CLONAGE DE GÈNES DE PETITS VERTÉBRÉS SUSCEPTIBLES DE VOIR LEUR EXPRESSION INDUITE PAR DES PESTICIDES ENVIRONNEMENTAUX ET SÉQUENÇAGE ET ASSEMBLAGE DU GÉNOME DE L'HIRONDELLE BICOLORE (*TACHYCINETA BICOLOR)*

par

Kathy Doyon

mémoire présenté au Département de biologie en vue
de l'obtention du grade de maître ès sciences (M.Sc.)

Faculté des sciences
Université de Sherbrooke

Sherbrooke, Québec, Canada, mai 2015

Le 21 mai 2015

le jury a accepté le mémoire de Madame Kathy Doyon
dans sa version finale.

Membres du jury

Professeur Luc Gaudreau
Directeur de recherche
Département de biologie

Professeur Fanie Pelletier
Codirectrice de recherche
Département de biologie

Professeur Benoit Leblanc
Évaluateur interne
Département de biologie

Professeur Pierre-Étienne Jacques
Président-rapporteur
Département de biologie

SOMMAIRE

Environ 3500 tonnes de pesticides sont étendues chaque année sur les terres agricoles du Québec. L'utilisation de plusieurs de ces substances a été interdite, car les ingrédients actifs qui les composaient avaient des effets toxiques sur les humains et/ou l'environnement. Malheureusement, certains qui sont toujours en vente ont aussi des effets secondaires non désirables. En effet, ces molécules exogènes ont le potentiel de moduler l'activation de protéines régulatrices comme le récepteur aux dioxines (AhR). AhR active, entre autres, l'expression de gènes faisant partie de la famille du cytochrome P450 (*CYP1A1* et *CYP1B1)* qui sont impliqués dans la détoxification. Or, dans certains cas, ces enzymes mènent à la production de molécules mutagènes en augmentant la toxicité des ingrédients actifs en les métabolisant.

Le projet global dans lequel s'insère ce projet de maîtrise vise à déterminer les effets génomiques des pesticides environnementaux sur des organismes vivants en milieu naturel dans les environs de la région de l'Estrie. Les espèces choisies comme modèles d'étude sont des insectivores, car les pesticides s'accumulent dans les lipides et que les insectes en sont une excellente source. Les consommateurs d'insectes sont donc d'excellents marqueurs du niveau de contamination de leur environnement par les pesticides. Les deux espèces sélectionnées sont l'hirondelle bicolore (*Tachycineta bicolor*) et la grande musaraigne (*Blarina brevicauda*).

De façon plus spécifique, le projet de maîtrise se divise en deux principaux objectifs. Le premier objectif est de valider que les pesticides présents dans l'environnement sont en concentration suffisante pour modifier la régulation génétique d'animaux en milieu naturel.

Cette validation se fera en comparant le taux d'expression de *CYP1A1* et *CYP1B1* (suite de leur activation par AhR) chez des bêtes ayant été exposées (ou non) à des pesticides. Comme le génome des deux modèles d'étude n'est pas encore séquencé, le clonage partiel des gènes à étudier (*AhR* et *CYP1*) a été entamé de même que la conception d'amorces qui permettra de quantifier le niveau d'expression de ces gènes par des réactions en chaîne par polymérase en temps réel (qPCR). À ce jour, une partie de la séquence d'*AhR*, de *CYP1B1* et de trois gènes contrôles ont été séquencés pour l'hirondelle, ce qui a permis de concevoir des amorces pour la quantification de l'expression d'*AhR* et de deux contrôles. Pour la musaraigne, ce sont les gènes *AhR*, potentiellement *CYP1A1* et trois contrôles qui ont été séquencés partiellement et ce sont les gènes *AhR* et les contrôles pour lesquels des amorces sont prêtes à être utilisées pour la quantification par qPCR.

Le deuxième objectif est d'observer les effets génétiques des pesticides de manière globale sur des organismes vivants. Un génome de référence est donc nécessaire pour identifier les régions qui vont être régulées (directement ou indirectement) par les polluants d'origine agricole. Pour la grande musaraigne, c'est le génome de la musaraigne commune (*Sorex araneus*) qui sera utilisé, car ces deux espèces sont très proches phylogénétiquement. Par contre, pour l'hirondelle bicolore, le séquençage, l'assemblage et l'annotation de son génome seront essentiels parce que les espèces d'oiseaux actuellement séquencées sont trop éloignées pour permettre l'identification des régions qui seront régulées différemment en présence de pesticides. Le séquençage a été fait et l'assemblage a permis de couvrir 55% du génome de l'hirondelle bicolore. Cependant, il est très fractionné avec un N50 de 1339 et un peu plus de 600 000 contigs. Quelques étapes restent à accomplir pour optimiser l'assemblage tel que l'élimination de la contamination (estimée à 5%). L'annotation pourra être faite lorsque l'étape précédente sera finie. Compte tenu de l'ampleur du projet, ce qui a été effectué dans le cadre de la maîtrise contribuera grandement à la bonne continuation de celui-ci. Le projet global permettra de mieux comprendre l'impact des pesticides sur des organismes sauvages.

AhR, CYP1A1, CYP1B1, régulation génique, pesticides, *Blarina brevicauda*, *Tachycineta bicolor*

REMERCIEMENTS

J’aimerais tout d’abord remercier mon directeur de recherche, le professeur Luc Gaudreau, de m’avoir acceptée en tant qu’étudiante à la maîtrise dans son laboratoire et de m’avoir permis de participer à un projet d’envergure avec des axes d’expérimentation aussi diversifiés. J’aimerais remercier la professeure Fanie Pelletier, ma codirectrice de maîtrise, grâce à qui j’ai eu accès à différentes ressources (matérielles et humaines) pour planifier certaines étapes de ma maîtrise. J’aimerais remercier mes deux conseillers Benoit Leblanc et le professeur Pierre-Étienne Jacques. Benoit, pour ces illustrations aussi divertissantes qu’utiles. Je ne sais pas ce que j’aurais fait sans la précieuse aide de Pierre-Étienne pour la partie bio-informatique du projet. Quand j’ai débuté, le seul aspect que je connaissais était la fameuse « boîte noire ». Évidemment, je remercie tous mes collègues passés et présents du laboratoire du Pr Gaudreau pour les conseils reçus tout au long de mon cheminement. Je remercie aussi tous les membres des laboratoires des professeurs Fanie Pelletier, Dany Garant et Marc Bélisle qui m’ont aidée à obtenir les spécimens d'hirondelles et/ou de musaraignes. Un énorme merci pour toutes les personnes qui m’ont aidée à différentes étapes du séquençage et/ou de l’assemblage du génome de l’hirondelle bicolore. Pour cette partie, j’aimerais souligner le professeur Sébastien Rodrigue, Charles Coulombe, Vincent Baby et Dominik Matteau. J’aimerais de plus remercier ma famille et mes amis qui m’ont soutenue tout au long de mes études que se soit pour nos petites discussions, pour jouer une partie de volley ou m’encourager à persévérer. Une mention toute spéciale pour Joannie, Sonia et Alexandre. Finalement, je remercie les organismes subventionnaires, soit l’Université de Sherbrooke, le CRSNG et le les Chaires de recherche du Canada.

TABLE DES MATIÈRES

SOMMAIRE ii
REMERCIEMENTS iv
TABLE DES MATIÈRES v
LISTE DES ABRÉVIATIONS vii
LISTE DES TABLEAUX ix
LISTE DES FIGURES xi
CHAPITRE 1 INTRODUCTION GÉNÉRALE 1
1.1 Les pesticides 1
1.2 Le mécanisme d'AhR 3
1.2.1. Les ligands 6
1.2.1.1 Les ligands exogènes 8
1.2.1.2 Les ligands endogènes 12
1.2.2 L'implication d'AhR dans les processus physiologiques 16
1.2.2.1 Le développement 17
1.2.2.2 Le système immunitaire et endocrinien 18
1.2.2.3 La détoxification 19
1.3 Le séquençage et l'assemblage d'un génome 21
1.3.1 Séquençage MiSeq 21
1.3.2 Évaluation de la qualité d'un séquençage 25
1.3.3 Évaluation de la qualité d'un assemblage 29
1.3.4 Le séquençage, l'assemblage et l'annotation de génomes d'oiseaux 29
1.3.4.1 Le séquençage 29
1.3.4.2 L'assemblage 30
1.3.4.3 L'annotation du génome 34
1.4 Le projet de recherche 35
1.4.1 Les objectifs globaux 36
1.4.2 Les objectifs spécifiques de maîtrise 37
CHAPITRE 2 DÉVELOPPEMENT 39
2.1 Matériels et méthodes 39
2.1.1 Les échantillons 39
2.1.2 Traitement des échantillons 44
2.1.2.1 Extraction d'ARN pour le clonage 44
2.1.2.2 Extraction de l'ADN pour le séquençage du génome de l'hirondelle bicolore 46
2.1.3 Clonage des gènes cibles *AhR* et *CYP* 46
2.1.4 Préparation de la quantification de l'expression des gènes 48
2.1.5 Préparation de la banque d'inserts pour le séquençage génomique de *Tachycineta bicolor* 50

2.1.6 Assemblage du génome de l'hirondelle bicolore 53
2.1.6.1 Combinaison des lectures d'une paire 54
2.1.6.2 Filtrer les adaptateurs 55
2.1.6.3 Vérification de la qualité des lectures 56
2.1.7 Assemblage 57
2.1.8 Contamination 59
2.2 Résultats 60
2.2.1 Extraction d'ARN 60
2.2.2 Clonage des gènes cibles *AhR* et *CYP* et qPCR 61
2.2.3 Séquençage génomique de *Tachycineta bicolor* 65
2.2.4 Assemblage 67
CHAPITRE 3 DISCUSSION GÉNÉRALE ET CONCLUSION 70
3.1 Préparation des échantillons 70
3.2 Clonage et préparation de la quantification de l'expression des gènes par qPCR 70
3.3 Séquençage et assemblage du génome de l'hirondelle bicolore 71
3.3 Conclusion 75
3.4 Perspectives 76
ANNEXE 1 79
ANNEXE 2 81
ANNEXE 3 82
ANNEXE 4 86
ANNEXE 5 88
BIBLIOGRAPHIE 90

LISTE DES ABRÉVIATIONS

2-OHE2	2-hydroxyestradiol
4-OHE2	4-hydroxyestradiol
βHLH	Facteur de transcription de la famille hélice-loupe-hélice basique
Å	Unité de longueur de 0,1 nanomètre
AA	Acide arachidonique
ADN	Acide désoxyribonucléique
ADNc	ADN complémentaire
ADNg	ADB génomique
AhR	Récepteur aux aryls hydrocarbones ou récepteur aux dioxines
AhRR	Répresseur d'AhR
ARN	Acide ribonucléique
ARNm	ARN messager
Arnt	Translocateur nucléaire d'AhR
ATP	Adénosine triphosphate
BLASTn	«Basic local alignment search tool» sur une banque de nucléotides
ChIP-Seq	Séquençage de la chromatine immunoprécipitée
CYP	Cytochrome P450
CYP1	Cytochrome P450, famille 1
CYP1A1	Cytochrome P450, famille 1, sous-famille A, polypeptide 1
CYP1A4	Cytochrome P450, famille 1, sous-famille A, polypeptide 4
CYP1A5	Cytochrome P450, famille 1, sous-famille A, polypeptide 5
CYP1B1	Cytochrome P450, famille 1, sous-famille B, polypeptide 1
DIM	3,3′-diindolylmethane
dNTP	désoxyribonucléotide triphosphate
DRE	Élément de réponse aux dioxines

EIF1	Facteur d'initiation de la traduction eucaryote 1
EIF4A1	Facteur d'initiation de la traduction eucaryote 4A2
FICZ	Formylindolo(3,2-b)carbazole
g	Intensité de la pesanteur artificielle
HAH	Hydrocarbures aromatiques halogénés
HAP	Hydrocarbures aromatiques polycycliques
Hsp90	Heat shock protein of 90 kDa
I3C	Indole-3carbinol
ICZ	Indolo[3,2-b]carbazole
kb	Kilobase
K_d	Constante de dissociation
kDa	Kilodalton
Kyr	Kynurénine
MeDIP-Seq	Séquençage de l'ADN méthylée immunoprécipitée
MTM 7	Mercator transverse modifiée 7
NLS	Signal de localisation nucléaire
p23	Prostaglandine E synthase 3
PAS	Domaine d'une protéine de type Per-Arnt-Sim
pb	paire de bases
PCB	Polychlorobiphényles
PCR	Réaction en chaîne par polymérase
qPCR	Réaction en chaîne par polymérase quantitative en temps réel
RNA-Seq	Séquençage global des ARN
RPLP0	«60S acidic ribosomal protein P0»
RT-PCR	Transcriptase inverse suivie d'un PCR
TCDD	2,3,7,8-tétrachlorodibenzo-p-dioxine
TP	Température de la pièce (~21 °C)
UVB	Ultraviolet
XAP2	Protéine associée au virus X de l'hépatite B
XRE	Élément de réponse aux xénobiotiques

LISTE DES TABLEAUX

Tableau 1 : Détails sur la provenance des hirondelles bicolores et sur la qualité de l'ARN obtenu. 41

Tableau 2 : Détails sur la provenance des grandes musaraignes et sur la qualité de l'ARN obtenu. 42

Tableau 3 : Amorces utilisées pour cloner les gènes *AhR*, *CYP1B1* et les gènes contrôles chez l'hirondelle bicolore. 47

Tableau 4 : Amorces utilisées pour cloner les gènes *AhR*, *CYP1A1* et les gènes contrôles chez la grande musaraigne. 48

Tableau 5 : Gènes clonés chez l'hirondelle bicolore. 62

Tableau 6 : Gènes clonés chez la grande musaraigne. 63

Tableau 7 : Amorces optimisées pour le qPCR pour les gènes de l'hirondelle bicolore. 63

Tableau 8 : Amorces optimisées pour le qPCR pour les gènes de la grande musaraigne. 64

Tableau 9 : Les caractéristiques des assemblages *de novo* obtenus via Newbler à partir des deux séquençages. 68

Tableau 10 : Les résultats de différentes combinaisons d'assemblages testées pour le premier séquençage. 68

Tableau 11 : Amorces optimisées pour le qPCR pour les gènes de l'hirondelle bicolore. 79

Tableau 12 : Amorces optimisées pour le qPCR pour les gènes de la grande musaraigne. 79

Tableau 13 : Séquences des portions d'ADNc des gènes clonés de l'hirondelle bicolore et la séquence amplifiée par qPCR (en gris) pour certains de ces gènes. 82

Tableau 14 : Séquences des portions d'ADNc des gènes clonés de la grande musaraigne et la séquence amplifiée par qPCR (en gris) pour certains de ces gènes. 84

LISTE DES FIGURES

Figure 1 : Structure du récepteur AhR. 3

Figure 2 : Mécanisme d'activation d'AhR. 6

Figure 3 : Classes de molécules qui lient AhR qui sont les plus étudiées. 7

Figure 4 : Structures des ligands d'AhR. 10

Figure 5 : Métabolisme de PAH menant à la formation d'époxydes. 20

Figure 6 : Méthode de séquençage de la plateforme Illumina. 22

Figure 7 : Étapes sommaires d'un assemblage de génome via un séquençage « paired-end ». 24

Figure 8 : Exemple des données d'une lecture dans un fichier fastq. 24

Figure 9 : La distribution des scores Phred de toutes les lectures à chaque position. 25

Figure 10 : Le nombre de lectures ayant un score Phred moyen de x. 26

Figure 11 : La répartition des bases selon les positions dans les lectures. 27

Figure 12 : Le taux de duplication des lectures. 28

Figure 13 : Étapes sommaires d'un assemblage de génome via des lectures « mate-paired ». 31

Figure 14 : Aire d'échantillonnage de l'hirondelle bicolore et de la grande musaraigne. 40

Figure 15 : La taille des inserts du premier séquençage suite à la sonication de l'ADNg. 49

Figure 16 : Taille de la banque d'inserts pour le séquençage 1. 51

Figure 17 : La taille des inserts du deuxième séquençage suite à la sonication de l'ADNg. 52

Figure 18 : Taille de la banque d'inserts pour le séquençage 2. 53

Figure 19 : Exemple de la qualité d'ARN obtenue. 61

Figure 20 : Le taux de duplication des lectures pour le premier séquençage. 66

Figure 21 : Le taux de duplication des lectures pour le deuxième séquençage. 67

Figure 22 : La moyenne du score Phred pour chaque position pour les lectures du séquençage 1. 86

Figure 23 : Le nombre de lectures ayant un score Phred moyen de x pour le séquençage 1. 86

Figure 24 : La répartition des bases selon les positions dans les lectures pour le séquençage 1. 87

Figure 25 : La moyenne du score Phred pour chaque position pour les lectures du séquençage 2. 88

Figure 26 : Le nombre de lectures ayant un score Phred moyen de x pour le séquençage 2. 88

Figure 27 : La répartition des bases selon les positions dans les lectures pour le séquençage 2. 89

CHAPITRE 1

INTRODUCTION GÉNÉRALE

1.1 Les pesticides

Les polluants sont retrouvés partout dans l'environnement et ils peuvent provenir de sources industrielles ou agricoles. Les pesticides représentent la majeure partie des polluants du secteur agricole. D'ailleurs, au cours des quinze dernières années, un minimum de 3500 tonnes de pesticides est utilisé chaque année dans la province du Québec seulement. La grande majorité, soit 80%, est épandue sur les terres agricoles pour lutter contre les organismes indésirables et maximiser le rendement des terres, alors que le restant est utilisé à des fins domestiques (Gorse & Blag, 2013). Les pesticides sont composés d'un mélange de différentes molécules dont la plus importante est l'ingrédient actif qui confère l'effet insecticide, herbicide ou fongicide. Les pesticides peuvent également avoir un résultat non désiré sur l'environnement et/ou sur des organismes vivants avec lesquels ils entrent en contact.

Dans les années 90, plusieurs pesticides qui étaient sur le marché québécois contenaient comme ingrédients actifs des molécules avec des effets secondaires nocifs. Les plus dangereux d'entre eux ont été retirés du marché à la fin du XXe siècle à la suite d'études démontrant leurs effets néfastes sur les humains et/ou l'environnement (Gorse & Blag, 2013). Le début des années 2000 a donc connu une nette amélioration en ce qui concerne le niveau de risques relié à l'utilisation de ces pesticides. Malheureusement, certains des pesticides encore utilisés aujourd'hui ont eux aussi un effet toxique sur la santé humaine et/ou l'environnement, mais il est moins notoire et/ou moins documenté (Hurd, Walker, & Whalen, 2012; Lindsay, Chasse,

Butler, Morrill, & Van Beneden, 2010; Liu, Wang, Zhang, Zhu, & Li, 2013; Montaño, Gutleb, & Murk, 2013; Zhu & Shan, 2009). Un plus grand nombre d'études doit être fait pour évaluer l'impact de ces pesticides sur les humains et/ou l'environnement pour s'assurer qu'ils ne mèneront pas à un ou des problème(s) de santé publique. Comme la composition des pesticides en ingrédients actifs et des autres molécules varie d'un pesticide à l'autre et que le potentiel de risque pour l'environnement ou pour la santé varie pour chaque ingrédient, il est essentiel d'étudier chaque pesticide. Une augmentation des chances de développer plusieurs types de cancers et de devenir infertile sont des exemples des effets secondaires des pesticides sur les humains lorsqu'ils sont administrés à haute concentration. De plus, à long terme, ils peuvent entrainer une perturbation de la régulation des hormones sexuelles (Craig et al., 2011; De Coster and van Larebeke, 2012). Au Canada, c'est Santé Canada qui décide si les pesticides sont homologués ou non selon les effets connus de ces produits. Chaque homologation est normalement réévaluée aux cinq-dix ans (Santé Canada).

La présence d'une molécule étrangère chez un organisme mène à l'activation de plusieurs voies cellulaires pour favoriser la dégradation et/ou l'élimination de cette molécule exogène. Dans le cas des pesticides, ce sont les ingrédients actifs qui jouent le rôle de ces xénobiotiques. Plus précisément, ils peuvent causer une altération de la régulation génique de processus cellulaires chez les organismes (De Coster & van Larebeke, 2012). Par exemple, le récepteur aux dioxines (AhR) peut être activé par certains ingrédients actifs lorsqu'ils sont présents dans le cytosol. Après son activation, il stimule l'expression de certains gènes impliqués dans plusieurs processus physiologiques tels que le développement, le système immunitaire, la tumorigenèse et la détoxification (discuté dans la section 1.3) (Stockinger, Di Meglio, Gialitakis, & Duarte, 2014; Tsuchiya, Nakajima, & Yokoi, 2005).

1.2 Le mécanisme d'AhR

Le récepteur aux dioxines, AhR, agit comme un facteur de transcription lorsqu'il est activé par un de ses multiples ligands. Il est impliqué dans plusieurs processus cellulaires tels que l'embryogenèse, l'inflammation et la détoxification (Opitz et al., 2011). Il serait apparu il y a environ 450 millions d'années, c'est-à-dire avant la divergence entre les poissons à cartilages et ceux à squelette (Hankinson, 1995). Il est donc retrouvé chez plusieurs espèces du règne animal. Il fait partie de la famille « basic Helix-loop-Helix/Per-Arnt-Sim » (βHLH/PAS). Situé à l'extrémité N-terminale d'AhR, le motif HLH est retrouvé chez plusieurs facteurs de transcription (Figure 1). Il est responsable de la liaison entre AhR, l'ADN et d'autres protéines avec lesquels le récepteur interagit (Hankinson, 1995; Mimura & Fujii-Kuriyama, 2003). Quant au domaine PAS, il est composé des régions répétées imparfaites PAS A et PAS B. Comme le motif βHLH, il est responsable de la liaison d'AhR avec d'autres protéines. Cependant, le domaine PAS a une deuxième fonction, il est impliqué dans la liaison entre AhR et les ligands qui sont responsables de son activation (Hankinson, 1995; Mimura & Fujii-Kuriyama, 2003). Le gène *AhR* est localisé sur le chromosome 7 sur le bras p15 chez l'humain. Il fait une longueur d'environ 50 kb, contient 11 exons et mène à la production d'une protéine de 96 kDa (Stejskalova, Dvorak, & Pavek, 2011).

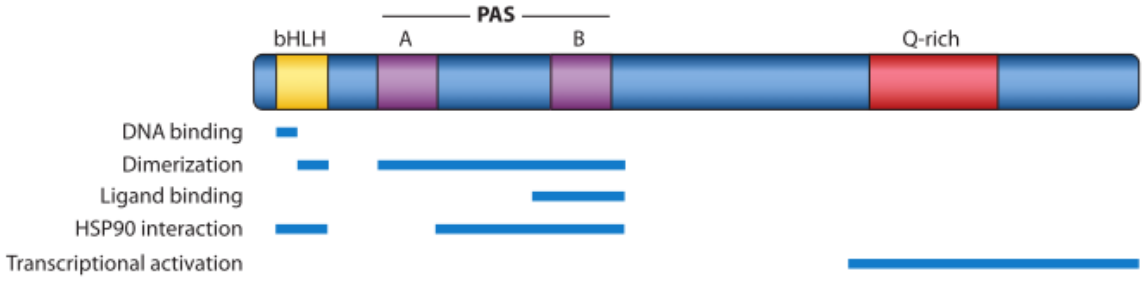


Figure 1 : Structure du récepteur AhR.
(Stockinger et al., 2014)

Lorsqu'inactif (sans la présence de ligand), AhR est maintenu dans le cytosol par un complexe de chaperonnes qui masque le signal de translocation nucléaire (NLS) situé dans le motif βHLH (Figure 1 et Figure 2). Le complexe chaperonne est composé de deux protéines de choc thermique de 90 kDa (Hsp90), d'une protéine associée au virus de l'hépatite B (XAP2) et d'une prostaglandine E synthase 3 (p23). Ces deux dernières seraient probablement nécessaires pour stabiliser Hsp90 (Kazlauskas, Sundström, & Poellinger, 2001; Mimura & Fujii-Kuriyama, 2003). Hsp90 et p23 seraient impliquées dans la liaison entre AhR et ses ligands. En effet, ils engendreraient un changement de conformation du récepteur et permettraient une bonne localisation de ce dernier dans le cytoplasme pour favoriser sa liaison avec un ligand. Hsp90 serait d'ailleurs nécessaire pour cette association. Il masque aussi la zone de liaison à l'ADN dans le domaine βHLH d'AhR (Stockinger et al., 2014). Après la liaison d'un ligand, le complexe de chaperonnes change la conformation d'AhR exposant le NLS à la machinerie d'importation nucléaire favorisant le transport du récepteur vers le noyau (Denison & Nagy, 2003; Kazlauskas et al., 2001; Mimura & Fujii-Kuriyama, 2003; Nguyen & Bradfield, 2008). Comme AhR est trop gros pour passer par les pores nucléaires pour aller au noyau, il est activement transporté à travers la membrane nucléaire par l'importine β (Stockinger et al., 2014).

Rendu dans le noyau, AhR y est séquestré en formant un dimer avec le translocateur nucléaire d'AhR (Arnt) qui fait aussi partie de la famille βHLH/PAS (Hankinson, 1995; Kazlauskas et al., 2001). C'est lors de cette association que le complexe chaperonne se dissocie d'AhR (Stockinger et al., 2014). Le complexe hétérodimèrique (AhR/Arnt) est stabilisé par une structure à quatre hélices due à l'interaction du domaine βHLH des deux protéines. Le complexe se lie ensuite à des éléments situés en amont du promoteur de gènes cibles, soit les éléments de réponses aux xénobiotiques (XRE). Ils sont aussi connus comme étant les éléments de réponse aux dioxines (DRE) (Mimura & Fujii-Kuriyama, 2003; Stejskalova et al., 2011; Tsuchiya et al., 2005). La séquence des XRE reconnue par le complexe hétérodimère AhR/Arnt est 5'-TNGCGTG-3' (Mimura & Fujii-Kuriyama, 2003). Cependant, la formation du complexe AhR/Arnt seule ne serait pas suffisante pour une liaison aux XRE, car un

traitement par une phosphatase empêche cette interaction. La dimérisation se fait vraisemblablement via la consommation d'énergie sous forme d'ATP (Mimura & Fujii-Kuriyama, 2003). Après à la liaison de l'hétérodimère AhR/Arnt aux XRE, le complexe engendre le remodelage de la structure de l'ADN en interagissant avec Brg1 qui est une sous-unité du complexe de remodelage SWI/SNF. Ces protéines sont impliquées dans l'expression des gènes (Stockinger et al., 2014). De plus, la partie C-terminale d'AhR facilite le recrutement de la machinerie de transcription en favorisant le recrutement de coactivateurs et de facteurs de transcription comme p300, SRC-1, TFIIB, le complexe P-TEFb et la protéine spécifique 1 (Sp1) (Mimura & Fujii-Kuriyama, 2003; Nguyen & Bradfield, 2008; Stockinger et al., 2014). Les gènes cibles vont ainsi être transcrits, les ARNm être traduits et les protéines exercer leur(s) fonction(s).

Lorsque non lié à AhR, Arnt est associé à la protéine répresseur d'AhR (AhRR) dans le noyau. Contrairement au complexe AhR/Arnt, le complexe Arnt/AhRR inhibe l'expression des gènes en se liant aux XRE. Dans le cas où AhR n'est pas lié à l'Arnt ou si le complexe AhR/Arnt n'est pas lié à de l'ADN, le signal d'exportation nucléaire du récepteur mène au retour d'AhR dans le cytoplasme où va il être dégradé après son ubiquitination (Denison & Nagy, 2003). Comme il y a des XRE en amont du gène *AhRR*, AhR et AhRR s'autorégulent en formant une boucle de rétroaction. Tous les deux peuvent contrôler l'expression du gène répresseur et ce dernier régule l'activité d'AhR en empêchant ou non l'Arnt de ce lier au récepteur en favorisant ou non son exportation du noyau (Mimura & Fujii-Kuriyama, 2003).

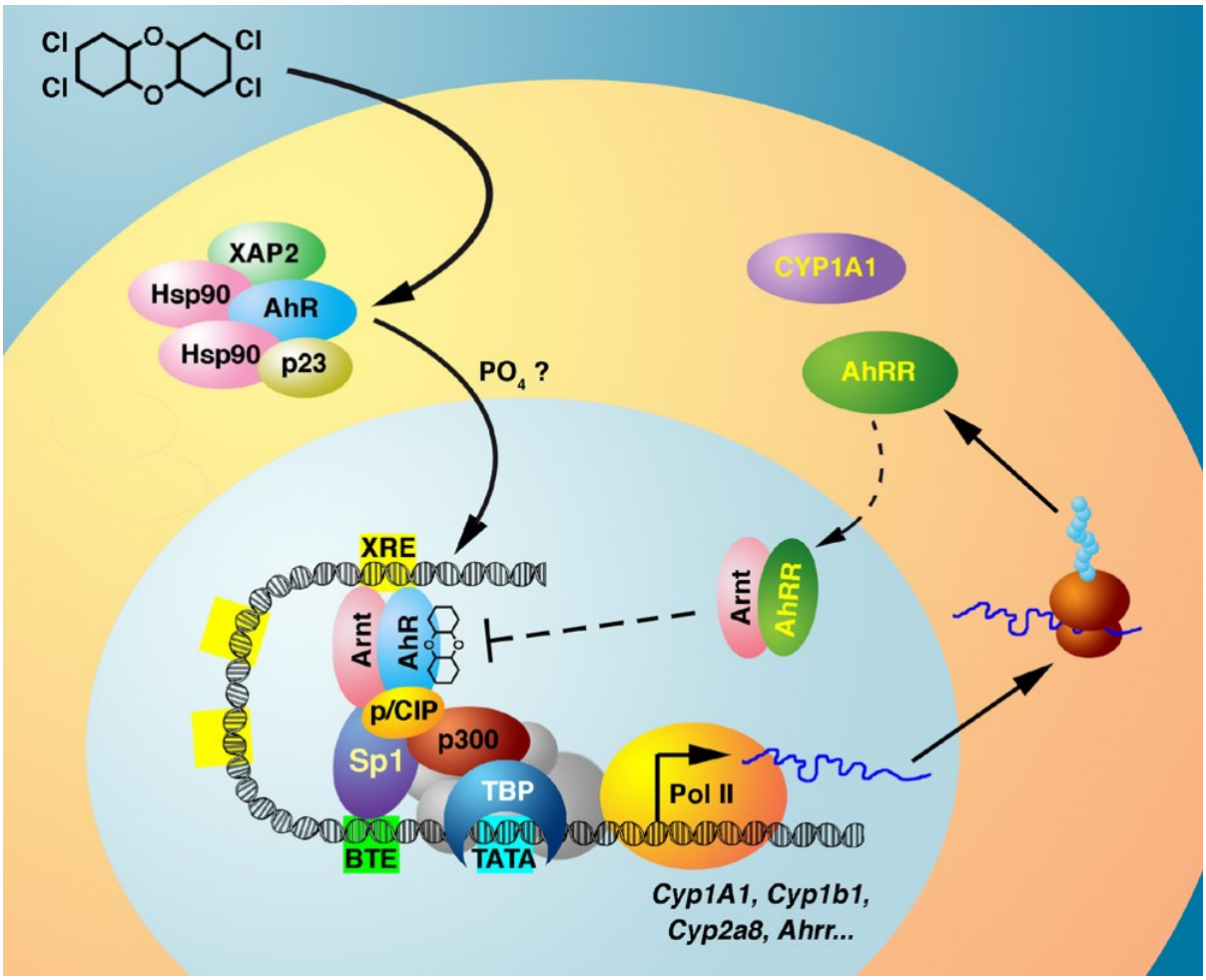
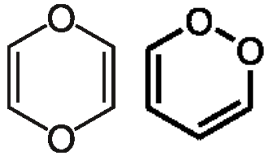


Figure 2 : Mécanisme d'activation d'AhR.
Gracieuseté de Benoît Leblanc.

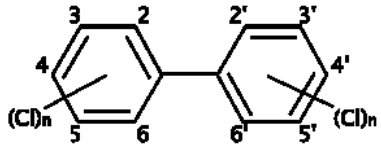
1.2.1. Les ligands

AhR peut être activé par diverses molécules. La plus grande classe de ces ligands est dite exogène, car ils ne sont pas produits par l'organisme lui-même. Ils sont soit des molécules produites naturellement par d'autres espèces, soit des contaminants environnementaux ou thérapeutiques qui ont été produits par l'activité humaine (Nguyen & Bradfield, 2008). Ces ligands exogènes entrent dans un organisme par contact après à une exposition à ces derniers.

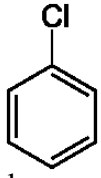
La seconde classe de ligands est composée de molécules endogènes qui sont produites directement par l'organisme en question (Nguyen & Bradfield, 2008). À ce jour, les ligands d'AhR les plus étudiés sont des molécules exogènes. Évidemment, l'ajout de ces molécules à des modèles *in vitro* ou *in vivo* pour voir l'effet qu'elles ont sur le récepteur AhR est plus facile que d'étudier l'effet d'une molécule X produite par les modèles surtout si elle n'a pas déjà été identifiée et que son métabolisme est plus ou moins connu. Comme AhR pouvait parfois être activé sans une exposition à un ligand exogène et que des anormalités développementales et physiologiques avaient été observées chez des souris où l'expression d'*AhR* avait été compromise, l'existence de ligands endogènes a longtemps été soupçonnée sans être validée (Denison & Nagy, 2003). Ce n'est que récemment qu'un d'entre eux a été reconnu (Opitz et al., 2011). Pour les autres, certaines preuves restent à être obtenues pour les classer (ou non) comme ligands officiels d'AhR (Denison & Nagy, 2003).



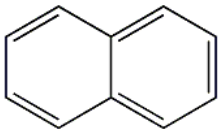
Dioxine



Polychlorobiphényles (PCB)



Hydrocarbure aromatique halogéné (HAH)



Hydrocarbure aromatique polycyclique (HAP)

Figure 3 : Classes de molécules qui lient AhR qui sont les plus étudiées.

Parmi les ligands identifiés, on retrouve principalement des molécules hydrophobes avec au moins un cycle aromatique. La majorité sont planaires bien que la liaison au récepteur peut encore ce faire si le ligand ne l'est pas complètement. Cependant, ce facteur influence le degré

d'affinité de la molécule pour le récepteur. Il a été déterminé que les ligands d'AhR ont une longueur de 12.0 à 14.0 Å, moins de 12 Å en largeur et un maximum de 5.0 Å d'épaisseur. Évidemment, le degré d'affinité des ligands pour AhR varie selon les caractéristiques de chacun (ex. : polarité, électronégativité) (Nguyen & Bradfield, 2008). Les dioxines, les hydrocarbures aromatiques polycycliques (HAP), les hydrocarbures aromatiques halogénés (HAH), les polychlorobiphényles (PCB) et d'autres composés ayant une structure semblable sont quelques exemples de ligands bien connus d'AhR (Figure 3). D'autres molécules peu similaires à celles présentées et contenant seulement un cycle de carbone peuvent aussi être des ligands d'AhR. Cependant, ces dernières molécules ont généralement une affinité plus faible pour le récepteur (Denison & Nagy, 2003; Hankinson, 1995; L'Héritier, Marques, Fauteux, & Gaudreau, 2014).

1.2.1.1 Les ligands exogènes

Comme mentionné précédemment, les ligands exogènes proviennent de deux sources, soit directement de l'environnement sans intervention humaine (produits par d'autres espèces) soit de l'activité humaine. Quelques exemples de ceux produits naturellement sont les flavonoïdes, l'indole-3-carbinol (I3C) et ses dérivés (Denison & Nagy, 2003; L'Héritier et al., 2014; Nguyen & Bradfield, 2008; Stejskalova et al., 2011). Les plus connus provenant de l'activité humaine sont les dioxines, les HAH, les HAP et les PCB.

Certains flavonoïdes, soit les flavones, les flavonols et les isoflavones, ont la capacité d'induire l'expression de certains gènes rapporteurs sous l'influence d'AhR chez des lignées cellulaires de mammifères. Par contre, on ne sait pas si l'activation d'AhR se fait directement par eux ou s'il s'agit d'un métabolite secondaire produit par leur métabolisme qui mène à cette activation. La quercetine, un flavonol, peut induire une activité « ethoxyresorufin-O-deethylase » (EROD). L'activité EROD est un biomarqueur utilisé pour confirmer l'exposition à des HAP

ou des HAH. Cette activité sert, par extension, d'indicateur d'une activation potentielle du récepteur AhR. De plus, la quercetine induit l'expression de *CYP1A1* et elle compétitionne avec le 2,3,7,8-tétrachlorodibenzo-p-dioxine (TCDD) pour se lier à AhR (Nguyen & Bradfield, 2008). Le TCDD est le meilleur ligand connu d'AhR à ce jour (discuté en détail un peu plus loin). D'après ces différents résultats, il est évident que certains flavonoïdes ont une incidence sur le récepteur AhR, mais il reste à déterminer si c'est de manière directe ou indirecte.

L'I3C est retrouvé naturellement dans les plantes. Lui-même n'active pas AhR, mais ses métabolites qui sont produits dans les conditions acides de l'estomac, soit l'indolo[3,2-b]carbazole (ICZ) et le 3,3′-diindolylmethane (DIM), la permettent (Denison & Nagy, 2003; Nguyen & Bradfield, 2008). Ils peuvent d'ailleurs induire l'expression de *CYP1A1* de même qu'une activité EROD. Entre le DIM et le ICZ, c'est ce dernier qui est le meilleur ligand d'AhR (Nguyen & Bradfield, 2008).

De tous les ligands exogènes, le TCDD est l'activateur par excellence du récepteur AhR. Sa constante de dissociation pour ce récepteur est effectivement très faible (Mimura & Fujii-Kuriyama, 2003). Le TCDD est une dioxine (Figure 4A). Ces dernières sont des molécules hétérocycliques aromatiques avec deux atomes d'oxygène dans un des cycles. Elles sont produites par les phénomènes de combustion et par la chloration de produits phénolés (Stejskalova et al., 2011). Elles sont toutes des ligands d'AhR, mais avec un degré d'affinité moins fort que la dioxine TCDD (Tsuchiya et al., 2005). Il est important de noter que la sensibilité d'AhR pour le TCDD varie énormément entre les espèces. En effet, une différence de toxicité 5000 fois moins élevée pour le cochon d'Inde que pour le hamster a été observée. Cet écart de toxicité a été attribué principalement au polymorphisme d'AhR entre différentes espèces (Mimura & Fujii-Kuriyama, 2003).

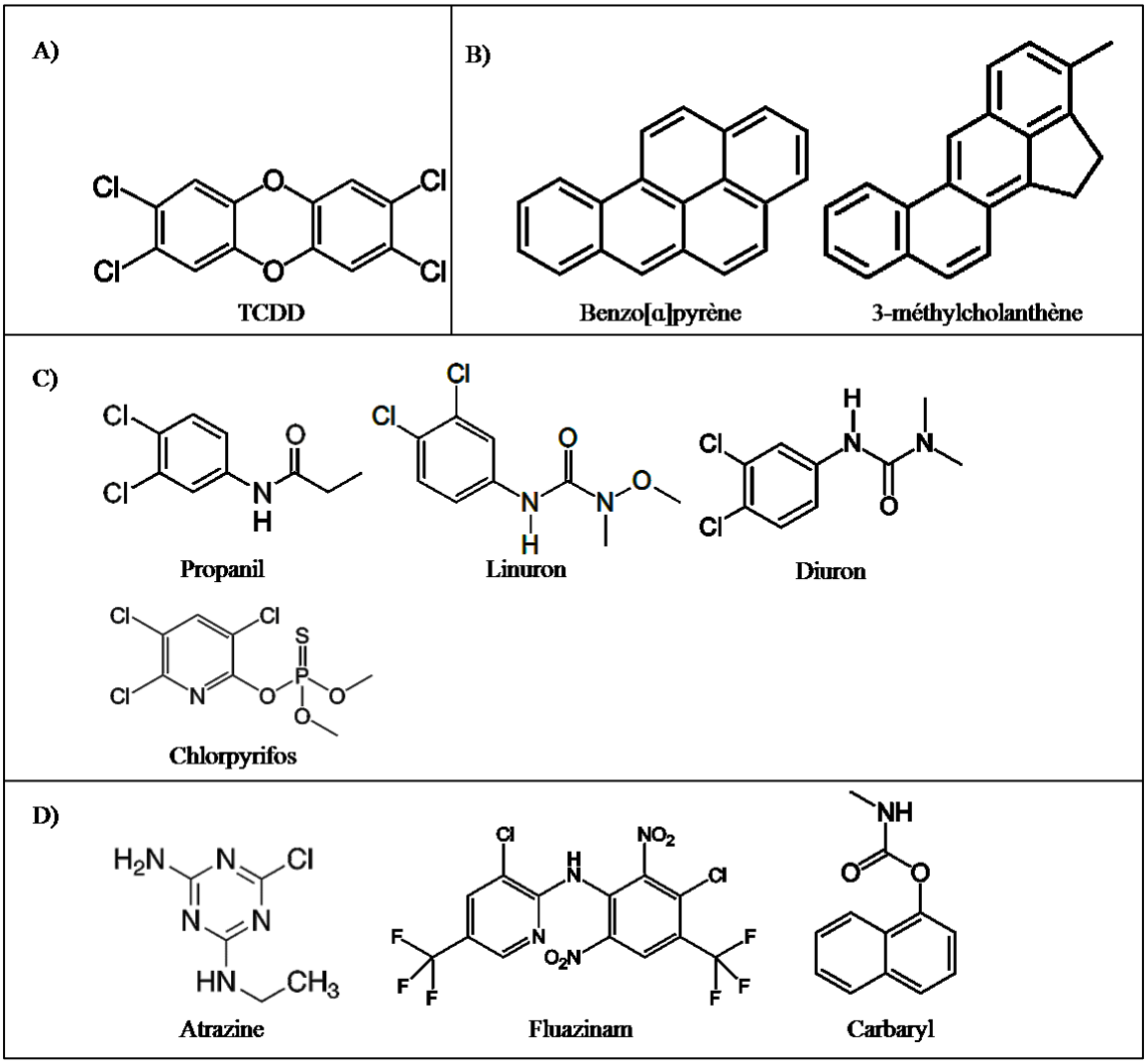


Figure 4 : Structures des ligands d'AhR.

Le plus fort activateur d'AhR connu à ce jour, le TCDD (A). Les PAH les plus connus (B). Quelques exemples d'ingrédients actifs activant le récepteur aux dioxines (C). D'autres exemples d'ingrédients actifs faisant partie des groupes chimiques activant AhR, mais qui n'ont pas encore été validés comme ligand (D).

Les HAH sont produits par la combustion, la chloration et le blanchissement du bois (Stejskalova et al., 2011). La structure de base des HAH est comme son nom l'indique un cycle aromatique sur lequel est lié au moins un atome halogène. La position du ou des halogène(s) dans la structure des ligands halogénés est très souvent latérale par rapport au plan

coplanaire de la molécule. C'est leur position dans la structure qui détermine le degré d'affinité entre ces molécules et AhR (Nguyen & Bradfield, 2008). Les HAP sont composés d'au moins deux cycles benzènes. On les retrouve naturellement dans les processus de combustion (éruptions volcaniques, feux de forêt) ou produits par l'activité humaine (fumée de cigarette, barbecue, énergie fossile, charbon). Les deux les plus connus sont le benzo[*a*]pyrène et le 3-méthylcholanthène (Figure 4B)) (Stejskalova et al., 2011).

Les PCB sont utilisés dans plusieurs procédés industriels tels que la fabrication d'huile et de plastique. Quelques-uns ressemblent à des dioxines planaires, alors que la majorité d'entre eux est similaire aux polychlorodibenzodioxines. La présence d'un halogène en position latérale sur le benzène aide à l'activation d'AhR. Ils sont depuis quelques années des contaminants environnementaux ubiquitaires (Stejskalova et al., 2011).

Depuis la révolution industrielle, AhR s'est retrouvé exposé à une panoplie de ligands exogènes avec lesquels il n'avait pratiquement pas de contact auparavant (comme les pesticides). L'augmentation du nombre et de la concentration de ces ligands exogènes pourraient avoir causé une nouvelle pression évolutive chez les espèces qui y sont particulièrement exposées, que se soit de manière aigüe et/ou à long terme. Un tel phénomène a d'ailleurs déjà été documenté dans une population de poulamon d'Atlantique (*Microgadus tomcod*) de la rivière Hudson où les installations électriques produisent des PCB. Cette population de poissons ont effectivement un variant d'*AhR* qui rend le récepteur moins sensible à ces ligands, réduisant ainsi le taux d'activation d'AhR (L'Héritier et al., 2014; Wirgin et al., 2011). Compte tenu de la très faible quantité d'informations disponibles concernant les effets à long terme des polluants et de tous les processus cellulaires dans lesquels AhR est impliqué (discuté dans la section 1.3), il est primordial de continuer les études sur le sujet.

Parlant de polluants, l'utilisation de pesticides fait maintenant partie intégrante des pratiques de l'agriculture moderne, car ils permettent de lutter contre des organismes non désirables comme des champignons, des insectes ou des plantes. Cependant, suite à une exposition à certains d'entre eux, des conséquences secondaires négatives comme des problèmes médicaux peuvent subvenir. Une augmentation de la probabilité de développer des cancers, des problèmes mentaux, neurologiques ou reproductifs ne sont que quelques exemples de ces effets secondaires (Stejskalova et al., 2011). Il est important de noter que plusieurs des ingrédients actifs retrouvés dans les pesticides ont une structure chimique qui fait d'eux de potentiels activateurs d'AhR. La Figure 4D représente trois ingrédients actifs qui sont de potentiels activateurs d'AhR en raison de leur structure qui rappelle celles de ligands d'AhR déjà validés. De plus, au moins 200 ingrédients actifs ont été identifiés comme agonistes du récepteur comme le chlorpyrifos, le propanil, le diuron et le linuron (Figure 4C) (Stejskalova et al., 2011). Ces quatre derniers sont d'ailleurs des HAH, il n'y a donc rien d'étonnant à ce qu'ils activent AhR.

1.2.1.2 Les ligands endogènes

Comme mentionnée précédemment, l'existence de ligands endogènes a longtemps été une hypothèse. Cette dernière était principalement basée sur la conservation d'AhR chez plusieurs espèces qui ne pouvait pas s'expliquer par la seule présence de ligands exogènes. Si toutes ces espèces vivaient dans le même environnement et étaient toutes exposées aux mêmes ligands exogènes, la conservation d'AhR pourrait se justifier seulement par leurs présences. Or, *AhR* est retrouvé chez des vertébrés vivants dans des milieux très diversifiés, qu'ils soient terrestres ou aquatiques. Toutes ces espèces sont donc exposées à des concentrations changeantes d'un milieu à l'autre de ces molécules exogènes différentes qui ont un potentiel d'activation d'AhR variable. De plus, la présence de la plupart des ligands exogènes est très récente due à la révolution industrielle. La conservation d'*AhR* ne pouvait logiquement pas s'expliquer par une pression sélective exercée par les ligands exogènes. L'hypothèse de ligands endogènes d'AhR

a donc été émise pour expliquer le maintien d'AhR chez toutes ces espèces (Nguyen & Bradfield, 2008). Plusieurs ligands endogènes ont été soupçonnés jusqu'à présent. Ils font partie de plusieurs classes de molécules telles que les indoles, les tétrapyrroles et les métabolites de l'acide arachidonique (AA). La variabilité de ligands endogènes hypothétiques reflète la vaste étendue des ligands exogènes identifiés à ce jour (Denison & Nagy, 2003).

Le tryptophane est un des acides aminés essentiels, c'est-à-dire qu'il ne peut être synthétisé par l'humain. Par contre, l'homme possède les enzymes nécessaires pour sa dégradation en métabolites secondaires. Quelques-uns d'entre eux sont des ligands endogènes potentiels d'AhR. Par exemple, la kynurénine (Kyn), un dérivé du tryptophane indolé, a été identifiée comme étant un ligand endogène du récepteur AhR en 2011 (Denison & Nagy, 2003; Opitz et al., 2011). Cette molécule, produite via le catabolisme du tryptophane par les indoleamine-2,3-dioxygenases 1 et 2, a été associée avec une augmentation du risque de développer un cancer chez des souris. En effet, elle provoquerait une diminution de la réponse immunitaire contre les cellules cancérigènes en inhibant la prolifération des cellules T qui sont, en partie, responsables de la protection contre le cancer. Cela aurait comme conséquence la prolifération de cellules cancéreuses. La production de Kyr serait faite par les cellules cancéreuses elle-même en quantité suffisante pour activer AhR (Frumento et al., 2002; Opitz et al., 2011). À ce jour, cette molécule est le seul ligand endogène d'AhR qui a été confirmé *in vivo* et qui a une implication au niveau de maladie physiologique comme le cancer et l'immunité (Opitz et al., 2011).

Un autre indole, le formylindolo(3,2-b)carbazole (FICZ), est dérivé de la photooxydation du tryptophane après une exposition aux ultraviolets (Denison & Nagy, 2003; Stockinger et al., 2014). Le degré d'affinité de ce métabolite pour AhR est similaire à celui du TCDD. Pour confirmer que ce ligand est bel et bien un ligand endogène d'AhR, il reste à déterminer si la formation de FICZ se fait *in vivo* (Nguyen & Bradfield, 2008). L'indigo et l'indirubine, eux aussi des indoles, sont métabolisés principalement chez les plantes, mais aussi par quelques

enzymes faisant partie de la famille du cytochrome P450 (Denison & Nagy, 2003). Leur capacité à activer AhR chez les levures est environ 50 fois plus forte que le ligand TCDD. Ils ont cependant un effet beaucoup moins extraordinaire chez les mammifères, soit de 50 000 à 100 000 fois moins d'activation d'AhR que la « super » dioxine (Denison & Nagy, 2003). Ces deux molécules ne sont pas catégorisées comme des ligands endogènes du récepteur AhR pour le moment, car elles sont présentes en très petites concentrations chez les mammifères, soit de l'ordre du picomolaire, alors que les quantités nécessaires pour activer AhR sont de l'ordre du nanomolaire. De plus, ces deux indoles ont seulement causé l'induction de *CYP1A1* chez le rat. Aucun changement d'expression chez un autre gène normalement régulé par AhR n'a été observé (Nguyen & Bradfield, 2008).

L'équilénine est une hormone estrogène qui est extraite de l'urine de juments (*Equus caballus)* en gestation. Elle induit de quinze fois l'expression de *CYP1A1* et environ cinq fois celle de gènes rapporteurs régulés par les XRE chez les cellules HepG2. Il reste cependant à confirmer que ces observations sont bien dues à l'activation d'AhR par une liaison directe avec l'équilénine, car l'affinité de cette hormone pour le récepteur est très faible. Il serait plus probable que l'équilénine active le récepteur à l'estrogène qui activerait à son tour le récepteur AhR. D'autres études permettront d'éclaircir le mécanisme d'activation d'AhR dans ce cas-ci (Nguyen & Bradfield, 2008).

Les produits de dégradation de l'hème, comme la biliverdine et la bilirubine, font partie de la classe des tétrapyrroles. Ils sont tous de potentiels ligands endogènes du récepteur AhR (Denison & Nagy, 2003). En effet, l'exposition de cellules hépatiques de souris à ces molécules induit, de manière dépendante à AhR, l'expression de *CYP1A1*, de gènes rapporteurs sous le contrôle de XRE en plus de mener à la détection d'une activité EROD. D'après d'autres études, l'hème et la biliverdine ne seraient que les précurseurs de la bilirubine qui serait le véritable activateur d'AhR. Cependant, aucune étude n'a encore confirmé la bilirubine (et/ou un des deux autres) comme un ligand endogène d'AhR. Cette

hypothèse est toujours à valider. Jusqu'à présent, les études vont dans les deux sens. Normalement, très peu d'hème est retrouvé dans le plasma chez les humains. Les concentrations ne seraient pas suffisantes pour mener à l'activation d'AhR. Or, chez les personnes souffrant du syndrome Crigler–Najjar, une concentration de plus de dix fois supérieure à celle nécessaire pour induire AhR est retrouvée dans le plasma. Chez ces personnes, l'hème et/ou ses dérivés pourraient avoir un rôle à jouer via l'activation d'AhR qui est peut-être impliqué dans cette maladie. La validation de l'implication de l'hème et ses dérivés comme ligands endogènes d'AhR pourraient avoir un impact majeur pour l'étude du syndrome Crigler–Najjar (Nguyen & Bradfield, 2008).

Pour finir, la lipoxine 4A, un métaboliste produit via l'acide arachidonique, active aussi AhR. Elle entre en compétition avec le TCDD, est capable d'induire des gènes régulés via les DRE comme *CYP1A1* et *CYP1A2*. L'implication de la lipoxine dans l'activation du récepteur serait probablement validée si on pouvait confirmer l'implication d'AhR dans le processus de développement par la lipoxine 4A (Nguyen & Bradfield, 2008)).

Au final, un seul ligand endogène d'AhR a été validé à ce jour, soit la kynurénine qui est dérivée du tryptophane. Cependant, plusieurs autres molécules endogènes pourraient elles aussi être des ligands endogènes d'AhR si on réussit à prouver que ces ligands potentiels sont produits *in vivo* en concentration suffisante et/ou activent AhR.

1.2.1.3 Activation sans ligands

Une troisième possibilité d'activation d'AhR, indépendante de la liaison d'un ligand, a aussi été proposée. Cette hypothèse a été formulée lorsqu'une liaison entre AhR et de l'ADN a été observée par un mécanisme indépendant du TCDD. Cette interaction a été obtenue chez des

cellules hépatiques de souris après une exposition à de l'AMPc. Cependant, les résultats n'ont pas été assez concluants pour confirmer cette hypothèse. L'activation d'AhR sans ligands reste donc une supposition (Nguyen & Bradfield, 2008).

1.2.2 L'implication d'AhR dans les processus physiologiques

Il est bien connu qu'AhR est impliqué dans la réponse aux xénobiotiques (des molécules étrangères) en stimulant l'expression de gènes de certaines protéines qui les métabolisent, ce qui favorise leur élimination de l'organisme (discuté dans la section 1.3.3) (Tsuchiya et al., 2005). Comme il a été mentionné précédemment, on pourrait croire que la conservation d'AhR chez plusieurs espèces dans l'évolution des vertébrés a un lien avec ces substances xénobiotiques. Or, ces espèces vivent dans des habitats très différents l'un de l'autre où ces molécules ne sont pas nécessairement toujours présentes, ce qui laisse penser qu'AhR pourrait avoir une autre fonction, par exemple d'ordre physiologique. En effet, plusieurs études sur AhR ont démontré qu'il a un rôle dans quelques processus cellulaires importants comme le développement et le système immunitaire. De plus, des protéines similaires à AhR et Arnt ont aussi été retrouvées chez des invertébrés comme *Drosophila melanogaster* et *Caenorhabditis* elegans venant appuyer l'importance d'AhR pour les organismes (Nguyen & Bradfield, 2008). Vu la grande variété de molécules pouvant agir comme ligand d'AhR et l'importance de ce dernier dans la régulation de processus cellulaires, il est impératif d'avoir une bonne compréhension de la manière dont il régule ces différentes voies. La section qui suit porte sur les principaux processus dans lesquels AhR a un rôle.

1.2.2.1 Le développement

Plusieurs facteurs sont impliqués dans le développement embryonnaire (différenciation cellulaire). Oct-4, c-Myc, Nanog et p53 sont les plus connus et étudiés, mais il semble que le récepteur aux dioxines ait également un rôle à jouer (Lindsey & Papoutsakis, 2012). Quelques études ont constaté des problèmes physiologiques lorsque l'expression ou le fonctionnement d'AhR était anormal. Par exemple, un taux de mortalité néonatale de 40 à 50% a été observé chez des souriceaux mutants pour l'exon 1 d'AhR. Cependant, les survivants avaient une diminution de 80% du nombre de lymphocytes dans la rate et une inflammation des voies biliaires. De plus, elles démontraient, en vieillissant, plusieurs indices de malformation. D'après ces résultats, AhR serait impliqué à plusieurs étapes de l'embryogenèse, soit dans la formation du cœur, de la peau et du système gastro-intestinal (Fernandez-Salguero et al., 1995).

Dans une autre étude, chez des souris mutantes pour le gène AhR, 100% d'entre elles avaient le canal d'Arantius qui restait ouvert, causant ainsi un dérèglement des métabolites produits dans le foie. Ce canal permet une bonne circulation du sang dans le foie du fœtus et se referme normalement après la naissance. Des souris sous-exprimant AhR avaient elles aussi tendance à avoir le canal d'Arantius ouvert après la naissance. Cependant, le canal de telles souris se fermait si AhR avait été stimulée par l'administration de TCDD au fœtus. De plus, dans ce dernier cas, le foie, à l'âge adulte, avait une taille normale. Ces résultats suggèrent fortement qu'AhR aurait un rôle à jouer dans le développement du foie et du système cardiovasculaire puisque l'expression normale du récepteur et/ou une activation de ce dernier (en cas d'inhibition d'expression) par le TCDD mènent à un phénotype normal (Lahvis et al., 2000). Finalement, les gènes régulés par les XRE sont exprimés durant le développement ce qui soutient l'implication d'AhR dans ce processus (Nguyen & Bradfield, 2008).

1.2.2.2 Le système immunitaire et endocrinien

AhR serait aussi impliqué dans plusieurs autres processus cellulaires. En effet, lorsqu'activé par des ligands, AhR mène à différents effets selon la nature et de la concentration de ces premiers. Par exemple, à de faibles concentrations (de l'ordre de ng/kg), le TCDD induit via AhR l'expression des gènes CYP dans le but de métaboliser les xénobiotiques en plus de réguler certains processus endocriniens. Cependant, à de plus fortes concentrations (µg/kg), des effets dommageables et non désirables sont observés comme une involution thymique (menant à une immunosuppression), des malformations néonatales et de plus fortes probabilités de développer un cancer (Stejskalova et al., 2011). Ces observations sont toutes dues à une suractivation d'AhR (via le TCDD), car chez des souris mutantes pour le gène *AhR*, les résultats obtenus après à une exposition au TCDD ne sont plus observés (Nguyen & Bradfield, 2008). Lorsqu'activé par des xénobiotiques, AhR inhibe les réponses immunitaires primaire et secondaire en plus d'être carcinogène en aidant la progression de tumeurs (Opitz et al., 2011). D'ailleurs, deux équipes ont démontré une corrélation positive entre l'activation d'AhR et le degré d'agressivité des tumeurs (Powell et al., 2013; Yang et al., 2008). Il est raisonnable de supposer que différentes concentrations d'autres ligands d'AhR mèneront à diverses réponses comme il a été observé avec le TCDD pour deux raisons principales; les ligands d'AhR se lient au même site sur ce dernier (soit le domaine PAS B) et ces ligands ont des structures similaires (voir Figure 4) (Mimura & Fujii-Kuriyama, 2003).

L'activation d'AhR peut aussi avoir un effet positif sur le système immunitaire. Par exemple, le ligand endogène Kyr, produit par le catabolisme du tryptophane dans les régions d'inflammation, empêche localement les réactions auto-immunes neuroinflammatoires en inhibant la réponse inflammatoire. Cependant, cette inhibition peut favoriser la croissance de tumeurs qui sont à proximité de ces régions où le système immunitaire est inhibé (Opitz et al., 2011).

1.2.2.3 La détoxification

Plusieurs enzymes sont impliquées dans le processus de détoxification. Lorsque le récepteur AhR est activé par un xénobiotique, ce sont les gènes régulé par un XRE qui sont régulés pour mener à la production de protéines de détoxification. Les sections suivantes discuteront de certaines de ces protéines, leurs fonctions et leurs effets.

Parmi les nombreux gènes qui sont régulés par la voie AhR, il est possible de compter certains gènes faisant partie de la famille du cytochrome P450 (CYP). Ces enzymes sont impliquées dans le processus de détoxification de molécules xénobiotiques par une réaction d'oxydation (Tsuchiya et al., 2005). Leur rôle est de transformer les molécules exogènes en composés secondaires moins toxiques pour l'organisme et plus facilement éliminables. CYP1A1 et CYP1B1 sont deux exemples de ces protéines et sont aussi impliquées dans le métabolisme de l'estradiol. Elles peuvent transformer le 17β-estradiol en 2-hydroxyestradiol (2-OHE2) et en 4-hydroxyestradiol (4-OHE2) respectivement (Coumoul, Diry, Robillot, & Barouki, 2001). Entre ces deux molécules, c'est le 2-OHE2 qui est considéré comme la molécule la moins dommageable, car elle inhibe la prolifération cellulaire de lignées du cancer du sein et induit l'apoptose chez des cellules immortelles provenant de la glande mammaire (Gupta, Mcdougal, & Safe, 1998; Hurh, Chen, Na, Han, & Surh, 2004; L'Héritier et al., 2014). Alors que le métabolisme du 4-OHE2 mène à la production d'estradiol-3,4-quinone qui cause la formation d'adduits à l'ADN. Ces adduits viennent déstabiliser la structure de l'ADN favorisant les mutations en causant des cassures des brins (Shimada et al., 1996; Volk et al., 2003; Zhao et al., 2006). Les deux molécules inhibent la prolifération cellulaire lorsqu'elles sont méthylées. Cependant, la méthylation du 4-OHE2 se fait plus lentement que celle du 2-OHE2 par la catéchol-O-méthyle-transférase (Tsuchiya et al., 2005). Bref, basé sur ces résultats, il est vraisemblablement préférable de produire plus de 2-OHE2 que du 4-OHE2. Le ratio CYP1A1/CYP1B1 est donc important considérant les effets que leurs métabolites respectifs

ont sur certains processus cellulaires. En conclusion, les produits de l'enzyme CYP1B1 sont plus génotoxiques pour un organisme que ceux de CYP1A1.

Dans le cas des pesticides, si certains favorisent l'expression du gène *CYP1B1* comparativement à celle de *CYP1A1* par l'activation d'AhR, il va s'en dire qu'ils deviennent préoccupants pour la santé humaine et l'environnement. D'autant plus qu'ils sont persistants et/ou utilisés à répétition année après année. Les effets des pesticides à court terme commencent à être bien documentés, mais c'est loin d'être le cas pour les effets à long terme (L'Héritier et al., 2014).

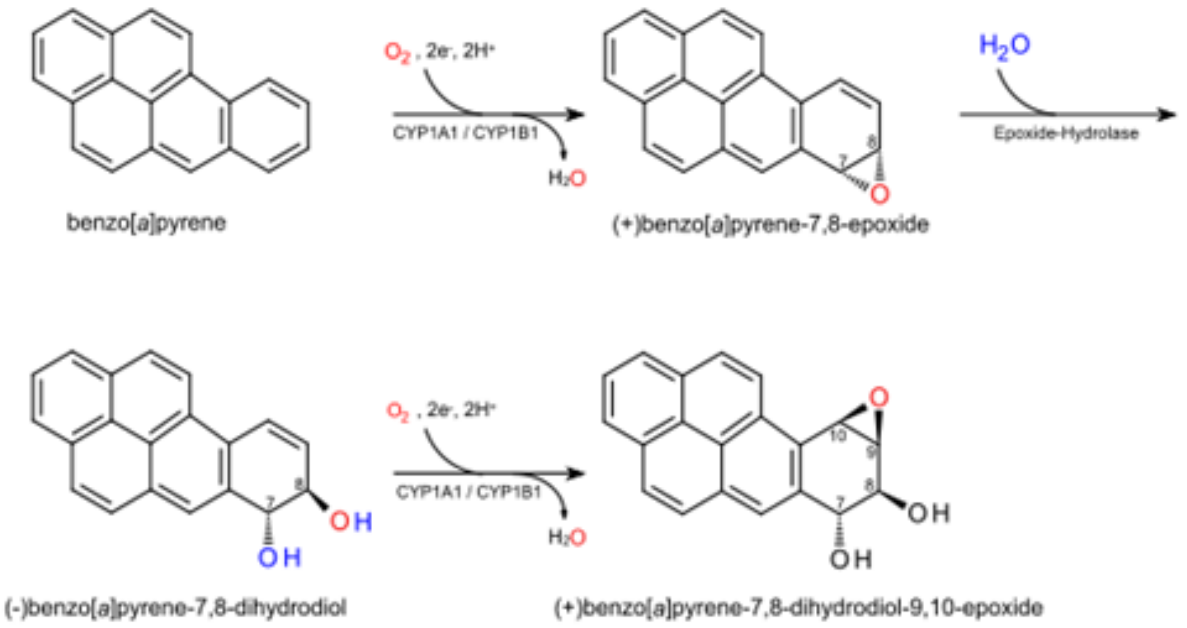


Figure 5 : Métabolisme de PAH menant à la formation d'époxydes.
L'implication des enzymes CYP1A1 et CYP1B1 dans deux étapes du métabolisme du benzo[a]pyrène en époxydes.

Comme mentionné, la fonction de la production de CYP est de détoxifier les cellules des substances pouvant être néfastes pour un organisme. Or, dans certains cas, les métabolites

produits sont plus toxiques que les originaux. Par exemple, ces enzymes sont impliquées dans la production d'espèces réactives de l'oxygène qui sont mutagènes (Mimura & Fujii-Kuriyama, 2003). De plus, lors du métabolisme des HAP, ces enzymes vont plutôt mener à une activation de ces molécules par l'ajout d'un groupe époxyde qui sera ensuite converti par l'époxyde hydrolase en dihydrodiol (Figure 5). Cette molécule sera finalement métabolisée en époxyde diol. Cette dernière engendre des dommages à l'ADN en se liant aux nucléotides, menant ainsi à la formation d'adduits à l'ADN (Stejskalova et al., 2011).

Chez les oiseaux, l'enzyme CYP1B1 est présente. Par contre, ce sont les enzymes CYP1A4 et CYP1A5 qui remplissent un rôle semblable à CYP1A1 chez les mammifères. La séquence d'acides aminés de CYP1A4 et CYP1A5 est d'ailleurs similaire à CYP1A1. Cependant, CYP1A4 ressemble plus enzymatiquement à CYP1A1 qu'à CYP1A2 alors que CYP1A5 est plus similaire à CYP1A2 d'après une comparaison enzymatique et immunologique. Comme il n'y a pas de relation orthologue entre CYP1A1 et CYP1A4 et/ou CYP1A5, ils ne sont pas considérés comme des gènes homologues et ont donc un nom différent. CYP1A4/5 auraient donc évolués indépendamment de CYP1A1/2, mais ils auraient été exposés à des pressions sélectives similaires (Gilday, Gannon, Yutzey, Bader, & Rifkind, 1996).

1.3 Le séquençage et l'assemblage d'un génome

1.3.1 Séquençage MiSeq

Plusieurs technologies sont maintenant disponibles pour séquencer un génome. Certaines permettre d'obtenir des bases avec un haut niveau de fiabilité, mais seulement pour de petits fragments, alors que d'autres permettre d'obtenir de plus longues séquences, mais avec moins

de fiabilité. Évidemment, le coût varie d'une technologie à l'autre dépendamment du service qui est demandé.

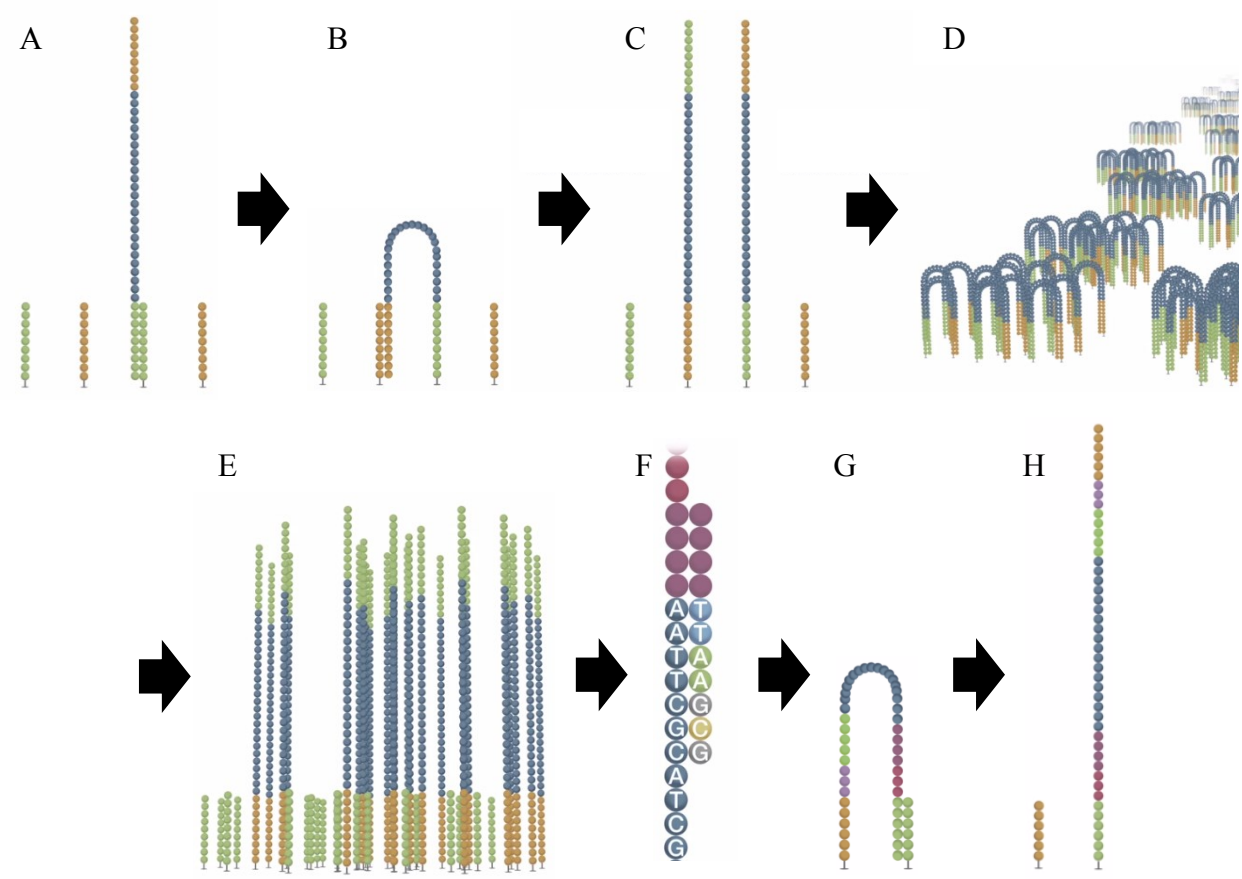


Figure 6 : Méthode de séquençage de la plateforme Illumina.
Chaque insert est fixé sur une plaque à l'aide d'adaptateurs complémentaires déjà présent sur celle-ci (A). Plusieurs cycles d'amplification permettent d'obtenir un grand nombre de copie de chaque insert dans une même région (B-E). Le séquençage se fait d'un côté de chaque fragment afin de générer la première lecture en ajoutant le dNTP complémentaire. Chaque dNTP est marqué à l'aide d'un marqueur fluorescent différent qui empêche l'ajout d'autres dNTP jusqu'à ce qu'il soit enlevé (le marquage est réversible) (F). De cette manière le séquenceur détermine quel dNTP a été ajouté à chaque cycle pour chaque groupe de fragment (correspondant à un insert). Par la suite, l'autre extrémité des fragments est liée au deuxième adaptateur afin de procédé à la deuxième lecture de tous les inserts (G-H) (Illumina).

La première étape est d'extraire l'ADN de cellules de l'espèce à séquencer et de le soniquer pour obtenir des fragments (inserts) de longueurs désirées pour créer la banque d'insert. Le système MiSeq de la plateforme d'Illumina permet d'obtenir de relativement longues lectures, soit de 300 pb. De plus, lorsque le séquençage est fait en « paired-end », chaque insert est séquencé à partir des deux extrémités à l'aide de deux adaptateurs (comprenant une amorce) différents présents à chaque extrémité (Illumina). Le séquençage est illustré à la Figure 6. Pour chaque insert, le séquençage MiSeq donne donc une paire de lectures couvrant jusqu'à 600 pb pour de longs fragments, et se rejoignant pour les plus courts (Figure 7, les 4 premières étapes). Pour les inserts de moins de 600 pb, l'alignement des deux lectures de la paire devrait permettre d'obtenir la séquence complète de chaque insert.

Les lectures obtenues suite à un séquençage sont présentées dans un fichier de format fastq. Ce type de fichier contient le nom des lectures, la séquence de ces lectures, un « + » et le score de qualité de chaque base. Ces quatre informations sont inscrites sur 4 lignes subséquentes pour une lecture, c'est-à-dire que les quatre premières lignes sont les informations pour la première lecture, les quatre suivantes celles du deuxième et ainsi de suite (Figure 8). Le score de qualité (score Phred; Q) attribué à chaque base d'une lecture est donné par le séquenceur. Il dépend du degré de certitude que la base à cette position ait été bien séquencée. L'équation suivante, $P = 10^{\frac{-Q}{10}}$, permet de déterminer la probabilité d'une erreur de séquençage (P) de la base à cette position. Donc, plus le score est élevé, moins il y a des chances d'erreur et inversement. Chaque score Phred est associé à un symbole ASCII afin de permettre l'inscription d'un seul caractère pour le score de qualité chaque base dans le fichier fastq (Figure 8) (Ewing & Green, 2005). À partir d'un score Q de 27 (soit 0,2% d'erreur), la qualité des bases est considéré de passable. Au-delà de 27, la qualité est considérée excellente (Babraham Bioinformatics).

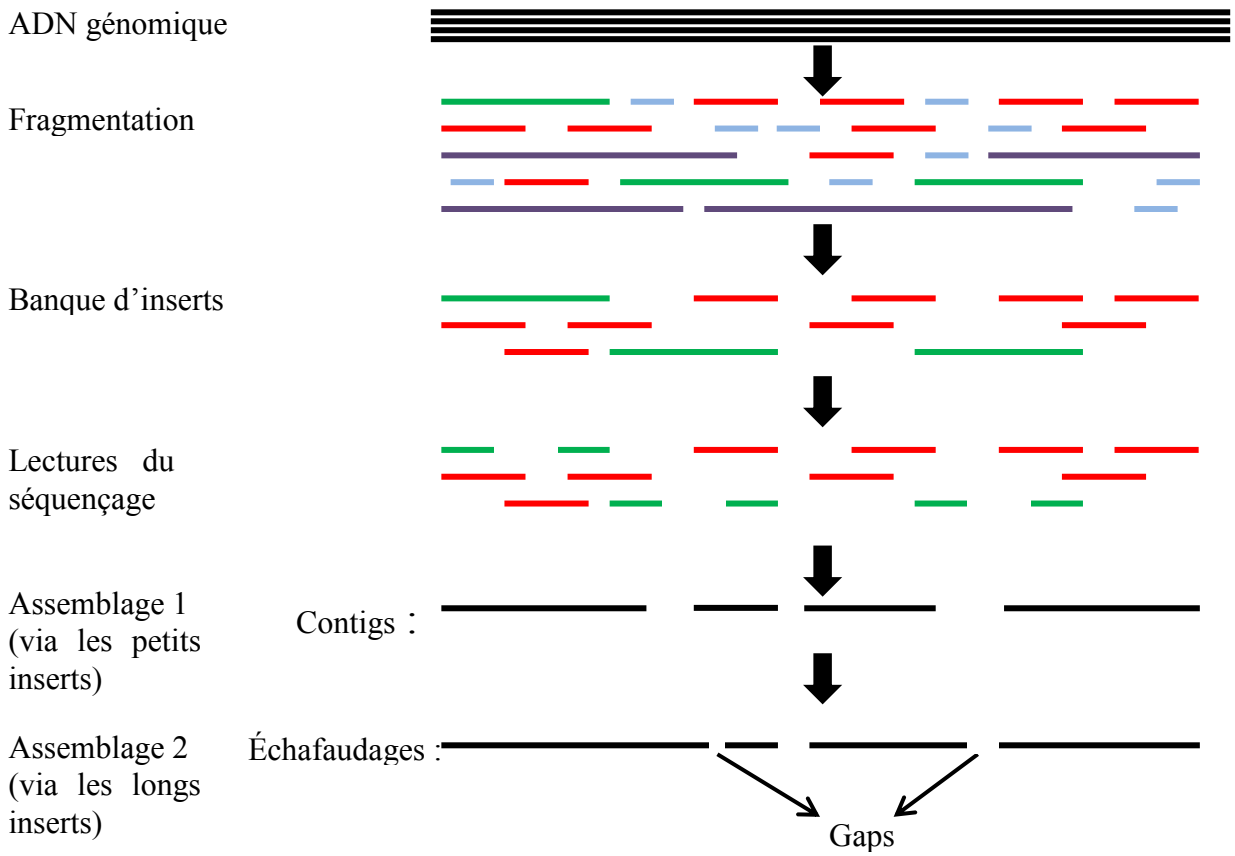


Figure 7 : Étapes sommaires d'un assemblage de génome via un séquençage « paired-end ». L'ADN génomique est isolé et soniqué en fragments de différentes longueurs. Ceux ayant une longueur désirée sont sélectionnés par migration sur gel et constituent la banque d'inserts pour le séquençage (les fragments trop courts ou trop longs ne sont pas conservés). Les inserts plus petits (rouges) sont séquencés au complet (chevauchement des deux lectures au centre de l'insert). Ils sont utilisés lors de la première étape de l'assemblage pour faire les contigs. Pour les inserts plus longs (verts), il n'est pas possible d'avoir une zone de chevauchement entre les deux reads d'une même paire. Ces derniers vont plutôt être utilisés pour faire les « échafaudages » lors de la deuxième étape (assemblage de contigs entre eux et création de trous (« gaps »)).

```
@M01893:10:000000000-A6AGA:1:1101:16010:1169
TTTCCCCTCTTTTTCTCTTTTTTTCTTCTGGGGGGATCTTCAATAAGCCCTTCACAGTGCGTC
+
-88@@CC6@6@;C@C6<CCCCC@@CC;C@,,,++++8::9?,,<,,,9:::9?,,,,:,:+8+
```

Figure 8 : Exemple des données d'une lecture dans un fichier fastq.
Ligne 1 : nom de la lecture; ligne 2 : la séquence; ligne 3 : + et ligne 4 : score de qualité.

1.3.2 Évaluation de la qualité d'un séquençage

Différentes données doivent être analysées après un séquençage pour évaluer la quantité et la qualité des lectures et la composition de la banque (en particulier le score Phred, le taux de GC et la quantité de duplications). L'outil FastQC permet d'analyser ces informations (Babraham Bioinformatics).

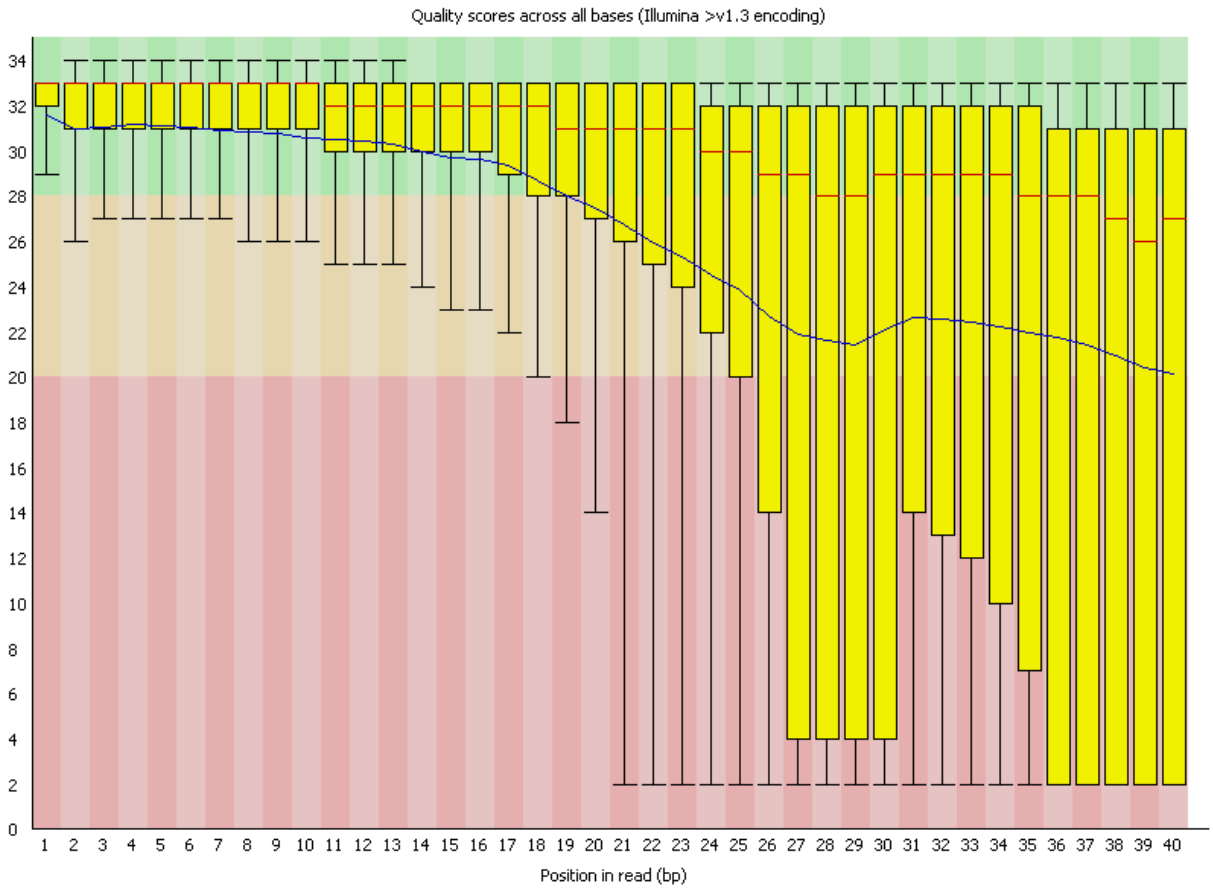


Figure 9 : La distribution des scores Phred de toutes les lectures à chaque position. (Babraham Bioinformatics)

Pour le score de qualité, différents graphiques sont fournis, soit pour donner la distribution du score Phred pour chaque position de toutes les lectures (Figure 9) ou pour dénombrer le nombre de lecture avec un tel score Phred moyen (Figure 10). Pour le premier graphique, la ligne rouge centrale représente la médiane, la bleue la moyenne, les boîtes jaune le deuxième et troisième quartile (25 à 75%) et les moustaches le 10^{ième} et 90^{ième} percentile. Normalement, la qualité diminue vers la fin des lectures et elle est un peu plus faible pour la deuxième lecture d'un insert comparativement à la première (Babraham Bioinformatics).

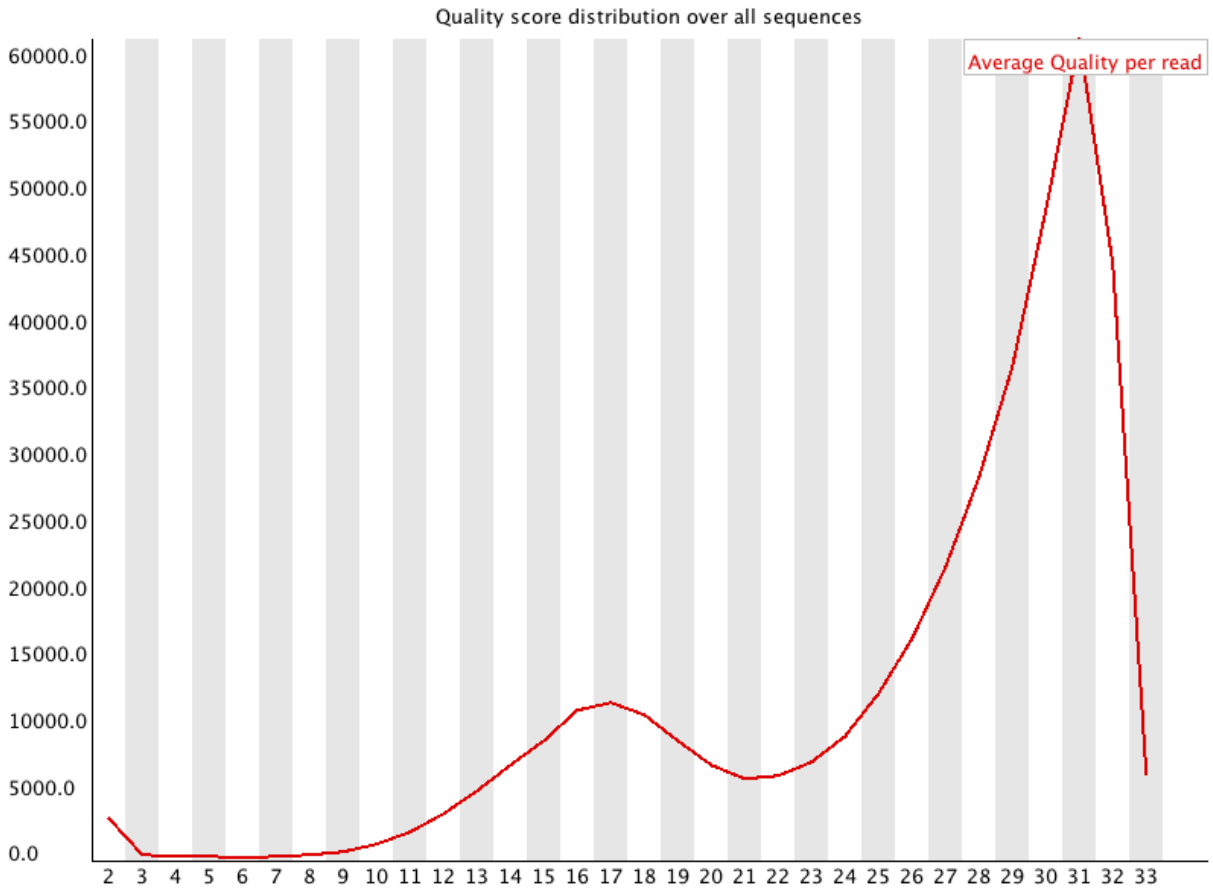


Figure 10 : Le nombre de lectures ayant un score Phred moyen de x.
(Babraham Bioinformatics)

Le graphique de distribution des bases dans les lectures permet de savoir s'il y avait un biais dans la librairie d'inserts (Figure 11). Normalement, chaque base devrait être distribuée similairement pour toutes les positions (Babraham Bioinformatics). De plus, le taux de GC des lectures (donné par un autre graphique non présenté) devrait être similaire à celui estimé pour l'espèce séquencée et/ou aux espèces proches phylogénétiquement.

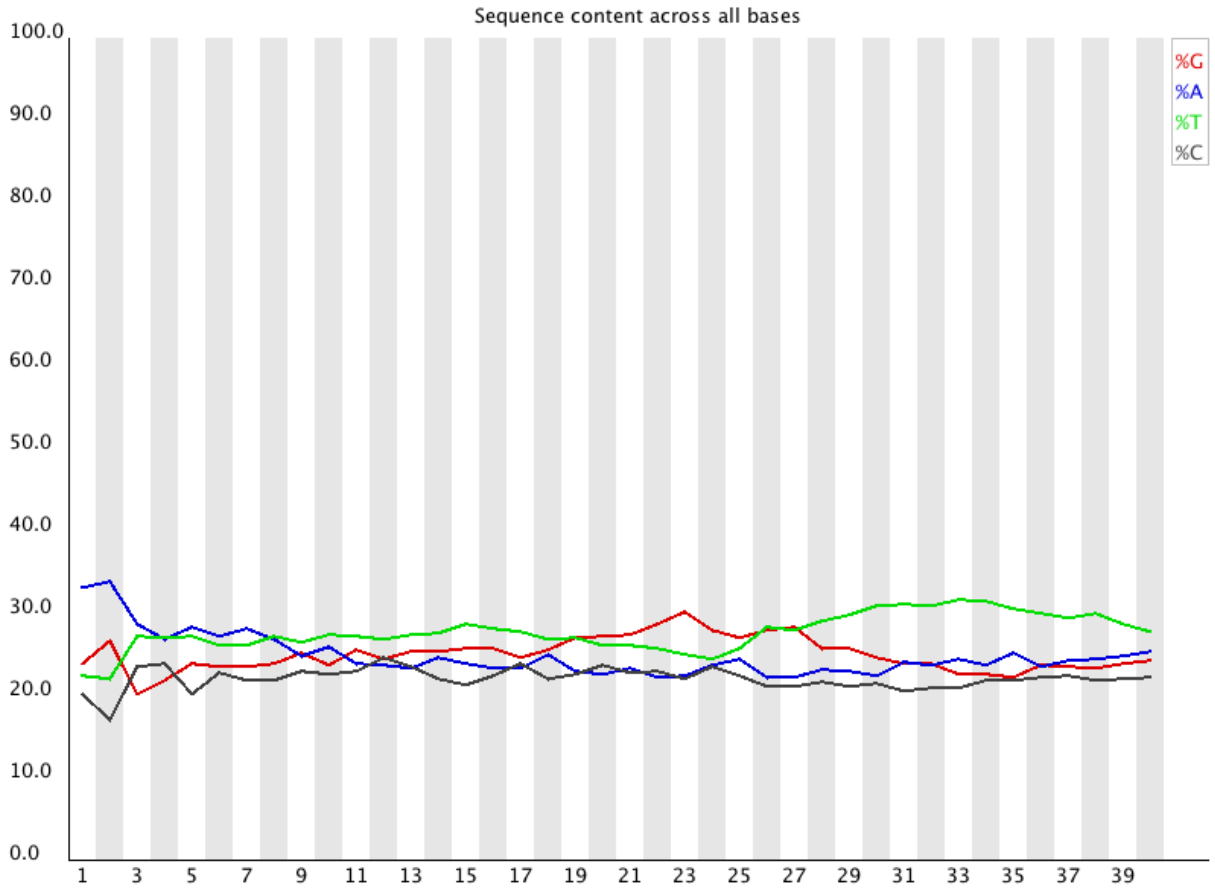


Figure 11 : La répartition des bases selon les positions dans les lectures.
(Babraham Bioinformatics)

La Figure 12 représente le taux duplication des lectures. Afin d'alléger cette analyse, seulement 100 000 lectures sont étudiées. De plus, ce ne sont que les cinquante premières

bases des lectures qui sont utilisées afin que les erreurs de séquençages ne causent pas un biais qui pourraient augmenter la diversité réelle des lectures (Babraham Bioinformatics).

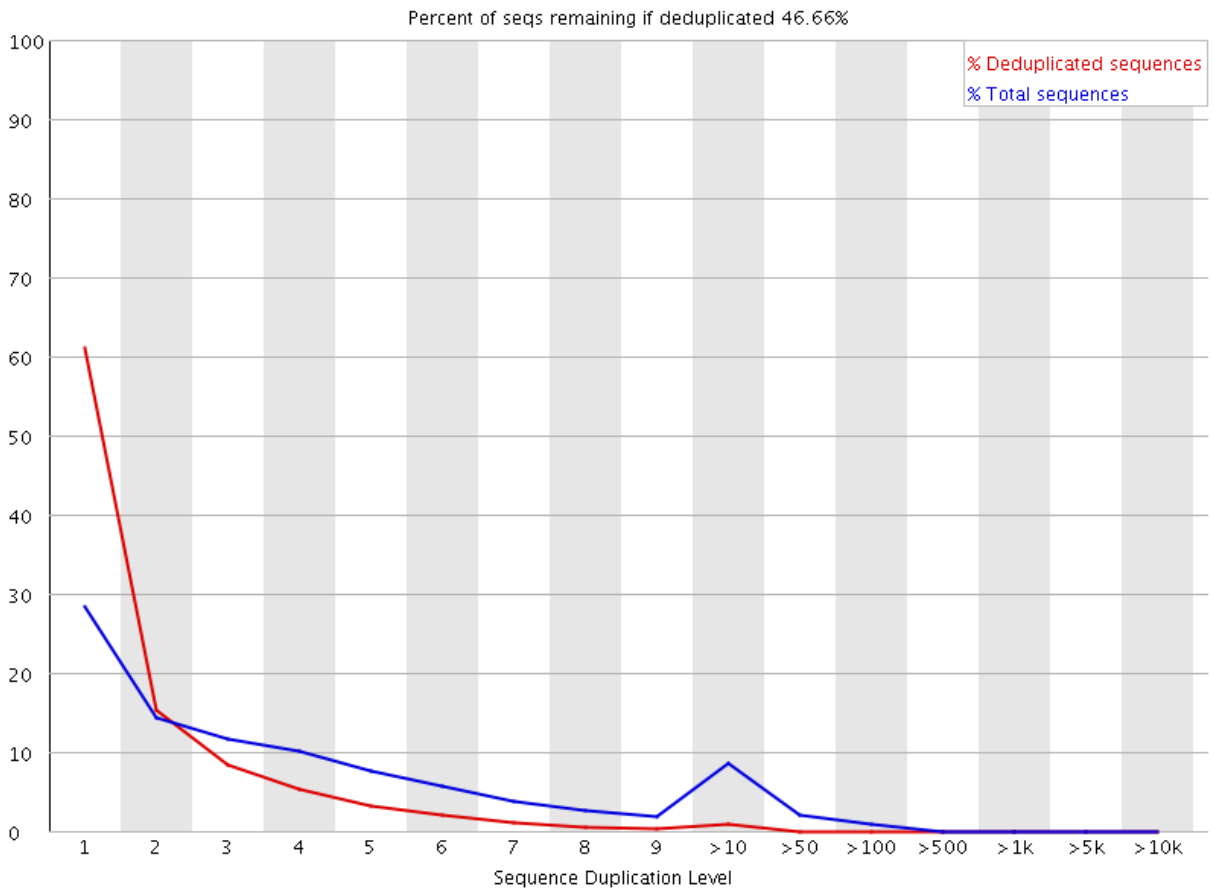


Figure 12 : Le taux de duplication des lectures.
(Babraham Bioinformatics)

D'autres informations sont également données par l'analyse FastQC comme le taux de bases inconnues (N) dans les séquences, la moyenne de longueur des lectures, les séquences surreprésentées et la présence d'adaptateur ou de k-mer (voir la section 1.4.4) qui ne seront pas présentées ici

1.3.3 Évaluation de la qualité d'un assemblage

Afin d'évaluer la qualité de l'assemblage d'un génome, plusieurs points sont à considérer. Le premier est le N50. Il se défini par la longueur à laquelle tous les contigs/échafaudages de cette longueur et plus correspondent à au moins 50% de la longueur totale de tous les contigs/échafaudages de cet assemblage (Bradnam et al., 2013). Un N50 élevé est normalement synonyme d'un bon assemblage, car il indique que beaucoup des contigs/échafaudages construits lors de l'assemblage sont longs et que peu d'entre eux sont courts. Les autres points à considérer sont le nombre de contigs et la quantité de bases couvertes dans cet assemblage. Un bon assemblage a peu de contigs, mais ils sont très longs et ils couvrent idéalement tout le génome. Il est important de considérer ces trois paramètres en parallèle pour bien interpréter la qualité d'un assemblage. Par exemple, un assemblage avec un N50 un peu moins élevé avec un nombre de contigs respectable et une quantité totale élevée de bases est préférable à un assemblage avec un N50 très élevé, mais avec moins de contigs et contenant peu de bases totales. Finalement, il est possible d'estimer quelle est la portion du génome qui a été séquencée lorsque la grosseur de ce dernier est connue. La taille du génome des oiseaux est d'environ 1,1 Gb (Ellegren et al., 2012).

1.3.4 Le séquençage, l'assemblage et l'annotation de génomes d'oiseaux

1.3.4.1 Le séquençage

Récemment, deux espèces d'oiseaux, soit le gobemouche à collier (*Ficedula albicollis*) et la mésange de Hume (*Pseudopodoces humilis)*, ont été séquencées par séquençage à haut débit («second-generation sequencing») (Ellegren et al., 2012; Hunter & Manter, 2012). L'assemblage et l'annotation du génome du gobemouche est depuis peu accessible sur la

banque de données « Ensembl Genome Browser ». Pour ce qui est de la mésange, l'assemblage du génome est disponible sur la banque de données NCBI.

Ces deux oiseaux ont été séquencés avec la technologie HiSeq d'Illumina qui permettait alors d'obtenir des lectures en paire de 150 bases chacune. Dans le cas du gobemouche, la banque d'inserts était composée de fragments d'une longueur de 200 à 5100 bases et celle de la mésange de fragments de 200 à 20 000 bases (Cai et al., 2013; Ellegren et al., 2012).

1.3.4.2 L'assemblage

Les lectures ont d'abord été séparées en différentes catégories (banques) selon la longueur pour faire des assemblages successifs. Le premier s'est fait en sélectionnant seulement les lectures courtes de 200 à 400 pb (séquencés au complet) (voir Figure 7, assemblage 1 seulement). Suite à cet assemblage, un deuxième a été effectué à partir des contigs obtenus via le reste du séquençage (fragments de 400 à 500 pb et les «mate-paired» (Figure 13). Cet assemblage permet de lier les échafaudages obtenus grâce à l'assemblage par les paired-end longs. GAPCLOSER de la suite SOAP a ensuite été utilisé pour fermer les trous avec toutes les lectures (Ellegren et al., 2012).

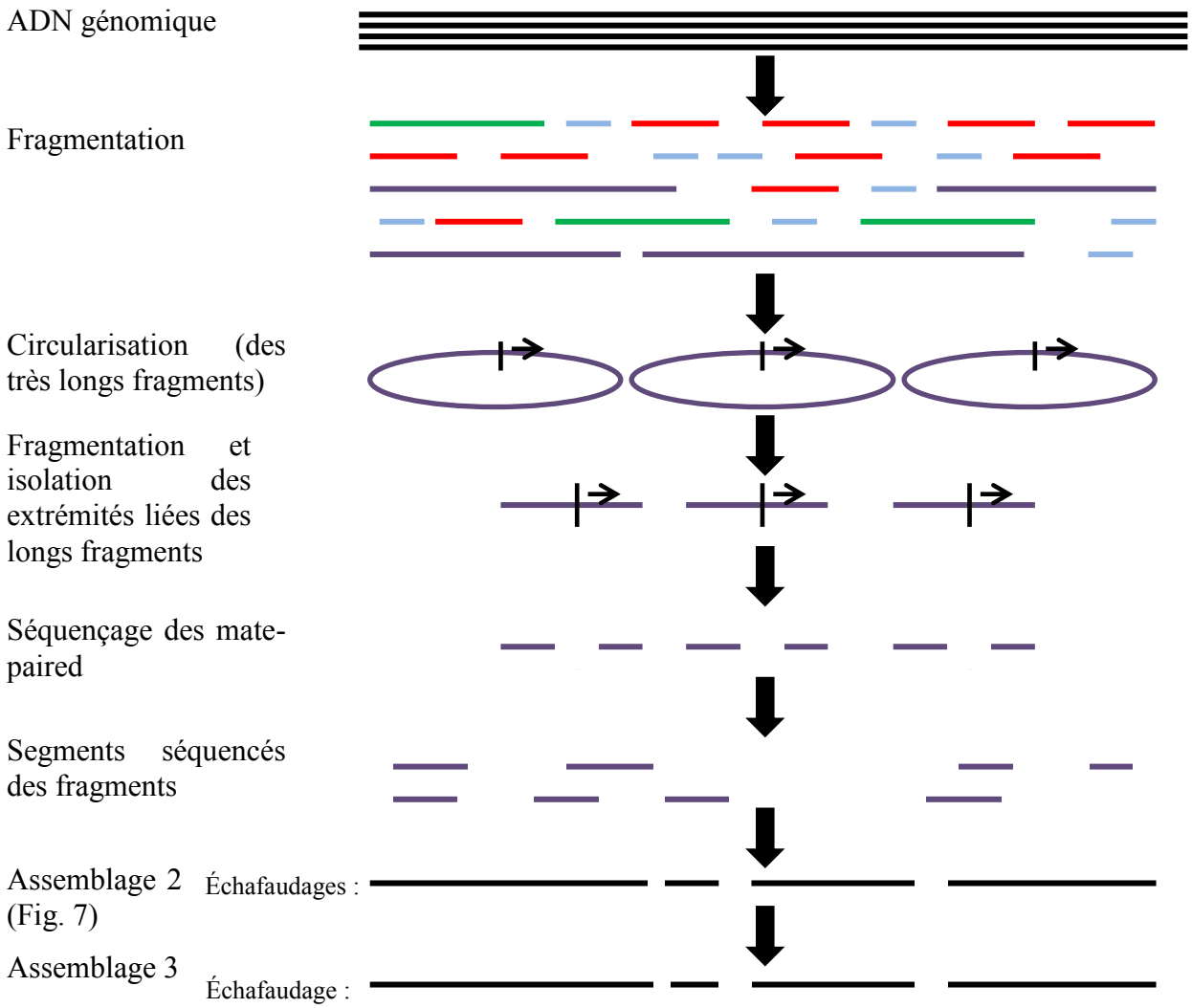


Figure 13 : Étapes sommaires d'un assemblage de génome via des lectures « mate-paired ».
De très longs fragments d'ADN sont circularisés puis fragmentés pour obtenir de petits inserts (composés seulement des deux extrémités de chaque fragment). Chaque insert est séquencé en paire et cela permet de former de plus longs échafaudages à partir des précédents (voir Figure 7).

SOAP est un assembleur de type graphe de Bruijn qui extrait des lectures des fragments d'une longueur déterminée. Ces fragments sont nommés « k-mer » et ils ont une longueur de 13 à 127 bases (à la discrétion de l'utilisateur) (Luo et al., 2012). Ils sont alignés entre eux pour construire un graphique de chevauchement. L'augmentation de la longueur des k-mer permet

de diminuer la probabilité qu'un k-mer soit aligné au mauvais endroit et ainsi engendrer un assemblage erroné (la séquence d'un k-mer plus long est retrouvée moins fréquemment dans un génome qu'un k-mer plus court). Cependant, l'utilisation de k-mer trop longs peut amener la perte de certaines régions sous-représentées dans la banque d'inserts, car les assembleurs du type graphe de Bruijn ne considèrent pas les k-mer moins représentés comparativement à la moyenne parce qu'ils pourraient être dus à une erreur de séquençage. Plus les k-mer sont longs, plus l'assemblage prend de la mémoire. Il existe d'ailleurs une nouvelle version du programme (2.04) qui permet d'utiliser deux éditions de ce dernier, soit la 63mer et la 127mer. Il est possible d'utiliser une longueur maximale des k-mer de 63 et 127 bases respectivement. Par conséquent, la version 127mer utilise deux fois plus de mémoire que la 63 pour des k-mer de même longueur. Pour cette raison, il est préférable d'utiliser la version 63mer quand la longueur des k-mer souhaitée est de 63 bases et moins. Évidemment, le principal paramètre utilisé pour cet assembleur est la longueur des k-mer (`K`). Il n'existe pas de longueur(s) prédéfinie(s) de k-mer pour une espèce et/ou pour un type de séquençage, il faut alors faire plusieurs tests pour obtenir le meilleur assemblage possible (Luo et al., 2012).

Par la suite, les séquences répétées ont été masquées. Comme le génome des oiseaux contient beaucoup moins de séquences répétées comparativement aux humain (environ 7% pour la mésange de Hume et 9% pour le gobemouche à collier), l'assemblage *de novo* est beaucoup moins laborieux (Ellegren et al., 2012; Hunter & Manter, 2012). Pour éliminer la contamination d'ADN provenant d'un autre organisme, 1% de chaque échafaudage été aligné sur la banque de données de NCBI. Tous ceux dont les deux premiers alignements ne faisaient pas partie d'un gène d'oiseaux ou de lézards ont été éliminés de l'assemblage, soit 0,2%. Une dernière étape de validation a ensuite été effectuée via la banque de gènes disponible sur NCBI pour l'espèce étudiée. Cela a permis de déterminer la qualité de l'assemblage du gobemouche, car un gène retrouvé en continu dans un échafaudage est signe d'un bon assemblage quand c'est le cas pour la plupart des gènes de référence. Dans le cas inverse, si

plusieurs gènes sont retrouvés divisés dans plusieurs échafaudages différents, c'est signe d'un assemblage de moins bonne qualité (Ellegren et al., 2012).

En plus du faible pourcentage de séquences répétées, les oiseaux ont un pourcentage de réarrangement plutôt bas et un nombre de chromosomes relativement stable, soit une dizaine de moyens à gros chromosomes (20-200 Mb) avec une trentaine de microchromosomes (10-20 Mb). D'ailleurs, la variation dans le nombre de chromosomes est principalement due à ces derniers (Ellegren et al., 2012). Grâce à ces particularités, l'assemblage final des chromosomes (alignement des échafaudages) du gobemouche et de la mésange s'est basé sur la présence de marqueurs et/ou l'alignement physique sur le génome du diamant mandarin (*Taeniopygia guttata*) et/ou de la poule (*Gallus glallus*). Pour estimer la grosseur des trous entre chaque échafaudage, la banque des mate-paired a été réutilisée (Ellegren et al., 2012; Hunter & Manter, 2012).

Newbler est un autre assembleur qui aurait pu remplacer SOAP qui travail un algorithme différent. Il utilise plutôt des « seeds » (au lieu de k-mer) d'une longueur déterminée par l'utilisateur. Il garde en mémoire leur emplacement dans les lectures et les aligne entre eux. Lorsqu'il y a un alignement entre deux seeds, Newbler rajoute des nucléotides aux extrémités (en revenant à la lecture d'où provient le seed) et vérifie s'ils s'alignent eux aussi et ainsi de suite. Ce type d'approche pour faire un assemblage par consensus (« overlap-layout-consensus ») (Margulies et al., 2005).

Au final, Ellegren et ses collègues ont réussi, avec un taux de couverture génomique de 85X (176 Gb de données générées), à assembler un génome incomplet de 1,13 Gb pour le gobemouche à collier. Le N50 des échafaudages est de 7,3 Mb et 89% de l'assemblage provenant d'environ 200 échafaudages plus gros que 1 Mb (Ellegren et al., 2012). L'équipe de Cai a quant à elle obtenu un génome incomplet de 1,04 Gb (95,4%) pour la mésange de Hume

à partir d'un séquençage ayant une couverture 96X (184,5 Gb de données générées). Le N50 des échafaudages était de 16,3 Mb (Hunter & Manter, 2012). Pour les deux génomes, une partie du caryotype reste inconnu principalement dû à la présence des microchromosomes qui sont plus ardus à assembler.

1.3.4.3 L'annotation du génome

Deux méthodes différentes ont été utilisées pour l'annotation du génome du gobemouche et de la mésange. Elles sont donc présentées séparément.

Pour le gobemouche, plusieurs logiciels ont été utilisés pour trouver les régions codantes *ab initio*. Parmi ceux-ci, MARKER, qui se base sur des programmes de prédictions, AUGUSTUS, qui se base sur les gènes déjà connu du gobemouche (disponible sur GenBank), GENEMARK, qui utilise les marqueurs de séquences exprimés (EST) (provenant du ARN-Seq du gobemouche) et SNAP, qui se base sur les gènes déjà trouvés par les autres logiciels. Finalement, MARKER a été réutilisé en incluant les prédictions des programmes précédents. Le taux de GC est d'environ 39% et 18 735 gènes codants potentiels ont été identifiés dont 99,5% ont été validé par RNA-Seq (Ellegren et al., 2012).

Pour la mésange, l'approche s'est basée sur les EST (marqueurs de séquences exprimées), les prédictions *ab initio* et l'alignement de gènes déjà identifiés chez d'autres espèces plus ou moins proches (le diamant mandarin, la poule, le lézard (*Pogona vitticeps*) et l'humain (*Homo sapiens*)). Dans ce cas-ci, les programmes TBLASTN/genBlastA/Genewise (qui utilisent les gènes d'autres espèces pour trouver des gènes homologues par alignement), BLAT/AUGUSTUS (qui utilisent les EST) et Genscan (qui utilise les paramètres créés pour identifier les gènes chez l'humain) ont été utilisés. L'annotation finale pour la mésange s'est faite via

InterProScan et une gamme de programmes contenant des banques de données sur les domaines et les motifs des familles de protéines. Le taux de GC obtenu est de 41,7%, ce qui est similaire aux autres oiseaux. Ils ont identifié 16 998 gènes codants potentiels dont 98,9% ont été validés par BLAST (Hunter & Manter, 2012).

1.4 Le projet de recherche

Quelques articles ont déjà démontré l'effet de certains pesticides *in vitro* et *in vivo*. Cependant, ces études ont été faites en laboratoire, c'est-à-dire en conditions contrôlées pour tester (Liu et al., 2013; Velisek, Kouba, & Stara, 2013). Dans le projet présenté ici, l'effet des pesticides sera observé chez des organismes vivants en milieu naturel. Ils sont donc naturellement exposés à plusieurs pesticides simultanément et à des concentrations variables.

Les deux espèces qui ont été sélectionnées sont des insectivores afin d'avoir des modèles animaux étroitement en contact avec les pesticides, car ceux-ci s'accumulent dans les lipides et que les insectes en sont une excellente source. L'hirondelle bicolore (*Tachycineta bicolor*) a été sélectionnée parce qu'elle vit dans les milieux agricoles et que plusieurs laboratoires à l'Université de Sherbrooke l'étudient. Il était donc facile d'avoir accès aux spécimens retrouvés sur le terrain en plus d'avoir un suivi individuel détaillé de la dynamique de population de cette espèce. D'après les données récoltées au cours de la dernière décennie dans la région de l'Estrie et ses environs (au Québec), cette espèce est en déclin. En effet, la masse moyenne des mâles a diminué d'environ 2%, celle des femelles de 8% et le nombre d'individus dans la population a diminué de 4% par année (données récoltées pour les années 2005-2011) (Rioux Paquette, Pelletier, Garant, & Bélisle, 2014). Évidemment, l'hirondelle bicolore est un oiseau migrateur et les individus nichant dans la région de l'Estrie passent les mois de septembre à avril majoritairement en Floride et possiblement un peu au Mexique. Aucun moyen n'est à notre disposition pour savoir si elles sont exposées ou non à des

pesticides durant cette période de l'année ni, dans un tel cas, auxquels et à quelle(s) concentration(s). Un deuxième modèle d'étude sédentaire, soit la grande musaraigne (*Blarina brevicauda),* a donc été sélectionné pour soutenir les résultats obtenus chez l'hirondelle. La musaraigne est un petit mammifère insectivore. Elle a été choisie parce qu'elle est une espèce facilement capturable à l'aide de pièges pour micromammifères et qu'elle est retrouvée de manière ubiquitaire au Québec. C'est d'ailleurs l'espèce de musaraigne majoritaire au Québec. Il a été facile de déterminer si l'animal avait été exposé ou non à des pesticides selon l'endroit où il a été capturé, car cette espèce a un petit habitat. Comme cette espèce vit près des cours d'eau, elle devrait être exposée à tous les polluants qui se retrouvent dans ces milieux suite au ruissellement des champs agricoles. Mon étude ne s'est pas seulement concentrée sur la grande musaraigne, car, contrairement à l'hirondelle, il existe très peu de données sur l'état des populations de cette espèce.

1.4.1 Les objectifs globaux

L'objectif 1 du projet est de s'assurer que les pesticides présents dans l'environnement sont en concentration suffisante pour induire des changements chez les modèles animaux sauvages. Pour faire cette vérification, le degré d'activation d'AhR via le niveau d'expression des gènes *CYP1* doit être quantifié par des réactions en chaîne par polymérase quantitative (qPCR). Après cette validation, il sera possible de procéder à l'objectif 2 qui permettra d'évaluer les effets génétiques et épigénétiques des pesticides sur les mêmes modèles. Un aspect à étudier sera la quantification de l'expression de tous les gènes chez les modèles en présence ou non de pesticides via un séquençage global d'ARN (RNA-Seq). Par la suite, le séquençage de la chromatine immunoprécipitée (ChIP-Seq) permettra d'identifier les régions avec lesquelles des protéines ayant un rôle dans l'expression ou l'inhibition des gènes interagissent. Finalement, les régions plus (ou moins) méthylées de l'ADN après à une exposition aux pesticides pourront être déterminées par un séquençage de l'ADN méthylé immunoprécipité (MeDIP-Seq). Cette étape va elle aussi servir à identifier les gènes actifs ou inhibés, car

l'ADN méthylé est normalement synonyme de régions qui sont plus condensées, donc moins exprimées. Finalement, toutes les informations obtenues par ces trois techniques vont pouvoir être analysées et se valider entre elles en plus d'apporter de l'information sur les gènes qui sont importants lors d'une exposition à des pesticides et sur la manière dont ils sont régulés.

Pour parvenir à l'objectif 1 (via qPCR), la séquence des gènes *AhR* et *CYP1* doit être connue et, pour l'objectif 2, la séquence du génome des espèces étudiées, ou d'une espèce phylogénétiquement proche d'elles, est nécessaire. Or, le génome des deux espèces sélectionnées n'est pas séquencé. Il aurait été plus simple d'utiliser la musaraigne commune comme modèle à la place de la grande musaraigne, car, contrairement à cette dernière, son génome a été séquencé. Or, elle vit en Europe. Heureusement, ces deux espèces sont très proche phylogénétiquement, il a donc été possible d'utiliser le génome de la musaraigne commune comme référence pour les différentes expériences du projet concernant ce mammifère. Dans le cas de l'hirondelle bicolore, les espèces d'oiseaux actuellement séquencées sont trop éloignées phylogénétiquement d'elle pour permettre une excellente identification des différentes régions du génome. Donc, un autre but du projet concerne le séquençage, l'assemblage et l'annotation du génome de l'hirondelle bicolore. Sommairement, l'objectif global poursuivi par le projet est d'évaluer l'effet des pesticides de manière globale, c'est-à-dire au niveau génétique et épigénétique pour identifier certains gènes, familles de gènes et/ou voies métaboliques importantes impliqués directement ou indirectement dans les effets toxiques des pesticides sur les organismes vivants.

1.4.2 Les objectifs spécifiques de maîtrise

Le projet a débuté par la préparation de l'objectif 1, soit le clonage d'une portion des gènes *AhR, CYP1A1, CYP1A4, CYP1A5* et *CYP1B1* de l'hirondelle bicolore et/ou de la grande

musaraigne. Suite à la validation que la séquence amplifiée était celle du gène désiré, la préparation des amorces pour le qPCR a été amorcée.

Après les difficultés à cloner certains gènes chez l'hirondelle bicolore et en considérant que le génome de l'hirondelle bicolore pourrait être utile pour plusieurs étapes du projet (en particulier pour l'objectif 2) et pour la communauté scientifique, il a été décidé que le séquençage de cette espèce était nécessaire. Évidemment, après le séquençage d'un génome, son assemblage de même que son annotation sont à faire. Ces deux premières étapes se sont basées principalement sur ce qui avait été fait pour le génome du gobemouche à collier et un peu pour celui de la mésange de Hume (Cai et al., 2013; Ellegren et al., 2012).

CHAPITRE 2

DÉVELOPPEMENT

2.1 Matériels et méthodes

2.1.1 Les échantillons

Les hirondelles bicolores qui ont été utilisées dans le cadre du projet ont toutes, à une exception près, été retrouvées mortes sur l'aire d'étude des professeurs Fanie Pelletier, Marc Bélisle et Dany Garant de l'Université de Sherbrooke du département de biologie (Figure 14 et Tableau 1). Elles ont succombé à des blessures produites par d'autres oiseaux pour défendre leur nichoir avant et/ou au début de la saison de reproduction. Pour ce qui est des musaraignes, elles proviennent de quelques fermes ciblées sur la même zone d'étude (Figure 14 et Tableau 2). Elles ont été piégées à l'aide de cage pour micromammifères de marque Sherman (9x3x3 pouces). Les analgésiques et les produits chimiques utilisés pour pratiquer l'euthanasie provoquent des changements métaboliques résultant à l'endormissement et la mort respectivement. En ce moment, aucune étude n'a étudié l'effet de ces produits xénobiotiques sur le récepteur AhR qui a une très grande variété de ligands. Pour s'assurer de la fiabilité des résultats, l'euthanasie a plutôt été faite de manière physique par dislocation cervicale sans l'administration analgésique.

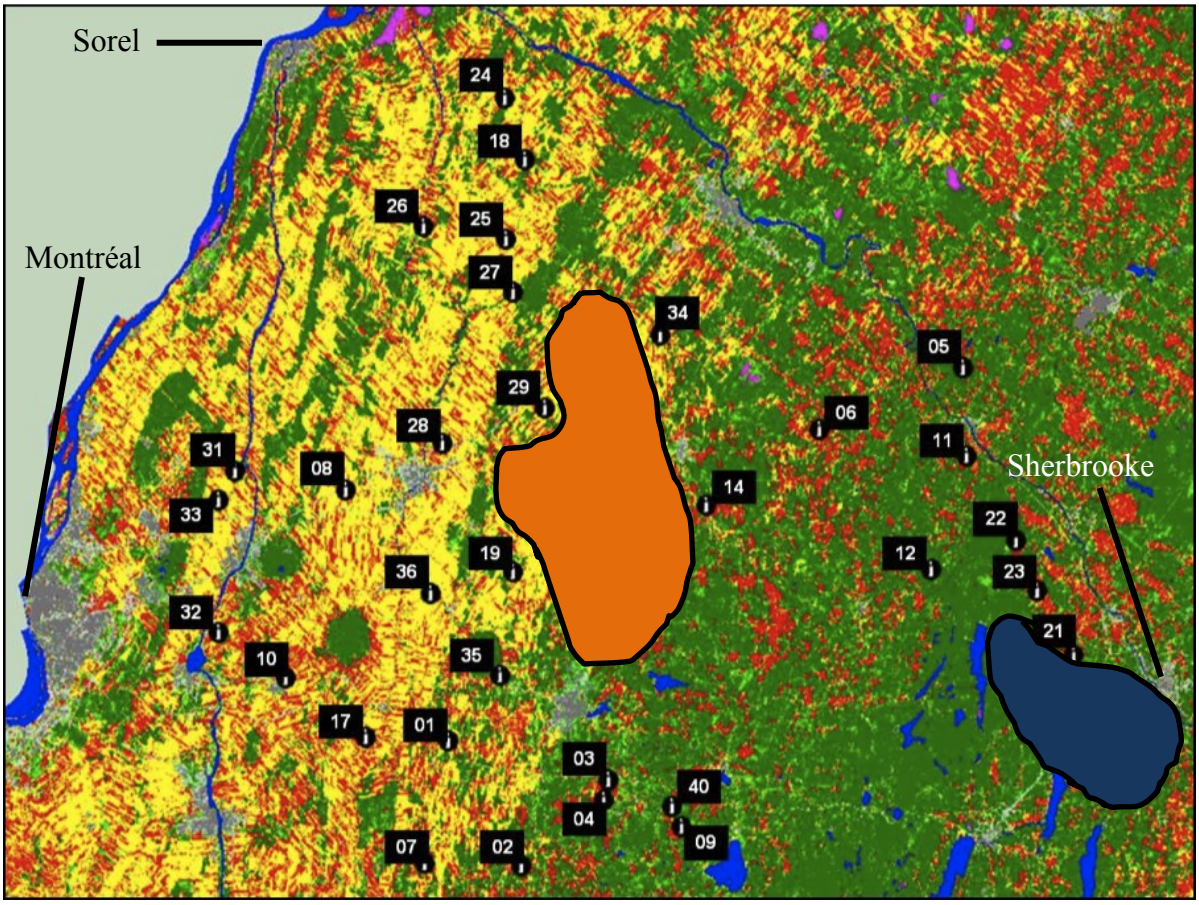


Figure 14 : Aire d'échantillonnage de l'hirondelle bicolore et de la grande musaraigne.

Représentation des 40 fermes de la région de l'Estrie qui sont étudiées par trois laboratoires à l'Université de Sherbrooke. En jaune : agriculture intensive (céréales); en rouge : agriculture extensive (animaux d'élevage, peu de pesticides); en vert : forêts. La surface orange est la zone d'échantillonnage de la musaraigne pour la région avec pesticides et la surface en bleu est la zone d'échantillonnage de la musaraigne pour la région sans pesticides (voir le Tableau 2). Les hirondelles proviennent des différentes fermes de cette carte (se référer au Tableau 1).

Tableau 1 : Détails sur la provenance des hirondelles bicolores et sur la qualité de l'ARN obtenu.

	Information sur l'hirondelle et son milieu					Information sur la manipulation de l'hirondelle		
No.	Récolte	Nichoir	Zone	Sexe	Bague	Prélèvement	Extraction d'ARN	État ARN
1	?	2909	P	F	2351-44306	11/04/13	12/04/13 16/04/13	Dégradé Dégradé
2	29/04/13	2009	P	?	-	30/04/13 4 °C, présence de vers	06/05/13	Dégradé
3	01/05/13	0707	P	F	-	02/05/13 4 °C	06/05/13 08/05/13	Dégradé Dégradé
4	02/05/13	3108	P	F	2351-42718	- Trop de vers	-	-
5	03/05/13	2702	P	F	-	06/05/13 4 °C, mangée	-	-
6	06/05/13	3104	P	F	-	06/05/13 4 °C	-	-
7	07/05/13	2909	P	F	2511-70371	08/05/13	08/05/13 15/05/13 03/06/13	Dégradé Partiel Partiel
8	07/05/13	2909	P	M	-	08/05/13	08/05/13 15/05/13 03/06/13 22/10/13	Dégradé Partiel Partiel Partiel
9	04/05/14	0210	P/CT	?	-	06/05/14 (5 jours à 4 °C)	?	Partiel moyen
10	04/05/14	2401	P	M	-	06/05/14	16/12/14	Partiel moyen
11	01/05/14	3005	P	M	2591-60644	06/05/14	18/06/14	Partiel
12	05/05/14	1905	P	F	-	06/05/14	16/12/14	Partiel
13	18/06/14	St-Denis	CT	M	-	19/06/14	16/12/14	Intact
14	19/05/14	2014	P	F	2311-95922	22/07/14	13/08/14	Partiel
15	14/06/14	3909	P	M	2351-44209	22/07/14	30/07/14	Intact
16	15/06/14	3004	P	F	-	22/07/14	23/07/14	Intact
17	24/05/14	2202	CT	F	2511-70717	22/07/14	30/07/14	Intact
18	19/05/14	2703	P	F	-	22/07/14	13/08/14	Partiel
19	16/06/14	3004	P	F	-	22/07/14	13/08/14	Dégradé

20	09/06/14	1010	P	F	2591-60534	22/07/14	23/07/14	Intact

Identification de chaque hirondelle bicolore récupérée sur le terrain avec la date à laquelle elle a été trouvée; à quel nichoir (XXYY; où XX représente le numéro de la ferme et YY représente le numéro du nichoir sur cette ferme); la caractérisation de la zone (contrôle [CT] ou avec pesticides [P]); le sexe et le numéro de bague de l'oiseau (si présente). Dans la partie manipulation, il y a les informations sur la date de prélèvement du tissu musculaire (fait sur une hirondelle conservée à -20ºC à moins d'une autre indication) et la température à laquelle l'hirondelle a été conservée avant et pendant le prélèvement; la date d'extraction de l'ARN et la qualité de ce dernier.

Tableau 2 : Détails sur la provenance des grandes musaraignes et sur la qualité de l'ARN obtenu.

	Information sur l'hirondelle et son milieu à la capture				Information sur la manipulation de l'hirondelle		
No.	Récolte	Ferme	Zone	État	Prélèvement	Extraction d'ARN	État ARN
1	14/03/13	St-Denis	CT	Morte	25/06/13	25/06/13	Partiel
2	01/12/13	St-Denis	CT	Vivante	06/01/14	06/01/14 10/06/14 18/06/14	Intact Partiel Intact
3	31/07/14	13	CT	Vivante	01/08/14	04/08/14 07/08/14	Intact Intact
4	31/07/14	13	CT	Vivante	01/08/14	04/08/14 07/08/14	Partiel* Intact
5	31/07/14	13	CT	Morte, fluides	01/08/14	04/08/14 07/08/14	Partiel* Intact
6	31/07/14	13	CT	Morte	01/08/14	04/08/14 07/08/14	Partiel* Intact
7	31/07/14	13	CT	Morte	01/08/14	04/08/14 07/08/14	Partiel* Intact
8	09/08/14	13	CT	Vivante	11/08/14	13/08/14	Intact
9	08/08/14	13	CT	Vivante	11/08/14	13/08/14	Intact
10	09/08/14	13	CT	Morte	11/08/14	13/08/14	Intact
11	09/08/14	13	CT	Morte, fluides	11/08/14	13/08/14	Intact
12	09/08/14	13	CT	Morte	11/08/14	13/08/14	Intact
13	09/08/14	13	CT	Morte	11/08/14	13/08/14	Intact

14	09/08/14	13	CT	Morte	11/08/14	13/08/14	Intact
15	10/08/14	13	CT	Morte	11/08/14	13/08/14 09/10/14	Partiel, non utilisé Intact
16	10/08/14	13	CT	Morte	11/08/14	13/08/14	Intact
17	30/08/14	JV	CT	Morte	03/09/14	04/09/14	Partiellement dégradé
18	30/08/14	JV	CT	Morte	03/09/14	04/09/14	Partiellement dégradé
19	03/09/14	St-Denis	CT	Morte	04/09/14	04/09/14	Intact
20	24/09/14	19	P	Vivante	29/09/14 – 30/09/14	08/10/14	Intact
21	25/09/14	19	P	Morte	29/09/14 – 30/09/14	08/10/14	Partiellement dégradé
22	26/09/14	15	P	Morte	29/09/14 – 30/09/14	08/10/14	Intact
23	25/09/14	19	P	Morte	29/09/14 – 02/10/14	08/10/14	Partiellement dégradé
24	25/09/13	20	P	Morte, raide	29/09/14 – 02/10/14	08/10/14	Intact
25	25/09/14	37	P	Morte	29/09/14 – 02/10/14	08/10/14	Intact
26	25/09/14	30	P	Morte	29/09/14 – 02/10/14	08/10/14	Intact
27	26/09/14	15	P	Vivante	29/09/14 – 02/10/14	08/10/14	Intact
28	25/09/14	39	P	Morte	29/09/14 – 02/10/14	09/10/14	Intact
29	25/09/14	37	P	Morte	29/09/14 – 02/10/14	09/10/14	Intact
30	25/09/14	16	P	Morte	29/09/14 – 03/10/14	09/10/14	Intact
31	25/09/14	16	P	Morte	29/09/14 – 03/10/14	09/10/14	Intact
32	24/09/14	39	P	Vivante	29/09/14 – 03/10/14	09/10/14	Intact
33	24/09/14	20	P	Morte	29/09/14 – 03/10/14	09/10/14	Intact
34	24/09/14	19	P	Morte, raide	29/09/14 – 03/10/14	09/10/14	Intact
35	25/09/14	16	P	Morte	29/09/14 – 03/10/14	09/10/14	Intact
36	24/09/14	30	P	Morte	29/09/14 –	09/10/14	Intact

					03/10/14		

Identification de chaque grande musaraigne avec la date à laquelle elle a été capturée; à quelle ferme; la caractérisation de la zone (contrôle [CT] ou avec pesticides [P]) et si la musaraigne a été retrouvée vivante ou morte. Dans la partie manipulation, il y a des informations sur la date de prélèvement du tissu musculaire (fait sur une musaraigne conservée à -20ºC); la date d'extraction de l'ARN et la qualité de ce dernier. *Une seule bande était visible lors de la migration de l'ARN sur gel d'agarose 1,5% (normalement, il y en a deux biens distinctes).

2.1.2 Traitement des échantillons

La récolte de tissus musculaires de poitrine chez l'hirondelle bicolore s'est faite soit quand elle avait été conservée à 4 °C ou à -20 °C (Tableau 1). Le prélèvement des tissus musculaires des musaraignes a toujours été fait quand elles étaient congelées à -20 °C (Tableau 2). Par la suite, tous les échantillons étaient immédiatement gelés dans l'azote liquide et conservés à -80 °C jusqu'à la prochaine étape, peu importe la méthode de préservation avant le prélèvement. Les tissus prélevés ont été broyés en une fine poudre à l'aide d'un mortier et d'un pilon préalablement lavé au savon et à l'éthanol puis refroidi à l'azote liquide.

2.1.2.1 Extraction d'ARN pour le clonage

Des aliquotes d'environ 50 mg de tissu moulus ont été répartis dans des tubes de 1,5 mL. 500 µL de Trizol ont été ajoutés à la poudre, mélangés à l'aide d'un vortex et écrasés avec de petits bâtons épousant la forme du fond des tubes pour 2-3 min (en alternance avec le vortex). Par la suite, les échantillons ont été centrifugés à vitesse maximale pour une minute et le surnageant a été récolté en ne touchant pas au culot. Après avoir réajusté le volume total à 500 µL avec du Trizol, 100 µL de chloroforme ont été ajoutés au surnageant. Le tout a été

mélangé 15 secs à intensité moyenne et incubé 2 min à température de la pièce (TP) avant d'être centrifugé 15 min à 12 000 g à 4 °C pour récupérer le surnageant en prenant soin de ne pas récupérer de l'interface (lipides et protéines). Ensuite, 100 µL de chloroforme ont été ajoutés à ce surnageant puis mélangé et centrifugé dans les mêmes conditions que l'étape précédente. Pour faire précipiter l'ARN, le surnageant récupéré (~300 µL) a été mélangé avec 200 µL d'isopropanol et incubé 10 min à TP avant d'être centrifugé 10 min à 12 000 g à 4 °C. Le culot d'ARN obtenu a été lavé avec 900 µL d'éthanol 75% suivit d'une centrifugation de 5 min à 7500 g à 4 °C trois fois pour éliminer toutes traces de phénol. Il a ensuite été séché à l'air avant d'être suspendu dans 20 µL d'eau moléculaire. L'intégrité des ARN obtenus a été vérifiée par migration sur gel d'agarose 1,5% avant de les conserver à -80 °C.

Après la validation de l'intégrité de l'ARN, les échantillons ont subi un traitement à la « DNase » via la trousse « RQ1 RNase-Free DNase » de Promega. Les réactions ont été faites dans un rapport 1 µg d'ARN : 1 unité de DNase. La plupart des réactions ont été faites dans un volume de 50 µL avec 15 µg d'ARN (sinon le même ratio a été respecté). Pour le reste du protocole, les spécifications du détaillant ont été suivies. Ces ARN, sans contamination d'ADN, ont été conservés à -80 °C.

Finalement, l'ADN complémentaire (ADNc) a été produit en effectuant une réverse-transcription suivie d'une réaction en chaîne par polymérase (RT-PCR). Pour chaque réaction, 1 µg d'ARN a été mélangé avec 0,5 µg d'hexamères aléatoires dans un volume total de 11 µL. Ce mélange a été incubé à 70 °C pendant 5 min avant que la température descende à 4 °C et que 14 µL d'une solution contenant 0,5 mM de dNTP, 0,5 µL de la réverse transcriptase (M-MuLV maison) et le tampon de cette enzyme soit ajoutée. Ensuite, le mélange a passé une heure à 42 °C puis quinze minutes à 85 °C. L'ADNc produit a été conservé à -20 °C.

2.1.2.2 Extraction de l'ADN pour le séquençage du génome de l'hirondelle bicolore

L'ADN de la première hirondelle séquencé provient du tissu musculaire d'un mâle qui a été retrouvé mort à la ferme 19 au nichoir 9 dans la région de l'Estrie (n. 8, voir Tableau 1). Il a été extrait à partir de 26 mg de tissus réduits en poudre dans l'azote liquide. L'extraction a été faite par le kit Quick-gDNA MiniPrep de Zymo Reseach en suivant les instructions du manufacturier.

2.1.3 Clonage des gènes cibles *AhR* et *CYP*

Comme les deux modèles d'étude ne sont pas des espèces dont le génome est disponible dans la littérature, l'étape suivante a été de cloner les gènes d'intérêts des deux modèles. Pour ce faire, les gènes homologues d'espèces proches phylogénétiquement de l'hirondelle bicolore, soit *Ficedulla albicollis* (gobe-mouche à collier), *Taeniogypia guttata* (diamant mandarin) et *Gallus gallus* (poule domestique), ou de la grande musaraigne, soit *Sorex araneus* (musaraigne commune) et *Tupaia belangeri* (toupaye de Belanger), ont été utilisés. Les séquences de ces gènes homologues ont été alignées entre elles à l'aide du logiciel GENtle, puis des amorces ont été conçues dans les régions les plus conservées pour amplifier par PCR une portion du gène désiré d'un des deux modèles (Tableau 3 et Tableau 4). Dû à la difficulté de cloner *CYP1B1* chez la musaraigne, plusieurs gènes homologues d'autres petits mammifères provenant de la banque d'« Ensembl Genome Browser » ont été ajoutés lors de l'alignement pour sélectionner les régions les plus conservées. Pour faciliter le clonage de ce gène, des amorces dégénérées ont aussi été utilisées avec un maximum de trois positions de nucléotide variables.

Tableau 3 : Amorces utilisées pour cloner les gènes *AhR*, *CYP1B1* et les gènes contrôles chez l'hirondelle bicolore.

Gènes	Amorce sens	Amorce antisens
AhR	5-Tachy-AHR1 GAATTCTTCCTTTATGGAAAGG	3-Tachy-AHR1 GGCAACTTCATGTTAGCTT
CYP1B1	5-Tchy_cCYP1B1-2 CACCGTCACCGACATCTTC	3-Tchy_cCYP1B1-2 TCCTCATTTGGATTAGCAGT
RPLP0	5-Tchy_RPLPO-1 GACAGGGCTACGTGGAAGTC	3-Tchy_RPLPO-1 GTCCAGCACTTCAGGGTTGT
EIF1	5-Tchy_EIF1-1 CCGGCACTGAGGACTACATC	3-Tchy_EIF1-1 TGGACTTTCAGCTGGTCGTC
EIF4A2	5-Tchy_EIF4A2-1 CCGCGGATTATAGCAGAGAC	3-Tchy_EIF4A2-1 ATAACATCCCGCTCCTTCTG

Après la réaction par PCR, le produit amplifié a migré sur gel d'agarose 1 à 2% et les bandes correspondant à la taille attendue (±100 pb) ont ensuite été purifiées par la trousse « QIAEX II Gel extraction Kit » produite par QIAGEN. Le produit purifié a ensuite été inséré dans le plasmide « pGEM(R)-T Easy » de Promega et transformé dans les bactéries MM294 par choc thermique à 42 °C. Les bactéries transformées ont été étalées sur milieu LB avec ampicilline et incubées toute la nuit à 37 °C. Le lendemain, des minipréparations de plasmides ont été faites sur quelques colonies sélectionnées. La validation que l'amplicon se trouvait bien dans les plasmides de ces clones a été faite par une double digestion avec les enzymes de restriction NdeI et SacII à 37ºC. Les clones positifs ont été séquencés par la plateforme Nanuq du Centre d'innovation de Génome Québec et de l'Université McGill. Pour valider que la séquence clonée chez l'organisme modèle était celle du gène désiré (*AhR* ou *CYP*), la séquence obtenue par le séquençage a été alignée sur la banque d'ADNc des animaux mentionnés précédemment (les trois oiseaux pour l'hirondelle ou les deux mammifères pour la musaraigne) via l'outil BLASTn d'« Ensembl Genome Browser ». Pour confirmer que la séquence était la bonne,

le(s) premier(s) résultat(s) devrait(ent) correspondre à la partie du gène cible hypothétiquement clonée, avoir un score élevé (c'est-à-dire une similarité élevée sur une certaine longueur) et un « e-value » faible. Le « e-value » indique le nombre d'alignement qu'une séquence de la même longueur que celle testée obtienne le même score dans une banque de données de taille équivalente à celle utilisée et composée de séquence aléatoire.

Tableau 4 : Amorces utilisées pour cloner les gènes *AhR*, *CYP1A1* et les gènes contrôles chez la grande musaraigne.

Gènes	Amorce sens	Amorce antisens
AhR	3-Racn_cAhR-1 TACAGAGTTGGACCGTTTGG	3-Shrew_cAhR-2 ATCATGCCACTCTCTCCAGTC
CYP1A1	3-Sara_CYP1A1-1 GATGGCCAGGAAAAGGAAG	5-Sara_CYP1A1-1 GCCCTGAAGAGTTTTGCC
RPLP0	5-Shrw_RPLPO-1 CTGATGGGCAAGAACACCA	3-Shrw_RPLPO-1 AAGCCTGGAAGAAGGAGGTC
EIF1	5-Shrw_EIF1-1 CGCTATCCAGAACCTCCACT	3-Shrw_EIF1-1 ATATGTTCTTGCGCTGGTCAC
EIF4A2	5-Shrw_EIF4A2-1 GTTCAAGGAGACCCAAGCAC	3-Shrw_EIF4A2-1 TTTCCTTCTGGTCCATGTCA

2.1.4 Préparation de la quantification de l'expression des gènes

À l'aide de la séquence obtenue des gènes cibles, des amorces ont été conçues pour faire des qPCR en respectant certaines conditions, soit une température de séparation de l'ADN double brin (Tm) située entre 57 et 63 °C, la longueur des amorces entre 18 et 23 bases, la longueur de l'amplicon entre 80 et 150 bases et aucune formation de structures secondaires dans les amorces. Au moins deux couples d'amorces ont été testés pour chaque gène pour augmenter la

probabilité d'en obtenir un satisfaisant. Les couples ont été testés pour déterminer leur température optimale et pour mesurer leur capacité à lier leur séquence complémentaire de manière spécifique et répétable via une courbe standard. Les amorces sélectionnées ont obtenu un taux d'efficacité entre 80 et 110% (Annexe 1 : Tableau 11 et Tableau 12).

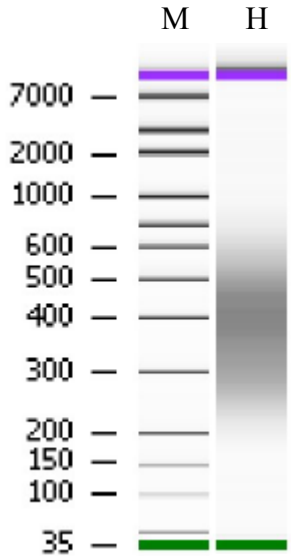


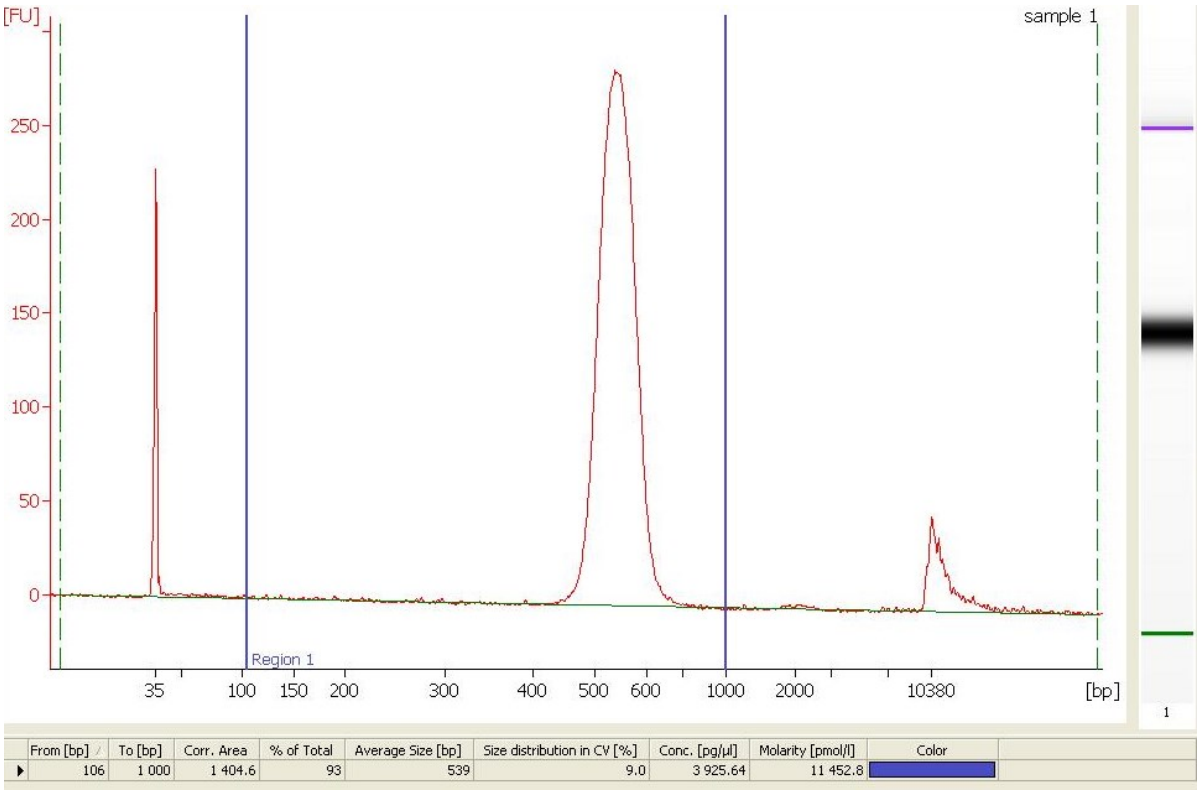
Figure 15 : La taille des inserts du premier séquençage suite à la sonication de l'ADNg. Longueur des fragments d'ADN de la première hirondelle séquencée après trente cycles de sonication (30 sec ON/30 sec OFF) à intensité moyenne. Le premier puit représente la taille des marqueurs et le puit H donne la taille des fragments.

La quantification de l'expression de gène(s) contrôle(s) en même temps que celle des gènes cibles est nécessaire pour valider les expériences de qPCR. Aucun gène contrôle n'a encore été identifié, à notre connaissance, comme de bons contrôles chez ces deux espèces dans les conditions à tester (soit en relation avec des pesticides). Les gènes qui ont été choisis sont impliqués dans la transcription de base. Ils devraient, théoriquement, être exprimés de manière égale en présence ou non de pesticides. Normalement, un seul gène contrôle est suffisant pour valider un qPCR. Cependant, dans l'étude présente, il a été jugé que trois gènes seraient nécessaires au départ pour s'assurer d'en avoir au moins un bon après toutes les étapes de vérification. Par exemple, le clonage peut s'avérer difficile pour un gène, les tests faits pour

trouver un couple d'amorces pour le qPCR peuvent échouer pour la partie du gène qui aura été amplifiée et il se peut qu'au final un gène jugé « contrôle » soit influencé par la présence de pesticides. Les trois gènes contrôle sélectionnés sont *RPLP0, EIF1* et *EIF4A2*. RPLP0 est une protéine faisant partie de la sous-unité 60S des ribosomes (Rich & Steitz, 1987). Les deux autres protéines sont des facteurs d'initiation de l'élongation chez les eucaryotes. Ils aident à la liaison entre le ribosome et l'ARNm en favorisant le reconnaissance de la coiffe sur ce dernier (Fletcher, Pestova, Hellen, & Wagner, 1999; Meijer et al., 2013)

2.1.5 Préparation de la banque d'inserts pour le séquençage génomique de *Tachycineta bicolor*

L'ADN extrait (voir la section 2.1.2.2) a été soniqué à intensité moyenne pour 30 secs suivit d'une pause de durée égale, et ce pour trente cycles (Figure 15). Cette étape a permis d'obtenir des fragments, nommés inserts, de différentes longueurs. La trousse « SPARK DNA Sample Prep » d'Enzymatics a été utilisée avec les « YIGA adaptors » pour préparer la banque d'inserts pour le séquençage. Ces adaptateurs, qui se lient aux extrémités des inserts, ont une structure en Y non symétrique (longueur de 20 et 33 pb) qui permet de diminuer la probabilité de formation de concatémères (liaison entre plusieurs adaptateurs et/ou inserts). La librairie a été amplifiée par qPCR avec les amorces « IGA-PCR-PE-F » et « TruSeq-b01 » (60 pb chaque) qui ont une région complémentaire avec les adaptateurs YIGA. Les fragments (adaptateurs + inserts) de 520 à 680 pb ont été sélectionnés par migration sur gel blue pippin 1,5%. La taille des fragments a ensuite été analysée (et indirectement celles des inserts) en utilisant le « DNA high sensitivity bioanalyzer » pour valider la qualité et la longueur des fragments de la banque varient entre 450 et 700 pb avec une moyenne de 539 pb (Figure 16). En soustrayant à ces fragments la longueur des amorces (comprenant les adaptateurs) de 60 pb chaque, il est possible de déterminer que la taille des inserts était entre 330 et 580 pb, soit un peu plus petit que ce qui avait été sélectionné théoriquement par le bioanalyseur.



From [bp]	To [bp]	Corr. Area	% of Total	Average Size [bp]	Size distribution in CV [%]	Conc. [pg/µl]	Molarity [pmol/l]	Color
106	1 000	1 404.6	93	539	9.0	3 925.64	11 452.8	

Figure 16 : Taille de la banque d'inserts pour le séquençage 1.

Analyse de la librairie pour le premier séquençage par « DNA high sensitivity bioanalyzer » après la sélection de la taille des fragments par migration sur blue pippin. Les deux marqueurs de taille sont représentés par les pics à 35 pb et 10380 pb. Le pic débutant approximativement à 450 pb et finissant à environ 700 pb représente la distribution de la taille de la banque d'inserts.

Le séquençage de la banque d'insert a été fait en utilisant un système MiSeq de la plateforme d'Illumina au Centre d'innovation Génome Québec et Université McGill via la méthode des « paired-end », c'est-à-dire que chaque insert est séquencé à partir des deux extrémités. En se basant sur le génome des oiseaux déjà séquencés, la grosseur attendue du génome de l'hirondelle bicolore est d'environ 1,1 Gb.

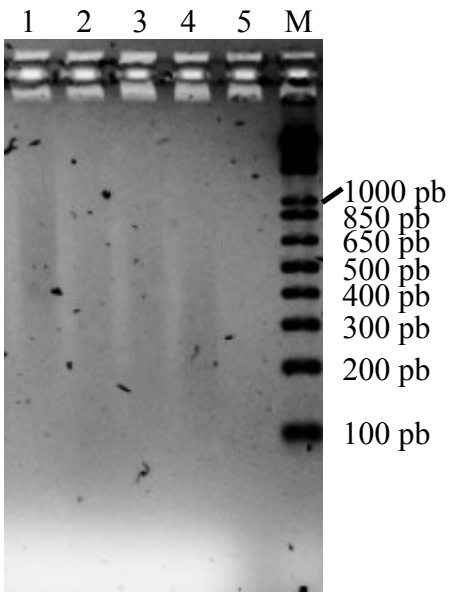
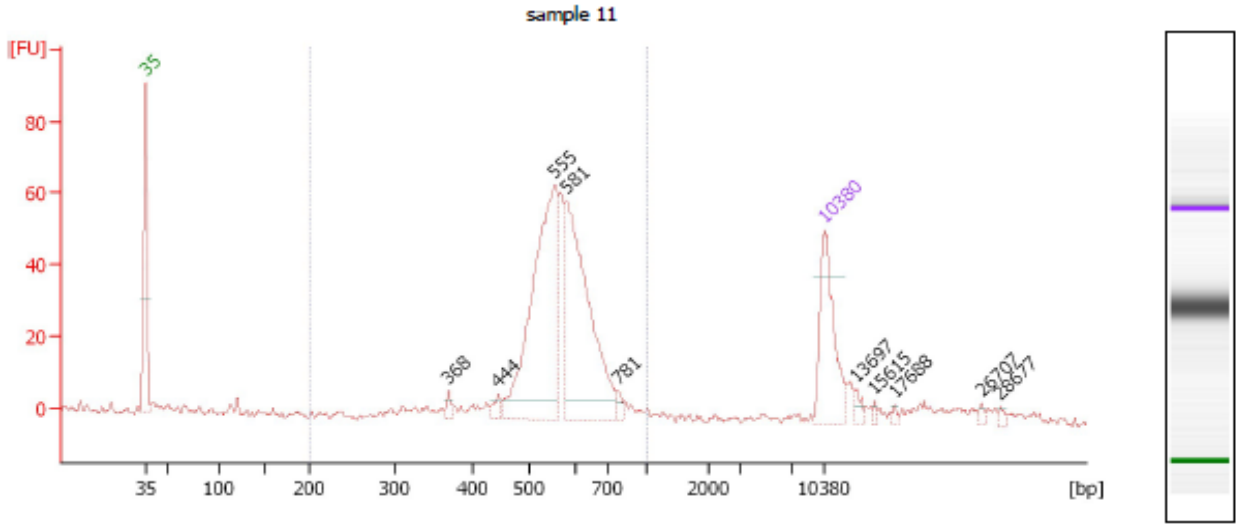


Figure 17 : La taille des inserts du deuxième séquençage suite à la sonication de l'ADNg. Longueur des fragments d'ADN de la deuxième hirondelle séquencée la sonication (30 sec ON/30 sec OFF) à intensité moyenne. Le puit 1 : un cycle de sonication; le puit 2 : deux cycles; le puit 3 : trois cycles; le puit 4 : quatre cycles; le puit 5 : cinq cycles et le puit marqueur.

Une deuxième hirondelle bicolore mâle (n.13, voir Tableau 1) a été séquencée pour des raisons qui seront expliquées dans la section 3.2. Contrairement à la première hirondelle (retrouvée morte sur le terrain), celle-ci a été capturée et euthanasiée par une surdose d'isoflurane en vase clos avant d'être congelée à -20ºC. Le même protocole a été utilisé pour procéder à l'extraction d'ADN à partir du tissu musculaire de la poitrine. La banque d'insert a été produite essentiellement dans les mêmes conditions que la première. Seulement l'étape concernant le nombre de cycles de sonication, passant de 30 à 4 a été modifiée pour conserver des fragments aux tailles souhaité. La même intensité de sonication et les mêmes paramètres de cycles ont été sélectionnés, mais les résultats étaient très différents (voir Figure 17 et Figure 15). Au final, la taille moyenne des fragments de cette librairie est approximativement de 555 pb et la taille des inserts est située entre 324 et 661 pb (Figure 18).



Peak table for sample 11 : sample 11

Peak	Size [bp]	Conc. [pg/µl]	Molarity [pmol/l]	Observations
1	35	125.00	5 411.3	Lower Marker
2	368	7.66	31.5	
3	444	9.45	32.2	
4	555	280.15	765.0	
5	581	238.56	622.2	
6	781	5.95	11.6	
7	10 380	75.00	10.9	Upper Marker
8	13 697	0.00	0.0	
9	15 615	0.00	0.0	
10	17 688	0.00	0.0	
11	26 707	0.00	0.0	
12	28 677	0.00	0.0	

Figure 18 : Taille de la banque d'inserts pour le séquençage 2.
Analyse de la librairie pour le deuxième séquençage par « DNA high sensitivity bioanalyzer » après la sélection de la taille des fragments par migration sur blue pippin. Les deux marqueurs de taille sont représentés par les pics à 35 pb et 10380 pb. Le pic débutant à 444 pb et finissant à 781 pb représente la distribution de la taille de la banque d'inserts. La taille moyenne de la banque est de 555 pb.

2.1.6 Assemblage du génome de l'hirondelle bicolore

Puisque majorité des inserts ont une longueur plus petite que 580 pb et qu'une lecture de 300 pb est générée à chaque extrémité des inserts, la plupart des inserts ont une paire de

lectures qui se chevauchent. Il a donc été possible de combiner la majorité des paires de lectures en l'équivalent d'une seule représentant l'insert au complet. Pour les inserts plus longs que 580 pb, la combinaison des lectures n'est pas réalisable dû à une zone d'appariement trop courte (moins de 20 pb) ou inexistante à l'étape de l'échafaudage. Ces lectures toujours paired-ends sont tout de même utiles pour l'assemblage à l'étape de l'échafaudage. La Figure 7 explique comment les programmes d'assemblage (assembleur) utilisent ces deux types de lectures pour assembler un génome.

2.1.6.1 Combinaison des lectures d'une paire

Combiner les lectures d'une paire facilite l'assemblage, car ils sont plus longs et ils sont deux fois moins nombreux que les lectures paired-end brutes. La première étape a donc été de fusionner les deux lectures d'une même paire lorsque leurs extrémités chevauchantes pouvaient être alignées en utilisant le programme FLASH 1.2.8 avec quelques paramètres spécifiques. Le premier paramètre est la longueur minimale d'alignement qui indique le nombre de bases qui doivent s'aligner pour qu'il y ait une fusion des deux lectures. Par défaut, il est à 10 bases, mais pour diminuer les fusions erronées, il a été fixé à 20 (`min-overlap`=20). Les banques ont été faites pour que la majorité des lectures d'une paire puissent s'aligner. Le deuxième paramètre qui a été utilisé est le maximum de mésappariements. Il représente le pourcentage maximal de bases non identiques acceptées dans la région d'alignement entre deux lectures d'une même paire pour qu'ils soient fusionnés. Si le pourcentage est supérieur, ils ne sont pas combinés. Dans le cas présent, entre 5 et 10% d'erreur ont été acceptés pour que les erreurs lors du séquençage influencent le moins possible le résultat (`max-mismatch-density`=0,05 ou 0,1). Il est fixé à 25% par défaut, mais cela représente une erreur de séquençage à chaque quatre bases et il a été jugé que c'était plutôt élevé (Magoc & Salzberg, 2011). Après la fusion, les lectures combinées ont été compilées

dans un fichier, alors que ceux ne l'ayant pas été ont été classées dans deux fichiers séparés, sous forme de lectures paired-end brutes.

2.1.6.2 Filtrer les adaptateurs

Tous les morceaux d'adaptateurs qui ont été séquencés doivent être éliminés des lectures, car ils pourraient engendrer un biais lors de l'assemblage. Dans le cas présent, ils ne devraient pas être séquencés, car la taille des inserts de la librairie avait au moins une longueur de 400 pb et que la longueur d'une lecture est de 300 pb. Par contre, il se peut que des inserts plus courts se soient retrouvés dans la banque et que, par conséquent, il y ait quelques lectures qui contiennent la séquence partielle ou complète d'un adaptateur. Comme il est très peu probable qu'il y ait de formation de concatémères (ex. : adaptateur-insert1-adaptateur-insert2-adaptateur) grâce à la structure en Y des adaptateurs, ces derniers ont été filtrés seulement aux extrémités des lectures, soit en 5' et 3', et pas dans la lecture au complet.

Pour éliminer les adaptateurs pouvant se retrouver dans la séquence des lectures, le programme CutAdapt v1.3 a été utilisé (Smeds & Künstner, 2011). Ses paramètres permettent de choisir l'endroit où chercher les adaptateurs (en 5', dans la lecture et/ou en 3'), c'est-à-dire la manière dont vont être filtrés les lectures. Le paramètre `a` a été utilisé pour faire une recherche du côté 3' de chaque lecture et le paramètre `g`, avec l'option `^` comme suffixe à l'adaptateur, pour en faire une du côté 5'. Pour cette étape, une erreur de 10% a été permise pour le repérage des adaptateurs (`e` 0,1). C'est le même taux qui a été permis pour la combinaison des paires avec FLASH. Ce programme a été utilisé seulement pour le premier séquençage, car un des assembleurs sélectionnés pour ce séquençage (SOAP) ne pouvait pas les filtrer lui-même (Martin, 2011).

2.1.6.3 Vérification de la qualité des lectures

La dernière étape avant de procéder à l'assemblage des lectures est de les filtrer pour éliminer les bases des extrémités ayant un score Phred jugé trop faible. Il aurait été possible de filtrer les scores de qualité avec CutAdapt, mais cette option n'a pas utilisée, car peu de paramètres sont disponibles avec ce programme (Martin, 2011). Un autre mieux adapté pour cette fonction a été choisi (Smeds & Künstner, 2011). Le filtrage a été effectué avec le programme ConDeTri v2.2, car il possède plusieurs options intéressantes qui sont indisponibles chez d'autres programmes du même type (Smeds & Künstner, 2011). C'est d'ailleurs ce dernier qui a été utilisé pour filtrer les lectures avant l'assemblage du génome du gobemouche à collier et de la mésange de Hume (Cai et al., 2013; Ellegren et al., 2012). ConDeTri est capable de traiter les lectures combinées et aussi les lectures en paire, ce qui est très pratique. Si une des lectures d'une paire est exclue par cette étape, l'autre sera considérée comme s'elle provenait d'un séquençage « single end », c'est-à-dire d'un seul côté. Cette lecture sera alors classée dans un fichier différent (Smeds & Künstner, 2011).

ConDeTri filtre en deux étapes. La première raccourcit les lectures à partir de l'extrémité 3' selon le score Phred accepté (bonne qualité = `hq`). Les bases ayant un score inférieur au `hq` peuvent être conservées à condition que leur somme n'excède pas le nombre maximal de bases consécutives (`ml`) ayant au moins le score minimum inférieur (faible qualité = `lq`). Dans ce cas, la séquence en 3' est gardée en mémoire jusqu'à la fin de la filtration et conservée s'il n'y a pas de dérogation aux paramètres de filtration pour le reste de la séquence vers le côté 5'. Dans le cas contraire, les bases antérieures sont supprimées de la lecture et la filtration continue. Pour que cette première étape de filtration arrête, il faut qu'il y ait un certain nombre de bases consécutives ayant un score `hq` (`mh`). Si après la filtration, une lecture a une longueur plus petite que la longueur minimum permise (`minlen`), cette lecture sera exclue des données.

À la deuxième étape de filtration, une certaine fraction des bases de la lecture (`frac`) doit avoir un score de qualité au-dessus du `hq` pour qu'elle soit approuvée. De plus, il ne faut pas qu'une des bases ait un score de qualité inférieur au minimum (`lq`). Les paramètres qui ont été spécifiés avec leur valeur respective pour le premier assemblage sont `hq`=25, `lq`=10, `frac`=[0,8], `minlen`=50 et `mh`=5 et `sc`=33. Ces valeurs sont toutes celles par défaut. Encore une fois, cette étape n'a pas été effectuée pour le deuxième séquençage parce que l'assembleur utilisé (Newbler) effectue lui-même le filtrage des bases (Margulies et al., 2005; Smeds & Künstner, 2011).

2.1.7 Assemblage

Deux assembleurs ont été utilisés dans le cadre du projet. SOAP*denovo* v2.04 a été sélectionné, car c'est ce programme qui a été utilisé pour l'assemblage du génome de deux oiseaux récemment séquencés, le gobemouche à collier et la mésange de Hume (Cai et al., 2013; Ellegren et al., 2012; Luo et al., 2012). Newbler, un autre programme d'assemblage, a été choisi parce qu'il utilise un algorithme différent de SOAP. Il a d'ailleurs été conseillé par le professeur Sébastien Rodrigue (Université de Sherbrooke, département de biologie) qui l'utilise dans son laboratoire. Il était alors possible de comparer les deux méthodes d'assemblage différentes. Ces deux assembleurs ont été utilisés pour le premier séquençage, mais seulement Newbler l'a été pour le deuxième. Voir la section 3.2 pour l'explication.

Différentes longueurs ont donc été testées, soit 31, 51, 61 (avec la version 63mer), 71, 100 et 127 (avec la version 127mer). La grosseur du génome est estimée à 1,1 Gb, donc le paramètre correspondant (`N`) a été fixé à 1100000000. La liste des lectures fusionnées a été utilisée pour former les contigs (`asm_flags`=1) et celle des lectures paired-end (non fusionnés) a été utilisée pour former les contigs et les échafaudages (`asm_flags`=3), car lors de la

construction des échafaudages, aucune base n'est ajoutée à l'assemblage. Le minimum de bases alignées nécessaire pour l'assemblage entre deux lectures a été fixé à 32 (`map_len`=32). Cette valeur a été utilisée pour l'assemblage du génome du gobemouche (Ellegren et al., 2012). Voir l'Annexe 2 pour la liste de tous les autres paramètres qui ont été testés et avec quelle(s) valeur(s).

Le programme Newbler a été utilisé de deux manières différentes pour faire un assemblage, soit avec un génome de référence, soit sans génome de référence (*de novo*). Les paramètres qui ont été utilisés pour ces deux méthodes d'assemblages sont `vt` pour spécifier la liste de la séquence des adaptateurs à filtrer et `o` pour spécifier le nom du fichier de sortie. Les paramètres utilisés seulement pour l'assemblage *de novo,* via la commande `runAssembly`, sont; `cpu` 48 pour procéder avec 48 processeurs (beaucoup de temps de calcul et de mémoire sont nécessaires pour un assemblage *de novo*); `large` qui permet de faire l'assemblage *de novo* plus rapidement en y allant de manière moins pointue, mais tout aussi rigoureuse et `urt` qui indique d'utiliser une lecture au complet même s'il y a seulement une partie de la lecture qui est alignée à l'extrémité d'un contig. Pour les assemblages avec génomes de référence, les deux génomes utilisés séparément, via la commande `runMapping`, sont ceux du gobemouche à collier et du diamant mandarin (*Taeniopygia guttata*). Ces génomes ont été obtenus par la banque de donnée Ensembl (Flicek et al., 2014). La seule autre option ajoutée pour ces assemblages est le `cpu` 24, car le temps de calcul de 24 processeurs était suffisant pour ce procédé.

Les fichiers contenant les lectures non combinées (paired-end) du séquençage ont dû être modifiés un peu, car Newbler n'est pas capable de reconnaître par lui-même ces types de lectures quand elles sont sous le format fastq. Les fichiers fastq paired-end ont été modifiés avec le programme fq_all2std.pl écrit par Vincent Baby (étudiant au doctorat sous la supervision du professeur Sébastien Rodrigue) pour les transformer en fichier seq et qual. Le

premier contient le nom des lectures et les séquences, alors que le deuxième contient le nom des lectures et le score de qualité des bases. Plusieurs assemblages ont été faits avec Newbler en utilisant les lectures brutes du séquençage (tous paired-end) et celles qui ont été combinées (lorsque c'était possible).

Pour les deux séquençages et les deux assembleurs, différentes combinaisons ont été faites pour optimiser au maximum le résultat de l'assemblage. Les tests faits comprennent des combinaisons entre les résultats d'assemblage avec et/ou sans génome de référence et entre le premier et/ou le deuxième séquençage.

2.1.8 Contamination

Lors du séquençage et de l'assemblage d'un nouveau génome, des problèmes de contaminations peuvent compliquer et/ou engendrer des erreurs dans le résultat final. La méthode d'extraction du matériel génétique permet de limiter au maximum toutes contaminations lorsque bien pratiquée. Par contre, il est impossible d'éliminer les contaminants qui étaient déjà présents dans l'échantillon. Les oiseaux sont des espèces qui sont normalement parasitées. En procédant à l'extraction de l'ADNg des hirondelles, l'ADNg de ces parasites a contaminé le matériel génétique et le tout a été séquencé. La technique utilisée pour éliminer ces lectures et/ou contigs de l'assemblage s'est basée sur ce qui avait été fait pour le gobemouche à collier par l'utilisation de BLAST (Altschul, Gish, Miller, Myers, & Lipman, 1990; Ellegren et al., 2012). Le fichier de la banque génomique « nt » du NCBI a été téléchargé en même temps que la liste de tous les *GenBank ID* (gi, numéro spécifique d'une séquence chez une espèce y) associés aux `taxid` (numéro taxonomique d'une espèce). Une liste de tous les *Taxonomic* ID (taxid) associés aux classes des *Aves* ou des *Reptilia* a été créée à partir de la liste du NCBI. Pour déterminer si les lectures/contigs provenaient de

contaminants ou non, ils ont été alignés sur la banque « nt » du NCBI en utilisant l'algorithme BLASTn avec les paramètres par défaut. Le taxid a été ajouté à chaque résultat du blast à partir du gi. Par la suite, les cinq premiers alignements et/ou les alignements ayant un score égal ou supérieur à 150 ont été conservés pour chaque lectures/contigs. La sélection des cinq meilleurs a été faite pour avoir un minimum d'alignement pour chaque contig même si leur score était inférieur à 150. C'est d'ailleurs la technique des meilleurs alignements qui a été utilisée pour retirer la contamination du génome du gobe mouche à collier (Ellegren et al., 2012). Ce n'est pas seulement les cinq premiers qui ont été sélectionnés parce qu'un contig peut avoir le même score d'homologie pour des séquences homologues d'espèces différentes. Donc, si pour un contig les huit meilleurs alignements ont le même score, seulement les cinq premiers auraient être conservés. Il aurait alors été possible de rejeter les alignements fait sur une séquence d'espèces aviaires et/ou reptiliennes. Un seuil minimal de 150 pour le score d'homologie a été fixé pour conserver les bons alignements. À partir de ce tri, tous les lectures/contigs qui n'avaient pas un alignement sur une partie de génome d'une espèce de reptiles ou d'oiseaux ont été retirés des données. Finalement, le niveau de contamination (en base) par rapport au nombre de bases totales du séquençage a été calculé.

2.2 Résultats

2.2.1 Extraction d'ARN

La qualité de tous les ARN, extraits des tissus musculaires de la manière décrite dans la section matériel et méthode (section 2.1.2), était soit intacte ou partiellement dégradée (Tableau 1, Tableau 2 et Figure 19). Pour être considéré comme intacte, l'ARN devait présenter deux bandes bien distinctes correspondant aux ARNr 18S et 28S. Pour tous ceux qui ont été produits différemment (c'est-à-dire lorsque l'échantillon n'avait pas déjà été congelé

ou avec une variation du protocole décrit dans la section 2.1.2), l'ARN obtenu était considéré comme complètement dégradé.

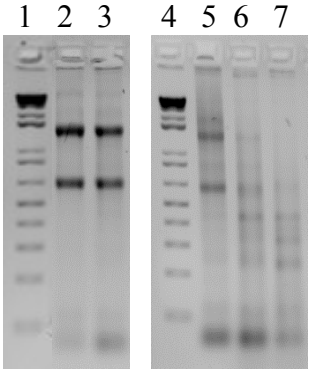


Figure 19 : Exemple de la qualité d'ARN obtenue.
Les puits 1 et 4 sont des marqueurs moléculaires (50 pb), les puits 2 et 3 représentent de l'ARN intact, les puits 5 et 6 de l'ARN partiellement dégradé et le puit 7 de l'ARN complètement dégradé.

2.2.2 Clonage des gènes cibles *AhR* et *CYP* et qPCR

Une portion des gènes *AhR*, *RPLP0*, *EIF1* et *EIF4A2* a été clonée pour l'hirondelle et la musaraigne (Tableau 5 et Tableau 6). En plus, le gène *CYP1B1* de l'hirondelle et *CYP1A1* de la musaraigne l'ont également été. Cependant, la séquence amplifiée du gène *CYP1A1* de la grande musaraigne ne s'est pas alignée au complet sur le même gène. Certains fragments correspondent au gène *CYP1A1* de cette espèce, mais d'autres correspondent au gène *CYP1A2*. Les séquences clonées sont données dans les Tableau 13 et Tableau 14 de l'Annexe 3.

Malgré l'ajout de gènes homologues d'autres espèces de mammifères un peu plus loin phylogénétiquement de la grande musaraigne que celles déjà présentées (la musaraigne commune et le toupaye de Belanger), le gène *CYP1B1* de cette espèce n'a pas été cloné. Les

gènes non clonés de l'hirondelle (*CYP1A4* et *CYP1A5*) seront probablement obtenus dans un futur proche via l'assemblage de son génome.

Tableau 5 : Gènes clonés chez l'hirondelle bicolore.

Gènes	Longueur d'ADNc amplifié (pb)	Exons amplifiés	Nombre total d'exons estimés	Longueur d'ADNc estimée (pb)	Score	E-value
AhR	576	4 à 8	9	6223	1078	0
CYP1B1	534	2 et 3	3	3836	940	0
RPLP0	603	1 à 5	8	1775	1053	0
EIF1	259	1 à 3	3	550	484	2e-136
EIF4A2	856	4 à 11	11	2687	1489	0

La longueur des fragments amplifiés pour l'hirondelle bicolore, le numéro des exons amplifiés correspondants aux gènes homologues de référence, le nombre total d'exons et la longueur de l'ADNc estimée à partir des mêmes informations de ce gène chez les espèces de référence. En plus, le score d'alignement et le e-value obtenus pour le gène amplifié de l'hirondelle bicolore sur le gène homologue du gobemouche à collier à l'aide de l'outil BLASTn d'Ensembl Genome Browser sur la banque d'ADNc de cette espèce.

À partir des séquences obtenues par clonage, des amorces ont été conçues pour procéder à la prochaine étape, soit le qPCR. Dans le Tableau 7 et le Tableau 8, la séquence des amorces utilisées est représentée de même que la température à laquelle ces amorces amplifient le mieux la séquence désirée. Des informations supplémentaires sur les résultats obtenus qui ont servi à valider le choix de ces couples d'amorces sont disponibles dans le Tableau 11 et le Tableau 12 de l'Annexe 1. Les séquences amplifiées par ces amorces par qPCR sont présentées dans le Tableau 13 et le Tableau 14 de l'Annexe 3.

Tableau 6 : Gènes clonés chez la grande musaraigne.

Gènes	Longueur d'ADNc amplifié (pb)	Exons amplifiés	Nombre total d'exons estimés	Longueur d'ADNc estimée (pb)	Score	E-value	% d'identité (avec *Shrew araneus*)
AhR	787	3 à 10	13	1974	787	0	
*CYP1A1**	999	5 et 6 205 pb			301	4^{e}-81	94.15
		1 54 pb			46,1	3^{e}-04	88,89
(CYP1A2)*		3 à 6 263 pb	8	1473	283	7^{e}-76	88,59
		1 64 pb			103	1^{e}-21	95,31
		1 32 pb			48,1	7^{e}-05	93,75
RPLP0	258	1 à 3	7	954	408	4^{e}-114	
EIF1	275	1	1	342	505	3^{e}-143	
EIF4A2	641	4 à 9	11	1224	1061	0	

La longueur des fragments amplifiés pour la grande musaraigne, le numéro des exons amplifiés (et, pour les gènes *CYP1A1* et *CYP1A2,* la longueur de l'alignement) correspondants aux gènes homologues de référence, le nombre total d'exons et la longueur de l'ADNc estimée à partir des mêmes informations de ce gène chez les espèces de référence. En plus, le score d'alignement, le e-value et le pourcentage d'identité (pour les gènes *CYP1A1* et *CYP1A2*) obtenus pour le gène amplifié de la grande musaraigne sur le gène correspondant de la musaraigne commune. *Il s'agit de la même séquence amplifiée qui s'aligne partiellement sur deux gènes.

Tableau 7 : Amorces optimisées pour le qPCR pour les gènes de l'hirondelle bicolore.

Gènes	Amorce sens	Amorce antisens	Température optimale	Longueur de l'amplicon	Exons amplifiés
AhR	5RT-Tchy_AhR1-4 GCTTTATTGTGCTGAAAACCA	3RT-Tchy_AhR1-4 AGGCCCATCGATTTTCTT	62 °C	95	7 et 8
EIF1	5-RT-Tchy_EIF1-	3-RT-Tchy_EIF1-1	60 °C	112	1 et 2

	1 GGAAGACCCT CACCACAGTC	TCGGGGTGCTC AATTACAG			
EIF4A2	5RT-Tchy_EIF4A2-1 CAGGGCGTGT GTTTGATATG	3RT-Tchy_EIF4A2-1 ATTTCATCGGCT TCATCCAG	60 °C	85	5 et 6

La séquence des amorces optimisées pour faire du qPCR pour différents gènes de l'hirondelle bicolore. La température optimale pour la réaction, la taille de l'amplicon et les exons amplifiés correspondant aux gènes homologues du gobemouche à collier sont donnés.

Tableau 8 : Amorces optimisées pour le qPCR pour les gènes de la grande musaraigne.

Gènes	Amorce sens	Amorce antisens	Température optimale	Longueur de l'amplicon	Exons amplifiés
AhR	5RT-Shrw_AhR-3 GTGCTTTGTATG CCGACTT	3RT-Shrw_AhR-3 AACAAGGCTA ACTGCGATG	60,5	144	7 et 8
RPLP0	5RT-Shrw_RPLP0-2 CTGAGATCCGG GACATGC	3RT-Shrw_RPLP0-2 GGCACCGTCA CCTCACAG	62 °C	85	2 et 3
EIF1	5-RT-Shrw_EIF1-1 CAGAGAAACGG CAGGAAGAC	3-RT-Shrw_EIF1-1 AATTACAGTA CCATTGCAGG CA	60 °C	114	1
EIF4A2	5RT-Shrw_EIF4A2-1 TTGTTGGTACA CCTGGGAGAG	3RT-Shrw_EIF4A2-1 AATCCCCGGC TTAACATCTC	60 °C	112	5 et 6

La séquence des amorces optimisées pour faire du qPCR pour différents gènes de la musaraigne bicolore. La température optimale pour la réaction, la taille de l'amplicon et les exons amplifiés correspondant aux gènes homologues de la musaraigne commune sont donnés.

2.2.3 Séquençage génomique de *Tachycineta bicolor*

Pour les deux séquençages, l'outil FastQC a été utilisé pour analyser la qualité des données du séquençage. La qualité des bases des lectures, représentée par le score Phred, a indiqué un profil de scores normal de même que la distribution des bases dans les lectures (voir Annexe 4 (Figure 22, Figure 23 et Figure 24) et Annexe 5 (Figure 25, Figure 26 et Figure 27)). Seulement quelques bases à la fin des lectures avaient un score un peu moins élevé indiquant que peu de lectures devraient être affectées lors de la filtration. Le taux de GC des lectures des séquençages tournait autour de 48%. C'est normalement ce qui est retrouvé chez les oiseaux. La couverture du génome représente le nombre de fois qu'une région X devrait être retrouvée dans le séquençage (Ellegren et al., 2012). Ce n'est qu'une approximation de la moyenne, c'est-à-dire que certaines régions sont plus représentées que ce taux et d'autres moins. D'après la quantité de lecture attendue à la suite du séquençage MiSeq, un taux de couverture d'environs 7X avait été estimé pour les deux séquençages. Or, pour le premier, le séquenceur a généré 11,5 Gb de données. Cela représente moins de lectures qu'espérées (9 012 352 inserts séquencés en paire), ce qui a diminué le taux de couverture du génome à 3,4X ($\frac{9\,012\,352 \text{ inserts} * 419 \text{ bases (taille moyenne des inserts)}}{1\,100\,000\,000\ bases\ (taille\ estimée\ du\ génome)}$). Pour le deuxième, la couverture obtenue a été de 7,7X grâce à la situation inverse (plus d'inserts ont été séquencées, $\frac{19\,348\,433 \text{ inserts} * 435 \text{ bases}}{1\,100\,000\,000\ bases}$, soit 24,1 Gb de données). Finalement, le taux de duplication des séquences pour le premier séquençage était de 21,1%, ce qui est un peu plus que la normale (Figure 20). Celui du deuxième séquençage était très élevé, soit de 85,8% (Figure 21). Cela indique qu'il n'y avait probablement pas assez de matériel au départ lors de la préparation de la banque d'inserts et/ou qu'un biais a favorisé certaines séquences. Le problème qu'il y a eu avec le sonicateur pour le deuxième séquençage pourrait expliquer le très au haut taux de duplication.

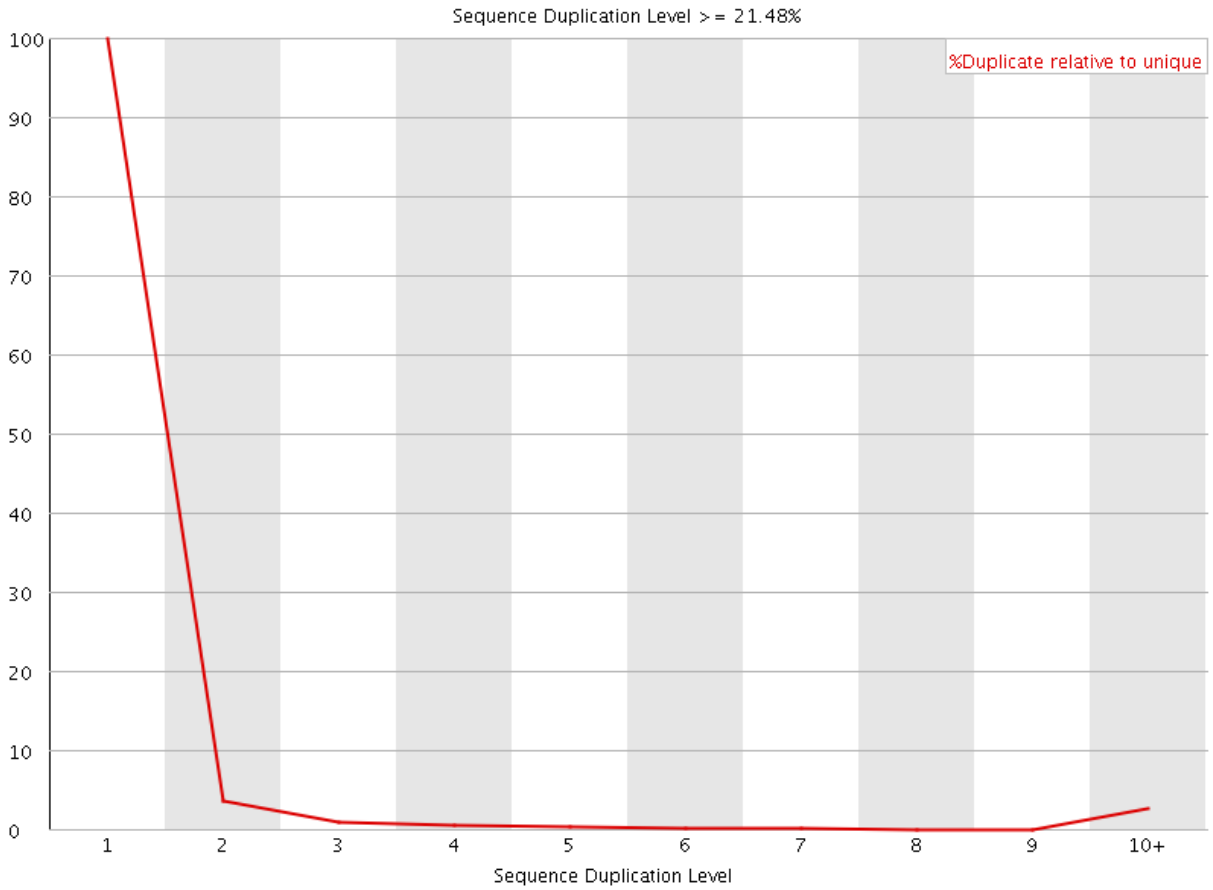


Figure 20 : Le taux de duplication des lectures pour le premier séquençage.

Pour la première lecture de chaque insert, le taux est de 21,5%. Pour la deuxième, il est de 20,8% (graphique non présenté). La moyenne est donc de 21,1%.

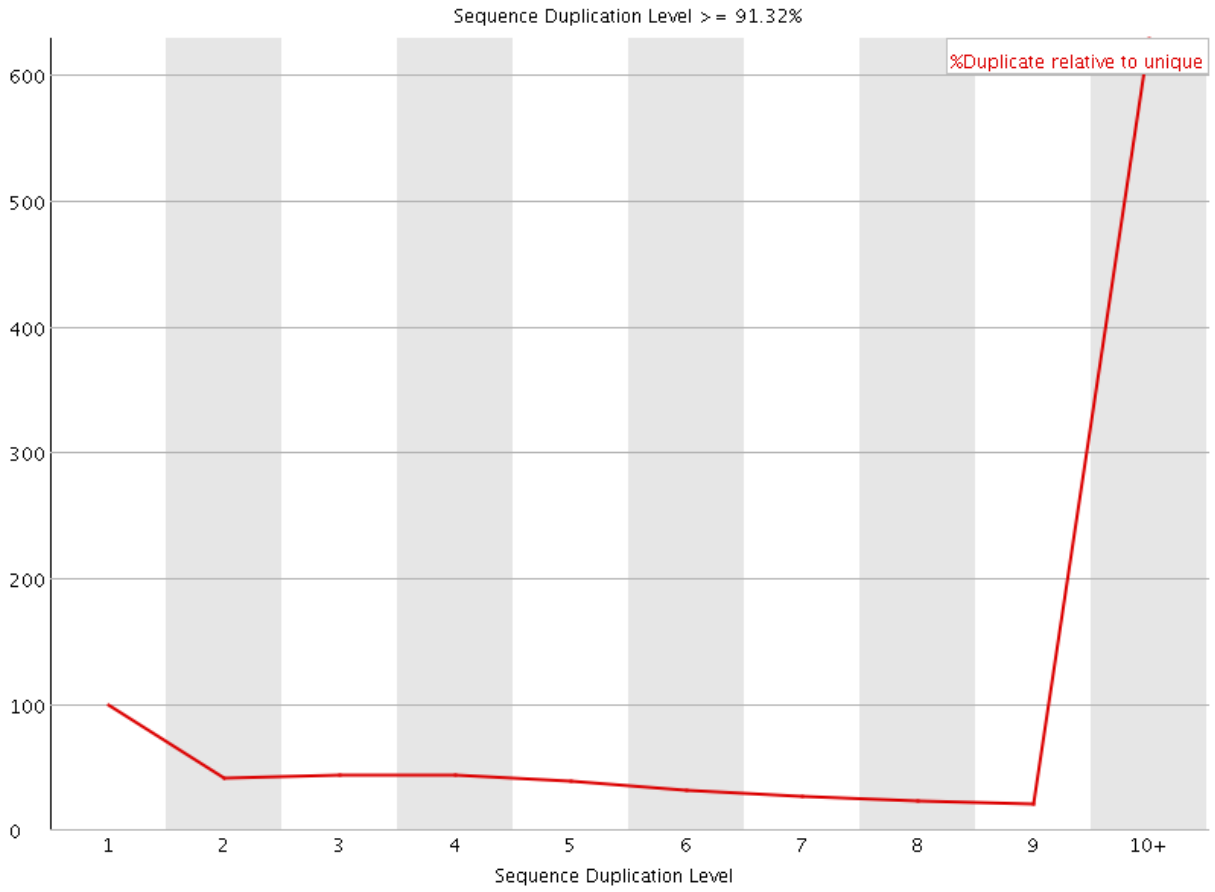


Figure 21 : Le taux de duplication des lectures pour le deuxième séquençage.

Pour la première lecture de chaque insert, le taux est de 91,3%. Pour la deuxième, il est de 80,3% (graphique non présenté). La moyenne est donc de 85,8%.

2.2.4 Assemblage

Le Tableau 9 donne le résultat de différentes caractéristiques importantes à considérer pour évaluer la qualité d'un assemblage pour les deux séquençages. Ces données sont celles de deux assemblages *de novo* simples (sans combinaison de plusieurs assemblages) avec Newbler. Aucun exemple de résultats n'a été présenté avec l'assembleur SOAP, car les assemblages faits par ce dernier n'étaient pas très concluants. Ils avaient un N50 de 500 pb dans le meilleur des cas, alors que ceux faits via Newbler en avaient un minimal d'environ 580 pb. SOAP a donc été mis de côté (voir l'explication complète à la fin de la section 3.3).

Différentes combinaisons d'assemblages ont été essayées. Quelques exemples sont présentés dans le Tableau 10.

.

Tableau 9 : Les caractéristiques des assemblages *de novo* obtenus via Newbler à partir des deux séquençages.

	N50 (bases)	Nombre de contigs	Nombre de bases (Gb)	Portion estimée du génome séquencé (%)
1er	1339	636 673	0,60	55
2ème	582	504 106	0,18	16

Tableau 10 : Les résultats de différentes combinaisons d'assemblages testées pour le premier séquençage.

	Les assemblages	N50 (bases)	Nombre de contigs	Nombre de bases (Gb)	Portion estimée du génome séquencé (%)
1	*de novo*	1339	636 673	0,60	55
2	Gobemouche à collier*	764	1489208	0,45	41
3	Diamant mandarin*	759	1469242	0,46	42
4	1+2	2553	48 386	0,11	10
5	1+3	2566	46 221	0,11	10
6	4+5	2837	26 133	0,08	7

*L'assemblage a été fait avec ce génome comme référence.

Afin d'évaluer l'échantillonnage des deux séquençages, les portions d'ADNc qui ont été clonées chez l'hirondelle bicolore (*AhR, CYP1B1, RPLP0, EIF1* et *EIF4A2*) ont été alignées sur les lectures brutes et sur un assemblage de chacun d'eux. Pour le premier séquençage, tous les ADNc partiels ont été retrouvés au complet. Pour le deuxième, seulement une partie de la

portion clonée de *RPLP0* a été retrouvée malgré une couverture du génome prétendument plus élevée que le premier.

L'élimination de la contamination des lectures et/ou des contigs n'a pas été complétée. Le niveau de contamination du premier séquençage semble être faible d'après les résultats préliminaires, soit moins de 5%. Cependant, il est beaucoup plus élevé que celui de l'assemblage du génome du gobemouche à collier (0,2%) (Ellegren et al., 2012). Pour le moment, le niveau de contamination du deuxième séquençage n'est pas connu, car ce séquençage est moins valide que le premier considérant le taux élevé de duplication des séquences de ce séquençage et l'absence des séquences clonées. Il n'a donc pas été jugé essentiel de faire ce test.

CHAPITRE 3

DISCUSSION GÉNÉRALE ET CONCLUSION

3.1 Préparation des échantillons

Il était essentiel de faire le prélèvement des tissus lorsque l'animal avait été préalablement congelé pour ne pas récolter de l'ARN dégradé (Tableau 1). De plus, il était plus facile de suspendre la poudre de tissu dans un plus petit volume (250 µL) de Trizol (le volume était complété à 500 µL après l'homogénéisation (voir la section 2.1.2.1)). C'est pour cette raison qu'un changement a été apporté au protocole de Life Technologies pour l'extraction d'ARN avec du Trizol. Au lieu de faire les extractions dans un volume standard de 1 mL avec 100 mg de tissu, elles ont été faites dans un volume de 0,5 mL avec 50 mg de tissu. Cette modification du protocole original a permis d'augmenter la qualité de l'ARN, il a passé de partiellement dégradé à intact dans la plupart des cas. De plus, ce changement a permis de récolter plus de matériel même s'il y avait deux fois moins de tissu à l'étape initiale.

3.2 Clonage et préparation de la quantification de l'expression des gènes par qPCR

Grâce au clonage, une partie de l'ADNc de certains gènes ciblés chez l'hirondelle bicolore et la grande musaraigne est maintenant connue. Pour l'hirondelle, seulement la séquence de *CYP1A4* et *CYP1A5* reste un mystère. Une séquence partielle a été obtenue pour *CYP1A5*, mais il était complètement impossible de déterminer si c'était celle de *CYP1A4* ou de *CYP1A5* (données non présentées), car leur séquence est trop similaire; leur degré d'identité est de 97%

chez le diamant mandarin. Par contre, le séquençage du génome de cette espèce facilitera la distinction entre la séquence de ces gènes, car il va permettre d'obtenir les introns et les exons et pas seulement une partie de l'ADNc. La diminution de la difficulté d'identification sera attribuée à la connaissance des introns qui sont des régions plus sujettes aux variations que les exons, qui ont une pression sélective plus élevée pour être conservés. Pour la musaraigne, le gène *CYP1B1* n'a pas été cloné malgré quelques tentatives avec des amorces dégénérées. Le clonage de *CYP1A1* reste quelque peu nébuleux. Une des séquences qui a été clonée et séquencée s'aligne sur différents gènes de la musaraigne commune, soit certains exons de *CYP1A1* et d'autres de *CYP1A2* (Tableau 6). Ces deux gènes se ressemblent quand on compare leur séquence respective (91% d'identité chez la musaraigne commune), il est donc plus difficile de bien les distinguer. D'autres tentatives seront à faire pour obtenir une portion de l'ADNc de ces deux gènes chez la musaraigne.

La préparation de la quantification de l'expression des gènes par qPCR a été complétée pour certains des gènes qui avaient été clonés. Les deux seuls pour lesquels des amorces n'ont pas été validées sont *CYP1B1* et *RPLP0* de l'hirondelle bicolore. Les deux couples d'amorces testés pour le premier et le seul pour le deuxième n'étaient pas satisfaisants. Le problème était soit leur non-spécificité, soit l'amplification de cette région n'avait pas un taux d'efficacité dans l'intervalle attendu (entre 80 et 110%).

3.3 Séquençage et assemblage du génome de l'hirondelle bicolore

Comme mentionné précédemment, le séquençage du génome de l'hirondelle bicolore a été jugé essentiel pour différentes étapes du projet. Chez les oiseaux, ce sont les femelles qui sont hétérogamétiques, c'est-à-dire qu'elles ont deux chromosomes sexuels différents (ZW). Les mâles sont homogamétiques (ZZ) (Ellegren et al., 2012). Afin de ne pas sous-représenter les chromosomes sexuels par rapport aux autosomes lors du séquençage, le choix du spécimen à

séquencer s'est arrêté sur un mâle. Celui qui a servi pour le premier séquençage a été retrouvé mort sur le terrain, mais il était dans de bonnes conditions (Tableau 1, hirondelle n.8).

Un taux de couverture du génome de 3,4X a été obtenu par le premier séquençage, ce qui est faible considérant que ces lectures allaient servir à assembler un génome à partir de zéro (très peu d'information est disponible sur la génétique de l'hirondelle bicolore). Il est important de noter qu'il s'agit d'une estimation de couverture qui ne prend pas en compte la présence de contaminants (parasites et/ou autres). Ces derniers viennent diminuer la couverture du génome. De plus, un taux aussi faible peut être problématique, car certains assembleurs éliminent les séquences qui sont présentes moins de trois fois dans un séquençage parce qu'ils considèrent qu'elles sont des erreurs de séquençage. Malgré tout, 55% du génome a été assemblé, mais il était très fragmenté (le N50 se situe environ entre 750 et 2900 pb) (Tableau 10). Un deuxième séquençage a été jugé pertinent afin d'accroitre le taux de couverture pour tâcher d'augmenter la portion du génome séquencée et le N50. Deux options étaient possibles; séquencer le même individu ou un deuxième. Chez les organismes diploïdes, comme les oiseaux, un individu peut avoir différents allèles. Ceux-ci compliquent un peu l'assemblage, car ils causent une variabilité génétique. Il est donc préférable de ne pas faire d'assemblage *de novo* via le séquençage de plusieurs individus en une seule étape, car il y aurait beaucoup trop de variabilité. Par contre, pour un assemblage fait avec un génome de référence, la présence d'une matrice sur laquelle le programme peut aligner les différents allèles facilite l'assemblage et la variabilité génétique devient moins problématique. En considérant seulement ce point, le séquençage de la même hirondelle serait préférable. Cependant, celle-ci a été retrouvée morte et il est possible que des organismes décomposeurs en pleine croissance aient contaminé l'échantillon (sans oublier les parasites déjà présents). La contamination a été estimée à environ 5% à l'aide de résultats préliminaires, mais chez le gobemouche à collier il était seulement de 0,2%. Afin de minimiser au maximum la contamination et, ainsi, faciliter l'assemblage, c'est une seconde hirondelle qui a été séquencée (Tableau 1, hirondelle n.13).

La couverture du génome par le deuxième séquençage était de 7,7X, ce qui laissait présager une nette amélioration de la portion du génome séquencée (avec un N50 plus élevé). Par contre, les résultats primaires ont indiqué que les assemblages obtenus étaient de mauvaise qualité avec seulement 16% du génome assemblé et un N50 de 585 pb (Tableau 9). Un biais lors de la préparation de la banque d'inserts qui aurait causé la surreprésentation (et/ou la sous-représentation) de certaines régions du génome pourrait expliquer ce résultat. Cette hypothèse est supportée par deux points. Le premier est le haut niveau de duplications des séquences dans les lectures brutes. Il était à 85,8%, alors qu'il aurait dû être autour du taux de couverture (Figure 21). Le deuxième est que l'ADNc de tous gènes clonés partiellement a seulement été retrouvé dans les lectures (et l'assemblage) du premier séquençage, alors que seulement une portion de la partie clonée de *RPLP0* a été retrouvée pour le deuxième. Ce dernier résultat apporte d'ailleurs de la valeur à l'hypothèse du biais lors de la préparation, mais elle n'a pas été validée et la cause reste inexpliquée. Cela pourrait d'ailleurs avoir un lien avec le changement du nombre de cycles de sonication nécessaire pour la fragmentation de l'ADNg. Le deuxième séquençage a tout de même été conservé pour ajouter un peu d'information à l'assemblage du premier en combinant les assemblages. Le Tableau 9 montre les résultats d'un assemblage simple pour chaque séquençage avec Newbler. Quelques exemples de combinaison d'assemblages ont été présentés dans le Tableau 10. Le N50 est meilleur pour un assemblage *de novo* qu'avec un des génomes de référence et il est près de deux fois supérieure en combinant des assemblages entre eux. Cependant, le nombre de contigs et de bases conservés diminuent de beaucoup lors de ces combinaisons. Au final, la portion du génome séquencé est plus élevée avec un assemblage simple qu'avec une combinaison d'assemblages dû à la perte de plusieurs contigs/bases. Il est toutefois important de noter que toutes les combinaisons d'assemblages effectuées ont été faites qu'avec les longs contigs (>500 pb). Il serait intéressant de refaire des combinaisons d'assemblages, mais avec tous les contigs construits par les assemblages simples.

Le choix d'un assembleur varie selon plusieurs paramètres. Par exemple, la méthode de séquençage (Illumina, Roche), le type de séquençage (lectures courtes ou longues), la taille du génome à assembler et le nombre de lectures peuvent influencer l'habileté d'un assembleur à fonctionner correctement selon l'algorithme qu'il utilise (Bradnam et al., 2013). Pour qu'un assemblage puisse être qualifié de bon, la plupart de ses contigs devraient avoir une longueur de beaucoup supérieure à celle des lectures fusionnées. Dans le cas des deux séquençages, le N50 devrait être plus élevé que 455 et 492 pb pour le premier et deuxième respectivement. Or, le N50 maximal obtenu par SOAP était inférieur à ces valeurs. Il était à environ 400 pb. Il a donc été décidé d'abandonner SOAP au profit de Newbler à cause de la piètre qualité des assemblages de ce premier.

La suppression de toute la contamination des lectures brutes et/ou des contigs a un peu été mise de côté pour focuser sur l'assemblage. Les résultats préliminaires semblent cependant indiquer que le niveau de contamination du premier séquençage était minime (moins de 5%). Le degré de contamination du deuxième n'est pas connu pour le moment.

Évidemment, le génome assemblé de l'hirondelle bicolore est beaucoup moins complet et plus fractionné (0,60 Gb avec un N50 de 1130 bases pour les contigs) que celui du gobemouche à collier (1,13 Gb avec un N50 de 7,3 Mb pour les échafaudages) et de la mésange de Hume (1,04 Gb avec un N50 de 16,3 Mb pour les échafaudages), car la quantité de données utilisées initialement était nettement supérieure pour ces deux dernières espèces; soit 176 Gb de données (85X) pour le gobemouche, 184,5 Gb de données (96X) pour la mésange comparativement à 11,5 Gb de données (3,4X, premier séquençage) et 24,1 Gb de données (7,7X, deuxième séquençage) pour l'hirondelle. De plus, les données fournis pour des analyses d'expression des gènes (RNA-Seq) ont servi à perfectionner l'assemblage du génome de ces deux espèces d'oiseaux (Cai et al., 2013; Ellegren et al., 2012). Afin d'arriver à un assemblage de qualité comparable, d'autre(s) séquençage(s) (ou les données provenant de RNA-Seq, ChIP-Seq) seraient nécessaire. Or, dans l'optique du projet présenté dans ce mémoire, il n'était

pas nécessaire d'avoir un assemblage de génome à ce niveau, bien qu'il aurait été souhaitable d'obtenir un N50 un peu plus élevé.

3.3 Conclusion

L'étude des effets des pesticides sur des organismes vivants permettra d'apporter beaucoup de connaissances sur l'incidence qu'ils peuvent avoir à court et long terme sur ces derniers. L'avantage des organismes modèles, c'est que les étapes préliminaires (clonage, préparation d'amorces, séquençage génomique) ne sont pas nécessaires, car leur génome est connu. De plus, pour certaines espèces des produits facilitant leur étude dans différentes conditions, tels des anticorps ou des puces, sont disponibles sur le marché. Les deux modèles choisis pour le projet ont été sélectionnés pour les différentes raisons mentionnées dans la section 1.4.1, mais ils ne sont pas des espèces classiquement étudiées. Évidemment, très peu d'information est disponible sur leur génétique et cela implique une préparation qui a résulté au recadrement du projet initial. Pour le faire avancer le plus efficacement et rapidement possible, plusieurs étapes ont été enclenchées simultanément. Au final, quelques-unes restent à compléter, soit la fin du clonage d'une portion de *CYP1B1* et de *CYP1A1* de la grande musaraigne, la préparation de la quantification de l'expression de *CYP1A4* et *CYP1A5* de l'hirondelle bicolore et de *CYP1A1* et *CYP1B1* de la grande musaraigne et la fin de l'assemblage du génome de l'hirondelle bicolore, avant de commencer à tester l'hypothèse de départ. Malgré ces quelques points, la préparation a été avancée et dans un futur proche, le projet arrivera aux objectifs globaux. Pour le moment, ce projet a été mis en suspens, mais d'autres étudiantes du laboratoire du Pr. Luc Gaudreau travaillent activement sur des sujets connexes, soit l'étude de l'effet d'un ingrédient actif précis (ou en combinaison) sur AhR (et d'autres gènes par la suite) dans des lignées cellulaires du cancer du sein ou de la prostate.

3.4 Perspectives

À court terme, il reste à compléter l'étape préliminaire du projet, soit quelques clonages, finir la préparation pour le qPCR et l'assemblage du génome de l'hirondelle bicolore. Afin de favoriser l'obtention de longs contigs, le repêchage de tous ceux formés par les assemblages primaires (et non seulement ceux de plus de 500 pb) lors des combinaisons d'assemblages sera à envisager et à tester. L'élimination de la contamination sera également à faire. Une analyse plus poussée des contigs permettra de valider la qualité de l'assemblage final. Elle pourra être faite en alignant tous les gènes de l'hirondelle bicolore disponibles dans la banque de données de NCBI sur les contigs. Comme l'hirondelle bicolore n'est pas un oiseau extrêmement étudié au niveau génétique, seulement 231 gènes sont séquencés (partiellement ou complètement). Dans le cas d'un bon assemblage, un gène va s'aligner complètement sur un seul contig (ou aux extrémités de deux). Idéalement, cette observation sera faite pour tous les gènes de la banque de données. Si au contraire, plusieurs gènes s'alignent sur différentes parties de contigs, signe que ces gènes sont discontinus dans l'assemblage, ce dernier sera considéré mauvais. Par contre, lorsque seulement quelques gènes sont dans cette situation, ils peuvent indiquer un bon assemblage avec quelques erreurs ou représenter les duplications de petits segments dans le génome de cette espèce. Pour pousser l'analyse de l'assemblage plus loin, la même chose pourra être faite avec le génome d'oiseaux proches phylogénétiquement. Dans ce cas, un réarrangement génomique pourrait expliquer l'alignement d'un contig sur plusieurs chromosomes différents. Le programme « REAPR analysis » pourra aussi permettre de vérifier la qualité de l'assemblage en utilisant des génomes de référence en plus d'être utilisé pour corriger les erreurs (Ellegren et al., 2012). Bien entendu, après la validation de l'assemblage du génome, il faudra procéder à son annotation. Elle pourra être faite par des programmes comme InterProScan (Ellegren et al., 2012). Ce dernier permet par exemple d'identifier différentes régions du génome via la banque de données de signatures protéiques.

Par la suite, différentes expériences seront à faire pour atteindre les objectifs globaux. Le premier, qui est de déterminer si les pesticides dans l'environnement ont un effet sur l'hirondelle bicolore et la grande musaraigne, pourra être vérifié par qPCR via l'activation des gènes *CYP1*. Si cette hypothèse est confirmée, les différentes expériences mentionnées dans la section 1.4.1 permettront d'observer les effets épigénétiques (RNA-Seq, ChIP-Seq et MeDIP-Seq) globaux des polluants environnementaux sur ces deux espèces. Le RNA-Seq permettra d'identifier quels gènes sont plus (ou moins) transcrits lorsqu'un modèle est exposé à des pesticides. Ce résultat mettra en évidence certaines familles de gènes qui seront intéressantes à étudier pour mieux comprendre l'impact des pesticides sur les organismes vivants. Différents ChIP-Seq pourront être faits contre certaines protéines déjà connues comme étant impliquées dans la réponse aux ingrédients actifs ou contre de nouvelles potentiellement impliquées qui auront été identifiées par le RNA-Seq. Ces résultats permettront d'apporter de nouvelles connaissances sur le mécanisme de régulation de ces protéines lorsqu'elles sont activées directement ou indirectement par ces polluants. Par exemple, un ChIP-Seq contre AhR permettra d'identifier avec quelles régions du génome il interagit lorsqu'il est activé par les pesticides. De plus, il sera possible de combiner cette information à celle du RNA-Seq pour savoir quel effet cette interaction a sur la région où elle a lieu (inhibition ou activation). Finalement, ces résultats pourront être supportés par un MeDIP-Seq qui indique le niveau de méthylation génomique. Les régions où l'ADN est plus méthylées sont normalement inactives, alors que celles qui ne le sont pas sont transcrites. Une activation de la transcription d'un gène n'est pas nécessairement synonyme d'une augmentation de la quantité de la protéine concernée, car d'autres régulations interviennent entre la transcription et production d'une protéine fonctionnelle. Des buvardages de type western permettront de s'assurer que les régions du génome activées mènent bien à l'augmentation de la quantité de ces protéines.

D'autres projets en préparation apporteront aussi des renseignements qui seront à analyser en parallèle avec la suite du projet qui vient d'être présenté. Par exemple, le dosage de boulettes de nourriture que les hirondelles donnent à leurs petits, par chromatographie en phase liquide couplée à la spectrométrie de masse, indiquera à quels pesticides les sujets ont été exposés et

leur quantité. Il sera alors possible de faire des liens entre des modulations de processus cellulaires et certains pesticides. De plus, cette technique pourra aussi être appliquée pour quantifier la présence de polluants et des métabolites dans les tissus des modèles. Cela apportera de l'information sur ce qui se passe métaboliquement chez ces animaux (dégradation de pesticides, présence de métabolites secondaires...). Ces renseignements pourront par ailleurs être associés à l'activation (ou l'inhibition) de certains gènes identifiés auparavant par RNA-Seq. Finalement, pour valider que tous ces résultats sont dus seulement aux pesticides présents dans l'environnement des sujets et non à un autre facteur (aucune condition environnementale n'est contrôlée), des expériences en laboratoire *ex vivo* pourront être faites à partir de cellules prélevées chez l'hirondelle et la musaraigne. Certaines des expériences sur les tissus des modèles seront répétées sur ces cellules en présence de certains pesticides. Les renseignements qui auront été obtenus par la nourriture des hirondelles serviront à sélectionner les pesticides à tester ainsi que leur concentration. En bref, toutes ces informations serviront à mettre en lumière les gènes importants impliqués dans la réponse aux pesticides.

ANNEXE 1

INFORMATIONS SUPPLÉMENTAIRES DES COUPLES D'AMORCES POUR LE QPCR

Tableau 11 : Amorces optimisées pour le qPCR pour les gènes de l'hirondelle bicolore.

Gènes	Amorces (couple)	%	R^2	Intervalle des CT	CT ØRT/ddH_2O	Nb pts
AhR	5RT-Tchy_AhR1-4 3RT-Tchy_AhR1-4	92,1	0,923	29 à 34	39/NA	3
EIF1	5-RT-Tchy_EIF1-1 3-RT-Tchy_EIF1-1	99,5	0,993	23 à 30	33/37	3
EIF4A2	5RT-Tchy_EIF4A2-1 3RT-Tchy_EIF4A2-1	85,7	0,998	25 à 33	35/NA	3

Séquence des amorces optimisées pour faire du qPCR pour différents gènes avec la température optimale pour la réaction, la taille des amplicons et le numéro des exons amplifiés correspondant aux gènes homologues de référence.

Tableau 12 : Amorces optimisées pour le qPCR pour les gènes de la grande musaraigne.

Gènes	Amorces (couple)	%	R^2	Intervalle des CT	CT ØRT/ddH-$_2$O	Nb pts
AhR	5RT-Shrw_AhR-3 3RT-Shrw_AhR-3	99	-	23-31	-	3
RPLP0	5RT-Shrw_RPLP0-2 3RT-Shrw_RPLP0-2	80	0,959	25 à 30	36/33	3
EIF1	5-RT-Shrw_EIF1-1 3-RT-Shrw_EIF1-1	82,8	0,996	29 à 37	NA/NA	3
EIF4A2	5RT-Shrw_EIF4A2-1 3RT-Shrw_EIF4A2-1	91,6	0,854	33 à 38	NA/NA	3

Séquence des amorces optimisées pour faire du qPCR pour différents gènes avec la température optimale pour la réaction, la taille des amplicons et le numéro des exons amplifiés correspondant aux gènes homologues de référence.

ANNEXE 2

LES AUTRES PARAMÈTRES DE SOAP UTILISÉS

Pour indiquer à SOAP à quel moment utiliser les différents types de lectures (paired-end et combinées), il faut créer un fichier texte contenant les différentes informations ci-dessous. Ce fichier est soumis dans la commande de SOAP par le paramètre « s ».

max_rd_len=600
[LIB]
avg_ins=480
reverse_seq=0
asm_flags=1
pair_num_cutoff=3
map_len=32
q=l'adresse des lectures combinées
[LIB]
avg_ins=480
reverse_seq=0
asm_flags=1
pair_num_cutoff=3
map_len=32
q=l'adresse des lectures considérées single-end (elles ne sont plus en couple, car leur paire a été supprimé)
[LIB]
avg_ins=480
reverse_seq=0
asm_flags=3
rank=1
pair_num_cutoff=3
map_len=32
q1=l'adresse des lectures sens qui sont toujours en paired-end
q2=l'adresse des lectures antisens qui sont toujours en paired-end

Les autres paramètres qui ont été testés directement dans la commande de SOAP : « p 48 », « d 0 ou 1 », « D 1 ou 2 », « u », « c », « C » et « g ».

ANNEXE 3

SÉQUENCES DES PORTIONS DE GÈNES CLONÉS ET LA SÉQUENCE AMPLIFIÉE PAR QPCR

Tableau 13 : Séquences des portions d'ADNc des gènes clonés de l'hirondelle bicolore et la séquence amplifiée par qPCR (en gris) pour certains de ces gènes.

Gène	ADNc
AhR	GAATTCTTCCTTTATGGAAAGGAATTTCATCTGTAGGTTACGATGTCTC CTGGATAATTCATCTGGATTCCTGGCTATGAATTTTCAAGGACGACTGA AGTTTCTCCATGGACAAAACAAGAAAGGGAAGGATGGTGCTACTTTGT CTCCTCAGCTTGCACTGTTTGCAGTAGCTACTCCCCTGCAGCCACCATC TATCCTTGAGATACGAACCAAAAACTTCATCTTCAGAACGAAACACAA ACTGGATTTCACACCTACTGGCTGTGATGCAAAAGGAAAGATTGTCCTG GGATACACTGAAGCAGAGCTGTGTATGAGAGGAACAGGATACCAGTTT ATTCATGCAGCTGATATGCTTTATTGTGCTGAAAACCACATCCGAATGA TGAAGACAGGTGAGAGTGGAATGACTGTATTTAGGCTTCTAACCAAAG AAAATCGATGGGCCTGGGTACAGGCAAATGCACGTCTTGTCTACAAAA ATGGAAGACCAGATTACATCATTGCCACACAAAGACCTCTTACAGATG AAGAAGGGGCAGAACATCTACGGAAGCGTAACATGAAGTTGCC
CYP1B1	CACCGTCACCGACATCTTCGGCGCCAGCCAGGACACCCTATCCACCGC CCTGCAGTGGCTGCTCATCTTCCTCATCAGGTATCCGAAAGTGCAGGCT AAAATGCAGGAAGAAGTGGATAGGATTGTTGGAAGAGACCGTCTGCCG TGCGTTGAAGATCAGCCTCACCTGCCCTACATCATGGCTTTCCTGTATG AATCCATGCGTTTCAGCAGCTTTGTGCCTGTGACTATCCCACATGCCAC CACAACCAACACCTTCATCATGGGCTACCTCATTCCCAAGGACACCGTC ATCTTTGTAAATCAGTGGTCAGTGAATCACGACCCAGCAAAATGGTCC AACCCGGAGGACTTTGATCCAACAAGATTCCTGGATGAGAATGGGTTC ATCAATAAAGATCTTACTAGCAGCGTGATGATTTTCTCATTGGGAAAGC GTCGGTGTATTGGAGAGGAGTTATCCAAGGTGCAGCTCTTTCTCTTTAC CTCCATACTGGTGCATCAGTGCAATTTTACTGCTAATCCAAATGAGGA
RPLP0	GACAGGGCTACGTGGAAGTCCAACTATTTTATGAAAATCATCCAACTCT TGGATGATTACCCAAAATGTTTCATTGTGGGAGCAGACAACGTGGGAT CCAAGCAGATGCAGCAGATCCGTATGTCCCTGCGTGGAAAAGCTGTTG TGCTGATGGGGAAGAACACGATGATGCGCAAAGCTATTCGTGGTCACC TGGAGAATAACCCTGCACTAGAAAAGCTGCTTCCTCATATCCGTGGGA ATGTAGGCTTTGTCTTCACTAAGGAGGATCTTACTGAGATCCGGGACAT

	GCTGCTGGCTAACAAGGTGCCAGCTGCTGCCCGTGCCGGTGCTATTGCT CCTTGTGATGTGACTGTGCCAGCCCAGAACACAGGCCTTGGACCTGAG AAGACCTCCTTTTTCCAGGCCTTGGGCATCACCACGAAGATTTCCAGAG GGACTATTGAAATTCTGAGCGATGTGCAGCTTATCAAGACTGGAGACA AAGTGGGTGCCAGCGAAGCCACCCTGCTCAACATGCTGAACATCTCCC CGTTCTCCTTTGGGTTGGTGATCCAGCAGGTCTTTGACAATGGCAGCAT TTACAACCCTGAAGTGCTGGAC
EIF1	CCGGCACTGAGGACTACATCCATATAAGGATCCAGCAGCGCAACGGCA GGAAGACCCTCACCACAGTCCAGGGCATCGCGGATGACTACGATAAAA AGAAACTGGTGAAGGCCTTCAAGAAGAAATTTGCCTGCAATGGTACTG TAATTGAGCACCCCGAGTATGGCGAAGTGATTCAGTTGCAGGGTGACC AGCGCAAGAACATCTGCCAGTTCCTCGTAGAGATTGGACTGGCTAAAG ACGACCAGCTGAAAGTCCA
EIF4A2	CCGCGGATTATAGCAGAGACCATGGCGGACCAGAGNNNNNNNNGCCC GATGGNNNCATCGAGAGCAATTGGAATGAGATTGTTGACANNTTCGAT GATATGAATTTAAAAGAATCCCTNNNAAGGGGCATTTACGCTTATGGTT TTGAGAAGNNNTCAGCTATTCAGCAGAGAGCTATTATTCCATGCATCA AAGGGTATGATGTGATTGCTCAAGCTCAGTCAGGTACTGGCAAGACAG CCACATTTGCTATTTCCATCCTGCAGCAGTTGGAGATTGATCTCAAGGA GTCCCAGGCACTAGTATTGGCCCCTACCAGAGAACTGGCTCAACAGAT TCAGAAGGTAATCCTGGCCCTTGGAGACTACATGGGAGCAACATGCCA CGCTTGTATTGGTGGTACAAATGTGCGCAATGAAATGCAGAAACTGCA GGCTGAGGCTCCACACATCGTGGTGGGAACTCCAGGGCGTGTGTTTGA TATGTTAAACAGGCGCTATCTTTCACCAAAATGGATCAAAATGTTTGTT CTGGATGAAGCCGATGAAATGTTGAGCCGTGGATTTAAGGATCAAATT TATGAGATCTTTCAAAAACTAAGCACAAACATCCAGGTTGTGTTGCTGT CAGCTACAATGCCAATGGATGTGTTGGAAGTGACCAAAAAGTTCATGA GAGATCCCATCCGTATTTTGGTGAAGAAGGAAGAACTGACTCTGGAGG GTATCAAGCAGTTCTACATTAACGTTGAGAGAGAGGAATGGAAACTAG ATACTCTCTGTGATCTGTATGAGACACTGACCATTACACAGGCTGTTAT TTTCCTGAATACAAGGAGAAAAGTAGACTGGCTTAC

Tableau 14 : Séquences des portions d'ADNc des gènes clonés de la grande musaraigne et la séquence amplifiée par qPCR (en gris) pour certains de ces gènes.

Gène	ADNc
AhR	GAGCCAAAAGCTTCTCTGATGTTGCATTAAAAGCATCTCCAGCTGACA GAGGCGGAGTACAGGATAACTGTAGAACAAAATTCAGAGAAAGCCTC AACTTACAAGAAGGTGAATTTCTATTACAGGCATTAAATGGCTTTGTGT TGGTTGTTACAACTGATGCCTTGGTTTTTATGCTTCTTCAACCATACAAG ATTACCTAGGATTTCAGCAGTCTGATGTCATACACCAGAGTGTATATGA ACTTAACCATACTGAAGATCGAGCTGAATTTCAACGTCGGCTACACTG GGCATTAGACCCTTCTCGGTGTACTGACTCTGGACAAAGAATTGATGA AGCTAATGGCCTCTCACAGCCATCAGTTTCTTTATAACCCAGACAAACT CCCTCCAGAAAACTCATCTTATATGGAGAGGTGCTTTGTATGCCGACTT AGATGTCTGCTGGATAATTCATCTGGTTTTCTGGCAATGAATTTCCAAG GAAGGTTAAAATATCTTCACGGACAGAACAAGAAAGGGAAAGATGGA TCAGTACTCCCATCGCAGTTAGCCTTGTTTGCAATAGCTACTCCACTTC AACCCCCATCCATCCTTGAAATCAGAACCAAATATTTCATCTTCAGAAC CAAGCACAAATTAGACTTCACACCTATTGGGTGTGATGCCAAAGGAAC AATTGTCTTAGGTTATACAGAGGCAGAGCTGTGCATGAAAGGCTCAGG ATATCAGTTCGTTCATGCTGCTGATATGCATCATTGTGCTGAGTGCCAT ATCCGAATGAT
*CYP1A1 (*ou *CYP1A2)*	GCCCTGAAGAGTTTTGCCATTGCCTCAGACCCGACATCCTCATCCTCCT GCTACCTGGAGGAGCATGTGAGCAAGGAGGCCCAATACCTTATTGGCA AGTTCCAGGAGCTGATGGCAGGGGCTGGGCACTTTGACCCCTATAAAT ATGTAGTGGTGTCTGTGGCCAATGTCATTTGTGCCATATGCTTCGGCCG ACGCTATGACCATGATGACCAGGAGCTGCTTAGCTTAGTGGACCTGAG TAATGAATTTGGGGAGATAGTCGCCTCGGGCAACCTAGCCGACTTCAT CCCCATCCTCCGTTACCTGCCCAATCCTGTCCTGGACAGCTTCAAGGAC TTGAATAAGAAGTTCTATGACTTCATGCAGAAGATGGTCAAGGAACAC TACAAAAAGTTTGAAAAGGGACACATCCGGGACATCACAGACAGCCTG ATTGAGCACTGTCAGGACAAGAAGCTGGATGAGAACGCCAACATCCAG CTGTCTGATGAGAAGATAGTCAATGTGGTCATGGATCTCTTTGGAGCTG GGTTTGACACGGTCACAACTGCCATCTCCTGGGCGCTCATCTACCTGGT GACAAGCCCCAATGTACAGAAAAAGATCCAGAAGGAGCTGGACATGG TGATTGGTGGAGCACGGCAGCCCCGGCTCTCGGACAGACCCCAGCTGC CCTACCTGGAGGCCTTCATCCTGGAGACCTTCCGACACTCCTCCTTCGT CCCCTTCACCATCCCCCATAGTACCACCAGAGACACCAGTCTGCTCGGT TTCTATATCCCCAAGGAACGCTGTGTCTTTGTGAACCAGTGGCAGATCA ACCATGACCCGCAGCTGTGGGGTGACCCATCTGTCTTCCGGCCAGAAA GGTTTCTCACTGCCACTGGCACCATTGACAAGACCCTGAGTGAGAAGA TGATGCTGTTTGGCCTGGGCAAGCGGAAGTGTGTAGGAGAAACCATTG CCCGCTGGGAGATCTTCCTTTTCCTGGCCATC
RPLP0	CTGATGGGCAAGAACACCATGATGCGCAAGGCCATCCGTGGGCACCTG GAGAACAACCCCGCGCTCGAGAAGTTGCTGCCGCACATCCGGGGAAAT

	GTGGGCTTCGTGTTCACCAAGGAGGACCTCACTGAGATCCGGGACATG CTGCTGGCCAATAAGGTGCCAGCTGCCGCCCGCGCAGGTGCCATTGCC CCCTGTGAGGTGACGGTGCCAGCCCAGAATACGGGCCTGGGGCCCGAG AAGACCTCCTTCTTCCAG
EIF1	CGCTATCCAGAACCTCCACTCTTTCGACCCCTTTGCTGATGCAAGTAAG GGTGATGATCTGCTTCCTGCTGGCACTGAGGATTATATCCATATAAGAA TTCAACAGAGAAACGGCAGGAAGACCCTTACCACCGTCCAAGGGATCG CTGATGATTACGATAAAAAGAAACTAGTAAAGGCGTTTAAGAAGAAAT TTGCCTGCAATGGTACTGTAATTGAGCATCCAGAATATGGAGAAGTGA TTCAGCTACAGGGTGACCAGCGCAAGAACATAT
EIF4A2	GTTCAAGGAGACCCAAGCACTAGTATTGGCCCCCACCAGAGAACTGNC TNAACAGATCCAAAAGGTCATTCTCGCACTTGGAGATTATATGGGAGC AACTTGTCATGCATGCATTGGTGGAACCAATGTTCGAAATGAAATGCA AAAACTCCAGGCTGAAGCACCACATATTGTTGTTGGTACACCTGGGAG AGTGTTTGATATGTTAAACAGAAGATATCNTTCTCCAAAATGGATCAA AATGTTTGTTTTGGATGAAGCAGNTGAGATGTTAAGCCGGGGATTTAA GGATNAAATCTATGAGATTTTTCNAAAATTAAATACAAGTATTCAGGTT GTGTTGCTTTCTGCCACAATGCCAACTGATGTGTTGGAAGTGACCAAAA AATTTATGAGAGATCCCATTCGAATCCTCGTGAAAAAGGAAGAGTTAA CCCTTGAAGGAATCAAACAATTTTATATTAATGTTGAGAGAGAGGAGT GGAAGCTGGATACGCTTTGTGACTTGTATGAGACATTGACGATTACGC AGGCTGTCATTTTTCTCAATACAAGACGCAAGGTAGACTGGCTTACAG AGAAAATGCACGCCAGGGACTTCACGGTTTCTGCTTTGCATGGTGACAT GGACCAGAAGGAAA

ANNEXE 4

ANALYSE DES LECTURES DU SÉQUENÇAGE 1 PAR FASTQC

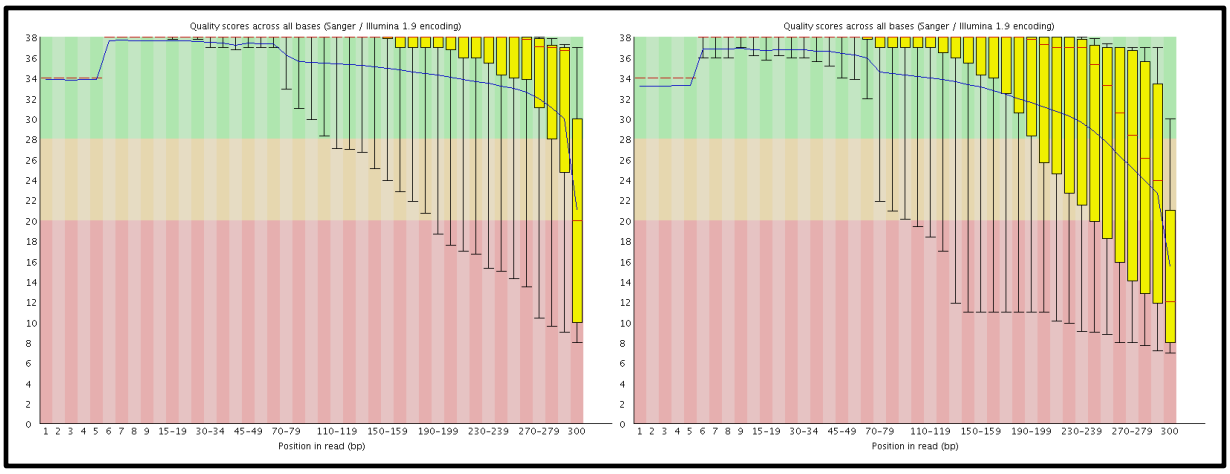


Figure 22 : La moyenne du score Phred pour chaque position pour les lectures du séquençage 1.

Lectures sens : gauche; lectures anti-sens : droite.

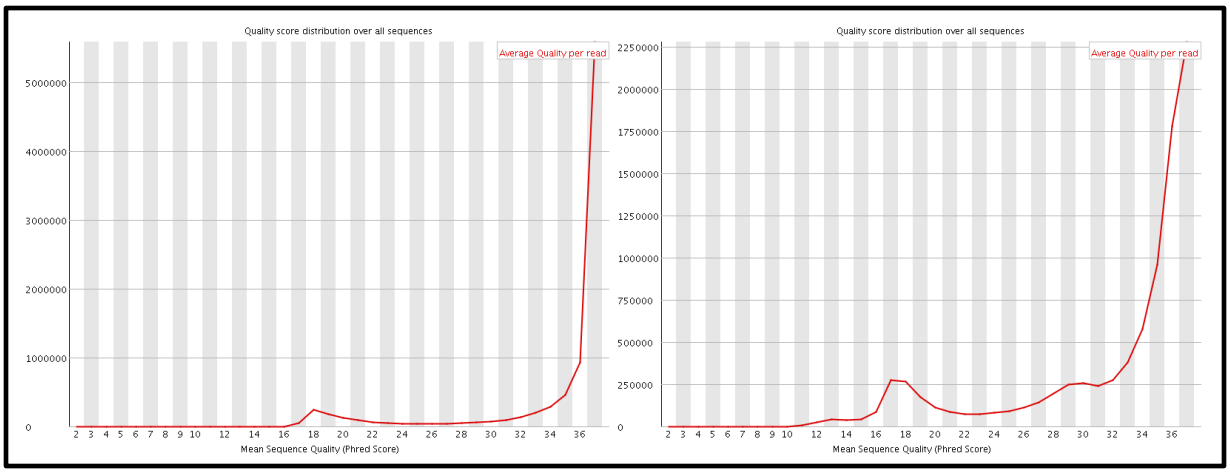


Figure 23 : Le nombre de lectures ayant un score Phred moyen de x pour le séquençage 1.

Lectures sens : gauche; lectures anti-sens : droite.

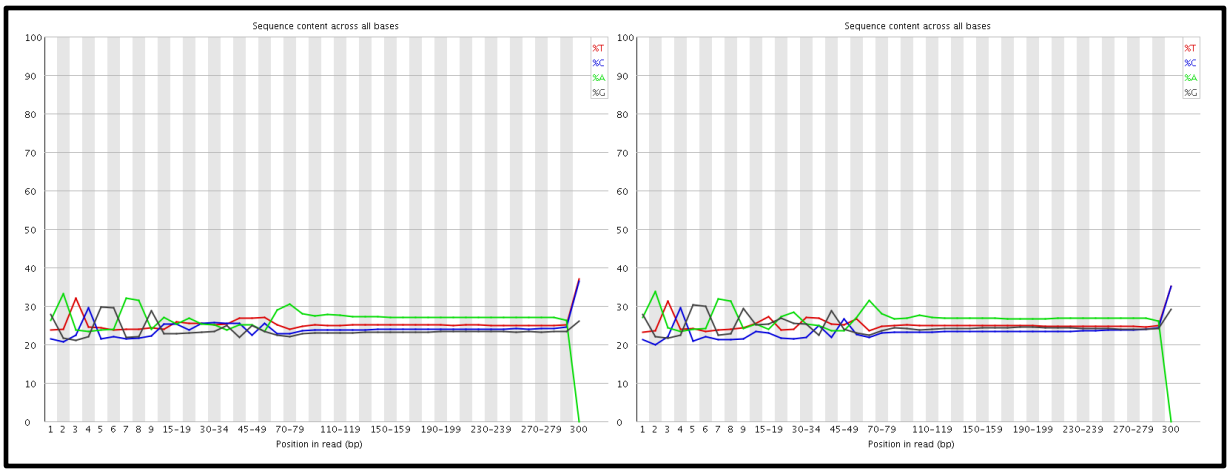


Figure 24 : La répartition des bases selon les positions dans les lectures pour le séquençage 1.

Lectures sens : gauche; lectures anti-sens : droite.

ANNEXE 5

ANALYSE DES LECTURES DU SÉQUENÇAGE 2 PAR FASTQC

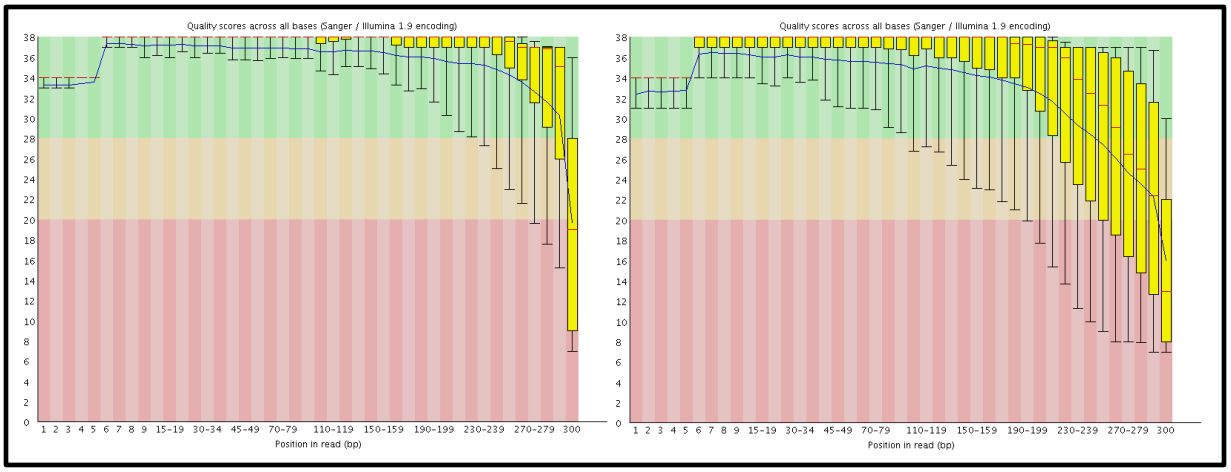


Figure 25 : La moyenne du score Phred pour chaque position pour les lectures du séquençage 2.

Lectures sens : gauche; lectures anti-sens : droite.

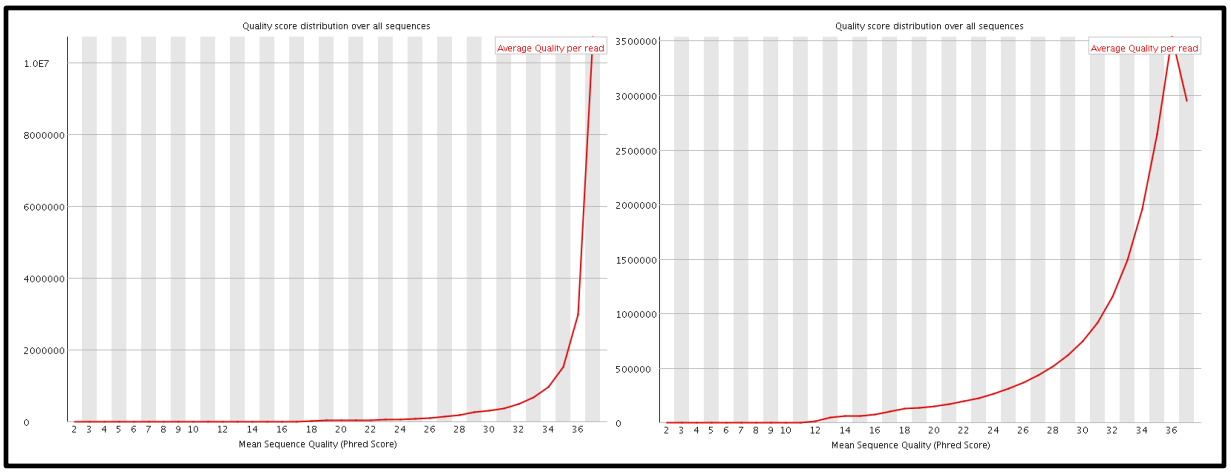


Figure 26 : Le nombre de lectures ayant un score Phred moyen de x pour le séquençage 2.

Lectures sens : gauche; lectures anti-sens : droite.

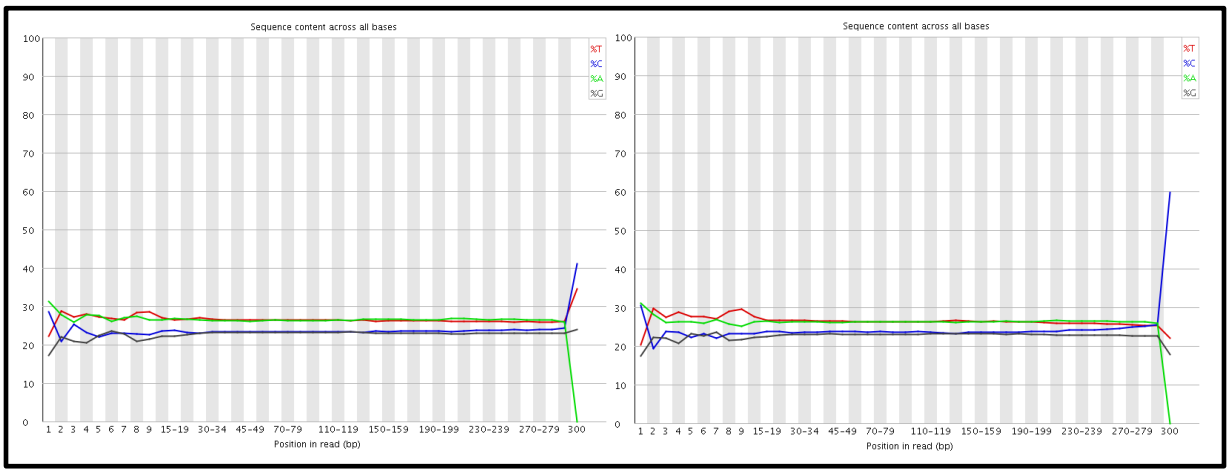


Figure 27 : La répartition des bases selon les positions dans les lectures pour le séquençage 2.

Lectures sens : gauche; lectures anti-sens : droite.

BIBLIOGRAPHIE

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic Local Alignment Search Tool. *Journal of Molecular Biology*, *215*, 403–410.

Bradnam, K. R., Fass, J. N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., … Korf, I. F. (2013). Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience*, *2*, 10.

Cai, Q., Qian, X., Lang, Y., Luo, Y., Xu, J., Pan, S., … Wang, J. (2013). Genome sequence of ground tit Pseudopodoces humilis and its adaptation to high altitude. *Genome Biology*, *14*, R29.

Coumoul, X., Diry, M., Robillot, C., & Barouki, R. (2001). Differential Regulation of Cytochrome P450 1A1 and 1B1 by a Combination of Dioxin and Pesticides in the Breast Tumor Cell Dioxin and Pesticides in the Breast Tumor Cell Line MCF-7 1. *Cancer Research*, 3942–3948.

De Coster, S., & van Larebeke, N. (2012). Endocrine-disrupting chemicals: associated disorders and mechanisms of action. *Journal of Environmental and Public Health*, *2012*, 713696.

Denison, M. S., & Nagy, S. R. (2003). Activation of the aryl hydrocarbon receptor by structurally diverse exogenous and endogenous chemicals. *Annual Review of Pharmacology and Toxicology*, *43*, 309–34.

Ellegren, H., Smeds, L., Burri, R., Olason, P. I., Backström, N., Kawakami, T., … Wolf, J. B. W. (2012). The genomic landscape of species divergence in Ficedula flycatchers. *Nature*, *491*, 756–60.

Ewing, B., & Green, P. (2005). Base-Calling of Automated Sequencer Traces Using. *Genome Research*, 175–185.

Fernandez-Salguero, P., Pineau, T., Hilbert, D., McPhail, T., Lee, S., Kimura, S., … Gonzalez, F. (1995). Immune system impairment and hepatic fibrosis in mice lacking the dioxin-binding Ah receptor. *Science*, *268*, 722–726.

Fletcher, C. M., Pestova, T. V, Hellen, C. U., & Wagner, G. (1999). Structure and interactions of the translation initiation factor eIF1. *The EMBO Journal*, *18*, 2631–2637.

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., … Searle, S. M. J. (2014). Ensembl 2014. *Nucleic Acids Research*, *42*, 749–755.

Frumento, G., Rotondo, R., Tonetti, M., Damonte, G., Benatti, U., & Ferrara, G. B. (2002). Tryptophan-derived Catabolites Are Responsible for Inhibition of T and Natural Killer Cell Proliferation Induced by Indoleamine 2,3-Dioxygenase. *Journal of Experimental Medicine*, *196*, 459–468.

Gilday, D., Gannon, M., Yutzey, K., Bader, D., & Rifkind, a B. (1996). Molecular cloning and expression of two novel avian cytochrome P450 1A enzymes induced by 2,3,7,8-tetrachlorodibenzo-p-dioxin. *The Journal of Biological Chemistry*, *271*, 33054–9.

Gorse, I., & Blag, C. (2013). *Bilan des ventes de pesticides pour l'année 2010*. Canada.

Gupta, M., Mcdougal, A., & Safe, S. (1998). Estrogenic and Antiestrogenic Activities of 16 a - and 2-Hydroxy Metabolites of 17 b -Estradiol in MCF-7 and T47D Human Breast Cancer Cells. *Journal of Steroid Biochemistry and Molecular Biology*, *67*, 413–419.

Hankinson, O. (1995). The aryl hydrocarbon. *Annual Review of Pharmacology and Toxicology*, *35*, 307–340.

Hunter, W. J., & Manter, D. K. (2012). Pseudomonas kuykendallii sp. nov.: a novel γ-proteobacteria isolated from a hexazinone degrading bioreactor. *Current Microbiology*, *65*, 170–5.

Hurd, T., Walker, J., & Whalen, M. M. (2012). Pentachlorophenol decreases tumor-cell-binding capacity and cell-surface protein expression of human natural killer cells. *Journal of Applied Toxicology*, *32*, 627–634.

Hurh, Y.-J., Chen, Z.-H., Na, H.-K., Han, S.-Y., & Surh, Y.-J. (2004). 2-Hydroxyestradiol induces oxidative DNA damage and apoptosis in human mammary epithelial cells. *Journal of Toxicology and Environmental Health. Part A*, *67*, 1939–53.

Kazlauskas, A., Sundström, S., & Poellinger, L. (2001). The hsp90 Chaperone Complex Regulates Intracellular Localization of the Dioxin Receptor. *Molecular and Cellulair Biology*. doi:10.1128/MCB.21.7.2594

L'Héritier, F., Marques, M., Fauteux, M., & Gaudreau, L. (2014). Defining Molecular Sensors to Assess Long-Term Effects of Pesticides on Carcinogenesis. *International Journal of Molecular Sciences*, *15*, 17148–17161.

Lahvis, G. P., Lindell, S. L., Thomas, R. S., McCuskey, R. S., Murphy, C., Glover, E., … Bradfield, C. a. (2000). Portosystemic shunting and persistent fetal vascular structures in

aryl hydrocarbon receptor-deficient mice. *Proceedings of the National Academy of Sciences of the United States of America*, *97*, 10442–7.

Lindsay, S., Chasse, J., Butler, R. a, Morrill, W., & Van Beneden, R. J. (2010). Impacts of stage-specific acute pesticide exposure on predicted population structure of the soft-shell clam, Mya arenaria. *Aquatic Toxicology (Amsterdam, Netherlands)*, *98*, 265–74.

Lindsey, S., & Papoutsakis, E. T. (2012). The evolving role of the aryl hydrocarbon receptor (AHR) in the normophysiology of hematopoiesis. *Stem Cell Reviews*, *8*, 1223–35.

Liu, S.-S., Wang, C.-L., Zhang, J., Zhu, X.-W., & Li, W.-Y. (2013). Combined toxicity of pesticide mixtures on green algae and photobacteria. *Ecotoxicology and Environmental Safety*, *95*, 98–103.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., … Wang, J. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, *1*, 18.

Magoc, T., & Salzberg, S. L. (2011). FLASH : Fast Length Adjustment of Short Reads to Improve Genome Assemblies Tanja Magoˇ. *Bioinformatics Advance Access*, 1–8.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. a, … Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, *437*, 376–80.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, *17*, 10.

Meijer, H. a, Kong, Y. W., Lu, W. T., Wilczynska, A., Spriggs, R. V, Robinson, S. W., … Bushell, M. (2013). Translational Repression and eIFA2 Activity Are Critical for MicroRNA-Mediated Gene Regulation. *Science*, *340*, 82–85.

Mimura, J., & Fujii-Kuriyama, Y. (2003). Functional role of AhR in the expression of toxic effects by TCDD. *Biochimica et Biophysica Acta - General Subjects*, *1619*, 263–268.

Montaño, M., Gutleb, A. C., & Murk, A. J. (2013). Persistent toxic burdens of halogenated phenolic compounds in humans and wildlife. *Environmental Science and Technology*, *47*, 6071–6081.

Nguyen, L. P., & Bradfield, C. a. (2008). The search for endogenous activators of the aryl hydrocarbon receptor. *Chemical Research in Toxicology*, *21*, 102–16.

Opitz, C. a, Litzenburger, U. M., Sahm, F., Ott, M., Tritschler, I., Trump, S., … Platten, M. (2011). An endogenous tumour-promoting ligand of the human aryl hydrocarbon receptor. *Nature*, *478*, 197–203.

Powell, J. B., Goode, G. D., & Eltom, S. E. (2013). The Aryl Hydrocarbon Receptor : A Target for Breast Cancer Therapy. *Journal of Cancer Therapy*, *2013*, 1177–1186.

Rich, B. E., & Steitz, J. a. (1987). Human acidic ribosomal phosphoproteins P0, P1, and P2: analysis of cDNA clones, in vitro synthesis, and assembly. *Molecular and Cellular Biology*, *7*, 4065–4074.

Rioux Paquette, S., Pelletier, F., Garant, D., & Bélisle, M. (2014). Severe recent decrease of adult body mass in a declining insectivorous bird population. *Proceedings. Biological Sciences / The Royal Society*, *281*. doi:10.1098/rspb.2014.0649

Shimada, T., Hayes, C. L., Yamazaki, H., Amin, S., Hecht, S. S., Guengerich, F. P., & Sutler, T. R. (1996). Activation of Chemically Diverse Procarcinogens by Human Activation of Chemically Diverse Procarcinogens by Human Cytochrome. *Cancer Research*, *56*, 2979–2984.

Smeds, L., & Künstner, A. (2011). ConDeTri--a content dependent read trimmer for Illumina data. *PloS One*, *6*, e26314.

Stejskalova, L., Dvorak, Z., & Pavek, P. (2011). Endogenous and Exogenous Ligands of Aryl Hydrocarbon Receptor: Current State of Art. *Current Drug Metabolism*, *12*, 198–212.

Stockinger, B., Di Meglio, P., Gialitakis, M., & Duarte, J. H. (2014). The aryl hydrocarbon receptor: multitasking in the immune system. *Annual Review of Immunology*, *32*, 403–32.

Tsuchiya, Y., Nakajima, M., & Yokoi, T. (2005). Cytochrome P450-mediated metabolism of estrogens and its regulation in human. *Cancer Letters*, *227*, 115–24.

Velisek, J., Kouba, A., & Stara, A. (2013). Acute toxicity of triazine pesticides to juvenile signal crayfish (Pacifastacus leniusculus). *Neuroendocrinology Letters*, *34*, 31–36.

Volk, D. E., Thiviyanathan, V., Rice, J. S., Luxon, B. A., Shah, J. H., Yagi, H., … Gorenstein, D. G. (2003). Solution Structure of a Cis-Opened (10R)-N6-Deoxyadenosine Adduct of (9S,10R)-9-10-Epoxy-7,8,9,10-tetrahydrobenzo[a]pyrene in a DNA Duplex. *Biochemistry*, 1410–1420.

Wirgin, I., Roy, N. K., Loftus, M., Chambers, R. C., Franks, D. G., & Hahn, M. E. (2011). Mechanistic basis of resistance to PCBs in Atlantic tomcod from the Hudson River. *Science (New York, N.Y.)*, *331*, 1322–5.

Yang, X., Solomon, S., Fraser, L. R., Trombino, A. F., Liu, D., Sonenshein, G. E., … Sherr, D. H. (2008). Constitutive regulation of CYP1B1 by the aryl hydrocarbon receptor (AhR) in pre-malignant and malignant mammary tissue. *Journal of Cellular Biochemistry*, *104*, 402–17.

Zhao, Z., Kosinska, W., Khmelnitsky, M., Cavalieri, E. L., Rogan, E. G., Chakravarti, D., … Guttenplan, J. B. (2006). Mutagenic activity of 4-hydroxyestradiol, but not 2-hydroxyestradiol, in BB rat2 embryonic cells, and the mutational spectrum of 4-hydroxyestradiol. *Chemical Research in Toxicology*, *19*, 475–9.

Zhu, B. Z., & Shan, G. Q. (2009). Potential mechanism for pentachlorophenol-induced carcinogenicity: A novel mechanism for metal-independent production of hydroxyl radicals. *Chemical Research in Toxicology*, *22*, 969–977.