Année 2019-2020

Mémoire de Master 2

Analyse de données transcriptomiques pour la caractérisation moléculaire des mécanismes régulant l'homéostasie des cellules symbiotiques chez le puceron du pois

Gaëtan Rongier

Encadrants: Nicolas Parisot et Federica Calevro

Laboratoire d'accueil : UMR 203 INRAE/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 la bactérie *Buchnera aphidicola.* Ces symbiotes vivent à l'intérieur de cellules spécialisées de l'hôte, les bactériocytes, pour lesquelles une étude récente a pu mettre en évidence un processus inédit de mort cellulaire qui se déclenche progressivement dès le stade adulte. L'étude a réussie à caractériser la cascade d'évènements cellulaires débutant par une hypervacuolisation provenant du réticulum endoplasmique qui aboutit à la mort des cellules. L'objectif de ce stage était la caractérisation moléculaire de cette cascade. Pour cela, j'ai réalisé une analyse de données transcriptomique à haut-débit (RNA-seq), complétée par une analyse d'enrichissement des annotations fonctionnelle des gènes différentiellement exprimés. Ces analyses ont été réalisées sur des bactériocytes de pucerons aux stades nymphaux N2, N3 et N4 ainsi qu'au stade adulte (9, 15 et 23 jours), pour pouvoir comparer l'expression de leurs gènes entre la phase de croissance du nombre de bactériocytes et de symbiotes et la phase de dégénérescence. Ces analyses ont permis d'identifier les bases moléculaires de ces phases de croissance et de dégénérescence avec notamment un grand nombre de gènes différentiellement exprimés associés à la régulation du cycle cellulaire. De plus on observe que la majorité de ces changements transcriptionnels se déroulent entre les stades N4, A9 et A15. Pour terminer on a pu mettre en évidence une régulation transcriptionnelle importante de l'apoptose, de l'autophagie ou encore de l'activité lysosomale dans la mise en place du processus de mort cellulaire unique identifié chez le puceron du pois.

Remerciements

Tout d’abord je tiens à remercier mes encadrants de stage, Mme Federica Calevro, directrice adjointe du laboratoire BF2i et responsable de l’équipe "SymT", et Nicolas Parisot enseignant-chercheur au laboratoire, pour leur accueil et le temps consacré pour m’aider, répondre à mes questions et la rédaction du rapport.

Je tiens également à remercier Mme Mélanie Ribeiro Lopes doctorante au laboratoire, pour son aide dans l’interprétation des résultats.

Je souhaite remercier spécialement les deux membres du bureau, Carole et Nicolas pour leur accueil et leur bonne humeur au quotidien.

Et enfin j’aimerais remercier toute l’équipe du BF2I pour leur accueil qui a grandement facilité mon intégration au laboratoire et la continuité des interactions au quotidien malgré des conditions difficiles.

Table des matières

Liste des logiciels

- Python 3.7.3
- R 3.6.2 et ses librairies:
  - DEseq2 1.26.0
  - TopGO 2.38.1
  - Rsubread 2.0.1
- HtseqCount 0.11.2
- Star 2.7.3
- Trimmomatic 0.39
- OrthoFinder 2.3.11

Liste des figures

Liste des tableaux

Liste des abréviations

NCBI = National center for biotechnology information

GO = Gene Ontology

PB = paire de bases

ACP = Analyse en composantes principales

KEGG = Kyoto Encyclopedia of Genes and Genome

1. Introduction

Les insectes représentent 90% des espèces animales connues, colonisant une grande partie des habitats terrestres [1]. Ce succès écologique peut s'expliquer en partie par leur capacité à se développer dans des environnements pourtant nutritionnellement carencés en acides aminés et autres nutriments essentiels comme le sang ou la sève des plantes. Leur survie au sein de ces niches trophiques est rendue possible grâce à leur fréquente association avec des microorganismes qui complémentent leur alimentation. Ces associations durables entre plusieurs organismes différents sont appelées symbioses [1]. Les termites champignonnistes ont par exemple établi une association symbiotique avec un champignon qui pré-digère des végétaux pour les rendre assimilables pour les termites. De façon plus intégrée, d'autres espèces d'insectes utilisent les capacités métaboliques de microorganismes internalisés dans leur tube digestif, voire à l'intérieur de cellules spécialisées, pour leur permettre de pallier leurs carences nutritionnelles. Certaines de ces associations symbiotiques sont devenues obligatoires pour la survie de l'hôte et du symbiote [1]. Des études ont notamment révélé que chez certains insectes, la symbiose est devenue indispensable pour leur développement embryonnaire ou encore leur immunité innée [2].

Au sein des insectes, l'ordre des hémiptères est un exemple de succès écologique avec plus de 100 000 espèces différentes malgré un régime alimentaire composé exclusivement de sève végétale [3]. C'est la fusion des capacités métaboliques de l'insecte hôte et de son ou ses symbiotes qui permet de complémenter son alimentation riche en sucres mais extrêmement pauvre en acides aminés essentiels et vitamines. Dans ces symbioses, le symbiote synthétise les acides aminés essentiels à l'hôte à partir des matières nutritives simples comme les sucres fournis par l'insecte hôte [4]. En retour, l'insecte lui assure une niche écologique souvent peu compétitive à l'intérieur de cellules spécialisées de l'hôte nommées bactériocytes [1].

Le puceron du pois (*Acyrthosiphon pisum*) est un de ces hémiptères se nourrissant exclusivement de sève élaborée. Il vit en association avec la bactérie *Buchnera aphidicola,* hébergée dans des bactériocytes, regroupés en *clusters* tapissant l'abdomen et entourant l'intestin des pucerons [4]. Les bactériocytes du puceron sont des cellules géantes, facilement isolables les rendant particulièrement adaptées aux études scientifiques. De plus, les génomes des deux partenaires sont disponibles rendant possible l'étude moléculaire de leurs interactions [5]. C'est pour cette raison que le puceron du pois est devenu le modèle par excellence en génomique des interactions.

Au cours d'une récente étude, mon laboratoire d'accueil a pu caractériser la dynamique du nombre de bactériocytes et des bactéries qu'ils hébergent tout au long de la vie du puceron. Cette étude a montré une coordination entre cette dynamique et les besoins physiologiques de l'hôte [6]. Au cours du développement nymphal du puceron, on observe une augmentation exponentielle du nombre de bactéries endosymbiotiques en parallèle à celle des bactériocytes. Une fois à l'âge adulte, la dynamique des cellules symbiotiques s'inverse. On observe une diminution du nombre de bactériocytes et de symbiotes en parallèle au vieillissement du puceron. Cette dynamique semble corrélée aux besoins métaboliques fluctuant en fonction de la période du cycle de vie, permettant d'atteindre un équilibre entre le coût énergétique et les avantages du symbiote pour l'hôte [7].

Cette étude a montré que chez les pucerons adultes, les bactériocytes subissent une dégénérescence progressive associée à d'important changements morphologiques. Ces changements se manifestent notamment par une hypervacuolisation progressive entraînant la mort des cellules [7]. De par ses caractéristiques, ce processus de mort cellulaire des bactériocytes du puceron est distinct de ceux connus jusqu'à présent.

Bien que ce processus de mort cellulaire unique soit maintenant bien caractérisé sur le plan cellulaire, les bases moléculaires sous-jacentes sont encore à découvrir. Mon travail de master a donc consisté en la caractérisation moléculaire, via l'analyse de données de séquençage à haut-débit du transcriptome (RNA-seq), du processus de mort cellulaire des bactériocytes du puceron du pois.

En analysant le transcriptome des bactériocytes issus de pucerons aux stades nymphaux N2, N3 et N4 ainsi qu'au stade adulte (9, 15 et 23 jours), nous avons identifié les gènes différentiellement exprimés au cours du processus de mort cellulaire. Ces analyses ont permis de mettre en évidence que la majorité des régulations transcriptionnelles ont lieu aux stades N4, A9 et A15 et concernent les processus de l'apoptose et de l'autophagie ainsi que les voies de signalisation qui les régulent.

2.Matériel et Méthodes

2.1 Données génomiques

Le génome d'*Acyrthosiphon pisum* a été initialement publié en 2010 [8] et a récemment été ré-assemblé (assemblage pea_aphid_22Mar2018_4r6ur; identifiant NCBI : GCF_005508785.1). Ce nouvel assemblage est composé de 21 920 scaffolds pour une taille totale de 541 Mb. Il a été annoté par le pipeline d'annotation des génomes eucaryotes du NCBI (NCBI *Acyrthosiphon pisum* Annotation Release 103) et a permis d'identifier 20 593 gènes.

2.2 Données transcriptomiques

Les données de transcriptomique à haut-débit (séquençage RNA-seq) proviennent de tissus bactériocytaires prélevés, en 5 réplicats, sur des pucerons de stades nymphaux N2, N3 et N4 ainsi que des pucerons adultes âgés de 9, 15 et 23 jours. Pour chaque échantillon, 200 bactériocytes ont été disséqués à partir de 5 individus. Après extraction et purification des ARN messagers, les échantillons ont été séquencés sur la plateforme de séquençage de Leuven (Belgique) en utilisant la technologie Illumina (HiSeq 4000) produisant en moyenne 50 601 451 de lectures de 50 pb en single-end par échantillon. Trente fichiers FASTQ ont ainsi été obtenus (tableau 1) pour un total de 1 518 043 537 de lectures. Le format FASTQ est un format de fichier texte permettant le stockage des lectures séquencées ainsi que des scores de qualité de prédiction du séquençage.

2.3 Environnement informatique

Les analyses bioinformatiques de ce stage ont été réalisées sur un serveur de calculs géré par le laboratoire BF2i. Il s'agit d'un serveur Dell R730 sous Debian 8 avec 14 CPUs et 120 Go de RAM.

2.4 Contrôle qualité des données

Le contrôle qualité des données de séquençage consiste en la suppression des régions où la qualité de prédiction des nucléotides est insuffisante pour réaliser une analyse robuste des résultats. Cette étape a été réalisée pour les 30 échantillons disponibles en utilisant la version 0.39 du logiciel Trimmomatic [9], outil rapide et parallélisable conçu pour nettoyer les jeux de données Illumina, avec les options suivantes :

SE/PE : indique si les données sont single-end (SE) ou paired-end (PE). Dans notre cas elles sont SE

-phred33 : Indique le format des scores de qualité de prédiction

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 de ses scores de qualité est inférieure à 25

SLIDINGWINDOW:4:15 : Parcourt la lecture à partir de l'extrémité 5' selon une fenêtre glissante de 4 nucléotides et tronque la lecture à l'extrémité 3' si le score de qualité moyen de la fenêtre est inférieur à 15.

MINLEN:36 : supprime la lecture si elle a une taille inférieure à 36 nucléotides

2.5 Positionnement des lectures sur le génome de référence

Une fois le contrôle qualité effectué la première étape de l'analyse consiste à positionner les lectures sur le génome de référence. Il faut au préalable indexer le génome de référence pour rendre l'opération réalisable dans un temps raisonnable. L'indexation du

génome et le positionnement des lectures a été réalisé en utilisant la librairie R/Bioconductor Rsubread version 2.0.1 [10].

L'indexation a été réalisé avec la fonction "buildindex" qui consiste à créer une table de hachage du génome, en utilisant les options suivantes :

- *indexSplit = FALSE* : N'autorise pas de partager l'index en plusieurs blocs.

Le mapping a, lui, été réalisé avec la fonction "Align" avec les paramètres suivants :
- *"index_Acyrthosiphon_pisum" :* Chemin vers le répertoire contenant l'index du génome
- *readfile1=readfile1 :* Chemin vers les fichiers FASTQ nettoyés
- *type = "rna" :* Les lectures correspondent au séquençage de molécules d'ARN
- *output_file = paste :* Chemin vers le fichier de sortie
- *useAnnotation=TRUE :* On utilise un fichier d'annotation des gènes
- *annot.ext :* Chemin vers le fichier d'annotation des gènes
- *isGTF=TRUE :* Le fichier d'annotation des gènes est au format GFF
- *GTF.featureType = "gene" :* Dans le fichier d'annotation on prend en compte les gènes
- *GTF.attrType="gene" :* On utilise les identifiants des gene
- unique = FALSE : Autorise à compter les gènes s'il s'aligne sur plusieurs *features* en même temps

2.6 Quantification de l'expression des gènes :

Le package Rsubread [10] utilisé pour l'alignement comprend également une fonction de quantification de l'expression des gènes par comptage des lectures alignées, appelée "featureCounts". La fonction compte le nombre de lectures issues de fichiers d'alignement au format SAM/BAM qui s'alignent sur un ensemble spécifié de caractéristiques génomiques (gène, transcrit ou exon). L'annotation utilisée dans cette étude est celle réalisée par le pipeline d'annotation des génomes eucaryotes du NCBI (NCBI *Acyrthosiphon pisum* Annotation Release 103). Cette fonction a été utilisée avec les paramètres suivants :

- *files :* Chemin vers les fichiers de mapping

- *countMultiMappingReads=TRUE :* Les lectures s'alignant à plusieurs positions sont comptées

- *useAnnotation=TRUE :* On utilise un fichier d'annotation des gènes

- *annot.ext :* Chemin vers le fichier d'annotation des gènes

- *isGTF=TRUE :* Le fichier d'annotation est au format GFF

- *GTF.featureType = "gene" :* Dans le fichier d'annotation on prend en compte les gènes

- *GTF.attrType="gene" :* On utilise les identifiants des gènes

- *allowMultiOverlap= TRUE :* On autorise le comptage des lectures qui s'alignent sur des *features* différents (ici sur plusieurs gène)

- *useMetaFeatures=TRUE :* Résume les comptages en utilisant GTF.attrType

2.7 Analyse d'expression différentielle :

A partir des tables de comptage du nombre de lectures par gène, la version 1.26.0 de la librairie R/Bioconductor DESeq2 a été utilisée [11] afin d'identifier les gènes différentiellement exprimés entre les différentes conditions analysées. La librairie DESeq2 permet dans un premier temps de normaliser les données de comptage entre les différents échantillons. Les données normalisées sont ensuite modélisées grâce à un modèle binomial négatif généralisé. Les *p-values* sont ensuite ajustées selon la procédure de Benjamini et Hochberg. Les gènes avec une *p-value* ajustée en dessous 0.01 sont considérés comme différentiellement exprimés.

2.8 Analyse d'enrichissement des annotations fonctionnelles :

Afin de faciliter l'analyse des listes de gènes différentiellement exprimés, il est possible d'identifier si certaines annotations fonctionnelles sont significativement enrichies dans ces listes. Pour cela, nous avons utilisé les annotations fonctionnelles des gènes du puceron du pois obtenues à partir de deux bases de données : Gene Ontology (GO) [12, 13] et Kyoto Encyclopedia of Genes and Genomes (KEGG) [14].

La significativité de l'enrichissement des annotations Gene Ontology a été évaluée en utilisant la version 2.38.1 de la librairie R/Bioconductor topGO [15] avec un seuil de p-value de 0.05. On a choisi parmi les algorithmes disponibles d'utiliser "elim". Il consiste à commencer par évaluer les termes GO les plus précis. On progresse alors vers les termes les moins précis en supprimant à chaque étape les gènes déjà évalués précédemment. A chaque évaluation, on applique le test statistique de Fisher afin de déterminer les GOs significativement enrichis.

La base de données KEGG [14] relie les informations génomiques aux voies métaboliques et voies de signalisation connues. Elle est séparée en deux bases : GENES (catalogue de gènes de tous les génomes séquencés) et PATHWAY (qui contient des représentations graphiques des processus biologiques tels que les voies de signalisation et métaboliques). La base PATHWAY est complétée par les relations d'orthologie entre les gènes présents dans la base de données. A partir du fichier de correspondance entre les protéines de notre génome de référence et les identifiants KEGG nous avons pu reporter les gènes différentiellement exprimés sur les représentations graphiques des voies métaboliques et de signalisation correspondant.

3. Résultats

Afin de caractériser les bases moléculaires du processus de mort cellulaire des bactériocytes du puceron du pois, l'équipe d'accueil a entrepris le séquençage à haut-débit (RNA-seq) du transcriptome des bactériocytes de pucerons aux stades nymphaux N2, N3 et N4 ainsi qu'au stade adulte (9, 15 et 23 jours). Ces stades correspondent à des étapes clés de la vie des bactériocytes. En effet, l'équipe d'accueil a précédemment montré que le nombre et la taille des bactériocytes augmentent depuis le stade embryonnaire jusqu'au jour 13 du stade adulte [8] parallèlement à la croissance du nombre de bactéries symbiotiques hébergées chez cet insecte. A partir du jour 13, on observe en revanche une réduction significative du nombre et de la taille des bactériocytes ainsi que du nombre de bactéries symbiotiques. Cette dégénérescence s'accompagne également d'importantes modifications morphologiques des bactériocytes avec l'apparition de zones hypervacuolisées à faible densité de symbiotes (Figure 1).

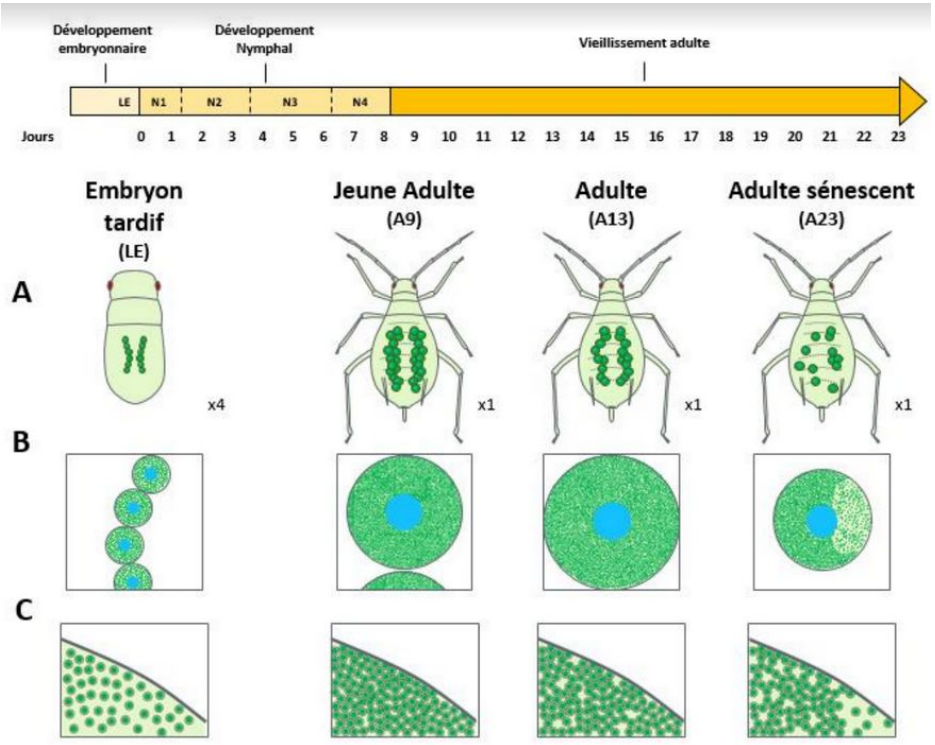

Figure 1 : Modèle schématique de la dynamique de *B. aphidicola* et des bactériocytes lors du développement parthénogénétique du puceron *A. pisum*. Evolution du nombre et de l'organisation des bactériocytes (A), de la taille des bactériocytes (B) ou de la densité des symbiotes (C) dans le corps du puceron, de l'embryon tardif à l'adulte sénescent. Dans

chaque panneau, l'échelle est constante, exceptée pour le panneau (A) où l'embryon est agrandi d'un facteur 4 par rapport aux pucerons adultes [3].

Ces modifications cellulaires ont été plus finement caractérisées dans une seconde étude [7]. Ces travaux ont notamment permis d'affiner la succession des évènements cellulaires avec une initiation du processus de mort cellulaire chez les adultes à jour 9 (figure 2) .

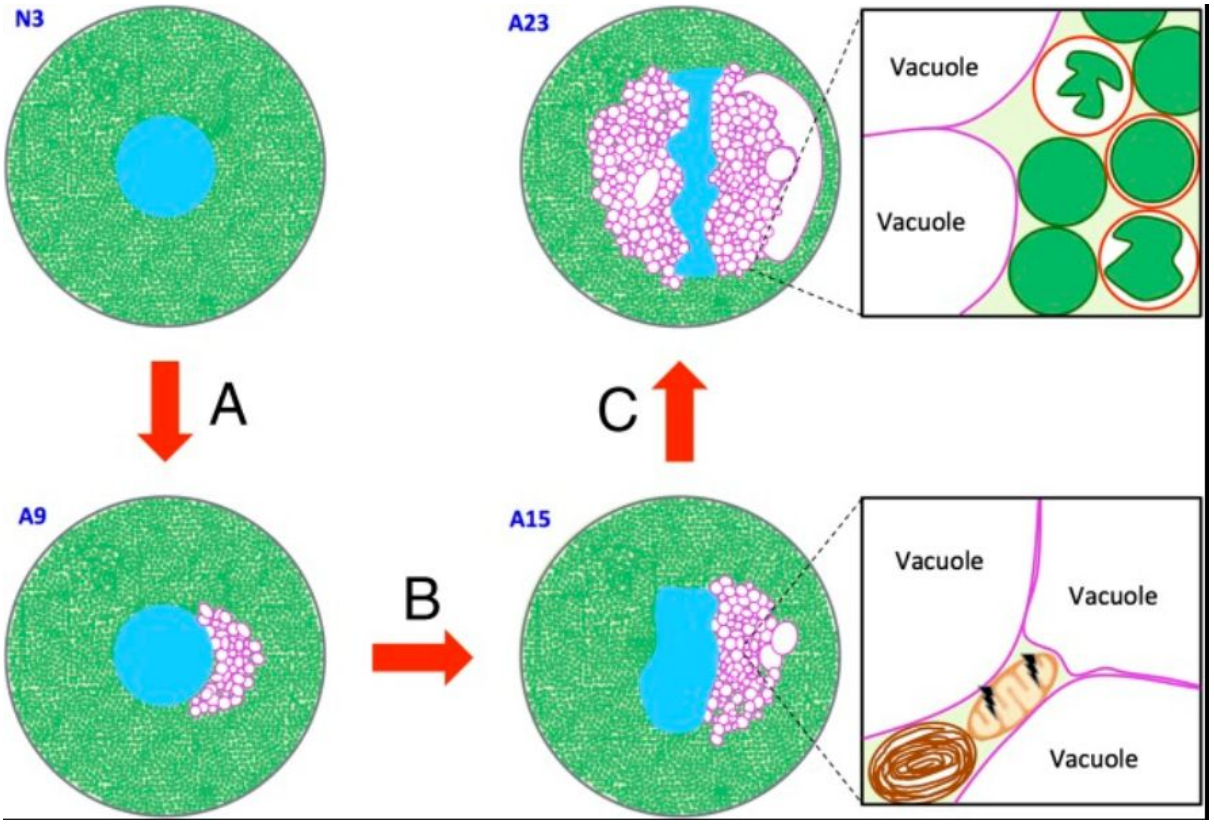

Figure 2 : Représentation schématique des phases principales de la mort cellulaire des bactériocytes du puceron du pois. (A) Phase I : induction de l'hypervacuolisation dérivé du réticulum endoplasmique. (B) Phase II : induction de la réponse au stress (activation de l'autophagie). (C) Phase III : dégradation de Buchnera par le lysosome. Bleu, noyau du bactériocyte; marron, autophagosome; vert foncé, *B. aphidicola*; vert clair, cytoplasme du bactériocyte; magenta, membrane des vacuoles; orange, mitochondrie; rouge, lysosome [9].

Les stades de développement analysées par RNA-seq au cours de ce stage correspondent donc à i) des stades de croissance des bactériocytes (N2, N3 et N4), ii) un

stade précoce (adulte à jour 9, A9), iii) un stade intermédiaire (A15) et enfin iv) un stade avancé (A23) du processus de mort cellulaire.

3.1 Identification des gènes différentiellement exprimés

3.1.1 Contrôle qualité des données

Avant de commencer l'analyse des données de séquençage il est important de vérifier la qualité des séquences afin de supprimer les séquences ou portions de séquences de mauvaise qualité ou non informatives (e.g. adaptateurs de séquençage).

Les 30 fichiers FASTQ (6 stades de développement en 5 réplicats) ont donc été traités avec l'outil Trimmomatic [9] (Tableau 1).

Tableau 1 : Nombre de lectures moyen supprimé lors du contrôle qualité des échantillons.

	Stade de développement					
	N2	**N3**	**N4**	**A9**	**A15**	**A23**
Nombre de lectures initial (Moyenne ± Ecart-Type)	41 563 652 ± 5 703 200	40 455 417 ± 3 678 953	40 359 215 ± 7 201 699	77 863 793 ± 30 534 972	44 373 662 ± 8 668 677	58 992 966 ± 14 858 182
Nombre de lectures supprimées (Moyenne ± Ecart-Type)	412 125 ± 741 155	78 495 ± 7 453	84 467 ± 18 891	165 760 ± 66 950	88 903 ± 15 466	105 624 ± 14 647
Pourcentage de lectures supprimées (%)	0,95	0,19	0,20	0,22	0,20	0,18

Cette étape a permis de montrer la qualité des données de séquençage avec environ 0.2% de lectures supprimées, excepté pour les échantillons N2 qui ont presque 1% de lectures éliminées. Cette différence pour la condition N2 provient en réalité d'un seul réplicat (N2.5) qui montre 3.97% de lectures supprimées alors que les 4 autres échantillons varient autour de 0.2%. Nous n'avons à ce jour pas identifié de facteur particulier pouvant

expliquer cette différence mais cette proportion de lectures supprimées reste tout de même très faible et ne remet pas en cause la qualité de ce réplicat.

3.1.2 Positionnement des lectures sur le génome de référence

Une fois les données de séquençage nettoyées, on va maintenant positionner chacune des lectures sur le génome du puceron du pois. Pour cela on utilise les fonctions “buildindex” et “Align” de la librairie R Rsubread [10]. Les résultats de cette étape, dite de *mapping*, sont présentées dans le tableau 2.

Tableau 2 : Résultats d’alignement des lectures pour les 6 conditions (N2, N3, N4, A9, A15, A23).

	Stade de développement					
	N2	**N3**	**N4**	**A9**	**A15**	**A23**
Nombre de lectures alignées (Moyenne ± Ecart-Type)	38 484 421 ± 5 415 948	35 517 900 ± 3 280 103	34 824 213 ± 5 562 523	64 063 262 ± 2 830 0830	40 103 182 ± 7 556 340	46 542 560 ± 9 116 242
Nombre de lectures alignées à une position unique (Moyenne ± Ecart-Type)	33894181 ± 5 042 865	30 013 086 ± 3 361 452	28 549 738 ± 4 492 032	49 662 226 ± 24 803 451	34 765 917 ± 6 623 251	39 665 016 ± 8 722 032
Nombre de lectures alignées à plusieurs positions (Moyenne ± Ecart-Type)	4 590 239 ± 746 475	5 504 813 ± 869 496	6 274 475 ± 1 709 153	14 401 036 ± 4 040 088	5 337 265 ± 984 479	6 877 543 ± 1 026 610
Nombre de lectures non alignées (Moyenne ± Ecart-Type)	2 997 478 ± 934 930	4 859 022 ± 1 900 248	5 450 534 ± 3 114 215	1 3634 770 ± 3 500 590	4 181 576 ± 1 228 348	12 344 781 ± 5 807 626

Lors de l’étape de *mapping*, en moyenne plus de 85% des lectures ont été alignées sur le génome de référence. Parmi celles-ci environ 73% sont des lectures alignées à une seule position génomique et 13% sont des lectures alignées à plusieurs positions

génomiques. La condition N2 qui a le pourcentage d'alignement le plus élevé est aussi celle qui le pourcentage le plus élevé d'alignement unique. A l'inverse, les échantillons A9 et A23 ont les plus faibles pourcentages d'alignement et également les plus faibles pourcentages d'alignements uniques.

3.1.3 Quantification de l'expression des gènes

Une fois que la position des lectures sur le génome de référence est connue, il est maintenant possible de compter le nombre de lectures qui s'alignent à une position génomique particulière, gène ou transcrit par exemple. Ce comptage représente une estimation de l'expression des gènes, nécessaire pour identifier les gènes différentiellement exprimés entre les différentes conditions expérimentales. Dans notre analyse, l'expression a été quantifiée au niveau des gènes en autorisant le comptage des lectures qui s'alignent à plusieurs positions génomiques. Ce choix a été rendu obligatoire car un certain nombre de gènes du puceron du pois ont subi des duplications au cours de leur histoire évolutive récente. Pour effectuer ce comptage la fonction "featureCounts" de la librairie R Rsubread [13] a été utilisé. Les résultats sont présentés dans le tableau 3.

Tableau 3 : Résultats de quantification de l'expression des gènes pour les 6 conditions (N2, N3, N4, A9, A15, A23).

	Stade de développement					
	N2	**N3**	**N4**	**A9**	**A15**	**A23**
Nombre de lectures comptées (Moyenne ± Ecart-Type)	33 227 300 ± 5 281 424	2 8480 901 ± 3 974 178	26 275 440 ± 4 593 920	43 887 223 ± 24 205 545	33 252 230 ± 6 569 789	37 820 733 ± 9 765 168
Nombre de lectures non comptées car non alignées sur un gène (Moyenne ± Ecart-Type)	5 257 120 ± 1 335 806	7 036 998 ± 1 689 431	8 548 773 ± 2 866 314	20 176 039 ± 4 831 064	6 850 952 ± 1 365 080	8 721 827 ± 1 422 100

Lors de la quantification de l'expression des gènes (tableau 3) environ 68,1% des lectures ont été assignées à au moins un gène et 18,4% ne s'alignent pas sur un gène.

On voit une nouvelle fois que les échantillons N2 ont des résultats différents avec un pourcentage de lectures comptées (80%) plus élevé que les autres conditions. Comme pour le *mapping*, on observe également que la condition A9 a elle aussi des résultats différents avec en moyenne un pourcentage de lectures comptées (53,43%) plus faible que les autres conditions (68,12% en moyenne).

3.1.4 Description globale du jeu de données

Une analyse préliminaire a permis la construction d'une analyse en composantes principales (ACP) (figure 3) et d'une *heatmap* (figure 4), de l'ensemble des données d'expression (comptages) des gènes entre toutes les conditions expérimentales.

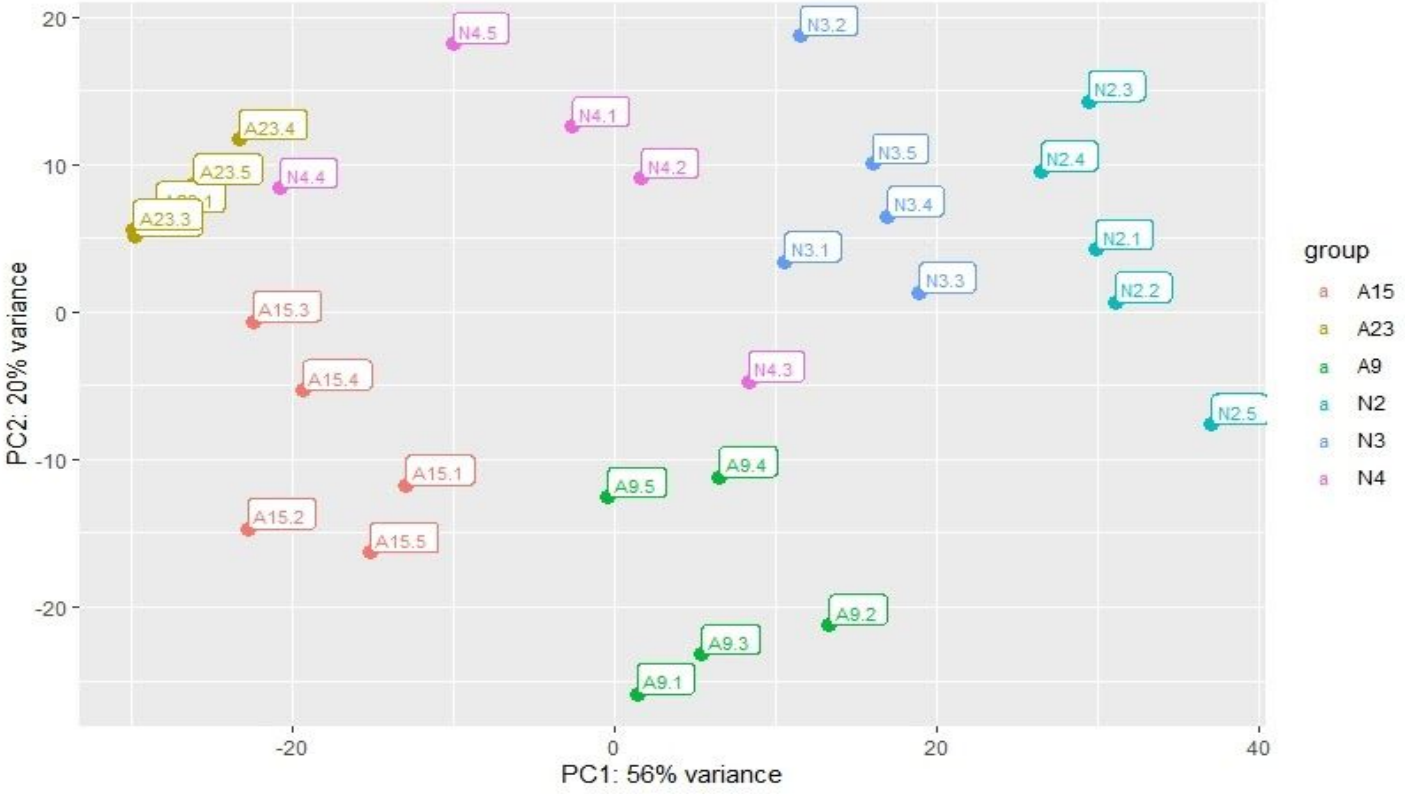

Figure 3 : Analyse en composantes principales des profils d'expressions de tous les échantillons des six conditions (N2, N3, N4, A9, A15, A23).

La première composante principale de cette ACP explique 56% de la variance observée dans le jeu de données alors que la seconde composante principale explique 20% de la variance. De façon intéressante, la première composante semble être fortement corrélée au stade de développement allant graduellement des stades de développement les plus avancés à gauche aux plus précoces à droite.

L'ACP permet également d'observer une faible variation intra-condition dans les deux principales composantes pour la plupart des échantillons à l'exception des échantillons N4. Ces échantillons montrent une variabilité plus importante qui pourrait témoigner d'une hétérogénéité du matériel biologique récolté à ce stade de développement.

La *heatmap* et le *clustering* hiérarchique des échantillons associé (figure 4) confirment également l'importante variabilité au sein de la condition N4, dont les réplicats ne sont pas regroupés. En effet, sur la *heatmap* les échantillons N4 sont mélangés aux échantillons N3 qui ne se regroupent donc pas eux aussi.

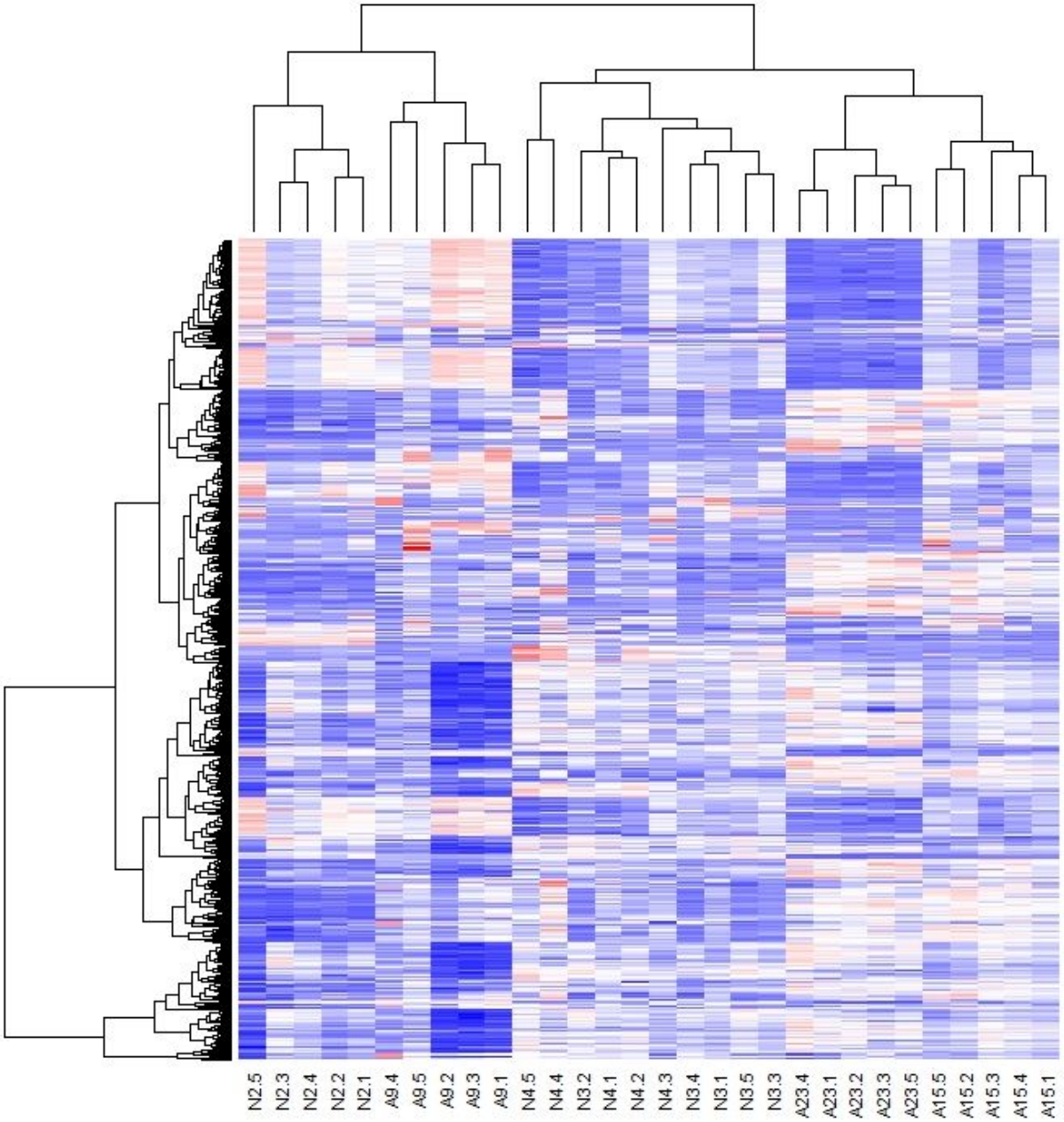

Figure 4 : *Heatmap* des 1000 gènes avec la plus grande variance entre les échantillons.

La condition N3 ne semblant pas très variable sur l'ACP il semble plus probable que ce soient les échantillons N4 qui perturbent le regroupement des échantillons N3. Ces résultats nous indiquent qu'il faudra être prudent quant aux conclusions tirées des futures analyses car elles pourraient être influencées par la variabilité intra-échantillon.

3.1.5 Analyse de l'expression différentielle

A partir des données de comptage, nous avons ensuite pu rechercher les gènes différentiellement exprimés entre les stades de développement successifs des pucerons (tableau 4). Les gènes sont considérés comme différentiellement exprimés s'ils sont en dessous d'un seuil de p-values corrigées de 0.01.

Tableau 4 : Nombre de gènes différentiellement exprimés entre les 5 comparaisons (N3 vs N2, N4 vs N3, A9 vs N4, A15 vs A9, A23 vs A15) avec le nombre de gènes sur- et sous-exprimés selon la première condition de la comparaison.

	Comparaison				
	N3 *vs* N2	**N4 *vs* N3**	**A9 *vs* N4**	**A15 *vs* A9**	**A23 *vs* A15**
Nombre de gènes différentiellement exprimés	318	186	2863	2013	569
- dont sur-exprimés	254	158	1384	839	110
- dont sous-exprimés	64	28	1479	1174	459

On observe alors que la plupart des régulations transcriptionnelles semblent avoir lieu entre les stades N4, A9 et A15. Ces résultats peuvent également être mis en évidence par des représentations graphiques de type VolcanoPlots pour les comparaisons réalisées (Figure 5).

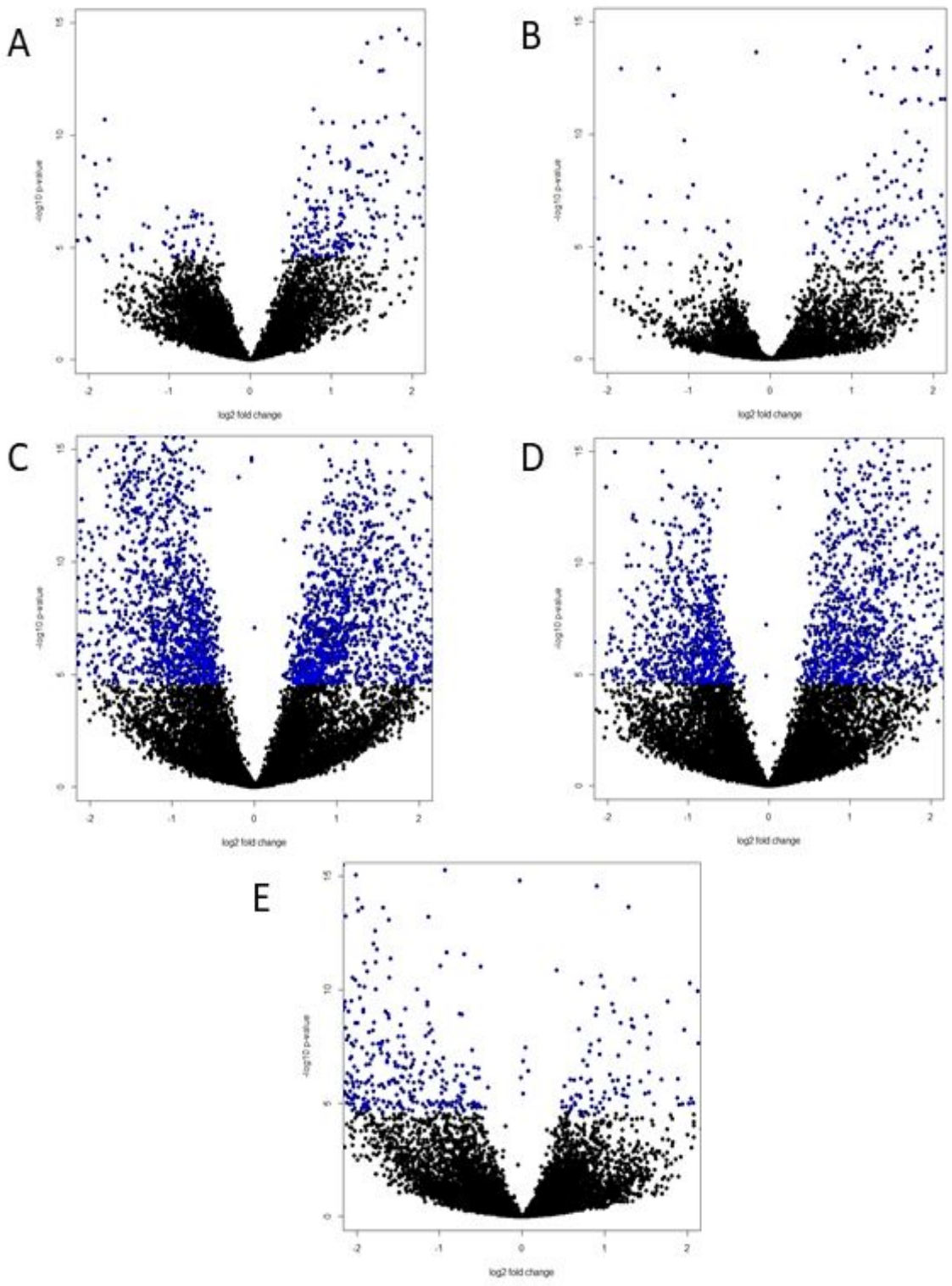

Figure 5 : Représentations Volcano Plots des gènes différentiellement exprimés entre les conditions testés. (A) Comparaison N3 et N2, (B) Comparaison N4 et N3, (C) Comparaison A9 et N4, (D) Comparaison A15 et A9, (E) Comparaison A23 et A15 . Les gènes sont considérés comme différentiellement exprimés si leur *p-value* ajustée est en dessous du seuil de 0.01.

Noir, les gènes non différentiellement exprimés; Bleu, les gènes différentiellement exprimés.

Ces représentations, ainsi que le tableau 4 nous indiquent donc que les régulations sont plutôt associées à une induction transcriptionnelle, avec plus de gènes sur-exprimés que de sous-exprimés, dans les stades nymphaux suivie d'une répression de l'activité transcriptionnelle des gènes concernés dans les stades adultes.

3.2 Caractérisation des bases moléculaires du processus de mort cellulaire

A partir de ces listes de gènes différentiellement exprimés, nous pouvons identifier si plusieurs d'entre eux sont impliqués dans les mêmes processus biologiques. On réalise alors une analyse d'enrichissement des annotations fonctionnelles afin de mettre en évidence les fonctions biologiques sur-représentées. Cette analyse peut-être réalisée sur différents types d'annotations : Gene Ontology (GO) [12, 13] ou KEGG [14].

3.2.1 Analyse d'enrichissement des annotations fonctionnelles Gene Ontology

Les résultats de l'analyse d'enrichissement des termes GO sont présentés dans le tableau 5. Les termes GO peuvent être regroupés en trois catégories : processus biologique (*Biological Process*), fonction moléculaire (*Molecular Function*) et compartiment cellulaire (*Cellular Compartment*).

Tableau 5 : Nombre de termes Gene Ontology significativement enrichis entre les 5 comparaisons (N3 vs N2, N4 vs N3, A9 vs N4, A15 vs A9, A23 vs A15) selon les 3 catégories *Biological Process, Molecular Function, Cellular Component*.

	Gene Ontology		
	Biological Process	**Molecular Function**	**Cellular Component**
N3 vs N2	85	78	37
N4 vs N3	28	41	10
A9 vs N4	162	105	62
A15 vs 9	145	97	35
A23 vs A15	79	42	18

On voit que le nombre de GO significativement enrichis dépend comme attendu du nombre de gènes différentiellement exprimés. Nous allons donc maintenant étudier ces listes d'annotations significativement enrichies afin d'identifier les bases moléculaires du processus de mort cellulaire.

Parmi les annotations significativement enrichies, on retrouve au stade N3 des GOs impliqués dans des processus biologiques liés aux lipides :

- GO:0016042 lipid catabolic process
- GO:0044255 cellular lipid metabolic process

Les gènes associés à ces GOs sont sur-exprimés dans N3 par rapport à N2 et pourraient être impliqués dans la régulation des membranes, ce qui semble cohérent avec les travaux de Simonet *et al.* (2016) [6] qui montrent une croissance et une augmentation du nombre de bactériocytes. La présence de termes GOs liés au cytosquelette pour lesquels les gènes associés sont eux-aussi sur-exprimés dans N3 par rapport à N2 conforte cette hypothèse :

- GO:0008017 microtubule binding

- GO:0005856 cytoskeleton

De manière intéressante, on relève également un terme GO impliqué dans la régulation du processus apoptotique (“GO:0042981 regulation of apoptotic process”) directement lié directement à la mort cellulaire. Les trois gènes différentiellement exprimés associés à cette annotation sont une nouvelle fois des gènes sur-exprimés dans N3 par rapport à N2. L’expression différentielle de tels gènes pourrait témoigner d’une mise en place précoce de certains processus de mort cellulaire dès le stade N3.

La régulation transcriptionnelle des gènes impliqués dans le cytosquelette des bactériocytes se maintient entre les stades N3 et N4 avec la présence des deux mêmes termes GO, associés à des gènes sur-exprimés dans N4 par rapport à N3, parmi la liste des annotations significativement enrichies. On remarque également, à partir du stade N4, la sur-expression de gènes impliqués dans l’autophagie :

- GO:0000045 autophagosome assembly
- GO:1905037 autophagosome organization
- GO:0016236 macroautophagy

Dans la continuité des processus de mort cellulaire initiés au stade N3, ces gènes impliqués dans l’autophagie pourraient notamment participer à l’élimination des symbiotes que l’on peut observer au niveau cellulaire [6, 7].

Au stade A9, les gènes impliqués dans le cytosquelette restent sur-exprimés :

- GO:0000226 microtubule cytoskeleton organization
- GO:0051013 microtubule severing
- GO:0007015 actin filament organization
- GO:0008017 microtubule binding
- GO:0030950 establishment or maintenance of actin cytoskeleton

Ils sont cette fois-ci associés à des gènes impliqués dans la mitose qui se retrouvent sur-exprimés dans A9 par rapport à N4 :

- GO:0005639 integral component of nuclear inner membrane
- GO:0005730 nucleolus
- GO:0005634 nucleus
- GO:0006260 DNA replication
- GO:1902412 regulation of mitotic cytokinesis
- GO:0007064 mitotic sister chromatid cohesion
- GO:0000070 mitotic sister chromatid segregation

La surexpression de ces gènes est cohérente avec la multiplication du nombre de bactériocytes observée au cours du développement du puceron [7]. On observe également que le stade A9 semble être caractérisée par un remodelage transcriptionnel des bactériocytes avec notamment des changements au niveau de la chromatine :

- GO:0006265 DNA topological change
- GO:0006338 chromatin remodeling

Une telle reprogrammation des bactériocytes a déjà pu être observée chez le puceron en cas de stress nutritionnel [25]. Il semblerait donc les bactériocytes du puceron du pois soient des cellules particulièrement plastiques.

Ce remodelage est cependant ponctuel et caractéristique du stade A9 dans notre étude car au stade A15 on observe cette fois une sous-expression des gènes impliqués dans ce processus. En revanche les processus d'élimination des symbiotes sont maintenus avec une sur-expression des gènes associés aux termes GO suivants :

- GO:0090382 phagosome maturation
- GO:0045335 phagocytic vesicle

De plus, on note également la présence du “GO:0034976 response to endoplasmic reticulum stress” qui semble directement liés au processus de mort cellulaire que l’on étudie. Il possèdent 7 sur-exprimés dans A15 et 5 sous-exprimés dans A15. Ces résultats sont cohérents avec les observations cellulaires menées dans l’étude de Simonet et al. 2018 [7] où il été montré que les vacuoles proviennent du réticulum endoplasmique.

Enfin, au stade A23, on observe une diminution de l’expression des gènes impliqués dans le cytosquelette qui témoigne que la croissance et l’augmentation du nombre de bactériocytes semble terminée :

- GO:0007017 microtubule-based process
- GO:0000226 microtubule cytoskeleton organization
- GO:0070507 regulation of microtubule cytoskeleton organization
- GO:0030865 cortical cytoskeleton organization
- GO:0030866 cortical actin cytoskeleton organization
- GO:0032506 cytokinetic process
- GO:0032886 regulation of microtubule-based process
- GO:0005200 structural constituent of cytoskeleton
- GO:0005874 microtubule

De plus, on voit également de nombreux termes GOs associés à des gènes sous-exprimés qui se rapportent à la prolifération et la division cellulaire :

- GO:0007346 regulation of mitotic cell cycle
- GO:0007088 regulation of mitotic nuclear division
- GO:0051783 regulation of nuclear division
- GO:0006275 regulation of DNA replication
- GO:1901990 regulation of mitotic cell cycle phase transition
- GO:0008283 cell proliferation
- GO:1901987 regulation of cell cycle phase transition

En revanche, aucune autre annotation ne semble se référer au processus de mort cellulaire.

3.2.2 Analyse des voies métaboliques et voies de signalisation impliquées dans le processus de mort cellulaire

On peut également compléter cette analyse Gene Ontology par une recherche des voies métaboliques et de signalisation impliqués dans le processus de mort cellulaire dans lesquelles les gènes différentiellement exprimés identifiés ont un rôle.

Grâce à la projection des listes de gènes différentiellement exprimés sur la base de données KEGG, on observe ainsi, au stade N3, 5 gènes impliqués dans la voie de signalisation PI3K-AKT qui intervient dans la régulation du cycle cellulaire et de l'autophagie. De même, 6 gènes sur-exprimés sont associés à la voie de l'apoptose et 6 autres impliqués dans l'activité lysosomale. Ces résultats sont cohérents avec les conclusions tirées lors de l'analyse Gene Ontology où l'on a pu montrer d'une part des processus en lien avec la croissance cellulaire et d'autre part une initiation des processus de mort cellulaire dès le stade N3.

Au stade N4, assez peu de voies métaboliques ou de signalisation semblent être impactées. On peut néanmoins observer une intensification de l'activité lysosomale et de l'autophagie car respectivement 3 et 2 des gènes différentiellement exprimés déjà présents au stade N3 sont aussi sur-exprimés au stade N4. On continue également de voir une régulation du cycle cellulaire et de l'apoptose avec deux nouveaux DEGs sur-exprimés dans N4 participant à l'activation de la de signalisation PI3K-AKT.

Au stade A9, la croissance et la prolifération cellulaire semblent très fortement impactées avec notamment :

- 30 gènes différentiellement exprimés impliqués dans la voie WNT
- 29 gènes différentiellement exprimés impliqués dans la voie HIPPO
- 24 gènes différentiellement exprimés impliqués dans la voie FOXO
- 23 gènes différentiellement exprimés impliqués dans la voie PI3K-AKT
- 10 gènes différentiellement exprimés impliqués dans la voie JAK-STAT

- 10 gènes différentiellement exprimés impliqués dans la voie HedgeHog

Toutes ces voies de signalisation sont connues pour réguler la croissance et la prolifération cellulaire.

Concernant la mort cellulaire, on retrouve également un nombre important (19) de gènes différentiellement exprimés impliqués dans l'apoptose (figure 6). Parmi ceux-ci, certains ont un contrôle négatif (XIAP/IAP) sur la voie alors que d'autres ont un contrôle positif (CASP7, cathepsines) sur l'apoptose

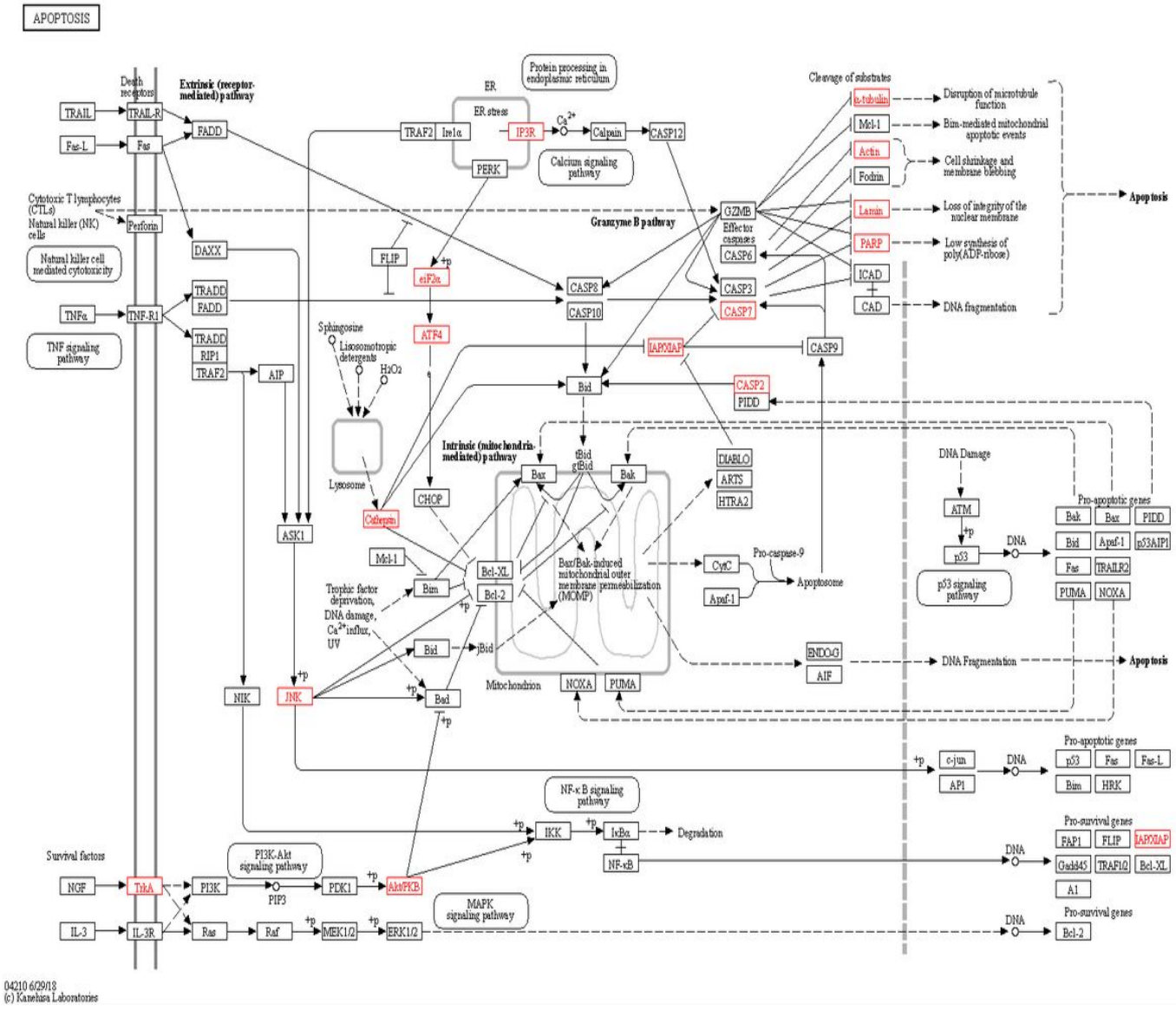

Figure 6 : Voie de l'apoptose issue de la base de données KEGG montrant en rouge les gènes différentiellement exprimés identifiés entre les conditions A9 et N4.

Le stade A9 se caractérise aussi par un nombre important de gènes différentiellement exprimés en lien avec réticulum endoplasmique. Ce qui est cohérent avec l'étude de Simonet *et al.* (2018) [7], qui montre les vacuoles observées proviennent de cette organite cellulaire. De manière cohérente, l'activité lysosomale est elle aussi impactée.

Au stade A15, l'autophagie (figure 7) s'active et l'activité lysosomale continue avec respectivement 23 et 22 gènes sur-exprimés associés. De même pour le réticulum endoplasmique pour lequel 20 gènes différentiellement exprimés ont pu être identifiés comme étant associés à la production de vésicules.

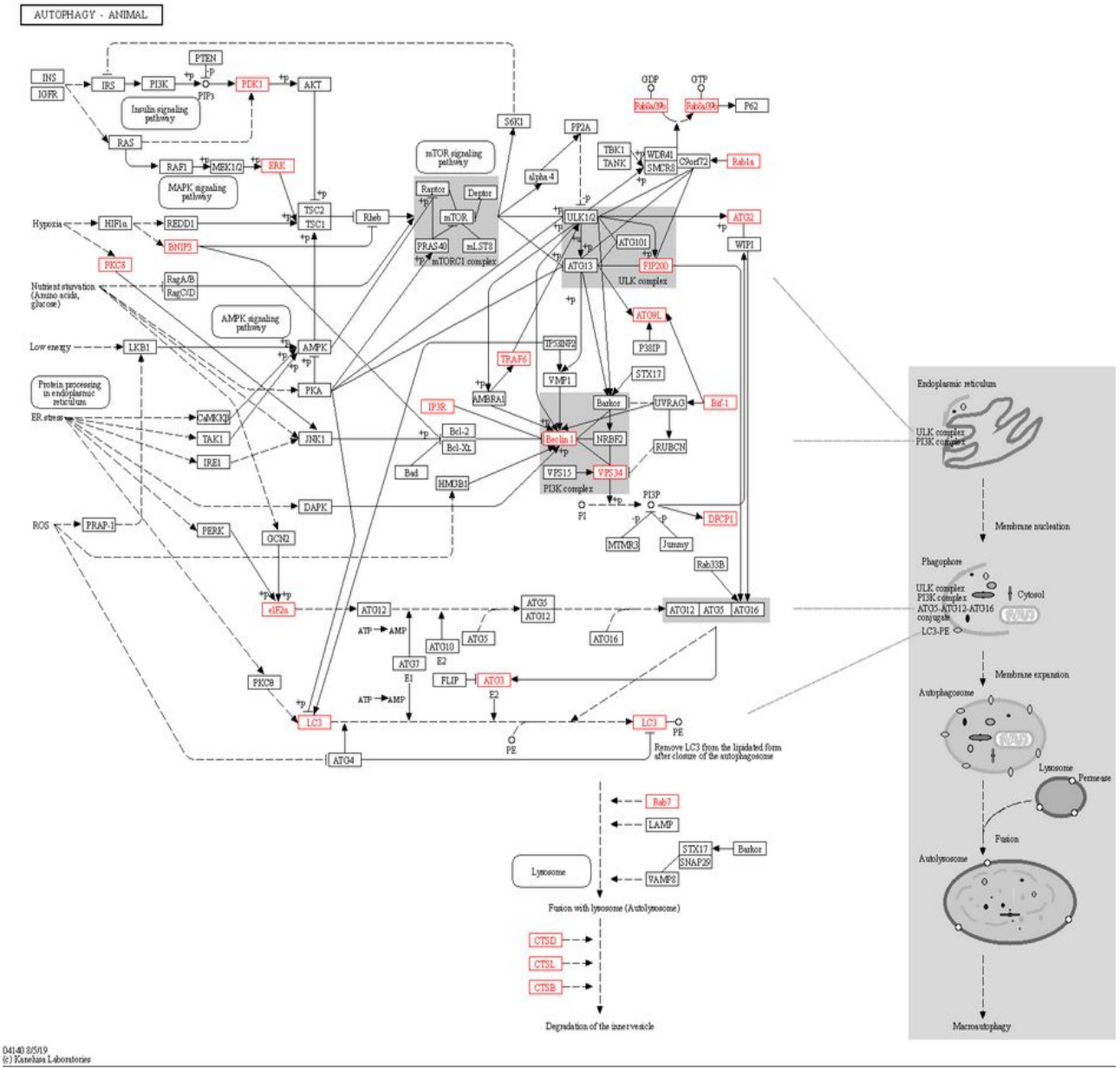

Figure 7 : Voie de l'autophagie issue de la base de données KEGG montrant en rouge les gènes différentiellement exprimés identifiés entre les conditions A15 et A9.

En ce qui concerne l'apoptose, de nombreux gènes sont une nouvelle fois différentiellement exprimés aux stades A15 et aux stades A23 sans pour autant montrer une activation ou une répression évidente de cette voie. De la même façon, la voie de signalisation PI3K-AKT est globalement réprimée à partir du stade A15, ce qui est cohérent avec une entrée des bactériocytes en sénescence.

La figure 8 résume l'ensemble de ces résultats en mettant en évidence la cascade d'évènements moléculaires identifiés au cours de cette étude.

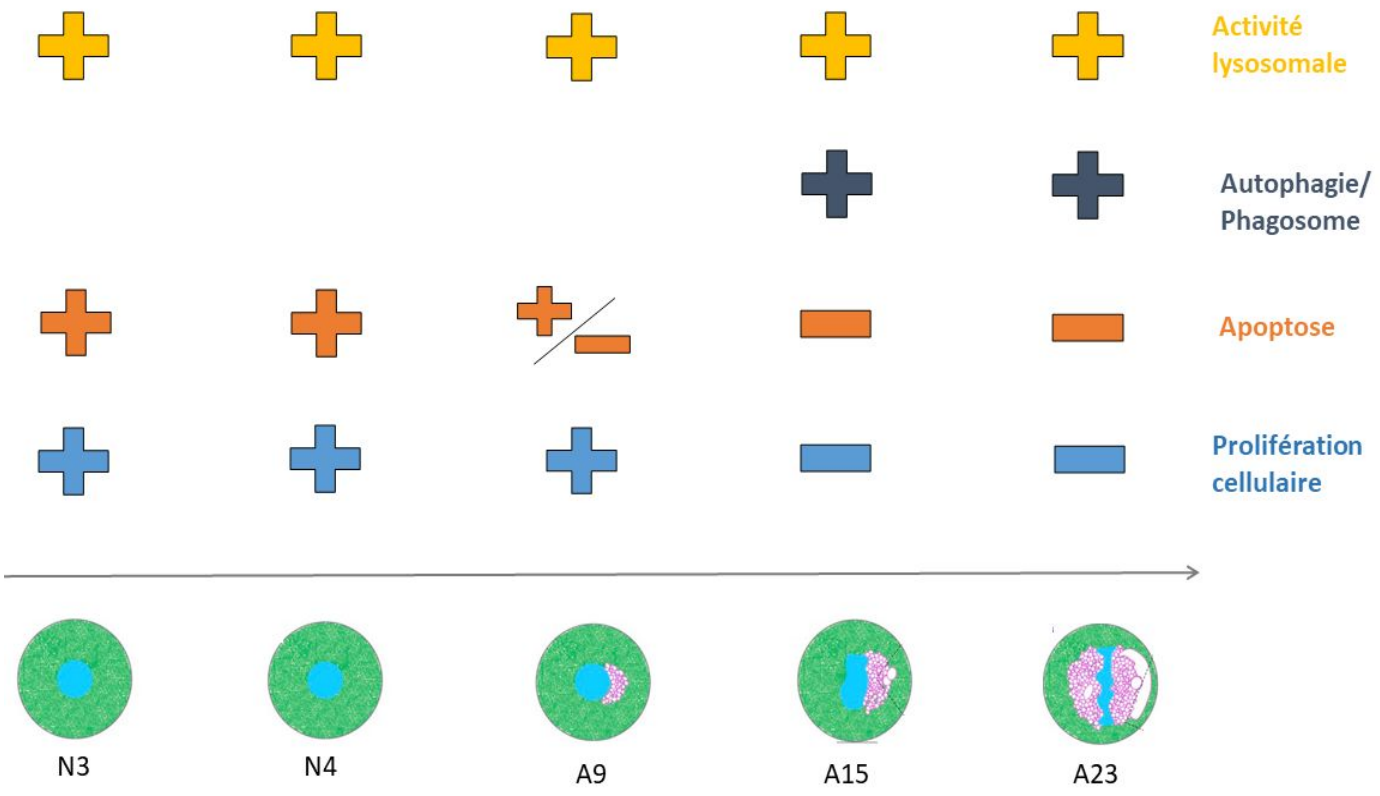

Figure 8 : Schéma récapitulatif des régulations mises en jeu tout au long du cycle de vie des bactériocytes.

4. Discussion

Cette analyse transcriptomique nous a permis de mettre en évidence les régulations transcriptionnelles impliquées dans l'homéostasie des cellules bactériocytaires du puceron du pois. Nous avons notamment pu observer que la majorité des modifications transcriptomiques semblent avoir lieu entre les stades N4, A9 et A15, lors du passage au stade adulte. Ces régulations sont cohérentes avec la physiologie de l'insecte qui passe d'une phase de croissance et de prolifération de ses bactériocytes et des symbiotes qu'ils hébergent à une phase de dégénérescence progressive caractérisée par une nouvelle forme de mort cellulaire.

On observe notamment une activation du cycle cellulaire par différentes voies de signalisation et la présence de nombreux gènes impliqués dans le cytosquelette, la production de membranes, et dans la mitose aux stades N3, N4 et A9, qui correspondent à la phase de croissance des bactériocytes. A l'inverse, ces voies de signalisation sont plutôt réprimées dans les stades A15 et A23, qui correspondent à la phase de dégénérescence des bactériocytes.

De plus, les analyses indiquent une activation de l'autophagie et la production de phagosomes à partir du stade A15, ce qui pourrait être important pour l'hypervacuolisation issue du réticulum endoplasmique qui est à la base de ce processus de mort cellulaire. On observe également une activité lysosomale croissante du stade N3 au stade A23 qui pourrait témoigner d'un contrôle du nombre de symbiotes dans les bactériocytes par l'intermédiaire de ces vacuoles.

Enfin, on constate également une régulation importante de l'apoptose tout au long de la vie du puceron sans pouvoir conclure sur une réelle tendance d'activation ou de répression. Ce processus semble donc être important pour dans la mise en place de cette

mort cellulaire bien que les travaux précédents de l'équipe aient pu montrer que cette mort cellulaire était non-apoptotique.

Ce travail a donc permis de poser les bases moléculaires du cycle de vie des bactériocytes du puceron du pois. Ces résultats confirment les observations cellulaires effectuées dans les études précédentes. Notamment celle de la dynamique d'évolution du nombre de bactériocytes et de bactéries symbiotiques [6]. Il permet également d'en apprendre plus sur le processus de mort cellulaire découvert lors d'une étude complémentaire [7]. On a également pu montrer l'importance de l'autophagie et de l'apoptose même si la régulation de ces processus dans le phénomène de mort cellulaire n'est pas encore totalement élucidée.

Pour de nombreux gènes différentiellement exprimés, nous n'avons cependant pas pu identifier de correspondance dans les annotations Gene Ontology et KEGG. L'analyse approfondie de ces gènes pourrait nous permettre d'identifier de nouveaux gènes candidats qui pourraient avoir un rôle important dans le processus étudié. De plus, les annotations Gene Ontology et la base de données KEGG proviennent d'organismes modèles et permettent l'annotation de voies métaboliques et de voies de signalisation canoniques alors que le processus de mort cellulaire étudié n'a encore jamais été décrit dans un aucun organisme. L'utilisation de telles ressources peut donc biaiser nos interprétations en associant certains gènes à des voies "classiques" alors que ceux-ci pourraient également avoir un rôle dans des processus encore inconnus. Enfin, en plus de travailler sur un organisme non-modèle, cette étude a aussi la particularité de se confronter à un modèle avec une variabilité biologique importante comme on a pu l'observer au stade N4. Cette variabilité pourrait nous empêcher de voir certaines régulations potentiellement importantes.

5. Conclusion

L'objectif de ce stage était de déterminer les processus moléculaires impliqués dans la mort cellulaire des bactériocytes du puceron du pois. Ainsi, l'équipe d'accueil disposait de banques RNA-seq obtenues obtenues à partir de bactériocytes de pucerons aux stades nymphaux N2, N3 et N4 ainsi qu'au stade adulte (9, 15 et 23 jours) correspondant aux différentes étapes clés du développement et de la dégénérescence de ces cellules.

Après avoir contrôlé la qualité des lectures de séquençage, j'ai pu identifier l'ensemble des gènes différentiellement exprimés entre les six stades de développement étudiés. Afin de faciliter l'interprétation de ces listes de gènes, j'ai ensuite procédé à une analyse d'enrichissement de leurs annotations fonctionnelles. Ces analyses montrent d'importants changements dans l'expression des gènes lors de la transition entre le stade nymphal et le stade adulte ce qui est cohérent avec la transition entre la phase de croissance des bactériocytes et des symbiotes qu'ils hébergent et leur phase de dégénérescence. On a pu également montrer l'importance des régulations de l'apoptose et de l'autophagie et des nombreuses voies de signalisation qui les régulent lors de cette transition. Ces changements sont vraisemblablement liés à l'initiation de l'hypervacuolisation des cellules qui est à la base de ce nouveau processus de mort cellulaire étudié. Cependant il reste encore beaucoup de mécanismes impliqués dans ce processus à découvrir.

Afin d'aller plus loin dans la caractérisation de ce processus de mort cellulaire, il est possible de procéder à des analyses complémentaires. On peut ainsi étudier plus en détail les gènes différentiellement exprimés qui n'ont pas de correspondance dans les bases de GO et KEGG. On pourrait également envisager d'identifier des gènes candidats, non plus au travers de comparaisons deux à deux entre chaque stade de développement, mais plutôt de façon globale en analysant leurs profils d'expression. Ces analyses pourraient nous permettre d'identifier des gènes avec des profils d'expression fortement corrélés et caractéristiques des observations cellulaires déjà réalisées.

BIBLIOGRAPHIE

1. Akman Gündüz E, Douglas AE. Symbiotic bacteria enable insect to use a nutritionally inadequate diet. *Proc Biol Sci*. 2009 doi:10.1098/rspb.2008.1476
2. Douglas AE. The molecular basis of bacterial-insect symbiosis. *J Mol Biol*. 2014; doi:10.1016/j.jmb.2014.04.005
3. Buchner Paul, *endosymbiosis of animals with plant microorganisms.* New York: interscience, 1965
4. Feng H, Edwards N, Anderson CMH, et al. Trading amino acids at the aphid-*Buchnera* symbiotic interface. *Proc Natl Acad Sci U S A*. 2019; doi:10.1073/pnas.1906223116
5. Zientz E., Silva F.J., Gross R. Genome interdependence in insect-bacterium symbioses. Genome Biology 2, 2001; doi:10.1186/gb-2001-2-12-reviews1032
6. Simonet, P., Duport, G., Gaget, K. *et al.* Direct flow cytometry measurements reveal a fine-tuning of symbiotic cell dynamics according to the host developmental needs in aphid symbiosis. *Sci Rep* 2016; doi:10.1038/srep19967
7. Simonet P, Gaget K, Balmand S, et al. Bacteriocyte cell death in the pea aphid/*Buchnera* symbiotic system. *Proc Natl Acad Sci U S A*. 2018; doi:10.1073/pnas.1720237115
8. International Aphid Genomics Consortium. Genome sequence of the pea aphid Acyrthosiphon pisum 2010 doi:10.1371/journal.pbio.1000313
9. Bolger, A. M., Lohse, M., & Usadel, B. Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*, 2014; doi:10.1093/bioinformatics/btu170
10. Liao Y, Smyth GK, Shi W The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Research*, **2019**; doi: 10.1093/nar/gkz114.
11. Love MI, Huber W, Anders S Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." *Genome Biology*, **2014;** doi:10.1186/s13059-014-0550-8.
12. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*. 2000; doi:10.1038/75556
13. The Gene Ontology Consortium, The Gene Ontology Resource: 20 years and still GOing strong, *Nucleic Acids Research* 2019; doi:10.1093/nar/gky1055
14. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. 2000 doi:10.1093/nar/28.1.27

15. Adrian Alexa, Jörg Rahnenführer, Thomas Lengauer, Improved scoring of functional groups from gene expression data by decorrelating GO graph structure, *Bioinformatics*, 2006; doi:10.1093/bioinformatics/btl140.
16. Colella S, Parisot N, Simonet P, et al. Bacteriocyte Reprogramming to Cope With Nutritional Stress in a Phloem Sap Feeding Hemipteran, the Pea Aphid *Acyrthosiphon pisum*. *Front Physiol*. 2018 doi:10.3389/fphys.2018.01498