# The complexity of mutagenesis: beyond the molecular clock

Kelley Harris
University of Washington

Workshop on Population and Speciation Genomics
January 24, 2020

# Forces that shape genomic diversity

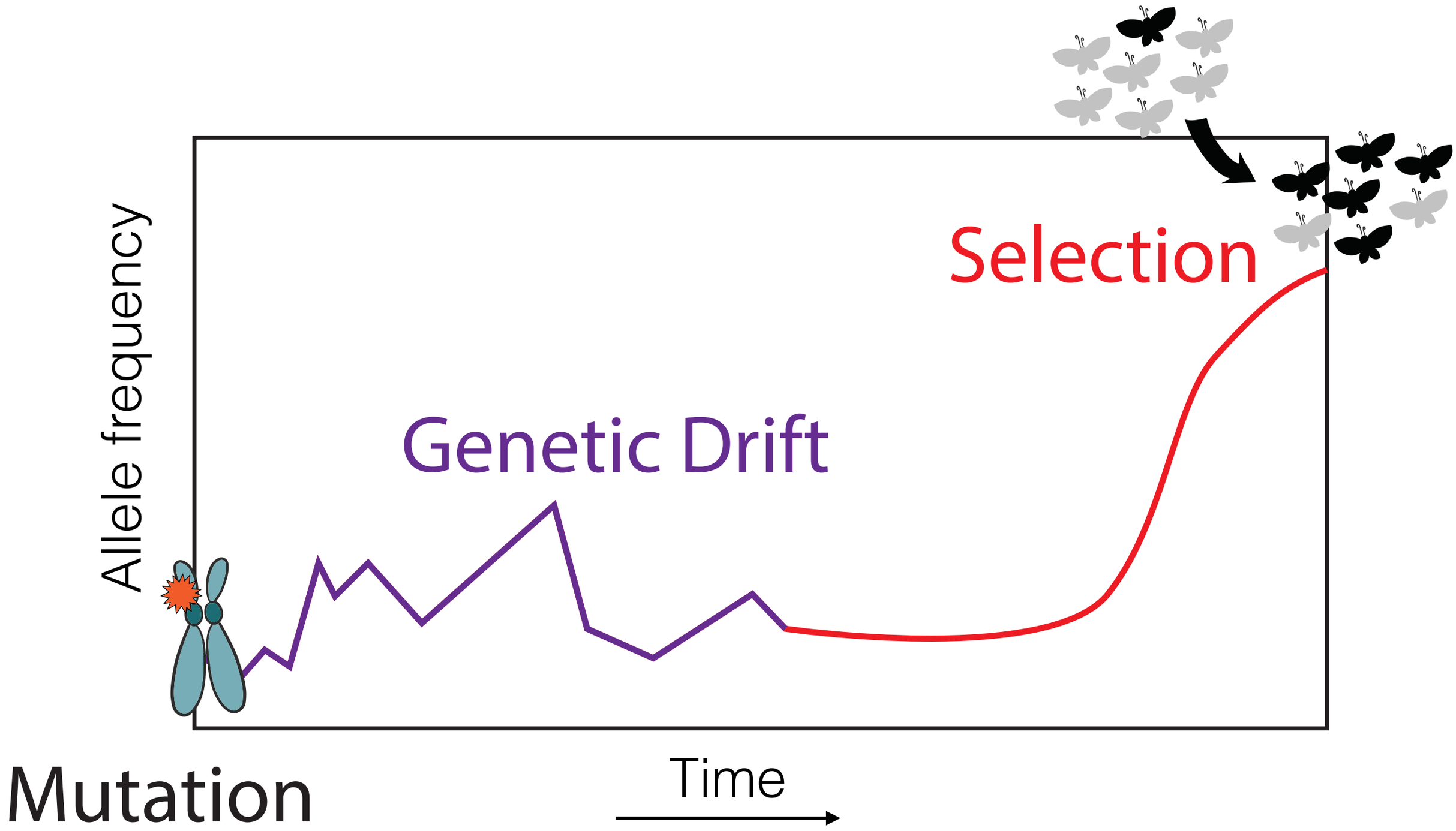
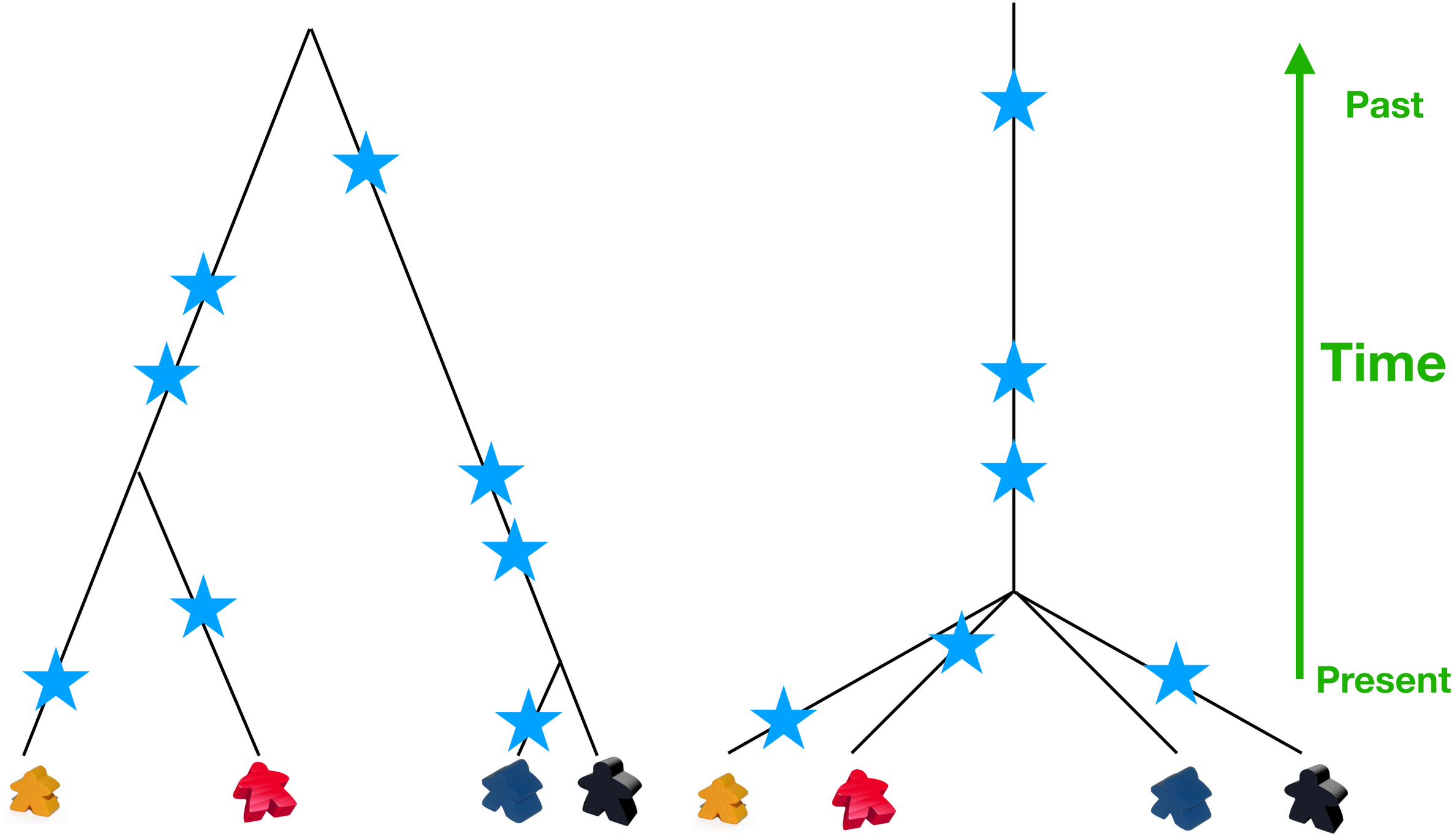



Image co-artist: Natalie Telis

# Mutations as a molecular clock



# When the clock breaks down (runs out of batteries?)

- Almost every population genetic method assumes that mutations accumulate at a constant rate per year within populations

- This assumption works fine until it doesn't

- The mutation process has complex features that can trip you up if you aren't looking out for them

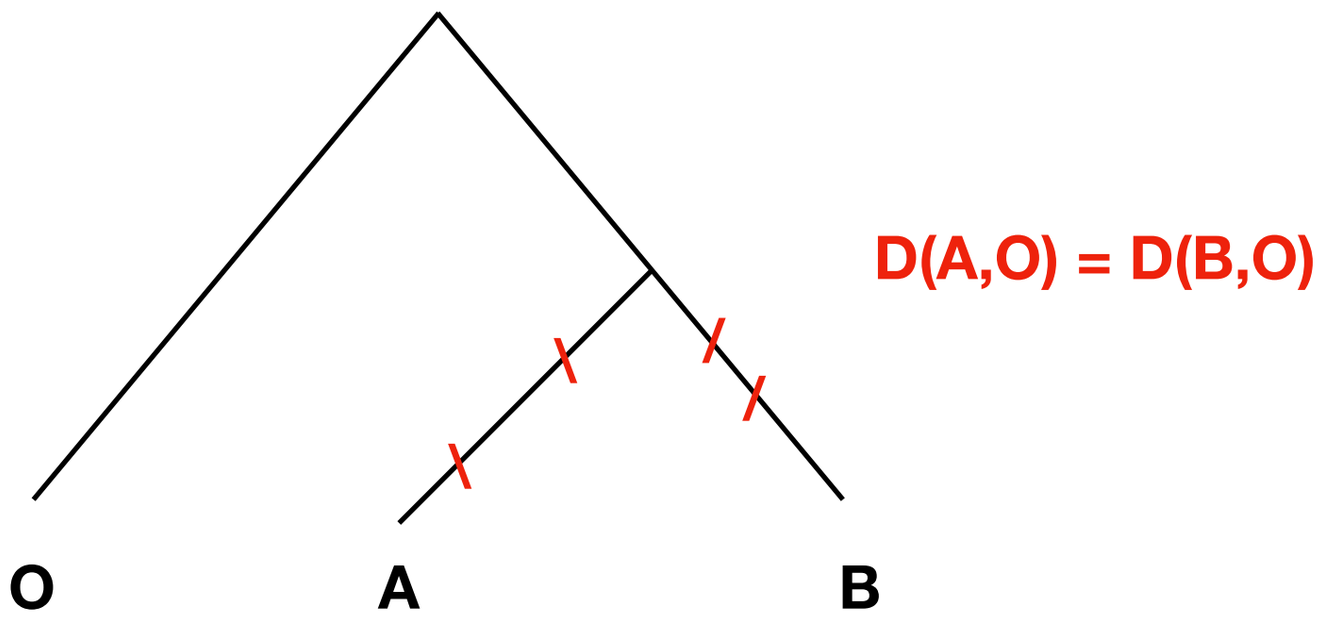
  - and are also interesting phenotypes to study in their own right

- Estimates of the mutation rate per year and generation time are needed to calibrate output of PSMC and other demographic inference methods

# Molecular clock 101

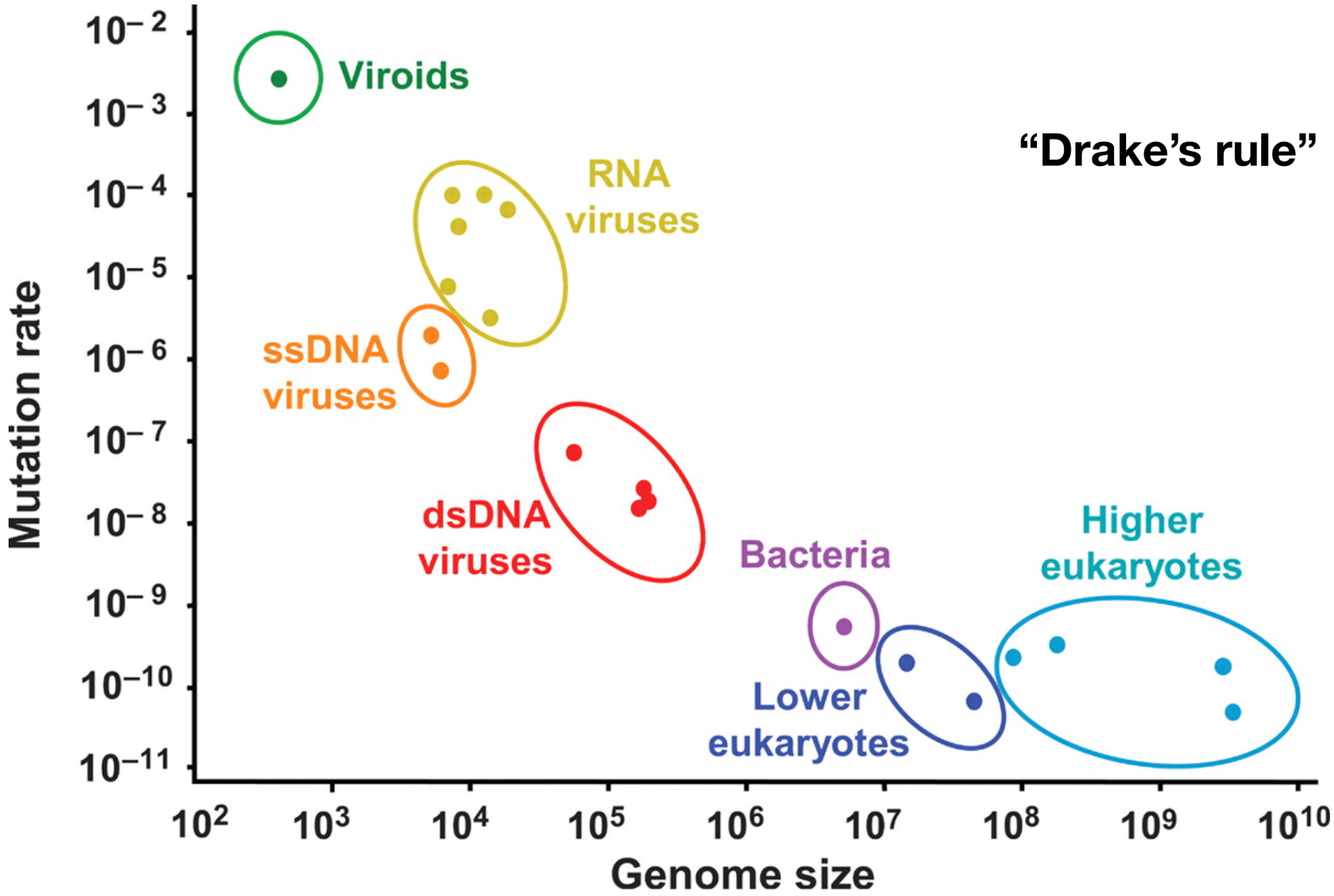
- Mutagenesis is more clock-like over short timescales compared to long time scales

- A simple branch length test can reveal whether mutagenesis is clock-ish in your data:




**Data can fail this test due to mutation rate variation, selection, or introgression**

# Violation of molecular clock over very long timescales




**“Drake’s rule”**

**Drake 1991**

Gago, et al. *Science* 2009

# The error threshold

- A simple model by Eigen & Schuster (1979) justifies Drake's rule

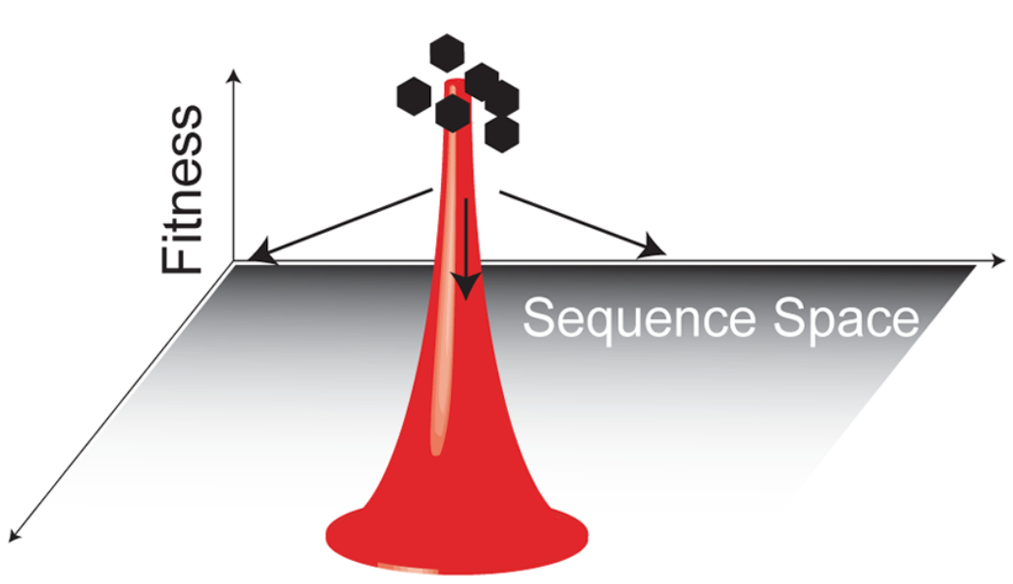
- Consider a "master" virus with fitness 1+*s* and genome length *L*

- All mutant viruses have fitness 1

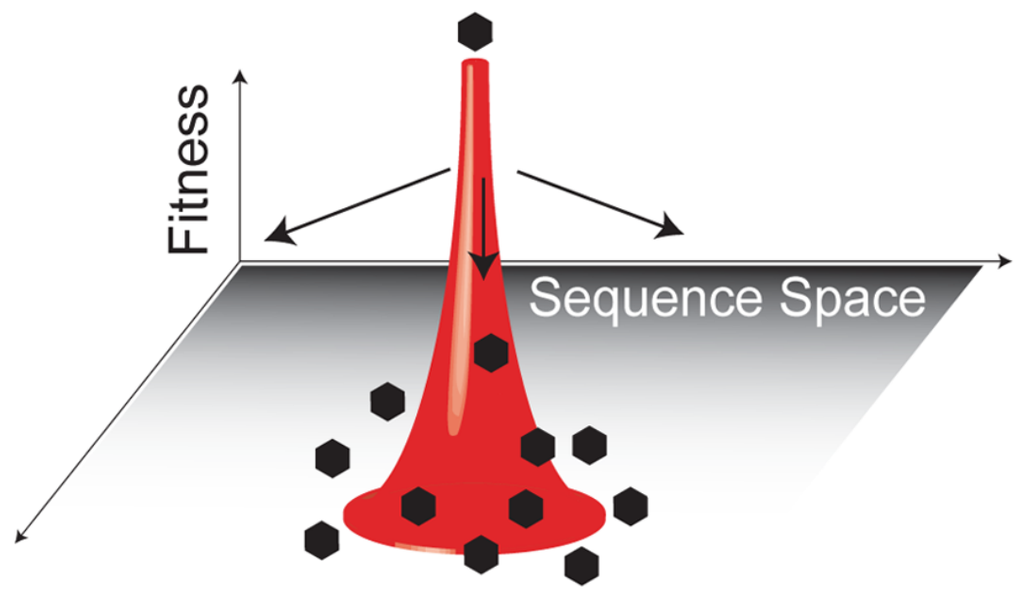
- The master sequence will die out due to Muller's rachet/"error catastrophe" if and only if the mutation rate mu is below a threshold:

- *mu* < log(s)/L

# Stable quasispecies vs error catastrophe




***mu*** **< error threshold**

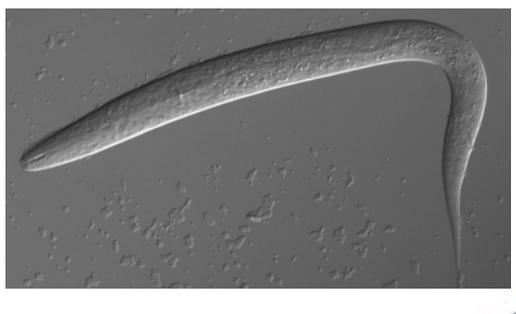


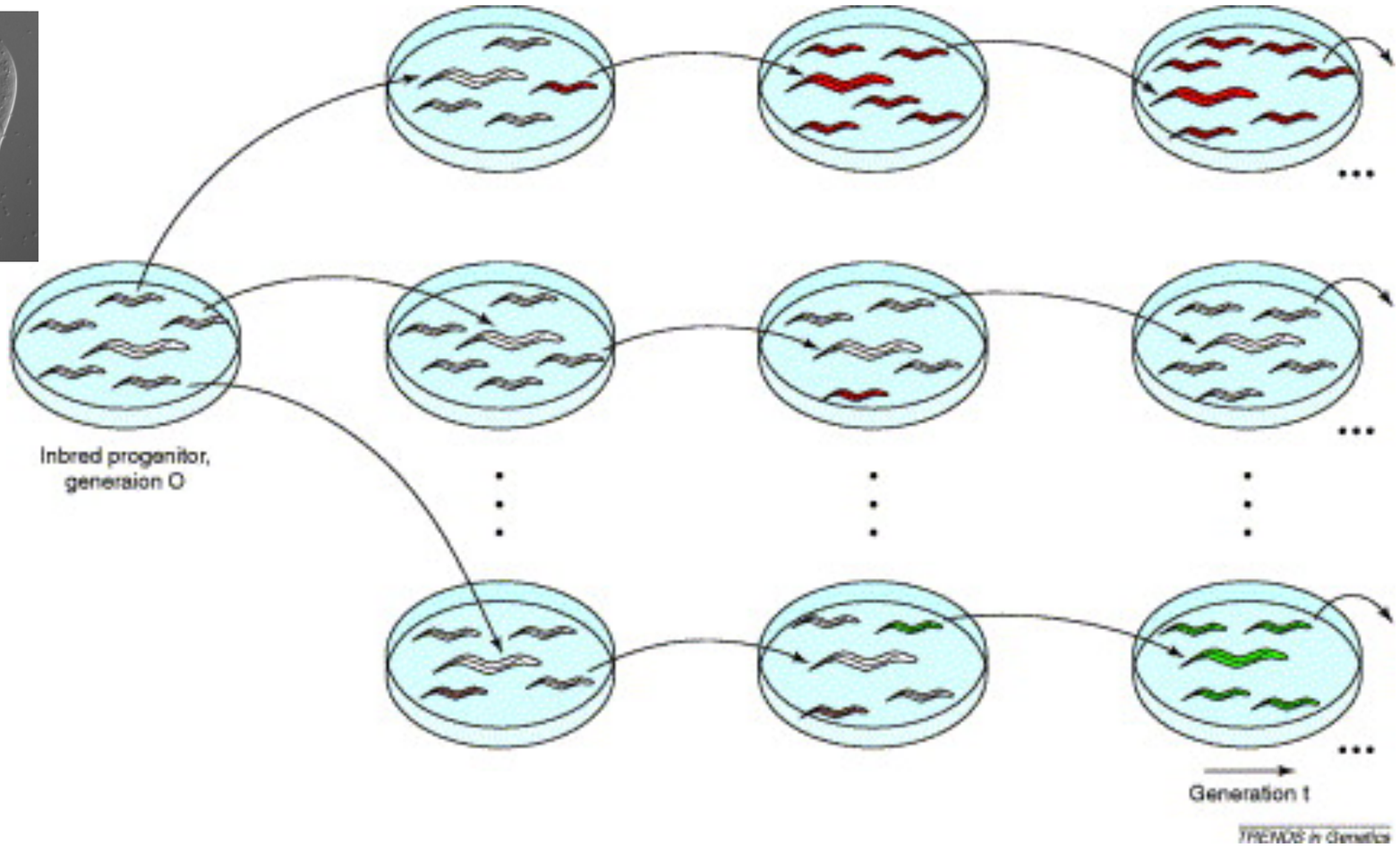

***mu*** **> error threshold**

**Lauring and Andino 2010**

# How can we gather mutation rate data to test these theories?

# Measuring mutation rates with mutation accumulation (MA) lines






**Keightly and Charlesworth 2005**

# MA with a reporter gene

Reporter gene
(e.g. encoding GFP or
luciferase)

**Inactive/ broken promoter**


**Point mutations can restore promoter function**


A reporter protein
Amount is easily measured
(e.g. GFP by fluorescence)

# Mutation rate estimates vary enormously in quality

- PSMC results, divergence time estimates, etc. depend heavily upon a mutation rate estimate. Where does that number come from?

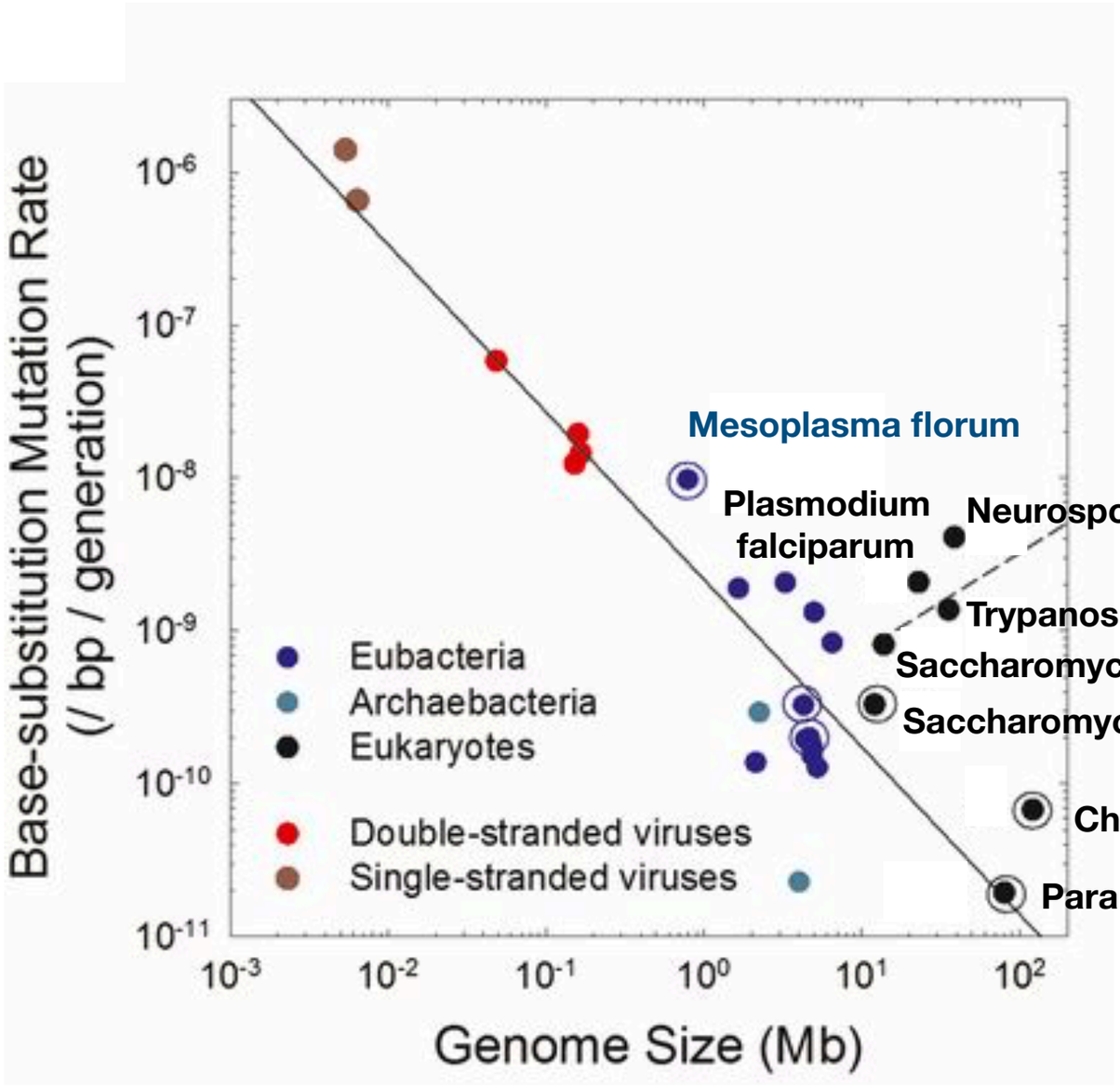
- Calculation from phylogenetic divergence data (substitutions / estimated divergence time)

- MA experiment + whole genome sequencing ($$-$$$$)

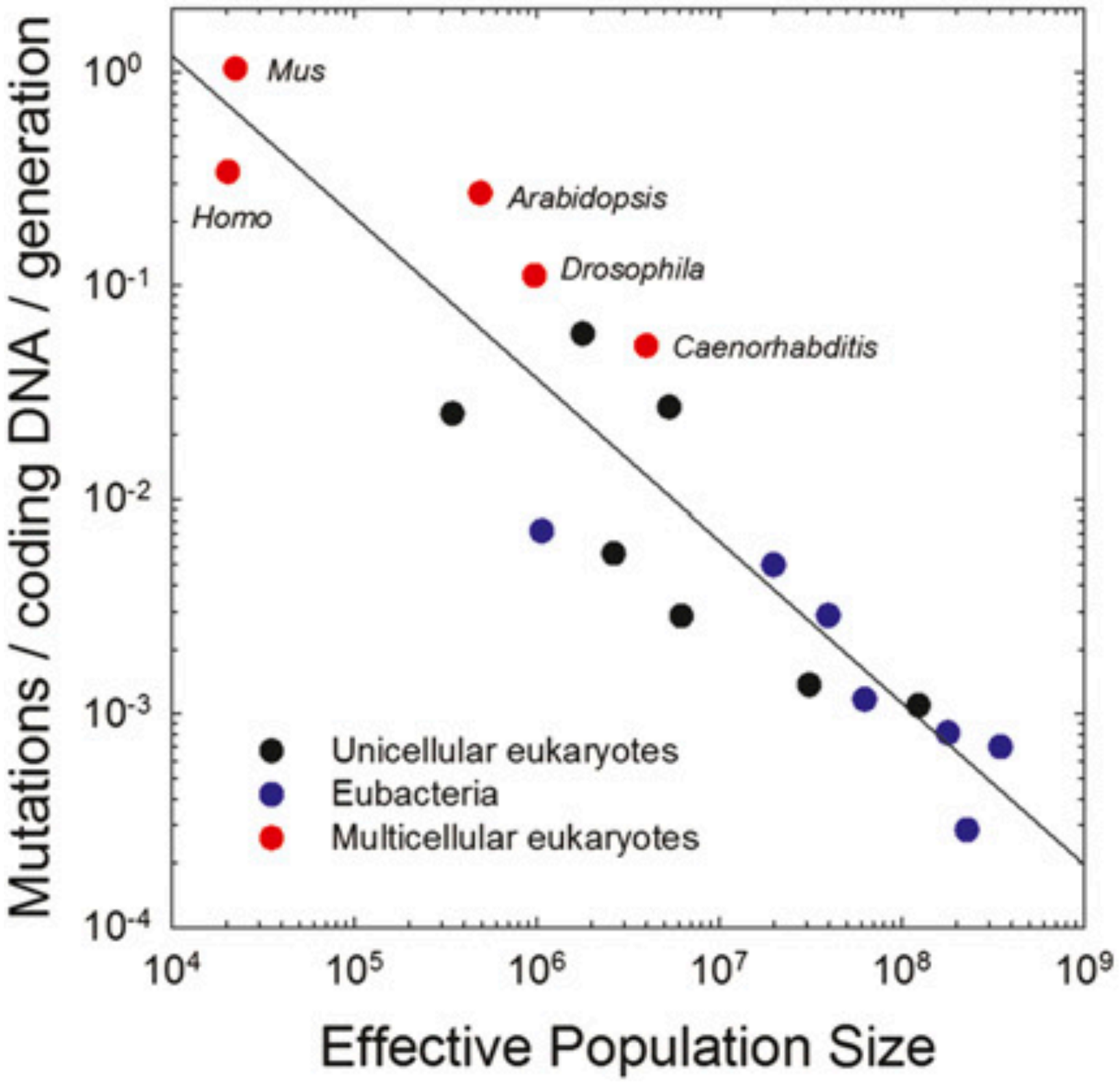
- MA experiment + reporter gene sequencing (cheap today, only reasonable direct estimate 10 years ago)

- Whole-genome trio sequencing ($$$$$$$$$$)

# Drake's rule driven mostly by viruses and bacteria




Sung, et al. 2012



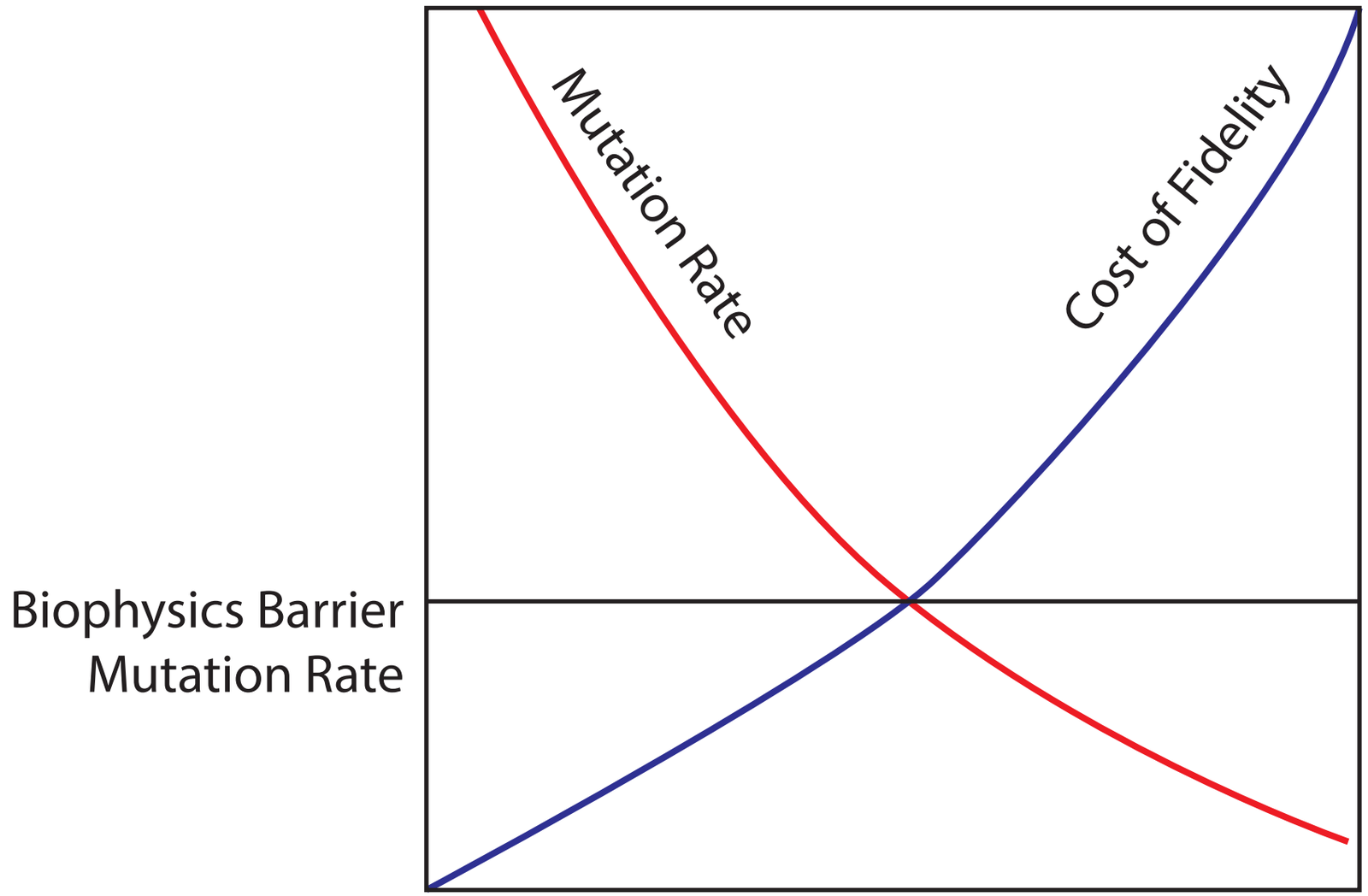

Lynch *Trends in Genetics* 2010

Sung, *et al*. *PNAS* 2012

# Why should effective population size affect mutation rate?

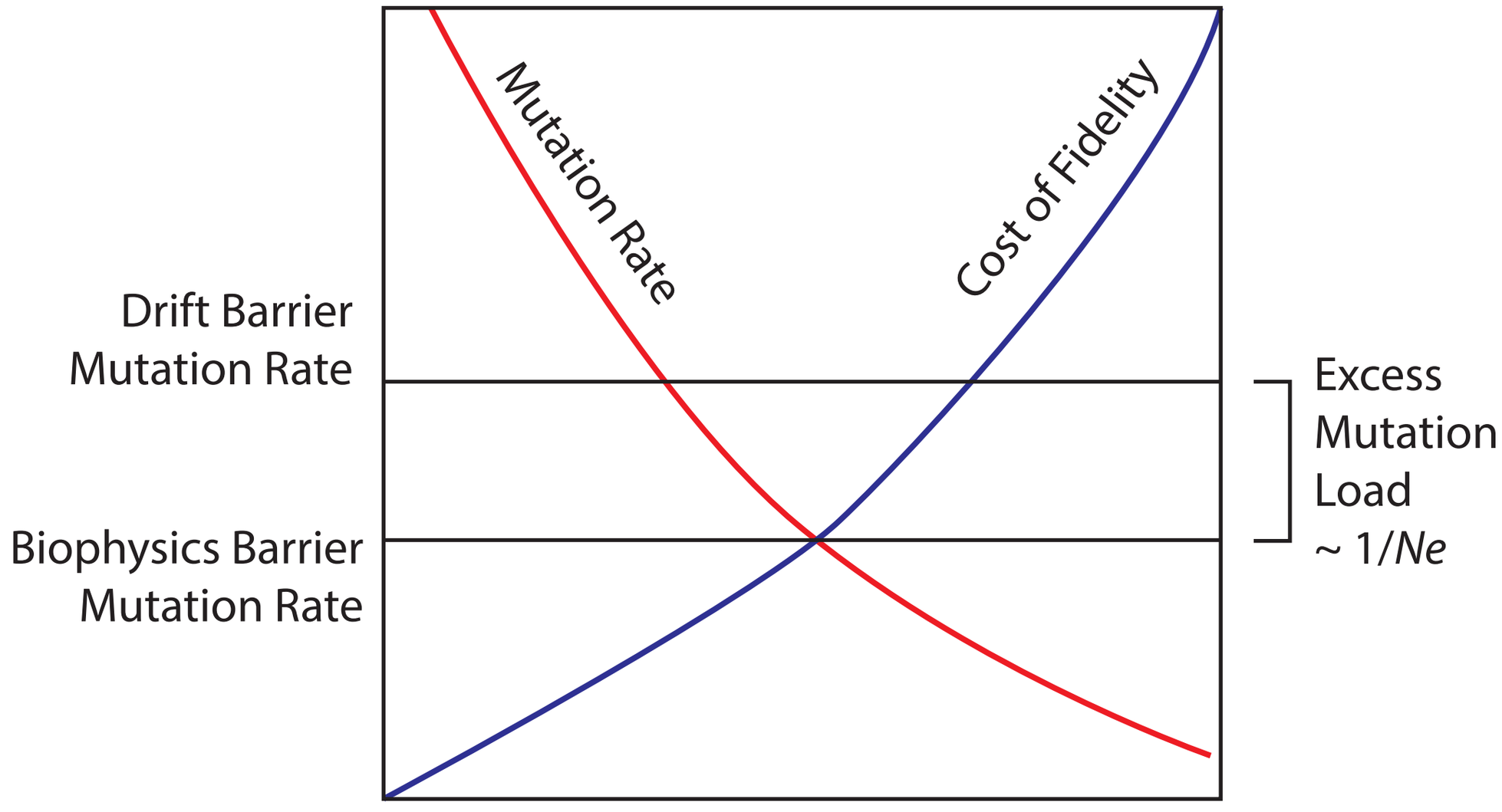
**Why is the mutation rate what it is?**

# 1.The Cost-of-Fidelity Model




Lynch *Trends in Genetics* 2010 Sung, *et al*. *PNAS* 2012

# 2.The Drift-Barrier Hypothesis




Lynch *Trends in Genetics* 2010 Sung, *et al*. *PNAS* 2012

# Mutators can be favored in asexual organisms

- Expected extra load of deleterious mutations must not exceed the expected benefit of beneficial mutations
- Robustness to environmental change
- Stress-induced mutagenesis?

# Stress-Induced Mutagenesis in Bacteria

Ivana Bjedov[1,*], Olivier Tenaillon[2,*], Bénédicte Gérard[2,*], Valeria Souza[3], Erick Denamur…

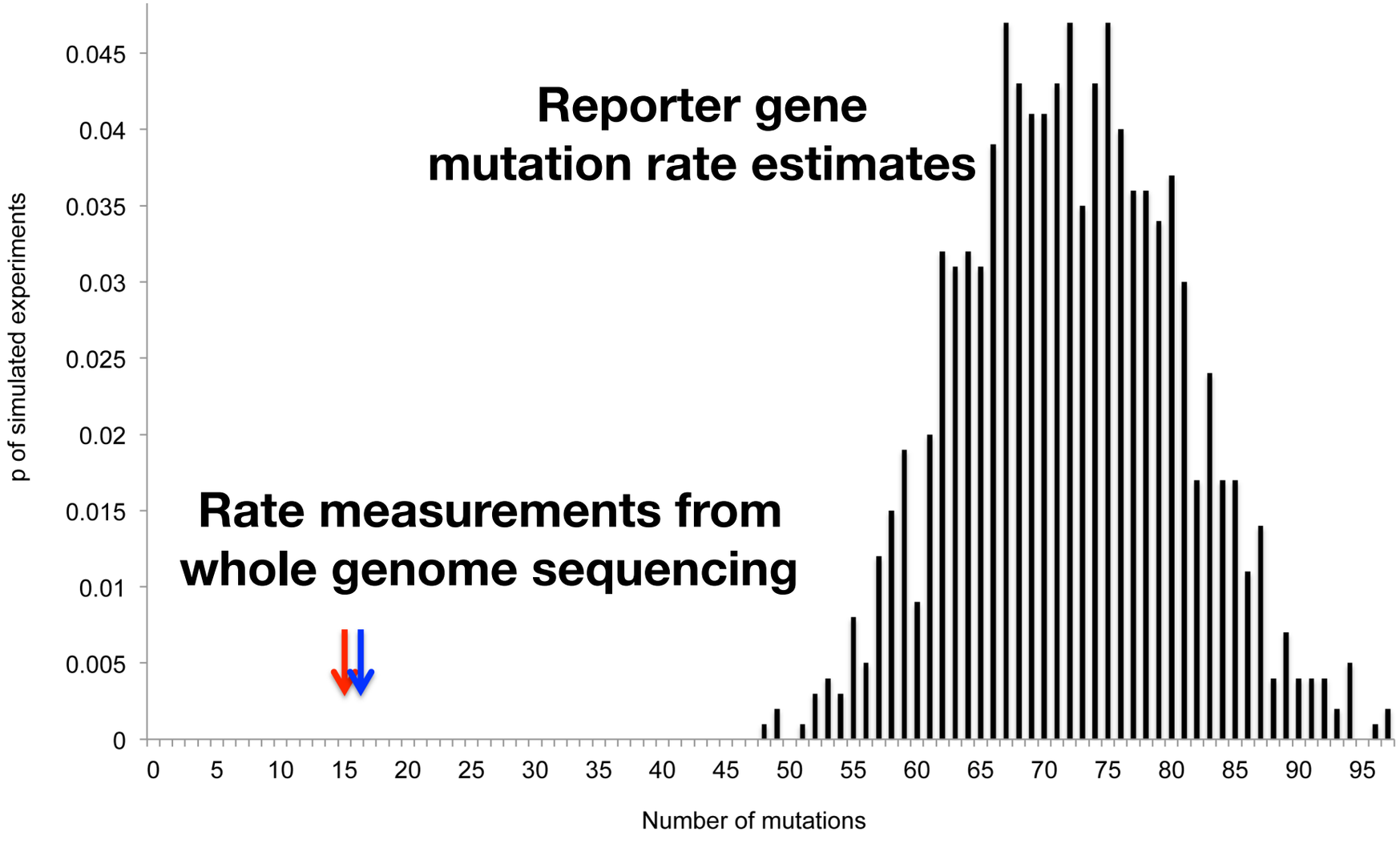
+ See all authors and affiliations

*Science* 30 May 2003:
Vol. 300, Issue 5624, pp. 1404-1409
DOI: 10.1126/science.1082240

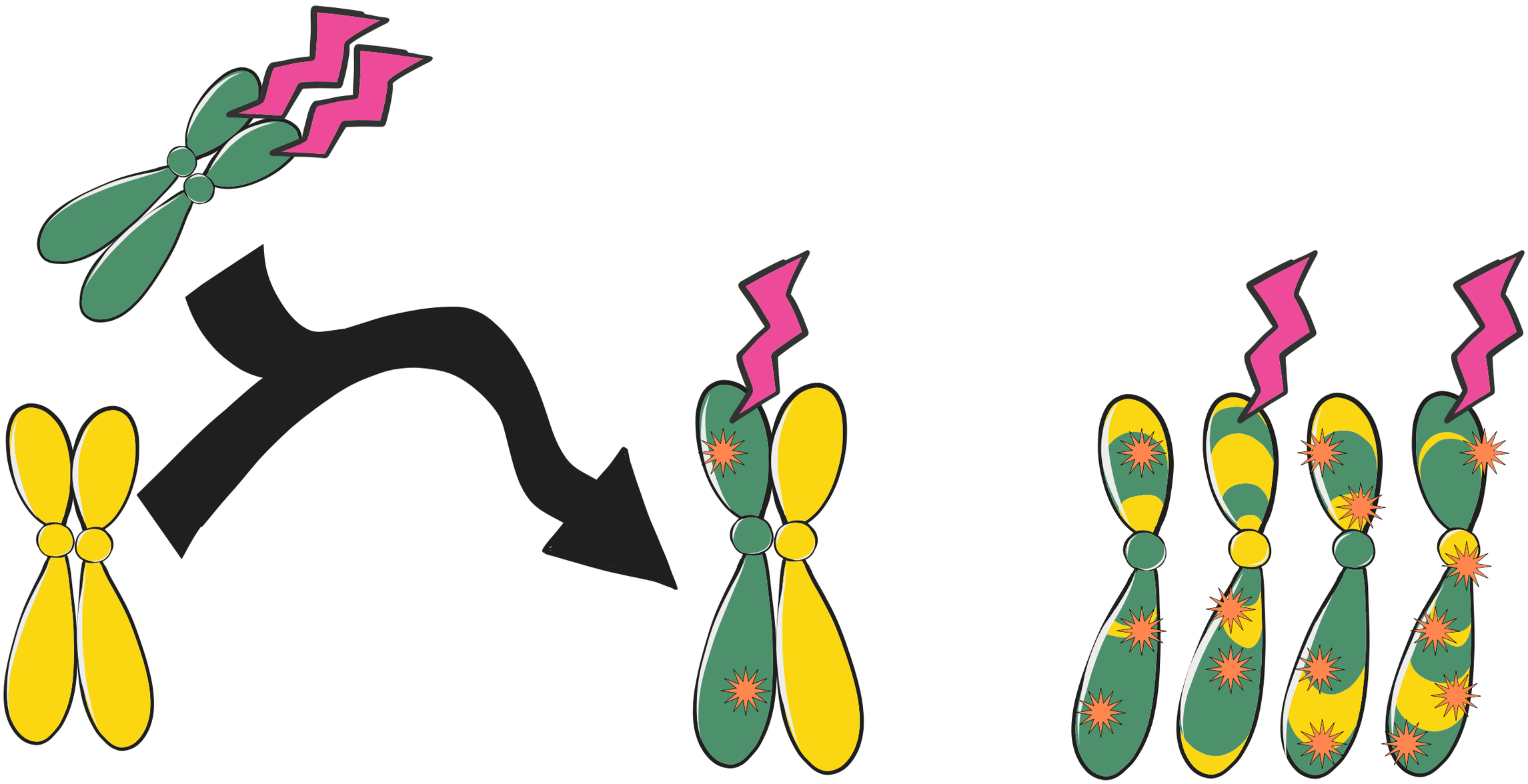
## Elevated Mutagenesis Does Not Explain the Increased Frequency of Antibiotic Resistant Mutants in Starved Aging Colonies

Sophia Katz, Ruth Hershberg

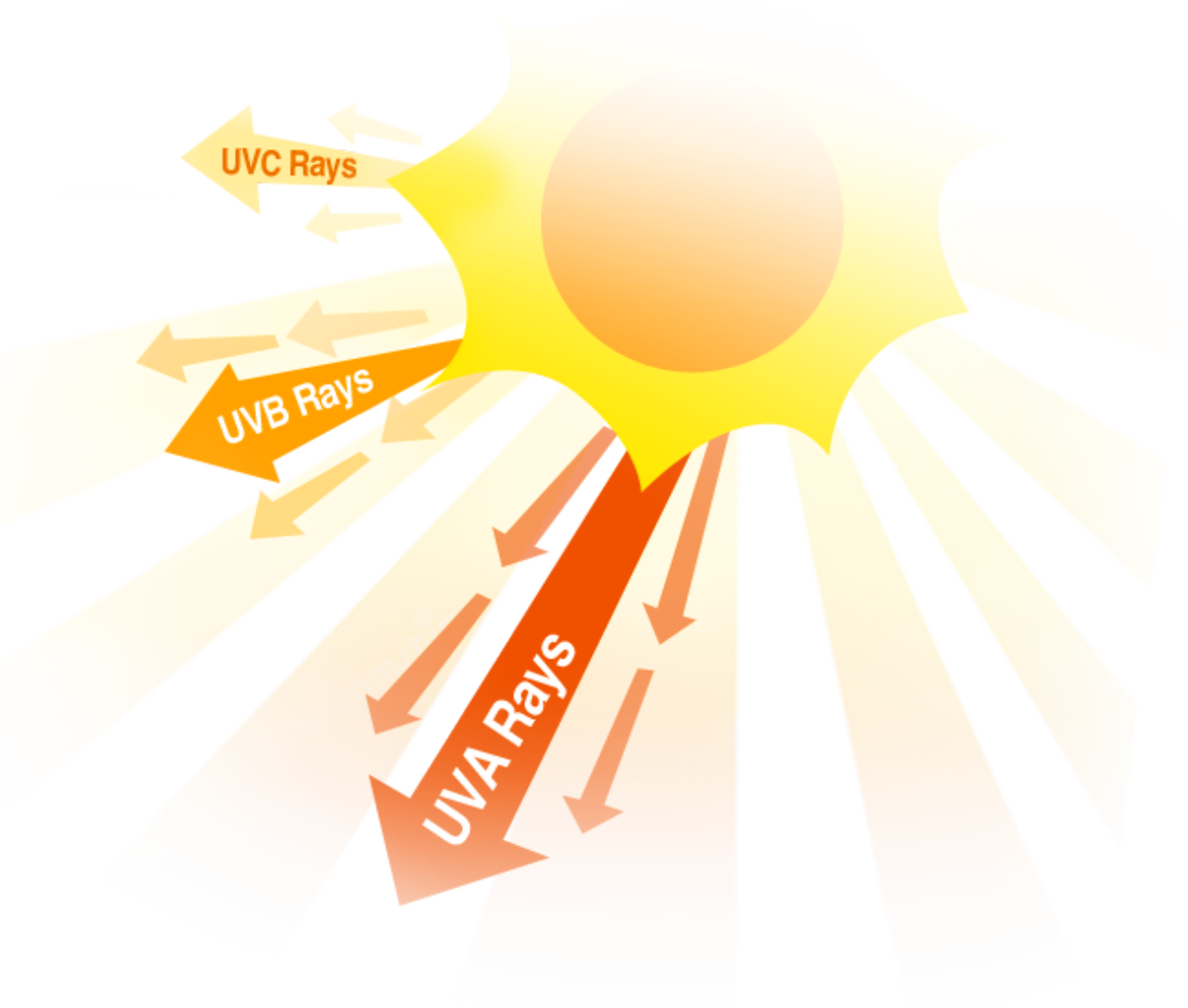
Published: November 14, 2013 • https://doi.org/10.1371/journal.pgen.1003968



# Selection against mutator alleles is weak in sexual organisms



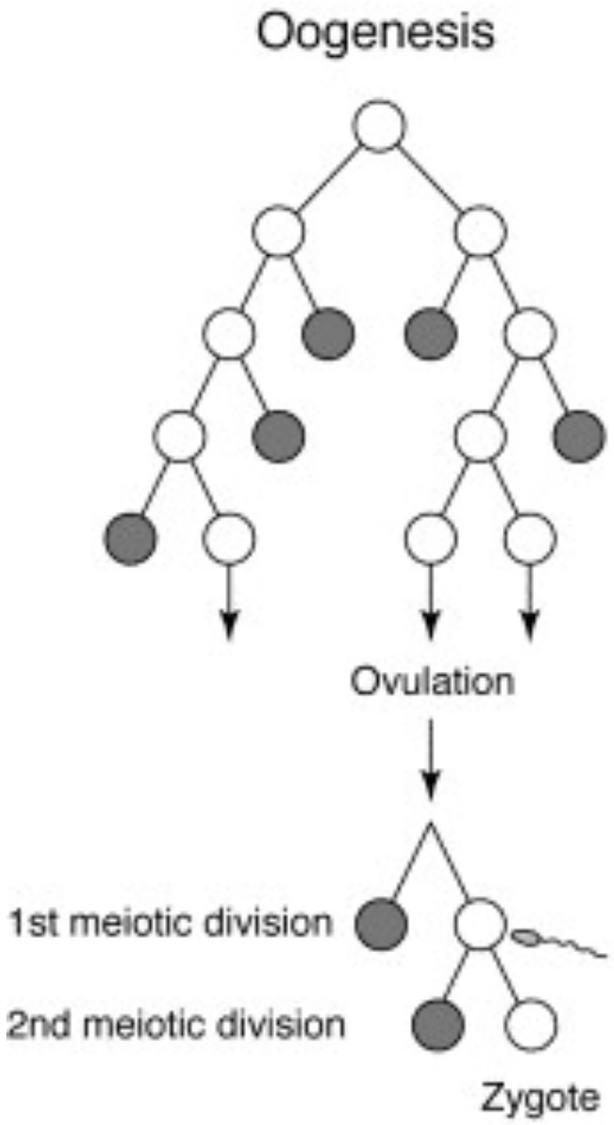
# Other factors affecting the mutation rate



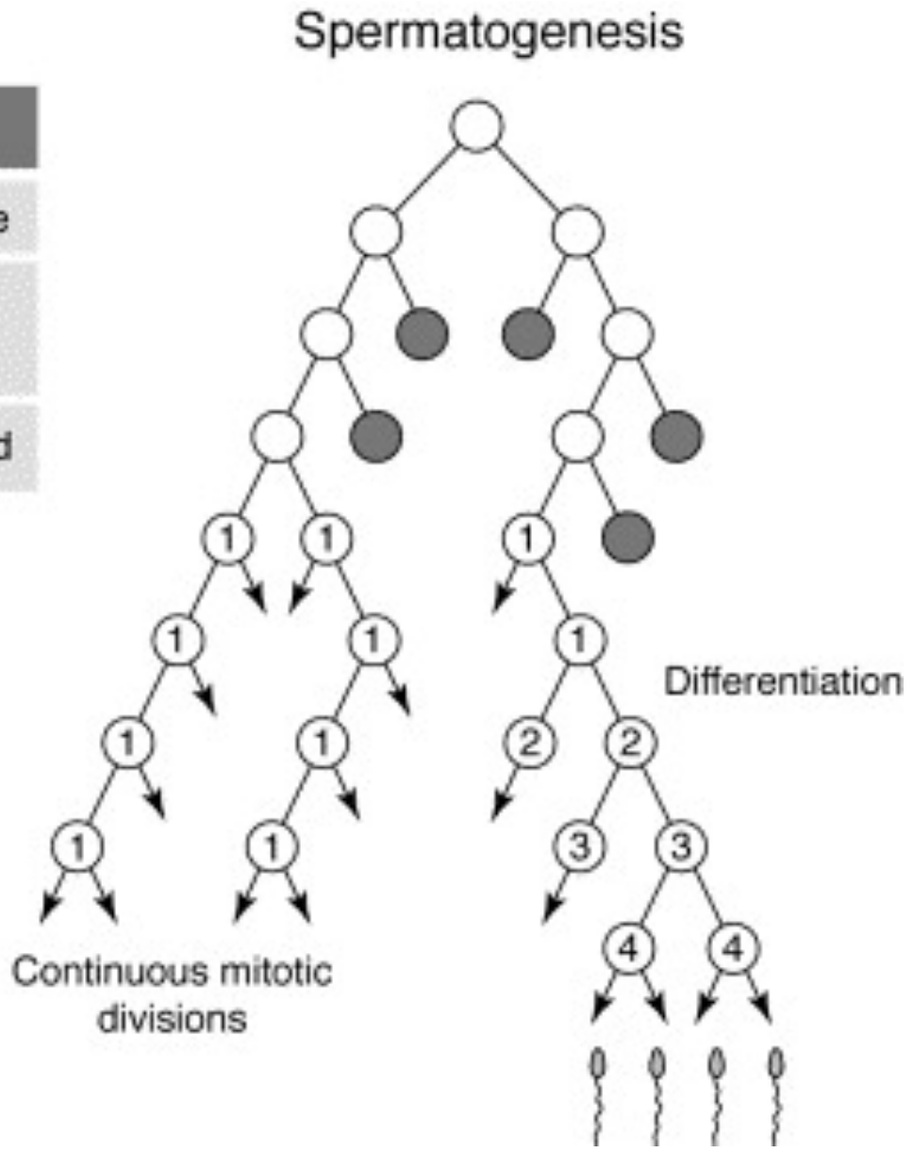

Environmental Mutagens

Life history

# Male mutation bias

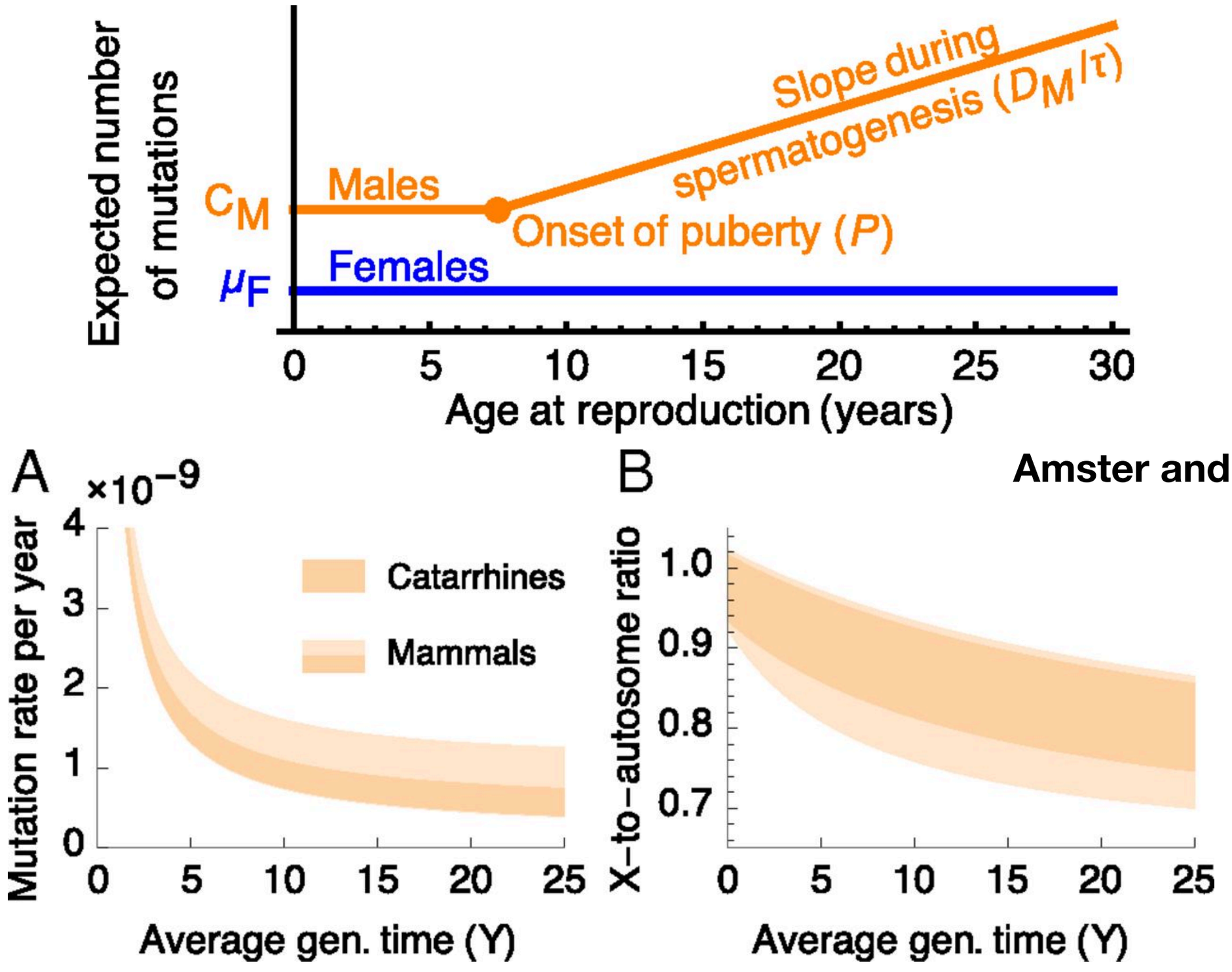



| ♀ | | ♂ | |
|---|---|---|---|
| Timetable | No. of cell divisions | | Timetable |
| 5th month of gestation | 22 | 30 | Puberty |
| Sexual maturity | 2 | 23 per year | Adulthood |
| Total: | 24 | 150 at 20 yr<br>380 at 30 yr<br>610 at 40 yr | |



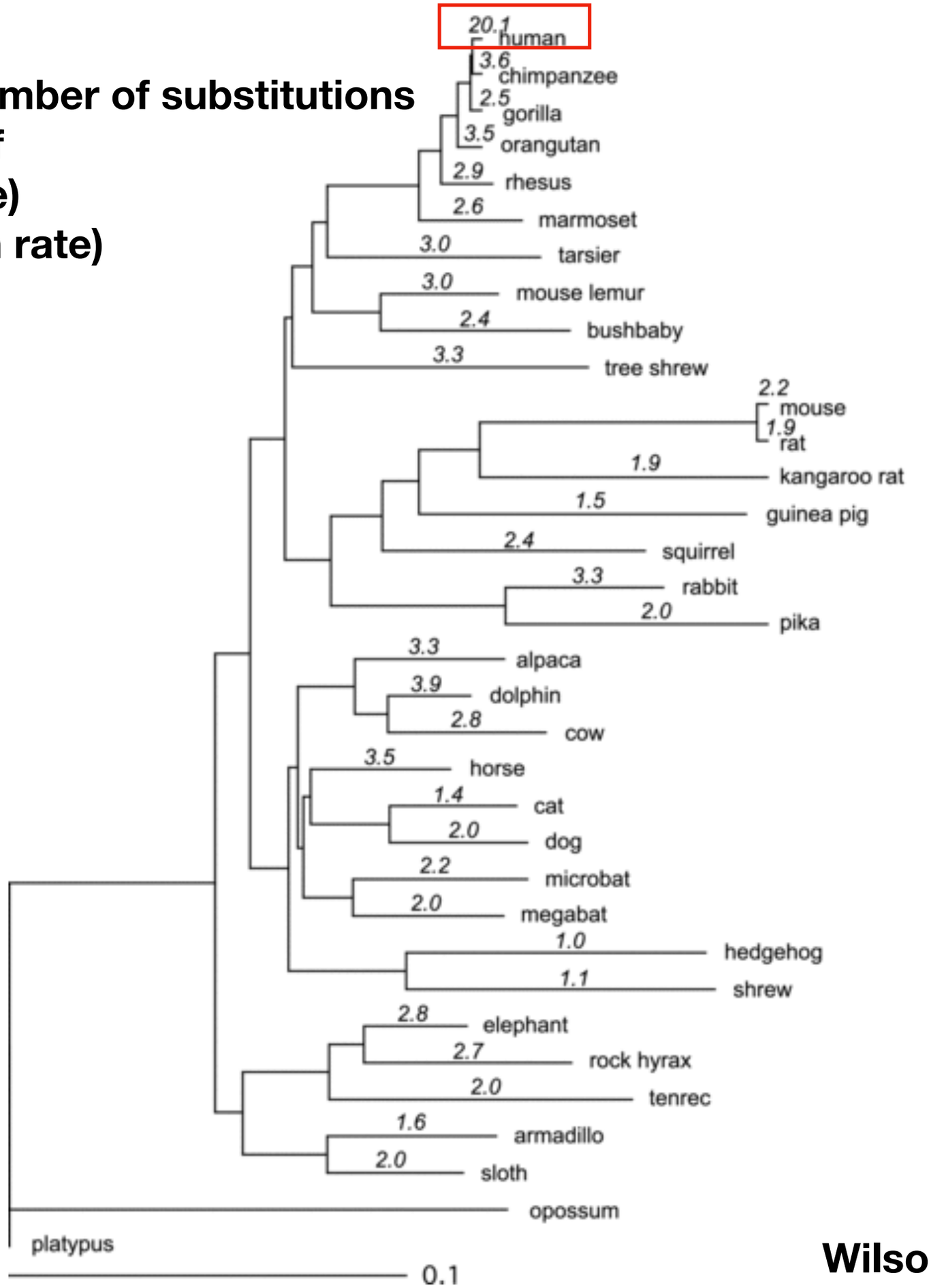

**Hurst and Ellegren 1998**

# Paternal age effect (the classical model)



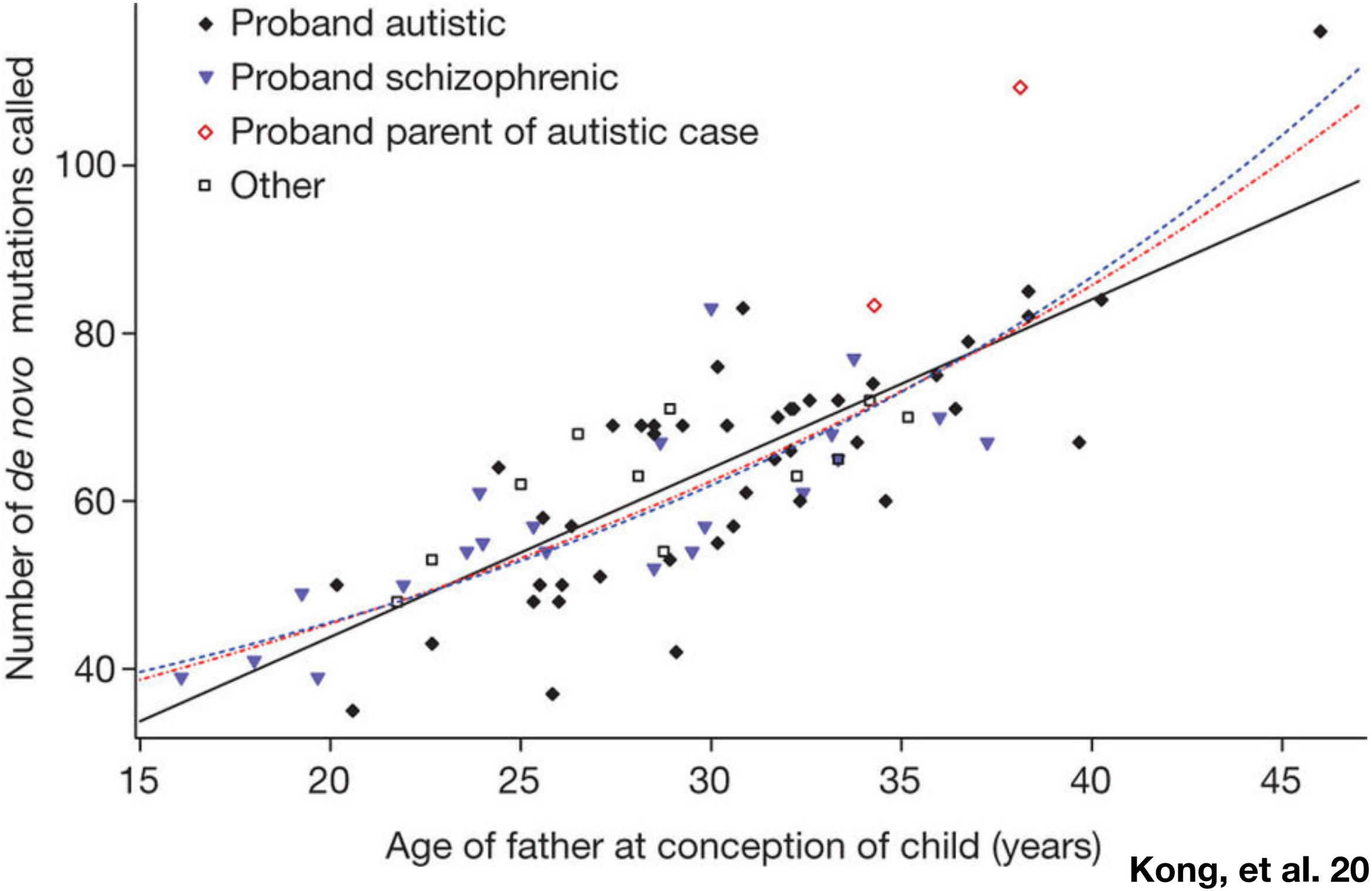

Amster and Sella 2016

**Branch length ~ number of substitutions**
**Label = Estimate of**
**(male mutation rate)**
**/(female mutation rate)**



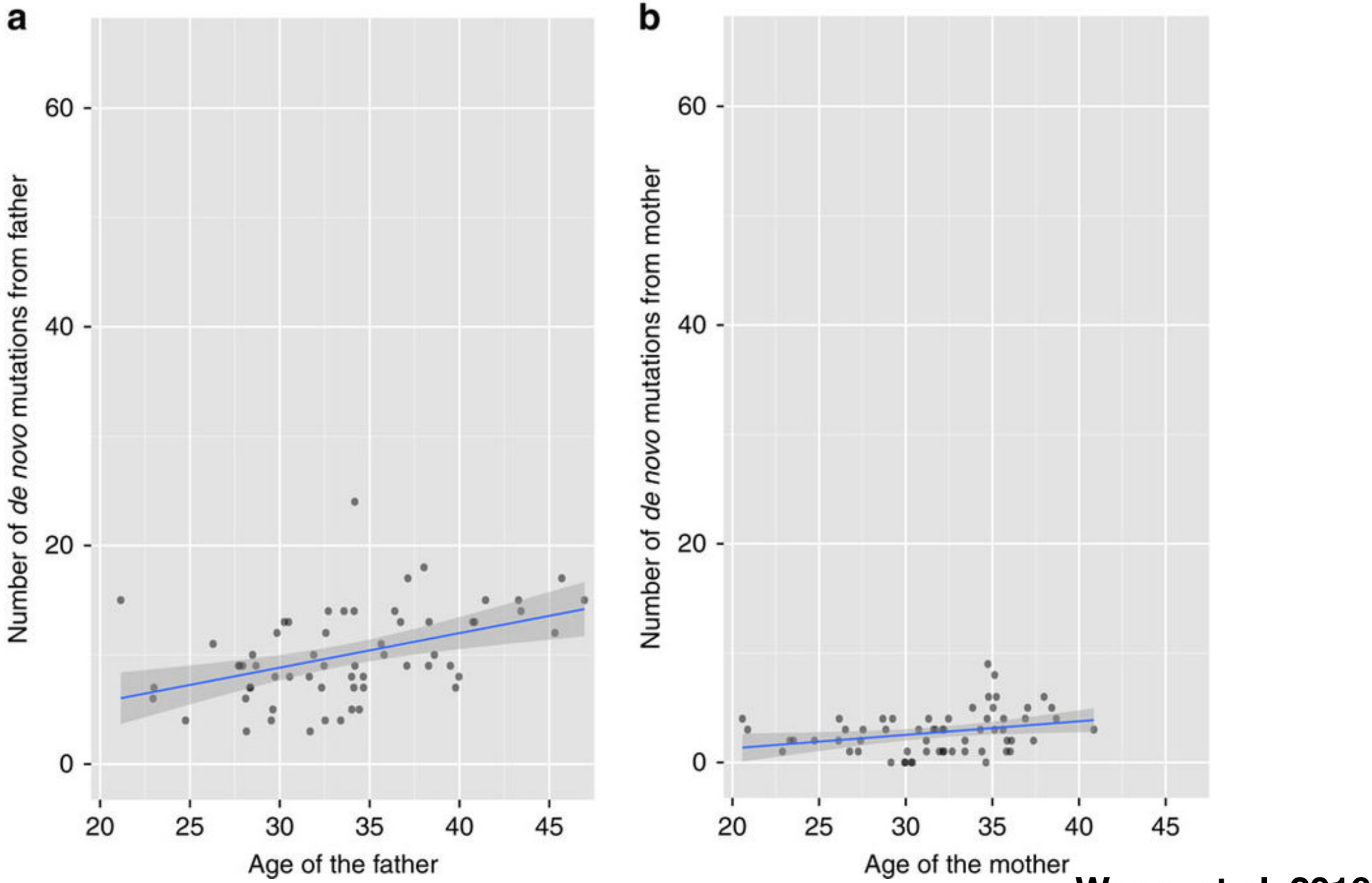

**Wilson Sayres, et al. 2011**

# Two additional *de novo* mutations per year of paternal age



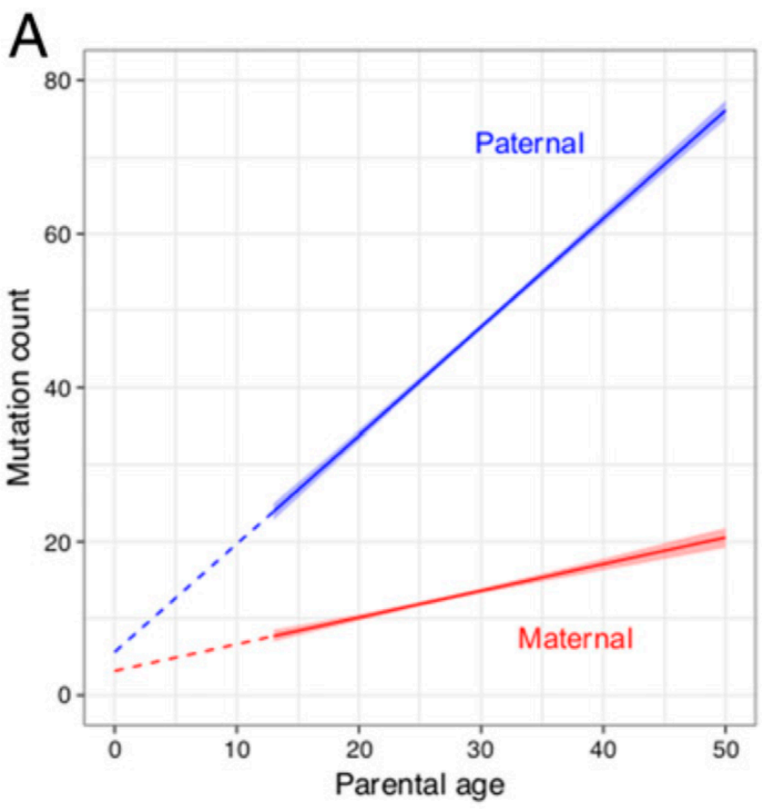

Kong, et al. 2012

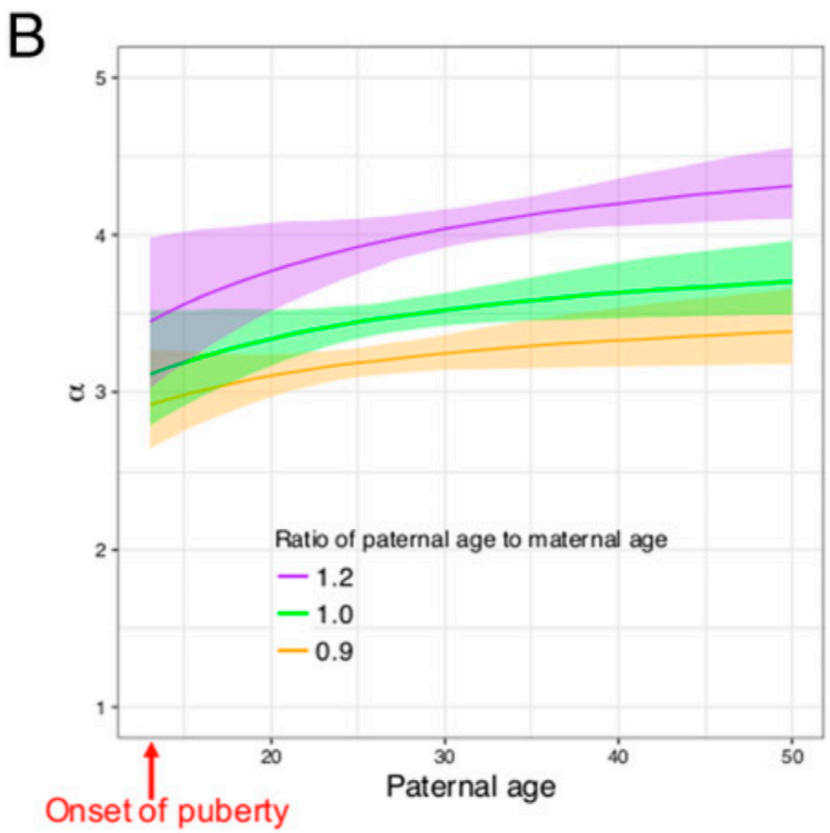
# A small but significant maternal age effect (0.5 muts/year)

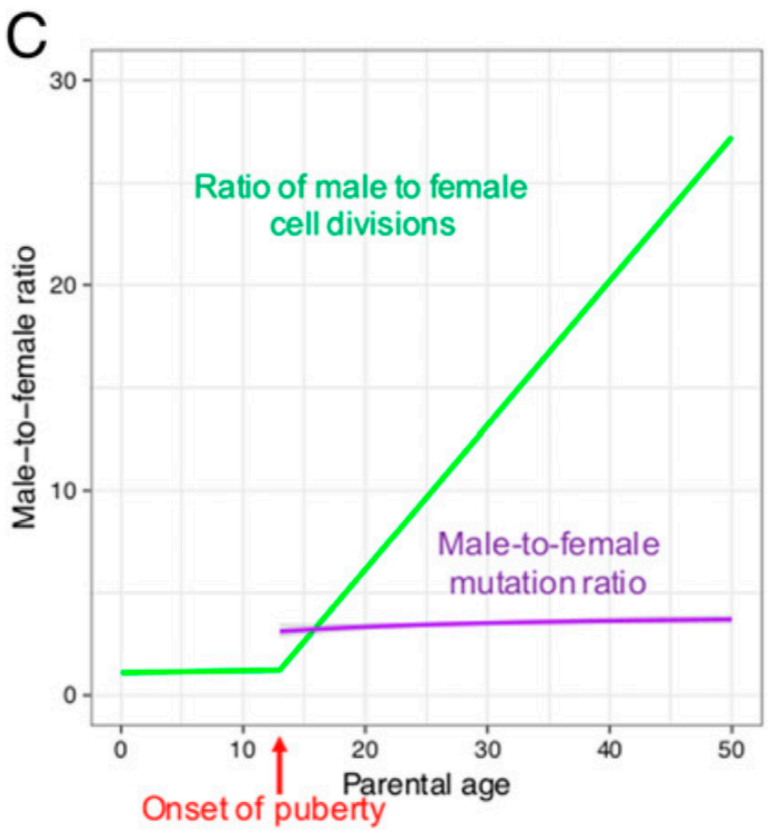



Wong, et al. 2016

# If spermatocyte replication causes the paternal age effect, the fraction of paternal mutations should increase with parental age

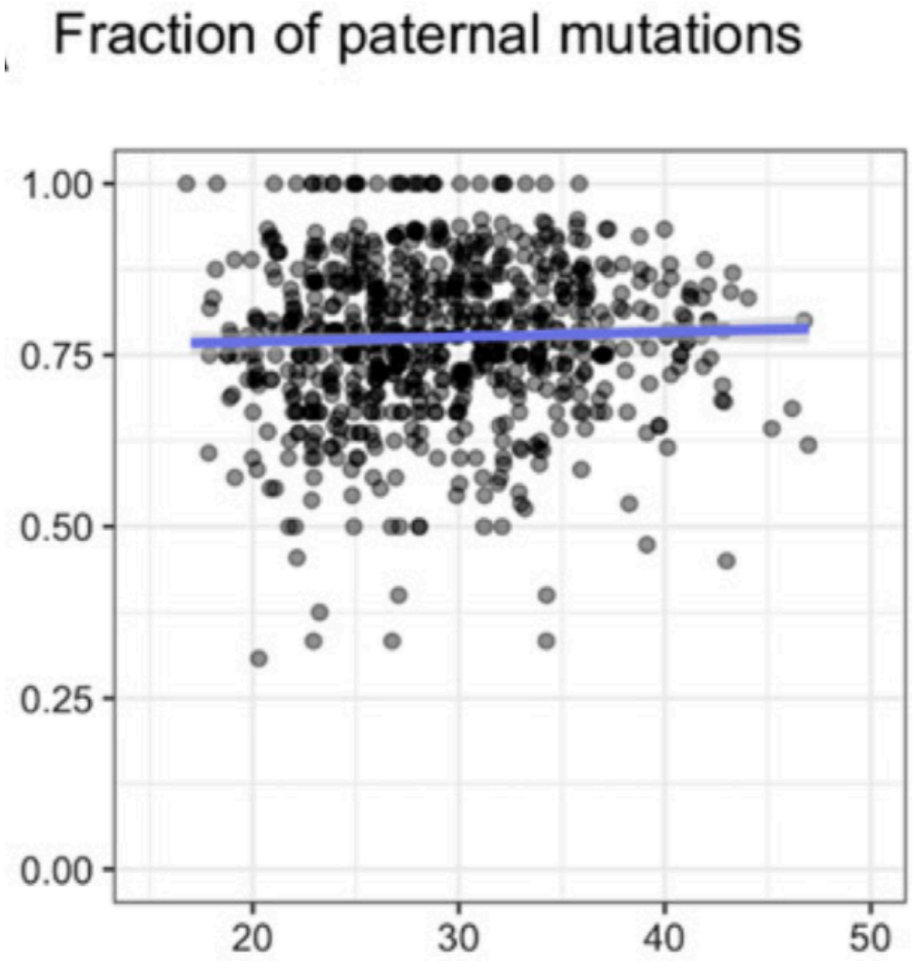



Gao, et al. 2019

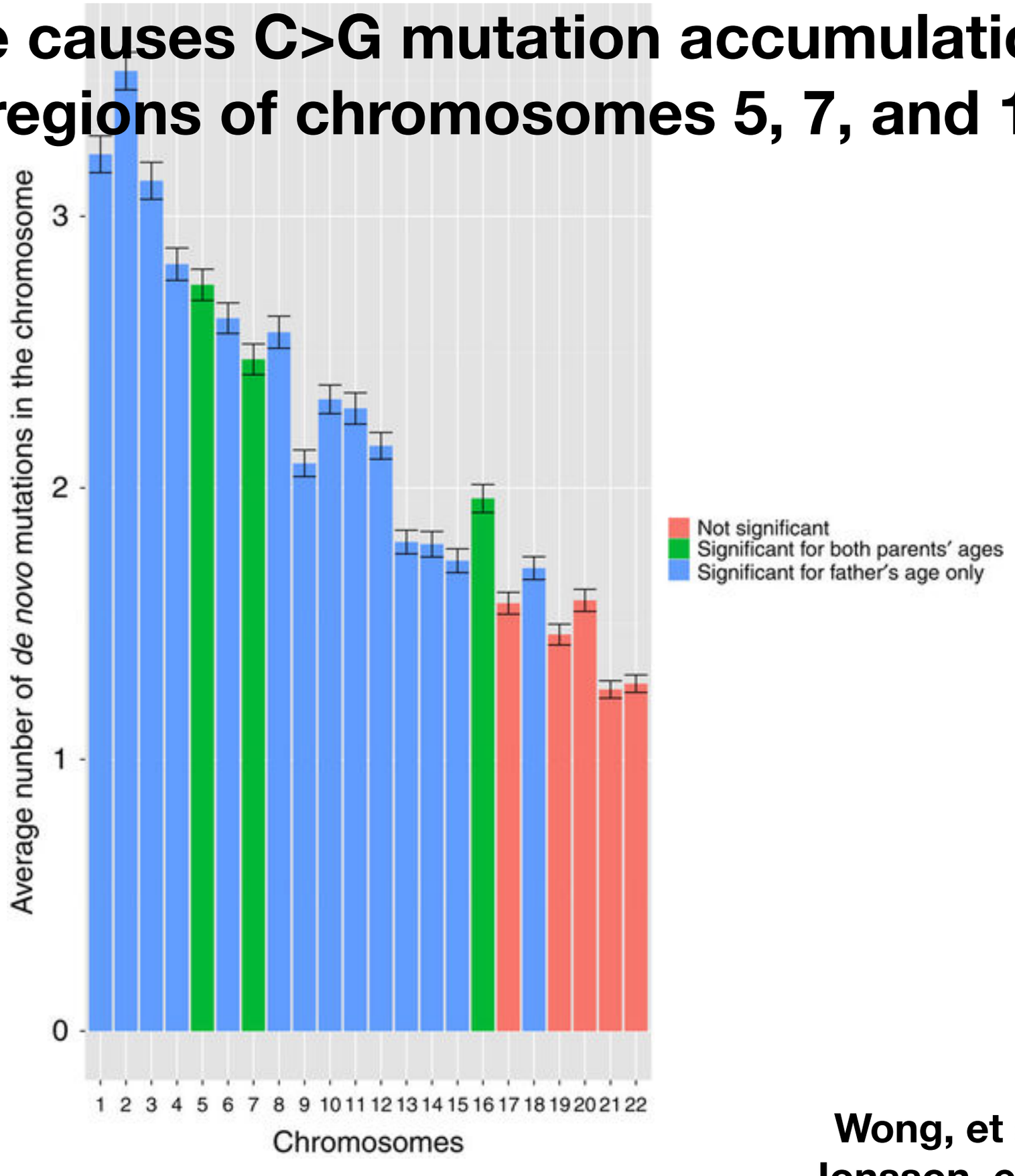
# Human trio data now contradict this prediction




## Overlooked roles of DNA damage and maternal age in generating human germline mutations

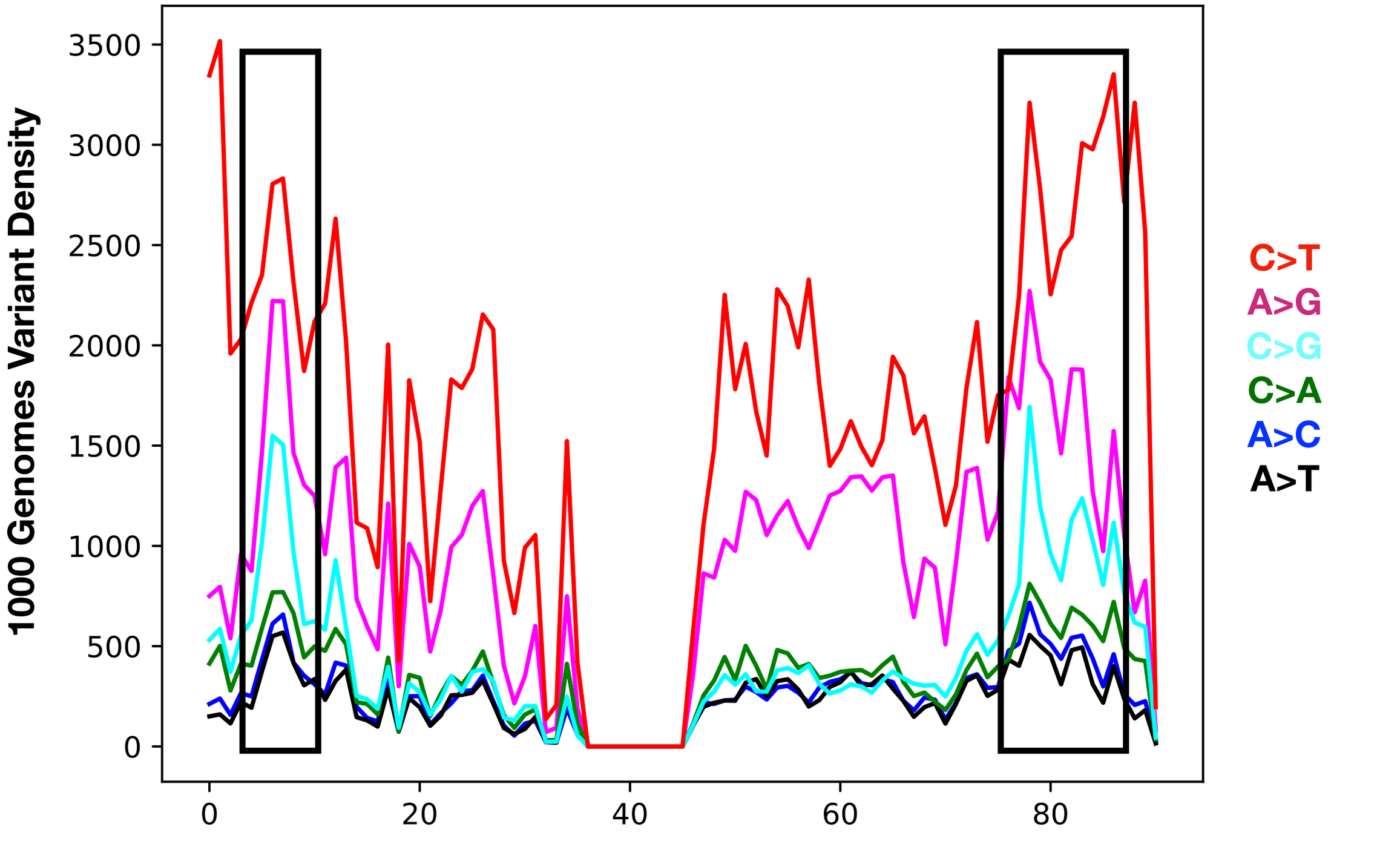
Ziyue Gao[a,b,1], Priya Moorjani[c,d], Thomas A. Sasani[e], Brent S. Pedersen[e], Aaron R. Quinlan[e,f], Lynn B. Jorde[e], Guy Amster[g,2], and Molly Przeworski[g,h,1,2]

# Maternal age causes C>G mutation accumulation in localized regions of chromosomes 5, 7, and 16



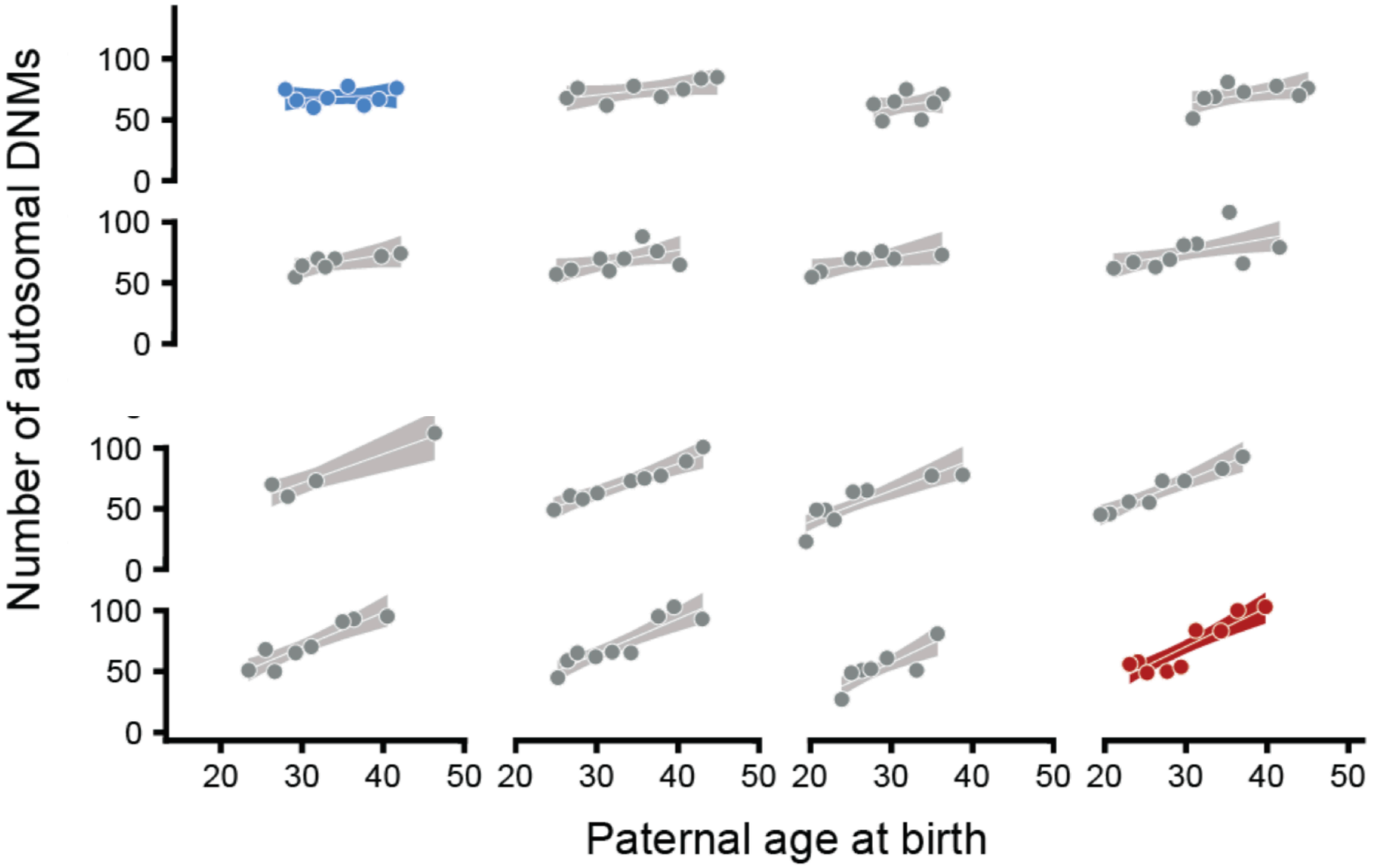

Wong, et al. 2016
Jonsson, et al. 2017

# Maternal age causes C>G mutation accumulation in localized regions of chromosomes 5, 7, and 16




Wong, et al. 2016

# Large CEPH families reveal variability in paternal age effect between families

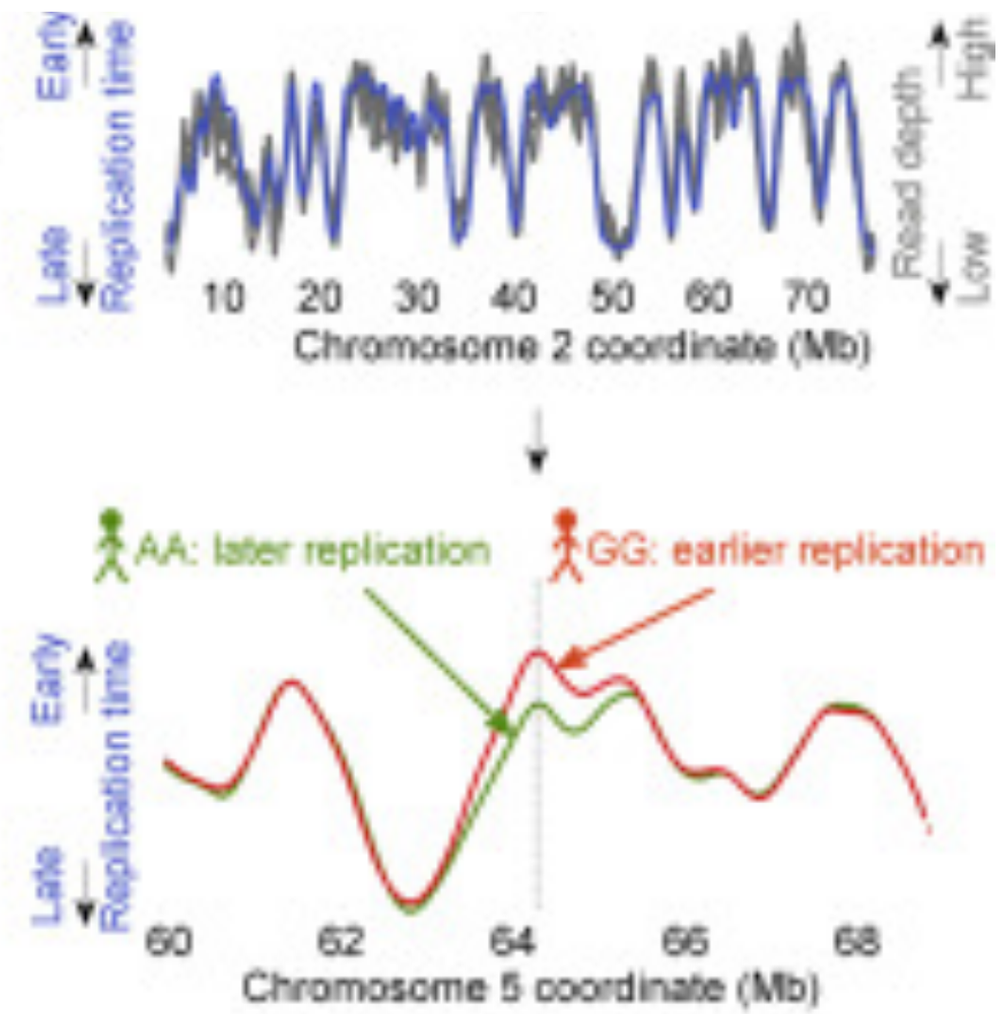



Sasani, et al. *eLife* 2018

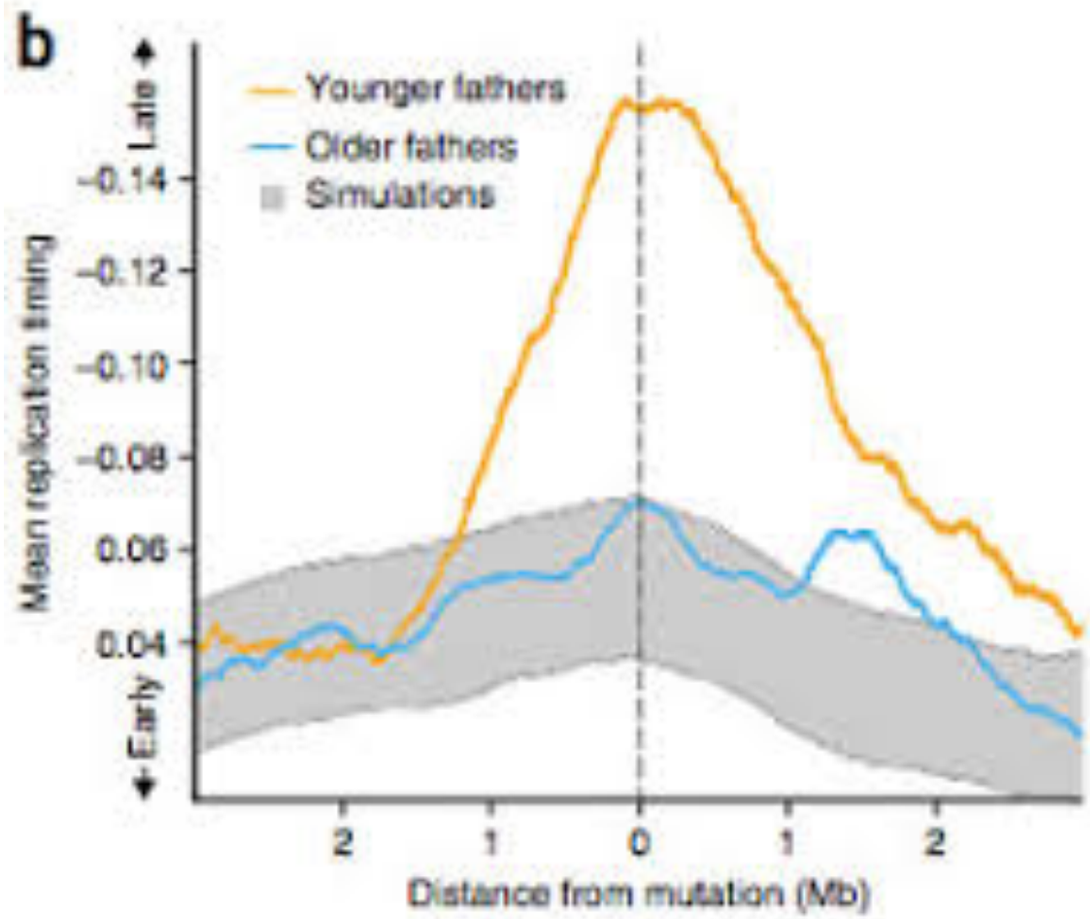
# Other causes of mutation rate variation along the genome

- Replication timing
- Transcription-associated-mutagenesis (TAM) and transcription-coupled-repair (TCR)
- Non-B-DNA structures and other DNA repeats
- Chromatin state

# Replication timing

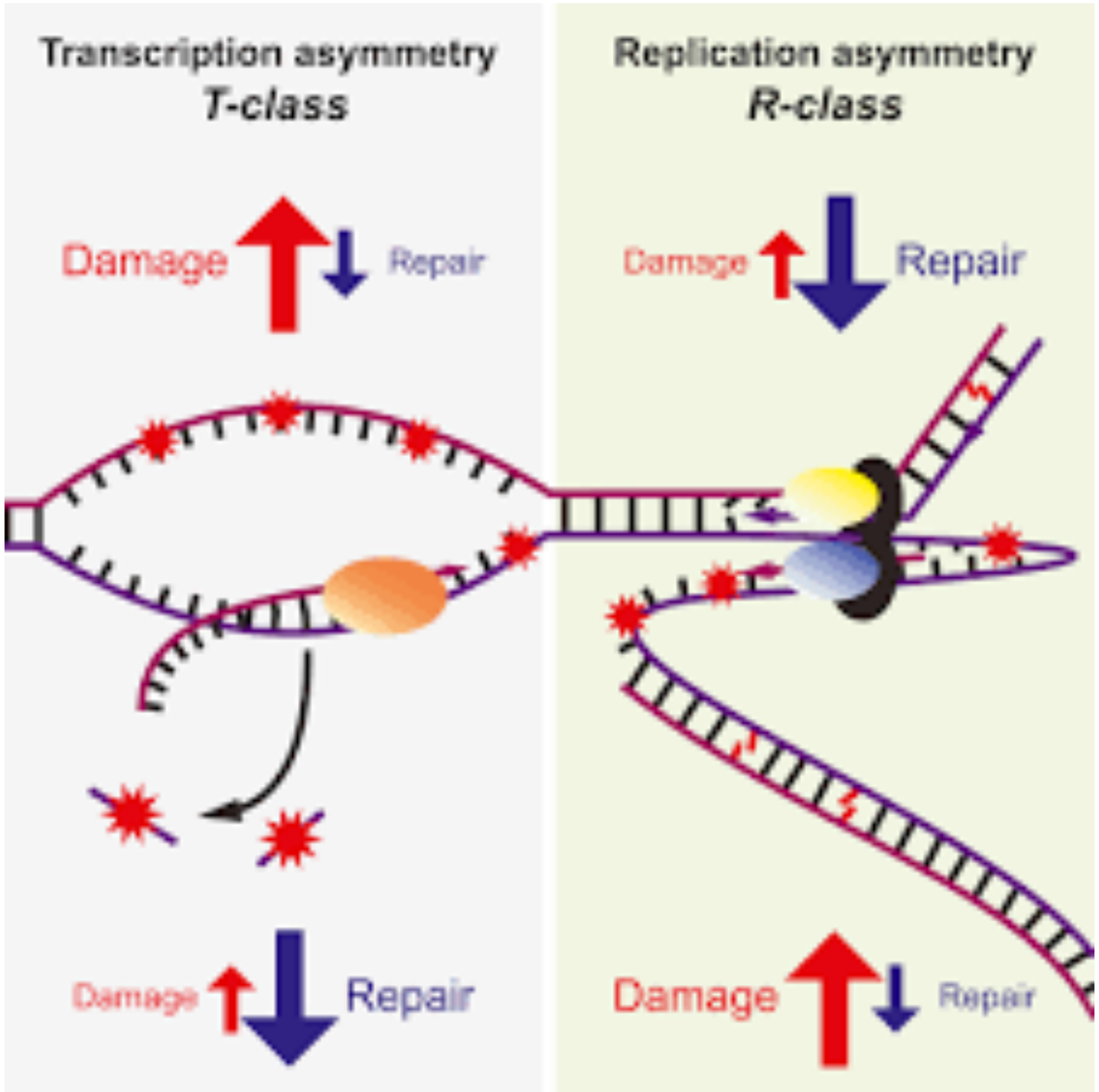



Koren, et al. 2012




Francioli, et al. 2015

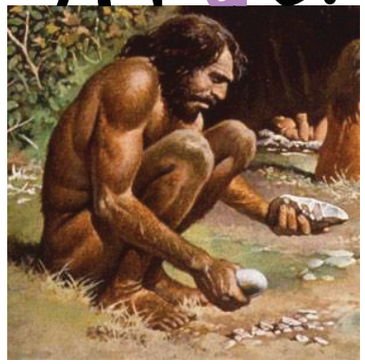
# Replication and transcription induce strand asymmetry



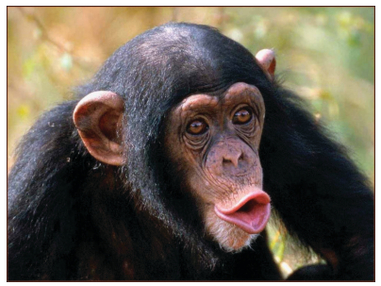

**Excess of G+T over A+C on coding strand of most genes**

**Haradhvala, et al. 2016**
**Green, et al. 2003**

# Measuring the human mutation rate

Human



Chimpanzee

Nachman and Crowell 2001

2.5*e*-8 mutations per site per gen

mgr.com.my

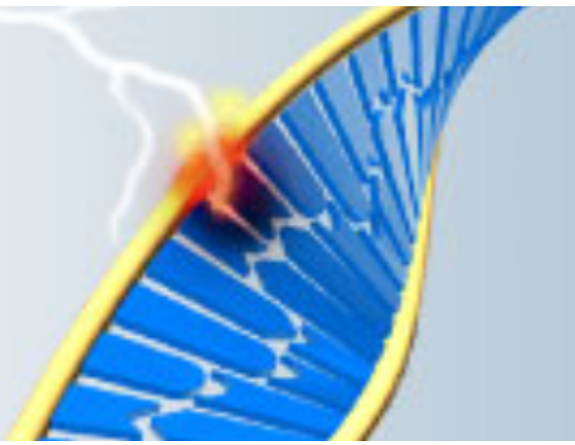
Parent-child trios

1000 Genomes Consortium 2010

1.0*e*-8 mutations per site per gen

The Human Mutation Rate Meeting

Leipzig, 25th - 27th February 2015



*NATURE* | NEWS

# DNA mutation clock proves tough to set

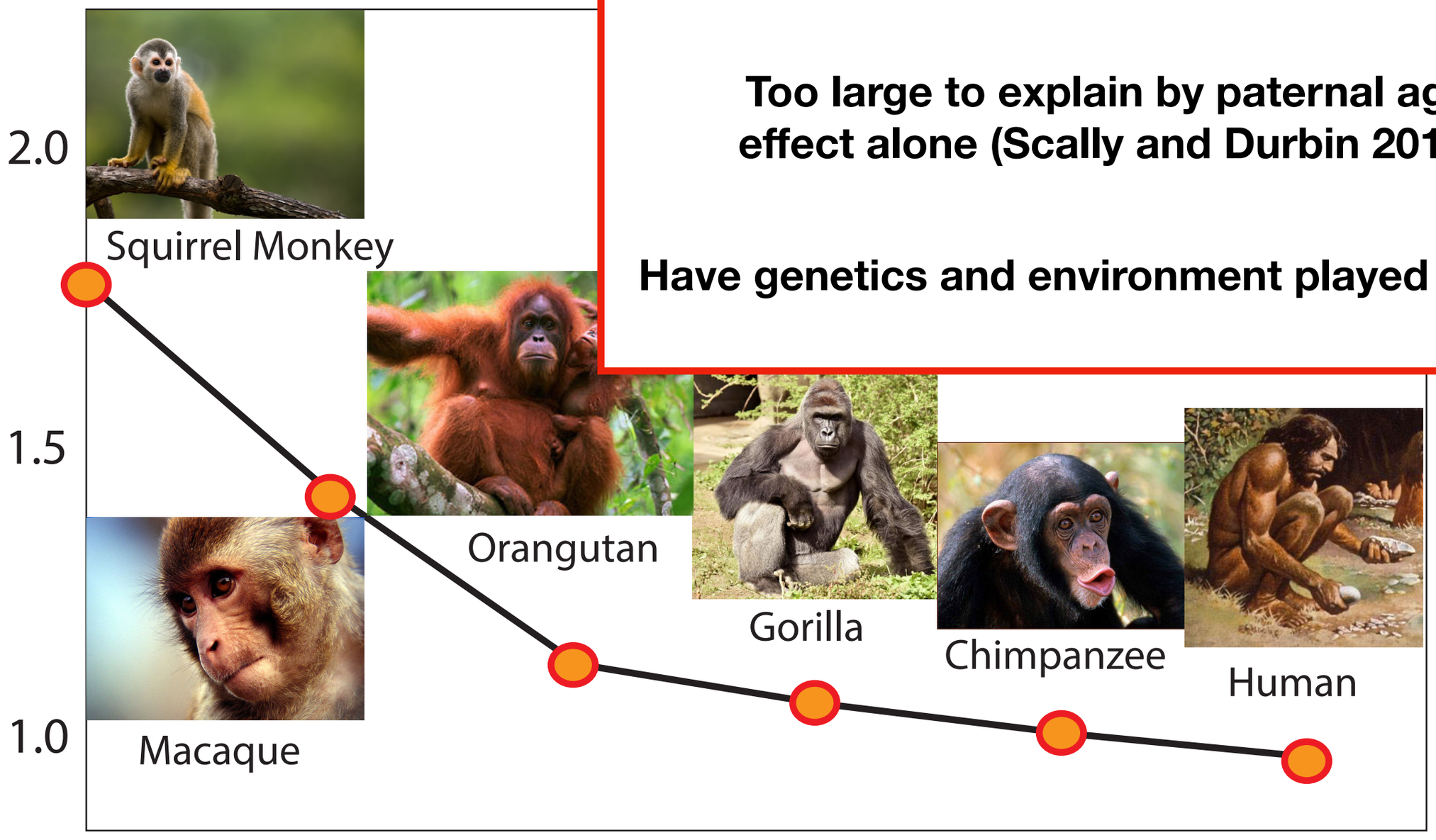
**Geneticists meet to work out why the rate of change in the genome is so hard to pin down.**

**Ewen Callaway**

10 March 2015

- What is the real human mutation rate?
- Has the mutation rate slowed down during recent human history?

# The Hominoid Mutation Rate Slowdown




**Too large to explain by paternal age effect alone (Scally and Durbin 2012)**

**Have genetics and environment played a role?**

Adapted from http://www.bio.indiana.edu/graduate/multidisciplinary/GCMS/trainees/thomas_gregg.php

Goodman *BioEssays* 1985 Moorjani, et al. *PNAS 2016*

# “The” mutation rate encompasses a menagerie of mutation types

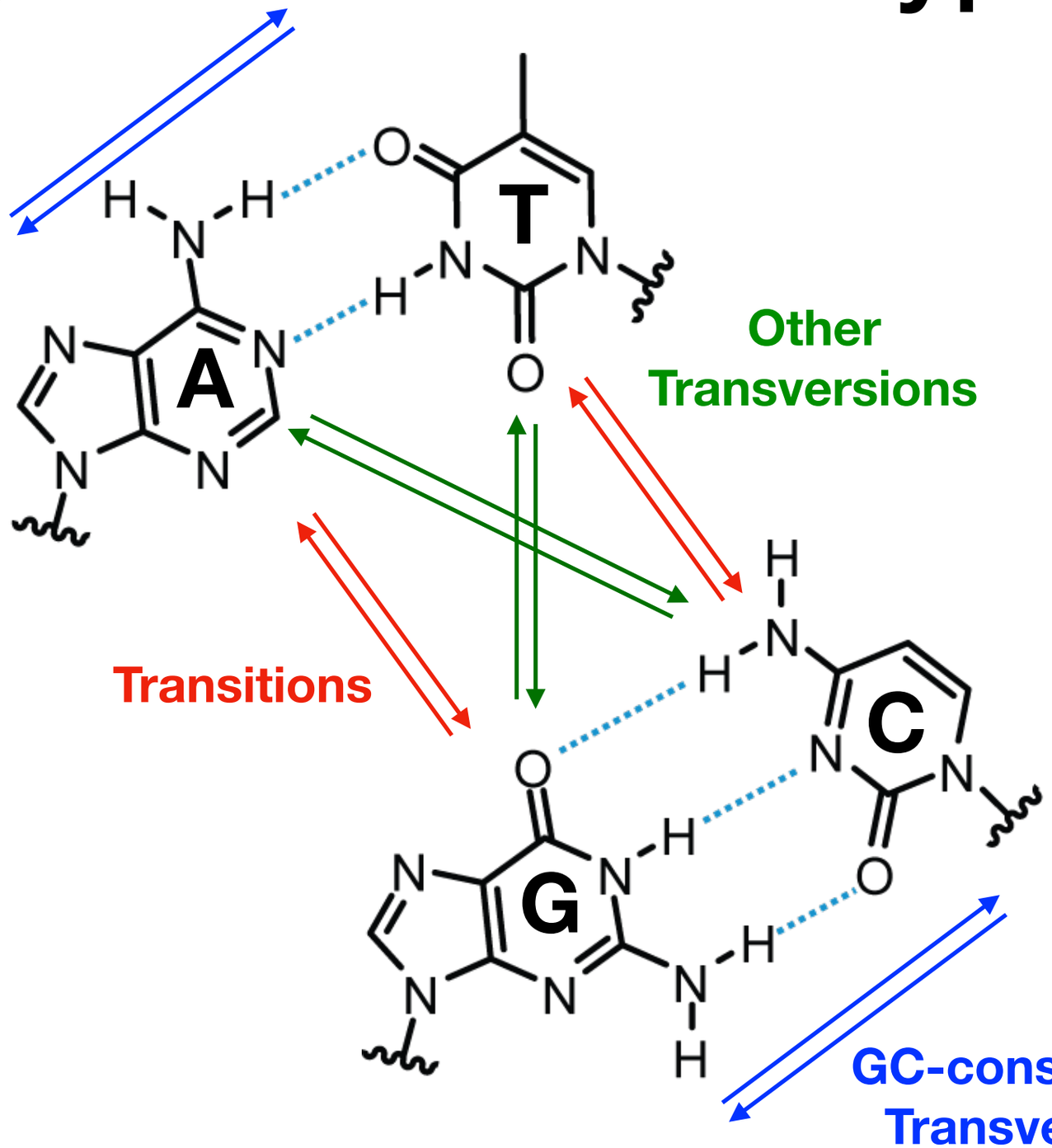
**Point Mutations**

**Multinucleotide Mutations**

CC ⟶ TT

**Small indels**

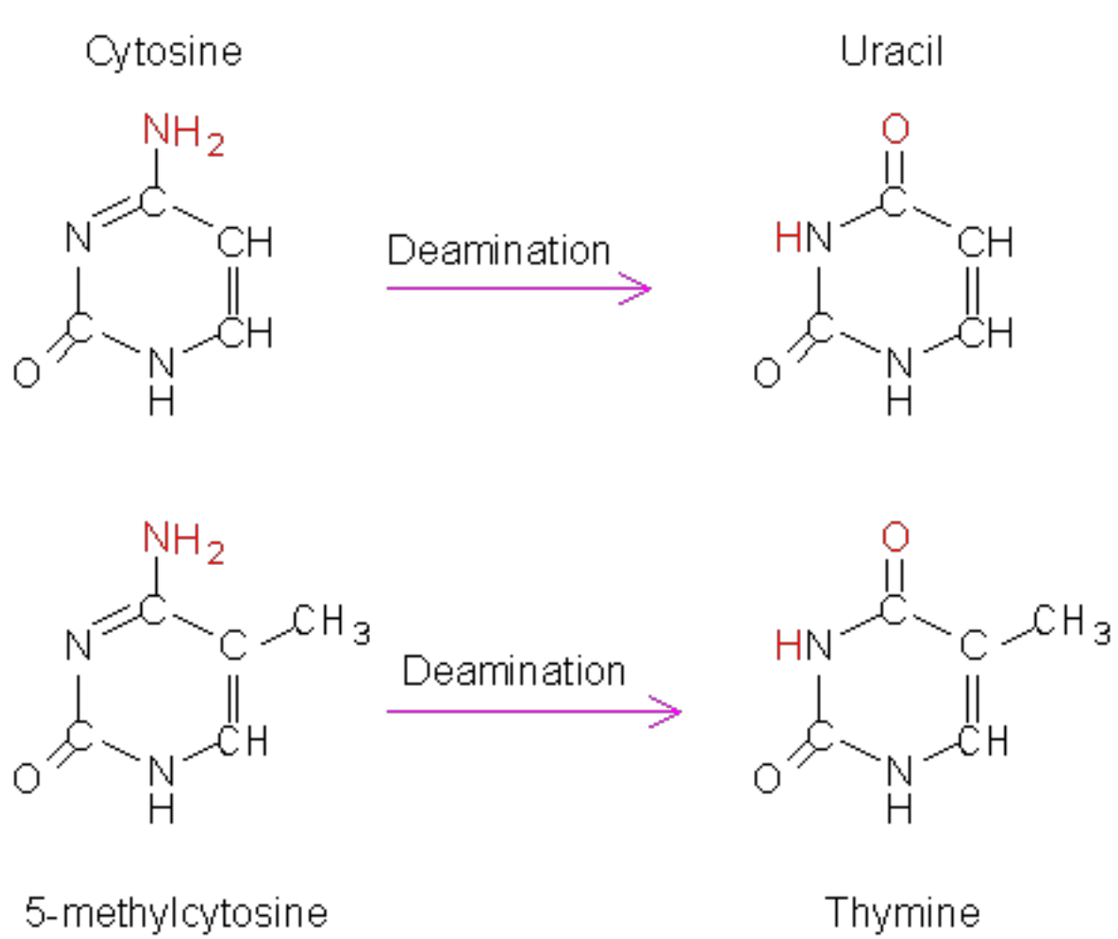
**Large Copy Number Changes**



# CpG Mutations

- Many species (incl humans, not incl *Drosophila*) methylate C when it's next to G (C-phosphate-G)

- CpG methylation regulates gene expression



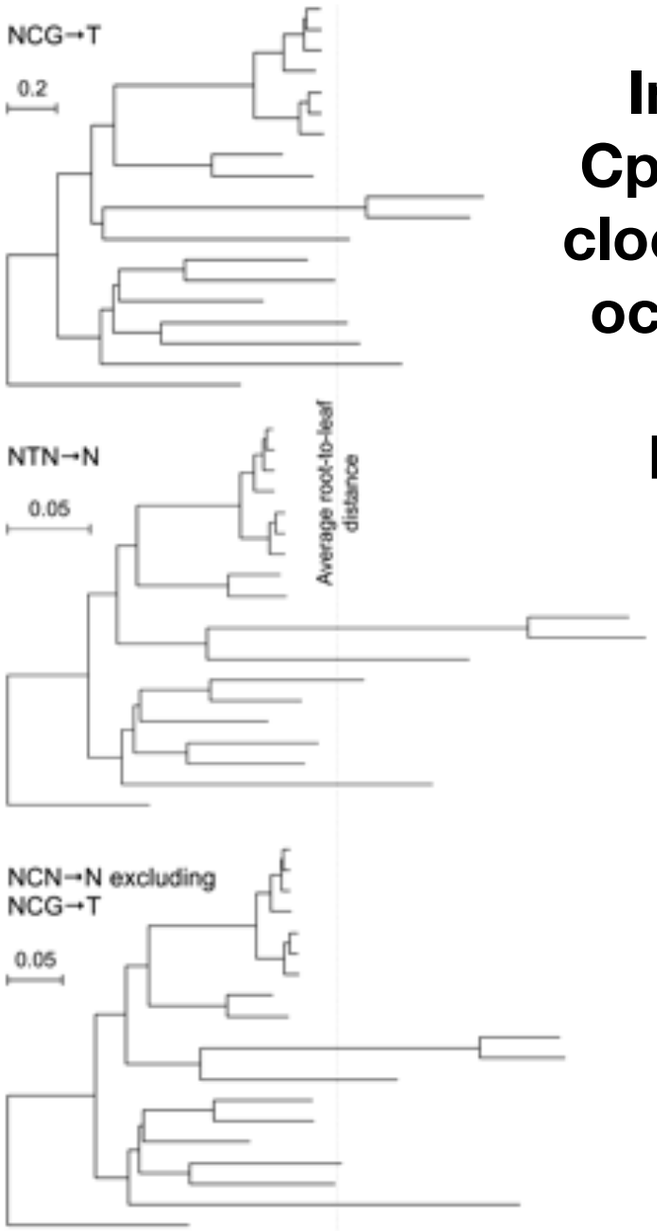
# CpG sites are hypermutable

- On average, CpG sites have a 30-fold higher mutation rate than other C's in the human genome

- 70-80% of CpGs are methylated in mammals; most unmethylated CpGs are part of CpG islands

- Fewer than 1% of dinucleotides in the human genome are CpGs, although the expected frequency is 0.21*0.21=4.41%

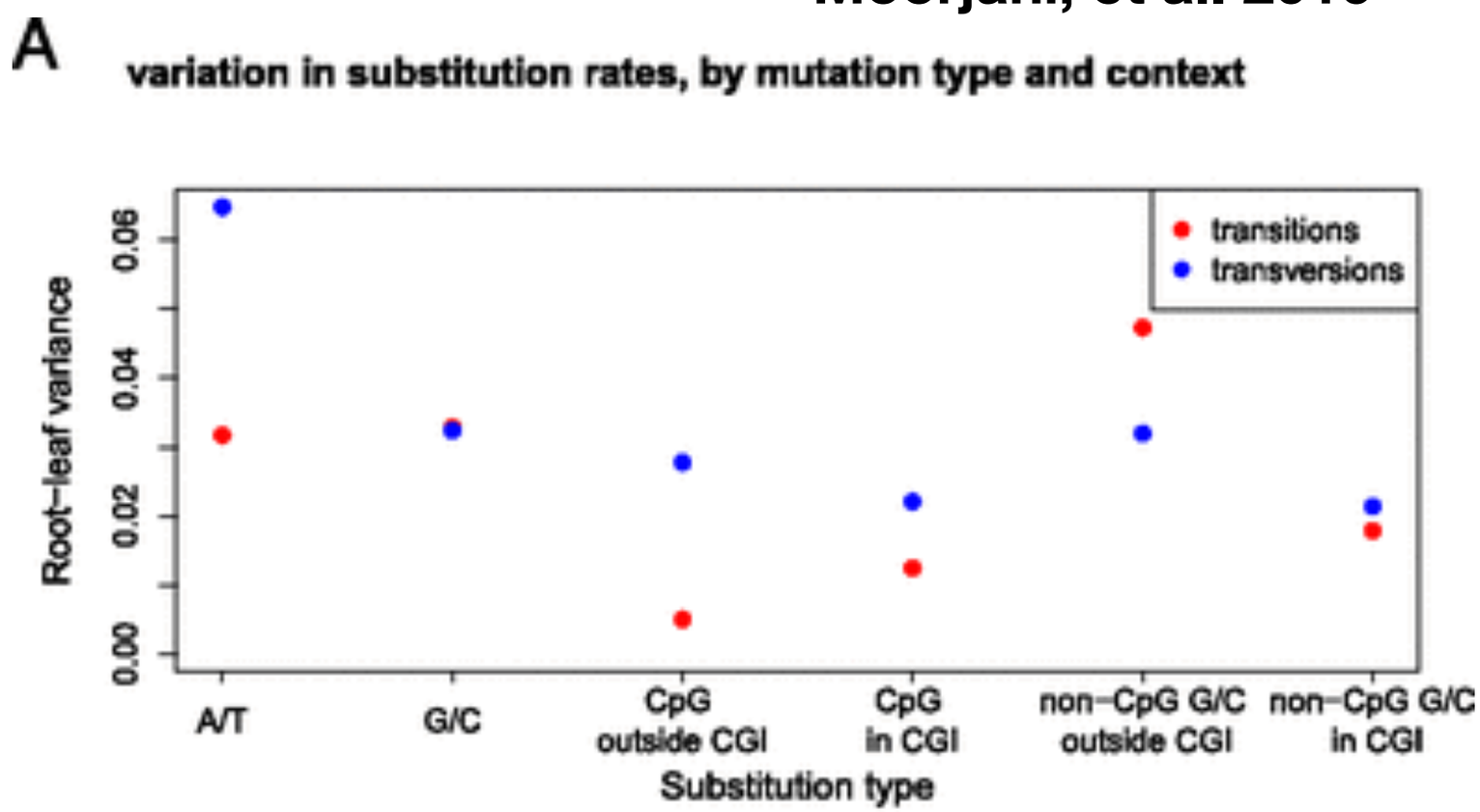
# CpG transitions are somewhat more clocklike than other mutations



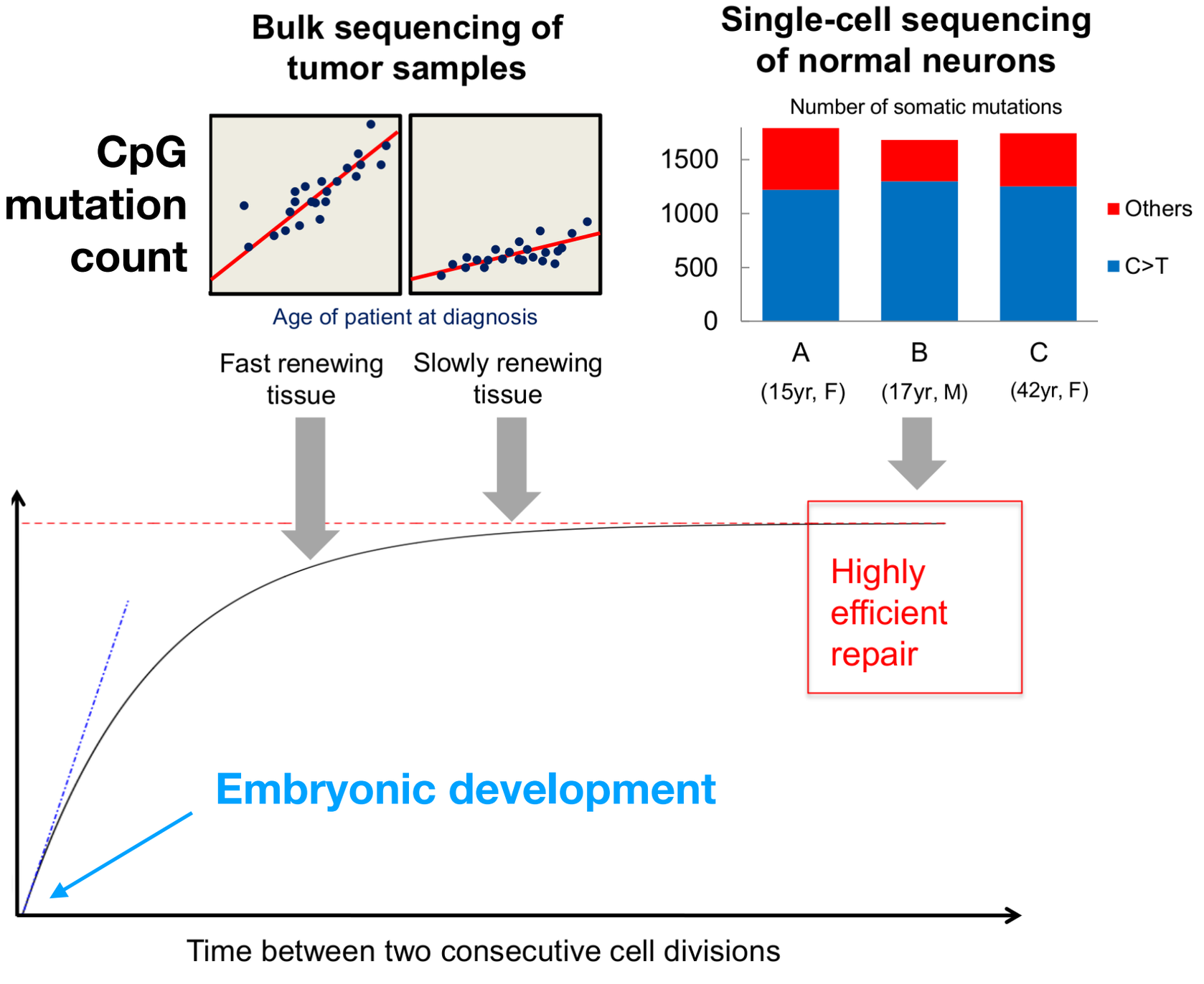

**In a tree of 19 mammals, CpG mutations yield a more clocklike tree than mutations occurring in other contexts**

**Hwang and Green 2004**

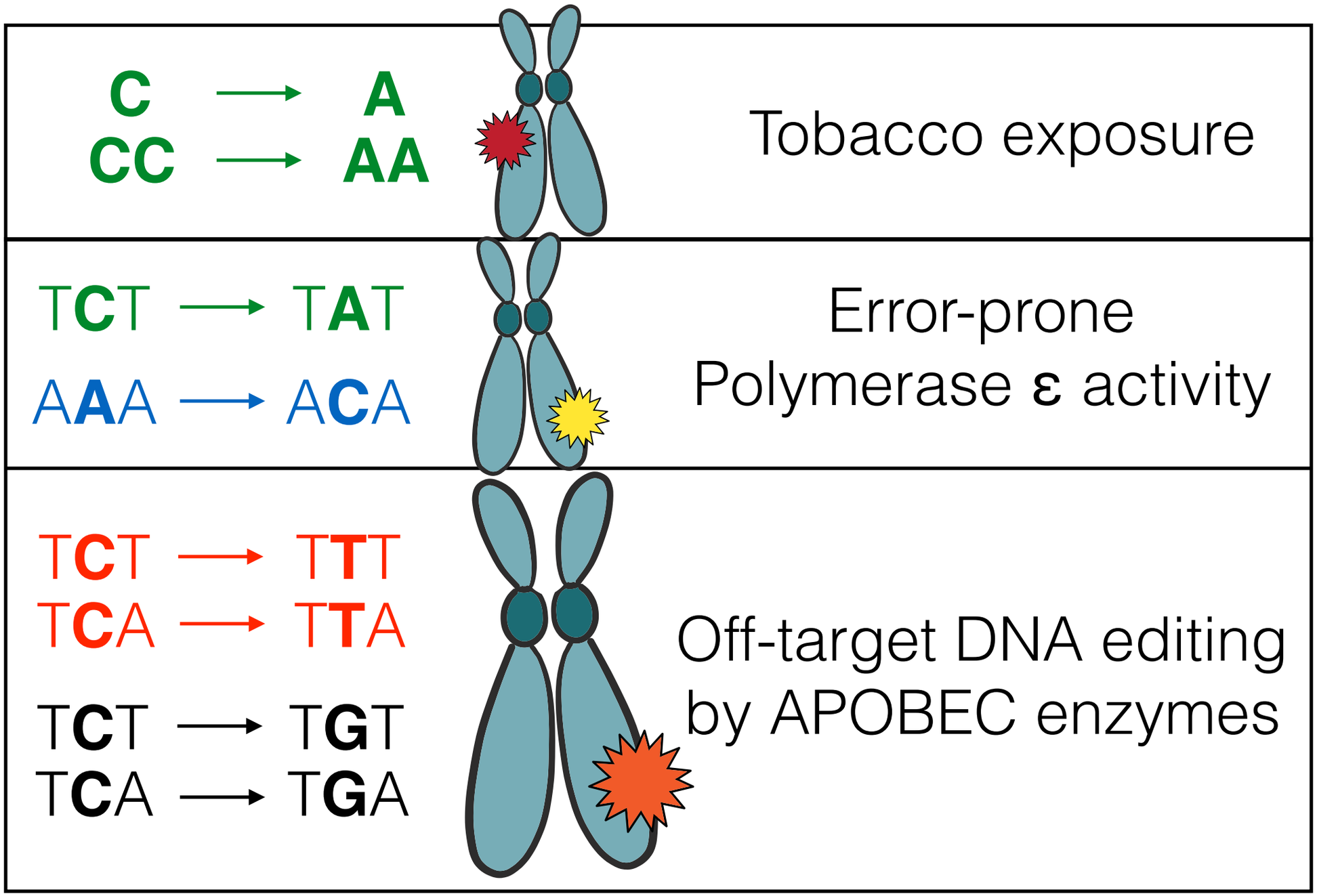
**CpG mutations also appear more clocklike than other mutations in great ape tree**
**Moorjani, et al. 2016**



# Limits to clock-like behavior of CpGs



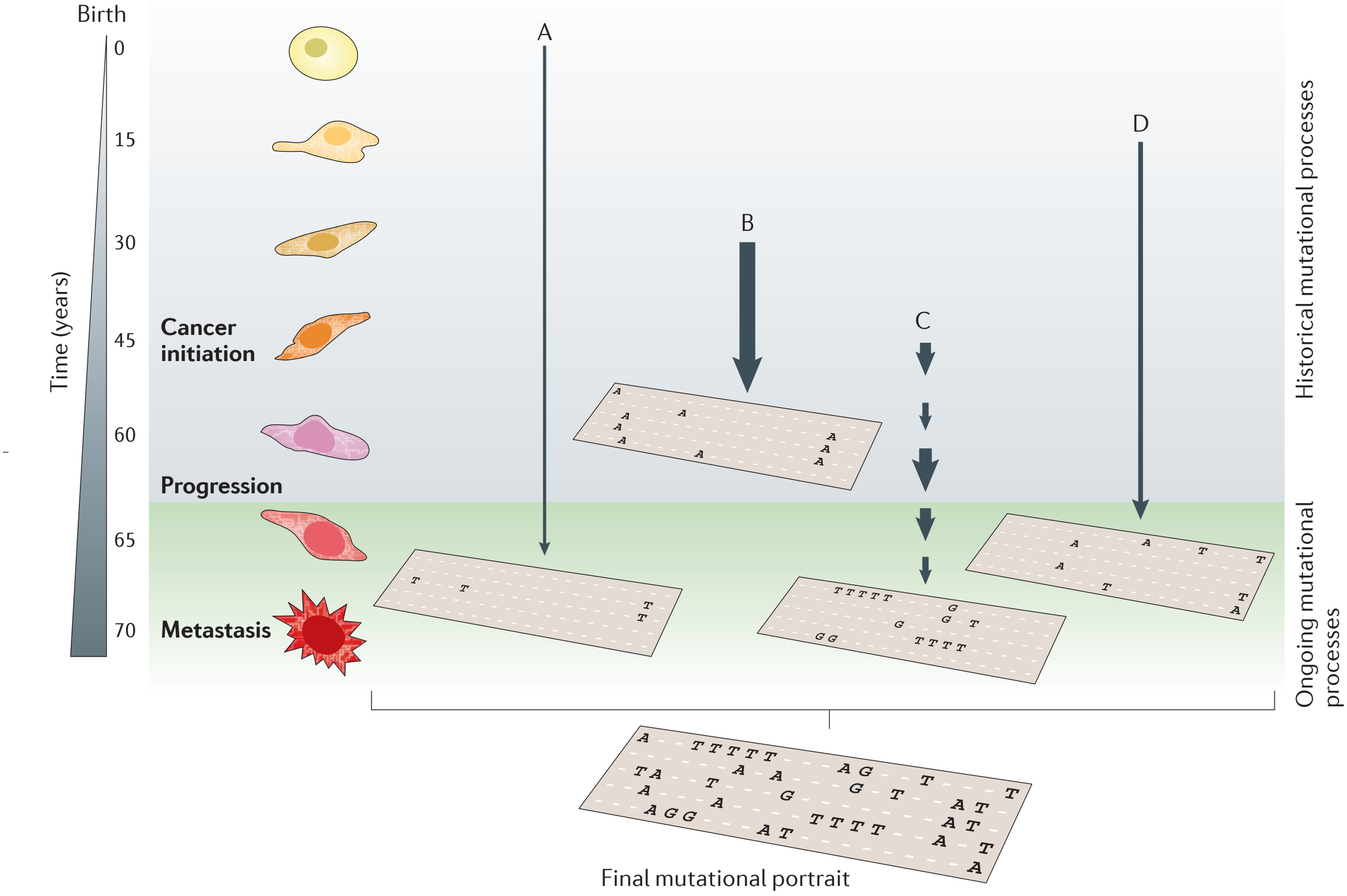
# “Mutational signatures” of types of DNA damage in cancer




Alexandrov, *et al. Nature* 2013

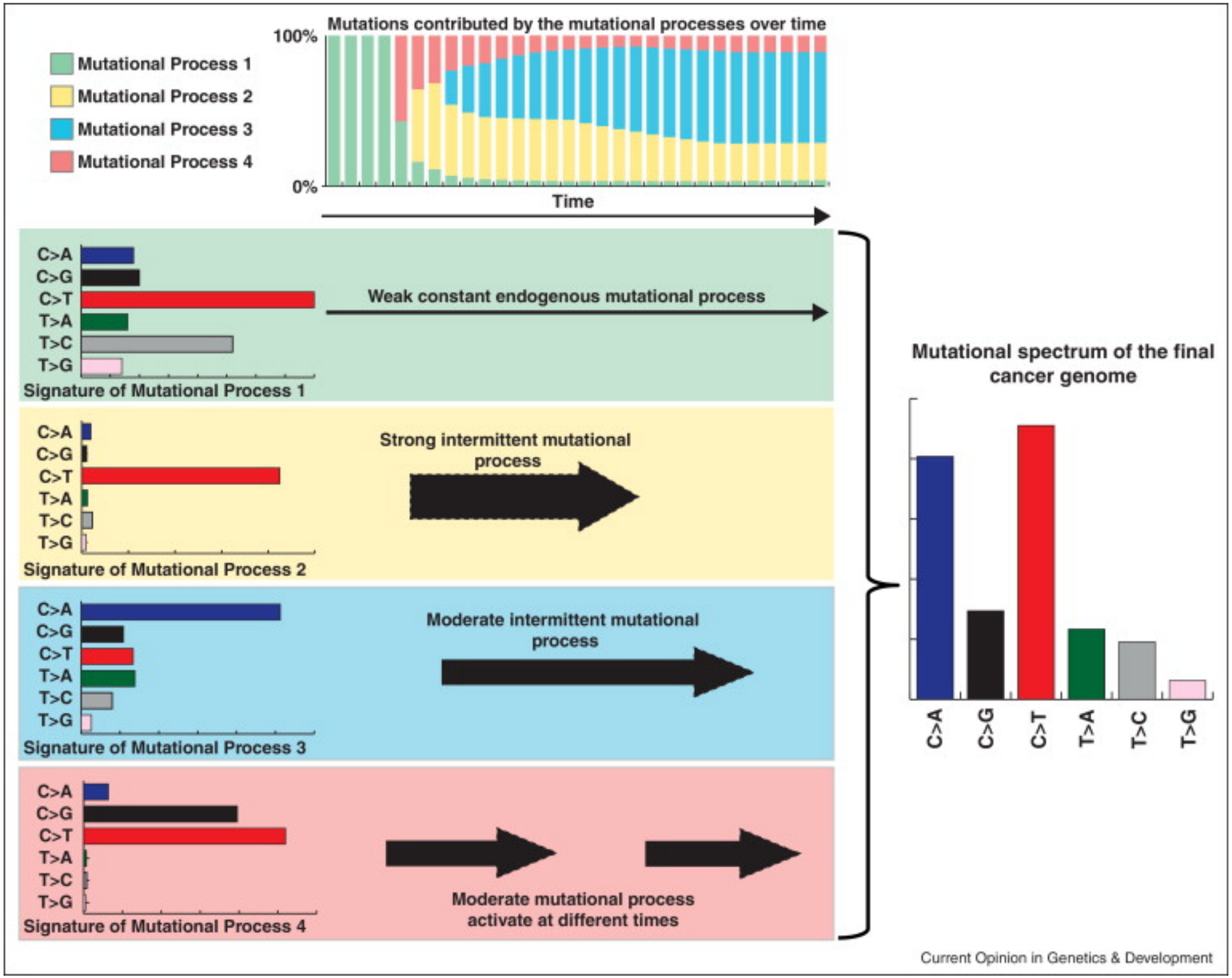
# APOBEC / AID deaminases

- APOBEC attacks RNA viruses, mutating TCA and TCT by deamination
- Its homologue AID hypermutates T cell receptors for proper immune function
- Both cause off-target germline mutations, especially in endogenous retroviral sequences
- APOBEC is erroneously switched on in many cancers (esp cervical), associated with poorer outcomes



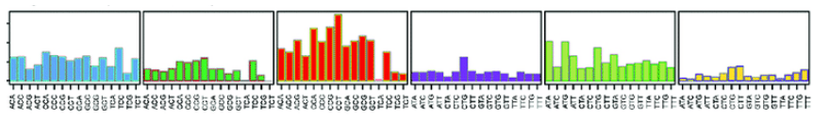

Helleday, *et al. Nature Review Genetics* 2014

Alexandrov, *et al. Nature* 2013



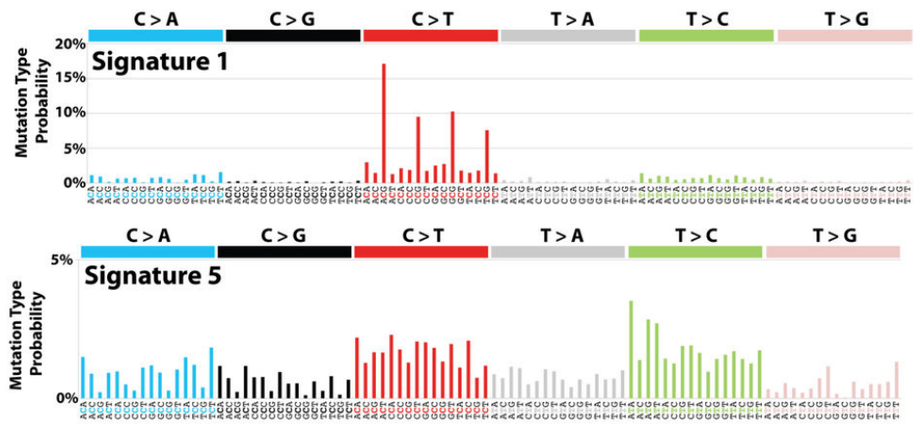

**Alexandrov and Stratton 2014**

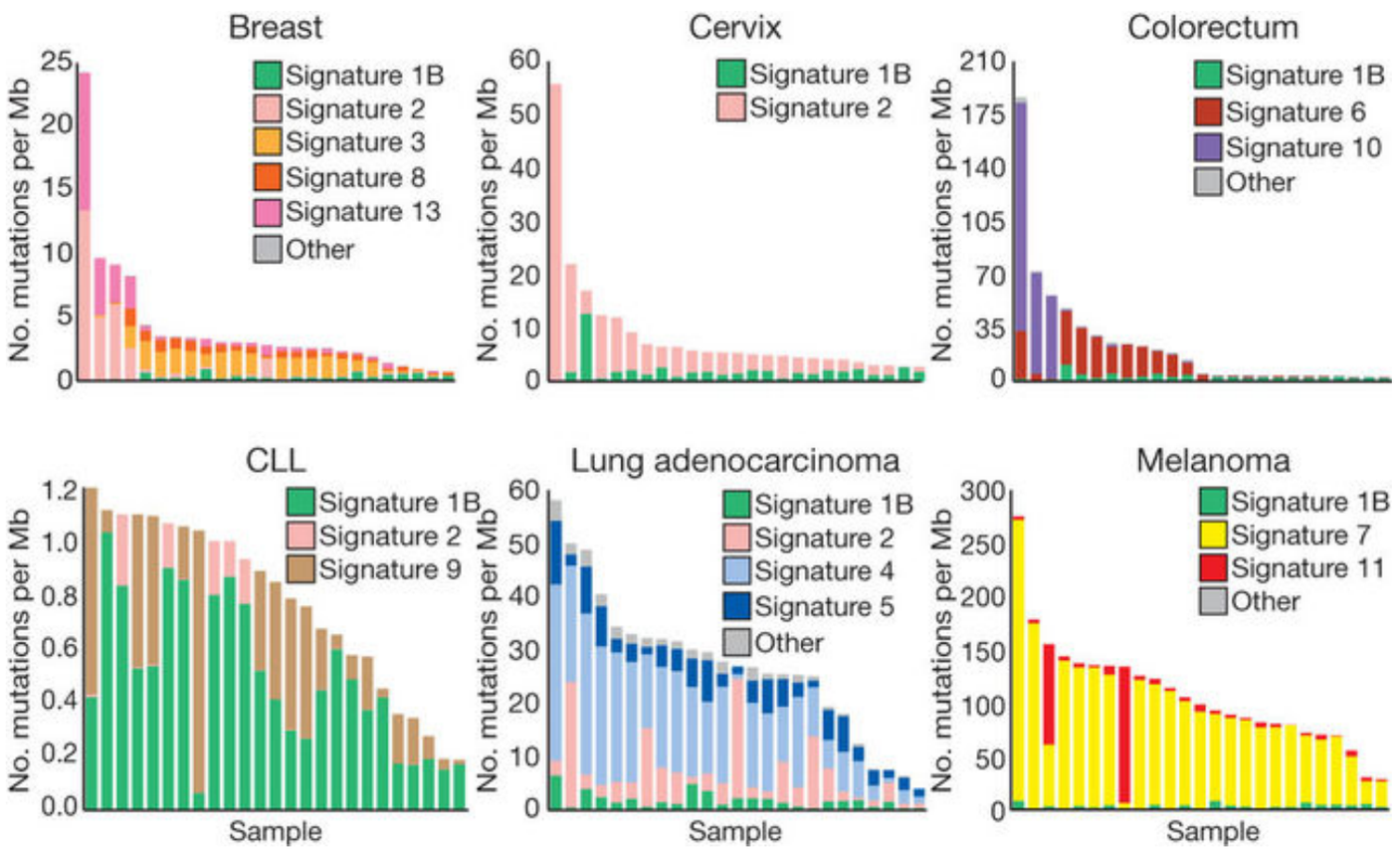
# Mutation signature analysis



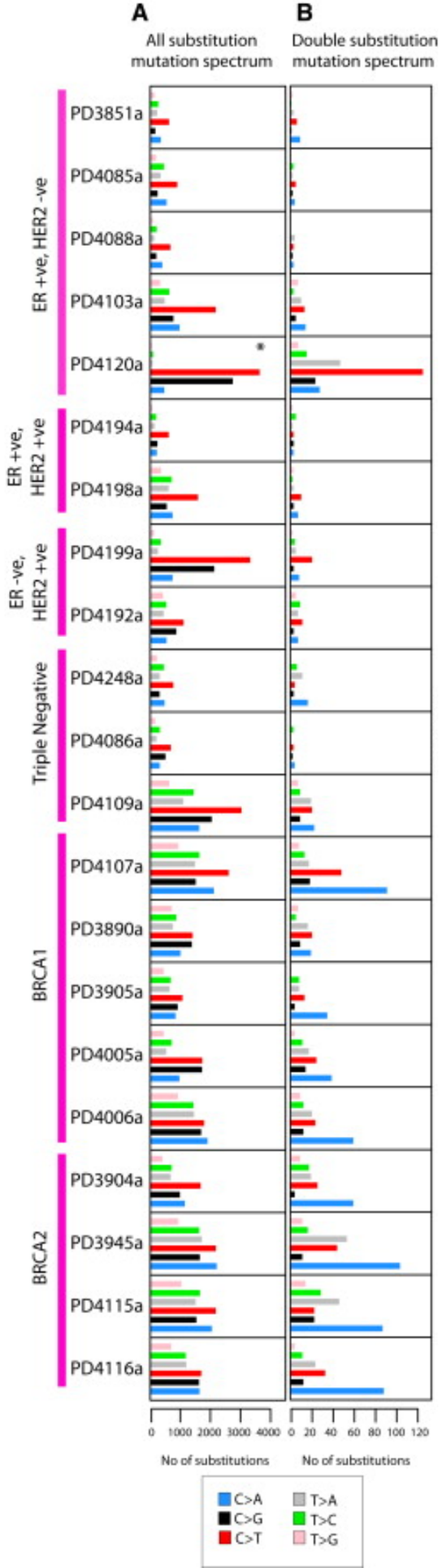
**Mutation counts in 96 triplet contexts across cancers**

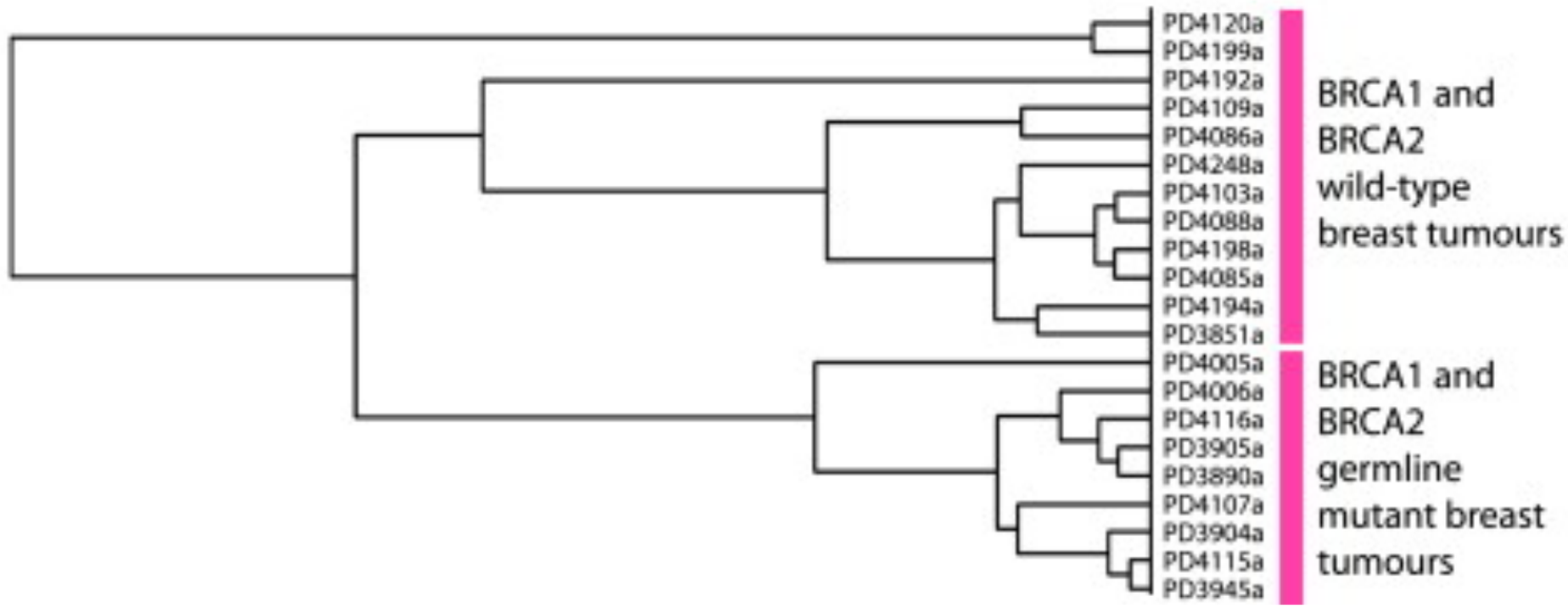
**Nonnegative matrix factorization**



# Effect of BRCA germline mutations on breast cancer mutation distribution



# Mutational signatures in the germline?

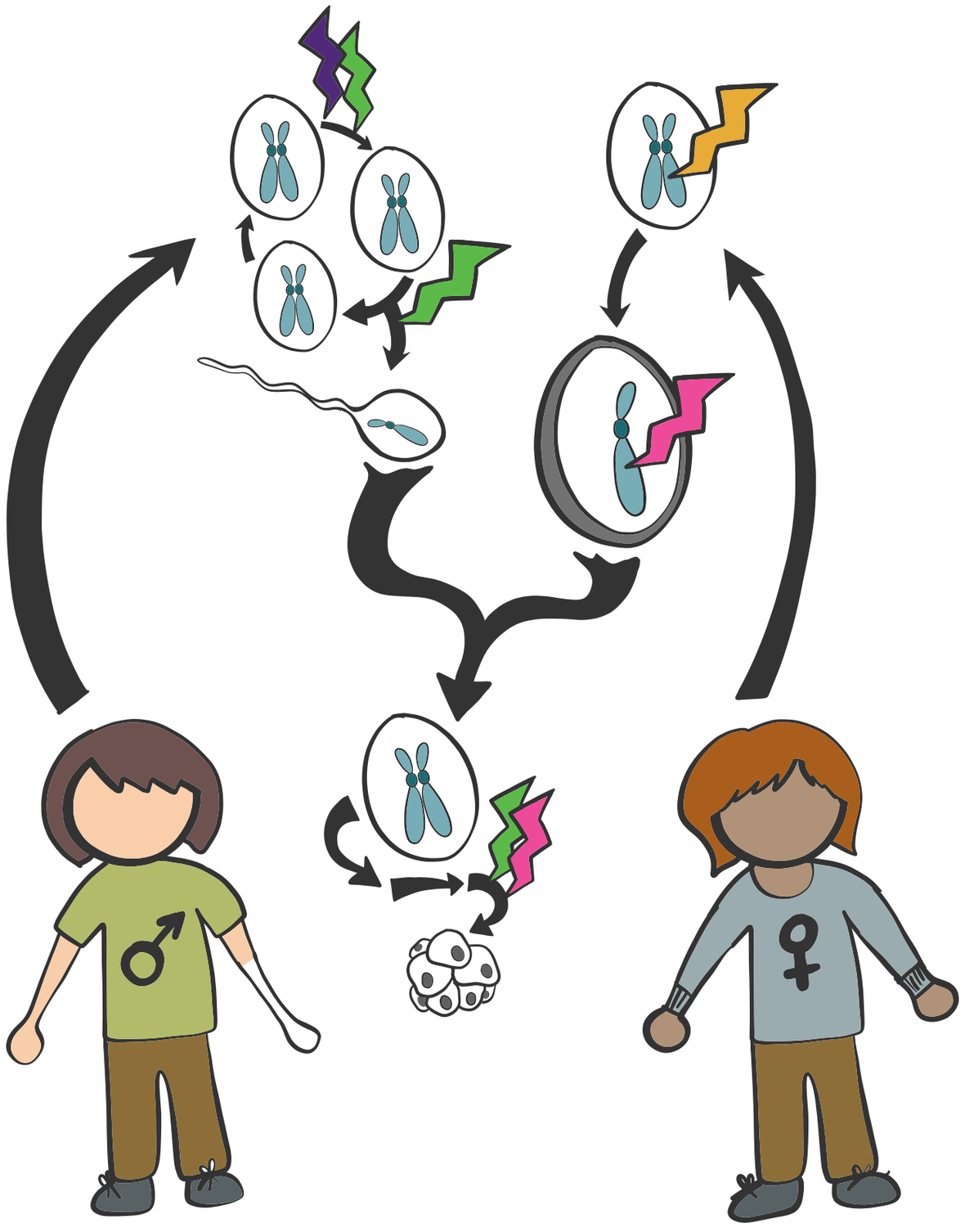
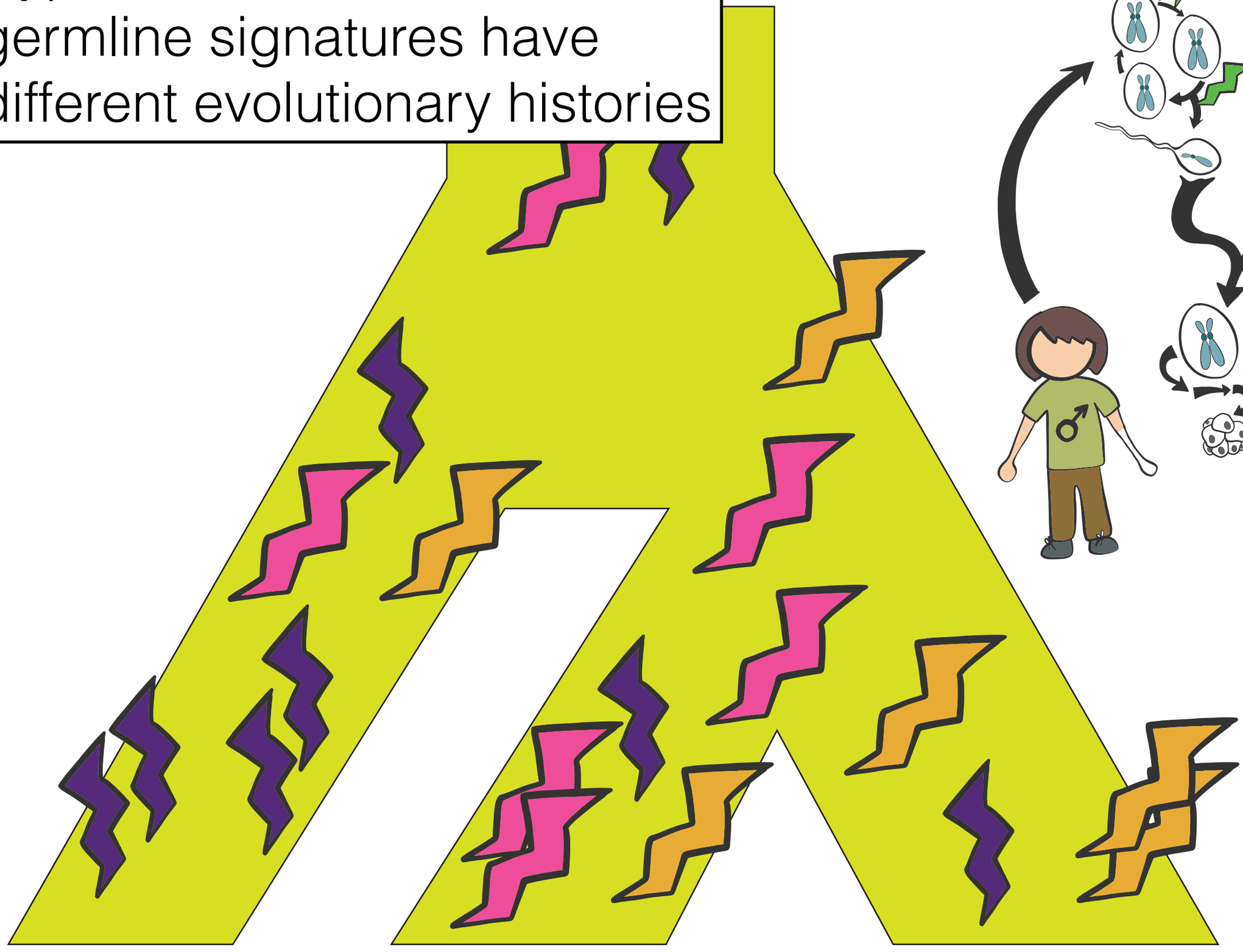
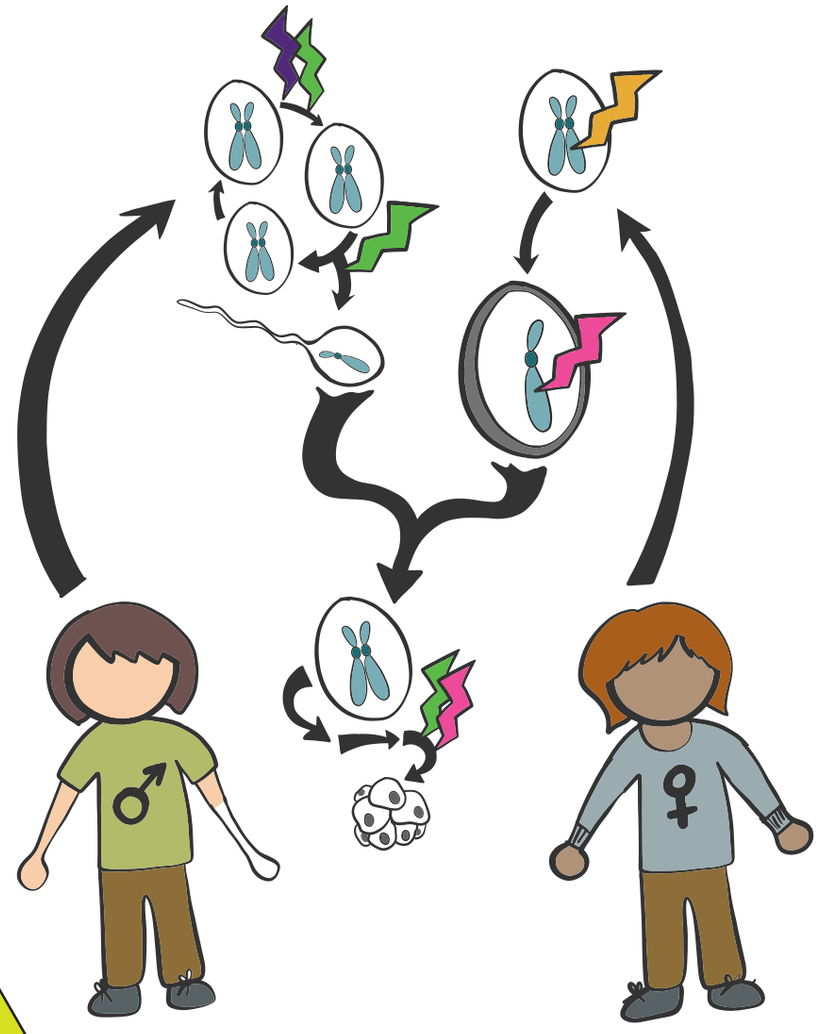


Image co-artist: Natalie Telis

Hypothesis: different germline signatures have different evolutionary histories



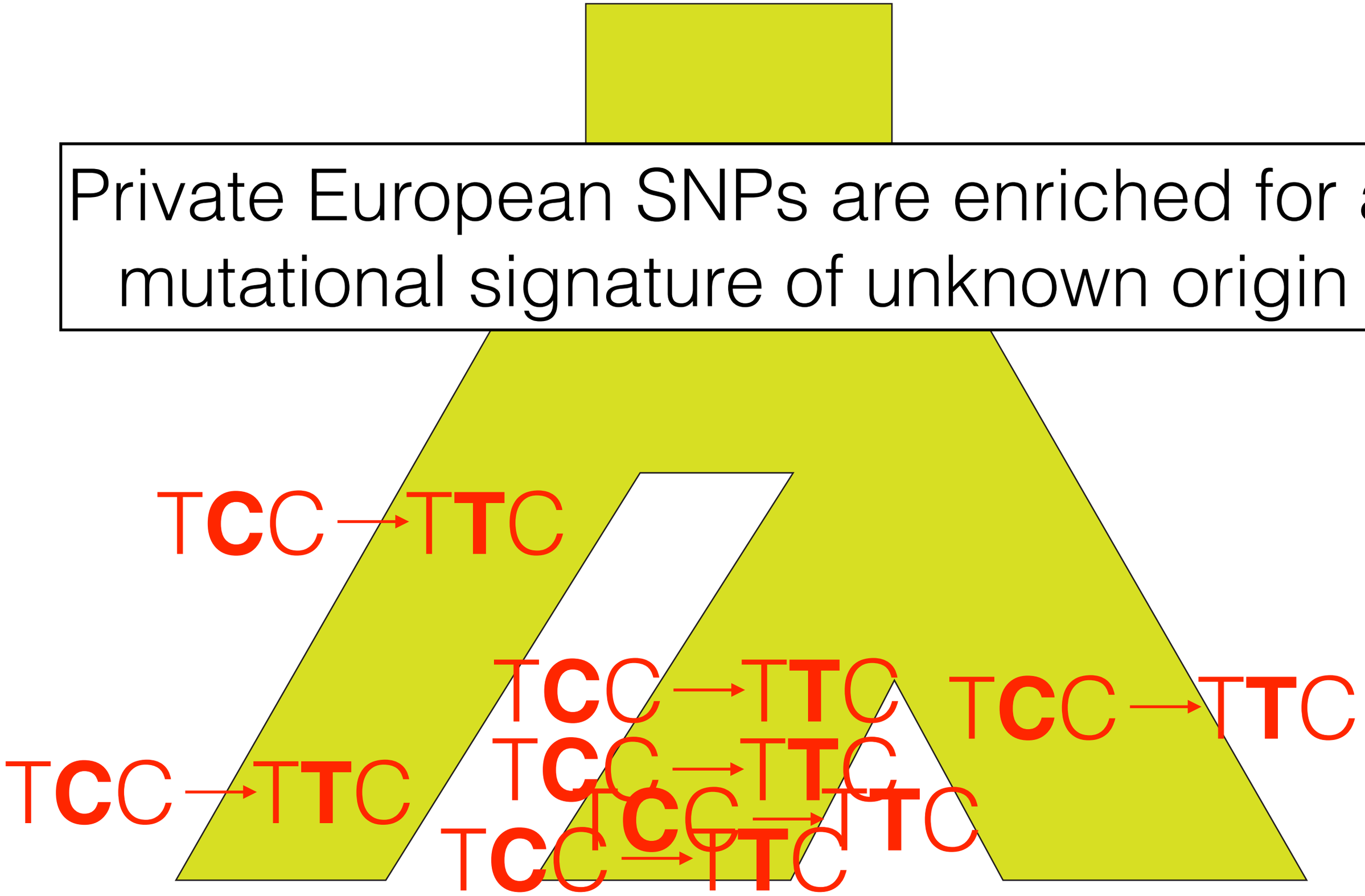


Africans

Europeans

East Asians

Private European SNPs are enriched for a mutational signature of unknown origin

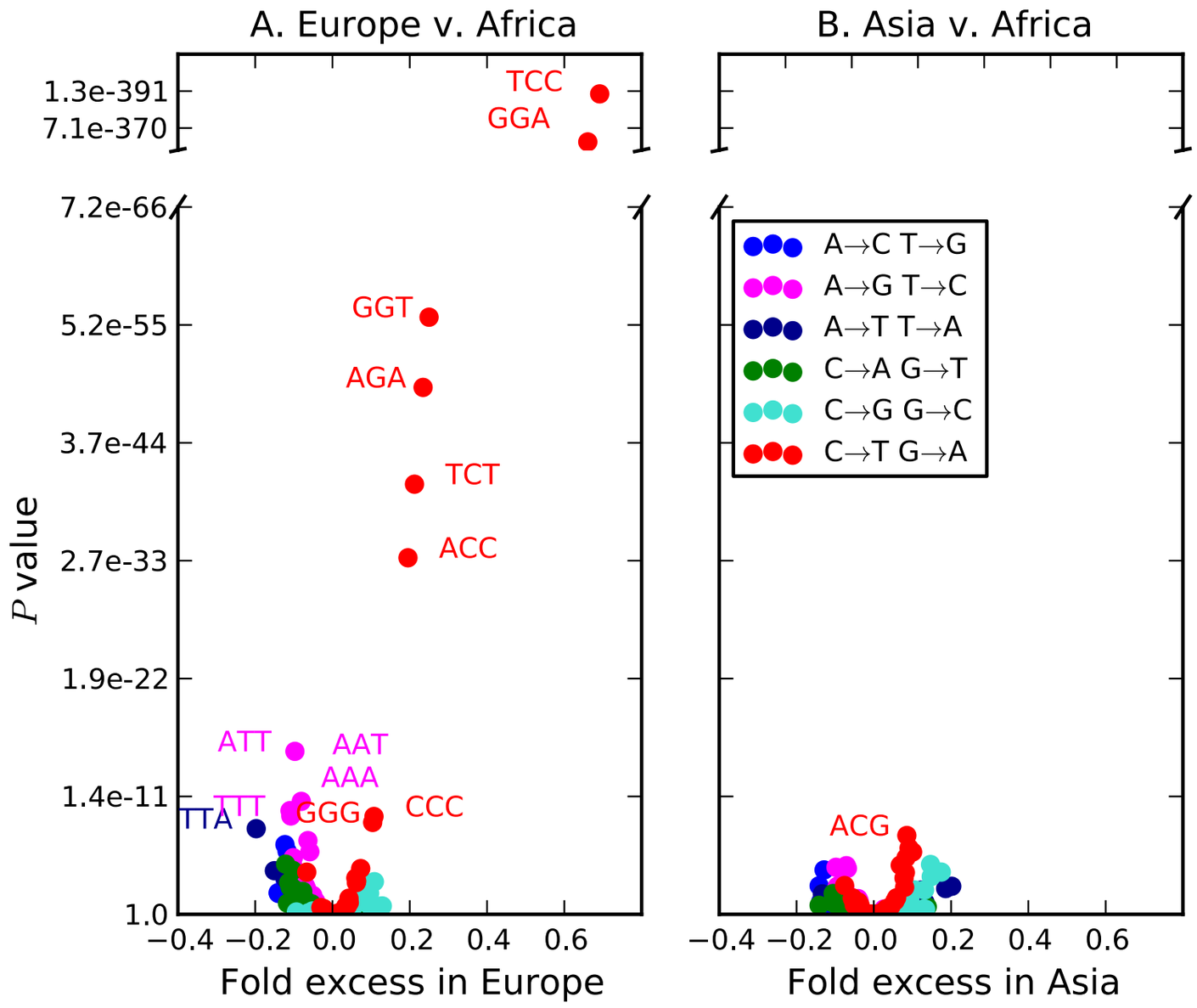



Harris *PNAS* 2015

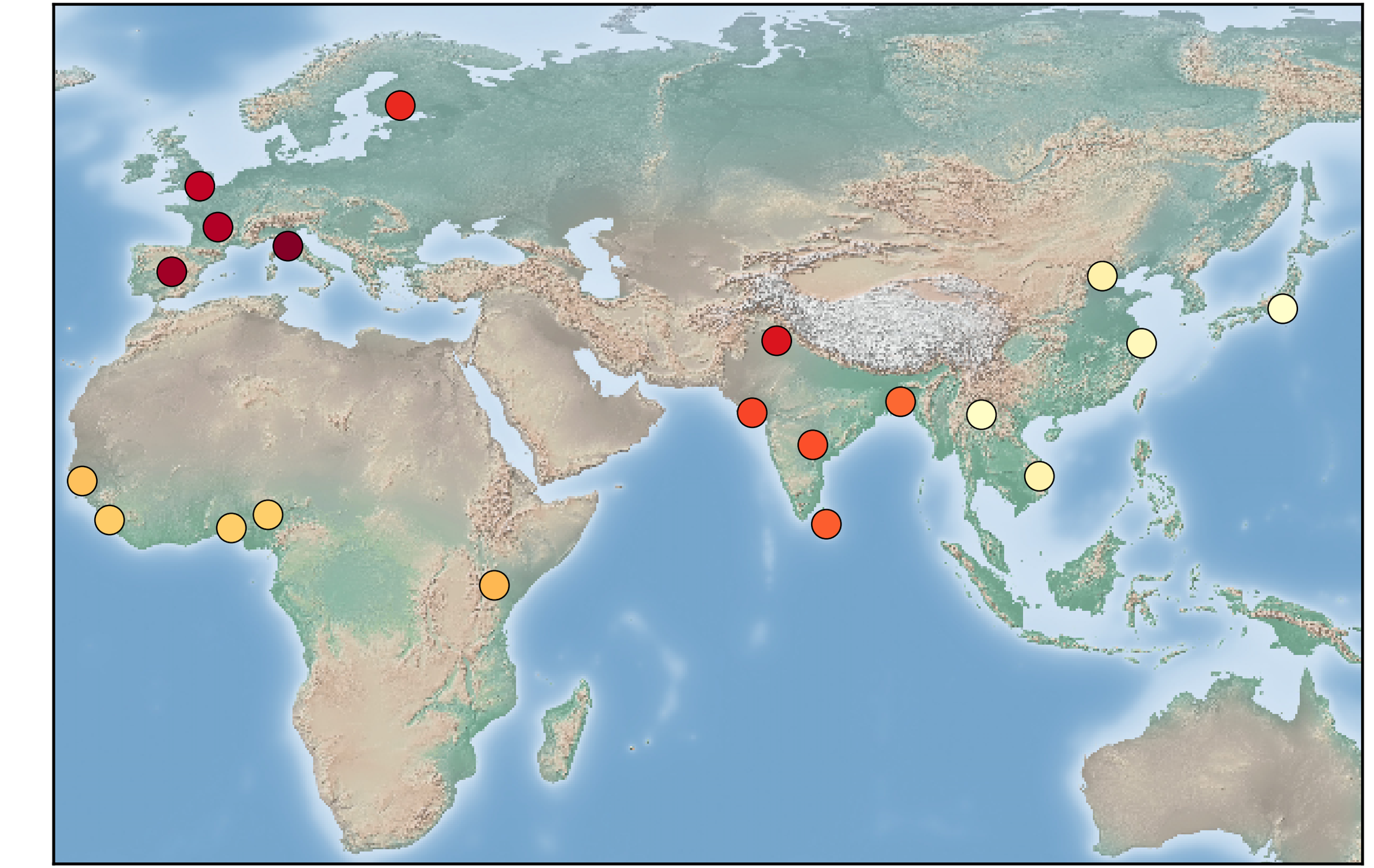
# A signature of elevated mutagenesis in the European germline

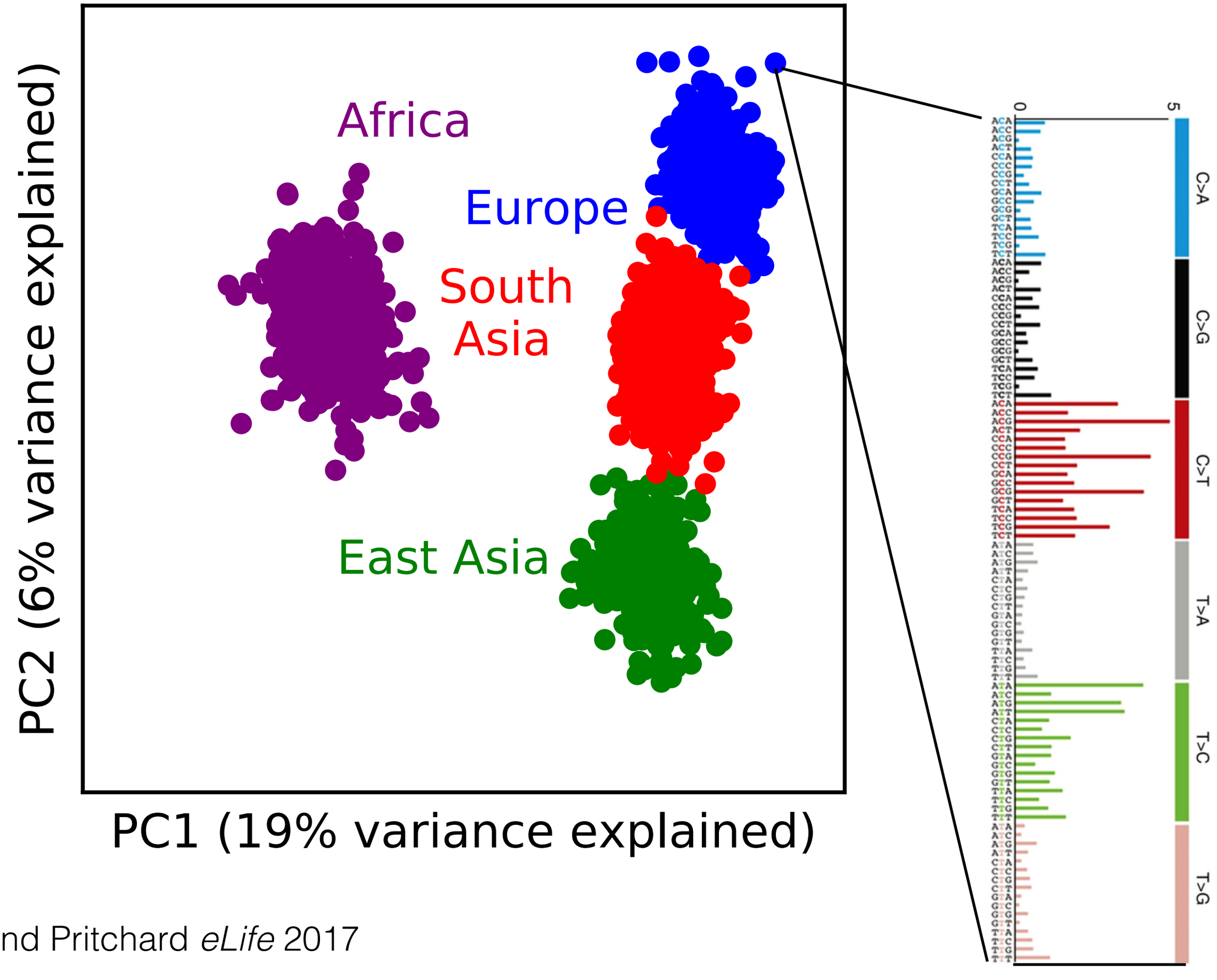
TCC→TTC

TCT → TTT
CCC → CTC
ACC → ATC




Harris *PNAS* 2015


0.017
0.018
0.019
TCC→TTC Mutation Fraction

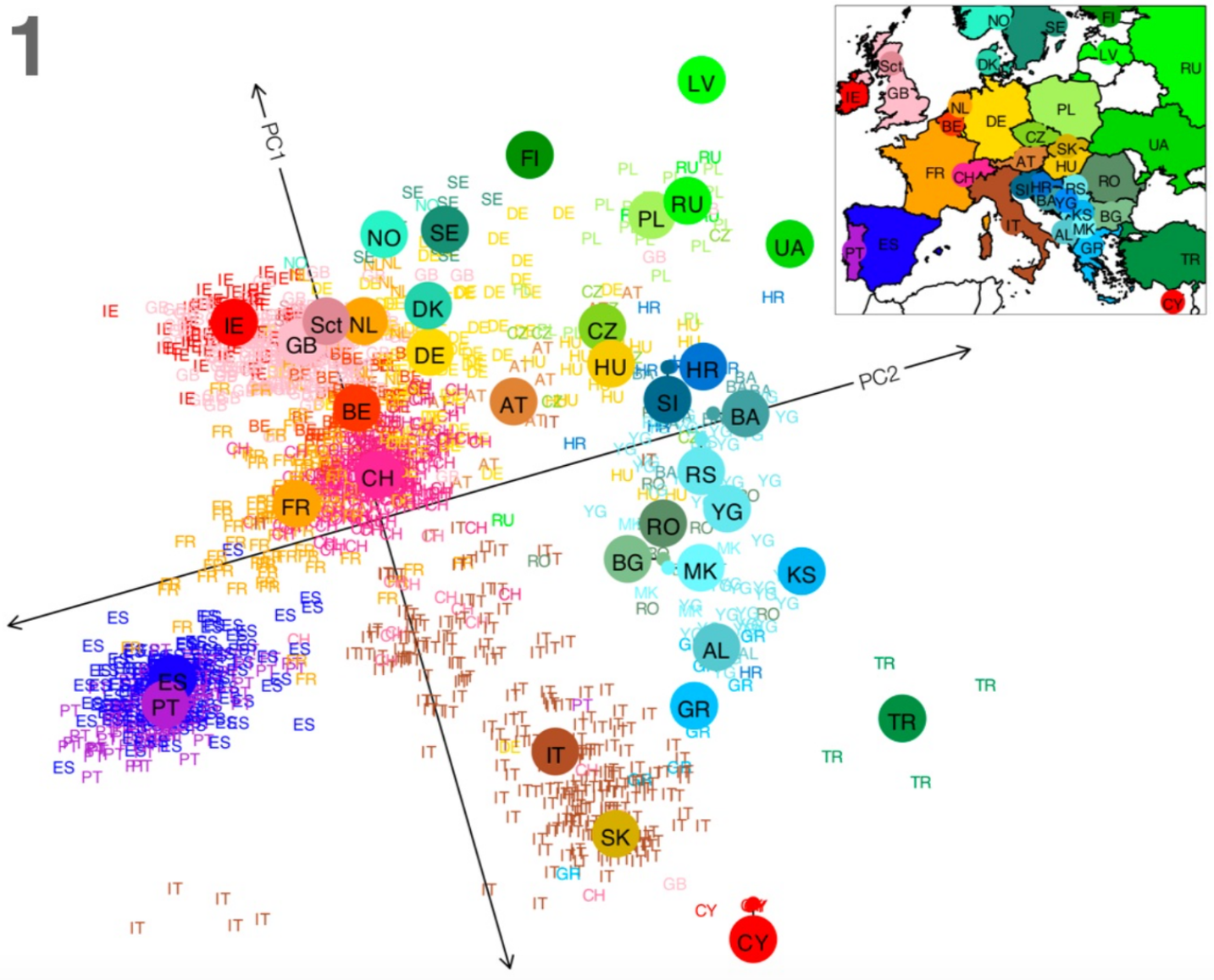



Harris and Pritchard *eLife* 2017

# Genes mirror geography within Europe

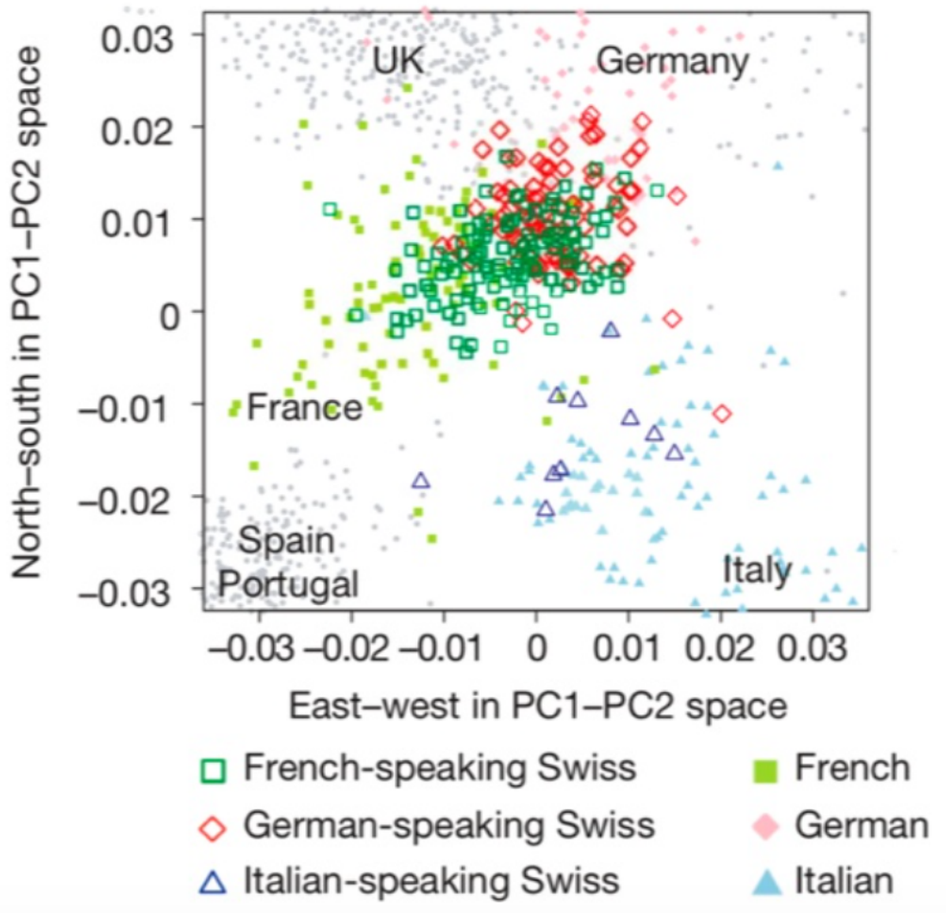
John Novembre[1,2], Toby Johnson[4,5,6], Katarzyna Bryc[7], Zoltán Kutalik[4,6], Adam R. Boyko[7], Adam Auton[7], Amit Indap[7], Karen S. King[8], Sven Bergmann[4,6], Matthew R. Nelson[8], Matthew Stephens[2,3] & Carlos D. Bustamante[7]

1

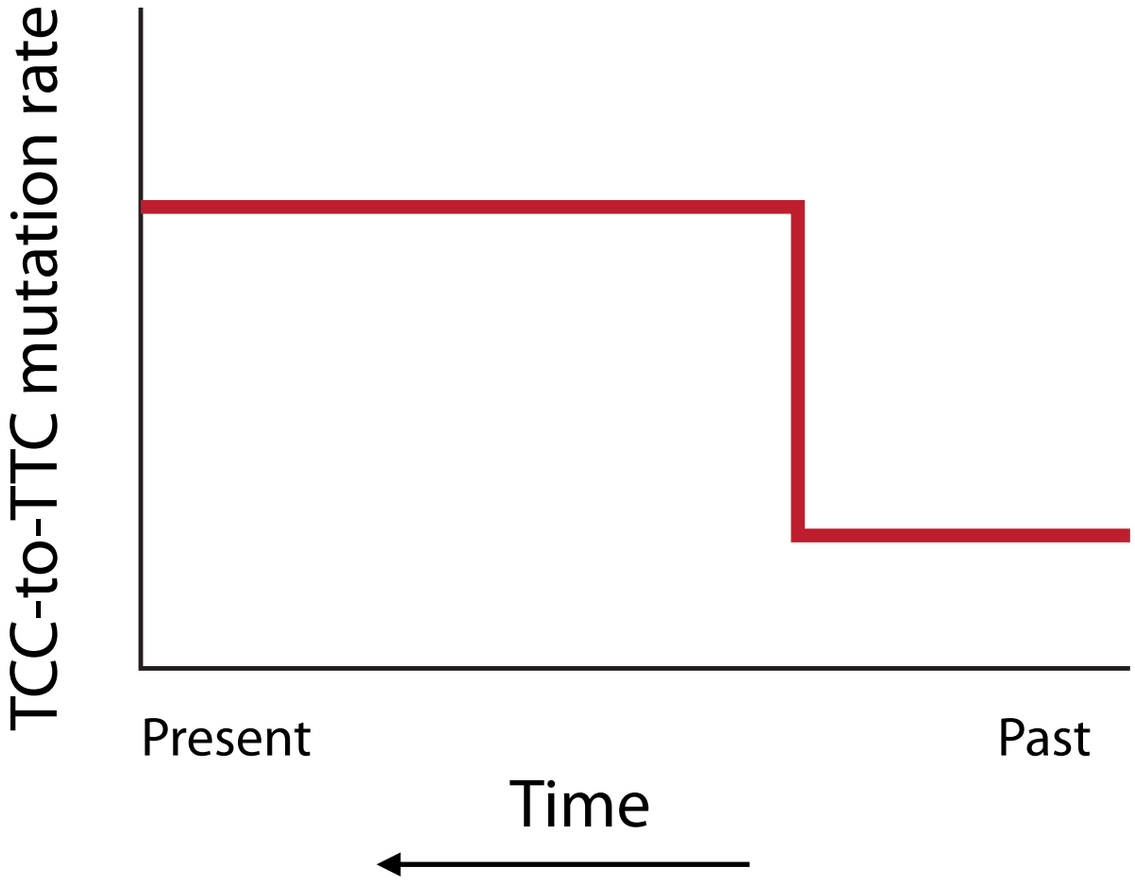


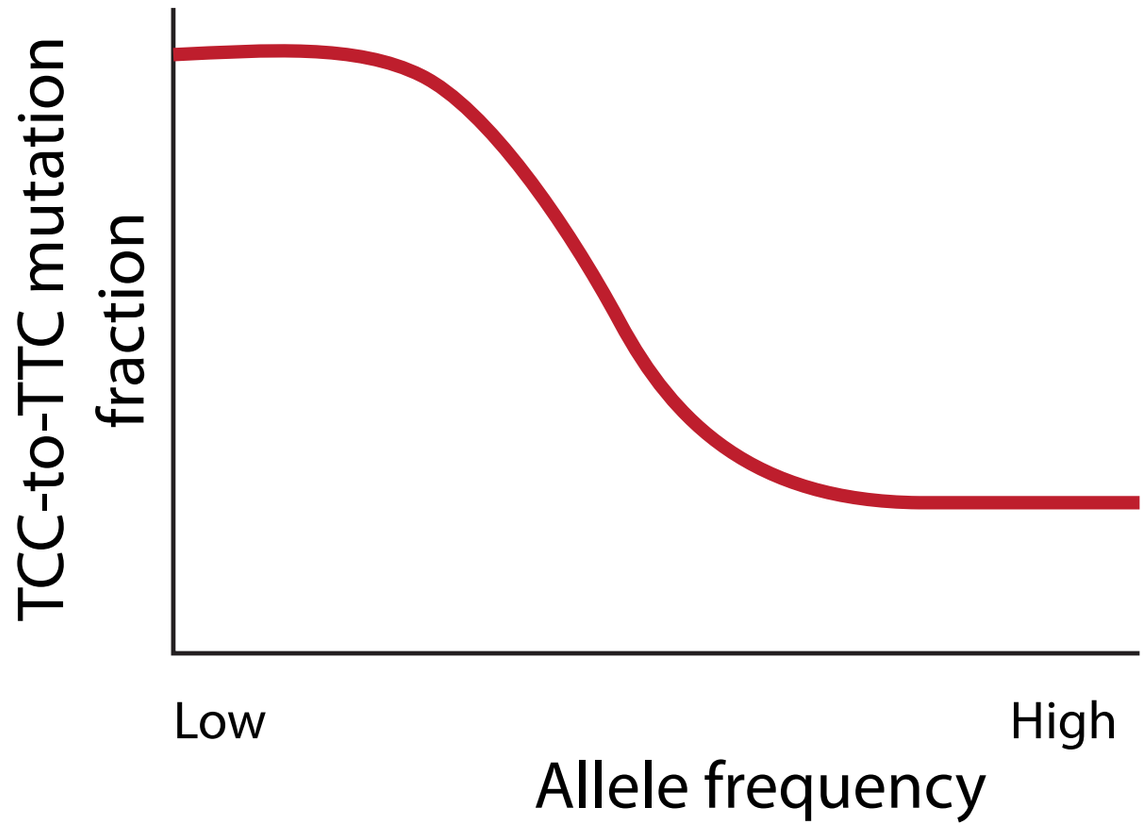


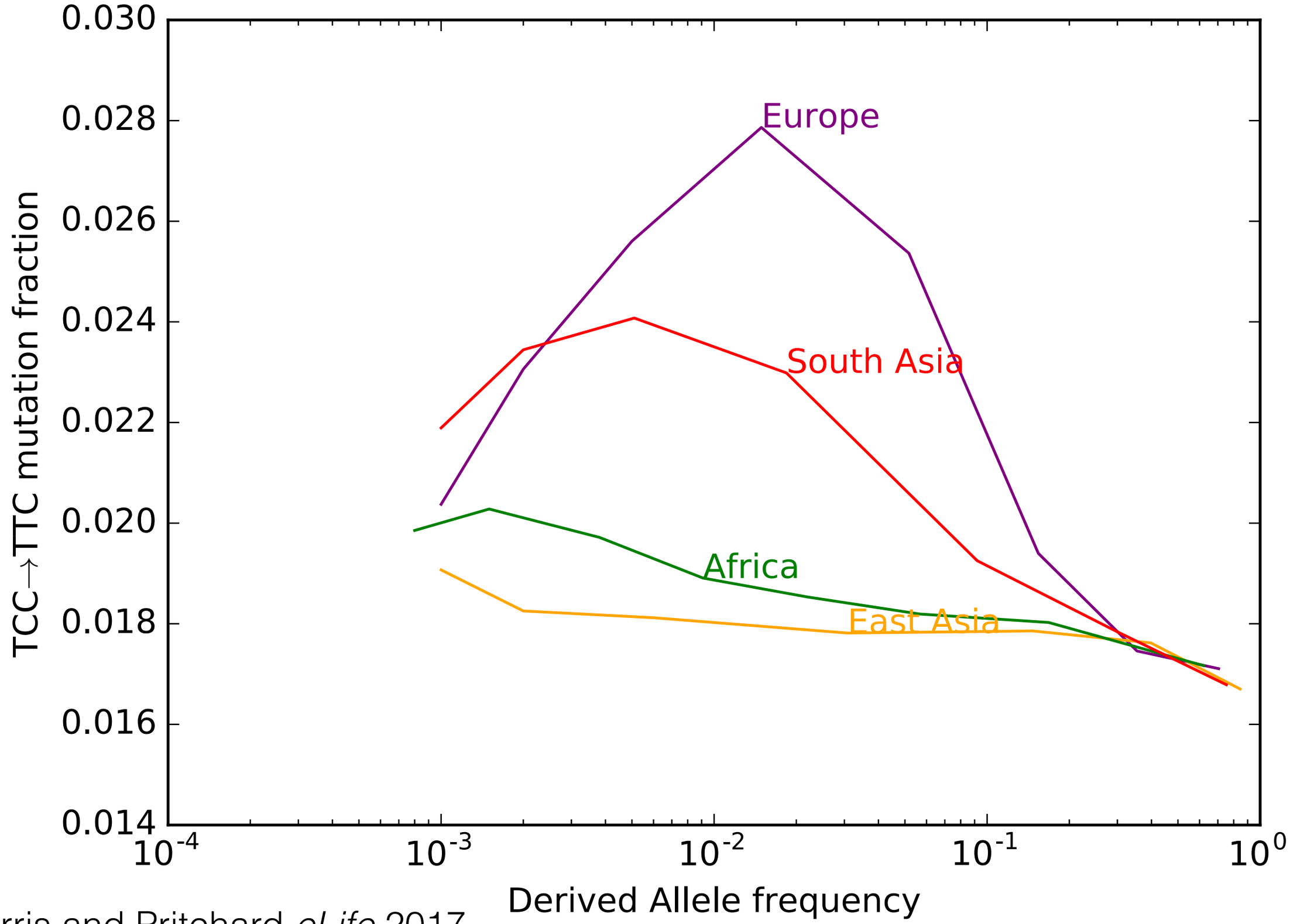




# Hypothetical Signature of a TCC-to-TTC mutation rate increase

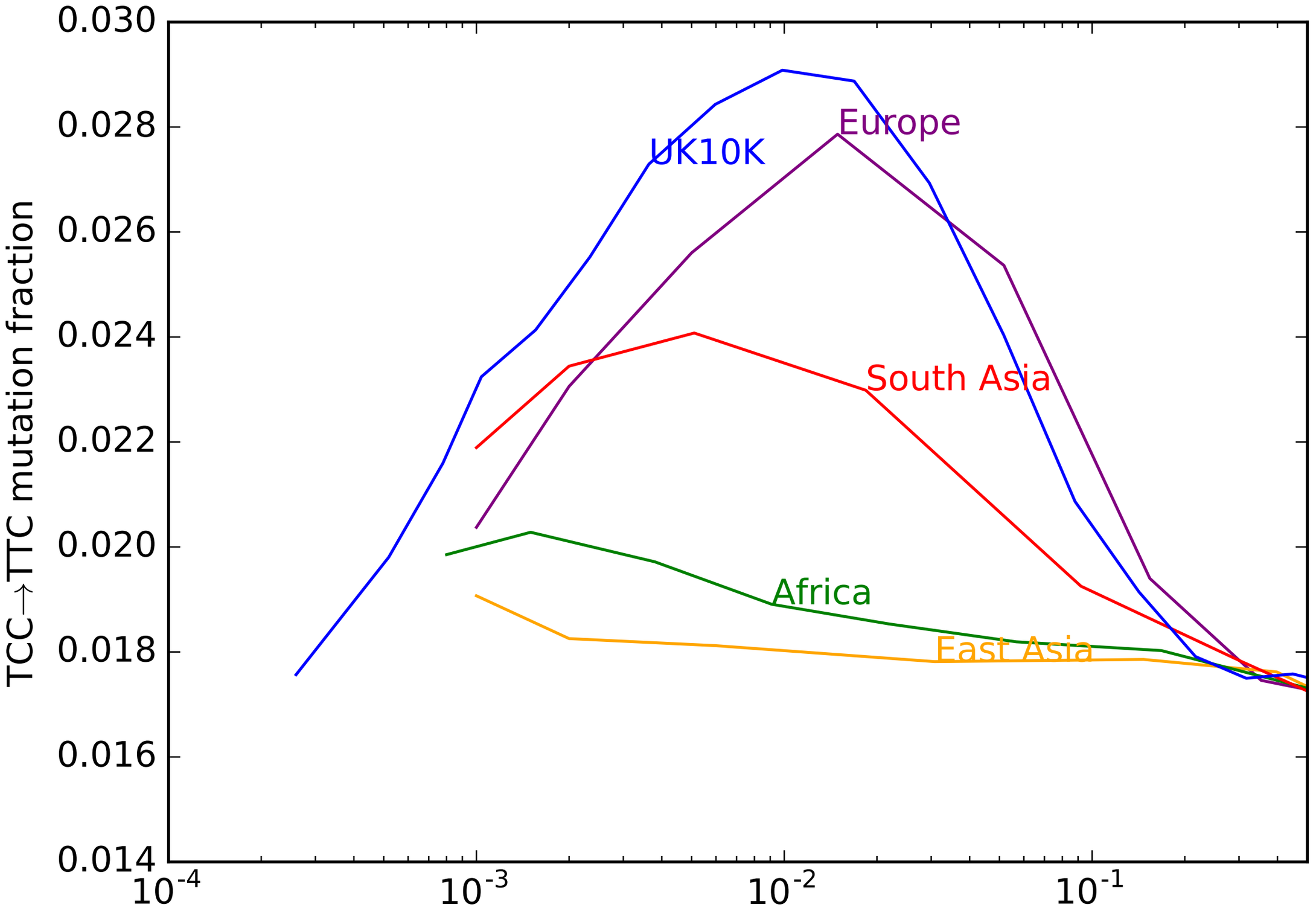




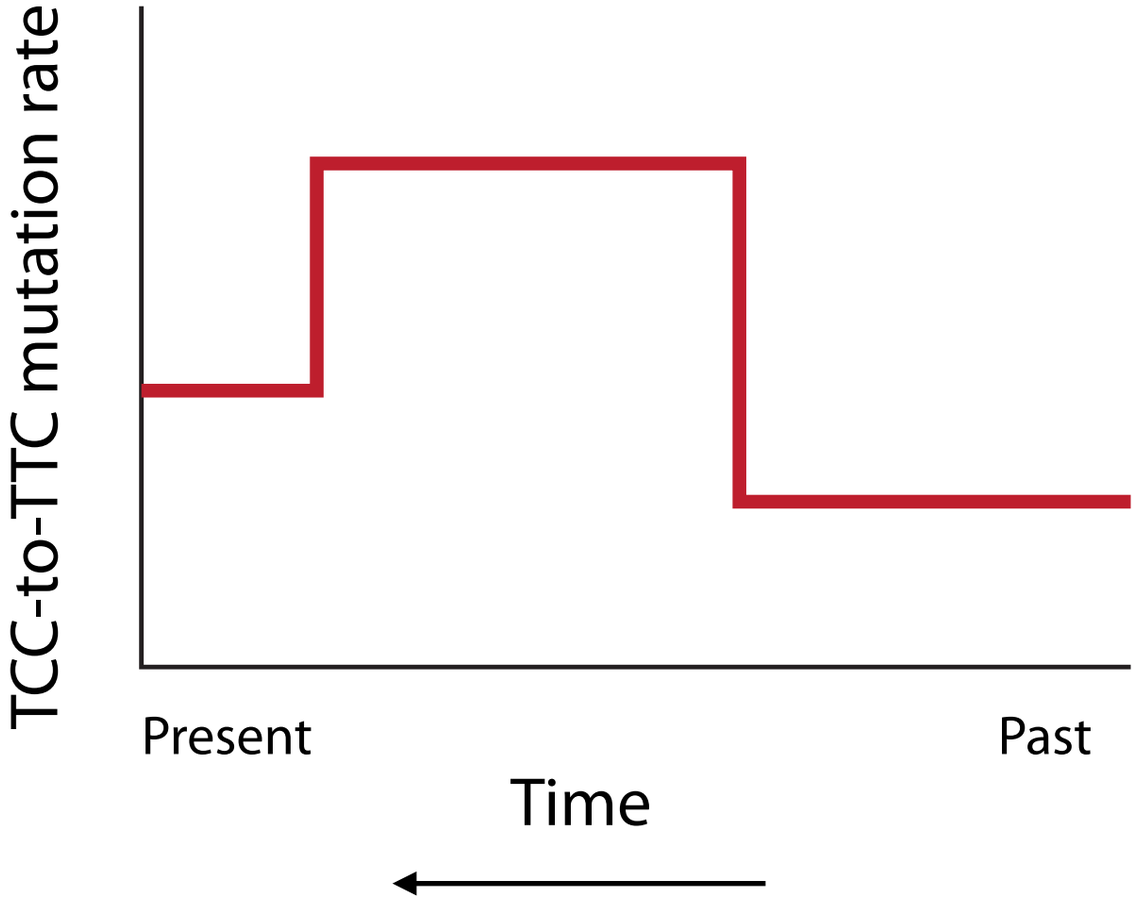

Harris and Pritchard *eLife* 2017

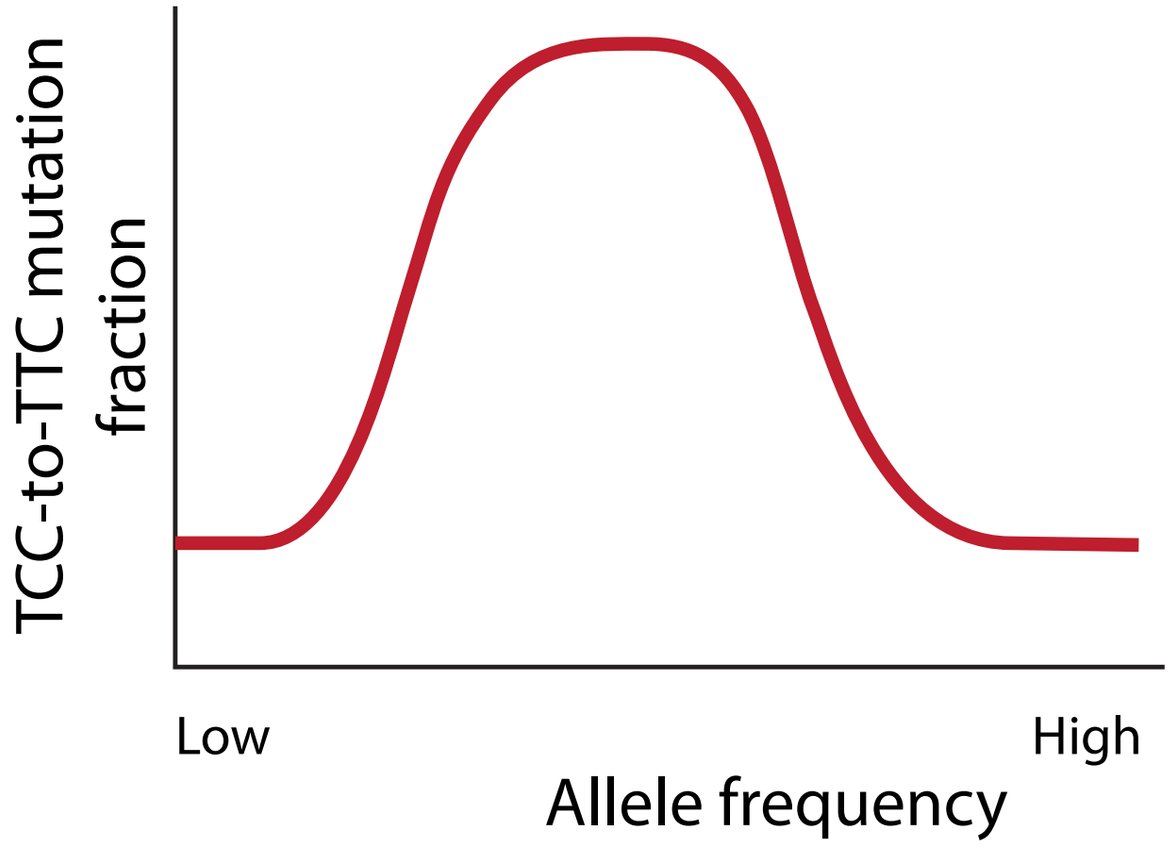
# Pulse replicates in the UK10K data



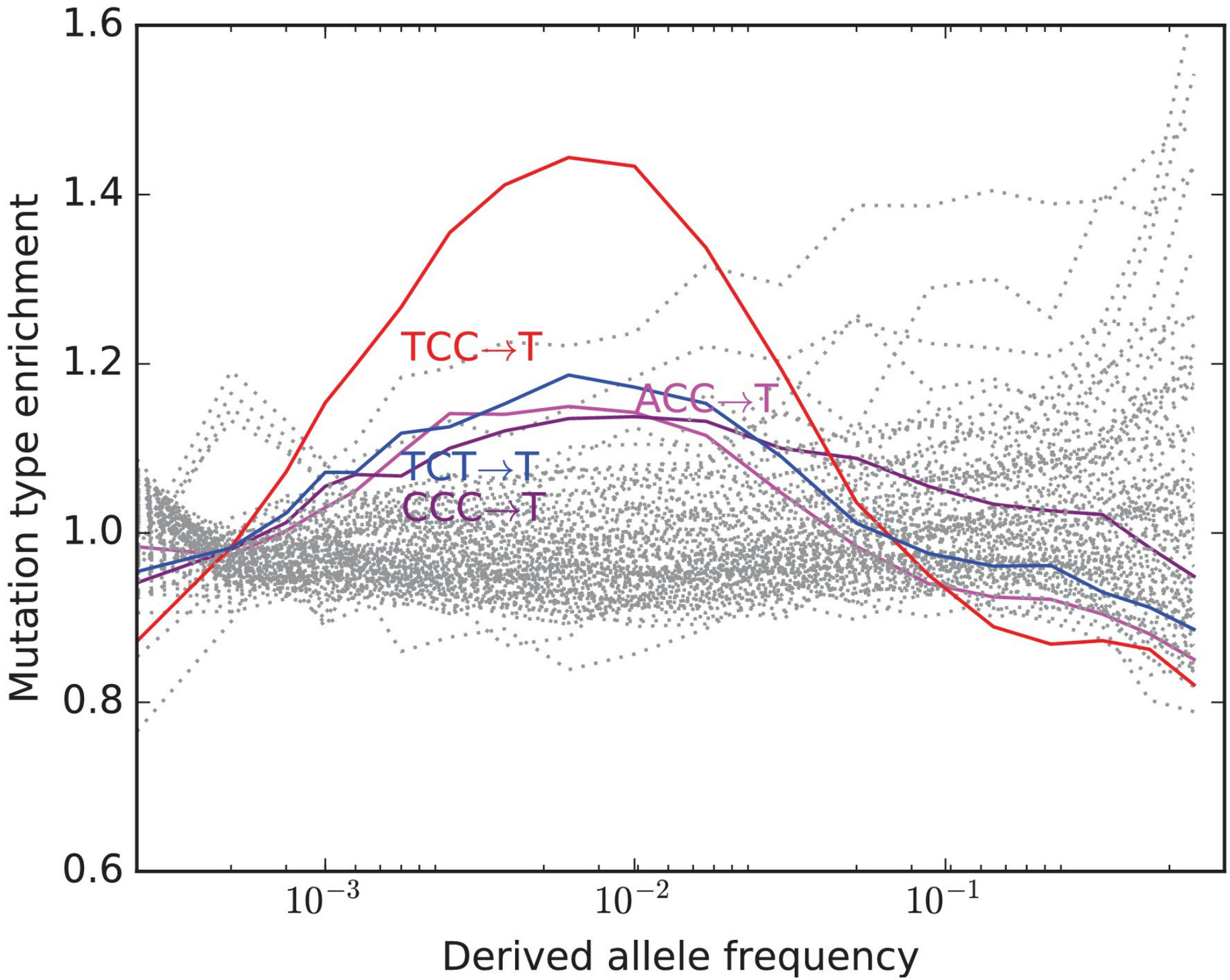

Harris and Pritchard *eLife* 2017

# A pulse of TCC-to-TTC mutations in Europe and South Asia?



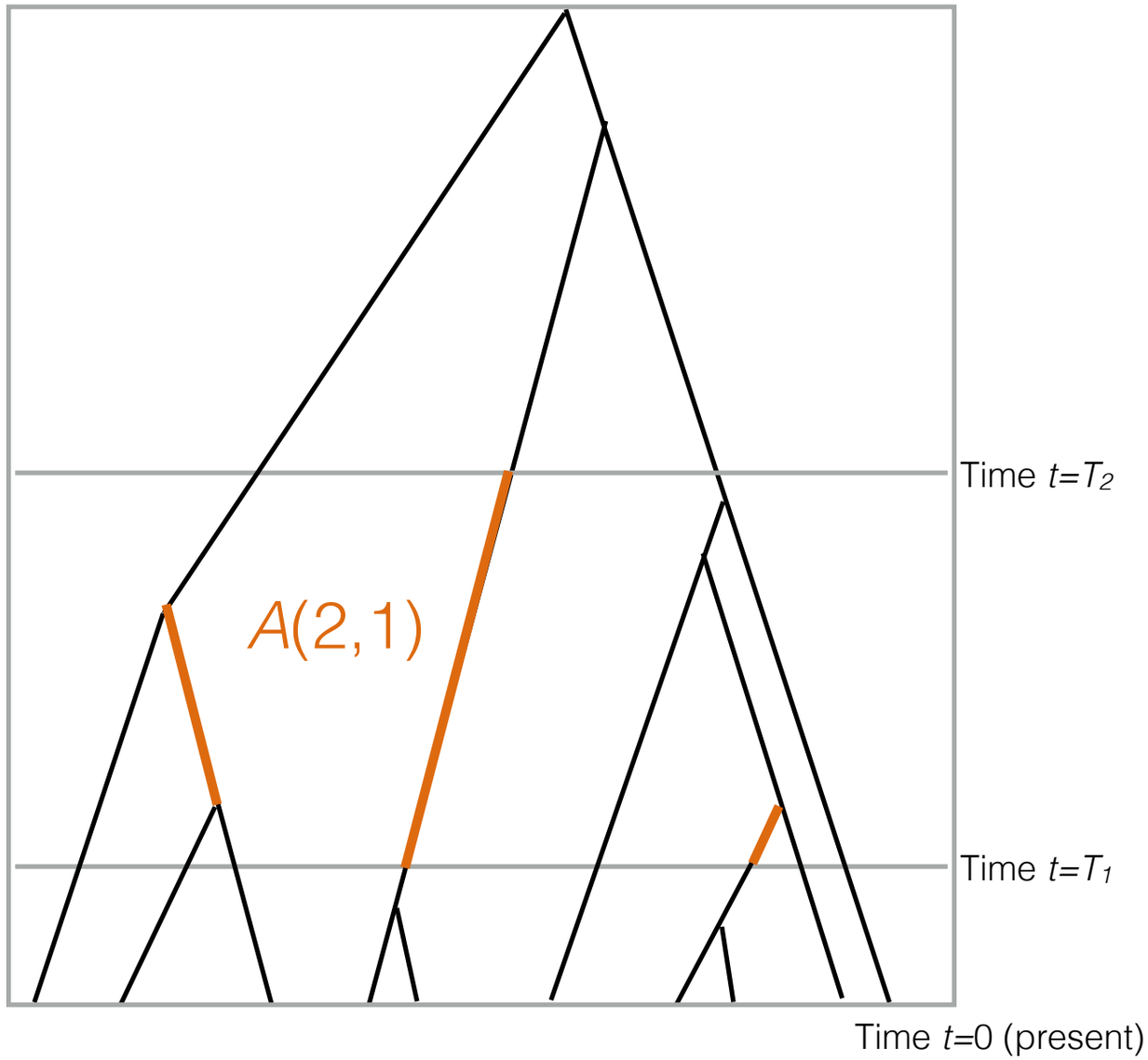



# Minor components of the pulse



# Expected TCC fraction as a function of allele frequency

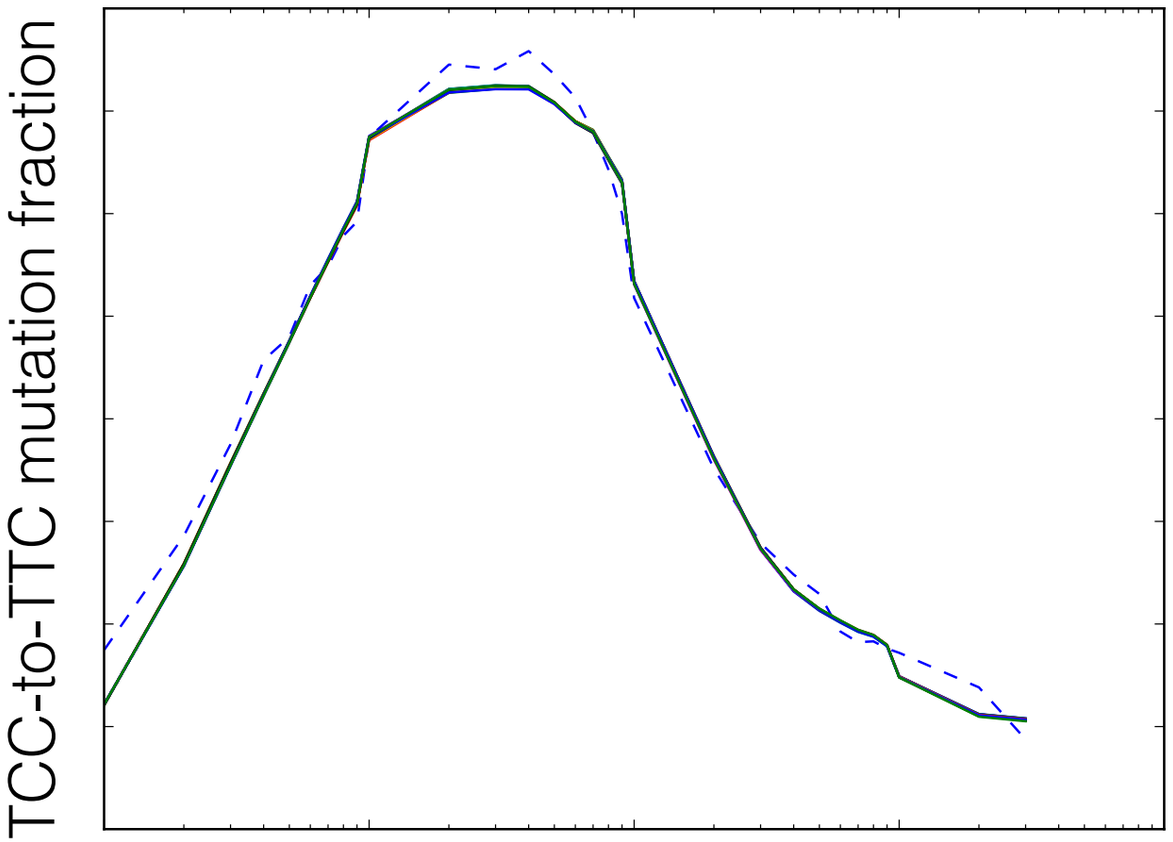
- Partition time into discrete intervals
- $A(k,i)$ = the total branch length subtending k lineages between times $T_i$ and $T_{i-1}$
- $r_i$ ~ the rate of TCC mutations between $T_i$ and $T_{i-1}$



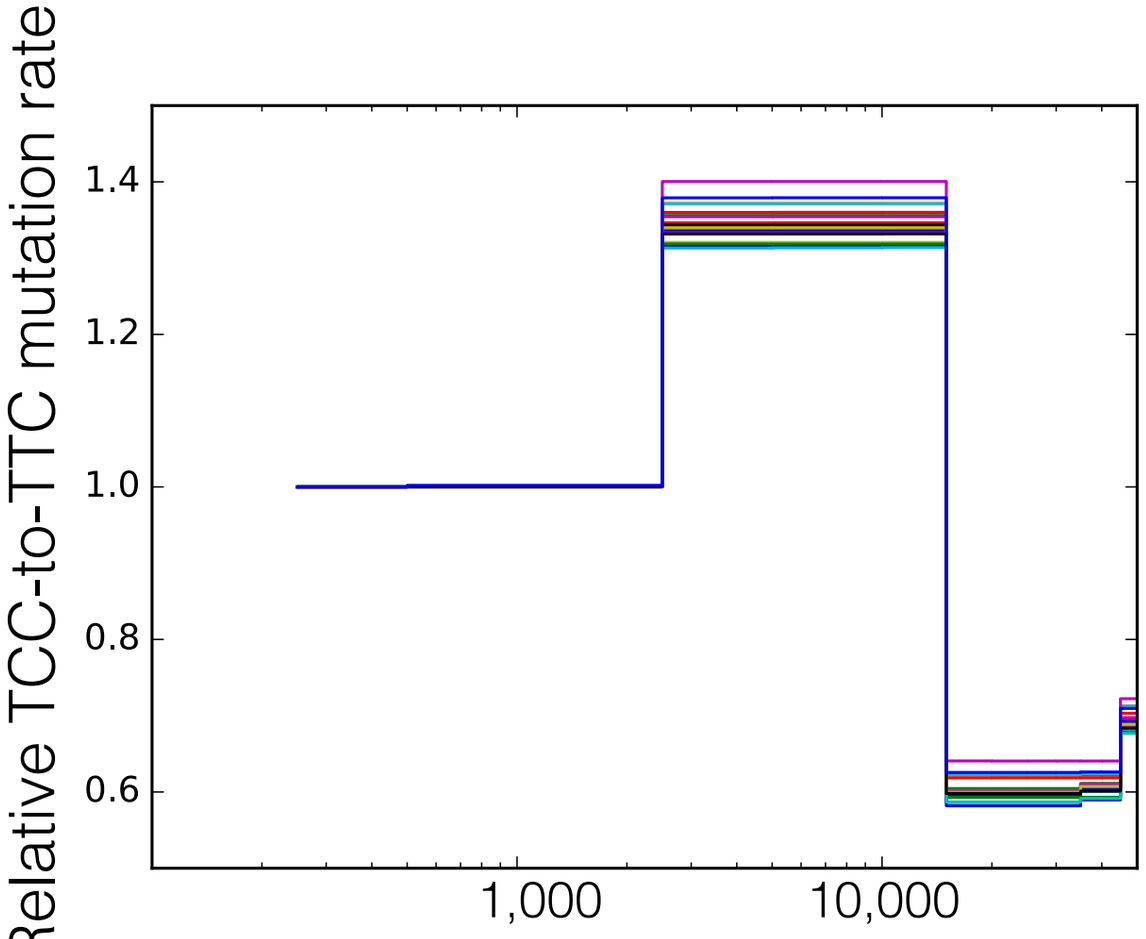

Expected TCC fraction as a function of allele frequency is

$$E[f(k)] \sim (\sum_i A(k,i)\ r_i) / \sum_i A(k,i)$$

# Inference of a mutation pulse lasting from 15,000 to 2,000 years ago




Allele frequency




Years before present

# Similar simultaneous mutation pulses in Europeans, South Asians, and…a dog STD??

RESEARCH ARTICLE

## Somatic evolution and global expansion of an ancient transmissible cancer lineage

Adrian Baez-Ortega[1], Kevin Gori[1,*], Andrea Strakova[1,*], Janice L. Allen[2], Karen M. Allum[3], Leontine Bansse-Issa[4], Thinlay N. Bhutia[5], Jocel…

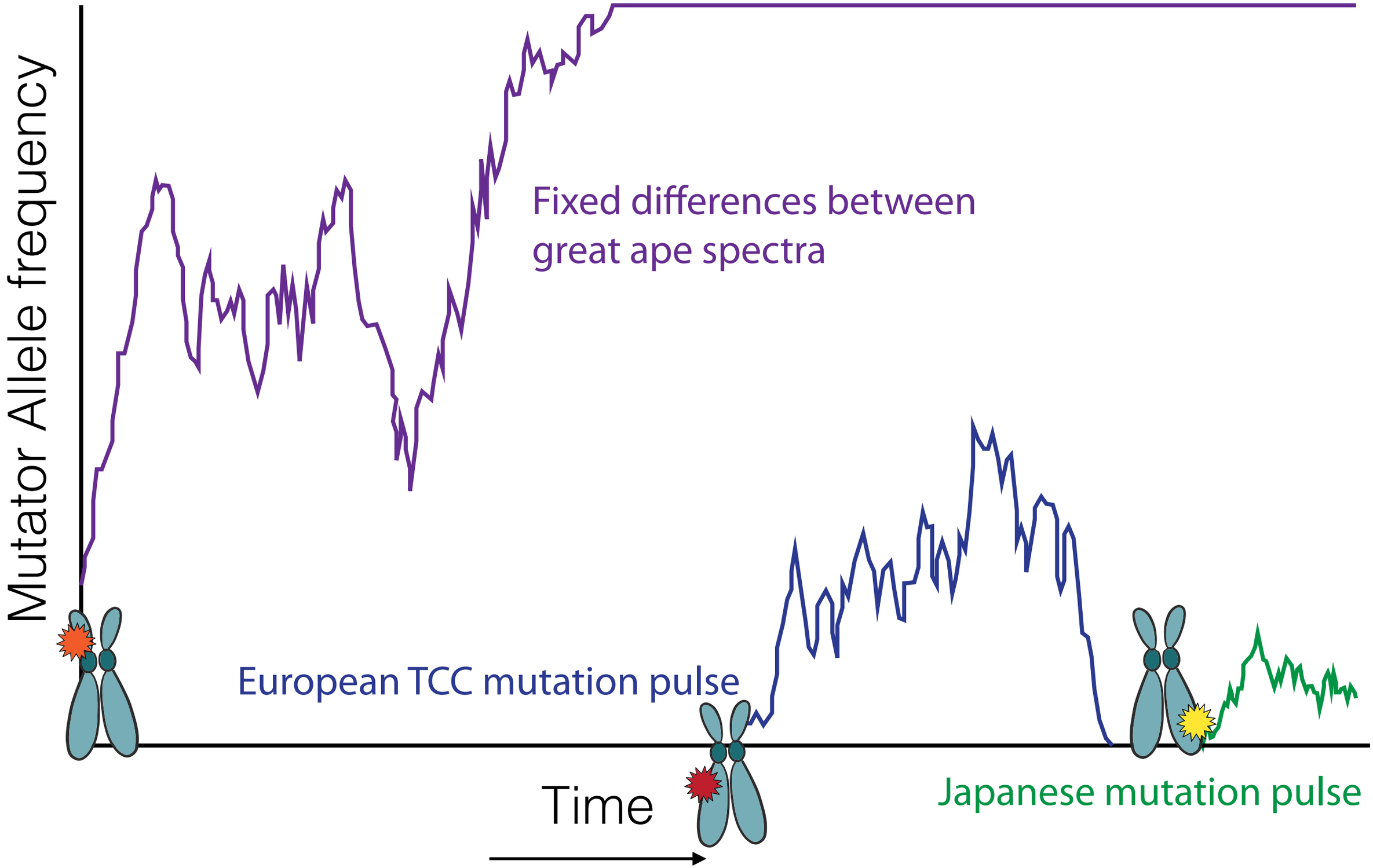
+ See all authors and affiliations

*Science* 02 Aug 2019:
Vol. 365, Issue 6452, eaau9923
DOI: 10.1126/science.aau9923

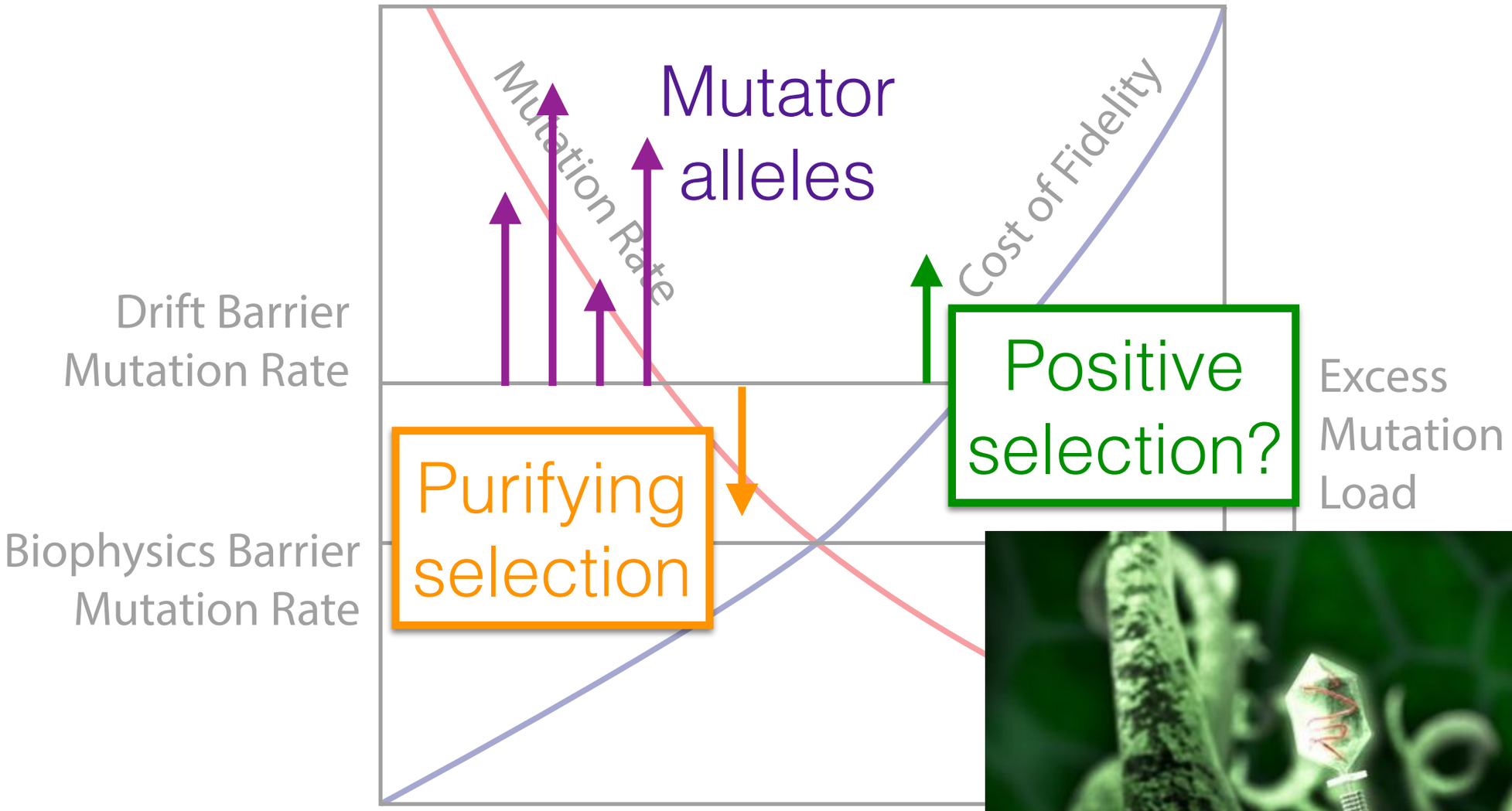
"A recent study (37) detected evidence for an excess of C>T mutations at TCC contexts, the mutation type most prevalent in signature A, accumulating in the human germ line between 15,000 and 2000 years ago. If this human mutation pulse is due to signature A, it could indicate a shared environmental exposure that was once widespread but has now disappeared."

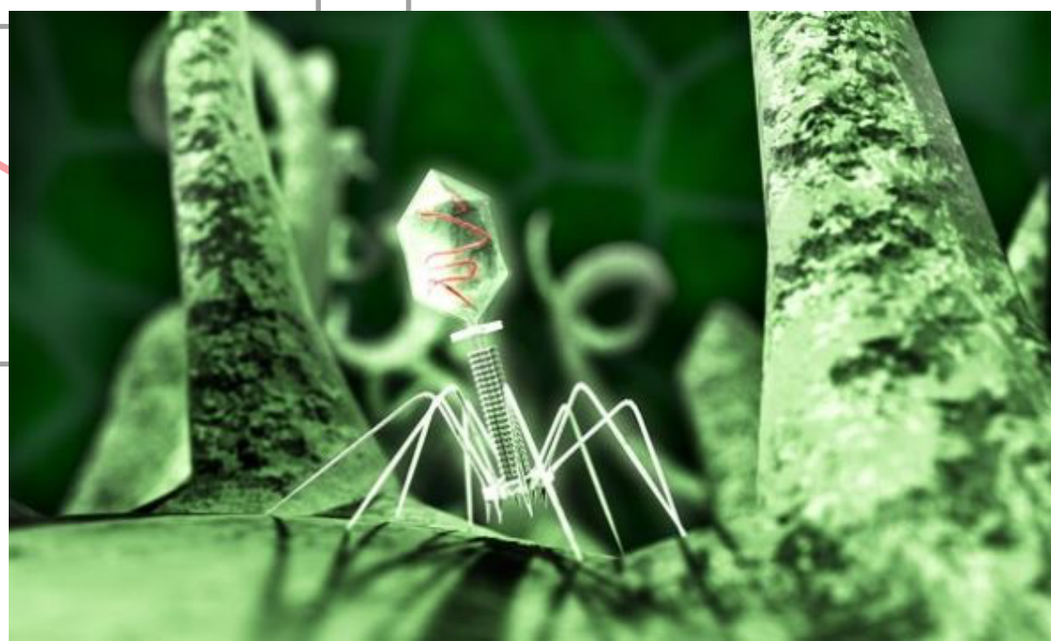
- Canine transmissible venereal tumors (CTVTs) all descend from an ancestral tumor in a dog who lived 4000 to 8500 years ago
- CTVTs experienced a high load of GTCCA>GTTCA mutations that ceased ~1,000 years ago
- Same timeframe as the European mutation pulse and similar (though not identical) sequence bias

# Future direction: are mutation pulses the relics of lost mutator alleles?



# How mutator alleles could promote rapid mutation spectrum turnover

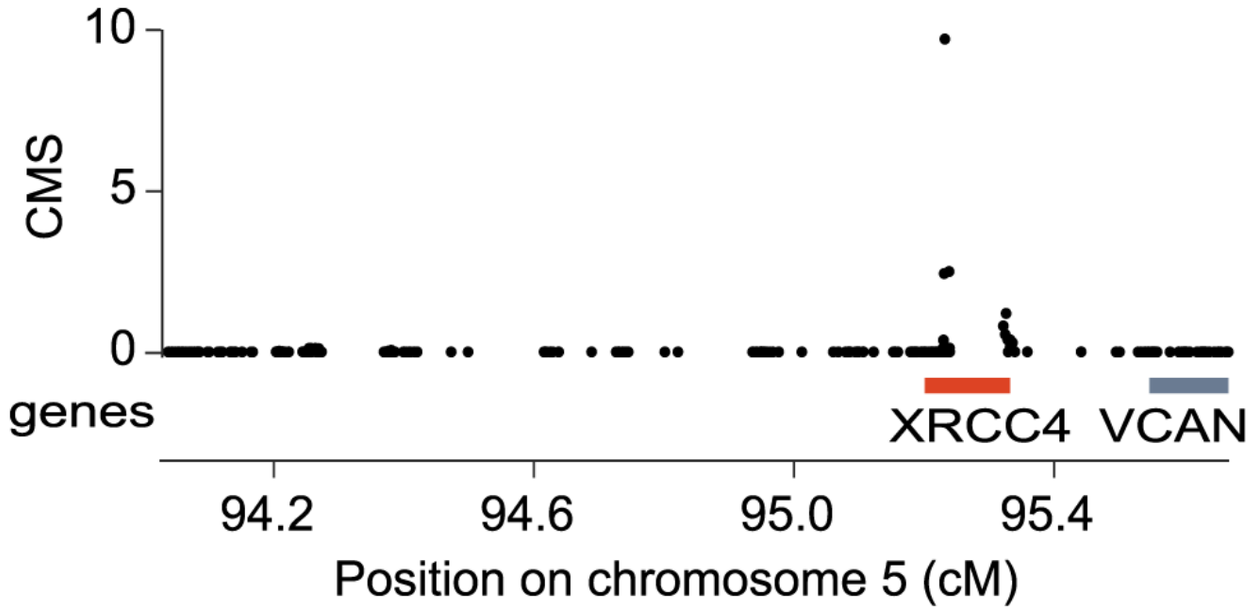


Sawyer and Malik *PNAS* 2006

# Positive selection in DNA repair genes and other housekeeping genes

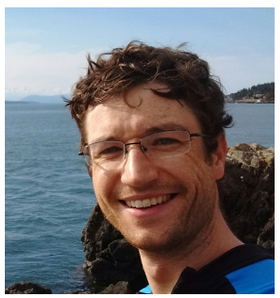
- BRCA1 & BRCA2 are under positive selection in primates
- 5 Nonhomologous end joining genes experienced positive selection during primate evolution, incl XRCC4 which has been under selection in Europeans
- Iron-uptake receptor TfR1 evolves under positive selection to avoid facilitating viral entry



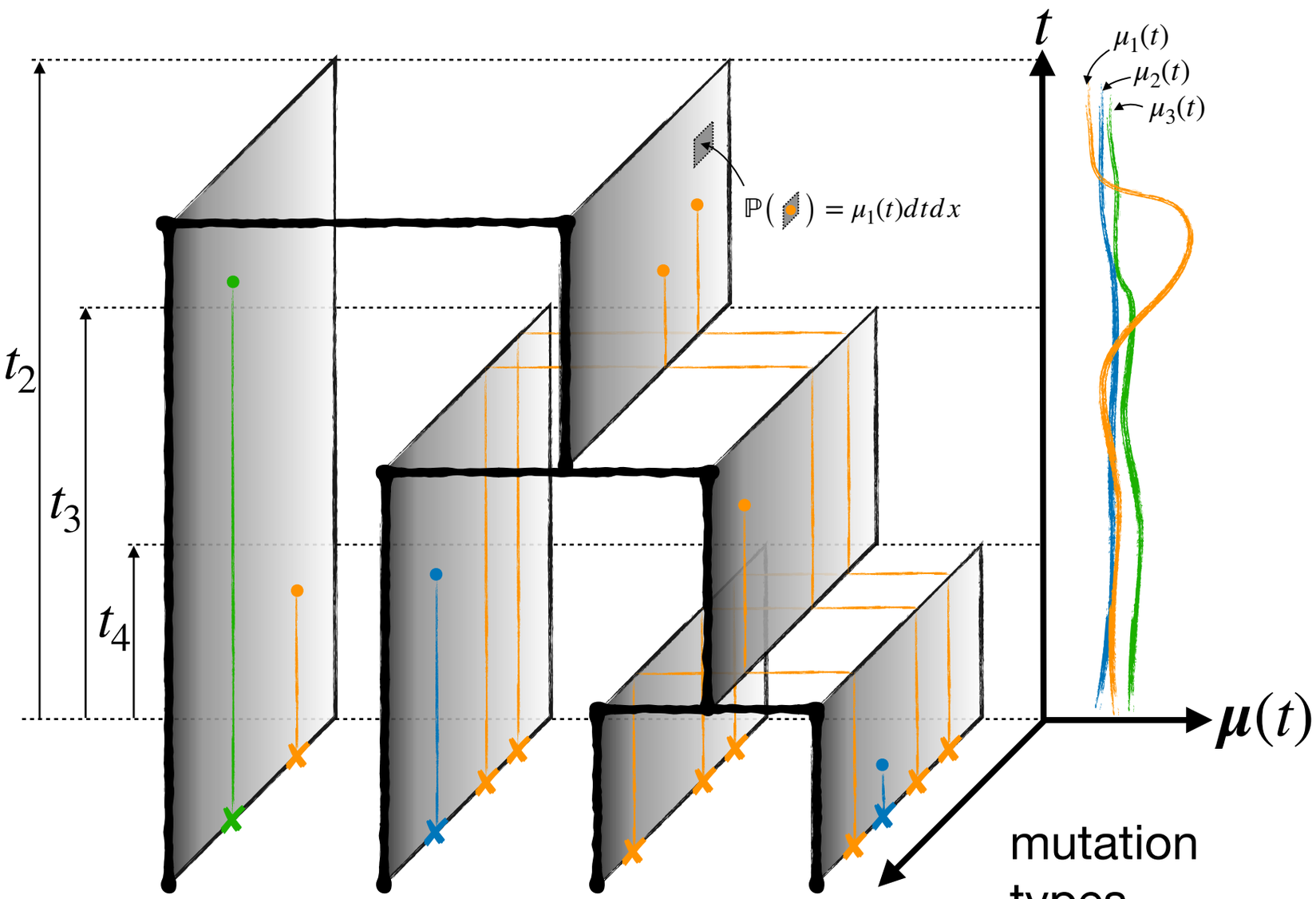

**Demogines, et al. 2010**
**Demogines, et al. 2013**

# MuSHI: Mutation Spectrum History Inference

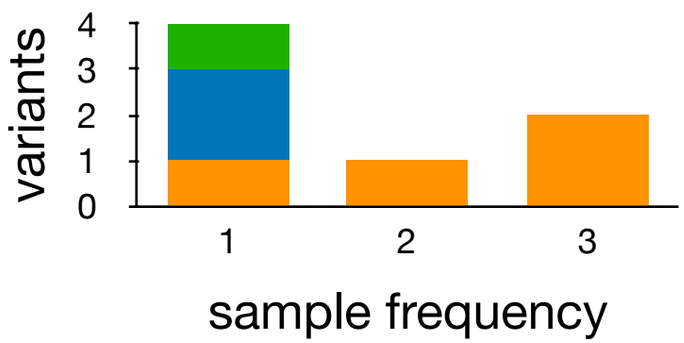
**Will Dewitt**



**Kameron Decker Harris**



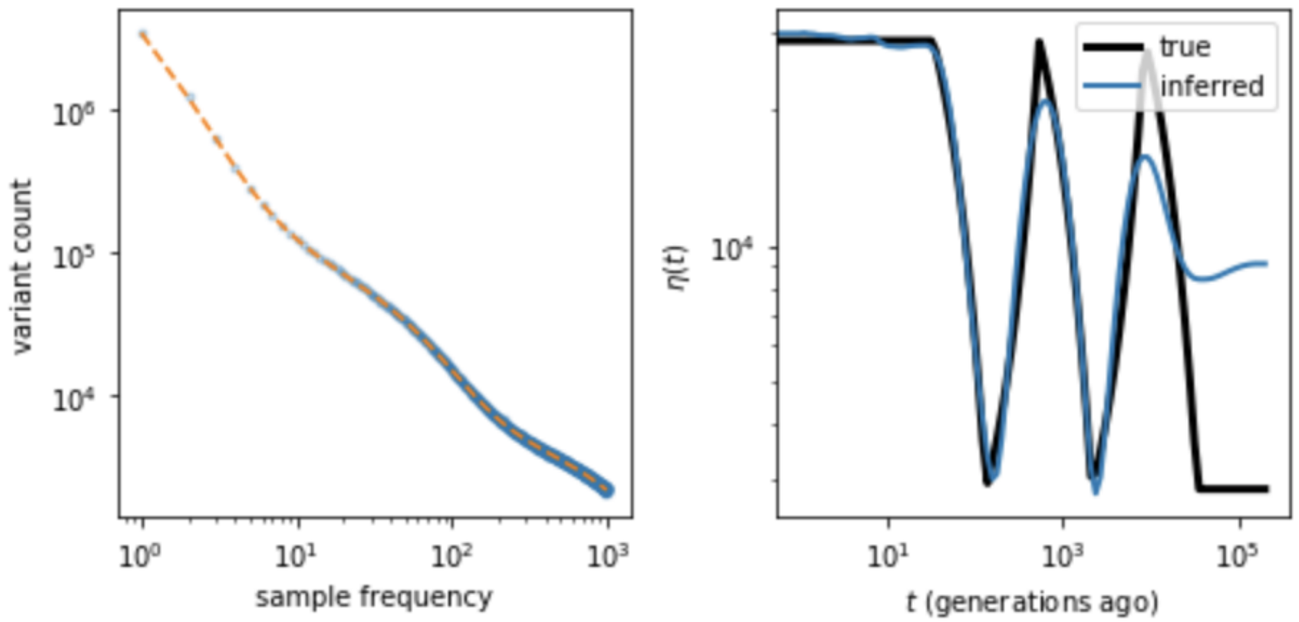

*k*-SFS:




$\longrightarrow$ $X = \begin{bmatrix} 1 & 2 & 1 \\ 1 & 0 & 0 \\ 2 & 0 & 0 \end{bmatrix}$ } sample frequency

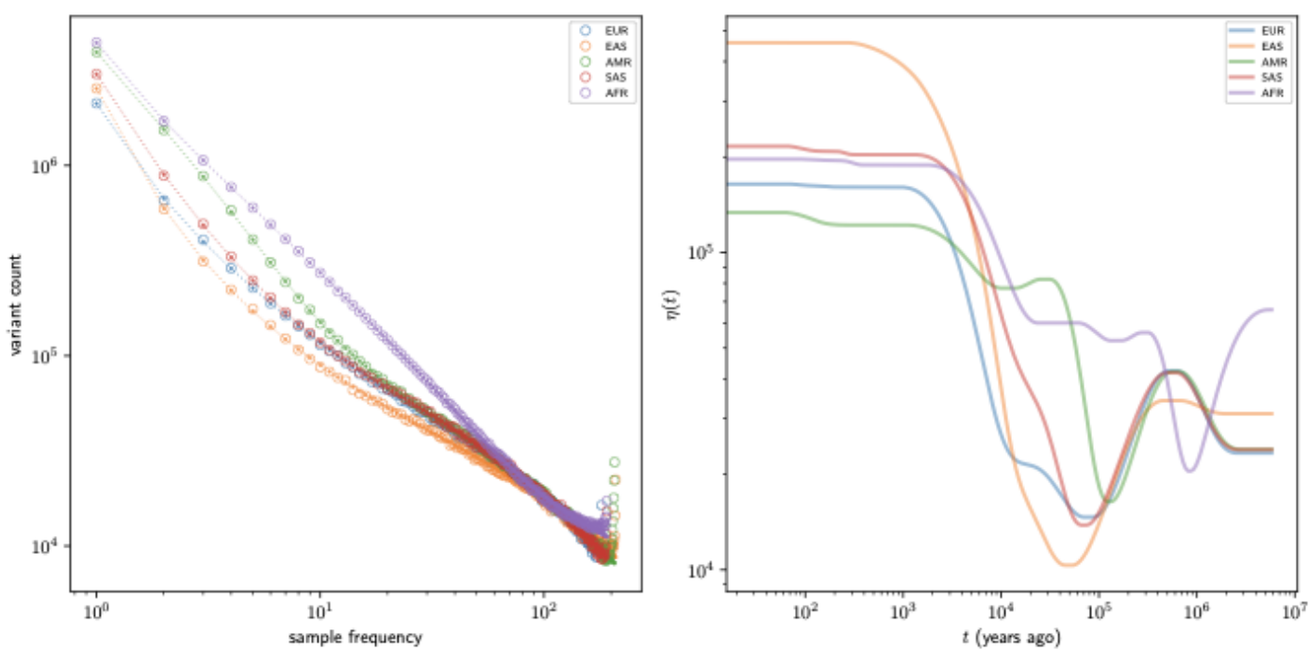
**Dewitt, Harris, and Harris, in prep**

# MuSHI estimates demographic history jointly with the mutation spectrum history (mush)

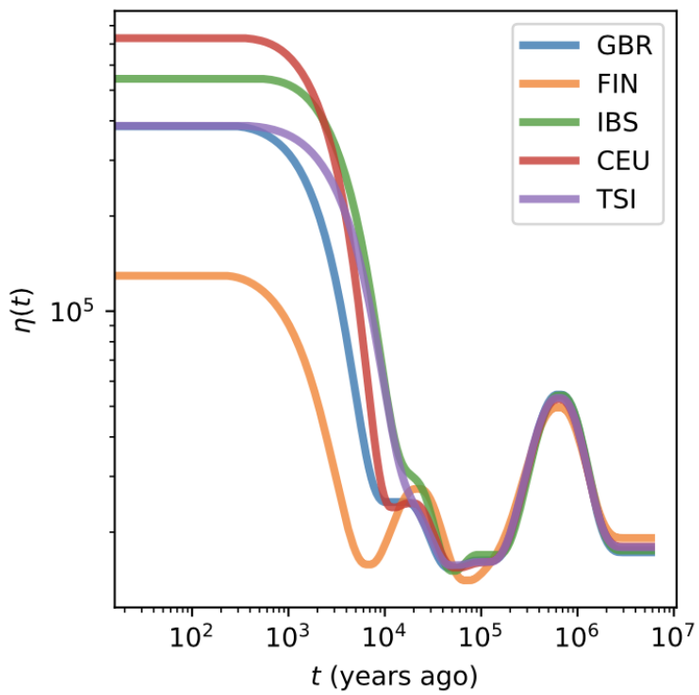



**Simulated data**

## 1000 Genomes Continental Groups




## 1000 Genomes Europeans

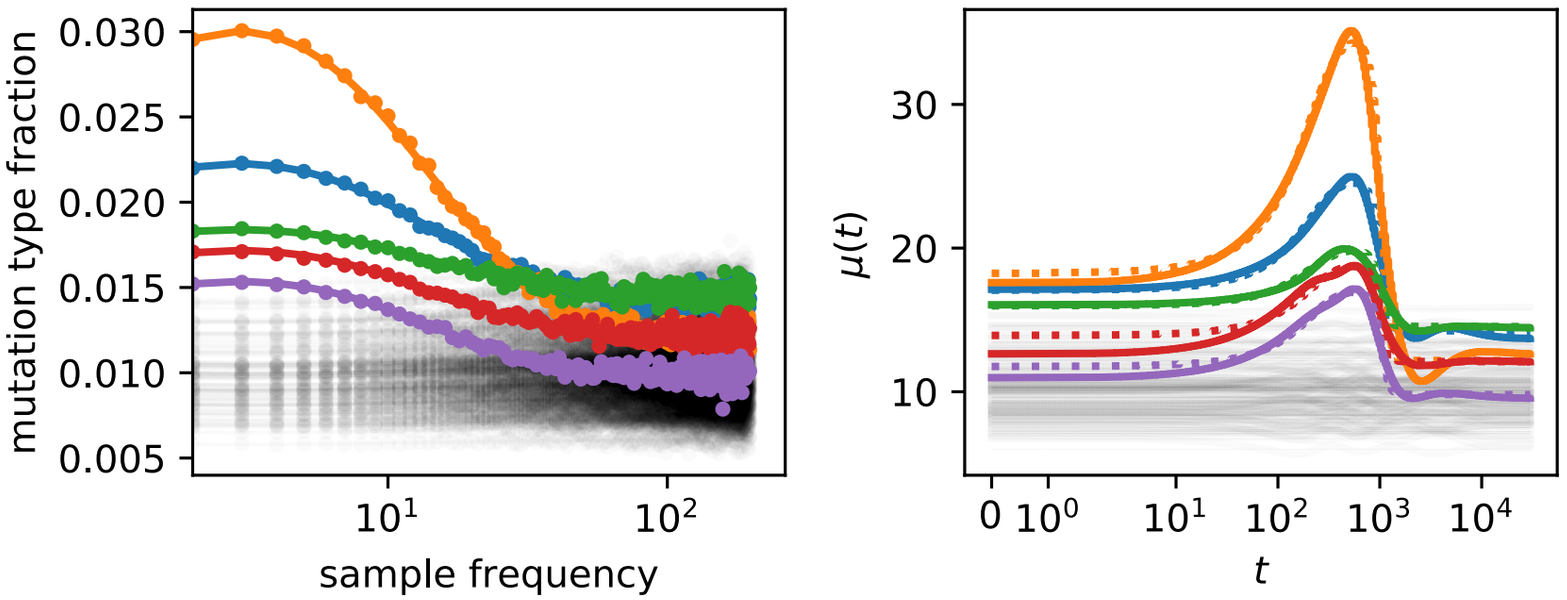



**Dewitt, Harris, and Harris, in prep**

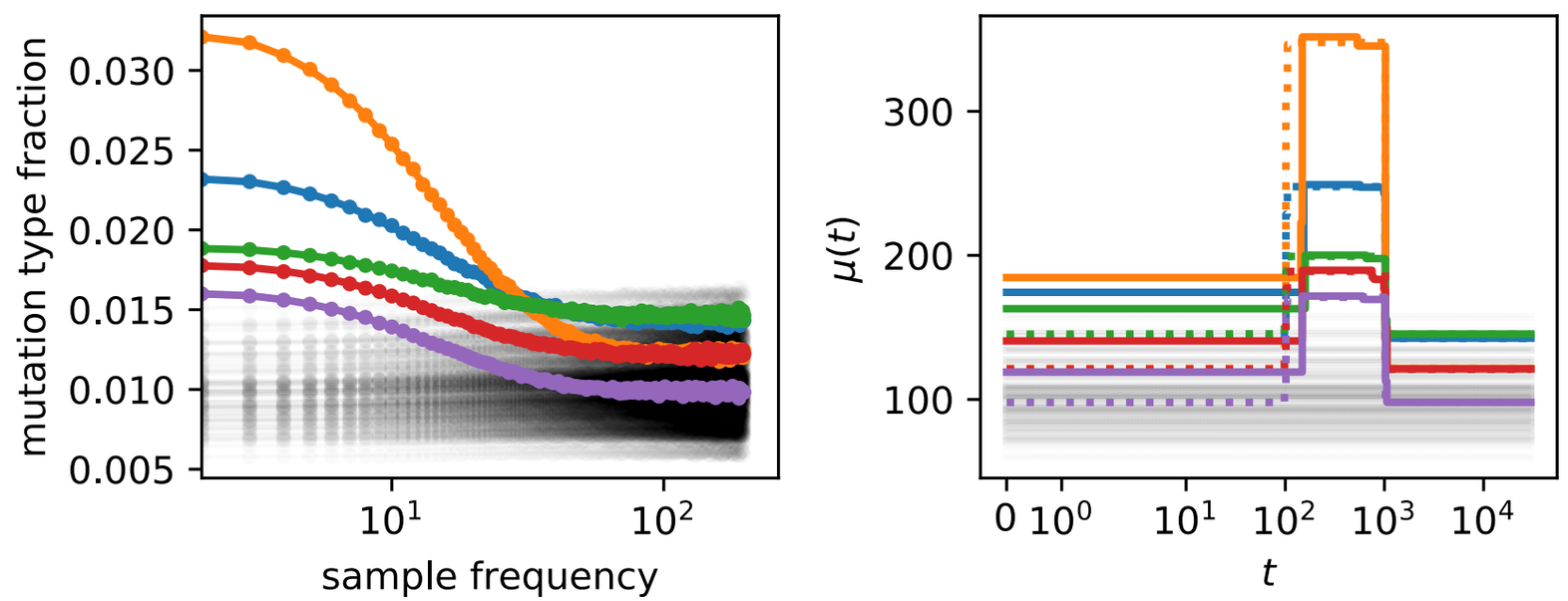
# A simulated example of pulse recovery

96 MUTATION TYPES WITH LATENT PULSE SIGNATURE AFFECTING 5

$\ell_2$-smooth pulse:

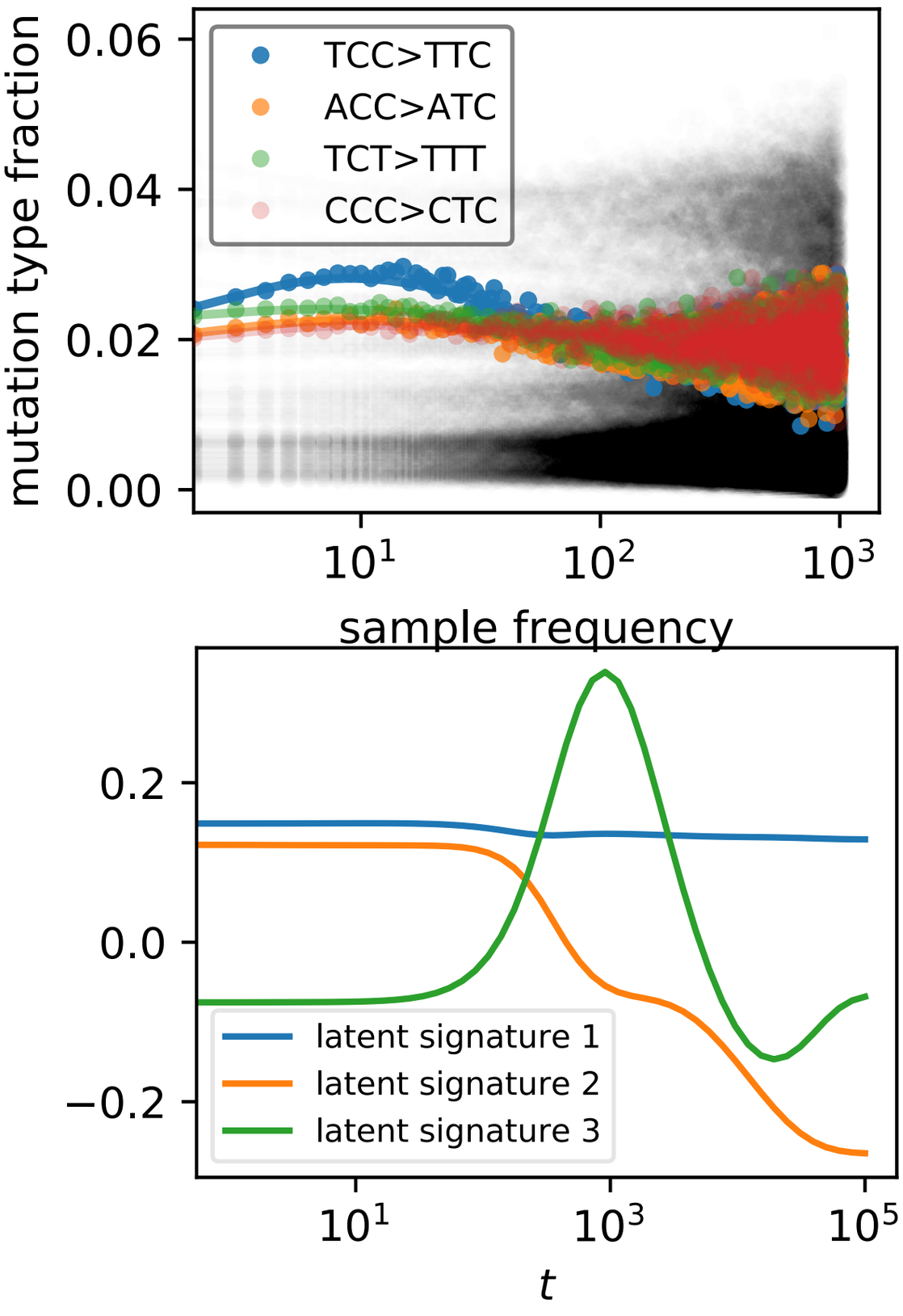


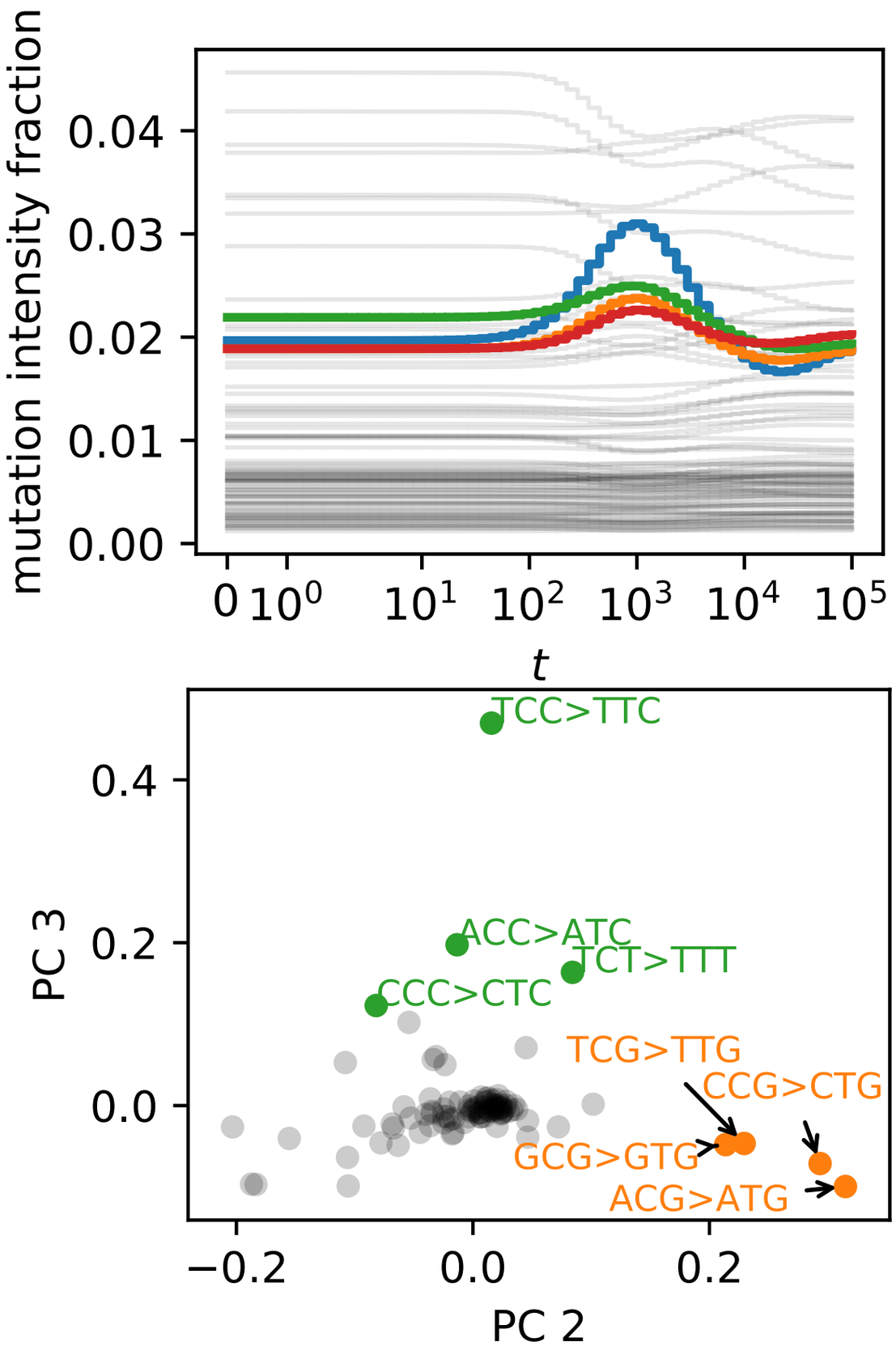

$\ell_1$-smooth pulse:



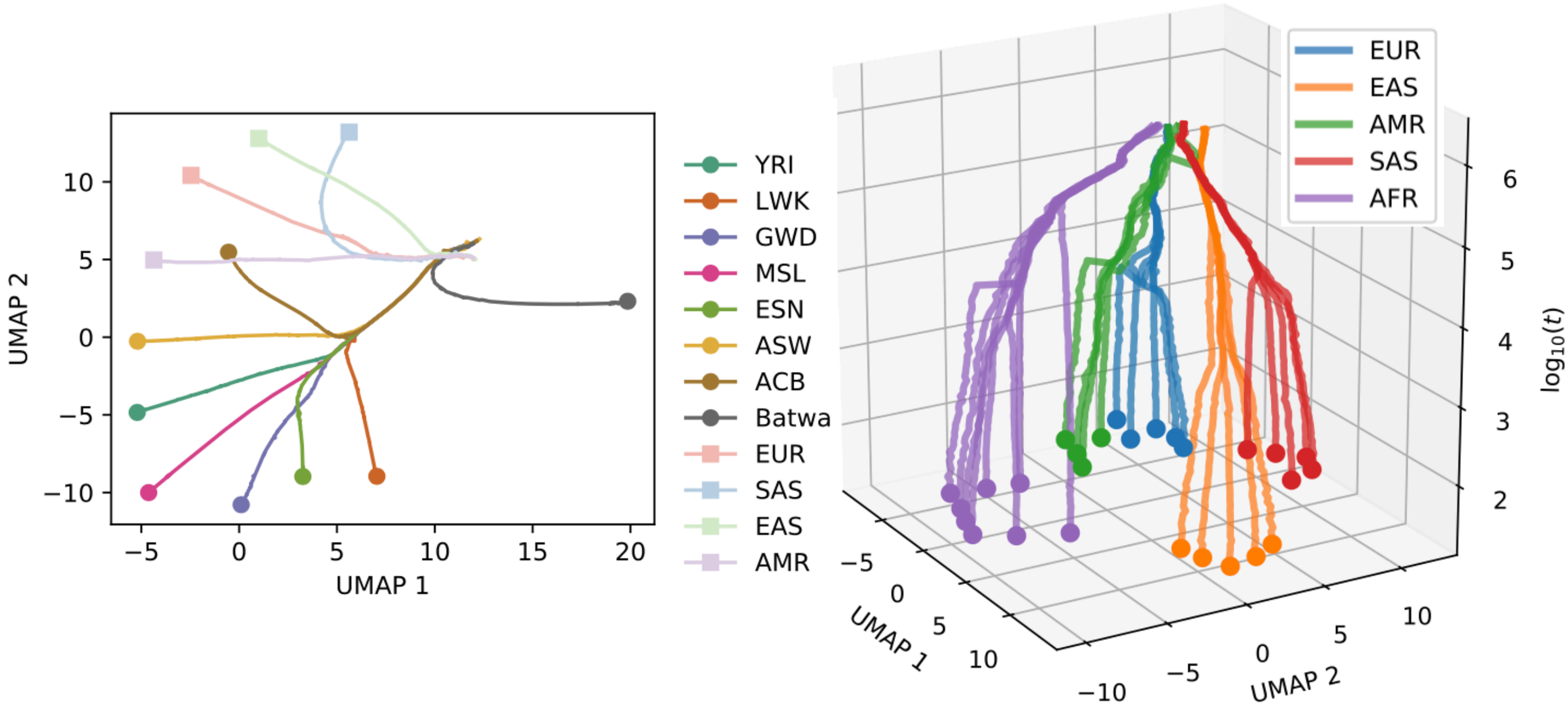

**Dewitt, Harris, and Harris, in prep**

# Automatic mutational signature extraction from Europeans (CEU)



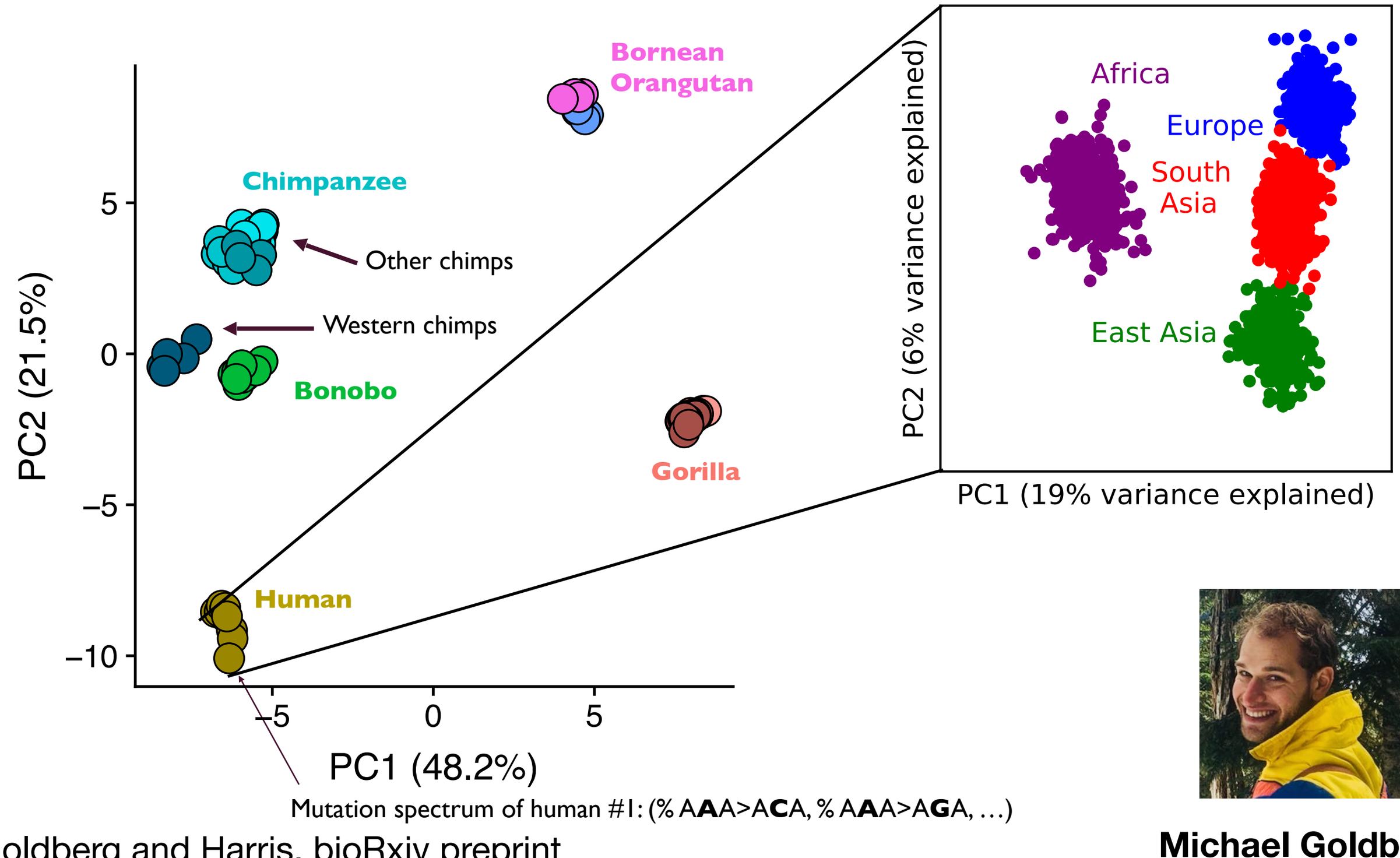



# UMAP visualization of mutation spectrum divergence over time




Dewitt, Harris, and Harris, in prep

# Great ape species display greater mutation spectrum drift than human populations do



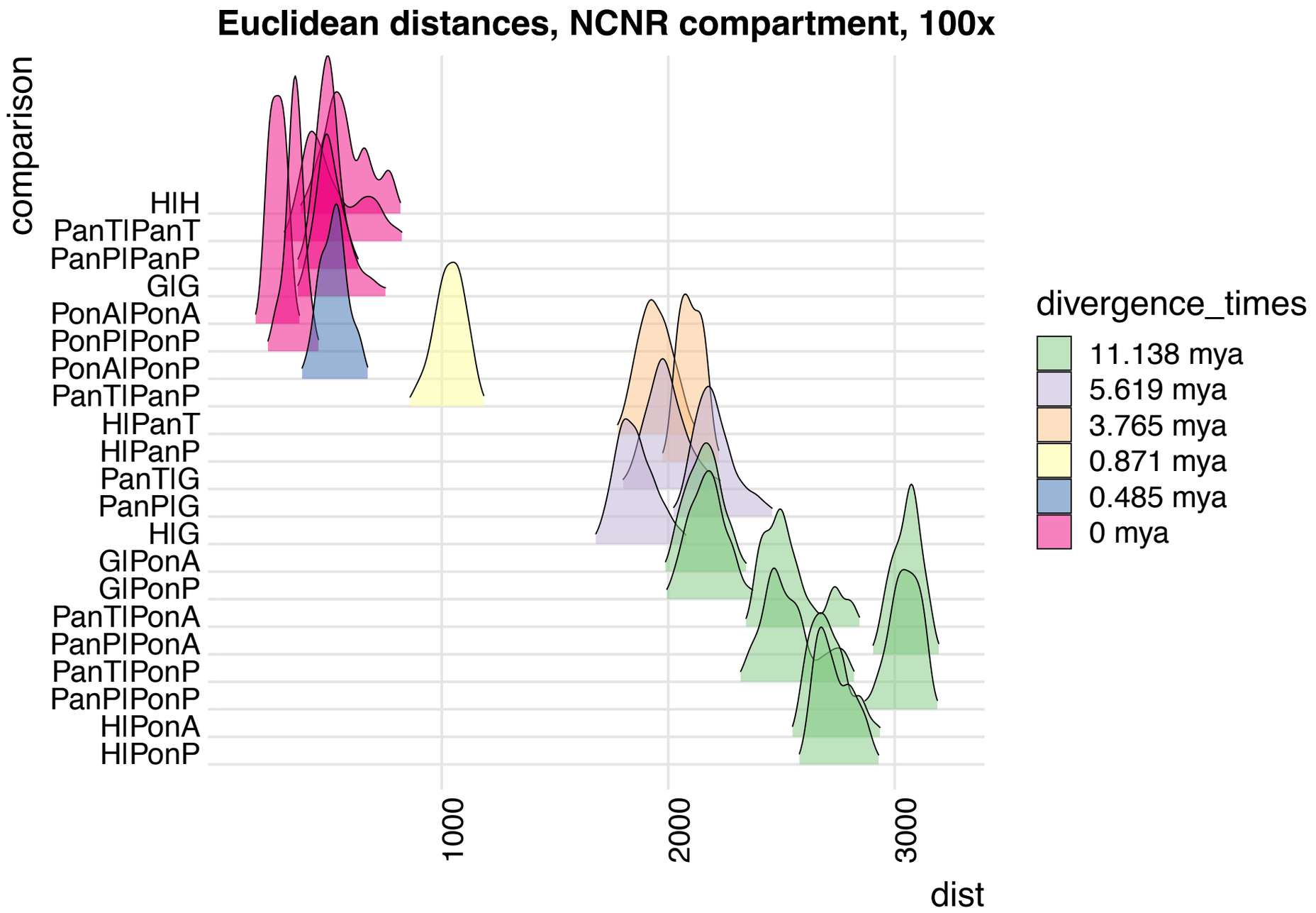

Mutation spectrum of human #1: (% A**A**A>A**C**A, % A**A**A>A**G**A, …)



**Michael Goldberg**

Goldberg and Harris, bioRxiv preprint

# Ape mutation spectra cluster by phylogeny, pointing to fixation of genetic mutators (not environmental mutagens)




Goldberg and Harris, bioRxiv preprint

# A case study of a mutational process that complicates population genetics

Multinucleotide mutations (MNM) are nearby SNPs that appear in the same generation

*AAAGTTAGCCGACAC*

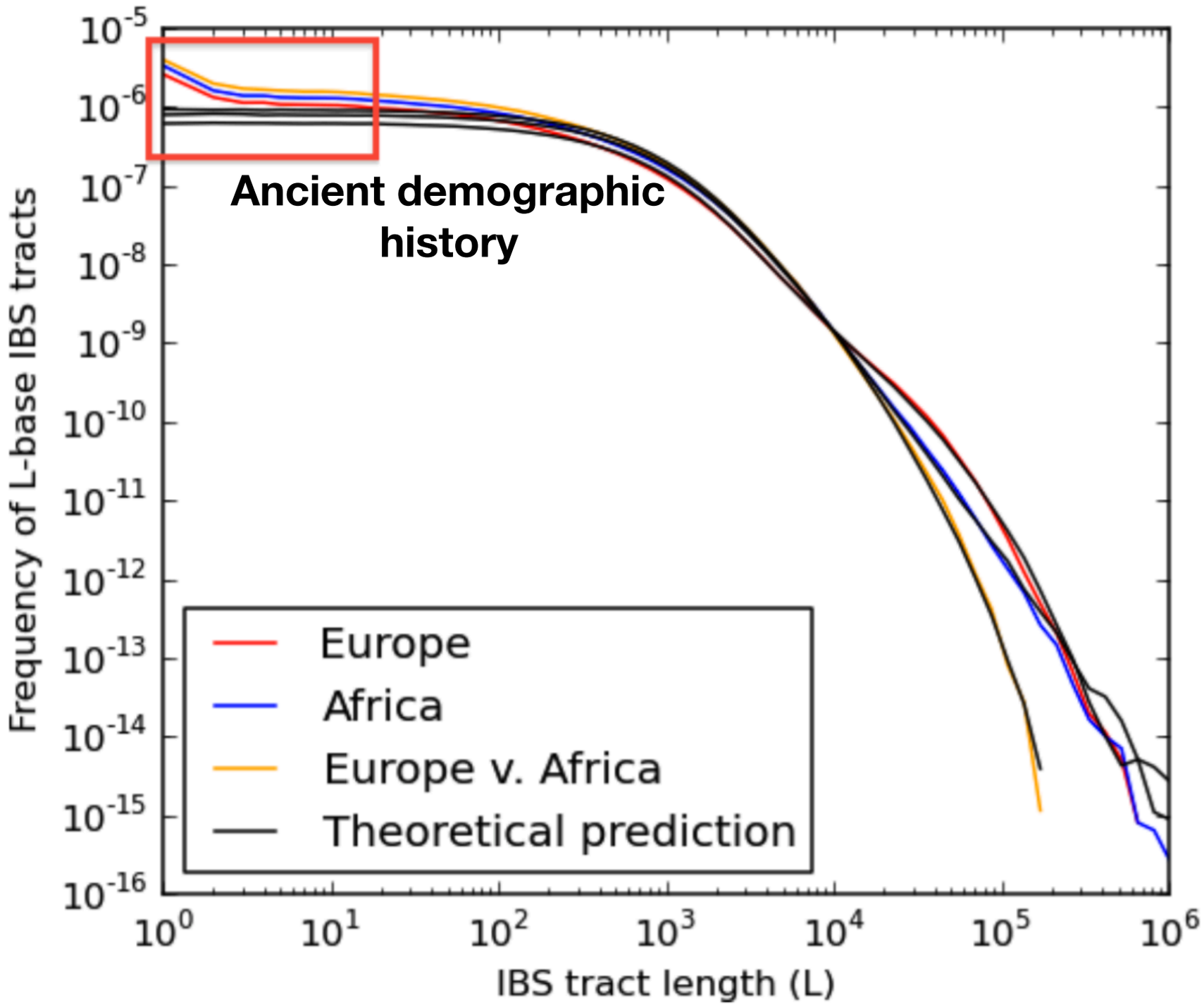
↓

*AAAGATAACCGACAC*

**Schrider, et al. 2011**

**Harris and Nielsen 2014**

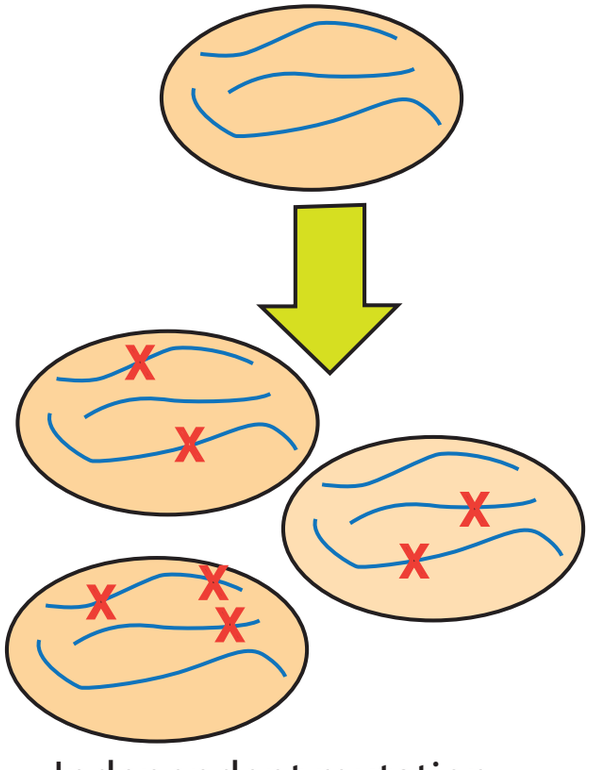
# Effect of MNMs in the distribution of tracts of identity by state



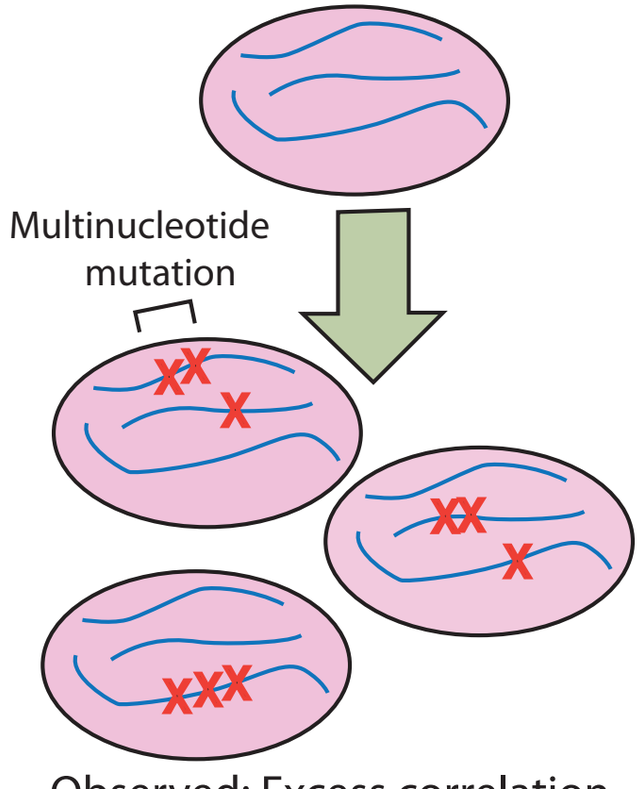
# Direct evidence for MNMs

- Most methods assume that all SNPs arise from rare, independent mutation events

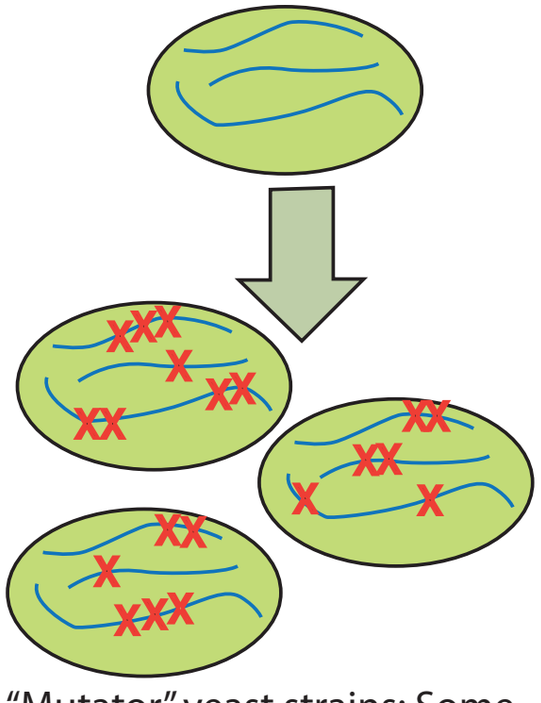
- MA experiments and trio sequences show that *de novo* mutations are too clustered for this to be true



Independent mutation hypothesis

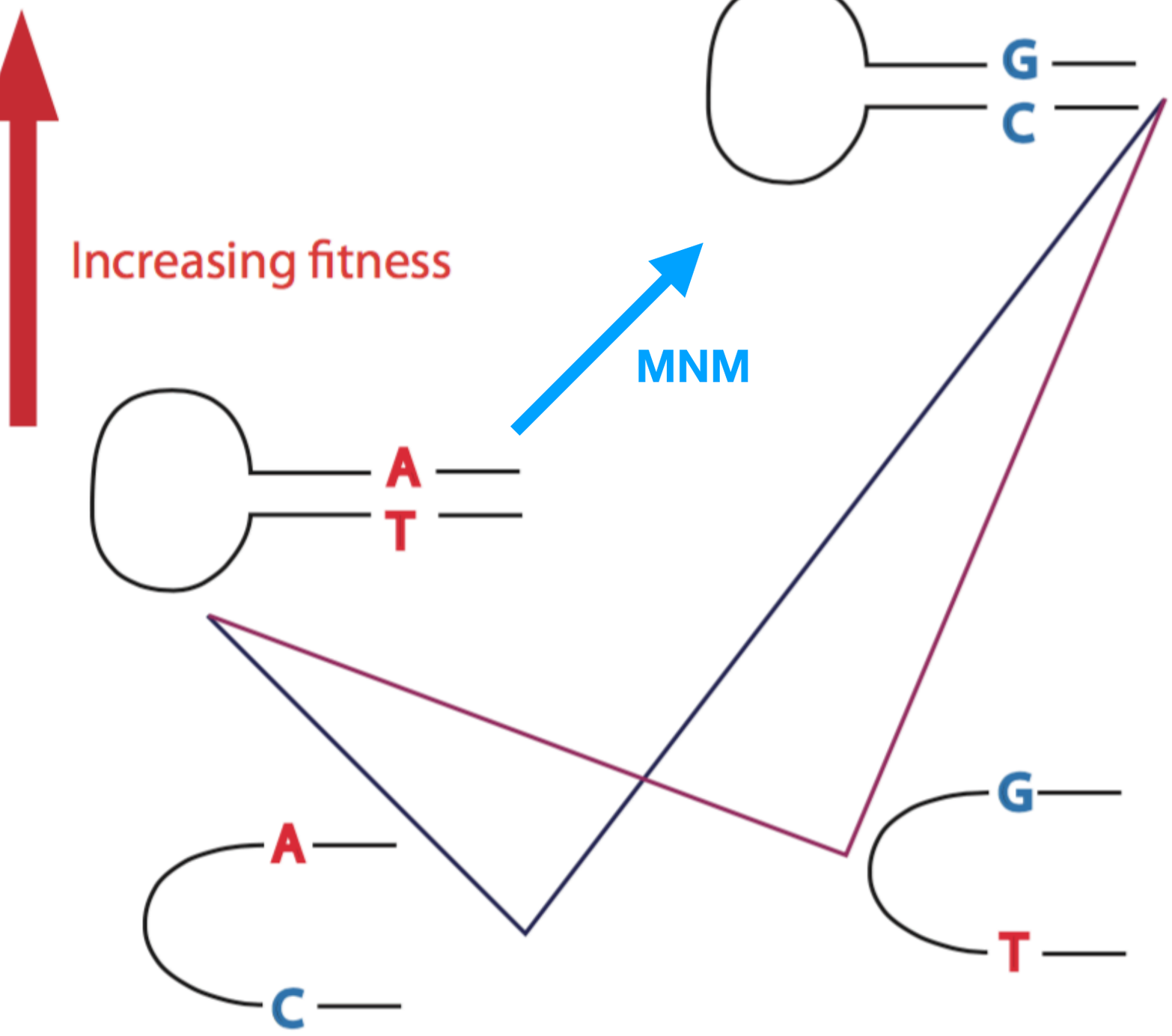



Observed: Excess correlation between *de novo* mutations

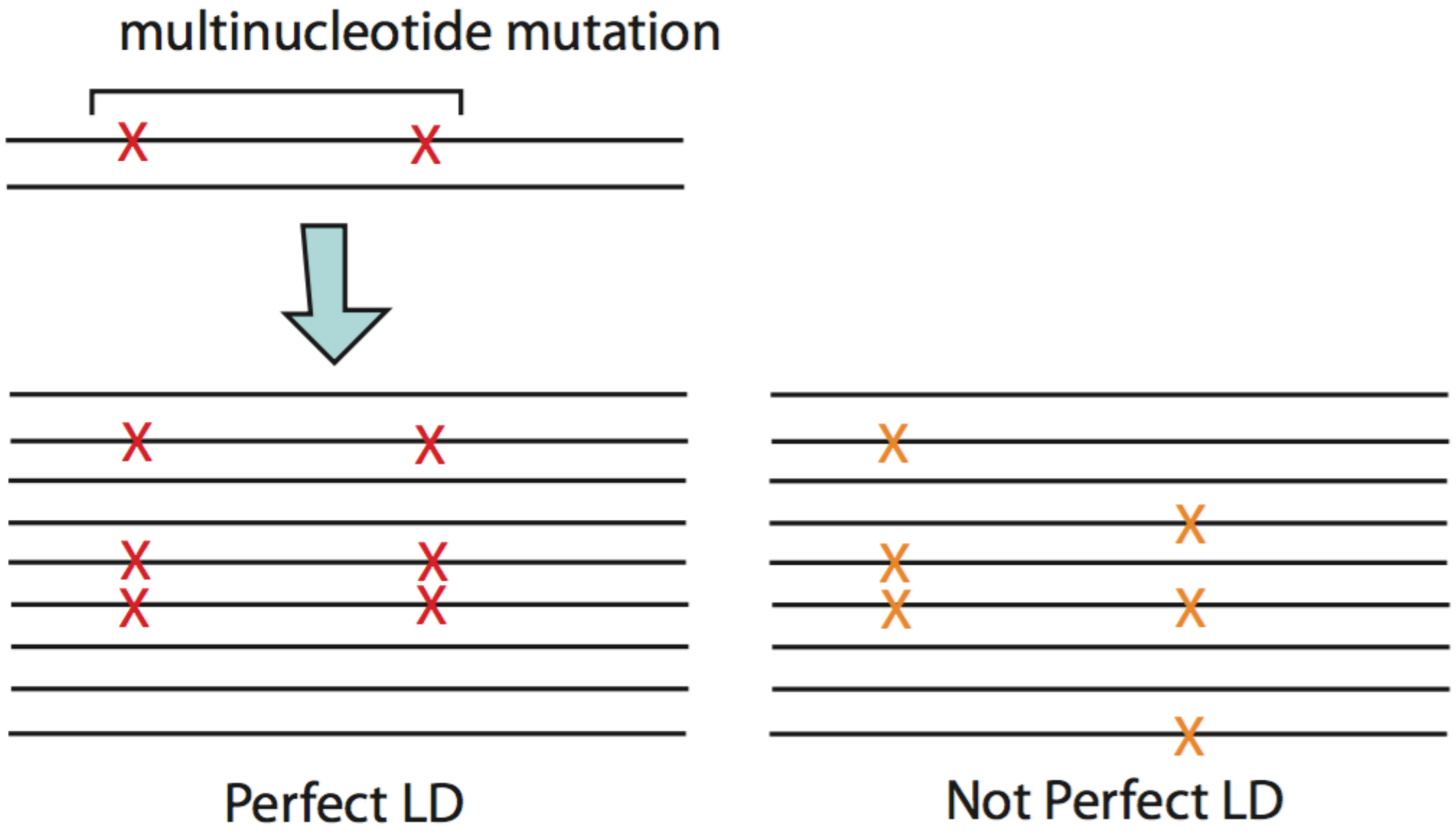


"Mutator" yeast strains: Some abnormal polymerases generate clustered mutations at a higher rate

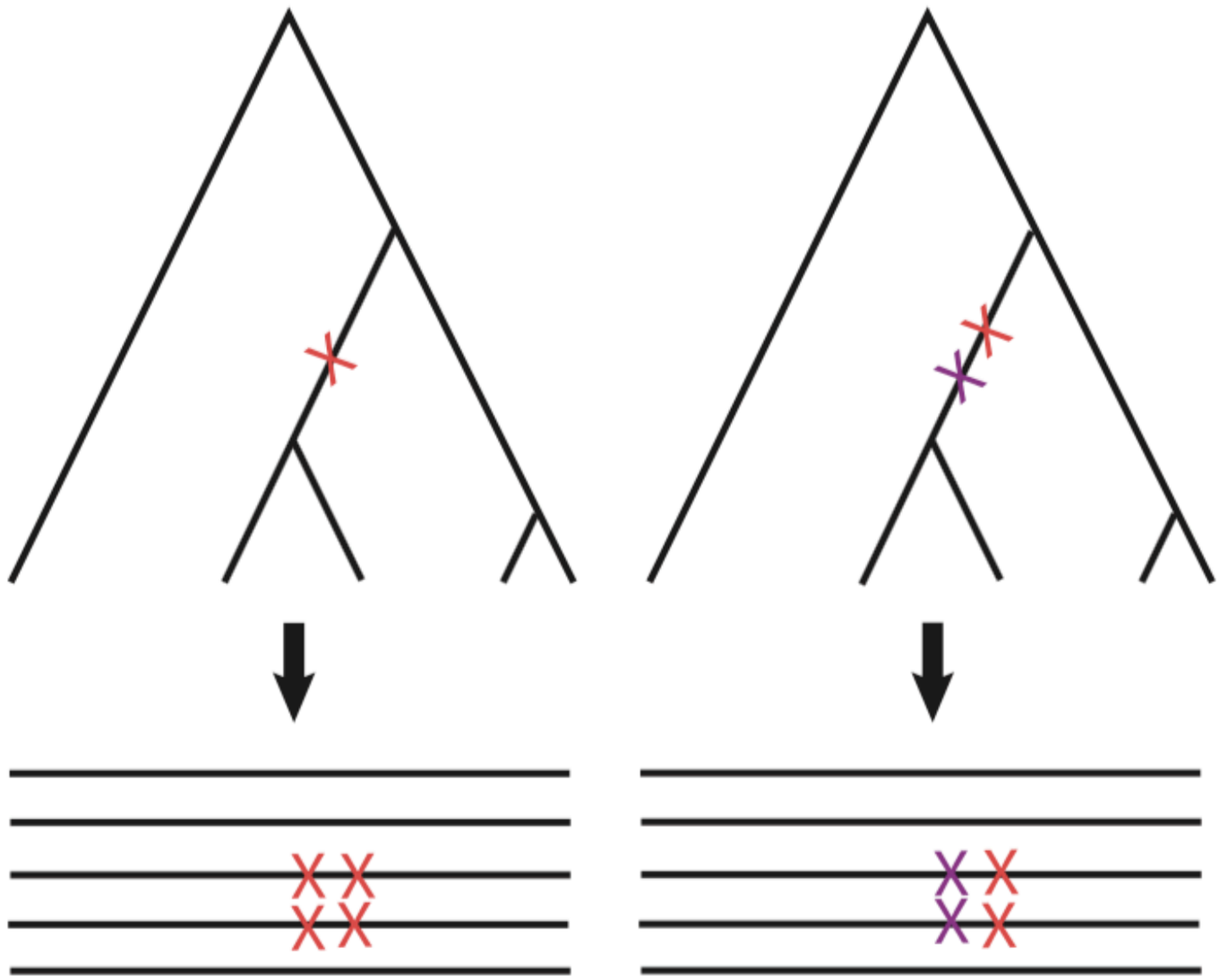
# MNMs could accelerate evolution across fitness valleys



Multinucleotide mutation should create pairs of SNPs in *perfect linkage disequilibrium (LD)*
(derived alleles occur in the same set of individuals)

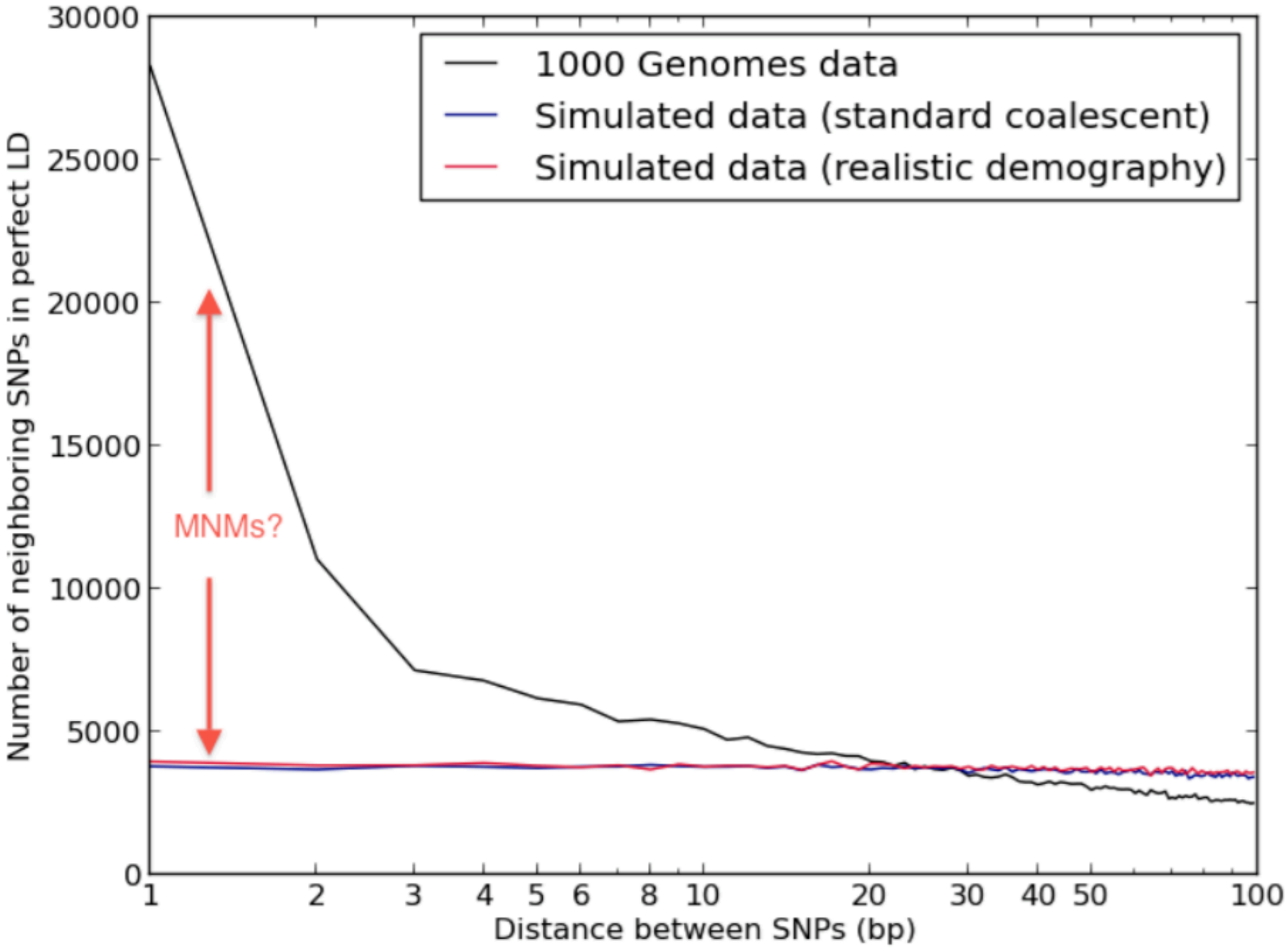


# Independent mutations at neighboring sites can also create SNPs in perfect LD



One MNM

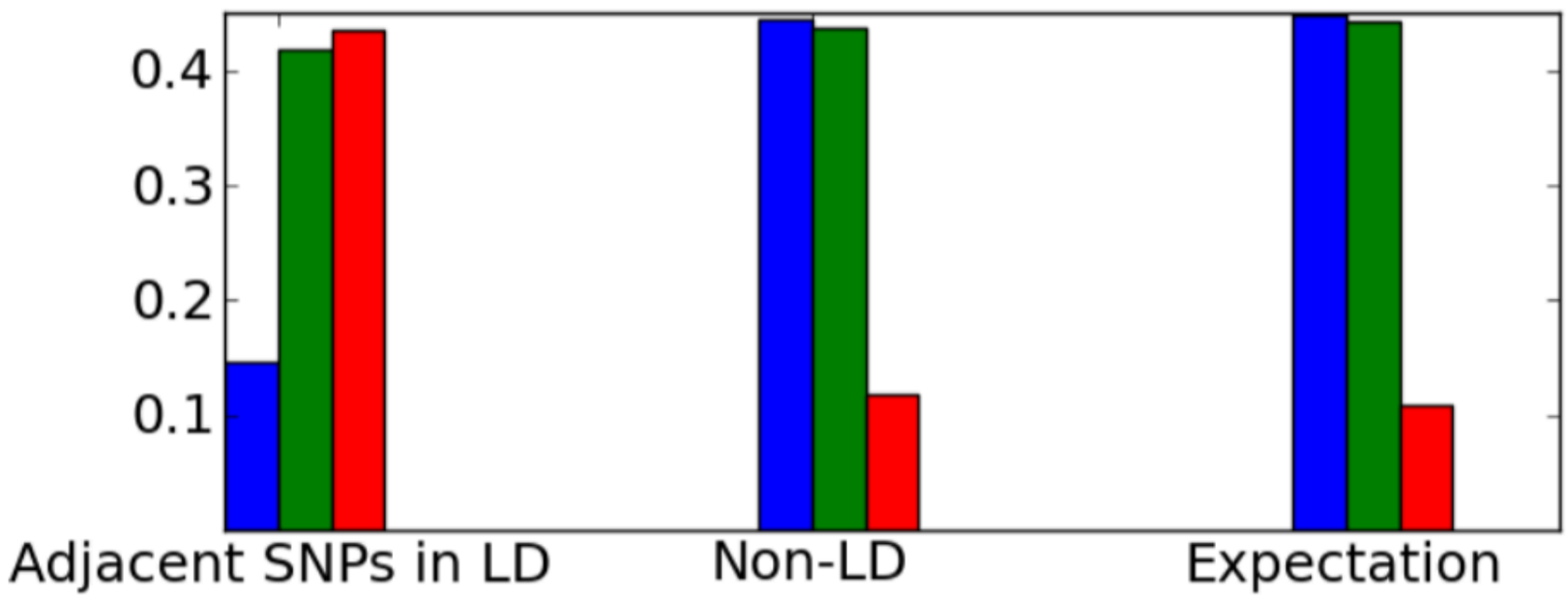
Two independent mutations



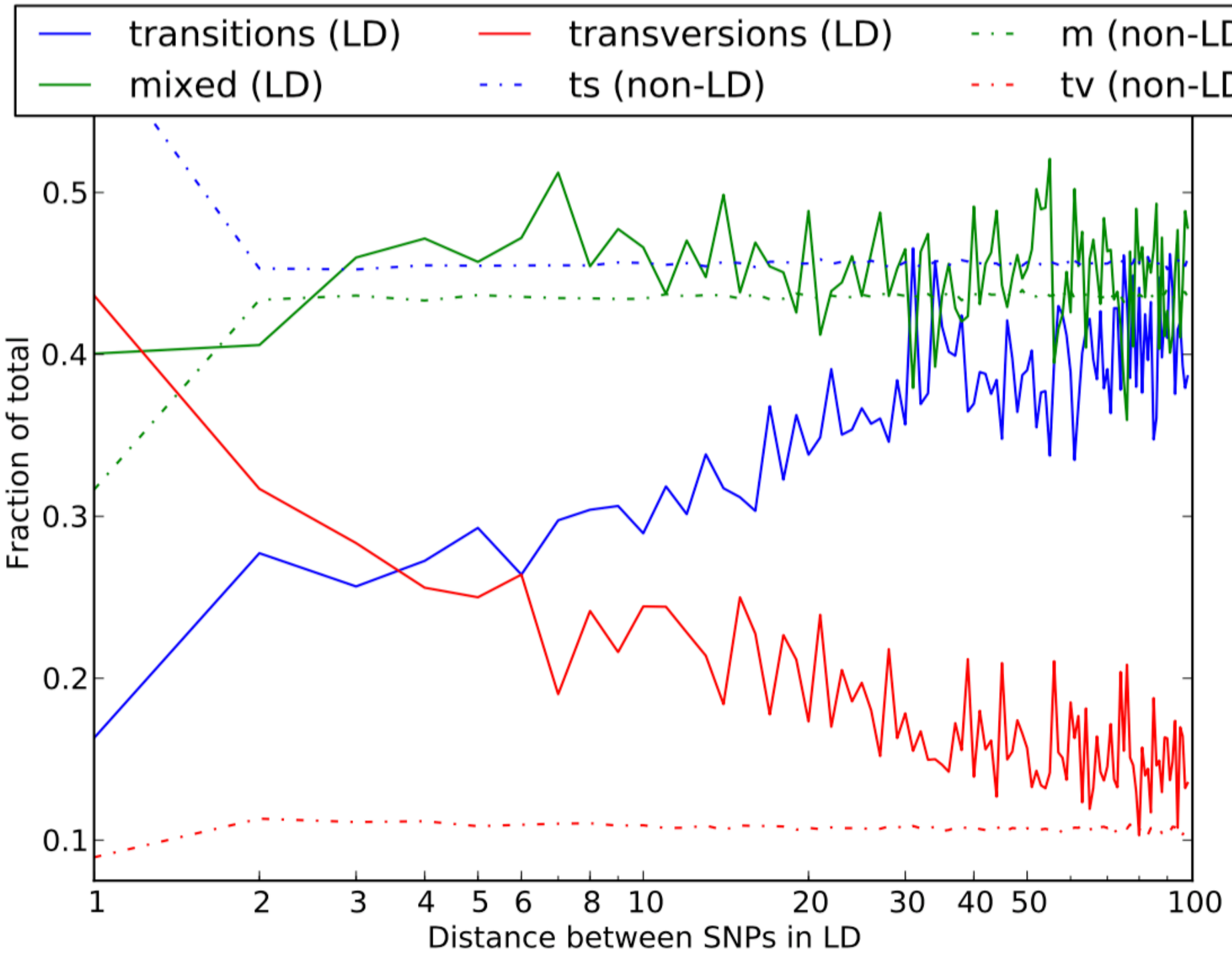

Compared to theoretical predictions, the 1000 Genomes Phase I data (1,092 humans from Africa, Europe, Asia, and the Americas) has excess close-together SNPs in perfect LD

# SNPs in perfect LD are enriched for transversions

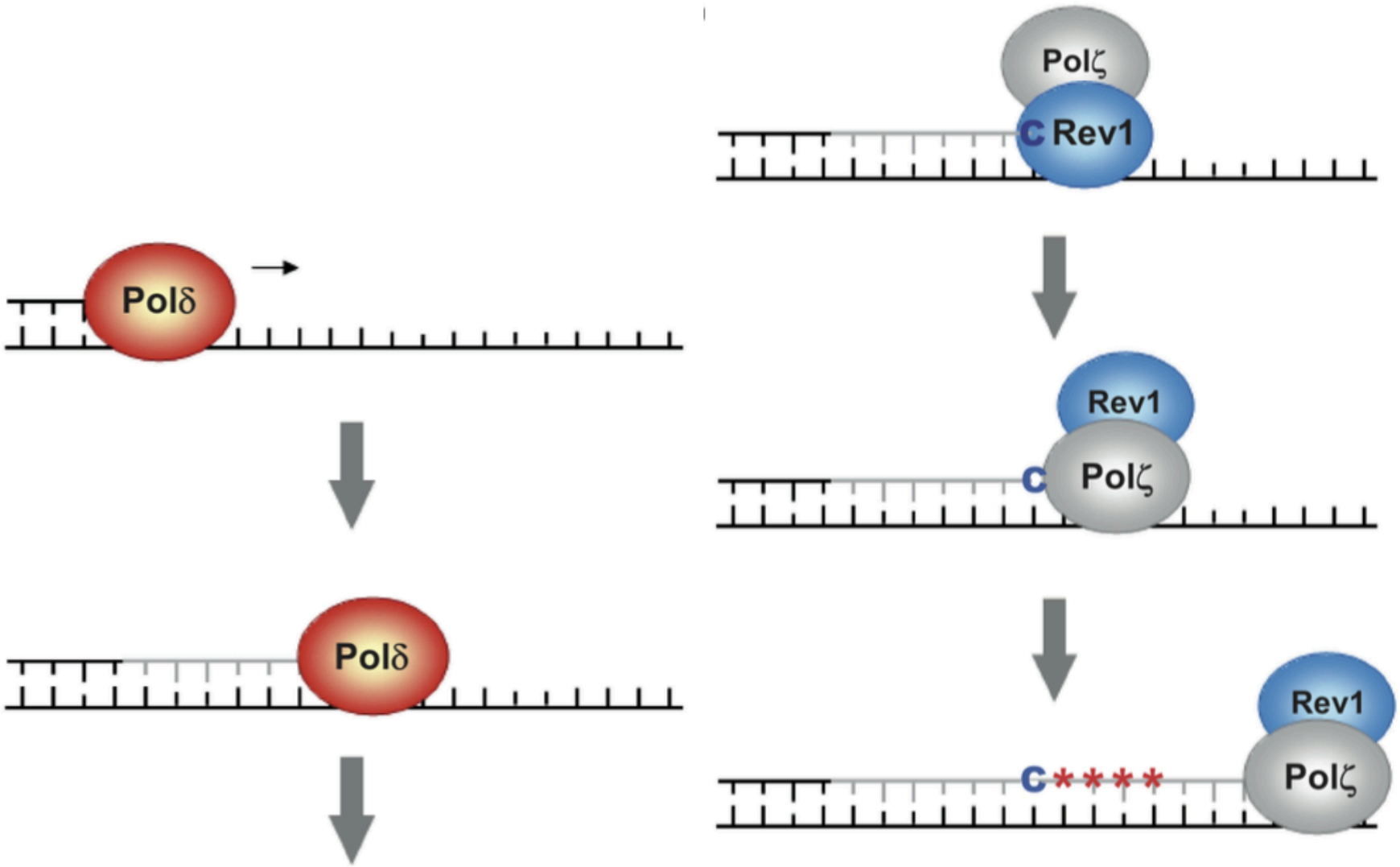
- 66% of human mutations are transitions (A>G, G>A, C>T, T>C)
- Pairs of SNPs in perfect LD are enriched for transversions, suggesting a different balance of mutational signatures



# Transversion-enrichment as a function of the distance between linked SNPs



# A candidate mechanism: error-prone translesion synthesis




Northam *et al.*, *Nucleic Acids Res.* 2014

# Matching mutational signatures between human variation and laboratory yeast

Environmental and Molecular Mutagenesis 53:777–786 (2012)

*Research Article*

## DNA Polymerase zeta Generates Clustered Mutations During Bypass of Endogenous DNA Lesions in *Saccharomyces cerevisiae*

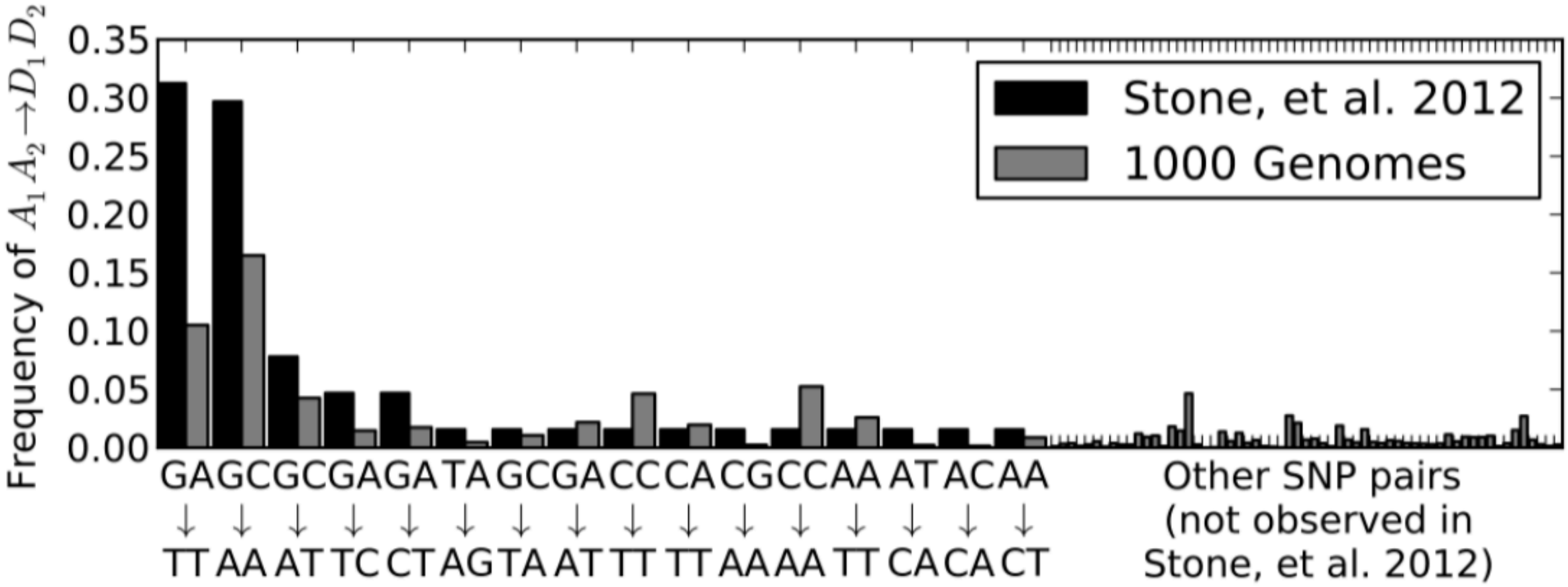
**Jana E. Stone, Scott A. Lujan, and Thomas A. Kunkel***

*Laboratory of Molecular Genetics and Laboratory of Structural Biology, National Institute of Environmental Health Sciences, NIH, DHHS, North Carolina*

- Stone, et al. created yeast deficient in nucleotide excision repair machinery and observed a high MNM rate

- Mechanism: increased translesion synthesis by Pol Zeta

# A matching dinucleotide mutational signature



# Further characterization of the Pol zeta mutational signature

- GC>AA mutations are concentrated in late-replicating regions of the genome

- Usually occur in GCG context, triggered by CpG deamination followed by polymerase stalling

- CpG deamination is triggered by transcription; usually occurs on transcribed strand

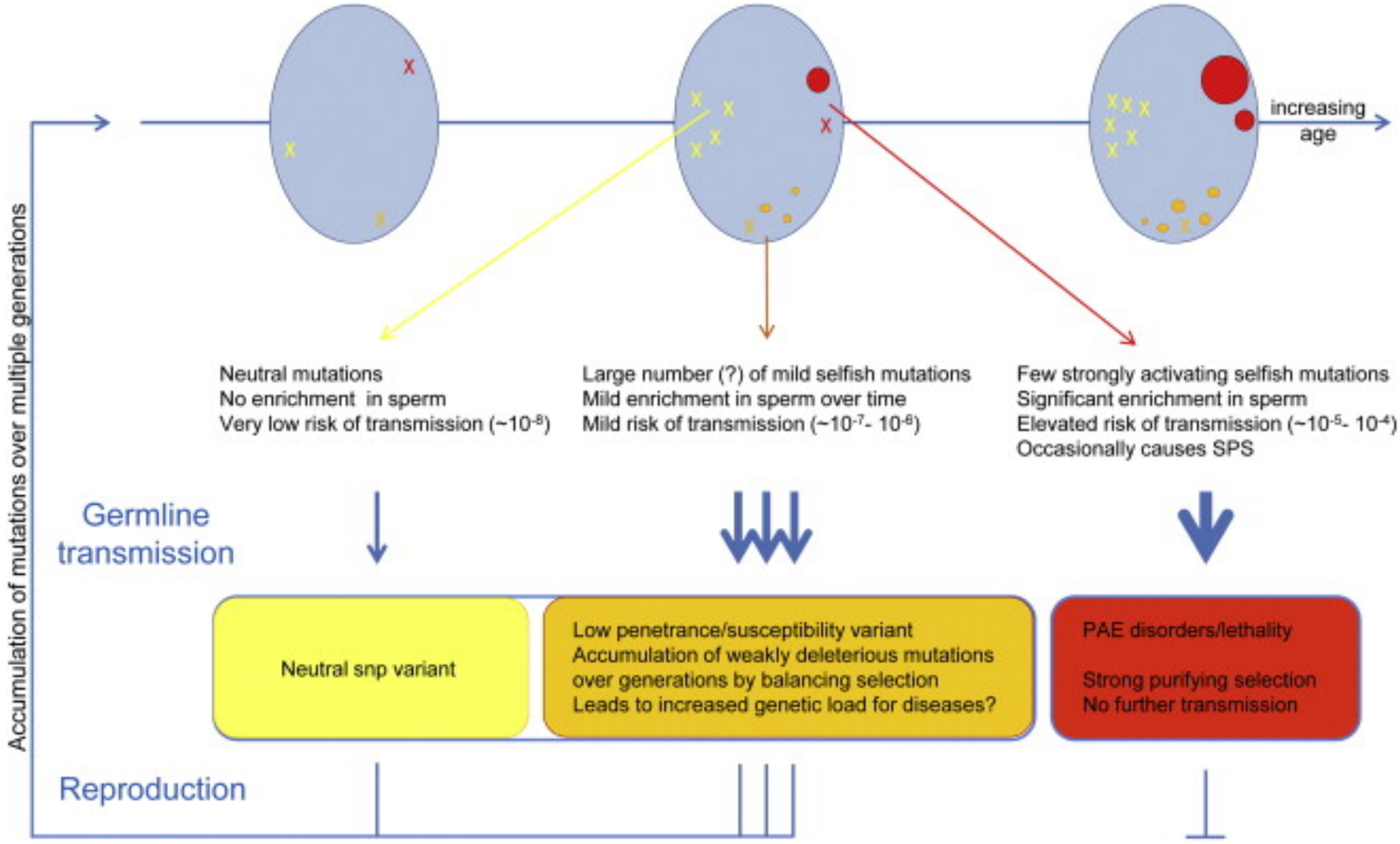
- Some genes contain GC>TT mutation hotspots, including HRAS where the mutation causes the Mendelian disorder Costello Syndrome

**Seplyarskiy, et al. 2015**

# Costello Syndrome is caused by selection within the aging testis

- A high penetrance Mendelian disease caused by a nonsynonymous point mutation in the HRAS oncogene
- Causes developmental delay and early childhood tumors
- Most commonly caused by a GC>TT mutation with a mutation rate of $10^{-5}$ per generation (normal mutation rate is $10^{-8}$ per site per generation)
- Biggest risk factor is paternal age

# HRAS mutations experience selfish selection within the testis




**Goriely and Wilkie 2012**