Why (how) does effective population size change ?

L. Chikhi, O. Mazet, W. Rodriguez

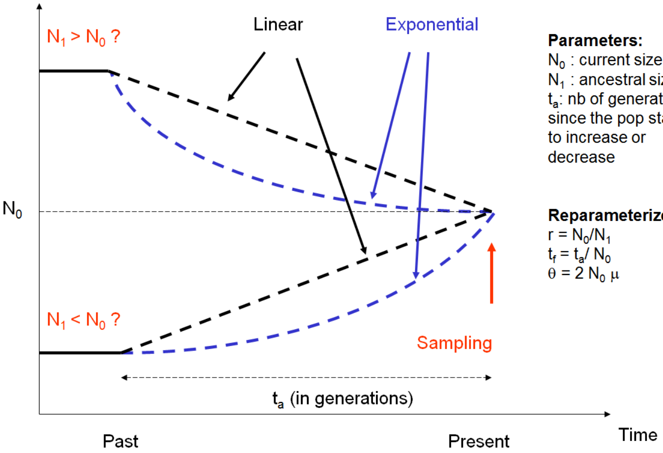
Institut Mathématique de Toulouse
Evolution et Diversité Biologique (Toulouse)
Instituto Gulbenkian de Ciência (Lisboa)

December 12, 2013

Problematic : Lounès Chikhi's talk in Bern, march 2012. Lounès caught Microsatellite data for Orangutan in Borneo. He used a model developped by Beaumont and Storz in 2002 :

Testing for Genetic Evidence of Population Expansion and Contraction: An Empirical Analysis of Microsatellite DNA Variation Using a Hierarchical Bayesian Model (*Evolution*)

Model of Beaumont 1999 and Storz et Beaumont, 2002



Parameters:
N_0 : current size
N_1 : ancestral size
t_a: nb of generations since the pop started to increase or decrease

Reparameterize:
$r = N_0/N_1$
$t_f = t_a/ N_0$
$\theta = 2 N_0 \mu$

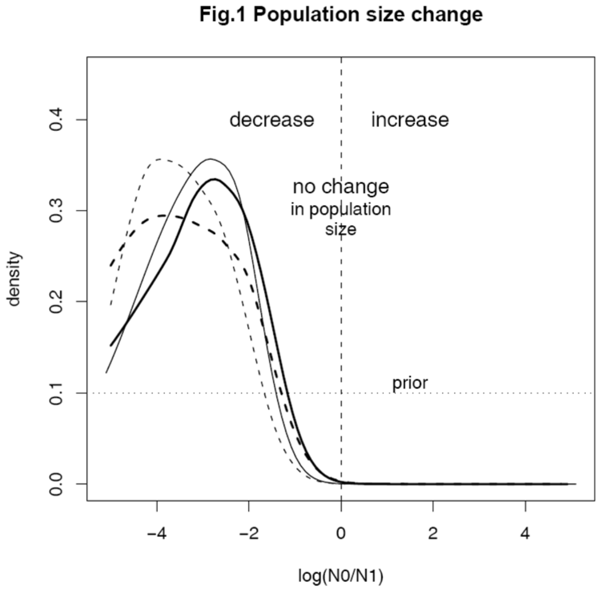
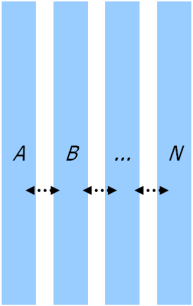
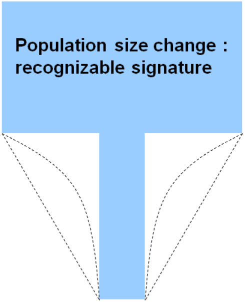


Fig.1 Population size change
decrease
increase
no change
in population
size
prior
density
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
log(N0/N1)

GENETIC DIVERSITY





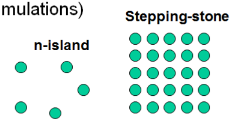
Population size change : recognizable signature

PROBLEM :

STRUCTURED POPULATIONS GENERATE A SIMILAR SIGNATURE

Effect of population structure on bottleneck signals

- Models of population structure (100 demes in all simulations)
  - n-island model (100 islands)
  - Stepping-stone (10 x 10) (toroidal)



- Parameters used
  - Stepwise mutation model assumed to simulate data
  - F_{ST} values used { 0.01 ; 0.05 ; 0.1 ; 0.25 }
  - θ values used { 1; 10 }
  - Number of loci { 5 ; 20 }
  - 50 individuals sampled (100 genes)

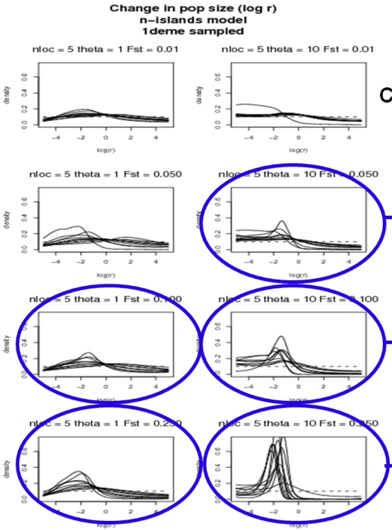
Differentiation

Diversity
(mutation and pop size)

- Sampling schemes :
  - n-islands model: samples from 1, 2, 10 and 50 demes
  - Stepping stone model: samples from 1, 2 neighbouring and 2 distant demes

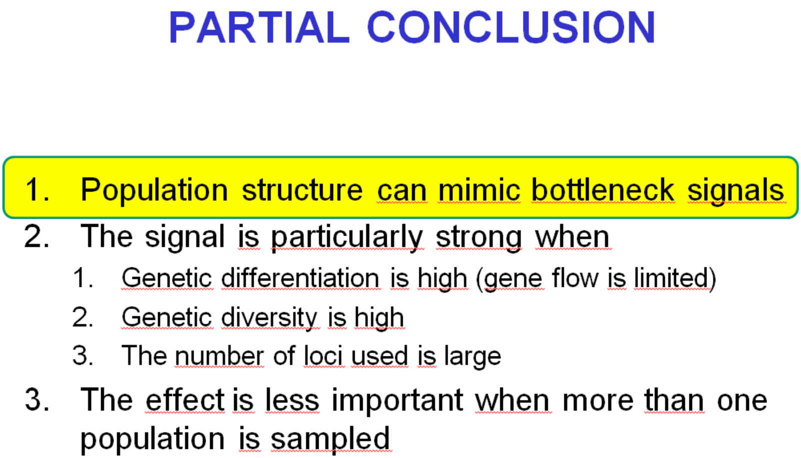
- 10 independent data sets for each parameter set (except 20 loci and 10 demes)

Effect of population structure on bottleneck signals



Can we separate population structure from population crash?

Bottleneck signals



PARTIAL CONCLUSION

1. Population structure can mimic bottleneck signals
2. The signal is particularly strong when
    1. Genetic differentiation is high (gene flow is limited)
    2. Genetic diversity is high
    3. The number of loci used is large
3. The effect is less important when more than one population is sampled

PARTIAL CONCLUSION

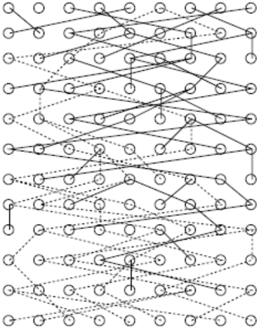
Need to

develop methods that can separate these two kinds of scenarios (structure *versus* bottleneck)

Modelling :

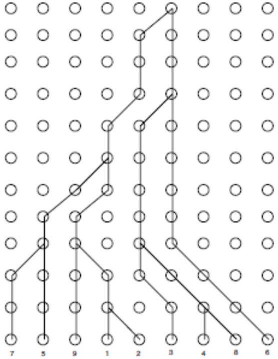
1. Wright-Fisher Model, rescaled continuous times to Kingman coalescent
2. Variable Population Size
3. Symmetric Island Model

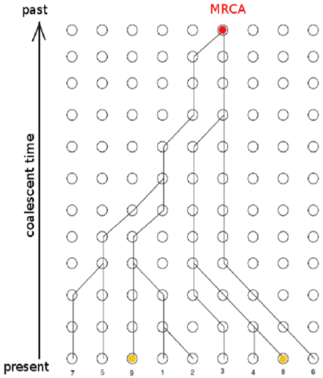
Hypotheses of Wright-Fisher Model :



- Asexual reproduction
- Constant Population Size
- Panmictic reproduction
- No selection
- Non overlapping generations

We now keep the Ancestors of the present population, and we can reach the Most Recent Common One (MRCA)





If T_k is the time (number of generations) needed to reach backward in the past the MRCA of k individuals, we have

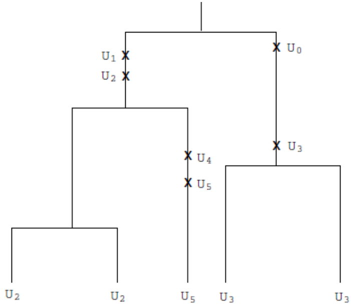
$$P(T_2 > n) = \left(1 - \frac{1}{N}\right)^n$$

then rescaling time when N is great

$$P(T_2 > [Nt]) = \left(1 - \frac{1}{N}\right)^{[Nt]} \simeq e^{-t}$$

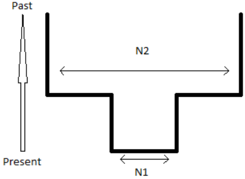
So $T_2 \sim \mathcal{E}(1)$, and similarly $T_k \sim \mathcal{E}\left(\frac{k(k-1)}{2}\right)$

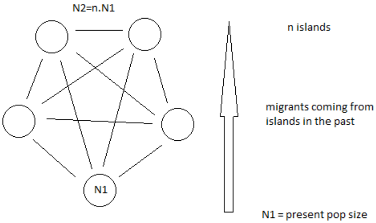
We obtain coalescent, on which we can put mutation events, each one occurring on each lineage with a time $\mathcal{E}\left(\frac{\theta}{2}\right)$ where $\theta = 2Nu$, u being the (small) probability of mutation by generation by gene, and each one producing new allele (Infinite Sites Model)



Looking for alleles (differences, number, repartition...) in the present will give the **information** about the ancestral tree (height, total length, shape...)

We will focus on the MRCA of 2 individuals T_2.
Question : Is it true that T_2 have a similar distribution if demographic population history underwent change of population size, or if population is embedded in a geographical structure ?



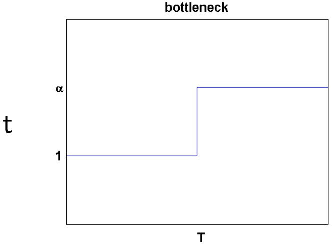


Let's have a first look on the mean and the variance.

Variable population size : $N(k)$ number of individuals in k-th generation before present ; let $f_N(x) = \frac{N([Nx])}{N(0)}$, assume $f_N(x) \to f(x) > 0$ for all x, and let $\lambda(x) = \frac{1}{f(x)}$. We then obtain

$$\mathbb{P}(T_2 > t) = e^{-\int_0^t \lambda(x)}\, dx$$

Simplest case : sudden bottleneck



$$\begin{aligned} \mathbb{E}(T_2) &= 1 + e^{-T}(\alpha - 1) \\ Var(T_2) &= 1 + 2Te^{-T}(\alpha - 1) \\ &\quad +2\alpha e^{-T}(\alpha - 1) \\ &\quad -(\alpha - 1)^2 e^{-2T} \end{aligned}$$

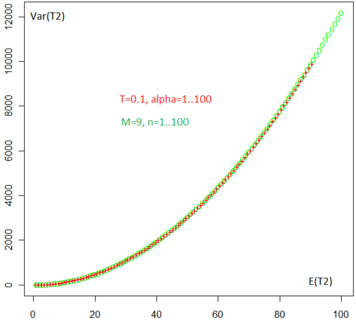
Geographical Structure : Symmetric Island Model. n islands of size N with constant scaled migration rate $M = 2Nm$. Thanks to $p_s(\theta)$ the probability that 2 lineages are identical by descent when they are picked from the *same* island, we can easily compute, since $p_s(\theta)$ is in fact the Laplace transform of the T_2 of those sampled two individuals :

$$\mathbb{E}(T_2) = n$$

and

$$Var(T_2) = n^2 + \frac{2(n-1)^2}{M}.$$

And indeed, for some range of parameters, it's not easy to distinguish geographical structure from variable population size



So we'll have to look further in the distributions

Variable Population Size, with a sudden bottleneck :

$$\boxed{f_{T_2}^{VPS}(t) = e^{-t}\mathbb{I}_{[0,T[}(t) + \frac{1}{\alpha}e^{-T-\frac{1}{\alpha}(t-T)}\mathbb{I}_{[T,+\infty[}(t)}$$

Symmetric Island Model, inversing Laplace transform $p_s(\theta) = \frac{a}{\theta+\alpha} + \frac{1-a}{\theta+\beta}$:

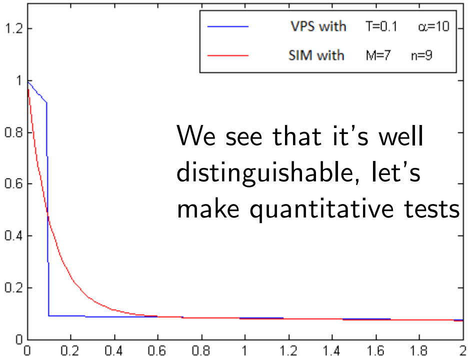
$$\boxed{f_{T_2}^{SIM}(t) = ae^{-\alpha t} + (1-a)e^{-\beta t}}$$

with

$$\alpha, \beta = \frac{1}{2}\left(1 + \frac{n}{n-1}M \pm \sqrt{\left(1 + \frac{n}{n-1}M\right)^2 - \frac{4M}{n-1}}\right),$$

$$a = \frac{1}{2} + \frac{1 + \frac{n-2}{n-1}M}{2\sqrt{\left(1 + \frac{n}{n-1}M\right)^2 - \frac{4M}{n-1}}},$$

Distribution comparison



We see that it's well distinguishable, let's make quantitative tests

Statistical tests with simulated data : $(X_i)_{i=1\dots K}$ iid following $f^{SIM}(M,n)$ (or $f^{VPS}(T,\alpha)$).

1. Estimation of (T,α) and (M,n) maximizing the likelihood of the data
2. Adequation test (KS) under $\mathcal{H}^{VPS}(\hat{T},\hat{\alpha})$ then $\mathcal{H}^{SIM}(\hat{M},\hat{n})$

Question : for a given couple (M,n) (or (T,α)), how great K has to be in order to be able to significatively choose the right model ?

Some results for $f^{SIM}(M, n)$ which is the most interesting in our problematic. Tested parameters : $M = 0.1, 0.2, 0.5, 1, 5, 10, 20, 50$ and $n = 2, 4, 10, 20, 50, 100$; 100 simulations for each couple.

- $K = 20$: no success at all.
- $K = 50, 100$: a tiny thrill with a high reject threshold (10%) : more than 50% (80% for $K = 100$) of success for $(M, n) = (0.5, 20), (1, 50), (1, 100)$...
- $K = 200, 500$: good then very good results for "good" parameters (roughly $M \leq 5$)
- $K = 1000$ is sufficient for most range ($M \leq 10$ or $M \leq 20$ for $n \geq 10$) even with strict threshold
- $K = 10000$ only needed to reach $M = 50$)

BUT T_2 values are not directly observable...

Number of mutations $\mathbb{P}(N = k | T_2 = t) = e^{-2t\mu}\frac{(2t\mu)^k}{k!}$ with μ mutation rate of the locus. Hence we can compute

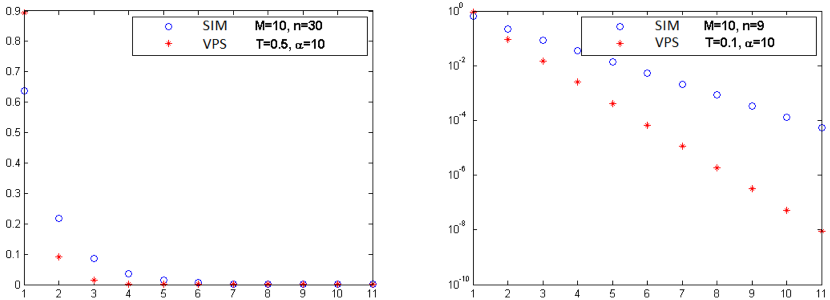
$$\mathbb{P}(N^{SIM} = k) = \frac{a}{\alpha + 2\mu}\left(\frac{1}{1+\frac{\alpha}{2\mu}}\right)^k + \frac{1-a}{\beta + 2\mu}\left(\frac{1}{1+\frac{\beta}{2\mu}}\right)^k$$

and

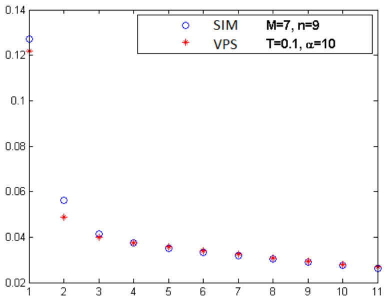
$$\mathbb{P}(N^{VPS} = k) = \frac{(2\mu)^k}{(2\mu+1)^{k+1}} +$$

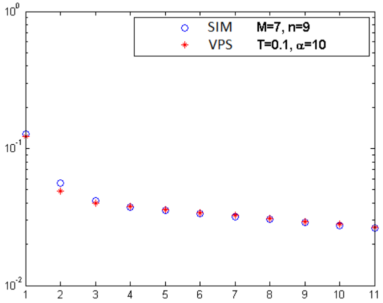
$$(2\mu)^k \sum_{i=0}^{k} \frac{e^{-T(2\mu+1)}T^{k-i}}{(k-i)!}\left(\frac{1}{\alpha(2\mu+\frac{1}{\alpha})^{i+1}} - \frac{1}{(2\mu+1)^{i+1}}\right)$$

Example with distinguishable parameters, distribution of number of mutations (semi-log scale on the right)



Example where it will be more difficult





Quantitative tests still to do !
And tests on real data at last !

Perspectives

- Other observable data (Microsat, allele repartition, SNPs repartition...)
- Other variations of population sizes (linear, exponential, several bottlenecks...)
- More information with T_k : explicit formulas ?

Akwnoledgments to Simon Boitard and Simona Grusea

Merci pour votre attention