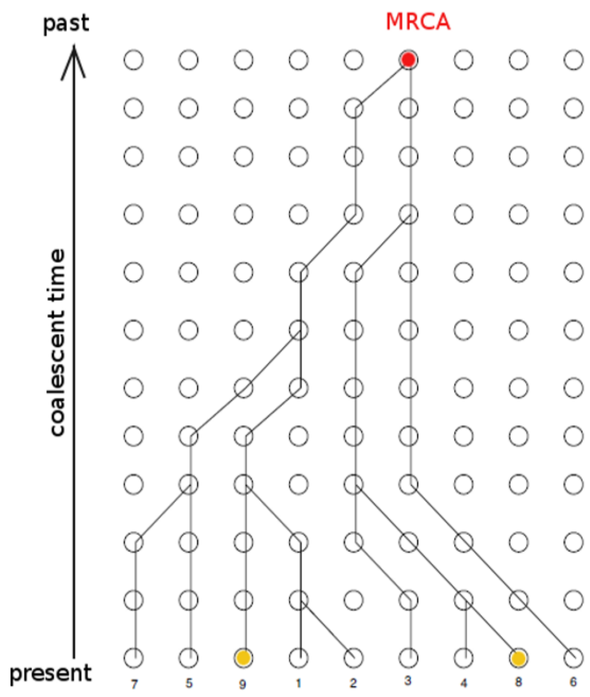
Exploring the estimation quality of PSMC using different sizes of input sequences.

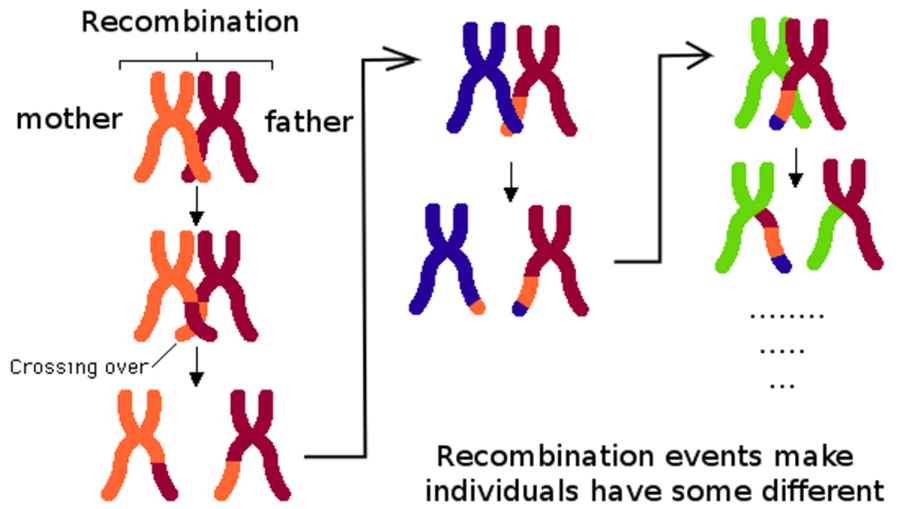
Rodríguez Willy, Chikhi Lounès, Mazet Olivier, Boitard Simon, Grusea Simona.

One of the aims of population genetics is to reconstruct key aspects of the demographic history of populations, such as expansions and contractions (bottlenecks). Therefore, most of population geneticists have used independent loci genotyped in several populations (from 10 up to a few hundreds of loci). Recently, Next Generation Sequencing (NGS) technologies have provided the entire genome for single individuals. Besides, the Pairwise Sequentially Markovian Coalescent model (PSMC) has been developed to infer the recent demographic history of populations, using the patterns of recombination and mutation in the genome of single diploid individuals.

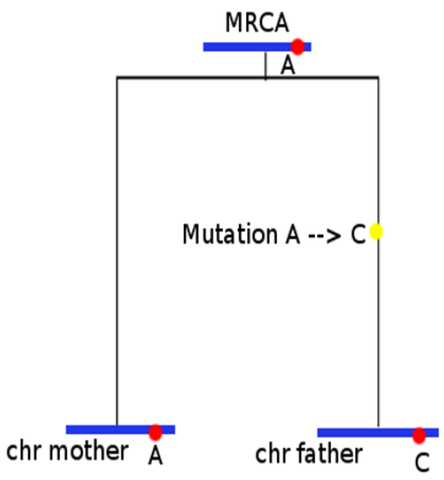
Most Recent Common Ancestor (MRCA)
Coalescence Time. TMRCA.



Recombination & Population History

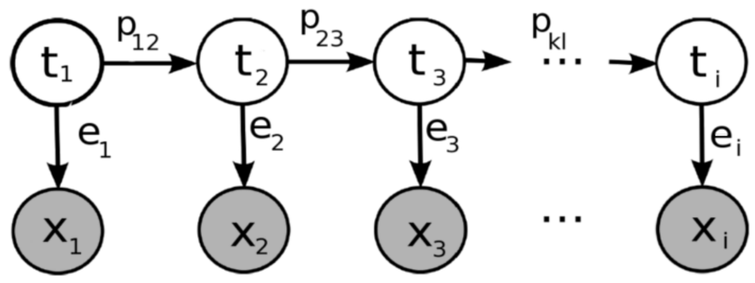


Heterozygous ⇒ Mutation

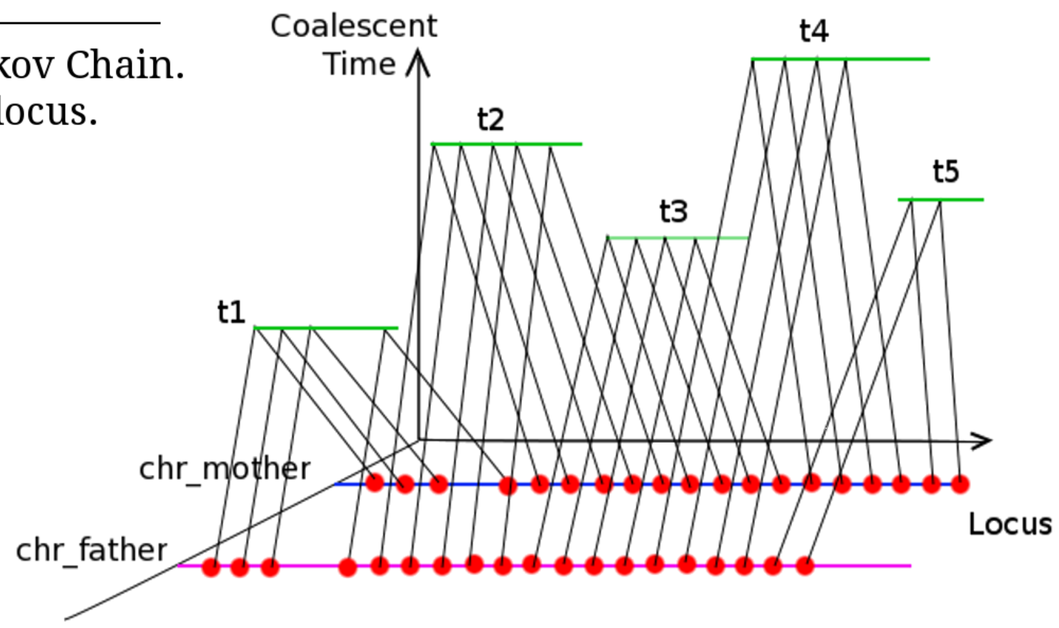


Hypothesis: The values of "t" are a Hidden Markov Chain. The observed states are the mutations at every locus.

Hidden States: Discretized Coalescent Time. T={t1, t2, ..., tn}



Observed States, homozygous, heterozygous. X={0, 1}



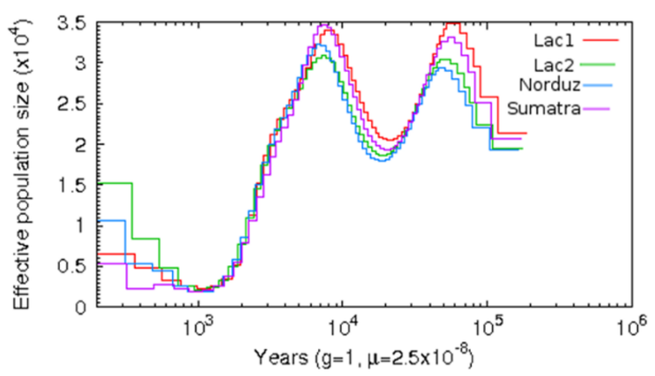
N(t) population size at time t

continuous case: $N(t) = \lambda(t)N_{ref}$

discrete case: $N_k = \lambda_k N_{ref}$

Objective: finding the values of λ_k

PSMC application to the full genome of sheeps



There are many species for which the entire genome is not available. What happens to the estimate of PSMC when we have less data input ?

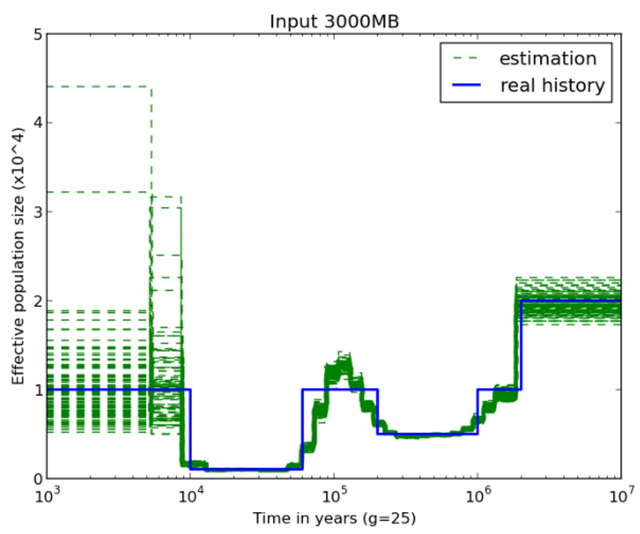
Simple exploratory analysis:

scaled mutation rate	0.001
bin size (S)	100
MS command	ms 2 100 -t 30000 -r 6000 30000000 -eN 0.01 \ 0.1 -eN 0.06 1 -eN 0.2 0.5 -eN 1 1 -eN 2 2
PSMC command	psmc -p 4+25*2+4+6 -t 15 -N 25 -r 5
Input sizes (in Megabases)	0.25, 0.5, 0.75, 1, 2.5, 3, 4, 5, 10, 25, 40, 50, 100, 250, 300, 400, 500, 1000, 2000, 3000

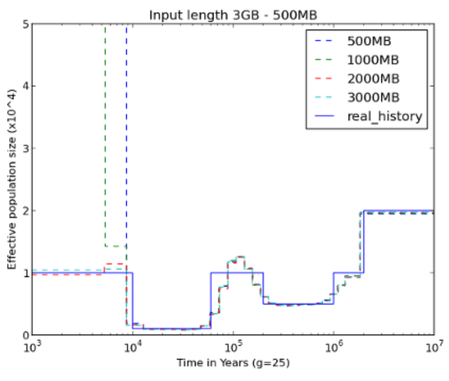
Let be [x0, x1] some time interval. N0 the simulated history and N1 the estimated history. We compute the accuracy. (Li, H., & Durbin, R. 2011. Supplementary Information) by:

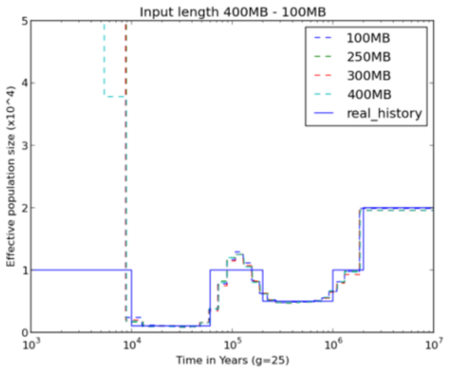
$$d(x_0, x_1) = \frac{1}{logx_1 - logx_0} \int_{x_0}^{x_1} \frac{|N_0(x) - N_1(x)|}{N_0(x) + N_1(x)} \frac{dx}{x}$$

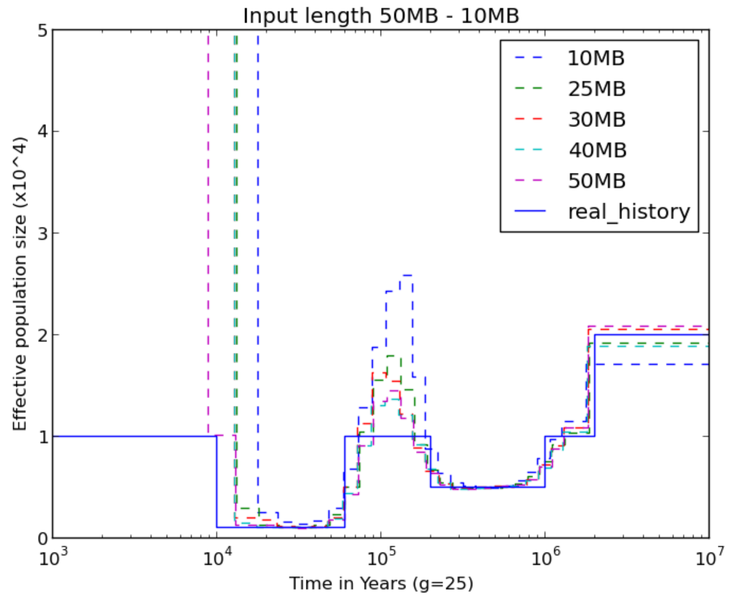
100 independent experiments

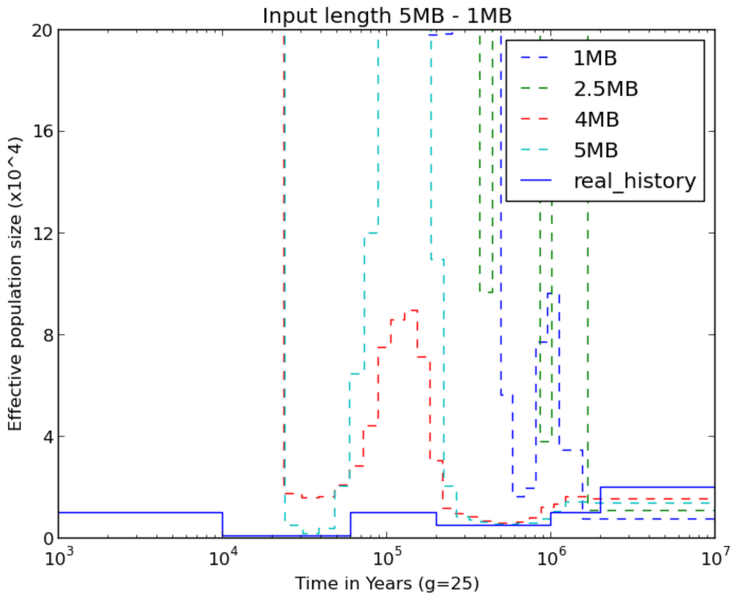


Average over the 100 data sets for each case



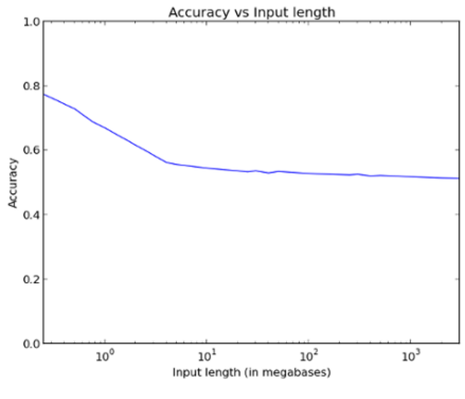






Accuracy = d(20kya, 2Mya)

For the average of the estimations in each case, we compute the accuracy as d(20kya, 2Mya)



In this experiment we can see that the accuracy of the method decreases when the size of the input data is reduced. However, we can obtain acceptable estimates, **even with 20 Megabases**. Under 10 MB, precision is rapidly lost.

Contact information:
willyrv@insa-toulouse.fr
http://willyrv.com