Année 2018-2019

Rapport de stage de Master 1

Caractérisation moléculaire du processus de dégénérescence des cellules symbiotiques chez le puceron du pois

Loïc Guille

Encadrants : Nicolas Parisot et Mélanie Ribeiro Lopes

Laboratoire d'accueil : UMR 203 INRA/INSA de Lyon – BF2i
Biologie Fonctionnelle, Insectes et Interactions
INSA, Bâtiment Louis Pasteur – 69621 Villeurbanne Cedex

Résumé

Le puceron du pois, *Acyrthosiphon pisum,* vit en association symbiotique obligatoire avec une bactérie *Buchnera aphidicola* localisée dans des cellules spécialisées : les bactériocytes. Une étude récente menée au laboratoire a permis de mettre en évidence un processus inédit de mort cellulaire chez ces bactériocytes au cours du vieillissement du puceron. Cette dégénérescence des bactériocytes se réalise via une cascade d'évènements cellulaires bien caractérisés se traduisant notamment par une hypervacuolisation des cellules. L'objectif de ce stage était la caractérisation au niveau moléculaire de ce nouveau processus de mort cellulaire par l'analyse de données transcriptomiques à haut-débit (RNA-seq). Pour cela, un *pipeline* d'analyse de données RNA-seq a été développé après avoir évalué les différentes solutions disponibles. Cette analyse comparative nous a conduit à sélectionner STAR comme outil d'alignement des lectures contre le génome de référence, HTSeq-count pour quantifier l'expression de chaque gène et enfin DESeq2 pour mener l'analyse d'expression différentielle. Cette méthode a permis d'identifier 1610 gènes différentiellement exprimés au cours du processus de mort cellulaire. Nous avons ensuite caractérisé cette liste de gènes, grâce à une analyse d'enrichissement des annotations Gene Ontology, qui nous a permis de d'identifier les différents processus biologiques mis en jeu lors de la dégénérescence bactériocytaire. Une application R Shiny a également été implémentée afin de faciliter la fouille de données de nos différents échantillons ainsi que des données publiques provenant de la base de données SRA du NCBI. En plus d'apporter des premiers éléments de caractérisation moléculaire de cette nouvelle forme de mort cellulaire, l'ensemble de ce travail a également permis de mettre en évidence l'hétérogénéité importante des populations de bactériocytes face à ce processus de dégénérescence.

Remerciements

Je tiens à remercier toutes les personnes qui ont contribué à la réussite de mon stage et qui ont participé lors de la rédaction de ce rapport.

Tout d'abord je tiens à remercier vivement mes deux encadrants de stage, Mme Mélanie Ribeiro Lopes doctorante au laboratoire BF2I, et M. Nicolas Parisot chercheur au laboratoire BF2I et enseignant au département BioSciences de l'INSA de Lyon, pour leur accueil, leur bienveillance et le temps qu'ils ont consacré pour répondre à mes nombreuses questions. J'ai beaucoup apprécié faire ce stage à vos côtés et cela m'as permis d'apprendre beaucoup de choses.

Je souhaitais aussi remercier toute l'équipe SymT que dirige Mme Federica Calevro pour leurs précieux conseils, leur retour concernant mon travail et leurs disponibilités pour moi au quotidien.

Je remercie également tous les membres du laboratoire pour leur accueil chaleureux et leur aide lorsque j'en avais besoin.

Enfin je souhaitais remercier spécialement les deux membres du bureau, Carole et Nicolas pour leur accueil, leur bienveillance et leur bonne humeur au quotidien.

Table des matières

Liste des figures

Liste des tableaux

Liste des abréviations

ACP : **A**nalyse en **C**omposantes **P**rincipales

bdd : **b**ase **d**e **d**onnées

BF2I : **B**iologie **F**onctionnelle, **I**nsecte et **I**nteractions

EC : **E**nzyme **C**ommission numbers

GO : **G**ene **O**ntology

NCBI : **N**ational **C**enter for **B**iotechnology **I**nformation

pb : paire de **b**ases

PE : Paired-**E**nd

RAM : **R**andom **A**ccess **M**emory

SAM : **S**equence **A**lignment/**M**ap

SE : Single-**E**nd

SRA : **S**equence **R**ead **A**rchive

Liste des logiciels utilisés

Cytoscape version 3.7.1

DESeq2 version 1.22.2

EdgeR version 3.24.3

FeatureCounts version 1.6.4

HISAT2 version 2.1.0

HTSeq-count version 0.11.1

R version 3.5.3

Shiny version 1.3.2

STAR version 2.4.0.1

Subread version 1.6.4

Trimmomatic version 0.39

1. Introduction

Les associations symbiotiques entre insectes et bactéries sont très répandues (Moran et Telang 1998) et constituent un moteur majeur du succès écologique des insectes qui représentent aujourd'hui 80% de la biodiversité. Ces relations symbiotiques ont par exemple permis le développement de certains insectes au sein de niches trophiques nutritionnellement déséquilibrées comme le sang ou la sève des plantes (Gündüz et Douglas 2009). La bactérie endosymbiotique complémente alors les voies métaboliques de l'insecte en lui fournissant des composés peu présents dans son milieu nutritif, comme des acides aminés essentiels et des vitamines. En retour, l'insecte procure à ses symbiotes un approvisionnement permanent en matières nutritives simples comme les sucres et leur assure une niche écologique souvent peu compétitive. Certaines de ces associations ont atteint un tel niveau d'intégration qu'elles sont qualifiées d'obligatoires, les insectes hôtes étant totalement dépendants de leurs symbiotes pour assurer leur développement et leur reproduction.

De manière fascinante, ces symbioses obligatoires d'insectes ont conduit, au cours de l'évolution, à l'émergence de cellules spécialisées dans lesquelles les symbiotes (alors appelés endosymbiotes) sont hébergés. Ces cellules, communément appelées bactériocytes (Buchner 1965, Matsuura, et al. 2015), demeurent une véritable source de questionnement quant à leur origine embryonnaire, les processus caractérisant leur développement, leur organogénèse ainsi que leur dégénérescence.

Au cours des dernières années, l'association symbiotique entre le puceron du pois *Acyrthosiphon pisum* et la protéobactérie *Buchnera aphidicola* a émergé comme un modèle particulièrement adapté à l'étude de ces cellules bactériocytaires. En effet, les bactériocytes du puceron sont des cellules géantes (pouvant excéder 100 µm de diamètre), facilement isolables par dissection. De plus, la disponibilité des génomes des deux partenaires symbiotiques (Shigenobu et al. 2000 ; IAGC 2010) permet de combiner approches cellulaires et moléculaires afin d'étudier les mécanismes régulant l'homéostasie bactériocytaire.

Dans une étude récente, le laboratoire BF2i a pu caractériser la dynamique couplée des bactériocytes et des endosymbiotes tout au long de la vie du puceron (Simonet, et al. 2016). Cette analyse a permis de mettre en évidence une coordination de ces deux populations en lien avec les besoins développementaux du puceron. Ainsi, il a été montré que le nombre ainsi que le volume des bactériocytes augmente parallèlement à celui de la population de *B. aphidicola* depuis le stade embryonnaire jusqu'au stade adulte. Cette phase de croissance répond aux

besoins nutritionnels du jeune puceron en matière d'acides aminés essentiels produits par le métabolisme symbiotique et nécessaires à son développement et sa reproduction. Ensuite, les bactériocytes d'*A. pisum* ne disparaissent pas au cours du vieillissement, contrairement à d'autres modèles de symbiose, mais leur nombre décroit progressivement jusqu'à la mort du puceron. Cette phase dégénérative se caractérise par une réduction du nombre et du volume des bactériocytes. L'étude menée au sein du laboratoire a également pu montrer que les bactériocytes, lors de cette phase dégénérative, subissent d'importantes modifications morphologiques liées à l'apparition de vacuoles suggérant ainsi l'existence, dans les bactériocytes d'*A. pisum*, d'un processus de mort cellulaire inédit chez les insectes.

Bien que ce processus de mort cellulaire ait été finement caractérisé à l'échelle cellulaire (Simonet, et al. 2018), les bases moléculaires de la dégénérescence bactériocytaire restent encore inconnues. C'est pourquoi le laboratoire BF2i a entrepris récemment la caractérisation moléculaire de cette mort cellulaire inédite chez les insectes par des approches de séquençage à haut-débit du transcriptome (RNA-seq). Ainsi, des bactériocytes ont été prélevés à partir de pucerons adultes âgés de 9, 15 et 23 jours correspondant respectivement aux stades précoces, intermédiaires et avancés de la dégénérescence bactériocytaire.

Mon travail au cours de ce stage a donc consisté à analyser les données RNA-seq obtenues afin de caractériser au niveau moléculaire le processus de dégénérescence des cellules symbiotiques du puceron du pois. Pour cela, j'ai dû, tout d'abord, mettre en place un *pipeline* d'analyse des données RNA-seq générées après avoir identifié et testé les outils les plus adaptés à notre jeu de données. Ensuite, j'ai pu appliquer le *pipeline* développé aux données RNA-seq afin de mettre en évidence les gènes différentiellement régulés au cours du processus de mort cellulaire. Enfin, j'ai développé une plateforme de visualisation et de fouille de ces résultats.

2. Matériels et méthodes

2.1 Données génomiques

Le génome d'*Acyrthosiphon pisum* a été publié en 2010 (IAGC 2010). Il a été assemblé en 23925 scaffolds pour une taille totale de 541 millions de nucléotides (assemblage Acyr_2.0, identifiant NCBI GCA_000142985.2). Il existe deux versions de l'annotation des gènes pour ce génome :

- la première provient du consortium de séquençage et d'annotation (IAGC 2010);
    - Annotation AphidBase v2.1b contenant 36 195 gènes.
- tandis que la deuxième correspond à un processus automatisé d'annotation réalisé par la banque de données du NCBI
    - Annotation NCBI Release 102 contenant 20 923 gènes.

2.2 Données transcriptomiques

Les données transcriptomiques RNA-seq ont été obtenues à partir de tissus bactériocytaires prélevés, en triplicats, sur des pucerons âgés de 9, 15 et 23 jours (échantillons J9, J15 et J23 respectivement). Pour chaque échantillon, 200 bactériocytes ont été prélevés à partir de 5 individus. Suite à un séquençage Illumina de lectures de 50pb en single-end, 9 fichiers au format FASTQ avec environ 40 millions de séquences pour chaque échantillon (Tableau 1) ont été obtenus. Le format FASTQ est un format de fichier permettant le stockage des lectures séquencées ainsi que des scores de qualité de prédiction associés.

2.3 Nettoyage du jeu de données

Avant de commencer l'analyse de nos fichiers de séquençage il est nécessaire de contrôler la qualité des séquences qu'ils contiennent afin de supprimer les séquences ou portions de séquences de mauvaise qualité ainsi que les éventuels adaptateurs de séquençage restants. Afin de nettoyer les fichiers FASTQ obtenus à partir des différentes conditions (J9, J15 et J23), nous avons utilisé Trimmomatic version 0.39. Trimmomatic est un logiciel rapide et parallélisable dédié au contrôle qualité des lectures de séquençage Illumina (Bolger, Lohse et Usadel 2014). Cet outil a été utilisé avec les options suivantes :

- `SE/PE` : qui indique si les fichiers sont en single-end (SE) ou en paired-end (PE). Dans notre cas, `SE` a été choisi.
- `LEADING:20` : tronque les premiers nucléotides de la lecture si leur score de qualité est inférieur à 20.

- `TRAILING:20` : tronque les derniers nucléotides de la lecture si leur score de qualité est inférieur à 20.
- `AVGQUAL:25` : supprime la lecture si la moyenne des scores de qualité de celle-ci est inférieure à 25.
- `SLIDINGWINDOW:10:30` : cette option génère une fenêtre glissante sur 10 nucléotides à partir de l'extrémité 5' de la lecture qui va tronquer l'extrémité 3' de la lecture dès lors qu'elle rencontrera une fenêtre de 10 nucléotides pour laquelle la qualité moyenne est inférieure ou égale à 30.
- `MINLEN:36` : supprime la lecture si sa taille est inférieure à 36 pb.

2.4 Positionnement des lectures sur le génome de référence

Une fois l'étape de nettoyage des séquences achevée, il est nécessaire de positionner les lectures sur le génome de référence. Cette étape est couramment appelée *mapping*.

2.4.1 Utilisation de HISAT2

HISAT2 est un outil de *mapping* développé par une équipe de l'université John Hopkins (Kim, Langmead et Salzberg 2015). Pour réaliser le *mapping*, la première étape consiste à construire un index du génome de référence. Un seul paramètre a été utilisé lors de sa création, il s'agit du nombre de processus à utiliser. Dans notre étude, nous utiliserons 8 processus. La deuxième étape consiste à aligner nos lectures sur le génome de référence grâce à l'index créé précédemment. Les paramètres qui ont été utilisés sont les suivants :

- `x` : spécifie le fichier d'index
- `q` : fichier de lectures au format FASTQ
- `phred33` : spécifie que les scores de qualité sont au format phred 33
- `k 20` : nombre maximum d'alignements primaires
- `no-unal` : ne retourne pas les lectures non mappées
- `S` : fichier de sortie au format SAM

Le format SAM est un format tabulé permettant de stocker les informations d'alignement des lectures sur une séquence génomique de référence.

Plusieurs jeux de paramètres ont été utilisés d'après l'article de Baruzzo et collaborateurs (Baruzzo, et al. 2017). Les paramètres additionnels sont les suivants :

- `pen-noncansplice 20` : spécifie la pénalité pour chaque paire de site d'épissage non canonique

- `mp 1,0` : spécifie le score maximum et minimum de pénalité pour un mésappariement
- `sp 3,0` : spécifie le score maximum et minimum de pénalité pour de l'alignement soft-clipping par base

2.4.2 Utilisation de STAR

Le logiciel STAR est un logiciel de *mapping* développé par Alexander Dobin (Dobin, et al. 2013). Comme pour HISAT2, la première étape consiste à construire un index. Les options utilisées pour construire cet index sont :

- `runThreadN 8` : utilisation de 8 processeurs pour construire l'index
- `runMode genomeGenerate` : utilisation du mode de création d'index
- `genomeDir` : nom du répertoire contenant l'index du génome
- `genomeFastaFiles` : chemin vers le fichier d'assemblage du génome
- `sjdbGTFfile` : chemin vers le fichier d'annotation du génome
- `sjdbGTFtagExonParentTranscript Parent` : précise le champ dans lequel est renseigné le nom du gène dans le fichier d'annotation du génome.
- `sjdbOverhang 50` : spécifie la longueur de la séquence génomique à utiliser pour construire la base de données de jonctions d'épissage.
- `genomeChrBinNbits 14` : permet de réduire la RAM utilisée.

Une fois que cet index a été généré, STAR passe à l'étape d'alignement de nos données de séquençage. Les options qui ont été utilisées sont les suivantes :

- `runThreadN 8`
- `genomeDir`
- `readsFilesIn` : chemin vers les fichiers de lectures à aligner.
- `outFileNamePrefix` : chemin vers les fichiers de résultats.
- `outSAMtype BAM SortedByCoordinate` : spécifie que les fichiers de sortie sont au format BAM trié par coordonnées génomiques.
- `outFilterMultimapMax 20` : définit un nombre maximum de positions maximum pour une lecture (*multimapping*) à 20 fois, au-dessus la lecture est supprimée.
- `outMultimapperOrder Random` : Ordonne aléatoirement les positions équivalentes (*multimapping*) pour une même lecture.
- `outFilterMismatchNmax 5` : Exclue les séquences dont le nombre de mésappariements est supérieur à 5.

Comme pour HISAT2, plusieurs jeux de paramètres ont été utilisés d'après l'article de Baruzzo et collaborateurs (Baruzzo, et al. 2017). En plus des paramètres précédents, les options qui ont été utilisées sont les suivantes :

- `outFilterMismatchNmax 17` : exclue les séquences si le nombre de mésappariements est supérieur à 17.
- `seedSearchStartLmax 30` : définit la taille de la *seed*.
- `alignSJoverhangMin 15` : minimum de taille de bloc pour l'épissage alternatif.
- `outFilterMatchNminOverLread 0` : sortie d'alignement présente seulement si le nombre d'alignements positifs est supérieur ou égal à cette valeur (normalisé sur la longueur de la lecture).
- `outFilterScoreMinOverLread 0.3` : alignement présent en sortie seulement si le score est supérieur ou égal à cette valeur (normalisé sur la longueur de la lecture).

Une fois ces alignements effectués, on obtient en sortie des fichiers au format BAM qui est un format compressé du format SAM.

2.5 Quantification de l'expression des gènes

Pour générer les matrices d'expression qui seront utilisées pour l'analyse d'expression différentielle, il est nécessaire d'utiliser des programmes qui vont permettre d'assigner chaque lecture à une région d'intérêt (ici un gène) grâce à sa position sur le génome obtenu via le *mapping*. On peut ainsi compter combien de lectures ont été alignées sur chaque gène et obtenir une estimation du niveau d'expression de chaque gène. Le problème majeur dans ce type d'analyse est la gestion des lectures pouvant être positionnées à plusieurs coordonnées sur le génome. Différents programmes sont utilisés à ce jour, et leur gestion de ce problème divergent.

2.5.1 Utilisation de FeatureCounts

FeatureCounts (Liao, Smyth et Shi 2014) est un logiciel de la suite Subread. Ce programme permet l'obtention de matrices de comptages qui pourront ensuite être utilisées pour l'analyse d'expression différentielle. Les paramètres qui ont été utilisés sont :

- `T` : nombre de processus à utiliser (8)
- `a` : fichier d'annotation
- `o` : fichier de sortie

Pour pouvoir utiliser featureCounts à partir de l'annotation provenant du NCBI il est nécessaire de rajouter le paramètre `-g` qui indique le nom du champ qui va être utilisé comme nom de gène pour le comptage. Dans notre cas, on utilisera `'transcript_id'`.

2.5.2 Utilisation de HTSeq-count

HTSeq-count (Anders, Pyl et Huber 2014) est un logiciel qui va permettre le comptage des lectures associées à chaque gène afin de faciliter l'analyse d'expression différentielle. Les paramètres qui ont été utilisés sont les suivants :

- `f` : spécifie le format du fichier d'entrée
- `s` : spécifie si les données en entrée sont issues d'un seul brin ou non
- `i` : spécifie le nom du champ à utiliser comme nom de gène pour le comptage (seulement pour le cas de l'annotation provenant du NCBI)

2.6 Analyse d'expression différentielle

Une fois les matrices de comptages obtenues, l'étape suivante consiste à réaliser un test statistique pour évaluer si l'expression de chacun des gènes de l'organisme étudié est significativement différente dans une condition par rapport à l'autre. Il existe plusieurs packages R pour cette étape, comme DESeq2 (Love, Huber et Anders 2014) ou EdgeR (Robinson MD 2010) qui ont été utilisés avec les paramètres par défaut.

2.7 Caractérisation de la liste de gènes différentiellement exprimés

Après avoir obtenu la liste de gènes différentiellement exprimés entre les conditions testées, il est nécessaire de les analyser d'un point de vue fonctionnel. On va alors chercher à savoir si ces gènes appartiennent à un processus biologique, ou une voie métabolique en particulier. Ainsi, la surreprésentation de chaque annotation fonctionnelle dans les listes de gènes différentiellement exprimés est évaluée par un test hypergéométrique sous R. Les annotations testées appartiennent au système Gene Ontology (GO) dont le vocabulaire est standardisé et structuré (Ashburner, et al. 2000). Les listes de termes GO significativement enrichis sont ensuite réduites en utilisant REVIGO (Supek, et al. 2011). La représentation graphique des termes GO est réalisée via Cytoscape (Shannon, et al. 2003).

3. Résultats

Le jeu de données à analyser correspond à trois stades de dégénérescence des bactériocytes du puceron : stade précoce (bactériocytes de pucerons âgés de 9 jours, J9), intermédiaire (J15) et avancé (J23). Ces stades ont été analysés en triplicats pour un total de 9 banques RNA-seq séquencées et disponibles au format FASTQ. Mon travail au cours de ce stage a consisté à mettre en place un chainage d'analyses de ces données RNA-seq afin de caractériser au niveau moléculaire le processus de dégénérescence des cellules symbiotiques du puceron du pois.

3.1 Analyse préliminaire du jeu de données

A mon arrivée au laboratoire une première analyse globale du jeu de données avait déjà été réalisée. Le contrôle qualité des séquences avec Trimmomatic (Bolger, Lohse et Usadel 2014) avait pu mettre en évidence la qualité des données de séquençage avec seulement environ 3% de lectures supprimées (Tableau 1).

Tableau 1 : Nombre de lectures par échantillon après contrôle qualité.

Échantillon	Nombre total de lectures après séquençage	Lectures supprimées après contrôle qualité (%)
J9.1	47 448 635	2.78
J9.2	44 561 907	3.13
J9.3	40 369 174	3.10
J15.1	39 951 674	3.19
J15.2	45 552 712	2.95
J15.5	42 967 008	3.14
J23.2	46 189 375	3.01
J23.3	31 551 743	3.09
J23.4	41 207 604	2.94

Cette analyse préliminaire avait également permis la construction d'une analyse en composantes principales de l'ensemble des gènes entre toutes les conditions (Figure 1) ainsi que d'une heatmap des 1000 gènes avec la plus grande variance entre les différentes conditions (Figure 2).

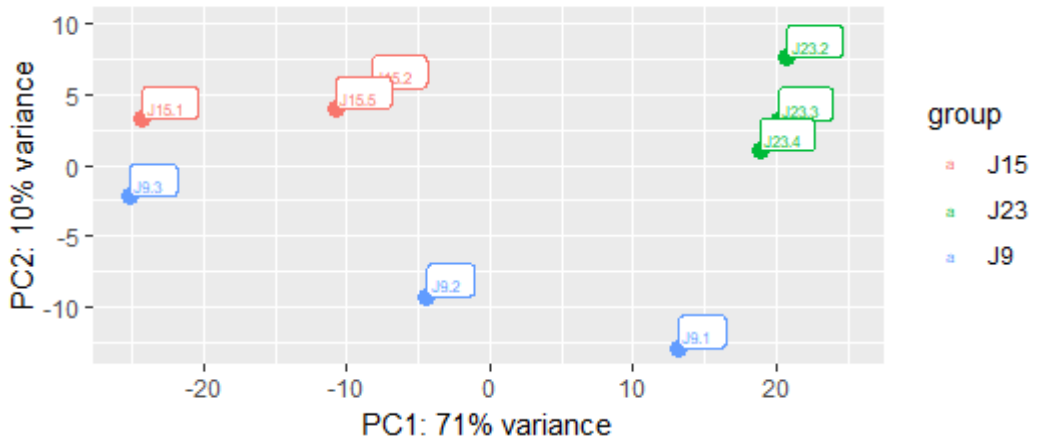


Figure 1 : Analyse en composantes principales des profils d'expression de tous les échantillons.

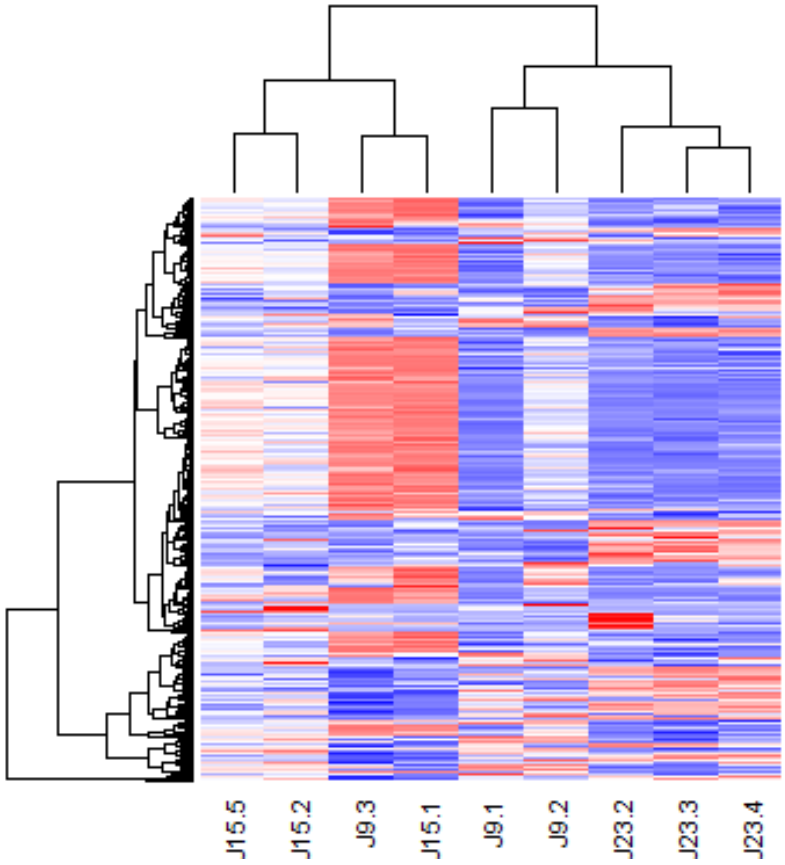


Figure 2 : Heatmap des 1000 gènes avec la plus grande variance entre tous les échantillons.

Sur ces deux représentations on peut observer une grande variabilité entre les réplicats des conditions J9 et J15 alors que les réplicats des échantillons J23 semblent très proches. Dans un premier temps, pour la mise en place du *pipeline* d'analyse de données RNA-seq, nous évaluerons les outils sur le jeu de données en excluant les trois échantillons J9 qui présentent

trop de variabilité ainsi que l'échantillon J15.1 qui semble aberrant par rapport aux autres réplicats de la condition J15.

3.2. Étude bibliographique des solutions disponibles

La figure 3 résume les différentes étapes d'une analyse de données RNA-seq. Afin de construire un *pipeline* optimal pour l'analyse de ce jeu de données, nous avons tout d'abord réalisé une étude bibliographique en essayant d'identifier, pour chaque étape, les outils les plus pertinents et les plus adaptés à notre jeu de données.

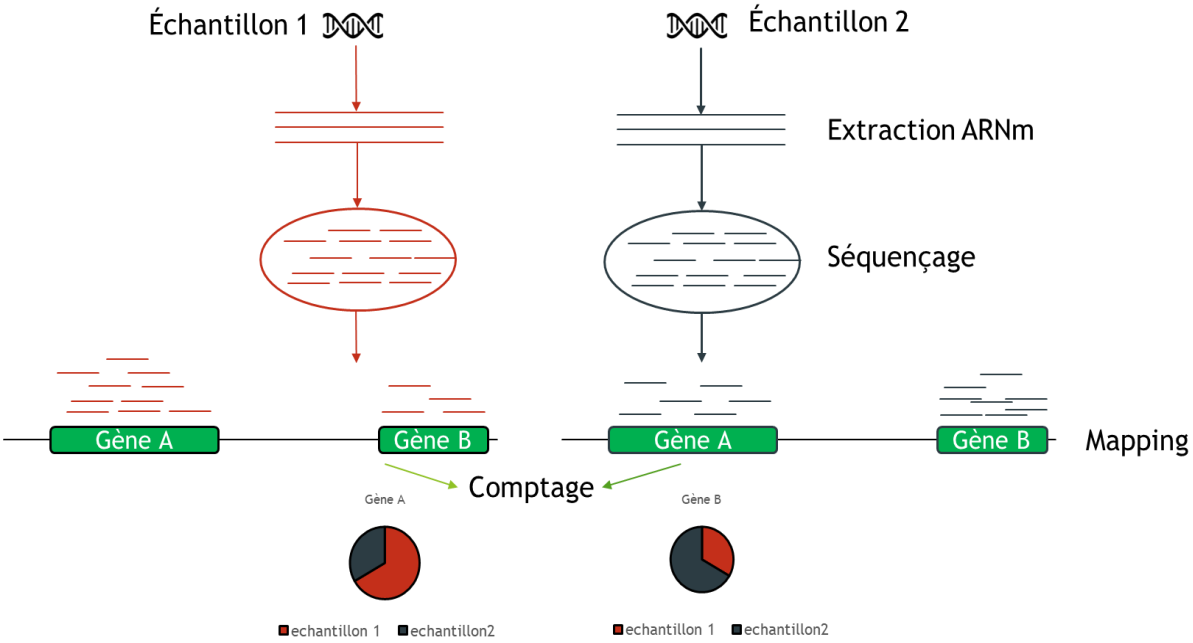


Figure 3 : Schéma représentant les différentes étapes d'une analyse de données RNA-seq.

3.2.1 Alignement contre le génome de référence

Premièrement, il est nécessaire de positionner les lectures des différentes banques RNA-seq sur le génome d'*A. pisum*.

D'après la littérature, HISAT2 (Kim, Langmead et Salzberg 2015) et STAR (Dobin, et al. 2013) semblent actuellement parmi les outils les plus performants et les plus utilisés (Grant, et al. 2011, Pär G Engström, et al. 2013, Baruzzo, et al. 2017).

HISAT2 est un outil d'alignement représentant un bon compromis entre performance et mémoire utilisée. STAR quant à lui est très rapide et dispose de paramètres par défaut très performants.

De plus, les travaux de Baruzzo et collaborateurs (Baruzzo, et al. 2017) sur la comparaison de différents outils de *mapping* sur des jeux de données simulés a permis

d'identifier des paramètres optimaux pour ces deux outils. Nous avons donc testé ces deux outils d'alignement avec leurs paramètres par défaut ainsi qu'avec des paramètres optimisés. L'outil le plus performant et son meilleur paramétrage sera intégré au *pipeline* d'analyse.

3.2.2 Quantification de l'expression des gènes

A partir des résultats de *mapping*, l'étape suivante consiste à compter le nombre de lectures alignées sur chaque gène afin d'estimer son niveau d'expression.

L'analyse bibliographique menée sur cette étape (Chandramohan, et al. 2013, Isaac A. Babarinde 2019) a permis d'identifier deux outils principaux : FeatureCounts (Liao, Smyth et Shi 2014) et de HTSeq-count (Anders, Pyl et Huber 2014).

FeatureCounts et HTSeq-Count sont deux programmes semblables dans leurs approches. Par défaut, HTSeq-count ne tient pas compte des séquences s'alignant à plusieurs positions. Cette stratégie aura donc tendance à sous-évaluer l'expression des gènes. Cependant, ce comportement par défaut peut être modifié grâce à l'option `--nonunique all` (Anders, Pyl et Huber 2014). FeatureCounts permet une gestion plus efficace des lectures s'alignant à plusieurs positions génomiques, notamment si les données en entrée sont des données paired-end (Liao, Smyth et Shi 2014).

La gestion des lectures s'alignant à plusieurs positions ne faisant pas consensus dans la communauté scientifique (Zytnicki 2017), et puisque leurs performances semblent équivalentes (Liao, Smyth et Shi 2014, Germain, et al. 2016, Gu 2016), les deux outils seront évalués pour notre analyse. L'outil le plus performant sur notre jeu de données sera ensuite intégré au *pipeline* d'analyse.

Le génome du puceron du pois dispose aujourd'hui de deux versions d'annotations concurrentes, celle du consortium d'annotation AphidBase et la version d'annotation automatisée du NCBI. Pour la comparaison des différents outils de comptage nous travaillerons donc en parallèle sur ces deux versions d'annotation.

3.2.3 Analyse d'expression différentielle

La dernière étape du *pipeline* d'analyse consiste à évaluer statistiquement la différence d'expression de chacun des gènes entre les conditions expérimentales à partir des comptages de lectures.

Une fois encore, nous avons sélectionné les deux outils les plus pertinents. Il s'agit de deux *packages* R (R Development Core Team 2005) : DESeq2 (Love, Huber et Anders 2014)

et edgeR (Robinson MD 2010). Ces deux outils se basent sur une loi binomiale négative pour réaliser l'analyse d'expression différentielle.

Nous comparerons donc ces deux outils avec leurs paramètres par défaut.

3.3. Évaluation des différents *pipelines* d'analyse RNA-seq

Après avoir sélectionné les outils les plus pertinents pour chacune des étapes de l'analyse, il est maintenant nécessaire d'évaluer leur performance sur notre jeu de données. Pour cela, toutes les combinaisons d'outils ont été testées (Annexe 1).

Au total, ce sont donc 32 combinaisons différentes qui ont été testées lors de cette analyse. Les résultats obtenus sont présentés dans le tableau 2.

Tableau 2 : Résultats de chaque étape des différents *pipelines* utilisés.

Pipeline	**Lectures alignées une fois (%)**	**Lectures alignées plusieurs fois (%)**	**Lectures comptées dans des gènes (%)**	**Nombre de gènes différentiellement exprimés avec DESeq2**	**Nombre de gènes différentiellement exprimés avec EdgeR**
S-AB-F	74.85	11.96	62.48	1610	1005
S-NCBI-F	74.91	11.91	38.98	970	583
S-AB-Ht	74.85	11.96	85.00	1610	1008
S-NCBI-Ht	74.91	11.91	53.00	970	569
H-AB-F	74.26	11.15	56.78	1594	999
H-NCBI-F	74.26	11.15	35.76	966	574
H-AB-Ht	74.26	11.15	85.00	1594	1000
H-NCBI-Ht	74.26	11.15	54.00	966	574
Spp-AB-F	78.34	19.78	47.34	1458	923
Spp-NCBI-F	78.38	19.73	29.46	904	538
Spp-AB-Ht	78.34	19.78	82.00	1458	908
Spp-NCBI-Ht	78.38	19.73	51.00	961	532
Hpp-AB-F	73.84	12.78	55.14	1616	971
Hpp-NCBI-F	73.84	12.78	34.4	964	569
Hpp-AB-Ht	73.84	12.78	83.00	1458	989
Hpp-NCBI-Ht	73.84	12.78	52.00	964	567

S : STAR ; H : HISAT2 ; pp : Paramètres publication ; AB : Annotation AphidBase ; NCBI : Annotation NCBI ; F : FeatureCounts ; Ht : HTSeq-count

Les résultats de *mapping* obtenus (Tableau 2) montrent que le taux de lectures alignées de façon unique est pratiquement le même que l'on aligne avec STAR ou avec HISAT2 (75% contre 74%) en gardant les paramètres par défaut. Si l'on optimise ces paramètres, le taux de lectures alignées une seule fois augmente avec STAR (78%) et reste stationnaire avec HISAT2. Néanmoins on s'aperçoit que pour STAR, cette augmentation de 4% du nombre de lectures alignées de manière unique s'accompagne d'une augmentation de presque 8% du nombre de lectures s'alignant à plusieurs positions.

En ce qui concerne l'étape de quantification d'expression, les résultats dépendent de deux facteurs. Le premier correspond à l'annotation qui est utilisée. En effet, si l'on utilise l'annotation AphidBase, les résultats sont meilleurs par rapport à l'annotation NCBI (26% de lectures comptées en plus). Cette différence s'explique par le fait que l'annotation du NCBI ne recense que 20 923 gènes pour le puceron alors que celle du consortium AphidBase en recense 36 195. Le deuxième facteur qui influence fortement le pourcentage de lectures comptées correspond à l'outil utilisé. En effet, HTSeq-count semble permettre de comptabiliser plus de lectures que FeatureCounts (23% d'écart).

Les résultats d'expression différentielle varient eux aussi en fonction de la version d'annotation mais aussi du *package* R utilisé. Ainsi, on observe 40% de gènes différentiellement exprimés en moins en utilisant l'annotation du NCBI par rapport à l'annotation AphidBase. De plus, on trouve globalement moins de gènes différentiellement exprimés avec edgeR (38% en moins) comparativement à l'utilisation de DESeq2.

En plus des critères présentés dans le tableau 2, le temps d'exécution des différents outils peut rentrer en jeu dans la sélection du *pipeline* le plus performant. Alors que les autres étapes sont relativement rapides, le *mapping* apparait comme l'étape la plus coûteuse en temps de calculs. Nous avons donc comparé les temps de calculs des deux outils de *mapping* avec leurs paramètres par défaut ainsi qu'avec les paramètres optimisés (Tableau 3).

Tableau 3 : Temps de calcul moyen pour les différents outils de *mapping*. Les analyses ont été réalisées sur un serveur Dell R730 sous Debian 8 avec 14 CPUs et 120Go de RAM. Chaque outil a été exécuté avec 8 processus en parallèle.

Techniques utilisés	Temps moyen par échantillon (min)
STAR	9
HISAT2	138
STAR paramètres optimisés	9
HISAT2 paramètres optimisés	323

On observe alors que le temps de calcul est multiplié par 15 entre STAR et HISAT2 dans le meilleur des cas. Nous avons donc choisi d'utiliser STAR dans notre *pipeline*. Notre choix s'est également porté sur l'utilisation des paramètres par défaut de STAR afin d'éviter un trop grand nombre de lectures s'alignant à plusieurs positions génomiques.

Concernant l'outil de quantification d'expression, nous avons choisi d'utiliser HTSeq-count qui offre les meilleurs résultats. Enfin, en ce qui concerne l'outil d'analyse d'expression différentielle, nous avons sélectionné DESeq2 afin de maximiser le nombre de gènes différentiellement exprimés identifiés.

Le *pipeline* final est présenté en figure 4.

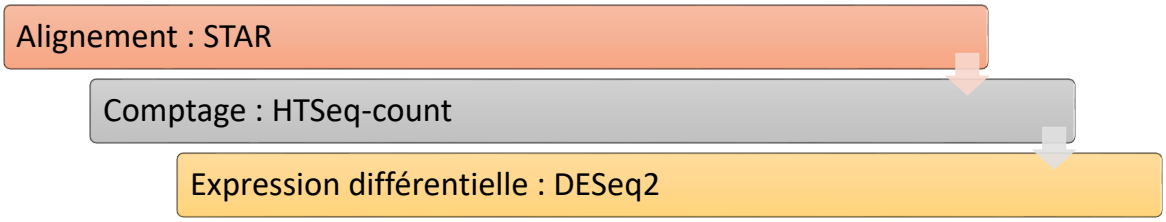


Figure 4 : Schéma représentant le *pipeline* sélectionné pour l'analyse des données RNA-seq.

3.4. Application à la caractérisation de la dégénérescence bactériocytaire

Après avoir déterminé le *pipeline* d'analyse RNA-seq le plus adapté, nous pouvons dorénavant l'appliquer à l'étude de la dégénérescence bactériocytaire chez le puceron du pois.

3.4.1 Analyse des trois stades d'avancement de la dégénérescence bactériocytaire

Dans un premier temps, nous appliquerons le *pipeline* à l'ensemble des 9 échantillons disponibles. Les résultats sont présentés dans le tableau 4.

Tableau 4 : Résultats obtenus sur les échantillons après application du *pipeline*.

Echantillon	Lectures alignées une fois (%)	Lectures alignées plusieurs fois (%)	Lectures comptées (%)
J9.1	68.64	18.47	51.80
J9.2	76.49	11.72	63.80
J9.3	71.14	14.94	56.70
J15.1	71.18	11.36	61.80
J15.2	75.69	14.01	60.60
J15.5	73.63	12.46	61.60
J23.2	72.99	11.20	63.30
J23.3	78.88	10.24	65.90
J23.4	73.08	11.89	62.50

Nous avons ensuite voulu identifier les gènes différentiellement exprimés entre les différentes conditions : J9 *versus* J15, J9 *versus* J23 et J15 *versus* J23 (Tableau 5).

Tableau 5 : Nombre de gènes différentiellement exprimés selon les conditions comparées.

Contraste : C1 *vs* C2	Nombre de gènes sous-exprimés en C1 *vs* C2	Nombres de gènes sur-exprimés en C1 *vs* C2
J9 vs J15	35	26
J9 vs J23	648	565
J15 vs J23	654	518

On observe alors que les comparaisons impliquant les échantillons J23 identifient plus de gènes différentiellement exprimés que la comparaison J9 versus J15. Outre une potentielle explication biologique, nous ne pouvons pas exclure à cette étape que ce résultat soit lié à la très forte variabilité observée au sein des échantillons J9 et J15 (Figure 1). Cette variabilité limite la sensibilité des analyses d'expression différentielle et ne permet donc d'identifier qu'un nombre limité de gènes différentiellement exprimés.

Ces résultats confortent donc le choix d'exclure les échantillons de la condition J9 et l'échantillon J15.1 pour notre analyse d'expression différentielle afin d'obtenir des résultats plus robustes.

3.4.2 Comparaison des stades intermédiaires et avancés de dégénérescence

Les résultats de l'expression des gènes montrent 1610 gènes différentiellement exprimés entre les échantillons des conditions J15 et J23 dont 812 gènes sous-exprimés au jour 15 par rapport au jour 23 et 798 gènes sur-exprimés.

Nous avons ensuite voulu caractériser cette liste de gènes en identifiant les processus biologiques différentiellement régulés au cours du processus de mort cellulaire. Pour cela nous avons entrepris une analyse d'enrichissement des annotations Gene Ontology (GO) disponibles pour le puceron du pois. Cette analyse a permis d'identifier 283 termes GO statistiquement sur-représentés parmi les gènes différentiellement exprimés. Après analyse avec l'outil REVIGO (Supek, et al. 2011) nous avons pu aboutir à une liste de 89 processus biologiques différentiellement régulés. On retrouve des processus biologiques tels que le transport membranaire, la régulation de la transcription et la phosphorylation des protéines (Figure 5).

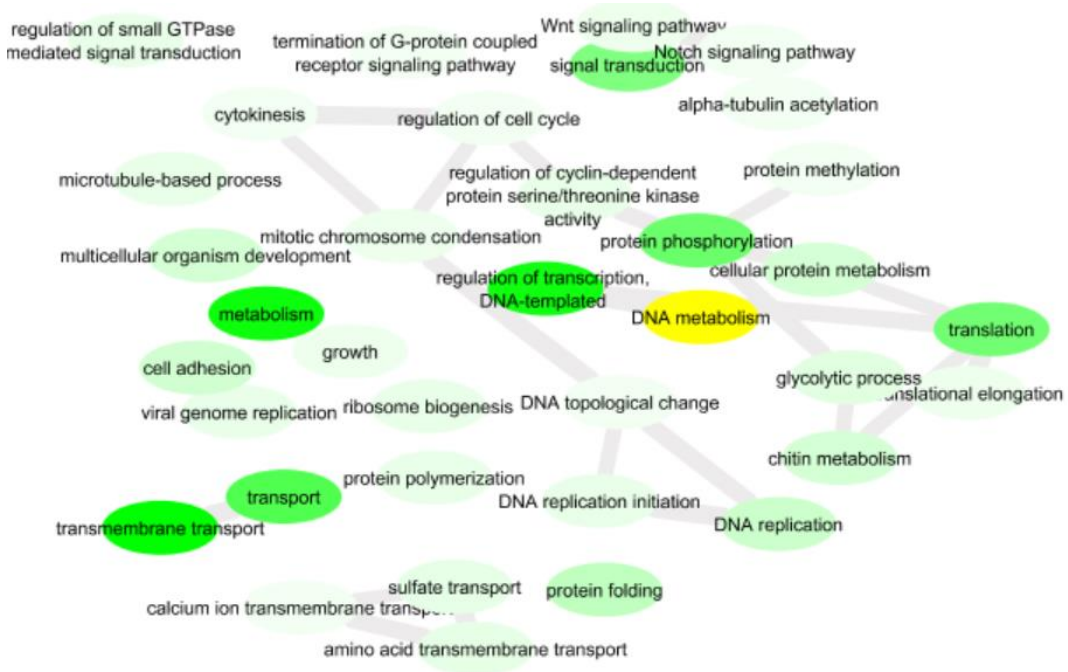


Figure 5 : Graphe des différents processus biologiques dont les gènes sont les plus différemment exprimés entre les conditions J15 et J23.

3.5. Développement d'une plateforme de visualisation des données RNA-seq

Un autre objectif de ce stage était le développement d'une plateforme de visualisation de ces données RNA-seq pour faciliter la fouille de données. Cette plateforme a été développée à l'aide du *package* Shiny (Chang, et al. 2019) de R développé par RStudio et permettant la

création de pages web dynamiques sur lesquelles il est possible de réaliser toutes les analyses/actions disponibles sous R.

L'objectif de cette plateforme est non seulement la visualisation des données acquises par l'équipe mais aussi de données RNA-seq publiques provenant de la banque de données SRA du NCBI (Leinonen, Sugawara et Shumway 2011). C'est pourquoi un script permettant l'analyse de ces données avec le *pipeline* mis en place a été implémenté (Figure 6). Ce script prend en entrée une liste de banques RNA-seq issues de la base de données SRA (*Short Read Archive*), télécharge les données à l'aide de l'outil fastq-dump (National Center for Biotechnology Information 2011) puis applique le *pipeline* d'analyse jusqu'à la génération des matrices de comptage de lectures par gène.

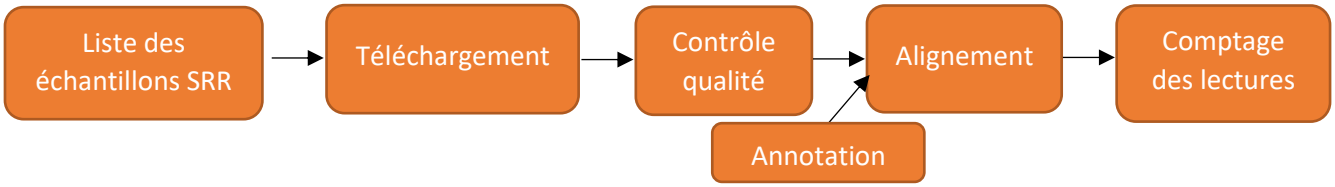


Figure 6 : Détails des différentes étapes du *pipeline* de récupération de données RNA-seq publiques pour le puceron du pois.

A partir de l'ensemble de ces résultats, nous avons développé une interface de visualisation qui se décompose en trois onglets différents.

Le premier est un onglet « information » sur lequel sont regroupées toutes les informations associées aux différents échantillons qu'il est possible de visualiser sur la plateforme (tissu, âge, sexe, stade de développement, souche, …). Une capture d'écran de cet onglet est présentée en Annexe 2.

Le deuxième onglet concerne la visualisation des analyses globales réalisées sur les données. L'utilisateur sélectionne les échantillons sur lesquels il souhaite travailler, puis le nombre de gènes maximum utilisés pour les analyses globales. Par exemple, si l'utilisateur sélectionne 5000 gènes alors les analyses globales se feront sur les 5000 gènes avec la plus grande variance entre les différents échantillons sélectionnés. L'interface propose ensuite d'afficher au choix : une ACP ou une *heatmap* sur les données sélectionnées ainsi qu'une représentation en boîtes à moustache des distributions d'expression de l'ensemble des gènes

pour chaque échantillon sélectionné. Une capture d'écran de cet onglet est présentée en Annexe 3.

Le troisième onglet permet la fouille des données de manière plus précise. Cette page permet de sélectionner le ou les échantillons d'intérêt, puis de sélectionner un ou plusieurs gènes afin d'afficher les valeurs d'expression de ces gènes au sein des échantillons sélectionnés. Cette représentation est disponible sous deux formes : un tableau téléchargeable ainsi qu'un diagramme à bâtons. Une capture d'écran de cet onglet est présentée en Annexe 4.

4. Discussion

La première étape de ce travail de stage consistait à mettre en place un *pipeline* d'analyse de données RNA-seq adapté au jeu de données acquis par l'équipe. L'analyse bibliographique préliminaire a permis de révéler plusieurs solutions performantes pour chacune des étapes de l'analyse. Après avoir testé ces différentes solutions sur notre jeu de données, il s'avère que les résultats peuvent être sensiblement différents d'un outil à l'autre voire même pour un même outil en fonction de son paramétrage. Chaque outil a évidemment ses forces et ses faiblesses mais il semblerait que certains outils aient fait le choix de proposer un paramétrage par défaut permettant de les appliquer avec succès à différents jeux de données alors que d'autres outils doivent être complètement re-paramétrés pour être performants sur chaque jeu de données. Dans notre étude, afin de mettre en place un *pipeline* d'analyse applicable aux futurs jeux de données de l'équipe, nous avons fait le choix de travailler avec les outils pour lesquels les paramètres par défaut sont les plus performants. Une évaluation complète et précise des différents outils et de chacun de leurs paramètres auraient nécessité l'utilisation de données simulées à l'image des travaux de Baruzzo et collaborateurs (Baruzzo, et al. 2017).

Le *pipeline* d'analyse a ensuite pu être appliqué à l'ensemble du jeu de données disponible au sein de l'équipe pour l'étude de la dégénérescence bactériocytaire chez le puceron du pois. Les analyses globales nous ont permis d'identifier un problème majeur dans ce jeu de données. En effet, il semble y avoir une variabilité importante dans les échantillons prélevés aux jours 9 et 15 alors que les trois réplicats des échantillons prélevés au jour 23 semblent relativement proches entre eux. D'un point de vue biologique, cela pourrait être dû au fait qu'à J9 et à J15 nos bactériocytes seraient très hétérogènes d'un point de vue morphologique, certains bactériocytes pourraient ne pas avoir encore entamé le processus de mort cellulaire alors que d'autres seraient déjà dans un stade avancé. Pour le moment, nous avons donc décidé d'enlever les échantillons provenant de la condition J9 ainsi que l'échantillon J15.1 qui semble

lui aussi présenter de grandes différences avec les deux autres réplicats prélevés au jour 15. Ce choix impacte fortement l'analyse puisqu'il n'est plus possible d'identifier les processus moléculaires mis en jeu au début de la mort cellulaire. De plus, le nombre de réplicats du stade intermédiaire (J15) est réduit à seulement deux réplicats ce qui diminue par conséquent la robustesse de nos analyses d'expression différentielle. Néanmoins, cette analyse nous a permis de mettre en évidence 1610 gènes différentiellement exprimés entre les conditions J15 et J23. L'analyse d'enrichissement des annotations de ces gènes différentiellement exprimés a révélé que les processus biologiques dans lesquels sont impliqués ces gènes sont : le transport transmembranaire, la régulation de la transcription et la phosphorylation des protéines.

5. Conclusion

Ces travaux avaient pour objectif de déterminer les processus moléculaires impliqués dans la mort cellulaire des cellules symbiotiques chez le puceron du pois. Ainsi, l'équipe disposait de banques RNA-seq obtenues à partir de bactériocytes prélevés sur des pucerons adultes âgés de 9, 15 et 23 jours correspondant respectivement aux stades précoces, intermédiaires et avancés de la dégénérescence bactériocytaire.

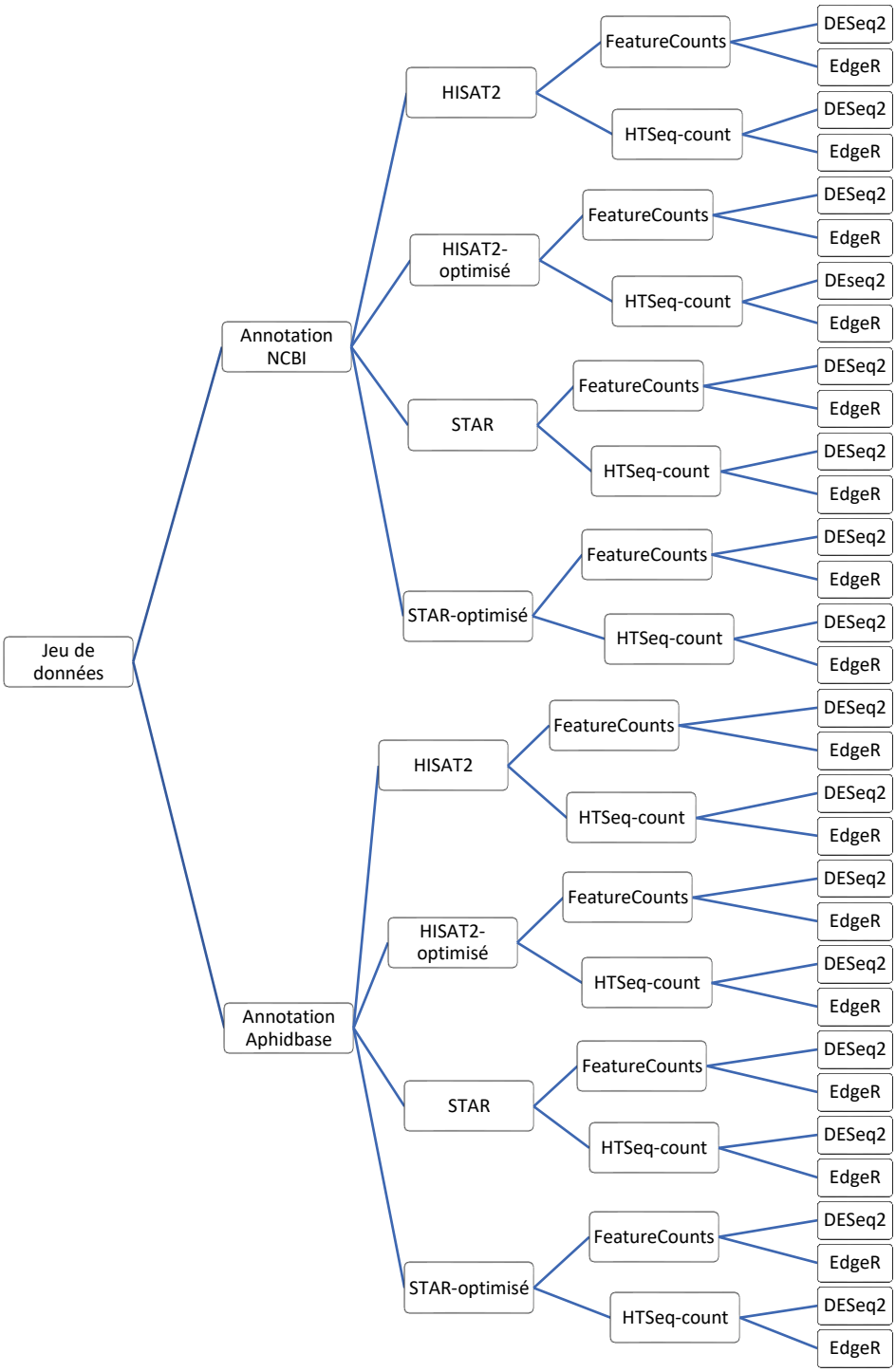
Après une analyse bibliographique des solutions disponibles, j'ai pu tester et sélectionner les outils les plus adaptés et les plus performants sur notre jeu de données. Une fois le *pipeline* d'analyse mis au point, j'ai pu l'appliquer aux différentes banques RNA-seq afin de mettre en évidence les gènes différentiellement régulés au cours du processus de mort cellulaire. En parallèle à ce travail, j'ai également développé une plateforme de visualisation et de fouille de ces résultats.

Les résultats ont permis de mettre en évidence une variabilité importante dans les échantillons correspondant aux stades précoces et intermédiaires du processus de mort cellulaire. Cette variabilité est vraisemblablement liée à une hétérogénéité importante des populations de bactériocytes au sein des pucerons adultes âgés de 9 et 15 jours. Au sein même d'un seul puceron, le processus de mort cellulaire ne semble pas synchronisé entre toutes les cellules bactériocytaires. Ainsi, afin de s'affranchir de cette hétérogénéité, il serait intéressant de pouvoir sélectionner les bactériocytes en fonction de leur stade d'avancement dans le processus de dégénérescence bactériocytaire et non plus en fonction de l'âge du puceron. Depuis l'acquisition du premier jeu de données, le laboratoire s'est d'ailleurs équipé d'une nouvelle loupe binoculaire permettant de sélectionner les différentes populations de bactériocytes.

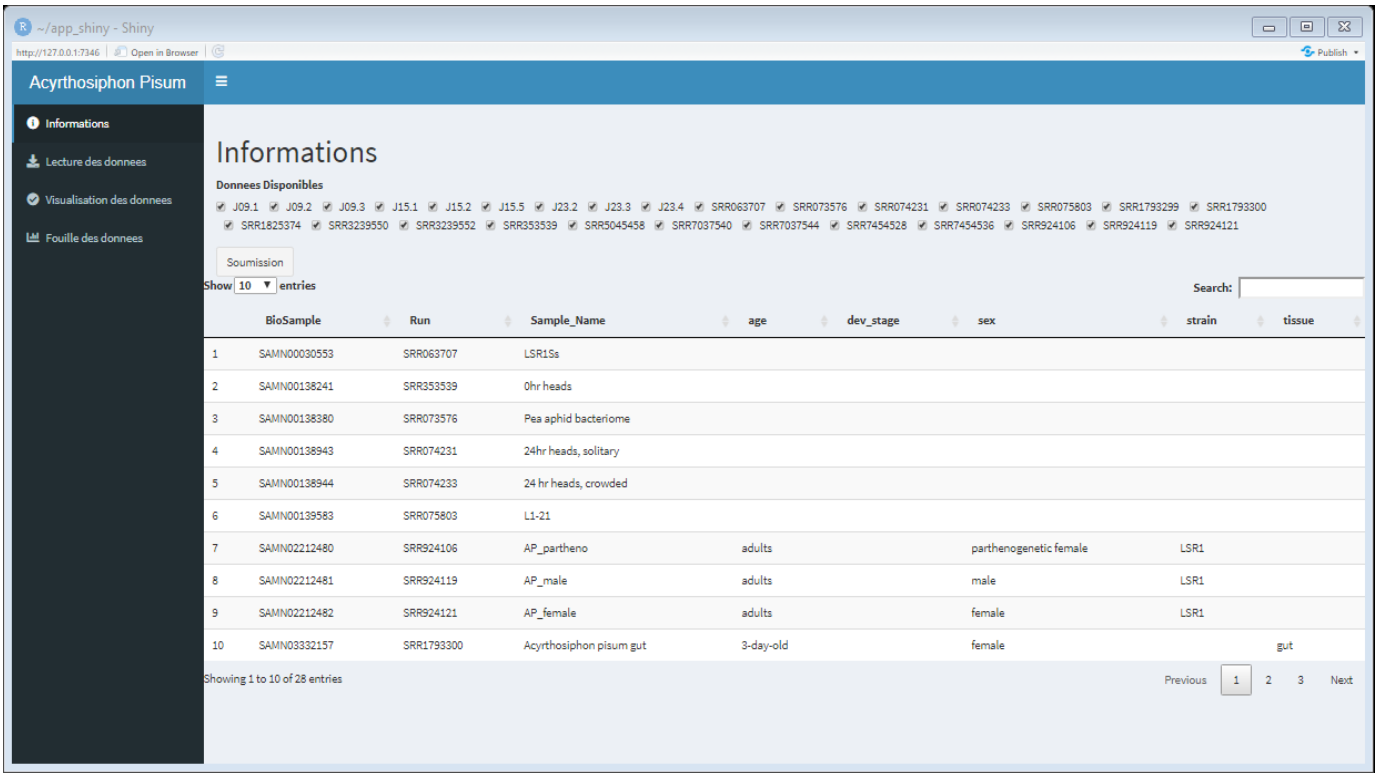
Les perspectives de ce travail s'orientent donc vers l'acquisition de nouvelles données pour mieux caractériser les étapes initiales du processus de mort cellulaire. Par ailleurs, l'analyse d'enrichissement des annotations des gènes différentiellement exprimés entre les conditions J15 et J23 couplée à l'utilisation de la plateforme de visualisation permettra de mettre en évidence les gènes clés impliqués dans cette nouvelle forme de mort cellulaire chez le puceron du pois.

Annexes

Annexe 1 : Schéma des différentes combinaisons d'outils testées pour l'analyse des données RNA-seq.



Annexe 2 : Capture d'écran de l'onglet « Informations » de l'interface de visualisation.



Annexe 3 : Capture d'écran de l'onglet « Visualisation » des données de l'interface.

Annexe 4 : Capture d'écran de l'onglet « Fouille de données » de l'interface.


~/app_shiny - Shiny
Acyrthosiphon Pisum
Informations
Lecture des donnees
Visualisation des donnees
Fouille des donnees
Fouille des donnees
Choisissez les conditions que vous souhaitez afficher
J09.1 J09.2 J09.3 J15.1 J15.2 J15.5 J23.2 J23.3 J23.4 SRR063707 SRR073576 SRR074231 SRR074233 SRR075803 SRR1793299 SRR1793300 SRR1825374 SRR3239550 SRR3239552 SRR353539 SRR5045458 SRR7037540 SRR7037544 SRR7454528 SRR7454536 SRR924106 SRR924119 SRR924121
Choisissez un mode de normalisation
Brut
TPM
RPKM
Brut
Entrez un identifiants de gene a afficher
ACYPI000030-RA ACYPI000007-RA
Download
Show 10 entries
Search:
V1 J09.1 J09.2 J09.3 J15.1 J15.2 J15.5 J23.2 J23.3 J23.4 SRR063707 SRR073576 SRR074231 SRR074233 SRR075803 SRR1793299 SRR1793300 SRR1825374
ACYPI000030-RA ACYPI000030-RA 131575 133795 111019 76832 116290 105765 93861 62913 92789 69405 22672 7729 6922 307 77911 94143 2658
ACYPI000007-RA ACYPI000007-RA 28256 21656 19973 17277 30115 22877 13554 18269 21904 38319 8797 3661 4404 88 25027 168326 192
Showing 1 to 2 of 2 entries
Previous 1 Next
ACYPI000030-RA
ACYPI000007-RA
Count
0 50000 100000 150000 200000
J09.1 J09.2 J09.3 J15.1 J15.2 J15.5 J23.2 J23.3 J23.4 SRR063707 SRR074231 SRR075803 SRR1793300 SRR3239550 SRR353539 SRR7037540 SRR7454528 SRR924106 SRR924121
Sample

Références

Anders, S, PT Pyl, et W Huber. «HTSeq - A Python framework to work with high-throughput sequencing data.» *Bioinformatics*, 2014.

Ashburner, M, et al. «Gene ontology : tool for the unification of biology. The Gene Ontology Consortium.» *Nature Genetics*, 2000.

Baruzzo, G, KE Hayer, EJ Kim, B Di Camillo, GA FitzGerald, et GR Grant. «Simulation-based comprehensive benchmarking of RNA-seq aligners.» *Nature Methods*, 2017.

Bolger, AM, M Lohse, et B Usadel. «Trimmomatic: a flexible trimmer for Illumina sequence data.» *Bioinformatics*, 2014.

Buchner, Paul. *Endosymbiosis of animals with plant microorganisms.* New York: Interscience, 1965.

Chandramohan, R., PY. Wu, JH. Phan, et MD. Wang. «Benchmarking RNA-Seq quantification tools.» *Med Biol Soc*, 2013.

Chang, Winston, Joe Cheng, JJ Allaire, Yihui Xie, et Jonathan Mc Pherson. «shiny : Web Application Framework for R.» 2019.

Cock, Pj, CJ Fields, N Goto, ML Heuer, et PM Rice. «The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants.» *Nucleic Acid Res*, 2010.

Dobin, A., Davies CA, Schlesinger F, et et al. «STAR: ultrafast universel RNA-seq aligner.» *Bioinformatics*, 2013.

Germain, Pierre-Luc, Alessandro Vitriolo, Antonio Adamo, Pasquale Laise, Vivek Das, et Giuseppe Testa. «RNAontheBENCH: computational and empirical resources for benchmarking RNAseq quantifiaction and differential expression methods.» *Nucleic Acids Research*, 2016.

Grant, Gregory R., et al. «Comparative analysis of RNA-Seq alignement algorithms and the RNA-Seq unified mapper (RUM).» *Bioinformatics*, 2011.

Gu, Quan. «featureCounts or htseq-count?» *Bioinformatics I/O.* 15 Février 2016. http://bioinformatics.cvr.ac.uk/blog/featurecounts-or-htseq-count/ (accès le Juin 4, 2019).

Gündüz, E. Akman, et AE. Douglas. «Symbiotic bacteria enable insect to use a nutritionally inadequate diet.» *Proceeding Biological sciences*, 2009.

Hansen, AK, et NA Moran. «Aphid genome expression reveals host-symbiont cooperation in the production of amino acids.» *Proc Natl Acad Sci U S A*, 2011.

Heddi, A Vallier, C Anselme, H Xin, Y Rahbé, et F Wackers. «Molecular and cellular profiles of insect bacteriocytes : mutualism and harm at the initial evolutionary step of symbiogenesis.» *Cell Microbiology*, 2005.

IAGC, The International Aphid Genomics Consortium. «Genome sequence of the pea aphid Acyrthosiphon pisum.» *PLoS Biology*, 2010.

Isaac A. Babarinde, Yuhao Li, Andrew P. Hutchins. «Computational Methods for Mapping Assembly and Quantification for Coding and Non-coding Transcripts.» *Computational and Structural Biotechnology Journal*, 2019.

Kim, Daehwan, Ben Langmead, et Steven L Salzberg. «HISAT: a fast spliced aligner with low memory requirements .» *Nature methods*, 2015.

Leinonen, Rasko, Hideaki Sugawara, et Martin Shumway. «The Sequence Read Archive.» *Nucleic Acids Research*, 2011.

Liao, GK Smyth, et W Shi. «featureCounts: an efficient general purpose program for assigning sequence reads to genomic features.» *Bioinformatics*, 2014.

Love, Michael I, Wolfgang Huber, et Simon Anders. «Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.» *Genome Biology*, 2014.

Matsuura, Y, Y Kikuchi, T Miura, et T Fukatsu. «Ultrabithorax is essential for bacteriocyte development.» *Proc Natl Acad Sci U S A*, 2015.

Moran, N.A, et A. Telang. «Bacteriocyte-asociated symbionts of insects.» *BioScience*, 1998.

National Center for Biotechnology Information. *SRA Knowledge Base.* 2011.

Pär G Engström, Tamara Steijger, et al. «Systematic evaluation of spliced alignment programs for RNA-seq data.» *Nature Methods*, 2013.

R Development Core Team. *R : A Language and Environment or Statistical Computing.* Vienne, 2005.

Robinson MD, McCarthy DJ, Smyth GK. «edgeR : a Bioconductor package for differential expression analysis of digital gene expression data.» *Bioinformatics*, 2010.

Shannon, Paul, et al. «Cytoscape : A software Environment for Integrated Models of Biomolecular Interaction Networks.» *Genome Research*, 2003.

Shigenobu, S, H Wanatabe, M Mattori, Y Sakaki, et H Ishikawa. «Genome sequence of the endocellular bacterial symbiont of aphids Buchnera sp.» *Nature*, 2000.

Simonet, P., et al. «Direct flow cytometry measurements reveal a finetuning of symbiotic cell dynamics according to the host developmental needs in aphid symbiosis.» *Sci Rep*, 2016.

Simonet, Pierre, et al. «Bacteriocyte cell death in the pea aphid/Buchnera symbiotic system.» *PNAS*, Février 2018.

Supek, F., M. Bosnjak, N. Skunca, et T. Smuc. «REVIGO summarizes and visualizes long lists of Gene Ontology terms.» *PLoS ONE*, 2011.

von Dohlen, CD, et NA Moran. «Molecular data support a rapid radiation of aphids in the Cretaceous and multiple origins of host alternation.» *Biological Journal of the Linnean Society*, 2000.

Wilson, A.C, et R.P Duncan. «Signature of host/symbiont genome coevolution in insect nutritional endosymbioses.» *Proc Natl Acad Sci U S A*, 2015.

Zytnicki, M. «mmquant: how to count multi-mapping reads ?» *Bioinformatics*, 2017.

Sitographie

http://www.spe.inra.fr/Toutes-les-actualites/association-bacterie-puceron

https://htseq.readthedocs.io/en/release_0.11.1/count.html

http://bioinf.wehi.edu.au/featureCounts/

https://samtools.github.io/hts-specs/SAMv1.pdf

http://bioinformatics.cvr.ac.uk/blog/featurecounts-or-htseq-count/

http://revigo.irb.hr/revigo.jsp – fragment-2a