Population structure
and
demography

Prepared by Claire Mérot & Anna Tigano
Physalia Course 15th September 2020

Why does population structure matter when studying adaptation?

Evolution (including adaptive evolution)

is the result of the interplay of

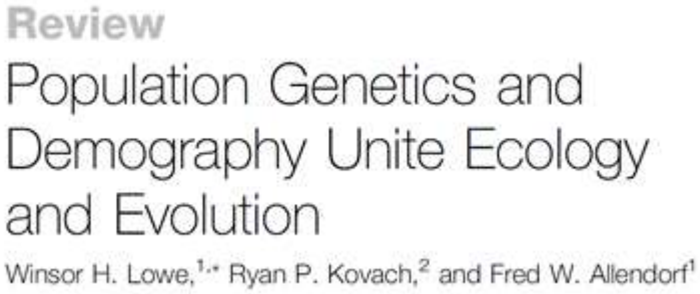
Selection

Drift

Mutation

Gene flow (migration + recombination)

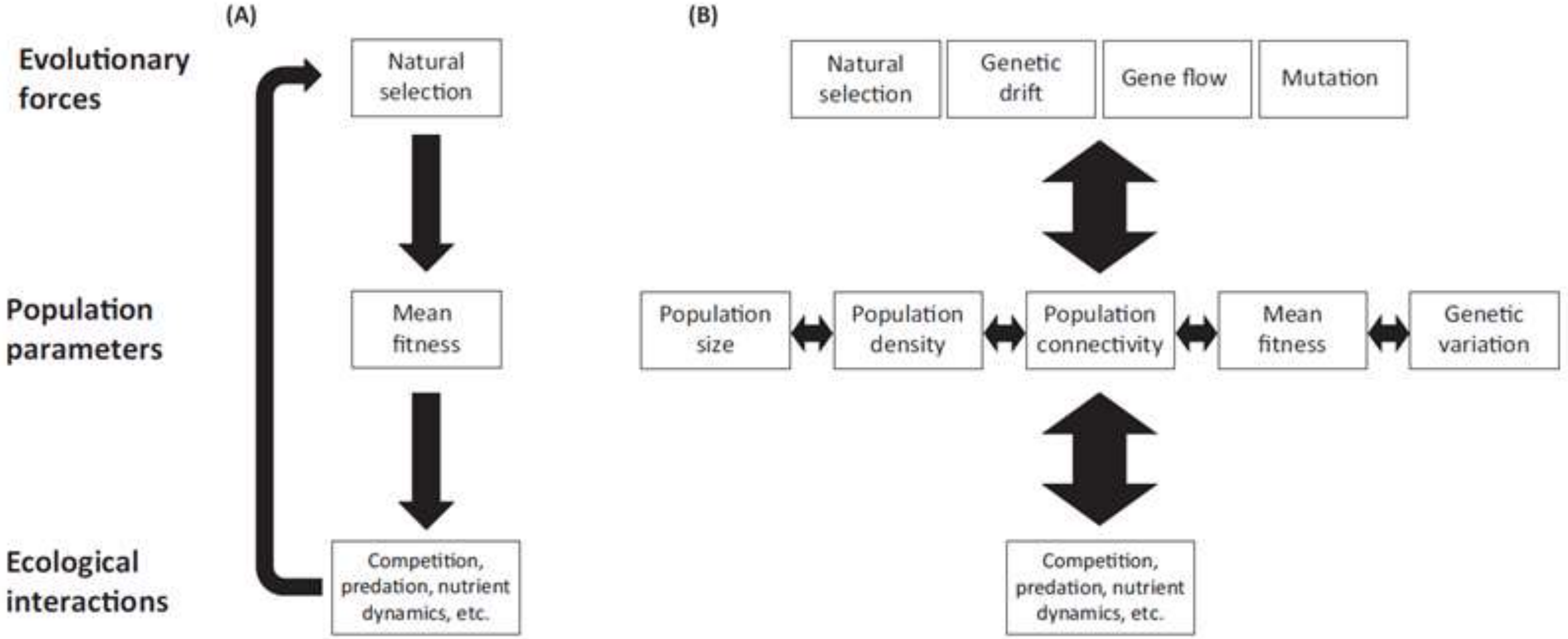
Evolutionary, demographic and ecological processes are inseparable



Review

Population Genetics and Demography Unite Ecology and Evolution

Winsor H. Lowe,[1,*] Ryan P. Kovach,[2] and Fred W. Allendorf[1]



Trends in Ecology & Evolution

Figure 1. Evolutionary and Ecological Processes Are Inseparable. Conceptual illustration of interconnections among evolutionary forces and ecological interactions (biotic and abiotic) through population-level demographic and genetic parameters. (A) represents those interconnections emphasized in current eco-evolutionary research. (B) represents a more comprehensive model of these interconnections, including the full suite of evolutionary forces and a range of population parameters that are themselves interdependent. We build our review around population demographic parameters (size, density, connectivity), but describe key interactions with genetic parameters (mean fitness, genetic variation). We define mean fitness according to population genetics theory as the sum of the fitnesses of genotypes in a population weighted by their proportions [88], thus representing the population-level effects of local adaptation.

Complementary objectives:

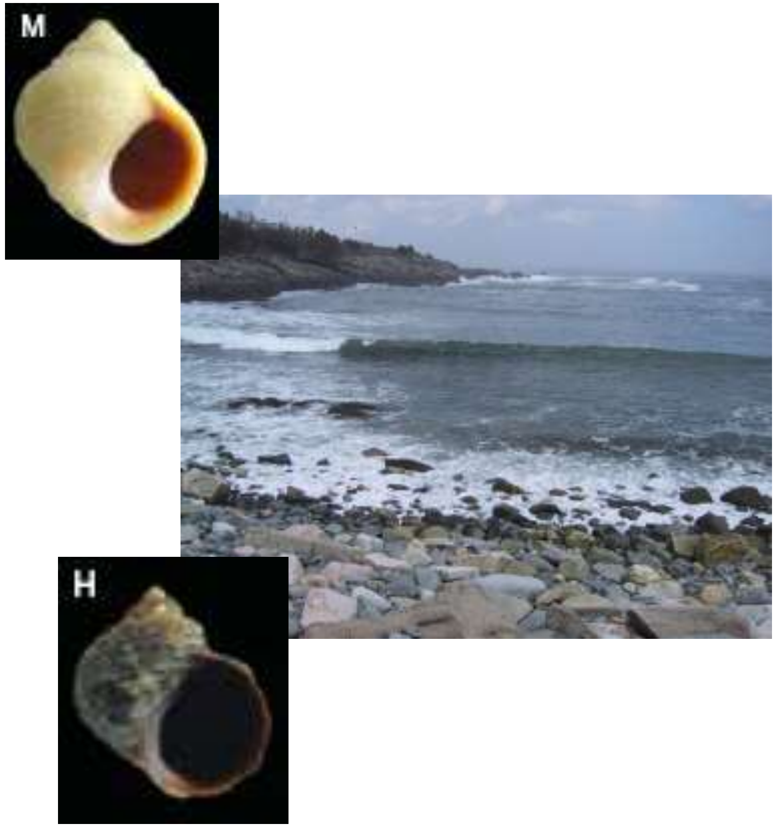
	Study selection and adaptation	**Demographic history and structure of populations**
Actions	Focus on (putatively) adaptive loci	Focus on neutral loci
Use	. Study ecological/functional diversity . Understand adaptative processes under divergent or balancing selection . Identify candidate genes	. Understand the past history of populations . Describe population connectivity . Assess general genetic diversity

Different loci tells a different story…

Exemple:

Wilding *et al.*, 2001. *Journal of Evolutionary Biology*, 14: 611-619

. Parapatric ecotype *Littorina saxatilis*

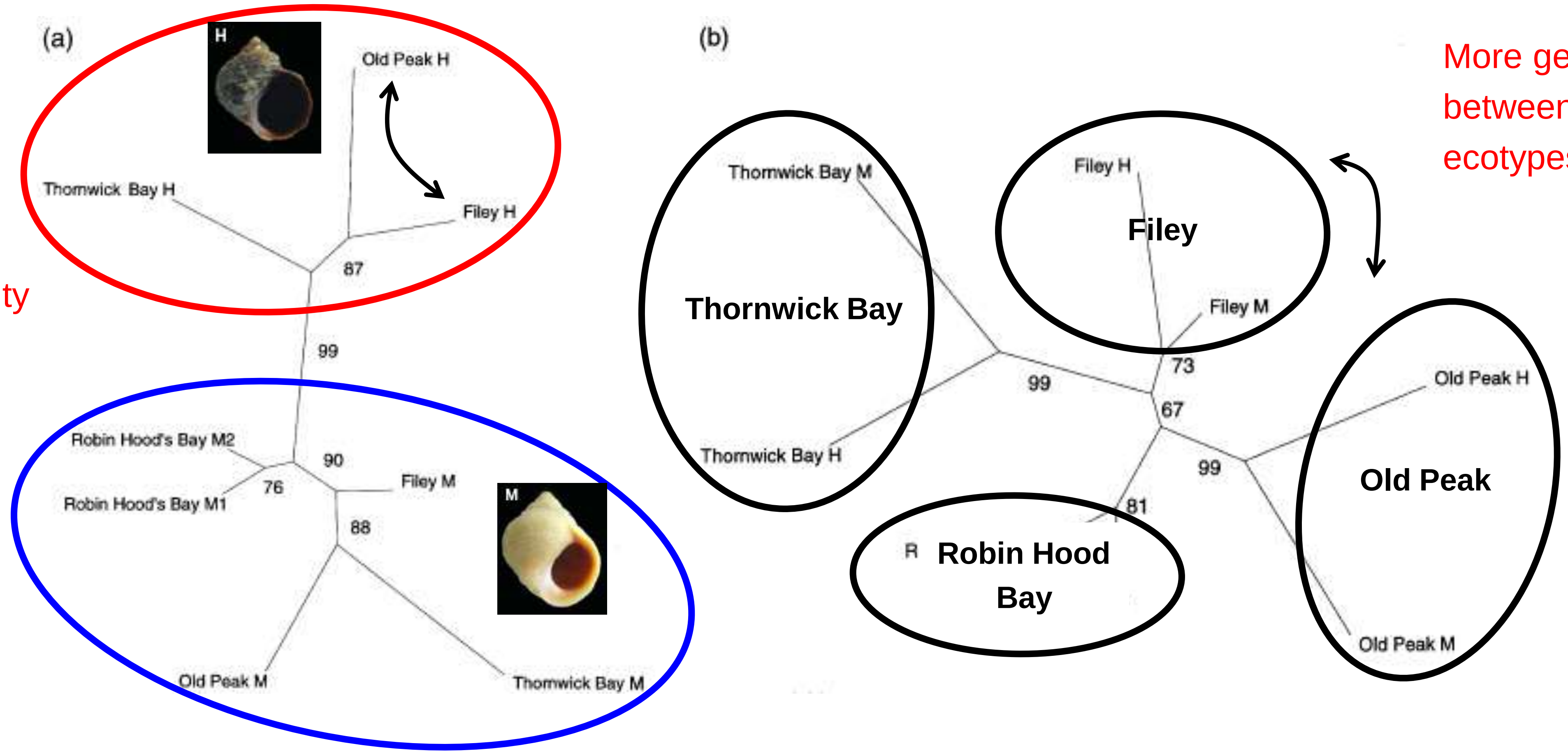


Different loci tells a different story…

Exemple:

Wilding *et al.*, 2001. *Journal of Evolutionary Biology*, 14: 611-619

All loci (290) **Neutral loci (275)**



More similarity between ecotypes of different regions ??

More gene flow between sympatric ecotypes?

Drift

= variation in allele frequency due to random processes

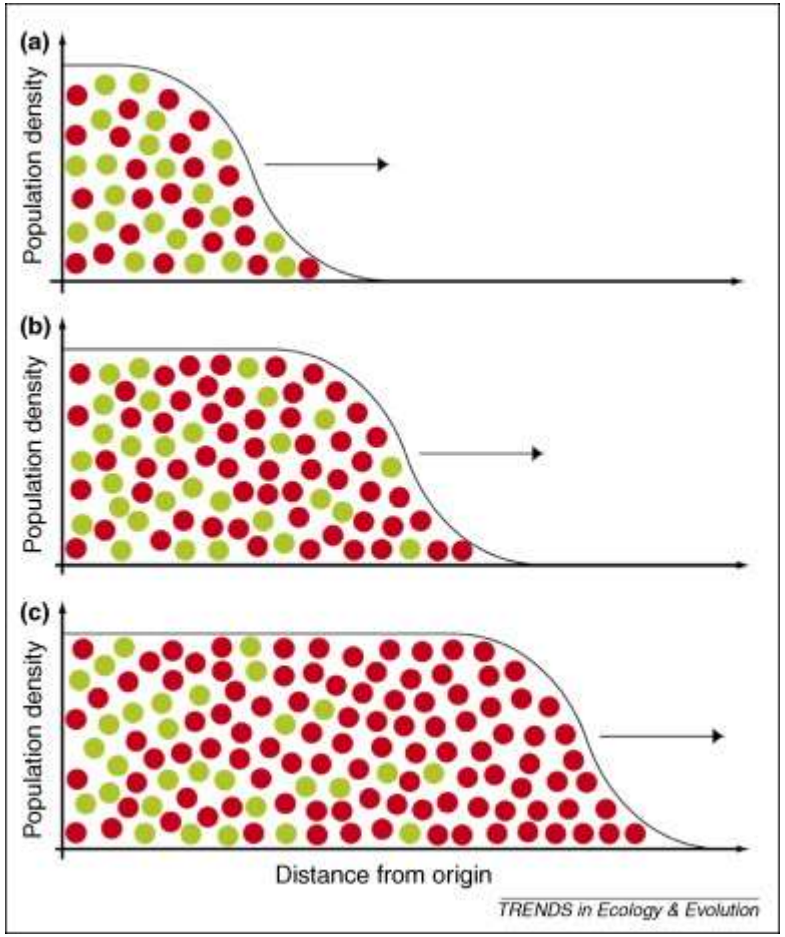
Drift is stronger in smaller populations and it can cause the loss or fixation of a variant due to random sampling of alleles.

Drift is the main driver of genetic population structure, and can generate a genetic footprint similar to that of selection.

Allele surfing

Populations on the leading edge of the expansion are small, and individuals from those populations contribute disproportionately to the propagating wave of expansion.

⇒Rapid drift of some alleles at the expanding edge and high differentiation in allele frequencies over the landscape for some loci, even in the absence of selection



Surfing during population expansions promotes genetic revolutions and structuration
Excoffier & Ray
TREE 2008
https://doi.org/10.1016/j.tree.2008.04.004

Spatial autocorrelation

Correlation between environmental variation & geographic distances
(e.g. climatic clines!)

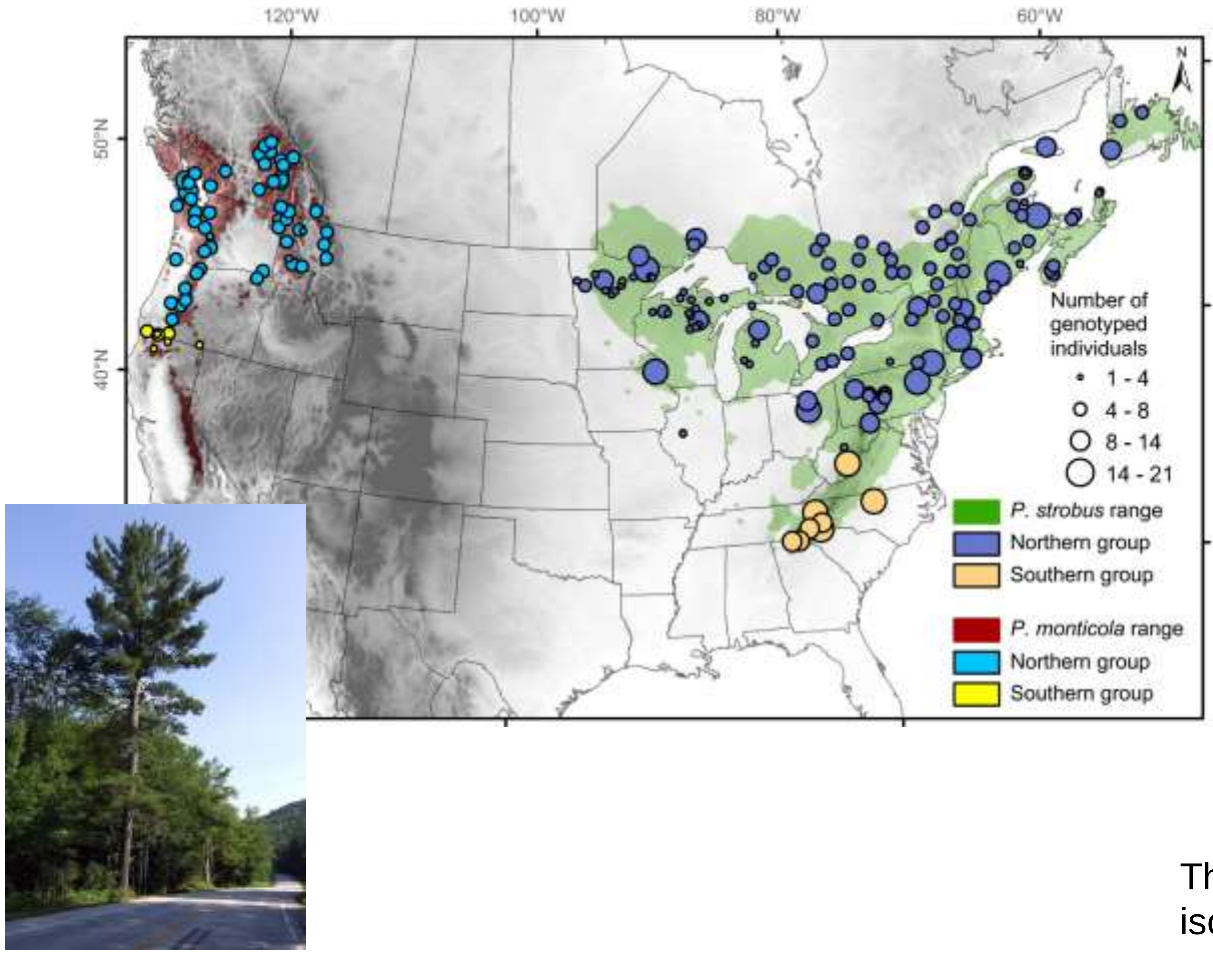
⇒What is adaptation? What is drift?

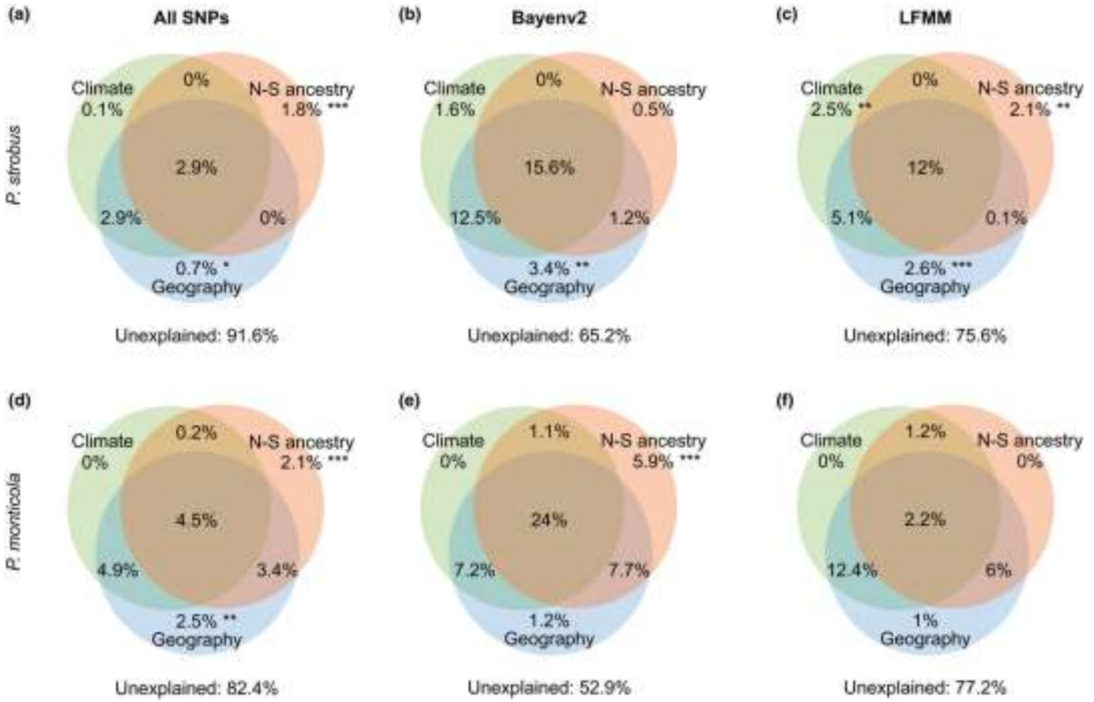
+ Residuals of past range expansion out of glacial refugees…

⇒What is adaptation? What is the results of past history?

Nearby locations are not statistically independent, strong correlations between neutral alleles and environmental variables are more likely to occur by chance than expected with some null models

Isolation-by-distance or adaptation along a gradient … or both?





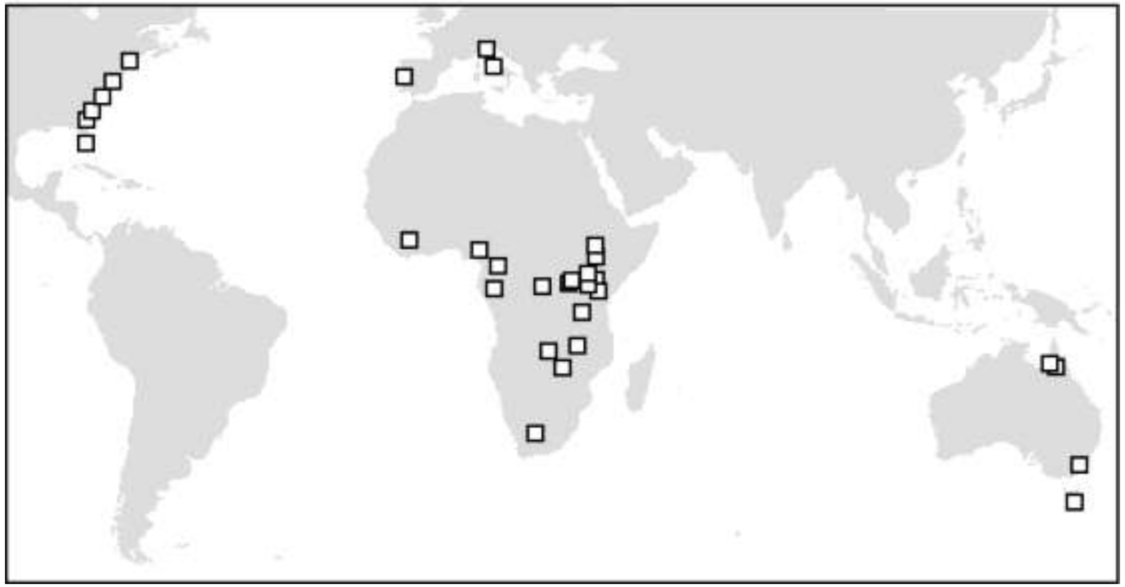
The challenge of separating signatures of local adaptation from those of isolation by distance and colonization history: The case of two white pines
Nadeau et al, 2016 https://doi.org/10.1002/ece3.2550

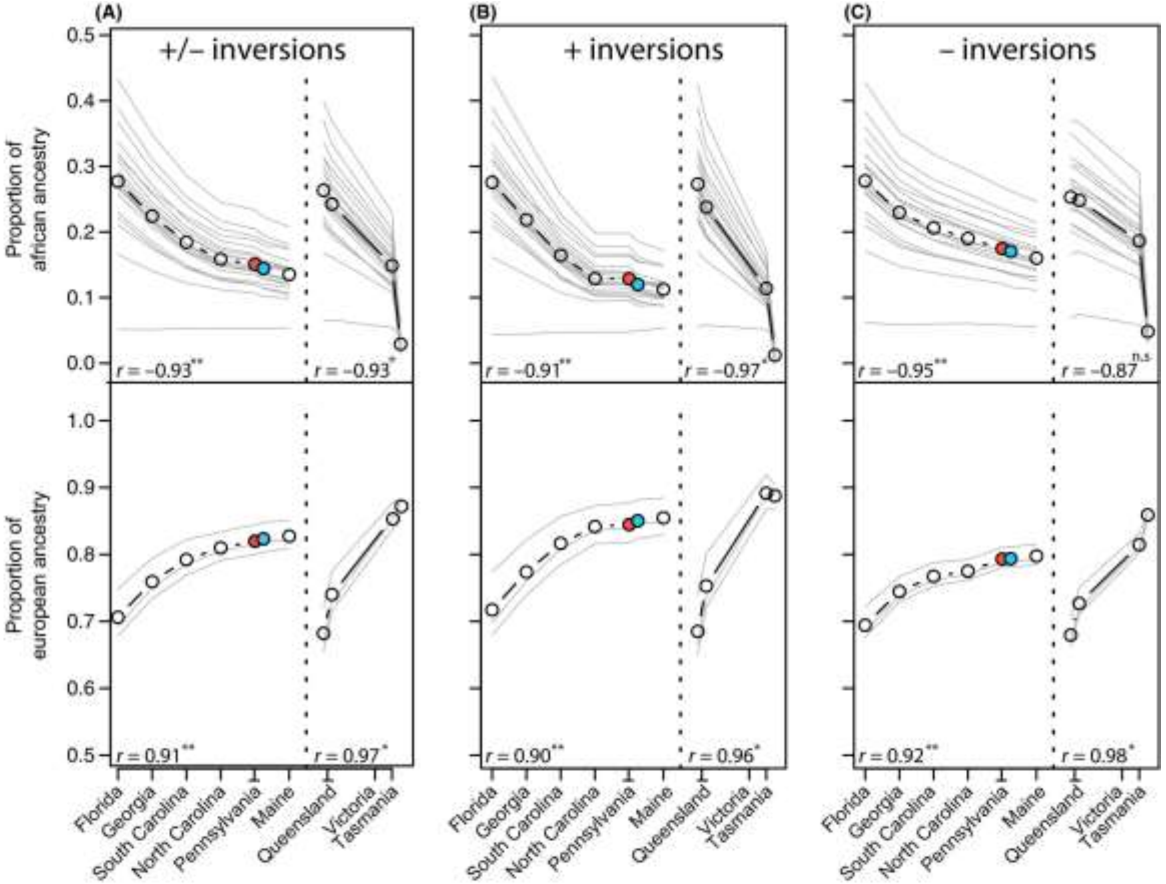
Contact between different lineages / hybridization

Signature of selection or of local adaptation are best detected in a context of (high) gene flow.

Any substructure (lineages, species, secondary contact, admixed populations) should be taken into account.

Clinal variation or secondary contact… Or both?





Bergland et al, 2015 MolEcol
https://doi-org /10.1111/mec.13455

How to characterise population structure?

Unsupervised methods:

- PCA

Semi-supervised methods (K = number of expected clusters)

- Bayesian clustering

Supervised methods (with location information for instance)

- DAPC
- Fst between pairs of populations

Principal Component Analysis (PCA)

- A common statistical tool that reduces matrix complexity by identifying the eigenvectors and ordering them

- The top PCs reflect axis of genetic variation along which individuals with same ancestry, or exchanging genetic material, are more similar to each other.

- Caution: can be strongly driven by few loci in linkage disequilibrium…

- For population structure purpose:

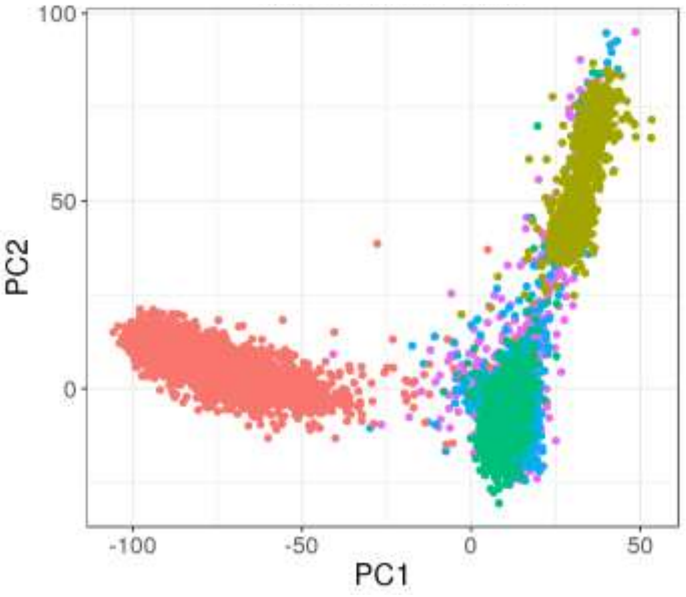
-> compare Pca on all SNPs vs. Pca on LD-pruned SNPs
-> look at loadings of the PCs: which fraction of the genome explains PC1? Explains PC2? Etc..

- There is lots of genetic variance, it can be relatively expected that even PC1 explains less than 1% of variance. (but it can also capture 20-50%... Depends on the dataset!

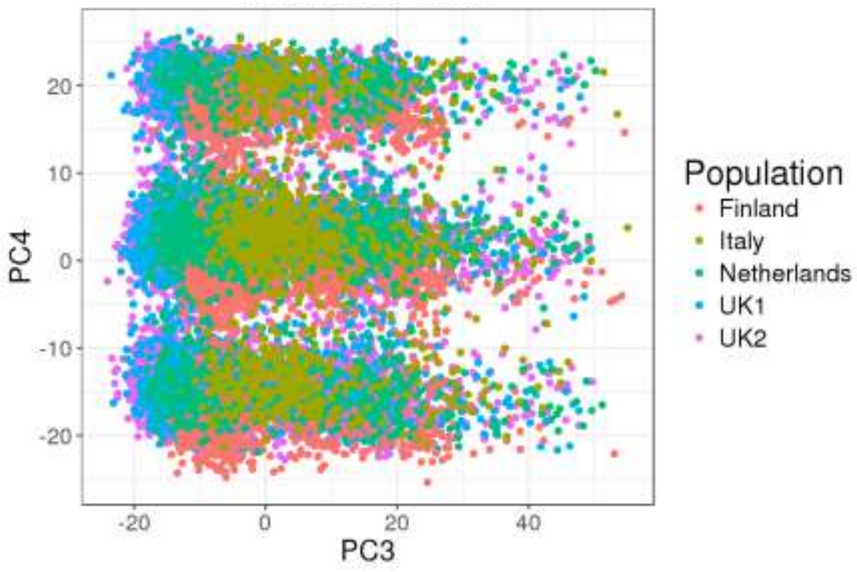
Principal Component Analysis (PCA)

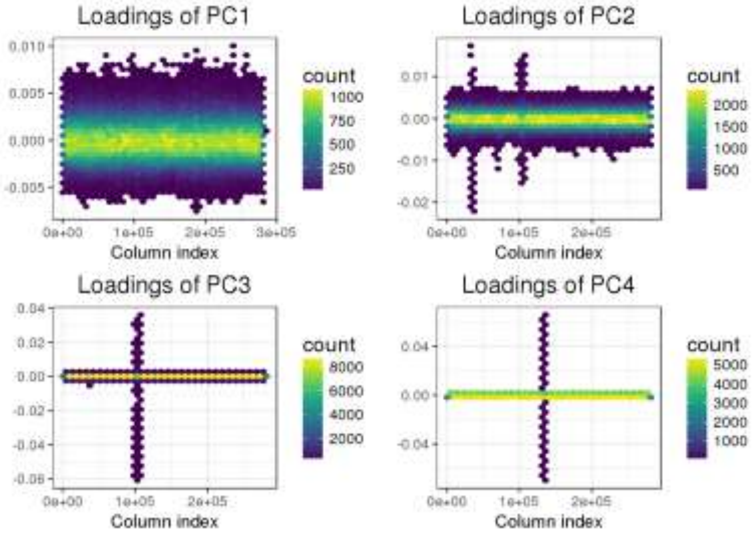
Each individual is a point with coordinates along all PCs
Each genetic marker contribute to all PCs with a different strength (loadings)

ALL SNPS



ALL SNPS



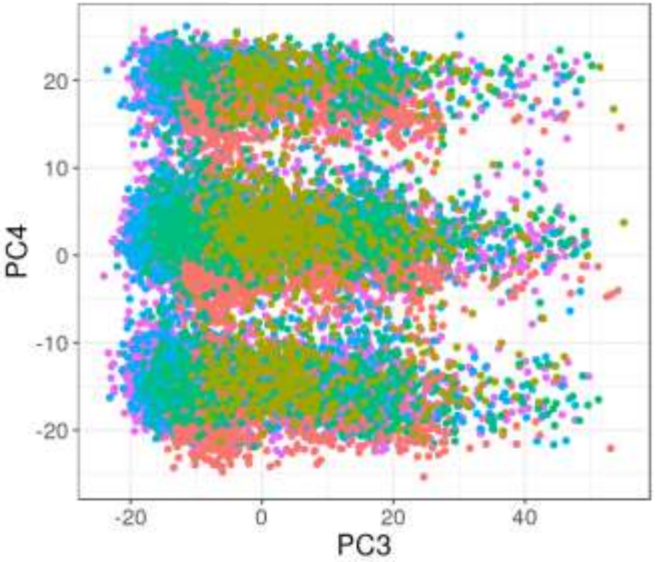


Packages *bigstatsr, bigsnpr* to remove short-range and long-range LD.
Nice tutorial about PCA for pop genomics!
Florian Privé
https://privefl.github.io/bigsnpr/articles/how-to-PCA.html

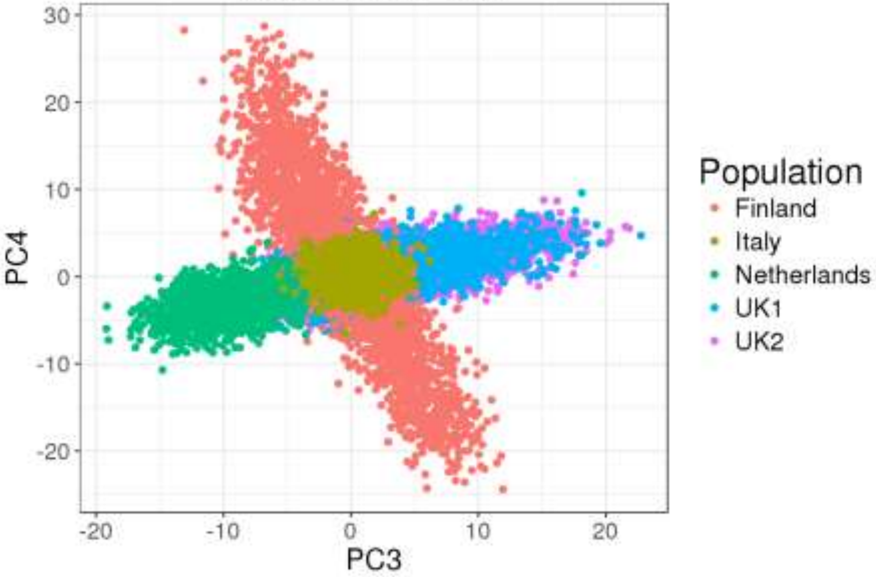
Principal Component Analysis (PCA)

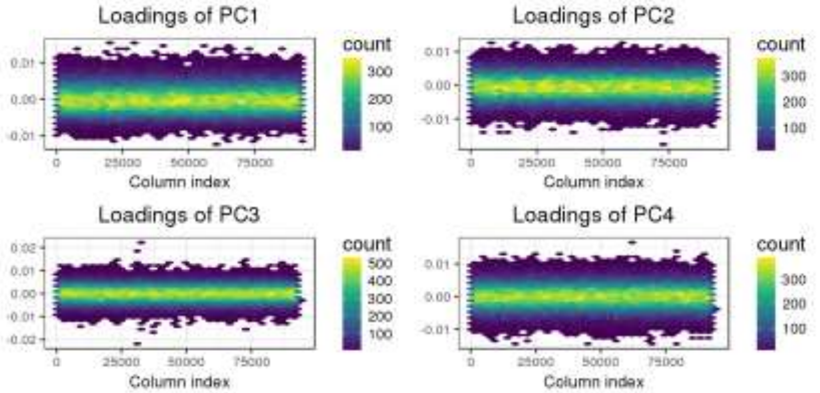
Each individual is a point with coordinates along all PCs
Each genetic marker contribute to all PCs with a different strength (loadings)

ALL SNPS



LD-pruned SNPs





Packages *bigstatsr, bigsnpr* to remove short-range and long-range LD.
Nice tutorial about PCA for pop genomics!
Florian Privé
https://privefl.github.io/bigsnpr/articles/how-to-PCA.html

Bayesian clustering (STRUCTURE, etc..)

- Aim to sort individuals into K clusters so as to minimize departures from Hardy-Weinberg equilibrium and linkage equilibrium

- Caution: can be strongly driven by few loci in linkage disequilibrium…

- For population structure purpose:

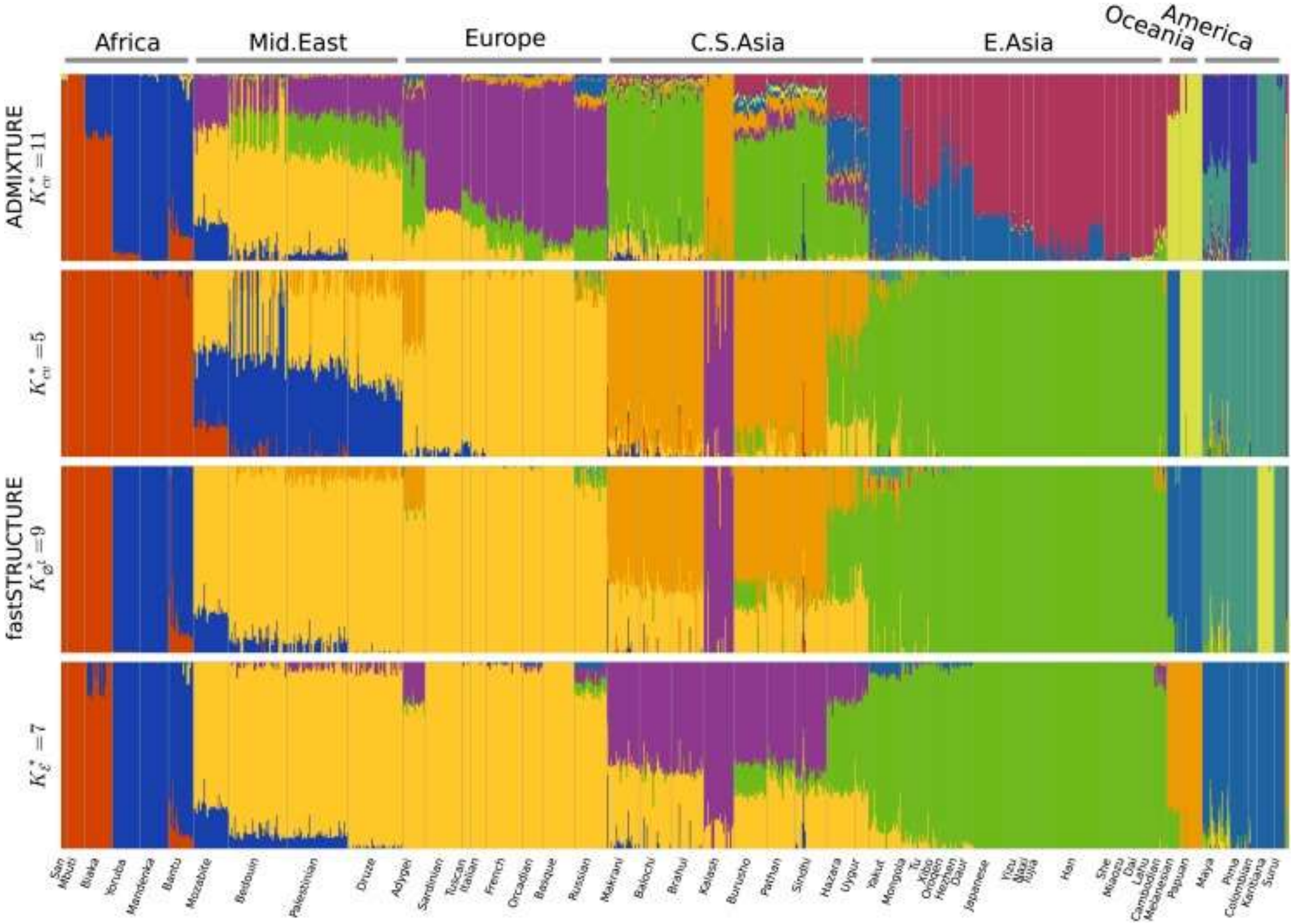
-> compare results on all SNPs vs. results on LD-pruned SNPs
-> explore many values of K

- Admixture or FastSTRUCTURE replace STRUCTURE for genome-wide data

- Evaluate the fit of the model

Evaluation of model fit of inferred admixture proportions
Genís Garcia-Erill Anders Albrechtsen
MER 2020
https://doi.org/10.1111/1755-0998.13171

Bayesian clustering (STRUCTURE, etc..)



Each individual is a thin vertical line that is partitioned into *K* colored segments according to its membership coefficients in *K* clusters.

fastSTRUCTURE: variational inference of population structure in large SNP data sets 2014 Genetics

Anil Raj[1], Matthew Stephens[2], Jonathan K Pritchard[3]

10.1534/genetics.114.164350

The advantage of unsupervised/semi-supervised methods:
=> Other surprises!!

Species lineage substructure

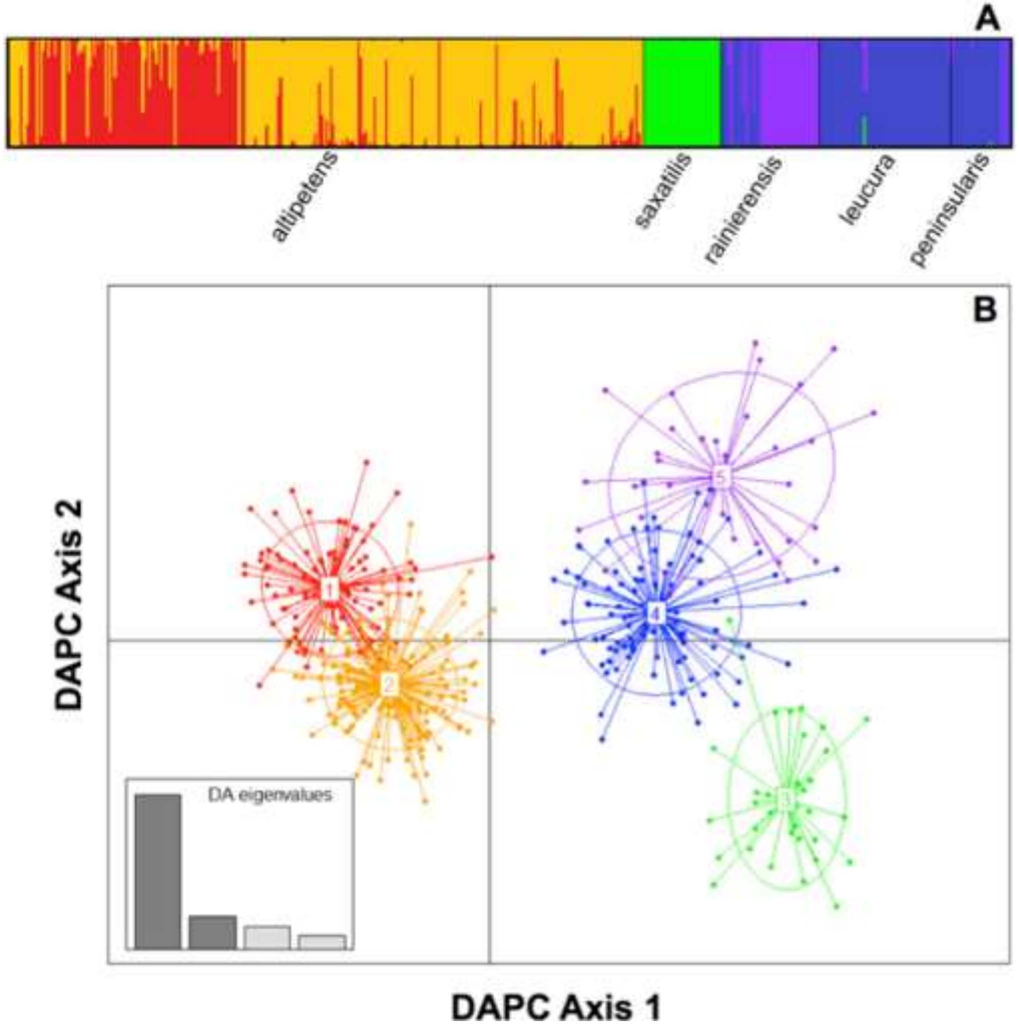
Hybridisation

Chromosomal
rearrangements….

DAPC (discriminant PCA)

- A mix of a discriminant analysis and a PCA
- It will try very hard to find axis of variation that discriminate the groups given *a priori*
- *WARNING:* A dangerous analysis when we have much more markers (SNPs) than groups (populations)… Be well aware of not over-fitting and not-overinterpreting the output.

Miller, J.M., Cullingham, C.I. & Peery, R.M. The influence of a priori grouping on inference of genetic clusters: simulation study and literature review of the DAPC method. *Heredity* (2020). https://doi.org/10.1038/s41437-020-0348-2



Languin et al, 2018,
Conservation genomics

Pairwise Fst

- Fst can only been computed between two groups.

- When sampling several populations, we will be interested in Fst between all pairs of individuals.

- A measure of genetic distance between all populations (does it correlate with ecological distances? Geographic distances? Etc?)

- Again, likely better on Ld-pruned SNPs to infer neutral structure…

- Absolute values are informative… (0,000x -> high gene flow, don’t bother too much about looking for structure. 0,01-0,1 -> consider carefully structure… Higher: do you really have one species?)

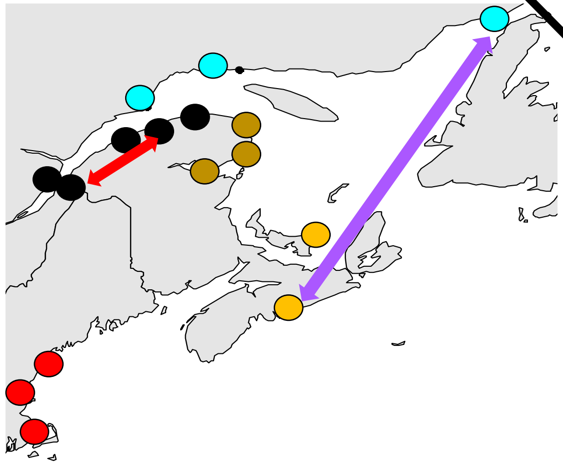
Pairwise Fst

ALL SNPs

North

	BS	RB	BT	SI	KA	ME	GM	RC	AG	CE	SS	NB	CB	BP	HA	MA
BS	0	0.003	0.006	0.005	0.009	0.005	0.01	0.007	0.006	0.007	0.009	0.007	0.006	0.011	0.015	0.019
RB		0	0.004	0.007	0.008	0.007	0.013	0.008	0.005	0.008	0.01	0.008	0.009	0.011	0.014	0.019
BT			0	0.011	0.004	0.011	0.021	0.014	0.005	0.012	0.015	0.007	0.012	0.01	0.01	0.02
SI				0	0.009	0.002	0.004	0.002	0.004	0.002	0.004	0.005	0.003	0.007	0.011	0.01
KA					0	0.009	0.02	0.013	0.004	0.011	0.013	0.004	0.01	0.006	0.005	0.016
ME						0	0.004	0.002	0.004	0.002	0.004	0.005	0.003	0.007	0.011	0.011
GM							0	0.003	0.011	0.004	0.004	0.013	0.008	0.014	0.021	0.013
RC								0	0.006	0.002	0.003	0.007	0.005	0.009	0.014	0.01
AG									0	0.005	0.006	0.003	0.006	0.005	0.007	0.012
CE										0	0.002	0.006	0.005	0.008	0.012	0.009
SS											0	0.008	0.007	0.008	0.012	0.007
NB												0	0.004	0.005	0.006	0.012
CB													0	0.008	0.012	0.014
BP														0	0.003	0.007
HA															0	0.008
MA																0

South



LD-pruned SNPs

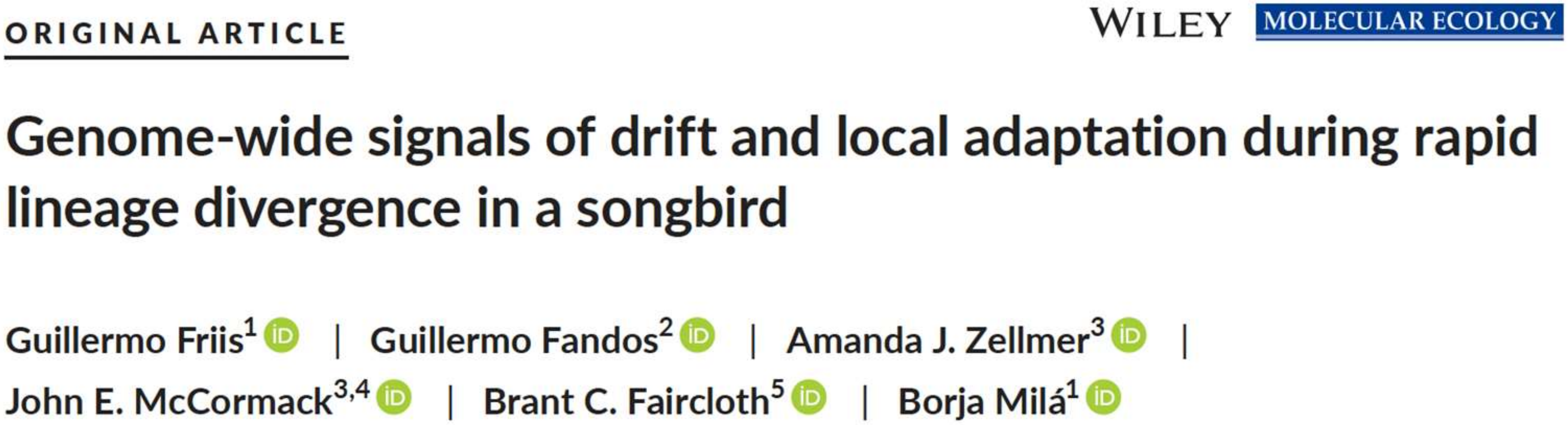
North

	BS	RB	BT	SI	KA	ME	GM	RC	AG	CE	SS	NB	CB	BP	HA	MA
BS	0	0.009	0.009	0.014	0.015	0.014	0.014	0.014	0.013	0.014	0.016	0.018	0.018	0.026	0.028	0.034
RB		0	0.008	0.017	0.017	0.017	0.017	0.017	0.015	0.016	0.018	0.023	0.024	0.029	0.031	0.037
BT			0	0.013	0.012	0.014	0.014	0.013	0.012	0.013	0.014	0.018	0.02	0.023	0.025	0.031
SI				0	0.008	0.007	0.008	0.007	0.008	0.008	0.008	0.009	0.01	0.013	0.014	0.02
KA					0	0.009	0.01	0.009	0.008	0.009	0.009	0.01	0.012	0.013	0.015	0.02
ME						0	0.008	0.007	0.008	0.008	0.009	0.01	0.01	0.012	0.014	0.02
GM							0	0.008	0.009	0.008	0.008	0.011	0.012	0.014	0.017	0.021
RC								0	0.007	0.007	0.008	0.01	0.011	0.014	0.016	0.02
AG									0	0.007	0.008	0.01	0.012	0.014	0.016	0.021
CE										0	0.007	0.01	0.012	0.014	0.016	0.02
SS											0	0.01	0.012	0.013	0.016	0.019
NB												0	0.008	0.013	0.016	0.021
CB													0	0.015	0.018	0.024
BP														0	0.007	0.012
HA															0	0.01
MA																0

South

LD-pruned SNPs
⇒ Isolation by
distance

Case study



ORIGINAL ARTICLE

WILEY MOLECULAR ECOLOGY

Genome-wide signals of drift and local adaptation during rapid lineage divergence in a songbird

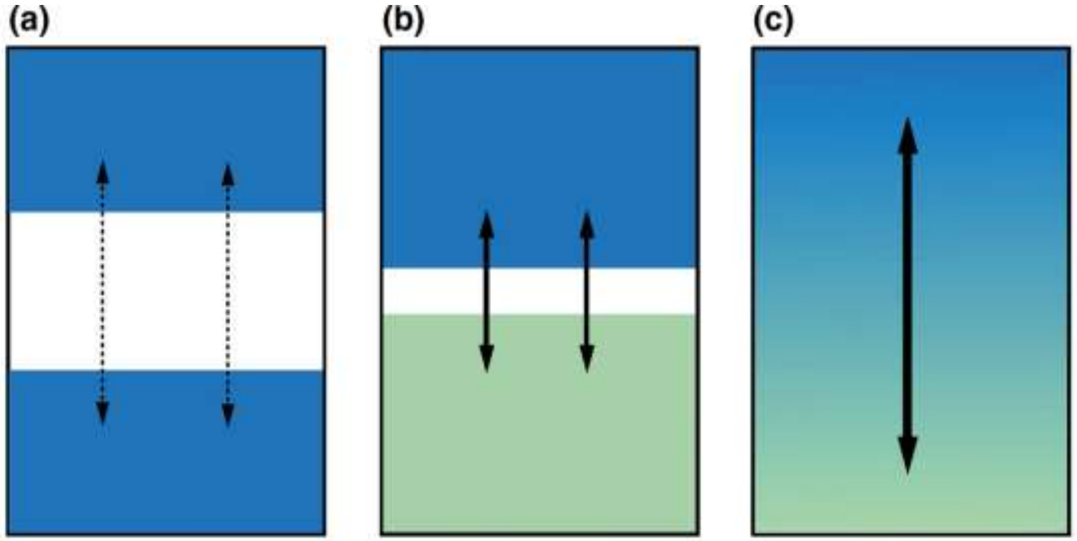
Guillermo Friis[1] | Guillermo Fandos[2] | Amanda J. Zellmer[3] | John E. McCormack[3,4] | Brant C. Faircloth[5] | Borja Milá[1]

Set expectations

(a) Geographically isolated populations in similar habitats.

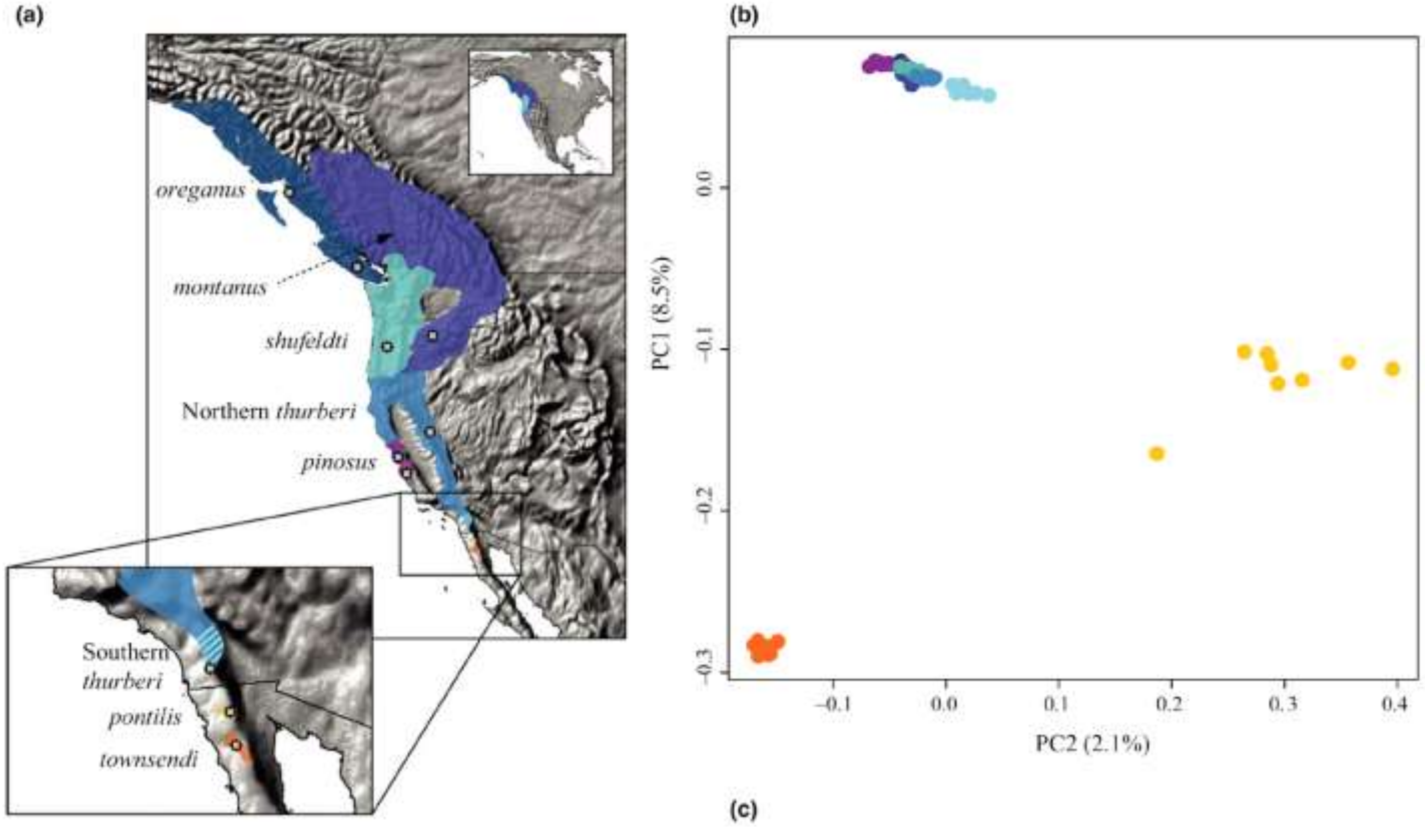
(b) Parapatric populations in ecologically
divergent habitats.

(c) Population continuum across a selective gradient



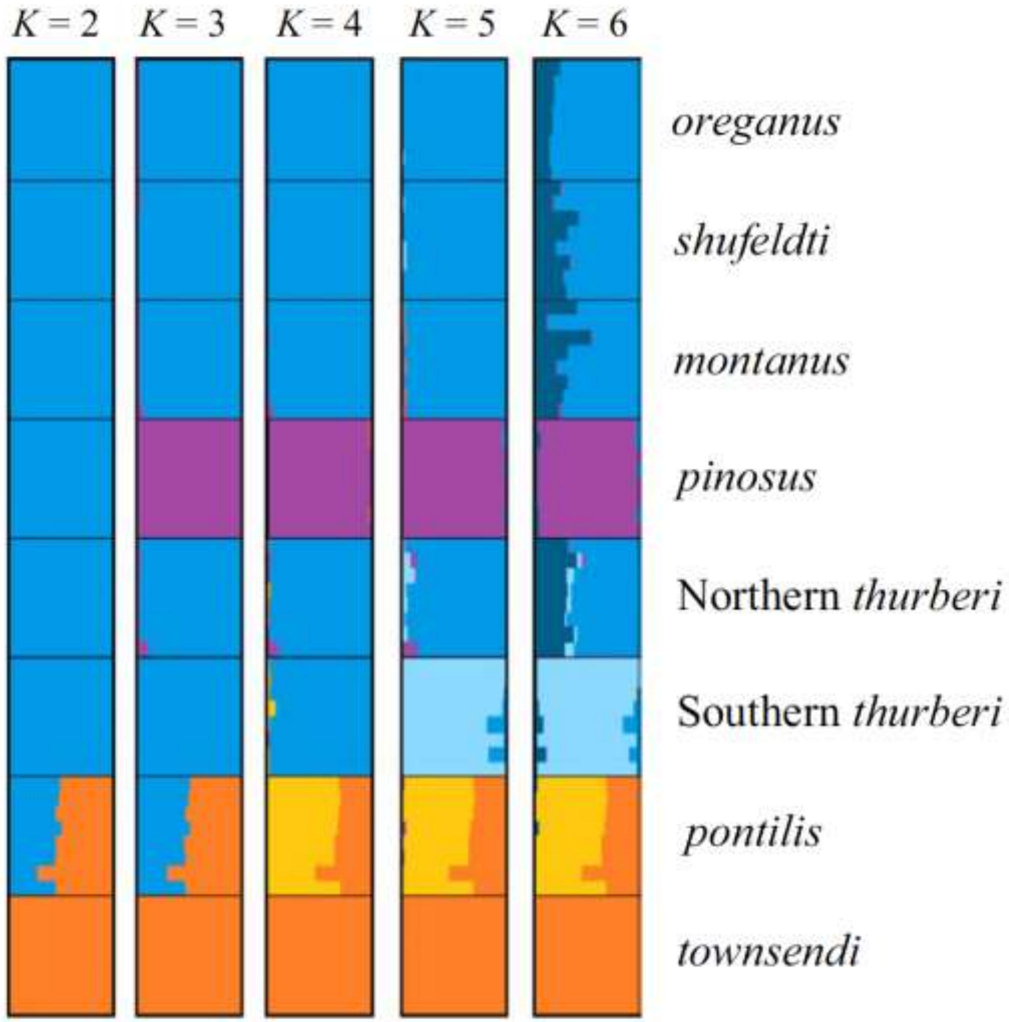
Habitat types:	ONE	TWO	GRADIENT
Gene flow:	LOW	MODERATE	HIGH
Neutral divergence:	HIGH	MODERATE	LOW
Adaptive divergence:	LOW	HIGH	HIGH

PCA for data exploration



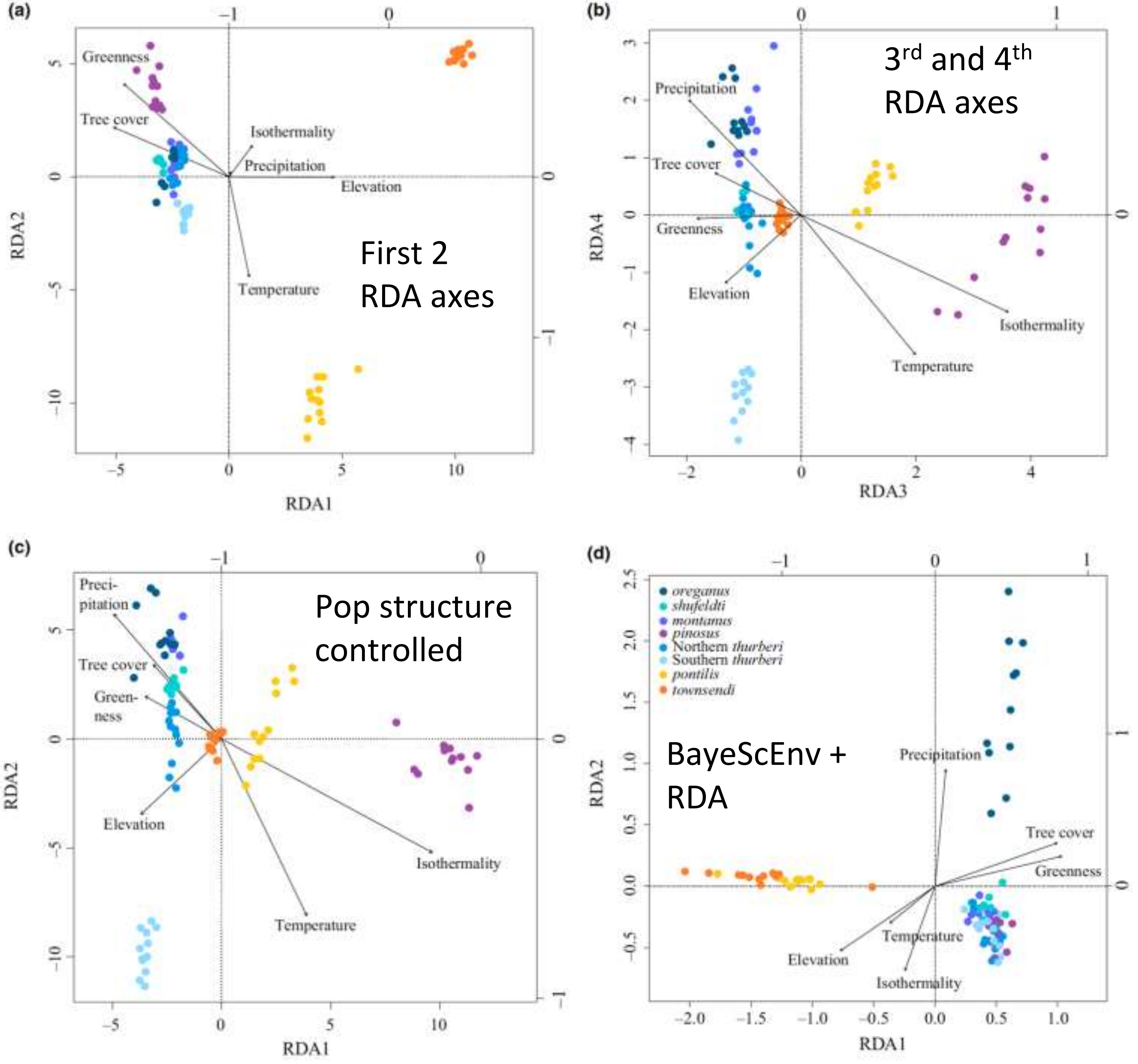
Prior-free data exploration

STRUCTURE for population structure



Always examine multiple K values as more than one K could be biologically informative

(STRUCTURE doesn't deal well with hierarchical population structure)



First 2 RDA axes

3rd and 4th RDA axes

Pop structure controlled

BayeScEnv + RDA

GEA

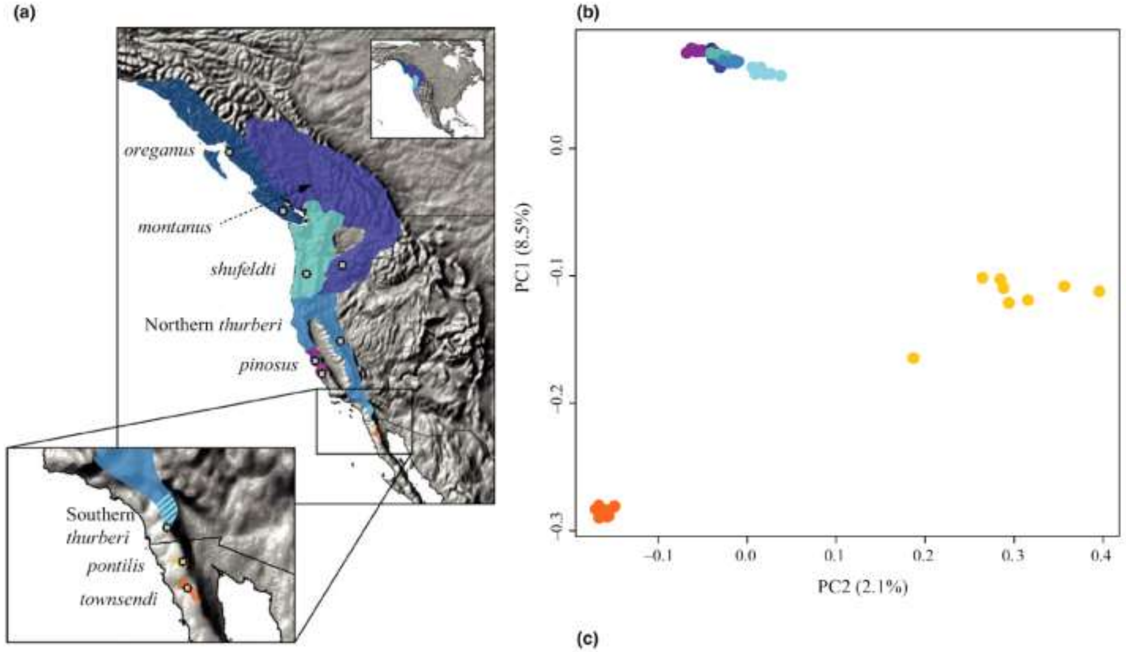
Genotype-environment associations with multiple approaches

Lecture tomorrow, we come back on that

Partition of genetic variation

- Environmental variables (controlling for population structure) 1.17%
- Environmental variable + pop structure 7.41%
- 92.59%?
  - Loci under balancing selection
  - Other selective pressures
  - Shared neutral variation due to relatedness and/or gene flow

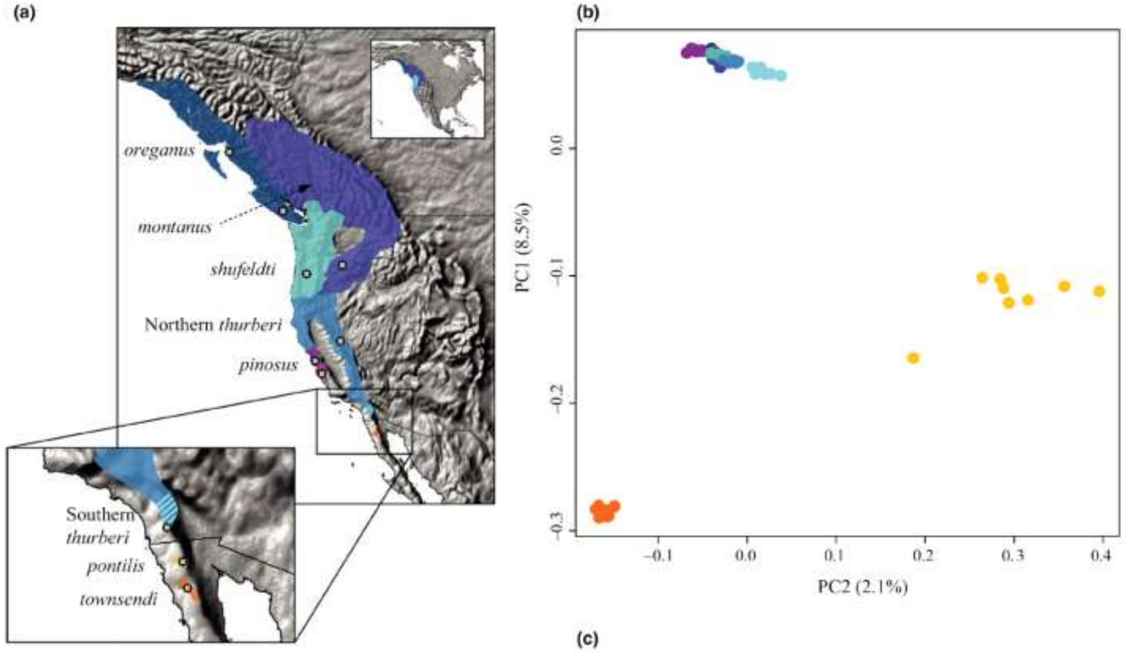
Environment + geography + demography



Strong population structure and weak GEA

Isolation-by-resistance and drift

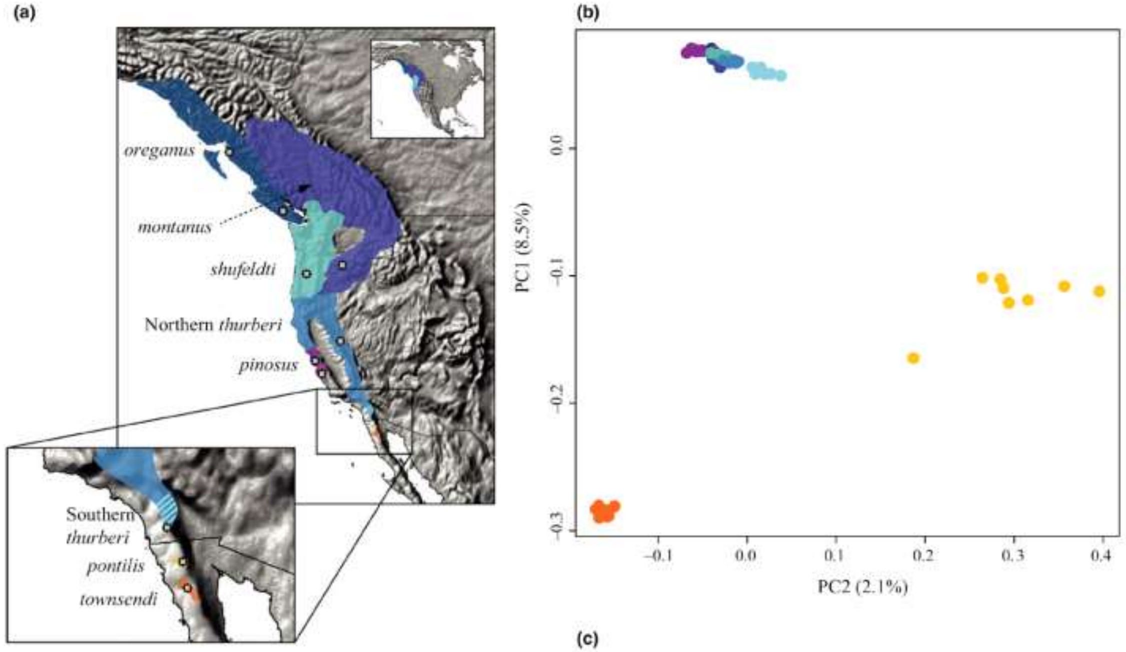
Environment + geography + demography



Weak pop structure and stronger differentiation on GEA

Isolation-by-adaptation?

Environment + geography + demography



No pop structure + environmental associations

⇒ Ongoing gene flow and local adaptation

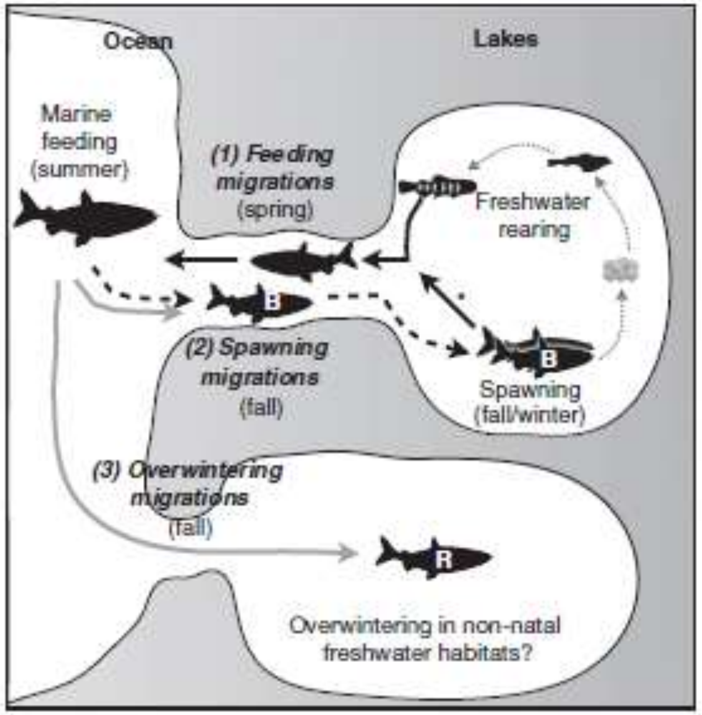
And help from experimental work /knowledge of natural history

Capture-Mark-Recapture

-> population size, dynamic and movement

Spatial ecology

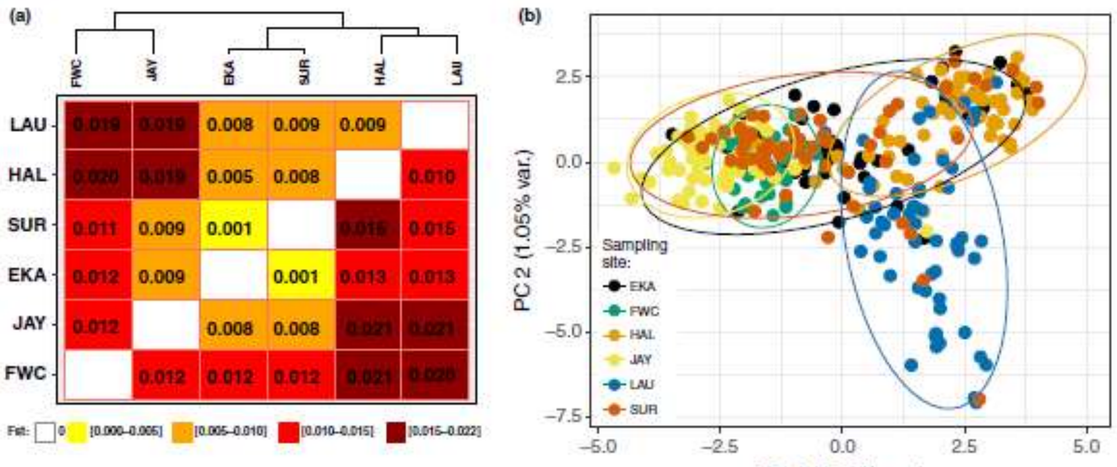
-> tracking, etc..

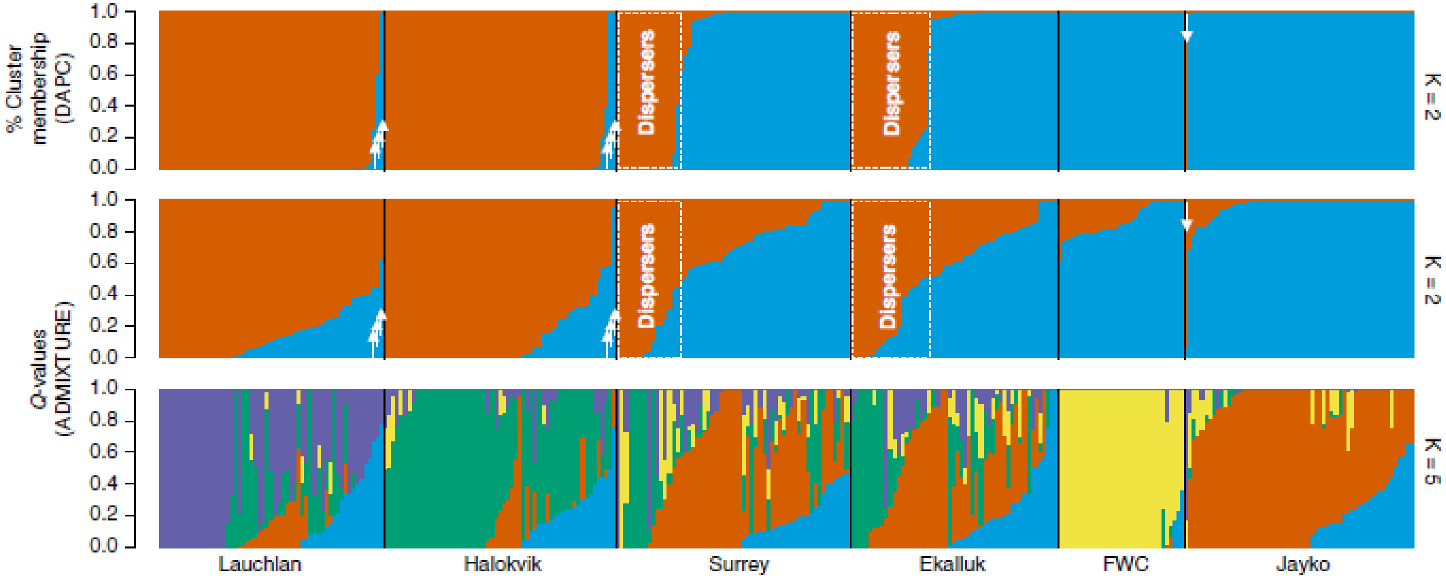

Ocean
Lakes
Marine feeding (summer)
(1) Feeding migrations (spring)
Freshwater rearing
B
(2) Spawning migrations (fall)
B
Spawning (fall/winter)
(3) Overwintering migrations (fall)
R
Overwintering in non-natal freshwater habitats?

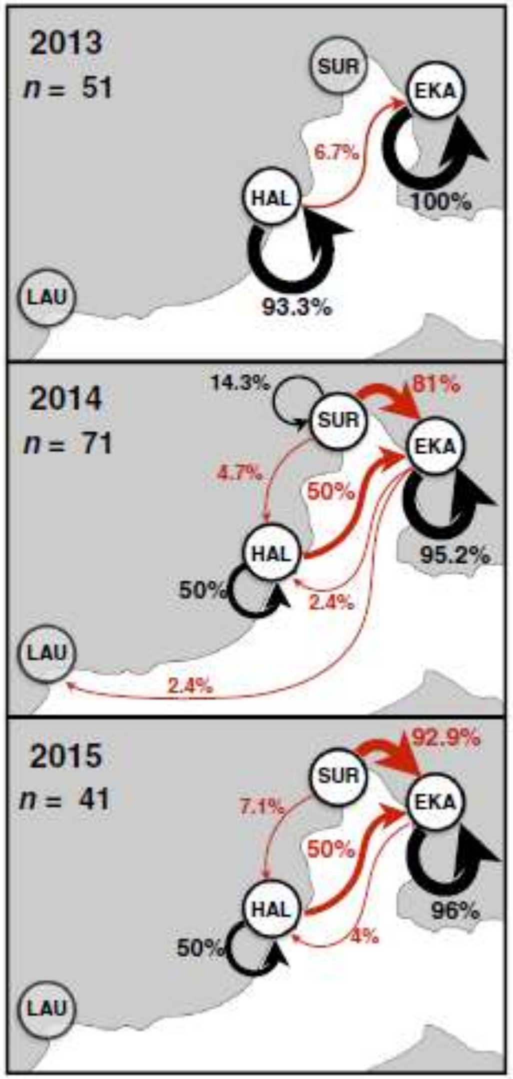
e.g : telemetry & genomics

Moore, J.-S., L.N. Harris, J. Le Luyer, B.J.G. Sutherland, Q. Rougemont, R.F. Tallman, A.T. Fisk & L. Bernatchez (2017) Genomics and telemetry suggest a role for migration harshness in determining overwintering habitat choice, but not gene flow, in anadromous Arctic Char. ***Molecular Ecology,*** 26(24): 6784-6800

And help from experimental work /knowledge of natural history





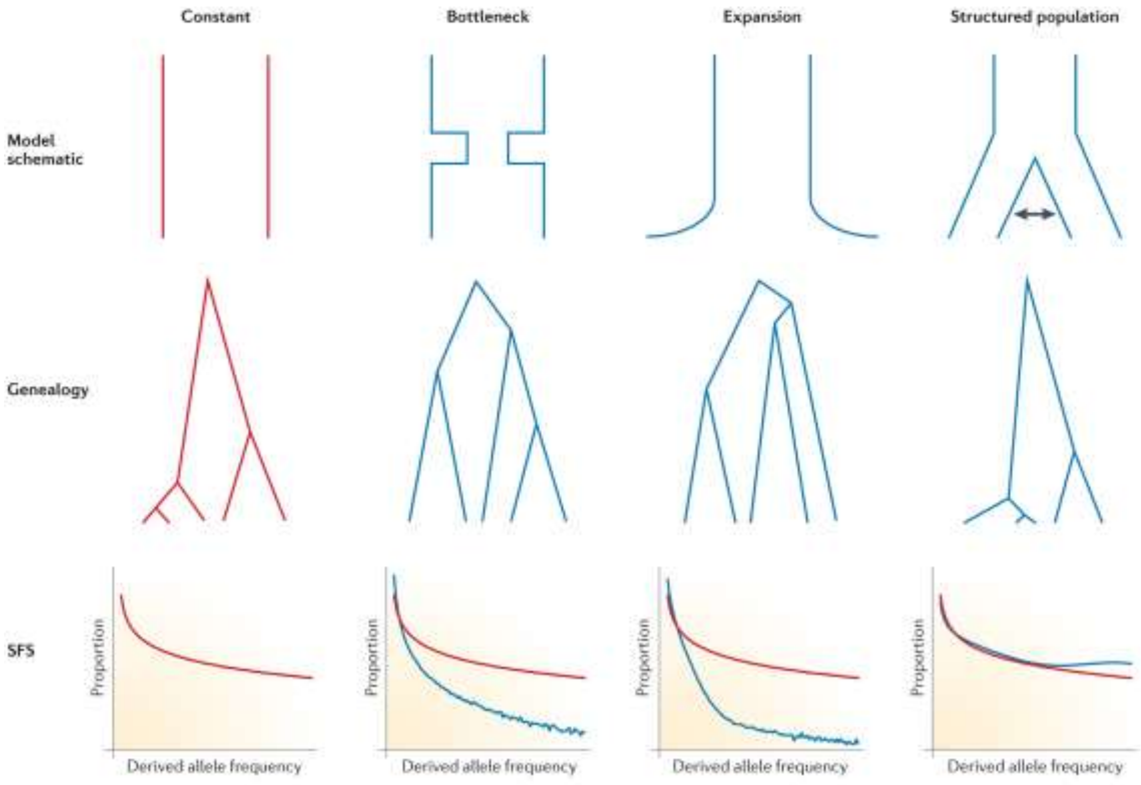


Moore, J.-S.,. *Molecular Ecology,* 26(24): 6784-6800

Beyond present structure…
How to know population history and demography?

Models:

- to understand population history, bottleneck, gene flow...
- demography can set a null model against which one can look for the effect of selection



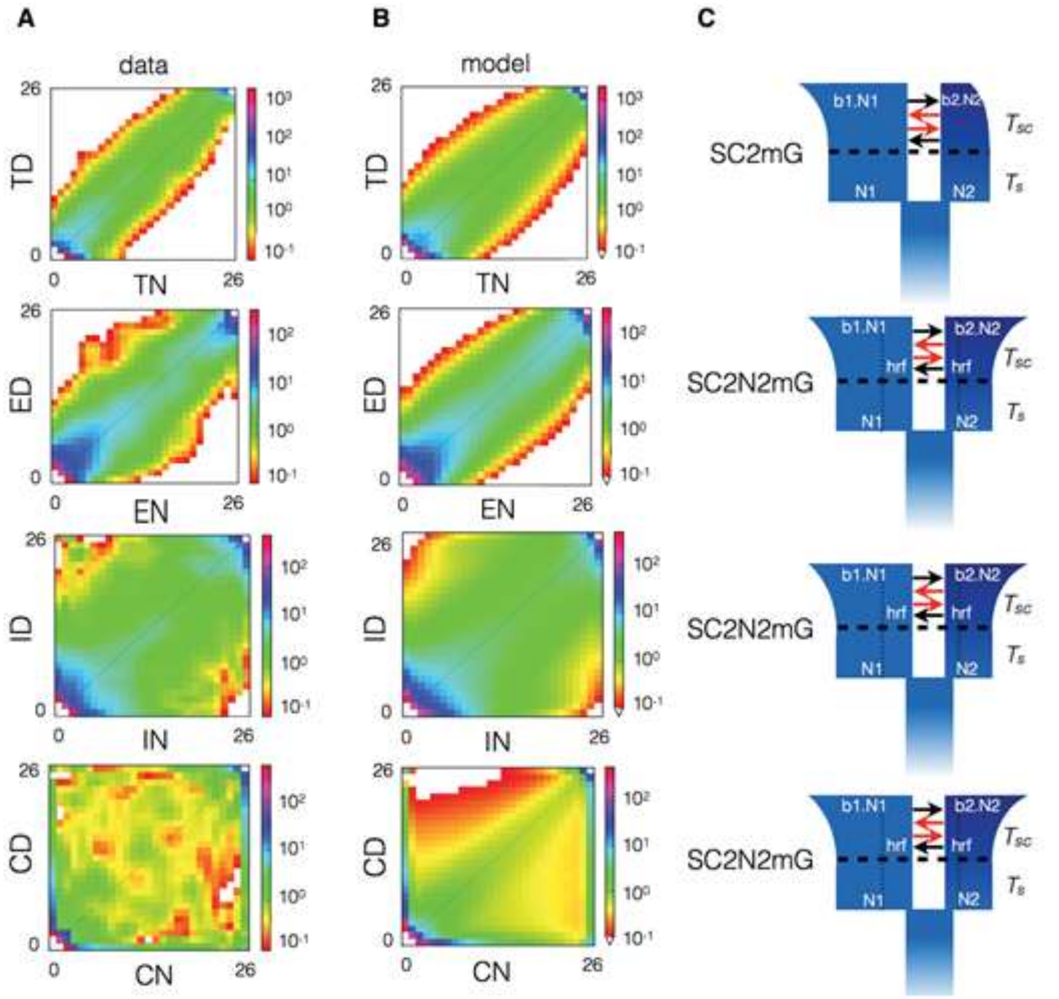
Nature Reviews | Genetics

Beyond present structure…
How to know population history and demography?

Based on coalescence theory

Compare SFS (site frequency spectrum) between real data and modelled data under different scenario

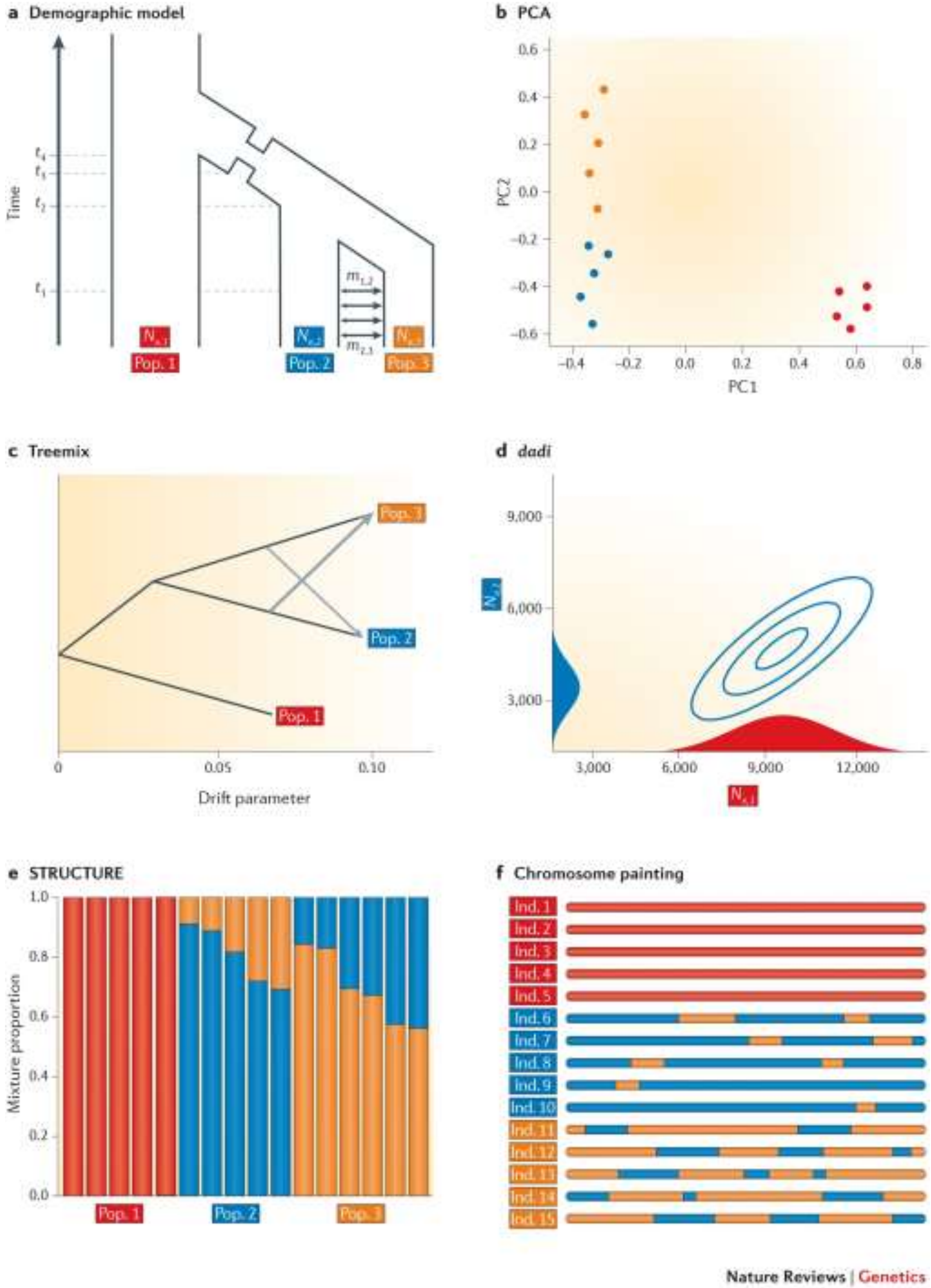
Common tools: dadi, FastSimCoal, ABC…



Population structure and demography

A good overview

Schraiber, J., Akey, J. Methods and models for unravelling human evolutionary history. Nat Rev Genet 16, 727–740 (2015). https://doi.org/10.1038/nrg4005



a Demographic model
Time
t_4
t_3
t_2
t_1
$N_{e,1}$
Pop. 1
$N_{e,2}$
Pop. 2
$m_{1,2}$
$m_{2,1}$
$N_{e,3}$
Pop. 3
b PCA
PC2
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
PC1
c Treemix
Pop. 3
Pop. 2
Pop. 1
0
0.05
0.10
Drift parameter
d dadi
9,000
6,000
3,000
$N_{e,2}$
3,000
6,000
9,000
12,000
$N_{e,1}$
e STRUCTURE
Mixture proportion
1.0
0.8
0.6
0.4
0.2
0.0
Pop. 1
Pop. 2
Pop. 3
f Chromosome painting
Ind. 1
Ind. 2
Ind. 3
Ind. 4
Ind. 5
Ind. 6
Ind. 7
Ind. 8
Ind. 9
Ind. 10
Ind. 11
Ind. 12
Ind. 13
Ind. 14
Ind. 15
Nature Reviews | Genetics