Inference of population history and mutation biology from genetic variation data

Kelley Harris

Evolutionary Cell Biology @ KITP
September 29, 2015

Acknowledgements

Rasmus Nielsen

INTEGRATIVE BIOLOGY

Stanford University

department of GENETICS

Jonathan Pritchard

EECS

Yun Song

wellcome trust sanger institute

Richard Durbin

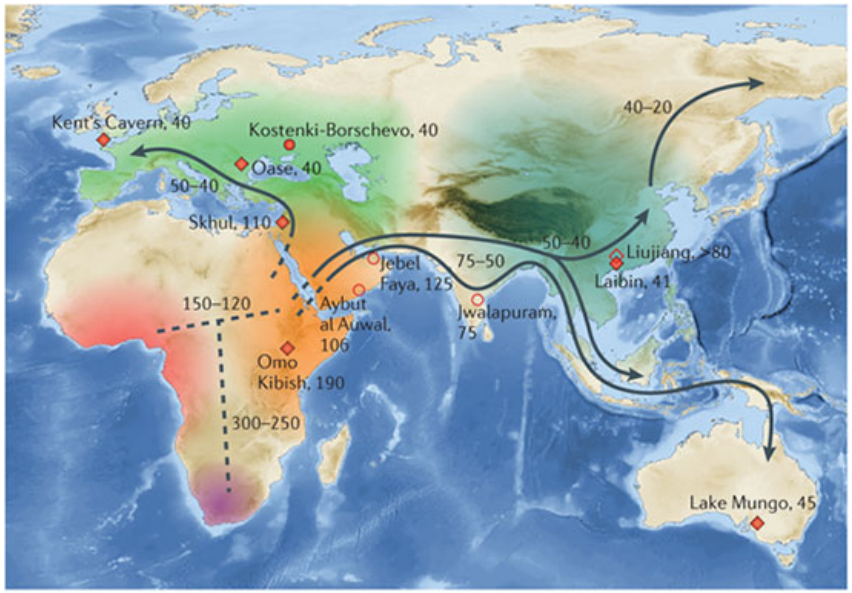
Part I: Using coalescent theory to infer demographic history

- Estimate past effective population size (N) from genomes sampled in the present
  - Usually smaller than census population size
  - Inversely proportional to speed of genetic drift
  - Directly proportional to effectiveness of natural selection
- Divergence & gene flow between populations

Part II: Inferring mutation biology

- Signatures of error-prone DNA polymerase activity in the germline
- Recent mutation rate evolution

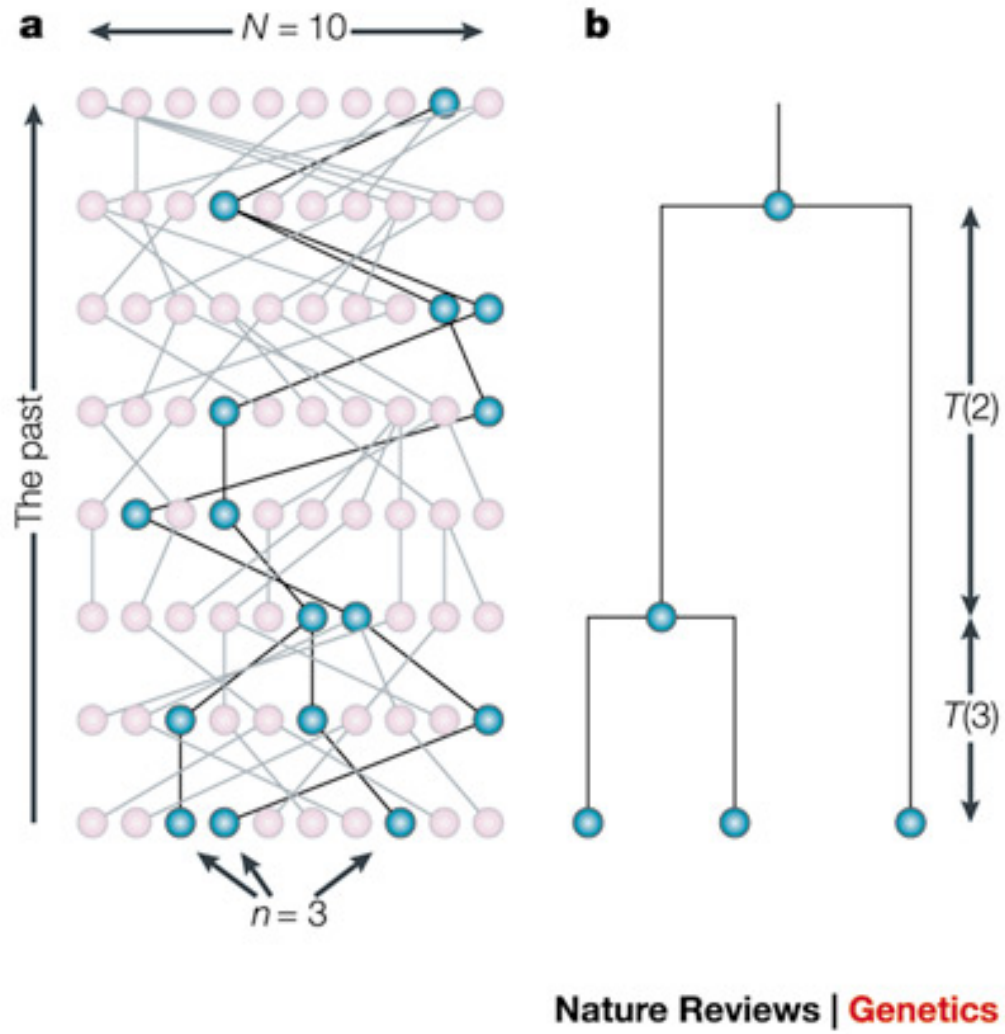
Motivating questions from human evolution



Nature Reviews | Genetics

- Are all non-Africans descended from a one population of migrants who left Africa together?
- Which human populations interbred with archaic hominids like Neanderthals? When and how often?
- How often do genetic adaptations cross species boundaries?
- Have population bottlenecks hurt our fitness and impeded adaptation?

The coalescent process (Kingman, 1982)



The Coalescent (sampling ancestors backward in time) is dual to Wright-Fisher evolution (having children forward in time)

Distribution of coalescence times

- The *coalescence time* at which two sequences find their common ancestor has distribution

$$\begin{aligned} T_2^{(\text{Same pop})}(t) &= \left(1 - \frac{1}{N}\right)^{t-1} \cdot \frac{1}{N} \approx \frac{1}{N} \exp\left(-\frac{t}{N}\right) \\ \mathbb{E}\left[T_2^{(\text{Same pop})}(t)\right] &= 2N \end{aligned}$$

- Sequences from different populations that diverged at time T must coalesce more anciently than T (if no migration):

$$\begin{aligned} T_2^{(\text{Diff pops})}(t) &\approx \frac{\mathbf{1}(t > T)}{N} \exp\left(-\frac{(t - T)}{N}\right) \\ \mathbb{E}\left[T_2^{(\text{Diff pops})}(t)\right] &= T + 2N \end{aligned}$$

Relating past population size to present genetic diversity

$$\begin{aligned} \theta &= 4N\mu \\ &= 2\mu \cdot [\text{Expected coalescence time of two sequences}] \\ D &= 2\mu T + 4N\mu \end{aligned}$$

- θ = Density of differences between two DNA sequences from the same population
- N = effective population size
- D = Density of differences between two DNA sequences from different populations
- T = Time these populations diverged

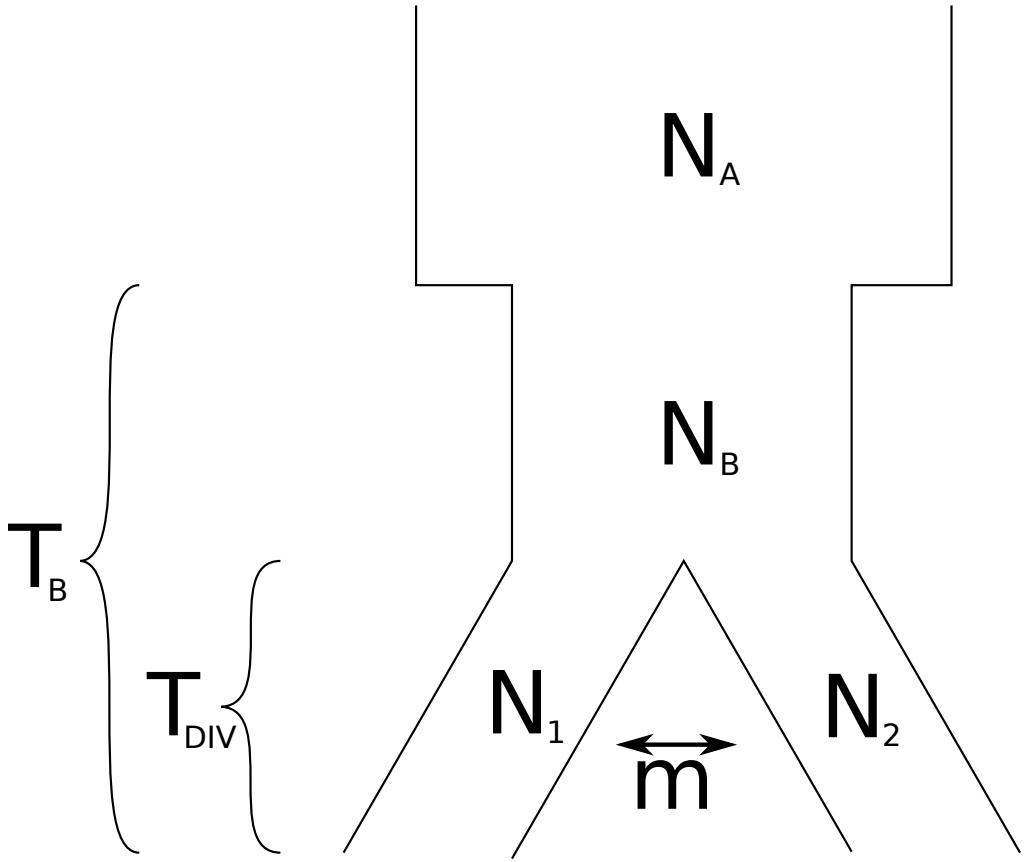
Assuming μ is a known constant, θ and D are sufficient statistics for estimating N and T

- In humans, $\theta \approx 0.001$ and $N \approx 10^4$
- Expected coalescence time of two sequences:

$$\begin{aligned} & \sim \quad 20{,}000 \text{ generations} \\ & \sim \quad 500{,}000 \text{ years} \\ & \sim \quad \text{Origin of anatomically modern humans} \end{aligned}$$

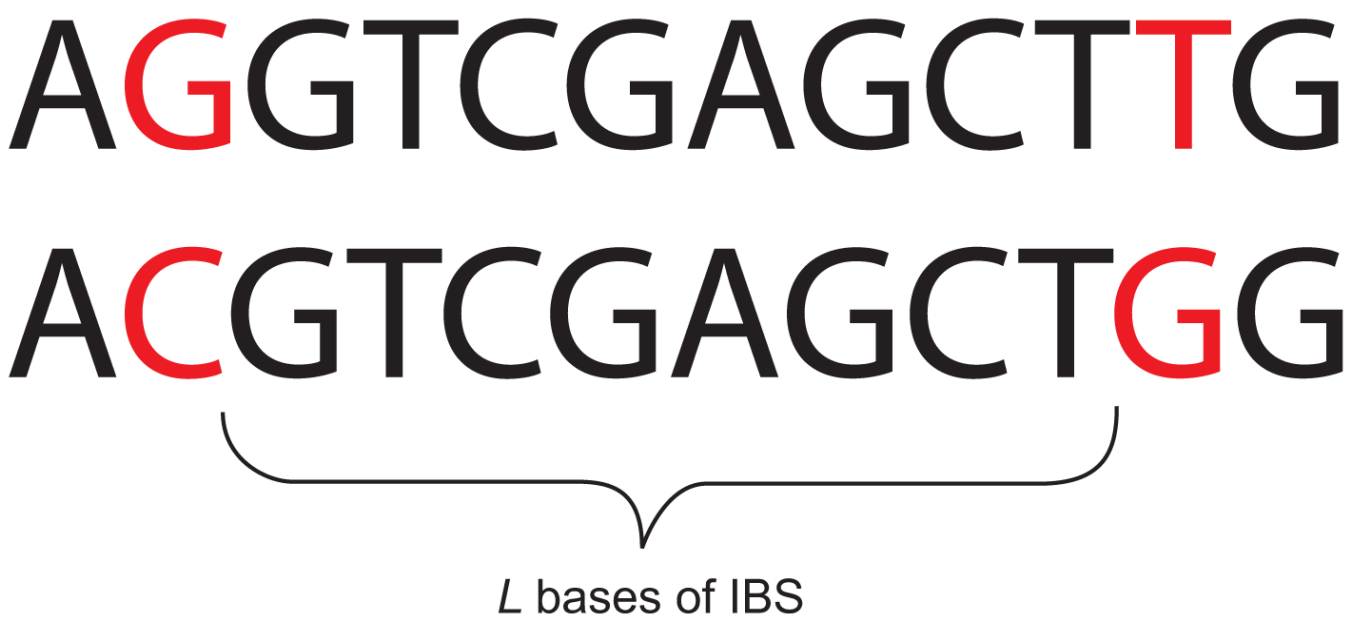
- A single diploid human genome is extremely informative about the entire ancestral human population (Li and Durbin *Nature* 2011)
- Major human populations diverged $< 100,000$ years ago, so most variation is shared between them

Adding more demographic complexity



If we allow population size changes and migration, a higher-dimensional set of summary statistics is needed to estimate all demographic parameters

Tracts of identity by state (IBS)



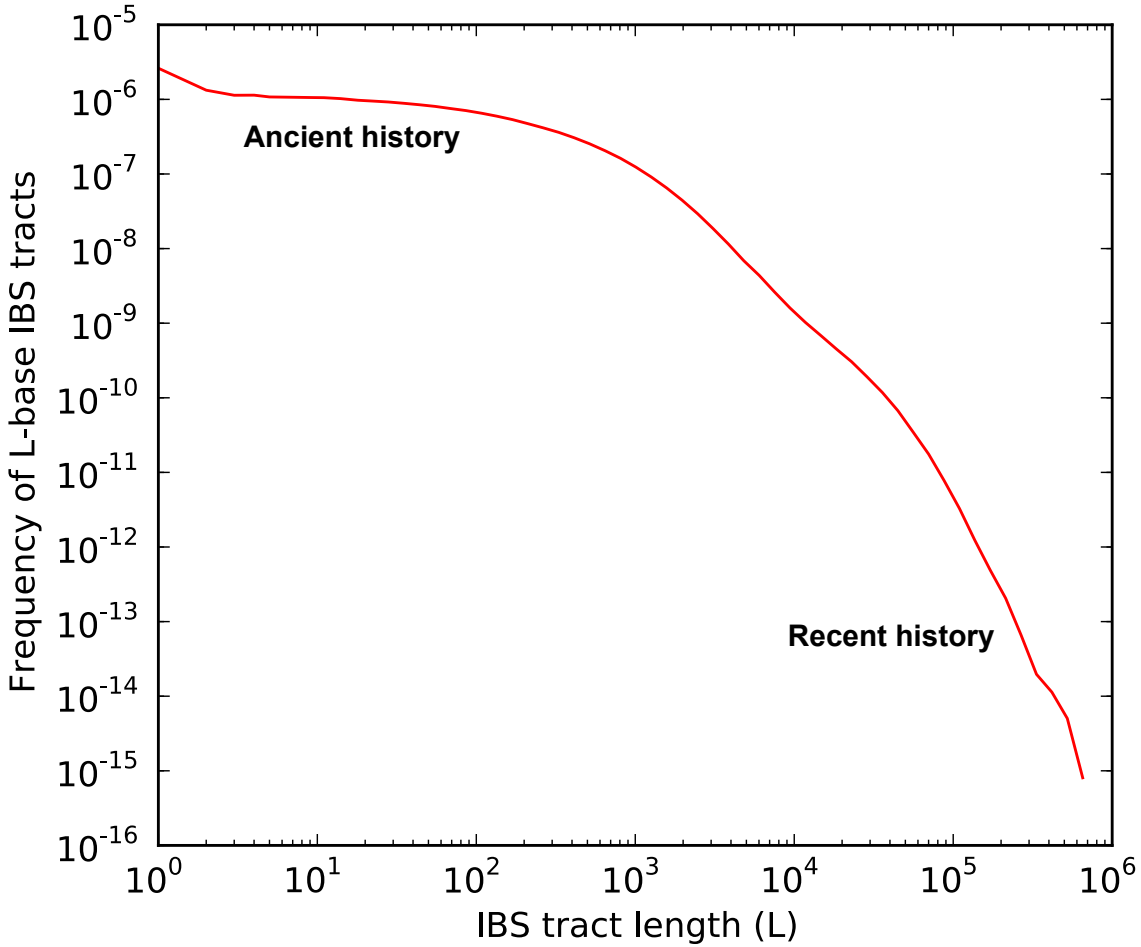
Complex demographic histories can be reconstructed from the length distribution of IBS tracts shared between DNA sequences

OPEN ACCESS Freely available online

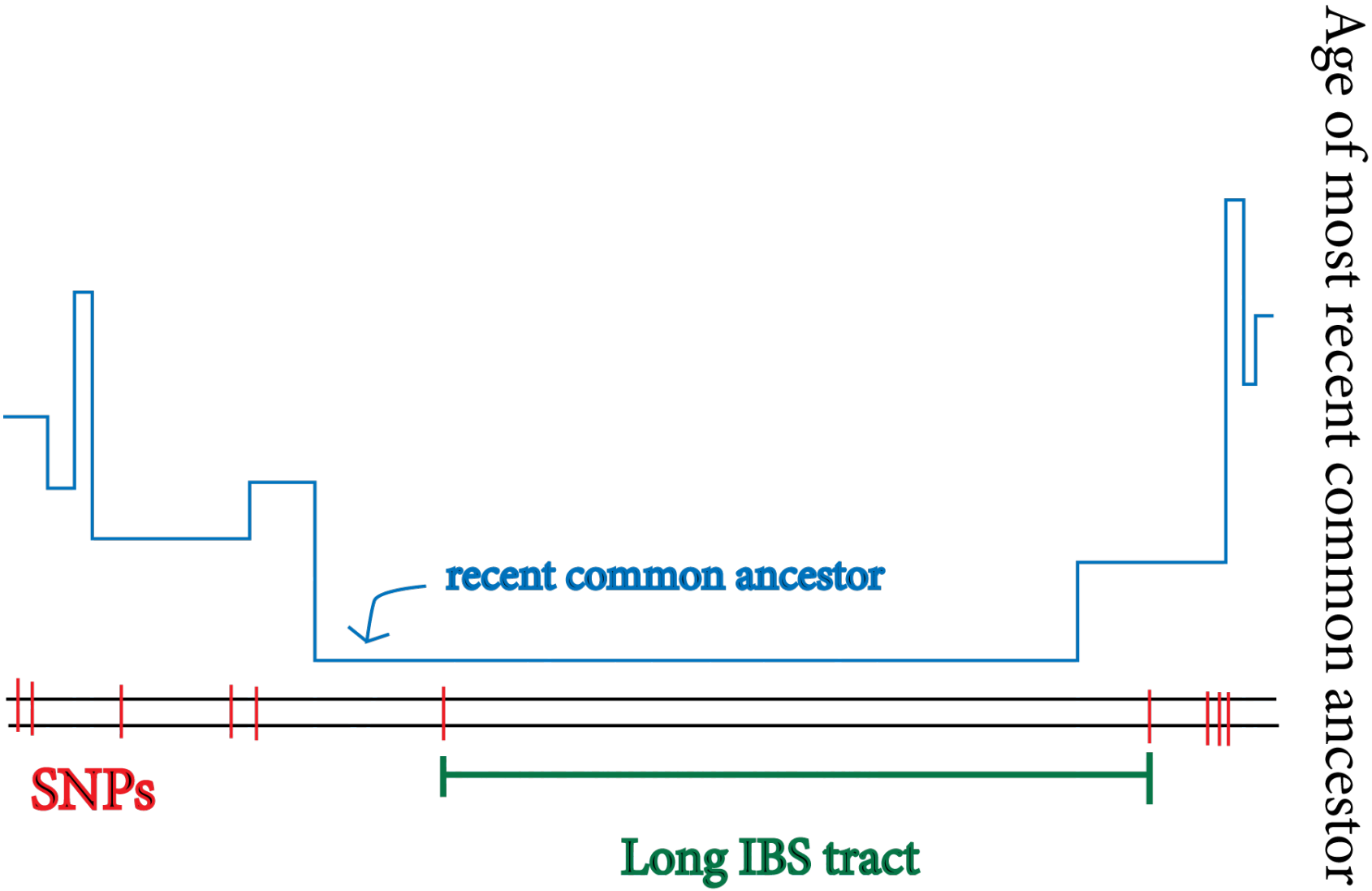
PLOS | GENETICS

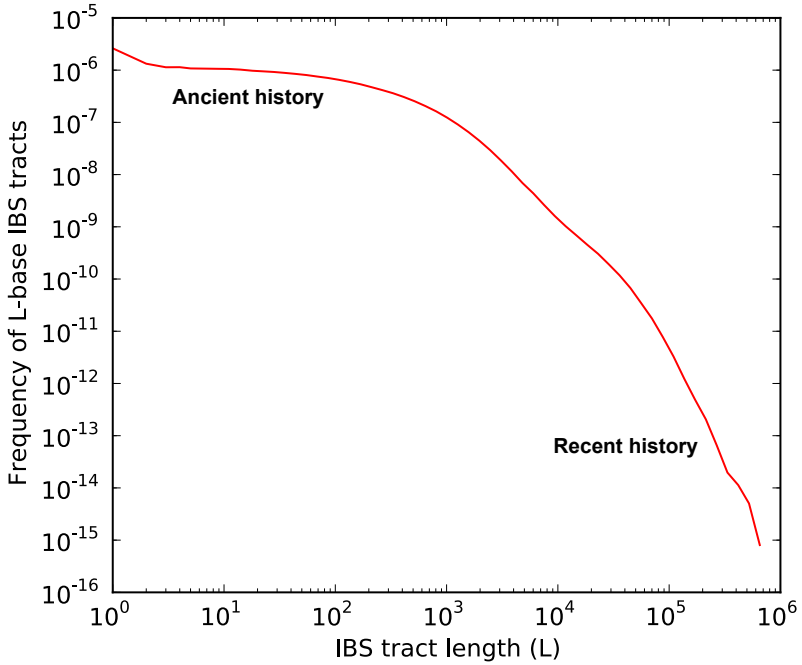
Inferring Demographic History from a Spectrum of Shared Haplotype Lengths

Kelley Harris[1]*, Rasmus Nielsen[2,3,4]



Relationship between recent history and long IBS tracts



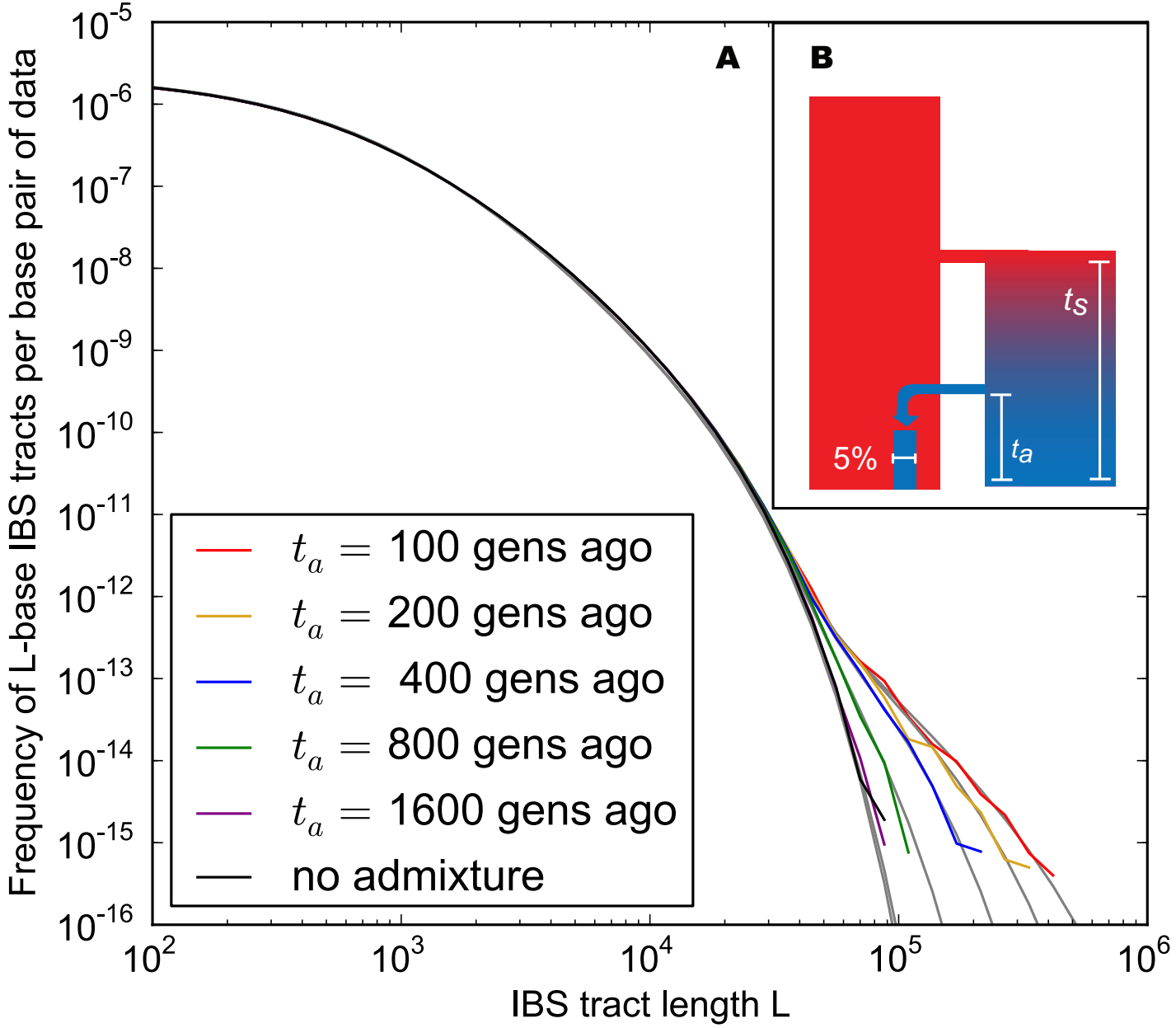


- $f_{\mathrm{IBS}}(L) :=$ observed frequency of L-base IBS tracts
- $H_{\Theta}(L) :=$ expected frequency of L-base IBS tracts under parametric model Θ
- Find parameters minimizing distance between $f_{\mathrm{IBS}}(L)$ and $H_{\Theta}(L)$

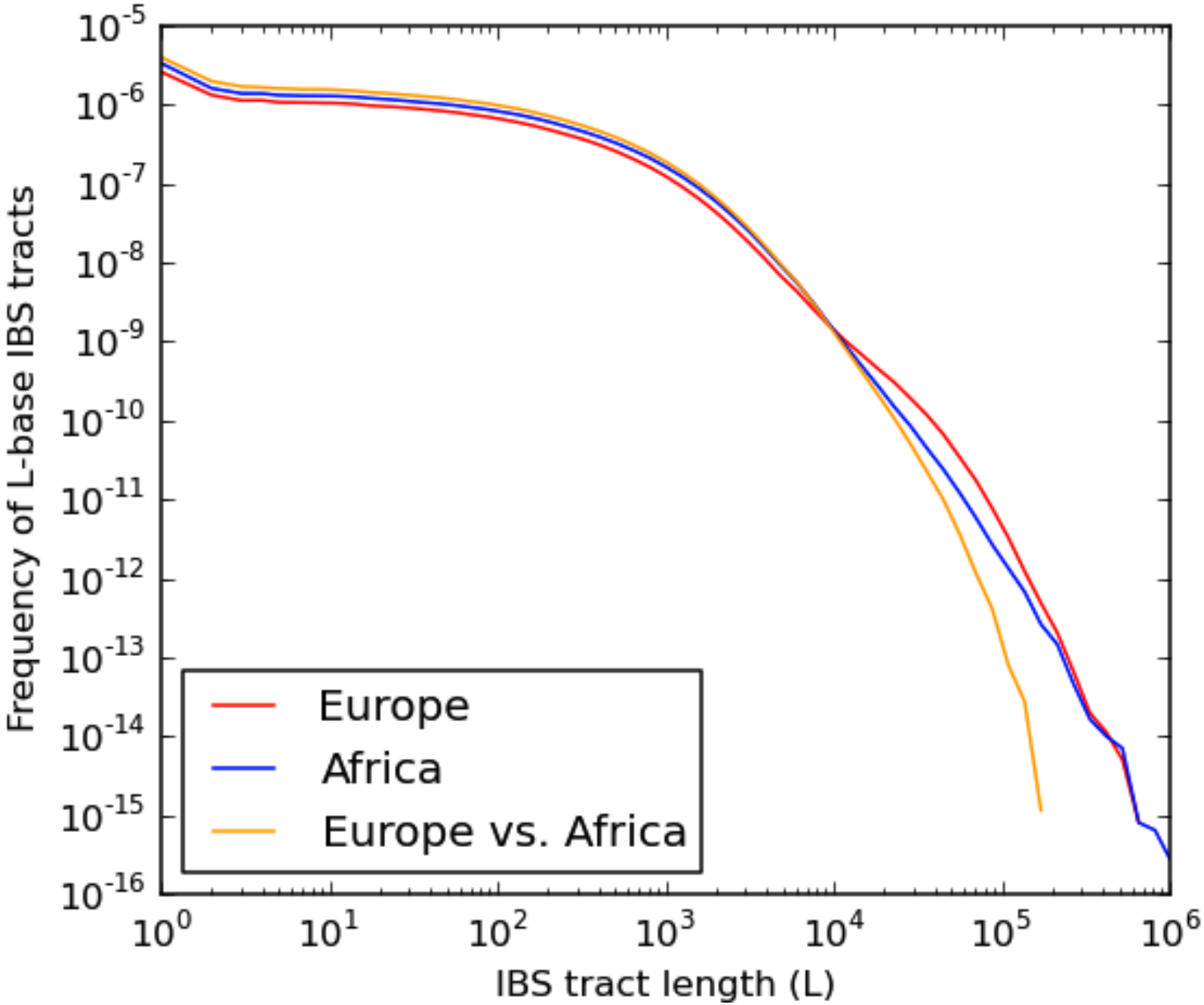
A simple case of the tract length formula

$$
\begin{aligned}
f_{\mathrm{IBS}}(L) \quad = \quad & \frac{1}{(1-\rho)(1+L(\rho+\theta))(2-\rho+L(\rho+\theta))}\left(\log\left(\frac{\rho+(L-2)(\rho+\theta)}{2\rho+\theta}\right)\log(1+(L-1)(\rho+\theta))\right. \\
& -\log(1+\theta)\log\left(\frac{1-\rho+(L-1)(\rho+\theta)}{1+2\rho+\theta}\right) \\
& \left.-\mathrm{Li}\left(\frac{\rho+(L-2)(\rho+\theta)}{1+(L-1)(\rho+\theta)}\right)+\mathrm{Li}\left(\frac{2\rho+\theta}{1+(L-1)(\rho+\theta)}\right)-\log(1+\theta)\log\left(\frac{\rho+(L-2)(\rho+\theta)}{2\rho+\theta}\right)\right) \\
& +\frac{1}{(1-\rho)(3-2\rho+L(\rho+\theta))(2-\rho+L(\rho+\theta))}\Bigg(\\
& -\log(2+\theta+(L-2)(\rho+\theta))\log\left(\frac{(1+(L-2)(\rho+\theta)}{1+\rho+\theta}\right) \\
& +\mathrm{Li}\left(\frac{1+(L-2)(\rho+\theta)}{2+\theta+(L-2)(\rho+\theta)}\right)-\mathrm{Li}\left(\frac{1+\rho+\theta}{2+\theta+(L-2)(\rho+\theta)}\right) \\
& -\log(1+\theta)\log\frac{1+\rho+2\theta}{\rho+\theta}+\log(1+\theta+(L-3)(\rho+\theta))\log\left(\frac{1+\theta+(L-2)(\rho+\theta)}{\rho+\theta}\right) \\
& -\mathrm{Li}\left(-\frac{1+\theta+(L-3)(\rho+\theta)}{1+\theta}\right)-\log(1+\theta+(L-3)(\rho+\theta))\log\left(\frac{2-2\rho+(L-1)(\rho+\theta)}{1+\theta}\right) \\
& +\mathrm{Li}(-1)+\log(1+t)\log(2)+\log(1+\theta)\log\left(\frac{2-2\rho+(L-1)(\rho+\theta)}{2+2\theta}\right) \\
& \left.-\mathrm{Li}\left(-\frac{1+\theta}{\rho+\theta}\right)+\mathrm{Li}\left(-\frac{1+\theta+(L-3)(\rho+\theta)}{\rho+\theta}\right)\right)+\cdots
\end{aligned}
$$

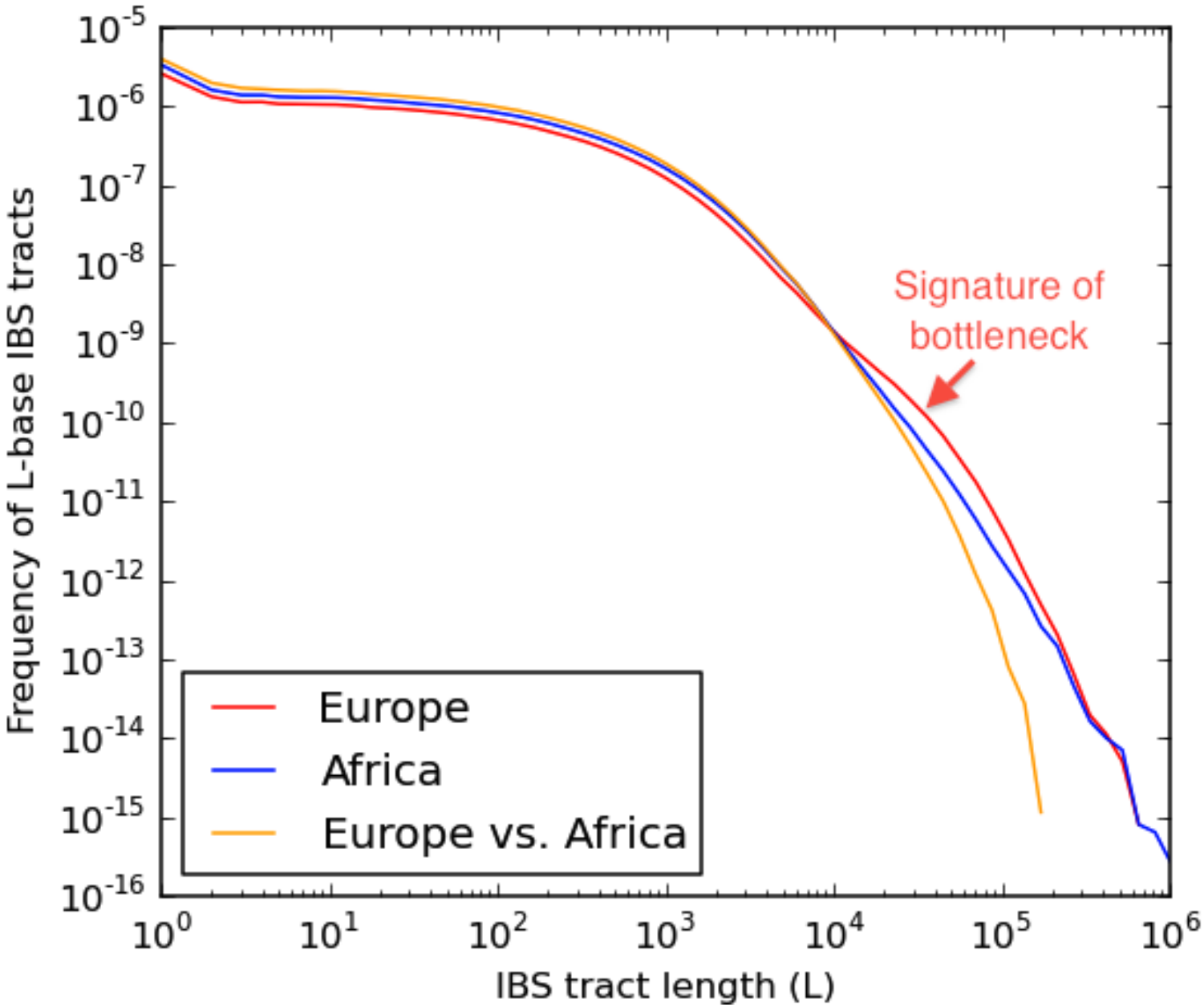
Power to date gene flow events



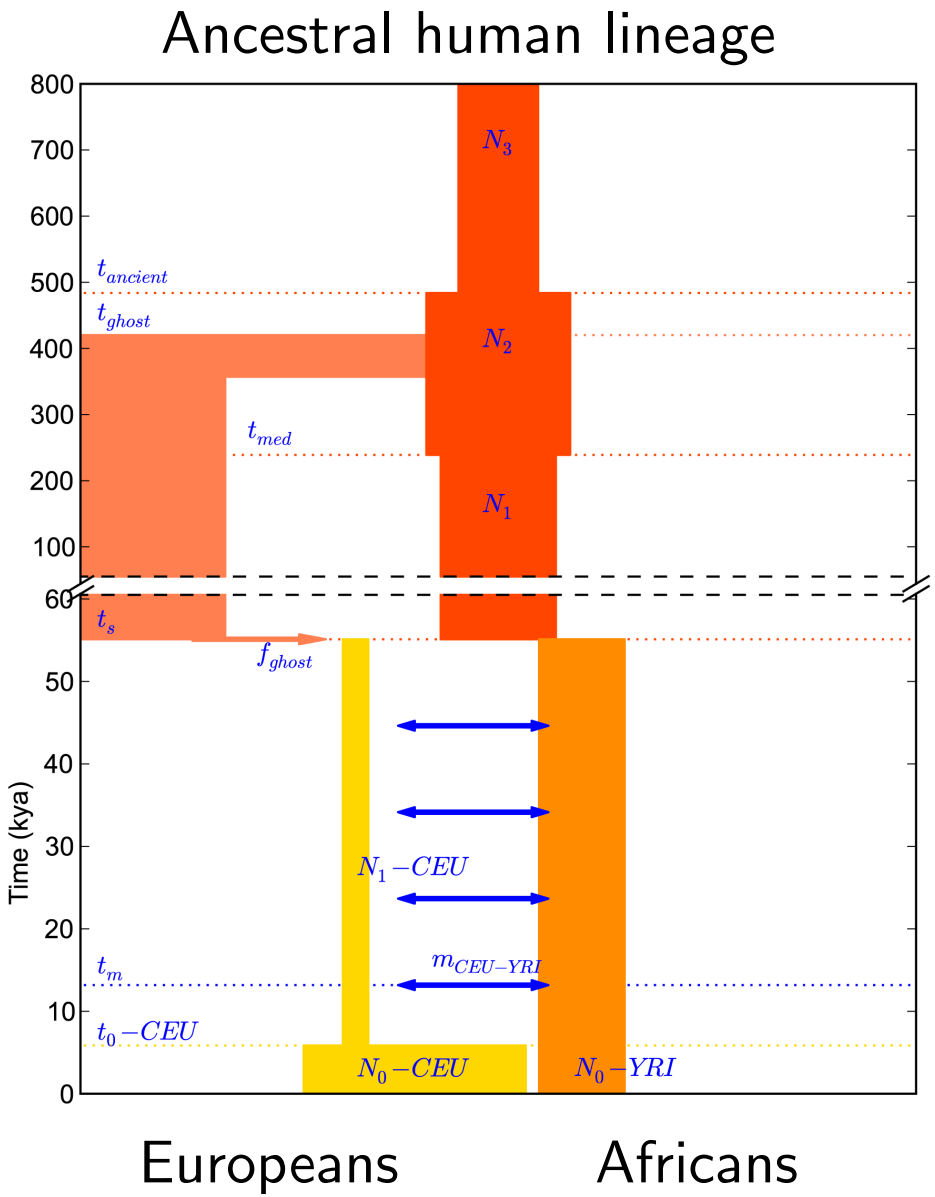
IBS tracts in human data



IBS tracts in human data



Inference of European/African divergence and migration



- Neanderthal-like admixture into Europeans
- Divergence 55,000 years ago*
- Out-of-Africa bottleneck
- Recent European-African migration

* Assuming 2.5×10^{-8} mutations per site per generation

Second application: Inference of rapid polar bear speciation

Eline Lorenzen

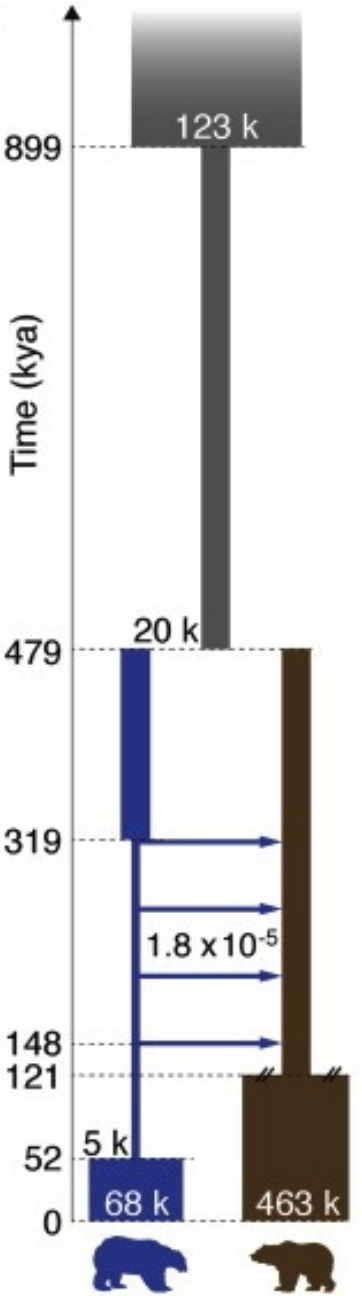
Matteo Fumagalli

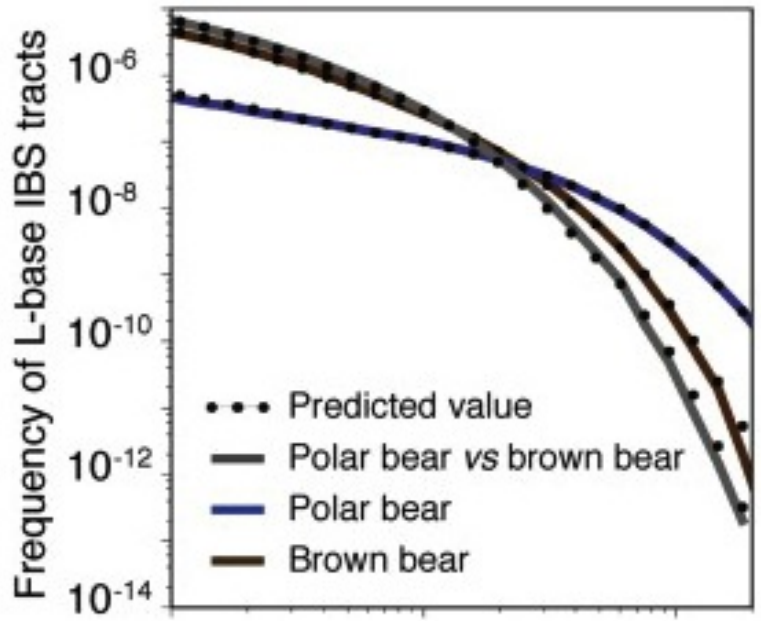
Cell

Population Genomics Reveal Recent Speciation and Rapid Evolutionary Adaptation in Polar Bears

Shiping Liu,[1,2,20] Eline D. Lorenzen,[3,4,20] Matteo Fumagalli,[3,20] Bo Li,[1,20] Kelley Harris,[5] Zijun Xiong,[1] Long Zhou,[1] Thorfinn Sand Korneliussen,[4] Mehmet Somel,[3,21] Courtney Babbitt,[6,7,22] Greg Wray,[6,7] Jianwen Li,[1] Weiming He,[1,2] Zhuo Wang,[1] Wenjing Fu,[1] Xueyan Xiang,[1,8] Claire C. Morgan,[9] Aoife Doherty,[10] Mary J. O'Connell,[9] James O. McInerney,[10] Erik W. Born,[11] Love Dalén,[12] Rune Dietz,[13] Ludovic Orlando,[4] Christian Sonne,[13] Guojie Zhang,[1,14] Rasmus Nielsen,[1,3,15,16,*] Eske Willerslev,[4,*] and Jun Wang[1,16,17,18,19,*]

- Positive select scans revealed changes along polar bear lineage related to fat metabolism and cardiovascular function
- Brown bears are omnivores, but polar bears subsist on marine mammal fat
- How quickly did these adaptations arise?





Liu, Lorenzen, Fumagalli, Li, Harris et al., *Cell* 2014

- Polar bears diverged from brown bears less than 500,000 years ago
- One-way barrier to gene flow: migration from polar into brown, but never the reverse

SIDE EFFECTS

How Brown and Polar Bears Split Up, but Continued Coupling

Greenpeace/European Pressphoto Agency

ANCIENT A new study extends the origin of polar bears back to 5 million years instead of 600,000.

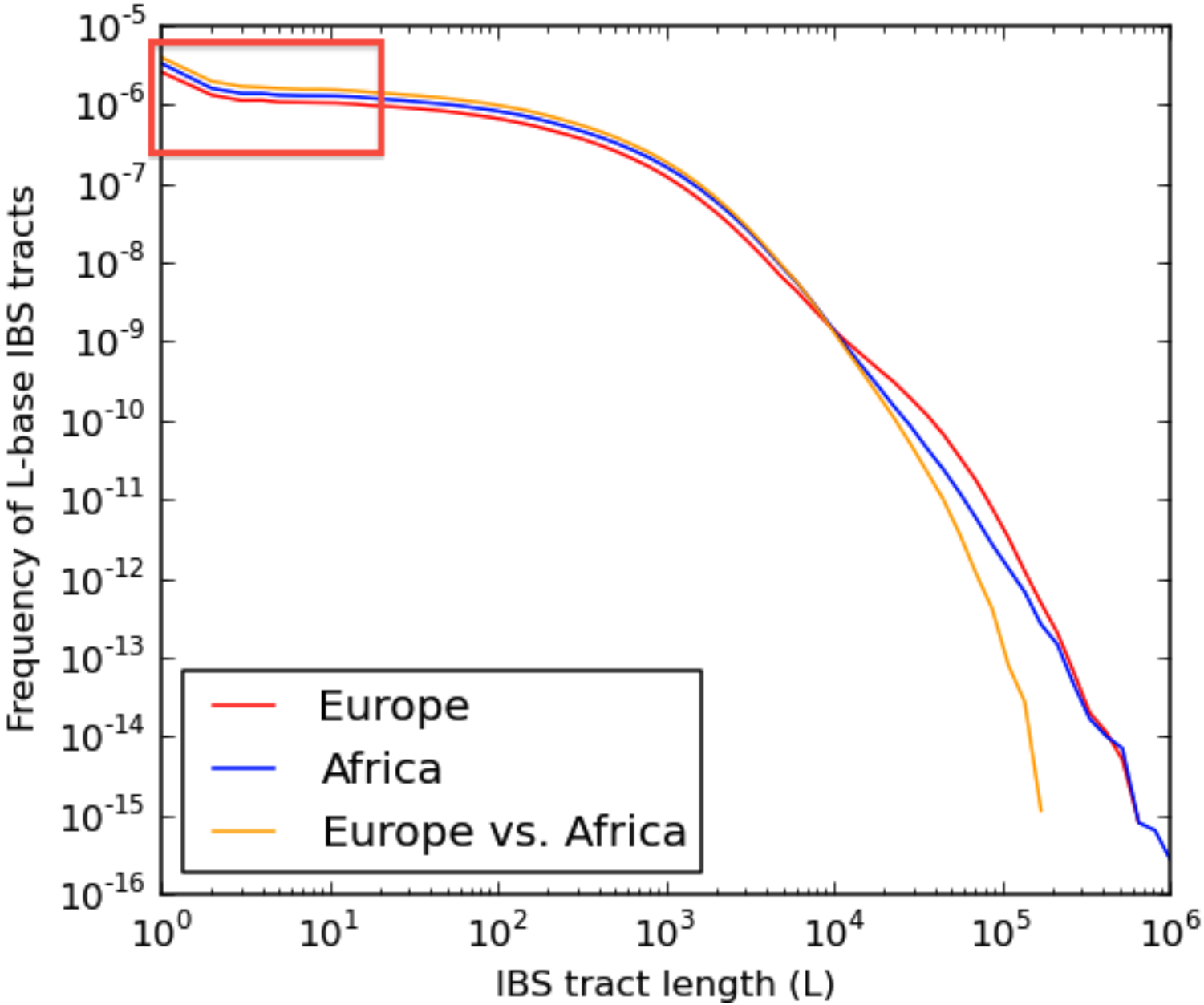
By JAMES GORMAN
Published: July 23, 2012 | 16 Comments

For many years, scientists who study the history of life on earth had to make do with fossils.

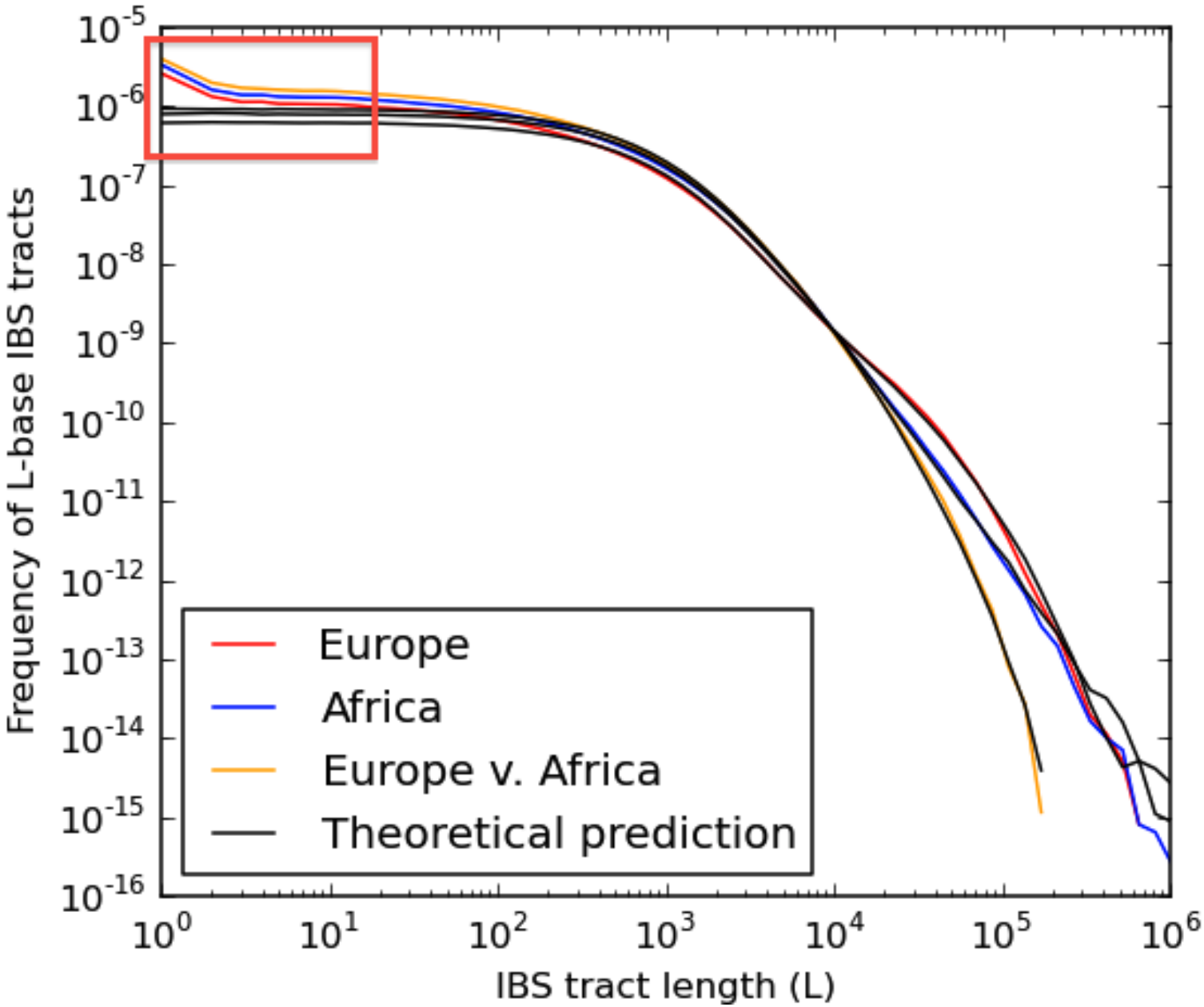
FACEBOOK

Strong polar bear bottleneck can confound some methods for inferring time of grizzly/polar bear divergence. Miller, *et al.* (*PNAS* 2012) estimated it occurred 4 million years ago!

From demography to mutation



From demography to mutation



- Source of excess short IBS tracts: multinucleotide mutations (MNMs)
- Complex mutations that create two or more SNPs at nearby sites in one generation

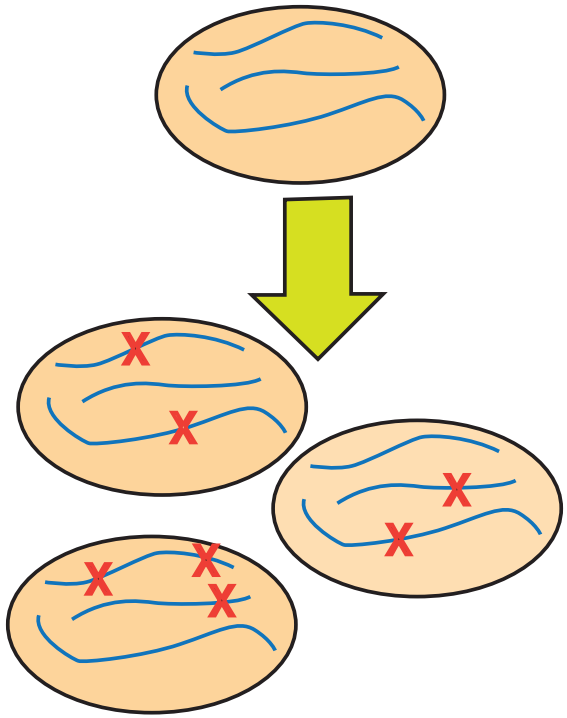
AAAGTTAGCCGACAC

↓

*AAAG**A**TA**A**CCGACAC*

Harris and Nielsen. *Genome Research* 2014.

Prior experimental evidence for multinucleotide mutation

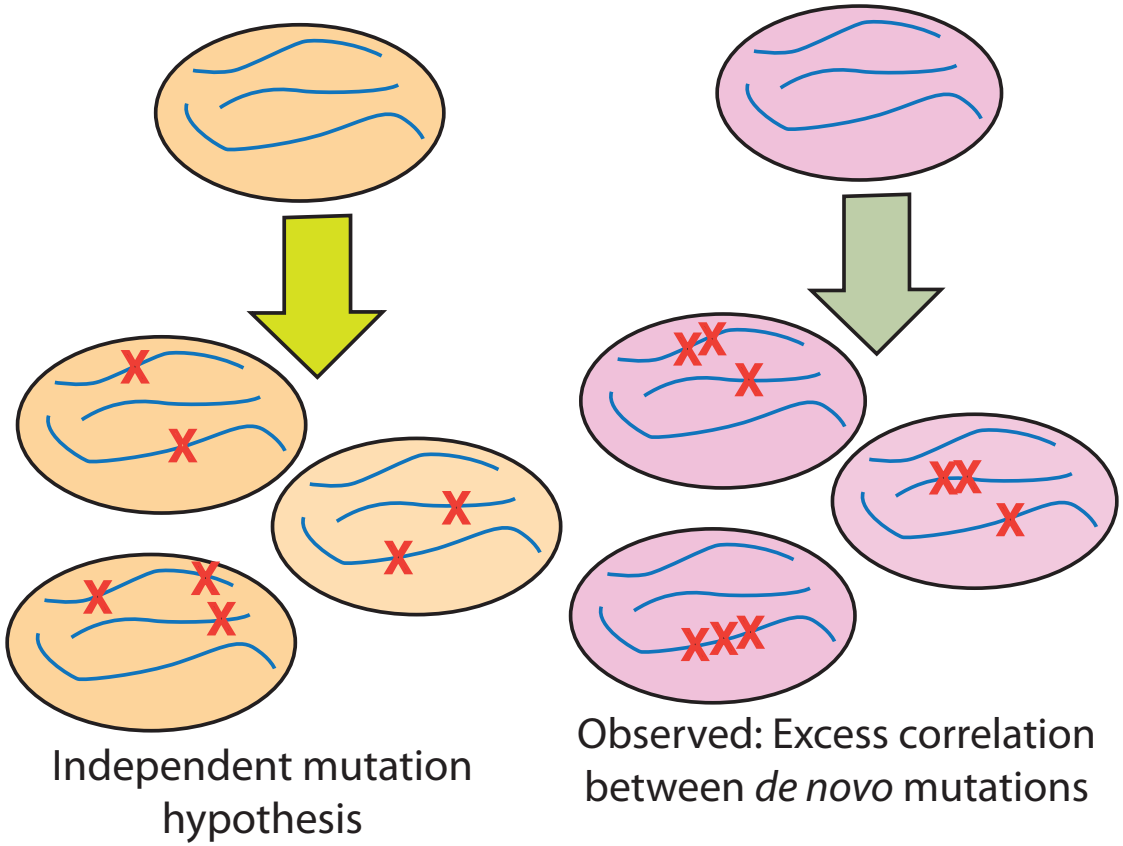


Independent mutation hypothesis

New mutations are directly observable in time series from yeast, *Drosophila*, etc

Schrider, *et al.*, *Current Biology* 2011; Schrider, *et al.*, *Genetics* 2013

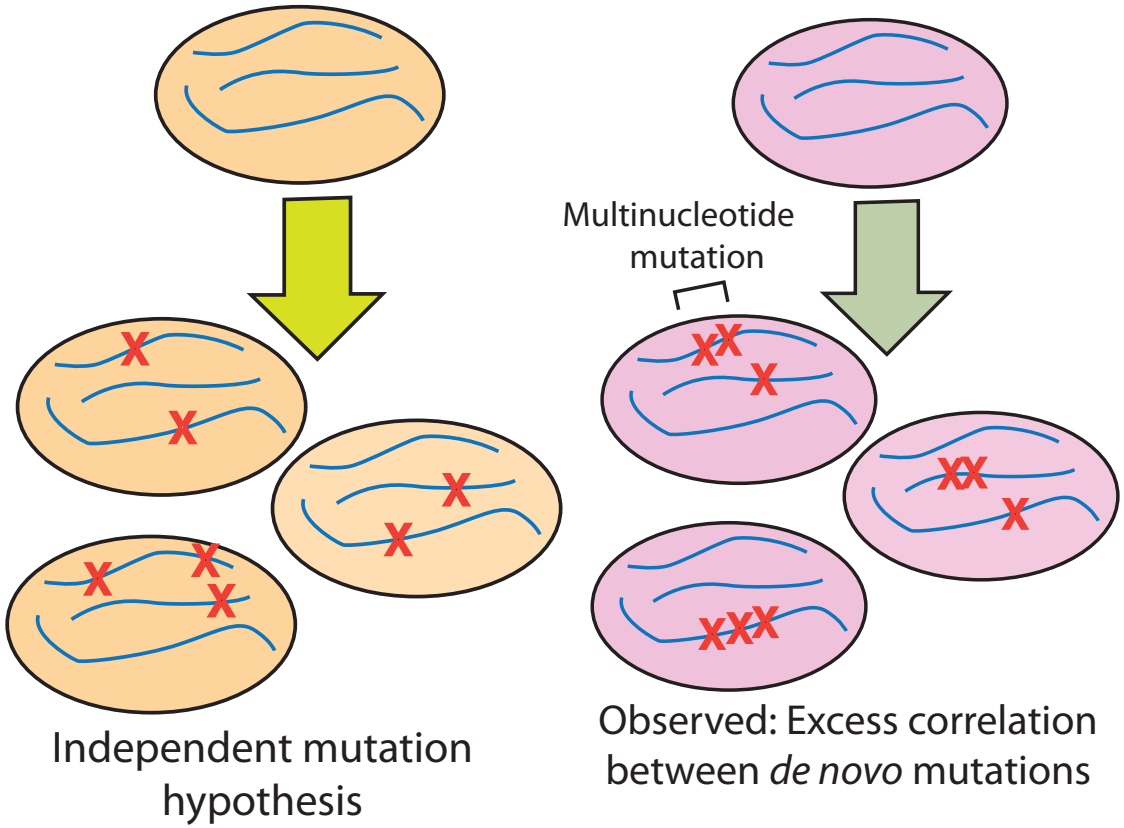
Prior experimental evidence for multinucleotide mutation



New mutations are directly observable in time series from yeast, *Drosophila*, etc

Schrider, *et al.*, *Current Biology* 2011; Schrider, *et al.*, *Genetics* 2013

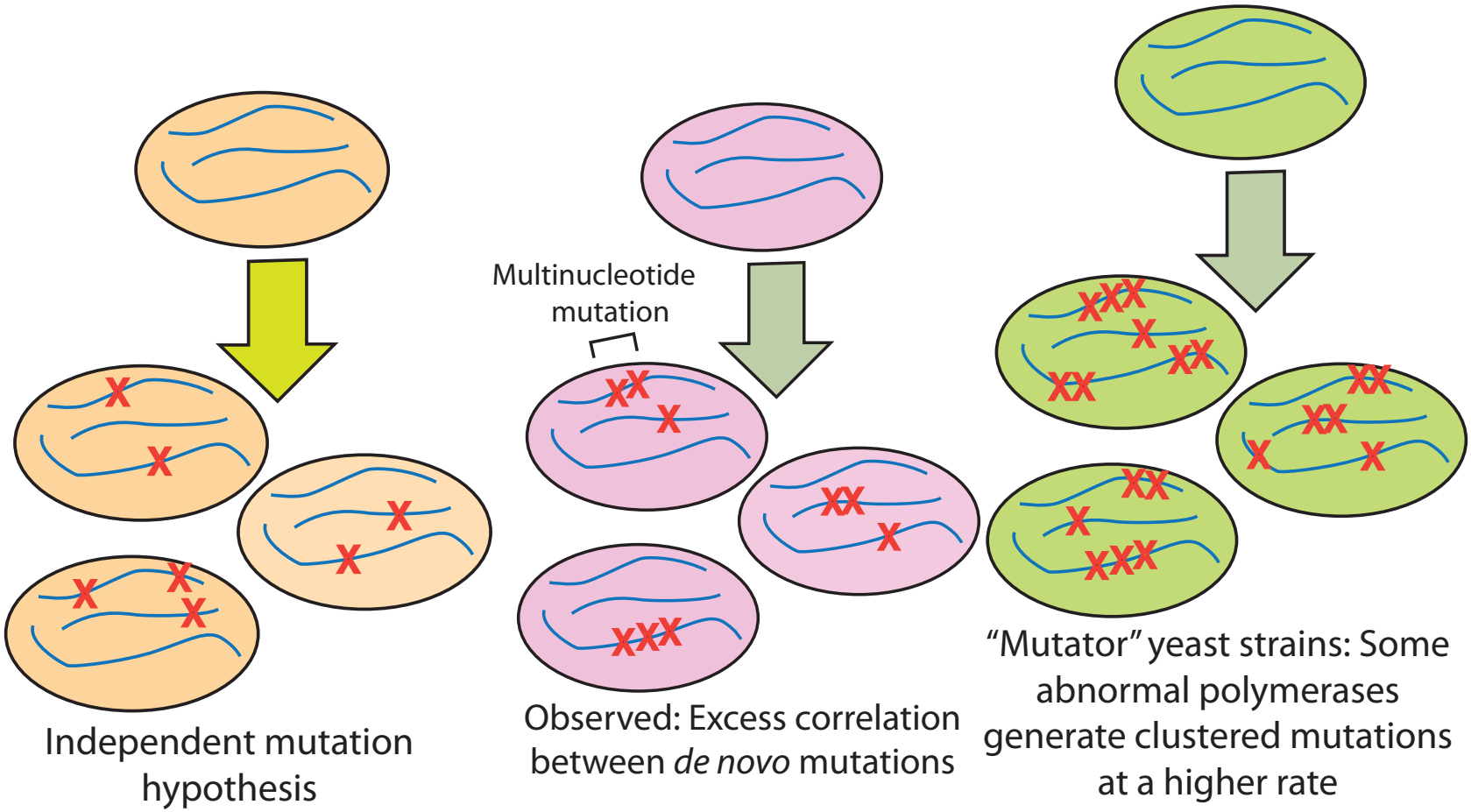
Prior experimental evidence for multinucleotide mutation



New mutations are directly observable in time series from yeast, *Drosophila*, etc

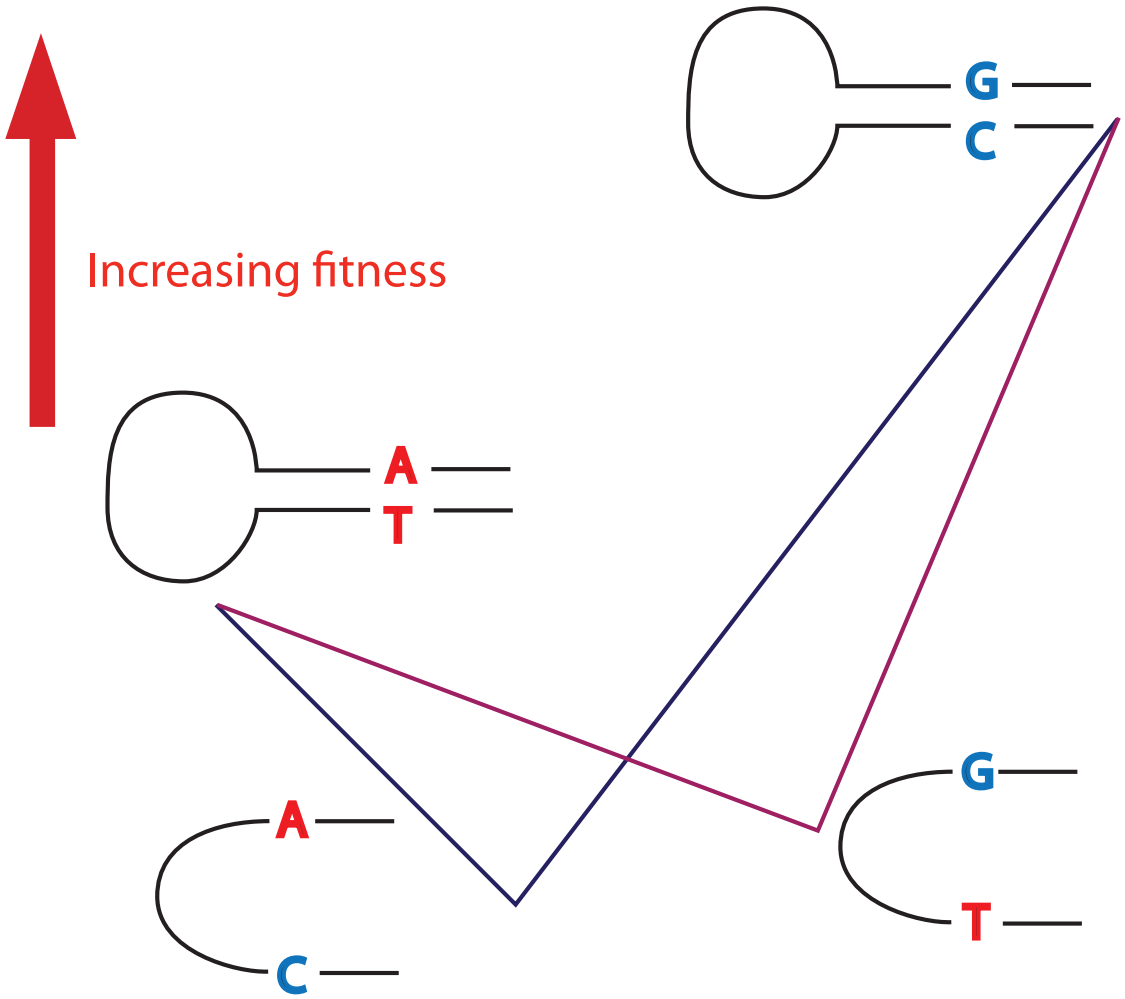
Schrider, *et al.*, *Current Biology* 2011; Schrider, *et al.*, *Genetics* 2013

Prior experimental evidence for multinucleotide mutation

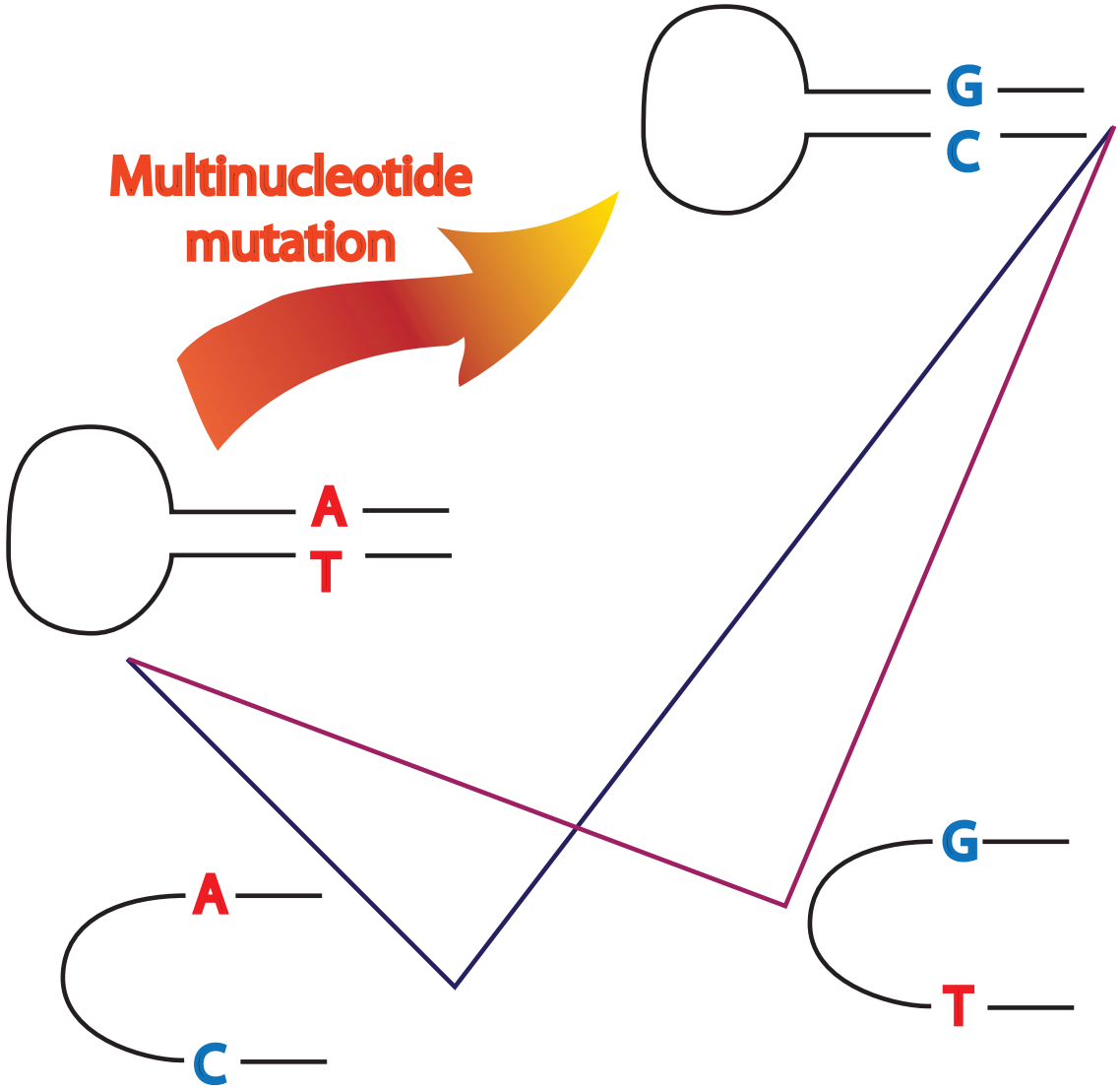


Sakamoto, *et al. DNA Repair* 2007; Stone, *et al. Environ and Mol Mut* 2012

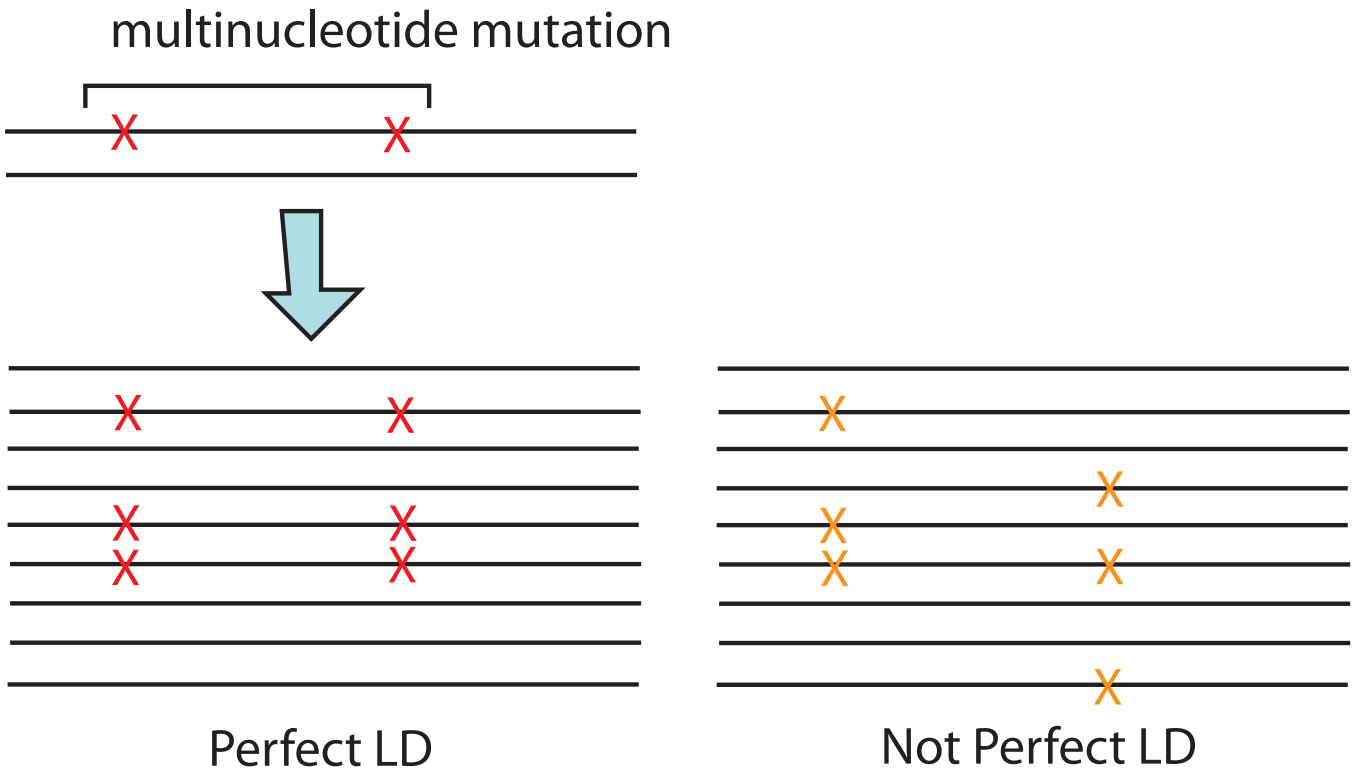
Widespread MNM could accelerate evolution across fitness valleys



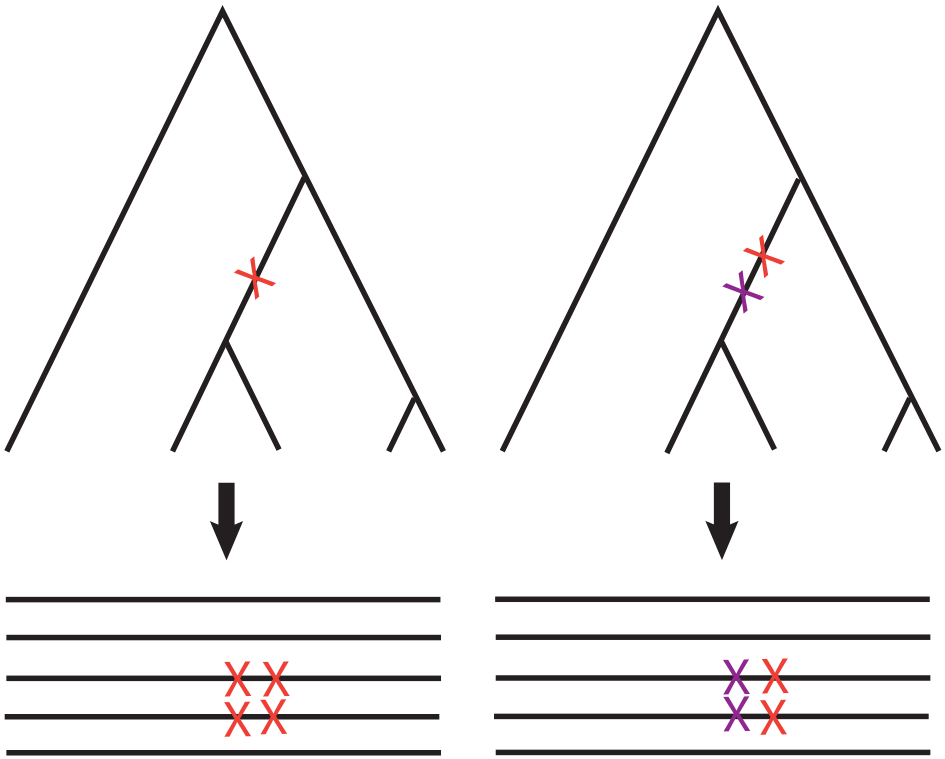
Widespread MNM could accelerate evolution across fitness valleys



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

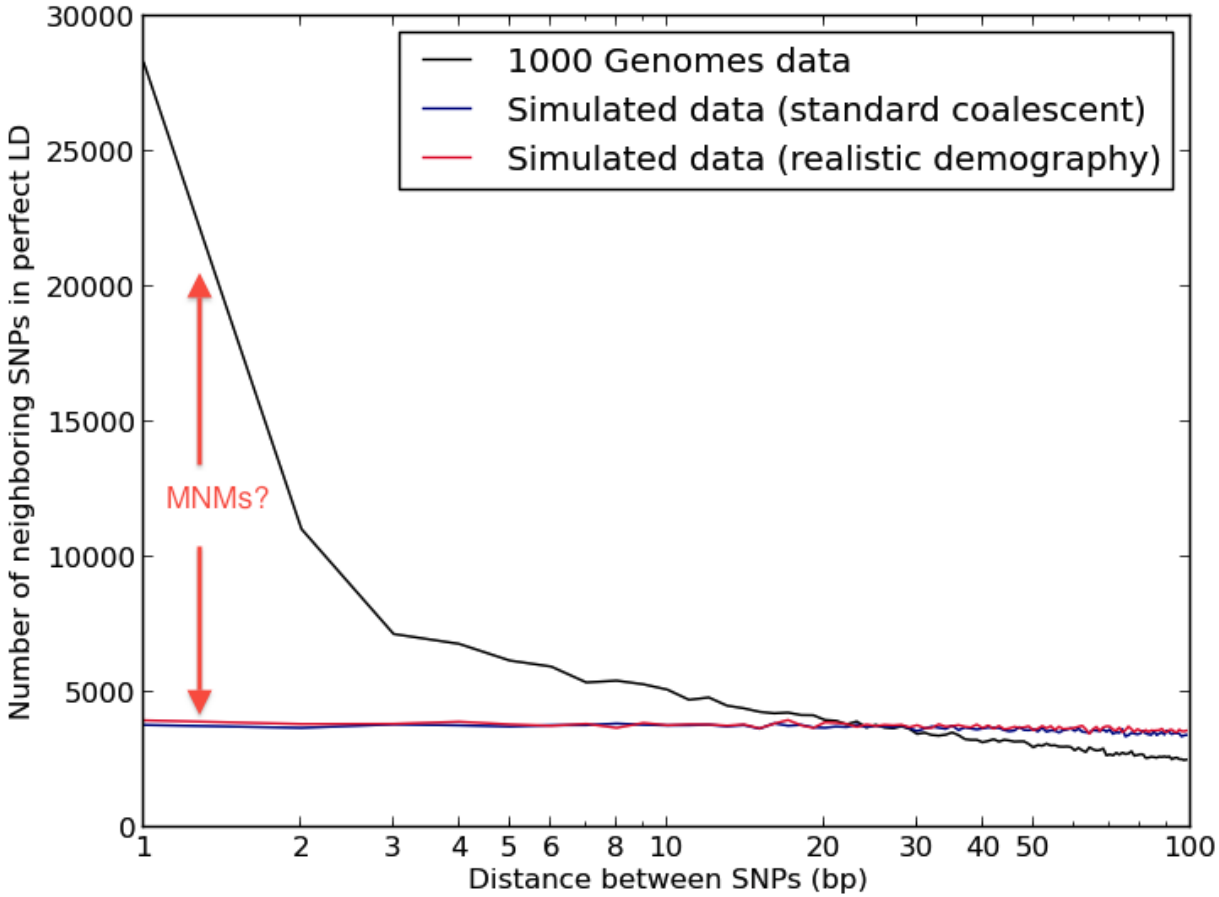


Independent mutations at neighboring sites can also create SNPs in perfect LD



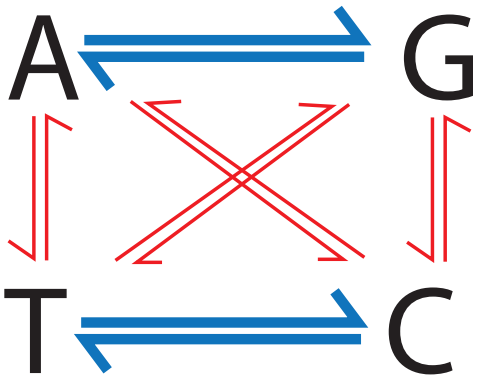
One MNM

Two independent mutations

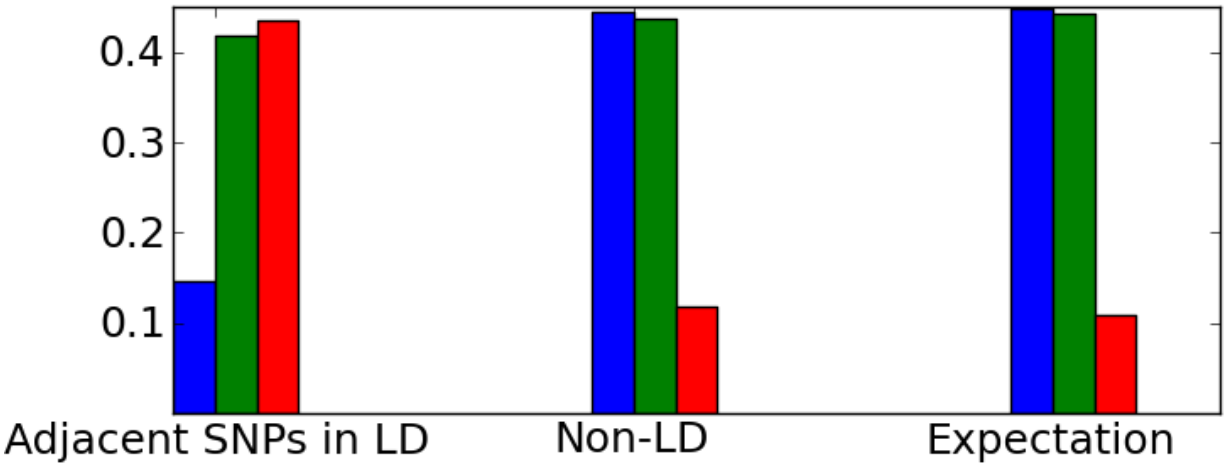


Compared to theoretical predictions, the 1000 Genomes Phase I data (1,092 humans from Africa, Europe, Asia, and the Americas) has excess close-together SNPs in perfect LD

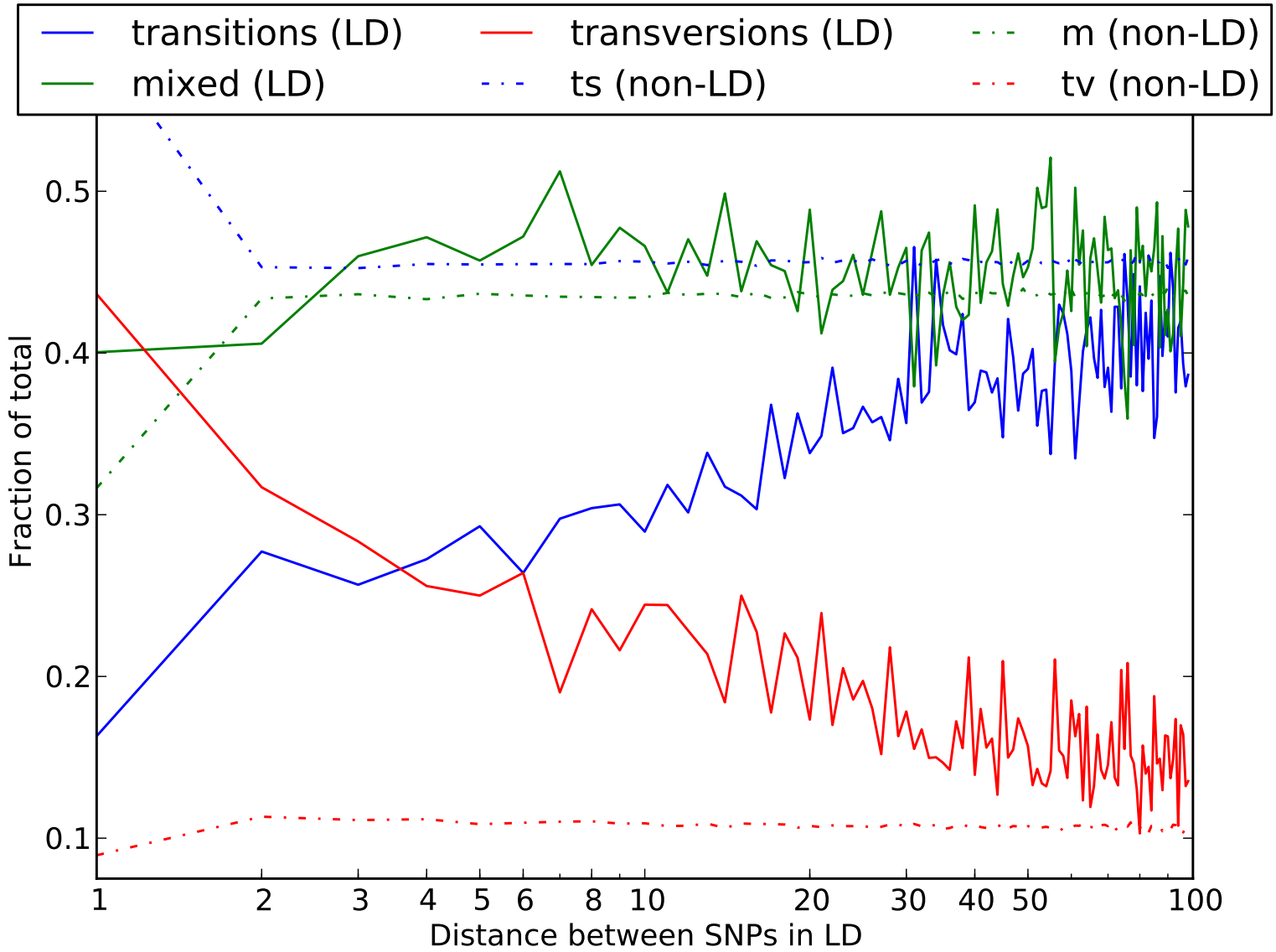
Perfect LD SNPs have excess transversions



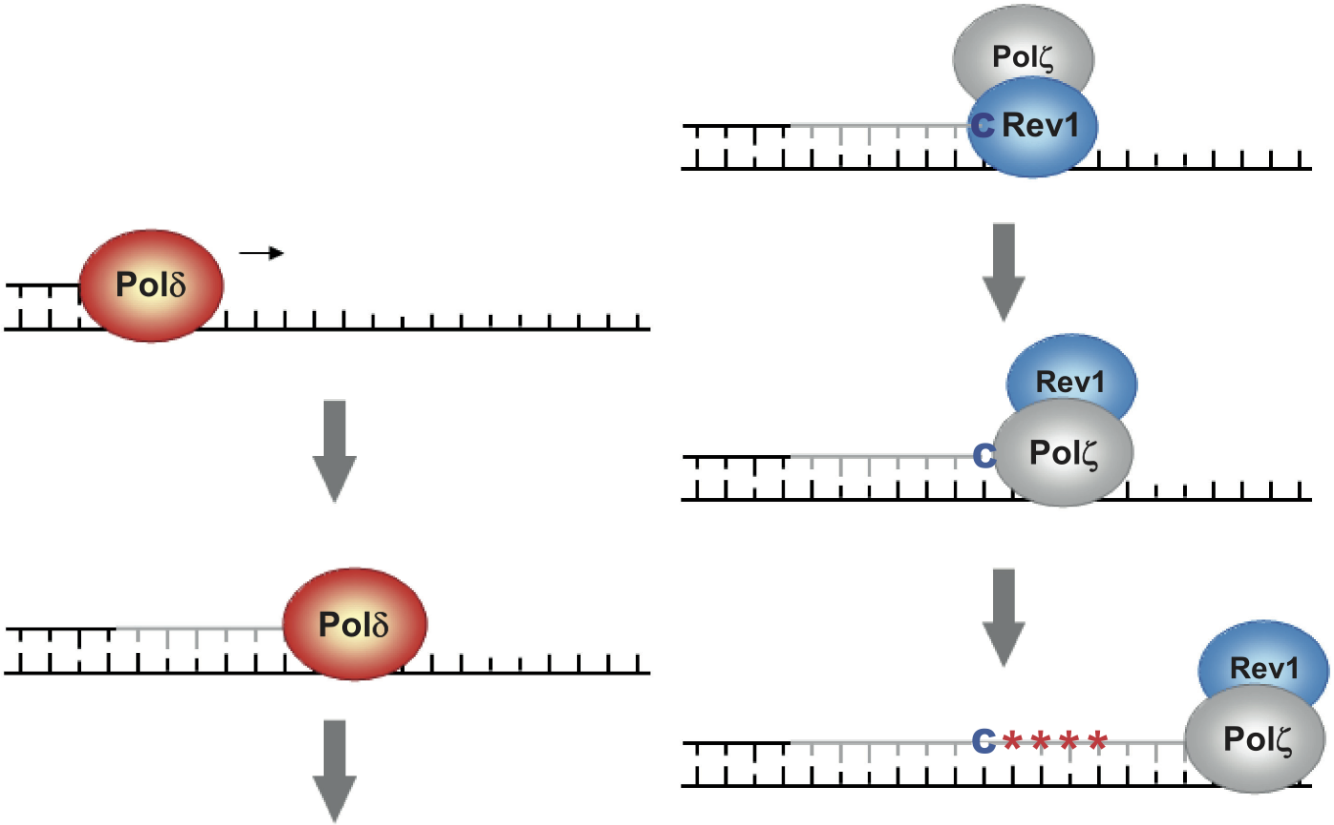
- 66% of human mutations are transitions ($A \rightleftharpoons G$, $C \rightleftharpoons T$)
- A SNP pair can consist of two transitions, two transversions, or one transition + one transversion (mixed)



Perfect LD SNPs have excess transversions



Error-prone translesion synthesis: a mechanism for MNM?



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

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

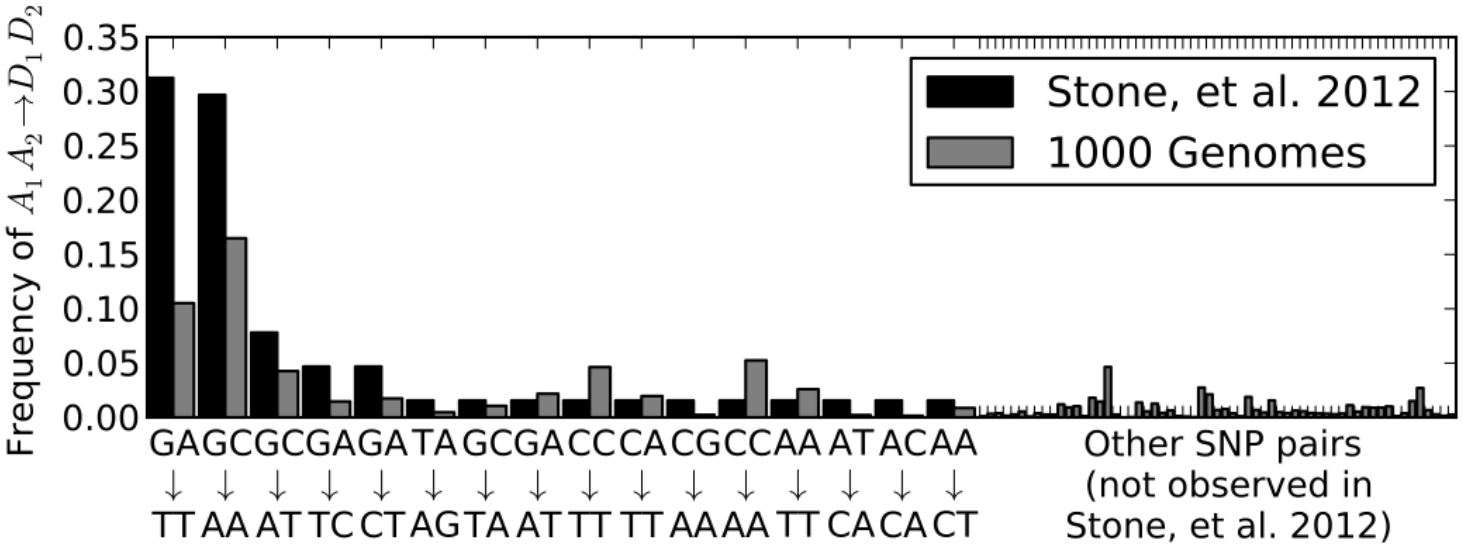
Research Article

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

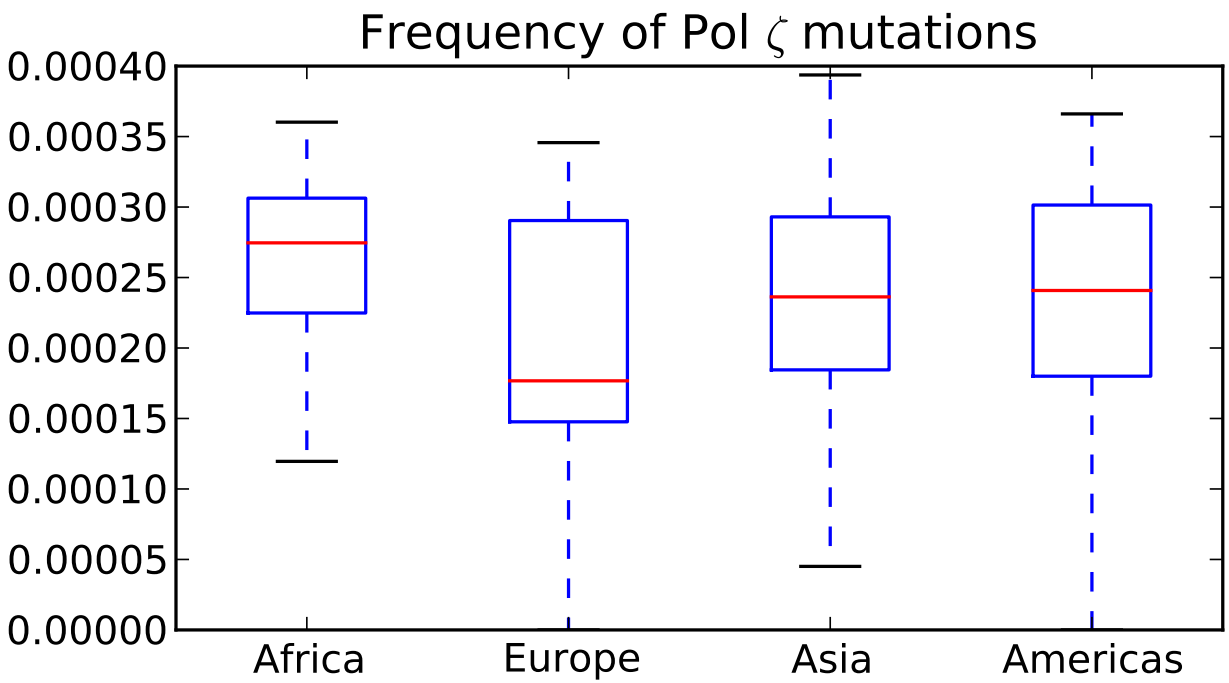
Jana E. Stone, Scott A. Lujan, and Thomas A. Kunkel*

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 rate of simultaneous mutation at nearby sites
- Increased translesion synthesis by Pol ζ

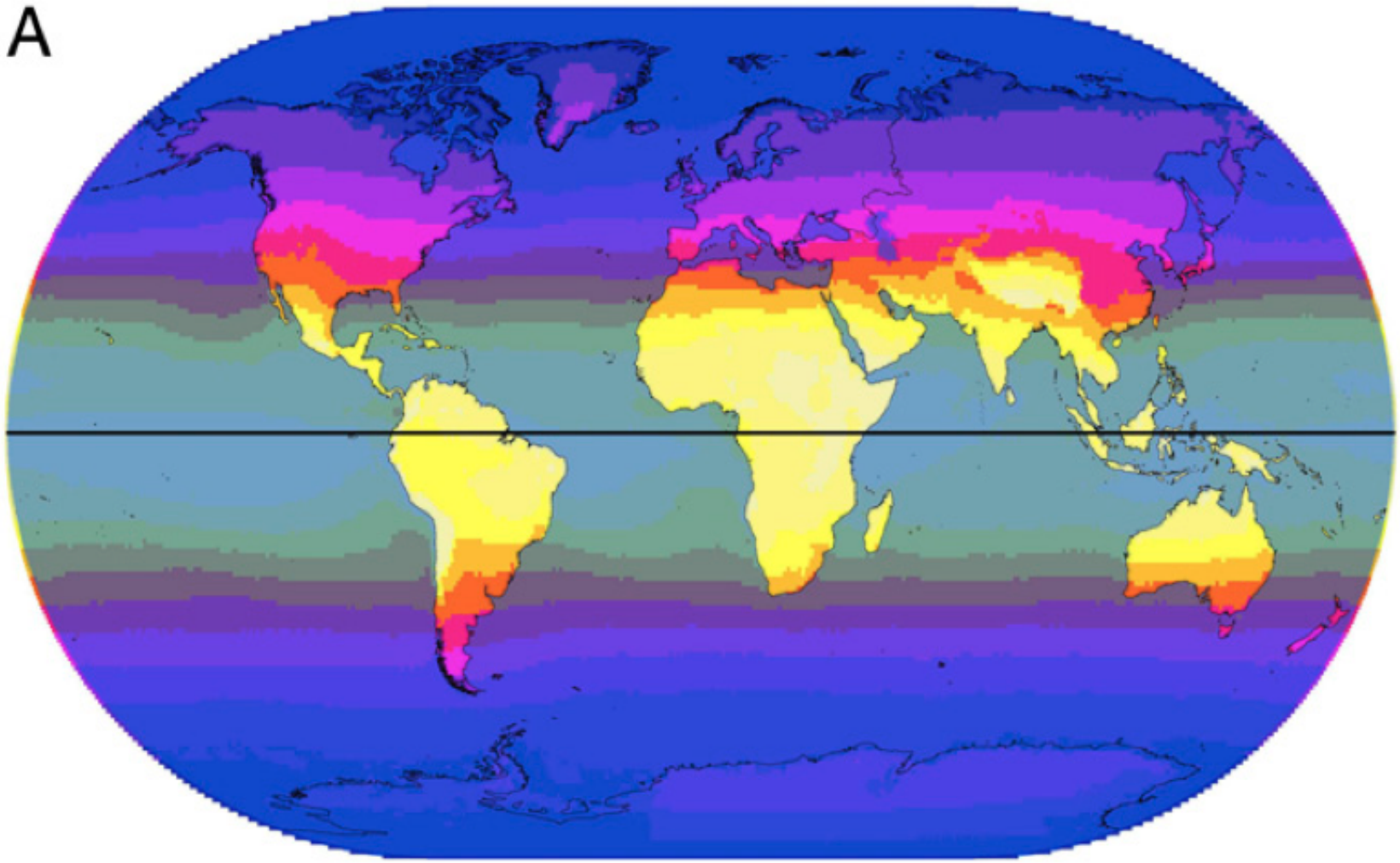


- Over 60% of the Stone, *et al.* tandem mutations were $\mathrm{GC} \to \mathrm{AA}$ or $\mathrm{GA} \to \mathrm{TT}$ (2 out of 78 possible $A_1A_2 \to D_1D_2$ combinations)
- $\mathrm{GC} \to \mathrm{AA}$ and $\mathrm{GA} \to \mathrm{TT}$ are by far the most common linked adjacent SNPs in the 1000 Genomes data
- A signature of Pol ζ activity in human population history



Pol ζ activity appears fairly uniform across populations

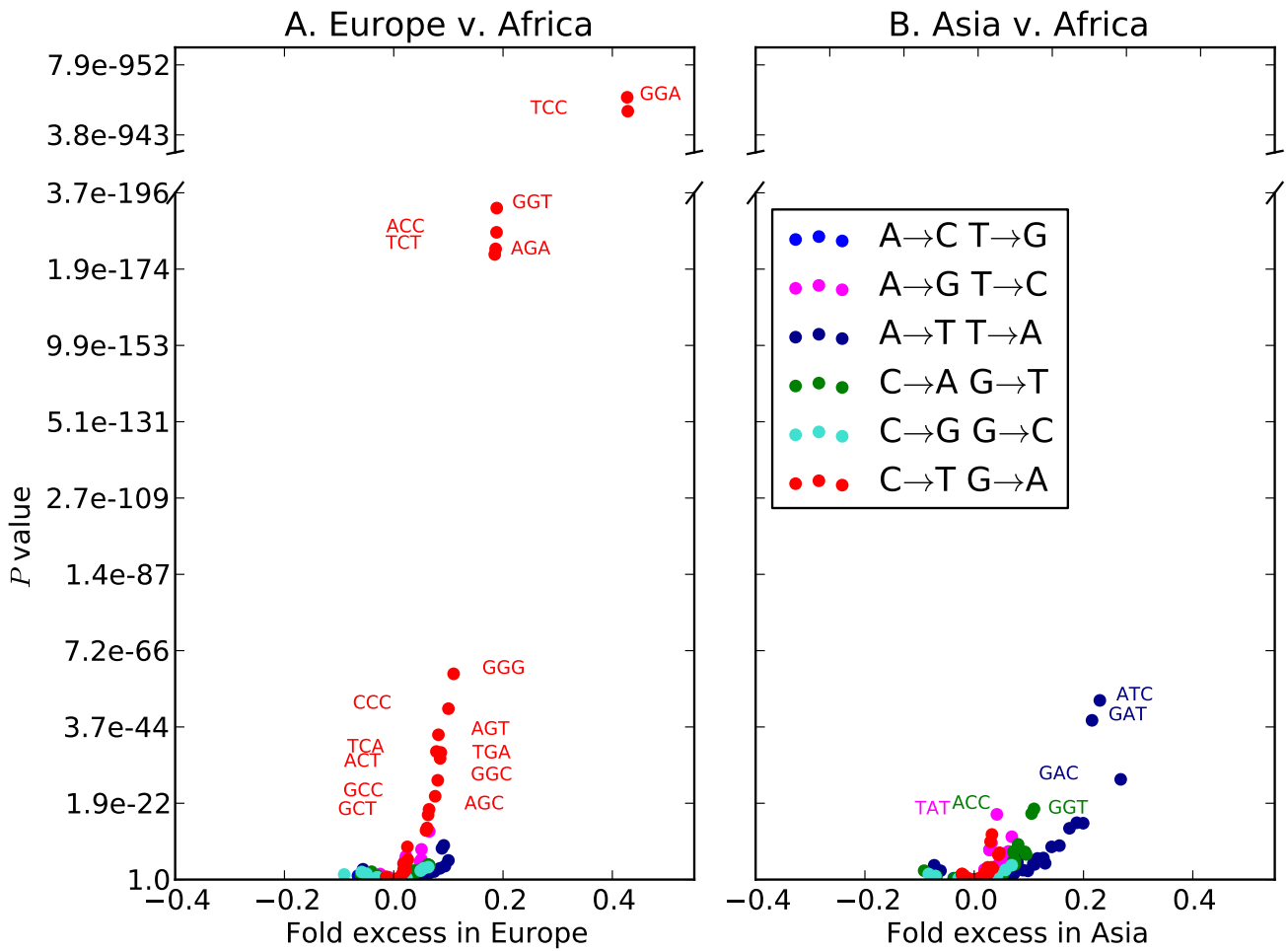
Are other mutagenic processes more variable? Under selection?



Jablonski and Chaplin *PNAS* 2010

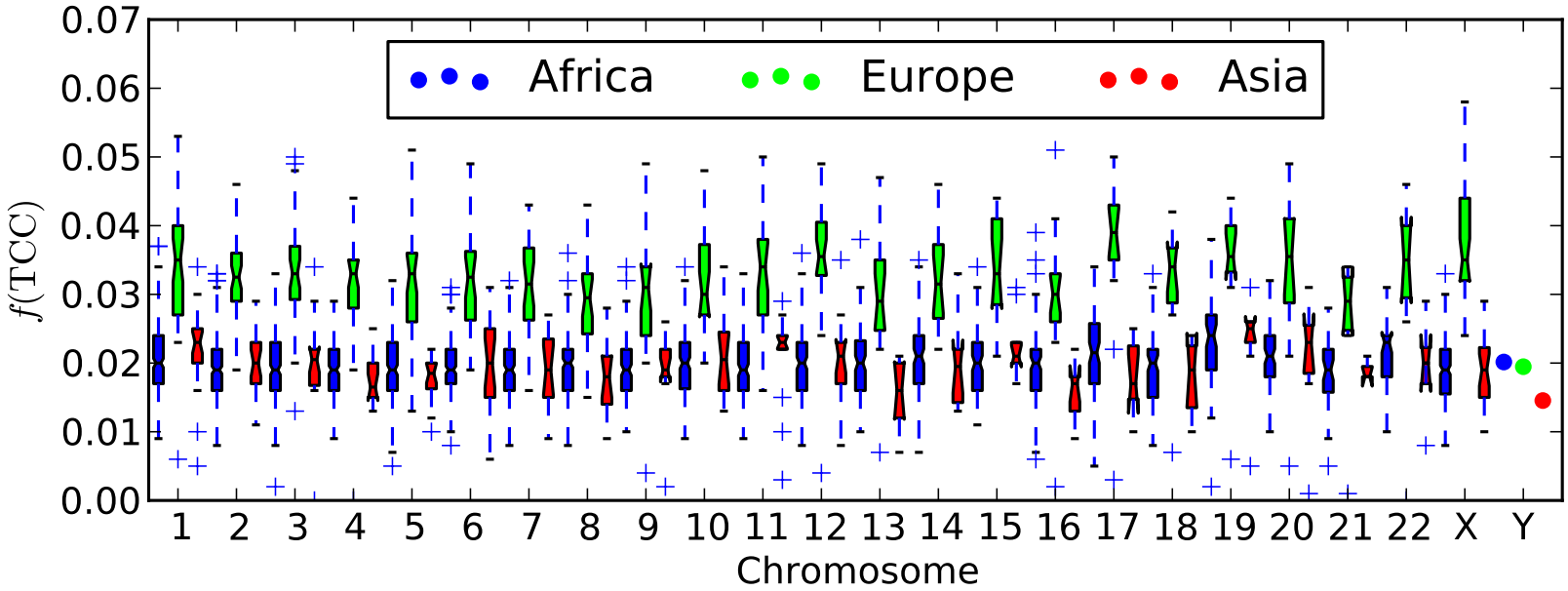
- Mutagen exposure is variable (e.g. UV radiation)
- Fraser (*Genome Res* 2013) found a strong signal of local adaptation of UV damage response regulation

Mutation spectra of continent-private variation

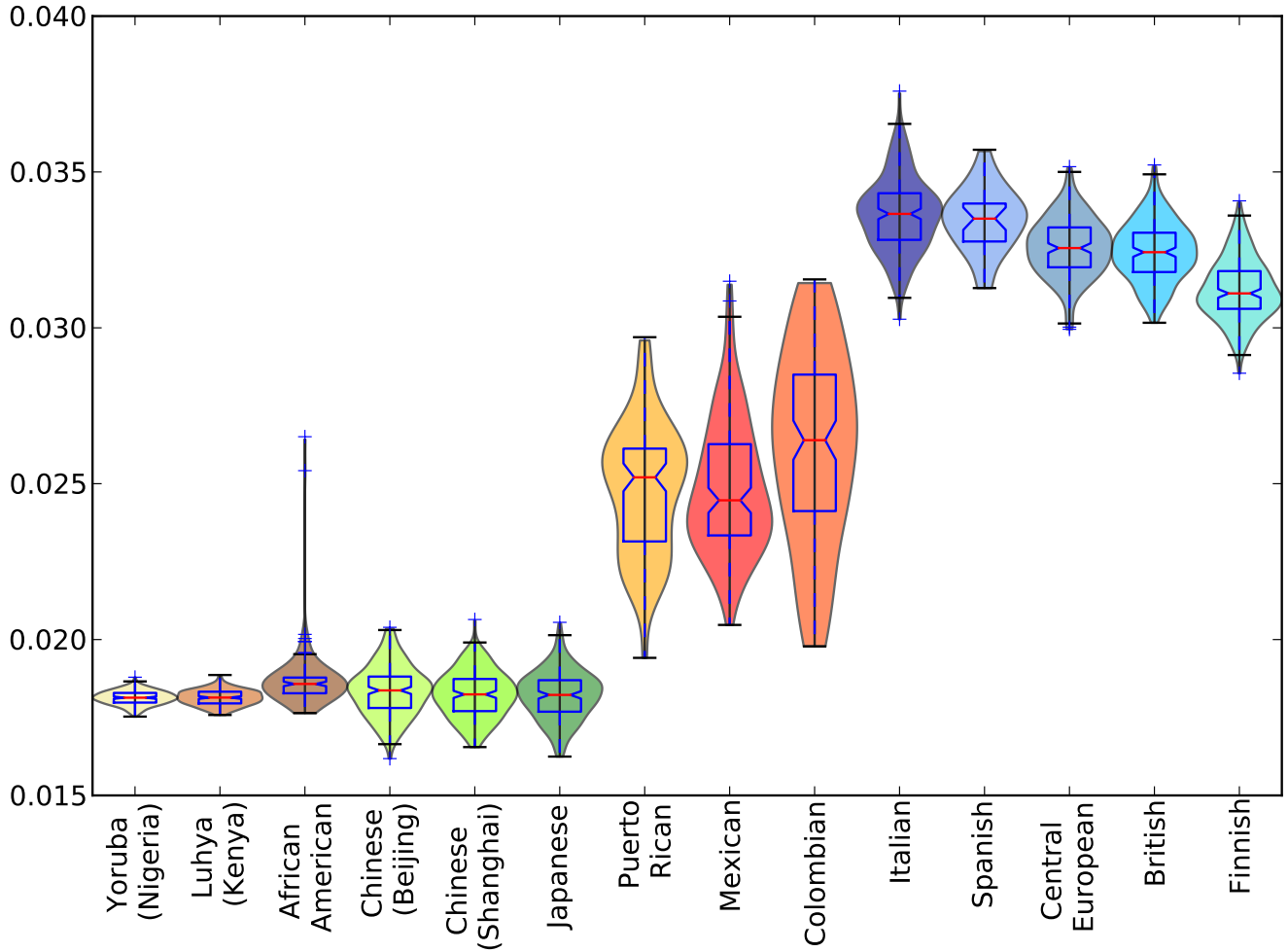


The frequency of 5'-TCC-3'→5'-TTC-3' is elevated in Europe

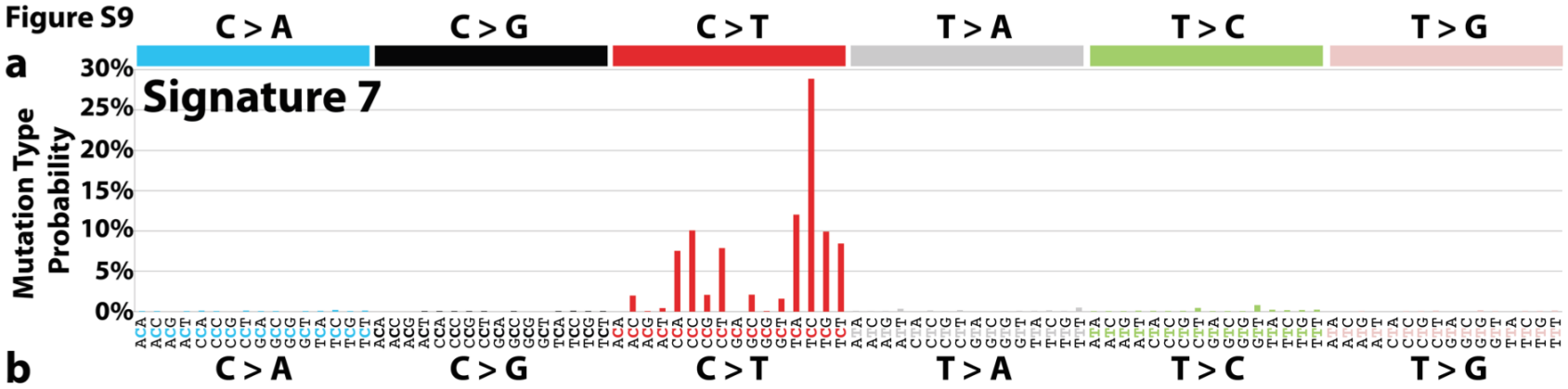
Harris, *PNAS* 2015



Harris, *PNAS* 2015



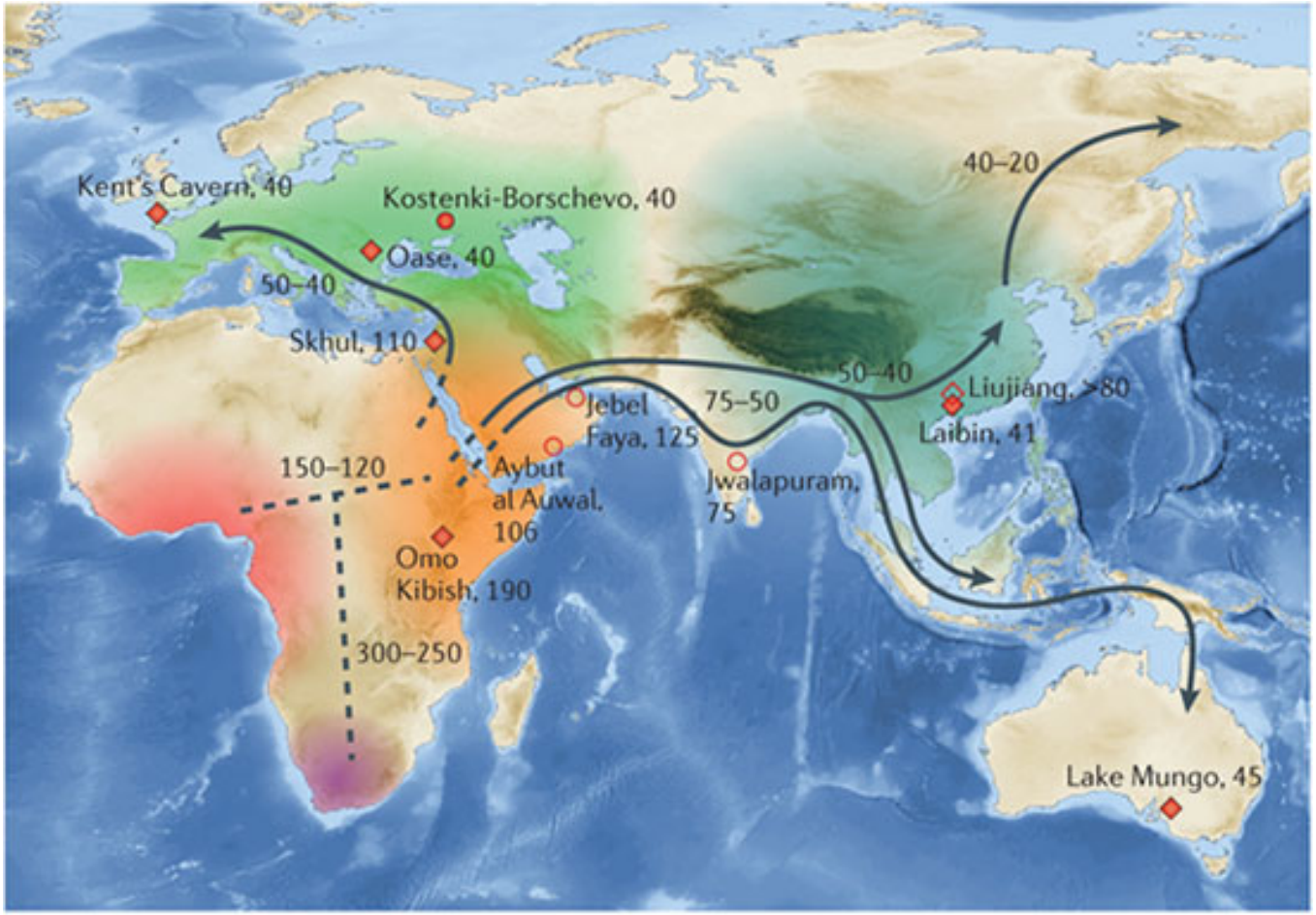
Harris, *PNAS* 2015



Alexandrov, et al. *Nature* 2013

- 5'-TCC-3'→5'-TTC-3' also dominates the mutational signature of melanoma
- Observed in early DNA sequencing of UV-irradiated cell cultures Drobetsky and Sage *Mutation Res* 1993

Mutation rate evolution could complicate efforts to infer human history



Scally and Durbin *Nature Rev Gen* 2012

Has the mutation rate slowed during human evolution?

Rates of Molecular Evolution: The Hominoid Slowdown

Morris Goodman

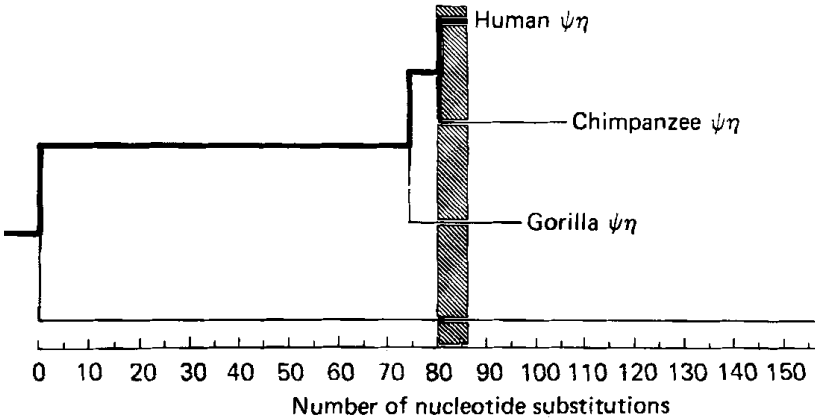


Fig. 5. *Evidence that among higher primates, at the genetic DNA sequence level, changed the least in descent from the Anthropoidea ancestor to the present. Thi phylogenetic reconstruction carried out by the maximum parsimony method, usi sequence data on the η-globin genetic locus.[8] This locus, which is an active emb*

Goodman *BioEssays* 1985

NATURE VOL. 326 5 MARCH 1987 LETTERS

The molecular clock runs more slowly in man than in apes and monkeys

Wen-Hsiung Li & Masako Tanimura

Center for Demographic and Population Genetics, University of Texas, PO Box 20334, Houston, Texas 77225, USA

Li and Tanimura *Nature* 1987