Exploration of sampling techniques

A contribution to the detection recent population decline using recombination events

Keywords : Population dynamics, genetic recombination, genomics, ecology

November 2021

Alix de Thoisy
Master 2 – Bio-informatics, Université de Paris

under the supervision of Thomas Forest,
Dr. Guillame Achaz &
Dr. Amaury Lambert
Stochastic Modelisation for the Inference of Life Evolution unit, Collège de France

Table of Contents

I - Introduction......3
II - Context......3
II – 1. Population decline using recombination events......3
II – 2. MLD blocks......4
II – 2. a) The four-gametes test......4
II – 2. b) Chopping and testing......5
II – 3. So far results......5
III - Problematic & solution proposed......5
VI - Implementation......7
VI – 1. Cut in MLD blocks......7
VI – 2. Time-measurement algorithm......7
VI – 3. Workflow......8
V - Results......8
VI - Limitations and perspectives......9
VII - References......9

I - Introduction

Industrial society is responsible for a rapid extinction of fauna and flora[1] species leading to an unprecented biodiversity crisis[2]. The decline of the species in a very few generations makes it difficult to monitor populations based on geographic dispersal or direct estimation of numbers of individuals, as used for Red List of the International Union for Conservation of Nature (IUCN)[3].

In order to reduce costs and allow monitoring of a larger number of species, genetic methods have been developed recently to quantify the genetic diversity of populations at different times and thus infer population sizes[4–6]. However, these approaches require time series data, which are challenging to obtain.

Methods have been proposed to use only a set of genomes collected at a single time point. They allow to detect variations in population size of the order of N_e generations in the past[7–9], N_e being the effective population size[10]. Markov chain approaches have been developed to infer recent variation but present a prohibitive computational cost for application to large sets of genomes[11].

II - Context

The work of the two-previous and the current PhD students at the *Stochastic Modelisation for the Inference of Life Evolution* unit focuses on detection of recent populations declines using the occurences of recombination events.

Genetic recombination, also called genetic reshuffling or crossing-over, is the exchange of genetic material between maternal and paternal chromosomes during meiosis, leading the offspring to carry traits that differ from either parent. In eukaryotes, it involves a pair of homologous chromosomes.

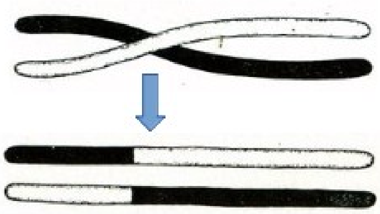

Figure 1: A simple representation of a crossing-over between chromosomes.

II – 1. Population decline using recombination events

Two segments of an alignment share the same coalescent tree[12] if no recombination event occured since their most recent commun ancester. Kerdoncuff and colleagues thus defined *Maximum Recombination Free (MRF)* blocks as maximal interval of sites with the same coalescent tree[13].

Classically, the length L of MRF blocks follows an exponential distribution with rate ρT, T being the total length of the alignment and ρ the recombination rate. For a constant recombination rate, the alignment length and the average length of MRF blocks are negatively correlated. Indeed, recombination event occur more often in deeper trees, their recombination-free blocs will be shorter. As the total length is proportional to the population size, large populations tend to have shorter MRF blocks too.

This dynamic is the core of the previously proposed methods : comparing the MRF block length distribution of a population of constant size to one that has undergone a size change by a factor of 1/κ at a time τ, the latter appears more dispersed, with more short and long blocks (Figure 2).

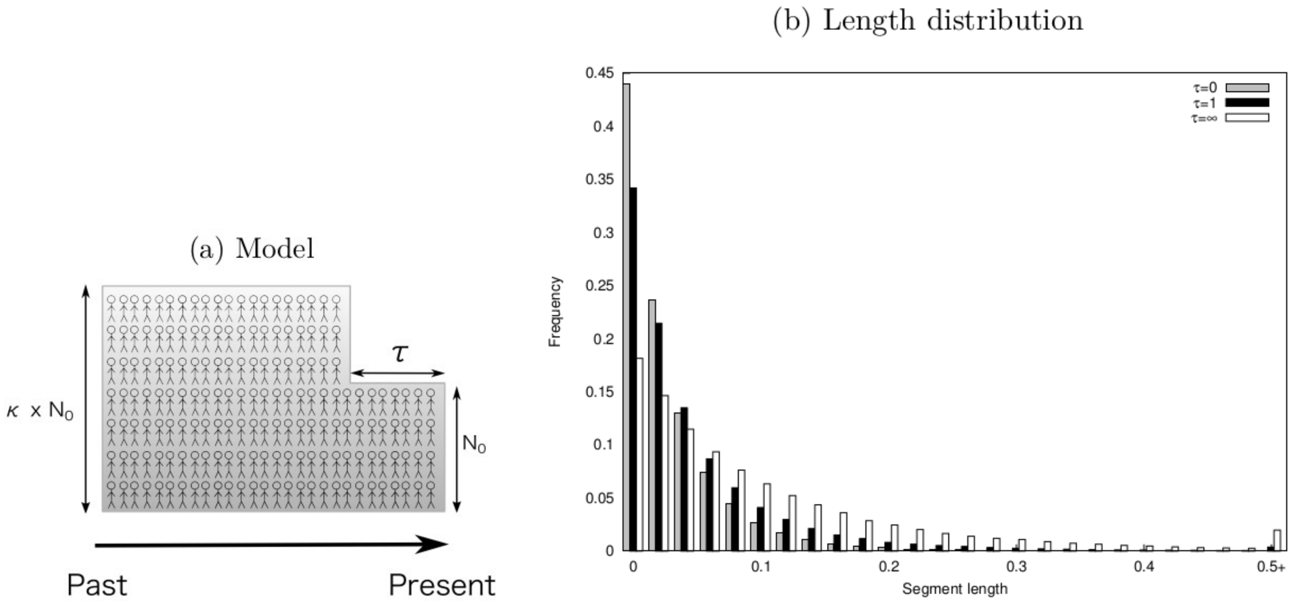

Figure 2 : Population model (a) and the corresponding distributions of MRF block lengths according to the time the size variaton happened (b), data simulated with msprime[14]. From Kerdoncuff, Lambert & Achaz, 2020.

II – 2. MLD blocks

One challenge is that recombination events are difficult to detect, leading them to use polymorphic site incompatibilities to place the minimal number of recombination events.

II – 2. a) The four-gametes test

The four-gametes test has been introduced in 1985 by Hudson and Kaplan to detect some recombination events[15]. It runs under the three hypothesis of the "infinite site model":

- the number of site able to mutate is infinite
- mutations can only occur once on a given position
- there are no recombination events.

Considering now two bi-allelics loci A/a - B/b, only three combinations of the alleles can be observed on a phylogenic tree (Figure 3). If four are observed, at least one hypothesis of the model is not respected. Given the short periods of time considered, two mutations on a same position is very unlikely, so it is in favour of a recombination event.

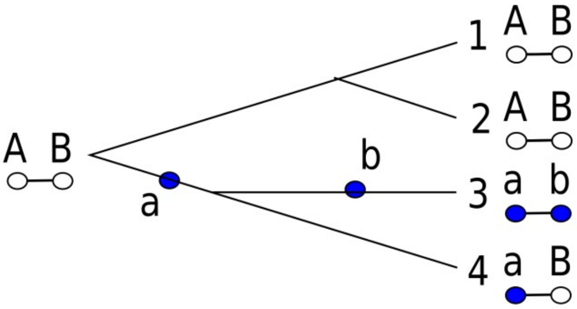

Figure 3: In this example, the combination A-b could only have been observed if a recombination event had occurred. Image from Kerdoncuff’s thesis work.

This test does not identify recombination events that do not modify the topology of a non-rooted tree. It will thus always be a lower estimation.

II – 2. b) Chopping and testing

The four-gametes test is run on 150-bases long segments along the genome, returning a list of pairwise incompatible polymorphisms, and an algorithm then chops the list to keep only the pairs of position required to explain the occurrence of recombination events. These events are placed in the middle and the interval between two are designated as *Maximum-linkage Desequilibrium (MLD)* blocks.

The team performed 1000 replicates and reported cases where the number of sifgnificantly larger and smaller blocks diverged from the distribution under the null hypothesis of constant population size.

Part of the current projet was to to reimplement this work, the code can be found on the Github respository.

II – 3. So far results

Kerdoncuff and colleagues ran this method on chromosome 1 of western lowland Gorillas (*Gorilla gorilla gorilla*) with a set of genomes from the Great Ape Genome Project[16]. With modifications to account for poor sequencing quality, they demonstrated the recent population decline, known from traditional monitoring methods.

III - Problematic & solution proposed

Tests on other data have shown the predominance of ancestral noise in genome-wide studies. This is because genomes have accumulated mutations from the early history of populations and these variations confound the inference of recent change. If a population has experienced a long period of growth and then a rapid recent decline, the former will take precedence and the methods developed will infer recent growth.

To overcome this problem, this project focuses on the detection of recent coalescent subtrees within a recombination-free block. The objective is to "catch" the set of sequences whose age of the most recent common ancestor in this segment is below a threshold (Figure 4).

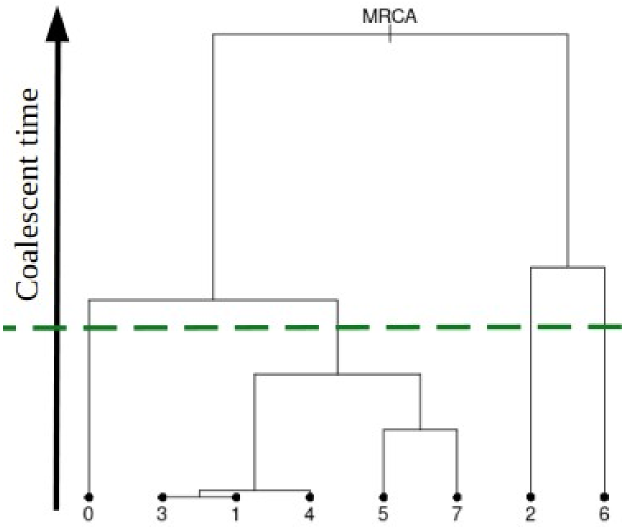

Figure 4 : A coalescent tree of 8 sequences, the age-threshold is represented by the dashed line in green. Here, sequences “3”, “1”, “4”, “5” and “7” would constitute a subtree of interest.

In regard of the nature of the data, two types of solutions can be considered.

The first one, similar to the one developed for the detection of population size variation by the lab team, could focus on the detection in deviations from a null hypothesis. This approach was particularly fruitful since it dispenses the necessity of costly computational method by using the comparison of large genetic trends at the population level, namely the on average shorter length of MLD blocks in large populations. This would face two major problems in our case, the identification of a significant trend that would allow dating and the size of the samples, the MLD blocks being of a few hundred bases while the previous analysis was done on whole genomes. It also makes the definition of a variable threshold more complicated.

The second type of solution is to propose a time estimation by numerical methods, which we explored in the following.

By calculation, only one element allows to link sequences and time: the mutation rate, or more precisely, the hypothesis that it is constant through time and along the genomes, assumed in the previous works. Determining the number of mutations within a sequence alignment can thus approximate the age of the most recent common ancestor.

Without falling back on phylogeny tools, it is not possible to get an idea (at reasonable cost) of the topology of the coalescent tree, not even with a very small number of sequences. The four-gametes test and the cutting in MLD blocks only tells us one things : that, within a block, a coalescent tree does exist.

This has two direct consequences:

- Polymorphic sites rather than mutations will be counted, since it is not possible to know where the mutation occurred (Figure 5). However, under the hypothesis of the infinite-site model, a locus cannot mutate twice, reducing our margin of error a bit.

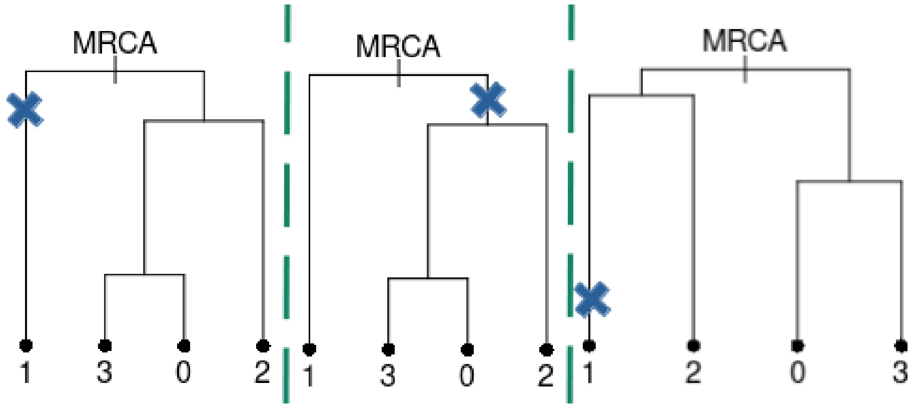

Figure 5 : Example of a case where, on one locus, the genome “1” carries a base different from the one shared by “0”, “2” and “3”. It is impossible to determine whether the mutation has occurred at a common ancester or on a single branch.

- It is necessary to assume an average topology for all trees, where the mutations are distributed in the same way. One could indeed imagine a population where mutations would accumulate in some individuals while they would be systematically counter-selected in others; the common ancestor inferred by the proposed method would be much younger than the actual one. However, under our assumptions of homogeneous and mixed populations, the approximation seems reasonable.

VI - Implementation

VI – 1. Cut in MLD blocks

The search for recent subtrees starts with the detection of MLD blocks by the methods proposed in the previous work. Since it uses the 4-gametes test for cutting, it is necessary to provide an input alignment of at least 4 sequences.

Part of the projet was to propose another implementation of it, which can be found on the Github repository and used alone. The program is implemented in two version :

- One that does not require the load of the entire alignment in memory and is thus suitable for large files. It can be run with the stand-alone main_alignment_to_mld.py script.
- One that does load the alignment in memory with the use of BioPython library[17]. Indeed, the computational cost of this algorithm is high and the computing time is often problematic before the RAM capacities. This version is then used in the workflow.

VI – 2. Time-measurement algorithm

The approximation of time is computed the simpliest possible way : the alignment is browsed and the polymorphic sites counted, the amount is then divided by the total length. This way, it does not return time in an directly interpretable metric but rather units of time.

It has a computing and time complexity of *O(T)*, *T* being the length of the alignment.

VI – 3. Workflow

Back to our overreaching goal, to detect all the possible recent subtrees of sequences in maxmimum linkage-desequilibrium.

A set of sequences aligned is first cut in MLD blocks, these blocks are then screened one-by-one, their "age" is returned. To have an acceptable running time, sequences of the global alignement are investigated four by four on their entire length.

V - Results

The workflow was run on a real-world data set of swallow (*Hironda rustica*) genomes. The samples were collected by Jérôme Fuchs in the 2010's through the veterinary school of Maisons-Alfort and sequenced by the Institut du Cerveau et de la Moelle at the Salpêtrière hospital in 2019 and 2020. The alignment was processed on European Bioinformatics Institute servers (EMBL-EBI) using the MAFFT algorithm[18].

The distribution of the MLD blocks lengths (Figure 6) has the expected exponantial-degrowth shape. No further interpreted is meant at this step, the cut in MLD blocs is necessary for continuing the work and the distribution plot only serves to verify the results so far.

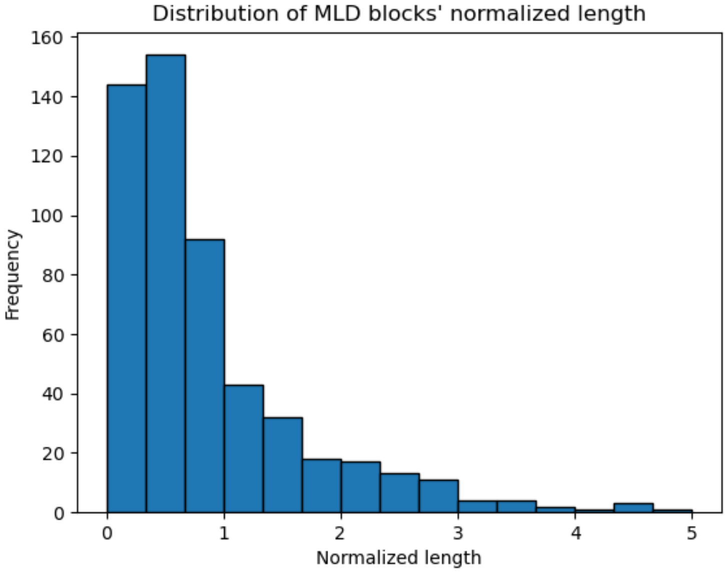

Figure 6 : Distribution of the MLD block normalized lengths, on *Hironda rustica* genomes.

The ratio of polymorphic sites distribution suggests that, by our method, many MLD blocks are inferred to be recent (Figure 7). This suggests that it will certainly be difficult to match a ratio threshold with a date. One approach would be to consider only segments of minimal length, since it is likely that the numerous tiny MLD blocks are to be inferred very recent.

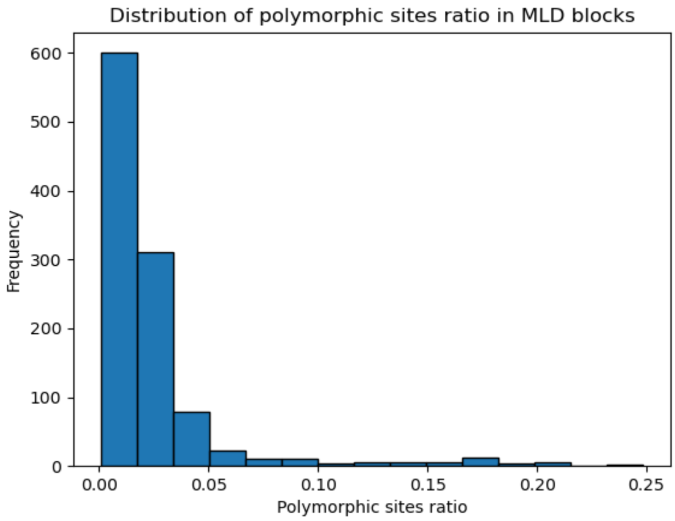

Figure 7 : Distribution of polymorphic sites ratio of the MLD blocks, on *Hironda rustica* genomes.

VI - Limitations and perspectives

Further investigation is required to find out whether the proposed solution could retrieve useful information and help to detect recent population decline.

A major limitation of previous studies and of this one is the assumption of a constant recombination rate along the genome, whereas it has been shown that in eukaryotes it can vary by a factor of ten depending on the localization[19]. Adressing this, however, would require some *a priori* knowledge about the genomes that the goal was to do without.

VII - References

1. Tollefson, J. Humans are driving one million species to extinction. *Nature* **569**, 171–171 (2019).
2. Barnosky, A. D. *et al.* Has the Earth's sixth mass extinction already arrived? *Nature* **471**, 51–57 (2011).
3. Rodrigues, A. S. L., Pilgrim, J. D., Lamoreux, J. F., Hoffmann, M. & Brooks, T. M. The value of the IUCN Red List for conservation. *Trends Ecol. Evol.* **21**, 71–76 (2006).
4. Régnier, C. *et al.* Mass extinction in poorly known taxa. *Proc. Natl. Acad. Sci.* **112**, 7761–7766 (2015).
5. van der Valk, T., Díez-del-Molino, D., Marques-Bonet, T., Guschanski, K. & Dalén, L. Historical Genomes Reveal the Genomic Consequences of Recent Population Decline in Eastern Gorillas. *Curr. Biol.* **29**, 165-170.e6 (2019).

6. Sánchez-Bayo, F. & Wyckhuys, K. A. G. Worldwide decline of the entomofauna: A review of its drivers. *Biol. Conserv.* **232**, 8–27 (2019).

7. Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H. & Bustamante, C. D. Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLOS Genet.* **5**, e1000695 (2009).

8. Sheehan, S., Harris, K. & Song, Y. S. Estimating Variable Effective Population Sizes from Multiple Genomes: A Sequentially Markov Conditional Sampling Distribution Approach. *Genetics* **194**, 647–662 (2013).

9. Beichman, A. C., Huerta-Sanchez, E. & Lohmueller, K. E. Using Genomic Data to Infer Historic Population Dynamics of Nonmodel Organisms. *Annu. Rev. Ecol. Evol. Syst.* **49**, 433–456 (2018).

10. Wright, S. Evolution in mendelian populations. *Genetics* **16**, 97–159 (1931).

11. Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat. Genet.* **46**, 919–925 (2014).

12. Kingman, J. F. C. The coalescent. *Stoch. Process. Their Appl.* **13**, 235–248 (1982).

13. Kerdoncuff, E., Lambert, A. & Achaz, G. Testing for population decline using maximal linkage disequilibrium blocks. *Theor. Popul. Biol.* **134**, 171–181 (2020).

14. Kelleher, J., Etheridge, A. M. & McVean, G. Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLOS Comput. Biol.* **12**, e1004842 (2016).

15. Hudson, R. R. & Kaplan, N. L. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111**, 147–164 (1985).

16. Prado-Martinez, J. *et al.* Great ape genetic diversity and population history. *Nature* **499**, 471–475 (2013).

17. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

18. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

19. Stapley, J., Feulner, P. G. D., Johnston, S. E., Santure, A. W. & Smadja, C. M. Variation in recombination frequency and distribution across eukaryotes: patterns and processes. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **372**, 20160455 (2017).