Statistical analysis for GWAS: Population structure

Alicia Martin
Postdoctoral Research Fellow

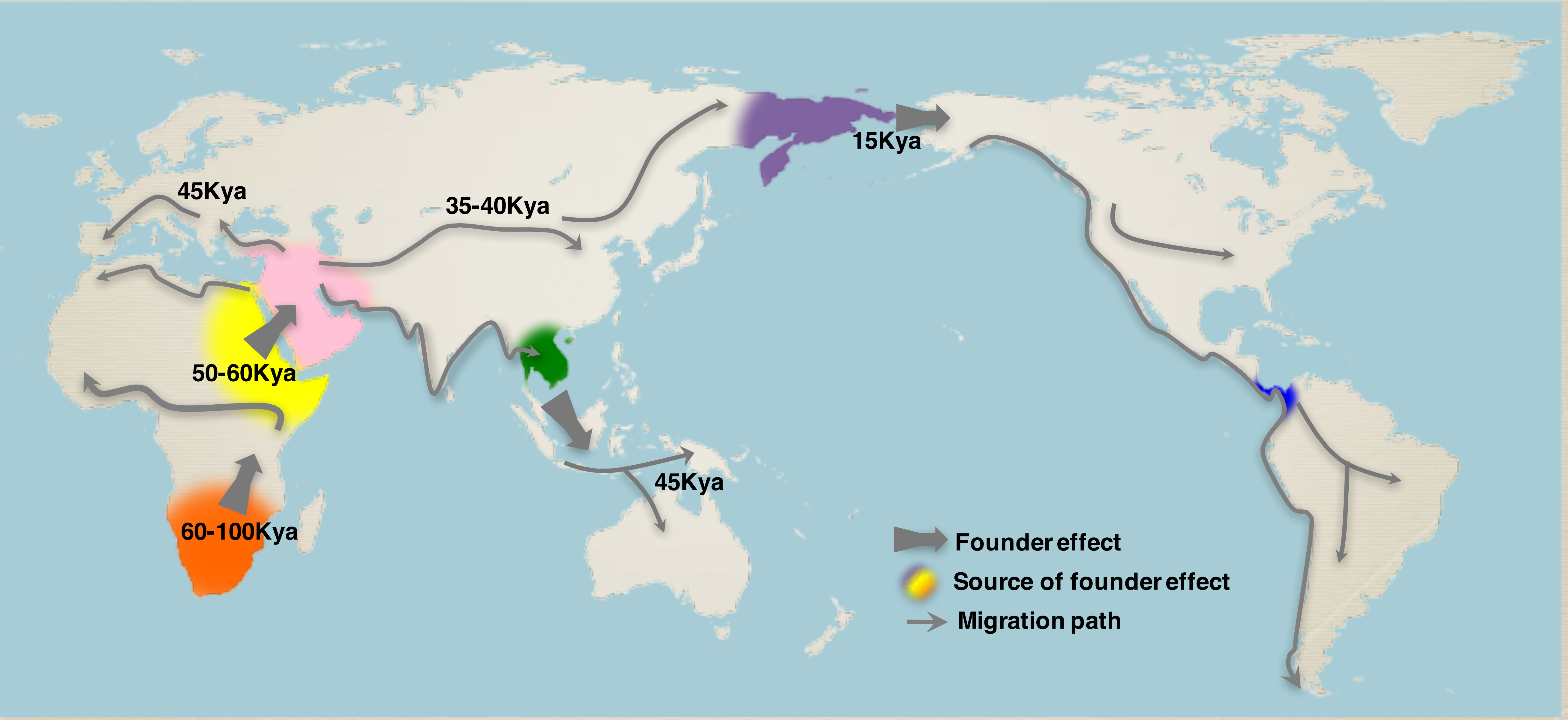
December 10, 2016

Modules

- Serial founder effects
- Basic population structure
- Hardy-Weinberg equilibrium
- How genetic structure changes
- Linkage disequilibrium
- Effective population size
- Demographic models
- African origins and population structure

Serial founder effects

Historical human migration routes



Henn, Cavalli-Sforza, and Feldman (2012) PNAS

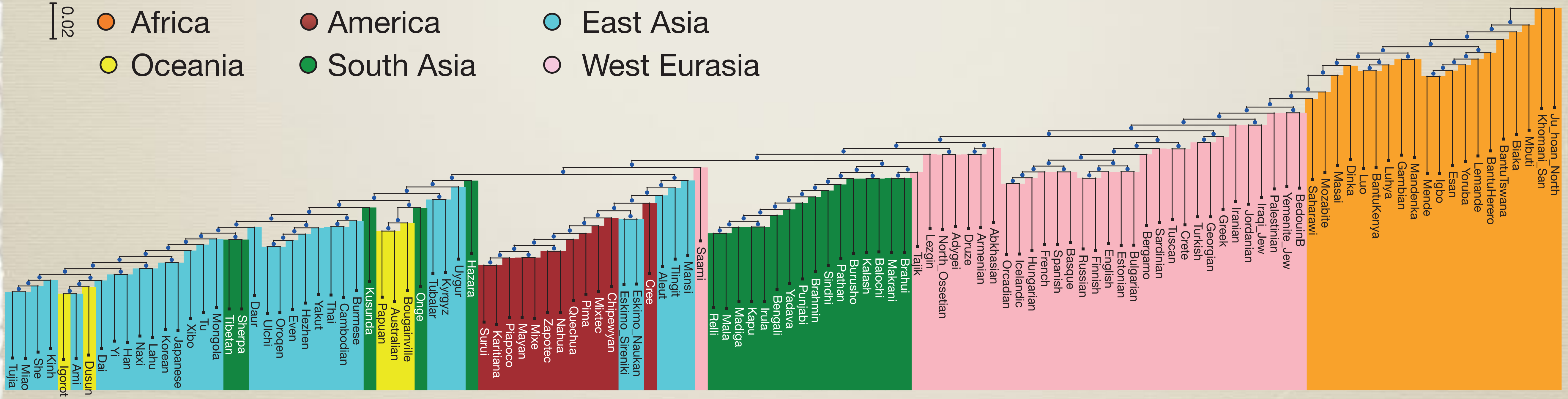
Genetic divergence across diverse human genomes

East Asians,
Americans, Oceania

South
Asians

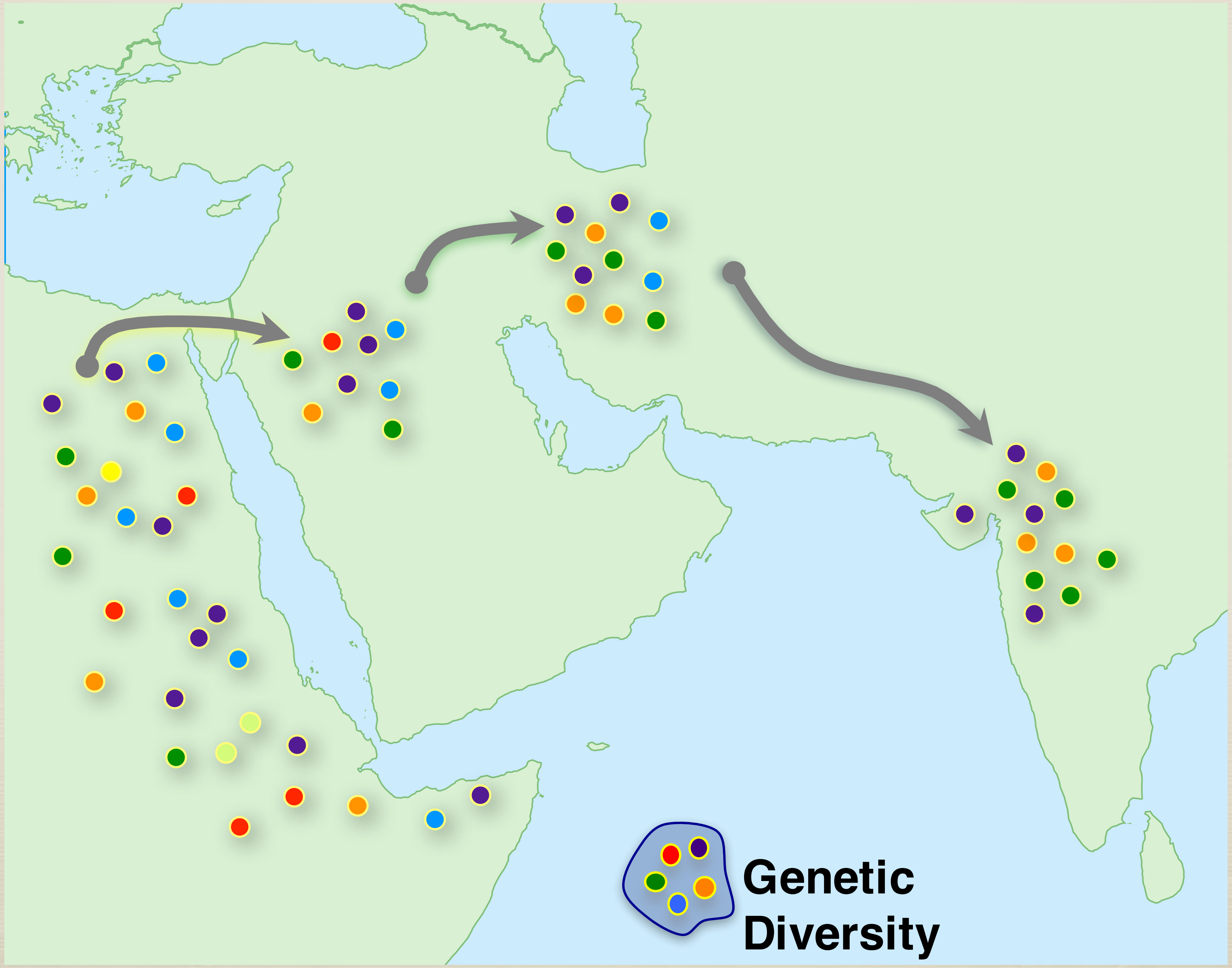
Europeans
& Near East

Africans



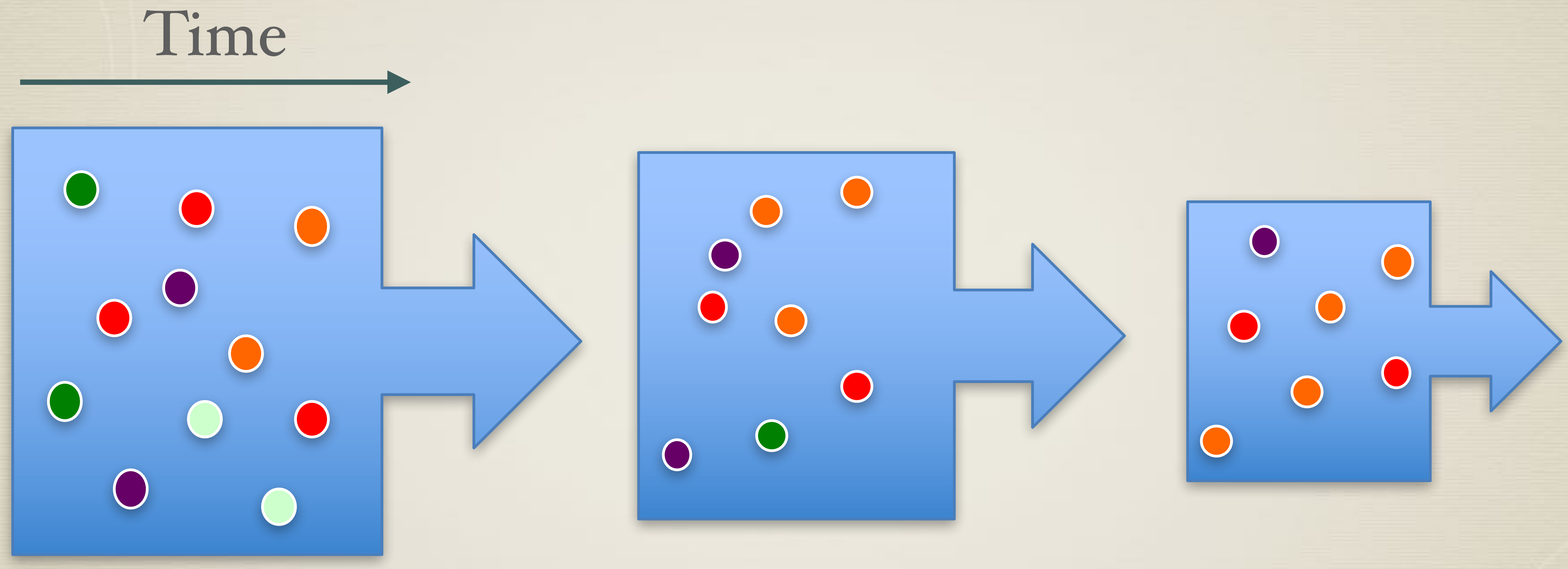
S Mallick et al. Nature 1–6 (2016) doi:10.1038/nature18964

Reduction in diversity due to serial founder effects



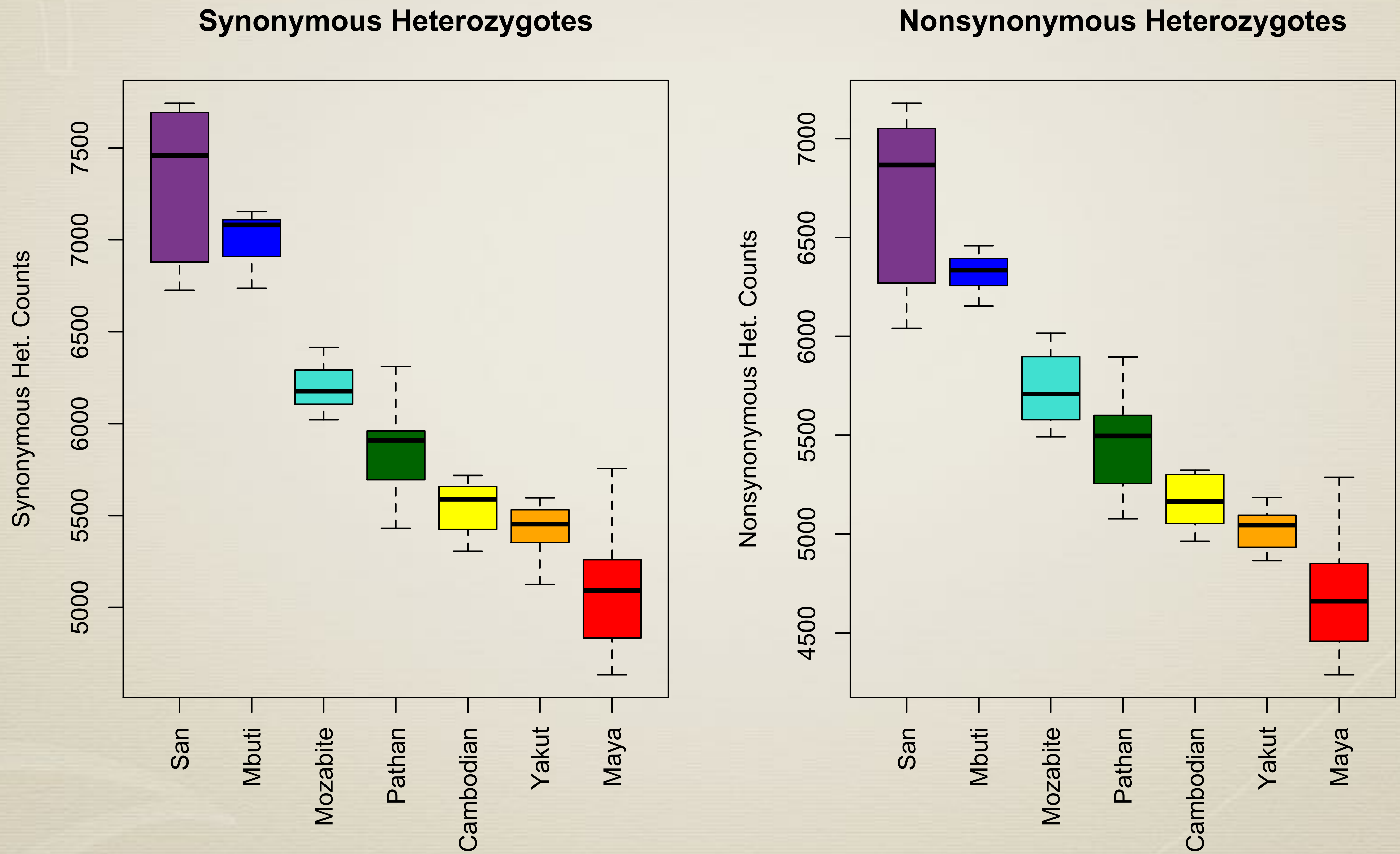
Henn, Cavalli-Sforza, and Feldman (2012) PNAS

Serial founder effect model and assumptions



- Migration after the initial founder expansion has been limited
- There has been no substantial admixture from another highly diverged population
- Post-expansion demographic fluctuations have not decreased diversity substantially

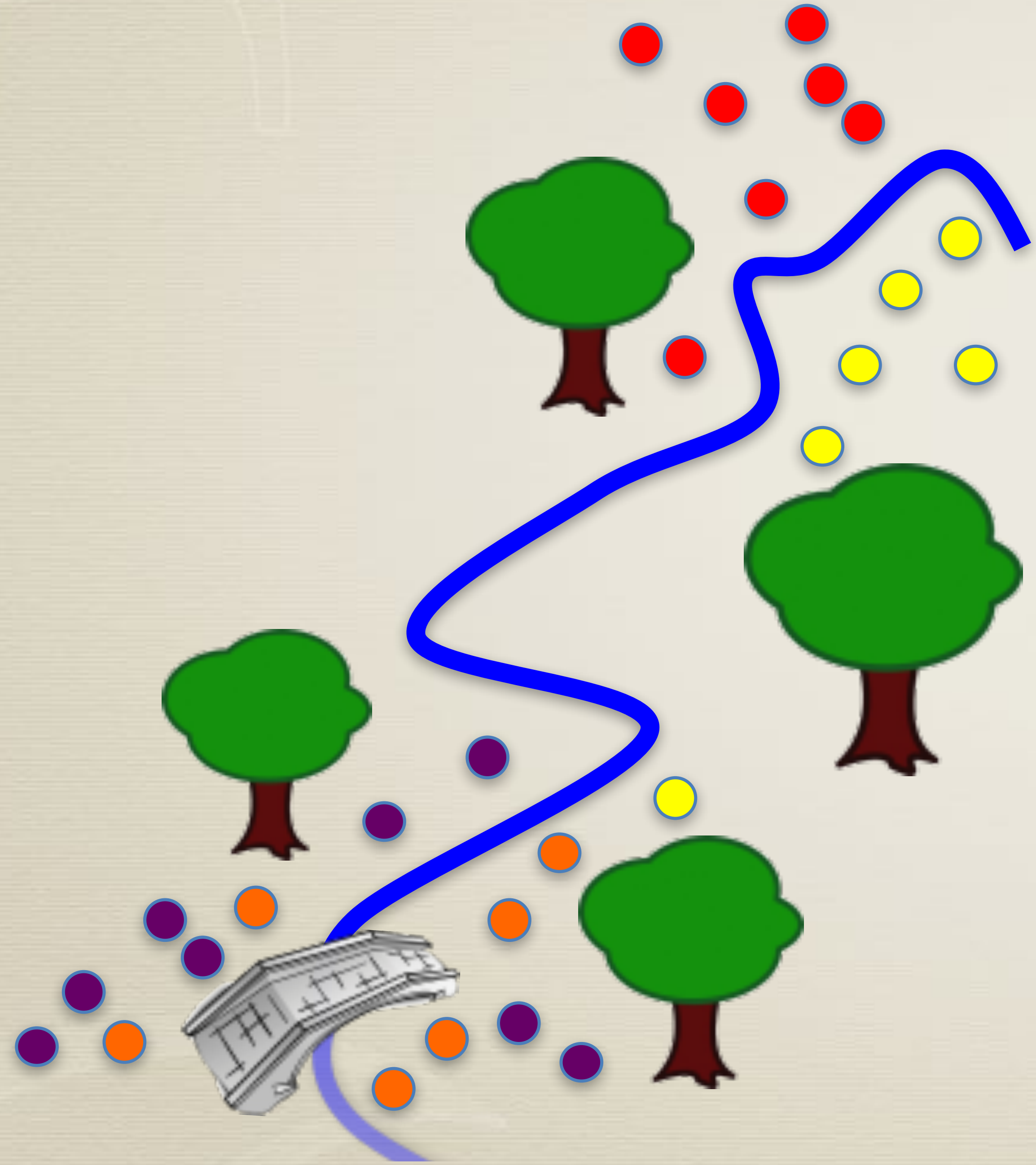
Decline in heterozygosity out-of-Africa



Henn, B.M., et al. (2016). PNAS. 113, E440-9.

Basic population structure

What is population structure?

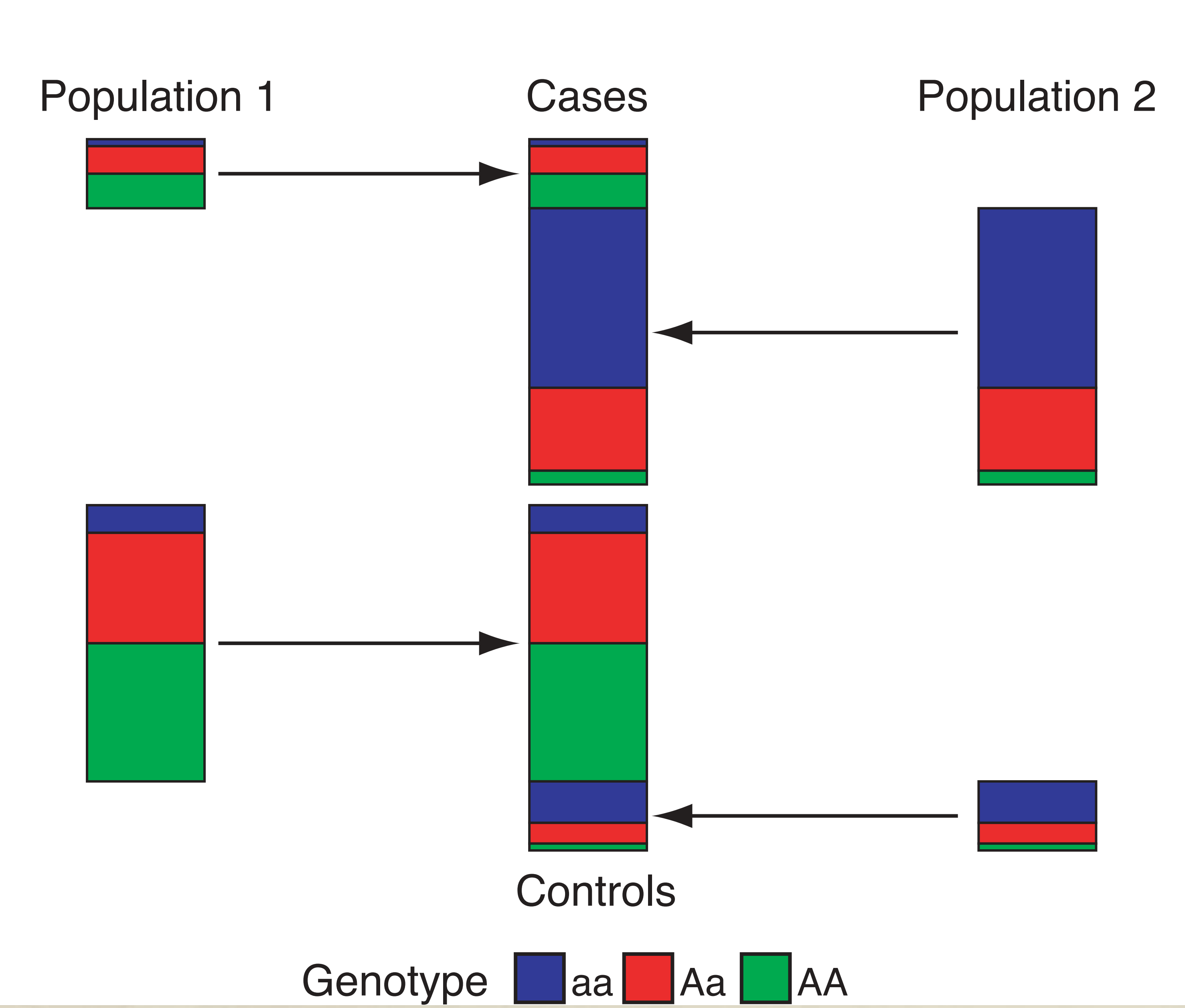


- Can be caused by multiple barriers to random mating: geography, language, ancestry

- Random mating is an important assumption in pop gen and stat gen models, usually assess population structure first

- Two commonly used methods of detecting structure are allele frequency-based clustering algorithms and principle component analysis

How does population stratification affect association analyses?

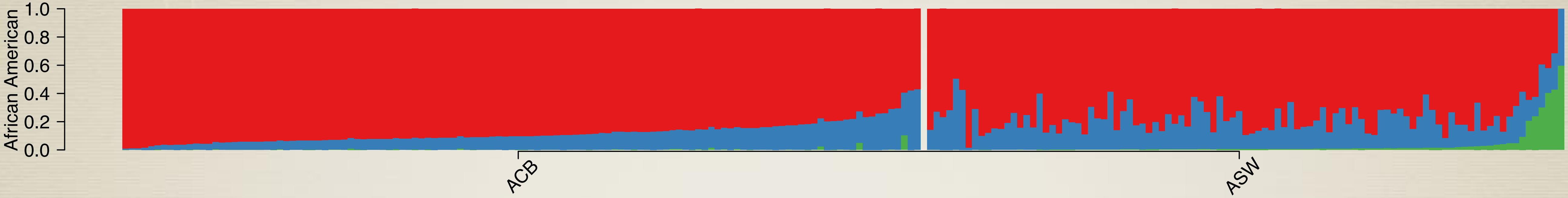


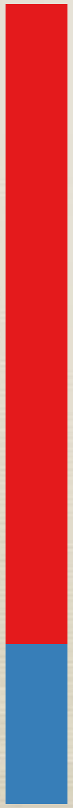
Disease more common in Population 2

- oversampling cases from this population relative to controls
- any allele that is more common in Pop 2 appears associated with the disease

Marchini et al., Nat Genet 2004

Population structure with clustering algorithms





I'm 80% red and 20% blue!

Each bar represents 1 individual. The number of colors is the number of potential ancestries. Proportion of different colors is the proportion of different ancestries for that individual

Continental ancestry

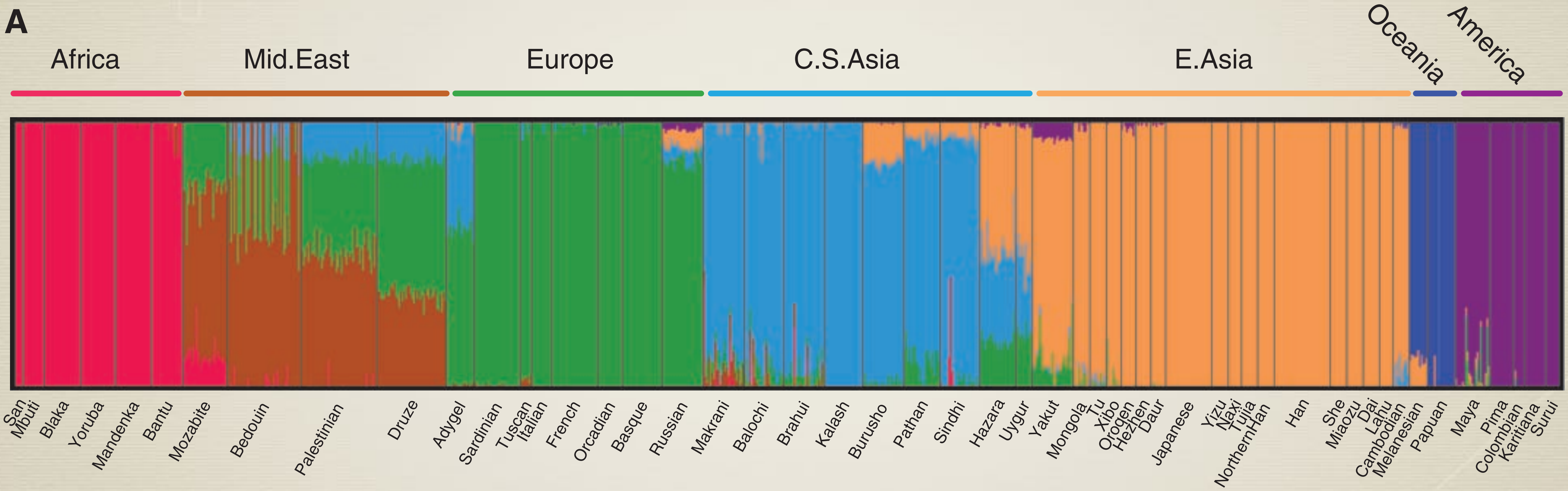
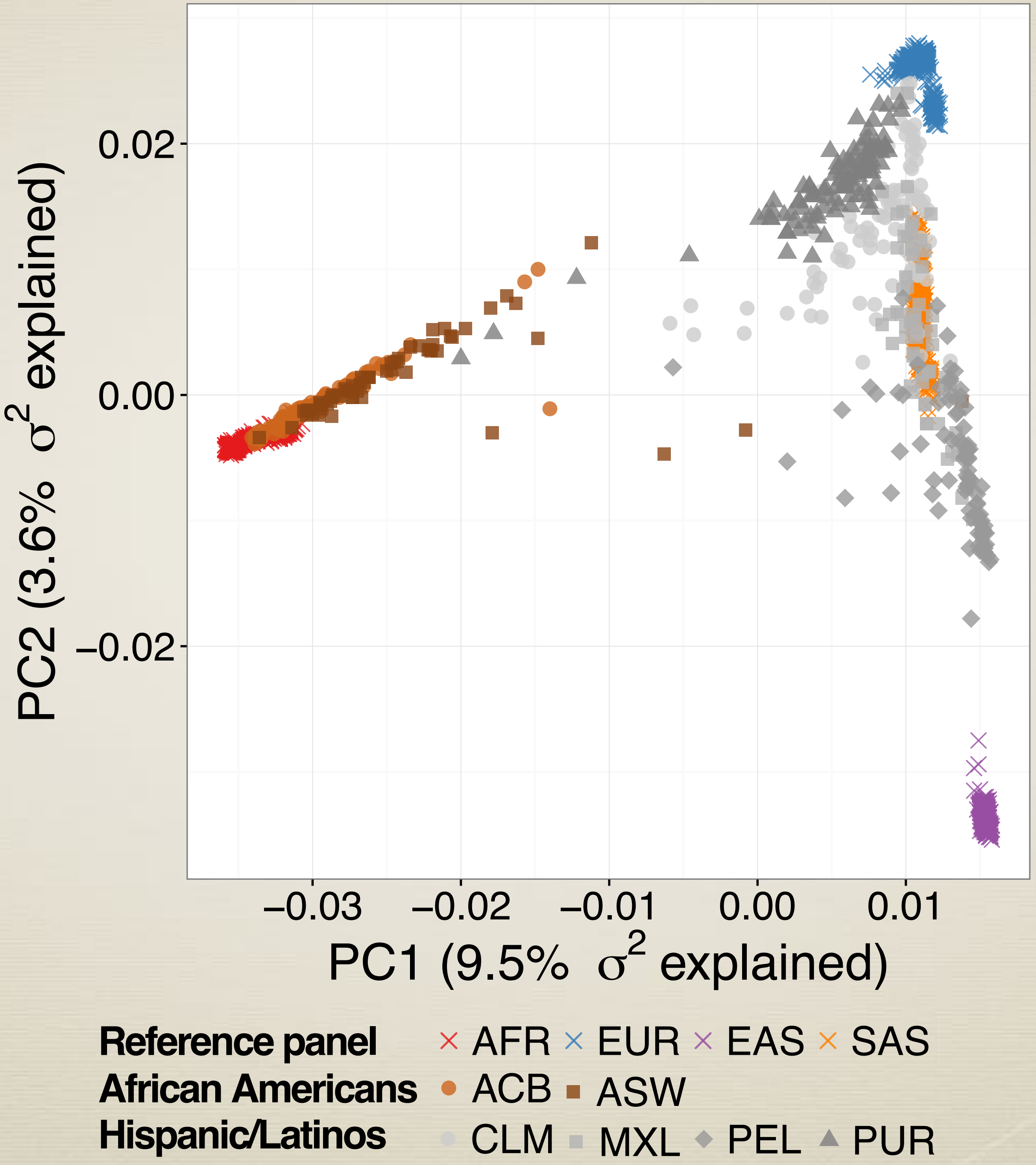


Fig. 1. Individual ancestry and population dendrogram. **(A)** Regional ancestry inferred with the *frappe* program at *K* = 7 (*13*) and plotted with the Distruct program (*31*). Each individual is represented by a vertical line partitioned into colored segments whose lengths correspond to his/her ancestry coefficients in up to seven inferred ancestral groups. Population labels were added only after each individual's ancestry had been estimated; they were used to order the samples in plotting.

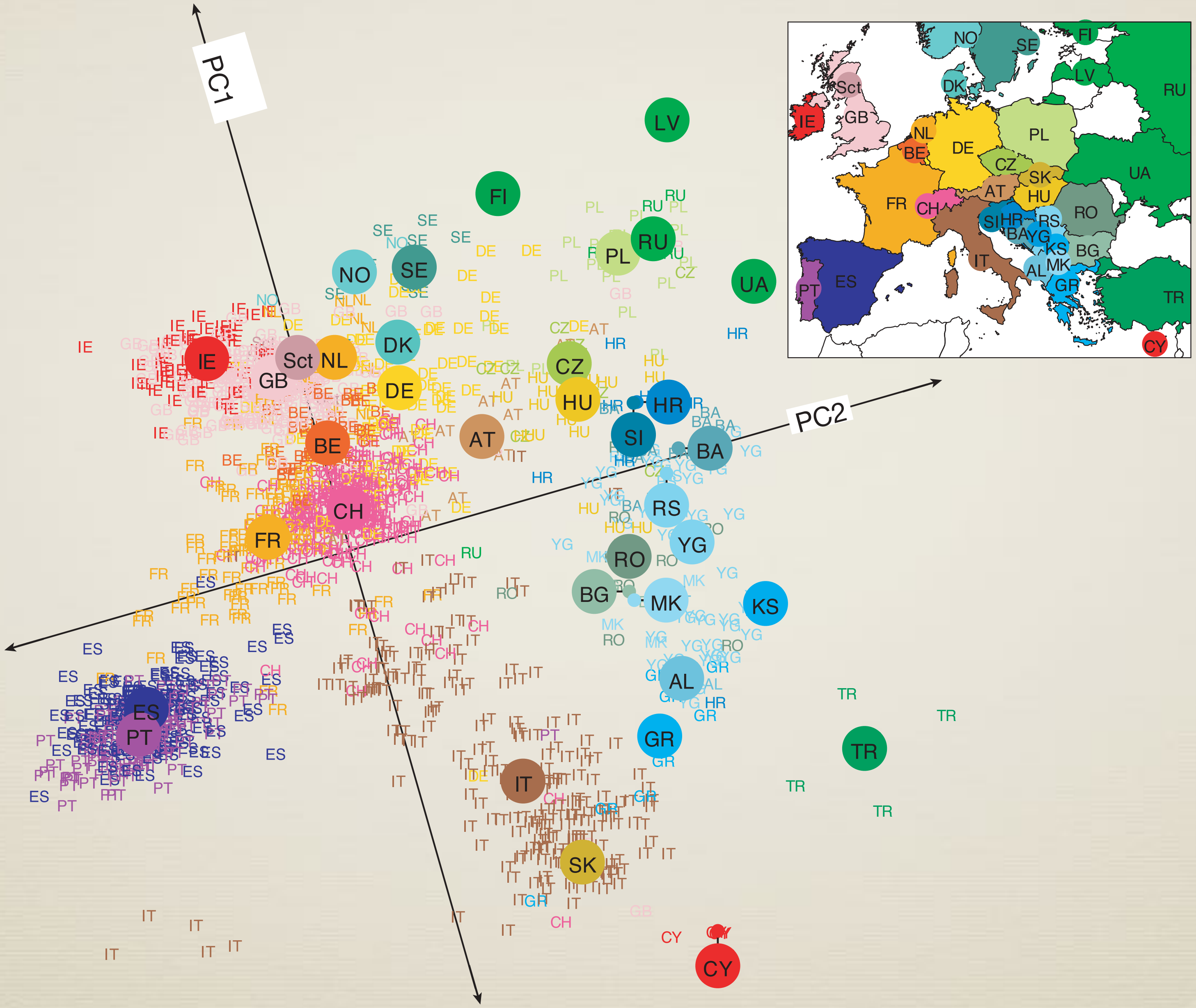
Li, J.Z., et al. (2008). Science 319, 1100–1104.

Global PCA



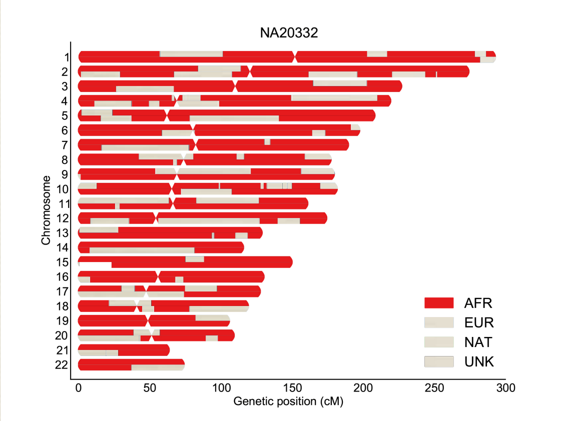
Martin, A.R., et al. bioRxiv. http://dx.doi.org/10.1101/070797

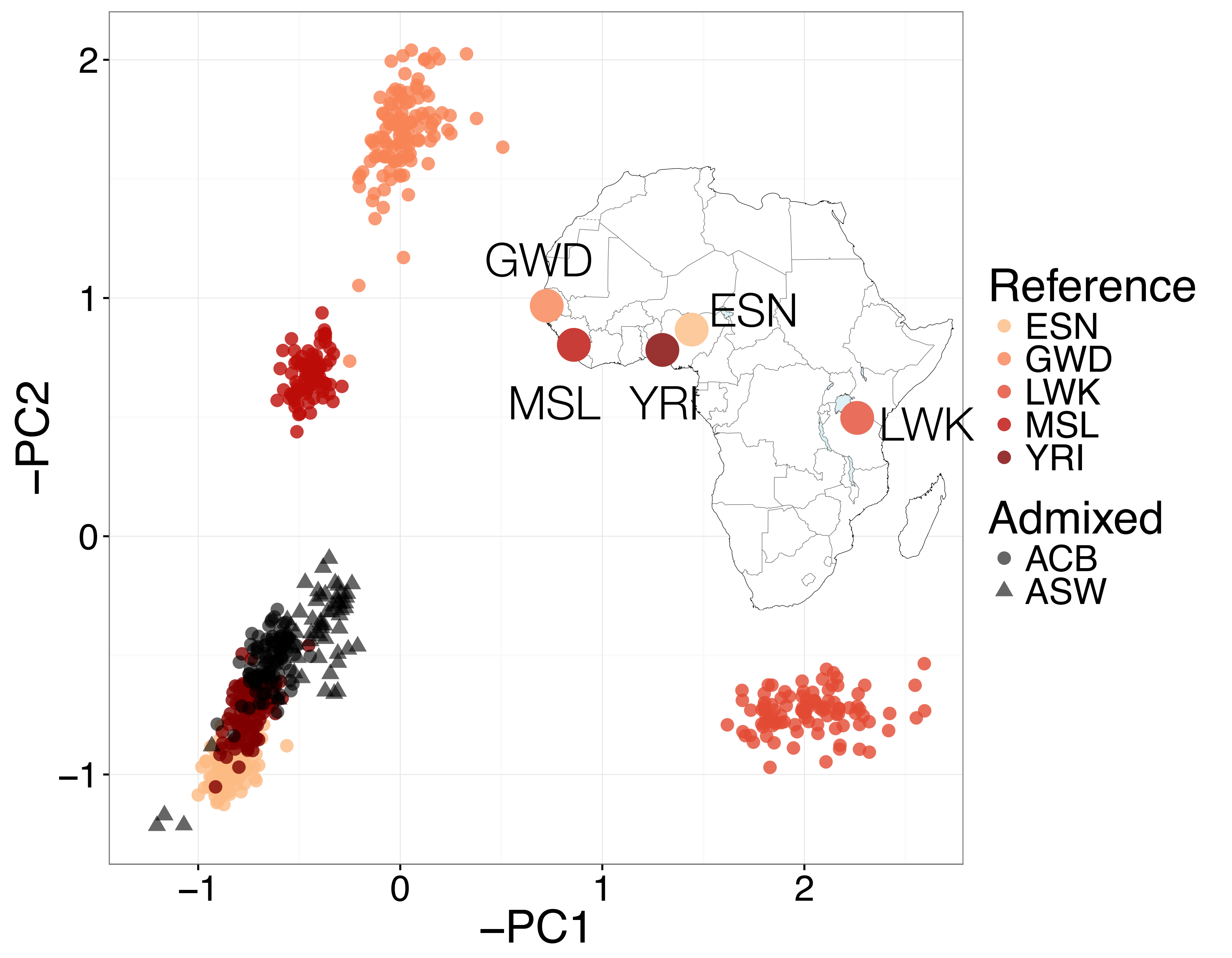
Genes mirror geography



Novembre, J., et al. (2008). Nature 456, 98–101.

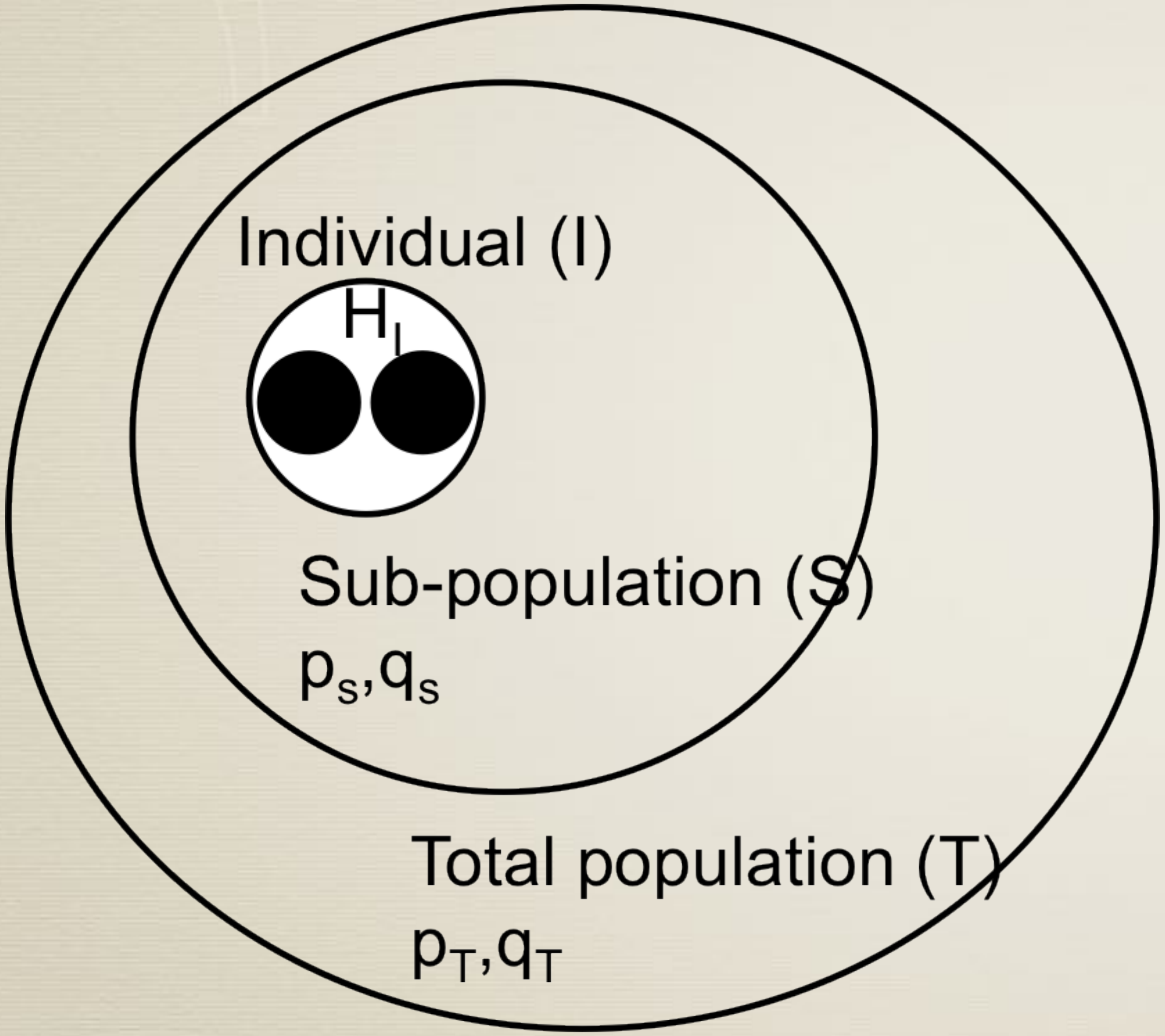
Ancestry-specific PCA provides insights into admixture origins





Martin, A.R., et al. bioRxiv. http://dx.doi.org/10.1101/070797

Fixation index (F_{ST})



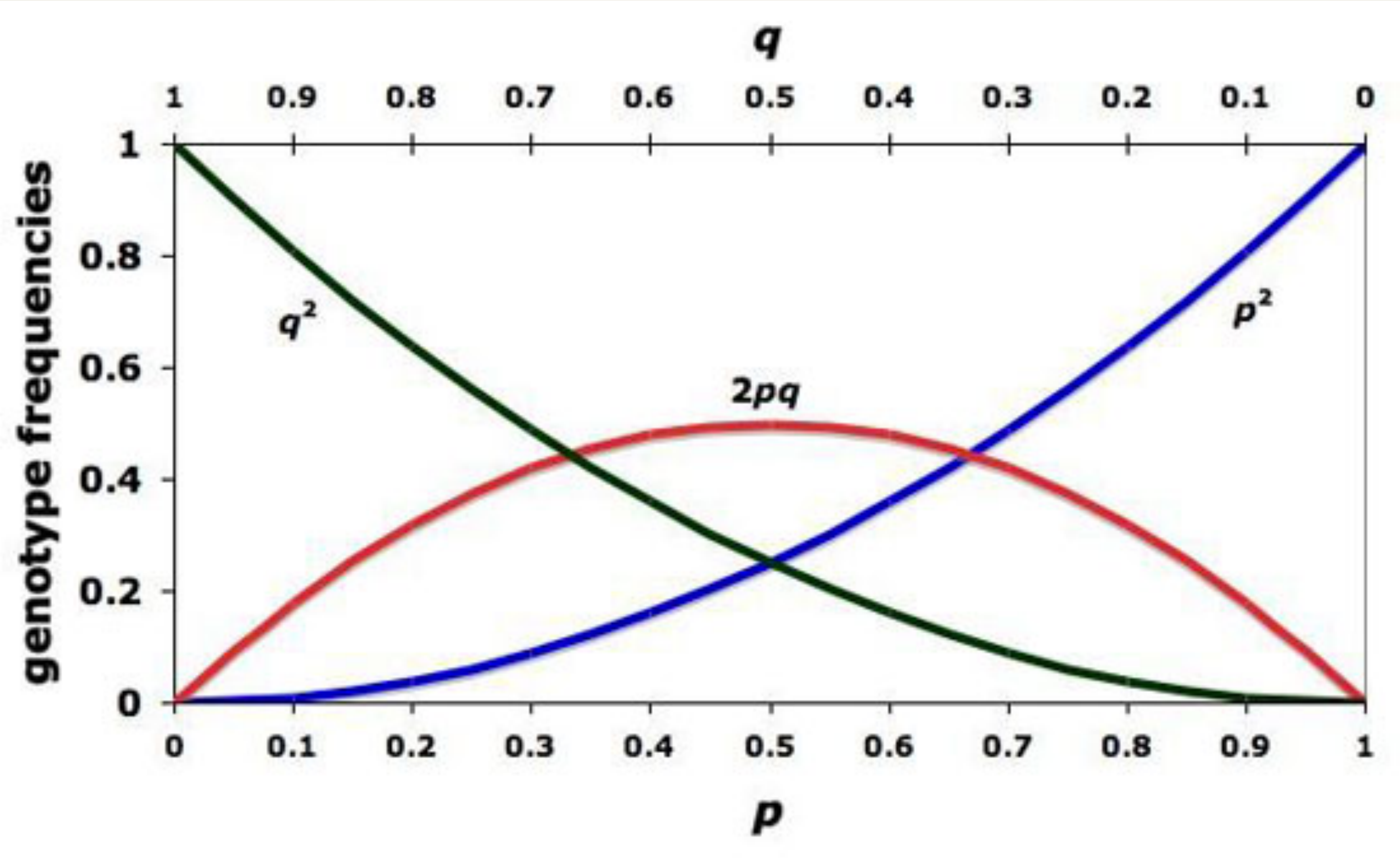
- Measures divergence across population pairs (S = subpopulations, T = total)
- H = heterozygosity

$$F_{ST} = 1 - \frac{H_S}{H_T} = 1 - \frac{2p_Sq_S}{2p_Tq_T}$$

Graham Coop's pop gen notes: http://bit.ly/2fEXzUe

Hardy-Weinberg Equilibrium

The **Hardy–Weinberg equilibrium** model states that allele and genotype frequencies in a population will remain constant from generation to generation in the absence of other evolutionary influences.



Parental allele frequencies

		Mom	
		A (p)	a (q)
Dad	A (p)	AA (p^2)	Aa (pq)
	a (q)	Aa (pq)	aa (q^2)

$p =$ frequency of A allele

$q =$ frequency of a allele

$P =$ frequency of AA genotype

$H =$ frequency of Aa genotype

$Q =$ frequency of aa genotype

Hardy-Weinberg equilibrium

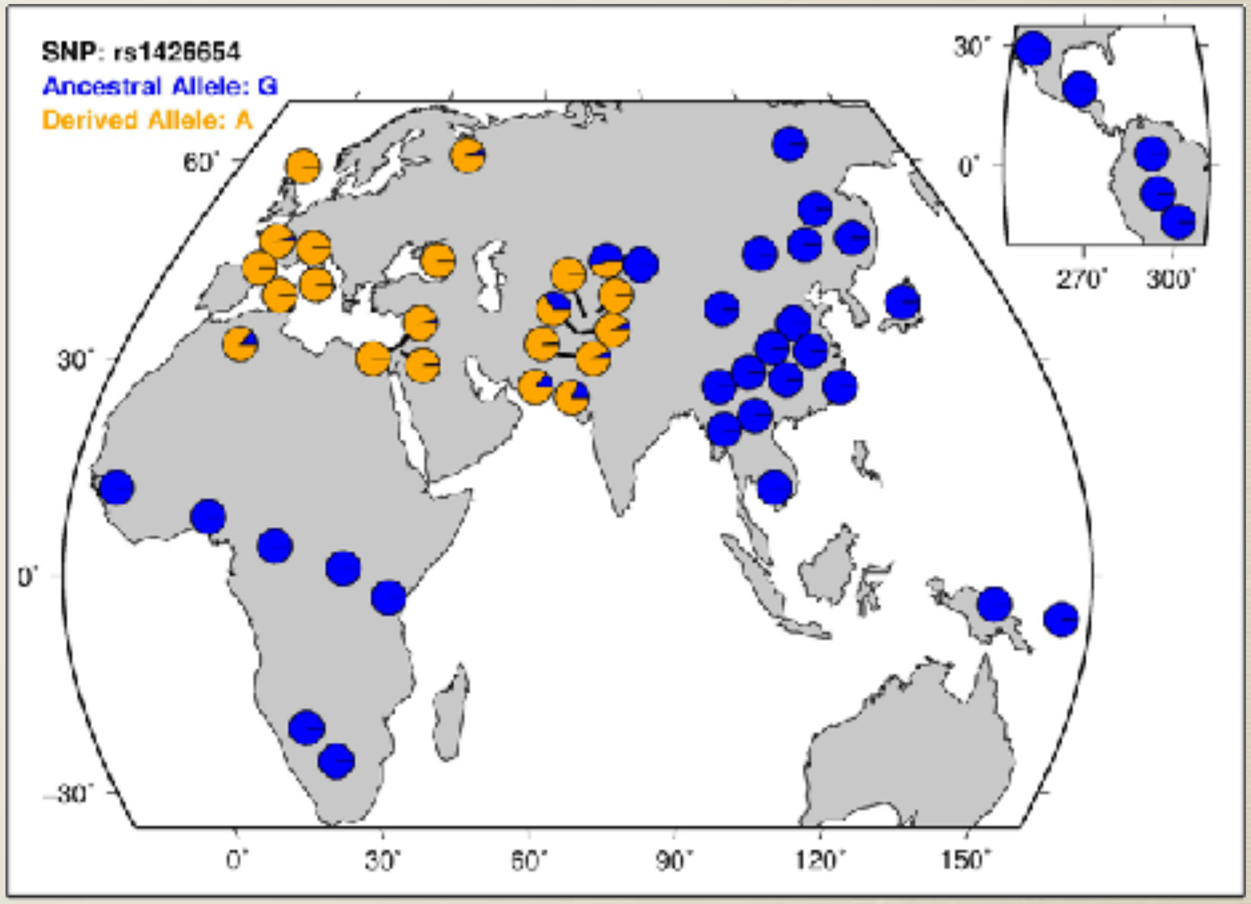
		Frequency of progeny		
Mating	Frequency (parents)	AA	Aa	aa
AA x AA	P^2	P^2		
AA x Aa	2PH	PH	PH	
AA x aa	2PQ		2PQ	
Aa x Aa	H^2	$H^2/4$	$H^2/2$	$H^2/4$
Aa x aa	2HQ		HQ	HQ
aa x aa	Q^2			Q^2
	$(P + H + Q)^2$	$(P+H/2)^2$	$2(P+H/2)^{2}$* $(Q+H/2)$	$(Q+H/2)^2$
	1	p^2	2pq	q^2

Hardy-Weinberg: assumptions and violations

Assumptions

✓ organisms are diploid
✓ only sexual reproduction occurs
? generations are non overlapping
? mating is random
? population size is infinitely large
? allele frequencies are equal in the sexes
? there is no migration, mutation or selection

SLC24A5 - skin color



Implications:

- Allele frequencies are constant, genetic diversity preserved
- HWE attained in just 1 generation of random mating

HWE in a realistic cohort

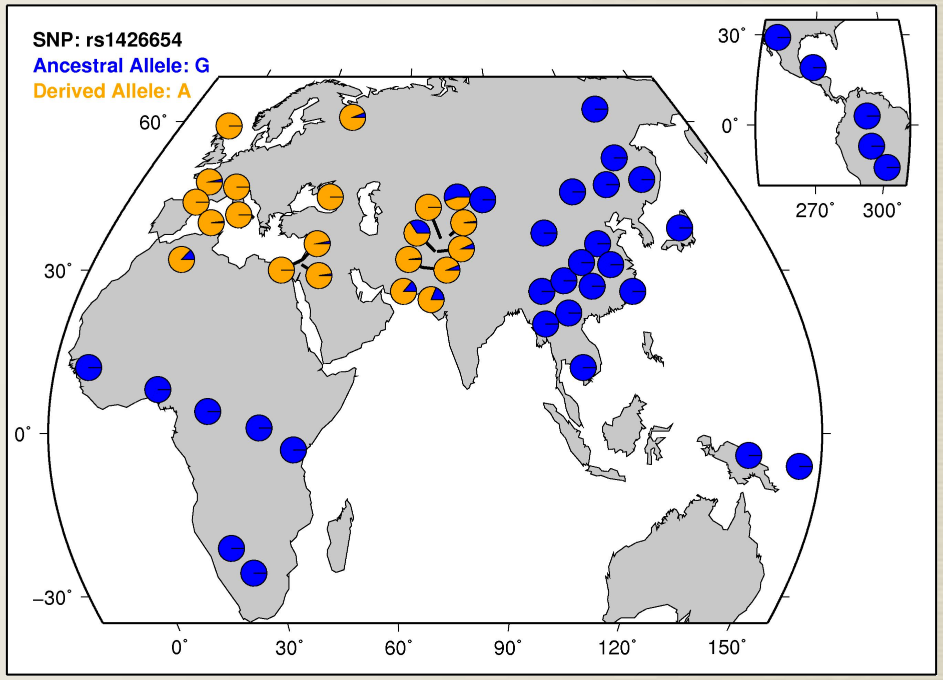
- Tennessen et al (ESP): 2439 individuals
  - 1351 Europeans, 1088 African Americans (80%, 20%)

$$P(\text{derived}|\text{European}) = 1$$

$$P(\text{derived}|\text{African American}) = 0.2$$

$$p_{cohort} = \frac{1*1351+0.2*1088}{2439} = 0.643$$

$$q_{cohort} = 1 - p_{cohort} = 0.357$$



allele	Observed	Expected
DD	$1^2 * 1351 + .2^2 * 1088 = 1395$	$2439 * 0.643^2 = 1088$
AD	$2 * .2 * .8 * 1088 = 348$	$2439 * 2 * 0.643 * .357 = 1120$
AA	$.8^2 * 1088 = 696$	$2439 * 0.357^2 = 311$

$\chi^2 = 615.08$ $P < 2.2\text{e-}16$

How genetic structure changes

How does population structure change?

Changes in allele frequencies through time

- mutation
- migration
- natural selection
- genetic drift
- non-random mating

How does population structure change?

Changes in allele frequencies through time

- mutation
- migration
- natural selection
- genetic drift
- non-random mating

spontaneous change in DNA

Human mutation rate:
~1.2 x 10^{-8} / bp

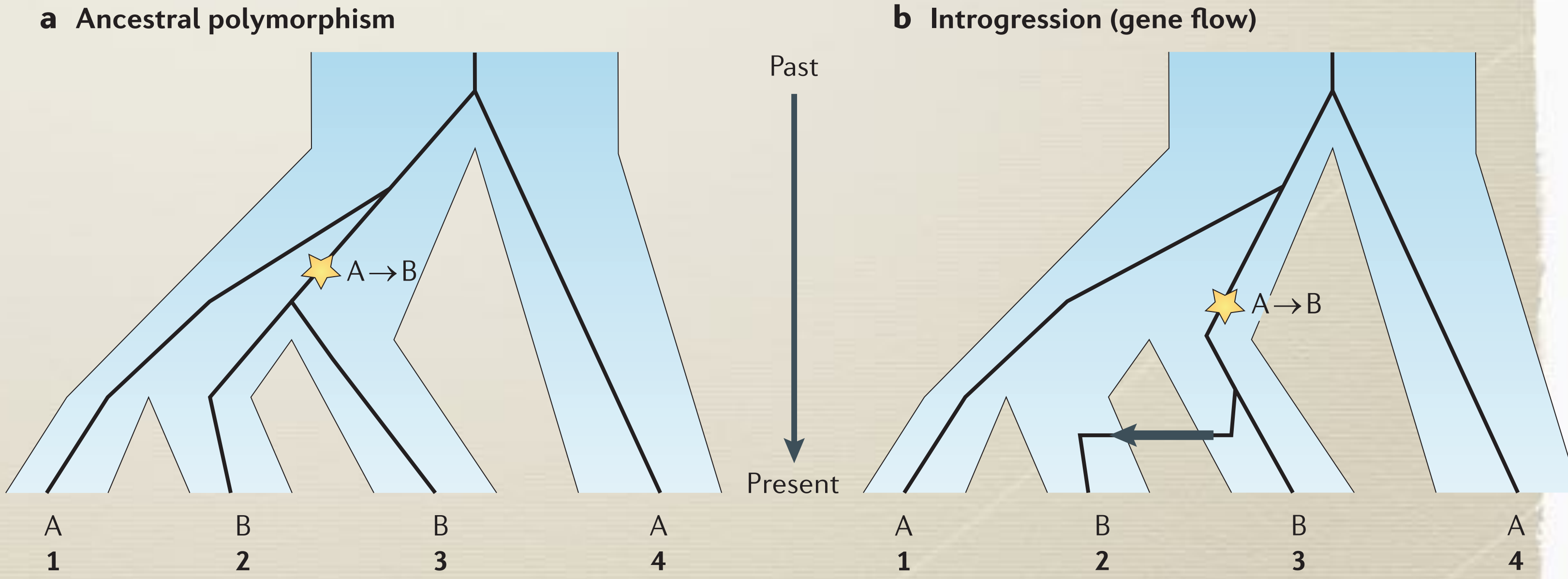
- ~80-100 total *de novo* variants
- <1 *de novo* coding variant

How does population structure change?

Changes in allele frequencies through time

- mutation
- migration
- natural selection
- genetic drift
- non-random mating

individuals moves into population, introduce new alleles (“gene flow”)



Sousa, V., and Hey, J. (2013). Nat. Rev. Genet. 14, 404–414.

How does population structure change?

Changes in allele frequencies through time

- mutation
- migration
- natural selection
- genetic drift
- non-random mating

certain genotypes produce more/less offspring

differences in survival and reproduction → differences in "fitness"

Many kinds: balancing (e.g. sickle-cell), positive (e.g. height), negative (most common), etc

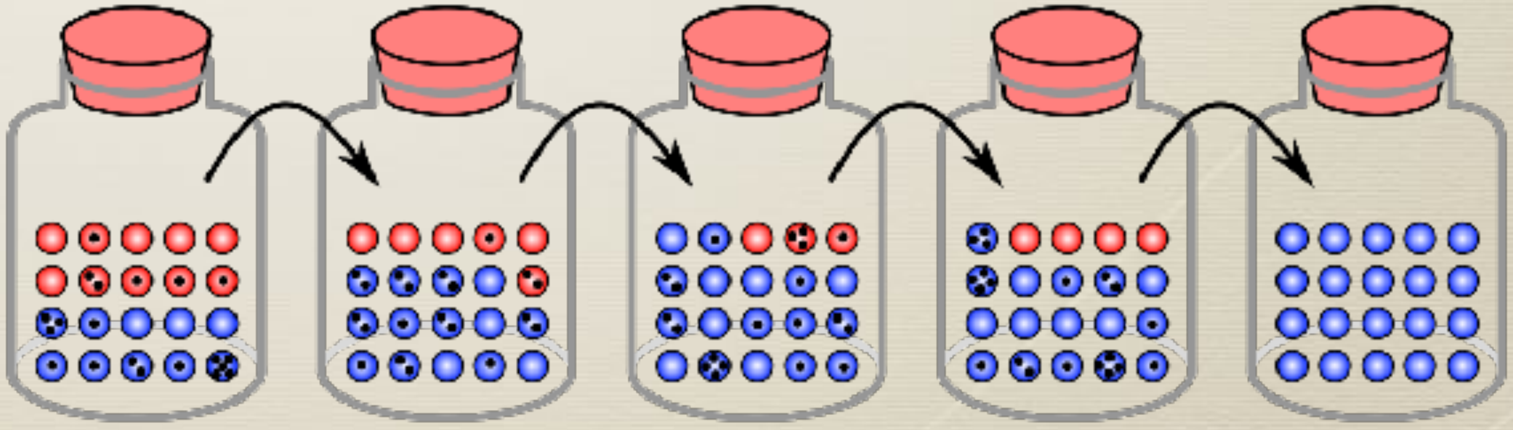
How does population structure change?

Changes in allele frequencies through time

- mutation
- migration
- natural selection
- genetic drift
- non-random mating

genetic change by chance alone

occurs in small populations



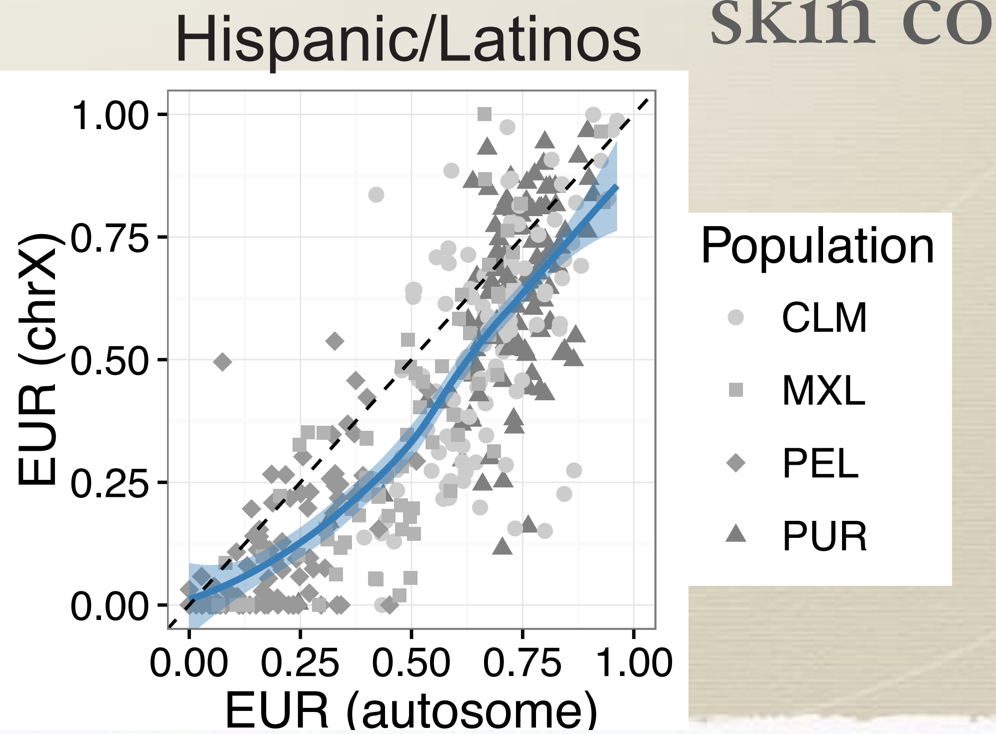
How does population structure change?

Changes in allele frequencies through time

- mutation
- migration
- natural selection
- genetic drift
- non-random mating

assortative mating: mate with similar type

Examples: education, height, skin color



Linkage disequilibrium

Linkage disequilibrium is the non-random association of alleles at different loci. Loci are said to be in LD when the frequency of association of their different alleles is higher or lower than what would be expected if the loci were independent and associated randomly.

Recombination is the process or act of exchanges of DNA between chromosomes, resulting in a different genetic combination and ultimately to the formation of unique gametes with chromosomes that are different from those in parents.

Calculation of linkage disequilibrium

Suppose we have the following sequences:

```
ACTTGTAT............GATCAACCAG
ACTCGTAT............GATCAACCAG
ACTCGTAT............GATCAGCCAG
```

SNP1 **SNP2**

Alleles	**1**	**2**
1	T	A
2	C	G

Calculation of linkage disequilibrium

- Covariance between A and B alleles at two loci:

$$D_{AB} = p_{AB} - p_A p_B$$

- Common statistic for summarizing LD:

$$r^2 = \frac{D^2}{p_A(1-p_A)p_B(1-p_B)}$$

- Decay of LD over time (t in generations):

$$D_t = (1-r)^t D_0$$

P(recombination in one generation)

Calculation of linkage disequilibrium

Haplotype	Symbol	Frequency
A_1B_1	x_{11}	0.6
A_1B_2	x_{12}	0.1
A_2B_1	x_{21}	0.2
A_2B_2	x_{22}	0.1

Allele	Frequency
A_1	$p_1=x_{11}+x_{12}=0.7$
A_2	$p_2=x_{21}+x_{22}=0.3$
B_1	$q_1=x_{11}+x_{21}=0.8$
B_2	$q_2=x_{12}+x_{22}=0.2$

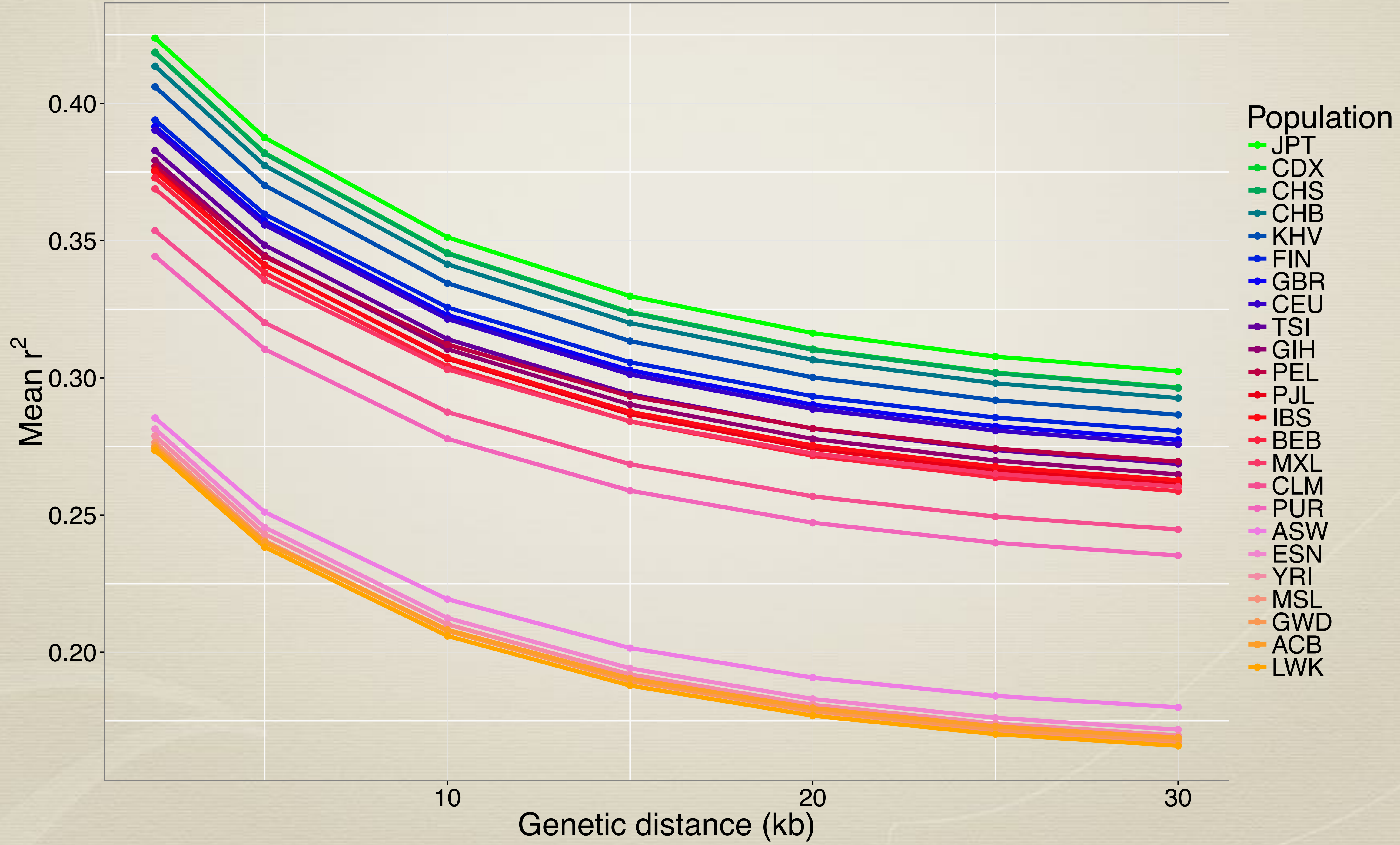
expected under equilibrium

observed

$$\begin{aligned} D &= \boxed{x_{11}} - \boxed{p_1 q_1} \\ &= 0.6 - 0.7 * 0.8 \\ &= 0.04 \end{aligned}$$

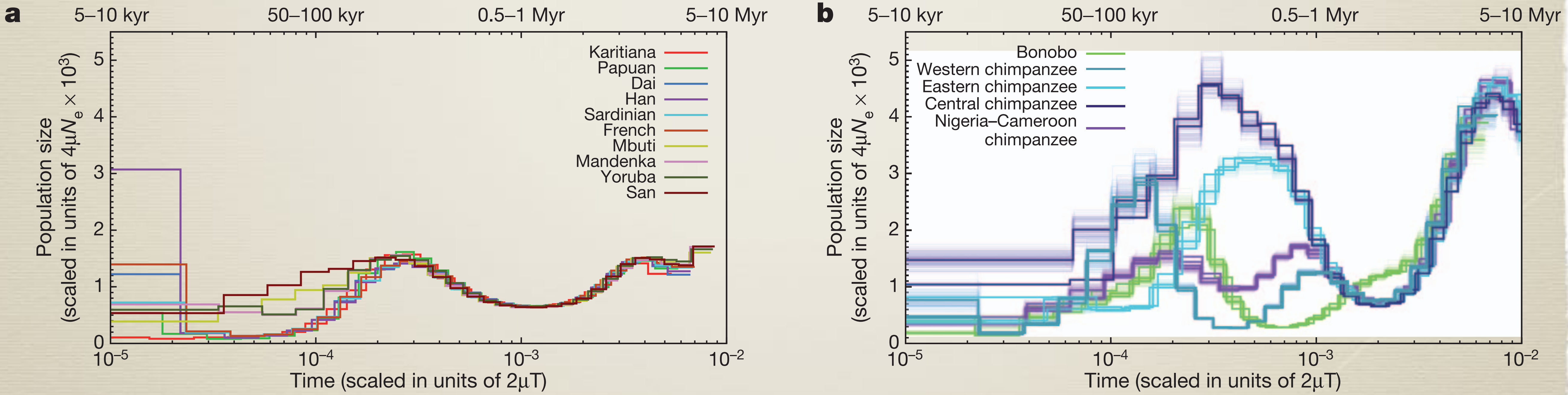
$$\begin{aligned} r^2 &= \frac{D^2}{p_1 p_2 q_1 q_2} \\ &= \frac{0.04^2}{0.7 * 0.3 * 0.8 * 0.2} \\ &= 0.048 \end{aligned}$$

LD decay across 1000 Genomes populations



Effective population size

The **effective population size** (N_e) is the population size that would result in the same rate of drift in an idealized constant population size, obeying our modeling assumptions, as that observed in our true population.



Prado-Martinez, J., et al. (2013). Nature 1–5.

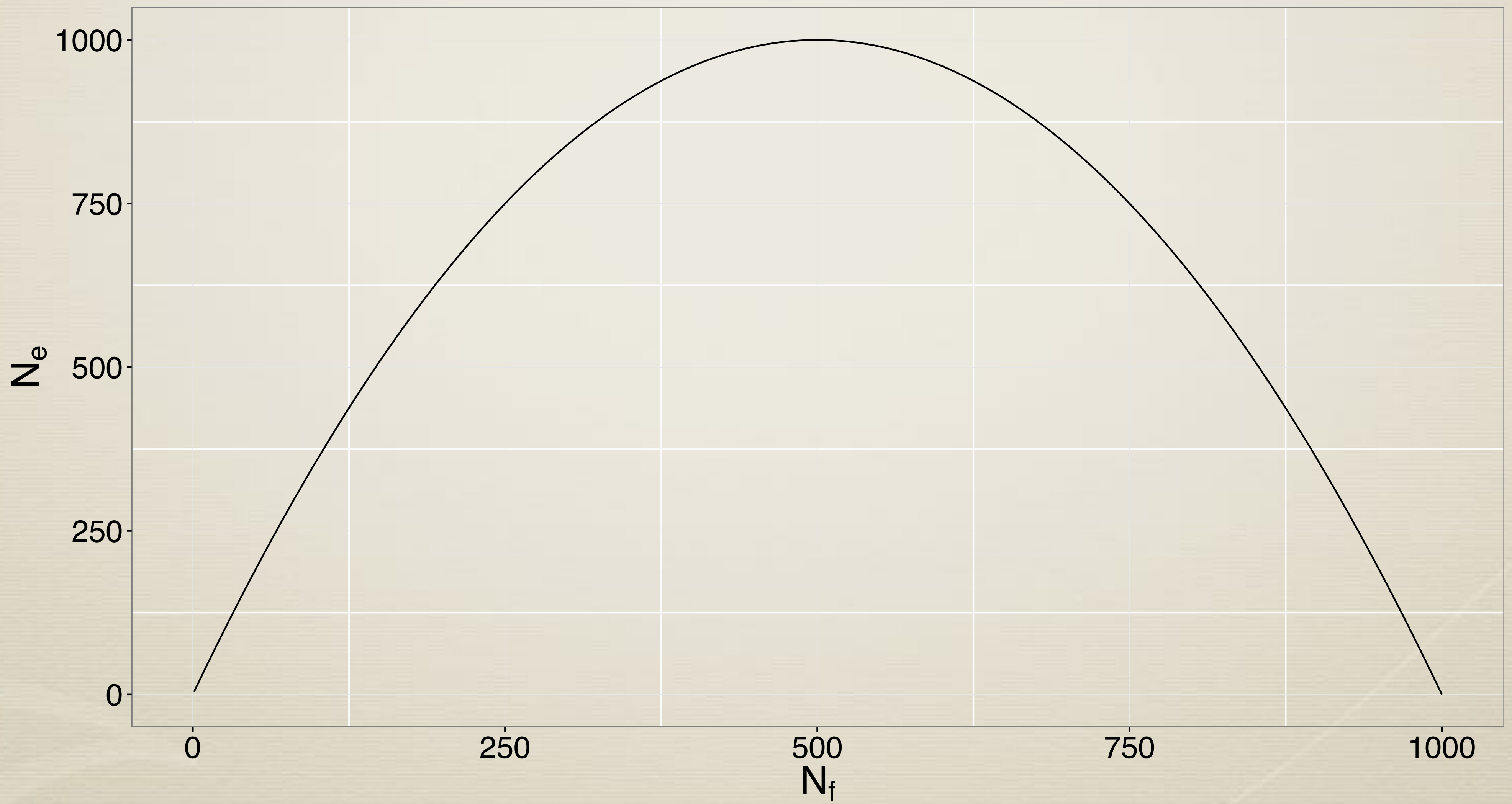
Assumptions of pop gen models affecting N_e when violated

- There are equal numbers of males and females, all of whom are able to reproduce
- All individuals are equally likely to produce offspring, and number of offspring the each produces varies no more than expected by chance
- Mating is random
- The number of breeding individuals is constant from one generation to the next.

Essentially all violations to pop gen models decrease N_e

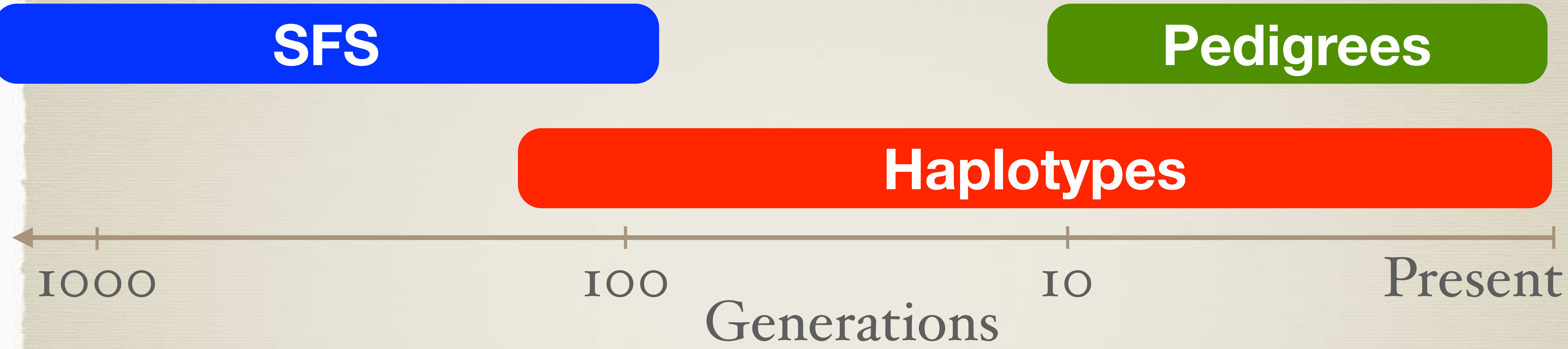
N_e with unequal numbers of breeding males and females

$$N_e = \frac{4N_mN_f}{N_m + N_f}$$



where $N_m + N_f = 1000$

Methodological timeline for human N_e inference



SFS

Pedigrees

Haplotypes

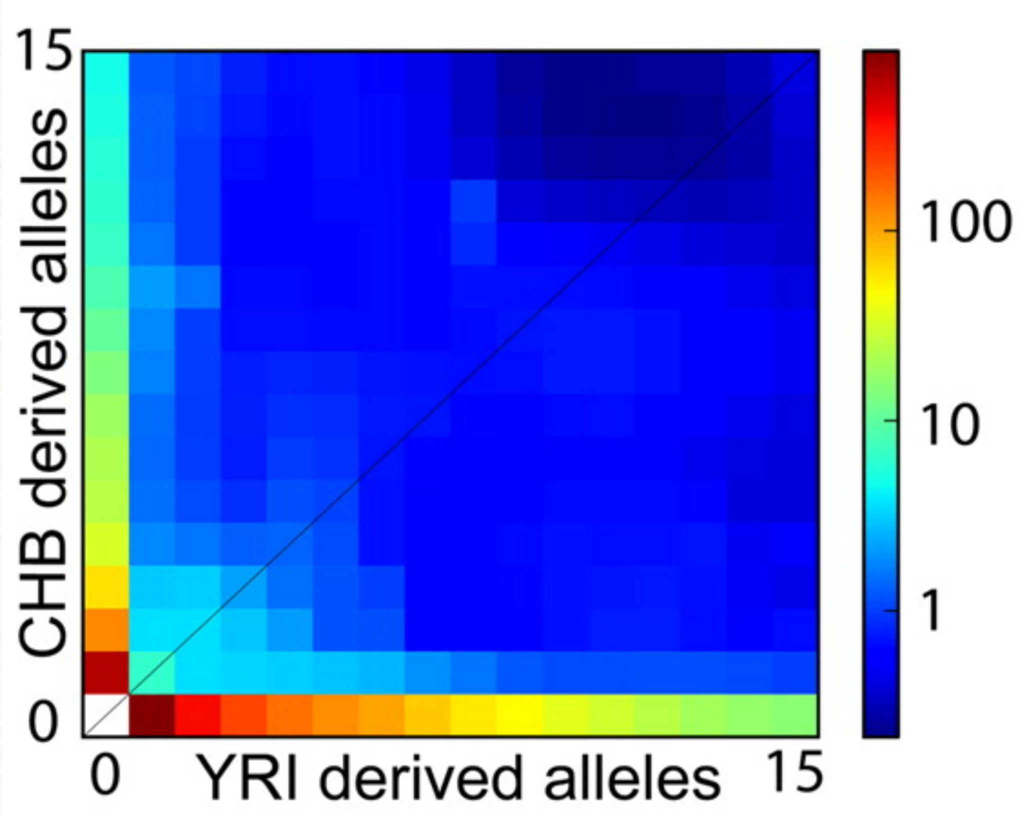
1000

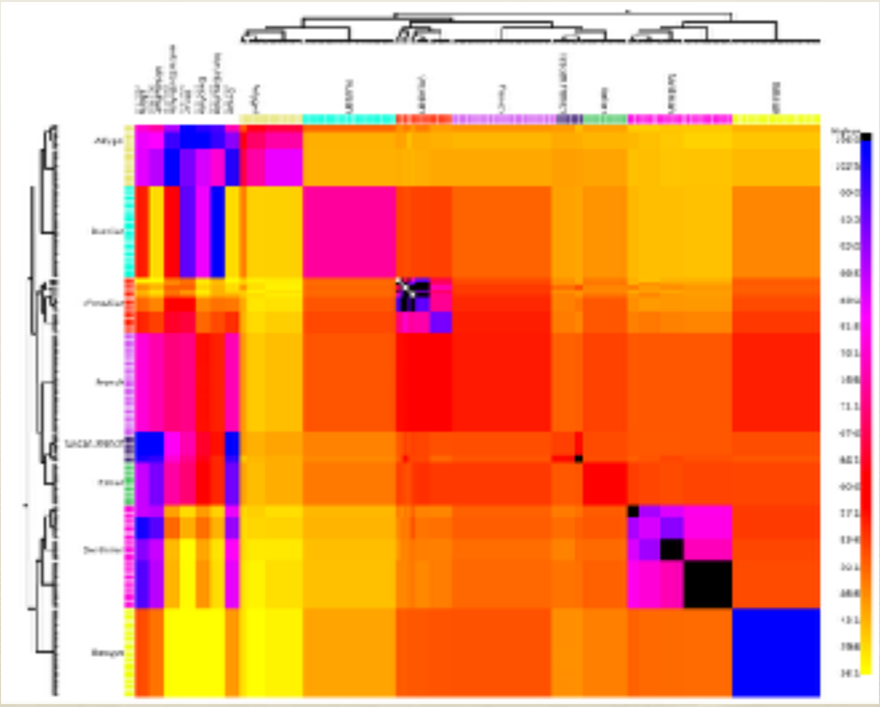
100

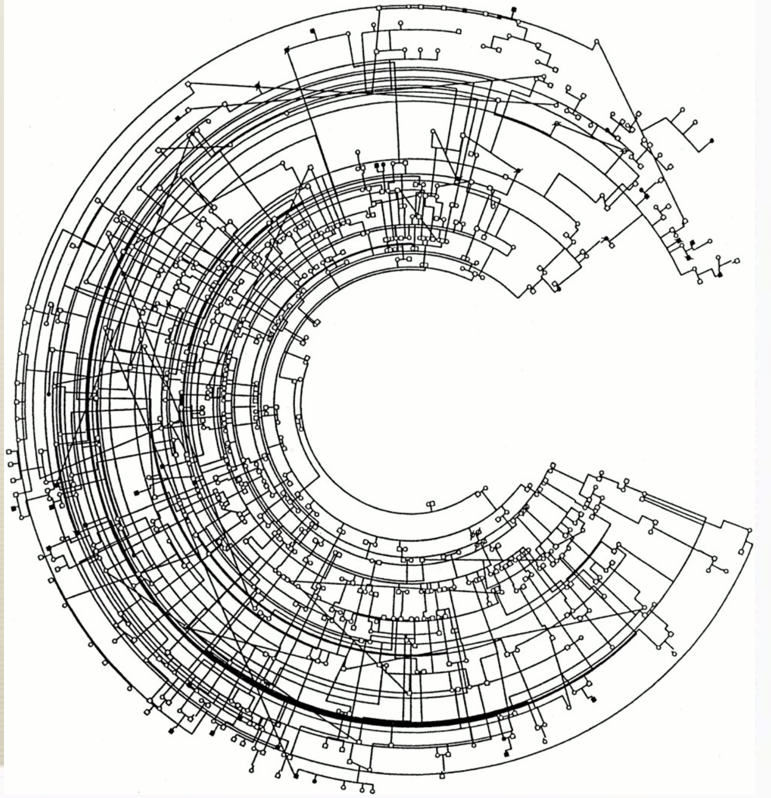
10

Present

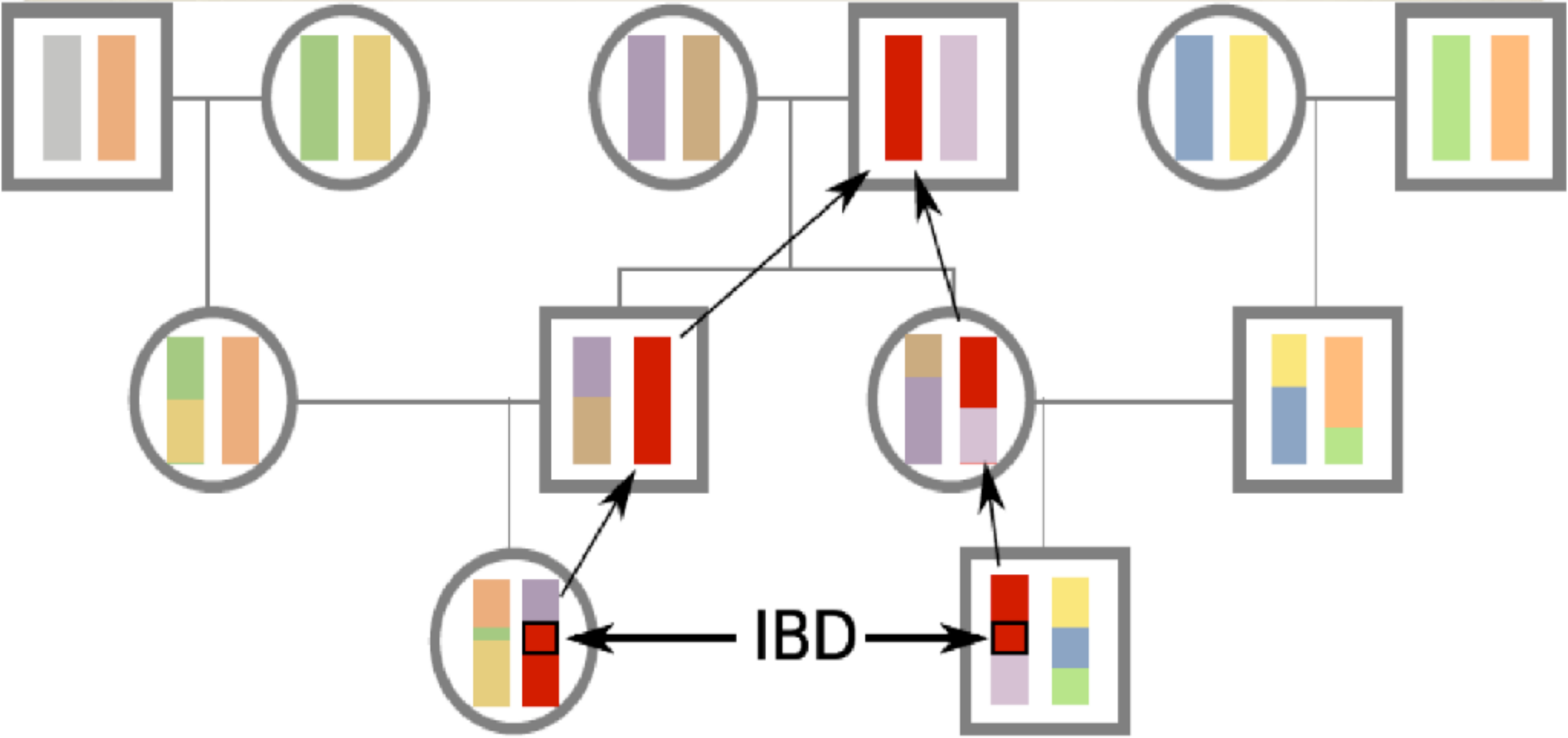
Generations



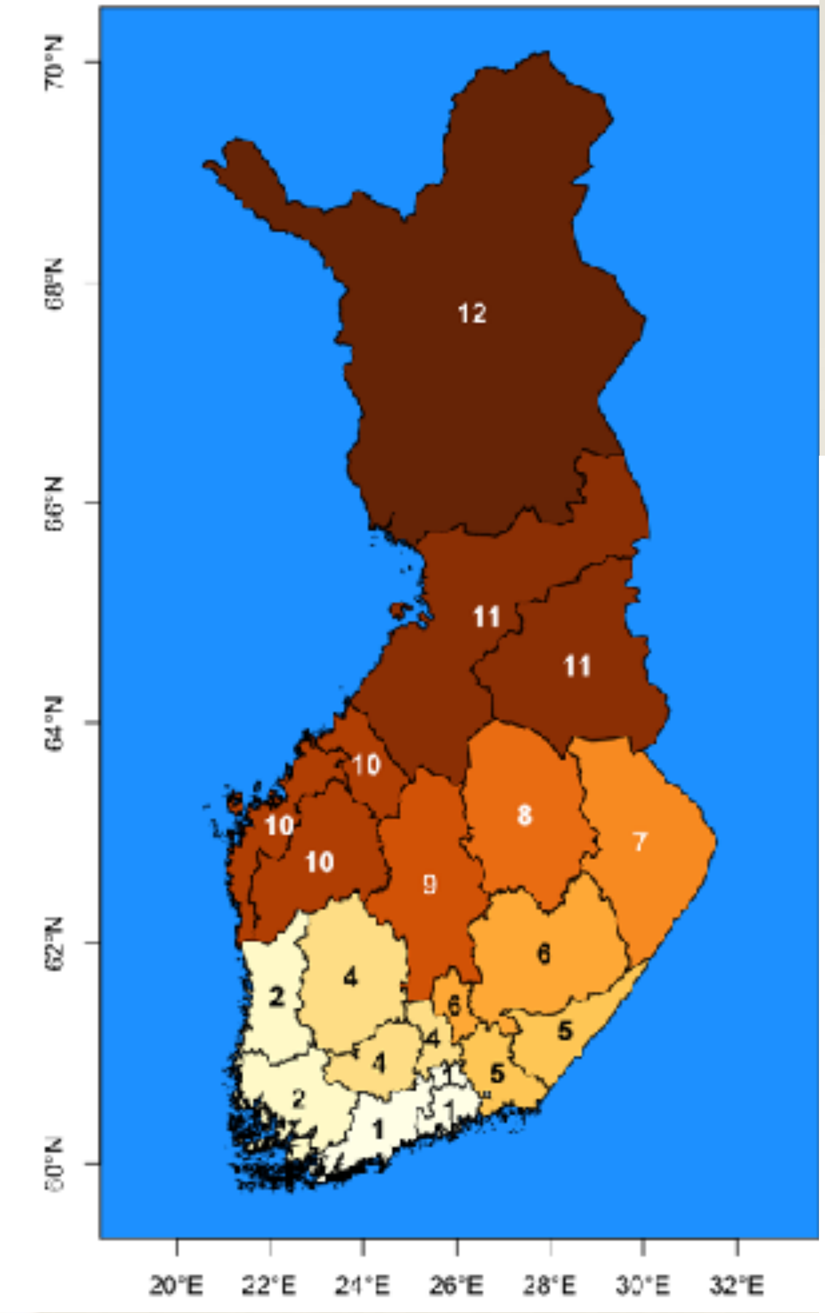




Identity-by-descent

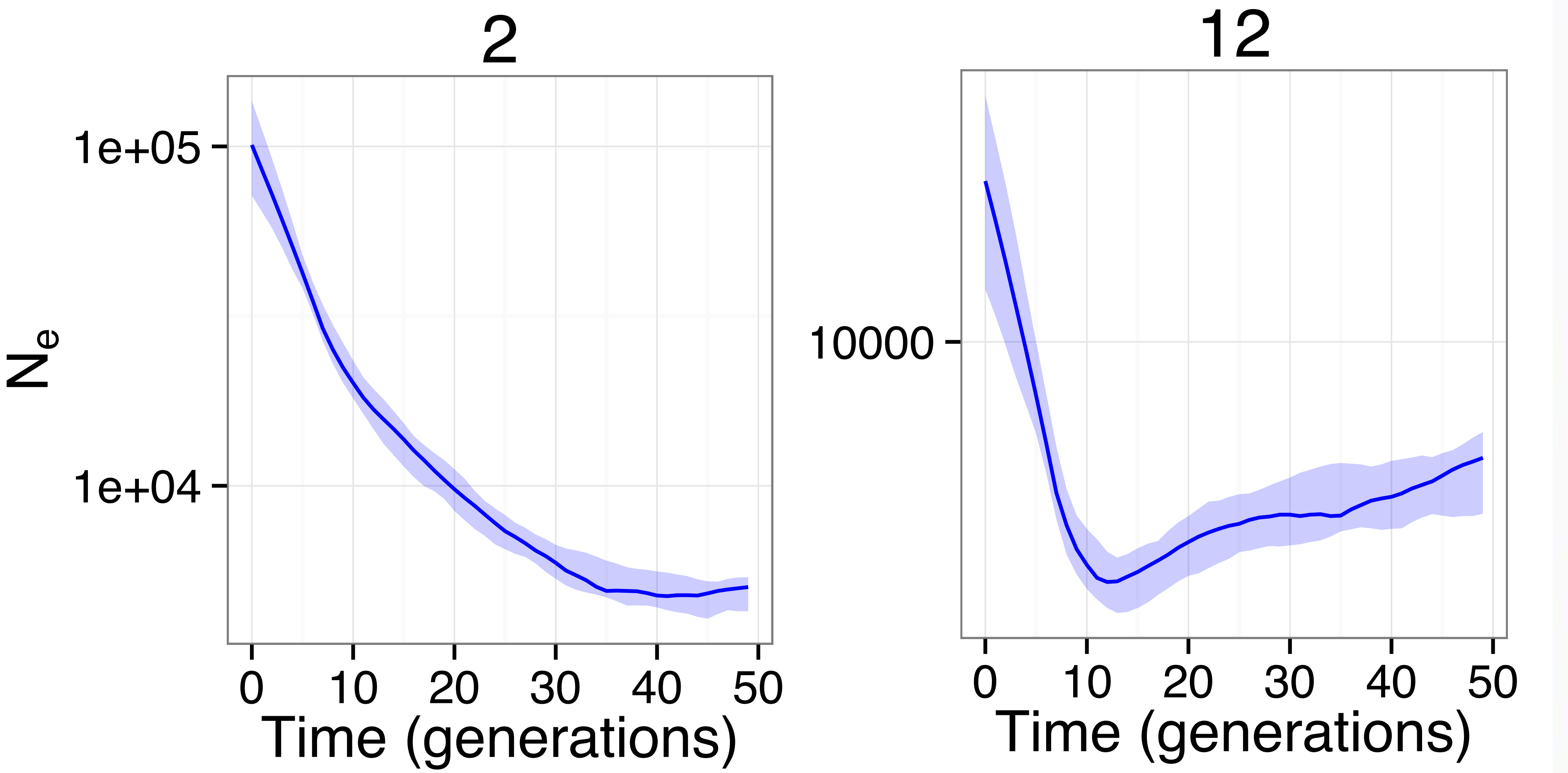


Fine-scale birth record data enables refined view of population history



2: Southwest coastal region started growing longer ago

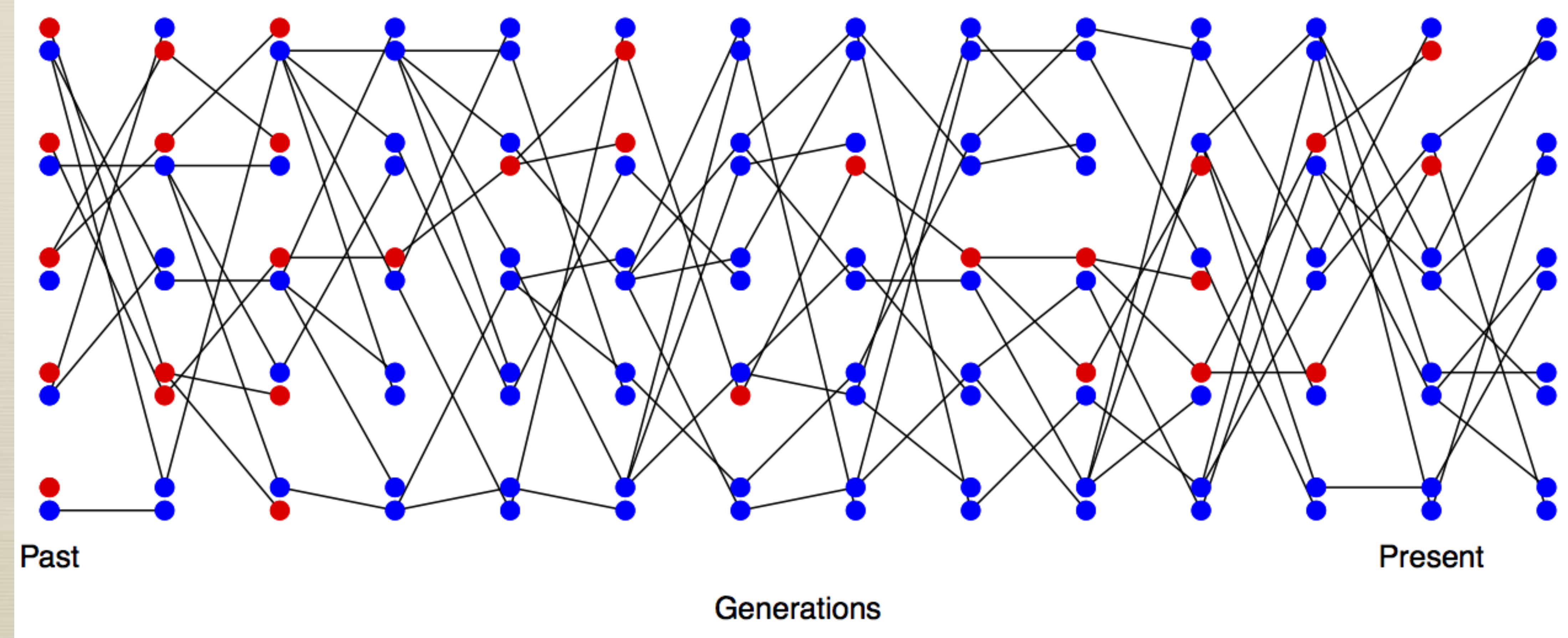
12: Lapland maintained very little growth for extended period



Demographic models

Wright-Fisher model

- Non-overlapping generations
- Finite, constant N
- Binomial sampling of alleles
- Basis of the coalescent



Graham Coop's pop gen notes: http://bit.ly/2fEXzUe

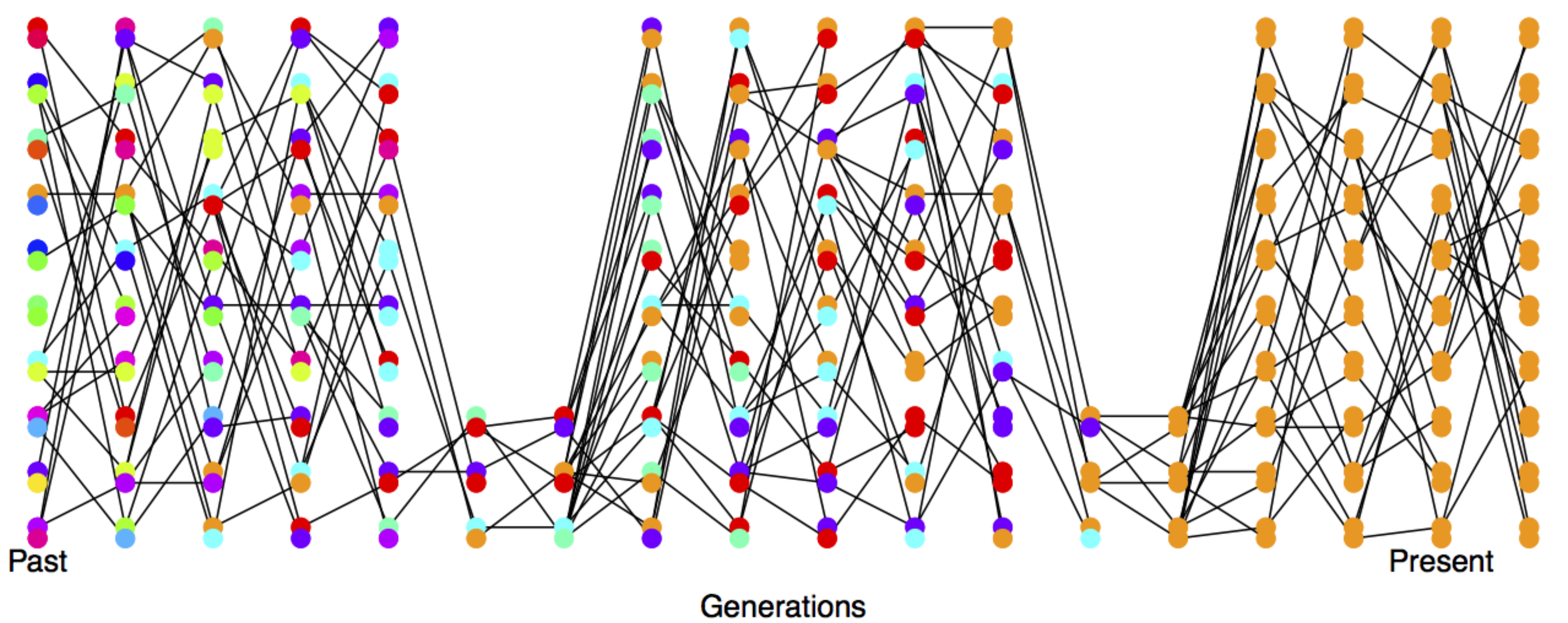
Wright-Fisher model

- Diploid population of size N has 2N alleles
- Probability that two alleles have same parent: 1/2N
- Probability different parent: 1-1/2N
- Probability two alleles coalesce before mutation:

$$\frac{1}{2N}\int_0^\infty e^{-t(2\mu+1/(2N))}dt = \frac{1/(2N)}{1/(2N)+2\mu} = \frac{1}{1+4N\mu}$$

- Population-scaled mutation rate: $\theta=4N_e\mu$
- From the binomial: $$\begin{aligned} E[K_1] &= Np \\ Var[K_1] &= Np(1-p) \end{aligned}$$

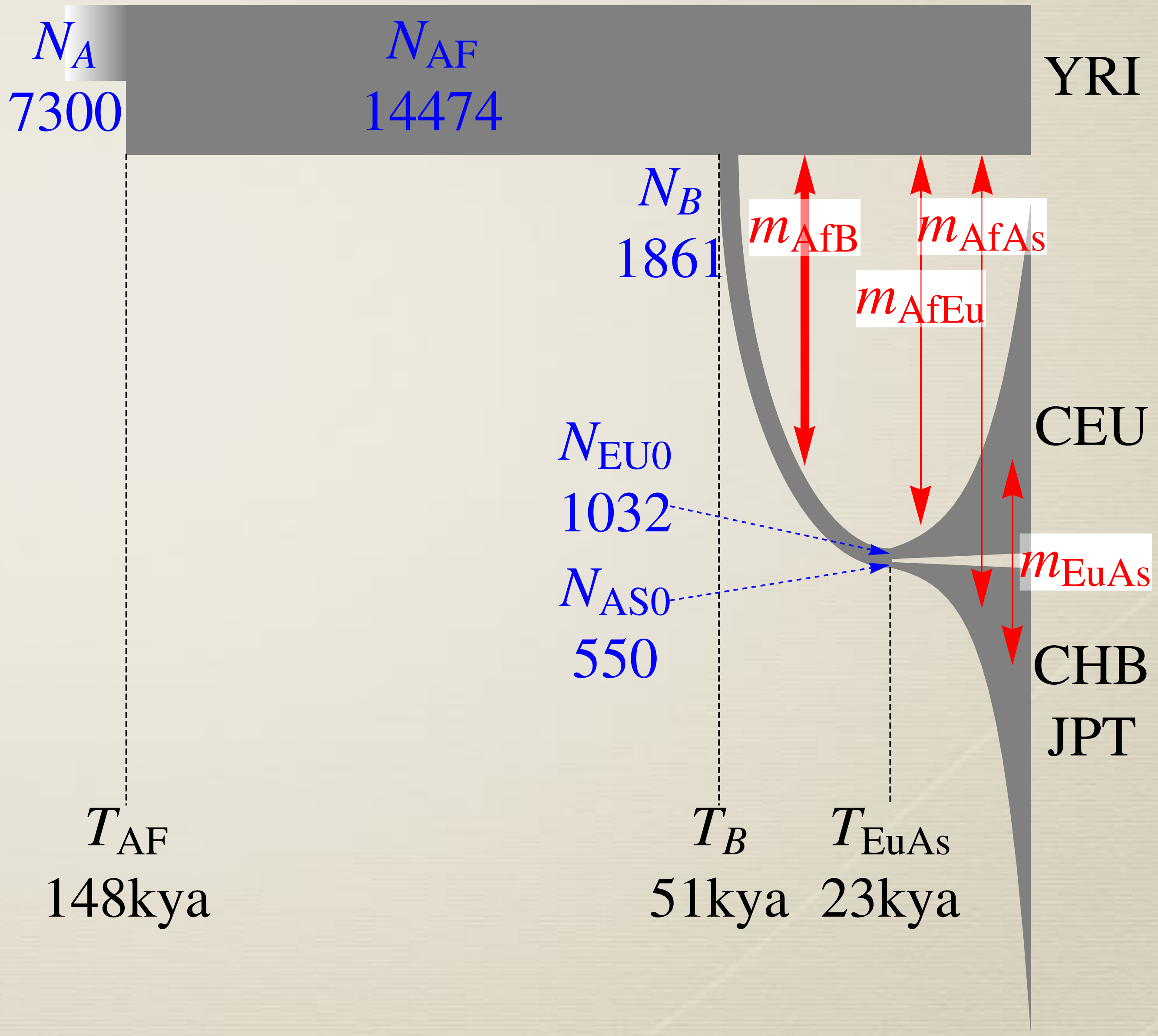
Loss of heterozygosity in a bottlenecking population



Graham Coop's pop gen notes: http://bit.ly/2fEXzUe

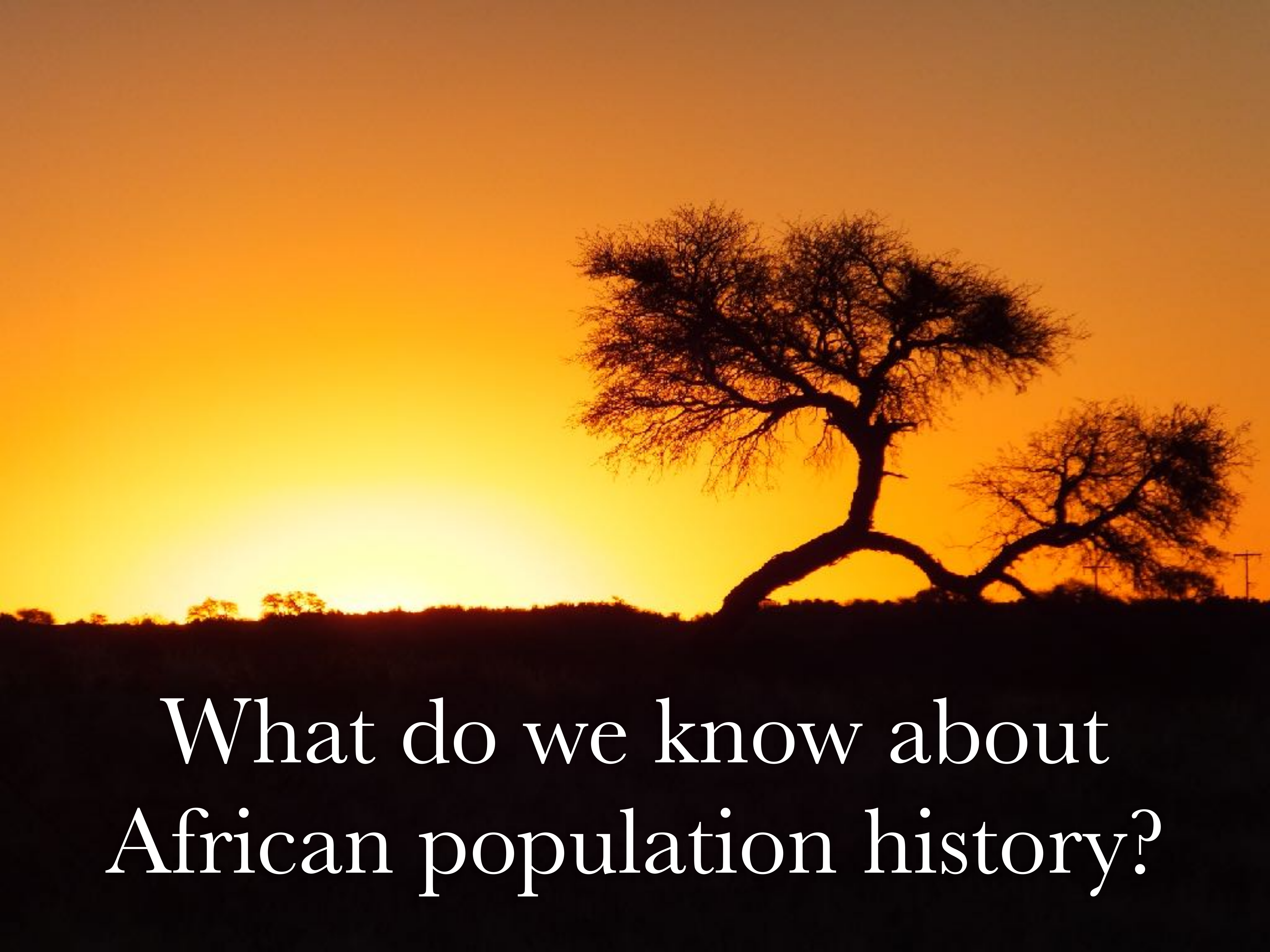
Demographic model from 1000 Genomes data

- Diffusion approximation, δaδi
- Site-frequency spectrum
- 1000 Genomes phase 1



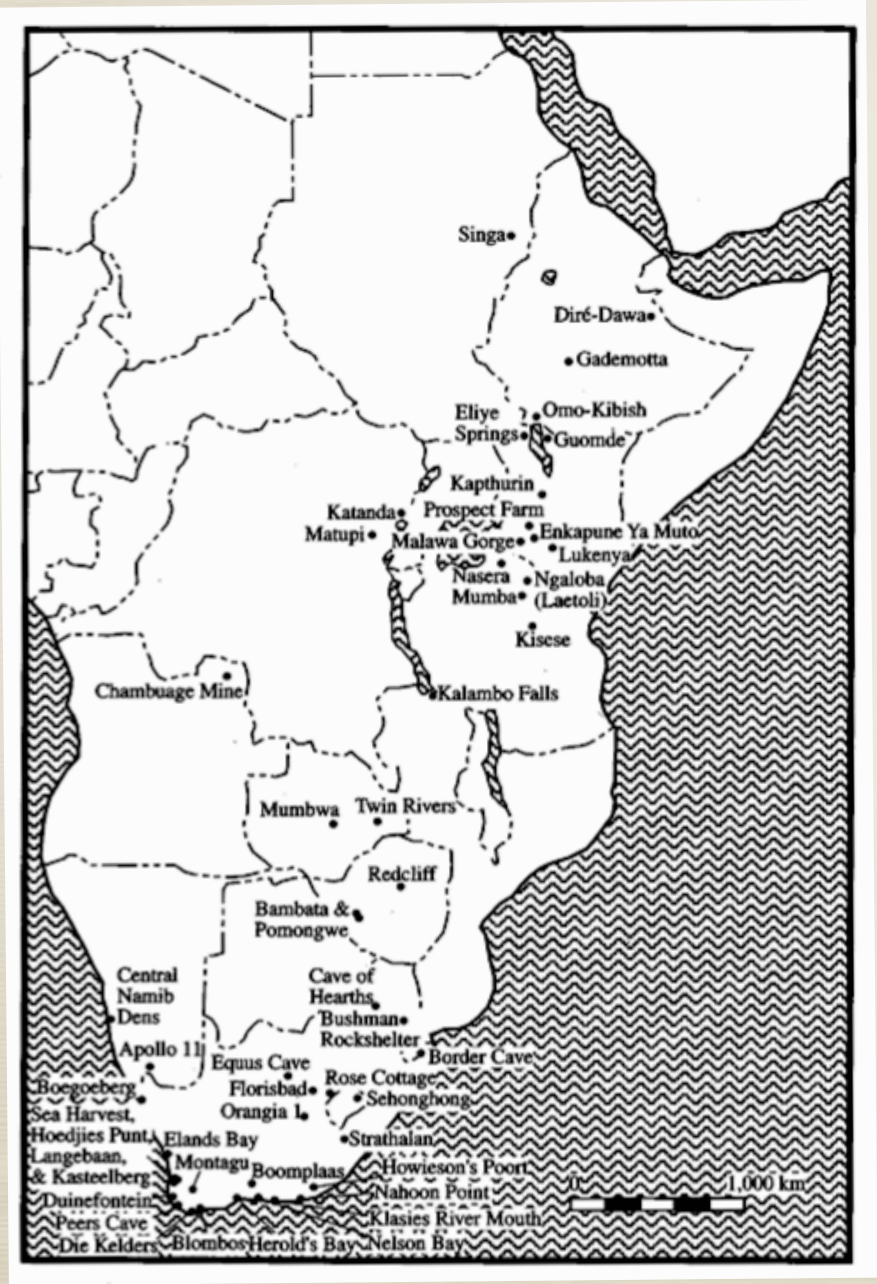
Gravel, S., et al. (2011). PNAS. 108, 11983–11988.

African origins and population structure

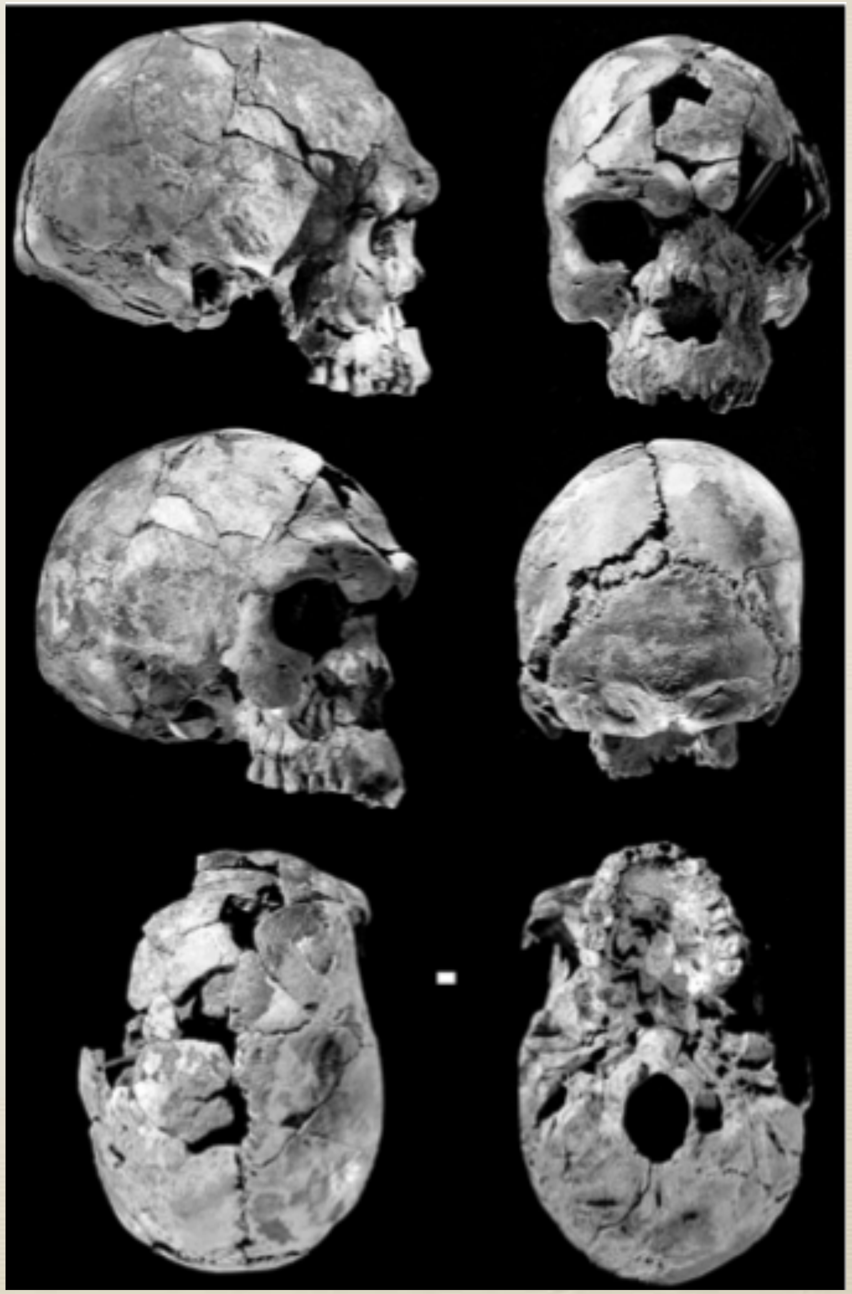


What do we know about African population history?

Anatomically modern humans originated in Africa

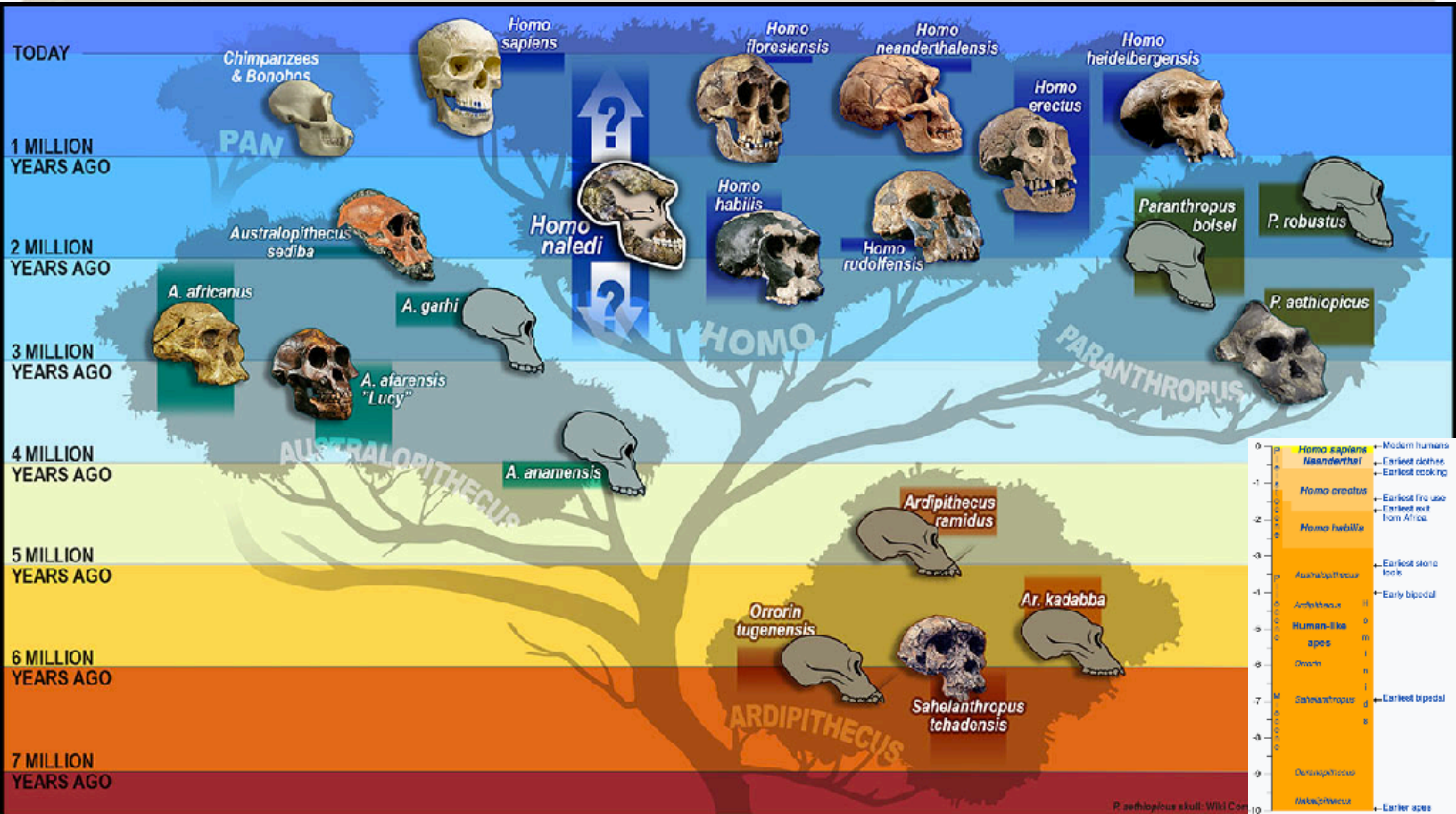


Klein 1999



White 2003

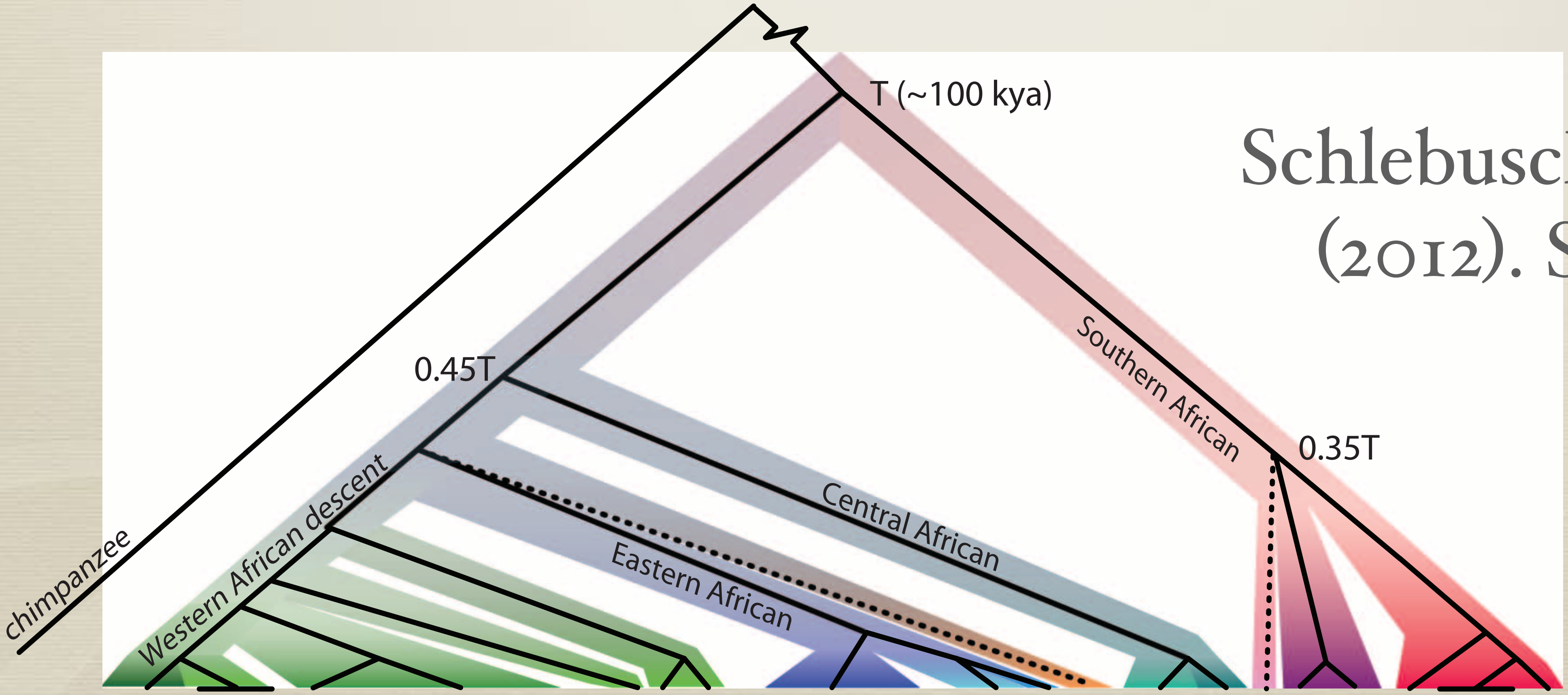
Hominid evolution



P. aethiopicus skull: Wiki Com

Timing of population divergence within Africa

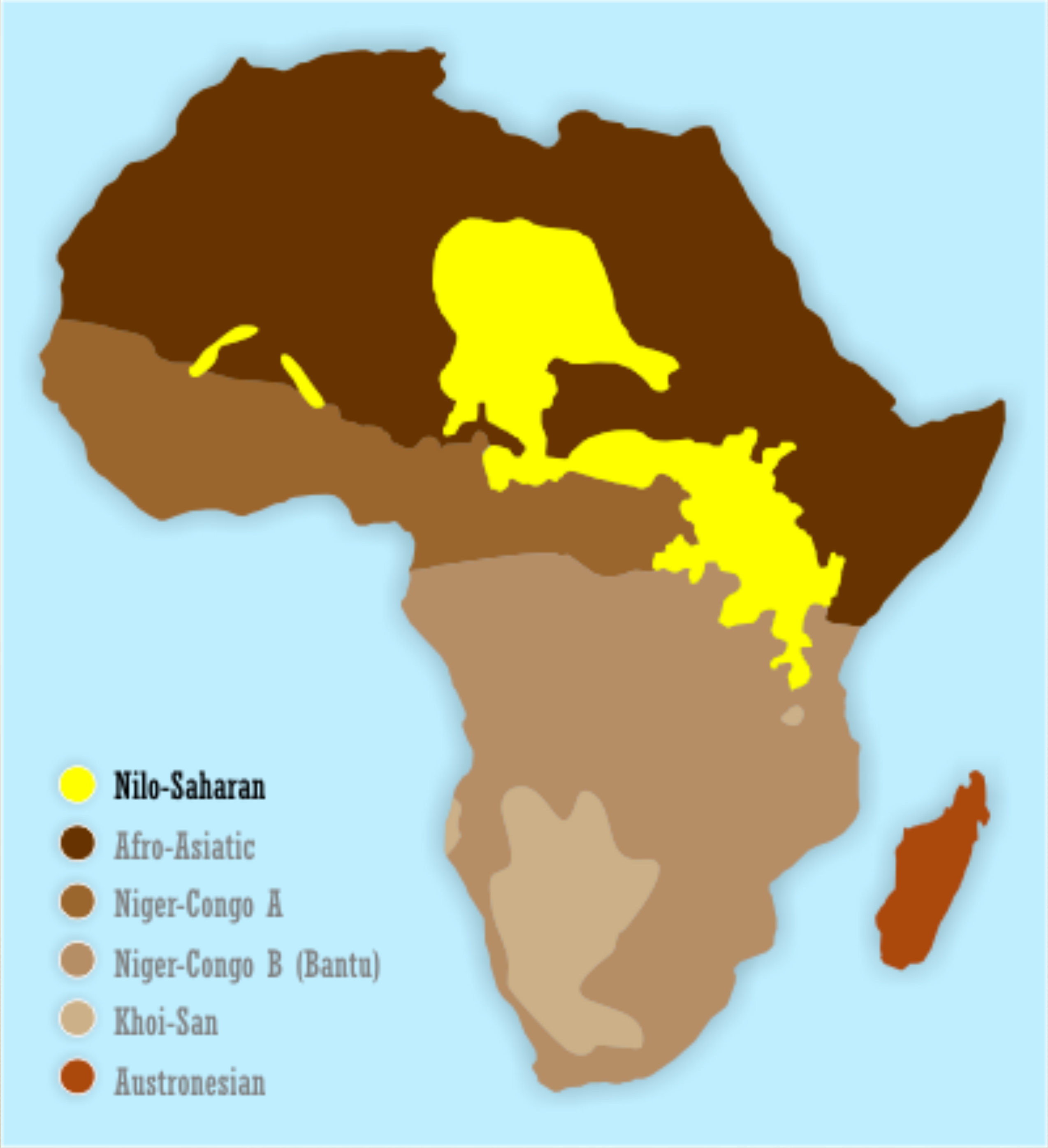
- Oldest divergence is between KhoeSan populations and everyone else (120-90 kya)
- Divergence between Central and Eastern Africans: 70-45 kya
- Eurasians derive from Eastern African populations



Schlebusch, C.M., et al. (2012). Science 374.

Linguistic structure

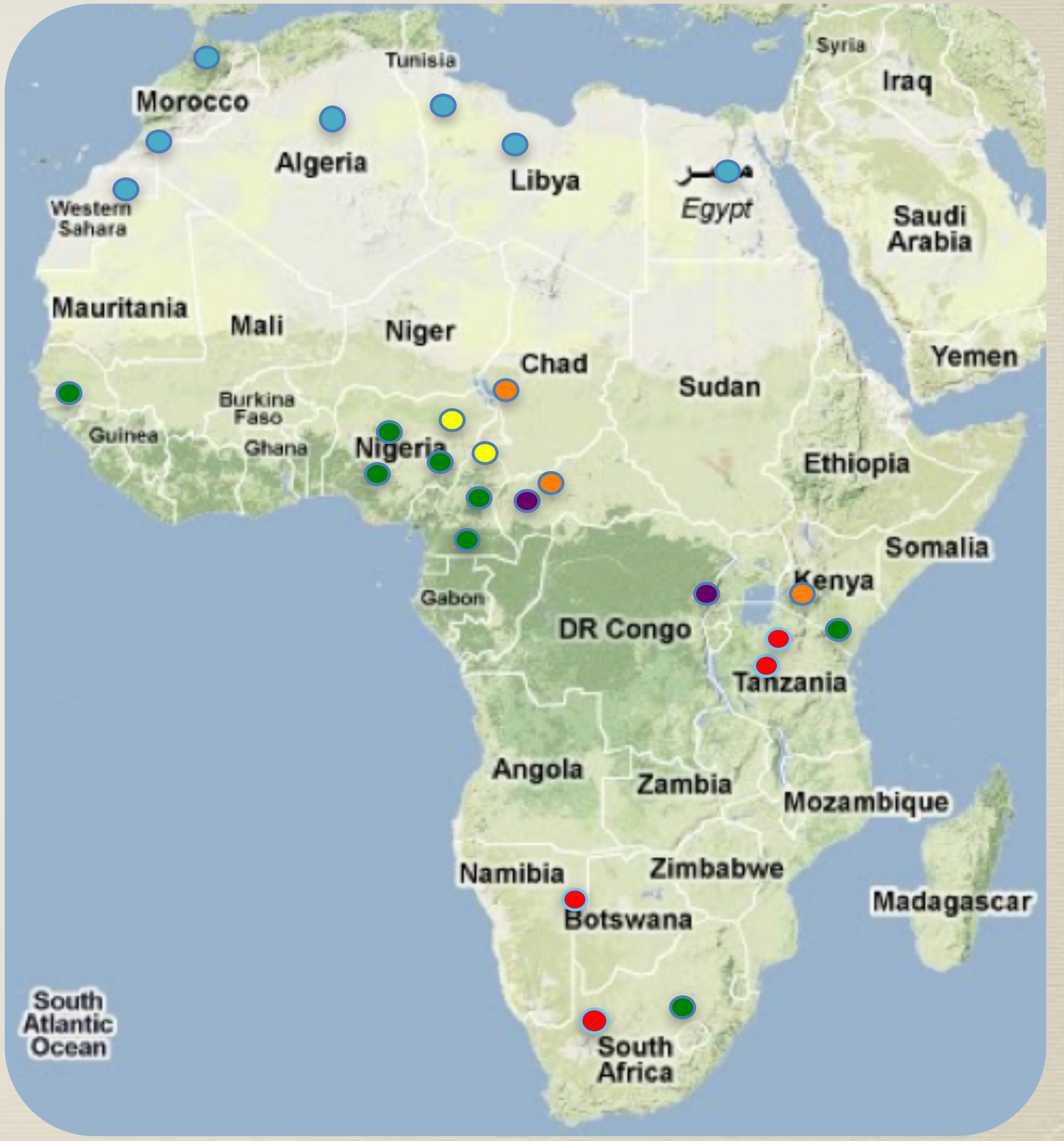
- 5 major language families in Africa
- Expansion of Niger-Congo language 4,000 years ago
- Most isolated and most controversial language family is Khoisan



Population structure

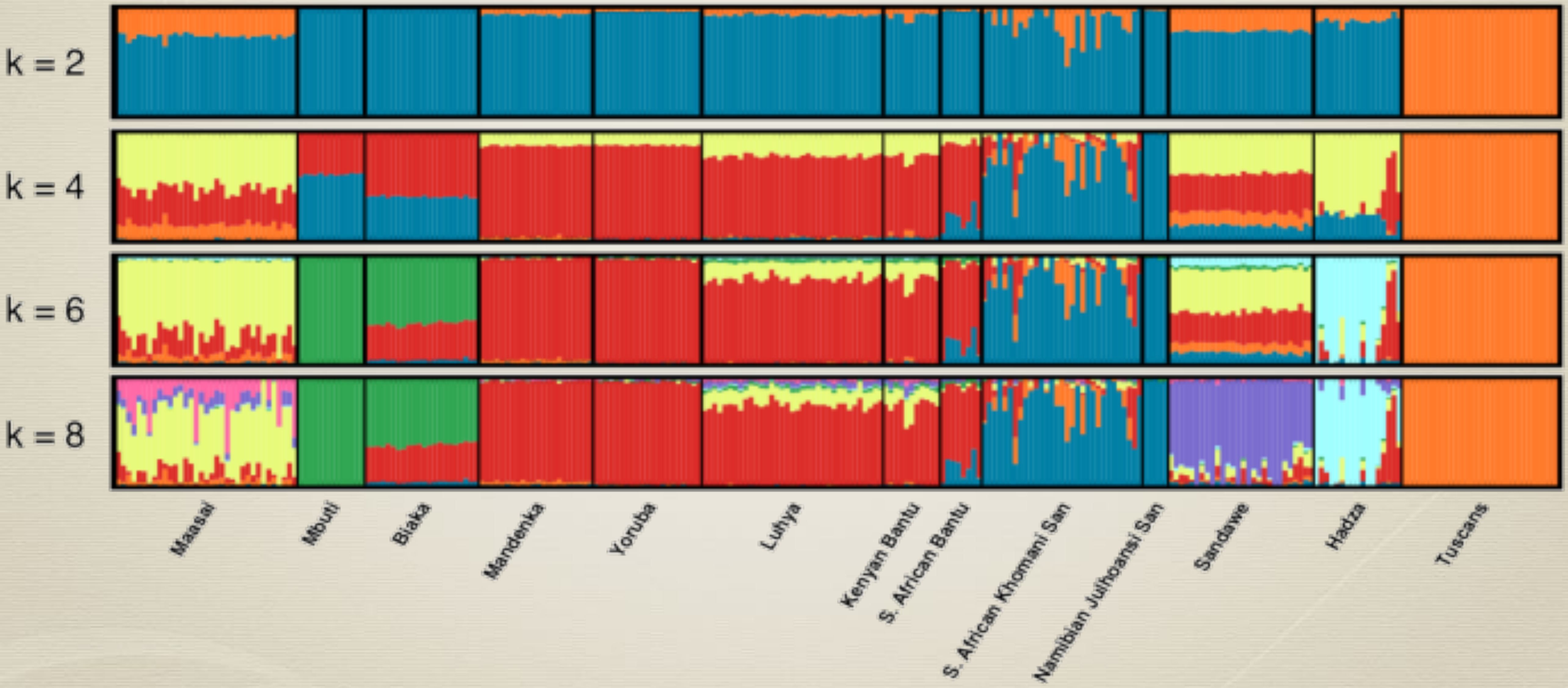
Population samples

- Samples assayed on multiple genotyping platforms: Illumina 550K.v2 & 600K, Affymetrix 6.0, HapMap3
- 50,000 - 500,000 SNPs across the genome
- Datasets are publicly available (http://www-evo.stanford.edu/repository/paper0002/)



Henn, B.M., et al. (2011). PNAS. 108, 5154–5162.

Structure within Africa



Henn, B.M., et al. (2011). PNAS. 108, 5154–5162.

Structure and F_{ST}

Cluster[1]	European	Sandawe	Hadza	Eastern Africa	Maasai[2]	Western African	Forest Pygmies
European							
Sandawe	0.135						
Hadza	0.256	0.158					
Eastern Africa	0.117	0.054	0.154				
Maasai[2]	0.172	0.108	0.218	0.104			
Western Africa	0.169	0.053	0.16	0.046	0.103		
Forest Pygmies	0.23	0.102	0.158	0.105	0.167	0.084	
Southern KhoeSan	0.25	0.122	0.222	0.131	0.194	0.115	0.107

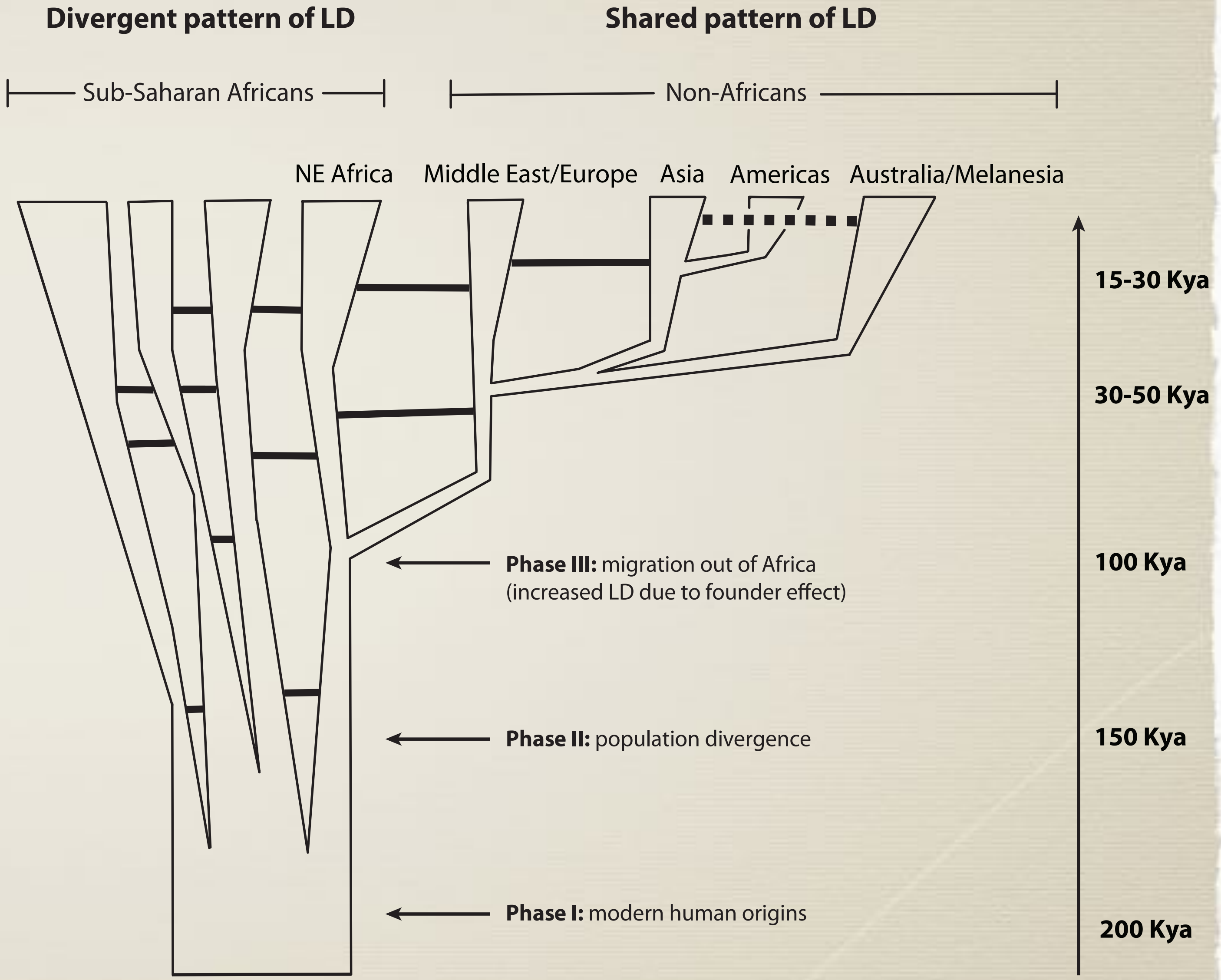
Henn et al. (PNAS, 2011)

Summary

- African populations are highly structured (pre-Bantu expansion)

- Time depth of structure is unresolved (~120-40 kya)

- Despite recent gene flow, underlying structure and diversity is detectable



Campbell, M.C., & Tishkoff, S.A. (2008). Annu Rev Genomics Hum Genet 9, 403–433.

Takeaways

- Complexities to population structure (LD, allele frequency differences, etc). Need to consider for ALL genetic methods:
  - GWAS - has potential to confound associations
  - RVAS - difficulty accounting for rare structure
  - Genetic risk prediction
  - ...many more