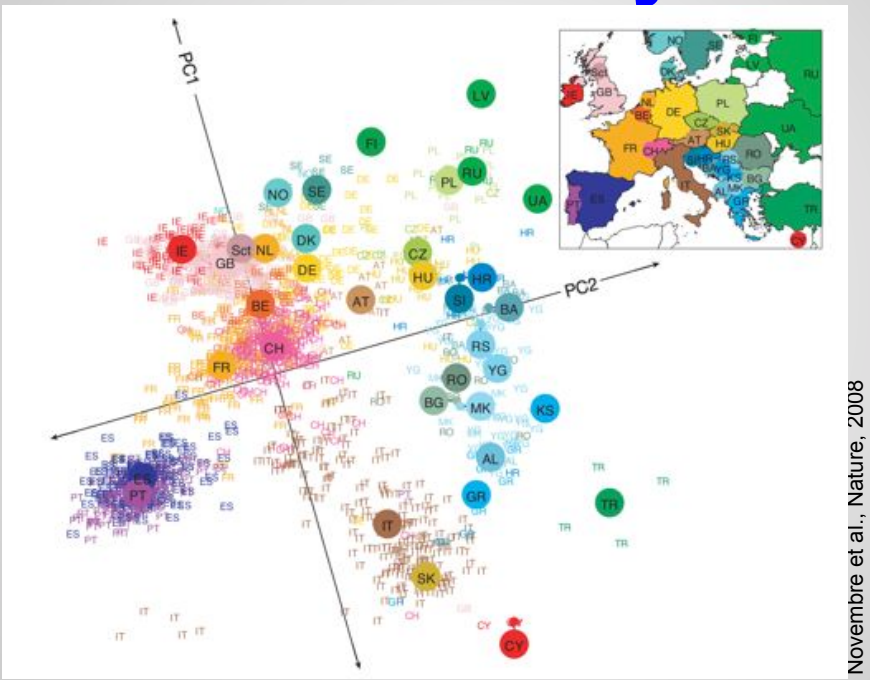
Advanced data analysis in population genetics

Demographic inference

under isolation by distance



Novembre et al., Nature, 2008

Raphael Leblois

Centre de Biologie pour la Gestion des Populations (CBGP), INRA, Montpellier

master B2E, Décembre 2012

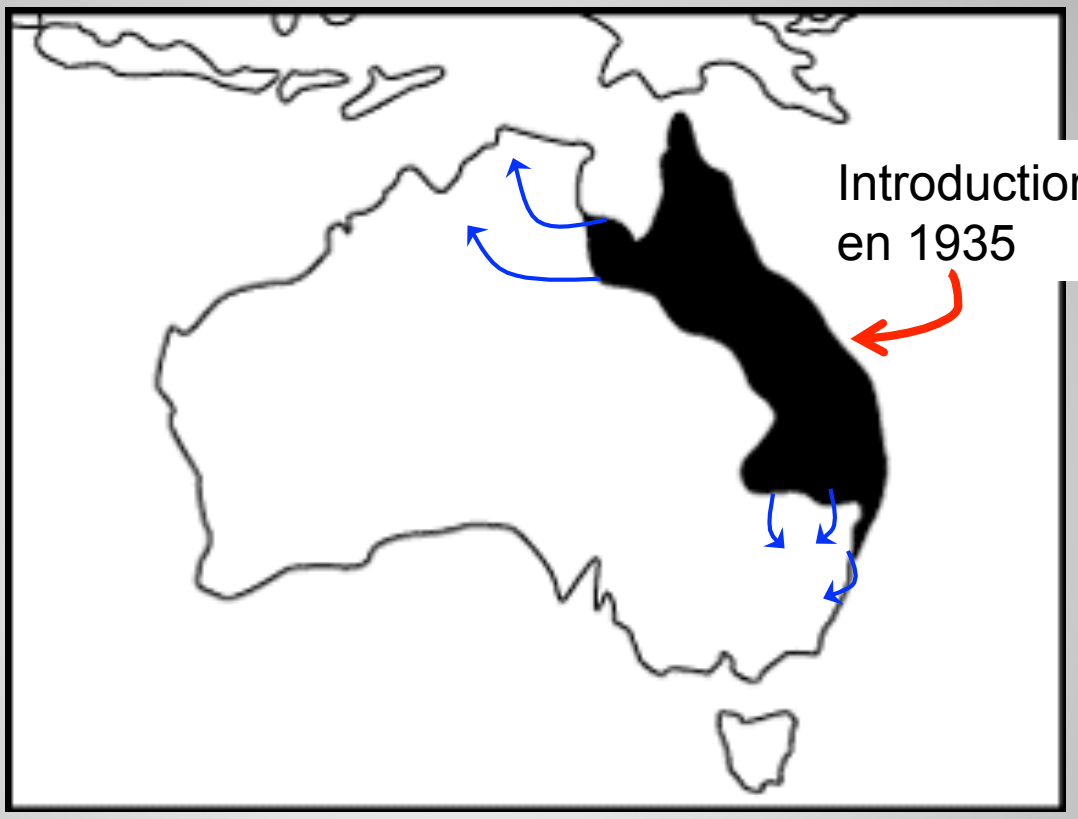
Advanced data analysis in population genetics

Demographic inference under isolation by distance

1. Demographic inference and population genetic models
2. IBD models
3. A simple inference method : Rousset's regression
4. Examples : some real data sets analyses (Pygmies and Damselflies)
5. Testing inference methods : application to the regression method
6. IBD between two habitats
7. Landscape genetics based on IBD
8. Other reasons to test and quantify IBD

Exemple d'une espèce invasive

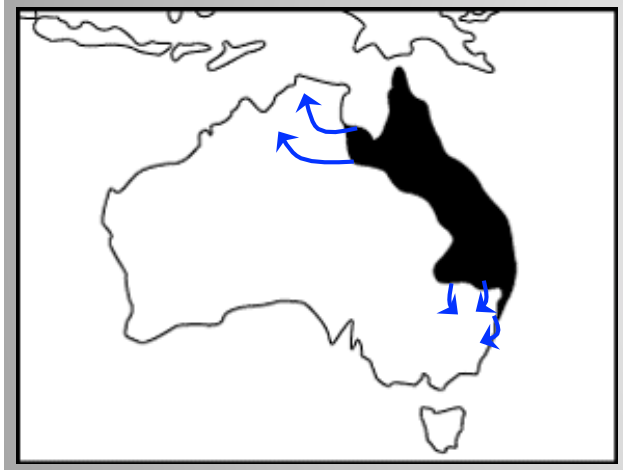

Crapaud de la canne à sucre



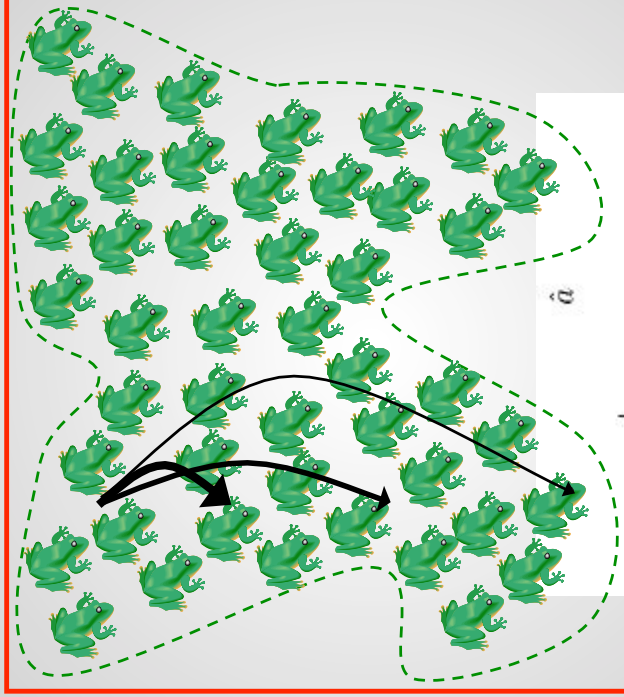
- colonisation extrêmement rapide de l'Australie, plus rapide au Nord qu'au Sud
- Comment : homme? transports?

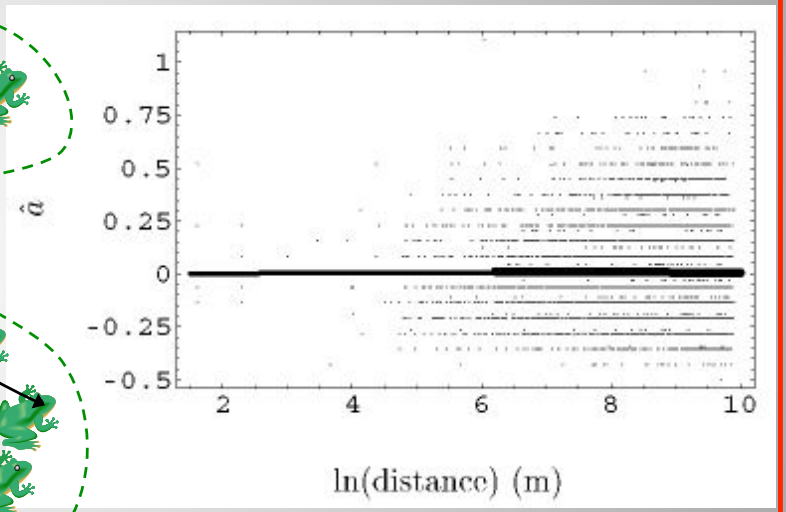
Exemple d'une espèce invasive


Crapaud de la canne à sucre



➢ pas d'isolement par la distance





➢ Pas d'isolement par la distance significatif dans les pops envahissantes -> Forte dispersion lors de l'invasion

➢ Peut expliquer la colonisation rapide de la côte Est de l'Australie (50km par an), dispersion par l'homme pas forcément en cause

"seascape" genetics on the North-Atlantic harbour porpoise

[Fontaine et al., 2007]

BMC Biology

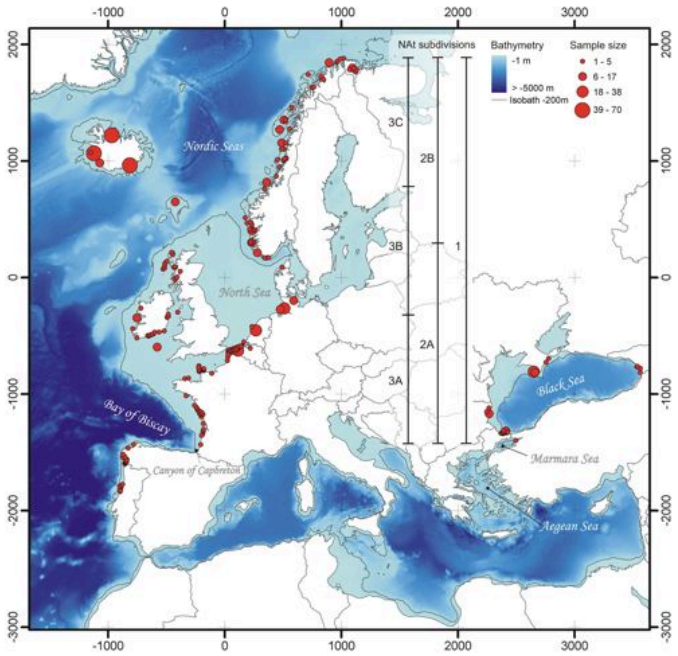
BioMed Central

Research article

Open Access

Rise of oceanographic barriers in continuous populations of a cetacean: the genetic structure of harbour porpoises in Old World waters

Michaël C Fontaine*[1,2], Stuart JE Baird[2], Sylvain Piry[2], Nicolas Ray[3],



"seascape" genetics on the North-Atlantic harbour porpoise

Rise of oceanographic barriers in continuous populations of a cetacean: the genetic structure of harbour porpoises in Old World waters

Abstract

Background: Understanding the role of seascape in shaping genetic and demographic population structure is highly challenging for marine pelagic species such as cetaceans for which there is generally little evidence of what could effectively restrict their dispersal. In the present work, we applied a combination of recent individual-based landscape genetic approaches to investigate the population genetic structure of a highly mobile extensive range cetacean, the harbour porpoise in the eastern North Atlantic, with regards to oceanographic characteristics that could constrain its dispersal.

Results: Analyses of 10 microsatellite loci for 752 individuals revealed that most of the sampled range in the eastern North Atlantic behaves as a 'continuous' population that widely extends over thousands of kilometres with significant isolation by distance (IBD). However, strong barriers to gene flow were detected in the south-eastern part of the range.

These barriers coincided with profound changes in environmental characteristics and isolated, on a relatively small scale, porpoises from Iberian waters and on a larger scale porpoises from the Black Sea.

Conclusion: The presence of these barriers to gene flow that coincide with profound changes in oceanographic features, together with the spatial variation in IBD strength, provide for the first time strong evidence that physical processes have a major impact on the demographic and genetic structure of a cetacean. This genetic pattern further suggests habitat-related fragmentation of the porpoise range that is likely to intensify with predicted surface ocean warming.

"seascape" genetics on the North-Atlantic harbour porpoise

Rise of oceanographic barriers in continuous populations of a cetacean: the genetic structure of harbour porpoises in Old World waters

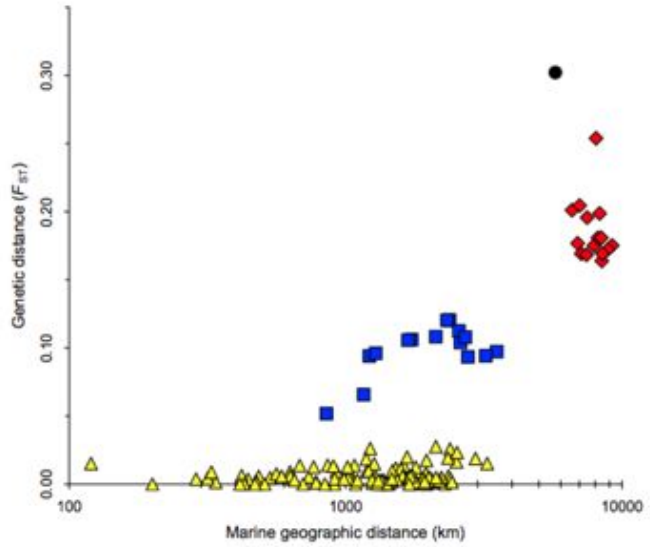


Figure 6
Genetic and geographic distance for pairs of sampled geographic areas. Yellow triangles indicate comparison between pairs of sampled localities within the same cluster; blue squares indicate pairs with one sampled locality in the NAt cluster and the IB cluster; red diamonds indicate pairs with one sampled locality in the NAt cluster and the BS cluster; and black circle indicate the comparison between the IB and the BS cluster.

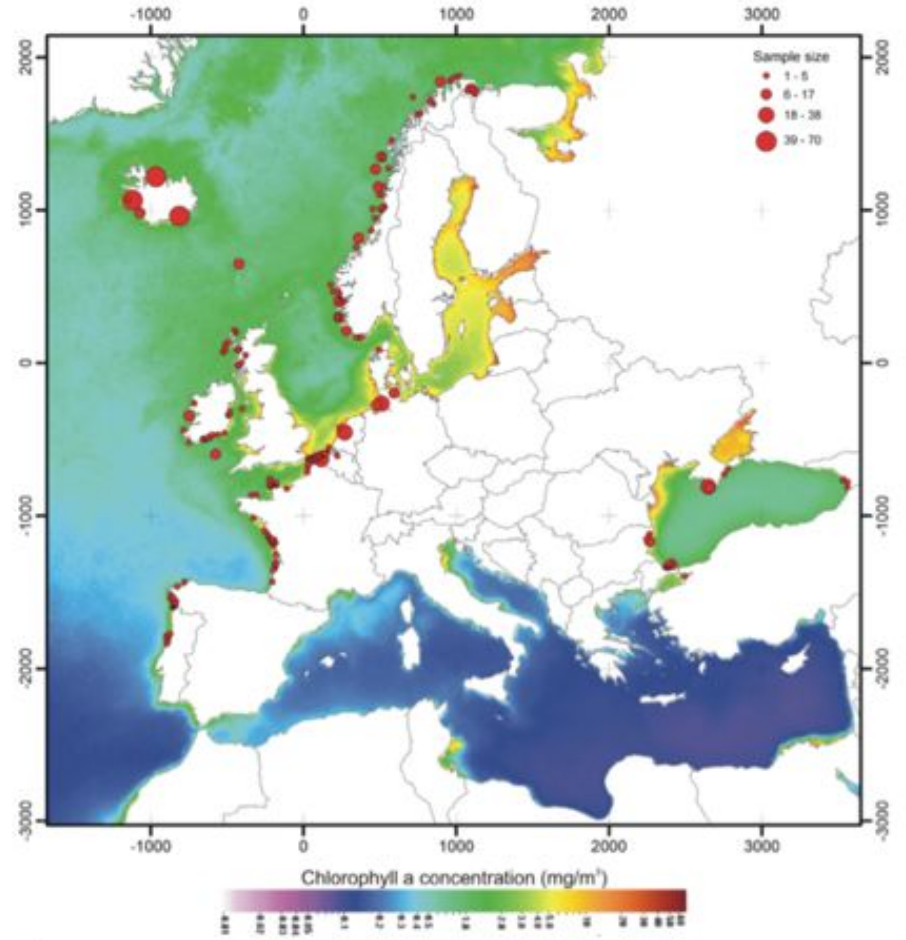
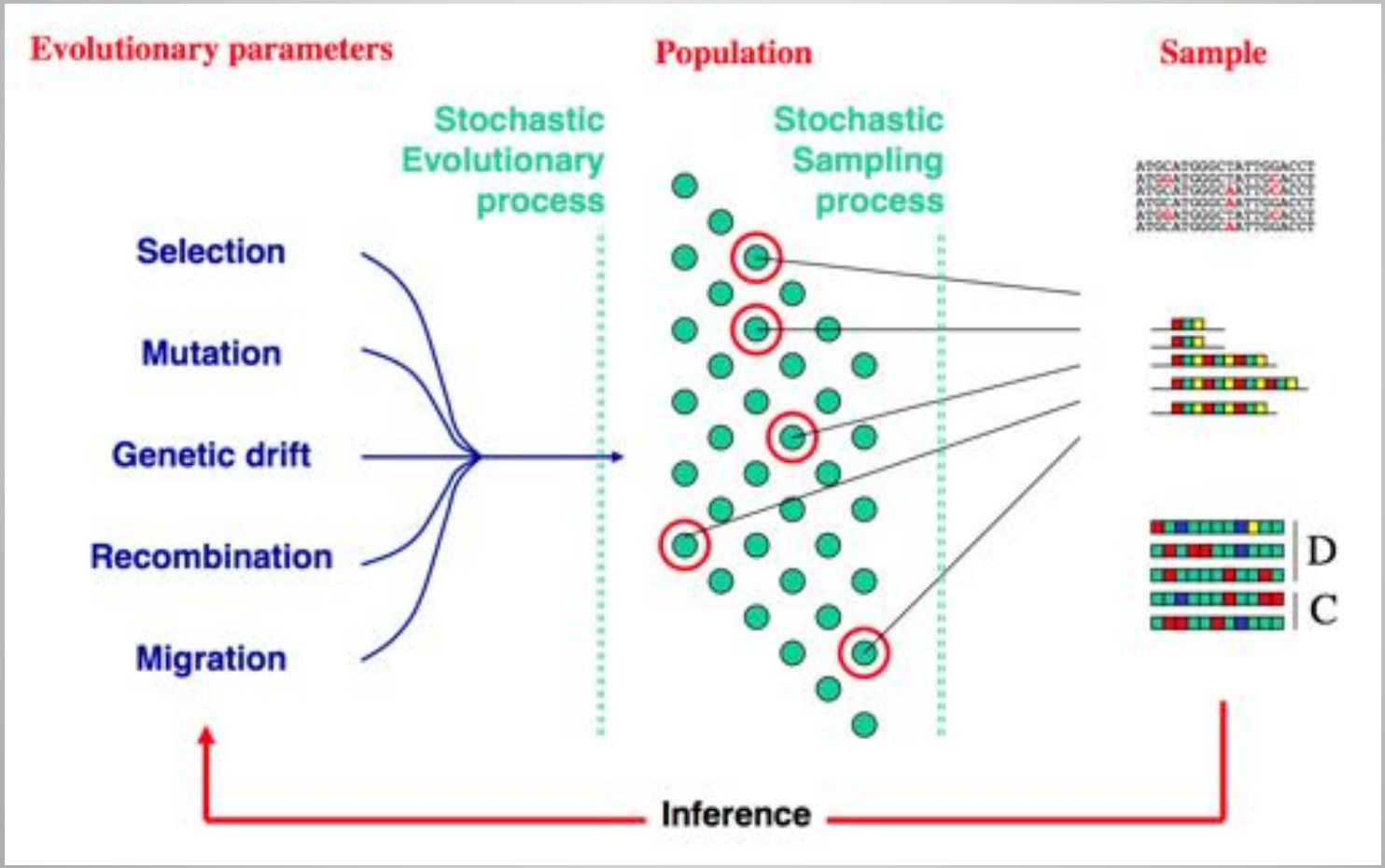


Figure 7
Climatological (1997–2006) annual sea surface chlorophyll concentrations. Data obtained with Sea-viewing Wide Field-of-view Sensor (SeaWIFS, modified from [80]).

Inference in population genetics

Using genetic markers to learn about evolutionary factors acting on natural populations



From McVean Courses : http://www.stats.ox.ac.uk/~mcvean/pgindex02.html

Demographic inference in population genetics

Demographic parameters (DP) are:

population sizes, migration rates, dispersal distances, divergence times, etc …

- General interest in evolutionary biology because DP are important factors for local adaptation of organisms to their environment

- Great interest also in ecology et population management ("Molecular ecology" : conservation biology, study of invasive species,…)

How to do demographic inferences?

- Direct methods, i.e. strictly demographic
  - ✓ tracking individuals: radio, GPS,…
  - ✓ Capture – Mark – Recapture studies (CMR)

  but do not account for temporal variability difficult and needs lots of time

- Indirect methods: neutral polymorphism and population genetics
  - ✓ more and more powerful because of recent advances in molecular biology and population genetic statistical analyses

Are those methods equivalent ?

How to make demographic inferences?

- Direct methods, i.e. strictly demographic

- Indirect methods: neutral polymorphism and population genetics

It is generally considered that :

Direct methods → "present-time and census" parameters

Indirect methods → "past and effective" parameters

How to make demographic inferences?

- Direct methods, i.e. strictly demographic
- Indirect methods: neutral polymorphism and population genetics

Direct methods → "present-time and census" parameters

Indirect methods → ~~"past and~~ effective" parameters

not always true… as we will see under IBD

How to make demographic inferences?

- Direct methods, i.e. strictly demographic

- Indirect methods: neutral polymorphism and population genetics

To make demographic inferences from genetic polymorphism, we need :

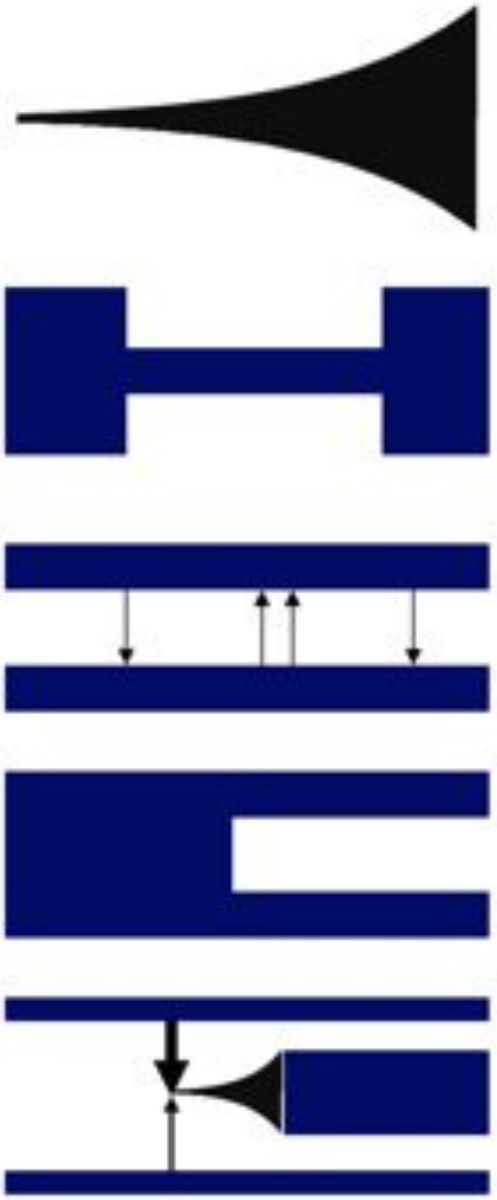
1 - Evolutionary models described by demographic parameters (DP)

2 - Some quantities (*F*-statistics), which can be
(i) expressed as a function of the DP of the model (migration, pop. size, etc.)
(ii) estimated on the genetic data

cf. course "Inference" by R. Vitalis : F_{ST} under the island model.

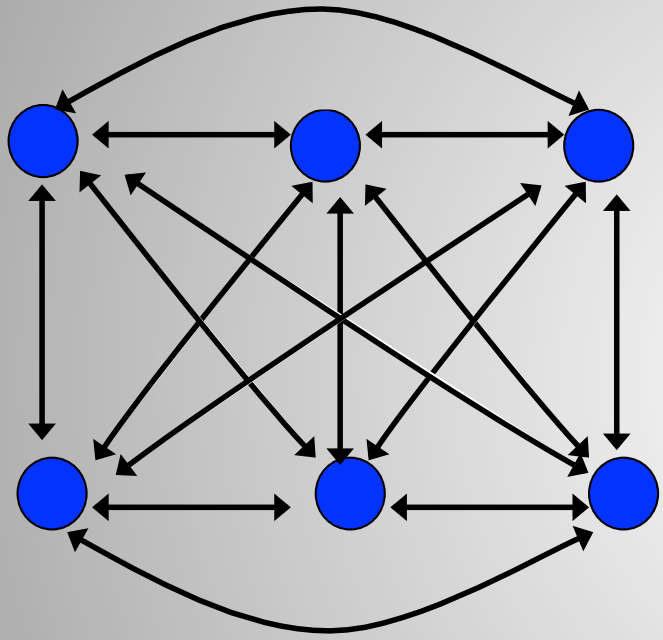
Demographic models

- Population growth
- Population bottlenecks
- Subdivided populations
- Population splits
- Admixture



Models for structured populations:

1 – the island model



Most simple structured model

2 to 3 demographic parameters :

d = sub-population number (or ∞)

N = sub-population size

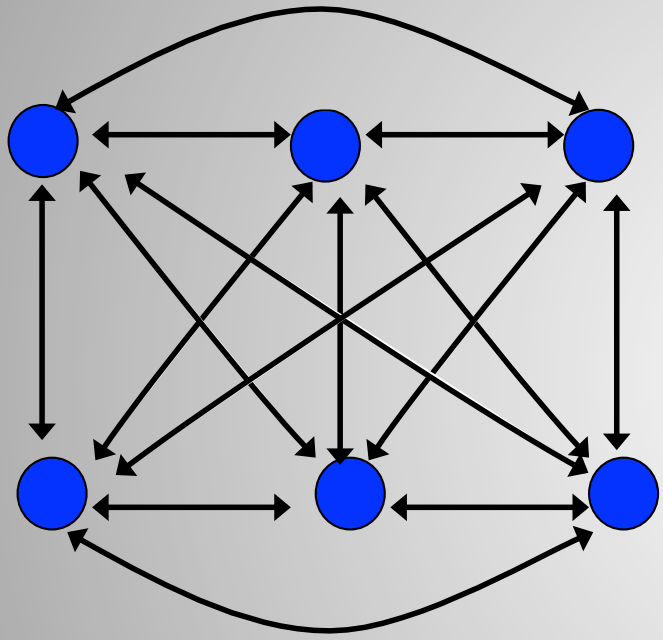
m = migration rate

Fully homogeneous and non-spatial

$$F_{ST} = 1 / (1 + 4Nm)$$

Models for structured populations:

1 – the island model



Most simple structured model

Fully homogeneous and non-spatial

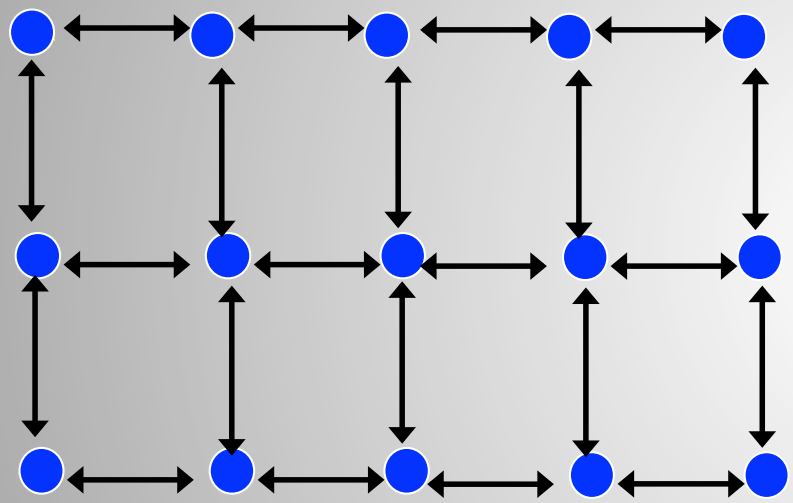
Extremely useful to study theoretical evolutionary effects of migration

but generally not realistic enough to allows precise demographic inferences

In practice $F_{ST} \neq 1 / (1 + 4Nm)$

Models for structured populations:

2 – the stepping stone model



also simple structured model but with

localized dispersal (1D, 2D or 3D)

the same 2 to 3 DP :

d = sub-population number (or ∞)

N = sub-population size

m = migration rate

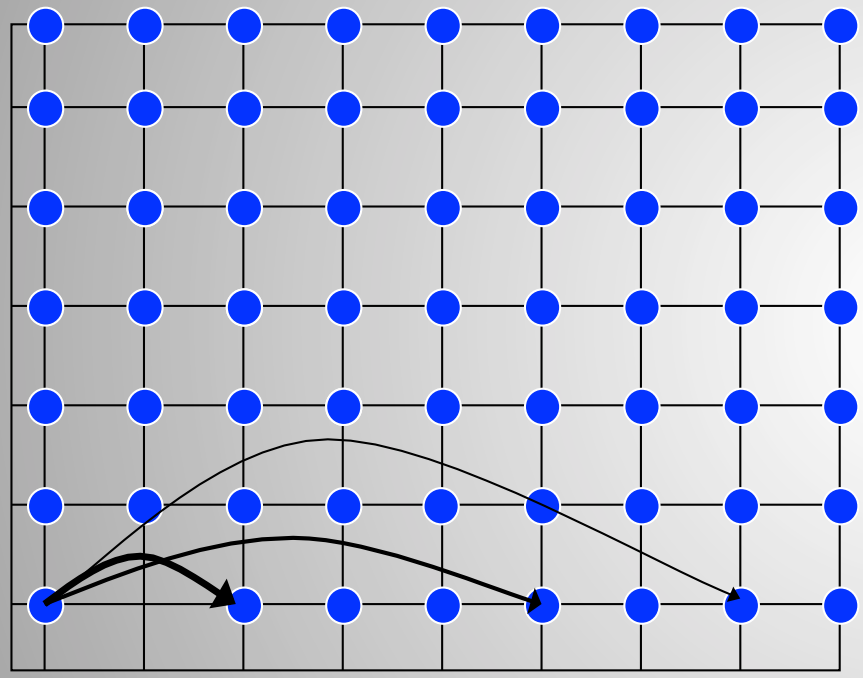
Fully homogeneous and "spatial"

Also extremely useful to study theoretical evolutionary effects of localized dispersal

but generally not realistic enough to allows precise demographic inferences

Models for structured populations:

3 – the general isolation by distance model



Based on the simple property that dispersal is localized in space

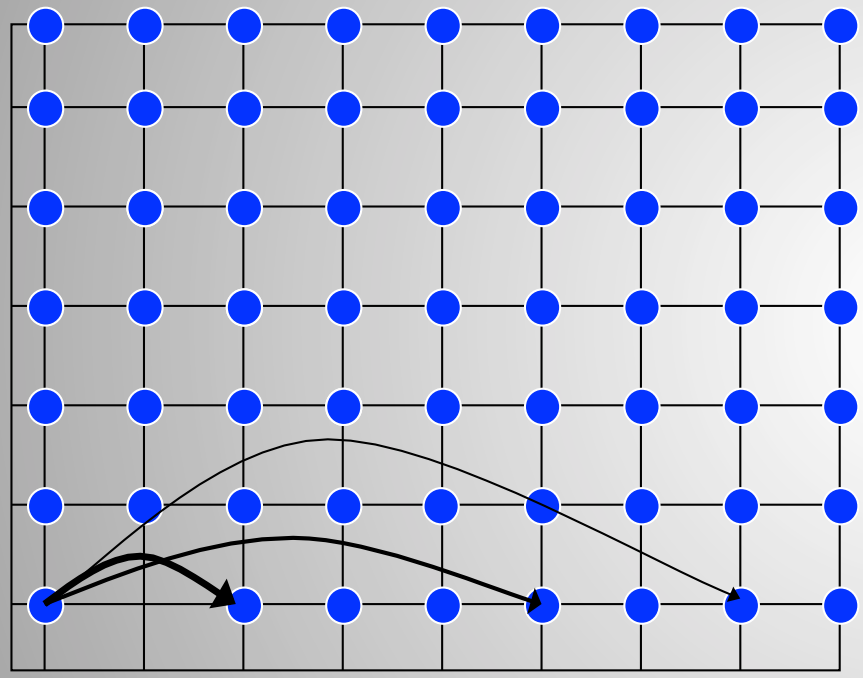
i.e., **2 individuals are more likely to mate if they live geographically close to each other**

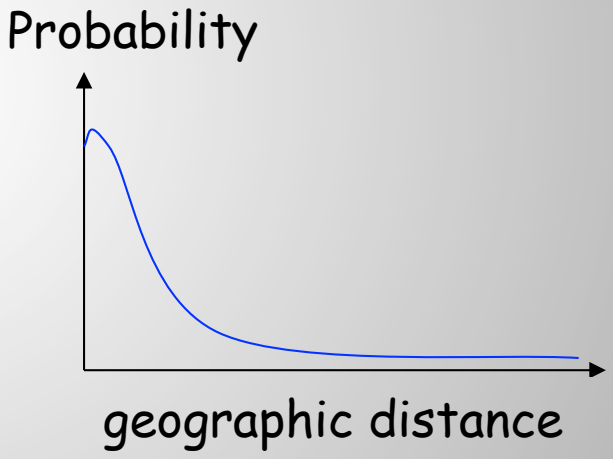
Endler (1977) first showed in a review that

the vast majority of species has geographically localized dispersal

Models for structured populations:

3 – the general isolation by distance model

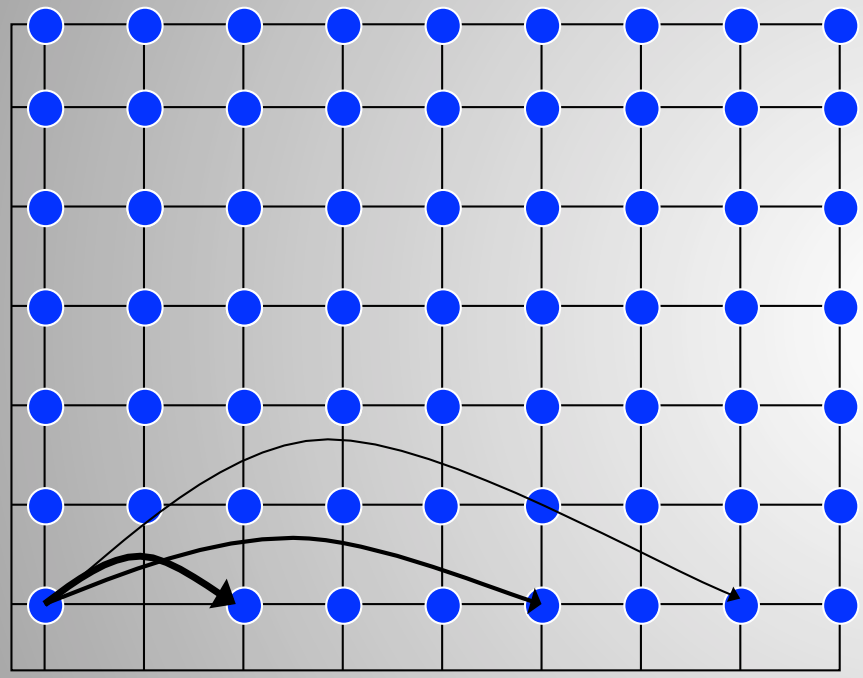


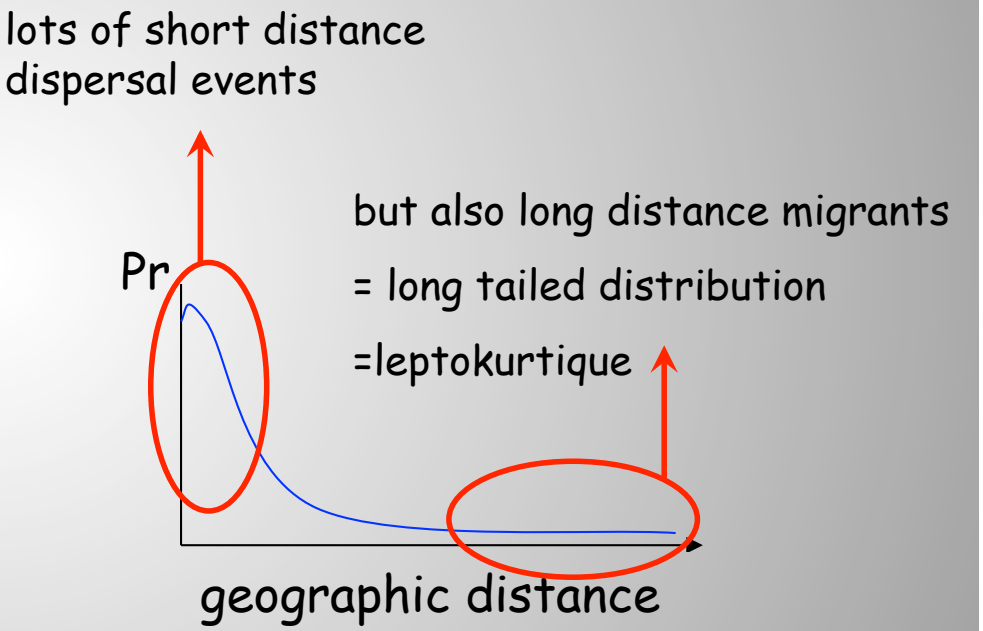


the migration rate between sub-populations is function of the geographic distance through a dispersal distribution

Models for structured populations:

3 – the general isolation by distance model



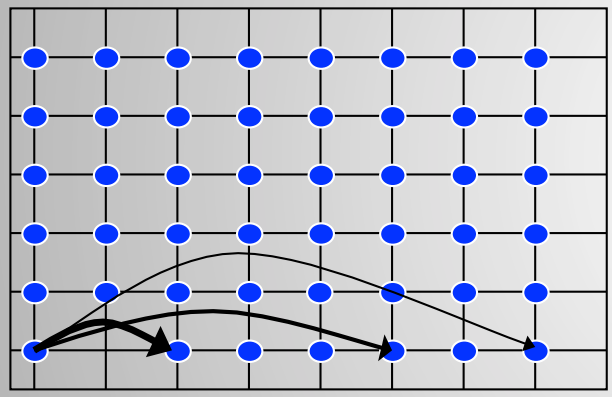


the migration rate between sub-populations is function of the geographic distance through a dispersal distribution

Models for structured populations:

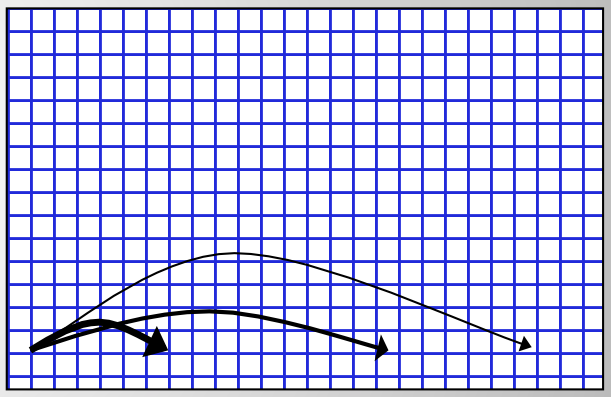
3 – the general isolation by distance model

2 models depending on individual spatial distribution in the landscape



Population with a demic structure

each node of the lattice corresponds to a panmictic sub-population of size N individuals



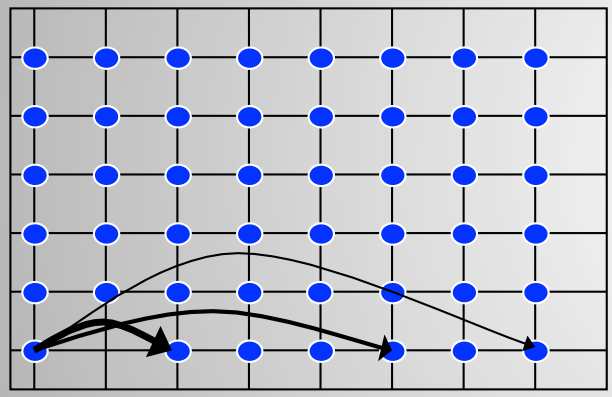
"continuous" population

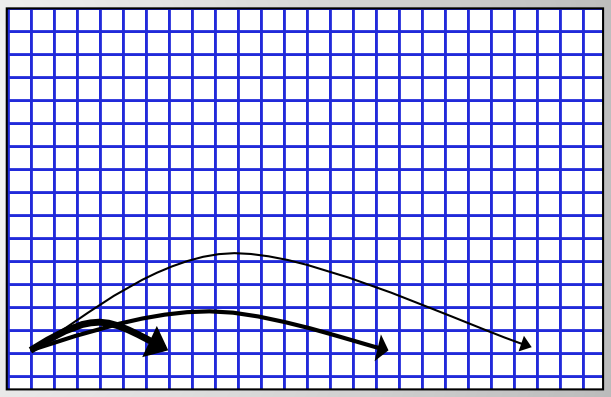
each node of the lattice is a single individual (N=1)

Models for structured populations:

3 – the general isolation by distance model

2 models depending on individual spatial distribution in the landscape





Fully homogeneous model :

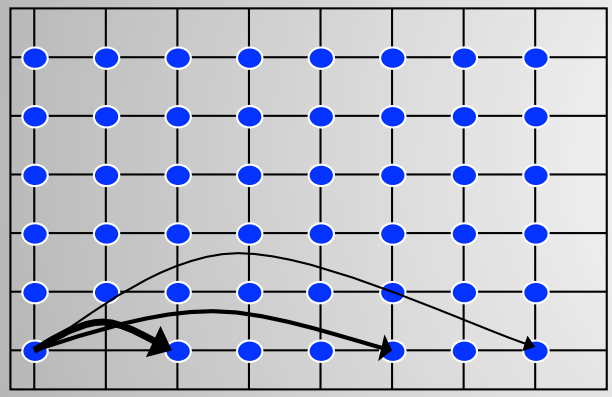
deme size or density of individuals is constant on the lattice

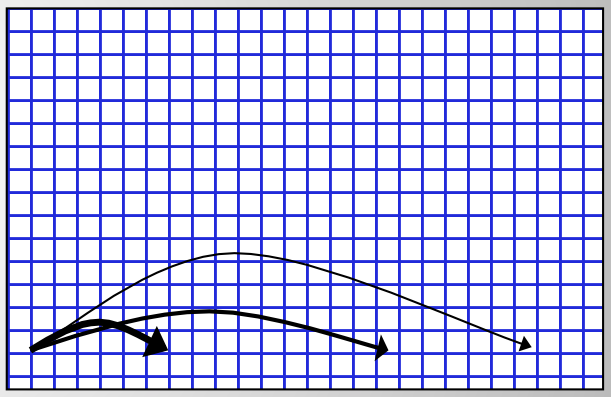
dispersal distribution is the same for all lattice nodes

Models for structured populations:

3 – the general isolation by distance model

2 models depending on individual spatial distribution in the landscape





2 (or more) demographic parameters :

N or D : sub-population size or density of individuals

σ^2 : mean squared parent-offspring dispersal distance

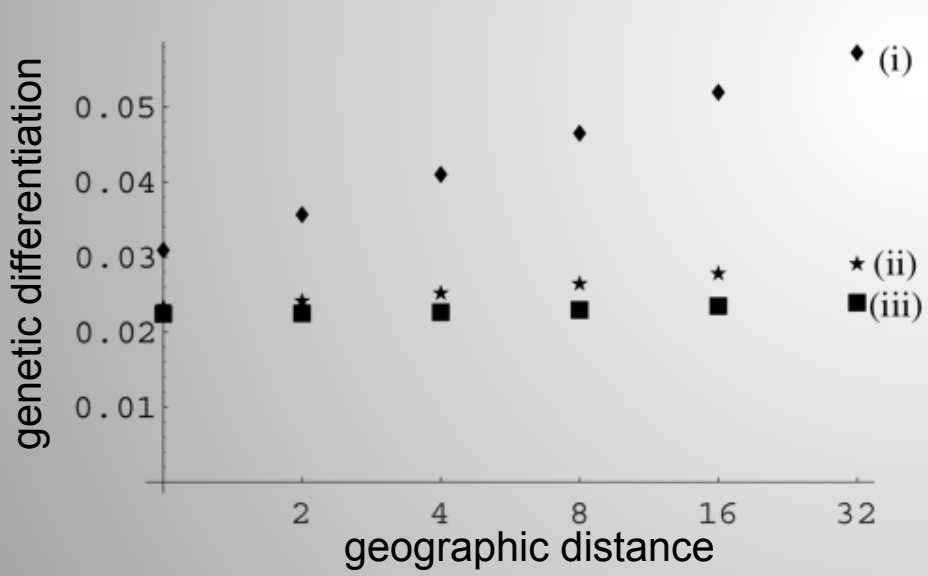
$D\sigma^2 \approx$ inverse of the "strength of IBD"

Models for structured populations:

3 – the general isolation by distance model

The main characteristic of IBD models is that

genetic differentiation increases with geographic distance



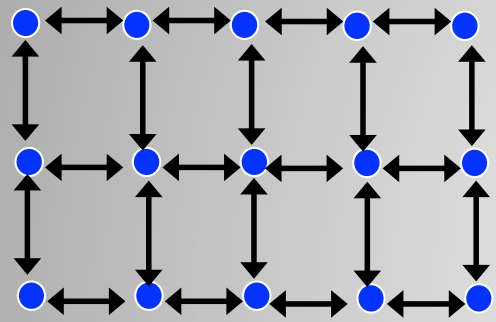
Strong IBD (**small $D\sigma^2$**)

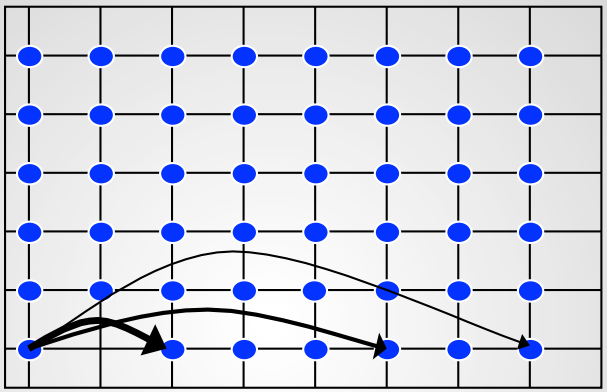
weak IBD (**large $D\sigma^2$**)

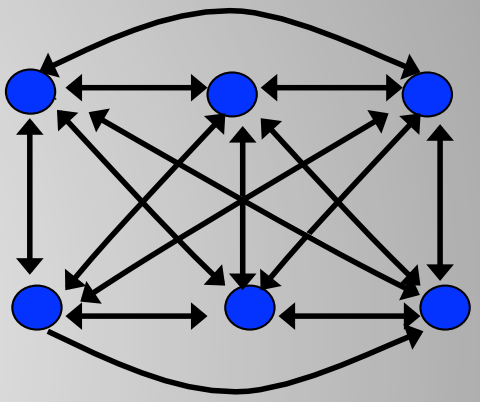
Island model, no IBD (**$D\sigma^2 = \infty$**)

Models for structured populations:

3 – the general isolation by distance model







IBD models are quite general depending on how localized dispersal is :

Stepping stone **>** **IBD** **>** **Island Model**

$\sigma^2 = m < 1$ $1 < \sigma^2 << \infty$ $\sigma^2 \approx \infty$

Dispersal inference under isolation by distance:

1 – the differentiation parameter : $F_{ST}/(1-F_{ST})$

The mathematical analysis is done in terms of probability of identity (cf Vitalis) and then expressed as relationship between F-statistics and DP

For the demic model :

Q_1 is the probability of identity of two genes taken within a deme,
Q_2, Q_r are prob. of identity of two genes taken in different demes (or at distance r),

$$\frac{Q_1 - Q_r}{1 - Q_1} = \frac{F_{ST}}{1 - F_{ST}} \text{ computed between demes at geographical distance } r$$

with $\frac{Q_1 - Q_2}{1 - Q_2} = F_{ST}$ and $Q_2 \Leftrightarrow Q_r$ to take distance into account

Dispersal inference under isolation by distance:

1 – the differentiation parameter : $F_{ST}/(1-F_{ST})$, a_r

The mathematical analysis is done in terms of probability of identity (cf Vitalis) and then expressed as relationship between F-statistics and DP

For the "continuous" model :

$$a_r \equiv \frac{Q_1 - Q_r}{1 - Q_1} \text{ computed between individuals at geographical distance } r$$

with Q_1 the probability of identity of two genes taken within an individual

and Q_r the prob. of id. of two genes taken in two individuals separated by a distance *r*

$$a_r \equiv \frac{Q_1 - Q_r}{1 - Q_1} \text{ is analoguous to } \frac{F_{ST}}{1 - F_{ST}} \text{ between individuals}$$

Dispersal inference under isolation by distance:

2 – relationship between differentiation and distance

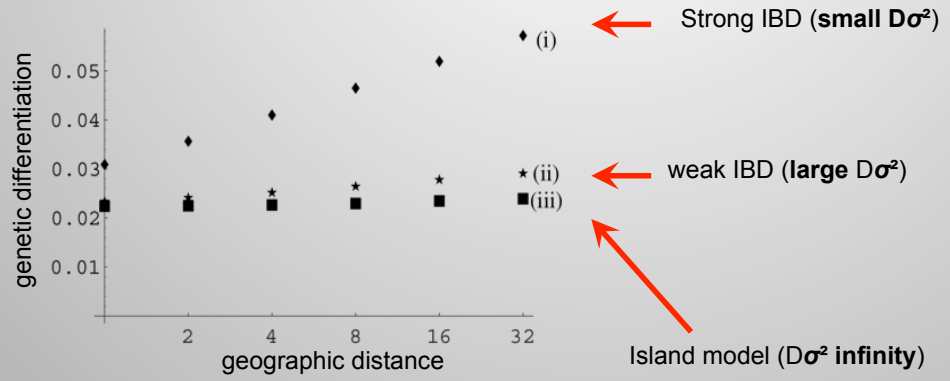
RECALL : 2 main demographic parameters :

N or ***D*** : sub-population size or density of individuals

σ^2 : mean squared parent-offspring dispersal distance

: inverse of the "strength of IBD"

+ ***μ*** the mutation rate (per locus per generation)



Dispersal inference under isolation by distance:

2 – relationship between differentiation and distance

The main result of the analysis of IBD models in terms of probabilities of identity is the following relationship between the differentiation parameter and the geographic distance and the different assumptions leading to it :

in one dimension IBD models with demes :

$$a_r \text{ or } \frac{F_{ST}}{1-F_{ST}} = \frac{Q_1 - Q_r}{1-Q_1} \approx \frac{1-e^{\frac{-\sqrt{2\mu r}}{\sigma}}}{4N\sigma\sqrt{2\mu}} + \text{constant}$$

$$a_r \text{ or } \frac{F_{ST}}{1-F_{ST}} \overset{\text{r et } \mu \text{ petit}}{\approx} \frac{r}{4N\sigma^2} + \text{constant}$$

Simple linear relationship between differentiation and distance but only for small distances and low mutation rates

Dispersal inference under isolation by distance:

2 – relationship between differentiation and distance

The main result of the analysis of IBD models in terms of probabilities of identity is the following relationship between the differentiation parameter and the geographic distance and the different assumptions leading to it :

in two dimension IBD models :

$$\frac{Q_1 - Q_r}{1 - Q_1} \underset{\text{r et } \mu \text{ petit}}{\approx} \frac{\ln(r)}{4\pi N\sigma^2} + \text{constant}$$

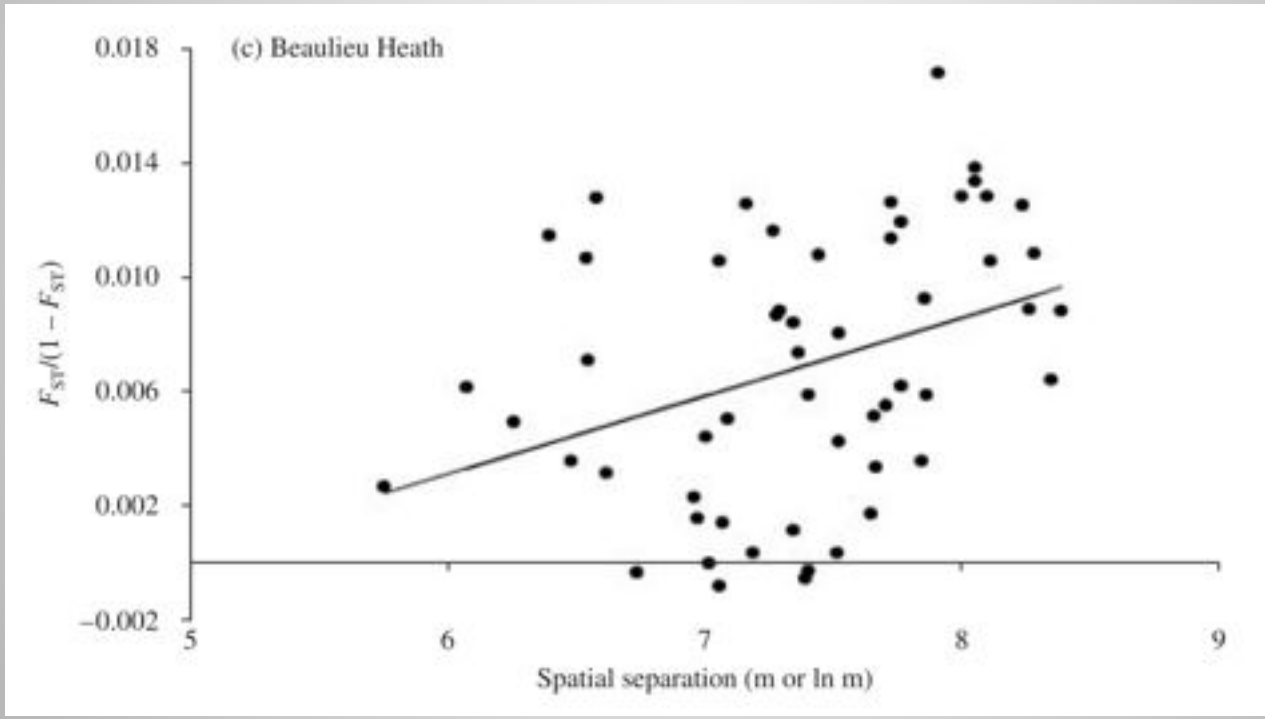
$$\underset{\text{N} \to \text{D}}{\approx} \frac{\ln(r)}{4\pi D\sigma^2} + \text{constant}$$

Simple linear relationship between differentiation and the logarithm of the distance
but only for small distances and low mutation rates

Dispersal inference under isolation by distance:

3 – the regression method of Rousset (1997, 2000)

The regression slope is expected to be $4\pi D\sigma^2$, thus a simple method to infer $D\sigma^2$ is to do the regression on the data and estimate the slope



➔ **1/slope is an estimator of $D\sigma^2$**

Dispersal inference under isolation by distance: 3 – the regression method of Rousset (1997, 2000)

The regression slope is expected to be $4\pi D\sigma^2$, thus a simple method to infer $D\sigma^2$ is to do the regression on the data and estimate the slope

In practice :

1 – go to field and sample 80-500 individuals on a given surface

2 – genotype them using a dozen or more of microsatellite markers

3 – Use Genepop : option IBD between individuals or demes

- it estimates $F_{ST}/(1-F_{ST})$ or a_r for all pairs of demes or individuals
- it regresses them against the geographic distance or its logarithm
- it infer the slope of the regression

Inference of $D\sigma^2$ under isolation by distance:

3 – the regression method of Rousset (1997, 2000)

- Point estimate : 1/slope ➔ estimate of $4\pi D\sigma^2$

- Significance :
  - ✓ **Mantel Test** (by permutations) :

    Test the correlation between the genetic and the geographic matrices

    by permuting rows and columns from one of the two matrices

    -> significant if the initial correlation is greater than

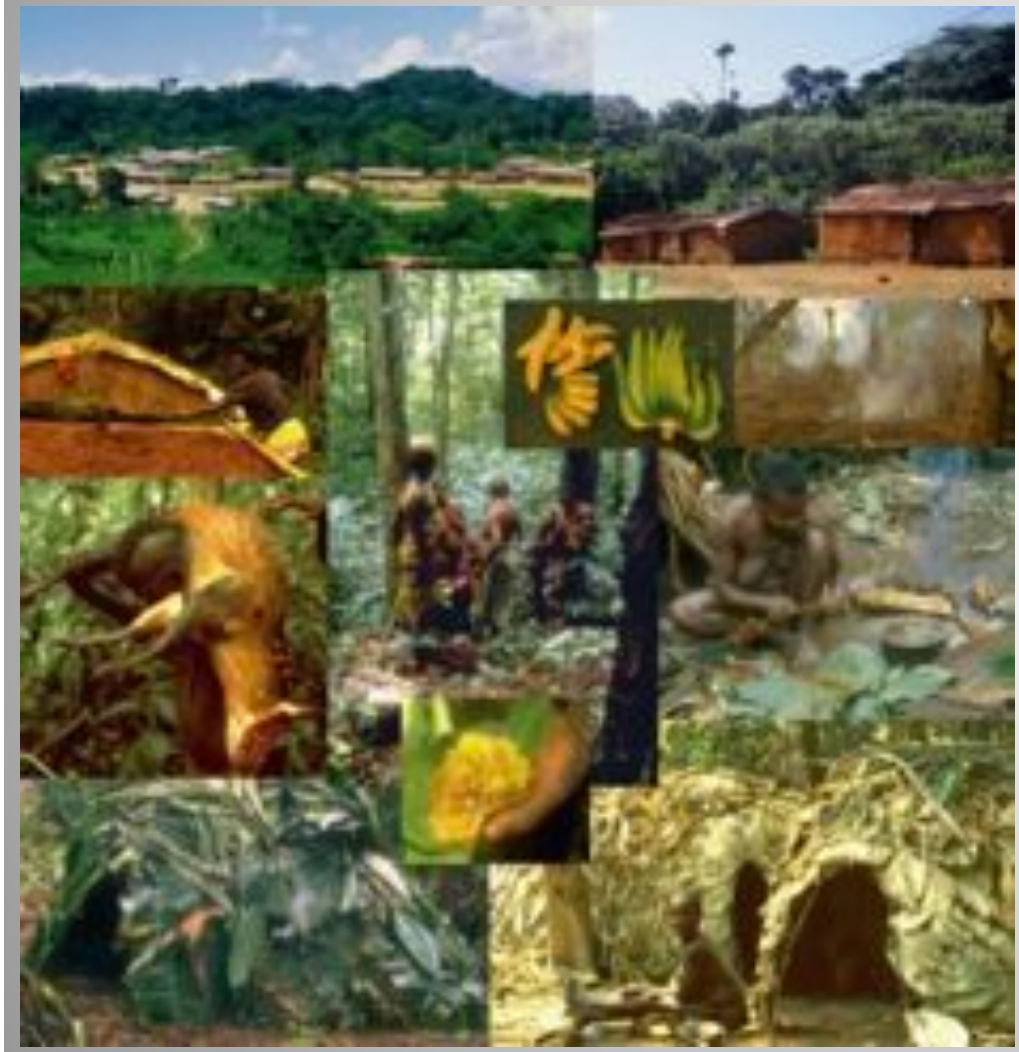
    the correlation on permuted matrices (e.g. in the higher 5%)
  - ✓ **Bootstrap** : re-sampling of loci (ok because they are independent)

    gives Confidence Intervals (CI) for the slope

    -> significant if the CI does not contain 0 (null slope, infinite $D\sigma^2$)

Inference of $D\sigma^2$ under isolation by distance:

4 – example on a Pygmy population



Paul Verdu PhD

National Museum of Natural History,

Paris :

History of the pygmy populations

from Western Africa

Inference of $D\sigma^2$ under isolation by distance:

4 – example on a Pygmy population

biology letters

Biol. Lett.
doi:10.1098/rsbl.2010.0192
Published online

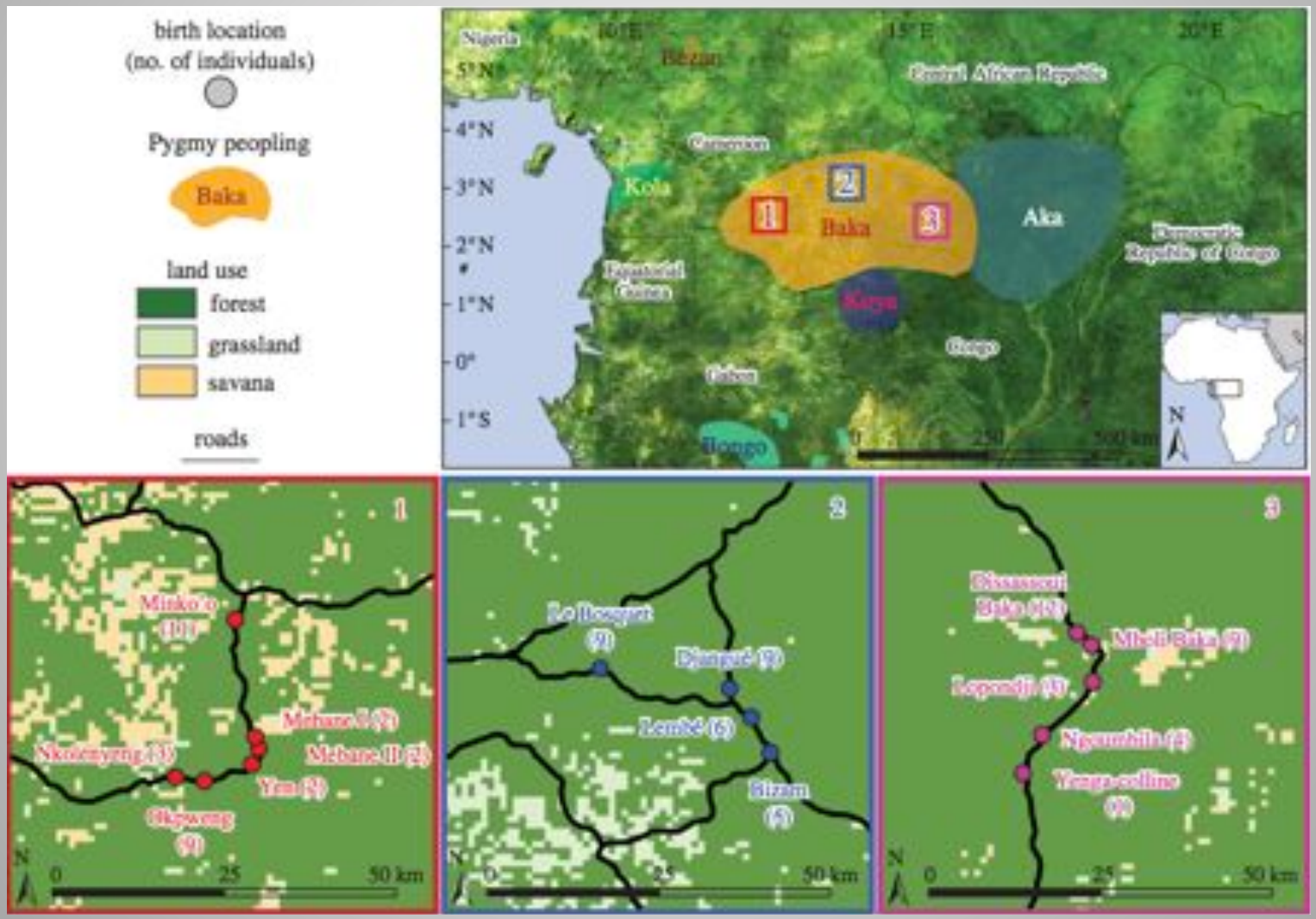
Population genetics

Limited dispersal in mobile hunter–gatherer Baka Pygmies

Paul Verdu[1,2,*], Raphaël Leblois[3], Alain Froment[4], Sylvain Théry[2], Serge Bahuchet[2], François Rousset[5], Evelyne Heyer[2] and Renaud Vitalis[2,†]

Inference of $D\sigma^2$ under isolation by distance:

4 – example on a Pygmy population



Inference of $D\sigma^2$ under isolation by distance:

4 – example on a Baka Pygmy population

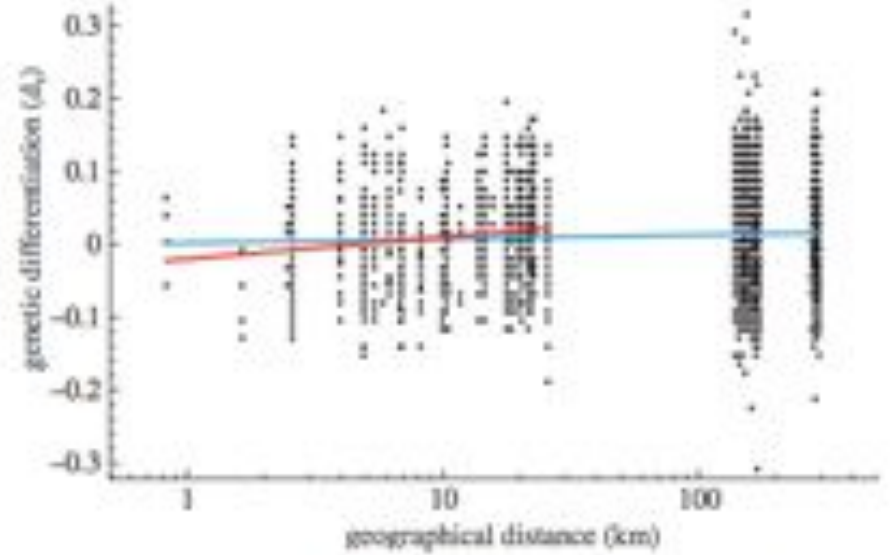


Figure 2. Correlation between genetic differentiation and the logarithm of geographical distances among Baka Pygmies. Multilocus estimates of pairwise differentiation ($\hat{d}_r$) are plotted against the logarithm of geographical distances (in kilometres). The linear regression considering all pairs of individuals is $y = 0.0027x - 0.0153$ (in blue). The linear regression considering only pairs of individuals born within the same group is $y = 0.0137x - 0.1138$ (in red).

Total sample : $\mathbf{4\pi D\sigma^2 = 373}$

within group (small scale) : $\mathbf{4\pi D\sigma^2 = 73}$

using D=0.47 ind/km^2

we have $12.4 < \sigma^2 < 63.2$ km^2

Cavalli-Sforza & Hewlett (1982) found
$\sigma^2 \approx 3683$ km^2
from a ethnological survey
in Aka pygmies !

Inference of $D\sigma^2$ under isolation by distance:

4 – example on a Pygmy population

indirect genetic estimate (regression method) : **$12.4 < \sigma^2 < 63.2$ km^2**

indirect ethnologic estimate **$\sigma^2 \approx 3683$ km^2**

Those discrepancies can be explained by:

- demographic/ethnologic data (distances between birthplaces and places of residence) may reflects exploration behavior rather than parent-offspring dispersal
- the two studies done in different pygmy groups (Aka vs Baka) which may have different dispersal behavior

Conclusions :

Although our results do not challenge the view that hunter–gatherer Pygmies have frequent movements in their socio- economic area, we demonstrate that extended individual mobility does not necessarily reflect extended dispersal across generations

Testing inference methods

1 – How to test an inference method ?

- Tests by simulations:

  = how close are estimates / values specified in simulations

  - simulations under the right model (i.e. the one used for inference)

    ➠ gives the precision of the inference in the best cases

  - simulations under a model that does not respect some assumptions

    ➠ gives the robustness / model assumptions

- Tests on real data sets for which we have "independent expectations"

  = For demographic parameter inference from genetic data, the only solution is to compare our indirect estimates with direct estimates obtain with demographic methods (CMR, tracking, …)

Testing inference methods

2 – Simulation test of the regression method

(1) Choice of mutational and demographic parameter values for simulations

↓

(2) Simulation : 1000 runs for 10 loci

↓

(3) Analysis of the 1000 simulated multilocus data sets

→1000 estimates of the regression slope

↓

(4) Comparison with the "expected" value of the slope :

Relative bias = $\sum$(Est-Exp)/Exp

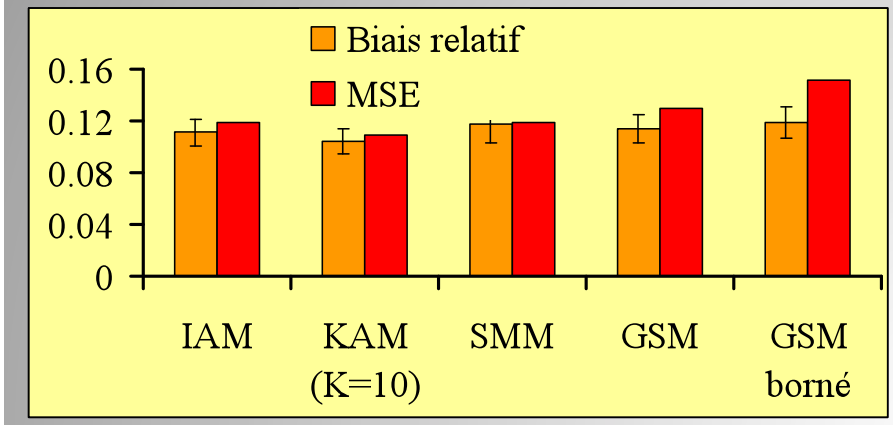
Mean squarre error MSE = $\sum(\text{Est-Exp})^2/\text{Exp}^2$

Proportion of estimates within a factor 2 from the expected value

i.e. in $[D\sigma^2_{exp} / 2 ; 2 \times D\sigma^2_{exp}]$

Testing inference methods

2 – Simulation test of the regression method

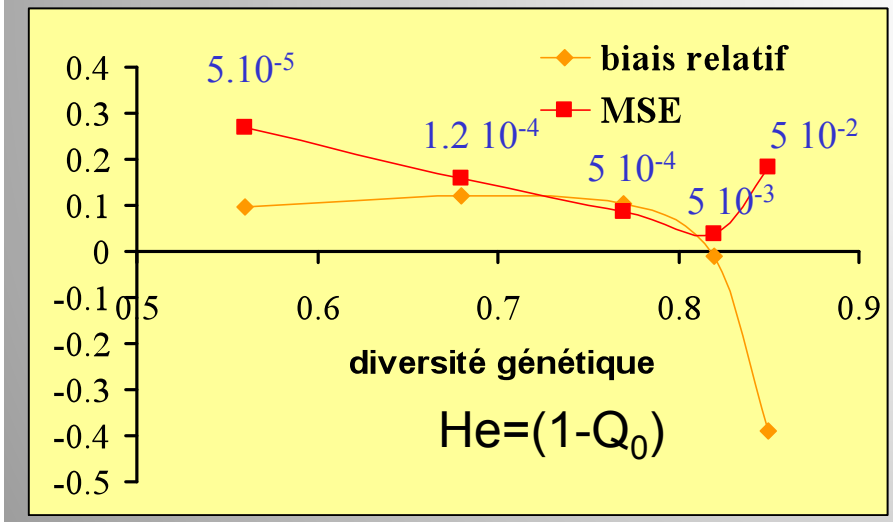


Influence of mutational processes

Method based on Identity by Descent (IBD)

Marker information is not by descent but by state: e.g. Stepwise mutations for microsats

Simulation results ➠ very robust method : small effects of different mutational models



Influence of mutation rate (genetic diversity)

Assumption: low ***μ*** ; but diversity is needed to have enough "genetic information"

Simulation results:

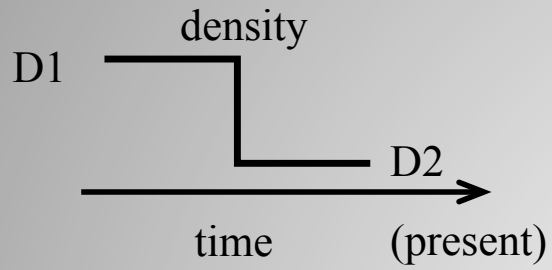
➠ better precision with high diversity (0.7-0.8)

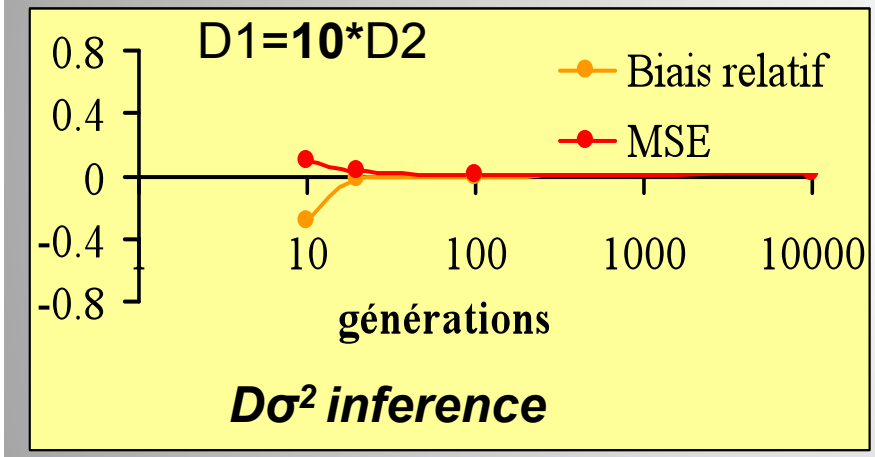
➠ strong bias for very high mutation rates

Microsatellites are good markers despite their complex mutational processes because they show high genetic diversity

Testing inference methods

2 – Simulation test of the regression method





Influence of past demographic processes:

Ex 1 : past decrease in density (bottleneck)

Simulations results ➠ robust method because the influence of past density is very weak

Other tests:

- past density increase
- spatial expansion
- spatial heterogeneity in density

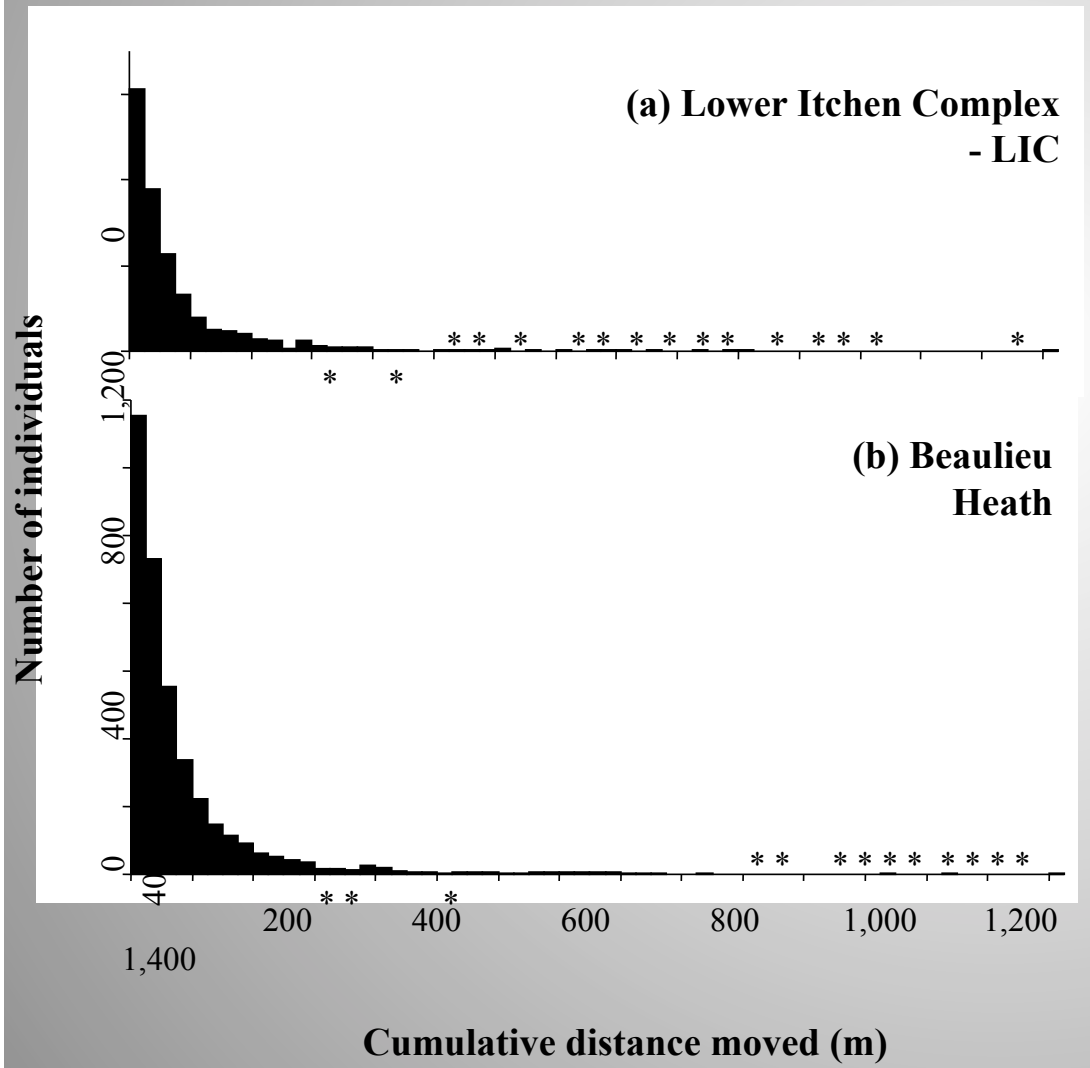
All simulation tests ➠ Global robustness of the regression method to temporal and spatial heterogeneities of demographic parameters :

➠ **the regression method infer the present-time and local *Dσ²* of the population sampled**

Testing inference methods

3 – Comparisons between genetic and demographic estimates

- example on damselfly populations (Watt et al. 2007 Mol.Ecol.)



Demographic data (CMR)

➠ Census density and distribution of dispersal

Testing inference methods

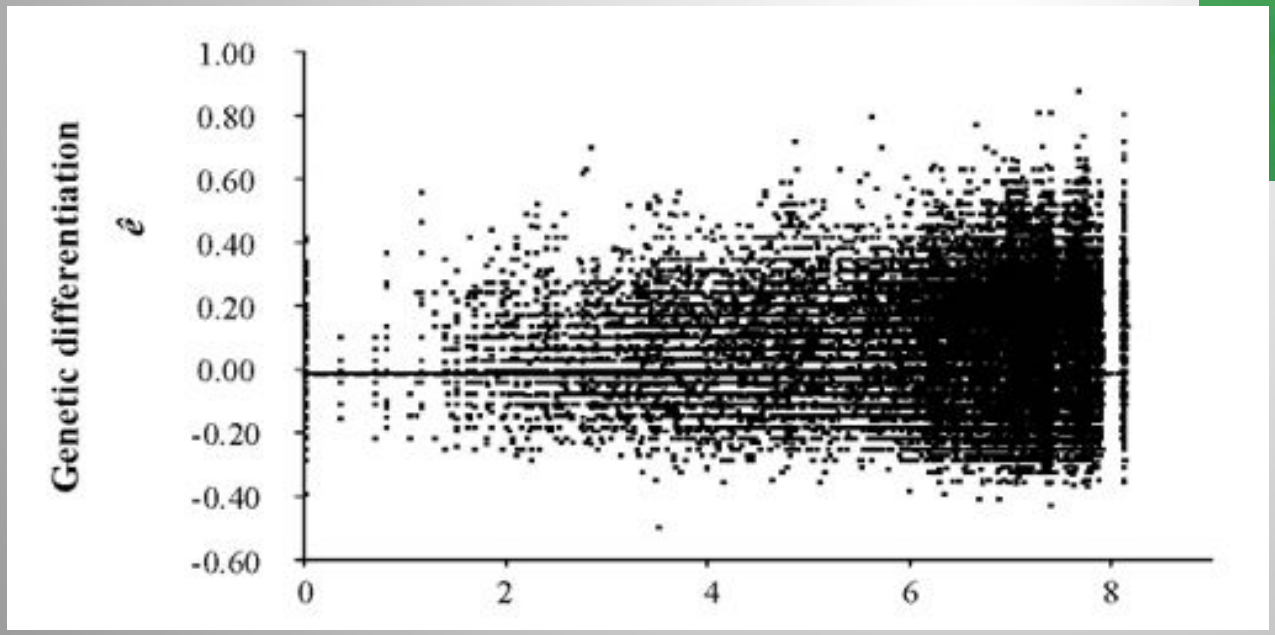
3 – Comparisons between genetic and demographic estimates

- example on damselfly populations (Watt et al. 2007 Mol.Ecol.)

Genetic data : 700 individuals genotyped

at 13 microsatellite loci

➠ indirect estimates of $D\sigma^2$



Testing inference methods

3 – Comparisons between genetic and demographic estimates

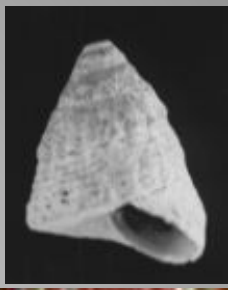
- example on damselfly populations (Watt et al. 2007 Mol.Ecol.)

	$D\sigma^2$ estimates	
	Direct (demographic)	Indirect (genetic)
Site 1	277	222
Site 2	249	259
Site 3	555	753

very good agreement between demographic and genetic estimates

Testing inference methods

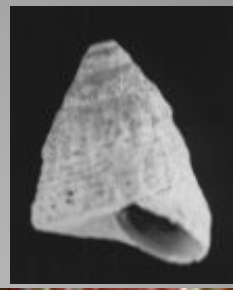
3 – Comparisons between genetic and demographic estimates



	Direct (Demography)	Indirect (genetic)
American Marten (Martes americana)	**7.5**	**3.8**
Kangaroo rats (Dipodomys)	**1.43**	**2.58**
intertidal snails (Bembicium vittatum)	**2.4**	**3.6**
Forest lizards (Gnypetoscincus queenslandiae)	**11.5**	**5.5**
Humans in the rainforest (Papous)	**29.3**	**21.1**
Legumin (Chamaecrista fasciculata)	**9.6**	**13.9**

Testing inference methods

3 – Comparisons between genetic and demographic estimates



	Direct (Demography)	Indirect (genetic)
American Marten	7.5	3.8
Kangaroo rats	1.43	2.58
intertidal snails	2.4	3.6
Forest lizards	11.5	5.5
Humans in the rainforest	29.3	21.1
Legumin	9.6	13.9

very good agreement between demographic and genetic estimates for all available data sets with demographic and genetic data at a local geographical scale

➠ **validate the regression method and isolation by distance models**

Usual (and often justified) critics on indirect demographic inferences

Main critics on demographic parameter inference from genetic data (Hasting et Harrison 1994, Koenig et al. 1996, Slatkin 1994) :

- Demo-genetic models are not realistic enough, especially dispersal modeling in the island model
- Natural population are often inhomogeneous and at disequilibrium, whereas most demo-genetic models assume spatial homogeneity and time equilibrium
- Assumptions on mutation rates and mutational models are oversimplified regarding complex mutational processes of genetic markers
- neutral markers do not really exist, there is always a form of selection

➠ Whitlock & McCauley (1999, Heredity) :

Indirect measure of gene flow and migration : Fst ≠1/(1+4Nm)

Usual (and often justified) critics on indirect demographic inferences

Main critics on demographic parameter inference from genetic data (Hasting et Harrison 1994, Koenig et al. 1996, Slatkin 1994) :

- no realistic models of dispersal
- too many assumptions on spatial homogeneity and time equilibrium
- oversimplified mutational models
- genetic markers are not neutral

➠ Whitlock & McCauley (1999, Heredity) :

Indirect measure of gene flow and migration : Fst ≠1/(1+4Nm)

So why do we have good results for $D\sigma^2$ inferences using the regression method on IBD models ?

Why $D\sigma^2$ inferences using the regression method on IBD models seems to work so well ?

- **The model : Isolation by Distance is a "relatively realistic" model**
  - Dispersal is well modeled (allows localized but also leptokurtic dispersal)
  - "Continuous" IBD models allows the consideration of continuous spatial distribution of individuals ➠ no need to a priori define sub-populations/demes
- **The inference method : the regression methods of Rousset (1997, 2000) is well designed, precise and robust**
  - the relationship between $F_{ST}/(1-F_{ST})$ and the distance is easier to interpret in terms of demographic parameters than Fstatistics alone (simple linear relationship)
  - No assumptions on the shape of the dispersal (allows leptokurtic distributions)
  - only valid for sampling at a local geographical scale (small distance assumption)
    ➠ less demographic and selective spatial heterogeneities
- **The genetic markers : microsatellites are good highly informative markers**

Why $D\sigma^2$ inferences using the regression method on IBD models seems to work so well ?

- **The model : Isolation by Distance is a "relatively realistic" model**
- **The inference method : the regression methods of Rousset (1997, 2000) is well designed, precise and robust**
- **The genetic markers : microsatellites are good highly informative markers**

➠ Both the demo-genetic model, the inference method, the sampling strategy and the genetic markers are important for the inference of demographic parameters to be accurate, i.e. to obtain precise and robust estimation of local and present-time demographic parameters

Why $D\sigma^2$ inferences using the regression method on IBD models seems to work so well ?

Quick interpretation of the robustness of the regression method to mutational processes and past demographic changes using the coalescent theory :

- small deme/sub-population sizes
- high migration rates
- sampling at small geographical scale

} short coalescence times

➠ **short coalescence times (i.e. most of the coalescent tree is in a recent past) decrease the influence of past factors acting on the distribution of polymorphism, such as past mutation processes et past demographic fluctuations**

Note that this effect is even more pronounced for the "continuous" IBD model because deme size is one individual and migration rates are very high (>0.3)

Extensions to classic isolation by distance models

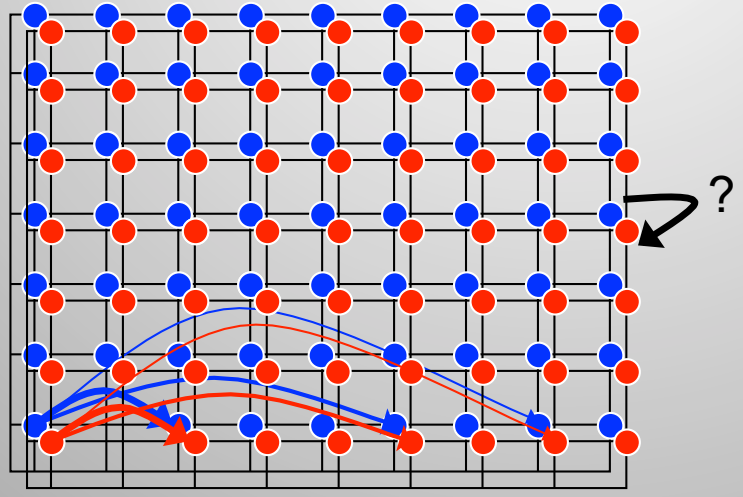
1 – IBD within and bewteen two habitats or groups

Using IBD models to test for potential gene flow between populations of organisms living in different habitats in sympatry (Rousset 1999)

Different habitats can be, for example :

- different hosts for a parasite
- agricultural vs natural populations

IBD within each habitat, but what could the signal of the differentiation between the habitats tell us about gene flow between those habitats



Extensions to classic isolation by distance models

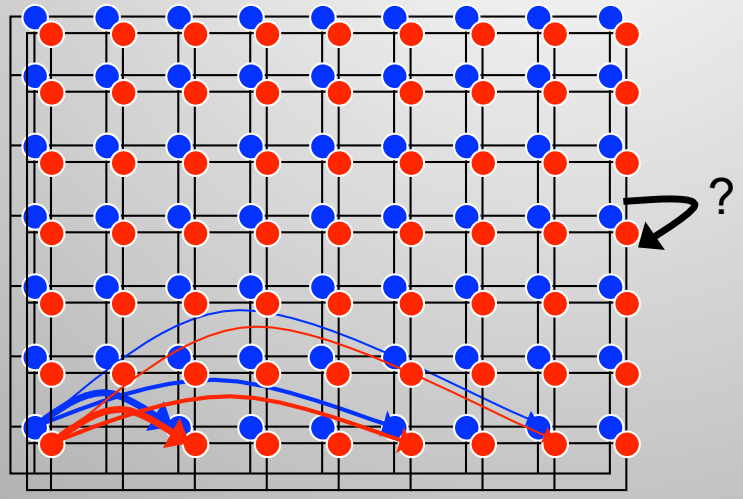
1 – IBD within and bewteen two habitats or groups

Using IBD models to test for potential gene flow between populations of organisms living in different habitats in sympatry (Rousset 1999)

Assumption : IBD in at least one of the habitats

The theory showed that if there is enough gene flow between the two habitats (m>0.001) then IBD should be observed between habitats, with a "intermediate" IBD pattern compared to IBD patterns within each habitat

if there is no gene flow between the two habitats (m<0.001) then the differentiation between habitats should be independent of the distance



Extensions to classic isolation by distance models

1 – IBD within and between two habitats or groups

Ex: European Corn Borer (Ostrinia Nubilalis), a major pest for corn plantations

Native in Europe, introduced in North America

Extensions to classic isolation by distance models

1 – IBD within and between two habitats or groups

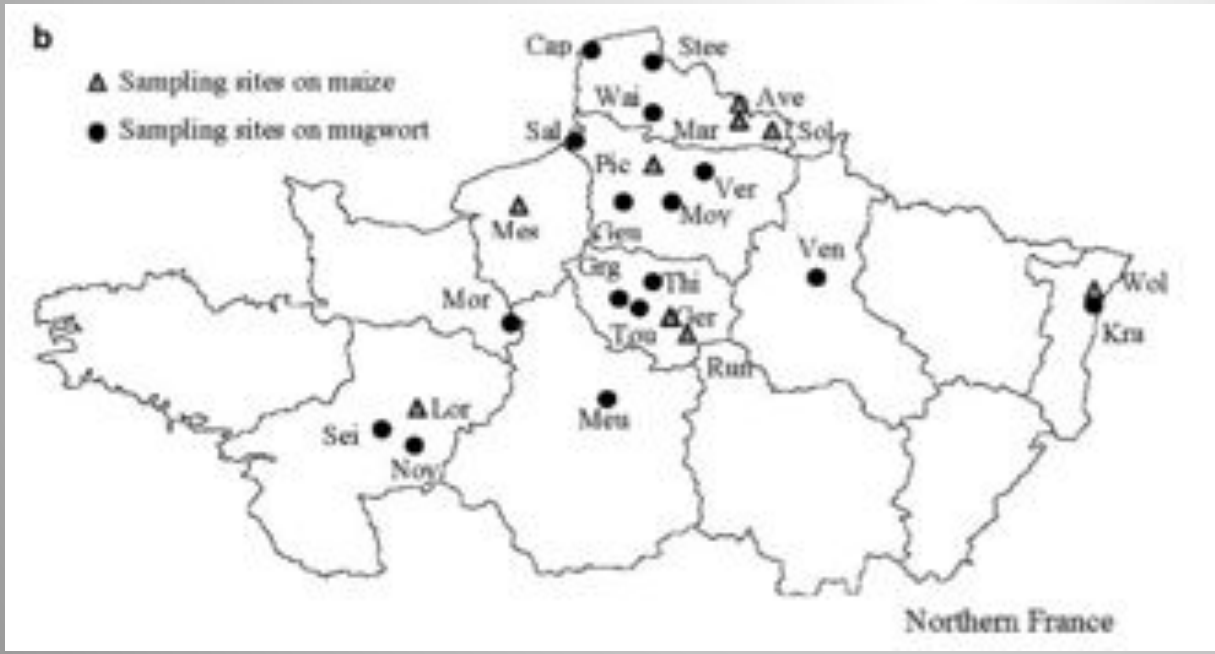
The European Corn Borer (Ostrinia Nubilalis)

naturaly feeds on mugwort (Asteraceae) in Europe

Extensions to classic isolation by distance models

1 – IBD within and between two habitats or groups

- GMO "Bt" maize plants are resistant to the European Corn Borer, but to manage the evolution of resistance to the B. thuringiensis toxins in the pest, there is a need to keep "refuge habitats" near the GMO plantations
- Refugia can theoretically be plant on which the insect can feed and reproduce, however, to be efficient, there should be enough gene exchanges between pest populations living on plantations and refuges



Martel et al (2003, Heredity) tested the usefulness of using mugwort natural populations as refuges

Extensions to classic isolation by distance models

1 – IBD within and between two habitats or groups

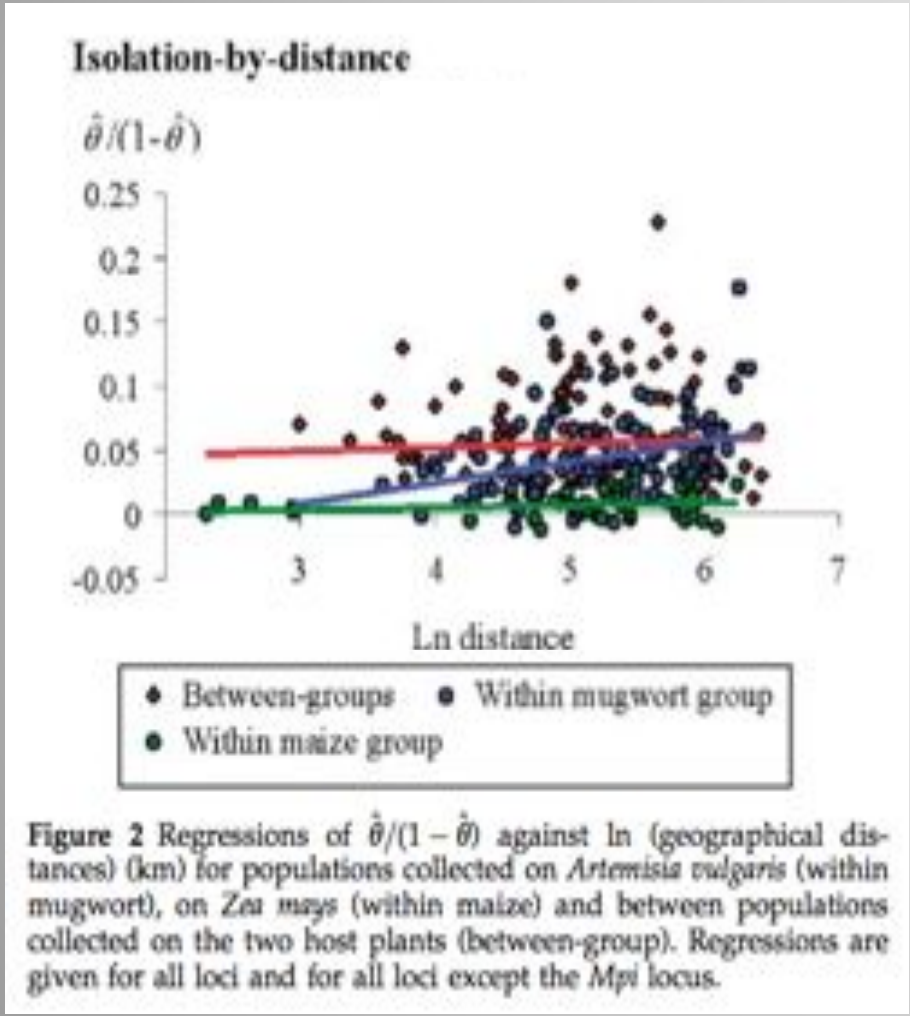


Figure 2 Regressions of $\hat{\theta}/(1-\hat{\theta})$ against ln (geographical distances) (km) for populations collected on *Artemisia vulgaris* (within mugwort), on *Zea mays* (within maize) and between populations collected on the two host plants (between-group). Regressions are given for all loci and for all loci except the *Mpi* locus.

Expectation :

No gene flow between habitats (m<0.01)

➠ differentiation between habitats independent of geographic distance

What is observed :

- Within mugwort-feeding pops ➠ slope is 0.0163 (significantly ≠ 0) and $D\sigma^2$=5 moths
- Within maize-feeding pops ➠ slope is 0.0020. (not ≠ 0) and $D\sigma^2$=40 moths
- Between Maize & Mugwort-feeding pops ➠ slope is 0.0029, (not ≠ 0)
- Differentiation is always higher between habitats than within each habitat

Extensions to classic isolation by distance models

1 – IBD within and between two habitats or groups

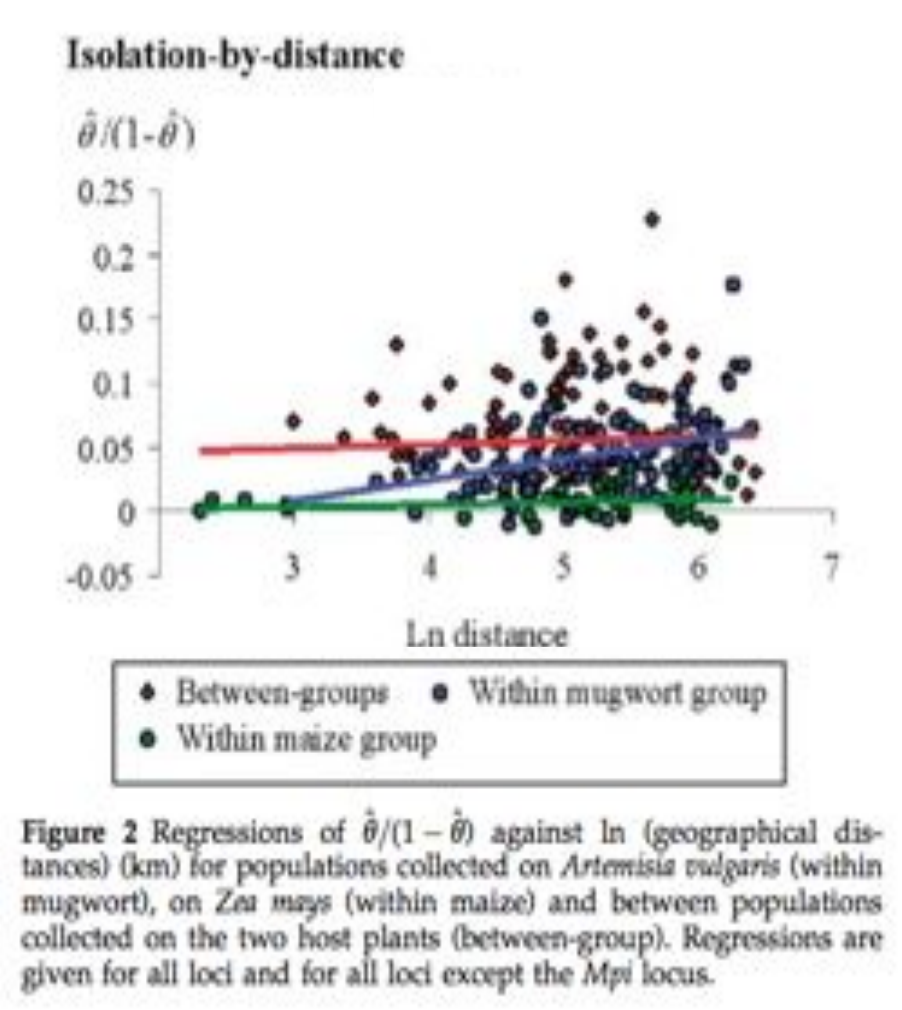


Figure 2 Regressions of $\hat{\theta}/(1-\hat{\theta})$ against ln (geographical distances) (km) for populations collected on *Artemisia vulgaris* (within mugwort), on *Zea mays* (within maize) and between populations collected on the two host plants (between-group). Regressions are given for all loci and for all loci except the *Mpi* locus.

Conclusions :

1. Difference in $D\sigma^2$ between the two host-plant groups probably due to higher densities in maize-feeding populations rather than differences in dispersal
2. there is clearly a strong barrier to gene flow between mugwort and maize-feeding populations of the European corn borer

➠ natural mugwort populations should not be used as refuges because it will not limit evolution of resistance within maize-feeding populations but only within mugwort-feeding populations

Extensions to classic isolation by distance models

2 – euclidian distance vs "least cost distance"

Habitat connectivity is often not homogeneous in space but strongly depends on landscape feature ➠ **using euclidian distance may not be optimal**

ex : Roe deers (*Capreolus capreolus*) in a patchy landscape (Coulon et al. 2004)

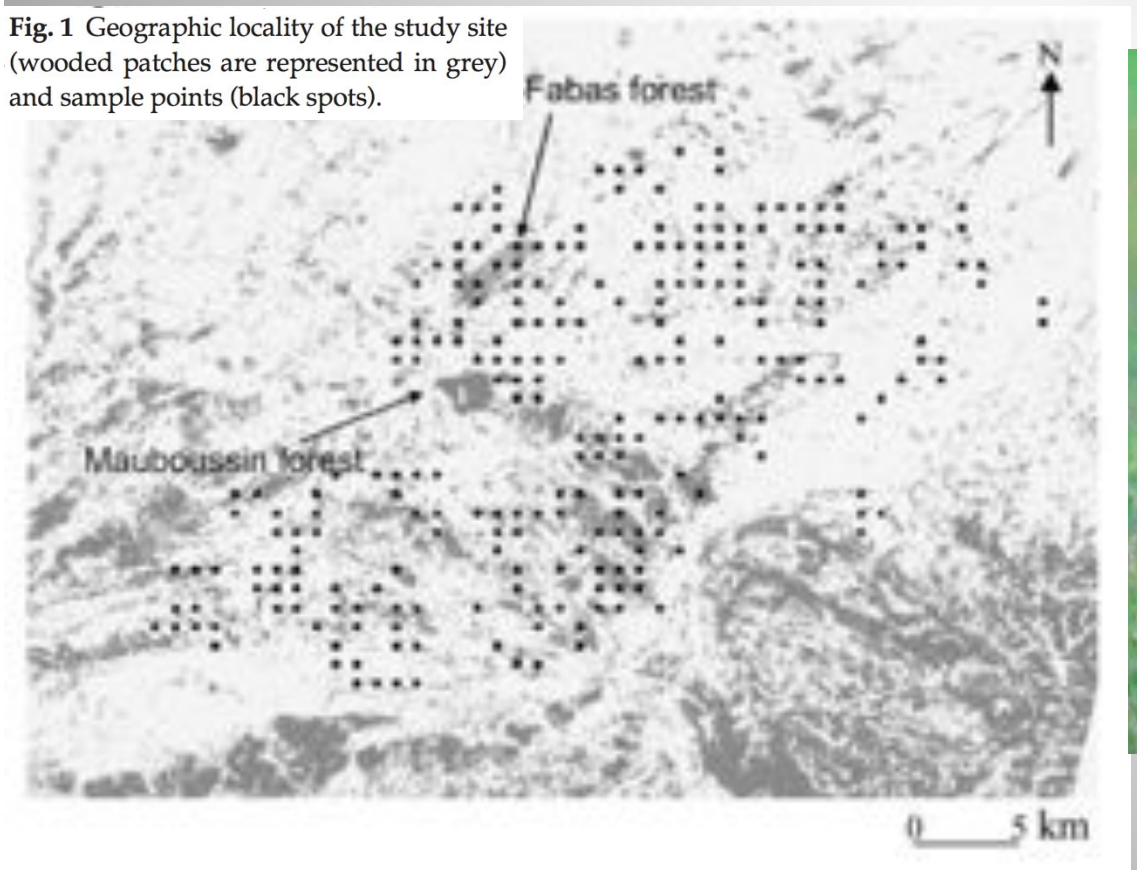


Fig. 1 Geographic locality of the study site (wooded patches are represented in grey) and sample points (black spots).

© Simon Fellous, cNature

Extensions to classic isolation by distance models

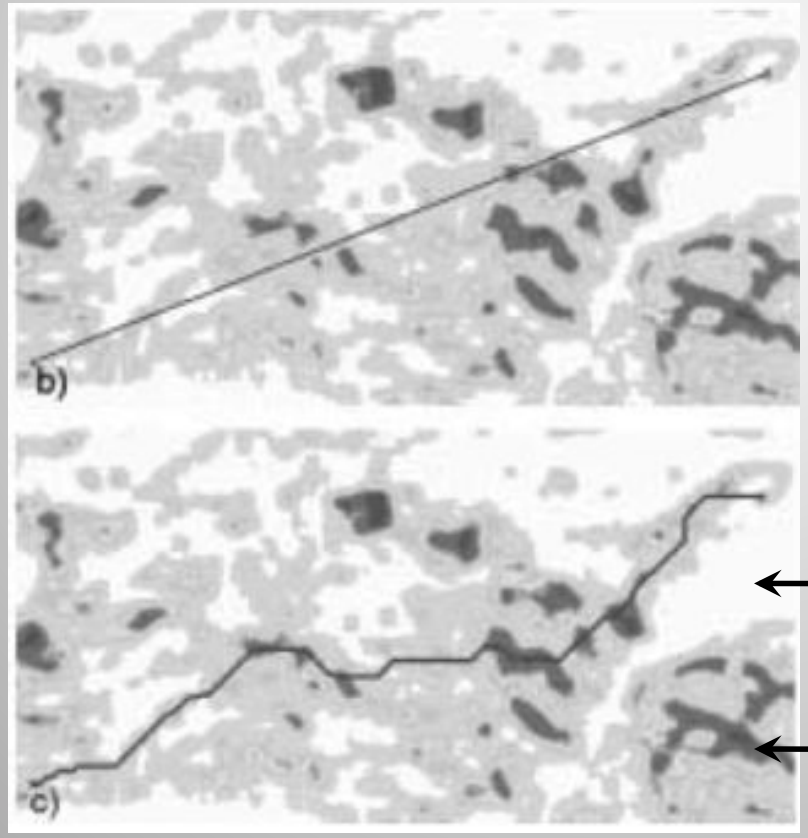
2 – euclidian distance vs "least cost distance"

ex : Roe deer population in a patchy landscape (Coulon et al. 2004)

the least cost distance is the trajectory that maximizes the use of wooded corridors

Euclidian distance

Least cost distance



Open land

Forest patches

Extensions to classic isolation by distance models

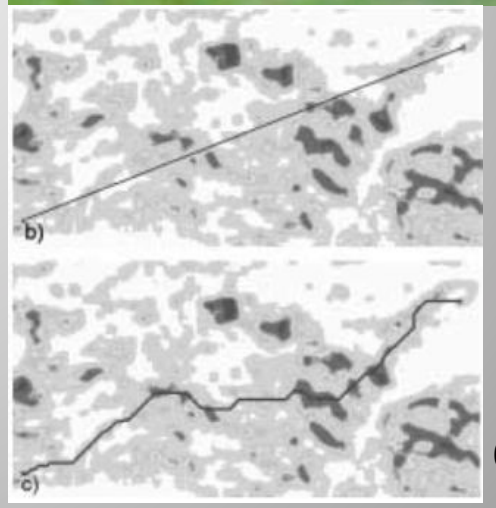
2 – euclidian distance vs "least cost distance"

ex : Roe deer population in a patchy landscape (Coulon et al. 2004)

Table 2 Correlations between genetic and (logarithmic) geographical distances for females and males roe deer. Values of the statistics *r* for Mantel tests are given for each relationship between genetic and geographical distances and the associated probabilities (in brackets) were calculated by carrying out 10 000 permutations of lines or columns of one of the two half-matrices

	Females	Males
ln Euclidean distance	0.019 (0.118)	−0.0001 (0.5)
ln least cost distance	0.031 (0.005)**	0.003 (0.401)

**$P < 0.01$.



✓ Better correlation between genetic differentiation and least cost distance

✓ IBD is only significant for females when considering the least cost distance

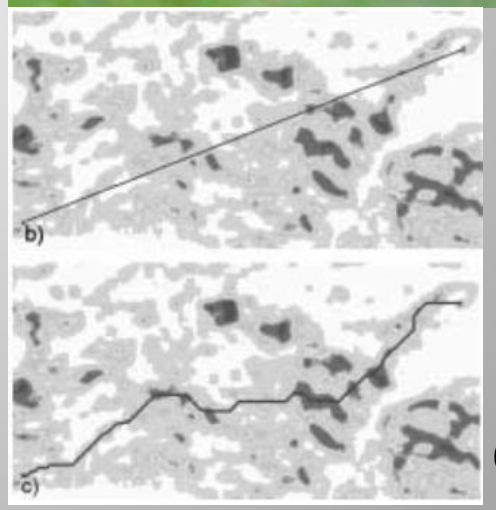
Extensions to classic isolation by distance models

2 – euclidian distance vs "least cost distance"

ex : Roe deer population in a patchy landscape (Coulon et al. 2004)

	Females	Males
ln Euclidean distance	0.019 (0.118)	−0.0001 (0.5)
ln least cost distance	0.031 (0.005)**	0.003 (0.401)

$^{**}P < 0.01$.



Limits and problems:

✓ What cost should we attribute to different landscape features?

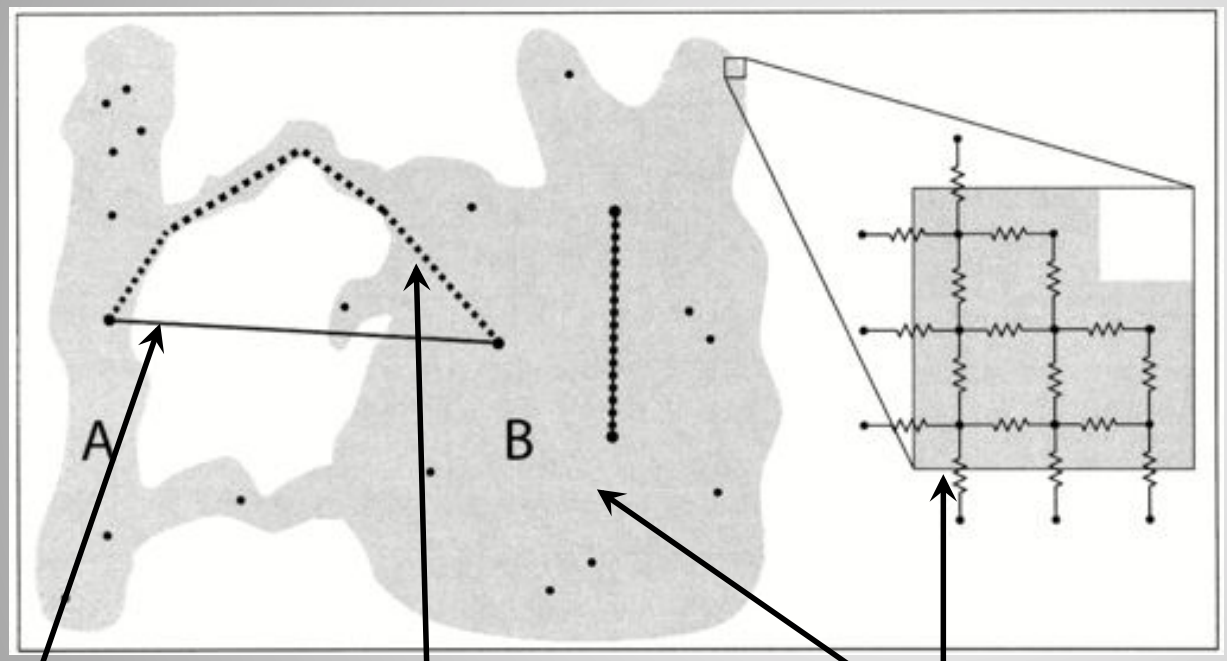
✓ Inference of the cost from genetic data may be really difficult (too many parameters)

✓ Does a better correlation really means a better model under IBD models?

Extensions to classic isolation by distance models

3 – euclidian distance vs resistance distance

Isolation by resistance (McRae 2006 Evolution) : analogy with circuit theory



Not a single path but all potential paths across the whole landscape surface

This "distance" is defined as the effective resistance that would oppose a conductive material displaying a topology similar to that of the study area.

euclidian distance

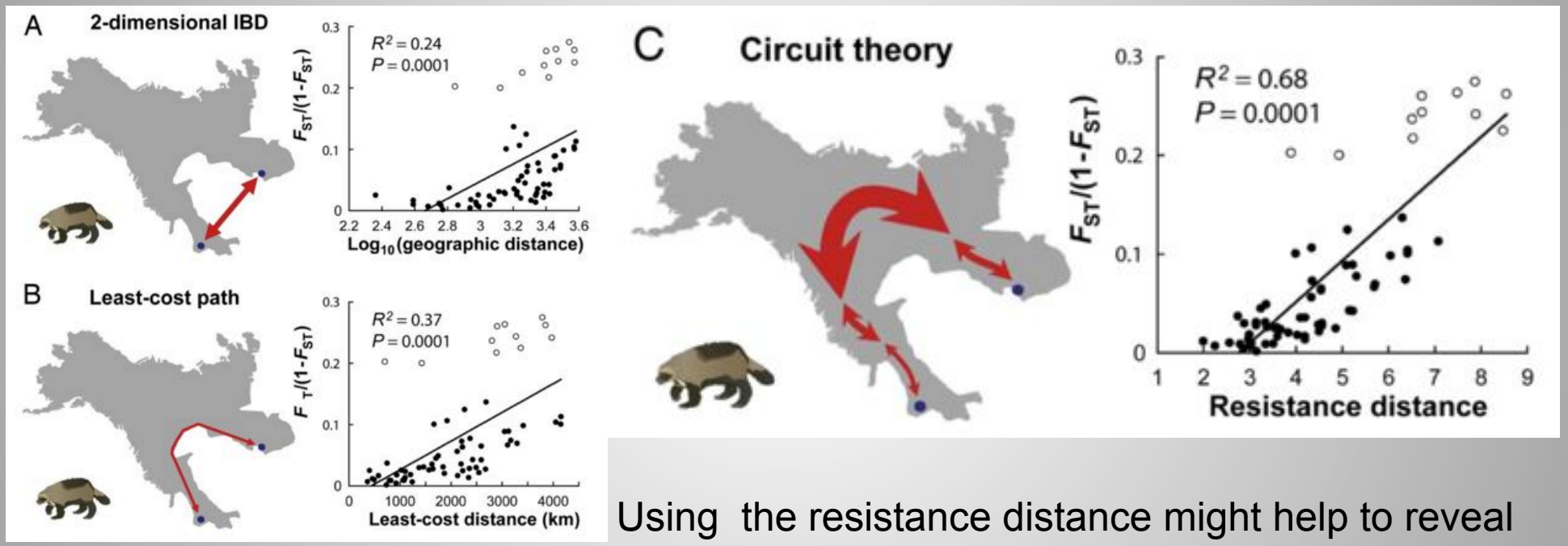
Least cost distance

Resistance surface

Extensions to classic isolation by distance models

3 – euclidian distance vs resistance distance

Isolation by resistance (McRae 2006 Evolution)



Using the resistance distance might help to reveal patterns of IBD in heterogeneous landscapes that would not have appeared with the use of Euclidean or least cost distances

However, as for the least cost methods, it is not straightforward to assign a resistance value for each of the different landscape features

Implications to real data set analyses:
ex: Conservation genetics of forest skinks

Molecular Ecology (2004) 13, 259–269 doi: 10.1046/j.1365-294X.2003.02056.x

Limited effect of anthropogenic habitat fragmentation on molecular diversity in a rain forest skink, *Gnypetoscincus queenslandiae*

JOANNA SUMNER,* TIM JESSOP,† DAVID PAETKAU‡ and CRAIG MORITZ§
*Department of Zoology and Entomology and the Rainforest CRC, University of Queensland, St Lucia, Qld 4072, Australia, †Center for Reproduction of Endangered Species, Zoological Society of San Diego, San Diego, CA, 92112, USA, ‡Wildlife Genetics International, Box 274, Nelson, BC, V1L 5P9, Canada, §Museum of Vertebrate Zoology, University of California, Berkeley, CA, 94720, USA

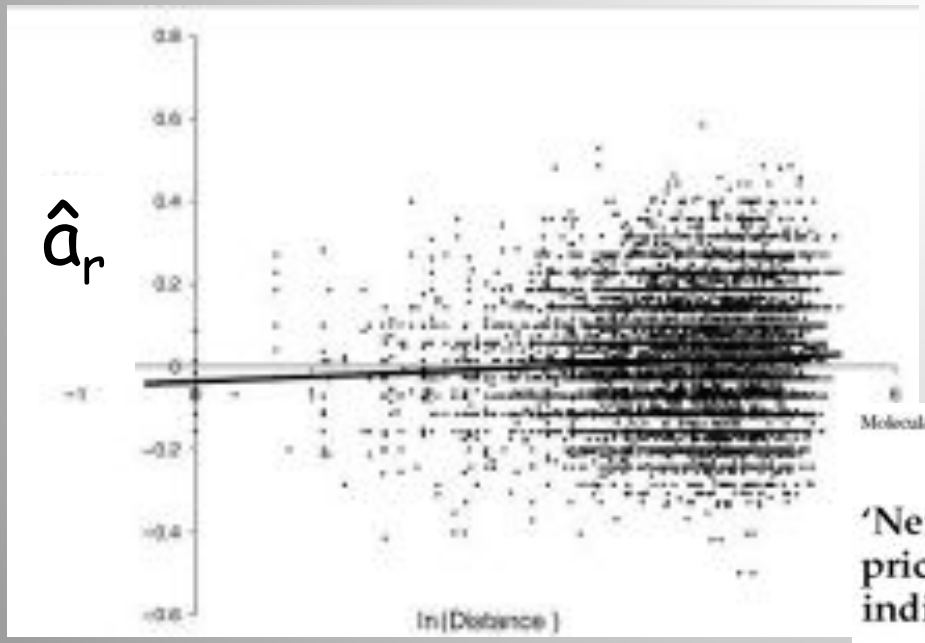
Documented habitat reduction, 10 skink generations ago
⟶ reduced genetic diversity ?

No decrease in N_a, H_e detected with 9 microsatellites…no signs of bottlenecks with specific methods…

Implications to real data set analyses: ex: Conservation genetics of forest skinks

Documented habitat reduction, 10 skink generations ago
⟶ reduced genetic diversity ?

No decrease in N_a, H_e detected with 9 microsatellites…no signs of bottlenecks with specific methods…



But strong isolation by distance $D\sigma^2$=7 [5.5 - 11.5]

Molecular Ecology (2001) 10, 1917–1927

'Neighbourhood' size, dispersal and density estimates in the prickly forest skink (*Gnypetoscincus queenslandiae*) using individual genetic and demographic methods

J. SUMNER,*† F. ROUSSET,‡ A. ESTOUP*§ and C. MORITZ*
*Department of Zoology and Entomology, University of Queensland, Qld 4072, Australia, ‡Laboratoire Génétique et Environnement, CNRS-UMR 5554, 34095 Montpellier, France, §Centre de Biologie et de Gestion des Populations, INRA, 34980 Montferrier/Lez, France

Implications to real data set analyses: ex: Conservation genetics of forest skinks

Documented habitat reduction, 10 skink generations ago
⟶ reduced genetic diversity ?

No decrease in N_a, H_e detected with 9 microsatellites…no signs of bottlenecks with specific methods…

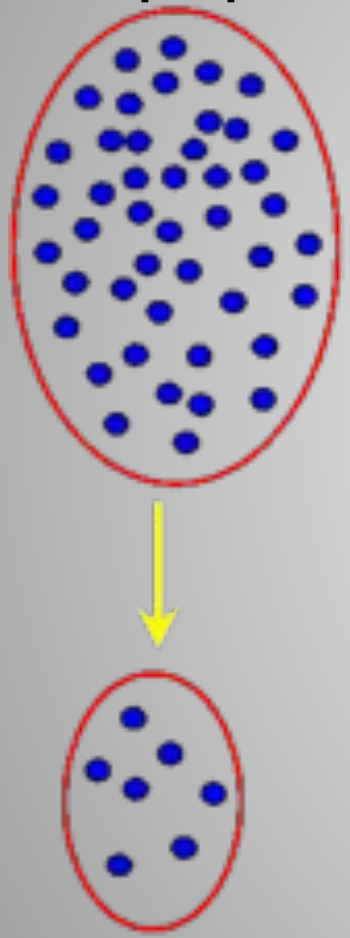
Molecular Ecology (2006) 15, 3601–3615 doi: 10.1111/j.1365-294X.2006.03046.x

Genetics of recent habitat contraction and reduction in population size: does isolation by distance matter?

RAPHAEL LEBLOIS,*†‡ ARNAUD ESTOUP† and REJANE STREIFF†
*Laboratoire Génétique et Environnement, CNRS-UMR 5554, 34095 Montpellier, France, †Centre de Biologie et de Gestion des Populations, INRA, Campus International de Baillarguet, CS 30016, 34988 Montferrier sur Lez cedex, France

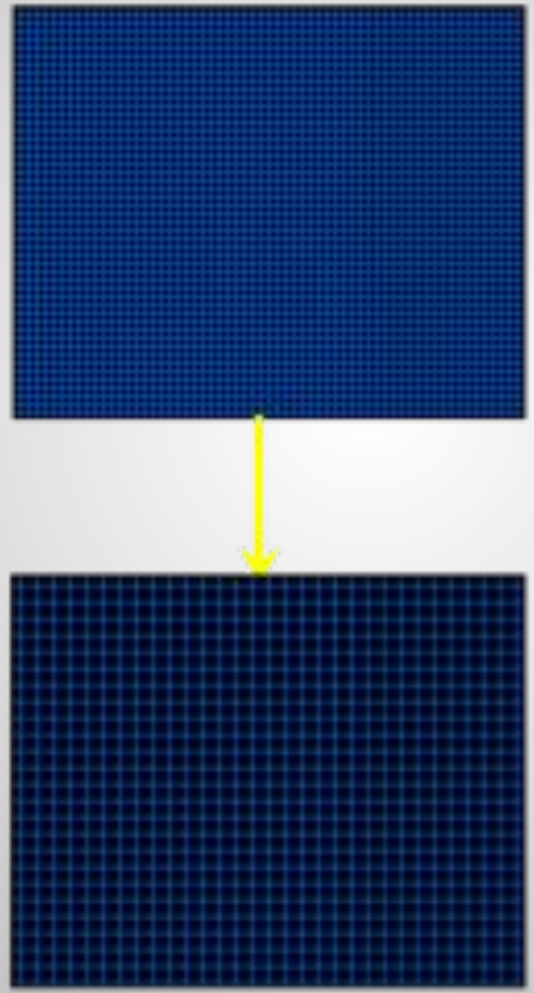
Bottleneck/Reduction in population size under **WF** vs. **IBD**

WF population

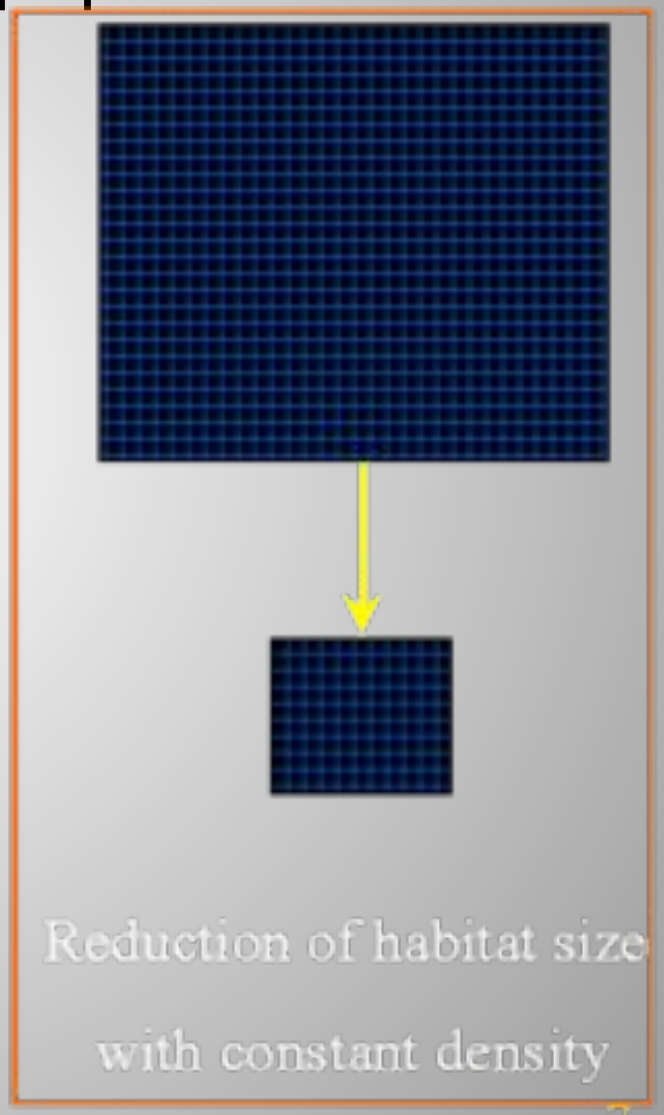


Reduction in population size

IBD population



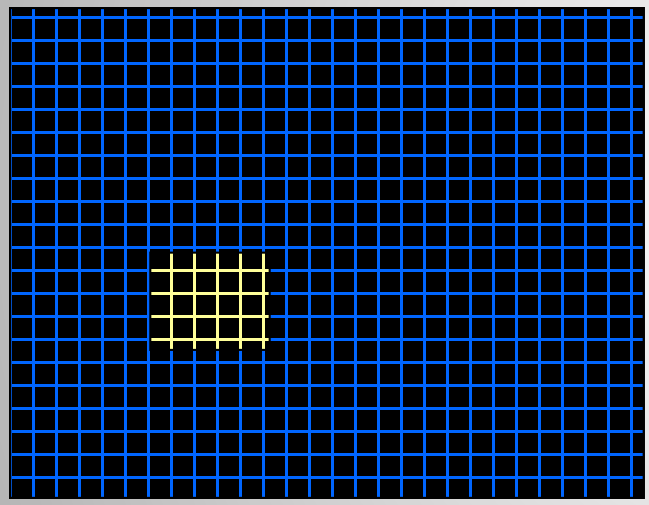
Density reduction



Simulation sampling design

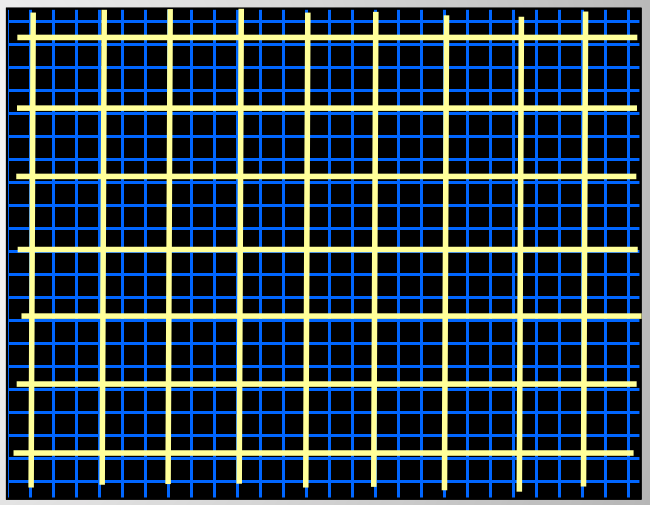
- 30 individuals, 10 loci

- 2 sampling designs for IBD:



Local sample

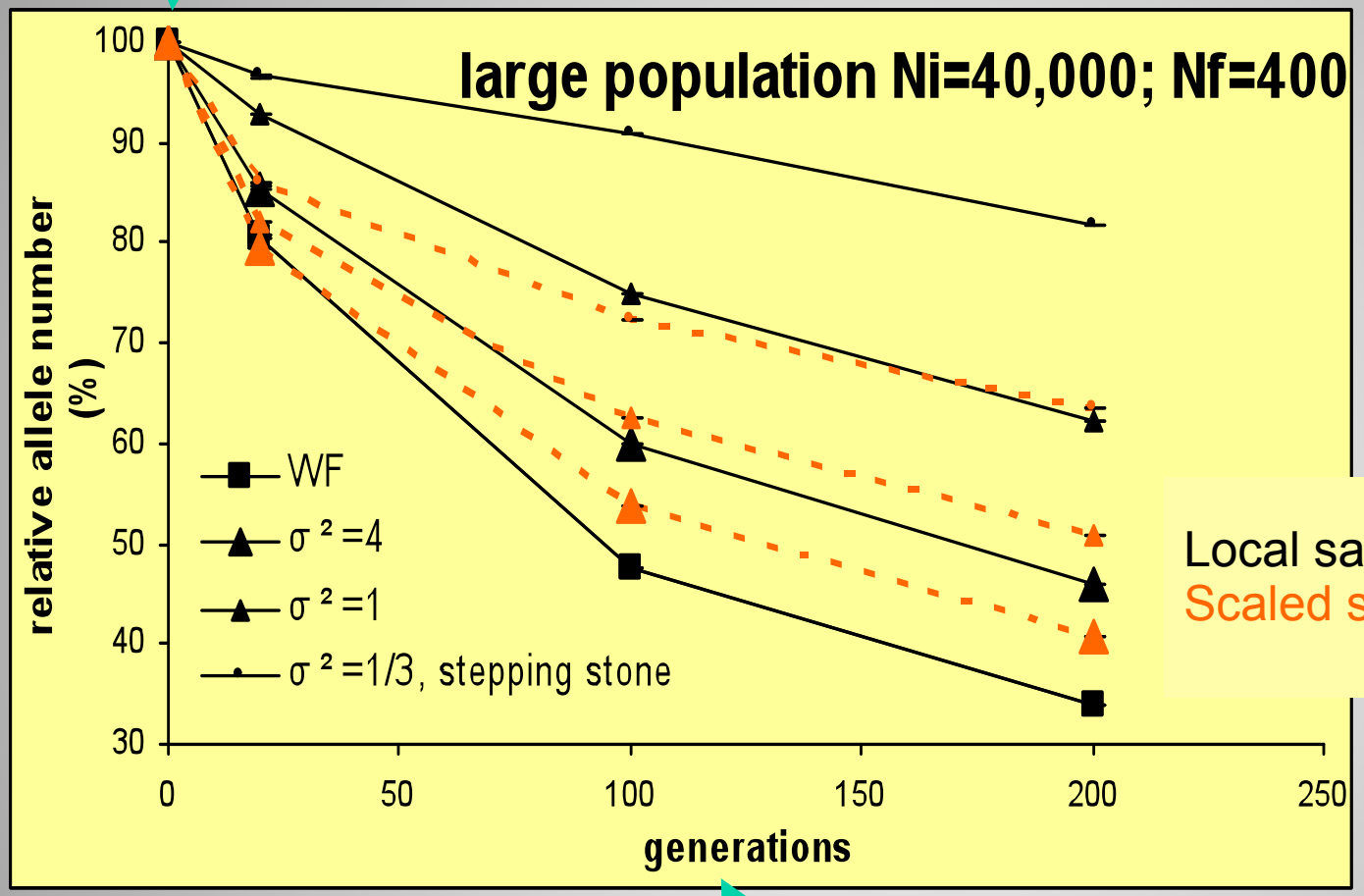
= at adajacent nodes



Scaled sample

= on the entire population surface

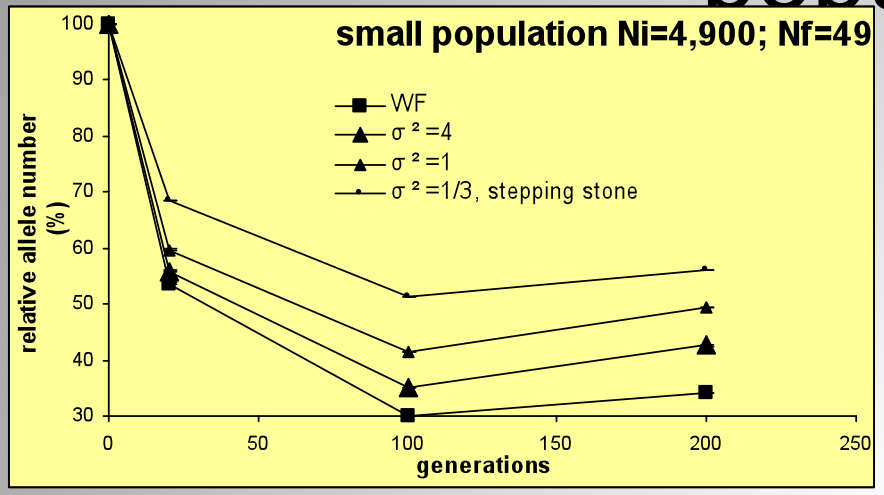
Control (i.e. without bottleneck, size=Ni)



Local sample (LS)
Scaled sample (SS)

Number of generations between bottleneck and sampling times

Results (2) : nA in bottlenecked populations

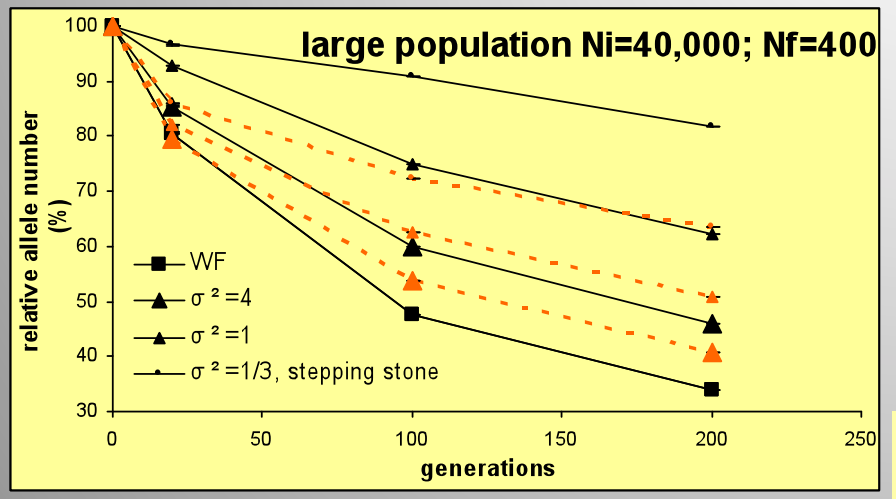


Medium size population = intermediate results…

Number of alleles :

- Influence of IBD is strong
- Influence of the sampling design is substantial in large population

→decrease the differences WF / IBD



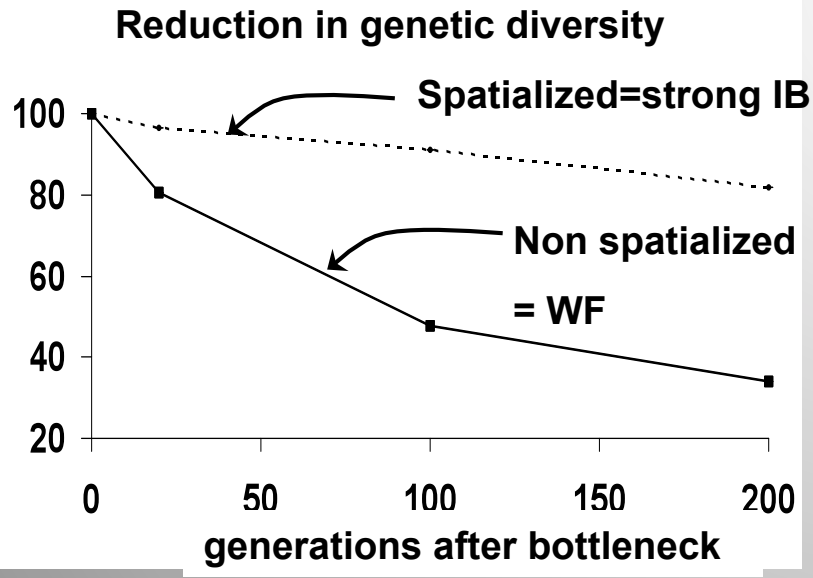
Local sample (LS)
Scaled sample (SS)

Implications to real data set analyses: ex: Conservation genetics of forest skinks

Documented habitat reduction, 10 skink generations ago
⟶ reduced genetic diversity ?

No decrease in N_a, H_e detected with 9 microsatellites…no signs of bottlenecks

Effect of spatial structure (IBD) :



➢ **Genetic diversity (N_a, H_e) is only weakly reduced under IBD after a bottleneck**

\+ **No bottleneck detection under IBD** (BOTTLENECK Cornuet & Luikart 1996, M Garza & Williamson 1996)

\+ **many false expansion signals!**

Importance of the spatial features and localized dispersal

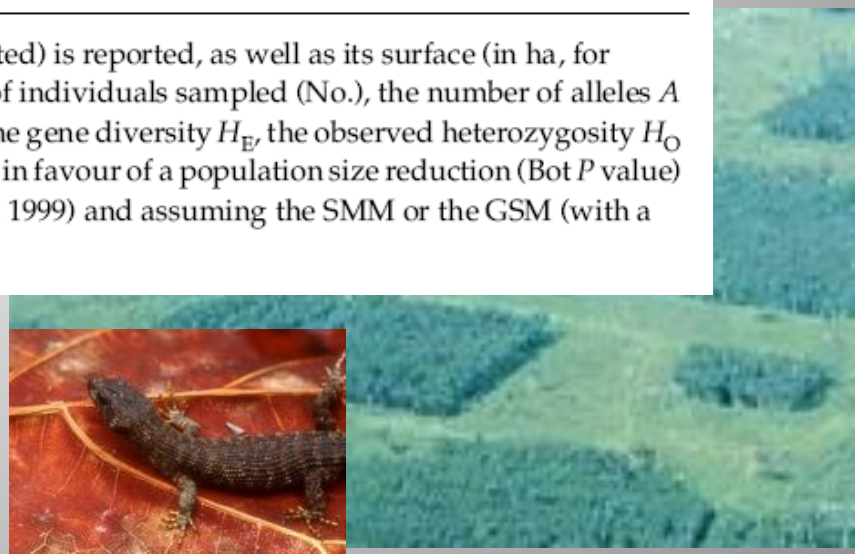
Implications to real data set analyses: ex: Conservation genetics of forest skinks

Table 2 Summary of skink population data set

Population name	Type	Area	Size	No.	A	H_E	H_O	M	Bot P value		Exp P value	
									SMM	GSM	SMM	GSM
Souita Falls (F1)	F	2.0	101–384	25	7.35	0.70	0.68	0.458	0.99	0.85	0.007**	0.18
Maalan Road (F2)	F	2.46	124–472	42	7.44	0.63	0.66	0.549	1.00	0.99	0.001**	0.01**
Waltham (F3)	F	2.64	133–507	27	7.75	0.68	0.67	0.496	0.99	0.90	0.007**	0.13
Pat Daley Park (F4)	F	5.96	300–1144	18	7.39	0.63	0.63	0.406	1.00	0.99	0.002**	0.01**
Nose Ring (F5)	F	24.19	1217–4645	29	9.31	0.70	0.68	0.524	1.00	0.99	0.005**	0.007**
Whiteing Road (F6)	F	36.31	1826–6972	30	8.77	0.71	0.69	0.522	1.00	0.99	0.005**	0.01**
Millaa Millaa Falls (F7)	F	65.06	3273–12497	27	7.88	0.66	0.65	0.465	1.00	1.00	0.001**	0.003**
Brotherton(C1)	C	NA	>> 40 000	28	8.26	0.69	0.68	0.523	0.99	0.98	0.019**	0.065
Cross-eye(C2)	C	NA	>> 40 000	30	8.00	0.69	0.70	0.526	1.00	1.00	0.001**	0.005**
Mount Father Clancy(C3)	C	NA	>> 40 000	32	8.77	0.69	0.72	0.575	1.00	1.00	0.002**	0.005**
Reynolds(C4)	C	NA	>> 40 000	28	9.18	0.66	0.71	0.566	1.00	1.00	0.002**	0.002**
Massey Creek(C5)	C	NA	>> 40 000	94	7.95	0.62	0.62	0.580	1.00	1.00	0.001**	0.002**

For each population, habitat type (F, fragmented and C, continuous, i.e. nonfragmented) is reported, as well as its surface (in ha, for fragments only), its approximate size in term of number of individuals, the number of individuals sampled (No.), the number of alleles A (adjusted for a sample size of 30 individuals using Ewens 1972's sampling formula), the gene diversity H_E, the observed heterozygosity H_O and the value of M statistics. The probability of rejecting the hypothesis of equilibrium in favour of a population size reduction (Bot P value) or expansion (Exp P value) was computed using the software BOTTLENECK (Piry *et al*. 1999) and assuming the SMM or the GSM (with a variance of 0.36) as mutation model.

Importance of the spatial features and localized dispersal



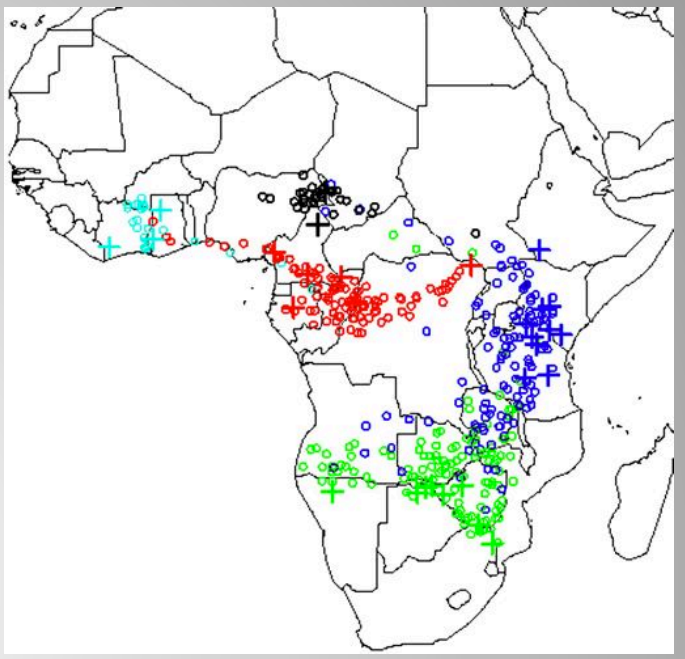
Extensions to classic isolation by distance models

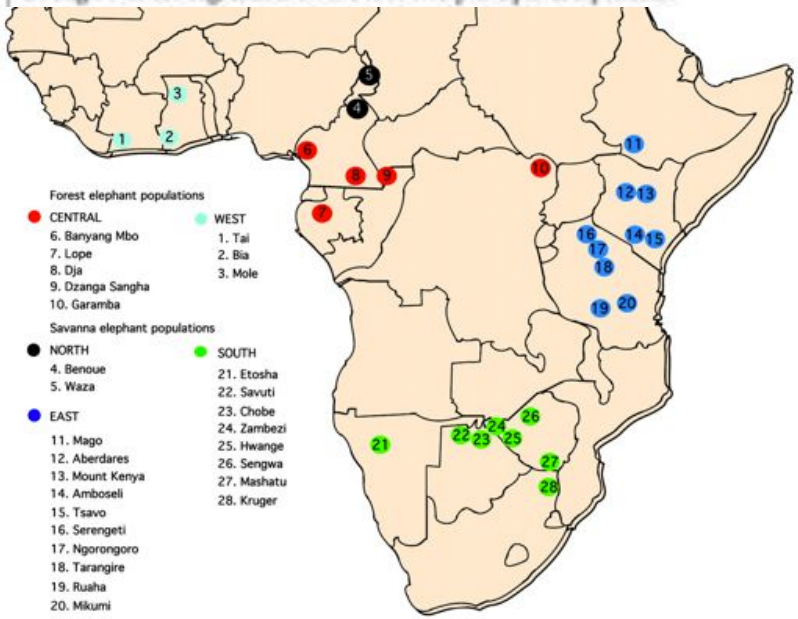
Box 1: Using isolation-by-distance patterns to perform spatially continuous assignment

Random genetic drift under IBD tends to produce smooth spatial variations of allele frequencies. Inferred maps of allele frequencies can be used to perform geographically explicit individual assignments. Wasser *et al.* (2004) and Wasser *et al.* (2007) developed a method that jointly estimates such maps and estimates the unknown geographic origin of a DNA sample by comparing its alleles with estimated allele frequencies. Rather than simply assigning individuals to predefined populations, the method can, in principle, assign individuals to any spatial location whose inferred allele frequencies best explains the genotype of the sample. Using this method, Wasser *et al.* (2007) showed that a large shipment of contraband ivory originated from a narrow region centred on Zambia. The accuracy of the assignment depends on the accuracy of the allele frequency map implicitly generated during the inference step, which in turn depends on the size of the training data set and on how much allele frequencies characterize a given region.

Pope *et al.* (2007) found that the individual spatial assignments generated by the method proposed by Wasser *et al.* (2004) could give ambiguous results (many possible locations). This might result from: (i) a lack of differentiation in the data; (ii) uncertainty about allele frequencies due in particular to the use of data with individuals continuously sampled over space; (iii) departure of data from the underlying statistical model; (iv) overparametrization compared with sample size; (v) MCMC convergence flaw. Pope *et al.* (2007) devised a simpler method based on the same rationale. They used their method to compare the movement of individual badgers before and after a culling operation performed in the context of bovine tuberculosis (*Mycobacterium bovis*) control. Even though they showed that the badgers moved, on average, further post- than pre-cull, it yet remains to be seen how accurate Pope *et al.*'s method is in the assignment of individuals to specific geographic localities.

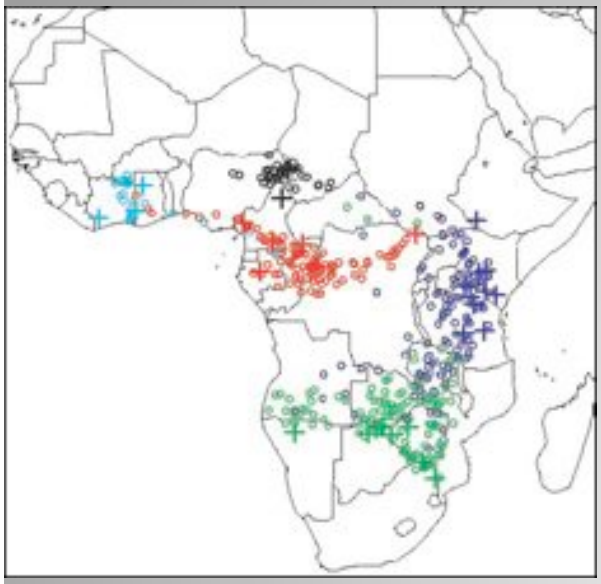
In a study in human genetics, modelling allele frequencies as a linear function of spatial coordinates at the synoptic scale, Arnos & Manica (2006) were capable of assigning individuals with an accuracy of 1200 miles. Novembre & Stephens (2008) proposed a method based on a PCA suitable for large SNPs data that predict spatial origin through a linear regression on the first two principal components.





(a) Map of Africa showing the collection sites divided into five regions: West Africa (cyan), Central forest (red), and Central (black), South (green) and East (blue) savanna. (b) Estimated locations of elephant tissue and faecal samples from across Africa when assignments are allowed to vary anywhere within the elephants' range. All tissue and scat samples ($n = 399$) were successfully amplified at seven or more loci. Sampling locations are indicated by a cross and are colour coded according to actual broad geographic region of origin: West Africa, Central forest, and Central, South and East savanna [colour coded as in (a)]. Assigned location of each individual sample is shown by a circle and is colour coded according to its actual region of origin. The closer each circle is to crosses of the same colour, the more accurate is that individual's assignment (figures and caption reprinted from Wasser *et al.* 2004).

Extensions to classic isolation by distance models



3 –

Assignment results for 37 tusks from a large seizure in Singapore. Circles represent the estimated origin of the 37 tusks analyzed. Plus signs coincide with the those in the figure above. [from Wasser et al. 2007]