Programme 2019/2022 – Action n° 4

Thème *Risques liés à la contamination chimique des milieux aquatiques* -

Protéomique ciblée pour la quantification multiplexée de biomarqueurs: perspectives pour la surveillance environnementale

Application d'une stratégie généralisable entre espèces pour l'identification de protéines d'intérêt en biosurveillance

Rapport d'étape

Arnaud Chaumot, Kévin Sugier, Anabelle Espeyte, Olivier Geffard, Davide Degli-Esposti

Décembre 2020

- **AUTEURS**

Laboratoire d'écotoxicologie, UR RiverLy, INRAE Lyon-Villeurbanne

Arnaud Chaumot, directeur de recherche (INRAE), arnaud.chaumot@inrae.fr

Kévin Sugier, post-doctorant (INRAE), kevin.sugier@inrae.fr

Anabelle Espeyte, ingénieur d'étude (INRAE), anabelle.espeyte@inrae.fr

Olivier Geffard, directeur de recherche (INRAE), olivier.geffard@inrae.fr

Davide Degli-Esposti, chargé de recherche (INRAE), davide.degli-esposti@inrae.fr

Laboratoires partenaires impliqués

UMR SEBIO Reims, contacts : A Geffard, M Palos, R Peden

UMR SEBIO Ineris, contacts : JM Porcher, A Bado-Nilles

UMR SEBIO Le Havre, contact : B Xuereb

UMR LIEBE Université de Lorraine, contact : S Devin

CEA Li2D, contact : J Armengaud

- **CORRESPONDANTS**

OFB : Olivier Perceval, DAST (OFB), olivier.perceval@ofb.gouv.fr

INRAE : Arnaud Chaumot, Directeur de recherche (INRAE), arnaud.chaumot@inrae.fr

- **RESUME**

Bien qu'une attente grandissante existe vis-à-vis de l'application d'outils biologiques pour l'évaluation de la qualité des milieux, l'utilisation des **outils moléculaires** pour la biosurveillance (**biomarqueurs**) reste aujourd'hui limitée, notamment chez les invertébrés, du fait de l'existence de plusieurs verrous techniques et scientifiques. Comme nous avons pu le démontrer par nos travaux chez l'espèce sentinelle *Gammarus fossarum*, la mise en place d'une **surveillance active** qui procède par l'encagement d'organismes calibrés provenant d'une unique population source permet de contrôler l'influence de la variabilité biologique des organismes sentinelles utilisés, facilitant ainsi l'interprétation des marqueurs suivis pour évaluer la qualité des milieux. Par ailleurs, nous avons récemment pu bénéficier chez cette espèce des dernières innovations en termes de biologie moléculaire et notamment pu développer l'approche dite de protéogénomique en collaboration avec le CEA (Marcoule). Ceci a conduit à définir chez ce crustacé une liste de plusieurs centaines de protéines associées à diverses fonctions physiologiques. Dans le même temps, nous avons exploité les performances de **la spectrométrie de masse** en collaboration avec l'ISA (Lyon) faisant la preuve de concept qu'à l'instar des approches multi-résidus développées pour le dosage des contaminants, ces méthodes analytiques de **protéomique ciblée** permettent de quantifier sur un même échantillon et en une unique analyse plusieurs dizaines de ces **protéines** possibles **biomarqueurs**, levant alors le verrou en surveillance de la multitude des méthodologies devant être mises en œuvre aujourd'hui pour la mesure de biomarqueurs moléculaires. Cette action a pour objectif de prolonger cette dynamique vers la mise en œuvre **opérationnelle** de ces méthodologies dans le cadre de la surveillance.

Pour cela, deux axes de travail sont poursuivis. Le premier axe qui a fait l'objet d'un premier rapport d'étape en 2019 est de développer une capacité d'acquisition massive de données en protéomique ciblée pour permettre notamment la définition de valeurs de référence et la construction d'un indicateur intégré de la toxicité des milieux basés sur ces multiples biomarqueurs chez le gammare. Le deuxième axe dont les premiers travaux font l'objet de ce rapport d'étape, vise à ouvrir la réflexion autour de la « protéomique pour la surveillance » à d'autres **espèces d'intérêt environnemental** déjà proposées comme sentinelles pour la surveillance des milieux aquatiques d'**eaux douces, de transition et marines**. Associant différents partenaires scientifiques (SEBIO Reims / Le Havre / Ineris ; LIEBE), l'objectif est de faire la démonstration de l'application d'une stratégie commune de développement de biomarqueurs protéiques mesurés sur organes chez six espèces (gammare, crevette bouquet, crevette blanche, dreissène, moule quagga, épinoche). Cette **stratégie** passe par une première phase de définition des marqueurs protéiques qui s'appuie sur une démarche de protéogénomique (couplage de séquençage du **transcriptome** et d'acquisition de données de **protéomique massive**) permettant de documenter le catalogue de protéines dans les organes d'intérêt chez les différentes espèces. Les peptides rapporteurs de protéines d'intérêt choisies parmi ces catalogues seront dosés *in fine* en spectrométrie de masse ciblée. Nous rapportons ici l'état des lieux des données moléculaires disponibles et manquantes pour l'établissement des catalogues protéiques chez les **six espèces d'étude**, la mise en place des travaux d'acquisition de nouveaux jeux de données de séquences moléculaires, ainsi qu'une première analyse des données moléculaires déjà disponibles.

MOTS CLES (CONTAMINATION CHIMIQUE, TOXICITE, BIOMONITORING ACTIF, PROTEOMIQUE, BIOMARQUEURS, INDICATEUR, BIODIVERSITE, ESPECES)

Table des matières

1. Contexte introductif

Cette action concerne une question nationale relative à la surveillance de la qualité chimique et toxique des milieux aquatiques. L'objectif de l'action qui s'inscrit dans le développement des outils de biosurveillance soutenu depuis une dizaine d'années par le partenariat AFB-Irstea aujourd'hui OFB-INRAE, est de lever certains verrous scientifiques devant permettre l'utilisation de biomarqueurs moléculaires à large échelle, ceci en bénéficiant des dynamiques actuellement mises en place d'une part autour de l'utilisation du gammare encagé dans le cadre de la surveillance chimique des milieux (par les agences de l'eau, suivi DCE) et d'autre part sur la question de l'intégration de la diversité des espèces pour une meilleure surveillance (e.g., projet OFB Sashimi).

L'application de la **DCE** pour surveiller la contamination chimique des eaux de surface consiste actuellement à déterminer si les niveaux de contamination sont conformes aux normes de qualité environnementale réglementaires (NQE). Les NQE sont des concentrations de polluants prioritaires dans l'eau ou dans le biote qui ne doivent pas être dépassées, afin de protéger la santé humaine et l'environnement. Pour la conformité aux NQE-biote, les approches passives (échantillonnage d'organismes résidents) ont été les premières à émerger pour les environnements côtiers avec, par exemple, le "Mussel Watch" initié en 1976 (Borja *et al*., 2008). Ces approches passives sont moins développées pour les masses d'eau continentales qui sont plus complexes au regard de la grande **diversité** des hydrosystèmes et des cortèges d'**espèces aquatiques** associés à ces différents habitats, rendant ainsi difficile le recours à un nombre faible d'espèces indicatrices (géographiquement représentatives et réparties largement sur le territoire). Ceci impose l'étude d'un très grand nombre de sites d'échantillonnage. La **biosurveillance active**, basée sur des organismes transplantés, a récemment été proposée comme approche alternative. Elle a l'avantage d'utiliser une seule espèce sur l'ensemble du territoire, de minimiser la variabilité biologique en utilisant des organismes calibrés (taille, sexe, etc.) provenant de la même population et enfin de maitriser le temps d'exposition (Bervoets *et al*., 2005). Dans ce contexte, le laboratoire d'écotoxicologie de Lyon a développé un outil de diagnostic de la contamination chimique des milieux, basé sur l'encagement de l'amphipode *Gammarus fossarum (*Geffard *et al.,* 2014), qui est actuellement proposé comme un outil de surveillance pour la conformité aux NQE biote[1]. Parallèlement à cette approche qui procède substance par substance, l'évaluation de la qualité des milieux aquatiques peut également s'appuyer sur l'utilisation d'**outils biologiques** permettant d'intégrer l'effet toxique de l'ensemble des contaminants auxquels sont co-exposés les organismes dans les écosystèmes contaminés. Aujourd'hui, les outils écotoxicologiques (bioessais et **biomarqueurs**) ne sont pas encore inclus dans la DCE, mais leur intérêt et leur utilisation ont récemment été interrogés et encouragés, aussi bien au niveau européen au travers des projets comme SOLUTIONS et DEMEAU, favorisant l'utilisation de tests *in vitro*, qu'au niveau national au travers du plan Micropolluants (2016-2021) et de l'appel à manifestation d'intérêt (AMI 2017) porté par l'Agence Française de la Biodiversité, mais également par la mise en place de groupes de travail comme celui porté par INERIS sur le rôle et l'intérêt que doivent avoir les bioessais dans la surveillance des milieux aquatiques. Comme l'illustrent nos travaux sur l'espèce sentinelle *Gammarus fossarum*, les approches *in situ*, basées sur l'encagement d'organismes offrent là aussi l'opportunité d'obtenir des données de biosurveillance dont l'interprétation en termes de qualité toxique des milieux est facilitée par le fait qu'elles limitent l'impact de

[1] EC. Technical Report 2014 – 083. Common implementation strategy for the Water Framework Directive (2000/60/EC). Guidance Document N°32 on biota monitoring (the implementation of EQS-Biota) under the Water Framework Directive.

facteurs biologiques de confusion (genre, statut reproductif, durée d'exposition et historique). Chez *G. fossarum*, différents marqueurs d'effets individuels (activité alimentaire, fécondité, mue) sont ainsi appliqués aujourd'hui dans le cadre de l'évaluation de la qualité des milieux aquatiques au niveau de stations du réseau de surveillance de différentes agences de l'eau françaises (Geffard *et al*. 2019).

Pour les **biomarqueurs moléculaires**, différentes limites techniques contraignent leur utilisation en routine pour la biosurveillance, tels que le manque de méthodes de quantification directe permettant d'assurer la répétabilité/reproductibilité dans le temps, et la comparaison des résultats entre études, avec des unités de mesures qui ne soient pas arbitraires ou dépendantes du protocole. De plus, l'approche multi-biomarqueurs est essentielle pour intégrer une large gamme de typologie de réponses biologiques / physiologiques induites par les différentes classes de contaminants présentes, visant à une évaluation exhaustive des facteurs de stress chimiques. Pour cela, certains indices ont été proposés pour interpréter les modulations de plusieurs biomarqueurs de manière intégrée, comme l'IBR (Beliaeff et Burgeot, 2002). Cependant, et malgré la simplification de l'analyse en intégrant plusieurs biomarqueurs dans un seul indice, il reste nécessaire de mettre en place une méthode spécifique pour chaque biomarqueur. Comme discuté par Trapp *et al*., (2014a), la plupart des biomarqueurs protéiques disponibles aujourd'hui chez les invertébrés reposent sur des méthodes spécifiques à chaque biomarqueur d'intérêt, multipliant le travail de laboratoire, le coût et le temps nécessaire à l'analyse d'un grand nombre d'échantillons, comme cela est imposé dans les programmes de surveillance.

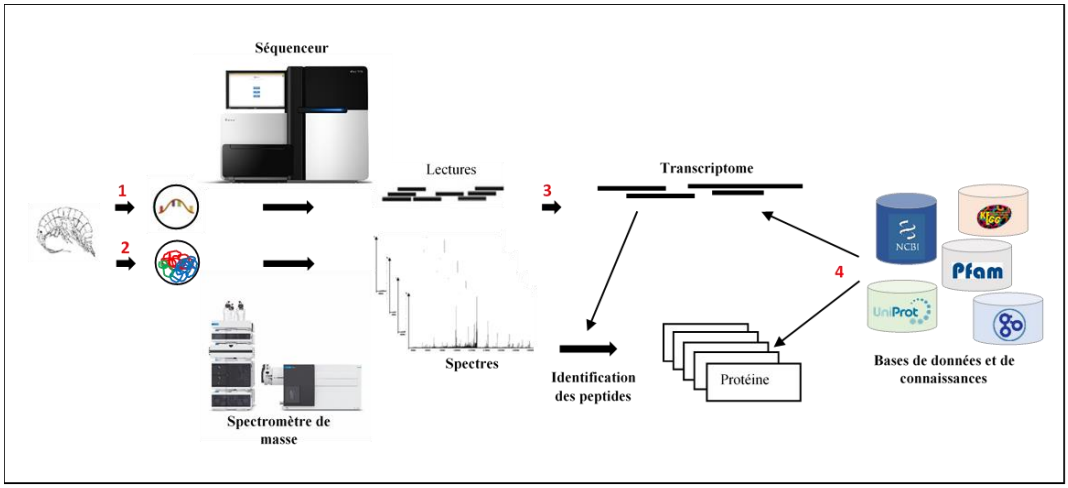

Fig. 1 : *Illustration de l'approche* ***protéogénomique*** *permettant d'établir un catalogue de protéines présentes chez une espèce sentinelle d'intérêt. Echantillonnage (individuel) pour* ***1.*** *extraire l'****ARN total*** *et séquençage après construction des banques Illumina; ou* ***2.*** *disséquer un organe, et extraire les* ***protéines*** *pour analyse en* ***spectromètrie de masse (shotgun****). ****3.*** *Les lectures obtenues par le séquenceur sont nettoyées et alignées entre elles par un assembleur (Trinity, Grabherr et al, 2011) pour construire le* ***transcriptome*** *de référence de l'espèce (****RNASeq****). Après assemblage, une prédiction des séquences codantes est effectuée en utilisant TransDecoder, et permet ainsi d'obtenir une prédiction du* ***protéome*** *(séquences de protéines prédites). Ces prédictions servent ensuite de référence pour interpréter les spectres protéomiques et les assigner à des séquences protéiques compatibles avec les masses enregistrées (****séquences de protéines observées****)* ***4.*** *Dans le but d'identifier les fonctions des protéines prédites ou identifiées, une annotation fonctionnelle peut être réalisée. Pour cela, les séquences protéiques sont alignées contre diverses banques de données composées de séquences connues : par homologie, les caractéristiques (nom, fonction…) des séquences alignées de la base de données sont transférées aux séquences nouvellement construites. Ceci permet de caractériser à un niveau moléculaire l'individu ou les organes échantillonnés.*

S'ajoute à cette lourdeur du développement de biomarqueurs « à façon » et de l'application de leur mesure en routine, la nécessité de pourvoir disposer de marqueurs sur un ensemble **d'espèces** représentatives de la **diversité** des communautés biologiques aquatiques. En effet, un des enjeux en écotoxicologie est de pouvoir intégrer la diversité des sensibilités des espèces cibles pour mieux évaluer l'impact des polluants dans les milieux aquatiques.

Au cours de la dernière décennie, les énormes progrès technologiques réalisés en chimie analytique ont fait émerger des méthodes par **spectrométrie de masse** hautement performantes en biochimie comme le dosage multiplexé de biomarqueurs protéiques dans le domaine du diagnostic médical. Mais, dans le domaine du diagnostic de la toxicité environnementale, l'absence de données génomiques et / ou protéomiques chez les espèces aquatiques sentinelles les plus couramment utilisés en écotoxicologie (notamment invertébrés) limite fortement le développement de biomarqueurs spécifiques et notamment l'application de telles méthodes. Une autre approche analytique, la **protéogénomique** a récemment été proposée comme une approche de rupture en écotoxicologie en réponse à ce verrou (Armengaud *et al*., 2014). Cette approche apparaît comme une alternative au difficile séquençage du génome chez chaque espèce d'intérêt et consiste à séquencer les ARNm matures (partie codante du génome) en parallèle de l'acquisition de données de protéomique massive (approche shotgun) pour obtenir des informations sur les séquences protéiques spécifiques à l'espèce d'intérêt (Figure 1). Aujourd'hui, la technologie RNAseq permet en effet un séquençage profond du **transcriptome** chez n'importe quelle espèce et le transcriptome d'une espèce d'intérêt peut ainsi être facilement établi et utilisé pour interpréter les spectres obtenus pour l'identification des protéines et établir un catalogue de séquences protéiques identifiées par spectrométrie de masse et spécifiques de l'espèce. Ces dernières années, nous avons ainsi développé en collaboration avec le CEA Marcoule (J. Armengaud) des travaux de biologie moléculaire et de découverte protéique chez l'espèce *Gammarus fossarum.* Il a été mis en place pour la première fois à cette échelle chez un invertébré dont le génome n'est pas disponible, une approche de protéogénomique pour la découverte non ciblée de protéines. Cette approche a permis d'établir le protéome d'organes cibles chez *G. fossarum*, et un total de 1800 protéines a été initialement certifié chez cette espèce (Trapp *et al*., 2014b). Aujourd'hui cette liste a été étendue à environ 3000 séquences protéiques expérimentales. Nous avons proposé ensuite de coupler ces approches de découverte aux approches de **protéomiques ciblées** qui offrent de nouvelles opportunités pour le dosage haut débit multiplexé de protéines par spectrométrie de masse (collaboration ISA CNRS Lyon1, A. Salvador) (Gouveia *et al.* 2019). A partir du protéome défini chez *G. fossarum*, et dans le cadre de la thèse de D. Gouveia, une méthode de dosage d'une cinquantaine de peptides biomarqueurs a pu être développée et appliquée sur le terrain à une échelle régionale en recourant à l'approche de biomonitoring actif développée chez notre espèce sentinelle (Gouveia et al 2017).

Objectifs de l'action

Le travail proposé dans cette action engagée depuis 2019 s'articule autour de deux axes

1 – Suite au développement méthodologique qui a abouti l'identification et la quantification simultanée de plusieurs dizaines de protéines biomarqueurs chez le gammare, le premier axe de travail de cette action est de développer chez le gammare une capacité d'acquisition massive de données en protéomique ciblée pour permettre notamment la définition de valeurs de référence et la construction d'un indicateur intégré de la toxicité des milieux basés sur ces multiples biomarqueurs chez le gammare. Ceci est passé par le

développement d'une nouvelle méthode multiplexe qui permet de quantifier simultanément la concentration d'une quarantaine de peptides rapporteurs de grandes fonctions biologiques chez ce crustacé, l'automatisation de la préparation des échantillons, et le test de la robustesse du protocole de préparation et d'analyse via l'acquisition des niveaux de ces biomarqueurs sur 325 échantillons. Ces développements ont fait l'objet d'un premier rapport d'étape en 2019 (Espeyte *et al.* 2019) et ont été poursuivis au cours de 2020.

2- Le deuxième axe dont les premiers travaux font l'objet de ce nouveau rapport d'étape, vise à ouvrir la réflexion « protéomique pour la surveillance » à d'autres **espèces d'intérêt environnemental** déjà proposées comme sentinelles pour la surveillance des milieux aquatiques d'**eaux douces, de transition et marines**. Associant différents partenaires scientifiques (SEBIO Reims, INERIS, LIEBE, SEBIO Le Havre), l'objectif est de faire la démonstration de l'application d'une même stratégie de développement de biomarqueurs protéiques chez six espèces : gammare, crevette bouquet, crevette blanche, dreissène, moule quagga, épinoche. Pour accompagner cet exercice de comparaison et de transférabilité entre espèces, une originalité méthodologique proposée pour cet axe est d'orienter ces développements vers le dosage de biomarqueurs par spectrométrie de masse sur **organes** /tissus, en visant des organes assurant des fonctions analogues chez les six espèces aquatiques choisies (branchies ; sang-hémolymphe; foie-hépatopancréas-glande digestive).

2. Stratégie mise en œuvre

2.1. Démarche générale

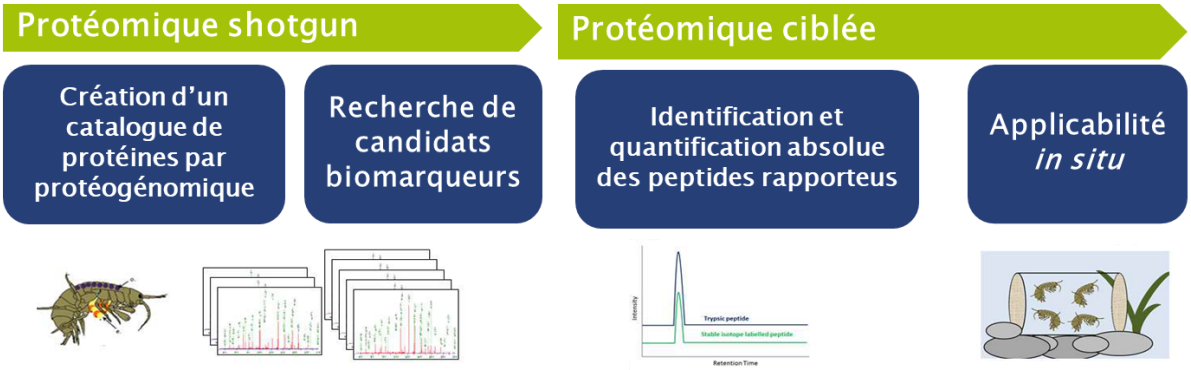

Fig. 2 : *Démarche en deux étapes pour le développement de biomarqueurs protéiques ; couplage des approches de protéomique massive (i.e. shotgun) et protéomique ciblée.*

L'objectif visé est d'appliquer la démarche suivie chez le gammare (démarche en 2 étapes ; Figure 2) pour aboutir au dosage d'une dizaine de biomarqueurs analogues pour six organismes sentinelles des milieux aquatiques dulçaquicoles, estuariens et marins. Les six espèces sélectionnées appartiennent à trois grands groupes animaux : **Poissons, Bivalves et Crustacés**. Pour aboutir à cette sélection, une réunion a rassemblé en mai 2019 les différents partenaires engagés dans cette proposition. Dans un esprit de démonstration et d'opérationnalité, il a semblé en effet pertinent de prendre en compte différents milieux (eaux douces, estuaires, littoral) pour lesquels le développement de la démarche du **monitoring actif** (transplantation d'une population de référence) est mené actuellement sur différentes espèces sentinelles. Le choix a été arrêté sur des marqueurs

d'état de santé (via le suivi de grandes fonctions physiologiques). Aussi nous avons choisi de cibler des protéines effectrices de grandes fonctions telles que l'osmo-régulation, l'immunité, la nutrition, la détoxication, … que l'on retrouve chez l'ensemble des grands groupes animaux (notamment invertébrés et vertébrés). De la même façon, nous avons opté pour une démarche de dosage de biomarqueurs dans 3 types d'organes aux fonctions globalement analogues entre les six espèces :

1/ foie – hépatopancréas – caeca – glande digestive
2/ sang - hémolymphe
3/ branchies

Espèces sentinelles et organes cibles retenus

- **Poissons** organes cibles : Foie Sang Branchie
  - *Gasterosteus*
    - *G. aculeatus* (épinoche)
- **Crustacés**
  - *Palaemon* organes cibles : Hépatopancréas Hémolymphe Branchie
    - *P. serratus* (bouquet)
    - *P. longirostris* (crevette blanche)
  - *Gammarus* (gammare) organes cibles : Caeca Hémolymphe Branchie
    - *G. fossarum*
- **Bivalves** organes cibles : Glande_digestive Hémolymphe Branchie
  - *Dreissena*
    - *D. polymorpha* (dreissène, moule zébrée)
    - *D. r. bugensis* (moule quagga)

Dans la stratégie proposée et rappelée sur la Figure 2, la première étape consiste à définir un catalogue de protéines exprimées dans le tissu ou l'organe cibles chez l'espèce d'intérêt. Pour pouvoir mettre en place l'approche de protéogénomique qui permet d'établir ces catalogues, deux types d'informations moléculaires sont nécessaires pour chacune des populations expérimentales dont sont originaires les organismes sentinelles utilisés: un **transcriptome de référence** qui permet de disposer des séquences protéiques spécifiques à la population d'étude, et une connaissance du sous-ensemble de ces protéines qui s'expriment dans chacun des organes cibles (**protéome d'organe**).

Dans notre étude, la technique de **RNA-seq** est utilisée sur les six espèces retenues afin de connaître les séquences codantes pour les protéines en se basant sur des données acquises **spécifiquement sur chaque population** étudiée. A ce jour, dans les banques de données publiques, le génome annoté de l'épinoche est disponible pour des populations dulçaquicoles (ex Aach, Suisse ; 47◦33′29.25″N, 9◦16′42.38″E) et marines (ex List, Sylt, Allemagne ; 55◦01′49.04″ N, 8◦25′37″ E) (Berner et al. 2019), ainsi que celui de la moule quagga (Calcino et al. 2019), isolée en Espagne. Cependant, notre retour d'expérience sur le développement de dosages ciblés par protéomique chez différentes espèces environnementales nous a amenés à conclure que s'appuyer sur des séquences publiques d'espèces proches, ou même sur la même espèce mais sur des populations distantes, amène à de nombreux « trous dans la raquette » du fait du polymorphisme de séquence possible au niveau populationnel. Ces variations de séquences protéiques empêchent ainsi la détection des peptides rapporteurs de certaines protéines d'intérêt. En effet, la variation d'un seul acide aminé rend inopérant la détection par spectrométrie de masse. C'est pourquoi nous allons nous appuyer pour nos travaux sur **un transcriptome**

spécifiquement acquis sur chacune des populations expérimentales de l'étude. Le RNA-seq de référence à utiliser ne sera pas nécessairement acquis sur les organes ciblés par le dosage final afin de s'appuyer pour chacune des espèces sur un RNA-seq profond réalisé sur des échantillons où l'extraction de matériel nucléotidique est le plus efficace (*e.g.*, organismes entiers ou échantillon composite de différents organes). Le séquençage de l'ARN en profondeur permet en effet l'identification de la plupart du génome codant le protéome de l'organisme.

Les transcriptomes de référence permettent ainsi de prédire de façon très large la grande majorité des séquences protéiques qui peuvent être retrouvées chez l'organisme d'intérêt. L'autre élément nécessaire à la démarche de sélection de biomarqueur est une connaissance du sous-ensemble de protéines spécifiquement exprimées dans l'organe cible. La caractérisation de ce protéome organe-spécifique s'appuie dans notre démarche sur l'acquisition de données de protéomique massive sur chacun des organes/tissus visés pour le dosage de biomarqueurs. La **protéomique shotgun** (couplée à la connaissance du transcriptome de l'espèce) permet en effet d'établir sans *a priori* la connaissance des protéines exprimées dans les tissus/organes cibles (Figure 1). Pour réaliser cette définition des sous-ensembles de protéines exprimés dans les organes ciblés dans chacune des six espèces, nous avons choisi de n'investiguer qu'une espèce par genre en faisant le pari que le protéome organe spécifique pourra être extrapolé entre espèces proches (ici entre Dreissènes ou entre Palaemons) en s'appuyant sur les relations d'orthologie entre les gènes codants identifiés dans les transcriptomes (on fait le pari que les fonctions sont conservées à ces échelles phylogénétiques faibles).

Le développement du dosage multiplexé en protéomique ciblée s'effectuera dans une deuxième étape (à partir du printemps 2021) sur des échantillons des trois tissus ciblés sur chacune des six espèces, en visant des biomarqueurs d'intérêt qui seront choisis parmi les catalogues protéomiques établis grâce à ces analyses protéogénomiques.

2.2. Données moléculaires disponibles et acquisition de nouvelles données nécessaires chez les 6 espèces d'étude

Espèce sentinelle	**Gammare**	**Dreissène**	**Moule quagga**
Groupe taxonomique Milieu	*Gammarus fossarum* Crustacés Eaux douces	*Dreissena polymorpha* Bivalves Eaux douces	*Dreissena r. bugensis* Bivalves Eaux douces
Population d'étude	population référence inrae Lyon (Ain)	population référence SEBIO Reims	population d'étude (Lorraine) LIEBE
Transcriptome sur la population d'étude	disponible (corps entier mâle / femelle)	disponible (hémolymphe / branchies / glande digestive)	disponible (glande digestive)
Catalogues protéomique massive sur organe (shotgun)	**disponible sur hépatopancréas** **à acquérir sur hémolymphe, branchies**	**disponible sur hémolymphe, glande digestive** **à acquérir sur branchies**	disponible sur glande digestive évalués à partir de *D. polymorpha* pour hémolymphe et branchies

Espèce sentinelle	**Epinoche**	**Crevette bouquet**	**Crevette blanche**
Groupe taxonomique Milieu	*Gasterosteus aculeatus* Poissons Eaux douces/marines	*Palaemon serratus* Crustacés Eaux marines	*Palaemon longirostris* Crustacés Eaux de transition
Population d'étude	élevage ineris	population d'étude en Cotentin SEBIO Le Havre	population d'étude en Seine SEBIO Le Havre
Transcriptome sur la population d'étude	**à acquérir**	**à acquérir**	**à acquérir**
Catalogues protéomique massive sur organe (shotgun)	disponible dans la littérature sur d'autres populations	**à acquérir sur hémolymphe, branchies, hépatopancréas**	évalués à partir de *P. serratus*

Tab. 1 : *Inventaire des ressources moléculaires mobilisables ou à acquérir chez les 6 espèces d'étude*

Le Tableau 1 fait le bilan des bases de données de séquences nécessaires et disponibles sur les populations choisies.

Un transcriptome pour chacune des deux populations de dreissènes étudiées est disponible (Péden et al. 2019) ainsi que pour la population de gammares utilisée par INRAE (Cogne et al. 2019). Les caractéristiques de ces transcriptomes sont détaillées dans le Tableau 2.

Espèces	*G. aculeatus (marin)*	*G. aculeatus (eau douce)*	*D. polymorpha*	*D. r. bugensis*	*G.fossarum B*
Technologie séquençage	HiSeq X		HiSeq 3000		HiSeq 3000
Méthode assemblage	Supernova v. 1.20		DRAP Oases 1.7		Trinity 2.4
Nbr bases	417 493 368	452 537 677	103 039 811	97 982 186	263 406 154
Nbr contigs	35 293	32 646	44 538	49 679	344 409
N50	396	3 636	3 094	2 674	1 354

Tab. 2 : *Caractéristiques des transcriptomes de référence disponibles chez les populations de dreissènes, moules quagga et de gammares de l'étude. Celui de deux populations d'épinoche de la littérature est également intégrés pour comparaison, le transcriptome spécifique de la population utilisée par l'ineris étant en cours de séquençage.*

Pour posséder un transcriptome de référence pour chacune des populations d'étude des six espèces sentinelles cibles, la construction des trois transcriptomes manquants (l'épinoche et les deux crevettes *Palaemon*) est en cours. Les échantillons en cours de séquençage sur la plateforme GenoToul (technologie Illumina NovaSeq) ont été construits à partir d'un seul individu de chacune des trois espèces en partant d'un mélange des ARN issus de différents tissus de ces individus (foie, muscle et rein pour l'épinoche ; hépatopancréas, muscle, céphalon pour le bouquet), dans un esprit de couverture large du transcriptome pour les bases de données de référence qui seront obtenues (Figure 3).

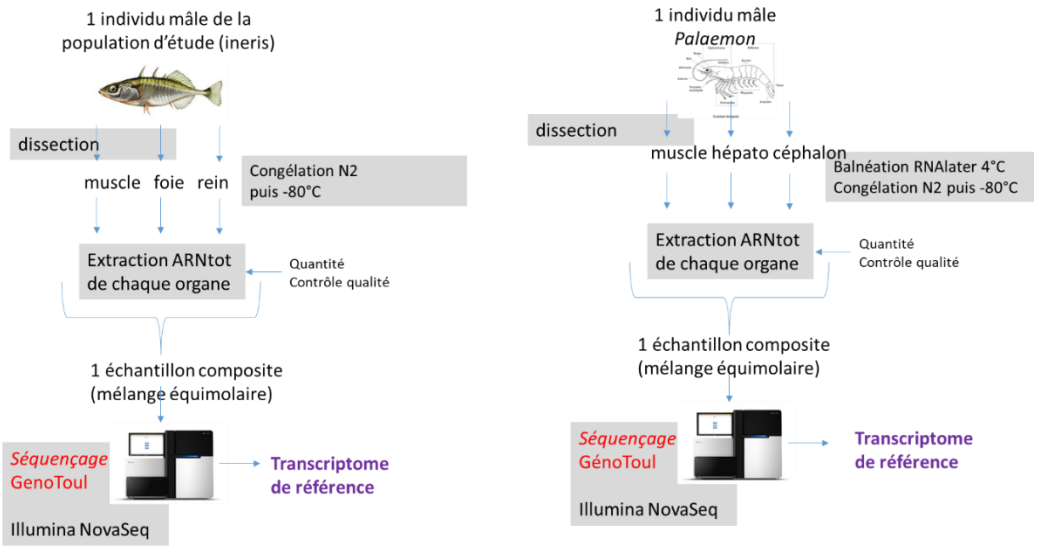

Fig. 3 : *Protocole suivi pour l'obtention des transcriptomes de référence pour la population d'épinoches de l'étude et pour les 2 populations (marine et estuarienne) de crevettes* Palaemon

Concernant le volet protéomique massive sur organe, des données sont déjà disponibles pour sept des couples espèce/organe de l'étude (Tableau 1). Pour l'épinoche, des connaissances protéomiques sont disponibles dans la littérature (Kültz et al. 2015; Li et al. 2018; Li et Kültz 2020). Quatre jeux de données shotgun proteomics sont dejà disponibles dans nos laboratoires : *Gammarus*-caeca, *Dreissena_polymorpha*-glande_digestive, *Dreissena_polymorpha*-hémolymphe, *Dreissena_quagga*-glande_digestive.

Et tenant compte de notre stratégie d'extrapolation au sein d'un même genre, l'acquisition de six nouveaux protéomes d'organes a été nécessaire à l'étude (Tableau 1). Cette acquisition est en cours : les échantillons d'organes obtenus sur gammare, palaemon et dreissène à l'automne (Figure 4) sont actuellement analysés par la plateforme ProGenoMix du CEA. L'analyse protéomique des échantillons d'extraits de protéines par spectrométrie de masse en tandem haute résolution permettra ainsi d'identifier les composants protéiques présents dans ces organes et d'estimer leurs quantités. Les protéines de chaque échantillon seront séparées sur gel SDS-PAGE en 5 fractions différentes selon leur masse moléculaire. Après digestion trypsique, chacune de ces fractions sera analysée séparément par spectrométrie de masse en tandem à l'aide d'un Q-Exactive HF (Thermo). L'analyse des données de spectrométrie de masse en tandem sera réalisée en se rapportant au transcriptome de référence de chaque population pour l'annotation des spectres obtenus et établir les catalogues protéiques de chaque organe.

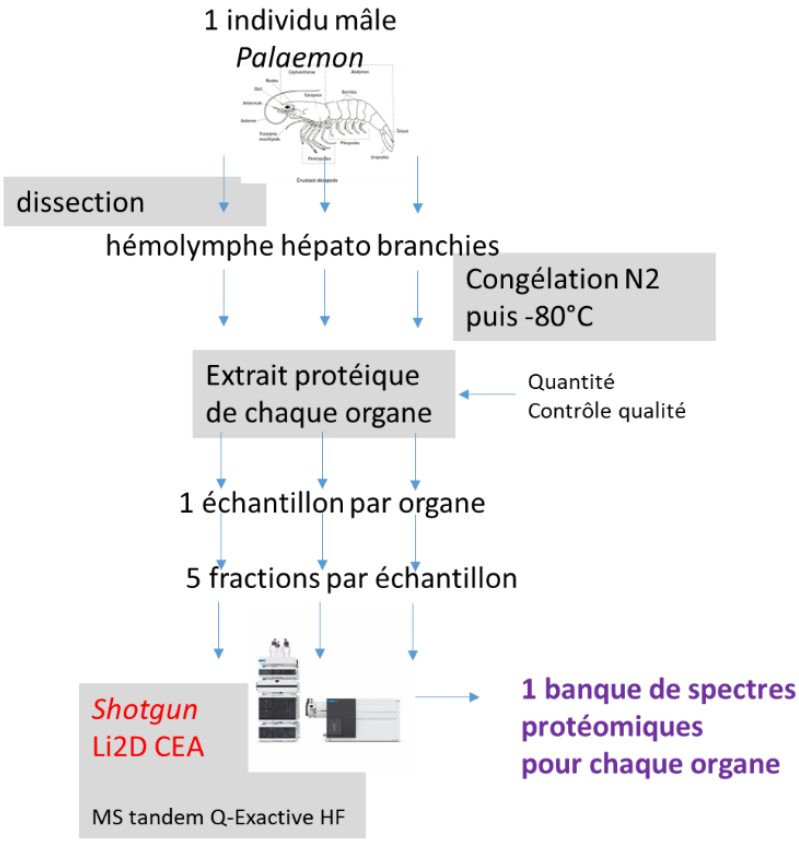

Fig. 4 : *Protocole mis en place pour l'obtention des protéomes d'organes pour protéomique shotgun : exemple de* Palaemon.

3. Premières analyses fonctionnelles des répertoires protéomiques

Différentes analyses préliminaires ont pu être conduites fin 2020 sur les transcriptomes et les protéomes d'organes disponibles. Elles ont visé tout d'abord à évaluer l'homogénéité de la qualité des transcriptomes obtenus par nos laboratoires respectifs sur gammares et moules zébrées et quagga (INRAE, SEBIO Reims, LIEBE), Une première analyse globale des répertoires fonctionnels représentés dans ces transcriptomes ainsi que dans le protéome obtenu antérieurement sur caeca de gammare a permis de sonder si l'approche envisagée permettra de cibler des biomarqueurs de différentes grandes fonctions biologiques.

3.1. Analyse des transcriptomes

En attendant le séquençage et les assemblages en cours, les transcriptomes d'épinoches de populations différentes de celles étudiés sont intégrées dans cette première analyse. Deux RNA-seq sont disponibles pour des épinoches d'Europe, l'un pour une population marine allemande, l'autre pour une population dulçaquicole suisse. Les RNA-seq des épinoches de cette analyse ne sont pas ceux qui seront utilisés *in fine* car provenant de populations trop éloignées de celle étudiée, mais ils servent ici d'indicateurs pour nos analyses futures. A l'écriture de ce rapport sont ainsi disponibles des transcriptomes pour quatre des six espèces de l'étude (Tableau 2). Un RNA-seq est disponible pour chacune des deux populations de moules étudiées ; et un RNA-seq pour la population de gammare étudiée.

A partir des données RNA-seq, des protéomes ont été prédits *in silico.* Un grand écart du nombre de protéines prédites peut être observé entre les cinq transcriptomes : de ~29 000 pour la moule quagga à ~177 000 pour le gammare (Tableau 3). Ceci peut s'expliquer notamment par deux facteurs : (i) les méthodes d'annotation qui ont été utilisés sont différentes; (ii) la taille des génomes prédite est très variable : environ 600 Mégabases pour l'épinoche, autour de 1,5 Gigabases pour les moules et plus de 6 Gb pour les gammares (Gregory 2005, données non publiées), impliquant potentiellement l'existence de nombreux pseudogènes donnant lieu à des faux positifs dans ces prédictions. Nous avons ensuite réalisée une annotation fonctionnelle de ces cinq protéomes prédits en utilisant InterProScan (Jones et al. 2014). En utilisant l'une des banques de données, Protein family (Pfam), entre 33% et 69% des protéines prédites n'ont pas de domaines protéiques identifiés (El-Gebali et al. 2019). Les espèces sentinelles sélectionnées n'étant que peu étudiées à l'échelle moléculaire, il existe ainsi chez ces espèces de nombreuses protéines inconnues pour les banques de données ou alors elles sont trop divergentes de celles des espèces modèles bien documentées dans ces bases de données publiques. Ceci renforce la nécessité de développer des ressources spécifiques sur ces taxons pour nos analyses et pour la communauté scientifique.

Néanmoins, pour évaluer la couverture du transcriptome global dans ces jeux de séquences RNA-seq disponibles, la méthode de recherche des gènes communs entre

espèces animales (BUSCO) a été utilisée sur les cinq transcriptomes (Seppey, Manni, et Zdobnov 2019). Ainsi, d'après cette approche, les protéomes sont complets entre 88% (moule zébrée) et 96% (gammare). Sachant que l'approche est plus adaptée aux arthropodes, ces résultats valident la qualité satisfaisante des ressources pour retrouver les protéines actrices des différentes fonctions qui seront recherchées dans les protéomes d'organes.

Espèces	*G. aculeatus (marin)*	*G. aculeatus (eau douce)*	*D. polymorpha*	*D. r. bugensis*	*G.fossarum B*
Nbr protéines prédites	55 347	54 954	30 842	29 066	177 384
Nbr protéines annotées (Pfam & 10^{-3})	20 811	20 578	21 566	19 629	55 977
Complet	87,5%	87,4%	86,5%	86,3%	90,7%
Fragment	4,9%	5,9%	1,5%	2,1%	5,3%
Non trouvé	7,6%	6,7%	12%	11,6%	4%

Tab. 3 : *Qualité des données transcriptomiques disponibles*.

Les analyses préliminaires des cinq annotations fonctionnelles de ces transcriptomes, notamment via la base de connaissance Gene Ontology (GO terms), permettent d'obtenir une vision des principaux grands processus biologiques représentés dans les transcriptomes de chacune de ces espèces (Figure 5). Les processus majeurs représentés sont presque identiques pour les cinq organismes, mais avec une répartition différente : ainsi est observée une plus grande présence de processus métaboliques chez le crustacé (autour de 40%) et les bivalves (autour de 38%) que chez les poissons (autour de 24%). En revanche, ces derniers ont une plus grande présence de processus de régulation (autour de 20%, contre moins de 10% pour les trois autres espèces). Chez *D. r. bugensis,* il semble qu'un plus grand nombre de processus du système immunitaire (>1% contre < 0,6% pour les quatre autres espèces) soient visibles.

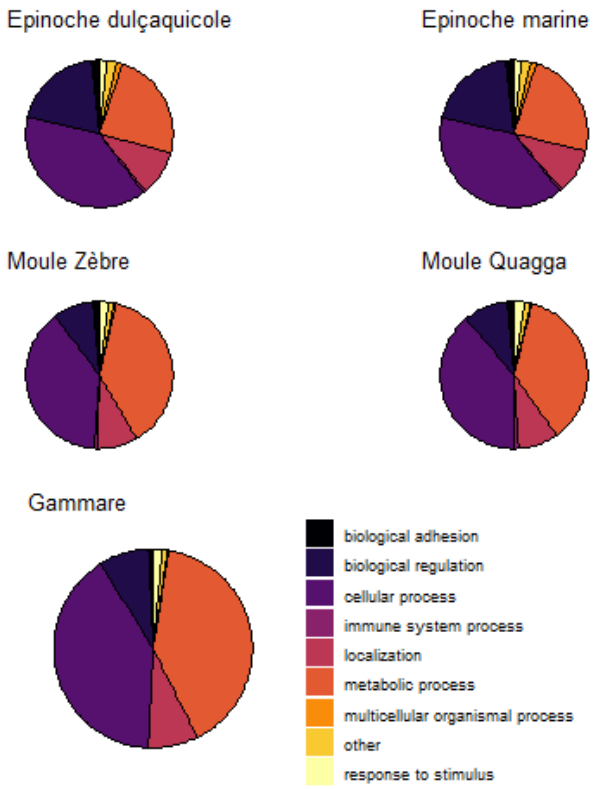

Fig. 5 : *Processus biologiques représentés dans l'annotation fonctionnelle des 5 transcriptomes analysés (GOterms).*

3.2. Premier profil de protéomes sur organe

Pour première exploration, nous avons analysé le jeu de données de protéomique shotgun obtenu sur caeca de gammares (~ 10 protéomes individuels), organe potentiellement impliqué dans les processus de digestion et de détoxication chez ce crustacé. Les analyses ont été effectuées à l'aide des bases de données KEGG (Kanehisa et Goto 2000) qui permettent d'affecter des fonctions biologiques aux protéines extraites et identifiées sur la base de leur similarité de séquence avec des protéines décrites chez d'autres espèces (Tableau 4). Ainsi, sur les 2 568 protéines uniques identifiées dans ce jeu de données, 1 520 (59%) sont associées à une annotation des bases de données KEGG. Comme attendu chez cet organe, ce sont surtout des fonctions du métabolisme et du catabolisme qui sont observés : enzymes de la glycolyse, métabolisme des lipides, des acides aminées, protéines connues pour jouer un rôle dans l'endocytose, le lysosome ou le phagosome… On peut noter que la voie de la biodégradation et du métabolisme des xénobiotiques rassemble des protéines qui ont un intérêt fort en écotoxicologie : cytochrome P450, enzymes dégradants les aminobenzoates, caprolactames ou des composés organochlorés. Ce sont de bons candidats à retrouver chez les autres espèces sentinelles.

KEGG pathway	Percentages
Carbohydrate metabolism	9.52381
Transport and catabolism	7.42646
Signal transduction	7.37897
Lipid metabolism	6.66403
Endocrine system	6.12584
Translation	5.64833
Amino acid metabolism	5.56655
Folding, sorting and degradation	4.5403
Immune system	4.1393
Energy metabolism	3.88603
Digestive system	3.39005
Xenobiotics biodegradation and metabolism	3.38478
Metabolism of other amino acids	3.18955
Nervous system	3.18428
Cell growth and death	2.89935
Metabolism of cofactors and vitamins	2.79119
Cellular community - eukaryotes	2.32951
Environmental adaptation	1.98655
Glycan biosynthesis and metabolism	1.91268
Aging	1.65677
Circulatory system	1.44308
Excretory system	1.4167
Nucleotide metabolism	1.39296
Cell motility	1.29007
Metabolism of terpenoids and polyketides	1.27424
Biosynthesis of other secondary metabolites	1.26369
Transcription	1.22147
Sensory system	0.825749
Development and regeneration	0.804643
Signaling molecules and interaction	0.720222
Cellular community - prokaryotes	0.358792

Tab. 4 : *Voies biologiques KEGG identifiées au sein du jeu de données protéomique shotgun acquis sur caeca de gammare (G. fossarum)*

4. Conclusion et suite des travaux

Les travaux engagés dans cette action autour de la démonstration de l'approche « protéomique en biosurveillance » ont permis d'établir avec un réseau de laboratoires nationaux une stratégie partagée et commune qui est appliquée sur six espèces sentinelles issues de différents groupes taxonomiques (poissons, bivalves, crustacés) et d'intérêt actuel pour la surveillance des milieux aquatiques le long du continuum eaux-douces, eaux de transition et eaux marines.

La première étape d'obtention de catalogues de protéines exprimées dans les organes d'intérêt dans lesquels seront dosés différents biomarqueurs de grandes fonctions biologiques est en cours de finalisation via l'acquisition de transcriptomes de référence pour l'ensemble des populations d'étude et de données de protéomique massive pertinentes pour chacune des six espèces considérées. Les premières analyses fonctionnelles des protéines des jeux de données déjà disponibles (transcriptomes et protéomes) montrent

une couverture assez grande des fonctions biologiques qui pourront être suivies par le choix de biomarqueurs qui sera opéré à partir de ces catalogues protéiques organe-spécifique.

Un travail de sélection des biomarqueurs à suivre sera réalisé avec les laboratoires partenaires de l'action en se basant sur ces catalogues et l'annotation fonctionnelle qui en sera faite chez chacune des espèces. Le développement de méthodes de dosage multiplexé de ces biomarqueurs en protéomique ciblée s'effectuera dans une deuxième étape sur des échantillons de trois tissus ciblés sur chacune des six espèces. Ceci aboutira à une démonstration de l'opérationnalité de la démarche proposée ici comme stratégie généralisable entre espèces pour l'identification de protéines d'intérêt en biosurveillance.

5. Références

Armengaud et al., (2014). Non-model organisms, a species endangered by proteogenomics. J. Proteo, 105 (2014): 5-18.

Beliaeff & Burgeot (2002). Integrated biomarker response: A useful tool for ecological risk assessment. Environ Toxicol Chem 21: 1316-1322.

Berner, Daniel, Marius Roesti, Steven Bilobram, Simon K Chan, Heather Kirk, Pawan Pandoh, Gregory A Taylor, Yongjun Zhao, Steven J M Jones, et Jacquelin DeFaveri. 2019. « De Novo Sequencing, Assembly, and Annotation of Four Threespine Stickleback Genomes Based on Microfluidic Partitioned DNA Libraries », 6.

Bervoets et al. (2005). Use of transplanted Zebra mussels (Dreissena polymorpha) to assess the bioavailability of microcontaminants in Flemish surface waters. Environ Sci Technol 39: 1492-505.

Borja et al. (2008). Overview of integrative tools and methods in assessing ecological integrity in estuarine and coastal systems worldwide. Mar. Pollut. Bull. 56: 1519-1537.

Calcino, Andrew D, André Luiz de Oliveira, Oleg Simakov, Thomas Schwaha, Elisabeth Zieger, Tim Wollesen, et Andreas Wanninger. 2019. « The quagga mussel genome and the evolution of freshwater tolerance ». DNA Research 26 (5): 411-22. https://doi.org/10.1093/dnares/dsz019.

Cogne, Yannick, Davide Degli-Esposti, Olivier Pible, Duarte Gouveia, Adeline François, Olivier Bouchez, Camille Eché, et al. 2019. « De Novo Transcriptomes of 14 Gammarid Individuals for Proteogenomic Analysis of Seven Taxonomic Groups ». Scientific Data 6 (1): 184. https://doi.org/10.1038/s41597-019-0192-5.

El-Gebali, Sara, Jaina Mistry, Alex Bateman, Sean R Eddy, Aurélien Luciani, Simon C Potter, Matloob Qureshi, et al. 2019. « The Pfam protein families database in 2019 ». Nucleic Acids Research 47 (D1): D427-32. https://doi.org/10.1093/nar/gky995.

Espeyte et al. (2019). Protéomique ciblée pour la quantification multiplexée de biomarqueurs: perspectives pour la surveillance environnementale - Cas d'étude chez *Gammarus fossarum.* Rapport d'étape, programme OFB-INRAE 2019-2022.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 15;29(7):644-52.

Geffard et al. (2019). Lien entre la toxicité, la contamination des milieux aquatiques mesurés chez Gammarus fossarum et la perturbation des communautés biologiques - Proposition de valeurs seuils au niveau national pour la mesure de marqueurs de toxicité chez Gammarus fossarum. Rapport AFB – programme 2016-2018.

Geffard et al. (2014). Développement d'une méthodologie pour l'amélioration du suivi chimique des eaux continentales Rapport de synthèse de l'étude pilote - Déploiement de l'outil gammare encagé au niveau national, résultats pour les métaux ciblés. Rapport Onema – programme 2013-2015.

Gouveia et al. (2017). Ecotoxico-proteomics for aquatic environmental monitoring: first in situ application of a new proteomics-based multibiomarker assay using caged amphipods. Environ Sci Technol 51: 13417-13426.

Gouveia D, et al (2019). Ecotoxicoproteomics: A decade of progress in our understanding of anthropogenic impact on the environment. Journal of Proteomics, 198:66-77.

Gregory, T.R. 2005. « Animal Genome Size Database:: Home ». 2005. http://www.genomesize.com.

Jones, P., D. Binns, H.-Y. Chang, M. Fraser, W. Li, C. McAnulla, H. McWilliam, et al. 2014. « InterProScan 5: Genome-Scale Protein Function Classification ». Bioinformatics 30 (9): 1236-40. https://doi.org/10.1093/bioinformatics/btu031.

Kanehisa, M., et S. Goto. 2000. « KEGG: Kyoto Encyclopedia of Genes and Genomes ». Nucleic Acids Research 28 (1): 27-30. https://doi.org/10.1093/nar/28.1.27.

Kültz, Dietmar, Johnathon Li, Xuezhen Zhang, Fernando Villarreal, Tuan Pham, et Darlene Paguio. 2015. « Population-Specific Plasma Proteomes of Marine and Freshwater Three-Spined Sticklebacks (Gasterosteus Aculeatus) ». Proteomics 15 (23-24): 3980-92. https://doi.org/10.1002/pmic.201500132.

Li, Johnathon, et Dietmar Kültz. 2020. « Proteomics of Osmoregulatory Responses in Threespine Stickleback Gills ». Integrative and Comparative Biology 60 (2): 304-17. https://doi.org/10.1093/icb/icaa042.

Li, Johnathon, Bryn Levitan, Silvia Gomez-Jimenez, et Dietmar Kültz. 2018. « Development of a Gill Assay Library for Ecological Proteomics of Threespine Sticklebacks (Gasterosteus Aculeatus) ». Molecular & Cellular Proteomics: MCP 17 (11): 2146-63. https://doi.org/10.1074/mcp.RA118.000973.

Péden, Romain, Pascal Poupin, Bénédicte Sohm, Justine Flayac, Laure Giambérini, Christophe Klopp, Fanny Louis, et al. 2019. « Environmental Transcriptomes of Invasive Dreissena, a Model Species in Ecotoxicology and Invasion Biology ». Scientific Data 6 (1): 234. https://doi.org/10.1038/s41597-019-0252-x.

Seppey, Mathieu, Mosè Manni, et Evgeny M. Zdobnov. 2019. « BUSCO: Assessing Genome Assembly and Annotation Completeness ». In Gene Prediction: Methods and Protocols, édité par Martin Kollmar, 227-45. Methods in Molecular Biology. New York, NY: Springer. https://doi.org/10.1007/978-1-4939-9173-0_14.

Trapp et al. (2014a). Next-generation proteomics: Toward customized biomarkers for environmental biomonitoring. Environ Sci Technol 48: 13560-13572

Trapp J, et al (2014b) Proteogenomics of Gammarus fossarum to Document the Reproductive System of Amphipods. Molecular & Cellular Proteomics, 13(12):3612-3625.