Evolutionary genomics
Data analysis module – Day 3

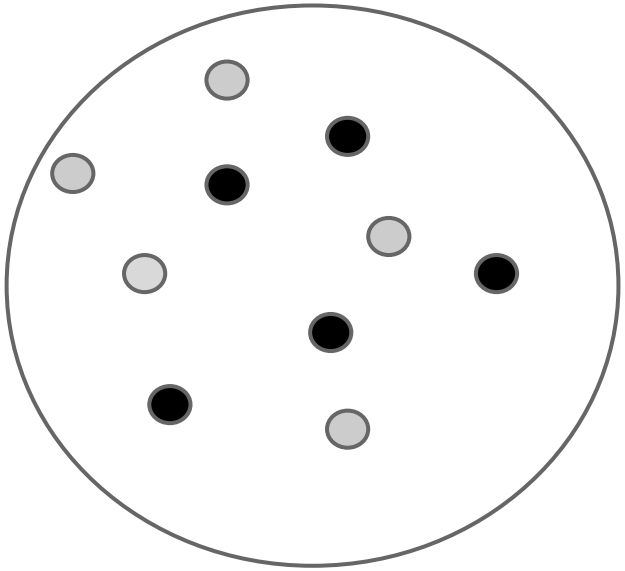
Population structure
and demographic inferences

April 15th 2015

Population subdivision

Population subdivision/structure: when the assumption of random mating (e.g. within a species) is not met

Important for demographic inferences, detection of natural selection, conservation, ...

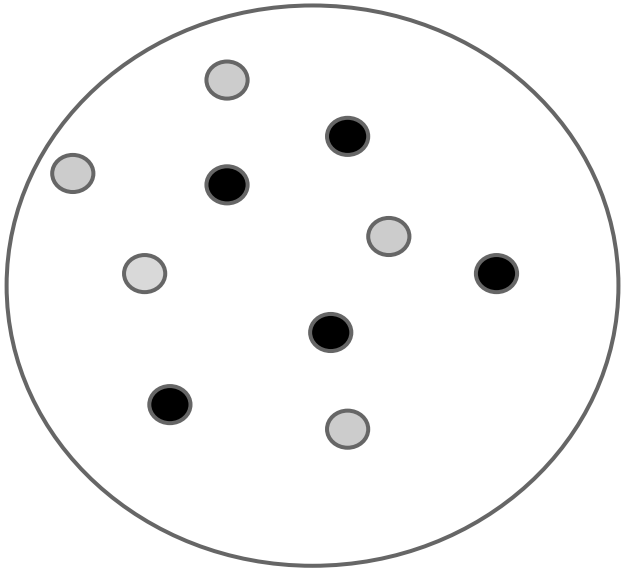


Expected Heterozygosity = 0.50

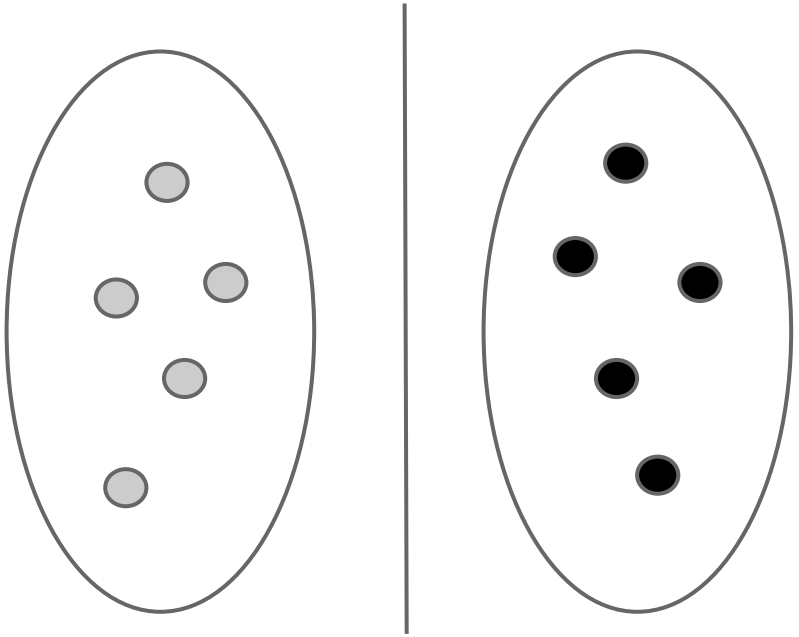
Population subdivision

Population subdivision/structure: when the assumption of random mating (e.g. within a species) is not met

Important for demographic inferences, detection of natural selection, **conservation**, ...



Expected Heterozygosity = 0.50



Expected Heterozygosity << 0.50

Population subdivision

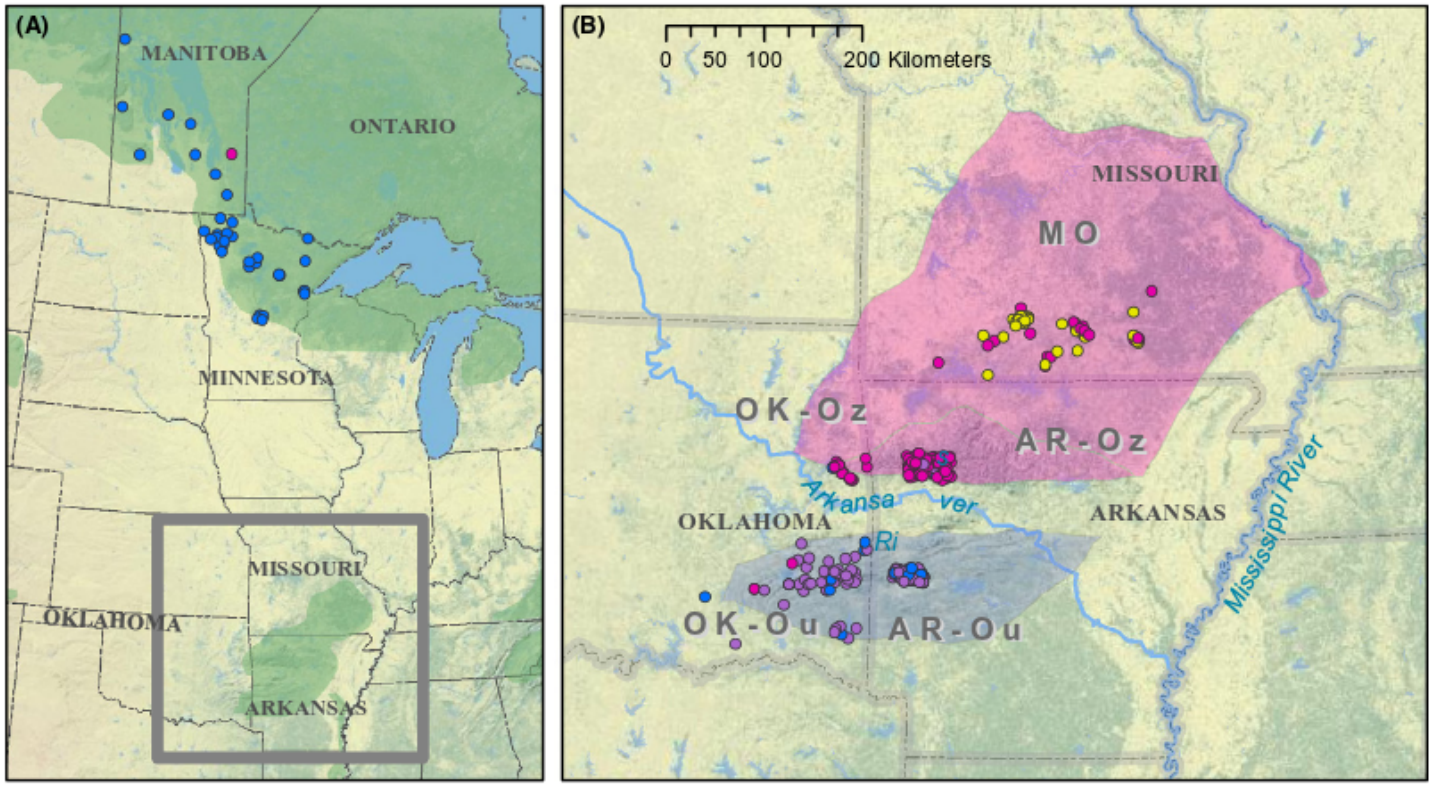


Fig. 1 (A) Map of American black bear distribution (green) in central North America with names of the states (Oklahoma, OK; Arkansas, AR; Missouri, MO; Minnesota, MN) and provinces (Manitoba, MB; Ontario, ON) discussed in this article. The grey box delineates the inset map (B) with the Ouachita (OU, purple) and Ozark (OZ, pink) land features and the five study areas of the CIH: OK-Ou, AR-Ou, OK-Oz, AR-Oz and MO. Sampling locations were coloured according to assignment of the four genetic clusters identified in Fig. 3.

"... subpopulations in OU and OZ cross state boundaries. Collaborative management among state agencies could help ensure that gene flow continues within and among subpopulations and future barriers (e.g. roads) are avoided or mitigated."

Puckett et al. Mol Ecol 2014

F_{ST}

Common measure for quantifying population subdivision.

$$F_{ST} = (H_T - H_S) / H_T$$

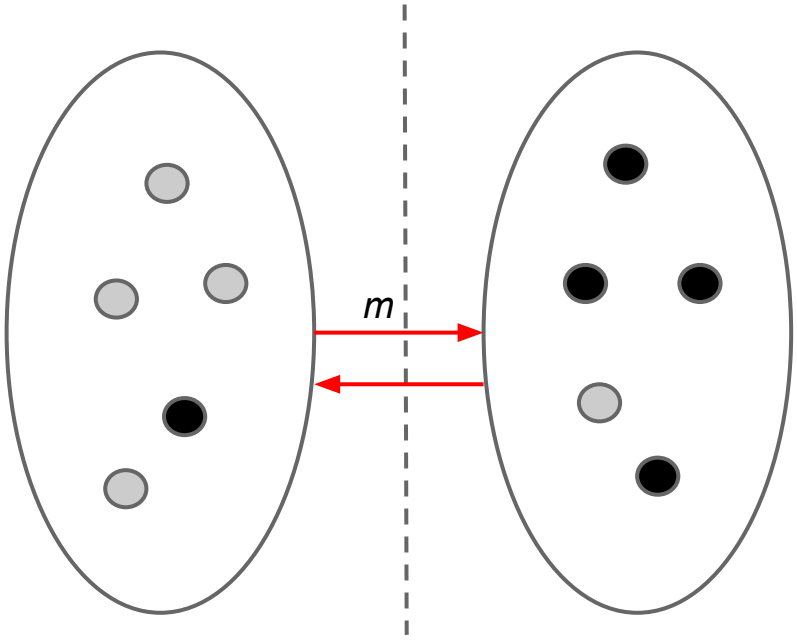
H_T: total (pooled) allele frequency from all populations

H_S: average across all populations

- if $H_T=H_S$ then $F_{ST}=0$
- if $H_S=0$ then $F_{ST}=1$

Migration rates

Two populations may be separated* but occasionally **migrants** may move and exchange genetic material.

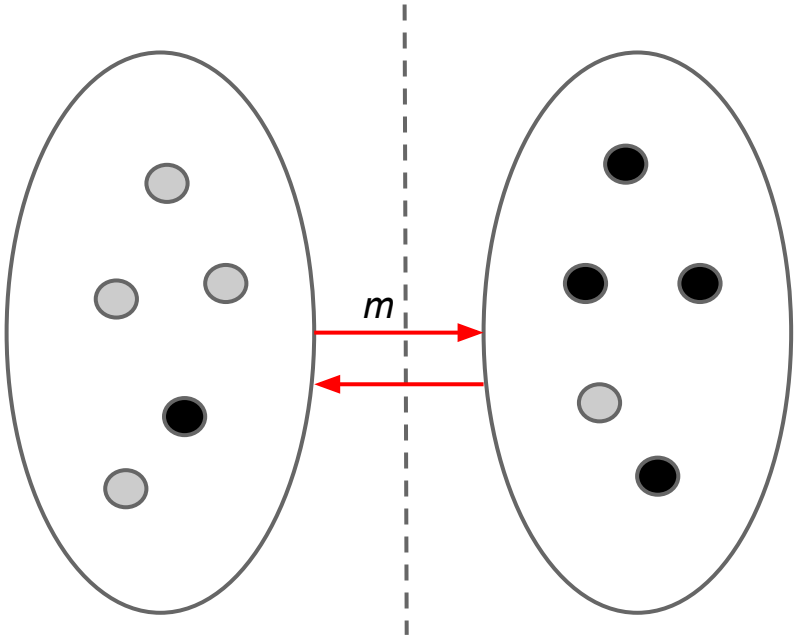


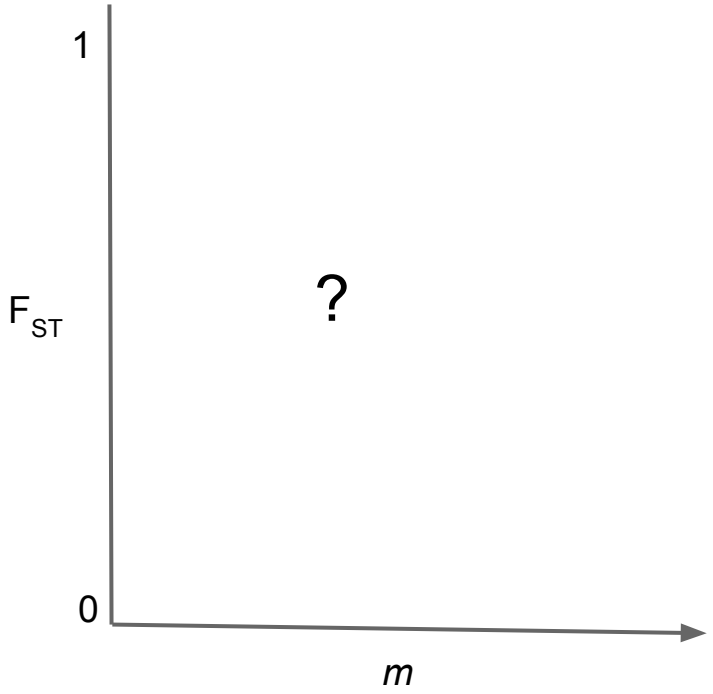
m = probability that an individual from one population is replaced with an individual from the other (per individual, per generation)

* island model: population subdivided for a long time

Migration rates

Two populations may be separated but occasionally **migrants** may move and exchange genetic material.

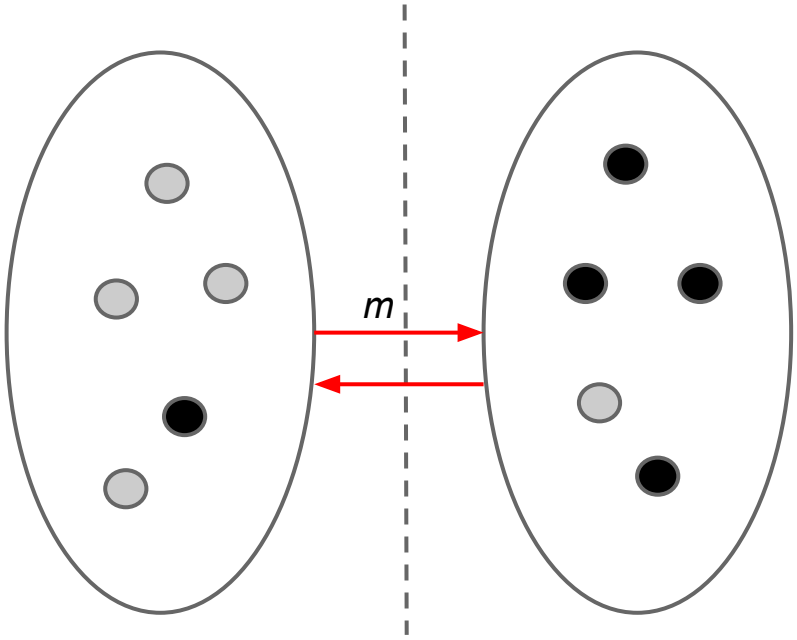


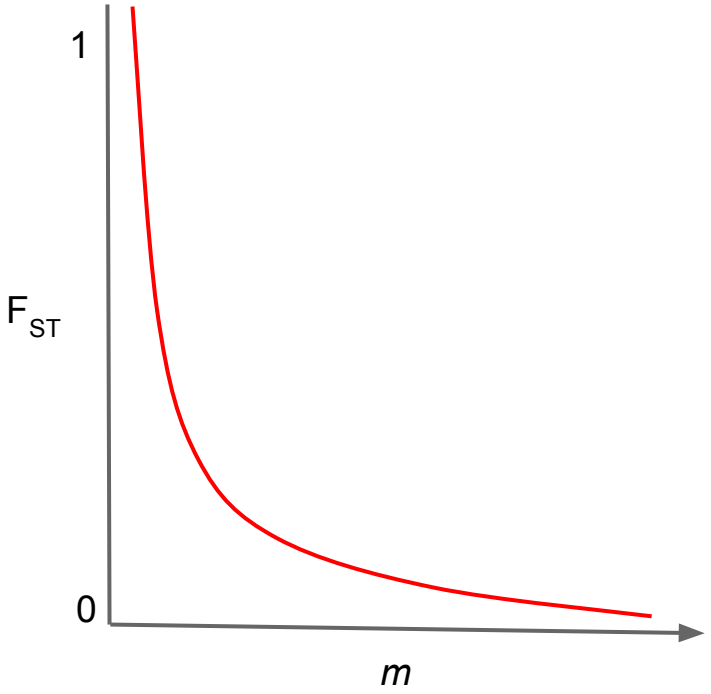


m = probability that an individual from one population is replaced with an individual from the other (per individual, per generation)

Migration rates

Two populations may be separated but occasionally **migrants** may move and exchange genetic material.

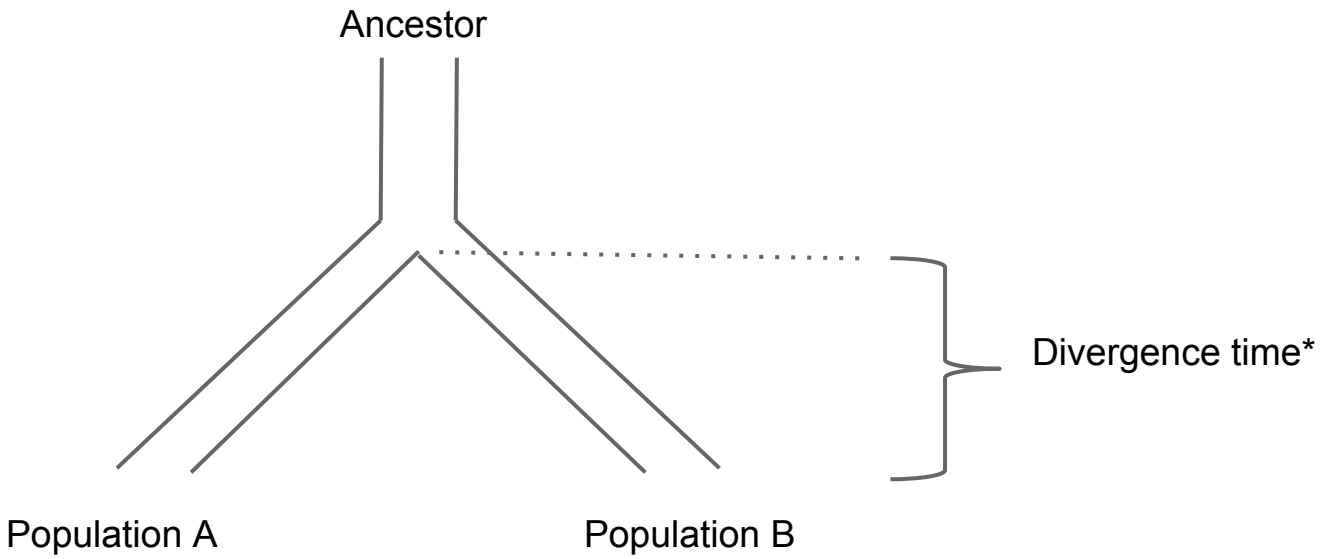




m = probability that an individual from one population is replaced with an individual from the other (per individual, per generation)

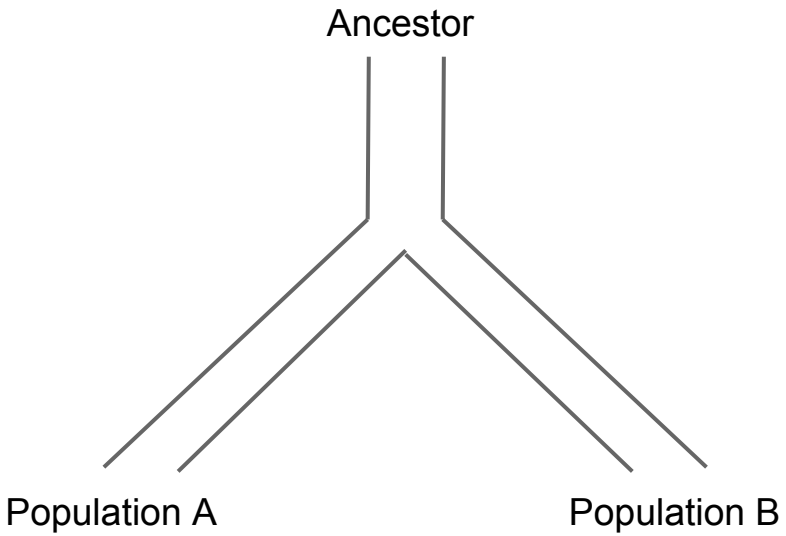
Divergence

Two populations share a common ancestor.



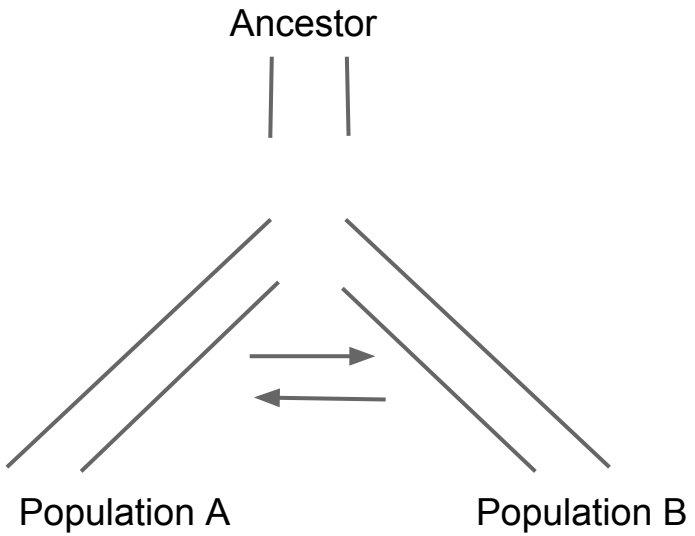
* under certain assumptions, the divergence time (in generations) can be approximated by $-\log(1-F_{ST})$

Divergence with (early/late/continuous) migration



Model 1

Estimated divergence time: T_1

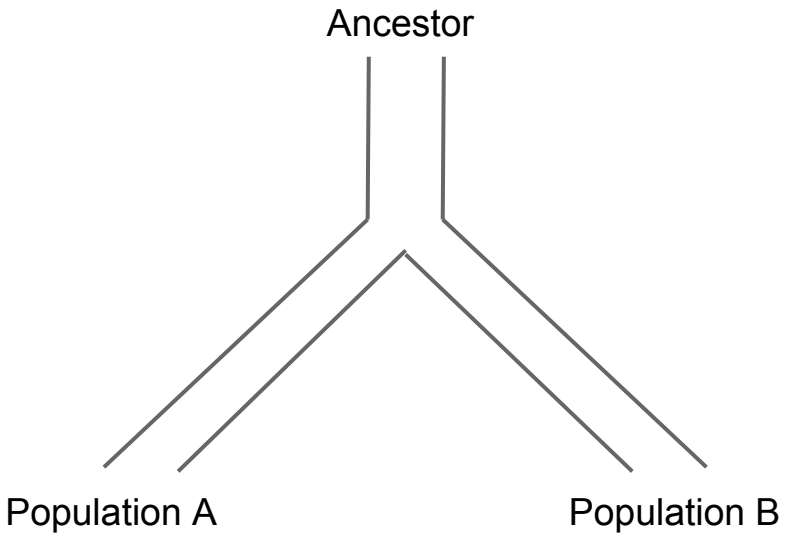


Model 2

Estimated divergence time: T_2

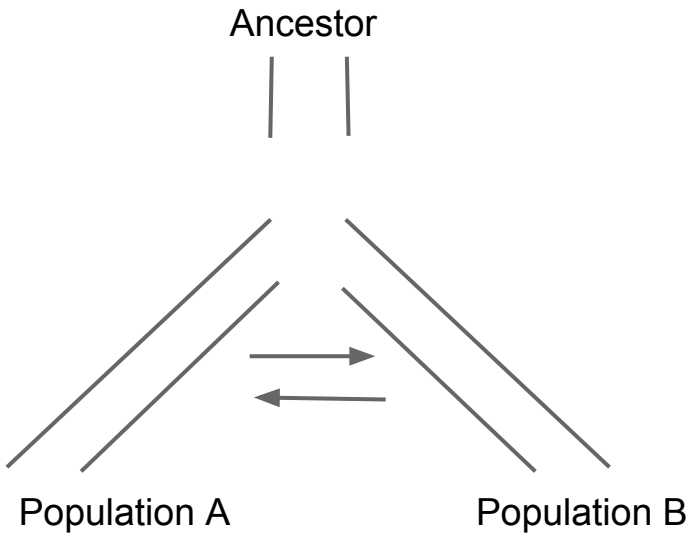
$T_1 <=> T_2$?

Divergence with (early/late/continuous) migration



Model 1

Estimated divergence time: T_1

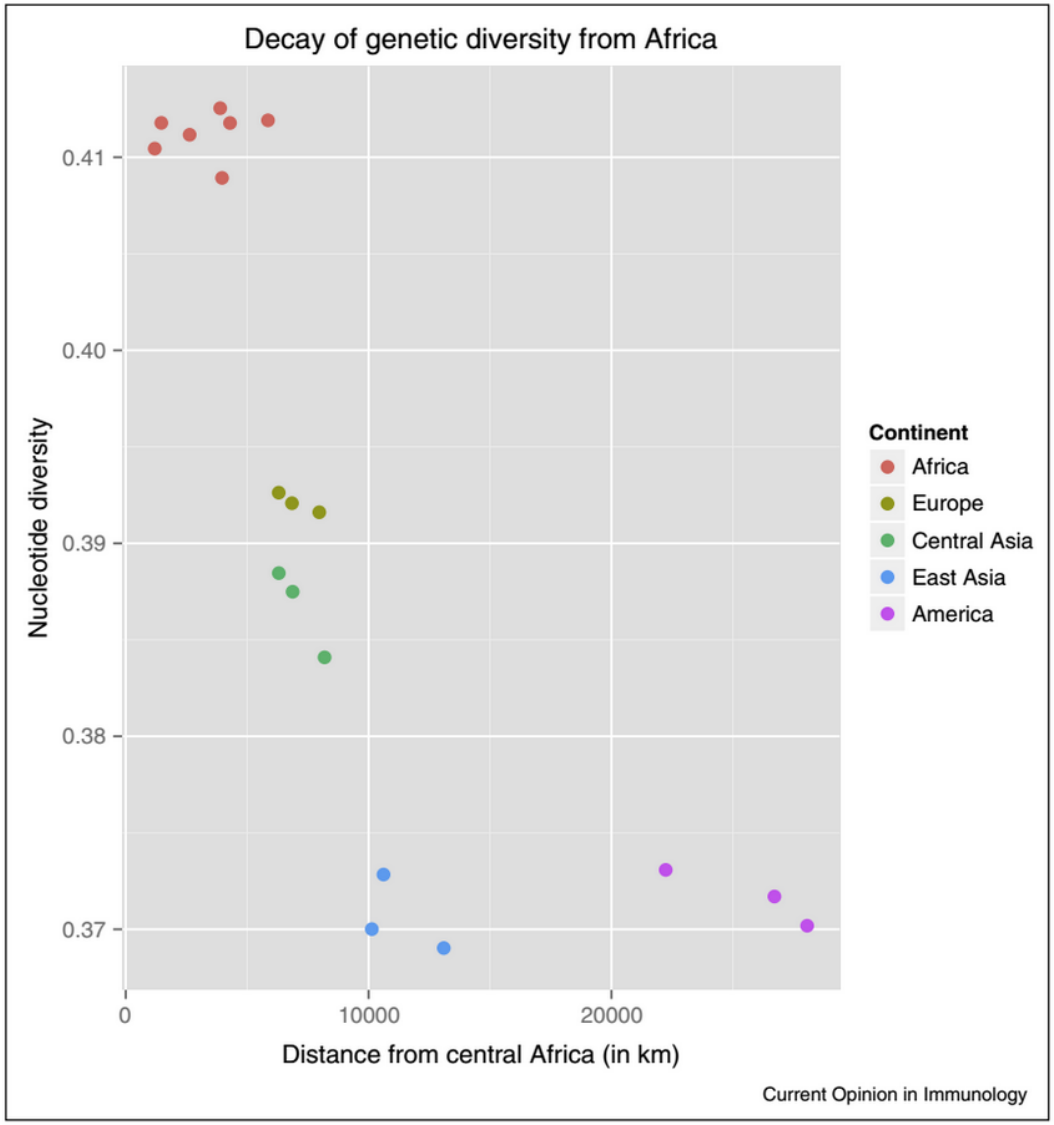


Model 2

Estimated divergence time: T_2

$$T_1 < T_2$$

Isolation by distance



Fumagalli and Sironi. 2014

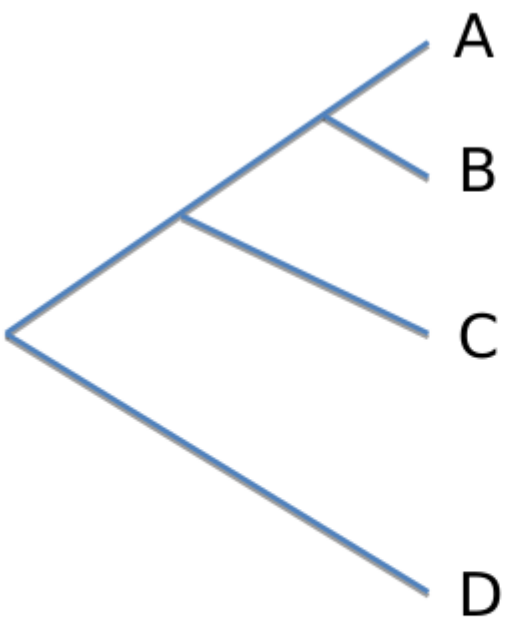
Representing population structure

Analysis of patterns of population subdivision for:

- samples clustering
- population assignment
- detection admixture events
- estimation of demographic parameters

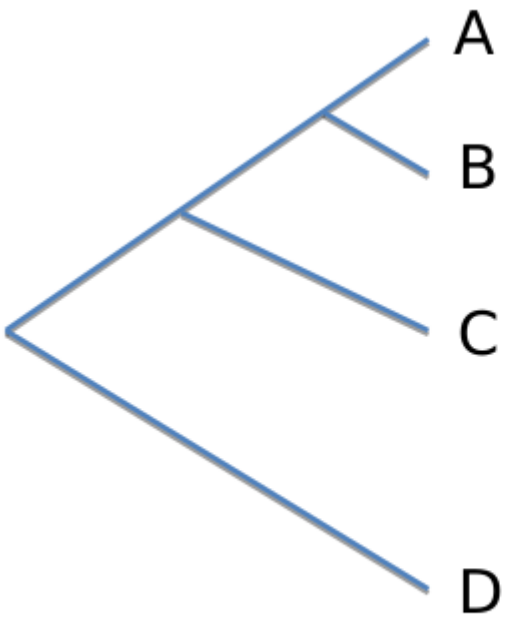
Most of these analyses are **descriptive** rather than quantitative!

Genetic distances



Genotype 1	Genotype 2	Distance
aa	aa	0
aa	aA	1
aa	AA	2
aA	aa	1
aA	aA	0
aA	AA	2
…	…	…

Genetic distances



Genotypes are {aa, aA, AA} as {0, 1 ,2}

For individuals i and j and N sites:

$$d(i,j) = -\log\left(1 - \frac{1}{N}\sum_{s=1}^{N}\frac{|g(i,s) - g(j,s)|}{2}\right)$$

genotype of *i* at site *s*

Genetic distances - case study

Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes

Xun Xu[1–3,12], Xin Liu[2,12], Song Ge[4,12], Jeffrey D Jensen[5,12], Fengyi Hu[6,12], Xin Li[1,12], Yang Dong[1,12], Ryan N Gutenkunst[7], Lin Fang[2], Lei Huang[3,4], Jingxiang Li[2], Weiming He[2,8], Guojie Zhang[1,2,4], Xiaoming Zheng[3,4], Fumin Zhang[3], Yingrui Li[2], Chang Yu[2], Karsten Kristiansen[2,9], Xiuqing Zhang[2], Jian Wang[2], Mark Wright[10], Susan McCouch[10], Rasmus Nielsen[1,9,11], Jun Wang[2,9] & Wen Wang[1]

Table 1 Summary of sequencing and variations for cultivated and wild rice

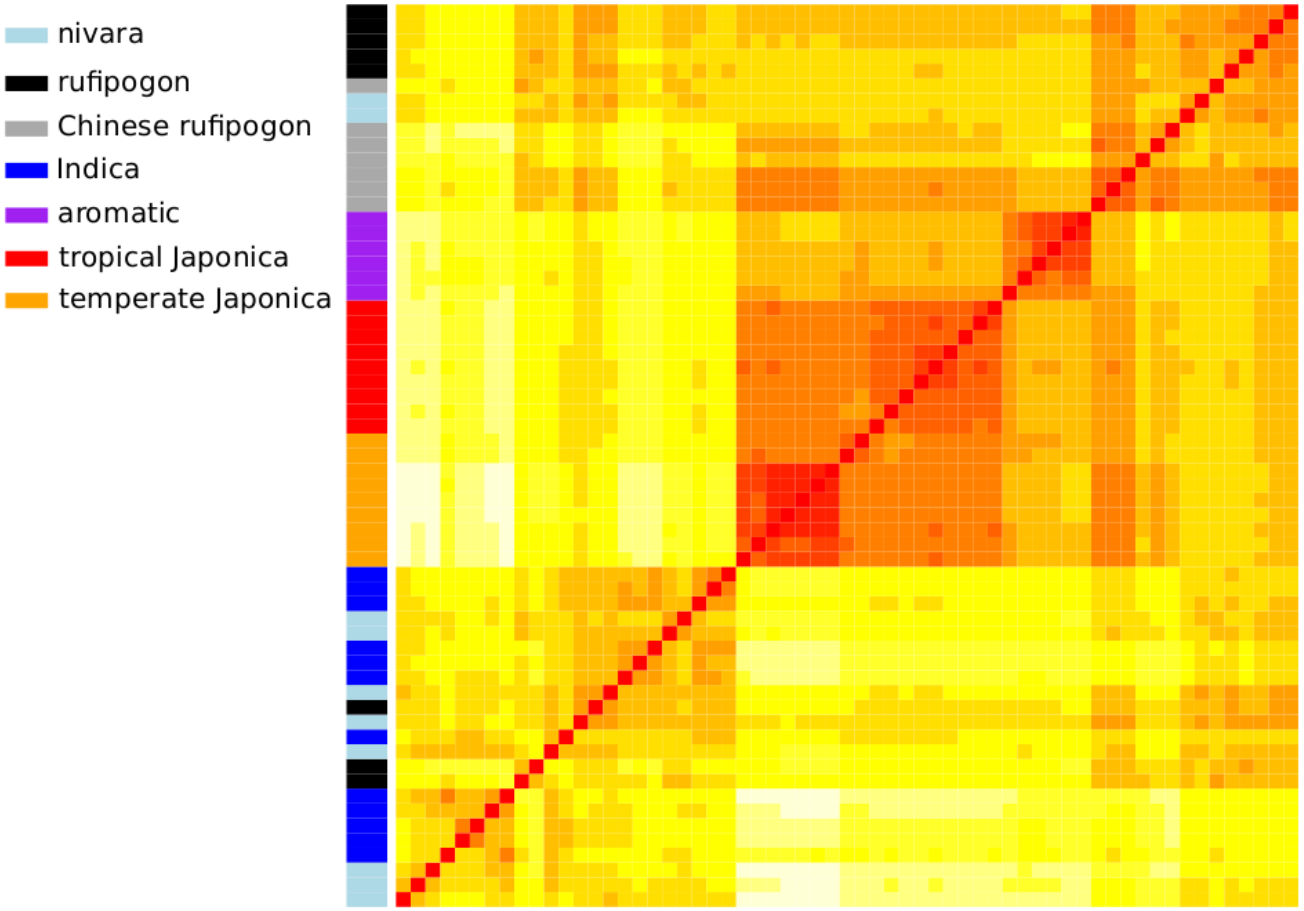
Group	Sample size	Raw data (Gbp)	Raw data depth	Uniquely mapping bases (Gbp)	Mean depth	Mean depth in gene region	SNP (M)	Indel (K)	SV (K)[a]
Total	50	319.2	18.7	182.4	11.8	14.3	6.5	808	94.7
Cultivar total	40[b]	249.8	18.7	163.0	12.2	14.5	4.4	612	62.1
Indica	12	82.5	18.5	47.4	10.6	13.3	3.0	441	38.7
AUS	2	13.4	18.0	7.9	10.6	13.0	0.8	183	17.4
IND	10	69.1	18.6	39.5	10.5	13.4	2.9	414	24.3
Japonica	24	167.3	18.7	115.6	12.9	15.1	2.5	355	28.5
ARO	6	45.6	20.4	29.7	13.3	15.9	1.1	183	15.3
TEJ	8	53.2	17.9	39.5	13.3	15.2	1.1	136	4.3
TRJ	10	68.5	18.4	46.4	12.5	14.6	1.5	208	11.1
Wild total	10	69.4	18.7	39.4	10.6	13.3	5.2	682	40.4
O. rufipogon	5	34.0	18.3	19.2	10.3	12.9	3.5	424	21.1
O. nivara	5	35.4	19.1	20.2	10.9	13.7	3.1	439	21.7

[a]Structural variation (insertion or deletion longer than 100 bp) relative to Nipponbare; the structural variations in the same region in different accessions count as one. [b]Only 36 cultivar accessions can be clearly put in a subgroup, and the other four accessions have admixed genetic backgrounds, as indicated in **Supplementary Table 1**.

Xu et al. Nature Biotech. 2012

Genetic distances - case study

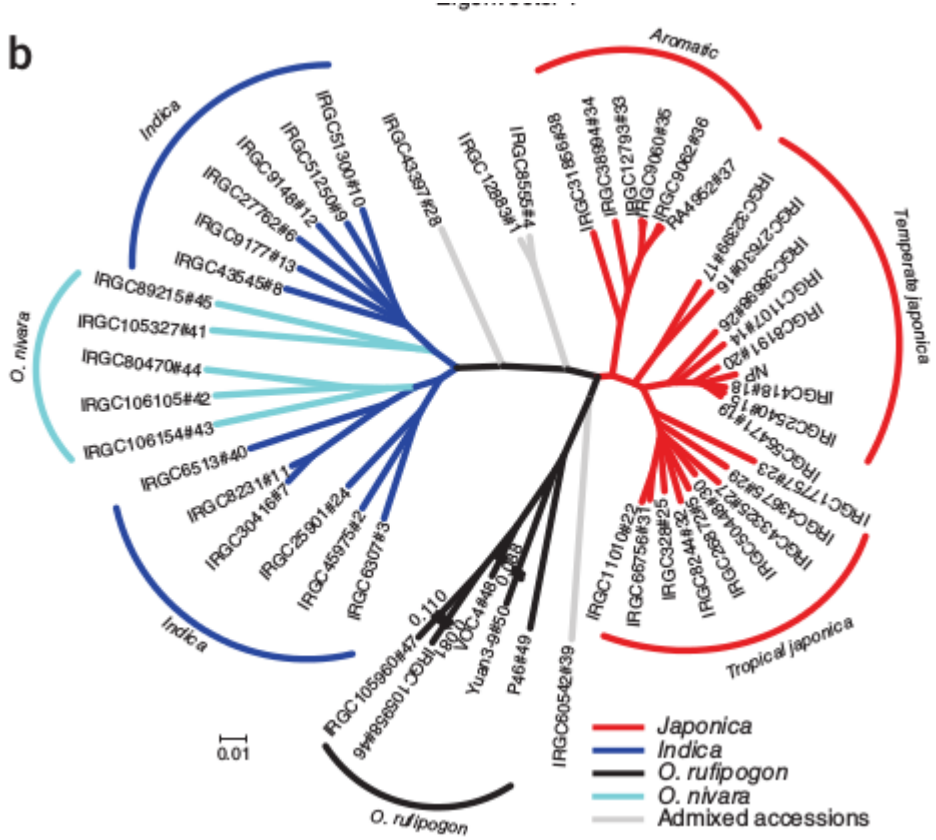
Clustering of samples to separate groups.



Vieira et al. 2015 in press

Genetic distances - case study

Graphical representation of genetic distance in form of a tree*.



* this should not be considered as a proper phylogenetic tree

Xu et al. Nature Biotech. 2012

PCA

Principal Component Analysis (PCA) is a statistical method to reduce data dimensionality.

In genetics, it summarises genotypes' relationships across all individuals into uncorrelated Principal Components (PCs, or eigenvectors).

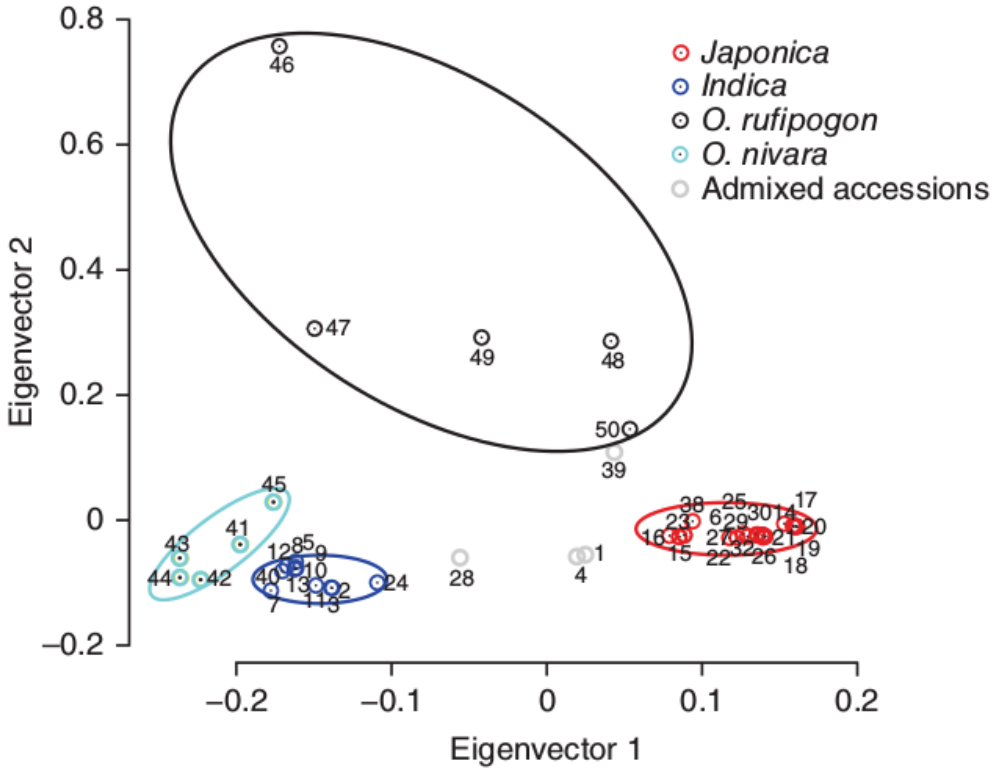
For n individuals and m sites a normalized covariance matrix C is calculated as

$$C_{(w,y)} = \frac{1}{m} \sum_{s=1}^{m} \frac{\left(G_{(w,s)} - 2\hat{p}_s\right)\left(G_{(y,s)} - 2\hat{p}_s\right)}{\hat{p}_s(1-\hat{p}_s)}$$

where $\widehat{p_i}$ is the derived allele frequency at site s (the labeling is again arbitrary) and $G_{(w,s)}$ is the number of derived alleles for individual w at site s ($G \in \{0, 1, 2\}$ in the diploid case).

PCA - case study

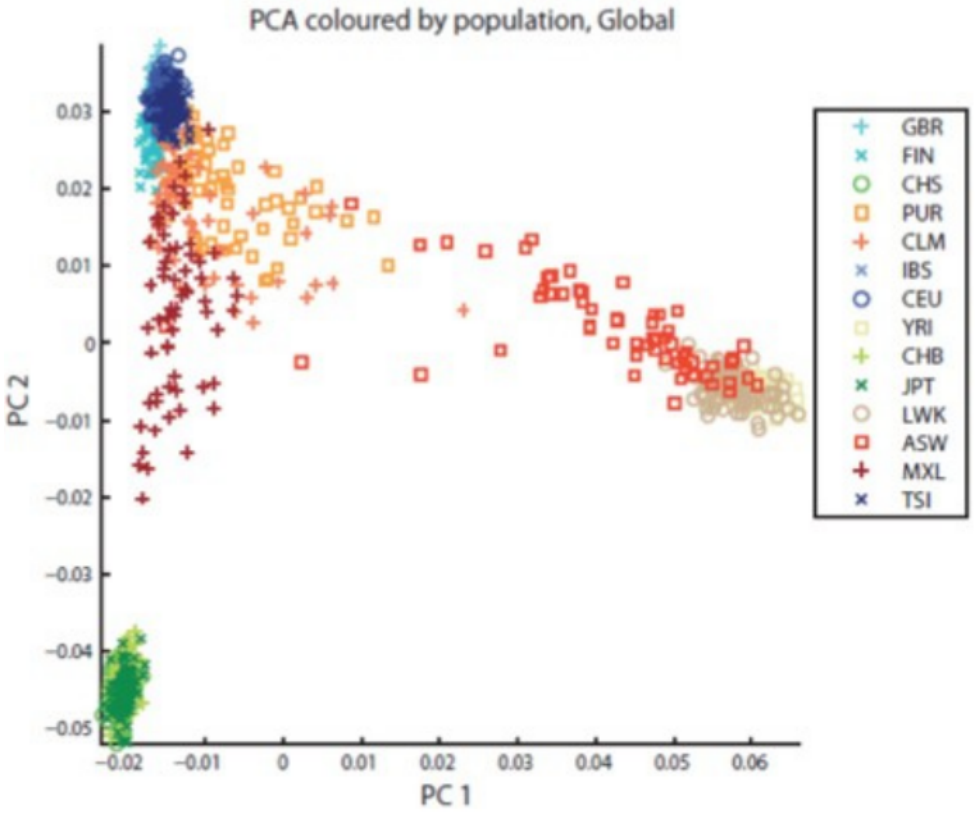
- An eigenvector decomposition of the covariance matrix is computed. Eigenvectors are then plotted.
- PCA is a **descriptive** analysis of your dataset, for clustering individuals (and identify population assignment).



Xu et al. Nature Biotech. 2012

PCA - case study

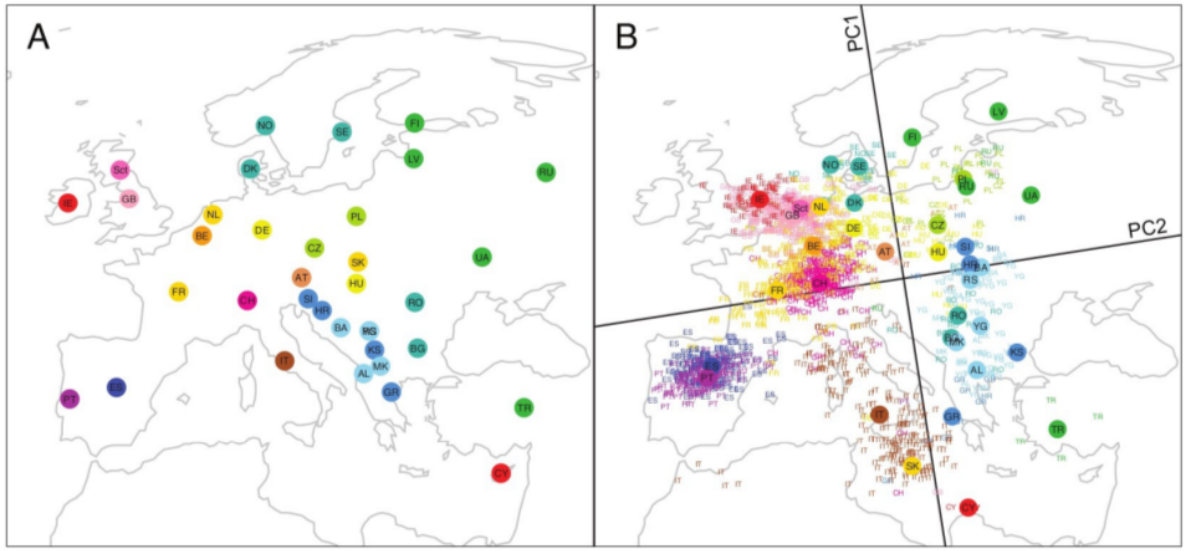
PCA should not be used to make inferences on historical events.



1000 Genomes Project Consortium. 2012

Other analyses

- Multi Dimensional Scaling (**MDS**) works from a matrix of similarities
- Discriminant Analysis of Principal Components (**DAPC**) finds clusters
- **Procrustes analysis** compares mapping coordinates of two (or more) data sets (e.g. genetic and geography)



Wang et al. 2012

Admixture analyses

For admixed populations: assign ancestry proportion to each individual from multiple (ancestral) populations.

$f_{A.1} = 0.6$

$f_{A.2} = 0.2$

AA: observed genotype

Pr(AA | pop=1) = ?

Pr(AA | pop=2) = ?

Admixture analyses

For admixed populations: assign ancestry proportion to each individual from multiple (ancestral) populations.

$f_{A.1} = 0.6$

$f_{A.2} = 0.2$

AA: observed genotype

Pr(AA | pop=1) = ?

Pr(AA | pop=2) = ?

Assuming HWE

P(AA)=f^2
P(AG)=2*f*(1-f)
P(GG)=(1-f)^2

Admixture analyses

For admixed populations: assign ancestry proportion to each individual from multiple (ancestral) populations.

$f_{A.1} = 0.6$

$f_{A.2} = 0.2$

$Pr(AA \mid pop=1) = f_{A.1}^2 / (f_{A.1}^2 + f_{A.2}^2) = 0.90$

$Pr(AA \mid pop=2) = f_{A.2}^2 / (f_{A.1}^2 + f_{A.2}^2) = 0.10$

Assuming HWE
P(AA)=f^2
P(AG)=2*f*(1-f)
P(GG)=(1-f)^2

$Pr(AG \mid pop=1) = ?$

$Pr(AG \mid pop=2) = ?$

Admixture analyses

For admixed populations: assign ancestry proportion to each individual from multiple (ancestral) populations.

$f_{A.1} = 0.6$

$f_{A.2} = 0.2$

$Pr(AA \mid pop=1) = f_{A.1}^2 / (f_{A.1}^2 + f_{A.2}^2) = 0.90$

$Pr(AA \mid pop=2) = f_{A.2}^2 / (f_{A.1}^2 + f_{A.2}^2) = 0.10$

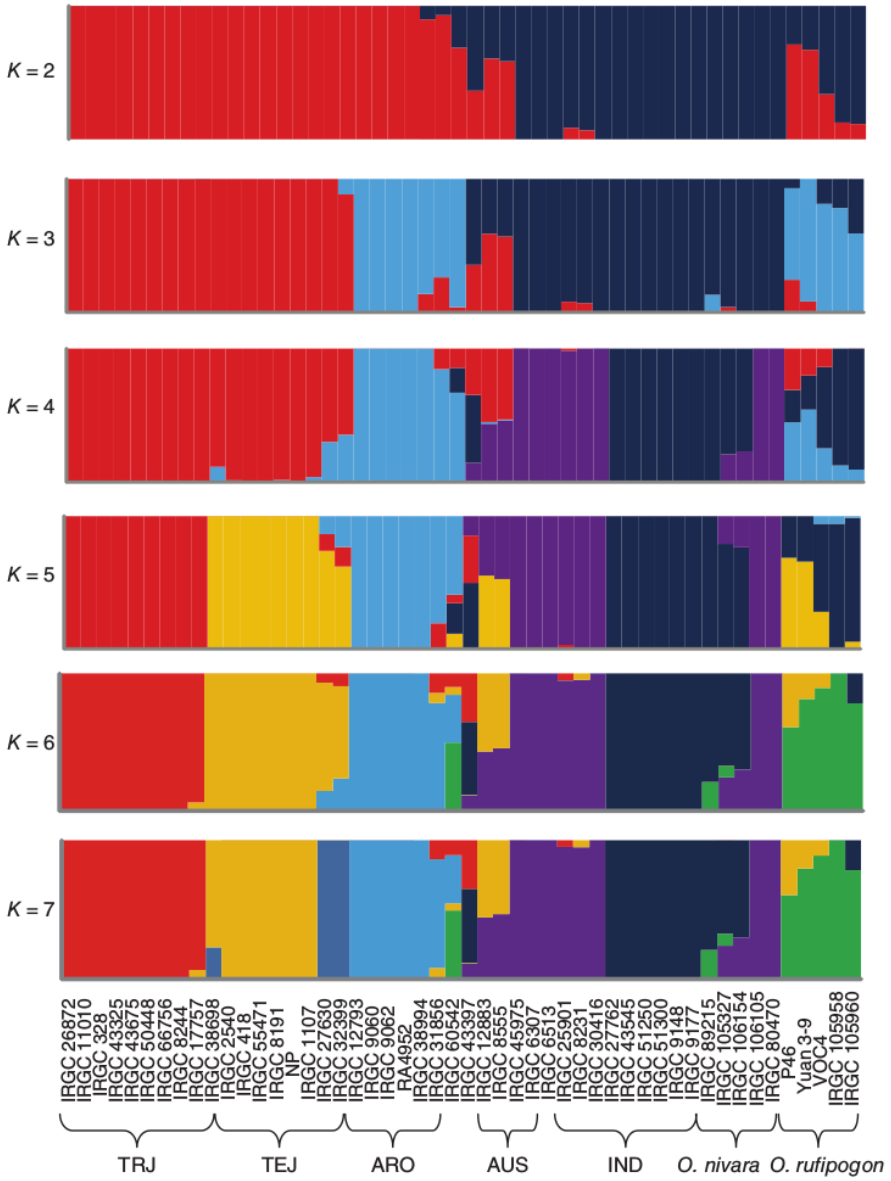
$Pr(AG \mid pop=1) = 2*f_{A.1}*(1-f_{A.1}) / ... = 0.60$

$Pr(AG \mid pop=2) = ... = 0.40$

Assuming HWE
P(AA)=f^2
P(AG)=2*f*(1-f)
P(GG)=(1-f)^2

Bayesian (STRUCTURE) or Maximum Likelihood (ADMIXTURE) approaches.

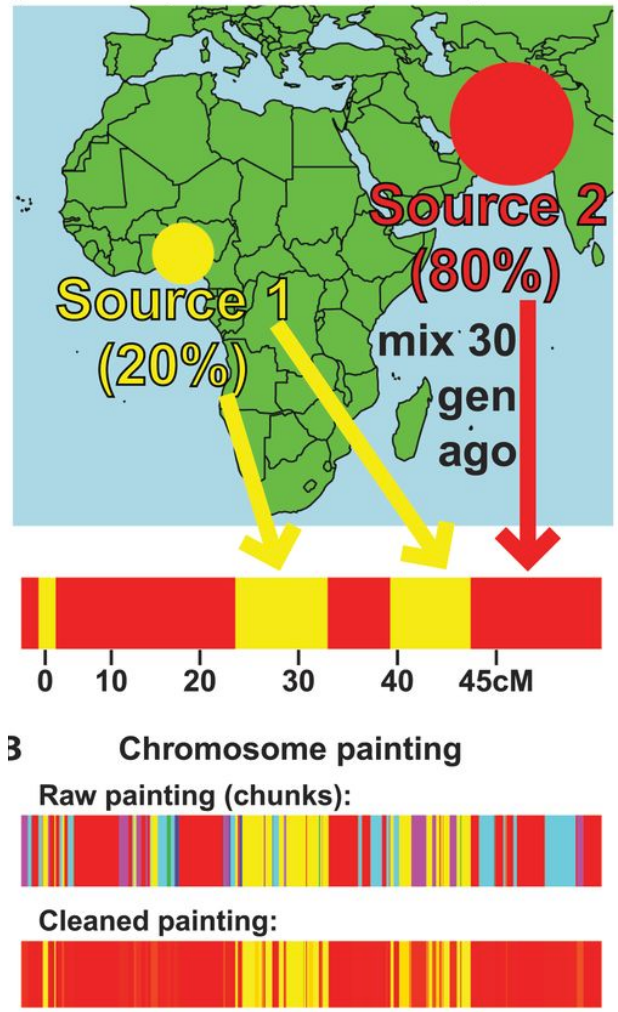
Admixture analyses



Xu et al. Nature Biotech. 2012

Admixture analyses

Local ancestry reconstruction

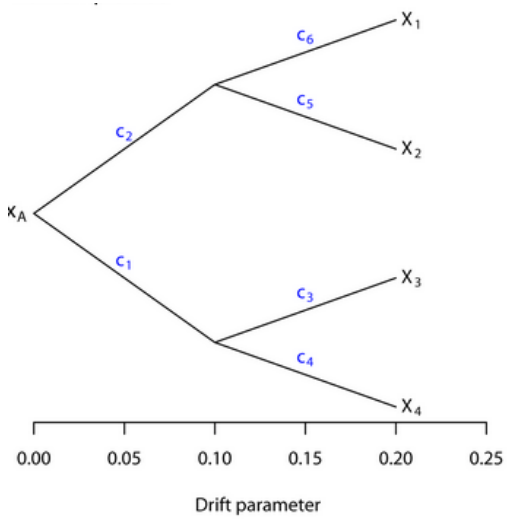


- Identify regions with excess of ancestry from a particular source population
- Demographic inferences

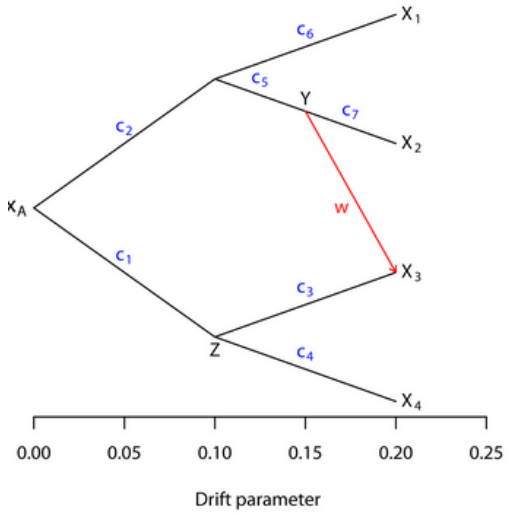
Hellenthal et al. 2014

TreeMix

Statistical model to infer population splits and admixtures in multiple populations.



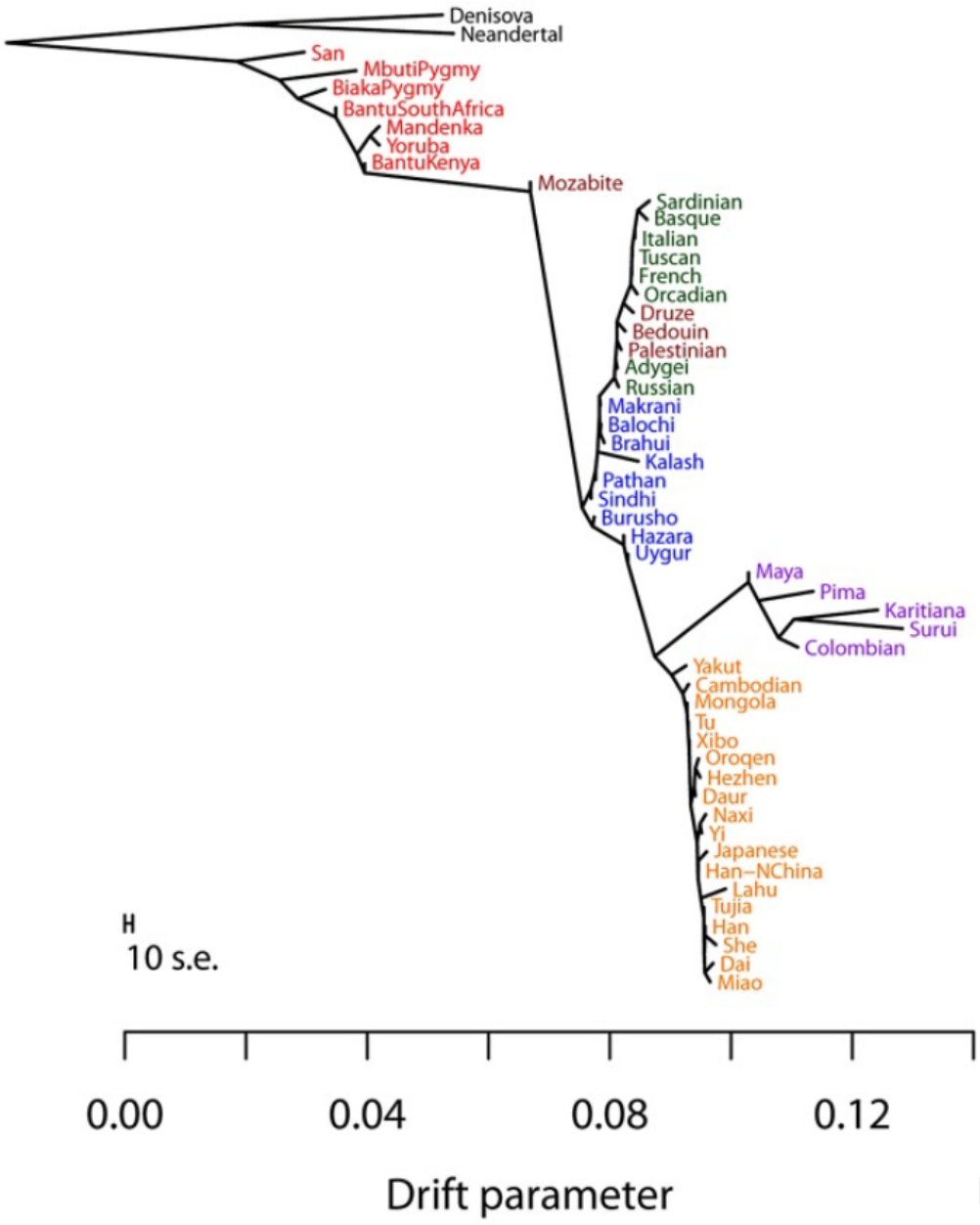
No migration events



1 migration event

Pickrell and Pritchard. 2012

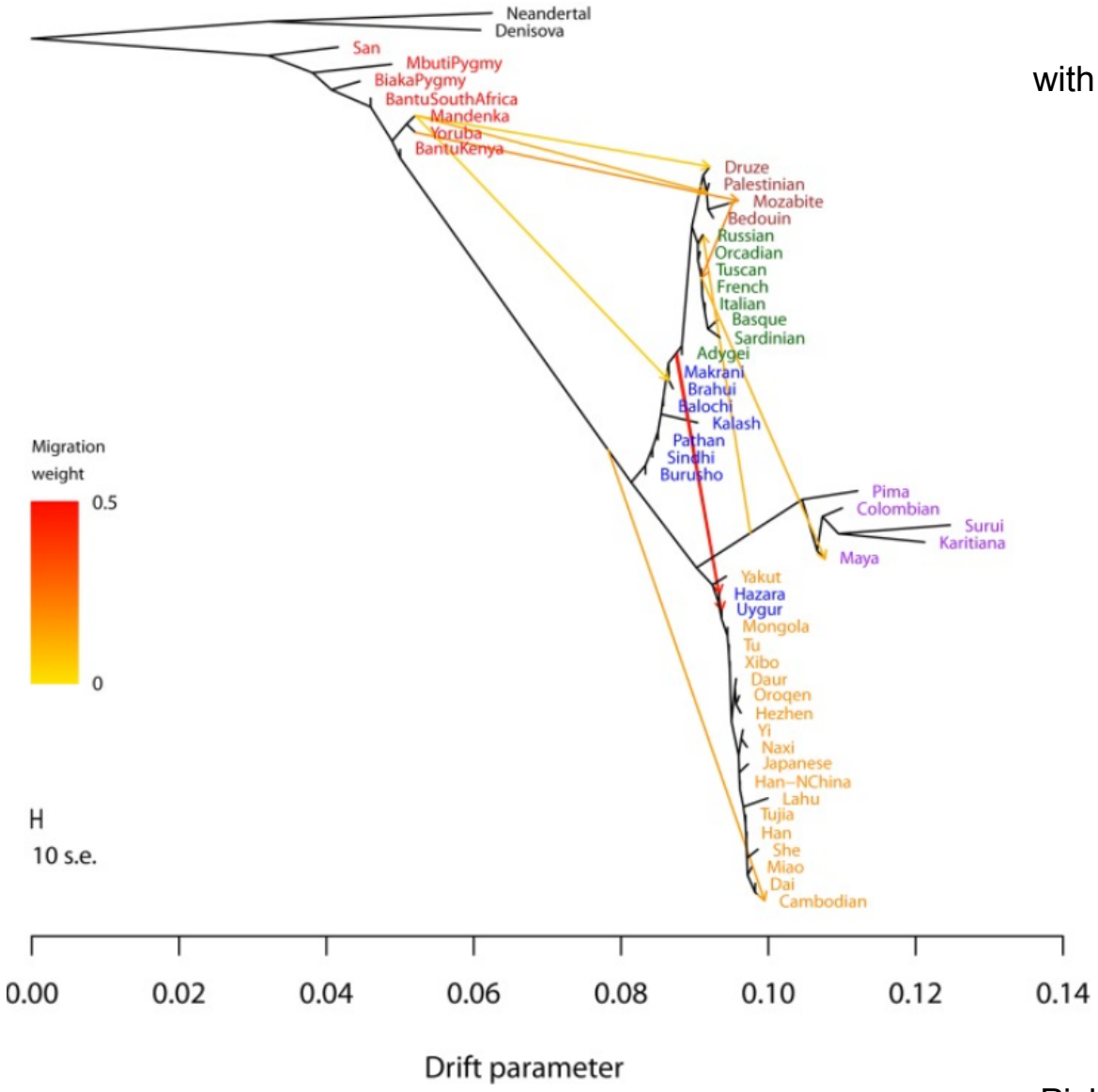
TreeMix



No migration events

Pickrell and Prichtard. 2012

TreeMix

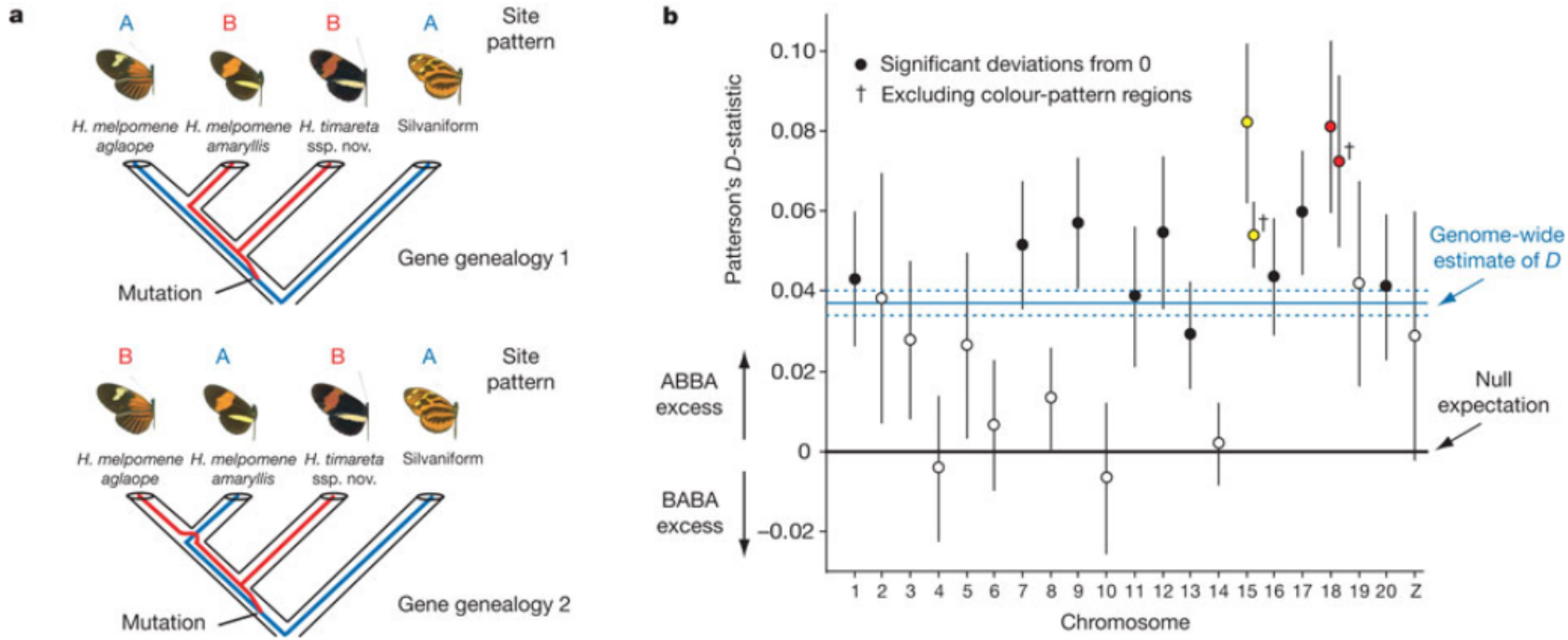


with migration events

Pickrell and Pritchard. 2012

Introgression / D-statistic

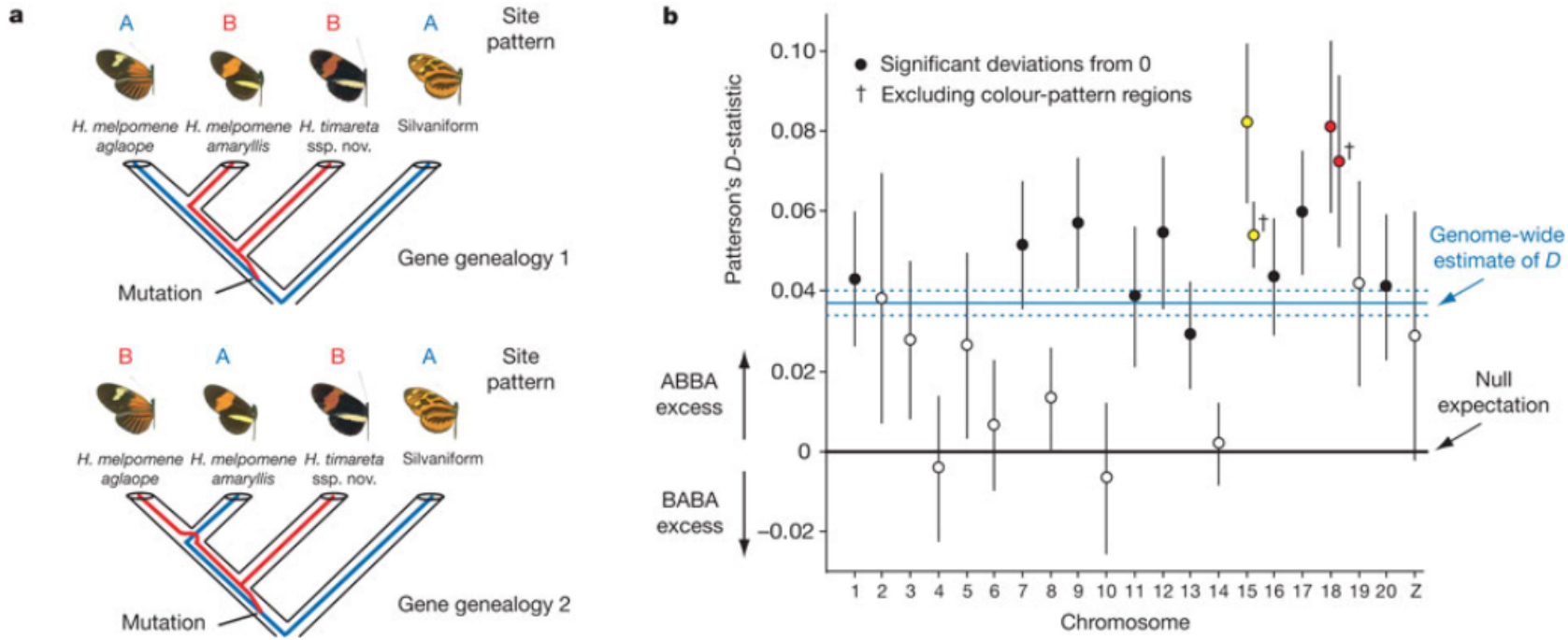
D ~ ABBA - BABA



Heliconius Genome Consortium. 2012

Introgression / D-statistic

D ~ ABBA - BABA



D>1 suggests an excess of “ABBA” sites that indicates introgression between “timareta” and “amaryllis”

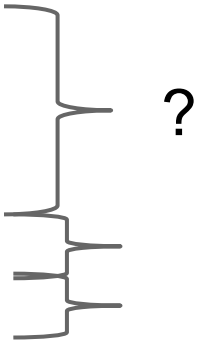
Heliconius Genome Consortium. 2012

Demographic inferences

Make use of **summary statistics** to characterise your data, and then use an analytical or simulations-based approach to infer demographic parameters of interest.

Summary statistics:

- allele frequency distribution
- nucleotide diversity
- haplotype diversity
- population genetic differentiation
- length of shared haplotypes
- ...

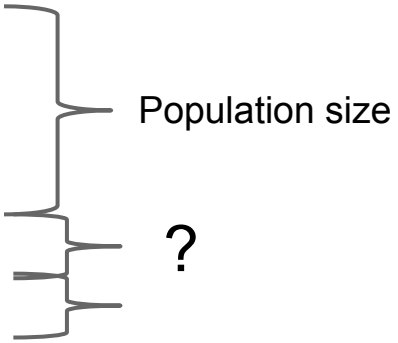


Demographic inferences

Make use of summary statistics to characterise your data, and then use an analytical or simulations-based approach to infer demographic parameters of interest.

Summary statistics:

- allele frequency distribution
- nucleotide diversity
- haplotype diversity
- population genetic differentiation
- length of shared haplotypes
- ...

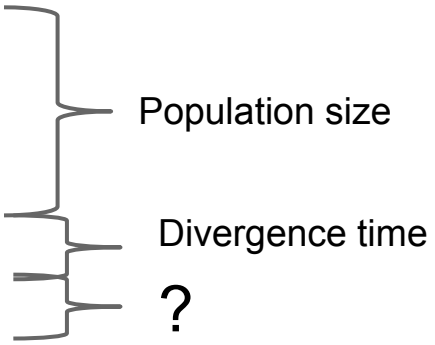


Demographic inferences

Make use of summary statistics to characterise your data, and then use an analytical or simulations-based approach to infer demographic parameters of interest.

Summary statistics:

- ❏ allele frequency distribution
- ❏ nucleotide diversity
- ❏ haplotype diversity
- ❏ population genetic differentiation
- ❏ length of shared haplotypes
- ❏ ...

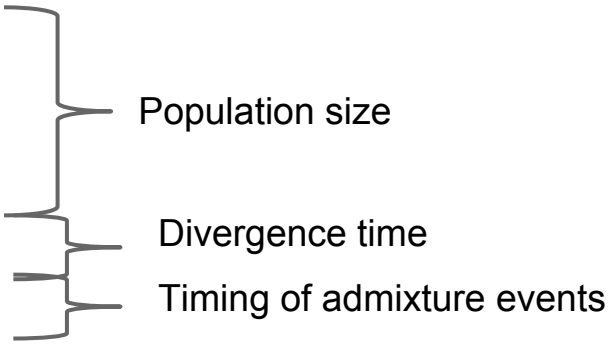


Demographic inferences

Make use of summary statistics to characterise your data, and then use an analytical or simulations-based approach to infer demographic parameters of interest.

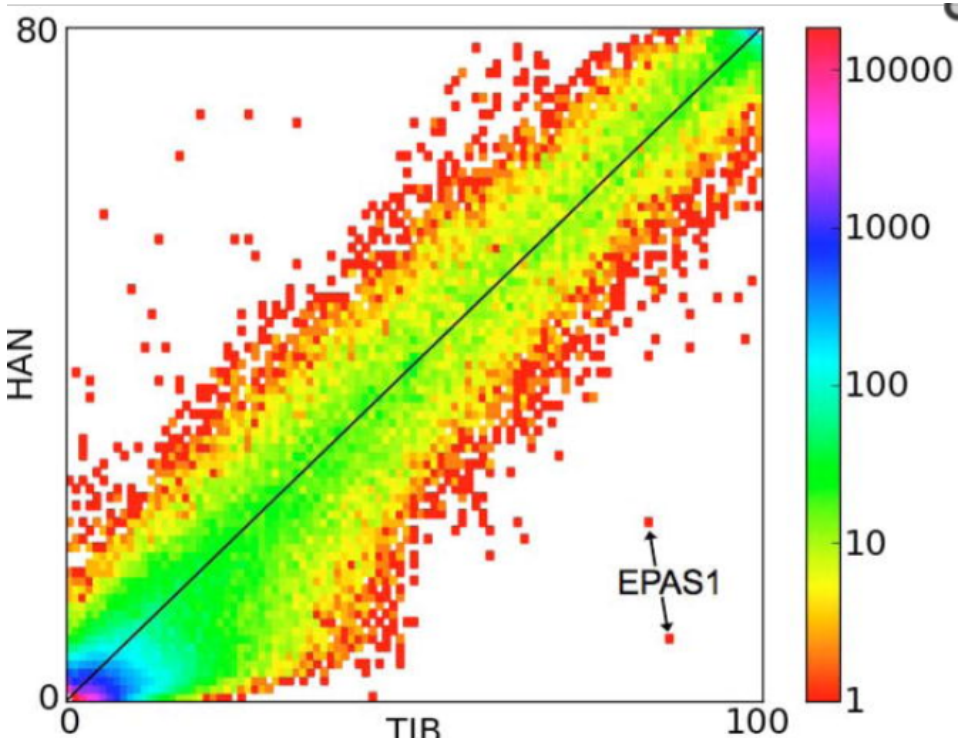
Summary statistics:

- ❏ allele frequency distribution
- ❏ nucleotide diversity
- ❏ haplotype diversity
- ❏ population genetic differentiation
- ❏ length of shared haplotypes
- ❏ ...



2D-SFS

The 2-dimension (or joint) site frequency spectrum (2D-SFS) shows the joint distribution of allele frequencies between 2 species.

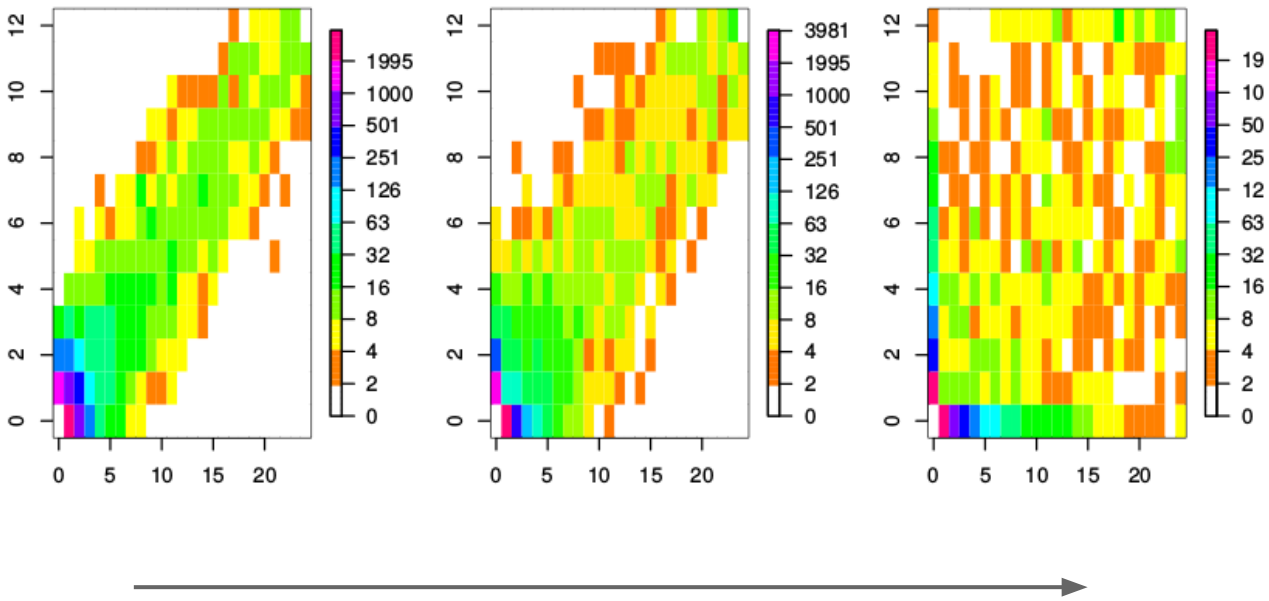


It is a valuable indicator for:

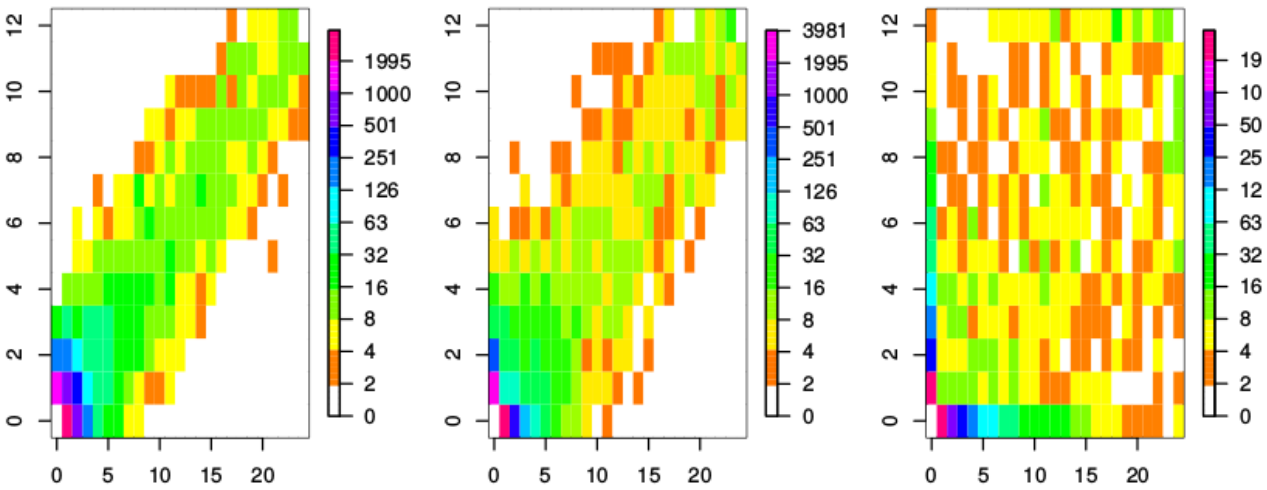
- ❏ population differentiation
- ❏ size changes
- ❏ migration rates
- ❏ ...

Yi et al. 2010

2D-SFS

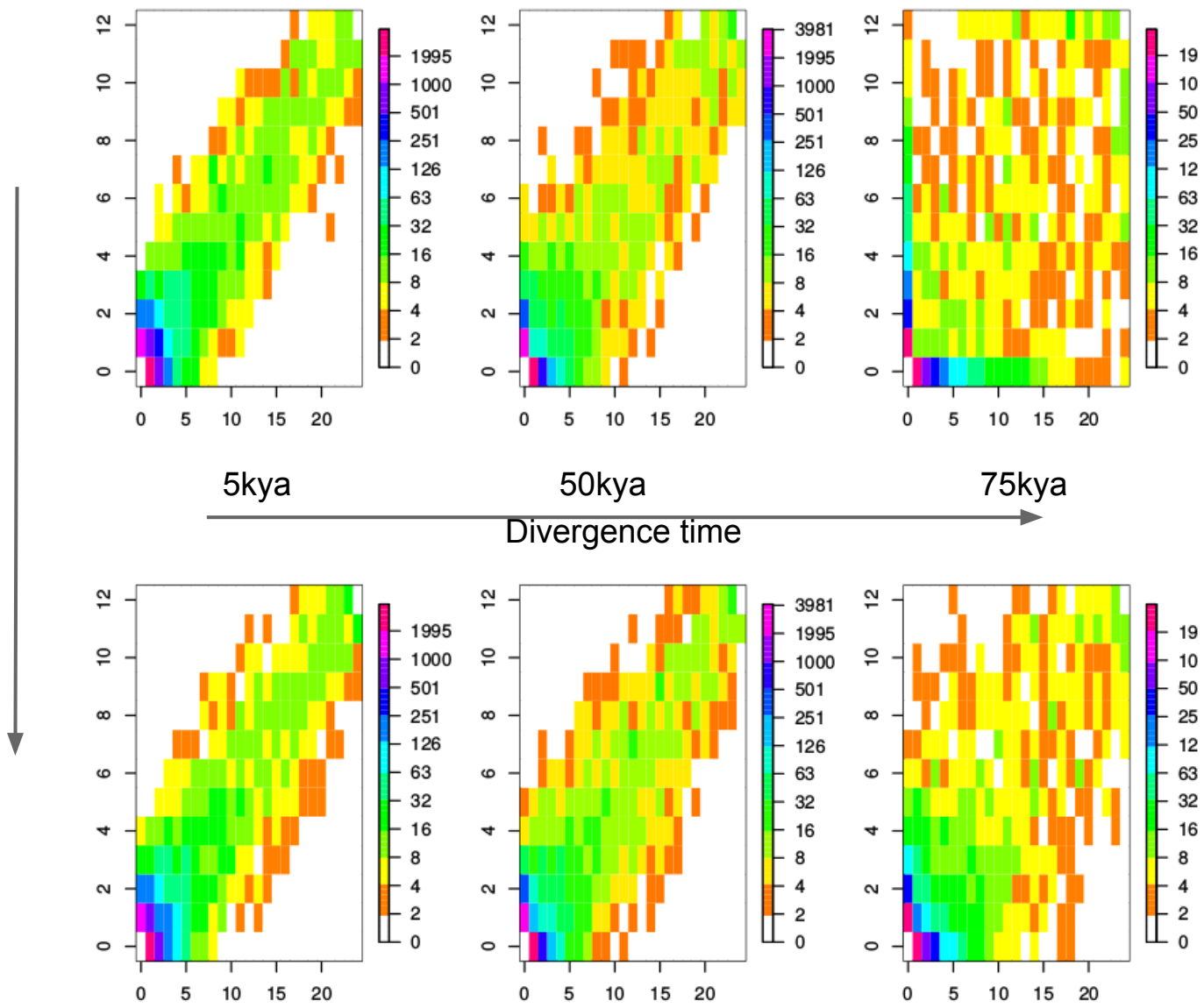

0 2 4 6 8 10 12
0 5 10 15 20
1995 1000 501 251 126 63 32 16 8 4 2 0
3981 1995 1000 501 251 126 63 32 16 8 4 2 0
19 10 50 25 12 63 32 16 8 4 2 0

2D-SFS

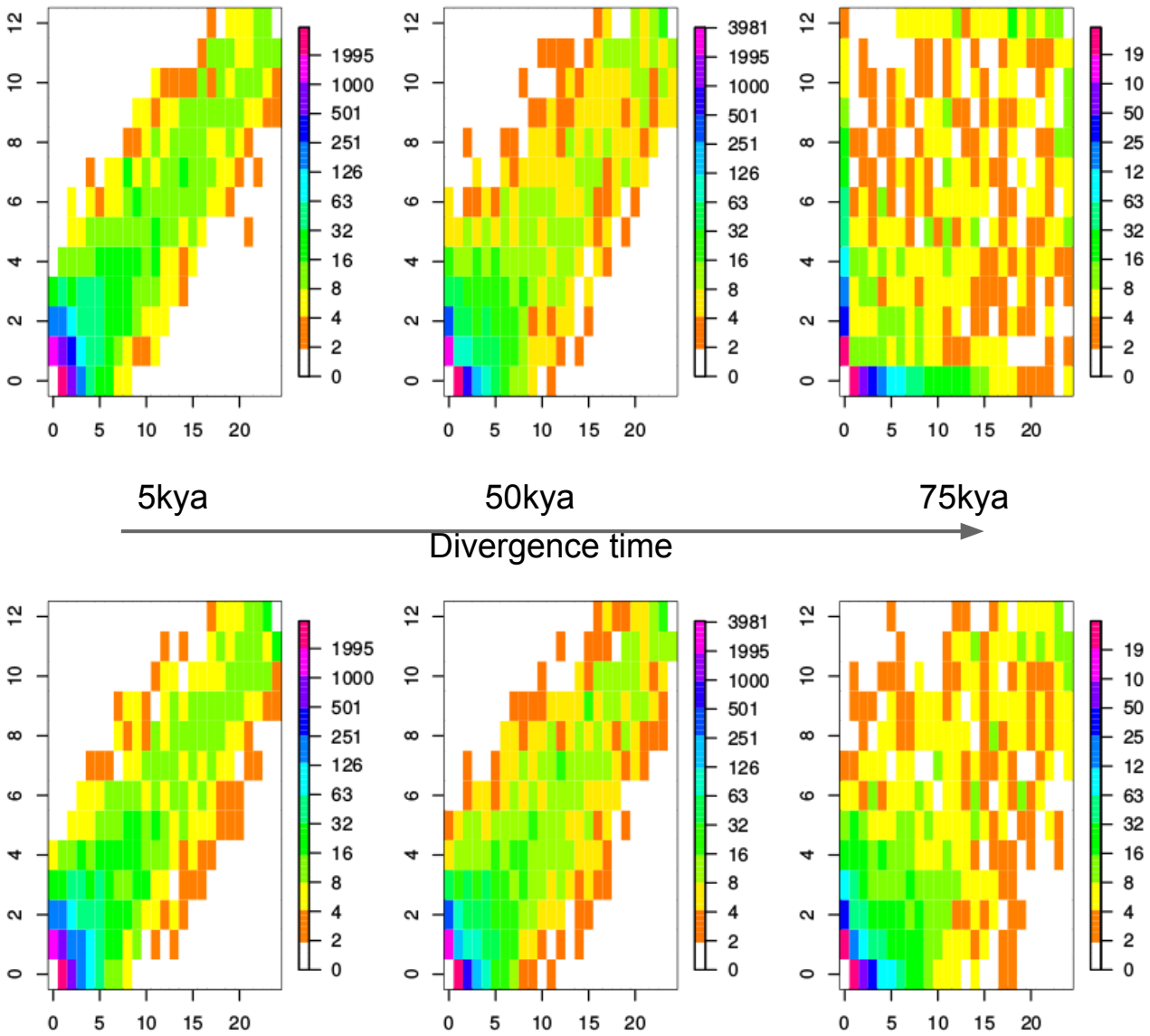


Divergence time

2D-SFS

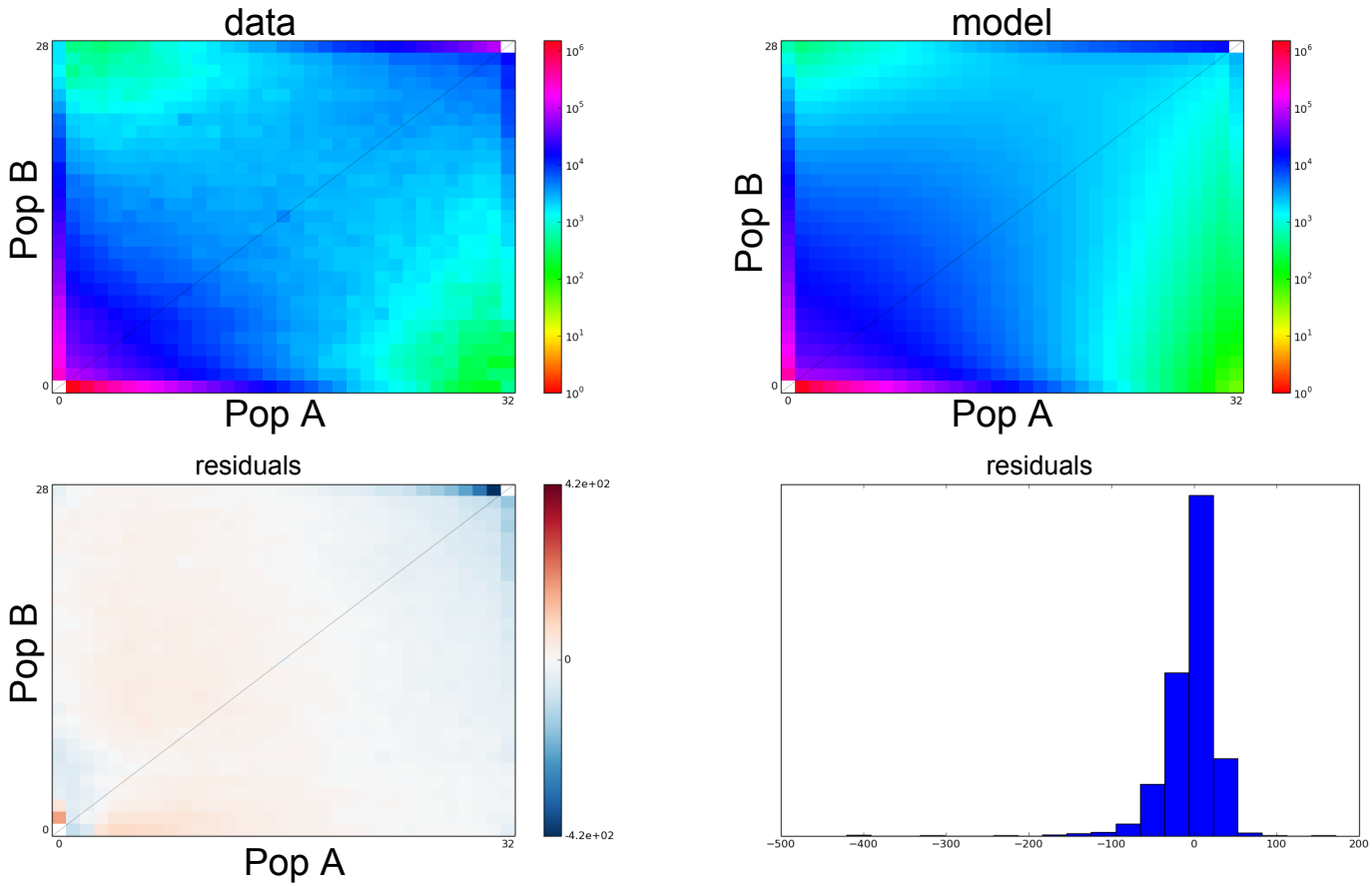


2D-SFS


Migration rate
5kya
50kya
75kya
Divergence time
0 5 10 15 20
0 2 4 6 8 10 12
1995 1000 501 251 126 63 32 16 8 4 2 0
3981 1995 1000 501 251 126 63 32 16 8 4 2 0
19 10 50 25 12 63 32 16 8 4 2 0

∂a∂i

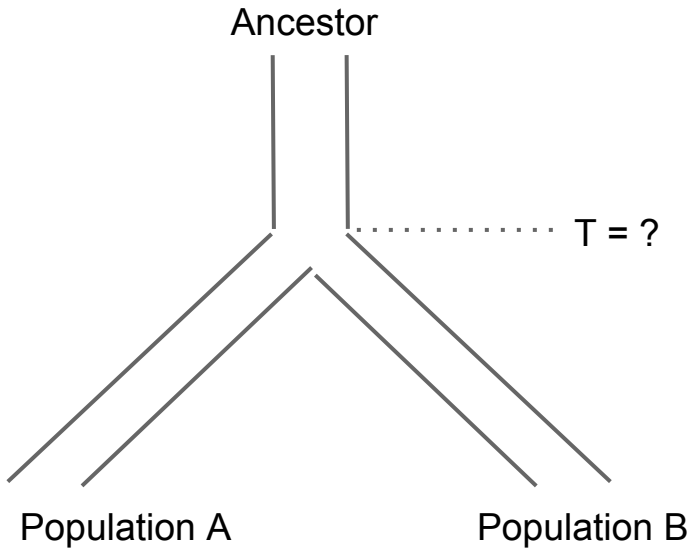
∂a∂i implements a method for demographic inference from genetic data, based on a diffusion approximation to the **allele frequency spectrum** for up to three simultaneous populations (no need to run simulations!).



Simulation-based approaches

Rejection algorithm:

1. simulate large numbers of data sets under an evolutionary model



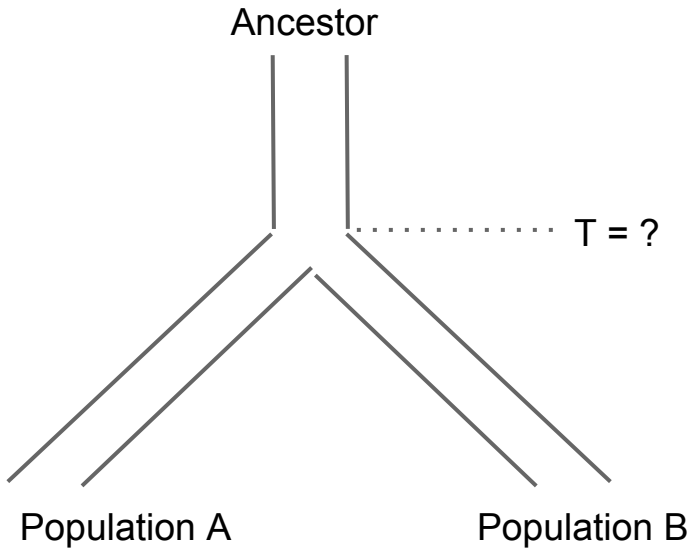
-> You need to specify a demographic model which will be used to simulate genetic data

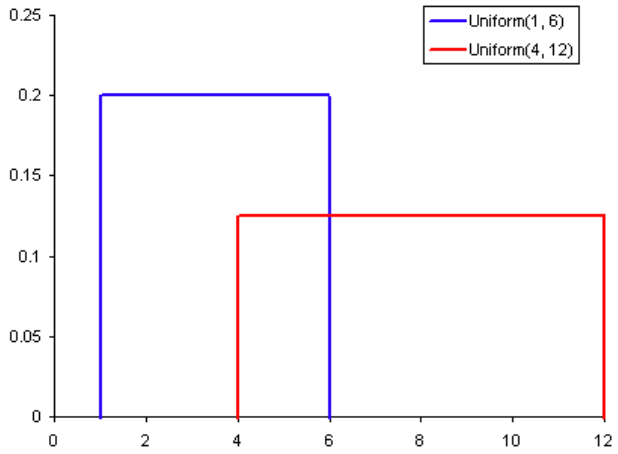
Simulation-based approaches

Rejection algorithm:

1. simulate large numbers of data sets under an evolutionary model

2. parameters of the scenario are from a (prior) probability distribution





T (divergence time)

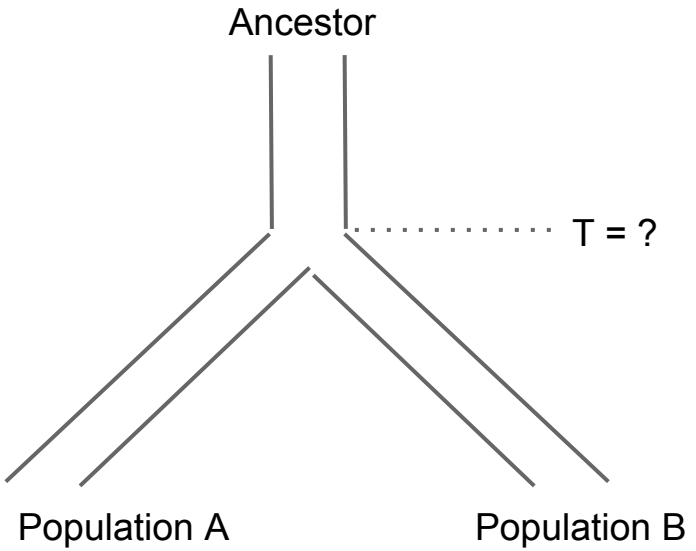
- Choose a distribution
- Define a range of values

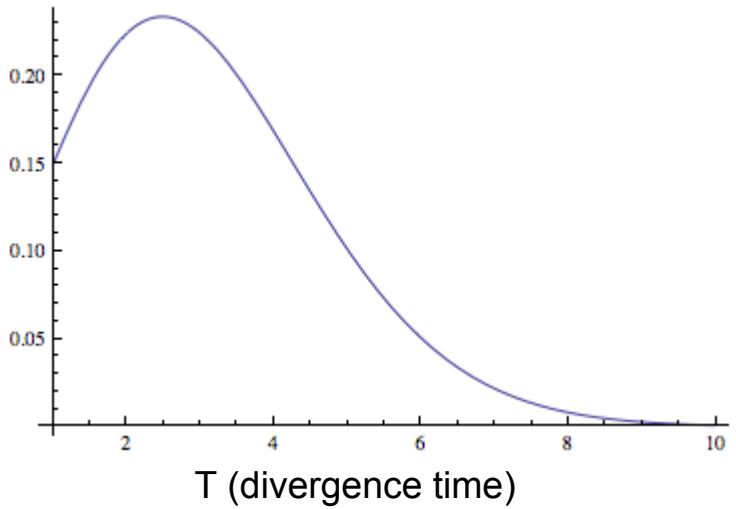
Simulation-based approaches

Rejection algorithm:

1. simulate large numbers of data sets under an evolutionary model

2. parameters of the scenario are from a (prior) probability distribution



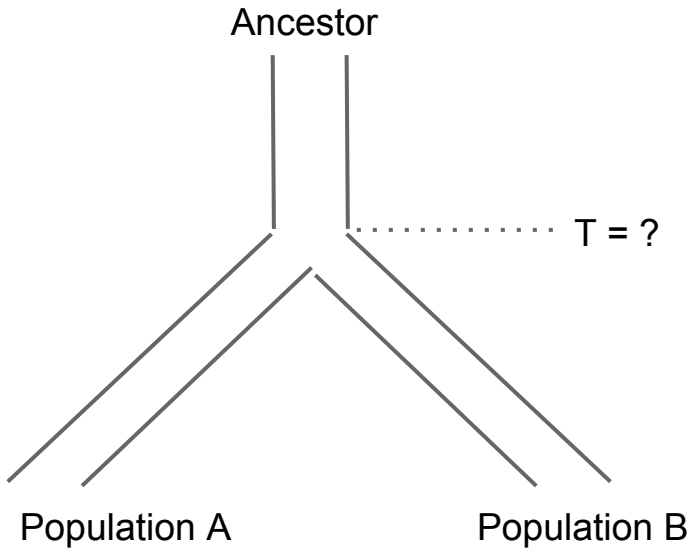


- Choose a distribution
- Define a range of values

Simulation-based approaches

Rejection algorithm:

1. simulate large numbers of data sets under an evolutionary model
2. parameters of the scenario are from a (prior) probability distribution
3. data generated by simulation are then reduced to summary statistics

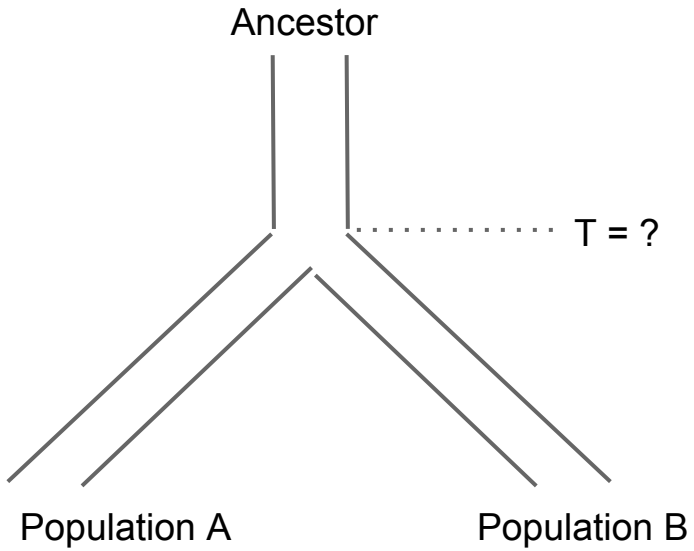


#Simulation	Sampled parameter (T)	Summary Statistics
1	10,345	0.123
2	20,213	0.219
3	5,890	0.098
...		

Simulation-based approaches

Rejection algorithm:

1. simulate large numbers of data sets under an evolutionary model
2. parameters of the scenario are from a (prior) probability distribution
3. data generated by simulation are then reduced to summary statistics
4. sampled parameters are accepted or rejected on the basis of the distance between the simulated and the observed summary statistics

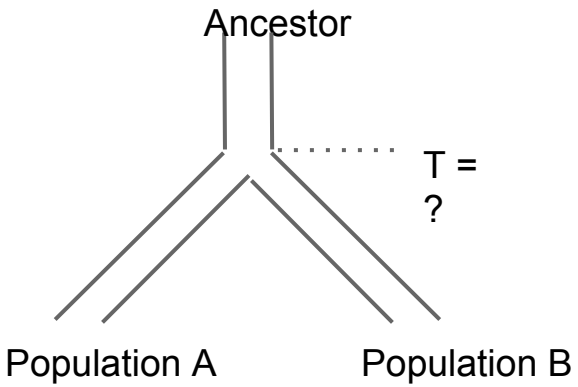


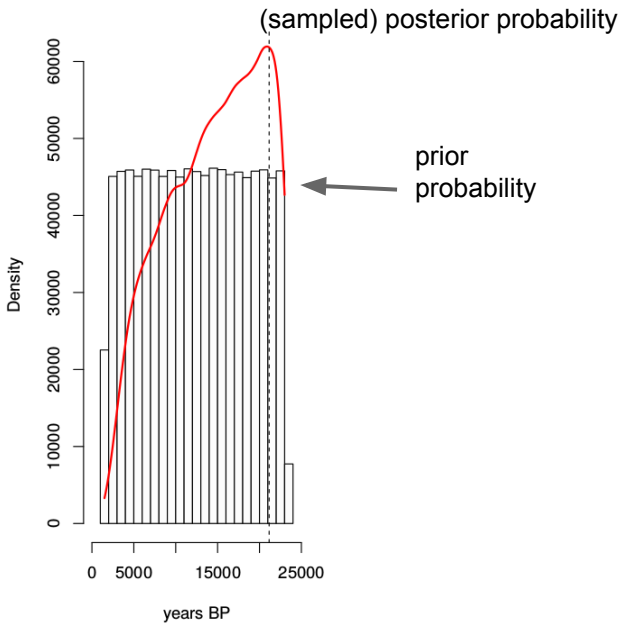
#Simulation	Sampled parameter (T)	Summary Statistic (SS)	Observed SS	Distance
1	10,345	0.12	0.20	0.08
2	**20,213**	**0.21**	**0.20**	**0.01**
3	5,890	0.09	0.20	0.11
...				

Simulation-based approaches

Rejection algorithm:

1. simulate large numbers of data sets under an evolutionary model
2. parameters of the scenario are from a (prior) probability distribution
3. data generated by simulation are then reduced to summary statistics
4. sampled parameters are accepted or rejected on the basis of the distance between the simulated and the observed summary statistics
5. sub-sample of accepted values contains the fitted parameter values, and allows us to evaluate uncertainty on parameters given the observed statistics

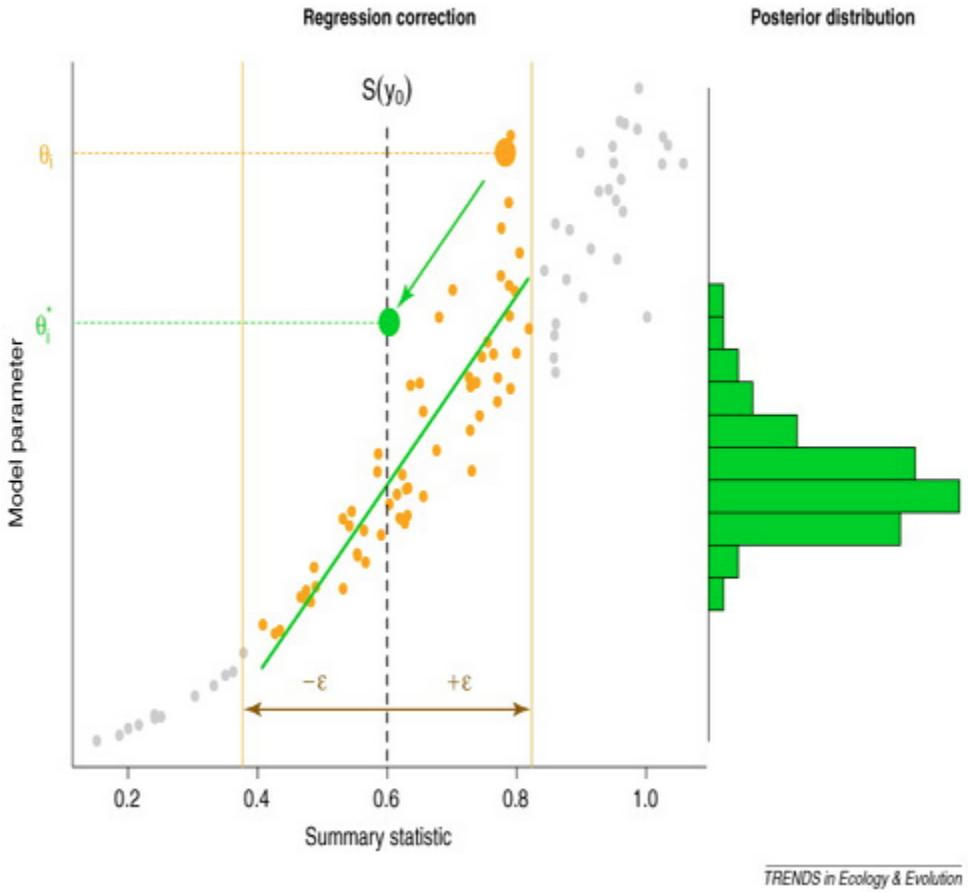




ABC

Approximate Bayesian Computation (ABC) approaches use summary statistics and simulations.

Applications of ABC are often based on improved versions of the basic rejection scheme.



Csilery et al. 2010

Software available

- EIGENSOFT (http://www.hsph.harvard.edu/alkes-price/software/
- fineSTRUCTURE and more (http://www.paintmychromosomes.com)
- dadi (https://code.google.com/p/dadi)
- TreeMix (https://code.google.com/p/treemix/)
- STRUCTURE (http://pritchardlab.stanford.edu/structure.html)
- ADMIXTURE (https://www.genetics.ucla.edu/software/admixture/)
- ABCtoolBox (http://www.cmpg.iee.unibe.ch/content/softwares__services/computer_programs/abctoolbox)
- ngsAdmix (http://www.popgen.dk/software/index.php/NgsAdmix)
- ngsDist (https://github.com/fgvieira/ngsDist)
- diCal (http://sourceforge.net/projects/dical/)
- PSMC (https://github.com/lh3/psmc)
- ...

References - 1

Puckett EE, Kristensen TV, Wilton CM, Lyda SB, Noyce KV, Holahan PM, Leslie DM Jr, Beringer J, Belant JL, White D Jr, Eggert LS. Influence of drift and admixture on population structure of American black bears (Ursus americanus) in the Central Interior Highlands, USA, 50 years after translocation. Mol Ecol. 2014 May;23(10):2414-27. doi: 10.1111/mec.12748. Epub 2014 May 5. PubMed PMID: 24712442.

Fumagalli M, Sironi M. Human genome variability, natural selection and infectious diseases. Curr Opin Immunol. 2014 Oct;30:9-16. doi:10.1016/j.coi.2014.05.001. Epub 2014 May 29. Review. PubMed PMID: 24880709.

Xu X, Liu X, Ge S, Jensen JD, Hu F, Li X, Dong Y, Gutenkunst RN, Fang L, Huang L, Li J, He W, Zhang G, Zheng X, Zhang F, Li Y, Yu C, Kristiansen K, Zhang X, Wang J, Wright M, McCouch S, Nielsen R, Wang J, Wang W. Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. Nat Biotechnol. 2011 Dec 11;30(1):105-11. doi:10.1038/nbt.2050. PubMed PMID: 22158310.

1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012 Nov 1;491(7422):56-65. doi: 10.1038/nature11632. PubMed PMID: 23128226; PubMed Central PMCID: PMC3498066.

Wang C, Zöllner S, Rosenberg NA. A quantitative comparison of the similarity between genes and geography in worldwide human populations. PLoS Genet. 2012 Aug;8(8):e1002886. doi: 10.1371/journal.pgen.1002886. Epub 2012 Aug 23. PubMed PMID: 22927824; PubMed Central PMCID: PMC3426559.

Hellenthal G, Busby GB, Band G, Wilson JF, Capelli C, Falush D, Myers S. A genetic atlas of human admixture history. Science. 2014 Feb 14;343(6172):747-51. doi: 10.1126/science.1243518. PubMed PMID: 24531965; PubMed Central PMCID: PMC4209567.

References - 2

Pickrell JK, Pritchard JK. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet. 2012;8(11):e1002967. doi: 10.1371/journal.pgen.1002967. Epub 2012 Nov 15. PubMed PMID: 23166502; PubMed Central PMCID: PMC3499260.

Heliconius Genome Consortium. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature. 2012 Jul 5;487(7405):94-8. doi: 10.1038/nature11041. PubMed PMID: 22722851; PubMed Central PMCID: PMC3398145.

Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZX, Pool JE, Xu X, Jiang H, Vinckenbosch N, Korneliussen TS, Zheng H, Liu T, He W, Li K, Luo R, Nie X, Wu H, Zhao M, Cao H, Zou J, Shan Y, Li S, Yang Q, Asan, Ni P, Tian G, Xu J, Liu X, Jiang T, Wu R, Zhou G, Tang M, Qin J, Wang T, Feng S, Li G, Huasang, Luosang J, Wang W, Chen F, Wang Y, Zheng X, Li Z, Bianba Z, Yang G, Wang X, Tang S, Gao G, Chen Y, Luo Z, Gusang L, Cao Z, Zhang Q, Ouyang W, Ren X, Liang H, Zheng H, Huang Y, Li J, Bolund L, Kristiansen K, Li Y, Zhang Y, Zhang X, Li R, Li S, Yang H, Nielsen R, Wang J, Wang J. Sequencing of 50 human exomes reveals adaptation to high altitude. Science. 2010 Jul 2;329(5987):75-8. doi: 10.1126/science.1190371. PubMed PMID: 20595611; PubMed Central PMCID: PMC3711608.

Csilléry K, Blum MG, Gaggiotti OE, François O. Approximate Bayesian Computation (ABC) in practice. Trends Ecol Evol. 2010 Jul;25(7):410-8. doi: 10.1016/j.tree.2010.04.001. Epub 2010 May 18. Review. PubMed PMID: 20488578.

Paper discussion

ARTICLE

Uncovering the Genetic History of the Present-Day Greenlandic Population

Ida Moltke,[1,2] Matteo Fumagalli,[3,4] Thorfinn S. Korneliussen,[5] Jacob E. Crawford,[3] Peter Bjerregaard,[6] Marit E. Jørgensen,[6,7] Niels Grarup,[8] Hans Christian Gulløv,[9] Allan Linneberg,[10,11,12] Oluf Pedersen,[8] Torben Hansen,[8,13] Rasmus Nielsen,[3,14,*] and Anders Albrechtsen[1,*]