Demographic modeling

Dealing with data

- Long branch attraction
  - homoplasy overwriting true signal (saturation)
- Gene duplication, gene extinction or paralogous sampling
  - Thinking you are analyzing homologous genes but some specimens have different gene copies
- Incomplete lineage sorting (ILS)
  - Gene trees differ from species tree due to variation shared among species/lineages
- Horizontal gene transfer
  - Hybridization
  - Introgression

- ALL HAVE TO DO WITH GENE TREES NOT MATCHING SPECIES TREES DUE TO NATURAL PHENOMENON

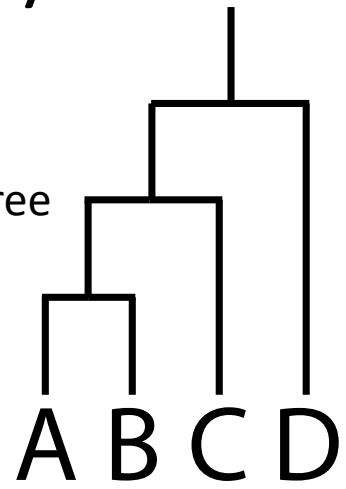
Dealing with data

- Concatenation
  - Genes can have entirely different evolutionary histories (mitochondrial vs nuclear)?
  - Assumes no hybridization
  - Assumes no incomplete lineage sorting
  - Can give high support for wrong relationships (ILS/hybridization)

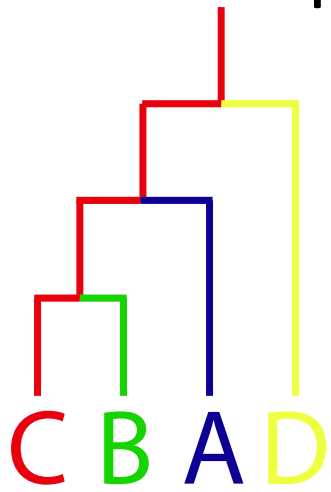
- Coalescent methods
  - Can deal with ILS but assumes no hybridization

Hybridization vs incomplete lineage sorting

Species Tree



Gene Tree



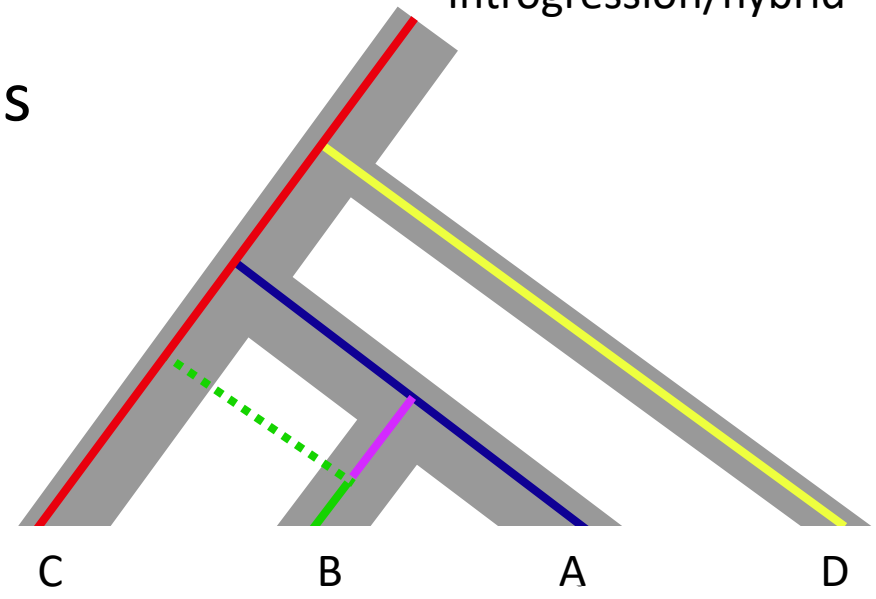
- Only difference is

Branch lengths

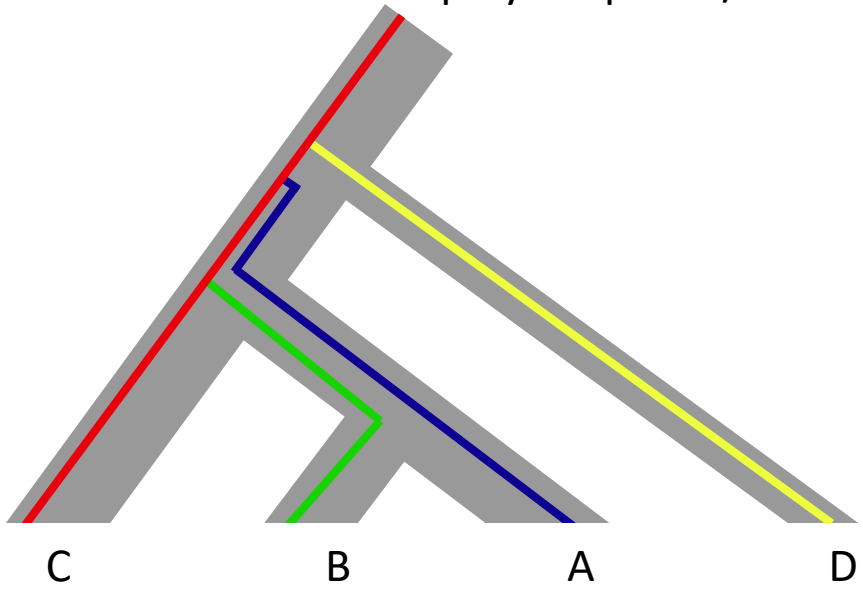
ILS = longer

Hybrid=shorter

Introgression/hybrid

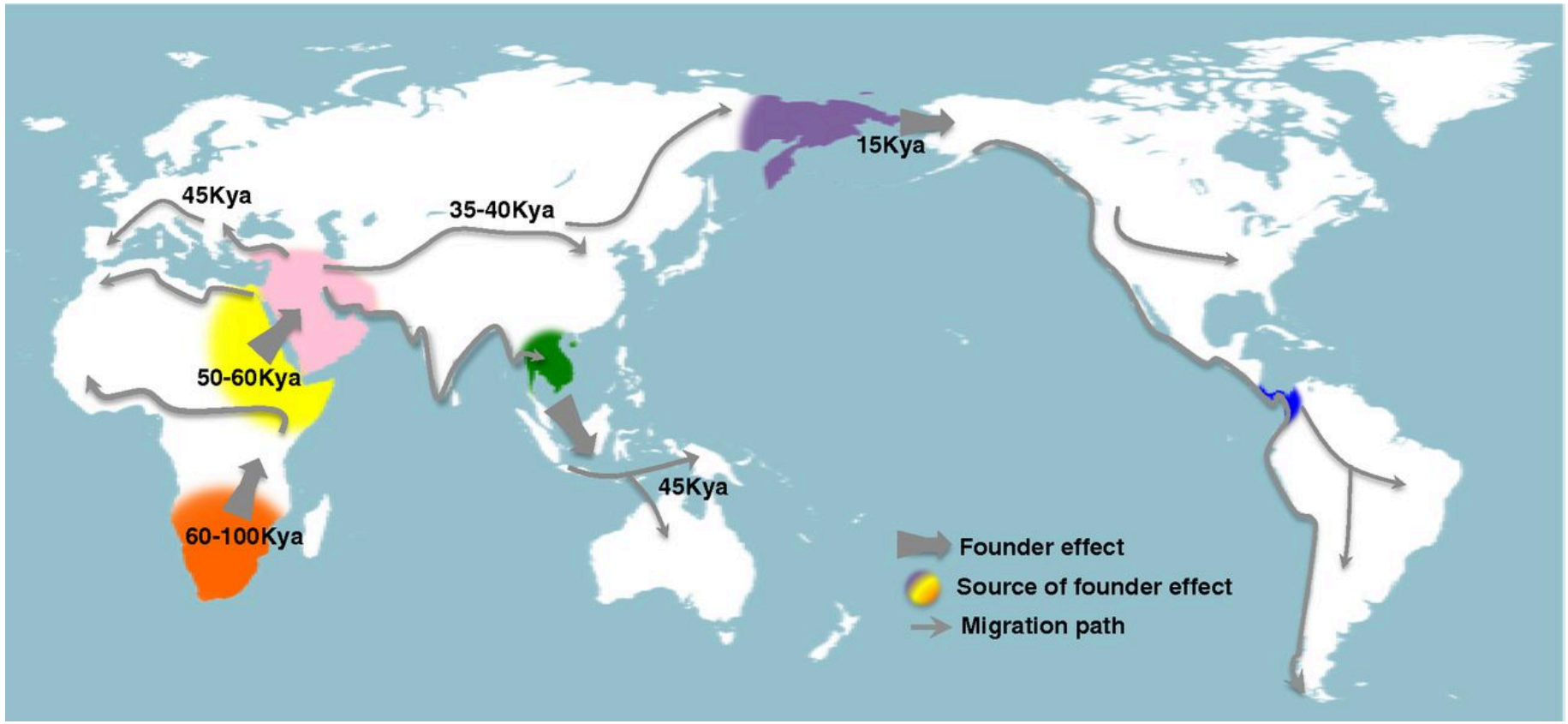


Ancestral polymorphism/ILS



Demographic modeling

- Accounts for gene flow but also used to understand population history



Demographic modeling/inference

- Population genetics/genomics
- Interface of species/populations

- Demographic inference refers to finding a particular model describing your species of interest over time
  - population size changes (Ne)
  - population split (divergence)
  - mixture events (migration)

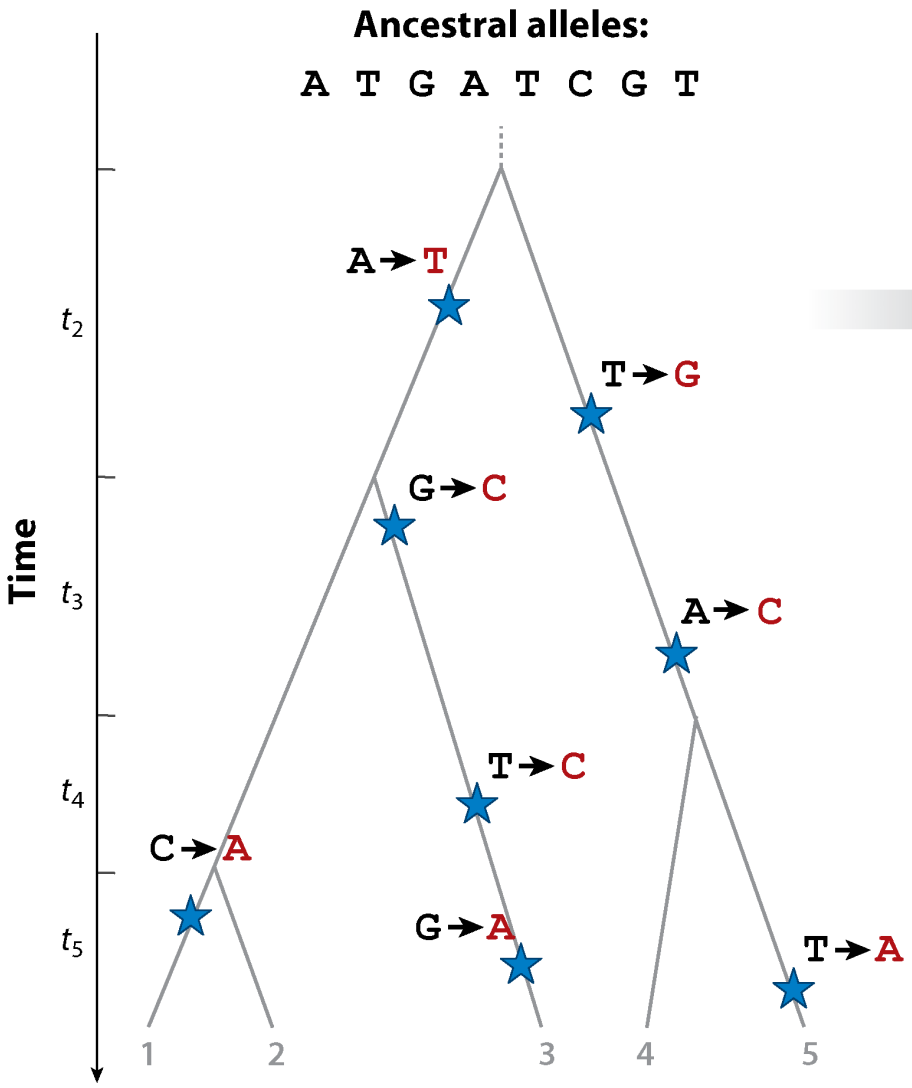
Site frequency spectrum

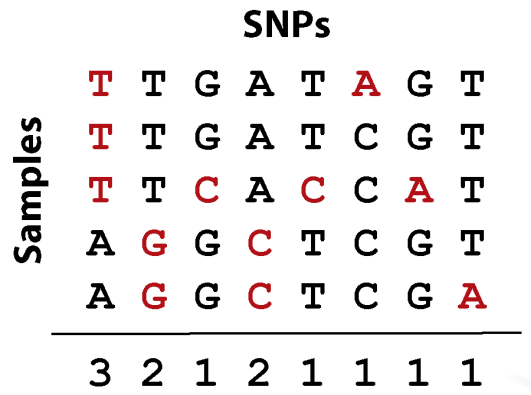
- The number of single-nucleotide polymorphisms (SNPs) at particular frequencies in a sample of individuals

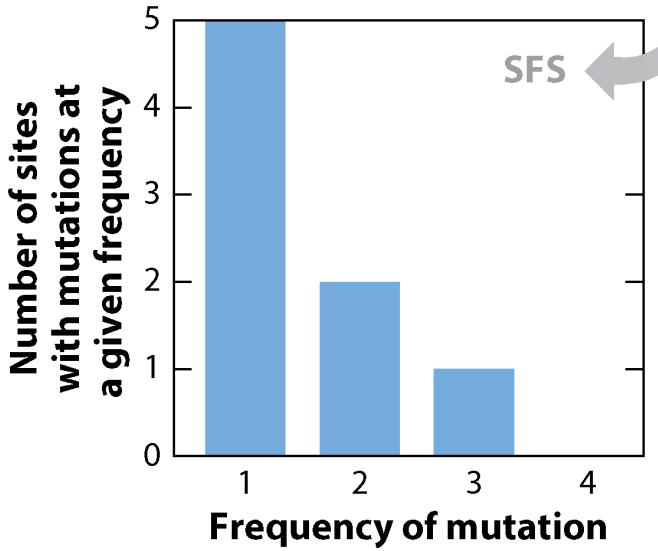
- The SFS can be constructed from a single genomic region, the entire genome, or a particular category of sites (non coding)
  - Rad-seq data
  - Whole genome data
  - Genome resequencing data

- The SFS treats all SNPs in the data set as independent of one another

Single species Site frequency spectrum

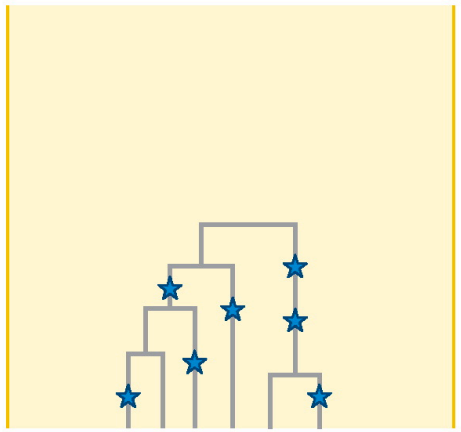


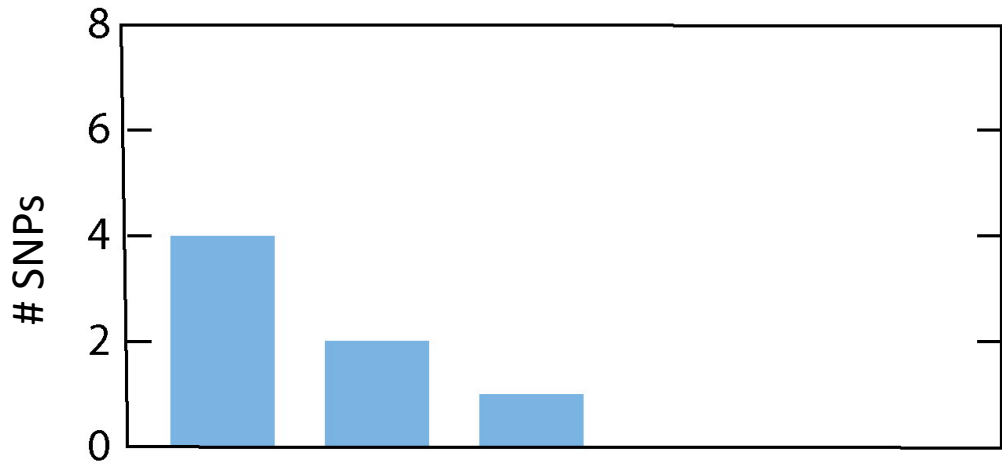




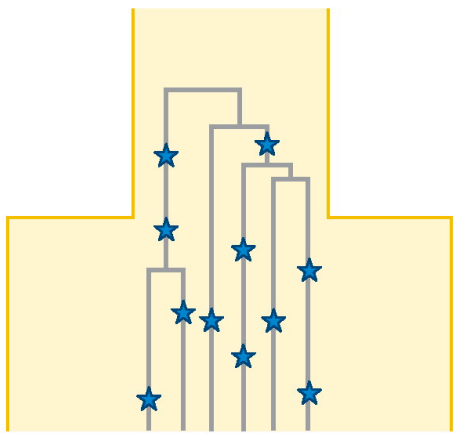
# of copies you have in a population

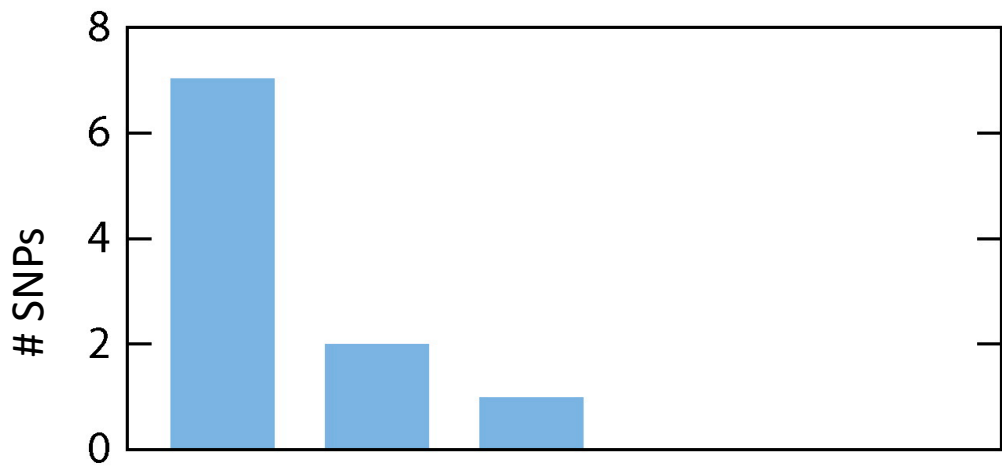
Standard neutral model





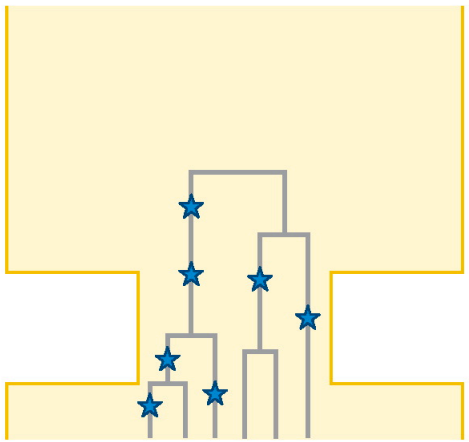
Population growth model

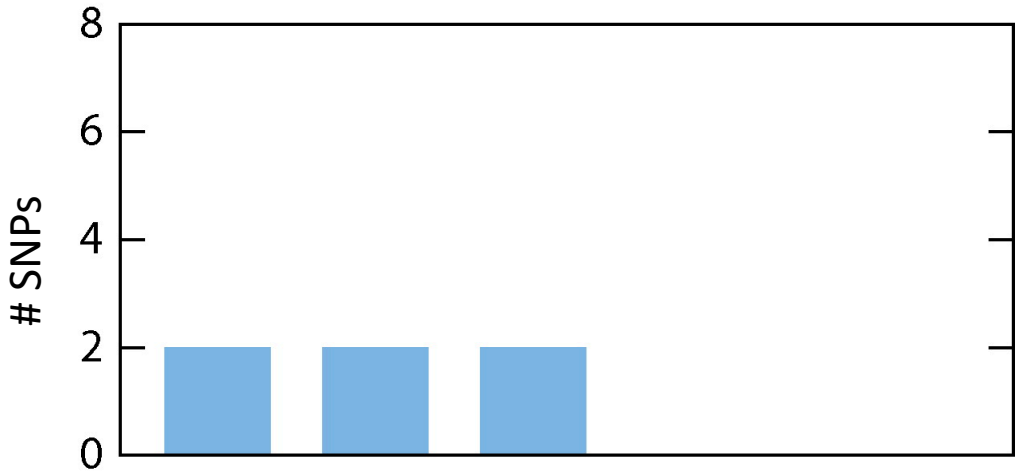




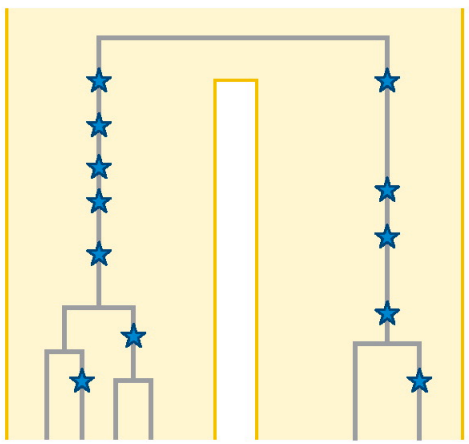
# of copies you have in a population

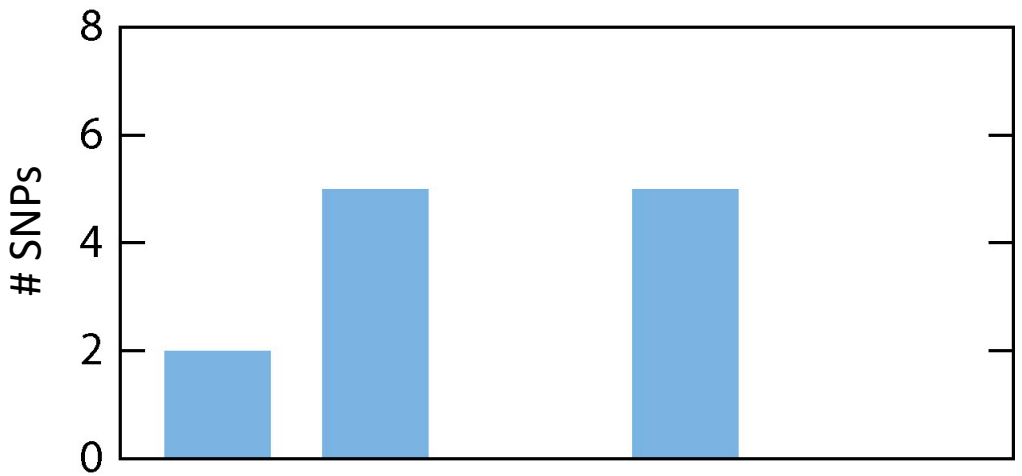
Bottleneck model



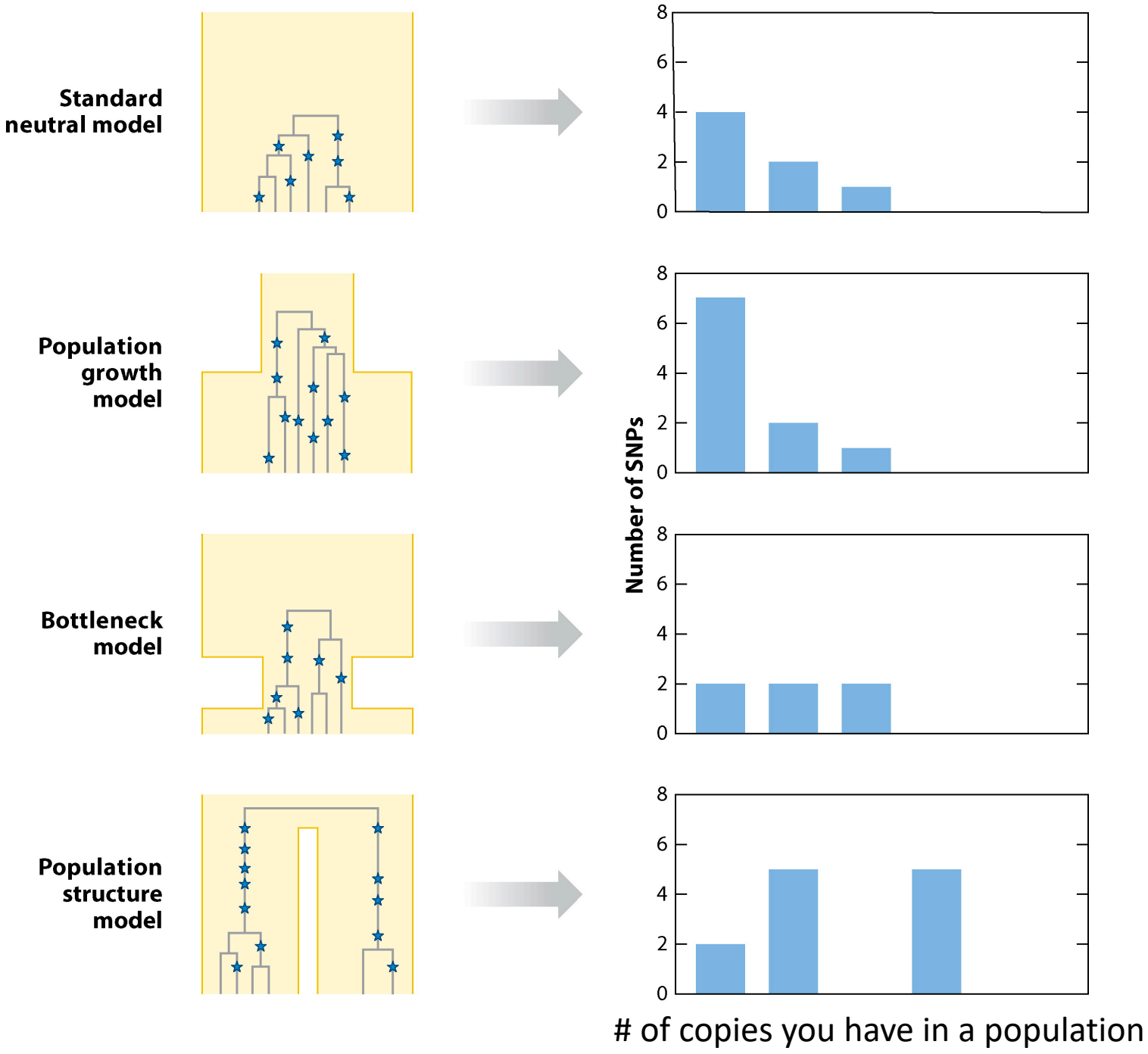


Population structure model





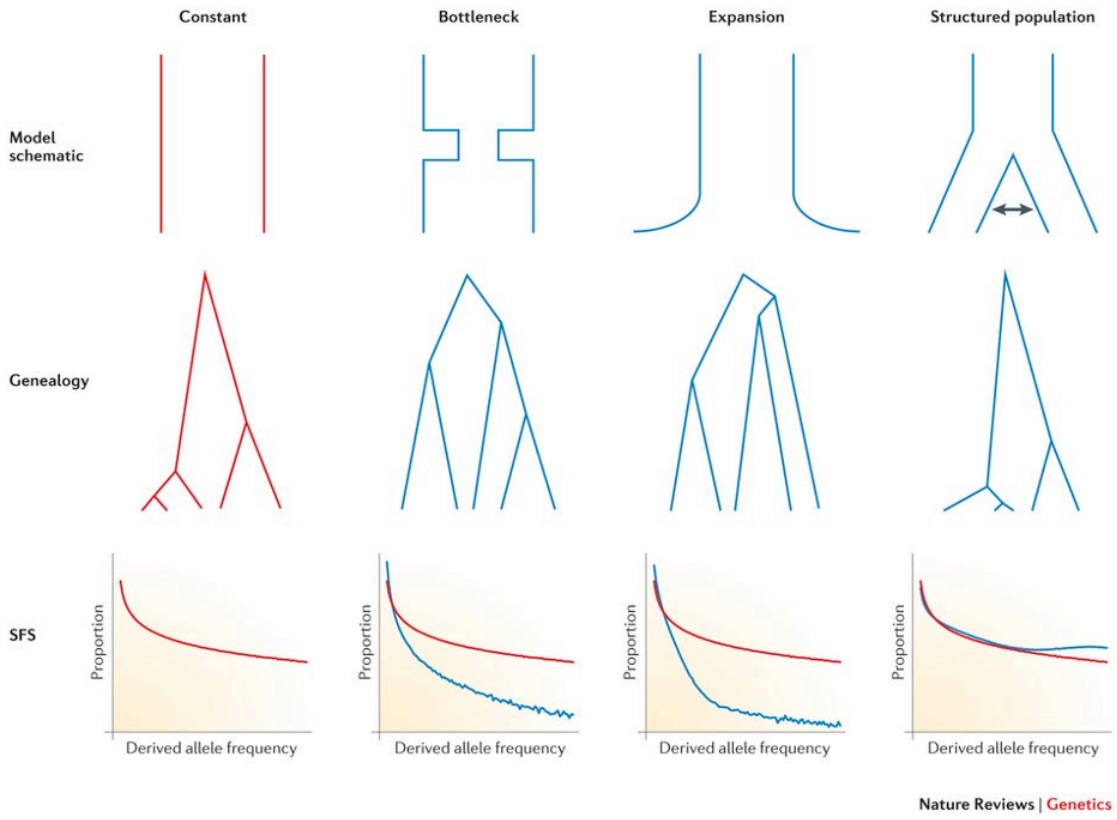
# of copies you have in a population


Standard neutral model
Population growth model
Bottleneck model
Population structure model
Number of SNPs
0
2
4
6
8
# of copies you have in a population

How to use SFS to infer demographic param

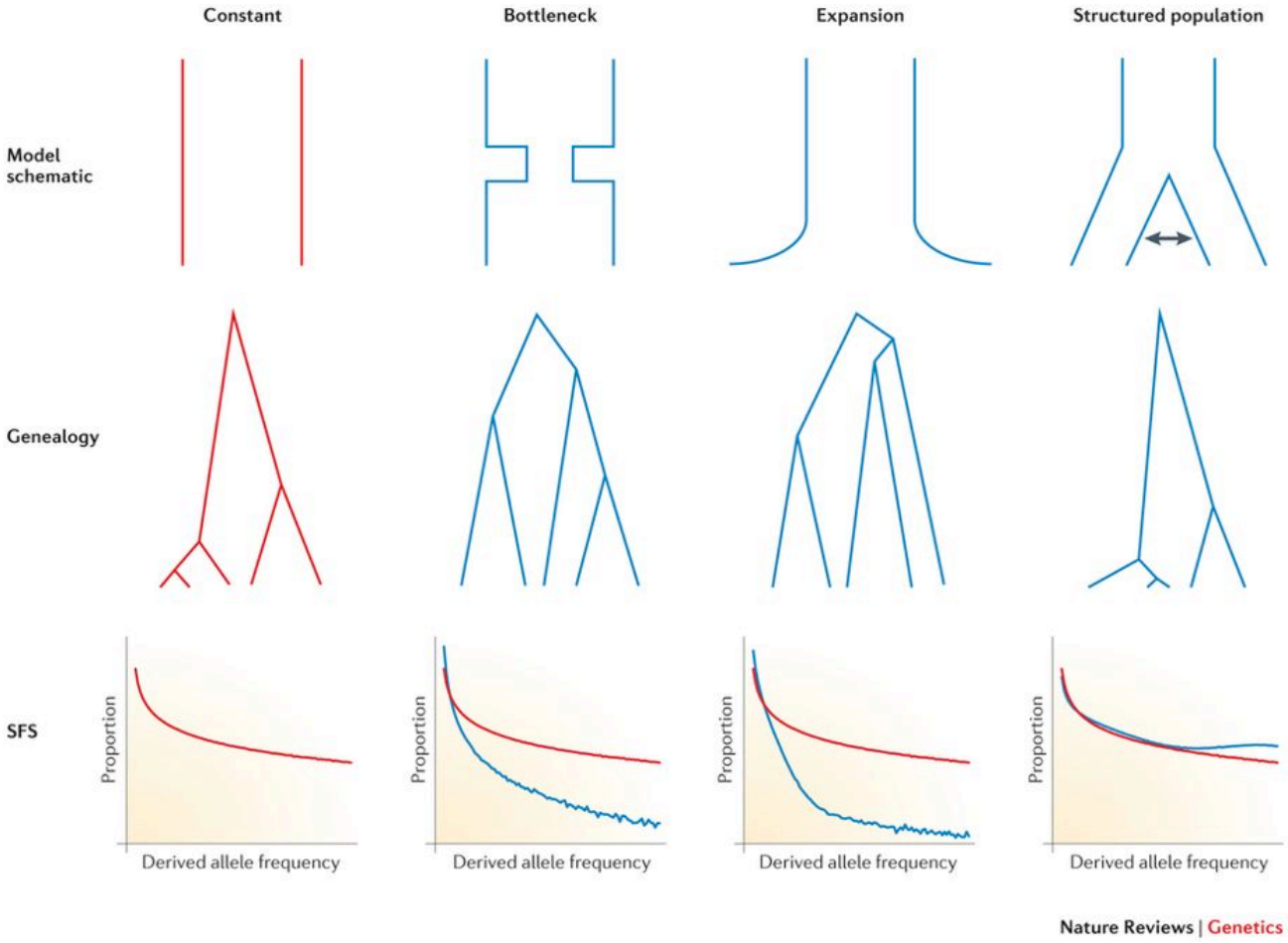
- The observed shape of the allele frequency spectrum
  - Is sensitive to demographic history
    - Population size changes
    - Migration
    - Substructure

- Construct various models to test
  - Constant
  - Bottleneck
  - Expansion
  - Structure
    - migration



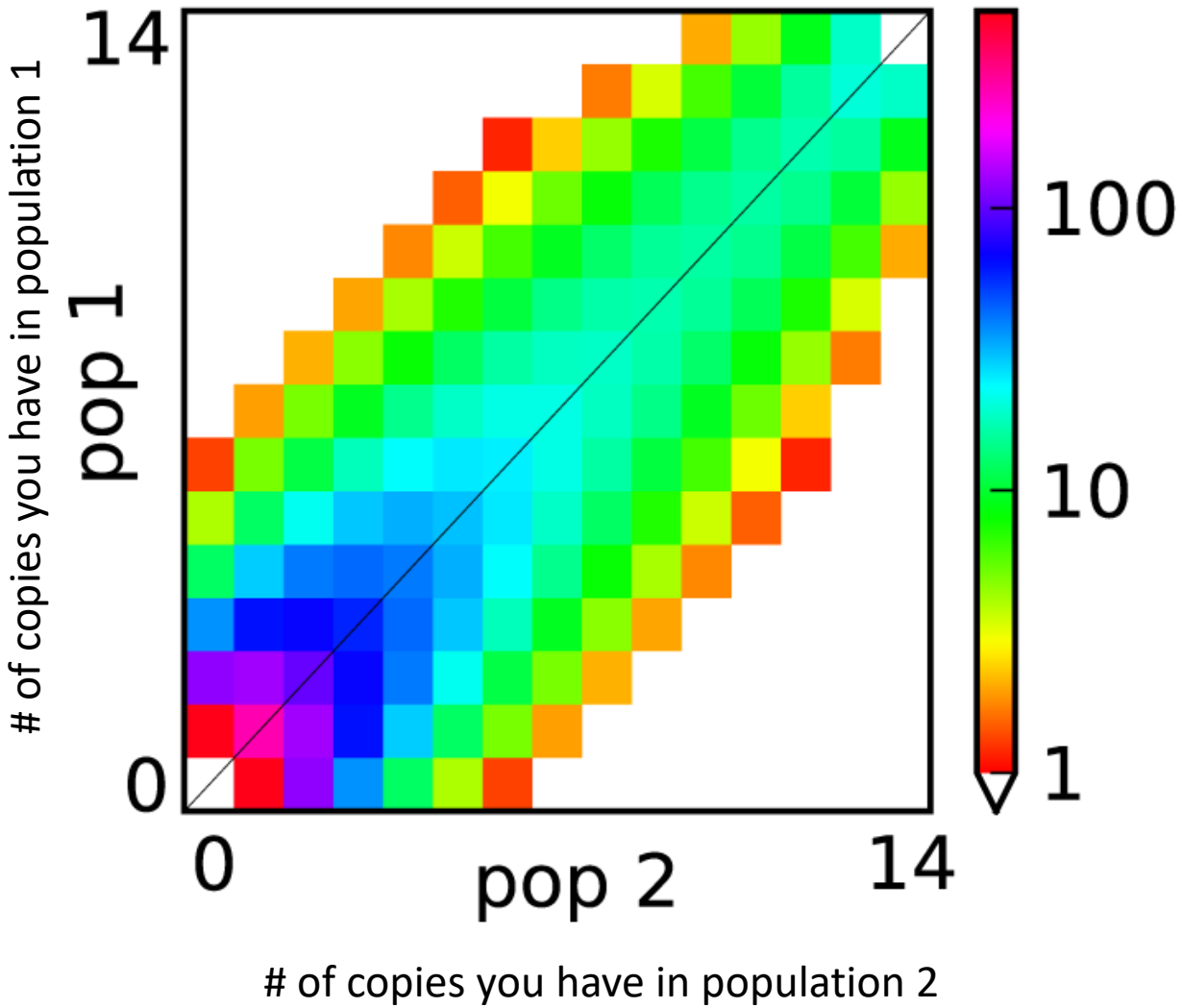
How to use SFS to infer demographic param

- Use likelihood function to compare observed SFS to simulated
  - L(Model|data)= likelihood score
- So you numerically solve
- MCMC is too complicated



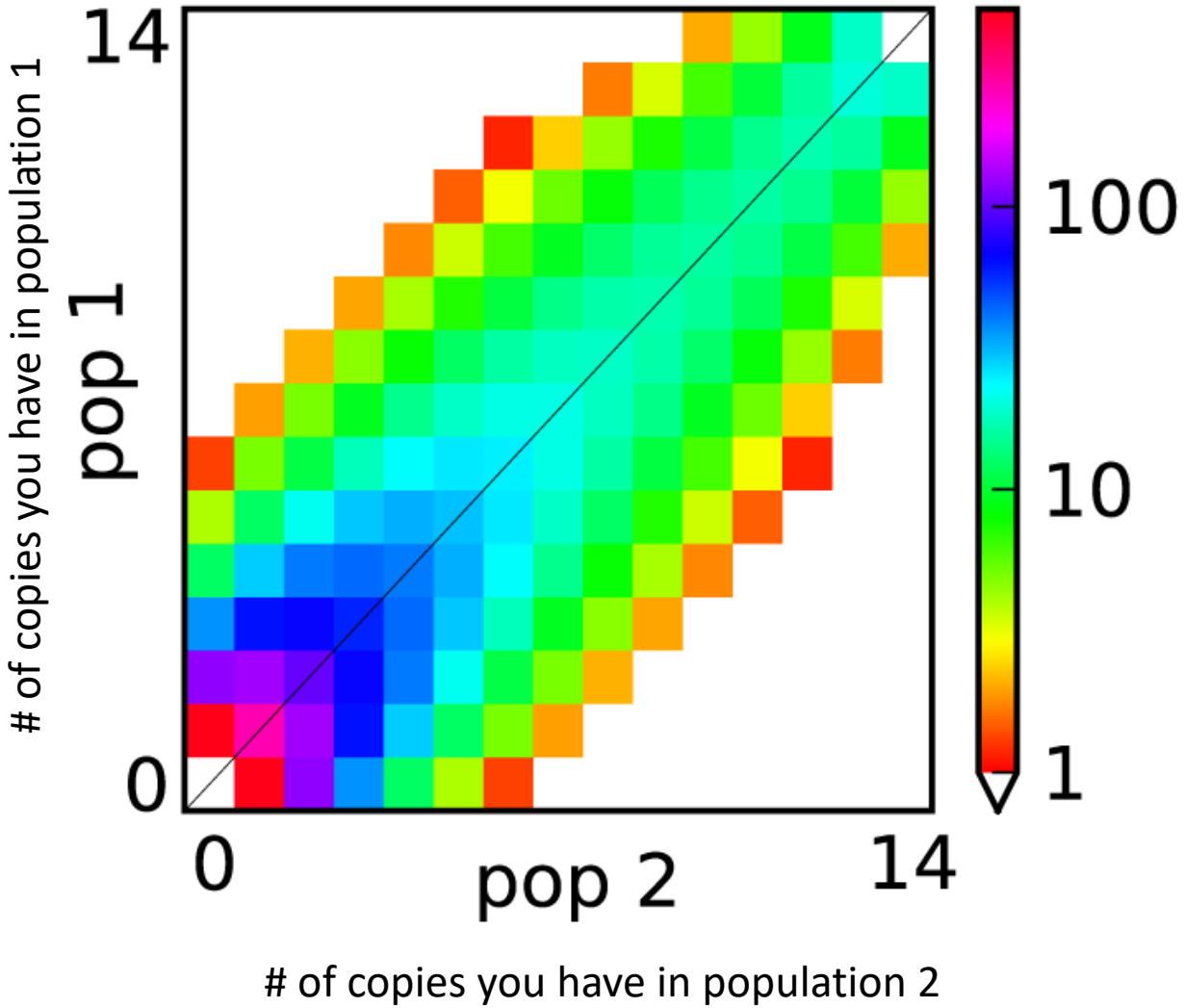
Joint SFS (2D-SFS)

- Basically a heatmap of allele frequency between two populations
- Showing how many sites in our data in which the allele frequencies in the 2 populations take on combination of values



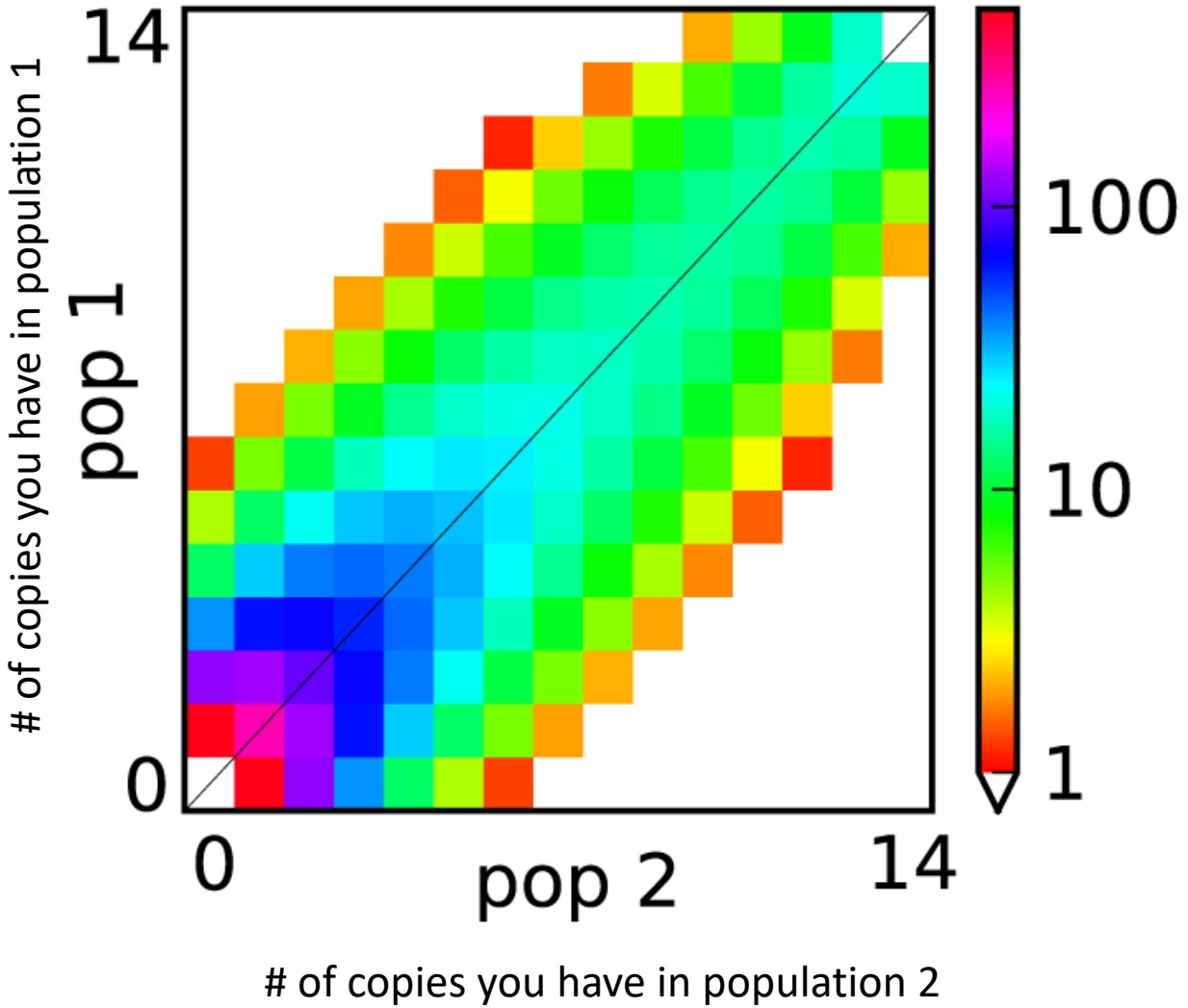
Joint SFS (2D-SFS)

- 7 seven individuals/ populations
- Diploid so 14 gene copies in each populations
- Typical SFS where lots of rare SNPs from either population with fewer common SNPs

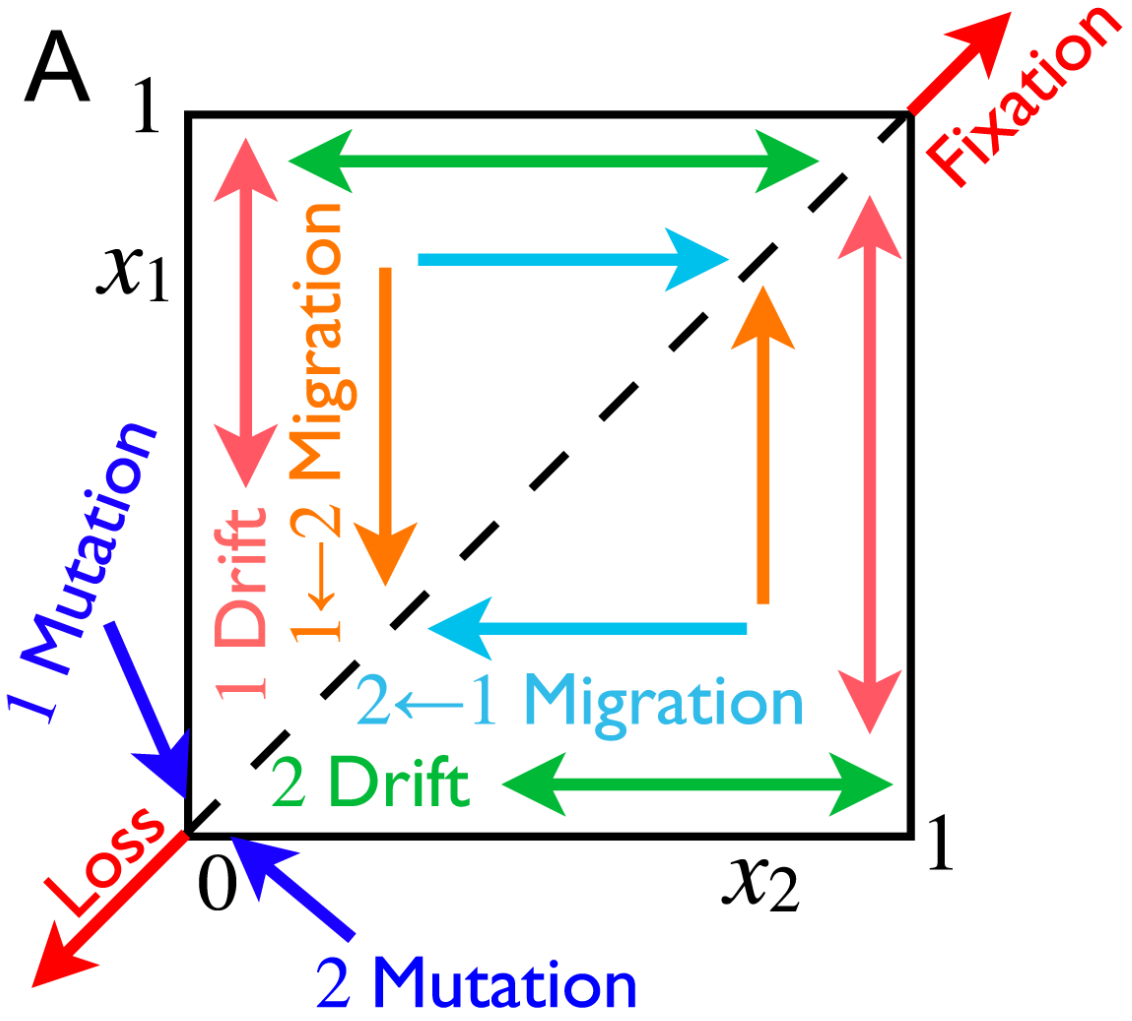


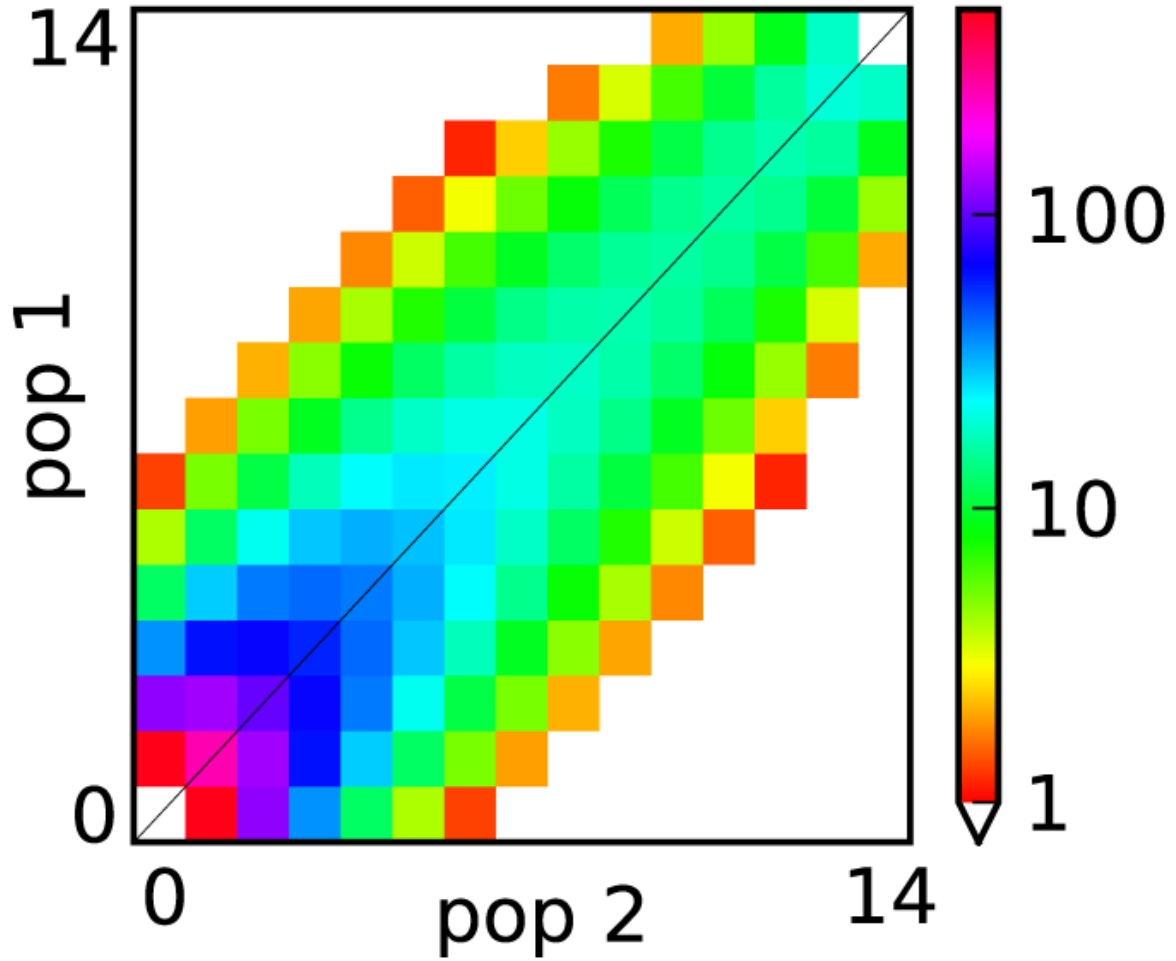
Joint SFS (2D-SFS)

- Mutation creates lots of rare alleles

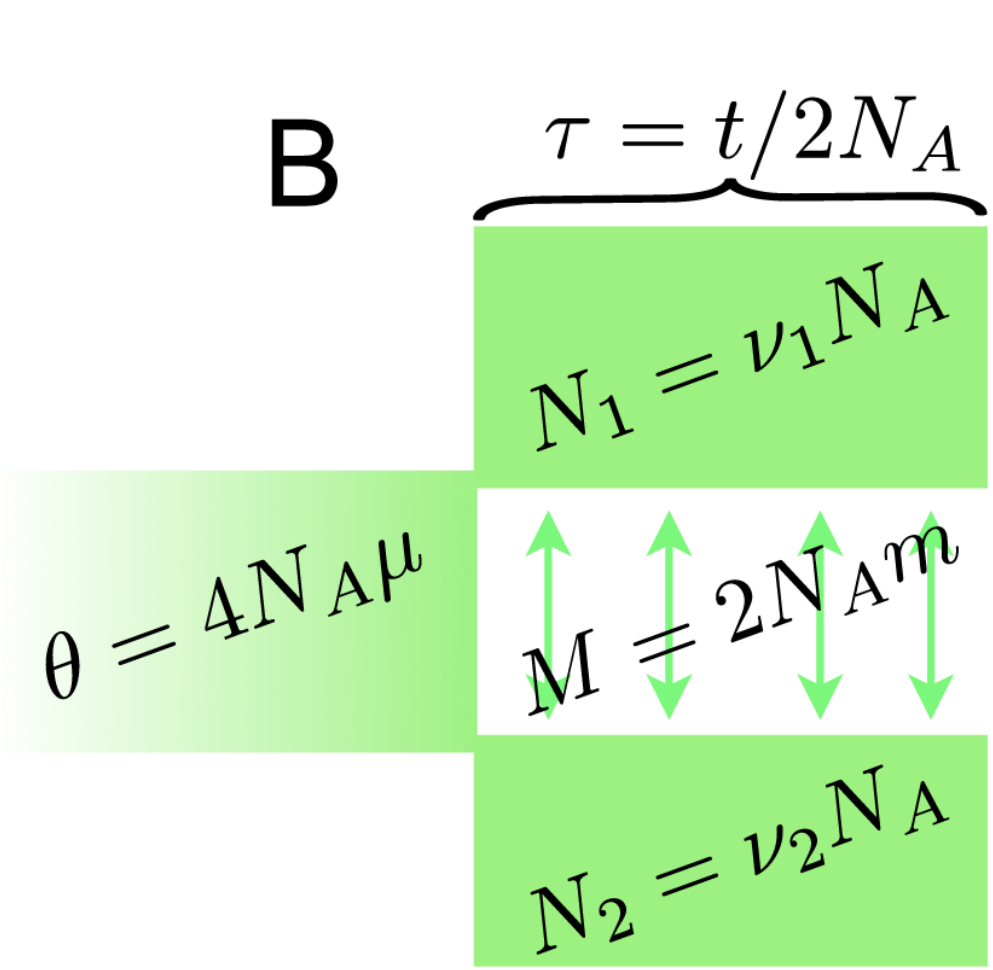


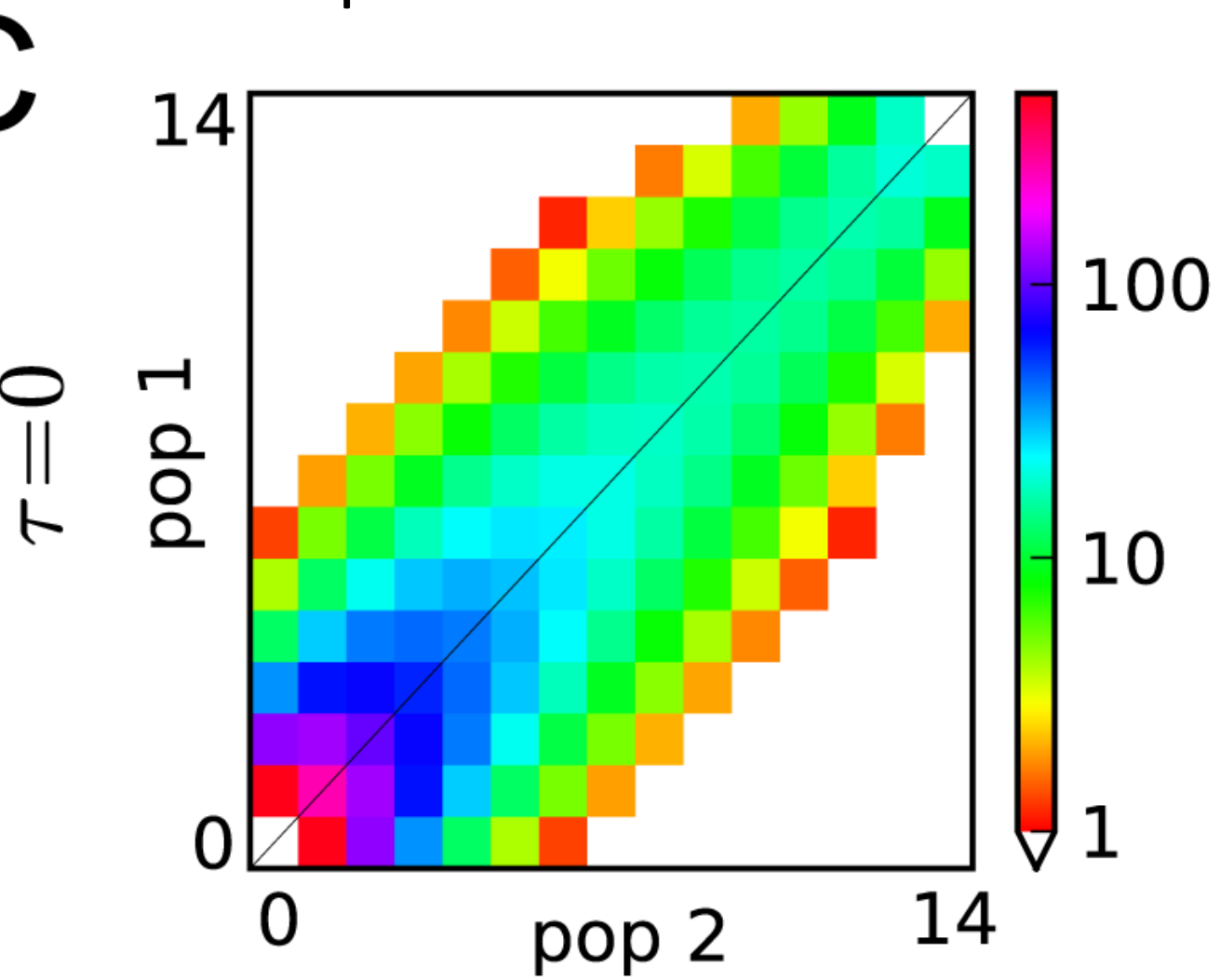
Joint SFS (2D-SFS)



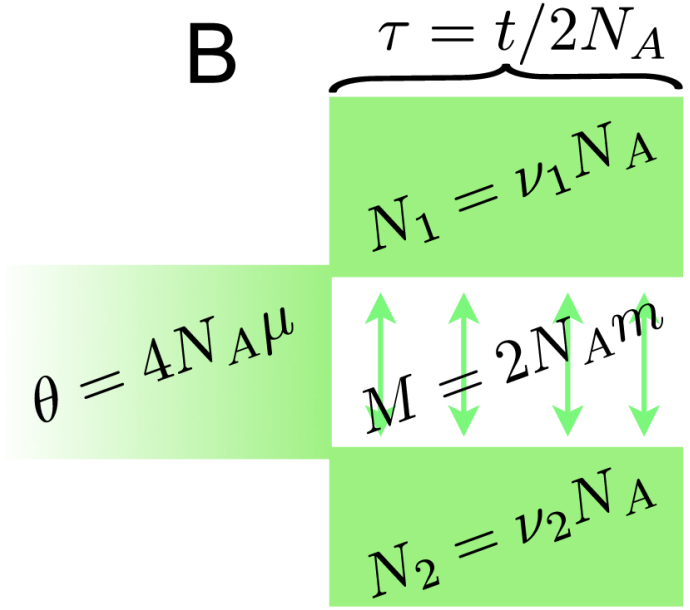


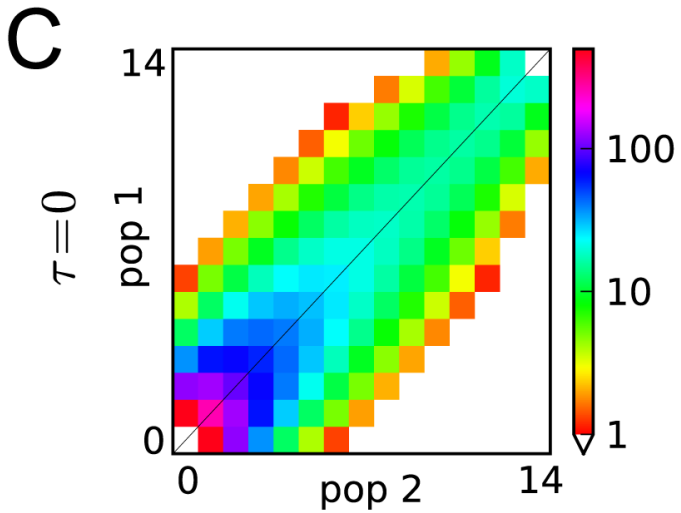
Construct models then compare to obs data

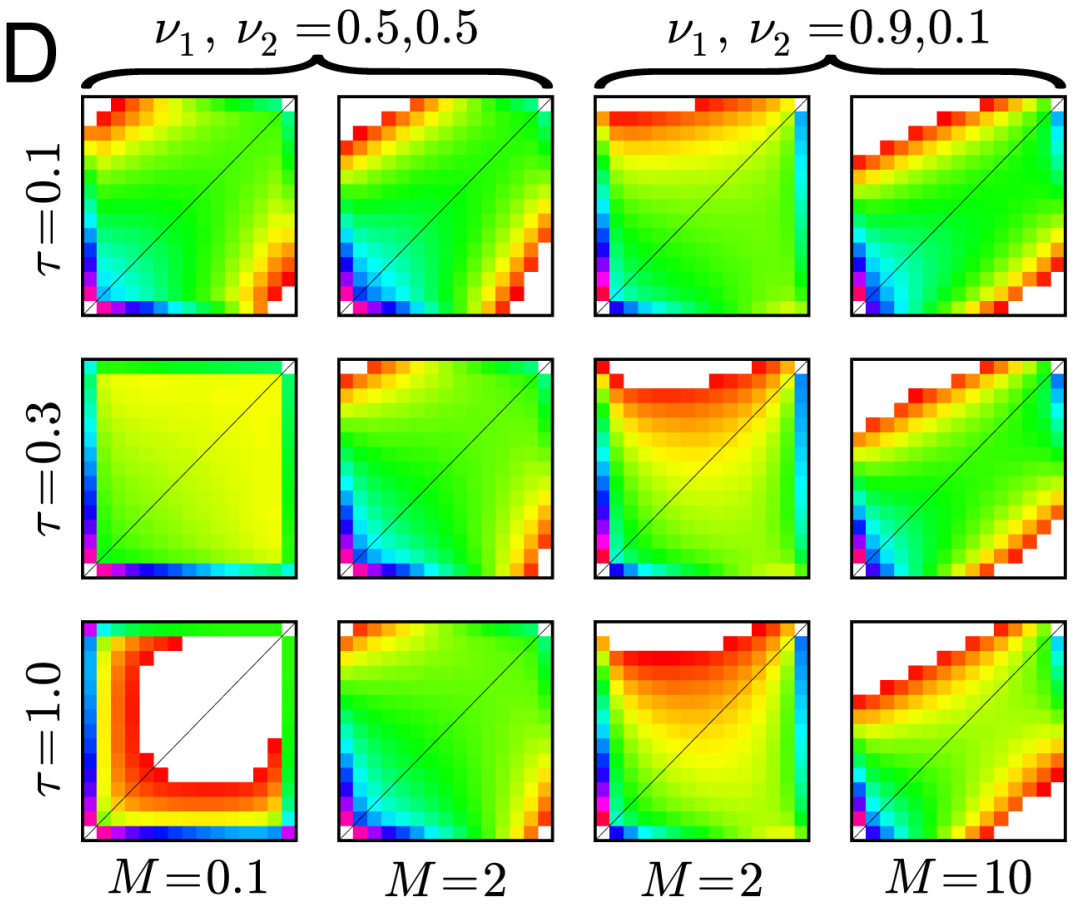




Construct models then compare to obs data





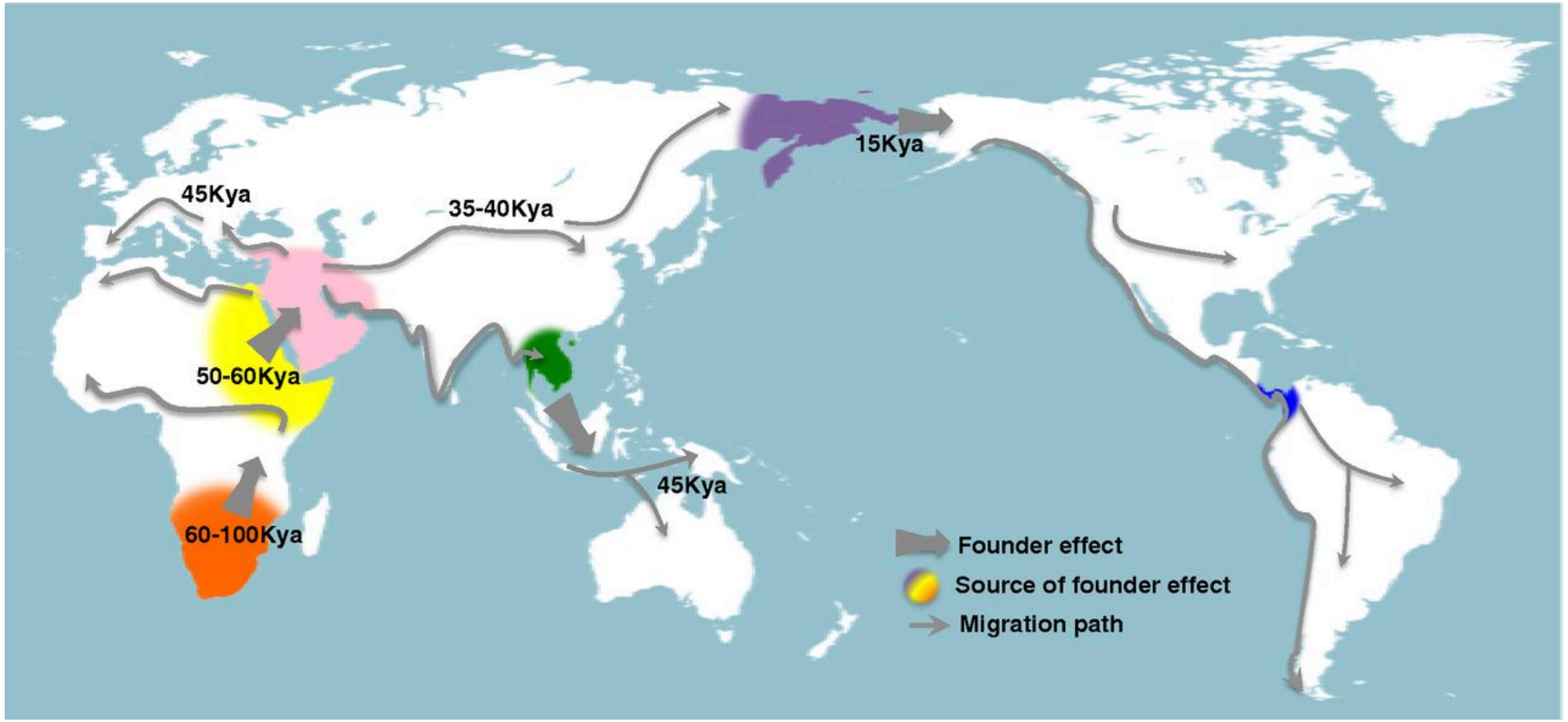


Other methods

- Treemix
  - Takes snp data and makes species trees
    - Adds migration edges and sees if likelihood scores improve

- Random forest methods
  - Takes SFS and model to calculate likelihood scores
  - Uses machine learning and random forest decision trees to determine which model best fits the data

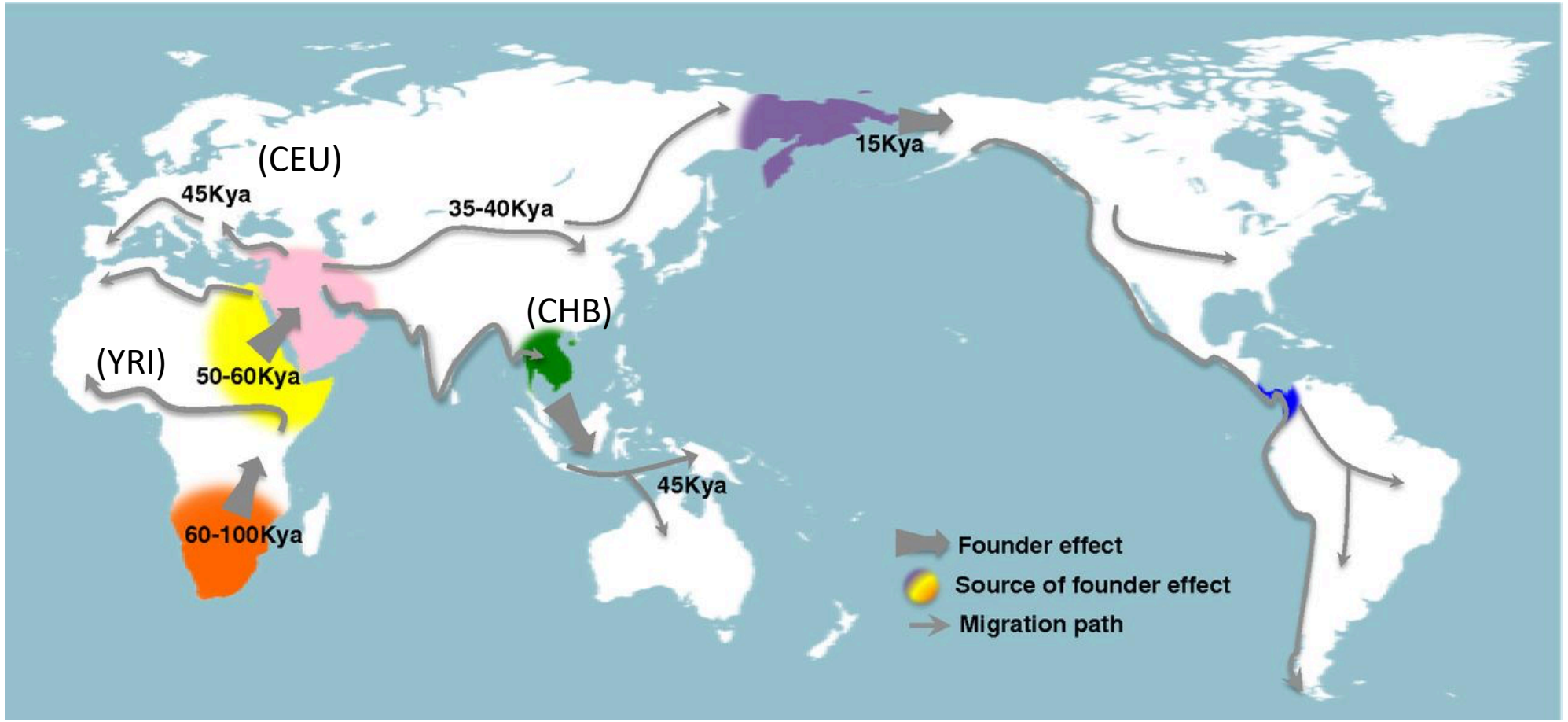
Example Out of Africa



Example Out of Africa

- Three populations
- 12 Yoruba individuals from Ibadan, Nigeria (YRI)
- 22 CEPH Utah residents with ancestry from northern and western Europe (CEU)
- 12 Han Chinese individuals sampled in Beijing, China (CHB).

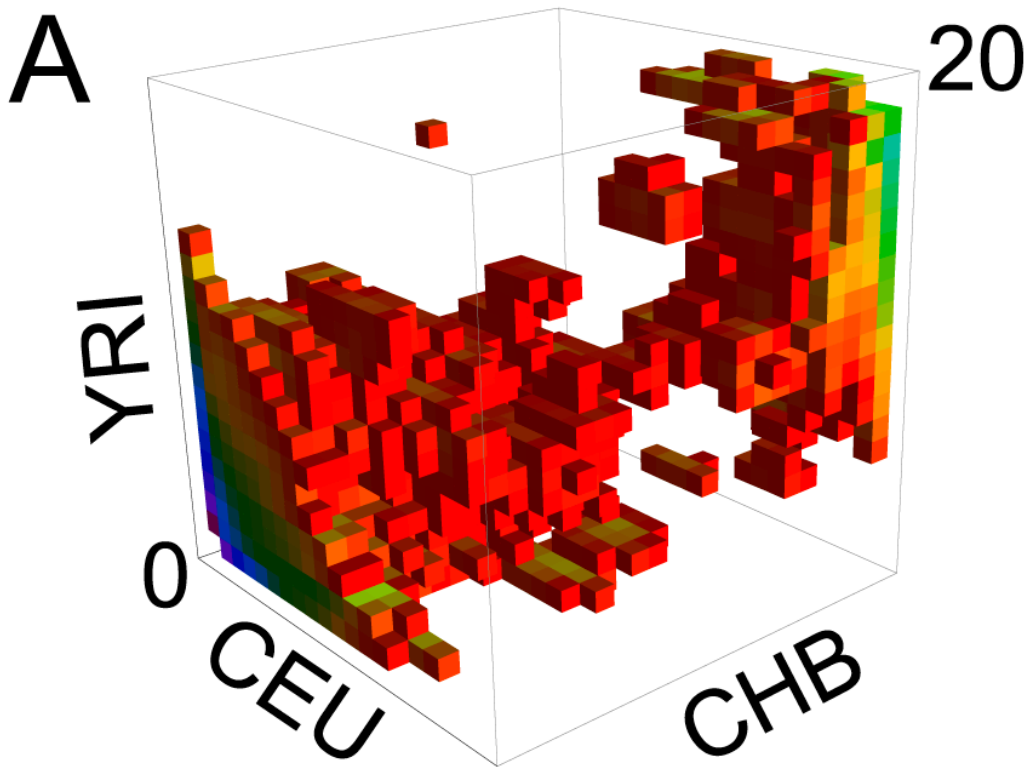
Example Out of Africa

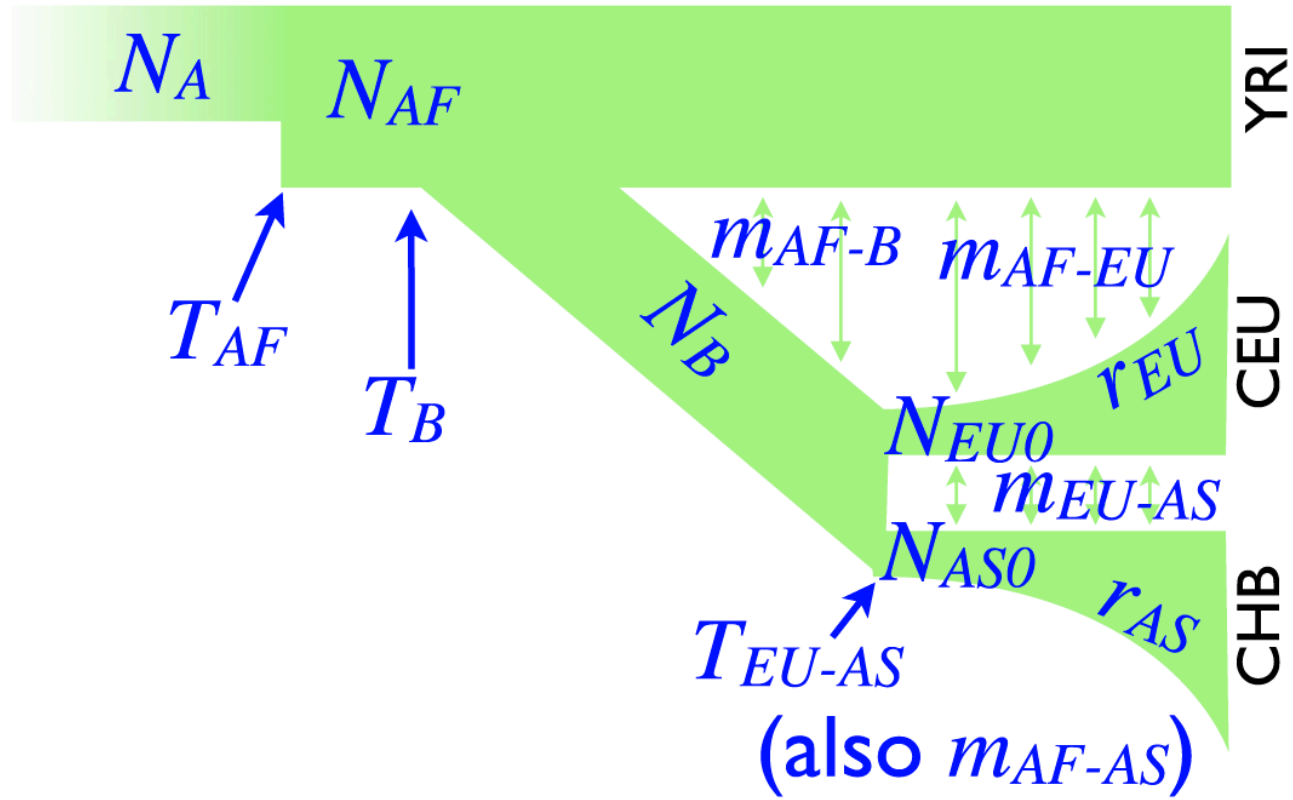


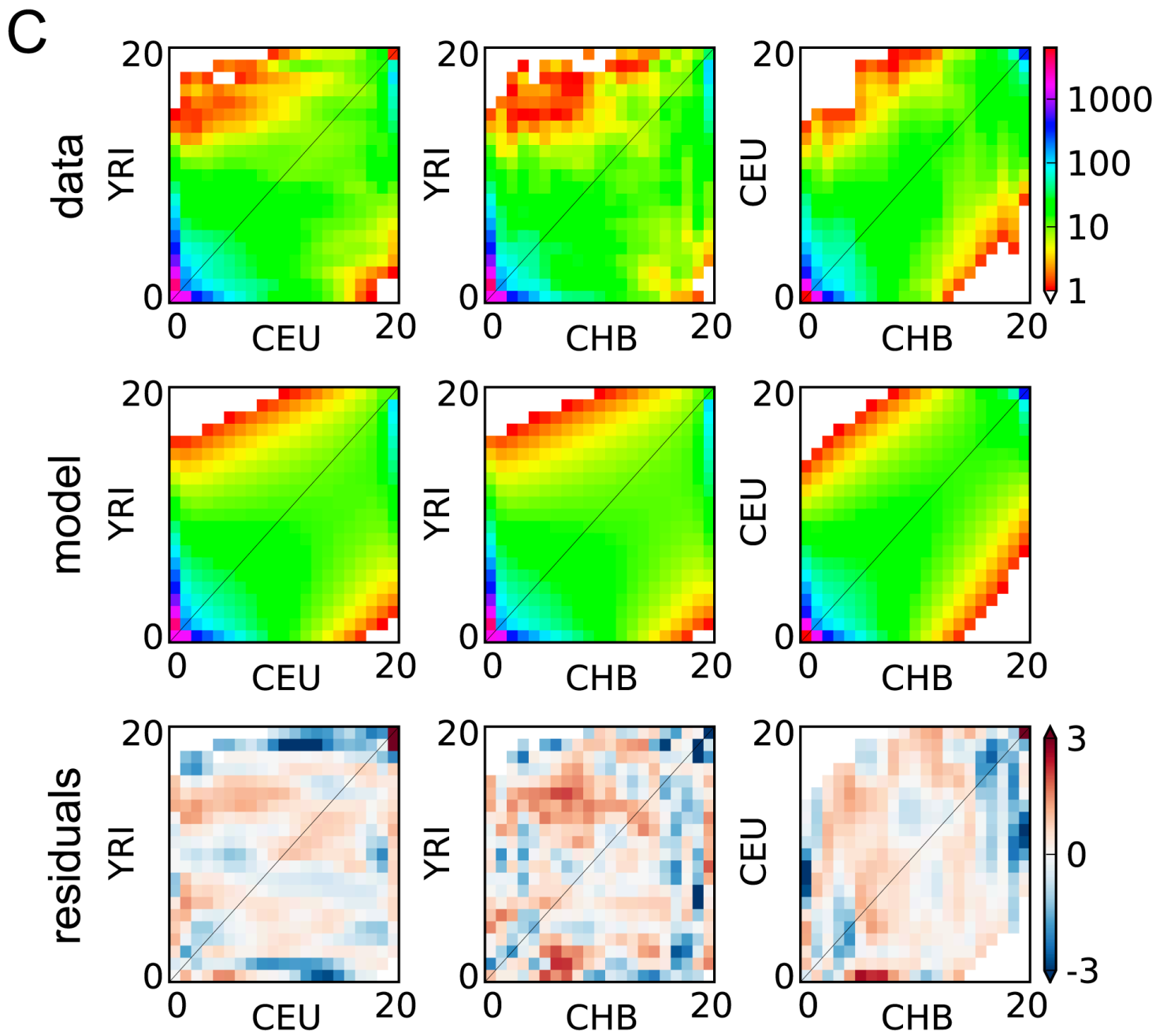
Example Out of Africa

- Data is from National Institute of Environmental Health Science's Environmental Genome Project SNPs database

- SNPs in 5.01 Mb of sequence from noncoding regions of 219 autosomal genes


A
20
YRI
0
CEU
CHB


B
N_A
N_{AF}
YRI
T_{AF}
T_B
N_B
$m_{AF\text{-}B}$
$m_{AF\text{-}EU}$
r_{EU}
CEU
N_{EU0}
$m_{EU\text{-}AS}$
N_{AS0}
r_{AS}
CHB
$T_{EU\text{-}AS}$
(also $m_{AF\text{-}AS}$)


C
data
model
residuals
YRI
CEU
CHB
0
20
1000
100
10
1
3
0
-3