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

Vitor Sousa

CE3C – center for ecology, evolution and environmental changes

2018 WSPG Cesky Krumlov
22 Jan 2020

vmsousa@fc.ul.pt

centre for ecology, evolution and environmental changes

Ciências ULisboa

Outline part II

Example of Applications:

- Human dispersal out of Africa (high quality whole-genome) – lessons on choice of models
- 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
- Inferring divergence times and gene flow in sawflies (ddRAD-seq data) – lessons from comparing models

Nourlangie, Kakadu National Park, NT, Australia

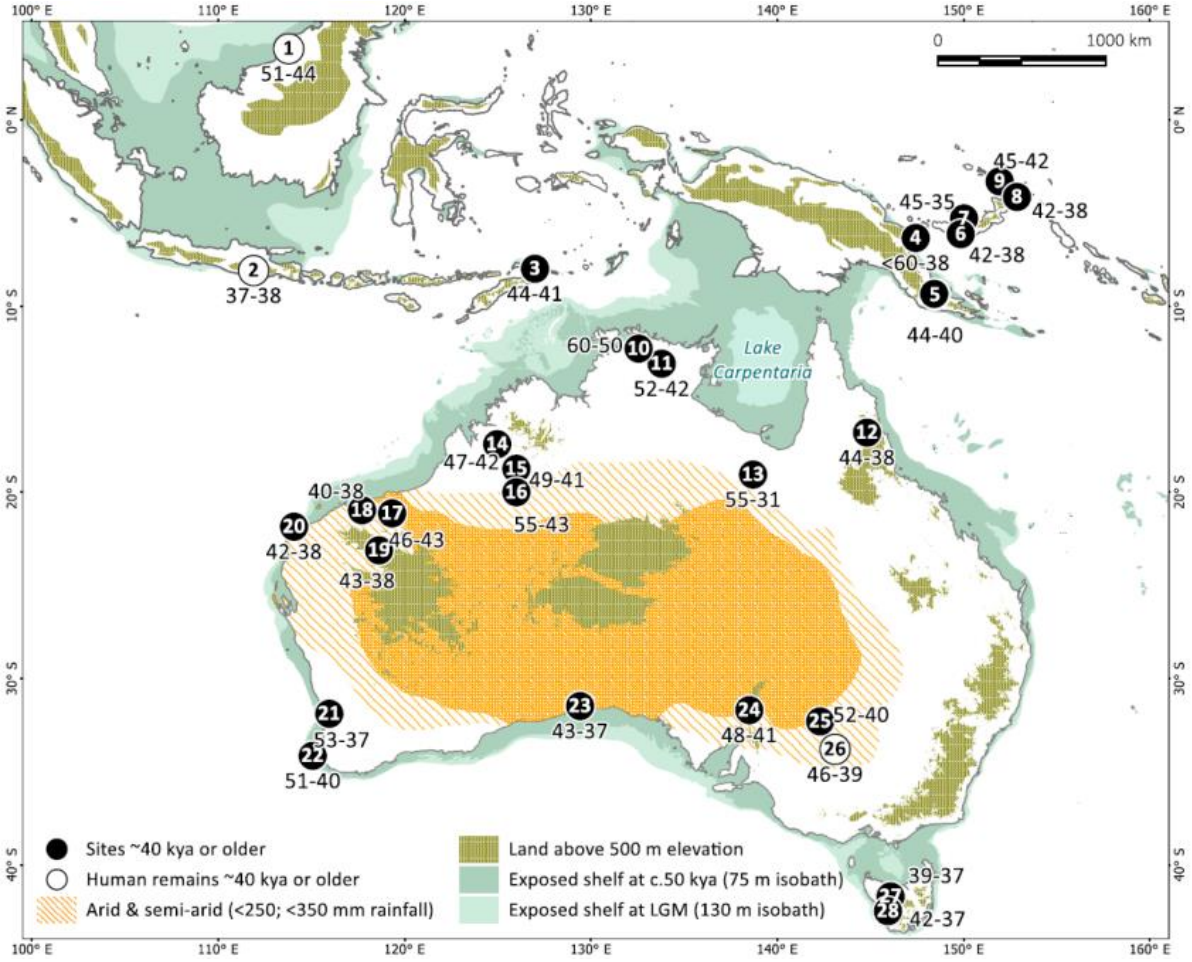
A genomic history of Aboriginal Australia

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)

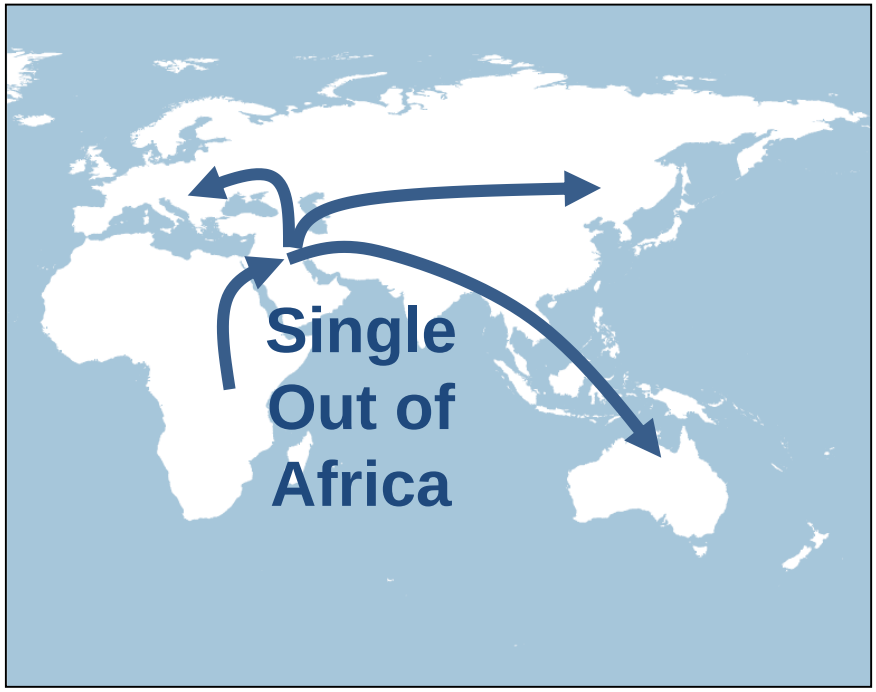
Ewaninga Rock Carvings Conservation Reserve, NT, Australia

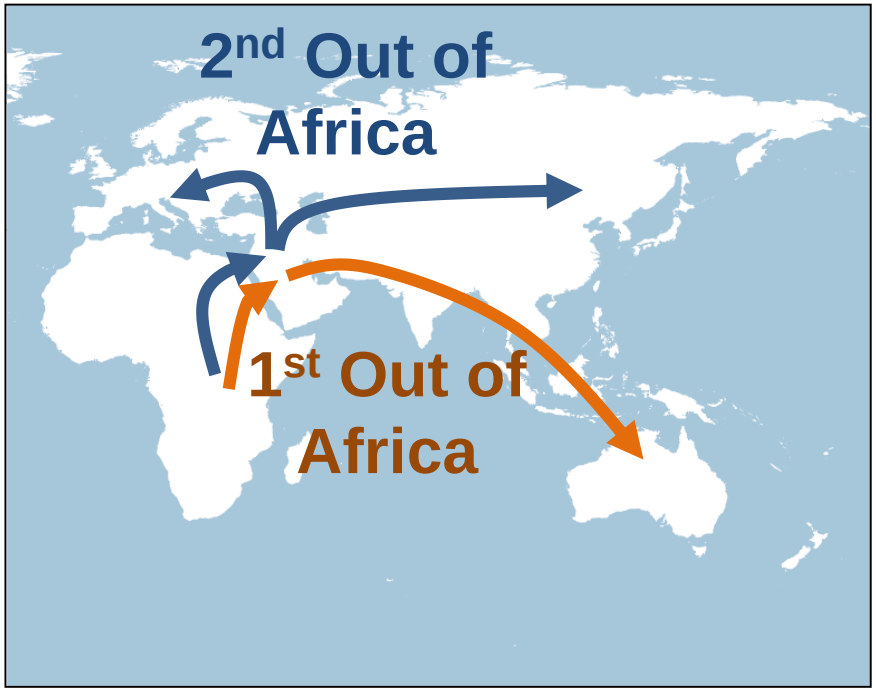
Australia harbors some of the oldest modern human remains outside Africa



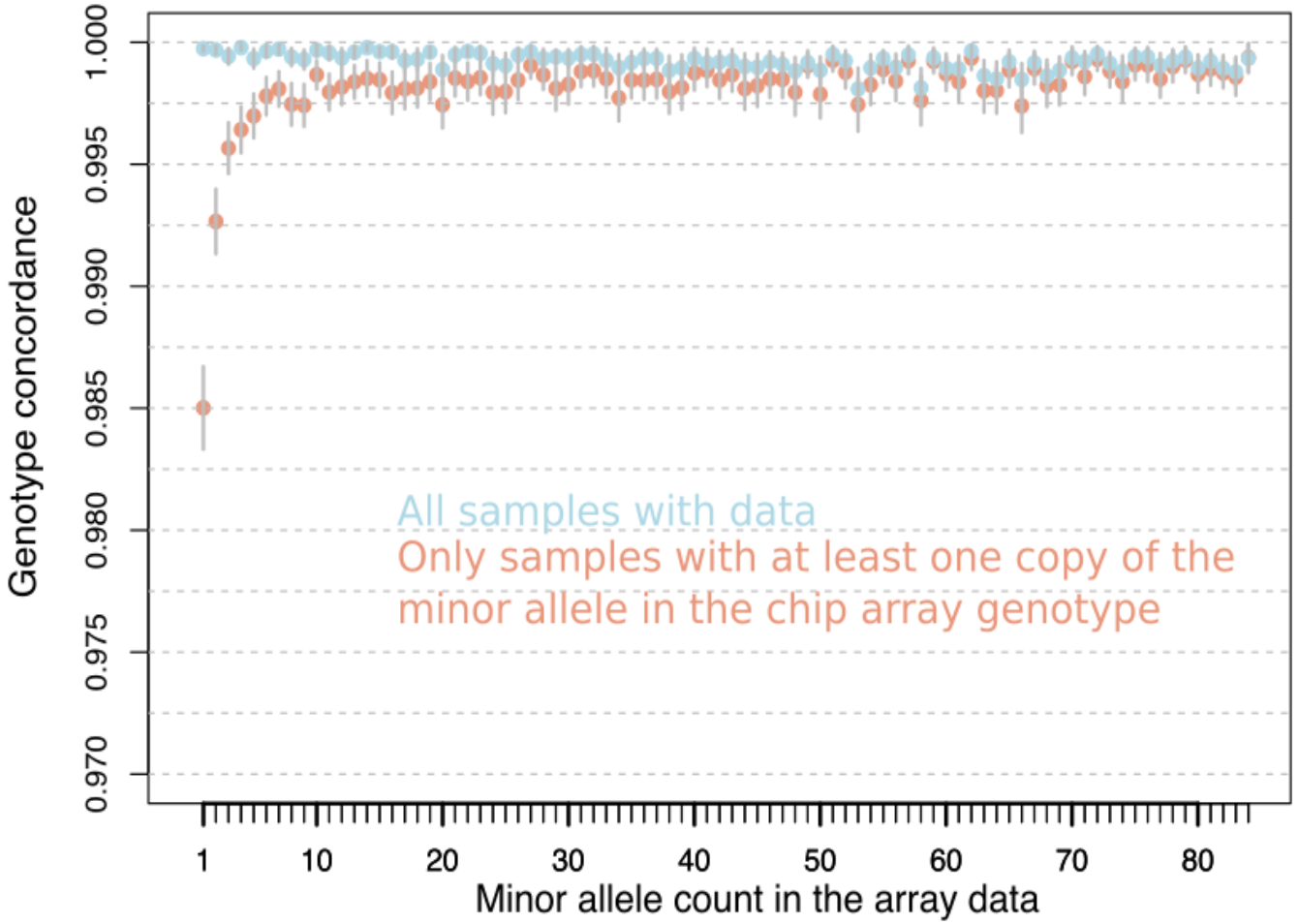
Many sites and remains dated to be older than 40 kya, suggesting a human settlement 47.5-55 kya

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





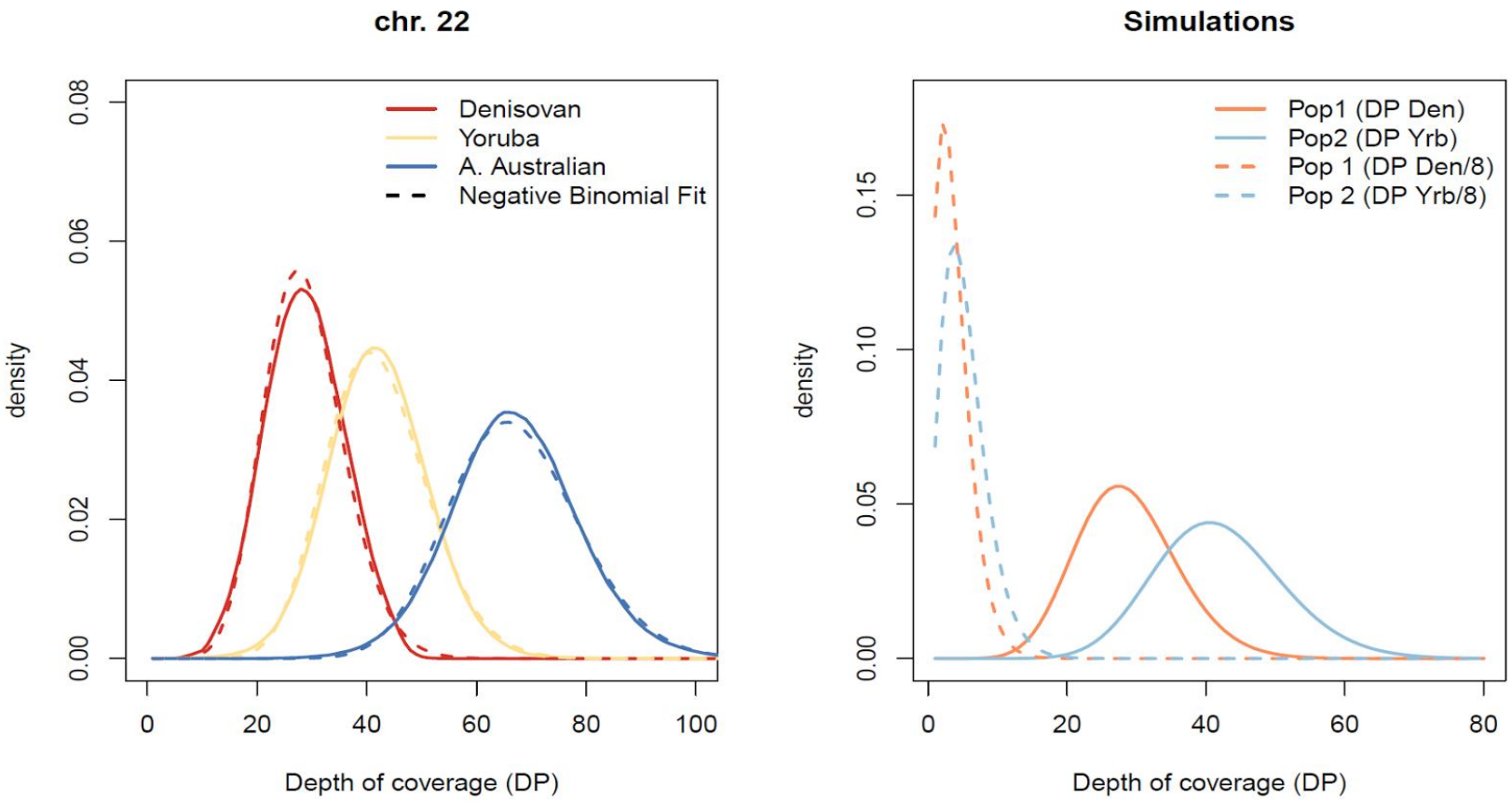
83 high-coverage Aboriginal Australians genomes



Average depth of coverage: 65x

Very good quality of genotype calls

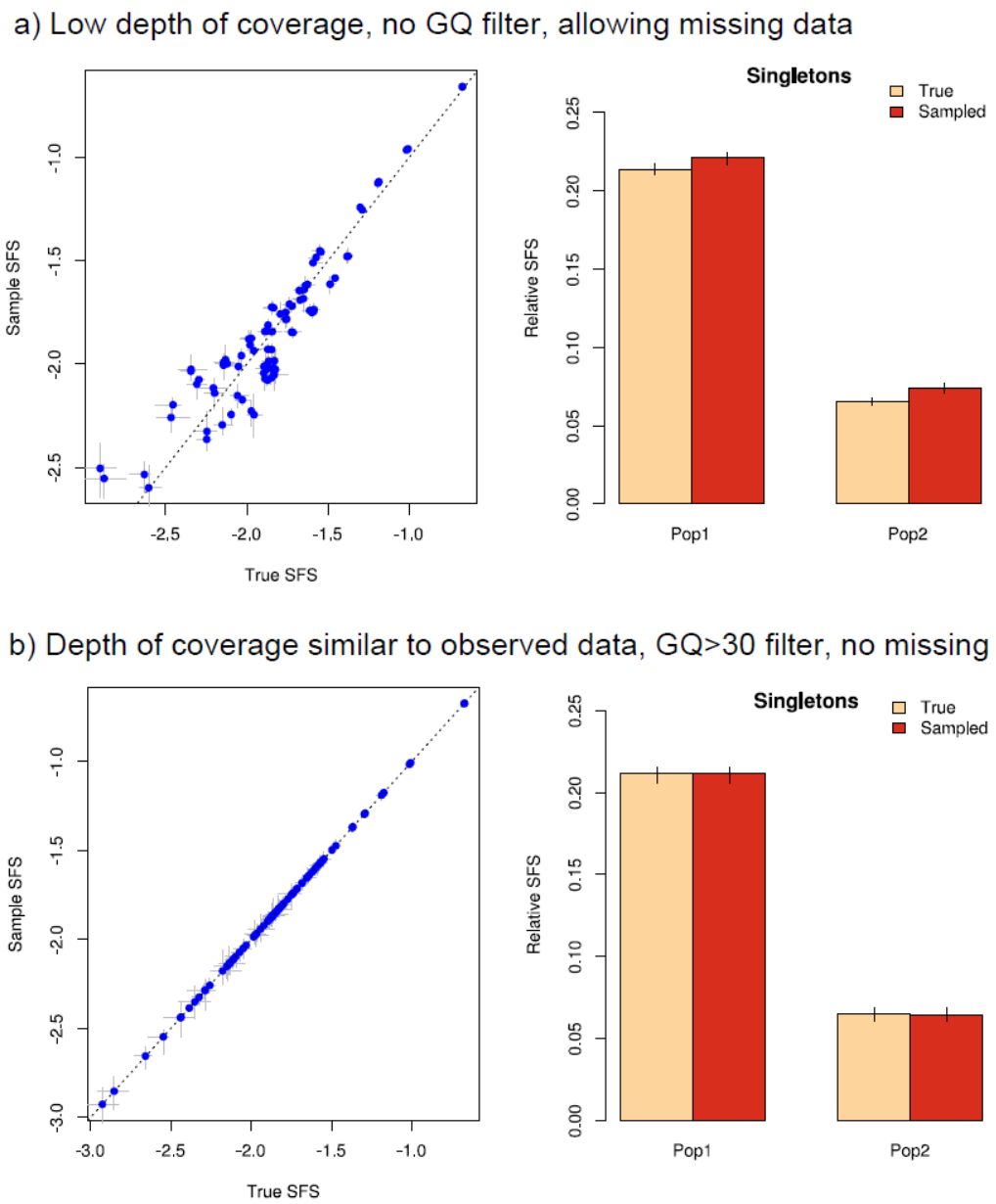
Effect of depth of coverage on SFS



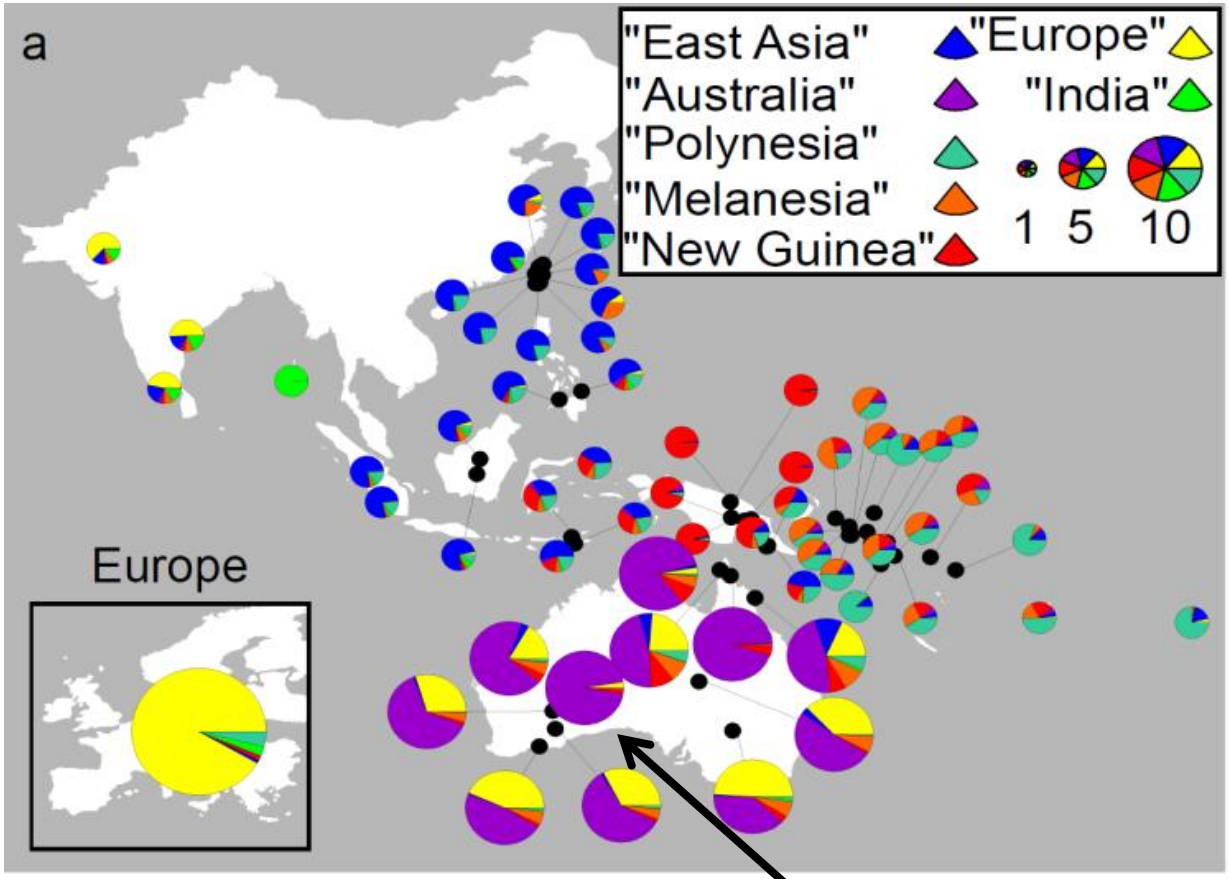
- Compared 2D SFS based on depth of coverage of observed data (mean larger than >20x), with a distribution 8 times smaller.

A note on recovering the SFS from genomic data

- Simulation study
- Low depth of coverage and missing data lead to biased SFS towards rare variants



83 high-coverage Aboriginal Australians genomes



Average depth of coverage: 65x

★ 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)

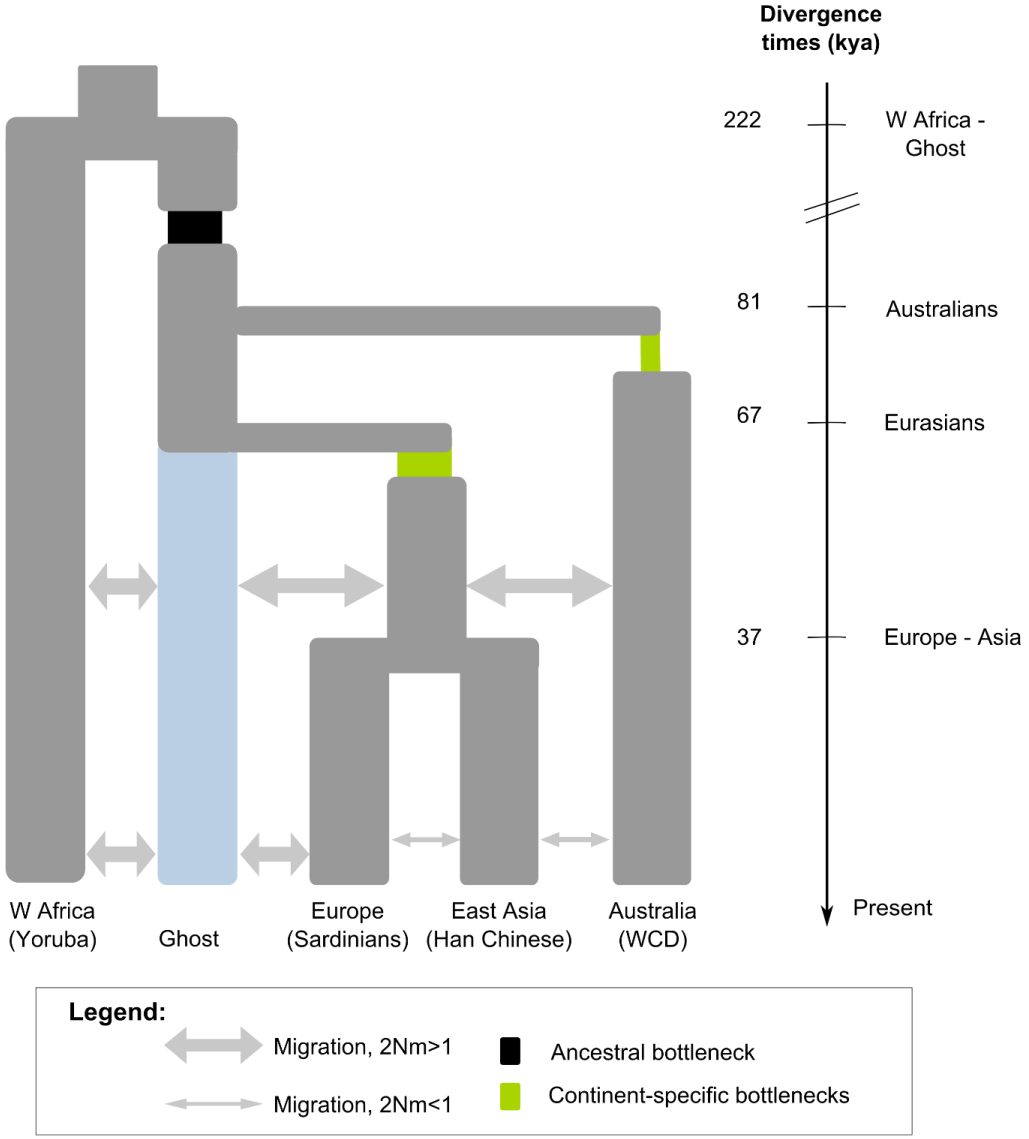
Do you have monomorphic sites?

- **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)

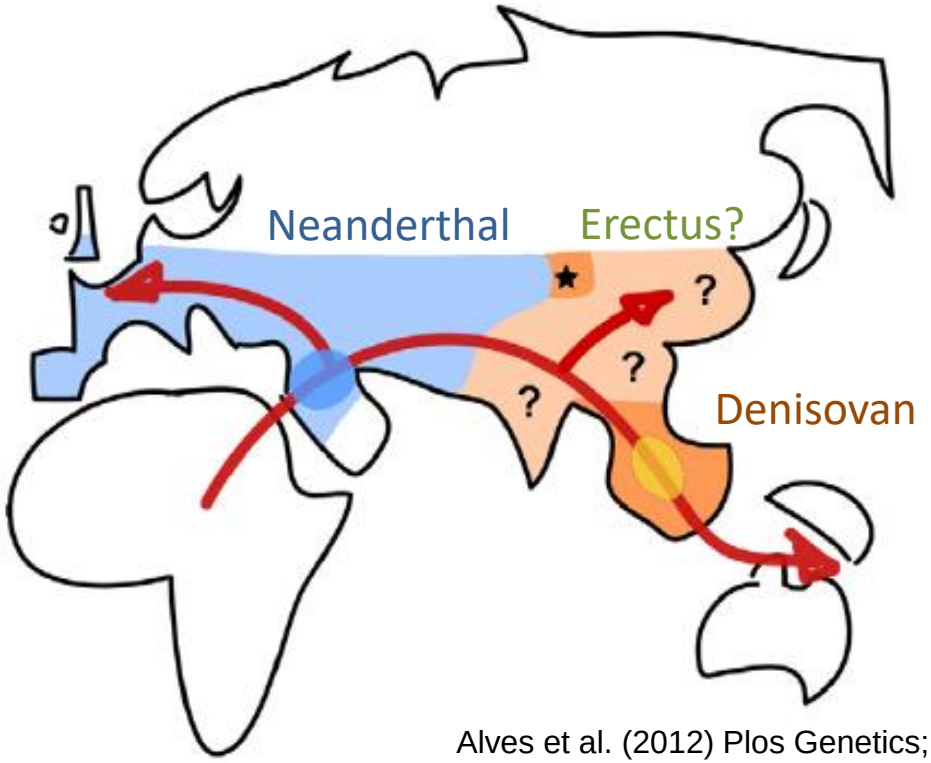
We always get results…

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)



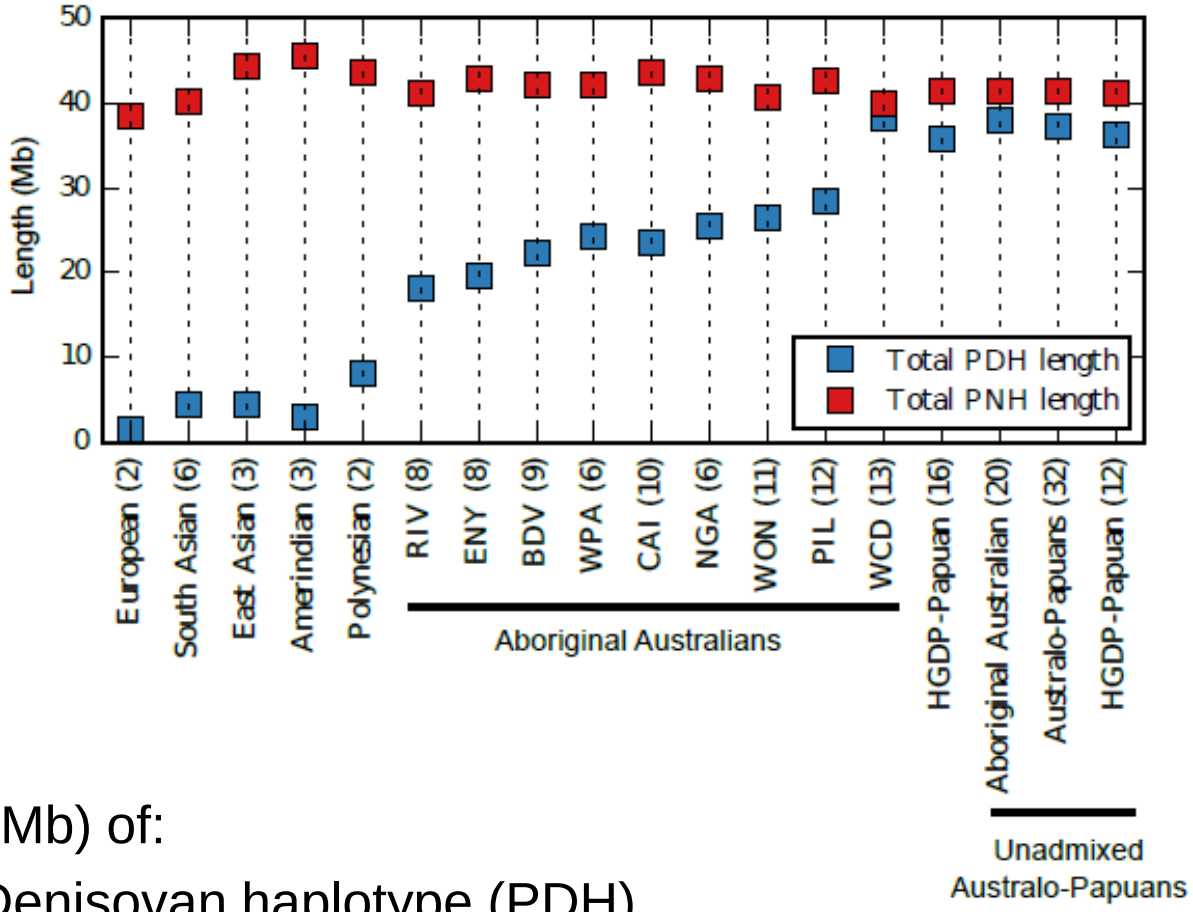
Towards a model incorporating Neanderthal and Denisovan admixture



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

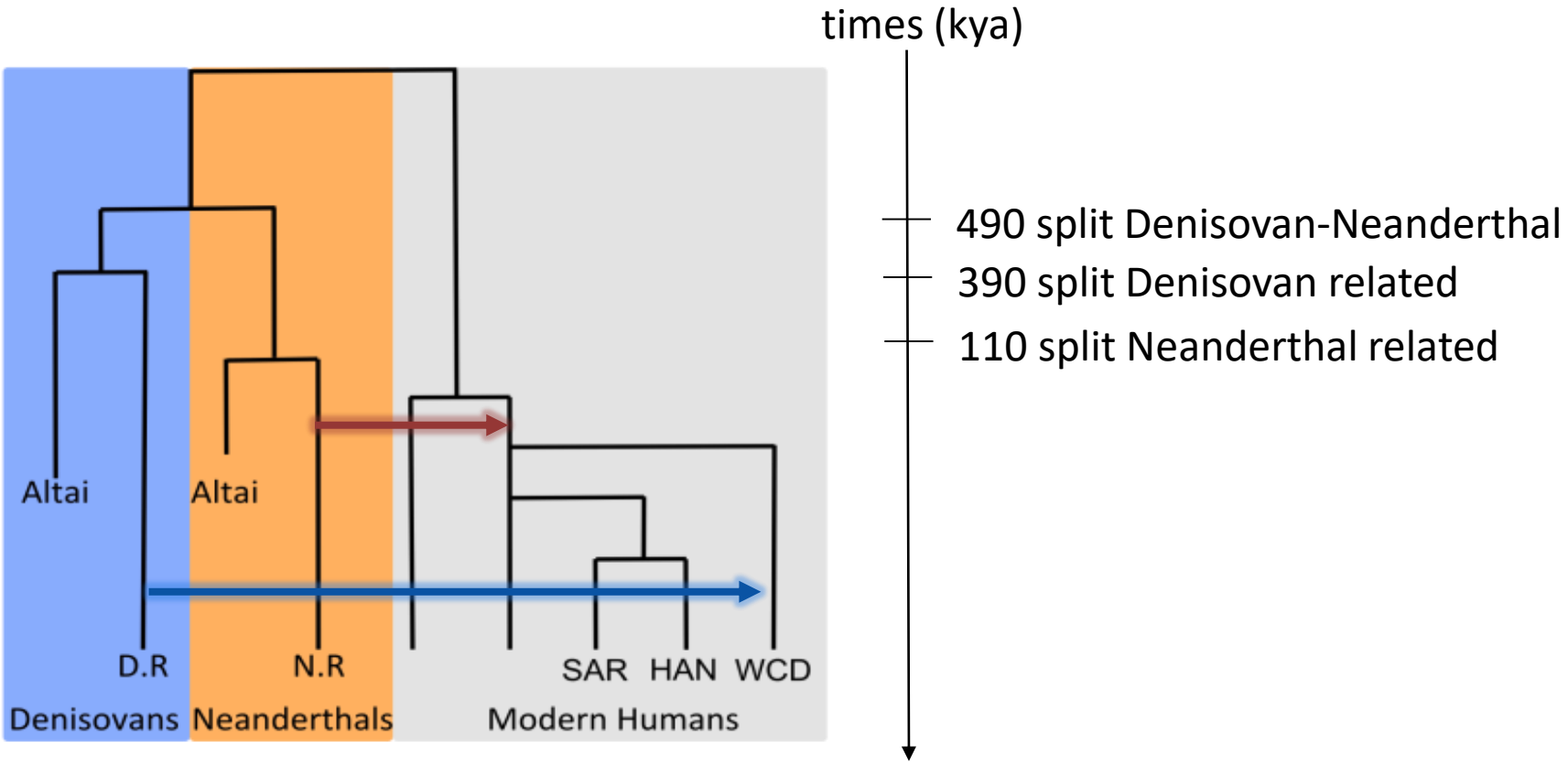
Evidence of archaic introgression



Total length (Mb) of:

- Putative Denisovan haplotype (PDH)
- Putative Neanderthal haplotypes (PNH)

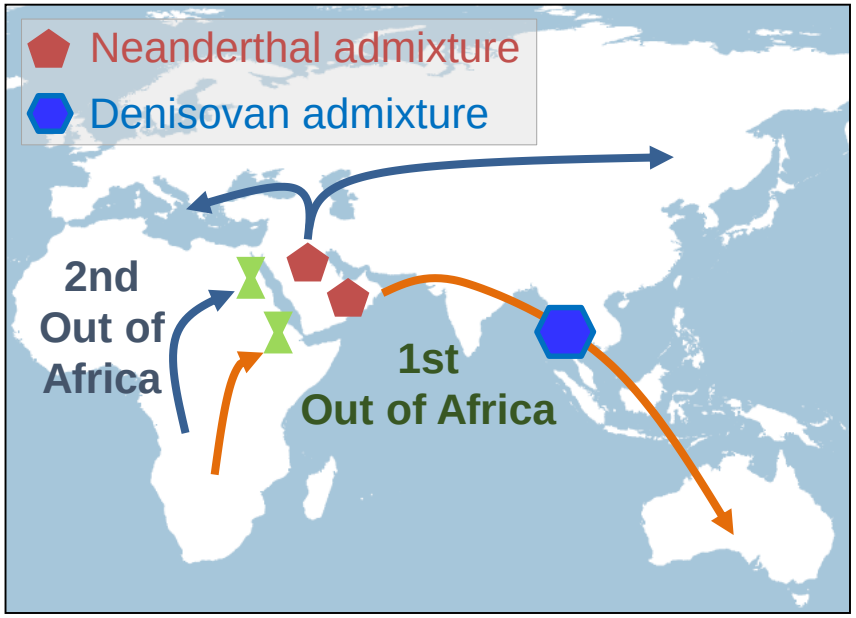
Accounting for shared ancestry of Neanderthal and Denisovan

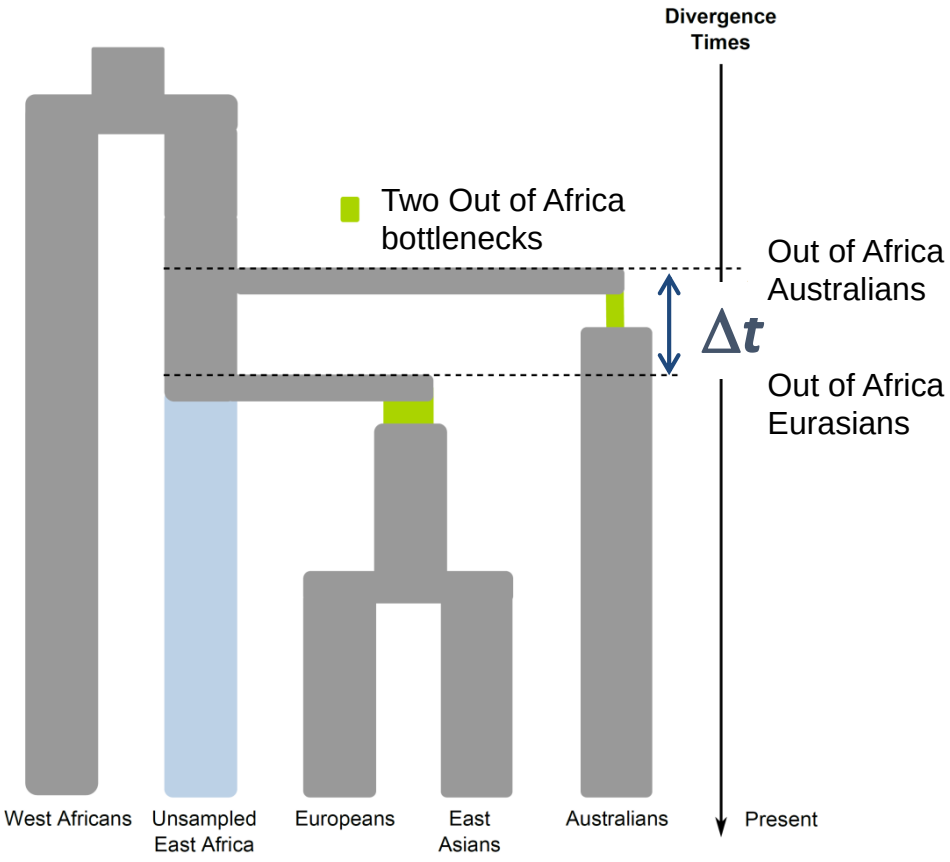


Admixture occurs between modern humans and:

- Denisovan-related (D.R.) population
- Neanderthal-related (N.R.) population

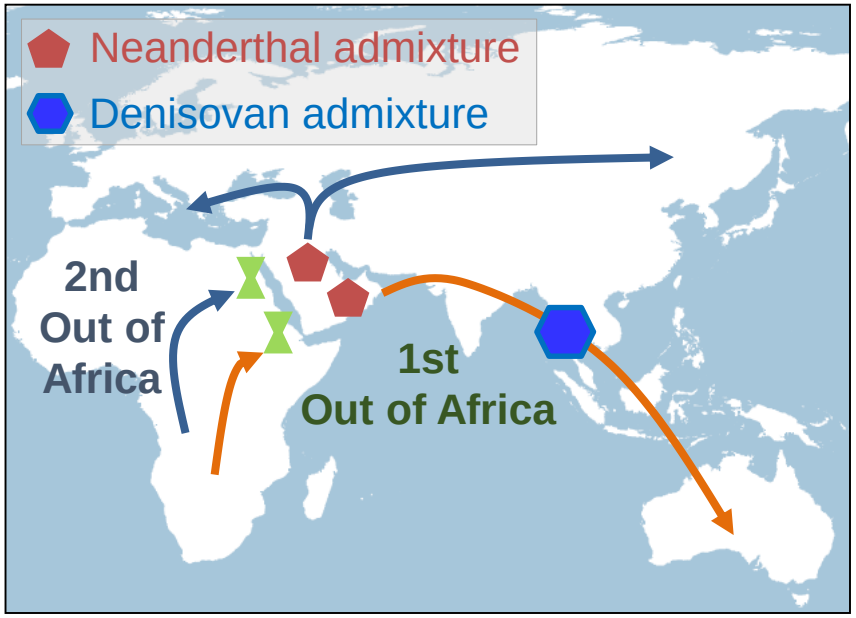
Two-waves out of Africa

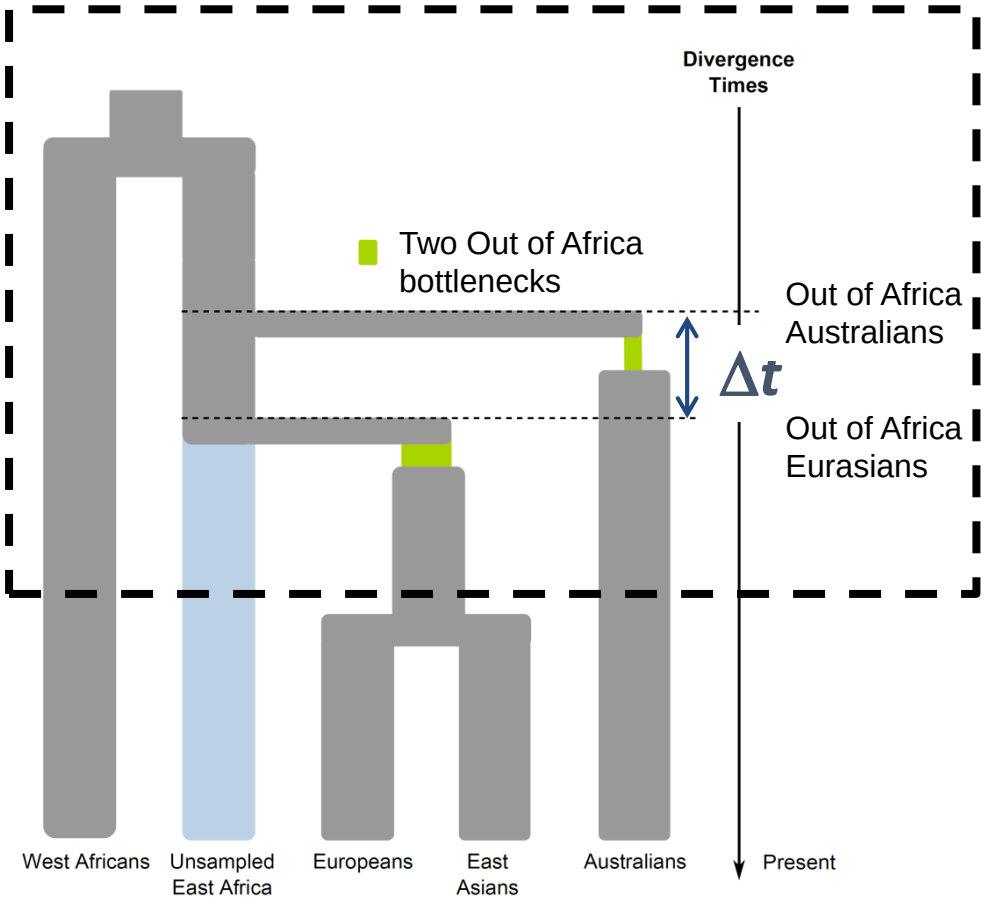




- Two different divergence times ($\Delta t >> 0$)
- Two independent bottlenecks associated with the two Out of Africa events

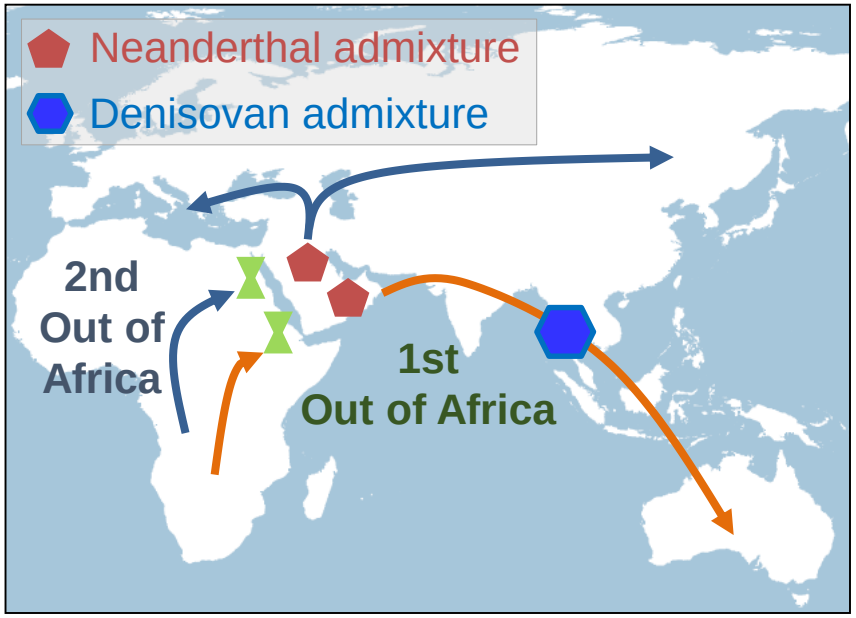
Two-waves out of Africa

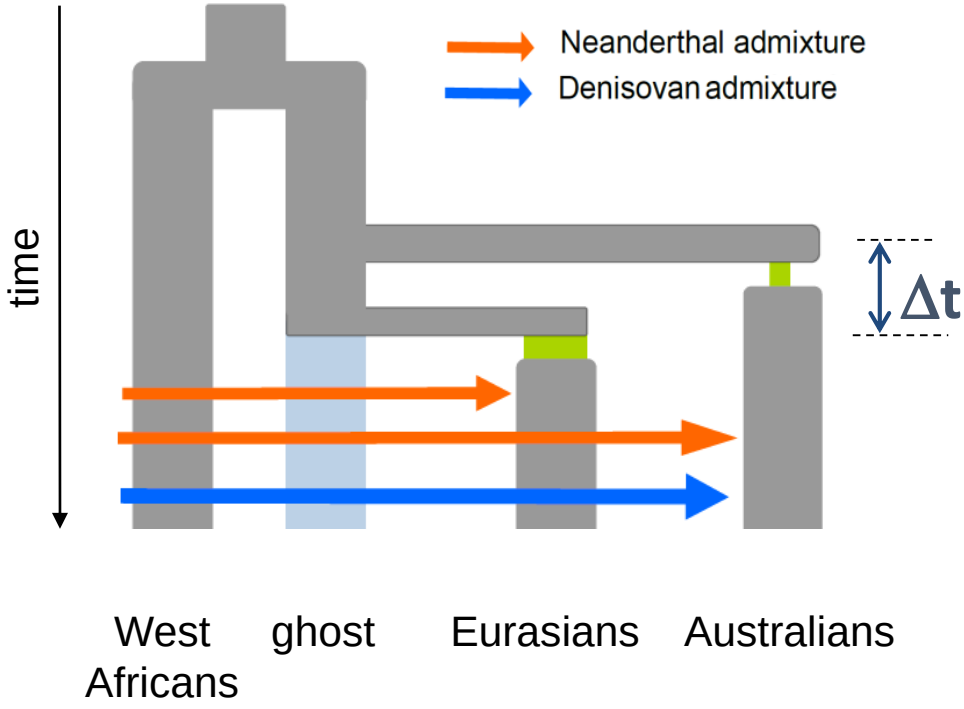




- Two different divergence times ($\Delta t >> 0$)
- Two independent bottlenecks associated with the two Out of Africa events

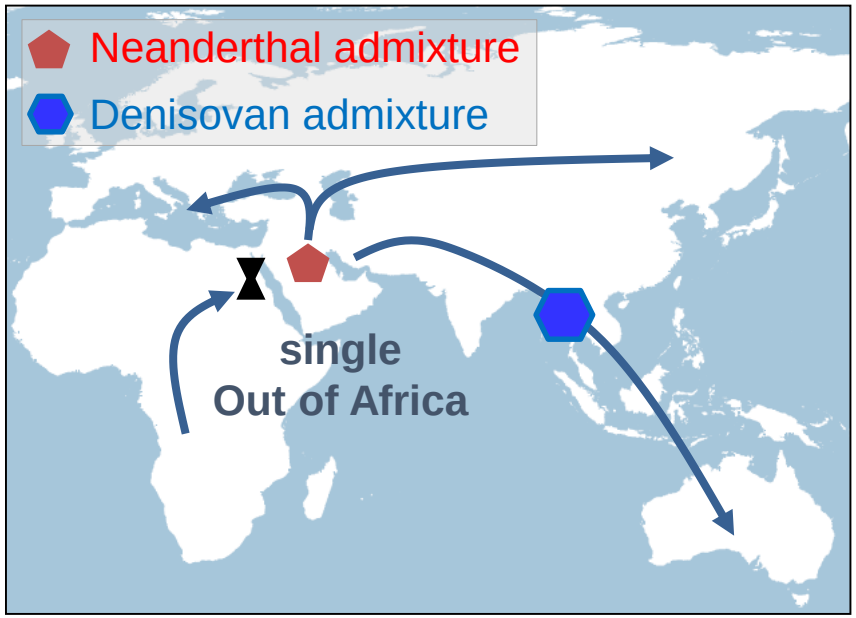
Two-waves out of Africa

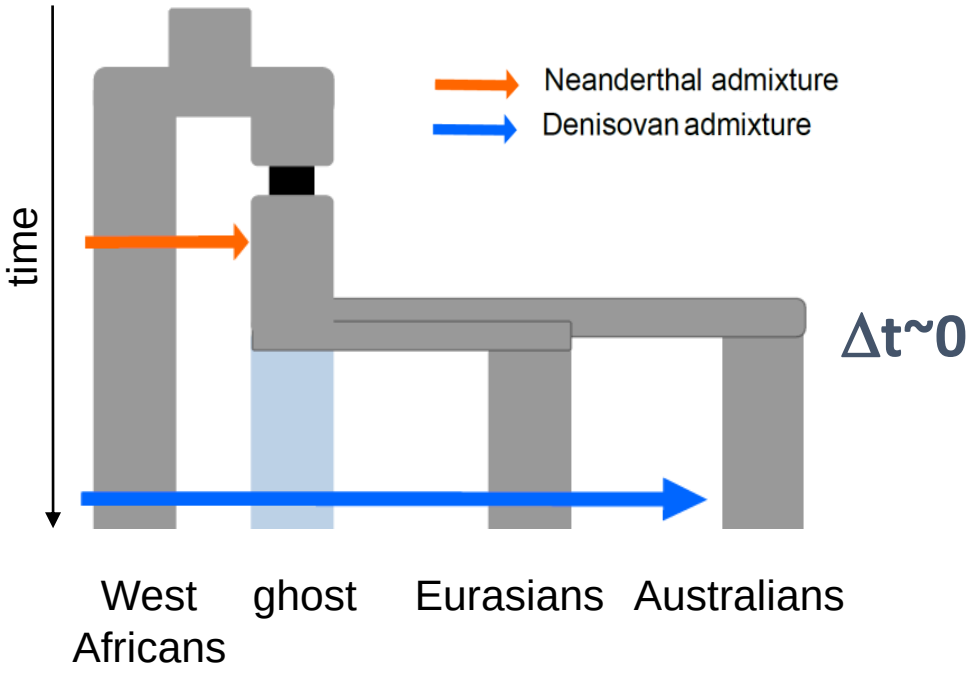




- Two different divergence times ($\Delta t >> 0$)
- Two independent bottlenecks associated with the two Out of Africa events

One wave out of Africa

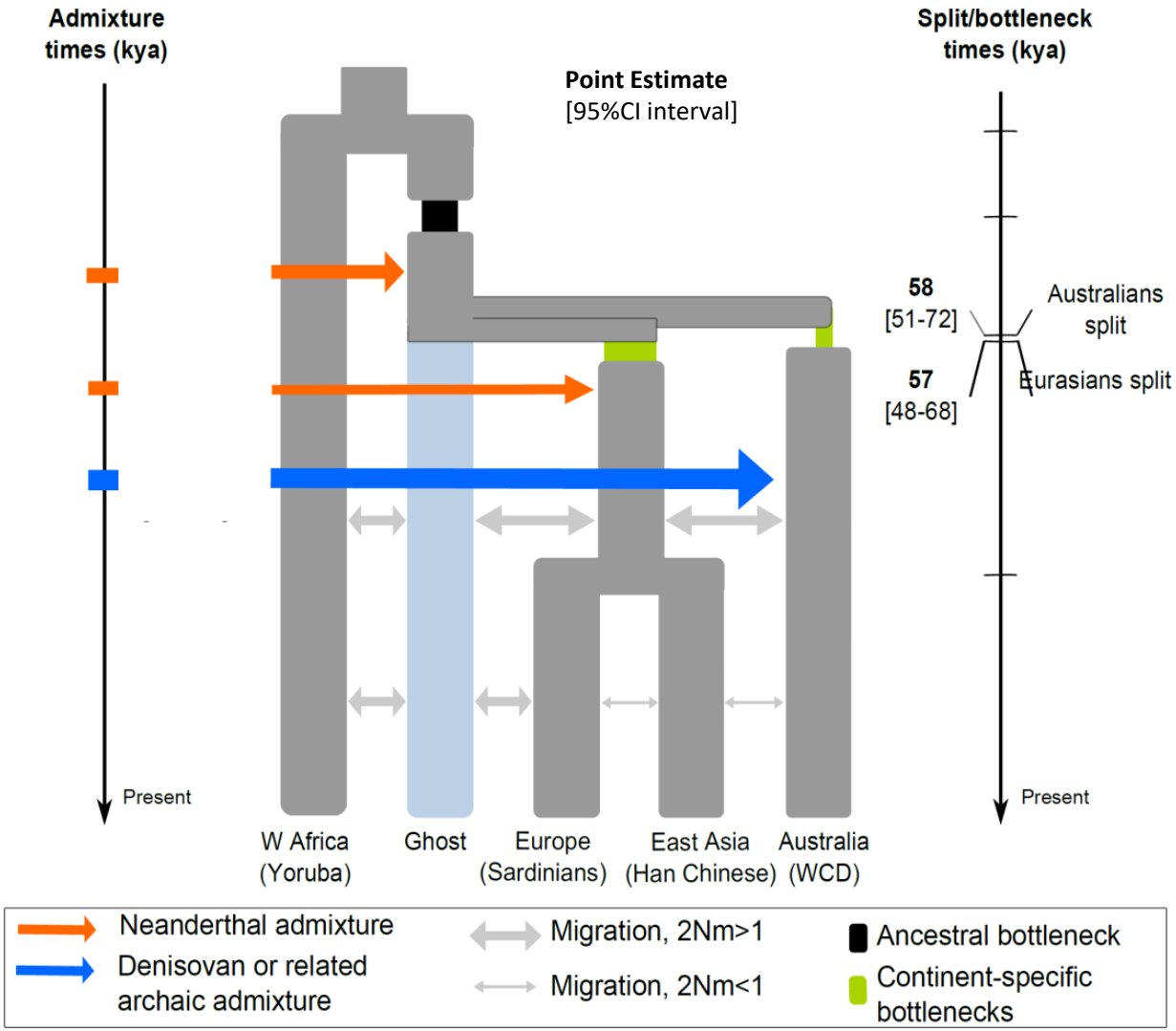




- Similar divergence times (Δt close to zero)
- One single bottlenecks associated with the Out of Africa events
- A major admixture pulse with Neanderthal

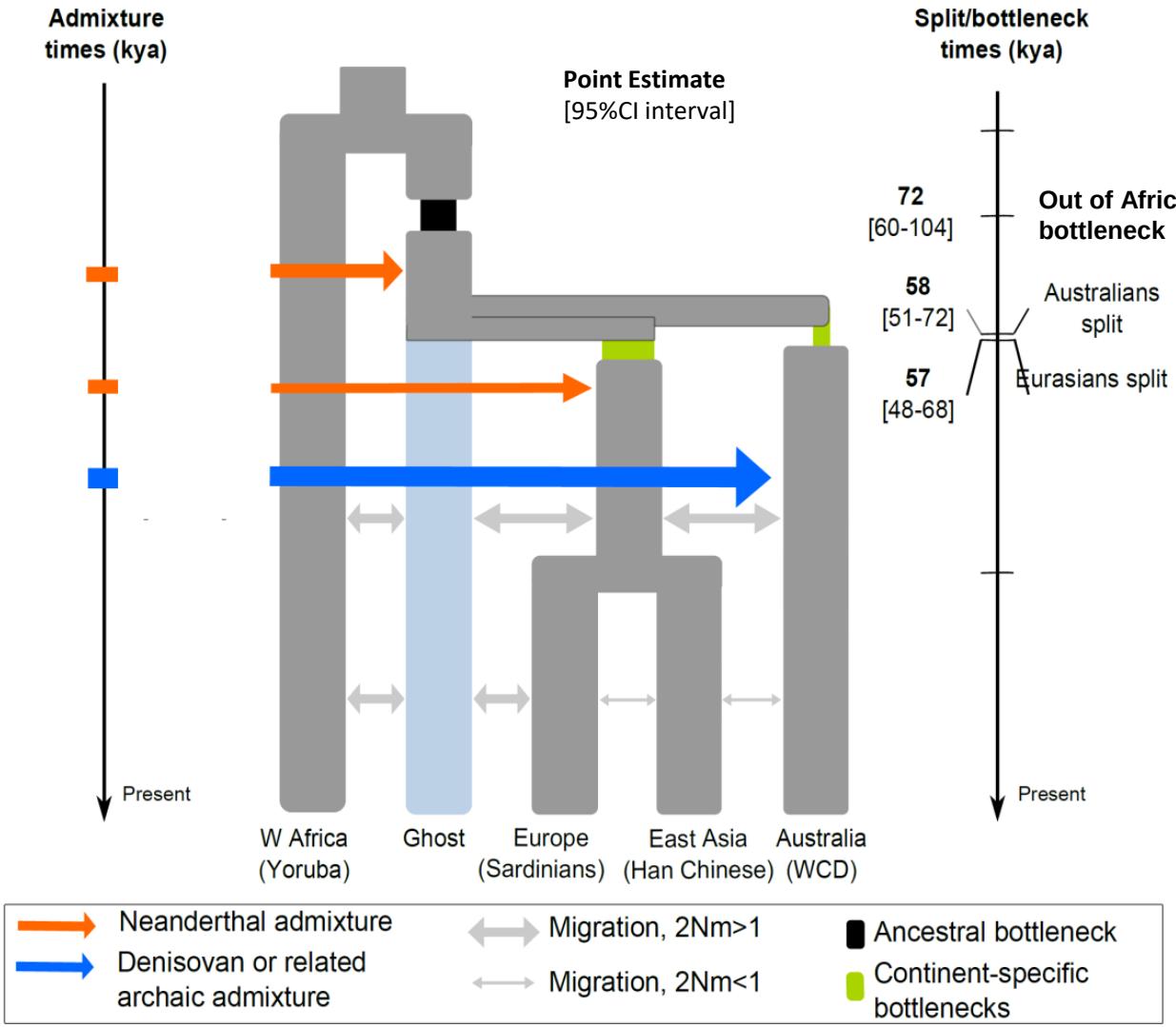
A single wave Out of Africa is consistent with our estimates when accounting for archaic admixture

- Similar divergence time (Δt close to zero)



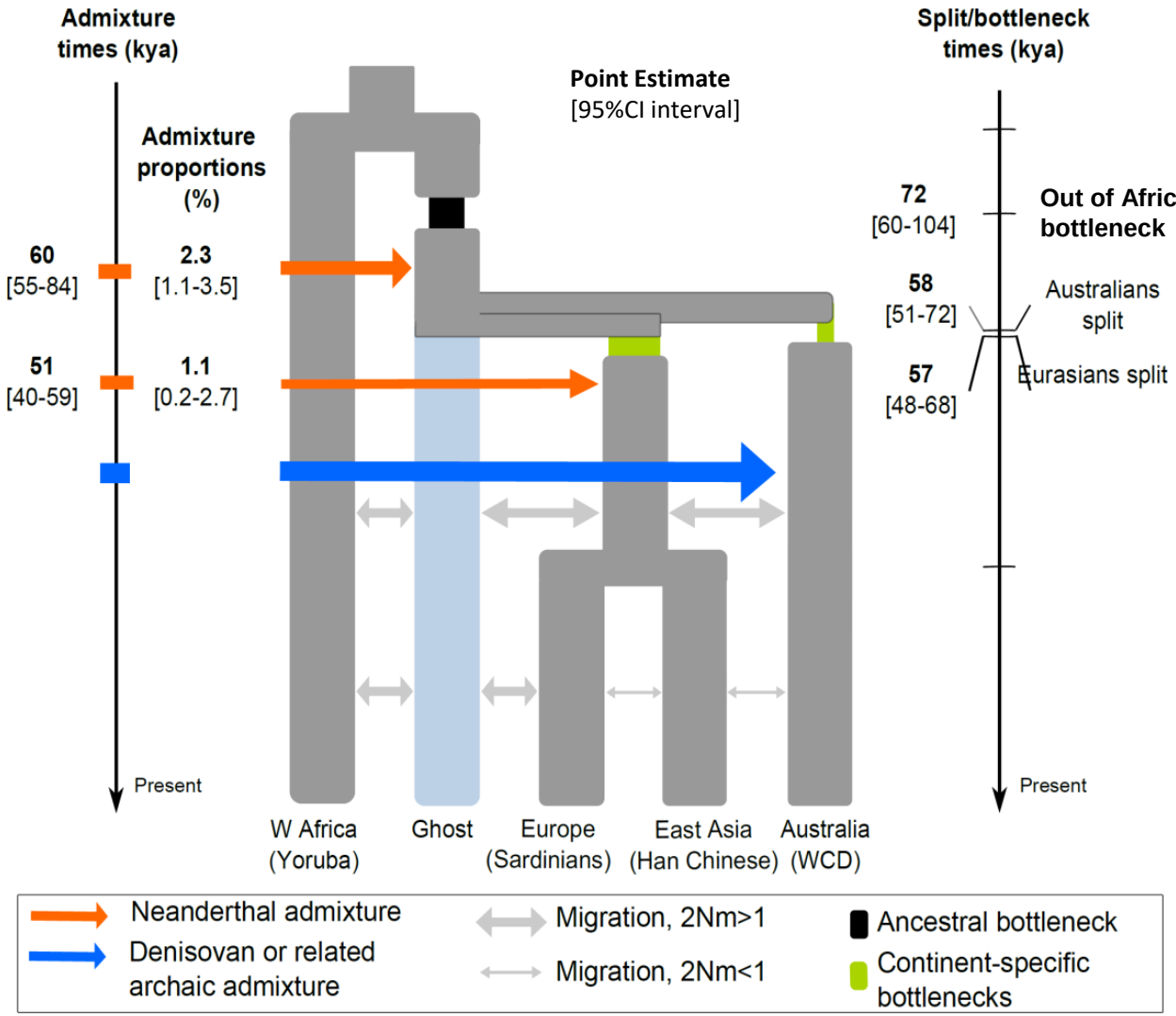
A single wave Out of Africa is consistent with our estimates when accounting for archaic admixture

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



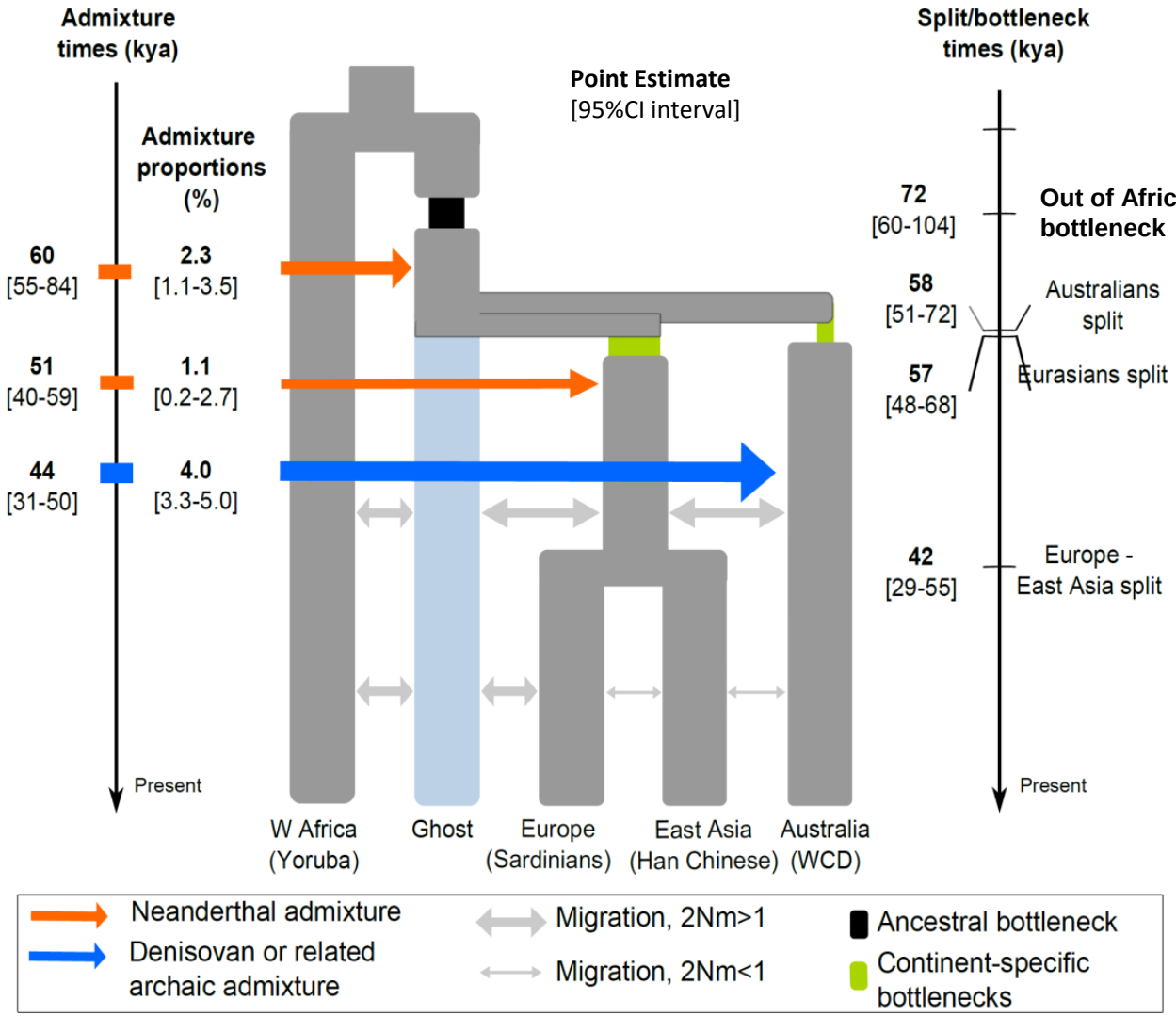
A single wave Out of Africa is consistent with our estimates when accounting for archaic admixture

- 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

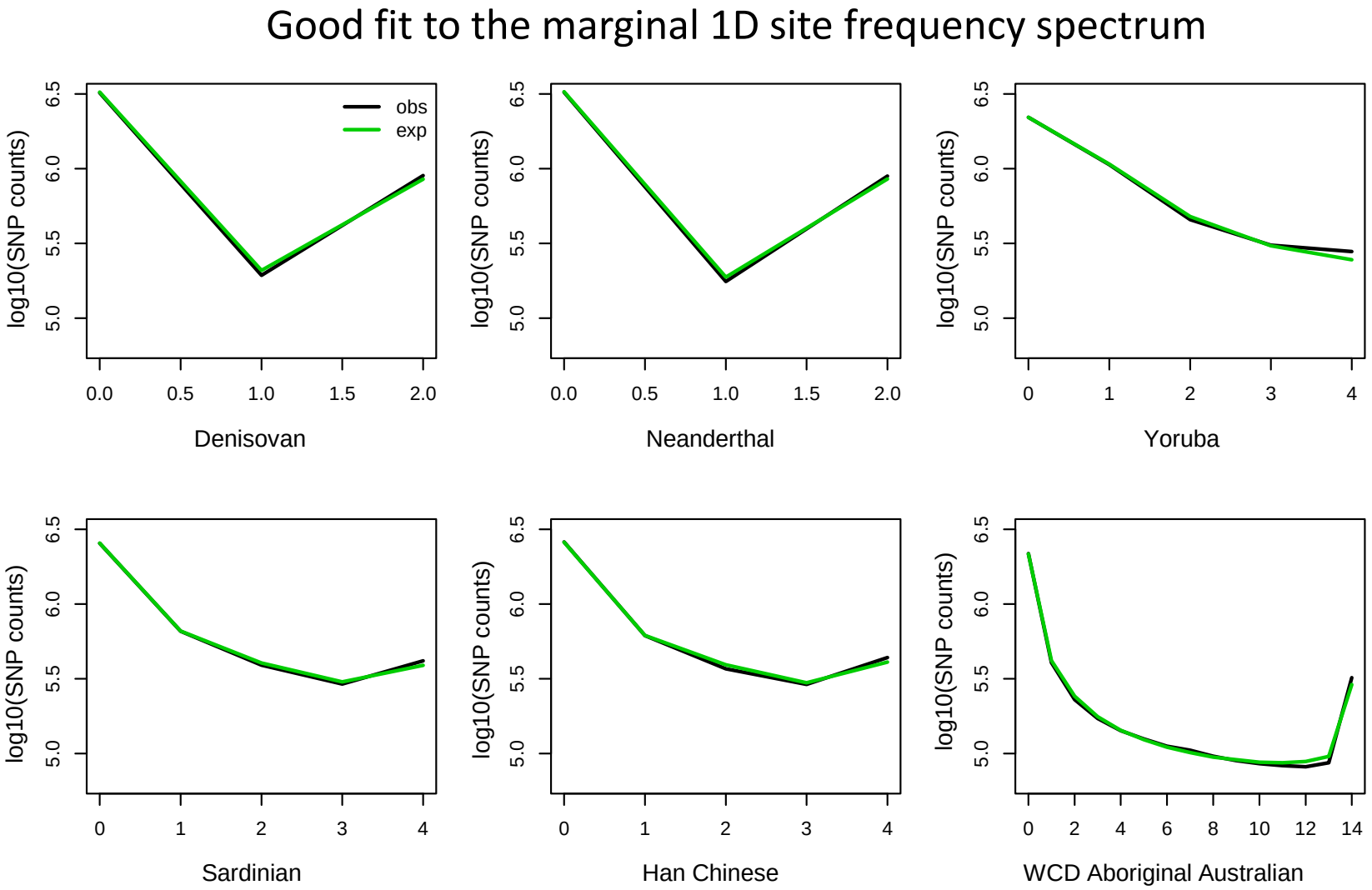


A single wave Out of Africa is consistent with our estimates when accounting for archaic admixture

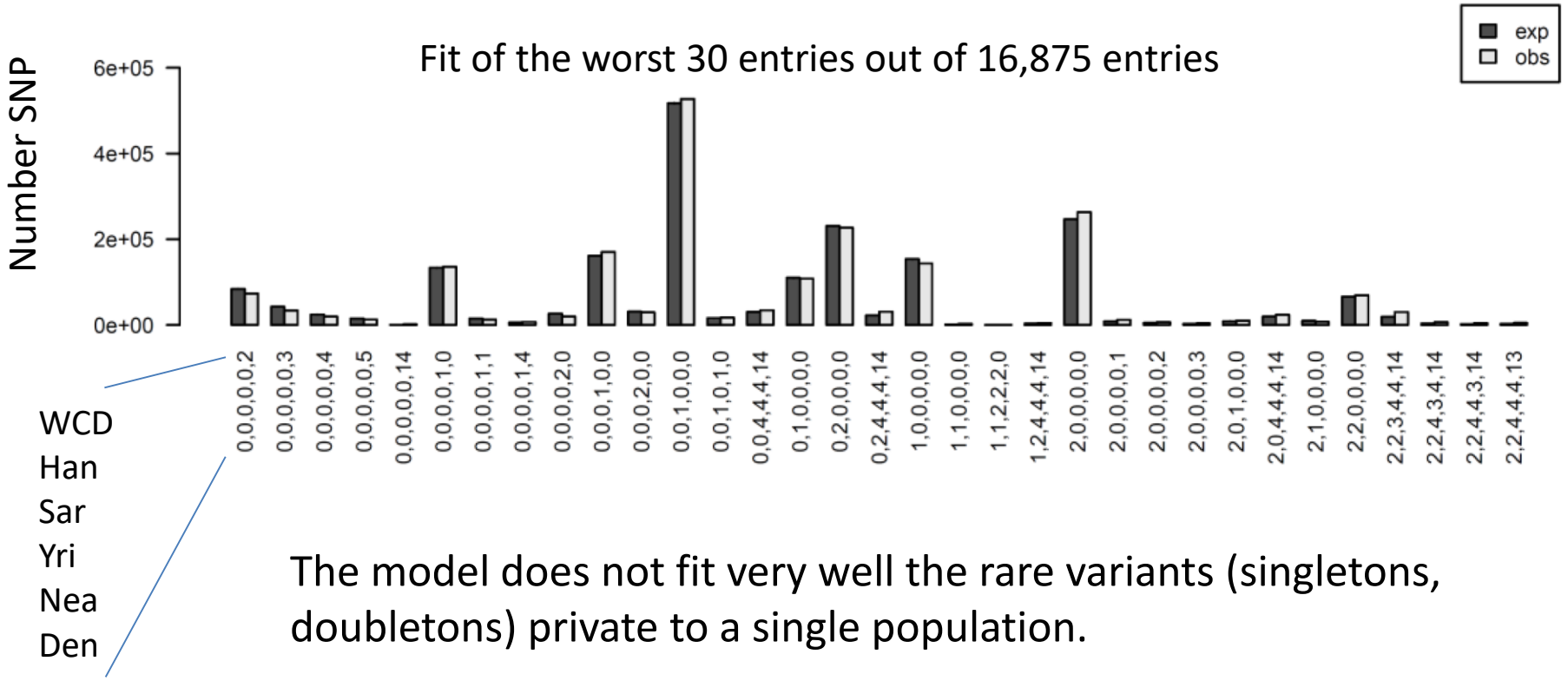
- Similar divergence time (Δt close to zero)
- Bottleneck associated with the Out of Africa event
- A major admixture pulse with Neanderthal in ancestors of all non-Africans



Model captures aspects about the observed data



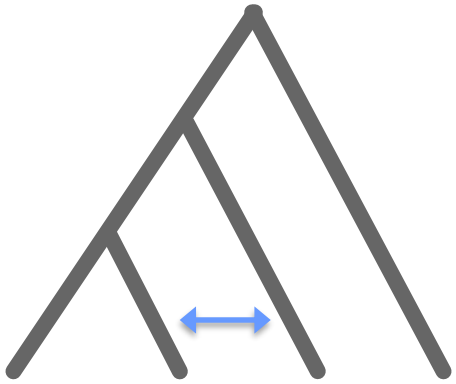
What entries are not well fitted?



The model does not fit very well the rare variants (singletons, doubletons) private to a single population.

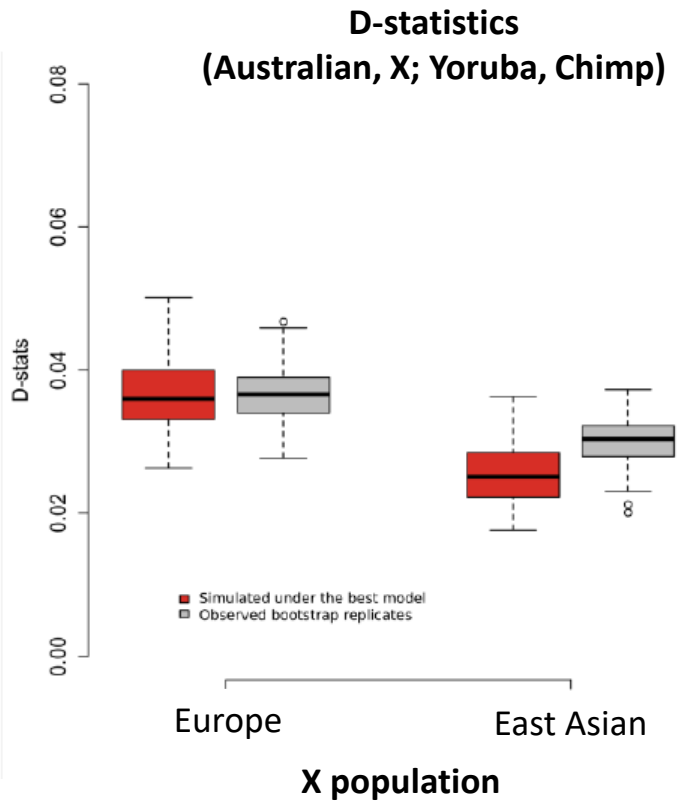
Pagani et al (2016) suggests two waves: Papuan genomes with signature of admixture with humans from first wave (at least 2% of their genome).

Model captures the higher derived allele sharing between Eurasians and Yoruba



Australia Europe or East Asian Yoruba Chimp

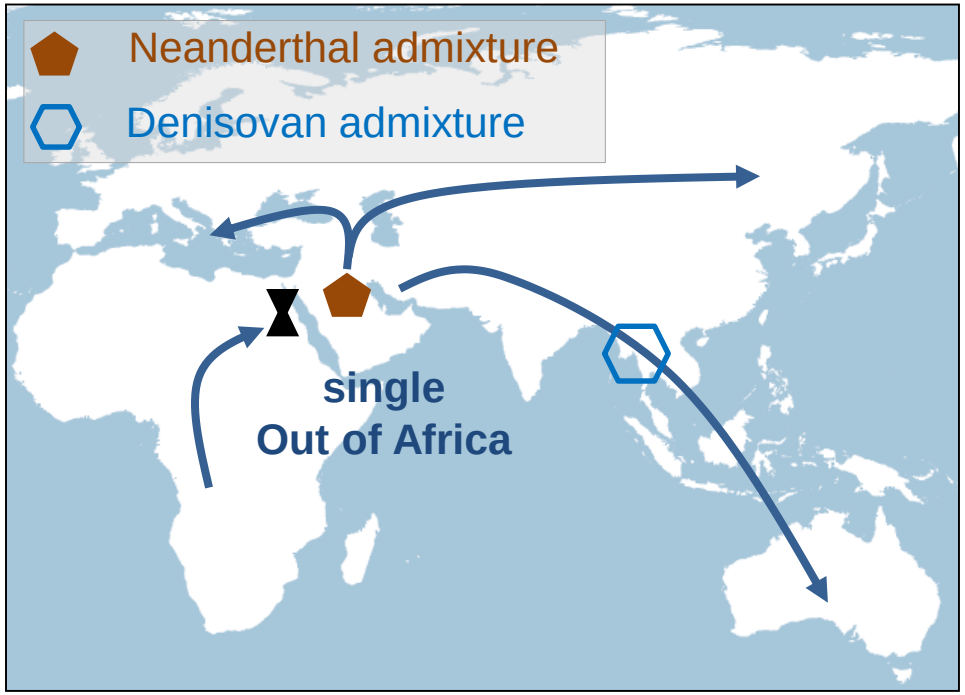
D-statistics suggest that Yoruba and Eurasians share more derived alleles than Yoruba and Australians



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)
- Two major dispersal waves into Asia: Aboriginal Australians diverged 51-72 kya from Eurasians



ARTICLE

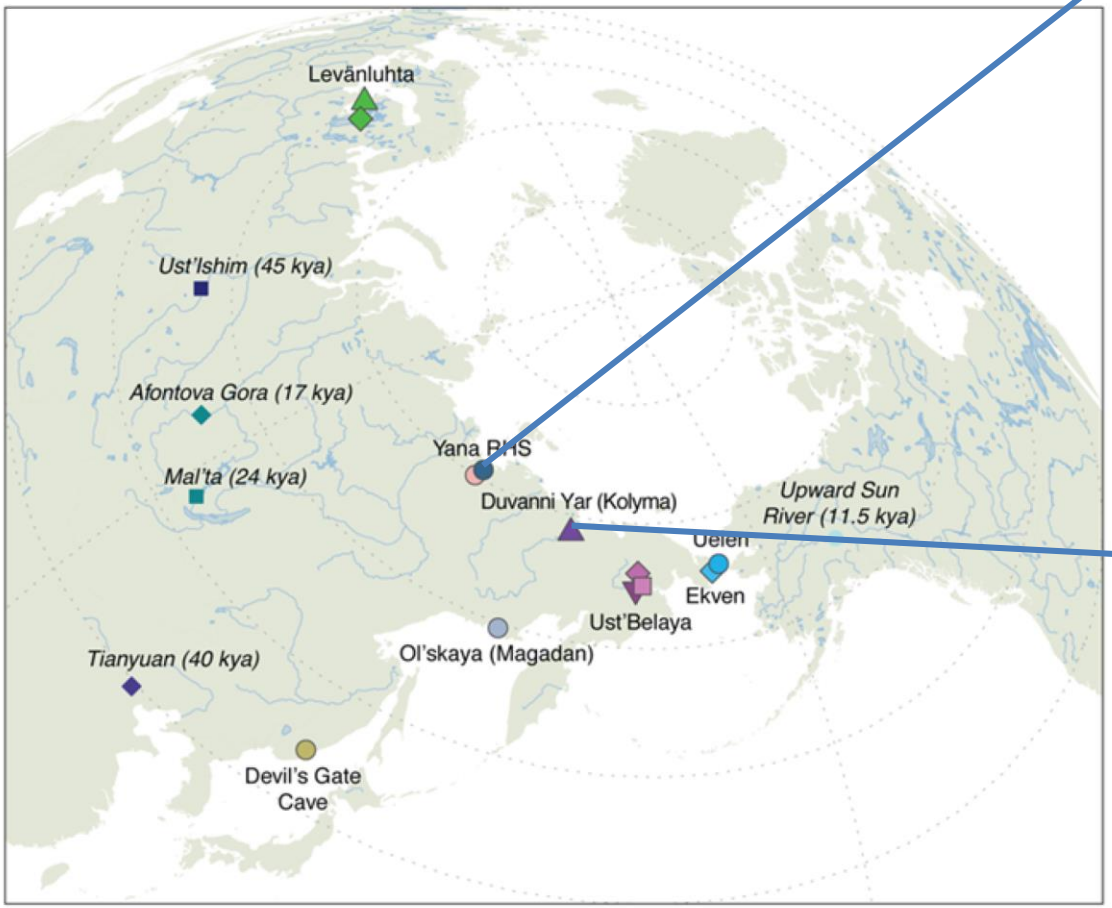
https://doi.org/10.1038/s41586-019-1279-z

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

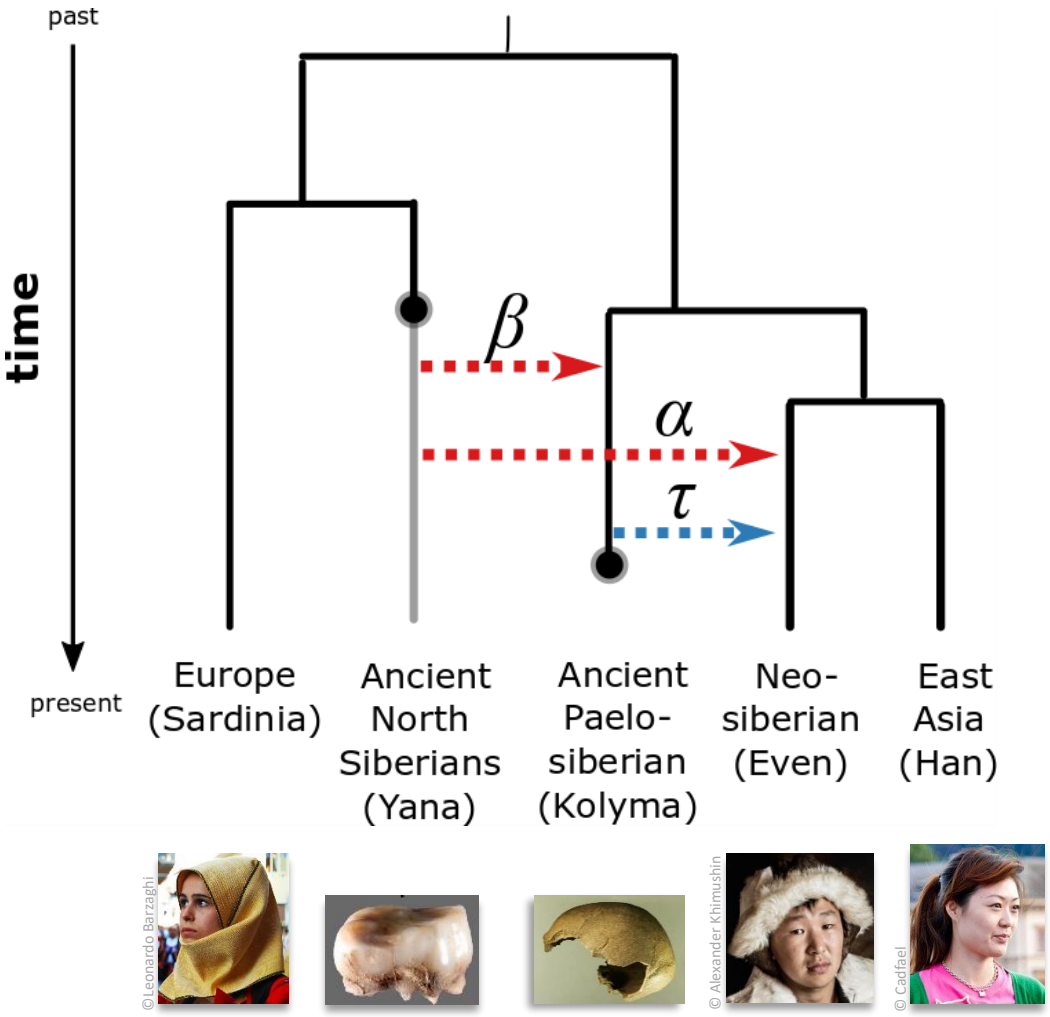
Method: Composite likelihood - *fastsimcoal2*
(Excoffier et al, 2013 Plos Genetics)

Hypothesis: Continuity vs Replacement of populations

For instance:

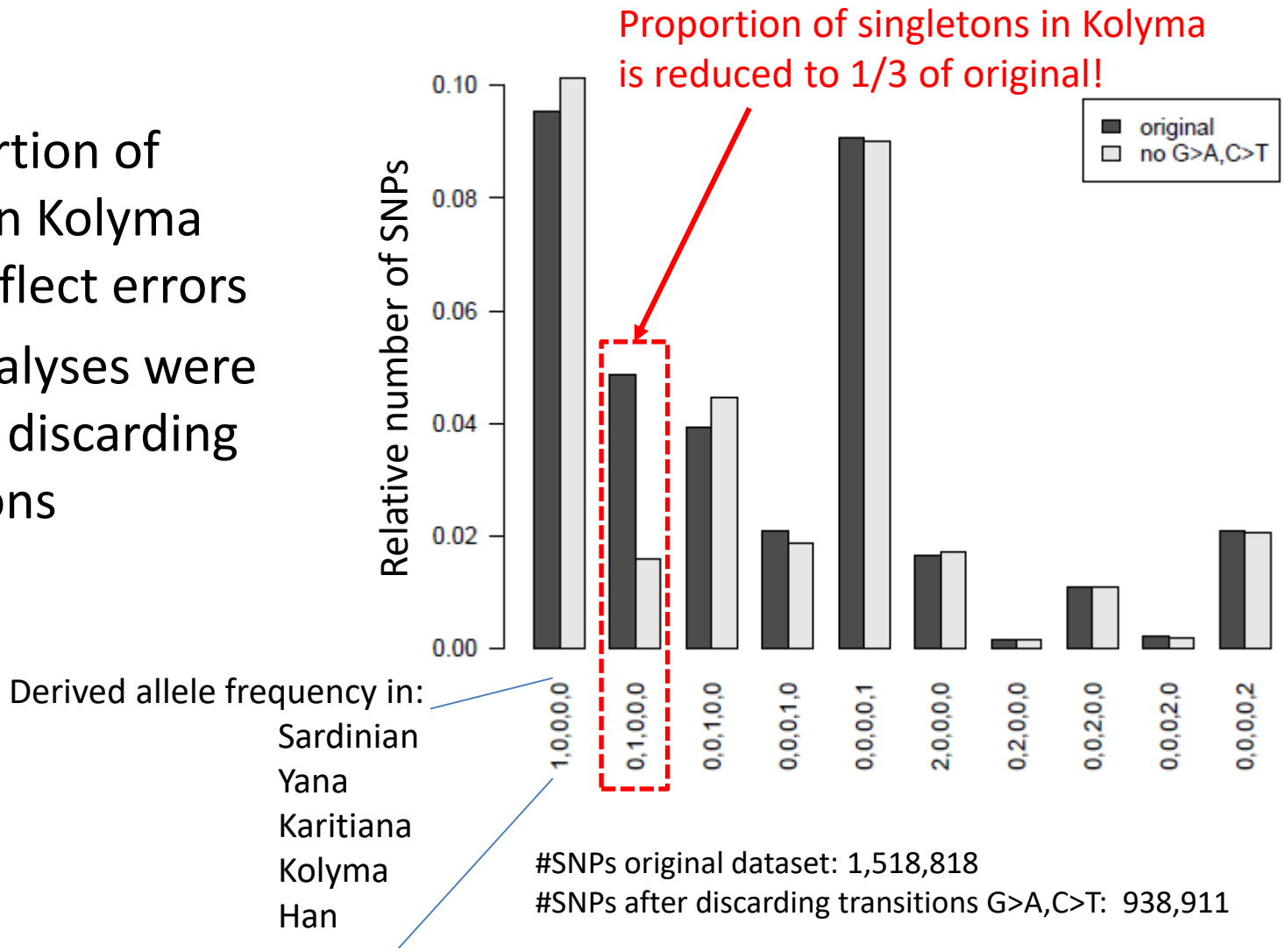
β = 1 indicates continuity: Kolyma descends from Yana

β = 0 indicates replacement of Yana by Kolyma

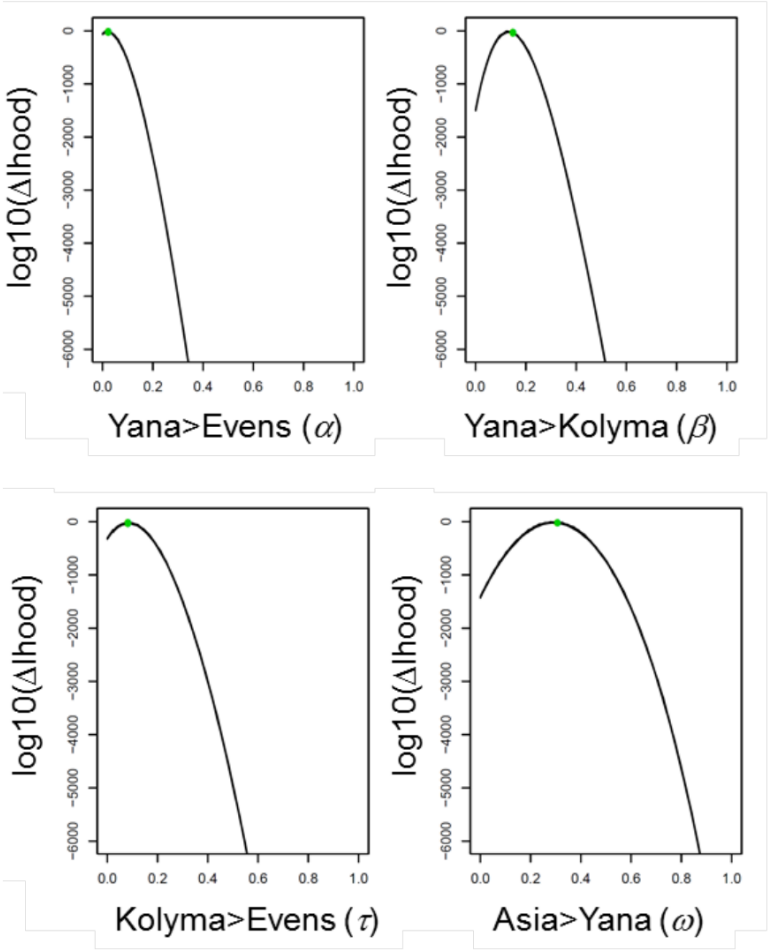


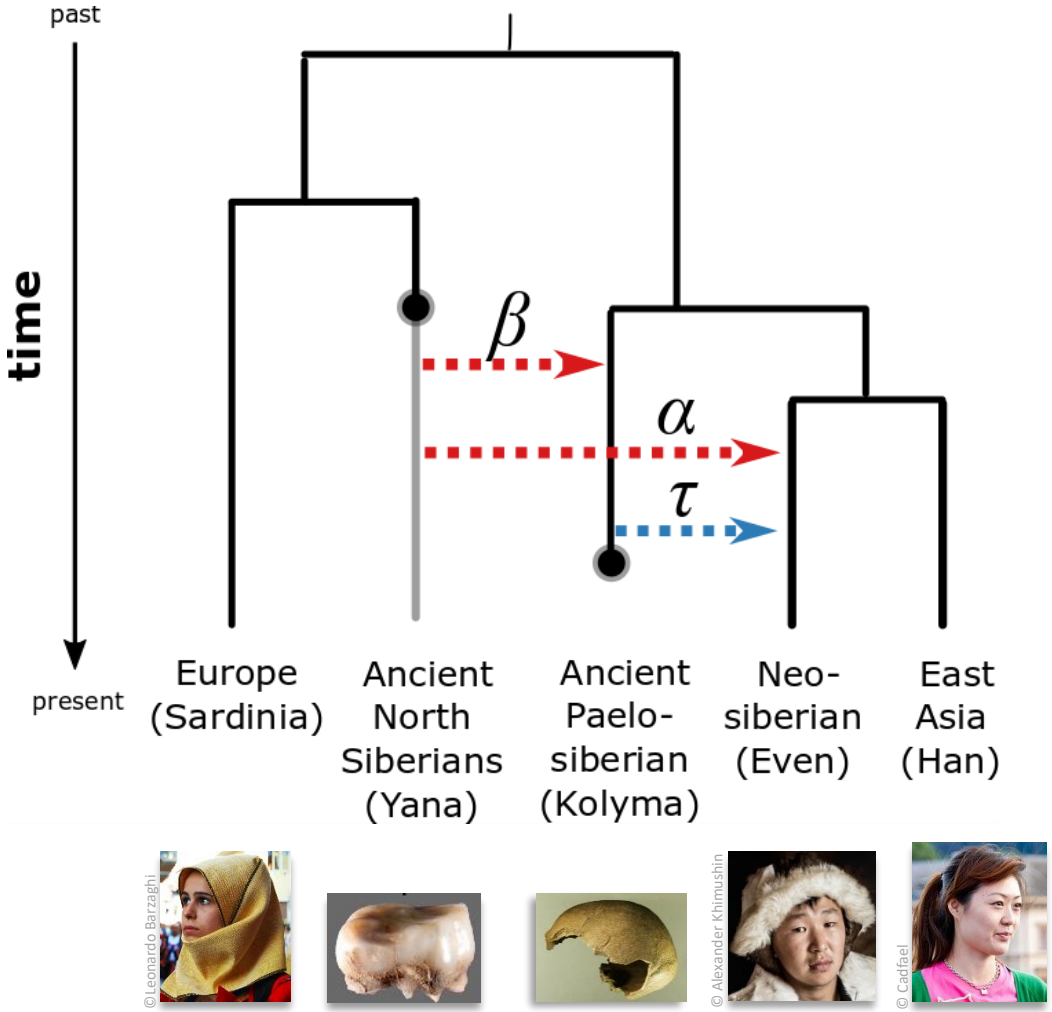
Site frequency spectrum is affected by damage patterns in ancient DNA

- High proportion of singletons in Kolyma probably reflect errors
- Thus, all analyses were performed discarding the singletons

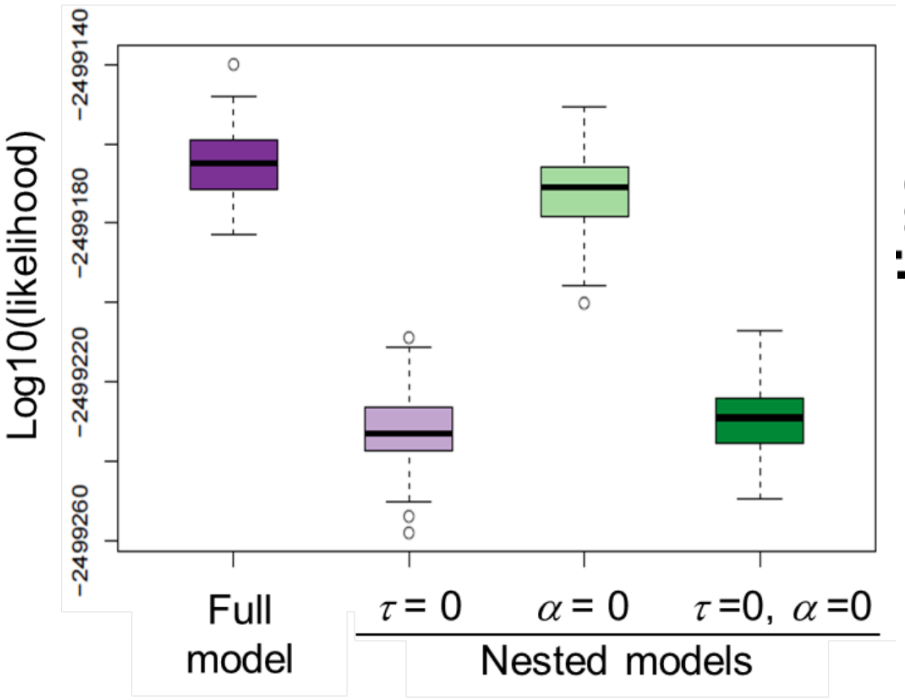


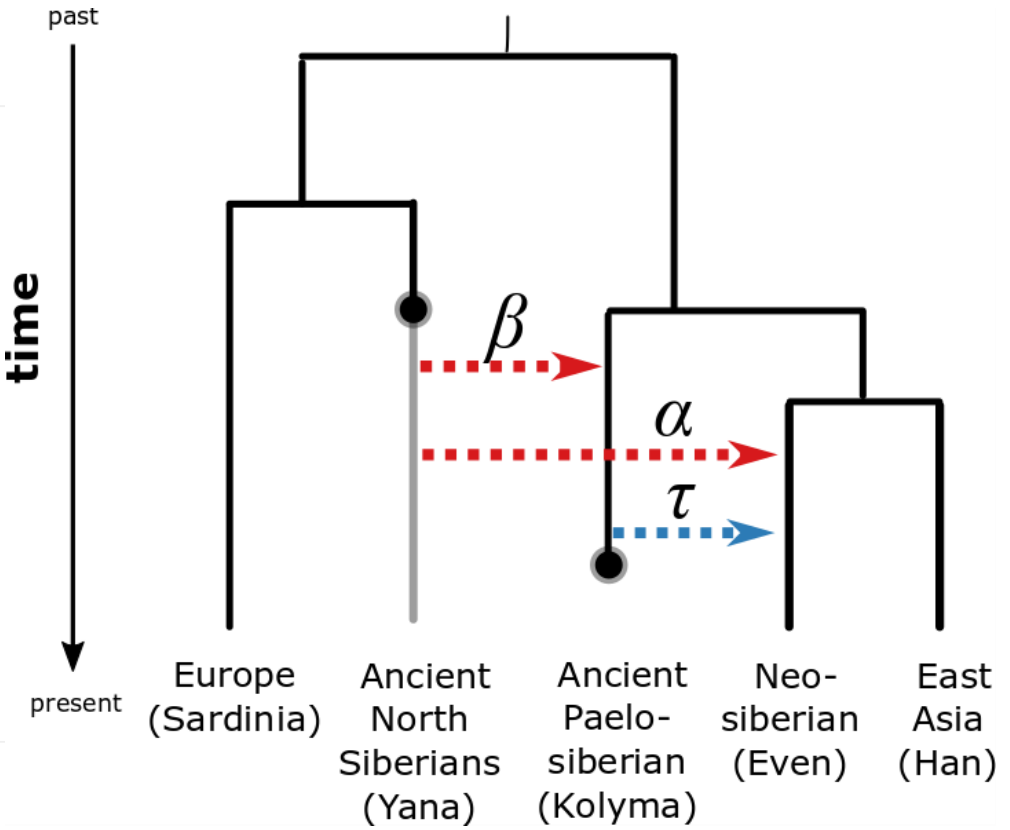
Model comparison and likelihood profiles consistent with replacement with gene flow



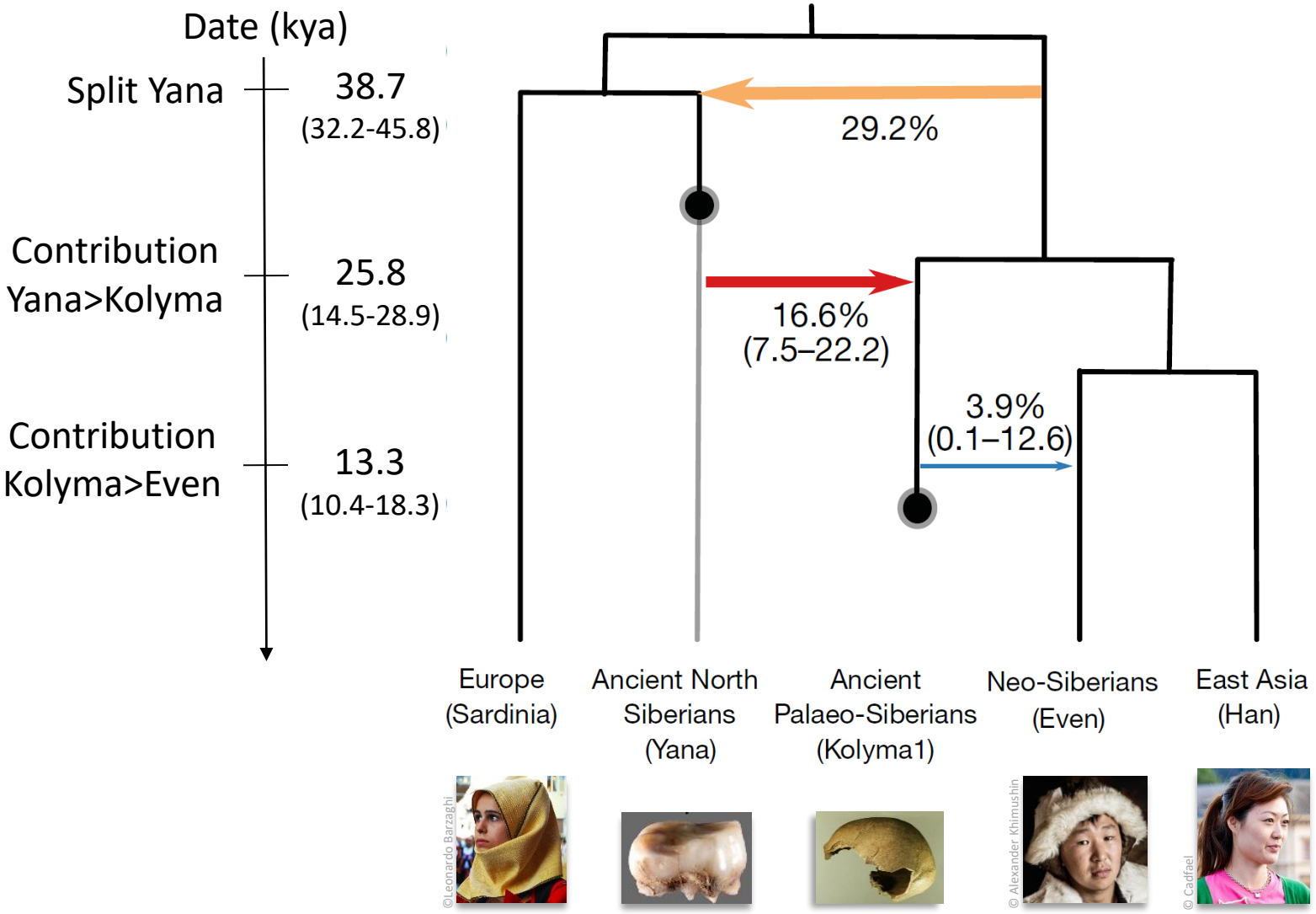


Model comparison and likelihood profiles consistent with replacement with gene flow



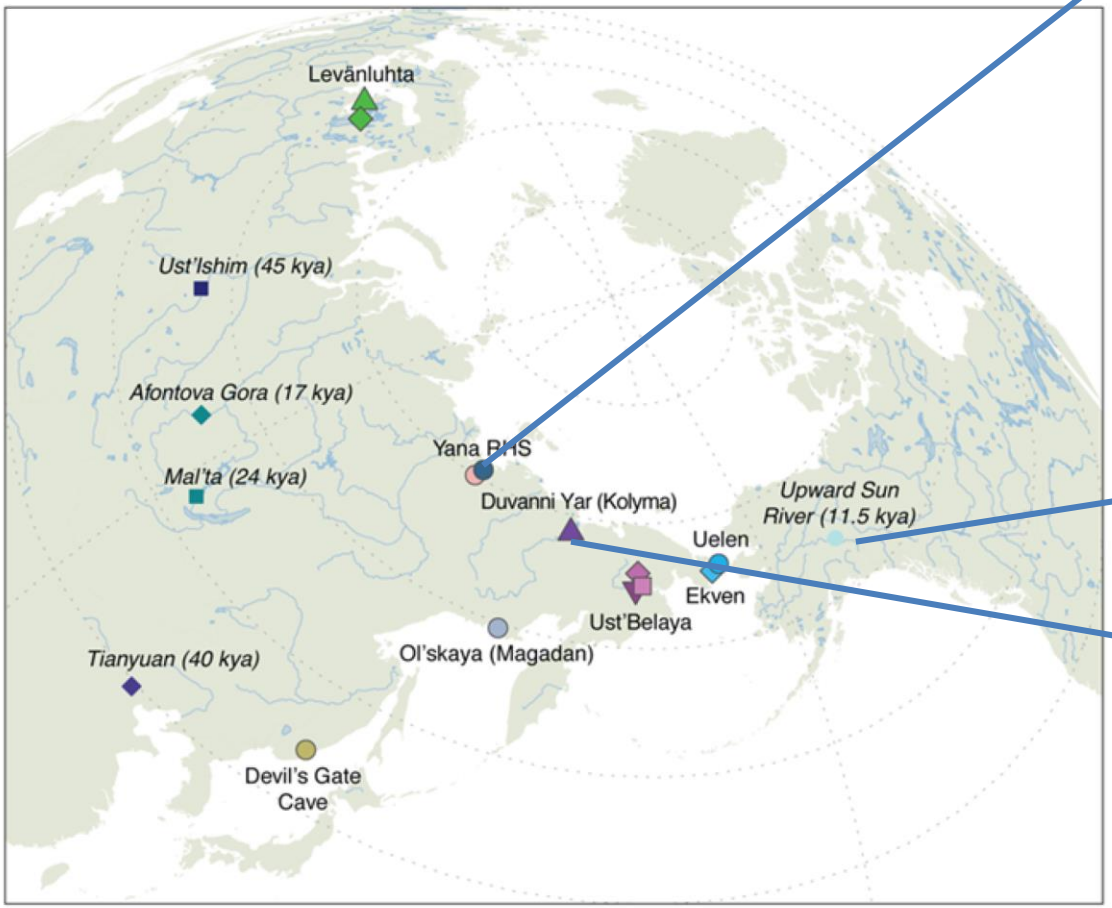


Estimates of best nested model indicate replacement with gene flow



Siberia and colonization of the Americas

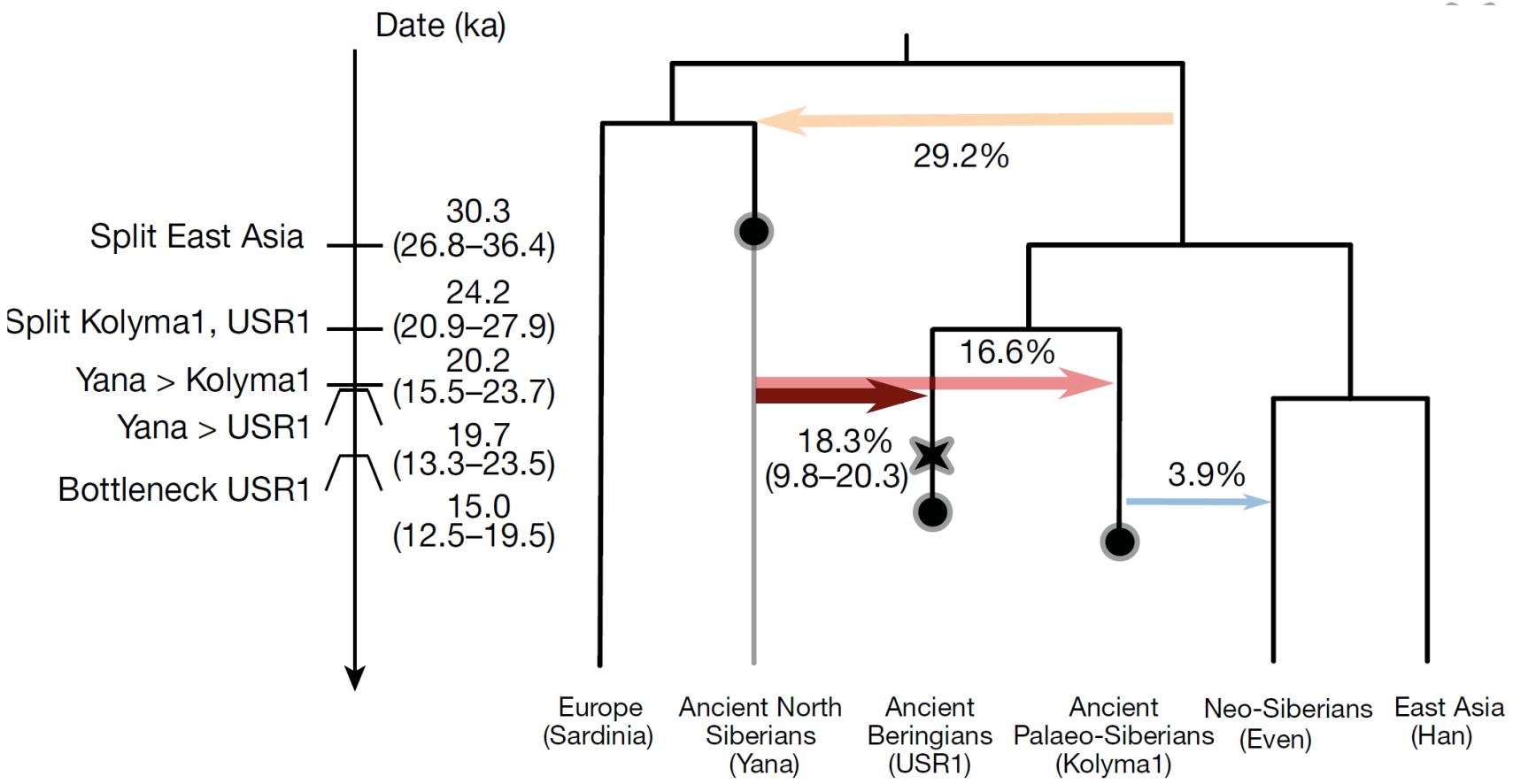
Yana RHS (31,600 years ago)
Whole-genome depth of coverage 25x



USR1 (11,500 years ago) Alaska

Kolyma (9,800 years ago)
Whole-genome depth of coverage 14x

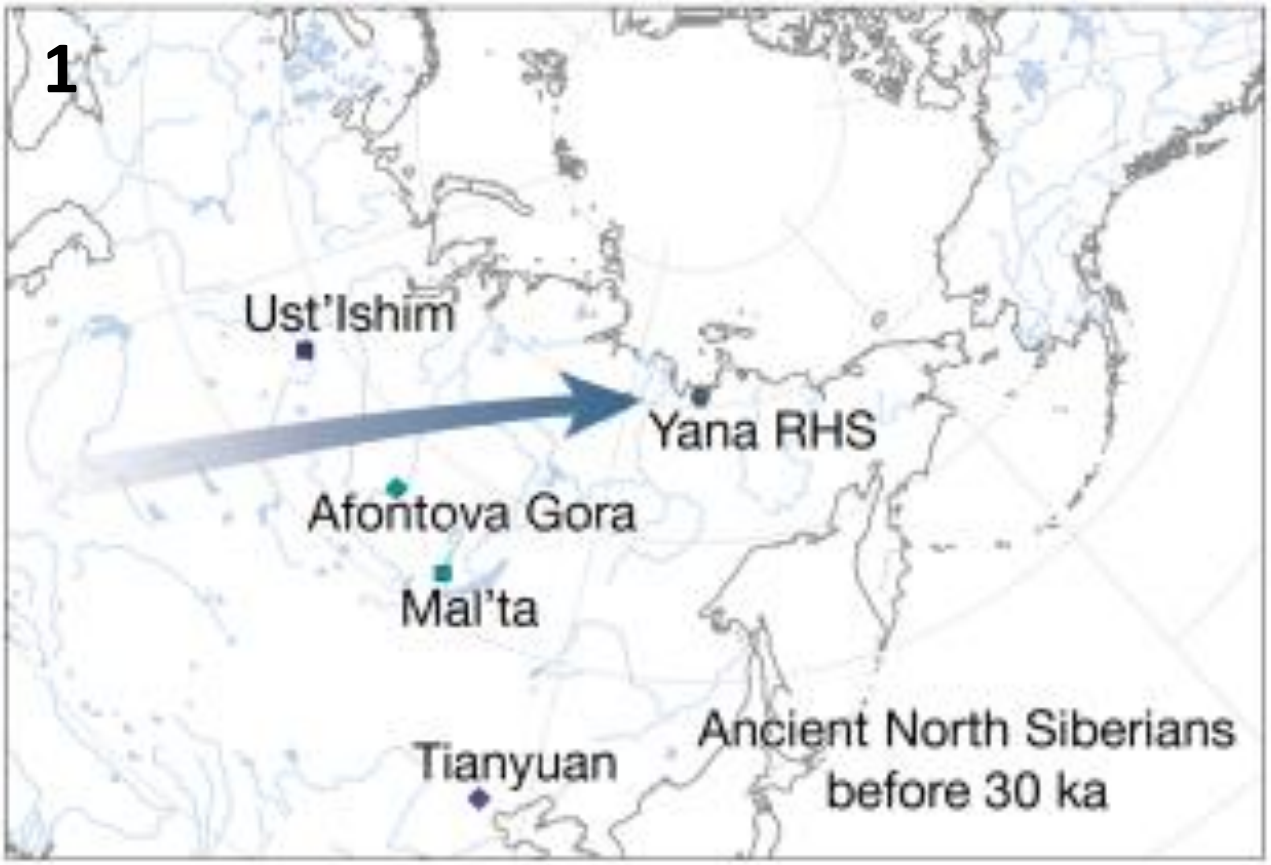
Estimates consistent with replacement with gene flow



- Kolyma is the closest population to Native Americans (USR1 and Karitiana)
- Native Americans with a contribution of up to 20% from Yana

Summary: 3 migration waves

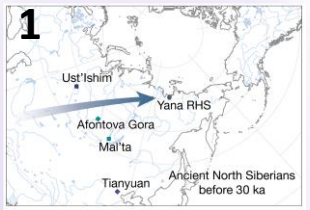
- Ancient North Siberians (Yana) reached Siberia before 30 ka (thousand-years ago)

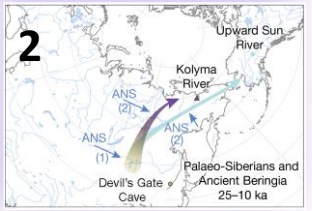


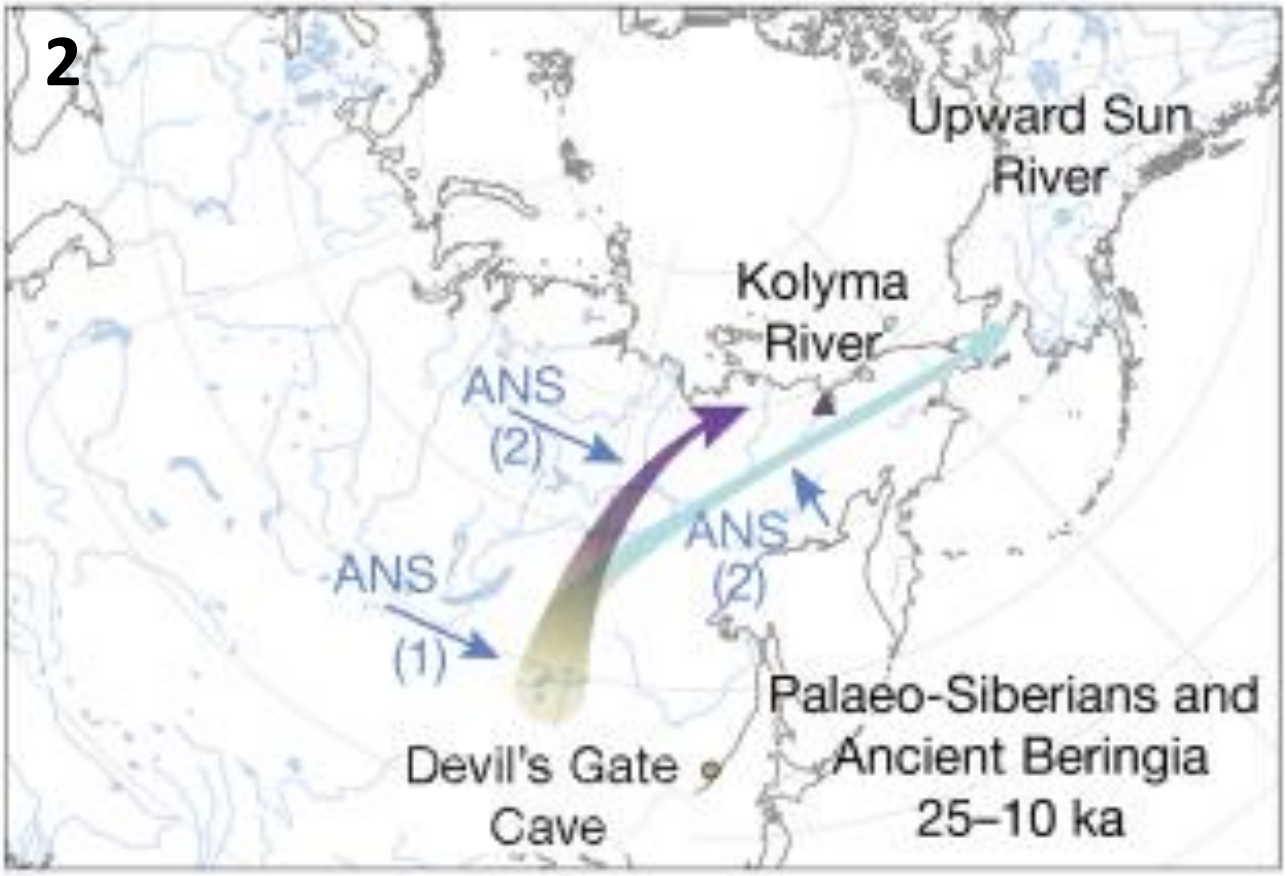
1st migration wave

Summary: 3 migration waves

- Ancient North Siberians (Yana) reached Siberia before 30 kya
- Paleo-Siberians (Kolyma) migrated after Last Glacial Maximum (26.5 ka)
- Native-Americans are closer to Kolyma, with 20% of Yana contribution



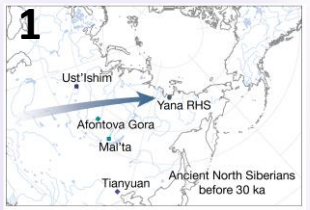


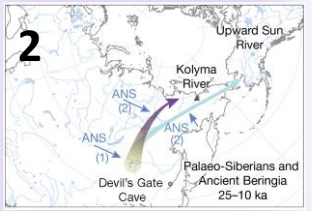


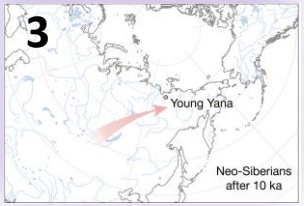
2nd migration wave

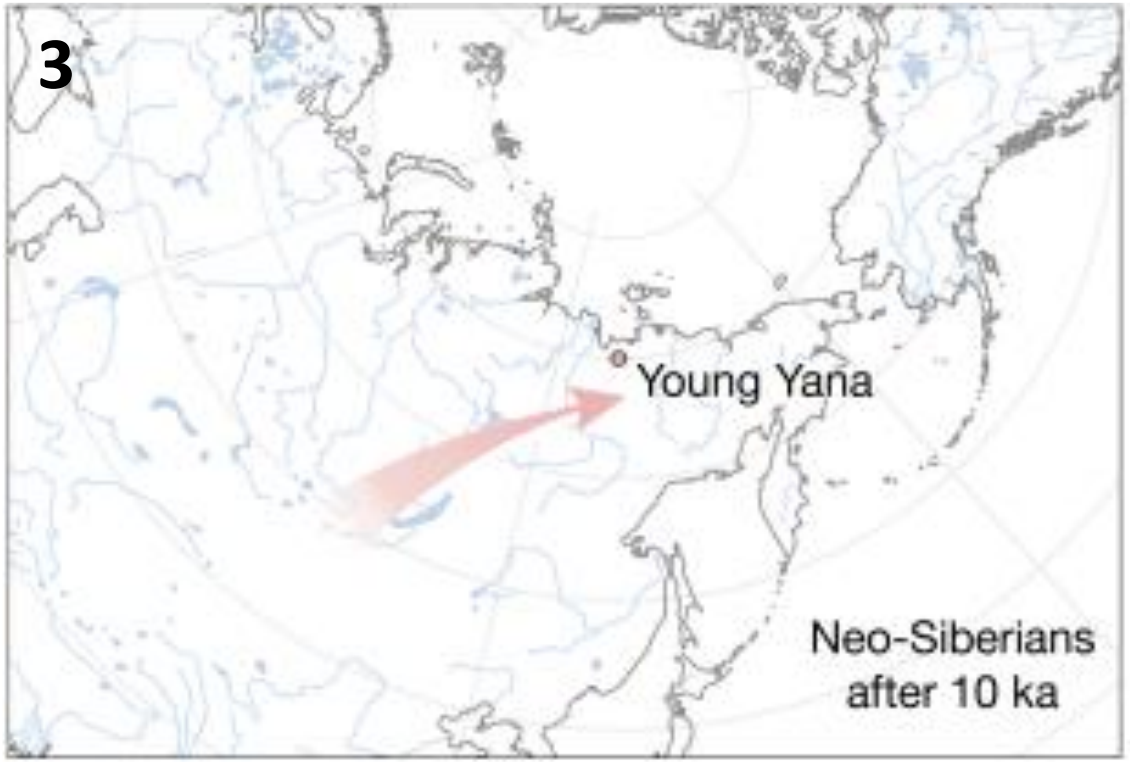
Summary: 3 migration waves

- Ancient North Siberians (Yana) reached Siberia before 30 ka
- Paleo-Siberians (Kolyma) likely migrated after Last Glacial Maxima
- Native-Americans are closer to Kolyma, with 20% of Yana contribution
- Paleo-Siberians (Kolyma) were replaced by Neo-Siberians, likely associated with the cooler period “Younger Dryas” (12.8-11.5 ka)









3rd migration wave

Deer mice from Nebraska Sand Hills

S. Pfeifer, S. Laurent, V. Sousa, C. Linnen, H. Hoekstra, L. Excoffier, J. Jensen

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

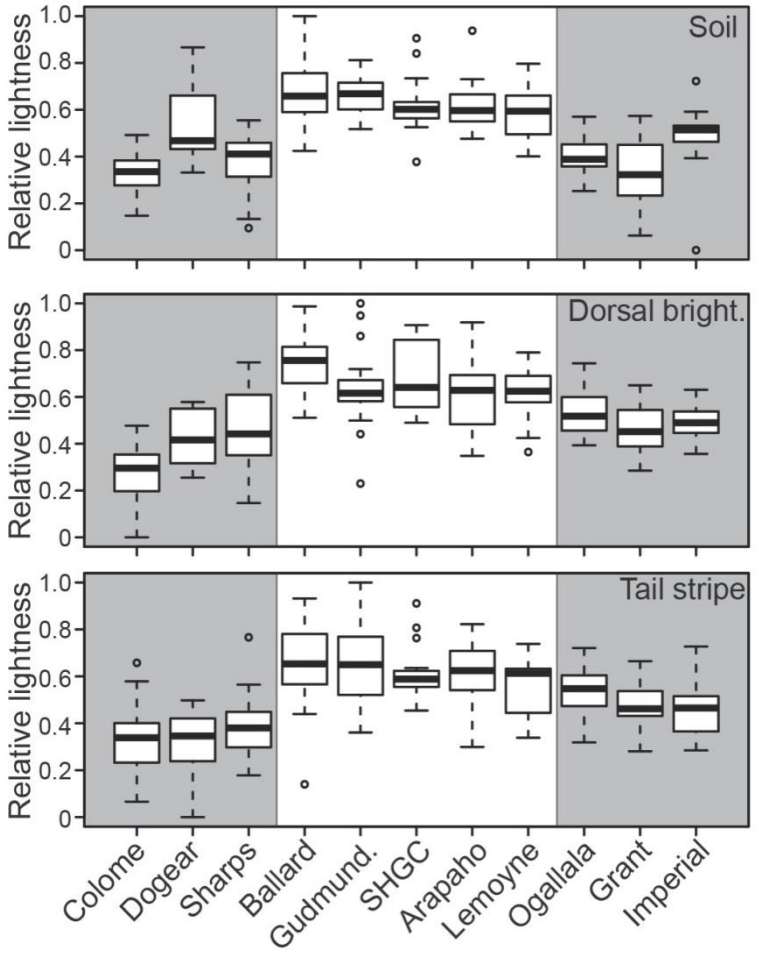
Linnen et al (2013) Science

Pfeifer*, Laurent*, Sousa* et al (in press) MBE

A transect across the Sand Hills (ON and OFF)

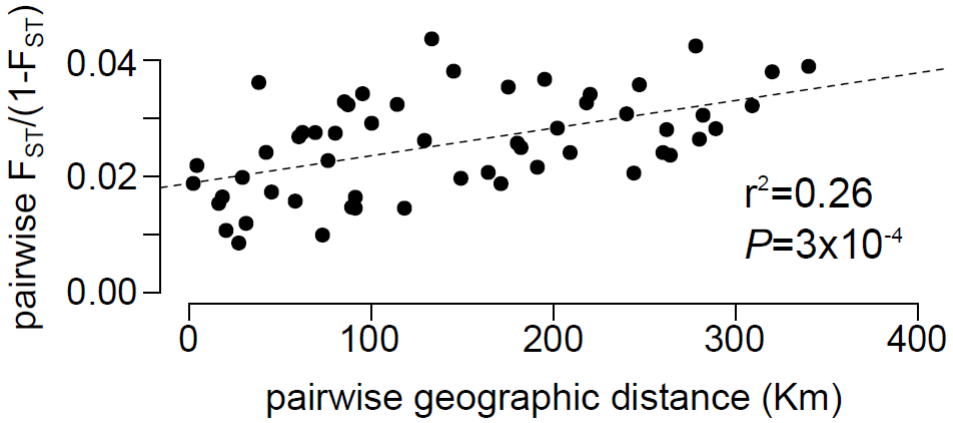
Sample locations “off” and “on” the Sand Hills

- 11 populations
- 330 individuals

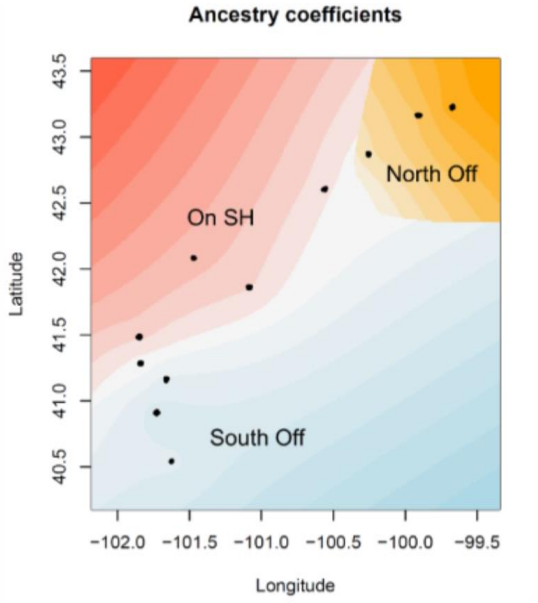


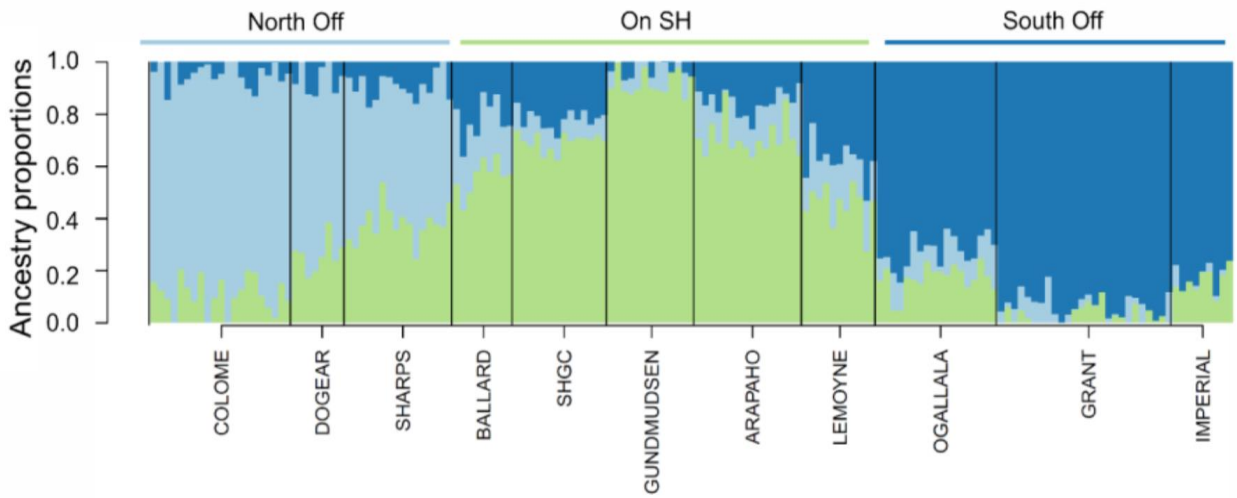
- Genomic data (NGS) data
  - Target 10,000 random 1.5kb regions
  - 185kbp region comprising the *Agouti* gene
- Phenotypic data for each individual

Evidence for isolation by distance but three groups



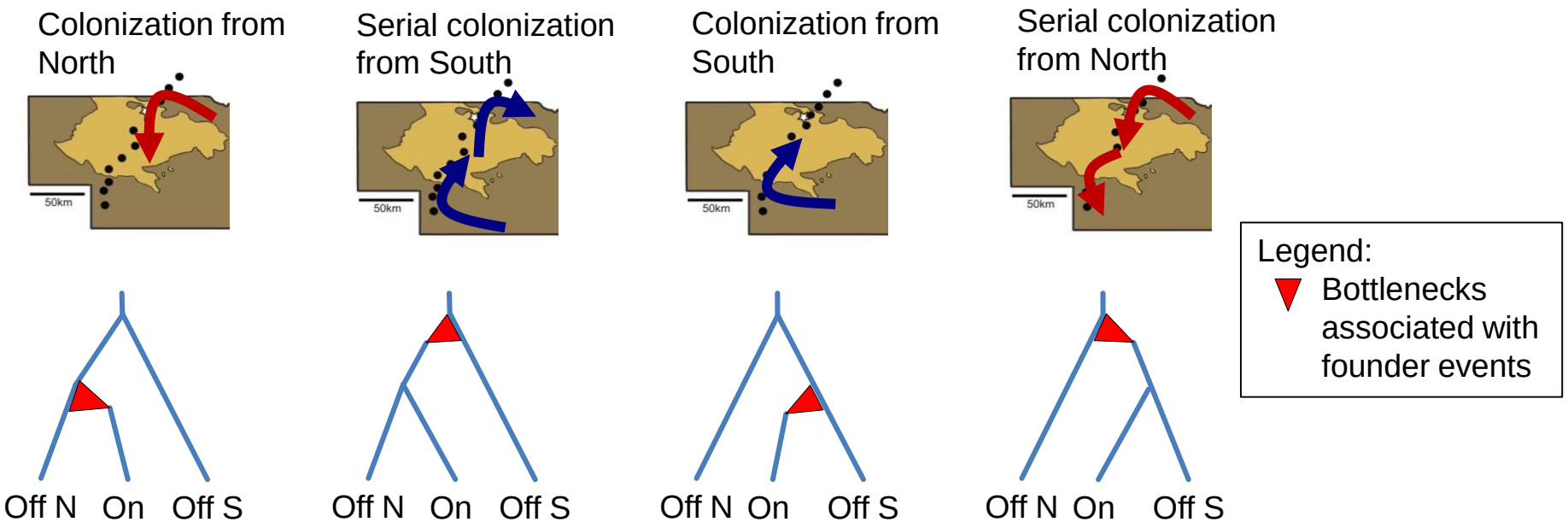
Geographically closer samples are genetically more similar





Model-based inference

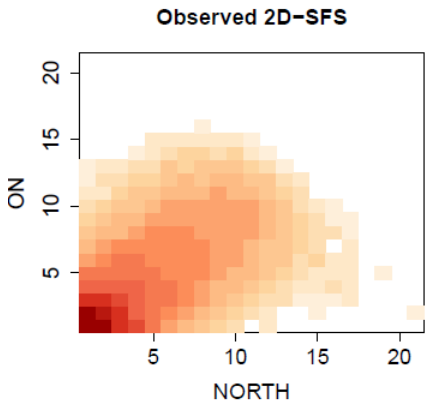
Is there evidence of gene flow between Off and On the Sand Hills?

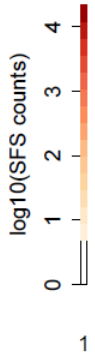


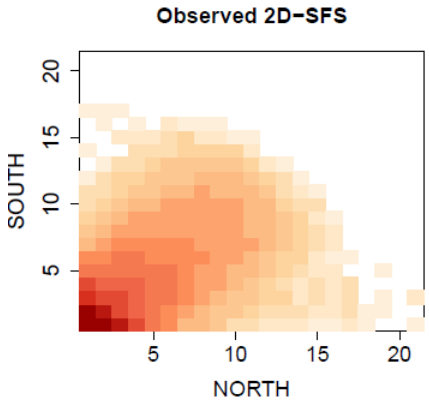
Estimates based on the joint **3D site frequency spectrum** (SFS):

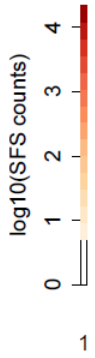
- folded SFS with 140,358 SNPs

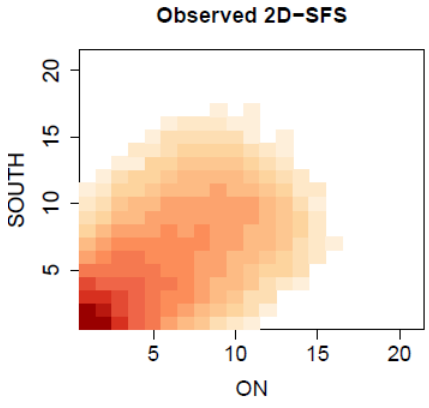
Deer mice: Pairwise marginal 2D SFS
Since we did not have an outgroup we used the folded SFS

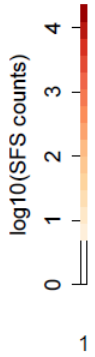








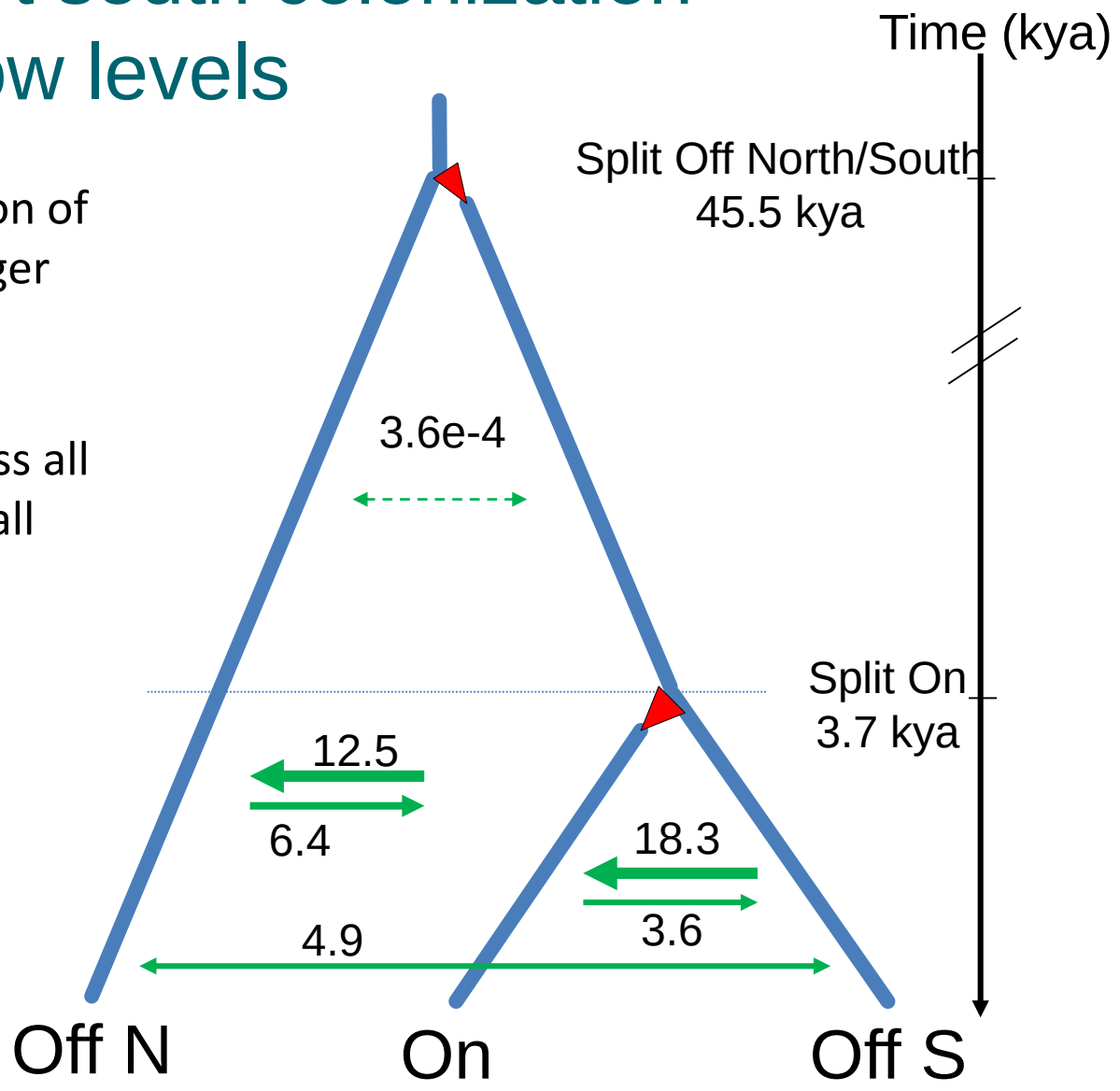




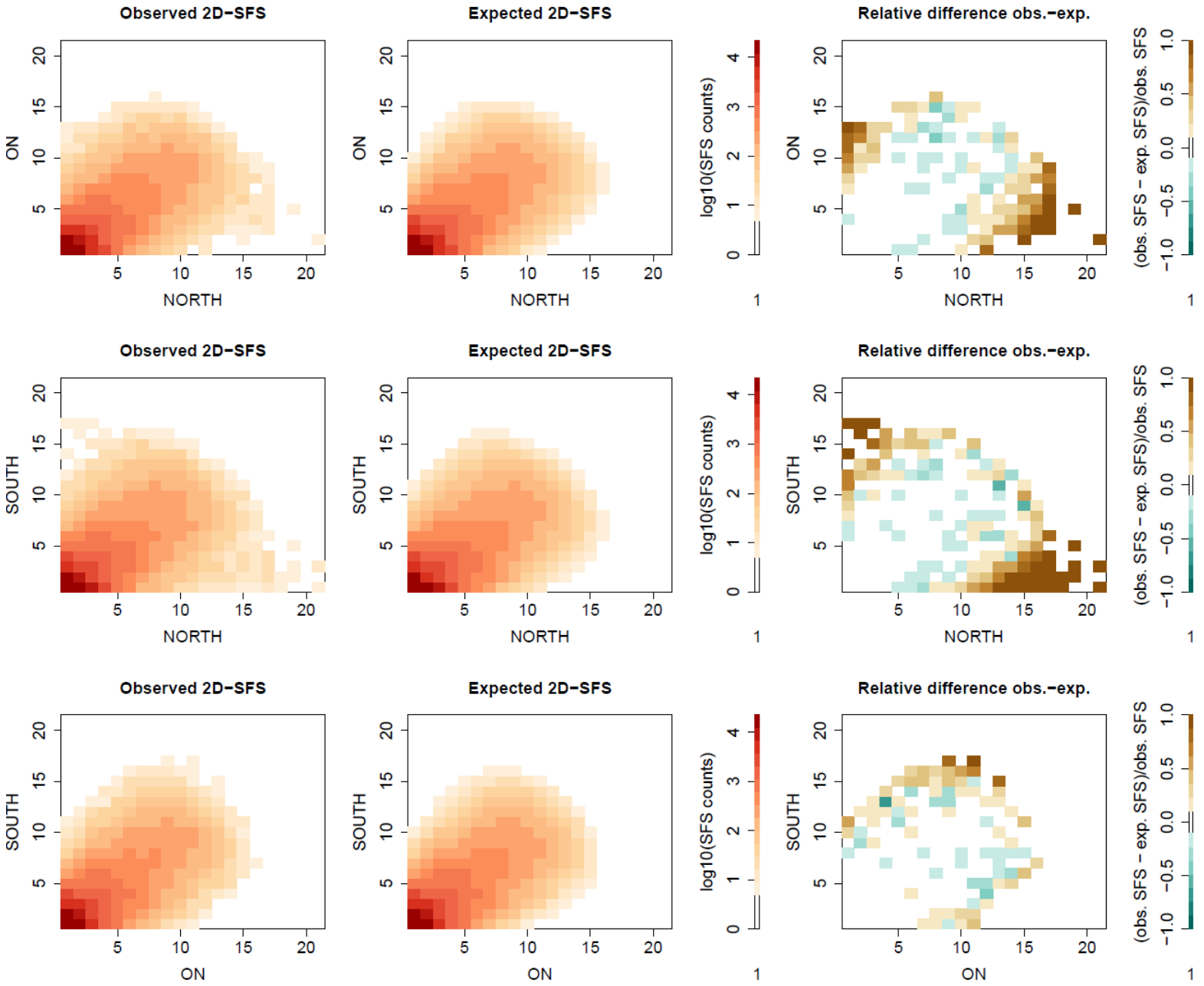
Estimates support south colonization and high gene flow levels

- 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.



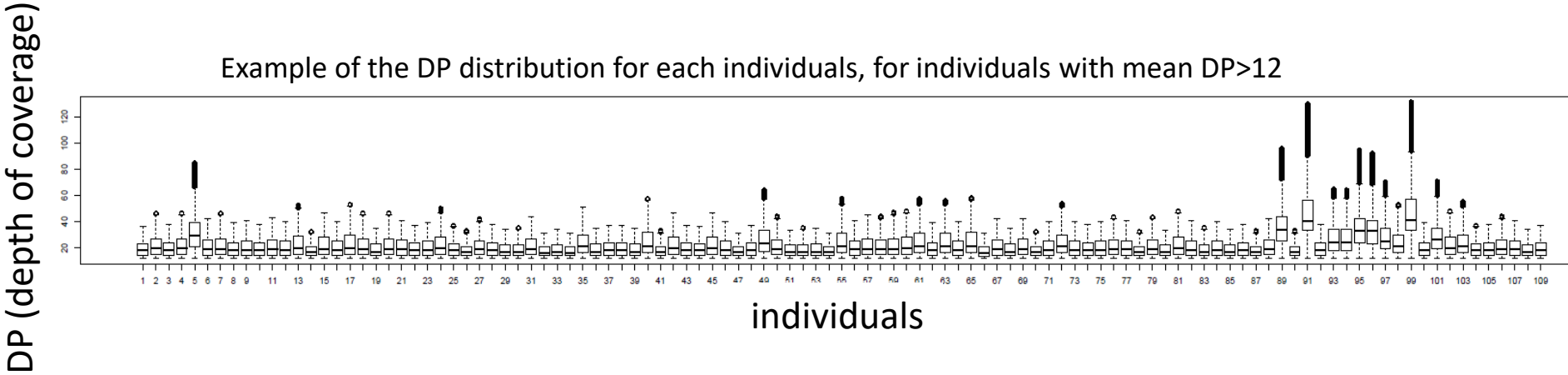
Deer mice: Model fit to marginal SFS



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 DP filters on the SFS

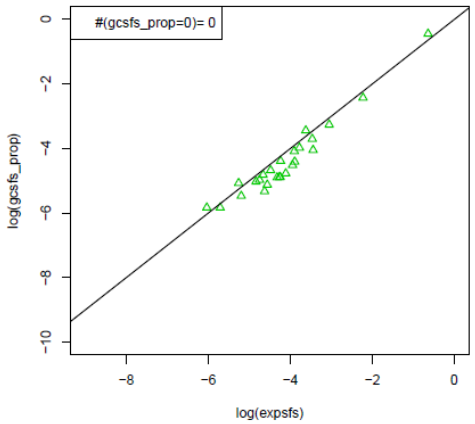
Simulation study

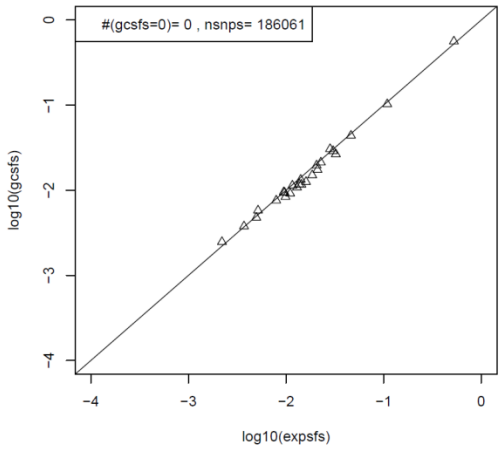
DP > 10

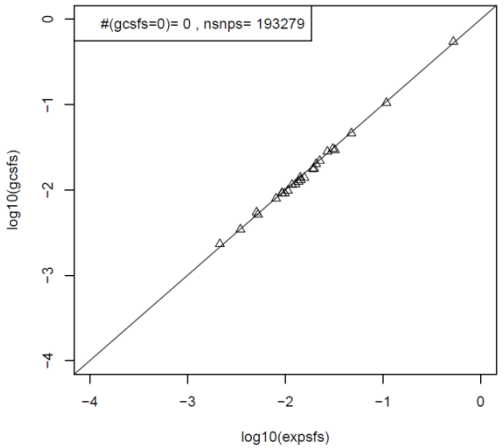
DP > 15

DP > 20

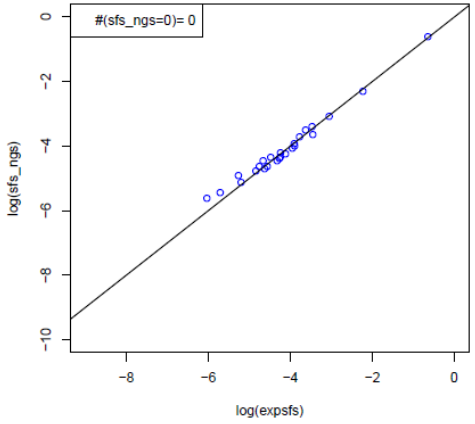
SFS based on called genotypes







SFS accounting for genotype uncertainty (ANGSD)

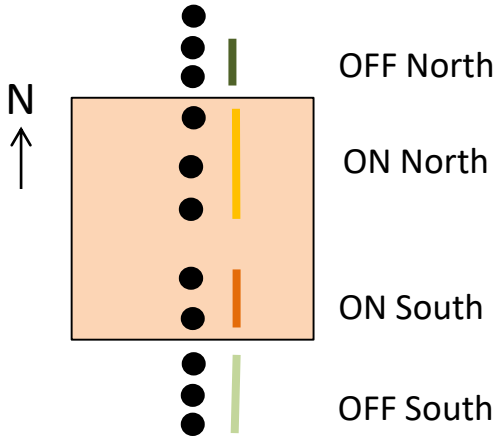


Simulated 2 pops SFS sampling 4 diploids from each pop, 200000 SNPs, mean coverage=**10x**, error rate=0.01. Simulated with correlated allele frequencies model (F_{ST}=(0.275, 0.01))

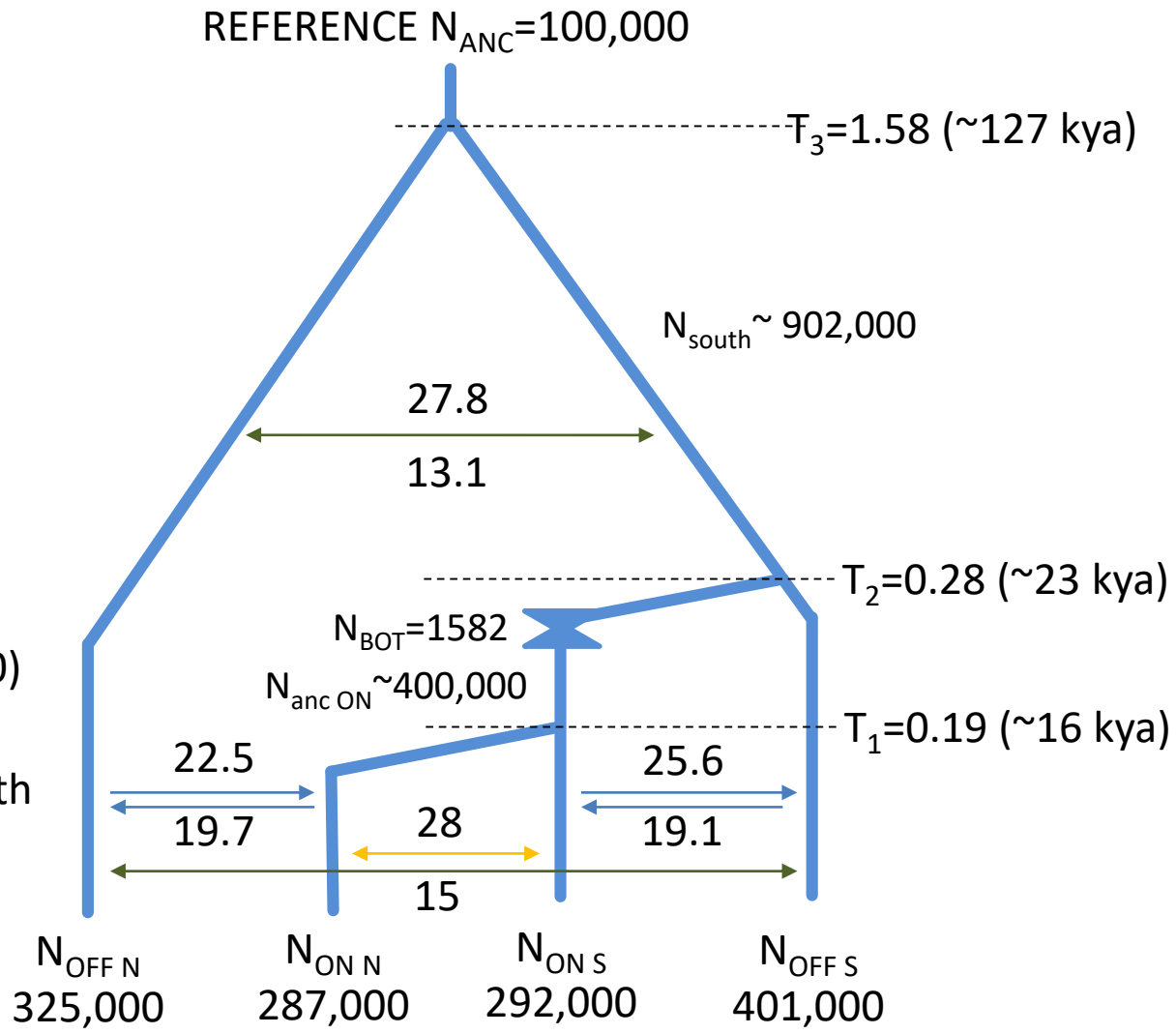
With DP>15 we have a very good approximation to the correct SFS, even when using the called genotypes

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

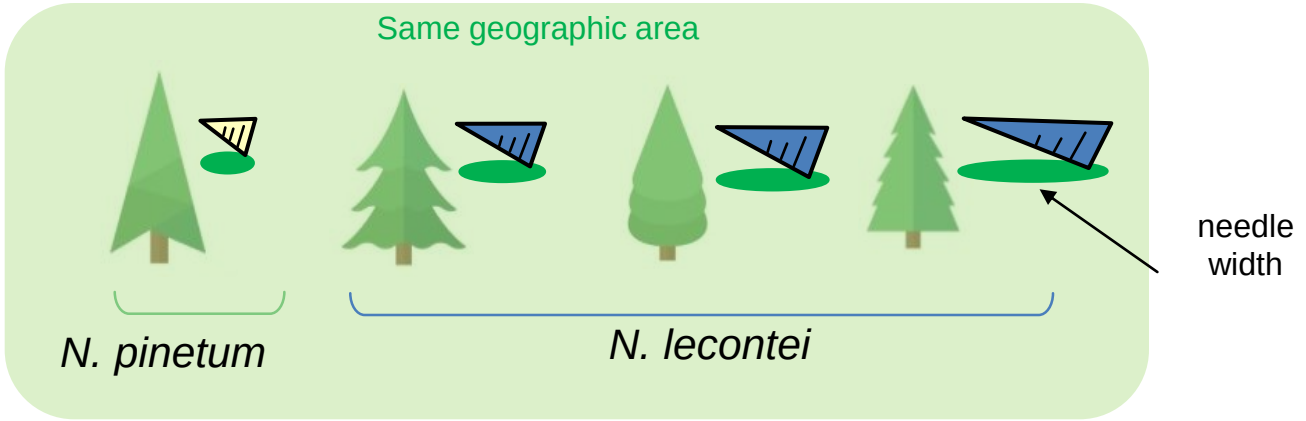
History, geography and host use shape genomewide patterns of genetic variation in the redheaded pine sawfly (*Neodiprion lecontei*)

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*

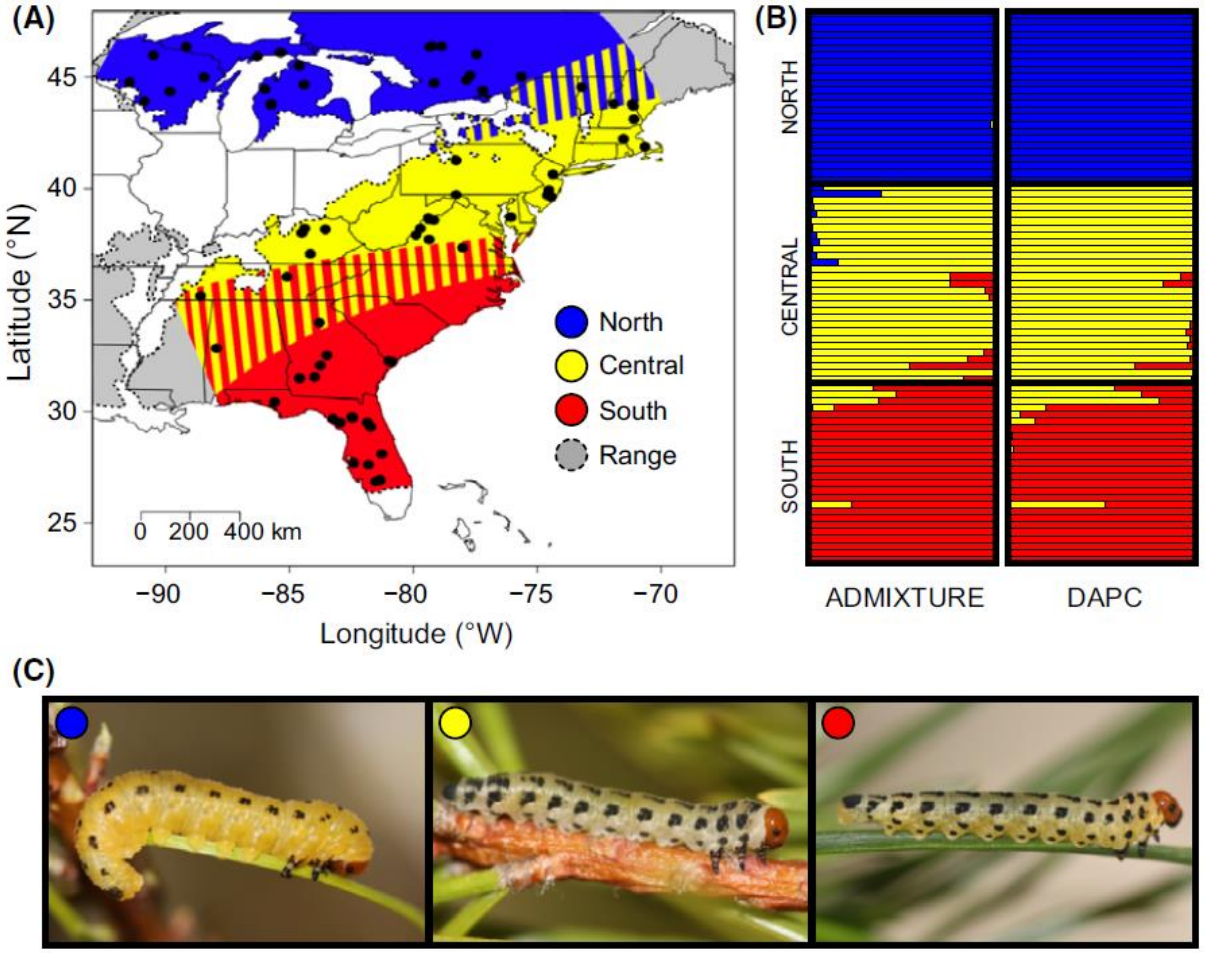
Sawflies *Neodiprion lecontei*

- 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

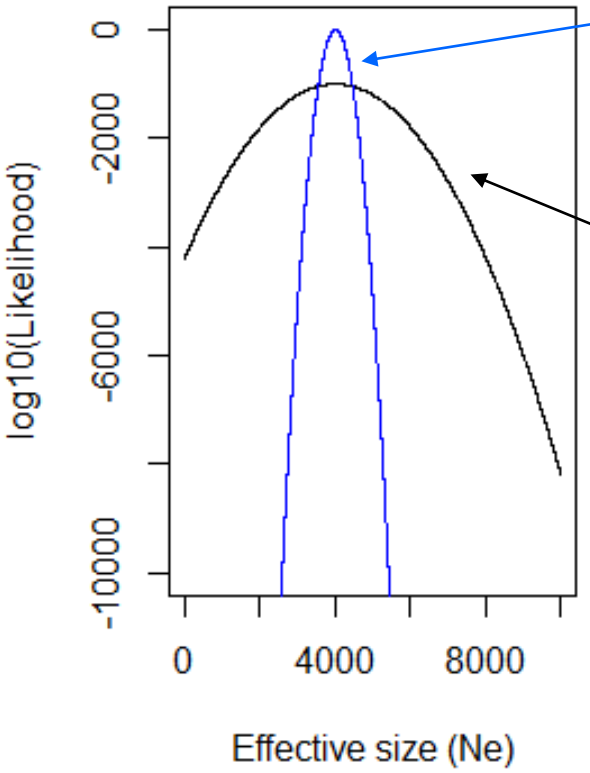


Given the detected three groups (North, Central, South):

- 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)



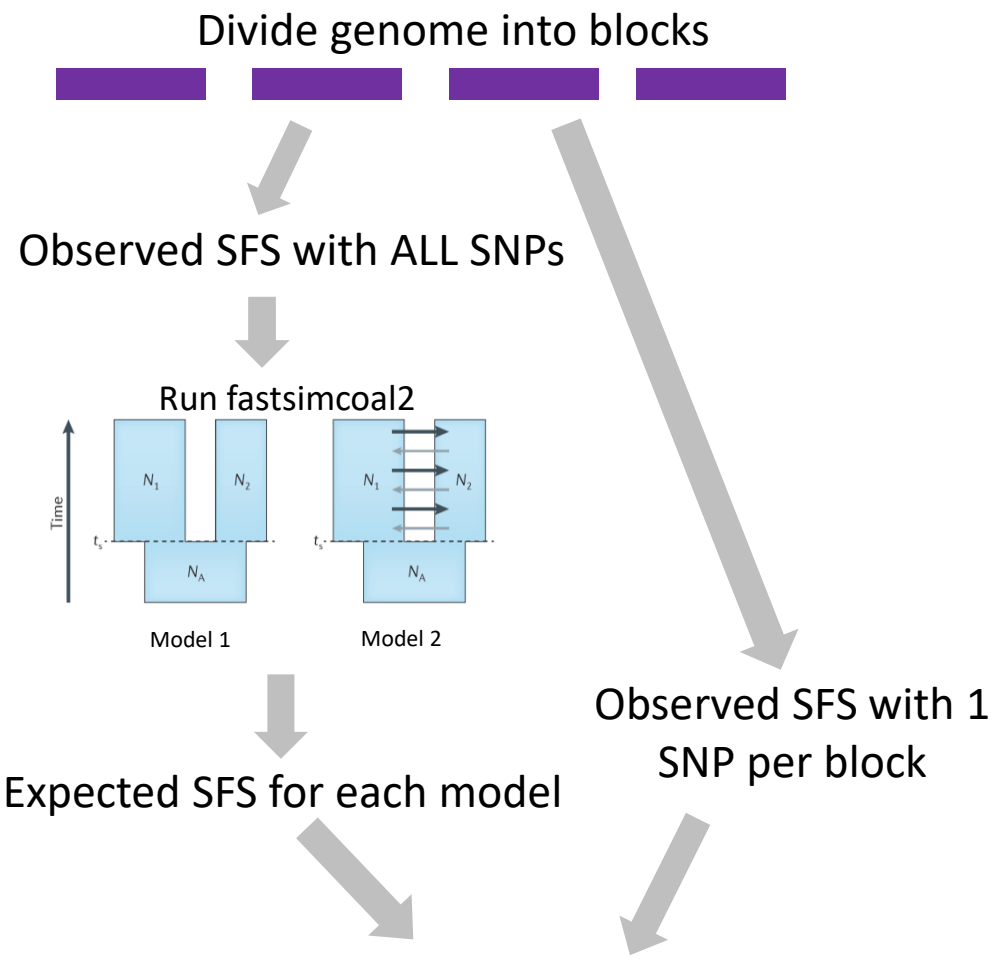
Composite likelihood (assuming linked sites are independent)

“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

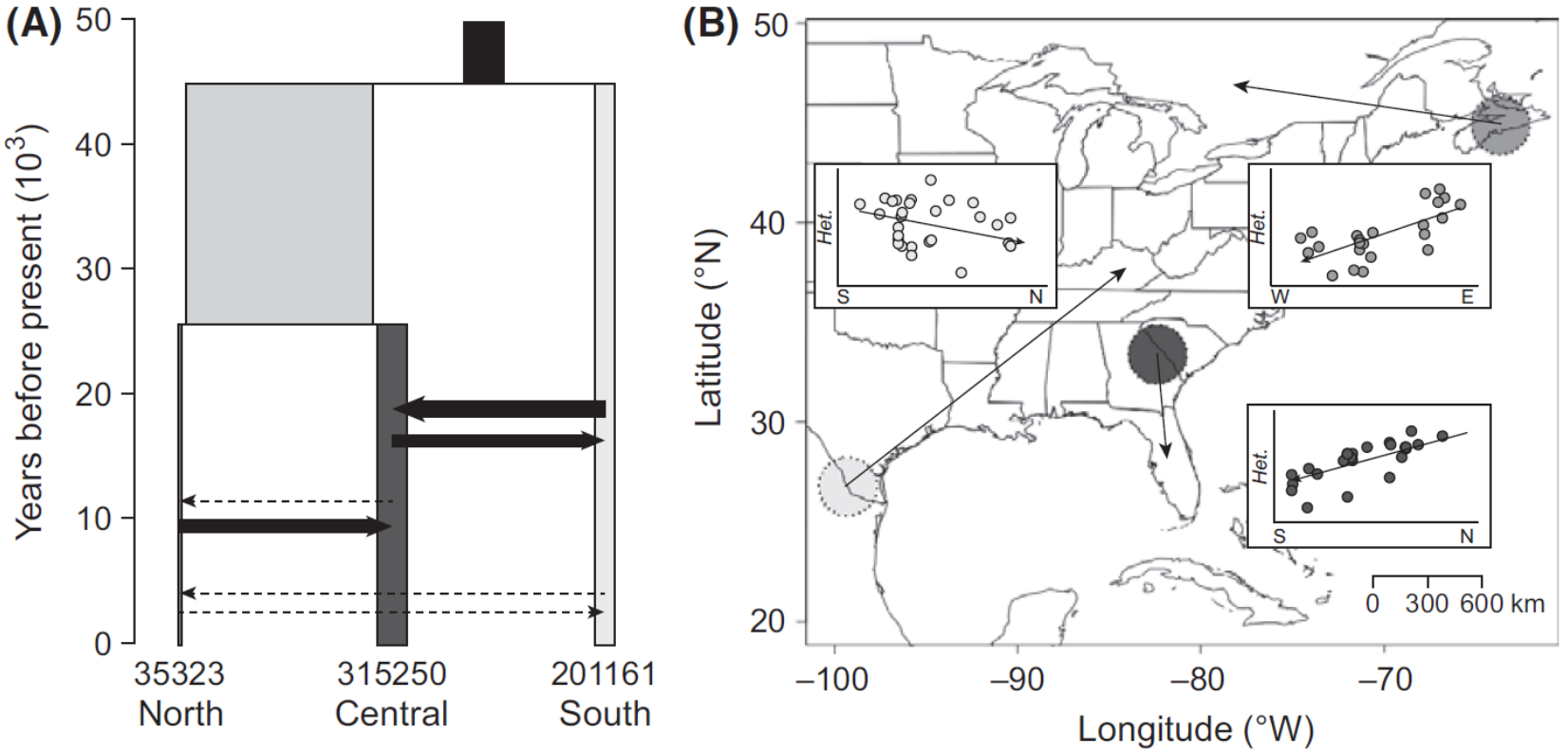


Comparing alternative models

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

- 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

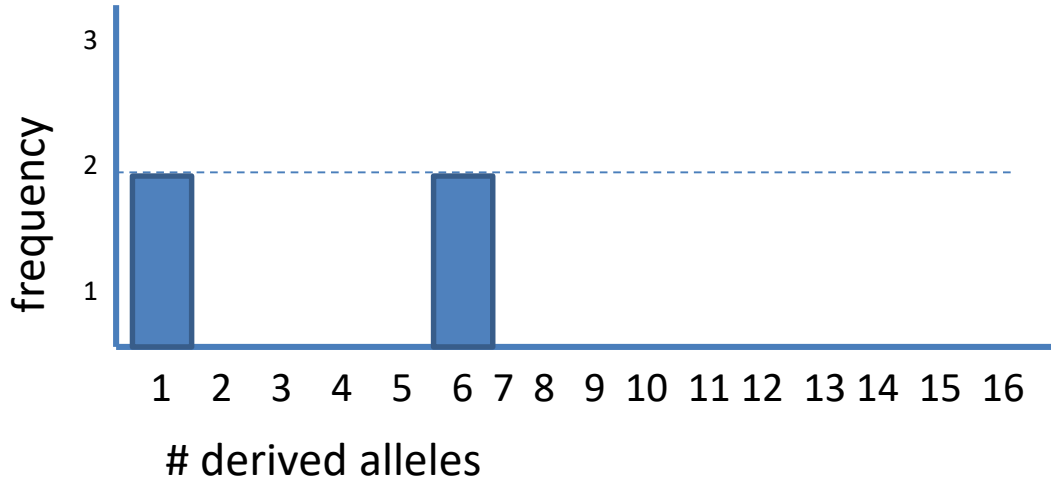
Protocol for model comparison based on AIC when we have independent SNPs

- 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
- Select the model with minimum AIC

Estimating SFS from observed data

- The sample size can vary across SNPs due to missing data
- How to deal with missing data?

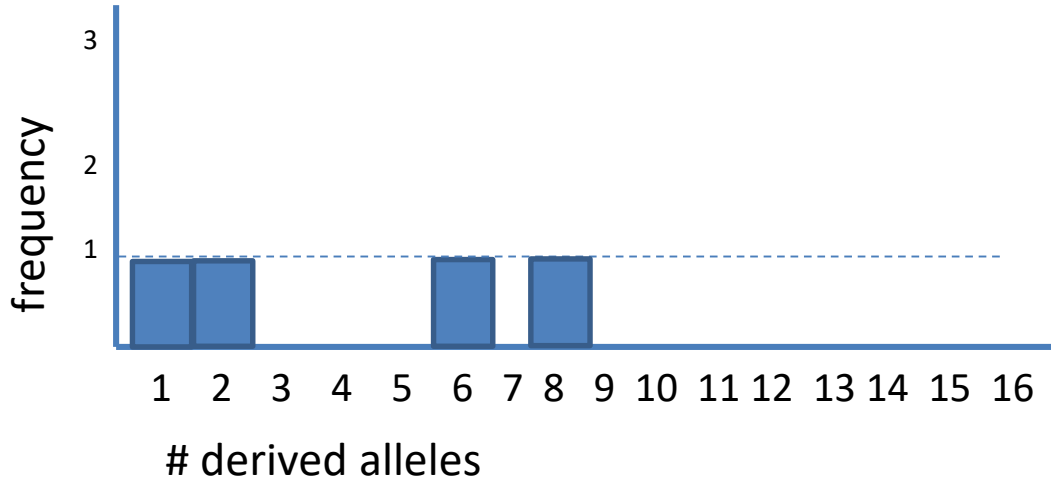
	Freq. derived	Sample size	Rel. freq
SNP1	1	16	1/16
SNP2	6	12	1/2
SNP3	1	12	1/12
SNP4	6	16	3/8



Estimating SFS from observed data

- The sample size can vary across SNPs due to missing data
- How to deal with missing data?

	Freq. derived	Sample size	Rel. freq
SNP1	1	16	1/16
SNP2	6	12	1/2
SNP3	1	8	1/12
SNP4	6	16	3/8

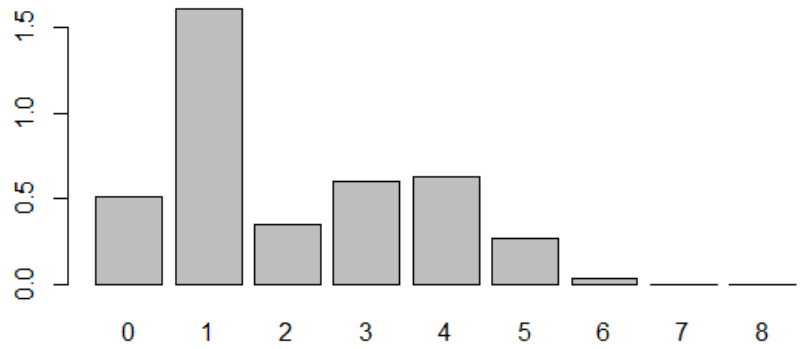


Estimating SFS from observed data

- The sample size can vary across SNPs due to missing data
- How to deal with missing data?

- Solution:
  - Find minimum sample size
  - Resample without replacement

	Freq. derived	Sample size	Rel. freq
SNP1	1	16	1/16
SNP2	6	12	1/2
SNP3	1	8	1/12
SNP4	6	16	3/8



Gavel et al. (2011) PNAS

Acknowledgements

Martin Sikora
Laurent Excoffier
Isabelle Dupanloup
Stephan Peischl
Eske Willerslev

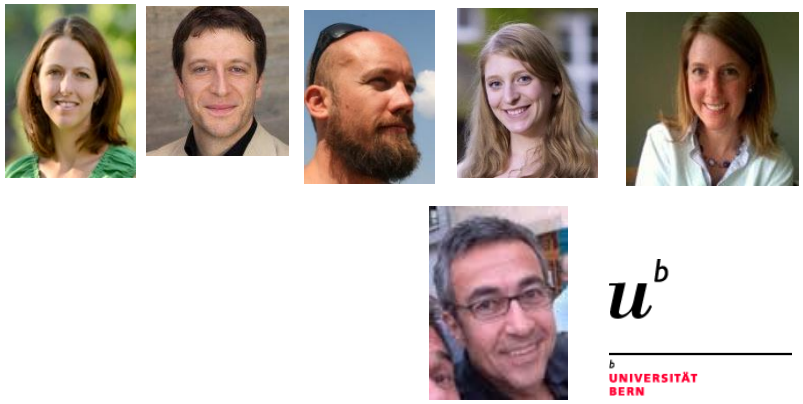
FNSNF Fonds national suisse Schweizerischer Nationalfonds Fondo nazionale svizzero Swiss National Science Foundation

Danmarks Grundforskningsfond Danish National Research Foundation

Centre for GeoGenetics

Thank you!

Catherine R. Linnen
Stefan Laurent
Jeffrey D. Jensen
Susanne Pfeifer
Hopi E. Hoekstra
Laurent Excoffier



Harvard University Museum of Comparative Zoology

CMPG

UK University of Kentucky

u b Universität Bern

SIB Swiss Institute of Bioinformatics

cE3c centre for ecology, evolution and environmental changes

FCT Fundação para a Ciência e a Tecnologia
Ministério da Ciência, Tecnologia e Ensino Superior

FC Ciências ULisboa

UID/BIA/00329/2015-2018
UID/BIA/00329/2019
CEECIND/02391/2017

Marie Curie Actions

MCSA 2018-2020:
MAPgenome
(N.799729)