Introduction to Population Genetics

Timothy O'Connor
timothydoconnor@gmail.com
Ryan Hernandez
ryan.Hernandez@mcgill.ca

Learning Objectives

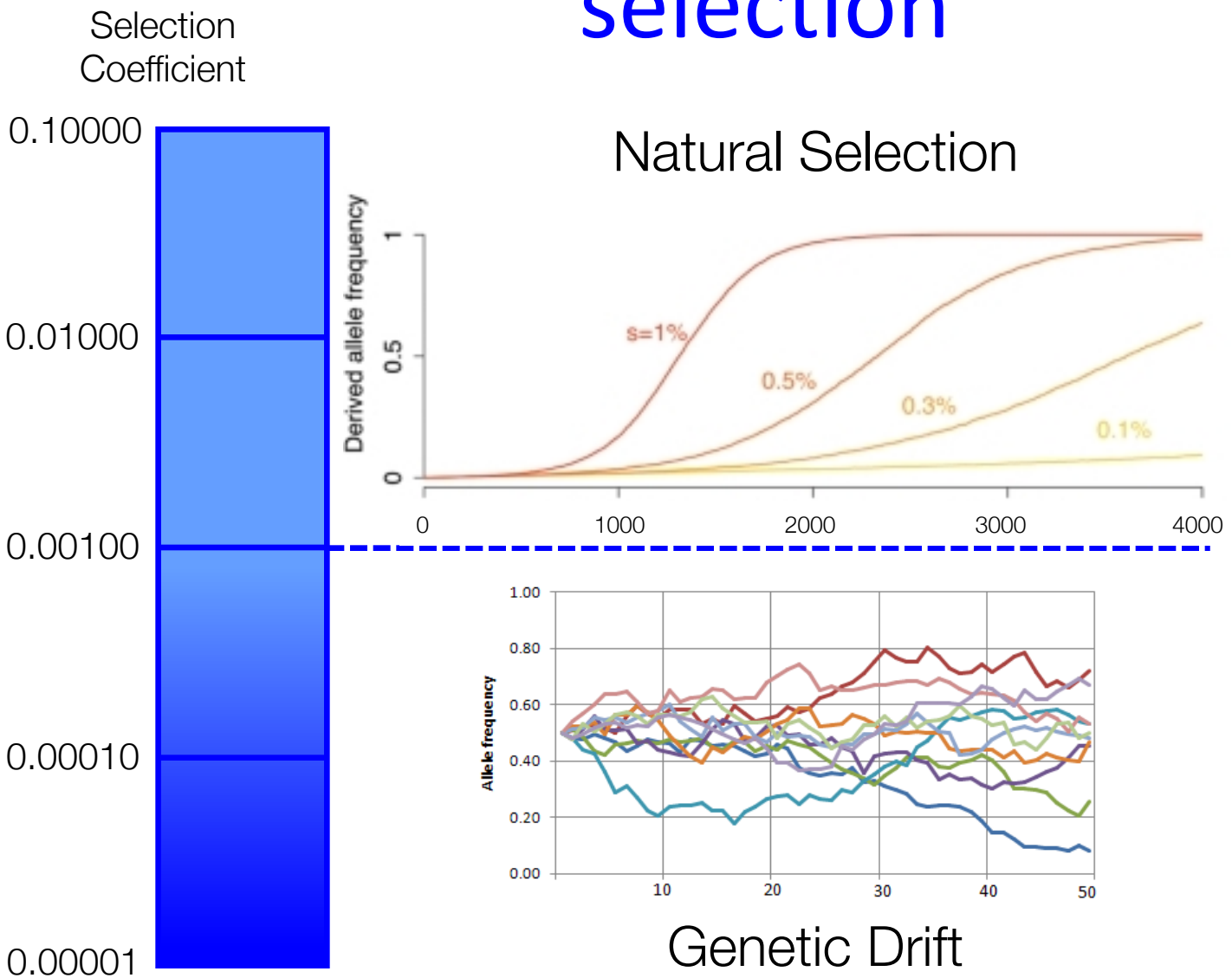
- Describe the key evolutionary forces
- How demography can influence the site frequency spectrum
  - Be able to interpret a site frequency spectrum
  - Understand how the SFS is affected by evolutionary forces
  - How we can use the SFS to understand evolutionary history of a population.

Review: What are the assumptions of Hardy-Weinberg?

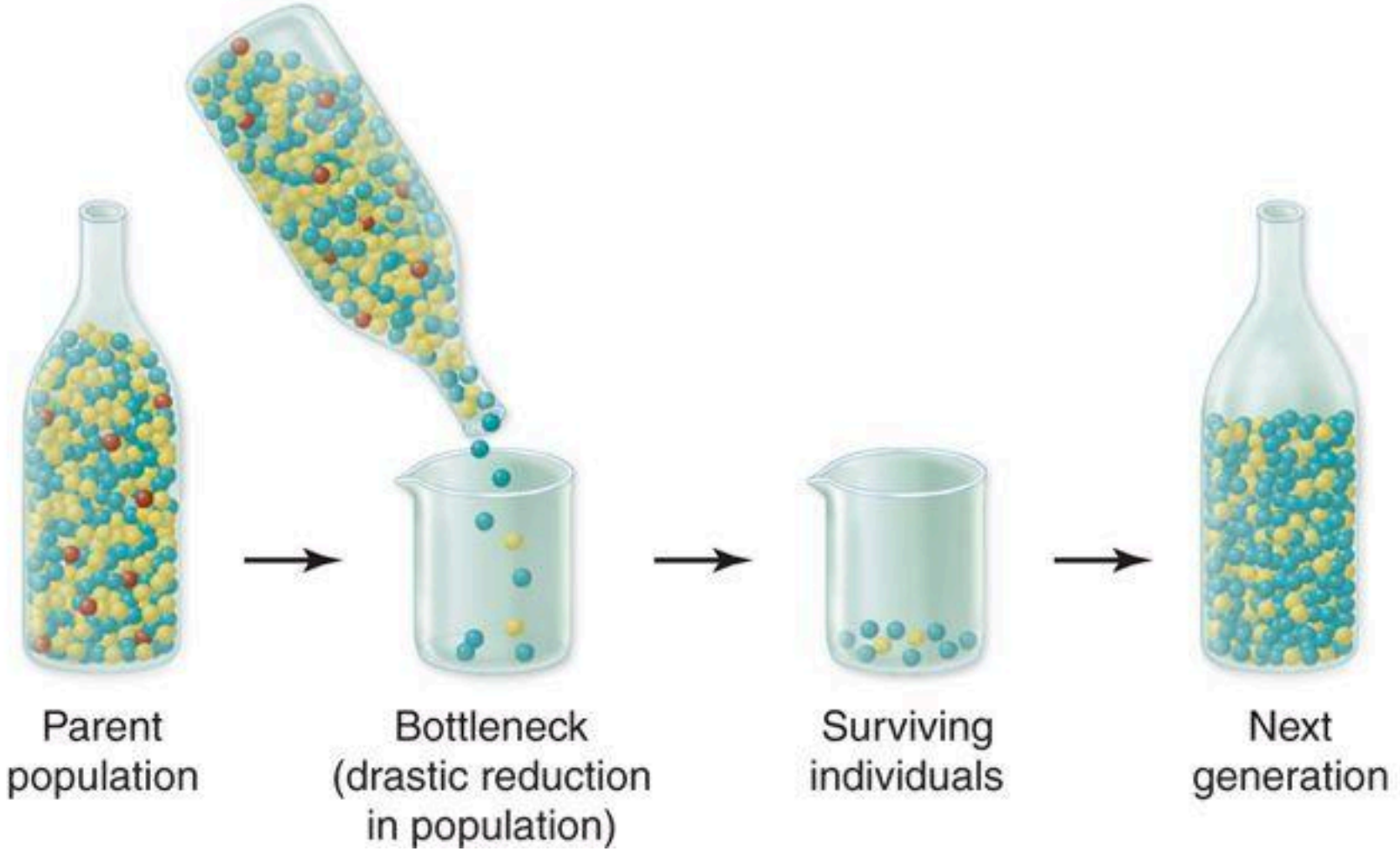
1) There must be no mutation
2) There must be no migration
3) Individuals must mate at random with respect to genotype
4) There must be no selection
5) The population must be infinitely large

How do these affect allele frequencies?

Drift, mutation, migration, and selection

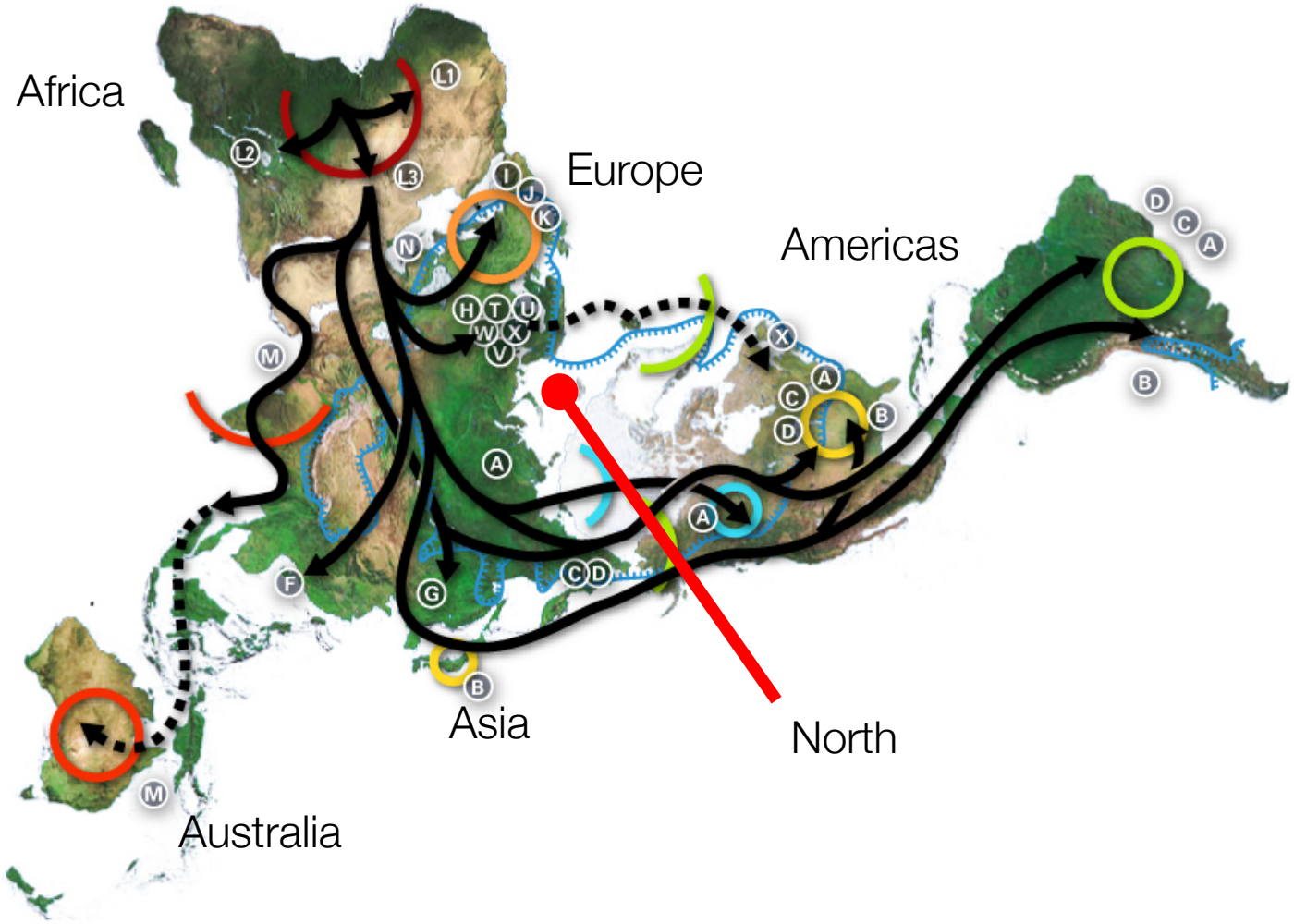


Genetic drift: Serial founder effect

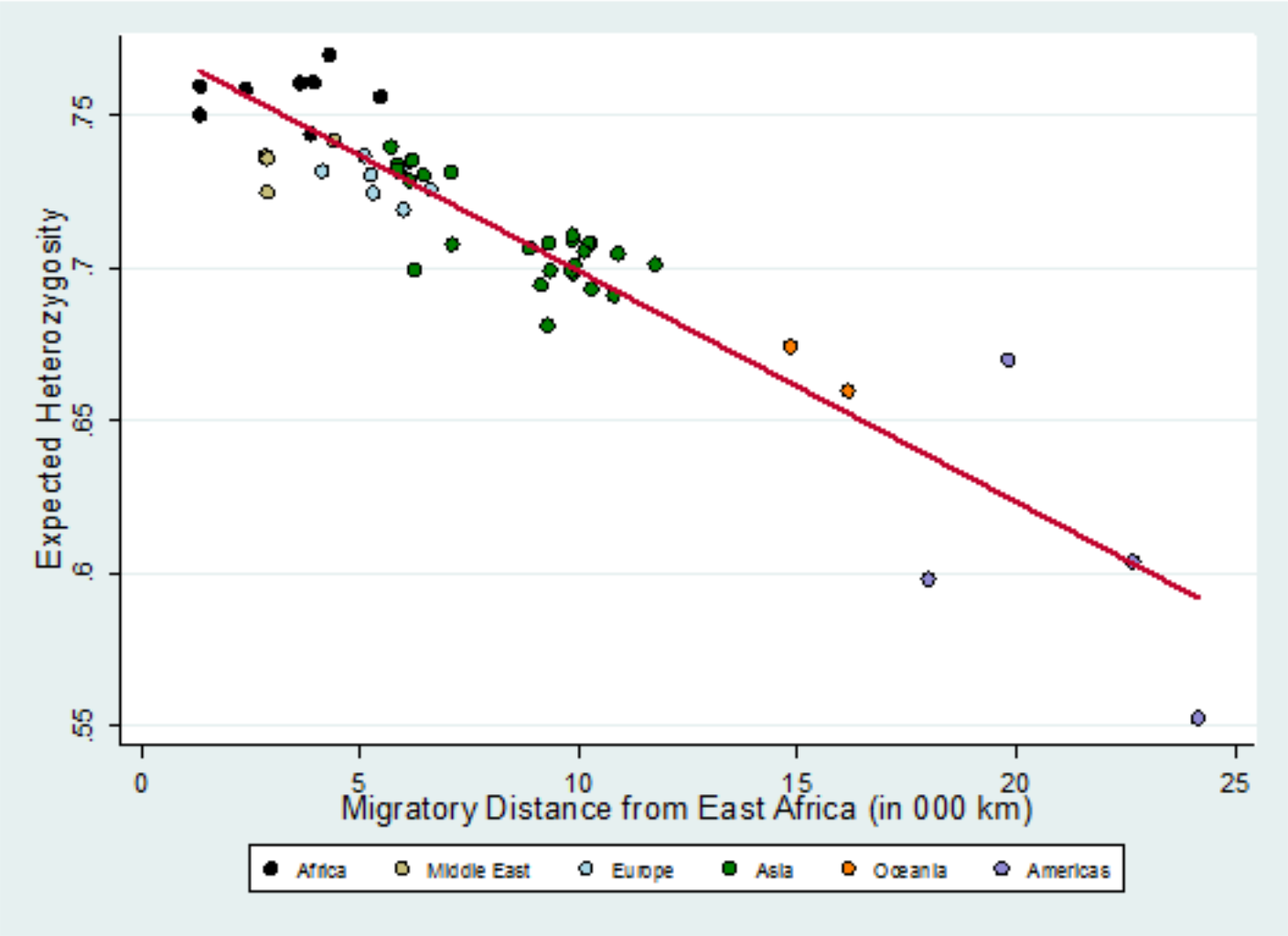


Out of Africa Model!

We now have an excellent “road map” of how humans evolved in Africa and migrated to populate the rest of the earth.



Heterozygosity is correlated with distance from East Africa



Ramachandran et al. 2005 PNAS

Mutation: How often do mutations arise?

Table 1. Estimated per generation mutation rates in mice

	Homozygous		Heterozygous		Final generation	Overall
	No.	Rate ($\times 10^{-9}$) (95% CI)	No.	Rate ($\times 10^{-9}$) (95% CI)	Rate ($\times 10^{-9}$)	Rate (95% CI)
SNV						
conA	63.3[a]	3.4 (2.6–4.4)[a]	101.5	5.7 (4.4–7.3)	6.9	5.4×10^{-9} ($4.6–6.5, \times 10^{-9}$)
conB	117.5[a]	5.1 (4.3–6.2)[a]	92.7	5.2 (4.0–6.7)	6.8	
mutC	1304	84.3 (76.1–94.5)	1944	110.6 (94.0–132.5)	150	9.4×10^{-8} ($9.0–9.8, \times 10^{-8}$)
mutD	1472	86.9 (79.1–96.6)	1633	92.3 (78.6–110.5)	90	
Indel						
conA	6.7[a]	0.57 (0.22–1.20)[a]	4	0.35 (0.10–0.91)	—	3.1×10^{-10} ($1.2–6.4, \times 10^{-10}$)
conB	4[a]	0.28 (0.08–0.71)[a]	3	0.26 (0.05–0.77)	—	
mutC	28	2.9 (1.9–4.1)	28	2.5 (1.7–3.6)	—	2.7×10^{-9} ($2.2–3.2, \times 10^{-9}$)
mutD	21	2.0 (1.2–3.0)	37	3.3 (2.3–4.5)	—	

Mutation rates per nucleotide per generation were estimated using the number of homozygous or heterozygous de novo mutations in conA, conB, mutC, and mutD. The estimates for SNVs were validated by counting newly arisen mutations in the final generation. The number of de novo mutations in conA and conB was partly adjusted for the frequency of true de novo variants; 95% confidence intervals (CI) were calculated by computer simulation or Poisson distribution error analysis of the number of mutations (details in Supplemental Methods).

[a]Note that homozygous variant numbers in control lines were uncertain due to the low ability to discriminate between de novo and initial variants; these values were not used in the estimates for the overall rate.

Uchimura et al. (2015) Genome Research

Mutation is a major source of phenotypic variation

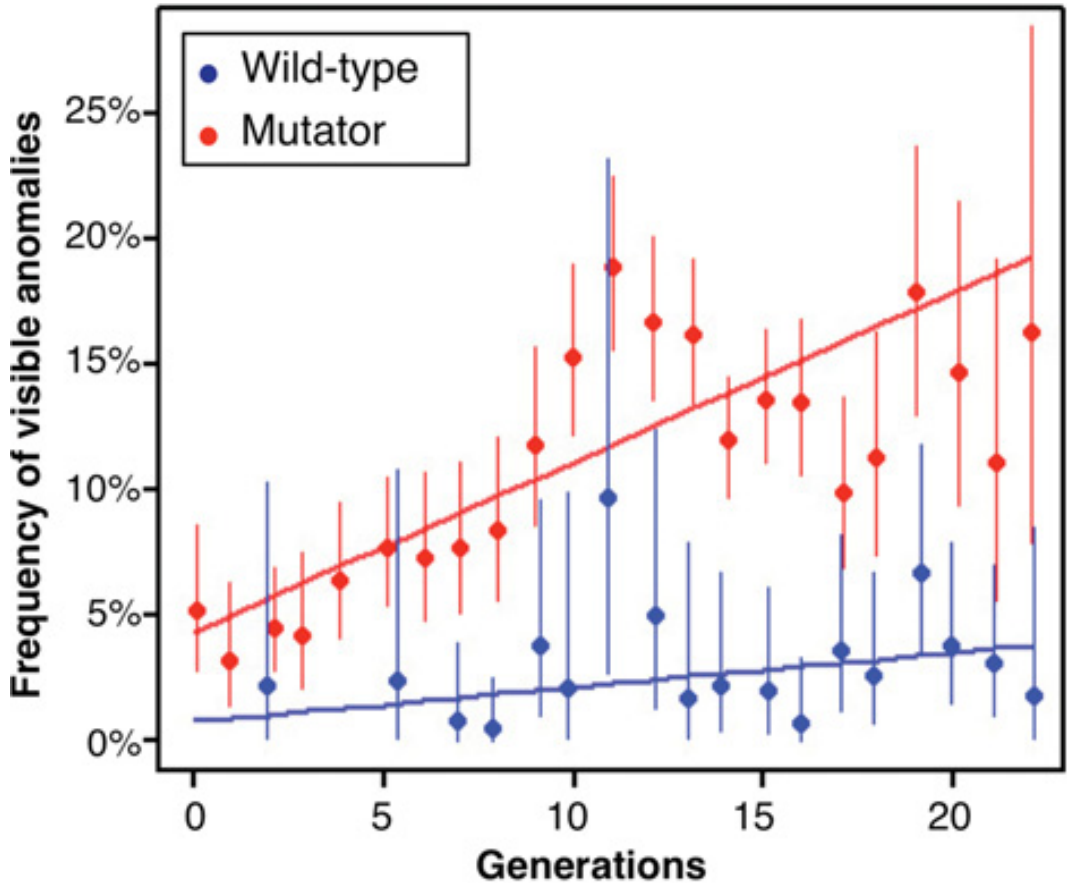


Figure 2. Frequency of visible phenotypic anomalies in breeding lines. Frequency of visible anomalies in each successive generation. Circles indicate observed frequencies with 90% CI, determined by Fisher's exact test. Since fewer than 20 mice were screened in the early-generation (fewer than seven generations) populations of control mice, mean phenotypic frequencies are shown for generations 0–3 and 4–6. Solid lines show the fit with a binomial linear model.

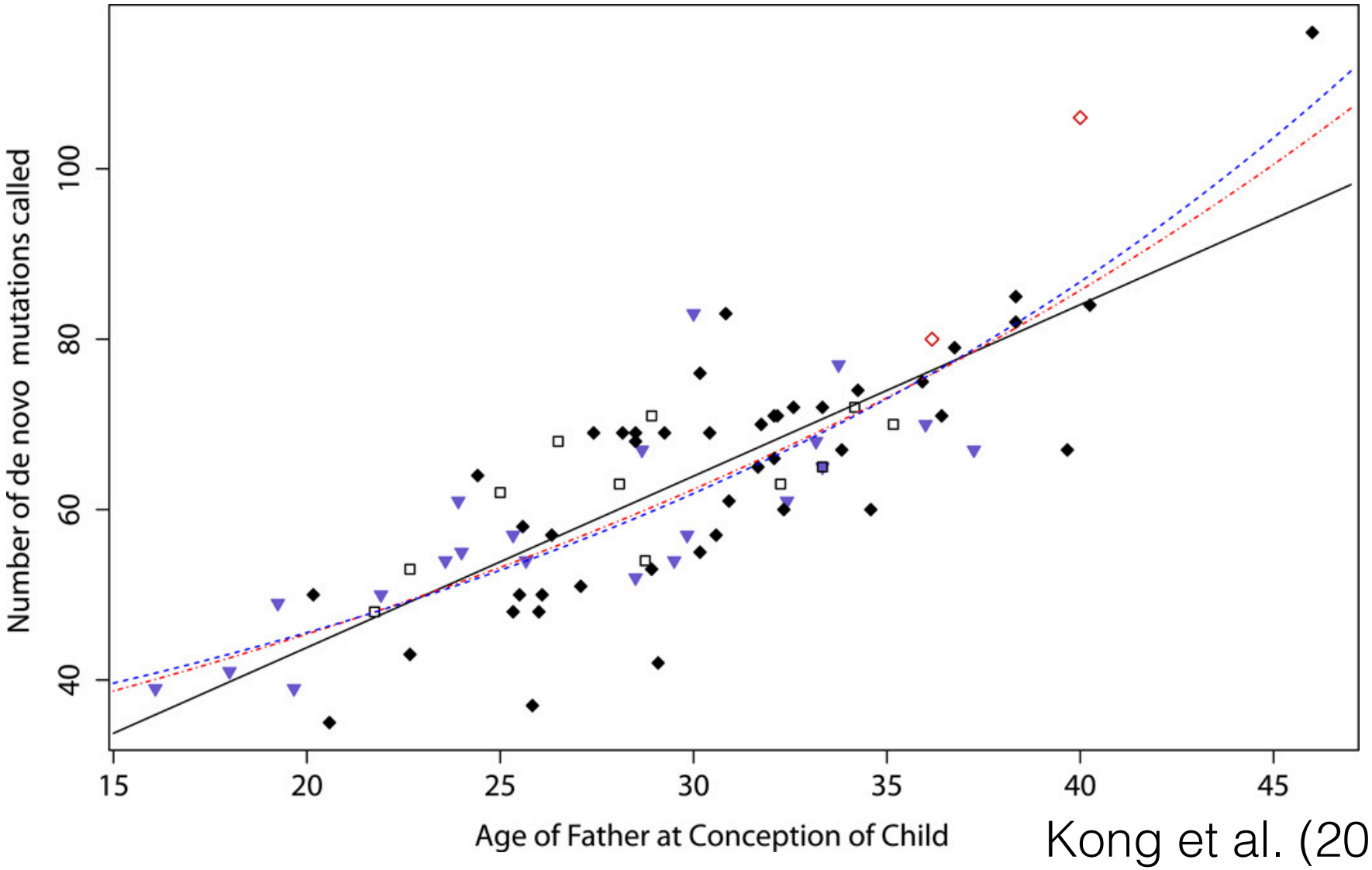
Uchimura et al. (2015) Genome Research

Mutation: How often do mutations arise in Humans?

study	loci considered	per-generation mean mutation rate (10^{-8} $bp^{-1}generation^{-1}$)	yearly mean mutation rate (10^{-9} $bp^{-1}y^{-1}$)	
			t_{gen} = 30 y	t_{gen} = 25 y
Kondrashov (2003)	disease	1.85 (0.00–3.65)	0.62 (0.00–1.22)	0.74 (0.00–1.46)
Lynch (2010)	disease	1.28 (0.68–1.88)	0.42 (0.23–0.63)	0.51 (0.27–0.75)
Roach *et al*. (2010)	WG	1.10 (0.68–1.70)	0.37 (0.23–0.57)	0.44 (0.27–0.68)
Awadalla *et al*. (2010)	WG	1.36 (0.34–2.72)	0.45 (0.11–0.91)	0.54 (0.14–1.09)
1000 Genomes Project (2010), CEU	WG	1.17 (0.94–1.73)	0.39 (0.31–0.57)	0.47 (0.38–0.69)
1000 Genomes Project (2010), YRI	WG	0.97 (0.72–1.44)	0.32 (0.24–0.48)	0.39 (0.29–0.58)
Sanders *et al*. (2012)	exome	1.28 (1.05–1.50)	0.43 (0.35–0.50)	0.51 (0.42–0.60)
O'Roak *et al*. (2012)	exome	1.57 (1.05–2.26)	0.52 (0.35–0.75)	0.63 (0.42–0.90)
Kong *et al*. (2012)	WG	1.20	0.40	0.48

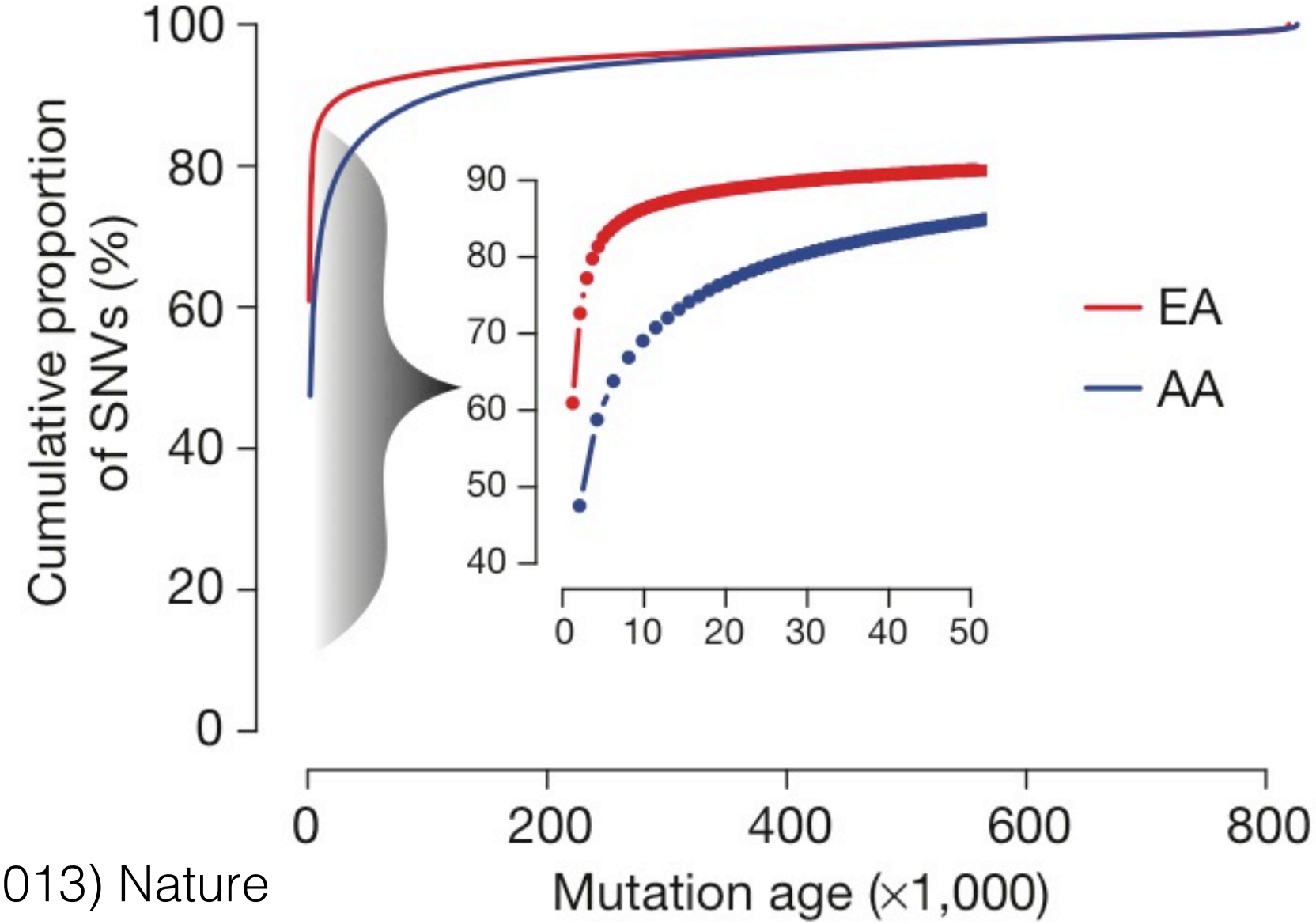
Scally and Durbin (2012) Nature Rev. Genet.

What are the effects of paternal age on mutation rate?



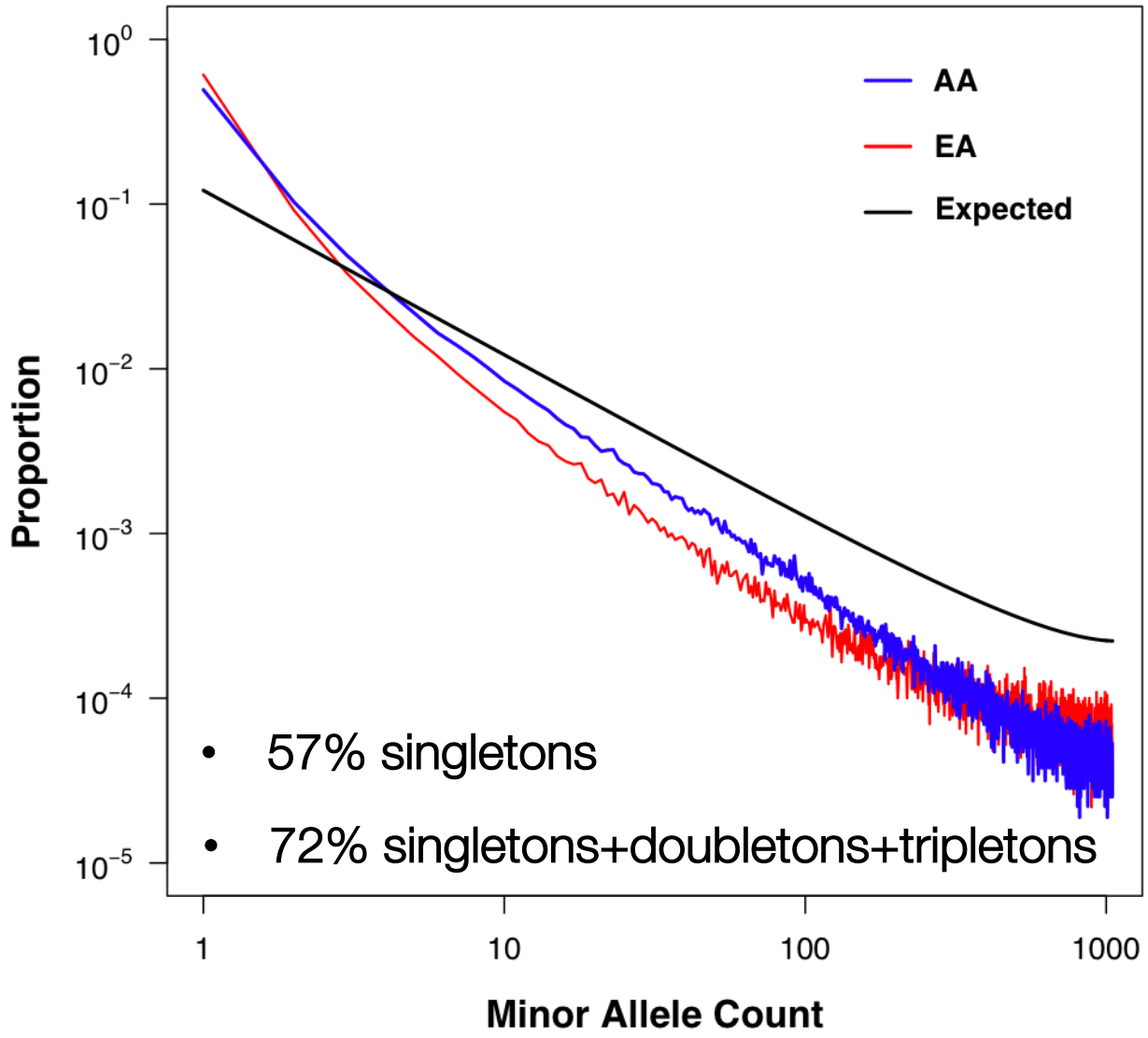
Kong et al. (2012) Nature

When did most variation arise?



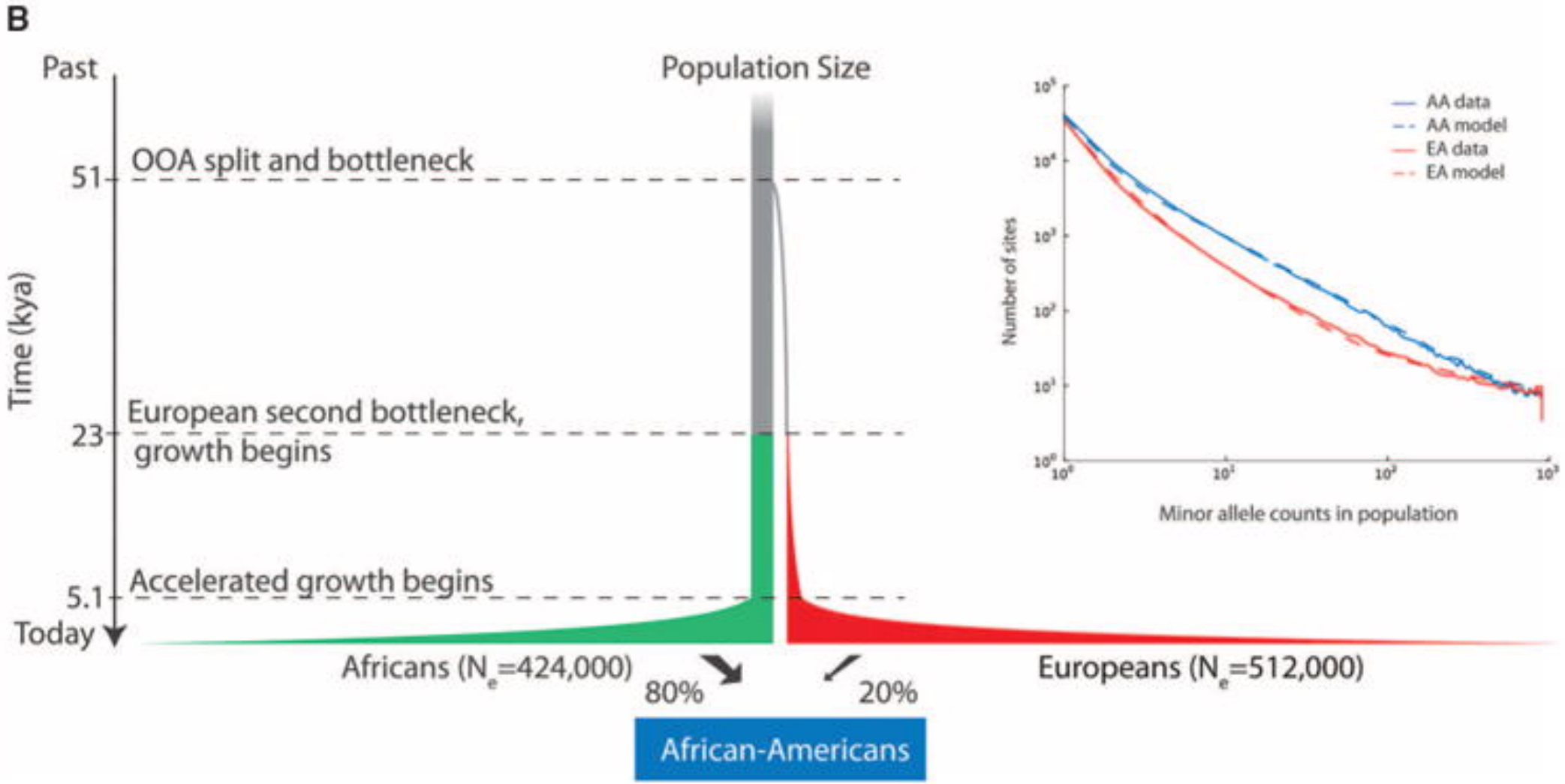
Fu et al. (2013) Nature

Most SNVs are very rare



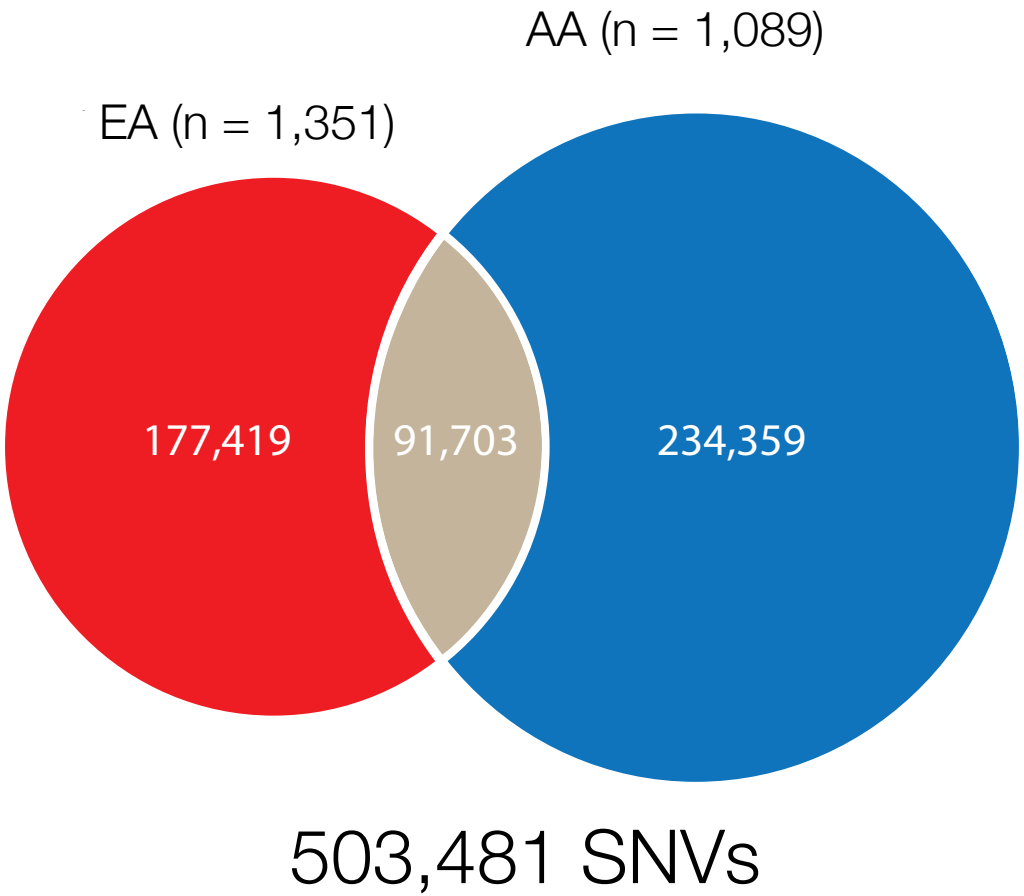
Tennessen et al. (2012) Science

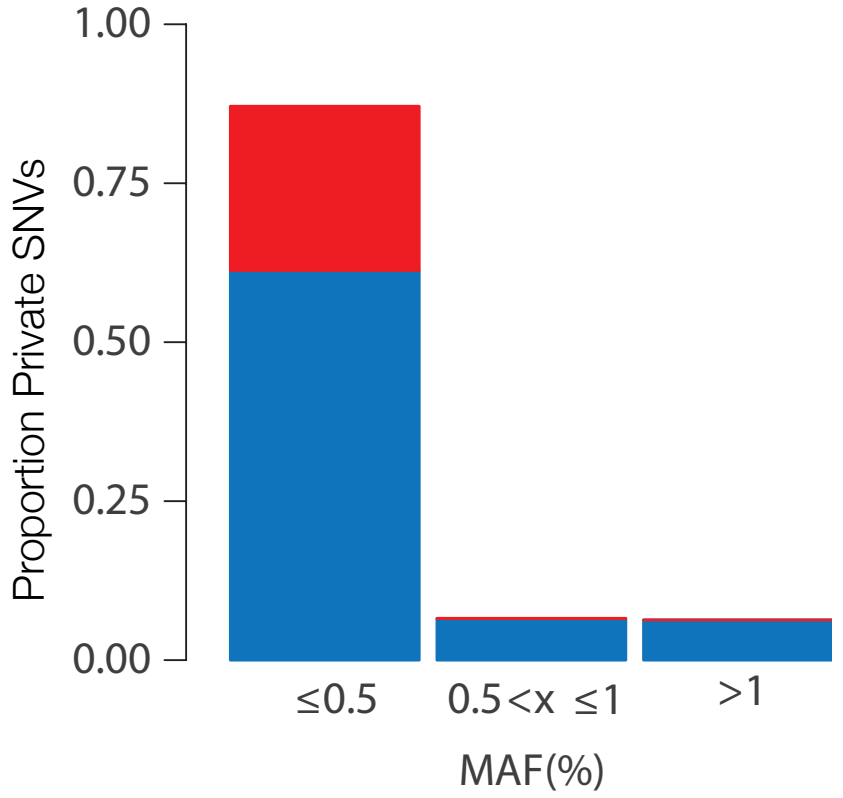
How has our population size grown?



Tennessen et al. (2012) Science

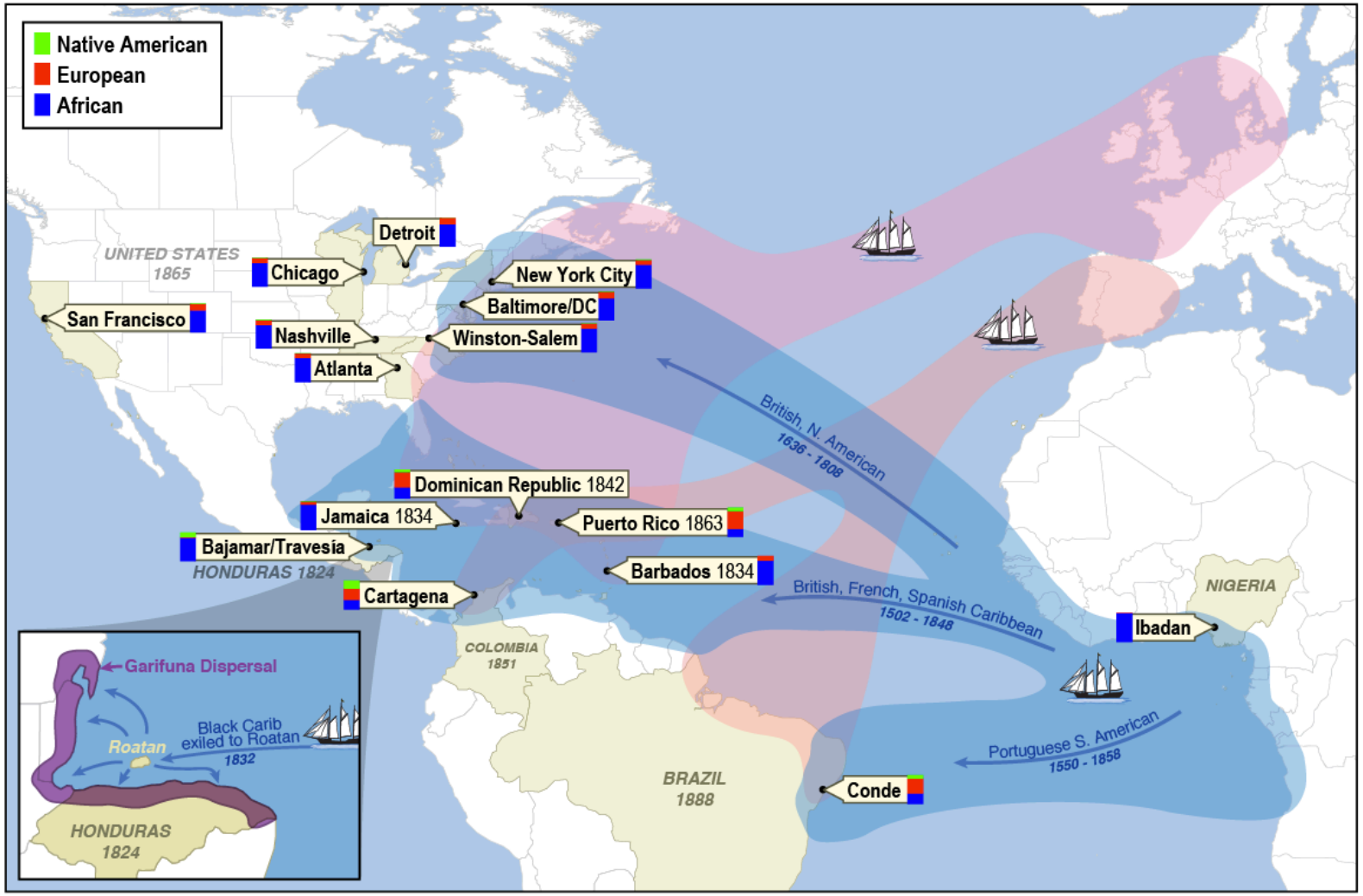
Most SNVs are population specific





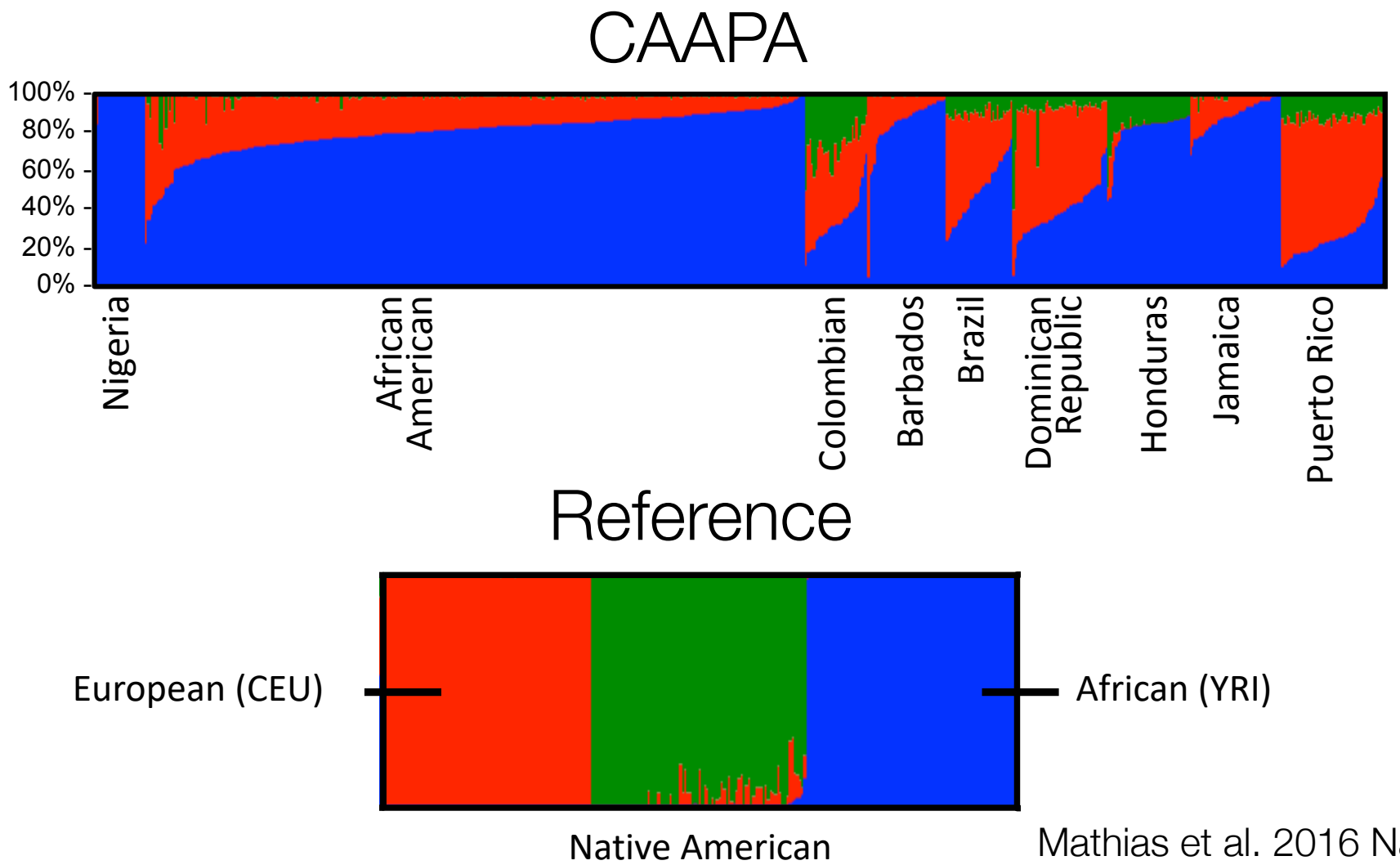
Tennessen et al. (2012) Science

Migration: Admixture is migration between diverged populations



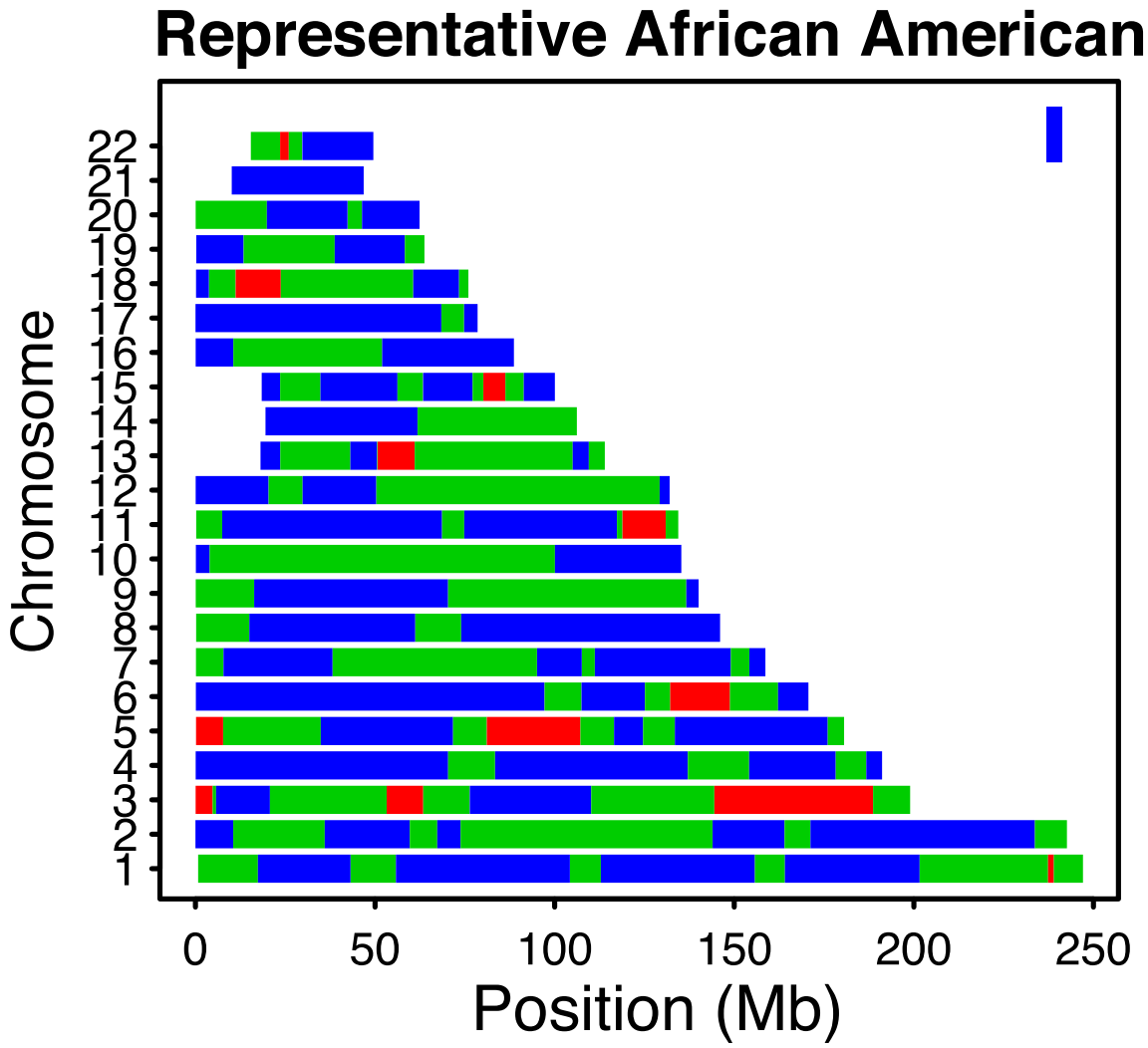
Mathias et al. (2016)
Nature Comm.

Estimates of global ancestry



Mathias et al. 2016 Nature Comm.

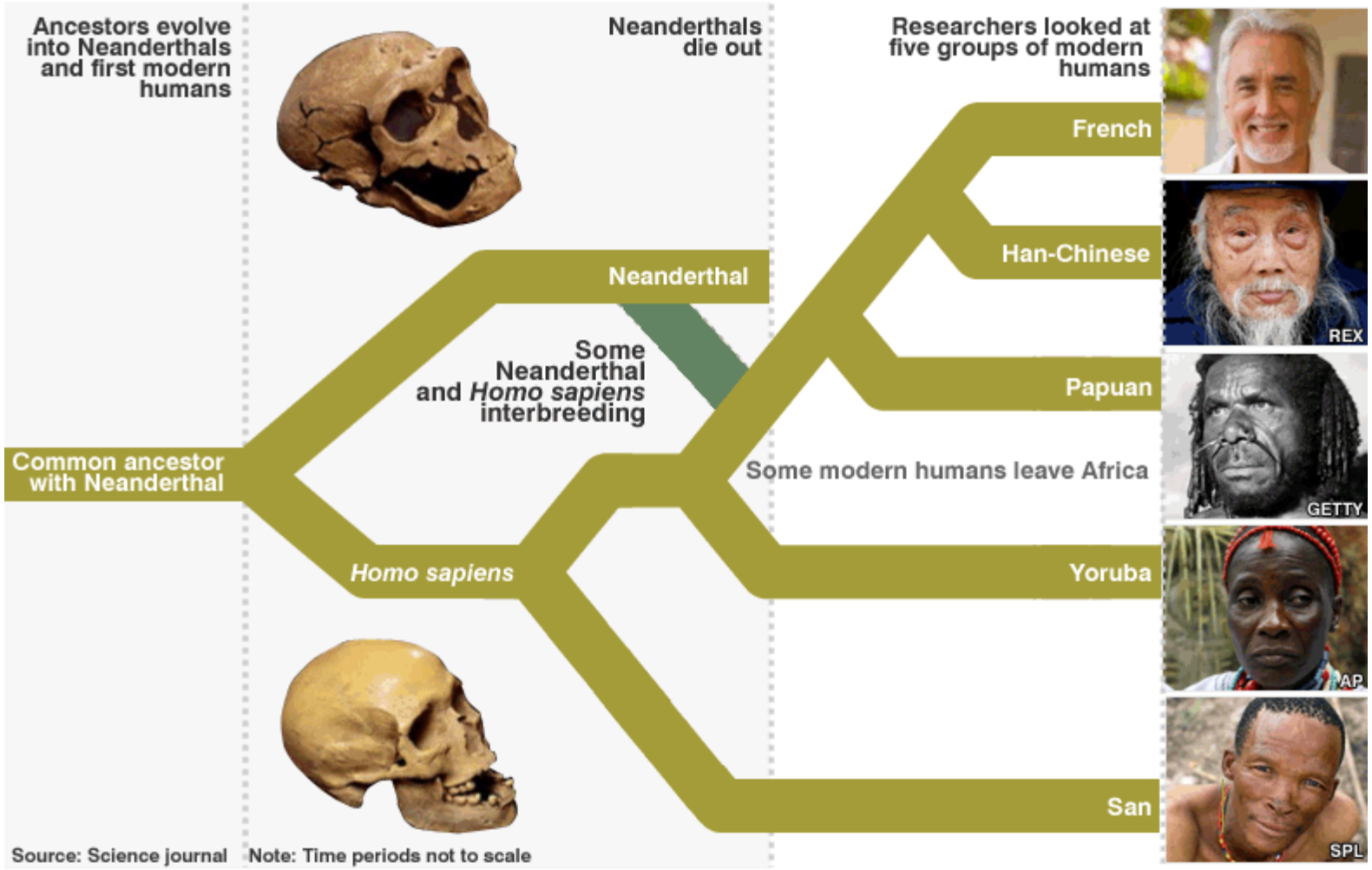
Local ancestry of a single individual



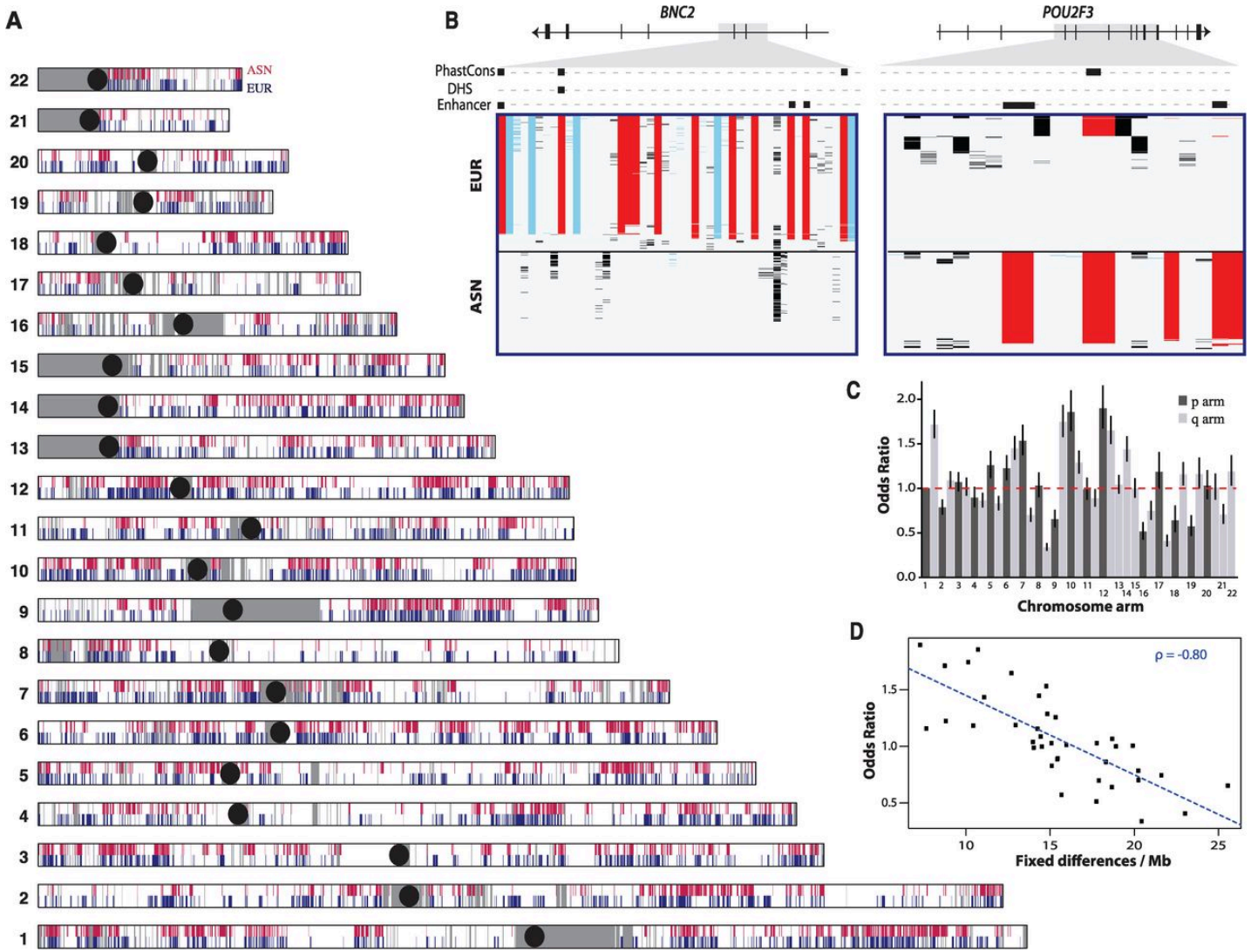
Bryc et al. 2009 PNAS

Ancient admixture: Neanderthals are still among us

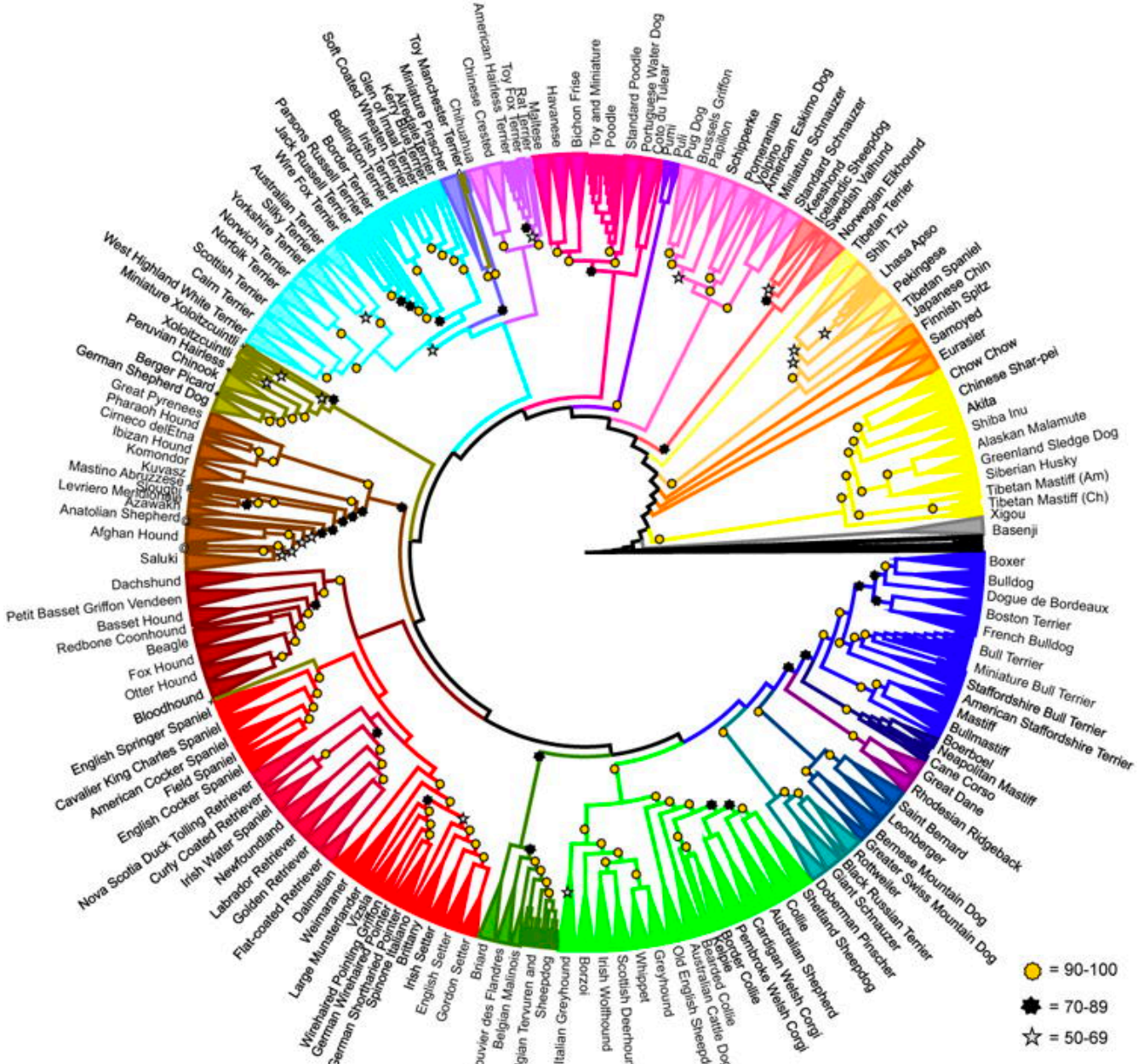
- Recent genetic data suggests that 1-4% of non-African genomes are derived from Neanderthals



Neanderthals are still among us



Dog breeding has produced both divergent groups



Parker et al. (2017) Cell Rep.

Dog breeding has produced both divergent groups and recent cross breeding is migration

Parker et al. (2017) Cell Rep.

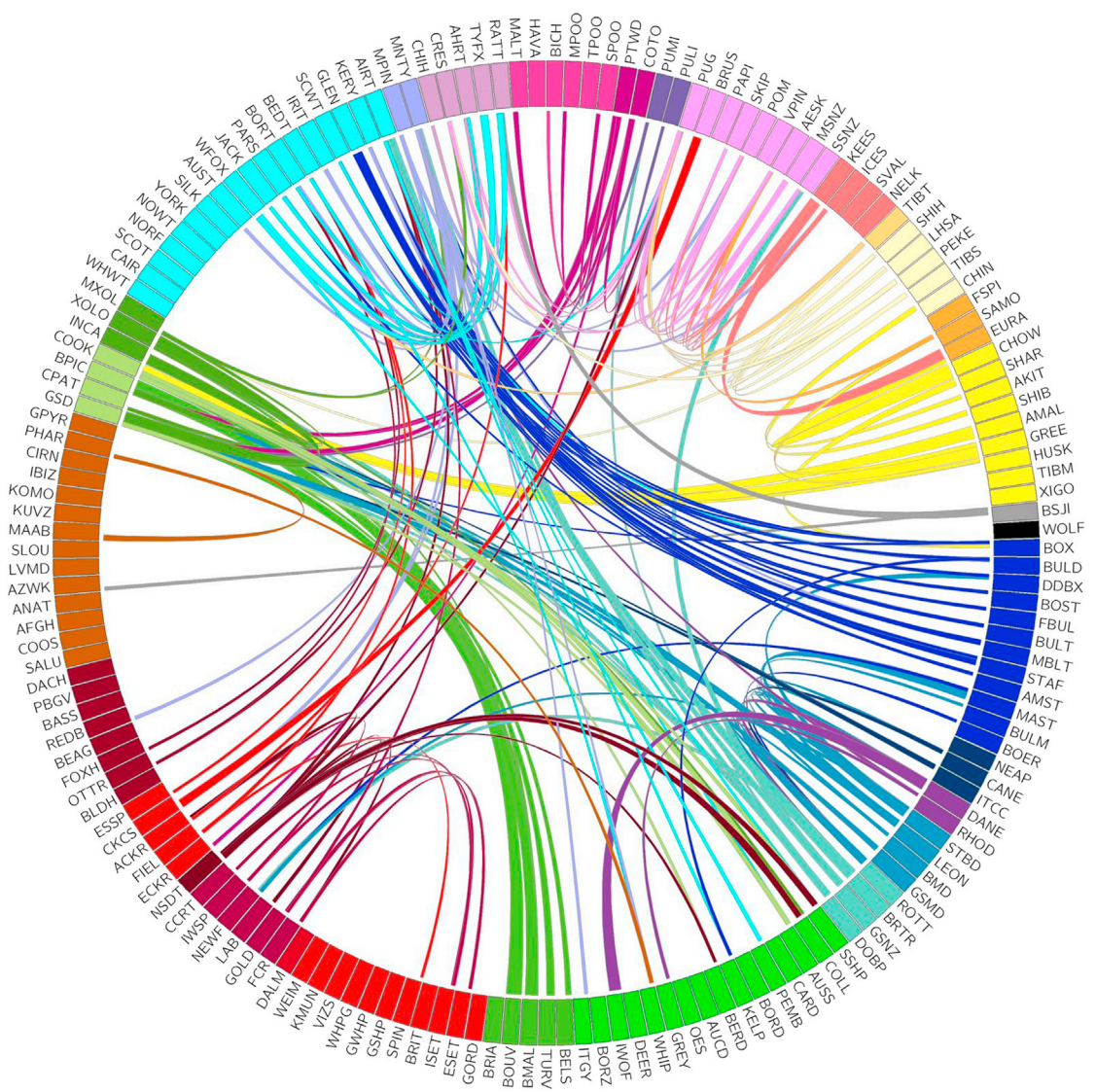


Figure 4. Haplotype Sharing between Breeds from Different Phylogenetic Clades

The circos plot is ordered and colored to match the tree in Figure 1. Ribbons connecting breeds indicate a median haplotype sharing between all dogs of each breed in excess of 95% of all haplotype sharing across clades. Definitions of the breed abbreviations can be found in Table S1.

Dog breeding has produced both divergent groups and recent cross breeding is migration

Parker et al. (2017) Cell Rep.

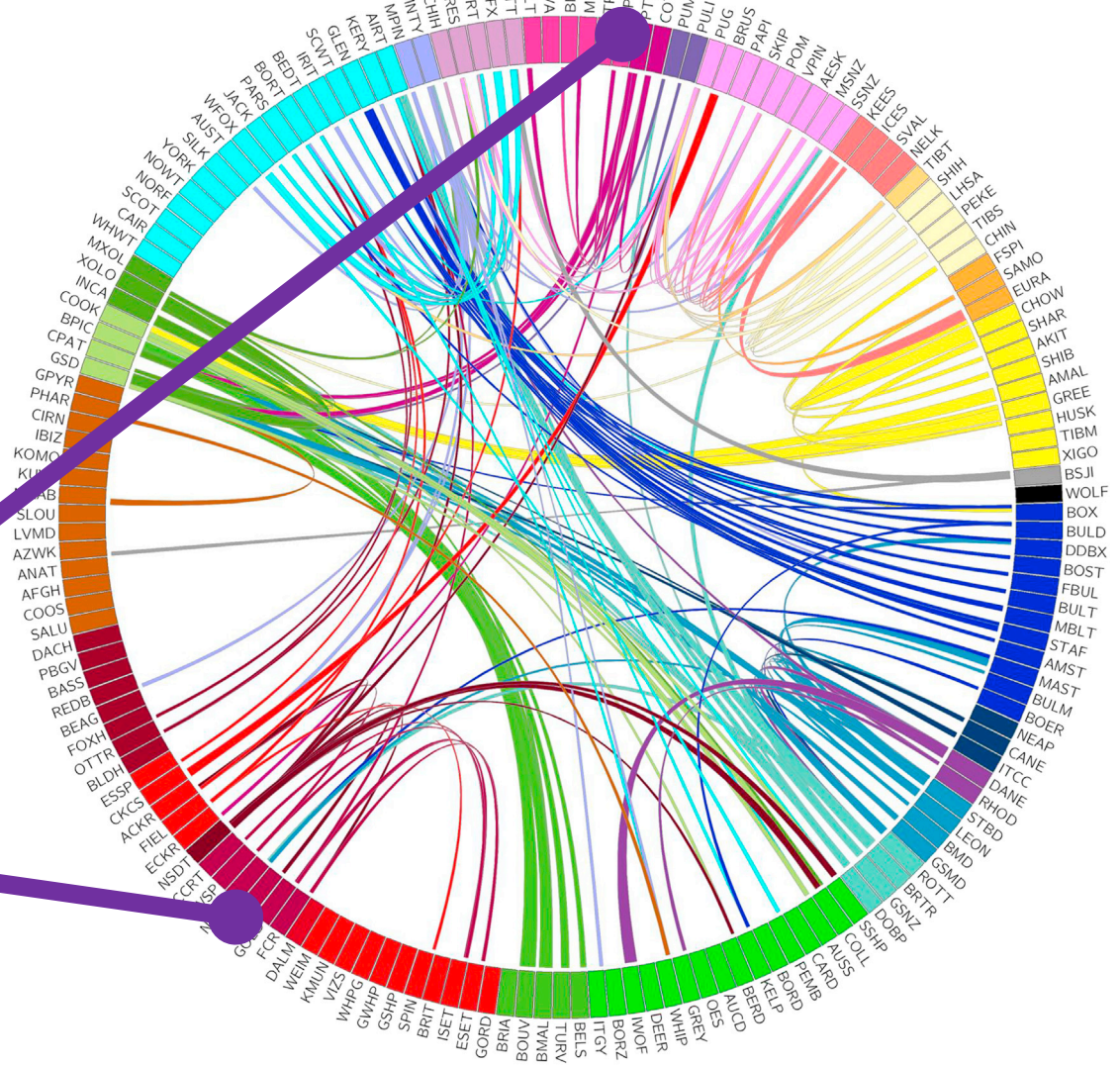
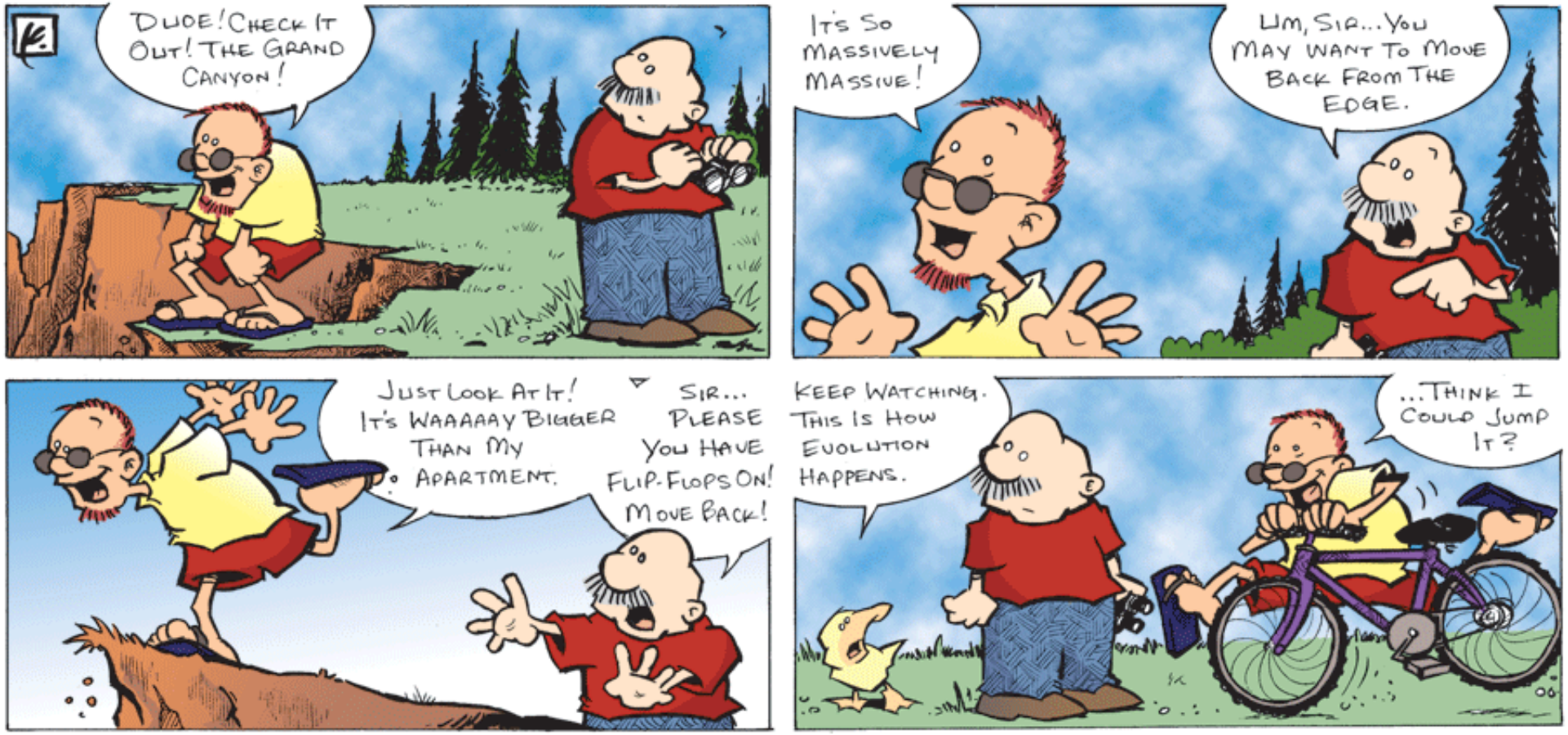


Figure 4. Haplotype Sharing between Breeds from Different Phylogenetic Clades

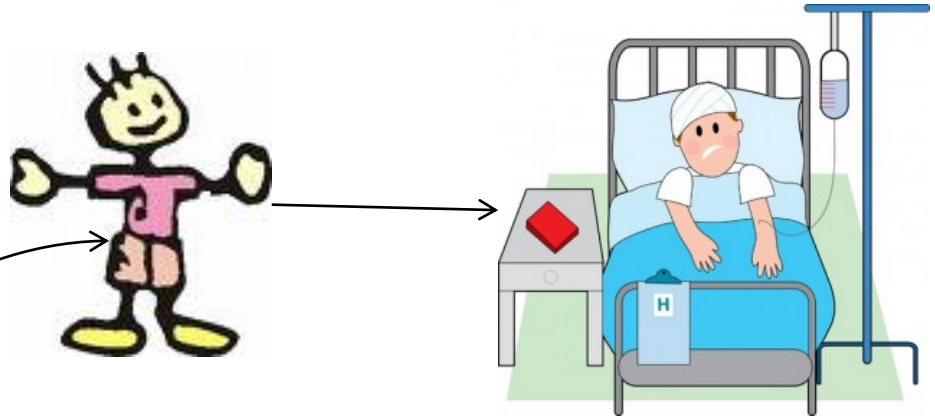
The circos plot is ordered and colored to match the tree in Figure 1. Ribbons connecting breeds indicate a median haplotype sharing between all dogs of each breed in excess of 95% of all haplotype sharing across clades. Definitions of the breed abbreviations can be found in Table S1.

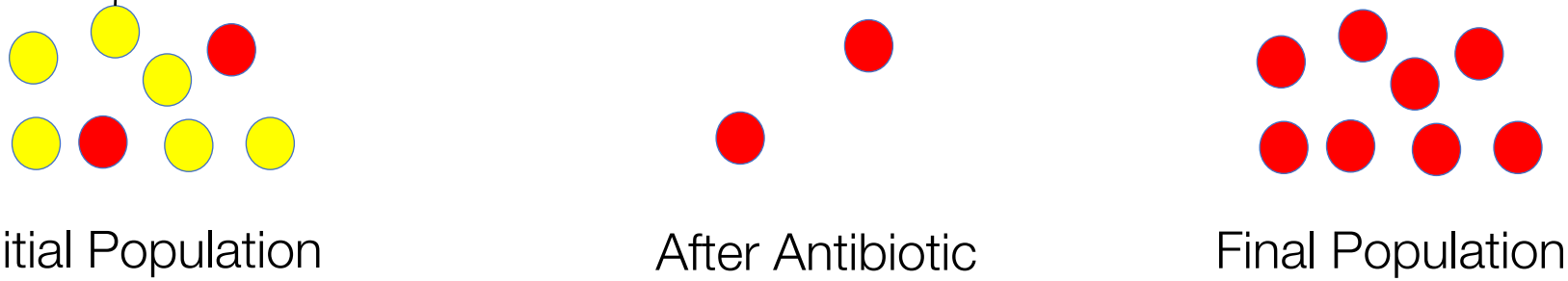
Adaptive (Darwinian) Selection

"I have called this principle, by which each slight variation, if useful, is preserved, by the term Natural Selection."—Charles Darwin from "The Origin of Species", 1859



Antibiotic resistance is an example of adaptive evolution





Initial Population

After Antibiotic

Final Population

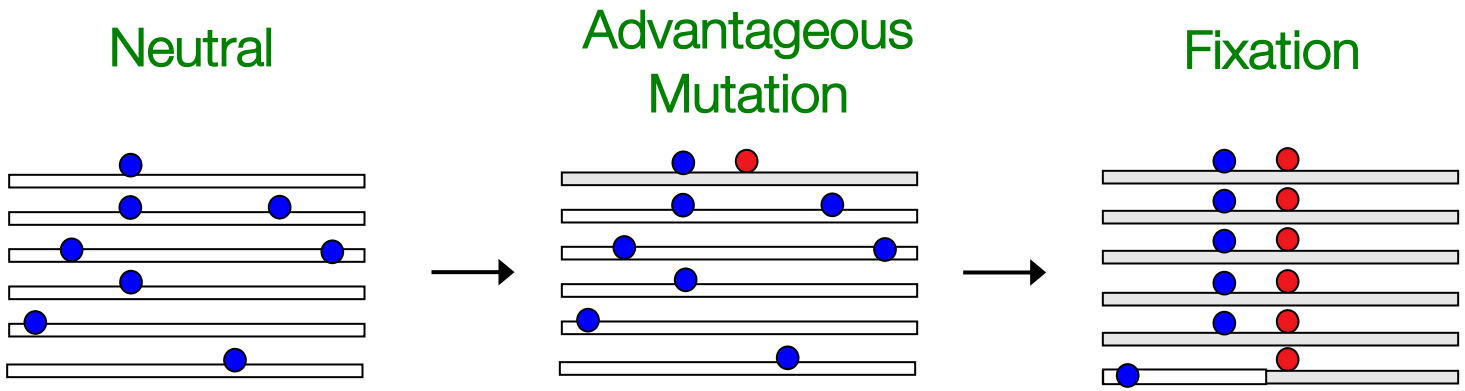
G

Not Resistant

T

Resistant

Reading the genome for signatures of positive selection



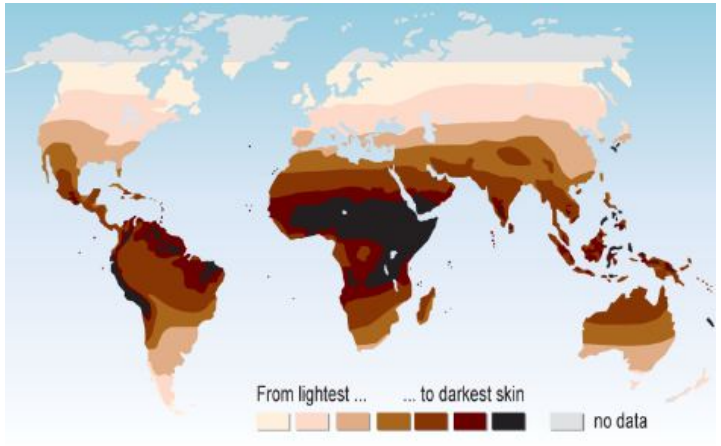
- This process imparts “signatures” on patterns of genetic variation that we can use to find adaptively evolving genes

Genes that influence physical traits have been targets of recent selection

Eye Color

HERC2

Skin Pigmentation



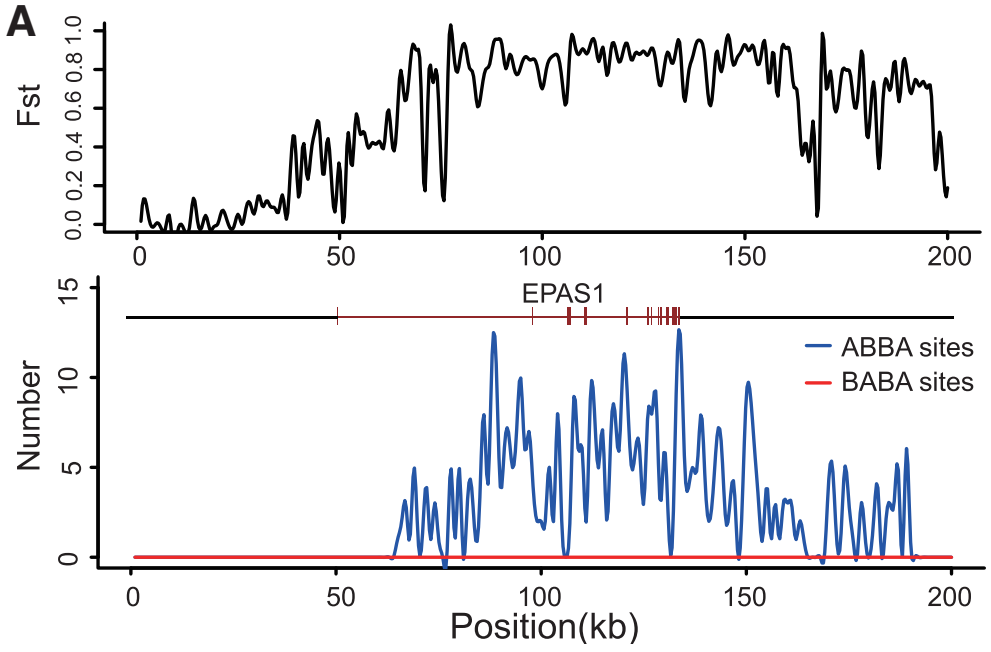
Source: Chaplin G.[©] , *Geographic Distribution of Environmental Factors Influencing Human Skin Coloration*, American Journal of Physical Anthropology 125:292–302, 2004; map updated in 2007.

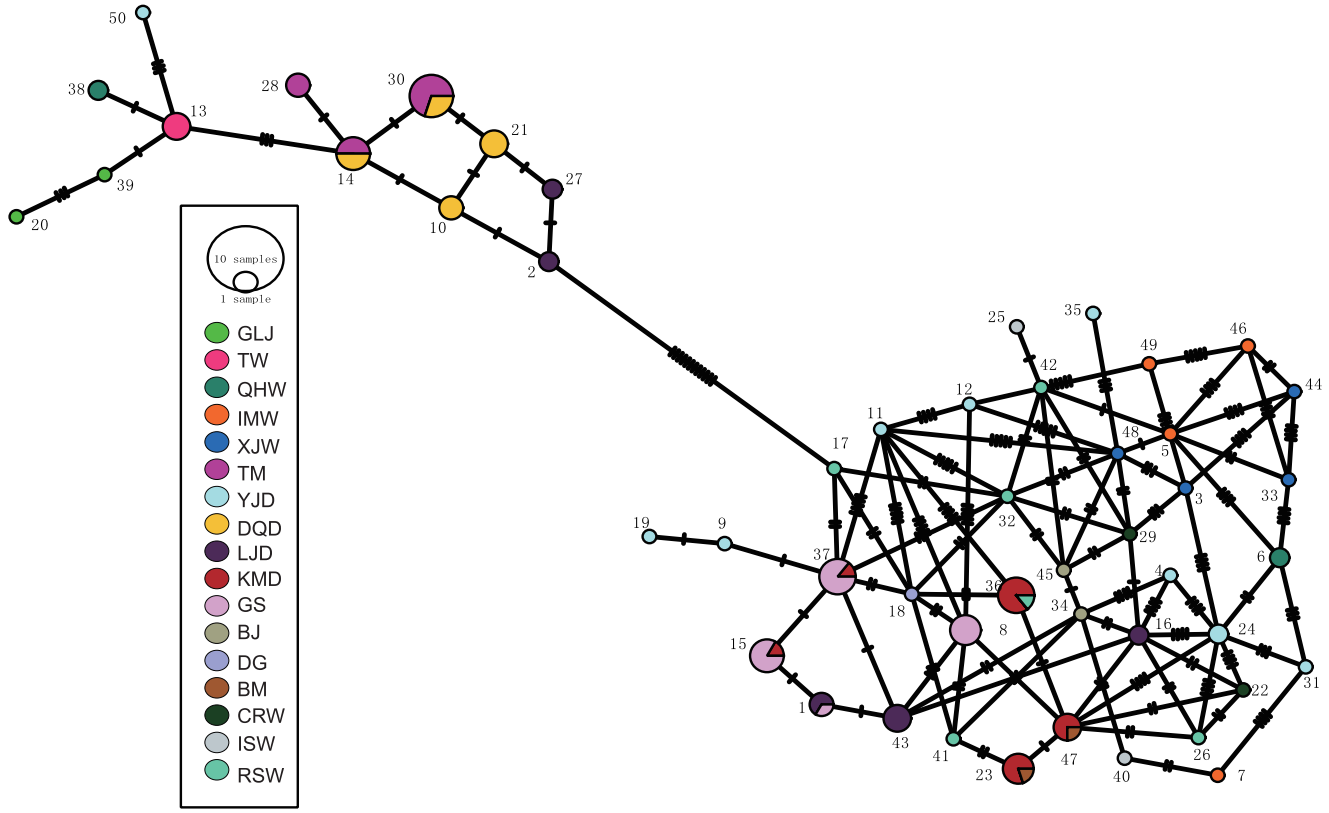
SLC24A5, OCA2, TYRP1

Hair Texture

EDAR

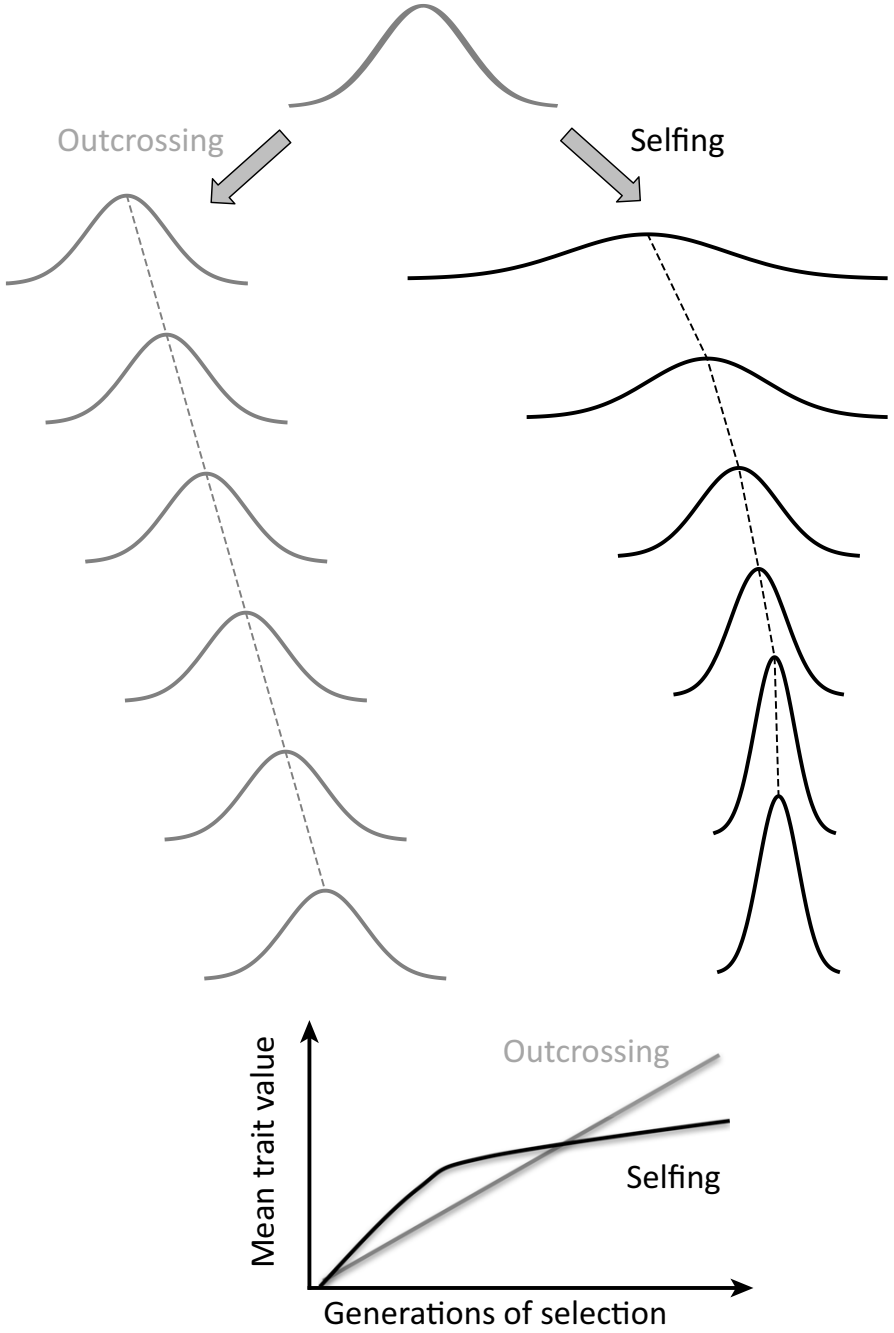
Non-independence of evolutionary forces: Adaptive-migration (introgression)



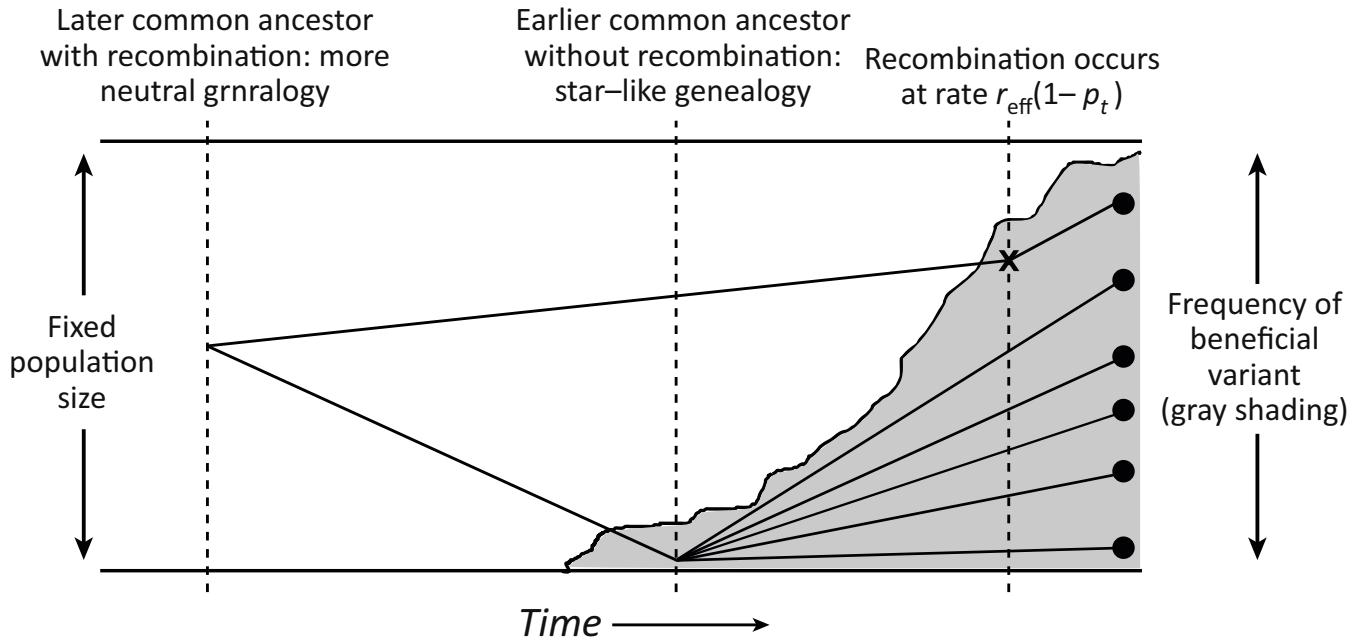


Mio et al. (2017) Molecular Biology and Evolution

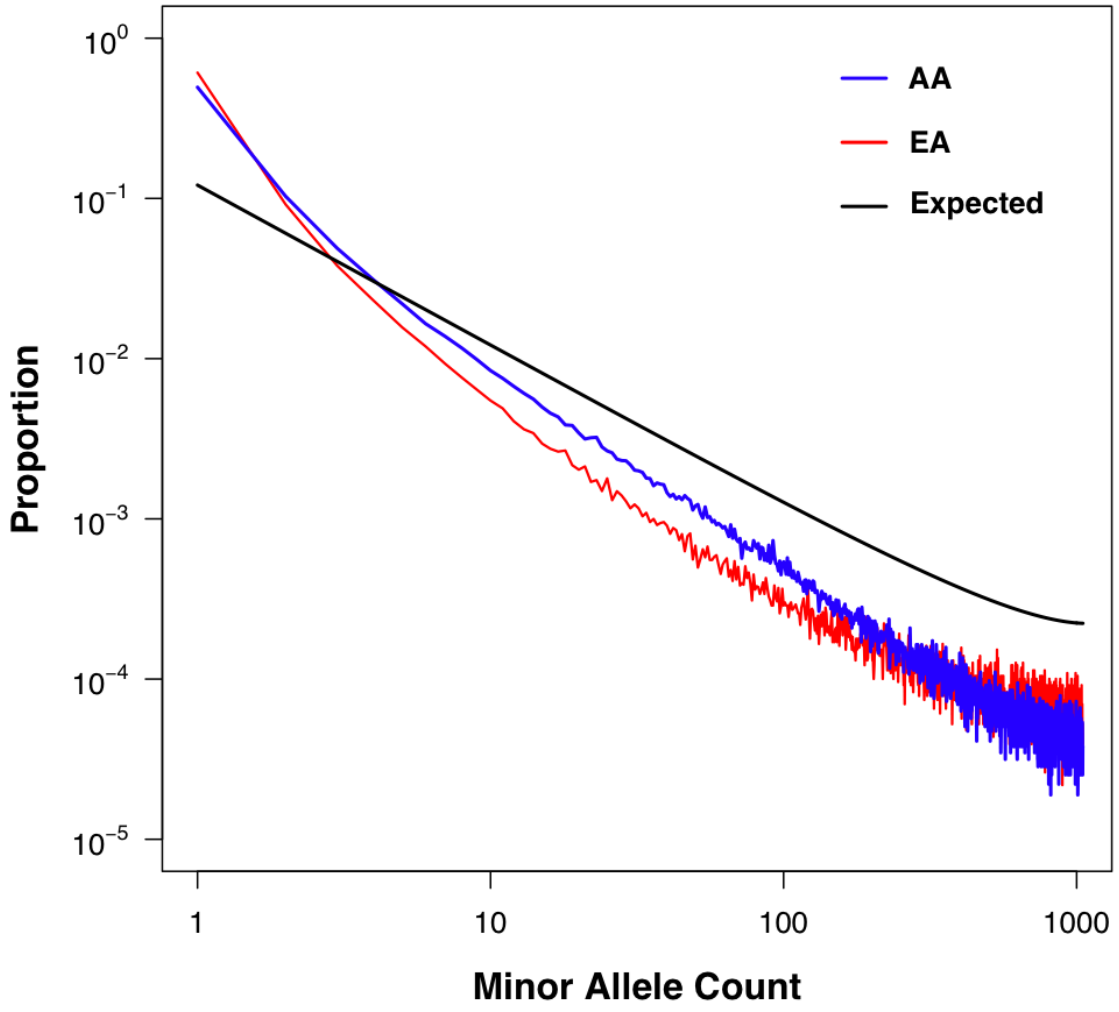
Non-independence of evolutionary forces: Drift (Selfing) and Selection



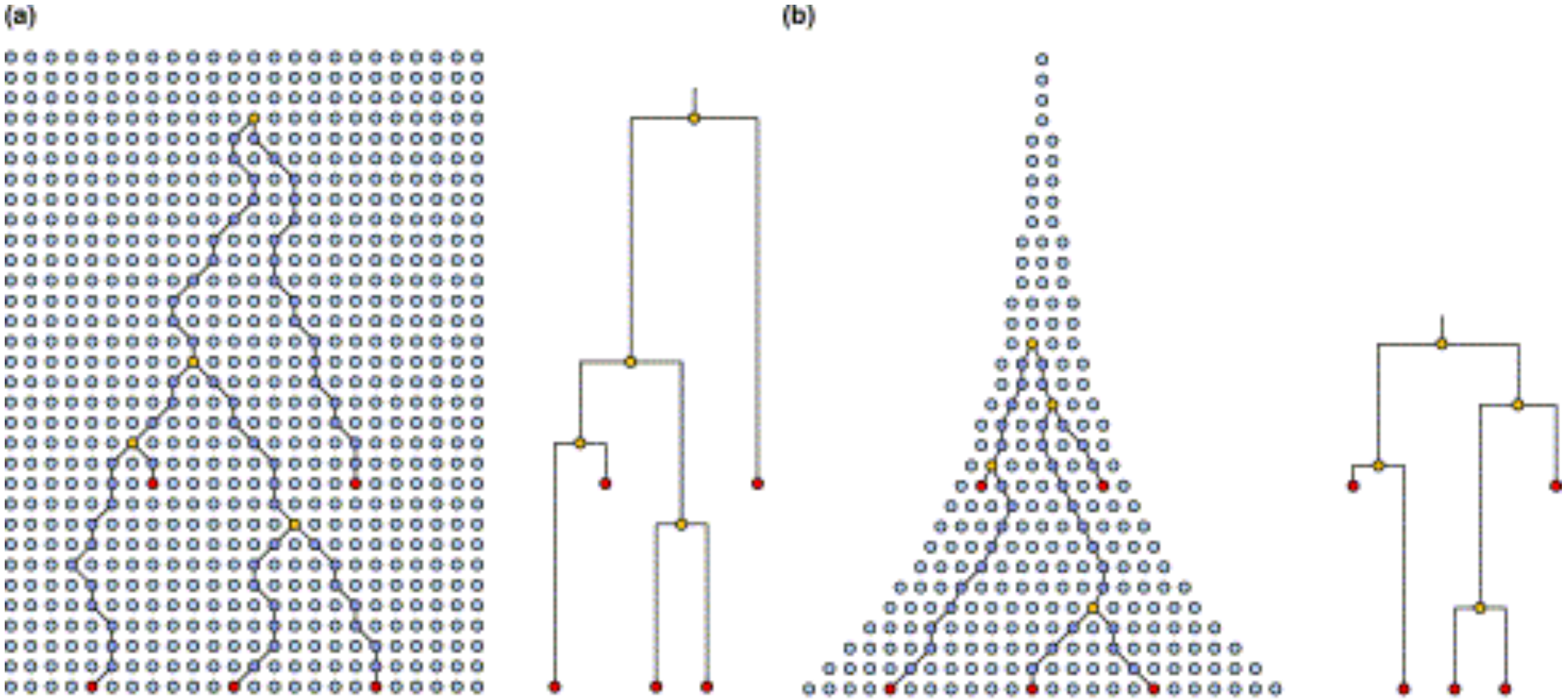
Hartfield et al. (2017) Trends in Genetics



These forces all affect the Site Frequency Spectrum (SFS)



Primer on coalescent



TREMDS in Ecology & Evolution

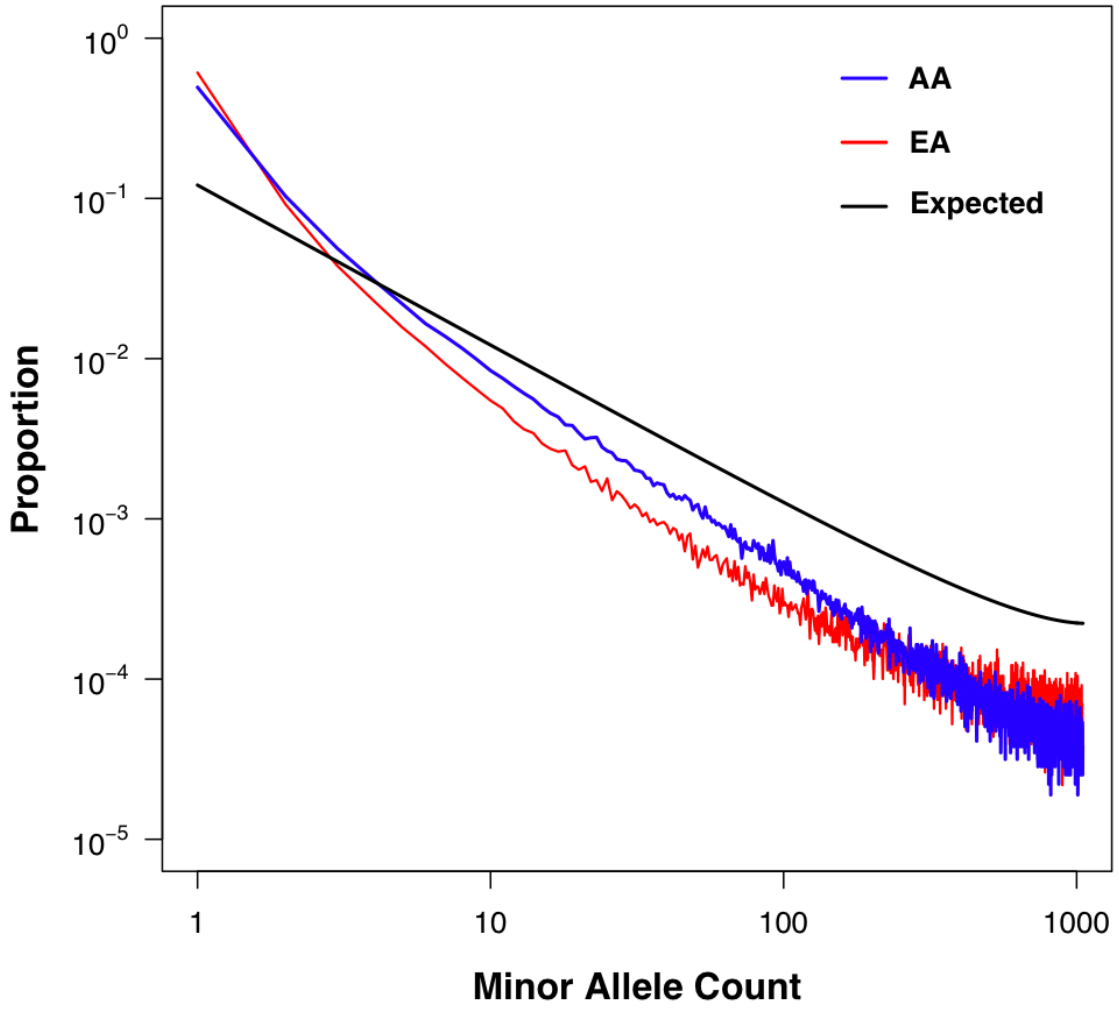
Primer on coalescent

$$E(T_i) = \frac{2}{i(i-1)} \qquad Var(T_i) = \left(\frac{2}{i(i-1)}\right)^2$$

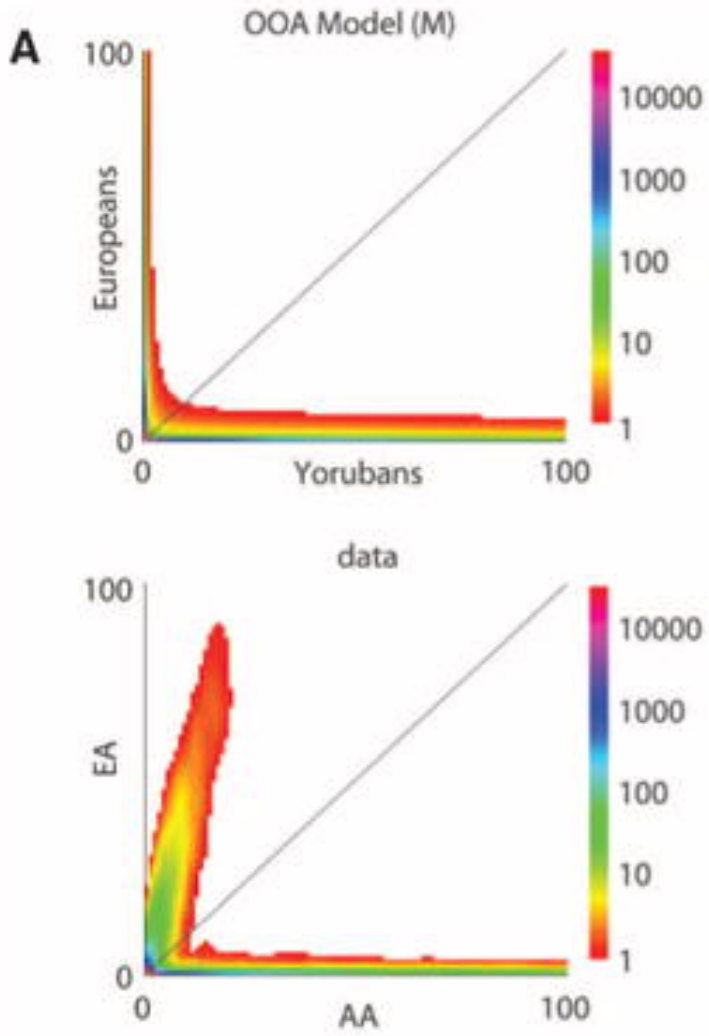
To generate a genealogy of i genes under Kingman's coalescent:

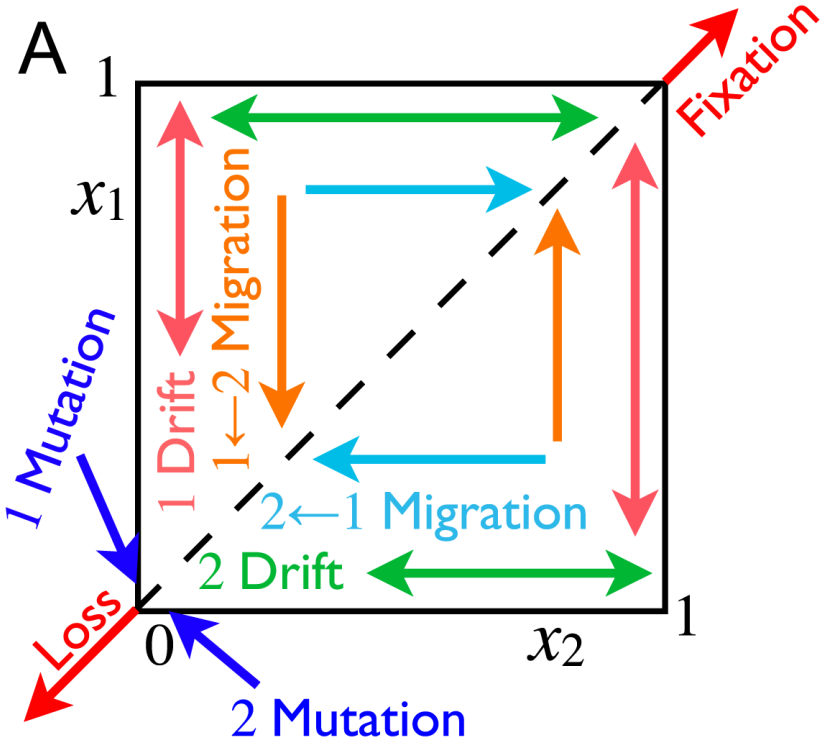
- Draw an observation from an exponential distribution with mean $\mu = 2/(i(i-1))$. This will be the time of the first coalescent event (looking from the present backwards in time).
- Pick two lineages at random to coalescence.
- Decrease i by 1.
- If $i = 1$, stop. Otherwise, repeat these steps [8, 9].

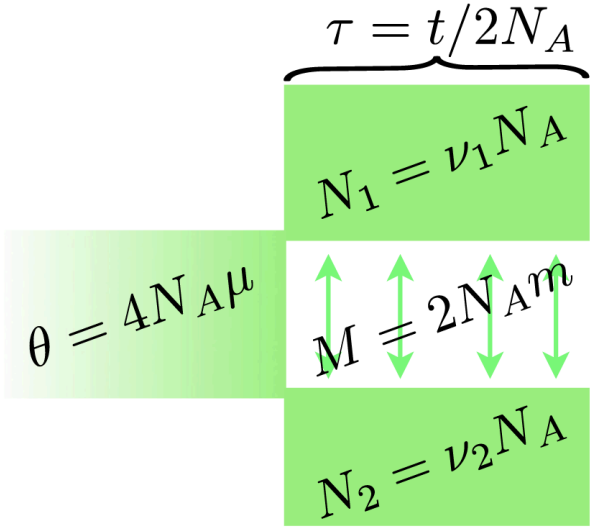
Site Frequency Spectrum (SFS)

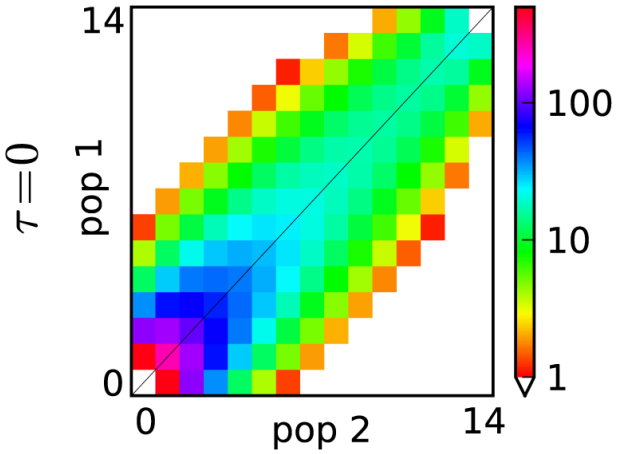


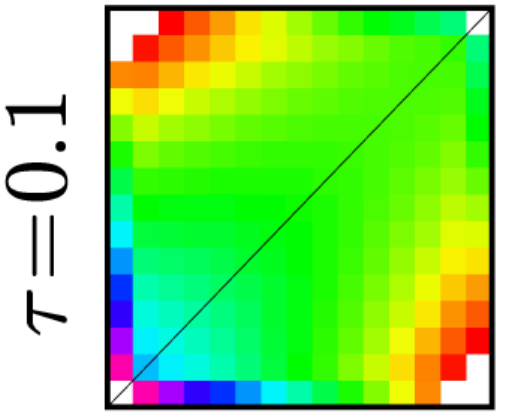
Joint Site Frequency Spectrum (JSFS)

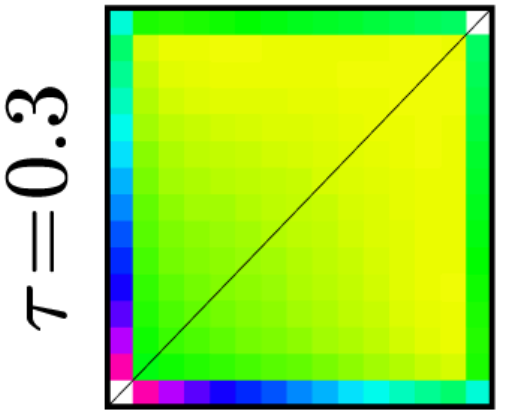


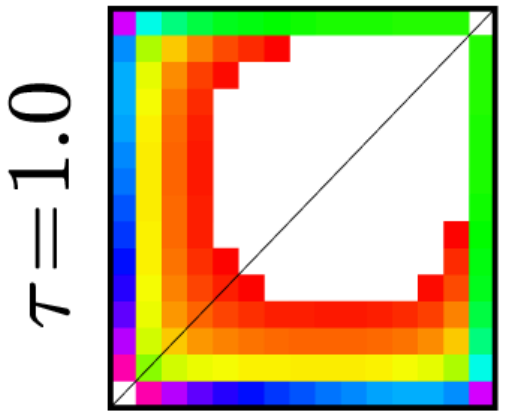












Gutenkunst et al. 2009 PLoS Genet.

Useful equations

Time: $t = T/(4*N_{ref}*Gen)$

- N_{ref} = reference or ancestral population size
- Gen = number of years per generation
- T = chronological years

$\theta = 4*N_{ref}*\mu*Length;$

- μ = mutation rate
- Length is the bp of the segment simulated (aka nsites for recombination)

Growth: $N(t) = N(0)e^{-t\alpha}$

Recombination: $\rho = 4N_{ref}r$

- r is the recombination rate between the ends of a unit length sequence

Migration: $M_{ij} = 4N_{ref}m_{ij}$

- m_{ij} is the fraction of subpopulation i that is made up of migrants from subpopulation j in forward time.

How can the SFS help us understand what happened?

- δaδi – Gutenkunst et al. (2009) – Using diffusion approximation to identify the maximum likelihood (ML) of the SFS given a demographic model.
- Moments – Jouganous et al. (2017) – Similar likelihood but uses alternative ordinary differential equation techniques to estimate model parameters making more complicated models possible.
- Approximate Bayesian Computation – Review: Csilléry et al. (2010) – A generalized framework to sidestep some of the difficulties in ML to enable the assessment of complex models by simulation.

Bayes' Rule

$$P(M|D) = \frac{P(D|M)P(M)}{P(D)}$$

P(M|D) = posterior probability of model M given data D

P(D|M) = likelihood of the data D given the model M

P(M) = prior probability of the model M

P(D) = probability of the data D

Likelihood is really hard!

$$\mathcal{L}(\Theta|S) = \prod_{i=1\ldots P} \prod_{d_i=0\ldots n_i} \frac{e^{-M[d_1,d_2,\ldots,d_P]} M[d_1,d_2,\ldots,d_P]^{S[d_1,d_2,\ldots,d_P]}}{S[d_1,d_2,\ldots,d_P]!},$$

$$M[d_1,d_2,\ldots,d_P] = \int_0^1 \cdots \int_0^1 \prod_{i=1,2,\ldots,P} \binom{n_i}{d_i} x_i^{d_i}(1-x_i)^{n_i-d_i}$$

$$\phi(x_1,x_2,\ldots,x_P)\mathrm{d}x_i.$$

Gutenkunst et al. (2009) PLoS Genet.

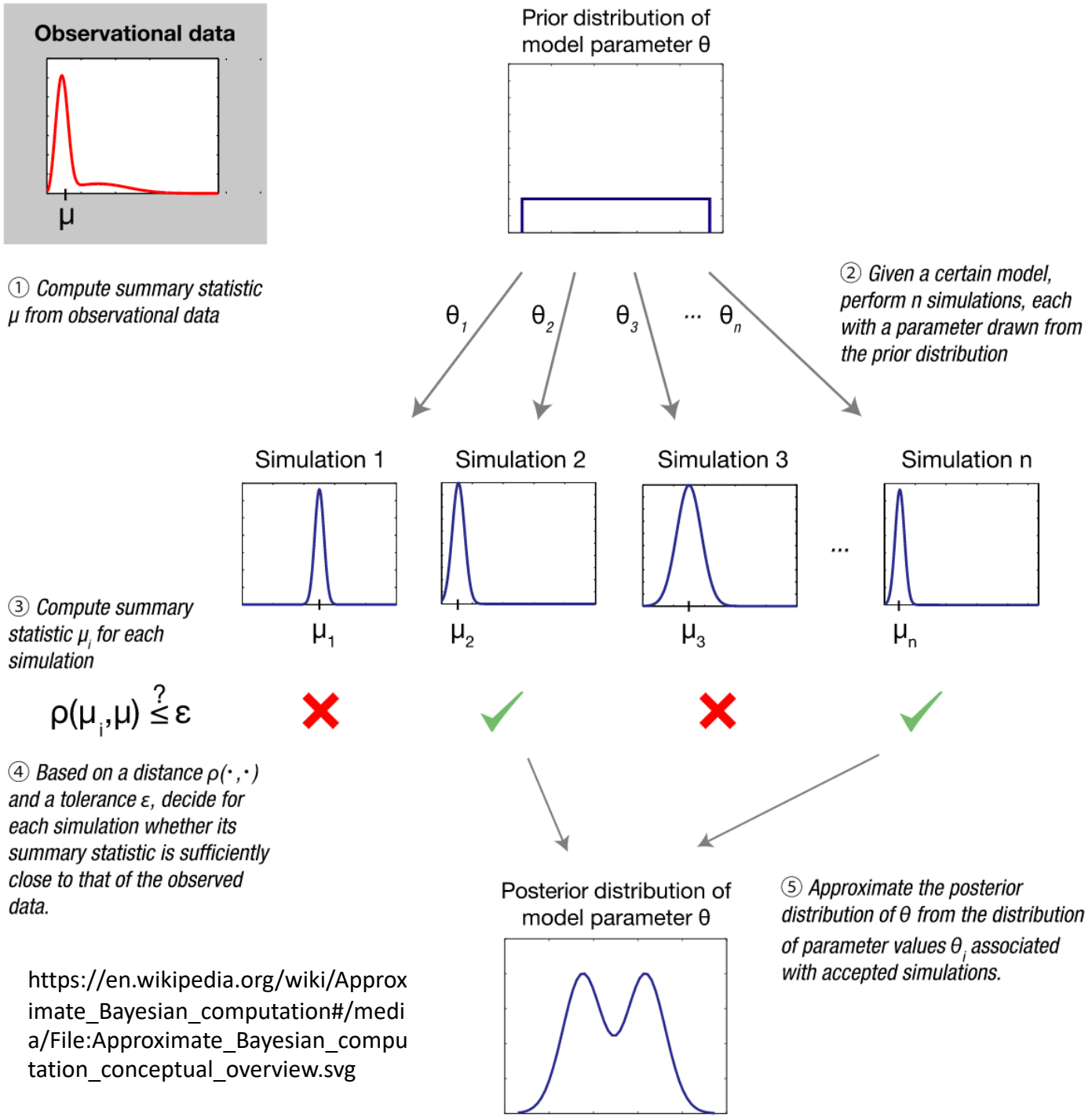
So is there a way around it with simulation?

Yes, yes there is 😄

$$\rho(\hat{D}, D) \leq \epsilon$$

$$\rho(S(\hat{D}), S(D)) \leq \epsilon$$

$$Set\ of\ j\ Simulations\ that$$
$$\min_j \sum_{i \in SFS} |SFS_{o,i} - SFS_{j,i}|$$

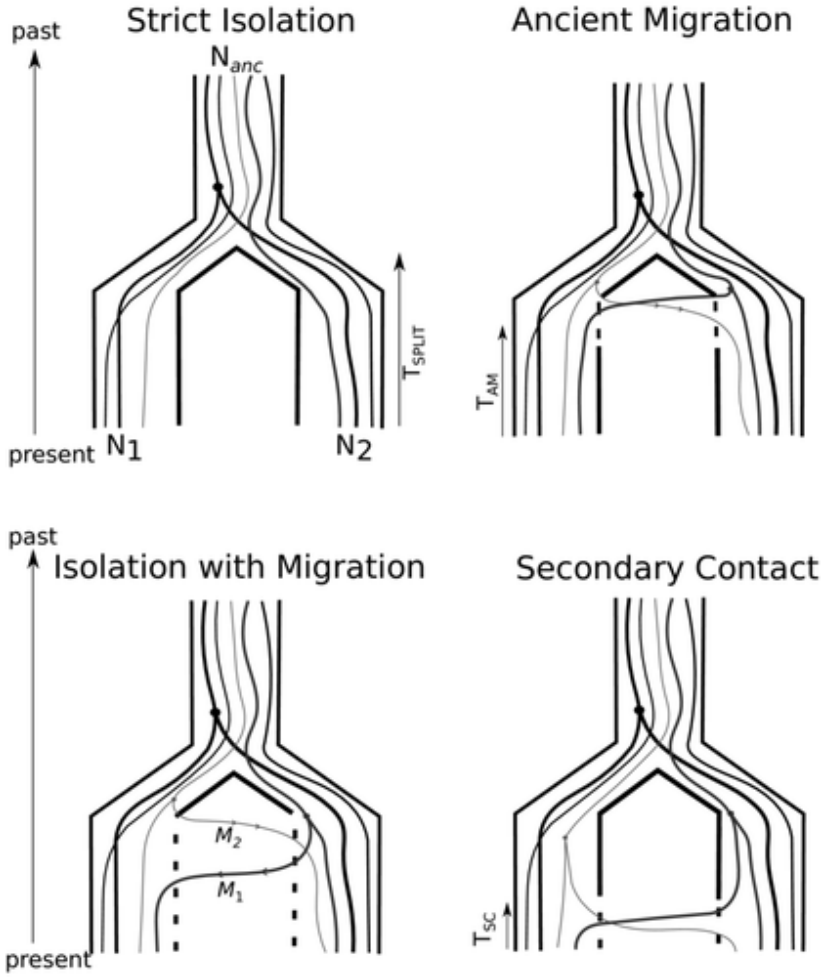


https://en.wikipedia.org/wiki/Approximate_Bayesian_computation#/media/File:Approximate_Bayesian_computation_conceptual_overview.svg

ABC in action

- Divergence models of Atlantic Salmon from North America and Eurasia
  - 2035 individuals from 77 locations
  - 5034 SNPs from a genotyping array
  - 19 summary statistics
  - 3500 best simulations (out of 14 × 1 million)

	All models			
	P(SI)	**P(AM)**	**P(IM)**	**P(SC)**
Within America	0.000	0.000	0.013	**0.984**
Between Continent	0.000	0.000	0.005	**0.993**
Within Europe	0.000	0.003	0.024	**0.967**



Rougemont & Bernatchez (2018) Evolution

Concluding Summary

- Four main evolutionary forces are: Mutation, migration, selection, and drift.
- These forces interact and rarely act independently.
- These forces change the site frequency spectrum in informative ways that we can use for both demographic analysis and simulation.