HAL
open science

Régulation par les miARNs des gènes régulant la fécondité et le développement embryonnaire précoce chez le poisson médaka

Fanny Casse Encadrée

To cite this version:

Fanny Casse Encadrée. Régulation par les miARNs des gènes régulant la fécondité et le développement embryonnaire précoce chez le poisson médaka. Bio-informatique [q-bio.QM]. 2019. hal-02192476

HAL Id: hal-02192476

https://inria.hal.science/hal-02192476

Submitted on 23 Jul 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Régulation par les miARNs des gènes régulant la fécondité et le développement embryonnaire précoce chez le poisson médaka

Juin 2019
Master 2 de Bioinformatique de Rennes 1

Fanny CASSE

Encadrée par Julien BOBE,
INRA Rennes Laboratoire Physiologie et Génétique des Poissons, équipe SOCs,

Emmanuelle BECKER,
INRIA, équipe DYLISS, et

Fabrice LEGEAI,
INRA du Rheu

Master bioinformatique

UNIVERSITÉ DE RENNES 1

ENGAGEMENT DE NON PLAGIAT

Je, soussigné(e) ....Fanny.....CASSE.....
étudiant(e) en..Master 2...Bioinformatique.....
déclare être pleinement informé que le plagiat de documents ou d'une partie de document publiés sur toute forme de support, y comprit l'internet, constitue une violation des droits d'auteur ainsi qu'une fraude caractérisée.

En conséquence, je m'engage à citer toutes les sources que j'ai utilisées pour la rédaction de ce document.

Date : 14 / 06 / 2019

Signature :

Casse

Document à compléter de manière manuscrite et à insérer obligatoirement en première page du rapport de stage.

Fanny CASSE

Master 2 Bioinformatique, de l'Université de Rennes 1
Année universitaire 2018-2019

Résumé

Ovogénèse, RNA-seq, expression différentielle, micro-ARNs, ARNs long non-codant

L'ovogenèse repose sur des processus biologiques hautement régulés et coordonnés impliquant des interactions géniques et la régulation de gènes. Au cours de l'ovogenèse, les cellules somatiques ovariennes subissent de nombreux changements transcriptionnels afin de préparer les cellules germinales non différenciées à former des gamètes. Dans ce contexte, le rôle régulateur des micro-ARNs (miARNs) est peu connu. Des études précédentes ont permis de découvrir des miARNs particulièrement exprimés dans l'ovaire, comme miR-202 dont le KO entraine une diminution de la quantité et de la qualité des gamètes. Dans le but de mieux comprendre les réseaux moléculaires impliqués, nous avons étudié le profil transcriptomique d'ovaires de medaka (*Oryzias latipes*) au cours du cycle de reproduction. Après alignement des lectures, des annotations ont été ajoutées à celles du génome de référence permettant de prédire 1131 nouveaux longs ARNs non codants et 539 nouveaux ARN messagers. Une analyse du différentiel d'expression (DE) au cours du cycle d'ovogénèse met en évidence 2412 gènes différentiellement exprimés. Un clustering a permis d'identifier des profils d'expression différentielle pertinents suggérant une nette différence entre les premiers temps de l'ovogenèse et les temps plus tardifs. En parallèle, l'analyse a permis d'identifier 37 miARNs particulièrement exprimés dans les tissus germinaux et 197 dans les tissus non germinaux. Leurs cibles ont été prédites et analysées, suggérant que les gènes différentiellement exprimés au cours de l'ovogénèse sont préférentiellement ciblés par les miARNs germinaux, en accord avec leur rôle de régulateur. Nos résultats permettent de proposer que certains miARNs réguleraient différemment l'expression de leurs cibles au cours du cycle d'ovogenèse.

Abstract

Oogenesis, differential expression, RNA-seq, micro-RNAs, long non-coding RNAs

Oogenesis is based on highly regulated and coordinated biological processes involving gene interactions and gene regulation. During oogenesis, ovarian somatic cells undergo many transcriptional changes to prepare undifferentiated germ cells to form gametes. In this context, the regulatory role of microRNAs (miRNAs) is not well known in fish. Previous studies have found miRNAs particularly expressed in the ovary, such as miR-202, whose KO causes a decrease in gamete quantity and quality. In order to better understand the molecular networks involved, we studied the transcriptomic profile of medaka ovaries *(Oryzias latipes)* during the reproductive cycle. After reads mapping, annotations were added to those of the reference genome to predict 1131 new long non-coding RNAs and 539 new messenger RNAs. An analysis of the differential expression (DE) reveals 2412 differentially expressed genes. Clustering identified relevant differential expression profiles suggesting a clear difference between the early stages of oogenesis cycle and later ones. In parallel, the analysis identified 37 miRNAs particularly expressed in germinal tissues and 197 in non-germinal tissues. Their targets were predicted and analyzed, suggesting that genes differentially expressed during oogenesis are preferentially targeted by germinal miRNAs, in accordance with their regulatory role. Our results suggest that some miRNAs would regulate the expression of their targets differently during the oogenesis cycle.

Abréviations

ARN : acide ribonucléique

lncARN : ARN long non codant

ARNm : ARN messager

miARN : micro-ARN

ARNpi : ARN piwi

ARNr : ARN ribosomique

ARNt : ARN de transfert

cpm : count of read per millions of read in library

GO : Gene Ontology

Table des matières

Résumé 1
Abstract 1
Abréviations 2
Introduction 5
Contexte 5
Objectifs 8
Mise en place 9
Matériel et Méthodes 10
Préparation des échantillons, extraction des ARNs et séquençage 11
Alignement des lectures séquencées (Figure 4a) 11
Nouvelle annotation du génome de référence (Figure 4b) 11
Analyse différentielle des gènes ovariens au cours du temps (Figure 4c) 12
Annotation fonctionnelle (Figure 4d) 13
Clustering des gènes différentiellement exprimés selon leur profil d'expression (Figure 4e) 13
Données de micro-ARNs et analyse différentielle des miARNs germinaux et somatiques (Figure 4g) 14
Prédiction des cibles de microARN (Figure 4h) 14
Clustering des gènes différentiellement exprimés cibles des miARNs germinaux (Figure 4i) 14
Résultats 15
Alignement des lectures sur le génome de médaka 15
Enrichissement des annotations du génome de référence et comptage des lectures assignées 15
Identification des gènes exprimés dans l'ovaire et annotation Gene Ontology 16
Identification des gènes différentiellement exprimés au cours de la cinétique dans l'ovaire et annotation Gene Ontology 18
Clustering des gènes différentiellement exprimés et annotations 18
Identification des micro-ARNs germinaux et somatiques 19
Identification des cibles des miARNs différentiellement exprimés 20
Clustering des gènes cibles différentiellement exprimés en fonction de leurs micro-ARNs germinaux régulateurs 21
Correspondance entre le clustering des gènes différentiellement exprimés au cours du cycle d'ovogénèse et le clustering des cibles des miARNs germinaux 22
Discussion et Conclusion 24
Nouvel assemblage 24
Les gènes différentiellement exprimés dans l'ovaire 24

Identification des gènes différentiellement exprimés 24
Classification des gènes différentiellement exprimés au cours du temps 25
Les miARNs s'expriment différement dans les tissus 26
Lien entre l'expression des gènes et leur régulation par les miARNs 27
Bibliographie 29
Annexes 31

Introduction

Contexte

Plus de 50% de la production mondiale de poisson est réalisée en aquaculture qui dépend en partie du succès de la reproduction des poissons. Dans ce cadre, le contrôle de la fécondité, la capacité à se reproduire, des poissons joue un rôle important en agronomie.

Chez les poissons, la fécondité des femelles est particulièrement liée à l'ovogénèse dans l'ovaire. Au cours de l'ovogénèse, les cellules somatiques ovariennes subissent de nombreux changement transcriptionnels et moléculaires aboutissant à la formation des gamètes femelles matures aptes à être fécondées (Iwamatsu 2004b). Ces processus biologiques, impliqués dans l'ovogénèse et l'embryogénèse précoces, sont finement régulés et coordonnées par des interactions géniques et la régulation de gènes. L'ovaire est un organe complexe comportant les différents stades ovocytaires en cours de maturation et de différentiation, et les cellules somatiques environnantes formant les follicules *(Figure 1).*

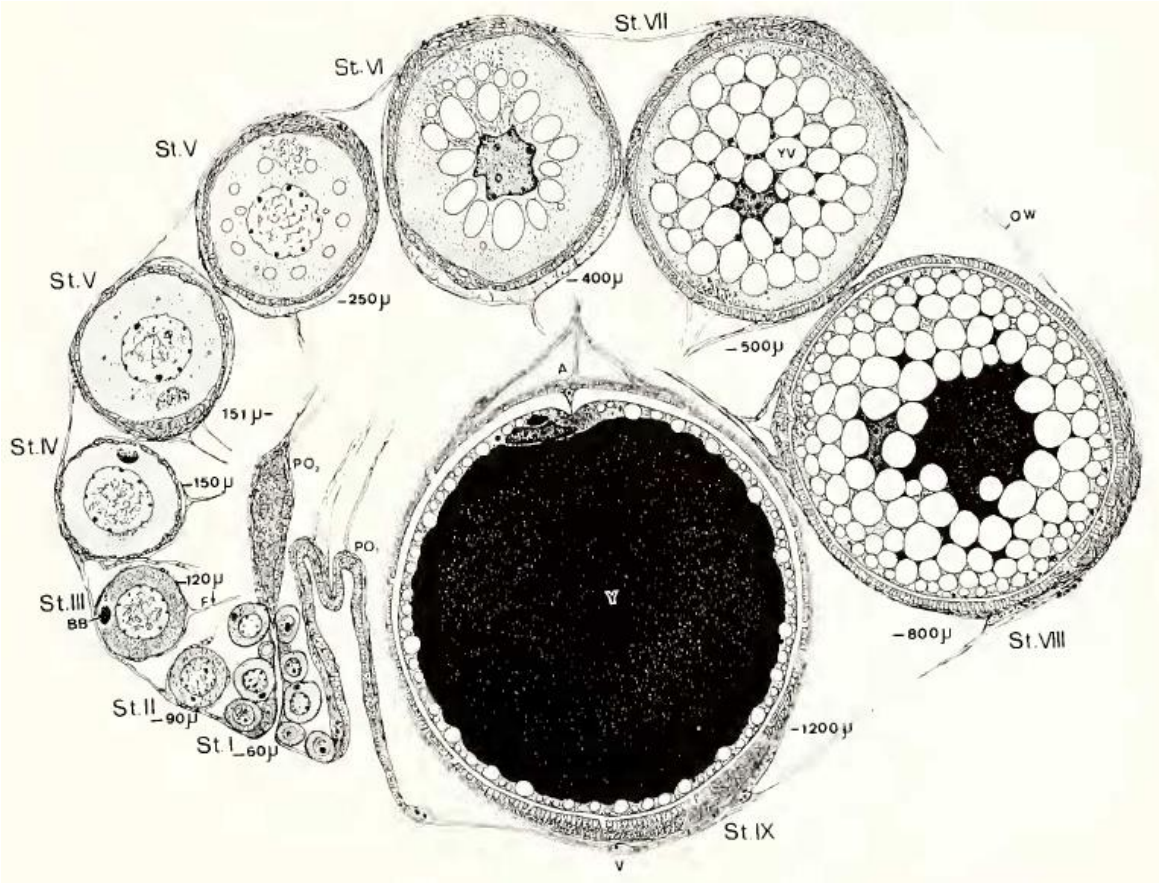


Figure 1: Ovogènese chez le médaka. L'ovogénèse chez les poissons téléostéens tel que le médaka est classifiée en 4 phases : prévitellogénèse (stade I à IV), vitellogénèse (stade V à VIII), postvotellogénèse (stade IX) et phase ovulatoire (stade X). La phase prévitellogénique est constituée de 4 stades : le stade nucléaire (st I) de la chromatone où l'ovocyte est de diamètre entre 20 et 60 µm et entourés de cellules folliculaires plates, le stade peri-nucléaire (st II) où l'ovocyte mesure 61 à 90 µm de diamètre, l'étape de l'enveloppe de chorion (st III) où l'ovocyte mesure entre 91 et 120 µm et où il développe leur enveloppe d'œuf, l'étape de fixation du filament et de la formation de gouttelette d'huile (st IV) où le diamètre de l'ovocyte est entre 120 et 150 µm et le diamètre de son noyau est entre 75 et 90 µm. La phase vitellogénique est constituée de 4 stades : le stade vésiculeux (st V) précoce où l'ovocyte mesure entre 151 et 250 µm et la vésicule vitelline se développe dans le cytoplasme, le stage vésiculaire tardif (st VI) où le diamètre de l'ovocyte est de 251 à 400 µm et où la vésicule vitelline est bien développée formant une couche autour du noyau et une couche folliculaire bien développée entoure l'ovocytes, le stade

précoce de formation du vitellus (st VII) où l'ovocyte mesure entre 401 et 500 µm et où les globules vésiculaires sont fusionnés, le stade de formation tardive du vitellus (st VIII) où l'ovocyte mesure entre 501 et 800 µm et les vésicules vitellines sont poussées vers la parties externes du cytoplasme de l'ovocyte. La phase postvitellogénique est le stade de maturation, l'ovocyte atteint un diamètre entre 801 et 1200 µm et le vitellus occupe la majorité de l'ovocyte. La phase d'ovulation où le diamètre de l'ovocyte est de 1200 µm et où l'ovocyte se sépare de la couche folliculaire pour passer dans la cavité ovarienne pour l'ovulation (Iwamatsu 2004)].

Chez le médaka (*Oryzias latipes*), la ponte des femelles est très régulière. En condition d'élevage, à partir de 3 mois (âge du début de la reproduction), les médakas femelles pondent chaque jour en moyenne une vingtaine d'œufs, dans l'heure qui suit le lever du soleil (Gay et al. 2018). De nombreuses études ont exploré les processus endocriniens impliqués dans l'ovogénèse chez le poisson (Lubzens et al. 2010), mais peu d'études ont jusqu'à présent analysé la dynamique de l'expression des gènes au cours du processus.

La transcriptomique fait référence à l'étude du transcriptome complet, incluant les ARNs messagers mais aussi les ARNs non codants dans une cellule, un tissu ou organisme spécifique pour un stade de développement ou état physiologique donné (Wang et al. 2019). Contrairement au génome relativement stable, le transcriptome varie avec le stade de développement et l'état physiologique. L'analyse du transcriptome est un outil puissant pour disséquer la relation entre le génotype et le phénotype et identifier des voies mécaniques qui contrôlent le développement des cellules.

Si les gènes codants pour des protéines ont été pendant de nombreuses années l'objet principal des études de transcriptomique, les ARNs non codants sont plus récemment devenus un nouveau centre d'intérêt, toujours dans le but de mieux comprendre et expliquer les phénomènes biologiques. Les ARNs non codants n'ont pas vocation à être traduit en protéines. Ils peuvent être classés en ARNs domestiques tel que les ARN ribosomique (ARNr) et les ARN de transfert (ARNt) et en ARNs régulateurs tel que les micro-ARN (miARN), les ARN long non codant (lncARN) et les ARN piwi (ARNpi) (Romano et al. 2017).

Les miARNs sont de courts ARNs non codants, simple brin, ayant une longueur comprise entre 20 et 24 nucléotides. Les gènes codant pour des miARNs sont transcrits sous la forme de longs précurseurs pri-miARNs. Chez les métazoaires, ils sont ensuite clivés pour devenir un pré-miARN mesurant environ 70 nucléotides et repliés en tige-boucle. Après transport dans le cytosol, le pré-miARN est clivé et libère deux brins d'ARN. L'un de ses brins deviendra plus abondant en nombre dans la cellule et sera appelé brin mature. Le brin mature sera le plus actif dans la cellule (Alberti et Cochella 2017) *(Figure 2).*

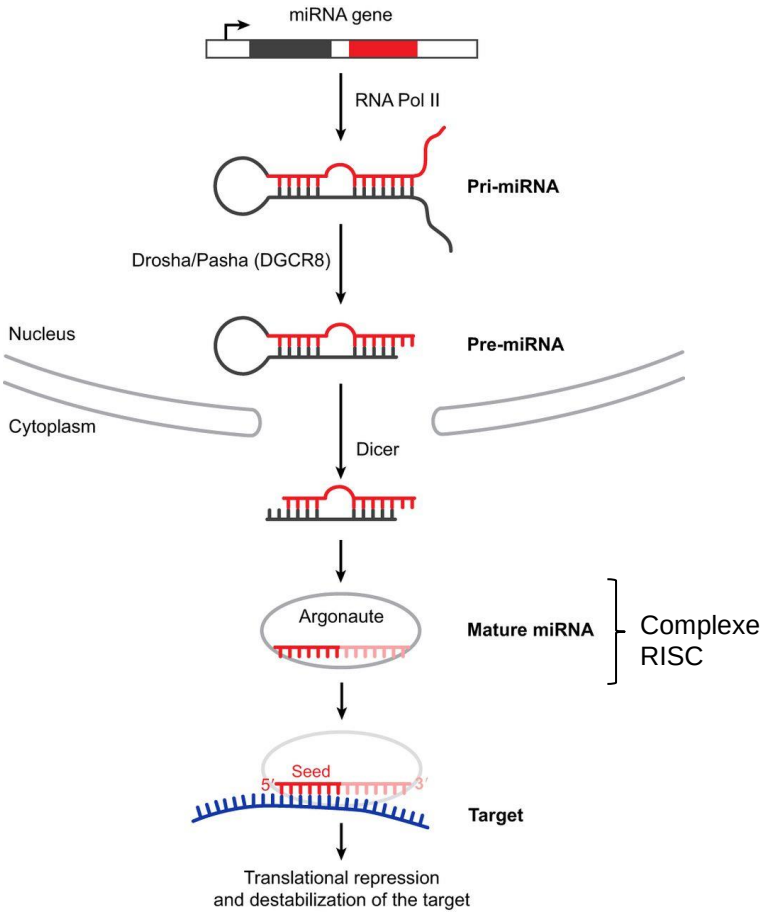


Figure 2: Biogénèse des microARNs. Les gènes codant pour des miARNs sont transcrits sous la forme de longs précurseurs pri-miARNs à l'aide de l'ARN polymérase II. Chez les métazoaires, ils sont ensuite clivés dans le noyau par un complexe Drosha/Pasha (DGCR8) pour devenir un pré-miARN mesurant environ 70 nucléotides et repliés en tige-boucle par complémentarité de base entre la première moitié et sa deuxième moitié de séquence. Après transport dans le cytosol, le pré-miARN est clivé par l'enzyme Dicer qui coupe la boucle et libère un petit ARN double-brin appelé duplexe miARN, ses deux brins se séparent (brin 5p et brin 3p). L'un de ses brins deviendra plus abondant en nombre dans la cellule et sera appelé brin mature. Le brin mature sera le plus actif dans la cellule en se couplant aux protéines Argonaute pour former le complexe RISC (RNA-induced silencing complex). Les miARNs guident ensuite le complexe RISC pour s'associer aux ARNms cibles **(Alberti et Cochella 2017).**

Le mode d'action des miARNs passe par une régulation post-transcriptionnelle des gènes, en se fixant sur leur ARNms cible, via le complexe RISC *(Figure2).* Les miARNs sont constitués d'un domaine « graine » composé généralement de 7 nucléotides, entre les nucléotides 2 et 8 du miARN (Brennecke et al. 2005), qui joue un rôle primordial dans son association avec les ARNms. Pour se fixer à son ARNm cible, le miARN s'hybride à sa cible au niveau de la graine, c'est-à-dire qu'il y a complémentarité de séquence entre le miARN et l'ARNm cible. Si l'hybridation est totale entre le miARN et l'ARNm cible, cela entraine la dégradation de l'ARNm, alors qu'une hybridation partielle induit son inactivation passant par la répression de sa traduction *(Figure 3).*

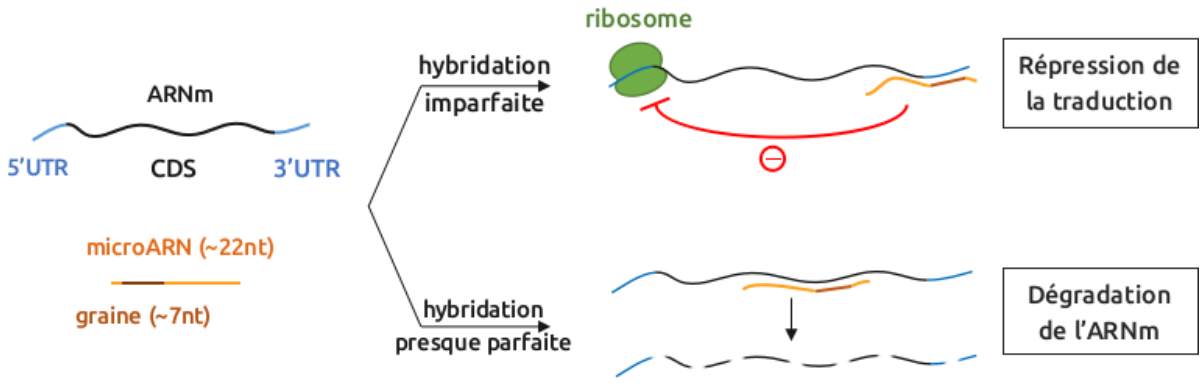


Figure 3: Régulation post-transcriptionnelle des ARN messagers par les micro-ARNs. Les miARNs d'environ 22 nucléotides sont constitués d'un domaine graine de 7 nucléotides, via ce domaine ils s'associent à leur ARNm cible par complémentarité de séquence. Les miARNs ont deux modes de régulation. L'hybridation imparfaite, c'est-à-dire une complémentarité de bases incomplète, du miARN à son ARNm cible induit une répression de la traduction de l'ARNm cible, une hybridation complète entraine la dégradation de l'ARNm cible.

Des études précédentes, chez les mammifères, suggèrent que les miARNs jouent un rôle important lors de l'ovogénèse. Chez la souris, la suppression de Dicer1 dans l'ovaire entraine l'infertilité des femelles (Otsuka et al. 2008). Il semblerait donc que le développement de l'ovaire soit régulé par les miARNs et nécessite leur maturation à l'aide de Dicer1. Dans les testicules de souris et chez l'homme, un miARN, le miR-202, est plus fortement exprimé dans les cellules de Sertoli et dans les cellules germinales (Wainwright et al. 2013; Dabaja et al. 2015), où il joue le rôle de régulateur clé dans la détermination des cellules souches spermatogoniales (Chen et al. 2017). Cependant, peu d'études fonctionnelles ont été réalisées chez le poisson. Chez le poisson zèbre, une étude montre que le même miARN, miR-202, est prédominant dans les gonades au cours du développement et à l'âge adulte (Presslauer et al. 2017). L'équipe de Julien Bobe au LPGP a montré que chez le médaka certains miARNs possèdent une expression ovarienne prédominante (Bouchareb et al. 2017), tel que le miR-202 (Gay et al. 2018). De plus, le knock-out de miR-202 induit une diminution de la fréquence de ponte, du nombre d'œufs et de leur qualité. Ainsi, chez les medakas knock-out miR-202, 85% des femelles sont substériles et 15% sont stériles (Gay et al. 2018). Ces différents résultats suggèrent un rôle primordial de ce miARN dans la formation des gonades et la fécondité.

Objectifs

Dans ce contexte de reproduction cyclique, mon projet de recherche comportait trois objectifs. Le premier était la caractérisation de l'expression des gènes par une cinétique couvrant le cycle journalier de reproduction chez le medaka et d'identifier des profils types de gènes. Le second objectif était de compléter l'étude des miARNs particulièrement exprimés dans les tissus germinaux en identifiant leurs gènes cible via des prédictions *in silico*. Le dernier était de lier ces deux résultats pour potentiellement identifier des cibles de miARNs germinaux dont l'expression évolue au cours d'un cycle d'ovogénèse.

Mise en place

Pour atteindre ces objectifs, une collaboration entre Julien Bobe, directeur de recherche au Laboratoire de Physiologie et Génomique des Poissons de l'INRA de Rennes, étudiant la fécondité des poissons, ainsi qu'Emmanuelle Becker, enseignante-chercheuse en bioinformatique dans l'équipe DYLISS à l'INRIA et Fabrice Legeai, bioinformaticien de l'INRA du Rheu rattaché à l'INRIA, a été mise en place.

Pour observer les changements transcriptionnels de l'ovaire au cours du temps, et les relier à la régulation des miARNs, le laboratoire LPGP de l'INRA a réalisé un séquençage à haut débit du transcriptome d'ovaire totale, à intervalles réguliers au cours d'un cycle d'ovogénèse (6 temps, 24h). Mon premier objectif était donc d'analyser les résultats de séquençage pour identifier les gènes exprimés dans l'ovaire, puis ceux différentiellement exprimés au cours du cycle. Cet objectif a nécessité une première étape visant à compléter les annotations du génome afin de produire un nouveau génome de référence. De nouveaux lncARNs et ARNm ont été découverts. A partir de comptages réalisés sur le nouvel assemblage, une analyse différentielle a été réalisée pour identifier les gènes différentiellement exprimés au cours du cycle, complétée par une analyse d'enrichissement en annotations Gene Ontology.

En parallèle de cette analyse transcriptomique de l'ovaire, le laboratoire LPGP avait préalablement réalisé un séquençage des miARNs dans différents tissus de médaka. Le second objectif était d'identifier les miARNs germinaux, nécessitant une analyse différentielle entre les tissus germinaux et somatiques. Cette identification a été suivie d'une recherche des cibles potentielles pour chacun de ces miARNs. Nous proposons ensuite un clustering des gènes cibles des miARNs, en fonctions des miARNs avec lesquels ils interagissent.

Le croisement de ces deux études nous permettra de répondre à la question suivante : Des gènes montrant un profil d'expression différentiel au cours du cycle d'ovogénèse peuvent-ils être reliés à des miARNs spécifiquement exprimés dans les tissus germinaux ?

Matériel et Méthodes

L'analyse effectuée pour répondre à ses problématiques est résumée dans le workflow de la figure 4. Chaque étape est plus amplement décrite dans la suite du matériel et méthodes.

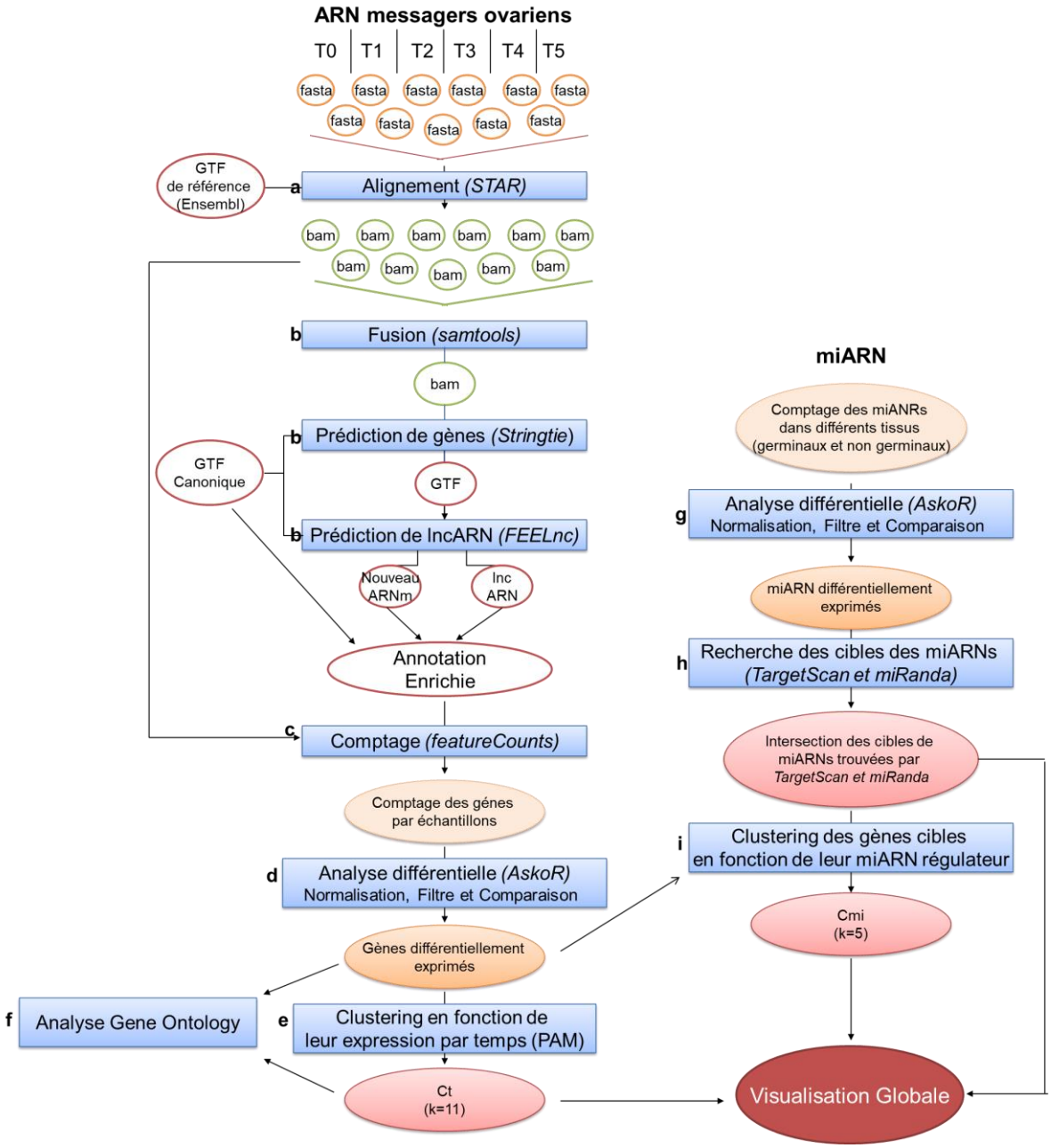


Figure 4: Workflow d'analyse. Les fichiers de séquençage, sous forme fasta contenant les lectures, des échantillons à 6 temps (T0, T1, T2, T3, T4 et T5) sont alignés sur le génome de référence du médaka, obtenu sur la plateforme ENSEMBL. Les fichiers d'alignement, sous forme bam sont ensuite fusionnés pour obtenir un unique fichier comportant tous les alignements. L'utilisation de Stringtie et de FEELnc permet respectivement d'identifier de nouvelle annotation d'ARNms et de lncARNs. Celles-ci sont ajoutées à l'annotation du génome de référence. Le comptage des lectures alignées est réalisé sur le génome nouvellement annoté, puis analysé différentiellement à l'aide d'AskoR pour identifier les gènes différentiellement exprimés au cours de la cinétique. Les gènes différentiellement exprimés sont analysés à l'aide des annotations Gene Ontology (GO). Ils sont clustérisés en fonction de leur niveau d'expression au cours du temps, puis ces clusters sont eux-mêmes analysé au moyen des annotations GO. En parallèle, les miARN comptages dans différents tissus (germinaux : ovaire, testicule, œufs, stades embryonnaires 1 et 8 cellules, follicules et non germinaux : yeux, cerveau, branchies, cœur, muscle, fois, rein, intestin, nageoire et stades embryonnaires 27, 31, 35 et 39) sont fournis par le LPGP. Une analyse d'expression différentielle entre tissus germinaux et non germinaux est réalisée. Les gènes cible des miARNs sont identifiés à l'aide de TargetScan et miRanda. Les deux études précédentes sont ensuite intégrées. Les gènes différentiellement exprimés cibles des

miARNs germinaux sont clusterisés suivant leurs miARNs régulateurs. Les clusters de gènes différentiellement exprimés au cours de la cinétique sont analysés afin de voir s'ils sont plus ciblés par les miARNs germinaux ou non germinaux. Une visualisation globale est proposée.

Préparation des échantillons, extraction des ARNs et séquençage

Les expériences ont été réalisées par Violette Thermes et Stéphanie Gay du Laboratoire de Physiologie et Génomique du Poisson de l'INRA de Rennes, dans le strict respect de la réglementation française et européenne sur les recommandations en matière de bien-être animal et ont été approuvées par le Comité de Protection et d'Utilisation des Animaux de l'INRA LPGP. Les médakas (Oryzias latipes) ont été élevés à 26°C. Les juvéniles se sont développés sous un régime de photopériode de croissance (12h jour/12h nuit) jusqu'à l'âge de 3 mois. A partir de 3 mois, les poissons adultes ont été élevés sous un régime de photopériode de jours longs (14h jour/10h nuit) permettant leur reproduction. Des dissections tissulaires ont été réalisées sur des poissons medakas adultes, après euthanasie par immersion dans une dose létale de tricaine à 30-50mg/L.

Les ARNs ont été séquencés par l'entreprise MGX-Montpelier GenomiX (Illumina HiSeq2500). Ils représentent le transcriptome ovarien de 36 poissons femelles médakas de type sauvage (wild type). Les ovaires ont été prélevés à six temps différents après la ponte (a.p.) (n=6 par temps) : T0 : 0h a.p., T1 : 4h a.p., T2 : 8h a.p., T3 : 12h a.p., T4 : 16h a.p. et T5 : 20h a.p, et broyés pour extraction l'ARNs dans le trizol *(Figure 5).*

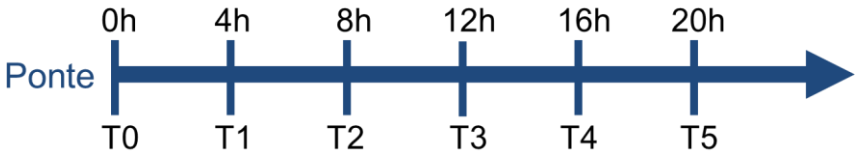


Figure 5: Extraction des ovaires de medaka au cours du cycle d'ovogénèse de 24h. Les ovaires sont prélevés juste après la ponte à T0, puis toute les 4h c'est-à-dire 4h après la ponte : T1, 8h après la ponte : T2, 12h après la ponte : T3, 16h après la ponte : T4 et 20h après la ponte : T5 . 6 ovaires ont été prélevés par temps.

Alignement des lectures séquencées (Figure 4a)

Le logiciel STAR (Dobin et al. 2013) a été utilisé pour aligner les lectures sur le génome de référence du medaka : Japanese medaka HdrR (ASM223467v1) version 95 (ENSEMBL). L'aligneur prend en entrée les fichiers fasta des reads séquencés et le fichier l'indexage du génome et rend en sortie un fichier d'alignement (BAM) par chaque échantillon. Il a été réalisé en autorisant 3 alignements multiples au maximum, 3 substitutions et une taille d'intron entre 10 et 50000 pb.

Nouvelle annotation du génome de référence (Figure 4b)

Pour compléter l'annotation du génome de référence de médaka : Japanese medaka HdrR (ASM223467v1) version 95 (EMBL), une étape de fusion des fichiers bam obtenus après alignement a été réalisée. Le résultat a été utilisé pour prédire de nouveaux transcrits (ARNms et ARN non codant), à l'aide de l'outil StringTie (Pertea et al. 2015). StringTie est un assembleur d'alignements de RNA-seq, il permet d'identifier la structure du génome (délimitation des transcrits) à partir des lectures alignées sur le génome. Cet assemblage est

comparé aux annotations déjà connues pour déterminer quels sont les nouveaux transcrits identifiés. Pour prédire de nouveaux lncARNs, l'outil FEELnc (FIExible Extraction of LncRNAs) (Wucher et al. 2017) a été utilisé. FEELnc réalise une première étape de filtrage des transcrits candidats en retirant les transcrits de moins de 200 nt. Cette étape est suivie de l'annotation précise des lncARNs à l'aide d'une forêt aléatoire analysant des critères tels que les fréquences de k-mers et les cadres de lecture ouverts (ORF).

Le comptage par gène a été effectué sur le génome nouvellement annoté à l'aide de featureCounts (Liao, Smyth, et Shi 2014). Le logiciel Subread featureCounts sert à quantifier les lectures alignées. Il prend en compte les fichiers de lectures alignées (format SAM ou BAM) produits lors de l'alignement et le fichier d'annotation du génome au format GTF pour compter le nombre de lectures par entité (gènes, promoteurs, exons, corps géniques, cellules génomiques et emplacements chromosomiques). Nous avons choisi de prendre les paramètres par défaut.

Analyse différentielle des gènes ovariens au cours du temps (Figure 4c)

Pour analyser l'expression des gènes, un seuil d'expression à 5 cpm (comptages par millions de lectures) dans au moins 4 échantillons a été fixé. Ce filtre permet de ne pas prendre en compte les faibles niveaux de comptage pouvant être dû à une incertitude de quantification (bruit) que l'on préfère ignorer. Pour réaliser l'analyse différentielle, le package AskoR (https://github.com/askomics/askoR) a été utilisé. AskoR est un package R créé pour l'analyse différentielle dans le laboratoire INRA par Sylvin Masanelli, au cours de son stage de Master2 encadré par Fabrice Legeai, et repris par Fabrice Legeai et Susete Alves-Carvalho. Il se base sur edgeR (Robinson, McCarthy, et Smyth 2010), un outil d'analyse de l'expression différentielle des gènes dans lequel la distribution des comptages est modélisée par une loi binomiale négative. La méthode de normalisation choisie est TMM (Trimmed Mean of M-values) qui exclut les gènes les plus fortement exprimés et se base sur l'hypothèse que la majorité des gènes ne sont pas exprimés de manière différentielle. Le facteur de normalisation est calculé par rapport à la moyenne géométrique entre les différents échantillons, ce qui permet d'être moins sensible aux valeurs élevées. La méthode de correction pour les tests multiples utilisée est celle de Benjamini et Hochberg (1995) ("BH"), de manière à ne pas être trop strict et de sélectionner un grand nombre de caractères potentiellement intéressants.

Pour comparer les différences d'expression des gènes dans les ovaires d'*O. latipes* pendant le cycle d'ovogenèse, 15 contrastes comparant deux à deux les différents points de la cinétique de l'ovogenèse (T0, T1, T2, T3, T4 et T5) ont été réalisés *(Figure 6).*

	T0	T1	T2	T3	T4	T5
T0		x	x	x	x	x
T1			x	x	x	x
T2				x	x	x
T3					x	x
T4						x
T5						

Figure 6 : Matrice de contrastes des 15 comparaisons effectuées deux à deux entre chaque temps (T0 à T5) du cycle ovarien.

Annotation fonctionnelle (Figure 4d)

Pour annoter fonctionnellement les gènes trouvés différentiellement exprimés, les annotations Gene Ontology ont été utilisées à l'aide du package R topGO (Alexa, Rahnenfuhrer, et Lengauer 2006) se servant du test statistique de Fisher. Les annotations Gene Ontology sont destinées à structurer la description fonctionnelle des gènes à l'aide d'un vocabulaire contrôlé commun à toutes les espèces et hiérarchisé sous forme de graphe acyclique orienté (DAG). Les propriétés fonctionnelles sont définies selon 3 axes : les composants cellulaires (répertoriant la localisation où les produits géniques remplissent une fonction), les fonctions moléculaires réalisées (activités qui se produisent au niveau moléculaire) et les processus biologiques (tel que la réparation de l'ADN ou la transduction de signal).

Clustering des gènes différentiellement exprimés selon leur profil d'expression (Figure 4e)

Un clustering des gènes différentiellement exprimés en fonction de leur profil d'expression au cours du cycle d'ovogénèse a été réalisé en se basant sur la méthode PAM (Partitioning Around Mediod) (Kaufman et Rousseeuw 1987).

PAM est une méthode de classification automatique, basée sur la notion de médoïde. Un médoïde est un point du cluster dont la dissimilarité moyenne par rapport aux autres points du cluster est minimale. Le principe de cette méthode est de réallouer itérativement chaque point, ici un gène, au groupe comportant les individus lui ressemblant le plus possible. PAM a pour objectif de former des groupes de manière à ce que la variance intragroupe soit la plus faible et par conséquent la variance intergroupe soit la plus grande. Cette méthode est plus robuste aux valeurs atypiques que la méthode des k-means, basée sur les moyennes des groupes.

Pour réaliser un clustering non supervisé avec la méthode PAM il est nécessaire de fixer un nombre de groupes préalablement. Différents indicateurs permettant de déterminer les nombres de cluster optimal ont été développés (voir packages Nbclust), leurs préconisations étant parfois contradictoire. Avec nos données, la valeur optimale du nombre de clusters n'est pas claire (les différents critères faisant ressortir par exemple k=4 avec la méthode Elbow ou k=8 avec la méthode Silhouette). Nous avons choisi k=11 groupes qui semblait l'un des meilleurs choix proposés et présentait des groupes pertinent d'un point de vue biologique.

Données de micro-ARNs et analyse différentielle des miARNs germinaux et somatiques (Figure 4g)

Les données des miARNs proviennent d'une étude précédente (Gay et al. 2018).

Après séquençage, alignement et comptage des lectures par miARNs dans différents tissus, une analyse d'expression différentielle a été réalisée comparant les tissus germinaux, c'est-à-dire contenant des cellules germinales ou des transcrits hérités maternellement (ovaire, testicule, œufs, stades embryonnaires 1 et 8 cellules, follicules) et les tissus non germinaux, c'est-à-dire ne contenant pas de cellules germinales (yeux, cerveau, branchies, cœur, muscle, fois, rein, intestin, nageoire et stades embryonnaires 27, 31, 35 et 39 cellules), permettant d'identifier les miARNs germinaux et non germinaux (somatiques). L'analyse différentielle des miARNs en fonction de leur expression dans les tissus a été aussi réalisée à l'aide d'AskoR. Les seuils de filtrage sont de 0,5 cpm dans au moins 3 échantillons sur les 19.

Prédiction des cibles de microARN (Figure 4h)

Nous avons analysé les 3'UTR des ARNms et les lncARNs de tous les gènes du génome de médaka. Afin de prédire si ces 3'UTR contenaient des sites de fixations pour chacun des 234 miARNs classés « germinaux » ou « somatiques », nous avons utilisé TargetScan (Lewis, Burge, et Bartel 2005) et miRanda (Enright et al. 2003).

TargetScan calcule un score permettant de prédire les cibles de miARNs. Ce logiciel utilise en entrée la séquence des miARNs matures, la graine de celui-ci et la séquence génomique de l'organisme. Il recherche dans le génome la présence de 6mers, 7mers et 8mers complémentaires à la région graine des miARNs donnés. Le score, prenant en compte différents paramètres (Grimson et al. 2013), donne une idée de la fiabilité de la prédiction. Plus le score est bas, plus la prédiction est fiable.

miRanda prend en entrée les séquences de miARNs mature et les séquences génomiques d'intérêt. Il calcule un score basé que la complémentarité des nucléotides entre la séquence du miARN et la séquence génomique et estime la stabilité thermodynamique des liaisons par un minimum d'énergie en kcal/mol. miRanda place par défaut un seuil de score à 140 (score≥140) et une énergie inférieure à 0 (E≤0 kcal/mol).

Les méthodes de prédiction des cibles de miARNs ont tendance à générer de nombreux faux positifs (Mockly et Seitz 2019). Pour réduire ce taux de faux positifs, seuls les cibles prédites obtenus par les deux algorithmes ont été conservées.

Clustering des gènes différentiellement exprimés cibles des miARNs germinaux (Figure 4i)

Un clustering des gènes différentiellement exprimés et cibles des miARNs germinaux a été réalisée en fonction de leurs miARNs germinaux régulateurs, afin d'identifier des groupes de gènes différentiellement exprimés ciblés par les mêmes groupes de miARNs. La méthode PAM (Partitioning Around Mediod)(Kaufman et Rousseeuw 1987) a été utilisée comme précédemment. Nous avons choisi à 5 groupes pour la pertinence biologique des groupes obtenus.

Résultats

Alignement des lectures sur le génome de médaka

L'étape d'alignement permet d'aligner les fragments de lectures sur le génome de référence. L'alignement des lectures réalisé sur le génome de référence de médaka v95 (ENSEMBL) a permis d'aligner en moyenne 89,6% (écart type : 0,6) des lectures. Selon les échantillons, le nombre de lectures séquencées varie entre 37 et 55 millions de lectures. Le taux d'alignement sur le génome de référence médaka est relativement stable dans les différents échantillons et varie entre 33,1 et 49,5 millions de lectures *(Figure 7).*

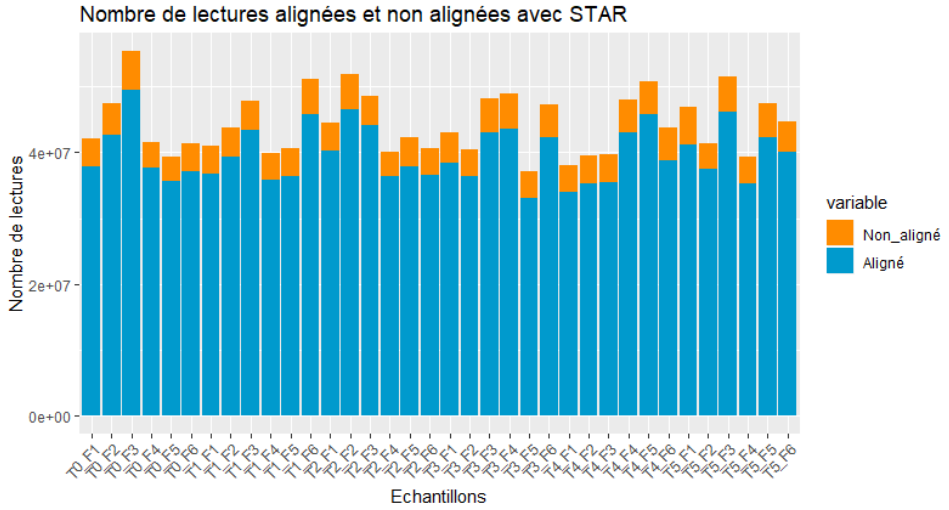


Figure 7: Nombre de lectures alignées et non alignées par échantillons avec l'outil d'alignement STAR. 89,9% (écart type : 0,6) des lectures sont alignées sur le génome de référence.

Enrichissement des annotations du génome de référence et comptage des lectures assignées

Le génome du médaka n'est pas aussi bien annoté que d'autres génomes comme celui de la souris, dans le but de compléter nos connaissances, une nouvelle annotation du génome de référence de médaka a été réalisée. 1131 nouveaux lncARNs et 539 nouveaux ARNm ont été annotés. Ces nouvelles annotations d'ARNs ont été ajoutées à l'annotation initiale du génome de référence de médaka *(Figure 8).*

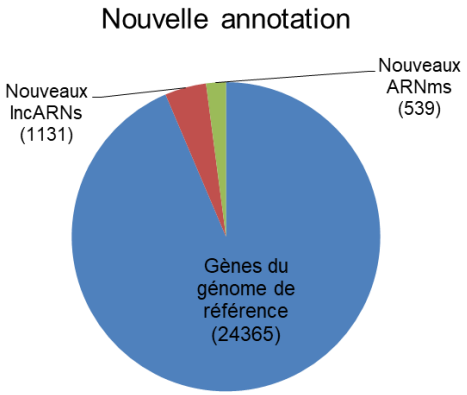


Figure 8 : Répartition des gènes de la nouvelle annotation. La nouvelle annotation réunie les annotations des gènes du génome de référence (24365), ainsi que les nouvelles annotations de transcrits d'ARN codants (539) et de lncARNs (1131).

Sur le génome de référence, 68,6% (écart type : 4,9) des lectures sont assignées à des exons. 31,4% (écart type : 5,6) des lectures ne sont pas assignées à des exons dont 4,2% (écart type : 0,3) ne sont pas assignées à des exons car leur assignation est ambiguë c'est-à-dire que la lecture peut s'aligner sur différent exons du génome. Selon les échantillons le nombre de lectures assignées varie entre 23,9 et 38,1 millions de lectures. Le taux d'alignement sur la première annotation du génome de référence médaka est relativement stable dans les différents échantillons et varie entre 33,1 et 49,5 millions de lectures *(Figure 9A).*

Sur le génome nouvellement annoté, le second alignement des lectures permet d'assigner 82,5% (écart type : 2) lectures à des exons sur le génome de référence. Cette étape permet de réduire à 17,4% (écart type : 2,5) les lectures non assignées à des exons dont 10,5% (écart type : 0,3) ne sont pas assignées à des exons car leur assignation est ambiguë. Selon les échantillons le nombre de lectures assigné varie entre 29,9 et 45 millions de lectures. Le taux d'alignement sur la nouvelle annotation du génome de référence médaka est relativement stable dans les différents échantillons et varie entre 33,1 et 49,5 millions de lectures *(Figure 9B).*

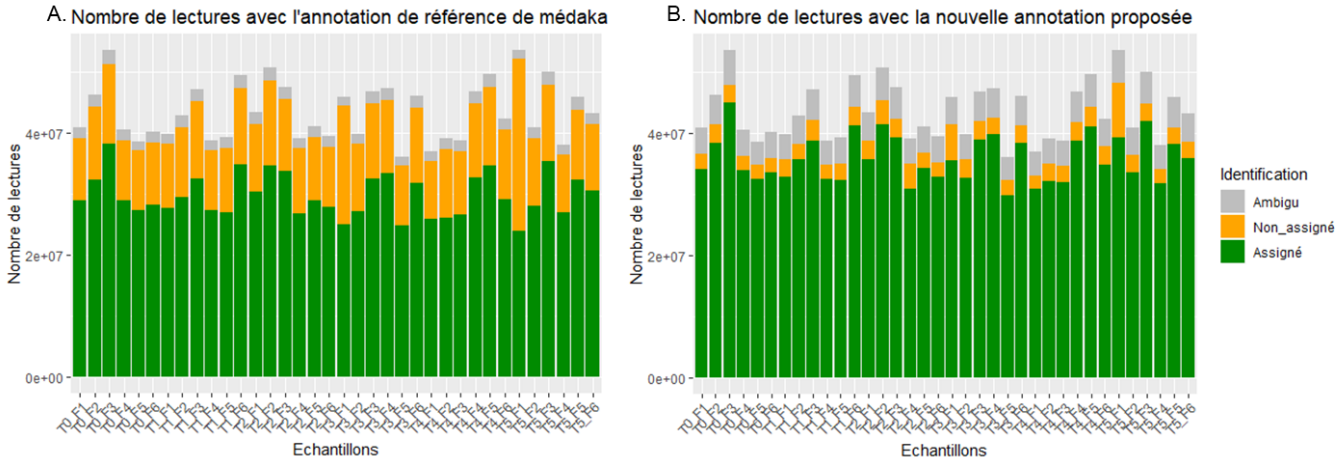


Figure 9: Nombre de lectures assignées et non assignées (Ambigu et Non assigné) à des gènes par échantillon avec l'outil de comptage Subread featureCounts. Les lectures sont classifiées en 3 catégories : celles assignés à un gène, celles ambiguës pouvant être assignées à différent gènes et ne sont assignés à aucun, et celles non assignées à un gène. **A. Comptage des lectures effectuées avec les annotations du génome de référence (v95.ENSEMBL). B. Comptage effectuées avec la nouvellement annotation.**

Quatre échantillons ont été retirés du jeu de donnée car leurs profils étaient très éloignés du reste des échantillons représentatifs de leur point de temps (Annexe 1 et 2 : MDS et clustering hiérarchique anormale). Une analyse différentielle de ces échantillons par rapport aux 31 autres suggère une expression anormale des gènes de la réponse immunitaire dans les quatre échantillons écartés.

Identification des gènes exprimés dans l'ovaire et annotation Gene Ontology

Après comptage du nombre de lectures par gènes et normalisation, nos données nous permettent d'identifier les gènes exprimés au cours de l'ovogénèse. Un total de 17150 gènes s'expriment au cours de l'ovogénèse *(Tableau 1).* Parmi ces gènes, 539 sont des transcrits d'ARN codants et 1131 sont des nouveaux transcrits d'ARN long non codant.

	31 samples
Filtering strategy	
Count per Million	1
Min. Nb. Replicate	4
Nb.Genes.Selected	17150
Differential Analysis	
Adjusted p.value (FDR)	0,05
Min. Fold-Change	-
Diff. Expressed Genes	2 412
New transcripts predicted long non-coding	69
New transcripts predicted coding	27

Tableau 1 : Analyse différentielle des gènes ovariens : Seuils et nombre de gènes par étape de filtrage et d'analyse différentielle. Le filtre utilisé pour sélectionné les gènes est de 1cpm dans au moins 4 échantillons par temps, 17150 gènes passe ces seuils. Les seuils utilisées pour l'analyse différentielle est une p-value ajustée de 0,05. 2412 gènes sont trouvés différentiellement exprimés entre au moins 2 temps dont 69 sont des transcrits de lcARNs prédits et 27 sont des transcrits d'ARN codant nouvellement prédits**.**

Une analyse des annotations fonctionnelles Gene Ontology de ces gènes ovariens a permis de mettre en évidence des processus cellulaires particulièrement actifs dans l'ovaire tel que des processus biologique impliqué dans la réplication de l'ADN et la division cellulaire (*DNA réplication, mitotic nuclear division, DNA transcription, cellular response to DNA damage stimulus, nucleocytoplasmic transport*), ainsi que des processus aboutissant à la dégradation des macromolécules (*cellulaire molecular catabolique process*). Les produits des gènes sont principalement dans le noyau, la lumière de la membrane, et la lumière des organelles intracellulaires. Les fonctions moléculaires enrichies sont des activités catalytiques sur l'ADN et l'ARN en accord avec le développement, ainsi que des activités de peptidase et ligase *(Figure 10A).*

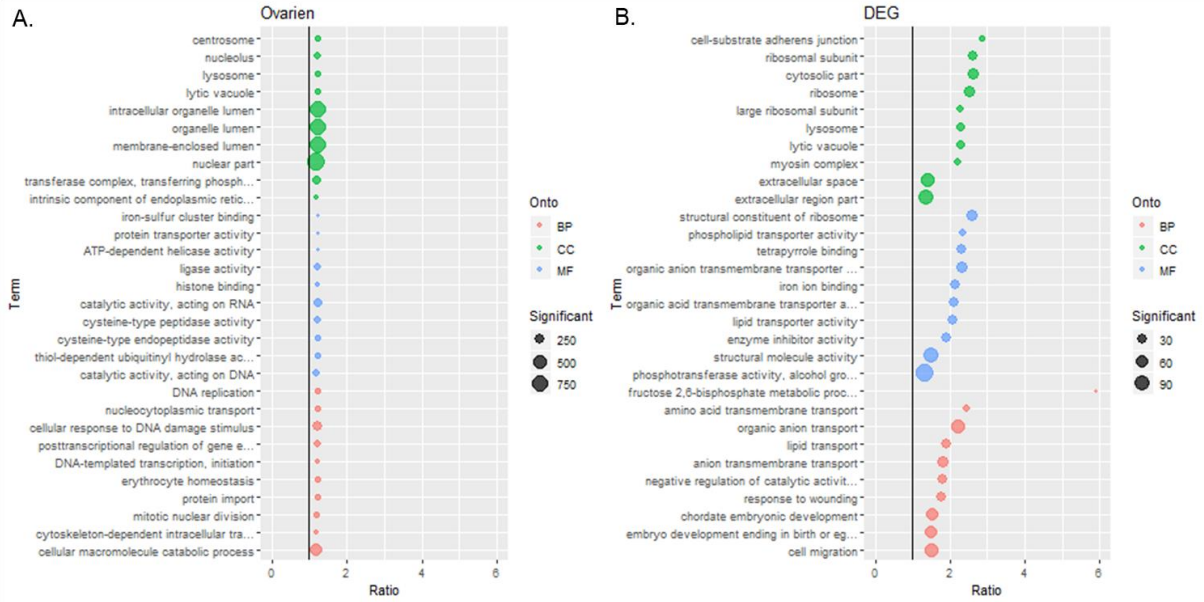


Figure 10 : Annotation Gene Ontology. A. Annotation GO des 17150 gènes ovariens. B. Annotation des 2112 gènes différentiellement exprimés au cours d'un cycle d'ovogénèse. Annotation des processus biologiques en rose, des composants cellulaires en bleu et des fonctions moléculaires en vert. La taille de significant représente le nombre de gènes dans ceux ovariens (A) et différentiellement exprimés (B) annoté par le terme correspondant. L'axe des abscisses représente le ratio entre le nombre de gènes significant, c'est-à-dire étant annoté par ce terme et le nombre de gènes étant annotés par ce

terme dans le génome total du médaka. Un seuil de p_value (test de Fisher) pour l'enrichissement des termes a été fixé à 0,05.

Identification des gènes différentiellement exprimés au cours de la cinétique dans l'ovaire et annotation Gene Ontology

L'analyse différentielle permet d'identifier les gènes différentiellement exprimés entre différentes conditions, dans notre cas entre différents temps au cours du cycle d'ovogénèse.

Pour les 15 contrastes étudiés, l'analyse identifie au total 2412 gènes différentiellement exprimés entre 2 temps du cycle d'ovogénèse. Entre les temps T0, T1 et T2, peu de gènes sont différentiellement exprimés, nous pouvons en déduire que ce sont en majorité les mêmes gènes qui s'expriment durant ces trois temps. Entre les temps T0 et T4, beaucoup de gènes sont différentiellement exprimés (1389). On constate que les temps T4 et T5 sont très différents des autres temps *(Figure 11).*

	T0	T1	T2	T3	T4	T5
T0		87	207	593	1389	726
T1			18	289	715	518
T2				167	550	502
T3					567	581
T4						536
T5						

Figure 11: Nombres de gènes différentiellement exprimés entres les 6 temps.

L'analyse des annotations GO des gènes différentiellement exprimés a permis d'identifier des processus biologiques enrichis comme des activités de migration cellulaire et des activités de transport d'anion, ainsi que des processus plus spécifique tel que le développement embryonnaire en accord avec la sélection de gènes évoluant au cours du temps pendant l'ovogénèse. Des activités de transports *(anion transport, lipid transport, cell migration)* sont enrichies dans les fonctions moléculaires pouvant être expliqué par la nécessité de mouvements et la maturation des cellules durant l'ovogénèse. Le milieu extracellulaire semble enrichi en produit des gènes différentiellement exprimés peut être expliqué par une interaction entre les cellules *(Figure 10B).*

Clustering des gènes différentiellement exprimés et annotations

La réalisation d'un clustering par la méthode PAM a permis de déterminer 11 clusters. La figure 12 montre la moyenne de l'expression des gènes par temps à gauche et l'expression des gènes par échantillons prélevé au cours du cycle d'ovogénèse à droite. Différents clusters présentent une expression forte de leurs gènes à un temps particulier.

Les clusters 1 et 2 regroupent des gènes s'exprimant fortement au temps T0, c'est-à-dire juste après la ponte. Le cluster 3 représente les gènes s'exprimant fortement au temps T0 et T5 ce qui fait penser à une expression cyclique. Les gènes du cluster 4 s'expriment en début de cycle (T0, T1, et T2) puis ne s'expriment plus. Pour le cluster 5, les gènes atteignent leur pic d'expression en T2, tandis que les gènes du cluster 6 l'atteignent en T3. Les clusters 7 et 8 sont fortement exprimés en T4, le niveau du cluster 8 étant bien plus élevé que celui du

cluster 7. Le cluster 9 regroupe des gènes de fin de cycle, fortement exprimés en T4 et T5. Les gènes du cluster 10 s'expriment fortement au temps T5, c'est-à-dire juste avant la ponte. Le cluster 11 regroupe les gènes s'exprimant particulièrement faiblement au temps T3.

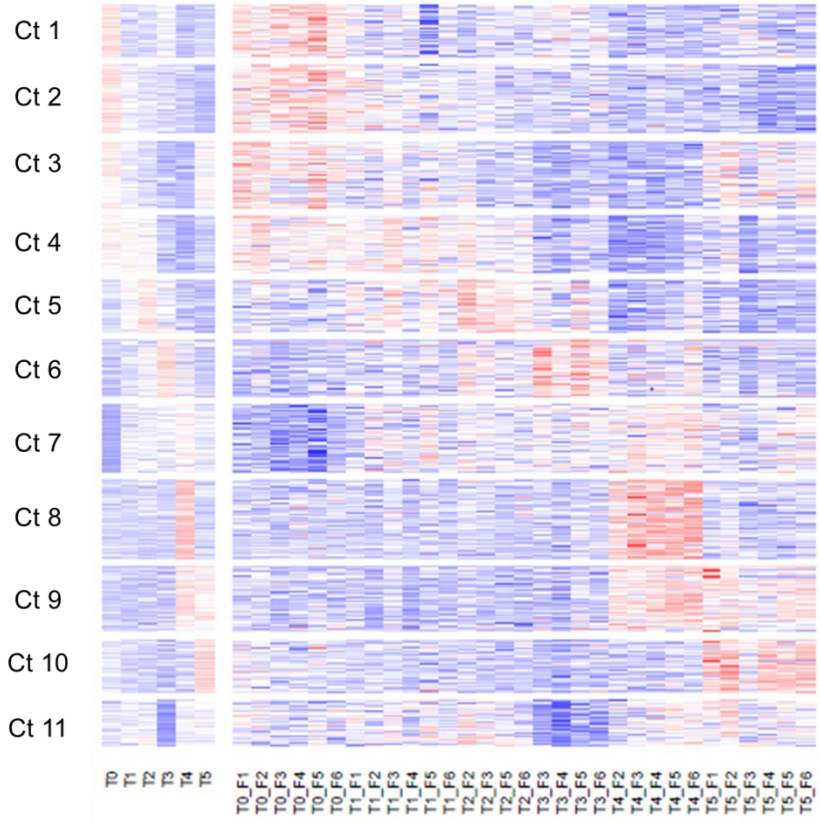


Figure 12: Clustering des gènes différentiellement exprimées suivant leur niveau d'expression au cours du cycle d'ovogénèse. Le clustering est constitué de 11 groupes (Ct1 à Ct11).

Identification des micro-ARNs germinaux et somatiques

De même que pour les ARNms, les miARNs ont un niveau d'expression qui varie selon le tissus et au cours du temps. L'analyse différentielle permet d'identifier les miARNs particulièrement exprimés dans les tissus germinaux (ovaire, testicule, œufs, stades embryonnaires 1 et 8 cellules, follicules) et ceux particulièrement présents dans les tissus non germinaux (yeux, cerveau, branchies, cœur, muscle, fois, rein, intestin, nageoire et stades embryonnaires 27, 31, 35 et 39 cellules).

457 miARNS sur 458 miARNs s'expriment assez pour pouvoir être analysé différentiellement. Cette analyse montre que 37 miARNs sont particulièrement exprimés dans les tissus germinaux (ARNs germinaux) et 197 dans les tissus non germinaux (ARNs somatiques) *(Tableau 2).*

	19 samples
Filtering strategy	
Count per Million	0,5
Min. Nb. Replicate	3
Nb.Genes.Selected	457
Differential Analysis	
Adjusted p.value (FDR)	0,05
Min. Fold-Change	-
Diff. Expressed Genes	234
Germinals	37
Somatics	197

Tableau 2: Analyse différentielle des miARNs : Seuils et nombre de gènes par étape de filtrage et d'analyse différentielle. Le filtre utilisé pour sélectionné les miARNs est de 0,5cpm dans au moins 3 échantillons par type de tissus (germinaux ou non germinaux), 457 gènes passe ces seuils. Le seuil de p-value ajusté utilisé est de 0,05. 234 gènes sont trouvés différentiellement exprimés entre les tissus germinaux et non germinaux dont 37 sont plus exprimés dans les tissus germinaux et 197 sont plus exprimés dans les tissus non germinaux.

Identification des cibles des miARNs différentiellement exprimés

Une identification des cibles des miARNs différentiellement exprimés (miARNs germinaux et somatiques) a été réalisée *via* TargetScan et miRanda. Seuls les gènes cibles prédites par les deux méthodes sont conservés.

Cette sélection a permis d'identifier 13000 gènes cibles des miARNs différentiellement exprimés dont 11722 sont cibles de miARNs trouvés différentiellement exprimés dans les tissus germinaux et 12945 sont cibles de miARNs somatiques.

Pour chaque cluster, le nombre de gènes cibles par miARN des miARNs germinaux (n=37) et somatiques (n =197) ont été calculé. La figure 13 montre, pour chaque cluster (Ct à Ct11), la moyenne des niveaux d'expression des gènes par temps (T0, T1, T2, T3 et T4) à gauche et une boite à moustache représentant le nombre de cibles par miARNs pour miARNs germinaux et somatiques à droite. Pour la majorité des clusters, Ct1, Ct2, Ct3, Ct4, Ct5, Ct8, Ct10 et Ct11, le nombre de gènes cible des miARNs germinaux, par cluster, est supérieur à celui des cibles des miARNs somatiques mais pas de manière statistiquement significative (test de wilcoxon). Concernant les clusters Ct6, Ct7 et Ct9, une différence significative (test de wilcoxon : $p<0,05$) entre le nombre de cible des miARNs germinaux et ceux des miARNs somatiques est détectée, les miARNs germinaux ont significativement plus de cibles dans ces clusters que les miARNs somatiques.

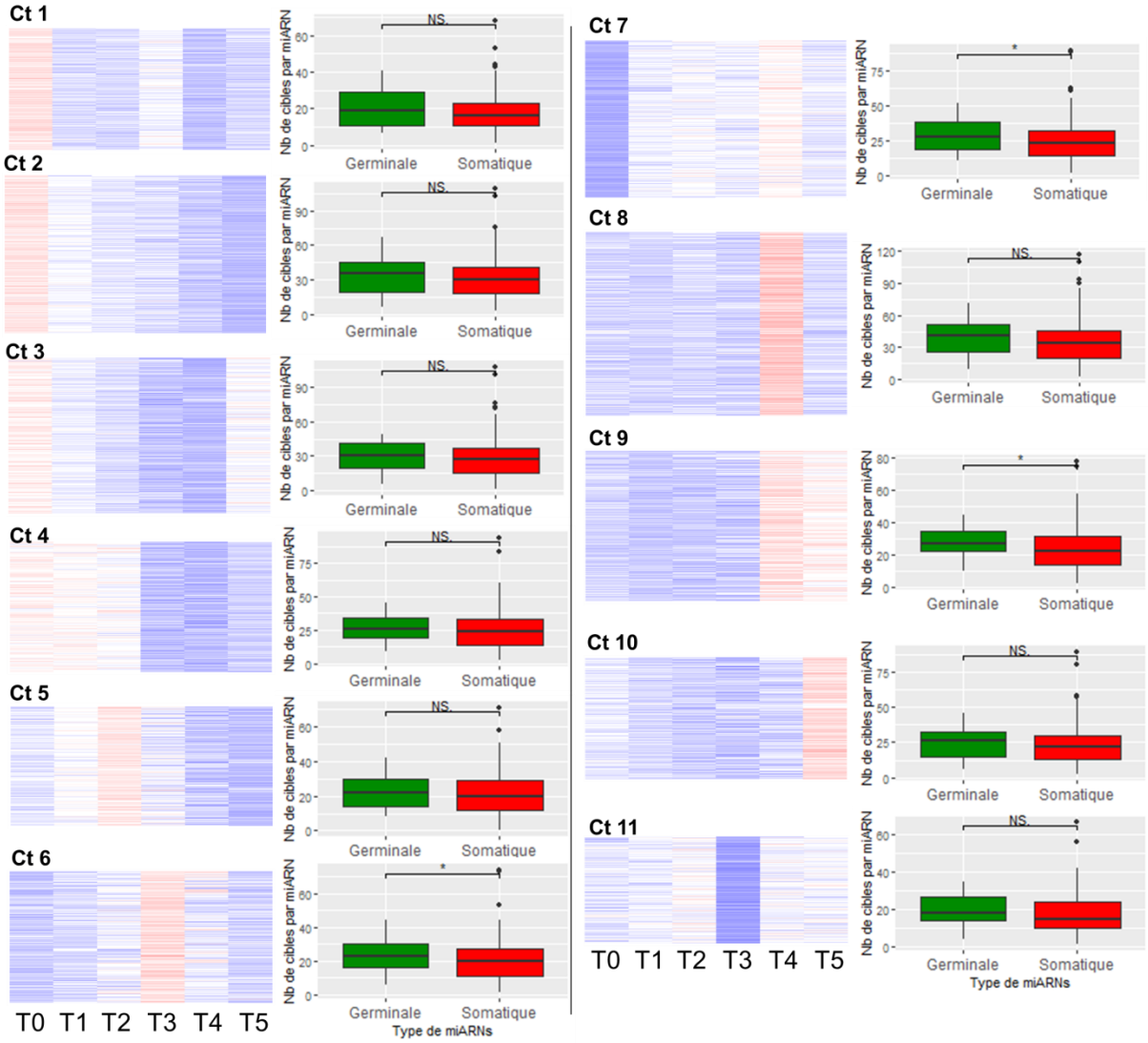


Figure 13 : Moyenne des niveaux d'expression des gènes par temps (gauche) et Nombre de cibles par miARNs (droite) pour les miARNs trouvés particulièrement exprimés dans les tissus germinaux (vert) et les miARNs trouvés particulièrement exprimés dans les tissus non germinaux (rouge) pour chaque cluster d'expression des gènes au cours du cycle (Ct1 à Ct11) *(Figure 12)*. Test de wilcoxon effectué par cluster entre les miARNs germinaux et somatiques (* p<0,05).

Clustering des gènes cibles différentiellement exprimés en fonction de leurs micro-ARNs germinaux régulateurs

Les 1476 des 2412 gènes présentant un profil d'expression différentiel au cours de l'ovogénèse sont des cibles potentielles des miARNs germinaux. Nous avons cherché à identifier des groupes de gènes ciblés par les mêmes miARNs. Pour cela, les 1476 gènes différentiellement exprimés cibles de miARNs ont été clusterisés en fonction des miARNs les ciblant. La réalisation d'un clustering des cibles des mi-ARNs germinaux par la méthode PAM a permis de déterminer 5 clusters. La figure 14 montre le classement des gènes en tant que cibles (vert) ou non-cibles (blanc) des miARNs germinaux (ax). Le cluster 1 (Cmi1) regroupe les gènes cibles de la majorité des miARNs germinaux et particulièrement ceux des miARNs a20, a310 et a314 mais pas par les a20, a310 et a314 qui sont des marqueurs de Cmi1. Les gènes du cluster 2 (Cmi2) sont cibles de peu de miARNs germinaux. Le cluster 3

(Cmi3) regroupe les gènes ciblés particulièrement par les miARNs a180, a181, a184 et a185. Le cluster 4 (Cmi4) regroupe les gènes cibles de miARNs germinaux mais sans être cible de miARNs préférentiels. Les gènes cibles des miARNs a420 et a422 sont regroupés dans le cluster 5 (Cmi5).

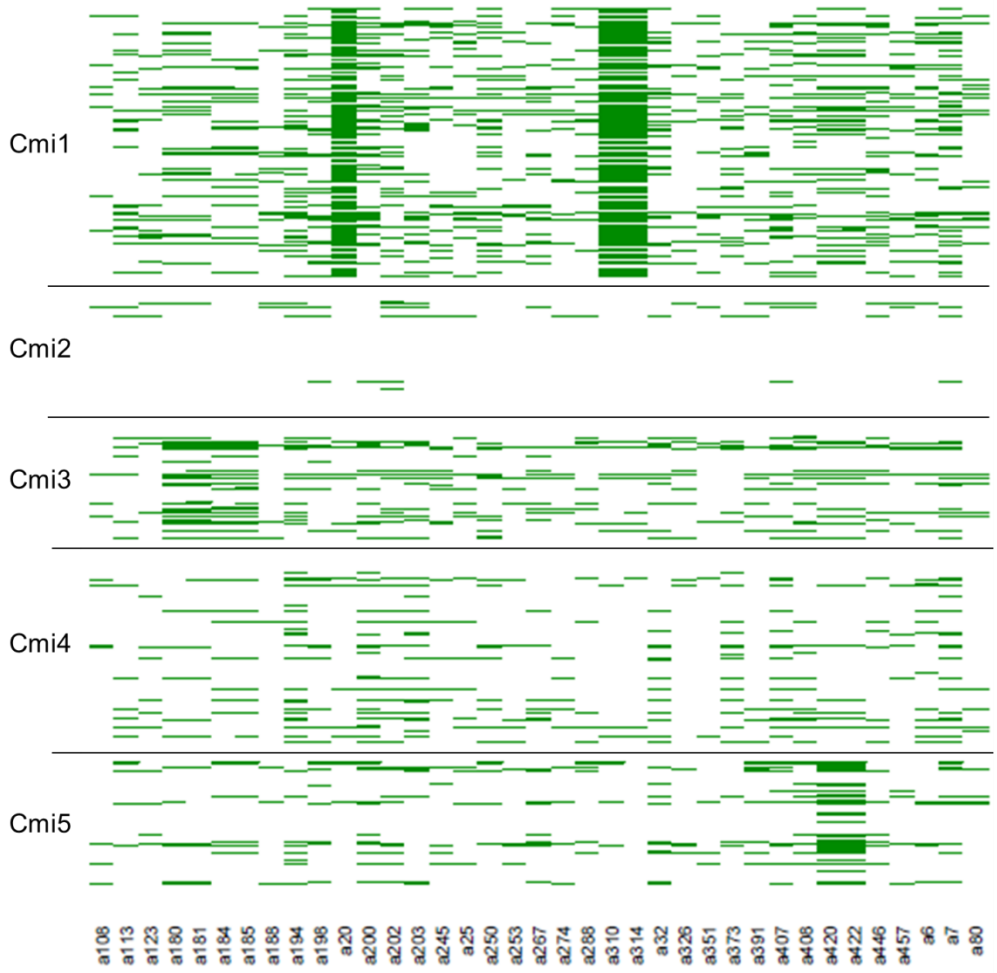


Figure 14 : Clustering des cibles de miARNs germinaux en fonction de leur miARN régulateur. 5 clusters ont été trouvé (Cmi 1 à Cmi 5).

Correspondance entre le clustering des gènes différentiellement exprimés au cours du cycle d'ovogénèse et le clustering des cibles des miARNs germinaux

Pour lier les résultats obtenus lors de l'analyse différentielle et du clustering (Ct1 à Ct11) de l'expression des gènes différentiellement exprimés au cours du temps (T0 à T5) avec ceux du clustering (Cmi1 à Cmi5) des gènes différentiellement exprimés et cibles de miARNs germinaux, une analyse du nombre de gènes se retrouvant au croisement de chaque cluster a été réalisée.

La figure 15 montre une importante correspondance entre les gènes du cluster Ct2, forte expression des gènes en T0, et Ct3, forte expression des gènes en T0 et T5, avec le cluster Cmi1, gènes ciblés par la majorité des miARNS germinaux, avec plus de 60 gènes trouvé identiques dans ces clusters. Les gènes du cluster Cmi2, cibles de peu de miARN germinaux,

se retrouvent dans les clusters Ct2 et Ct3, dont leurs gènes sont fortement exprimés à T1 et dans les clusters Ct6 à Ct9 dont l'expression de leurs gènes est tardive au cours du cycle d'ovogénèse (T3, T4 et T5). Cela fait penser à une régulation cyclique. Les gènes du cluster Cmi3 sont répartis dans les différents clusters de la cinétique d'ovogénèse. Le cluster Cmi4, regroupant les gènes cibles de miARNs germinaux mais sans être cible de miARNs préférentiels, a des gènes en commun avec le cluster Ct2, dont les gènes s'expriment fortement en T0 et Ct7 et Ct8, dont les gènes s'expriment fortement en T4. Les clusters Ct3, dont les gènes s'expriment fortement en T0 et T5, et Cmi5, regroupant particulièrement les gènes cibles de a20 et a22, ont de nombreux gènes en commun, 40.

	Cmi1	Cmi2	Cmi3	Cmi4	Cmi5
Ct1	40	12	9	19	13
Ct2	62	22	22	41	15
Ct3	42	29	16	33	40
Ct4	41	19	22	24	24
Ct5	31	9	16	34	21
Ct6	39	27	16	28	14
Ct7	47	30	18	48	19
Ct8	65	24	26	39	33
Ct9	43	23	14	25	18
Ct10	44	19	16	32	11
Ct11	31	18	12	22	19

Figure 15 : Identification du nombre de gènes différentiellement exprimés à la fois dans les clusters (Ct1 à Ct11) en fonction de l'expression des gènes au cours du cycle ovarien (T0 à T5) et à la fois dans les clusters des cibles des miARNs germinaux.

Discussion et Conclusion

Nouvel assemblage

L'alignement des lectures sur le génome de référence est de bonne qualité puisque 86% des lectures ont été alignés sur le génome de référence.

L'identification de nouveaux transcrits dans le génome de medaka a permis d'identifier 529 nouveaux transcrits d'ARNs codants et 1131 d'ARNs non codants. Ces nouvelles annotations ont été ajoutées à celles du génome de référence.

L'identification de nouvelles annotations, en plus de compléter les informations connues sur le génome du médaka, permet d'augmenter, lors de l'étape de comptage, de 68% à 82% le nombre de lectures assignées à des exons. Le poisson medaka est particulièrement étudié dans le domaine de la reproduction et son annotation n'est pas complète. Compléter l'annotation du génome de référence en prédisant de nouveaux ARNms et de nouveaux lncARNs permet d'assigner plus des lectures à leur gène. De plus, ajouter ces nouvelles annotations aux annotations du génome de référence permet d'étudier de manière plus précise le génome et particulièrement les lncARNs, ayant eux aussi une fonction régulatrice.

Pour continuer d'étudier les nouveaux transcrits et mieux les annotés, il pourrait être intéressant de vérifier leurs présences dans les autres tissus de médaka. Particulièrement pour les nouveaux ARNms, une caractérisation de leur rôle par homologie de séquence avec d'autres espèces comme le poisson zèbre peut être réalisés à l'aide de BLAST (Basic Local Alignment Search Tool).

Les gènes différentiellement exprimés dans l'ovaire

Identification des gènes différentiellement exprimés

L'analyse différentielle a permis identifier 17150 gènes dans l'ovaire dont 2412 sont différentiellement exprimés, cela montre bien que tous les gènes du génome ne sont pas impliqués dans l'ovogénèse, du moins ne varient pas au cours d'un cycle d'ovogénèse. D'autre part, 69 nouveaux transcrits long non codants et 27 nouveaux transcrits codants sont aussi trouvés différentiellement exprimés parmi les 2412 gènes différentiellement exprimés. 14% des gènes de l'ovaire montre un profil particulier d'expression au cours du cycle, cela suggère que de nombreux processus sont mobilisés au cours de l'ovogénèse.

Cependant, bien que l'analyse des annotations Gene Ontology montre que ce ne sont pas les même processus trouvés enrichis dans les gènes ovariens et dans les gènes différentiellement exprimés, nous n'observons pas de processus très précis. Cela peut être expliqué par la très faible couverture des annotations GO pour le génome de médaka comparé à d'autre génome comme celui de la souris. De plus, une importante partie des annotations est certainement

transposée par homologie des gènes annotés du poisson zèbre, expliquant les annotations très généraliste obtenues.

Cependant il ne faut pas oublier que les gènes différentiellement exprimés proviennent d'ovaire, un organe complet comportant de nombreux follicules constitués de nombreux types de cellules à différent stade de maturation. Pour avoir une vision plus précise des mécanismes sous-jacents et établir le profil d'expression des gènes par type cellulaires, des méthodes plus poussées incluant la spectrométrie de masse ou le single cell pourraient être utilisées. Deux méthodes pourraient être mise en place, une séparation des différents types cellulaires par spectrométrie de masse après extraction de l'ovaire suivie d'une RNA-seq par type cellulaire, ou une analyse RNA-seq single cell permettant d'étudier le transcriptome par cellule, ce qui permettrait même d'analyser l'évolution du transcriptome des cellules de follicules au cours du temps.

Classification des gènes différentiellement exprimés au cours du temps

Le nombre de gènes trouvés différentiellement exprimés n'est pas le même entre chacune des comparaisons des temps. Les temps T0, T1 et T2 sont proches et T4 est très différents des autres temps, bien qu'ils soient plus proche en terme d'expression des gènes de T5 que des autres temps ce qu'on a pu voir lors de la classification des échantillons (Annexe 1 et 2). Les T0, T1, T2 et T3 présentent des processus graduels. Alors que T4 et T5 présente des profils très différents des autres. Ces résultats suggèrent que l'expression des gènes change peu au début du cycle d'ovogénèse et qu'un mécanisme se met en place entre le temps T3 et T4.

Le clustering des gènes différentiellement exprimés, en fonction de leur expression au cours du temps, a permis de d'identifier 11 clusters. Ces clusters montrent la variation de l'expression des gènes au cours du temps. Le cluster Ct1 regroupe les gènes qui auraient un rôle dans le début du cycle d'ovogénèse. Les résultats nous montrent des patterns cyclique de l'expression des gènes pour le cluster Ct2 et Ct3, cela suggèrent que les gènes du cluster Ct2 sont impliqué dans les premiers stades de l'ovogénèse et que leur action diminue au cours du temps. Les gènes du cluster Ct3 seraient utiles au début du cycle et leurs rôles diminuerait au cours du temps jusqu'à la phase ovulatoire où ils pourraient jouer un rôle dans la mise en place d'un nouveau cycle d'ovogénèse. Les gènes du cluster Ct4 serait impliqué uniquement durant le début du cycle mais n'aurait pas de rôle à partir du temps T3 jusqu'à la fin du cycle. Les gènes du cluster Ct5 et Ct6 joueraient respectivement des rôles aux temps T2 et T3. Les gènes du cluster Ct7 ne seraient pas impliqués dans la mise en place précoce lors d'un cycle d'ovogénèse mais jouerait des rôles au cours du reste du cycle. De même pour le cluster Ct11 dont ces gènes joueraient des rôles tous le long du cycle sauf au temps T3. Les gènes du cluster Ct8 et Ct10 seraient impliqué respectivement aux temps T4 et T5, alors que les gènes du cluster Ct9 joueraient des rôles durant toute la fin du cycle d'ovogénèse. Ces listes de gènes par cluster vont permettre aux biologistes de réaliser de plus ample expérience concernant ces gènes.

Pour continuer notre étude, une recherche des motifs des promoteurs et facteurs de transcriptions des gènes pourrait être faite en complément. Cette analyse complémentaire pourrait potentiellement faire ressortir des groupes de gènes étant sous le contrôle de même facteurs de transcription, ce qui pourrait expliquer leur expression différentielle au cours du temps.

Les miARNs s'expriment différemment dans les tissus

L'analyse différentielle a permis d'identifier 37 miARNs particulièrement exprimés dans les tissus germinaux (miARNs germinaux) et 197 miARNS particulièrement exprimés dans les tissus non germinaux (miARNs somatiques). Ce résultat suggère que des miARNs exprimés plus spécifiquement dans les tissus germinaux pourraient jouer un rôle dans l'ovogénèse en régulant la traduction de gènes impliqués dans ce processus.

Une analyse du nombre de gènes cibles des miARNs germinaux et des miARNs somatiques parmi les gènes différentiellement exprimés indique que 1476 gènes sur les 2412 gènes DEG sont ciblés par au moins un miARN germinal.

On observe que les gènes différentiellement exprimé au cours du cycle d'ovogénèse sont en moyenne plus ciblés par les miARNs germinaux que somatiques. Cependant, cette tendance n'est pas statistiquement significative, sauf dans les clusters Ct6, Ct7 et Ct9 qui regroupent les gènes s'exprimant plus en fin de cycle d'ovogénèse. Cette observation sur les clusters Ct6, Ct7 et Ct9 pourraient suggérer une régulation plus importante de ces gènes pas des miARNs germinaux au début du cycle d'ovogénèse, puisque les miARNs jouent le rôle de répresseur de la traduction ou dégradent leurs gènes. Pour aller plus loin dans cette analyse, il serait intéressant de coupler les données d'expression des gènes avec une cinétique d'expression des miARNs pendant l'ovogénèse, afin de corréler les niveaux des gènes cibles avec ceux des miARNs régulateurs potentiels.

Cependant, l'identification des cibles par analyse bioinformatique induit de nombreux faux positifs (Mockly et Seitz 2019). Bien que nous ayons cherché à limiter ce nombre de faux positif en utilisant dans notre étude l'intersection des résultats de TargetScan et de miRanda, nous nous somme interrogées sur la pertinence d'ajouter des seuils plus stricts pour prédire et sélectionner les cibles, en sélectionnant les scores de prédiction de meilleure qualité. Cette hypothèse est actuellement à l'étude, mais son intérêt semble limité par le manque de fiabilité des scores proposés par TargetScan et miRanda.

Le clustering des gènes différentiellement exprimés en fonction des miARNs dont ils sont cibles a montré des résultats intéressants. En effet, certains groupes de gènes sont ciblés par de nombreux miARNs germinaux alors que d'autres sont cibles de peu de miARNs. Le clustering regroupe aussi des gènes spécifiquement ciblés par certains miARNs. Les gènes du

cluster Cmi1 sont préférentiellement ciblés par les miARNs a20, a310 et a314. Le cluster Cmi3 réunie les gènes ciblés particulièrement pas les miARNs a180, a181, a184 et a185. Le cluster Cmi5 regroupe les gènes ciblés par les miARNS germinaux a420 et a422. Concernant ces miARNs ciblant les mêmes groupes de gènes, on peut se demander si leurs graines, servant à l'hybridation, ne seraient pas similaires ou très ressemblant.

Lien entre l'expression des gènes et leur régulation par les miARNs

Nos résultats montrent que des gènes montrant un profil d'expression différentiel au cours du cycle d'ovogénèse pourraient être reliés à une régulation spécifique des miARNs particulièrement exprimés dans les tissus germinaux. Pour aller plus loin dans notre étude, les deux clusterings effectués ont été comparés. Un comptage des gènes présents à la fois dans les clusters déterminés à partir de leur expression différentielle au cours du temps (Ct1 à Ct11) et dans les clusters déterminés à partir de leurs miARNs régulateurs potentiels (Cmi1 à Cmi5) a été effectué.

On peut constater que les gènes sont ciblés tous un large spectre de miARNs et plus particulièrement les miARNs a20, a310 et a314 (Cmi1) se retrouvent dans tous les clusters en particulier Ct2 et Ct8, dont les gènes s'expriment fortement à T0 et T4. Ce qui suggèrent que la régulation de ces gènes par les miARNs s'effectuerait entre T1 et T3 et à T5. Les gènes régulés par peu de miARNs (Cmi2) s'expriment particulièrement au temps T4 et T5. On peut émettre l'hypothèse d'une régulation des miARNs pendant la première phase du cycle d'ovogénèse. Des gènes particulièrement régulés par a180, a181, a184 et a185 (Cmi3) se retrouvent dans tous les clusters de la cinétique d'ovogénèse. Leur présence diffuse dans tous les clusters ne semble par cohérente avec un mode de régulation dépendant de la présence d'un miARN régulée au cours du temps. Les gènes étant cibles de miARNs de manière non préférentielle sont exprimés dans la majorité des clusters de la cinétique d'ovogénèse mais plus particulièrement dans ceux dont les gènes s'expriment au temps T0 et T4. On peut émettre l'hypothèse d'une régulation cyclique par les miARNs avec une plus faible régulation au temps T0 et T4. Les gènes particulièrement ciblés par a420 et a422 (Cmi5) s'expriment préférentiellement aux temps T4 et T5 ou T0. Cela pourraient suggérer un rôle répresseur de a420 et a422 aux temps T1, T2 et T3, c'est-à-dire plutôt pendant la première phase du cycle d'ovogénèse.

Tous ces résultats nécessitent une confirmation biologique par l'expérimentation pour vérifier si les cibles déterminés bioinformatiquement le sont bien biologiquement. De plus une analyse de l'expression des miARNs au cours d'un cycle d'ovogénèse pourrait venir renforcer nos résultats et compléter nos connaissances sur le mode d'action des miARNs. Cependant les miARNs ont deux modes de régulations, l'un état l'inhibition de la traduction sans dégradation des ARNms cibles, le rôle régulateur des miARNs pourraient ne pas être visible avec l'analyse du transcriptome.

Pour compléter cette étude, les travaux en cours s'orientent vers une identification des cibles des lncARNs s'exprimant dans l'ovaire et découverts au cours de mon stage. Plus

particulièrement, une identification des cibles des lncARNS différentiellement exprimés au cours de l'ovogénèse (n=69) sera réalisé à l'aide de lncTar (Li et al. 2015). L'idée sous-jacente est de rechercher des cibles de lncARNs dont l'expression est inversement corrélée à l'expression de son lncARN régulateurs.

D'autre part, des études au sein du LPGP ont permis de s'intéresser plus particulièrement à miR-202 (a80) (Gay et al. 2018). Ils ont montré que le knock-out de miR-202 induit une diminution de la fréquence de ponte, du nombre d'œufs et de leur qualité suggérant un rôle primordial de ce miR-202 dans la fécondité et la production des ovocytes. Pour identifier les mécanismes et gènes ciblés par miR-202, une cinétique RNA-seq sur des ovaires du KO au cours d'un cycle d'ovogénèse pourrait être intéressante, et sa comparaison avec la cinétique étudiée chez le médaka sauvage et présentée dans ce rapport serait pertinente. Toutefois, des résultats préliminaires par qPCR n'ont montré aucune variation significative de l'expression de miR-202 au cours du cycle d'ovogénèse, ce qui pourrait indiquer une régulation indirecte faisant intervenir un autre partenaire ou une régulation plus précoce lors du développement de l'ovaire. Pour vérifier la seconde hypothèse, une analyse en amont de la période de reproduction pourrait être réalisée pour étudier les mécanismes impliqués dans le développement de l'ovaire.

Bibliographie

Alberti, Chiara, et Luisa Cochella. 2017. « A Framework for Understanding the Roles of MiRNAs in Animal Development ». *Development* 144 (14): 2548-59. https://doi.org/10.1242/dev.146613.

Alexa, A., J. Rahnenfuhrer, et T. Lengauer. 2006. « Improved Scoring of Functional Groups from Gene Expression Data by Decorrelating GO Graph Structure ». *Bioinformatics* 22 (13): 1600-1607. https://doi.org/10.1093/bioinformatics/btl140.

Bouchareb, Amine, Aurélie Le Cam, Jérôme Montfort, Stéphanie Gay, Thaovi Nguyen, Julien Bobe, et Violette Thermes. 2017. « Genome-Wide Identification of Novel Ovarian-Predominant MiRNAs: New Insights from the Medaka (Oryzias Latipes) ». *Scientific Reports* 7 (1). https://doi.org/10.1038/srep40241.

Brennecke, Julius, Alexander Stark, Robert B Russell, et Stephen M Cohen. 2005. « Principles of MicroRNA–Target Recognition ». *PLoS Biology* 3 (3). https://doi.org/10.1371/journal.pbio.0030085.

Chen, Jian, Tanxi Cai, Chunwei Zheng, Xiwen Lin, Guojun Wang, Shangying Liao, Xiuxia Wang, et al. 2017. « MicroRNA-202 Maintains Spermatogonial Stem Cells by Inhibiting Cell Cycle Regulators and RNA Binding Proteins ». *Nucleic Acids Research* 45 (7): 4142-57. https://doi.org/10.1093/nar/gkw1287.

Dabaja, Ali A, Anna Mielnik, Brian D Robinson, Matthew S Wosnitzer, Peter N Schlegel, et Darius A Paduch. 2015. « Possible germ cell-Sertoli cell interactions are critical for establishing appropriate expression levels for the Sertoli cell-specific MicroRNA, miR-202-5p, in human testis ». *Basic and Clinical Andrology* 25 (mars). https://doi.org/10.1186/s12610-015-0018-z.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, et Thomas R. Gingeras. 2013. « STAR: Ultrafast Universal RNA-Seq Aligner ». *Bioinformatics* 29 (1): 15-21. https://doi.org/10.1093/bioinformatics/bts635.

Enright, Anton J, Bino John, Ulrike Gaul, Thomas Tuschl, Chris Sander, et Debora S Marks. 2003. « MicroRNA Targets in Drosophila ». *Genome Biology*, 14.

Gay, Stéphanie, Jérôme Bugeon, Amine Bouchareb, Laure Henry, Clara Delahaye, Fabrice Legeai, Jérôme Montfort, et al. 2018. « MiR-202 Controls Female Fecundity by Regulating Medaka Oogenesis ». Édité par Manfred Schartl. *PLOS Genetics* 14 (9): e1007593. https://doi.org/10.1371/journal.pgen.1007593.

Grimson, Andrew, Kyle Kai-How Farh, Wendy K Johnston, Philip Garrett, Lee P Lim, et David P Bartel. 2013. « MicroRNA Targeting Specificity in Mammals: Determinants Beyond Seed Pairing », 28.

Iwamatsu, Takashi. 2004. « Stages of Normal Development in the Medaka Oryzias Latipesq ». *Mechanisms of Development*, 14.

Kaufman, Leonard, et Peter J. Rousseeuw. 1987. « Clustering By Means of Medoids », 1987.

Lewis, Benjamin P., Christopher B. Burge, et David P. Bartel. 2005. « Conserved Seed Pairing, Often Flanked by Adenosines, Indicates That Thousands of Human Genes Are MicroRNA Targets ». *Cell* 120 (1): 15-20. https://doi.org/10.1016/j.cell.2004.12.035.

Li, Jianwei, Wei Ma, Pan Zeng, Junyi Wang, Bin Geng, Jichun Yang, et Qinghua Cui. 2015. « LncTar: A Tool for Predicting the RNA Targets of Long Noncoding RNAs ». *Briefings in Bioinformatics* 16 (5): 806-12. https://doi.org/10.1093/bib/bbu048.

Liao, Y., G. K. Smyth, et W. Shi. 2014. « FeatureCounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features ». *Bioinformatics* 30 (7): 923-30. https://doi.org/10.1093/bioinformatics/btt656.

Lubzens, Esther, Graham Young, Julien Bobe, et Joan Cerdà. 2010. « Oogenesis in Teleosts: How Fish Eggs Are Formed ». *General and Comparative Endocrinology* 165 (3): 367-89. https://doi.org/10.1016/j.ygcen.2009.05.022.

Mockly, Sophie, et Hervé Seitz. 2019. « Inconsistencies and Limitations of Current MicroRNA Target Identification Methods ». In *MicroRNA Target Identification*, édité par Alessandro Laganà, 1970:291-314. New York, NY: Springer New York. https://doi.org/10.1007/978-1-4939-9207-2_16.

Otsuka, Motoyuki, Min Zheng, Masaaki Hayashi, Jing-Dwan Lee, Osamu Yoshino, Shengcai Lin, et Jiahuai Han. 2008. « Impaired MicroRNA Processing Causes Corpus Luteum Insufficiency and Infertility in Mice ». *Journal of Clinical Investigation* 118 (5): 1944-54. https://doi.org/10.1172/JCI33680.

Pertea, Mihaela, Geo M Pertea, Corina M Antonescu, Tsung-Cheng Chang, Joshua T Mendell, et Steven L Salzberg. 2015. « StringTie Enables Improved Reconstruction of a Transcriptome from RNA-Seq Reads ». *Nature Biotechnology* 33 (3): 290-95. https://doi.org/10.1038/nbt.3122.

Presslauer, Christopher, Teshome Tilahun Bizuayehu, Martina Kopp, Jorge M. O. Fernandes, et Igor Babiak. 2017. « Dynamics of miRNA transcriptome during gonadal development of zebrafish ». *Scientific Reports* 7 (mars). https://doi.org/10.1038/srep43850.

Robinson, Mark D., Davis J. McCarthy, et Gordon K. Smyth. 2010. « edgeR: a Bioconductor package for differential expression analysis of digital gene expression data ». *Bioinformatics* 26 (1): 139-40. https://doi.org/10.1093/bioinformatics/btp616.

Romano, Giulia, Dario Veneziano, Mario Acunzo, et Carlo M. Croce. 2017. « Small Non-Coding RNA and Cancer ». *Carcinogenesis* 38 (5): 485-91. https://doi.org/10.1093/carcin/bgx026.

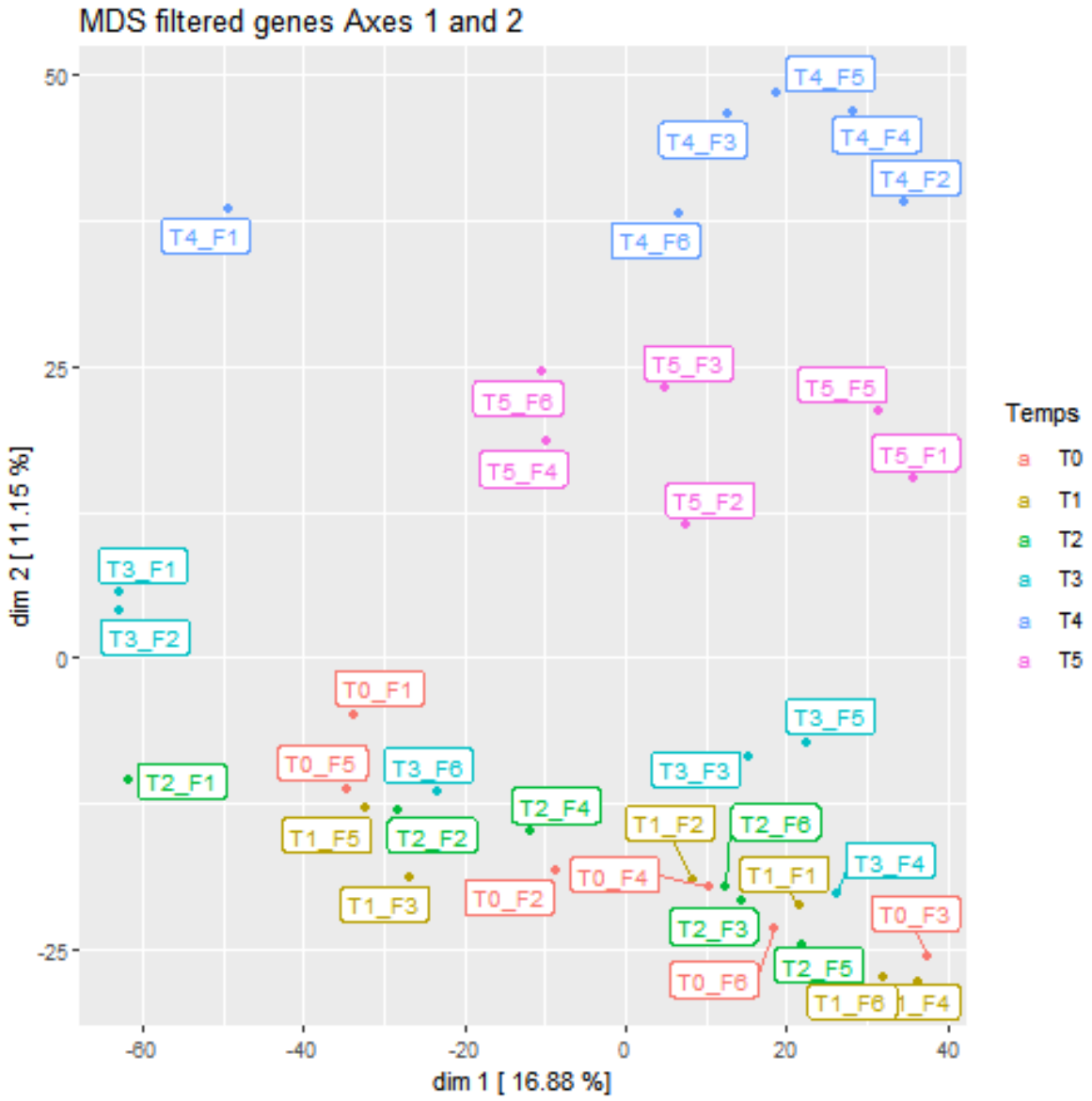
Wainwright, Elanor N., Joan S. Jorgensen, Youngha Kim, Vy Truong, Stefan Bagheri-Fam, Tara Davidson, Terje Svingen, et al. 2013. « SOX9 Regulates MicroRNA MiR-202-5p/3p Expression during Mouse Testis Differentiation ». *Biology of Reproduction* 89 (2): 34. https://doi.org/10.1095/biolreprod.113.110155.

Wang, Ti-Tai, Chien-Yueh Lee, Liang-Chuan Lai, Mong-Hsun Tsai, Tzu-Pin Lu, et Eric Y. Chuang. 2019. « AnamiR: Integrated Analysis of MicroRNA and Gene Expression Profiling ». *BMC Bioinformatics* 20 (1). https://doi.org/10.1186/s12859-019-2870-x.

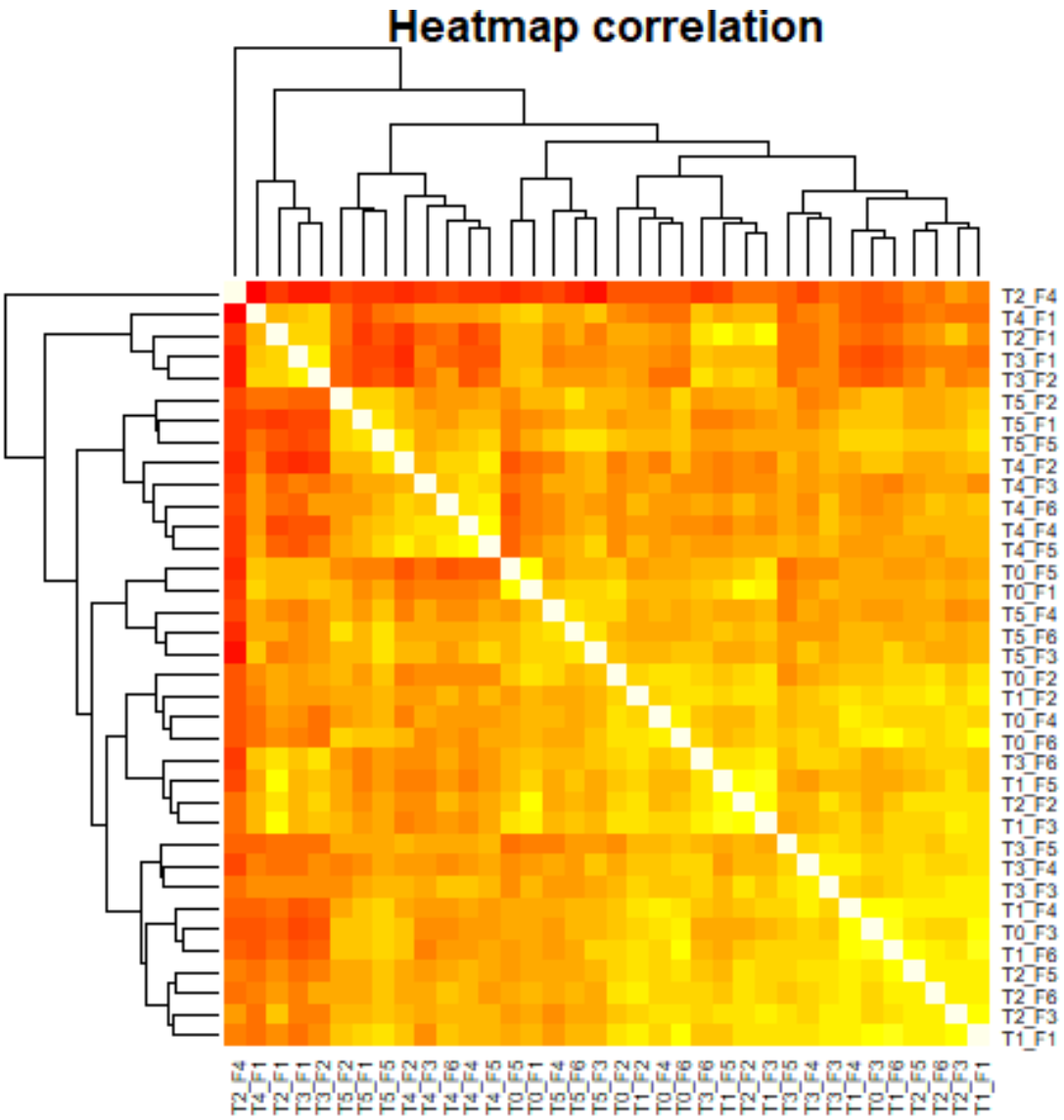
Wucher, Valentin, Fabrice Legeai, Tosso Leeb, Vidhya Jagannathan, Edouard Cadieu, Audrey David, Hannes Lohi, et al. 2017. « FEELnc: A Tool for Long Non-Coding RNA Annotation and Its Application to the Dog Transcriptome ». *Nucleic Acids Research* 45 (8): 12.

https://github.com/askomics/askoR

Annexes



Annexe 1: MDS des 36 échantillons (6 ovaires par temps T0, T1, T2, T3, T4 et T5) prélevés au cours du cycle d'ovogénèse. Les échantillons du temps T4 se regroupent bien, sauf T4_F1. Les échantillons du temps T5 semble former un groupe. Les échantillons des autres temps semblent être réunis entre eux.



Annexe 2: Heatmap et classification hiérarchique des 36 échantillons (6 ovaires par temps T0, T1, T2, T3, T4 et T5) prélevés au cours du cycle d'ovogénèse. L'échantillon T2_F4 n'est pas classifié avec les autres échantillons. Les échantillons T2_F1, T3_F1, T3_F2 et T4_F1 semblent avoir une expression différente des autres temps qui se regroupent.