Inference of demographic parameters from genetic data

Raphael Leblois, CBGP – Montpellier

December 2022

CBGP meeting day: “Dynamique et Génétique des Populations ”

Indirect demographic inferences

1 - Genetic data carry information about evolutionary (demographic?) parameters

2 - First historical developments of indirect demographic inference and their limits

3 - Are these limitations a real barrier to indirect demographic inference

4 - Introduction to spatial models in population genetics : Isolation By Distance (IBD)

5 - Historical developments to infer demographic parameters under IBD

6 - IBD : relevant models for local demographic inferences

Discussion…

Indirect demographic inferences

1 - Genetic data carry information about evolutionary (demographic?) parameters

2 - First historical developments of indirect demographic inference and their limits

3 - Are these limitations a real barrier to indirect demographic inference

4 - Introduction to spatial models in population genetics : Isolation By Distance (IBD)

5 - Historical developments to infer demographic parameters under IBD

6 - IBD : relevant models for local demographic inferences

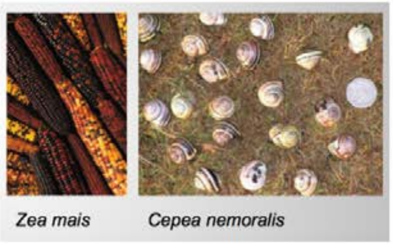
Discussion…

Population genetics aims at analyzing the **processes** controlling **genetic polymorphism** (= variability) in populations

- Describe the genetic polymorphism and its distribution within and between individuals and populations
- Infer the processes (evolutionary forces) that shape(d) the genetic polymorphism

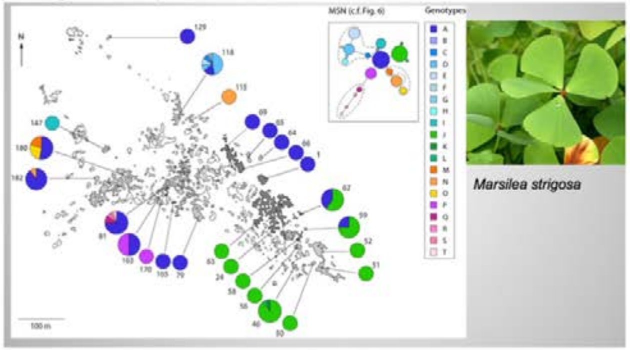
→ Understand **how evolution works**

Repartition of the genetic polymorphism:



Zea mais *Cepea nemoralis*

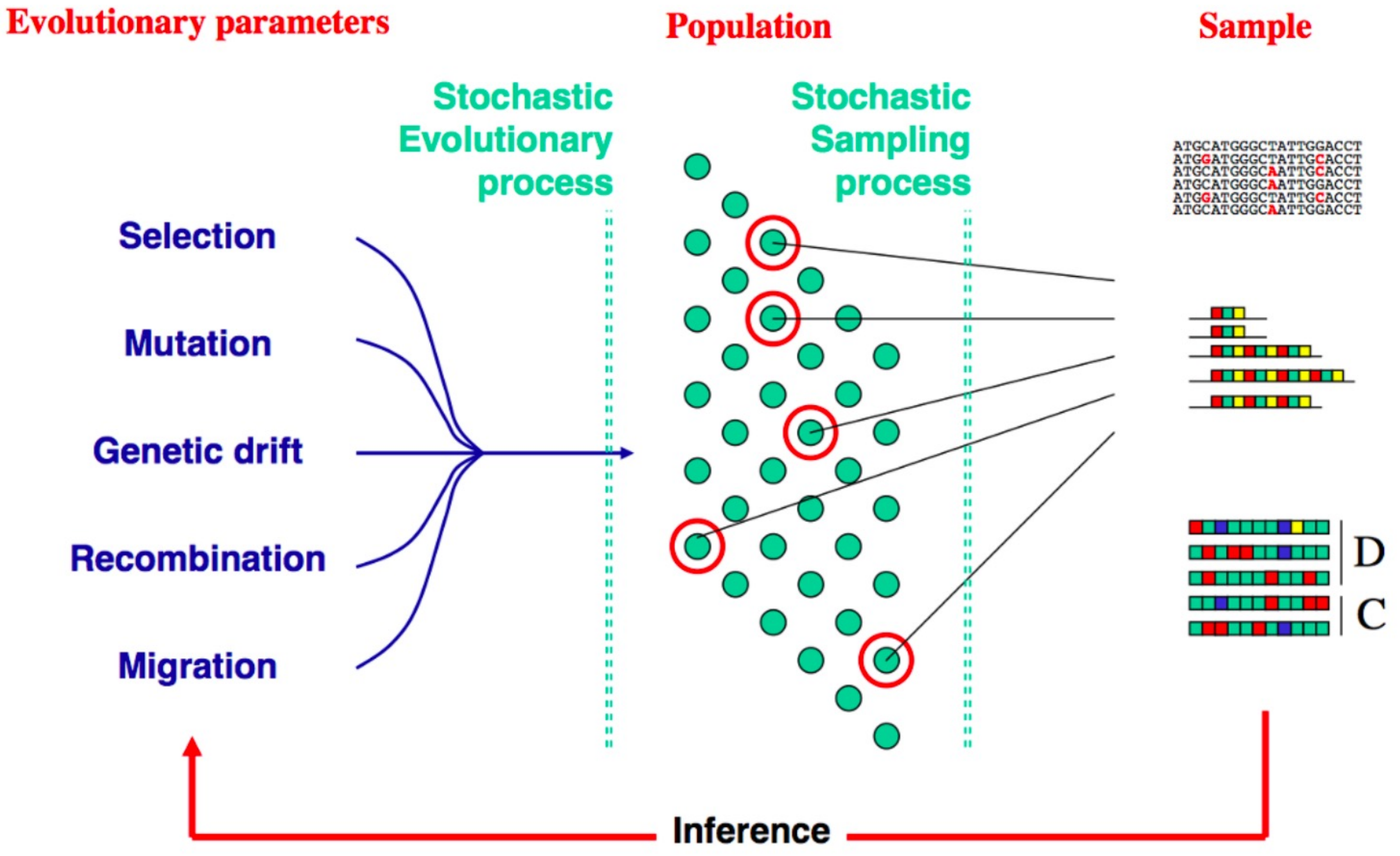
within and between individuals



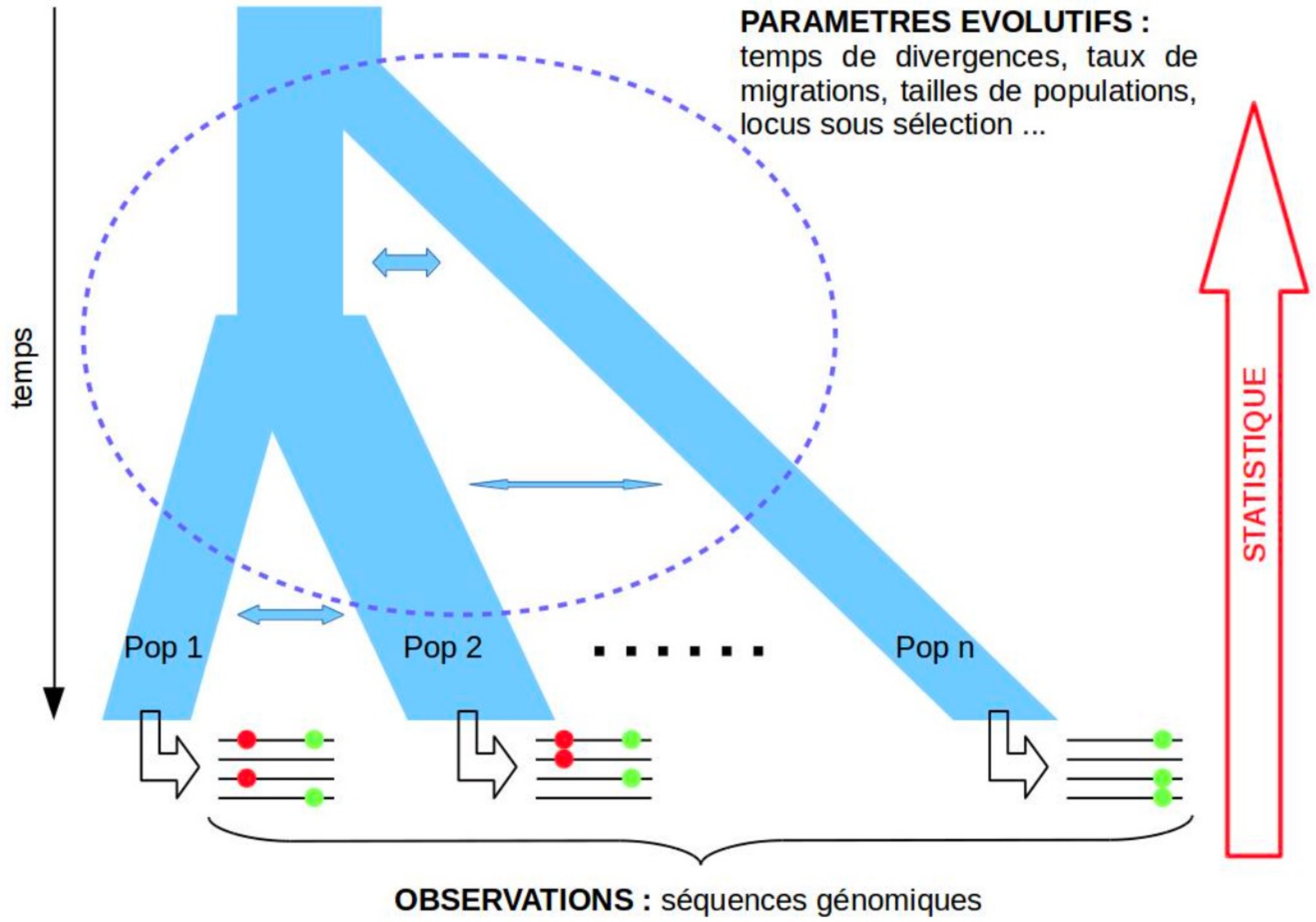
Marsilea strigosa

within and between populations

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



Inference of evolution history at short time scale (within species) from **molecular data**.


PARAMETRES EVOLUTIFS :
temps de divergences, taux de migrations, tailles de populations, locus sous sélection ...
temps
STATISTIQUE
Pop 1
Pop 2
. . . . . .
Pop n
OBSERVATIONS : séquences génomiques

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" approaches for conservation biology, study of invasive species, agro-ecology…

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 ?

Evolutionary vs. demographic parameters

Classical evolutionary forces / parameters

Drift (population size *N*)

Mutation μ ($N*\mu$)

Selection *s* (*N*s*)

Recombination r (N*r)

Migration m (N*m)
dispersal m (N*m) + others ($g_{\_geom}$, …)

“Classical” (?) demographic parameters

Population size

Dispersal/Migration

More “individual parameters”

Survival / mortality

Fecondity

Growth (Age classes)

….

effective parameters vs. census parameters
(i.e., with a successful reproduction) vs. (i.e., followed or not by a success- ful reproduction)

Evolutionary vs. demographic parameters

Classical evolutionary forces / parameters

Drift (population size *N*)

Mutation μ ($N*\mu$)

Selection s ($N*s$)

Recombination r (N*r)

Migration m (N*m)
dispersal m (N*m) + others (g$_{\_geom}$, …)

And their variation through time

“Classical” (?) demographic parameters

Population size

Dispersal/Migration

More “individual parameters”

Survival / mortality

Fecondity

Growth (Age classes)

….

effective parameters vs. census parameters
(i.e., with a successful reproduction) vs. (i.e., followed or not by a success- ful reproduction)

Indirect demographic inferences

1 - Genetic data carry information about evolutionary (demographic?) parameters

2 - First historical developments of indirect demographic inference and their limits

3 - Are these limitations a real barrier to indirect demographic inference

4 - Introduction to spatial models in population genetics : Isolation By Distance (IBD)

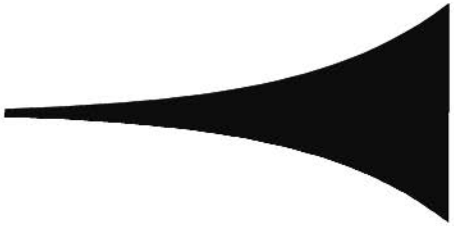
5 - Historical developments to infer demographic parameters under IBD

6 - IBD : relevant models for local demographic inferences

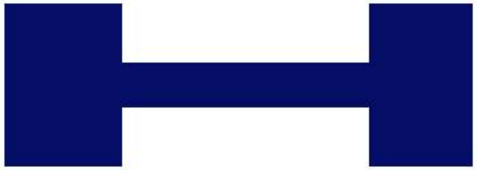
Discussion…

Demographic models classically used in population genetics

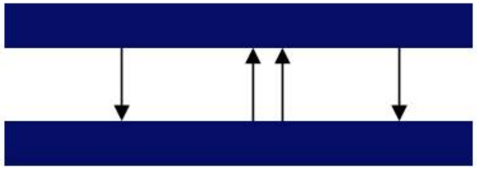
- Population growth



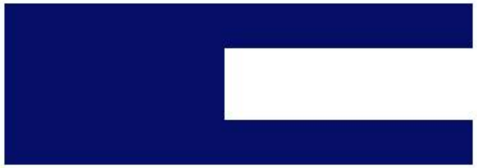
- Population bottlenecks



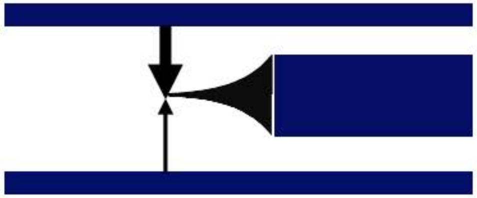
- Subdivided populations



- Population splits

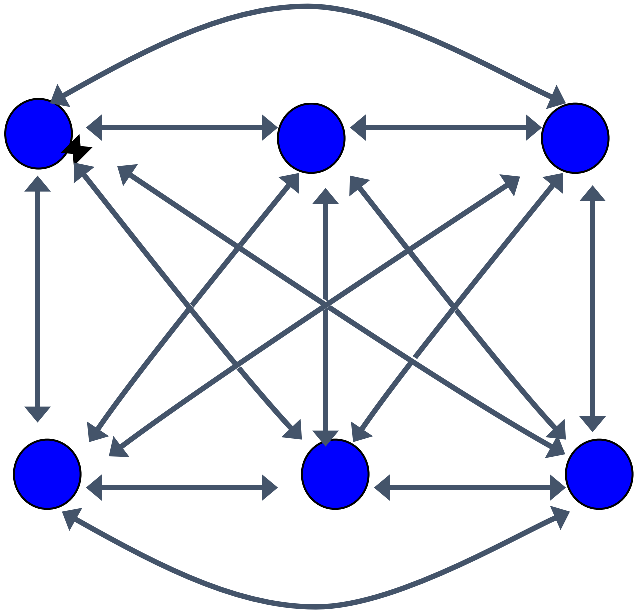


- Admixture



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

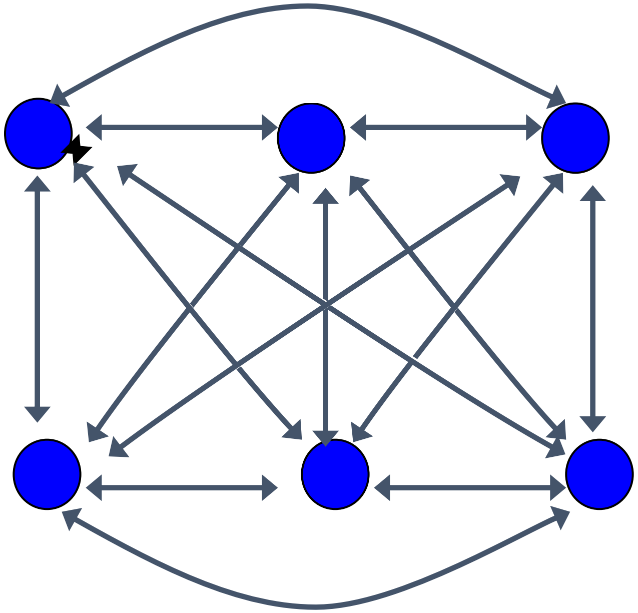
m = migration rate

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

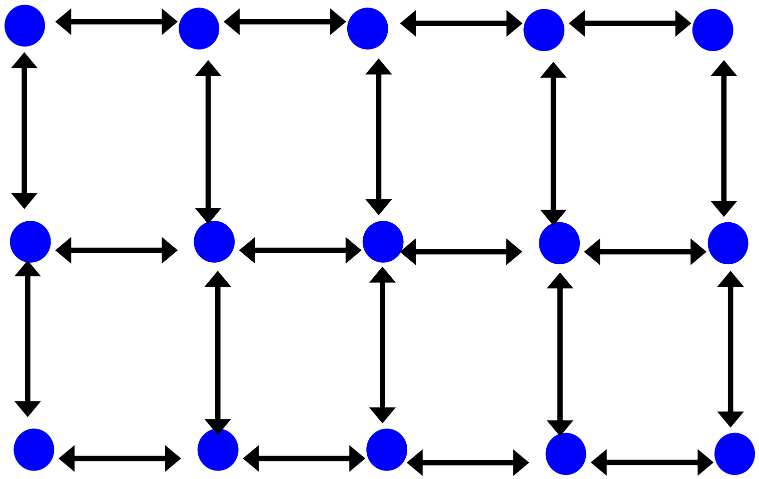
Extremely useful to study theoretical evolutionary effects of migration

and widely used until 2000 (with low number of genetic markers)

but generally not realistic enough to allows precise demographic inferences …

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 between adjacent demes

Fully homogeneous and "spatial"

Extremely useful to study theoretical evolutionary effects of migration

and widely used until 2000 (with low number of genetic markers)

but generally not realistic enough to allows precise demographic inferences …

Before the numeric (and genomic) area, inferences were based on

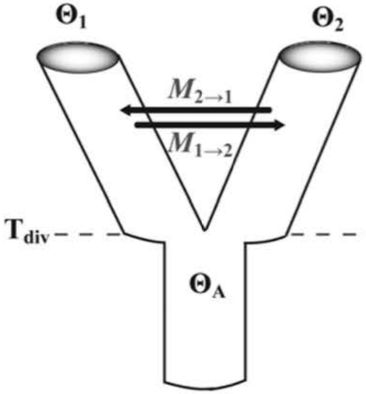
- single **summary statistics**, related to a model parameter, e.g.

$$F_{st} \approx \frac{1}{(1+2Nm)} \rightarrow \hat{Nm} \approx \frac{1}{4}(\frac{1}{\hat{F}_{st}-1}) \text{ (island model)}$$

$$F_{st} \approx 1-(1-\frac{1}{(2N)})^t \approx 1-\exp(-t/(2N) \rightarrow$$

$$t/\hat{2}N \approx -log(1-\hat{Fst}) \text{ (pure divergence model)}$$

strong limitation : can not consider more complex models, e.g.



Divergence with Migration

Before the numeric (and genomic) area, inferences were based on

- single **summary statistics**, related to a model parameter, e.g.

$$F_{st} \approx \frac{1}{(1+2Nm)} \rightarrow \hat{Nm} \approx \frac{1}{4}(\frac{1}{\hat{F_{st}}-1})$$

$$F_{st} \approx 1-(1-\frac{1}{(2N)})^{t} \approx 1-\exp(-t/(2N) \rightarrow$$

$$t/\hat{2}N \approx -log(1-\hat{Fst})$$

- single **summary statistics**, related to a caracteristic of the model, e.g.

  excess or deficit of H_e → signal of a bottleneck or an expansion, respectively
- very few more sophisticated inferences based on :
  - simple (oversimplified) models with few parameters
  - with mathematical and/or biological approximations (e.g. of the **likelihood**, no mutations,...)

Many estimations in model and non-model species from 1980 to 2010, but with two major obvious limitations :

- Limited information in few markers
- use only a small fraction of the information carried by the genetic data
- Non-realistic / oversimplistic demographic 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

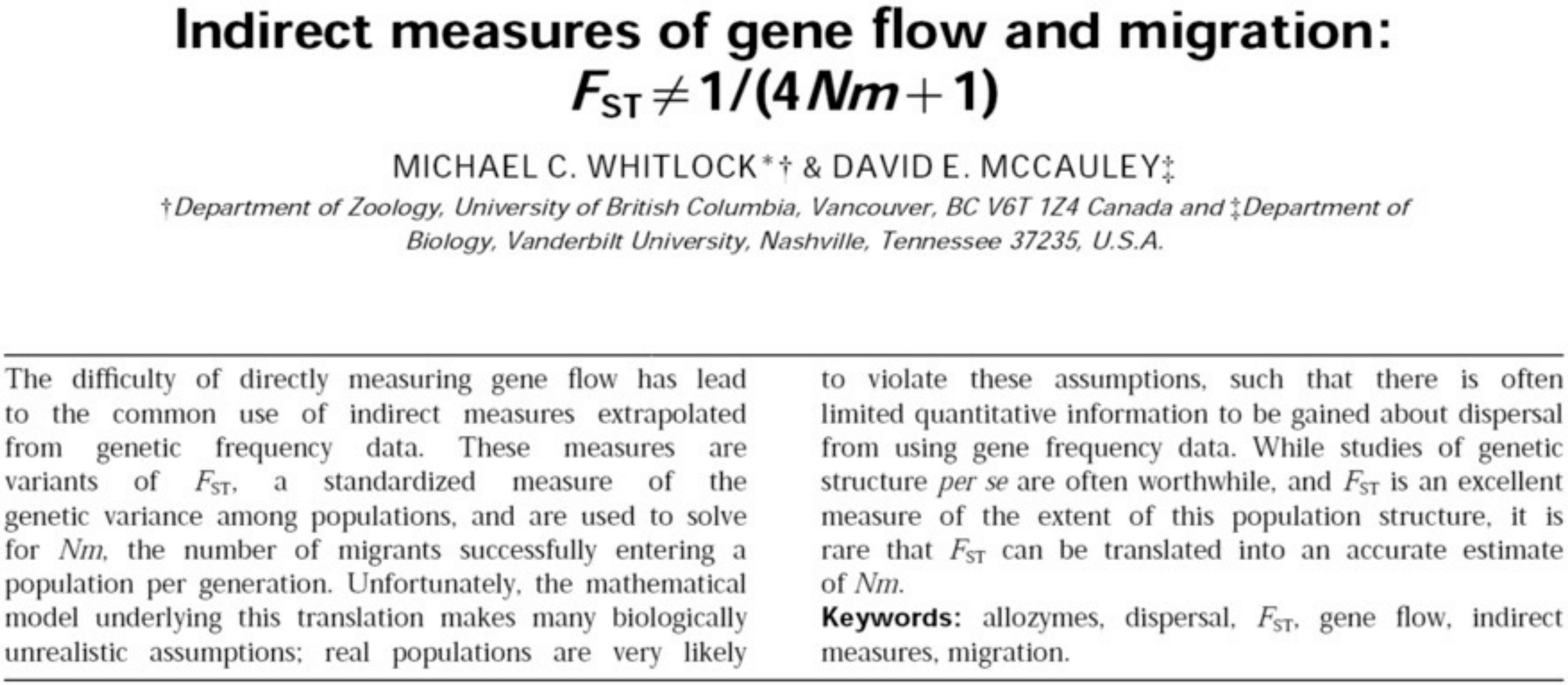
➢ Natural population are often inhomogeneous and at disequilibrium, whereas most demo-genetic models assume spatial homogeneity and time equilibrium

➢ 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

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) :



Indirect measures of gene flow and migration: $F_{ST} \neq 1/(4Nm+1)$

MICHAEL C. WHITLOCK*† & DAVID E. MCCAULEY‡

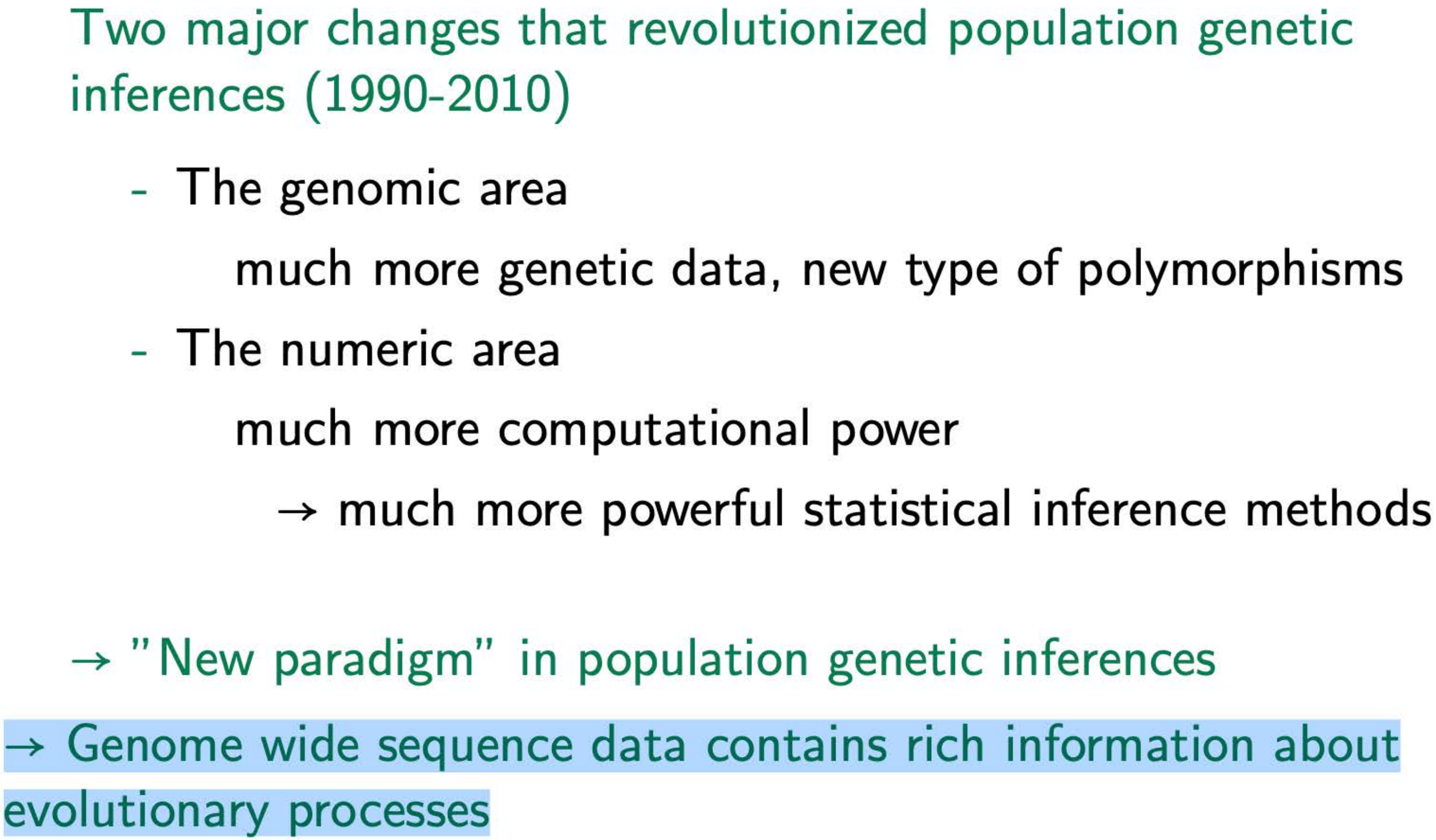
†Department of Zoology, University of British Columbia, Vancouver, BC V6T 1Z4 Canada and ‡Department of Biology, Vanderbilt University, Nashville, Tennessee 37235, U.S.A.

The difficulty of directly measuring gene flow has lead to the common use of indirect measures extrapolated from genetic frequency data. These measures are variants of F_{ST}, a standardized measure of the genetic variance among populations, and are used to solve for *Nm*, the number of migrants successfully entering a population per generation. Unfortunately, the mathematical model underlying this translation makes many biologically unrealistic assumptions; real populations are very likely to violate these assumptions, such that there is often limited quantitative information to be gained about dispersal from using gene frequency data. While studies of genetic structure *per se* are often worthwhile, and F_{ST} is an excellent measure of the extent of this population structure, it is rare that F_{ST} can be translated into an accurate estimate of *Nm*.

Keywords: allozymes, dispersal, F_{ST}, gene flow, indirect measures, migration.

Many estimations in model and non-model species from 1980 to 2010, but with two major obvious limitations :

- Limited information in few markers
- use only a small fraction of the information carried by the genetic data
- Non-realistic / oversimplistic demographic models



Two major changes that revolutionized population genetic inferences (1990-2010)

- The genomic area

  much more genetic data, new type of polymorphisms

- The numeric area

  much more computational power

  → much more powerful statistical inference methods

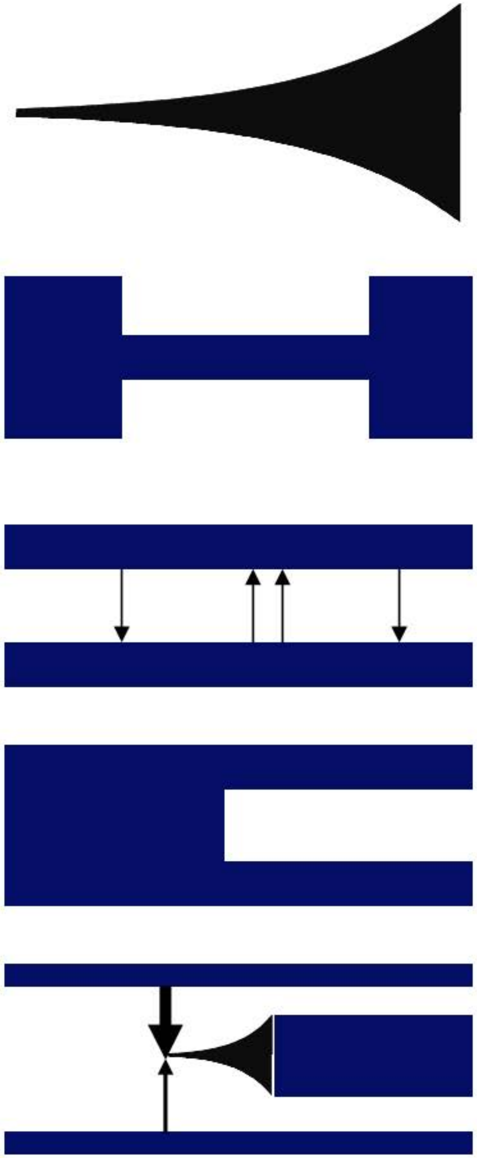
→ "New paradigm" in population genetic inferences

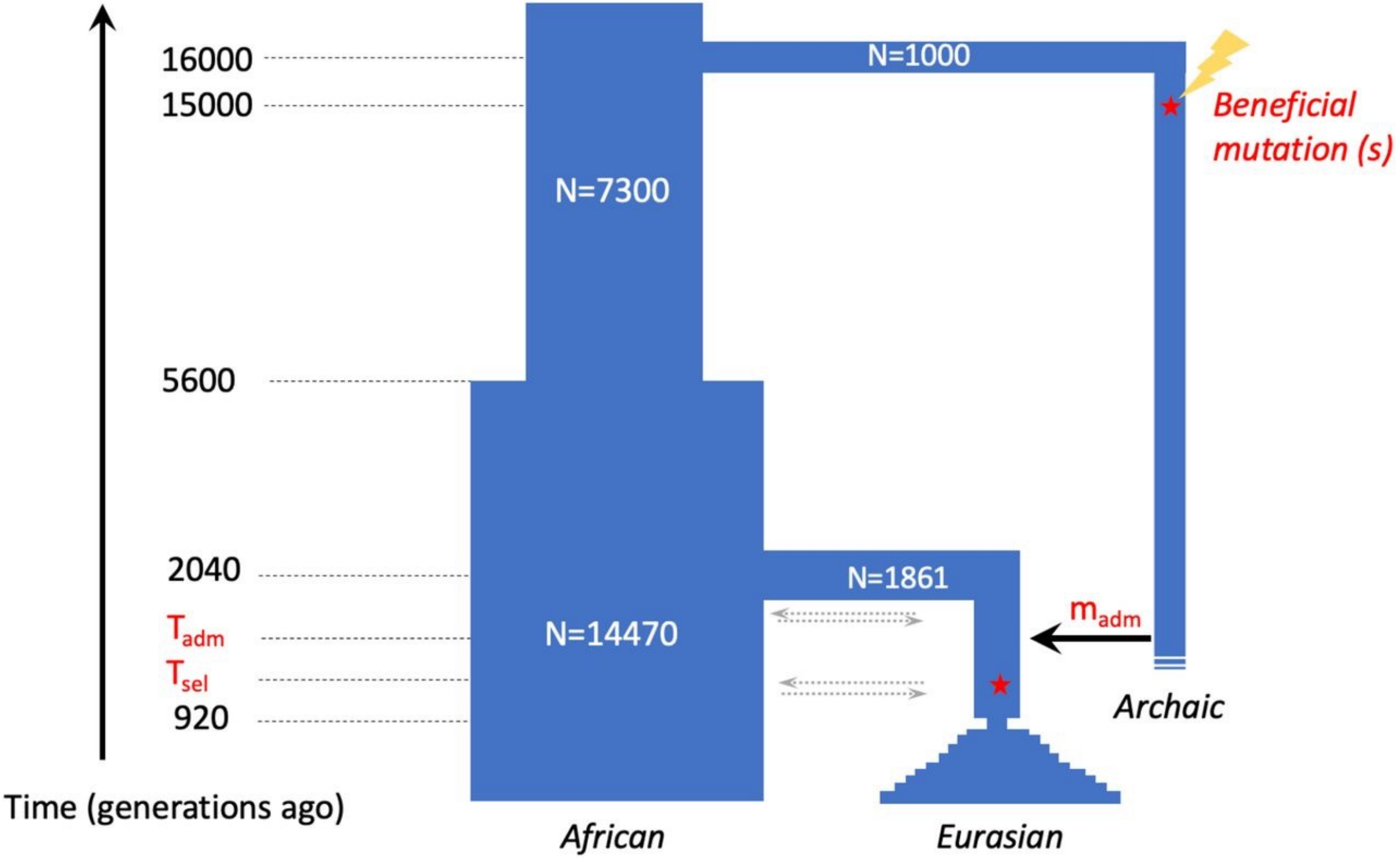
→ Genome wide sequence data contains rich information about evolutionary processes

More markers, more computers -> we can now consider more complex models made of combination of :

Demographic models classically used in population genetics

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





e.g. demographic and adaptative scenario for human evolution, Zhang_et_al_2022

Now, we have more and more powerful computers and clusters, allowing computationally intensive statistical inferences using:

- Monte Carlo simulation (to explore large parameter space), among which Monte Carlo Markov Chains (MCMC)
- Bayesian inferences (often coupled with MCMCs)
- Maximum likelihood (or Bayesian inference) with estimation of the likelihood by simulations (e.g. coalescent)
- Hidden Markov Models along the genome (HMM)
- **Simulation-based** inference methods using sumary statistics
- and more recently using **artificial intelligence AI** : machine learning, deep learning, neural networks, ...

→ allow inferences of all parameters of more realistic models (thanks also to the increase of genetic information)

Indirect demographic inferences

1 - Genetic data carry information about evolutionary (demographic?) parameters

2 - First historical developments of indirect demographic inference and their limits

3 - Are these limitations a real barrier to indirect demographic inference

4 - Introduction to spatial models in population genetics : Isolation By Distance (IBD)

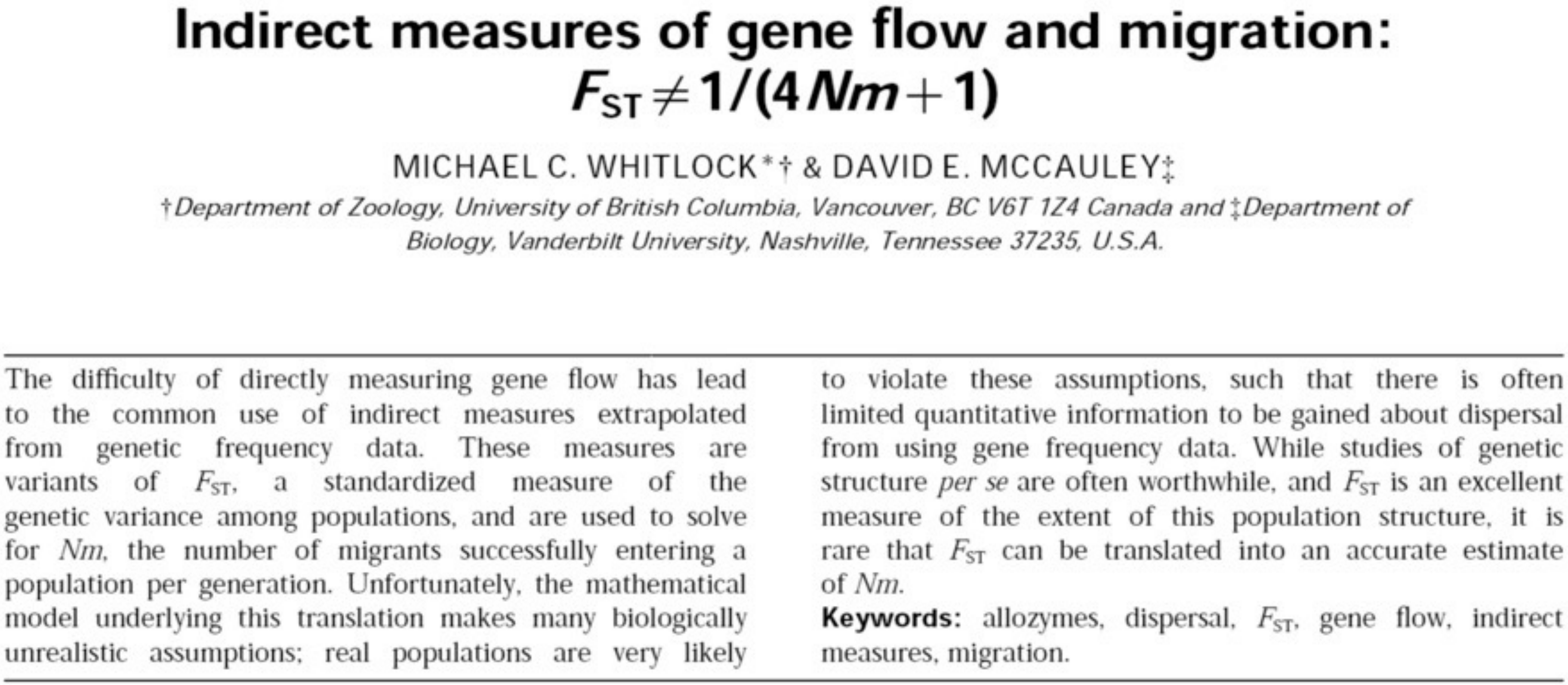
5 - Historical developments to infer demographic parameters under IBD

6 - IBD : relevant models for local demographic inferences

Discussion…

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) :



Indirect measures of gene flow and migration: $F_{ST} \neq 1/(4Nm+1)$

MICHAEL C. WHITLOCK*† & DAVID E. MCCAULEY‡

†*Department of Zoology, University of British Columbia, Vancouver, BC V6T 1Z4 Canada and* ‡*Department of Biology, Vanderbilt University, Nashville, Tennessee 37235, U.S.A.*

The difficulty of directly measuring gene flow has lead to the common use of indirect measures extrapolated from genetic frequency data. These measures are variants of F_{ST}, a standardized measure of the genetic variance among populations, and are used to solve for *Nm*, the number of migrants successfully entering a population per generation. Unfortunately, the mathematical model underlying this translation makes many biologically unrealistic assumptions; real populations are very likely to violate these assumptions, such that there is often limited quantitative information to be gained about dispersal from using gene frequency data. While studies of genetic structure *per se* are often worthwhile, and F_{ST} is an excellent measure of the extent of this population structure, it is rare that F_{ST} can be translated into an accurate estimate of *Nm*.

Keywords: allozymes, dispersal, F_{ST}, gene flow, indirect measures, migration.

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

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 $\neq$1/(1+4Nm)

This is still true for studies after the genomic and numeric revolution with more markers and more computers…

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 $\neq$1/(1+4Nm)

This is still true for studies after the genomic and numeric revolution with more markers and more computers… but is it true for all situations/methods/models/species/samples/… ?

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

Indirect demographic inferences

1 - Genetic data carry information about evolutionary (demographic?) parameters

2 - First historical developments of indirect demographic inference and their limits

3 - Are these limitations a real barrier to indirect demographic inference

4 - Introduction to spatial models in population genetics : Isolation By Distance (IBD)

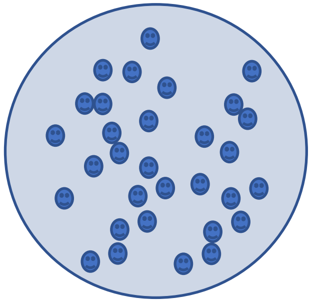
5 - Historical developments to infer demographic parameters under IBD

6 - IBD : relevant models for local demographic inferences

Discussion…

Isolation By Distance (IBD) models

- **Derived from the classical Wright-Fisher model :**



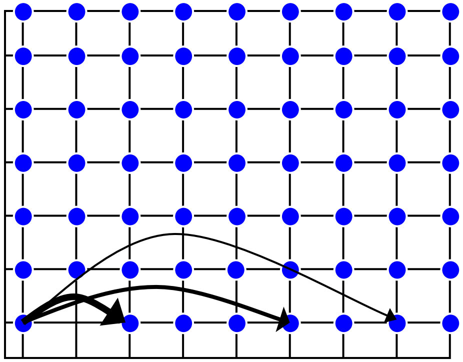
  - **isolated panmictic** population
  - **finite** and constant (relaxable) population size
  - **Non-overlapping generations**
  - **Same** expected **reproductive success** for all individuals (E[offspring nbr per adult] = 1)

- **But with a spatial population structure and (potentialy) limited dispersal :**

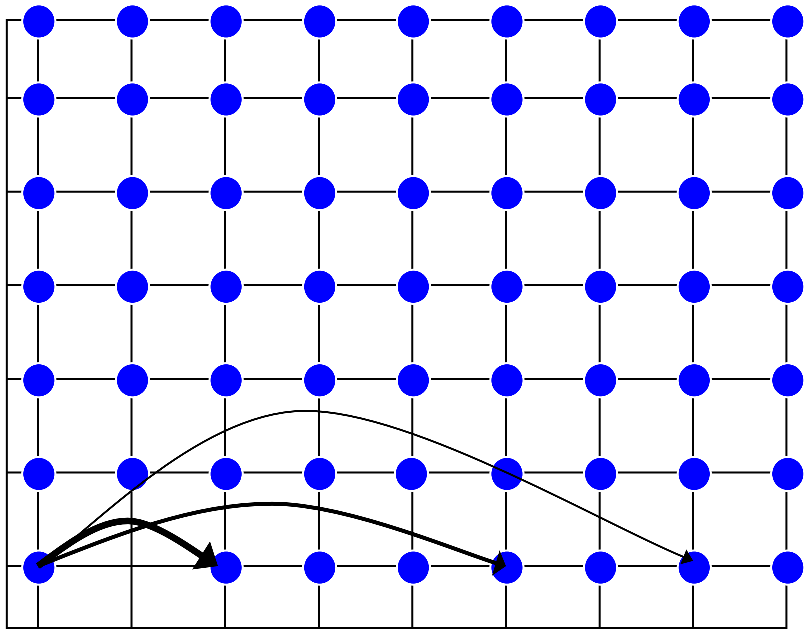
  - **finite** and constant (relaxable) population sizes
  - **Non-overlapping generations**
  - **Same** expected **reproductive success** for all individuals E(offspring nbr per adult) = 1

  - **set of panmictic sub-populations** (patchy habitat)

    or **individuals/couples** (continuous habitat)
  - **homogeneously distributed** over the habitat (on a lattice)
  - **spatially limited dispersal** (dispersal distribution)
  - but **isolated** from other populations



Isolation By Distance (IBD) models

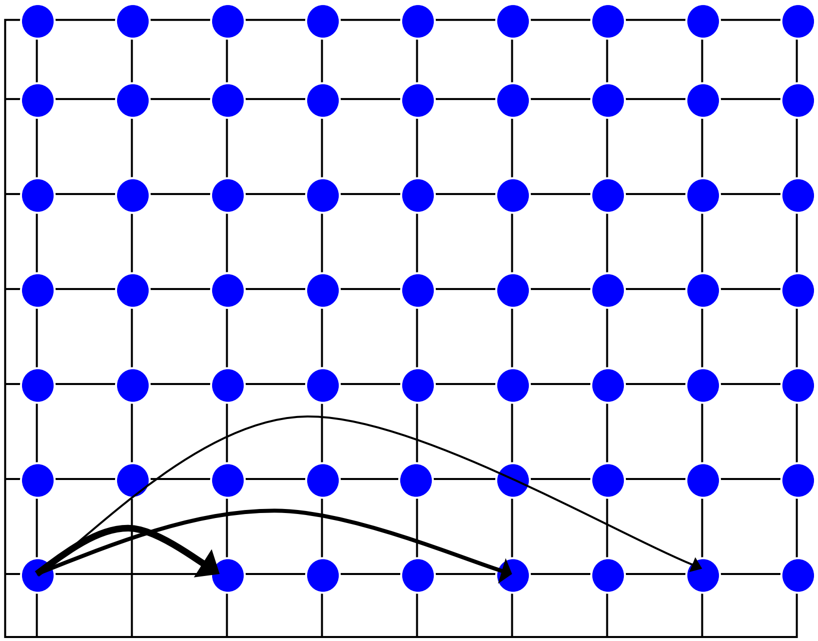


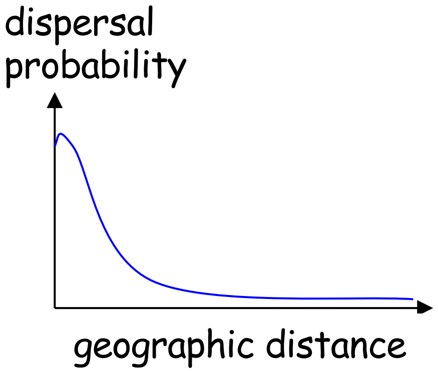
Based on the simple property that dispersal between generations (Parent-Offspring dispersal) is localized in space i.e., **2 individuals are more likely to be close relatives if they live geographically close to each other**

Endler (1977) first showed in a review that

the vast majority of species has geographically localized dispersal

Isolation By Distance (IBD) models

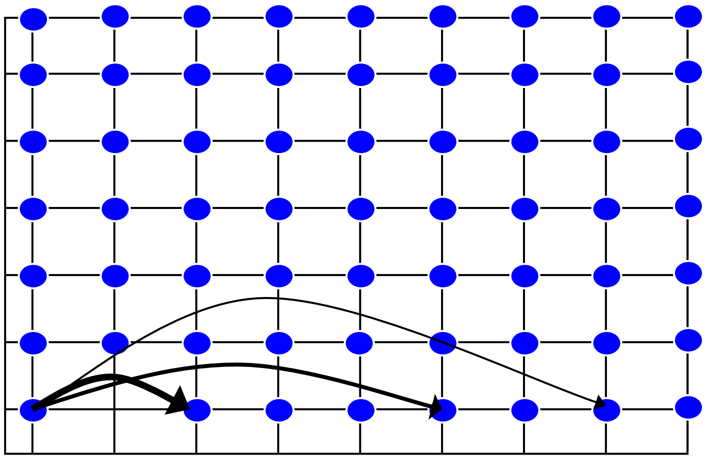




the parent-offspring dispersal (migration) rate over the habitat is decreasing function of the geographic distance, modelled through a dispersal distribution

Isolation By Distance (IBD) models

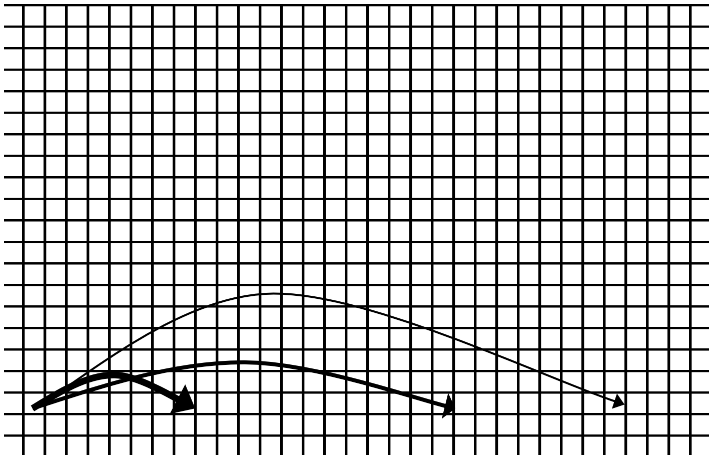
2 variants of IBD models depending on individual spatial distribution in the landscape, which general depends on the repartition of suitable habitats in the landscape



Patchy favorable habitat or population clusters

IBD between demes

each node of the lattice corresponds to a panmictic sub-population (deme) of size *N* individuals

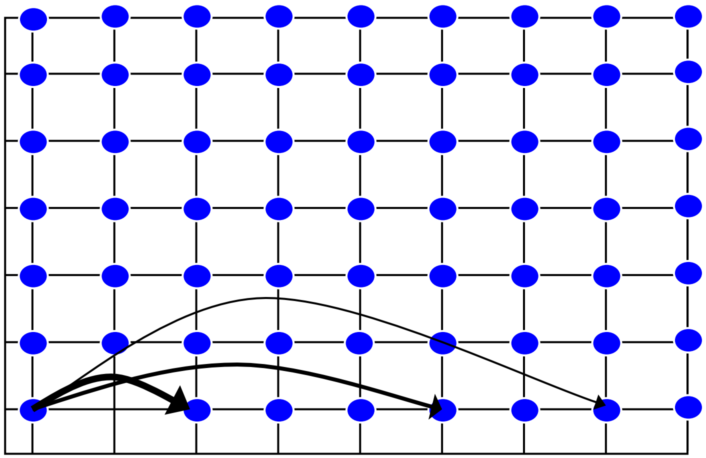


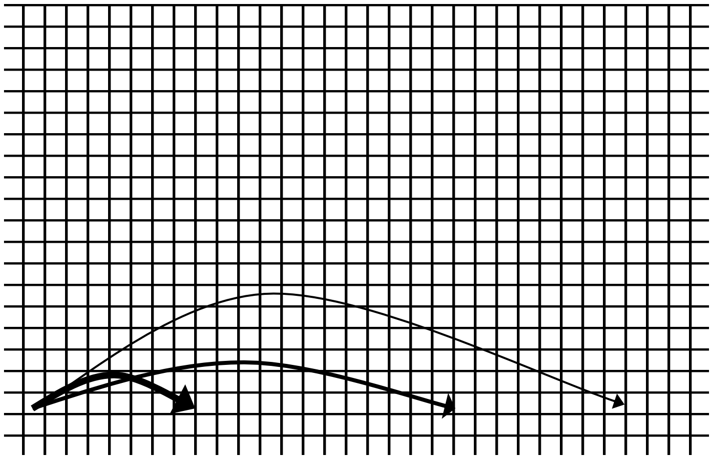
Continuous habitat

IBD between individuals

each node of the lattice corresponds to a single individual (N=1) or a couple (N=2)

Isolation By Distance (IBD) models





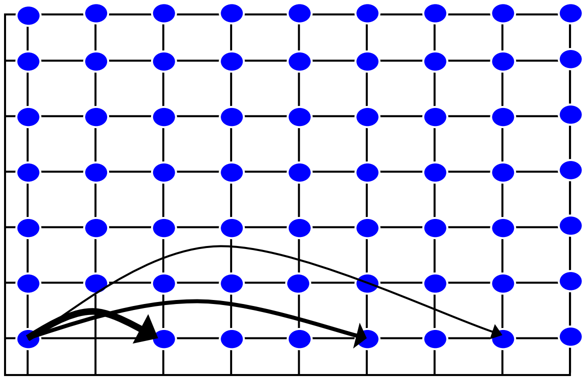
Fully homogeneous model :

Same deme size / density of individuals over the lattice

Same dispersal distribution for all lattice nodes

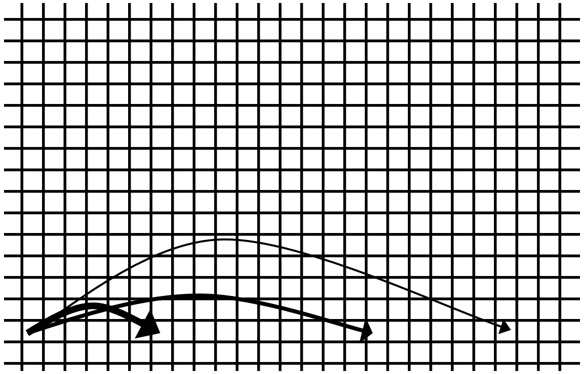
…but can be relaxed if we want to consider spatially (and temporally) heterogeneous IBD models…

Isolation By Distance (IBD) models



Fully homogeneous model

implies few parameters:



Canonical parameters :

Lattice size: n_x (n_y), sometimes infinite

Deme size: N

Migration rate : m

Dispersal distribution: any (e.g. geometric)

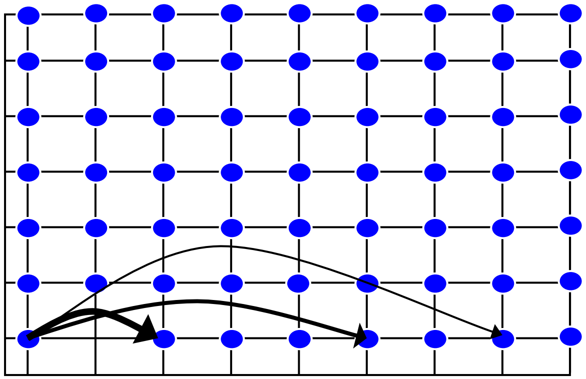
Dispersal shape: 1 to 3 parameters (e.g. g_{geom})

Lattice Unit (= mesh length) : L

Mutation model = any

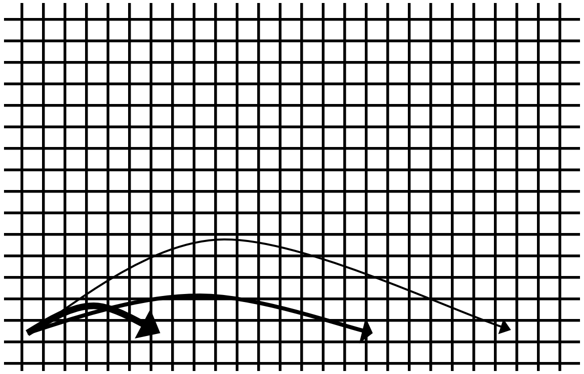
Mutation rate = μ

Isolation By Distance (IBD) models



Fully homogeneous model

implies few parameters:



Canonical parameters :

Lattice size: n_x (n_y), sometimes infinite

Deme size: N

Migration rate : m

Dispersal distribution: any (e.g. geometric)

Dispersal shape: 1 to 3 parameters (e.g. g_{geom})

Lattice Unit (= mesh length) : L

Mutation model = any

Mutation rate = μ

Composite parameters :

σ^2 = mean square parent-offspring distance

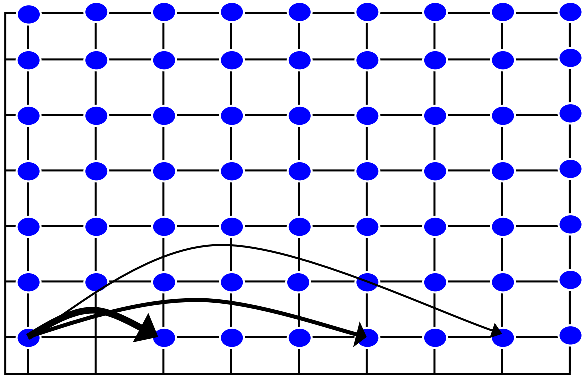
= $m(1+\mathrm{g})/(1-\mathrm{g})^2$ for geometric dispersal

$D\sigma^2$ $(N\sigma^2)$ or $2 * ploidy * \pi D\sigma^2$

$4\pi D\sigma^2$ $(4\pi N\sigma^2)$ is classically called the "**neighborhood size**"

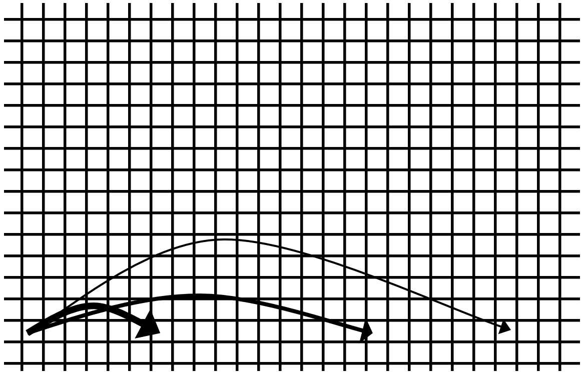
$(\mathbf{4\pi})\boldsymbol{D\sigma^2}$ [or $(\mathbf{4\pi})\boldsymbol{N\sigma^2}$] is the **inverse of the strength the isolation by distance pattern**

Isolation By Distance (IBD) models



Fully homogeneous model

implies few parameters:



Canonical parameters :

Lattice size: n_x (n_y), sometimes infinite

Deme size: *N*

Migration rate : *m*

Dispersal distribution: any (e.g. geometric)

Dispersal shape: 1 to 3 parameters (e.g. g_{geom})

Lattice Unit (= mesh length) : *L*

Mutation model = any

Mutation rate = μ

Composite parameters :

σ^2 = mean square parent-offspring distance

= $m(1+\mathrm{g})/(1-\mathrm{g})^2$ for geometric dispersal

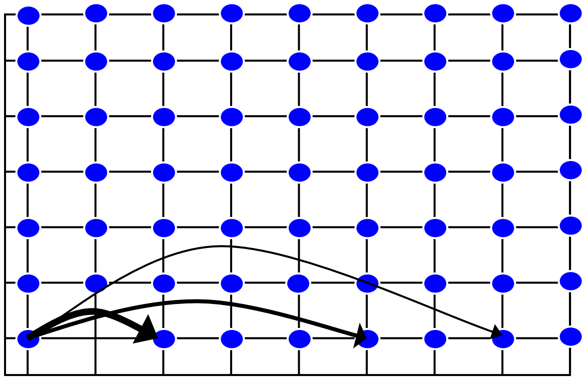
$D\sigma^2$ ($N\sigma^2$) or $2 * ploidy * \pi D\sigma^2$

$4\pi D\sigma^2$ ($4\pi N\sigma^2$) is classically called the "**neighborhood size**"

In 2D, $D\sigma^2$ is a number of individuals, and σ^2 can be expressed (and interpreted) in “mean inter-individual distance” unit (e.g. D=1)

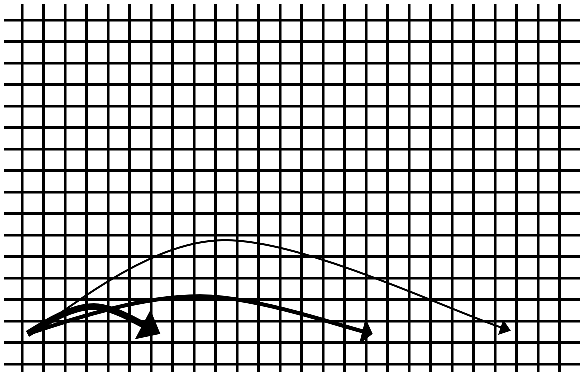
$(\boldsymbol{4\pi})\boldsymbol{D\sigma^2}$ [or $(\boldsymbol{4\pi})\boldsymbol{N\sigma^2}$] is the **inverse of the strength the isolation by distance pattern**

Isolation By Distance (IBD) models



Fully homogeneous model

implies few parameters:



Canonical parameters :

Lattice size: n_x (n_y), sometimes infinite

Deme size: *N*

Migration rate : *m*

Dispersal distribution: any (e.g. geometric)

Dispersal shape: 1 to 3 parameters (e.g. g_{geom})

Lattice Unit (= mesh length) : *L*

Mutation model = any

Mutation rate = μ

Composite parameters :

σ^2 = mean square parent-offspring distance

= $m(1+\mathrm{g})/(1-\mathrm{g})^2$ for geometric dispersal

$D\sigma^2$ $(N\sigma^2)$ or $2 * ploidy * \pi D\sigma^2$

$4\pi D\sigma^2$ $(4\pi N\sigma^2)$ is classically called the "**neighborhood size**"

$\theta_{d(eme)}$ = $2 * ploidy * N\mu$

$\theta_{g(lobal)}$ = $2 * ploidy * n_x * n_y * N\mu$

2Nm

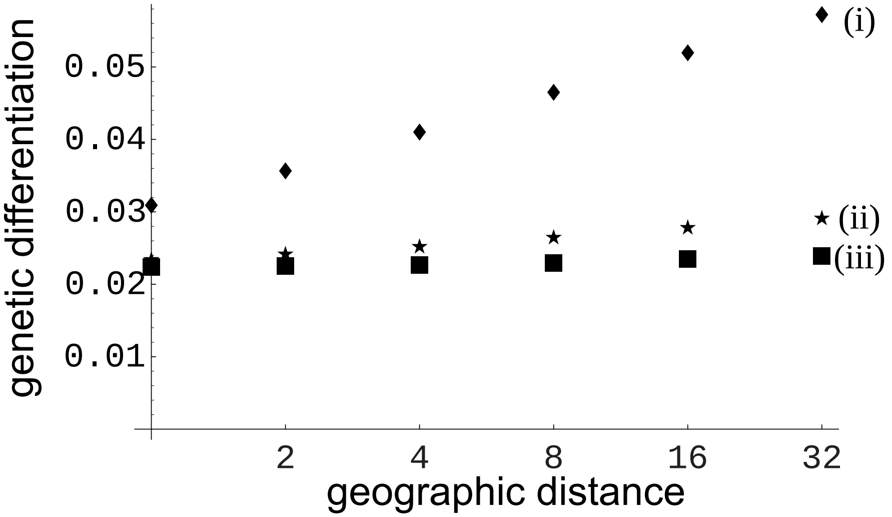
Density D = N/L^2

$(\boldsymbol{4\pi})\boldsymbol{D\sigma^2}$ [or $(\boldsymbol{4\pi})\boldsymbol{N\sigma^2}$] is the **inverse of the strength the isolation by distance pattern**

Isolation By Distance (IBD) models

One of the main characteristic of IBD models is that

genetic differentiation increases with geographic distance



Strong IBD (**small Dσ^2**)

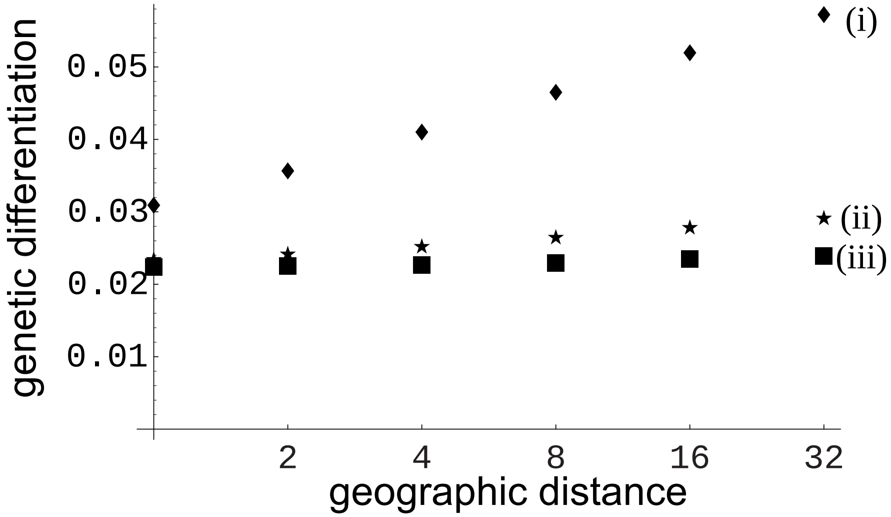
weak IBD (**large Dσ^2**)

Island model, no IBD (**Dσ^2 = ∞**)

Isolation By Distance (IBD) models

One of the main characteristic of IBD models is that

genetic differentiation increases with geographic distance



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

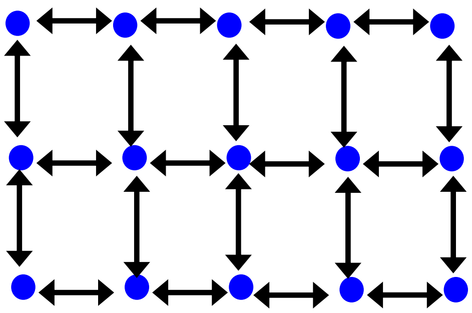
weak IBD (**large $D\sigma^2$**)

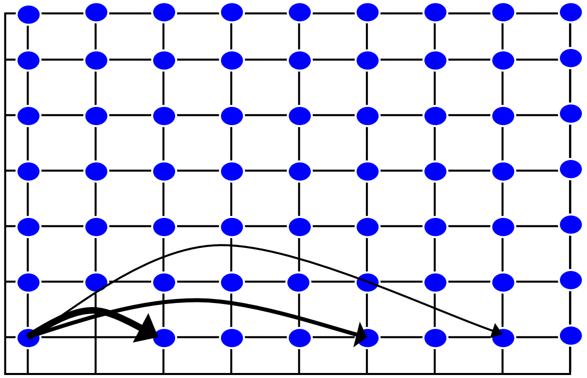
Island model, no IBD (**$D\sigma^2$ = ∞**)

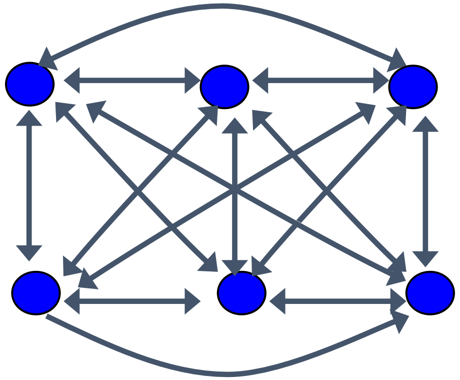
-> Mantel tests are used to test the presence of IBD

= the correlation between genetic and geographic distances

Isolation By Distance (IBD) models





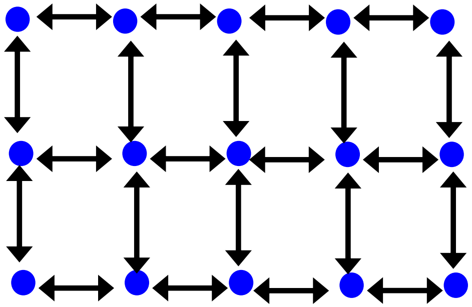


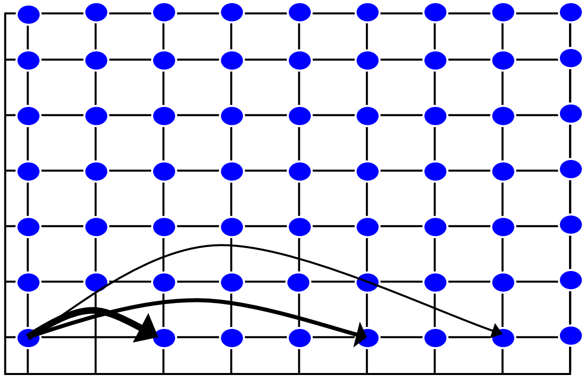
IBD models, which include the stepping stone and the island model as "limit cases", are quite general depending on how localized dispersal is :

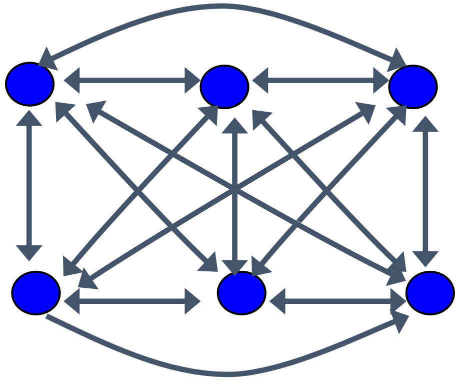
Stepping stone **>** **IBD** **>** **Island Model**

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

Isolation By Distance (IBD) models







IBD models, which include the stepping stone and the island model as “limit cases”, are quite general depending on how localized dispersal is :

Stepping stone > IBD > Island Model

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

Geometric
dispersal -> g_geom=0.0 g_geom=0.x g_geom=1.0

Indirect demographic inferences

1 - Genetic data carry information about evolutionary (demographic?) parameters

2 - First historical developments of indirect demographic inference and their limits

3 - Are these limitations a real barrier to indirect demographic inference

4 - Introduction to spatial models in population genetics : Isolation By Distance (IBD)

5 - Historical developments to infer demographic parameters under IBD

6 - IBD : relevant models for local demographic inferences

Discussion…

Demographic inferences under IBD

Historical developments :

- Wright 1943 : the idea of limited parent-offspring dispersal among homogeneously distributed individuals or sub-populations (misleading “Neighborhood size”)
- 1950-1980 : test of positive correlation between various measures of genetic differentiation and geographic distance
- 1980-1997 : Mantel tests + regression differentiation vs distance (Slatkin 1993) but not a good inference method (only valid to infer 2*Nm* under a stepping-stone dispersal model)
- Rousset 1997 : Mantel Test + regression $\frac{F_{st}}{1-F_{st}}$ vs log(distance)

Demographic inferences under IBD

Rousset 1997 main theoretical result :

mathematical analysis of **IBD models with demes** (in terms of probabilities of identity) is the following **linear relationship between** the **differentiation** parameter and the **geographic distance** and the different assumptions leading to it :

$$\frac{Fst}{1-Fst} \equiv \frac{Q_0 - Q_r}{1-Q_0} \approx \frac{\ln(r)}{4\pi N\sigma^2} + constant$$

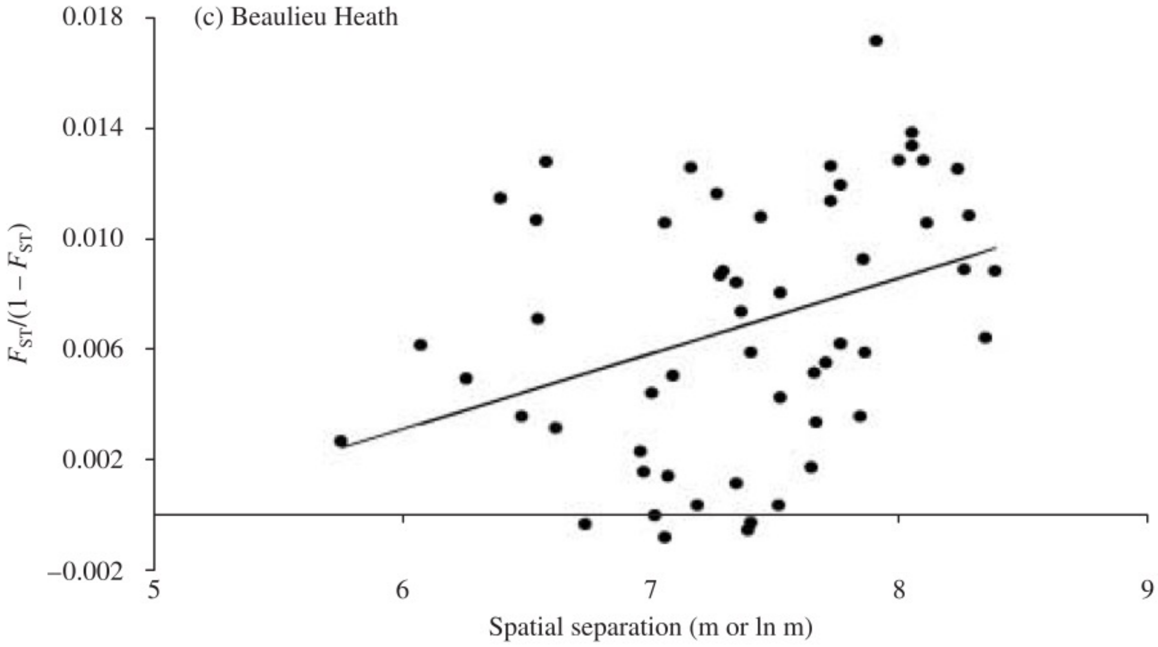
Linear relationship between differentiation and ln(geog. distance) in 2 dimension IBD

only valid at a small geographical scale (10 - 100 σ^2) and for low mutation rates

Demographic inferences under IBD

Rousset 1997 main practical result : The regression method

The regression slope is expected to be 1/ $4\pi N\sigma^2$, thus a simple method to infer $N\sigma^2$ is to do the regression on the data and estimate the slope



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

Demographic inferences under IBD

Historical developments :

- Wright 1943 : the idea of limited dispersal among homogeneously distributed individuals or populations (misleading “Neighborhood size”)
- 1950-1980 : test of positive correlation between various measures of genetic differentiation and geographic distance
- 1980-1997 : Mantel tests + regression differentiation vs distance (Slatkin 1993) but not a good inference method (only for stepping-stone dispersal)
- Rousset 1997 : Mantel Test + regression $\frac{F_{st}}{1-F_{st}}$ vs log(distance)

-> first method **to infer $D\sigma^2$ under IBD with demes**

Demographic inferences under IBD

Historical developments :

- Wright 1943 : the idea of limited dispersal among homogeneously distributed individuals or populations (misleading “Neighborhood size”)

- 1950-1980 : test of positive correlation between various measures of genetic differentiation and geographic distance

- 1980-1997 : Mantel tests + regression differentiation vs distance (Slatkin 1993) but not a good inference method (only for stepping-stone dispersal)

- Rousset 1997 : Mantel Test + regression $\frac{F_{st}}{1-F_{st}}$ vs log(distance)

  -> first method **to infer $\boldsymbol{D\sigma^2}$ under IBD with demes**

- Rousset 2000 : extension of the regression method to analyse the **differentiation between individuals living in a continuous habitat**

Demographic inferences under IBD

Extension of Rousset's (1997) results to analyse the **differentiation between individuals living in a continuous habitat** (no panmictic sub-populations, N=1 individual or a couple)

Definition of $\mathrm{a_r}$ (an equivalent of $\frac{F_{st}}{1-F_{st}}$) to compute the differentiation between individuals (and not demes)

$$a_r \equiv \frac{Q_0 - Q_r}{1 - Q_0} \approx \frac{\ln(r)}{4\pi D\sigma^2} + constant$$

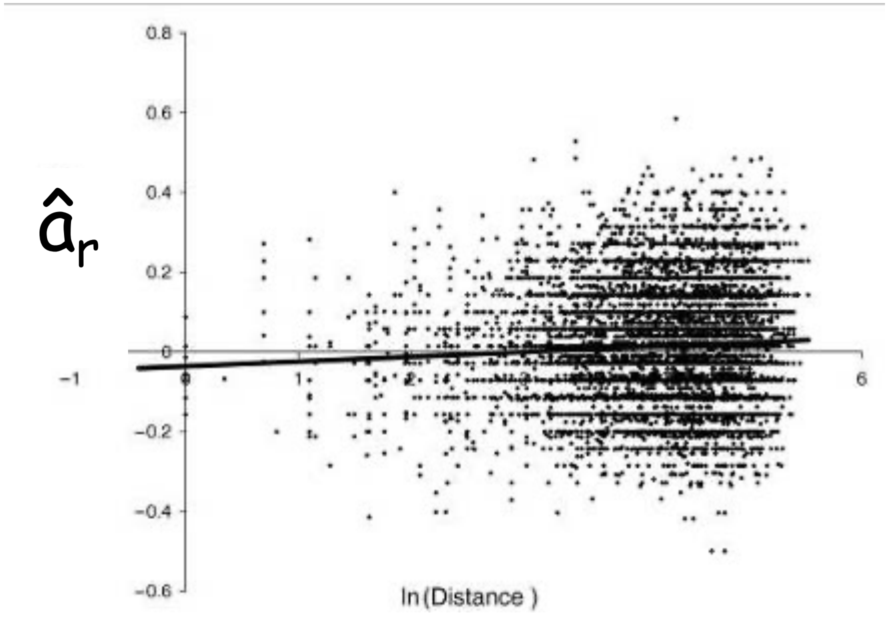
Linear relationship between differentiation and ln(geog. distance) in 2 dimensional IBD

Only valid at a small geographical scale (10 - 100 σ^2) and for low mutation rates.

Demographic inferences under IBD

Rousset 2000 main practical result : The regression method between individuals

The regression slope is expected to be 1/ $4\pi D\sigma^2$, 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 $D\sigma^2$**

Demographic inferences under IBD

Historical developments :

- Wright 1943 : the idea of limited dispersal among homogeneously distributed individuals or populations (misleading “Neighborhood size”)
- 1950-1980 : test of positive correlation between various measures of genetic differentiation and geographic distance
- 1980-1997 : Mantel tests + regression differentiation vs distance (Slatkin 1993) but not a good inference method (only for stepping-stone dispersal)
- Rousset 1997 : regression $\frac{F_{st}}{1-F_{st}}$ vs log(distance)

-> first method **to infer $\boldsymbol{D\sigma^2}$ under IBD with demes**

- Rousset 2000 : regression a_r vs log(distance) for a continuous habitat

-> Inference of $\boldsymbol{D\sigma^2}$ **under IBD between individuals in a continuous habitat**

Demographic inferences under IBD

Historical developments :

- Wright 1943 : the idea of limited dispersal among homogeneously distributed individuals or populations (misleading “Neighborhood size”)

- …

- Rousset 1997 : regression $\frac{F_{st}}{1-F_{st}}$ vs log(distance)

  -> first method **to infer $\boldsymbol{D\sigma^2}$ under IBD with demes**

- Rousset 2000 : regression a_r vs log(distance) for a continuous habitat

  -> Inference of **$\boldsymbol{D\sigma^2}$ under IBD between individuals in a continuous habitat**

  Both between-individual and between-demes regression methods have been **extensively used** (Rousset 1997: 2800 citations, Rousset 2000: 500 citations)

  but most applications only considered the result of the mantel test to show a significant (or not) IBD signal and do not use the slope to infer **$\boldsymbol{D\sigma^2}$ ….**

Demographic inferences under IBD

Historical developments :

- Wright 1943 : the idea of limited dispersal among homogeneously distributed individuals or populations (misleading "Neighborhood size")

- …

- Rousset 1997 : regression $\frac{F_{st}}{1-F_{st}}$ vs log(distance)

  -> first method **to infer $\boldsymbol{D\sigma^2}$ under IBD with demes**

- Rousset 2000 : regression a_r vs log(distance) for a continuous habitat

  -> Inference of **$\boldsymbol{D\sigma^2}$ under IBD between individuals in a continuous habitat**

- Leblois et al 2003, 2004 : tests of the performance of the regression method to estimate $\boldsymbol{D\sigma^2}$

Demographic inferences under IBD

Simulation tests of the regression method between individuals in a continuous habitat (Rousset, 2000)

- Development of **IBDSim** a genetic data simulator under IBD
  - “exact” coalescence algorithm (backward generation-by-generation)
  - flexible potentially heterogeneous in space and time IBD models

- Test of expected precision and robustness of the estimation of $D\sigma^2$ from a classical microsatellite data set (10x10 individuals genotyped at 10 loci)

  - Good precision (bias<20%, RMSE<30%, >95% estimates within a factor 2)

  - Robust to recent installation/expansion : IBD patterns establish quickly

  - Robust to recent (>20 generations) and moderate (10-20X) changes in density and dispersal

➔ **1/slope sems to be a robust estimator of local and present-time Dσ^2**

Demographic inferences under IBD

Many applications , e.g. :

- marginated tortoise $D\sigma^2$ = 6 – 10 (individual-based IBD)
- marbled newt $D\sigma^2$ = 5.5 – 45 depending on ponds density (demic IBD)
- greater horseshoe bat $D\sigma^2$ = 20 - 32 (individual-based IBD)
- pollen beetle $D\sigma^2$ = 50 – 100 (large scale demic IBD)

Some of them giving “unexpected” results

- the house mouse within Senegalese villages $D\sigma^2$ = 5.0 – 7.4 (demic IBD)
- Procesionnary moth $D\sigma^2$ =0.4 - 1.5 (individual and demic IBD)

But without expectation on the “real” $D\sigma^2$, we can not say much more than that…

Indirect demographic inferences

1 - Genetic data carry information about evolutionary (demographic?) parameters

2 - First historical developments of indirect demographic inference and their limits

3 - Are these limitations a real barrier to indirect demographic inference

4 - Introduction to spatial models in population genetics : Isolation By Distance (IBD)

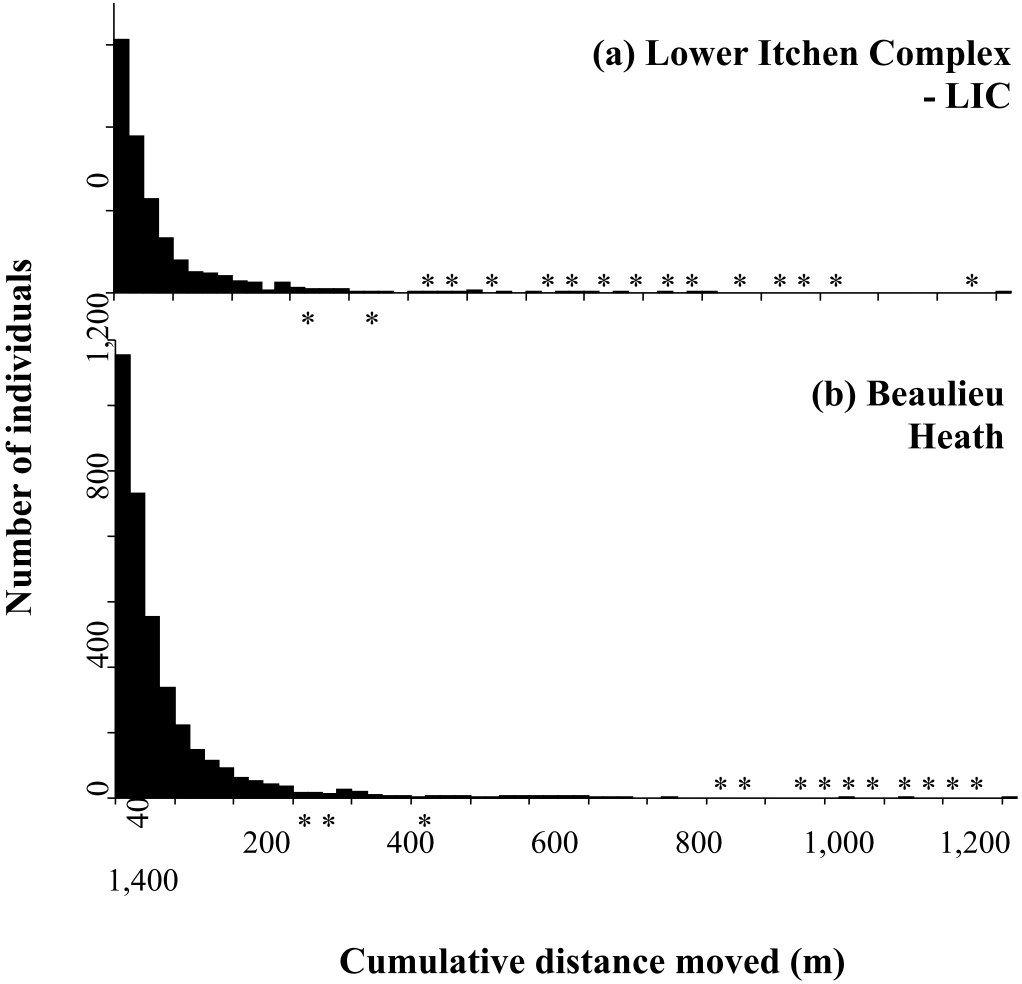
5 - Historical developments to infer demographic parameters under IBD

6 - IBD : relevant models for local demographic inferences

Discussion…

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

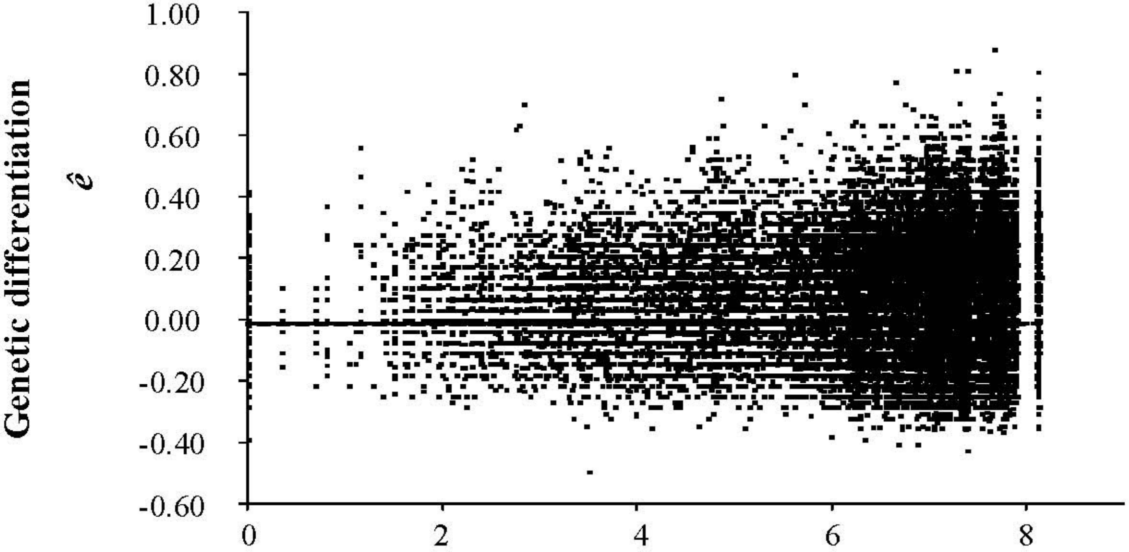
Comparisons between genetic and demographic estimates

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

Genetic data : 700 individuals genotyped

at 13 microsatellite loci

➠ indirect estimates of $\boldsymbol{D\sigma^2}$



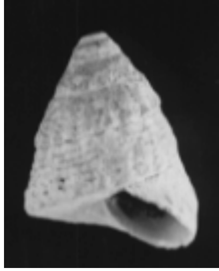
Comparisons between genetic and demographic estimates

- 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

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

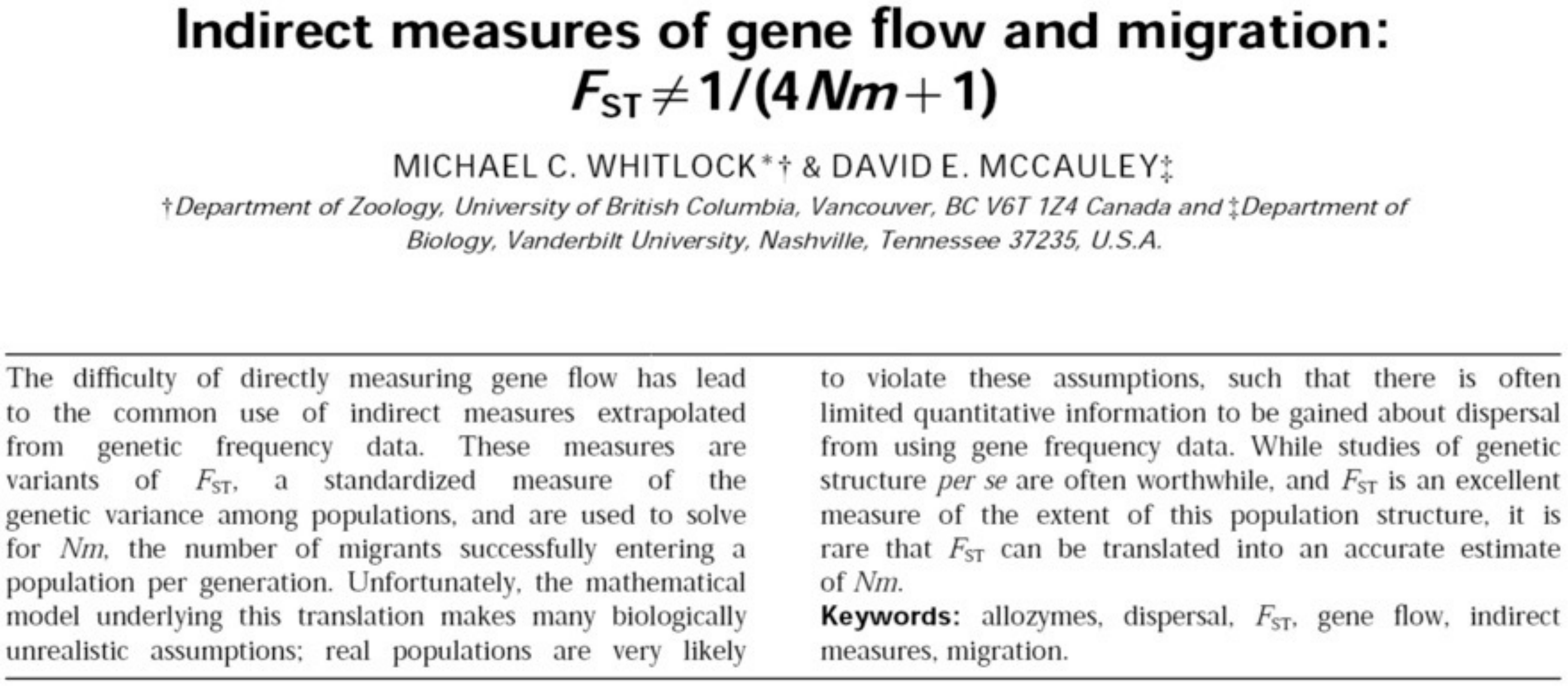
demographic and genetic data at a local geographical scale

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

IBD seems to be relevant models for the inference of demographic parameters at small geographic and temporal scale

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) :



Indirect measures of gene flow and migration: $F_{ST} \neq 1/(4Nm+1)$

MICHAEL C. WHITLOCK*† & DAVID E. MCCAULEY‡

†Department of Zoology, University of British Columbia, Vancouver, BC V6T 1Z4 Canada and ‡Department of Biology, Vanderbilt University, Nashville, Tennessee 37235, U.S.A.

The difficulty of directly measuring gene flow has lead to the common use of indirect measures extrapolated from genetic frequency data. These measures are variants of F_{ST}, a standardized measure of the genetic variance among populations, and are used to solve for *Nm*, the number of migrants successfully entering a population per generation. Unfortunately, the mathematical model underlying this translation makes many biologically unrealistic assumptions; real populations are very likely to violate these assumptions, such that there is often limited quantitative information to be gained about dispersal from using gene frequency data. While studies of genetic structure *per se* are often worthwhile, and F_{ST} is an excellent measure of the extent of this population structure, it is rare that F_{ST} can be translated into an accurate estimate of *Nm*.

Keywords: allozymes, dispersal, F_{ST}, gene flow, indirect measures, migration.

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 $\neq$1/(1+4Nm)

So why do we have good results for *Dσ^2* inferences using the regression method on IBD models ?

Why *Dσ²* 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)
- ”pseudo-continuous" lattice 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 *Dσ²* 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*σ^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 ”pseudo-continuous" lattice model because deme size is one individual and migration rates are very high (>0.3)

Demographic inferences under IBD

Historical developments :

- Wright 1943 : the idea of limited dispersal among homogeneously distributed individuals or populations (misleading “Neighborhood size”)
- …
- Rousset 1997 : regression $\frac{F_{st}}{1-F_{st}}$ vs log(distance)

  -> first method **to infer $\boldsymbol{D\sigma^2}$ under IBD with demes**

- Rousset 2000 : regression a_r vs log(distance) for a continuous habitat

  -> Inference of $\boldsymbol{D\sigma^2}$ **under IBD between individuals in a continuous habitat**

IBD seems to be relevant models for the inference of demographic parameters at small geographic and temporal scale

Since 2000, many developements in landscape/ statistical spatial population genetics
Mostly visualization/correlation tools but not much on demograhic parameter inference

e.g. Mapi (Piry et al. 2016), EEMS (Petkova et al. 2015, Al-Asadi et al. 2019) and many others…

Indirect demographic inferences

1 - Genetic data carry information about evolutionary (demographic?) parameters

2 - First historical developments of indirect demographic inference and their limits

3 - Are these limitations a real barrier to indirect demographic inference

4 - Introduction to spatial models in population genetics : Isolation By Distance (IBD)

5 - Historical developments to infer demographic parameters under IBD

6 - IBD : relevant models for local demographic inferences

7 – Our work to go further than the regression method

Discussion…

Demographic inferences under IBD

Historical developments :

- Wright 1943 : the idea of limited dispersal among homogeneously distributed individuals or populations (misleading “Neighborhood size”)
- …
- Rousset 1997 : regression $\frac{F_{st}}{1-F_{st}}$ vs log(distance)

  -> first method **to infer $D\sigma^2$ under IBD with demes**

- Rousset 2000 : regression a_r vs log(distance) for a continuous habitat

  -> Inference of $D\sigma^2$ **under IBD between individuals in a continuous habitat**

- Rousset & Leblois 2007 and 2011 : Coalescence-based maximum likelihood inferences under IBD (coalescent approx.) in 1D and 2D
  - ML ideal statistical framework : takes all the information carried by the genetic data (many developments, eg. MCMC coa-based 1995-2010)
  - Adaptation of the Importance Sampling algorithms of Griffiths et al. implemented in the software MIGRAINE

What's in the Migraine software?

C++ core SIS computations

Point sampling, Likelihood estimations, Write R code, launch R analysis

R (automated interaction between C++, R code and R package 'blackbox')

Likelihood surface interpolation, MLEs and CIs, Plots, next points

Migraine can automatically run iterative analysis by considering a sequence of (C++, R) computations.

This procedure allows to obtain better inferences by maximizing the number points in the good zone of the parameter space.

Demographic models implemented in Migraine: IBD

Linear or planar isolation by distance (IBD) models (**Eq.**)

- Fully homogeneous IBD model → four parameters $(+\ \mu)$:
  - $*$ *d*: nb of subpopulations (usually larger than nb of sampled subpop)
  - $*$ *N*: subpop size (nb of genes, $N_T = d \times N$)
  - $*$ *m*: the emigration rates from any subpopulation
  - $*$ *g*: shape of the geometric dispersal distribution
    ($g = 0$ → Stepping stone; $g = 1$ → Island)

- Availlable mutation models : KAM

- **Inference of 3 scaled parameters**:
  - $*$ $\theta = 2N\mu$
  - $*$ $M = 2Nm$
  - $*$ g
  - $+$ one composite parameter: the neighborhood size $Nb = 4\pi D\sigma^2$

Isolation by distance: Parameters

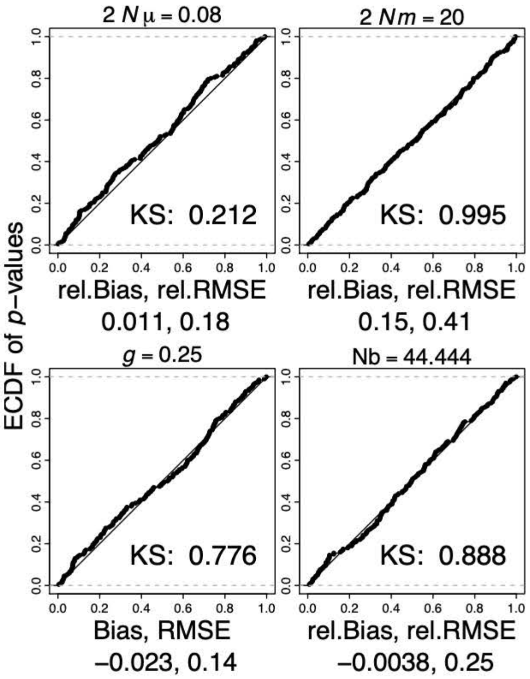
Deme size N, dispersal probability m, mutation probability μ distribution of dispersal distance: geometric decrease with distance, with scale parameter g.

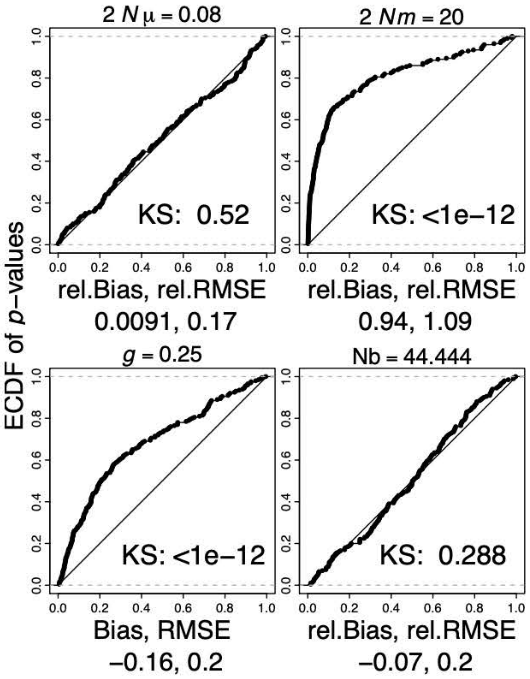
special interest in the neighborhood size $\propto D\sigma^2$ where D is population density and σ^2 is second moment of dispersal distance

Likelihoods computed under the classical limit $N \to \infty$, $\mu \to 0$ for given $N\mu$; and likewise $m \to 0$ for given Nm (“diffusion limit”)

Results under ideal conditions: validating the whole inference process and finding limits...

N: 40000 → 40; m: 0.00025 → 0.25; μ: $10^{-6} \to 10^{-3}$





Diffusion approximations ($N \to \infty$, $\mu \to 0$; $m \to 0$)
→ bias in Nm estimation increases with m

Results under ideal conditions: validating the whole inference process and finding limits...

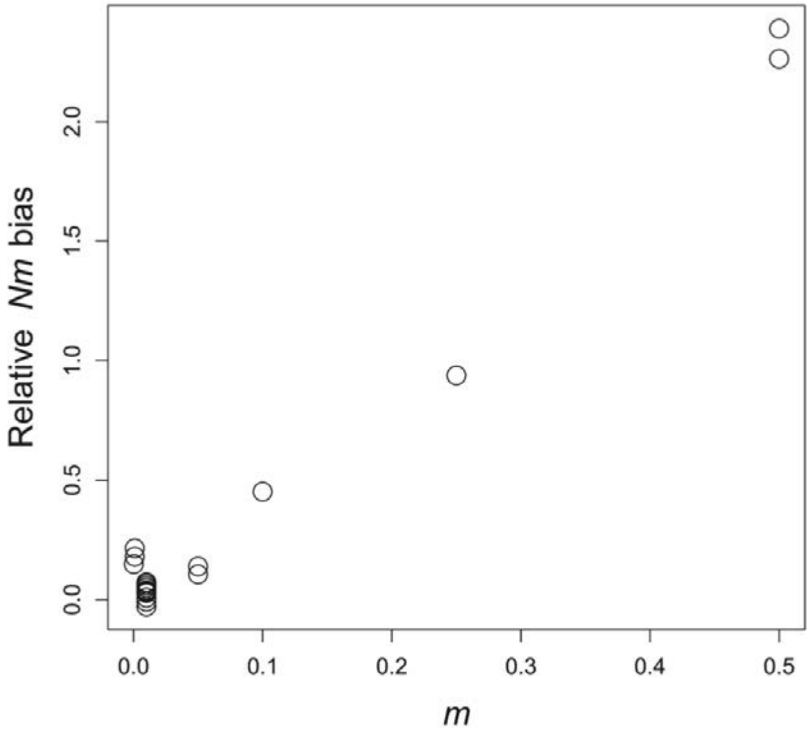
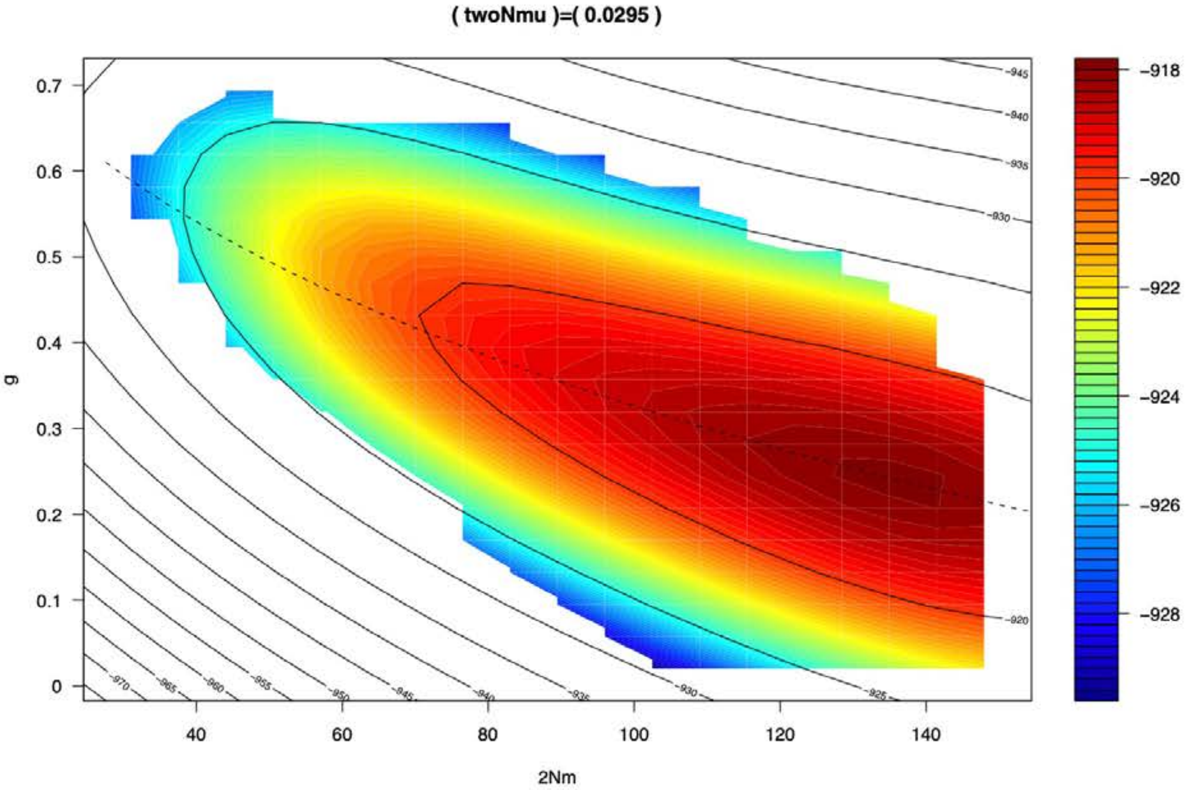


FIG. 4. Relationship between dispersal probability and bias of estimated number of migrants for all cases in table 1.

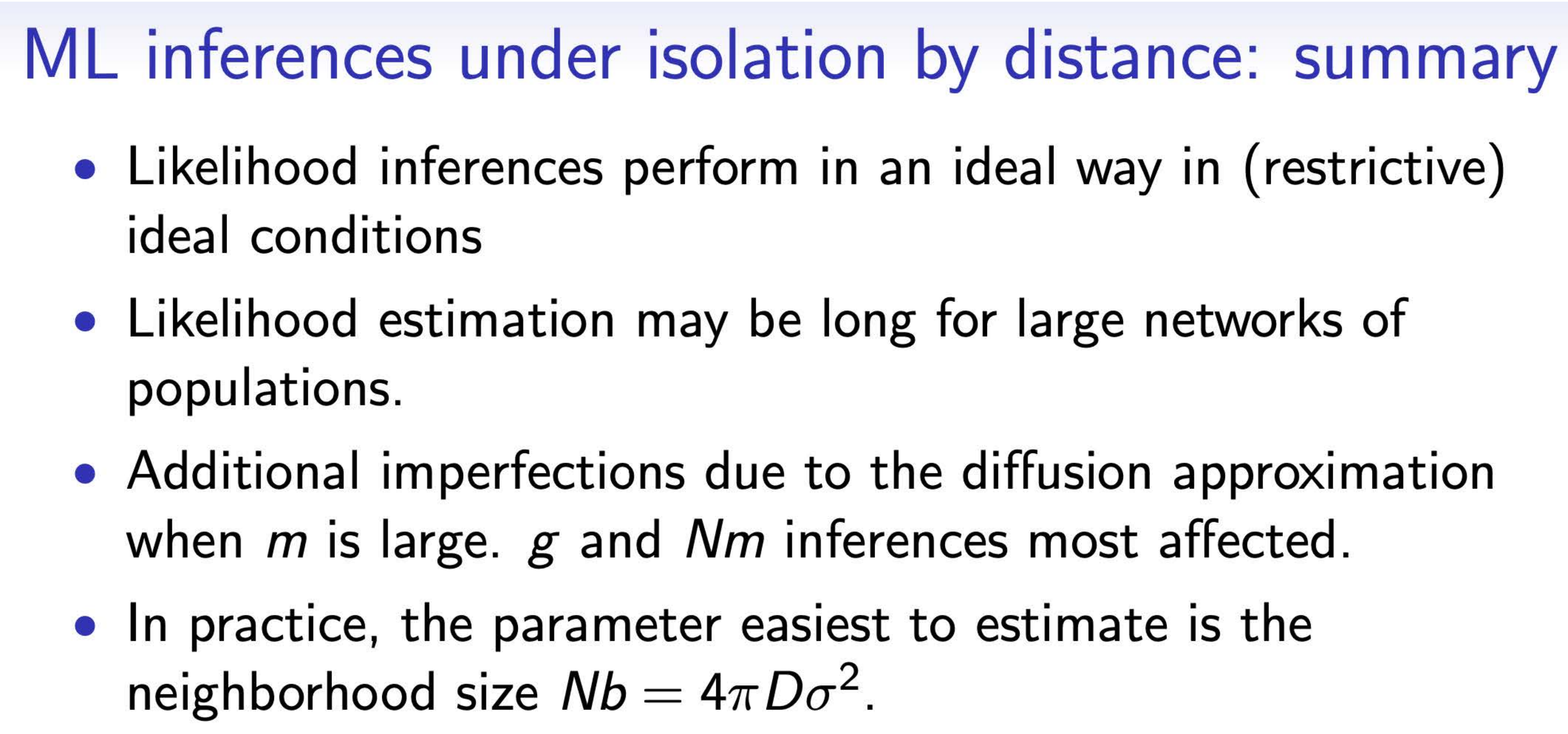
Diffusion approximations ($N \to \infty$, $\mu \to 0$; $m \to 0$)
→ bias in *Nm* estimation increases with *m*

Results under ideal conditions: another limit du to Nm, g covariance

3d main result: not much information to infer Nm and g separately



Demographic inferences under IBD



ML inferences under isolation by distance: summary

- Likelihood inferences perform in an ideal way in (restrictive) ideal conditions
- Likelihood estimation may be long for large networks of populations.
- Additional imperfections due to the diffusion approximation when m is large. g and Nm inferences most affected.
- In practice, the parameter easiest to estimate is the neighborhood size $Nb = 4\pi D\sigma^2$.

Demographic inferences under IBD

Comparison regression method VS Maximum-Likelihood in MIGRAINE :
only a slight improvement of$\boldsymbol{D\sigma^2}$ estimation…

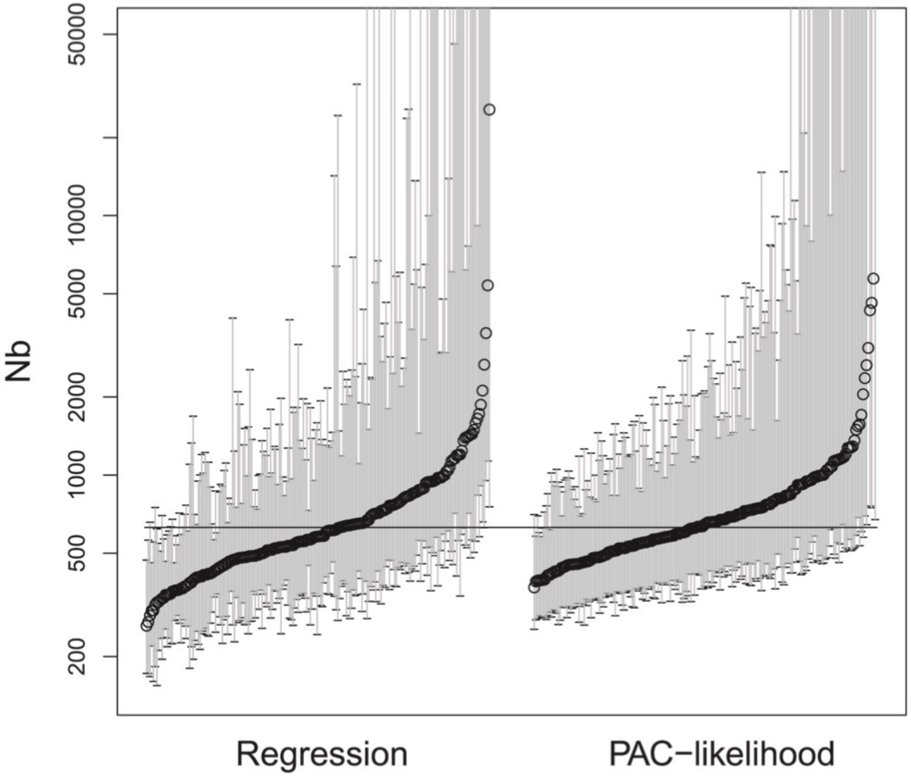


FIG. 7. Distributions of estimates and confidence intervals for Nb, by the spatial regression method and by PAC-likelihood, for case [46]. The horizontal line marks the true parameter value.

Demographic inferences under IBD

Historical developments :

- Wright 1943 : the idea of limited dispersal among homogeneously distributed individuals or populations (misleading “Neighborhood size”)
- …
- Rousset 1997 : regression $\frac{F_{st}}{1-F_{st}}$ vs log(distance)

  -> first method **to infer $\boldsymbol{D\sigma^2}$ under IBD with demes**

- Rousset 2000 : regression a_r vs log(distance) for a continuous habitat

  -> Inference of $\boldsymbol{D\sigma^2}$ **under IBD between individuals in a continuous habitat**

- Rousset & Leblois 2007 and 2011 : Coalescence-based maximum likelihood inferences under IBD (coalescent approx.) in 1D and 2D
  - Inference of $D\sigma^2, \theta_d = 2N_d\mu,$ $\text{and to a lesser extent } 2Nm$ and g
  - but can not deal with IBD between individuals in a continuous habitat, nor with small demes or large migration rates
    - ➢ quite strong practical limits…

Recent developments towards simulation-based inference under IBD

- The regression method is limited to the inference of $D\sigma^2$ only
- Coalescence-based maximum likelihood methods are limited due coalescent approximations and not much flexibility in the models.

- Aim : use the power of simulation-based inference methods (e.g. ABC Approximate Bayesian Computation or similar methods) = Inference can be done under any model from which data can be simulated in reasonable times.

  but need to find good summary statistics that carry information about the parameter of interest

- OK for any IBD model because exact (generation-by-generation) coalescence algorithms allows "fast" simulations :

  - Existing simulator IBDSim (Leblois et al. 2007) but no recombination

  - -> developpement of a new simulator Gspace (PhD T. Virgoulay 2018-2022)
    - More efficient
    - Cleaner code
    - Recombination

Recent developpements on simulation-based inference under IBD

- Aim : use the power of simulation-based inference methods to try to infer all parameters of an IBD model

- Development of a pipeline for such inference and to test the performance of the inferences :

  - two C++ simulators (**IBDSim** / **GSpace**)

  - A C++ library (**GSumstat**) to compute summary statistics on the simulated data sets
    - non-spatial : $N_a, H_e, H_o, F_{st}, F_{is},$
    - spatial : $Q_r, \frac{F_{st}}{1-F_{st}}, a_r$ and e_r regression slope and intercept
    - recomb & spatial : exponential 2D regression of η (Vitalis & Couvet 2001) with geographic and genetic (chromosomal) distance. η = differentiation based on joint probability of identity at 2 loci separated by a given genetic distance between 2 individuals separated by a given geographic distance.

Recent developpements on simulation-based inference under IBD

- Aim : use the power of simulation-based inference methods to try to infer all parameters of an IBD model

- Development of a pipeline for such inference and to test the performance of the inferences :

    - two C++ simulators (**IBDSim** / **GSpace**)

    - A C++ library (**GSumstat**) to compute summary statistics on the simulated data sets

    - A R package (**gspace2infr**) to link the simulators, the summary statistics library and the inference methods (**ABC-RF** and **Infusion**) also designed to facilitate performance tests of the inference

Recent developpements on simulation-based inference under IBD

- Aim : use the power of simulation-based inference methods to try to infer all parameters of an IBD model

- Development of a pipeline for such inference and to test the performance of the inferences :

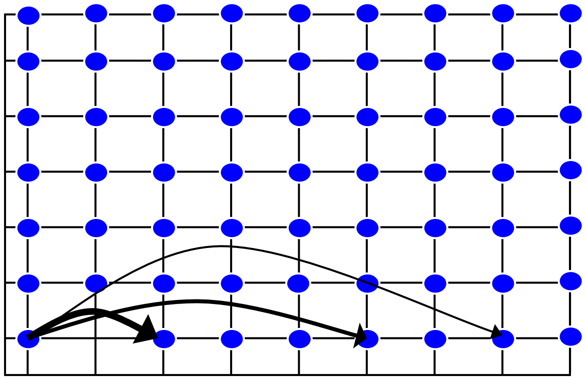
  - two C++ simulators (**IBDSim** / **GSpace**)

  - A C++ library (**GSumstat**) to compute summary statistics on the simulated data sets

  - A R package (**gspace2infr**) to link the simulators, the summary statistics library and the inference methods (**ABC-RF** and **Infusion**)

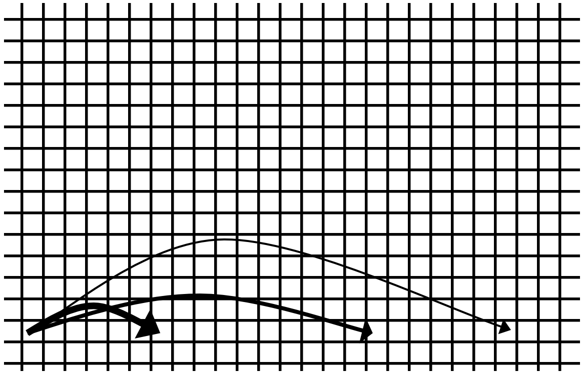
We just got our first encouraging results over the last months !

Isolation By Distance (IBD) models



Fully homogeneous model

implies few parameters:



Canonical parameters :

Lattice size: n_x (n_y), sometimes infinite

Deme size: *N*

Migration rate : *m*

Dispersal distribution: any (e.g. geometric)

Dispersal shape: 1 to 3 parameters (e.g. g_{geom})

Lattice Unit (= mesh length) : *L*

Mutation model = any

Mutation rate = μ

Composite parameters :

σ^2 = mean square parent-offspring distance

= $m(1+\mathrm{g})/(1-\mathrm{g})^2$ for geometric dispersal

$D\sigma^2$ $(N\sigma^2)$ or $2 * ploidy * \pi D\sigma^2$

$4\pi D\sigma^2$ $(4\pi N\sigma^2)$ is classically called the "**neighborhood size**"

$\theta_{d(eme)}$ = $2 * ploidy * N\mu$

$\theta_{g(lobal)}$ = $2 * ploidy * n_x * n_y * N\mu$

2Nm

Density D = N/L^2

$(\boldsymbol{4\pi})\boldsymbol{D\sigma^2}$ [or $(\boldsymbol{4\pi})\boldsymbol{N\sigma^2}$] is the **inverse of the strength the isolation by distance pattern**

Recent developments on simulation-based inference under IBD

- Aim : use the power of simulation-based inference methods to try to infer all parameters of an IBD model

- Our first results for IBDF between individuals in a continuous habitat (1 couple par lattice node, 20 independant microsats or 10 chromosomes with 50 SNPs on each) :

  - Very good inference (bias & var < 1-5%) with a small nbr of markers for :
    - Canonical parameters : *m* , *g,* (*D* with less precision, to be verified)
    - *Composite parameters : θ_d, θ_g,* $D\sigma^2$

  - To be confirmed : some information (order of magnitude) but not precise inference for:
    - Canonical parameters : square_*lattice_size_nx,* μ

Futur developments on simulation-based inference under IBD in the DevOcGen project

- Aim : use the power of simulation-based inference methods under IBD to try to **infer local and present density, dispersal, population sizes**
  but also **their recent changes (e.g. in the last 5?-10?-20-50 generations)**
  - Currently in an stable and homogeneous habitat

Futur developments for the DevOCGen PhD student:
sept 2022-2025, co-funding SPE-INRAe

- Implementation and test of :
  - demographic changes in time (next PhD student)
  - heterogeneous habitat (probably after…)
    e.g. barriers/corridors to dispersal
    e.g. high vs low density zones

- Implementation of new Sumary Statistics or replace them by IA (CNN, Flora Jay)

Also need for code optimization to decrease computation times… for simulations and summary statistic computations…

Thanks for your attention !

Questions / Discussion,

this afternoon because I’ve been too long…