Latent variable models in population genetics

olivier.francois@imag.fr, Flora.jay@imag.fr

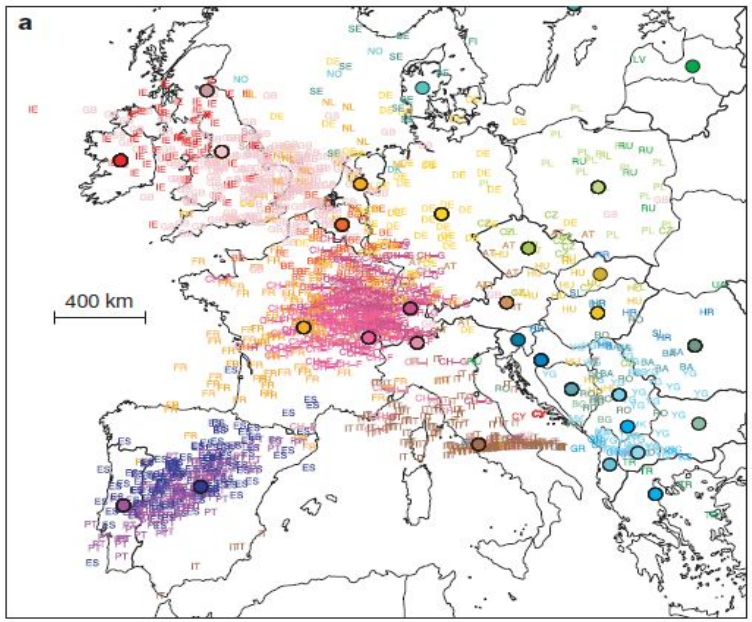
Statlearn'11 - Grenoble, March 17-18, 2011

Outline

- Population genetic structure
- Mixture and admixture models
- Latent variable regression models
- Applications to humans and plants

Population genetic structure

- Many organisms form genetically differentiated subpopulations (herds, colonies, schools, prides, packs).
- Importance of geographic scales (regional, local scales)
- Importance of culture, social rules, habitat (environmental covariates)



Novembre et al (Nature 2008)

Population genetic structure

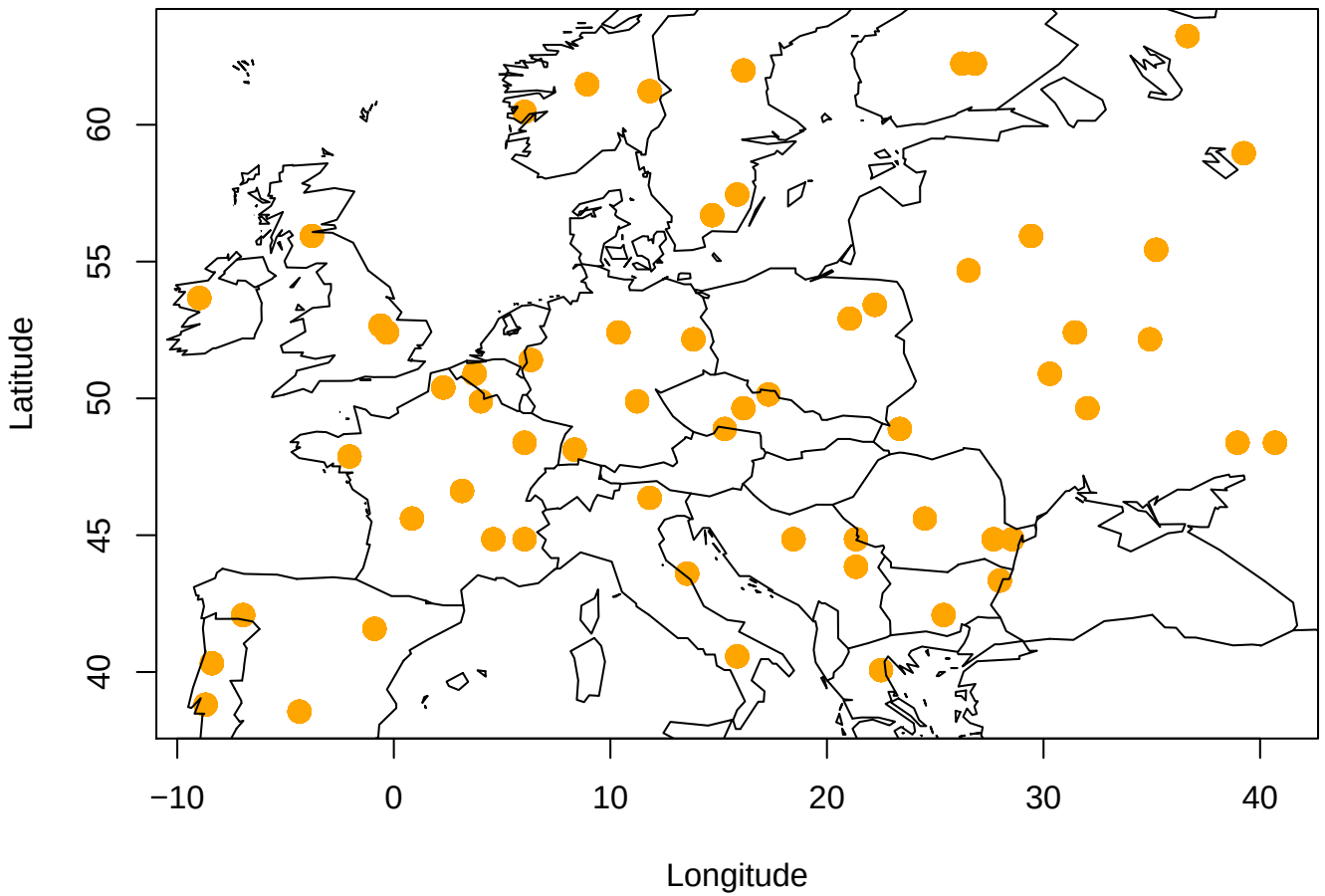
- Natural populations are not random mating populations. There is a geographic range within which individuals are more closely related to one another than to those far apart.
- Population structure is influenced by the demographic history of a species, past events of population fission and fusion, migrations, etc.
- A clear understanding of population structure is useful for detecting genes under selection.
- A clear understanding of population structure is useful for detecting genes associated with particular phenotypes (for example, diseases).

Genotypic data sets

- n (diploid) individuals, L loci
- Standard genetic markers (10-100 polymorphic markers)
- SNPs ($2.5\ 10^5$ - $2.5\ 10^6$ binary markers)
- Nuclear DNA sequences, full genomes

	Loc1.1	Loc1.2	Loc2.1
ind1	11	8	3
ind2	11	7	5
...			

Geographical sampling



Cultural, ecological or environmental covariates

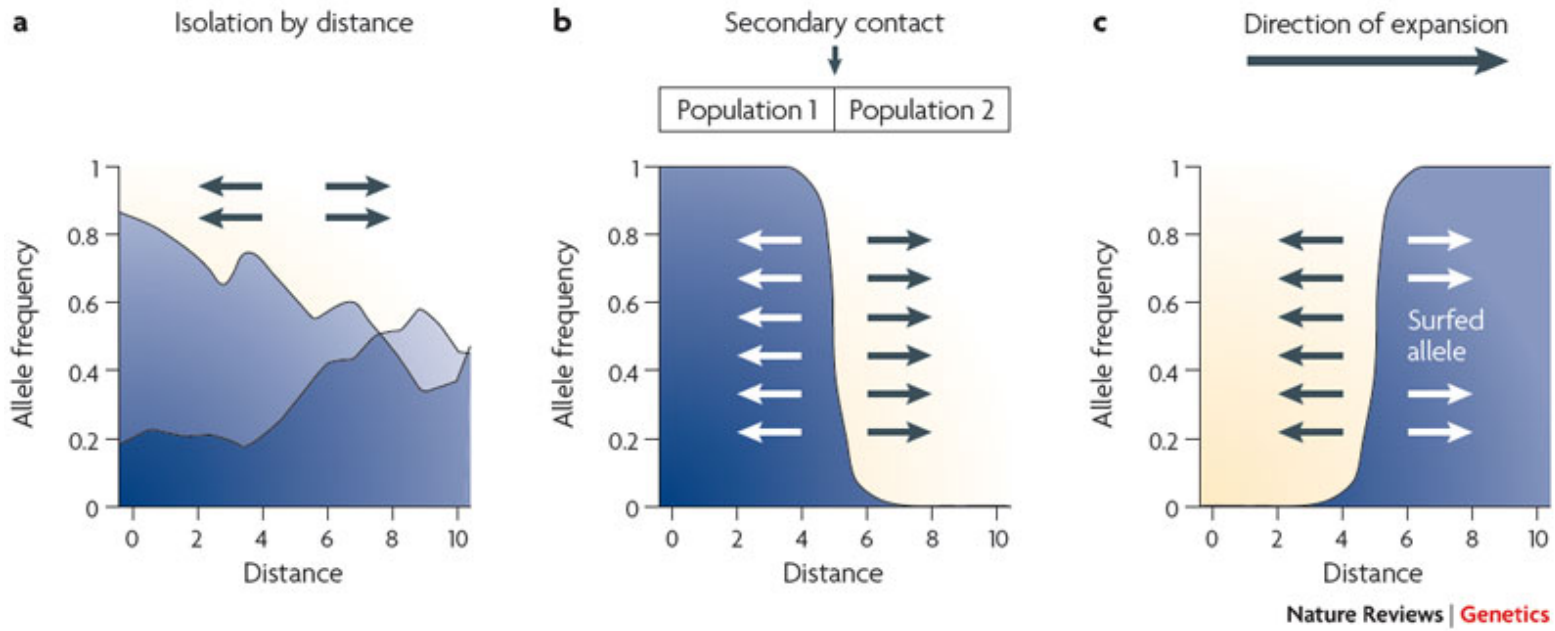
- Language
- Habitat
- Climate
- ...

Regional and local scales

- At the regional scale, clusters and clines are the consequences of demographic processes such as colonization, admixture, reproductive isolation, or selection
- At the local scale, restricted dispersal creates local patches of identical genotypes, spatial autocorrelation of allele frequencies, and long-range isolation by distance patterns

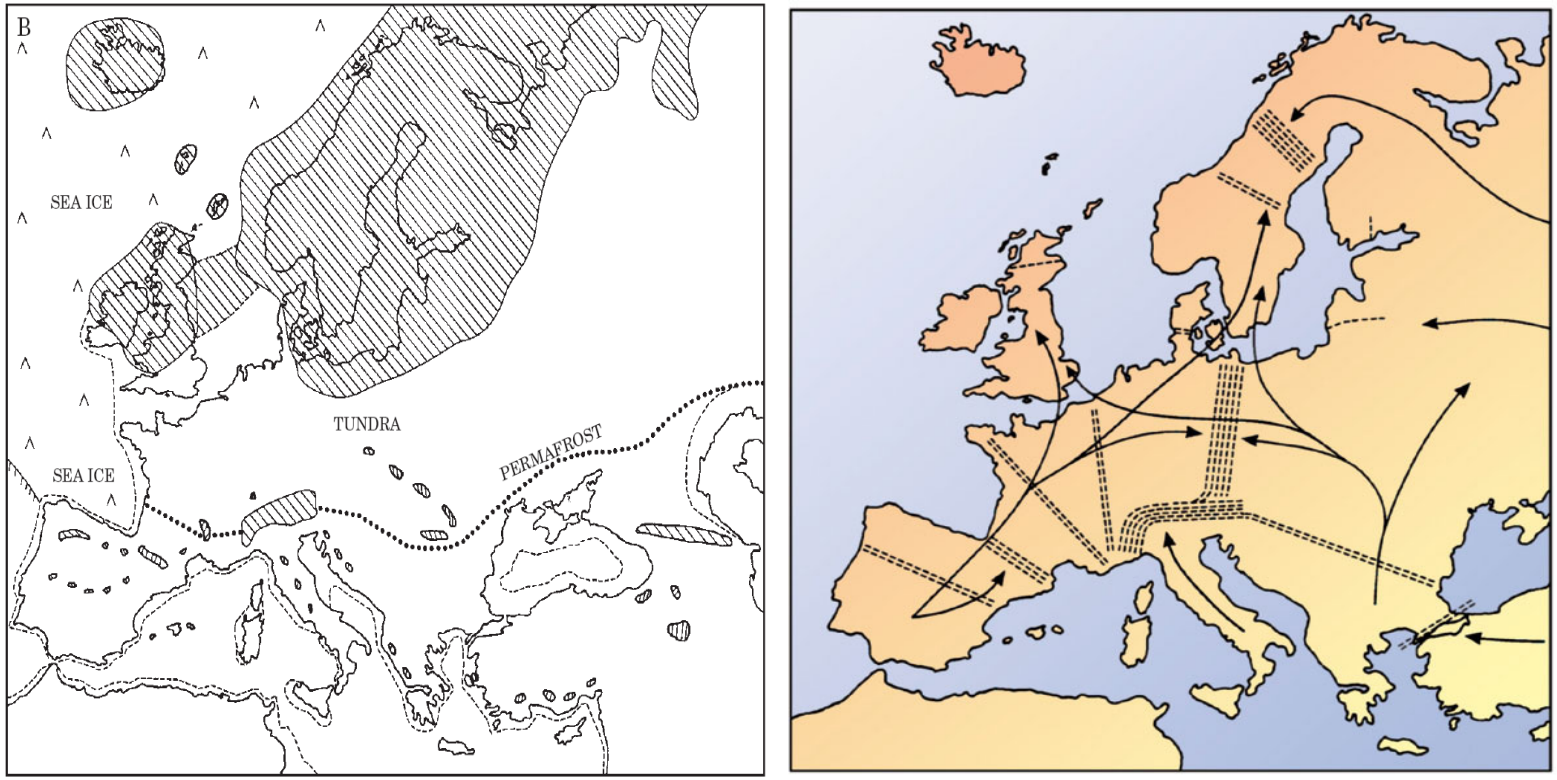
Genetic clusters and clines

- Genetic clusters: Genetically divergent groups of individuals that arise when gene flow is impeded by physical or behavioral obstacles
- Clines: Large-scale spatial trends in allele frequencies or genetic diversity



Novembre and Dirienzo (Nature Review Genetics 2009)

Clusters and clines are not mutually exclusive patterns



Hewitt (Nature 2000)

Basics of population genetics: The Hardy-Weinberg equilibrium model

- Allele and genotype frequencies in a population remain constant from generation to generation.
- It assumes an infinitely large population size, random mating, no mutation, no migration, and selective neutrality.
- Genotype frequencies are deduced from the allele frequencies.

	0 (q)	1 (p)
0 (q)	q^2	pq
1 (p)	pq	p^2

Genotype frequencies at a bi-allelic locus (Single Nucleotide Polymorphism). Heterozygosity $H = 2pq$.

Linkage Disequilibrium (LD)

- Non random association of alleles at two or more loci
- Considering two bi-allelic loci, A and B, LD can be measured by

$$D = p_{AB} - p_A p_B = \mathrm{cov}(A, B)$$

- In the absence of evolutionary forces other than random mating, the linkage disequilibrium measure D converges to zero at a rate equal to the recombination rate between the two loci.

Population structure creates LD at unlinked loci

- Suppose our sample contains two populations in equal proportions
- In population 1, we have $p_A^1 = 1$ and $p_B^1 = 0$ $(D = 0)$
- In population 2, we have $p_A^2 = 0$ and $p_B^2 = 1$ $(D = 0)$
- In the sample,

$$p_A = 1/2 \quad \text{and} \quad p_B = 1/2$$

Thus, because $p_{AB} = 0$, the linkage disequilibrium measure is non zero, and maximal in absolute value

$$|D| = 1/4\,.$$

Population structure creates Hardy-Weinberg disequilibrium

- Suppose our population contains two subpopulations in equal proportions, each in HW equilibrium
- Consider a bi-allelic (0/1) locus and let p_1 and p_2 denote the allele frequencies in subpopulations 1 and 2 (frequencies of 1).
- In the total population, we have

$$p = \frac{p_1 + p_2}{2} \quad \text{and} \quad H = p_1 q_1 + p_2 q_2$$

Thus, $H \neq 2pq$.

Bayesian clustering algorithms

- Assume K unknown subpopulations, n individuals genotyped at L loci.
- Principle: Clusters should maximize HWE and minimize LD
- Mixture model: For each individual i and each cluster k, compute the probability that individual i originates in cluster k.
- Admixture model: For each individual i and each cluster k, compute the fraction of genome of individual i that originates in cluster k.

Mixture model

- Data: $(y_{i\ell})_{i\leq n,\ell\leq L}$ is a matrix of 0 and 1, where each individual i is coded with 2 rows.
- Clusters: $(z_i)_{i\leq n}$, $z_i \in \{1,\ldots,K\}$
- Allele frequencies: Given $z_i = k$,

$$\Pr(y_{i\ell} = 1 \mid z_i = k, p) = p_{\ell,k}$$

and

$$\Pr(y|\ z, p) = \prod_{i=1}^{n}\prod_{\ell=1}^{L}\prod_{j=1}^{2} p_{\ell,z_i}^{y_i^j}(1-p_{\ell,z_i})^{1-y_i^j}$$

Prior distributions

- Independent allele frequencies are sampled at each locus ℓ and in each cluster k

$$p_{\ell k} \sim \mathrm{beta}(\lambda_1, \lambda_2) \quad (\text{default value } \lambda_i = 1).$$

- Uniform distribution on individual cluster labels

$$\Pr(z_i = k) = \frac{1}{K}, \quad k = 1, \ldots, K.$$

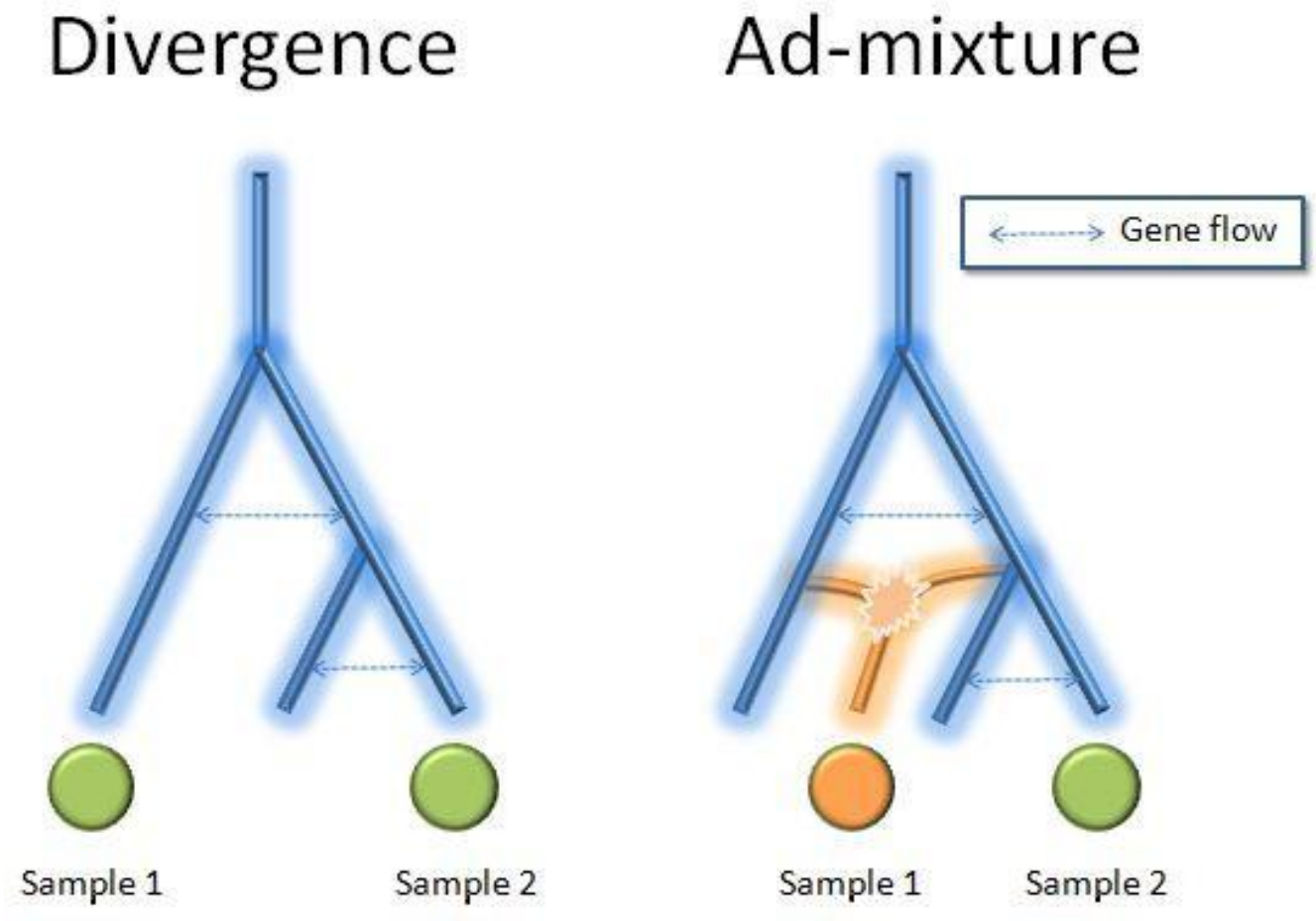
Statistical mixture model

- Known for long in statistics as the *latent class model* (Lazarsfeld and Henry 1968, Goodman 1974)
- Bayesian implementation popularized by the software `structure` (Pritchard et al 2000)

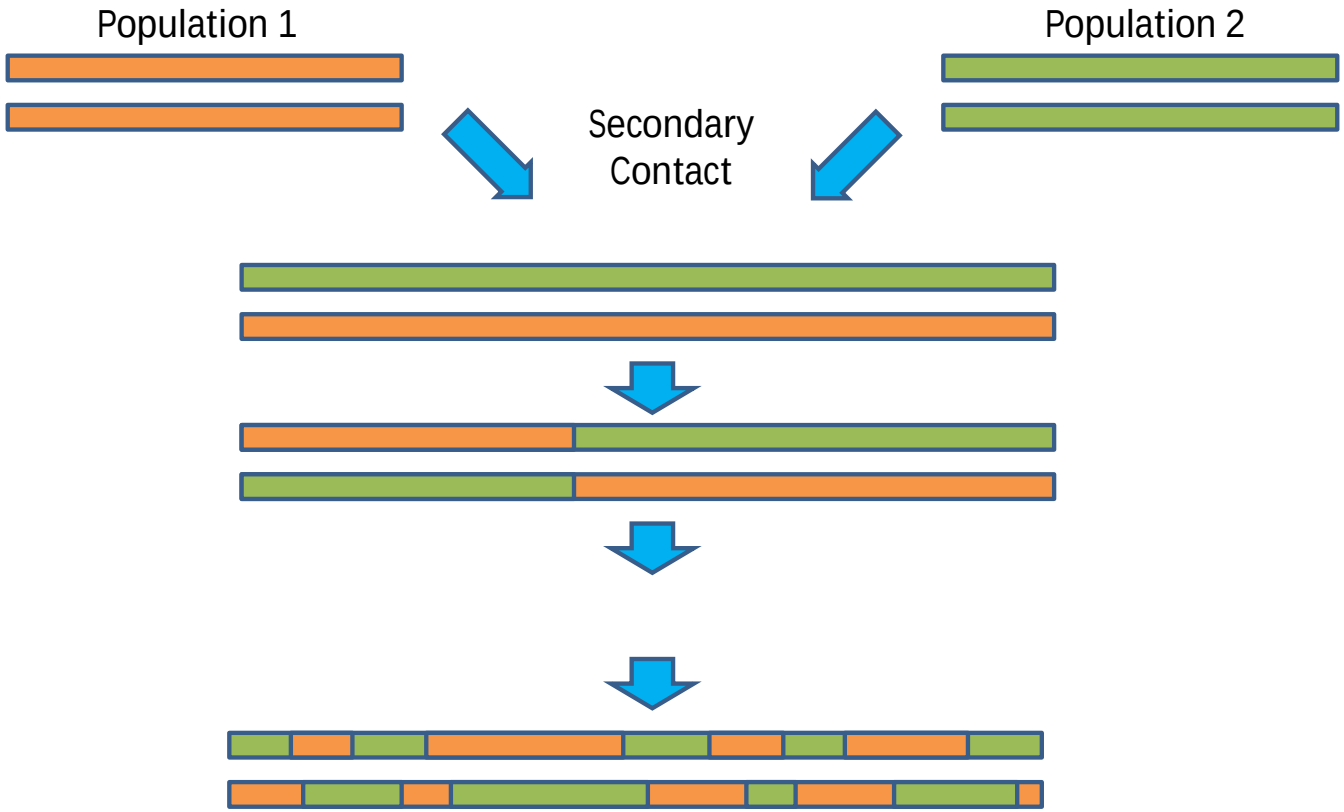
Admixture models

- Genetic admixture is the process by which a hybrid population is formed from contributions by two or more parental (or ancestral) populations
- In an admixed population, individual genomes are themselves (to a greater or a lesser extent) admixed.
- Bayesian clustering methods are capable of calculating individual admixture proportions where the ancestral populations are not imposed by the sampling process.

Divergence vs Admixture of populations



Admixture and LD



Admixture model

- Introduction of additional parameters: Q-matrix ($n \times K$ dimensions)

  $q_{i,k}$ = proportion of individual i's genome from population k

- One cluster for each allele copy

  $$z_{i,\ell} = \text{population of origin of allele copy } y_{i,\ell}$$

  and

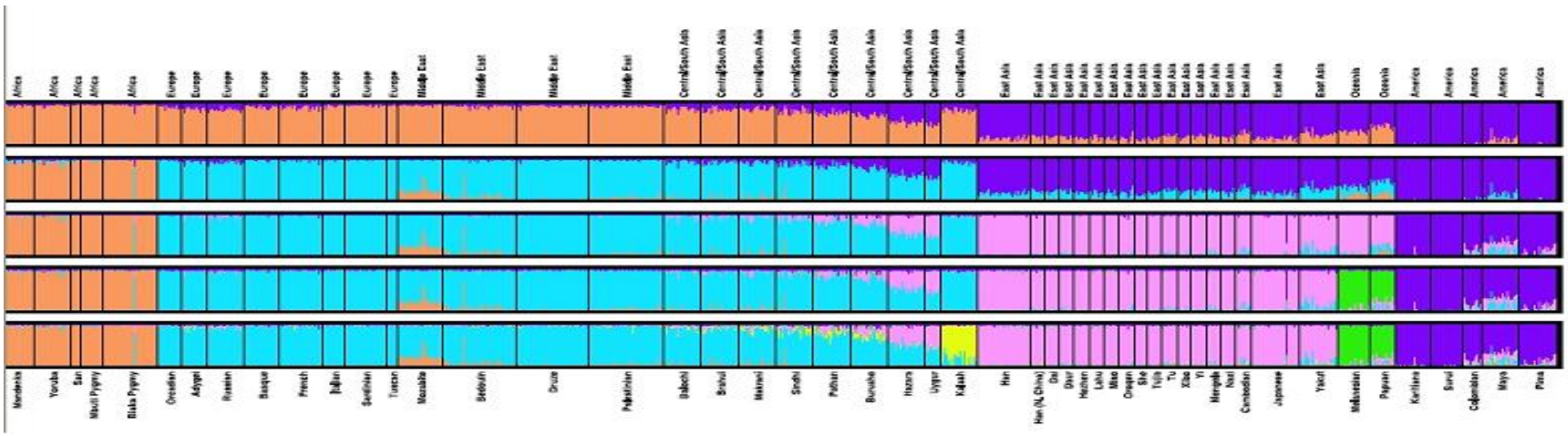
  $$\Pr(z_{i,\ell} = k \mid p, q) = q_{i,k}$$

  where

  $$q_{i,\cdot} \sim \mathcal{D}(\alpha_1, \ldots, \alpha_K) \quad \text{Dirichlet distribution.}$$

Genetic structure of human populations

- Each individual is represented by a segment of total length 1.
- Admixture proportions are represented by colored segments.



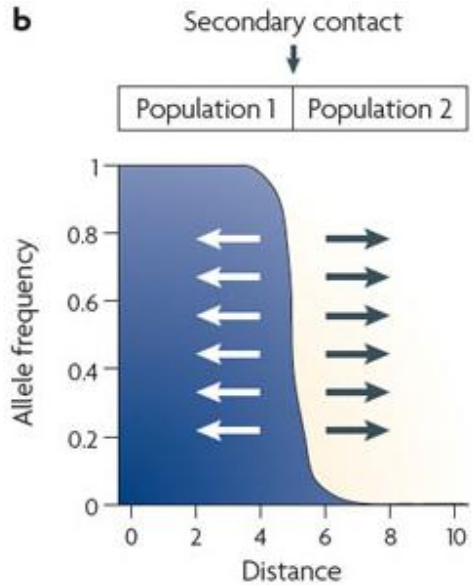
Rosenberg et al (Science 2003)

Connections

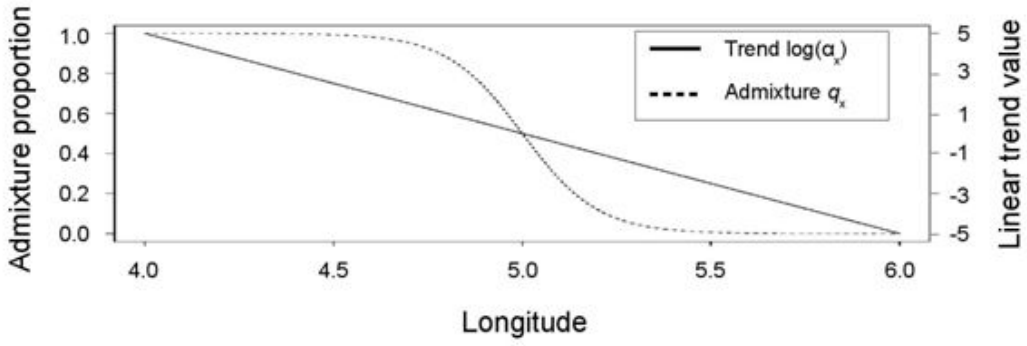
Population genetics	Statistical learning
`structure` Pritchard et al 2000 Mixture model Admixture model	Latent class models Lazarsfeld and Henry 1968 Latent Dirichlet Allocation Blei et al 2003
`structurama` Dirichlet process Hueselenbeck et al 2007	Hierarchical topic models Blei et al 2004
Hidden probit model Jay et al 2011	Latent class regression models Bandeen-Roche et al 1999
`pca` Patterson et al 2006 `sfa` Stephens 2010	probabilistic PCA Non-negative matrix factorization

How can we account for geography in admixture models?

- Gradients in gene frequencies are created by the contact of two or more populations.
- Their shape is sigmoidal (Barton and Hewitt 1986)



Modeling the cline



$$\log(\alpha_{i,k}) = f(x_i)^T \beta_k$$

$$\mathrm{E}[q_{i,k}] = \frac{\alpha_{i,k}}{\Sigma_{k'=1}^{K} \alpha_{i,k'}}$$

Admixture proportions vary logistically

Non-linear trend surface

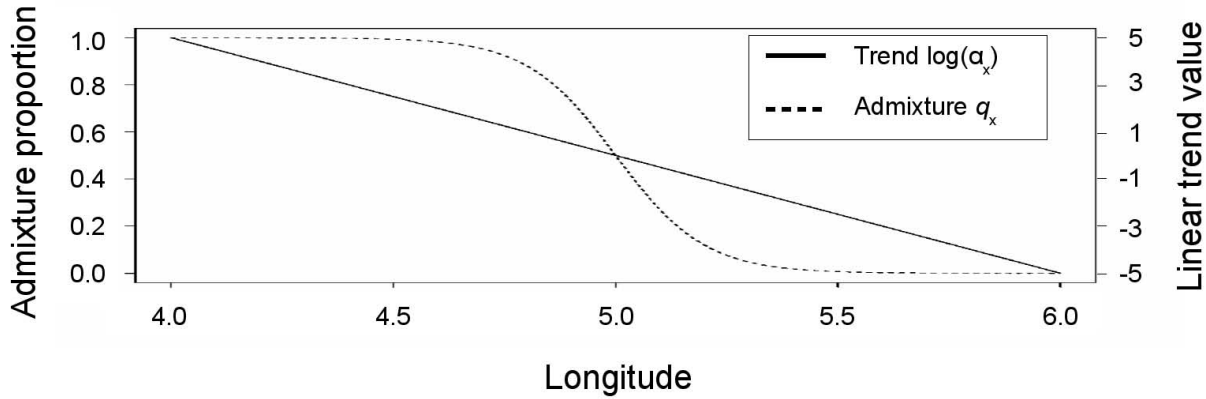
Estimated from the spatial and genetic data

Extension of the structure algorithm: including spatial information

- Population genetic structure is spatially structured. Let x_i be the spatial coordinates of individual i. We assume the a kriging model for the prior distribution

$$\log \alpha_{i.} = f(x_i)^T \beta_{.} + \epsilon_{i.}$$

where the hyper-prior distribution on β is non-informative and ϵ is a spatially autocorrelated Gaussian noise.

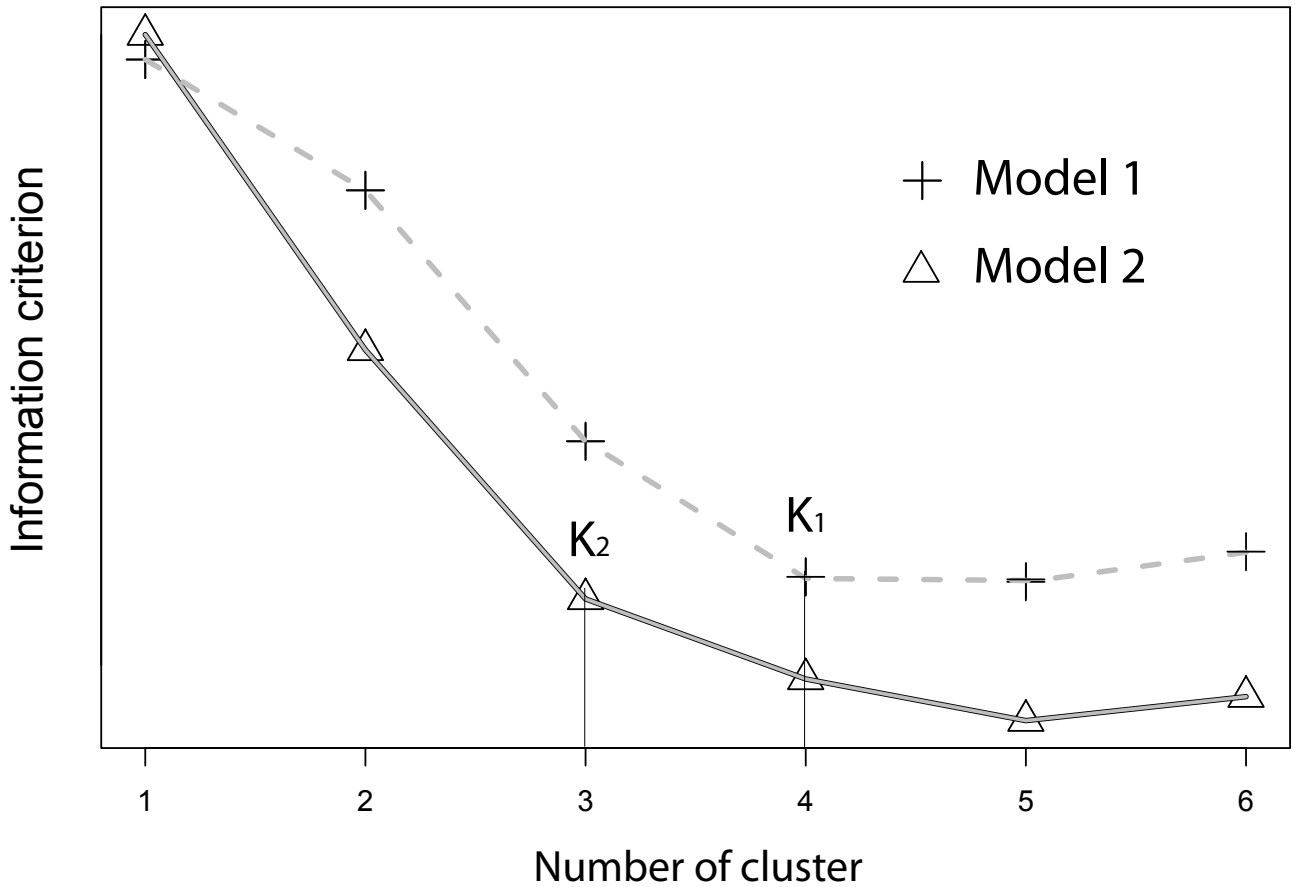


- tess model (Durand et al Mol Biol Evol 2009).

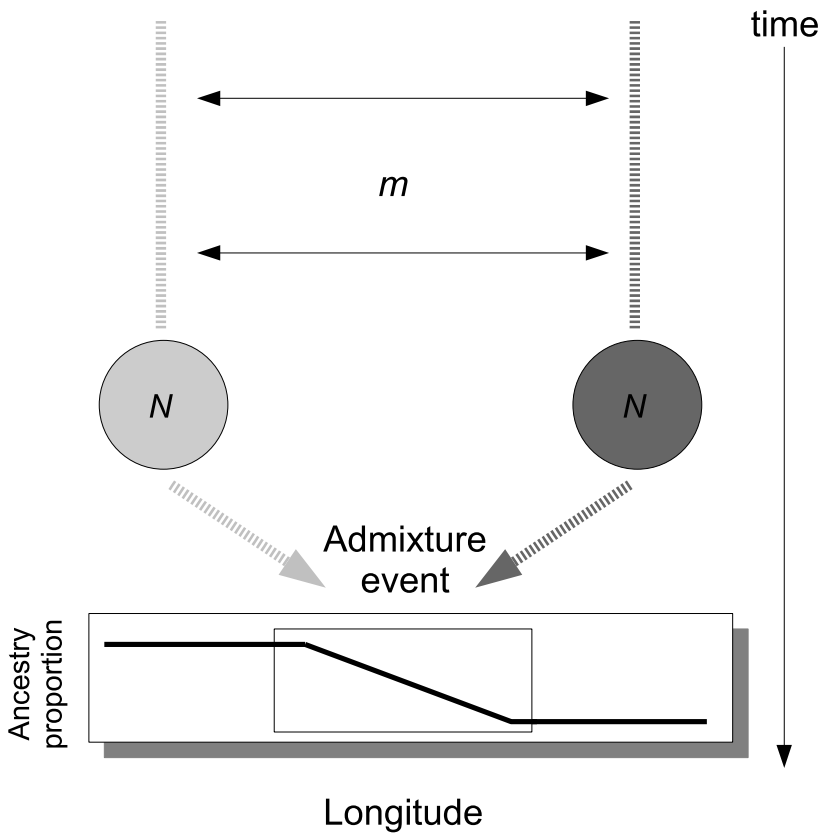
Implementation and choice of K

- Inference based on Gibbs samplers
- Model choice (K) is a difficult issue
- Deviance Information Criterion
- Cross Validation

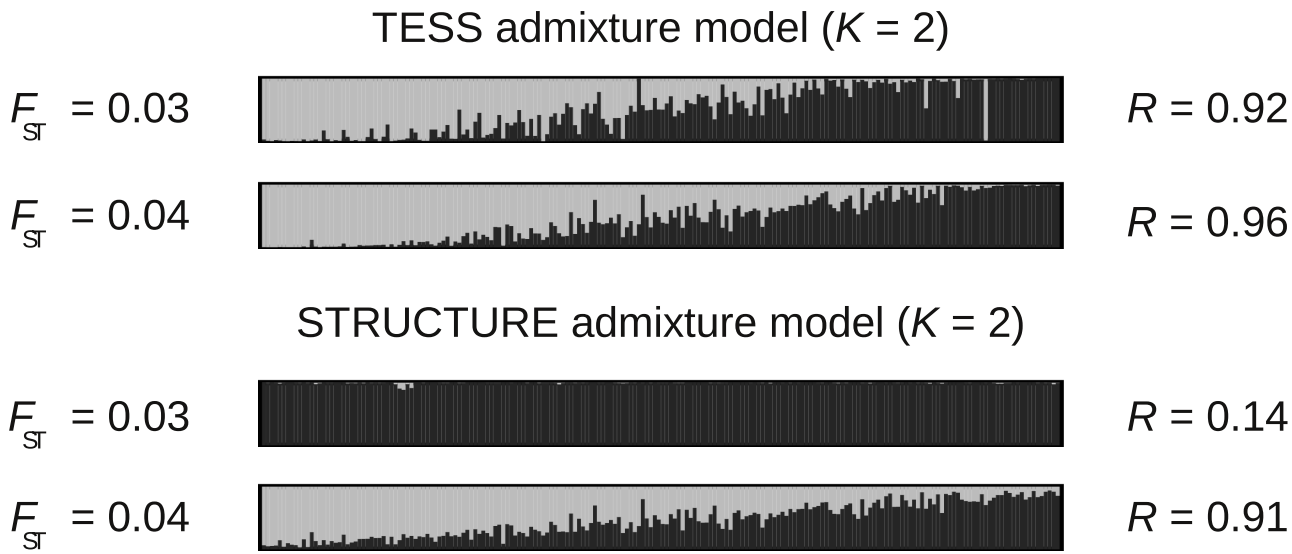
Choosing K – DIC curves



Geographic admixture of 2 parental populations (simple model)

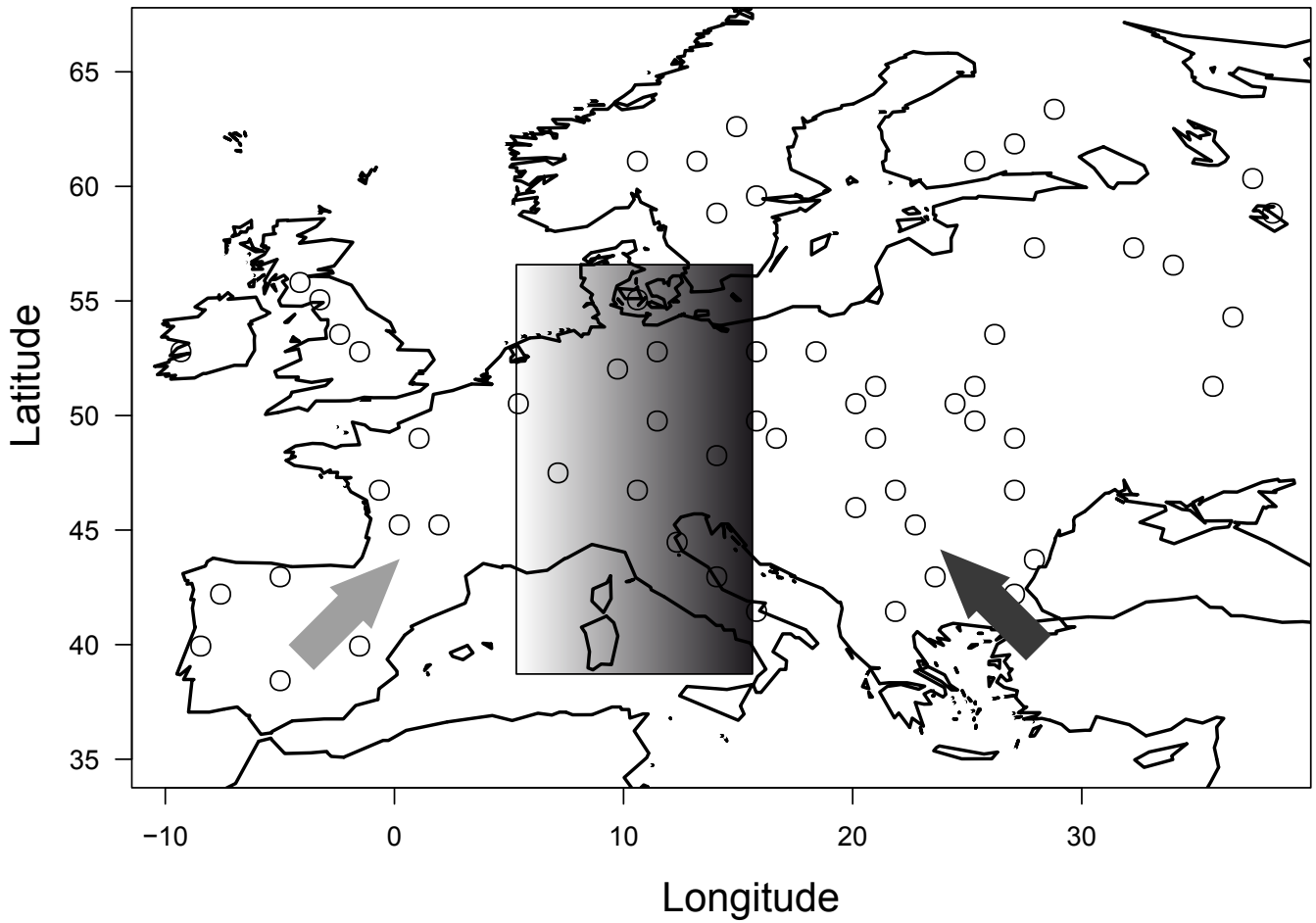


Results of `structure` and `tess`



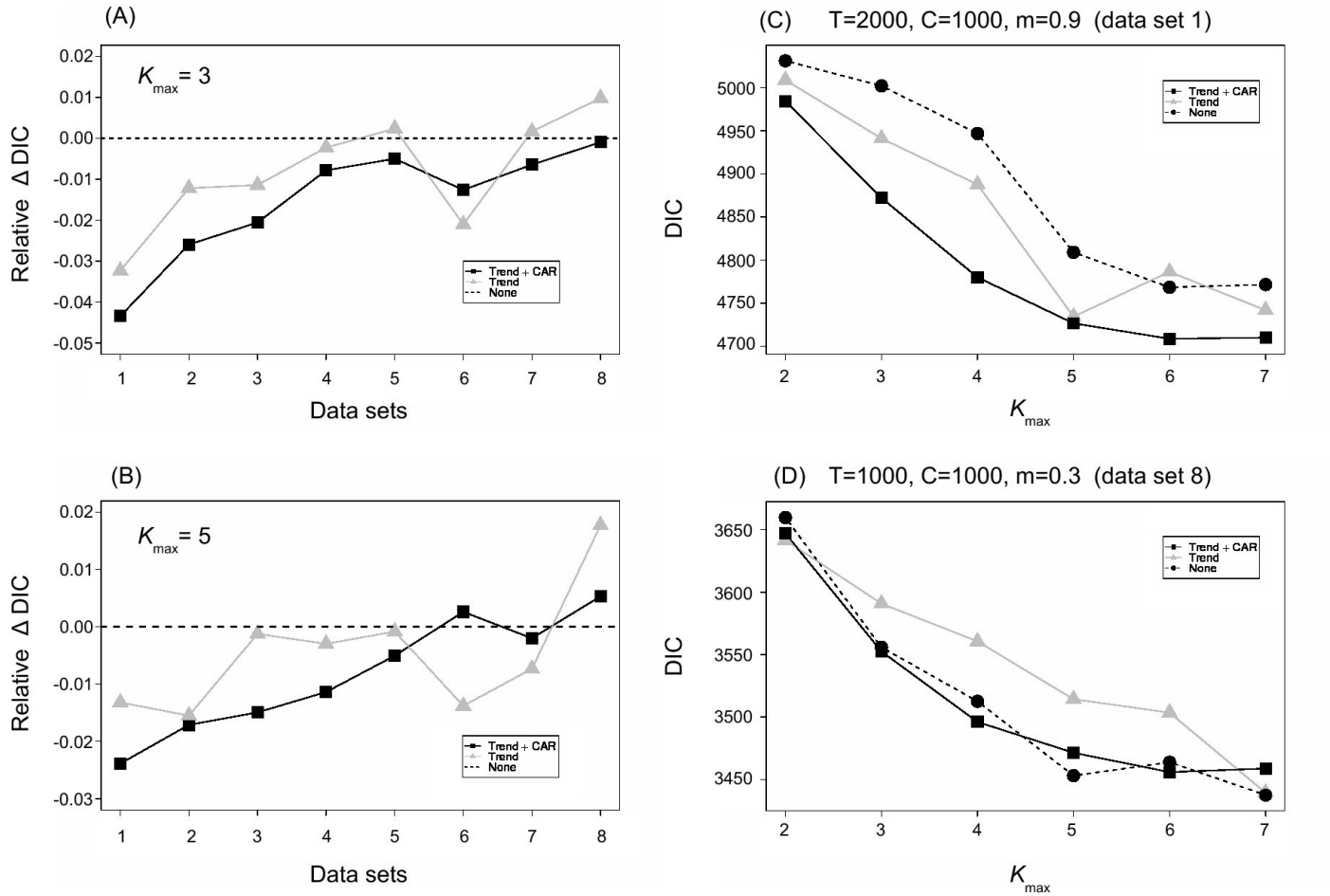
- F_{ST} (of the pooled parental populations) is a measure that quantifies the departure from HWE in the ancestral population

A realistic scenario for a contact zone (Non equilibrium stepping-stone simulation) (Movie)

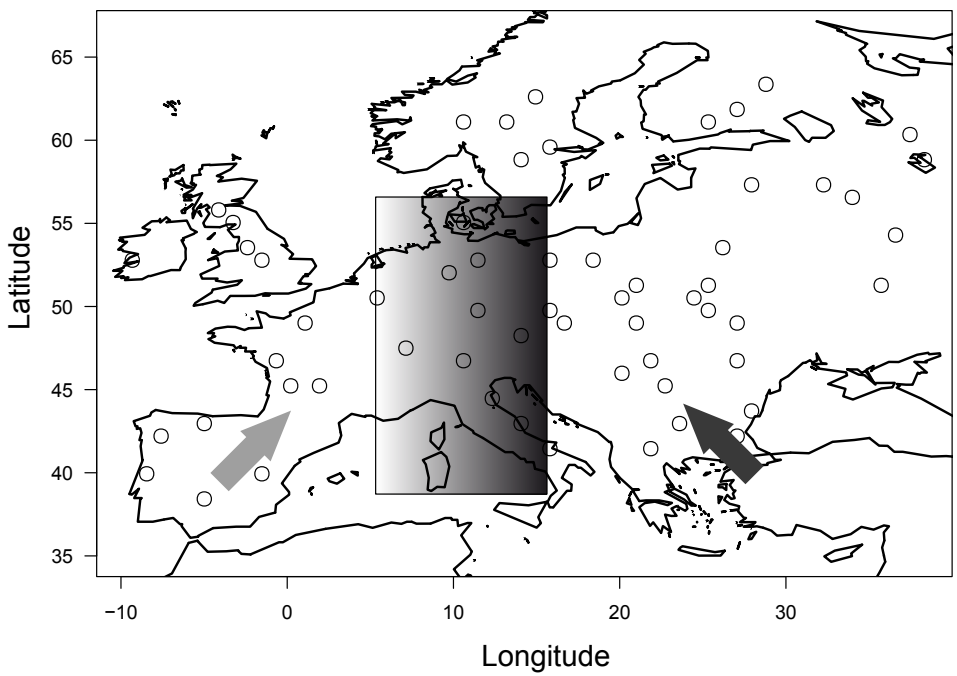


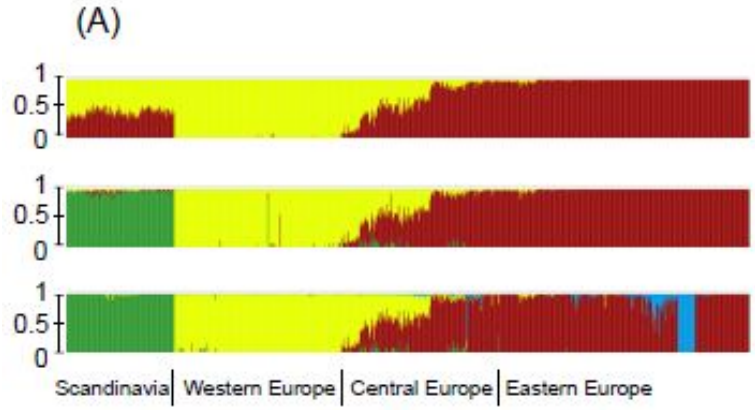
Results

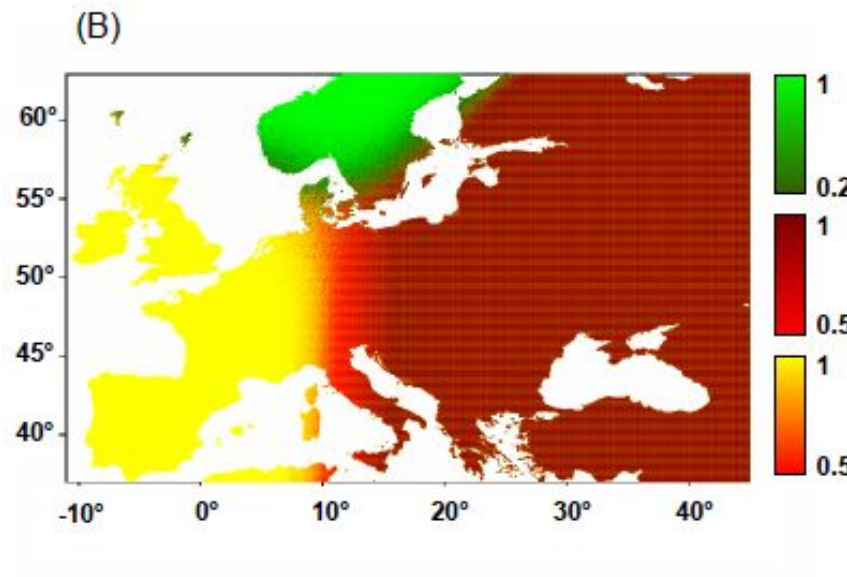
8 data sets – Each with 100 polymorphic markers (20 alleles) for 1200 individuals (20 populations)



Results of inference: Visualizing the contact zone



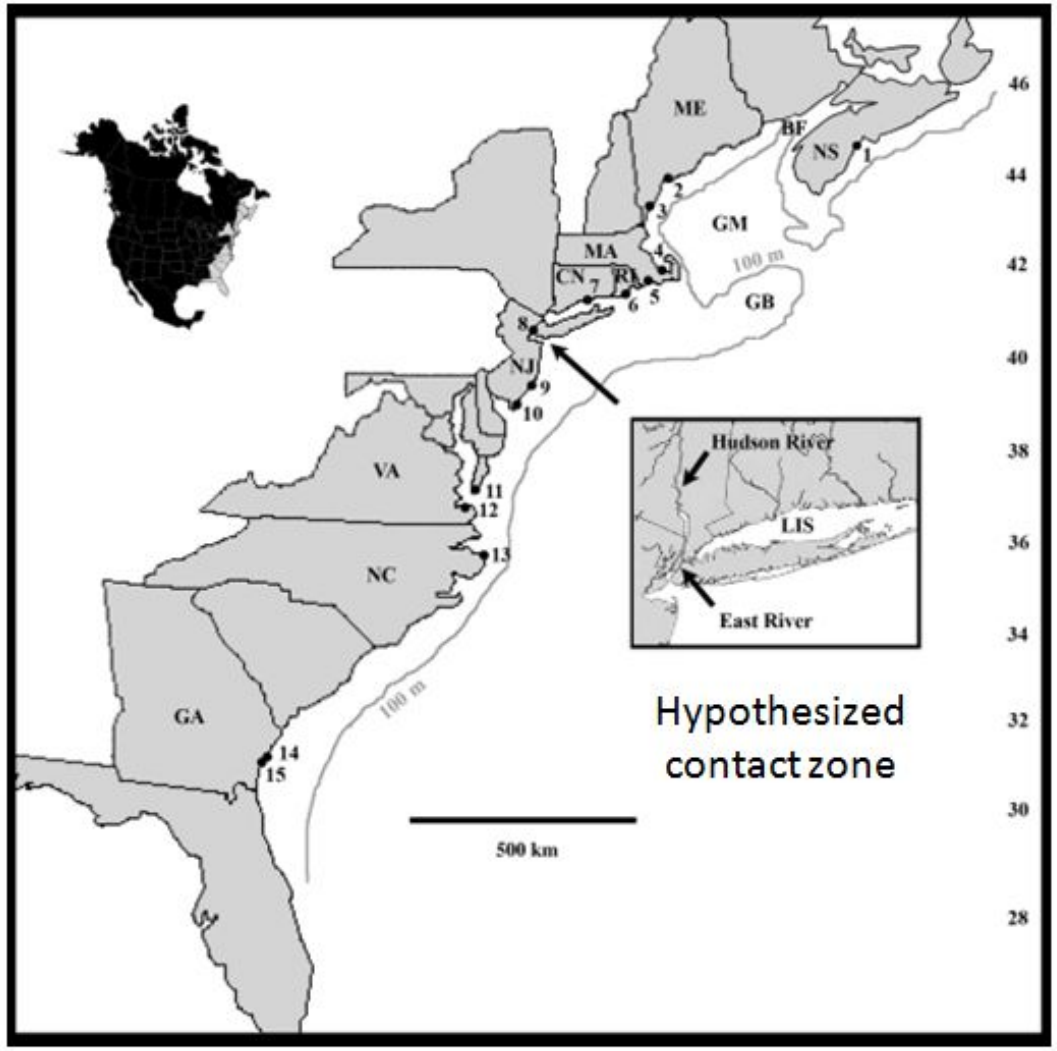




Application to *Fundulus heteroclitus*

Fundulus heteroclitus
722 individuals
15 populations

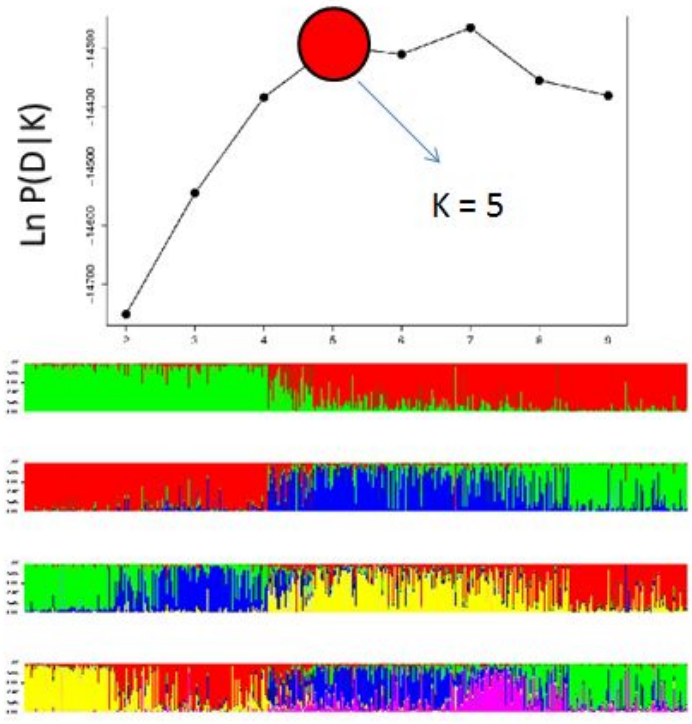
Adams et al 2006



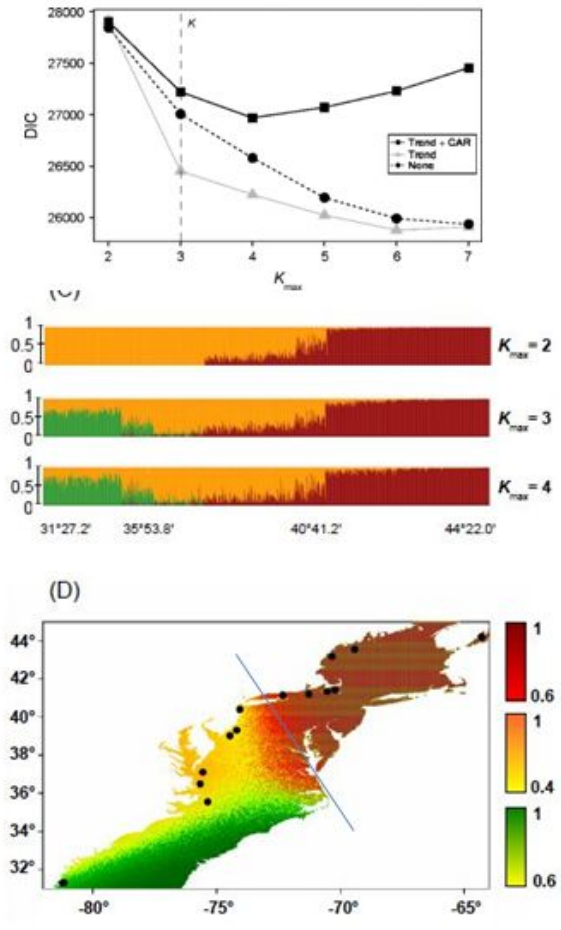
Application to *Fundulus heteroclitus*

Data: 722 individuals, 8 markers

STRUCTURE



TESS (CAR Model)



Using covariates: Latent class probit regression model (Jay et al 2011)

- In the multinomial probit model, there $K-1$ regression equations

$$W_{ik} = \tilde{X}_i \beta_k + \epsilon_{ik}$$

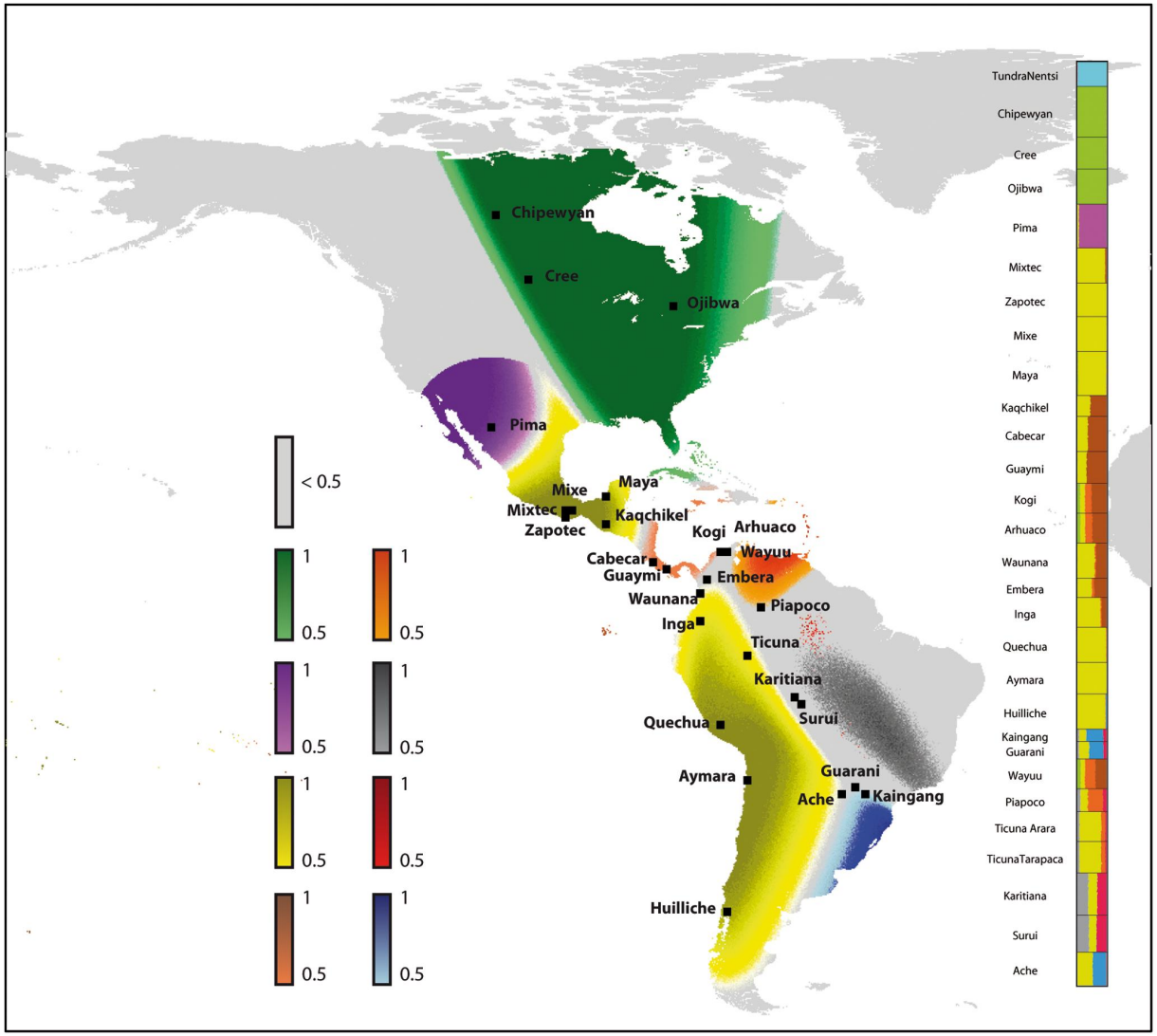
- A hidden cluster label variable can be obtained as follows

$$z_i = \begin{cases} K & \text{if } \max W_{i\ell} < 0 \\ k & \text{if } W_{ik} = \max W_{i\ell} > 0 \end{cases}$$

- Bayesian inference

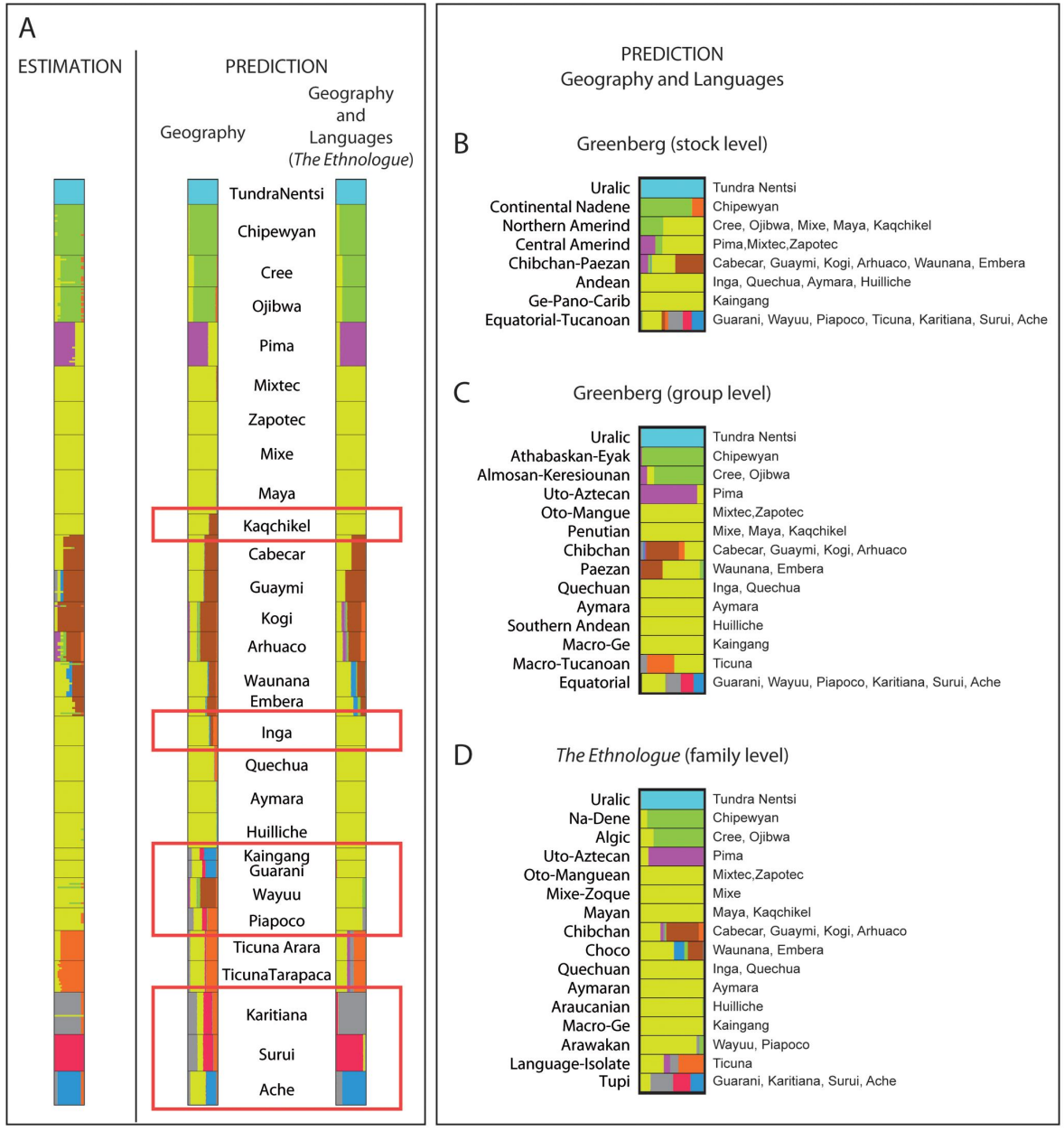
$$\Pr(z, \beta, p|y) \propto \Pr(y|z, p)\Pr(z|\beta)\Pr(\beta)\Pr(p)$$

Hidden regression model (Jay et al 2011)



Population structure of Native American populations (HGDP data set, 512 individuals – 678 markers)

Genes and languages in the Americas



Admixture inference with covariates

- `pops admixture model`. Let x_i^s be the spatial coordinates of individual i, and x_i^e a set of environmental covariates (eg, climatic variables).

$$\log \alpha_{i.} = f(x_i^s)^T \beta_.^s + f(x_i^e)^T \beta_.^e + \epsilon_{i.}$$

  where the prior distribution on β is non-informative and ϵ is a spatially autocorrelated Gaussian noise.

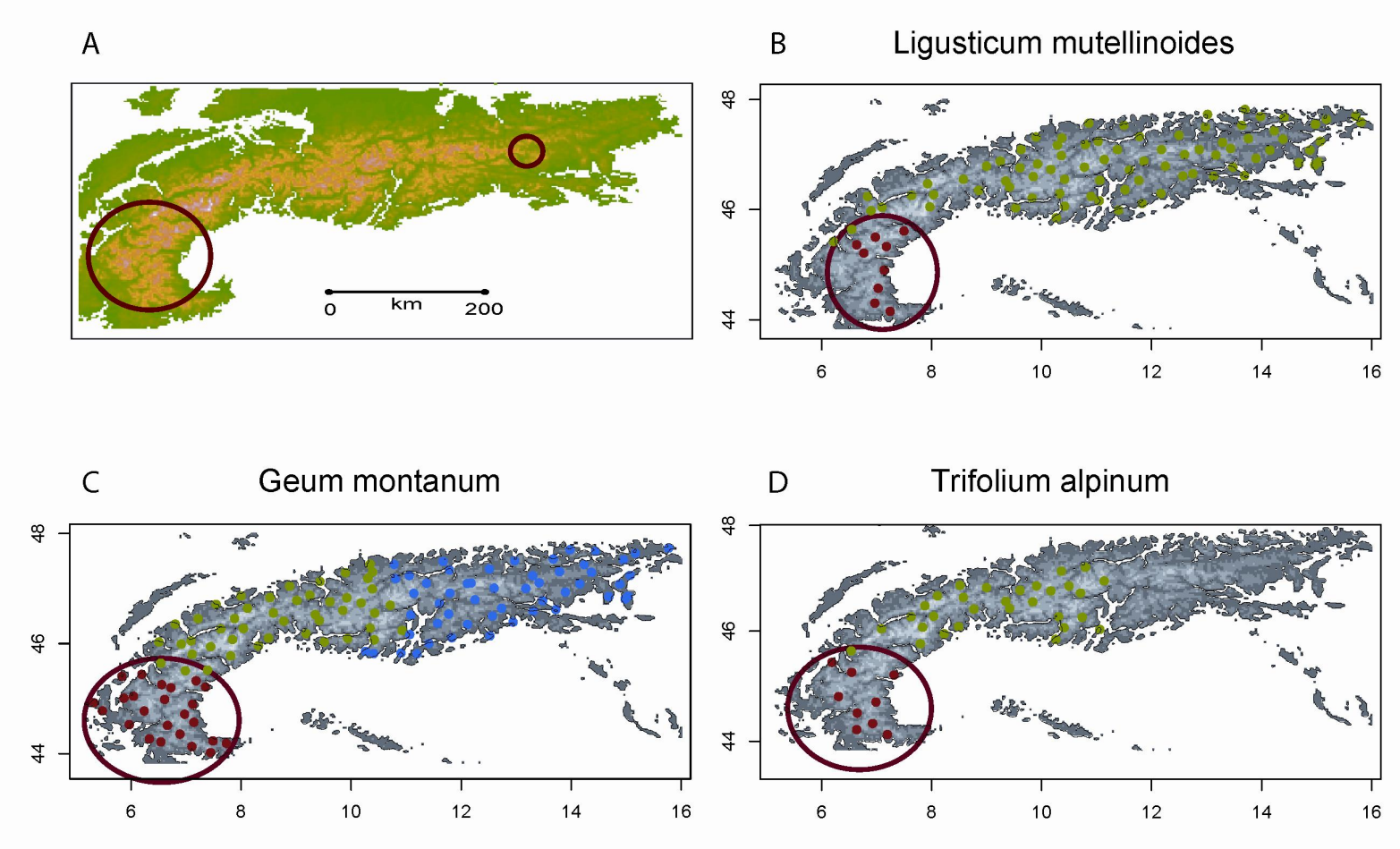
- α_i is proportional to the 'average' individual admixture coefficient.
- Meets classical ecological modeling assumptions (Lichstein et al 2002)

Application to alpine plants (Intrabiodiv project, collab. LECA)

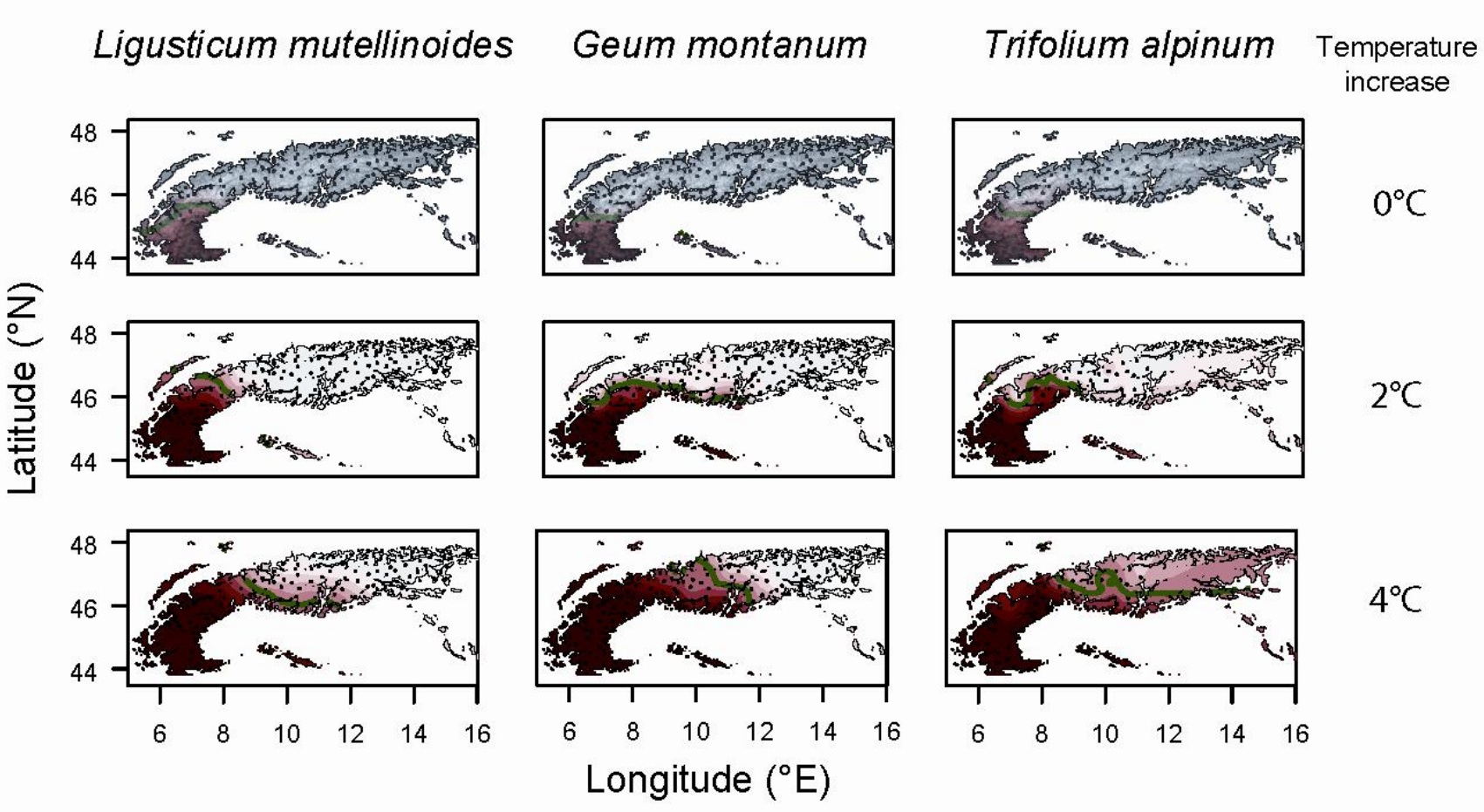
20 species sampled in the Alps ($\approx$ 300 individuals, 200 markers for each species)

Ligusticum mutellinoides – *Geum montanum* – *Trifolium alpinum*

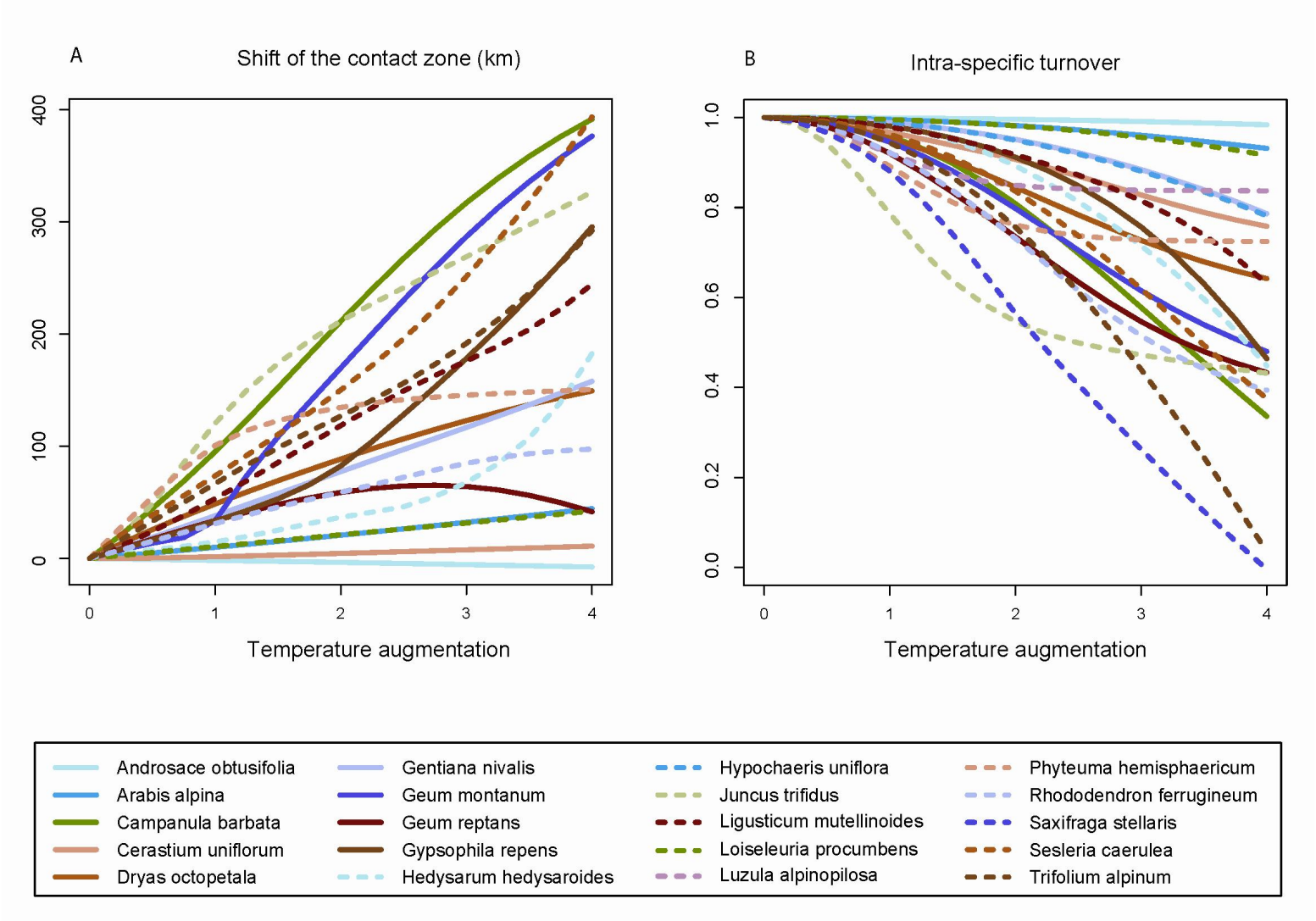
Population structure



Prediction of intra-specific turnover under scenarios of climate warming (Movie)



Prediction of intra-specific turnover under scenarios of climate warming



Concluding messages

- Bayesian algorithms detect population structure and individual admixture levels without a need to predefine ancestral populations
- Landscape genetics: Include ecological and geographic covariates in the inference of population structure
- Current and future developments: Explore connections with topic and probabilistic PCA models
- More developments: Include local adaptation models in the inference of allele frequencies

Acknowledgments

- Eric Durand and Michael Blum (Computational and Mathematical Biology, TIMC-IMAG/BCM)
- Chibiao Chen, Florence Forbes (INRIA)
- Oscar Gaggiotti, Stephanie Manel (LECA)
- ANR MAEV, Complex Systems Institute (IXXI) and PEPS CNRS project *Computational landscape genetics*.