Demographic inference based on Site frequency spectrum (SFS) – Part I

Vitor C. Sousa

CE3C – center for ecology, evolution and environmental changes
Department of Animal Biology
Faculdade de Ciências da Universidade de Lisboa

2025 WSPG Cesky Krumlov
22 Jan 2025

vmsousa@fc.ul.pt

ce3c centre for ecology, evolution and environmental changes

CHANGE Global Change and Sustainability Institute

EGB Evolutionary Genomics and Bioinformatics group

FC Ciências ULisboa

Outline

Part I

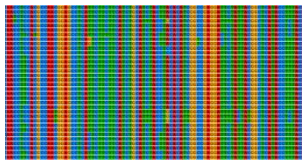
- Modeling demographic history: Population trees vs gene trees
- The SFS and coalescent trees
- Fastsimcoal2 principles – composite likelihood
- Approximate Bayesian Computation

Part II

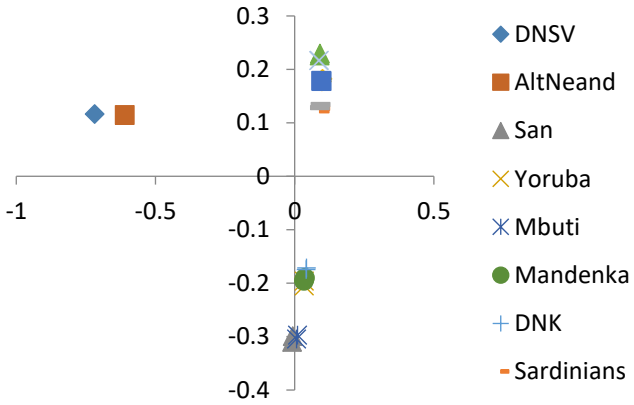
- Example of applications to different problems and types of data

What can we learn from population genomic data?

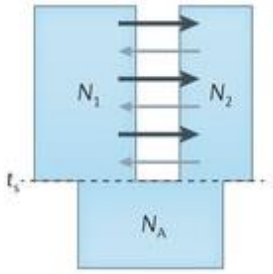
Genomic data



«Model-free» methods e.g. PCA



Model-based methods



Summary statistics:

- Characterize **genetic diversity** within and among populations

- Characterize **genetic differentiation** among populations

Evolutionary Processes:

- Demography
- Selection
- Mutation
- Recombination

Patterns

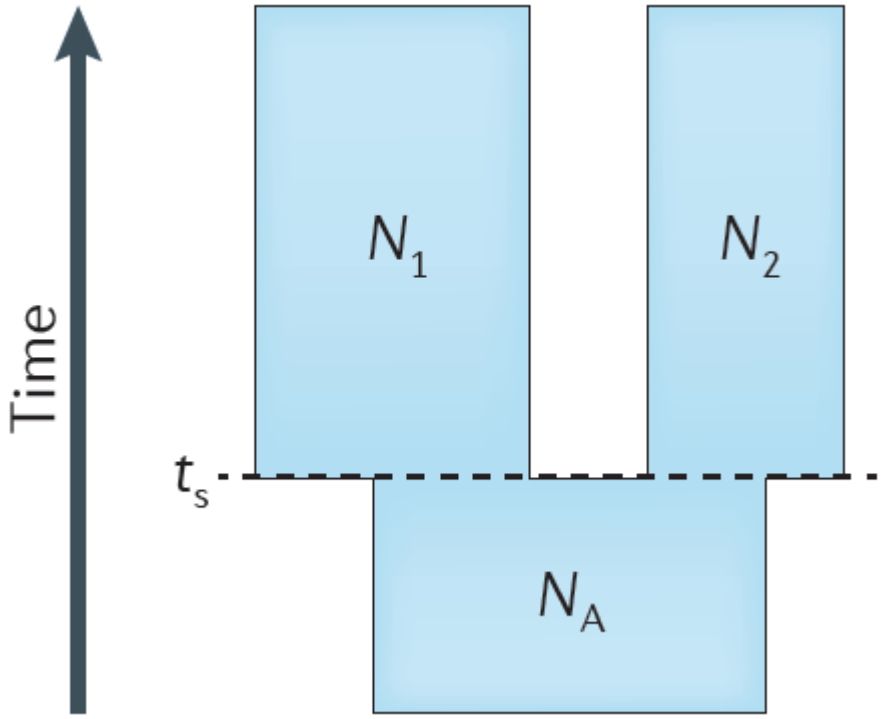
Processes

Demographic history of populations

Past demographic events:

- Population split
- Migration events
- Changes in effective population sizes (expansions or bottlenecks)
- Temporal changes in migration rates and effective sizes

Population tree



Why do we care about demographic history?

It is often interesting in itself

- What is the amount of gene flow? Time of split?

Demography affects the efficiency of natural selection

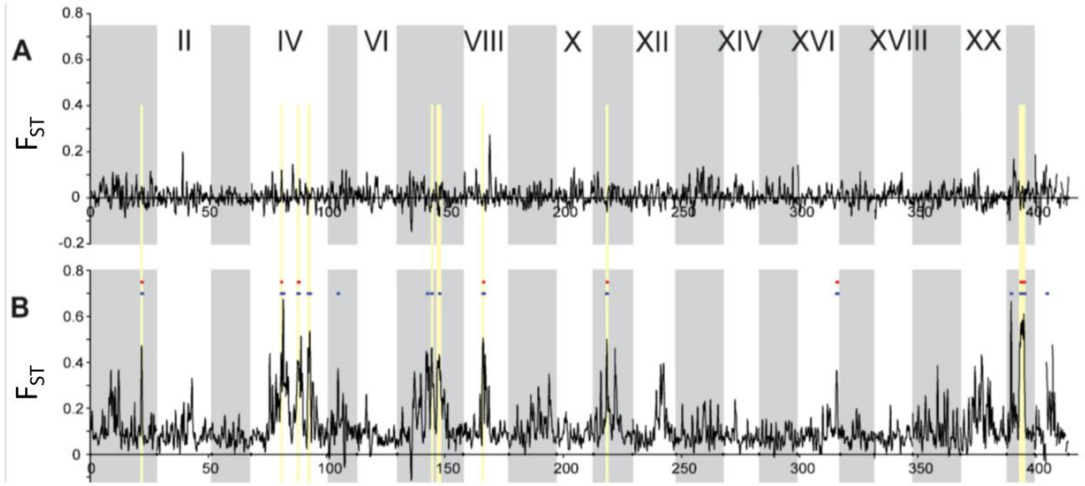
- Response to selection is different in small *vs* large populations, with *vs* without gene flow, etc.

Demographic history affects genome-wide patterns

- "null" model: regions under selection detected as outliers.

Between a pair of marine populations

Between marine and freshwater populations



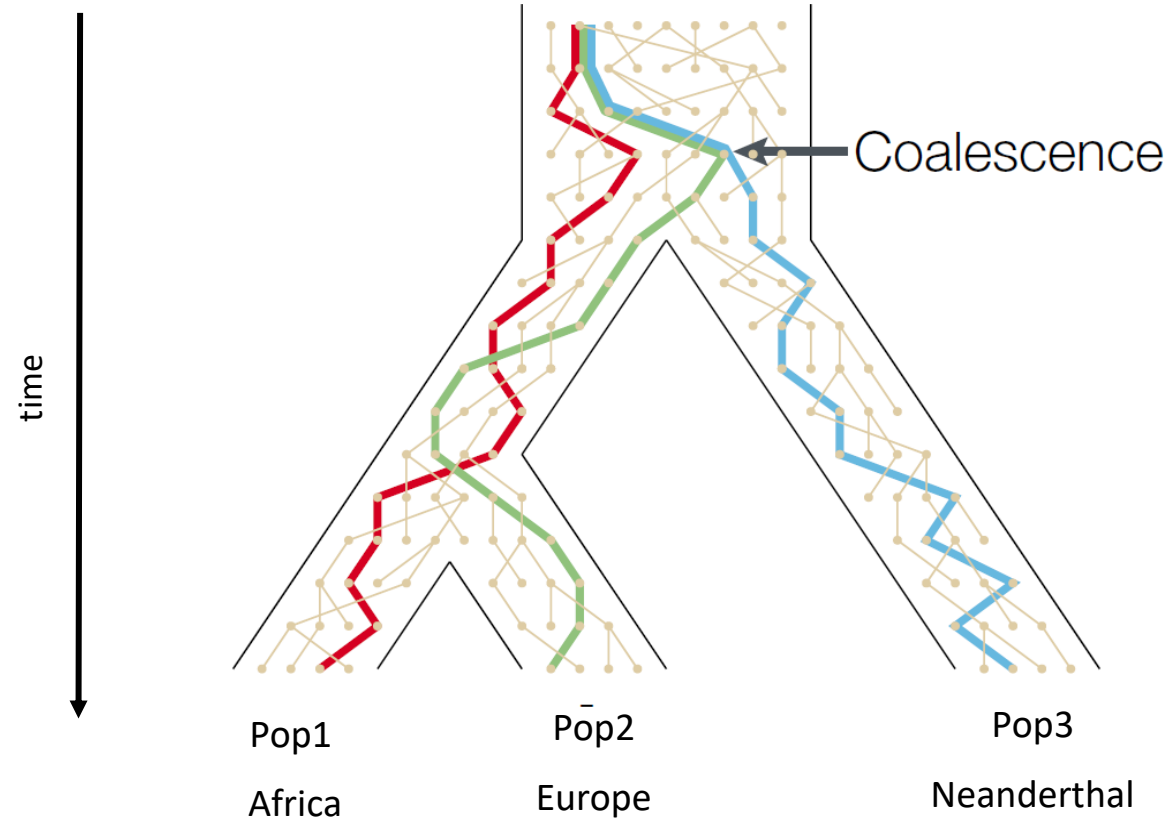
Chromosome position

Alternating white and grey indicate linkage groups (chromosomes)

Hohenlohe et al (2010) Plos Genetics

Coalescent trees "link" population history to observed genetic patterns

Coalescent theory describes the expected gene trees, accounting for mutation, recombination and demographic history



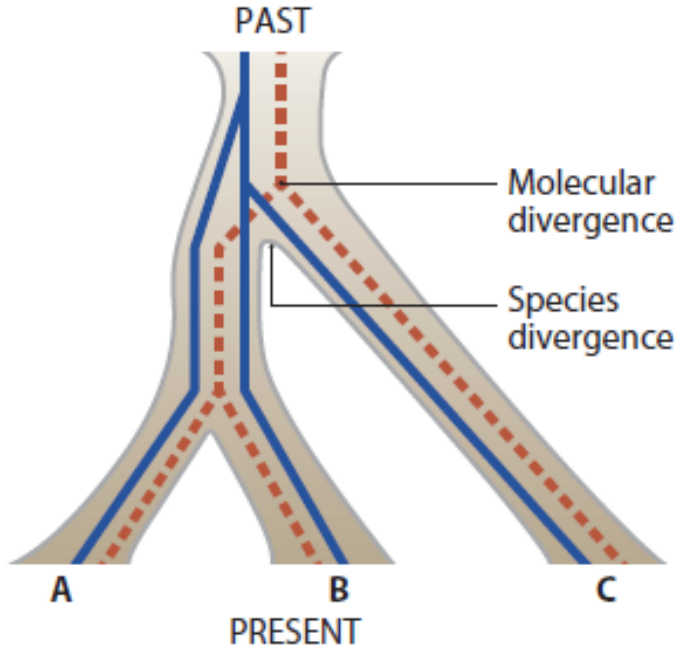
Rosenberg and Nordborg (2002) *Nat Rev Genetics*

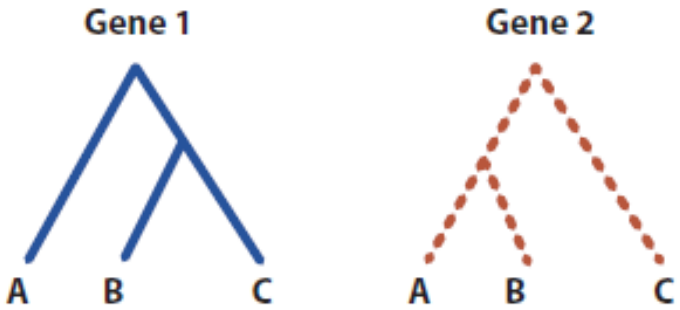
Gene trees vs. Population trees

Gene trees reflect the ancestral relationship of sampled gene copies/chromosomes (before adding mutations).

The relationship between populations is given by the **population tree**.

In phylogenetics it is usually assumed that the gene tree reflects the population/species tree, but that is not the case in population genetics.





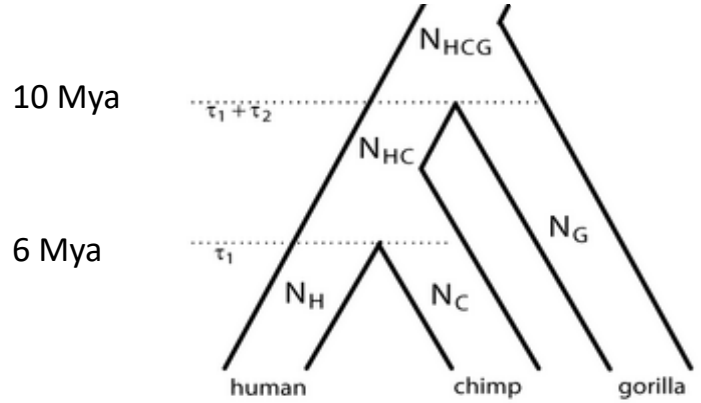
Nichols (2001) TREE

Jobling et al (2014) Human Evolutionary Genetics,2nd edition

Activity

Population tree vs species trees in Human, Chimp and Gorilla

Species tree



What is the proportion of polymorphic sites in the genomes that fall into each gene tree?

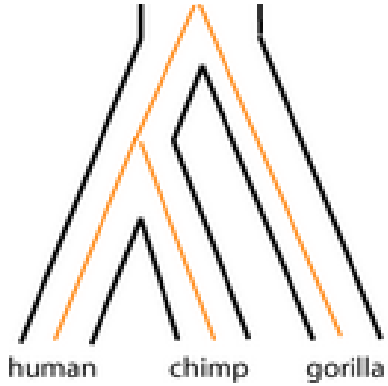
A) 99.0% (H,C); 0.5% (H,G); 0.5% (C, G)

B) 90.0% (H,C); 5.0% (H,G); 5.0% (C, G)

C) 70.0% (H,C); 15.0% (H,G); 15.0% (C, G)

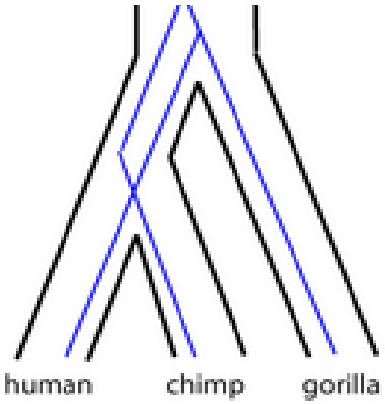
1 single species tree
3 possible gene tree topologies

Gene tree 1
(H,C)



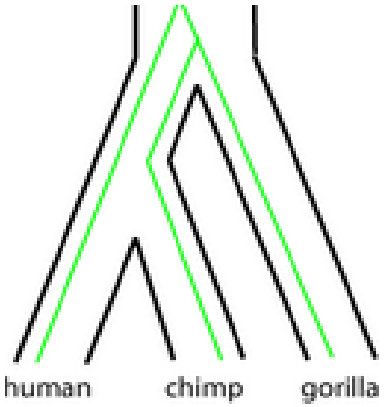
(H,C):G

Gene tree 2
(H,G)



(H,G):C

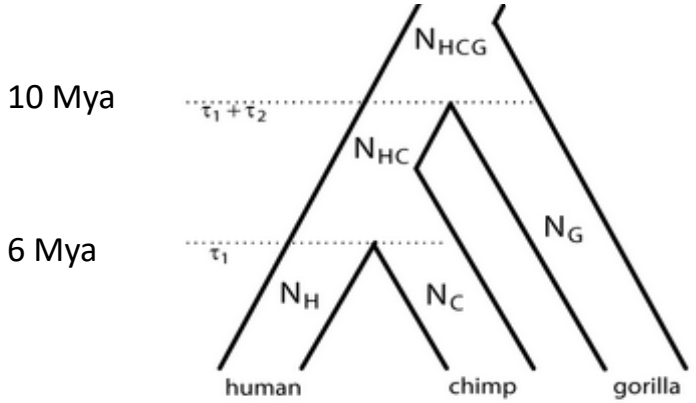
Gene tree 3
(C,G)



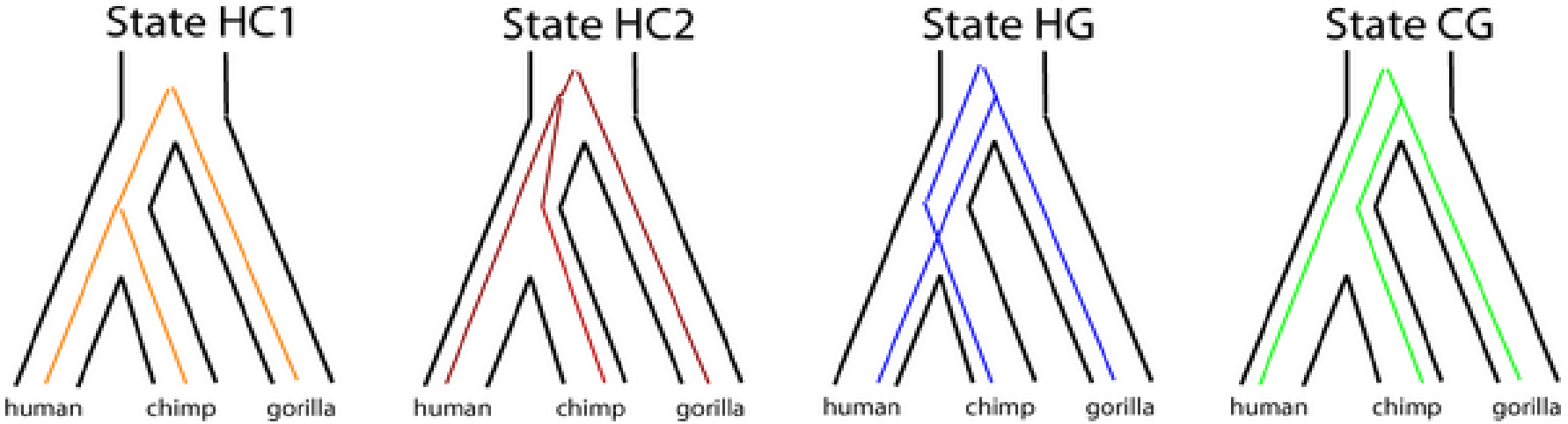
(C,G):H

Incomplete lineage sorting: Gene trees at a particular gene favor a topology different from the species tree.

Species tree



1 single species tree
3 possible gene tree topologies



(H,C):G
70%

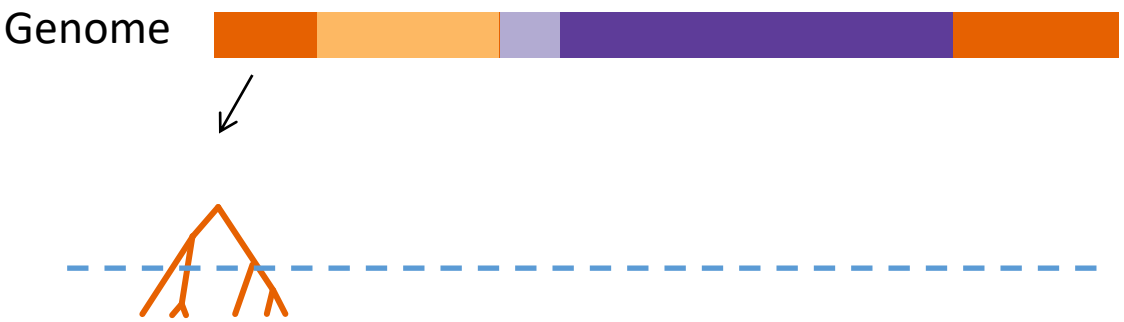
(H,G):C
15%

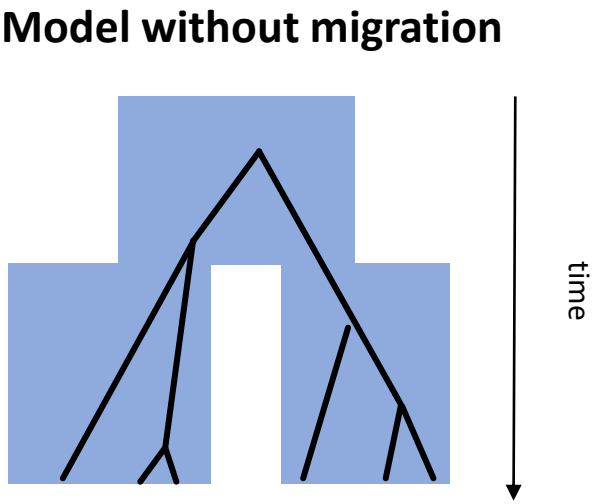
(C,G):H
15%

Scally et al. (2012) Nature 483, 169–175
Hobolth et al (2007) PLOS Genetics 3(2): e7. doi:10.1371/journal.pgen.0030007

Reconstructing the demographic history from genomic data

Because of recombination, different regions of the genome can have different gene trees



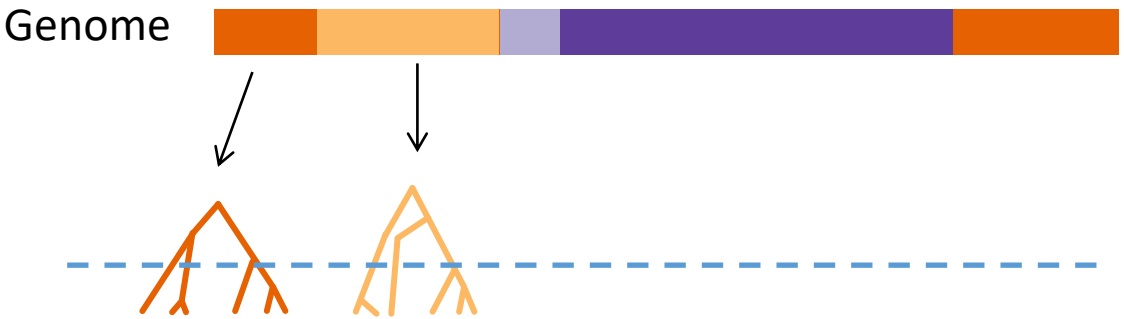


- Demography is expected to affect the entire genome
- Natural selection acts on specific functional regions

All gene trees are consistent with the population tree. Independent gene trees can be seen as independent replicates of the same population tree.

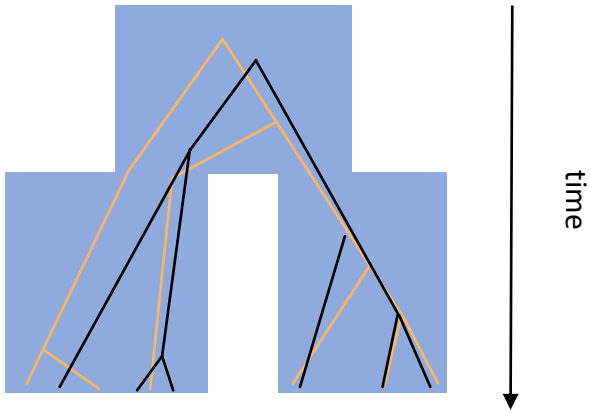
Reconstructing the demographic history from genomic data

Because of recombination, different regions of the genome can have different gene trees



- Demography is expected to affect the entire genome
- Natural selection acts on specific functional regions

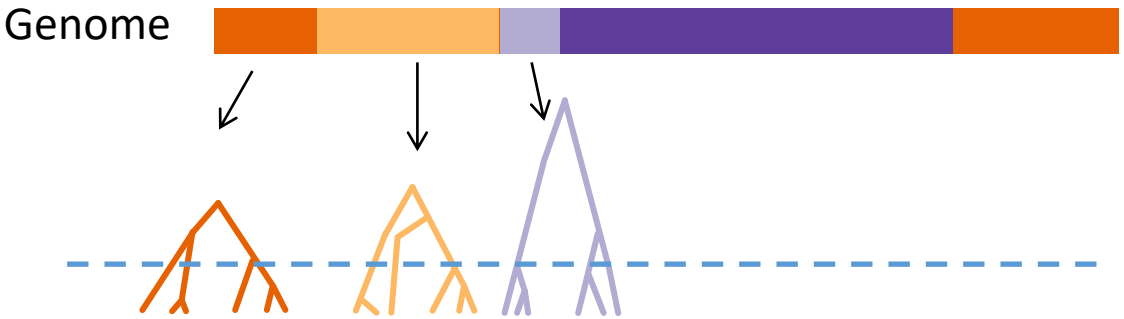
Model without migration



All gene trees are consistent with the population tree. Independent gene trees can be seen as independent replicates of the same population tree.

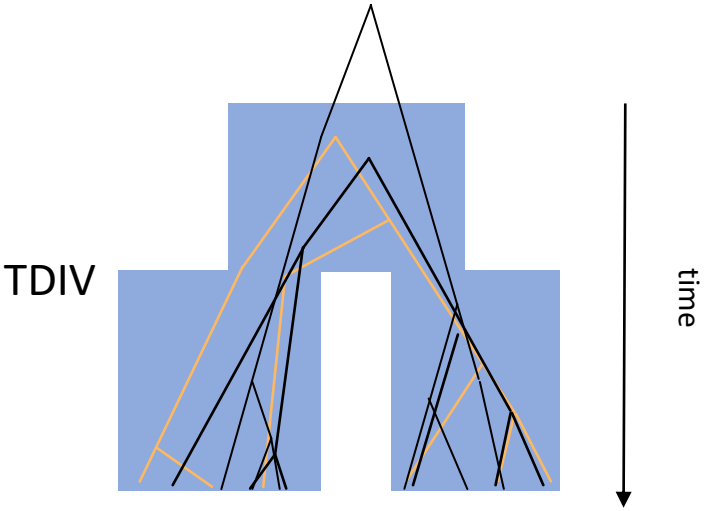
Reconstructing the demographic history from genomic data

Because of recombination, different regions of the genome can have different gene trees



- Demography is expected to affect the entire genome
- Natural selection acts on specific functional regions

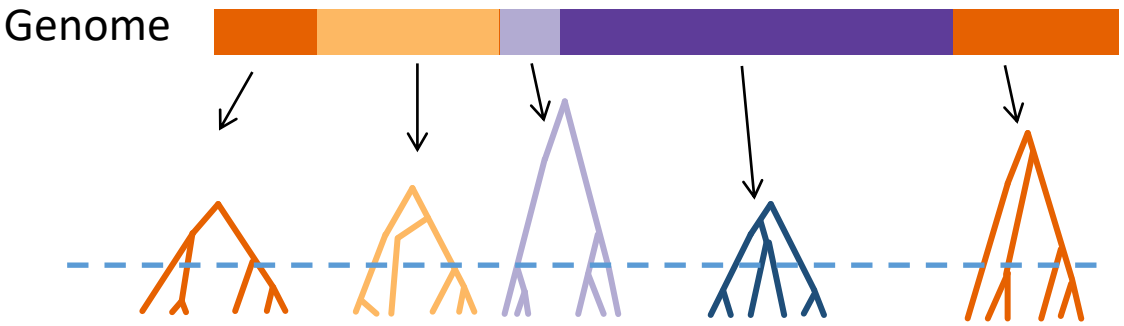
Model without migration



All gene trees are consistent with the population tree. Independent gene trees can be seen as independent replicates of the same population tree.

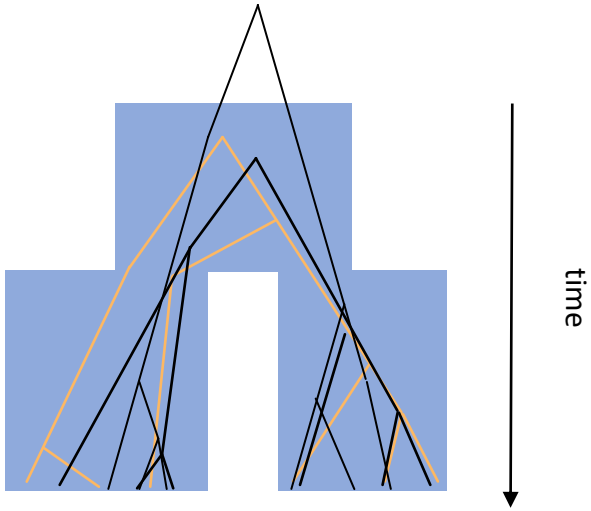
Reconstructing the demographic history from genomic data

Because of recombination, different regions of the genome can have different gene trees



- Demography is expected to affect the entire genome
- Natural selection acts on specific functional regions

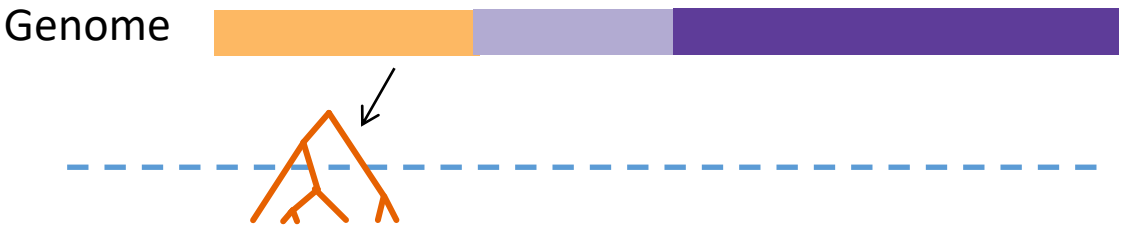
Model without migration



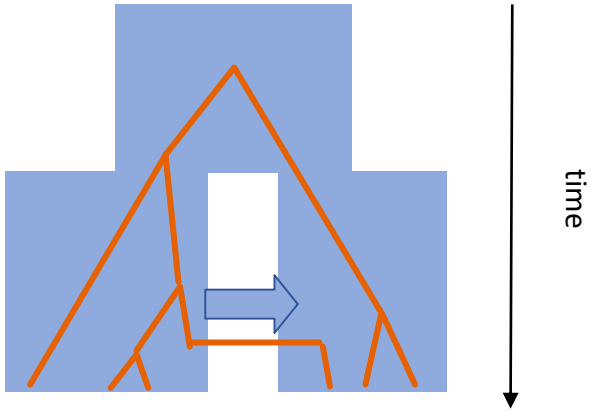
All gene trees are consistent with the population tree. Independent gene trees can be seen as independent replicates of the same population tree.

Reconstructing the demographic history from genomic data

Because of recombination, different regions of the genome can have different gene trees



Model with migration

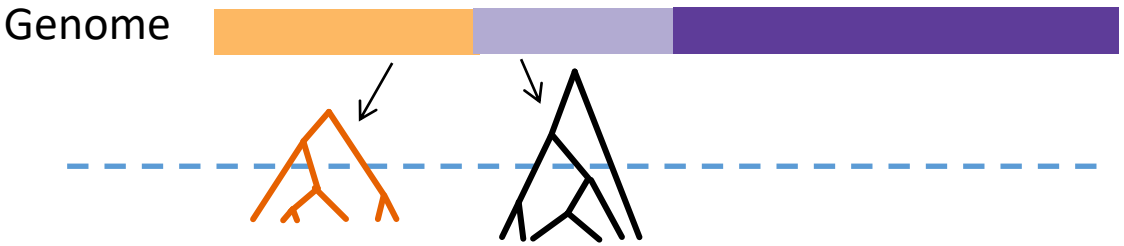


- Demography is expected to affect the entire genome
- Natural selection acts on specific functional regions

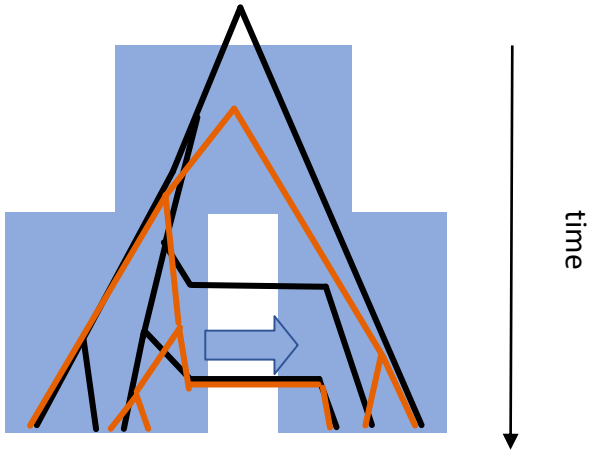
All gene trees are consistent with the population tree. Independent gene trees can be seen as independent replicates of the same population tree.

Reconstructing the demographic history from genomic data

Because of recombination, different regions of the genome can have different gene trees



Model with migration

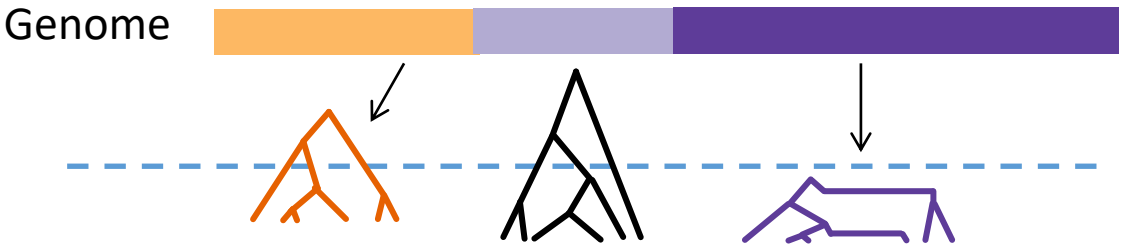


- Demography is expected to affect the entire genome
- Natural selection acts on specific functional regions

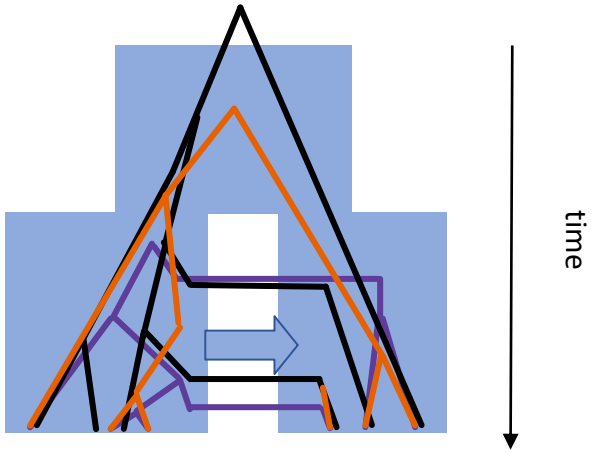
All gene trees are consistent with the population tree. Independent gene trees can be seen as independent replicates of the same population tree.

Reconstructing the demographic history from genomic data

Because of recombination, different regions of the genome can have different gene trees



Model with migration

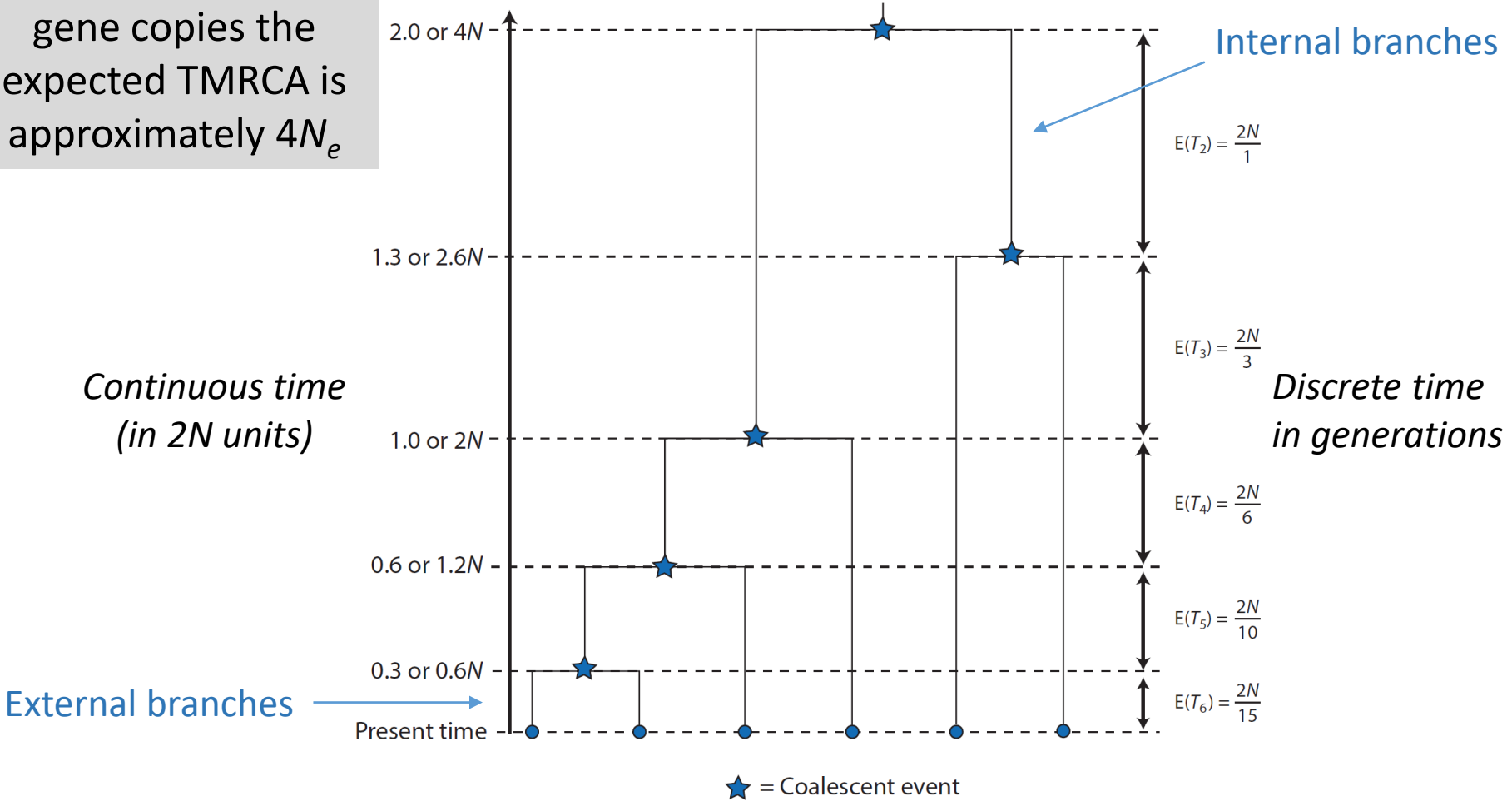


- Demography is expected to affect the entire genome
- Natural selection acts on specific functional regions

All gene trees are consistent with the population tree. Independent gene trees can be seen as independent replicates of the same population tree.

Expected coalescent times in a single constant size population

For a sample of *n* gene copies the expected TMRCA is approximately $4N_e$



Activity

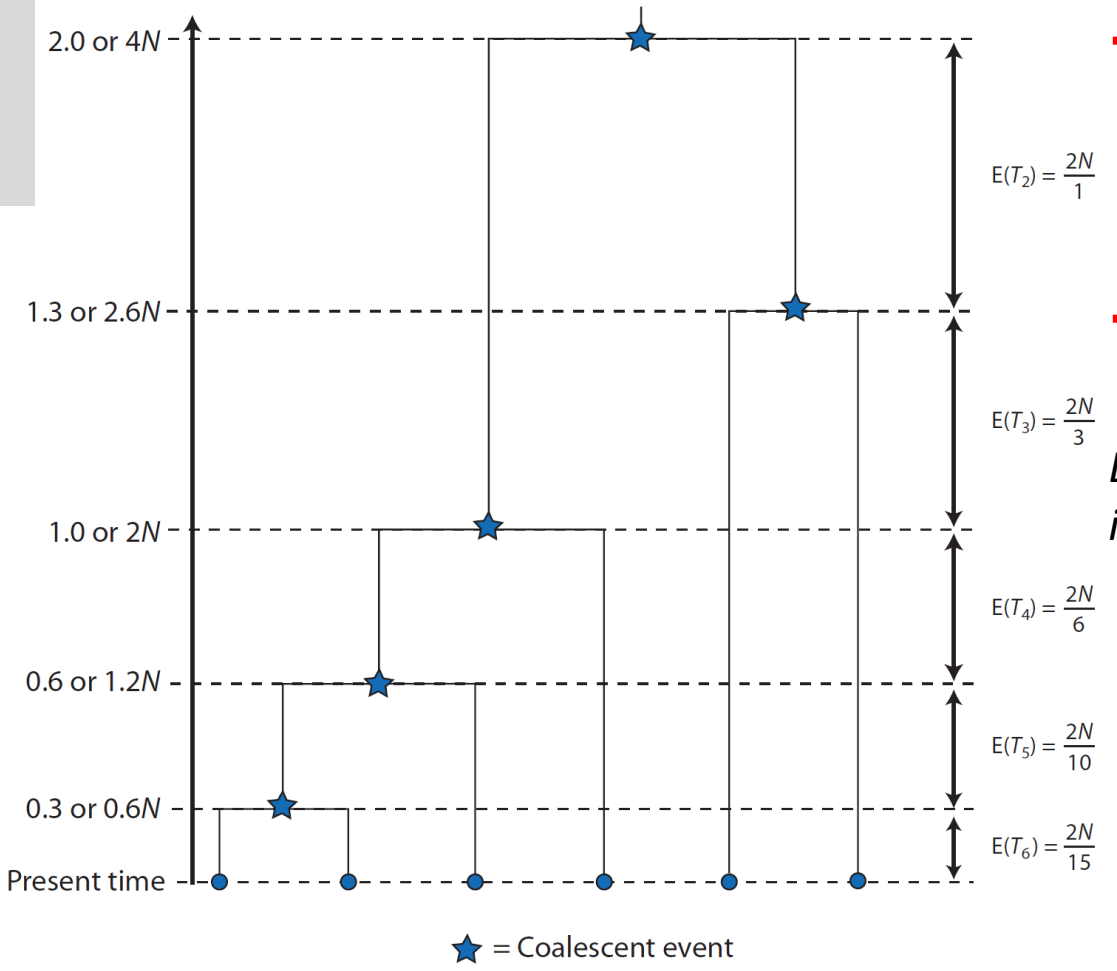
Check your intuition about coalescent gene trees

- What are the longest branches we expect in a single constant size population?
  - A: External branches (tips of gene tree)
  - B: Internal branches

- Do we expect the relative branch length to differ in large and small populations?
  - Yes
  - No

Expected coalescent times in a single constant size population

For a sample of *n* gene copies the expected TMRCA is approximately $4N_e$



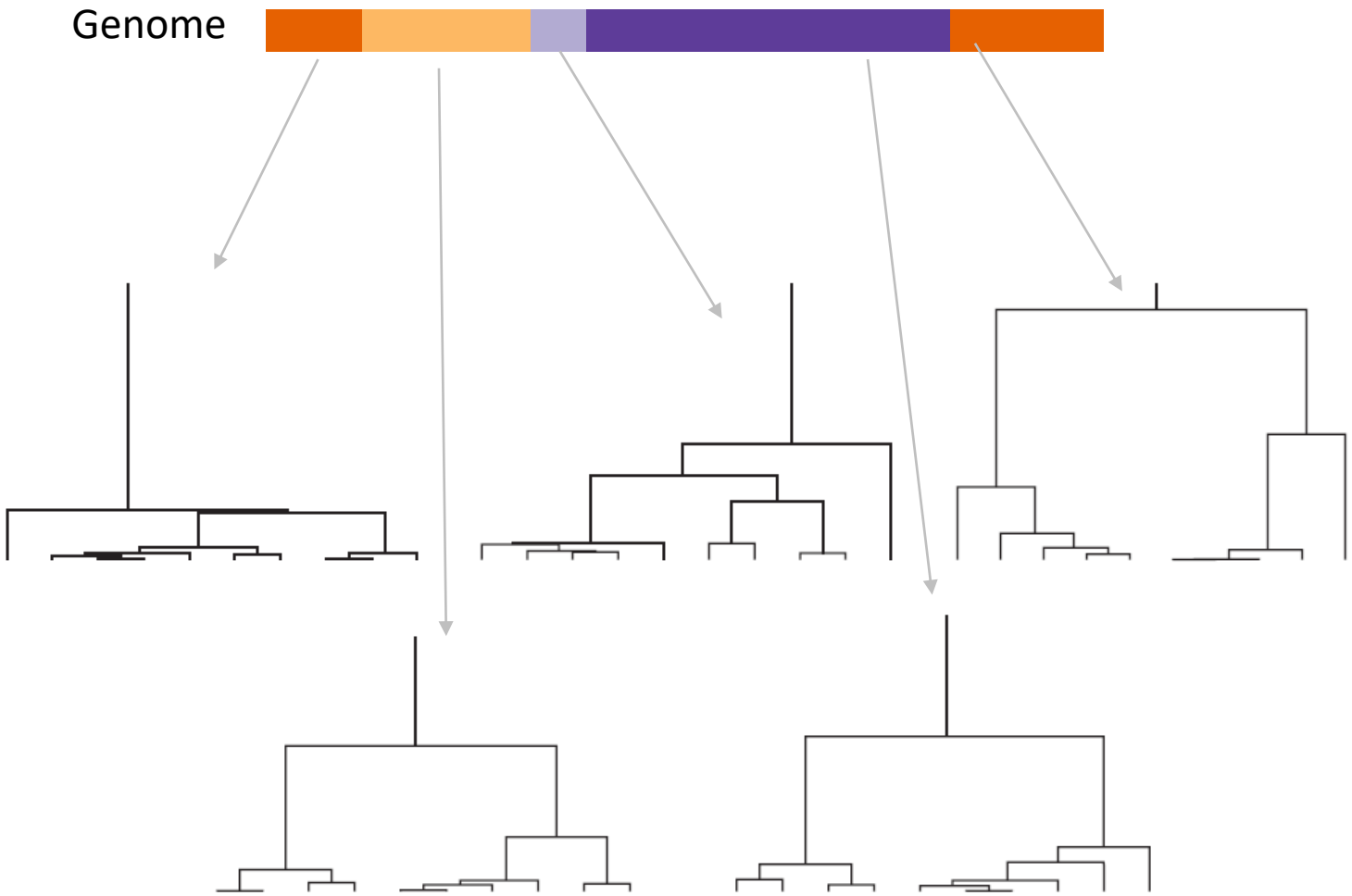
E[T_2]=2Ne

Continuous time (in 2N units)

Discrete time in generations

Long internal branches - when there is only 2 lineages left, we expect them to take 2Ne generations to coalesce

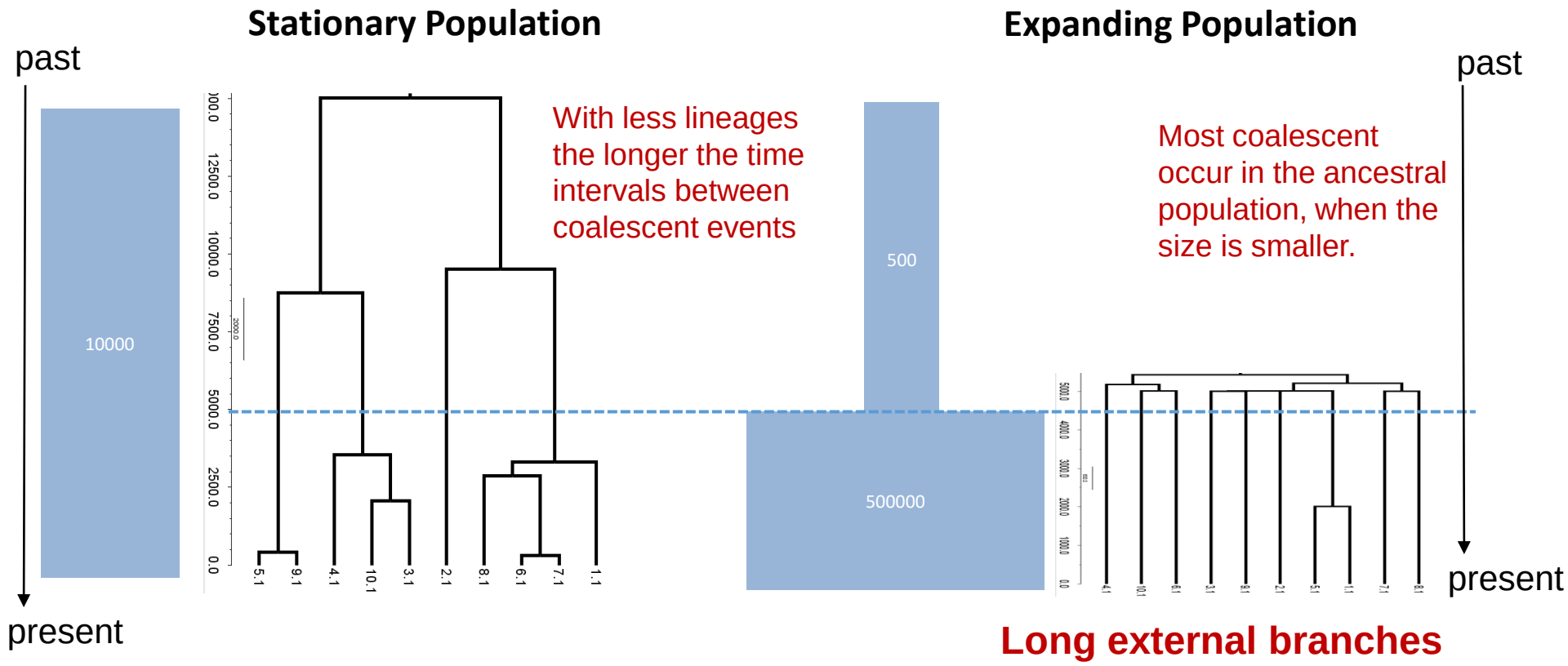
The expected TMRCA is 4Ne, but there is a large variance!



Five independent genomic regions from the **same constant size population**.

Hein et al (2004) Gene Genealogies, Variation and evolution

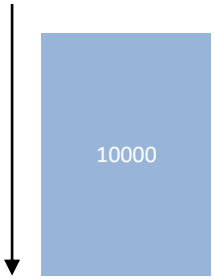
Gene trees in expanding populations



- Coalescent rate is larger in smaller populations, and so we expect smaller intervals between coalescent events in smaller populations
- Coalescent rate is lower with a lower number of lineages, and so we expected larger intervals between coalescent events as the number of lineages decrease

Stationary population

gene trees at five genome regions
(all share same population history!)



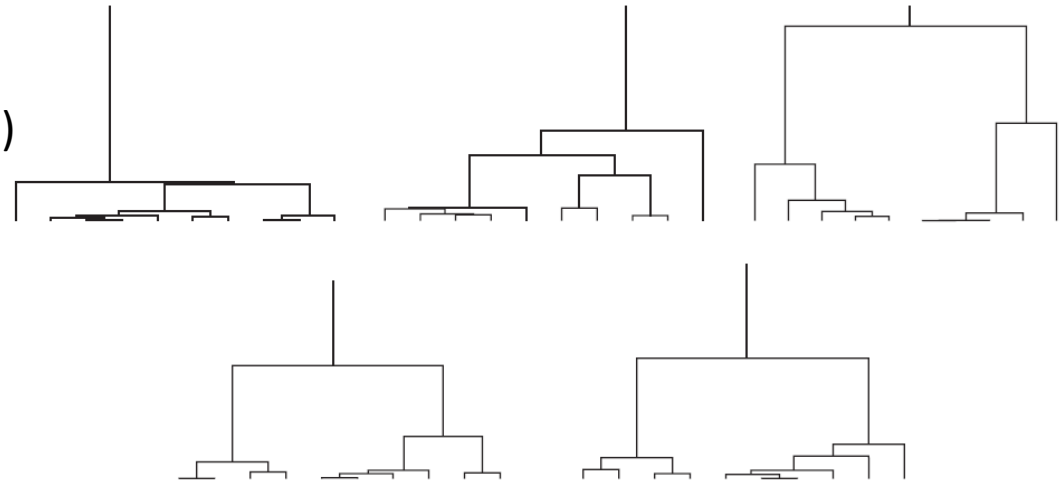
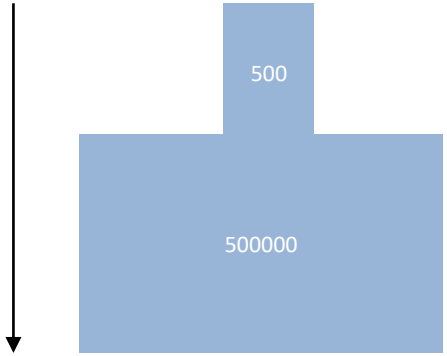


Figure 4.2 Five replicates of the coalescent process with constant population size for a sample of ten genes. Note the large variance in the time of the MRCA among replicates.

Expanding population

gene trees at five genome regions
(all share same population history!)



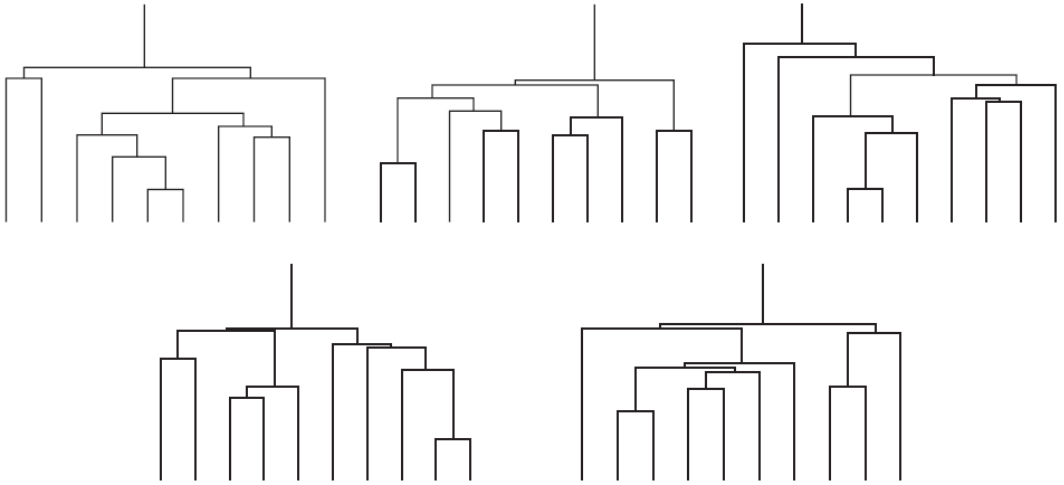
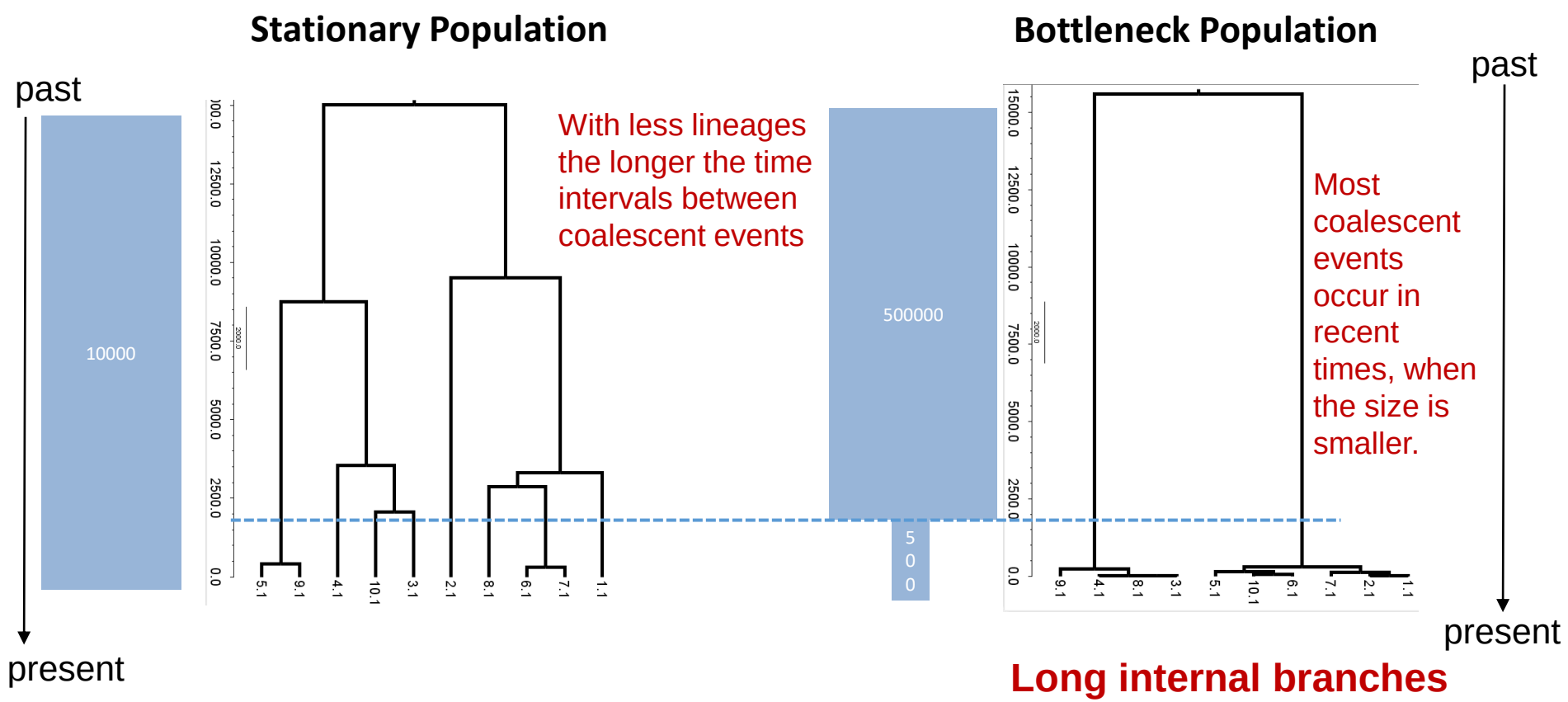


Figure 4.3 Five replicates of the coalescent with exponential growth, $\beta = 1000$, for a sample of $n = 10$ genes. Note the smaller variance in the time until the MRCA compared to the same quantity in Figure 4.2.

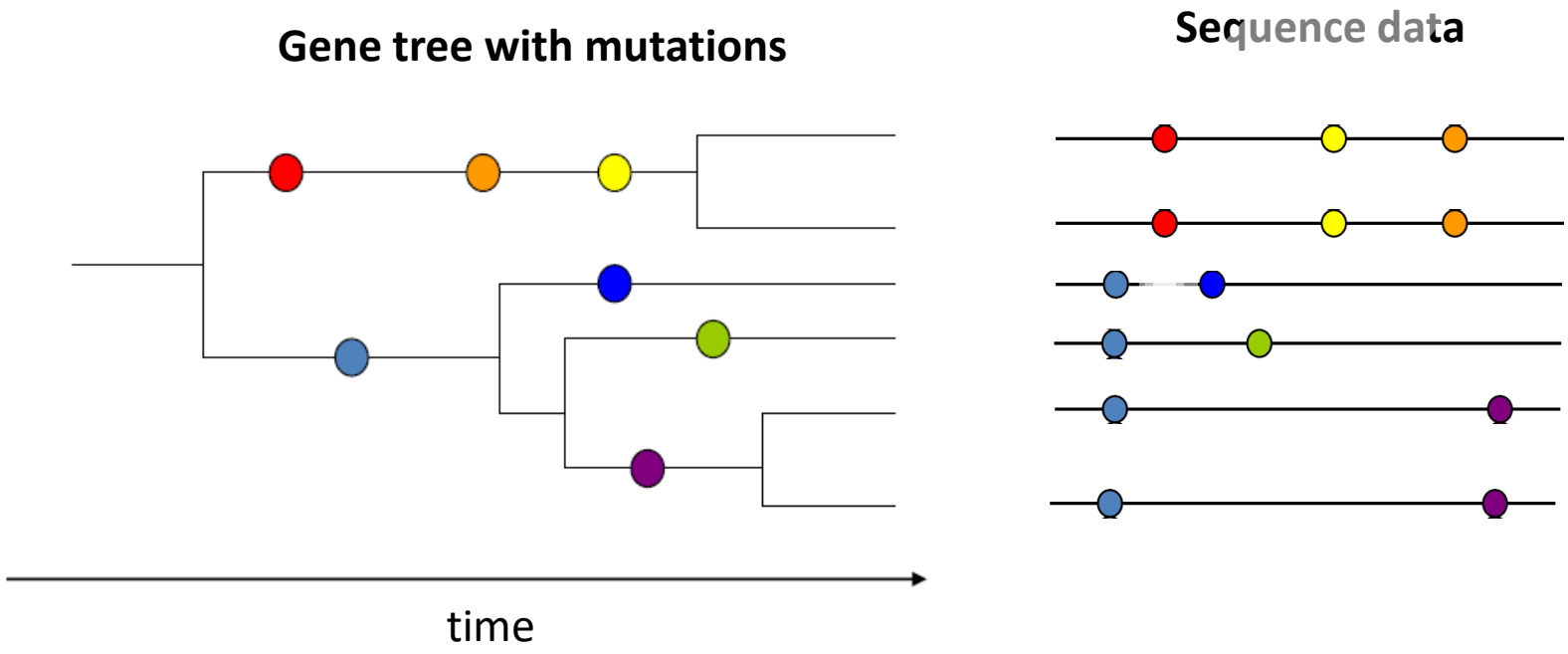
Hein et al (2004) Gene Genealogies, Variation and evolution

Gene trees for decreasing populations



- If we could observe directly the gene trees, we could easily reconstruct the population tree and the demographic history.
- But we do not observe gene trees…
- We can still learn about gene trees from the observed mutations and the allele frequencies in samples

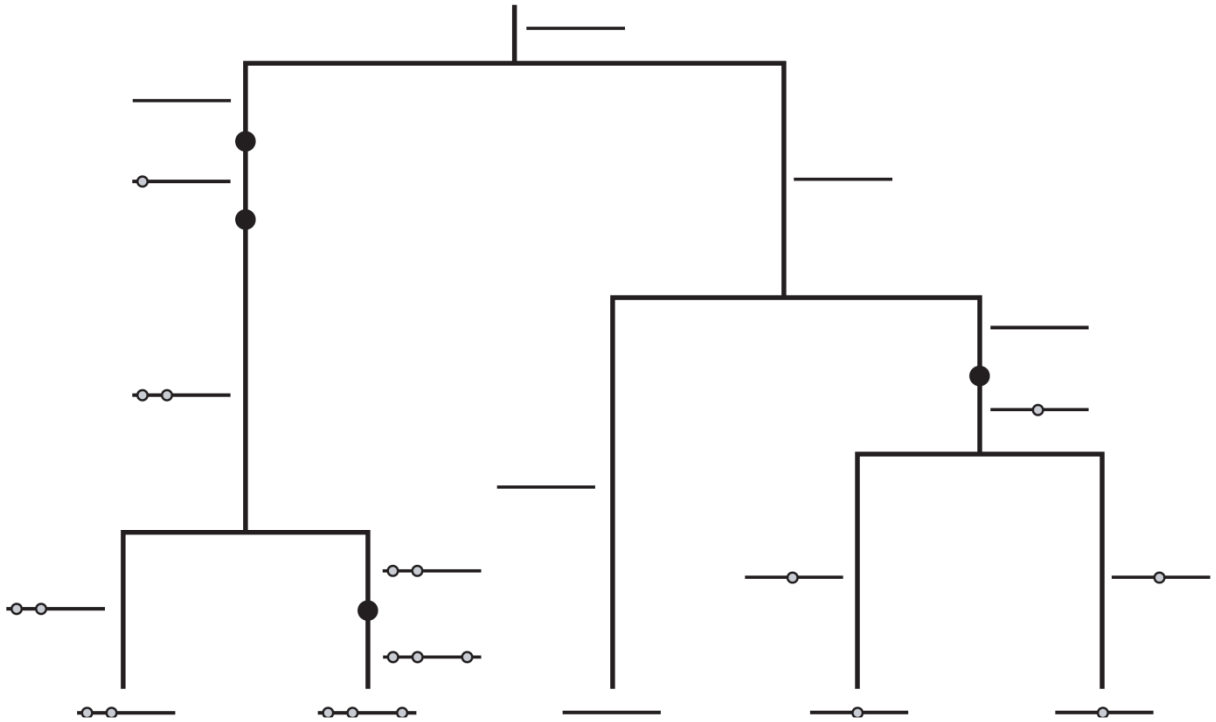
Adding neutral mutations to gene trees under the Infinite sites model



No back mutations, no multiple mutations on the same site.

Adding neutral mutations

The shape of neutral coalescent trees only depend on the population demography, and not on the mutational process. Assuming that all alleles have the same fitness, the mutational process can be modeled as an independent process superimposed on a realized coalescent tree.

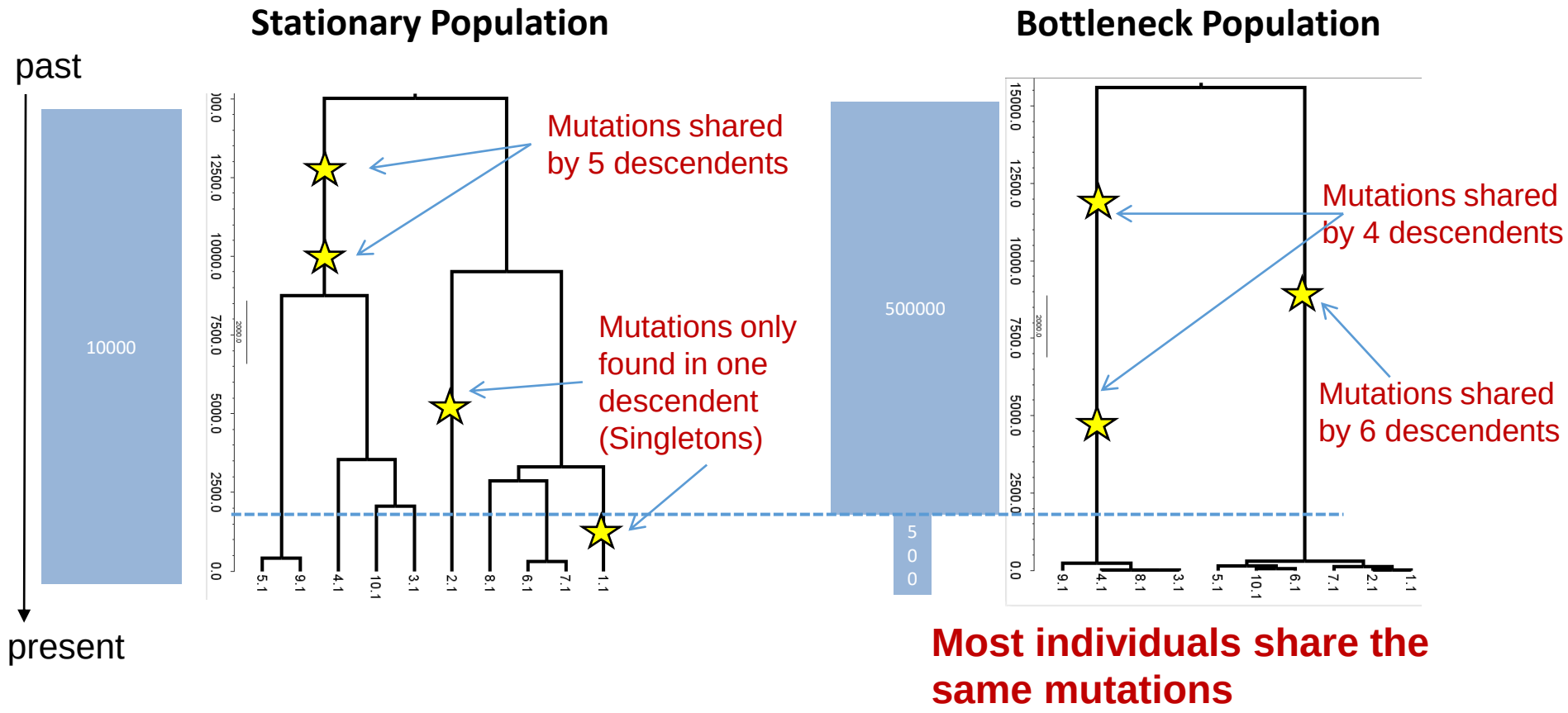


Mutations just accumulate along the branches of the tree according to a **Poisson process** with rate $\lambda_i = \mu t_i$ for the *i*-th branch of length t_i. The Poisson process is stochastic but it should be immediately **obvious that long branches will carry more mutations than short branches**

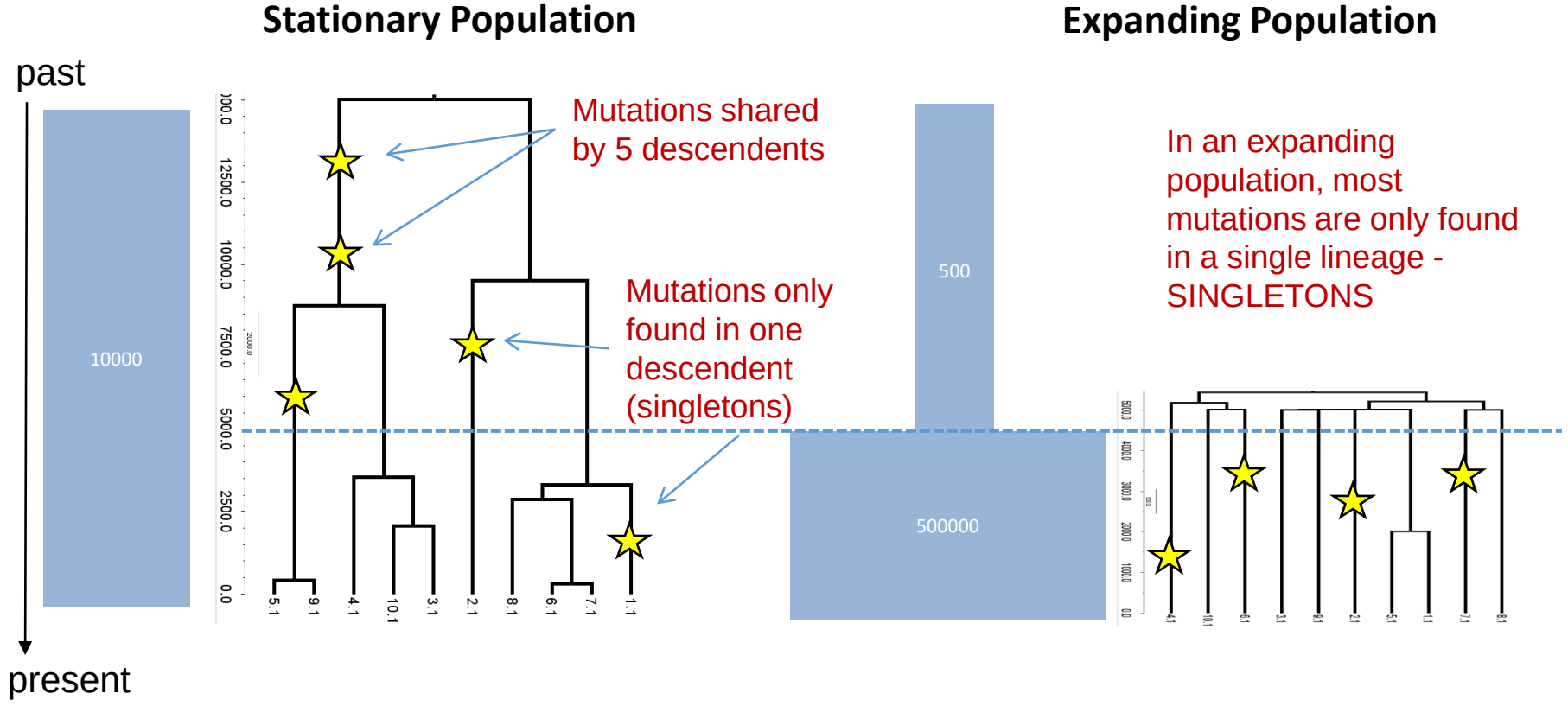
Hein et al (2004) Gene Genealogies, Variation and evolution

We expect less rare variants in populations that went through a bottleneck

- Mutations accumulate along the branches.
- The longer a given branch the more likely it becomes that a mutation have happened on it.

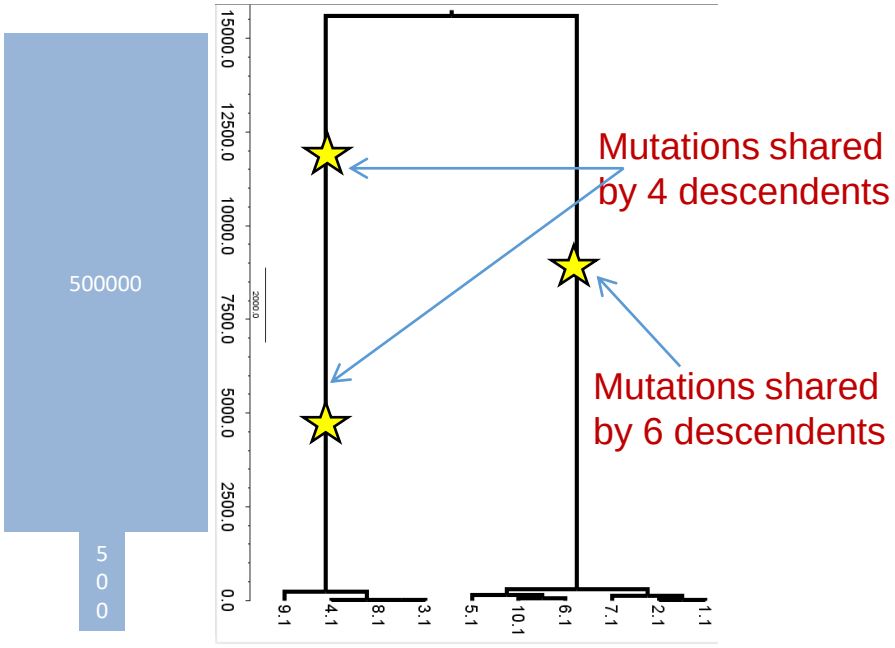


We expect more rare variants in expanding populations than in populations with a constant size



Activity
What would be the Tajima's D for these bottleneck and expansion scenarios?

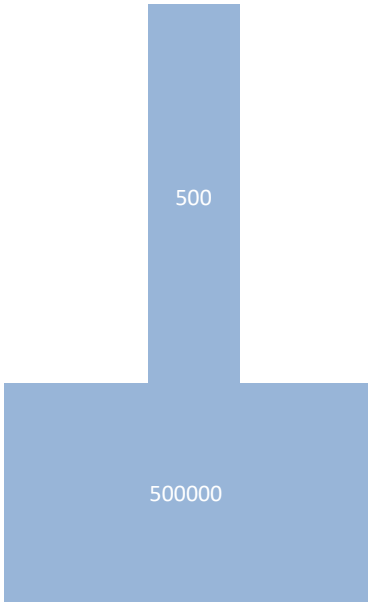
Bottleneck Population



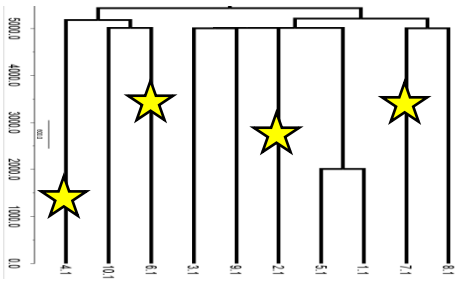
Mutations shared by 4 descendents

Mutations shared by 6 descendents

Expanding Population

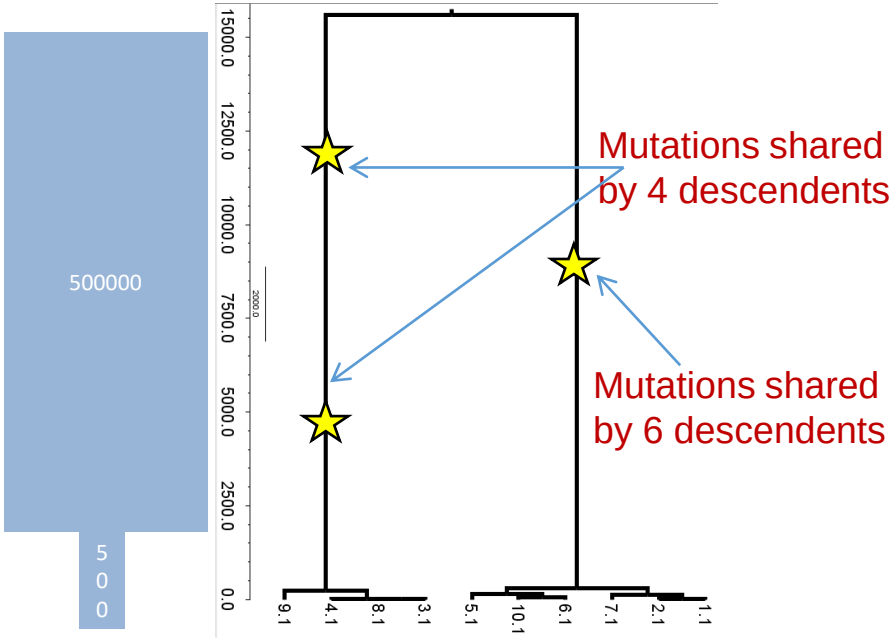


In an expanding population, most mutations are only found in a single lineage - SINGLETONS



Activity
What would be the Tajima's D for these bottleneck and expansion scenarios?

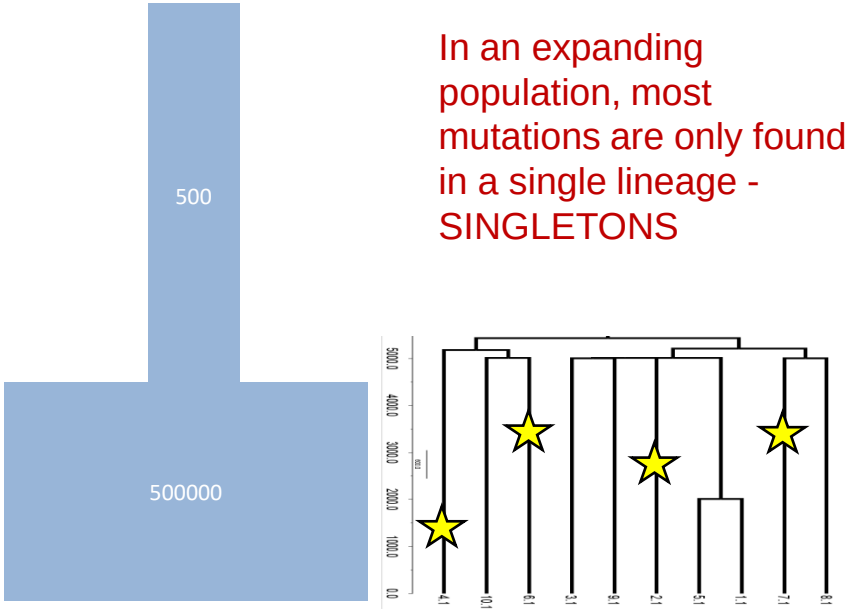
Bottleneck Population



Tajima's D > 0

$\pi > \Theta_w$

Expanding Population



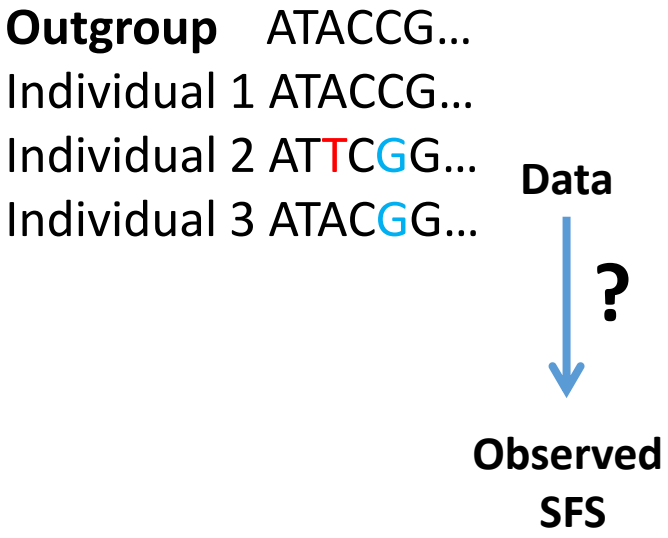
Tajima's D < 0

$\pi < \boldsymbol{\theta_w}$

What is the expected site frequency spectrum?

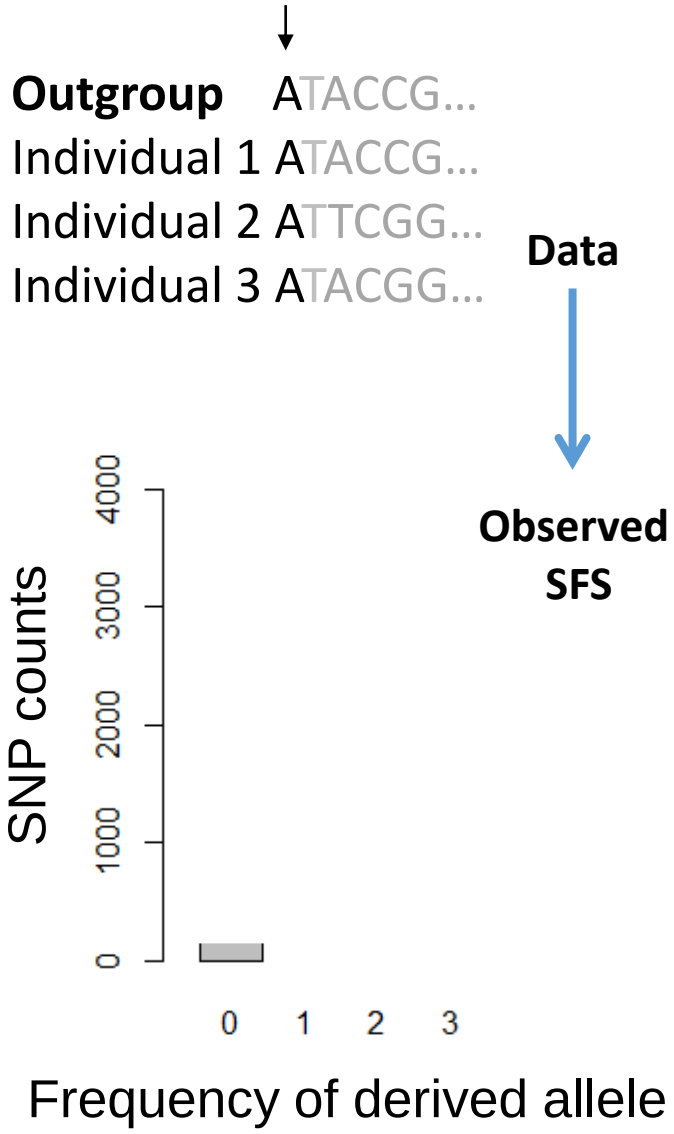
Site frequency spectrum (SFS)

- The SFS summarizes efficiently genome-wide data
- Assuming a single population – 1Dimensional SFS



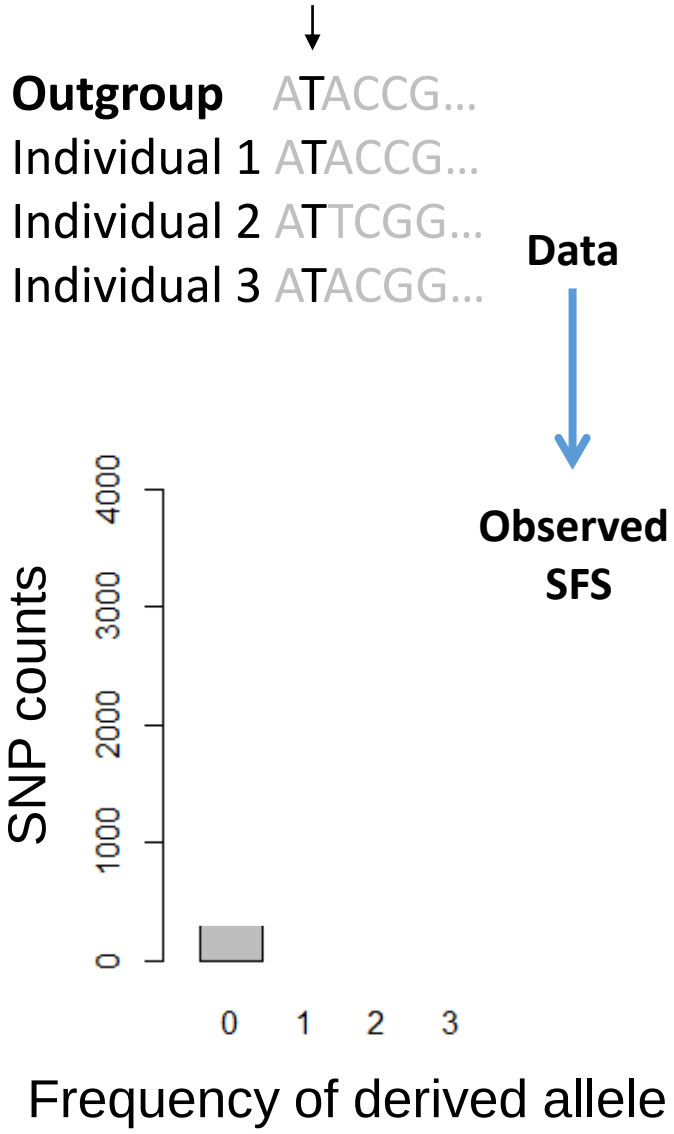
Site frequency spectrum (SFS)

- The SFS summarizes efficiently genome-wide data
- Assuming a single population – 1Dimensional SFS



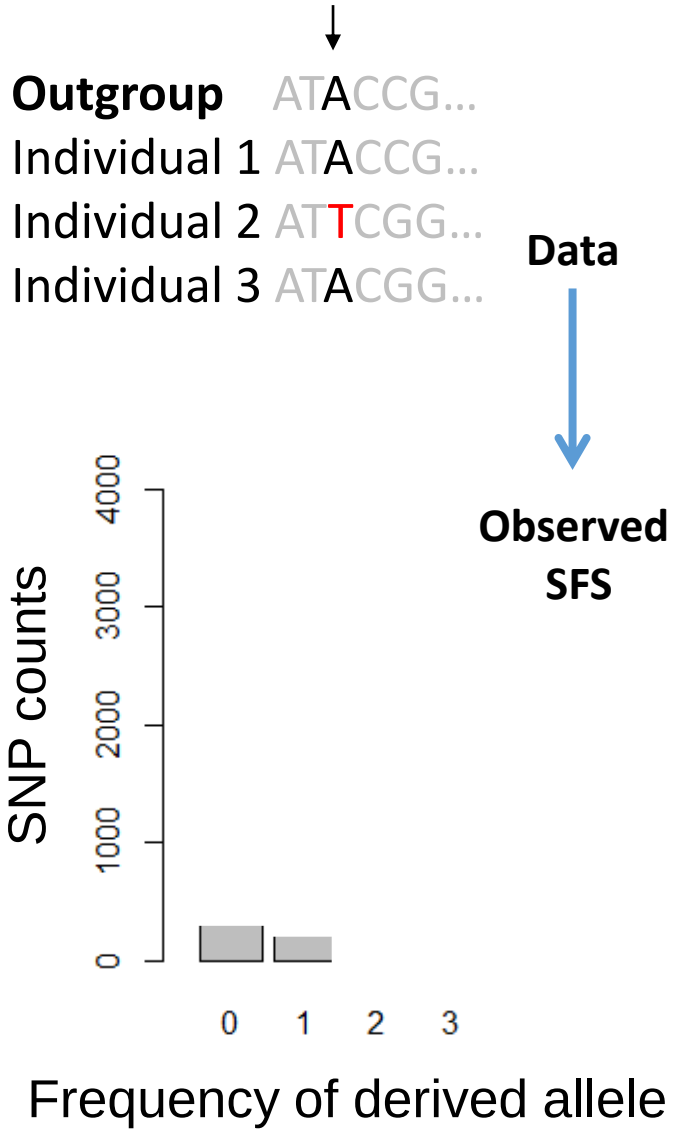
Site frequency spectrum (SFS)

- The SFS summarizes efficiently genome-wide data
- Assuming a single population – 1Dimensional SFS



Site frequency spectrum (SFS)

- The SFS summarizes efficiently genome-wide data
- Assuming a single population – 1Dimensional SFS



Site frequency spectrum (SFS)

- The SFS summarizes efficiently genome-wide data

- Assuming a single population – 1Dimensional SFS

The SFS ignores information about linkage. It is best suited for the study of many unlinked (or recombining) DNA sequences.

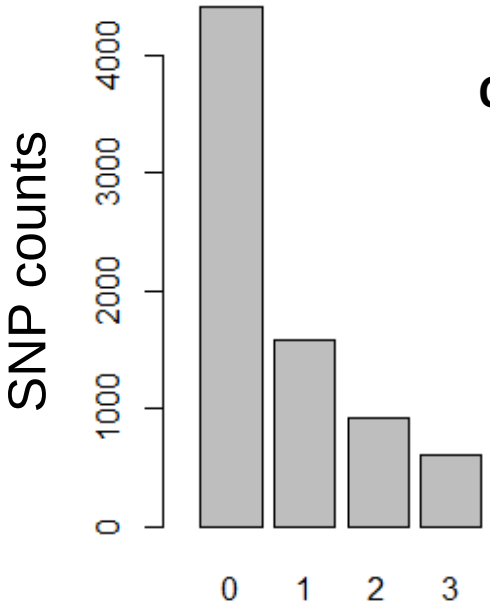
In a stationary population, the expected SFS relative frequencies are given by:

$$E(\xi_i) = \frac{\theta}{i}$$

Fu and Li, 1993

Outgroup ATACCG…
Individual 1 ATACCG…
Individual 2 ATTCGG…
Individual 3 ATACGG…

Data



Observed SFS

VCF (variant call format) files

CHROM	POS	ID	REF	ALT	QUAL	FILTER	FORMAT	BL2009P4_us23
"Supercontig_1.50"	"2"	NA	"T"	"A"	"44.44"	NA	"GT:AD:DP:GQ:PL"	"0|0:62,0:62:99:0,190,2835"
"Supercontig_1.50"	"246"	NA	"C"	"G"	"144.21"	NA	"GT:AD:DP:GQ:PL"	"1|0:5,5:10:99:111,0,114"
"Supercontig_1.50"	"549"	NA	"A"	"C"	"68.49"	NA	"GT:AD:DP:GQ:PL"	NA
"Supercontig_1.50"	"668"	NA	"G"	"C"	"108.07"	NA	"GT:AD:DP:GQ:PL"	"0|0:1,0:1:3:0,3,44"
"Supercontig_1.50"	"765"	NA	"A"	"C"	"92.78"	NA	"GT:AD:DP:GQ:PL"	"0|0:2,0:2:6:0,6,49"
"Supercontig_1.50"	"780"	NA	"G"	"T"	"58.38"	NA	"GT:AD:DP:GQ:PL"	"0|0:2,0:2:6:0,6,49"

https://grunwaldlab.github.io/Population_Genetics_in_R/reading_vcf.html

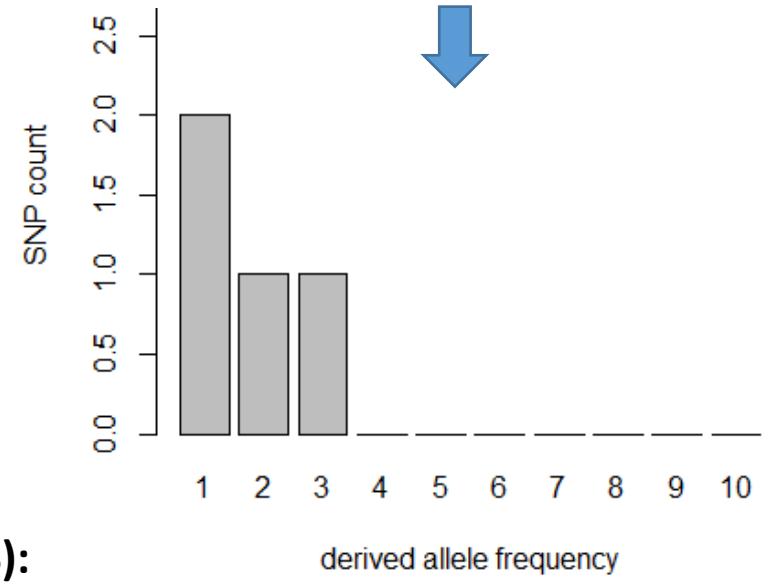
We can obtain the SFS from genotype call data

Genotypes:

- 0 homozygote for reference allele
- 1 heterozygote
- 2 homozygote for alternative allele

This can be done if we have enough depth of coverage (>10x)

	SNP1	SNP2	SNP3	SNP4
Individual 1	0	2	0	1
Individual 2	0	0	1	0
Individual 3	1	0	0	0
Individual 4	0	1	0	0
Individual 5	0	0	1	0



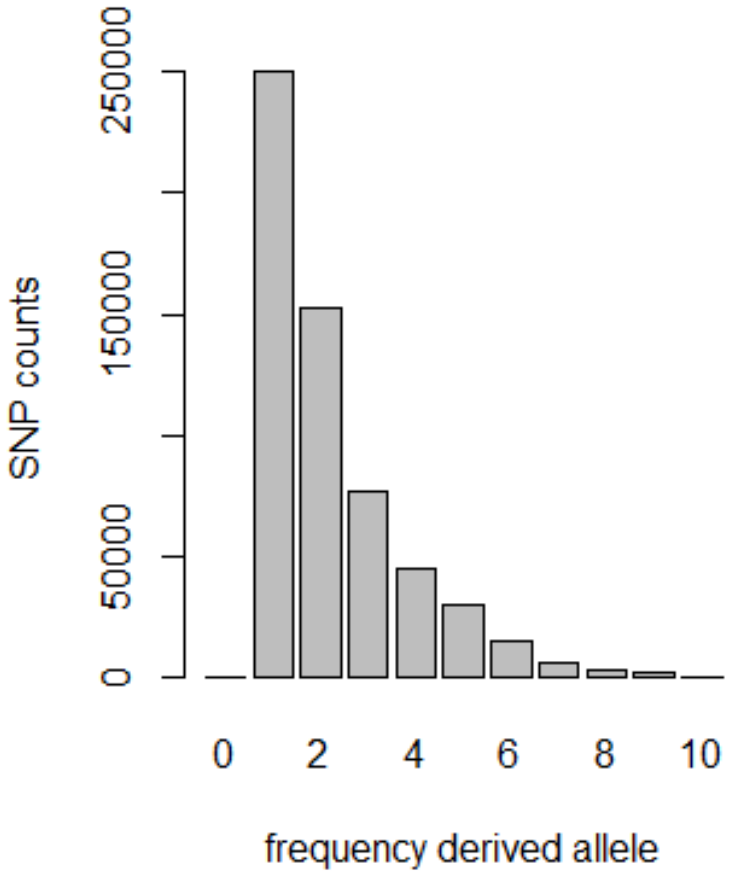
Observed SFS is a vector (1 dimensional SFS):

Frequency	0	1	2	3	4	5	6	7	8	9	10
SNP count	0	2	1	1	0	0	0	0	0	0	0

SFS from genotype call data

Even if we have millions of SNPs we can summarize the genomic data to 10 numbers with the SFS!

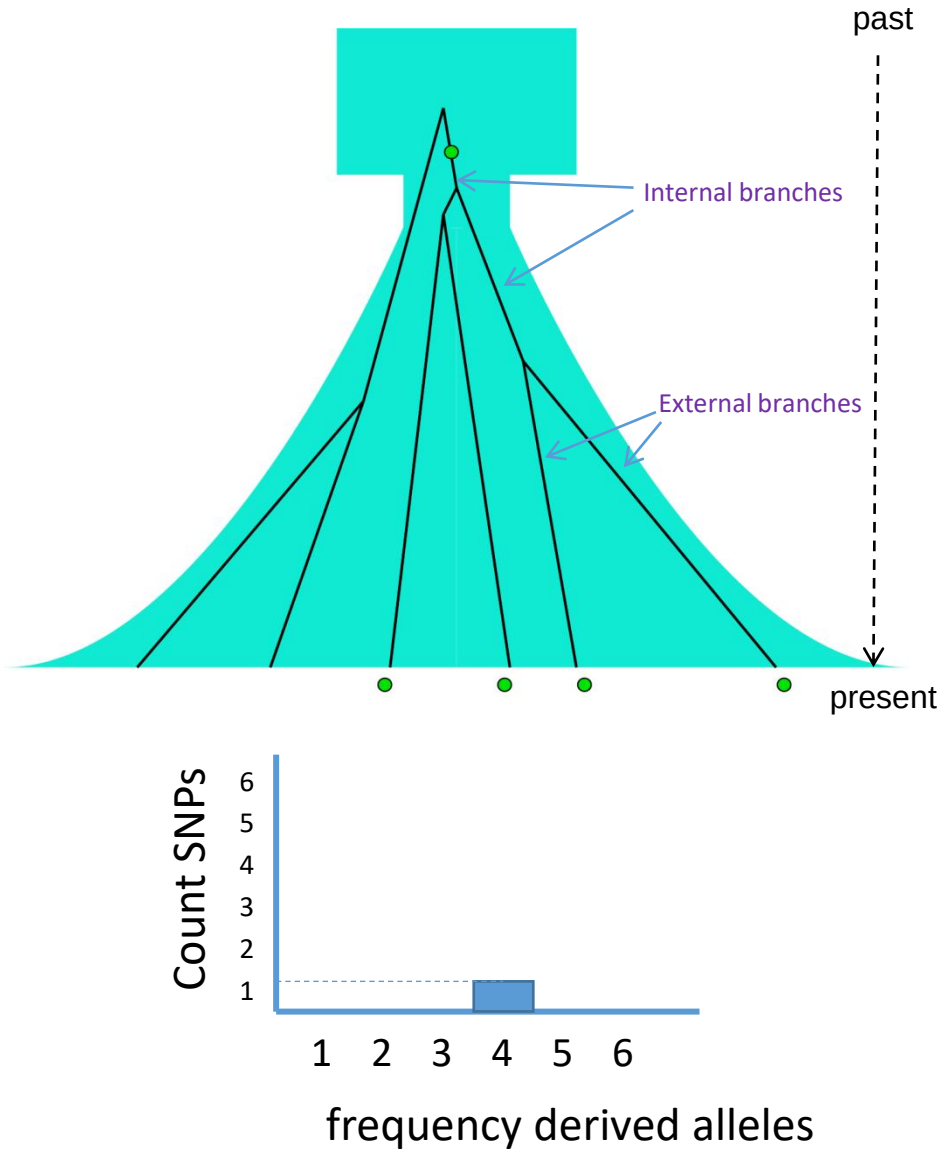
The size of the SFS depends on the number of sampled individuals.



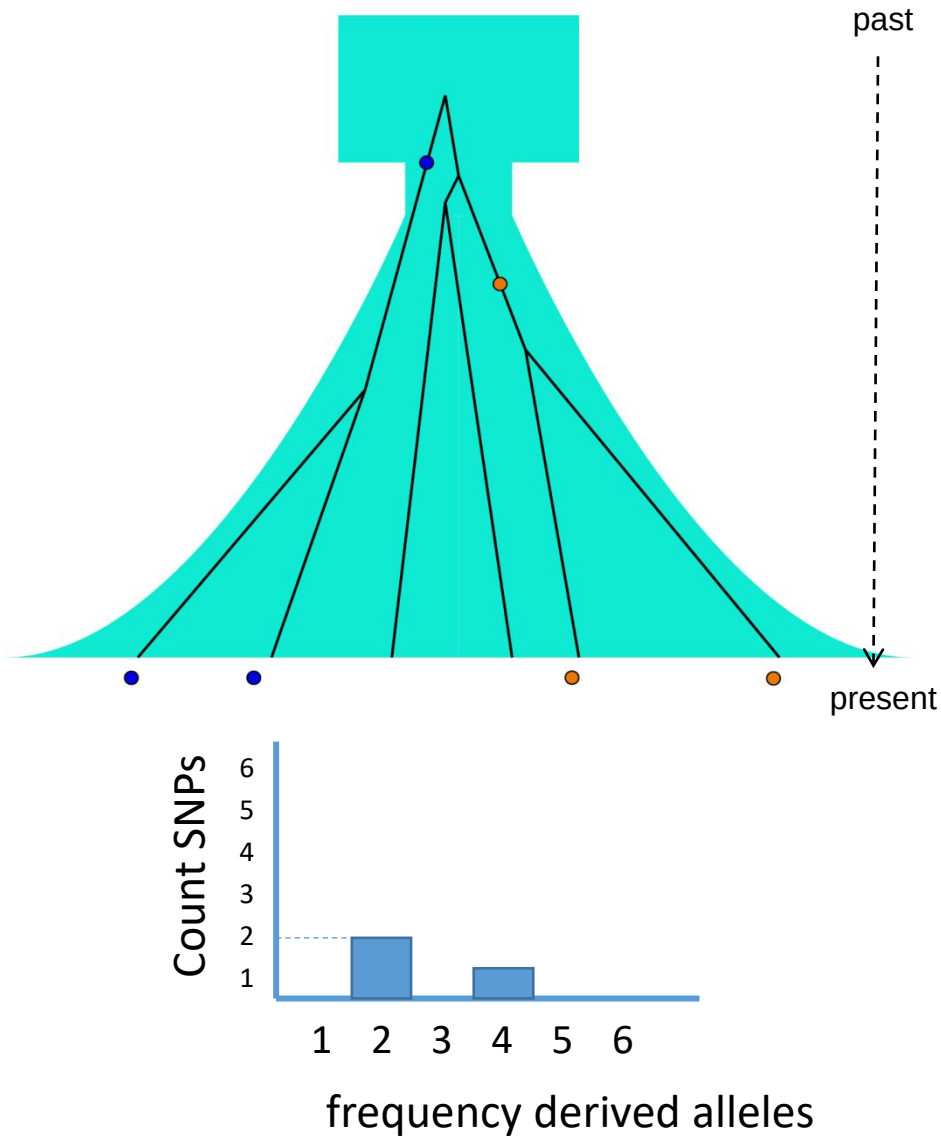
Observed SFS is a vector (1 dimensional SFS):

Frequency	0	1	2	3	4	5	6	7	8	9	10
SNP count	0	250,032	152,300	76,504	45,362	30,210	15,329	5,642	3,524	2,123	0

Coalescent and the SFS

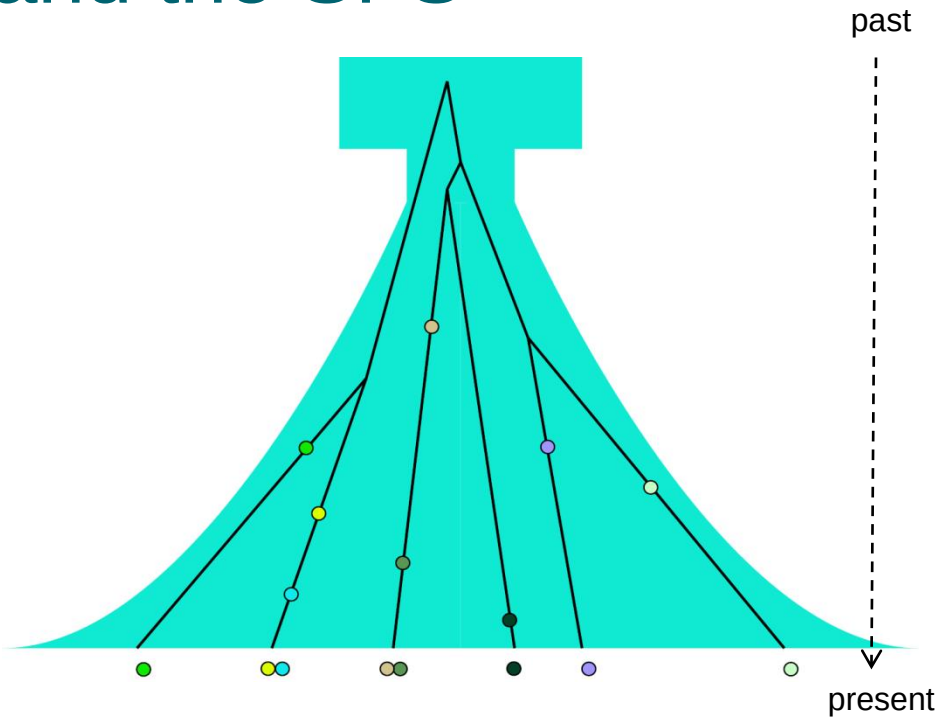


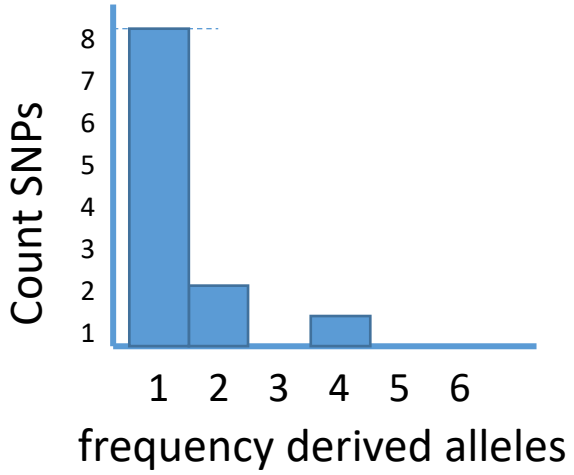
Coalescent and the SFS



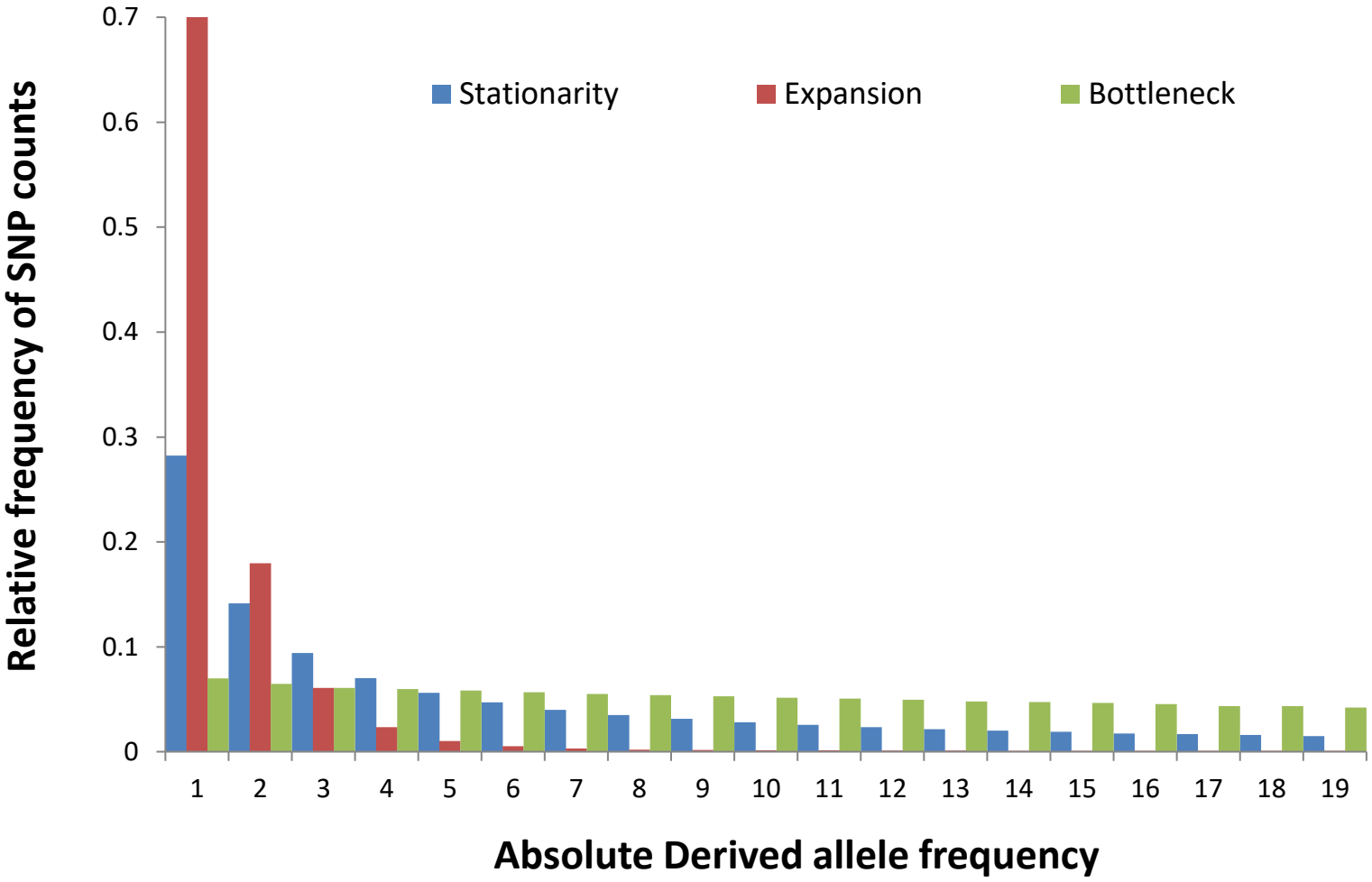
Coalescent and the SFS

- A recent population growth following a bottleneck leads to gene trees with long external branches
- Very few mutations in the internal branches
- Most mutations in long external branches are only found in one lineage, resulting in an excess of singletons

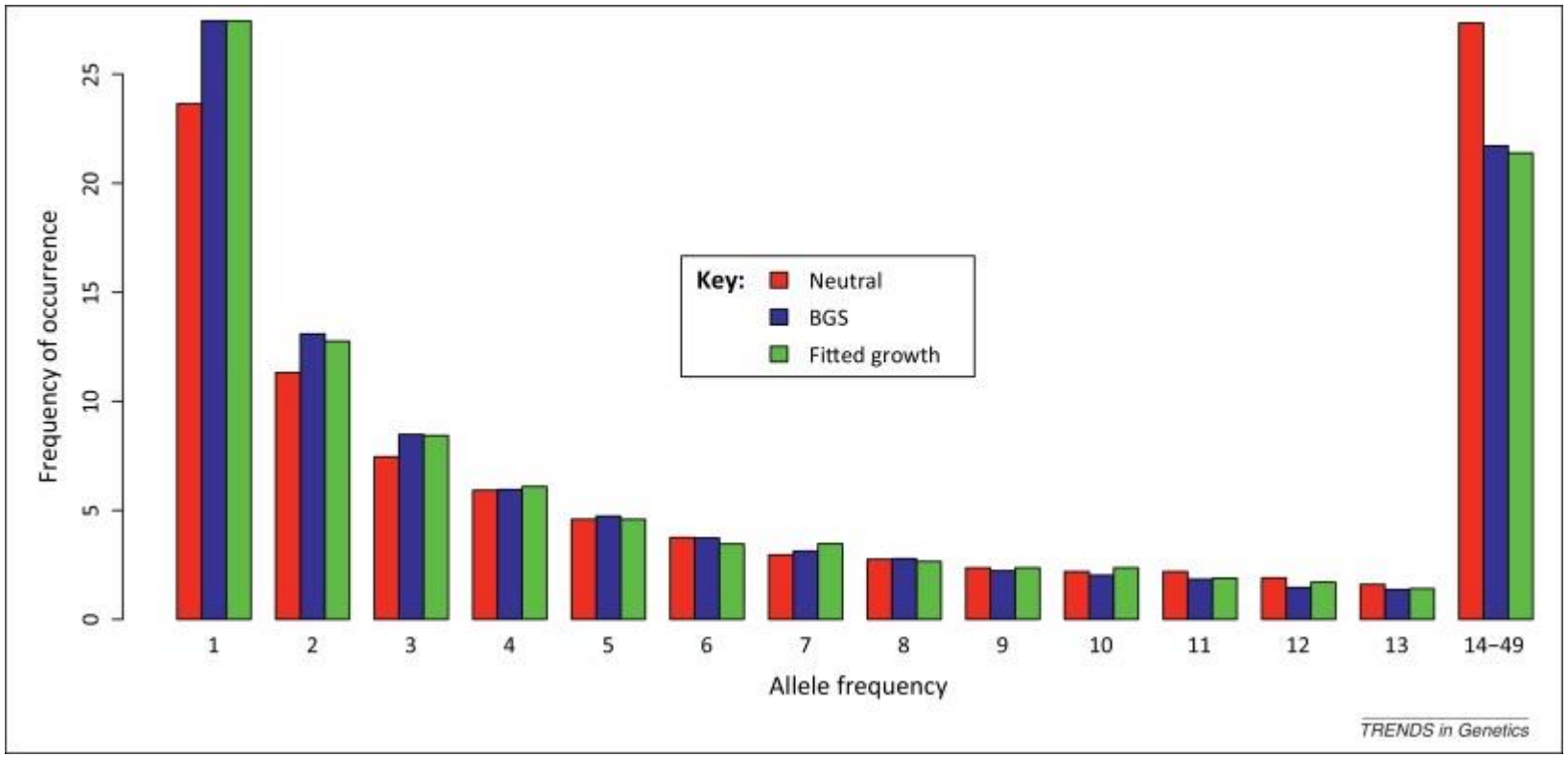




SFS depends on past demography



Natural selection also affects the SFS



Background selection (BGS) leads to patterns similar to population expansion.

Bank et al (2014) Trends in Genetics

Population structure

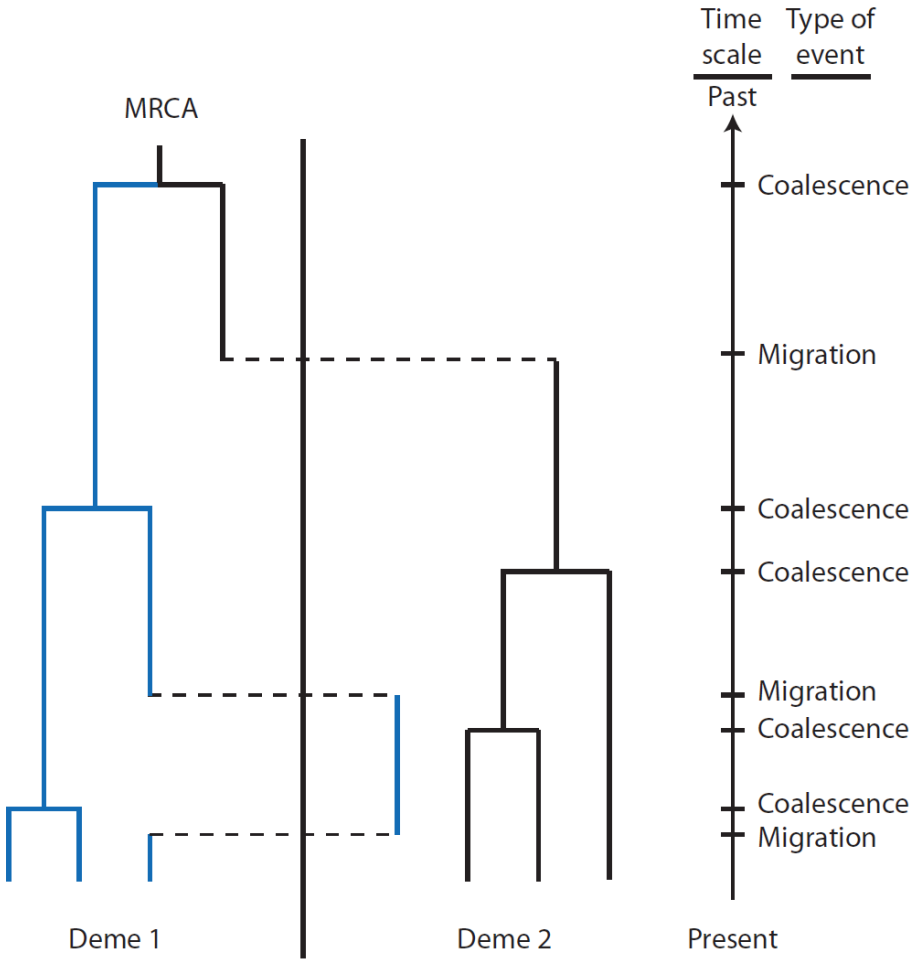
Migration events can be incorporated into gene trees.

Migration from Pop 2 to Pop 1, leads to lineages moving from Pop 1 to pop 2 backward in time.

At each generation, the probability of immigration into population 1 from population 2 is given by:

Pr(migrate) = $n_1 * m$

Where n_1 is the number of lineages in population 1, and m is the immigration rate.



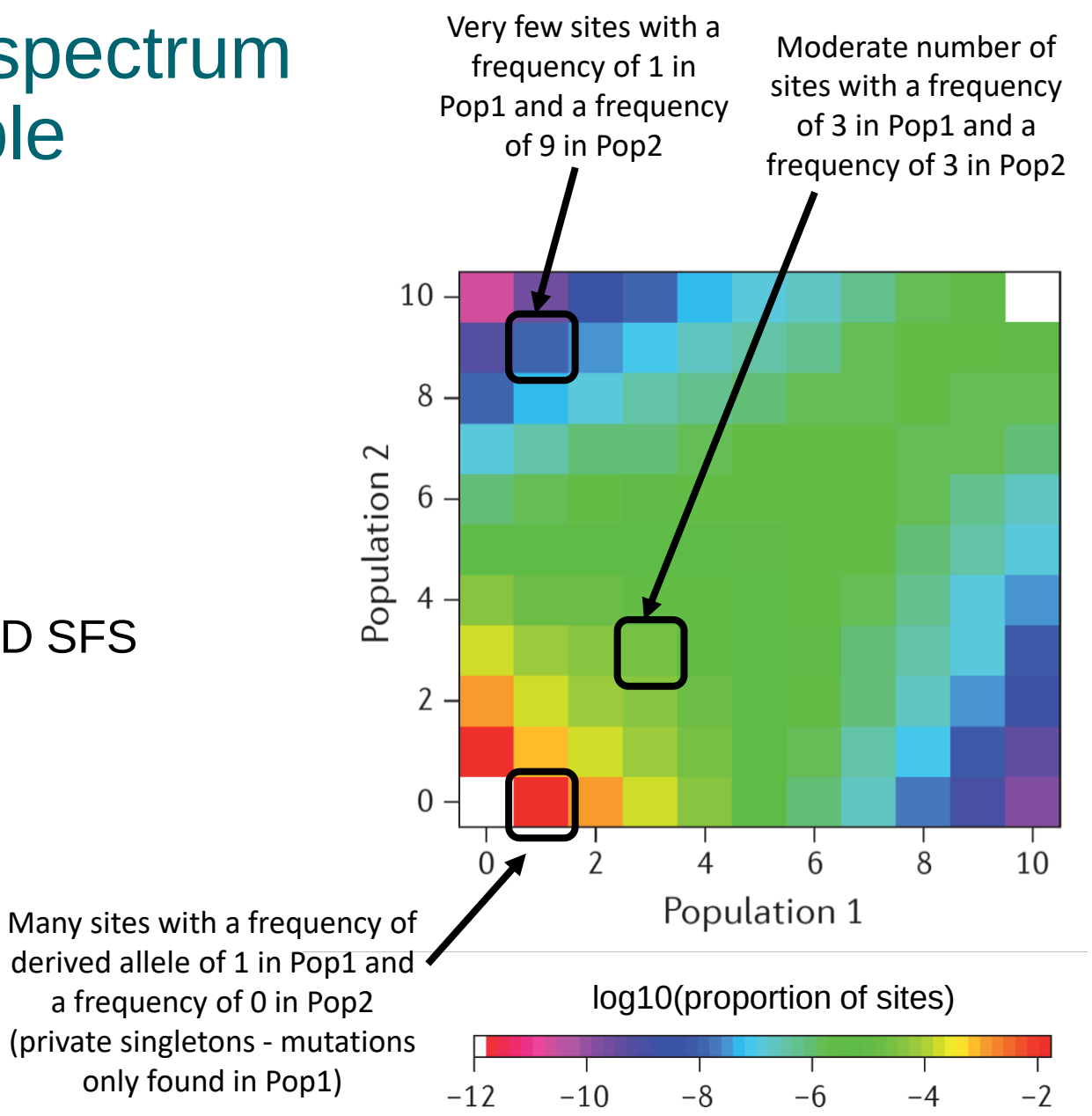
Site frequency spectrum (SFS) for multiple populations

- Single population:

  1D SFS

- Multiple populations:

  2D, 3D, …, n_{pop}D SFS



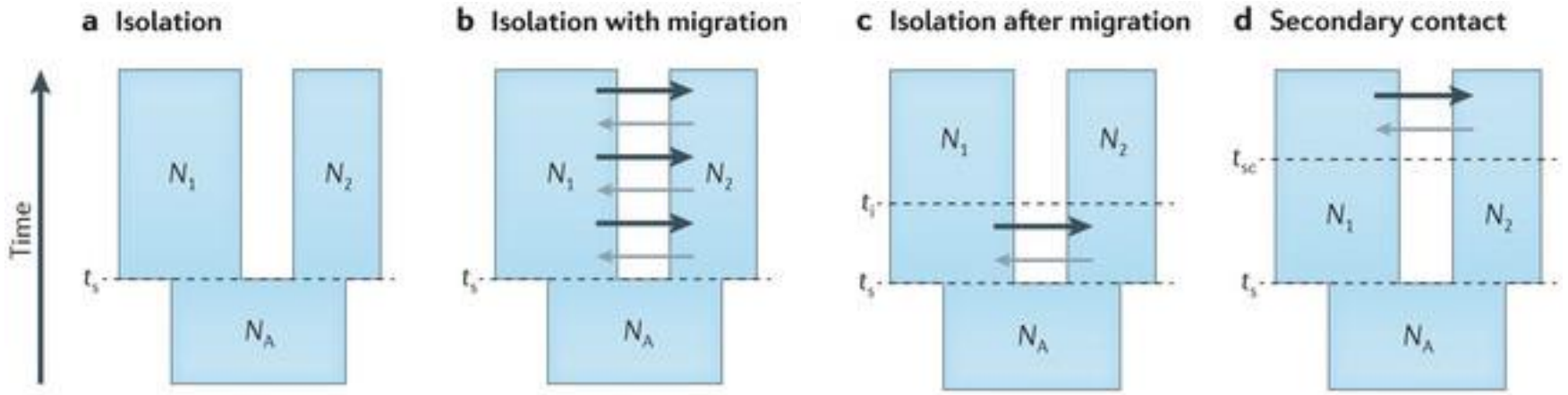
Model based inference

Sample genetic markers

What processes generated the data?

Define models to test specific hypotheses

- What is the model that best fits the data?
- What are the most likely parameters of each model?



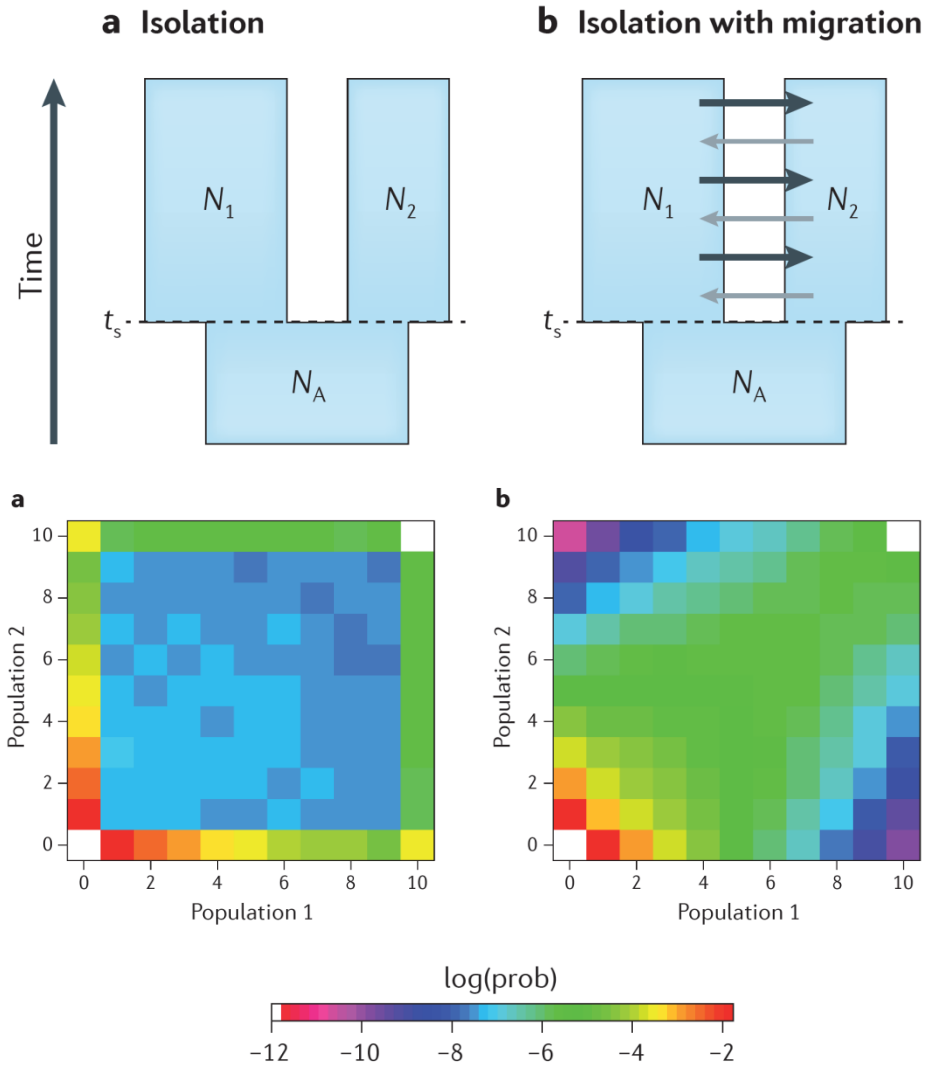
Sousa and Hey (2013) Nat. Rev. Gen.

"All models are wrong but some are useful"

George Box

Site frequency spectrum (SFS)

The SFS contains information about the demographic history of populations

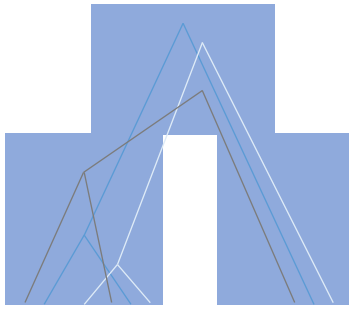


Sousa and Hey (2013) Nat. Rev. Genet.

Inferring the demographic history from the SFS

Genomic Data

Model



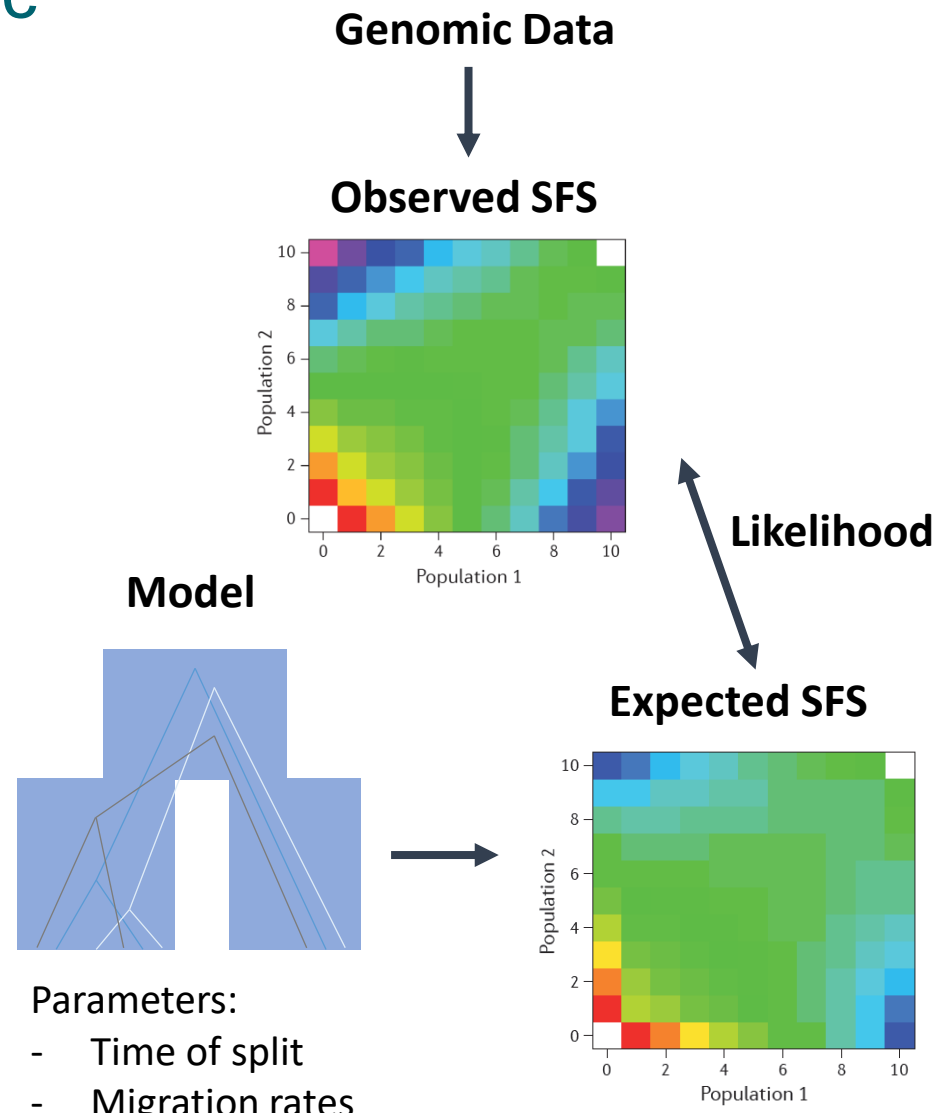
Parameters:

- Time of split
- Migration rates
- Effective sizes

Inferring the demographic history from the SFS

- The likelihood is easily computed based on the expected SFS under a given model

- There are different ways to obtain the expected SFS
  - Diffusion (forward in time)
  - Coalescent (backward in time)



Excoffier et al. (2013) *PloS Genetics*

Composite likelihood

Even though we can have linked sites, we assume that all sites are independent. Given ***S*** polymorphic sites (SNPs) out of ***L*** sites (Adams and Hudson, 2004) the composite likelihood is:

$$CL = \Pr(X \mid \theta) \propto P_0^{L-S} (1-P_0)^S \prod_{i=1}^{n-1} \hat{p}_i^{\,m_i}$$

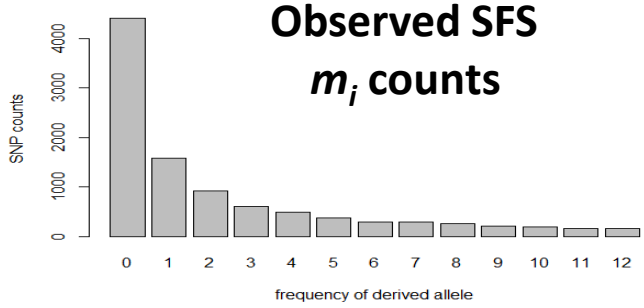
probability of no mutation on the tree

probability of at least one mutation in the tree

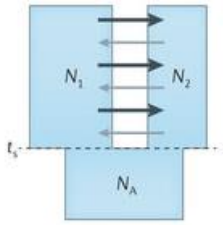
The 3 ingredientes for likelihood

Observed SFS

***m_i* counts**

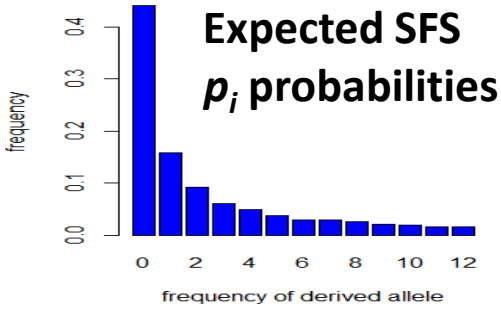


Model



Expected SFS

***p_i* probabilities**



Composite likehood

These probabilities depend:

- Number of monomorphic sites
- A fixed and mutation rate

Excoffier et al. (2013) *PloS Genetics*

Expected SFS under a given model using coalescent

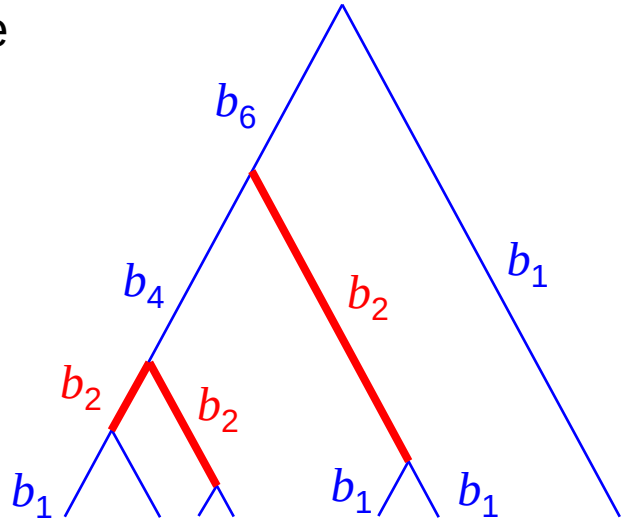
The probability of a SFS entry *i* can be estimated under a specific model θ from its expected coalescent tree as (Nielsen 2000)

$$p_i = \frac{E(t_i \mid \theta)}{E(T \mid \theta)}$$

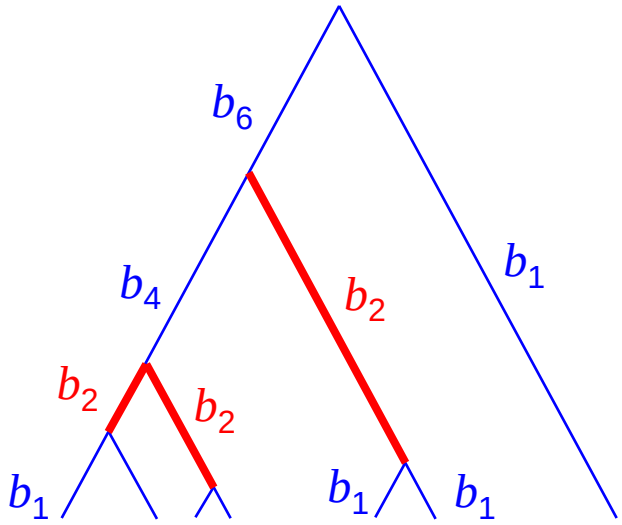
Where t_i is the total length of all branches directly leading to i terminal nodes, and T is the total tree length.

It gives the relative probability that if a mutation occurs on one of these b_i branches, it will be observed *i* times in the sample

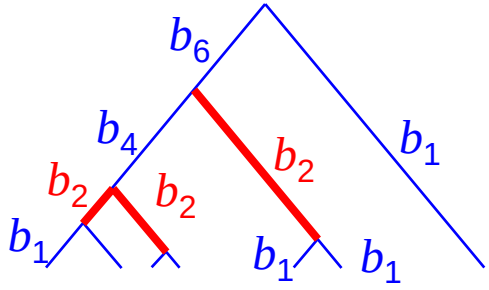
This is true under the limit of low mutation rate. No more than 1 mutation per site, back mutations not allowed!



Everything is relative if we do not know the mutation rate and number of monomorphic sites



T_L = total branch length



Frequency	0	1	2	3	4	5	6	7
SNP probability p_i	0	Sum(b_1)/ T_L	Sum(b_2)/ T_L	Sum(b_3)/ T_L	Sum(b_4)/ T_L	Sum(b_5)/ T_L	Sum(b_6)/ T_L	0

- The same expected SFS can be obtained in a large or small tree
- We need a mutation rate and the number of monomorphic sites to distinguish among the two!

Many methods based on the SFS

Different ways to obtain the expected SFS $\boldsymbol{p}_i$ under different demographic models

- Coalescent-based
  - Multiple populations

    Fastsimcoal2 (Excoffier et al 2013 PLoS Genetics)

    Momi (Kamm et al 2015) and Momi 2 (Kamm et al 2021)

    Rarecoal (Schiffels et al 2016 Nat Genetics)
  - Single population

    Stairway plot (Liu and Fu, 2015 Nat Genetics)
- Diffusion-based

  Dadi (Gutenkunst et al 2009 PLoS Genetics)

  Multipop (Lukic and Hey 2012 Genetics)

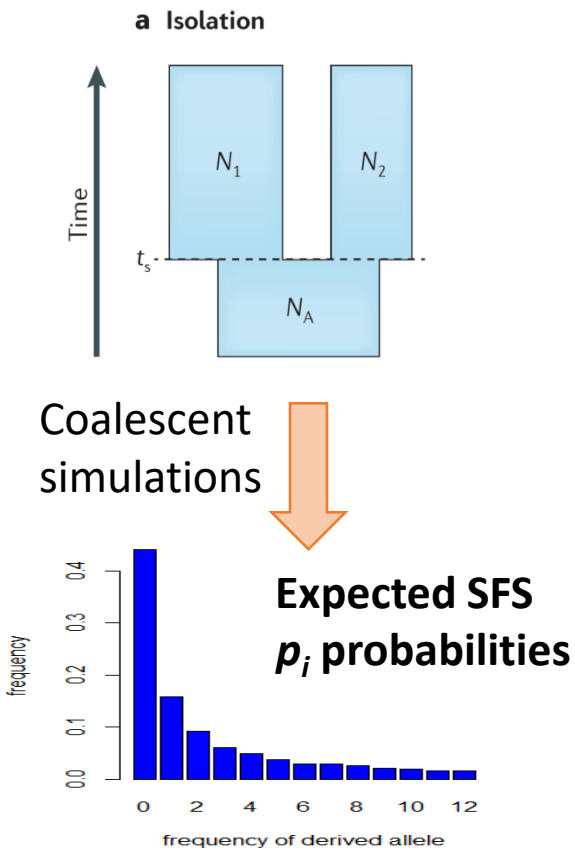
  Moments (Jouganous et al 2017 Genetics)

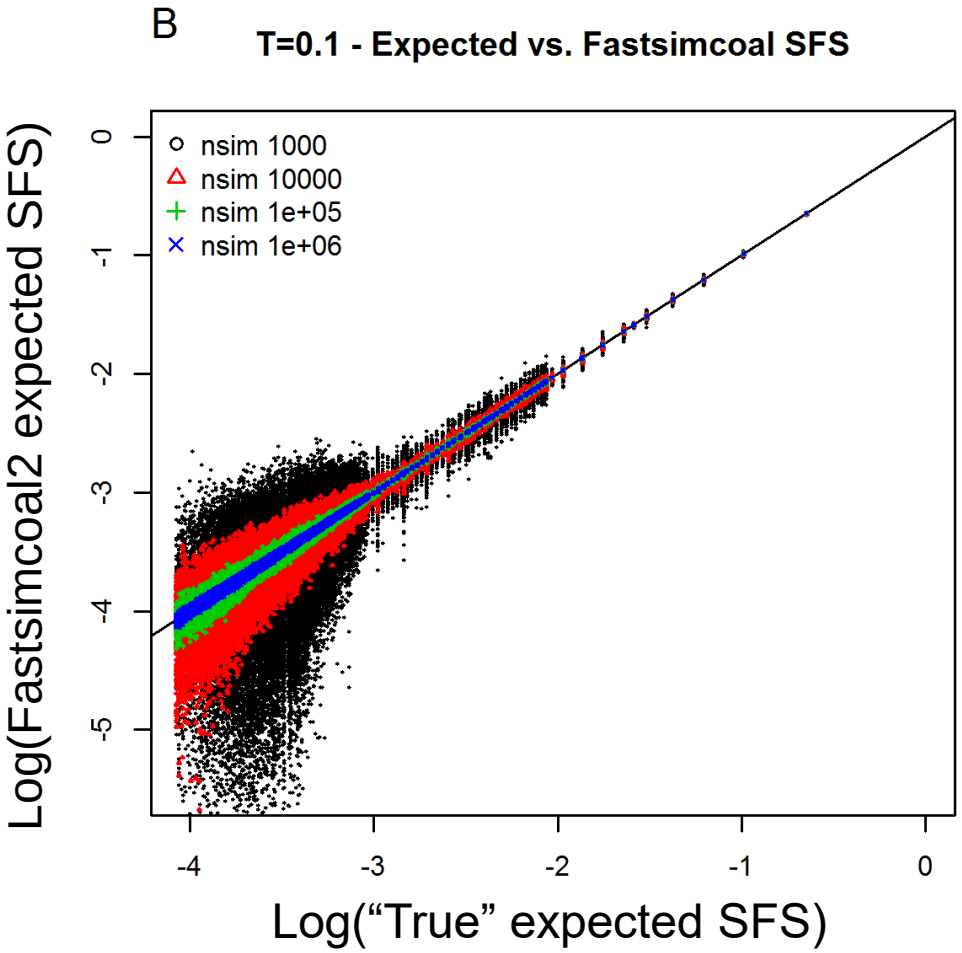
Inferring demographic history with fastsimcoal2 based on the SFS

- Fastsimcoal2 can estimate parameters from the SFS using coalescent simulations
- Maximum (composite) likelihood method
- Uses a conditional expectation (CEM) maximization algorithm to find parameter combinations that maximize the likelihood
- **It approximate the expected SFS** by performing coalescent simulations (>100,000)

Fastsimcoal2 principle: approximate the expected SFS with coalescent simulations

Use at least 100,000 coalescent simulations



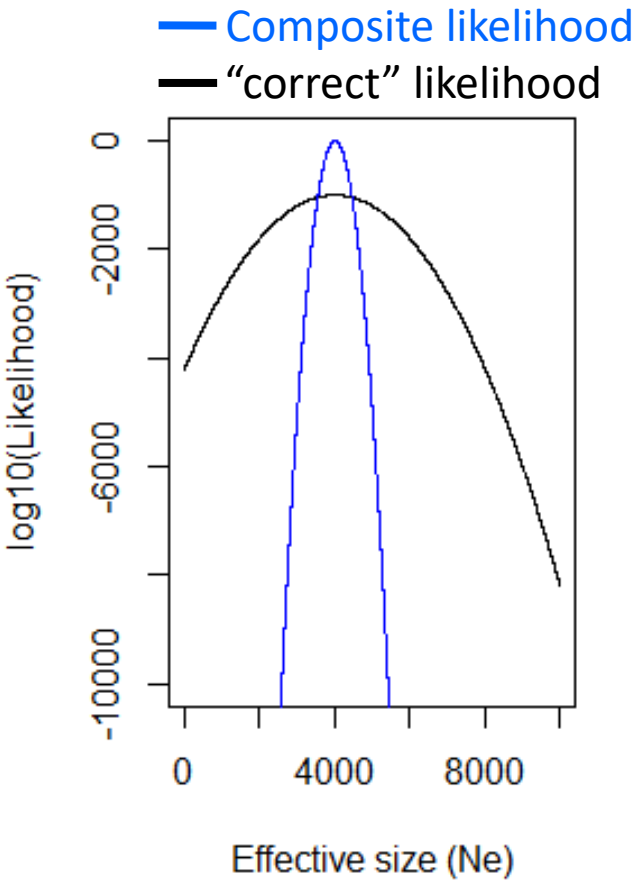


Excoffier et al. (2013) PloS Genetics; Chen (2012) TPB

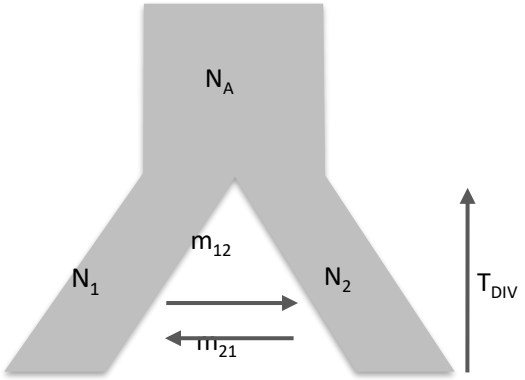
Properties of composite likelihoods

This composite likelihood (CL) is not a proper likelihood due to the non-independence of allele frequencies at linked sites.

- CL is maximized for the same parameters as full likelihood
- Can be used for parameter estimation
- Confidence intervals cannot be estimated from likelihood profile, need to bootstrap
- CL surface might be more complex than likelihood surface, and thus more difficult to explore and get the global maximum
- CL ignores information on linkage disequilibrium (recombination) between sites

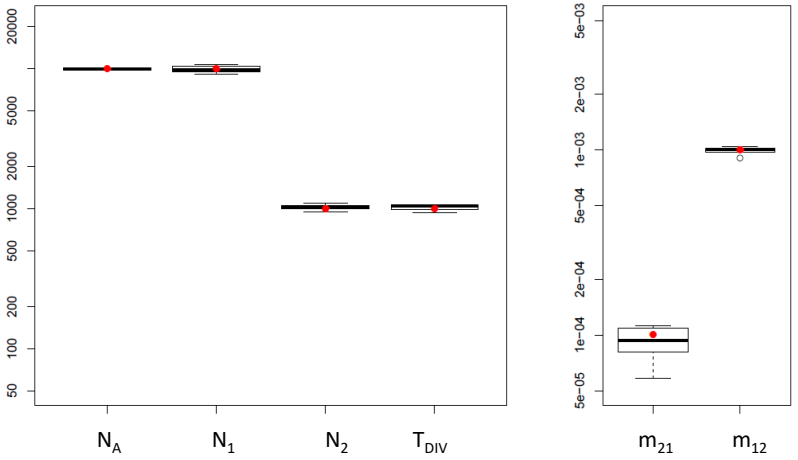


Comparisons of *fastsimcoal2* with *dadi*

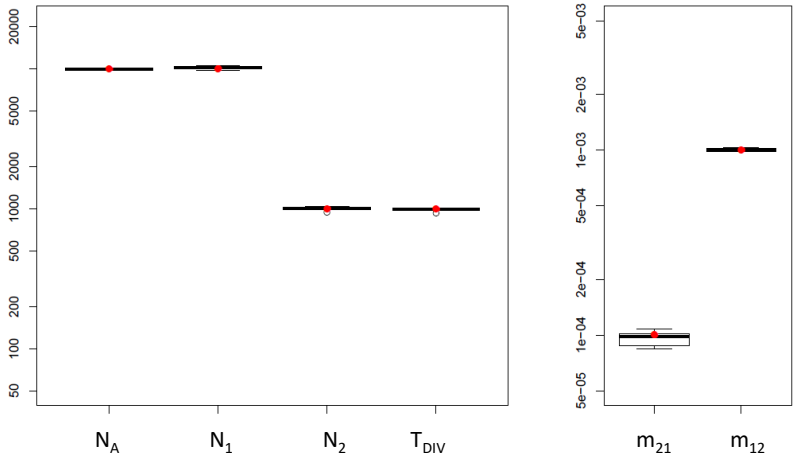


Simulation of 20 Mb data

fastsimcoal2



$\partial a \partial i$



Protocol for parameter estimation with fastsimcoal2 using the SFS

1. **Get the observed SFS**:
   - derived SFS (DAF or unfolded SFS), when the ancestral state is known;
   - minor allele frequency SFS (MAF or folded SFS) when the ancestral state is unknown
2. Define the **demographic model**
3. **Estimate the parameters** – repeat 50-100 runs, and selecting the run with maximum likelihood
4. **Bootstrap** to obtain confidence intervals for each parameter – bootstrap 10-100 datasets, by repeating a few runs for each dataset
   - For datasets with linked sites use block-bootstrap, diving the genome into blocks

Potential problems

- Maximization of the CL is not trivial (precision of the approximation and convergence problems)
- Need to repeat estimations to find maximum CL
- Needs genomic data (several Mb), difficult to have gene-specific estimates
- Next-generation sequencing data must have high coverage (>10x) to correctly estimate SFS

Limitations of estimating demographic parameters from SFS

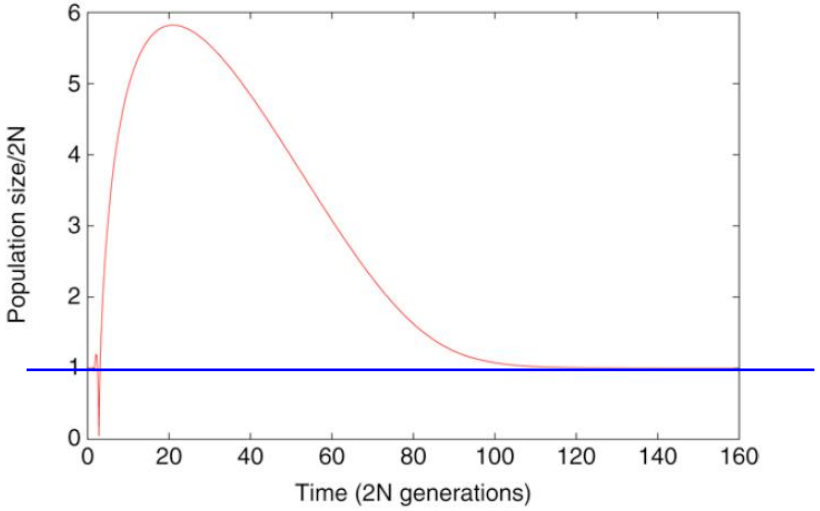
Can one learn history from the allelic spectrum?

Simon Myers[a], Charles Fefferman[b], Nick Patterson[a,*]

[a] *Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge MA 02142, United States*
[b] *Deptartment of Mathematics, Fine Hall, Washington Road, Princeton, NJ 08544, United States*

Received 17 March 2007
Available online 30 January 2008

A demographic history with the same spectrum as a constant size population



ELSEVIER

Theoretical Population Biology
Volume 120, March 2018, Pages 42-51

On the decidability of population size histories from finite allele frequency spectra

Soheil Baharian, Simon Gravel

Geometry of the Sample Frequency Spectrum and the Perils of Demographic Inference

Zvi Rosen, Anand Bhaskar, Sebastien Roch and Yun S. Song

GENETICS *October 1, 2018 vol. 210 no. 2 665-682;*
https://doi.org/10.1534/genetics.118.300733

Fundamental limits on the accuracy of demographic inference based on the sample frequency spectrum

Jonathan Terhorst and Yun S. Song

PNAS June 23, 2015 112 (25) 7677-7682; first published June 8, 2015 https://doi.org/10.1073/pnas.1503717112

Demographic inference based on Site frequency spectrum (SFS) – Part II

Vitor C. Sousa

CE3C – center for ecology, evolution and environmental changes
Department of Animal Biology
Faculdade de Ciências da Universidade de Lisboa

2022 WSPG Cesky Krumlov
09 Jun 2022

vmsousa@fc.ul.pt

Outline part II

Example of Applications:

- Human dispersal out of Africa (high quality whole-genome) – lessons on model comparison with linked SNPs
- Human colonization of Siberia and America (ancient whole-genome data) - lessons on dealing with sequencing errors
- Deer mice colonization of Nebraska Sand Hills (targeted re-capture data) – lessons on effects of filtering
- Divergence times and gene flow in sawflies (ddRAD-seq data) – lessons from model comparison with ddRAD
- Hybridization in freshwater fish (GBS data) - lessons from inferring relative parameters with limited data

Nourlangie, Kakadu National Park, NT, Australia

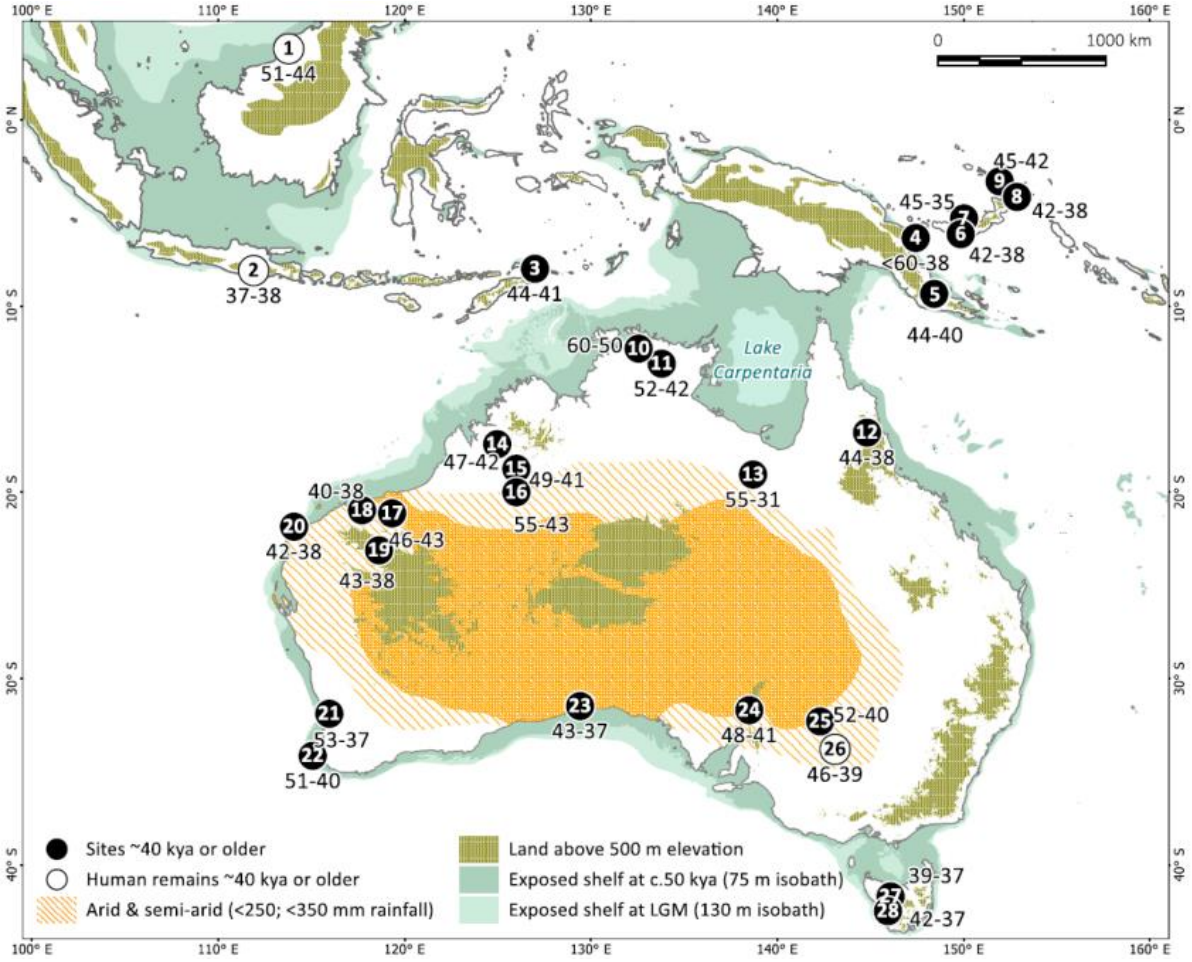
A genomic history of Aboriginal Australia

Anna-Sapfo Malaspinas[1,2,3]*, Michael C. Westaway[4]*, Craig Muller[1]*, Vitor C. Sousa[2,3]*, Oscar Lao[5,6]*, Isabel Alves[2,3,7]*, Anders Bergström[8]*, Georgios Athanasiadis[9], Jade Y. Cheng[9,10], Jacob E. Crawford[10,11], Tim H. Heupink[4], Enrico Macholdt[12], Stephan Peischl[3,13], Simon Rasmussen[14], Stephan Schiffels[15], Sankar Subramanian[4], Joanne L. Wright[4], Anders Albrechtsen[16], Chiara Barbieri[12,17], Isabelle Dupanloup[2,3], Anders Eriksson[18,19], Ashot Margaryan[1], Ida Moltke[16], Irina Pugach[12], Thorfinn S. Korneliussen[1], Ivan P. Levkivskyi[20], J. Víctor Moreno-Mayar[1], Shengyu Ni[12], Fernando Racimo[10], Martin Sikora[1], Yali Xue[8], Farhang A. Aghakhanian[21], Nicolas Brucato[22], Søren Brunak[23], Paula F. Campos[1,24], Warren Clark[25], Sturla Ellingvåg[26], Gudjugudju Fourmile[27], Pascale Gerbault[28,29], Darren Injie[30], George Koki[31], Matthew Leavesley[32], Betty Logan[33], Aubrey Lynch[34], Elizabeth A. Matisoo-Smith[35], Peter J. McAllister[36], Alexander J. Mentzer[37], Mait Metspalu[38], Andrea B. Migliano[29], Les Murgha[39], Maude E. Phipps[21], William Pomat[31], Doc Reynolds[40], Francois-Xavier Ricaut[22], Peter Siba[31], Mark G. Thomas[28], Thomas Wales[41], Colleen Ma'run Wall[42], Stephen J. Oppenheimer[43], Chris Tyler-Smith[8], Richard Durbin[8], Joe Dortch[44], Andrea Manica[18], Mikkel H. Schierup[9], Robert A. Foley[1,45], Marta Mirazón Lahr[1,45], Claire Bowern[46], Jeffrey D. Wall[47], Thomas Mailund[9], Mark Stoneking[12], Rasmus Nielsen[1,48], Manjinder S. Sandhu[8], Laurent Excoffier[2,3], David M. Lambert[4] & Eske Willerslev[1,8,18]

Nature(2016)

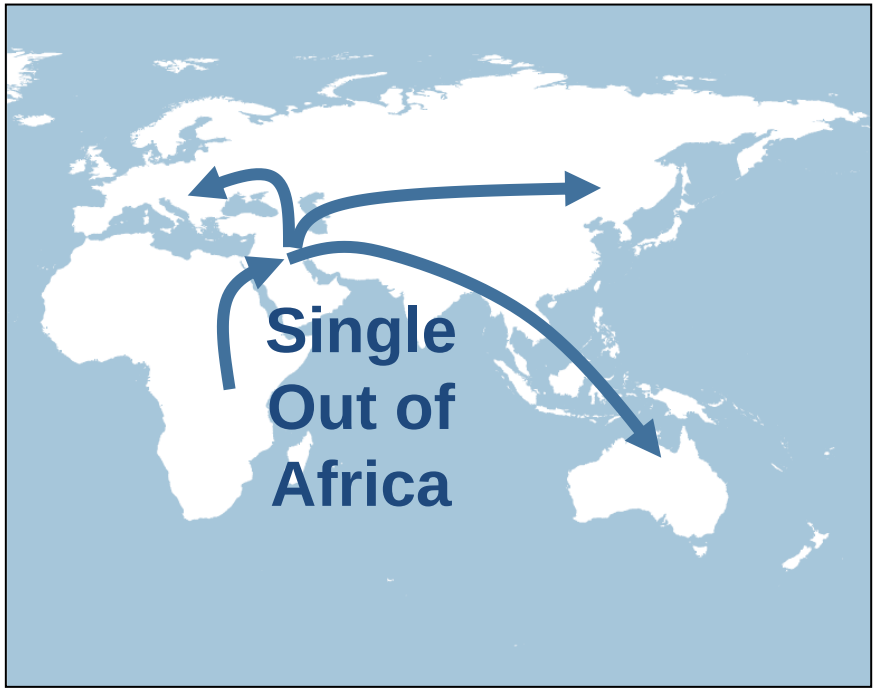
Ewaninga Rock Carvings Conservation Reserve, NT, Australia

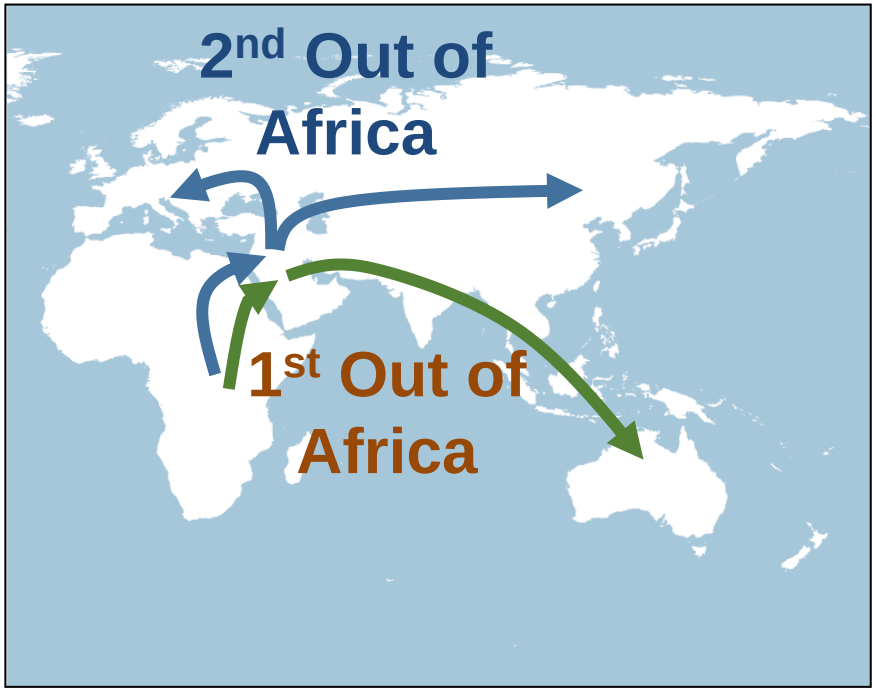
Australia harbors some of the oldest modern human remains outside Africa



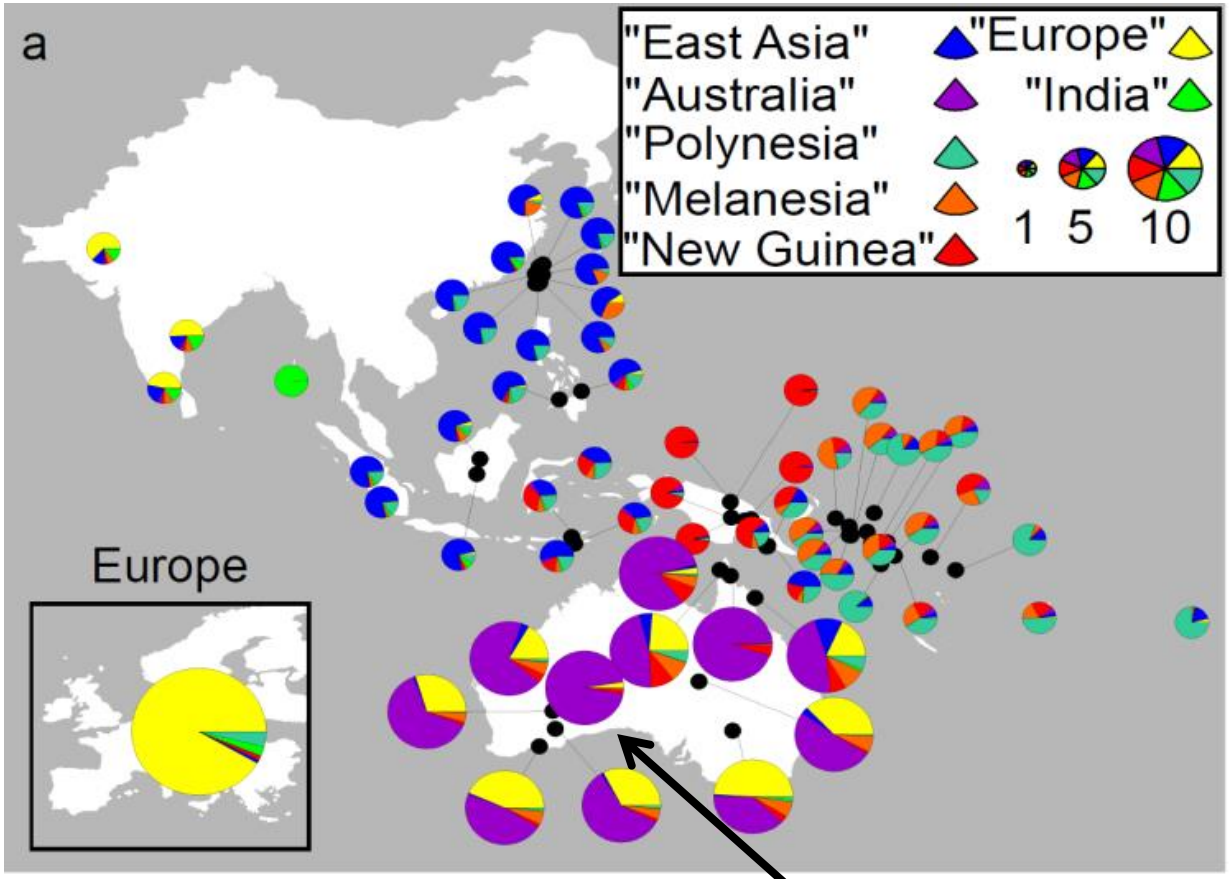
Many sites and remains dated to be older than 40 kya, suggesting a human settlement 47.5-55 kya

One wave out of Africa *vs* Two waves out of Africa





83 high-coverage Aboriginal Australians genomes

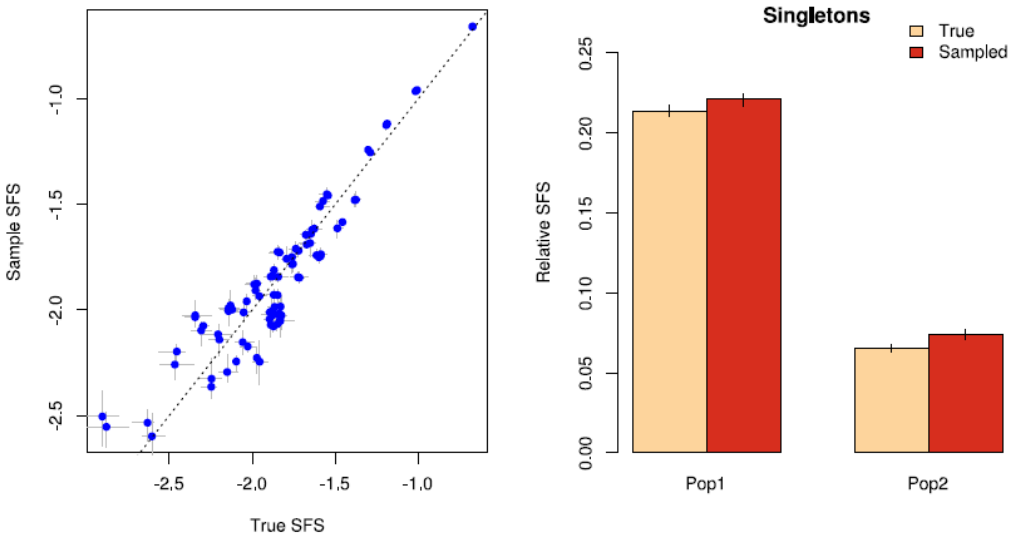


Average depth of coverage: 65x

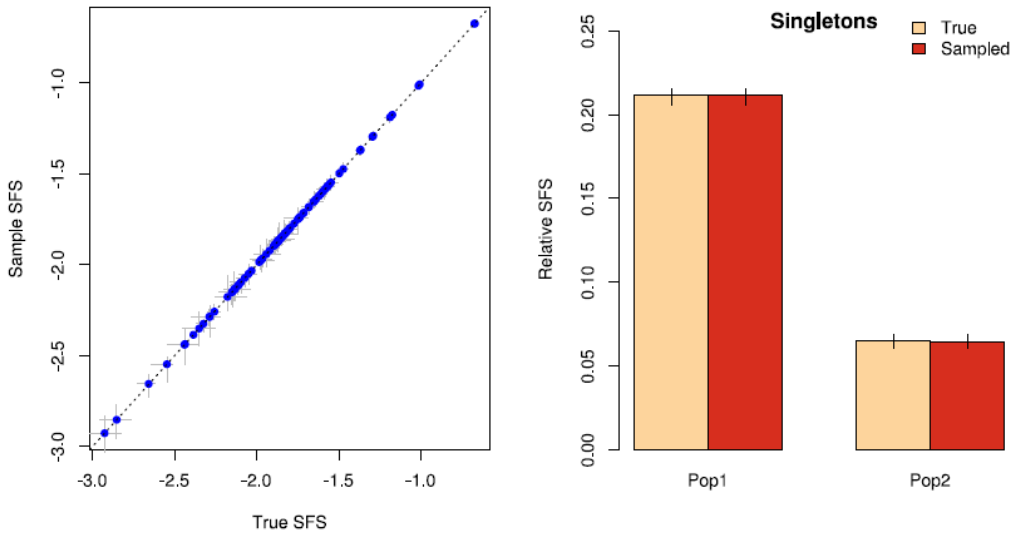
A note on recovering the SFS from genomic data

- Simulation study
- Low depth of coverage and missing data leads to biases towards rare variants

a) Low depth of coverage, no GQ filter, allowing missing data



b) Depth of coverage similar to observed data, GQ>30 filter, no missing dat



★ Archaic human genomes:
- 1 Neanderthal (~66 kya)
- 1 Denisovan (~52 kya)

Mutation rate assumed
1.25×10^{-8} /site/gen
Scally and Durbin (2012) *Nat. Rev. Genet.*

Generation time
29 years/gen
Fenner (2005) *Am. J. Phys. Anthropol.*

Since we want to infer demography we tried to minimize the number of sites affected by selection:

- 985 1Mb blocks outside genic regions and CpG islands (~4.3 Million SNPs)
- 5 dimensional SFS (16,875 entries)
- Confidence intervals obtained using block-bootstrap

Towards a model to test the hypotheses: One *vs* Two waves Out of Africa

- Data (SFS)
- (Re-)Define model (hypotheses to test)
- Run fastsimcoal2
- Estimates!
  - Assess the fit to the data

Do you have an outgroup?

- **Yes** – use the derived (unfolded) SFS
- **No** – use the minor allele frequency spectrum (folded)

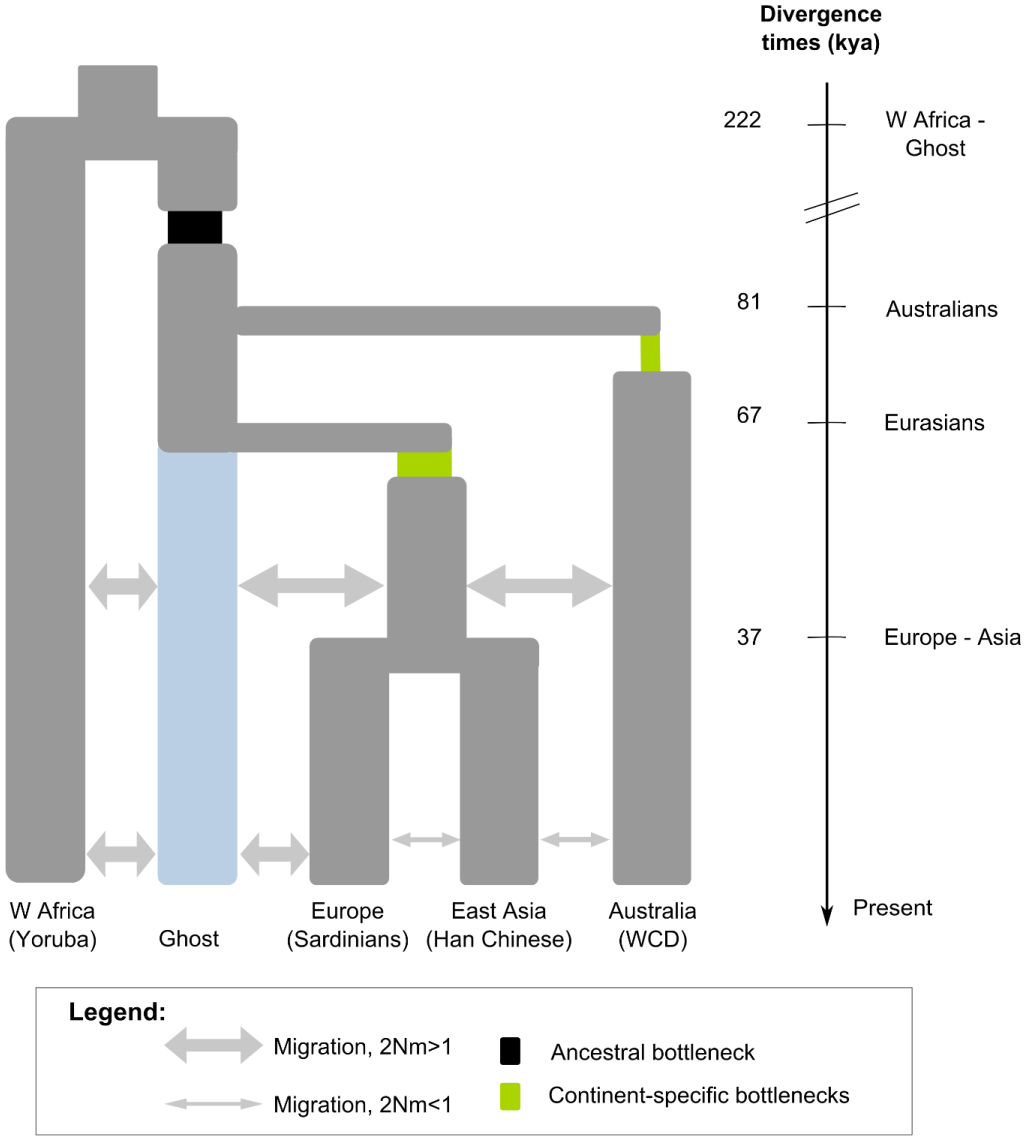
Do you have monomorphic sites?

- **Yes** - then, given a mutation rate you can infer the absolute times and effective sizes
- **No** – then all your estimates need to be relative to a fixed parameter (fixed Ne or fixed time)

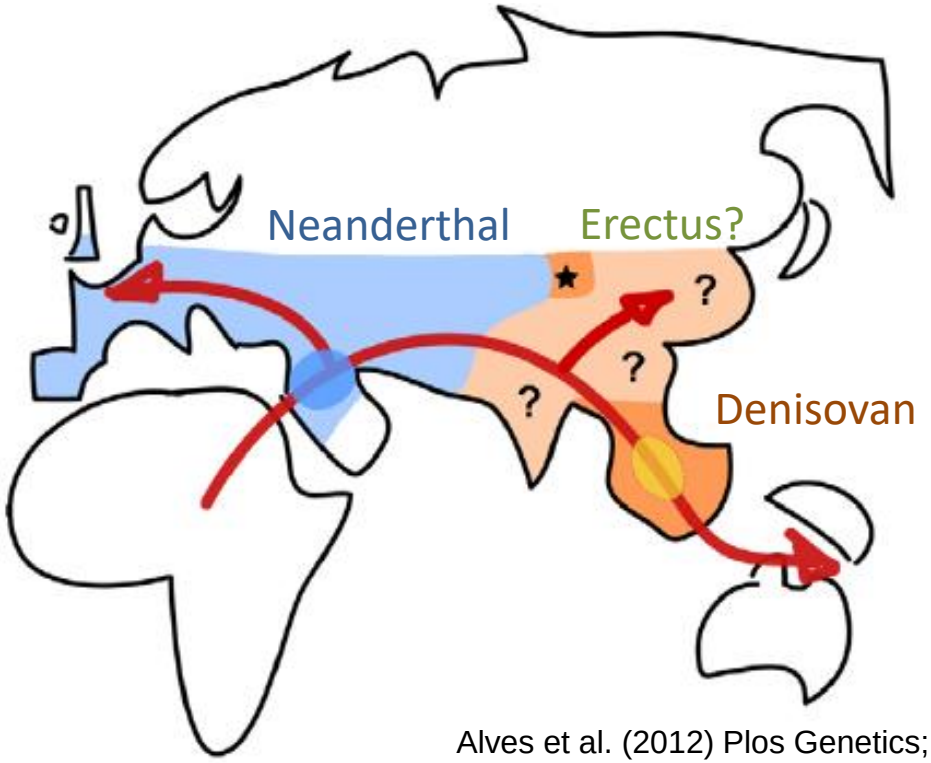
We always get results…

Evidence of two waves Out of Africa:

- Old split leading to colonization of Australia (81kya)
- More recent split leading to colonization of Eurasia (67 kya)



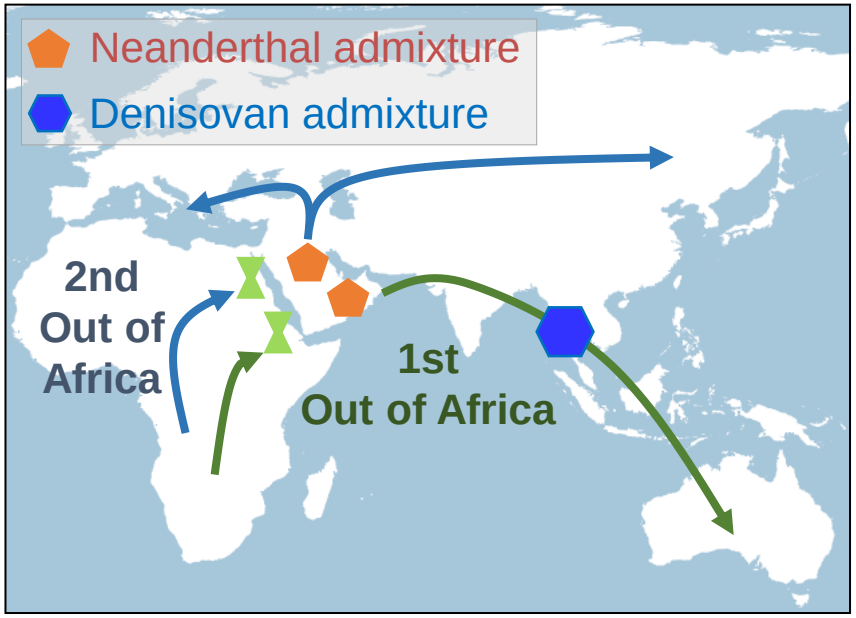
Towards a model incorporating Neanderthal and Denisovan admixture

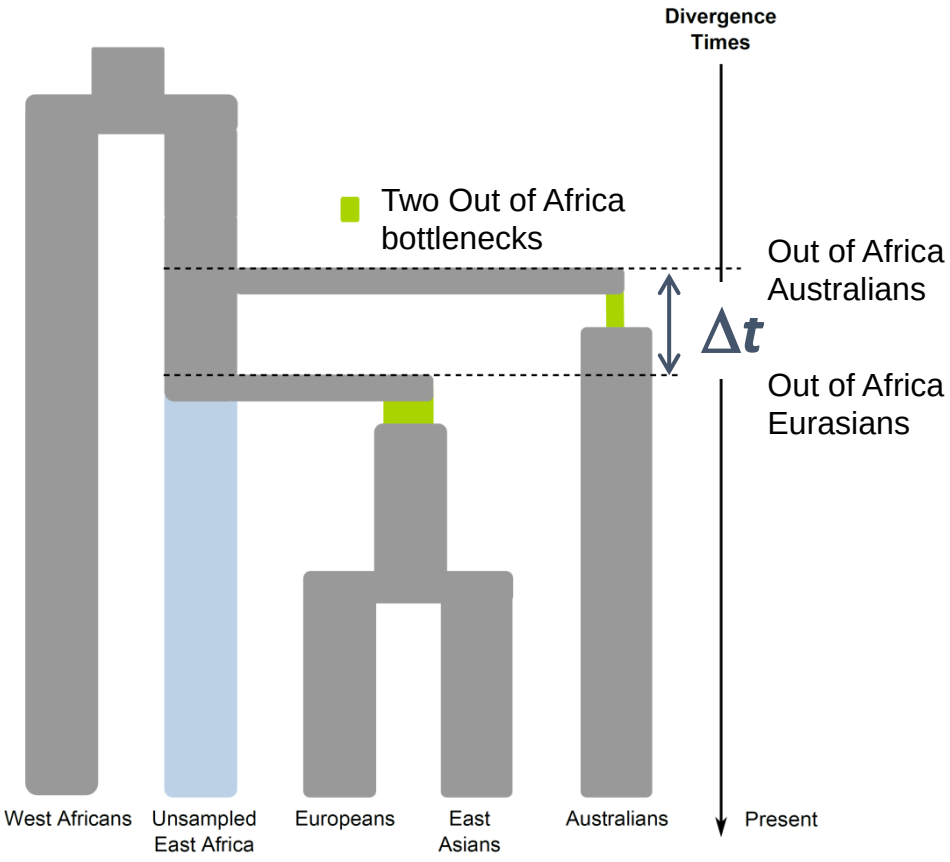


Alves et al. (2012) Plos Genetics;

- Non-African populations: 1-4% estimated Neanderthal admixture
- Aboriginal Australians and New Guineans: 3-6% estimated Denisovan admixture
- Archaic admixture can affect times of split estimates

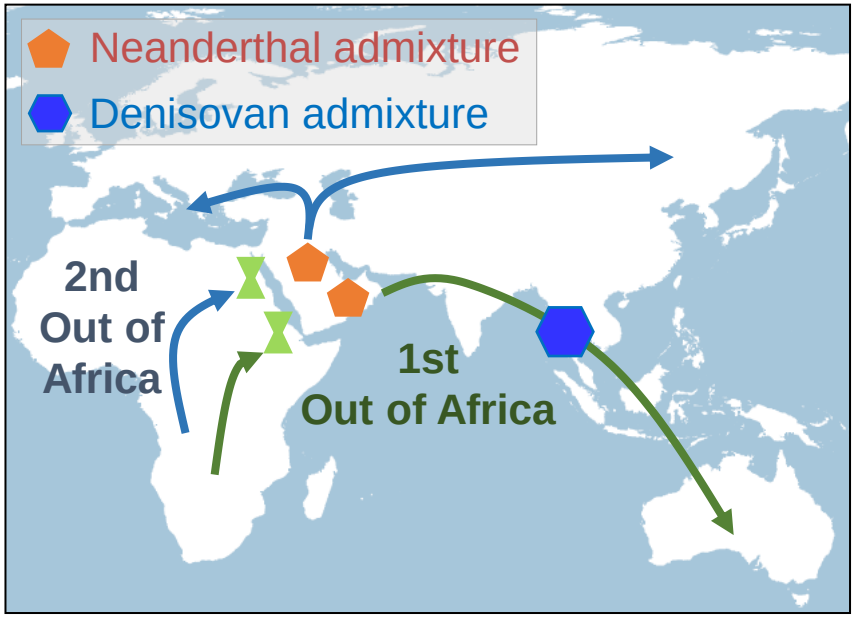
Two-waves out of Africa

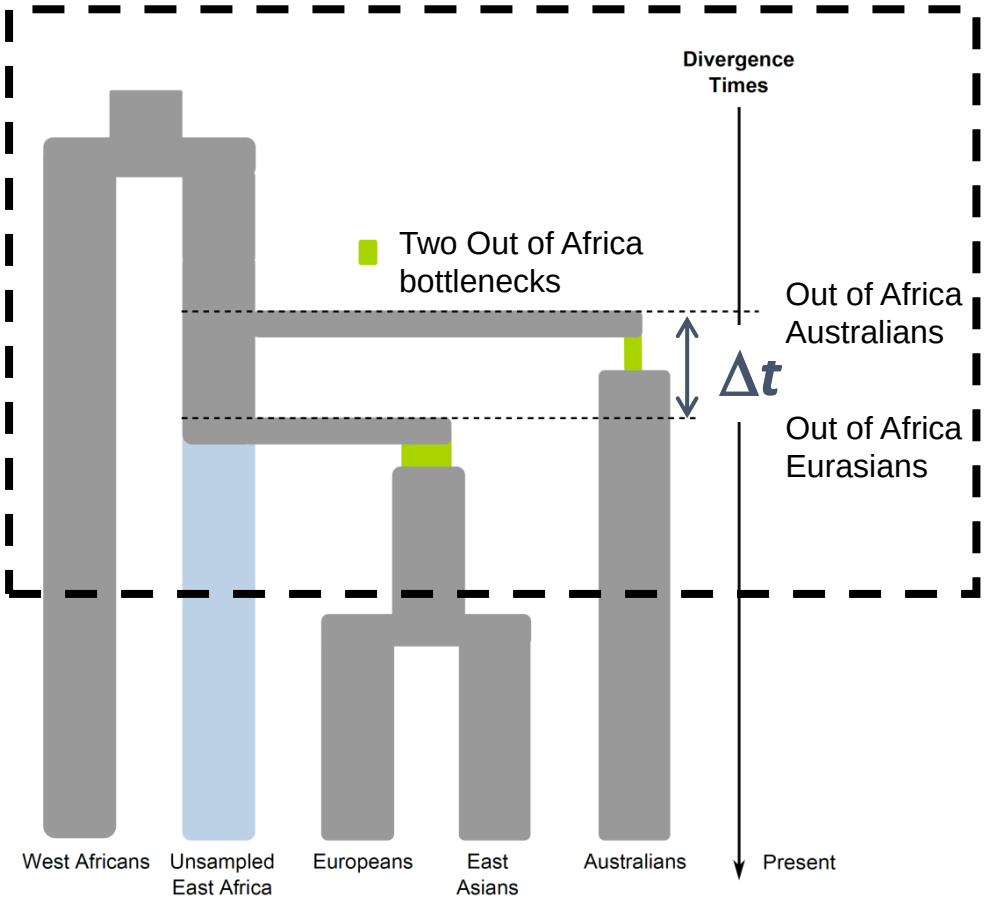




- Two different divergence times ($\Delta t >> 0$)
- Two independent bottlenecks associated with the two Out of Africa events

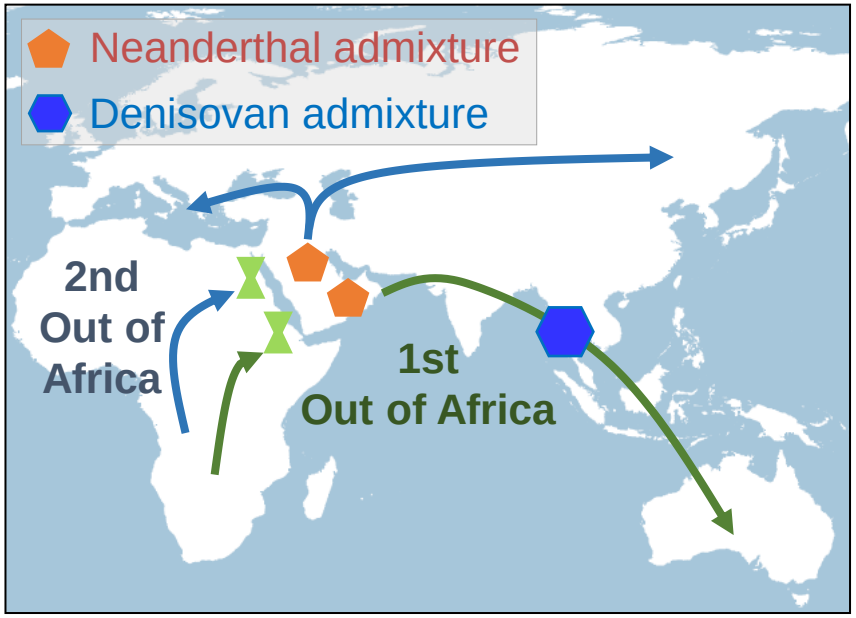
Two-waves out of Africa

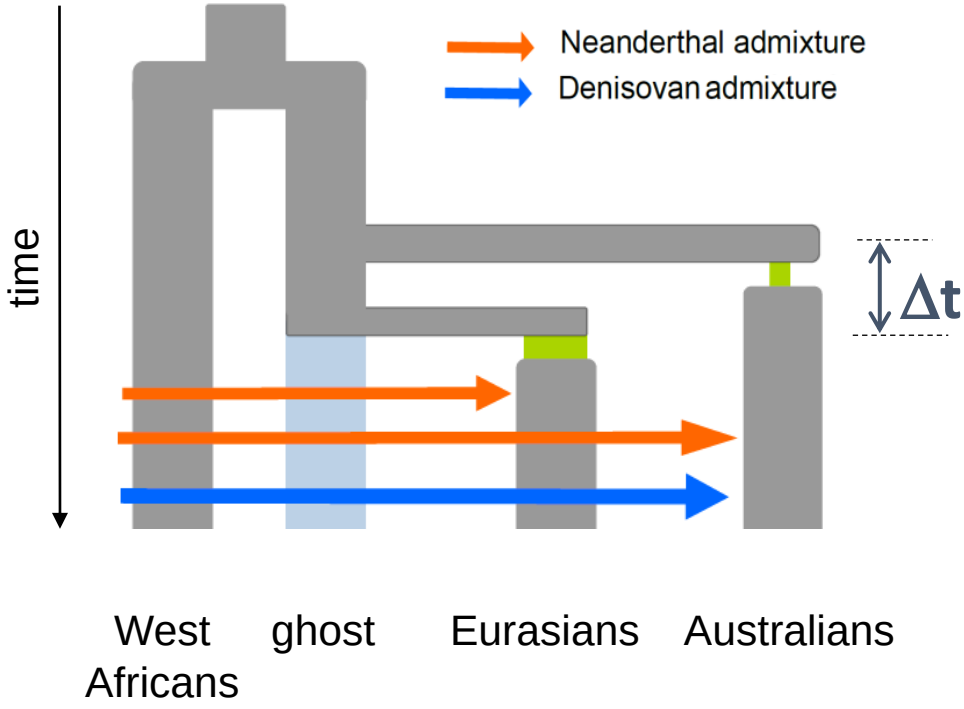




- Two different divergence times ($\Delta t >> 0$)
- Two independent bottlenecks associated with the two Out of Africa events

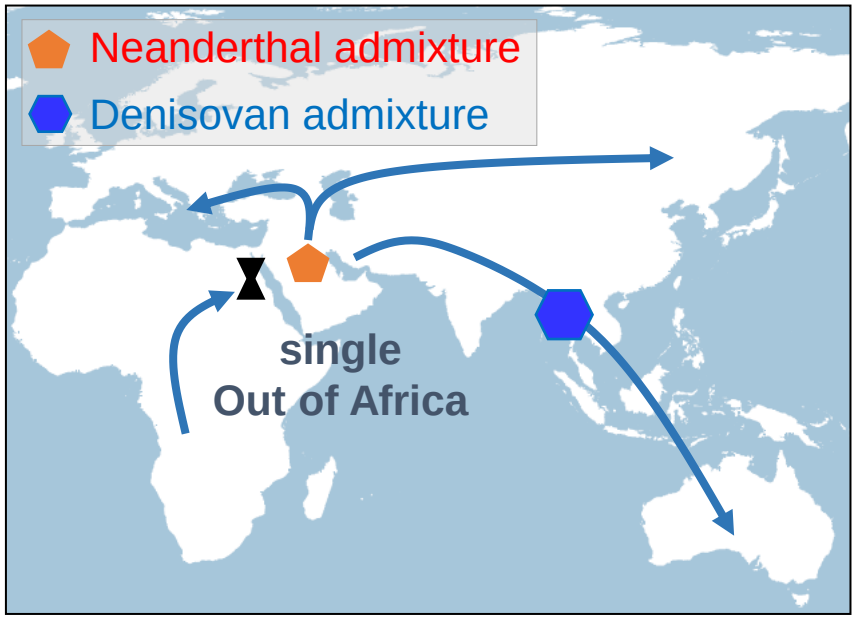
Two-waves out of Africa

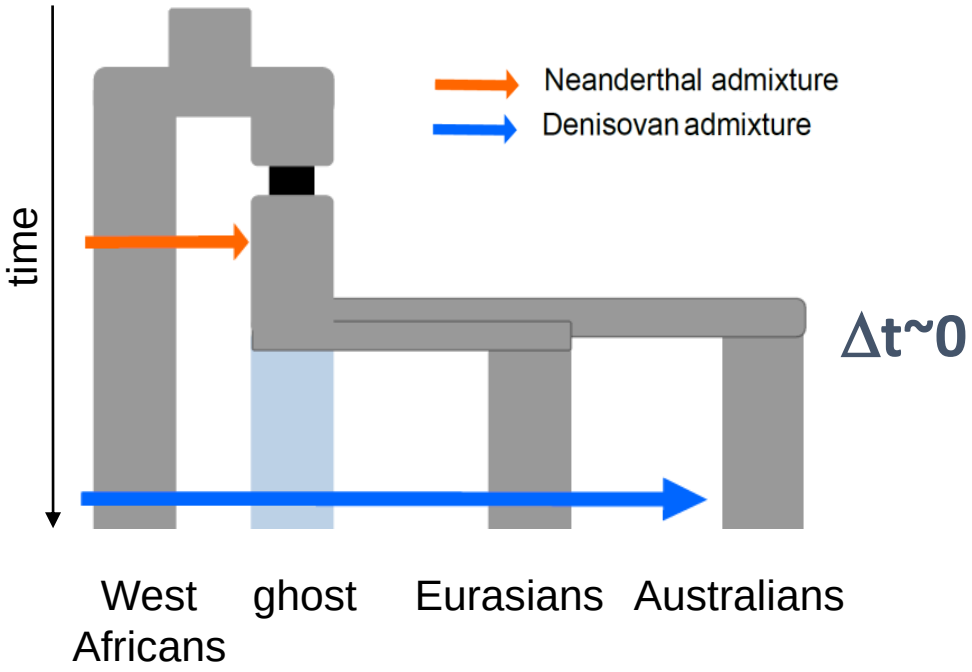




- Two different divergence times ($\Delta t >> 0$)
- Two independent bottlenecks associated with the two Out of Africa events

One wave out of Africa

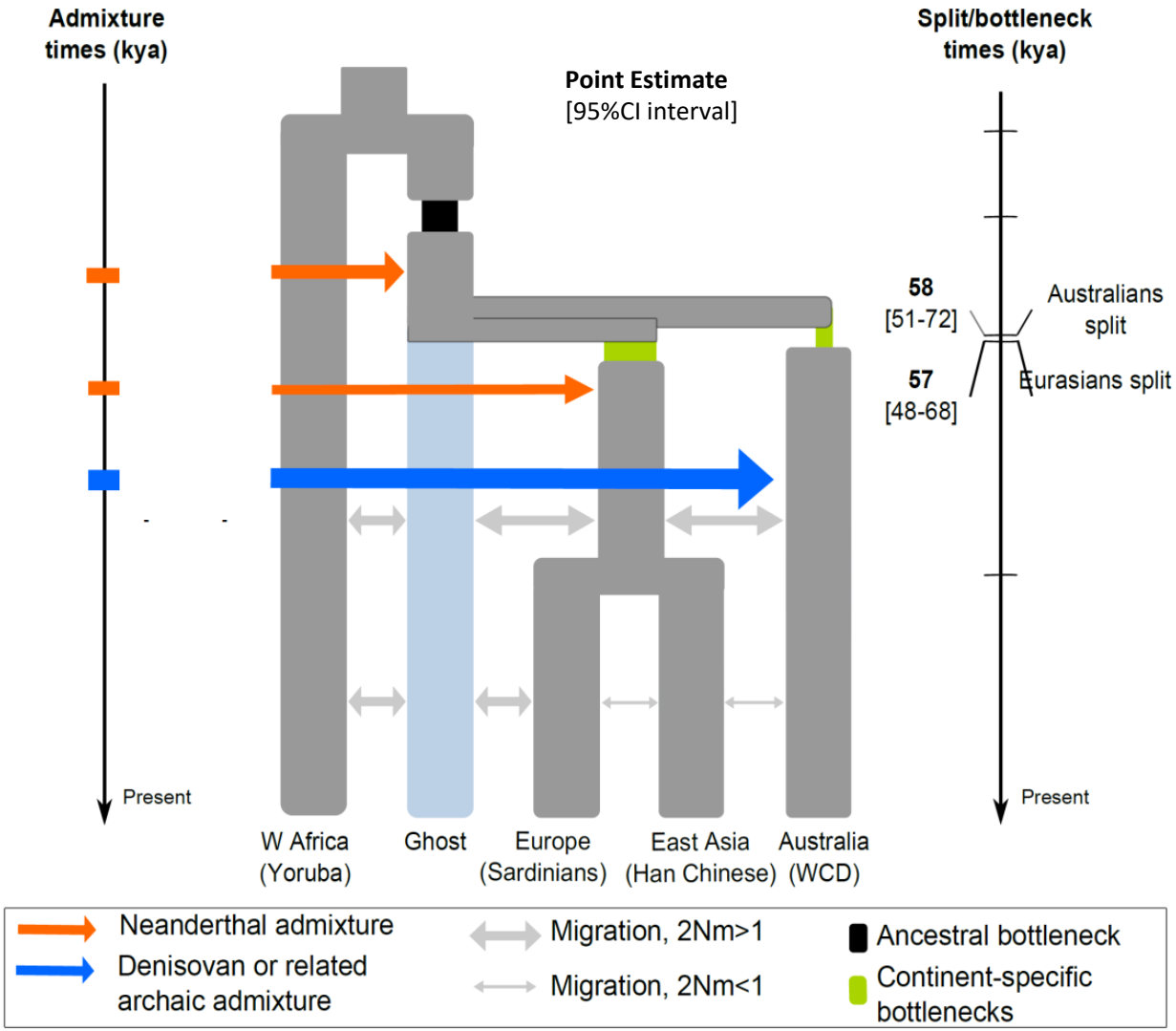




- Similar divergence times (Δt close to zero)
- One single bottlenecks associated with the Out of Africa events
- A major admixture pulse with Neanderthal

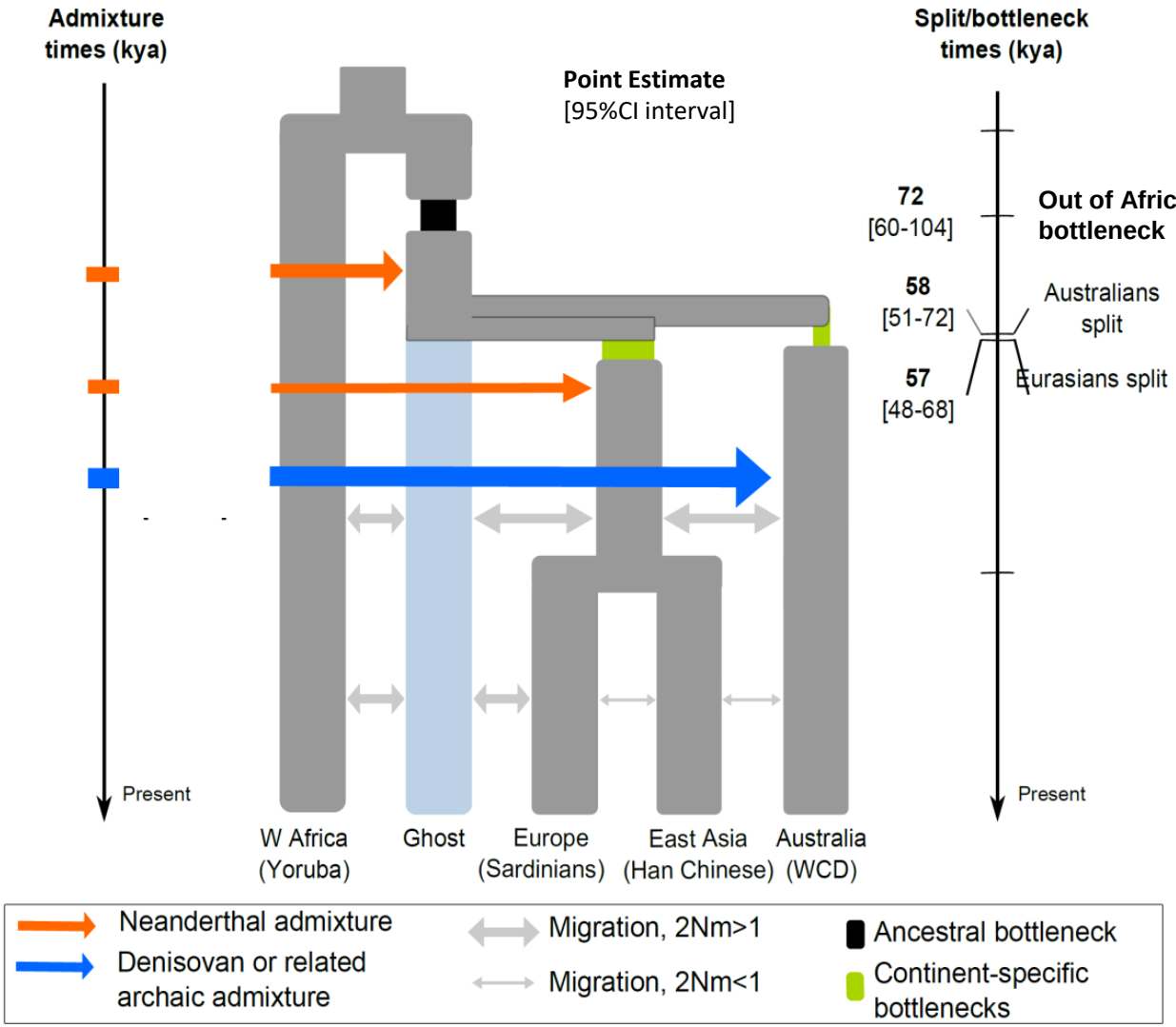
A single wave Out of Africa is consistent with our estimates when accounting for archaic admixture

- Similar divergence time (Δt close to zero)



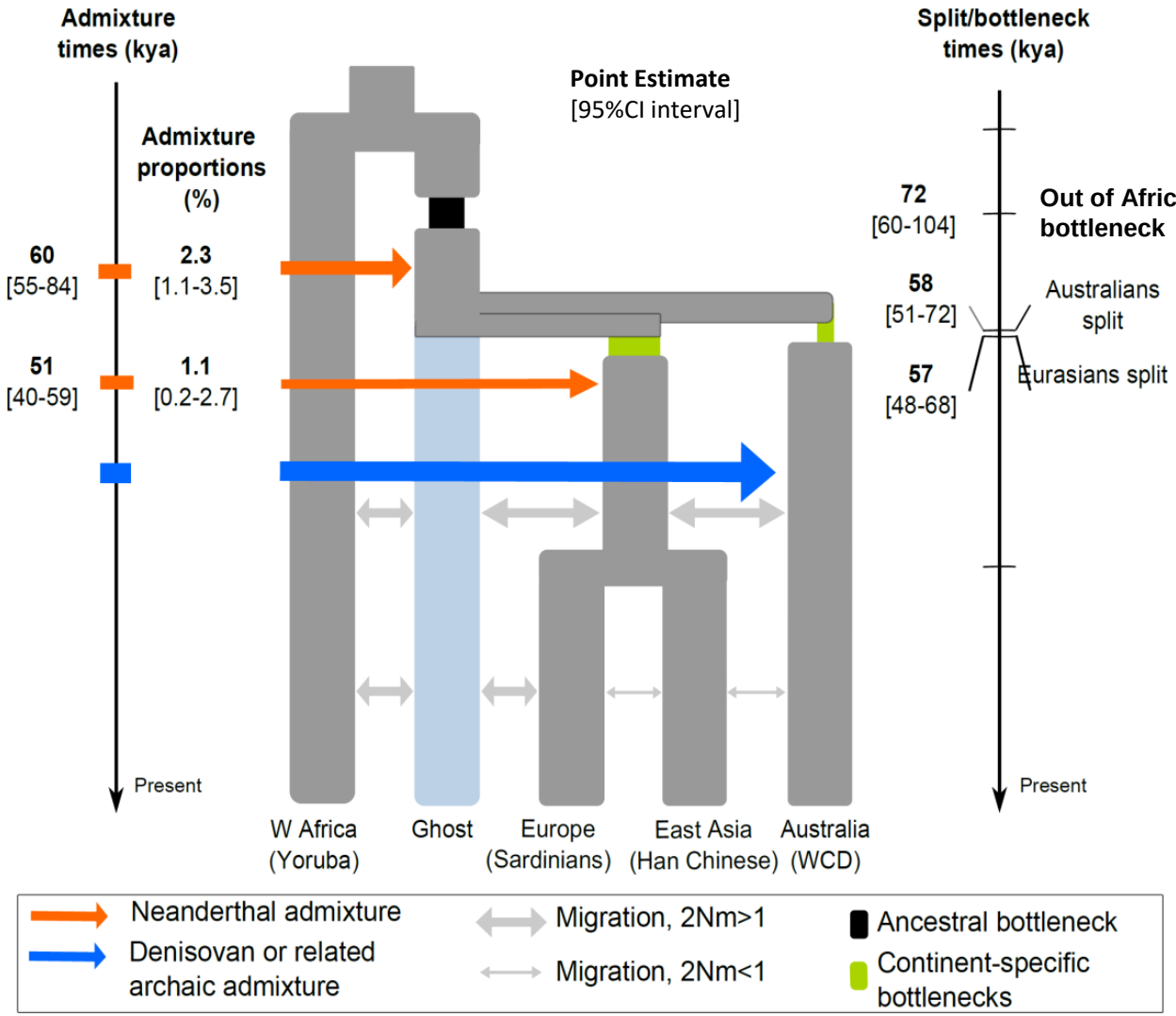
A single wave Out of Africa is consistent with our estimates when accounting for archaic admixture

- Similar divergence time (Δt close to zero)
- Bottleneck associated with the Out of Africa event



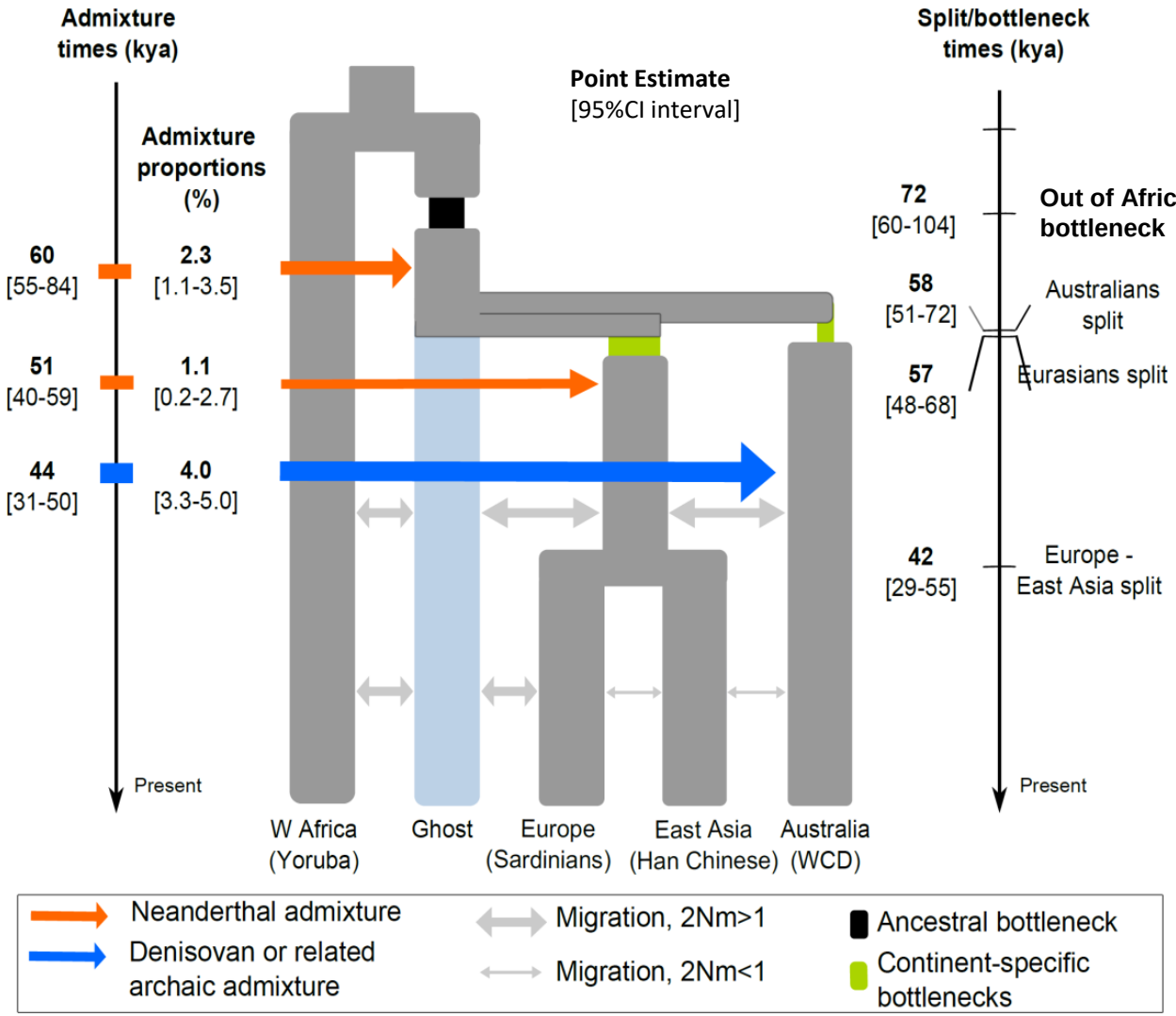
A single wave Out of Africa is consistent with our estimates when accounting for archaic admixture

- Similar divergence time (Δt close to zero)
- Bottleneck associated with the Out of Africa event
- A major admixture pulse with Neanderthal in ancestors of all non-Africans

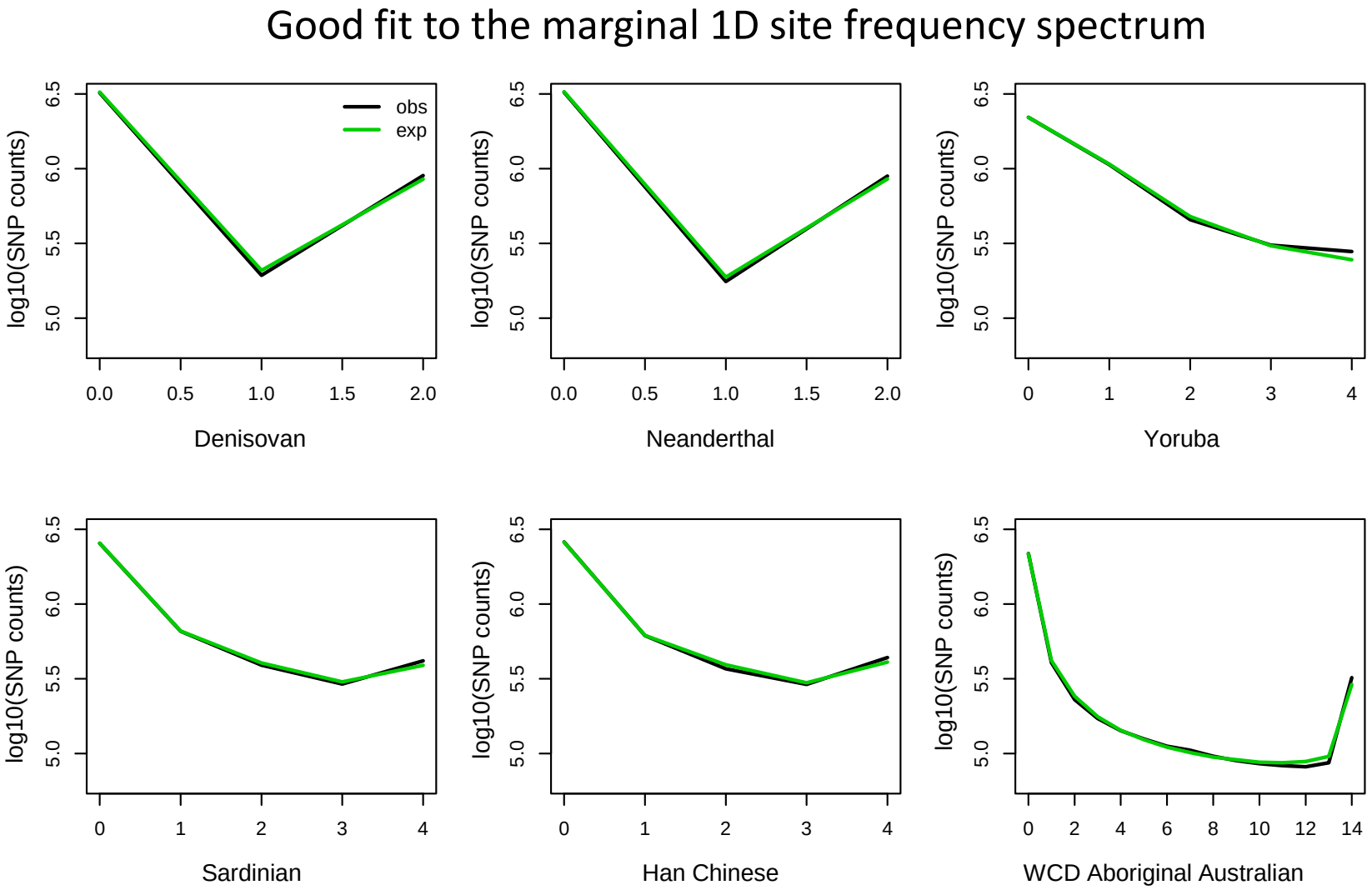


A single wave Out of Africa is consistent with our estimates when accounting for archaic admixture

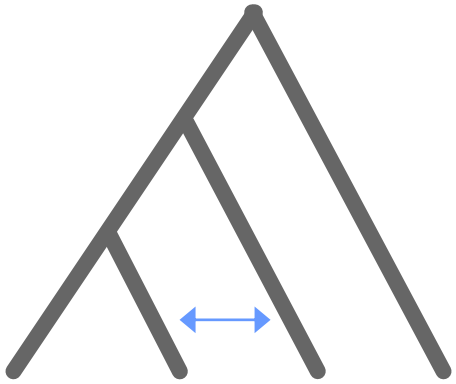
- Similar divergence time (Δt close to zero)
- Bottleneck associated with the Out of Africa event
- A major admixture pulse with Neanderthal in ancestors of all non-Africans



Model captures aspects about the observed data

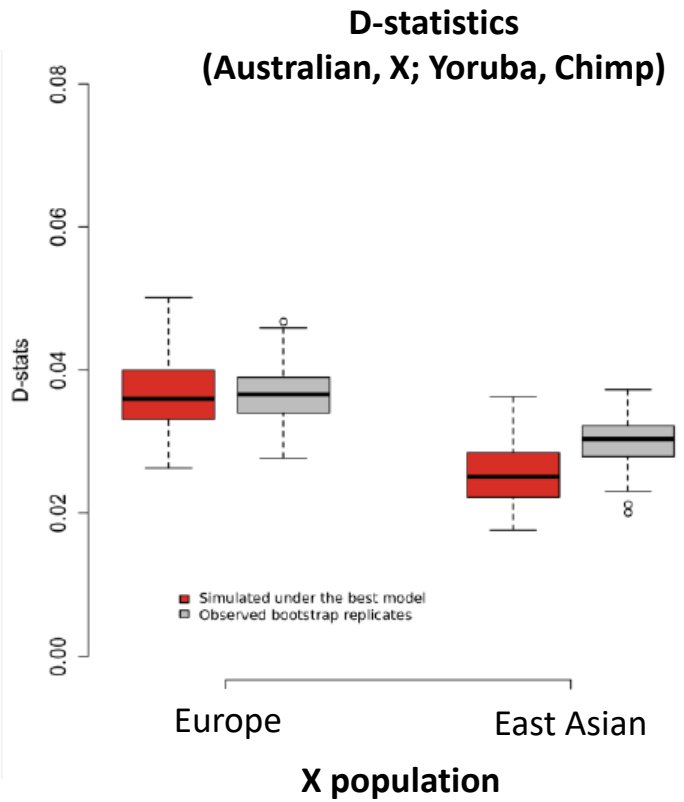


Model captures the higher derived allele sharing between Eurasians and Yoruba



Australia Europe or East Asian Yoruba Chimp

D-statistics suggest that Yoruba and Eurasians share more derived alleles than Yoruba and Australians

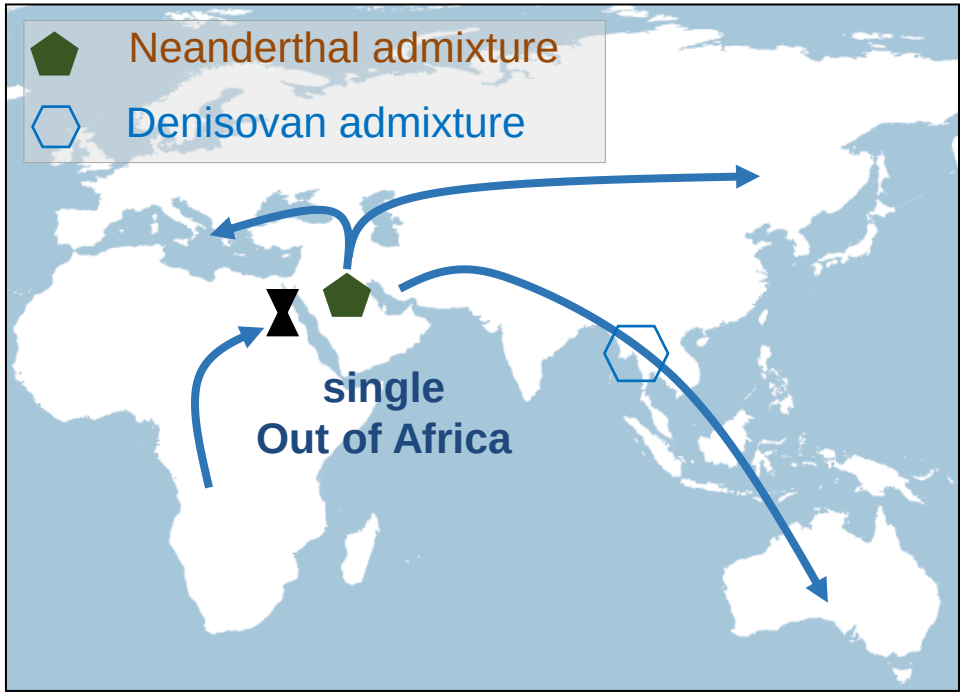


Summary

Aboriginal Australians genomes support a single major wave out of Africa

- Accounting for archaic admixture with Neanderthal and Denisovan was crucial to understand population divergence

- Genomic data consistent with a single major dispersal event out of Africa (60-104 kya)



ARTICLE

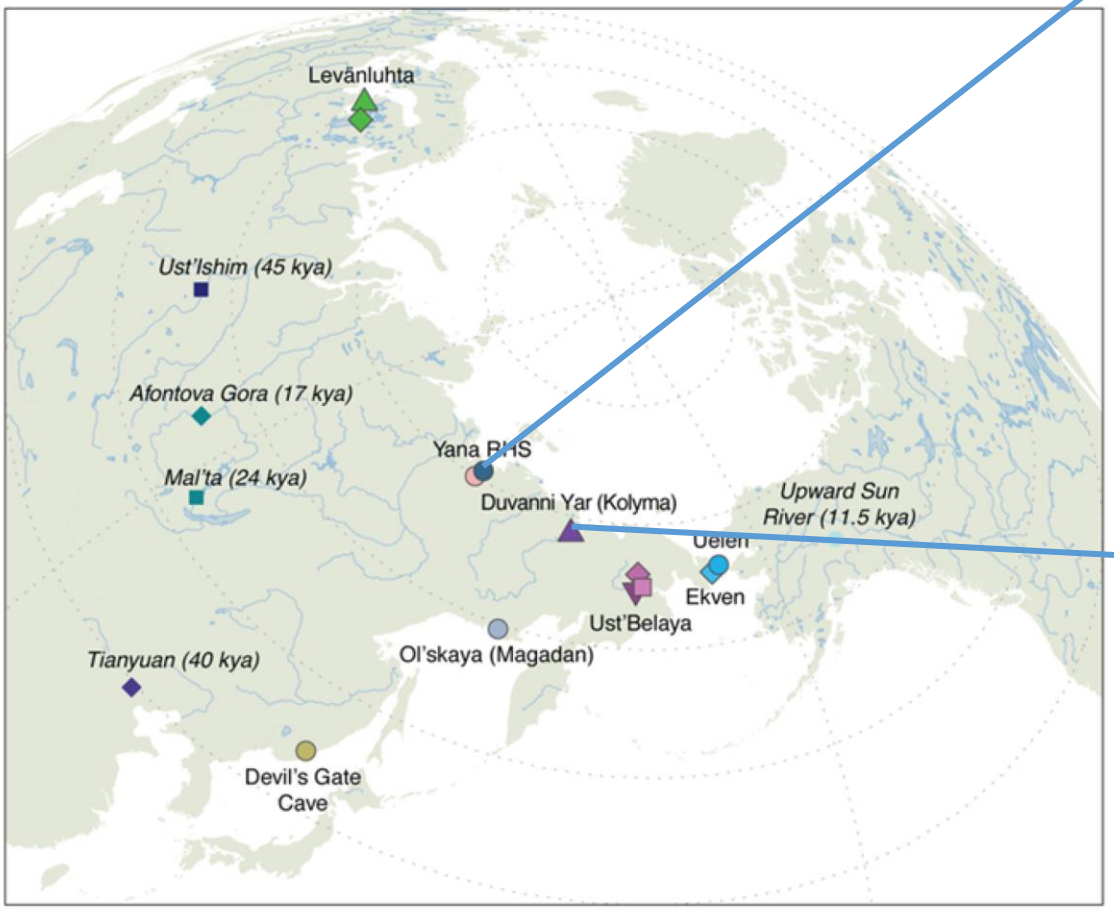
https://doi.org/10.1038/s41586-019-1279-z

The population history of northeastern Siberia since the Pleistocene

Martin Sikora[1,43]*, Vladimir V. Pitulko[2,43]*, Vitor C. Sousa[3,4,5,43], Morten E. Allentoft[1,43], Lasse Vinner[1], Simon Rasmussen[6,41], Ashot Margaryan[1], Peter de Barros Damgaard[1], Constanza de la Fuente[1,42], Gabriel Renaud[1], Melinda A. Yang[7], Qiaomei Fu[7], Isabelle Dupanloup[8], Konstantinos Giampoudakis[9], David Nogués-Bravo[9], Carsten Rahbek[9], Guus Kroonen[10,11], Michaël Peyrot[11], Hugh McColl[1], Sergey V. Vasilyev[12], Elizaveta Veselovskaya[12,13], Margarita Gerasimova[12], Elena Y. Pavlova[2,14], Vyacheslav G. Chasnyk[15], Pavel A. Nikolskiy[2,16], Andrei V. Gromov[17], Valeriy I. Khartanovich[17], Vyacheslav Moiseyev[17], Pavel S. Grebenyuk[18,19], Alexander Yu. Fedorchenko[20], Alexander I. Lebedintsev[18], Sergey B. Slobodin[18], Boris A. Malyarchuk[21], Rui Martiniano[22], Morten Meldgaard[1,23], Laura Arppe[24], Jukka U. Palo[25,26], Tarja Sundell[27,28], Kristiina Mannermaa[27], Mikko Putkonen[25], Verner Alexandersen[29], Charlotte Primeau[29], Nurbol Baimukhanov[30], Ripan S. Malhi[31,32], Karl-Göran Sjögren[33], Kristian Kristiansen[33], Anna Wessman[27,34], Antti Sajantila[25], Marta Mirazon Lahr[1,35], Richard Durbin[22,36], Rasmus Nielsen[1,37], David J. Meltzer[1,38], Laurent Excoffier[4,5]* & Eske Willerslev[1,36,39,40]*

Nature (2019)

Colonization of Siberia



Yana RHS (31,600 years ago)

Whole-genome depth of coverage 25x

Kolyma (9,800 years ago)

Whole-genome depth of coverage 14x

Hypothesis: Continuity vs Replacement of populations

Data: Ancient and present-day samples; 625 blocks of 1Mb (~1.5 Million SNP), far from genic regions and CpG islands

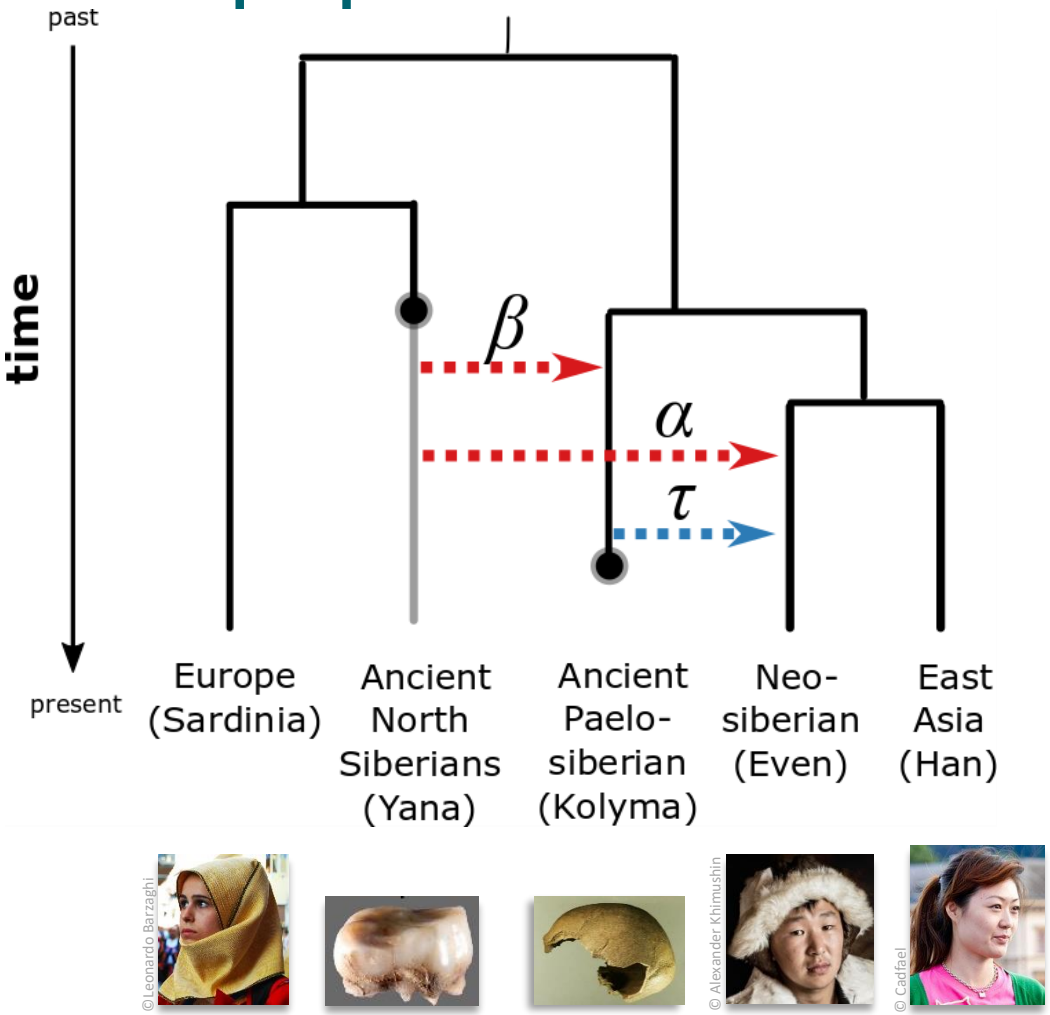
Method: Composite likelihood - *fastsimcoal2*
(Excoffier et al, 2013 Plos Genetics)

Hypothesis: Continuity vs Replacement of populations

For instance:

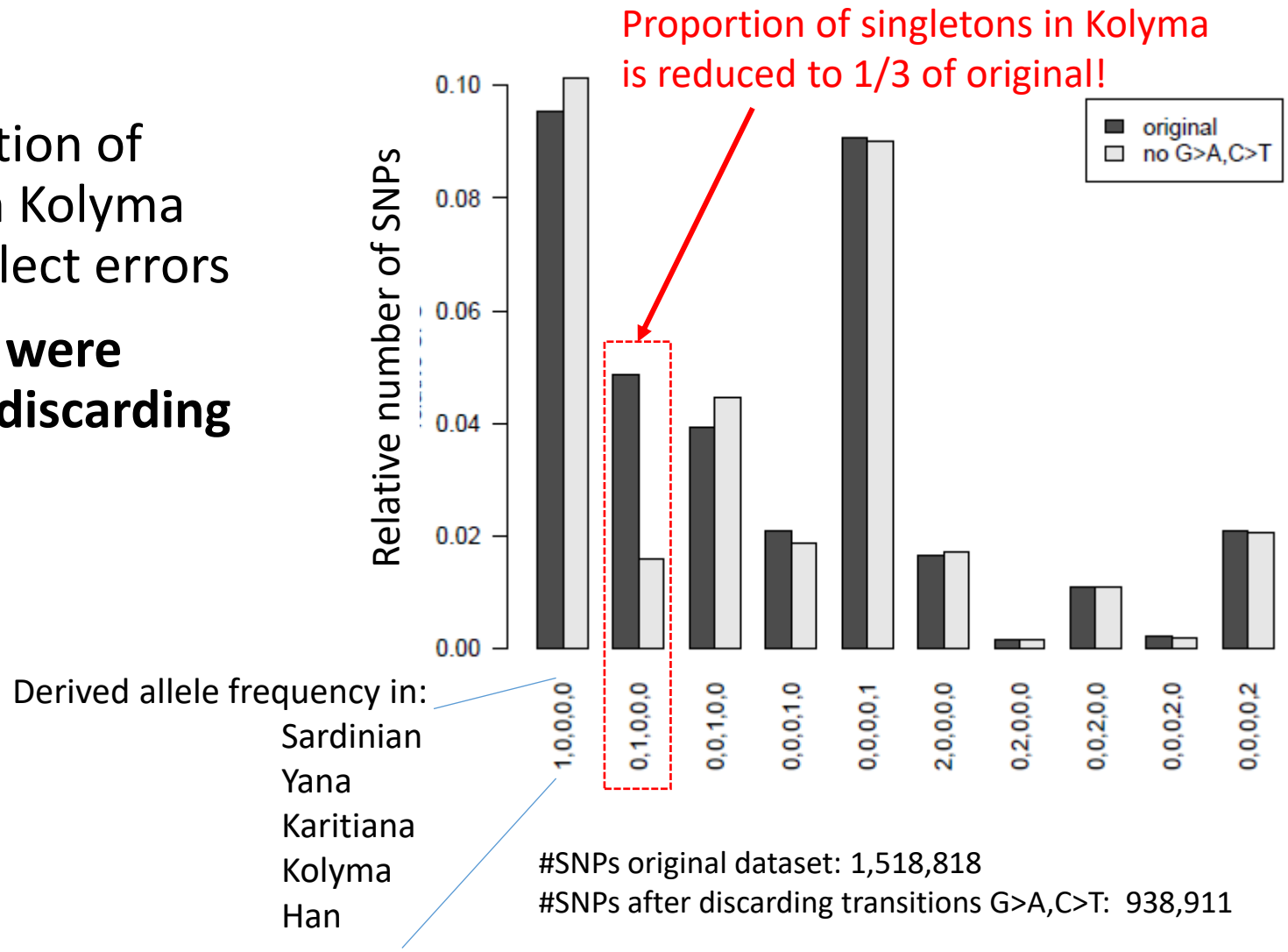
β = 1 indicates continuity: Kolyma descends from Yana

β = 0 indicates replacement of Yana by Kolyma

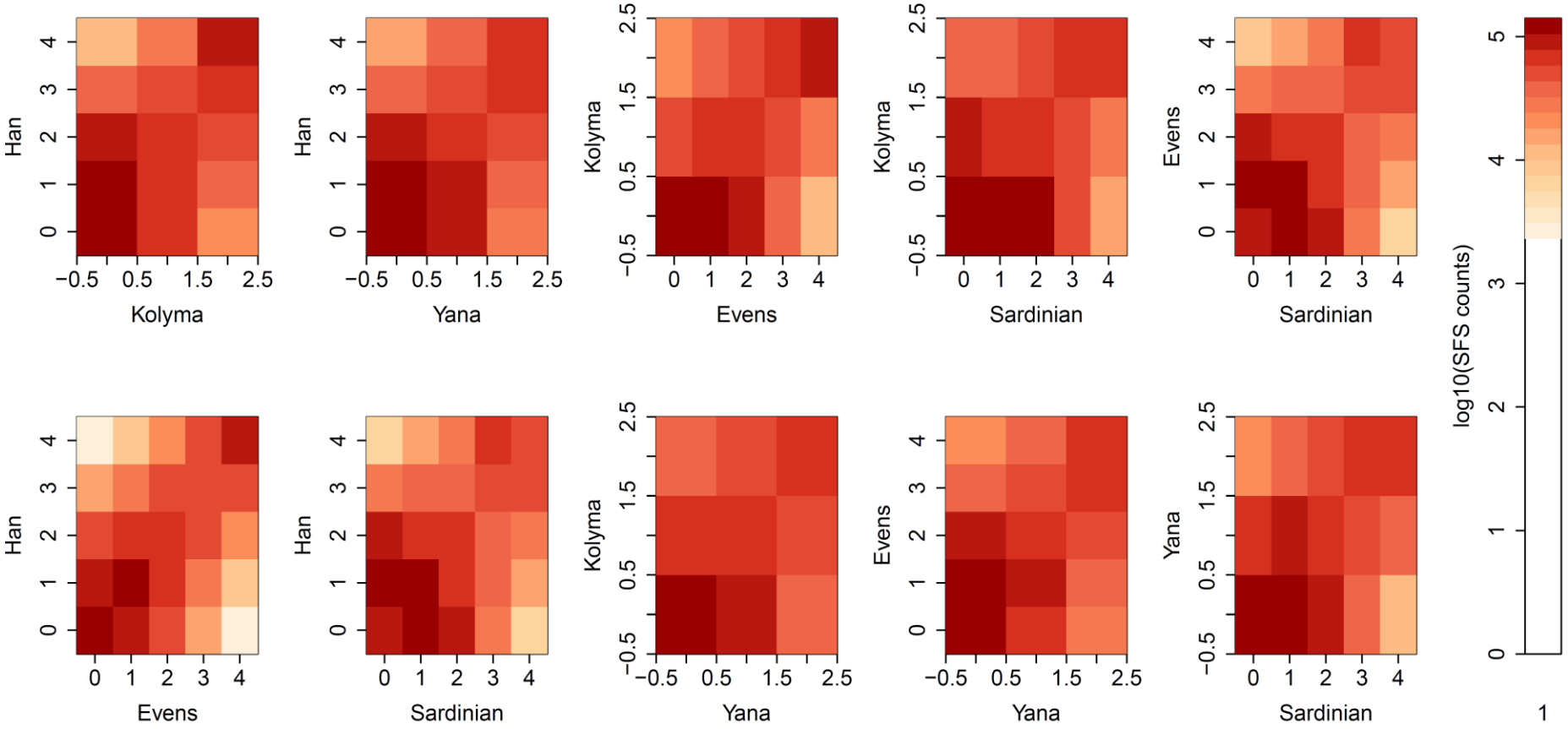


Site frequency spectrum is affected by damage patterns in ancient DNA

- High proportion of singletons in Kolyma probably reflect errors
- **All analyses were performed discarding singletons**

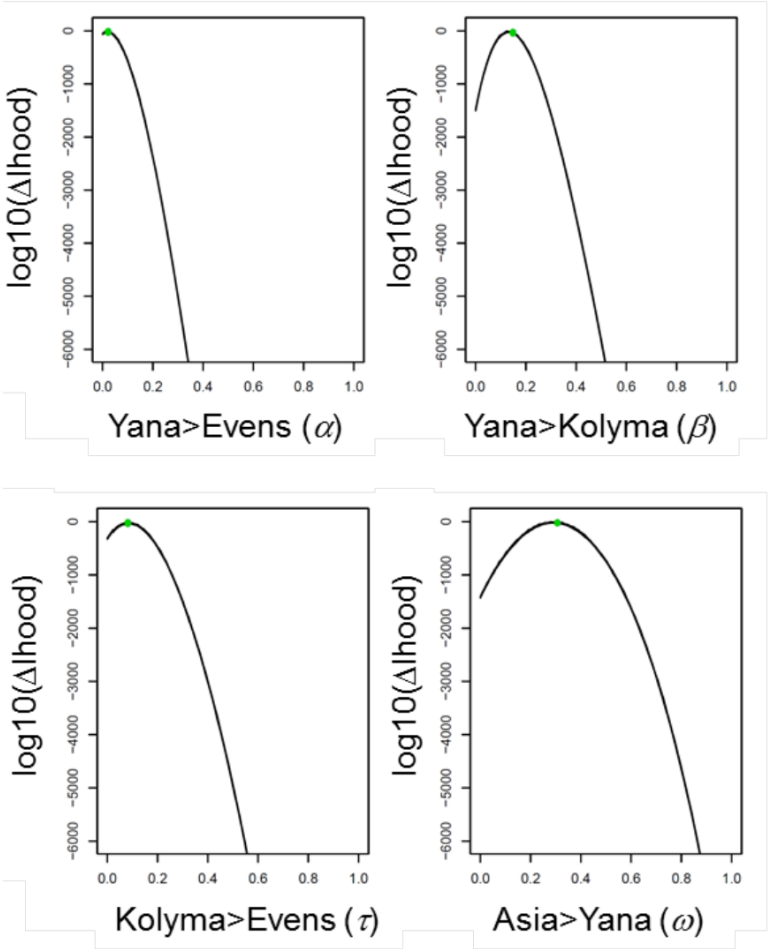


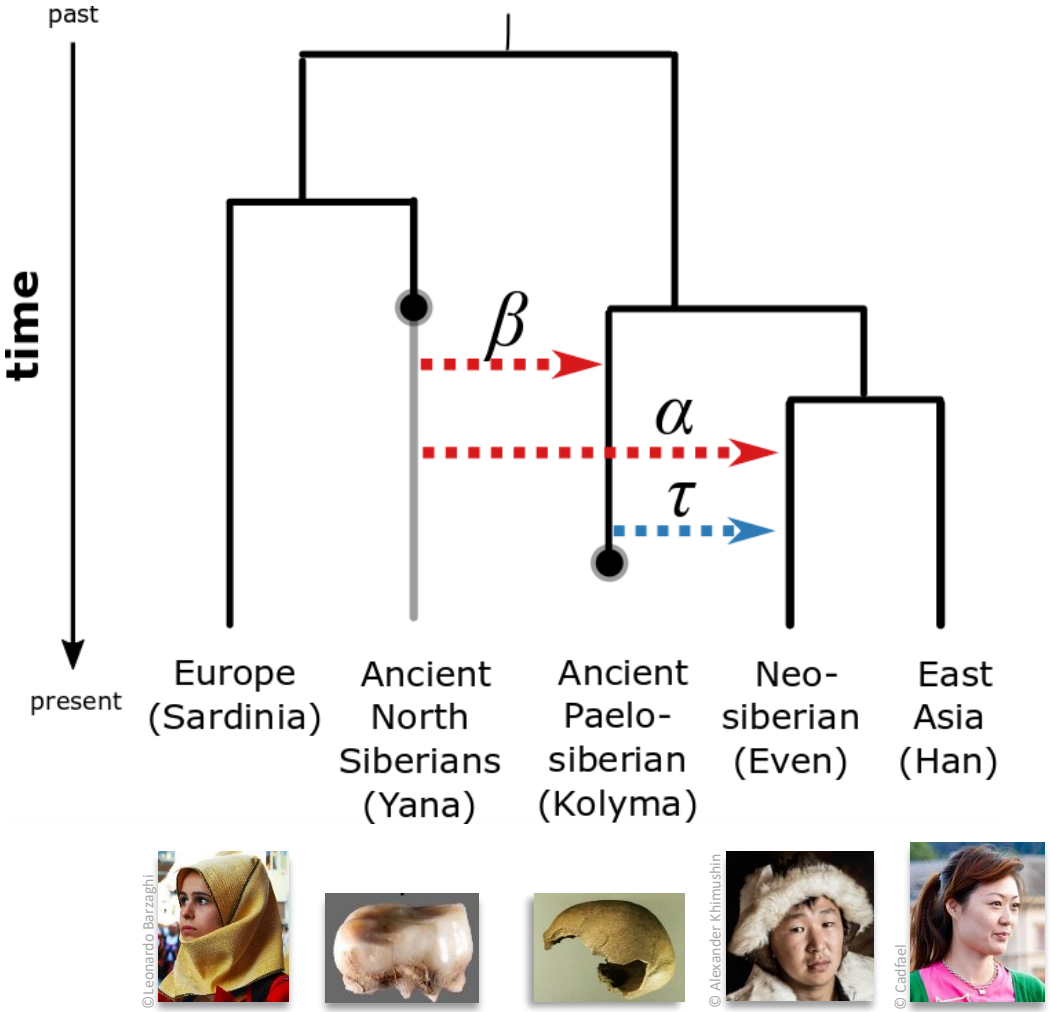
Data: Marginal 2D-SFS



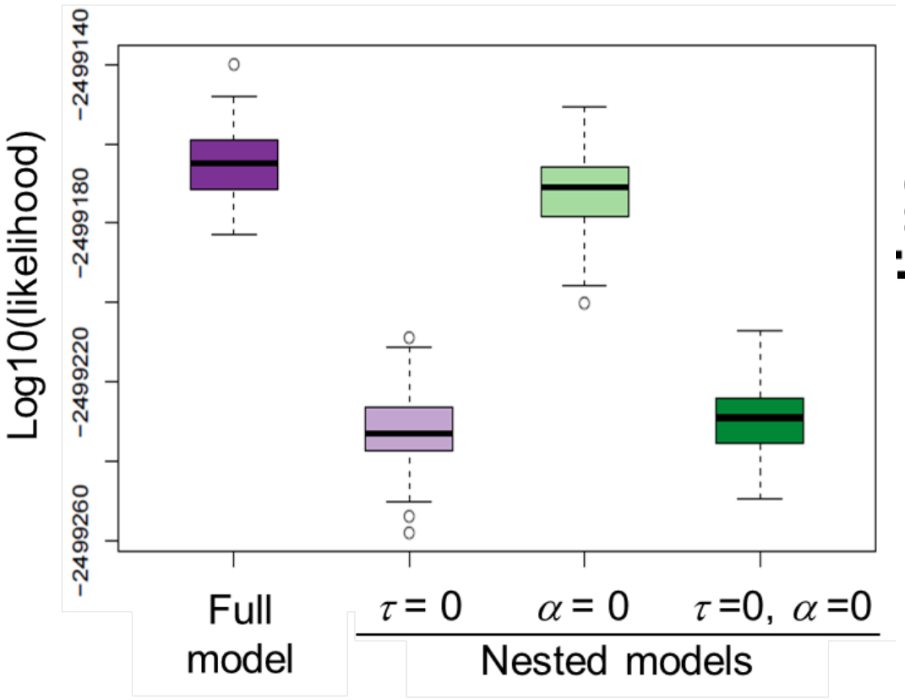
Observed Data: Joint 5 population site-frequency spectrum (1125 entries) obtained from 625 blocks of 1Mb (~1.5 Million SNP)

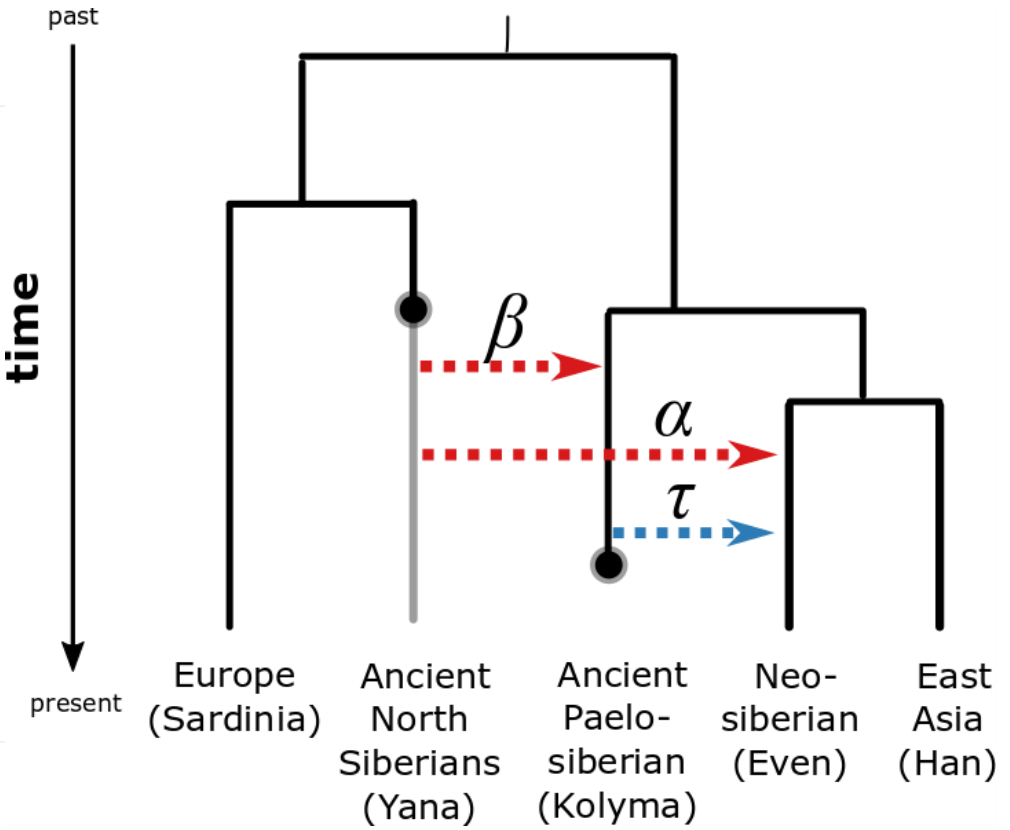
Model comparison and likelihood profiles consistent with replacement with gene flow



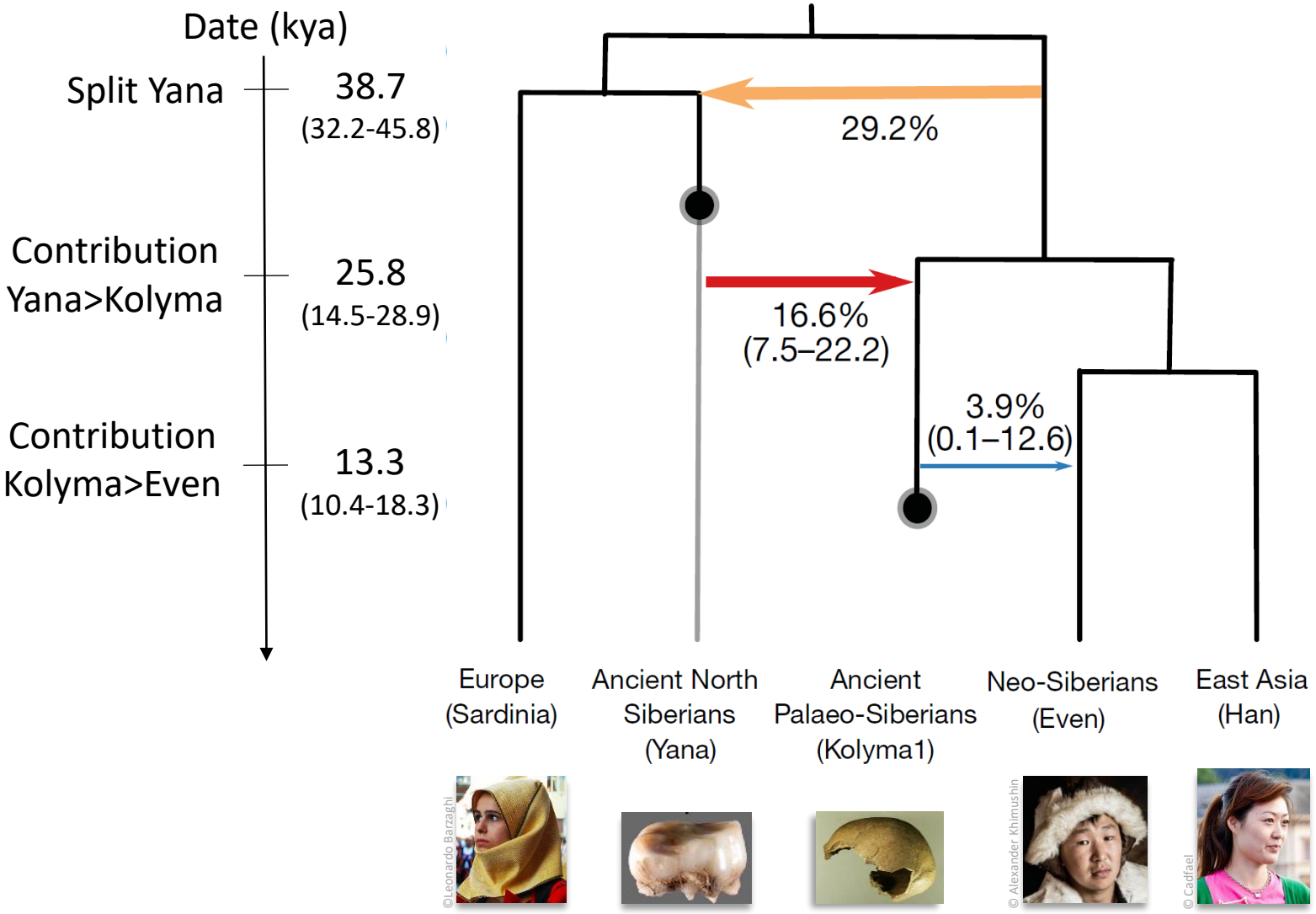


Model comparison and likelihood profiles consistent with replacement with gene flow

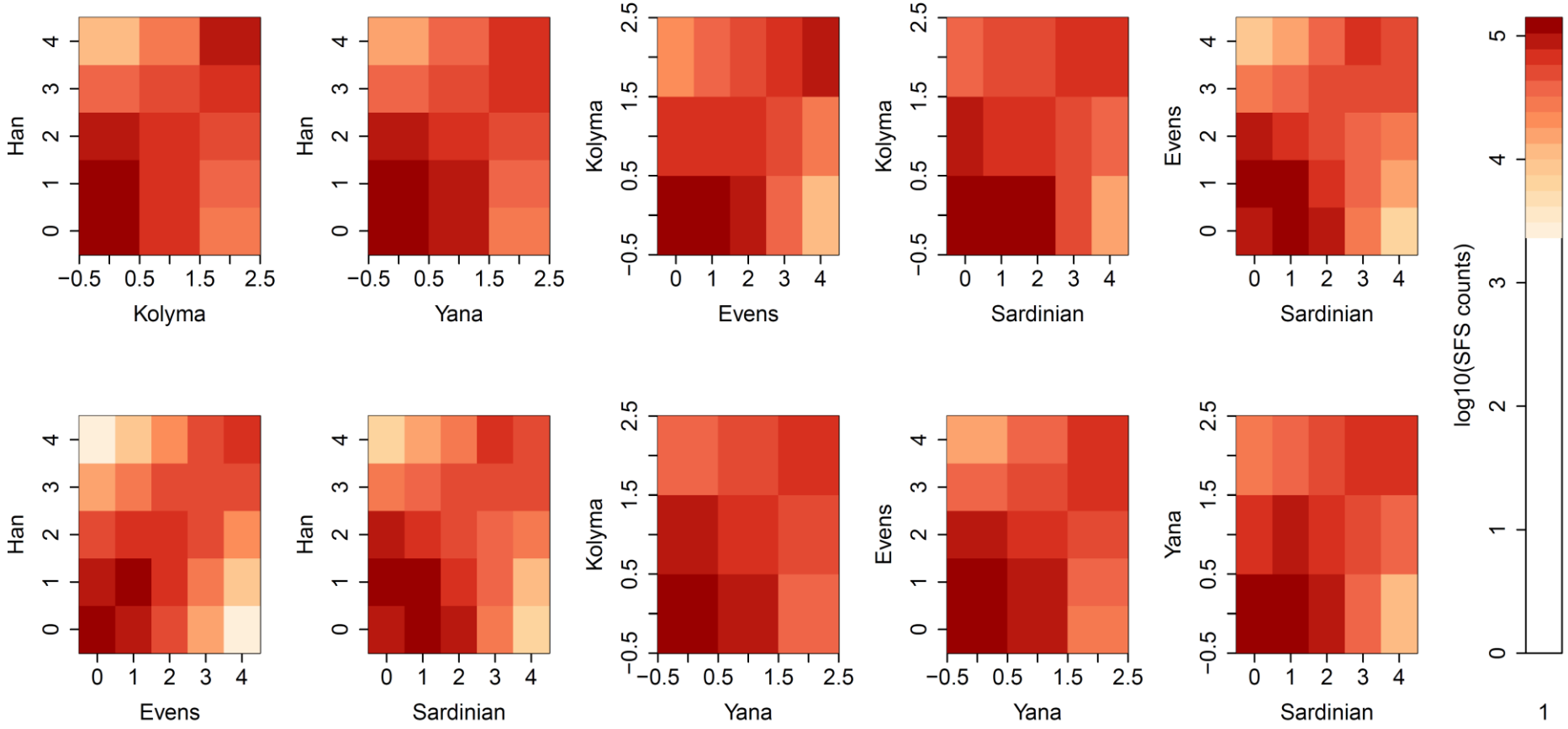




Estimates of best nested model indicate replacement with gene flow

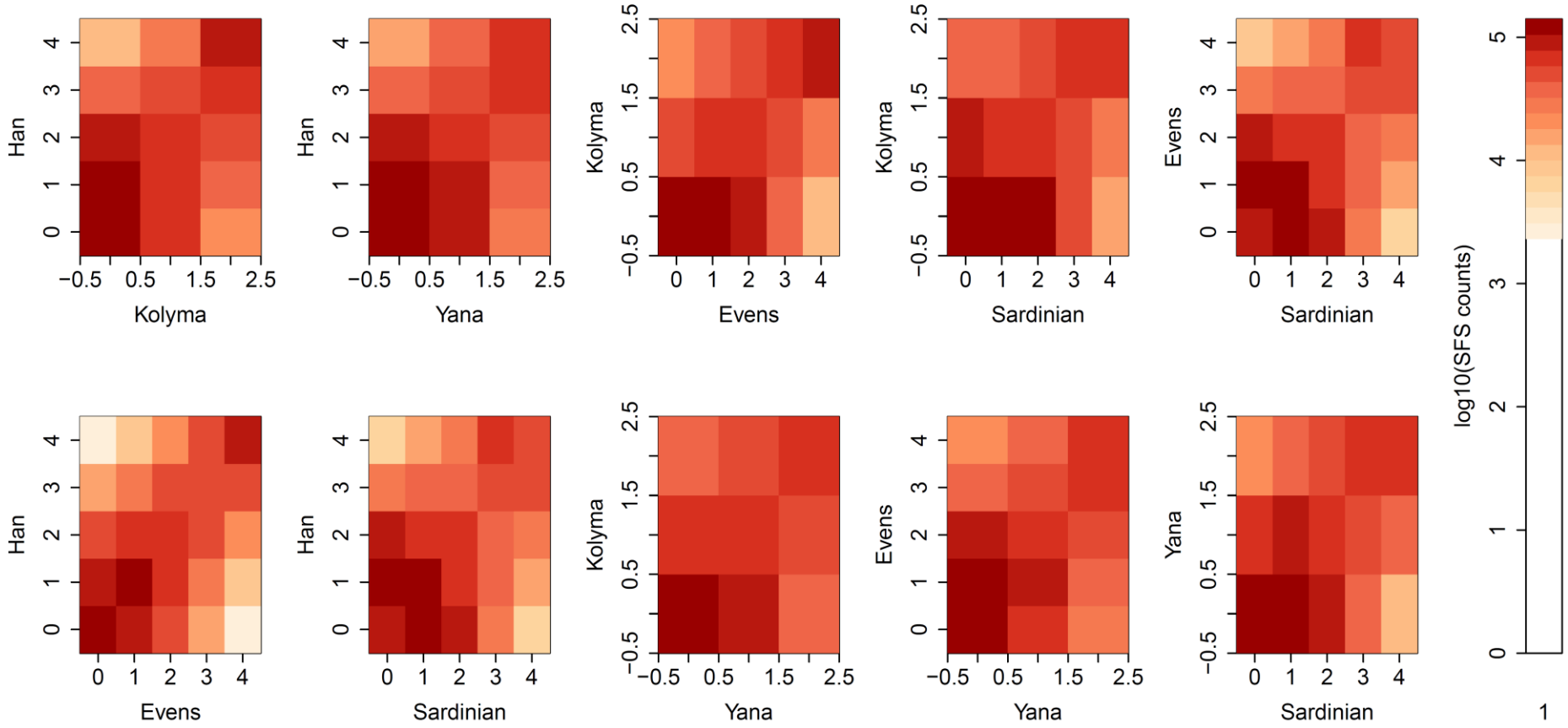


Fit of expected SFS to observed data



Expected SFS according to the parameters that maximize the likelihood

Fit of expected SFS to observed data



Observed SFS

Coat color adaptation in deer mice *Peromyscus maniculatus*

- Habitat (soil color) correlated with coat phenotype
- Field experiments suggest that light color confers selective advantage against visually hunting predators
- Nebraska Sand Hills were formed 8000 to 15,000 years ago

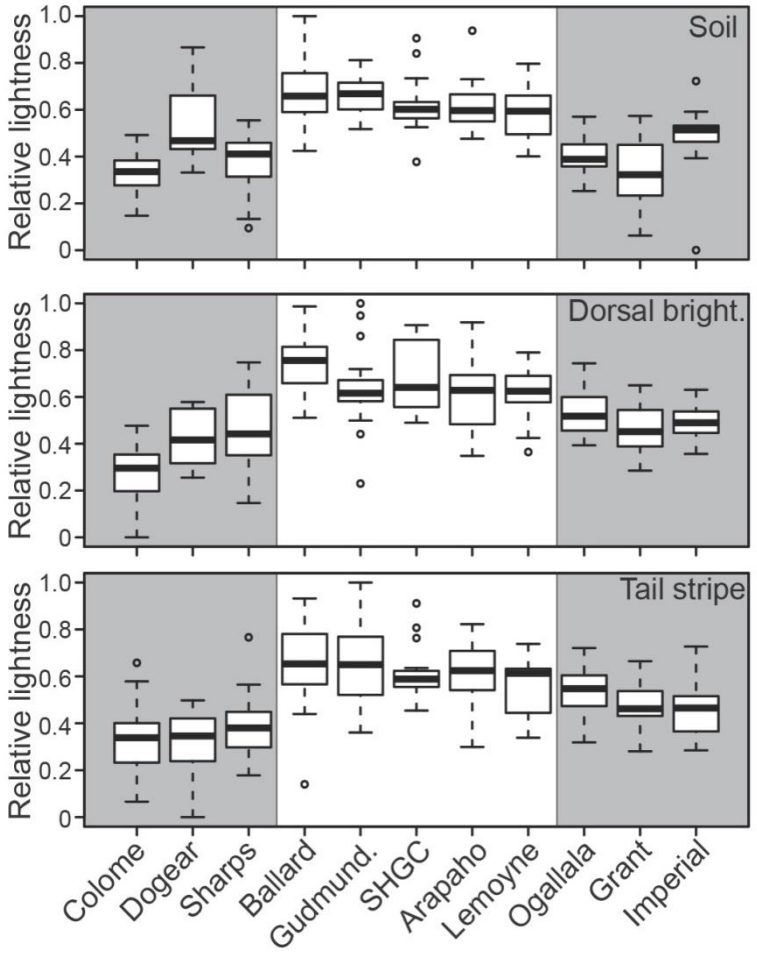
Linnen et al (2013) Science

Pfeifer*, Laurent*, Sousa* et al (2018) MBE

A transect across the Sand Hills (ON and OFF)

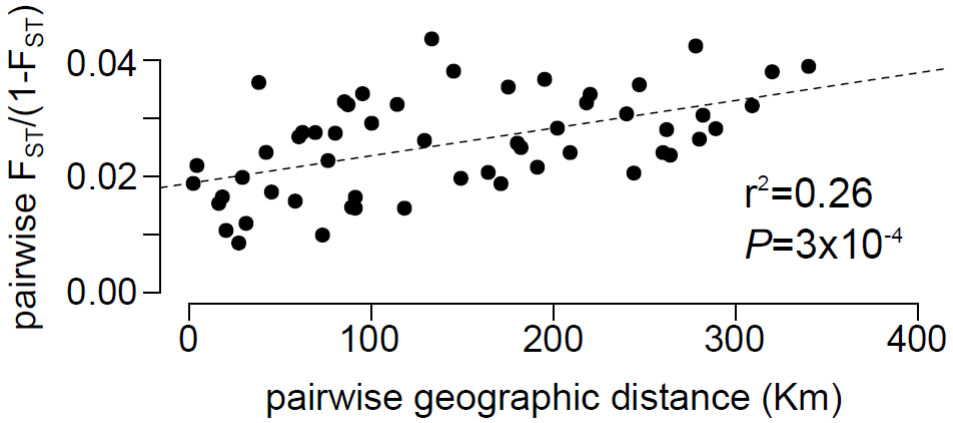
Sample locations “off” and “on” the Sand Hills

- 11 populations
- 330 individuals

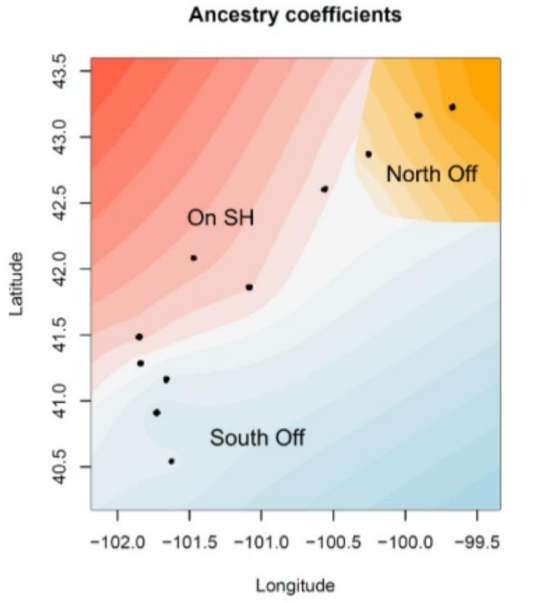


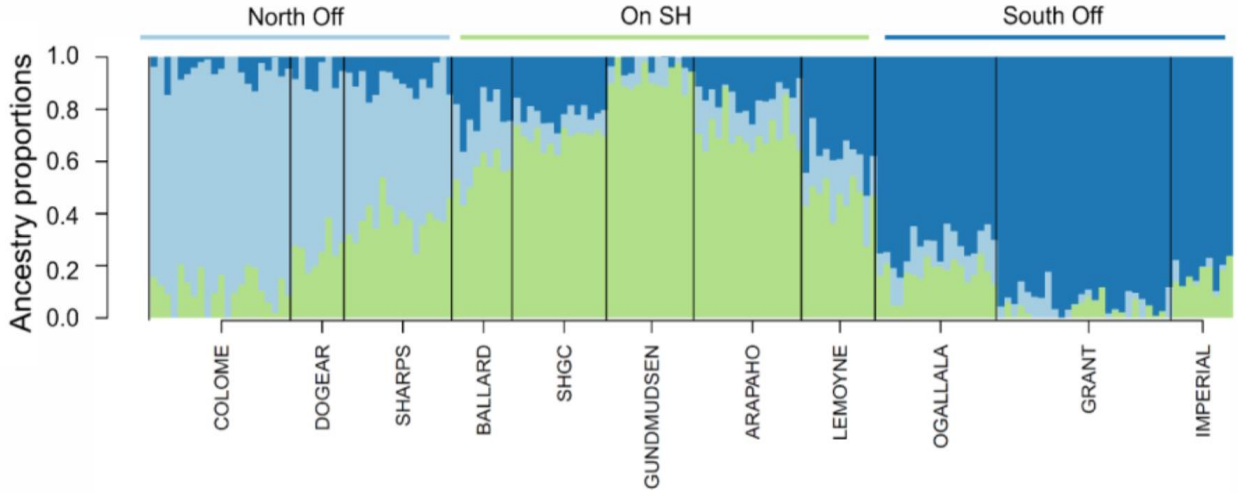
- Genomic data (NGS) data
  - Target 10,000 random 1.5kb regions
  - 185kbp region comprising the *Agouti* gene
- Phenotypic data for each individual

Evidence for isolation by distance but three groups



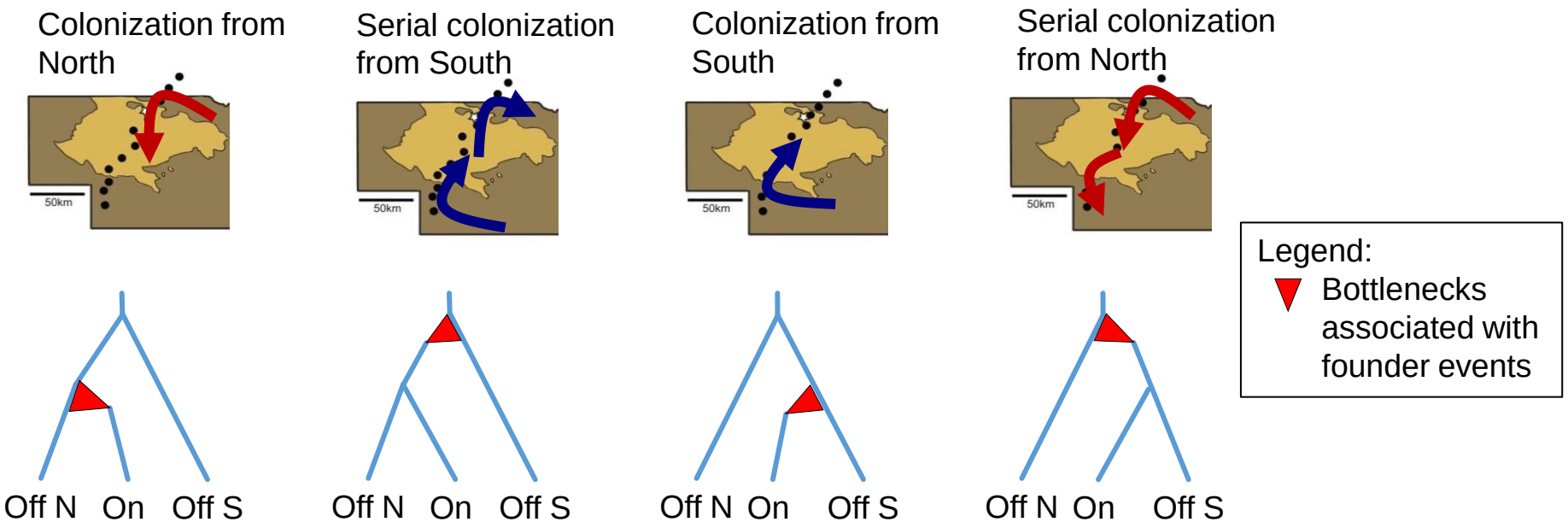
Geographically closer samples are genetically more similar





Model-based inference

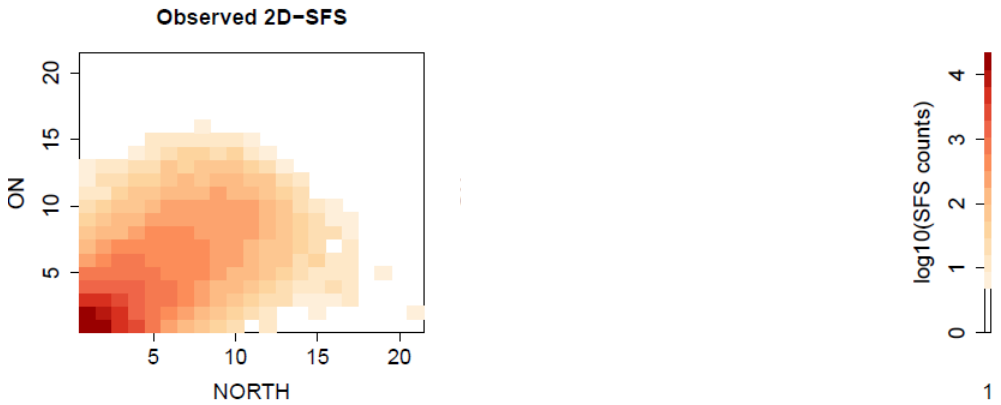
Is there evidence of gene flow between Off and On the Sand Hills?

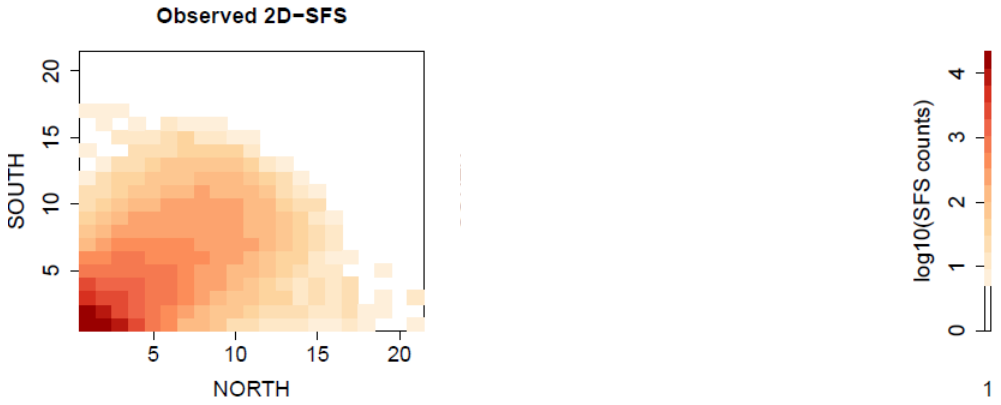


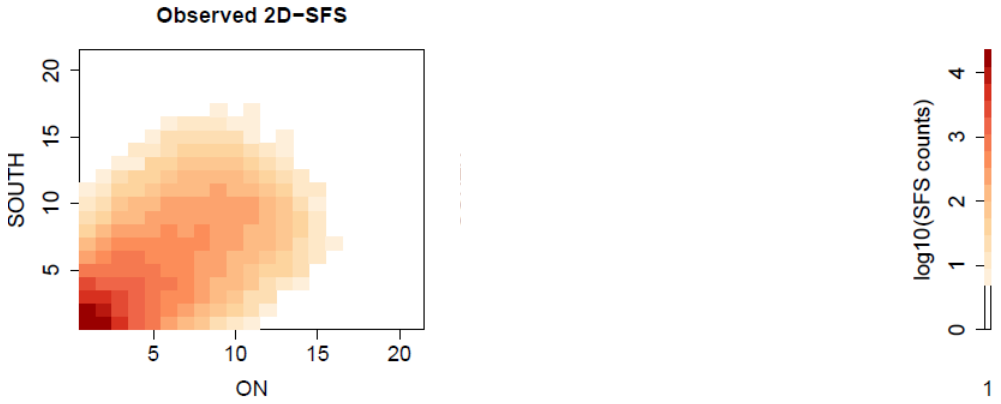
Estimates based on the joint **3D site frequency spectrum** (SFS):

- folded SFS with 140,358 SNPs

Deer mice: Pairwise marginal 2D SFS
Since we did not have an outgroup we used the folded SFS



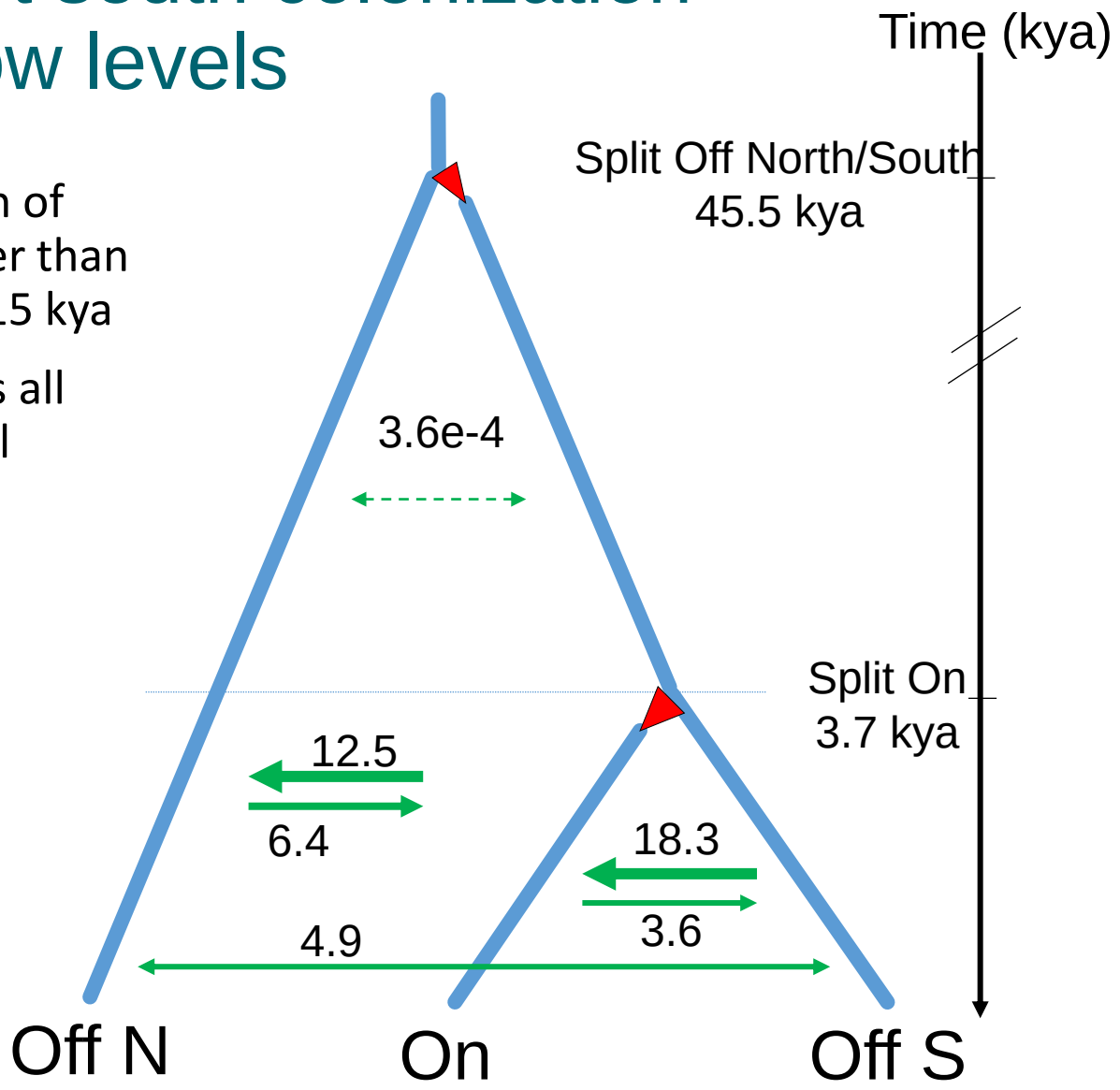




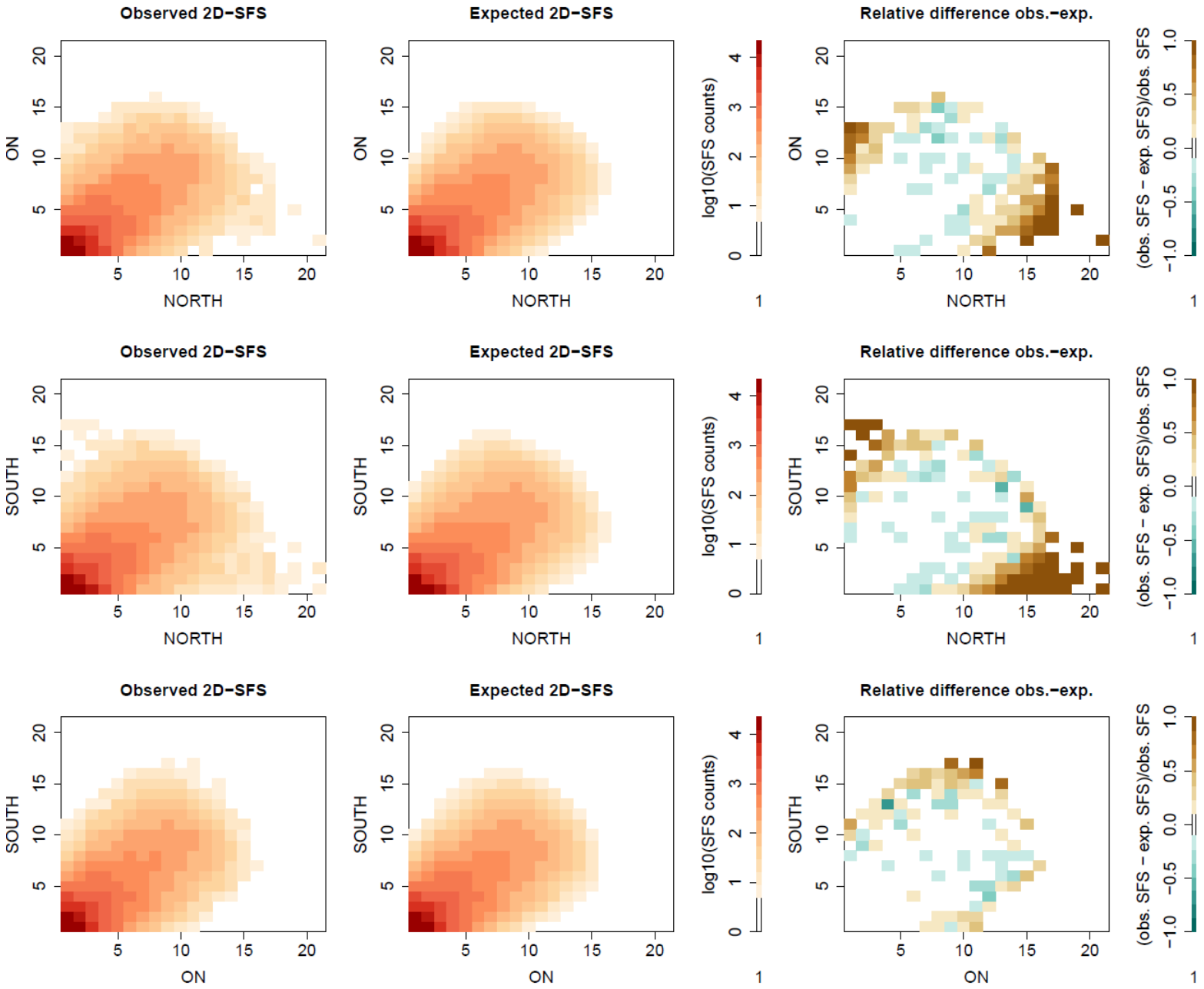
Estimates support south colonization and high gene flow levels

- Recent time of colonization of Sand Hills ~3-5 kya, younger than formation of Sand Hills 8-15 kya
- High migration rates across all populations, inferred for all models

Migration rates above/below arrows in units of 2Nm, i.e. average number of immigrants per generation.



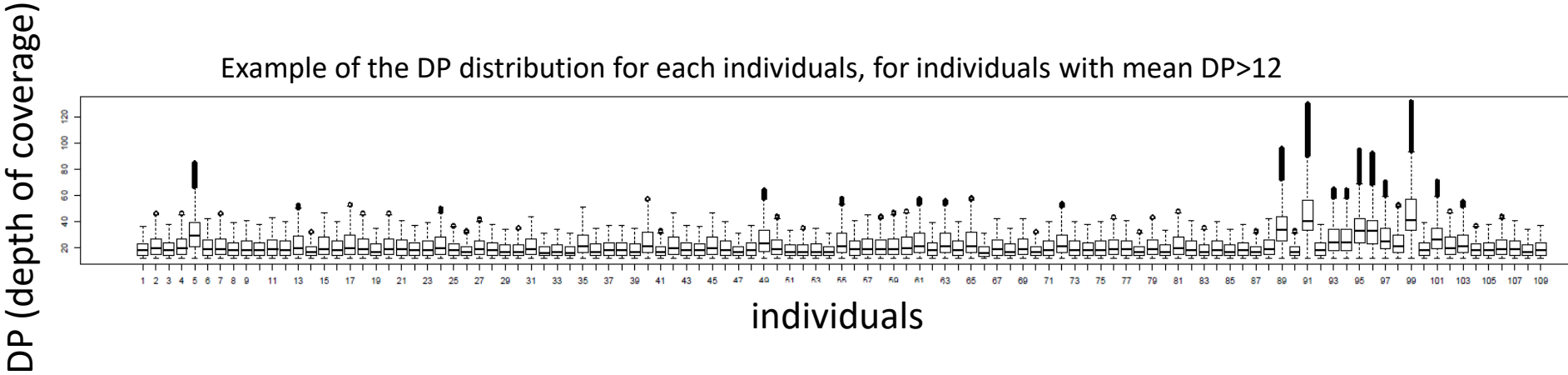
Deer mice: Model fit to marginal SFS



Some lessons I learned working with the deer mice data

- Be carefull when applying Hardy-Weinberg filters to your data!
- Be carefull when filtering on depth of coverage applying the same thresholds for all individuals!

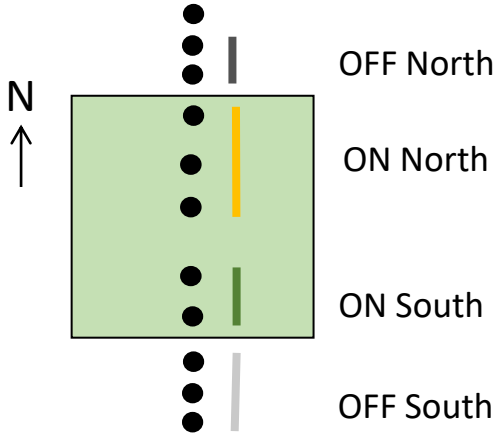
The depth of coverage varied considerably across individuals



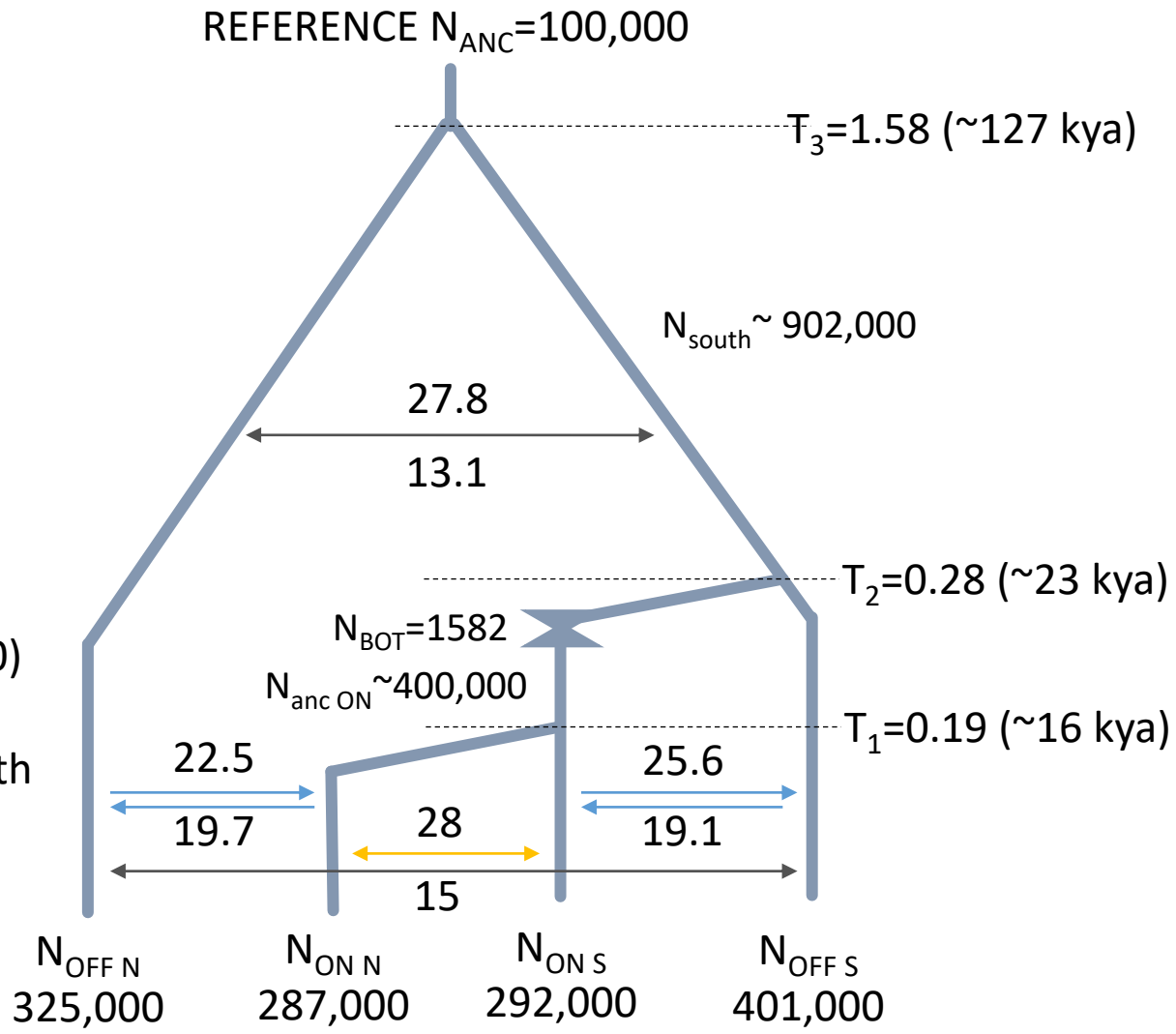
- Applying the same threshold for all individuals can lead to biases
- Apply a filter on DP for each individual

Effect of HW filtering on demographic estimates

Removing sites with HWE excess and deficit leads to different estimates



- High migration between all groups of populations (2Nm~20)
- No evidence of a strong bottleneck signal associated with colonization of SH



Sawflies and RAD data

MOLECULAR ECOLOGY

Molecular Ecology (2016) doi: 10.1111/mec.13972

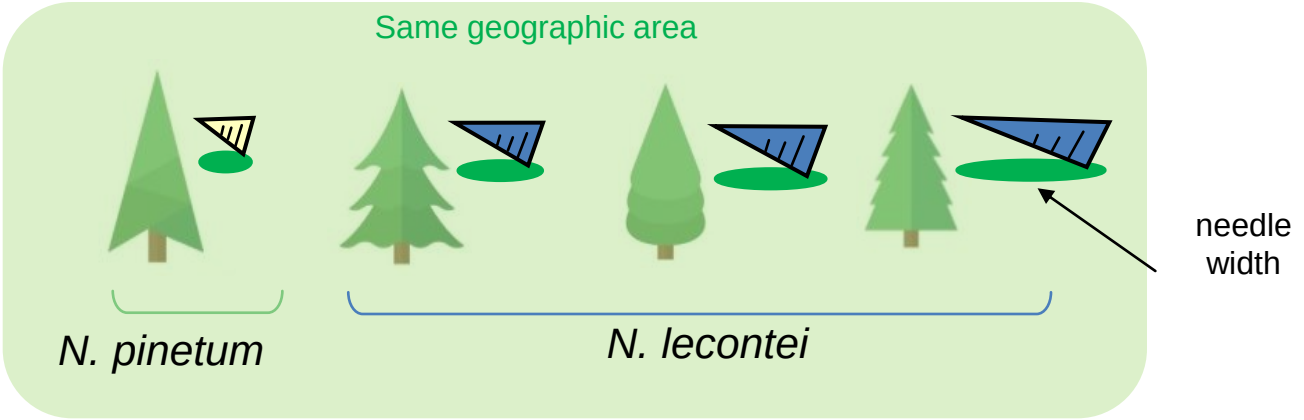
History, geography and host use shape genomewide patterns of genetic variation in the redheaded pine sawfly (*Neodiprion lecontei*)

ROBIN K. BAGLEY,* VITOR C. SOUSA,† MATTHEW L. NIEMILLER‡ and CATHERINE R. LINNEN*

**Department of Biology, University of Kentucky, Lexington, KY 40506, USA, †cE3c - Centre for Ecology, Evolution and Environmental Changes, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal, ‡Illinois Natural History Survey, Prairie Research Institute, University of Illinois Urbana-Champaign, Champaign, IL 61820, USA*

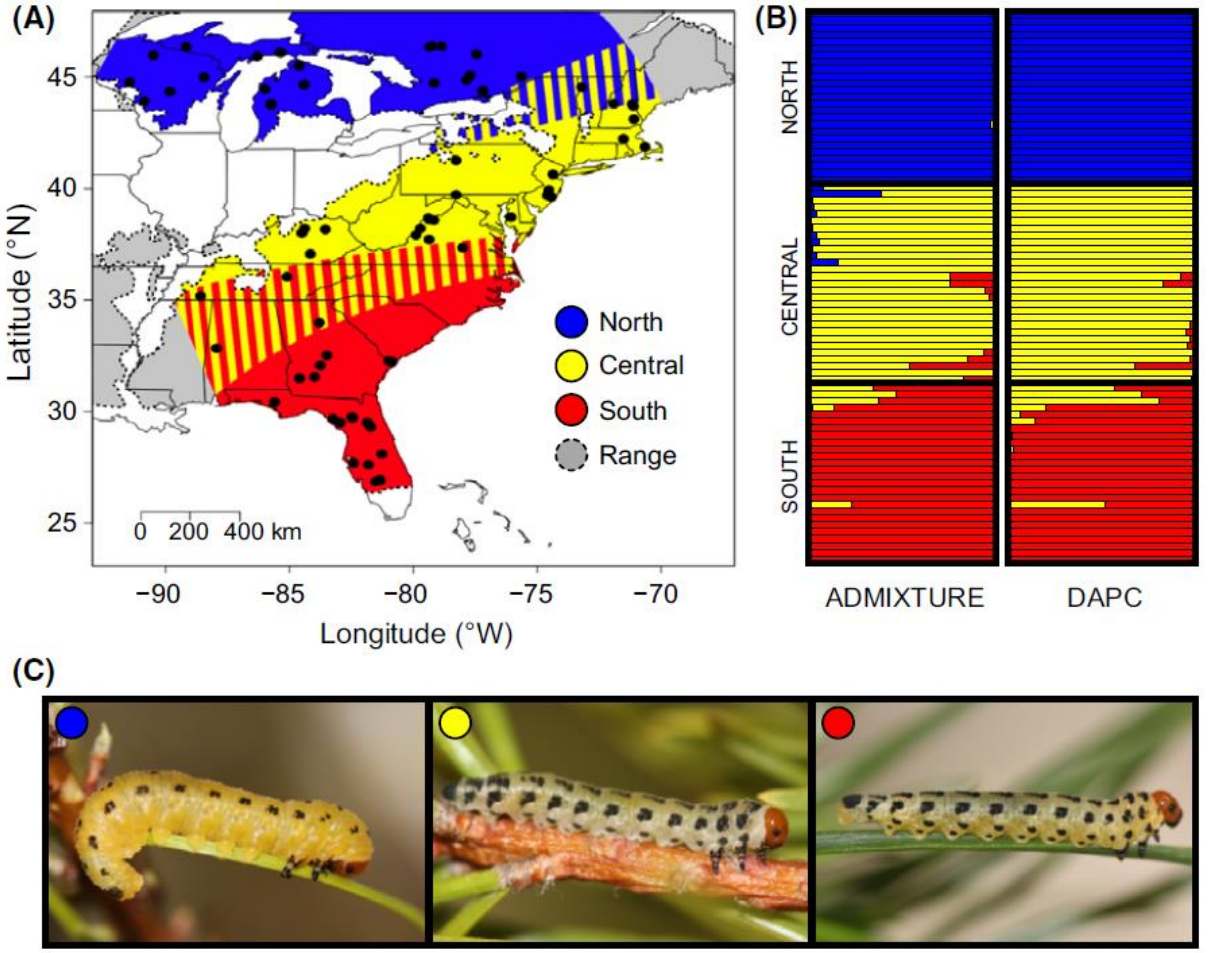
Sawflies *Neodiprion lecontei*

- Hymenoptera
- Plant-feeding insects
- Pine tree specialists



ddRAD seq data

- 80 individuals from 77 localities and 13 host species
- 100 bp paired-end reads, mapped to reference genome of *N. lencontei*
- Depth of coverage filter DP>10

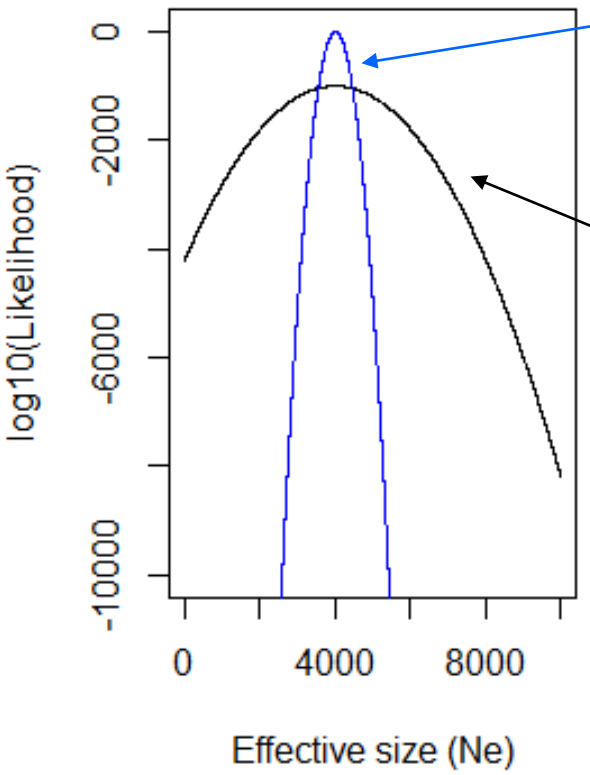


Given the detected three groups (North, Central, South):

- What is the the population tree topology?
- What are the split times?
- What are the migration levels among groups?

Comparing models with composite likelihoods

- Fastsimcoal2 likelihood is “correct” if all SNPs are independent
- We can then compare the model likelihoods using Akaike Information Criterion (AIC)



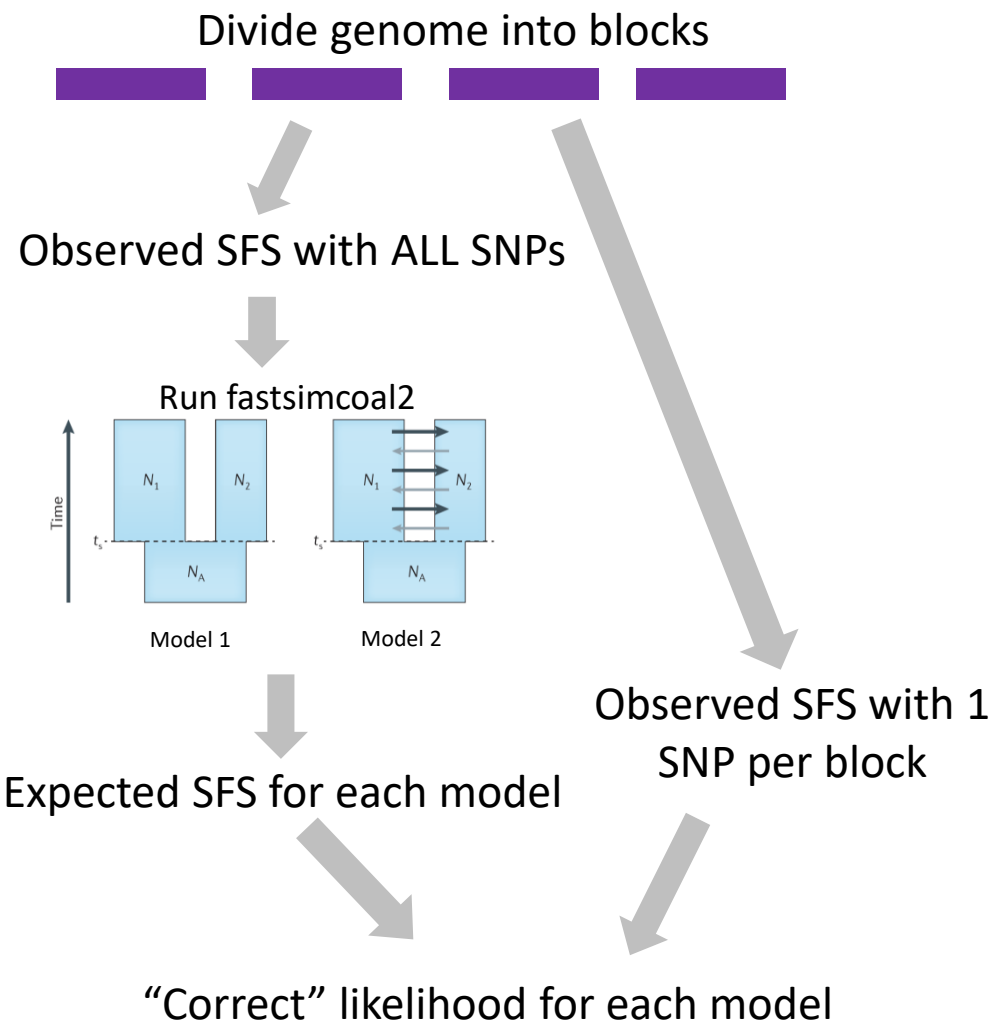
Composite likelihood (assuming linked sites are independent)

“correct” likelihood (all sites are actually independent)

Composite likelihood provide unbiased maximum likelihood parameter estimates, but the likelihoods are inflated

A strategy to compare models

1. Divide the dataset into LD blocks.
2. Create a dataset with all SNPs (including linked SNPs)
3. For each model, obtain the parameters that maximize the likelihood (this is ok even with linked sites!) and the corresponding expected SFS
4. Create a dataset with "independent" SNPs (1 SNP per RAD tag)
5. Given the expected SFS of each model, compute the "correct" likelihood for each model with the dataset with independent SNPs
6. Compare models with AIC

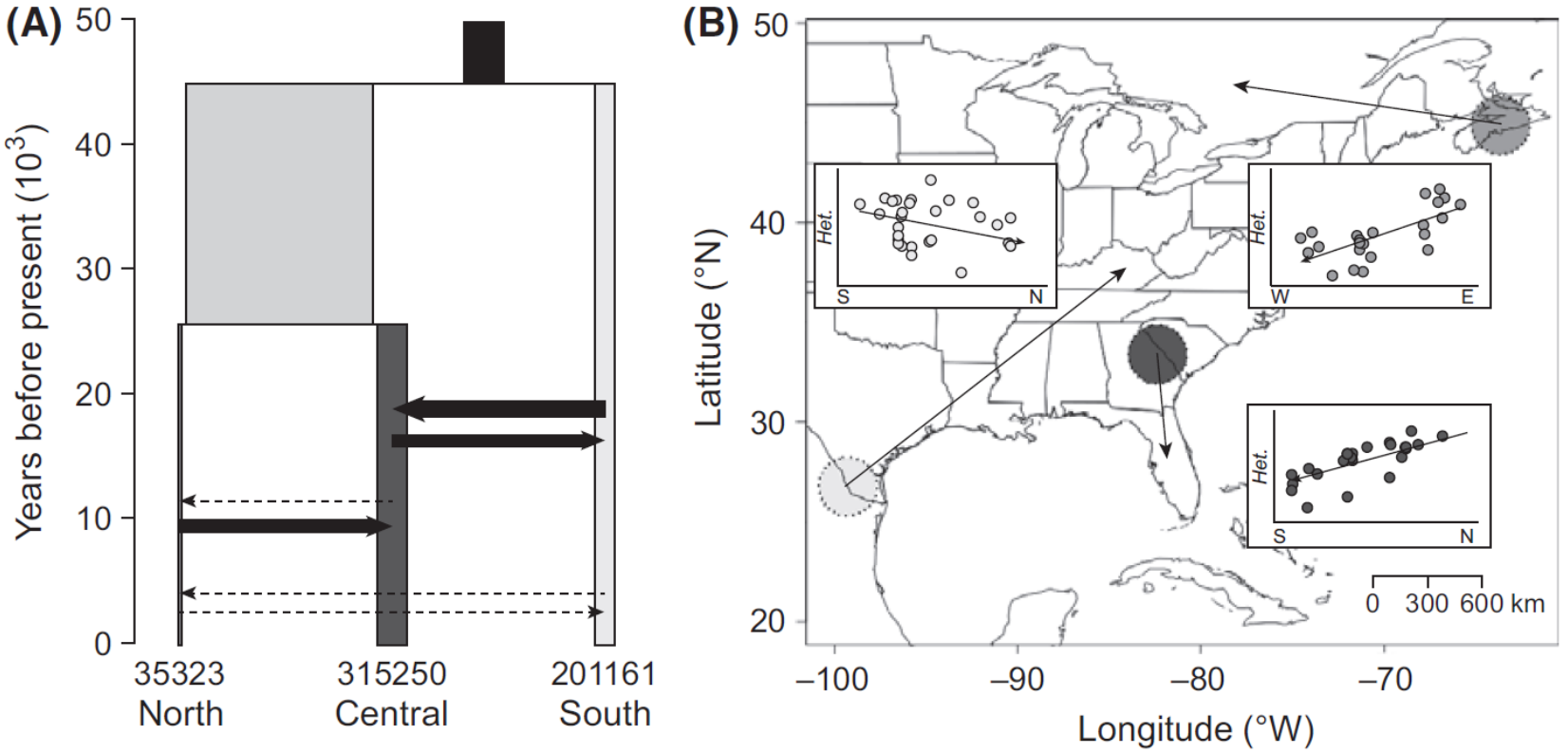


Comparing alternative models

Table 2 Summary of the likelihoods for the sixteen demographic models tested. Lhood (ALL SNPs) and Lhood (1 SNP) correspond to the mean likelihood computed with the data sets containing 'all SNPs' (including monomorphic sites) and a 'single SNP' (without monomorphic sites) per RAD locus, respectively. Mean likelihoods were computed based on 100 expected site frequency spectra simulated according to the parameters that maximized the likelihood of each model. Topology names for each model are as indicated in Fig. S1 (Supporting information). AIC scores and relative likelihoods (Akaike's weight of evidence) were calculated based on the 'single SNP' data set following Excoffier *et al.* 2013.

Topology	Migration allowed?	Exponential growth?	North bottleneck?	$\log_{10}$(Lhood) ALL SNPs	$\log_{10}$(Lhood) 1 SNP	# Parameters	AIC	ΔAIC	Relative likelihood
North–South	No	No	No	−46502.02	−7381.4	7	34006.70	75.69	0.000
North–Central	No	No	No	−46475.82	−7369.0	7	33949.44	18.43	0.000
South–Central	No	No	No	−46502.18	−7381.6	7	34007.60	76.59	0.000
Trifurcation	No	No	No	−46501.54	−7380.4	5	33998.07	67.06	0.000
North–South	Yes	No	No	−46470.49	−7365.0	15	33947.25	16.24	~0.000
North–Central	Yes	No	No	−46462.24	−7361.5	15	33931.01	0.00	0.851
South–Central	Yes	No	No	−46467.69	−7363.8	15	33941.57	10.56	0.004
Trifurcation	Yes	No	No	−46470.28	−7364.7	11	33937.93	6.91	0.027
North–South	Yes	Yes	No	−46469.48	−7362.8	18	33942.91	11.90	0.002
North–Central	Yes	Yes	No	−46461.17	−7361.7	18	33937.82	6.80	0.028
South–Central	Yes	Yes	No	−46463.73	−7363.9	18	33948.15	17.13	~0.000
Trifurcation	Yes	Yes	No	−46467.72	−7363.3	14	33937.39	6.37	0.035
North–South	Yes	Yes	Yes	−46467.45	−7361.5	20	33940.86	9.85	0.006
North–Central	Yes	Yes	Yes	−46461.25	−7362.1	20	33943.82	12.81	0.001
South–Central	Yes	Yes	Yes	−46463.58	−7364.1	20	33953.08	22.07	0.000
Trifurcation	Yes	Yes	Yes	−46466.06	−7362.4	16	33936.93	5.92	0.044

Estimates favors a scenario where North and Central diverged more recently with asymmetric gene flow



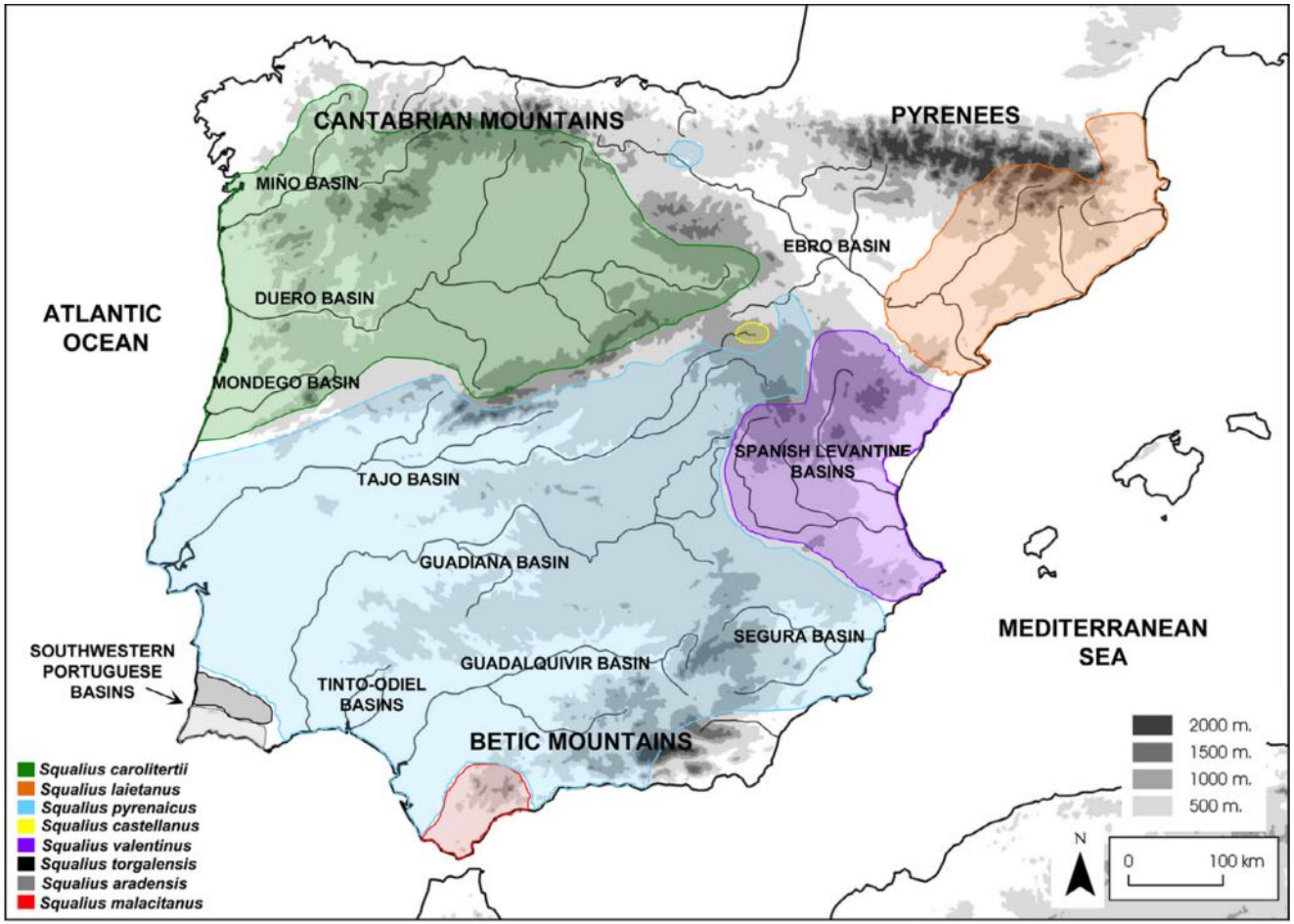
The inferred population tree topology and divergence times are consistent with divergence and range expansion from different refugia after LGM

Summary sawflies

- Fastsimcoal2 can be applied to RAD seq data
- We used a strategy to obtain (as close as possible) the “correct” likelihood by dividing the data into blocks, inferring the expected SFS for each model with ALL SNPs, and then re-computing the “true” likelihood with independent SNPs (1 SNP per block)
- Despite the reduced number of SNPs we were able to discriminate models based on their likelihoods

Inferring admixture in freshwater fish species

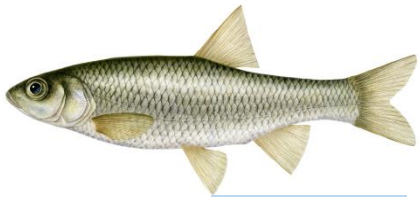
Sofia Mendes
(EGB, CE3C)



Mendes et al (2021) Heredity

GBS data (48 individuals, 23,562 SNPs with ~37% missing data)

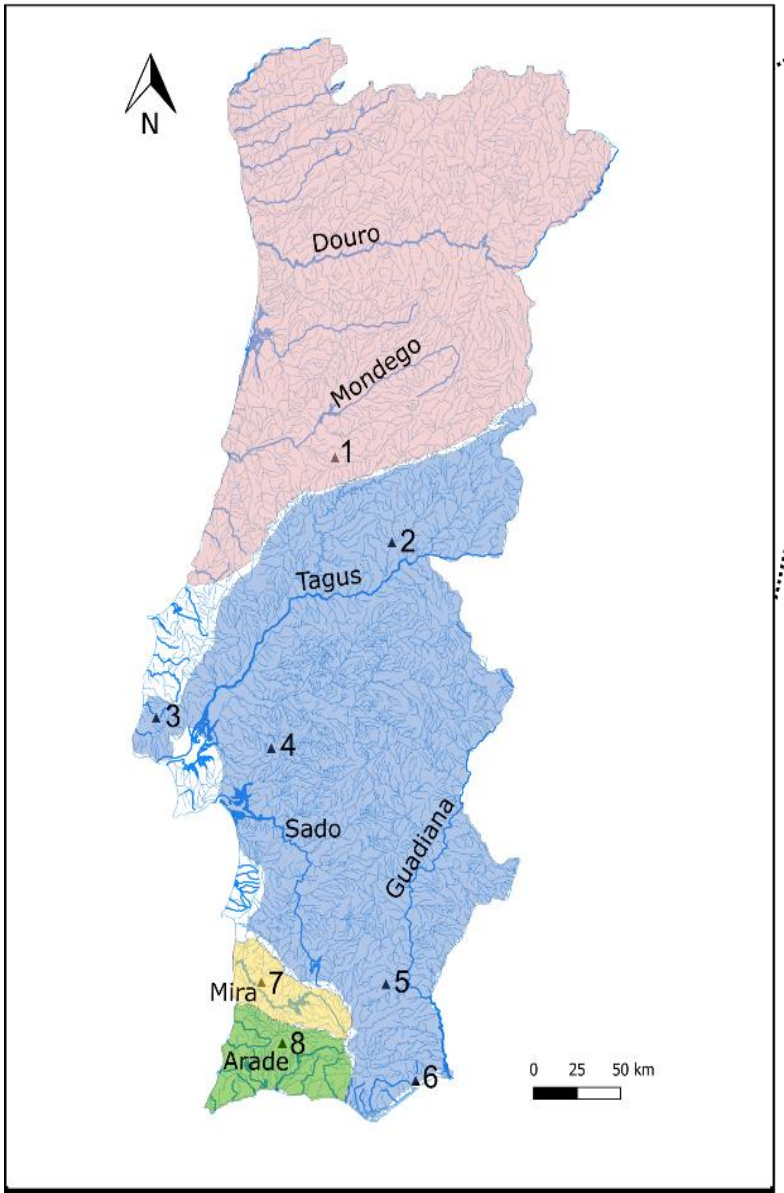
S. carolitertii



S. pyrenaicus

S. torgalensis

S. aradensis



Sampling locations:

1. Mondego — *S. carolitertii*
2. Ocreza
3. Lizandro — *S. pyrenaicus North*
4. Canha
5. Guadiana — *S. pyrenaicus South*
6. Almargem
7. Mira — *S. torgalensis*
8. Arade — *S. aradensis*

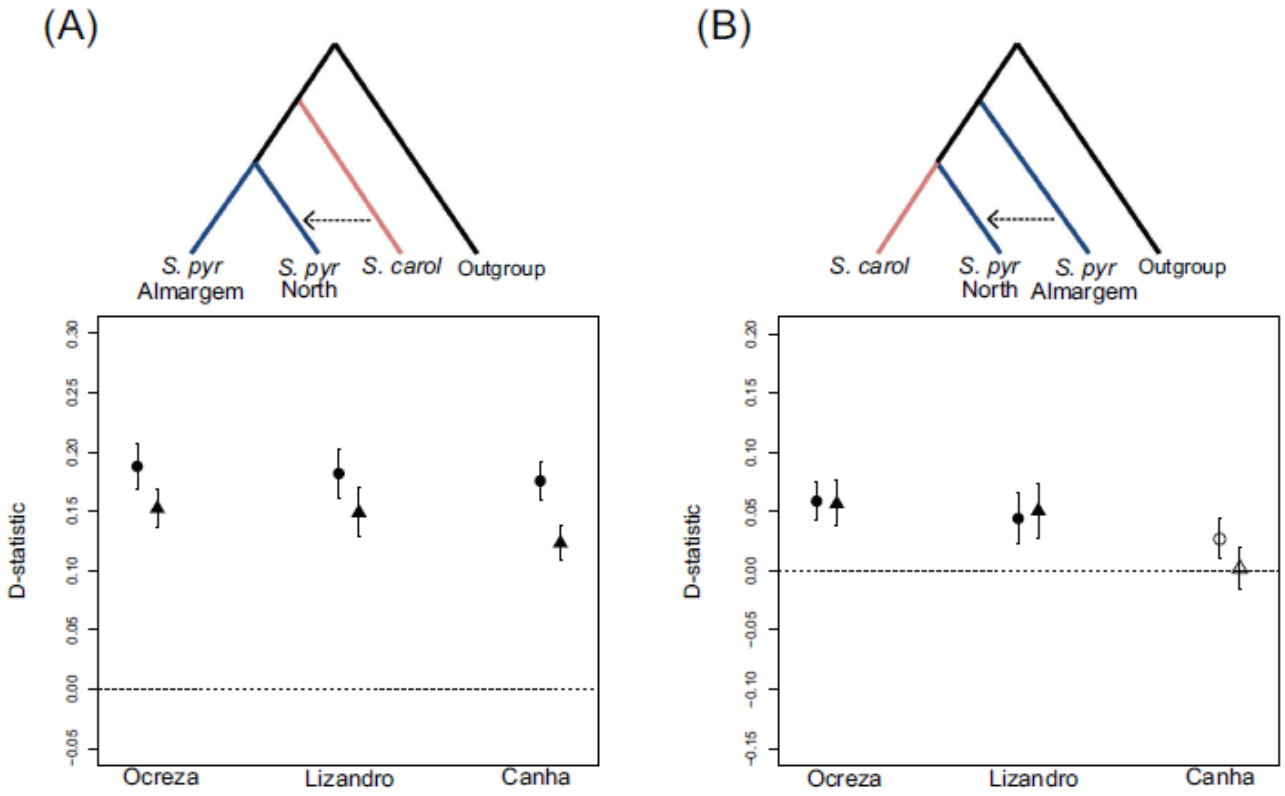
S. carolitertii

S. pyrenaicus

S. torgalensis

S. aradensis

D-statistic (ABBA-BABA)



D-statistic indicates that the relationship cannot be described by a bifurcating tree

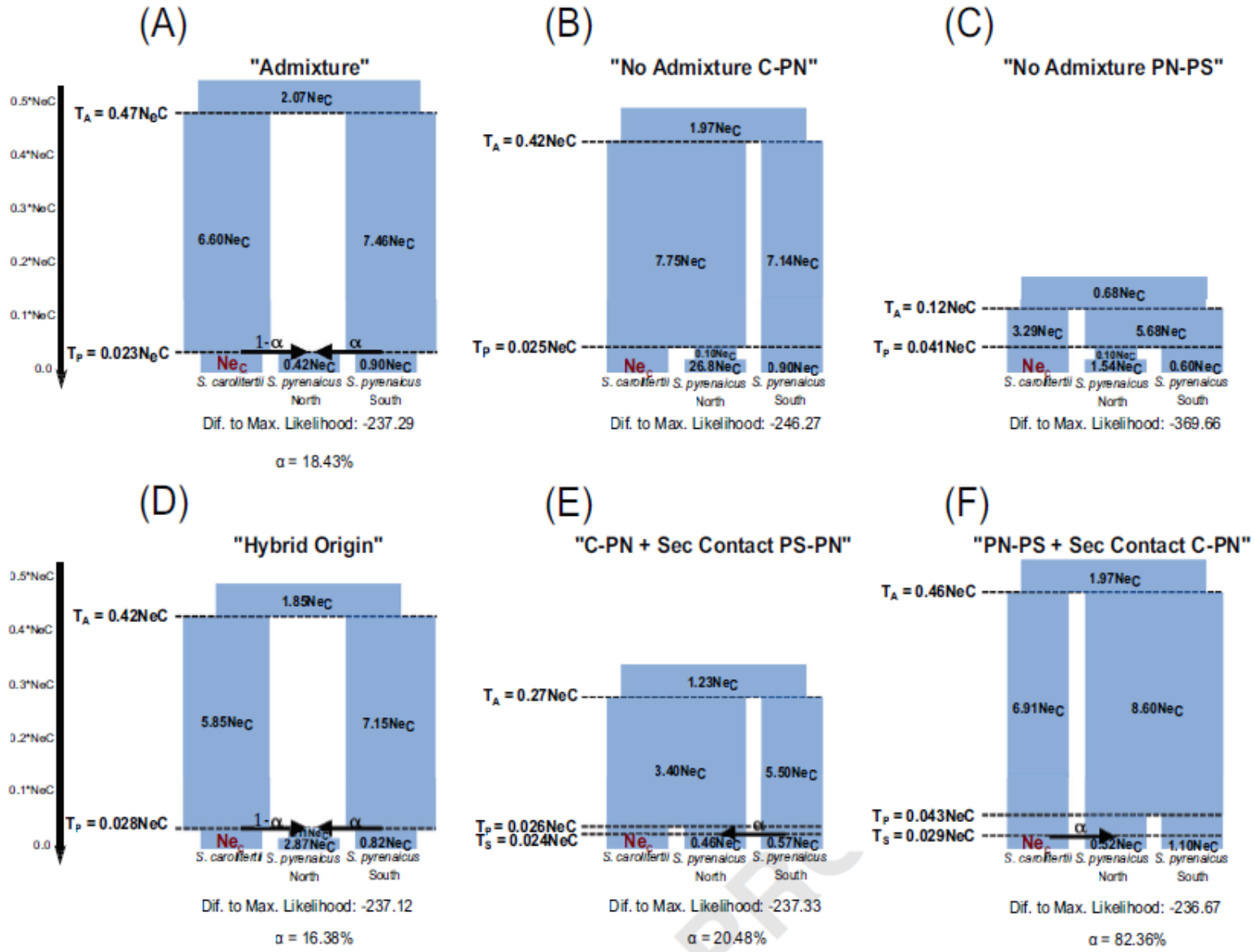
Mendes et al. (2021) Heredity

Pairwise 2D-folded SFS without monomorphic sites – inference based on relative parameters

- DP>10x
- Dowsampling 3 individuals from P1 (*S. carolitertii*), 4 individuals from H (*S. pyrenaicus North*), 3 individuals from P2 (*S. pyrenaicus* South)
- 8,758 SNPs – very small dataset!
- Folded SFS according to minor allele across the 3 populations
- Size of the three pairwise 2D-SFS: 175 entries

Mendes et al. (2021) Heredity

Relative parameter estimates



Mendes et al. (2021) Heredity

3D-SFS including linked sites

3D-SFS 1SNP per GBS locus

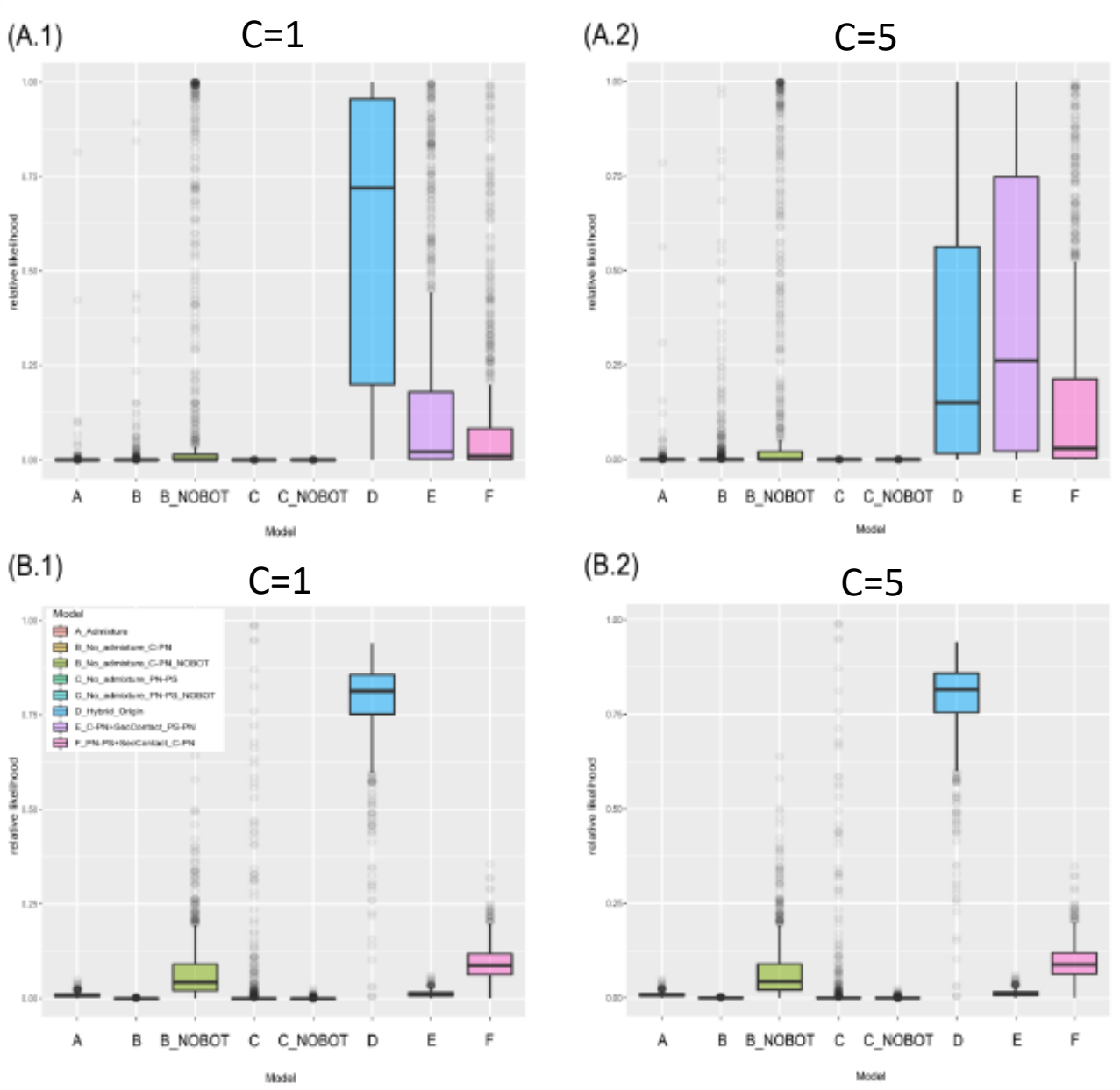


Figure S11– Relative likelihoods of demographic models based on AIC. Comparison of 8 models, including models B and C without bottlenecks, with 1000 bootstrap replicates. (A) Boxplots of relative likelihoods based on 3D-SFS with all SNPs per block with n=3 *S. carolitertii*, n=4 *S. pyrenaicus* North and n=3 *S. pyrenaicus* South individuals. Size of the joint 3D-SFS is SFSsize=441.

Protocol for model comparison based on AIC when we have independent SNPs

- Get the observed SFS
- Define the alternative models
- Perform 50-100 runs under each model
- Select the runs with maximum likelihood under each model
- Compute the AIC (Akaike information critera) for each model based on dataset with unlinked SNPs
- Select the model with minimum AIC

Demography and linked selection

Research article

Background selection as null hypothesis in population genomics: insights and challenges from *Drosophila* studies

Josep M. Comeron

Published: 06 November 2017 https://doi.org/10.1098/rstb.2016.0471

The Impact of Purifying and Background Selection on the Inference of Population History: Problems and Prospects

Parul Johri, Kellen Riall, Hannes Becher, Laurent Excoffier, Brian Charlesworth, Jeffrey D. Jensen

Molecular Biology and Evolution, Volume 38, Issue 7, July 2021, Pages 2986–3003, https://doi.org/10.1093/molbev/msab050

Published: 16 February 2021

Research Article

Genetics and Genomics

Background selection and biased gene conversion affect more than 95% of the human genome and bias demographic inferences

Fanny Pouyet, Simon Aeschbacher, Alexandre Thiéry, Laurent Excoffier

University of Bern, Switzerland; Swiss Institute of Bioinformatics, Switzerland; University of Zurich, Switzerland

Aug 20, 2018 · https://doi.org/10.7554/eLife.36317

BROWSE PUBLISH ABOUT

PLOS BIOLOGY

OPEN ACCESS

CONSENSUS VIEW

Recommendations for improving statistical inference in population genomics

Parul Johri, Charles F. Aquadro, Mark Beaumont, Brian Charlesworth, Laurent Excoffier, Adam Eyre-Walker, Peter D. Keightley, Michael Lynch, Gil McVean, Bret A. Payseur, Susanne P. Pfeifer, Wolfgang Stephan, Jeffrey D. Jensen

Published: May 31, 2022 • https://doi.org/10.1371/journal.pbio.3001669

Acknowledgements

Laurent Excoffier

FONDS NATIONAL SUISSE
SCHWEIZERISCHER NATIONALFONDS
FONDO NAZIONALE SVIZZERO
SWISS NATIONAL SCIENCE FOUNDATION

UNIVERSITÄT BERN

SIB Swiss Institute of Bioinformatics

Thank you!

Fundação para a Ciência e a Tecnologia

UID/BIA/00329/2020
PTDC/BIA-EVL/43457/2021

International Human Frontier Science Program Organization

RGY0081/2020

centre for ecology, evolution and environmental changes

Global Change and Sustainability Institute

Ciências ULisboa