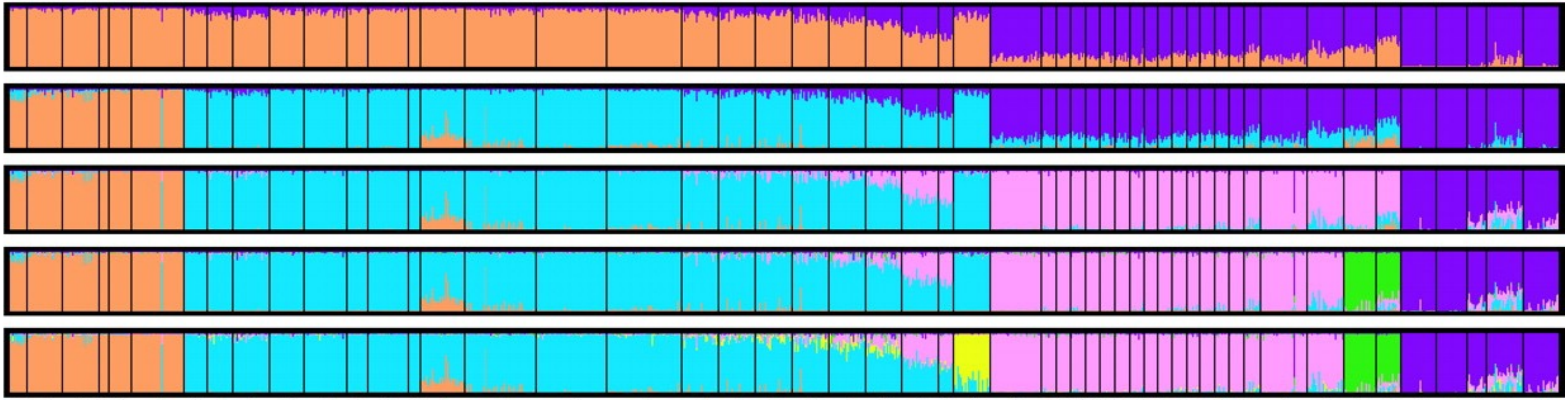
Population Structure

Hancock

March 21, 2024

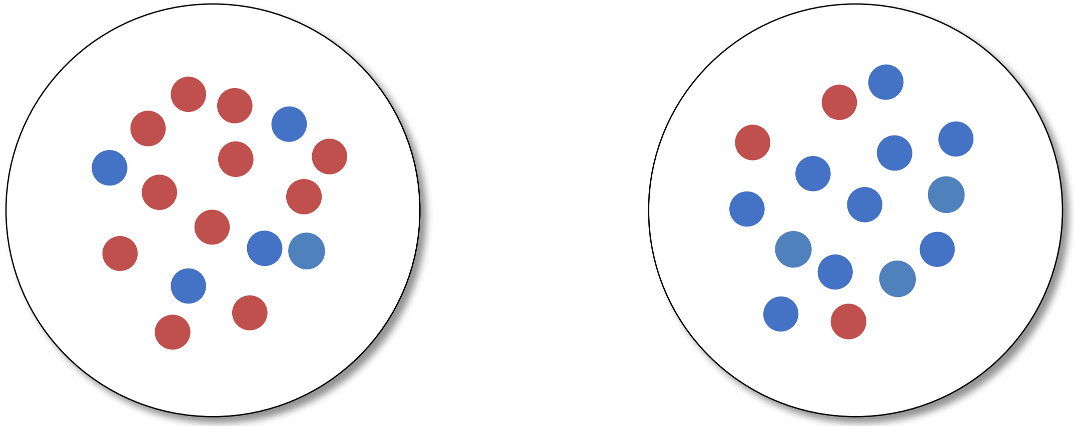


Population structure: what is it and why should we care?

- Real populations are at least somewhat spatially structured
- Structure arises when individuals living close to one another tend to mate more often than those living farther away
- When populations are not randomly mating due to geographic structure, there is ‘population subdivision’ or ‘population structure’
- Population structure can impact diversity and the rate of adaptation because (with low or no migration) drift and selection act at the level of individual sub-populations (demes)
- Similarly, population structure is relevant for conservation because isolated populations are prone to lose diversity via drift, and lower levels of genetic variation may impede selection – this is why migration corridors are important in conservation programs

Allele frequency in subdivided populations

- A simple model of population structure
- Two populations (or “demes”)



Population 1

Population 2

The mean frequency of the A allele overall is a weighted average:

$$f_A = \frac{2N_1 f_{A1} + 2N_2 f_{A2}}{2N_1 + 2N_2}$$

If populations are of the same size, it would reduce to the more familiar:

$$f_A = \frac{f_{A1} + f_{A2}}{2}$$

Heterozygosity in subdivided populations

Average heterozygosity across the sub-populations, simplifies if the two populations are of equal size and HWE:

$$H_S = \frac{2f_{A1}(1-f_{A1})+2f_{A2}(1-f_{A2})}{2} = \mathrm{f}_{\mathrm{A1}}(1-\mathrm{f}_{\mathrm{A1}})+\mathrm{f}_{\mathrm{A2}}(1-\mathrm{f}_{\mathrm{A2}})$$

Heterozygosity in the total population if the two populations are of equal size and in HWE:

$$H_T = 2\frac{f_{A1}+f_{A2}}{2}\,(1-\frac{f_{A1}+f_{A2}}{2})$$

Expected heterozygosity if the pooled population is in HWE

Heterozygosity in subdivided populations

Average heterozygosity across the sub-populations, simplifies if the two populations are of equal size and HWE:

$$H_S = \frac{2f_{A1}(1 - f_{A1}) + 2f_{A2}(1 - f_{A2})}{2} = \mathrm{f}_{\mathrm{A1}}(1 - \mathrm{f}_{\mathrm{A1}}) + \mathrm{f}_{\mathrm{A2}}(1 - \mathrm{f}_{\mathrm{A2}})$$

Heterozygosity in the total population if the two populations are of equal size and in HWE:

$$H_T = 2\frac{f_{A1} + f_{A2}}{2}\,(1 - \frac{f_{A1} + f_{A2}}{2})$$

Expected heterozygosity if the pooled population is in HWE

If the frequency difference between the two populations is $\delta = |f_{A1} - fA_2|$, then the above equation can be written as:

$$H_T = f_{A1}(1 - f_{A1}) + \mathrm{f}_{\mathrm{A2}}(1 - \mathrm{f}_{\mathrm{A2}}) + \frac{\delta^2}{2}$$

The Wahlund effect

$$H_T = f_{A1}(1 - f_{A1}) + \mathrm{f}_{\mathrm{A2}}(1 - \mathrm{f}_{\mathrm{A2}}) + \frac{\delta^2}{2}$$

As the frequencies in the two sub-populations diverge, δ increases and $H_T > H_S$, so that the population contains fewer heterozygozotes than expected, given the pooled allele frequencies.

Therefore, population subdivision will always lead to a reduction in heterozygosity (and an increase in homozygosity) relative to a randomly mating population.

This is a general result that also holds if the two populations are of different sizes.

This decrease in heterozygosity is referred to as the Wahlund effect

Wahlund effect

- An inevitable consequence of drift among subpopulations is a deviation from the expected heterozygosity under Hardy-Weinberg for the population as a whole
- The more different the gene frequencies among subpopulations, the greater the overall loss of heterozygotes in the total population
- The Wahlund effect is a consequence of this variance in the change in allele frequencies

Quantifying population differentiation: F_{ST}

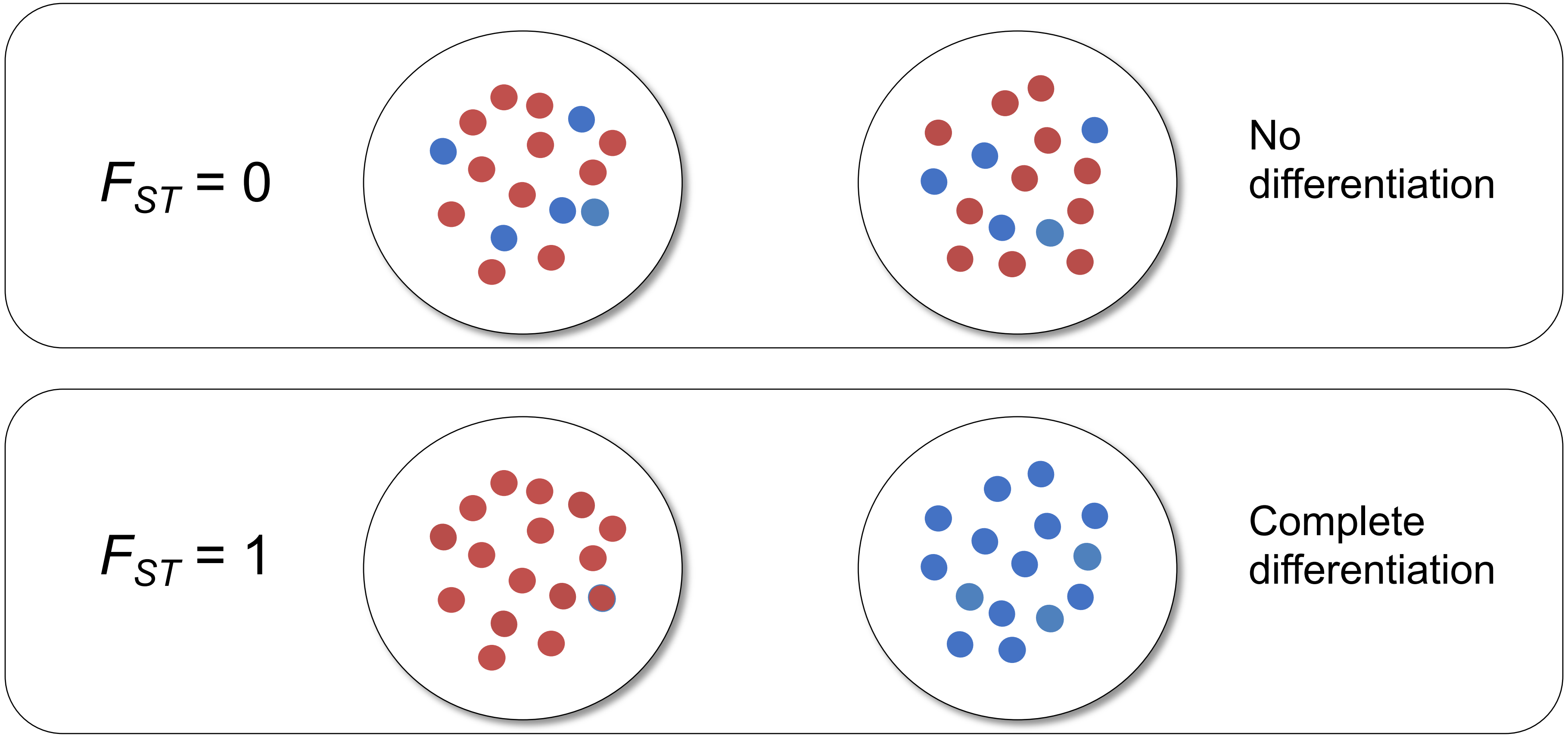
- First introduced by Sewell Wright
- F_{ST} is also referred to as "Wright's fixation index" and is related to his statistics to measure degree of inbreeding

$$F_{ST} = \frac{H_T - H_S}{H_T}$$

- Our example in previous slides was with two populations, but F_{ST} can also be calculated with more populations
- Whenever $H_T = H_S$ then F_{ST} = 0
- As populations become more differentiated, H_T increases relative to H_S , and F_{ST} increases

F_{ST} : Wright's fixation index

F_{ST} measures the amount of genetic variance that can be explained by population structure



$F_{ST} = 0$

No differentiation

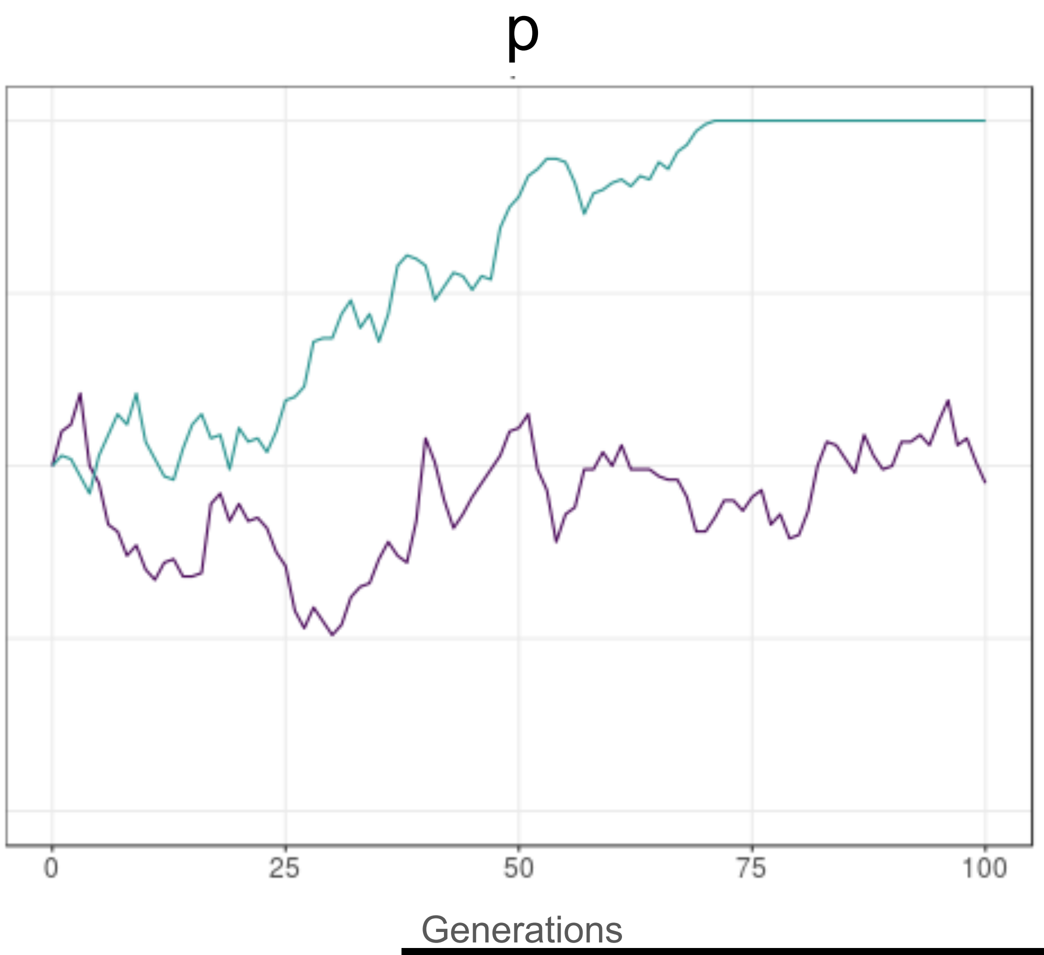
$F_{ST} = 1$

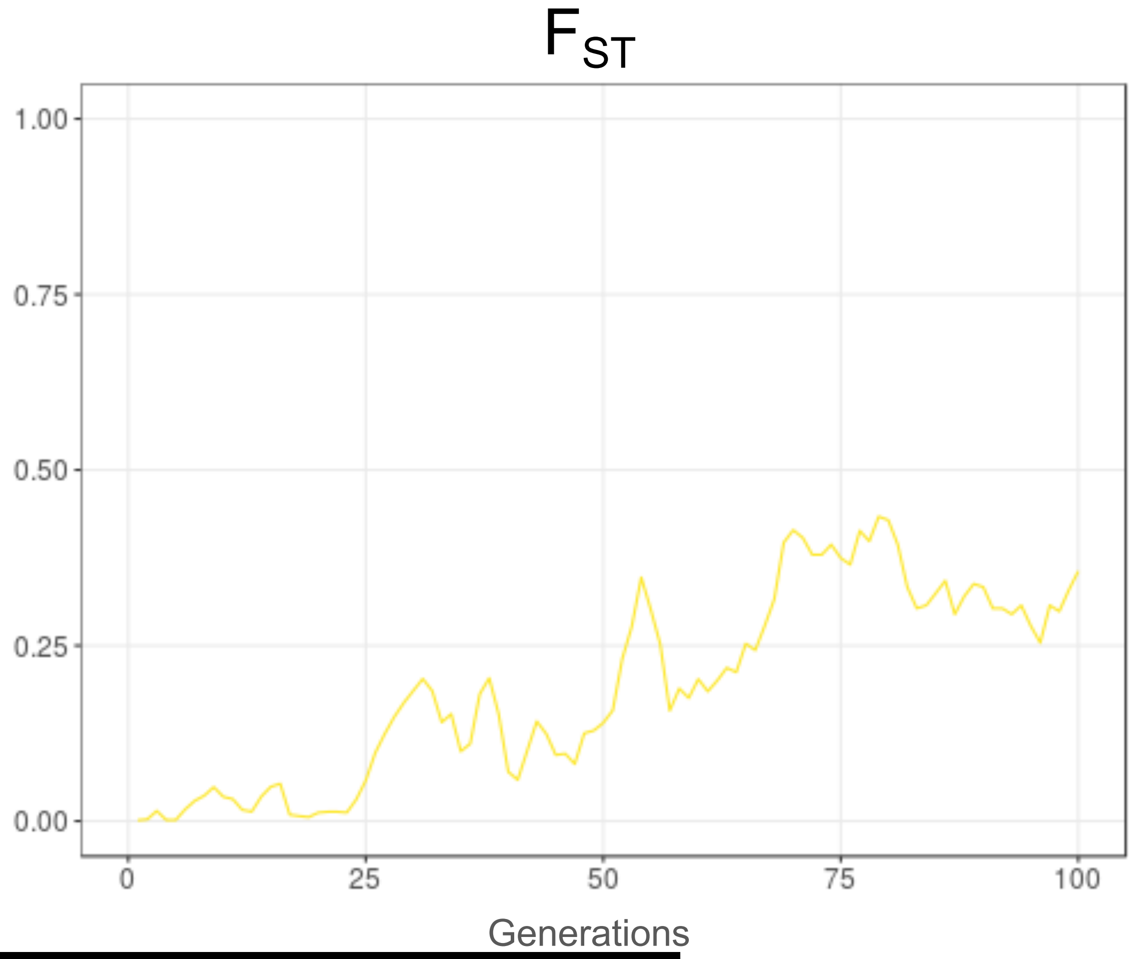
Complete differentiation

Simulations of two populations

- Simulations to examine how F_{ST}, H_S and H_T change with p under structure and migration
- Here is the link to this Rshiny app in case you want to play with the simulation model yourself: https://cjbattey.shinyapps.io/driftR/

Simulation of two populations

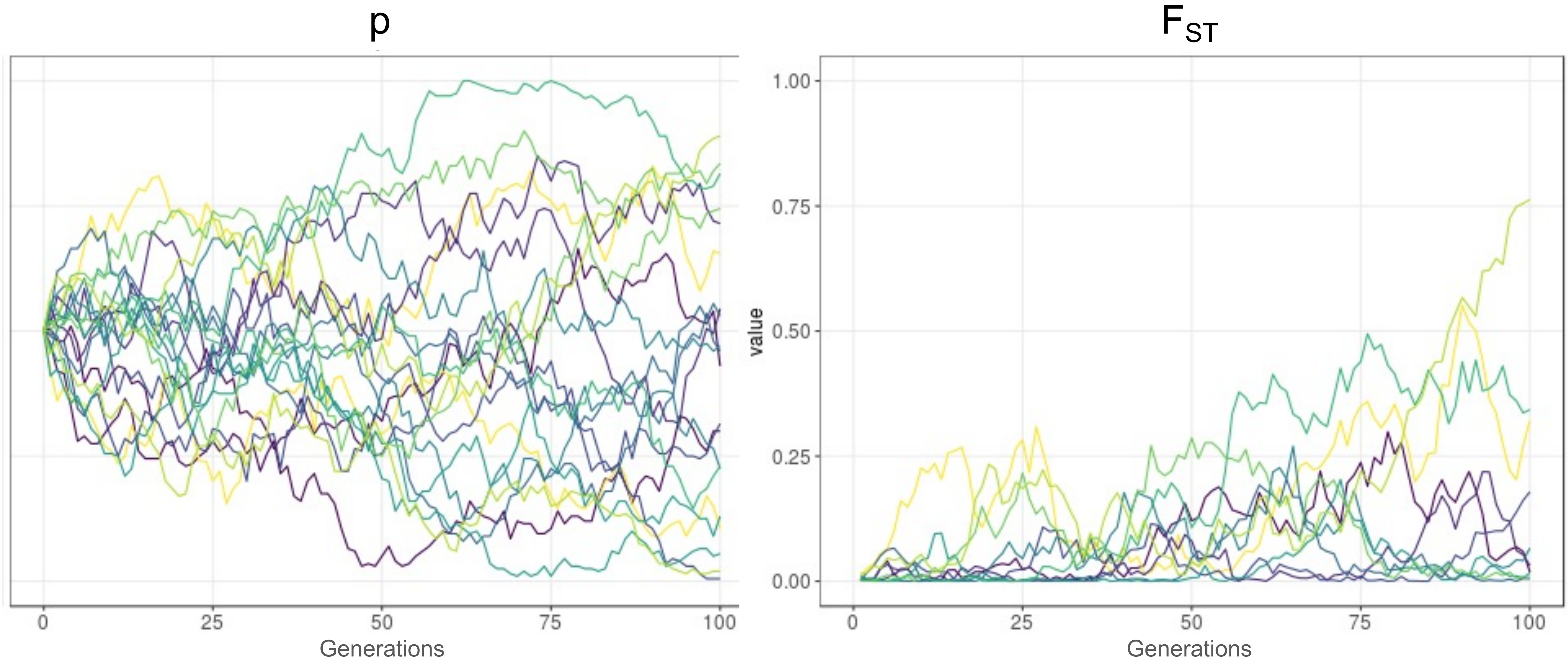




As allele frequencies diverge, F_{ST} increases

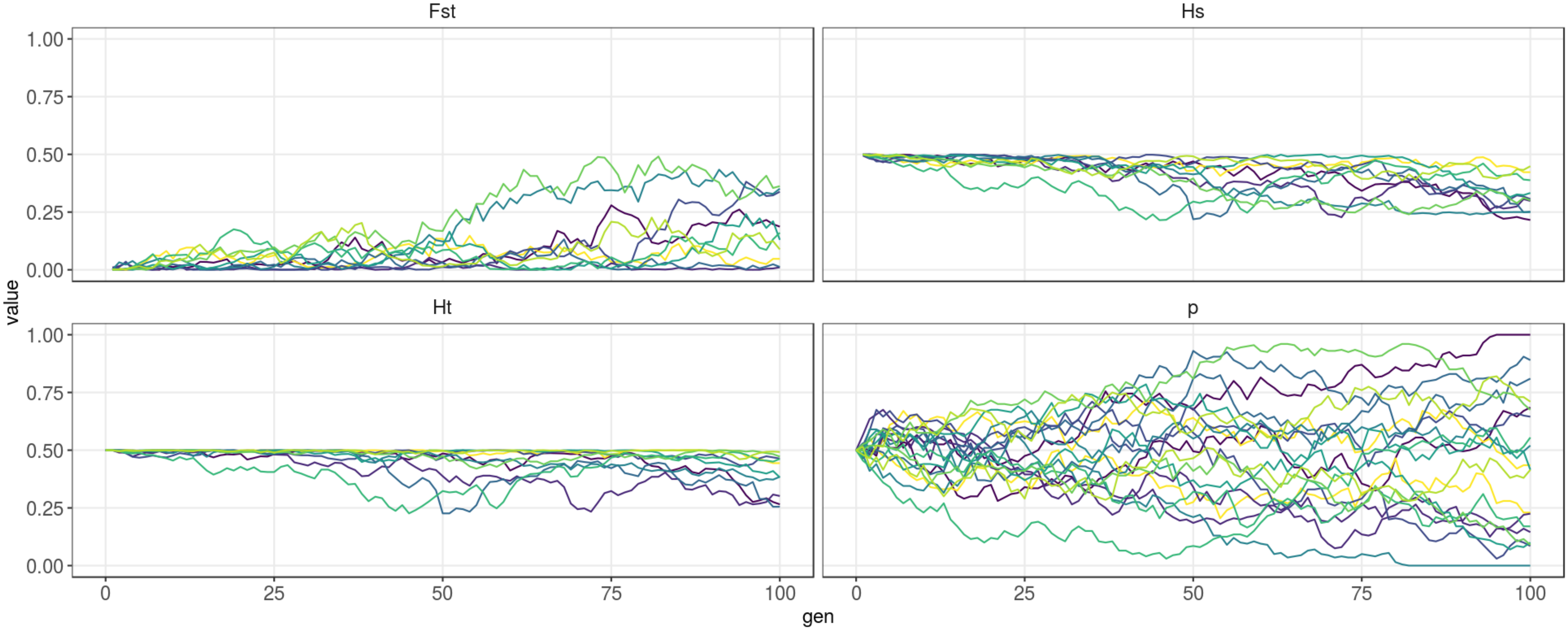
Two populations starting at the same frequency p=q=0.5, 2N=100, no migration, no selection

Simulations of two populations with replicates show variation but on average, F_{ST} increases over time



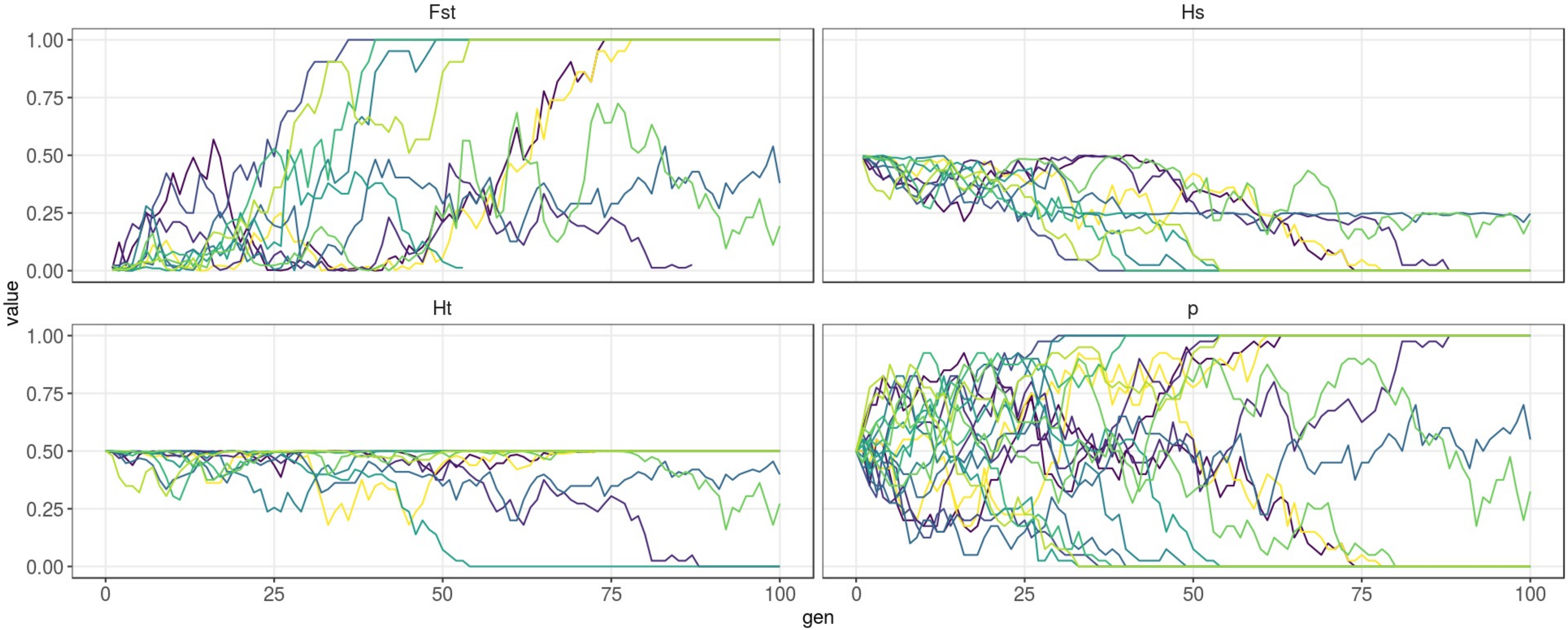
Two populations (with replication) starting at the same frequency p=q=0.5, 2N=100, no migration, no selection

Simulation of two populations with replicates



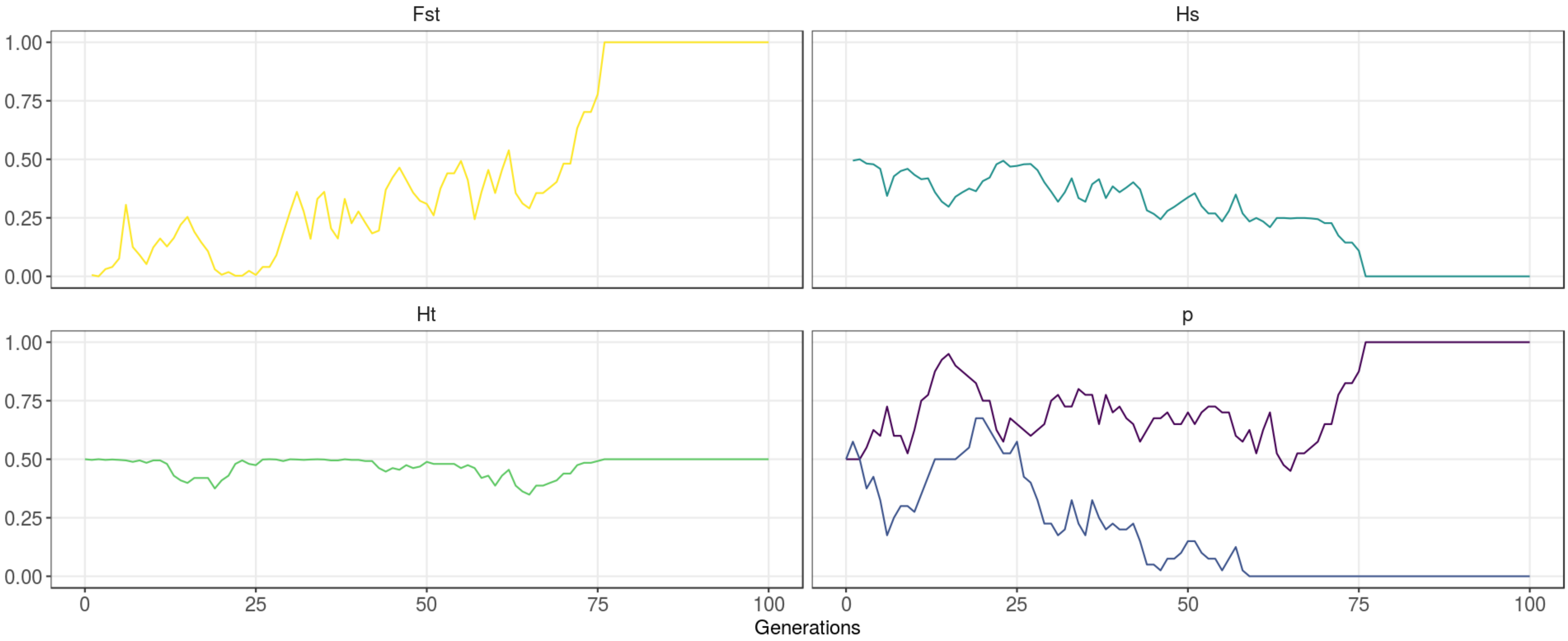
Two populations (with replication) starting at the same frequency p=q=0.5, 2N=100, no migration, no selection

Larger effect on heterozygosity when population size is small (2N=20)



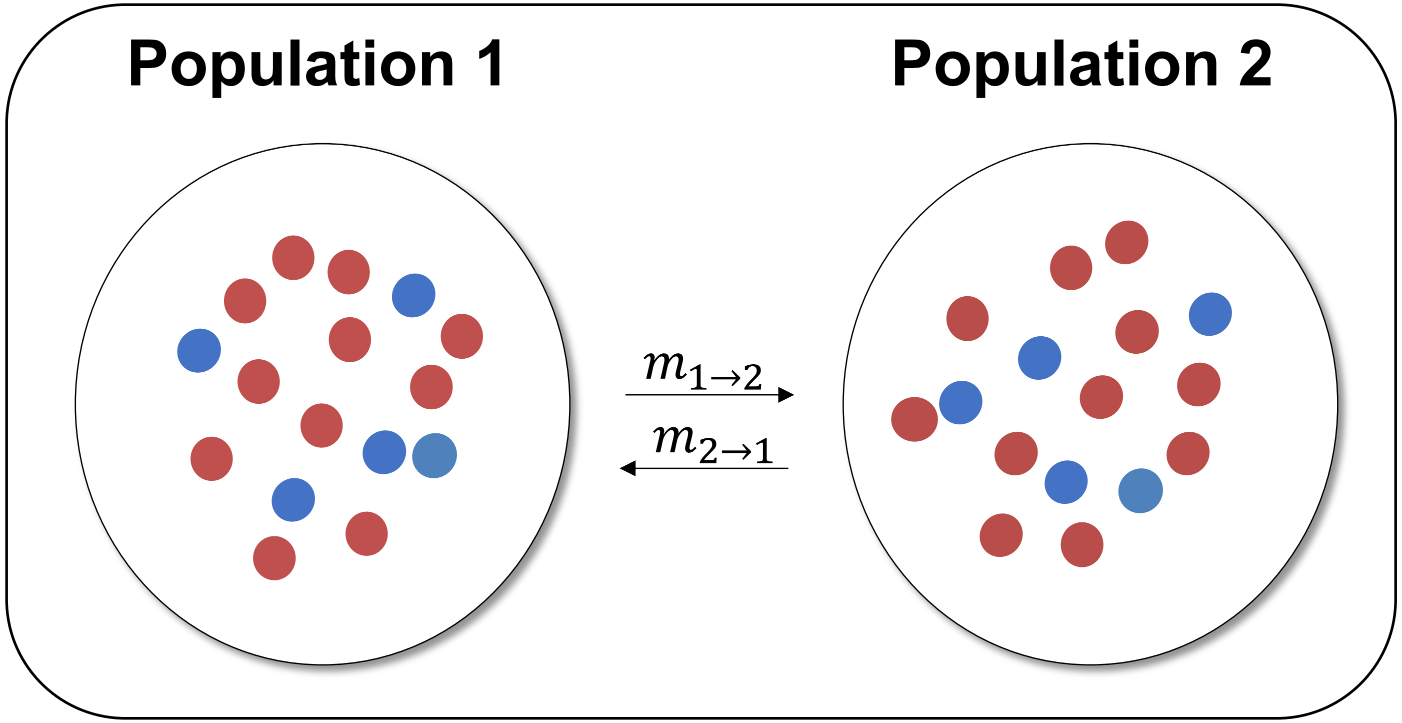
Two populations starting at the same frequency p=q=0.5, 2N=20, no migration, no selection

Larger effect on heterozygosity when population size is small (2N=20)



Two populations starting at the same frequency p=q=0.5, 2N=20, no migration, no selection

How does F_{ST} change with migration between populations?



Wright-Fisher model with migration:

- 2 population model
- Each population fits a W-F model
- Occasionally an individual from one population is replaced with an individual from the other population

Wright-Fisher model with migration

Two populations (1 and 2)

Allele frequencies in the two populations at time *t*+1:

Frequency of A
in pop1 at time t

Frequency of A
in pop2 at time t

$$E[f_{A1}(t+1)] = (1 - m_{2 \to 1}) f_{A1}(t) + m_{2 \to 1} f_{A2}(t)$$

Probability of
no migration

Probability of
migration from
pop2 into pop1

M = number of migrants per generation = 2Nm

Wright-Fisher model with migration

Two populations (1 and 2)

Allele frequencies in the two populations at time *t*+1:

$$E[f_{A1}(t+1)] = (1 - m_{2\,\rightarrow\,1})f_{A1}(t) + m_{2\,\rightarrow\,1}f_{A2}(t)$$

$$E[f_{A2}(t+1)] = (1 - m_{1\,\rightarrow\,2})f_{A2}(t) + m_{1\,\rightarrow\,2}f_{A1}(t)$$

Wright-Fisher model with migration

Two populations (1 and 2)

Allele frequencies in the two populations at time *t*+1:

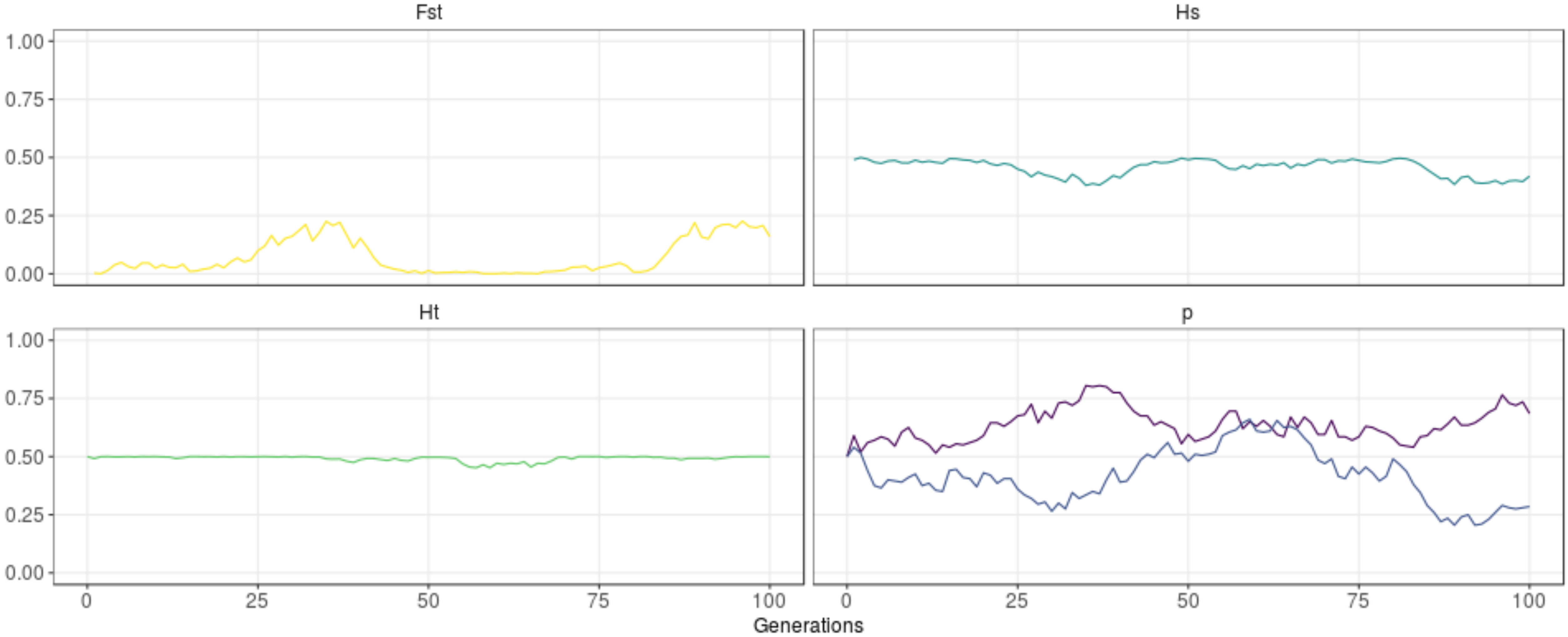
$$E[f_{A1}(t+1)] = (1 - m_{2\,\rightarrow\,1})f_{A1}(t) + m_{2\,\rightarrow\,1}f_{A2}(t)$$

$$E[f_{A2}(t+1)] = (1 - m_{1\,\rightarrow\,2})f_{A2}(t) + m_{1\,\rightarrow\,2}f_{A1}(t)$$

Equilibrium frequency occurs when $E[f_{A1}(t+1)] = f_{A1}(t) = f_{A1}$

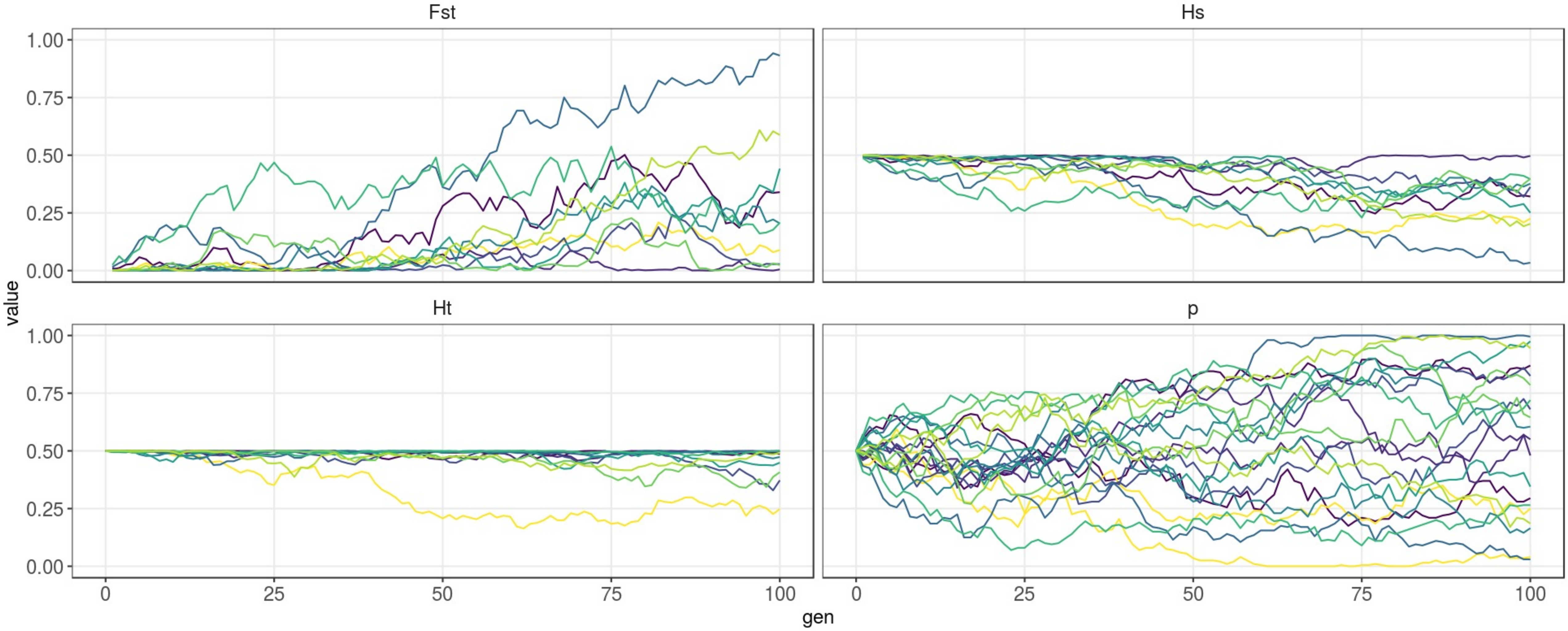
And the two islands have equal frequencies $f_{A1} = f_{A2}$

Two populations with migration



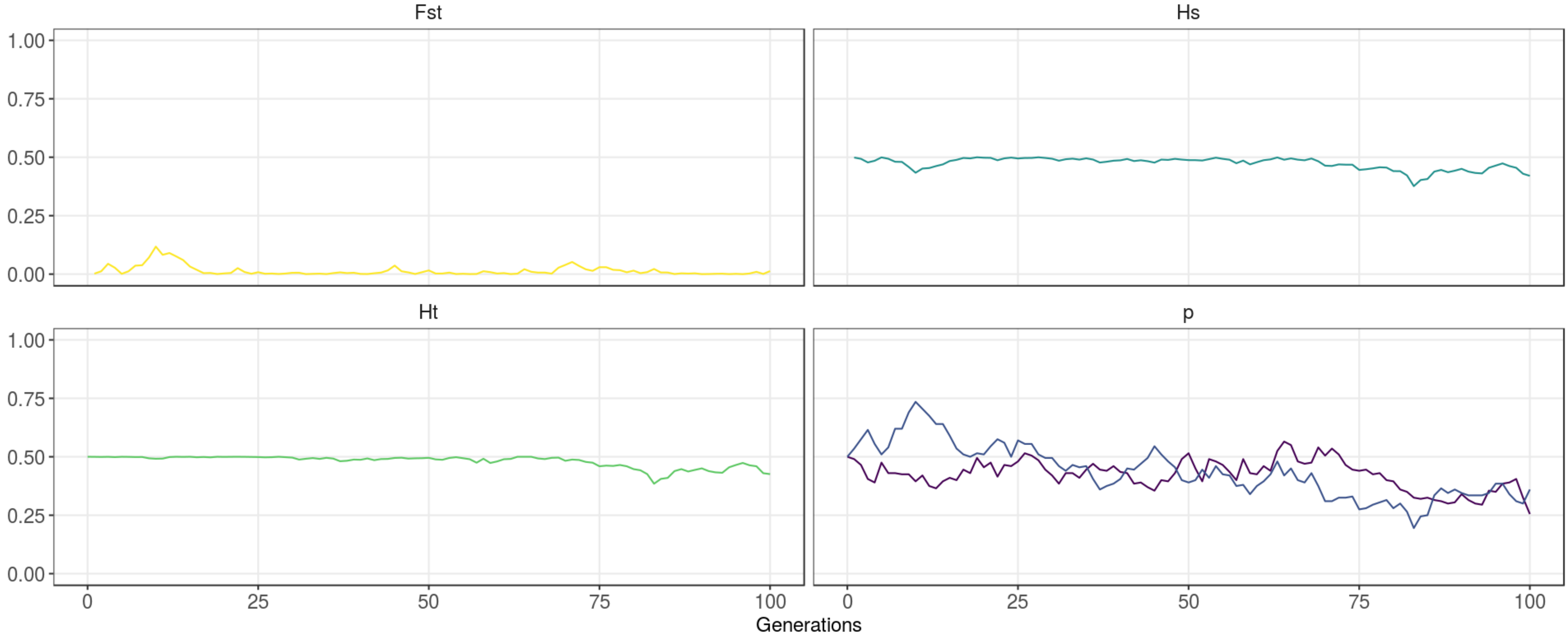
Two populations starting at the same frequency p=q=0.5, 2N=100, 0.8 migrants per generation, no selection

Two populations with migration, 10 replicates



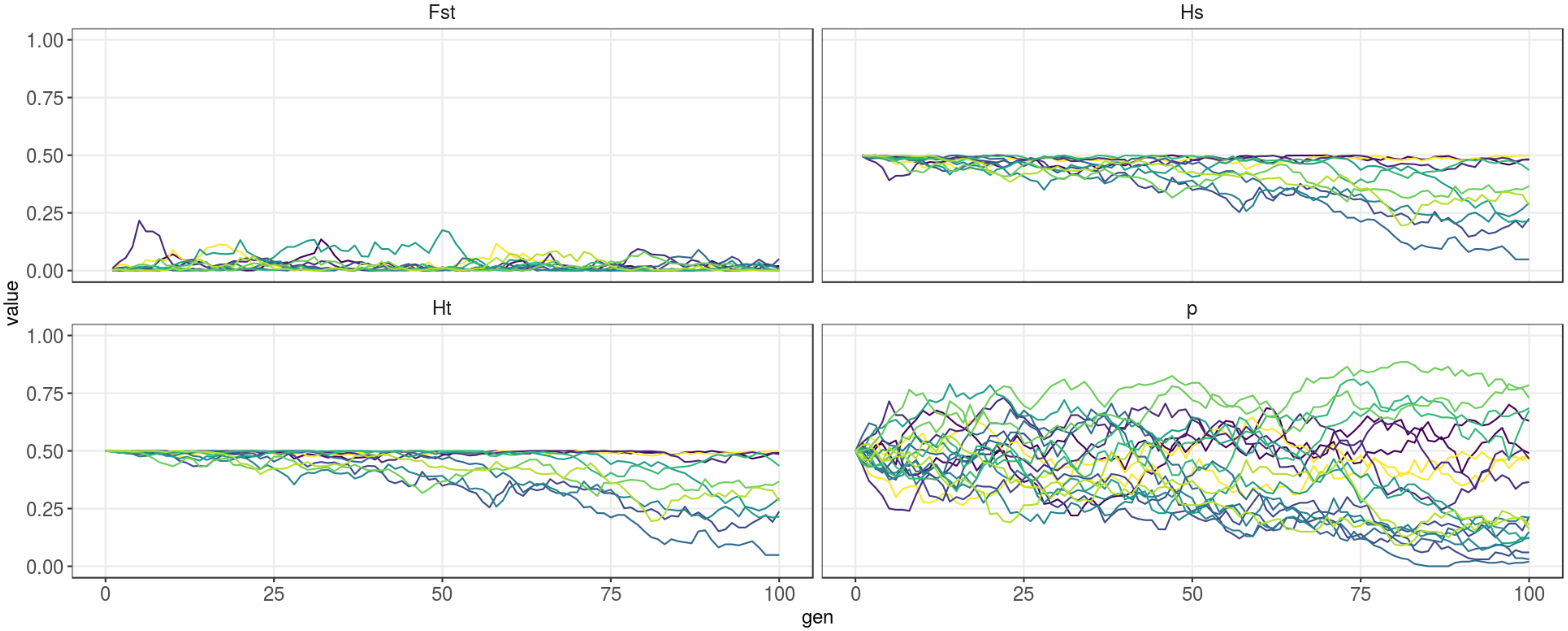
Two populations starting at the same frequency p=q=0.5, 2N=100, 0.8 migrants per generation, no selection

Two populations with high migration



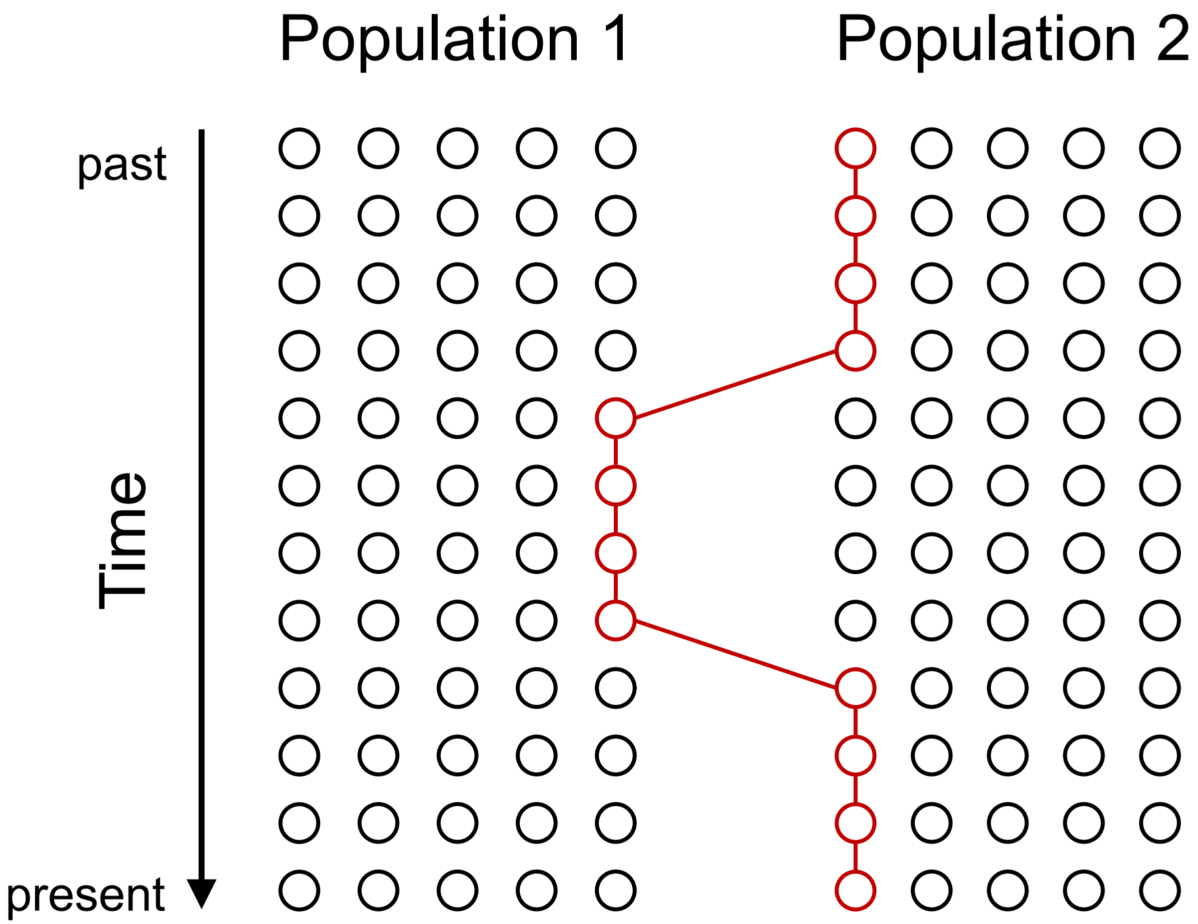
Two populations starting at the same frequency p=q=0.5, 2N=100, 10 migrants per generation, no selection

Two populations with high migration, with replicates



Two populations starting at the same frequency p=q=0.5, 2N=100, 10 migrants per generation, no selection

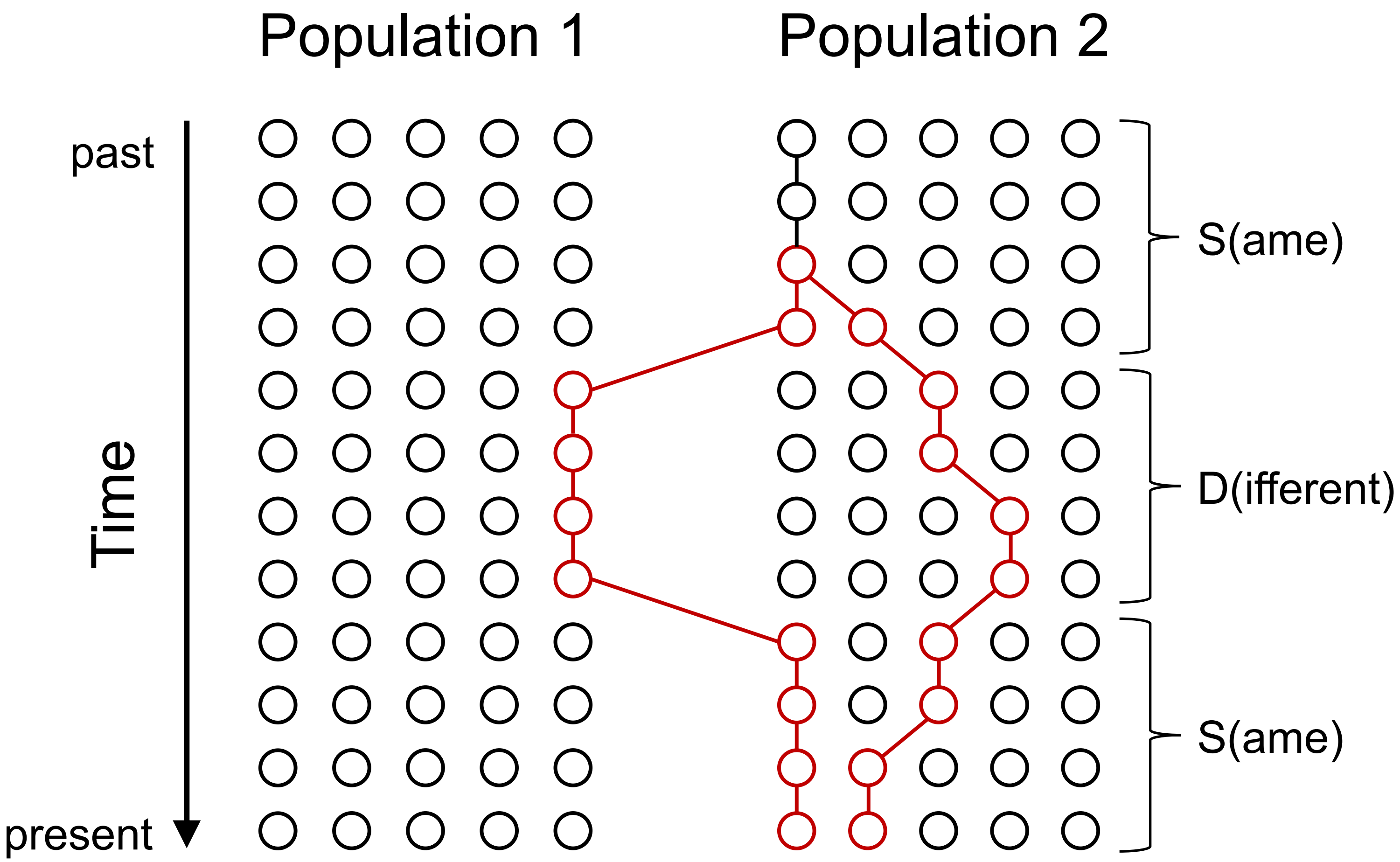
Coalescent with migration



The ancestry of a gene copy from population 2 traced backwards in time in a 2-population W-F model

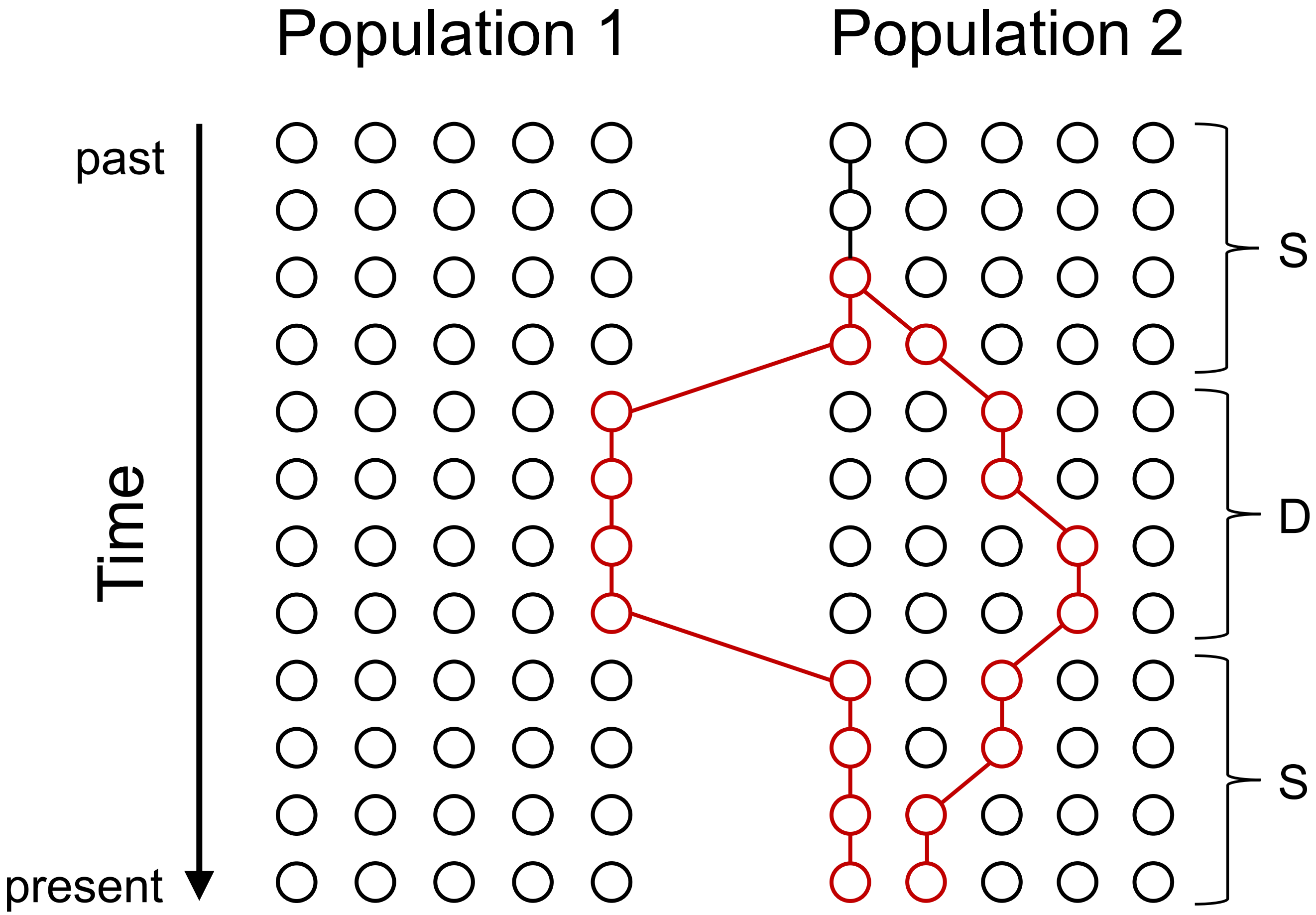
Each individual migrates at a rate M, so the expected time to the first migration event from pop1 to pop2 is the waiting time or 1/M. And the expected time in any direction is 1/2M.

Coalescent with migration



Coalescence of two gene copies sampled from population 2. Note that there are periods of time in which the ancestral gene copies are in the same (S) versus different (D) populations.

Coalescent with migration

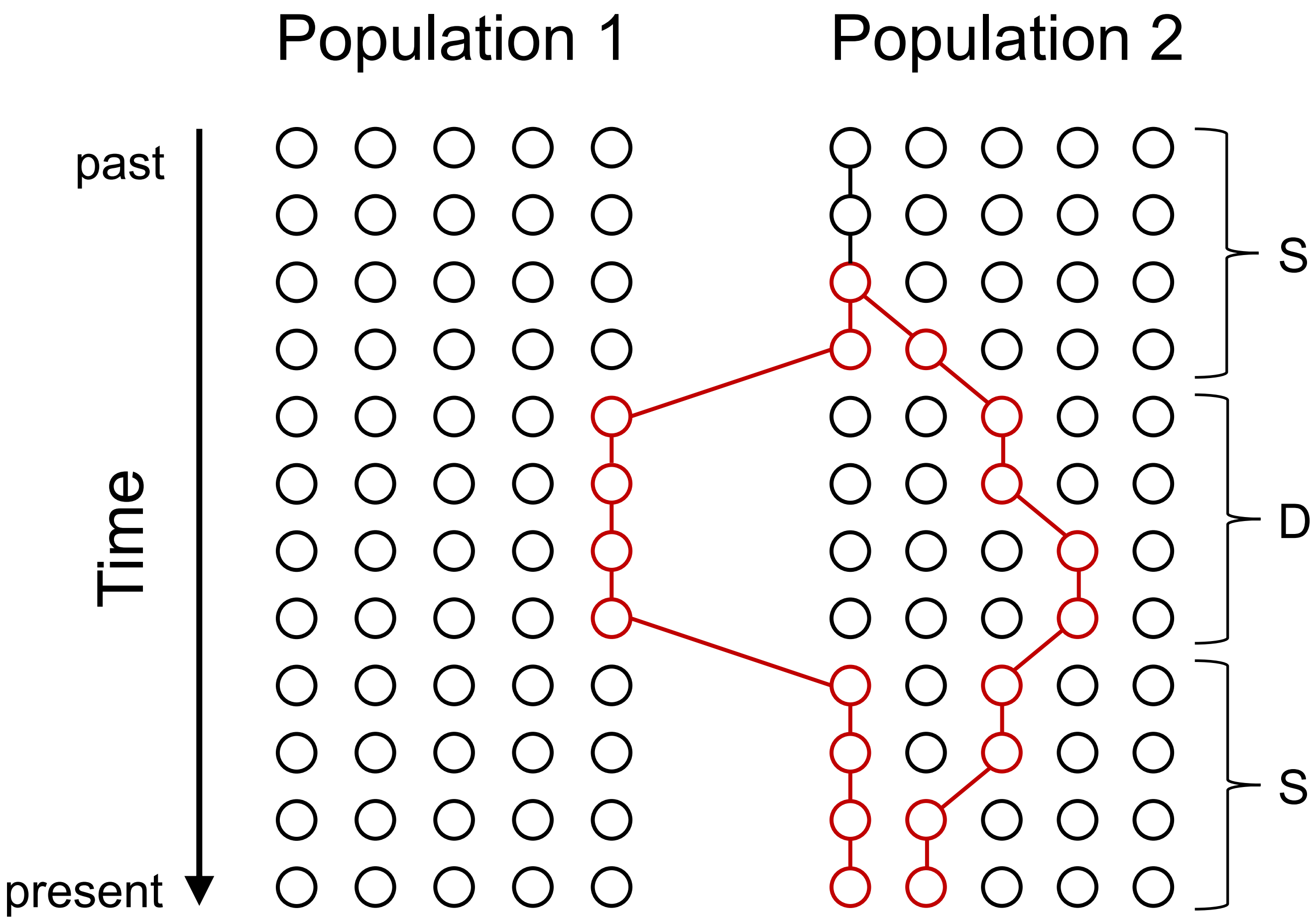


The expected coalescence time for two lineages that are in different populations is:

$$E_D[t] = \frac{1}{2M} + E_S[t]$$

The waiting time is the sum of the time it takes to migrate and the time to coalesce once in the same population

Coalescent with migration



The expected coalescence time for two gene copies in the same population (in units of 2N) is:

$$E_S[t] = 2$$

4N generations to coalesce

And the expected time for two copies in different populations (in units of 2N) is:

$$E_D[t] = \frac{1}{2M} + 2$$

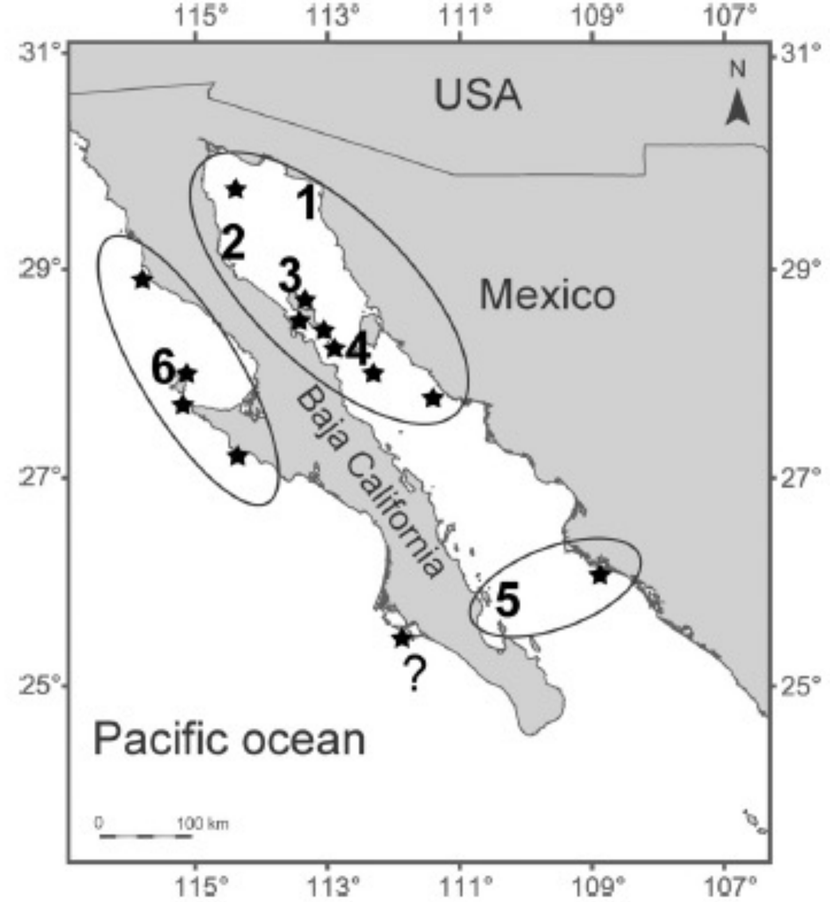
4N + N/M generations to coalesce

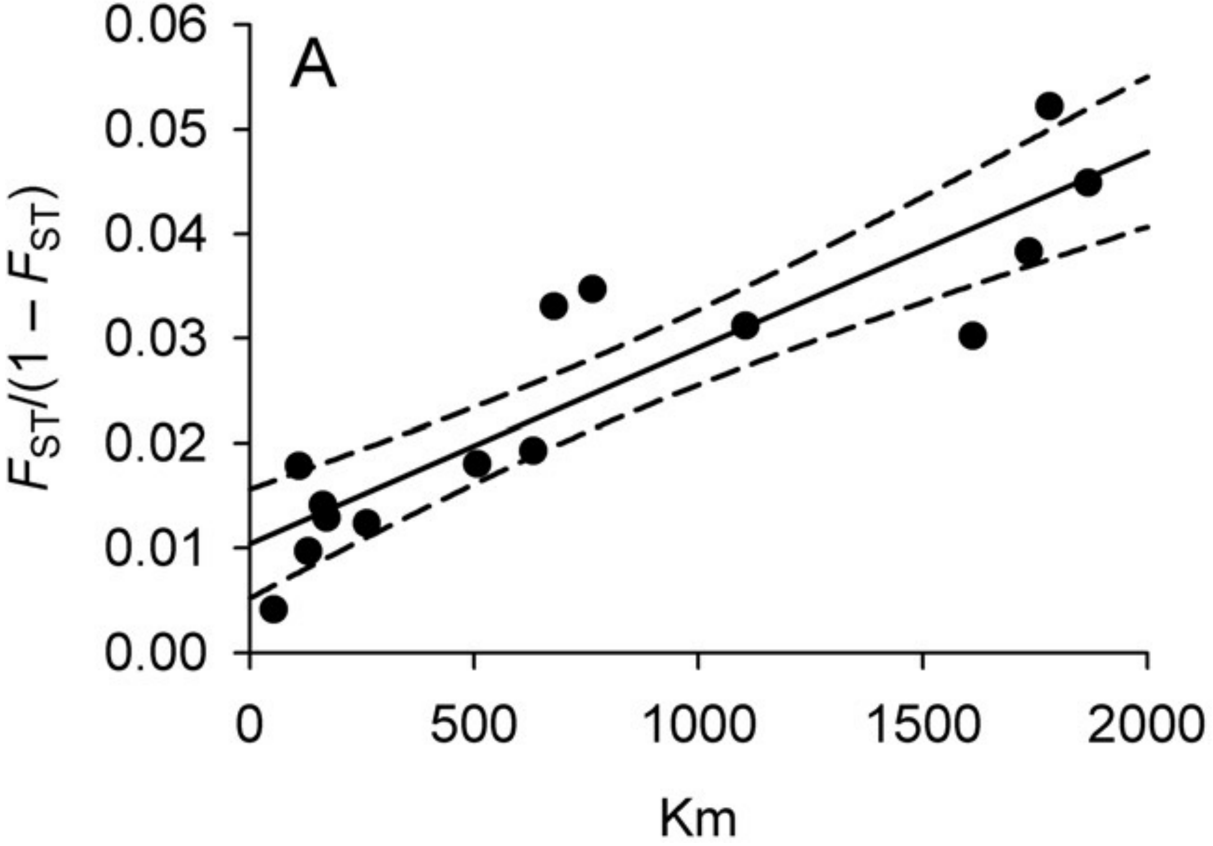
So, the expected coalescence time is longer for alleles in different populations

Isolation by distance

- As populations migrate away from their origin, divergence is expected to increase
- More complex patterns are possible depending on population history, but it is possible to test for a pattern of isolation by distance
- In cases of isolation by distance, the degree of population subdivision increases with physical distance from the species origin

Isolation by distance in sea lions





Gonzalez-Suárez et al., 2009

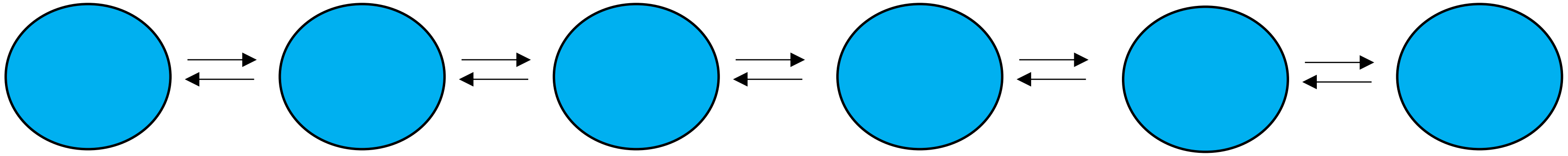
What model(s) can explain such a pattern?

Some possibilities

- A stepping stone model, in which migration occurs only between pairs of adjacent populations
- Sequential founder effect model
- Divergence followed by secondary contact (admixture between diverged groups) can cause variation to appear to be continuous

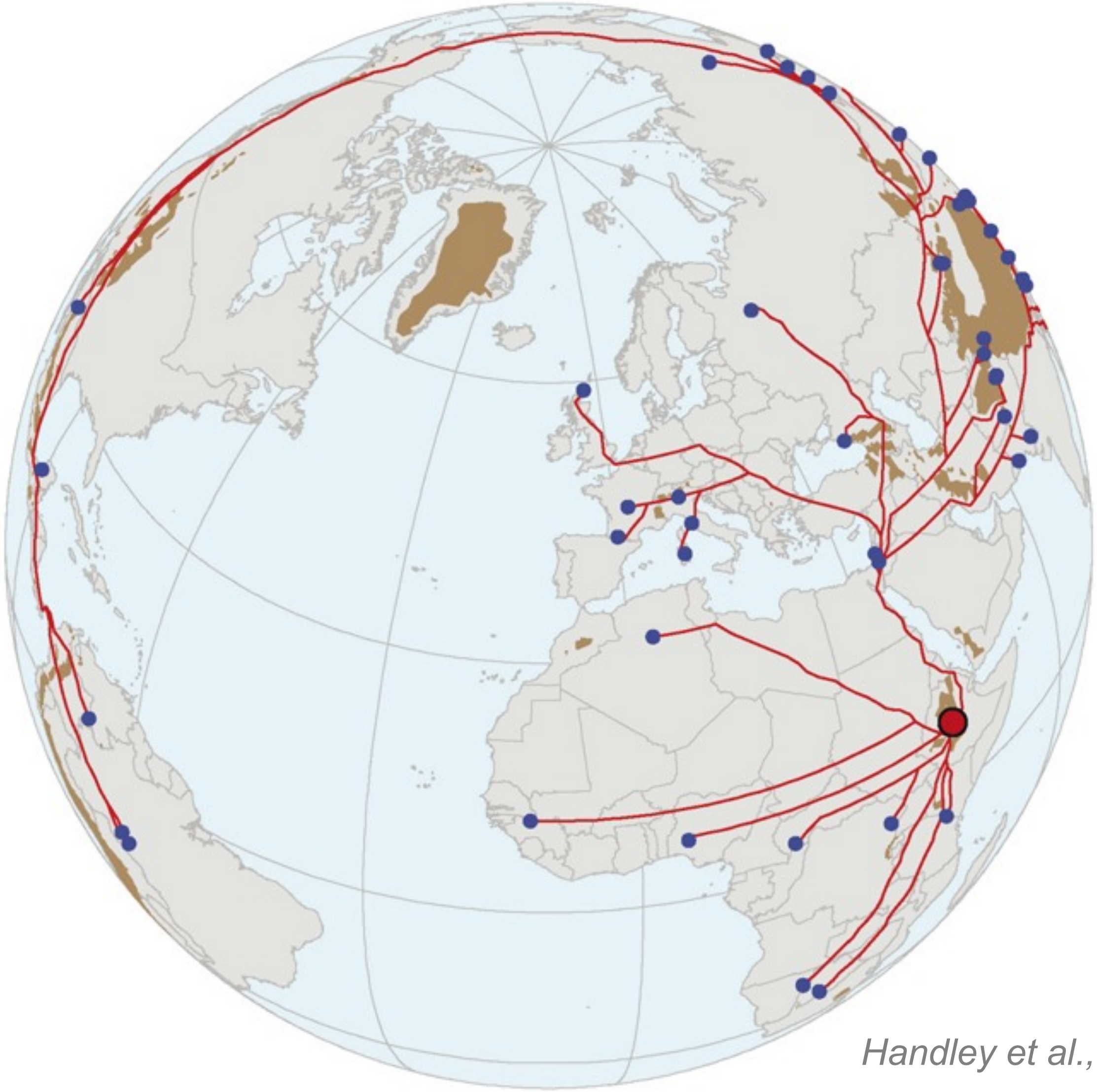
Stepping stone model

Migration among nearest neighbor sub-populations



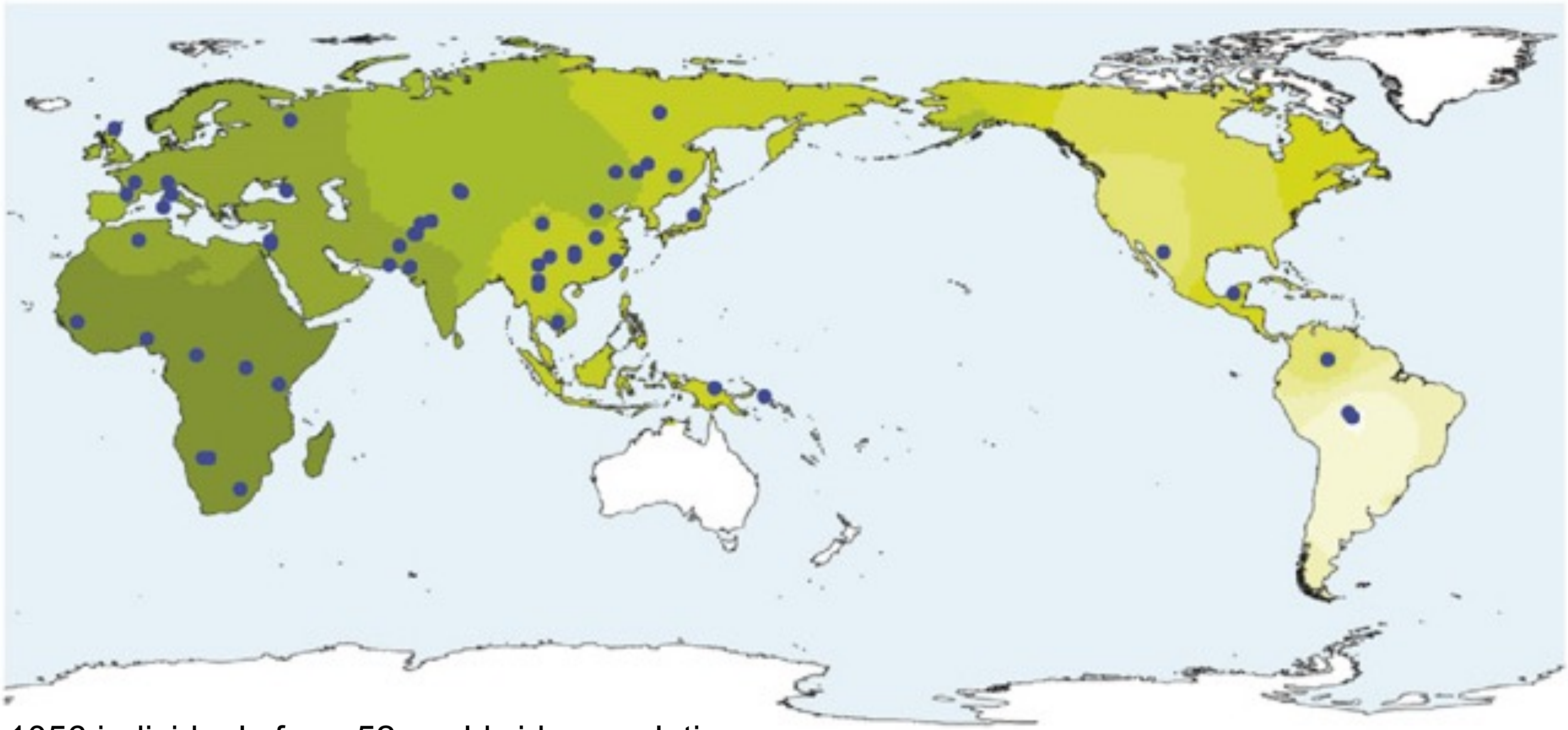
Stepping stone models assume that M is a function of distance between populations

Isolation by distance in human populations



Handley et al., Trends in Genetics, 2007

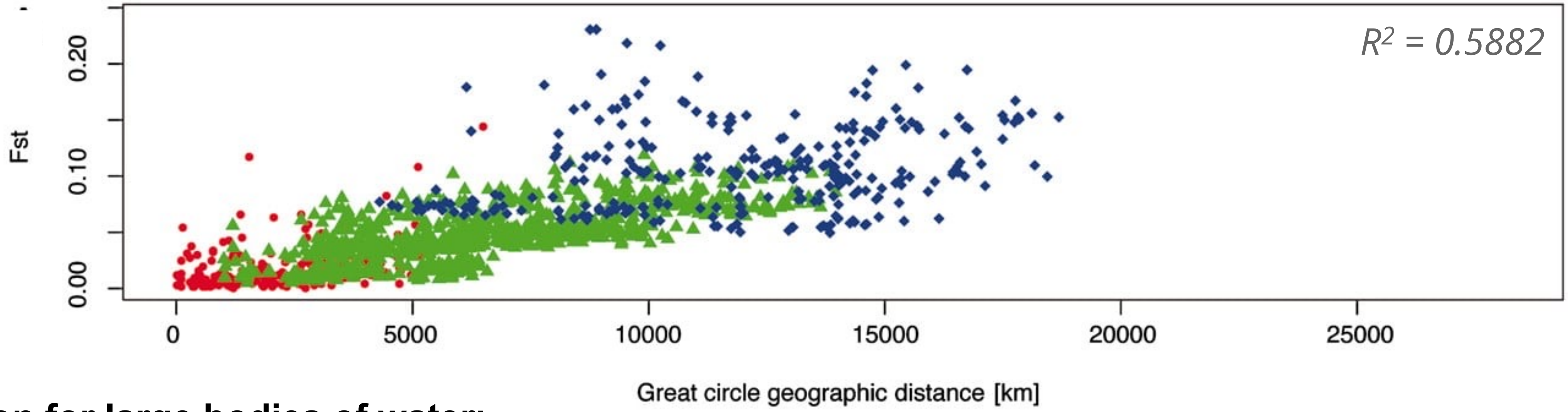
HGDP: Human Genome Diversity Panel



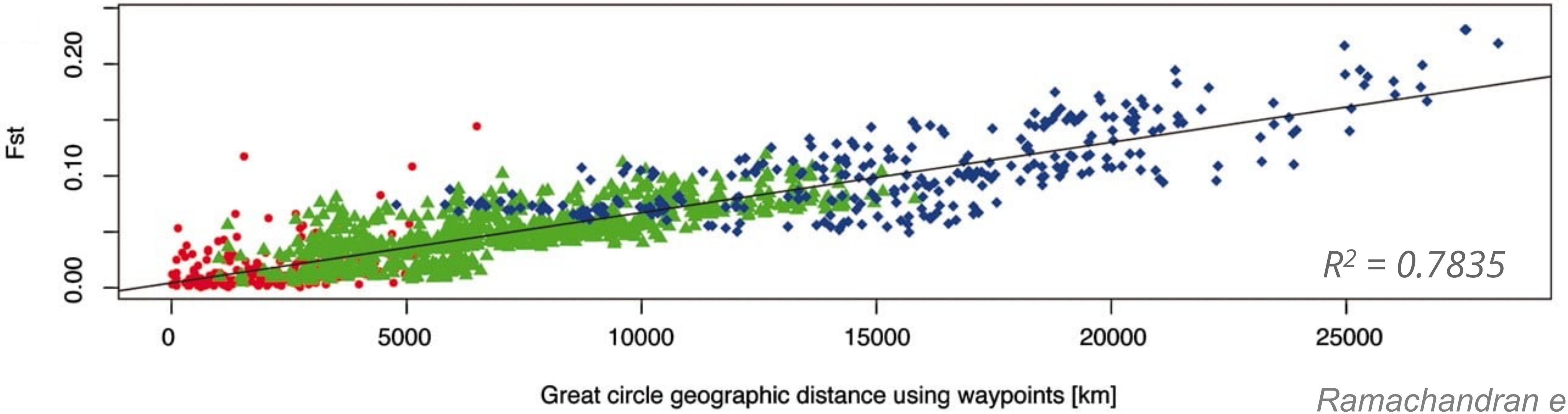
1056 individuals from 52 worldwide populations

TRENDS in Genetics

Isolation by distance in human populations

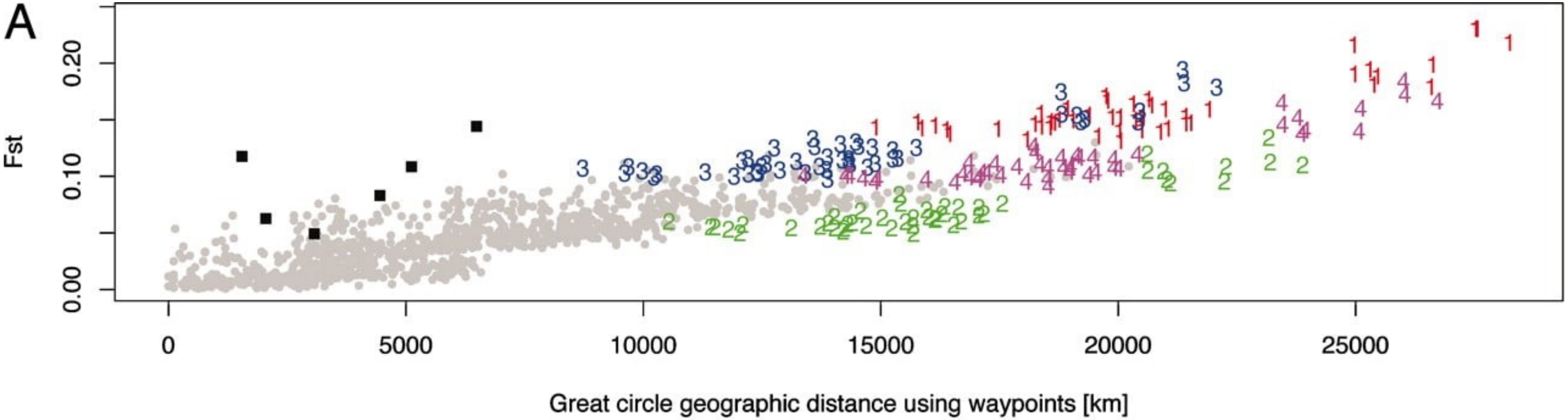


Correction for large bodies of water:



Ramachandran et al., 2005

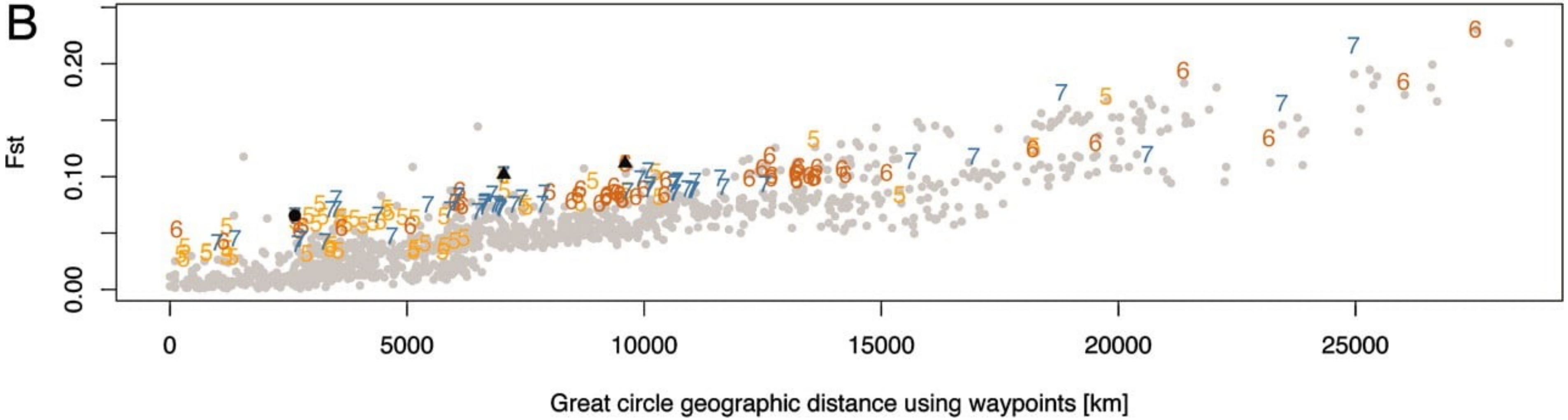
Influence of individual populations on the regression



The number representing each population is the rank of its influence on the regression, with 1 indicating the population whose removal from the data alters the regression by the greatest amount (see Materials and Methods and Table 2). All other points not involving comparisons with the populations of greatest influence are in gray. (A) Red 1 denotes comparisons including Karitiana; green 2, Maya; navy blue 3, Pima; and purple 4, Colombia. Black squares show comparisons between the American populations. Comparisons involving the Maya (labeled as 2) tend to produce smaller F_{ST} values than are predicted by the regression line, and excluding the Maya from analysis increases R^2 to 0.8183. The slight increase in the error sum of squares of the regression when the Maya are included in the data set shows that they have little influence on the observed pattern.

Ramachandran et al., 2005

Influence of individual populations on the regression

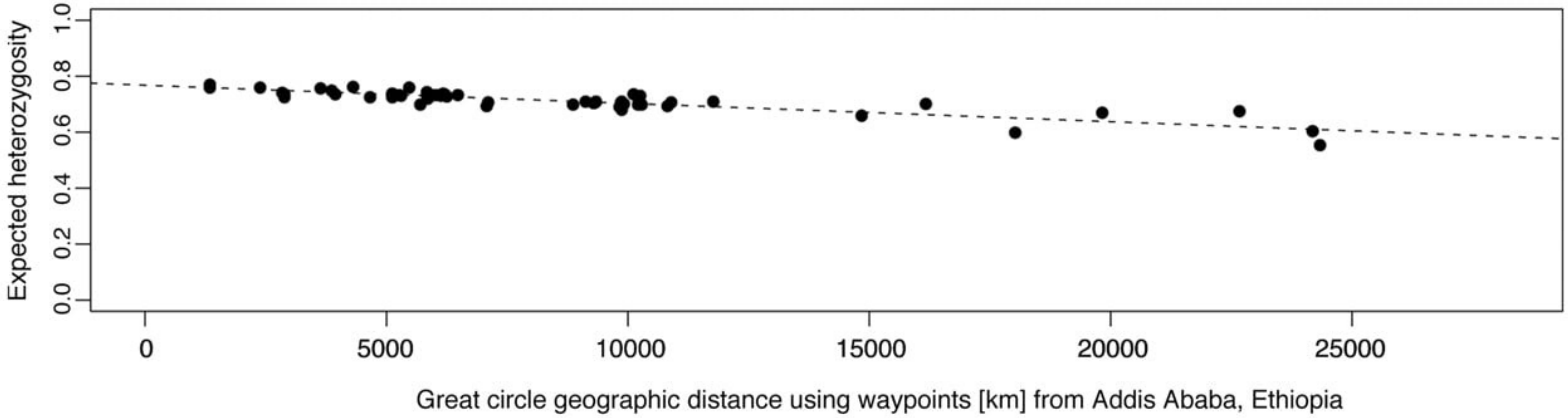


Orange 5 denotes comparisons including Kalash; brown 6, San; and blue 7, Mbuti Pygmy. The black circle is the comparison between the San and Mbuti Pygmies. The black triangles are comparisons of the Kalash to the San and Mbuti. The Kalash have been identified as a genetic isolate from the rest of Pakistan; here, comparisons of the Kalash with other Central/South Asian and East Asian populations produce large residuals, whereas comparisons with European and Middle Eastern groups do not, consistent with the closer relationships of the Kalash to groups in these regions than to groups in East Asia or to other groups in Pakistan. The high F_{ST} values observed in comparisons with the Mbuti Pygmies or the San, both hunter-gatherer populations, are likely to be a consequence of the deep genetic structure believed to exist in Africa and of the amount of genetic isolation these groups have experienced from other African populations.

Ramachandran et al., 2005

Heterozygosity declines with distance from Ethiopia

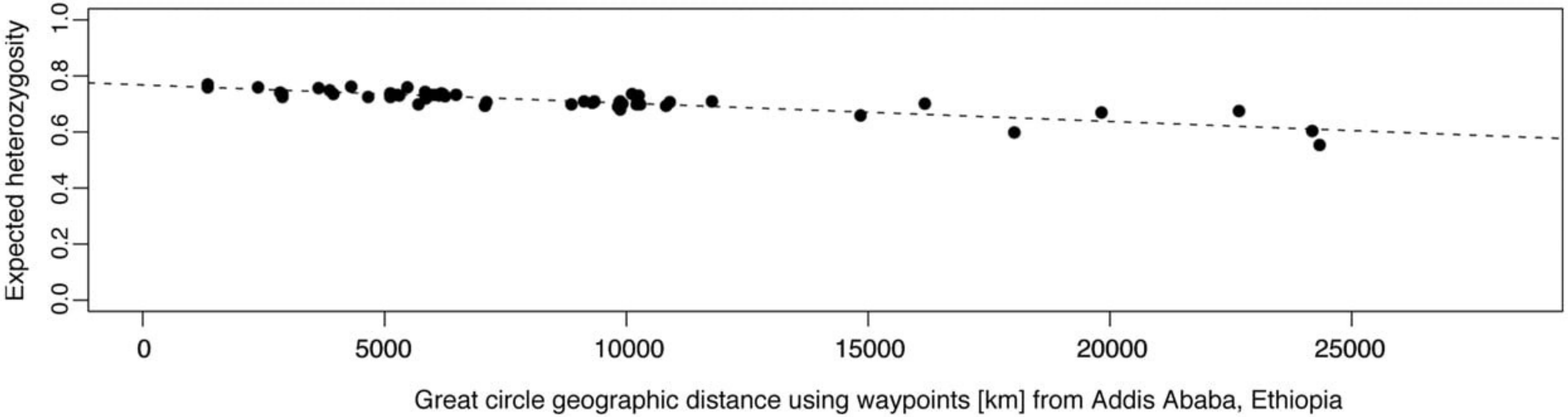
Data



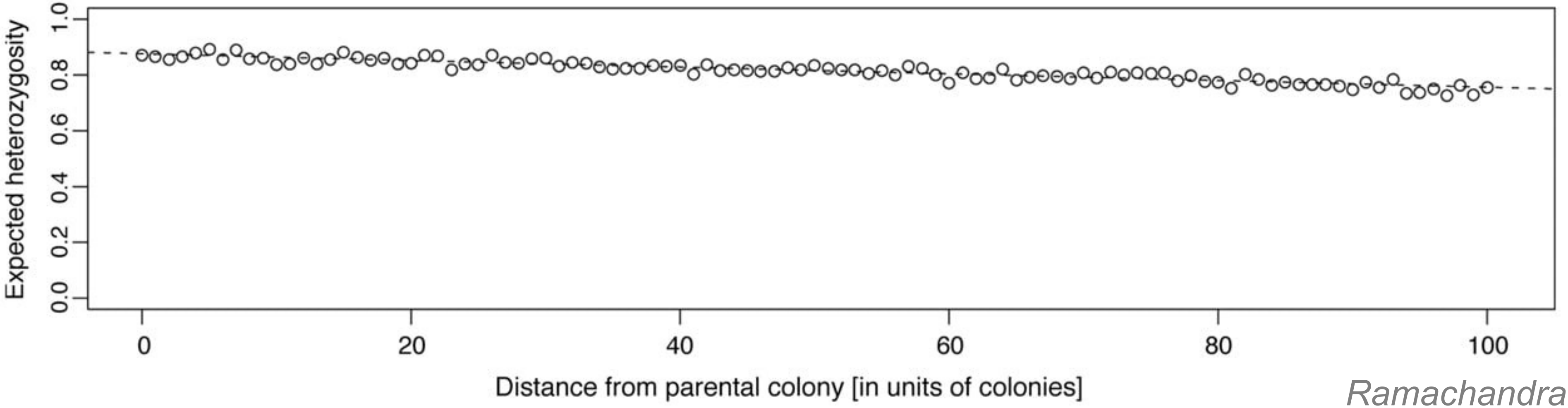
Ramachandran et al., 2005

Heterozygosity declines with distance from Ethiopia and fits expectations from simulations

Data



Simulations



Ramachandran et al., 2005

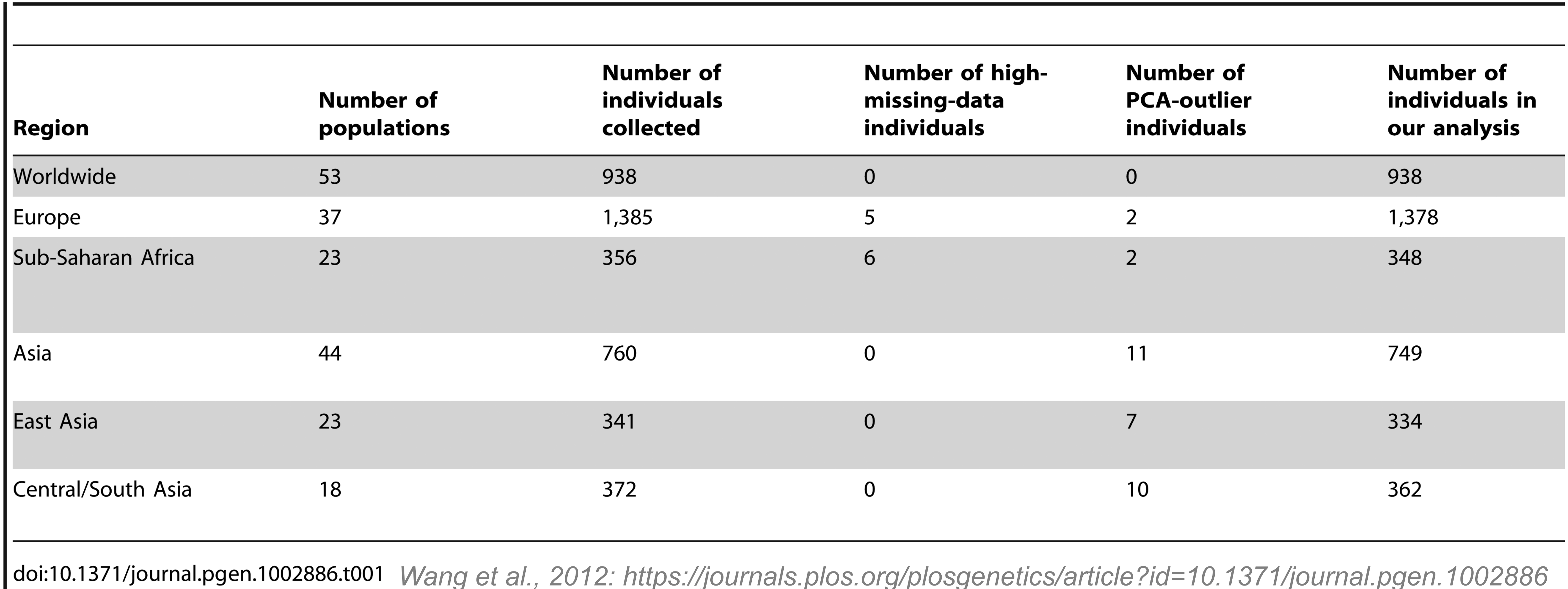
Multi-dimensional similarity between populations and its relationship to geographic distance

- In PCA, samples are projected onto a series of orthogonal axes (principal components or PCs) that are constructed from a linear combination of genotypic values across genetic markers, such that each PC sequentially maximizes the variance among samples projected on it
- A Procrustes analysis approach can be used to quantify the similarity between statistical maps of genetic variation and geographic maps
- In Wang et al., the authors apply the Procrustes approach together with PCA to study the geographic structure of human genetic variation across different geographic regions.

Wang et al., 2012: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1002886

Multi-dimensional similarity between populations and its relationship to geographic distance

Based on a common set of autosomal SNP markers shared by datasets collected from different studies, the authors evaluate the similarity between genes and geography in examples from Europe, Sub-Saharan Africa, Asia, East Asia, and Central/South Asia, as well as in a worldwide sample.

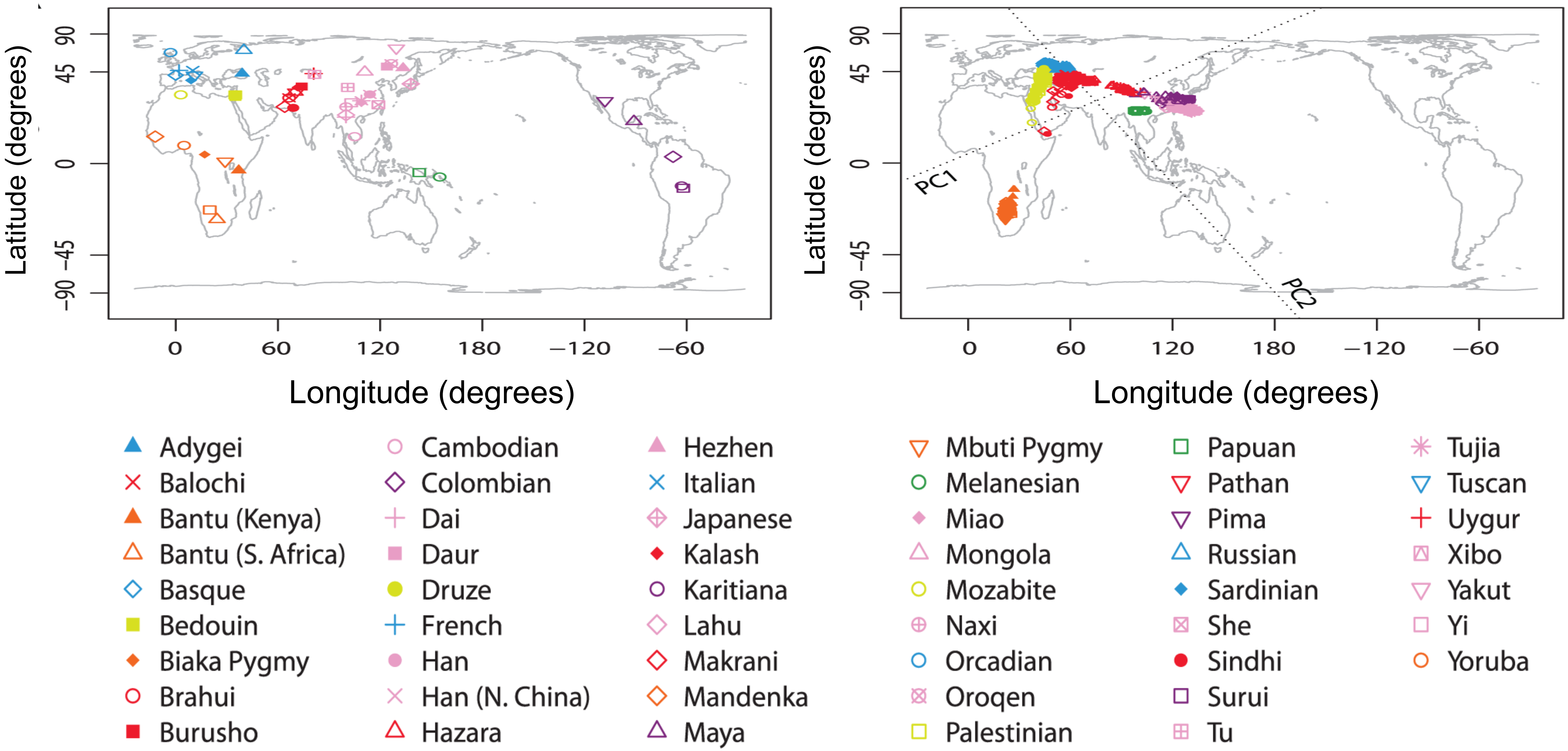


Region	Number of populations	Number of individuals collected	Number of high-missing-data individuals	Number of PCA-outlier individuals	Number of individuals in our analysis
Worldwide	53	938	0	0	938
Europe	37	1,385	5	2	1,378
Sub-Saharan Africa	23	356	6	2	348
Asia	44	760	0	11	749
East Asia	23	341	0	7	334
Central/South Asia	18	372	0	10	362

doi:10.1371/journal.pgen.1002886.t001

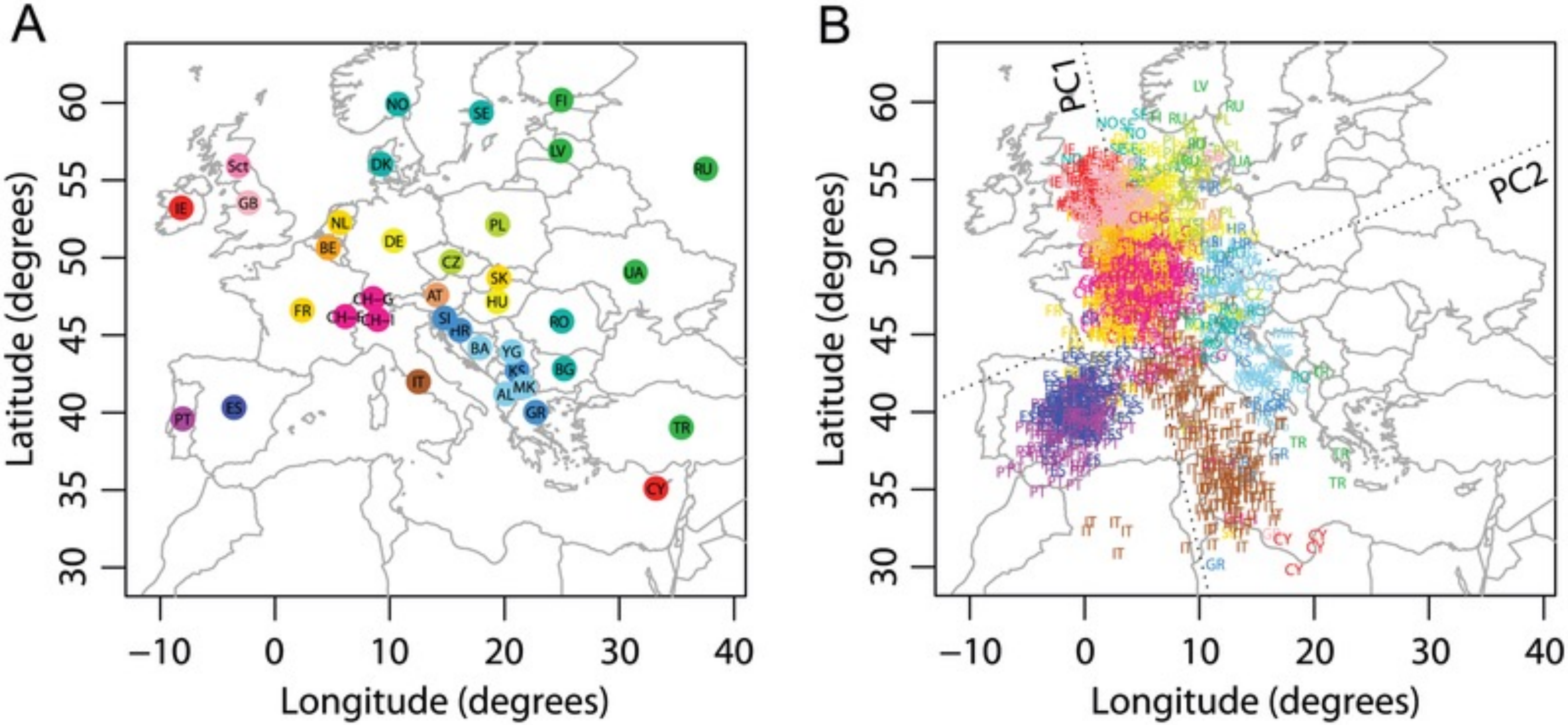
Wang et al., 2012: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1002886

Worldwide variation



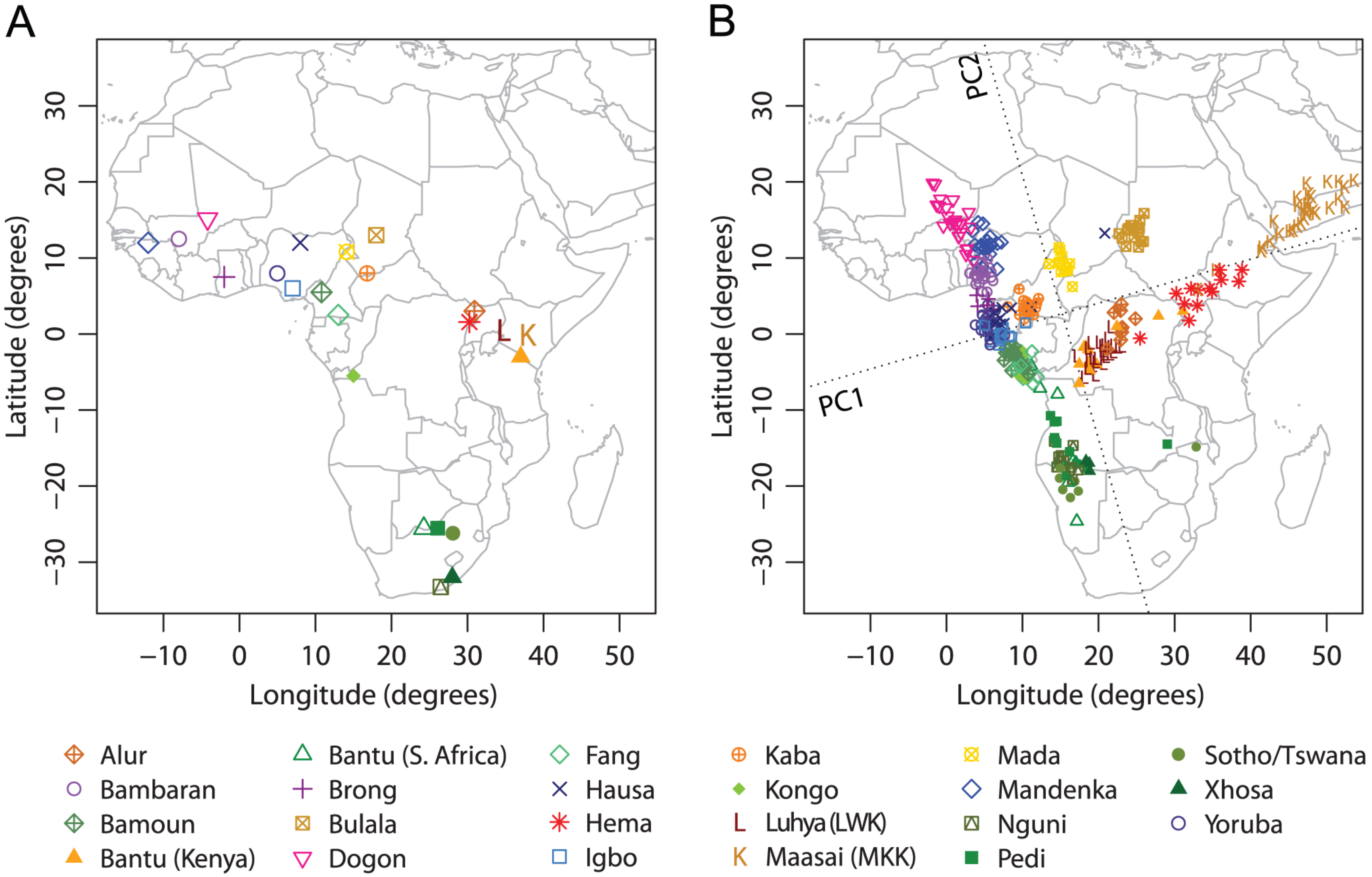
Wang et al., 2012: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1002886

Genetic variation mapped to geographic variation across Europe



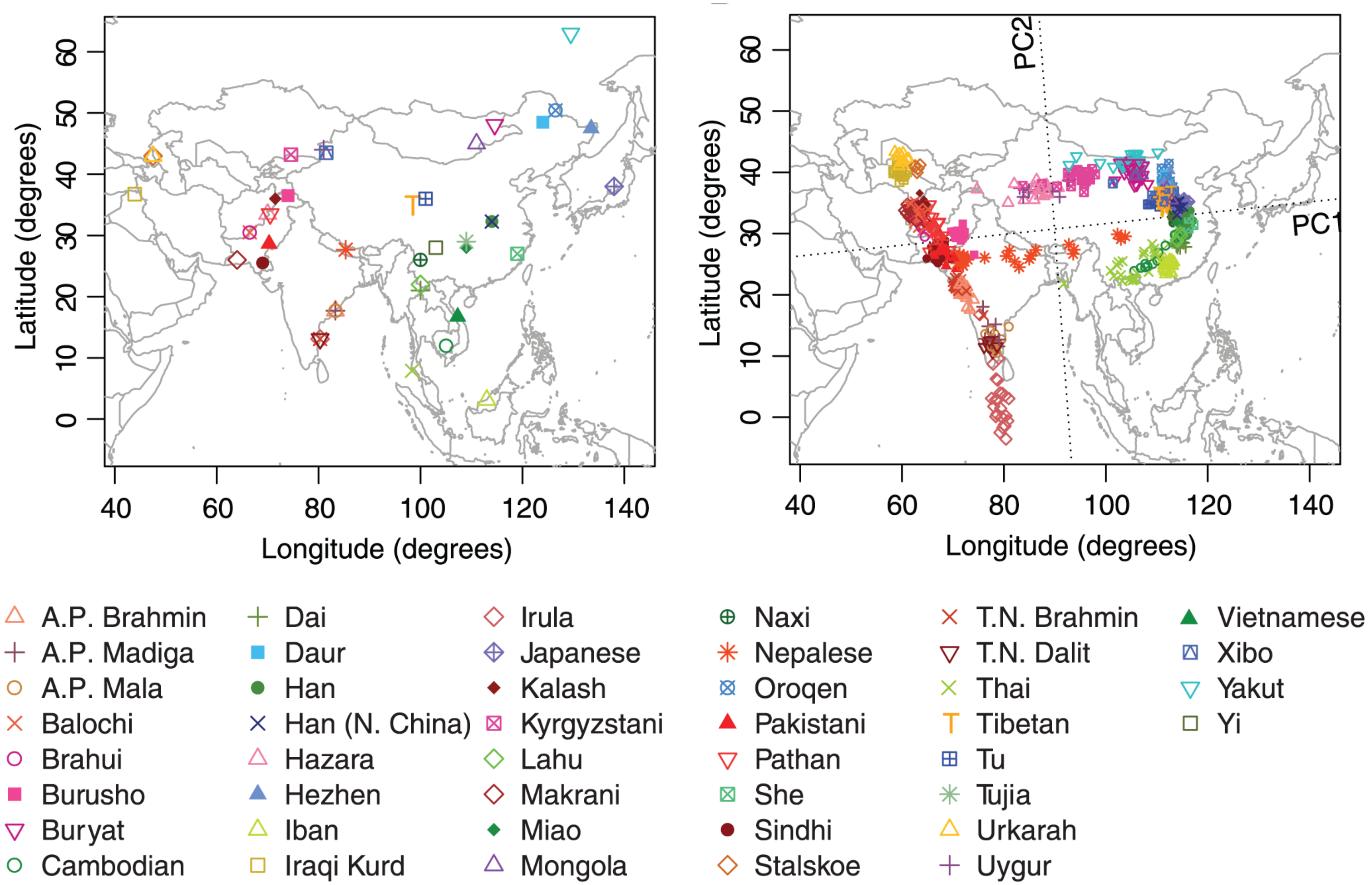
Wang et al., 2012: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1002886

Genetic variation mapped to geographic variation in Africa



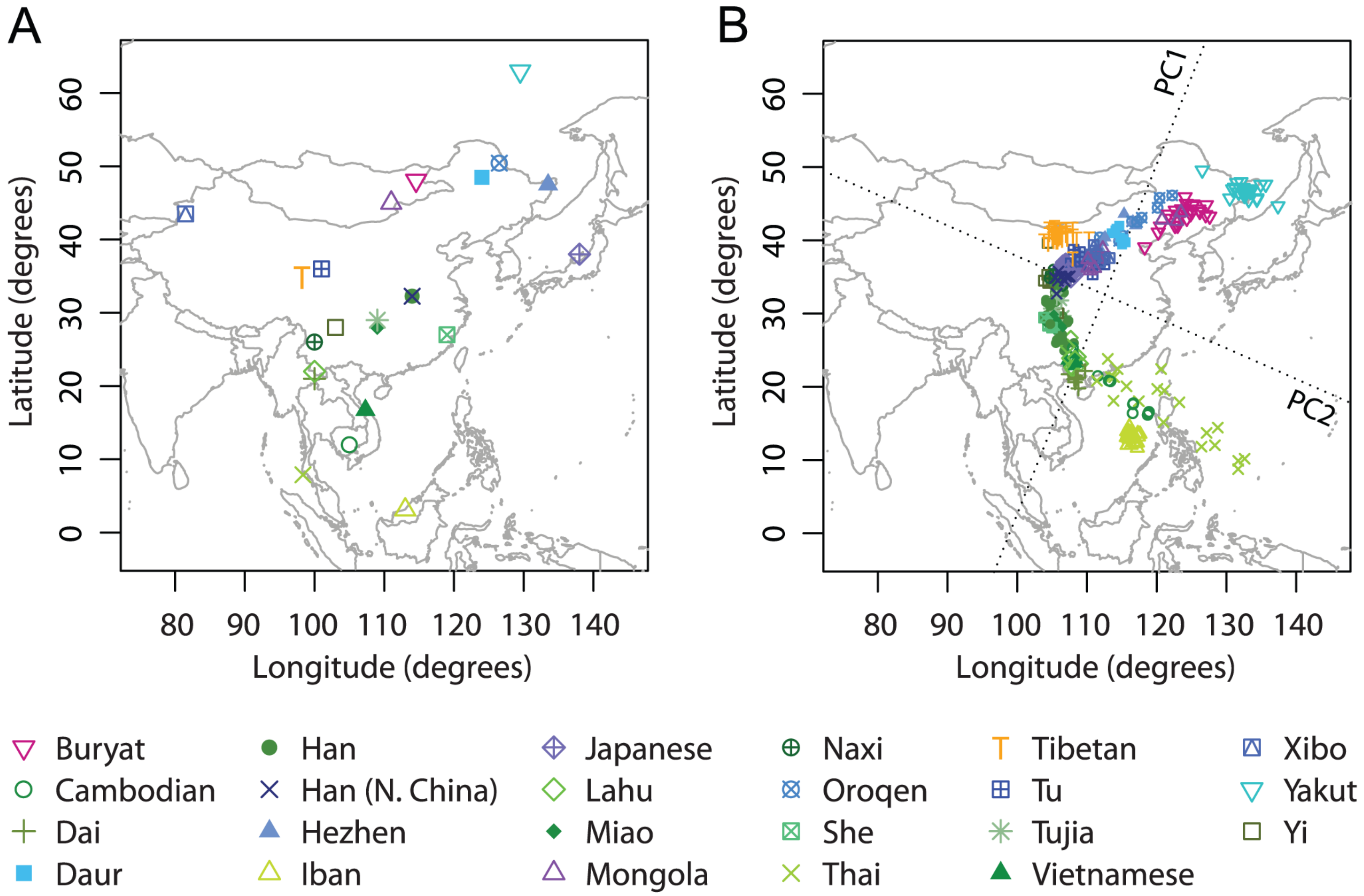
Wang et al., 2012: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1002886

Genetic variation mapped to geographic variation in Asia



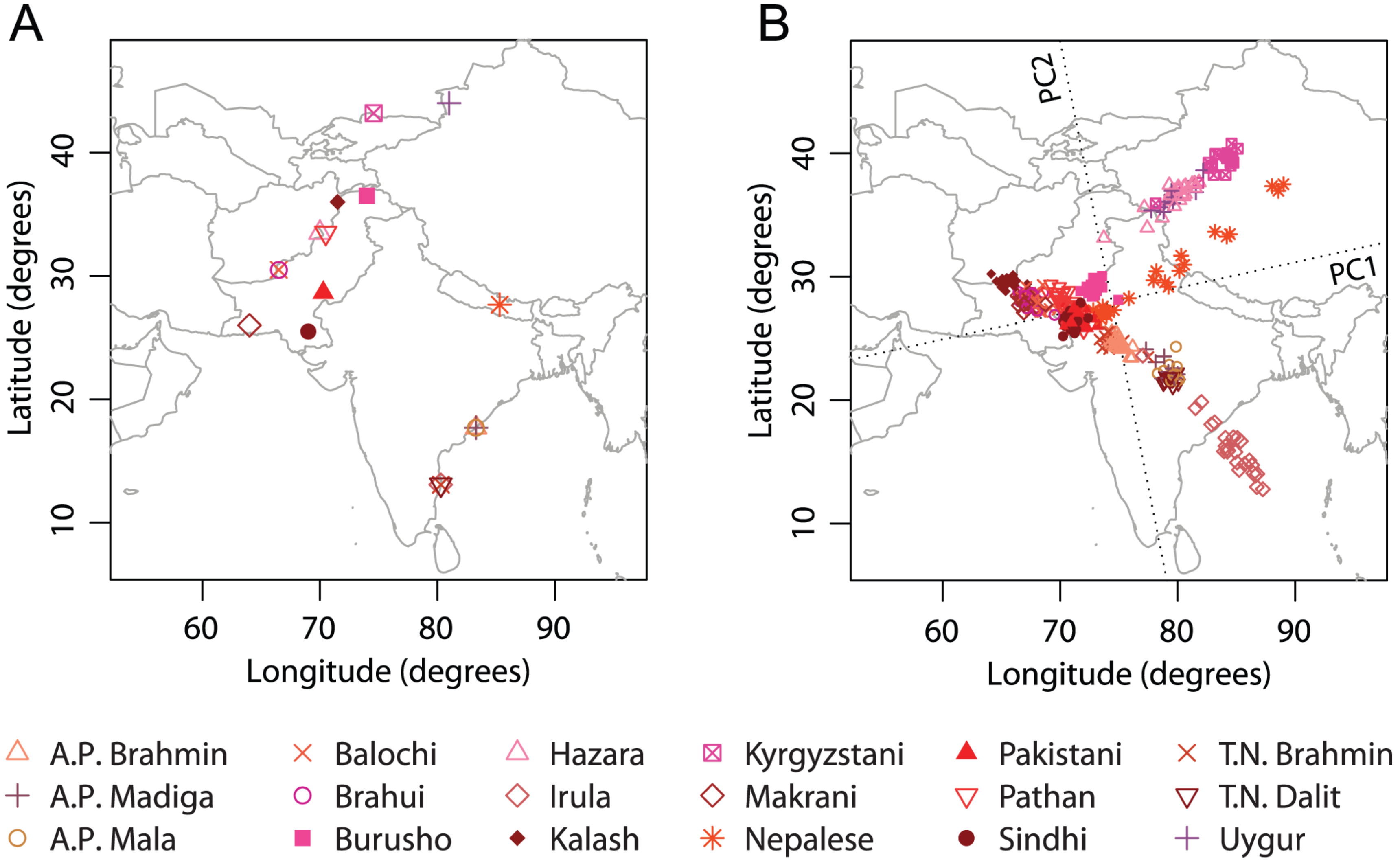
Wang et al., 2012: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1002886

Genetic variation mapped to geographic variation East Asia



Wang et al., 2012: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1002886

Genetic variation mapped to geographic variation south Asia

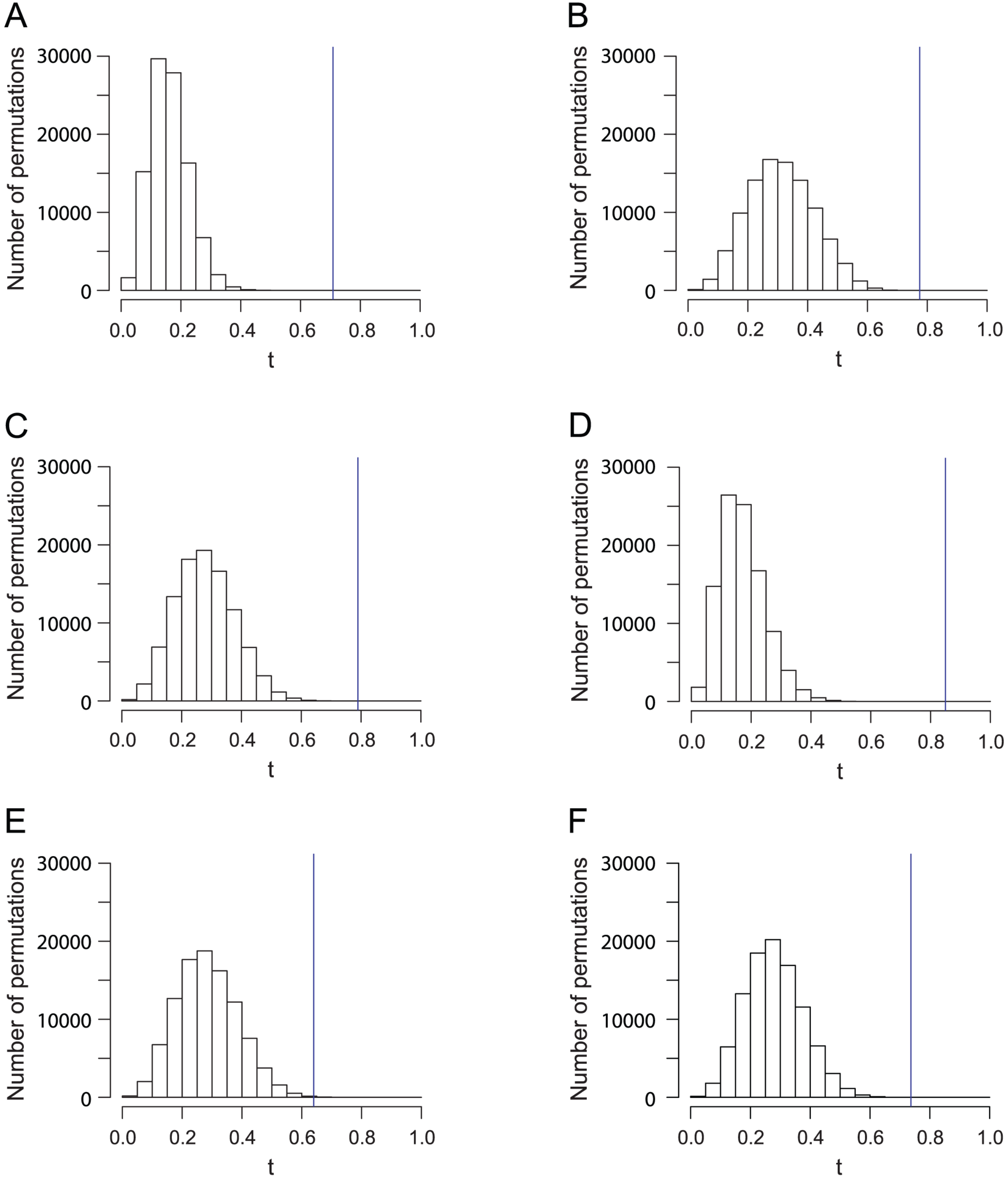


Wang et al., 2012: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1002886

Genetic and geographic distance are significantly more similar than permuted (randomized) genetic and geographic distances

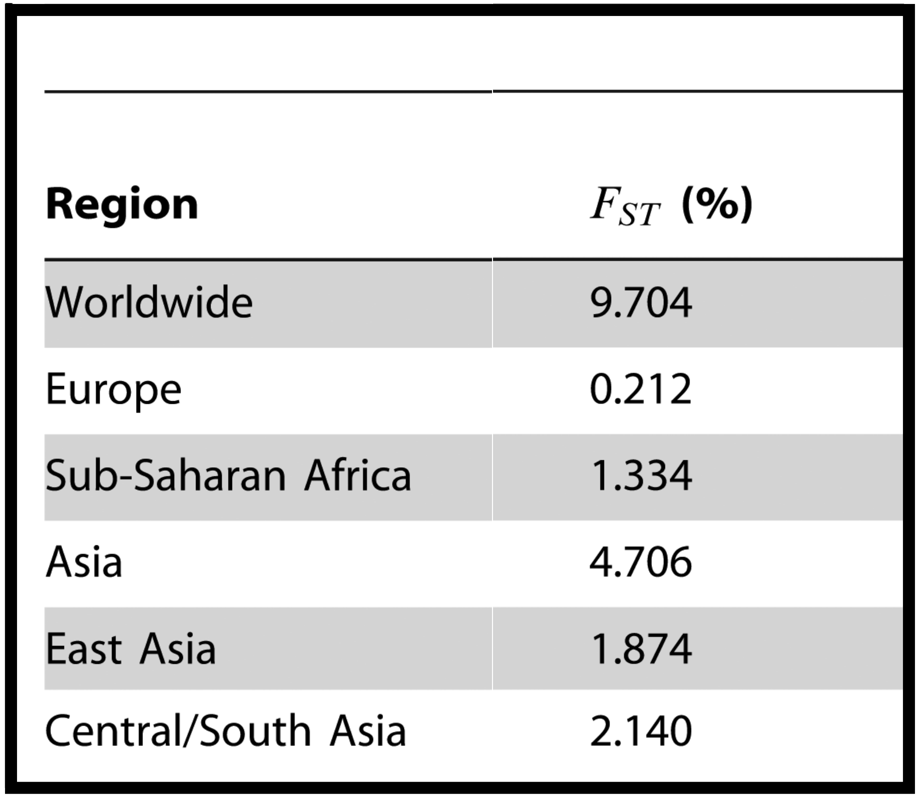
Panels refer to the following data sets:

A. Worldwide
B. Europe
C. Sub-saharan Africa
D. Asia
E. East Asia
F. Central/South Asia



https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1002886

F_{ST} in human populations is low



Region	F_{ST} (%)
Worldwide	9.704
Europe	0.212
Sub-Saharan Africa	1.334
Asia	4.706
East Asia	1.874
Central/South Asia	2.140

As a whole, human populations are more similar than they are different. Within-population differences among individuals account for 93 to 95% of genetic variation; differences among major groups constitute only 3 to 5%.

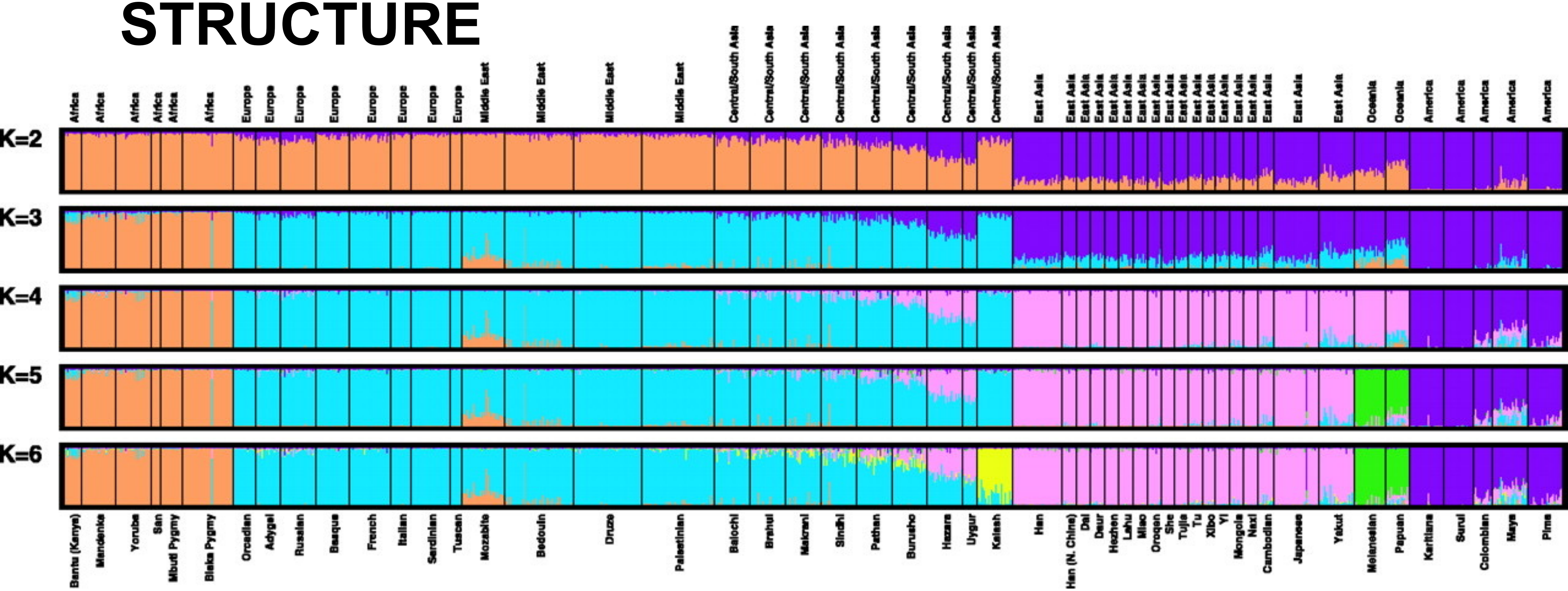
Wang et al., 2012: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1002886

Another approach: decomposing a SNP x population matrix into K factors

STRUCTURE clustering approach

- 377 autosomal microsatellite loci
- Genotyped in the HGDP

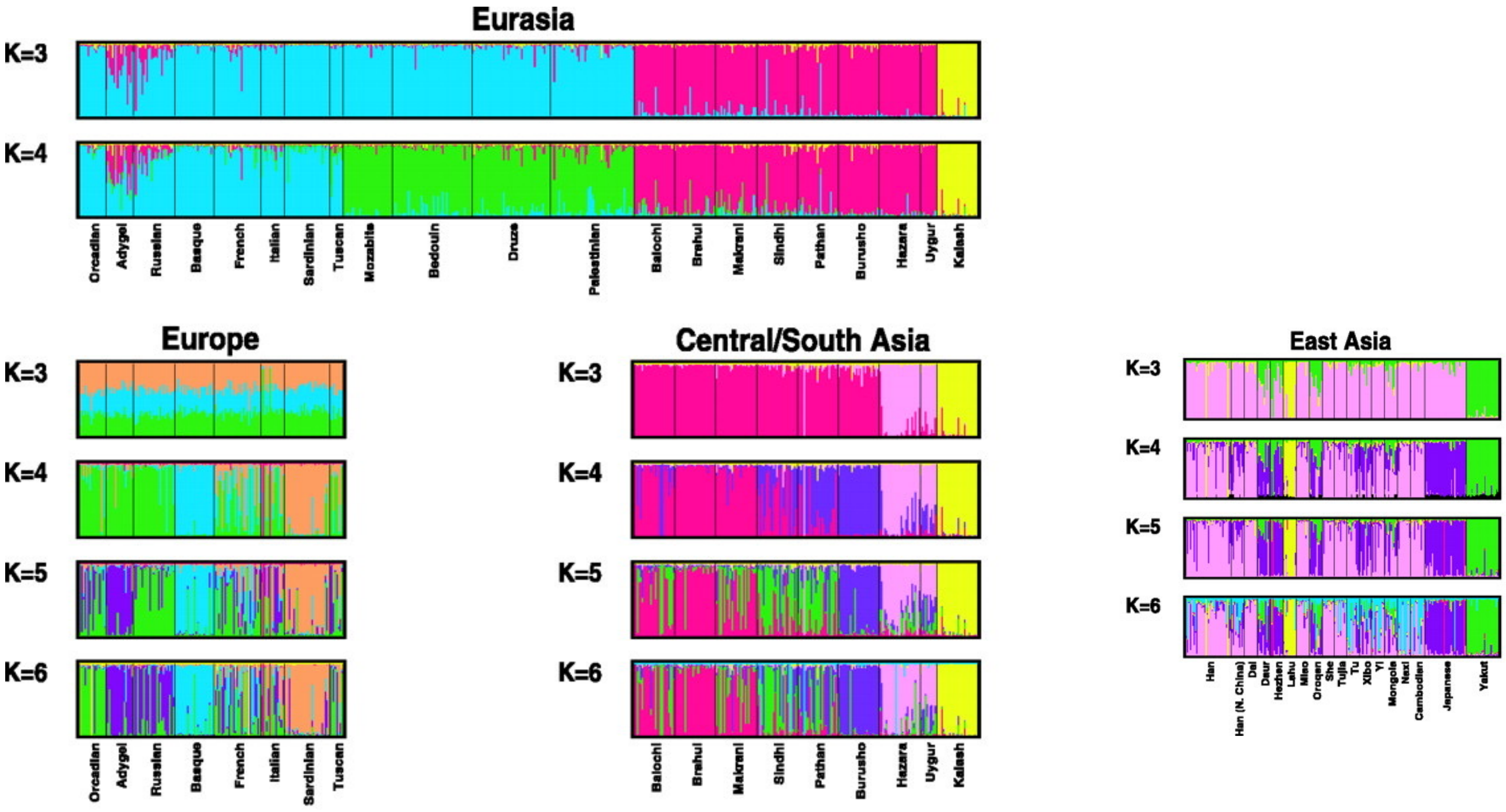
STRUCTURE



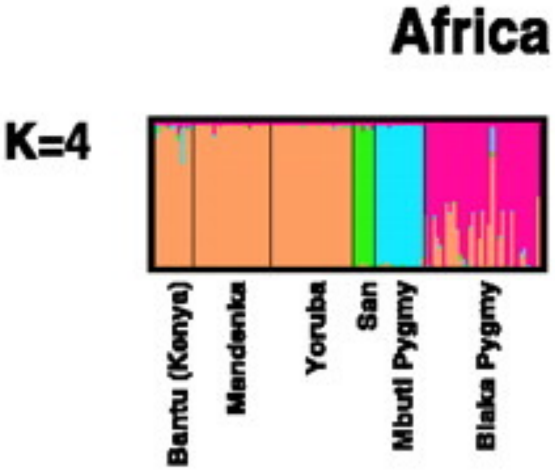
Within-population differences among individuals account for 93-95% of genetic variation
Differences among major groups account for only 3-5%
Even with the small proportion of differences between groups, with a large number of loci major geographic regions separate into difference clusters

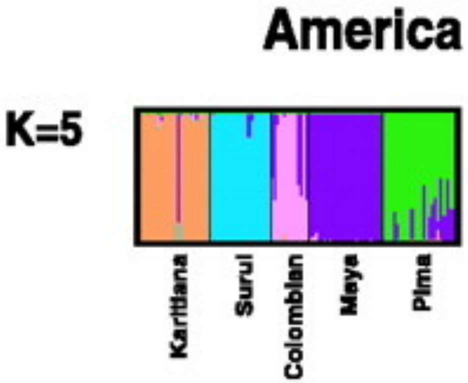
Rosenberg et al., 2002

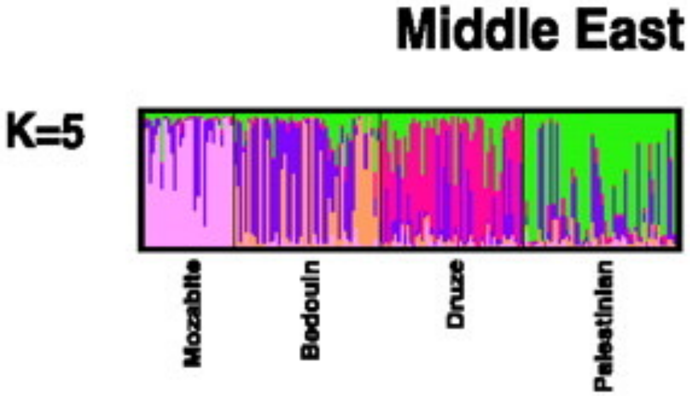
Structure in Eurasia

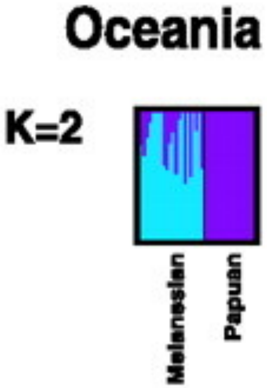


Structure in other geographic regions


Africa


America


Middle East


Oceania

What do these results mean?

- Even though there are very few fixed genetic differences between human populations, it is often possible to distinguish related groups based on genetic variation
- This is because of the additive effect of slight differences in allele frequencies across many genomic loci

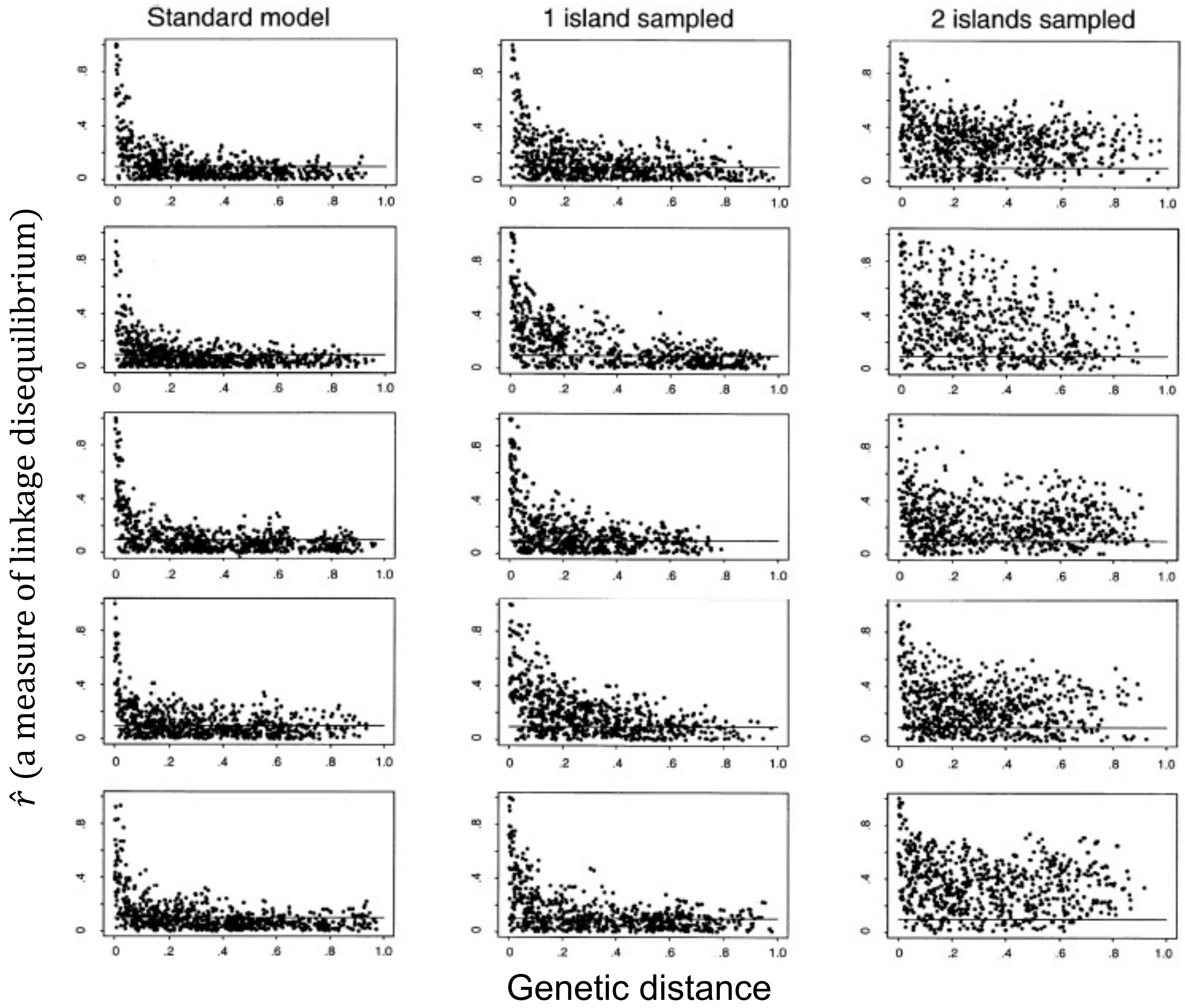
Population structure can confound associations over space

- To deal with this, the structure of a population can be included as part of a linear model
- A ***linear mixed model*** includes a kinship matrix based on the covariance of populations based on SNP variation
- Alternatively, the first several principal components can be included as covariates in a linear model
- We will come back to these types of models in future lectures

LD and population structure

How does population structure affect linkage disequilibrium?

- Simulations of linkage disequilibrium under three simple models of population structure (Pritchard and Przeworski 2001)
- Simulated genetic data under a standard Wright Fisher model, and an island model (simple split)
- They took 5 random subsets of bi-allelic markers in sample sizes of 400 individuals, calculated pairwise linkage disequilibrium based on r and plotted this



The measure of LD used here is the square root of the commonly used statistic:

$$r^2 = \frac{(p_{AB} - p_A p_B)^2}{p_A p_a p_B p_b}$$

Standard model: randomly mating population with $2N_e = 2\text{x}10^4$

1 island sampled: randomly mating population with $2N_e = 1\text{x}10^4$

2 islands sampled: 2 diverging randomly mating populations with $2N_e = 1\text{x}10^4$ in each

Population structure causes LD to increase

Pritchard and Przeworski, 2001

Population structure impacts linkage disequilibrium

- LD decays more rapidly in the larger population
- LD is high when the two islands are analysed together relative to when they are analysed separately. This tells us that
  1. Treating structured populations as one randomly mating population will result in a combined population with high LD, so we need to be careful about how we define populations.
  2. Populations that recently united (in a secondary contact scenario) or for which gene flow links populations, will have higher LD than expected for a single randomly mating population