The Migraine project : A user- friendly software for likelihood-based inferences of spatial structure and demographic history from genetic data

Raphaël Leblois, Champak Beeravolu Reddy, François Rousset

The Migraine project :

A “user-friendly” software for likelihood-based inference of spatial structure and demographic history from genetic data

Summer Research School
"Software and Statistical Methods for Population Genetics”
June 2013

Raphaël Leblois[1] Champak Beeravolu[1] François Rousset[2]

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

[2]Institut des Sciences de l’Évolution (ISE-M), Université Montpellier 2

Overview

1. Introduction
2. Algorithms used in Migraine
3. Demographic & mutational models
4. The Migraine software
5. Demo
6. Simulation studies
    - Inferences under Isolation By Distance
    - Inferences of past changes in population sizes
7. Future directions
8. Conclusions

Migraine project: objectives and methods

FOCUS

Inference by ML of demographic and historical parameters from genetic data :

- Migration rates, dispersal distributions, changes in population size, divergence events,...
- Allelic data (microsatellites), short DNA sequence data, SNPs

Migraine project: objectives and methods

FOCUS

Inference by ML of demographic and historical parameters from genetic data :

- Migration rates, dispersal distributions, changes in population size, divergence events,...
- Allelic data (microsatellites), short DNA sequence data, SNPs

AIM

Assess validity and robustness of the method :

- Bias, RMSE, coverage properties of confidence intervals
- robustness to realistic but “uninteresting” mis-specifications

→ provide an “easy to use” software based on a validated method

Migraine project: objectives and methods

Methods

Estimation of likelihood by an absorbing MC algorithm using Importance Sampling (IS) technics :

- first described by Griffiths & Tavaré (1994)
- further improved by Stephens & Donnelly (2000) for single pop.
- and generalized by de Iorio & Griffiths (2004 *Adv. Appl. Probability*)

This approach uses coalescent simulation to estimate the likelihood of a genetic sample, but is very different from the more common MCMC approaches (e.g. LAMARC, IM, MsVar)

Overview

1. Introduction
2. Algorithms used in Migraine
3. Demographic & mutational models
4. The Migraine software
5. Demo
6. Simulation studies
   - Inferences under Isolation By Distance
   - Inferences of past changes in population sizes
7. Future directions
8. Conclusions

Coalescent-based algorithms to estimate the likelihood

IS algorithms:

- Griffiths et al.
- absorbing Markov chain on the genealogical space
- Independent exploration of the parameter space

MCMC algorithms:

- Felsenstein et al.
- Monte Carlo Markov Chain on the genealogical and parameter spaces

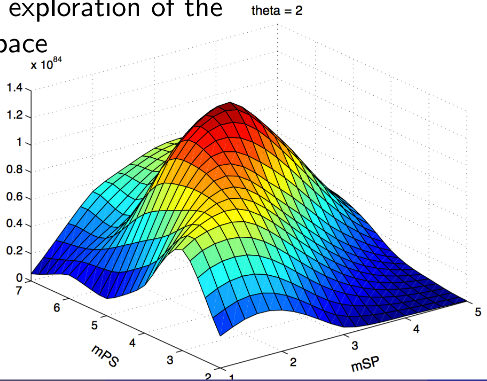
Coalescent-based algorithms to estimate the likelihood

IS algorithms:

- absorbing Markov chain on the genealogical space
- Independent exploration of the parameter space

MCMC algorithms:

- Monte Carlo Markov Chain on the genealogical and parameter spaces



Coalescent-based algorithms to estimate the likelihood

IS algorithms:

- Griffiths et al.
- absorbing Markov chain on the genealogical space
- Independent exploration of the parameter space
- difficult to implement, only simple models
- not much used : GeneTree and Migraine only

MCMC algorithms:

- Felsenstein et al.
- Monte Carlo Markov Chain on the genealogical and parameter spaces
- Easier to implement, can consider complex models
- Commonly used and implemented in many softwares : e.g. Lamarc, Migrate, Batwing, IM, MsVar,...

IS Coalescent-based algorithms used in Migraine

- Let **n** be the sample configuration:
  $\mathbf{n} = \{n_{\alpha i}\}$ (allele/haplotype counts in each location sampled)
- Denote $\mathcal{H}$ an ancestral history (i.e. a coalescent tree with mutations) from the present configuration, $H_0 = \mathbf{n}$, to the MRCA, H_{-m}:

$$\mathcal{H} = \{H_k; k = 0, -1, \ldots, -m\}$$

- Then for any given state H_k of the history :

$$p(H_k) = \sum_{\{H_{k-1}\}} p(H_k|H_{k-1})p(H_{k-1})$$

IS Coalescent-based algorithms used in Migraine

- $\mathbf{n} = \{n_{\alpha i}\}$: sample configuration
- $\mathcal{H} = \{H_k; k = 0, -1, \ldots, -m\}$: ancestral history of the sample
- $p(H_k) = \sum_{\{H_{k-1}\}} p(H_k|H_{k-1})p(H_{k-1})$
- Expending the recursion over all ancestral histories compatible with the sample, leads to :

$$p(H_0) = \sum_{(H_0,\ldots,H_{-m})} p(H_0|H_1)\ldots p(H_{-m+1}|H_{-m})p(H_{-m})$$
$$= E_p\left[p(H_0|H_1)...p(H_{-m+1}|p(H_{-m})\right]$$

IS Coalescent-based algorithms used in Migraine

- $\mathbf{n} = \{n_{\alpha i}\}$: sample configuration
- $\mathcal{H} = \{H_k; k = 0, -1, \ldots, -m\}$: ancestral history of the sample
- $p(H_k) = \sum_{\{H_{k-1}\}} p(H_k|H_{k-1})p(H_{k-1})$
- Expending the recursion over all ancestral histories compatible with the sample, leads to :

$$p(\mathbf{n}) = p(H_0) = E_p\left[p(H_0|H_1)...p(H_{-m+1}|p(H_{-m})\right]$$

However:

- Forward transition prob. $p(H_k|H_{k-1})$ can not be directly used in a backward process
- Backward transition prob. $p(H_{k-1}|H_k)$ are generaly unknown (except for parent independent mutations (PIM) in a single panmictic population)

IS Coalescent-based algorithms used in Migraine

- $\mathbf{n} = \{n_{\alpha i}\}$: sample configuration
- $\mathcal{H} = \{H_k; k = 0, -1, \ldots, -m\}$: ancestral history of the sample
- $p(H_k) = \sum_{\{H_{k-1}\}} p(H_k|H_{k-1})p(H_{k-1})$
- Importance Sampling (IS) technic is used:
  Let $Q(H_{k-1})$ be a proposal distribution such that

$$p(H_k) = \sum_{\{H_{k-1}\}} p(H_k|H_{k-1}) \frac{p(H_{k-1})}{Q(H_{k-1})} Q(H_{k-1})$$

$$= \mathrm{E}_Q \left[p(H_k|H_{k-1}) \frac{p(H_{k-1})}{Q(H_{k-1})} \right]$$

but need an efficient proposal distribution...

IS Coalescent-based algorithms used in Migraine

The ideal proposal: $Q(H_{k-1}) = p(H_{k-1}|H_k)$

- The ideal proposal is the backward transition probability $p(H_{k-1}|H_k)$, then

$$p(H_k|H_{k-1})\frac{p(H_{k-1})}{Q(H_{k-1})} = \frac{p(H_k \cap H_{k-1})}{p(H_{k-1}|H_k)} = p(H_k)$$

and a single tree reconstruction allows exact likelihood computations (null variance).

- $p(H_{k-1}|H_k)$ is unknown, instead we use approximations $\hat{p}(H_{k-1}|H_k)$:

$$\mathrm{E}_Q\left[p(H_{k-1})\frac{p(H_k|H_{k-1})}{\hat{p}(H_{k-1}|H_k)}\right] = p(H_k)$$

but then many trees are necessary to get a good estimation of the likelihood.

IS Coalescent-based algorithms used in Migraine

- The likelihood of the present configuration can then be written as a product of importance weights:

$$p(\mathbf{n}) = p(H_0) = \mathrm{E}_{\hat{p}} \underbrace{\left[\frac{p(H_0|H_{-1})}{\hat{p}(H_{-1}|H_0)} \cdots \frac{p(H_{-m+1}|H_{-m})}{\hat{p}(H_{-m}|H_{-m+1})} p(H_{-m}) \right]}_{\mathcal{W}_r}$$

$$= \mathrm{E}_{\hat{p}} \left[\frac{p(\mathcal{H}_{\rightarrow})}{\hat{p}(\mathcal{H}_{\leftarrow})} \right]$$

- Then we use Monte Carlo simulations on the absorbing backward Markov chain process describe above, using the IS transition probabilities, to infer the likelihood for a given parameter point Θ

$$L(\Theta) = p_{\Theta}(\mathbf{n}) \approx \frac{1}{R} \sum_{r=1}^{R \gg 1} \mathcal{W}_r$$

IS Coalescent-based algorithms used in Migraine

- Griffiths & Tavaré 1994, Nath & Griffiths 1996, Bahlo & Griffiths 2000 : "uniform" IS proposal, not very efficient (millions of trees).
- Stephens & Donnelly 2000 : much more efficient IS proposal for a single isolated population (1-100 trees).
- deIorio & Griffiths 2004a, b : generalization of SD2000 proposal for structured population models (30-100 trees).

IS Coalescent-based algorithms used in Migraine

Additional approximate but fast algorithm : the PAC-likelihood

Migraine also uses an heuristic approximation known as PAC-likelihood defined by Li and Stephens 2003, Cornuet and Beaumont 2007

- Based on $\hat{\pi}$ an approximation of $\pi(j, \alpha|\mathbf{n})$ the probability that, given an observed sample configuration **n**, the next sampled gene is of type j and from population α (same approx. than SD2000 & DIG2004)
- No tree reconstruction, only based on the different type of gene observed in the sample

IS Coalescent-based algorithms used in Migraine

Additional approximate but fast algorithm : the PAC-likelihood

Migraine also uses an heuristic approximation known as PAC-likelihood defined by Li and Stephens 2003, Cornuet and Beaumont 2007

- Based on $\hat{\pi}$ an approximation of $\pi(j, \alpha|\mathbf{n})$ (same approx. than SD2000 & DIG2004)
- No tree reconstruction,

Basic idea : each sampled genes is added one by one with associated probability $\hat{\pi}(j, \alpha|\mathbf{n})$ to reconstruct the whole sample

$$\begin{aligned} p(\mathbf{n}) &= p(\mathbf{n}-\mathbf{1})\pi(j, \alpha|\mathbf{n}-\mathbf{1}) \\ &\approx p(\mathbf{n}-\mathbf{1})\hat{\pi}(j, \alpha|\mathbf{n}-\mathbf{1}) \end{aligned}$$

IS Coalescent-based algorithms used in Migraine

Additional approximate but fast algorithm : the PAC-likelihood

Migraine also uses an heuristic approximation known as PAC-likelihood defined by Li and Stephens 2003, Cornuet and Beaumont 2007

- Based on $\hat{\pi}$ an approximation of $\pi(j, \alpha|\mathbf{n})$ (same approx. than SD2000 & DIG2004)
- No tree reconstruction,

$$\hat{L}_{PAC}(\theta) = \underbrace{\frac{1}{R}\sum_{r=1}^{R\gg 1}\mathcal{M}_n\prod_{n}^{i=2}\hat{\pi}(gene_i|\mathbf{n}_i = \{gene_k\}_{k<i})}_{R\ \ random\ \ sample\ \ reconstruction}$$

IS Coalescent-based algorithms used in Migraine

Additional approximate but fast algorithm : the PAC-likelihood

Migraine also uses an heuristic approximation known as PAC-likelihood defined by Li and Stephens 2003, Cornuet and Beaumont 2007

- Based on $\hat{\pi}$ an approximation of $\pi(j, \alpha|\mathbf{n})$ (same approx. than SD2000 & DIG2004)
- No tree reconstruction,
- **Pros**: very fast, very accurate
- **Cons** : can only be applied for equilibrium models (IBD, OnePop, NPop)

IS Coalescent-based algorithms : conclusion

- Very different from classical coalescent-based MCMC
- Very efficient since the work of SD2000, and DIG2004
- PAC-likelihood is a good fast approximation for equilibrium models
- But it is not always straightforward to add new mutational or demographic features

Overview

1. Introduction
2. Algorithms used in Migraine
3. Demographic & mutational models
4. The Migraine software
5. Demo
6. Simulation studies
   - Inferences under Isolation By Distance
   - Inferences of past changes in population sizes
7. Future directions
8. Conclusions

Mutational models implemented in Migraine

PIM = KAM (allelic data, Crow and Kimura 1970)

Parent independent mutation :

each mutation → one of the K (or K-1) possible allelic states

Allows to consider the most efficient proposal distributions for any demographic model (optimal IS proposal distribution under a single population model, i.e. a single tree give the exact likelihood)

most basic approximation for microsatellite mutation processes

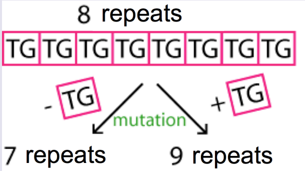
Mutational models implemented in Migraine

SMM (allelic data, Ohta and Kimura 1973)

Strict stepwise model :

each mutation adds or removes a motif

better approximation for microsatellite mutation processes than KAM



GSM (allelic data, Pritchard et al. 1999)

Generalized stepwise model :

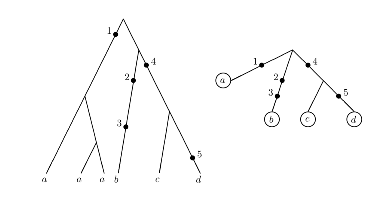
each mutation adds or removes X motif, with $X \sim \mathcal{G}\text{eom}(pGSM)$

better approximation for microsatellite mutation processes than SMM but adds a parameter, $pGSM$ ($\nearrow$ computation times)

Mutational models implemented in Migraine

ISM (DNA sequence data, Kimura 1969)

- The most simple model of sequence evolution
- Polymorphisms at a base pair correspond to a unique mutation in the coalescent
- New mutations only occur at sites never previously mutant
- Each mutation produces a new haplotype
  → The haplotypes in a sample define a unique perfect phylogeny



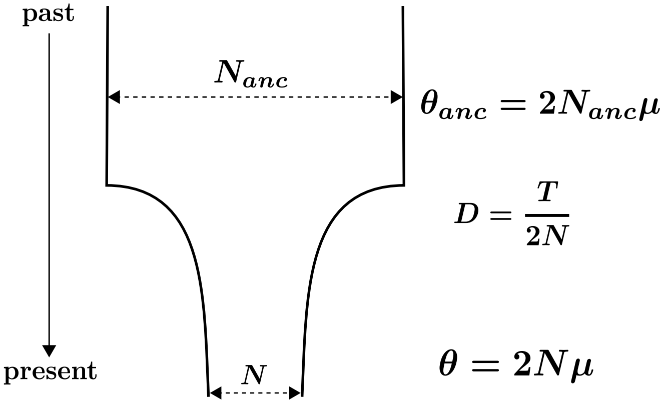
Demographic models implemented in Migraine: OnePop

One stable WF population (Eq.)

- One demographic parameter ($+ \mu$, mutation rate/locus/generation):
  - $*$ N: pop size (nber of genes)
- Availlable mutation models : KAM/PIM, SMM, GSM, ISM
- Inference of one or two scaled parameters:
  - $*$ [$pGSM$] if GSM
  - $*$ $\theta = 2N\mu$

One population with single past change in size : The OnePopVarSize model

Ex: a single population undergoing an exponential contraction that started T generation ago



$$\theta_{anc} = 2N_{anc}\mu$$

$$D = \frac{T}{2N}$$

$$\theta = 2N\mu$$

Demographic models implemented in Migraine: OnepopVarSize

One WF population with variable size : single past change (Diseq.)

- Three parameters (+ μ, mutation rate/locus/generation):
  - N_{act}: pop size at sampling time (nber of genes)
  - T: Time in the past when demographic change starts,
  - N_{anc}: ancestral population size
- Availlable mutation models : KAM/PIM, SMM, GSM, ISM
- Inference of 3-4 scaled parameters:
  - [$pGSm$] if GSM
  - $\theta = 2N_{act}\mu$
  - $D = \frac{T}{2N_{act}}$
  - $\theta_{anc} = 2N_{anc}\mu$

Demographic models implemented in Migraine: OnepopVarSize

One WF population with variable size : single past change (Diseq.)

- Three parameters: N_{act}, T, N_{anc}
- Availlable mutation models : KAM/PIM, SMM, GSM, ISM
- Inference of 3-4 scaled parameters:
  - $[pGSm]$ if GSM
  - $\theta = 2N_{act}\mu$
  - $D = \frac{T}{2N_{act}}$
  - $\theta_{anc} = 2N_{anc}\mu$
- Tested with exponential decrease in population size (section OPVS), but can consider discret, linear or logistic growths and declines.

Demographic models implemented in Migraine: Npop

Two populations connected by migration (Eq.)

- Four parameters (+ μ, mutation rate/locus/generation):
  - N_T: total pop size (nber of genes, $N_1 + N_2$)
  - $q_1 = N_1/N_2$: relative pop sizes,
  - $m_{1\to2}$ and $m_{1\to2}$, the migration rates
- Availlable mutation models : KAM/PIM, SMM, GSM, ISM
- Inference of 4-5 scaled parameters:
  - [$pGSm$] if GSM
  - $\theta = 2N_T\mu$
  - q_1
  - $M_1 = 2N_1m_{1\to2}$
  - $M_2 = 2N_2m_{2\to1}$

Demographic models implemented in Migraine: Npop

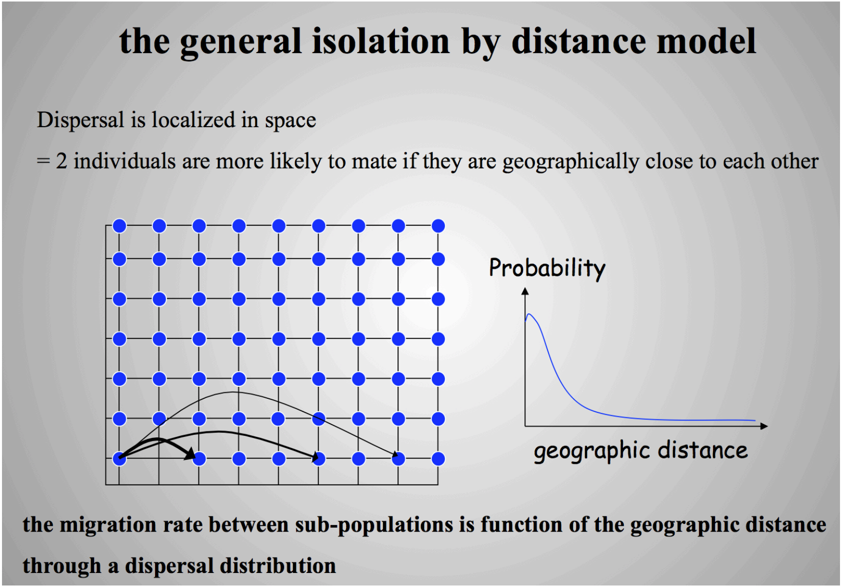
Two populations connected by migration (Eq.)

- Four parameters: N_T, $q_1 = N_1/N_2$, $m_{1\to2}$ and $m_{1\to2}$
- Availlable mutation models : KAM/PIM, SMM, GSM, ISM
- Inference of 4-5 scaled parameters: $[pGSm]$, $\theta = 2N_T\mu$, q_1, $M_1 = 2N_1m_{1\to2}$, $M_2 = 2N_2m_{2\to1}$

More Populations?

- Migraine should be able to consider up to four populations connected by migration, but only under PIM,
- but it has never been tested
- Main potential problem: high nber of parameters, e.g. 15 param for 4 populations

Demographic models implemented in Migraine: IBD



the general isolation by distance model

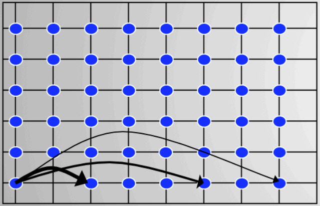
Dispersal is localized in space

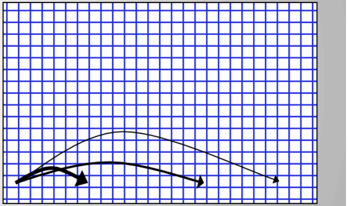
= 2 individuals are more likely to mate if they are geographically close to each other

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

Demographic models implemented in Migraine: IBD

2 models depending on individual spatial distribution in the landscape





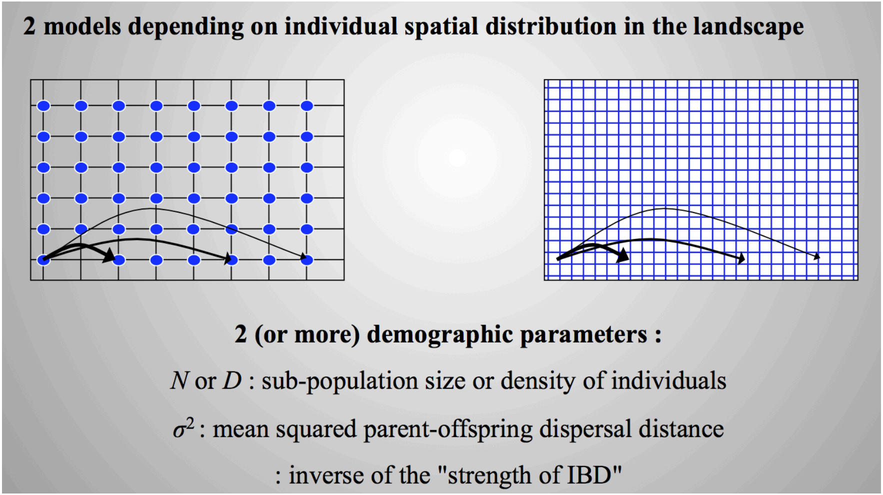
Population with a demic structure

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)

Demographic models implemented in Migraine: IBD



2 models depending on individual spatial distribution in the landscape

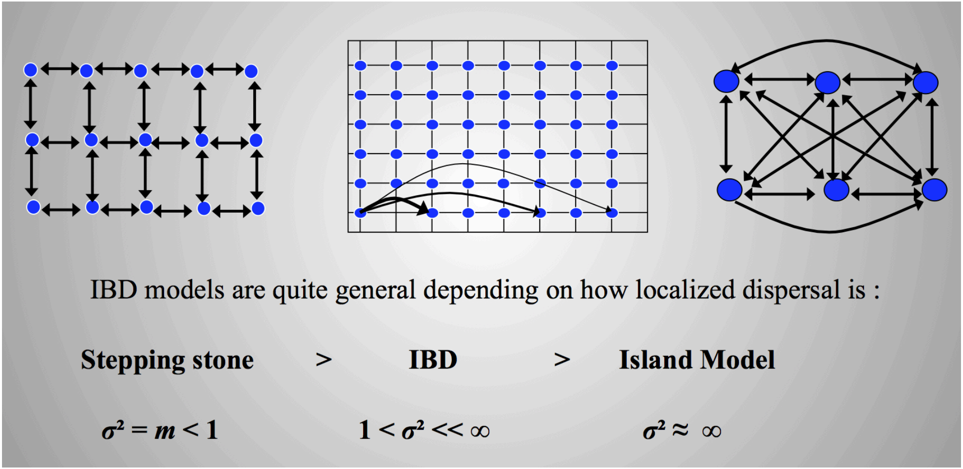
2 (or more) demographic parameters :

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

σ^2 : mean squared parent-offspring dispersal distance

: inverse of the "strength of IBD"

Demographic models implemented in Migraine: IBD



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

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

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

Demographic models implemented in Migraine: IBD

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

- Fully homogeneous model → four parameters (+ μ):
  - ∗ d: nb of subpopulations
  - ∗ N: sub pop size (nber of genes, $N_T = dN$)
  - ∗ m: the emmigration rates from any subpopulation
  - ∗ g: shape of the geometric dispersal distribution in the inference algorithmn
- Availlable mutation models : KAM/PIM
- Inference of 3 scaled parameters:
  - ∗ $\theta = 2N\mu$
  - ∗ $M = 2Nm$
  - ∗ g
  - + one composite parameter $Nb = 4\pi D\sigma^2$

Mutational & demographic models: summary

Mutational models:

- KAM/PIM, SMM, GSM, ISM (and soon SNPs...)
- Migraine allows multimarker analyses e.g. SMM/GSM, ISM/GSM, ...

Demographic models:

- At equilibrium : OnePop, N(2-4)pop, IBD
- Disequilibrium models : OnePopVarSize, (and soon FounderFlush, IM between 2 pops,...)

Overview

1. Introduction
2. Algorithms used in Migraine
3. Demographic & mutational models
4. The Migraine software
5. Demo
6. Simulation studies
   - Inferences under Isolation By Distance
   - Inferences of past changes in population sizes
7. Future directions
8. Conclusions

What's in the Migraine software?

C++ core IS computations

- Stratified random sampling of parameter points (Bounds provided by user)
- Estimation of the likelihood at each point using IS
- write R code

R (automated interaction between C++ and R codes)

- Likelihood surface interpolation by Kriging
- Inference of MLEs and CIs
- (Nice) Plots of 1D and 2D Likelihood profiles
- Computation of a list of new points inside the convex 99.9% envelope
- Computation of LRT-Pvalue (e.g. to test an hypothesis = Nratio<1)

parts of R code written by C++,
others more constant parts compiled in a packge called "Rmigraine"

What’s in the Migraine software?

C++ core IS computations

Point sampling, LIkelihood estimation, Write R code

R scripts (automated interaction between C++ and R codes)

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.

How does the Migraine software work?

- One (or many) Genepop data files
  associated with a nexus files for DNA sequence data sets

- Parametrization of C++ and R analysis using a text file
  or using the graphical interface (Soon)

- Run Migraine

- Outputs :
  - Results text file (ML, CI, LR tests)
  - Graphics in a ps / eps / pdf file

How does the Migraine software work?

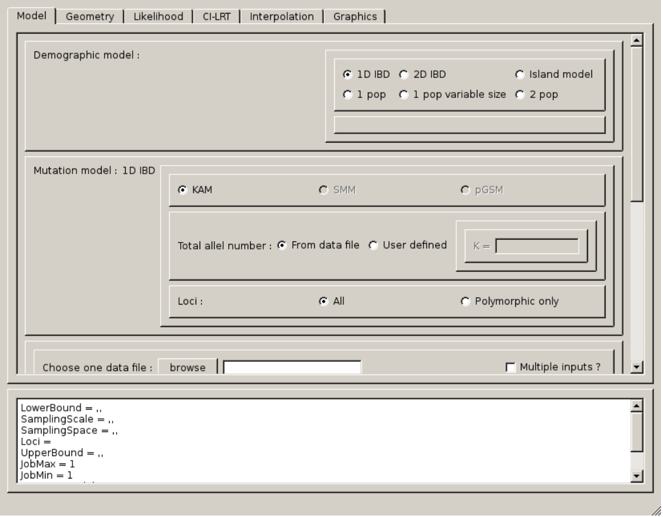
most complex parameters have good default values and ... we provide a very detailed and comprehensive documentation with:

- Basic theory (IS + kriging)
- How to install Migraine (C++ code and R package)
- Complete description by key words of all parameters
- description and interpretation of all outputs
- Simple examples to run (good to start with)

Moreover, the GUI will include a "What's this" button linked to all keyword description of the documentation

How does the Migraine software work?

GUI under construction (should be finished for July!)



Demographic model :

● 1D IBD ○ 2D IBD ○ Island model
○ 1 pop ○ 1 pop variable size ○ 2 pop

Mutation model : 1D IBD

● KAM ○ SMM ○ pGSM

Total allel number : ● From data file ○ User defined

Loci : ● All ○ Polymorphic only

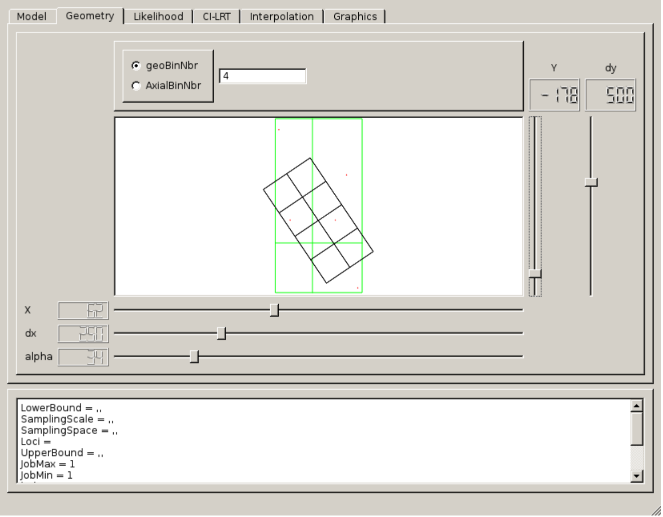
Choose one data file :

□ Multiple inputs ?

```
LowerBound = ,,
SamplingScale = ,,
SamplingSpace = ,,
Loci =
UpperBound = ,,
JobMax = 1
JobMin = 1
```

How does the Migraine software work?

GUI under construction (should be finished for July!)



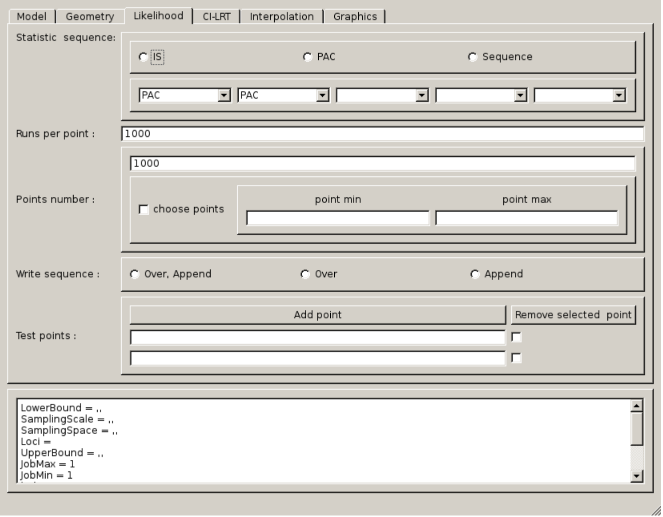
● geoBinNbr
○ AxialBinNbr
4

Y -178
dy 500
x 62
dx 250
alpha 34

```
LowerBound = ,,
SamplingScale = ,,
SamplingSpace = ,,
Loci =
UpperBound = ,,
JobMax = 1
JobMin = 1
```

How does the GUI of Migraine look like?

GUI under construction (should be finished for July!)



Statistic sequence: PAC PAC

Runs per point : 1000

Points number : 1000

Write sequence :

Test points :

```
LowerBound = ,,
SamplingScale = ,,
SamplingSpace = ,,
Loci =
UpperBound = ,,
JobMax = 1
JobMin = 1
```

Overview

1. Introduction
2. Algorithms used in Migraine
3. Demographic & mutational models
4. The Migraine software
5. Demo
6. Simulation studies
   - Inferences under Isolation By Distance
   - Inferences of past changes in population sizes
7. Future directions
8. Conclusions

demo on two examples...

Lets look in details into two examples of concrete data analyses :

IBD and OnePopVarSize....

Overview

1 Introduction

2 Algorithms used in Migraine

3 Demographic & mutational models

4 The Migraine software

5 Demo

6 Simulation studies
- Inferences under Isolation By Distance
- Inferences of past changes in population sizes

7 Future directions

8 Conclusions

Overview

1 Introduction

2 Algorithms used in Migraine

3 Demographic & mutational models

4 The Migraine software

5 Demo

6 Simulation studies

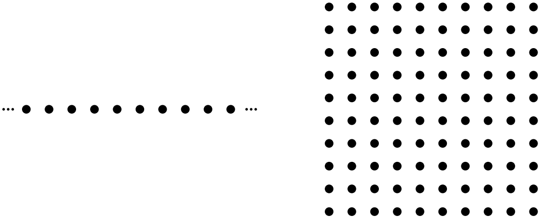
- Inferences under Isolation By Distance
- Inferences of past changes in population sizes

7 Future directions

8 Conclusions

Isolation by distance: biological context

- Localized dispersal
- Ecological studies of dispersal in non-model organisms
- Small data sets, ∼ 10-20 microsatellites, ∼ 200–300 individuals



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.

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 (marginal 1D distribution in 2D model).

Isolation by distance: Parameters

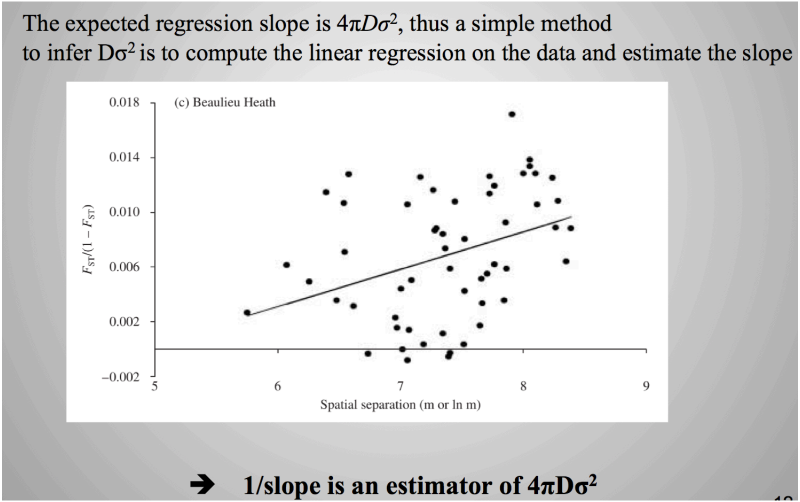
Deme size N, dispersal probability m, mutation probability μ
distribution of dispersal distance: geometric decrease with distance, with scale parameter g.

special interest in the neighborhood size $\propto D\sigma^2$ where D is population density and σ^2 is second moment of dispersal distance (marginal 1D distribution in 2D model).

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")

Previous method: Rousset's regression (1997)

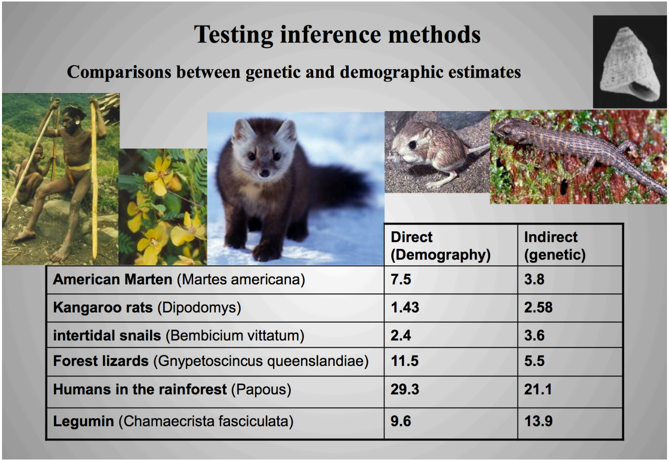
F_{ST}-based method implemented in Genepop



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

➔ **1/slope is an estimator of $4\pi D\sigma^2$**

Special interest in IBD models



Testing inference methods

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

good agreement between genetic and demographic estimates
→ quite realistic model for fine scale population genetics

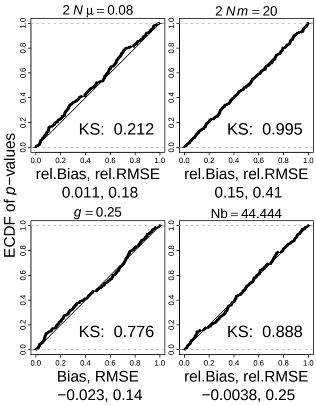
Migrainevalidation procedure

- Check ideal performance under ideal conditions
- Check robustness under non-ideal conditions (various mis-specifications)

Migrainevalidation procedure

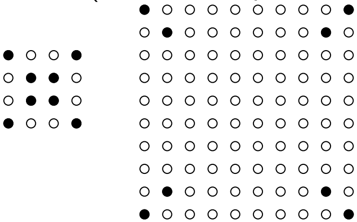
- Check ideal performance under ideal conditions

Ideal performance := valid confidence intervals ⇔ uniform distribution of *p*-values of (profile) LR tests of true simulation parameters



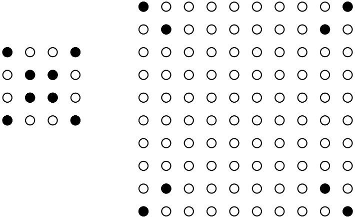
IBD simulation design: ideal conditions

- 40 gene copies at each of 10 loci in each of 8 demes (sometimes 10) demes (smallish sample size for ecological studies).



IBD simulation design: ideal conditions

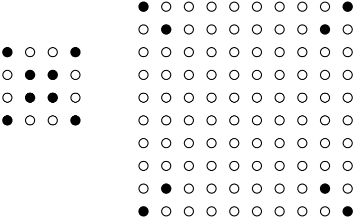
- 40 gene copies at each of 10 loci in each of 8 demes (sometimes 10) demes (smallish sample size for ecological studies).



- 200 simulated data sets

IBD simulation design: ideal conditions

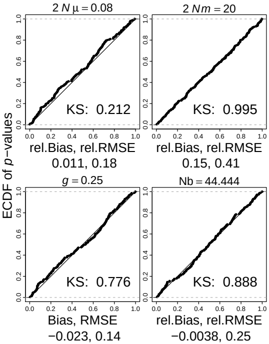
- 40 gene copies at each of 10 loci in each of 8 demes (sometimes 10) demes (smallish sample size for ecological studies).



- 200 simulated data sets
- 100 demes: each data set takes ≈ 6 CPU hours by PAC-likelihood, ∼1 CPU year by true likelihood (though easy to distribute over different CPUs)

Results under ideal conditions: validating the whole inference process

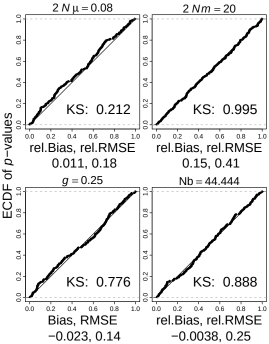
ex: N: 40000; m: 0.00025; μ: 10^{-6}

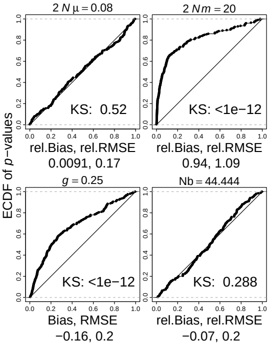


First result: very good LRT distributions
→ validation of the method

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

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

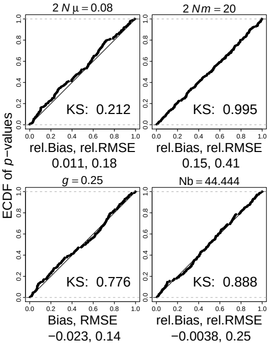


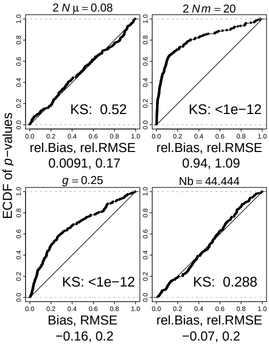


Something wrong ?

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

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

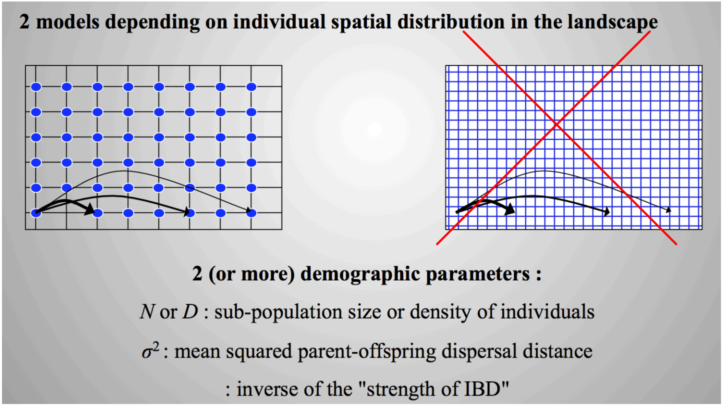




Diffusion approximation→ bias in Nm estimation increases with m

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

2d main result: Diffusion approximation strongly limits the consideration of “continuous populations” models with Migraine



2 models depending on individual spatial distribution in the landscape

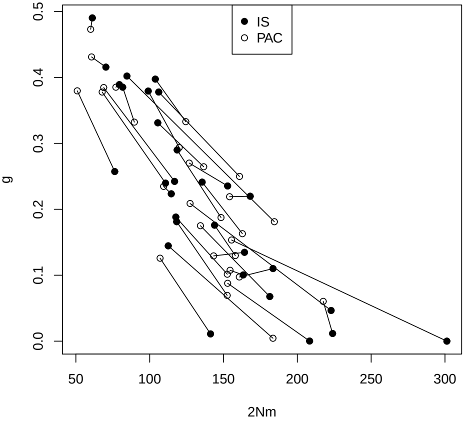
2 (or more) demographic parameters :

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

σ^2 : mean squared parent-offspring dispersal distance

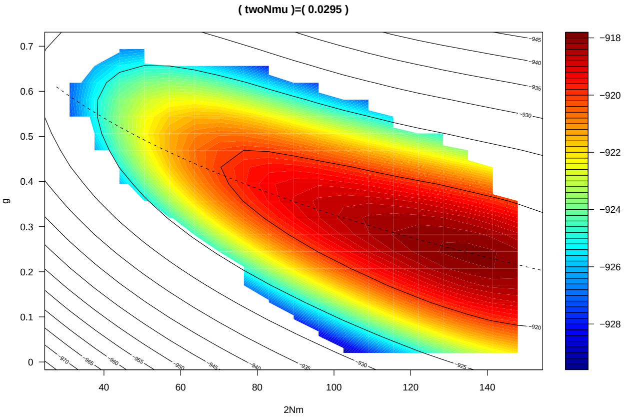
: inverse of the "strength of IBD"

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



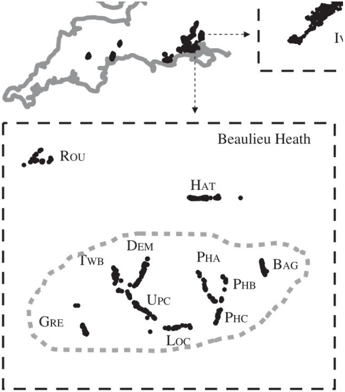
Results under ideal conditions: another limit du to Nm, g covariance

3d main result: no information to infer Nm and g separately



A realistic setting

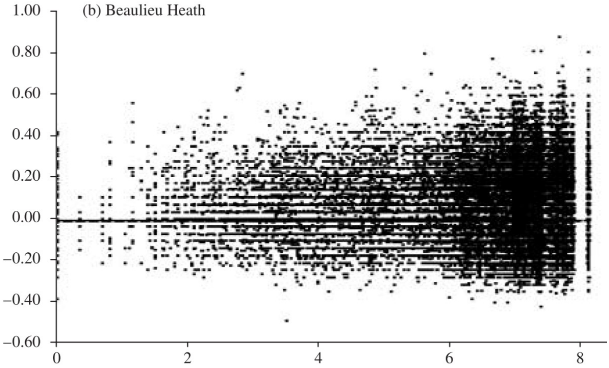
A realistic setting



Demographic estimate $D_e\sigma_e^2 \hat{=} 555$ ind ($D_e \hat{=} 0.003$ ind.m^{-2}, $\sigma_e \hat{=} 125$ m)

Watts et al. *Mol. Ecol.* 2007

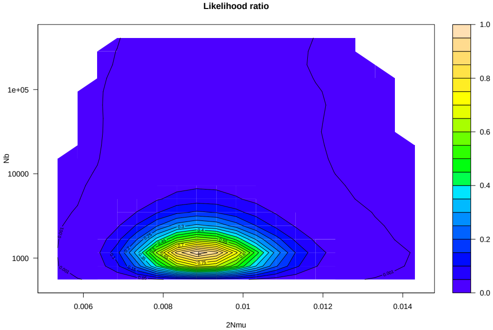
A realistic setting



Demographic estimate $D_e\sigma_e^2 \hat{=} 555$ ind ($D_e \hat{=} 0.003$ ind.m^{-2}, $\sigma_e \hat{=} 125$ m)
Genetic regression estimate $D_e\sigma_e^2 \hat{=} 753$ ind (CI 319 – 3162).

Watts et al. *Mol. Ecol.* 2007

A realistic setting



Demographic estimate $D_{\mathrm{e}}\sigma_{\mathrm{e}}^2 \hat{=} 555$ ind ($D_{\mathrm{e}} \hat{=} 0.003$ ind.m^{-2}, $\sigma_{\mathrm{e}} \hat{=} 125$ m)
Genetic regression estimate $D_{\mathrm{e}}\sigma_{\mathrm{e}}^2 \hat{=} 753$ ind (CI 319 – 3162).
Genetic PAC-likelihood estimate $D_{\mathrm{e}}\sigma_{\mathrm{e}}^2 \hat{=} 1110$ ind (CI 600 – 3125)

Performance in messy/realistic conditions

- Unknown mutation model
- Unknown dispersal distribution
- Cannot consider continuous populations (i.e. N=1)

Performance in messy/realistic conditions

- Unknown mutation model

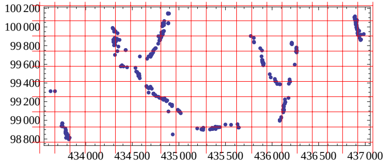
  Simulations of samples under SMM, analysis under KAM

- Unknown dispersal distribution

  Simulation of samples under “Sichel” model (Chesson & Lee, 2005)

  Analysis under the geometric dispersal model

- Cannot consider continuous populations (i.e. N=1)



A binning step is incorporated

Performance in messy/realistic conditions

- Unknown mutation model

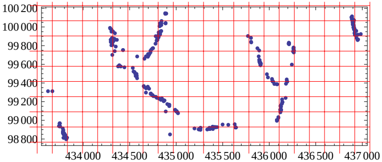
  Simulations of samples under SMM, analysis under KAM

- Unknown dispersal distribution

  Simulation of samples under "Sichel" model (Chesson & Lee, 2005)

  Analysis under the geometric dispersal model

- Cannot consider continuous populations (i.e. N=1)

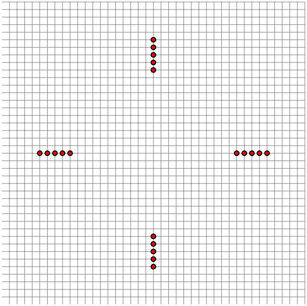


A binning step is incorporated

→ Many things can go wrong, but neighborhood estimation is relatively robust

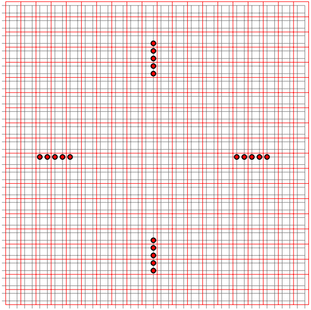
Performance in messy/realistic conditions

- **Simulations settings**:
  40×40 array, $N = 50$, m=0.5, $g = 0.5$, $\mu = 10^{-4}$
  200 individuals, 10 loci



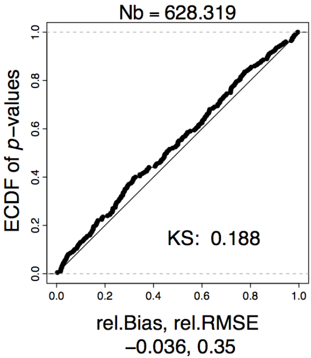
Performance in messy/realistic conditions

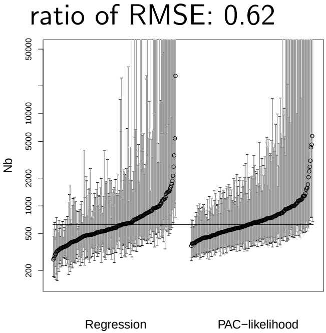
- **Simulations settings**:
  40×40 array, $N = 50$, m=0.5, $g = 0.5$, $\mu = 10^{-4}$
  200 individuals, 10 loci
- **Analysis settings**: 20×20 grid of bins, (few CPU days per sample)



Performance in messy/realistic conditions

- **Simulations settings**:
  40 × 40 array, $N = 50$, m=0.5, $g = 0.5$, $\mu = 10^{-4}$
  200 individuals, 10 loci
- **Analysis settings**: 20 × 20 grid of bins, (few CPU days per sample)





Performance in messy/realistic conditions

- Complex effects of binning on Nm and g estimation
  Bad: depend on the number of samples per bin → difficult to infer dispersal rates and shape
- Expected $> 50\%$ negative bias of $N\mu$ estimates under the SMM
  (no bias under correctly specified mutation model)
- Neighborhood estimation is more robust
- Gains in efficiency relative to the spatial regression method: ratios of RMSE from 0.27 to 0.62

ML inferences under isolation by distance: summary

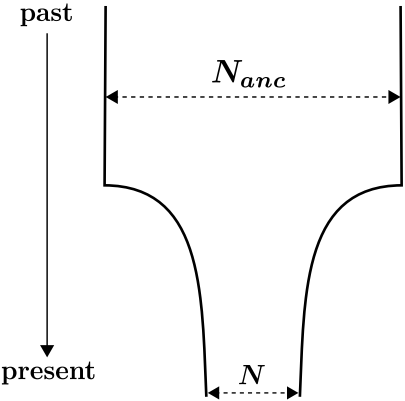
- Likelihood inferences perform in an ideal way in (restrictive) ideal conditions
- Likelihood estimation still prohibitively long in large networks of populations. PAC-likelihood more feasible.
- Additional imperfections (Likelihood and PAC-likelihood) due to the diffusion approximation when m is large.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

Overview

1. Introduction
2. Algorithms used in Migraine
3. Demographic & mutational models
4. The Migraine software
5. Demo
6. Simulation studies
   - Inferences under Isolation By Distance
   - Inferences of past changes in population sizes
7. Future directions
8. Conclusions

The model

A single population undergoing an exponential contraction



$$\theta_{anc} = 2N_{anc}\mu$$

$$D = \frac{T}{2N}$$

$$\theta = 2N\mu$$

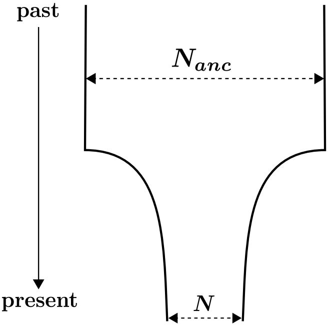
The bottleneck model (OnePopVarSize)

Demographic model

- Single isolated panmictic population
- Population size started to change
  T generations in the past, exponentially
  until present $=$ sampling time

Biological vs. scaled parameters

- Population sizes :
  N genes ($\theta_{act} = 2N\mu$),
  N_{anc} genes ($\theta_{anc} = 2N_{anc}\mu$)
- Time (change duration) :
  T generations ($D = T/2N$)



$$\theta_{anc} = 2N_{anc}\mu$$

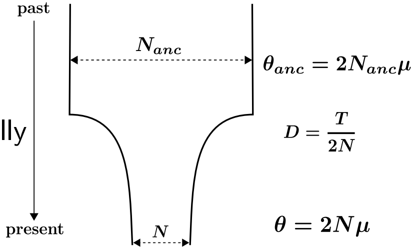
$$D = \frac{T}{2N}$$

$$\theta = 2N\mu$$

The bottleneck model (OnePopVarSize)

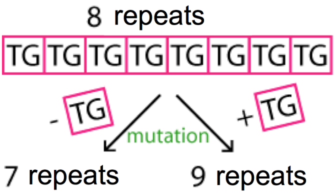
Demographic model

- Single isolated panmictic population
- Population size started to change T generations in the past, exponentially until present



Mutational model

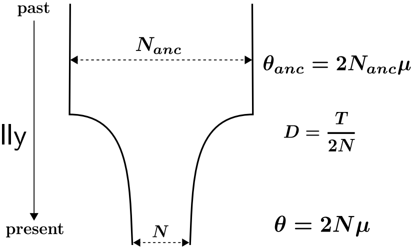
- Allelic data (microsatellites)
- simple model : SMM
- mutation rate : $\mu = 10^{-3}$



The bottleneck model (OnePopVarSize)

Demographic model

- Single isolated panmictic population
- Population size started to change T generations in the past, exponentially until present



Mutational model

- Allelic data (microsatellites)
- simple model : SMM
- mutation rate : $\mu = 10^{-3}$

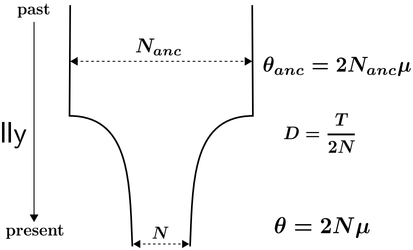
Other available method for such model : MsVar (M. Beaumont)

- Coalescent-based
- MCMC algorithm
- Bayesian implementation

The bottleneck model (OnePopVarSize)

Demographic model

- Single isolated panmictic population
- Population size started to change T generations in the past, exponentially until present



Mutational model

- Allelic data (microsatellites)
- simple model : SMM
- mutation rate : $\mu = 10^{-3}$

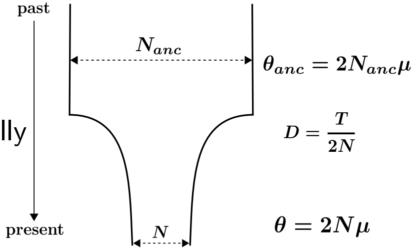
Genetic sample (small)

- 100 gene copies sampled
- 10 loci genotyped

The bottleneck model (OnePopVarSize)

Demographic model

- Single isolated panmictic population
- Population size started to change T generations in the past, exponentially until present

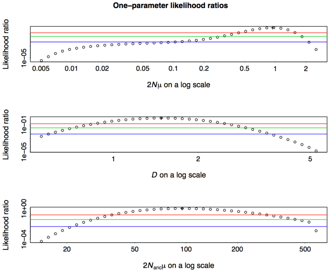


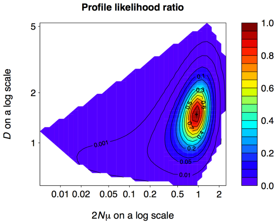
Mutational model

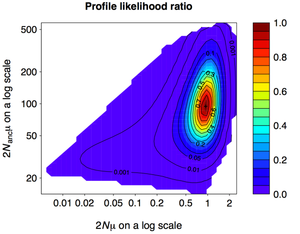
- Allelic data (microsatellites)
- simple model : SMM
- mutation rate : $\mu = 10^{-3}$

outputs for OnePopVarSize for a single data set analysis

most importantly : 1D and 2D Likelihood ratio profiles

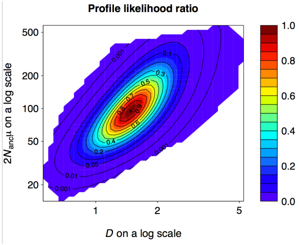






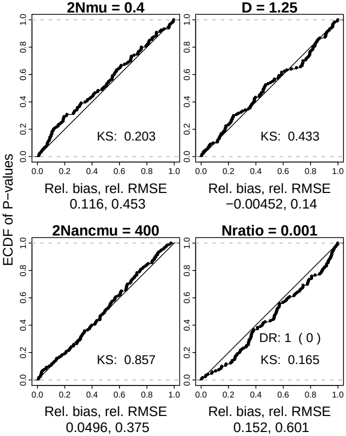
Results summary :

```
*** Confidence intervals ***

95%-coverage confidence interval for twoNmu : [ 0.441 -- 1.573 ]
95%-coverage confidence interval for D : [ 0.857 -- 2.502 ]
95%-coverage confidence interval for twoNancmu : [ 36.76 -- 295.6 ]
95%-coverage confidence interval for Nratio : [ 0.00329 -- 0.0268 ]

*** Point estimates ***

     twoNmu            T          D  twoNancmu
      0.937            0       1.48      94.67
```



OnePopVarSize : Bias, MSE, LRT on simulated data

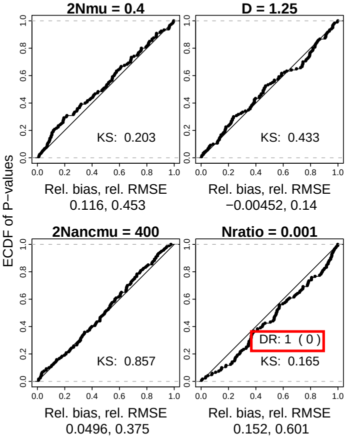
Same analyses as for IBD :



(usually)GOOD

OnePopVarSize : Bias, MSE, LRT on simulated data

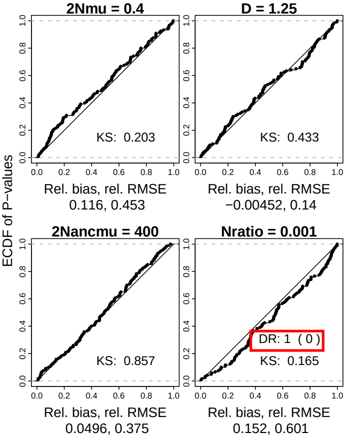
Same analyses as for IBD : (+ bottleneck detection rate)

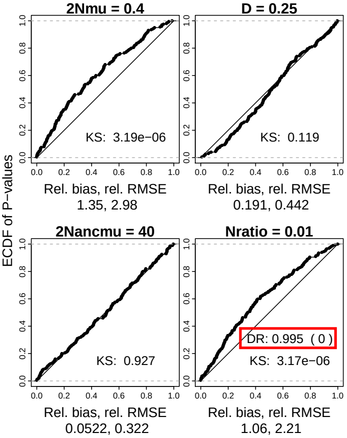


(usually)GOOD

OnePopVarSize : Bias, MSE, LRT on simulated data

Same analyses as for IBD : (+ bottleneck detection rate)



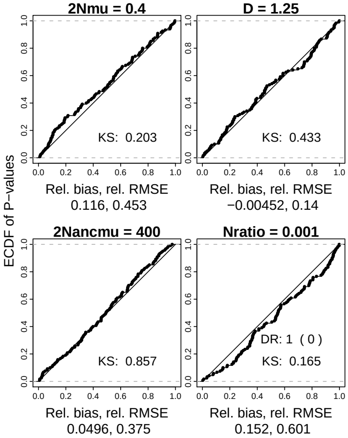


(usually)GOOD

(sometimes) LESS GOOD

OnePopVarSize : Bias, MSE, LRT on simulated data

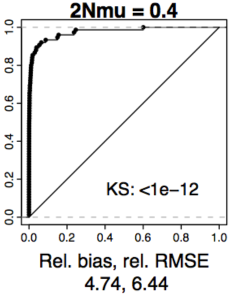
Same analyses as for IBD : (+ bottleneck detection rate)



Extremely recent and strong

10 Generations, $D = 0.025$

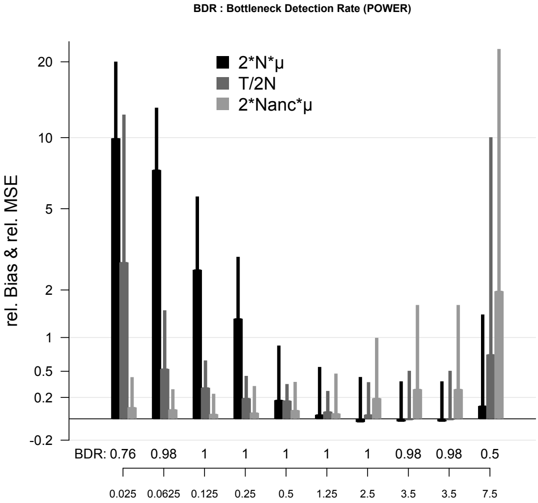
$N_{ratio} = 0.001$ ($\theta_{anc} = 400.0$)



(usually)GOOD

(very rarely) BAD

OnePopVarSize : influence of the timing of the population size change



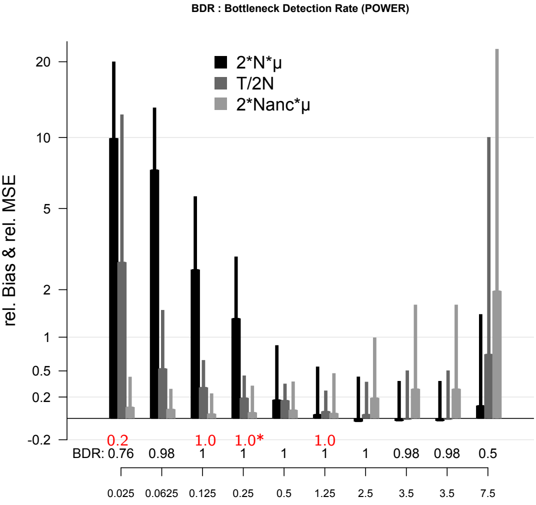
Expected performances for very recent to very ancient change

- T varies from 10 to 3000 generations ($D = T/2N$ from 0.025 to 7.5)

Results

- Very good bottleneck detection rate
- Precise parameter inference, at least for some parameters
- Strong dependance on the scenarios (as expected)

OnePopVarSize : influence of the timing of the population size change



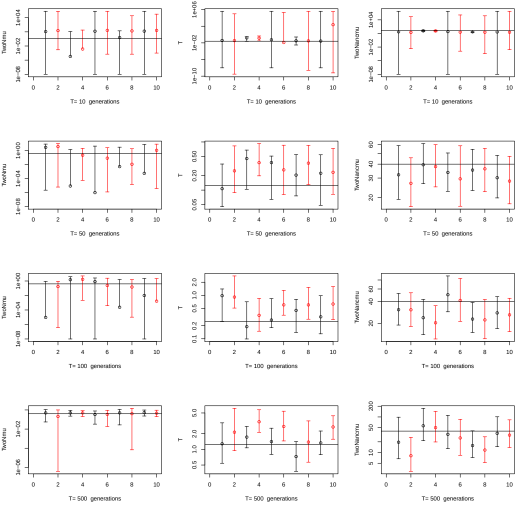
Expected performances for very recent to very ancient change

- T varies from 10 to 3000 generations ($D = T/2N$ from 0.025 to 7.5)

Some comparison with MsVar

- Similar performances for “good” scenarios
- Better bottleneck detection rate for “non-optimal” scenarios

OnePopVarSize : influence of the timing of the population size change



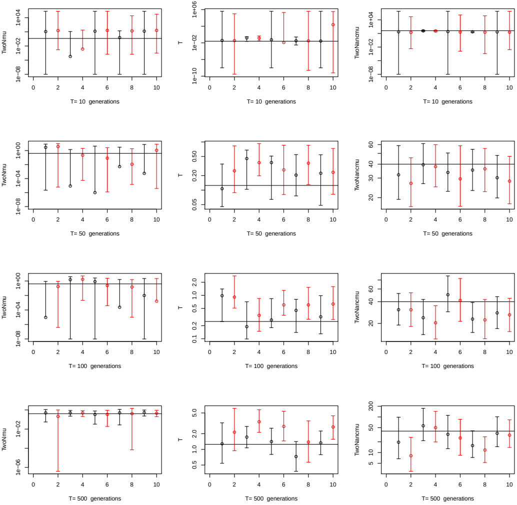
Expected performances for very recent to very ancient change

- T varies from 10 to 3000 generations ($D = T/2N$ from 0.025 to 7.5)

Some comparison with MsVar

- Parameter inference seems more accurate

OnePopVarSize : influence of the timing of the population size change



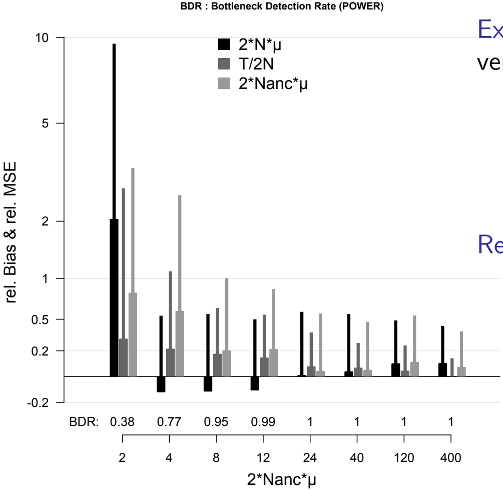
Expected performances for very recent to very ancient change

- T varies from 10 to 3000 generations ($D = T/2N$ from 0.025 to 7.5)

Comparison with MsVar is not easy

- Frequentist vs. bayesian approaches
- very long computation times for MCMC

OnePopVarSize : influence of the strength of the population size change



Expected performances for very weak to very strong changes

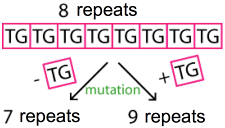
- Nratio varies from 5 to 1000,
- $N = 200,\ N_{anc} = \{400, , 200\ 000\}$
- fixed $D = 1.25$ (good case))

Results

- Very good bottleneck detection rate for $N_{ratio} \geqslant 10$
- Precise parameter inference when bottlenecks are detected
- better for stronger bottlenecks

OnePopVarSize : mis-specification of mutation processes

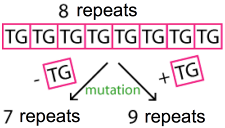
Microsatellite markers show complex mutation processes



- Mutations do not fit SMM,
  indels of more than one repeat often occur

OnePopVarSize : mis-specification of mutation processes

Microsatellite markers show complex mutation processes

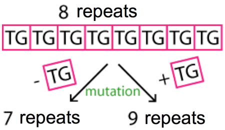


- Mutations do not fit SMM,
  indels of more than one repeat often occur

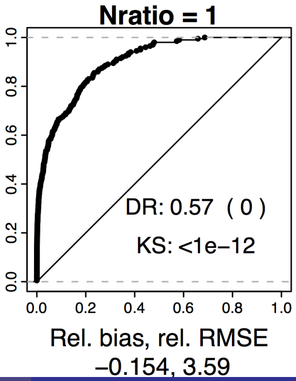
- Better mutation model = Generalized Stepwise Model (GSM)
  indels of X (geometric) repeats
  commonly found value in "natura" :
  $pGSM \approx 0.22$

OnePopVarSize : mis-specification of mutation processes

Microsatellite markers show complex mutation processes



- Mutations do not fit SMM, indels of more than one repeat often occur
- Better mutation model = GSM indels of X (geometric) repeats commonly found value in "natura" : $pGSM \approx 0.22$
- Problem : Analyses under the SMM of data simulated under a GSM in a stable population often show signs of bottleneck (57% of false detection with $pGSM = 0.22$)



OnePopVarSize : mis-specification of mutation processes

Solution : bottleneck model includes GSM (work with P. Pudlo)

- One more parameter (pGSM) ⇒ 4 param. to infer
- Longer runs are needed because of larger param. space

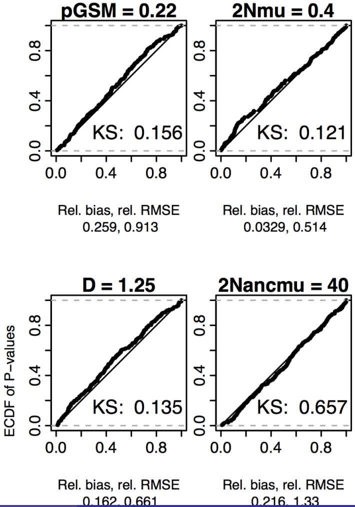
OnePopVarSize : mis-specification of mutation processes

Solution : bottleneck model includes GSM (work with P. Pudlo)

- One more parameter (pGSM)
- Longer runs are needed

results : inferences under a GSM

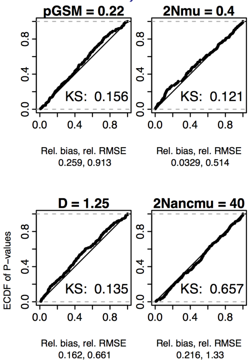
- pGSM infered with limited precision
- Other parameters well inferred
- Not much loss of precision with GSM vs. SMM



OnePopVarSize : mis-specification of mutation processes

Solution : bottleneck model includes GSM (work with P. Pudlo)

- One more parameter (pGSM)
- Longer runs are needed



results : inferences under a GSM

- pGSM infered with limited precision
- Other parameters well inferred
- Not much loss of precision with GSM vs. SMM

- GSM itself too simple: persistent mis-specification of mutational model
- Either robust model or need to consider other type of data, such as SNPs

OnePopVarSize : conclusions and perspectives

- Very efficient for bottleneck detections
- Accurate inferences for most demographic scenarios
- Relatively robust to fine scale population structure (i.e. local IBD)
- Much faster and more accurate than the MCMC equivalent (MsVar)

But :

- Not robutst to mutational processes
- Not robust to large scale population structure (e.g. island structure)
- Inaccurate for very strong demographic disequilibrium situations

what remains to do :

- Distinguishing between immigration and pop. size variation
- Adapting IS for disequilibrium models (not an easy task...)

Overview

1. Introduction
2. Algorithms used in Migraine
3. Demographic & mutational models
4. The Migraine software
5. Demo
6. Simulation studies
   - Inferences under Isolation By Distance
   - Inferences of past changes in population sizes
7. Future directions
8. Conclusions

Perspectives

Mutational models being currently implemented :

- Short DNA sequences (ISM)
- SNPs

Perspectives

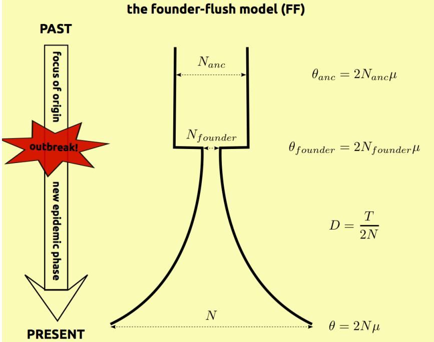
Mutational models being currently implemented :

- Short DNA sequences (ISM)
- SNPs

Demographic models which we plan to implement "shortly":

- Founder-Flush

(first tests are running, see Poster 15)



Perspectives

Mutational models being currently implemented :

- Short DNA sequences (ISM)
- SNPs

Demographic models which we plan to implement “shortly”:

- Founder-Flush
- Pure divergence 2-4 populations (with C. Beeravolu)
- Isolation with Migration 2-3 populations (with C. Beeravolu)
- Island population structure with past size variations
- IBD in two habitats (ecological barrier)(with A. Coulon)
- IBD with a geographic barrier (with A. Coulon)

Overview

1 Introduction

2 Algorithms used in Migraine

3 Demographic & mutational models

4 The Migraine software

5 Demo

6 Simulation studies
- Inferences under Isolation By Distance
- Inferences of past changes in population sizes

7 Future directions

8 Conclusions

MIGRAINE general conclusions

Very encouraging results

- Relatively easy to use (iterative analyses, no need for fine tuning)
- Reasonnable computation times (3h to 3 days for a classical data set),
  except for large IBD and strong disequilibrium
- Easy to paralellise
- Competitive compared to "MCMC-coalescent-based" approaches

some limits

- Strong bias on $N\mu$ for very strong disequilibrium situations
- Limited number of parameters

...more and more models will be added, be patient...

MIGRAINE general conclusions

Very encouraging results

- Relatively easy to use (iterative analyses, no need for fine tuning)
- Reasonnable computation times (3h to 3 days for a classical data set), except for large IBD and strong disequilibrium
- Easy to paralellise
- Competitive compared to “MCMC-coalescent-based” approaches

some limits

- Strong bias on $N\mu$ for very strong disequilibrium situations
- Limited number of parameters

...more and more models will be added, be patient...

Thank you for your attention