Mémoire

Auteur : Corato, Amélie
Promoteur(s) : Meyer, Patrick
Faculté : Faculté des Sciences
Diplôme : Master en bioinformatique et modélisation, à finalité approfondie
Année académique : 2021-2022
URI/URL : http://hdl.handle.net/2268.2/15423

Université de Liège
Faculté des Sciences
Janvier 2022

Utilisation de sources de carbone organique chez les microalgues: analyses de données RNA-Seq

Amélie Corato

Promoteur: **Patrick Meyer**

Co-Promotrice: **Claire Remacle**

Mémoire en vue de l'obtention du grade

Master en Bioinformatique et Modélisation

Département des Sciences de la Vie

Laboratoire de Génétique et Physiologie des microalgues

UR InBioS

Remerciements

Tout d'abord, je tiens à remercier mon promoteur, Patrick Meyer pour m'avoir poussée à réaliser ce master en bioinformatique. Sans vous, je n'aurais probablement pas eu l'idée de me lancer dans un nouveau master. Merci également à vous pour votre suivi tout au long du master pendant ces 2 années et demi.

Je tiens ensuite à remercier ma co-promotrice, Claire Remacle. Merci à vous pour votre implication dans mon travail pendant cette année riche en rebondissements. Merci également d'avoir cru en moi et de m'avoir poussée à terminer ce master alors que ma thèse était finie.

Un grand merci également à Marc Hannikenne pour son aide et ses conseils dans le traitement des données RNA-Seq, mais aussi pour son cours de "practical genomics" si précieux.

Toute cette aventure n'aurait pas été la même sans les personnes présentes, de près ou de loin. Merci aux personnes du Laboratoire de Génétique et Physiologie des microalgues, dont notamment Pablo, Edmée et Tom, pour la bonne humeur, les temps de midi et soirées partagés.

Je tiens à faire un remerciement spécial à mon binôme de travaux, ma bioinformaticienne de coeur, Pauline. Merci pour toute ton aide, ton soutien, tes corrections, tes encouragements, tes messages et nos moments partagés. Sans toi, je n'aurais probablement pas tenu le coup aussi bien et autant apprécié ce master.

Un tout grand merci aussi à mes amis et mes parents pour le soutien qu'ils m'ont apporté alors que je me lançais dans cette aventure un peu folle. Un merci particulier à Pierre pour ta présence au quotidien et tes encouragements.

" Un programme informatique fait ce que vous lui avez dit de faire,
pas ce que vous voulez qu'il fasse. "

Loi de Murphy

Amélie Corato Laboratoire de Génétique et Physiologie des microalgues

Promoteur: Patrick Meyer

Co-Promotrice: Claire Remacle Janvier 2022

Utilisation de sources de carbone organique chez les microalgues: analyses de données RNA-Seq

Les microalgues sont, à l'heure actuelle, principalement utilisées en tant que source importante de biomasse et de biomolécules à haute valeur ajoutée. La croissance hétérotrophe de microalgues se déroule à l'obscurité en présence d'une source de carbone organique et gagne de plus en plus d'intérêt en industrie. En effet, elle permet notamment d'atteindre une productivité volumétrique bien supérieure à l'autotrophie. Dans ce contexte est né le projet DARKMET dont une des parties vise à obtenir une meilleure compréhension du métabolisme algal en hétérotrophie en utilisant différentes sources de carbones organiques. Pour ce faire, la microalgue rouge poly-extrémophile *Galdieria sulphuraria* 074W a été étudiée et sa croissance hétérotrophe a montré une différence dans sa pigmentation selon la source de carbone organique utilisée, *i.e.* elle reste verte-bleue en présence de glycérol et devient jaune-pâle en présence de glucose. Cette observation phénotypique intéressante nous a poussé à étudier son métabolisme hétérotrophe plus en détail et de réaliser un pipeline d'analyse complet de l'obtention de données RNA-Seq jusqu'à l'identification de gènes différentiellement exprimés.

Galdieria sulphuraria 074W a donc été cultivée en hétérotrophie dans 6 conditions d'ajout de source de carbone organique différentes. Les ARNs ont été extraits et séquencés via un séquençage Illumina. La qualité des séquences ainsi obtenues a été analysée graĉe à au programme *FastQC*, et le logiciel *Trimmomatic* a permis une filtration des séquences afin d'en augmenter la qualité finale. Les séquences filtrées ont ensuite été alignées sur le génome de référence de *G. sulphuraria* en utilisant *HiSat2* et les alignements ainsi obtenus ont été comptés grâce à *Htseq-count*. Enfin, des analyses statistiques ont été menées sur ces comptes grâce au package R *DESeq2*. Bien que les données utilisées dans ce mémoire soient très hétérogènes au sein des réplicats biologiques, le pipeline d'analyse RNA-Seq a été mené à son terme et pourra être réutilisé sur de futures données de séquençage.

Table des matières

1 Introduction

1.1 Contexte de l'étude RNA-Seq: Le projet DARKMET

Les microalgues sont des organismes unicellulaires eucaryotes ou procaryotes photosynthétiques produisant des métabolites exploitables et valorisables. A l'heure actuelle, elles trouvent principalement leur intérêt en tant que source importante de biomasse (pour les industries de la nutraceutique ou du biodiesel) et de biomolécules à haute valeur ajoutée telles que les caroténoïdes et acides gras [17]. Depuis les années 1950, la majorité de la production de microalgues est réalisée en autotrophie, qui est un mode de croissance basé sur une assimilation du CO_2 dépendante de la lumière. Bien que majoritaire, ce mode de croissance présente de nombreux désavantages dont principalement: une productivité biomassique restreinte et une limitation des vitesses de division cellulaire causées par la faible disponibilité de la lumière. De plus, les systèmes de cultures pour ces croissances sont onéreux à construire et à entretenir [35]. Dès lors, depuis quelques dizaines d'années, de plus en plus d'études se tournent vers un mode de croissance à l'obscurité totale et en présence d'une source de carbone organique: l'hétérotrophie. Du fait de l'absence de limitation par la lumière, la productivité volumétrique atteinte par ce type de cultures est plus importante, ce qui le rend donc intéressant d'un point de vue économique [23, 6]. Bien que l'hétérotrophie soit une condition de croissance très prometteuse pour l'industrie, les connaissances fondamentales sur le métabolisme hétérotrophe des microalgues sont encore insuffisantes. Ces connaissances sont néanmoins importantes, non seulement, pour optimiser les vitesses de croissance des microalgues à l'obscurité mais également pour trouver des conditions permettant d'orienter le métabolisme algal vers la production de produits valorisables.

C'est dans ce contexte qu'est né le projet "DARKMET" (*i.e.* "Metabolism of microalgae in heterotrophy"). Ce projet a pour but d'améliorer la production hétérotrophe de composés valorisables excrétés ou non par certaines souches de microalgues. La première étape de ce projet est donc d'obtenir une meilleure compréhension du métabolisme algal hétérotrophe en utilisant différents substrats carbonés tels que le glucose, le glycérol

ou encore des hydrolysats de biomasse lignocellulosique, une ressource renouvelable facilement disponible et étant un déchet de diverses industries [22, 14]. En effet, une des problématique importante de la croissance hétérotrophe est le choix du substrat carboné. A l'heure actuelle, le glucose est le sucre à 6 carbones le plus souvent utilisé tant à grande échelle qu'en laboratoire. D'autres substrats carbonés tels que l'acétate, plus petit dérivé carboné métabolisé par une grande majorité de microalgues, et le glycérol, produit dérivé de beaucoup d'industries telles que celle de la canne à sucre, sont également très souvent utilisés. Cependant, par soucis tant environnemental qu'économique, des ressources renouvelables ou dérivées d'autres industries doivent être explorées.

Un deuxième volet de ce projet concerne la production de composés valorisables par les souches sélectionnées. Les composés ciblés lors des croissances hétérotrophes des microalgues incluent généralement les acides gras, les lipides et les pigments. Cependant, une recherche de produits à plus haute valeur ajoutée, tels que de nouveaux composés bioactifs, est nécessaire pour rendre ces cultures hétérotrophes compétitives. Ainsi, le projet DARKMET s'est intéressé à la production de nouvelles molécules antibactériennes ou antifongiques bioactives produites par les microalgues sélectionnées, de sorte à en faire une nouvelle voie de valorisation de la biomasse. D'autre part, le projet s'est aussi intéressé à l'étude plus approfondie de la croissance hétérotrophe des microalgues sélectionnées et notamment de la souche *Galdieria sulphuraria*. C'est dans ce deuxième pan du projet que s'inscrit ce mémoire.

1.2 *Galdieria sulphuraria*

1.2.1 *G. sulphuraria*, une microalgue poly-extrémophile

Galdieria sulphuraria est une microalgue rouge acidophile et thermophile, décrite pour la première fois en 1981 par Merola et son équipe [30]. Comme d'autres algues appartenant aux Cyanidiales, elle vit dans des habitats chauds (jusqu'à 50°C) et acides (pH compris entre 0.05 et 3), tels que des volcans ou encore des sources chaudes acides [33, 43]. Les Cyanidiales s'étant adaptées à des environnements extrêmes, ce sont des organismes eucaryotes possédant des caractéristiques physiologiques assez inhabituelles telles que, par exemple, la résistance à un large éventail d'ions métalliques toxiques [2, 1].

Alors que la majorité des rhodophytes (*i.e.* algues rouges) se développent sous un mode de croissance strictement autotrophe, *Galdieria sulphuraria* a la particularité de croître dans 3 modes différents: autotrophie, hétérotrophie et mixotrophie [19, 20, 32]. L'autotrophie est un mode de croissance dans lequel la seule source d'énergie est la lumière avec le CO_2 comme source de carbone minéral. L'hétérotrophie, quant à elle, est un mode de croissance sans apport de lumière (*i.e.* à l'obscurité totale) nécessitant la présence de constituants organiques carbonés assimilables. Enfin, la mixotrophie consiste en une croissance sous une lumière modérée en présence d'une source de carbone organique. Dans ce cas, l'assimilation photosynthétique du carbone inorganique et la transformation du carbone organique peuvent avoir lieu en même temps. Ainsi, lors d'une croissance hétérotrophe, il a été montré que *G. sulphuraria* est capable d'assimiler plus de 50 sources de carbone organique différentes telles que des sucres, des alcools de sucres, des intermédiaires du cycle des acides tricarboxyliques, et des acides aminés [2, 33, 19, 43]. Gross et son équipe ont également montré que *Galdieria* ne semble pas avoir de préférence pour un groupe spécial de sucres tels que les sucres en conformation L/R ou encore uniquement des sucres avec le même nombre d'atomes de carbone [19]. Ainsi, sa croissance en hétérotrophie est une caractéristique forte démontrant sa grande flexibilité métabolique et fait de cette microalgue un organisme idéal pour l'étude des enzymes impliqués dans le catabolisme des sucres rares et du métabolisme des cellules végétales au sens large [19, 2].

Grâce à leurs capacités métaboliques uniques, les organismes extrémophiles tels que *G. sulphuraria* gagnent de plus en plus d'intérêt dans l'industrie biotechnologique et énergétique [33]. En effet, à l'heure actuelle, les études concernant *G. sulphuraria* se multiplient [13] et concernent, par exemple, l'étude de son métabolisme hétérotrophe [3, 5, 40], sa capacité à bio-éliminer ou accumuler des métaux lourds tels que le palladium ou encore des complexes de platine [46, 1], sa croissance sur des substrats tels que les eaux usées [15, 47] ou encore sa production de lipides utiles dans les biofuels [44].

1.2.2 La souche *G. sulphuraria* 074W

Lors d'une étude sur la croissance hétérotrophe d'une souche de *Galdieria sulphuraria* en 1995, Gross et son équipe ont constaté que leurs résultats concernant la croissance et la pigmentation de la souche n'étaient pas reproductibles [19]. En cultivant la souche sur une boîte de Pétri supplémentée en 25 mM de glucose, ils ont constaté qu'elle était en

fait constituée d'un mélange de deux souches. En effet, en les conservant à l'obscurité, certaines colonies restaient vertes alors que d'autres devenaient plus pâles, voire jaunes. Il ont alors décidé d'isoler ces colonies et ont nommé les vertes "la souche 074G", et les plus pâles "la souche 074W". Excepté cette différence de pigmentation dans le noir, les 2 clones sont très similaires au niveau de leur morphologie, de leur taille et de leur croissance [19]. De plus, alors que les 2 souches montrent, en hétérotrophie, des contenus en chlorophylles différents, aucune différence dans leur pigmentation ni croissance n'a été constatée en autotrophie [19, 18]. Chez *G. sulphuraria*, il a également été rapporté qu'à l'obscurité, le glucose semble, non seulement, sous-réguler le nombre de membranes thylakoïdiennes et de pigments photosynthétiques produits, mais également réprimer la synthèse de la phycocyanine (complexe pigmentaire intéressant) [33, 43].

Depuis lors, le métabolisme de *Galdieria sulphuraria* 074W a été bien étudié et sa génomique a également émergé avec le séquençage et l'assemblage de son génome jusqu'au niveau des scaffolds (plus d'informations sur ce génome, *cf.* 2.2) [13, 41].

1.2.3 Etudes génomiques chez *G. sulphuraria*

Les génomes de 11 souches de *Galdieria sulphuraria* ont été séquencés et assemblés. La taille médiane de tous ces génomes est de 14,3176 Mbp, le nombre de protéines médian est de 7174 protéines et le %GC médian de 39,2%. Les études sur *G. sulphuraria* et son génome se sont ainsi multipliées et concernent différents domaines comme, par exemple, l'étude de ses gènes probablement acquis par transfert horizontal de gènes (HGT), qui est le processus d'intégration du matériel génétique provenant d'un autre organisme sans en être le descendant [38]. En effet, l'analyse des génomes de 10 Cyanidiales a permis à Rossoni et son équipe de déterminer qu'environ 1% des gènes de leur génome ont été acquis par HGT [37]. Par ailleurs, d'autres études se concentrent sur l'analyse des mécanismes menant à l'adaptation de *G. sulphuraria* à des environnements extrêmes et les conséquences fonctionnelles de l'extrémophilie sur les génomes mitochondrial et plastidial [21, 10, 2, 9]. Barbier et son équipe ont, en effet, identifié des gènes uniques et essentiels à la flexibilité métabolique de *G. sulphuraria*, dont notamment, des gènes codant pour des transporteurs membranaires et des enzymes du métabolisme des carbohydrates [2]. Enfin d'autres études se portent, non seulement, sur l'analyse plus globale du génome de *G. sulphuraria* et dans l'identification de voies métaboliques [48], mais également dans

l'analyse de l'évolution de son génome [36, 9]. En effet, Qiu et son équipe ont récemment identifié, dans le génome de *Galdieria sulphuraria*, 155 gènes associés à la machinerie splicéosomale qui étaient également présents de manière putative chez l'ancêtre commun [36].

1.2.4 L'étude de *G. sulphuraria* 074W dans le cadre du projet DARKMET

La souche de *G. sulphuraria* utilisée dans l'étude RNA-Seq de ce manuscrit provient de la collection d'algues de l'université Frédérique II de Naples (Algal Collection University Federico II, ACUF). C'est un isolat du Mont Lawu de l'île de Java (Indonésie), de type 074. Etant donné qu'elle blanchit en présence de glucose en conditions de croissance hétérotrophe, elle est appelée 074W (*cf.* 1.2.2). Les détails d'expérience et observations ayant mené à l'élaboration de l'étude RNA-Seq vont être décrites dans la suite de ce paragraphe, alors que les détails techniques de l'expérience RNA-Seq seront présentés dans la partie Matériel (*cf.* 2.1).

Etant donné sa capacité à croître en hétérotrophie, la souche de *G. sulphuraria* a ainsi été mise en culture à l'obscurité, sur deux sources de carbone organique: le glucose et le glycérol.Ces deux sucres sont les éléments limitants de la croissance hétérotrophe de *G. sulphuraria* 074W (l'azote et le phosphore n'étant pas épuisés en fin de croissance). Lors de la mise en

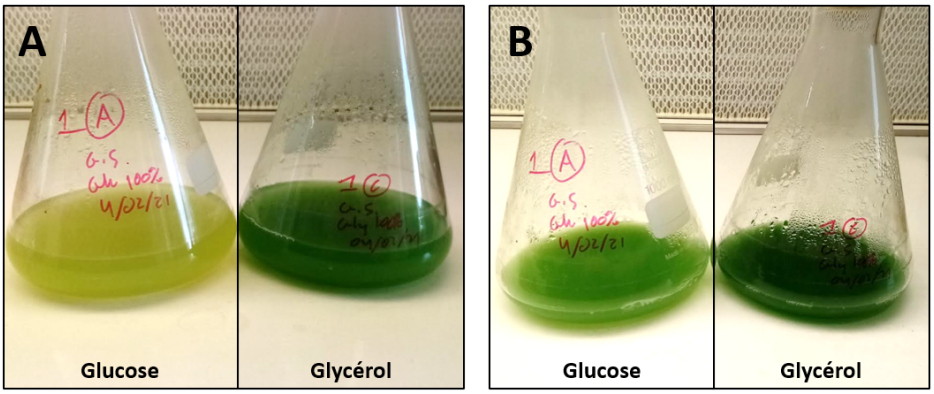

Figure 1.1: Photographies de cultures hétérotrophes de *Galdieria sulphuraria* 074W sur glucose et glycérol. A – Après 4 jours à l'obscurité. Les sources de carbone organique sont encore présentes dans le milieu. B – Après 9 jours à l'obscurité. Le glucose du milieu est épuisé et les microalgues sous glucose retrouvent une couleur plus verte.

culture de la deuxième génération (*i.e.* premier repiquage après un passage de conditions autotrophes à hétérotrophes) et après quelques jours de croissance, l'observation majeure et directe réalisée était la perte de coloration de *Galdieria* cultivée sur glucose (Figure 1.1). En effet, cette culture est devenue rapidement "jaune" alors que la culture sur glycérol maintenait une belle couleur verte-bleutée (Figure 1.1). Ainsi, et ce pour toutes les cultures faites depuis lors, tant que du glucose est présent dans le milieu, le contenu en chlorophylles est équivalent à environ 40% de celui en présence de glycérol seul. Lorsque la source de carbone est épuisée, le contenu en pigments augmente jusqu'à atteindre une valeur égale à celle est des cultures sur glycérol. Il semblerait donc que chez *Galdieria sulphuraria* 074W, le glucose agisse comme répresseur de la biosynthèse de chlorophylles, comme l'ont déjà mentionné Stadnichuk et son équipe en 1998 [45]. Excepté cette différence de pigmentation, la production des autres métabolites d'intérêt (acides gras et glycogène) ne semble pas être affectée par la présence d'une source de carbone organique ou l'autre.

Cette différence phénotypique entre les conditions "glucose" et "glycérol" ainsi que la possible répression de la biosynthèse de chlorophylles, nous ont menés à étudier plus en détail, non seulement, les mécanismes métaboliques impliqués dans la perte de pigments induite par le glucose, mais aussi de considérer une analyse RNASeq. Cette dernière nous permettra d'analyser les variations dans le transcriptome des cellules suivant l'ajout de glucose dans le milieu et la diminution du contenu en pigments.

1.3 Analyse des données RNA-Seq

Comme mentionné ci-dessus, l'analyse de l'expression différentielle de gènes chez *Galdieria sulphuraria* 074W, cultivée en hétérotrophie sur deux sources de carbone organique distinctes, a été choisie, de sorte à mieux comprendre les variations phénotypiques observées lors de sa croissance sur glucose. Cette approche permet de mettre en évidence certains gènes sous- ou sur-exprimés selon différentes conditions, qui sont, dans notre cas, les ajouts successifs de glucose et glycérol à une culture hétérotrophe. Pour la réalisation de ce type d'analyses, c'est la méthode "RNA-seq" qui a été choisie. En effet, c'est une technique très puissante pour les études transcriptomiques et elle permet d'abolir certaines limitations qui étaient présentes dans les méthodes antérieures (*e.g.* les microarrays, la PCR), telles que, par exemple, la nécessité préalable de connaître l'organisme [34, 12]. Les différentes

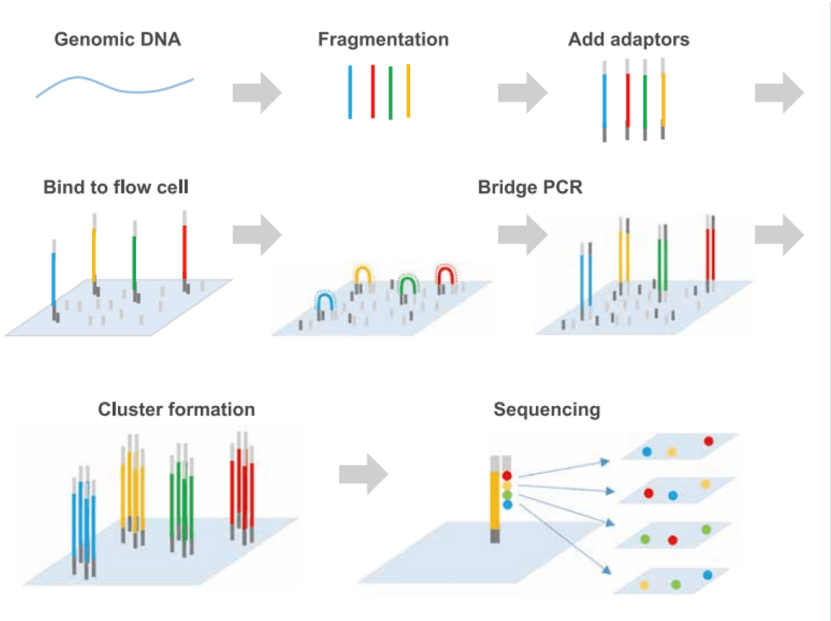

Figure 1.2: Les différentes étapes du séquençage Illumina.

étapes de la méthode RNA-Seq vont donc être décrites dans la suite de cette partie.

1.3.1 La préparation des échantillons pour le séquençage

Les ARNs sont tout d'abord extraits des échantillons et la préparation du matériel génétique pour le séquençage de *Galdieria sulphuraria* (Cyanidiales) commence par la création d'une libraire d'ADN complémentaires (ADNc) sur base des transcrits présents dans les échantillons [34, 28]. Pour ce faire, les ARN messagers (ARNm) sont sélectionnés grâce à une bille sur laquelle sont fixés des oligonucléotides poly(T). Les ARNm sont ensuite fragmentés et rétro-transcrits en ADN en utilisant des amorces aléatoires. Les ADNc ainsi obtenus sont donc pourvus, à leurs deux extrémités, d'une queue poly(A) et sont ainsi prêts à être séquencés (Figure 2.1).

1.3.2 Le séquençage de nouvelle génération Illumina

Depuis son introduction en 2007, le séquençage utilisé dans le cadre des analyses RNA-Seq est majoritairement le séquençage Illumina (Figure 1.2) [29, 42]. Cette méthode est basée sur un système de séquençage par synthèse, où la terminaison est cyclique et réversible. Suite à la première étape décrite ci-dessus (*cf.* 1.3.1), l'ADN est lié, à ses extrémités, à des adaptateurs d'oligonucléotides complémentaires à d'autres, qui sont, eux, présents à la surface d'une "flowcell", *i.e.* une plaque de verre où les réactions de séquençage et la

visualisation des nucléotides incorporés ont lieu. Les brins d'ADN vont ainsi se lier aux adaptateurs complémentaires de la surface et vont être amplifiés par PCR sous forme de ponts. Une dénaturation sépare ensuite les deux brins et une amplification complémentaire à lieu. Ce processus se répète de nombreuses fois, ce qui donne naissance à des "clusters", *i.e.* un groupe d'un millier de séquences identiques au brin matrice d'origine, qui vont être séquencés (Etape "Bridge PCR" sur la Figure 1.2).

Le séquençage des clusters va donc être réalisé sur un des deux brins et commence par l'ajout d'un oligonucléotide. Quatre bases d'ADN, possédant un fluorophore lié à leur extrémité 3', et une ADN polymérase sont donc ajoutées dans la flowcell. L'ADN polymérase incorpore la base appropriée et la surface est lavée des bases non-incorporées. Ensuite, une image de la surface est prise et la base est identifiée grâce à son groupement fluorophore. Ce groupement est clivé, ce qui permet de libérer l'extrémité 3' afin qu'une nouvelle base et son fluorophore soient fixés. Le cycle recommence ainsi un certain nombre de fois déterminant la longueur des séquences [26].

Une technique utilisée afin d'obtenir des informations supplémentaires est le séquençage "paired-end" (pairé) (Annexe A1 - Figure A1.1). Dans ce type de séquençage, deux adaptateurs avec deux amorces de séquençage différentes sont utilisés. Une première amplification en pont va avoir lieu et un des deux brins va être séquencé. Le cluster est ensuite régénéré en conservant l'autre brin pour un second séquençage. Cette méthode permet donc, non seulement, d'obtenir deux séquences amplifiées provenant du même fragment, mais aussi qu'elles soient à une distance semblable des autres fragments. Le séquençage pairé permet donc, non seulement, de positionner plus efficacement les différentes séquences les unes par rapport aux autres qu'un séquençage dit "non-pairé", mais également de réduire les problèmes de multi-alignement [34].

Le séquençage Illumina permet ainsi d'obtenir des millions voir des dizaines de millions de séquences, aussi appelées "reads". Afin d'utiliser les données RNA-Seq pour, par exemple, la comparaison de niveaux d'expression génique, l'information donnée par ces millions de séquences va devoir être "convertie" et la première étape consiste en leur alignement[8].

1.3.3 L'alignement des séquences obtenues

Cette étape est un des processus les plus couramment utilisé sur des données de séquençage à haut débit, dont celles obtenues par Illumina [34]. Le but majeur de l'alignement des séquences est leur cartographie afin de trouver l'emplacement où chaque read correspond le mieux à la référence, tout en tenant compte des erreurs et variations (*e.g. indels*, SNPs).

L'alignement post-séquençage peut être réalisé de plusieurs façons différentes en utilisant, soit un génome de référence et des modèles de gènes annotés existants pour l'espèce concernée, soit un génome de référence d'une espèce proche ou encore, la reconstruction des transcrits à partir des séquences issues d'un assemblage *de novo* [11, 29]. De ces différentes techniques, seul l'alignement sur base d'un génome de référence va être détaillé ici car pertinent pour la suite du manuscrit.

Ce type d'alignement est principalement réalisé via l'utilisation d'un programme sensible à l'épissage, prenant en compte la présence d'exons dans les séquences génomiques et permettant, si nécessaire, d'interrompre l'alignement entre deux exons. Ces programmes permettent également de détecter l'expression de séquences introniques et ainsi de laisser de gros espacements entre les introns [8]. Des transcrits ou variants non-décrits auparavant peuvent donc être détectés par ce type d'alignement. Un désavantage majeur de cette technique est cependant sa limitation aux espèces dont un génome de référence de bonne qualité est disponible. En effet, la présence d'erreurs de séquençage ou d'annotations peuvent affecter l'alignement [8].

Presque tous les programmes d'alignement de séquences courtes utilisent une stratégie en 2 étapes [34]. Premièrement, des techniques "heuristiques" sont utilisées. Elles permettent une recherche rapide donnant une liste réduite de localisations possibles du read sur le génome de référence. Ensuite, des algorithmes d'alignement plus lents et plus précis sont utilisés sur le sous-ensemble de possibilités préalablement déterminé [16]. Les aligneurs actuels permettent une correspondance heuristique rapide en utilisant soit des tables de hachage soit des transformées de Burrows-Wheeler (BWT) [16, 29, 8]. Les méthodes basées sur les tables de hachage sont rapides dans le classement des séquences candidates et peuvent détecter des différences complexes entre la référence et la lecture. Les aligneurs BWT, quant à eux, peuvent aligner des séquences très proches de la référence de façon

très efficace, mais sont néanmoins moins sensibles que les tables de hachage. Ils se basent sur des index FM qui sont des fichiers index compressés basés sur la transformée de BW et couvrant l'ensemble du génome. En combinaison avec d'autres stratégies, ils permettent ainsi de déterminer efficacement la localisation et le nombre d'occurrences d'un motif. Différents programmes utilisent cette technique, c'est le cas de *TopHat* ou *HiSat2*, développé plus récemment.

1.3.4 Le compte des séquences alignées

Pour les analyses RNA-Seq, le niveau d'expression des gènes ou des transcrits est relié au nombre de séquences alignées sur eux. En effet, si la différence observée dans le compte d'un read pour un gène/un transcrit entre deux conditions expérimentales différentes est statistiquement significatif, ce gène/ce transcrit peut être considéré comme exprimé différentiellement dans les données RNA-Seq [8]. C'est donc pour cette raison que les séquences venant d'être alignées sur le génome vont être comptées.

Cette étape de comptage, relativement basique, est réalisée par différents programmes dont un des plus connus et utilisé dans la suite de ce manuscrit: *Htseq-count*. Le choix du programme dépend de plusieurs paramètres tels que si l'alignement est pairé ou non, ou encore de la ploïdie de l'organisme. Ce programme nécessite en entrée, d'une part, le fichier d'annotation et, d'autre part, le fichier obtenu à la fin de l'alignement. Le fichier d'annotation peut se trouver sous différents formats ayant différentes versions, dont les plus connus sont les formats: GFF ("General Feature Format") et GTF ("Gene Transfert Format"). Ce fichier contient les coordonnées et attributs de gènes, transcrits et autres motifs (*e.g.* les codons START/STOP ou les séquences codantes) [49]. Le fichier obtenu à la fin de l'alignement, quant à lui, est sous le format SAM ("Sequence Alignment/Map") [49, 24]. Celui-ci contient les données de séquençage ainsi que des informations relatives à l'alignement telles que sa cartographie ou sa qualité. Ce type de fichier peut être converti sous un format BAM ("Binary Alignment/Map") étant une version compressée et binaire du fichier SAM [24]. Ces formats SAM/BAM sont des formats standards pour l'alignement de reads courts, ce qui en facilite l'analyse et la lecture [29]. Ainsi, le programme de comptage va produire, grâce à ces deux fichiers, une matrice reprenant les comptes des séquences alignées pour chaque gène, qui pourra être ensuite utilisée pour des analyses telles que l'expression différentielle de gènes.

1.3.5 Les analyses d'expression différentielle de gènes

Une fois les séquences alignées sur le génome de référence et comptées, il est possible de mesurer l'expression différentielle des gènes (DEG). Le but d'une analyse d'expression différentielle est de mettre en évidence des gènes dont l'abondance change significativement entre différentes conditions expérimentales. Ainsi, en général, cela signifie d'utiliser une table résumant les comptes de séquences pour chaque condition et de réaliser des tests statistiques entre les échantillons d'intérêt [34].

Afin d'effectuer ce type d'analyse, le package R *DESeq2* est utilisé [25]. Dans ce package, les DEG sont détectés en utilisant une approche paramétrique, *i.e.* en utilisant des modèles linéaires généralisés à distribution binomiale négative (GLM) [8, 12]. *DESeq2* va ainsi d'abord modéliser les données pour estimer les coefficients du modèle et ensuite déterminer y pour chaque x, c'est-à-dire, l'expression différentielle en fonction des comptes normalisés ainsi que de la dispersion pour chaque gène [27].

1.4 Objectifs du mémoire

Galdieria sulphuraria 074W est une microalgue poly-extrémophile pouvant croître en hétérotrophie sur de nombreuses sources de carbone organique différentes. Ses grandes adaptabilité et flexibilité métaboliques lui confèrent un statut d'organisme idéal pour l'étude de la croissance hétérotrophe et du catabolisme des sucres. Lors d'une croissance hétérotrophe, une différence de pigmentation majeure peut être observée selon la source de carbone organique utilisée. En effet, alors qu'elle va conserver une couleur verte-bleue en hétérotrophie sur glycérol, la présence de glucose va lui conférer une couleur plus pâle voire jaune. Cette différence phénotypique ayant été observée au laboratoire, une analyse RNA-Seq a été designée en vue de comparer les variations dans le transcriptome des cellules suivant l'ajout de l'une ou l'autre source de carbone organique.

Le but de ce mémoire est donc de suivre un pipeline d'analyse de données RNA-Seq sur des données provenant d'un organisme haploïde cultivé, en hétérotrophie, dans 6 conditions différentes. Ce pipeline ira de l'analyse des séquences brutes provenant du séquençage jusqu'à celle des gènes différentiellement exprimés (DEG) par l'utilisation d'un package R.

2 Matériels

2.1 Conditions de culture et expérience détaillée

En Juin 2021, des cultures de *Galdieria sulphuraria* 074W ont été réalisées en hétérotrophie pure sur deux sources de carbone organiques: le glucose et le glycérol. Les différentes cultures ont été ensemencées à une densité optique de 0.2 à 800 nm dans un milieu Allen modifié dont la composition est la suivante : $(NH_4)_2SO_4$ 11.35 mM, $MgSO_4.7H_2O$ 1.22 mM, KH_2PO_4 2.20 mM, $CaCl_2.2H_2O$ 0.14 mM, NaCl 0.34 mM, Fe-Na-EDTA 36 µM, H_3BO_3 92 µM, $MnCl_2.4H_2O$ 18 µM, $ZnSO_4$ 15 µM, $(NH_4)_6Mo_7O_24.4H_2O$ 1.7 µM, $CuSO_4.5H_2O$ 0.64 µM, $NaVO_3.4H_2O$ 0.41 µM, $CoCl_2.6H_2O$ 0.34 µM. De plus, 12.5 mM de Glucose ou 25 mM de Glycérol y ont été ajoutés (ce qui assure un même nombre d'atomes de carbone disponibles dans les deux conditions). Elles ont été placées dans un incubateur à 42°C sous une agitation orbitale constante. Après 67 heures de croissance, des échantillons ont été récoltés et ceux-ci constituent les Condition 1 et Condition 2 (Figure 2.1). 50 mL de culture ont été récoltés, centrifugés 3 min à 3000 g et les culots conservés à -80°C. 1 heure après la première récolte, une deuxième source de carbone a été ajoutée aux cultures (identique (Condition 3 et 6) ou différente (Condition 4 et 5) à la première) (Figure 2.1). Après 6 heures supplémentaires de culture (74h au total), les échantillons des 4 conditions supplémentaires ont ainsi été récoltés comme mentionné précédemment. L'expérience a été réalisée en triplicata.

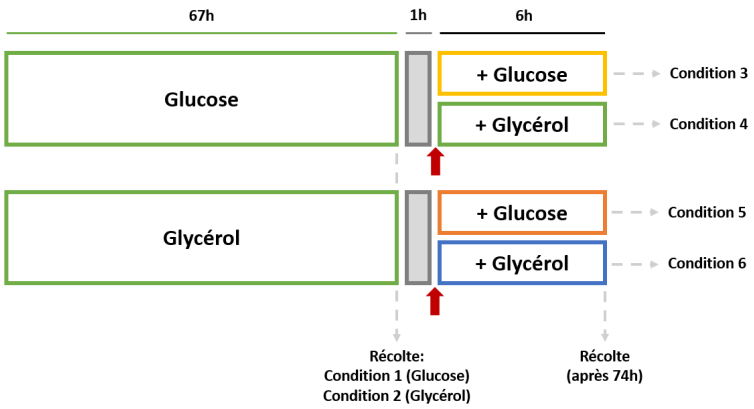

Figure 2.1: Schéma explicatif de l'expérience réalisée. Les flèches rouges représentent le moment d'ajout de la 2ème source de carbone.Les flèches grises pointillées représentent les moments de récolte.

Ensuite, les ARN totaux des échantillons des 6 conditions ont été extraits suivant un protocole adapté de Newman 1990 [31] et envoyés à la société Novogene Co., LTD (Beijing, China) qui en a réalisé le séquençage. La compagnie a tout d'abord construit la librairie ARN grâce au kit "NEB Next® Ultra™ RNA Library Prep Kit" et ensuite le séquençage via la méthode Illumina. Ce dernier a produit en moyenne 11663173 séquences pairées d'une longueur de 150 bp pour tous les échantillons.

Les échantillons seront notés comme suit : GSRNA_XY_* où X est la condition utilisée (Figure 2.1), Y est le réplicat biologique et peut prendre les valeurs 1, 2 ou 3 et * est le sens de lecture (1 ou 2).

2.2 Génome de référence et fichier d'annotations

Le 17 novembre 2021, le génome de référence de *Galdieria sulphuraria* 074W (N° NCBI: *GCA_000341285.1, ASM34128v1*), le transcriptome au format ".fna" ainsi que le fichier d'annotations au format GFF, ont été téléchargés sur le site NCBI à l'url suivante : https://www.ncbi.nlm.nih.gov/genome/405. Ce génome fait une longueur totale de 13,712 Mbp, contient 7174 protéines et possède un contenu en GC de 37,9 % . Le niveau d'assemblage le plus élevé sont les scaffolds dont le nombre est de 433.

2.3 *FastQC*

FastQC est un outil visant à fournir un moyen simple de réaliser des contrôles de la qualité des séquences brutes et/ou filtrées issues du séquençage à haut débit[1]. Il permet de déterminer rapidement si les séquences ont des anomalies et ainsi de visualiser la présence de problèmes ou de biais provenant du séquenceur ou de la librairie de départ, pouvant affecter la suite des analyses.

Après analyse des séquences, ce programme fournit des graphiques résumant les résultats ainsi qu'un rapport au format HTML comprenant des informations telles que la longueur des séquences, leur taux de duplication, leur contenu en GC, le contenu en base de séquences ou encore la qualité du séquençage pour chaque base. FastQC peut être utilisé dans deux modes différents[2]:

[1]https://rtsf.natsci.msu.edu/genomics/tech-notes/fastqc-tutorial-and-faq/
[2]https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/

- De façon autonome et interactive pour les analyses immédiates d'un petit nombre de fichiers FASTQ, ou,
- De façon non-interactive où il doit être intégré dans un pipeline d'analyses plus large pour le traitement systématique d'un grand nombre de fichiers.

Ce logiciel est téléchargeable à l'url suivante : https://www.bioinformatics.babraham.ac.uk/projects/download.html#fastqc.

2.4 *Trimmomatic*

Trimmomatic est un outil rapide servant, non seulement, à filtrer les séquences brutes provenant du séquençage Illumina mais aussi à en ôter les adaptateurs [4]. Ce programme est utilisé en ligne de commande et successivement à *FastQC* de sorte à minimiser les problèmes et biais détectés par ce dernier. Il est capable de traiter des séquences pairées et non-pairées sur des fichiers FASTQ. Lors de l'utilisation avec des séquences pairées, comme ce sera le cas dans ce manuscrit, la correspondance entre les paires de séquences est maintenue grâce au rejet des paires incomplètes post-filtration et coupure des séquences.

Ce logiciel est téléchargeable à l'url suivante : http://www.usadellab.org/cms/?page=trimmomatic.

2.5 *HiSat2*

HiSat2 est un programme rapide et sensible d'alignement utilisé pour le mapping de reads issus d'un séquençage de nouvelle génération contre un génome de référence. Sur base d'une extension de BWT pour les graphes, les développeurs ont implémenté un graphe d'index FM. Celui-ci est ainsi généré par le programme *hisat2-build*[3] et a été développé sur base d'implémentations de *HiSat2* et *Bowtie2*. Il permet d'obtenir, *in fine*, des alignements au format *SAM* peuvant être utilisés ensuite par d'autres programmes.

Ce logiciel est téléchargeable à l'url suivante: http://daehwankimlab.github.io/hisat2/download/.

[3]https://daehwankimlab.github.io/hisat2/manual/

2.6 *Samtools*

Samtools est une suite de programmes pour le traitement de données de séquençage à haut débit. Il s'agit d'une suite d'outils permettant de lire, écrire, éditer, indexer et regarder des fichiers aux formats *SAM/BAM/CRAM* [24]. Les commandes *samtools view* et *samtools sort* seront utilisées dans la suite du manuscrit[4]. La première est un outil permettant de convertir les formats entre eux (de *SAM* vers *BAM* et inversement) et la deuxième de trier les alignements au sein des fichiers *SAM/BAM*.

Ce logiciel est téléchargeable à l'url suivante: http://www.htslib.org/download/.

2.7 *Htseq-count*

Htseq-count est un programme permettant le comptage du nombre de séquences alignées sur chaque motif[5]. Dans le cas des analyses RNA-Seq, les motifs sont souvent des gènes, ou chaque gène est considéré comme l'union de tous ses exons.

Ce logiciel est téléchargeable à l'url suivante: https://pypi.org/project/HTSeq/.

2.8 *DESeq2*

DESeq2 est un package R utilisé pour la détection de gènes différentiellement exprimés (DEG) au sein d'une matrice de comptage [27, 25].

Ce package s'installe sur R ou Rstudio grâce aux commandes:

```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")

BiocManager::install("DESeq2")
```

[4]http://www.htslib.org/doc/samtools.html
[5]https://htseq.readthedocs.io/en/release_0.11.1/count.html

3 Méthodes

3.1 Analyse de la qualité et filtration des séquences brutes (*FastQC* et *Trimmomatic*)

Préalablement à toute analyse de qualité des reads, le nombre de séquences par fichier est déterminé grâce à la commande *grep* (Annexe A2). La qualité des séquences brutes est ensuite évaluée grâce au programme *FastQC*. Ce programme a été utilisé sur l'ensemble des données de séquençage. Pour ce faire, un fichier contenant les noms d'échantillons a été créé (Annexe A3.1) et un job array a ensuite été compilé (Annexe A3.2).

Une fois les résultats de qualité de *FastQC* analysés, les séquences sont filtrées et coupées grâce au programme *Trimmomatic*. La société Novogene avait réalisé un pré-traitement des données brutes qui impliquait notamment la suppression des reads contenant: les adaptateurs du séquençage Illumina, un pourcentage de base N supérieur à 10% et une qualité de base, calculée sur 50% du nombre total de bases, inférieure à 5 (Q≤5). Dans ce contexte, les valeurs suivantes ont été utilisées pour la filtration avec *Trimmomatic*:

- leading et trailing : 26
- slidingwindow : 5:26
- crop : 144
- minlen : 100

Pour réaliser la filtration des séquences, un fichier comprenant les noms d'échantillons et les valeurs des paramètres est créé (Annexe A4.1). Un job array faisant appel à *Trimmomatic* est ensuite créé et compilé (Annexe A4.2).

Une fois cette étape de filtration et traitement des séquences réalisée, le programme *FastQC* est à nouveau utilisé pour analyser ces séquences et en déterminer la qualité. L'analyse *FastQC* est donc réalisée en utilisant un job array, comme précédemment (Annexe A3.2). Un autre paramètre important calculé après une filtration des séquences par *Trimmomatic* est le calcul du pourcentage de perte de séquences. Celui-ci est réalisé sur base des séquences comptées avant et après filtration.

3.2 Alignement des séquences via *HiSat2*

La première étape pour réaliser le mapping a été de convertir le fichier d'annotation du génome du format GFF au format GTF2, indispensable pour le bon fonctionnement des programmes suivants (Annexe A5.1). De ce fichier GTF2 obtenu sont extraits les listes d'exons et de sites d'épissage présents dans le génome, informations qui seront utiles à l'alignement des séquences (Annexe A5.2).

Ensuite, le génome de référence de *Galdieria sulphuraria* a été indexé grâce à *hisat2-build* (Annexe A5.3) et l'indexation a été vérifiée grâce à *hisat2-inspect* (Annexe A5.4). Ce dernier programme va créer un fichier FASTA contenant les séquences de référence initiales sur base de l'index créé précédemment ainsi que les séquences de référence utilisées pour le construire. Les séquences brutes vont enfin pouvoir être alignées sur le génome de référence (Annexe A5.5). Le résumé des résultats d'alignement sont enregistrés dans les rapports d'activité de *HiStat2* (.e*) et sont analysés.

3.3 Préparation des fichiers pour *Htseq-count* avec *SAMtools*

Une fois l'alignement des séquences réalisé par *HiSat2*, l'outil *SAMtools* a été utilisé pour transformer les fichiers obtenus au bon format pour la suite des analyses. Ainsi, un nouvel index a été créé à partir du génome de référence (Annexe A6.1). Celui-ci a ensuite été utilisé par *SAMtools* afin, d'une part, de transformer les fichiers *SAM*, résultant de *HiSat2*, en fichiers *BAM* et, d'autre part, de les trier afin de faciliter l'utilisation de *Htseq-count* (Annexe A6.2).

3.4 Compte des séquences avec *Htseq-count*

Les fichiers étant prêts, les alignements de séquences ont été comptés grâce à *htseq-count*. Ce comptage a été réalisé via l'utilisation par défaut qui ne tient pas compte des séquences qui s'alignent plusieurs fois sur le génome (Annexe A7.1). Les informations sur les alignements de séquences, (*i.e.* sans motif, ambigu, non aligné et alignement non-unique) se trouvent dans les rapports d'activité d'*htseq-count* (.o*), sont étudiées

attentivement et répertoriées dans un fichier tabulaire. Dans ces rapports d'activités se trouvent également les résultats du comptage. Ceux-ci sont également récupérés dans une matrice de comptage (Annexe A7.2)

3.5 *Analyses statistiques des comptes de séquences*

Les analyses statistiques des comptes de séquences sont effectuées via le package R *DESeq2* (Annexe A8). Ces analyses ont été réalisées en deux temps : premièrement, sur le dataset complet (*i.e.* les 3 réplicats biologiques par condition), et deuxièmement, sur un dataset tronqué ne contenant que 2 réplicats biologiques (Annexes A8.1 et A8.3). Un modèle a donc été réalisé sur l'entièreté du dataset puis une analyse en composantes principales ainsi qu'une heatmap ont été construites pour rendre compte de la variabilité intrinsèque aux réplicats biologiques (Annexes A8.1 et A8.3). Une liste de gènes différentiellement exprimés a été extraite à titre d'exemple en faisant le contraste entre les conditions 1 et 2 (Annexe A8.2).

4 Résultats

4.1 Analyse de qualité et filtration des séquences

La qualité des séquences issues du séquençage Illumina a été analysée avant de commencer leur alignement sur le génome de référence. Cette analyse a pour but, non seulement, de minimiser les biais mais également d'augmenter la qualité des séquences en vue des prochaines analyses. Elle a permis, entre autres, de déduire les valeurs des paramètres à utiliser pour filtrer les séquences via le programme *Trimmomatic*.

Le séquençage Illumina a généré 36 fichiers de séquences, contenant une moyenne de 11663173 séquences pairées, correspondant aux 18 échantillons de *Galdieria sulphuraria* (chacun ayant été séquencé en sens et anti-sens). Ces 18 échantillons correspondent aux triplicata des 6 conditions énoncées dans la partie Matériel (*cf.* 2.1 et Figure 2.1). Les séquences brutes analysées dans ce mémoire sont des séquences qui ont été obtenues grâce à un pré-traitement des données réalisé l'entreprise Novogene. Ce pré-traitement impliquait la suppression des reads contenant : les adaptateurs du séquençage Illumina, un pourcentage de base N supérieur à 10% et une qualité de base, calculée sur 50% du nombre total de bases, inférieure à 5 ($Q \leq 5$).

Le nombre de séquences (pré-traitées) par fichier a donc été calculé (Table 4.1) et la qualité de ces séquences évaluée grâce au programme *FastQC*. Ce programme fourni un rapport au format HTML par fichier de séquences (*cf. 2.3*), duquel les premières informations peuvent être extraites. Ainsi, la totalité des séquences (sens et anti-sens) ont une taille de 150 bp et un %GC moyen de 40.5%. De façon globale, il est clair que le pré-traitement réalisé par Novogene nous a permis d'obtenir des séquences de bonne qualité et un nombre

Table 4.1: Compte des séquences avant et après filtration avec *Trimmomatic*. La 3ème colonne représente le pourcentage de séquences perdues au cours de la filtration. Les données présentées sont les trois réplicats de la condition 1.

	Nombre de séquences pré-filtration	**Nombre de séquences post-filtration**	**Pourcentage de séquences perdues**
GSRNA_11_sens	9857250	8517589	13.6%
GSRNA_12_sens	14726243	12691217	13.8%
GSRNA_13_sens	11552309	9847067	14.8%

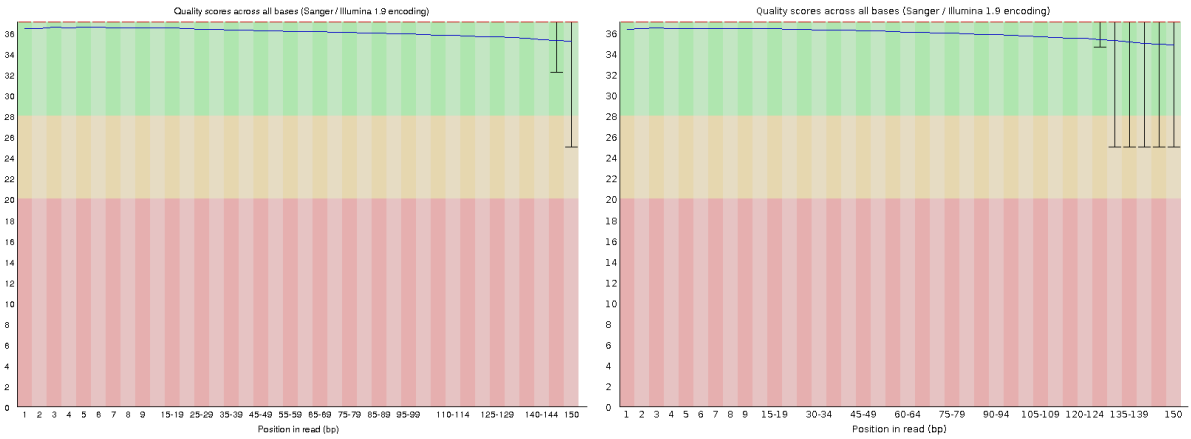

Figure 4.1: Qualité par base avant filtration. Gauche: Séquence sens de l'échantillon GSRNA_31. Droite: Séquence anti-sens de l'échantillon GSRNA_53.

de séquences par fichier, dont un exemple est présenté au Tableau 4.1, homogène entre réplicats biologiques et échantillons. Néanmoins, la totalité des séquences a quand même été analysée et refiltrée pour espérer gagner en qualité sans engendrer une perte de séquences trop importante.

Prenons 2 échantillons représentatifs correspondant à une séquence sens et une séquence anti-sens (Figure 4.1). Comme montré sur la Figure 4.1, le score de qualité par base Q est d'environ 36 tout au long des séquences, sens comme anti-sens, et diminue légèrement à leur extrémité. Cette qualité est raisonnable et ne demandera pas de réaliser une filtration trop stricte. Un deuxième point à analyser dans le rapport de *FastQC* est le contenu en base des séquences (Annexe A9.1 – Figure A9.1). Dans tous les fichiers, la quantité relative de chaque base (A/T et G/C) en début de séquence est déséquilibrée et se stabilise ensuite. Ce déséquilibre est dû à l'amorçage aléatoire des bases lors de la production de la librairie du séquençage Illumina et n'est pas problématique pour la suite des analyses. En

fin de séquence, par contre, un très léger déséquilibre est visible sur certains échantillons après la base 140 (Annexe A9.1 – Figure A9.1).

L'analyse des résultats de *FastQC* a ainsi permis de déterminer les valeurs des paramètres nécessaires à la filtration et la coupe des séquences via *Trimmomatic*. Dans ce programme, l'option *illuminaclip* permet de couper les adaptateurs et séquences spécifiques du séquençage Illumina. Bien que probablement non nécessaire dans notre cas, au vu des conditions de pré-traitement des séquences, cette option a néanmoins été maintenue pour s'assurer qu'aucun adaptateur et séquence spécifique ne viennent perturber les analyses futures. Les paramètres *leading*, *trailing* et *slidingwindow* sont fixés à un score Q de 26, score minimal permettant de conserver un nombre élevé de séquences ayant une bonne qualité. Le paramètre *leading* permet de couper, en début de séquence, les bases ayant un score en-dessous du score Q fixé (*i.e.* 26). *trailing*, quant à lui, permet de couper des bases similaires en fin de séquence. Le paramètre *slidingwindow* scanne la séquence et coupe la fenêtre de lecture si sa qualité moyenne est inférieure au score Q fixé. L'option *crop* a, quant à elle, été définie à une valeur de 144 et correspond à la taille à laquelle les séquences vont être coupées. Enfin, l'option *minlen* a été fixée à 100 et permet de conserver uniquement les séquences d'une taille plus grande ou égale à 100 bp, *i.e.* d'éliminer les séquences devenues trop courtes suite à la filtration. Ainsi, après l'utilisation de *Trimmomatic*, toutes les séquences auront une taille entre 100 et 144 bp ainsi qu'un score Q d'au moins 26.

Une fois les séquences filtrées et coupées par *Trimmomatic*, leur nombre est déterminé (Tableau 4.1) et leur qualité est de nouveau analysée grâce au programme *FastQC*. Comme attendu, les rapports HTML post-filtration indiquent une taille de séquence entre 100 et 144 bp. Le score de qualité est, quant à lui, bien supérieur à 26 et plus homogène tout au long des séquences qu'avant filtration (Annexe A9.2 – Figure A9.2). Le contenu en base par séquence est, quant à lui, toujours en déséquilibre en début de séquences, *i.e.* entre la première et la dixième base, mais la quantité relative de chaque base est plus stable, même en fin de séquence (Annexe A9.2 – Figure A9.3). Enfin, comme le montre le Tableau 4.1, le pourcentage de perte de séquences causé par l'utilisation de *Trimmomatic* a été calculé sur base des nombres de séquences avant et après filtration. Les paramètres de filtration utilisés résultent en une perte de séquences d'environ 15%

pour tous les échantillons. Cette perte n'étant pas excessive et la qualité des séquences étant globalement plus élevée, les paramètres utilisés (score de qualité de 26, longueur de séquence entre 100 et 144 bp) ont été jugés adéquats pour permettre la suite des analyses.

4.2 Alignement des séquences sur un génome de référence

La deuxième étape de l'analyse des séquences brutes consiste en leur alignement sur un génome de référence. Cette étape permet ensuite de pouvoir évaluer l'expression des gènes dans les différents échantillons. Afin de réaliser cette étape, le génome de référence de *G. sulphuraria* 074W (Numéro d'accession: GCA_000341285.1, ASM34128v1) et le programme *HiSat2* ont été utilisés (Annexe A5).

Le programme *HiSat2* nécessitant, en entrée, un fichier d'indexation du génome de référence, celui-ci a été généré par le programme *hisat2-build* (Annexe A5.3). Ce dernier utilise le génome de référence ainsi que deux fichiers contenant chacun les sites d'épissages et les exons, informations extraites du fichier GTF2 (Annexe A5.2). Ce fichier GTF2 a été obtenu par conversion du fichier GFF téléchargé sur le site du NCBI, en utilisant le programme *gffread* (Annexe A5.1). L'option "-T" a dû être utilisée de sorte à obtenir le fichier au format GTF2, utilisable par la suite des programmes (une image de l'organisation de ce fichier GTF2 est présenté en Annexe A10 - Figure A10.1).

Lors de l'alignement avec *HiSat2* (Annexe A5.5), le paramètre *–threads* permet d'effectuer des alignements en parallèle et de synchroniser les résultats. Les séquences étant pairées, il prend une valeur de 2 ce qui permet d'accélérer de façon presque linéaire l'alignement. Le paramètre *–secondary* permet d'obtenir les alignements multiples dans le rapport d'alignement créé grâce à l'utilisation de l'option *–summary-file*.

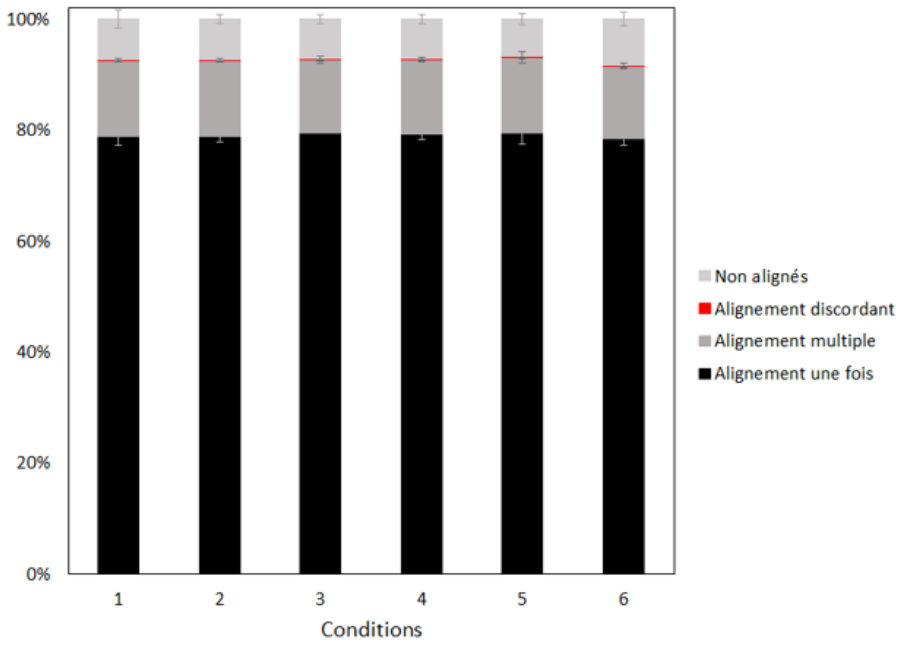

Figure 4.2: Résultats d'alignement pour les différentes conditions. Les résultats sont présentés en pourcentage de séquences et moyennés sur les réplicats biologiques. Conditions : 1 – Glucose, 2 – Glycérol, 3 – Glucose + Glucose, 4 – Glucose + Glycérol, 5 – Glycérol + Glucose, 6 – Glycérol + Glycérol.

Enfin, l'option *-q* indique à *HiSat2* que les fichiers de séquences apportés sont au format FASTQ.

Le rapport d'alignement créé a ensuite été utilisé afin d'évaluer la qualité des alignements ainsi que la fréquence d'alignement (alignement unique, multiple, de façon discordante ou non aligné, Figure 4.2). Pour des séquences pairées, des séquences s'alignant de façon discordante représentent un alignement pour lequel l'orientation d'une des deux séquences n'est pas celle espérée. Ainsi, la Figure 4.2 représente les résultats des différentes catégories d'alignement sous forme de pourcentage moyen pour les réplicats biologiques de chaque condition. Pour tous les échantillons, sans exception, la totalité des reads pairés s'alignent sur le génome de référence et l'alignement total des séquences varie entre 91 et 95%. Il y a donc, en moyenne, 7.5% de séquences non alignées. Ensuite, une moyenne de 13.5% des séquences sont alignées plus d'une fois sur le génome de référence et de 79% sont alignées strictement une fois. Enfin, moins d'un pourcent de séquences s'alignent de façon discordante. Malgré un peu de variabilité entre les réplicats biologiques (Figure 4.2), les résultats d'alignement semblent homogènes entre les différents échantillons. De plus, les pourcentages de séquences non alignées et alignées de façon discordante sont faibles ce

qui tant à montrer que les options du programme *HiSat2* ont été bien sélectionnées.

4.3 Compte des séquences

Une fois l'alignement sur le génome de référence effectué, les séquences alignées sont comptées grâce au programme *Htseq-count*. Cette étape est indispensable pour déterminer la quantité de séquences alignées par gène, quantité qui sera ensuite utilisée pour réaliser les analyses de gènes différentiellement exprimés (DEG).

Dans le cas des données pairées, le programme *Htseq-count* nécessitant des fichiers d'alignements triés suivant les noms ou les positions dans l'alignement, les fichiers *SAM* obtenus lors de l'alignement avec *HiSat2* ont été convertis et triés pour remplir ce critère. Cette étape de préparation des fichiers a été réalisée avec *Samtools* (Annexe A6). Un fichier d'indexation a d'abord été créé au format adéquat pour être utilisé par *Samtools* (Annexe A6.1). Ensuite *samtools view* et *samtools sort* ont été utilisés pour transformer les fichiers *SAM* en *BAM* et pour trier les fichiers *BAM* obtenus. L'option *-n* de *samtools sort* permet ainsi de trier les alignements par nom (Annexe A6.2).

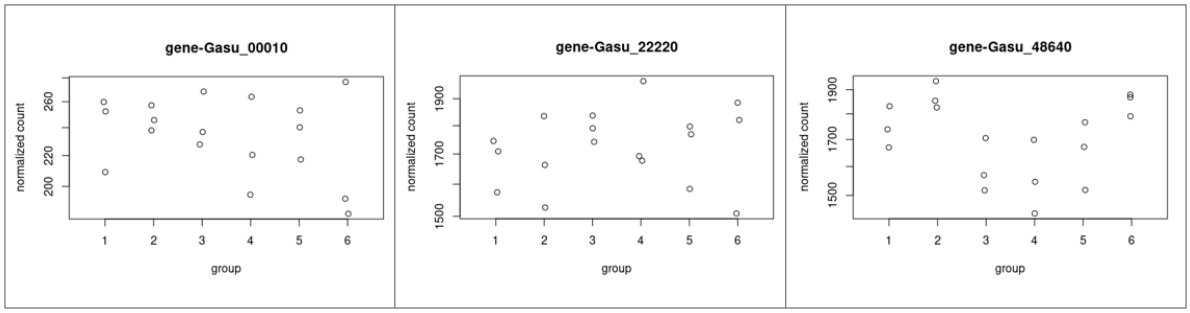

Figure 4.3: Compte des séquences pour 3 gènes aléatoires dans les différentes conditions. Group = Conditions : 1 – Glucose, 2 – Glycérol, 3 – Glucose + Glucose, 4 – Glucose + Glycérol, 5 – Glycérol + Glucose, 6 – Glycérol + Glycérol

Le programme *Htseq-count* a ensuite été utilisé avec les options *-f*, permettant de spécifier le format du fichier d'entrée (*i.e.* bam), et *-r* qui indique la base du tri (*i.e.* par nom). Concernant, l'option *-s* (ou *–stranded*), elle a été définie sur "no" puisque le kit utilisé pour créer la librairie d'ADN complémentaire utilise un protocole qui est "non-stranded specific". Les résultats de *Htseq-count* sont récupérés dans une seule et même matrice et sont analysés. Cette matrice, dont un extrait est présent en Annexe (Annexe 11 - Figure A11.1), est importée dans R pour les analyses suivantes.

4.4 Analyses statistiques des comptes de séquences

Les gènes différentiellement exprimés sont obtenus en utilisant le package R *DESeq2* et la matrice de comptage obtenue grâce à *Htseq-count*. Ainsi, une fois le modèle créé grâce à la fonction DESeq, les résultats obtenus ont été observés de différentes manières (Annexe A8.1). Premièrement, la fonction *PlotCounts* a permis de mettre en graphique le compte de séquences pour un gène particulier en fonction des différentes conditions (Figure 4.3). Ce graphique, réalisé pour 3 gènes sélectionnés complètement au hasard, montre une assez grande disparité dans les comptes de gènes entre réplicats biologiques.

Deuxièmement, les résultats sont utilisés pour créer un graphique PCA ("Principal Component Analysis") (Figure 4.4). Ce graphique, regroupant les réplicats biologiques sous la même couleur, permet d'analyser la variance du niveau d'expression des gènes au travers des conditions, et donc de mettre en évidence les clusters de génotypes similaires. La PCA essaie donc de décrire comment la variance se réparti entre échantillons et identifie des composantes qui expliquent une majeure partie de cette variance. Ainsi, les coordonnées PC1 et PC2 correspondent aux pourcentages de variation des données les plus importantes. Comme le montre la Figure 4.4, les données analysées démontrent une grande hétérogénéité au sein des réplicats biologiques. En effet, les "clusters" obtenus sont

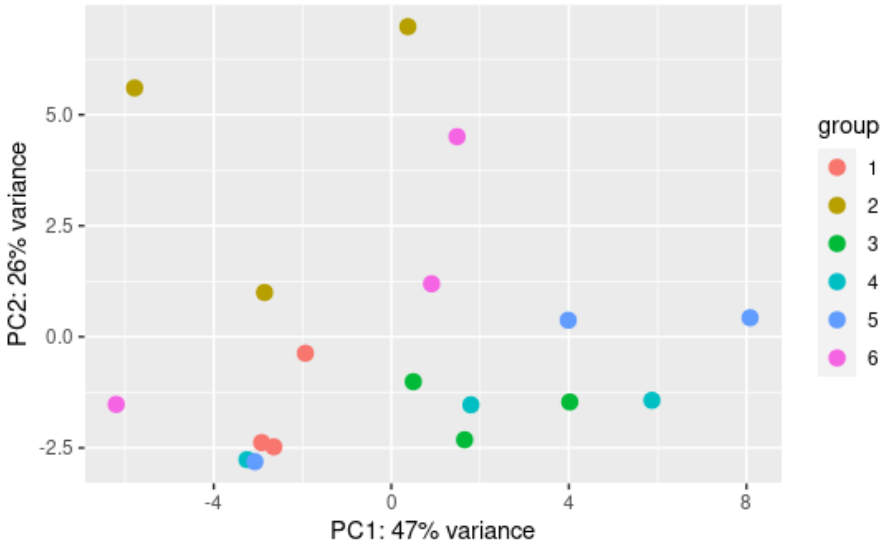

Figure 4.4: Graphique PCA (Analyse des composantes principales) des différentes conditions. Les réplicats biologiques au sein d'une condition sont représentés par la même couleur. Conditions : 1 – Glucose, 2 – Glycérol, 3 – Glucose + Glucose, 4 – Glucose + Glycérol, 5 – Glycérol + Glucose, 6 – Glycérol + Glycérol.

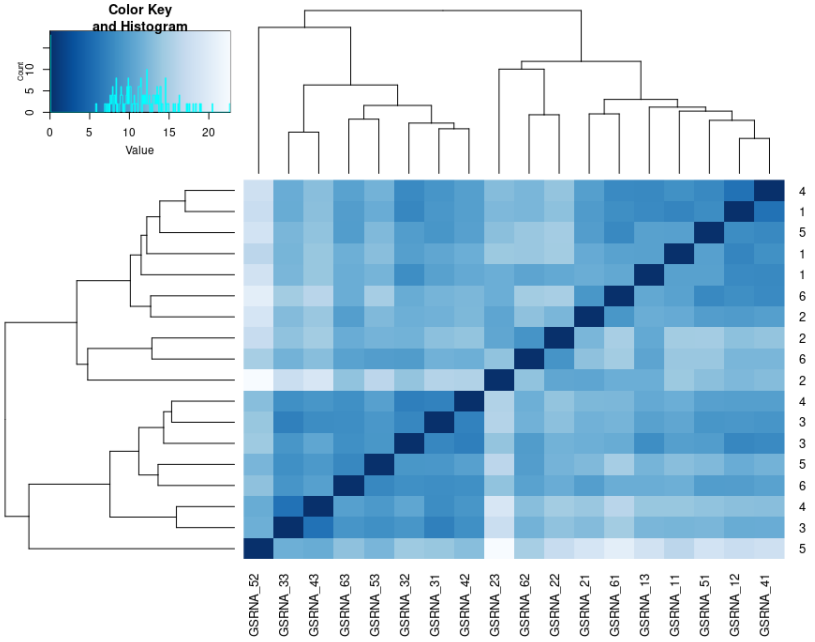

Figure 4.5: Heatmap des différents échantillons.

très disparates et ne permettent pas de mettre en évidence une séparation sur base des conditions expérimentales sur l'une ou l'autre des composantes principales (PC).

Troisièmement, une heatmap peut être réalisée afin d'exprimer différemment cette même constatation (Figure 4.5). Celle-ci permet de visualiser les distances entre échantillons

et les échantillons similaires doivent donc se regrouper en cluster, ce qui n'est pas le cas pour la heatmap présentée à la Figure 4.5. En effet, sur celle-ci, les échantillons sont relativement disparates, alors que les réplicats biologiques devraient se trouver groupés 3 par 3. Ils se trouvent même éparpillés dans chacun des deux clusters, à l'exception des échantillons des conditions 1, 2 et 3 qui eux, sont, éparpillés au sein du même cluster.

Outre ces informations graphiques, il est tout de même possible de récupérer une liste de gènes différentiellement exprimés. Cette liste pourrait être utilisée par la suite, par exemple, pour la création d'un diagramme de Venn afin de comparer les gènes sur- et sous-exprimés entre différentes conditions. Pour ce faire, seuls les gènes ayant un facteur de changement (log2 fold-change) supérieur à 1 ou inférieur à -1 ainsi qu'une p-valeur inférieure à 0.05 sont considérés comme différentiellement exprimés. Dans le cas des analyses de ce mémoire, il est possible de récupérer une liste de gènes pour certaines conditions uniquement, les autres comparaisons menant à des listes vides (Annexe A8.2). Un exemple de liste de gènes sous-régulés[6] dans le cas de la comparaison entre les conditions 1 et 2 (Glucose vs Glycérol) est présentée en Annexe A12 - Figure A12.1. Bien que cette liste comporte quelques gènes remplissant les conditions, elle n'a pas été exploitée au vu de la variabilité générale des échantillons.

Enfin, ces résultats montrent tous un problème d'hétérogénéité au sein des échantillons. En effet, la variance au sein des réplicats biologiques est tellement grande qu'il semble compliqué de tirer des conclusions quant à l'expression différentielle des gènes. Bien qu'il ne s'agisse pas d'une solution statistiquement très appropriée, nous avons néanmoins décidé d'ôter un réplicat biologique pour espérer retrouver des clusters et tirer quelques conclusions de l'expérience (Annexe A8.3). Il est néanmoins important d'insister sur le fait que la perte de pouvoir statistique engendrée par cette manipulation des données est grande et risque de rendre les résultats obtenus inexploitables en l'état.

Ainsi, de façon arbitraire et en se basant sur le graphique PCA (Figure 4.2), le réplicat biologique le plus éloigné des deux autres au sein d'une condition a été ôté du dataset. Et des analyses similaires à celles du dataset complet ont été réalisées (Annexe A8.3).

Les *PlotCounts* des mêmes gènes choisis aléatoirement ont été réalisés et sont présentés en Annexe (Annexe A13 - Figure A13.1). Sur ceux-ci, on remarque que la variabilité

[6]La liste des gènes sur-régulés étant vide, elle n'a pas été incluse dans ce manuscrit.

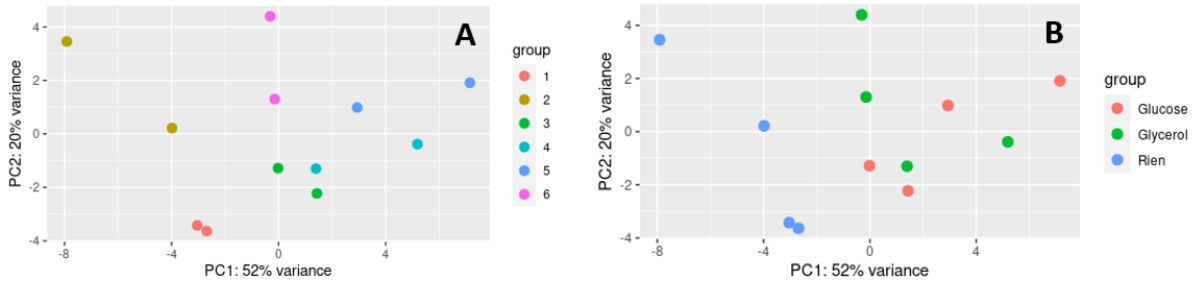

Figure 4.6: Graphique PCA (Analyse des composantes principales) des différentes conditions. A : Sur base des conditions. Les réplicats biologiques au sein d'une condition sont représentés par la même couleur. B : Sur base de la 2eme source de carbone organique ajoutée. Conditions : 1 – Glucose, 2 – Glycérol, 3 – Glucose + Glucose, 4 – Glucose + Glycérol, 5 – Glycérol + Glucose, 6 – Glycérol + Glycérol.

est encore bien présente au sein des conditions. Ensuite le graphique PCA exprimé par rapport aux différentes conditions a pu être réalisé et est présenté à la Figure 4.6A. Sur cette figure, on remarque également la variabilité importante mais que suivant la PC1, deux conditions se détachent des autres (*i.e.* les conditions 1 et 2 d'une part, et les 4

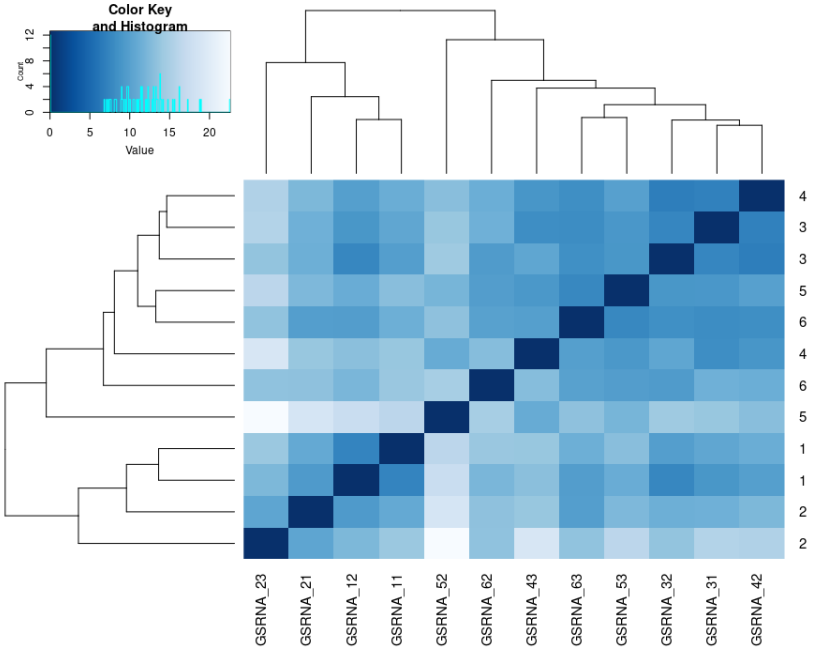

Figure 4.7: Heatmap des différents échantillons.

autres conditions d'autre part). Cette distinction est liée aux conditions d'ajout de la deuxième source de carbone organique. En effet, si l'on affiche le graphique PCA non pas en fonction des 6 conditions mais en fonction des conditions de la 2ème source de carbone ajoutée (Aucune (Rien), Glucose ou Glycérol), on remarque effectivement un

"cluster" comprenant les conditions 1 et 2 auxquelles aucune deuxième source n'a été ajoutée (Figure 4.6B).

Enfin, la heatmap a elle aussi été implémentée et est présentée à la Figure 4.7. Celle-ci permet, comme attendu, de montrer les mêmes résultats que le graphique PCA. En effet, un groupe, contenant les conditions 1 et 2 semble se dégager des quatre autres conditions, qui, elles, montrent encore une grande variabilité.

5 Discussion et perspectives

Un des buts majeurs du projet DARKMET a été l'étude plus approfondie du métabolisme hétérotrophe de microalgues sélectionnées sur base de leur bonne croissance dans ces conditions. Une des souches de microalgues étudiée dans ce projet est *Galdieria sulphuraria*, une microalgue rouge poly-extrémophile. Comme d'autres Cyanidiales, *G. sulphuraria* peut croître dans des habitats très acides (à pH inférieur à 3) comme très chauds (jusqu'à 50°C), et possède la particularité de se développer en mixotrophie et en hétérotrophie sur une grande diversité de sources de carbone organique. Cette particularité lui confère une grande flexibilité métabolique, ce qui en fait un organisme idéal pour l'étude du catabolisme des sucres et du métabolisme hétérotrophe.

La souche de *Galdieria sulphuraria* étudiée dans le cadre de ce projet est la souche 074W. Lorsqu'elle est cultivée en hétérotrophie, cette dernière possède une différence de pigmentation selon la source de carbone organique utilisée. En effet, elle conserve une couleur vert-bleu lorsqu'elle est en présence de glycérol et prend une couleur plus pâle-jaune en présence de glucose. De plus, une fois la source de carbone (*i.e.* le glucose) consommée, elle reprend une couleur verte proche de celle de la condition sous glycérol. Cette différence phénotypique, observée au laboratoire, nous a poussé à considérer une analyse RNA-seq afin de mettre en évidence d'éventuelles variations d'expression génique entre différentes conditions en présence de glucose et/ou de glycérol (6 conditions au total). Le séquençage des transcrits a ainsi été réalisé par séquençage Illumina et les séquences obtenues ont été alignées sur le génome de référence de *G. sulphuraria* 074W, assemblé au niveau des scaffolds. De plus, ces alignements ont été assignés à des gènes et ensuite comptés afin de pouvoir analyser statistiquement ces comptes de séquences.

Ainsi, après une analyse de la qualité des séquences brutes par *FastQC*, elles ont été filtrées afin d'augmenter leur qualité et réduire les biais éventuels. Ces séquences filtrées ont ensuite été alignées sur le génome de référence grâce au programme *HiSat2*. L'alignement montre ainsi un pourcentage de séquences alignées une fois (80%) et alignées de multiples fois (13.5%) cohérent avec d'autres données de la littérature pour la microalgue verte de référence *Chlamydomonas reinhardtii* [7].

Ensuite, les séquences alignées ont été comptées avec le logiciel *Htseq-count* afin d'utiliser

les comptes pour l'analyse des gènes différentiellement exprimés. Au cours de l'analyse des comptes de gènes, il s'est avéré clair que les données ont un problème d'hétérogénéité au sein des triplicata biologiques. En effet, les représentations graphiques (PCA et Heatmap) montrent qu'il est impossible de construire des clusters reprenant tous les réplicats biologiques du même échantillon indépendamment des autres échantillons. La suppression d'un réplicat biologique est une solution imparfaite, soumise à des décisions arbitraires sur le choix du réplicat à éliminer, qui n'a pas réussi à améliorer les problèmes de variance. De plus, le pouvoir statistique perdu au cours de cette opération est trop important pour tirer la moindre conclusion sur les données.

Deux hypothèses sont alors possibles pour expliquer cette variabilité. Premièrement, il pourrait y avoir trop peu de variabilité entre les échantillons et, dans ce cas, nous essayons de comparer des points dont la distribution est aléatoire sur des données variant très faiblement. Cette première hypothèse semble néanmoins peu probable étant donné la grande différence de phénotype observée (ne serait-ce qu'entre les conditions glucose (1) et glycérol (2)). Deuxièmement, la variance au sein des échantillons est effectivement très grande et pourrait être due à un problème d'échantillonnage. Quoi qu'il en soit, cette constatation rend impossible la suite des analyses et, ce faisant, la détermination de gènes différentiellement exprimés entre différentes conditions d'une façon statistiquement valable.

Plusieurs pistes d'explications peuvent être données afin d'améliorer un futur échantillonnage. Les croissances ayant été réalisées un réplicat biologique à la fois à une semaine d'intervalle, il est possible que l'état physiologique de la pré-culture hétérotrophe n'ait pas été identique entre les expériences. Si c'est le cas, le point de départ de tous les réplicats biologiques est différent et cela amène de la variabilité entre eux. Il est également possible que les délais au cours de l'échantillonnage même aient été différents, *i.e.* que certaines cultures aient été exposées plus longtemps à la lumière ou soient tombées en anoxie pendant la récolte. Enfin, le design de l'expérience et le choix des timings de récolte ont été élaborés sur base d'un article sur *Chromochloris zofingiensis* [39]. Il est possible que laisser *G. sulphuraria* durant 6 heures en présence de la deuxième source de carbone organique soit trop long ou trop court. En effet, le pic d'adaptation de la souche à la deuxième source de carbone est peut-être déjà passé ou pas encore en place, au niveau des

ARNs. *G. sulphuraria* ayant un métabolisme plus flexible que *C. zofingiensis* et encore assez méconnu, le design de l'expérience devrait peut-être être réadapté. Cette hypothèse va en faveur d'une trop petite variabilité entre les échantillons soumis aux deux sources de carbone organique successives.

Afin de pouvoir analyser plus en détail les gènes différentiellement exprimés chez *Galdieria sulphuraria* dans ces conditions, la suggestion évidente est donc de recommencer la partie de croissance et échantillonnage, au moins en duplicata biologique. Ce faisant, il serait alors possible d'obtenir des données avec moins de variabilité et de pouvoir les analyser complètement.

References

[1] E. Adams, K. Maeda, T. Kato, and C. Tokoro. Mechanism of gold and palladium adsorption on thermoacidophilic red alga *Galdieria sulphuraria*. *Algal Research*, 60: 102549, 2021. doi: 10.1016/j.algal.2021.102549.

[2] G. Barbier, C. Oesterhelt, M. D. Larson, R. G. Halgren, C. Wilkerson, R. M. Garavito, C. Benning, and A. P. M. Weber. Comparative genomics of two closely related unicellular thermo-acidophilic red algae, *Galdieria sulphuraria* and *Cyanidioschyzon merolae*, reveals the molecular basis of the metabolic flexibility of *Galdieria sulphuraria* and significant differences in carbohydrate metabolism of both algae. *Plant Physiology*, 137:460–474, 2 2005. doi: 10.1104/pp.104.051169.

[3] R. Barone, L. D. Napoli, L. Mayol, M. Paolucci, M. G. Volpe, L. D'Elia, and et al. Autotrophic and heterotrophic growth conditions modify biomolecule production in the microalga *Galdieria sulphuraria* (cyanidiophyceae, rhodophyta). *Marine drugs*, 18:169, 2020.

[4] A. M. Bolger, M. Lohse, and B. Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30:2114–2120, 8 2014. doi: 10.1093/bioinformatics/btu170.

[5] C. Bottone, R. Camerlingo, R. Miceli, G. Salbitani, G. Sessa, G. Pirozzi, and S. Carfagna. Antioxidant and anti-proliferative properties of extracts from heterotrophic cultures of *Galdieria sulphuraria*. *Natural Product Research*, 33:1659–1663, 6 2019. doi: 10.1080/14786419.2018.1425853.

[6] M. Carone, A. Corato, T. Dauvrin, T. L. Thanh, L. Durante, B. Joris, F. Franck, and C. Remacle. Heterotrophic growth of microalgae. *Grand Challenges in Algae Biotechnology*, pages 71–109, 2019. doi: 10.1007/978-3-030-25233-5_3.

[7] M. Castruita, D. Casero, S. J. Karpowicz, J. Kropat, A. Vieler, S. I. Hsieh, W. Yan, S. Cokus, J. A. Loo, C. Benning, M. Pellegrini, and S. S. Merchant. Systems biology approach in *Chlamydomonas* reveals connections between copper nutrition and multiple metabolic steps . *The Plant Cell*, 23:1273–1292, 4 2011. doi: 10.1105/tpc.111.084400.

[8] G. Chen, C. Wang, and T. Shi. Overview of available methods for diverse rna-seq data analyses. *Science China Life Sciences*, 54:1121–1128, 2011. doi: 10.1007/s11427-011-4255-x.

[9] C. H. Cho, S. Park, C. Ciniglia, E. Yang, L. Graf, D. Bhattacharya, and H. Yoon. Potential causes and consequences of rapid mitochondrial genome evolution in thermoacidophilic *Galdieria*(rhodophyta). *BMC Evolutionary Biology*, 20:112, 2020. doi: 10.1186/s12862-020-01677-6.

[10] M. Cloeter and G. Schonknecht. Temperature dependent changes in glycolysis in the thermoacidophilic red alga *Galdieria sulphuraria*. *Research Reports from Life Science Freshmen Research Scholars*, 2, 2016.

[11] A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, and et al. A survey of best practices for rna-seq data analysis. *Genome Biology*, 17: 13, 2016. doi: 10.1186/s13059-016-0881-8.

[12] J. Costa-Silva, D. Domingues, and F. Lopes. Rna-seq differential expression analysis: An extended review and a software tool. *PloS one*, 12:e0190152, 2017.

[13] S. J. Davis. Using minion nanopore sequencing to generate a de novo eukaryotic draft genome: preliminary physiological and genomic description of the extremophilic red alga *Galdieria sulphuraria* strain sag 107.79. 2016.

[14] B. C. B. de Freitas, E. H. Brächer, E. G. de Morais, D. I. P. Atala, M. G. de Morais, and J. A. V. Costa. Cultivation of different microalgae with pentose as carbon source and the effects on the carbohydrate content. *Environmental Technology*, 40: 1062–1070, 4 2019. doi: 10.1080/09593330.2017.1417491.

[15] H. Delanka-Pedige, S. Munasinghe-Arachchige, J. Cornelius, S. Henkanatte-Gedera, and et al. Pathogen reduction in an algal-based wastewater treatment system employing *Galdieria sulphuraria*. *Algal Research*, 39:101423, 2019. doi: 10.1016/j.algal.2019.101423.

[16] P. Flicek and E. Birney. Sense from sequence reads: methods for alignment and assembly. *Nature Methods*, 6:S6–S12, 2009. doi: 10.1038/nmeth.1376.

[17] M. Gong and A. Bassi. Carotenoids from microalgae: A review of recent developments. *Biotechnology Advances*, 34:1396–1412, 12 2016. doi: 10.1016/J.BIOTECHADV.2016.10.005.

[18] O. Graverholt and N. Eriksen. Heterotrophic high-cell-density fed-batch and continuous-flow cultures of *Galdieria sulphuraria* and production of phycocyanin. *Applied Microbiology and Biotechnology*, 77:69–75, 2007. doi: 10.1007/s00253-007-1150-2.

[19] W. Gross and C. Schnarrenberger. Heterotrophic growth of two strains of the acido-thermophilic red alga *Galdieria sulphuraria*. *Plant and Cell Physiology*, 36:633–638, 6 1995. doi: 10.1093/oxfordjournals.pcp.a078803.

[20] W. Gross, J. Küver, G. Tischendorf, N. Bouchaama, and W. Büsch. Cryptoendolithic growth of the red alga *Galdieria sulphuraria* in volcanic areas. *European Journal of Phycology*, 33:25–31, 1998. doi: 10.1080/09670269810001736503.

[21] K. Jain, K. Krause, F. Grewe, G. Nelson, A. Weber, A. Christensen, and J. Mower. Extreme features of the *Galdieria sulphuraria* organellar genomes: A consequence of polyextremophily? *Genome Biology and Evolution*, 7:367–380, 1 2015. doi: 10.1093/gbe/evu290.

[22] A. Karnaouri, A. Chalima, K. G. Kalogiannis, D. Varamogianni-Mamatsi, A. Lappas, and E. Topakas. Utilization of lignocellulosic biomass towards the production of omega-3 fatty acids by the heterotrophic marine microalga crypthecodinium cohnii. *Bioresource Technology*, 303:122899, 2020. doi: 10.1016/j.biortech.2020.122899.

[23] N. Khatoon and R. Pal. Microalgae in biotechnological application: A commercial approach. *Plant Biology and Biotechnology: Volume II: Plant Genomics and Biotechnology*, pages 27–47, 2015. doi: 10.1007/978-81-322-2283-5_2.

[24] H. Li, B. Handsaker, A. Wysoker, G. P. D. P. Subgroup, and et al. The sequence alignment/map format and samtools. *Bioinformatics*, 25:2078–2079, 8 2009. doi: 10.1093/bioinformatics/btp352.

[25] M. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, 15:550, 2014. doi: 10.1186/s13059-014-0550-8.

[26] H. Marc. Practical genomics | module 1 – de novo assembly of a small artifical genome. pages 1–24, 11 2019.

[27] H. Marc. Practical genomics | genome analysis module. November:1–23, 2019.

[28] S. Marguerat and J. Bähler. Rna-seq: from technology to biology. *Cellular and Molecular Life Sciences*, 67:569–579, 2010. doi: 10.1007/s00018-009-0180-6.

[29] P. A. McGettigan. Transcriptomics in the rna-seq era. *Current Opinion in Chemical Biology*, 17:4–11, 2013. doi: 10.1016/j.cbpa.2012.12.008.

[30] A. Merola, R. Castaldo, P. D. Luca, R. Gambardella, A. Musacchio, and R. Taddei. Revision of cyanidium caldarium. three species of acidophilic algae. *Giornale botanico italiano*, 115:189–195, 7 1981. doi: 10.1080/11263508109428026.

[31] S. M. Newman, J. E. Boynton, N. W. Gillham, B. L. Randolph-Anderson, A. M. Johnson, and E. H. Harris. Transformation of chloroplast ribosomal rna genes in chlamydomonas: molecular and genetic characterization of integration events. *Genetics*, 126:875–888, 12 1990. doi: 10.1093/genetics/126.4.875.

[32] V. Náhlík, V. Zachleder, M. Čížková, K. Bišová, A. Singh, D. Mezricky, T. Řezanka, and M. Vítová. Growth under different trophic regimes and synchronization of the red microalga *Galdieria sulphuraria*. *Biomolecules*, 11:939, 2021.

[33] C. Oesterhelt, E. Schmälzlin, J. M. Schmitt, and H. Lokstein. Regulation of photosynthesis in the unicellular acidophilic red alga *Galdieria sulphuraria*. *The Plant Journal*, 51:500–511, 8 2007. doi: 10.1111/j.1365-313X.2007.03159.x.

[34] A. Oshlack, M. Robinson, and M. Young. From rna-seq reads to differential expression results. *Genome Biology*, 11:220, 2010. doi: 10.1186/gb-2010-11-12-220.

[35] M. Placzek, A. Patyna, and S. Witczak. Technical evaluation of photobioreactors for microalgae cultivation. *E3S Web Conf.*, 19, 2017. doi: 10.1051/e3sconf/20171902032.

[36] H. Qiu, A. Rossoni, A. Weber, H. S. Yoon, and D. Bhattacharya. Unexpected conservation of the rna splicing apparatus in the highly streamlined genome of galdieria sulphuraria. *BMC Evolutionary Biology*, 18:41, 2018. doi: 10.1186/s12862-018-1161-x.

[37] A. Rossoni, D. Price, M. Seger, D. Lyska, P. Lammers, D. Bhattacharya, and A. Weber. The genomes of polyextremophilic cyanidiales contain 1% horizontally transferred genes with diverse adaptive functions. *Elife*, 8:e45017, 2019.

[38] A. Rossoni, G. Schönknecht, H. Lee, R. Rupp, S. Flachbart, and et al. Cold acclimation of the thermoacidophilic red alga *Galdieria sulphuraria*: Changes in gene expression and involvement of horizontally acquired genes. *Plant and Cell Physiology*, 60:702–712, 3 2019. doi: 10.1093/pcp/pcy240.

[39] M. S. Roth, S. D. Gallaher, D. J. Westcott, M. Iwai, K. B. Louie, M. Mueller, A. Walter, F. Foflonker, B. P. Bowen, N. N. Ataii, and et al. Regulation of oxygenic photosynthesis during trophic transitions in the green alga *Chromochloris zofingiensis*. *The Plant Cell*, 31:579–601, 3 2019. doi: 10.1105/tpc.18.00742.

[40] G. Salbitani and S. Carfagna. Different behaviour between autotrophic and heterotrophic *Galdieria sulphuraria*; (rhodophyta) cells to nitrogen starvation and restoration. impact on pigment and free amino acid contents. *International Journal of Plant Biology*, 11, 7 2020. doi: 10.4081/pb.2020.8567.

[41] G. Schönknecht, W. Chen, C. Ternes, G. Barbier, R. Shrestha, M. Stanke, A. Bräutigam, B. Baker, and et al. Gene transfer from bacteria and archaea facilitated evolution of an extremophilic eukaryote. *Science*, 339:1207–1210, 3 2013. doi: 10.1126/science.1231707. 10.1126/science.1231707.

[42] J. Shendure and H. Ji. Next-generation dna sequencing. *Nature Biotechnology*, 26: 1135–1145, 2008. doi: 10.1038/nbt1486.

[43] J. K. Sloth, M. G. Wiebe, and N. T. Eriksen. Accumulation of phycocyanin in heterotrophic and mixotrophic cultures of the acidophilic red alga *Galdieria sulphuraria*. *Enzyme and Microbial Technology*, 38:168–175, 2006. doi: 10.1016/j.enzmictec.2005.05.010.

[44] M. Somers, P. Chen, J. Clippinger, J. Cruce, R. Davis, P. Lammers, and J. Quinn. Techno-economic and life-cycle assessment of fuel production from mixotrophic *Galdieria sulphuraria* microalgae on hydrolysate. *Algal Research*, 59:102419, 2021. doi: 10.1016/j.algal.2021.102419.

[45] I. Stadnichuk, M. Rakhimberdieva, Y. Bolychevtseva, and et al. Inhibition by glucose of chlorophyll a and phycocyanobilin biosynthesis in the unicellular red alga *Galdieria partita* at the stage of coproporphyrinogen iii formation. *Plant Science*, 136:11–23, 1998. doi: 10.1016/S0168-9452(98)00088-0.

[46] Y. Sun, M. Shi, T. Lu, D. Ding, Y. Sun, and Y. Yuan. Bio-removal of ptcl62 complex by *Galdieria sulphuraria*. *Science of The Total Environment*, 796:149021, 2021. doi: 10.1016/j.scitotenv.2021.149021.

[47] D. Tchinda, S. Henkanatte-Gedera, I. Abeysiriwardana-Arachchige, H. Delanka-Pedige, and et al. Single-step treatment of primary effluent by *Galdieria sulphuraria*: Removal of biochemical oxygen demand, nutrients, and pathogens. *Algal Research*, 42:101578, 2019. doi: 10.1016/j.algal.2019.101578.

[48] A. Weber, C. Oesterhelt, W. Gross, A. Bräutigam, L. Imboden, I. Krassovskaya, N. Linka, J. Truchina, J. Schneidereit, and H. Voll. Est-analysis of the thermo-acidophilic red microalga *Galdieria sulphuraria* reveals potential for lipid a biosynthesis and unveils the pathway of carbon export from rhodoplasts. *Plant molecular biology*, 55:17–32, 2004.

[49] H. Zhang. Overview of sequence data formats. *Statistical Genomics*, pages 3–17, 2016.

Annexes

A1 Séquençage paired-end

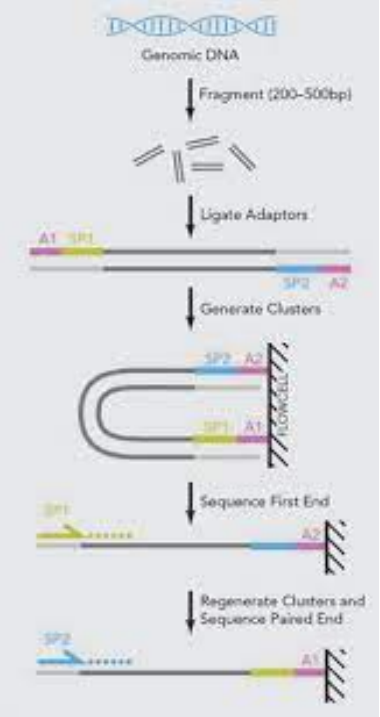

Figure A1.1: Schéma du séquençage paired-end.

A2 Détermination du nombre de reads

```
1 grep -c \@A00 *.fq
```

A3 *FastQC*

A3.1 Création d'un fichier avec les noms d'échantillon

```
1 for f in *.fq; do echo `basename $f .fq`; done | cut -f1,2 \
2 -d '_' | uniq > "nom de fichiers"
```

A3.2 Job Array

```
#!/bin/bash

#$ -S /bin/bash
#$ -V
#$ -cwd
#$ -q smallnodes.q

#$ -N fastqc_AC

#$ -t 1-18
#$ -tc 18

samples=`sed -n "${SGE_TASK_ID} p" "nom de fichier"`
samplesArray=($samples)
names=${samplesArray[0]}

fastqc ../reads/${names}_*.fq
```

A4 *Trimmomatic*

A4.1 Création d'un fichier avec les noms d'échantillon

```
for f in `cat samples_name.txt`; do echo -e "$f\t26\t5\t144\t100"; \
done > "nom complet"
```

A4.2 Job Array

```
#!/bin/bash

```

```
#$ -S /bin/bash
#$ -V
#$ -cwd

#$ -q smallnodes.q

#$ -t 1-18
#$ -tc 18

samples=`sed -n "${SGE_TASK_ID} p" "nom complet"`
samplesArray=($samples)
basename=${parameterArray[0]}
quality=${parameterArray[1]}
sliding=${parameterArray[2]}
crop=${parameterArray[3]}
minlen=${parameterArray[4]}

java -jar /media/vol1/apps/Trimmomatic-0.32/trimmomatic-0.32.jar PE \
-threads 2 -phred33 ../reads/${basename}_1.fq ../reads/${basename}_2.fq \
../reads/${basename}_1_paired.fq ../reads/${basename}_1_unpaired.fq \
../reads/${basename}_2_paired.fq ../reads/${basename}_2_unpaired.fq \
ILLUMINACLIP:/media/vol1/apps/Trimmomatic-0.32/adapters/TruSeq3-PE. \
fa:2:30:20 LEADING:${quality} TRAILING:${quality} \
SLIDINGWINDOW:$sliding:${quality} CROP:${crop} MINLEN:${minlen}
```

A5 HiSat2

A5.1 Conversion du fichier GFF en GTF

```
gffread -E "nom fichier GFF" -T -o "nom fichier GTF"
```

A5.2 Extraction de la liste d'exons et des sites d'épissage

```
eval "$(pyenv init --path)" | hisat2_extract_splice_sites.py \
"Fichier GTF" > "GS_splicing_sites.txt"

eval "$(pyenv init --path)" | hisat2_extract_exons.py "Fichier GTF" \
> "GS_exons.txt"
```

A5.3 Indexation du génome avec *hisat2-build*

```
#!/bin/bash

#$ -S /bin/bash
#$ -V
#$ -cwd
#$ -q fatnodes.q

#$ -N AC_hisat2-build

hisat2-build --ss "GS_splicing_sites.txt" --exon "GS_exons.txt" \
-f "nom fichier du génome de référence (format FNA)" "nom fichier index"
```

A5.4 Vérification de l'indexation du génome avec *hisat2-inspect*

```
#!/bin/bash

#$ -S /bin/bash
#$ -V
#$ -cwd
#$ -q smallnodes.q

#$ -N AC_hisat2-inspect
```

```

hisat2-inspect "nom fichier index"
```

A5.5 Alignement avec *hisat2*

```
#!/bin/bash

#$ -S /bin/bash
#$ -V
#$ -cwd

#$ -q fatnodes.q

#$ -t 1-18
#$ -tc 18


samples=`sed -n "${SGE_TASK_ID} p" "nom de fichier"`
samplesArray=($samples)
names=${samplesArray[0]}

hisat2 -q --threads 2 --secondary --summary_file ${names}.align.txt \
-x "nom fichier index" -1 ../reads/${names}_1_paired.fq -2 \
../reads/${names}_2_paired.fq  -S ${names}.sam
```

A6 Préparation des fichiers pour HtSeqCount avec SAMtools

A6.1 Indexation du génome de référence avec *samtools faidx*

```
#!/bin/bash

```

```
#$ -S /bin/bash
#$ -V
#$ -cwd

#$ -q smallnodes.q

#$ -N AC_ST_index

samtools faidx ../genome_ref/Galdieria_genomic.fna
```

A6.2 Conversion fichier SAM en BAM et tri des fichiers avec *samtools view et sort*

```
#!/bin/bash

#$ -S /bin/bash
#$ -V
#$ -cwd

#$ -q smallnodes.q

#$ -N AC_samtools_sort

#$ -t 1-18
#$ -tc 18

samples=`sed -n "${SGE_TASK_ID} p" nom de fichier`
samplesArray=($samples)
names=${samplesArray[0]}

samtools view -bt "fichier index SAMtools (format fai)" ${names}.sam \
> ${names}.bam

samtools sort -n ${names}.bam -o ${names}.sorted.bam
```

A7 HtSeqCount

A7.1 Job array pour *htseq-count*

```
#!/bin/bash

#$ -S /bin/bash
#$ -V
#$ -cwd

#$ -q smallnodes.q
#$ -N AC_htseq_count

#$ -t 1-18
#$ -tc 18

samples=`sed -n "${SGE_TASK_ID} p" samples_name.txt`
samplesArray=($samples)
names=${samplesArray[0]}

htseq-count -f bam -r name -s no -o ${names}.bamout ${names}.sorted.bam \
" fichier GTF"
```

A7.2 Récupération des informations et création de la matrice

```
# Récupération de la liste des gènes
cut -f1 "fichier de sortie .o" > gene_list

# Récupération des comptes de gènes
for f in `cat "nom de fichier"`; do cut -f2 "fichier de sortie .o" \
> $f.count; done

# Création de la matrice
paste gene_list *.count > htseq_count.matrix
```

A8 Le package DESeq2

A8.1 Analyse de tous les réplicats

```
# Set environnement and librairies loading
setwd("Chemin vers le fichier")
library("DESeq2")
library("limma")
library("gplots")
library("RColorBrewer")
library("genefilter")

# Load data and construct dataset
countData=read.table("htseq_count.matrix",header=FALSE, \
row.names=1,sep="\t")
colnames(countData) <- c("GSRNA_11","GSRNA_12","GSRNA_13","GSRNA_21", \
"GSRNA_22","GSRNA_23","GSRNA_31","GSRNA_32","GSRNA_33","GSRNA_41", \
"GSRNA_42","GSRNA_43","GSRNA_51","GSRNA_52","GSRNA_53","GSRNA_61", \
"GSRNA_62","GSRNA_63")

# Dataset description
condition=rep(c("1", "2", "3", "4", "5", "6"),each=3)
source1=rep(c("Glucose", "Glycerol", "Glucose", "Glucose", \
"Glycerol", "Glycerol"),each=3)
source2=rep(c("Rien", "Rien", "Glucose", "Glycerol", "Glucose", \
"Glycerol"),each=3)

# Load dataset description in a data frame
colData=data.frame(condition, source1, source2, \
col.names=names(countData))

# Model construction using DESeq
# Load the data unsing DESeqDataSetFromMatrix command
condDesign=DESeqDataSetFromMatrix(countData = countData, \
colData = colData, design = ~ condition)

# Build model using the DESeq command
cond_DESeq <- DESeq(condDesign)
```

```

# Observe parameters of the model
res <- results(cond_DESeq)
resOrdered <- res[order(res$padj),]

# Count the resulsts with padj < 0.1
sum(resOrdered$padj < 0.1, na.rm=TRUE)
# [1] 33

# Plot the normalized count for three random genes
plotCounts(cond_DESeq, gene="gene-Gasu_00010", intgroup="condition")
plotCounts(cond_DESeq, gene="gene-Gasu_22220", intgroup="condition")
plotCounts(cond_DESeq, gene="gene-Gasu_48640", intgroup="condition")

# Summary statistics of the data with PCA
rld<-rlog(cond_DESeq)
plotPCA(rld, intgroup="condition")

# Build a heatmap of sample distances
# Build sample distance
sampleDist <- dist(t(assay(rld)))

# Build heatmap
sampleDistMatrix<-as.matrix(sampleDist)
rownames(sampleDistMatrix)<-paste(rld$condition)
colnames(sampleDistMatrix)<-colnames(countData)
colours=colorRampPalette(rev(brewer.pal(9, "Blues")))(300)
heatmap.2(sampleDistMatrix, dendrogram = "both", trace = "none", \
col = colours, margin=c(6, 8))
```

A8.2 Liste de gènes différentiellement exprimés

```

# Find and export differentially expressed genes

# For Condition 1 and 2 comparison
res12=results(cond_DESeq, contrast = c("condition", "1", "2"),\
lfcThreshold = 1, alpha = 0.05)
```

```

# Down-regulated genes
fc_cond12L<- res12[which(res12$log2FoldChange < -1 & res12$padj<0.05),]

# Export data into table
write.table(fc_cond12L,"down_reg_condition_1_vs_2.txt",sep="\t")
```

A8.3 Analyse avec seulement deux réplicats

```
# Load data
countData=read.table("htseq_count.matrix",header=FALSE,\
row.names=1,sep="\t")

# New dataset with only duplicates
countData_d=countData[,-c(3,5,9,10,13,16)]
# On enlève les réplicats GSRNA_13, 22, 33, 41, 51 et 61

colnames(countData_d) <- c("GSRNA_11","GSRNA_12","GSRNA_21",\
"GSRNA_23","GSRNA_31","GSRNA_32","GSRNA_42","GSRNA_43","GSRNA_52",\
"GSRNA_53","GSRNA_62","GSRNA_63")

# Describe the dataset for each variable
conditiond=rep(c("1", "2", "3", "4", "5", "6"),each=2)
source1=rep(c("Glucose", "Glycerol", "Glucose", "Glucose",\
"Glycerol", "Glycerol"),each=2)
source2=rep(c("Rien", "Rien", "Glucose", "Glycerol", \
"Glucose", "Glycerol"),each=2)

# Load dataset description in a data frame
colData=data.frame(conditiond, source1, source2,\
col.names=names(countData_d))

# Model construction using DESeq
# Load the data unsing DESeqDataSetFromMatrix command
condDesign_d=DESeqDataSetFromMatrix(countData = countData_d,\
colData = colData, design = ~ conditiond)

```

```
# Build model using the DESeq command
cond_DESeq_d <- DESeq(condDesign_d)

# Observe parameters of the model
res_d <- results(cond_DESeq_d)
resOrdered <- res_d[order(res_d$padj),]
summary(resOrdered)

#Combien de pval < 0.1
sum(resOrdered$padj < 0.1, na.rm=TRUE)
#[1] 193

# Plot the normalized count for a random gene (same as triplicates)
plotCounts(cond_DESeq_d, gene="gene-Gasu_00010", intgroup="conditiond")
plotCounts(cond_DESeq_d, gene="gene-Gasu_22220", intgroup="conditiond")
plotCounts(cond_DESeq_d, gene="gene-Gasu_48640", intgroup="conditiond")

# Summary statistics of the data with PCA
rld_d<-rlog(cond_DESeq_d)
plotPCA(rld_d, intgroup="conditiond")
plotPCA(rld_d, intgroup="source2")

# Build a heatmap of sample distances
# Build sample distance
sampleDist <- dist(t(assay(rld_d)))
# Build heatmap
sampleDistMatrix<-as.matrix(sampleDist)
rownames(sampleDistMatrix)<-paste(rld_d$conditiond)
colnames(sampleDistMatrix)<-colnames(countData_d)
colours=colorRampPalette(rev(brewer.pal(9, "Blues")))(300)
heatmap.2(sampleDistMatrix, dendrogram = "both", trace = "none",\
col = colours, margin=c(6, 8))
```

A9 Rapports FastQC

A9.1 Avant filtration

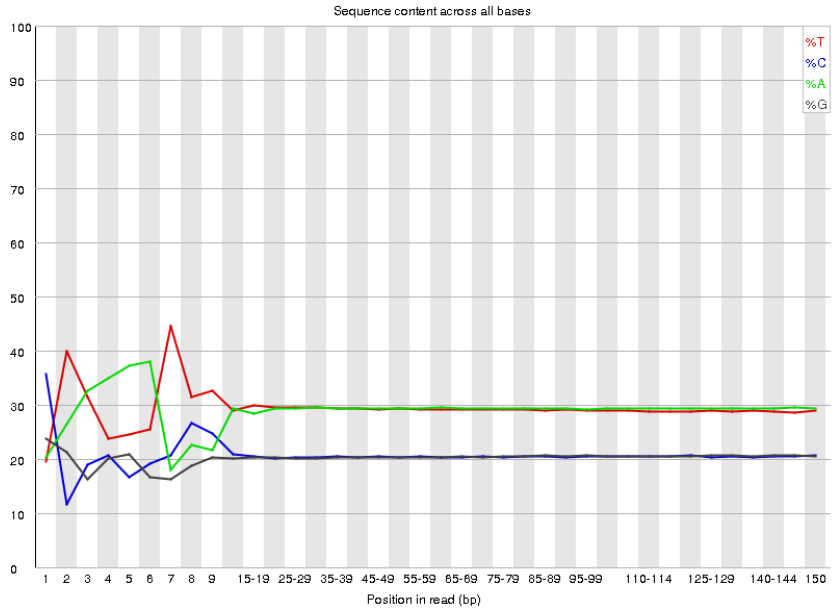

Figure A9.1: Contenu en bases.
Séquence sens de l'échantillon GSRNA_33.

A9.2 Après filtration

A9.2.1 Qualité par base

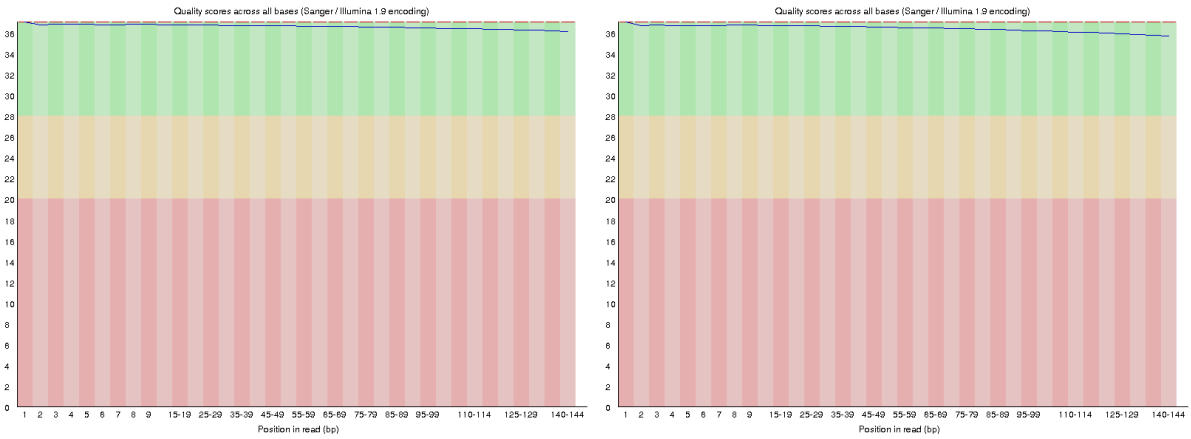

Figure A9.2: Qualité par base après filtration. Gauche: Séquence sens de l'échantillon GSRNA_31. Droite: Séquence anti-sens de l'échantillon GSRNA_53.

A9.2.2 Contenu en bases

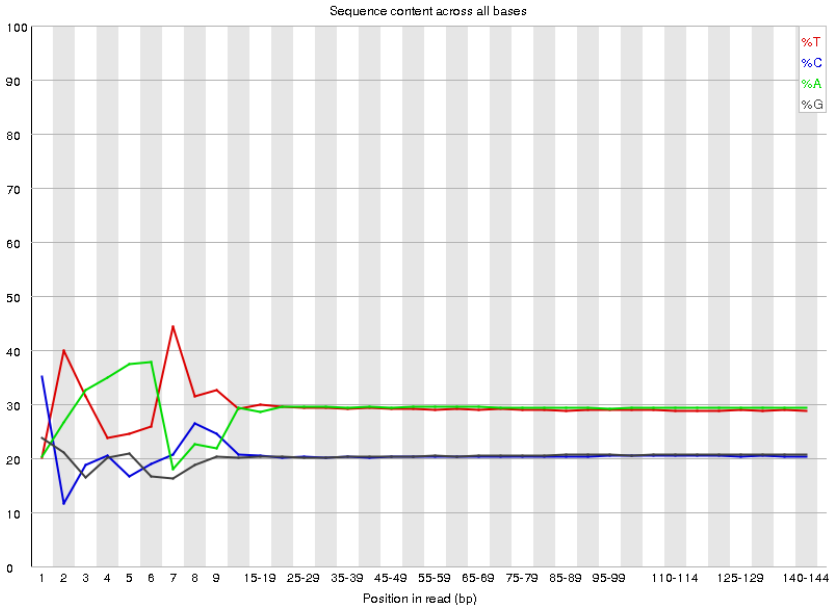

Figure A9.3: Contenu en bases après filtration.
Séquence sens de l'échantillon GSRNA_33.

A10 Fichier GTF2

```
NW_005178043.1  RefSeq  exon    201     1090    .       +       .       transcript_id "rna-XM_005674655.1"; gene_id "gene-Gasu_66230"; gene_name "Gasu_66230";
NW_005178043.1  RefSeq  CDS     215     1090    .       +       0       transcript_id "rna-XM_005674655.1"; gene_id "gene-Gasu_66230"; gene_name "Gasu_66230";
NW_005178046.1  RefSeq  exon    2       400     .       -       .       transcript_id "rna-XM_005702181.1"; gene_id "gene-Gasu_66220"; gene_name "Gasu_66220";
NW_005178046.1  RefSeq  CDS     2       378     .       -       0       transcript_id "rna-XM_005702181.1"; gene_id "gene-Gasu_66220"; gene_name "Gasu_66220";
NW_005178048.1  RefSeq  exon    1       665     .       -       .       transcript_id "rna-XM_005702182.1"; gene_id "gene-Gasu_66210"; gene_name "Gasu_66210";
NW_005178048.1  RefSeq  exon    712     820     .       -       .       transcript_id "rna-XM_005702182.1"; gene_id "gene-Gasu_66210"; gene_name "Gasu_66210";
NW_005178048.1  RefSeq  CDS     463     665     .       -       2       transcript_id "rna-XM_005702182.1"; gene_id "gene-Gasu_66210"; gene_name "Gasu_66210";
NW_005178048.1  RefSeq  CDS     712     802     .       -       0       transcript_id "rna-XM_005702182.1"; gene_id "gene-Gasu_66210"; gene_name "Gasu_66210";
NW_005178050.1  RefSeq  exon    2       868     .       +       .       transcript_id "rna-XM_005702183.1"; gene_id "gene-Gasu_66200"; gene_name "Gasu_66200";
NW_005178050.1  RefSeq  CDS     2       779     .       +       1       transcript_id "rna-XM_005702183.1"; gene_id "gene-Gasu_66200"; gene_name "Gasu_66200";
NW_005178053.1  RefSeq  exon    1       611     .       -       .       transcript_id "rna-XM_005702184.1"; gene_id "gene-Gasu_66190"; gene_name "Gasu_66190";
NW_005178053.1  RefSeq  exon    659     1685    .       -       .       transcript_id "rna-XM_005702184.1"; gene_id "gene-Gasu_66190"; gene_name "Gasu_66190";
NW_005178053.1  RefSeq  exon    1737    2130    .       -       .       transcript_id "rna-XM_005702184.1"; gene_id "gene-Gasu_66190"; gene_name "Gasu_66190";
NW_005178053.1  RefSeq  CDS     4       611     .       -       2       transcript_id "rna-XM_005702184.1"; gene_id "gene-Gasu_66190"; gene_name "Gasu_66190";
NW_005178053.1  RefSeq  CDS     659     1685    .       -       0       transcript_id "rna-XM_005702184.1"; gene_id "gene-Gasu_66190"; gene_name "Gasu_66190";
NW_005178053.1  RefSeq  CDS     1737    2117    .       -       0       transcript_id "rna-XM_005702184.1"; gene_id "gene-Gasu_66190"; gene_name "Gasu_66190";
NW_005178054.1  RefSeq  exon    1       1517    .       -       .       transcript_id "rna-XM_005702185.1"; gene_id "gene-Gasu_66180"; gene_name "Gasu_66180";
NW_005178054.1  RefSeq  exon    1571    1663    .       -       .       transcript_id "rna-XM_005702185.1"; gene_id "gene-Gasu_66180"; gene_name "Gasu_66180";
NW_005178054.1  RefSeq  exon    1707    1955    .       -       .       transcript_id "rna-XM_005702185.1"; gene_id "gene-Gasu_66180"; gene_name "Gasu_66180";
NW_005178054.1  RefSeq  exon    1988    2090    .       -       .       transcript_id "rna-XM_005702185.1"; gene_id "gene-Gasu_66180"; gene_name "Gasu_66180";
NW_005178054.1  RefSeq  CDS     106     1517    .       -       2       transcript_id "rna-XM_005702185.1"; gene_id "gene-Gasu_66180"; gene_name "Gasu_66180";
NW_005178054.1  RefSeq  CDS     1571    1663    .       -       2       transcript_id "rna-XM_005702185.1"; gene_id "gene-Gasu_66180"; gene_name "Gasu_66180";
NW_005178054.1  RefSeq  CDS     1707    1955    .       -       2       transcript_id "rna-XM_005702185.1"; gene_id "gene-Gasu_66180"; gene_name "Gasu_66180";
NW_005178054.1  RefSeq  CDS     1988    2000    .       -       0       transcript_id "rna-XM_005702185.1"; gene_id "gene-Gasu_66180"; gene_name "Gasu_66180";
NW_005178055.1  RefSeq  exon    1       530     .       -       .       transcript_id "rna-XM_005702186.1"; gene_id "gene-Gasu_66170"; gene_name "Gasu_66170";
```

Figure A10.1: Extrait du fichier GTF2.

A11 Matrice résultat de *Htseq-count*

gene-Gasu_00010	185	320	251	286	321	280	209	240	260	273	182	218	197	269	231	161	246	167
gene-Gasu_00020	127	182	156	157	217	170	102	144	140	138	116	107	111	135	139	121	133	116
gene-Gasu_00030	388	525	334	485	598	401	410	408	655	403	496	521	381	676	395	362	336	435
gene-Gasu_00040	4	15	3	8	5	2	5	3	4	4	7	12	0	6	1	7	6	5
gene-Gasu_00050	5	18	15	11	13	14	7	11	5	9	0	14	10	13	2	9	11	11
gene-Gasu_00060	239	404	307	364	447	401	341	260	389	363	294	382	365	391	391	274	348	325
gene-Gasu_00070	1	3	1	1	2	2	1	4	3	3	4	1	2	1	0	1	1	3
gene-Gasu_00080	0	1	0	2	4	6	2	1	1	1	0	4	2	1	0	0	1	0
gene-Gasu_00090	444	573	440	578	606	659	393	422	438	473	353	349	418	470	413	414	351	361
gene-Gasu_00100	1044	1405	1051	1222	1204	1050	1160	1150	1483	1145	1197	1221	1033	1717	1133	1036	937	1037
gene-Gasu_00110	680	1106	839	894	923	988	770	857	916	965	846	792	872	810	709	733	694	695
gene-Gasu_00120	789	1086	623	957	875	821	745	666	996	820	846	861	678	952	847	733	637	711
gene-Gasu_00130	1111	1804	1206	2180	1389	1730	1185	1100	1354	1471	1216	1229	1524	985	1257	1380	896	1189
gene-Gasu_00140	525	738	420	831	464	526	548	456	584	651	554	614	676	483	497	622	381	525
gene-Gasu_00150	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure A11.1: Extrait du fichier résultat de *Htseq-count*.

A12 Exemple d'une liste de gènes sous-régulés

	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
gene-Gasu_05610	110.04659	-2.758480	0.3730542	-4.713739	2.432123e-06	2.672092e-03
gene-Gasu_10210	976.28141	-4.799035	0.6100828	-6.227081	4.752069e-10	1.566282e-06
gene-Gasu_42070	453.20500	-2.695401	0.2846812	-5.955439	2.593752e-09	5.699337e-06
gene-Gasu_50510	14.72399	-3.192758	0.5437283	-4.032820	5.511158e-05	4.541194e-02
gene-Gasu_54440	88.65726	-3.991205	0.6057945	-4.937656	7.906730e-07	1.042423e-03
gene-Gasu_57950	49.42571	-7.010121	1.1233836	-5.350017	8.794606e-08	1.449351e-04
gene-Gasu_57960	221.34345	-3.133079	0.4653738	-4.583582	4.570785e-06	4.304373e-03
gene-Gasu_57970	216.24704	-9.136267	1.1030436	-7.376198	1.628741e-13	1.073666e-09

Figure A12.1: Liste de gènes sous-régulés lors d'une comparaison entre les Conditions 1 (Glucose) et 2 (Glycérol).

A13 *PlotCounts* de 3 gènes aléatoires

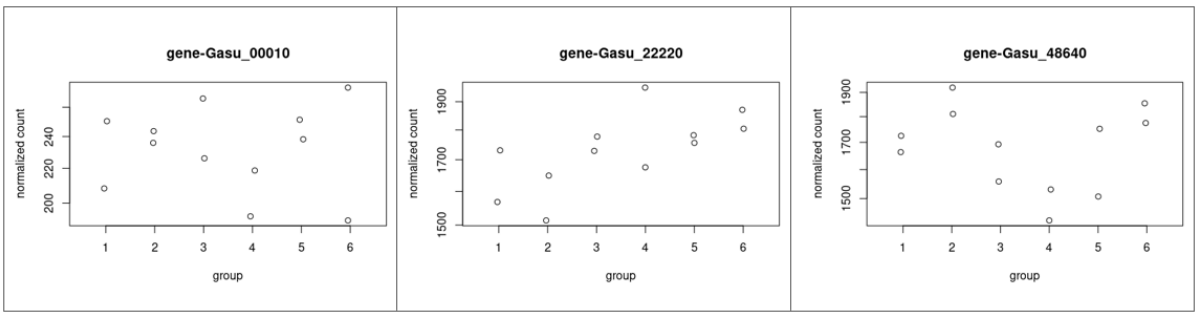

Figure A13.1: Compte des séquences pour 3 gènes aléatoires dans les différentes conditions. Group = Conditions : 1 – Glucose, 2 – Glycérol, 3 – Glucose + Glucose, 4 – Glucose + Glycérol, 5 – Glycérol + Glucose, 6 – Glycérol + Glycérol