UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE BIOLOGIA ANIMAL

Comparative demography of endangered species using genomic data

Miguel Cisneiros e Faria Lourenço

Mestrado em Biologia Evolutiva e do Desenvolvimento

Dissertação orientada por:
Lounès Chikhi, PhD
Margarida Matos, PhD

2024

Acknowledgements

Only the best is good enough
Só o melhor é bom o suficiente

Calouste Sarkis Gulbenkian

When I first entered the Instituto Gulbenkian de Ciência (IGC), I only knew that this was *the* place to be if you want to be a scientist or, more specifically, a biologist. Little did I know about Calouste Gulbenkian's motto and its probable impact on the creation of what made IGC a pioneer and renowned science institute, and a sub-product of one of the most endowed charitable foundations in the world, the Calouste Gulbenkian Foundation.
Gulbenkian's 'controversial' actions in life have certainly enabled the current environmental state of our planet, but his wishes upon his passing have touched many lives through art, charity, science, and education, and tried to provide a better world. For me, the ultimate proof of this was being able to do my MSc project at IGC, especially during the institute's final years.

Holding this thesis on a physical material or reading it through a digital platform would not be possible without Lounès Chikhi. More than my advisor and more than the leader of the Populations and Conservation Genetics (PCG) group, Lounès was a surprise. Not only was his immense knowledge of biodiversity and ecology crucial for developing this project, his openness, his humor, and his understanding of the impact of several dimensions in science, like social justice (and vice versa) have, together, exponentially boosted my experience at IGC for this thesis project. Naturally, I can say the same about all the members of the PCG group, like Margarida (my practical tutor), Camille, Filipa, Inês, Ravi, Joana, and Diogo. Our scientific and non-scientific discussions at lunch or during snack time have enriched me immensely and formed a special bond tying us all together. In many of these occasions, we were happily joined by a "borrowed group member", Carolina, and another ex-member of the PCG group, Rita. All were essential for my stay at the IGC, sometimes for my mental health, and for the scientific output that I now make available. An enormous thank you and farewell hug to all of you. The PCG group will be missed.

The Faculty of Sciences of the University of Lisbon (FCUL), with this MSc, has provided me lectures with some of the most brilliant researchers in the Portuguese landscape of Evolutionary Biology. These include Inês Fragata, Telma Laurentino, Filipe Sousa, Vítor Sousa, Patrícia Beldade, and many others. Also, Professor Margarida Matos deserves a special mention because, despite having a more symbolic and bureaucratic role as my supervisor from FCUL, she assumed a proactive position by meeting with me and Lounès to receive updates on my progress. Though sometimes we clashed in red-tape issues like timings and others, Professor Margarida has demonstrated a 'tough love' approach that ultimately led me to present today the best possible version of my thesis. To all, my sincerest thank-you.

I would also like to thank Professor Elio Sucena, coordinator of the MSc in Evo-Devo Biology, for these years in this MSc program. Professor Elio and the team of his lab (Evo-Devo) at IGC – Priscilla, Patricia, Tânia, and Diogo - have also kindly hosted me for some weeks of practical courses and have

contributed to the improvement of my wet-lab and *Drosophila* maintenance skills. Their support and hospitality made a big difference in my first experience at IGC, and I'm very grateful for that.

I've been lucky to meet so many different people during my MSc, at my part-time job at Fundação Champalimaud, and at the *Residência Universitária Professor Egas Moniz* (RUEM-SASUL). Some of these amazing folks became close friends who've been there through the highs and lows, whether it was to celebrate the good times or lend an ear when things got tough. A huge thanks to Júlia, Inês G., Marco, Maria Beatriz, André, Mariana, Rita, Sara, Teresa, Tiago, and many others – I couldn't have done this without you.

Over this last year of thesis I have also had great support from people that have seen my potential and have contributed to this chapter of my life in an exceptional way. Starting with Francisco Pina-Martins, the person who encouraged me to pursue this specific MSc and whose mentorship and support turned into friendship. I'm also deeply grateful to Professor Helena Caria, whose regular "check-ins" on my progress - whether through a quick text or a casual lunch - made a big difference, both academically and personally. In addition, I would like to thank my boss at BioData.pt | ELIXIR-PT, Inês Chaves, for taking a chance on me and for her encouragement and comprehension towards the task of finishing and submitting this thesis in time. Also, a big hug to Gil, Mariana, Luciana, and Savannah.

Finally, I want to thank my brothers, Inês, Pedro, and Rita, and my "family from Algarve", Iman, Sofia and Ema. You have been there for me with advice, support, and encouragement whenever I needed it. This MSc feels like something we accomplished together, and I'm truly grateful to have you in my corner.

Last but not least, in the words of Snoop Dogg, "I want to thank me for believing in me, I want to thank me for doing all this hard work. I wanna thank me for having no days off. I wanna thank me for never quitting. I wanna thank me for always being a giver and trying to give more than I receive. I wanna thank me for trying to do more right than wrong. I wanna thank me for being me at all times". Jokes aside, these have been challenging years at a personal level: whether through health scares, financial difficulties, or even other hiccups of a quarter-life crisis. To be here is a great accomplishment and so, my present-self would like to thank my past-self and remind my future-self to always see the glass half-full.

For the financial support, in the form of scholarships and student housing, I hereby proper thank the *Direção-Geral do Ensino Superior* (DGES) and the *Serviços de Ação Social da Universidade de Lisboa* (SASUL).

Resumo alargado

Madagáscar, uma das áreas com maior biodiversidade do mundo, é conhecido por abrigar inúmeras espécies endémicas, incluindo o género *Microcebus* (lémures-rato), um grupo de pequenos primatas. Este género tem passado por um processo de diversificação significativa, resultando no reconhecimento de mais de 20 espécies ao longo das últimas décadas. Estes lémures-rato habitam diversas regiões da ilha e apresentam uma variedade de adaptações ecológicas, sendo uma peça central para a conservação da biodiversidade de Madagáscar. Este estudo centra-se em três espécies: *Microcebus arnholdi*, *Microcebus tavaratra*, e *Microcebus murinus*, que foram selecionadas devido aos seus diferentes nichos ecológicos, distribuições geográficas, e relevância para os esforços de conservação. O objetivo principal desta tese é reconstruir as suas histórias demográficas e estruturas populacionais com base em dados genómicos, avaliando como a conectividade entre populações e o fluxo génico mudaram ao longo do tempo.

Para tal, este trabalho utiliza duas abordagens principais: o método *Pairwise Sequentially Markovian Coalescent* (PSMC) e o *Structured Non-stationary Inference Framework* (SNIF). O PSMC, desenvolvido por Li e Durbin em 2011, é uma ferramenta que permite modelar o tamanho efetivo de uma população ao longo do tempo com base nos tempos de coalescência inferidos a partir dos genomas de indivíduos diploides. Este método é particularmente útil para estimar eventos demográficos, como expansões populacionais ou *bottlenecks* genéticos. No entanto, o PSMC depende de genomas de alta qualidade para inferir a história demográfica e, quando se utilizam genomas de espécies relacionadas, podem surgir enviesamentos que afetam a precisão das inferências.

Para melhorar as inferências do PSMC, esta tese utiliza o SNIF, uma ferramenta mais recente introduzida por Arredondo et al. em 2021, que tem em conta a estrutura populacional das espécies. O SNIF ajusta um modelo *n-island*, que divide a população em várias subpopulações (ou ilhas/demes), permitindo a estimação de variações nas taxas de migração e mudanças na conectividade ao longo do tempo. Desta forma, o SNIF oferece uma análise mais detalhada da estrutura populacional, permitindo inferir eventos demográficos complexos que o PSMC tradicional poderia não detetar.

Os resultados desta tese revelam padrões demográficos distintos entre as três espécies estudadas. *Microcebus arnholdi*, uma espécie microendémica restrita à região montanhosa e húmida da Montagne d'Ambre, no norte de Madagáscar, apresenta uma história de declínio acentuado no fluxo génico, que começou por volta de 500 mil anos atrás. Este declínio é provavelmente consequência de mudanças climáticas e de fragmentação do habitat, possivelmente associadas à atividade vulcânica na região norte de Madagáscar. As erupções vulcânicas no Norte, que criaram a Província Alcalina de Madagáscar do Norte, formaram terrenos férteis e habitats favoráveis ao desenvolvimento de florestas húmidas, onde *M. arnholdi* se adaptou e especializou-se. Este isolamento geográfico e a especialização ecológica desta espécie podem explicar a sua história demográfica distinta, marcada por uma redução mais precoce na conectividade em comparação com outras espécies.

Por outro lado, *Microcebus murinus*, uma espécie com uma distribuição mais ampla em florestas secas no oeste de Madagáscar, mostra um padrão diferente. Embora também tenha sofrido um declínio na conectividade por volta de 500 mil anos atrás, este declínio foi menos acentuado e o fluxo génico foi mantido por um período mais longo. Isto sugere que *M. murinus* teve uma maior capacidade de adaptação a diferentes condições ambientais e que as florestas secas do oeste de Madagáscar proporcionaram habitats mais estáveis durante as mudanças climáticas ocorridas no Pleistoceno.

Microcebus tavaratra, que habita as florestas secas do norte de Madagáscar, apresenta uma história demográfica ainda mais complexa. O declínio no fluxo génico desta espécie ocorreu mais tardiamente, entre 200 mil e 300 mil anos atrás, o que sugere que ela foi inicialmente menos afetada pelas pressões ambientais que impactaram outras espécies. Esta diferença temporal no declínio da conectividade pode ser explicada pela ampla distribuição geográfica de *M. tavaratra* e pela maior estabilidade dos seus habitats em comparação com as áreas de florestas húmidas. Além disso, a diversidade genética relativamente elevada observada em *M. tavaratra* pode estar associada ao seu maior tamanho populacional e distribuição mais ampla, o que lhe permitiu manter uma maior conectividade durante períodos de mudanças ambientais.

O uso do SNIF permitiu a identificação de eventos demográficos específicos e mudanças na conectividade que não foram capturadas com o PSMC sozinho. Por exemplo, o SNIF foi capaz de inferir o número de subpopulações, ou demes, e as mudanças nas taxas de migração entre estas ao longo do tempo, oferecendo uma visão mais detalhada de como as pressões ambientais e o isolamento geográfico moldaram a evolução destas espécies. A comparação entre as inferências feitas com o SNIF e os dados paleoclimáticos também ajudou a identificar potenciais correlações entre eventos demográficos e mudanças climáticas em Madagáscar.

Contudo, este estudo também destaca os desafios associados ao uso de dados genómicos em organismos não-modelo, como *Microcebus*. Uma das principais dificuldades encontradas foi a falta de genomas de referência de alta qualidade para espécies como *M. arnholdi* e *M. tavaratra*. A utilização de genomas de espécies relacionadas, como *M. murinus*, introduz enviesamentos nas inferências demográficas, uma vez que a divergência filogenética entre as espécies pode distorcer os resultados. Para mitigar esses problemas, esta tese utilizou alinhamentos *de novo* específicas para estas espécies, o que melhorou a precisão das inferências demográficas.

Apesar destas limitações, os resultados apresentados nesta tese têm implicações importantes para a biologia da conservação. As inferências sobre a estrutura populacional e as mudanças na conectividade ao longo do tempo fornecem informações críticas para os esforços de conservação, particularmente em regiões como Madagáscar, onde a fragmentação do habitat e a perda de biodiversidade são questões prementes. O conhecimento sobre as histórias demográficas das espécies de *Microcebus* pode ajudar a identificar populações em risco e orientar as estratégias de conservação, como a criação de áreas protegidas e a gestão da fragmentação do habitat.

Além disso, o SNIF demonstrou ser uma ferramenta poderosa para inferir padrões demográficos complexos, permitindo uma análise mais robusta da dinâmica populacional de espécies ameaçadas. A sua capacidade de modelar estruturas populacionais *non-stationary* e modelar também mudanças na conectividade oferece uma nova perspetiva sobre como as espécies responderam às pressões ambientais ao longo do tempo. No futuro, o uso de genomas de alta qualidade ou alinhamentos *de novo* com maior qualidade poderão contribuir para melhorar ainda mais a precisão destas inferências, permitindo uma compreensão mais completa das histórias evolutivas das espécies de *Microcebus*.

Em suma, esta tese contribui significativamente para o campo da genómica populacional ao aplicar ferramentas inovadoras, como o SNIF, para estudar a estrutura populacional de espécies não-modelo. Os resultados avançam o conhecimento sobre as dinâmicas populacionais de *Microcebus* e oferecem um quadro para estudos futuros em outros organismos. Ao alinhar dados genómicos com registos paleoclimáticos, este trabalho de investigação evidencia uma ligação entre a diversidade genética e as

mudanças ambientais, destacando os fatores que impulsionam a evolução das espécies em ecossistemas isolados como o de Madagáscar.

Assim, as descobertas desta tese não apenas ampliam a nossa compreensão da história evolutiva das espécies de *Microcebus*, mas também oferecem informações valiosas para a conservação da biodiversidade única de Madagáscar. A aplicação de ferramentas genómicas para estudar a estrutura e conectividade populacional representa um passo importante para o desenvolvimento de estratégias de conservação mais eficazes e para a proteção dos habitats destas espécies ameaçadas.

Palavras-chave

Lémures-rato, Genómica populacional, História demográfica, Alterações na conectividade, Biologia da conservação

Abstract

Madagascar, a biodiversity hotspot, is home to *Microcebus* (mouse lemurs), with over 20 species. This study focuses on three species: *Microcebus arnholdi*, *M. tavaratra*, and *M. murinus*, selected for their distinct ecological niches and conservation importance. Using genomic data, the research reconstructs their demographic histories, assessing changes in gene flow and connectivity over time.

The study employs the Pairwise Sequentially Markovian Coalescent (PSMC) and the Structured Non-stationary Inference Framework (SNIF). PSMC models effective population size from genomic data, while SNIF incorporates population structure, accounting for migration and connectivity shifts. SNIF offers a more detailed analysis of demographic changes, revealing shifts that PSMC alone might miss, and estimates subpopulation sizes, migration rates, and connectivity patterns, providing insights into how environmental factors and geographic isolation have influenced lemur evolution.

A key challenge is the absence of high-quality reference genomes for *M. arnholdi* and *M. tavaratra*, complicating the use of alignment-based methods like PSMC. To improve accuracy, the study uses *de novo* assemblies, though some biases persist, especially when using *M. murinus* as a reference.

The results indicate differing demographic patterns. *M. arnholdi*, confined to Montagne d'Ambre, experienced a significant decline in gene flow around 500,000 years ago, likely due to climatic changes. *M. murinus* also shows reduced connectivity but maintained gene flow for longer. *M. tavaratra*, inhabiting drier northern forests, saw a delayed reduction in gene flow around 300,000 years ago, possibly due to distinct ecological pressures. Its higher genetic diversity may be linked to its broader distribution.

This research demonstrates the utility of SNIF in conservation genomics, offering insights into population structure and connectivity changes, advancing conservation strategies for these species of mouse lemurs.

Keywords

Mouse lemurs, Population genomics, Demographic history, Connectivity changes, Conservation biology

Index

List of Acronyms and Abbreviations

BP – Before Present

IICR – Inverse Instantaneous Coalescence Rate

MRCA – Most Recent Common Ancestor

PSMC – Pairwise Sequentially Markovian Coalescent

SNP – Single Nucleotide Polymorphism

SNIF – Structured Non-stationary Inference Framework

Kya – Thousand years ago

Mya – Million years ago

NMAP – Northern Madagascar Alkaline Province

List of Tables

List of Figures

1. Introduction

1.1 Madagascar: a case-study for Population Genomics and Conservation

Madagascar is an island off the southeastern coast of Africa and a biodiversity hotspot due to its extremely high levels of species endemism and to the fact that its biodiversity is highly threatened (Myers et al., 2000). It is believed that this arose as a consequence of its isolation from both Africa and India, to which it was connected more than 120 and 80 MY ago, respectively. Lemurs are an emblematic example of primates only found in Madagascar. Mouse lemurs (*Microcebus spp.*), with more than 20 recognized species, are an example of the diversification that took place in Madagascar (van Elst et al., in press/2024). Mouse lemurs are morphologically cryptic - they share a similar appearance - which explains why the number of species identified and described has increased from four to more than 20 within 2.5 decades. They are present across all regions and habitats of Madagascar, and they generally exhibit allopatric distributions, with only a few cases of sympatry. Their generation time is short and thought to be around 2.5 years (Radespiel et al., 2019).

In this thesis, we will focus on three species, namely *Microcebus arnholdi*, *M. tavaratra*, and *M. murinus*. The divergence between these species is not known with certainty, as the estimated times based on mitochondrial DNA differ from those obtained from nuclear genes. Based on the latter, van Elst et al. (2023) estimated that *M. arnholdi* and *M. tavaratra* diverged around 490 Kya, while they diverged from *M. murinus* around 1.5 Mya, a time that corresponds also to the estimated MRCA of all *Microcebus* species.

Figure 1.1 - Map of Madagascar with *Microcebus spp.* distribution; orange: M. arnholdi, green: M. murinus, blue: M. tavaratra)

Microcebus murinus was the first species of Microcebus to be described (in 1795). It is widely distributed across the West of Madagascar and is the most studied species of *Microcebus* - and the only with a High-Quality reference available on the standard genome databases.

While *M. arnholdi* is microendemic and currently thought to only survive in the Montagne d'Ambre, a humid forest habitat in the northern region of Madagascar, the other two species prefer dry forest environments and are located in wider regions in the north-northeast (*M. tavaratra*) and west (*M. murinus*) of Madagascar (Fig. 1.1). The population density of *Microcebus* varies across species, but there is a consensus concerning the decline of populations due to habitat loss(Louis et al., 2008). The biodiversity of Madagascar has been endangered mostly due to human activities, habitat fragmentation, and climate change (Quéméré et al., 2013). Woodlands and forests, the habitat of 90% of Malagasy species, have lost an estimated 44% of its forest cover from 1953 to 2014 (Goodman & Benstead, 2005; Vieilledent et al., 2018).

The International Union for Conservation of Nature (IUCN) Red List update of 2020 stated that 98% of all lemurs are threatened and that 38% are Critically Endangered (IUCN, 2020). Moreover, global reports on biodiversity indicate unprecedented rates of biodiversity loss and suggest that we are facing the sixth mass extinction (Ceballos et al., 2015). Thus, there is an increased significance in studying the population structure and demographic histories of these species towards conservation efforts (Hohenlohe et al., 2021).

1.2 Highlights on Population Genomics and Evolution

Since Charles Darwin's discovery of evolutionary changes through natural selection, there has been increasing interest in quantifying biological variation, especially in the genetic context (Charlesworth, 2010). The later integration of Mendelian genetics and the structure of DNA, revealed through the work of Franklin, Watson, and Crick, set the stage for modern genomics and the study of genetic inheritance (Giani et al., 2020). This shift has provided new tools to measure genetic diversity and model evolutionary processes, crucial for understanding population dynamics.

Population Genetics is the branch of evolutionary biology that studies the genetic composition of populations and how it changes over time, primarily as a result of evolutionary forces. This now-centennial field arose mainly from the works of Ronald Fisher, Sewall Wright, and J.B.S. Haldane (Gillespie, 2004). More specifically, the fathers of Population Genetics have defined evolutionary processes - i.e., drift, selection, migration, mutation, and recombination –, developed concepts, and developed models to try to shed some light on genetic variation among individuals (Okazaki et al., 2021).

Population Genetics itself has evolved into Population Genomics, where the research on the effects of those previously mentioned evolutionary forces and variability is not restricted to some loci, as the whole-genome sequencing approach has increased marker density and produces more accurate population genetics estimates (Black IV et al., 2003). Distinguishing between loci-specific and genome-wide effects is particularly useful to delve into different scopes: while the former is better to study fitness and adaptation, the latter enables a focus on the demography, history, and structure of a population (Luikart et al., 2003). Given this range of perspectives, we will use Population Genomics as an umbrella term from now on.

One of the key objectives of Population Genomics is to explore how life's diversity is distributed, how it has evolved, and how it may adapt to future environmental changes (Ellegren, 2014). The processes of selection, migration, genetic drift, mutation, and recombination play crucial roles in shaping population histories, leaving detectable signatures in species' genomes. The most famous model in Population Genomics is the Wright-Fisher Model, which provides a simple and tractable framework for understanding the genealogical history of alleles within a population (Wakeley, 2013). This model describes genetic drift in a single, well-mixed population while assuming non-overlapping generations, random mating, and constant population size (Fisher, 1922; Wright, 1931).

In 1982, John Kingman introduced a mathematical framework that traces the genealogical history of genetic material within a population, describing how alleles in present-day individuals trace back to shared common ancestors: the Coalescent Theory. The moments in history to which those common ancestors can be traced are called coalescent times (Kingman, 1982; Wakeley, 2004).

One critique of the standard Coalescent Theory and the models for inference of the demographic history of species is their reliance on a single parameter, population size (N), to depict the complex evolutionary history. These models typically emphasize fluctuations in population size over time without adequately accounting for factors like population structure and spatial differentiation between geographical locations, which are crucial for understanding demographic events. Also, since the standard Coalescent Theory assumes panmixia - individuals within a population mate randomly with one another -, several researchers have highlighted that this assumption can lead, for example, to the detection of false bottlenecks, thus misrepresenting the real history of those species (Chikhi et al., 2010; Wakeley, 2001).

It is known that most species are structured, and it has been proven that their demographic history is shaped by events of connectivity like fragmentation and expansion - such as habitat fragmentation in *Pan troglodytes* (chimpanzees), which reduces gene flow between populations over time (Gagneux & Varki, 2001). Representations of some population structures are possible by using models such as the *n-island* model. In this model, a population is partitioned into multiple subpopulations (islands or demes), each with a consistent size (N) and interconnected by a constant migration rate (m) (Wright, 1931).

1.3 PSMC, IICR, and SNIF

Recently, the advances in computing power, the growth in access to genomic data of multiple species, and the development of new statistical models have advanced the estimation of parameters of interest such as migration or admixture rates, and the dates of putative bottlenecks, expansions or splitting events (Rodríguez et al., 2018). For example, the Pairwise Sequentially Markovian Coalescent (PSMC), developed in 2011 by Li and Durbin, estimates demographic history (i.e., effective population size; *Ne*) by modeling coalescence times with a hidden Markov model, across the genome of a single diploid individual (Li & Durbin, 2011). Recently, Liu & Hansen (2017) found that the method could be used with reduced representation libraries, such as restriction site-associated DNA sequencing (RAD-seq) - as an alternative to using the whole genome.

Like most genomic tools, PSMC requires the alignment of the genomic sequence of the target species to a reference genome (Li & Durbin, 2011). Since reference genomes are often unavailable for non-model species, the typical approach is to align the data of the species of interest to the genome of a closely related species. However, using a reference genome from a closely related species can introduce

significant biases in downstream analyses. Prasad et al. (2022) found that when the phylogenetic divergence between the target species and the reference genome exceeds approximately 3%, the inferred demographic histories using the Pairwise Sequentially Markovian Coalescent (PSMC) model are notably impacted. For instance, estimates of heterozygosity, runs of homozygosity (ROH), and demographic events may differ substantially depending on the divergence between the reference and target species. This can obscure true evolutionary signals, particularly in non-model organisms where high-quality conspecific reference genomes are not available.

Mouse lemurs are an example of those non-model organisms, and a study on *Microcebus* has shown that even smaller phylogenetic divergences (~2%) can affect PSMC inferences and other genetic analyses. The use of a short-read *de novo* assembly specific to the target species often yields more accurate results compared to relying on a distantly related reference genome. These findings underscore the importance of minimizing phylogenetic distance between the reference and target species to avoid biases that can mislead evolutionary and population genetic interpretations (Henrique, n.d.) (*in prep*).

Mazet et al. (2016) have introduced the Inverse Instantaneous Coalescence Rate (IICR). Simply put, IICR is a sample and time-dependent function that can be interpreted as an effective size in a panmictic population. The previously mentioned PSMC method, though typically viewed by most authors as a way of detection of population size changes, actually infers the IICR for two samples (Mazet et al., 2015). Thus, IICR curves can be interpreted as summaries of genomic information (Chikhi et al., 2017).

In 2021, Arredondo et al. presented the Structured Non-stationary Inference Framework (SNIF) method. SNIF assumes *a priori* a non-stationary *n-island* model or "piecewise stationary *n-island* model" and automatically fits, through several iterations, the best model to a PSMC/IICR curve (Arredondo et al., 2021; Couloigner, 2022). This framework enables the inference of the following:

- Number of subpopulations (n)
- Size of those sub/populations (N)
- Time of change in connectivity (t_i)
- Migration rates/gene flow (m_i)

SNIF has been validated using simulations - with "ms" - and has been applied to real data from species like humans, and a species of orangutans, *Pongo abelii* (Arredondo et al., 2021; Couloigner, 2022) and chimpanzees (Steux, 2023). Arredondo et al. (2021) have highlighted that further work is still necessary, and we will come back to this in the Discussion.

1.4 Aims and Goals

In this thesis, we intend to study the population structure and the history of connectivity of three species of *Microcebus* (mouse lemurs) using whole genome sequence data obtained from a small number of individuals from each of those species. To do so, in this project we revisit the PSMC method of Li & Durbin (2011) applied in the literature to *Microcebus*, we use SNIF to infer n-island models that can explain the PSMC curves obtained for each species, and we confront those models with paleo climatic events.

With these goals in mind, the following research strategy is proposed:

1. To replicate the PSMC analysis of *Microcebus arnholdi* from specific literature

2. To optimize inference parameters on SNIF analyses on *Microcebus spp.* and assess the impact of reference divergence on PSMC outputs
3. To infer demographic histories under n-island models for three species of mouse lemurs using SNIF;
4. To validate the method of inference by re-running SNIF on model-based simulated genomic data using the *ms* coalescent simulation software;
5. To crossmatch patterns of connectivity changes from the inferred demographic histories with paleo climatic events

2. Materials and Methods

2.1 *Microcebus spp.* genomic data

Microcebus arnholdi was the principal subject of interest and *Microcebus murinus* was used for analyses and comparison due to the fact that higher quality genomic data, including HQ reference genome, was available for the latter species. The purpose of *M. tavaratra*'s data was to further extend the analyses and to add another element of comparison with *M. arnholdi*'s results, due to the scarcity of available information and the short genetic distance of both species.

The basis for the PSMC replication was a study on *Microcebus arnholdi* by Teixeira et al. (2021) that used *Microcebus murinus* High Quality genomic sequences as a reference for alignments. The genomic data from that study that was used in this thesis is available in NCBI's BioSample database (SAMN14909741) - a *Microcebus arnholdi* individual from Fantany, Madagascar.

The data (PSMC files) used for the inference of *n-island* models is mainly classified into two categories: alignments made with High Quality database-stored and *de novo* – available from previous studies carried out within the group - genomic sequences (references) of *Microcebus* species. Table 2.1 describes some properties of the genomic data, for each species of *Microcebus* that was used for SNIF inferences.

Table 2.1 - Selected characteristics of the genomic data used for SNIF inferences

Alignment	Species	Aligned with	Reference type	File type	Minimum depth	Maximum depth
1	*M. arnholdi*	*M. murinus*	High Quality	PSMC	10	60
2	*M. arnholdi*	*M. arnholdi*	*de novo*	PSMC	10	60
3	*M. arnholdi*	*M. murinus*	*de novo*	PSMC	10	58
4	*M. murinus*	*M. murinus*	*de novo*	PSMC	11	63
5	*M. murinus*	*M. murinus*	High Quality	PSMC	10	61
6	*M. tavaratra*	*M. murinus*	High Quality	PSMC	10	60
7	*M. tavaratra*	*M. tavaratra*	*de novo*	PSMC	11	63

2.2 Data Treatment and Analyses

2.2.1 PSMC Replication

An adapted version of the protocol – with steps, software (and versions) - that was used for generating the PSMCs from the genomic data is available in the Supplementary Information I. The parameters used for exact replication were the ones available in the Supplementary Information of the article by Teixeira et al. (2021), and the recommended parameters included in our protocol were also used to observe the differences. A mix of parameters was also used to observe the role each parameter had in

the generation of the PSMCs. All these parameters combinations are available at Table 3.1 in the *Results* section.

2.2.2 SNIF Inferences

The Structured Non-stationary Inference Framework (SNIF) assumes *a priori* a non-stationary *n-island* model - with constant deme size, and in which migration rates between populations are constant during specific periods (called *components; c*) - and automatically fits that model to a PSMC/IICR curve. Though this fitting process is iterative and automatic, there is a need to define several parameters *a priori* to set the parameter space that SNIF will explore in all analyses – a non-exaustive list of parameters is available at the Supplementary Information II. The weight (ω) parameter can be given as an example: the role of this variable is to tell SNIF to focus the analysis more towards the more recent past (ω>1.0), the more ancient past (ω<1.0), or in a uniform distribution along all the curve (ω=1.0).

Since the parameters of SNIF differ – due to the input data quality and other factors – from species to species, the process of identifying the optimal parameters is based on trial and error. This optimization task is described in more detail for the parameters c and ω in the *Results* section. Nevertheless, the definite value for the distance parameter (weight, ω) was set at 0.5, and the number of components (c) was 7. Each repetition of the analysis encompassed a variable number of optimization rounds or iteration, ranging from at least 10 to a maximum of 200, ensuring a comprehensive exploration of the parameter space. To mitigate potential stochastic effects and ensure robust findings, the number of repetitions was set to five.

A mutation rate (u) of 1.52×10^{-8} governed the frequency of genetic variations within the simulated population. As for the parameter of generation time, we considered 2.5 years for the *Microcebus* genera. Both the mutation rate and generation time parameters were retrieved from the literature (Teixeira et al., 2021).

2.2.3 *ms* Validations

After performing the demographic history inferences with SNIF, it is recommended to simulate pseudo-observed data sets (*pods*) based on those inferences and re-run SNIF on these *pods* to determine if the new inferred scenarios are similar to the scenarios used to simulate the *pods*.

From all the repetitions of a SNIF inference, the repetition with the smallest/lower distance value – closest inference to the real PSMC - defines a scenario that is typically chosen to be simulated using "ms", a software that allows the simulation of sequences and coalescence times under many complex demographic models and histories (Hudson, 2002). The *ms* program simulates T2 values (coalescence time for a sample size of two) from that repetition, and those values are used by SNIF to create a simulated IICR curve as a target for inference. These simulation-based inferences used the same parameters optimized for the original PSMC-based inferences, including weight (ω=0.5) and the number of components (c=7). Each simulated IICR was analyzed using 5 repetitions of SNIF, mirroring the iterative optimization process used with empirical genomic data. More information on specific parameters and arguments is available on the "ms" literature (Hudson, 2002).

A validated event refers to a demographic event that is detected consistently both in the SNIF analysis using empirical genomic data and in the validation analysis using simulated pseudo-IICR curves generated by "ms". This ensures that the inferred event is not an artifact of the original data or model assumptions but instead reflects a robust demographic signal.

3. Results

3.1. PSMC Replication

The most recent literature on the demographic history of *Microcebus arnholdi*, using PSMC, is a 2021 article by Teixeira et al. (2021). Before proceeding to the inference of *n-island* models for this species, we decided to compare (Figure 3.1 - images A and B) the PSMC/IICR published in that study with the PSMC we produced (using our data) for the same species and alignment - *M. arnholdi* aligned with an *M. murinus* High Quality reference.

After reading the Supplementary Information of that article and concluding that the parameters used for the production of that PSMC/IICR were not the same as the ones used in the protocol used by our group, we proceeded to replicate the results (*M. arnholdi* individual from the Fantany population) from Teixeira et al. with our protocol (Figure 3.1 - images C to I) and with the parameters we changed to detect putative differences that could explain the disparities between PSMCs (Table 3.1).

Table 3.1 - Listing of parameters used for generating each PSMC/IICR image of Figure 3.1. The parameters in bold are the ones that differ in each image from the original parameters used by (Teixeira et al., 2021)

Figure 3.1	**Max number of iterations (-N)**	**Max 2N0 coalescent time (-t)**	**Maximum read depth/ coverage (-d)**	**Minimum read depth/ coverage (-d)**	**Quality score filter (-q)**
A	30	5	100	3	30
B	**25**	**15**	**60**	**10**	**20**
C	30	5	100	3	30
D	**25**	**15**	**30**	**5**	**20**
E	**25**	5	**30**	**5**	30
F	30	**15**	**30**	**5**	30
G	30	5	**30**	**5**	30
H	30	5	**30**	**5**	**20**
I	**25**	**15**	100	3	**20**

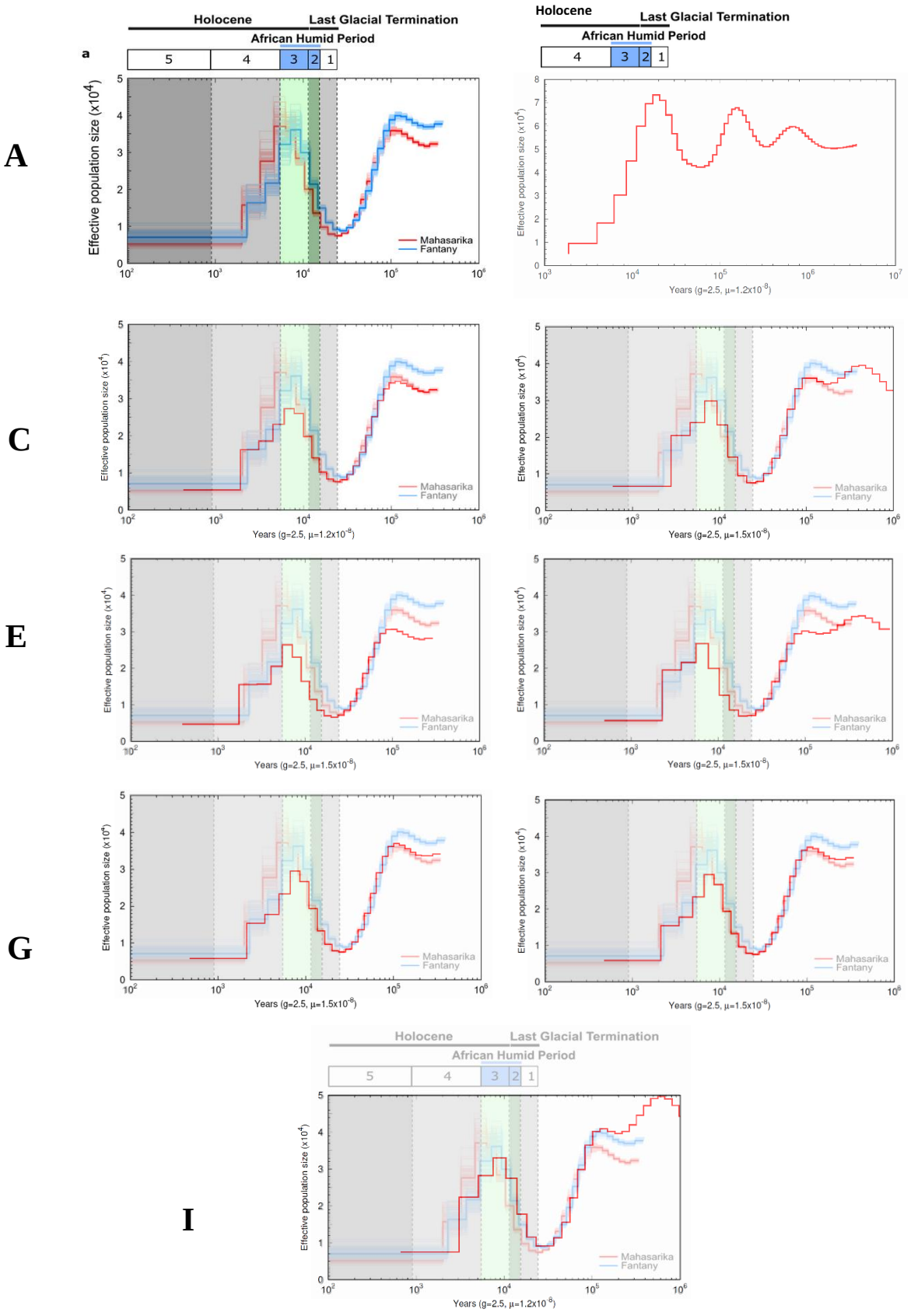

Figure 3.1 - Original PSMC/IICR published in Teixeira et al., 2021 (A) and the PSMC generated by us (B) for an *M. arnholdi* individual aligned with an *M. murinus* High Quality reference. Images C and D depict, respectively, a replication attempt with the parameters from the article and a reproduction with our parameters. The latter images (E, F, G, H and I) represent the impact assessment of changing specific parameters. All images were generated using the genomic data from Teixeira et al., except for image B.

No image in Figure 3.1 can exactly replicate the PSMC/IICR from Teixeira et al. (2021). The images we generated never overlap the original curves, with the better result being from image I. In images B, D, F, and I, an increased -t (maximum 2N0 coalescent time) parameter value enables a deeper temporal analysis, revealing PSMC/IICR changes further into the past. The similarity between images G and H shows that different values for the -q (base quality filtering) parameter do not appear to impact the outputs.

3.2. SNIF Parameter Optimization

As mentioned previously in the *Materials and Methods* section, for each species of interest, there is a need to find and optimize the parameters of the parameter space of SNIF. As we intend to focus later (in Section 3.3) on the impact of same/different species alignments on SNIF's inferences, here we will focus on the optimization of the parameters c and ω for inferences on a *Microcebus arnholdi* individual aligned with a reference sequence of the same species (*de novo*) and with a sequence from *Microcebus murinus* (high quality). In order to avoid "visual noise", only the best repetition of the SNIF inference is shown in Figures 3.2, 3.3, 3.4.

3.2.1. Components (c) parameter

The number of components (c) corresponds to the periods during which the n-island model parameters are assumed to be constant, but different between components. Increasing the number of components increases the number of parameters of the model - because the migration rate varies between components - and thus improves the fit of the model to the observed PSMC. Here we chose $c \in \{5, 6, 7, 8\}$ to allow for a reasonably large number of components but also tried to favour the models with the smallest number of parameters. Figure 3.2 shows the results of the inference with SNIF for ω=0.5, but similar results were obtained for ω=1.

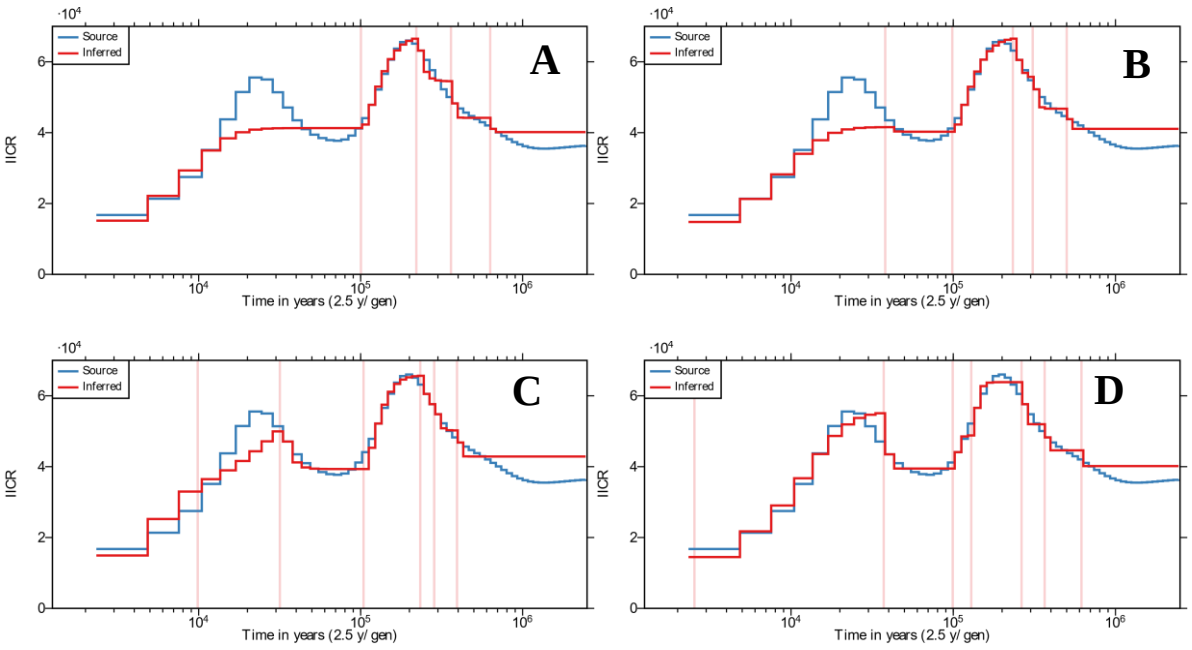

Figure 3.2 - SNIF inference outputs for the components (c) – 5 (A), 6 (B), 7 (C), 8 (D) - parameter optimization for an *M. arnholdi* individual aligned with a *de novo* reference sequence from the same species.

Figure 3.2 shows that, as expected, the SNIF inference improves with the number of components, with SNIF performing best for c=7 and c=8 components. For 5 and 6 components (Fig. 3.2 - A and 3.2 - B), there are no obvious differences between the inference, and SNIF appears to miss the first hump. In these cases, approximation of the target PSMC is good for the ancient past, though the most recent increase in the IICR, around 20 Kya, is not well-fitted. When adding one more component, to 7 (Fig. 3.2-C), that recent increase is better fitted than previously, with an event being placed by SNIF exactly on 10 Kya. When working with 8 components (Fig. 3.2 - D), the fitting is improved but placing the component at the beginning of the IICR curve.

Looking at the results for an *M. arnholdi* individual aligned with an *M. murinus* High Quality reference, Figure 3.3, we observe again the inference improvement with the increase of the number of components. While 5 and 6 components seems to do a relatively adequate job, using 7 components displays a better fitting, reducing the visual distance between the inference (in red) and the target IICR (in blue). When SNIF attempts to place events considering 8 components (Fig. 3.3 - D), the result is not as good as previously since it places 2 events at the most recent past - with one right at the beginning of the source IICR.

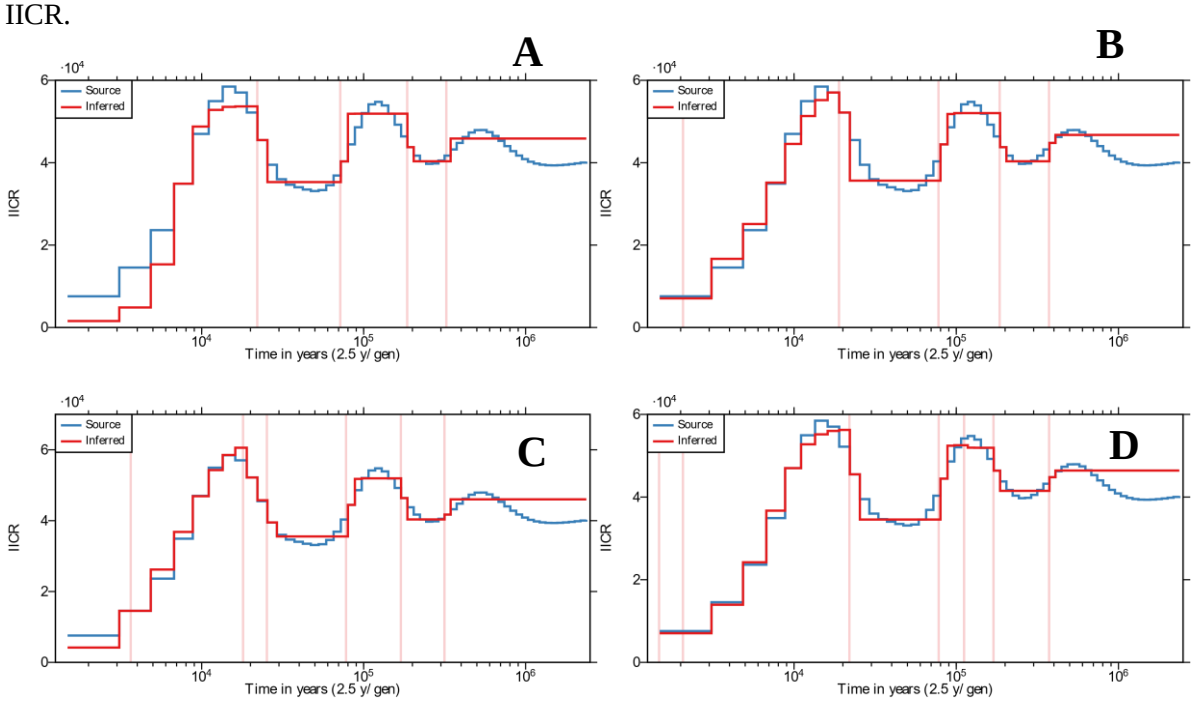

Figure 3.3 - SNIF inference outputs for the components (c) – 5 (A), 6 (B), 7 (C), 8 (D) - parameter optimization for an *M. arnholdi* individual aligned with a High Quality reference sequence from *M. murinus*.

3.2.2. Weight (ω) parameter

For the Weight (ω) parameter, we tested the values 0.5 and 1.0. To include all humps in the analyses, testing was limited to values lower than 1.0. Values above that would shift SNIF's focus more toward the recent past, by forcing the placing of events in that period, and thus diminish even further the representation of the third hump.

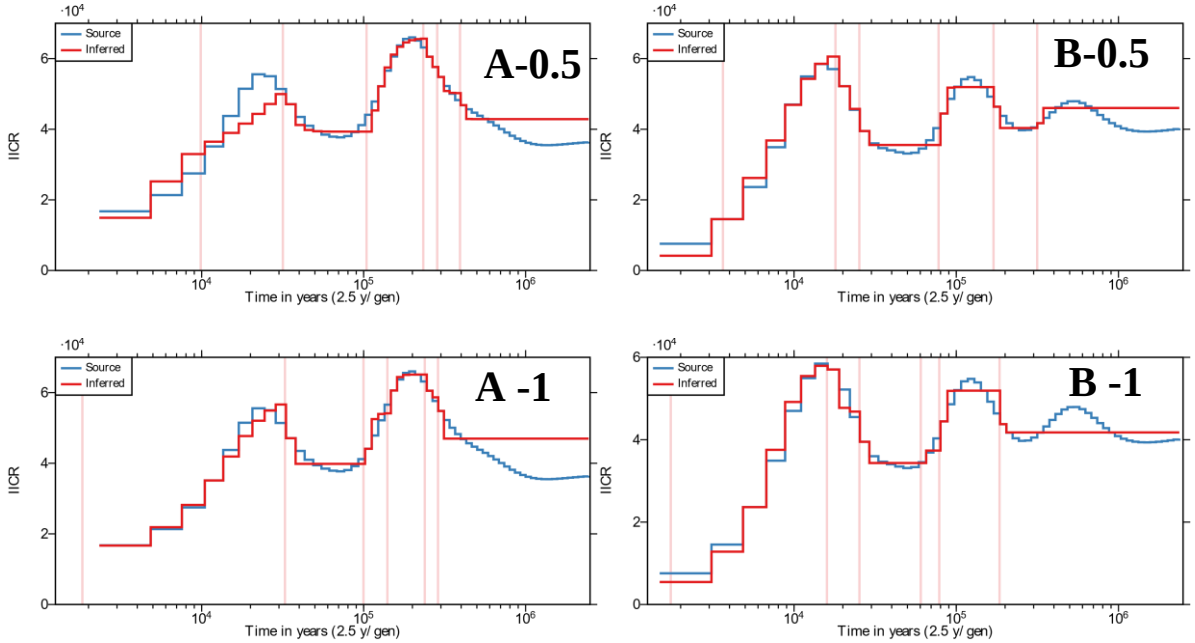

Figure 3.4 - SNIF inference outputs for the weight (ω) – 0.5 and 1.0 - parameter optimization for an *M. arnholdi* individual aligned with a *de novo* reference sequence from the same species (A) and with a High Quality reference sequence from *M. murinus* (B).

Despite efforts in choosing lower values for ω, in Figure 3.4 it is evident that SNIF cannot place any events further into the past beyond 400 Kya. When ω = 1.0, an event is always placed at around 2000 years BP. From Fig. 3.4 A-1.0 to A-0.5 this issue disappears with the tradeoff of reducing the fitting of the first curve. This tradeoff is unnoticed in B but a weight of 0.5 presents a slightly better fitting of the third curve than with 1.0.

3.3. SNIF Inferences

Before proceeding to use SNIF to infer the different demographic histories of *M. arnholdi* it was necessary to assess how reliable those inferences would be if we were aligning the genomic sequence of interest with a High-Quality reference sequence or with a *de novo* reference sequence. In Figure 3.5 we make that distinction, using the most widely studied species of *Microcebus*, *M. murinus*. Despite subtle differences between the PSMC of each alignment, the simulated IICR not only shows close fits to each PSMC but also shows demographic events in almost the exact time periods, e.g. at around 50, 120, and between 300 and 700 Ky BP. The same patterns are evident in the connectivity graphs (B) where, for each time period, the migration rate is the same between alignments, except for 3 repetitions in which differences are visible starting from around 20 Kya. In the histogram of islands (C), it is interesting to verify that the number of demes for the *de novo* alignment averages between 20 and 40, while the number of islands for the High-Quality alignment ranges lower between 18 and 30.

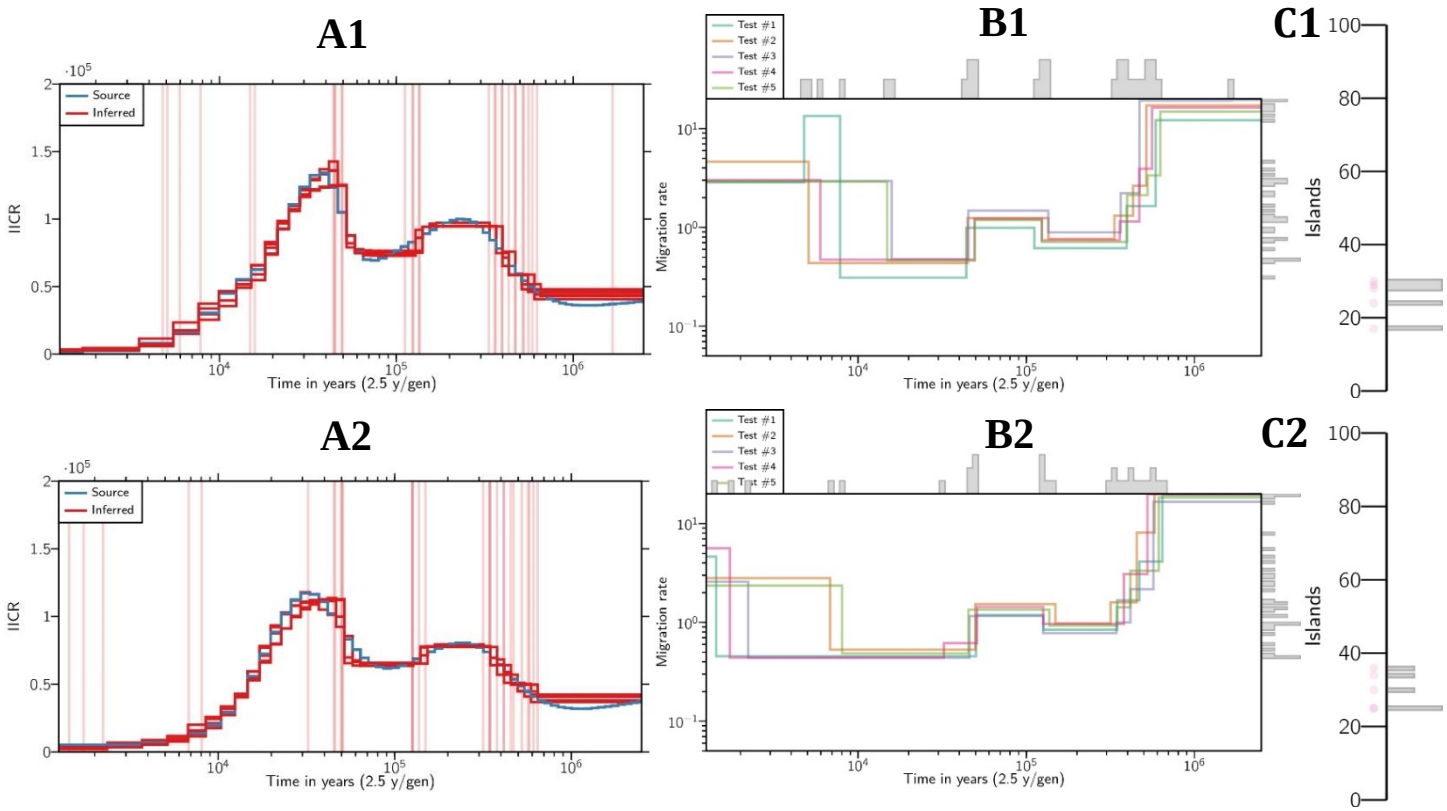

Figure 3.5 - SNIF inferences for *M. murinus* aligned with High Quality (A1, B1, C1) or *de novo* (A2, B2, C2) reference sequences from the same species. Images A correspond to the IICR plots, B to the connectivity graphs and C to the histograms with the number of islands/demes.

For the sake of comparison, we decided to also confront the inferences of *M. arnholdi* aligned with *M. murinus* HQ and with *M. murinus de novo* references (Figure 3.6). Looking at the results we obtained (Fig. 3.6-A), we can see that the placing of events by SNIF is much more consistent among repetitions for the *de novo* alignment than for the one with the HQ reference. There are some putative periods for demographic events that are common to both alignments: 20 to 30 Kya, 60-80 Kya, 160-200 Kya, 350-400 Kya, and 650 to 800 Kya. SNIF has also identified events in the last millennia for Fig. 3.6-A1 (circa 500 to 700 years ago) and Fig. 3.6-A2 (circa 700 to 900 years ago). Regarding connectivity, periods of change in migration rate are quite similar between alignments. Finally, the number of inferred islands is steadier for the *de novo* alignment (10 to 15) than for the High Quality (15 to 35).

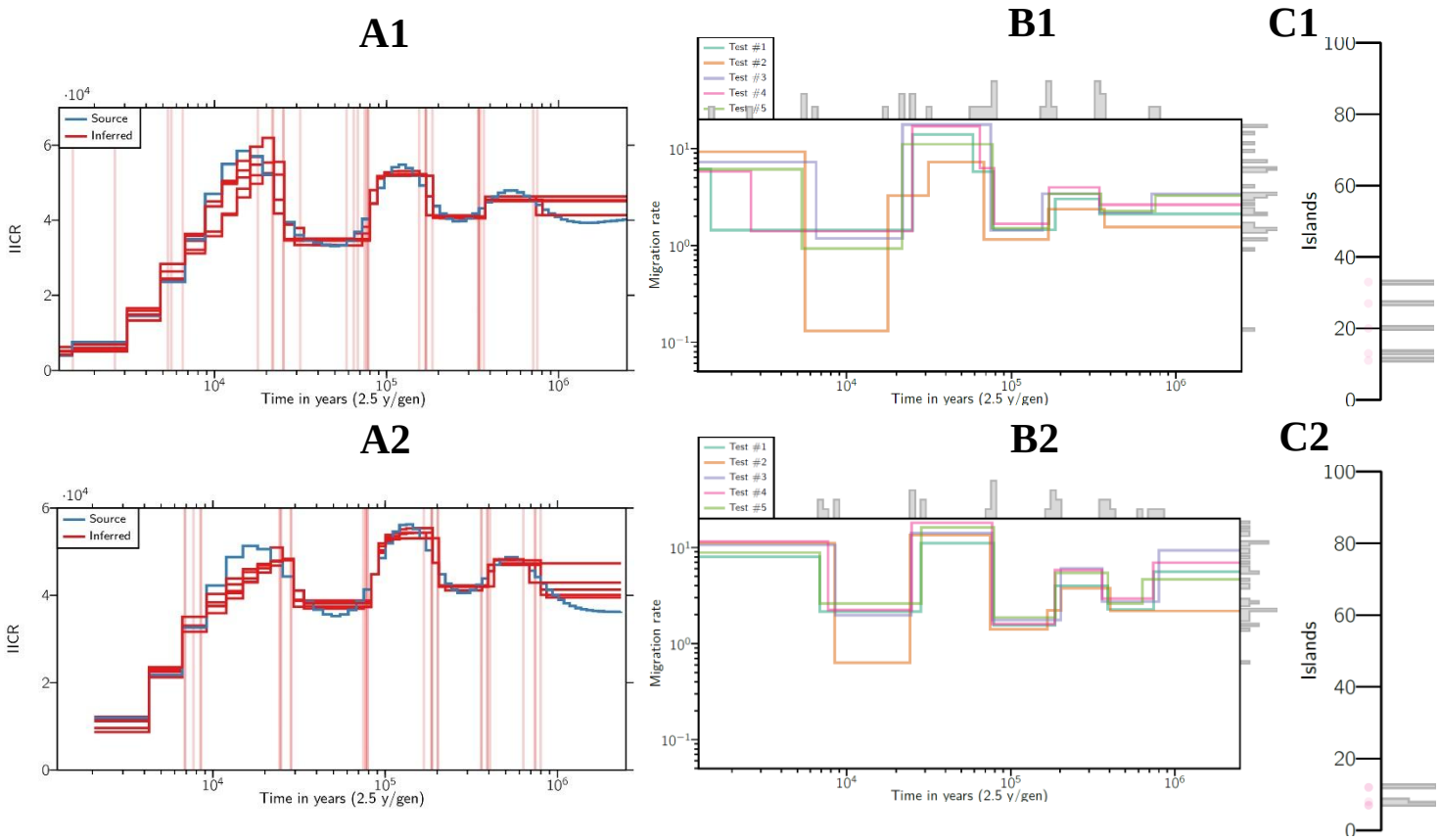

Figure 3.6 - SNIF inferences for *M. arnholdi* aligned with High Quality (1) or *de novo* (2) reference sequences from *M. murinus*. Images A correspond to the IICR plots, B to the connectivity graphs and C to the histograms with the number of islands/demes

Using different species as references for the alignments, nn contrast to what we have seen in Figure 3.5 and 3.6, *Microcebus arnholdi* does not follow the similarities between High Quality (Figure 3.7-A1) and *de novo* (Fig. 3.7-A2) alignments displayed previously. While in Fig. 3.7-A1 we observe three humps with maximum heights at approximately 15, 125, and 500 Kya, in Fig. 3.7-A2 are only represented two curves with vertices at 20 and 200 Kya. These maximum heights, though similar for the first curve of Fig. 3.7-A1 and Fig. 3.7-A2 – at an approximate inverse instantaneous coalescence rate of 6 – differ from the second curve onwards in the two alignments. This different number of curves, and their position in time, had implications for SNIF's inference of connectivity changes: while migration rates (Fig. 3.7-B) seem to share a similar pattern of increase followed by a decrease - between circa 100 and 10 Kya, adding a generous margin - there is much uncertainty regarding the most recent and the most ancient past. For the *M. arnholdi de novo* alignment (Fig. 3.7-B2) one can observe a decrease on the migration rate until about 200 Kya, while the same cannot be said for the High Quality reference alignment (Fig. 3.7-B1) due to its apparent and relatively stable migration rates during the same period. The number of demes (Fig. 3.7-C) is also incongruent between alignments with Fig. 3.7-C1 presenting about [10, 32] and Fig. 3.7-C2 being around [5, 18] islands, with 2 repetitions pointing towards this lowest number.

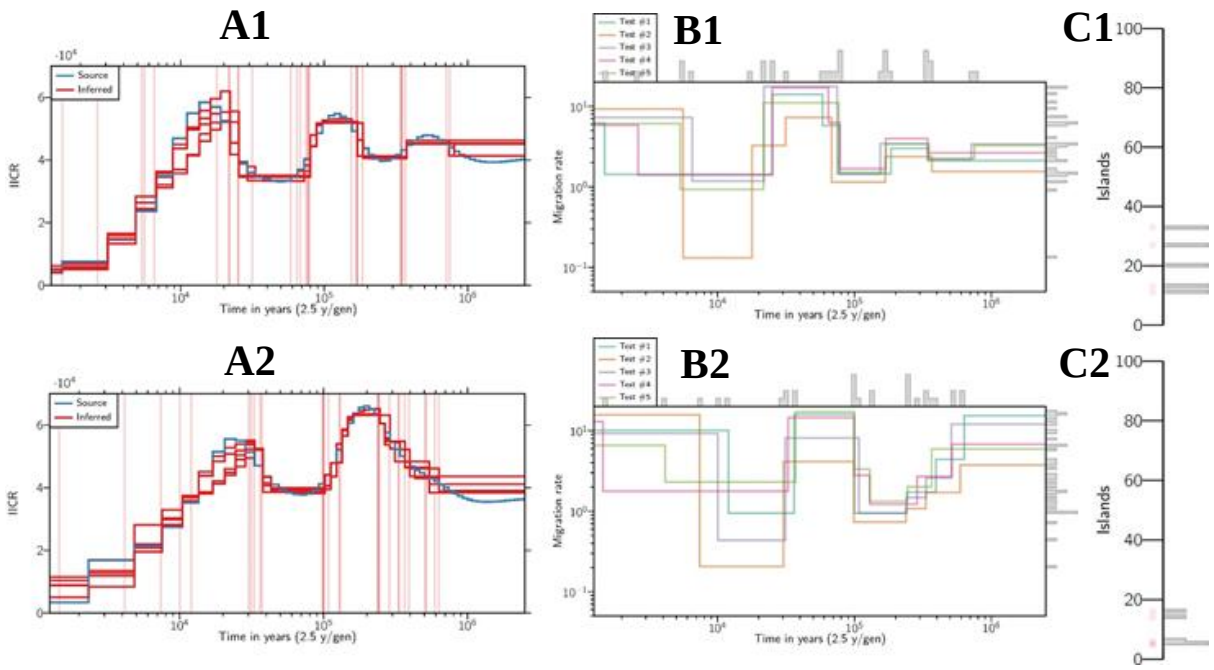

Figure 3.7 - SNIF inferences for *M. arnholdi* aligned with High Quality (A1, B1, C1) or *de novo* (A2, B2, C2) reference sequences – from *M. murinus* (1) and *M. arnholdi* (2), respectively. Images A correspond to the IICR plots, B to the connectivity graphs and C to the histograms with the number of islands/demes.

Finally, for *M. tavaratra* aligned with *M. murinus* HQ (Figure 3.8 - A1), SNIF found 3 clusters of possible demographic events - between 2 and 4 Kya, 70 to 200 Kya, and 400 to 500 Kya. The PSMC has a visible rise in the most ancient end of the plot: though probably meaningless, it had an impact on SNIF's inference of the latter hump (400 Kya to approx. 1.5 Mya). Regarding migration patterns, in Fig. 3.8 - B1, a period of high connectivity is noticed between 200 Kya and 400/500 Kya, which corresponds with the events detected in A1.
Analyzing the results from the alignment with the *M. tavaratra de novo* reference, the placement of demographic components in Fig. 3.8 - A2 is less clustered compared to Fig. 3.8 - A1, which used the High-Quality reference from *M. murinus*. Notably, the period of higher-probability demographic events (200 to 400 Kya) observed in Fig. 3.8 - A2 contrasts with the corresponding periods identified in Fig. 3.8 - A1. Furthermore, the prominent hump visible in Fig. 3.8 - A1, spanning 400 to 500 Kya, is markedly flattened and nearly absent in Fig. 3.8 - A2. These differences highlight the potential influence of reference genome choice on demographic inference results. In the connectivity plot (Fig. 3.8 - B2), we see a stable and high migration rate until around 300 Kya, where a decline is evident until reaching a plateau around 100 Kya.
The number of inferred demes is similar for both alignments - 10 to 15 islands.

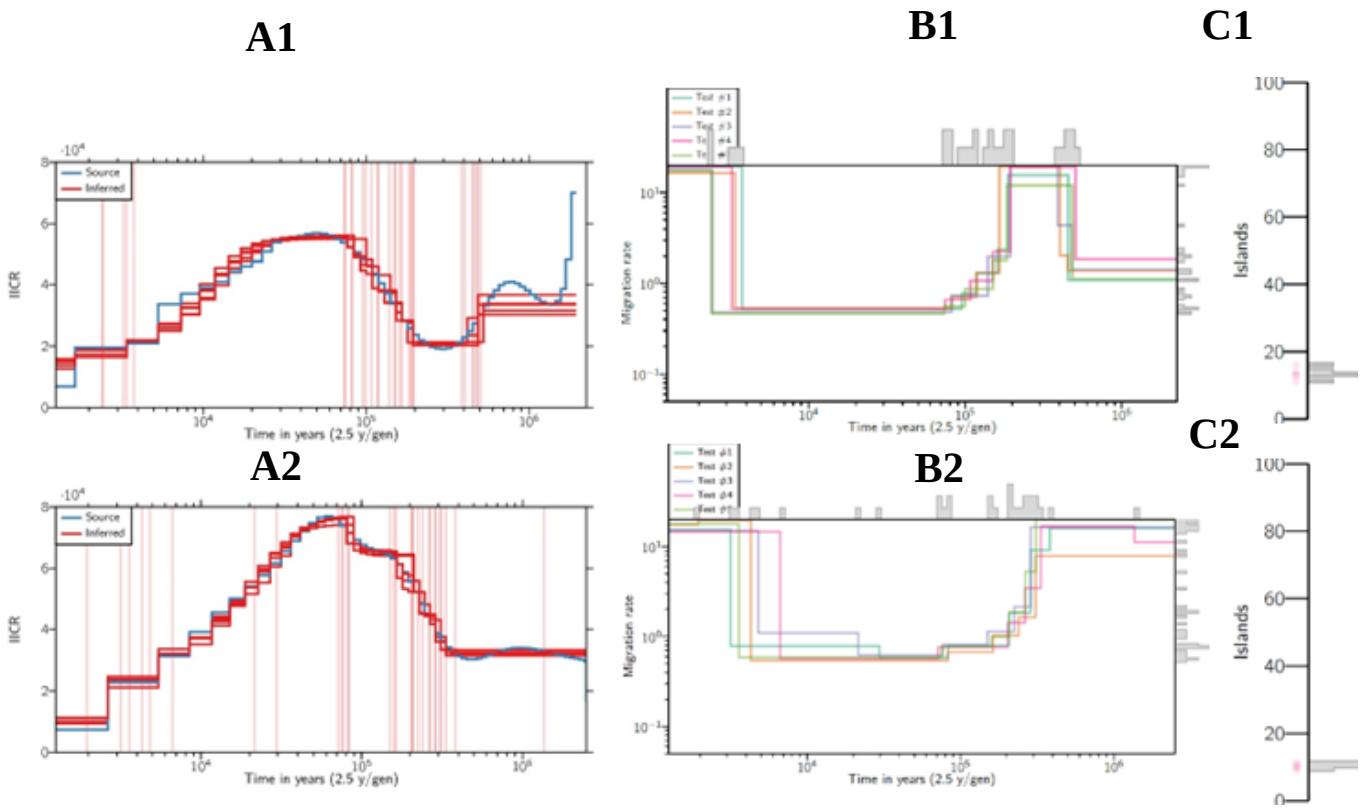

Figure 3.8 - SNIF inferences for *M. tavaratra* aligned with High Quality (A1, B1, C1) or *de novo* (A2, B2, C2) reference sequences – from *M. murinus* (1) and *M. tavaratra* (2), respectively. Images A corresponds to the IICR plot, B to the connectivity graph and C to the histogram with the number of islands/demes.

3.4. *ms* Validations

In order to validate the previous results, we simulated T2 values (coalescence time for a sample size of two) with the *ms* program and used these values to construct an IICR curve that is used by SNIF as a target for inference (the same way it uses a PSMC file). If SNIF identifies demographic events in the simulation that match the ones from the original PSMC, those will be the only events considered validated for Discussion - Table 4.1.

In Figure 3.9 we observe that *ms* succeeded in its simulations, though the simulated PSMCs should not be trusted beyond 900 Kya, further into the ancient past. For the alignment with *M. murinus* HQ (Fig. 3.9-A) we found 4 periods with validated events - at around 20 Kya, 70 to 80 Kya, around 200 Kya, and between 300 and 400 Kya - from which the latter 3 were also found for the *M. murinus de novo* alignment (Fig. 3.9-B). This different-species *de novo* alignment also validated 2 other events: one around 30 Kya and another between 700 and 900 Kya. Looking at the validation using the *M. arnholdi de novo* alignment (Fig. 3.9-C) we can see 4 major periods of events at 30-40 Kya, circa 100 Kya, another at 250 Kya, and lastly around 350-400 Kya. This last event is the only that seems to be common to all alignments of Figure 3.9.

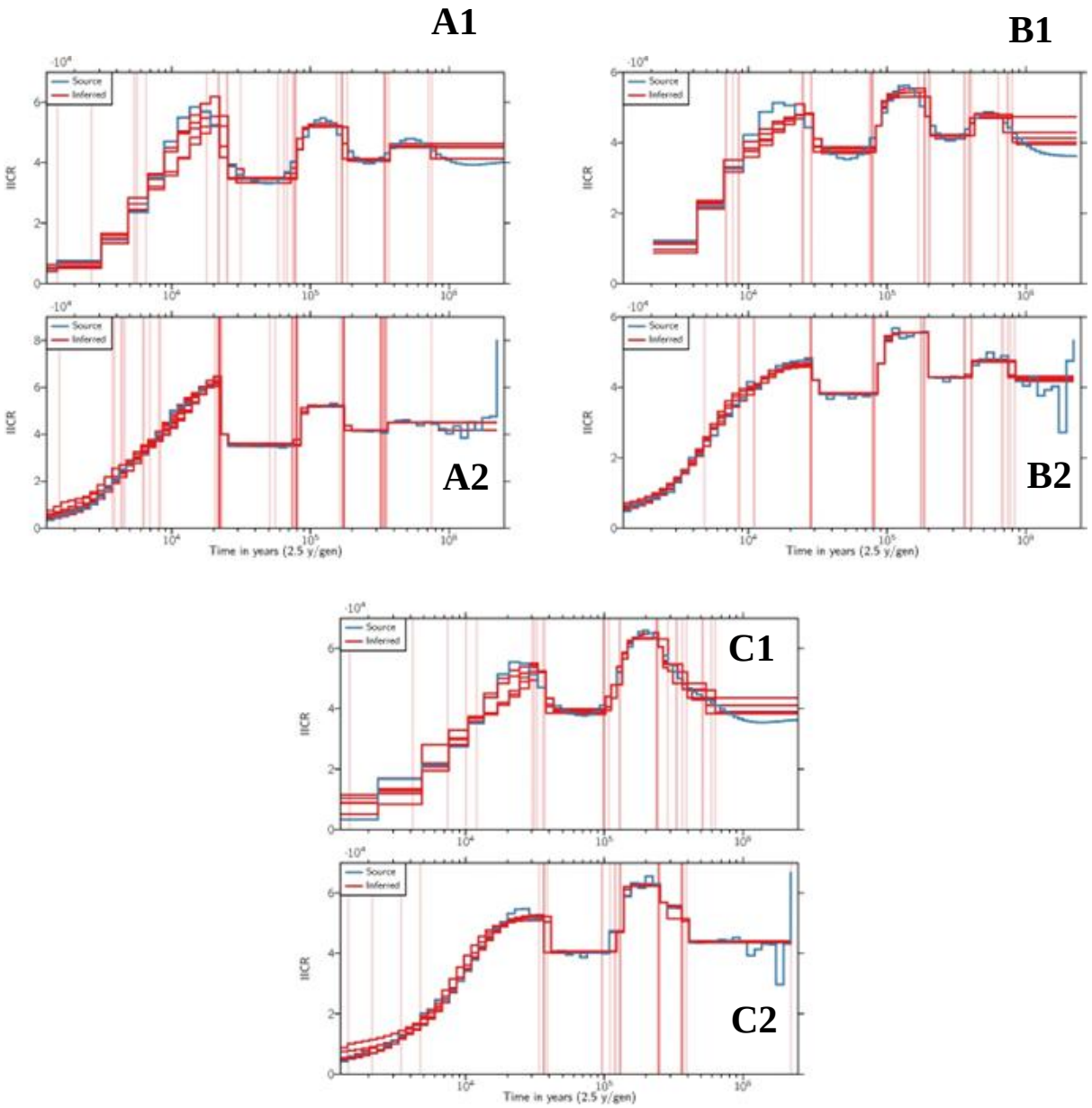

Figure 3.9 - SNIF inferences (1) and corresponding ms validations (2) for an *M. arnholdi* individual aligned with a *M. murinus* High Quality (A) and a *de novo* (B) references, and with a *M. arnholdi de novo* (C) reference sequence

Regarding the validations for the *M. murinus* individual, in Figure 3.10, there are no exact matches between events detected by SNIF in both alignments. For the validations with the HQ reference (Fig. 3.10-A) we can see 4 periods of events - circa 50 Kya, 125 Kya, 300-380 Kya, and 470-600 Kya. We found again 4 validated events but at different periods for the *de novo* alignment (Fig. 3.10-B): 35 Kya, 110 Kya, 400 Kya, and 550-620 Kya.

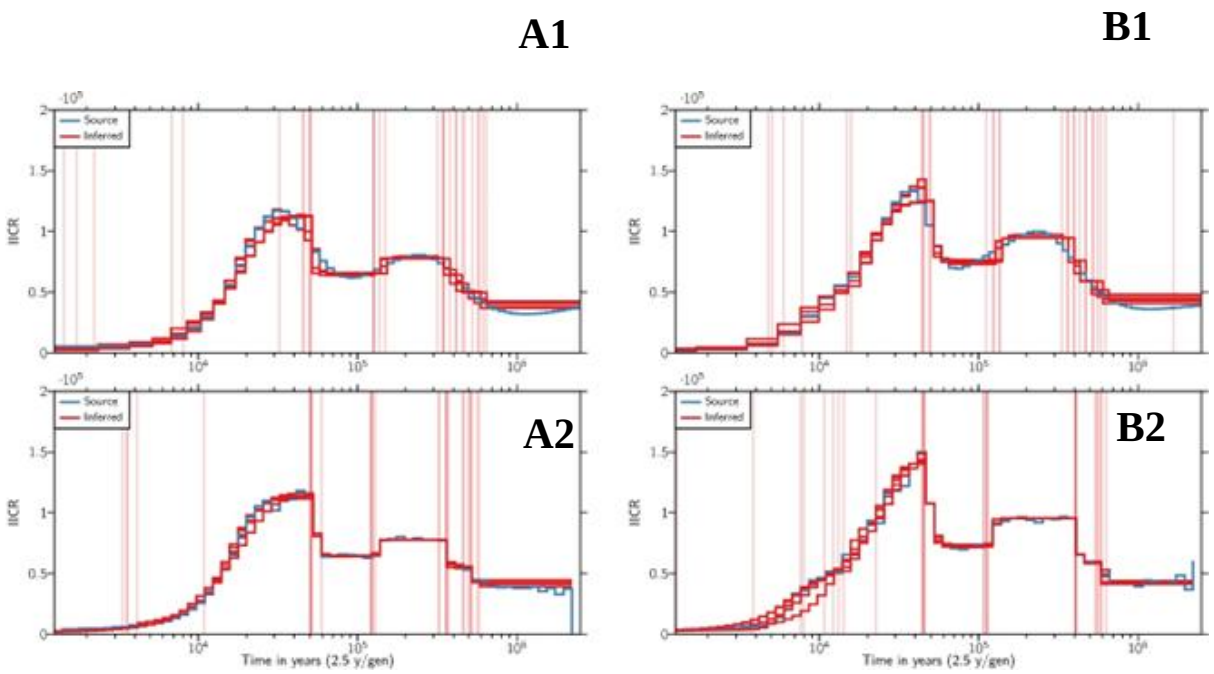

Figure 3.10 - SNIF inferences (1) and corresponding ms validations (2) for an *M. murinus* individual aligned with a *M. murinus* High Quality (A) and a *de novo* (B) reference sequences

In Figure 3.11, it is clear that the ms simulation failed when simulating the IICR for *M. tavaratra* using the *M. tavaratra de novo* reference (Fig. 3.11 - B2). Although the simulation appears to just be shifted to the left and there may be some overlap between events, we decided to disregard the simulation results and consider all the putative events in the discussion.

For the alignment with the high-quality reference from a different species (Fig. 3.11 - A2), the placement of events is scattered, especially in the more recent time periods. This makes it difficult to draw clear conclusions compared to the original inference (Fig. 3.11 - A1). However, the two periods with a high frequency of probable demographic events observed in the IICR inference—between 70 Kya and 200 Kya, and from 400 to 500 Kya—are closely reflected in the simulated pseudo-IICR (Fig. 3.11 - A2).

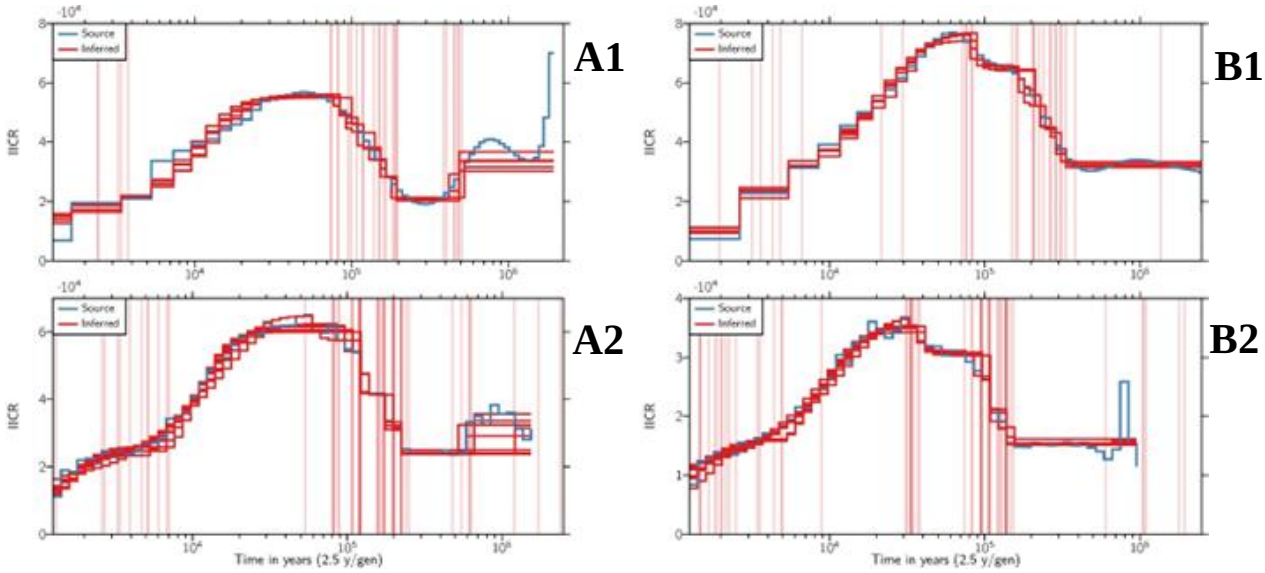

Figure 3.11 - SNIF inferences (1) and corresponding ms validations (2) for an *M. tavaratra* individual aligned with a *M. murinus* High Quality (A) and an *M. tavaratra de novo* (B) reference sequences.

4. Discussion

This discussion begins with our attempt to replicate the PSMC analysis of *Microcebus arnholdi* from specific literature, providing a foundational understanding of the species' demographic history. Next, we optimized the inference parameters for SNIF analyses on *Microcebus* spp., assessing the impact of reference divergence on the PSMC outputs, which highlighted key challenges in reconciling genomic signals across species. Our primary focus was to infer the demographic histories of *M. arnholdi*, *M. murinus*, and *M. tavaratra* under n-island models using SNIF, uncovering patterns of connectivity among these species over time. To validate our method, we re-ran SNIF on model-based simulated genomic data using the ms coalescent simulation software, ensuring the robustness of our inferences. Finally, we crossmatched the inferred demographic histories with paleoclimatic events, shedding light on how external environmental factors may have influenced historical gene flow and population structure in *Microcebus* spp.

4.1. PSMC Replication

The Federation of American Societies for Experimental Biology (FASEB), in its 2016 report "*Enhancing Research Reproducibility*", defined the concepts of *Replicability* and *Reproducibility* (FASEB, 2016). From that report, Barba (2018) has synthesized *Replicability* as "*the ability to duplicate (i.e., repeat) a prior result using the same source materials and methodologies. This term should only be used when referring to repeating the results of a specific experiment rather than an entire study*". Also, the concept of *Reproducibility* was defined as "*the ability to achieve similar or nearly identical results using comparable materials and methodologies. This term may be used when specific findings from a study are obtained by an independent group of researchers*" (Barba, 2018). From another point of view, *in extremis*, one can argue that making 100% reproducible science is not possible due to several factors, such as time, resources, systemic pressure to publish (or perish) or even human nature.

When looking at the results we obtained in the process of using the published parameters and ours for generating PSMCs with the original genomic data from Teixeira et al. (2021) we concluded that that experiment was not easily replicable or reproducible on the basis of the information published. We should also stress that the differences between the published results and ours - while visible - were not serious, the main issue being the use of *M. murinus* reference sequence, as we discuss below. The images from Figure 3.1 attest our attempts to find if any of the different parameters - from Table 3.1, whether individually or combined among them - could directly explain the differences in results. If the difference between image A (the original result) and image C from Fig. 3.1 indicates some challenges in replicating the experiment, the other images (D to I) from the same figure provide additional context for our observations about non-reproducibility.

The proximity of the generated PSMC/IICR curves to the curves of the Mahasarika individual and not to the Fantany individual raised doubts. A hypothetical explanation could be that there was a mix-up attributing the accession numbers of the genomic data of the individuals in NCBI's BioSample database, but we have not confirmed this with the authors of the paper who performed the analyses, nor did we try to replicate the PSMC of the individual from Mahasarika to test for similar results.

Our findings in this thesis do not contradict the results obtained by Teixeira et al. (2021); instead, they illustrate the importance of peer review and reproduction of results in strengthening scientific research.

4.2. Methods and Limitations on Inferences

In the section 3.2, we exemplified the parameter optimization process for SNIF inferences with focus on the number of components (c) and weight (ω) parameters. If the number of components impacts the inference in the number of demographic events and their placing in time, the weight parameter changes SNIF's focus on reducing the distance between the observed and simulated IICR in specific periods, depending on the weight/value the user gave to the recent or ancient past.

We concluded that the best inferences from SNIF used c = 7 and ω = 0.5. Our results support evidence found by (Steux, 2023) that increasing the number of components improves the fit of observed PSMC data, especially in more complex demographic histories. While Arredondo et al. (2021) recommend using user discretion to choose the optimal number of components based on the specific dataset, our findings suggest that a higher number of components may provide a better fitting in certain cases. However, in Figures 3.2 and 3.3 we also tested an 8th component that SNIF always placed at the beginning of the IICR curve. This raised doubts about SNIF's ability to place more than 7 demographic events for these species and, thus, the existence of an 8th component. In addition, we have seen that the improvement of the results with the number of components was independent of the species that was used as reference for the alignment. Regarding the weight (ω) parameter, choosing the value of 0.5 served better our interests of not giving much relevance to the recent past.

Nevertheless, this parameter optimization process improved the fitting of the PSMCs - as our results in section 3.2 confirm - despite some periods remaining ill-fitting. This may be due to our model not capturing the complexity of real populations: their structure is probably not exactly an n-island model, which does not include population size changes. The stochastic aspect of coalescence can also sometimes create false signals in the PSMC, which can then be difficult to fit. The clearest example of this is the PSMC of *M. tavaratra* aligned with the high quality reference of *M. murinus* (Figure 3.8), where an exaggerated rise in the ancient end of the PSMC curve (Fig. 3.8-A1) and flattened features in the *de novo* alignment (Fig. 3.8-B1) illustrate how stochastic noise can distort demographic inferences and its consequential effects on the placing of events by SNIF. Thus, it becomes evident that this manual process of parameter optimization is essential to guide SNIF on what-to and not-to infer for each species.

The validations we obtained using "ms" cannot be read as definite proof that the demographic events identified by SNIF occurred exactly at that period or happened at all. What these validations are supposed to give us is confidence that the model inferred by SNIF using the pseudo-observed IICR (IICRpods) also has consistency for the real genomic data. In other words, "ms" validates the SNIF approach as a method, not the veracity of its results. Thus, the main goal of this thesis is to propose demographic histories by matching known paleo climatic events with the results inferred using SNIF, as the results of our validation step showed that SNIF was able to recover scenarios as complex as those originally inferred with real data, except for the de novo alignment of *M. tavaratra*.

The PSMC method is an imperfect estimate with limitations - some addressed in this thesis, some addressed in other studies (Henrique, n.d.) (*in prep*) - such as the quality of the reference genome, the divergence of the species to which that reference genome belongs, or even its reliance (by definition) on a single individual for presenting a demographic history for the whole species. For this latter question, we must recognize that sampling individuals from the same or proximal locations might impact the PSMC results by presenting a possibly false harmony among observations when specifically studying population structure.

Regarding the quality of the reference genome, Figures 3.5 and 3.6 showed that the number of humps in the PSMCs did not vary in function of the type of genomic reference - *de novo* or HQ. Thus, it becomes a question of divergence of the species used as reference for the alignment. As mentioned in the introduction, a critical factor affecting the accuracy of PSMC inferences is the phylogenetic distance between the target species and the reference genome. We used *de novo* reference genomes from *Microcebus murinus, Microcebus tavaratra* and *Microcebus arnholdi* to align our target sequences in an attempt to overcome the divergence limitations. Furthermore, SNIF-estimated events should be interpreted cautiously when falling within a lower-confidence region of the PSMC curve, as suggested by (Li & Durbin, 2011), when transposed to a generation time of 2.5 years and a mutation rate of 1.5×10^{-8} per generation. Specifically, PSMC is less reliable for events older than approximately 500,000 years in *Microcebus*, corresponding to 200,000 generations given these parameters.

Finally, an *n-island* model is useful to infer population structure in a simpler fashion due to the consistent size (N) of demes/islands and constant migration rate (m). This must also be recognized as a limitation, since other models like the *stepping-stone* model account for variable deme sizes and migration rates.

We must emphasize that this discussion, like all scientific discussions, is based on the results we obtained, while recognizing the aforementioned limitations. Consequently, we offer several perspectives and potential demographic histories that, while they may not fully reflect the exact sequence of events that shaped these species' current state, have been inferred to the best of our knowledge using innovative tools like SNIF.

4.3. Demographic history of *Microcebus*

In this thesis, we aimed to infer demographic histories under *n-island* models for three species of mouse lemurs using SNIF. We considered important to address the issue of a High Quality reference versus a *de novo* reference, and it's impact on the PSMC and, thus, on SNIF. This question arises due to the scarcity of published High Quality genomic references for endangered species like *M. arnholdi* or *M. tavaratra*. When there is a need for a sequence alignment/mapping for these species, the rule of thumb has been to use a High Quality genomic reference from a sister-species - in this case, the species *M. murinus*. Since the literature and works from our group (Henrique, n.d.) (*in prep*) showed that using a *de novo* reference seems to be more trustworthy than using a divergent HQ reference, we have decided to mainly focus this discussion on the results we obtained using the *de novo* references for each species with periodic remarks to put the results that used a High Quality reference in perspective - e.g. comparing of SNIF inferences of demographic events, IICR differences, or number of demes.

Before delving in the demographic histories, we will briefly review what we know regarding the phylogenies and current spatial distribution of these *Microcebus* species. Phylogenies from nuclear genes showed that *M. arnholdi* and *M. tavaratra* diverged around 490 Kya, while their MRCA diverged from *M. murinus* around 1.5 Mya (van Elst et al., 2024). *M. arnholdi* is microendemic to the Montagne d'Ambre, a humid forest habitat in the northern region of Madagascar, the other two species prefer dry forest environments and are located in wider regions in the north-northeast (*M. tavaratra*) and west (*M. murinus*) of Madagascar. Given that *M. tavaratra* is the only dry environment species in its clade, (van Elst et al., 2024) assumed that the MRCA of *M. arnholdi* and *M. tavaratra* probably lived in a humid environment habitat.

In Table 4.1 we summarized the validated demographic events that SNIF identified for each species, considering only same-species *de novo* alignments. One of the goals of this thesis was to identify patterns of common genomic signatures that could point towards shared demographic events among different *Microcebus spp.* Knowing that the MRCA of *Microcebus arnholdi* and *Microcebus tavaratra* was present around 490 Kya, we used that date as a reference for where to discern putative events in the PSMCs of those species. In Figure 4.1, that speciation time is graphically displayed with a purple vertical line, whereas the red vertical lines show possible demographic change events.

Table 4.1 - Summary of validated demographic events identified by SNIF for each species (same-species *de novo* alignments)

Alignment \ Events (Kya)	30	35	40	70	80	90	100	110	200	250	300	350	400	550	620
M. arnholdi	■	■	■				■			■		■	■		
M. murinus								■						■	■
M. tavaratra				■	■	■			■	■	■	■	■		

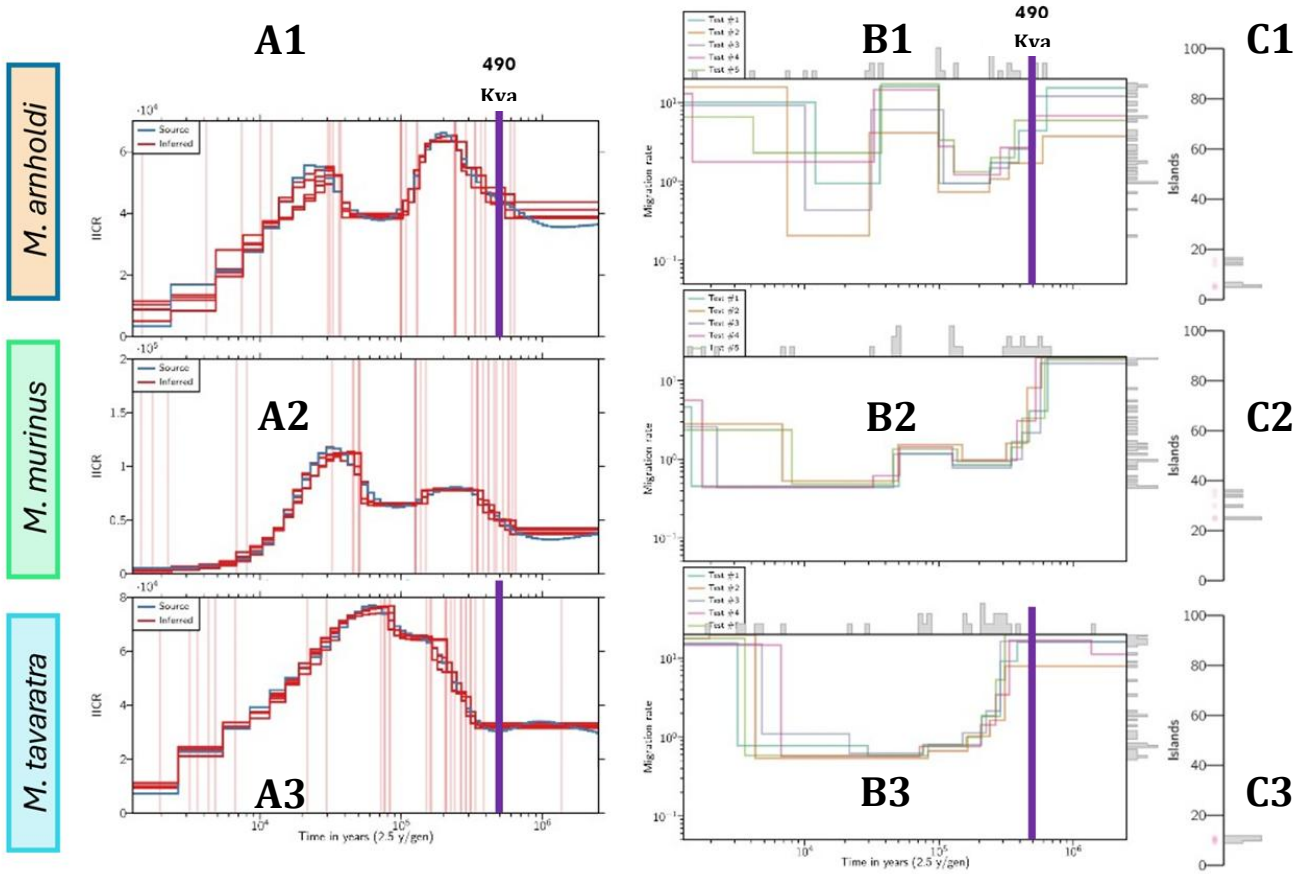

Figure 4.1 – SNIF inferences of *d*emographic events (A), connectivity change times (B) and number of demes (C) for *Microcebus arnholdi* (1), *M. murinus* (2), and *M. tavaratra* (3) aligned with same-species *de novo* references

Regarding connectivity, *Microcebus arnholdi* exhibits a marked decrease in migration rates starting around the speciation time of 500 Kya (Fig. 4.1-B1). This trend is similarly observed in *M. murinus* (Fig. 4.1-B2), suggesting a shared demographic event impacting both species. However, *M. tavaratra*

(Fig. 4.1-B3) shows a delayed response, with its decline in migration rates occurring between 200 and 300 Kya. This temporal gap raises questions about the distinct demographic pressures acting on *M. tavaratra* and why it remained more connected for a longer period. One possible explanation lies in their differing habitats - *M. arnholdi* is confined to the humid forests of Montagne d'Ambre, whereas *M. tavaratra* inhabits a broader range in drier environments, which may have shielded it from earlier environmental changes.

The divergence in migration trends among these species can also be understood by examining their phylogenetic histories in relation to current spatial distributions. The geographic isolation of *M. arnholdi* in Montagne d'Ambre, coupled with a wider range of stable dry habitats for *M. tavaratra*, likely influenced their respective demographic histories. This isolation could explain why *M. arnholdi* experienced earlier changes in connectivity, while *M. tavaratra* remained unaffected until much later.

When clustering the species based on SNIF-inferred demographic events, *M. arnholdi* and *M. murinus* display a synchronous decline in connectivity around the same period, suggesting they were similarly impacted by environmental or demographic changes. In contrast, *M. tavaratra* forms a separate cluster due to its delayed migration rate decline, possibly reflecting slower or different ecological pressures in its habitat. Environmental factors, particularly volcanic activity in the Northern Madagascar Alkaline Province (NMAP) during the Cenozoic, offer valuable insights into the demographic shifts observed in *M. arnholdi*. Volcanic eruptions less than 1 million years ago may have created varied altitudes and fertile soils, promoting the development of humid forests, especially in Montagne d'Ambre (Pratt et al., 2017). This unique habitat likely provided a favorable environment for *M. arnholdi* to settle and adapt, potentially driving its earlier demographic shift compared to *M. tavaratra*, which occupies a broader, drier range. The geographic isolation of *M. arnholdi*, alongside its habitat specialization, may have contributed to its earlier decline in migration rates, while the less isolated *M. tavaratra* experienced later changes in connectivity. This volcanic activity and the resulting ecological conditions may have played a significant role in shaping the population structures and demographic histories of these species.

It is also essential to consider the relationship between migration rates and effective population sizes. As gene flow decreases, the IICR (or, in some cases, effective population size) of structured populations tends to increase (Mazet et al., 2015; Nei & Takahata, 1993; Wakeley, 2001). For both *M. arnholdi* and *M. murinus*, this dynamic is evident, with decreased connectivity aligning with periods of population structuring. Conversely, *M. tavaratra* maintained higher connectivity for longer, which may explain its higher genetic diversity, and larger distribution area (Sgarlata et al., 2018). Interestingly, *M. murinus*, despite its wider distribution, exhibits the lowest inferred IICR among the three species. This outcome may seem counterintuitive but highlights the importance of considering population structure. *M. murinus* has more inferred demes than the other species, indicating considerable spatial structuring. This prevents assuming panmixia, meaning the IICR, influenced by restricted gene flow among demes, cannot be equated to the metapopulation Ne.

Furthermore, it is crucial to integrate insights from paleoenvironmental shifts, particularly the significant population decline of *M. arnholdi* around the mid-Holocene (~5.5 Kya), which coincided with a reduction in precipitation following the termination of the African Humid Period (AHP) (11.8–5.5 Kya). On the other hand, the connectivity of *M. tavaratra* and *M. murinus* appears to increase precisely at the end of the AHP, which could be explained by their preference for drier environments. These changes occurred long before significant human impacts, which are believed to have started influencing the environment more substantially around 1–2 Kya in northern Madagascar (Crowley, 2010).

5. Conclusions and Perspectives

The field of Population Genomics is a vast world in constant update. This thesis project shows exactly how what is taught in lectures is just the tip of the iceberg on the array of methods, perspectives and knowledge that an Evolutionary Biologist must master for this specific branch of science. But not only one must master those skills, one must remain aware of the forwards and backwards that global scientific efforts output on the latest literature and communications. This should be no surprise, since "Questioning" - whether about observed phenomena or the fundaments of current knowledge - is the very first step of the Scientific Method. Thus, inquiring on the validity of the "how" (the current methods) and its consequences on the "what" (the objects of scientific interest) is crucial towards the evolution of the field itself.

We have found that the amount of information on the methods and data used in the literature can play a great role when one tries to validate that literature, whether through replication or reproduction of the experiments and results. Efforts towards reproducible/replicable science should aim towards continuous improvement and not immediate excellence, as the latter increases scientific work's vulnerability to errors and missing details.

The decrease in gene flow for *M. arnholdi, M. murinus,* and *M. tavaratra* aligns with periods of habitat fragmentation and climatic fluctuations in Madagascar. These findings underscore the impact of environmental changes on the genetic diversity and population structure of endangered species. The sharp decline in connectivity observed in *M. arnholdi*, particularly, highlights the vulnerabilities of microendemic species to environmental shifts, while *M. tavaratra*'s delayed demographic response suggests resilience linked to its broader habitat range.

Our study also demonstrates the utility of the Structured Non-stationary Inference Framework (SNIF) in inferring demographic histories from genomic data. The ability to model non-stationary population structures and connectivity changes offers a powerful tool for conservation biology, enabling more accurate assessments of population dynamics and extinction risks.

Future research should focus on obtaining High-Quality genomic sequences - or mainstream the use of *de novo* references as recommended by Henrique et al. (*in prep*) - for a wider range of *Microcebus* species to improve the accuracy of demographic inferences. Additionally, developing models that account for variable deme sizes and migration rates - e.g. stepping-stone - could provide a more nuanced understanding of population dynamics than the *n-island* model. To further develop SNIF, it would be of interest to create a Graphical User Interface (GUI) for researchers without programming skills and to potentially automatize the optimization of SNIF's parameter space in a fashion similar to Bayesian inferences. Studies like Steux's (2023) have tried to shed some light on the previously mentioned matter of single individual bias on PSMCs, thus augmenting the number and location of sampled individuals to be of interest for the research on *Microcebus* population structure.

Ultimately, the findings of this thesis contribute to the broader understanding of conservation genomics, emphasizing the need for interdisciplinary approaches that incorporate ecological, climatic, and genetic perspectives. By employing innovative tools like SNIF, researchers can gain deeper insights into the complex dynamics of population structure and demographic history, which are critical for informing effective conservation strategies.

References

Arredondo, A., Mourato, B., Nguyen, K., Boitard, S., Rodríguez, W., Noûs, C., Mazet, O., & Chikhi, L. (2021). Inferring number of populations and changes in connectivity under the n-island model. *Heredity 2021 126:6*, *126*(6), 896–912. https://doi.org/10.1038/s41437-021-00426-9

Barba, L. A. (2018). *Terminologies for Reproducible Research*. https://arxiv.org/abs/1802.03311v1

Black IV, W. C., Baer, C. F., Antolin, M. F., & DuTeau, N. M. (2003). POPULATION GENOMICS: Genome-Wide Sampling of Insect Populations. *Https://Doi.Org/10.1146/Annurev.Ento.46.1.441*, *46*, 441–469. https://doi.org/10.1146/ANNUREV.ENTO.46.1.441

Ceballos, G., Ehrlich, P. R., Barnosky, A. D., García, A., Pringle, R. M., & Palmer, T. M. (2015). Accelerated modern human-induced species losses: Entering the sixth mass extinction. *Science Advances*, *1*(5). https://doi.org/10.1126/SCIADV.1400253/SUPPL_FILE/1400253_SM.PDF

Charlesworth, B. (2010). Molecular population genomics: a short history. *Genetics Research*, *92*(5–6), 397–411. https://doi.org/10.1017/S0016672310000522

Chikhi, L., Rodríguez, W., Grusea, S., Santos, P., Boitard, S., & Mazet, O. (2017). The IICR (inverse instantaneous coalescence rate) as a summary of genomic diversity: insights into demographic inference and model choice. *Heredity 2018 120:1*, *120*(1), 13–24. https://doi.org/10.1038/s41437-017-0005-6

Chikhi, L., Sousa, V. C., Luisi, P., Goossens, B., & Beaumont, M. A. (2010). The Confounding Effects of Population Structure, Genetic Diversity and the Sampling Scheme on the Detection and Quantification of Population Size Changes. *Genetics*, *186*(3), 983–995. https://doi.org/10.1534/GENETICS.110.118661

Couloigner, C. (2022). *Comparative demographic histories of endangered Southeast Asian primate species from genomic data: real data, inferences and simulations*.

Crowley, B. E. (2010). A refined chronology of prehistoric Madagascar and the demise of the megafauna. *Quaternary Science Reviews*, *29*(19–20), 2591–2603. https://doi.org/10.1016/J.QUASCIREV.2010.06.030

Ellegren, H. (2014). Genome sequencing and population genomics in non-model organisms. *Trends in Ecology & Evolution*, *29*(1), 51–63. https://doi.org/10.1016/J.TREE.2013.09.008

FASEB. (2016). *National Academies Releases Report on Reproducibility and Replicability in Science | FASEB*. https://www.faseb.org/journals-and-news/washington-update/national-academies-releases-report-on-reproducibility-and-replicability-in-science

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *222*(594–604), 309–368. https://doi.org/10.1098/RSTA.1922.0009

Gagneux, P., & Varki, A. (2001). Genetic Differences between Humans and Great Apes. *Molecular Phylogenetics and Evolution*, *18*(1), 2–13. https://doi.org/10.1006/MPEV.2000.0799

Giani, A. M., Gallo, G. R., Gianfranceschi, L., & Formenti, G. (2020). Long walk to genomics: History and current approaches to genome sequencing and assembly. *Computational and Structural Biotechnology Journal*, *18*, 9–19. https://doi.org/10.1016/J.CSBJ.2019.11.002

Gillespie, J. H. (2004). *Population Genetics*. https://doi.org/10.56021/9780801880087

Goodman, S. M., & Benstead, J. P. (2005). Updated estimates of biotic diversity and endemism for Madagascar. *Oryx*, *39*(1), 73–77. https://doi.org/10.1017/S0030605305000128

Henrique, M. (n.d.). *in prep*.

Hohenlohe, P. A., Funk, W. C., & Rajora, O. P. (2021). Population genomics for wildlife conservation and management. *Molecular Ecology*, *30*(1), 62–82. https://doi.org/10.1111/MEC.15720

Hudson, R. R. (2002). Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics*, *18*(2), 337–338. https://doi.org/10.1093/BIOINFORMATICS/18.2.337

IUCN. (2020, July). *Almost a third of lemurs and North Atlantic Right Whale now Critically Endangered - IUCN Red List | IUCN*. https://www.iucn.org/news/species/202007/almost-a-third-lemurs-and-north-atlantic-right-whale-now-critically-endangered-iucn-red-list

Kingman, J. F. C. (1982). The coalescent. *Stochastic Processes and Their Applications*, *13*(3), 235–248. https://doi.org/10.1016/0304-4149(82)90011-4

Li, H., & Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature 2011 475:7357*, *475*(7357), 493–496. https://doi.org/10.1038/nature10231

Liu, S., & Hansen, M. M. (2017). PSMC (pairwise sequentially Markovian coalescent) analysis of RAD (restriction site associated DNA) sequencing data. *Molecular Ecology Resources*, *17*(4), 631–641. https://doi.org/10.1111/1755-0998.12606

Louis, E. E., Engberg, S. E., McGuire, S. M., McCormick, M. J., Randriamampionona, R., Ranaivoarisoa, J. F., Bailey, C. A., Mittermeier, R. A., & Lei, R. (2008). Revision of the Mouse Lemurs, Microcebus (Primates, Lemuriformes), of Northern and Northwestern Madagascar with Descriptions of Two New Species at Montagne d'Ambre National Park and Antafondro Classified Forest. *Https://Doi.Org/10.1896/052.023.0103*, *23*(1), 19–38. https://doi.org/10.1896/052.023.0103

Luikart, G., England, P. R., Tallmon, D., Jordan, S., & Taberlet, P. (2003). The power and promise of population genomics: from genotyping to genome typing. *Nature Reviews Genetics 2003 4:12*, *4*(12), 981–994. https://doi.org/10.1038/nrg1226

Mazet, O., Rodríguez, W., Grusea, S., Boitard, S., & Chikhi, L. (2015). On the importance of being structured: instantaneous coalescence rates and human evolution—lessons for ancestral population size inference? *Heredity 2016 116:4*, *116*(4), 362–371. https://doi.org/10.1038/hdy.2015.104

Myers, N., Mittermeler, R. A., Mittermeler, C. G., Da Fonseca, G. A. B., & Kent, J. (2000). Biodiversity hotspots for conservation priorities. *Nature 2000 403:6772*, *403*(6772), 853–858. https://doi.org/10.1038/35002501

Nei, M., & Takahata, N. (1993). Effective population size, genetic diversity, and coalescence time in subdivided populations. *Journal of Molecular Evolution*, *37*(3), 240–244. https://doi.org/10.1007/BF00175500

Okazaki, A., Yamazaki, S., Inoue, I., & Ott, J. (2021). Population genetics: past, present, and future. *Human Genetics*, *140*(2), 231–240. https://doi.org/10.1007/S00439-020-02208-5/METRICS

Prasad, A., Lorenzen, E. D., & Westbury, M. V. (2022). Evaluating the role of reference-genome phylogenetic distance on evolutionary inference. *Molecular Ecology Resources*, *22*(1), 45–55. https://doi.org/10.1111/1755-0998.13457

Pratt, M. J., Wysession, M. E., Aleqabi, G., Wiens, D. A., Nyblade, A. A., Shore, P., Rambolamanana, G., Andriampenomanana, F., Rakotondraibe, T., Tucker, R. D., Barruol, G., & Rindraharisaona, E. (2017). Shear velocity structure of the crust and upper mantle of Madagascar derived from surface wave tomography. *Earth and Planetary Science Letters*, *458*, 405–417. https://doi.org/10.1016/J.EPSL.2016.10.041

Quéméré, E., Hibert, F., Miquel, C., Lhuillier, E., Rasolondraibe, E., Champeau, J., Rabarivola, C., Nusbaumer, L., Chatelain, C., Gautier, L., Ranirison, P., Crouau-Roy, B., Taberlet, P., & Chikhi, L. (2013). A DNA Metabarcoding Study of a Primate Dietary Diversity and Plasticity across Its Entire Fragmented Range. *PLOS ONE*, *8*(3), e58971. https://doi.org/10.1371/JOURNAL.PONE.0058971

Radespiel, U., Lutermann, H., Schmelting, B., & Zimmermann, E. (2019). An empirical estimate of the generation time of mouse lemurs. *American Journal of Primatology*, *81*(12), e23062. https://doi.org/10.1002/AJP.23062

Rodríguez, W., Mazet, O., Grusea, S., Arredondo, A., Corujo, J. M., Boitard, S., & Chikhi, L. (2018). The IICR and the non-stationary structured coalescent: towards demographic inference with arbitrary changes in population structure. *Heredity 2018 121:6*, *121*(6), 663–678. https://doi.org/10.1038/s41437-018-0148-0

Sgarlata, G. M., Salmona, J., Aleixo-Pais, I., Rakotonanahary, A., Sousa, A. P., Kun-Rodrigues, C., Ralantoharijaona, T., Jan, F., Zaranaina, R., Rasolondraibe, E., Zaonarivelo, J. R., Andriaholinirina, N. V., & Chikhi, L. (2018). Genetic Differentiation and Demographic History of the Northern Rufous Mouse Lemur (Microcebus tavaratra) Across a Fragmented Landscape in Northern Madagascar. *International Journal of Primatology*, *39*(1), 65–89. https://doi.org/10.1007/S10764-018-0015-0/TABLES/2

Steux, C. (2023). *Comparative demography of endangered primate species in fragmented habitat using genomic data: inference, simulation and real data*. PSL Université Paris.

Teixeira, H., Montade, V., Salmona, J., Metzger, J., Bremond, L., Kasper, T., Daut, G., Rouland, S., Ranarilalatiana, S., Rakotondravony, R., Chikhi, L., Behling, H., & Radespiel, U. (2021). Past environmental changes affected lemur population dynamics prior to human impact in Madagascar. *Communications Biology 2021 4:1*, *4*(1), 1–10. https://doi.org/10.1038/s42003-021-02620-1

van Elst, T., Sgarlata, G. M., Schüßler, D., Tiley, G. P., Poelstra, J. W., Scheumann, M., Blanco, M. B., Aleixo-Pais, I. G., Evasoa, M. R., Ganzhorn, J. U., Goodman, S. M., Hasiniaina, A. F., Hohenlohe, P. A., Ibouroi, M. T., Jan, F., Kappeler, P. M., Pors, B. Le, Manzi, S., Olivieri, G., … Radespiel, U. (2023). An integrative and generalizable approach to elucidate cryptic diversifications sheds

light on mouse lemur taxonomy and evolution. *Journal/Source Name*, *Volume*(Issue), pages. https://doi.org/DOI (if available)

van Elst, T., Sgarlata, G. M., Schüßler, D., Tiley, G. P., Poelstra, J. W., Scheumann, M., Blanco, M. B., Aleixo-Pais, I. G., Rina Evasoa, M., Ganzhorn, J. U., Goodman, S. M., Hasiniaina, A. F., Hending, D., Hohenlohe, P. A., Ibouroi, M. T., Iribar, A., Jan, F., Kappeler, P. M., Le Pors, B., … Salmona, J. (2024). Integrative taxonomy clarifies the evolution of a cryptic primate clade. *Nature Ecology & Evolution 2024*, 1–16. https://doi.org/10.1038/s41559-024-02547-w

Vieilledent, G., Grinand, C., Rakotomalala, F. A., Ranaivosoa, R., Rakotoarijaona, J. R., Allnutt, T. F., & Achard, F. (2018). Combining global tree cover loss data with historical national forest cover maps to look at six decades of deforestation and forest fragmentation in Madagascar. *Biological Conservation*, *222*, 189–197. https://doi.org/10.1016/J.BIOCON.2018.04.008

Wakeley, J. (2001). The Coalescent in an Island Model of Population Subdivision with Variation among Demes. *Theoretical Population Biology*, *59*(2), 133–144. https://doi.org/10.1006/TPBI.2000.1495

Wakeley, J. (2004). Metapopulations and Coalescent Theory. *Ecology, Genetics and Evolution of Metapopulations*, 175–198. https://doi.org/10.1016/B978-012323448-3/50010-6

Wakeley, J. (2013). Coalescent theory has many new branches. *Theoretical Population Biology*, *87*(1), 1–4. https://doi.org/10.1016/J.TPB.2013.06.001

Wright, S. (1931). EVOLUTION IN MENDELIAN POPULATIONS. *Genetics*, *16*(2), 97–159. https://doi.org/10.1093/GENETICS/16.2.97

Supplementary Information I

PSMC Inference Tutorial

Adapted from Margarida Henrique's "Introduction to bioinformatic and genomic analyses for PSMC inference". Original version available by request to Lounès Chikhi or Margarida Henrique or via GitHub: https://github.com/PopConGen/intro_genomics_PSMC

- Software:
  - **FastQC** - For quality control of raw sequence data.
  - **Trim Galore** - For quality trimming and adapter removal in sequencing reads.
  - **Cutadapt** - Used by Trim Galore for adapter trimming.
  - **BWA (Burrows-Wheeler Aligner)** - For aligning sequence reads to a reference genome.
  - **Picard** - For sorting BAM files and marking duplicate reads.
  - **Qualimap** - For quality control of alignment sequencing data (e.g., coverage, mapping statistics).
  - **Samtools** - For manipulating SAM/BAM files (e.g., creating pileups).
  - **Bcftools** - For processing VCF/BCF files and consensus sequence calling.
  - **vcfutils.pl** - For filtering variants and converting VCF files to FASTQ.
  - **PSMC (Pairwise Sequentially Markovian Coalescent)** - For inferring population size changes over time.
  - **Gnuplot** - For plotting PSMC results.

- Versions used – at 25th May 2023:

Software	**Version**
FastQC	0.11.9
Trim Galore	0.6.10
BWA	0.7.17
Picard	2.27.9
Qualimap	2.2.2
Samtools	1.17
Bcftools	1.17
PSMC	2020 release
Gnuplot	5.4.5
Seqtk	1.3

- Steps:

1. Download and Prepare Data

Retrieve raw sequencing reads (e.g., paired-end data) from a public database (e.g., ENA, NCBI). Download files in .fastq.gz format.

Unzip Files:

gunzip *.fastq.gz

2. Quality Control (FastQC)

Install FastQC:

sudo apt-get update

sudo apt install fastqc

Run FastQC:

fastqc <read1.fastq> <read2.fastq>

Evaluate sequencing quality using the generated HTML reports.

3. Data Cleaning (Trim Galore)

Install Trim Galore:

Install dependencies:

sudo apt install cutadapt python3-dev

Download and configure Trim Galore:cd /opt

cd /opt

sudo mkdir trim_galore

sudo wget <trim_galore_download_link>

sudo unzip TrimGalore-<version>.zip

sudo ln -s /opt/trim_galore/TrimGalore-<version>/trim_galore /usr/local/bin/trim_galore

Run Trim Galore:

trim_galore --paired --fastqc <read1.fastq.gz> <read2.fastq.gz>

4. Align Reads to the Reference Genome (BWA)

Download Reference Genome:

Retrieve the reference genome from a public repository (e.g., GenBank or Assembly databases).

Decompress the genome file:

gunzip <reference_genome>.fna.gz

Install BWA:

sudo apt install bwa

Index the Reference Genome:

bwa index <reference_genome>.fna

Align Reads:

bwa mem -M -t <number_of_threads> <reference_genome>.fna <read1.fastq.gz> <read2.fastq.gz> > alignment.bam

5. Process Alignment (Picard)

Install Picard:

wget -O picard.jar "https://github.com/broadinstitute/picard/releases/latest/download/picard.jar"

Sort BAM File:

java -jar picard.jar SortSam I=alignment.bam O=sorted_alignment.bam SORT_ORDER=coordinate

Mark Duplicates:

java -jar picard.jar MarkDuplicates I=sorted_alignment.bam O=dedup_alignment.bam M=metrics.txt

6. Quality Control of Alignment (Qualimap)

Install Qualimap:

Ensure Java and R are installed:

java -version

Rscript --version

Download Qualimap:

wget <qualimap_download_link>

unzip <qualimap_version>.zip

Run Qualimap:

./qualimap bamqc -bam dedup_alignment.bam --java-mem-size=12G

7. Generate Consensus Sequence

Install Samtools and Bcftools:

sudo apt install samtools bcftools

Generate VCF File and Consensus Sequence:

samtools mpileup -C 50 -uf <reference_genome>.fna dedup_alignment.bam | \

bcftools call -c | \

vcfutils.pl vcf2fq -d <min_depth> -D <max_depth> | gzip > consensus.fq.gz

Choose depth thresholds (-d and -D) based on your dataset:

Minimum depth: Approximately a third of the average coverage.

Maximum depth: Approximately twice the average coverage.

8. Infer Population History (PSMC)

Install PSMC:

git clone https://github.com/lh3/psmc.git

cd psmc

make

cd utils

make

Convert FASTQ to PSMCFA:

./fq2psmcfa -q <quality_threshold> consensus.fq.gz > consensus.psmcfa

Run PSMC:

./psmc -N <max_iterations> -t <theta_rho_ratio> -r <recombination_rate> -p "<time_intervals_pattern>" -o output.psmc consensus.psmcfa

Adjust parameters according to research objectives:

-N: Maximum number of iterations (e.g., 25).

-t: Initial theta/rho ratio (e.g., 5).

*-p: Pattern of time intervals (e.g., "4+25*2+4+6").*

Plot Results:

./psmc_plot.pl -u <mutation_rate> -g <generation_time> -p output_plot output.psmc

Supplementary Information II

SNIF Parameters

More information on SNIF and its parameters can be found in https://github.com/arredondos/snif

SNIF's parameters include:

- **data_source**: (relative) path for the folder containing the input file(s).
- **source_type**: specifies the type of input file(s)
    - PSMC
    - ms command
- **IICR_type**: specifies the type of IICR
    - Exact - if the input data is a PSMC from genomic data
    - T_sim - if the input data is a ms command to simulate T2s (SNIF does the simulation by calling ms and will produce the corresponding IICR)
    - Seq_sim - if the input data is a ms command to simulate a genetic sequence (SNIF does the simulation by calling ms and will produce the corresponding IICR)
- **ms_reference_size**: refers to the population effective size. Required only in a ms command input file for scaling purposes
- **ms_simulations**: corresponds to the number of T2 samples in a T_sim simulation
- **psmc_mutation_rate**: applied mutation rate of the PSMC or simulation
- **psmc_number_of_sequences**: corresponds to the number of chromosomes to be simulated (only used if the source file is a simulated sequence)
- **psmc_length_of_sequences**: the length of simulated sequences to be simulated (only used if the source file is simulated sequences).
- **infer_scale**: whether to scale the results
    - True - results will be presented in generation times
    - False - results will be presented as T2 times.
- **data_cutoff_bounds**: specifies the time period where inferences will occur
- **distance_computation_interval**: (optional) subset of the **data_cutoff_bounds** time period where the distance between the source IICR and the inferred IICR is calculated.
- **bounds_event_times**: (optional) subset of the **distance_computation_interval** that defines time periods where the demographic events should be inferred. The end goal is to allow the user to ignore ancient or recent events that should not be inferred.
- **data_time_intervals**: number of inferring time intervals (or components). SNIF will infer the dates between these time intervals and the demographic parameters for each time interval (one value of migration rate "M" and one value of relative deme size "s" per component).
- **distance_function**: the distance function (which returns the fitting) can be calculated in two different ways:

    - ApproximatePDF: it uses a weighted method. According to the value of the distance_parameter, more weight will be given to the recent events or the ancient events. This is the preferred approach.
    - Visual: it calculates the best fit considering all time periods have the same weight.

- **distance_parameter**: it regulates the "AproximatePDF" method. 1 is the default value. distance_parameter < 1 gives less weight to the recent past, allowing inferences in the most ancient past. distance_parameter > 1 gives more weight to the recent past, biasing towards the most recent events.
- **distance_max_allowed**: to reach the best fitting, the algorithm tries to minimize a distance function calculated between the input IICR and the inferred IICR. distance_maw_allowed is a threshold, below which the algorithm stops (if the lower bound of the number of **rounds_per_test_bounds** is already reached). If this value is not met, the program runs as many iterations as allowed (upper bound of the number of rounds_per_test_bounds).
- **rounds_per_test_bounds**: an interval of number of rounds of the algorithm. The lower bound is a minimum number of round, always reach even though the distance is below the **distance_max_allowed**. The upper bound is a maximum number of rounds, beyond which the algorithm stops even though the **distance_max_allowed** is not reached.
- **repetitions_per_test**: number of repetitions per given input file and per given set of parameters.
- **number_of_components**: a component designates a time period in which no demographic changes occur. For n components, n-1 events of changes in migration and/or size occurs. This value is data dependent. The user should analyse what is the best minimal number of components that can best explain its data. Usually, it is advisable not to exceed 5 or 6 components (beyond that, the inference becomes more uncertain, and the gain in fit becomes very small).
- **bounds_islands**: range of values for n, the number of islands. The lower bound must be > 1 (i.e there must be at least two islands).
- **bounds_migrations_rates**: range of values for Mi, the scaled migration rates. The lower bound must be > 0.
- **bounds_deme_sizes**: range of values for N the haploid size of the demes. The algorithm infers only one absolute effective deme size value, for the first component. The deme sizes for the other components are given as multiplicative factors of this first deme size value. **bounds-deme_sizes** must correspond to this type of relative value. The size of the demes also depends on the inferred number of islands.
- **bounds_effective_size**: interval of values for the effective size (of one deme) that can be tested.
- If infer_scale = True, the interval must be given in number of generations.

If the **infer_scale** is set for "True", the intervals in the parameters **data_cutoff_bounds, distance_computation_interval,** and **bounds_event_times** must be inputted as the number of generations.