Guide méthodologique pour la réalisation d'une expérience de métabarcoding ADN à partir d'échantillons de lait ou de fromages à destination des experts et laboratoires

Eric Dugat-Bony, Blandine Polturat, Céline Delbès, Christine Achilleos, Cresciense Lecaudé, Hélène Tormo, Marion Dalmasso, Nicolas Orieux, Sarah Chuzeville, Sébastien Theil, et al.

ADAMOS

Un nouveau regard sur les écosystèmes laitiers et fromagers : adaptation, développement et appropriation des méthodes omiques à des fins d'écologie microbienne (2019-2023)

GUIDE METHODOLOGIQUE

pour la réalisation d'une expérience de métabarcoding ADN à partir d'échantillons de laits ou de fromages
à destination des experts et laboratoires

Version 2 - 21 novembre 2024

Introduction

La gestion des écosystèmes microbiens laitiers et fromagers participe largement à l'élaboration des caractéristiques organoleptiques finales des fromages, produits par les filières fromagères valorisant leur terroir. Avec le développement des outils d'analyse haut débit, de nouvelles méthodes, dites « omiques » (pour métagénomique, transcriptomique, métabolomique…) sont ou vont devenir accessibles. Parmi elles, le séquençage massif d'amplicons, également appelée métabarcoding ADN, permet de décrire la composition des communautés microbiennes présentes dans les échantillons analysés.

Le principe consiste à extraire l'ADN d'un échantillon puis à amplifier par PCR un fragment cible à l'aide d'un couple d'amorces prédéfini. Ces produits PCR, après ajouts de barcodes (oligonucléotides uniques pour chaque échantillon) et d'adaptateurs de séquençage, sont ensuite séquencés tous en même temps. Après le séquençage, les séquences sont triées par échantillon grâce aux barcodes puis assignées à des taxons par comparaison avec des séquences de référence.

Beaucoup de méthodes et outils d'analyse ont été développés pour obtenir une vision la plus précise possible des écosystèmes étudiés. Les techniques de préparation puis d'analyse des échantillons dépendent de l'écosystème, des questions auxquelles on souhaite répondre et de la technologie de séquençage utilisée. Nous proposons dans ce guide des conseils pratiques tant sur le plan de la préparation des échantillons que sur l'analyse bioinformatique des données.

Ce document s'appuie en particulier sur une publication de référence concernant les méthodes d'extraction d'ADN à partir d'échantillons de laits et de fromages (Quigley *et al.,* 2012), et d'un guide pratique sur l'analyse des données de ce type (Falentin *et al.*, 2019). Il a été élaboré dans le cadre de l'action 1 du projet ADAMOS (CASDAR 2019-2023) et vise à constituer un point d'entrée pour les professionnels souhaitant entreprendre des expériences de métabarcoding ADN.

Pour être sûr de consulter la dernière version de ce guide, rendez-vous sur Fromic, dans la partie "Appliquer" :

https://tinyurl.com/fromic-guide-methodologique

FROMIC

Pour en savoir plus, vous pouvez consulter la partie "Appliquer les méthodes omiques à des cas concrets" de l'outil Fromic :

3.D Choisir la méthodologie de prélèvements
https://tinyurl.com/prelevement1

5. Réaliser les prélèvements :
https://tinyurl.com/prelevement2

Partie 1 · Prélèvement et conservation des échantillons

1.1. Échantillons liquide (exemple : lait, lactosérum)

Principe

Il consiste à centrifuger rapidement entre 120 et 160 ml d'échantillon, éliminer le surnageant puis congeler les culots cellulaires à -20°C.

Comment réaliser cette étape ?

D'un point de vue technique, les centrifugations devront être réalisées à 4°C si le matériel à disposition le permet. De même, en fonction du matériel à disposition, les centrifugations seront réalisées soit pendant 30 minutes à 5000 × *g* soit pendant 15 minutes à 9000 × *g*.

La quantité à prélever peut varier selon les échantillons, il est important d'échanger à ce sujet avec le laboratoire qui réalisera les extraction.

En cas de difficulté pour acheminer les échantillons au laboratoire dans un délai raisonnable (inférieur à 24h), une alternative à cette procédure consiste à congeler directement et le plus rapidement possible la quantité nécessaire pour les analyses et à réaliser la centrifugation dans un second temps, après décongélation. Attention cependant : certains types d'échantillon ne peuvent pas être congelés (à voir avec le laboratoire).

1.2. Echantillons solides (exemples : fromages, caillés)

Principe

La conservation des échantillons se fait par congélation à -20°C.

Comment réaliser cette étape ?

Pour chaque fromage, trois échantillons de 250 mg chacun seront placés dans des tubes eppendorf de 2 ml puis congelés à -20°C avant extraction de l'ADN.

Une alternative à cette procédure consiste à réaliser une séparation de cellules avant la congélation, et sera plutôt indiquée si une extraction d'ADN par un kit commercial est envisagée par la suite. Pour cela, 5 à 10 g de fromage seront dilués au 1/10ème dans une solution stérile de citrate de sodium 20 g/l puis mélangés deux fois 30 secondes au Stomacher. Le mélange obtenu sera ensuite traité comme un échantillon liquide, conformément aux indications présentes dans le paragraphe précédent.

Il est nécessaire de bien réfléchir en amont à la procédure de prélèvement mise en œuvre afin que l'échantillon analysé soit le plus représentatif possible de la partie du fromage qu'il représente (exemple : cœur, surface) et pour limiter les transferts d'un compartiment à l'autre. Il est important de travailler avec du matériel et un support de découpe propres et désinfectés, le mieux étant d'avoir à disposition du matériel stérile. La méthode idéale est cependant difficile à codifier puisqu'elle dépendra de la variété de fromage considérée, notamment de sa géométrie, et du matériel à disposition de la personne réalisant l'échantillonnage.

Partie 2 · Extraction de l'ADN

L'extraction de l'ADN est cruciale pour la réussite de l'expérience de métabarcoding. Elle doit être conduite avec l'objectif de récupérer l'ADN de l'ensemble des microorganismes présents dans l'échantillon, qu'il s'agisse de bactéries (Gram+ et Gram-), de levures ou de champignons filamenteux. Il est important de noter que la fragmentation de l'ADN lors de l'étape d'extraction n'est pas un obstacle à la suite de la procédure puisque les fragments d'ADN à amplifier par PCR sont de taille modeste (<500 pb). Des protocoles incluant des étapes de lyse agressives (par exemple mécaniques) sont donc à privilégier.

Principe

Il consiste à réaliser une lyse des parois cellulaires (une combinaison de lyses chimique, enzymatique et mécanique est recommandée) afin de libérer l'ADN contenu dans les cellules, suivie d'une extraction de l'ADN au phénol/chloroforme et d'une purification sur colonne à l'aide d'un kit commercial. L'ADN purifié doit être conservé à -20°C dans de l'eau de qualité biologie moléculaire, du Tris 10 mM (pH 8,0) ou du TE 1X (Tris 10 mM pH 8,0 EDTA 1 mM).

Comment réaliser cette étape ?

Un exemple de protocole détaillé est disponible en annexe 1. Cette méthode a été testée et validée sur une grande variété de fromages et de laits différents dans le cadre du projet MétaPDOCheese.

Une alternative à cette procédure consiste à utiliser un kit commercial dédié pour réaliser l'ensemble de la procédure d'extraction d'ADN. Cette procédure est plus rapide et plus facile à mettre en œuvre dans la mesure où elle permet d'éviter l'utilisation de composés classés cancérogènes, mutagènes et toxiques pour la reproduction (CMR, dans le cas présent le phénol). Elle évite donc la nécessité de disposer d'une sorbonne et d'une procédure de gestion des déchets chimiques appropriée. Cependant, l'efficacité d'extraction est très variable d'un type de fromage à l'autre et peut s'avérer insuffisante pour les échantillons peu chargés en microorganismes (exemples : laits, caillés, fromages frais). Il est donc recommandé de faire des pré-essais sur des échantillons représentatifs de ceux qui seront analysés dans le projet avant de choisir la méthode d'extraction et le kit utilisé. Un exemple de protocole détaillé est disponible en annexe 2.

Une fois l'ADN extrait, il faut réaliser un dosage pour estimer sa concentration. Les méthodes fluorométriques spécifiques de l'ADN (exemple : Qubit ou équivalent) sont recommandées pour obtenir une mesure précise. Les méthodes spectrophotométriques (exemple : Nanodrop ou équivalent) sont moins précises et ont tendance à surestimer de manière importante les concentrations d'ADN, en particulier du fait qu'elles détectent toutes les molécules absorbant à 260 nm de manière indifférenciée (ADN, ARN, contaminants).

Partie 3 · Amplification par PCR

Principe

L'amplification par PCR permet d'isoler et de multiplier sélectivement une portion d'ADN donnée. Dans le cas du métabarcoding ADN, la portion visée est un marqueur phylogénétique c'est-à-dire une séquence qui possède les caractéristiques suivantes :

1

Elle est **ubiquitaire**, donc présente dans le génome de l'ensemble des espèces procaryotes ou dans l'ensemble des espèces eucaryotes ;

2

Elle possède **des régions suffisamment conservées** pour permettre de définir des amorces dites « universelles » capables de l'amplifier chez toutes les espèces visées ;

3

Elle possède des **régions suffisamment variables** pour permettre de discriminer les différentes espèces présentes sur la base de leurs séquences ;

4

Elle est **suffisamment utilisée** pour que des bases de données publiques importantes soient disponibles pour permettre de réaliser les identifications.

Il n'existe pas de marqueur phylogénétique permettant de détecter simultanément les procaryotes et les eucaryotes. Pour les procaryotes, le gène codant pour l'ARNr 16S, et plus particulièrement la portion de ce gène encadrant les régions V3 et V4, est actuellement le marqueur phylogénétique le plus utilisé pour les expériences de métabarcoding. Concernant les eucaryotes, plusieurs marqueurs phylogénétiques sont actuellement fréquemment utilisés (gène codant pour l'ARNr 28S, espaces intergéniques ITS1 et ITS2, gène RPB2). Pour le cas particulier de l'analyse de produits laitiers, un projet interne mené à INRAE indique que le second espace intergénique (ITS2) est plus performant que les autres pour détecter les espèces caractéristiques de ce biotope.

Comment réaliser cette étape ?

Afin d'amplifier par PCR le marqueur choisi (par exemple le marqueur 16S pour les bactéries et l' ITS2 pour les champignons), il est recommandé d'utiliser 10 ng d'ADN comme matrice de départ (dosage fluorométrique) et de réaliser 30 cycles d'amplification. Concernant le choix de l'enzyme permettant de réaliser l'amplification, nous recommandons l'utilisation d'une Taq Polymerase certifiée sans ADN contaminant (exemple : MTP Taq (Sigma ref : D7442)). Un protocole détaillé est disponible en annexe 3.
Une fois l'amplification réalisée et avant d'envoyer le produit PCR à la plateforme de séquençage ou au prestataire, il faut contrôler qu'un amplicon de la taille attendue a été produit grâce à une électrophorèse sur gel d'agarose à 1,5%.

Pour être compatible avec les séquenceurs de nouvelle génération et plus particulièrement l'Illumina MiSeq – le plus utilisé à l'heure actuelle pour cette application – la taille de l'amplicon généré doit être idéalement comprise entre 200 et 500 pb.

Partie 4 · Importance et rôle des contrôles

Principe

Les réactifs utilisés lors de l'extraction d'ADN ou l'amplification par PCR peuvent parfois être source de contamination. Celle-ci est non négligeable en particulier lorsque les échantillons d'ADN sont très faiblement concentrés. Il est donc fortement recommandé d'introduire au sein des plans d'expérience des contrôles négatifs afin de repérer les potentielles espèces contaminantes provenant des réactifs et ainsi pouvoir les éliminer du jeu de données.

Comment réaliser cette étape ?

Le minimum recommandé consiste à réaliser un témoin négatif d'extraction c'est-à-dire une extraction d'ADN à partir d'eau qualité biologie moléculaire portant la mention « DNA-free ». L'ajout de témoins négatifs supplémentaires, introduits aux différentes étapes de la procédure permet de déterminer quelles étapes sont les principales sources de contamination. Il s'agit par exemple du témoin négatif colonne (eau DNA-free passée sur la colonne des kits de purification d'ADN puis amplifiée), ou du témoin négatif PCR (amplification par PCR de l'eau DNA-free).

Que se passe-t-il ensuite ?

Les échantillons et les contrôles sont envoyés à la plateforme de séquençage ou au prestataire. Des barcodes leur seront ajoutés (oligonucléotides uniques pour chaque échantillon) puis ils seront mélangés de façon équimolaire (création d'une librairie) avant d'être séquencés tous ensembles. Les données seront finalement fournies à l'utilisateur sous la forme de fichiers de séquences, généralement au format fastq (voir paragraphe 6.2).

Il est également fortement recommandé d'intégrer dans les plans d'expériences, un ou plusieurs témoins positifs appelés témoins Mocks. Ces témoins consistent en l'amplification par PCR de mélanges d'ADN d'une ou plusieurs communautés reconstituées. La comparaison des résultats obtenus sur ces témoins positifs avec l'attendu permet de valider l'ensemble de la procédure de métabarcoding (amplification, séquençage, analyse bioinformatique).

NB : des témoins Mocks déjà reconstitués sont distribués par certains fournisseurs de réactifs de biologie moléculaire (ex: ZymoBIOMICS Microbial Community DNA Standard).

Exemple en pratique

Dans le cadre du projet Adamos :

- 1 témoin négatif d'extraction pour chaque série d'extraction (1 pour 23 échantillons)
- 1 témoin positif par run (référence : ZymoBIOMICS Microbial Community DNA Standard, ZD6306).

Partie 5 · Choix du pipeline bioinformatique

Principe

Il existe une multitude de suites bioinformatiques - appelés pipelines car elles sont en réalité chacune composée de plusieurs outils organisés en chaîne - permettant de traiter les données de métabarcoding.

Quel pipeline choisir ?

Les pipelines les plus populaires sont actuellement QIIME2 (Bolyen *et al.*, 2019) et Mothur (Schloss, 2020). Cependant, leur utilisation nécessite un certain niveau d'expertise informatique. Des outils tout aussi performants mais plus simples à utiliser pour des non-initiés en bioinformatique sont également disponibles, comme FROGS (Escudié *et al.*, 2017) (http://frogs.toulouse.inra.fr) dont les différentes étapes seront décrites ci-dessous.

Partie 6 · Analyse bioinformatique des données de métabarcoding ADN avec FROGS

1. Accès à Frogs
2. Import des données sur Galaxy
3. Démultiplexage
4. Preprocessing
5. Définitions des unités taxonomiques opérationnelles
6. Elimination des séquences chimériques
7. Filtres sur la base de l'abondance
8. Extraction des séquences ITS
9. Affiliation taxonomique
10. Export des données et contrôle manuel des affiliations

6.1. Accès à Frogs

FROGS est distribué sous un environnement Galaxy (https://usegalaxy.org), soit une interface web qui a pour objectif de rendre la bioinformatique accessible aux utilisateurs n'ayant pas de compétences en programmation informatique. En France, FROGS est disponible sur les instances Galaxy de plusieurs plateformes d'analyses bioinformatiques listées dans le tableau 1.

Plateforme bioinformatique	Lien de connexion
Genotoul (INRAE, Toulouse)	http://bioinfo.genotoul.fr
Migale (INRAE, Jouy-en-Josas)	https://migale.inrae.fr
ABIMS (CNRS, Roscoff)	http://abims.sb-roscoff.fr
Institut Pasteur (Paris)	https://research.pasteur.fr/fr/team/bioinformatics-and-biostatistics-hub
IFB-Core (Orsay)	https://www.france-bioinformatique.fr/cluster-ifb-core
South Green (CIRAD, Montpellier)	https://www.southgreen.fr

Tableau 1. *Liste non exhaustive des plateformes bioinformatiques hébergeant FROGS sur leur instance Galaxy.*

Comment ?
Suite à la création d'un compte sur l'une des plateformes, l'utilisateur a accès à un ensemble d'outils bioinformatiques – dont le pipeline FROGS – ainsi que des ressources de stockage et de calcul lui permettant de réaliser ses analyses.

Une version de FROGS en ligne de commandes est également disponible (https://github.com/geraldinepascal/FROGS) et peut-être installée localement sous Linux.

6.2. Import des données sur Galaxy

La plateforme ou le prestataire de séquençage peut vous envoyer les résultats sous deux formes :

Non démultiplexées

Vous aurez alors deux fichiers de séquences au format fastq contenant les lectures R1 (forward) et R2 (reverse), puis un fichier tabulé contenant les barcodes et noms d'échantillons associés.

Que faire ?

Chargez directement les deux fichiers de séquences au format fastq sur Galaxy dans votre historique de travail à l'aide de l'outil intitulé « Upload Data » (Figure 1), chargez également un fichier tabulé contenant dans la première colonne le nom des échantillons et dans la deuxième les barcodes correspondants (si une stratégie de barcoding en « dual index » a été utilisée, une troisième colonne contenant un second barcode par échantillon doit être ajoutée), puis passez à l'étape de démultiplexage.

Démultiplexées

Vous aurez alors deux fichiers de séquences au format fastq par échantillon contenant les lectures R1 (forward) et R2 (reverse).

Que faire ?

Compressez d'abord avec gzip les fichiers fastq de chacun des échantillons. Ils auront alors l'extension suivante : « .fastq.gz » (attention : certaines plateformes de séquençage envoient les fichiers déjà compressés, dans ce cas vous n'avez pas à le faire). Pour cela, sous Windows vous pouvez utiliser l'outil 7-Zip (https://www.7-zip.fr), et sous Linux vous pouvez utiliser la commande gzip (http://www.linux-france.org/article/memo/node129.html). Ensuite, créez une archive tar contenant tous les fichiers ayant l'extension « .fastq.gz » (attention : l'archive doit contenir directement les fichiers de séquences et non un dossier contenant les fichiers). Pour cela, sous Windows vous pouvez utiliser l'outil 7-Zip (https://www.7-zip.fr/), et sous Linux vous pouvez utiliser la commande tar (https://doc.ubuntu-fr.org/tar). Enfin, chargez l'archive tar sur Galaxy dans votre historique de travail à l'aide de l'outil intitulé « Upload Data » (Figure 1) puis passez directement à l'étape de pre-processing (pas de démultiplexage).

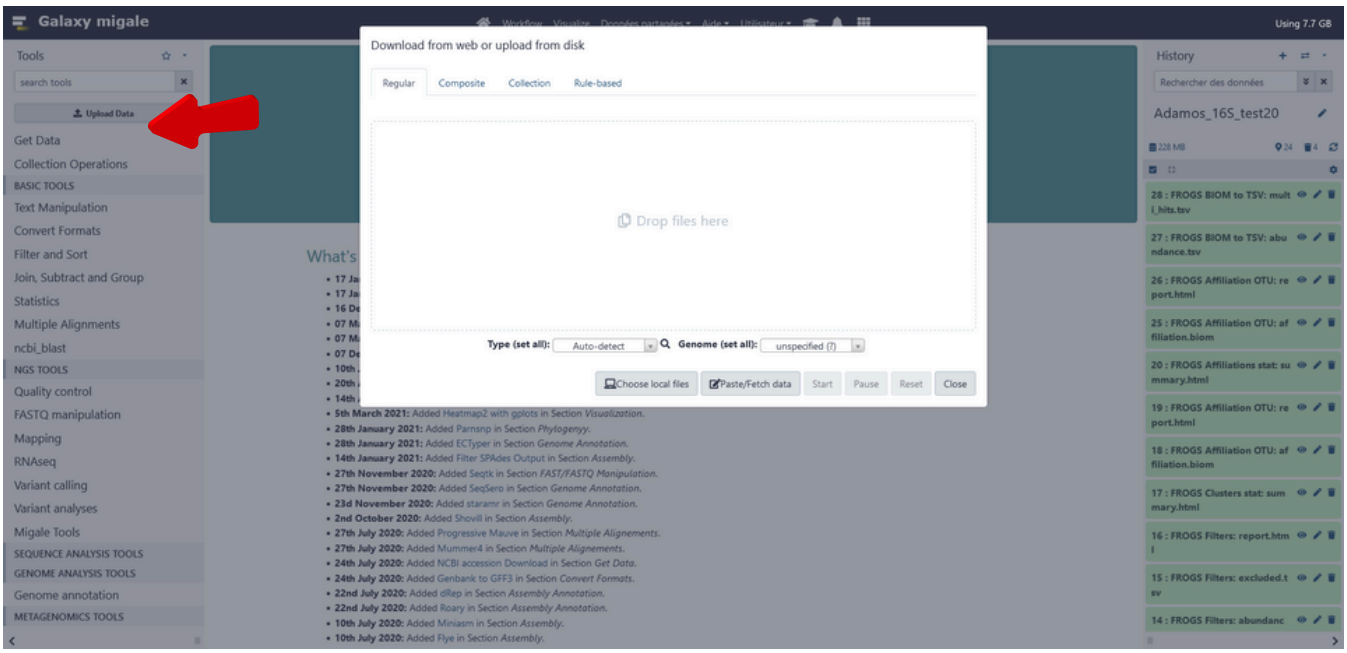

Figure 1. *Capture d'écran de l'étape d'import des données sur Galaxy*

6.3. Démultiplexage

Principe

Le démultiplexage consiste, grâce au repérage des barcodes dans les séquences, à associer les lectures à l'échantillon d'origine quand plusieurs échantillons ont été séquencés en même temps sur une même piste d'Illumina.

Comment réaliser cette étape avec Frogs ?

Il faut utiliser l'outil « FROGS Démultiplex reads » (Figure 2) qui demande en entrée le fichier tabulé contenant les noms des échantillons et les barcodes correspondants («Barcode File ») et le(s) fichier(s) de séquences au format fastq. Vous devez préciser si votre séquençage a été réalisé dans un sens (single-end data) ou dans les deux (paired-end data). Vous pouvez aussi préciser le nombre de mismatches que vous autorisez sur les barcodes (par défaut la valeur est à 0), et la localisation des barcodes (par défaut en position Forward). Les valeurs par défaut sont celles qui sont recommandées dans la plupart des cas mais vous avez tout de même la main pour les modifier

Dans votre historique de travail, vous obtiendrez en sortie, un fichier tabulé contenant le nombre de séquences par échantillon et deux archives TAR.GZ : la première contenant les données de séquences démultiplexées (à utiliser pour le preprocessing), la seconde contenant celles qui n'ont pas pu être démultiplexées et qui ne pourront donc pas être utilisées par la suite.

Sur le panneau central, en dessous des paramètres à renseigner, se trouve une aide pour chaque outil disponible sur Galaxy

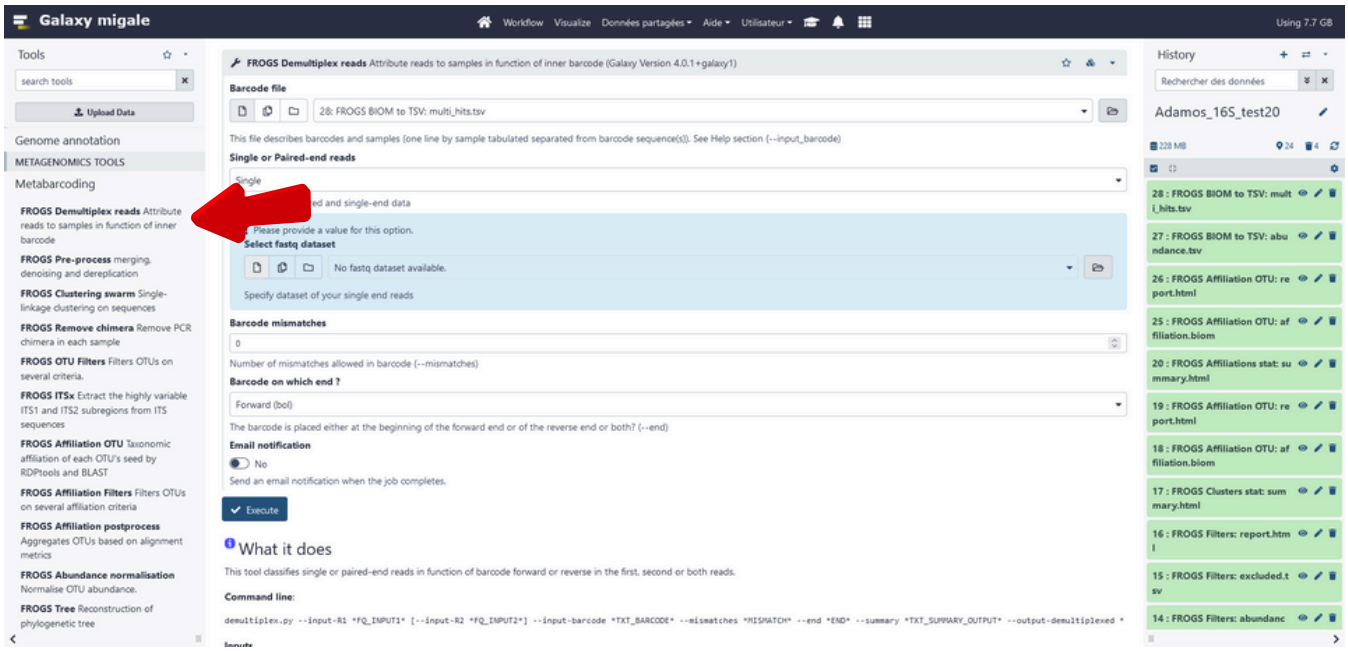

Figure 2. *Capture d'écran de l'étape de démultiplexage avec FROGS sur Galaxy*

6.4. Preprocessing

Principe

Le preprocessing consiste à éliminer du jeu de données les séquences dont la qualité est inférieure à ce qui est attendu. Par exemple, il est souhaitable d'éliminer celles contenant des bases inconnues ("N"), celles qui ne font pas la taille attendue (soit trop courtes soit trop longues) et celles ne contenant pas les amorces ou avec des erreurs dans celles-ci. FROGS profite également de cette étape pour éliminer les séquences des amorces et pour contiguer les séquences R1 et R2, c'est-à-dire aligner les séquences forward (R1) et reverse (R2) et produire une séquence consensus. Enfin l'outil va également dérépliquer les séquences, c'est-à-dire ne conserver qu'un seul exemplaire de chaque séquence unique tout en gardant l'information de comptage de ces séquences dans les différents échantillons (ceci permet de réduire drastiquement les temps de calculs lors des étapes suivantes).

Comment réaliser cette étape avec Frogs ?

Il faut utiliser l'outil « FROGS Pre-process » (Figure 3) qui demande en entrée soit les fichiers de séquences au format fastq par échantillon soit une archive contenant ces fichiers (cette dernière solution est recommandée, voir section import des données et démultiplexage).

Vous devez ensuite indiquer la taille de vos séquences (lectures R1 et R2, en général 250 ou 300 bases, information à vérifier auprès du prestataire de séquençage ou en ouvrant un fichier fastq) et la façon dont sera réalisé le contiguage. Les valeurs par défauts conviennent dans la plupart des cas. Pour les données fongiques pour lesquelles les marqueurs choisis peuvent être occasionnellement d'une taille importante (ITS1 ou ITS2 par exemple), il est recommandé d'activer l'option « *Would you like to keep unmerged reads*? ».

Il faut également préciser les tailles minimale et maximale de l'amplicon attendu. Exemples : pour les régions V3 et V4 du gène codant pour l'ARNr16S, l'amplicon attendu fait environ 460 paires de bases et sa taille ne varie pas beaucoup d'une espèce à l'autre. Vous pouvez donc préciser comme valeur minimum 400 et comme valeur maximum 520, ce qui permettra d'exclure, a priori, des séquences qui ne correspondent pas à ces régions. Pour la région ITS2, puisque sa taille est très variable d'une espèce à l'autre, il est recommandé de renseigner une gamme de valeurs beaucoup plus étendue, par exemple 40 à 1000.

Le protocole de séquençage, par défaut « Illumina standard », convient dans la plupart des cas. Cependant, si vos séquences sont déjà dépourvues des amorces vous pouvez sélectionner « Custom protocol (Kozich *et al.*, 2013).

Enfin, il faut renseigner les séquences des amorces que vous avez utilisées lors de l'amplification par PCR (celle indiquée en annexe 3 par exemple). Concernant l'amorce reverse, elle doit être reverse-complémentée. Vous pouvez effectuer cette opération avec l'outil en ligne suivant :
https://www.bioinformatics.org/sms/rev_comp.html.

Dans votre historique de travail, vous obtiendrez en sortie :

1

Un fichier au format fasta contenant les séquences dérépliquées ayant été conservées suite au pre-processing.

2

Un fichier de comptage au format tsv contenant le nombre de fois que chaque séquence unique a été observée dans chacun des échantillons.

3

Un rapport au format html contenant un compte-rendu de cette étape

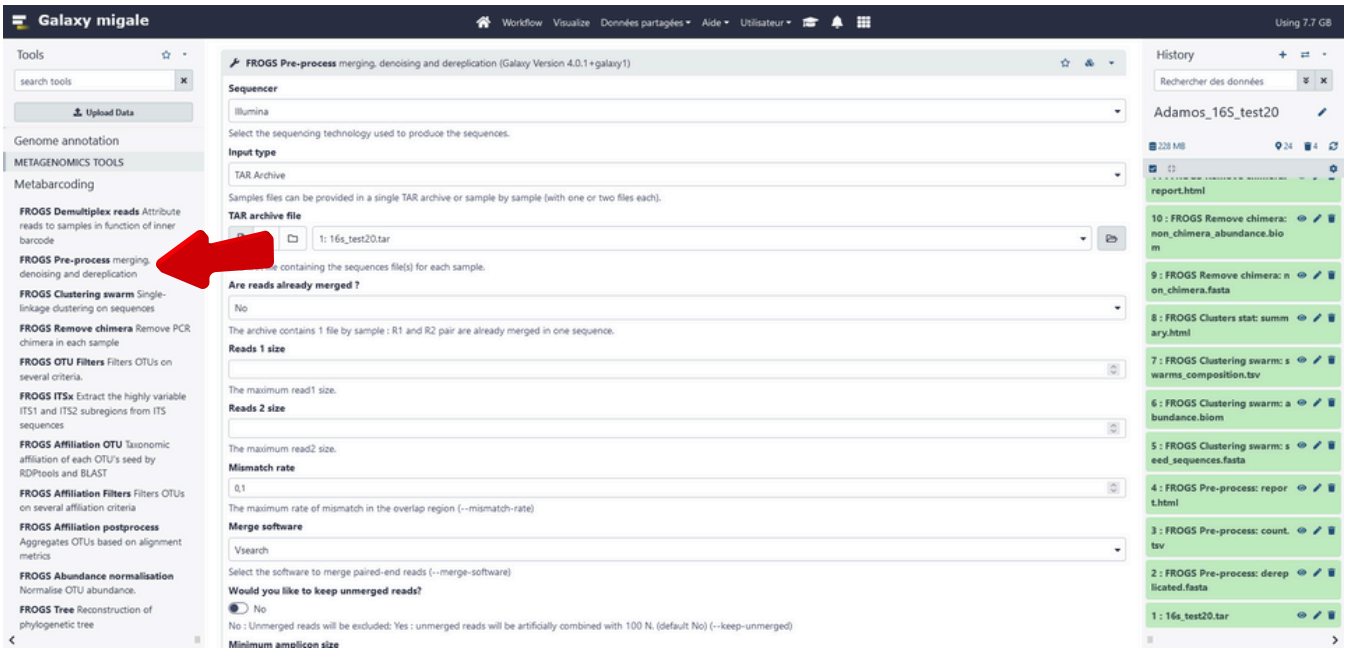

Figure 3. *Capture d'écran de l'étape de pre-processing avec FROGS sur Galaxy*

6.5. Définition des unités taxonomiques opérationnelles (OTUs)

Principe

Cette étape consiste à regrouper les séquences qui se ressemblent fortement en unité taxonomiques opérationnelles (OTUs ou Clusters).

Dans les étapes ultérieures, seule une séquence représentative de chaque OTU, souvent la plus abondante, sera assignée taxonomiquement. En revanche, toutes les séquences de l'OTU serviront au calcul d'abondance de ce taxon. Cette procédure permet de réduire le nombre de séquences à traiter, et donc le temps de calcul, mais aussi de masquer certaines erreurs introduites lors de l'amplification par PCR et du séquençage.

Il existe plusieurs approches bioinformatiques pour effectuer ce regroupement de séquences. A l'heure actuelle, l'agrégation progressive de séquences (une distance d'aggrégation de 1 combinée à l'option "Refine OTU clustering" permet généralement de bien séparer des séquences provenant d'espèces différentes) proposée par l'outil SWARM (Mahé *et al.*, 2014) et la définition d'ASVs (*Amplicon Sequence Variants*) proposée par l'outil *DADA2* (Callahan *et al.*, 2016) qui conserve tous les variants de séquences possibles après avoir éliminé les possibles erreurs de séquençage grâce à un modèle dédié, sont les plus utilisées.

Comment réaliser cette étape avec Frogs ?

FROGS utilise SWARM pour réaliser cette étape. Il faut utiliser l'outil « *FROGS Clustering swarm* » (Figure 4) qui demande en entrée un fichier de séquences au format fasta et le fichier de comptage correspondant au format tsv, comme ceux produits par l'étape de pre-processing.

Vous devez ensuite choisir la distance d'agrégation, c'est-à-dire la valeur X définissant le nombre de différences autorisées pour que deux séquences appartiennent à la même OTU. Par défaut cette valeur est fixée à 1 et est combinée à l'option "Refine OTU clustering". Dans ces conditions, elle convient assez bien pour discriminer des espèces bactériennes sur la base des séquences des régions V3-V4 de l'ADNr16S et des espèces fongiques sur la base de la région ITS2.

Normalement, FROGS sélectionne par défaut les fichiers produits à l'étape de pre-processing dans votre historique de travail mais vérifiez que ce sont bien ceux-là qui ont été sélectionnés.

Dans votre historique de travail, vous obtiendrez en sortie, un fichier de séquences au format fasta (nommé *seed_sequences.fasta*) contenant les séquences représentatives des OTUs ainsi qu'une table d'abondance au format biom (nommée *abundance.biom*).

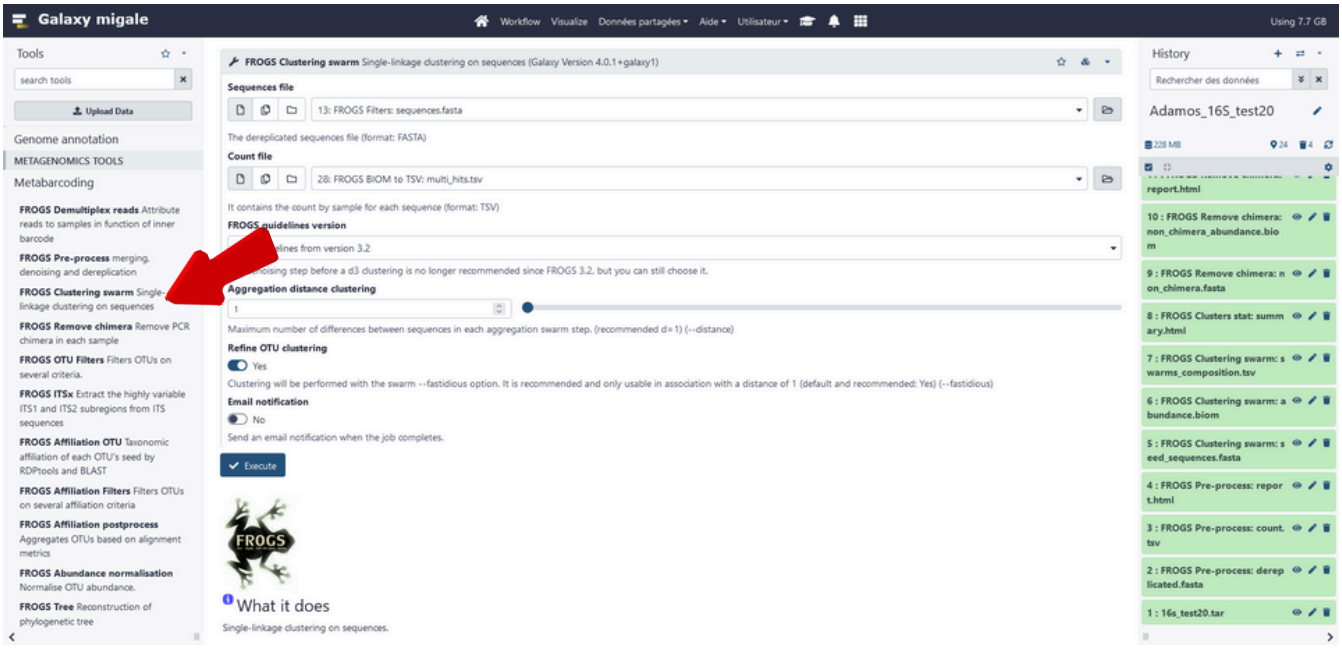

Figure 4. *Capture d'écran de l'étape de définition des OTUs avec FROGS sur Galaxy*

6.6. Elimination des séquences chimériques

Principe

Les séquences chimériques sont des séquences artéfactuelles, produites au cours de l'amplification par PCR. Lorsqu'un brin d'ADN en cours d'élongation se détache de sa matrice, il peut s'hybrider à un brin matrice différent lors des cycles suivants, et ainsi former une molécule chimérique. Cette chimère, composée de la séquence de deux (ou plus) brins matrices, n'a aucune existence dans l'échantillon de départ, mais devient présente dans la librairie et sera donc séquencée. Il est donc indispensable de repérer ce type de séquences et de les éliminer du jeu de données.

Il existe une multitude d'outils bioinformatiques permettent de réaliser ce travail. Certains utilisent une méthode dite « sur référence » qui est basée sur le principe que les séquences 'parents' de la chimère sont présentes dans les bases de données. La chimère est détectée par alignement des séquences représentatives des OTUs contre une base de données de séquences de référence. Une lecture mosaïque qui génère des alignements partiels sur ≥2 séquences de référence appartenant à des taxons différents sera éliminée. D'autres outils utilisent une méthode dite « *de novo* » qui part du principe que les séquences 'parents' de la chimère sont présentes dans l'échantillon séquencé en plus grande quantité que la chimère elle-même. La chimère est repérée suite à l'alignement de chaque séquence représentative d'un OTU avec les autres qui sont détectées dans le même échantillon. Une séquence mosaïque qui génère des alignements partiels (sur une partie de la longueur) sur ≥2 séquences différentes présentes dans l'échantillon considéré sera éliminée.

Comment réaliser cette étape avec Frogs ?

FROGS utilise la méthode *de novo* implémentée dans l'outil VSEARCH (Rognes *et al.*, 2016) pour réaliser cette étape. Il faut utiliser l'outil « *FROGS Remove chimera* » (Figure 5) qui demande en entrée un fichier de séquences au format fasta et la table d'abondance au format biom correspondante, comme ceux produits par l'étape de définition des OTUs.

Dans votre historique de travail, vous obtiendrez en sortie :

1. Un fichier de séquences au format fasta (nommé non_chimera.fasta) contenant les séquences non chimériques
2. Une table d'abondance au format biom (nommée non_chimera_abundance .biom) qui ne contiendra plus que l'information correspondante à ces séquences.
3. Un rapport au format html qui résume les principaux résultats obtenus lors de cette étape.

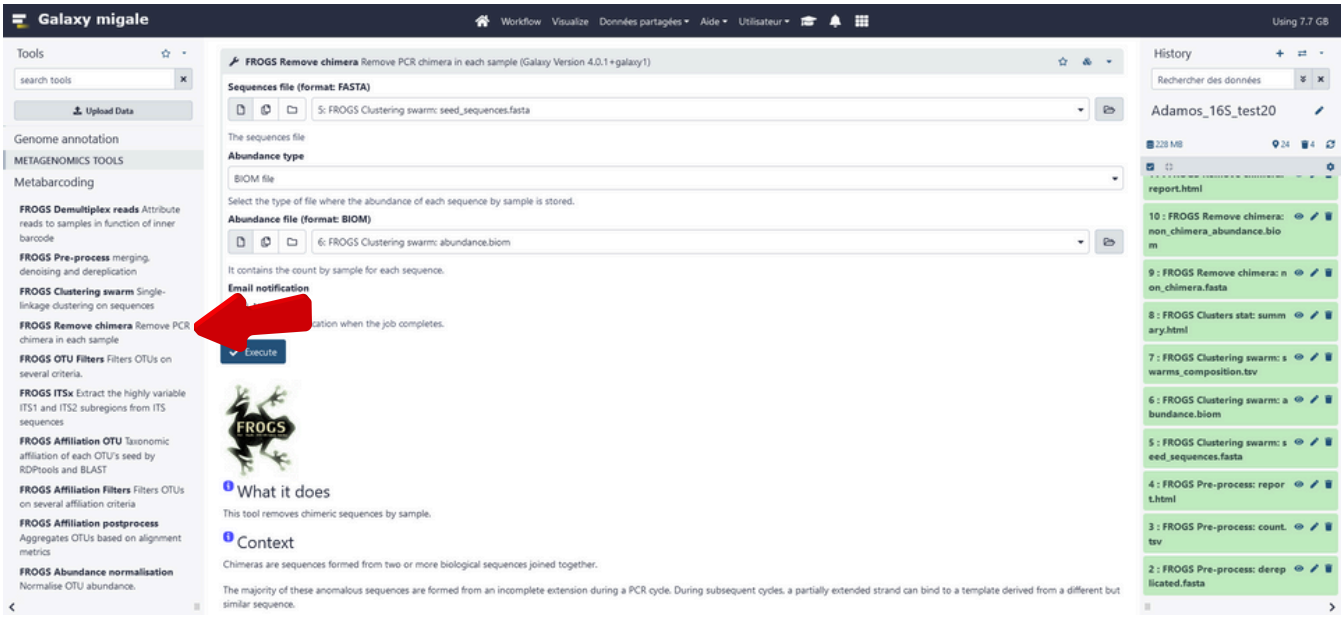

Figure 5. *Capture d'écran de l'étape d'élimination des séquences chimériques avec FROGS sur Galaxy*

6.7. Filtres sur la base de l'abondance

Principe

Les OTUs présents en très faible abondance sont majoritairement des chimères ou des séquences comportant des erreurs de séquençage, qui n'ont pas été détectées lors des étapes précédentes. Il est donc très courant de les supprimer du jeu de données. En particulier, une étude de référence dans le domaine recommande de supprimer les OTUs dont l'abondance relative dans le jeu de données est inférieure à 0,005 % (Bokulich *et al.*, 2013).

Si le plan expérimental inclu des réplicats biologiques ou techniques, il peut aussi être intéressant d'utiliser cette information pour ne retenir que les OTUs dont la présence est répétable. Ces filtres sont essentiels car ils permettent de réduire drastiquement le nombre d'OTUs générés en éliminant la majorité du bruit de fond généré par l'approche de métabarcoding. Ils facilitent ainsi l'interprétation biologique des résultats et leur robustesse.

Comment réaliser cette étape avec Frogs ?

Il faut utiliser l'outil « FROGS Filters » (Figure 6) qui demande en entrée un fichier de séquences au format fasta et la table d'abondance au format biom correspondante, comme ceux produits par l'étape d'élimination des séquences chimériques.

Pour appliquer le filtre recommandé par Bokulich *et al.* (2013) à l'ensemble du jeu de données, rentrez la valeur 0,00005 pour le paramètre « *Minimum proportion of sequences abundancy to keep OTU* ».

Vous pouvez également indiquer le nombre d'échantillons minimum dans lequel une OTUs doit être détectée pour être conservée grâce à l'option « *Minimum prevalence* » ou choisir de ne conserver que les X OTUs les plus abondants du jeu de données avec l'option « *N biggest OTU* » (cette dernière option n'est pas recommandée).

Dans votre historique de travail, vous obtiendrez en sortie :

1. Un fichier de séquences au format fasta (nommé sequences.fasta) contenant les séquences filtrées
2. Une table d'abondance au format biom (nommée abundance.biom) qui ne contiendra plus que l'information correspondante à ces séquences
3. Un rapport au format html qui résume les principaux résultats obtenus lors de cette étape
4. Un fichier tabulé au format tsv (nommé *excluded.tsv*) contenant les identifiants des OTUs exclues lors de cette étape

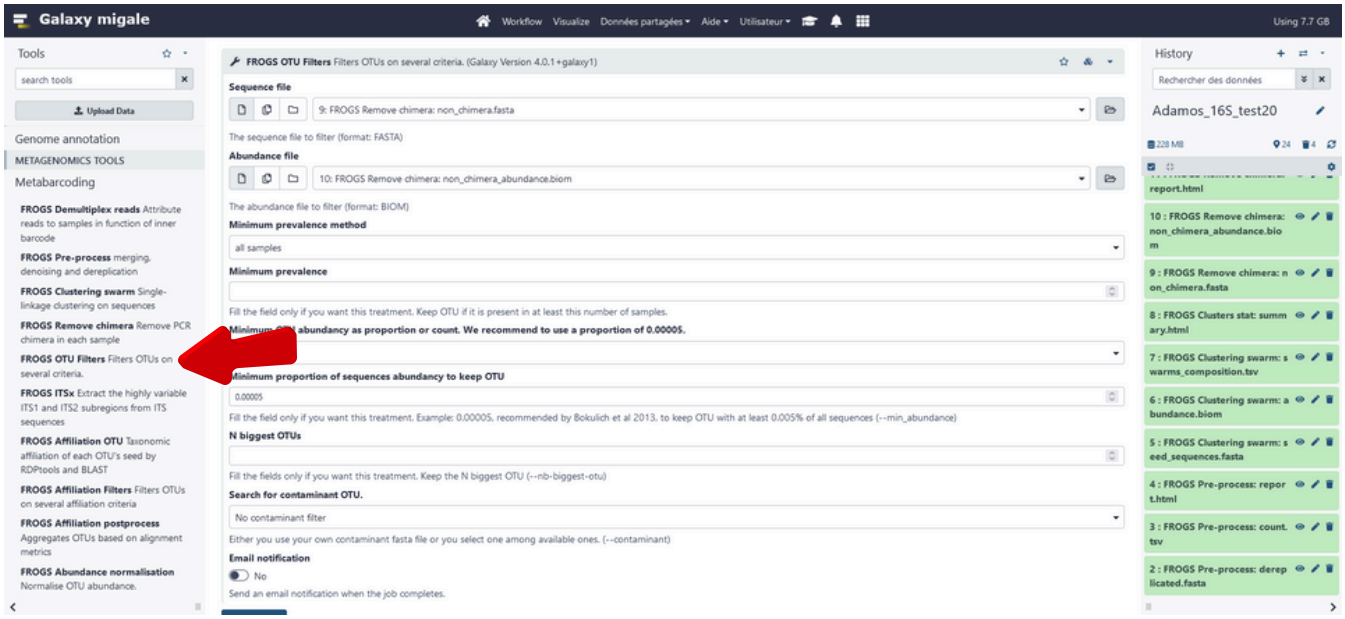

Figure 6. *Capture d'écran de l'étape de filtration sur la base de l'abondance*

6.8. Extraction des séquences ITS

(à réaliser uniquement pour les données fongiques)

Principe

Les marqueurs ITS1 et ITS2 sont des séquences intergéniques hypervariables situées entre les gènes codants pour des ARN ribosomaux des organismes eucaryotes (18S ; 5,8S ; 28S). Les amorces utilisées pour les amplifier par PCR ciblent des régions conservées de ces gènes. Afin de faciliter l'assignation taxonomique des séquences, il convient d'extraire uniquement les portions correspondantes à la région ITS ciblée (ITS1 ou ITS2) et d'éliminer les petits morceaux de séquences correspondantes aux gènes codants pour les ARN ribosomaux.

L'outil ITSx (Bengtsson-Palme *et al.*, 2013) a été développé pour répondre à cet objectif. Il permet en outre d'éliminer toutes les séquences qui ne correspondent pas à des séquences ITS et dont la présence dans le jeu de données peut être liée à des amplifications aspécifiques.

Comment réaliser cette étape avec Frogs ?

Il faut utiliser l'outil « FROGS ITSx » (Figure 7) qui demande en entrée un fichier de séquences au format fasta et la table d'abondance au format biom correspondante, comme ceux produits par l'étape d'élimination des séquences chimériques et l'étape de filtration sur la base de l'abondance.

Vous avez la possibilité d'activer l'option « Trim conserved sequence (SSU, 5.8S, LSU) ? » (recommandé, par défaut elle est désactivée), qui permet d'éliminer les petits morceaux de séquences correspondantes aux gènes codants pour les ARN ribosomaux. Vous avez également la possibilité de choisir de conserver seulement les séquences correspondantes à certains organismes (Fungi par défaut).

Dans votre historique de travail, vous obtiendrez en sortie :

1. Un fichier de séquences au format fasta (nommé itsx.fasta) contenant les séquences ITS filtrées.
2. Une table d'abondance au format biom (nommée itsx.biom) qui ne contiendra plus que l'information correspondante à ces séquences
3. Un rapport au format html qui résume les principaux résultats obtenus lors de cette étape .
4. Un second fichier de séquences au format fasta (nommé *non_ITS_sequence.fasta*) contenant les séquences des OTUs exclues lors de cette étape et qui ne correspondent pas, a priori, à des séquences ITS.

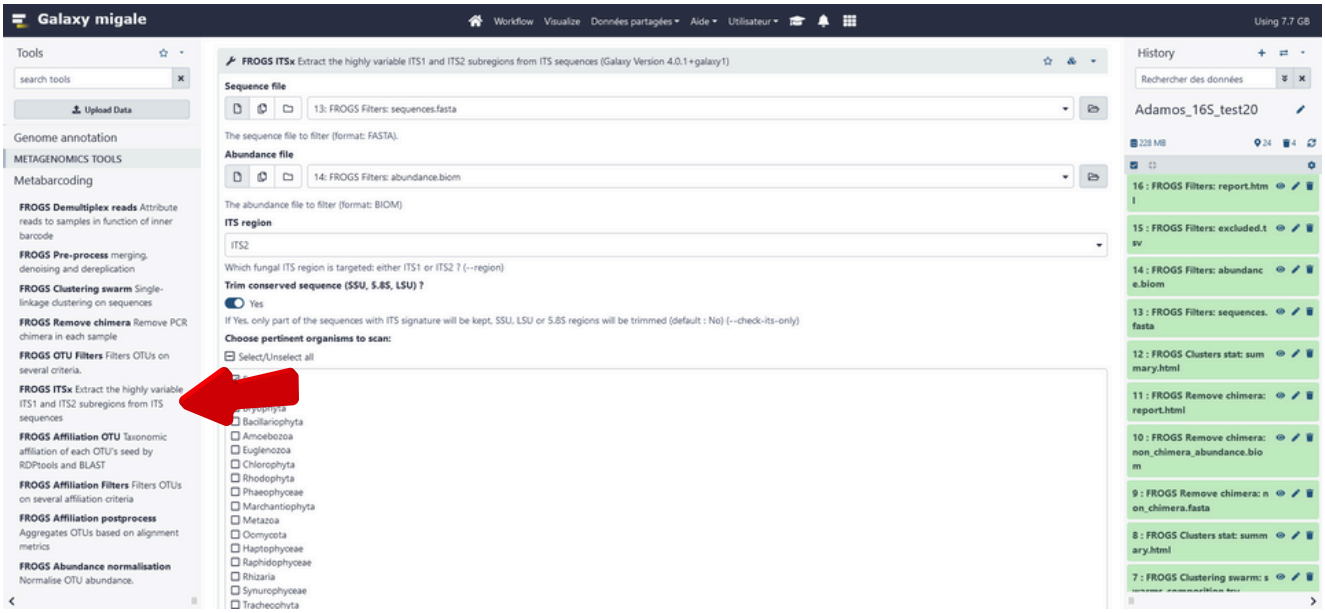

Figure 7. *Capture d'écran de l'étape d'extraction des séquences ITS avec FROGS sur Galaxy*

6.9. Affiliation taxonomique

Principe

L'assignation taxonomique se fait par comparaison des séquences représentatives de OTUs avec une base de données de séquences de référence qui peut être soit générale (ex : *Silva*, *Greengenes*, *LTP*, *EzTaxon*) soit spécifique à un écosystème (ex : *DairyDB* pour les produits laitiers).

Comment réaliser cette étape avec Frogs ?

FROGS utilise par défaut l'outil Blast (Altschul *et al.*, 1990) pour réaliser cette étape et plusieurs bases de données de références sont accessibles (les plus couramment utilisées sont SSU_SILVA pour les séquences ADNr16S et ITS_UNITE_FUNGI pour les séquences ITS1 et ITS2 de levures et champignons). FROGS offre également la possibilité, en option, d'effectuer une seconde affiliation avec l'outil RDP Classifier (Wang *et al.*, 2007). Il faut utiliser l'outil « FROGS Affiliation OTU » (Figure 8) qui demande en entrée un fichier de séquences au format fasta et la table d'abondance au format biom correspondante, comme ceux produits par l'étape d'élimination des séquences chimériques ou d'extraction des séquences ITS.

Dans votre historique de travail, vous obtiendrez en sortie :

1. Une table d'abondance au format biom (nommée *affiliation_abundance.biom*) qui contiendra les données de comptage des OTUs dans les différents échantillons et la taxonomie de chacune des OTUs (selon 7 rangs : *Kingdom*, *Phylum*, *Order*, *Class*, *Familly*, *Genus*, *Species*)

2. un rapport au format html qui résume les principaux résultats obtenus lors de cette étape.

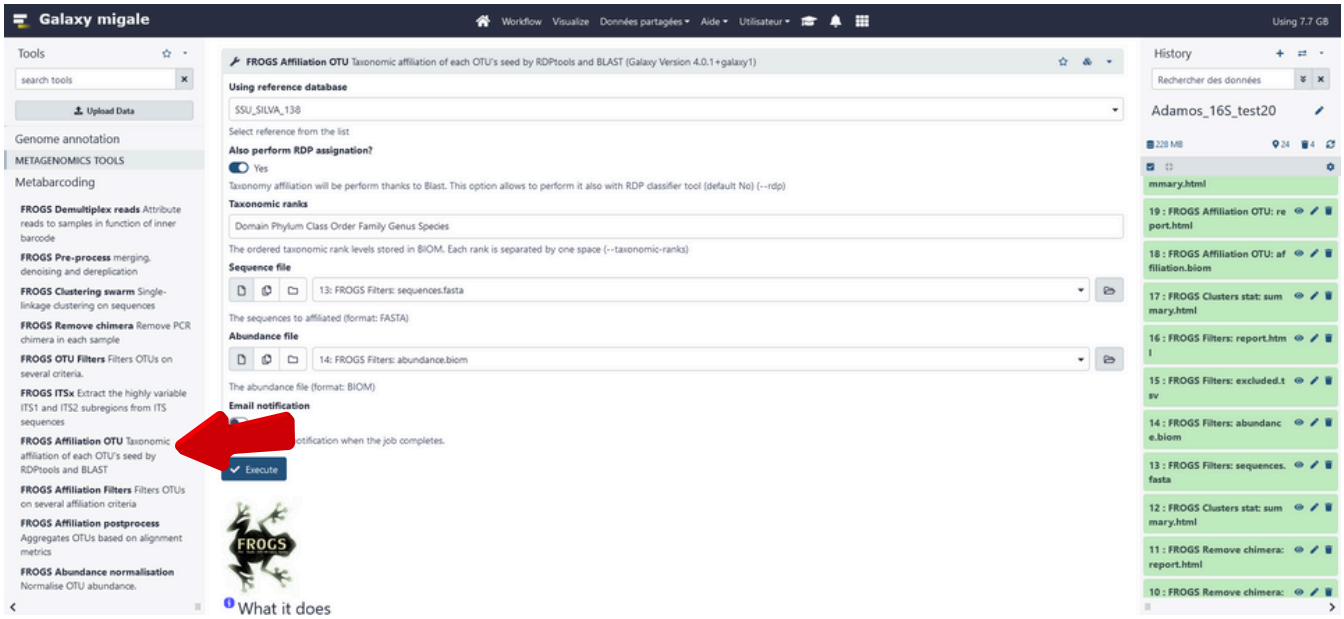

Figure 8. *Capture d'écran de l'étape d'affiliation taxonomique avec FROGS sur Galaxy*

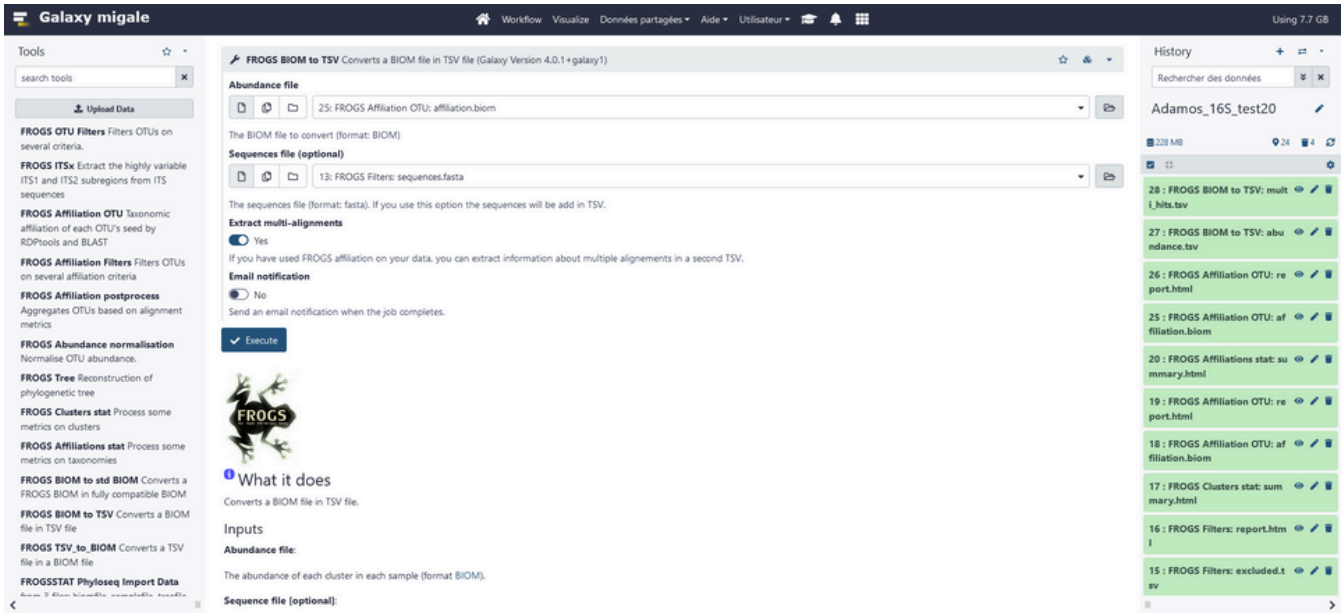

Figure 9. *Capture d'écran de l'étape de transformation de la table d'abondance du format biom au format tsv avec FROGS sur Galaxy*

6.10. Export des données et contrôle manuel des affiliations

Principe

La table d'abondance est au format biom (http://biom-format.org), le format standard mis en place par la communauté pour manipuler et stocker des données de comptages sous forme compacte. C'est un format destiné à une utilisation par des machines qui n'est donc pas facilement lisible par l'être humain. Il est néanmoins possible de convertir un fichier biom en fichier tabulé (tsv) qui peut être ouvert avec un tableur classique.

Pour certaines OTUs, les affiliations taxonomiques proposées méritent d'être contrôlées et expertisées : cela concerne au moins les OTUs les plus abondantes, celles présentant des multi-affiliations ou celles pour lesquelles les pourcentages d'identité et de couverture sont <98%.

Comment réaliser cette étape avec Frogs ?

Pour convertir la table au format biom en fichier tabulé il faut utiliser l'outil « *FROGS biom to tsv* » (Figure 9) qui demande en entrée un fichier de séquences au format fasta (ex : sortie de l'étape de filtre) et la table d'abondance au format biom correspondante, comme celle produite par l'étape d'affiliation taxonomique. L'option « *Extract multi-alignments* », lorsqu'elle est activée permet de générer en sortie, en plus de la table d'abondance au format tabulé (nommée *abundance.tsv*), une seconde table au format tsv (nommée *multi_affiliations.tsv*). Certaines OTUs peuvent en effet être multi-affiliées, c'est-à-dire que plusieurs résultats identiques ont été obtenus par Blast contre la base de données de séquences de référence. Cette seconde table contient la liste de ces meilleurs résultats pour vous aider à choisir l'affiliation à conserver pour vos analyses.

Il est également possible de récupérer les séquences représentatives des OTUs dans le fichier abundance.tsv et de les comparer à celles présentes dans d'autres bases de données que celle utilisée lors de l'étape d'affiliation (pour les bactéries EzBioCloud est un choix intéressant car cette base de données est expertisée). Sur la base de ces résultats supplémentaires, vous pourrez faire le choix de conserver ou non les OTUs dans le jeu de données, et de conserver ou de modifier l'affiliation taxonomique proposée.

Vous pouvez enfin exploiter les résultats des contrôles (positifs et négatifs) à ce stade pour consolider vos données. Par exemple, vous pouvez décider d'éliminer du jeu de données les OTUs qui sont majoritaires dans les contrôles négatifs et qui sont donc probablement le reflet d'une contamination. Si les résultats des contrôles positifs (mock) sont très différents de l'attendu, vous pouvez également le garder à l'esprit pour l'interprétation des résultats.

Si des modifications manuelles sont apportées à la table d'abondance, il faudra l'enregistrer au format tsv et la réimporter sur galaxy à l'aide de l'outil « *FROGS tsv to biom* ».

Une fois la table d'abondance vérifiée et consolidée, des tests statistiques seront conduits afin de répondre à ou aux question(s) posée(s) au départ. FROGS dispose d'une série d'outils statistiques dont le nom débute par « FROGSSTAT » que vous pouvez explorer librement en vous laissant guider par les aides disponibles. Ces outils sont principalement basés sur les fonctions disponibles dans le package R intitulé PhyloSeq (McMurdie and Holmes, 2013).

Partie 7 · Analyse bioinformatique des données de métabarcoding ADN avec FROGS

1. Normalisation des données
2. Diversité alpha
3. Diversité bêta
4. Analyses différentielles

Une fois la table d'abondance produite, l'exploration et l'exploitation des données passent par l'utilisation de méthodes statistiques. Dans la communauté scientifique, cette analyse est très majoritairement réalisée sous l'environnement de travail R (https://www.r-project.org). Pour faciliter le transfert de ces méthodes vers un public non informaticien/statisticien, des fonctions ont été implémentées dans FROGS et des applications shiny dédiées ont été développées.

On peut citer par exemple :

- ExploreMetabar
https://shiny.migale.inrae.fr/app/exploremetabar
- Easy16S
https://shiny.migale.inrae.fr/app/easy16S

Pour utiliser ces outils il convient d'avoir à disposition au minimum les deux fichiers suivants :

1. La table d'abondance vérifiée et consolidée (format biom).
2. Un fichier tabulé contenant les données associées aux échantillons (metadata).

Comment réaliser cette étape avec Frogs ?

Il est possible de créer avec FROGS un objet de type RData à l'aide de l'outil « *FROGSSTAT Phyloseq Import Data* » qui permettra de poursuivre les analyses sous les différents environnements de travail cités précédemment (R, FROGS, ExploreMetabar, Easy16S).

Que se passe-t-il ensuite ?

Dans cette partie sont décrites les différentes étapes classiquement réalisées pour l'analyse d'une table d'abondance (peu importe l'environnement de travail choisi) et les méthodes associées les plus courantes.

7.1. Normalisation des données

Principe

La profondeur de séquençage n'est pas nécessairement homogène entre tous les échantillons provenant du même projet. Pour corriger ce biais, il convient de normaliser les données présentes dans la table d'abondance avant de poursuivre les analyses.

Principales méthodes utilisées

- **Total Sum Scale** (**TSS**) : somme des reads par échantillon ramenée en %
- **Raréfaction** : sous-échantillonnage par tirage aléatoire du même nombre de reads pour chaque échantillon, en se basant sur le nombre de reads de l'échantillon le moins profond

Les deux méthodes sont toutes les deux très utilisées, soit dès le début de l'analyse ou seulement pour les analyses de béta-diversité.

Il n'est pas nécessaire de normaliser les données pour le calcul de certains indices de diversité alpha (ex : Shannon et Inverse Simpson).

La méthode TSS est simple et conservative, et est mieux adaptée que la raréfaction pour le calcul de l'indice de diversité alpha Chao1. En effet, elle conserve l'information sur les ASVs/OTUs rares alors qu'elle est perdue avec la méthode de raréfaction. En revanche, elle est sensible au biais de profondeur. Dans certains cas, notamment lorsque la profondeur de séquençage est très hétérogène, il peut donc s'avérer plus judicieux d'utiliser la raréfaction.

7.2. Diversité alpha

Principe

Les analyses dites de diversité alpha sont utilisées pour évaluer le niveau de diversité présent dans chacun des échantillons pris indépendamment. Pour cela différents indices peuvent être calculés puis ensuite comparés d'un échantillon à l'autre ou entre plusieurs groupes d'échantillons (exemple en Figure 10).

Principales méthodes utilisées

- **Richesse observée** : nombre d'OTUs détectés dans l'échantillon
- **Chao1** : estimation du nombre d'OTUs présents dans l'échantillon
- **Shannon** et **Inverse Simpson** : indices plus complexes prenant en compte non seulement la richesse mais aussi l'équitabilité des abondances observées. Plus ils sont élevés, plus il y a d'OTUs avec des abondances fortes.

Les tests statistiques

- Une analyse de variance (**ANOVA**) peut être utilisée en première approche pour tester l'effet d'un ou deux facteurs sur la diversité mesurée par un des indices ci-dessus. Lorsque le ou les facteurs testés séparent les échantillons en plus de deux groupes, un **test de Tukey** peut être utilisé pour comparer les groupes d'échantillons deux à deux. L'ANOVA et le test de Tukey sont des tests paramétriques. Ils supposent donc que les variances sont homogènes.
- L'alternative non-paramétrique consiste à effectuer un test de **Kruskal-Wallis** suivi d'un **test post-hoc** pour les comparaisons deux à deux. Cette approche est préférable si les variances ne sont pas homogènes

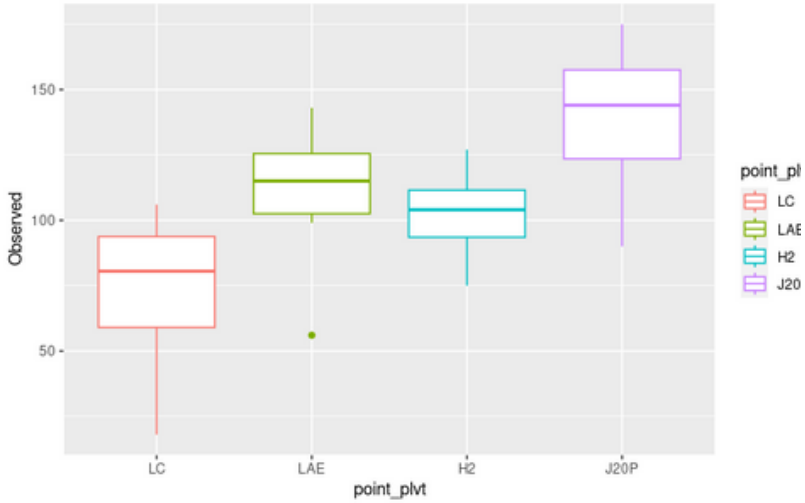

Figure 10. *Comparaison de la richesse en OTU observée dans les échantillon d'un jeu de données issu du projet Adamos (LC = lait cru, LAE = lait avant emprésurage, H2 = fromage 2h après moulage ; J20P = pâte 20 jours après moulage)*

7.3. Diversité bêta

Principe

Les analyses dites de diversité beta sont utilisées pour comparer la composition des communautés microbiennes entre échantillons. Il est alors possible de repérer quels échantillons ont des communautés qui sont similaires ou au contraire différentes (exemple en Figure 11).

Ces analyses sont basées sur le calcul de distances entre chaque paire d'échantillons et des méthodes d'ordination sont ensuite appliquées sur les matrices de distances pour représenter les relations entre les échantillons.

Il est indispensable d'avoir normalisé la table d'abondance avant de réaliser des analyses de diversité bêta.

Les mesures de distance

Il en existe plusieurs et toutes ne prennent pas en compte les mêmes informations dans leur calcul. Les plus utilisées sont les suivantes :

- **Jaccard** : prend en compte la présence/absence des taxa seulement
- **Bray-Curtis** : prend en compte la présence/absence et l'abondance des taxa en donnant plus de poids aux taxa abondants (= méthode pondérée)
- **UniFrac** : prend en compte la présence/absence des taxa et donne un poids plus important aux taxa éloignés dans un arbre phylogénétique (attention : pour utiliser cette distance, il faut fournir un arbre phylogénétique pour la construction de l'objet de type RData)
- **UniFrac pondérée** : version pondérée de Unifrac, qui tient donc compte en plus de l'abondance des taxa (attention : pour utiliser cette distance, il faut fournir un arbre phylogénétique pour la construction de l'objet de type RData)

Plus une valeur de distance entre deux échantillons est proche de 0 plus elle témoigne d'une structure de communauté différente.

Les mesures d'ordination

- **Multidimensional Scaling** (MDS), autrement appelée Principal Coordinate Analysis (PCoA) : cette méthode préserve les valeurs brutes des distances pour représenter les relations entre les échantillons dans un espace à faible dimension
- **Non-Metric Multidimensional Scaling** (NMDS) : cette méthode non paramétrique conserve uniquement les rangs des distances (méthode non paramétrique) pour représenter les relations entre les échantillons dans un espace à faible dimension

Une fois représentés dans un espace à deux ou trois dimensions, les échantillons (points dont la forme ou la couleur permettent d'afficher leur appartenance à un groupe) sont plus ou moins éloignés les uns des autres en fonction de la similarité de leur profil microbien.
Pour identifier des structures dans les données, il est tout à fait possible (et même recommandé), d'explorer la table d'abondance avec plusieurs combinaisons de distances et de méthodes d'ordination.

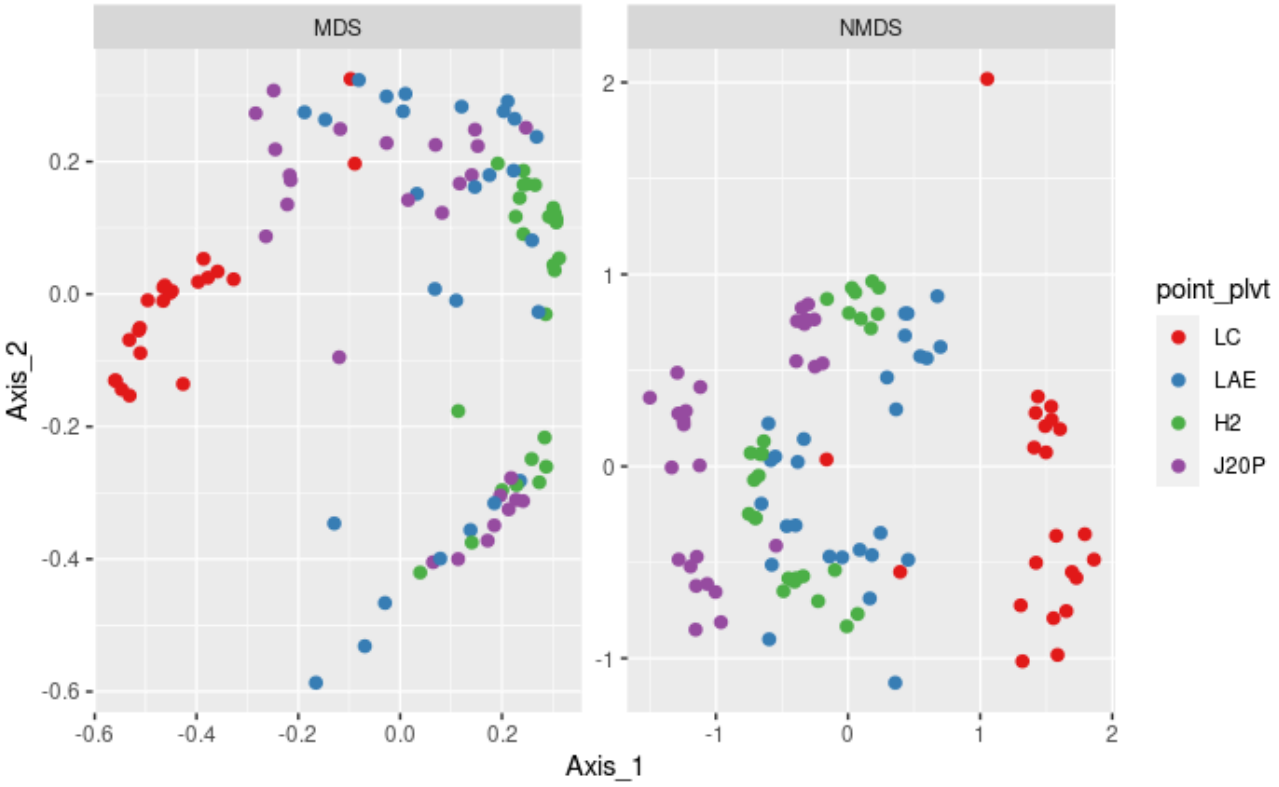

Figure 11. *Représentation ordinale en MDS et NMDS d'un jeu de données issu de prélèvements du projet Adamos (points de prélèvements : LC = lait cru, LAE = lait avant emprésurage, H2 = fromage 2h après moulage ; J20P = pâte 20 jours après moulage)*

Les tests statistiques

- L'**analyse de variance permutationnelle** (PERMANOVA) est utilisée pour tester l'effet de plusieurs facteurs sur une matrice de distances.
- La méthode **Pairwise ADONIS** est utilisée pour compléter l'analyse avec des comparaisons de groupes d'échantillons deux à deux.

7.4. Analyses différentielles

Principe

Les analyses différentielles permettent de repérer les taxa (OTUs) différentiellement présents ou abondants dans un groupe d'échantillons par rapport à un autre.

Méthodes les plus courantes

- **DeSeq2** (Love *et al.*, 2014)
- **Metagenome** Seq (Paulson *et al.*, 2013)
- **Metacoder** (Foster *et al.*, 2017)
- **ANCOM BC** (Lin and Peddada, 2020)

Concernant ces méthodes, la communauté scientifique n'a pas le recul nécessaire pour promouvoir l'une plutôt qu'une autre. A l'heure actuelle, il est donc recommandé d'en utiliser plusieurs et de comparer les résultats. Si des OTUs sont repérés comme différentiellement abondants avec plusieurs méthodes, cela donne de la robustesse au résultat.

Références

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., **1990**. Basic local alignment search tool. J. Mol. Biol. 215, 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Bengtsson-Palme, J., Ryberg, M., Hartmann, M., Branco, S., Wang, Z., Godhe, A., Wit, P.D., Sánchez-García, M., Ebersberger, I., Sousa, F. de, Amend, A., Jumpponen, A., Unterseher, M., Kristiansson, E., Abarenkov, K., Bertrand, Y.J.K., Sanli, K., Eriksson, K.M., Vik, U., Veldre, V., Nilsson, R.H., **2013**. Improved software detection and extraction of ITS1 and ITS2 from ribosomal ITS sequences of fungi and other eukaryotes for analysis of environmental sequencing data. Methods Ecol. Evol. 4, 914–919. https://doi.org/10.1111/2041-210X.12073

Bokulich, N.A., Subramanian, S., Faith, J.J., Gevers, D., Gordon, J.I., Knight, R., Mills, D.A., Caporaso, J.G., **2013**. Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. Nat. Methods 10, 57–59. https://doi.org/10.1038/nmeth.2276

Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C.C., Al-Ghalith, G.A., Alexander, H., Alm, E.J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J.E., Bittinger, K., Brejnrod, A., Brislawn, C.J., Brown, C.T., Callahan, B.J., Caraballo-Rodríguez, A.M., Chase, J., Cope, E.K., Da Silva, R., Diener, C., Dorrestein, P.C., Douglas, G.M., Durall, D.M., Duvallet, C., Edwardson, C.F., Ernst, M., Estaki, M., Fouquier, J., Gauglitz, J.M., Gibbons, S.M., Gibson, D.L., Gonzalez, A., Gorlick, K., Guo, J., Hillmann, B., Holmes, S., Holste, H., Huttenhower, C., Huttley, G.A., Janssen, S., Jarmusch, A.K., Jiang, L., Kaehler, B.D., Kang, K.B., Keefe, C.R., Keim, P., Kelley, S.T., Knights, D., Koester, I., Kosciolek, T., Kreps, J., Langille, M.G.I., Lee, J., Ley, R., Liu, Y.-X., Loftfield, E., Lozupone, C., Maher, M., Marotz, C., Martin, B.D., McDonald, D., McIver, L.J., Melnik, A.V., Metcalf, J.L., Morgan, S.C., Morton, J.T., Naimey, A.T., Navas-Molina, J.A., Nothias, L.F., Orchanian, S.B., Pearson, T., Peoples, S.L., Petras, D., Preuss, M.L., Pruesse, E., Rasmussen, L.B., Rivers, A., Robeson, M.S., Rosenthal, P., Segata, N., Shaffer, M., Shiffer, A., Sinha, R., Song, S.J., Spear, J.R., Swafford, A.D., Thompson, L.R., Torres, P.J., Trinh, P., Tripathi, A., Turnbaugh, P.J., Ul-Hasan, S., van der Hooft, J.J.J., Vargas, F., Vázquez-Baeza, Y., Vogtmann, E., von Hippel, M., Walters, W., Wan, Y., Wang, M., Warren, J., Weber, K.C., Williamson, C.H.D., Willis, A.D., Xu, Z.Z., Zaneveld, J.R., Zhang, Y., Zhu, Q., Knight, R., Caporaso, J.G., **2019**. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat. Biotechnol. 37, 852–857. https://doi.org/10.1038/s41587-019-0209-9

Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., Holmes, S.P., **2016**. DADA2: High-resolution sample inference from Illumina amplicon data. Nat. Methods 13, 581–583. https://doi.org/10.1038/nmeth.3869

Escudié, F., Auer, L., Bernard, M., Mariadassou, M., Cauquil, L., Vidal, K., Maman, S., Hernandez-Raquet, G., Combes, S., Pascal, G., **2017**. FROGS: Find, Rapidly, OTUs with Galaxy Solution. Bioinformatics. https://doi.org/10.1093/bioinformatics/btx791

Falentin, H., Auer, L., Mariadassou, M., Pascal, G., Rué, O., Dugat-Bony, E., Delbès, C., Nicolas, A., Rifa, E., Mondy, S., Boulch, M.L., Cauquil, L., Hernandez, G., Terrat, S., Abraham, A.-L., **2019**. Guide pratique à destination des biologistes, bioinformaticiens et statisticiens qui souhaitent s'initier aux analyses métabarcoding 23.

Foster, Z.S.L., Sharpton, T.J., Grünwald, N.J., **2017**. Metacoder: An R package for visualization and manipulation of community taxonomic diversity data. PLOS Comput. Biol. 13, e1005404. https://doi.org/10.1371/journal.pcbi.1005404

Lin, H., Peddada, S.D., **2020**. Analysis of compositions of microbiomes with bias correction. Nat. Commun. 11, 3514. https://doi.org/10.1038/s41467-020-17041-7

Love, M.I., Huber, W., Anders, S., **2014**. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 550. https://doi.org/10.1186/s13059-014-0550-8

Mahé, F., Rognes, T., Quince, C., de Vargas, C., Dunthorn, M., **2014**. Swarm: robust and fast clustering method for amplicon-based studies. PeerJ 2, e593. https://doi.org/10.7717/peerj.593

McMurdie, P.J., Holmes, S., **2013**. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLOS ONE 8, e61217. https://doi.org/10.1371/journal.pone.0061217

Paulson, J.N., Stine, O.C., Bravo, H.C., Pop, M., **2013**. Differential abundance analysis for microbial marker-gene surveys. Nat. Methods 10, 1200–1202. https://doi.org/10.1038/nmeth.2658

Quigley, L., O'Sullivan, O., Beresford, T. p., Paul Ross, R., Fitzgerald, G. f., Cotter, P. d., **2012**. A comparison of methods used to extract bacterial DNA from raw milk and raw milk cheese. J. Appl. Microbiol. 113, 96–105. https://doi.org/10.1111/j.1365-2672.2012.05294.x

Rognes, T., Flouri, T., Nichols, B., Quince, C., Mahé, F., **2016**. VSEARCH: a versatile open source tool for metagenomics. PeerJ 4, e2584. https://doi.org/10.7717/peerj.2584

Schloss, P.D., **2020**. Reintroducing mothur: 10 Years Later. Appl. Environ. Microbiol. 86. https://doi.org/10.1128/AEM.02343-19

Wang, Q., Garrity, G.M., Tiedje, J.M., Cole, J.R., **2007**. Naive Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. Appl. Environ. Microbiol. 73, 5261–5267. https://doi.org/10.1128/AEM.00062-07

Annexes

1. Extraction d'ADN à partir d'échantillon de fromage (méthode phénol/chloroforme)
2. Protocole d'amplification par PCR

Annexe 1 : Extraction d'ADN à partir d'échantillon de fromage (méthode phénol/chloroforme)

1. Matériel nécessaire

- Tubes de broyage 2 ml à vis PCR grade (réf. 72694406 Sarstedt)
- Billes de broyage zirconium 0.1 mm (réf : 11079101 fournisseur : BioSpec)
- Billes de broyage zirconium 0.5 mm (réf : 11079105 fournisseur : BioSpec)
- Vortex
- Bain-Marie
- Sorbonne
- Broyeur : Precellys Evolution de Bertin (fournisseur : Ozyme)
- Bain-Marie à sec avec bloc pour tubes de 2 mL
- Centrifugeuse de paillasse
- Tubes PhaseLockGel de 2 mL (fournisseur : Quantabio)
- Tubes Eppendorf de 2 mL et 1.5 mL
- Kit Genomic DNA Clean & Concentrator-10 (Zymo research)

2. Solutions à préparer

- Tris-HCl 1 M pH7.8
- Solution GT = Thiocyanate de guanidium 4M, Tris-HCl 0.1 M pH 7.8
- Solution LS = N-Lauryl sarcosine 10% (w/v)
- TES = Tris 50 mM – EDTA 1 mM – Sucrose 6.7 % pH8
- Mélange lysosyme-lyticase-TES = lysozyme (40 mg/mL) + lyticase (1000 U/mL) dans du TES
- Sodium Dodecyl Sulfate (SDS) 20% (w/v)
- Protéinase K 14 mg/mL
- Tampon SP = 47.35 mL de Na2HPO4.12H2O 0.2M + 2.65 mL de NaH2PO4.H2O 0.2M + 50 mL d'eau
- Tampon AE = acétate de sodium 50 mM, EDTA 10 mM
- RNaseA 20 mg/mL

3. Mode opératoire

3.1. Lyse :

- Placer 250 mg de fromage dans un tube de broyage de 2 mL à vis contenant aussi 200 mg du mélange de billes de zirconium (mélange 50/50 de diamètres 0.1 et 0.5 mm)
- Ajouter : 250 µL de solution GT + 40 µL de solution LS + 200 µL de tampon SP + 200 µL de tampon AE
- Agiter au broyeur Precellys Evolution 1 fois 20 secondes à vitesse 6500 m/s afin d'homogénéiser le mélange
- Ajouter 75 µL de mélange lysosyme-lyticase-TES
- Mélanger au vortex puis incuber à 37°C (Bain-Marie) pendant 30 min, mélanger au vortex au bout de 15 min
- Ajouter 40 µL de Protéinase K (14 mg/mL) et 100 µL de SDS 20 %
- Mélanger au vortex puis incuber à 55°C (Bain-Marie à sec) pendant 30 min, mélanger au vortex au bout de 15 min

3.2. Extraction

- Se placer sous une Sorbonne et ajouter 500 mL phénol-chloroforme équilibré à pH 8
- Broyer 45 secondes à vitesse 10000 m/s (broyeur Precellys Evolution)
- Incuber 2 min à 55°C (Bain-Marie à sec placé sous la Sorbonne)
- Refroidir dans de la glace mélangée d'eau
- Broyer de nouveau 45 secondes à vitesse 10000 m/s (broyeur Precellys Evolution)
- Incuber 2 min à 70°C (Bain-Marie à sec placé sous la Sorbonne)
- Centrifuger 30 min à 14000 x g à température ambiante
- Récupérer la phase aqueuse supérieure dans un tube PhaseLockGel de 2 mL (préalablement centrifugé 30 sec), et ajouter 500 µL de phénol-chloroforme équilibré à pH 8
- Mélanger par quelques retournements vigoureux
- Centrifuger 5 min à 14000 x g à température ambiante
- Récupérer la phase aqueuse supérieure dans un nouveau tube PhaseLockGel de 2 mL (préalablement centrifugé 30 sec), et ajouter 500 µL de chloroforme
- Centrifuger 5 min à 14000 x g à température ambiante
- Récupérer la phase aqueuse (minimum 500 µL) dans un tube Eppendorf de 2 mL
- Ajouter 2 µl de RNase A (20 mg/mL), incuber 30 min à 37°C (Bain-Marie à sec)

3.3. Purification (kit Genomic DNA Clean & Concentrator-10)

- Ajouter 2 volumes de ChIP DNA Binding Buffer pour 1 volume d'échantillon (phase aqueuse récupérée en fin d'extraction)
- Transférer le mélange dans la colonne Zymo-Spin IC-XL elle-même placée dans le tube collecteur (1 mL maximum à la fois, si besoin recommencer plusieurs fois afin de faire passer l'ensemble de l'échantillon)
- Centrifuger 30 secondes à 14000 x g à température ambiante
- Eliminer l'éluat
- Ajouter 200 µL de DNA Wash Buffer sur la colonne
- Centrifuger 1 min à 14000 x g à température ambiante
- Eliminer l'éluat
- Ajouter à nouveau 200 µL de DNA Wash Buffer sur la colonne
- Centrifuger 1 min à 14000 x g à température ambiante
- Eliminer l'éluat et transférer la colonne sur un tube Eppendorf de 1.5 mL
- Ajouter 40 µL d'eau directement sur la membrane au fond de la colonne et incuber 1 min à température ambiante
- Centrifuger 20 secondes à 14000 x g à température ambiante pour éluer l'ADN
- Déterminer la concentration d'ADN au Qubit puis stocker à 4°C à court terme (quelques jours à quelques semaines) ou à -20°C à long terme.

Annexe 2 : Protocole d'amplification par PCR

Procaryotes (régions V3-V4 du gène codant pour l'ARNr 16S) :

16SV3F 5'-ACGGRAGGCWGCAG-3'
16SV4R 5'-TACCAGGGTATCTAATCCT-3'

Mix PCR:

ADN	**10.0**	**ng**
Buffer	5.0	µL
dNTP mix (10 mM)	1.0	µL
Forward Primer (10 µM)	1.0	µL
Reverse Primer (10 µM)	1.0	µL
Taq polymerase	0.5	µL
H2O	q.s.p 50.0	µL

Programme PCR : 94°C - 60s ; 30 cycles (94°C - 60s; 65°C - 60s ; 72°C - 60s) ; 72°C - 10min

Eucaryotes (ITS2) :

ITS3f 5'-GCATCGATGAAGAACGCAGC-3'
ITS4_KYO1 5'-TCCTCCGCTTWTTGWTWTGC-3'

PCR Mix:

ADN	**10.0**	**ng**
Buffer	5.0	µL
dNTP mix (10 mM)	1.0	µL
Forward Primer (10 µM)	1.0	µL
Reverse Primer (10 µM)	1.0	µL
Taq polymerase	0.5	µL
H2O	q.s.p 50.0	µL

Programme PCR : 94°C - 60s;30 cycles (94°C - 60s; 65°C - 60s ; 72°C - 60s); 72°C - 10min

Remarque : des adaptateurs pour le séquençage doivent être rajoutés en 5' sur les séquences des amorces avant de réaliser l'amplification. Il faut se rapprocher de la plateforme de séquençage ou du prestataire pour les connaître au préalable et commander les amorces complètes (adaptateur suivi de l'amorce).

Remerciements

Cet ouvrage a été réalisé dans le cadre du **projet ADAMOS**, porté par CERAQ, dont nous remercions l'ensemble des partenaires.

Ce projet a bénéficié de financements du Ministère de l'agriculture et de l'alimentation (CASDAR) et des conseils départementaux de Savoie et Haute-Savoie (Plan filière lait cru), que nous remercions pour leur soutien.

Le projet ADAMOS est affilié au **RMT "Filières fromagères valorisant leur terroir".** L'objectif de ce réseau est d'accompagner, par la R&D, les filières fromagères valorisant leur terroir dans les changements auxquels elles doivent faire face : les évolutions tant sociétales que règlementaires, les innovations technologiques ou encore le dérèglement climatique. Le RMT favorise l'échange et l'émergence de projets entre les acteurs de la recherche et du développement et les filières fromagères de terroir. Il permet ainsi de mettre en débat, de faire-valoir et de cultiver les fondamentaux sur lesquels ces filières s'appuient et fondent leur différenciation. Ses travaux portent notamment sur l'étude des microflores des laits et des fromages, la gestion de la ressource herbagère, les savoir-faire traditionnel et la durabilité des filières.

Cet ouvrage est le fruit d'un travail collectif, coordonné par Eric Dugat-Bony (INRAE).

Tous nos remerciements aux personnes ayant contribué à la préparation de cet ouvrage :

Eric Dugat-Bony (INRAE), Blandine Polturat (CERAQ), Céline Delbès (INRAE), Christine Achilleos (INRAE), Cresciense Lecaudé (CERAQ), Hélène Tormo (INP PURPAN), Marion Dalmasso (Université de Caen), Nicolas Orieux (ENILV), Sarah Chuzeville (ACTALIA), Sébastien Theil (INRAE), Yvette Bouton (CIGC).

Contacts : *eric.dugat-bony@inrae.fr ; blandine.polturat@ceraq.fr*

Avec la contribution financière du compte d'affectation spéciale développement agricole et rural CASDAR

MINISTÈRE DE L'AGRICULTURE ET DE L'ALIMENTATION
Liberté
Égalité
Fraternité

Les méthodes omiques sont des outils importants pour mieux comprendre les communautés microbiennes des laits et des fromages. Parmi elles, la métagénétique est désormais considérée comme une méthode de routine dans les laboratoires de recherche.

Ce guide a pour objectif de favoriser l'appropriation de cette méthode dans les filières fromagères en proposant des conseils méthodologiques depuis la préparation des échantillons jusqu'à l'analyse statistique des données.

ADAMOS

INRAE