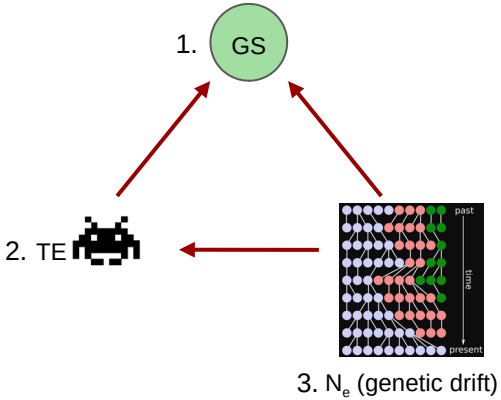


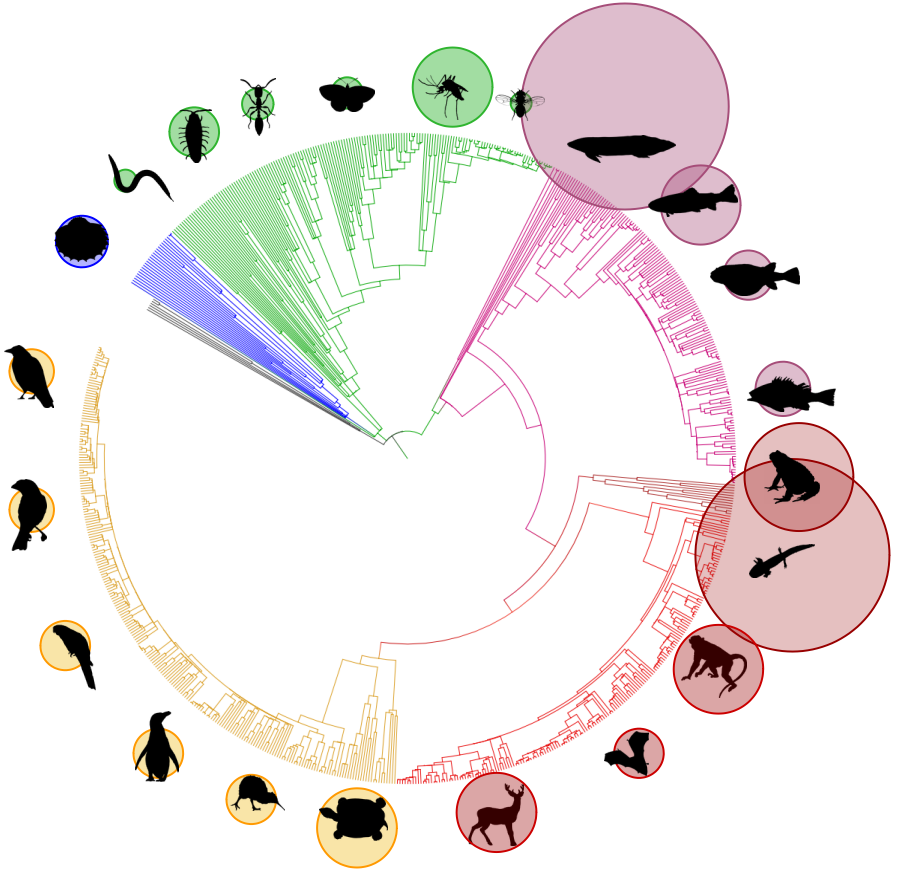
Genome size variation in animals: impact of effective population size and transposable elements

Animals show a remarkable variation in their genome sizes (GS) and one of the major contributors to this variation are transposable elements (TEs). Genetic drift has been posited to be the non-adaptive force driving the patterns of TEs expansion and ultimately those of GS variation.

Objective

Test the general validity of the hypothesis of neutral variation of GS at broader and smaller phylogenetic distance in multiple animal groups





1. Determinants of genome size

Genome size variation: an overview

Genome size is extremely variable across all the tree of life:

Prokaryotes

Carsonella ruddii - 160 Kb

Plants

Paris japonica - 150 Gb

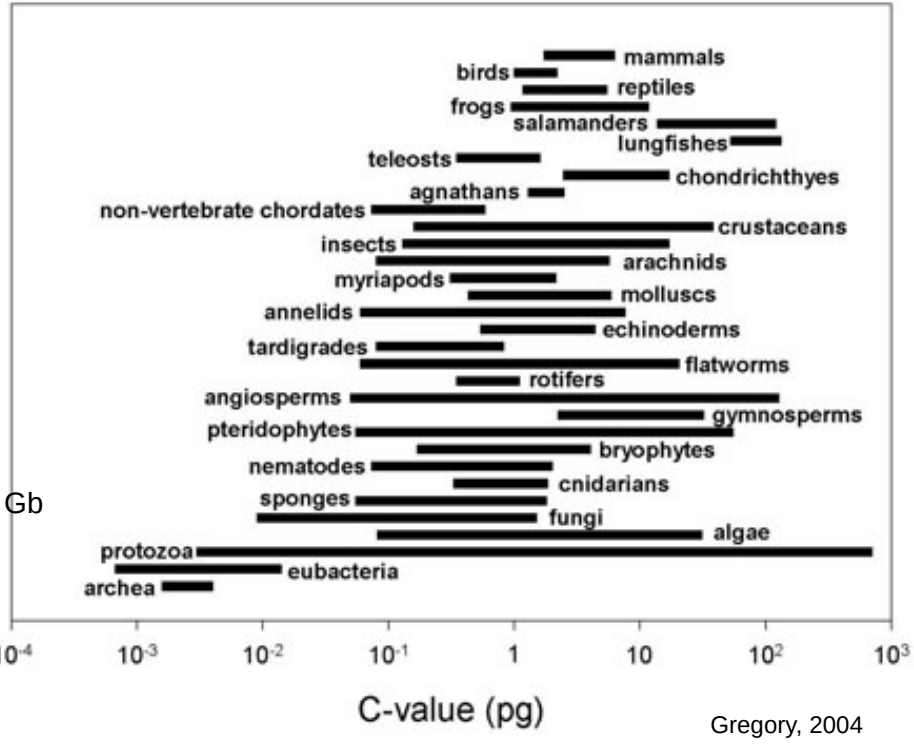
Metazoa

Pratylenchus coffeae - 20 Mb

Protopterus aethiopicus - 130 Gb

Necturus lewisi - 118 Gb

Necturus maculosus - 79 Gb

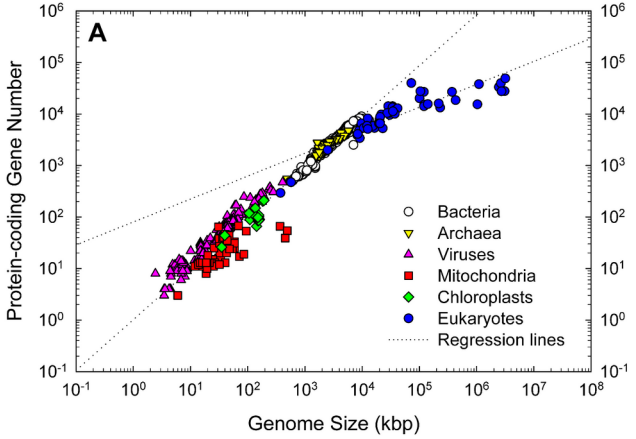


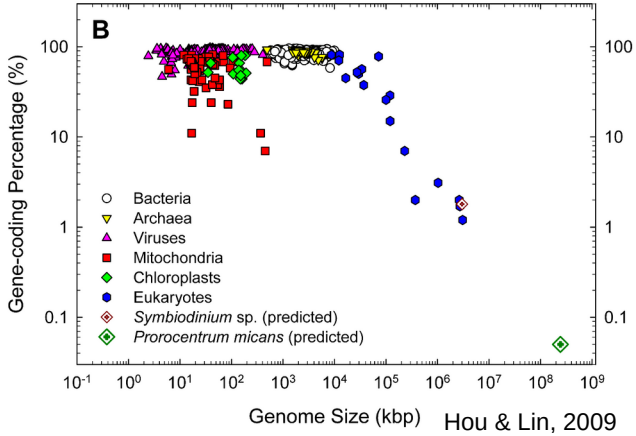
Gregory, 2004

1 pg = 978 Mb

Determinants of genome size

- In *eukaryotes* there is no correlation between GS and the apparent complexity (number of cells, number of cell types) → C-value enigma
- Gene number is not related to complexity → G-value enigma
- GS variation is not related to the quantity of gene-coding DNA or to the number of genes
- Genomes are mostly made up of non-coding DNA (ncDNA)



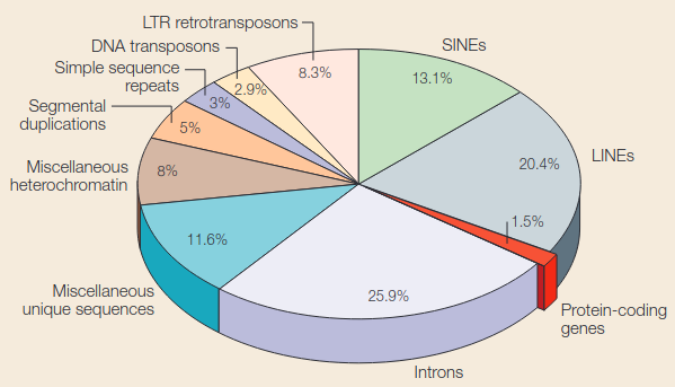


Hou & Lin, 2009

Determinants of genome size

- In *eukaryotes* there is no correlation between GS and the apparent complexity (number of cells, number of cell types) → C-value enigma
- Gene number is not related to complexity → G-value enigma
- GS variation is not related to the quantity of gene-coding DNA or to the number of genes
- Genomes are mostly made up of non-coding DNA (ncDNA)

Composition of human genome after the first whole genome sequencing



Gregory, 2005

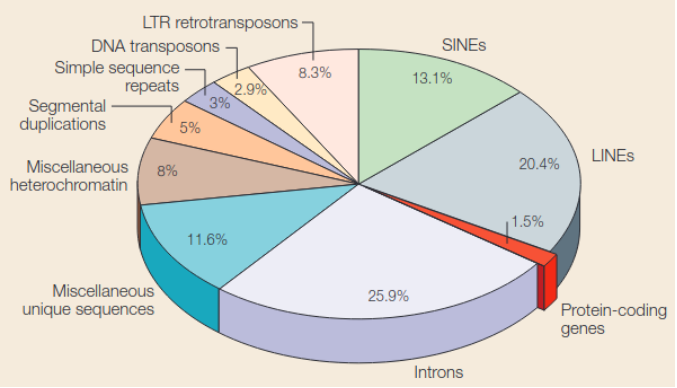
Determinants of genome size

- In *eukaryotes* there is no correlation between GS and the apparent complexity (number of cells, number of cell types) → C-value enigma
- Gene number is not related to complexity → G-value enigma
- GS variation is not related to the quantity of gene-coding DNA or to the number of genes
- Genomes are mostly made up of non-coding DNA (ncDNA)

Why is ncDNA accumulated and how is it maintained in the genome?

1) Selectionist hypotheses
2) Nucleotypic hypotheses
3) Neutralist hypotheses

Composition of human genome after the first whole genome sequencing

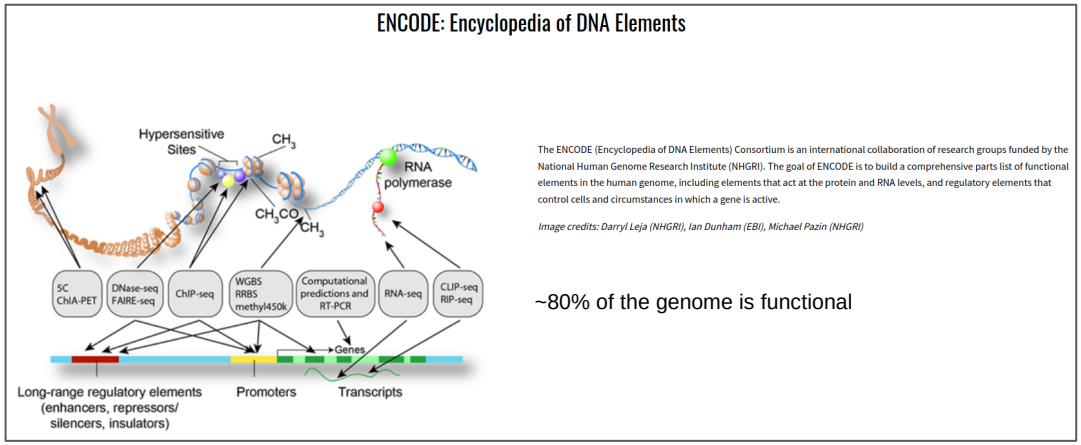


Gregory, 2005

Determinants of genome size

1) Selectionist hypothesis:

Most part of ncDNA is “literal” and functional, meaning that it should be undergoing negative selection



~80% of the genome is functional

But…

Use of “causal role”, instead of “selected effect function”

Estimates of conserved sequences are ~ 5-10% of the human genome (Ward & Kellis, 2012)

Mutational load

Determinants of genome size

2) **Nucleotypic hypothesis:**

ncDNA is mainly “indifferent” DNA and genome size is indirectly under selection because of its nucleotypic effects → nuclear volume, cellular volume, cell division time, metabolic rate, generation time, development time

Nucleoskeletal hypothesis:

ncDNA is mainly “indifferent” DNA and genome size is indirectly selected because of its effect on cell size → optimal ratio between nuclear volume and cytoplasmic volume

Examples of such correlations support these hypotheses, anyway it is hard to prove causation and its direction

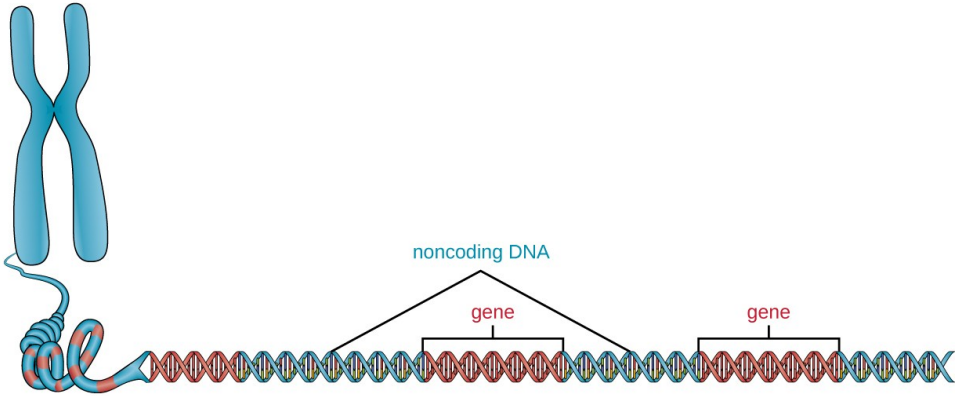
Correlation between genome size and cell size are observed even in prokaryotes which do not have nucleus

The same correlations could be explained either with population genetics principles (neutralist Mutational Hazard hypothesis)

Determinants of genome size

3) **Neutralist hypothesis:**

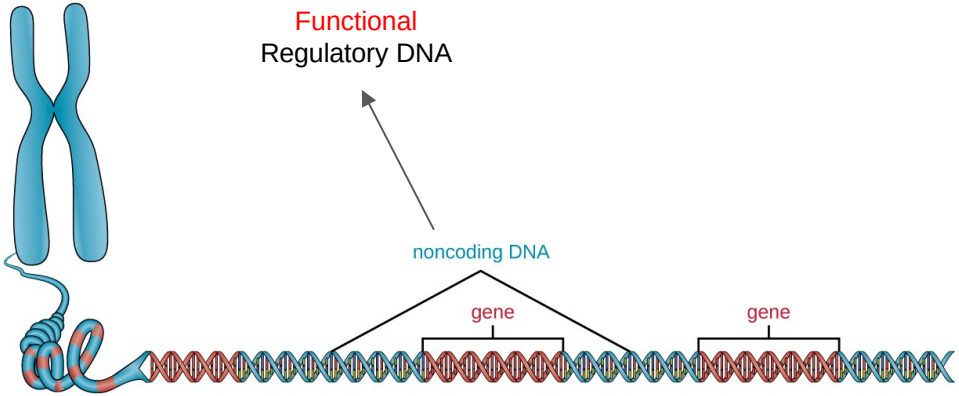
The ncDNA that makes up the bulk of genomes is mostly *nonfunctional*, or “junk”: it does not contribute to the fitness of the organism and its accumulation is generally neutral



Genome size and junk

3) **Neutralist hypothesis:**

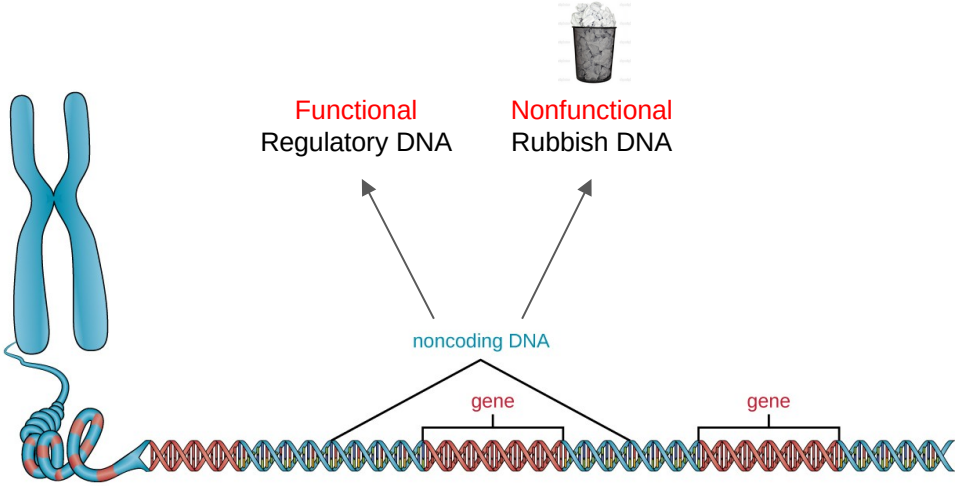
The ncDNA that makes up the bulk of genomes is mostly *nonfunctional*, or “junk”: it does not contribute to the fitness of the organism and its accumulation is generally neutral



Genome size and junk

3) **Neutralist hypothesis:**

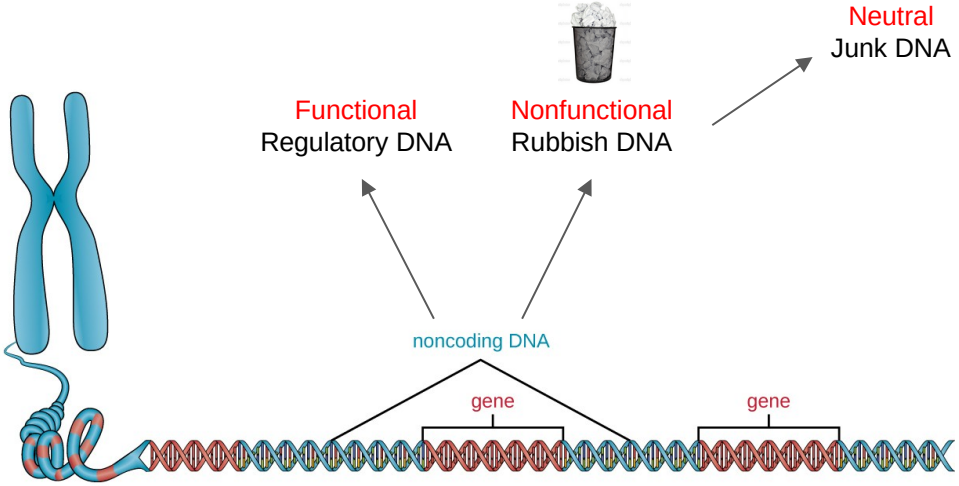
The ncDNA that makes up the bulk of genomes is mostly *nonfunctional*, or “junk”: it does not contribute to the fitness of the organism and its accumulation is generally neutral



Genome size and junk

3) **Neutralist hypothesis:**

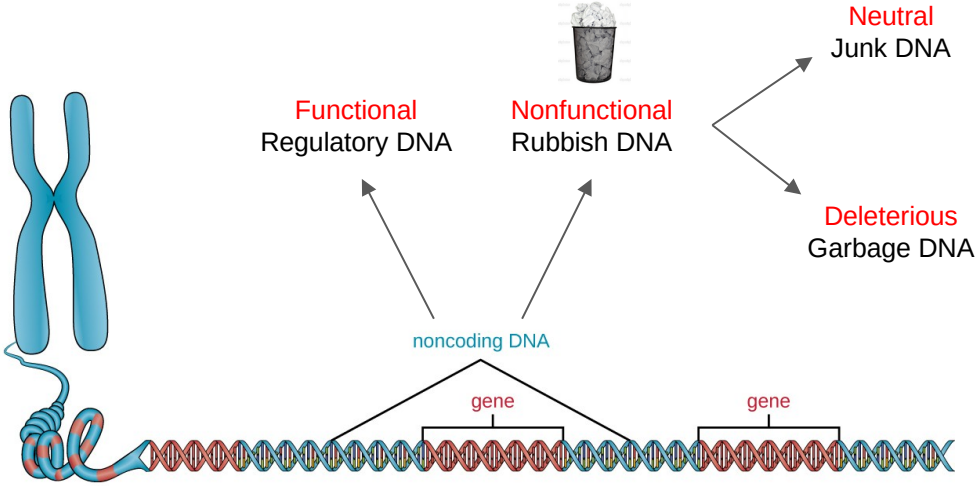
The ncDNA that makes up the bulk of genomes is mostly *nonfunctional*, or “junk”: it does not contribute to the fitness of the organism and its accumulation is generally neutral



Genome size and junk

3) **Neutralist hypothesis:**

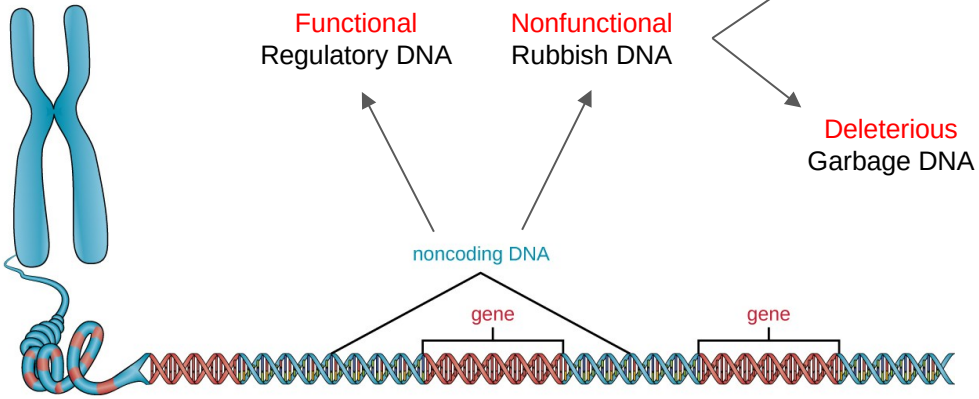
The ncDNA that makes up the bulk of genomes is mostly *nonfunctional*, or “junk”: it does not contribute to the fitness of the organism and its accumulation is generally neutral



Genome size and junk

3) **Neutralist hypothesis:**

The ncDNA that makes up the bulk of genomes is mostly *nonfunctional*, or “junk”: it does not contribute to the fitness of the organism and its accumulation is generally neutral

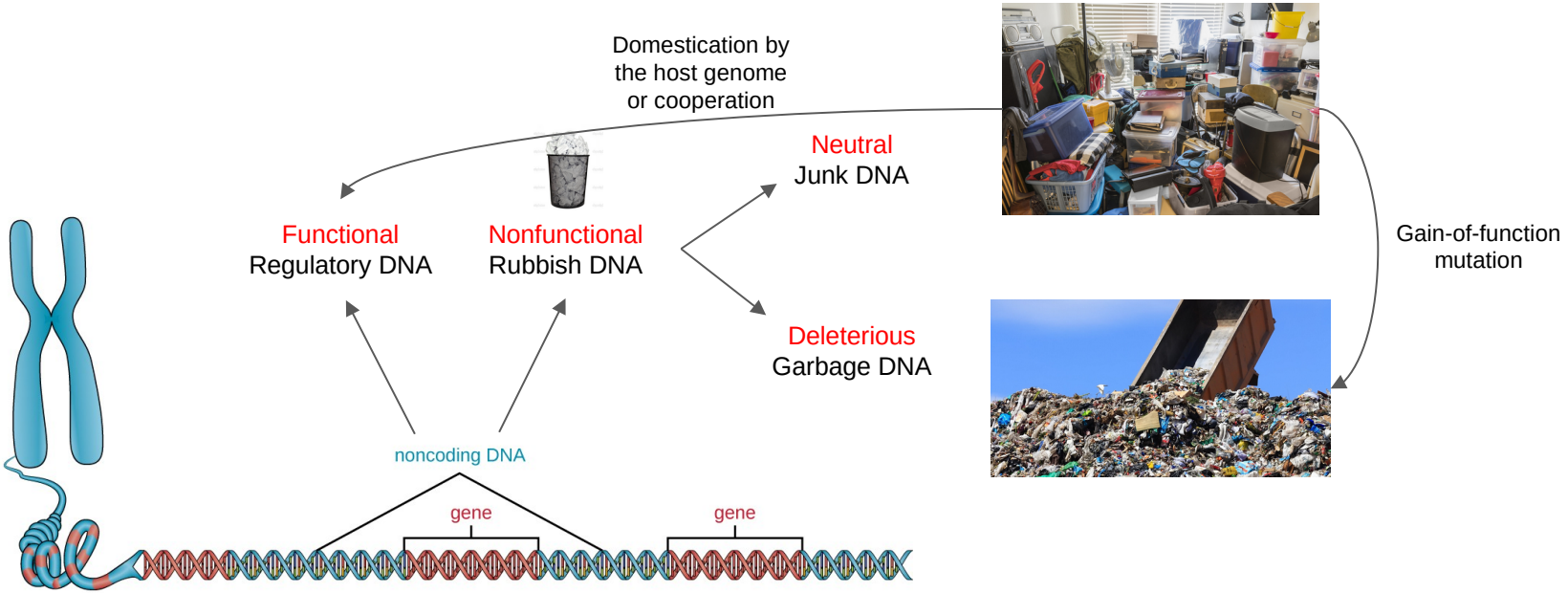


Gain-of-function mutation

Genome size and junk

3) **Neutralist hypothesis:**

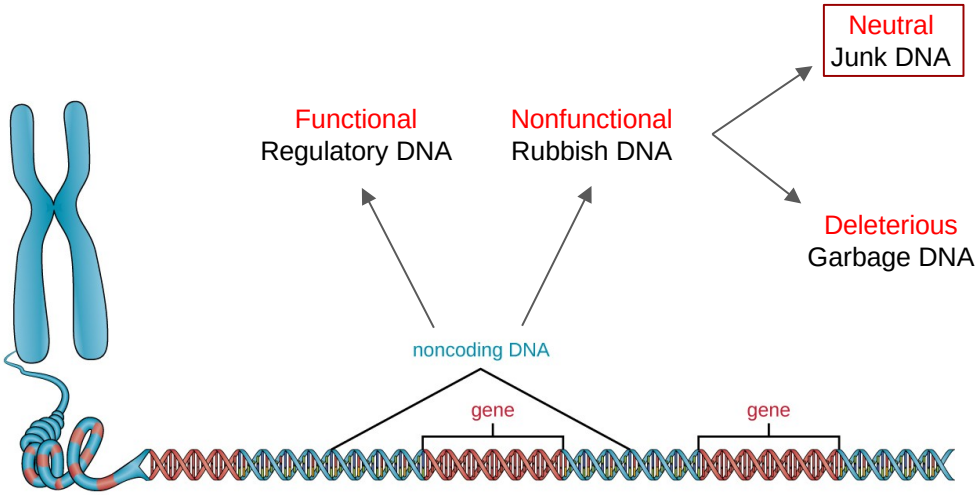
The ncDNA that makes up the bulk of genomes is mostly *nonfunctional*, or “junk”: it does not contribute to the fitness of the organism and its accumulation is generally neutral



Genome size and junk

3) **Neutralist hypothesis:**

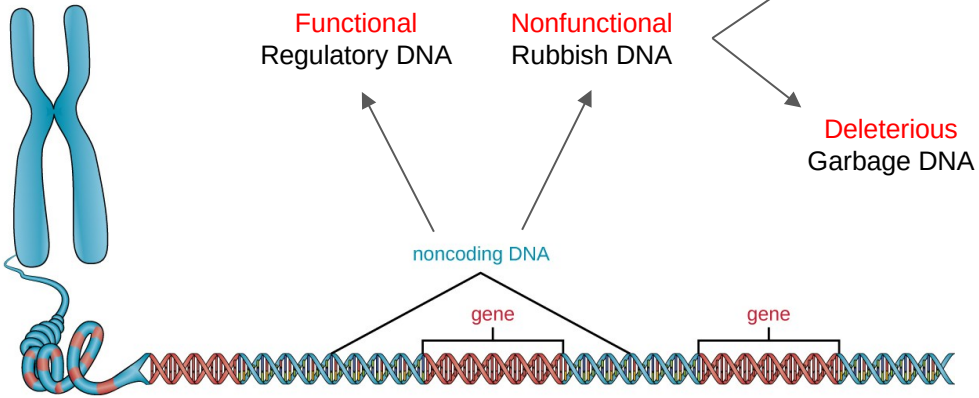
The ncDNA that makes up the bulk of genomes is mostly *nonfunctional*, or “junk”: it does not contribute to the fitness of the organism and its accumulation is generally neutral



Genome size and junk

3) **Neutralist hypothesis:**

The ncDNA that makes up the bulk of genomes is mostly *nonfunctional*, or “junk”: it does not contribute to the fitness of the organism and its accumulation is generally neutral

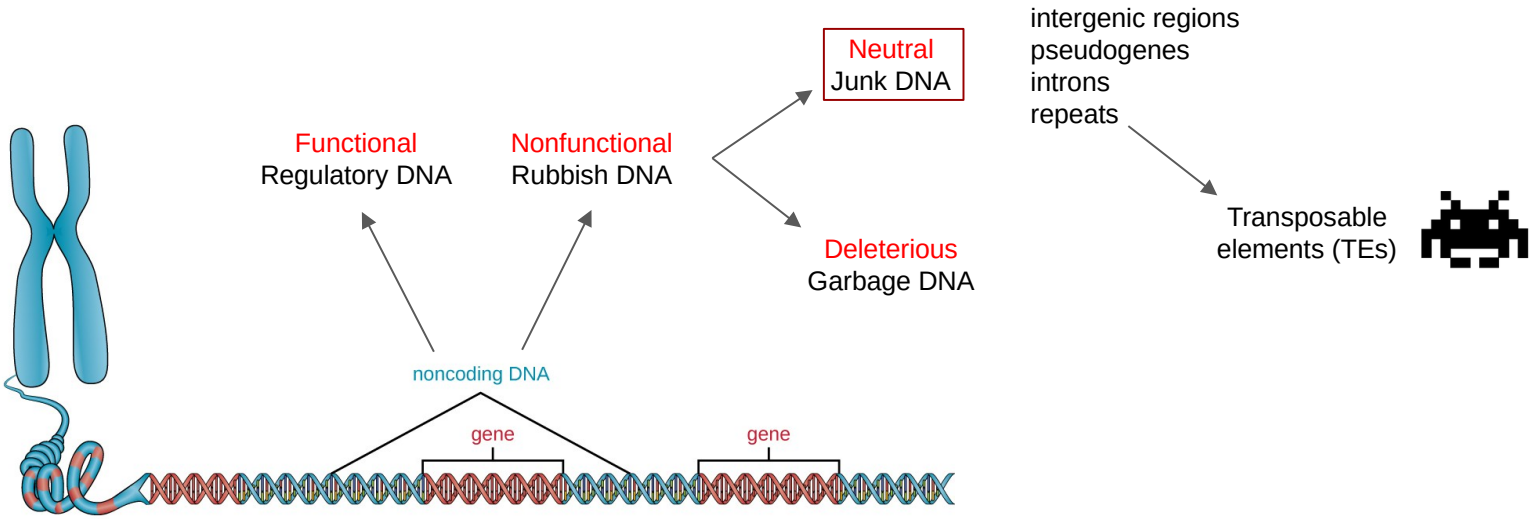


intergenic regions
pseudogenes
introns
repeats

Genome size and junk

3) **Neutralist hypothesis:**

The ncDNA that makes up the bulk of genomes is mostly *nonfunctional*, or “junk”: it does not contribute to the fitness of the organism and its accumulation is generally neutral



2. Transposable elements and their impact on genome architecture

Transposable elements: main features

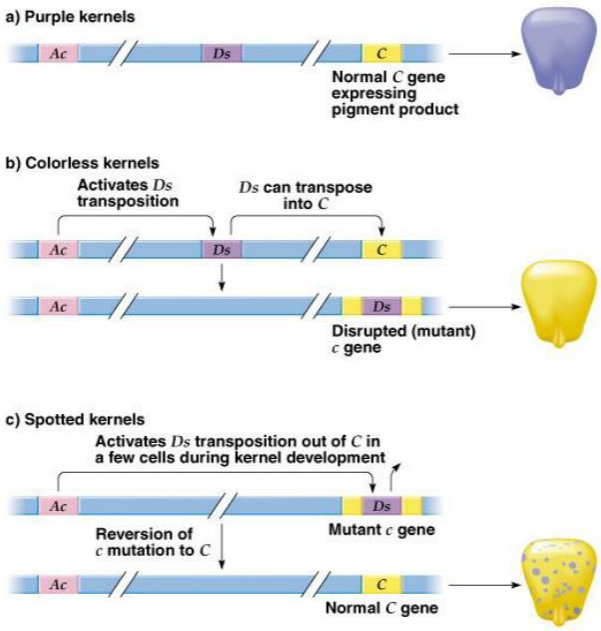
Discovery by Barbara McClintock between the ‘40s and ‘50s in *Zea mays* where she studied the genetic basis of the kernel unstable mosaicism.

Dissociation (Ds, non-autonomous element) was the responsible of a chromosome break occurring always in the same point

Ds would not cause the mutation without *Activator* (Ac, autonomous element) being present

The transposition of Ac and Ds would happen patchily in the cells, causing the color mosaicism of maize kernels

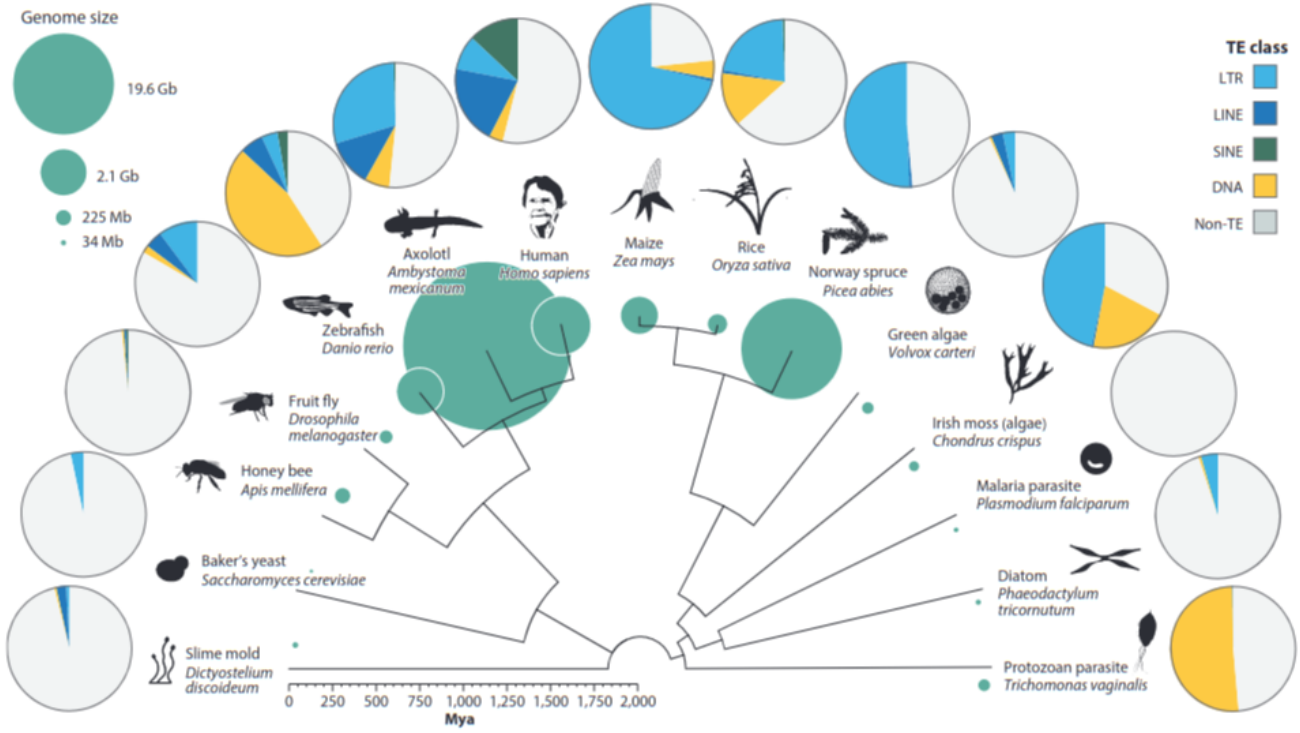
Fig. 7.24, Transposon effects on corn kernel color.



Russell, 2014

Transposable elements: main features

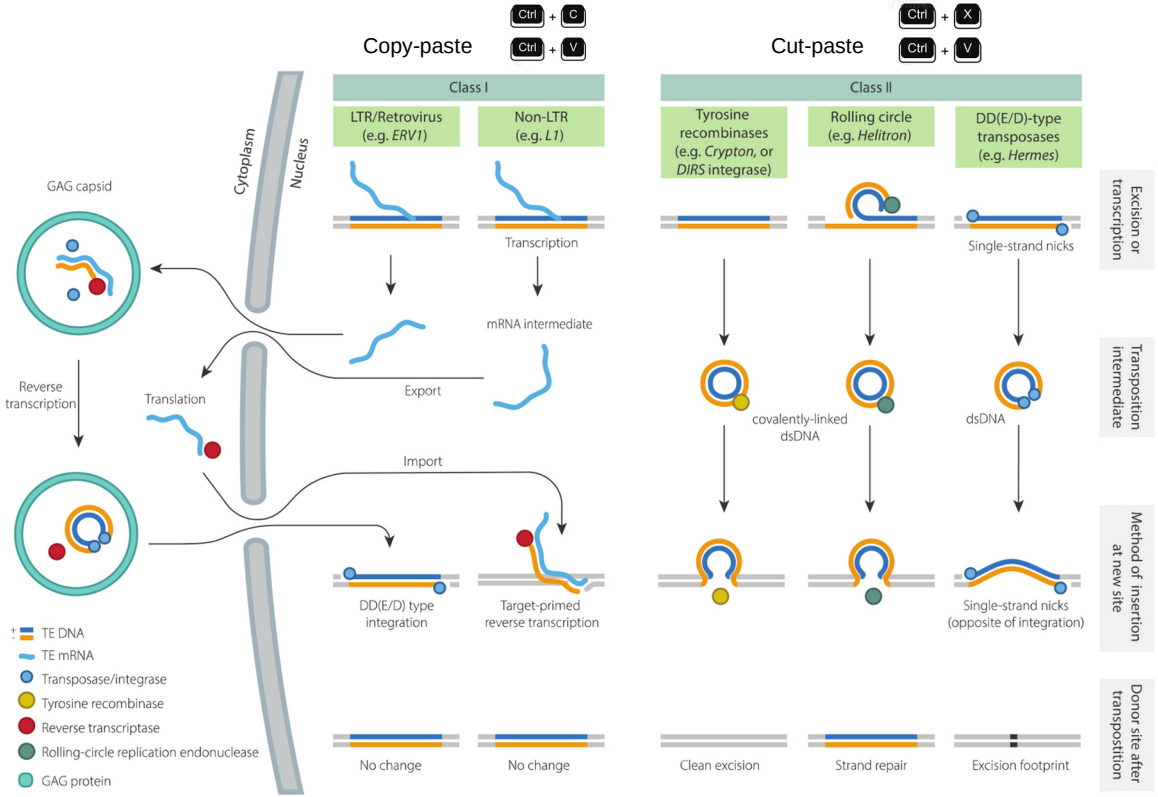
- Genetic sequences able to self-replicate and spread within the genome that they colonize
- Ubiquitous in all living organisms, prokaryotes and eukaryotes
- From few hundreds of bp to tens of thousands of bp



Wells & Feschotte, 2020

Transposable elements: main features

- Genetic sequences able to self-replicate and spread within the genome that they colonize
- Ubiquitous in all living organisms, prokaryotes and eukaryotes
- From few hundreds of bp to tens of thousands of bp



Wells & Feschotte, 2020

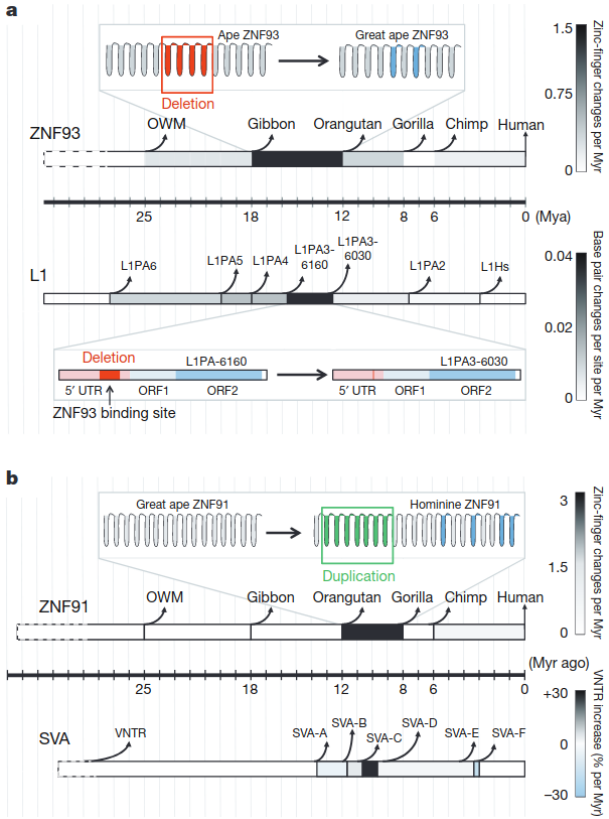
Transposable elements: from selfish to fuel of genome evolution

- *Jumping genes*, *molecular parasites*, *selfish DNA* → TEs proliferate and accumulate in the genome, causing its expansion and mostly not contributing to the fitness of the organism or being deleterious for their host

- **Selfish DNA Hypothesis**
  The TE landscape in a genome is the result of contrasting selection processes on two different levels: at the **genome level,** the most efficient replicators are positively selected in the "population" of selfish elements; at the **organism level**, negative selection acts on TEs bearing deleterious impact on the host fitness

- Arms race and genomic conflicts: TEs can evolve to overcome the host defences, and the host genome can undergo selection for new mechanisms to restrain TEs expansion → host-parasite co-evolution

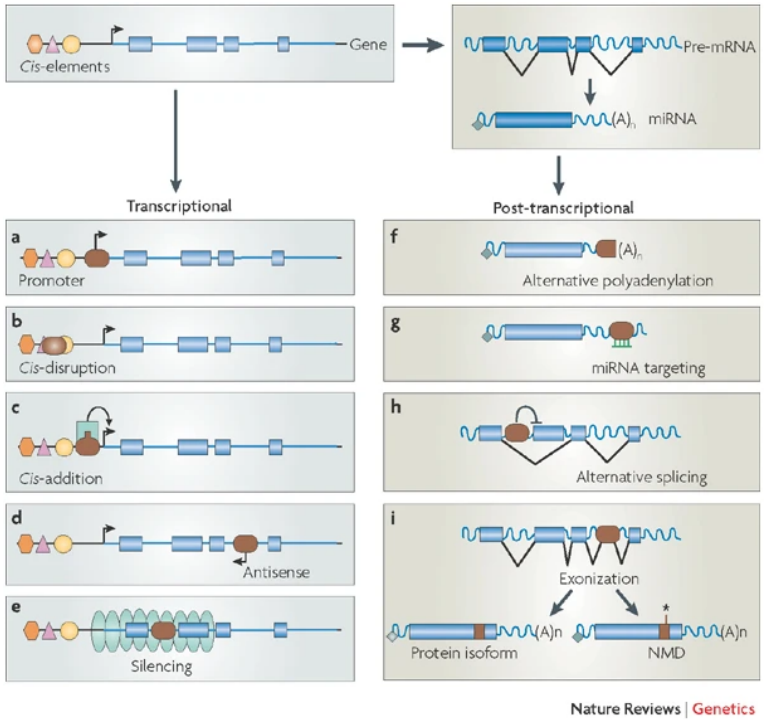
- Example: co-evolution of KRAB-ZNF sequences and SVA and L1 TE families in the primate lineage

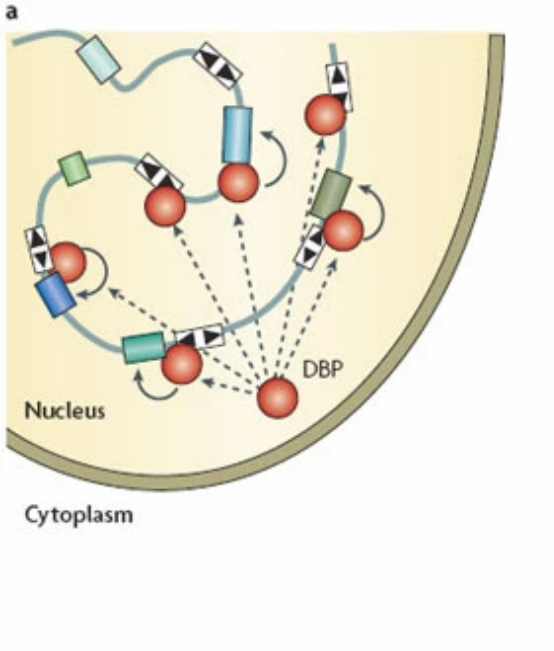


Jacobs et al., 2014

Transposable elements: from selfish to fuel of genome evolution

The activity of TEs is an important mutagenic source: they can affect **gene expression**, rewire existing **regulatory networks** and provoke genomic rearrangements (e.g. Non Allelic Homologous Recombination)



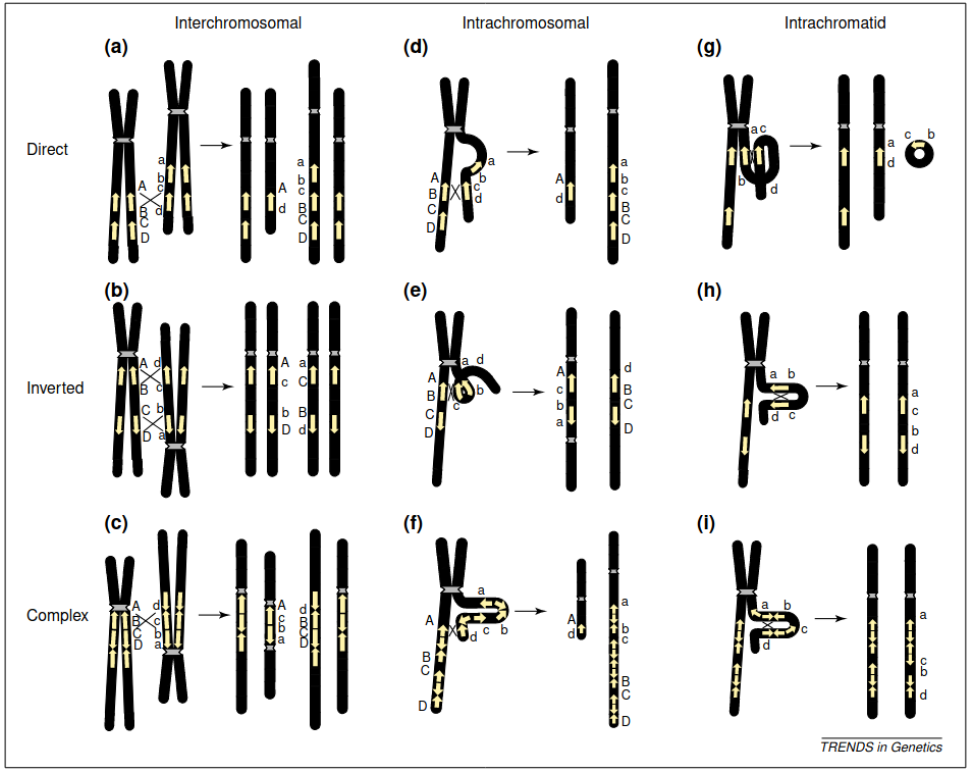


DBP = DNA Binding Protein
RBP = RNA Binding Protein

Feschotte, 2008

Transposable elements: from selfish to fuel of genome evolution

The activity of TEs is an important mutagenic source: they can affect gene expression, rewire existing regulatory networks and provoke **genomic rearrangements** (e.g. Non Allelic Homologous Recombination)



Stankiewicz & Lupski, 2002

Transposable elements: from selfish to fuel of genome evolution

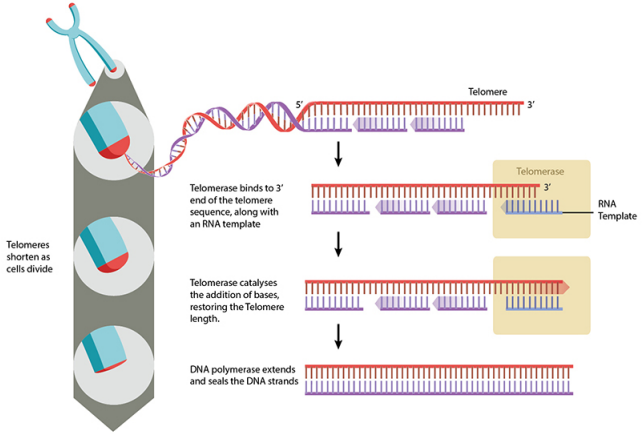
Although less frequent, TEs can be co-opted as genes or regulatory sequences (promoters, enhancers, etc.), give rise to new genes, and be integrated in the regulatory networks of the host

Protection of the telomeres

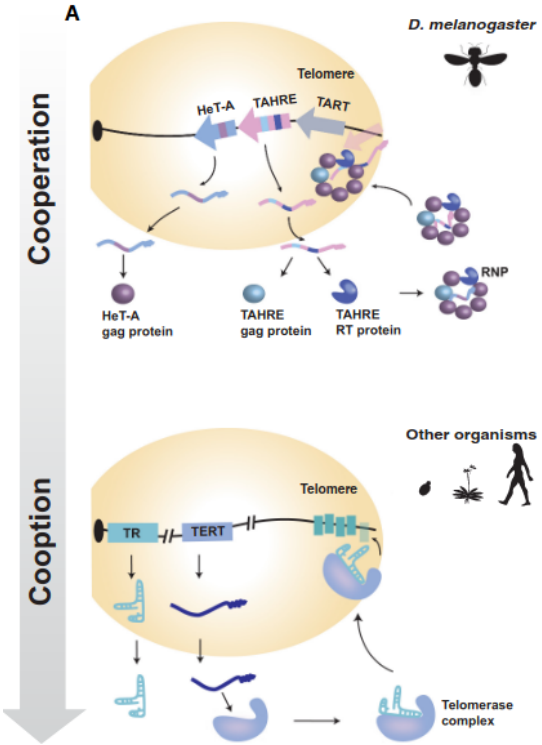
At each duplication round the last fragment of the chromosome is lost. Telomeres are terminal portions of chromosomes made of highly repetitive DNA protecting from the loss of important genetic information.
To restore telomeres most organisms use the telomerase, a retrotranscriptase that supplies a RNA template allowing for the extension of the terminal part with a repeated sequence.

Diptera lost telomerase and rely on the activity of 3 families of “telomeric retroelements” that reconstitute the telomeres by retrotransposing right in the telomeric regions.

Even the retrotranscriptase component of the telomerase is thought to have originated from a TE



Cosby et al., 2019



Transposable elements: from selfish to fuel of genome evolution

Although less frequent, TEs can be co-opted as genes or regulatory sequences (promoters, enhancers, etc.), give rise to new genes, and be integrated in the regulatory networks of the host

Birth of new transcription factors

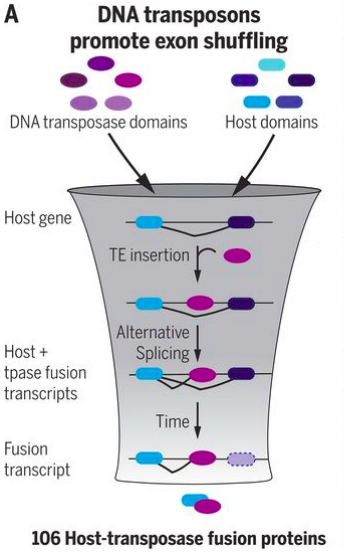
106 genes derived by 94 distinct events of fusion between a host and a transposase domain (HTFs) were identified in tetrapods.

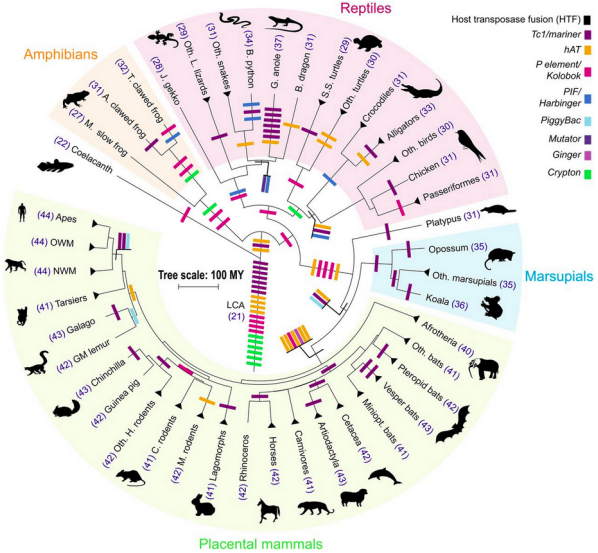
After insertion of the TE within the host gene sequence, a splice site provided by the TE could induce alternative splicing and the fusion of the host and the transposase domains into a new transcript.

All analyzed HTFs were found under purifying selection

KRAB-transposase domains (sequence-specific repressive factors)

Pax genes (metazoan-conserved developmental transcription factors)





Cosby et al., 2021

Transposable elements: from selfish to fuel of genome evolution

TEs' nature is originally selfish as any of their components is related to their proliferating activity, regardless of its effect on the genome environment where they thrive
Their pervasiveness inevitably causes interactions of different nature between their sequences and those of the genome:

Neutral evolution → TEs duplicate and colonize genomes as long as they do not impact the host fitness

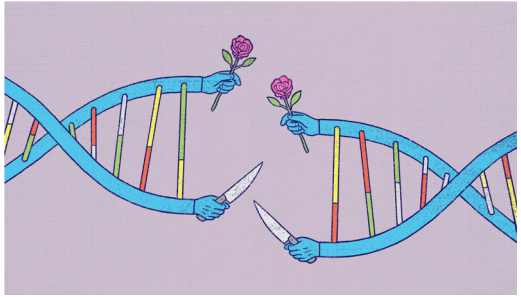
Genomic conflict → New defence and invasion mechanisms are evolved through time

Cooperation → A mutualism between favouring both mobile elements and host biology is found

Cooption → TEs, initially there because of their proliferative function, are "spandrels" supplying raw material that can be exapted for new functions useful for the host

TEs dynamics shape the genomic architecture and the phenotypes of virtually all organisms, affecting their evolutionary trajectories

The Spandrels of San Marco and the Panglossian Paradigm (Gould & Lewontin, 1979)



3. Effective population size and genetic drift: the importance of non-adaptive forces

What’s effective population size?

The effective population size is the number of individuals that actually contribute to the genetic variability of the population and measures the force of genetic drift

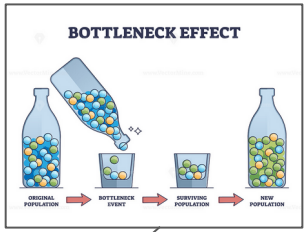
In an ideal population, $N_e = N_c$

Effective population size — Census population size

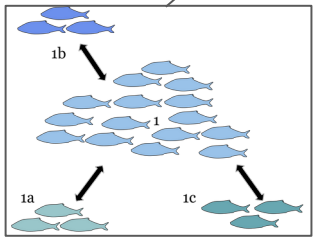
- Finite diploid population of constant size N
- Sex ratio 50:50 with all individuals able to reproduct
- Absence of selection: any fitness variation between individuals is due to chance
- No overlapping generations
- Random mating
- Constant number of reproducing individuals at each generation

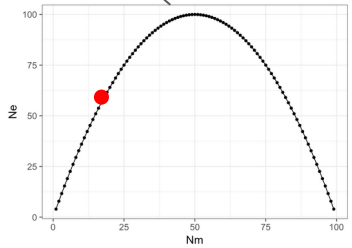
In real populations, deviation from these conditions will always cause $N_e < N_c$:

- Abrupt variations of population size like bottlenecks or founder effects
- Inbreeding
- Biased sex ratio
- Population structured in semi-isolated subpopulations



↓ N_e



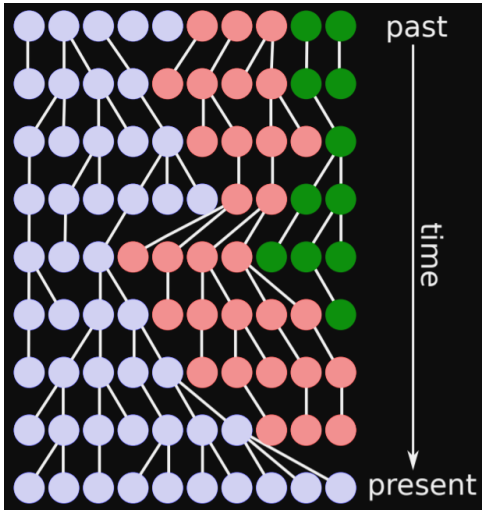


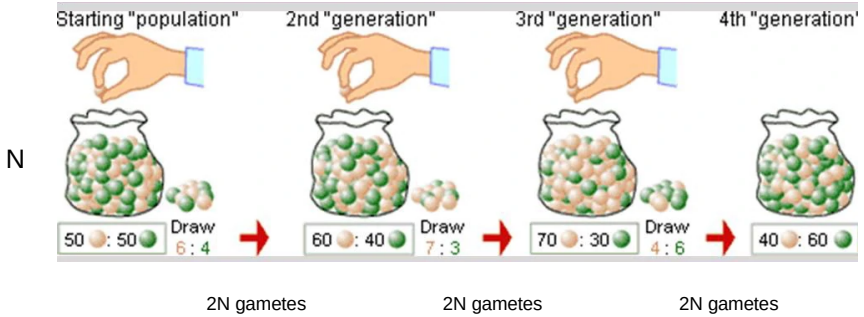
Genetic drift: why does effective population size matter?

Genetic drift is the change of allelic frequencies through generations due to stochastic effects

Non-adaptive evolutionary force

The effects of drift through time are similar to those of sampling error: the higher the number of samples, the better is the representation of the original variability of the population at each generation; a small sample will cause the variability to shift away more quickly from the original condition





Genetic drift: why does effective population size matter?

Genetic drift is the change of allelic frequencies through generations due to stochastic effects

Non-adaptive evolutionary force

The effects of drift through time are similar to those of sampling error: the higher the number of samples, the better is the representation of the original variability of the population at each generation; a small sample will cause the variability to drift away more quickly from the original condition. The power of genetic drift is proportional to $1 / N_e$

The *population size* determines how quickly neutral alleles are fixed or lost in the population as an effect of *chance*

$P = 1 / 2N_e$

Initial fixation probability of a neutral mutation

$K = 2N_e\mu (1 / 2N_e) \rightarrow K = \mu$

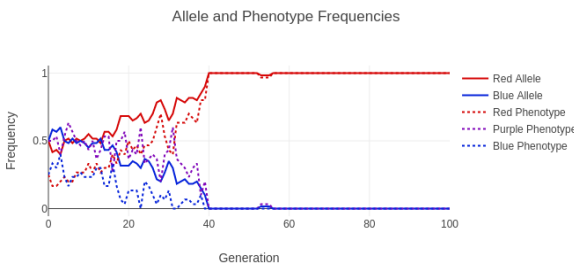
Rate of fixation of a neutral mutation

Mutation rate

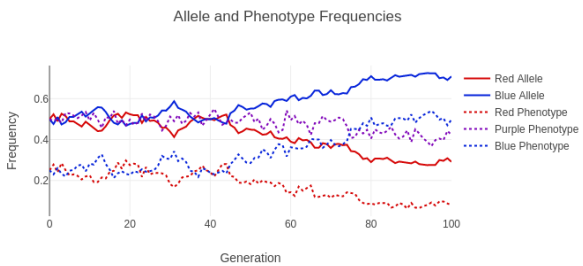
$t = 4N_e$ generations

Average time for a neutral mutation to reach fixation

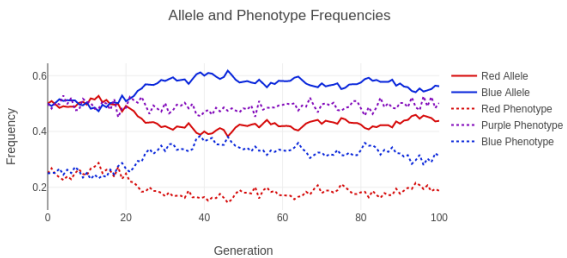
N = 30



N = 400



N = 1000



https://www.biologysimulations.com/population-genetics

Genetic drift: why does effective population size matter?

Nearly-neutral hypothesis

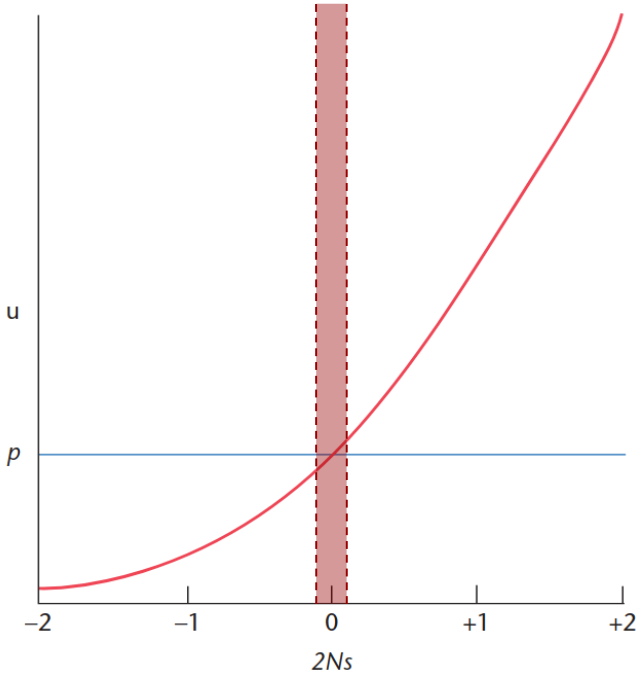
In the immediate range of neutral mutations (s = 0), there is a fraction of slightly advantageous or slightly deleterious mutations that will evolve as if they had no selection effect → *effective neutrality*

$| 4 N_e s | << 1$ → allele behaves as a neutral one and evolves under genetic drift

$| 4 N_e s | >> 1$ → allele behaves as a selected one and evolves under selection

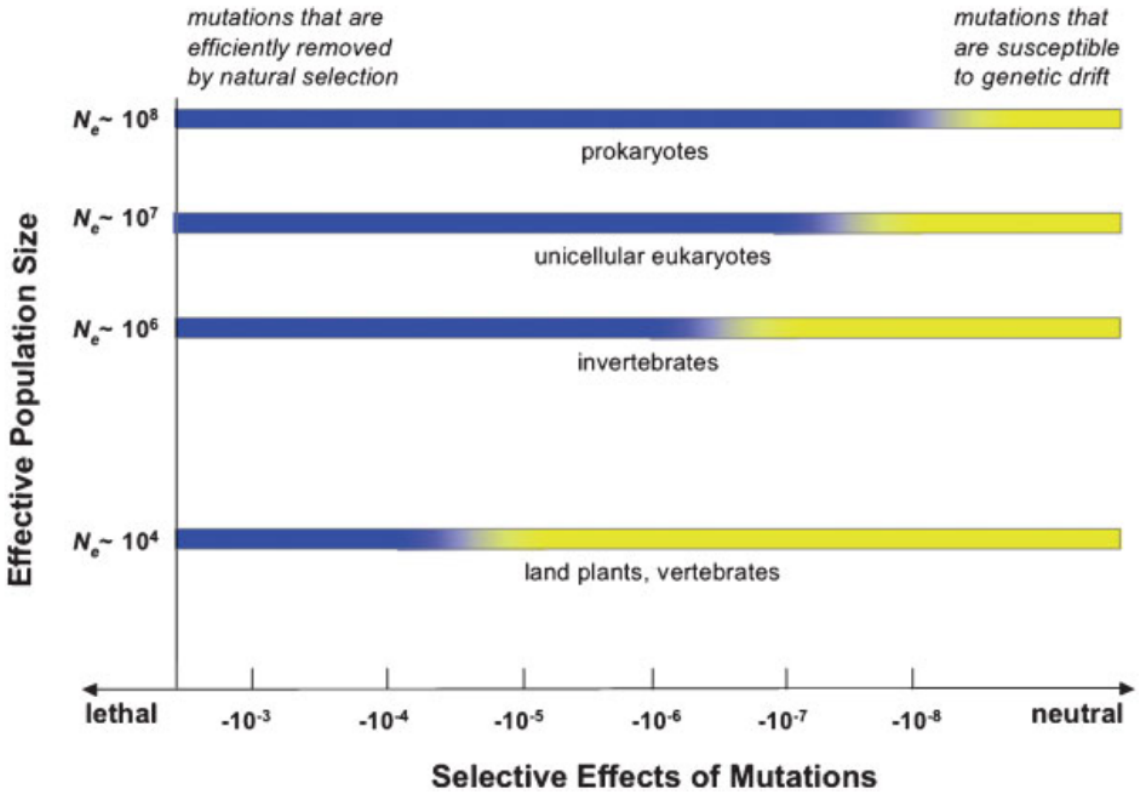
The efficacy with which deleterious mutations are eliminated or promote advantageous ones depends on the effective population size.

A slightly deleterious mutation with the same selective effect can be maintained or removed from a population depending on its effective population size, i.e. depending on the relative strength of selection and genetic drift acting in the population



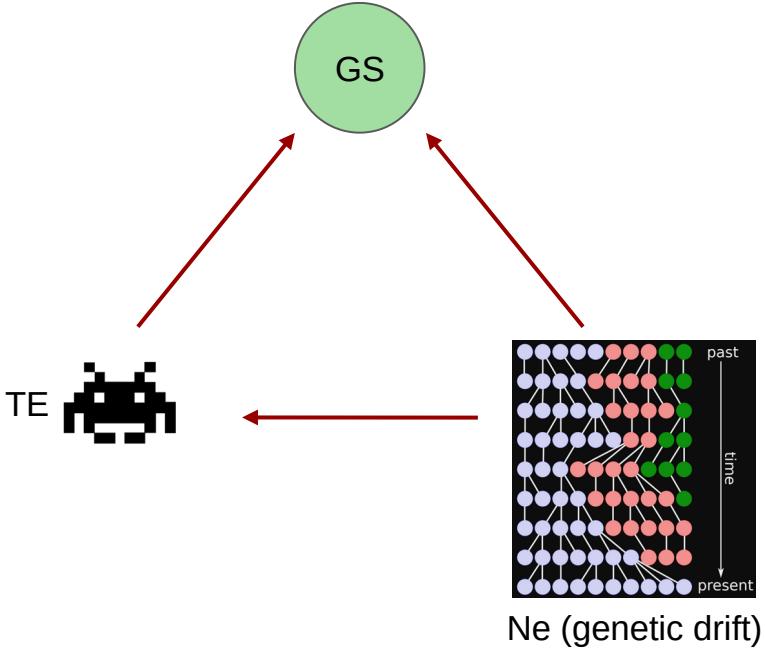
Ohta, 2008

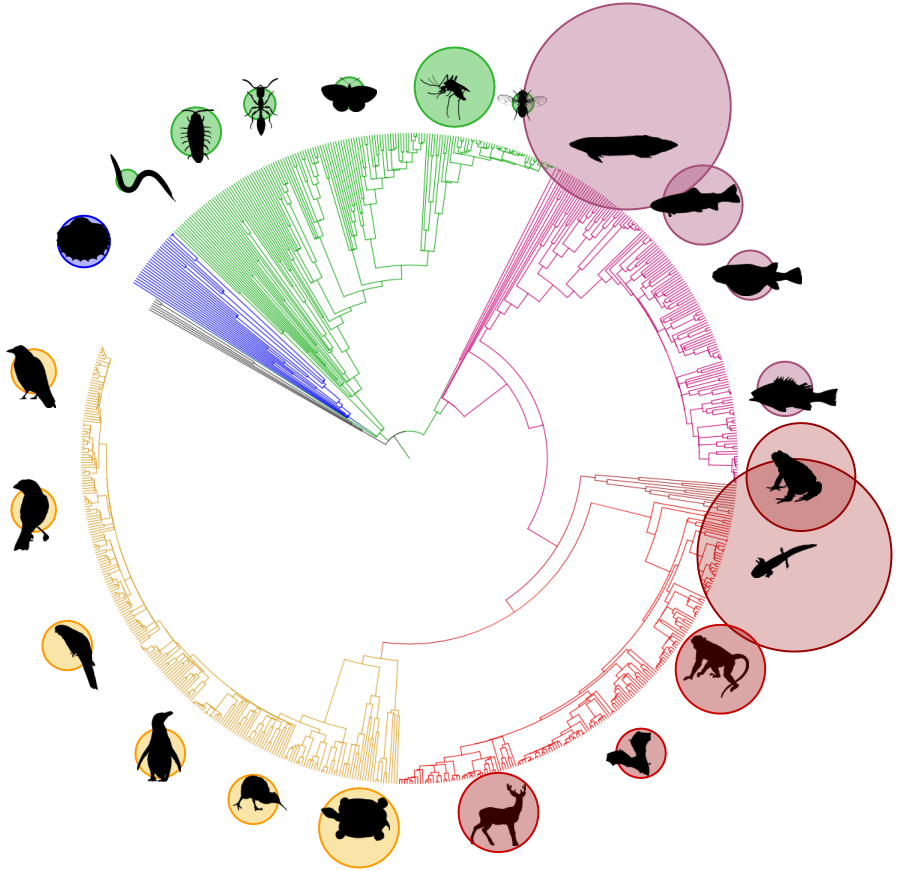
Genetic drift: why does effective population size matter?



Yi, 2006

Genome size variation in animals: impact of effective population size and transposable elements





Genome size variation in animals

Animals display a ~ 6000-fold variation in genome size

Pratylenchus coffeae - 20 Mb

Protopterus aethiopicus - 130 Gb

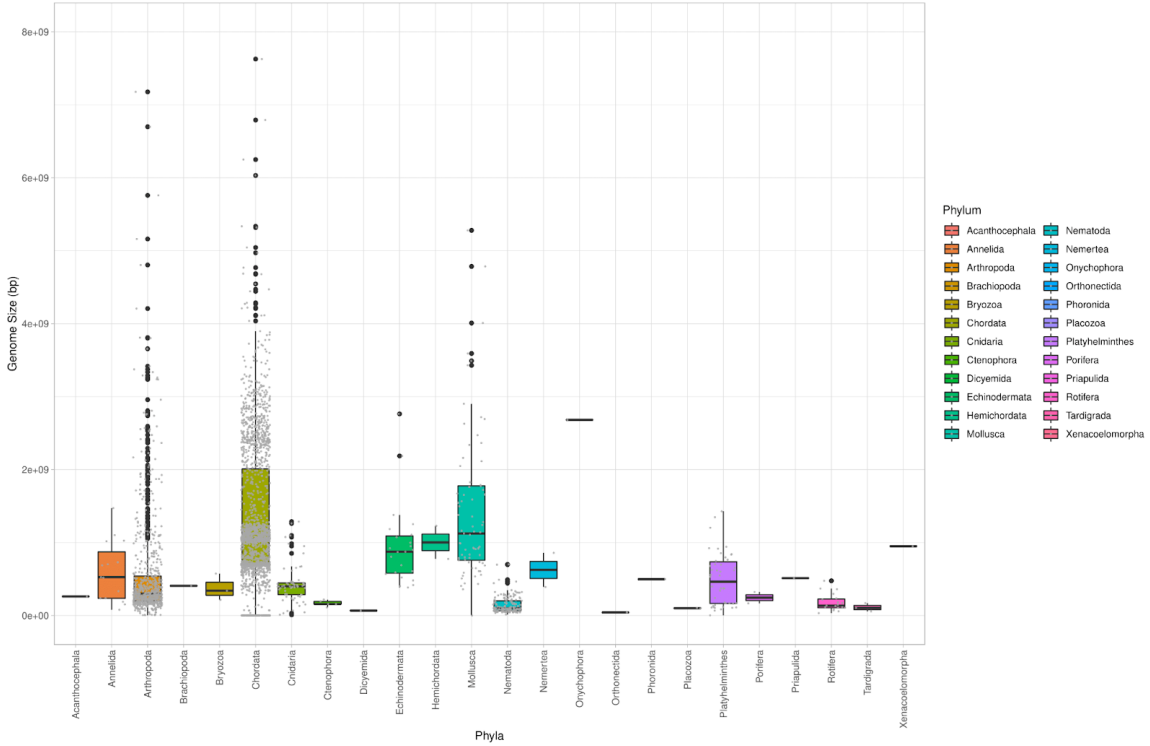
Junk DNA is the main responsible of this huge variation

↓

Introns
Pseudogenes
Intergenic sequences (spacers)
Repetitive sequences

↓

Transposable elements

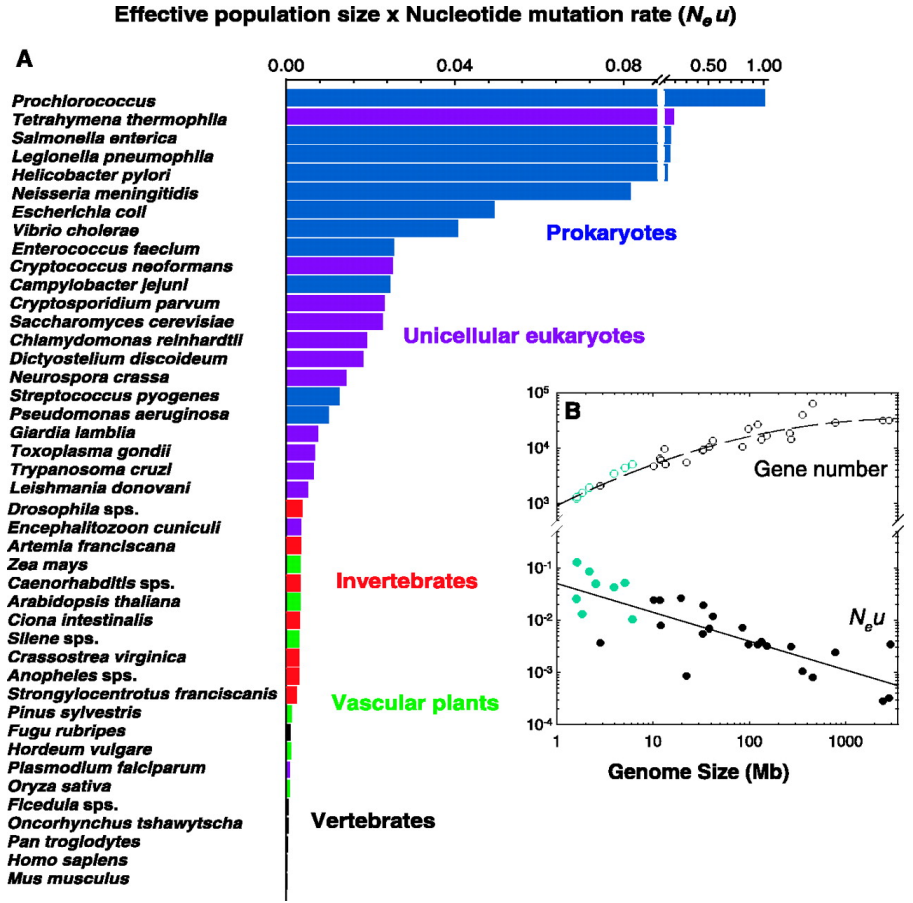


A neutralist explanation for genome size variation

Mutational Hazard Hypothesis (MHH)

Lynch and Conery (2003) observed an inverse correlation on a large scale between genome size and effective population size (Ne)

MHH: all the extra ncDNA filling genomes is potentially deleterious because of its increased liability to new mutations: new genetic material that is initially neutral can accumulate changes that can lead to gain deleterious functions



Lynch & Conery, 2003

A neutralist explanation for genome size variation

Mutational Hazard Hypothesis (MHH)

Lynch and Conery (2003) observed an inverse correlation on a large scale between genome size and effective population size (Ne)

MHH: all the extra ncDNA filling genomes is potentially deleterious because of its increased liability to new mutations: new genetic material that is initially neutral can accumulate changes that can lead to gain deleterious functions

We can consider the new insertion of a TE as a slightly deleterious mutation. According to the nearly-neutral theory, the destiny of the new TE copy depends on the N_e, measure of the power of genetic drift:

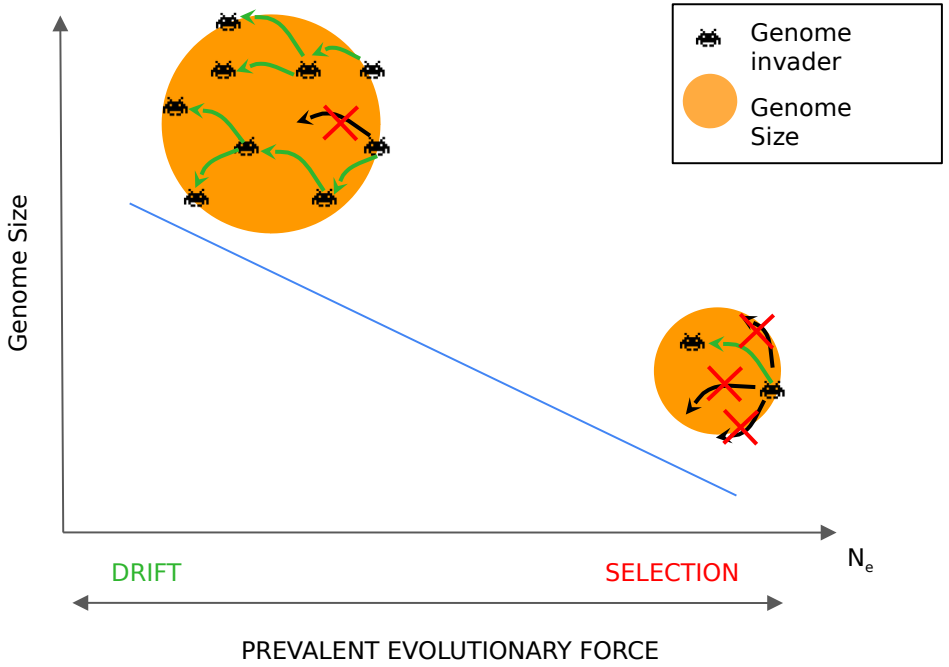
A neutralist explanation for genome size variation

Mutational Hazard Hypothesis (MHH)

Lynch and Conery (2003) observed an inverse correlation on a large scale between genome size and effective population size (Ne)

MHH: all the extra ncDNA filling genomes is potentially deleterious because of its increased liability to new mutations: new genetic material that is initially neutral can accumulate changes that can lead to gain deleterious functions

We can consider the new insertion of a TE as a slightly deleterious mutation. According to the nearly-neutral theory, the destiny of the new TE copy depends on the N_e, measure of the power of genetic drift:



A neutralist explanation for genome size variation

Mutational Hazard Hypothesis (MHH)

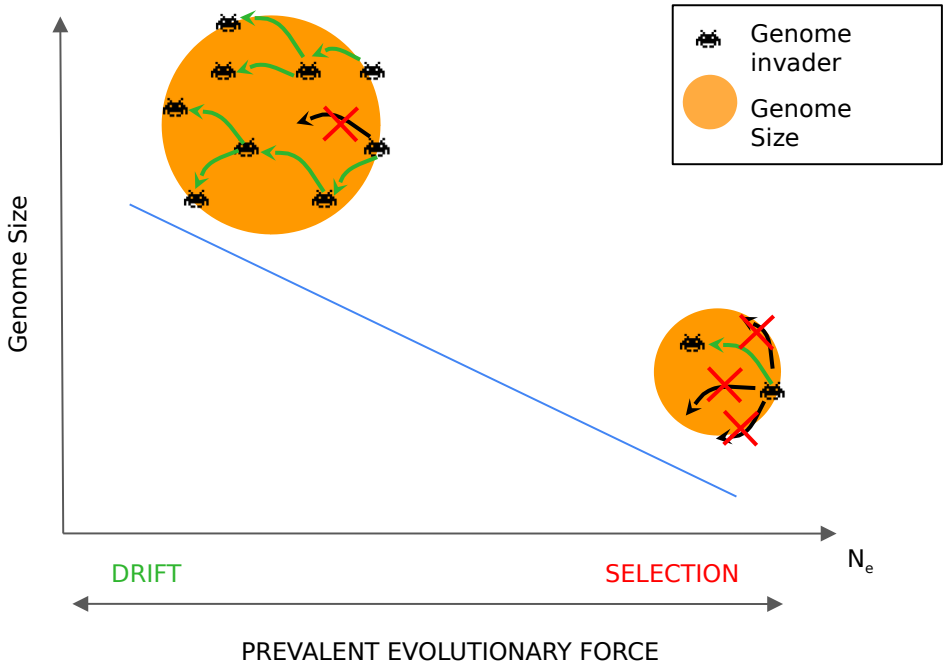
Lynch and Conery (2003) observed an inverse correlation on a large scale between genome size and effective population size (Ne)

MHH: all the extra ncDNA filling genomes is potentially deleterious because of its increased liability to new mutations: new genetic material that is initially neutral can accumulate changes that can lead to gain deleterious functions

We can consider the new insertion of a TE as a slightly deleterious mutation. According to the nearly-neutral theory, the destiny of the new TE copy depends on the N_e, measure of the power of genetic drift:

- In organisms with high N_e, the new TE will be effectively counter-selected and TEs overall will not contribute to the genome enlargement
- In organisms with low N_e, drift tends to overwhelm the effect of selection, allowing for a TE insertion with the same fitness effect to quickly drift to fixation, and to transposons to more easily invade the genome.

Genome size increases are led by random accumulation of nearly neutral

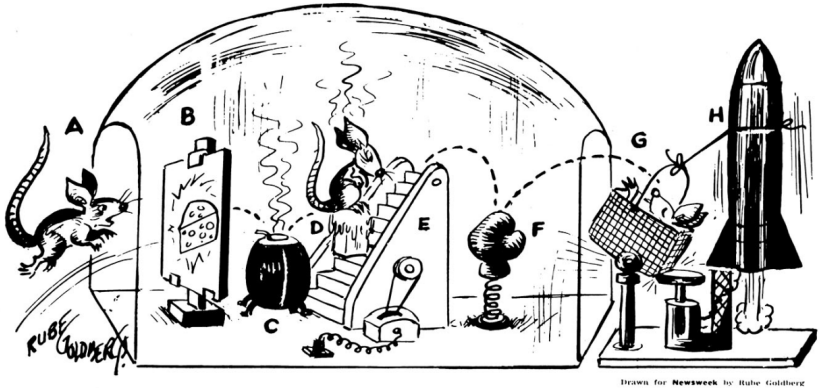


A neutralist explanation for genome size variation

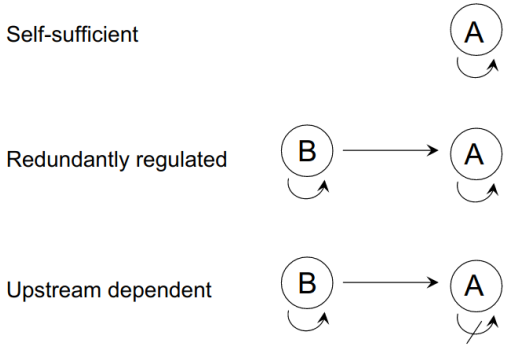
The MHH uses general principles of population genetics to explain the emergence of the features of *complex genome architecture* (genome size, TEs dynamics, introns).

Extra (deleterious) DNA is initially accumulated by randomness and supplies raw material for subsequent adaptation → building on of complex and often redundant biochemical pathways and genomic signatures (regulatory networks, protein complexes).

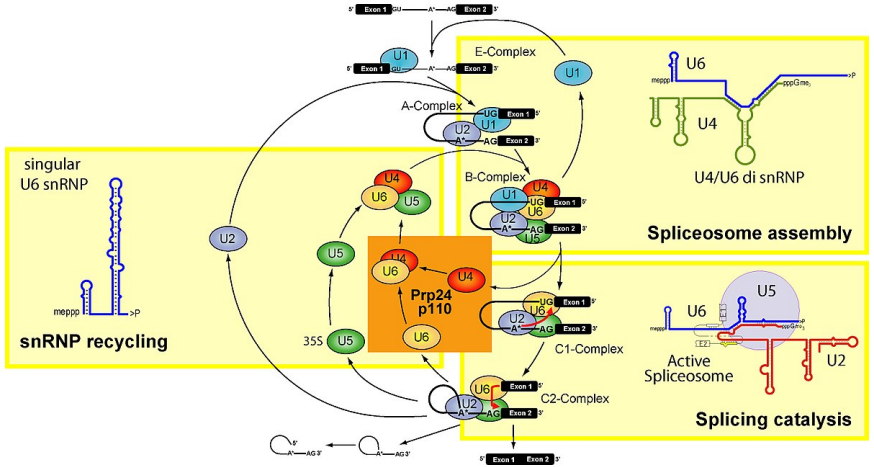
Drift and selection are equally important forces driving genome and organismal evolution



The best mousetrap by Rube Goldberg: Mouse (A) dives for painting of cheese (B), goes through canvas and lands on hot stove (C). He jumps on cake of ice (D) to cool off. Moving escalator (E) drops him on boxing glove (F) which knocks him into basket (G) setting off miniature rocket (H) which takes him to the moon.



Lynch, 2007



A neutralist explanation for genome size variation

Even if attractive, it is hard to establish the universal ability of MHH to predict a complex character such genome size that can be potentially influenced by many different factors

The data supporting the theory are based on very distantly related taxa across all the tree of life, however studies encompassing diversity at smaller scale are few and gave so far contrasting results

Can MHH also explain GS variation as a general trend and between phylogenetically closer lineages?

ARTICLE

https://doi.org/10.1038/s41467-019-11308-4 OPEN

The determinants of genetic diversity in butterflies

Alexander Mackintosh[1], Dominik R. Laetsch[1], Alexander Hayward[2], Brian Charlesworth[1], Martin Waterfall[1], Roger Vila[3] & Konrad Lohse[1]

Under the neutral theory, genetic diversity is expected to increase with population size. While comparative analyses have consistently failed to find strong relationships between census population size and genetic diversity, a recent study across animals identified a strong correlation between propagule size and genetic diversity, suggesting that r-strategists that produce many small offspring, have greater long-term population sizes. Here we compare genome-wide genetic diversity across 38 species of European butterflies (Papilionoidea), a group that shows little variation in reproductive strategy. We show that genetic diversity across butterflies varies over an order of magnitude and that this variation cannot be explained by differences in current abundance, propagule size, host or geographic range. Instead, neutral genetic diversity is negatively correlated with body size and positively with the length of the genetic map. This suggests that genetic diversity is determined both by differences in long-term population size and the effect of selection on linked sites.

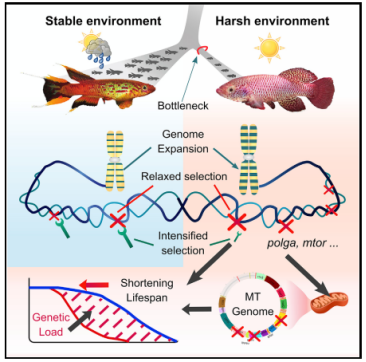
Lynch and Conery[63] have put forward analogous arguments for the evolution of genome sizes: genomes may expand in populations with low N_e, if selection against transposable element proliferation and intron expansion becomes inefficient. While the large genome size and TE content of *Leptidea* species[62] is consistent with this, we find no support for any relationship between genome size and neutral diversity across our set of species. Instead, our analyses clearly show that genome size has significant phylogenetic signal across butterflies ($n = 37$, Pagel's $\lambda = 1.000$, $p = 6.1 \times 10^{-7}$) and so must evolve slowly, whereas variation in genetic diversity has little phylogenetic structure (Fig. 1).

Article

Cell

Relaxed Selection Limits Lifespan by Increasing Mutation Load

Graphical Abstract



Authors
Rongfeng Cui, Tania Medeiros, David Willemsen, ..., Martin Graef, Martin Reichard, Dario Riccardo Valenzano

Correspondence
rcui@age.mpg.de (R.C.), dvalenzano@age.mpg.de (D.R.V.)

In Brief

Studying the genomic changes accompanying the repeated evolution of short lifespan in 45 African killifish species reveals that neutral genetic drift, rather than adaptive evolution, explains the accumulation of deleterious gene variants affecting lifespan and aging.

Research

Less effective selection leads to larger genomes

Tristan Lefébure,[1] Claire Morvan,[1] Florian Malard,[1] Clémentine François,[1] Lara Konecny-Dupré,[1] Laurent Guéguen,[2] Michèle Weiss-Gayet,[3] Andaine Seguin-Orlando,[4] Luca Ermini,[4] Clio Der Sarkissian,[4] N. Pierre Charrier,[1] David Eme,[1] Florian Mermillod-Blondin,[1] Laurent Duret,[2] Cristina Vieira,[2,5] Ludovic Orlando,[4,6] and Christophe Jean Douady[1,5]

[1]Université de Lyon, Université Claude Bernard Lyon 1, CNRS UMR 5023, ENTPE, Laboratoire d'Ecologie des Hydrosystèmes Naturels et Anthropisés, F-69622 Villeurbanne, France; [2]Université de Lyon, Université Claude Bernard Lyon 1, CNRS UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, F-69622 Villeurbanne, France; [3]Université de Lyon, Université Claude Bernard Lyon 1, CNRS UMR 5310, INSERM, Institut NeuroMyoGène, F-69622 Villeurbanne, France; [4]Center for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, 1350K Copenhagen, Denmark; [5]Institut Universitaire de France, F-75005 Paris, France; [6]Université de Toulouse, University Paul Sabatier (UPS), CNRS UMR 5288, Laboratoire AMIS, F-31073 Toulouse, France

The evolutionary origin of the striking genome size variations found in eukaryotes remains enigmatic. The effective size of populations, by controlling selection efficacy, is expected to be a key parameter underlying genome size evolution. However, this hypothesis has proved difficult to investigate using empirical data sets. Here, we tested this hypothesis using 22 de novo transcriptomes and low-coverage genomes of asellid isopods, which represent 8 independent habitat shifts from surface water to resource-poor groundwater. We show that these habitat shifts are associated with higher transcriptome-wide d_N/d_S. After ruling out the role of positive selection and pseudogenization, we show that these transcriptome-wide d_N/d_S increases are the consequence of a reduction in selection efficacy imposed by the smaller effective population size of subterranean species. This reduction is paralleled by an important increase in genome size (25% increase on average), an increase also confirmed in subterranean decapods and mollusks. We also control for an adaptive impact of genome size on life history traits but find no correlation between body size, or growth rate, and genome size. We show instead that the independent increases in genome size measured in subterranean isopods are the direct consequence of increasing invasion rates by repeat elements, which are less efficiently purged out by purifying selection. Contrary to selection efficacy, polymorphism is not correlated to genome size. We propose that recent demographic fluctuations and the difficulty of observing polymorphism variation in polymorphism-poor species can obfuscate the link between effective population size and genome size when polymorphism data are used alone.

Can we link genome size and TE activity to effective population size?

The purpose of the project is to systematically test the MHH in relatively closely related lineages and across diverse animal groups

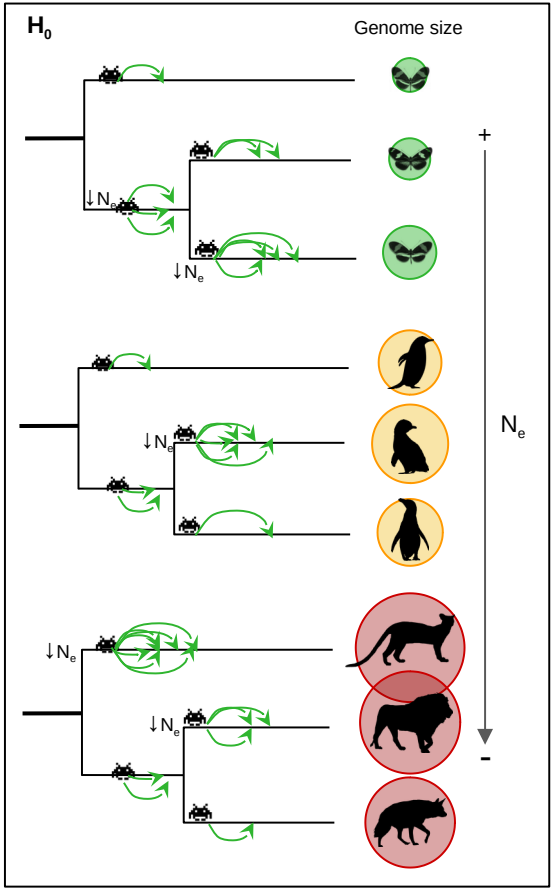
Main steps:

1)Acquiring genomic data spanning both distantly diverged and closely related animal lineages from public repositories
2)Estimation of the genome sizes (assembly size, k-mer frequency, c-value)
3)Estimation of proxy parameters for N_e (polymorphism data, divergence data)
4)TE annotation for both a qualitative (TE type) and quantitative (% of genome) assessment of the repetitive content of genomes

Does variation of long-term Ne explain genome size variation?

Is genome size significantly affected by its repeat content (TEs)?

Can N_e variation explain the observed bursts of TE activity ?



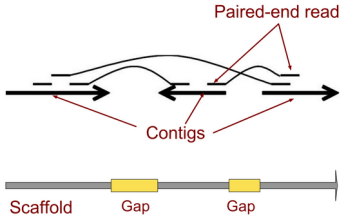
A bit of methods

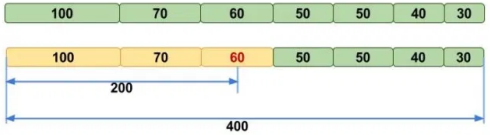
1) Acquiring genomic data

Sampling of reference genome assemblies of all metazoan groups in NCBI (as of october-november 2021) → 3214 species

Not all the species have the same data quality → filter the dataset based on assembly quality statistics, conserved genes completeness and availability of raw reads

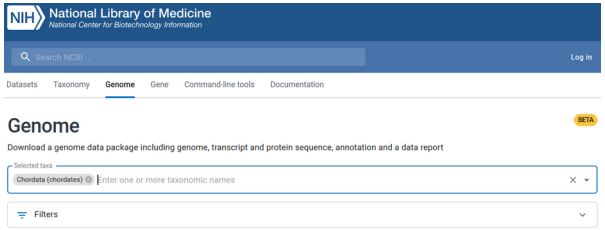
Contig N50 >= 50 Kb

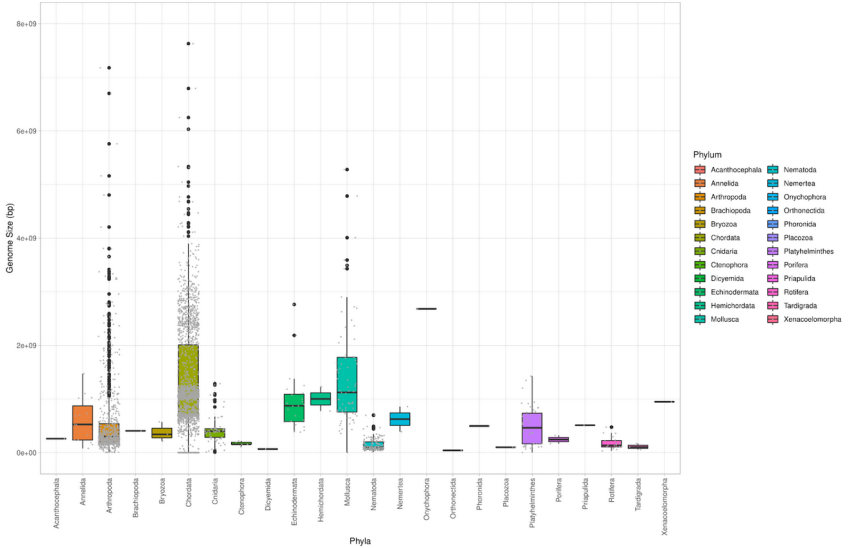




BUSCO **>= 70 %**

Benchmarking Universal Single-Copy Orthologs: evaluates assembly completeness based on the detection of single copy metazoan-conserved gene sets (954 markers)



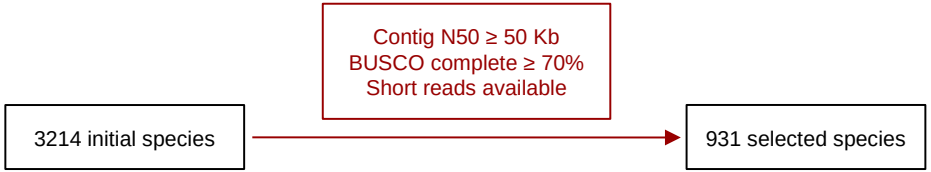


A bit of methods

1) Acquiring genomic data

Sampling of reference genome assemblies of all metazoan groups in NCBI (as of october-november 2021) → 3214 species

Not all the species have the same data quality → filter the dataset based on assembly quality statistics, conserved genes completeness and availability of raw reads




```
  4 Annelida
230 Arthropoda
  1 Brachiopoda
  2 Bryozoa
632 Chordata
  7 Cnidaria
  4 Echinodermata
 28 Mollusca
 12 Nematoda
  1 Phoronida
  1 Placozoa
  1 Platyhelminthes
  5 Rotifera
  2 Tardigrada
```

A bit of methods

1) Estimating genome size

Which method(s) to use?

- Assembly size → it is available for all the species but can be less reliable for size estimation of large highly repetitive genomes
- C-values (flow cytometry, Feulgen densitometry) from the Animal Genome Size Database → available for 269 species (less than ⅓ of the dataset)
- K-mer frequencies approach → exploitable if the original short reads are available (829 species)

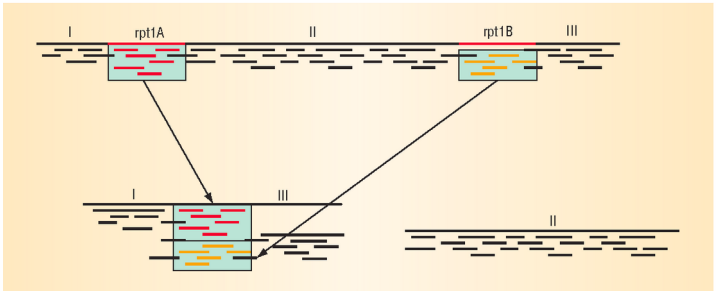
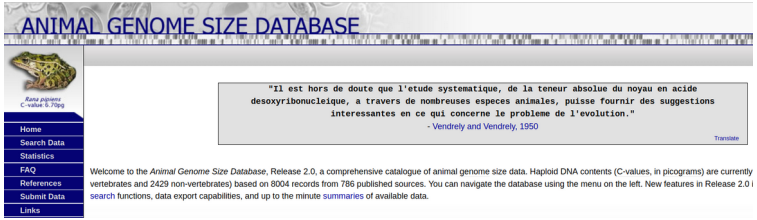
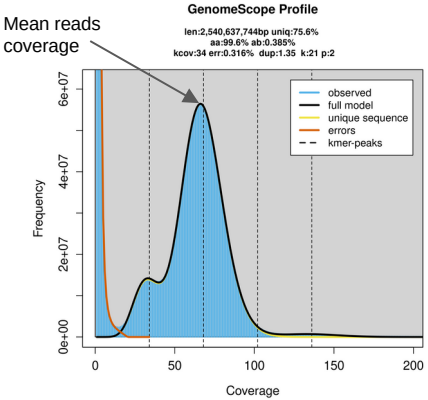


Figure 2. Repeat sequence. The top represents the correct layout of three DNA sequences. The bottom shows a repeat collapsed in a misassembly.

Pop et al., 2002



21-mers, 2N

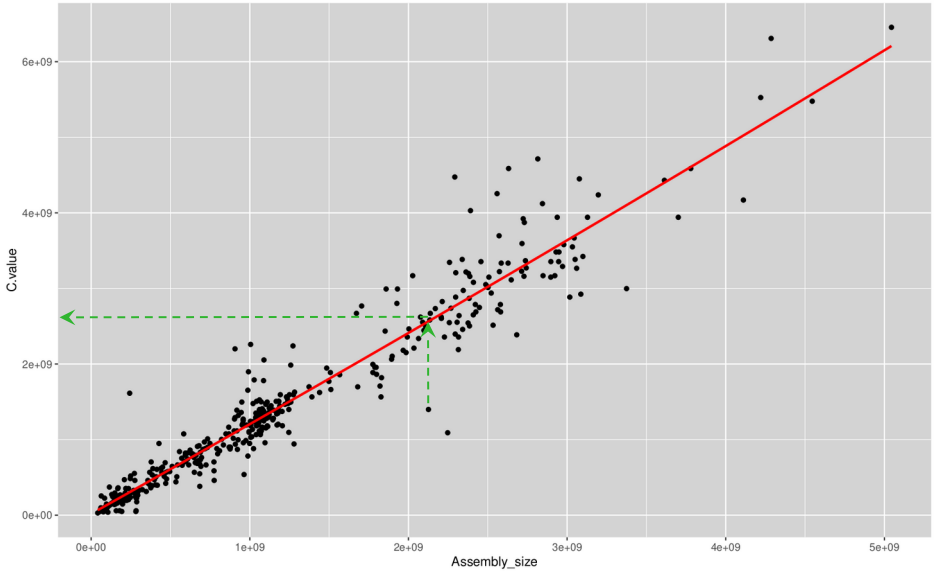


A bit of methods

1) Estimating genome size

Assembly sizes overall tend to underestimation compared to the corresponding C-values

From the linear regression, a “corrected GS” is estimated for all the species starting from their original assembly size



A bit of methods

2) Estimating effective population size

Genetic divergence data give a proxy of Ne and is calculated on the 954 busco single-gene set

dN / dS = rate of nonsynonymous substitutions / rate of synonymous substitutions

aa sequence changes

aa sequence is unchanged

dN << dS → few fixed nonsynonymous mutations are index of purifying selection

dN >> dS → an excess of fixed nonsynonymous mutations is index of positive selection

dN ≈ dS → sequence synonymous and nonsynonymous changes occurring approximately at the same rate is index of neutral evolution

A bit of methods

3) TE annotation

Correctly annotating the actual amount and age of TEs is particularly challenging because of their repetitive nature

3.1 Starting from the assembly

- *Homology searches* of known TE family sequences in the genome assembly (public libraries like Dfam, RepBase)

Issue
Sequences in public libraries come so far from a limited number known species: using only this method will work well with species represented in the libraries but will miss a lot of lineage specific TEs in unrepresented species. The higher the phylogenetic distance, the more the undetected repeated sequences

- *De novo* discovery of new elements with algorithms based on repetitivity detection (RepeatModeler)

The integration of these two methods currently give a more complete estimate of the quantity and types of TEs in a genome

Issue
The completion of a pipeline merging homology search, *de novo* discovery and curation of the outputs is very time consuming and computational demanding (more than 900 species!!!)

A bit of methods

3) TE annotation

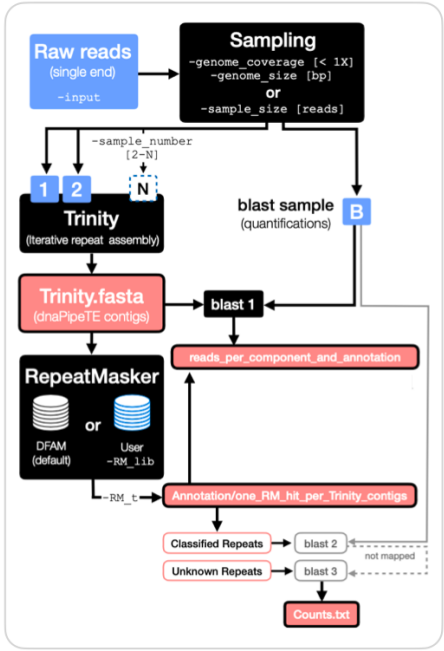
3.2 Starting from the unassembled reads

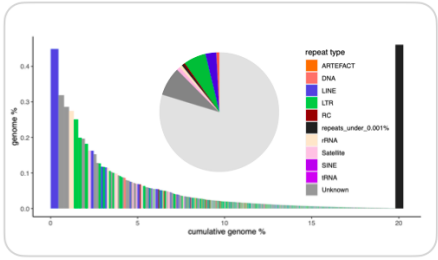
Given that most of the genomes are highly repeated, it is assumed that an undersampling of the short reads (e.g. 0.25x) will only leave repeated sequences.

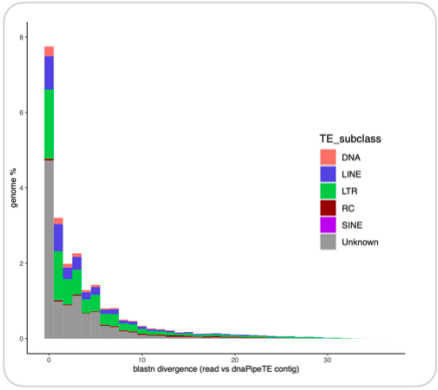
The reads of multiple undersamplings are assembled into contigs, blasted against the public TE databases, and used to quantify the TE types proportions.

Much faster than the previous method

Compare the outcomes of the two pipelines on a subset of species to test their consistency and use the faster method







References

Gregory, T. R. (2004). Macroevolution, hierarchy theory, and the C-value enigma. Paleobiology, 30(2), 179-202.
Yi, S. V. (2006). Non-adaptive evolution of genome complexity. Bioessays, 28(10), 979-982.
Russell, P. J., Cicchini, C., Marchetti, A., & Antoccia, A. (2014). Genetica: un approccio molecolare. Pearson.
Gregory, T.R. (2005). Animal Genome Size Database. http://www.genomesize.com.
Lynch, M., & Walsh, B. (2007). The origins of genome architecture (Vol. 98). Sunderland, MA: Sinauer Associates.
Lynch, M., & Conery, J. S. (2003). The origins of genome complexity. science, 302(5649), 1401-1404.
Hou, Y., & Lin, S. (2009). Distinct gene number-genome size relationships for eukaryotes and non-eukaryotes: gene content estimation for dinoflagellate genomes. PloS one, 4(9), e6978.
Wells, J. N., & Feschotte, C. (2020). A field guide to eukaryotic transposable elements. Annual review of genetics, 54, 539.
Orgel, L. E., & Crick, F. H. (1980). Selfish DNA: the ultimate parasite. Nature, 284(5757), 604-607.
Gregory, T. R. (2005). Synergy between sequence and size in large-scale genomics. Nature Reviews Genetics, 6(9), 699-708.
Ward, L. D., & Kellis, M. (2012). Evidence of abundant purifying selection in humans for recently acquired regulatory functions. Science, 337(6102), 1675-1678.
Jacobs, F. M., Greenberg, D., Nguyen, N., Haeussler, M., Ewing, A. D., Katzman, S., ... & Haussler, D. (2014). An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. Nature, 516(7530), 242-245.
Feschotte, C. (2008). Transposable elements and the evolution of regulatory networks. Nature Reviews Genetics, 9(5), 397-405.
Cosby, R. L., Chang, N. C., & Feschotte, C. (2019). Host–transposon interactions: conflict, cooperation, and cooption. Genes & development, 33(17-18), 1098-1116.
Cosby, R. L., Judd, J., Zhang, R., Zhong, A., Garry, N., Pritham, E. J., & Feschotte, C. (2021). Recurrent evolution of vertebrate transcription factors by transposase capture. Science, 371(6531), eabc6405.
Stankiewicz, P., & Lupski, J. R. (2002). Genome architecture, rearrangements and genomic disorders. TRENDS in Genetics, 18(2), 74-82.
Gould, S. J., & Lewontin, R. C. (2020). The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. In Shaping Entrepreneurship Research (pp. 204-221). Routledge.
Ohta, T. (2008). Molecular Evolution: Nearly Neutral Theory. eLS.
Lynch, M., & Conery, J. S. (2003). The origins of genome complexity. science, 302(5649), 1401-1404.
Lynch, M. (2007). The frailty of adaptive hypotheses for the origins of organismal complexity. Proceedings of the National Academy of Sciences, 104(suppl_1), 8597-8604.