# Maximum likelihood inference comparing sequence and allelic markers in a population of variable size: an Importance Sampling approach

Champak Beeravolu Reddy, Francois Rousset, Pierre Pudlo, Raphaël Leblois

# Maximum likelihood inference comparing sequence and allelic markers in a population of variable size: an Importance Sampling approach

**Champak Beeravolu Reddy[1], François Rousset[2], Pierre Pudlo[1,3] & Raphaël Leblois[1]**

[1] **CBGP**, *INRA, Campus International de Baillarguet CS 30016 34988 Montferrier-sur-Lez cedex*
[2] **ISEM**, *CNRS, Université de Montpellier II - CC 065 34095 MONTPELLIER Cedex 05*
[3] **I3M**, *Université de Montpellier II - CC 051 34095 MONTPELLIER Cedex 05*

## Introduction & Objectives

Building on the **Griffiths & Tavaré (1994a,b,c)** approach, **Stephens & Donnelly (2000)** introduced an efficient importance sampling scheme for computing the likelihood of a genetic sample. Their method consists of approximating the conditional probability of the allelic type of an additional gene given those currently in the sample. This technique was further extended by **De Iorio & Griffiths (2004a,b)** to a subdivided population framework for various mutation models between gene types.

For sequence loci, the results of **De Iorio & Griffiths (DIG)** represent an improvement in terms of efficiency over those of **Bahlo & Griffiths (2000)** but this hasn't been tested on more than a single panmictic population. DIG consider the DNA sequences to be evolving under an **Infinitely-many Sites Mutation model (ISM)**. For allelic loci (microsatellites) evolving under a **Step-wise Mutation model (SMM)**, we use the Importance Sampling proposal distributions of **De Iorio *et al.* (2005)**.

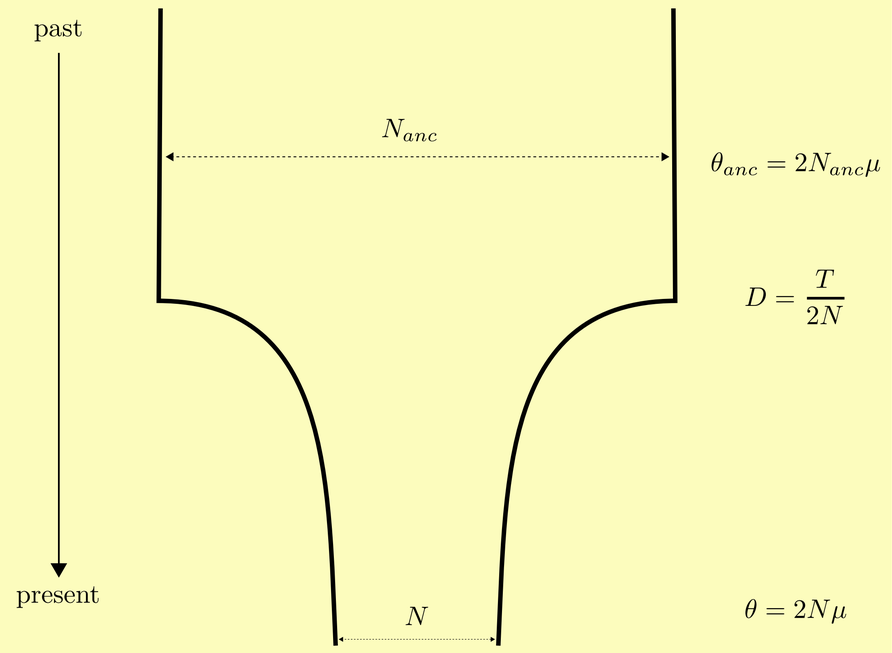
Here, we used results from **Griffiths & Tavaré (1994c)** in order to implement the case of a single exponentially contracting population (*see box: The demographic model*) and contrasted the performance of estimation from sequence and allelic markers. This approach gave us an insight into the potential complementarity of the considered genetic markers for improving the precision of the demographic and mutational estimates.

## Software & Simulations

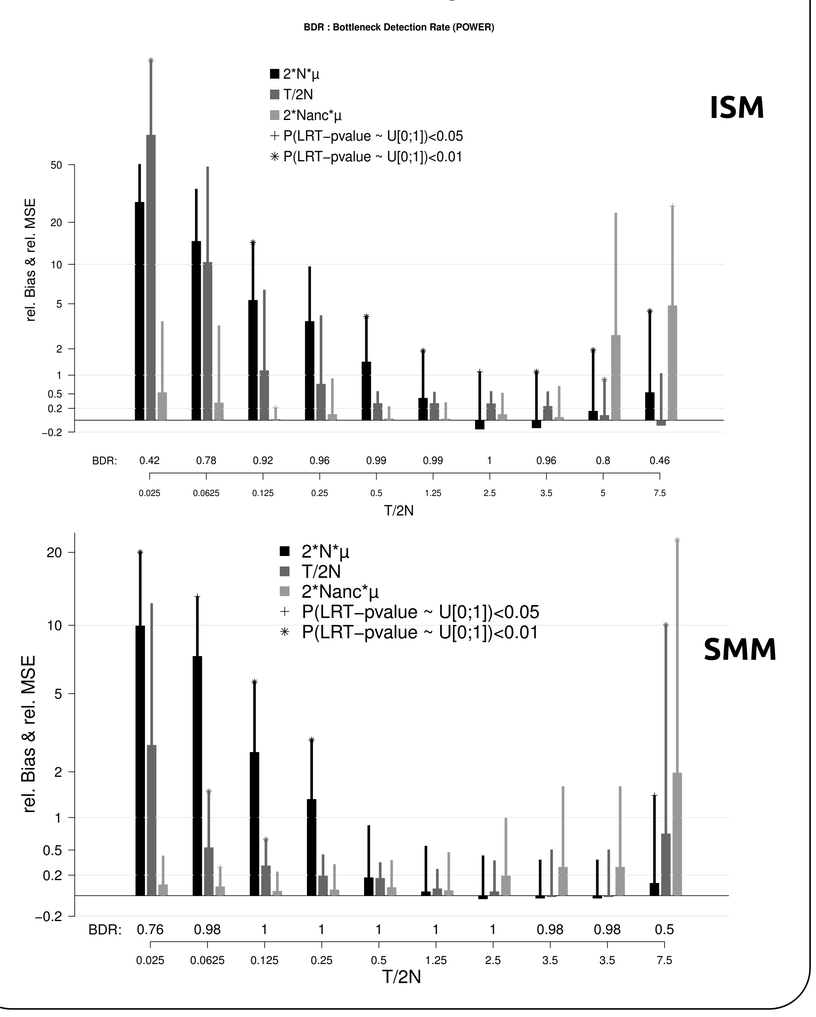
We have developed software which simulate data in the well known Genepop (for allelic data) and the Nexus (for sequence data) formats. These simulations were generated using **IBDSim** (http://raphael.leblois.free.fr/#softwares) which follows an exact coalescent algorithm. This data was then analysed by the **Migraine** software (http://kimura.univ-montp2.fr/~rousset/Migraine.htm) for inferring the demographic and mutational parameters (using DIG's method). Note that the current version of Migraine can also handle more complex demographic models than the one presented here such as **Isolation by Distance (IBD)** based on DIG's method.

We choose a simulation strategy which consisted of varying the time since the contraction began ($D = T/2N$) keeping the other parameters constant. A total of 10 demographic scenarios were thus obtained, each of which was repeated 200 times. The initial size of the population was set to 200 000 haploid individuals which progressively diminishes down to 200 individuals. Each sample consists of 100 individuals genotyped at 10 loci.

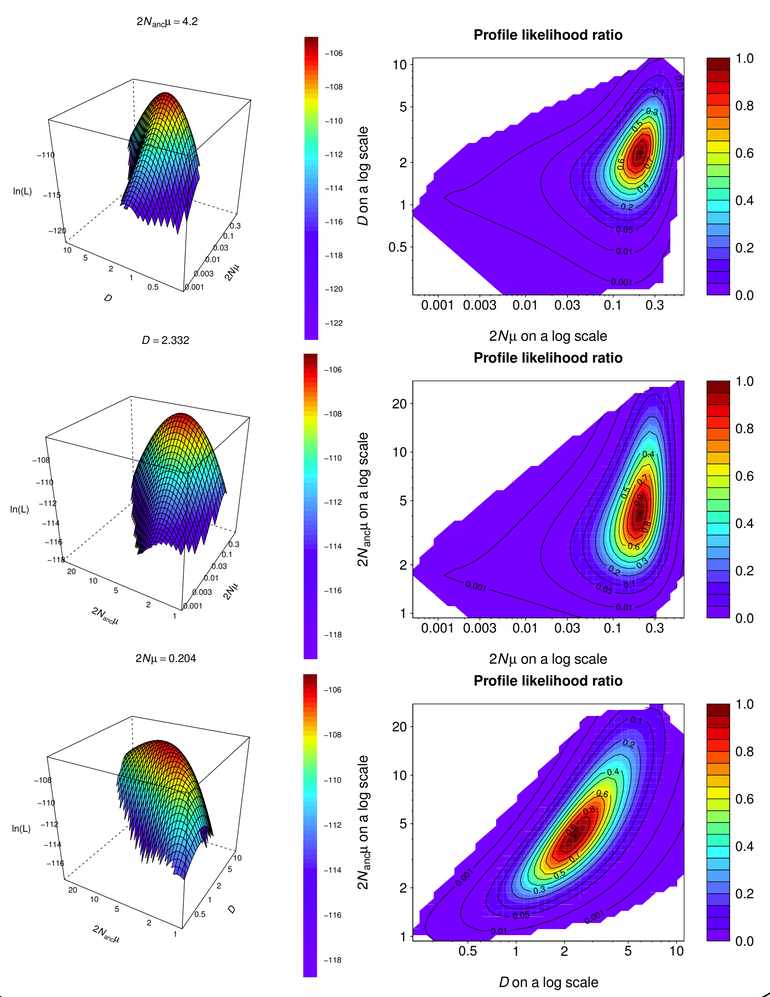
## The demographic model : a single exponentially contracting population




## Overall estimation performance




## Example of Migraine's ouput (ISM, $D$ = 1.25)




## Differences between markers

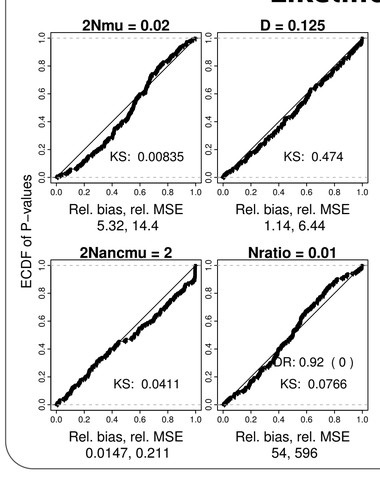
We accounted for the differences in the evolution of the genetic markers as follows: for microsatellites under a SMM, we assumed a mutation rate of $10^{-3}$ while for DNA sequences under an ISM we choose $5\times10^{-5}$**/sequence**. The simulated data had the following number of haplotypes/alleles ($N_a$) and expected heterozygosity ($H_{exp}$) for the cases presented here:

**ISM**
$D = 1.25$ : $N_a = 2.6$, $H_{exp} = 0.39$
$D = 0.025$ : $N_a = 7.44$, $H_{exp} = 0.65$
**SMM**
$D = 1.25$ : $N_a = 4.4$, $H_{exp} = 0.58$
$D = 0.025$ : $N_a = 13.4$, $H_{exp} = 0.87$

## An extreme scenario (ISM, $D$ = 0.025)

Very recent and extreme contractions result in flat profile likelihoods (*right*) or saddle point like situations (*left*). Here, the surfaces indicate that the markers have little information on the bottleneck and thus the current population size. This results in very wide confidence intervals for the parameters $D$ and $N$. Finally large $N$ values are confounded with the true values of $N_{anc}$ resulting also in wide confidence intervals.


## Likelihood ratio tests (LRT)




LRT's were performed on each of the estimated parameters for a given bottleneck scenario. This consists of calculating 200 LR statistics (one for each repetition) using the likelihood values at the true parameters.

As the LR statistic approximately follows a chi-square distribution, we calculated the corresponding p-values. The figures on the left plot the empirical cdf of the obtained p-values. This helps us check whether the distribution of the p-values for each parameter is uniform.

## Results & Conclusions

The bottleneck scenario implemented here:
1° was successfully detected over a wide range of scenarios;
2° was inferred with reasonable precision for intermediate scenarios. The SMM estimates performed better than their ISM counterparts but this can be understood in terms of heterozygosity and the number of alleles/haplotypes present in the simulated data (*see box: Differences between markers*);
3° the time taken in comparision to typical "MCMC" approaches was very short (~4hrs for a single repetition). Slightly longer runs improve estimation performance (*not shown here*);
4° for all scenarios, the LRT's indicate a very good conformity of the likelihood surface generated by Migraine to the true likelihood.

## Perspectives

The current version of Migraine includes implementations of 1D and 2D **IBD** models (including the island model as a subcase), as well as the two-populations model described in De Iorio *et al.* (2005) and the elementary one-population model for allelic data. We are currently testing novel Importance Sampling proposals for the **Generalized Stepwise Mutation model (GSM)** and DIG's algorithms for sequence data (ISM). Other mutation models such as **Single Nucleotide Polymorphisms (SNP's)** will be available shortly. Planned developments also include **Isolation with Migration** models and **continent-island** models. Finally, a user friendly graphical interface is also underway in order to ease the use of Migraine (*no pun intended*).

## References:

Griffitths, R.C. & Tavaré, S. (1994a) Simulating probability distributions in the coalescent. Theor. Popn. Biol., 46,131-159.
Griffitths, R.C. & Tavaré, S. (1994b) Ancestral inference in population genetics. Stat. Sci., 9, 307-319.
Griffitths, R.C. & Tavaré, S. (1994c) Sampling theory for neutral alleles in a varying environment. Phil. Trans. R. Soc. Lond. B, 344, 403-410.
Bahlo, M. & Griffiths, R.C. (2000) Inference from gene trees in a subdivided population. Theor. Popn. Biol. 57, 79-95.
Stephens & M., Donnelly, P. (2000) Inference in molecular population genetics. J. Roy. Statist. Soc. B 62, 605–655.
De Iorio & M. & Griffiths, R.C. (2004a) Importance sampling on coalescent histories, I. Adv. Appl. Probab. 36, 417–433.
De Iorio M. & Griffiths, R.C. (2004b) Importance sampling on coalescent histories, II. Subdivided population models. Adv. Appl. Probab. 36, 434-454.
De Iorio M, Griffiths R.C., Leblois R. & Rousset F. (2005) Stepwise mutation likelihood computation by sequential importance sampling in subdivided population models. Theor Popul Biol. 68, 41–53.
Leblois R., Estoup A. & Rousset F. (2009) IBDSIM: a computer program to simulate genotypic data under isolation by distance. Molecular Ecology Resources, 9, 107–109.

## Affiliations & Sponsors:

CBR & RL are supported by the IM-MODEL@CORALFISH project while CBR, FR, PP & RL are supported by the EMILE project, both funded by the French National Research Agency.