Année académique
2023-2024

Formations de Bacheliers et de Masters organisées par la Fédération Wallonie-Bruxelles

Etude génétique des Chrysomèles du genre Oreina : séquençage rad et tentative de modélisation spatiotemporelle

TFE RÉALISÉ PAR BELLANCA ALESSIA

BLOC 3 DU CURSUS BACHELIER EN BIOTECHNIQUE HEH – DÉPARTEMENT DES SCIENCES ET TECHNOLOGIESHEPH - CONDORCET

Etude génétique des Chrysomèles du genre Oreina : séquençage rad et tentative de modélisation spatiotemporelle

TFE RÉALISÉ PAR BELLANCA ALESSIA

BLOC 3 DU CURSUS BACHELIER EN BIOTECHNIQUE HEH – DÉPARTEMENT DES SCIENCES ET TECHNOLOGIESHEPH - CONDORCET

Remerciement

Je tiens à exprimer ma gratitude envers M. Mardulyn, ainsi que sa doctorante Maeva Sorel et son mémorant Logan Capizi, pour leur suivi et l'aide apporter durant la conception de ce travail. Je souhaite également remercier chaleureusement Mme. Besanger pour la bienveillance dont elle a fait preuve envers moi durant tout mon bachelier ainsi que M. Coornaert pour ses précieux conseils.

Résumé

Ce travail de fin d'étude se concentre sur l'analyse génétique des chrysomèles du genre *Oreina*, en particulier les espèces *Oreina cacaliae*, *Oreina speciosissima*, et *Oreina speciosa*. L'objectif principal était d'étudier les structures de population et les scénarios démographiques de ces espèces. Pour ce faire, la technique de séquençage RADseq et l'outil bioinformatique STACKS ont été employée, permettant de capturer la diversité génétique des populations.

L'analyse en composantes principales (ACP) et les diagrammes de structure de population ont révélé une différenciation génétique prononcée entre les différentes populations.

Malgré des contraintes techniques rencontrées, notamment avec l'utilisation du programme fastsimcoal pour modéliser les scénarios démographiques, il a été possible de formuler plusieurs hypothèses concernant l'évolution des populations. Les résultats obtenus suggèrent que l'isolement géographique dans les environnements montagnards a joué un rôle crucial dans la différenciation locale des populations étudiées. Ces conclusions offrent de nouvelles perspectives sur l'impact de l'isolement géographique et les dynamiques évolutives dans des environnements complexes comme les montagnes.

Summary

This thesis focuses on the genetic analysis of leaf beetles of the Oreina genus, particularly the species Oreina cacaliae, Oreina speciosissima, and Oreina speciosa. The main objective was to study the population structures and demographic scenarios of these species. To achieve this, the RADseq sequencing technique and the bioinformatics tool STACKS were employed, enabling the capture of the genetic diversity of the populations.

Principal Component Analysis (PCA) and population structure diagrams revealed a pronounced genetic differentiation between the different populations. Despite the technical challenges encountered, particularly with the use of the fastsimcoal program for modeling demographic scenarios, it was possible to formulate several hypotheses regarding population evolution. The results suggest that geographic isolation in mountainous environments has played a crucial role in the local differentiation of the studied populations. These conclusions offer new insights into the impact of geographic isolation and the evolutionary dynamics in complex environments such as mountains.

Table des matières

Table des figures

Lexique

ACP (Analyse en Composantes Principales) : Méthode statistique utilisée pour réduire la dimensionnalité d'un ensemble de données tout en conservant la majeure partie de la variation présente dans les données. Elle est souvent utilisée pour visualiser la variation génétique entre les populations ou les individus.

Barrière géographique : Obstacles naturels tels que les montagnes, les rivières ou les vallées profondes qui limitent les mouvements et les échanges génétiques entre populations.

Chrysomèles : Famille de coléoptères comprenant plus de 35 000 espèces.

Dérive génétique : Processus évolutif par lequel les fréquences alléliques dans une population changent au hasard d'une génération à l'autre, particulièrement dans les petites populations.

Diagramme de structure de population : Représentation graphique utilisée pour visualiser la distribution génétique des individus au sein de différentes populations, souvent en montrant la proportion d'appartenance de chaque individu à différents clusters génétiques.

Diversité génétique : Variabilité génétique présente au sein d'une population ou d'une espèce, résultant des variations dans les séquences d'ADN entre les individus.

Gène CO1 (cytochrome c oxydase sous-unité 1) : Gène mitochondrial couramment utilisé dans les études de phylogénie et d'identification des espèces en raison de sa séquence conservée parmi les espèces apparentées.

Homoplasie : Phénomène où des traits ou des séquences d'ADN similaires apparaissent indépendamment chez deux espèces ou individus sans origine commune, souvent dû à des mutations convergentes.

Modélisation spatio-temporelle : Technique utilisée pour comprendre comment les populations se sont dispersées et ont évolué dans l'espace et le temps, en tenant compte des événements historiques comme les glaciations.

Nunatak : Zone non recouverte de glace entourée de glaciers, qui a servi de refuge pour certaines espèces durant les périodes glaciaires.

Oreina : Genre de chrysomèles montagnardes présentes dans les régions alpines d'Europe, caractérisées par leur incapacité à voler et leur spécialisation sur certaines plantes hôtes.

Phylogéographie : Domaine de recherche qui se concentre sur les principes et processus régissant la distribution géographique des lignées génétiques, en particulier au sein et entre des espèces étroitement liées.

Polymorphisme nucléotidique (SNP, Single Nucleotide Polymorphism) : Variation d'une seule paire de bases dans une séquence d'ADN, utilisée comme marqueur génétique pour étudier la diversité génétique au sein des populations.

RADseq (Restriction-site Associated DNA Sequencing) : Technique de séquençage qui permet de cibler et séquencer des régions spécifiques du génome, notamment pour identifier les SNPs dans les études de génétique des populations.

Refuges post-glaciaires : Zones où les espèces ont survécu durant les périodes glaciaires et à partir desquelles elles ont recolonisé les habitats lorsque les conditions climatiques se sont améliorées.

Séquençage RAD : Voir RADseq.

Structuration génétique : Répartition non aléatoire des génotypes dans une population, souvent due à des facteurs comme la sélection naturelle, la dérive génétique, ou les barrières géographiques.

t-SNE (t-distributed Stochastic Neighbor Embedding) : Méthode de réduction de dimensionnalité non linéaire utilisée pour visualiser des données complexes en deux ou trois dimensions, en préservant les relations locales entre les points.

UMAP (Uniform Manifold Approximation and Projection) : Algorithme de réduction de dimensionnalité qui permet de visualiser la structure globale des données tout en conservant les relations locales, souvent utilisé en complément de t-SNE.

1. Introduction

1.1. Chrysomèles

Fig. 1 : Chrysomèle oreina cacaliae

Les Chrysomèles sont présentes dans le monde entier et comprennent plus de 35 000 espèces différentes, ce qui en fait l'une des familles les plus diverses parmi les coléoptères (Jolivet, 2002). Les chrysomèles sont souvent utilisées en tant qu'organismes modèles dans les études sur la coévolution plante-insecte, la chimie des défenses des insectes, et la dynamique des populations. C'est ce dernier aspect qui sera le sujet de cette étude (Pasteels, 1989).

Au sein de cette famille, le genre *Oreina* se distingue par plusieurs caractéristiques. Les chrysomèles du genre *Oreina* sont principalement montagnardes et se retrouvent dans les régions alpines et subalpines d'Europe. Ce sont des coléoptères non volants, une caractéristique qui limite leurs capacités de dispersion (Triponez, 2011) (Kalberer, 2005). Ce genre comprend 28 espèces décrites principalement sur la base de caractères liés aux organes génitaux mâles. Toutes ces espèces sont phytophages, c'est-à-dire qu'elles se nourrissent exclusivement de plantes et plus précisément des familles Apiacées ou Astéracée. Chaque espèce a un petit nombre de plantes hôtes spécifiques, ce qui fait que la répartition géographique des espèces de chrysomèles est fortement liée à la répartition géographique de leurs plantes hôtes (Jolivet P. &., 1986).

L'étude génétique des espèces du genre *Oreina* présente un intérêt particulier pour mieux comprendre les mécanismes évolutifs et adaptatifs qui ont permis à ces coléoptères de se diversifier et de prospérer dans des environnements montagneux difficiles.

1.2. Diversité génétique

La diversité génétique est le résultat de variations au niveau de l'ADN des individus d'une espèce. La caractérisation de la diversité génétique d'une espèce vise à déterminer la répartition génétique de l'espèce, en analysant les différences et les similitudes entre les individus sur le plan génétique (Frankham, 2002)

Pour caractériser la diversité génétique d'une espèce, l'utilisation des SNPs (Single Nucléotide Polymorphisms) comme marqueur moléculaire est préconisée. Les « single-nucleotide polymorphisms » ou polymorphismes nucléotidiques sont des nucléotides au sein du génome pour lesquels il existe différents allèles. La propriété intéressante de ces SNP est leur faible taux de mutation, ce qui permet d'éviter l'homoplasie, un phénomène qui peut entraîner un biais dans l'analyse phylogénétique. En effet, on peut retrouver un site nucléotidique identique chez deux individus sans qu'ils aient une origine commune mais parce qu'ils ont atteint cet état indépendamment, par mutation (Hedrick, 2005).

Il est possible de mettre en évidence plusieurs dizaines de milliers de SNPs pour une espèce sans nécessité d'information sur le génome à l'aide de la méthode RADseq ou « Restriction site-Associated DNA Sequencing » (Baird, 2008). Dans cette méthode, le génome de chaque individu est digéré par des enzymes de restriction. Les fragments liés à un code-barres servant à l'identification sont ensuite sélectionnés en fonction de leur taille par électrophorèse afin de garder des fragments dont la taille est comprise entre 100 et 200 paires de bases. Les fragments peuvent alors être séquencés (Miller, 2007).

Cette technique de séquençage est particulièrement utile pour des études de génétique des populations et de phylogéographie, où l'on cherche à comprendre comment les espèces se sont dispersées et ont évolué au fil du temps (Davey, 2010). En combinant ces données génétiques avec une modélisation spatiotemporelle, il devient possible de reconstituer l'histoire évolutive des populations de *Oreina*.

1.3. Modélisation spatiotemporelle

La phylogéographie concerne l'observation et l'analyse de la distribution spatiale des génotypes et l'inférence de scénarios historiques sur ces espèces. En fait, la phylogéographie ajoute à l'analyse phylogénétique une composante géographique (Kidd, 2006). La phylogéographie peut aussi être décrite comme un domaine de recherche qui se concentre sur les principes et les processus qui régissent la distribution géographique des lignées apparentées, en particulier au sein et entre des espèces étroitement liées (Avise, 2009).

Le dernier maximum glaciaire (LGM) se situe entre 23 000 et 18 000 ans et a eu un impact majeur sur la répartition actuelle des espèces, poussant probablement des espèces à se réfugier aux altitudes les plus basses où les conditions climatiques étaient moins extrêmes. Les Alpes étaient recouvertes de glace. Pourtant, la diversité génétique dans les Alpes est bien plus élevée que dans d'autres régions, suggérant ainsi que cette région était un refuge pour certaines espèces (Pinceel, 2005).

Les deux grandes hypothèses de refuge sont : l'hypothèse des zones périphériques, qui correspondrait à une recolonisation des Alpes à partir de l'extérieur, avec des espèces réfugiées dans les péninsules méditerranéennes de la péninsule ibérique, de l'Italie et des Balkans, et la recolonisation des espèces qui s'est faite via le déplacement de celles-ci vers le nord, au fur et à mesure que le climat se réchauffait.

L'autre hypothèse est celle des refuges intérieurs appelés "nunataks", selon laquelle la recolonisation des Alpes après le maximum glaciaire se serait déroulée à partir d'une expansion de ce refuge pour éventuellement se reconnecter. Holderegger décrit un troisième type de refuge : les plaines. Ce refuge postule qu'il y aurait des refuges sans glace au sein même des glaciers couvrant les Alpes. Cette dernière hypothèse se différencie de l'hypothèse des Nunataks par le fait que ces derniers se situent au-dessus des glaciers, et non dans les glaciers. Mais dans le cadre de cette étude, l'hypothèse des plaines et celle des Nunataks seront considérées comme une seule et même hypothèse : la recolonisation des Alpes à partir de l'intérieur. À elles trois, ces hypothèses pourraient résumer l'ensemble des scénarios existants (Holderegger, 2009).

2. But du travail

L'objectif principal de ce travail est d'explorer la diversité génétique des chrysomèles du genre *Oreina*.

Plus spécifiquement, ce travail propose :

- D' identifier les espèces présente dans le genre grâce au séquençage RAD-seq et au pipeline Stacks.
- D'analyser la structuration des espèces à travers des méthodes statistiques telles que l'analyse en composantes principales (ACP) et la structure de populations.
- De réitérer l'analyse Stacks sur 3 espèces (*Oreina cacaliae*, *Oreina speciosissima*, et *Oreina speciosa)*
- D'analyser la structuration des populations au seins de ses espèces.
- D'examiner l'impact des barrières géographiques et de l'histoire glaciaire sur la répartition actuelle des populations, en explorant comment ces facteurs ont pu limiter la dispersion et favoriser la différenciation génétique.

En réalisant ces objectifs, ce travail cherche à enrichir les connaissances sur la biogéographie et l'évolution des espèces alpines, tout en fournissant des bases solides pour des études futures dans le domaine de la génétique des populations.

3. Matériel et méthode

3.1. Échantillonnage

391 Chrysomèles du genre *Oreina* ont été récoltés sur les plantes *Peucedanum ostruthium*, *Adenostyles sp.*, *Petasites sp*. et *Cirsium sp.* dans les Alpes françaises, suisses et autrichiennes durant les mois de juillet et aout 2021 ainsi qu'en aout 2022 par le professeur P. Mardulyn.
Pour chaque individu des informations concernant la localité ainsi que la plante sur lequel l'insectes a été récolté ont été consigné dans un tableau. (voir annexe 1).

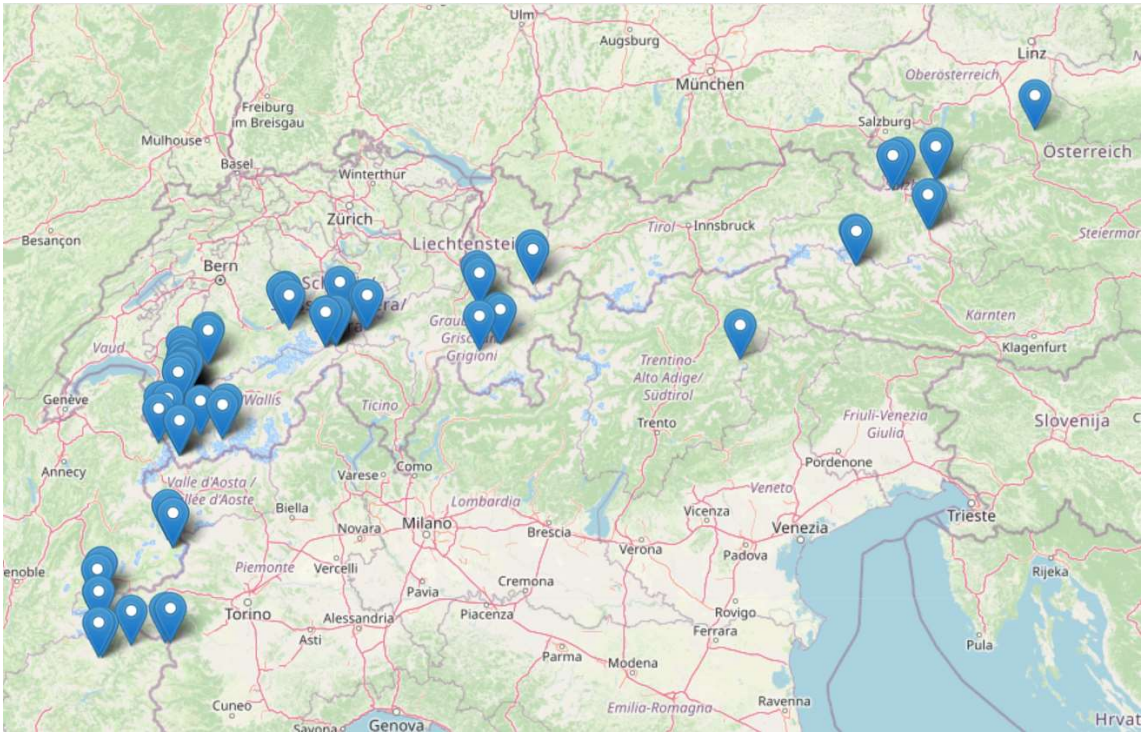


Fig. 2 : Mise en évidence des localités de récolte (RÉALISÉ À PARTIR DU SCRIPT PYTHON EN ANNEXE2)

Sur ces 391 insectes, 368 ont été séquencé. Les échantillons ont été ramenés au laboratoire de l'Université Libre de Bruxelles dans le service d'évolution et écologie ou a eu lieu l'extraction d'ADN. L'ADN a ensuite été envoyé à une société spécialisée (floragenex) qui a réalisé le séquençage RAD-Seq sur un séquenceur illumina. Le protocole suivi par Floragenex est décrit par Baird et al. (2008).

3.2. Séquençage RAD

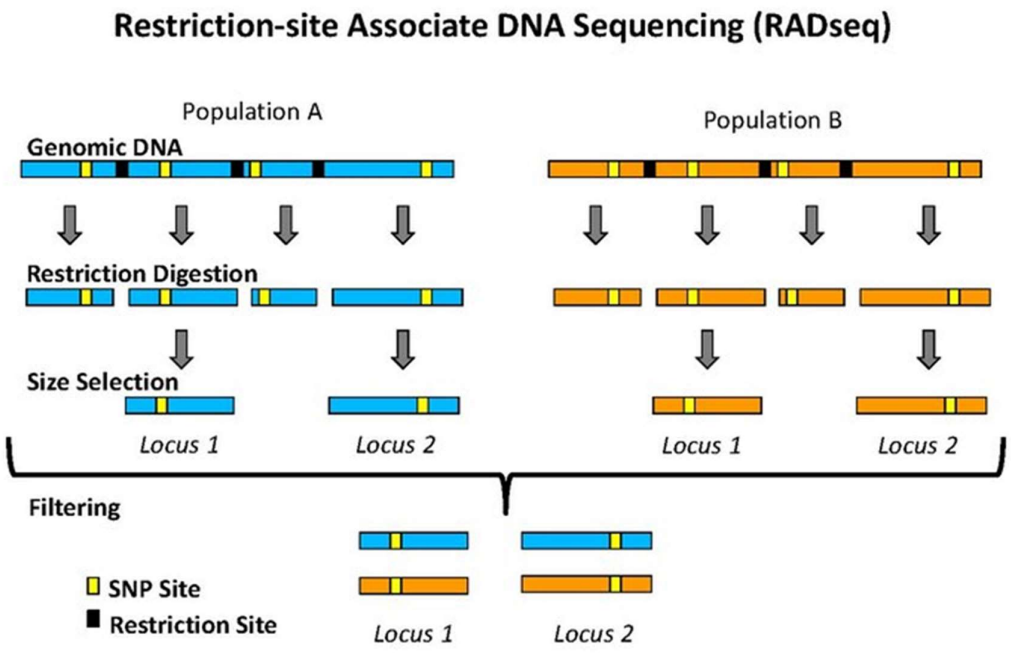


Fig. 3 : Illustration Expliquant le séquençage RAD-SEQ

Le séquençage RAD (Restriction-site Associated DNA sequencing) est une technique de génomique qui permet de séquencer des régions spécifiques du génome, appelées loci, chez plusieurs individus pour identifier des variations génétiques, comme des SNPs (Single Nucleotide Polymorphisms) (Davey, 2010).

L'ADN génomique des individus de deux populations (Population A et Population B) est coupé en fragments à des endroits spécifiques appelés sites de restriction (indiqués en noir sur l'image). Ces sites sont reconnus par des enzymes de restriction qui coupent l'ADN à ces emplacements précis (Baird, 2008). Les fragments d'ADN sont de différentes tailles. Une sélection de taille est effectuée pour isoler des fragments d'une longueur spécifique, correspondant aux régions d'intérêt du génome. Ces fragments sont appelés loci. Les loci sélectionnés contiennent des SNPs (indiqués en jaune sur l'image), qui sont les points de variation entre les différents individus ou populations (Peterson, 2012).

Les fragments d'ADN de taille appropriée (contenant les SNPs) sont filtrés et préparés pour le séquençage. Les loci sont ensuite séquencés pour identifier les SNPs et analyser les différences génétiques entre les populations (Andrews 2016).

3.3. Données initiales

Les données ont été reçues sous forme de 4 fichier compressé.

Après avoir été décompressé et démultiplexé, les reads au format fastq sont obtenu.

```
@336_1_2101_10267_1016/1
TGCAGGAATTTAAATAGACCAGCGTCCTACTATCAGTGTACAAGAAAAACTCATTTGCCAACAAATAAGTTCGCCAACGACGAATACACCATAAGGCTGCTAGTACTTCCTTCTCGTTAGCG
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:
@336_1_2101_10339_1016/1
TGCAGGGATTTATCAGGGTATGCCGCTAACTAATTATTTTTTAATTTCAAATCCTACCAACCATCGAACATCATCGAATTCGGGCCACAGCCAACTTGATGTTTTTCAATAAAATGCAATTC
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFF
```

Fig. 4 : Read au format fastQ

Structure des reads :

ligne 1 : identifiant de la séquence de lecture commençant par @

ligne 2 : séquence nucléotidique

ligne 3 : séparateur "+"

ligne 4 : score de qualité au format phred, chaque caractère ASCII correspond à un score numérique indiquant la fiabilité de chaque base

La longueur des séquences est uniforme (122 bases) et les scores de qualité sont élevés (36 en moyenne ce qui correspond sur une échelle phred à une probabilité d'erreur d'environ 1 sur 4000).

L'enzyme utilisé pour le séquençage Rad est ici **sbfI.**

L'enzyme SbfI reconnaît et coupe une séquence d'ADN spécifique composée de 8 paires de bases, généralement **5'-CCTGCA|GG-3'**. SbfI coupe l'ADN de manière asymétrique, laissant des extrémités cohésives (sticky ends) après la coupure. Cette propriété est utile pour le séquençage RAD-seq car elle permet l'ajout de codes-barres pour l'identification des fragments d'ADN.
Le choix de SbfI est influencé par sa séquence de reconnaissance longue (8 bases), qui se produit moins fréquemment dans le génome comparé à des enzymes reconnaissant des séquences plus courtes. Cela permet de générer un nombre gérable de fragments d'ADN, adaptés à la capacité de séquençage du RAD-seq. Ce type de coupure est particulièrement utile dans les études où l'on souhaite couvrir une partie représentative du génome sans obtenir un nombre excessif de fragments à séquencer. De plus le génome des chrysomèles *Oreina* présente des séquences riches en GC, pour lesquelles SbfI est particulièrement adapté.

Le nombre de séquence pour chaque individu a été calculé grâce à un script présent en annexe 3. L'axe X représente chaque fichier .fq.gz dans mon dossier. Le nombre de fichiers est trop grand pour pouvoir afficher le nom de chaque fichier de façon lisible sur le graphe. L'axe Y indique le nombre de séquences trouvées dans chaque fichier. Les valeurs de l'axe Y vont de 0 à environ 4 millions de séquences. On peut donc voir sur ce graphe que le nombre de séquences n'est pas uniforme. Il est essentiel de normaliser les données pour éviter que les individus avec plus de données n'influencent davantage l'ACP. Cette standardisation est faite sur R au moment de calculer l'ACP.

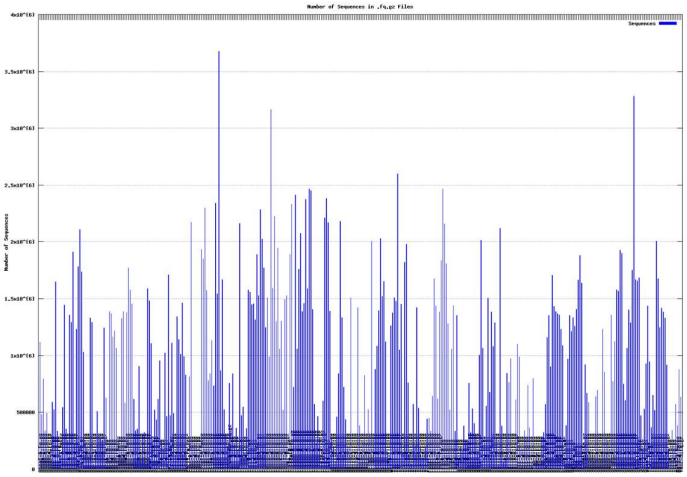


Fig. 5 : Visualisation du nombre de séquence par individu

3.4. STACKS

STACKS est un pipeline de logiciels d'analyse de données génomiques qui a été conçue pour analyser des données provenant du séquençage ADN associé à un site de restriction (RAD). Les points forts de STACKS sont sa flexibilité et son manuel clair et détaillé. (Catchen J., 2013)

STACKS est accessible uniquement à partir d'un environnement de type Unix. Linux est une famille de systèmes d'exploitation de type Unix. Ce système d'exploitation est indispensable pour l'analyse de données biologique car de nombreuses applications bio-informatiques sont développées pour Linux et inexistante sur d'autre système d'exploitation. C'est le cas de tous les logiciels utilisés dans le cadre de cette étude à l'exception de R. Parmi les nombreuses distributions Linux disponibles, la distribution Gentoo a été utilisée pour effectuer les analyses de ce travail.

Pour réaliser ces analyses, l'accès à Céci a été obtenu. Le Céci est un ensemble de clusters dont l'objectif est de fournir aux chercheurs des équipements informatiques performants. Les clusters sont installés et gérés localement sur les différents sites des universités participantes (ULiège, UNamur, UMons, UCL et ULB) et ils sont accessibles à tous les chercheurs des universités membres.

Tous fonctionnent sous Linux et utilisent Slurm comme gestionnaire de tâches.

En informatique un cluster est un groupe d'ordinateurs interconnectés qui travaillent ensemble pour résoudre des problèmes de calcul complexes ou effectuer des tâches de calcul haute performance (HPC). Un cluster se compose généralement de plusieurs ordinateurs (appelés nœuds), qui sont connectés par un réseau à haut débit et travaillent ensemble pour effectuer des tâches de calcul en parallèle. Cela permet à un cluster d'atteindre une puissance de traitement beaucoup plus élevée et des temps d'exécution plus rapides qu'un ordinateur pourrait atteindre seul. Les clusters sont couramment utilisés pour les applications gourmandes en données, telles que les données de séquençage.

L'accès au cluster a nécessité la création d'une clé SSH et le partage de la clé publique. Ensuite, un fichier ~/.ssh/config a été configuré à l'aide de l'assistant proposé sur le site."

```
ssh-keygen -f ~/.ssh/id_rsa.ceci -t
rsa -b 2048
Chmod 600 ~/.ssh/config
Ssh lemaitre3
```

Le cluster lemaitre3 a été initialement choisi pour effectuer le travail, car le programme STACKS y était déjà installé. Cependant, la limite de temps autorisée pour un job sur lemaitre3 étant de 2 jours, ce délai s'est révélé insuffisant pour compléter l'analyse. Par conséquent, STACKS a été installé sur le cluster Dragon2, qui permet des jobs d'une durée maximale de 10 jours. L'analyse STACKS a duré 7 jours.

Commande utiliser pour uploader des données sur le cluster :

```
Scp /data/stage/base/ abellanc@dragon2:
```

Les scripts envoyés sur le cluster doivent contenir des en-tête slurm. Slurm est un système de gestion de tâches utilisé pour la planification et l'exécution de travaux sur des clusters. Voir annexe 4.

Trois fichiers sont nécessaires pour réaliser une analyse avec stacks.

- Les données RAD-Seq (c'est-à-dire les séquences générées par un séquenceur Illumina, qui peuvent être compressées)
- Les codes-barres : ce fichier est formé de deux colonnes séparées par une tabulation, la première contient le code-barre et la deuxième le nom de l'échantillon qui lui est associé.
- La carte de populations : ce fichier est formé de deux colonnes séparées par une tabulation, la première contient le nom de l'échantillon et la deuxième le nom de la population qui lui est associée.

STACKS contient 7 programmes :

- Process_radtags
- Ustacks
- Cstacks
- Sstacks
- Tsv2bam
- Gstacks
- Populations

3.4.1. Process_radtags

Ce programme vérifie que le code-barre et le site de restriction sont intacts et peut corriger des erreurs mineures dans ceux-ci. Il démultiplexe ensuite les données et vérifie le score de qualité moyen dans une fenêtre de lecture prédéfinies. Si le score est inférieur à 90 % de probabilité d'être correct la lecture est rejetée ce qui permet d'éliminer certaines erreurs de séquençage.

Les données sont séparées en 4 fichiers, un par plaque, il en est de même pour les fichiers de barcodes. Il y a un barcode en index et un inline, le barcode en index sers à identifier la plaque et ne doit donc pas être utilisé dans le démultiplexage, seul le barcode inline qui permet d'identifier l'échantillon est intéressant. Autre particularité le fichier de barcode contient le barcode et le site de restriction tout deux l'un à la suite de l'autre, il est donc nécessaire d'enlever les 6 derniers nucléotides de chaque ligne afin d'obtenir uniquement le barcode. En fournissant le nom de l'enzyme (ici sbfI) process_radtags retrouve lui-même le site de restriction.

Le fichier de barcodes a dû être converti d'un format Excel vers un format .txt séparé par des tabulations

Exemple d'utilisation :

```
process_radtags -p /data/stage/base/U0_C659_1.fasta.gz -o
/data/stage/stacks/samples -b /data/stage/barcodes -e sgrAI -c -q
```

```
-p : chemin vers le fichier contenant les données
-o : chemin de stockage des fichiers traités
-b : chemin du fichier contenant les code-barres.
-e : nom de l'enzyme de restriction utilisée
-c : nettoyer les données, supprimer toute lecture avec une base
          inconnues
-q : ignorer les lectures avec des scores de qualité faible
```

3.4.2. Ustacks

A partir des données de séquençage RAD, ce programme identifie et regroupe dans un premier temps les lectures parfaitement identiques. Dans un deuxième temps il va fusionner les piles précédemment formées en fonction de la distance autorisée entre les piles. Il identifie les polymorphismes de nucléotides simples (SNP) et génotype les échantillons à chaque locus.

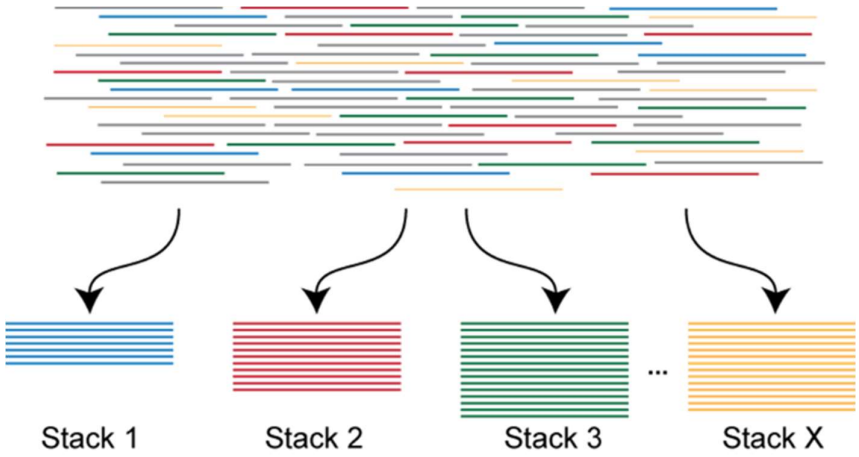


Fig. 6 : Illustration du fonctionnement de Ustacks

Exemple d'utilisation :
Le programme Ustacks ne permettant pas de lancer l'analyse pour plusieurs échantillons simultanément, un script shell a été écrit pour automatiser ce processus.

```
# ! /bin/sh
j = 1
for i in ../samples/*.fq.gz
     do
          ustacks -f ''$i'' -o ./ustacks -i '$'j'' -m 3 - M 5
          ((j++)
     Done
```

-f : chemin vers l'échantillon
-o : chemin de stockage des fichiers traités
-i : index unique par échantillon traités
- m : profondeur minimale d'une pile, sachant que seules les lectures parfaitement identiques sont empilées.
- M : distance autorisée entre les piles. Si le nombre de nucléotides qui différent entre deux empilements est inférieur à M, les deux piles seront fusionnées.

3.4.3. Cstacks

Ce programme établi un catalogue de loci à partir de l'ensemble des échantillons traités par le programme Ustacks. Il identifie ensuite les loci qui sont partagés entre plusieurs échantillons.

Example d'utilisation :

```
cstacks -P . -M ../../base/popmap -n 1
```

```
-P : chemin vers les échantillons traités par ustacks
-M : chemin vers un fichier contenant la carte de population
-o : chemin de stockage des fichiers traités
-s : chemin vers un échantillon traité par ustacks
-n = Distance autorisée entre les loci du catalogue. Si le nombre de
nucléotides qui différent entre deux loci du catalogue est inférieur
à n, les deux loci seront fusionnée.
```

3.4.4. Sstacks

Ce programme fait correspondre les loci de chaque échantillon à ceux du catalogue généré par Cstacks. Il associe le génotype de chaque individu à un locus.

Example d'utilisation :

```
stacks -P . -M ../../base/popmap
```

```
-P : chemin vers les échantillons traités par ustacks et les catalogues
-M : chemin vers un fichier contenant la carte de population
-o : chemin de stockage des fichiers traités
-s : chemin vers un échantillon traité par ustacks
```

3.4.5. Tsv2bam

Ce programme transpose les données afin qu'elles soient orientées par locus, plutôt que par échantillon.

Example d'utilisation :

```
tsv2bam -P . -M ../../base/popmap
```

```
-P : chemin vers les données
-M : chemin vers un fichier contenant la carte de population
```

3.4.6. Gstacks

Ce programme examine tous les individus d'une population pour un locus à la fois. Il identifie les SNP au sein de la population pour chaque locus, puis génotype chaque individu à chaque SNP identifié.

Example d'utilisation :

```
gstacks -P . -M ../../base/popmap
```

```
-P : chemin vers les données
-M : chemin vers un fichier contenant la carte de population
```

3.4.7. Populations

Ce programme calcule un certain nombre de statistiques sur la génétique des populations, il permet d'appliquer différents filtres sur les résultats et d'exporter les données en une variété de formats.

Example d'utilisation :

```
populations -P . --popmap ../../base/popmap –vcf
```

```
-P : chemin vers les données
--popmap : chemin vers un fichier contenant la carte de population
--vcf : les fichiers de sorties sont en formats vcf
-r : seuls les locus présents dans au moins x% des échantillons sont
retenus
```

3.5. Délimitation des espèces et populations

R est un langage de programmation. Il peut être utilisé via l'interface graphique RStudio ou via l'interface de ligne de commande. Dans l'ensemble, R est un outil puissant et polyvalent pour l'analyse statistique et la visualisation de données, et il est largement utilisé dans le domaine de la bio-informatique. R a été utilisé afin de réaliser divers graphe. (voir script en annexe 5.)

FastStructure est un logiciel permettant de déduire la structure de lapopulation à partir de données génétiques, en particulier à partir de données de polymorphisme mononucléotidique (SNP). Il utilise un algorithme bayésien pour modéliser la structure de la population. Il faut fournir à Faststructure un nombre de population (K), il affectera ensuite les individus à un cluster.

Pour pouvoir utiliser faststructure plusieurs modifications doivent être apporté au fichier de données.

- Le fichier doit être au format structure (ce format peut être demander à la dernière étape de stacks)
- Les lignes d'en-tête comportant la version de stacks doivent être supprimé
- 4 colonnes de # doivent être ajouté en début de fichier
- Les 0 symbolisant les données manquantes doivent être remplacée par des -9. Cette manipulation a été effectuée avec le programme Sed.
  Sed 's/ « 0 »/ « -9 »/g' population.str > population2.str

Principaux paramètres :

```
-K <int> : nombre de populations
--input=<file> : chemin vers le fichier .Q
--output=<file> : chemin vers le dossier ou écrire le résultat
--format={bed,str} format du fichier d'entré (bed par défaut)
```

La **méthode d'Evanno** a été utilisée pour estimer le nombre optimal de populations (K) à partir des données, en se basant sur les valeurs de vraisemblance obtenues avec faststructure. (G. Evanno, 2005).

L'(K) correspond à la différence entre le Ln P(D) (le logarithme de la probabilité postérieure) calculé pour la valeur K et celui calculé pour la valeur K-1 => L'(K) = L(K) - L (K-1)

| L'' (K) | = | L'(K+1) - L'(K) |

Pour calculer le delta K, il faut lancer le run 20 fois pour chaque valeur de K ce qui permet de calculer une moyenne (m) et une déviation standard (s) pour L(K). (voir détails des calculs en annexe 6)

delta K = m (| L''(K) |) / s (L(K) => delta K = m(| L(K+1) - 2L(K) + L(K-1) |) / s (L(K))

Pophelper est un outil logiciel conçu pour analyser les données génétiques des populations et faciliter l'interprétation des résultats de diverses analyses génétiques des populations. Il fournit une interface graphique conviviale pour la visualisation des données.
Pophelper a été utilisé à partir de Rstudio pour obtenir une représentation graphique des résultats fournis par Faststructure. Le script R permettant d'utiliser pophelper se trouve en annexe 7.

Les labels sont ajoutés en joignant un fichier contenant la liste des populations pour chaque échantillon, dans le bon ordre. (sélectionner la colonne population du fichier au format structure avec awk et l'écrire dans un nouveau fichier)

Option cochée pour trier les populations.

3.6. Modélisation spatio-temporelle

EasySFS (Site Frequency Spectrum) est un logiciel utilisé afin de générer un SFS à partir d'un fichier vcf. Un SFS est une statistique qui résume la distribution des fréquences alléliques pour un ensemble de variants, en une matrice par population. Ce programme est nécessaire pour que les données soient dans un format accepté par Fastsimcoal.

Les données manquantes sont une caractéristique importante des ensembles de données de type RAD-seq et la suppression des sites manquants supprimerait une grande partie des données. Une autre solution serait d'imputer les valeurs manquantes mais l'imputation ne sera pas fiable s'il y a beaucoup de valeurs manquantes. La méthode de projection vers le bas est une sorte de compromis entre ces deux extrêmes. Une matrice de données complète est construite sur base d'une projection vers une plus petite taille d'échantillon et une moyenne de tous les rééchantillonnages possibles.

Exemple d'utilisation :

```
/opt/easySFS-master/easySFS.py
 -i /data/stage/stacks/s321/populations/populations.snps.vcf
-p /data/stage/base/popmap –preview

/opt/easySFS-master/easySFS.py
 -i /data/stage/stacks/s321/populations/populations.snps.vcf
-p /data/stage/base/popmap –proj=32,10,18
```

Le nombre de projection à choisir est celui qui maximise le nombre de sites pour chaque population.

FastSimCoal est un logiciel utilisé pour simuler les données génétiques des populations selon divers scénarios évolutifs, y compris des histoires démographiques complexes, des expansions de population et des événements de migration. Il peut également simuler divers types de marqueurs génétiques, notamment les polymorphismes mononucléotidiques (SNP), les microsatellites et l'ADN mitochondrial (ADNmt). L'une des principales caractéristiques de FastSimCoal est sa capacité à générer des ensembles de données adaptées à des questions de recherche spécifiques. Par exemple, les utilisateurs peuvent spécifier le nombre de populations, les taux de migration, la taille effective des populations et d'autres paramètres importants pour l'étude de la structure génétique des populations. De plus, il est capable d'estimer des paramètres de l'évolution des populations à partir de données génétiques. FastSimCoal génère une sortie dans une variété de formats, y compris les fichiers VCF, MS et Arlequin, qui peuvent être utilisés avec divers outils logiciels de génétique des populations pour une analyse plus approfondie (L. Excoffier, 2013) (C. Kastally, 2021).

4. Résultats

4.1. Délimitation des espèces du genre *Oreina*

L'analyse en composante principale est une méthode statistique utilisée pour réduire la dimensionnalité d'un ensemble de données tout en conservant la majeure partie de la variation des données. (analyse en composante principale, s.d.) L'ACP fonctionne en transformant l'ensemble de données d'origine en un nouvel ensemble de variables, appelées composantes principales, qui ne sont pas corrélées et classées selon leur importance pour expliquer la variabilité des données. La première composante principale représente la plus grande proportion de la variance des données, la deuxième composante représente la deuxième plus grande proportion, et ainsi de suite. L'ACP peut être utilisée pour la visualisation des données, car elle permet de tracer des données de grande dimension dans un espace de dimension inférieure. En génétique, l'ACP peut être utilisée pour visualiser la variation génétique entre les populations ou les individus.

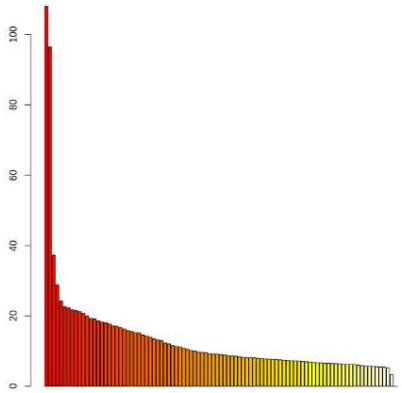


Fig. 7 : Graphe des proportions des composantes

Le nombre de composantes principales est choisi en fonction du graphe des proportions des composantes avec la « méthode du coude » qui consiste à repérer l'endroit à partir duquel le pourcentage d'inertie diminue beaucoup plus lentement lorsque l'on parcourt le diagramme de gauche à droite. On considère ici 2 composantes principales.

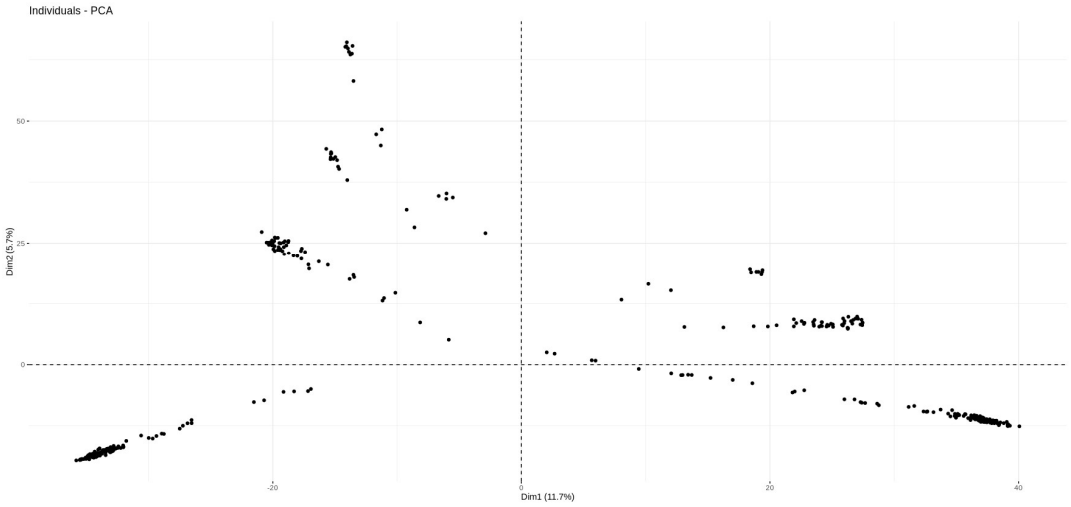


Fig. 8 : ACP du genre Oreina

Chaque point sur le graphique représente un individu. Ces individus sont projetés dans l'espace des deux composantes principales.
La position d'un point dans ce plan reflète les similarités ou différences par rapport aux autres points : des points proches les uns des autres sont similaires selon les composantes principales.
On peut observer des clusters (groupes de points) sur le graphique. Ces groupes peuvent potentiellement représenter différentes espèces d'insectes ou différents groupes au sein d'une espèce.

Des études antérieures ont montré que des gènes mitochondriaux uniques peuvent être utilisés pour délimiter les espèces (E. Pante, 2015). Pour pouvoir attribuer une espèce à chaque cluster le gène CO1 de quelques individus a été séquencé et comparé au gène CO1 connu dans les banques de données biologiques.
Le gène CO1 (cytochrome c oxydase sous-unité 1), est un gène mitochondrial dont la séquence d'ADN est conservée dans une certaine
mesure entre les espèces apparentées. C'est pour cette raison qu'il est largement utilisé dans les études de phylogénie et d'identification des espèces.
Le séquençage du gène CO1 a donc permis d'associer certains individus à une espèces.

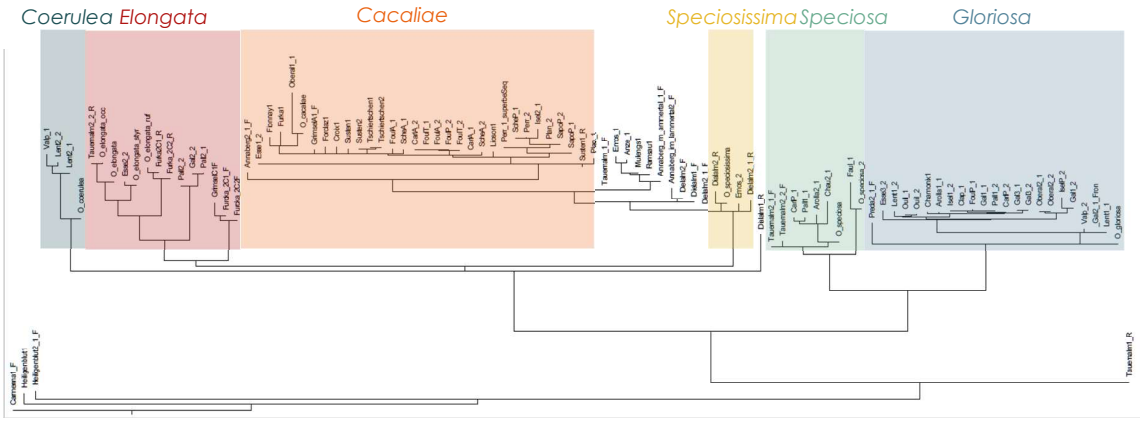


Fig. 9 : Arbre phylogénétique construit sur base du séquençage du gène CO1

Les individus se rapprochant de ces « repères » sur l'ACP sont associés à la même espèce.

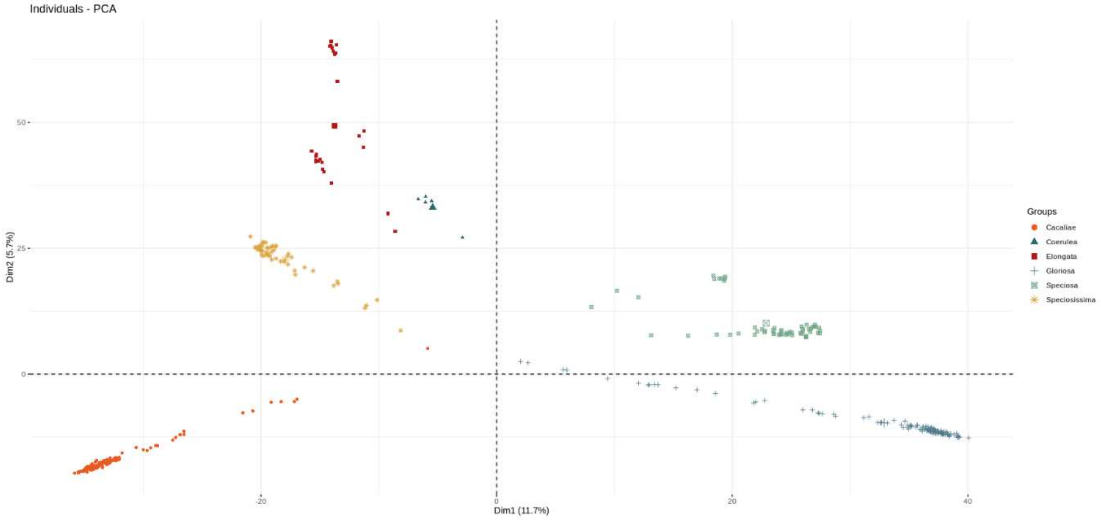


Fig. 10 : ACP du genre Oreina avec mise en évidence des espèces

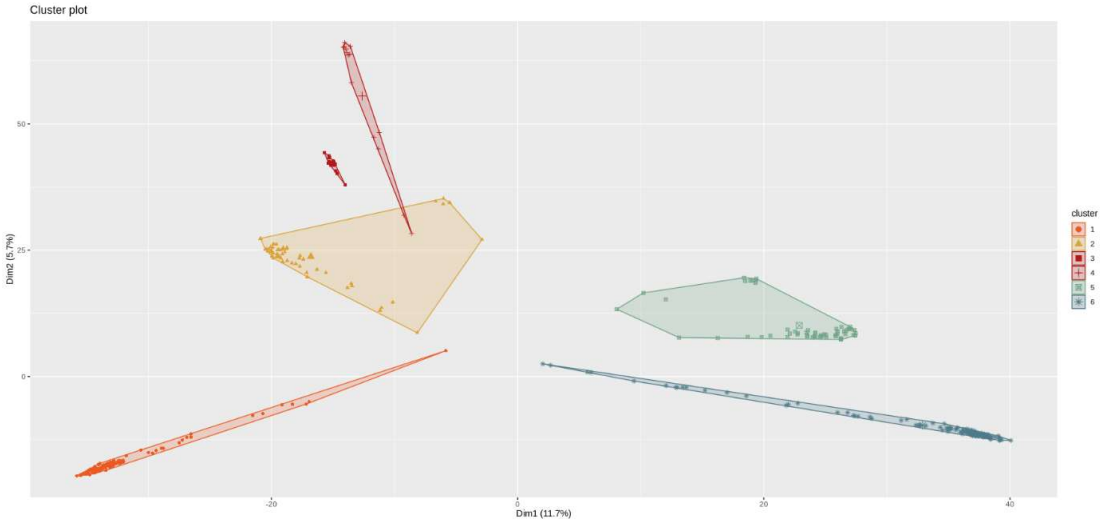


Fig. 11 : ACP avec mise en évidence des clusters

Chaque point représente un individu, et la couleur/symbole du point pourrait correspondre à une espèce. Les clusters sont bien différenciés dans l'espace des composantes principales, suggérant une variation génétique significative.

Les formations en ligne suggèrent une distribution non uniforme des individus selon les deux premières composantes principales. Les individus d'une même espèce pourraient être très homogènes pour certaines caractéristiques, mais varier systématiquement pour d'autres, créant ainsi ces formations linéaires. Les chrysomèles d'une même espèces peuvent avoir un aspect physique (la couleur notamment) différent.

Notons que l'espèces *O.coerulea* est représenté par peu d'individus, ce qui rend sont identifications en tant que espèces à part entière difficile. Elle est dans cette analyse de cluster associé à l'espèces *O.speciosissima.* L'especes *O.elongata* est elle divisé en deux clusters, il y a certainement une grande diversité génétique au sein de cette espèce.

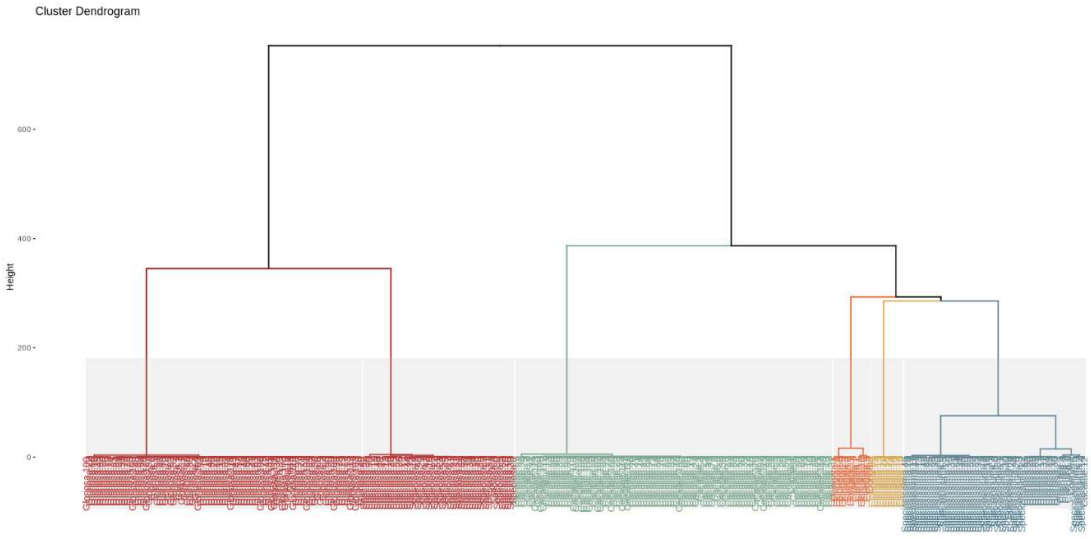


Fig. 12 : Dendrogramme avec mise en évidence des espèces du genre Oreina

Le dendrogramme est un arbre hiérarchique qui montre comment les individus sont regroupés en fonction de leurs similarités.

Les branches qui se rejoignent représentent des groupes d'individus similaires. Plus les branches se rejoignent bas dans le dendrogramme, plus les individus sont similaires.

Les branches principales qui se rejoignent à des niveaux très élevés peuvent représenter des regroupements majeurs, potentiellement correspondant à des espèces distinctes.

Inversement, des branches qui se rejoignent à des niveaux bas sont probablement des individus appartenant à la même espèce ou à des groupes très similaires au sein d'une espèce.

Cela complète bien l'analyse en ACP, offrant une perspective différente mais complémentaire sur la façon dont les individus se regroupent.

L'ACP est un outil statistique qui fonctionne avec tout type de données et n'est pas spécifique à l'analyse de données biologique. C'est la raison pour laquelle Le logiciel Faststructure et la librairie Pophelper sur R ont ensuite été utilisés. Contrairement à l'ACP, Faststructure utilise un modèle d'évolution des populations pour analyser les données, ce qui en fait un outil plus spécifique. Pophelper permet de visualiser les proportions de mélanges au sein des différentes populations.
Une analyse à l'aide de faststructure et de pophelper permet d'approfondir les résultats.

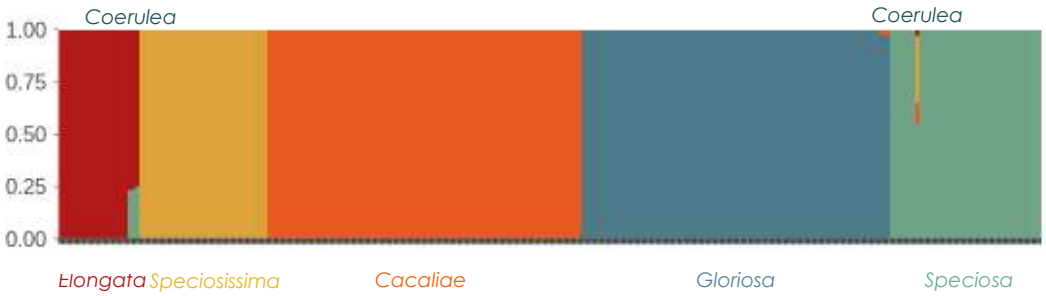


Fig. 13 : Diagramme de structure de population du genre Oreina

Ce graphe est un diagramme de structure de population utilisé pour visualiser la distribution génétique de différentes espèces (ou populations).

- L'Axe X représente les individus classés par espèces. Chaque barre verticale représente un individu, et les différentes couleurs dans chaque barre représentent les proportions des clusters génétiques attribués à cet individu.
- L'Axe Y représente la proportion d'appartenance de chaque individu à différents clusters génétiques. L'échelle varie de 0 à 1, ce qui signifie que chaque individu peut avoir une composition génétique mixte (appartenant à plusieurs clusters) ou être entièrement assigné à un cluster.

Le graphe obtenu avec pophelper montre un cluster par espèces. Un seul cluster signifie une seule variété génétique, ces espèces ont donc été correctement identifié. L'espèce *O.coerulea* présente un mélange important de cluster, ceci est certainement dû au petit nombre d'individus identifié à cette espèces (5 sur 368). Cette différence de nombres d'individus fait que faststructure n'associe pas ces individus à une espèce mais à un mélange des autres espèces présentes dans l'échantillonnage. *O.coerulea* en raison de ce petit nombre d'individus ne pourra donc pas être étudier sur base de cet échantillonnage.

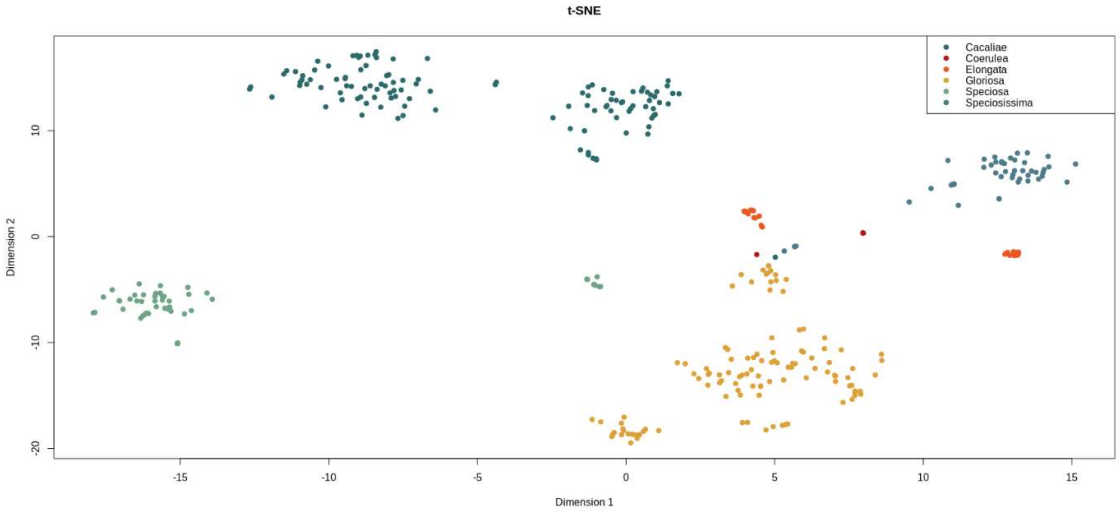


Fig. 14 : Graphe TSNE du genre Oreina

t-SNE (t-distributed Stochastic Neighbor Embedding) est une technique de réduction dimensionnelle principalement utilisée pour visualiser des données en 2 ou 3 dimensions. Contrairement à l'ACP, t-SNE est non linéaire et excelle à préserver les relations locales, c'est-à-dire qu'il maintient les points proches dans l'espace d'origine proches dans l'espace réduit.
t-SNE commence par calculer les probabilités de proximité entre chaque paire de points dans l'espace haute dimension. Ensuite, il tente de trouver une représentation dans un espace de dimension inférieure qui préserve ces proximités.

Les espèces sont clairement séparées en plusieurs clusters. Quelques points semblent légèrement plus dispersés, ce qui pourrait indiquer une certaine variabilité au sein de ces individus qui partagent des caractéristiques avec plusieurs groupes.
Les *Cacaliae* sont ici divisée en deux clusters majeurs, ce qui pourrait indiquer une sous-structure ou une variabilité importante au sein de cette espèce. L'études des populations nous on apprendra plus sur les différences génétiques présentes au sein de cette espèce.

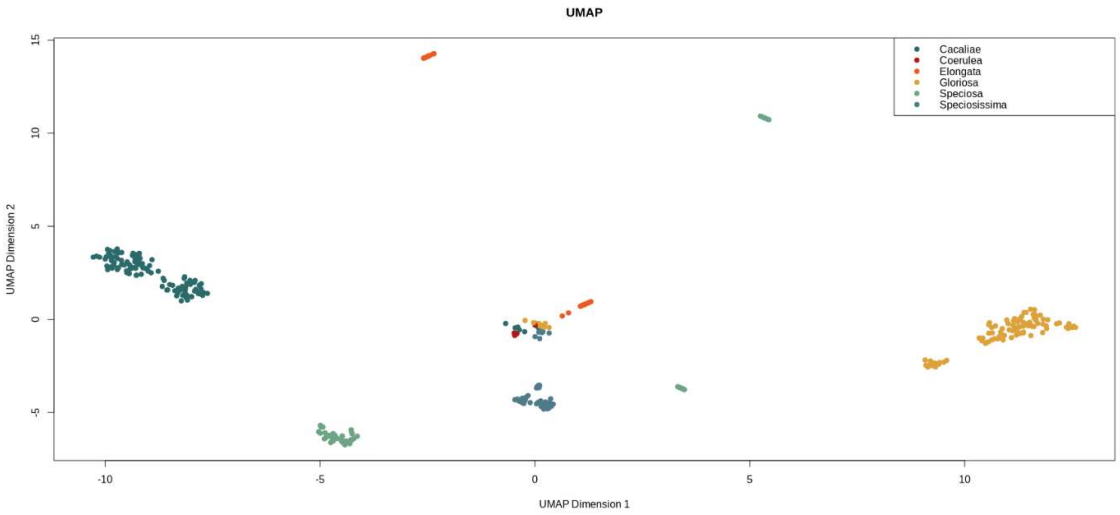


Fig. 15 : Graphe UMAP du genre Oreina

UMAP(Uniform Manifold Approximation and Projection) est une autre technique non linéaire de réduction de la dimension.
UMAP construit un graphe de proximité des données dans l'espace d'origine, en connectant chaque point à ses plus proches voisins.
Ensuite, il essaie de trouver une représentation dans un espace de dimension réduite qui préserve autant que possible les connexions et distances du graphe original.

L'ACP est linéaire et peut manquer certaines structures complexes présentes dans les données. t-SNE et UMAP sont non linéaires et peuvent révéler des patterns et des sous-groupes moins visible sur l'ACP.
UMAP et t-SNE se complètent, car UMAP donne une meilleure idée des relations globales entre les groupes, tandis que t-SNE révèle des nuances locales au sein de chaque groupe. Leur utilisation conjointe offre une vue plus complète.

4.2. Délimitation des populations

4.2.1. Cacaliae

Fig. 16 : Oreina cacaliae male(gauche) femelle(droite).

Oreina cacaliae est une espèce de chrysomèle qui se distingue par sa répartition dans les régions alpines. Comme toutes les espèces du genre *Oreina*, *O. cacaliae* est phytophage et se nourrit principalement de plantes de la famille des Astéracées. Cette espèce, non volante, est limitée dans ses capacités de dispersion, ce qui influence fortement sa structure génétique et sa distribution géographique.

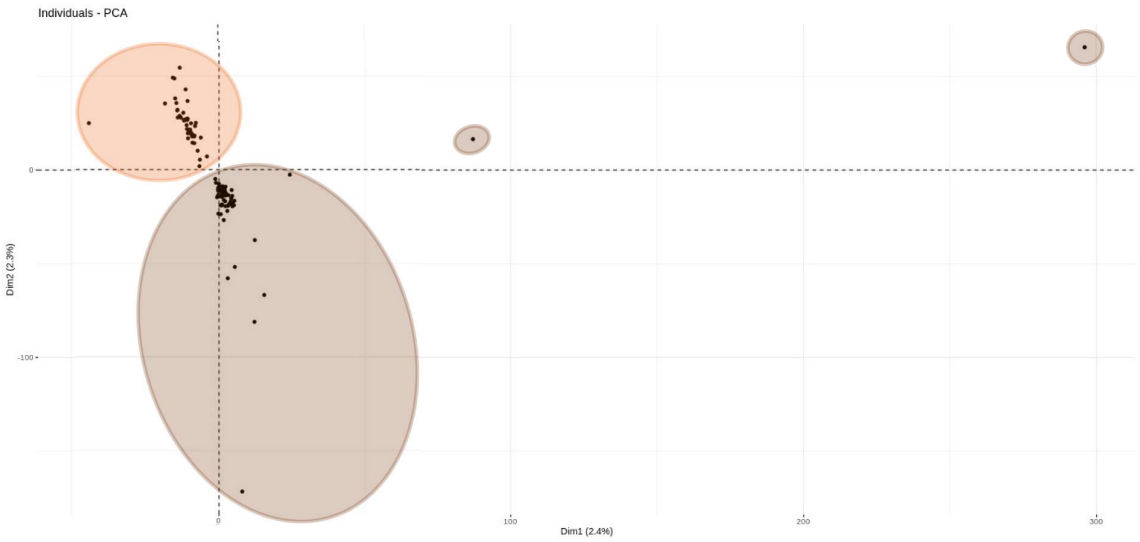


Fig. 17 : ACP avec mise en évidence des différentes populations de O.cacaliae

L'analyse en composantes principales (ACP) pour *O. cacaliae* montre une structuration claire des populations, avec deux groupes distincts qui reflètent une différenciation génétique significative. Chaque point sur le graphique représente un individu, et les groupes qui se forment suggèrent une homogénéité au sein des populations, mais une diversité génétique importante entre elles.

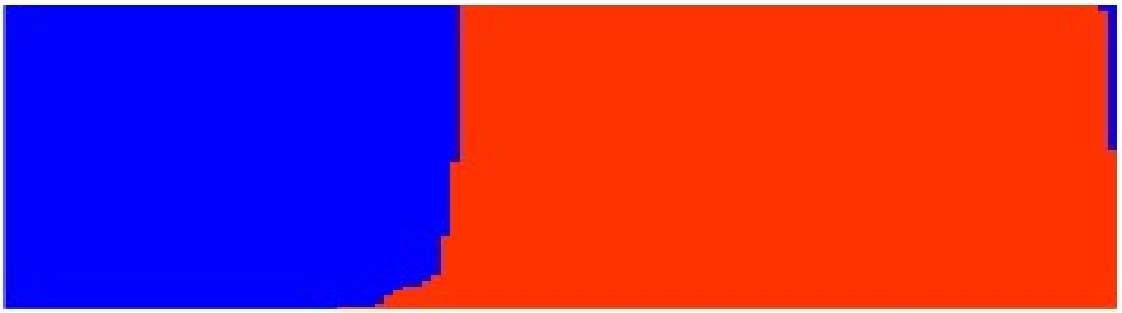


Fig. 18 : Diagramme de structure de population de O.cacaliae

Le diagramme de structure de population indique que les individus de *O. cacaliae* appartiennent principalement à deux cluster génétique, ce qui suggère une faible diversité génétique intra-espèces.

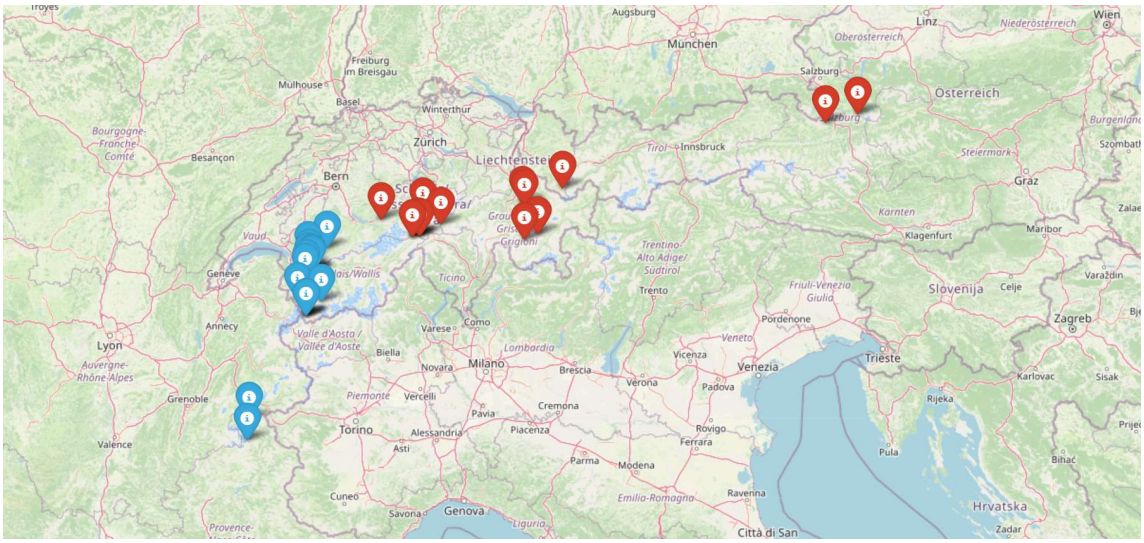


Fig. 19 : Mise en évidence de la répartition géographique des populations au sein de O.Cacaliae

La répartition géographique des populations montre que les deux populations de *O. cacaliae* sont distribuées dans deux zones distinctes des Alpes, ce qui pourrait expliquer la structuration observée dans les analyses génétiques. Les barrières géographiques naturelles telles que les montagnes peuvent limiter les échanges entre populations, menant à une différenciation génétique locale.

La vallée du Rhône en Suisse et le Valais forment une grande vallée entre les massifs montagneux qui pourrait également jouer un rôle dans la séparation des populations.

L'histoire glaciaire des Alpes, avec des refuges post-glaciaires potentiels, pourrait également jouer un rôle clé dans cette répartition.

4.2.2. Speciosissima

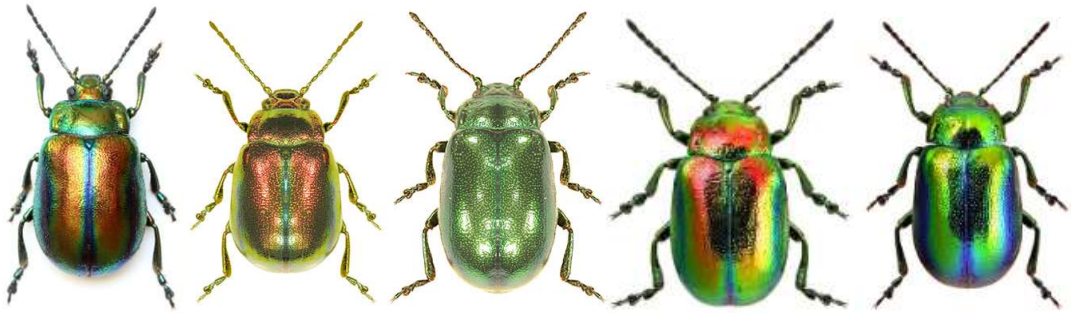


Fig. 20 : Oreina speciosissima

Oreina speciosissima est une autre espèce montagnarde, qui partage plusieurs traits écologiques avec *O. cacaliae*, notamment son régime alimentaire et son incapacité à voler. Cela conduit à une dynamique similaire en termes de dispersion et de structuration génétique.

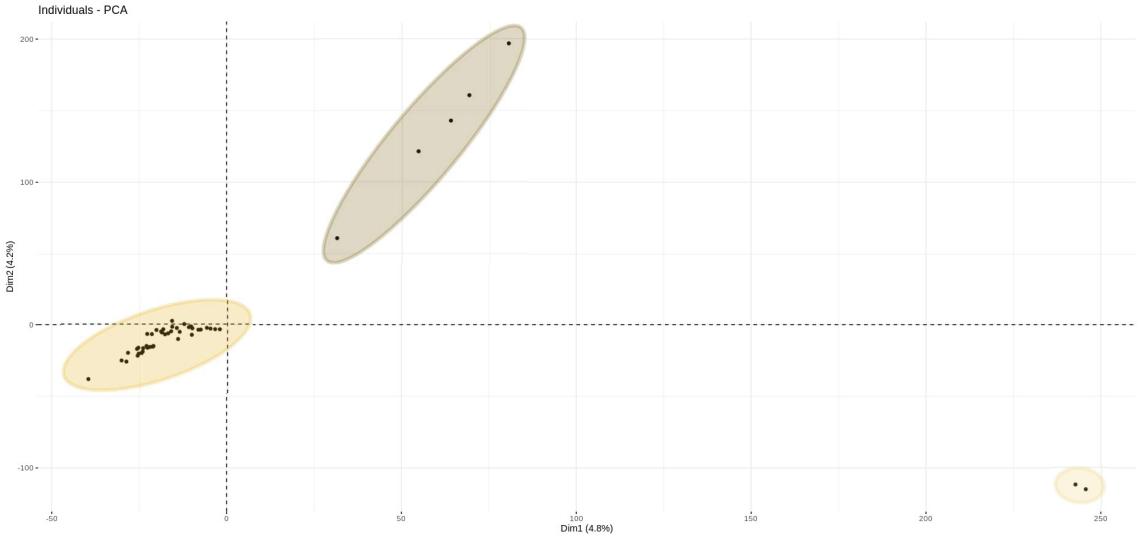


Fig. 21 : ACP avec mise en évidence des différentes populations de O.speciosissima

L'ACP pour O. speciosissima révèle une structuration génétique marquée, avec 3 populations bien différenciées.

Fig. 22 : Diagramme de structure de population de O.speciosissima

Le diagramme de structure de population montre également une faible diversité génétique au sein des populations, similaire à ce qui est observé pour *O. cacaliae*.

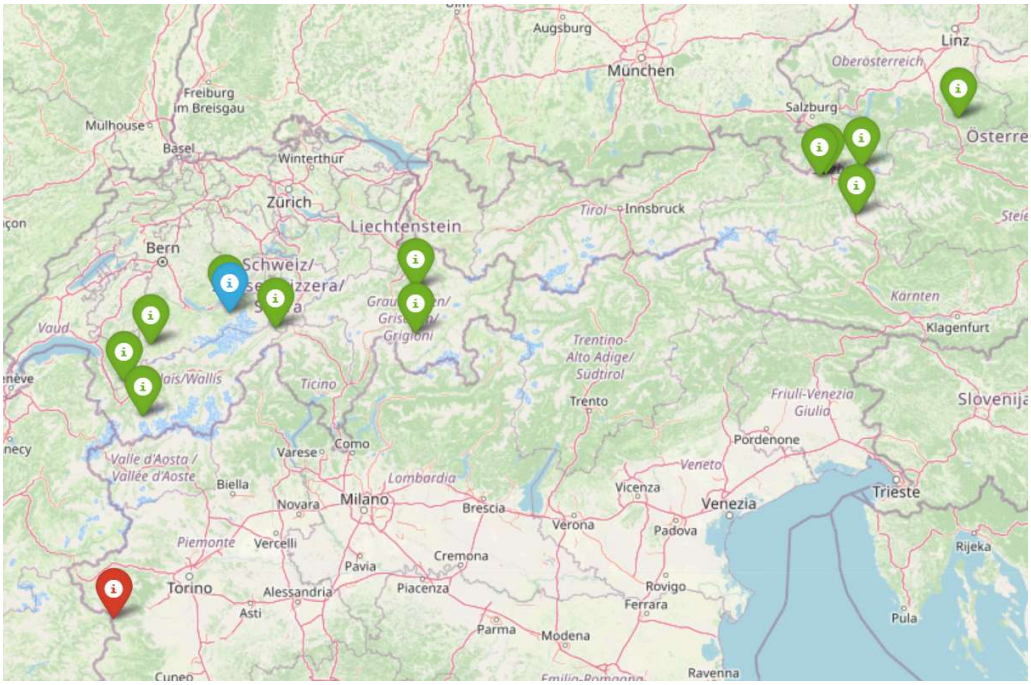


Fig. 23 : Mise en évidence de la répartition géographique des populations au sein de O.speciosissima

La répartition géographique des populations de *O. speciosissima* montre que les individus des populations représentées en rouge et en bleu ont tous été retrouvés au même endroit. La population représentée en rouge est isolée géographiquement des autres individus de l'espèce. La population représentée en bleu est isolée par les Alpes orientales et la vallée de l'Inn. Ces isolements limitent les rencontres et échange entre populations et peuvent expliquer la différence génétique de ses individus.

4.2.3. *Speciosa*

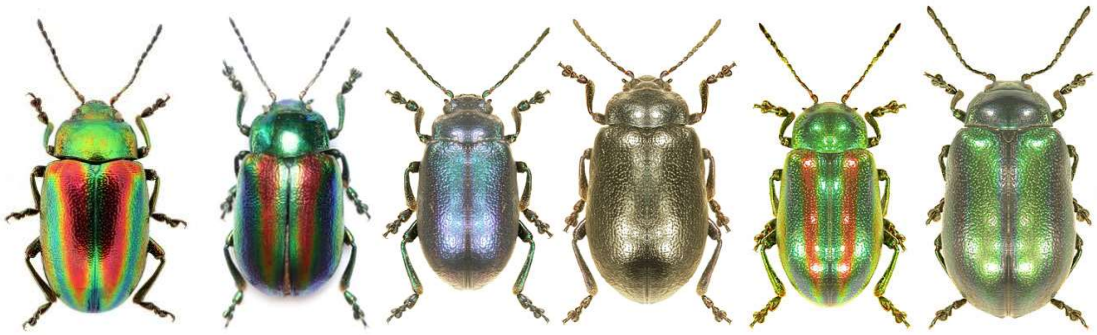


Fig. 24 : Oreina specioa

Oreina speciosa est également une espèce non volante du genre *Oreina,* adaptée aux environnements montagnards des Alpes. Comme ses congénères, elle présente une spécialisation écologique qui limite ses capacités de dispersion, influençant sa structure génétique.

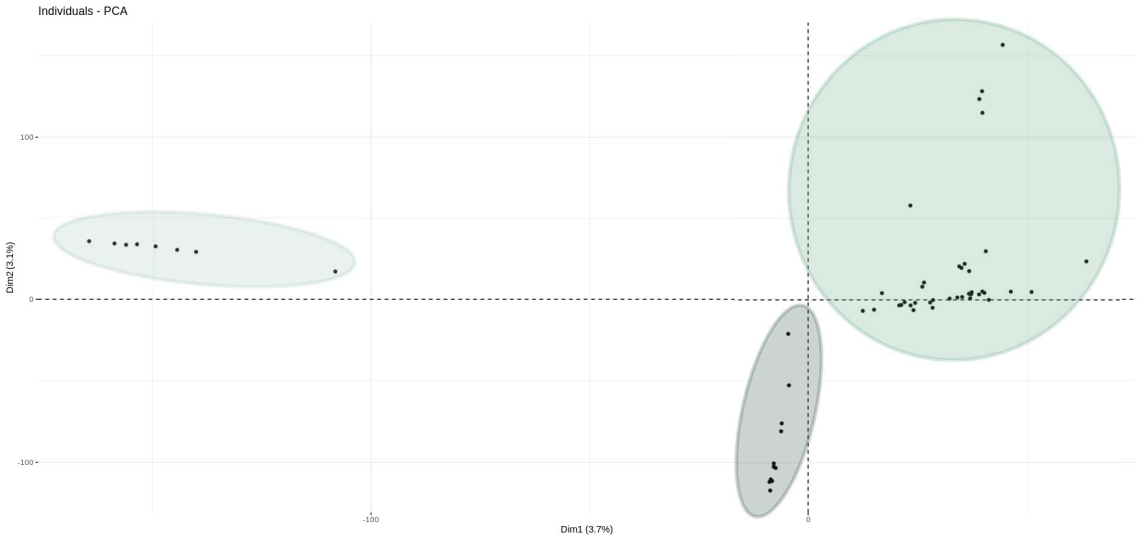


Fig. 25 : ACP avec mise en évidence des différentes populations de O.speciosa

L'ACP pour *O. speciosa* montre 3 populations distinctes, avec une séparation nette entre les différents groupes.

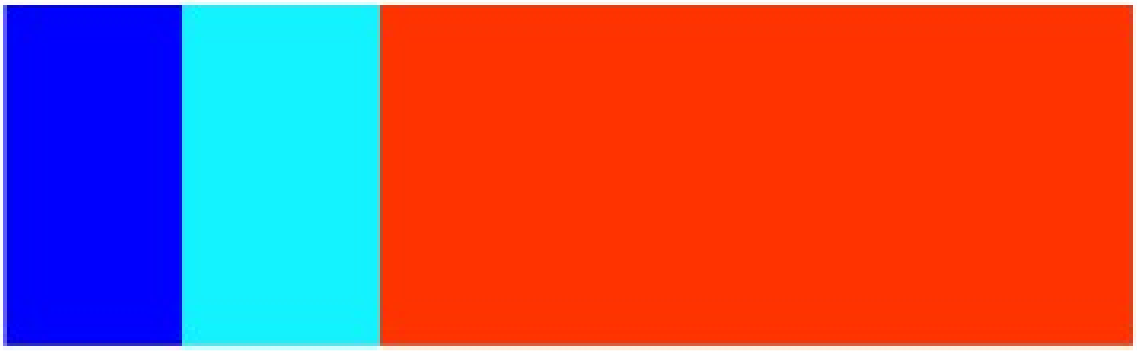


Fig. 26 : Diagramme de structure de population de O.speciosa

Le diagramme de structure de population pour *O. speciosa* révèle une situation similaire à celle des deux autres espèces, avec 3 clusters principaux.

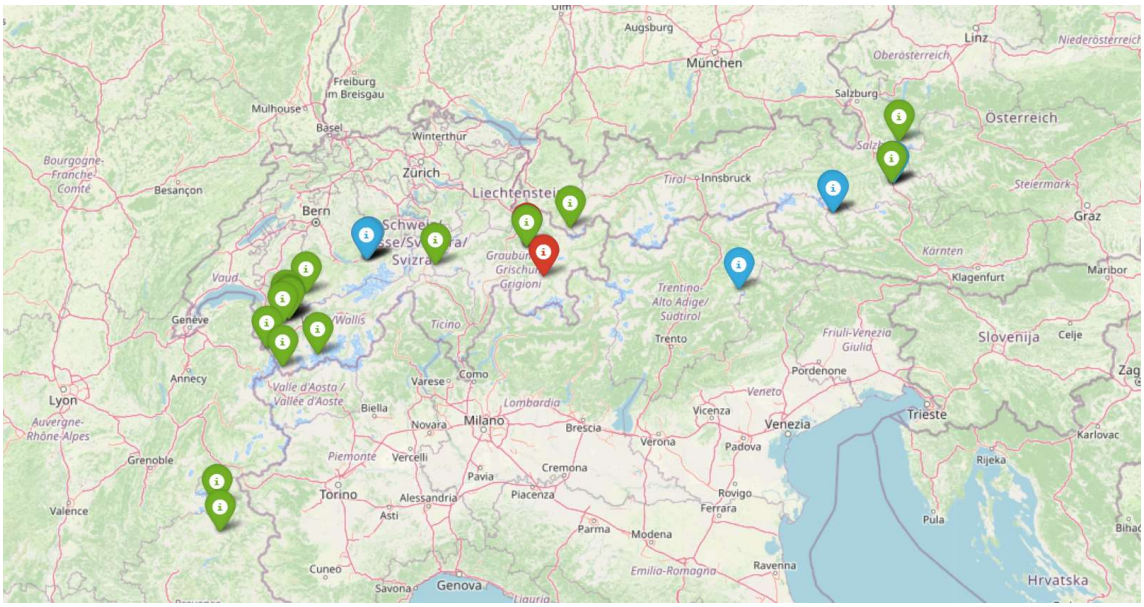


Fig. 27 : Mise en évidence de la répartition géographique des populations au sein de O.speciosa

Les individus représenté en rouge sont principalement isolées par les Alpes centrales et les profondes vallées des Grisons, tandis que les populations bleues sont isolées par les Hautes Alpes et la vallée de l'Inn dans le Tyrol. Ces barrières naturelles jouent un rôle crucial dans l'isolement de ces groupes par rapport aux autres populations. Ce qui explique leurs différence génétique.

4.3. Modélisation spatio-temporelle

L'analyse structurelle des population de l'espèce *O.cacaliae* à révélé la présence de deux populations dont la séparation géographique est nette.

Huit hypothèses ont été formulées concernant les origines de ces deux populations.

Scénario 1	Division Ancestrale	Hypothèse du Refuge
Scénario 2	Division Récente	
Scénario 3	Colonisation Ancestrale de la population 1 vers la population 2	
Scénario 4	Colonisation Ancestrale de la population 2 vers la population 1	
Scénario 5	Colonisation Récente de la population 1 vers la population 2	
Scénario 6	Colonisation Récente de la population 1 vers la population 2	
Scénario 7	2 Colonisation Ancestrale à partir de l'extérieur	Hypothèse de la recolonisation
Scénario 8	2 Colonisation Récente à partir de l'extérieur	

Le diagramme de coalescence est utilisé pour illustrer les hypothèses concernant l'histoire démographique des populations, Il met en évidence les changements de taille des populations au cours du temps.

Légende :

NA	Nombre d'individu dans la population ancestrale
N1	Nombre d'individu dans la population 1
N2	Nombre d'individu dans la population 2
NMAXGLAC	Nombre d'individu durant le maximum glaciaire
NFON	Nombre d'individu fondateur d'une nouvelle colonie
T1	Fin de l'ère glaciaire 11000
T2	Fin du maximum glaciaire18000
T3	Début du maximum glaciaire23000
T4	Début de l'ère glaciaire 115000

Scénario 1 :

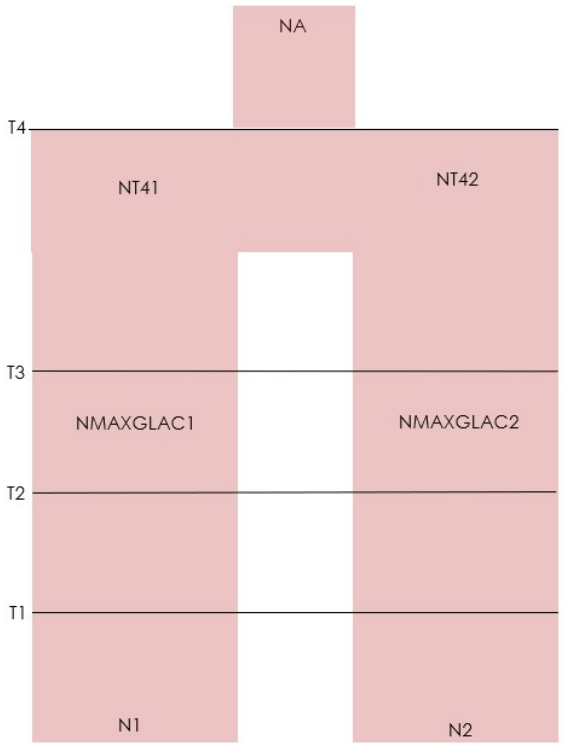


Fig. 28 :Diagramme de coalescence scénario 1

Les diagrammes 7 autres scénario sont présent en annexe 8.

Le logiciel fastsimcoal a été utilisé pour tenter de modéliser les scénarios démographiques des populations étudiées. Fastsimcoal est un outil couramment utilisé pour effectuer des simulations coalescentes sous différents scénarios démographiques et pour estimer des paramètres tels que les tailles de population ancestrales, les événements de migration, et les changements dans les tailles de population au cours du temps.

Malgré plusieurs tentatives, l'exécution du programme n'a pas produit les résultats escomptés. Le logiciel fonctionnait sans générer de messages d'erreur, mais les fichiers de sortie étaient vides, indiquant que le programme n'a pas réussi à réaliser les simulations prévues.

L'absence de résultats peut être due à plusieurs facteurs, notamment :

- Problèmes de Paramétrage
- Problèmes Techniques : Il est possible que des limitations matérielles ou des configurations incorrectes du système aient entravé le bon fonctionnement du programme.

Le script et les différents fichiers de paramètre utilisé peuvent être consulté en annexe 9 .

5. Discussion

Les résultats de cette étude mettent en lumière l'impact des facteurs géographiques sur la structuration génétique des chrysomèles du genre *Oreina*. Les analyses génétiques, incluant l'ACP et les diagrammes de structure de population, montrent une différenciation significative entre les populations des trois espèces étudiées. Cette différenciation est probablement le résultat de la combinaison de plusieurs facteurs :

Barrières géographiques naturelles : Les chaînes de montagnes, les vallées profondes, et d'autres obstacles physiques ont limité les mouvements et les échanges génétiques entre les populations, favorisant ainsi une forte structuration génétique.

Histoire glaciaire des Alpes : Le dernier maximum glaciaire a forcé les espèces à se réfugier dans des zones spécifiques, où elles ont survécu et se sont diversifiées. La recolonisation des habitats alpins après la fonte des glaciers a suivi des chemins variés, contribuant à la diversité génétique observée aujourd'hui.

Limitation des capacités de dispersion : En tant qu'espèces non volantes, les chrysomèles du genre *Oreina* sont particulièrement sensibles à la fragmentation de leur habitat. Cette faible capacité de dispersion a probablement renforcé l'isolement des populations, exacerbant les différences génétiques locales.

Ces facteurs combinés expliquent la structuration génétique observée et suggèrent que les populations actuelles sont le résultat d'un long processus de diversification.

6. Conclusion générale

Ce travail de fin d'étude a exploré les structures de population de plusieurs espèces du genre *Oreina*. À partir de ce genre, six espèces distinctes ont été identifiées à l'aide des outils bio-informatiques STACKS, de l'analyse en composantes principales (ACP), et du programme faststructure. Parmi ces six espèces, trois ont été sélectionnées pour une étude approfondie des structures de population : *Oreina cacaliae*, *Oreina speciosissima*, et *Oreina speciosa*.

Les résultats obtenus suggèrent que l'isolement géographique dans les environnements complexes comme les montagnes joue un rôle crucial dans la différenciation locale de ses espèces non volantes comme. Cette différenciation, couplée à la faible capacité de dispersion de ces espèces, a conduit à une répartition des populations étroitement liée aux caractéristiques topographiques des Alpes.

Malgré les défis techniques, notamment avec l'utilisation de fastsimcoal pour la modélisation démographique, plusieurs hypothèses sur l'évolution des populations ont été formulées.

Ces conclusions apportent des perspectives nouvelles sur les dynamiques évolutives dans les environnements montagnards, soulignant l'importance de la géographie physique dans la diversification génétique.

7. Perspectives

Les résultats obtenus ouvrent plusieurs possibilités de recherche :

Études comparatives avec d'autres espèces alpines : Il serait intéressant de comparer les résultats obtenus pour le genre *Oreina* avec ceux d'autres espèces montagnardes, afin de déterminer si les mêmes processus évolutifs et géographiques influencent leur structuration génétique.

Modélisation démographique: Il n'a pas été possible d'intégrer les résultats de modélisation démographique issus de fastsimcoal dans cette étude. Pour des travaux futurs, il pourrait être utile de

- Revoir et ajuster les paramètres d'entrée.
- S'assurer que le système utilisé pour exécuter le logiciel est adapté aux exigences techniques de fastsimcoal.
- Explorer des alternatives ou des outils complémentaires pour la modélisation démographique.

Impact du changement climatique : Les effets du réchauffement climatique sur la répartition des populations d'*Oreina* pourraient être explorés, en particulier en ce qui concerne les possibilités de recolonisation de nouveaux habitats ou la réduction des zones habitables.

Études génétiques supplémentaires : Des analyses génétiques plus approfondies, incluant des marqueurs microsatellites ou des séquençages de génomes complets, pourraient révéler des aspects encore méconnus de la diversité génétique au sein de ces espèces.

8. Bibliographie

analyse en composante principale. (s.d.). Récupéré sur centre de la science et de la biodiversité du quebec: https://r.qcbs.ca/workshop09/book-fr/analyse-en-composantes-principales.html

Andrew, K. J. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, pp. 81-92.

Avise, J. C. (2009). Phylogeography: retrospect and prospect. *Journal of Biogeography, 36(1)*, pp. 3-15.

Baird, N. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLOS ONE, 3(10), e3376.*

C. Kastally, S. D. (2021). Estimating Migration of Gonioctena quinquepunctata (Coleoptera: Chrysomelidae) Inside a Mountain Range in a Spatially Explicit Context.

Catchen J., H. P. (2013). Stacks: an analysis tool set for population genomics.

Davey, J. W. (2010). RADSeq: next-generation population genetics. *Briefings in Functional Genomics, 9(5-6)*, pp. 416-423.

E. Pante, J. A. (2015). Use of RAD sequencing for delimiting species.

Frankham, R. B. (2002). Introduction to Conservation Genetics. *Cambridge University Press.*

G. Evanno, S. R. (2005). Detecting the number of clusters of individuals using the software STRUCTURE : a simulation study.

Hedrick, P. W. (2005). Genetics of Populations. *Jones & Bartlett Learning.*

Holderegger, R. &.-E. (2009). Impacts of climate change on the recolonization and persistence of species in alpine areas. *Biological Conservation, 142(8)*, pp. 1487-1497.

Jolivet. (2002). Biology of Leaf Beetles. *Intercept Limited*.

Jolivet, P. &. (1986). Host-plants of Chrysomelidae of the world: an essay about the relationships between the leaf-beetles and their food-plants. *Backhuys Publishers*.

Kalberer, N. M. (2005). An alternative, biologically relevant method for the investigation of chemically mediated tritrophic interactions: The use of rooted host plants. *Journal of Chemical Ecology, 31(9)*, pp. 1877-1884.

Kidd, D. M. (2006). Phylogeographic information systems: putting the geography into phylogeography.. *Journal of Biogeography, 33(11)*, pp. 1851-1865.

L. Excoffier, I. D.-S. (2013). Robust Demographic Inference from Genomic and SNP Data.

Miller, M. R. (2007). Rapid and cost-effective polymorphism identification and genotyping using sequenced RAD markers. *Public Library of Science, 3(10), e3376.*

Pasteels, J. M.-R. (1989). Defensive mechanisms of chrysomelid beetles. *In J. G. Pasteels & A. C. Beevor (Eds.), Insect Chemical Ecology. Wiley.*

Peterson, B. K. (2012). Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLOS ONE, 7(5), e37135.*

Pinceel, J. J. (2005). The genetic structure of alpine populations of the land snail Albinaria coerulea: evidence for survival in ice-free refugia. *Molecular Ecology, 14(9)*, pp. 2415-2423.

Triponez, Y. S. (2011). Population structure and limited dispersal in the flightless alpine leaf beetle Oreina elongata (Coleoptera: Chrysomelidae). *European Journal of Entomology, 108(4)*, pp. 607-615.

9. Table des Annexes

Annexe 1 : Tableau de données de récolte

Localité	Altitude (m)	Coordonnées	Date	Plantes hôtes	n	extraction1	extraction2	extraction3	Noms
Col du Galibier 1	2000	45.085552 N - 6.436986 E	6 juillet 2021	Peucedanum ostruthium	15	2	5		Gali1-3 à Gali1-7
Col du Galibier 2	2000	45.085552 N - 6.436986 E	6 juillet 2021	Adenostyles sp.	9	2	5	5	Gali2-3 à Gali2-7
Col du Galibier 3	2438	45.053494 N - 6.404769 E	6 juillet 2021	Peucedanum ostruthium	14	2	3		Gali3-3 à Gali3-5
Pré de Mme Carle	1870	44.916816 N - 6.415810 E	7 juillet 2021	Peucedanum ostruthium	16	2	7		CarlP-3 à CarlP-9
Pré de Mme Carle	1870	44.916816 N - 6.415810 E	7 juillet 2021	Adenostyles sp. (leucophylla?)	16	2	5		CarlA-3 à CarlA-7
Palluel 1	1800	44.733380 N - 6.449525 E	7 juillet 2021	Peucedanum ostruthium	16	2	7		Pall1-3 à Pall1-9
Valpréveyre	1940	44.813906 N - 6.974973 E	8 juillet 2021	Peucedanum ostruthium	5	2	3		Valp-3 à Valp-5
Col de Bouchet	2500	44.817918 N - 7.017292 E	8 juillet 2021	Senecio sp.		2	3		Bouc-1 à Bouc-3
Clapeyto		44.799167 N - 6.691389 E	9 juillet 2021	Peucedanum ostruthium et Adenostyles	6	2	4		Clap-3 à Clap-6
Palluel 2	2400	44.728053 N - 6.417923 E	9 juillet 2021	Cirsium sp.	11	2	5		Pall2-3 à Pall2-7
La Lenta 1	2140	45.386358 N - 7.039131 E	10 juillet 2021	Peucedanum ostruthium	15	2	5		Lent1-3 à Lent1-7
La Lenta 2	2140	45.386358 N - 7.039131 E	10 juillet 2021	Centaurea uniflora (3) et Adenostyles sp. (1)	4	2	5		Lent2-3 à Lent2-7
Ouillette	2409	45.432201 N - 7.001632 E	10 juillet 2021	Peucedanum ostruthium	11	2	5		Ouil-3 à Ouil-7
La Fouly	1700 - 2000	45.930364 N - 7.096628 E	6 août 2021	Adenostyles sp.	15	2	3		FoulA-3 à FoulA-5
La Fouly	1850 - 2000	45.927969 N - 7.102654 E	6 août 2021	Peucedanum ostruthium	20	2	7		FoulP-3 à FoulP-9
La Fouly	> 2000	45.927969 N - 7.102654 E	6 août 2021	Petasites sp.	8	2	3		FoulT-3 à FoulT-5
Scheidback	1400 -1800	46.466178 N - 7.336962 E	7 août 2021	Adenostyles sp.	10	2	5		ScheA-3 à ScheA-7
Scheidback	1600 -1900	46.450306 N - 7.336962 E	7 août 2021	Peucedanum ostruthium	10	2	6		ScheP-3 à ScheP-8
Iseltwald1	1400	46.695907 N - 7.969272 E	8 août 2021	Petasites sp.	9	2	6		Isel1-3 à Isel1-8
Iseltwald2	1700	46.705969 N - 7.991015 E	8 août 2021	Petasites sp.	13	2	5		Isel2-3 à Isel2-7
Isetlwald	1700	46.705969 N - 7.991015 E	8 août 2021	Peucedanum ostruthium	20	2	5		IselP-3 à IselP-7
Faulhorn	2500	46.660108 N - 8.022879 E	9 août 2021	Senecio grandiflora	6	2	4		Faul-3 à Faul-6
Col des Essets 1	1900 - 2000	46.274298 N - 7.159264 E	10 août 2021	Adenostyles sp. (présence Peucedanum)	11	2	5		Esse1-3 à Esse1-7
Col des Essets 2	2000	46.275602 N - 7.161782 E	10 août 2021	Peucedanum ostruthium (isolés)	5	2	3		Esse2-3 à Esse2-5
Col des Essets 3	2000	46.277391 N - 7.165223 E	10 août 2021	Adenostyles sp. (isolés)	11	2	5		Esse3-3 à Esse3-7
Anzeinde	2000	46.285420 N - 7.167495 E	10 août 2021	Adenostyles sp. (isolés)	10	2	5		Anze-3 à Anze-7
Plans	1400	46.251233 N - 7.107846 E	11 août 2021	Petasites sp.	7	2	5		Plan-3 à Plan-7
Sapolaires	1800 - 2200	46.244251 N - 7.096687 E	11 août 2021	Adenostyles sp.	12	2	5		SapoA-3 à SapoA-7
Sapolaires	1800 - 2200	46.244251 N - 7.096687 E	11 août 2021	Peucedanum ostruthium	20	2	7		SapoP-3 à SapoP-9
Perris	2150 - 1800	46.212698 N - 7.087400 E	11 août 2021	Adenostyles sp.	14	2	7		Perr-3 à Perr-9
Emosson	1900	46.068695 N - 6.937859 E	12 août 2021	Peucedanum ostruthium	10	2	5		Emos-3 à Emos-7
Chaume1	2000	46.293777 N - 7.118768 E	12 août 2021	Adenostyles sp.	14	2	5		Chau1-3 à Chau1-7
Chaume2	2000	46.297110 N - 7.127493 E	12 août 2021	Peucedanum ostruthium	11	2	5		Chau2-3 à Chau2-7
Col de la Forclaz, CH	1527	46.0576 N, 7.0010 E	9 mai 2022	Petasites	15			5	Forc1 à Forc5
Fionnay, CH	1370	46.0418 N, 7.2766 E	10 mai 2022	Petasites	14			5	Fionn1 à Fionn5
Tschiertschen, CH	1350 - 1450	46.8203 N, 9.6012 E	13 mai 2022	Petasites	15			5	Tsch1 à Tsch5
Mulengs, CH	1470	46.5374 N, 9.62371 E	14 mai 2022	Petasites	9			5	Mule1 à Mule5
Chamonix	1780	45.9998 N - 6.9170 E	17 juillet 2022	Peucedanum ostruthium ?	10			5	Chamo1 à Chamo5
Arolla1	1820 - 2100	46.0489 N - 7.4916 E	18 juillet 2022	Peucedanum ostruthium	> 20			5	Arol1-1 à Arol1-5
Arolla2	2080 - 2410	46.0201 N - 7.4640 E	18 juillet 2022	Peucedanum ostruthium	> 10			5	Arol2-1 à Arol2-5
Lioson	1650 - 1950	46.3877 N - 7.1211 E	19 juillet 2022	Adenostyles	> 20			5	Lios1 à Lios5
Croix	1780	46.3244 N - 7.1269 E	19 juillet 2022	Adenostyles, Peucedanum	7			5	Croi1 à Croi5
Furka	2044	46.5702 N - 8.3983 E	20 juillet 2022	Adenostyles	15			5	Furk1 à Furk5
Furka2A	2400 - 2500	46.5635 N - 8.4143 E	20 juillet 2022	Adenostyles	8			5	Furk2A-1 à Furk2A-5
Furka2C	2400 - 2500	46.5635 N - 8.4143 E	20 juillet 2022	Cirsium spinosissimum	14			7	Furk2C-1 à Furk2C-7
Furka2P	2400 - 2500	46.5635 N - 8.4143 E	20 juillet 2022	Peucedanum ostruthium	> 10			7	Furk2P-1 à Furk2P-7
GrimselA	2200	46.5598 N - 8.3320 E	20 juillet 2022	Adenostyles	> 10			5	GrimA-1 à GrimA-5
GrimselC	2200	46.5598 N - 8.3320 E	20 juillet 2022	Cirsium spinosissimum	10			7	GrimC-1 à GrimC-7
Susten	2224	46.7296 N - 8.4471 E	21 juillet 2022	Adenostyles	7			5	Sust1 à Sust5
Oberal	2027	46.6589 N - 8.6535 E	21 juillet 2022	Adenostyles	1			1	Ober1 à Ober5
Oberal2	2000	46.6560N - 8.6768 E	21 juillet 2022	Peucedanum ostruthium, Adenostyles	15			7	Ober2-1 à Ober2-7
Löser	1610 - 2100	46.7998 N - 9.6203 E	22 juillet 2022	Adenostyles, Petasites, Senecio fuchsii	> 30			5	Lose1 à Lose5
Carmenna	1986 m	46.7859 N - 9.6218 E	22 juillet 2022	Peucedanum ostruthium	> 10			7	Carm1 à Carm7
Preda	1900 m	46.5849 N - 9.7786 E	22 juillet 2022	Adenostyles	> 10			5	Preda1 à Preda5
Preda2	2072 m	46.5766 N - 9.8014 E	22 juillet 2022	Peucedanum ostruthium	15			5	Pred2-1 à Pred2-5
Gaschurn	1940 m	46.9166 N - 10.0740 E	5 août 2022	Peucedanum ostruthium	1			1	Gasc1-1
Gaschurn2	1690 m	46.9405 N 10.0590 E	5 août 2022	Adenostyles	15			5	Gasc2-1 à Gasc2-5
Passo Pordoi	2228 m	46.4807 N 11.8192 E	6 août 2022	Peucedanum ostruthium	1			1	Pord1
Heiligenblut	2000	47.0255 N 12.7987 E	8 août 2022	Adenostyles	1			1	Heil1-1
Heiligenblut2	1660	47.0304 N 12.7920 E	8 août 2022	Adenostyles	1			1	Heil2-1
Annaberg im lammertal	1080	47.5098 N 13.4675 E	8 août 2022	Petasites	2			2	Anna1-1 à Anna1-2
Annaberg2	1630	47.5274 N 13.4754 E	9 août 2022	Adenostyles	8			5	Anna2-1 à Anna2-5
Ramsau	1130 - 1470	47.7966 N 14.3039 E	10 août 2022	Adenostyles, Senecio	> 30			5	Rams1 à Rams5
Dielalm	1160 - 1450	47.4724 N 13.1648 E	10 août 2022	Senecio, Petasites, Adenostyles	15			5	Diel1-1 à Diel1-5
Dielalm2	1670	47.4598 N 13.1079 E	11 août 2022	Senecio, Adenostyles	5			5	Diel2-1 à Diel2-5
Tauernalm	1256 - 1660	47.2372 N 13.4206 E	11 août 2022	Petasites, Senecio, Adenostyles	> 30			5	Taue1-1 à Taue1-5
Tauernalm2	1950	47,2361 N 13,4078 E	11 août 2022	Peucedanum ostruthium	6			5	Taue2-1 à Taue2-5
				TOTAL		**66**	**163**	**162**	**391**

Annexe 2 : Script Python permettant de réaliser les cartes

```python
import folium

# Liste des coordonnées (latitude, longitude)
coordinates = [
(45.085552, 6.436986),
(45.085552, 6.436986),
(45.053494, 6.404769),
(44.916816, 6.415810),
(44.916816, 6.415810),
(44.733380, 6.449525),
(44.813906, 6.974973),
(44.817918, 7.017292),
(44.799167, 6.691389),
(44.728053, 6.417923),
(45.386358, 7.039131),
(45.386358, 7.039131),
(45.432201, 7.001632),
(45.930364, 7.096628),
(45.927969, 7.102654),
(45.927969, 7.102654),
(46.466178, 7.336962),
(46.450306, 7.336962),
(46.695907, 7.969272),
(46.705969, 7.991015),
(46.705969, 7.991015),
(46.660108, 8.022879),
(46.274298, 7.159264),
(46.275602, 7.161782),
(46.277391, 7.165223),
(46.285420, 7.167495),
(46.251233, 7.107846),
(46.244251, 7.096687),
(46.244251, 7.096687),
(46.212698, 7.087400),
(46.068695, 6.937859),
(46.293777, 7.118768),
(46.297110, 7.127493),
(46.0576, 7.0010),
(46.0418, 7.2766),
(46.8203, 9.6012),
(46.5374, 9.62371),
(45.9998, 6.9170),
(46.0489, 7.4916),
(46.0201, 7.4640),
```

```
    (46.3877, 7.1211),
    (46.3244, 7.1269),
    (46.5702, 8.3983),
    (46.5635, 8.4143),
    (46.5635, 8.4143),
    (46.5635, 8.4143),
    (46.5598, 8.3320),
    (46.5598, 8.3320),
    (46.7296, 8.4471),
    (46.6589, 8.6535),
    (46.6560, 8.6768),
    (46.7998, 9.6203),
    (46.7859, 9.6218),
    (46.5849, 9.7786),
    (46.5766, 9.8014),
    (46.9166, 10.0740),
    (46.9405, 10.0590),
    (46.4807, 11.8192),
    (47.0255, 12.7987),
    (47.0304, 12.7920),
    (47.5098, 13.4675),
    (47.5274, 13.4754),
    (47.7966, 14.3039),
    (47.4724, 13.1648),
    (47.4598, 13.1079),
    (47.2372, 13.4206),
    (47.2361, 13.4078),
]

# Créer la carte centrée sur la première paire de coordonnées
m = folium.Map(location=coordinates[0], zoom_start=8)

# Ajoute des points pour chaque coordonnée
for coord in coordinates:
    folium.Marker(location=coord).add_to(m)

# Enregistrer la carte en tant que fichier HTML
m.save("map.html")
```

Annexe 3 : Script nombre de reads par individus

```
#!/bin/bash

output_data="sequence_counts.dat"

echo "# Filename Number_of_Sequences" > $output_data


for file in *.fq.gz; do
  if [ -f "$file" ]; then
    count=$(zcat "$file" | grep -c "^@")
    echo "$file $count" >> $output_data
  fi
done


# Script gnuplot pour générer le graphe

gnuplot_script="plot_script.gp"

cat << 'EOF' > $gnuplot_script

set terminal png size 800,600

set output 'sequence_counts.png'

set title 'Number of Sequences in .fq.gz Files'

set xlabel 'File'

set ylabel 'Number of Sequences'

set xtics rotate by -45

set grid

set style data histogram

set style fill solid 1.00 border -1

set boxwidth 0.9

plot 'sequence_counts.dat' using 2:xtic(1) title 'Sequences' with
histogram

EOF


gnuplot $gnuplot_script
```

Annexe 4 : Script envoyé sur le cluster

```
#!/bin/bash

# job parameters

#SBATCH --job-name=qpstacks
#SBATCH --mail-user=alessia.bellanca@std.heh.be
#SBATCH --mail-type=ALL
#SBATCH --output=%j_out

# job resources

#SBATCH --time=4:00:00
#SBATCH --partition=batch
#SBATCH --ntasks=1
#SBATCH --ntasks-per-node=1
#SBATCH --cpus-per-task=8
#SBATCH --mem-per-cpu=8000

module load releases/2018b
module load Stacks/2.41-foss-2018b

cd stacks

process_radtags -f ../base/UO_C659_1.fastq.gz -o ./samples -b ../base/barcodes -e sgrAI -r -c -q
```

```
#!/bin/bash

# job parameters

#SBATCH --job-name= qpstacks
#SBATCH --mail-user=alessia.bellanca@std.heh.be
#SBATCH --mail-type=ALL
#SBATCH --output=%j_out

# job resources

#SBATCH --time=5:00:00
#SBATCH --partition=batch
#SBATCH --ntasks=1
#SBATCH --ntasks-per-node=1
#SBATCH --cpus-per-task=1
#SBATCH --mem-per-cpu=1000


module load Stack/2.41-foss-2018b

cd stacks

j=1
for i in ./samples/*.fq.gz
        do
                ustacks -f "$i" -o ./result -i "$j"  -m 3 -M 2
                ((j++))
        done
```

Annexe 5 : Script R réalisant les divers graphe

```r
# Installation des packages
install.packages("stringi")  # Pour installer adegenet
install.packages("adegenet")  # Pour manipuler des objets genlight
install.packages("vcfR")      # Pour lire des fichiers VCF
install.packages("FactoMineR")  # Pour l'ACP
install.packages("factoextra")  # Pour visualiser l'ACP
install.packages("Rtsne")       # Pour t-SNE
install.packages("umap")        # Pour UMAP

# Chargement des bibliothÃ¨ques
library(adegenet)
library(vcfR)
library(FactoMineR)
library(factoextra)
library(Rtsne)
library(umap)

# RÃ©pertoire de travail
setwd("/data/stage/")

# Chargement des donnÃ©es VCF
datavcf <- read.vcfR('populations.snps.vcf')

# Conversion du fichier VCF en format genlight
datagen <- vcfR2genlight(datavcf)

# Remplacement des valeurs manquantes (NA) par la moyenne
datagen.na <- tab(datagen, NA.method = "mean", freq = TRUE)
datagen <- as(datagen.na, "genlight")

# Lecture de la table de population
popmap <- read.table("/data/stage/popmapsp")

# Exclusion du control
datagen <- datagen[indNames(datagen) != "FGXCONTROL"]

# Association des noms de population aux individus
datagen_pop <- datagen
datagen_pop$ind.names <- popmap$V2

# Calcul de l'ACP avec standardisation
res.acp_pop <- PCA(datagen_pop, scale.unit = TRUE, graph = FALSE)

# Visualisation de l'ACP
fviz_pca_ind(res.acp_pop,
             label = 'none',
             habillage = as.factor(datagen_pop$ind.names),
             repel = TRUE
)

```

```r
# Analyse de clustering
res.hcpc <- HCPC(res.acp_pop, graph = FALSE)
fviz_dend(res.hcpc)
fviz_cluster(res.hcpc, ellipse.type = "convex", geom = "point")

# t-SNE
# PrÃ©paration des donnÃ©es
datagen_matrix <- as.matrix(datagen.na)
set.seed(123)
tsne_result <- Rtsne(datagen_matrix, dims = 2, perplexity = 30, verbose = TRUE, max_iter = 1000)

# Visualisation du t-SNE
plot(tsne_result$Y, col = as.factor(datagen_pop$ind.names), pch = 19,
     xlab = "Dimension 1", ylab = "Dimension 2",
     main = "t-SNE")
legend("topright", legend = levels(as.factor(datagen_pop$ind.names)),
       col = 1:length(levels(as.factor(datagen_pop$ind.names))), pch = 19)

# UMAP
# PrÃ©paration des paramÃ¨tres UMAP
umap_config <- umap.defaults
umap_config$n_neighbors <- 15
umap_config$min_dist <- 0.1

# Application de UMAP
set.seed(123)
umap_result <- umap(datagen_matrix, config = umap_config)

# Visualisation de UMAP
plot(umap_result$layout, col = as.factor(datagen_pop$ind.names), pch = 19,
     xlab = "UMAP Dimension 1", ylab = "UMAP Dimension 2",
     main = "UMAP")
legend("topright", legend = levels(as.factor(datagen_pop$ind.names)),
       col = 1:length(levels(as.factor(datagen_pop$ind.names))), pch = 19)
```

Annexe 6 : Calcul du K Optimal

K		L(K)	L(K)	L'(K)	L''(K)	|L''(K)|		
	4	-0,131475699	-0,1314757	-0,131475699	0,13474753	0,13474753		
	5	-0,128203869	-0,1282039	0,003271831	0,002877216	0,002877216		
	6	-0,127809254	-0,1278093	0,000394614	0,000251165	0,000251165		
	7	-0,127665805	-0,1276658	0,000143449	0,05720576	0,05720576		
	8	-0,184728116	-0,1847281	-0,057062311	0,114477743	0,114477743		
	9	-0,127312684	-0,1273127	0,057415432	0,057383453	0,057383453		
	10	-0,127280705	-0,1272807	3,19791E-05	3,19791E-05	3,19791E-05		
							écart à la moy2	
K4		-0,131469085	-0,1314691	-0,131469085	3,50038E-05	3,50038E-05	4,37509E-11	
		-0,131504089	-0,1315041	-0,131504089	2,2648E-06	2,2648E-06	8,05955E-10	
		-0,131501824	-0,1315018	-0,131501824	3,27461E-05	3,27461E-05	6,82492E-10	
		-0,131469078	-0,1314691	-0,131469078	6,5427E-06	6,5427E-06	4,38449E-11	
		-0,131475621	-0,1314756	-0,131475621	7,334E-07	7,334E-07	6,21653E-15	
		-0,131474887	-0,1314749	-0,131474887	2,1617E-06	2,1617E-06	6,59742E-13	
		-0,131477049	-0,131477	-0,131477049	8,9133E-06	8,9133E-06	1,82103E-12	
		-0,131468136	-0,1314681	-0,131468136	7,6019E-06	7,6019E-06	5,72118E-11	
		-0,131475737	-0,1314757	-0,131475737	2,4389E-06	2,4389E-06	1,44818E-15	
		-0,131478176	-0,1314782	-0,131478176	3,4065E-06	3,4065E-06	6,13531E-12	
		-0,13147477	-0,1314748	-0,13147477	3,8712E-06	3,8712E-06	8,64054E-13	
		-0,131470899	-0,1314709	-0,131470899	6,1098E-06	6,1098E-06	2,30472E-11	
		-0,131477008	-0,131477	-0,131477008	9,2483E-06	9,2483E-06	1,71362E-12	
		-0,13146776	-0,1314678	-0,13146776	2,0191E-06	2,0191E-06	6,30316E-11	
		-0,131469779	-0,1314698	-0,131469779	1,2105E-06	1,2105E-06	3,50481E-11	
		-0,13147099	-0,131471	-0,13147099	4,841E-07	4,841E-07	2,21808E-11	
		-0,131471474	-0,1314715	-0,131471474	3,4553E-06	3,4553E-06	1,78552E-11	
		-0,131468019	-0,131468	-0,131468019	6,9792E-06	6,9792E-06	5,89954E-11	
		-0,131474998	-0,131475	-0,131474998	3,875E-07	3,875E-07	4,92306E-13	
		-0,13147461	-0,1314746	-0,13147461			1,18624E-12	
MOY		-0,131475699	-0,1314757	-0,131475699	7,13569E-06	7,13569E-06	1,86629E-09	9,33147E-11

K5	-0,130069264	-0,1300693	-0,130069264	0,132285563	0,132285563	3,4797E-06		
	-0,127852965	-0,127853	0,002216299	0,002218173	0,002218173	1,23133E-07		
	-0,127854839	-0,1278548	-1,8739E-06	4,1397E-06	4,1397E-06	1,21822E-07		
	-0,127860853	-0,1278609	-6,0136E-06	0,001334481	0,001334481	1,1766E-07		
	-0,129201347	-0,1292013	-0,001340495	0,002685741	0,002685741	9,94964E-07		
	-0,127856101	-0,1278561	0,001345246	0,001355411	0,001355411	1,20942E-07		
	-0,127866267	-0,1278663	-1,01653E-05	1,96291E-05	1,96291E-05	1,13975E-07		
	-0,127856803	-0,1278568	9,4638E-06	1,50348E-05	1,50348E-05	1,20455E-07		
	-0,127862374	-0,1278624	-5,571E-06	1,69502E-05	1,69502E-05	1,16619E-07		
	-0,127850995	-0,127851	1,13792E-05	1,51582E-05	1,51582E-05	1,2452E-07		
	-0,127854774	-0,1278548	-3,779E-06	6,783E-05	6,783E-05	1,21867E-07		
	-0,127926383	-0,1279264	-7,1609E-05	0,001084802	0,001084802	7,69985E-08		
	-0,129082794	-0,1290828	-0,001156411	0,000389277	0,000389277	7,7251E-07		
	-0,129849928	-0,1298499	-0,000767134	0,002765273	0,002765273	2,70951E-06		
	-0,127851788	-0,1278518	0,00199814	0,002004719	0,002004719	1,23961E-07		
	-0,127858367	-0,1278584	-6,5792E-06	3,57783E-05	3,57783E-05	1,19371E-07		
	-0,127900725	-0,1279007	-4,23575E-05	8,3598E-05	8,3598E-05	9,18962E-08		
	-0,127859484	-0,1278595	4,12405E-05	8,86299E-05	8,86299E-05	1,18601E-07		
	-0,127906874	-0,1279069	-4,73894E-05	9,98139E-05	9,98139E-05	8,8206E-08		
	-0,127854449	-0,1278544	5,24245E-05			1,22094E-07		
MOY	-0,128203869	-0,1282039	-0,006392722	0,007714211	0,007714211	9,7788E-06	4,8894E-07	0
K6	-0,127803478	-0,1278035	-0,127803478	0,127047587	0,127047587	3,33689E-11		
	-0,128559369	-0,1285594	-0,000755891	0,001626835	0,001626835	5,62672E-07		
	-0,127688426	-0,1276884	0,000870943	0,000765744	0,000765744	1,45995E-08		
	-0,127583226	-0,1275832	0,0001052	5,66696E-05	5,66696E-05	5,10887E-08		
	-0,127421357	-0,1274214	0,000161869	0,000156727	0,000156727	1,50464E-07		
	-0,127416215	-0,1274162	5,1425E-06	0,000164615	0,000164615	1,5448E-07		
	-0,127575688	-0,1275757	-0,000159473	5,94604E-05	5,94604E-05	5,45535E-08		
	-0,1276757	-0,1276757	-0,000100012	0,000146428	0,000146428	1,78368E-08		
	-0,127629284	-0,1276293	4,64156E-05	9,3493E-06	9,3493E-06	3,23893E-08		
	-0,127573519	-0,1275735	5,57649E-05	0,001470727	0,001470727	5,5571E-08		
	-0,128988482	-0,1289885	-0,001414963	0,002830708	0,002830708	1,39058E-06		
	-0,127572736	-0,1275727	0,001415746	0,001587548	0,001587548	5,59408E-08		
	-0,127744538	-0,1277445	-0,000171802	0,000647282	0,000647282	4,18817E-09		
	-0,128563623	-0,1285636	-0,000819085	0,001809838	0,001809838	5,69072E-07		
	-0,127572869	-0,1275729	0,000990753	0,002164794	0,002164794	5,58779E-08		
	-0,12874691	-0,1287469	-0,00117404	0,002356589	0,002356589	8,79197E-07		
	-0,127564361	-0,1275644	0,001182549	0,001032722	0,001032722	5,99731E-08		
	-0,127414534	-0,1274145	0,000149827	0,0002527	0,0002527	1,55804E-07		
	-0,127517408	-0,1275174	-0,000102874	4,69147E-05	4,69147E-05	8,51746E-08		
	-0,127573367	-0,1275734	-5,59591E-05			5,5643E-08		
MOY	-0,127809254	-0,1278093	-0,006378668	0,007591223	0,007591223	4,40514E-06	2,20257E-07	0

K7	-0,127610249	-0,1276102	-0,127610249	0,128124661	0,128124661	3,08645E-09		
	-0,127095838	-0,1270958	0,000514411	0,001135636	0,001135636	3,24862E-07		
	-0,127717063	-0,1277171	-0,000621225	0,000988923	0,000988923	2,62737E-09		
	-0,127349365	-0,1273494	0,000367698	9,84409E-05	9,84409E-05	1,00134E-07		
	-0,127080108	-0,1270801	0,000269257	0,001781713	0,001781713	3,43041E-07		
	-0,128592564	-0,1285926	-0,001512456	0,000439843	0,000439843	8,58881E-07		
	-0,129665177	-0,1296652	-0,001072613	0,003487106	0,003487106	3,99749E-06		
	-0,127250684	-0,1272507	0,002414493	0,003035906	0,003035906	1,72326E-07		
	-0,127872097	-0,1278721	-0,000621413	0,001043736	0,001043736	4,25564E-08		
	-0,127449775	-0,1274498	0,000422322	0,00068365	0,00068365	4,66692E-08		
	-0,127711103	-0,1277111	-0,000261328	0,00050679	0,00050679	2,05186E-09		
	-0,12746564	-0,1274656	0,000245462	0,000290611	0,000290611	4,0066E-08		
	-0,127510789	-0,1275108	-4,51489E-05	0,000300153	0,000300153	2,403E-08		
	-0,127255785	-0,1272558	0,000255004	0,000449149	0,000449149	1,68117E-07		
	-0,12744993	-0,1274499	-0,000194145	6,69651E-05	6,69651E-05	4,66023E-08		
	-0,12771104	-0,127711	-0,00026111	0,000260934	0,000260934	2,04616E-09		
	-0,127711216	-0,1277112	-1,762E-07	0,00010254	0,00010254	2,06214E-09		
	-0,127608853	-0,1276089	0,000102363	5,6853E-05	5,6853E-05	3,24362E-09		
	-0,127563342	-0,1275633	4,55105E-05	0,000127656	0,000127656	1,04987E-08		
	-0,127645487	-0,1276455	-8,21453E-05			4,12819E-10		
MOY	-0,127665805	-0,1276658	-0,006382274	0,00752533	0,00752533	6,1908E-06	3,0954E-07	0
K8	-0,127269998	-0,12727	-0,127269998	0,127259518	0,127259518	0,003301435		
	-0,127280477	-0,1272805	-1,04794E-05	0,000353997	0,000353997	0,003300231		
	-0,127644953	-0,127645	-0,000364476	0,000719875	0,000719875	0,003258487		
	-0,127289555	-0,1272896	0,000355399	1,14611118	1,14611118	0,003299188		
	-1,273045336	-1,2730453	-1,145755781	2,291302614	2,291302614	1,184434371		
	-0,127498503	-0,1274985	1,145546833	1,145333822	1,145333822	0,003275229		
	-0,127285492	-0,1272855	0,000213011	0,000560068	0,000560068	0,003299655		
	-0,127632549	-0,1276325	-0,000347057	0,000844254	0,000844254	0,003259904		
	-0,127135352	-0,1271354	0,000497197	0,000666531	0,000666531	0,003316927		
	-0,127304685	-0,1273047	-0,000169334	0,000199888	0,000199888	0,00329745		
	-0,127274131	-0,1272741	3,05544E-05	0,000131677	0,000131677	0,00330096		
	-0,1271119	-0,1271119	0,000162231	0,000325467	0,000325467	0,003319628		
	-0,127275135	-0,1272751	-0,000163236	0,000326302	0,000326302	0,003300845		
	-0,127112069	-0,1271121	0,000163066	0,000698626	0,000698626	0,003319609		
	-0,12764763	-0,1276476	-0,00053556	0,000820136	0,000820136	0,003258182		
	-0,127363054	-0,1273631	0,000284576	0,001017312	0,001017312	0,00329075		
	-0,12809579	-0,1280958	-0,000732736	0,001554725	0,001554725	0,00320722		
	-0,127273802	-0,1272738	0,000821988	0,000945726	0,000945726	0,003300998		
	-0,12739754	-0,1273975	-0,000123737	0,001103098	0,001103098	0,003286795		
	-0,128624375	-0,1286244	-0,001226836			0,00314763		
MOY	-0,184728116	-0,1847281	-0,006431219	0,248435517	0,248435517	1,246775496	0,062338775	0

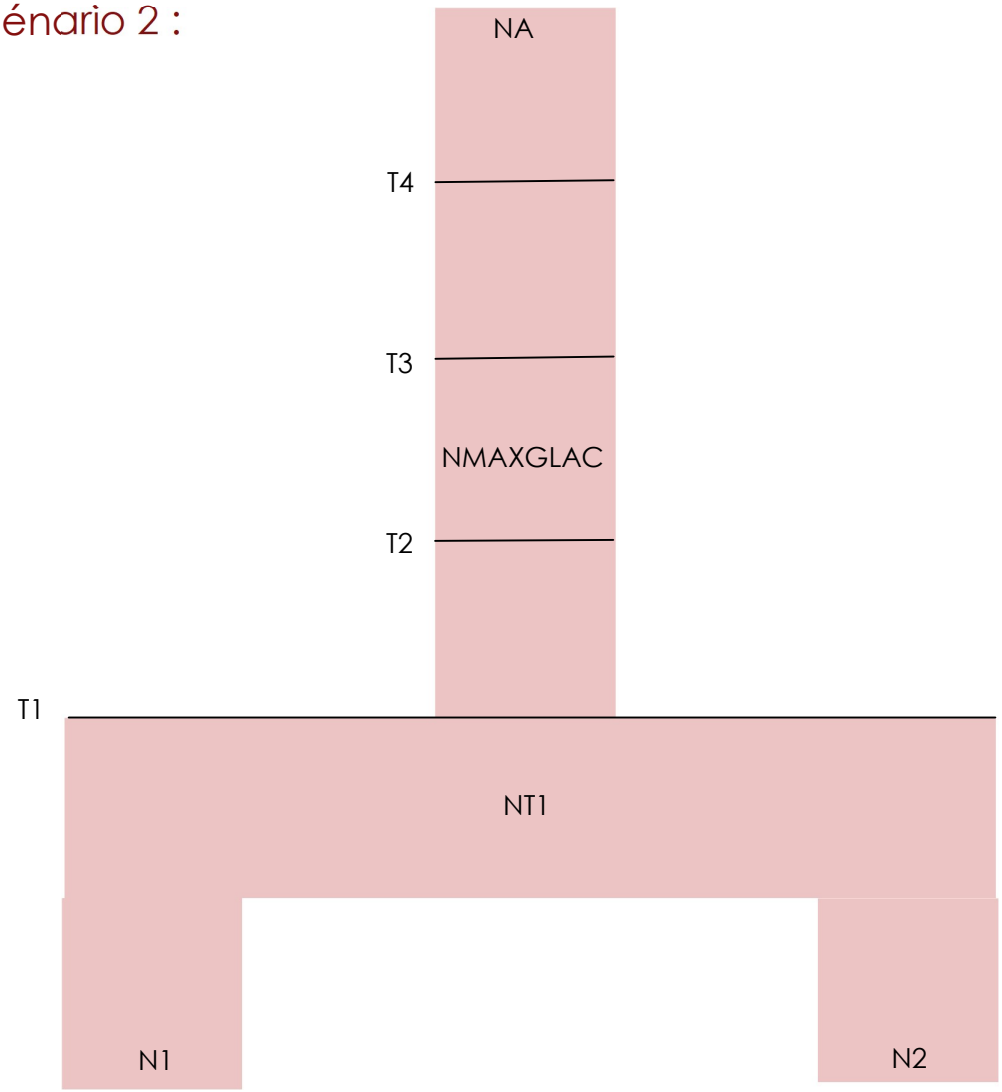
K9	-0,127312214	-0,1273122	-0,127312214	0,127483408	0,127483408	2,2105E-13		
	-0,127141019	-0,127141	0,000171194	0,000172259	0,000172259	2,94687E-08		
	-0,127142084	-0,1271421	-1,0643E-06	0,000518253	0,000518253	2,91044E-08		
	-0,127661401	-0,1276614	-0,000519318	0,000789165	0,000789165	1,21604E-07		
	-0,127391554	-0,1273916	0,000269848	0,000188042	0,000188042	6,22045E-09		
	-0,127309749	-0,1273097	8,18052E-05	0,000188656	0,000188656	8,61634E-12		
	-0,1274166	-0,1274166	-0,000106851	0,00019315	0,00019315	1,07985E-08		
	-0,127330301	-0,1273303	8,62986E-05	0,000178864	0,000178864	3,10364E-10		
	-0,127422867	-0,1274229	-9,25657E-05	0,00038091	0,00038091	1,21403E-08		
	-0,127134522	-0,1271345	0,000288345	0,000618328	0,000618328	3,17416E-08		
	-0,127464505	-0,1274645	-0,000329983	0,000257526	0,000257526	2,30498E-08		
	-0,127536963	-0,127537	-7,24574E-05	0,000207865	0,000207865	5,0301E-08		
	-0,127401555	-0,1274016	0,000135408	0,000132288	0,000132288	7,89804E-09		
	-0,127133859	-0,1271339	0,000267696	0,000293054	0,000293054	3,19785E-08		
	-0,127159216	-0,1271592	-2,53576E-05	0,000142142	0,000142142	2,35524E-08		
	-0,127326716	-0,1273267	-0,000167499	0,000183522	0,000183522	1,96887E-10		
	-0,127310693	-0,1273107	1,60223E-05	0,000134761	0,000134761	3,96273E-12		
	-0,12715991	-0,1271599	0,000150783	0,000355389	0,000355389	2,33398E-08		
	-0,127364517	-0,1273645	-0,000204606	0,000435689	0,000435689	2,68664E-09		
	-0,127133434	-0,1271334	0,000231082			3,21304E-08		
MOY	-0,127312684	-0,1273127	-0,006356672	0,006992277	0,006992277	4,36534E-07	2,18267E-08	0
K10	-0,127346166	-0,1273462	-0,127346166	0,127272417	0,127272417	4,28511E-09		
	-0,127419914	-0,1274199	-7,37485E-05	0,00016765	0,00016765	1,93792E-08		
	-0,127326013	-0,127326	9,39015E-05	0,000188118	0,000188118	2,05279E-09		
	-0,127420229	-0,1274202	-9,42165E-05	0,000161886	0,000161886	1,9467E-08		
	-0,12735256	-0,1273526	6,76695E-05	0,000155307	0,000155307	5,1631E-09		
	-0,127440197	-0,1274402	-8,76377E-05	0,000377177	0,000377177	2,54378E-08		
	-0,127150658	-0,1271507	0,000289539	0,000307404	0,000307404	1,69122E-08		
	-0,127168522	-0,1271685	-1,78642E-05	2,42454E-05	2,42454E-05	1,2585E-08		
	-0,127162141	-0,1271621	6,3812E-06	2,20625E-05	2,20625E-05	1,40574E-08		
	-0,127177822	-0,1271778	-1,56813E-05	2,22724E-05	2,22724E-05	1,05849E-08		
	-0,127171231	-0,1271712	6,5911E-06	0,000173444	0,000173444	1,19845E-08		
	-0,127338083	-0,1273381	-0,000166852	0,000105769	0,000105769	3,2923E-09		
	-0,127399167	-0,1273992	-6,10838E-05	0,000280302	0,000280302	1,40333E-08		
	-0,127179949	-0,1271799	0,000219218	0,000399987	0,000399987	1,01517E-08		
	-0,127360719	-0,1273607	-0,00018077	0,000194839	0,000194839	6,40228E-09		
	-0,12734665	-0,1273467	1,4069E-05	0,000173303	0,000173303	4,34877E-09		
	-0,127159278	-0,1271593	0,000187372	0,000211792	0,000211792	1,47446E-08		
	-0,127183697	-0,1271837	-2,44194E-05	0,000139283	0,000139283	9,41051E-09		
	-0,127347399	-0,1273474	-0,000163702	0,0003474	0,0003474	4,44817E-09		
	-0,127163702	-0,1271637	0,000183698			1,36897E-08		
MOY	-0,127280705	-0,1272807	-0,006358185	0,006880245	0,006880245	2,2243E-07	1,11215E-08	0

Annexe 7 : Script R permettant d'utiliser pophelper

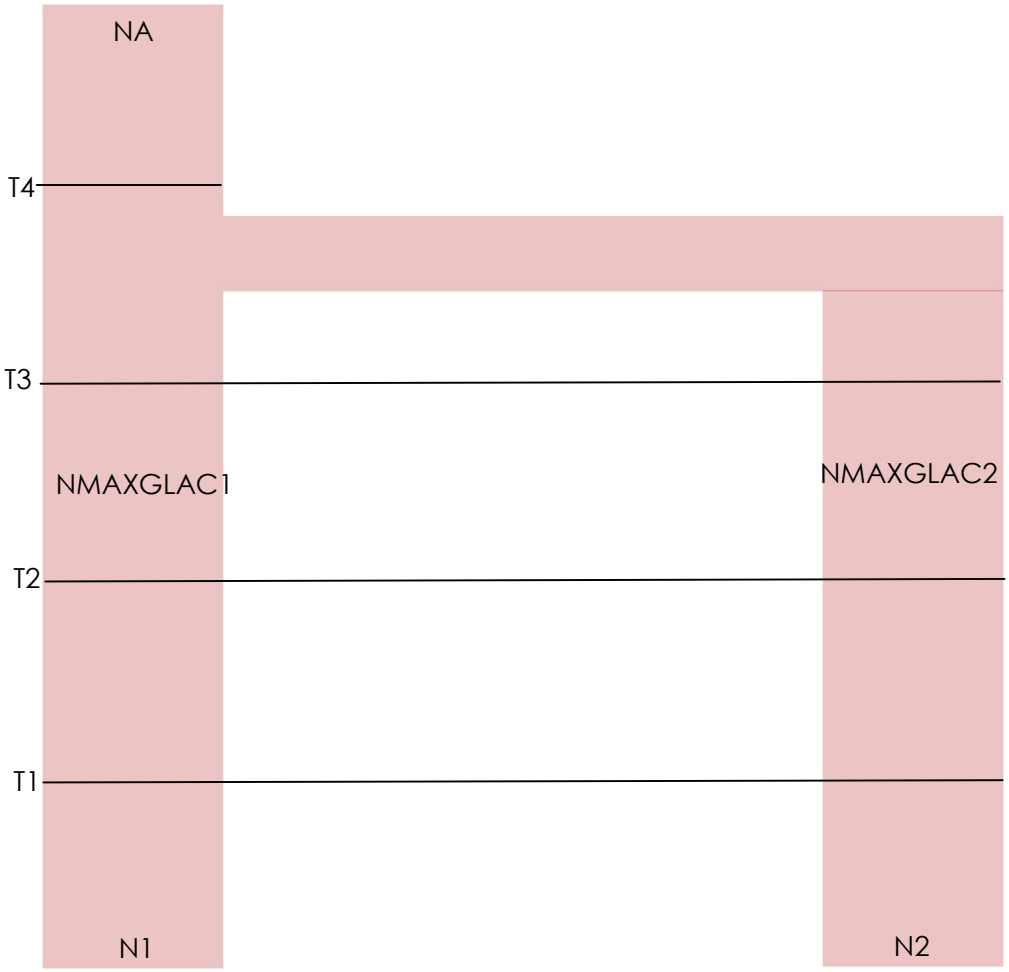
```
#il existe une version en ligne sans passer par R mais elle est
moins complète


Install.packages (c(« ggplot2 », « gridExtra », « label.switching »,
« tidyr », « remotes », « colourpicker », « DT », « highcharter », «
htmlwidgets », « magrittr », « markdown », « RColorBrewer », « shiny
», « shinnyAce », « shinyBS », « shinythemes », « shinyWidgets », «
viridislite », « writexl »), repos= http://cran.us.r-project.org)

Remotes :: install_github('royfrancis/pophelper')

Remotes :: install_github('royfrancis/pophelperShiny')

Library(pophelperShiny)

runPophelper()


#Copier l'adresse d'écoute donné par R dans un navigateur internet
pour avoir accès à l'interface graphique.
```

Annexe 8 : Diagramme de coalescence

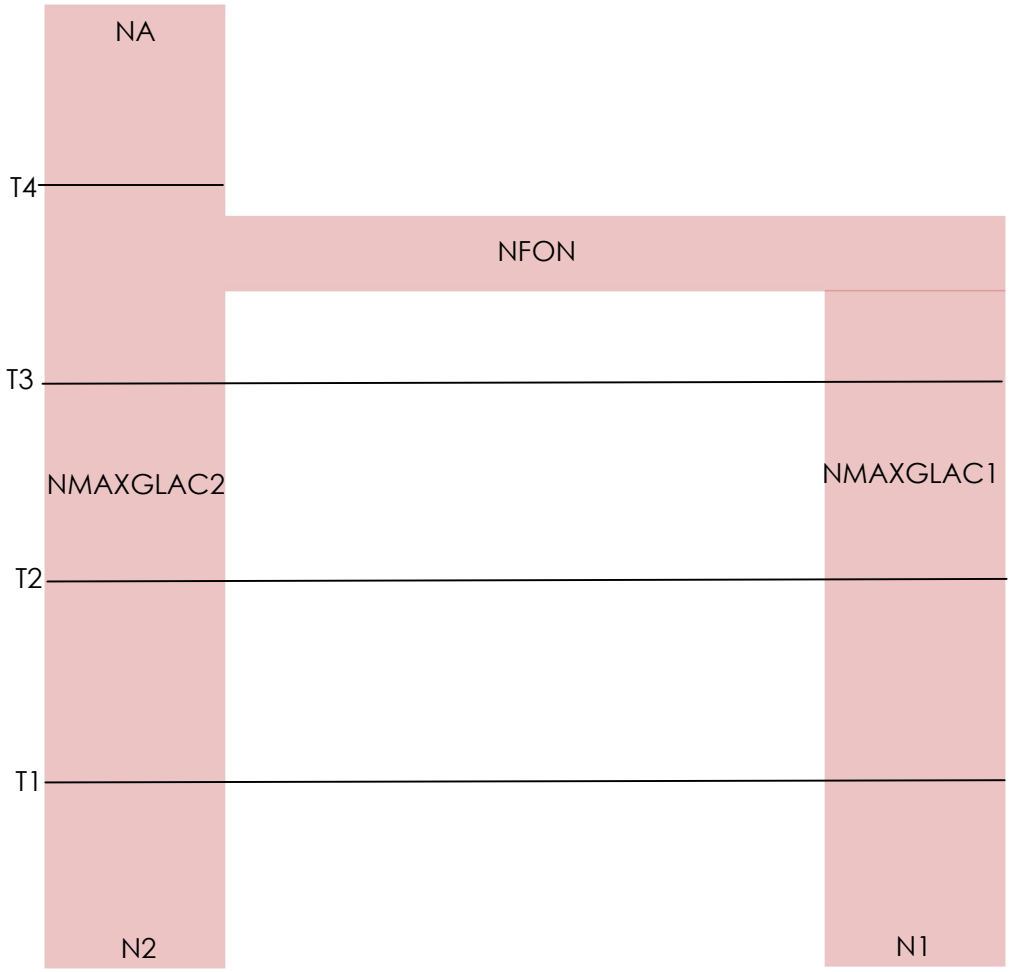
Scénario 2 :



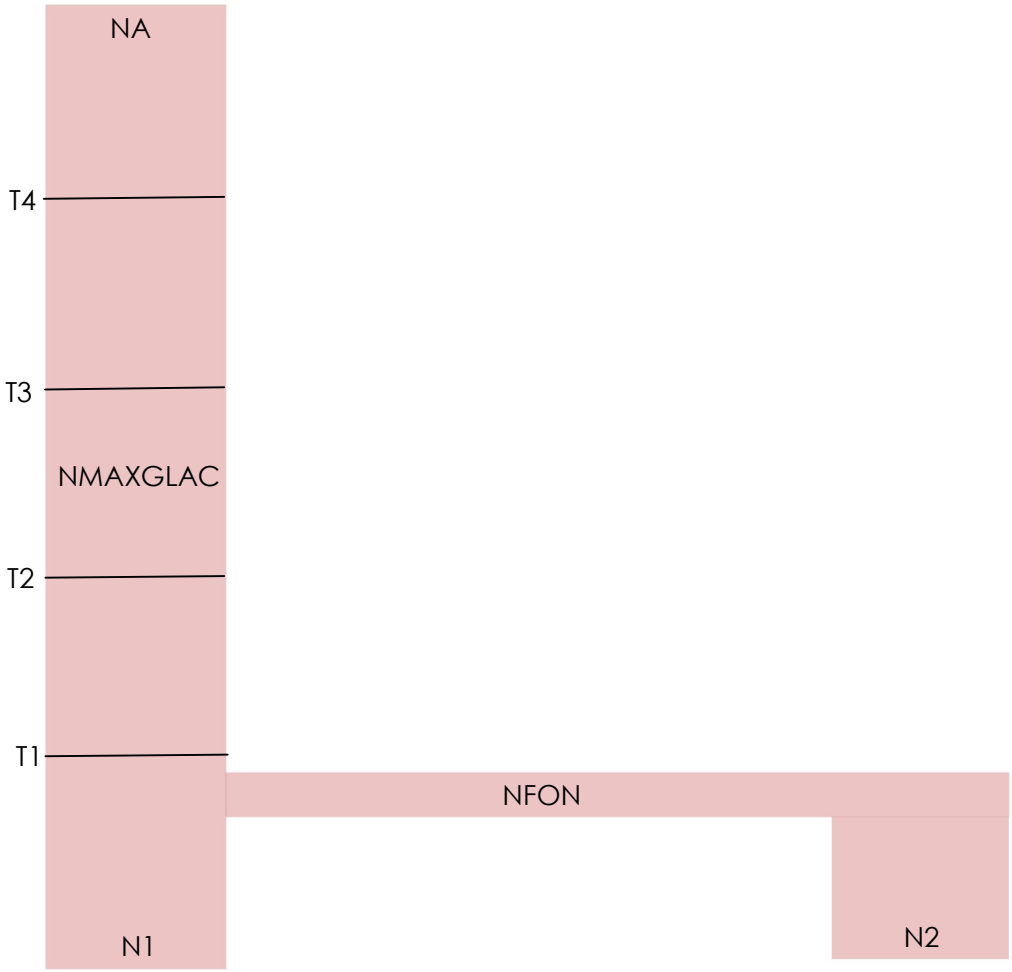
Scénario 3 :



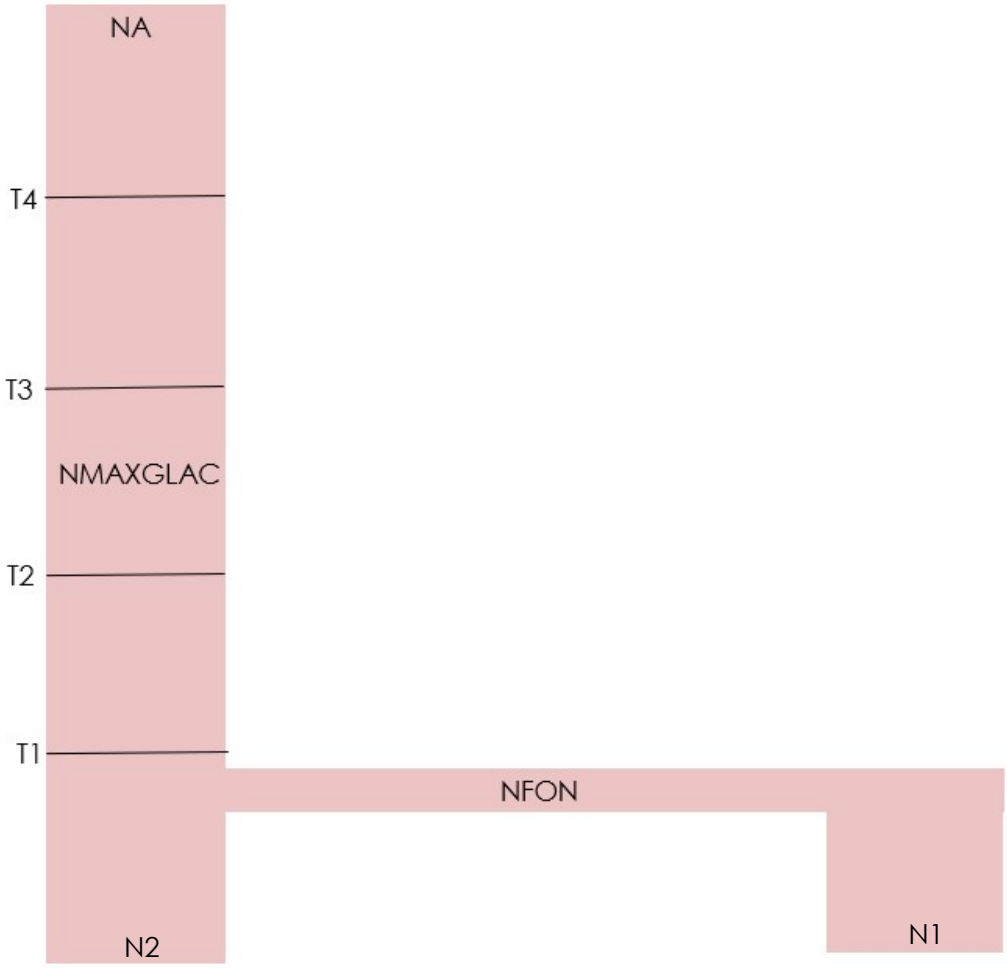
Scénario 4 :



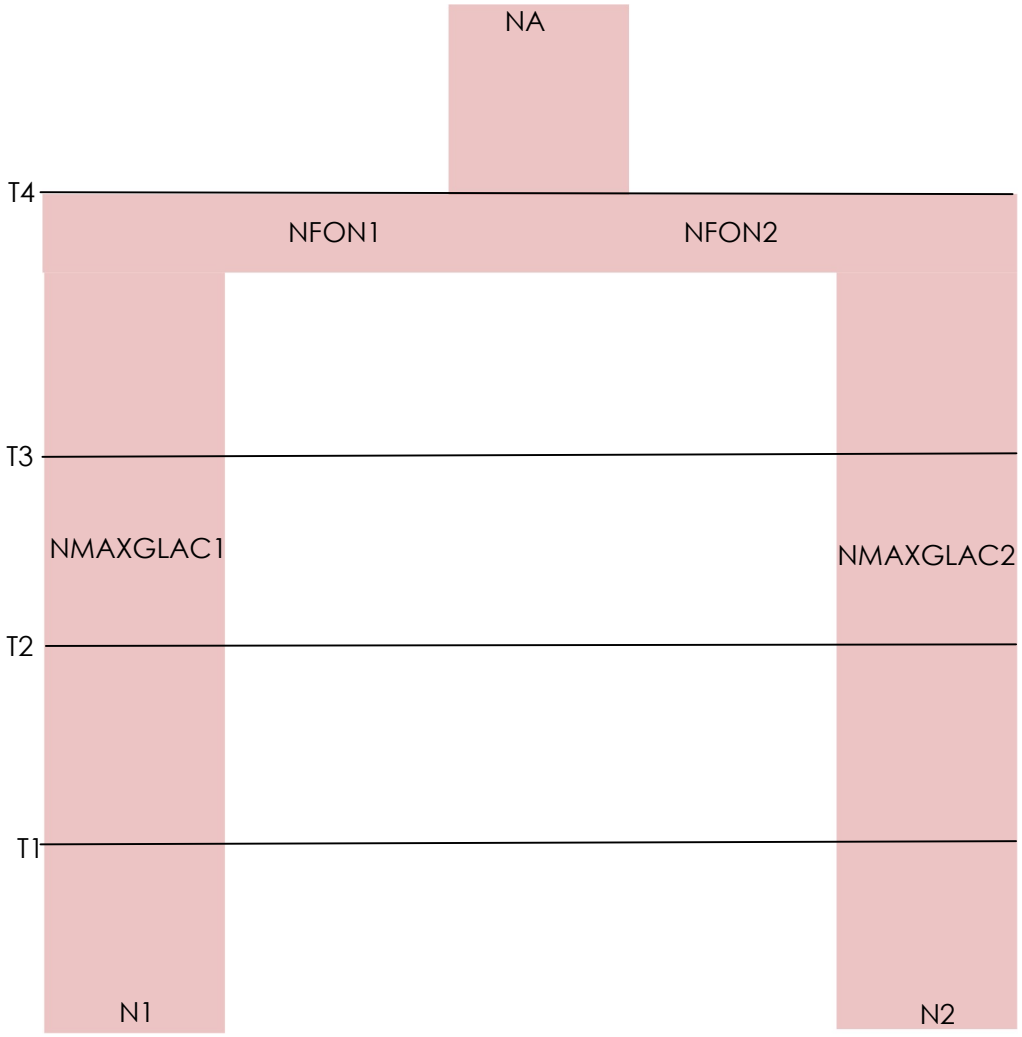
Scénario 5 :



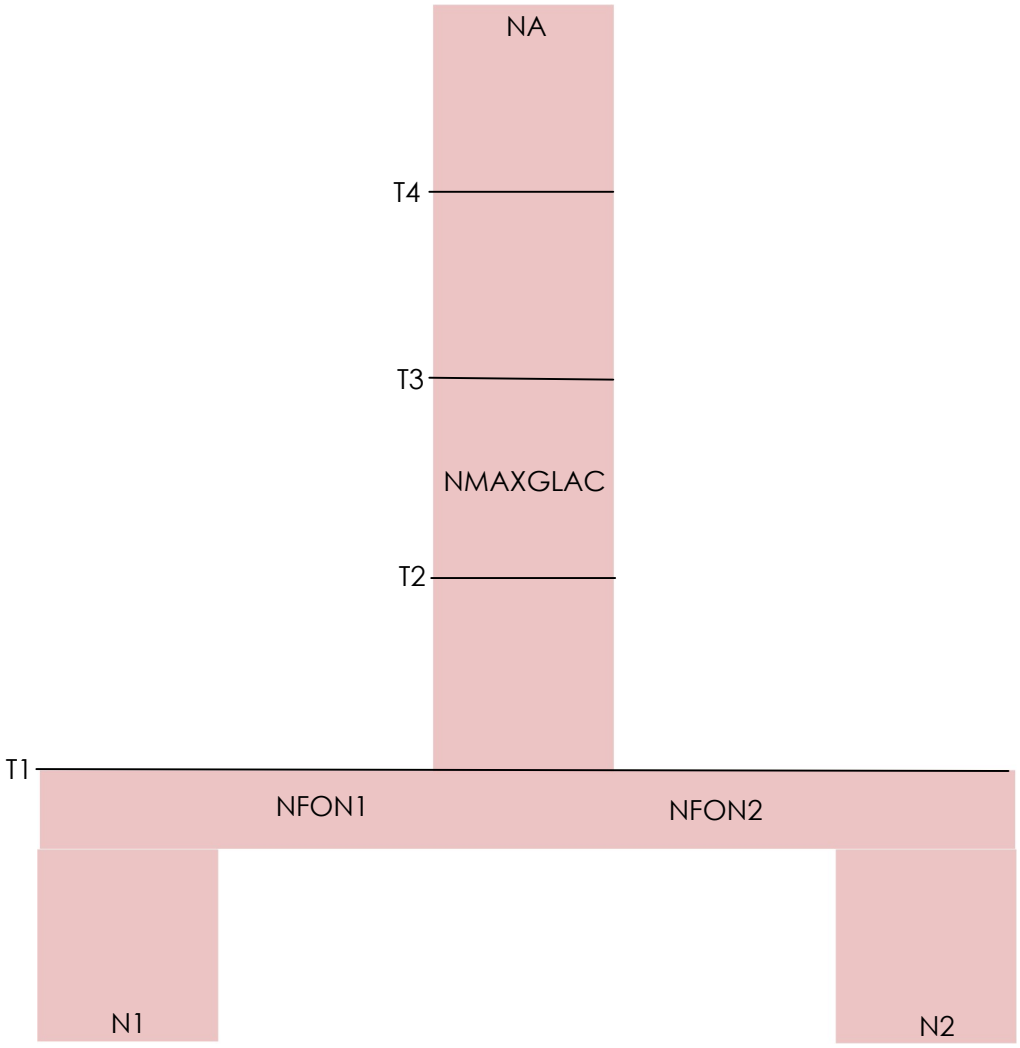
Scénario 6 :



Scénario 7 :



Scénario 8 :



Annexe 9 : utilisation de Fastsimcoal

Script d'exécution :

```
echo "=== VARIABLES... ==="

PREFIXES=scenario1


echo "=== FASTSIMCOAL ==="

/opt/fsc27_linux64/fsc27093 -t ${PREFIXES}.tpl -e ${PREFIXES}.est -
m -0 -n 10 -L 40 -s 0 -M -c 8


echo "=== DONE. ==="
```

Ficher de paramètre :

```
//Parameters for the coalescence simulation program :
fastsimcoal.exe

2 samples to simulate :

//Population effective sizes (number of genes)

17705

28590

//Samples sizes and samples age

98

136

//Growth rates : negative growth implies population expansion

0

0

//Number of migration matrices : 0 implies no migration between
demes

2

//Migration matrix 0

0.0000 2.660657511568891e-04

2.660657511568891e-04 0.0000

//Migration matrix 1

0 0

0 0
```

```
//historical event: time, source, sink, migrants, new deme size,
growth rate, migr mat index

5 historical event

4193 1 0 1 46295 0 1

11000 0 0 0 368579 0 1

18000 0 0 0 12989 0 1

23000 0 0 0 368579 0 1

115000 0 0 0 18162 0 1

//Number of independent loci [chromosome]

1

//Per chromosome: Number of contiguous linkage Block: a block is a
set of contiguous loci

1

//per Block:data type, number of loci, per gen recomb and mut rates

FREQ 1 0 1e-8

// Priors and rules file

// **********************


[PARAMETERS]

//#isInt? #name #dist.#min #max

//all N are in number of haploid individuals

1     NDEBGLAC   logunif    1000  1000000    output bounded

1     NFINLGM    logunif 100     10000000   output bounded

1     NT    logunif    100   10000000   output bounded

1     N1    logunif    1000  1000000 output bounded

1     N2    logunif    1000  1000000 output bounded

0     MIG   unif  0     1e-3  output bounded

1     TDIV  unif  1000  11000 output bounded


[COMPLEX PARAMETERS]

1     NBEFOREDIV = N1+N2    output
```