Bioinformatique pour l'exploration de la diversité inter-espèces et inter-populations : hétérogénéité & données multi-omiques

Yannick Cogne

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITÉ DE MONTPELLIER

En Bioinformatique

École doctorale : Sciences Chimiques et Biologiques pour la Santé (CBS2 ED n°168)

Unité de recherche : Laboratoire d'Innovations technologiques pour la détection et le diagnostic (Li2D)

Bioinformatique pour l'exploration de la diversité inter-espèces et inter-populations: hétérogénéité & données multi-omiques

Présentée par Yannick Cogne
Le 7 Octobre 2019

Sous la direction de Jean Armengaud
et encadré par Christine Almunia

Devant le jury composé de

Mme Ana Maria Varela Coelho, Investigador Auxiliar, Université de Lisbonne	Rapporteur
Mme Christine Carapito, Chargée de Recherche, Université de Strasbourg	Rapporteur
Mr Jacques Colinge, Professeur, Université de Montpellier	Président
Mr Michel Hébraud, Directeur de Recherche, INRA Clermont-Ferrand	Examinateur
Mr Jean Armengaud, Directeur de Recherche, CEA Marcoule	Directeur de thèse
Mme Christine Almunia, Chargée de recherche, CEA Marcoule	Encadrante de thèse

UNIVERSITÉ DE MONTPELLIER

« Les espèces qui survivent ne sont pas les espèces les plus fortes, ni les plus intelligentes, mais celles qui s'adaptent le mieux aux changements. »

Charles Darwin

Table des matières

Remerciements

Tout d'abord, je tiens à remercier tout particulièrement Jean Armengaud et Christine Almunia qui ont respectivement dirigé et encadré cette thèse. Christine, je n'ai pas de mots pour t'exprimer à quel point je te suis reconnaissant pour tous les savoirs aussi bien scientifiques qu'humains que tu as su partager avec moi. Etre à ton contact m'a permis d'évoluer chaque jour, de partager et d'apprendre au quotidien. Ton soutien et ton optimisme sans faille jusqu'au bout de cette thèse me permettent aujourd'hui de voler vers de nouveaux horizons en gardant en tête l'ensemble de tes conseils. Jean, tu as su me conseiller, me diriger et tes conseils ont toujours été d'une vraie aide. Grâce à toi j'ai eu l'occasion d'apprendre énormément sur la valorisation du travail de recherche et sur les différentes tâches d'un chercheur. Tu m'as toujours tiré vers le haut et encourager à repousser mes limites. Tu m'as fait confiance depuis le début pour cette thèse et m'a laissé évoluer au sein de ton équipe que tu diriges d'une main de maître et qui a beaucoup de chance de t'avoir.

Je souhaite remercier les membres du jury d'avoir accepté d'évaluer mon travail. Je remercie les Dr. Christine Carapito et Dr. Ana Varela Coelho pour avoir accepté d'être rapporteurs ainsi que les Dr. Jacques Colinge et Dr Michel Hébraud d'être mes examinateurs.

Je remercie le CEA de m'avoir financé ainsi que l'ANR via le projet « PROTEOGAM ».

Je souhaite remercier tout particulièrement les membres de l'équipe protéomique du LI2D grâce à qui j'ai pu effectuer l'ensemble de ces travaux. Tout d'abord, Olivier Pible qui est un bioinformaticien à l'esprit vif qui n'a eu de cesse de m'inspirer, de me guider et de m'apprendre le monde de la protéomique tout au long de cette thèse. A Jean Charles Gaillard, mon cher colloc de bureau des derniers temps avec qui j'ai passé de longs moments à rire, il a su me conseiller et m'apprendre énormément en spectrométrie de masse, et sans qui je n'aurais toujours ni permis, ni A, ni gilet jaune, j'ai au moins les deux derniers grâce à toi. A Guylaine Miotello et Gérard Steinmetz de qui j'ai pu recevoir de nombreux conseils sages et avec qui j'ai eu le plaisir de collaborer quelque fois. A Béatrice Alpha Bazin avec qui ça a été un plaisir de discuter et de se chambrer à de nombreuses reprises. A Lucia Grenga notre jeune maman à qui je souhaite plein de bonheur. A Charlotte Mappa qui a toujours été de bon conseil et à qui je souhaite plein de réussite dans sa nouvelle vie. A Duarte Gouveia qui a été mon prédécesseur au sein de PROTEOGAM et, je l'espère, réussira au plus vite son projet de carrière. A Karim Hayoun, toi qui a été mon camarade de licence et de thèse, je te souhaite la plus belle fin de thèse imaginable. A toi, Karen Culotta alias S.A.V. qui a toujours su m'impressionner autant de par tes qualités humaines que professionnelles. Et enfin à toi, Virginie Jouffret, tu es une personne exceptionnelle dont je souhaite le plus grand bonheur et la plus belle réussite dans ta thèse et tes projets futurs. Merci encore Karim, Karen et Virginie d'avoir su être là dans les soirées difficiles !!

Je souhaite vous remercier vous, les ex-Berti Boulettes, Céline Guigue et Laetitia Pinto, pour avoir toujours été là chaque jour de cette thèse, m'avoir soutenu et aidé à passer outres les moments difficiles. Je te remercie Céline d'avoir enrichi mon vocabulaire chaque jour et de me permettre de ne plus dire j'ai fait une gaffe mais de dire j'ai fait une Céline.

Je souhaite remercier tous les autres membres du laboratoire LI2D qui ont toujours été présents pour moi et prêts à me soutenir. A Fabrice Gallais, mon ancien colloc de bureau alias le CHEF avec qui j'ai pu énormément apprendre sur les ~~BITCOINS~~ virus. A la discrète Noémie Allemand, ma colloc de bureau de ces derniers temps, ne te laisse pas faire par JC dans le dernier mois qu'il te reste et je te souhaite plein de réussite pour le futur !! A Laurent Bellanger de m'avoir accueilli au sein de son laboratoire. A Sylvie Ruat et Fabienne Gas pour les quelques pauses cigarettes partagées à discuter. A Yves Brignon

pour ses nombreuses attentions le matin. A notre super dynamique Joëlle Illiano grâce à qui la moindre demande se transforme en banalité. A Anastasia Dewolf qui aura été là du début à la fin toujours avec sa bonne humeur. Merci à Virginie Nouvel, Yannick Delcuze, Martine Colomp, Anne Desplan, Marie Anne Roncato, Pascale Richard, Stéphanie Debroas et Hélène Batina qui m'ont toujours accueilli avec de grands sourires et grâce à qui chaque venue dans le laboratoire a été un plaisir. Et merci à vous les jeunes Niza Bazaline, Constance Frolich, Basile Leduque et Oumayma El Kaddouri pour toute la bonne ambiance que vous apportez au laboratoire. Et pour finir un remerciement particulier à Charlotte Foissard alias la hyène qui ne mord pas et qui a dédié de précieux moments pour corriger ce manuscrit et a toujours su apporter rigueur et bonne humeur au laboratoire.

Je tiens à remercier les membres de l'IRSTEA, Olivier Geffard, Arnaud Chaumot, Davide Degli-Esposti qui ont su me recevoir et m'aiguiller tout au long de cette thèse afin de toujours conserver en tête la vision écotoxicologique. Arnaud et Olivier se fut un plaisir de travailler avec d'aussi grands spécialistes que vous en écotoxicologie aquatique !!!

Un grand merci à toutes les personnes ayant su motiver mon envie de faire de la recherche. Plus particulièrement à Florin Grigorescu qui a été un de mes principaux mentors dans cette traversée, tu m'as tellement enrichi scientifiquement je ne saurai jamais assez te remercier. Je remercie Alban Mancheron de m'avoir martyrisé en stage, d'avoir été un exemple de pédagogie, de réussite et d'investissement. Je te remercie aussi pour le plus beau des dictons que j'ai appris et qui m'a énormément servi en thèse. « Régle N°1… Règle N°2… !! ». Je remercie aussi Charles Romieu qui a su m'apprendre nombre de méandres de la biologie et qui reste aujourd'hui un exemple de rigueur !! Ta confiance et ton aide m'ont permis de me construire dans un environnement de liberté total guidé par ton esprit vif et je t'en remercie. Enfin Corinne Lautier, bien plus qu'une guide tu as été une de mes mamans de sciences avec Christine !! Toujours à prendre soin de moi sur le plan professionnel comme personnel, je vous remercie toutes les deux et je vous dois énormément. Alban disait il y a deux types de chercheurs, ceux qui savent répondre et ceux qui posent les bonnes questions. Ce qui est sublime c'est d'avoir eu à travers vous tous comme guides, des chercheurs sachant être les deux. Je tiens à remercier l'ensemble de l'équipe pédagogique du master SNS BCD et plus particulièrement Anne Muriel Chiffolleau, Isabelle Mougenot, Jacques Colinge, Annie Chateau et Emmanuel Douzery qui ont su être des exemples de pédagogie et qu'il sera toujours un plaisir de recroiser en congrès !! Enfin pour finir je souhaite remercier Anne-Marie Freyria qui m'a fait découvrir la recherche en biochimie et à qui je dois mon goût de la recherche depuis mes 15 ans.

Je tiens à remercier mes amis qui ont su me soutenir tout le long de cette traversée de thèse. Johan et Yann... Une histoire de CU tout ça... Et quelle histoire !! 8 ans après vous êtes toujours là et j'ai pu tout le long compter sur vous. Merci aux autres amis de la cité U, et notamment à Aurélien Velay à qui je souhaite une fulgurante carrière de médecin !! Merci aussi aux anciens camarades de promotion, Zinédine, Karim, Clément et Axel qui ont chaque jour rendu ces moments d'études agréables. Merci à mes deux vieux amis Nicolas² qui ont su être là depuis toujours et à qui je souhaite plein de réussite. A tous les jeunes ASTICO et plus particulièrement Matthieu, je suis sûr que tu seras un président digne des meilleurs !! A Gonché et Marion mes deux meilleures amies et confidentes qui trouvent toujours les mots pour réconforter mes maux. Et enfin Jules, je t'ai connu à mon inscription à la fac et depuis on ne s'est plus lâché, tu as été là chaque jour, chaque moment difficile et tu seras là le dernier jour de mes études enfin !! Je te souhaite plein de réussite en tant que bioinformaticien et souhaite que tu t'épanouisses à la hauteur de ta générosité. Enfin Elise, tu as su être là tout le long de cette thèse et m'apporter énormément de précieux conseils. Tu as su me comprendre et me soutenir dans les plus difficiles moments et je t'en remercie. Bats-toi comme tu as toujours su le faire et deviens la brillante chercheuse que tu mérites d'être. Bonne chance pour ta thèse !

Enfin je souhaite remercier ma famille. A ma maman et mon papa que j'espère rendre fier, à la hauteur de leurs nombreux sacrifices pour me permettre d'être celui que je suis aujourd'hui. Je suis conscient de la chance que j'ai eu de vous avoir et du miracle que vous avez réalisé à partir de là où vous êtes partis. A ma sœur à qui je souhaite de fulgurantes réussites pour la suite de ses études, j'ai hâte du jour où je ne comprendrai plus rien quand tu parles de math ce qui ne devrait pas tarder à arriver !! Tu es une battante et tu t'en sortiras toujours, je n'en doute pas, alors continues comme ça je suis fière de toi petite grande sœur. A mon oncle Sylvain qui a été un modèle en sciences même s'il vend des séquenceurs pourris !! Tu as su m'orienter vers ton amie Anne Marie alors que j'étais encore jeune et ainsi me permettre de découvrir ma passion et ma voie très tôt. Je ne saurais jamais assez te remercier. Enfin à toi ma chérie, je souhaite que tu réussisses au mieux ta thèse et te fait pleinement confiance pour affronter les difficultés. Je serais toujours là pour te soutenir autant que ce que tu l'as été pour moi. Ton bonheur et sourire seront toujours une priorité pour moi et je ferai tout pour que tu conserves ton éclatante joie de vivre !! Pour finir je souhaite dédier cette thèse à mes arrière grands-parents maternels disparus dans les dernières années. Vous avez toujours été un exemple de force de vie et d'amour !! Merci pour tous ces moment d'été à vos côtés !

Liste des figures:

Liste des tables:

Liste des fichiers supplémentaires (disponible au lien suivant : https://figshare.com/projects/SupplementaryFilesYC/66737) :

Chapitre 1 :

- Tables complémentaires
- 3Frame_Translate : https://github.com/YannickCogne/3Frame_Translate

Chapitre 2 :

- Fichiers de sortie complémentaires, contrôle qualité (FASTQC)
- Fichiers de sortie complémentaires, traduction (Transdecoder)
- Fichiers de sortie complémentaires, annotation (Trinnotate)
- Fichier de sortie complémentaire, évaluation (Transrate)
- Qfiltering : https://github.com/YannickCogne/Qfiltering

Chapitre 3 :

- Figure 1 complémentaires
- BAITS : https://github.com/YannickCogne/BAITS
- Tables complémentaires, sortie de BAITS
- Alignement multiple complémentaires (Cellulase)
- Fichiers de sortie complémentaires (Mascot)

Chapitre 4 :

- Tables complémentaires
- Publication en préparation complémentaire (Data in Brief)
- Tables complémentaires Data in Brief

Liste des abréviations :

- ADN : acide désoxyribonucléique
- EST : marqueurs de séquences exprimées
- ARN : acide ribonucléique
- p-value : probabilité d'obtenir le résultat par hasard
- WGS : séquençage de génome entier
- WXS : séquençage d'exome entier
- Rna-seq : séquençage du transcriptome entier
- LCQ : loci de caractère quantitatif
- CDS : cadre ouvert de lecture
- HPP : projet protéome humain
- SNV/SNP : variants d'un nucléotide
- CNV : variation de nombre de copies de gènes
- ADNc : ADN codant
- ESI : source d'ionisation par électronébuliseur
- HDC : Cellule de collision à haute énergie
- G . : Gammarus
- E. : Echinogammarus

Introduction bibliographique:

I-La protéogénomique

1) Protéogénomique : émergence du concept et définitions

a) Origine de la protéogénomique

Depuis la mise au point du séquençage de l'acide désoxyribonucléique (ADN), comme la méthode Sanger, la génomique a rapidement évolué puis a été automatisée pour permettre la génération rapide de base de données de séquences génomiques pour les organismes vivants (Smith et al. 1986). En parallèle, la protéomique et les interprétations des spectres générés par la spectrométrie de masse sont devenues plus complexes à cause de l'augmentation du nombre de données expérimentales.

Les méthodes permettant de déterminer les peptides, par lecture de spectres de masse, utilisées pour identifier les protéines sont devenues insuffisantes et limitantes en temps de calcul. Les protéomistes et génomiciens jusqu'alors séparés se sont mis à travailler de concert afin de mettre au point de nouvelles méthodologies d'analyse. Les outils d'interprétation des spectres en utilisant la traduction des séquences nucléotidiques ont alors vu le jour. Un des premiers exemples est apparu en 1994, les séquences protéiques disponibles pour l'Homme furent utilisées pour interpréter des spectres (Eng et al. 1994). Puis l'année suivante, c'est directement une traduction en 6 cadres de lecture des marqueurs de séquences exprimées (EST) qui est utilisée (Yates et al. 1995). C'est la naissance du domaine qui sera par la suite nommée la protéogénomique (Jaffe et al. 2004). La protéogénomique s'applique alors aux organismes possédant les bases de données les plus complètes tels que *Drosophilia melanogaster* (Brunner et al. 2007), *Arabidopsis thaliana* (Baerenfaller et al. 2008) ou encore *Deinococcus deserti* (de Groot et al. 2009).

Dans les années 2005, la révolution du séquençage a lieu. Aujourd'hui, les nouvelles technologies permettent pour un coût moindre de séquencer un débit inégalable de données. Le séquençage du génome humain, ayant pris 12 années et près de 3 milliards d'euros au début des années 2000, devient même accessible à un particulier pour quelques centaines d'euros et quelques semaines d'attente de nos jours. Ainsi, la protéogénomique se développe grâce à l'arrivée du séquençage haut débit de l'ARN permettant de générer des bases de données de plus en plus complètes pour l'interprétation des données protéomiques (Wang et al. 2009). Cette révolution permet aujourd'hui d'appliquer en routine la protéogénomique aux espèces sans base de données de référence dites « non-modèles » (Armengaud et al. 2014).

b) Définition de la protéogénomique

La définition de la protéogénomique au sens littéral définie par Jaffe et al. en 2004, consiste en l'amélioration des annotations structurelles des génomes grâce à l'utilisation de la protéogénomique. Cette approche a été utilisée aussi bien sur des génomes bactériens (Venter et al. 2011, Gallien et al. 2009) que des génomes eucaryotes (Castellana et al. 2008, Castellana et al. 2013). Grâce à cette méthode de nouveaux gènes ayant échappés au système d'annotation automatique ont pu être mis en évidence grâce aux données protéomiques. Cependant les travaux protéogénomiques utilisant des données de séquençage de génome pour l'interprétation protéomique se sont ensuite développés

n'ayant plus pour unique objectif l'annotation du génome (Rubiano-Labrador et al. 2014, Wilmes et al. 2008, Christie-Oleza et al. 2013). L'application de l'approche protéogénomique s'est ensuite étendue à l'analyse des données multi-omiques alliant la protéomique, la transcriptomique et la génomique. Par exemple, à partir des données de séquençage des acides nucléiques, les bases de données protéiques sont construites après la traduction *in silico* des séquences nucléotidiques en 6 cadres de lectures (Armengaud et al. 2014). Actuellement la protéogénomique repose sur trois éléments clés :

- Génération des données protéomiques

La spectrométrie de masse permet la mesure de séquences peptidiques. Cette méthodologie permet d'obtenir les empreintes des séquences peptidiques analysées. Ces empreintes, appelées spectres, correspondent à la masse sur charge (m/z) de chacun des peptides analysés (ion parent). Pour une interprétation plus performante de la séquence peptidique, le peptide est ensuite fragmenté par collision et permet d'obtenir la mesure de sous fragments correspondants (ion fils). Les spectres à analyser correspondent donc à l'ensemble des masses sur charges d'ion parent et de la masse sur charge de l'ensemble de ses sous-fragments générés.

Dans ce cas, il est préférable d'utiliser la technique protéomique ascendante (bottom-up) décrite par Yates et al. 1995. Cette technique repose sur l'analyse des peptides par spectrométrie de masse d'un mélange après une protéolyse des protéines par la trypsine. Cette méthodologie permet de tenir compte notamment des modifications post-traductionnelles et d'identifier même partiellement des séquences de protéines potentiellement mal assemblées ou incomplètes. C'est donc la méthodologie de génération de données protéomiques propice pour la protéogénomique se basant sur des bases de données contenant une information parfois partielle et/ou erronée. A l'inverse la méthodologie descendante (top-down) permet l'analyse en direct de la masse sur charge de la protéine sans digestion trypsique. Dans le cas de la protéogénomique, les séquences protéiques dans la base de données étant partielles ou incorrectes dans la majorité des cas, elles ne peuvent permettre cette analyse.

- Génération d'une base de données à partir du génome

La génération de la base de données, à l'origine, proposait une lecture dans les 6 cadres de lecture des séquences nucléotidiques provenant des marqueurs de séquences exprimées (EST) (Yates et al. 1995). Ces 6 cadres de lecture correspondent aux différentes possibilités de traduction du transcrit en protéine soit dans les cadres de lecture du brin sens (1,2,3) soit dans les cadres de brin anti sens (-1,-2,-3). Suite à l'évolution technologique de 2005, les EST sont remplacées par le résultat de séquençage à haut débit (NGS). Aujourd'hui, en fonction de la question biologique posée, une traduction en 3 cadres de lecture dans le cas de séquençage orienté permet de construire la base de données protéiques. Enfin, dans l'objectif de créer une première base de données spécifique d'un organisme non-modèle par exemple, une recherche des séquences codantes peut être effectuée (Armengaud et al. 2014).

- Analyses des données protéomiques

L'attribution des séquences peptidiques aux spectres générés par la spectrométrie de masse se fait en comparant les spectres expérimentaux mesurés aux masses théoriques calculées à partir de la base de données. A chaque peptide candidat pour un même spectre, un score est ensuite calculé. Le meilleur score d'attribution indique la faible probabilité d'obtenir ce dernier par hasard (p-value). Une des

principales particularités de la protéogénomique dans cette étape se trouve dans le contrôle des faux positifs. Notamment lors de l'attribution des séquences peptidiques aux spectres, le nombre de peptides faux positifs augmente avec la taille des bases de données. En effet, la nature de la base de données est dans ce cas d'exploitation particulière car elle contient en majeure partie des séquences protéiques n'existant pas, en plus de posséder un nombre de séquences relativement élevé. De ce fait, le nombre de candidats aléatoires pour l'attribution de chaque spectre augmente ce qui force les scores seuil à être plus élevés pour ne pas augmenter le nombre de faux positifs. Il faut donc un contrôle du taux de faux positif adapté à ce type de requête qui permet de prendre en compte la nature de la base de données tout en n'étant pas trop limitant sur les scores de validation des séquences peptidiques attribuées aux spectres.

La protéogénomique se concentre aujourd'hui principalement sur plusieurs applications (Nesvizhskii 2014) :

- La correction de l'annotation du génome avec notamment des applications étendues à la :
    - Validation des variants d'épissage
    - Validation de variants nucléotidiques
    - Recherche de chimères
- La recherche de marqueurs pour le cancer
- L'analyse des organismes non-modèles
- L'inclusion dans les analyses métaprotéomiques de mélanges d'organismes

Les différentes méthodologies propres aux différentes applications sont représentées sur la Figure 1. Dans la suite de ce manuscrit, l'application de la protéogénomique pour la correction de l'annotation, la recherche de marqueurs pour le cancer et l'analyse des organismes non-modèles sera détaillée.

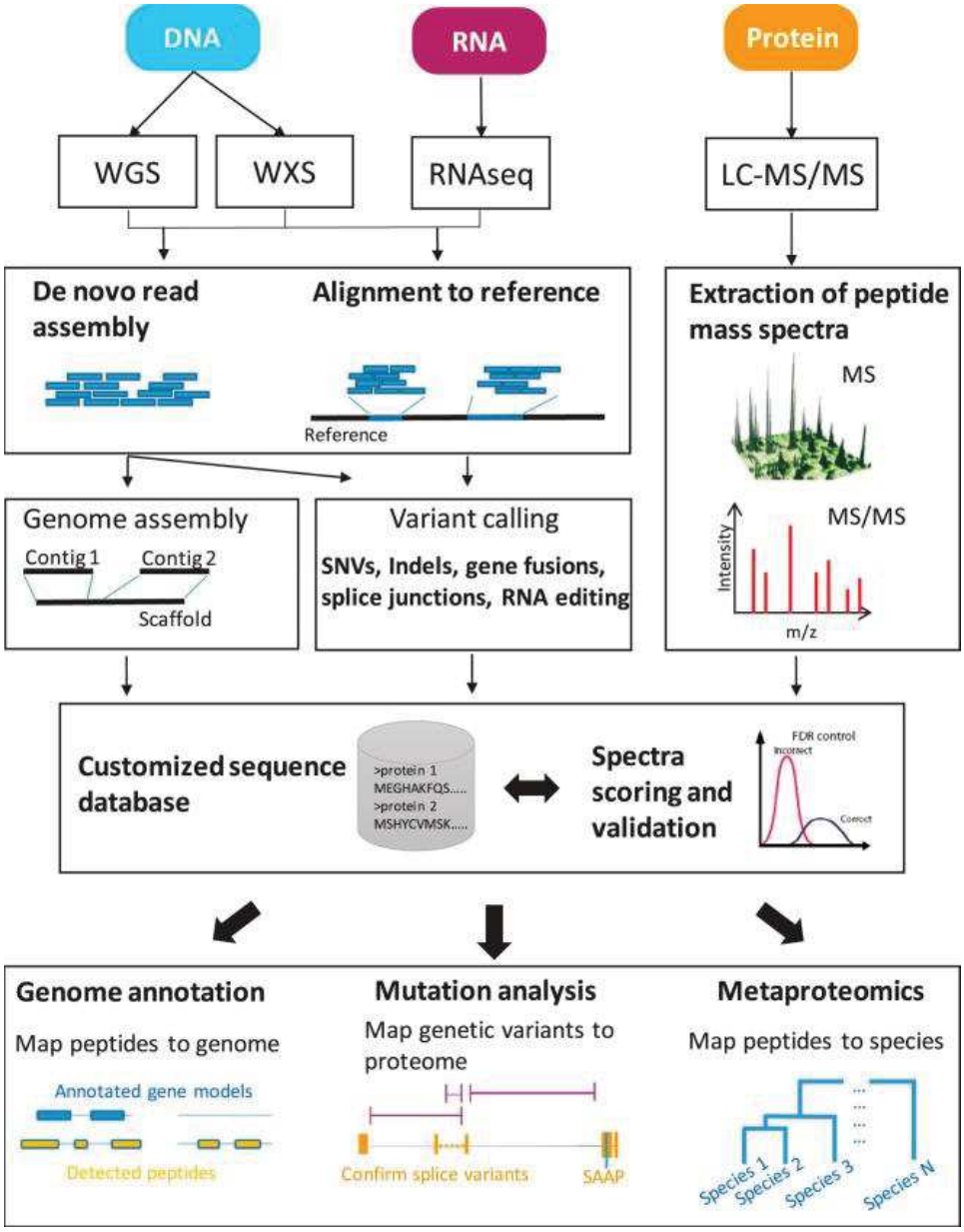


Figure 1 Présentation des différentes applications de la protéogénomique. Les technologies basées sur le séquençage pour séquencer l'ADN (séquençage de génome entier, WGS; séquençage d'exome entier, WXS) et l'ARN (ARN-seq) génèrent des millions de lectures de séquençage courtes assemblées en génomes, exomes ou transcriptomes par assemblage de novo ou par localisation sur une référence. Les modifications de séquences spécifiques à l'échantillon sont déterminées et les séquences de nucléotides sont transformées en bases de données personnalisées, à base de séquences centrées sur les acides aminés. Les spectres de masse des peptides obtenus par analyse LC-MS / MS à partir d'un échantillon correspondant sont ensuite analysés et validés par rapport à la base de données personnalisée, ce qui permet la détection de séquences peptidiques spécifiques à l'échantillon. En fonction de la portée du projet de protéogénomique, ces peptides peuvent ensuite être utilisés pour (1) faciliter l'annotation du génome par la détection de peptides dans des régions du génome non annotées ; (2) identifier des mutations spécifiques de la tumeur traduites dans le protéome ainsi que de nouveaux variants d'épissage de protéines ; et (3) détecter des peptides spécifiques à une espèce dans les communautés microbiennes par exemple. Figure issue de l'article de Ruggles et al. 2017.

2) Correction de l'annotation génomique à l'aide de données protéomiques

a) Quand la protéogénomique aide à la compréhension du génome

L'annotation peut être faite au niveau fonctionnel ou alors au niveau structural pour le génome. Dans le cas de l'annotation fonctionnelle, il s'agit d'attribuer une fonction à une portion génomique. Cette dernière peut être effectuée à l'aide de manipulations génétiques permettant d'identifier notamment des loci de caractère quantitatif (LCQ) (Zeng et al. 1999). Ces loci, une fois précisément identifiés, peuvent être caractérisés, au niveau fonctionnel, en utilisant des techniques biologiques (mutagénèse, synthèse des protéines, etc.) qui attribuent une annotation fonctionnelle à chaque portion génomique. D'une façon globale, ce travail n'est effectué que sur des organismes dits « modèles ». Ces organismes ont soit un intérêt scientifique ou économique direct (l'Homme, la vigne, le blé, virus ou bactérie pathogène, etc.) soit sont des organismes plus facilement exploitables en laboratoire qui servent de modèles d'étude (la souris, le poisson zèbre, la levure, etc.).

Cependant, il existe une relation séquence/fonction qui permet aujourd'hui d'attribuer les fonctions d'une région génomique à une autre région similaire en terme de séquence. On utilise alors les propriétés des séquences proches en enchaînement nucléotidique (homologues) pour définir des séquences possédant la même fonction dans un autre organisme (orthologues). Toutefois, des séquences similaires peuvent présenter une histoire évolutive différente et à terme une fonction distincte (paralogues). Dans ce cas, notamment, il existe des erreurs d'annotations possibles. De plus, il existe des gènes « orphelins » n'étant pas définis dans d'autres organismes que celui étudié. Dans le cas des organismes non-modèles ces gènes orphelins sont complexes à annoter. Il est de plus évident que plus ces organismes non-modèles s'éloignent des organismes modèles plus le nombre de gènes orphelins augmente (Trapp et al. 2015). Enfin, les technologies actuelles nous fournissent des fragments de séquençage qui doivent être assemblés pour recréer les génomes complets. Ces méthodes, utilisant notamment les chevauchements des séquences, peuvent conduire à des erreurs et à la création de combinaisons de séquences inexistantes (chimères).

Un autre niveau possible d'annotation est l'annotation structurale du génome. En effet, en plus d'identifier la fonction d'une portion génomique, on peut détecter les portions correspondant à des régions codantes ou non. Par exemple pour les eucaryotes, une portion génomique peut être définie comme étant exonique, intronique, intergénique ou encore régulatrice. Dans un premier temps, les résultats d'analyse en spectrométrie de masse permettent de vérifier ces annotations structurales sur des bactéries, archaes et des organismes eucaryotes (Venter et al. 2011, Rison et al. 2007, Gupta et al. 2007, Armengaud et al. 2009, de Groot et al. 2009). Au cours du temps, les premières propositions d'utilisation systématique de la protéogénomique pour l'amélioration de la qualité de l'annotation apparaissent (Ansong et al. 2008, Castellana & Bafna 2010). Il est alors aussi proposé d'utiliser la protéomique pour valider la détection de régions codantes (cadre ouvert de lecture, CDS) au sein des nouveaux génomes. En effet, la détection de peptides dans l'échantillon correspondant à ces séquences est une preuve de leur expression en transcrits puis de leur traduction. Des études démontrent alors la possibilité de codon initiateur alternatif de la traduction au sein des procaryotes (Nielsen et al. 2005, Baudet et al. 2010). Ces résultats sont complémentaires à une autre étude qui

démontre que le codon d'arrêt de la traduction (TGA) peut aussi induire la synthèse d'un acide aminé, la sélénocysteine (Low et al. 1996). La solution adoptée est alors la recherche des protéines en utilisant la traduction en 6 cadres de lectures permettant de ne plus se limiter aux régions codantes définies entre un codon initiateur et terminateur putatif. Enfin, des travaux mettent en évidence la possibilité de corriger les erreurs d'assemblages et la détection des chimères en utilisant les attributions protéomiques. En effet, l'absence de certains peptides ou les couvertures non-homogènes d'attribution pour une protéine peuvent indiquer une erreur d'assemblage ou la présence de chimère. Grâce aux progrès effectués et le développement des outils dédiés, la protéogénomique s'intègre dans les études d'annotations actuelles afin d'en améliorer la qualité. L'ensemble des exemples de correction de l'annotation est résumé dans la Figure 2.

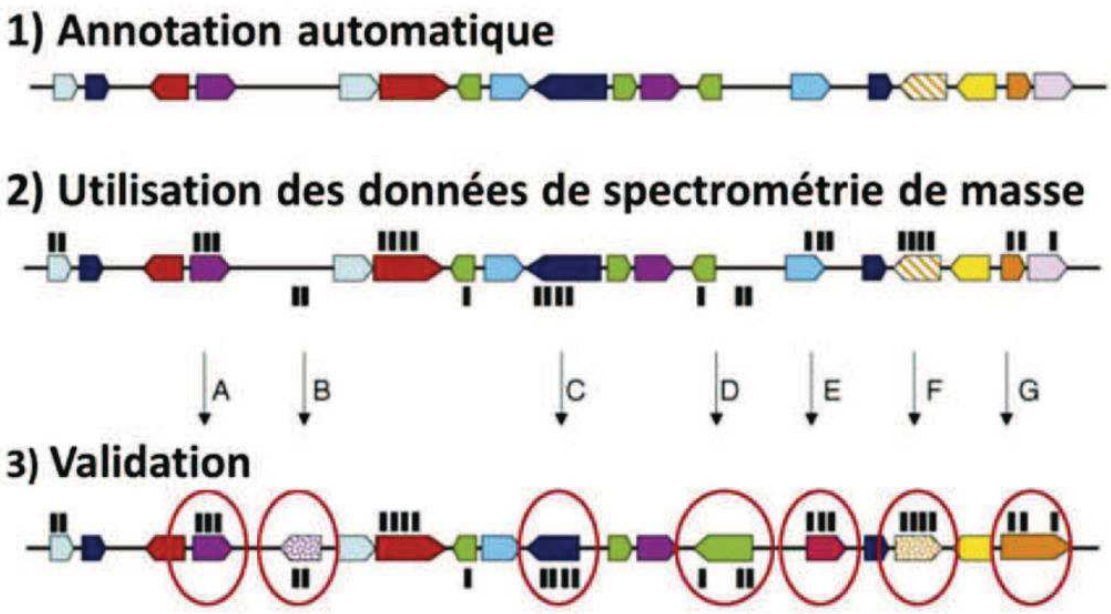


Figure 2 : Schéma des 3 étapes de correction de l'annotation par protéogénomique. 1) Annotation originelle automatique. 2) Localisation des peptides sur le génome. 3) Correction proposée suite à la localisation des peptides. Les peptides sont indiqués par les rectangles noirs. Les peptides au-dessus de la séquence sont dans le sens positif de lecture (cadre 1,2 ou 3) et ceux en dessous dans le sens négatif de lecture (cadre -1,-2 ou -3). Dans le cas A) la protéogénomique valide l'annotation. Dans le cas B) un nouveau gène est détecté. Dans le cas C) ou D) la position du codon start est décalée en aval ou en amont respectivement de l'annotation originelle. Le cas E) un décalage du cadre de lecture est détecté. Dans le cas F) les peptides indiquent que le sens de lecture est inversé. Enfin dans le cas G) les peptides mettent en évidence la considération d'un codon stop non exprimé ou d'un décalage de cadre de lecture dans l'annotation originelle. Figure issue de Armengaud 2009.

b) Apport actuel de la protéogénomique dans l'annotation

Dans le cadre d'organismes modèles tels que l'Homme ou *Arabidopsis thaliana* la protéogénomique permet de valider une majeure partie des gènes. Par exemple pour l'Homme, Kim et al. 2014 annoncent près de 84 % du protéome théorique total comme étant identifiés. Il existe cependant de nombreuses controverses au sujet de cette publication. En effet, le projet protéome humain (HPP) a maintenant mis en place des conditions pour accepter la présence d'une protéine plus stringente. Par

exemple, il est désormais nécessaire de présenter deux peptides différents par protéine avec un contrôle du taux de faux positifs à 1%, au niveau spectres, peptides et protéines avant d'en affirmer la présence. En ce qui concerne *Arabidopsis thaliana* il s'agirait de près de 40% des gènes annotés qui seront validés par protéomique et près de 13 % de l'annotation de gènes supplémentaires manquants qui seraient mis en évidence par la protéogénomique (Castellana et al. 2014).

De nos jours, la protéogénomique est même utilisée pour permettre une meilleure annotation d'organismes non-modèles tel que le panda géant (Chen et al. 2015). Dans le cadre de cette étude, l'annotation est définie en utilisant aussi bien les méthodes automatiques reposant sur la transcriptomique ainsi que la méthodologie protéogénomique. En effet, la sensibilité accrue du séquençage d'ARN permet de mettre en évidence les régions du génome pouvant être exprimées (exons). Des outils permettent ensuite de générer une annotation théorique en partant de la localisation des transcrits sur le génome (*ab initio*) suivie de l'utilisation des méthodes par similarité de séquences pour l'attribution des fonctions.

Dans ce type d'étude, la protéogénomique assure la validation biologique pour les nombreuses annotations théoriques produites par ces logiciels, par exemple Augustus (Stanke et al. 2006), apparaît comme étant la méthodologie la plus rapide et la moins coûteuse pour cette étape. Dans le cadre de cette publication par exemple, il s'agit de près de 13 % de nouveaux gènes détectés comme codant issus du panda géant qui ont pu être validés directement par la protéomique. Même s'il ne s'agit pas encore de la majeure partie de l'annotation, cela représente près de 1400 gènes dont la validation aurait nécessité un coût et un temps largement supérieur pour une évaluation ciblée de leur expression. Dans ce cas, la protéogénomique sans *a priori* peut s'appliquer directement.

3) Onco-protéogénomique

a) Quand la protéogénomique aide à comprendre les tumeurs

Définie pour la première fois par Helmy et al. 2010, l'onco-protéogénomique repose sur l'enrichissement des bases de données de séquences protéiques humaines standards avec des séquences peptidiques spécifiques des cellules cancéreuses. En effet, les cellules cancéreuses possèdent leur propre génome dérivé du génome de l'hôte. Ce génome possède de nombreuses mutations qui permettent de les différentier des cellules saines de l'hôte. Ainsi les données de séquençage permettent de mettre en évidence les mutations spécifiques, les variants d'épissage anormaux, les variants d'un nucléotide non-synonyme (SNV), les additions ou suppressions de nucléotides (indel), les fusions de gènes ou le nombre aberrant de copies de gènes (CNV).

L'ajout des résultats provenant du séquençage aux bases de données générique de l'humain permettent, dans le cas des cellules cancéreuses, de mettre en évidence de nouveaux peptides comme décrit dans la Figure 3. En effet, la détection de peptides contenant des SNV permet de confirmer leur présence et de ne pas confondre avec une éventuelle erreur de séquençage. De plus, les peptides à la jonction de deux exons peuvent confirmer la présence d'un variant d'épissage particulier ou encore la présence de fusion de gènes. Enfin avec des techniques analogues à celle de la correction de l'annotation, des décalages de cadres de lecture peuvent être mis en évidence afin de détecter les insertions et délétions.

De nos jours, une information supplémentaire peut être ajoutée en utilisant les bases de données de variants spécifiques répertoriés comme par exemple la base de données dbSNP, largement alimentée par le projet 1000 génomes humains (1KGP, Sudmant et al. 2015). Cette base de données contient les

variants nucléotidiques détectés sur plus de 2000 humains en considérant plus de 3 origines ethniques différentes. Enfin la protéomique permet aussi de mieux comprendre les mécanismes d'action de certains médicaments sur les cellules cancéreuses afin dans le futur d'adapter au mieux les traitements (Dalzon et al. 2019).

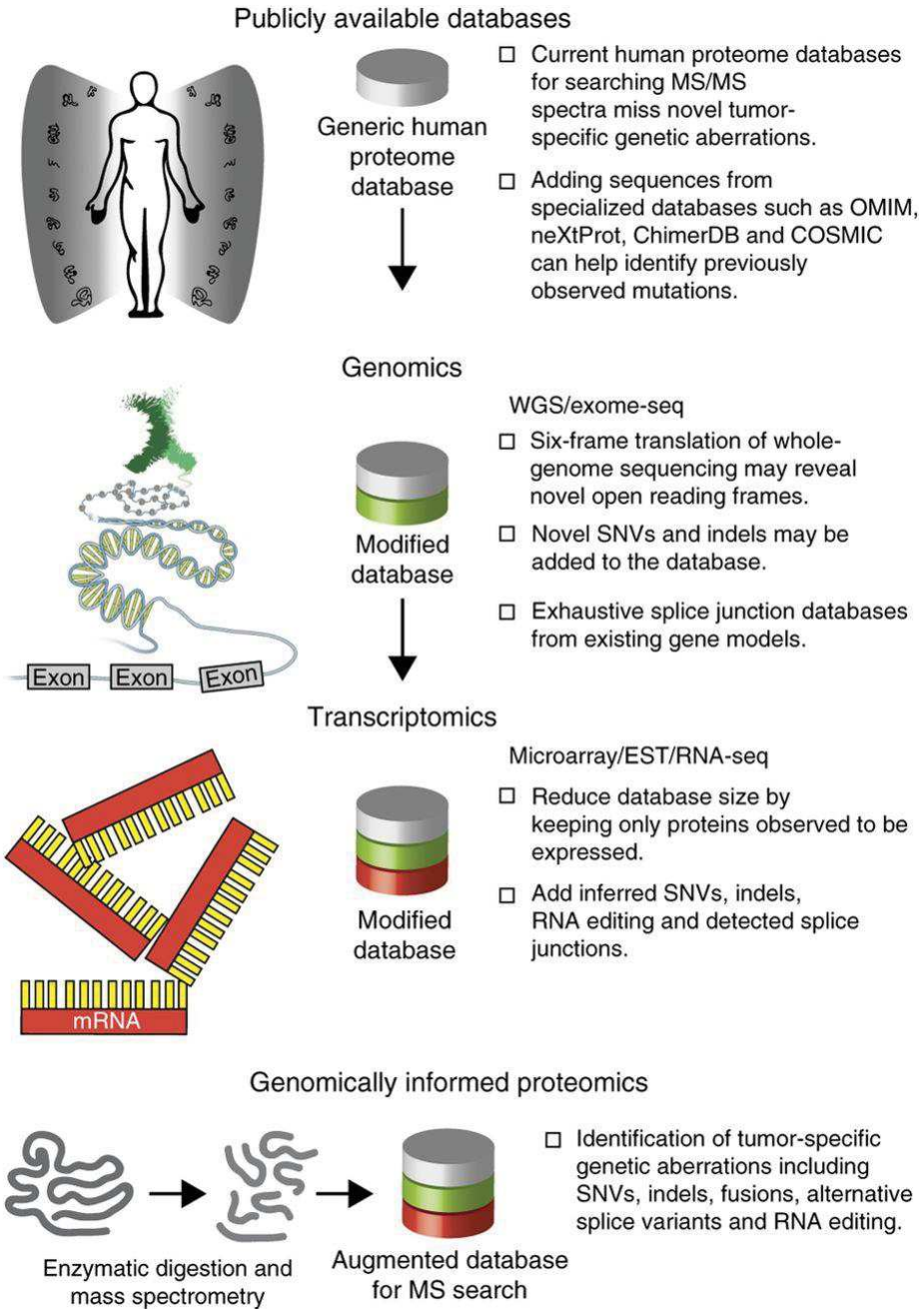


Figure 3 : Méthodologie de génération de base de données protéogénomiques intégrant les modifications détectées au niveau génomique et transcriptomique. Détection des mutations génétiques spécifiques à la tumeur en protéomique du cancer. Les mutations génétiques non caractérisées auparavant ne peuvent pas être identifiées en recherchant les spectres protéomiques MS / MS dans une base de données de protéines générale. Pour détecter ces anomalies, des bases de données modifiées doivent être générées en appliquant d'autres «technologies omiques» à la tumeur en question. Les protéines prédites issues du séquençage du génome, du séquençage de l'exome ou du profil de transcription peuvent contribuer à la constitution de ces bases de données enrichies et à la prise en charge de la «protéomique informée par la génomique». EST, séquence de transcrits exprimées ; WGS, séquençage du génome entier. Figure issue de Alfaro et al. 2014.

b) *Apport actuel de l'onco-protéogénomique*

Dans le premier cas d'utilisation de Helmy et al. 2010, l'analyse de 15 expériences de protéomique shotgun de cellules HeLa S3 en utilisant leur transcriptome spécifique a permis de mettre en évidence 25 peptides spécifiques de cellules cancéreuses. Quelques années plus tard, une étude sur le même type cellulaire a permis de mettre en évidence 450 nouveaux peptides (Evans et al. 2012).

La protéogénomique a été appliquée à ce jour pour l'étude de divers cancers tels que le cancer du poumon (Sun et al 2013), du foie (Mo et al. 2008), du sein (Menon & Omenn 2010) ou encore colorectal (Wang et al. 2012, Halvey et al. 2014). Ces différentes études permettent de mettre en évidence plusieurs phénomènes biologiques. L'un des premiers objectifs est de mettre en évidence de nouveaux produits de gènes impliqués dans les différents cancers. On distingue les variants d'épissage spécifiques de cellules cancéreuses, les fusions de gènes et les variants nucléotidiques spécifiques aux cancers qui peuvent être mis en évidence par la présence de peptides spécifiques. En ce qui concerne la mise en évidence de nouveaux gènes impliqués dans le cancer, l'étude de la co-expression entre les gènes déjà définis comme étant impliqués et les nouvelles cibles apporte une information supplémentaire (Johansson et al. 2019). Ces variants et cibles identifiés ont pu permettre la mise en évidence de nombreux biomarqueurs potentiels permettant de prédire par exemple la résistance à un traitement ou le risque de présence de cancer chez un patient (Fu et al. 2017).

4) Amélioration de l'interprétation des données protéomiques grâce aux séquençages

a) *Interprétation des données protéomiques d'organismes non modèle*

Dans le cas de l'absence de base de données, par exemple pour un organisme non modèle, les protéines ne peuvent pas être identifiées autrement que par la méthodologie *de novo*. Cette méthodologie ne permet cependant pas une attribution optimale des spectres bruités, provenant des peptides co-élués, en spectrométrie de masse. De plus, aucune information sur la séquence complète de la protéine ne peut être fournie et l'annotation des peptides pour déterminer leur fonction n'est pas suffisante. Dans un premier temps, les organismes les plus proches ont été utilisés afin de permettre l'attribution. Cependant, cette solution reste limitante. Pour pallier à cela, la solution proposée est l'utilisation combinée des données d'un organisme de référence complétées par les EST spécifiques de l'espèce disponible permettant une amélioration de l'interprétation des spectres (Grimplet et al. 2005).

Cependant l'utilisation d'une base de données composée principalement de séquences provenant de l'organisme modèle le plus proche ne permet qu'une interprétation partielle des spectres (Trapp et al. 2015). L'interprétation peut alors être améliorée en utilisant des génomes même incomplets (Nanduri et al. 2005). Les progrès en séquençage permettent de séquencer les transcrits sous forme de fragments appelés lectures (Wang et al. 2009). Grâce à cette nouvelle technologie de nombreuses espèces non modèles sont alors séquencées à l'instar du projet Tara Oceans ayant pour ambition de séquencer les organismes des fonds marins (Carradec et al. 2018). Contrairement aux EST, ces séquençages permettent d'avoir une vision large du transcriptome. Une traduction de ces transcrits

assemblés permet de générer une base de données de séquences de protéines théoriques pour l'attribution des spectres permettant d'améliorer significativement l'attribution. Le premier exemple de la protéogénomique d'organismes non modèles est alors proposé par de Groot et al. 2009 sur *D. deserti.* Quelques années plus tard Alexeev et al. 2012 effectuent une étude à la fois génomique et protéomique de *Spiroplasma melliferum*. Ensuite, c'est Rubiano-Labrador et al. 2014 qui l'appliquent en utilisant le génome incomplet de *Tislia consotensis*.

b) Utilisation sur des eucaryotes pour mieux comprendre l'environnement

La protéomique informée par transcriptomique peut permettre de comprendre et d'analyser des modèles biologiques d'intérêt ne possédant jusqu'alors aucun séquençage. Il s'agit du contexte dans lequel se déroule cette thèse. En effet, nous travaillons principalement sur le Gammare, qui est un organisme sentinelle permettant la détection de polluants dans l'eau. Cependant, il y a de celà encore 5 ans, aucune donnée de séquençage n'était disponible pour cette espèce. Les études écotoxicologiques étaient alors limitées aux observations physiologiques ou traits de vie sans avoir accès aux mécanismes moléculaires. Afin d'explorer plus en profondeur les mécanismes moléculaires impliqués au sein de *Gammarus fossarum,* Trapp et al. 2014 ont effectué une première étude protéogénomique permettant d'identifier près de 1873 protéines dont 218 spécifiques de l'espèce. Cette étude a permis notamment d'améliorer la compréhension du système reproductif du Gammare. Suite à ces premiers travaux, l'analyse différentielle des protéomes d'organismes exposés ou non à des polluants a permis de sélectionner des biomarqueurs pertinents, témoins de pollution. Cette approche par protéomique différentielle a permis d'identifier un grand nombre de biomarqueurs de pollution à partir d'autres organismes sentinelles non modèles comme la Dreissène par exemple (Leprêtre et al. 2019).

5) Outils informatiques pour la protéogénomique

a) Bases de données de variants

De nombreuses bases de données de variants ont été construites à partir des données extraites de publications récentes. Une liste de celles-ci est proposée par Yashwant et al. 2015. Parmi les principales pouvant être explorées par protéogénomique, il existe :

- OMIM : Base de données sur les mutations ponctuelles induisant des maladies héréditaires chez l'Homme.
- dbSNP : Base de données sur les variants d'un seul nucléotide (SNP) détectés chez l'homme dans le cadre du projet 1000 génomes.
- COSMIC : Base de données de variants somatiques spécifiques du cancer
- Human Protein Mutant Database : Traduction des gènes de protéines contenant des codons non synonymes provenant de dbSNP ou OMIM.

b) *Outils de génération de bases de données*

Une des étapes clefs de la protéogénomique se trouve dans la construction d'une base de données. CustomProDB développé par Wang & Zhang 2013 est un package R permettant la construction de base de données protéogénomiques. Ce logiciel prend en entrée les résultats d'un alignement et permet de générer des bases de données contenant soit les protéines dont les ARNm ont été détectés, soit les protéines dont les séquences nucléotidiques correspondantes sont modifiées ou encore les peptides signatures d'une nouvelle jonction d'épissage. L'outil PROTEOFORMER permet à partir de profils ribosomiques de détecter les régions codantes des gènes et de construire une base de données spécifique en utilisant ces informations (Crappé et al. 2014). Cet outil est aussi capable d'intégrer les informations des variants nucléotidiques. D'autres outils supplémentaires sont indiqués dans la Figure 4.

c) *Recherche automatique protéogénomique*

Afin de rendre accessible la protéogénomique, il est nécessaire d'obtenir des outils permettant une analyse complète partant de la création de la base de données jusqu'à l'interprétation des spectres. La suite d'outils PGTools permet notamment de réaliser la création de base de données à façon (Nagaraj et al. 2015). Cette suite d'outils permet ensuite l'attribution, l'analyse et la visualisation des données protéogénomiques. Peppy (Risk et al. 2013) et ENOSI (Woo et al. 2015) sont aussi deux outils disponibles permettant la construction de la base de données et son analyse, en intégrant le calcul du taux de faux positifs. Cependant, il est à noter que la totalité de ces outils fonctionne uniquement avec des génomes de référence disponibles et ne peut donc pas être appliqué directement à la protéogénomique d'organisme non modèle.

d) *Correction de l'annotation*

Dans le cadre de la correction de l'annotation par protéogénomique des outils permettent de localiser les peptides identifiés sur le génome puis de modifier cette annotation. Les outils ProteoAnnotator (Ghali et al. 2014), PGP (Tovchigrechko et al. 2014), SpliceVista (Zhu et al. 2014), Genome Peptide Finder (Specht et al. 2011), le workflow Peptimapper (Guillot et al. 2019) et Proteogenomic Mapping Tool (Sanders et al. 2011) permettent d'effectuer ce traitement en localisant les peptides sur le génome une fois ces derniers identifiés.

e) *Format protéogénomique*

Une des limitations pour la correction de l'annotation se trouve notamment dans la différence des formats de localisation entre protéomique et transcriptomique. Pour pouvoir rendre ce type d'analyse interopérable malgré l'intégration de données omiques hétérogènes, les formats proSAM et proBAM ont donc été développés (Wang et al. 2016, Menschaert et al. 2018). Ces formats ont pour objectif l'uniformisation des formats de sorties habituels obtenus par les logiciels d'analyse des données transcriptomiques afin de rendre les sorties protéogénomiques plus simples à intégrer au sein d'études telle que l'annotation. Ces formats sont de plus compatibles avec les outils de visualisation ce qui permet d'améliorer une interprétation manuelle qui, jusqu'alors, restait fastidieuse. Ce format permet une lecture par l'utilisateur dans la version proSam sous un format tabulé. Ce format comprend notamment le nom de la séquence contenant le peptide identifié et sa position au sein de la séquence. Des indicateurs supplémentaires sont ajoutés à la suite permettant notamment l'identification de la portion variante d'un peptide par rapport à la référence ou encore la possibilité d'identifications

multiples. Dérivé du format SAM, le format BAM est un format compressé permettant une analyse plus rapide par les logiciels, il en va de même pour le format proBAM.

De plus, afin de considérer les différentes protéoformes, il est nécessaire de créer un nouveau format permettant d'intégrer les modifications post-traductionnelles aux séquences d'acides aminées. A cet effet un nouveau format, le format ProForma a été proposé par LeDuc et al. 2018. Ce format permet de représenter l'ensemble des modifications en les intégrant avec une nomenclature contenant notamment la masse de la modification au sein des séquences pour simplifier l'exploitation.

f) Outils de visualisation de données

La correction automatique de l'annotation peut être effectuée automatiquement avec les outils précédents mais pour plus de précision une vérification manuelle peut être effectuée. Etant donnée la masse de données générées par la méthodologie protéogénomique, des outils de visualisation doivent être utilisés. Parmi ceux-ci, le logiciel VESPA (Peterson et al. 2012) permet de localiser les données transcriptomique et protéomique ainsi que l'annotation sur le génome afin de visualiser les concordances ou discordances entre les données. D'autres logiciels permettent la localisation de données multi-omiques sur le génome tels que Circos (Krzywinski et al. 2009) et IGV (Robinson et al. 2012). Ce dernier peut même être utilisé grâce à sa lecture des formats BAM et SAM et à la sortie du format proBAM/proSAM. Enfin, afin de comprendre les interactions entre protéines ou gènes, la visualisation des interactions connues peut être utile. En ce qui concerne les organismes modèles, des outils tels que GeneMania (Warde-Farley et al. 2010) ou encore STITCH (Szklarczyk et al. 2016) peuvent être intégrés en tant que plugins au logiciel Cytoscape (Shannon et al. 2003) spécialisé dans l'étude des réseaux d'interaction protéique et génique.

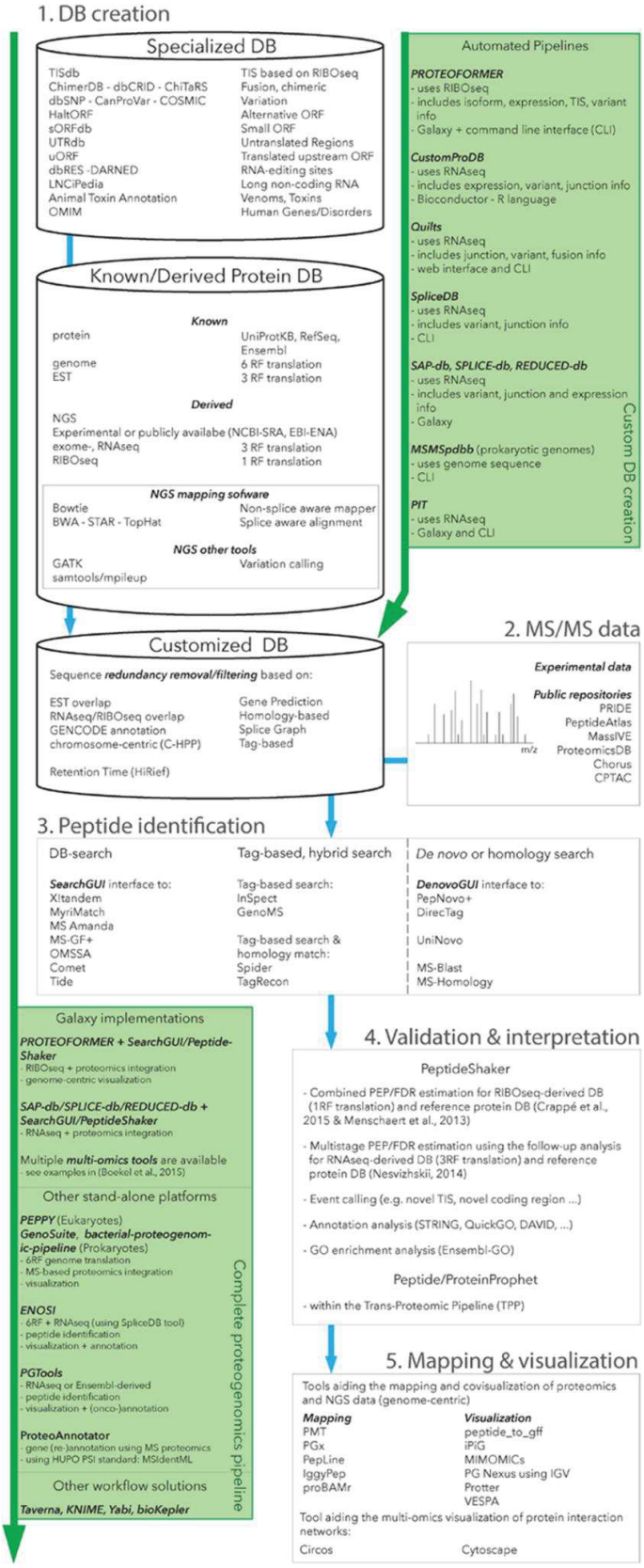


Figure 4 : Représentation des différents outils pour la protéogénomique. 5 types d'outils différent sont ici présenté, 1) les outils de génération de base de données, 2) les base de données de stockage de données protéomique, 3) les outils d'identification de peptides, 4) les outils de validation et d'interprétation des résultat et enfin 5) les outils de localisation et visualisation des résultats protéogénomiques. Figure issue de Menshaert & Fenyö 2017

II-Omiques pour la protéogénomique d'organismes non-modèles

1) Technologie de séquençage d'acides nucléiques

a) Le séquençage d'acides nucléiques nouvelle génération

Les nouvelles technologies de séquençage permettent actuellement de produire une grande quantité d'information pour un coût réduit. La présentation de la technologie se concentrera sur la technologie développée par Bentley et al. 2008 pour les séquenceurs Illumina et utilisée dans cette thèse pour l'ensemble des séquençages d'ARNm. Le séquenceur utilisé est l'Hiseq3000 et ses valeurs de débit et coût seront utilisées comme référence pour les explications qui vont suivre.

L'ARNm et l'ADN sont séquencés avec la même technique. La différence entre les 2 types de séquençage réside dans la rétrotranscription de l'ARN en ADN codant (ADNc). Les ADN obtenus sont ensuite fragmentés afin d'obtenir de petits fragments de 100 à 200 paires de bases environ. Des adaptateurs sont ensuite ajoutés à l'extrémité des fragments permettant leur fixation à une puce. Le séquençage ensuite effectué est un séquençage par synthèse. Une polymérase vient ajouter tour à tour un nucléotide à une séquence d'adaptateur complémentaire à celle utilisée pour la fixation des fragments. Cette polymérase agit en ajoutant les nucléotides par complémentarité de séquence. Les nucléotides possèdent un terminateur de chaîne réversible et un groupement fluorescent, le détecteur mesure la fluorescence et le terminateur est ensuite retiré puis le prochain nucléotide est inséré et ce jusqu'au séquençage complet du fragment.

Cette méthodologie permet aujourd'hui de générer en 1 à 3 jours près de 2.5 milliards de lectures de fragments qui peuvent représenter jusqu'à 750 Giga paires de base pour un prix d'environ 22 $ par Giga paires de base (d'après les données constructeurs) (Goodwin et al. 2016). Pour donner un ordre d'idée, sur un génome humain, cette technologie permettrait en 3 jours d'obtenir suffisamment de lecture pour couvrir en moyenne 250 fois le génome humain pour un prix d'environ 17000 $. Ce prix représente un coût environ 150 000 fois inférieur à celui du premier séquençage humain.

Cette technologie a été choisie afin de produire les données nécessaires pour répondre à la question posée dans le cadre de la thèse, non seulement pour le débit important de données, mais aussi pour son faible taux d'erreur. En effet contrairement à d'autres technologies de séquençage proposant des prix encore inférieurs, le séquençage par l'Hiseq3000 assure un taux d'erreur de 0.1 % annoncé par le constructeur. C'est non seulement ce taux d'erreur faible, mais aussi la nature des erreurs de séquençage qui ont suscité notre intérêt pour cette technologie. En effet, les principales erreurs de séquençage observées ne sont pas des décalages du le cadre de lecture lors de la traduction en protéines mais sont principalement des substitutions.

b) Vers une nouvelle génération de séquençage

Une nouvelle méthodologie de séquençage est désormais disponible. Deux technologies concurrentes sont actuellement disponibles Pacific Biosciences (PacBIO) et Oxford nanopore. Ces deux technologies, considérées aujourd'hui comme la 3ème génération de séquençage, permettent de générer des

fragments de plusieurs dizaines de kilos paires de bases voir même de plusieurs centaines dans le cas de la technologie nanopore. De plus, ce qui est mis en avant par la société nanopore, c'est la portabilité de leur produit se présentant sous forme de clef USB permettant des analyses directement sur le terrain réduisant ainsi le risque d'altération de l'échantillon par exemple.

La principale limitation de cette technologie, aujourd'hui, réside dans le fort taux d'erreur encore présent lors du séquençage qui reste limitant pour la protéogénomique. En effet, les erreurs actuelles encore estimées à plus de 5 % pour les deux technologies se traduisent globalement par des insertions ou suppressions de séquences entrainant le décalage du cadre de lecture après traduction en protéine. Des approches hybrides peuvent être utilisées, mais le coût de l'utilisation conjointe des lectures longues et courtes reste limitante. Ces erreurs dans le cas de la technologies nanopore sont cependant en grande partie dûes à l'analyse du signal et à l'attribution de la séquence après la lecture du signal. Ces erreurs pourraient être corrigées par des algorithmes d'interprétation du signal, basés sur des méthodes par apprentissage. Ces derniers, en constante évolution pourraient permettre, à terme, d'exploiter ce type de technologies directement pour la protéogénomique.

2) Concepts et outils informatiques associés au séquençage (Illumina)

a) Méthodes et outils de prétraitement des données

Les données brutes de séquençage Illumina possèdent des biais de séquençage connus qui peuvent être corrigés avant tout traitement futur des données. Un outil, FASTQC, a été développé afin de contrôler ces différents biais (Andrews, 2010). Ce logiciel fournit les statistiques de contrôle de qualité des lectures et vérifie la présence de ces biais. Aujourd'hui dans la majorité des publications basées sur ce type de données, certaines étapes sont faites en routine.

- Baisse de la qualité de séquençage :

La qualité des dernières bases de chaque fragment de lecture nucléotidique (~10 % finale) est généralement moins élevée que les premières. Ce biais provient de la spécificité probablement moins forte pour l'élongation et l'intégration d'une base nucléotidique lorsque le fragment séquencé devient trop long. Pour s'affranchir du risque d'erreur d'interprétation des nucléotides, la solution employée est d'éliminer la fin des lectures. Ce traitement peut être effectué par des outils tel que Trimmomatic (Bolger et al. 2014).

- Présence d'adaptateur dans les lectures :

Lorsque le fragment est plus court que la taille prévue lors de la préparation, la fin de la lecture correspond à une partie de l'adaptateur. Les séquences d'adaptateur étant connues, il est possible de vérifier leur présence en fin de séquence et de les retirer. L'outil de référence pour ce type de traitement est CUTADAPT.

- Présence de séquences contaminantes :

La préparation d'échantillons d'ARNm est sensible à toute contamination par des nucléotides. L'extraction des ARN à partir des organismes et la préparation de la librairie sont des étapes présentant un risque de contamination d'origine humaine. De plus, dans le cas de l'étude des ARNm, la déplétion des ARN ribosomaux est nécessaire avant le séquençage. Cette déplétion peut s'avérer incomplète et induire la présence de lecture correspondant aux ARN ribosomaux. Dans les deux cas, l'utilisation d'une base de données incluant toutes les séquences nucléotidiques prédites comme potentiellement

contaminantes (Humain) ainsi que l'ARNr de l'espèce étudiée, peut être utilisée pour identifier et quantifier la contamination qui sera par la suite retirée les lectures correspondantes. Cette étape peut être effectuée en utilisant l'outil SortMeRNA (Kopylova et al. 2012).

- Erreur de séquençage :

Il arrive que le séquenceur produise des lectures incorrectes pour un nucléotide donné. Dans le cadre du séquenceur utilisé, cela a 0.1 % de chance de se produire (selon le constructeur). Pour corriger toutes les erreurs engendrées par le séquençage, la stratégie se base sur la grande couverture en lecture d'une position donnée. En effet, l'erreur est supprimée grâce à la superposition des lectures et en fonction des versions les plus fréquentes qui sont alors conservées. Rcorrector (Song & Florea 2015) permet notamment de corriger les lectures Illumina.

- Erreur de séquençage sur les premières bases nucléotidiques :

Ce biais identifié par Kasper et al. 2010 interviendrait notamment lors de l'utilisation d'hexamères aléatoires pour la rétro-transcription. Ce processus peut générer des mutations artificiellement sur les nucléotides en position 1 à 13 des lectures. Dans le cadre d'étude reposant sur la recherche de variants nucléotidiques, ces nucléotides doivent être supprimés afin de réduire le risque de fausses interprétations.

- Surreprésentation de séquences :

Dans le cas du RNA-seq, certains transcrits majoritaires peuvent représenter une proportion non-négligeable des lectures. Une étude de la fréquence des séquences des lectures peut permettre de l'identifier et de supprimer les lectures considérées comme doublon pour faciliter l'analyse par la suite.

En conclusion, ces différents traitements impliquent en général une perte importante de données. Il est donc essentiel de vérifier la proportion de données brutes concernée par les différents biais avant de les appliquer. De plus, certains traitements peuvent ne pas être compatibles avec une question biologique donnée. Par exemple, l'assemblage peut être largement impacté par l'utilisation d'un paramétrage trop strict lors du retrait des bases de mauvaise qualité ou encore des adaptateurs (Macmanes 2014).

b) Exploitation des données de séquençage pour la protéogénomique

Dans le cadre de la protéogénomique, plusieurs voies d'exploitation des résultats de séquençage sont possibles. Le but de l'utilisation de ces données reste cependant toujours le même, il s'agit d'alimenter une base de données afin de détecter de nouvelles séquences présentes dans l'échantillon.

- Détection des variations de séquence :

Afin d'effectuer la détection de variation, les lectures doivent être localisées sur un génome de référence. Pour se faire, des outils tels que BWA (Li & Durbin 2009) ou BOWTIE (Langmead et al. 2009) ont été développés. Ils permettent de localiser les lectures sur le génome malgré les erreurs de séquençage ou la variation possible des séquences des lectures sur le génome. Ces algorithmes sont principalement basés sur l'utilisation de sous-fragments des lectures. La tolérance, préalablement définie, du nombre de sous-fragments s'alignant parfaitement avec le génome permet d'accepter ou non la localisation sur une position génomique donnée. Dans le cadre de l'utilisation de données issues

de transcriptome, le paramétrage de ces méthodologies permet de considérer les introns présents sur le génome et absents des transcrits implémentés par exemple dans TOPHAT (Trapnell et al. 2009) ou encore STAR (Dobin et al. 2013).

Afin de détecter les variations de séquences dues aux épissages alternatifs, la localisation des lectures de transcrits est comparée à l'annotation afin de fournir une nouvelle annotation. Cette nouvelle annotation peut être effectuée par des outils tel que Cufflinks (Trapnell et al. 2010) et permet de générer par la suite les nouveaux transcrits. Ces derniers, en considérant les exons nouvellement détectés, pourront être intégrés aux bases de données protéogénomiques.

Un autre type de variation de séquence peut être dû à la présence de variants nucléotidiques, par exemple, dans le cadre de maladies tel que le cancer. Afin de détecter ces variations des logiciels comparent les séquences des lectures alignées à la séquence de référence. En cas de discordance entre les deux séquences, une étude de la profondeur de détection de cette variation et une application d'algorithme dit d'apprentissage sont utilisées pour différencier le bruit (erreur de séquençage) des réelles variations. Ces méthodologies sont développées dans la suite logicielle GATK (McKenna et al. 2010). Les variations peuvent ensuite être intégrées aux séquences originelles afin de les intégrer à une base de données protéogénomique.

- Constitution d'une base de données dans le cas d'organismes non-modèles:

Pour les organismes n'ayant pas de génome séquencé ou ayant un génome de mauvaise qualité les données peuvent être directement exploitées pour générer par assemblage *de novo* les séquences du génome ou des transcrits. Les méthodes d'assemblages *de novo* reposent principalement sur l'utilisation des chevauchements entre séquences afin de les réunir pour créer des morceaux contigus de séquences. Les algorithmes du logiciel d'assemblage utilisé dans la thèse reposent sur l'utilisation des graphes de De Bruijn (Trinity). Ces graphes orientés permettent de représenter les chevauchements entre les sous mots présents dans les lectures et ainsi de reconstruire l'enchainement des séquences. Pour la reconstruction de génome le logiciel nommé Velvet (Zerbino & Birney 2008) est basé sur cette approche. Dans le cas des lectures de transcrits, les variants d'épissage doivent être pris en compte comme le fait le logiciel Trinity (Grabherr et al. 2013). Une base de données dérivée de la traduction de ces assemblages permet ensuite de produire la base de données protéogénomique.

c) Outils d'évaluation de l'assemblage

La qualité des assemblages est un point clef pour les études protéogénomiques. En effet, l'interprétation des spectres étant dépendante de la traduction des séquences, un assemblage comportant de nombreuses erreurs peut induire une interprétation erronée par la suite. Afin de vérifier la qualité des assemblages, plusieurs méthodologies sont possibles.

- Mesure des paramètres d'un assemblage pour en évaluer sa qualité :

L'analyse des longueurs des séquences contigües assemblées est le premier paramètre statistique indicateur de la qualité de l'assemblage. Il est considéré que plus la longueur des séquences assemblées est grande plus l'assemblage semble être de bonne qualité. Pour cela, la taille moyenne, la taille de la plus grande séquence et le nombre de séquences dont la taille est supérieure à une longueur donnée sont utilisés. En complément de ces valeurs, des valeurs spécifiques à l'assemblage

appelées Nx (x étant un nombre donné entre 0 et 100) ont été mises en place. Ce nombre correspond à la taille de la plus grande séquence possible telles que toutes les séquences ayant une taille supérieure ou égale à celle-ci représentent x % du nombre de bases du séquençage. Ces différents paramètres peuvent être calculés grâce à l'outil Transrate (Smith-Unna et al. 2016). Transrate propose en plus de calculer le taux de lectures localisées sur les transcrits après assemblages et propose de fournir un score.

- Utilisation des séquences orthologues de référence
- L'utilisation de séquences attendues comme conservées entre tous les organismes vivants à un niveau taxonomique donné est aussi possible. En effet, on s'attend à ce qu'un certain nombre de séquences soit conservé, notamment celles codant pour des protéines impliquées dans les mécanismes biologiques de bases pour un niveau taxonomique donné. Ces séquences sont alors considéré gènes orthologues. L'outil BUSCO (Simão et al. 2015) utilise les données de la base de OrthoDB pour valider le bon assemblage de ces gènes homologues au sein d'un transcriptome assemblé. Ainsi BUSCO permet de connaitre parmi les homologues à un niveau taxonomique donné le nombre de gènes retrouvés complets, fragmentés ou même absents. Cette information permet ensuite de comparer les assemblages et d'en évaluer la qualité. Ainsi, la proportion de gènes orthologues retrouvée indique la qualité de l'assemblage. La protéomique, une solution récente pour l'évaluation de la qualité d'un assemblage

La protéomique a été récemment proposée afin de comparer les résultats de différents outils d'assemblages dans une optique de protéogénomique (Ma et al. 2018, Luge et al. 2016). Dans ces deux publications, le taux d'attribution de spectres MS/MS, en utilisant les bases de données générées à partir des logiciels, sert d'évaluateur de la qualité des bases. Cette méthodologie permet une validation plus spécifique de l'échantillon analysé. Un apport essentiel de cette analyse protéomique est la capacité d'évaluer de façon intégrale le transcriptome en ne se concentrant pas uniquement sur les gènes orthologues. En effet les gènes orthologues peuvent constituer un biais dans l'analyse de la qualité car ils sont pour la plupart des gènes essentiels dans les mécanismes biologiques de bases pour le maintien de l'intégrité des organismes. Or, ces gènes possèdent un niveau de conservation plus important que les gènes spécifiques de la lignée taxonomique considérée et probablement un taux d'expression relativement élevé. Dans ces deux cas leur évaluation perd en pertinence car il s'agit de gènes plus simples à assembler que la majorité, car abondants. Donc se baser sur leur bon assemblage pour évaluer l'assemblage de l'ensemble des gènes constitue un biais dans l'évaluation de la qualité globale de l'assemblage.

3) La protéomique shotgun

a) La protéomique shotgun, une protéomique de découverte

La protéomique ascendante dites protéomique shotgun repose sur l'étude complète d'un protéome après digestion par une protéase des protéines. Cette approche permet d'étudier plus facilement l'intégralité d'un protéome. L'approche repose sur la mesure par un spectromètre de masse du rapport masse/charge de peptides libérés par la protéolyse des protéines. La protéase la plus utilisée est la trypsine qui permet de couper les protéines après les acides aminés avec une chaîne latérale chargée positivement (Lysine ou Arginine). Les peptides sont ensuite séparés sur une chromatographie en phase inverse, couplée au spectromètre de masse, qui trie les peptides en fonction de leur niveau d'hydrophobicité. Les peptides en sortie de chromatographie sont ensuite ionisés par une source et introduits dans le spectromètre de masse. Une fois que le rapport masse/charge des peptides ionisés

est défini, ces derniers sont fragmentés dans une cellule de collision et les rapports masses/charges des sous fragments correspondants sont analysés à leur tour. Cette approche fournit donc des données à deux niveaux, celui des ions dit « parents », MS1, et des ions dit « fils » correspondants, MS2.

b) Le Q-Exactive HF

Le spectromètre de masse utilisé lors de mes travaux de thèse est le Q-Exactive HF qui est commercialisée par la société TermoFisher scientific. Cet appareil repose sur l'utilisation de 4 composants principaux. En amont du spectromètre de masse un système de chromatographie liquide haute performance permet de séparer les différents peptides afin de réduire le nombre de peptides analysés en parallèle dans le spectromètre de masse. Le flux de peptides à analyser est donc assuré en continu jusqu'au premier composant du spectromètre de masse.

- Source ionisante, source ESI :

Afin de pouvoir analyser les peptides, se trouvant à l'initiale en solution, il est nécessaire de les transformer en phase gazeuse. La source d'ionisation par électronébuliseur (ESI) permet de faire passer les peptides de la phase aqueuse à une phase gazeuse adaptée à l'analyse par le spectromètre de masse. Il s'agit d'une source ionisante dite douce car elle permet de ne pas fragmenter les peptides lors de leur entrée dans le spectromètre de masse.

- Analyseur de type quadripôle :

L'analyseur sert de filtre et nous permet de sélectionner parmi les peptides en phase gazeuse uniquement les peptides ayant la masse sur charge souhaitée pour l'étude. Ce filtre est appliqué par la modulation des champs électriques qui permet de trier les peptides en fonction de leur masse sur charge.

- Chambre de fragmentation, cellule de collisions HCD :

Cette chambre de fragmentation permet de produire les fragments à partir des ions parents. Cette fragmentation est effectuée grâce à une collision avec un gaz inerte (Azote) qui va casser l'ion parent en 2 sous fragments.

- Analyseur couplé au détecteur, Orbitrap :

Cet analyseur utilise le principe d'attraction afin de faire effectuer des rotations aux peptides autour d'un axe ovale chargé. Des champs électriques sont appliqués autour de cet axe et permettent aux ions d'effectuer des oscillations dépendantes à la fois de la charge et de la masse de l'ion. Cette méthode permet de minimiser le nombre de molécules effectuant des orbites similaires et donc d'éviter le risque de collision. Chaque espèce d'ion effectue sa propre oscillation et tourne autour de l'axe principal. Ce mouvement est ensuite enregistré par les détecteurs et par une transformation de Fourier ce mouvement sinusoïdal est transformé en spectre correspondant à l'intensité en fonction de la masse sur charge d'un ion.

Dans le cas de l'analyse d'un échantillon le spectromètre de masse fonctionne en deux étapes principales.

- Analyse des peptides tryptiques :

La totalité des peptides co-élué sont ionisés. Le quadripôle permet la sélection des peptides positifs dans la gamme de masse sur charge définie par l'utilisateur. L'ensemble de ces peptides est ensuite

analysé et mesuré par l'orbitrap. Cette première analyse permet d'obtenir le spectre MS1 des peptides tryptiques. Elle permet aussi de définir l'abondance de chaque peptide tryptique à partir de laquelle les ions à analyser après fragmentation sont sélectionnés en fonction du paramétrage défini par l'utilisateur. Ce nombre est limité car il dépend de la fréquence de l'analyseur (20 Hz pour le QExactive). Ici pour l'exemple considérons que seuls les 20 peptides les plus abondants seront étudiés.

- Analyse des fragments de peptides :

Cette étape s'effectue en plusieurs sous-cycles. La méthode d'analyse des protéomes par spectrométrie de masse s'appuie sur la sélection des 20 peptides les plus abondants (Méthode Top20) qui se divise en 20 sous-cycles. Pour chacun des sous-cycles, la totalité des peptides co-élués seront ionisés mais, cette fois ci, le quadripôle assure uniquement la sélection d'un peptide qui correspond chacun son tour aux 20 peptides abondants définis dans l'étape précédente. Ce peptide sera ensuite envoyé dans la chambre de fragmentation et les sous fragments seront envoyés dans l'orbitrap pour être analysés et produire les spectres MS2.

c) Amélioration possible en spectrométrie de masse

Il existe quatre axes principaux d'amélioration en spectrométrie de masse. D'une part la diminution des temps de transfert d'un composant à l'autre du spectromètre et de sélection des ions parents à analyser permet un gain notable dans le nombre de spectres générés et donc dans le nombre de peptides détectables. Un autre gain de temps potentiel repose sur la rapidité et le nombre d'ion que l'analyseur et les détecteurs peuvent analyser. Une amélioration de ce système permet aussi de détecter plus de spectres et donc à terme de détecter aussi plus de peptides. La troisième possibilité provient d'une observation qui peut être effectuée sur les données après analyse. En effet même si le nombre de spectres détectés augmente, ce n'est pas toujours le cas de la proportion de spectres interprétés et attribués à une séquence peptidique. En effet, la qualité des spectres analysés est elle aussi déterminante de l'interprétation de l'analyse. La co-élution d'un trop grand nombre de peptides ou la mauvaise séparation de ces derniers produisent des spectres difficilement analysables. Dans ce cas, les spectres ne peuvent pas être interprétables et de ce fait ne permettent pas l'identification de séquences peptidiques. Pour résoudre ce problème, une séparation supplémentaire des ions en fonction de leur mobilité ionique permet d'augmenter le nombre de spectres interprétables. La spectrométrie de masse à mobilité ionique consiste à soumettre les ions à un champ électrique. Le déplacement des ions en fonction de sa charge et de sa masse assure leur séparation.

4) Interprétation des données en protéomique shotgun

a) Pré-traitement des fichiers de sortie du spectromètre de masse

Les données brutes du spectromètre de masse (RAW) sont ensuite analysées par le logiciel ProteomeDiscoverer. Ce logiciel permet de faire un tri sur les spectres de mauvaise qualité. De plus ce logiciel permet de simplifier les données de spectrométrie de masse. Par exemple, les spectres des ions parents sont réduits au pic le plus intense du massif isotopique.

Le logiciel Excalibur permet de visualiser non seulement l'ensemble des spectres mais aussi le détail des massifs isotopiques des ions parents qui ne sont plus visibles après leur conversion en MGF par ProteomeDiscoverer.

b) Attribution d'une séquence à un spectre MS/MS

Afin d'attribuer une séquence aux spectres générés par le spectromètre de masse il existe deux méthodologies principales.

- Attribution *de novo* des peptides :

L'analyse des spectres MS/MS peut permettre de reconstruire la séquence du peptide en utilisant la masse du peptide parent ainsi que celle des ions fils générés après fragmentation. Des outils tels que Novor ; PepNovo ou PEAKS assurent ce type de traitement. Cependant, avec la méthode *de novo*, uniquement 35% des peptides identifiés possèdent une séquence correcte, ce qui rend complexe l'analyse des résultats produits par ce type de méthode (Muth et al. 2018).

- Utilisation d'un moteur de recherche et d'une base de données :

Afin de simplifier l'attribution des spectres à une séquence peptidique la solution est d'utiliser une banque de données de protéines potentiellement présentes dans l'échantillon. Les protéines de la banque sont ensuite digérées *in silico* afin d'obtenir pour chacune une liste de peptides. Les peptides sont ensuite fragmentés *in silico*. Ainsi, pour l'interprétation, les spectres générés par spectrométrie de masse sont comparés avec les données de digestion et fragmentation *in silico* intégrant une erreur de masse tolérée qui dépend de la technologie de mesure de masse utilisée. Grâce à cette approche, il devient possible d'intégrer l'analyse des modifications post-traductionnelles. Parmi les outils les plus répandus pour faire cette recherche il existe des logiciels libres d'utilisation tels que Xtandem, OMSSA, Andromeda ou bien MS-GF+ ainsi que des logiciels payants tel que Mascot. Mascot est le logiciel choisi par l'équipe pour cette thèse de par son antériorité ainsi que de par le support informatique proposé par Matrix Science. Mascot sera donc l'outil présenté plus en détail dans la suite de ce manuscrit.

c) Fonctionnement du moteur de recherche Mascot

Le logiciel Mascot est un logiciel sous licence commerciale dont l'ensemble des calculs des scores et des seuils de validation des spectres n'est pas accessible aux utilisateurs. Ainsi certains calculs ne pourront être présentés dans ce manuscrit. Le moteur de recherche fonctionne en trois étapes principales :

- Digestion de la base de données :

La base de données utilisée par l'utilisateur est entièrement digérée *in silico* par la protéase utilisée pour l'analyse en spectrométrie de masse. La masse des peptides parents générés *in silico* est alors comparée à la masse des peptides parents obtenus avec les spectres expérimentaux (MS). Cette première étape permet d'établir un premier paramètre nommé « Qmatch ». Le Qmatch correspond au nombre de candidats pour chaque spectre en se basant uniquement sur le nombre de peptides parents *in silico* ayant la même masse que le peptide parent mesuré expérimentalement.

- Attribution des séquences de peptides aux spectres :

Pour chacun des peptides candidats pour un spectre donné les fragments *in silico* sont alors générés. Ces fragments sont alors comparés au spectres expérimentaux MS2. Le nombre de spectres MS2 concordant avec les masses des fragments générés *in silico* permettent d'établir un score nommé Ion Score. En fonction du Qmatch et du risque d'obtenir le même résultat par hasard (p-value) accepté par l'utilisateur, une valeur seuil nommée « Mascot Identity Threshold » peut être fixée. Cette valeur seuil correspond à l'Ion Score minimal accepté pour l'attribution d'une séquence peptidique à un spectre (« peptide spectra match » PSM). Mascot propose d'utiliser un second seuil nommé « Mascot Homology Threshold » qui permet une attribution moins restrictive en étudiant la distribution des meilleurs scores d'attribution pour un spectre donné.

- Attribution des peptides aux protéines :

Les protéines peuvent posséder de nombreux peptides communs, notamment dans le cas des protéines isoformes ou partageant des fonctions communes. Afin d'attribuer les séquences peptidiques identifiées aux protéines, celles-ci sont d'abord regroupées en famille en fonction du

nombre de peptides communs qu'elles possèdent. Dans un premier temps, au sein des familles sont définit 3 types de protéines. Les protéines « leader » possédant le plus de peptides « discriminants » (bold-red). Dans un second temps, sont définies les protéines avec les mêmes ensembles de peptides (same-set) et celles possédant des sous-ensembles de peptides (sub-set). Par la suite lors de l'analyse, le nombre de spectres des peptides peut être partagé entre les protéines possédant le plus de peptides par parcimonie. Cette manipulation peut cependant être problématique lors de la présence de nombreux isoformes de protéines possédant un grand ensemble de peptides communs.

d) Gestion des faux positifs

Le grand nombre de comparaisons et de tests effectués lors de l'attribution des peptides aux spectres augmente le risque d'obtenir de fausses attributions. Par exemple en prenant 1% de risque d'attribution aléatoire pour un test donné on se retrouve avec un taux de faux positifs (FDR) supérieur à 1% si ce test est répété de multiple fois. Afin de contrôler ce risque, il existe une méthodologie de contrôle de ce taux de faux positif au niveau des spectres, des peptides et des protéines.

- Spectres

En ce qui concerne le contrôle de la FDR dans le cas des attributions des spectres à une séquence de peptide, Mascot utilise le paramètre Qmatch. En effet plus le Qmatch sera élevé plus la chance d'avoir une attribution aléatoire est grande. Dans ce cas le seuil de validation doit donc être plus restrictif et le seuil MIT ou MHT devient plus élevé.

- Spectres ou peptides ou protéines

En ce qui concerne la gestion de la FDR, il existe deux solutions proposées par Mascot. La plus répandue est l'utilisation d'une base de données leurre par lecture inverse de la base de données cible utilisée (Elias et Gygi 2011). Le nombre d'identifications à une p-value donnée permet d'estimer le taux de faux positifs présent dans la base et d'ajuster la p-value afin que ce taux de faux positif corresponde à une valeur définie par l'utilisateur. La seconde solution pour établir le taux de faux positifs est l'utilisation d'un programme utilisant un algorithme d'apprentissage nommé Percolator (The et al. 2016). L'algorithme repose sur un entrainement permettant d'établir les paramètres de validation en utilisant les identifications possédant les scores les plus élevés. Ces méthodologies permettent d'estimer le taux de faux positifs que ce soit au niveau spectre, peptide ou protéine. Dans le cas des protéines, un filtre usuellement appliqué est l'utilisation des protéines identifiées avec au moins deux peptides différents.

e) Quantification par spectrométrie de masse

Deux unités de mesure sont possibles pour la quantification de l'expression des protéines. D'une part le comptage le plus simple consiste à compter le nombre de spectres attribués aux protéines. Cependant les spectres sont reliés à une donnée d'intensité qui permet une quantification plus précise. L'utilisation des pics d'intensité élué, à un temps « t » donné de l'échantillon, permet d'augmenter la précision de la mesure car elle permet de considérer des spectres n'ayant pu être attribués. La valeur ainsi calculée se nomme comptages du chromatogramme des ions extraits (XIC) et est implémentée notamment dans le logiciel Maxquant. Cependant, cette valeur n'est pas optimum dans le cadre de la protéogénomique possédant des bases de données potentiellement partielles et contenant une quantité anormale de faux peptides. L'utilisation du XIC, dans ce cas-là, induit des erreurs potentielles en augmentant le risque de cumuler le comptage de peptides différents élués à un temps similaire entre les échantillons.

5) Concepts et outils bioinformatiques pour l'analyse des données de protéomique

a) Librairie pour l'analyse de données

Certaines librairies permettent l'analyse et la conversion de données protéomiques. Ces librairies permettent d'accélérer le traitement non seulement des tables de peptides et protéines mais aussi les différents calculs en ligne de commande. Les principaux outils ou librairies sont présentés dans la liste suivante :

- DecoyPyrat : Ce logiciel permet la création d'une base de données leurre à partir du fichier Fasta de la base de données cible par mélange aléatoire. L'avantage de ce logiciel développé en Python est qu'il vérifie que les séquences ainsi générées ne produisent pas de peptides provenant de la base de données cible.

- Pyteomics : Développé par Levitsky et al. 2018, cette librairie permet de créer une interface de traitement en python pour les données protéomiques. Les fonctionnalités de cette librairie permettent notamment la conversion des formats protéomiques ou l'interprétation statistique des résultats.

- Biopython : Cette librairie python est une librairie libre permettant l'analyse et l'interprétation de la majorité des formats bioinformatiques. De plus cette librairie donne accès aux systèmes de requête des plus grandes bases de données telles que NCBI ainsi qu'à leurs outils par requête web.

- RforProteomics : Développé par Gatto & Christoforou 2014. Cette librairie permet d'avoir accès au différents packages R permettant l'analyse ou le traitement de données protéomiques. Par exemple, l'outil permet d'avoir accès aussi bien aux packages d'illustration des résultats qu'au contrôle qualité des fichiers issus de spectrométrie de masse.

- Pandas : Pandas est un package python de traitement des données bioinformatiques. Son utilisation permet notamment la lecture et l'interface avec les formats propriétaire EXCEL. De plus la librairie possède de nombreuses fonctions mathématiques dérivées de la librairie numpy, optimisée pour les calculs.

- GOOEY : Ce package python a pour objectif de pouvoir générer une interface graphique en quelques lignes de code. Il s'agit aujourd'hui de l'alternative la plus rapide et simple pour produire une interface utilisateur d'un script codé en python.

b) Plateforme graphique pour analyser les données protéomiques

Une plateforme graphique permet l'accès aux outils informatique sans ligne de commande via une interface graphique ou une page web. Ainsi, l'ensemble des outils sont accessibles aux biologistes. Il existe trois plateformes graphiques gratuites principales pour l'analyse de données protéomiques :

- Galaxy : Galaxy-P développée par l'équipe américaine (Sheynkman et al. 2014) et Proteore (http://www.proteore.org/) développée par l'équipe grenobloise sont deux plateformes dérivée de Galaxy. L'avantage de ces deux plateformes est de rendre disponible graphiquement les scripts dérivés des librairies d'analyse. De plus, cette plateforme permet d'installer automatiquement tout nouvel outil ou de mettre à jour les outils existants. Enfin cette plateforme met en place des outils de visualisation des résultats protéomiques. Il s'agit de la seule plateforme en accès libre de droit permettant de visualiser et modifier le code de la plateforme.

- Maxquant & Perseus : Maxquant a été développé par Jürgen Cox & Matthias Mann en 2008. Il permet d'attribuer des séquences aux spectres, et de quantifier les peptides par l'analyse du XIC. Ce logiciel d'exploration produit des données sous un format compatible avec le logiciel Perseus (Tyanova et al. 2016) auquel il est associé pour l'analyse statistique. Ces outils tirent l'avantage de l'exploitation transversale des répliquas des échantillons afin d'analyser le plus de spectres possibles en utilisant le moteur de recherche Andromeda. De plus l'utilisation du comptage XIC permet d'obtenir une quantification plus précise que l'analyse par comptage de spectres dans le cas de l'utilisation de base de données possédant peu d'erreur de séquence. L'analyse statistique par Perseus est simplifiée et permet, avec une interface centrée utilisateur, de mettre en place des méthodologies de traitements hautement reproductibles. De plus, Perseus permet de nombreuses analyses statistiques sans être uniquement centré sur l'analyse de l'expression différentielle.

- PatternLab : PatternLab est un logiciel avec interface graphique développée par Carvalho et al. 2008. Ces fonctionnalités initiales sont similaires aux fonctionnalités de Perseus mais ne s'appliquent principalement qu'à l'étude de l'expression différentielle en terme d'analyse statistique. PatternLab possède des modules complémentaires permettant l'attribution des séquences au spectres par exemple. Il regroupe en un seul logiciel les fonctionnalités de Maxquant et Perseus. L'avantage supplémentaire est la possibilité d'utiliser une version serveur de PatternLab pour les analyses. Cela permet d'installer un second logiciel client sur les ordinateurs d'une équipe de recherche et de forcer l'utilisation d'une et une seule version de PatternLab pour toute l'équipe.

6) Infrastructures informatiques

a) Choix des environnements de travail

Dans le cadre de ma thèse pour des raisons de politique de sécurité le système d'exploitation Windows a dû être utilisé pour l'ensemble des traitements. Cependant les systèmes d'exploitation peuvent être une limitation importante en bioinformatique. Par exemple, la majorité des logiciels ne fonctionnent que sous des systèmes d'exploitation possédant un noyau UNIX (MAC ou Linux). De plus, parfois, les logiciels peuvent répondre différemment en fonction de l'état du système d'exploitation (version, logiciels concurrents installés) et nuire à la reproductibilité scientifique. Pour lutter contre ces limitations une approche de container peut être utilisée. Cette approche, implémentée par Docker (Boettiger 2015) ou Singularity (Kurtzer et al. 2017) par exemple, permet de créer une boite virtuelle contenant uniquement le système d'exploitation et le logiciel. Cette approche permet de fournir ce container pour pouvoir reproduire les traitements de données. L'avantage de Docker est qu'il est

possible de l'installer sous Windows. En utilisant une machine virtuelle (Hyper-V), Docker permet de générer un sous-système linux communiquant avec Windows et permettant d'exécuter les containers et donc l'ensemble des logiciels bioinformatiques. Reposant sur cette technologie de container, une initiative nommée Biocontainer permet de retrouver une grande partie des logiciels bioinformatiques sous forme de container prêt à l'emploi (Da Veiga Leprevost et al. 2017).

Une autre solution existe depuis 2018. Elle permet d'installer un sous-système linux sur Windows (WSL). Il était déjà possible depuis longtemps d'installer des machines virtuelles simulant un ordinateur sur lequel on pouvait installer un autre système d'exploitation. Cependant, dans le cas des machines virtuelles, une partie seulement des ressources de l'ordinateur pouvait être utilisée. Grâce à WSL un environnement Linux complet est disponible sous Windows et peut aussi permettre d'exécuter les logiciels bioinformatiques.

b) Installation des logiciels

Les logiciels bioinformatiques nécessitent parfois l'installation d'autres programmes pour utiliser leurs fonctionnalités. On nomme cela des dépendances. Certains logiciels tels que la suite logiciel GATK nécessite plusieurs dizaines de dépendances qui elles-mêmes nécessitent d'autres dépendances. La gestion de ces dépendances peut être fastidieuse et la suite d'outils GATK peut donc être installée en plusieurs jours voire mêmes semaines. Pour pallier à ce problème une initiative nommée Conda et plus particulièrement Bioconda permet d'installer automatiquement les logiciels bioinformatiques ainsi que leurs dépendances (Grüning et al. 2018). Grâce à cette initiative, les logiciels peuvent être simplement installés grâce à la commande « conda install », par exemple en quelques minutes pour GATK. De plus, Conda permet le maintien et le choix des mises à jour pour les différents logiciels, ce qui permet de rendre plus reproductible les différents traitements bioinformatique. Enfin, Conda permet de gérer les limitations administrateur afin d'installer aisément sans les droits d'administrateur un logiciel bioinformatique sur son espace local.

c) Gestion des enchainements d'outils

Les méthodologies de traitement des données biologiques pour répondre aux questions biologiques demandent l'utilisation combinée de plusieurs outils. Ces outils ont chacun leurs paramètres, leurs formats de fichier d'entrée ou de sortie qui rendent difficile leur enchainement directement dans un script simple. De plus l'accès aux outils se fait dans la plupart du temps en ligne de commande ce qui est parfois un frein à leur utilisation par les biologistes. Afin de rendre plus reproductible et accessible ces méthodologies, des méthodes de gestion de ces enchainements ont été développées.

- Gestionnaire en ligne de commande

Afin de pouvoir enchaîner facilement les outils des gestionnaires en ligne de commande tels que Snakemake ont été mis en place (Köster & Rahmann 2012). Snakemake est inspiré de la structure des Makefile. Il permet de lister une liste d'outils de fichier d'entrée et de paramètres pour chacun et gère la répartition des tâches aussi bien sur un ordinateur qu'un regroupement d'ordinateurs (cluster). Ce logiciel permet aussi de générer automatiquement une figure de l'enchainement des outils afin de l'intégrer à un rapport de traitement des données.

- Gestionnaire en interface graphique

Pour rendre accessible les logiciels bioinformatiques aux biologistes, une interface graphique est nécessaire. A cet effet le projet Galaxy a été mis en place (Afgan et al. 2018). Il s'agit probablement du

projet bioinformatique réunissant le plus grand nombre de chercheurs à ce jour avec près de 7500 publications reliées au projet. L'idée est de créer une interface graphique simple d'utilisation et simple à alimenter avec de nouveaux outils. Pour cela, l'ajout d'un outil nécessite un simple fichier de configuration avant d'être ajouté à une librairie d'outil (toolshed) et d'être accessible à tous les chercheurs de la communauté, ce qui est le cas de plus de 5500 outils disponibles à ce jour. Un autre apport essentiel de ce projet est la gestion de l'historique de traitement publiable qui permet de garder une traçabilité parfaite des traitements effectués sur les données. Ce projet tire parti de l'utilisation de Bioconda et de Docker pour rendre l'installation des outils la plus simple et reproductible possible. Il existe même une installation de Galaxy possible par Docker ce qui permet son utilisation sous Windows. C'est donc l'interface Galaxy qui a été choisie pour la majorité des traitements de cette thèse.

7) Limites actuelles et enjeux en protéogénomique

a) Limitations actuelles

La principale limitation de la protéogénomique se trouve dans la nature des bases de données générées qui ne sont pas toujours adaptées aux logiciels d'attribution protéomique. D'une part la base de données protéogénomique diffère de par le fait qu'une base de données habituelle possède uniquement des séquences « vraies » pouvant être présentes dans l'échantillon. Dans le cas de la protéogénomique une traduction systématique en 6 cadres de lecture génère une proportion importante de « fausses » protéines ne pouvant être observées en protéomique. De plus la taille des bases de données ainsi générée est d'une taille largement supérieure aux bases de données actuelles.

Ces deux phénomènes rendent plus complexe le contrôle de la FDR par une approche cible/leurre. Cette méthodologie repose sur le fait que la base de données cible ne contient que des séquences « vraies » ce qui n'est pas le cas en protéogénomique. De plus, la taille de la base de données cible induit la création d'une base de données leurre de même taille dans laquelle des séquences de peptides réellement présentes peuvent être générées aléatoirement. Or cette méthodologie repose aussi sur le fait de ne posséder que des séquences fausses dans la base de données leurre. La taille imposante de ces bases de données induit aussi une multiplication des peptides candidats concurrents possédant la même masse et induit un seuil de score de validation plus élevé à obtenir lors de l'attribution d'un spectre à un peptide.

Enfin au niveau protéique les outils d'assemblage RNA-seq génèrent de nombreux isoformes pour un même transcrit ce qui induit la création de plusieurs protéines possédant une grande partie de leur peptide en commun. Lors de l'attribution des peptides à une protéine, les peptides se différentiant d'un isoforme à l'autre peuvent permettre d'appliquer une parcimonie et de faire un choix sur l'isoforme présent dans l'échantillon. Cependant lors d'études sur de multiples échantillons ces peptides discriminants peuvent ne pas être détectés et ainsi les peptides et comptage de spectres peuvent ne pas être répartis correctement pour une étude d'expression différentielle en aval de cette attribution.

b) Aller au-delà des limites

A partir des observations suivantes des méthodologies ont été mises en œuvre afin de permettre une meilleure attribution dans les études protéogénomique. Une première solution est proposée par Jagtap et al. 2013. Le concept est d'effectuer la recherche en deux tours sur la base de données. Une

première recherche, en utilisant la base de données complète, permet la détection des protéines présente dans l'échantillon. Une seconde recherche est effectuée sur une sous base de données contenant uniquement les protéines observées dans le premier tour. Cette solution permet de mieux attribuer les peptides aux différentes protéines détectées dans l'échantillon en utilisant un espace de recherche restreint. Une autre solution proposée est celle de contrôler la FDR différemment lors de l'attribution. Pour cela un algorithme de classification par apprentissage peut permettre d'évaluer la proportion de bonne et mauvaise attribution et ainsi établir le taux de faux positif (The et al. 2016). Une autre solution proposée par Sheynkman et al. 2016 est d'utiliser des outils de recherche de cadre ouvert de lecture afin de réduire la base de données. Ces outils se basent sur un entrainement à partir de protéines existantes proposées ou en sélectionnant un sous ensemble des cadres ouverts de lecture les plus long afin de valider ou d'invalider les plus cours. Enfin des outil graphique d'intégration des données protéomique pour des analyses complémentaires sont nécessaire pour rendre ces analyses plus accessible à tous à l'instar du logiciel MASPECTRAS 2 qui permet de rendre accessible les données de protéines identifiés (Ceereena et al. 2010).

III-La Protéogénomique en écotoxicologie

1) Ecotoxicologie

a) Définition de l'écotoxicologie

L'écotoxicologie à comme origine la toxicologie appliquée à l'écologie. Pour bien comprendre ce concept il faut donc remonter à l'origine de la toxicologie. La toxicologie a pour objectif d'évaluer le danger d'une substance ainsi que la probabilité d'exposition à cette substance. Le danger d'une substance peut être étudié via l'impact de celle-ci sur des organismes vivants afin d'en définir la dose minimale effective et d'évaluer les réponses de l'organisme vivant. La probabilité d'exposition quant à elle va concerner le temps d'exposition, la voie d'exposition et prendre en compte les différents paramètres qui concernent l'organisme exposé (développement, sexe).

L'écologie quant à elle concerne l'étude de l'environnement. Cette science prend en compte l'étude de population et d'écosystèmes (mélanges d'organismes) pouvant être présents dans un biome donné. On étudie dans cette science par exemple les paramètres physiologiques impactant la santé ou bien la démographie des différents organismes vivants. Son alliance à la toxicologie permet donc de comprendre et d'analyser la présence de substances néfastes dans un milieu naturel en observant les organismes vivants.

La notion d'écotoxicologie en elle-même provient de la toxicologie de l'environnement. Le terme lui-même est créé par le français René Truhaut en 1977. De nos jours, l'écotoxicologie permet de mettre en évidence l'impact de l'Homme sur son environnement. Ces résultats répondent à une préoccupation sociétale majeure. En effet, dans le cadre du projet « un monde, une santé », l'organisation mondiale de la santé (OMS) définit comme primordiale pour la santé de l'homme, l'étude de son environnement et des organismes y vivant.

b) Apport de l'écotoxicologie

De nos jours les techniques de dosages moléculaires sont suffisamment précises pour détecter la présence d'une molécule dans l'eau ou l'air par exemple. Cependant son dosage ne permet pas de connaitre l'impact sur le vivant et d'en établir le risque (Connon et al. 2012). L'observation de la bioaccumulation de molécules chimiques dans les êtres vivants et son étude trans-générationnelle peut quant à elle mettre en évidence les impacts physiologiques (Thompson et al. 2004). Si ces impacts physiologiques peuvent dans un premier temps être sans conséquences apparentes sur les organismes étudiés, les technologies actuelles d'analyses du vivant peuvent mettre en avant des mécanismes moléculaires activés. Ces mécanismes peuvent par la suite être étudiés en laboratoire pour caractériser les paramètres liés aux probabilités d'exposition et mieux estimer le danger lié à ces molécules (Bambino & Chu 2018).

De nos jours, les molécules utilisées par l'industrie, le milieu médical ou encore l'agriculture sont soumises à de nombreux contrôles avant commercialisation. Cependant, pour définir leur cadre d'utilisation et caractériser leurs modifications après utilisation, les expériences sur les modèles de laboratoire peuvent ne pas être suffisantes. Dans ces cas notamment, l'écotoxicologie complète les premières données car c'est le domaine qui va plus loin en s'intéressant au décryptage de l'impact sur le vivant de ces produits.

c) *Utilisation d'organisme sentinelle en ecotoxicologie*

Un des points fort de l'écotoxicologie est la mise en place de biomarqueurs permettant le suivi et la détection de polluant de l'environnement (Lopez-Barea 1995). L'eau est un des milieux le plus impacté par les activités humaines. Diverses contaminations sont présentes telles que des métaux lourds, des perturbateurs endocriniens ou encore même des pesticides. Afin d'explorer la contamination en milieu naturel il faut des organismes vivant présent dont on peut évaluer la santé. Historiquement, bien avant l'écotoxicologie moderne les mineurs utilisaient déjà des canaries qui mourraient en présence de monoxyde de carbone. Cette méthodologie utilisait donc des animaux sentinelles permettant la détection du danger. Aujourd'hui en écotoxicologie, on utilise toujours le même principe.

Afin de pouvoir comprendre les mécanismes biologiques impactés par l'environnement, dans un premier temps, il est plus simple de travailler sur des organismes très étudiés en laboratoire. Dans le cas des organismes aquatiques étudiés en laboratoire, les connaissances scientifiques sur le poisson zèbre sont telles, qu'il est une cible de choix pour les études d'écotoxicologie aquatique (Dai et al. 2014). Aujourd'hui le poisson zèbre reste même l'animal recommandé par l'union européenne pour les expériences alternatives aux tests sur les animaux (EURL-ECVAM). Ce poisson est alors encagé en milieux naturel et ces paramètres physiologiques et moléculaires contrôlés par rapport à des standards établis en laboratoire. Ces différentes mesures ont permis d'établir d'une part des critères physiologiques de bonne santé de l'environnement et d'une autre, d'établir les molécules à doser indicatrices de son état de santé, appelées biomarqueurs. L'évolution des technologies permet de ne pas se concentrer uniquement sur ces espèces dites modèles mais aussi d'explorer d'autres organismes. Ces méthodologies permettent aussi l'analyse de maladies pour des espèces d'intérêt pour le domaine agroalimentaire (Soares et al. 2012, Marques et al. 2019).

2) Objectifs et enjeux des omiques pour l'écotoxicologie

a) *Apport de la génomique pour l'écotoxicologie*

Les études génomiques peuvent être exploitées en écotoxicologie pour démontrer et évaluer l'impact transgénérationnel (Zhang et al. 2018) des produits contaminants. En effet, l'analyse des modifications de la conservation en nucléotide des gènes liée à une réponse à un polluant peut démontrer l'implication de certains gènes, voire même une cascade de gènes, dans la réponse au stress induit. De plus, une étude des modifications épigénétiques telles que la méthylation de l'ADN permet de divulguer la régulation de l'expression au niveau génomique, en réponse au stress. Enfin au sein d'organismes, tels que les plantes, l'étude du nombre de copies de gènes et l'analyse de variants mettent en évidence des systèmes alternatifs mis en place, en réponse au stress. Enfin la sur-sélection de variants géniques dans les régions régulatrices ou codantes révèle une pression sélective positive pour ces variations induites par la présence des polluants.

b) *Apport de la transcriptomique pour l'écotoxicologie*

L'avantage de l'étude transcriptomique est de pouvoir intégrer les données d'expressions. En effet en transcriptomique, les analyses d'expressions différentielles mettent en évidence là sur ou sous expression d'ARNm, en réponse aux modifications environnementales. Dans ce cas, les variations génétiques, présentes dans les régions codantes, qui sont observées, indiquent ensuite quelles seront les protéines qui seront potentiellement plus efficientes dans la réponse. Enfin, en réponse au stress environnemental, les études transcriptomiques peuvent mettre en évidence de nouveaux variants d'épissage.

c) *Apport de la protéomique pour l'écotoxicologie*

En écotoxicologie, les études protéomiques s'intéressent principalement à la détection de sur ou sous production des protéines d'organismes exposés aux contaminants en comparaison avec des organismes de références et non exposés. En plus de l'analyse par protéomique différentielle, la protéomique permet l'analyse des modifications post-traductionnelles qui indique les voies de signalisation activées ou non par l'exposition de l'organisme aux contaminants. Par exemple, l'étude conjointe de la modulation de l'expression des protéines et du phosphoprotéome entier montre en détail quelles sont les protéines impliquées dans les divers processus cellulaires perturbés, reflétant les effets toxiques des produits contaminants. Associée à la génomique ou à la transcriptomique, la protéomique détecte les peptides marqueurs de mutations détectées et permettent d'affirmer ou d'infirmer la présence de ces variants (Gouveia et al. 2019). Un des principaux intérêts de l'étude protéomique est l'observation de la molécule effective de la réponse au polluant. Cette observation permet non seulement de voir l'effet sur l'abondance des protéines mais aussi par l'étude des modifications post-traductionnelles d'avoir une information sur leur état d'activation par exemple. De plus, en protéomique l'observation est faite après l'ensemble des régulations d'expressions et traduction ce qui permet de voir l'état réel de l'expression des différents gènes. Enfin, par protéomique ciblée il est possible d'obtenir une quantification quasi absolue et précise de la quantité d'une protéine donnée et donc d'atteindre un niveau de précision dans le suivi de l'expression des protéines, nécessaires, pour l'exploitation de biomarqueurs en écotoxicologie.

3) Les organismes non modèles : Une nouvelle source d'information

a) *Intérêt des organismes non modèles*

A l'origine le concept même de l'écotoxicologie repose sur l'étude la plus complète possible d'un environnement donné. Cependant, parmi l'ensemble des organismes vivants il existe, même aujourd'hui, qu'une faible portion dont le génome est séquencé rendant ainsi l'analyse des mécanismes moléculaires internes impossible. Pour autant, il est essentiel de comprendre l'impact sur l'ensemble d'un écosystème des molécules rejetées dans l'environnement. L'analyse d'une grande diversité de population peut révéler les mécanismes moléculaires qui permettraient de découvrir de nouveaux mécanismes de réponse aux différents polluants (Armengaud et al. 2014).

Et pourtant, Il est essentiel d'utiliser des organismes non modèles pour analyser directement dans l'environnement un maximum de paramètres tels que des biomarqueurs. Ces derniers permettront de mettre en place des indicateurs de qualité d'un milieu dont la confiance dépend du nombre de marqueurs et d'espèces utilisés.

b) *La solution protéogénomique pour obtenir rapidement des données moléculaires*

Afin de pouvoir analyser les organismes non modèles et comprendre les mécanismes moléculaires induits par l'exposition aux polluants, il est nécessaire d'analyser leur protéome. Cependant la protéomique classique ne permet que partiellement l'exploration des protéomes d'organismes ne possédant pas de base de données de référence. La protéomique informée par transcriptomique apparait alors comme une solution à l'utilisation d'organismes non modèles en écotoxicologie (Armengaud et al. 2014). En effet les séquençages des ARNm de ces organismes méconnus permettent alors de générer les premières bases de données de transcrits. Par similarité avec des organismes modèles ces derniers peuvent être annotés afin d'avoir une orientation sur les fonctions potentielles des transcrits. Puis une traduction de ces transcrits permet de constituer une première base de

données protéomique qui peut alors servir à l'interprétation des protéomes de ces organismes dans des conditions standards ou exposées aux polluants.

IV-Contexte de l'équipe de recherche et enjeux de la thèse

1) Historique des travaux protéogénomique de l'équipe d'accueil

a) Utilisation de la protéogénomique pour la correction de l'annotation

L'équipe du Dr Jean Armengaud (Institut Joliot/DRF/LI2D-CEA Marcoule), directeur de cette thèse, est spécialisée en protéomique et plus particulièrement en protéogénomique. En effet, cette équipe a fait partie des pionniers sur les travaux en protéogénomique pour la correction de génome bactérien. Notamment en 2009 de Groot et al. s'illustrent en utilisant la protéogénomique afin de produire l'annotation de *Deinococcus deserti*, première étude au niveau mondial présentant le génome et le protéome d'un organisme vivant en même temps. Cette étude a permis la validation et détection de la moitié des gènes annotés ainsi que la détection de 15 gènes non prédits et la réorientation de 11 gènes prédits. S'ensuit une seconde publication de Jean Armengaud en 2009 se proposant de poser les bases de la correction de l'annotation aidée par la protéomique. Enfin, en 2010 avec les travaux de Baudet et al., l'équipe s'illustre une nouvelle fois avec la correction de 73 codons d'initiations de traduction pour *Deinococcus deserti,* et la découverte étonnante que parmi ces codons d'initiations certains sont extrêmement rares mais possibles chez les bactéries. Enfin en 2013, de nouveaux travaux de l'équipe mettent en évidence par protéogénomique la présence de plusieurs codons d'initiations de la traduction pour plusieurs gènes chez une bactérie marine (Bland et al., 2014). Ces travaux qui portent sur des développements méthodologiques pour identifier systématiquement les Nter des protéines ont permis la soutenance de la thèse de Céline Bland dans le domaine de la protéogénomique.

2) Historique des travaux d'ecotoxicologie des collaborateurs

a) Etudes de Gammarus fossarum en tant que sentinelle en ecotoxicologie

L'équipe du Dr Olivier Geffard, maintenant dirigée par Dr Arnaud Chaumot, (IRSTEA) est spécialisée dans les études écotoxicologiques de terrains. Les travaux de cette équipe ont notamment permis de mettre en place le contrôle de la qualité de l'eau en utilisant pour sentinelle le Gammare. L'équipe s'est illustrée dans l'élaboration de traits de vie du Gammare permettant d'identifier la pollution au sein des eaux douces (Dedourge-Geffard et al. 2009, Lacaze et al. 2011). Ces traits de vie concernent notamment la reproduction et l'évaluation de la reproduction et chez la femelle, des paramètres supplémentaires sont mesurés tels que la mue. C'est en 2009 que l'équipe définit comme biomarqueurs l'acetylcholinesterase qui entraine, lors de l'altération de son activité, une modification comportementale du Gammare (Xuereb et al. 2009). L'expérience de l'équipe dans la connaissance des paramètres physiologiques impactés en présence de polluants au sein du gammare a permis la création d'une entreprise nommée BIOMAE. Cette société assure aujourd'hui l'encagement de gammare et contrôle la qualité des eaux en utilisant comme paramètres évaluateurs les traits de vie définis par l'équipe.

C'est en collaboration entre cette équipe spécialiste de la physiologie du Gammare et l'équipe de Jean Armengaud précurseur dans les travaux protéogénomique que des travaux de caractérisation moléculaire avancés sur le Gammare ont été initiés. Ces travaux ont été effectués par le biais de la thèse de Judith Trapp puis de Duarte Gouveia sous la co-direction d'Olivier Geffard et Jean Armengaud.

3) Les travaux protéogénomiques sur le Gammare : bientôt 10 ans

a) Développement de biomarqueurs pour la surveillance en écotoxicologie chez Gammarus fossarum:

La thèse de Judith Trapp soutenue en 2014 a initié les premiers travaux protéogénomiques sur *Gammarus fossarum*. Durant cette thèse, une première base de données spécifique du Gammare a été générée par le séquençage des ARNs en utilisant des organes reproducteurs de mâles et femelles, des céphalons et l'ampoule hépatopancréatique de multiples Gammares (>200). Ce séquençage a permis de produire une base de données de plus de 200,000 séquences contiguës après séquençage produisant plus d'un million d'ORF potentiels après traduction en 6 cadres de lecture. Cette base de données nommée GFOSS constitue la première base de données spécifique de *Gammarus fossarum*. En parallèle 192 échantillons ont été analysés en spectrométrie de masse provenant du fractionnement des protéomes extraits des cephalons et des appareils reproducteurs males et femelle de gammares. C'est plus d'un million de spectres qui sont ensuite exploités durant la thèse de Judith Trapp. Une première analyse a permis de mettre en évidence 43,505 séquences peptidiques différentes regroupées en 1873 protéines ainsi validées par l'approche protéogénomique (Trapp et al. 2014). Parmi celles-ci 218 ont été définies comme spécifiques de *Gammarus fossarum*.

Suite à ces premiers résultats, une première étude a permis la sélection de séquences peptidiques biomarqueurs (Trapp et al. 2014). Judith s'est ensuite plus particulièrement focalisée sur l'étude des modulations de l'expression de protéines au sein du système reproductif en réponse à l'environnement (Trapp et al. 2015, Trapp et al. 2016, Trapp et al. 2016). C'est une liste finale de 177 peptides potentiels biomarqueurs représentant 55 protéines qui a pu être établie. Ces protéines sont réparties dans 4 fonctions biologiques clefs que sont la mue et la régulation hormonale ; les biomarqueurs généraux ; les biomarqueurs spécifiques du sexe et enfin les biomarqueurs d'immunité.

Une autre étude clef menée par Judith concerne l'impact de l'utilisation d'une base spécifique de l'espèce pour l'interprétation des données protéomiques (Trapp et al. 2016, Figure 5). Judith montre notamment dans cette étude que l'attribution des spectres provenant des protéines extraites de *Gammarus pulex* ont une attribution de seulement 10% contre 20% lorsqu'il s'agit de l'attribution de protéines de Gammarus fossarum en utilisant la base de données GFOSS. Il est alors mis en évidence le risque au sein de l'environnement d'avoir des biais en fonction de l'espèce dans l'utilisation des biomarqueurs potentiels.

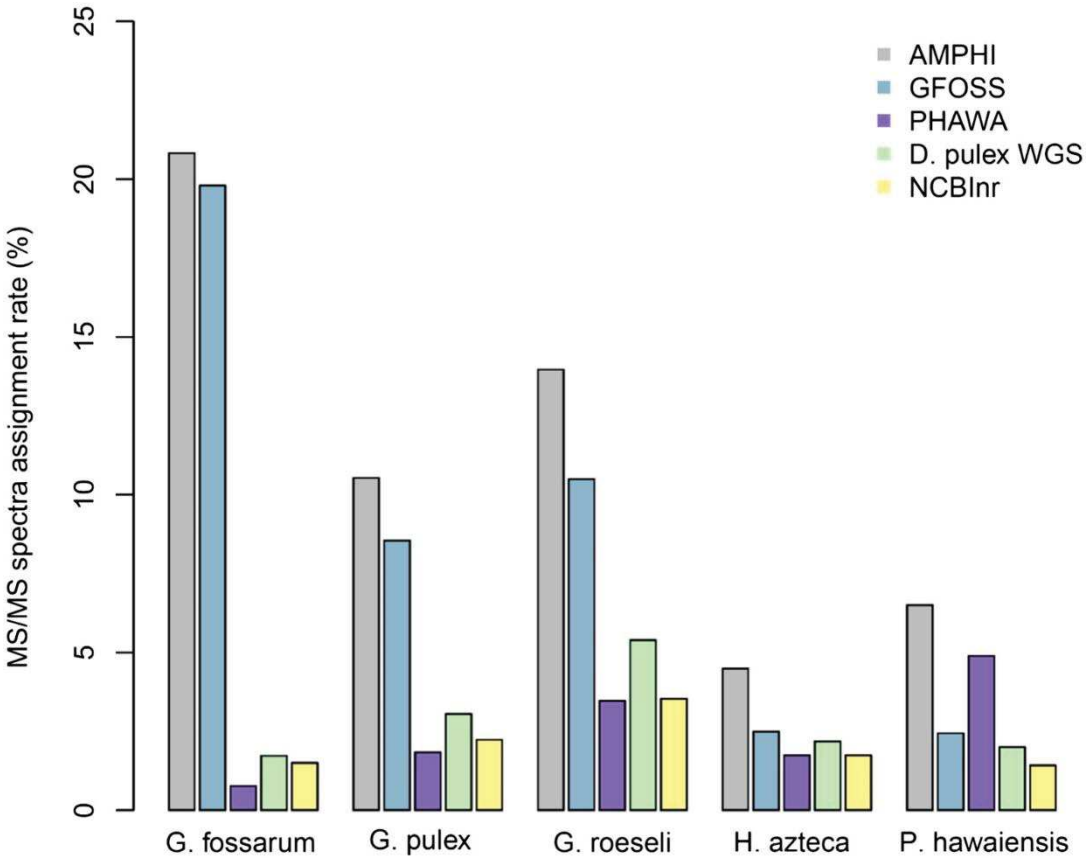


Figure 5 : Les données de spéctrométrie de masse générées pour *Gammarus fossarum, Gammarus pulex, Gammarus roeseli, Hyallela azteca* et *Parahyale hawaiensis* ont été interprété sur 5 bases de données différentes. NCBInr correspond à l'ensemble des données provenant de NCBInr, PHAWA au transcriptome de *Hyallela azteca*, *Daphnia pulex* WGS au génome séquencé de *Daphnia pulex*, GFOSS au transcriptome assemblé de *Gammarus fossarum* et AMPHI à la concaténation des 3 dernière bases de données. Figure issue de Trapp et al. 2016.

b) Sélection des biomarqueurs pertinents pour l'évaluation de l'environnement chez Gammarus fossarum

La thèse de Duarte Gouveia soutenue en 2017 a permis la validation et la mise au point d'une méthode analytique pour doser spécifiquement les biomarqueurs mis en évidence. En effet, à partir des 177 peptides biomarqueurs établis suite aux travaux de Judith, Duarte a réalisé des étapes de validation d'une part analytique puis physiologique de ceux-ci. La validation analytique a consisté à établir les limites de détection de ces peptides biomarqueurs potentiels par une méthodologie de protéomique quantitative basée sur la méthodologie « Selecter Reaction Monitoring » (SRM) (Charnot et al. 2017).

Le suivi dans le cycle de vie de ces biomarqueurs validés analytiquement a permis ensuite de les valider physiologiquement au sein de *Gammarus fossarum* comme les biomarqueurs pertinents pour le suivi en écotoxicologie (Gouveia et al. 2017). Suite à ces étapes de validation, 38 peptides biomarqueurs ont pu être certifiés, représentant 25 protéines. Ces biomarqueurs ont ensuite été appliqués sur l'étude terrain de 17 sites dont 4 dits de références (sites considérés propres) et 13 contaminés. Cette étude a permis la validation *in situ* de l'utilisation des biomarqueurs et propose d'utiliser des diagrammes indicateurs de santé pour chaque site à partir du dosage des biomarqueurs (Gouveia et al. 2017).

Afin de mettre en avant l'ensemble de cette méthodologie de découverte de biomarqueurs, de validation, puis d'utilisation pour des échantillons de terrain, Gouveia et al. proposent une revue en « écotoxicoprotéomique » (Gouveia et al. 2019).

4) Enjeux de la thèse dans l'amélioration des études en écotoxicologie

a) Le projet ANR PROTEOGAM

Ma thèse se déroule dans le cadre du projet ANR (ANR-14-CE21-0006) « Protéomique pour de nouveaux biomarqueurs en écotoxicologie chez les gammares : challenge de la biodiversité et immunoanalyse multiplexée comme outil de diagnostic ». Ce projet ANR est en collaboration entre l'IRSTEA Villeurbanne, le CEA Marcoule et l'ICN situé à Nice. Les principaux acteurs de ce projet sont Olivier Geffard (porteur du projet), Arnaud Chaumot, et Davide Degli-Esposti du laboratoire d'écotoxicologie pour l'IRSTEA, Stephane Azoulay et Cecile Becquart de l'équipe molécules bioactives pour l'ICN, et Jean Armengaud, Christine Almunia, Olivier Pible du laboratoire «Innovative technologies for Detection and Diagnostics» pour le CEA, ainsi que Duarte Gouveia de l'IRSTEA puis du CEA. L'équipe de Stephane Azoulay est chargée dans ce projet d'établir des méthodologies analytiques de suivi des biomarqueurs par immunologie.

Le projet ANR Proteogam se propose de définir des marqueurs moléculaires au sein des gammares pour sonder la qualité biologique de l'eau douce à l'aide d'un système de surveillance actif basé sur l'encagement des organismes dans l'environnement. Suite aux travaux précédemment réalisés entre l'IRSTEA et le CEA, une première base de données issue du séquençage de *Gammarus fossarum* a été générée (GFOSS). Cependant une grande diversité d'espèces est observée au sein du genre Gammare à travers toute l'Europe. Les travaux de Judith Trapp (Trapp et al. 2016) démontrant que l'utilisation d'une base de données non spécifique de l'espèce induit une baisse du taux d'attribution des spectres, il est donc nécessaire de travailler sur la transversalité des biomarqueurs entre les espèces. De ce fait, l'application des biomarqueurs en conditions environnementales en Europe, implique la prise en compte et l'étude de la variabilité intra et inter populations au sein des espèces de Gammares les plus répandues en Europe. C'est donc l'objectif de mon projet de thèse.

b) Enjeux et moyen mise en œuvre pour la thèse

Afin de prendre en compte la variabilité inter et intra espèces présente dans l'environnement il a été décidé de générer des données multi-omiques pour 7 groupes de Gammaridés. Les sept groupes d'animaux sont répartis de façon à être représentatif des espèces les plus courantes, rencontrées en Europe. Pour intégrer l'échelle de diversité la plus large possible, les groupes choisis sont les suivants dont la phylogénie est représentée Figure 6:

- 3 espèces cryptiques *Gammarus fossarum A, B et C*
- *Gammarus wautieri*, *proche* phylogénétiquement de *Gammarus fossarum*
- *Gammarus pulex,* plus éloignée phylogénétiquement de *Gammarus fossarum*
- *Echinogammarus berilloni*, gammare d'eau douce et *Echinogammarus marinus* aussi nommé *Gammarus marinus*, gammare d'eau salé. Il s'agit d'un autre genre que le genre Gammarus, plus éloigné phylogénétiquement de *Gammarus fossarum*.

Pour chacun de ces groupes le séquençage du transcriptome d'un male et d'une femelle a été obtenu avec succès (14 transcriptomes). Pour chacun de ces groupes l'analyse en spectrométrie de masse de 10 individus par groupes et sexe était initialement prévue (140 protéomes). Pour des raisons

stratégiques durant la thèse ces répartitions ont été modifiées du fait de la disponibilité de chaque espèce de gammare prélevée et sont indiquées dans la Table 1 suivante :

Tableau 1 : Répartition du nombre d'analyse de spectrométrie de masse par espèce :

Espèces	Sexe	Nombre de run MS/MS
Gammarus fossarum B	Femelle	30
Gammarus fossarum B	Male	30
Gammarus pulex	Femelle	20
Gammarus pulex	Male	20
Echinogammarus berilloni	Femelle	10
Echinogammarus berilloni	Male	10
Echinogammarus marinus	Femelle	10
Echinogammarus marinus	Male	10
Gammarus wautieri	Femelle	3
Gammarus wautieri	Male	3
Gammarus fossarum A	Femelle	7
Gammarus fossarum A	Male	7
Gammarus fossarum C	Femelle	2
Gammarus fossarum C	Male	2

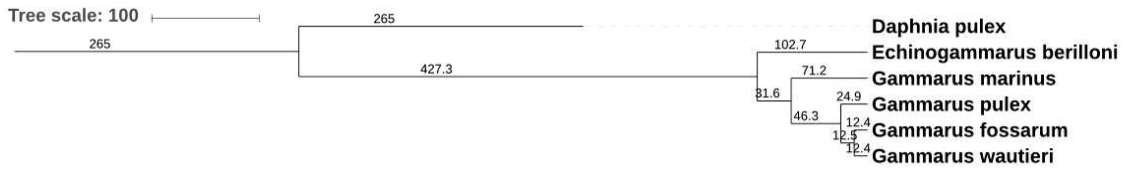


Figure 6 : Arbre phylogénétique des différents groupes d'animaux utilisés au sein du projet PROTEOGAM avec pour branche externe *Daphnia pulex.* Arbre basé sur l'alignement multiple de la catalase par clustalW et arbre généré par Phyml après filtration de l'alignement par GBLOCKS.

c) Objectifs de la thèse

Cette thèse s'inscrit au sein du projet ANR Protéogam ayant pour objectif final la mise en place de biomarqueurs pour le suivis de la qualité des eaux en Europe par une méthode de biosurveillance active basée sur l'utilisation de gammare encagés. Afin de réaliser ce suivi les études en écotoxicologie menées par Judith Trapp et Duarte Gouveia ont permis de sélectionner des peptides biomarqueurs exploitables pour leur l'utilisation avec des Gammare d'une espèce contrôle *Gammarus fossarum B*,

encagés, dont le stade de développement est synchronisé. Au vu de la forte diversité d'espèces au sein des Gammares et de la difficulté de les différencier par des caractères morphologiques il est essentiel d'évaluer la transversalité des biomarqueurs potentiels précédemment définis. Afin de répondre à cette nécessité le transcriptome de chacune de ces espèces doit être déterminé afin d'interpréter par la suite les données protéomique. Cette réalisation correspond donc aux deux premiers objectifs de ma thèse :

i) Optimisation de la méthodologie d'assemblage des transcriptomes et de génération de la base de données protéogénomique avec l'objectif d'améliorer l'interprétation protéogénomique. Voir Chapitre I.
ii) Mettre à disposition de la communauté écotoxicoprotéomique les assemblages des transcriptomes des 7 espèces sélectionnées. Voir Chapitre II.

Une fois la base de données protéogénomiques de chaque espèce obtenue, il est possible de valider la présence des séquences de biomarqueurs définie pour *Gammarus fossarum B*. Cette réalisation a conduit au troisième objectif de la thèse :

iii) Mise au point d'une méthodologie d'étude de la variabilité inter-espèce au sein des Gammares avec pour cible les séquences peptidiques biomarqueurs potentiels précédemment définis pour *Gammarus fossarum B*. Voir Chapitre III.

Enfin le projet a eu pour ambition de générer l'analyse protéomique individuelle pour l'ensemble des échantillons. La production de tels résultats permet de mettre au point des analyses au niveau individuel au sein de populations de Gammare. L'exploitation des protéomes individuels des 40 échantillons de *Gammarus pulex* répartis entre deux rivières (Référence vs polluée) nécessite la mise en place de nouvelles méthodes d'analyses de données, individu-centrée, en protéomique. Cette réalisation correspond au dernier objectif de ma thèse :

iv) Mise au point d'une méthodologie d'analyse individu-centrée, permettant d'observer la variabilité intra-population au sein d'une analyse *in situ* de l'impact environnemental sur le vivant. Voir Chapitre IV.

Chapitre 1: Optimisation de la méthodologie d'assemblage de transcriptomes et de traduction pour générer une base de données protéogénomique

Pour la protéogénomique d'organismes non modèles, il est essentiel d'avoir une base de données la plus adéquate possible afin de maximiser les découvertes de peptides ou de protéines d'intérêt. Il faut qu'elle soit exhaustive en terme de séquences mais réduite en terme de taille. A cet effet, optimiser les étapes d'assemblages et de traduction permettant la génération de la base de données utilisée pour l'interprétation peut présenter un gain en terme de résultats ou de temps de calcul. Les données transcriptomiques possèdent de nombreux biais connus à ce jour qui sont dans la plupart des cas systématiquement traités par des paramétrages par défaut ou des *a priori* méthodologiques.

A ce jour, il n'existe que peu d'informations sur l'impact réel final de l'utilisation systématique de ces méthodes sur la qualité de l'assemblage. De plus, les critères d'évaluation des assemblages reposant sur des méthodes d'évaluation principalement basées sur la taille ou encore le nombre des séquences générées ne sont pas suffisamment fonctionnellement informatives sur l'assemblage. Les autres méthodologies reposant sur le principe d'utilisation de gènes orthologues ne permettent pas non plus une évaluation de l'ensemble du transcriptome et se concentre sur des séquences relativement conservées et potentiellement plus exprimées, et donc plus simple à assembler.

Dans ce chapitre, nous utilisons un critère d'évaluation de la qualité de l'assemblage reposant sur le taux d'interprétation des spectres MS/MS acquis en protéomique shotgun et interprétés en interrogeant les bases de données de séquences polypeptidiques dérivées de ces assemblages. L'impact de la taille de la base de données est un élément essentiel dans le contrôle du taux de faux positif. Pour en tenir compte, nous proposons une solution de réduction du nombre de séquences polypeptidiques en intégrant une méthodologie automatique de recherche de cadre ouvert de lecture. Afin d'optimiser les paramètres de cette méthodologie nous avons utilisé le taux d'interprétation des spectres MS/MS acquis en protéomique shotgun. Le manuscrit a été soumis au journal Proteomics.

Proteogenomics-guided evaluation of RNAseq assembly and protein database construction for non-model organisms

Yannick Cogne[1], Duarte Gouveia[1], Arnaud Chaumot[2], Davide Degli-Esposti[2], Olivier Geffard[2], Olivier Pible[1], Christine Almunia[1], Jean Armengaud[1#]

[1]Laboratoire Innovations technologiques pour la Détection et le Diagnostic (Li2D), Service de Pharmacologie et Immunoanalyse (SPI), CEA, INRA, F-30207 Bagnols-sur-Cèze, France.

[2]Irstea, UR RiverLY Laboratoire d'écotoxicologie, centre de Lyon-Villeurbanne, F-69625 Villeurbanne, France.

[#]**Corresponding author**

Jean Armengaud, CEA-Marcoule, DRF-Li2D, Laboratory "Innovative technologies for Detection and Diagnostics", BP 17171, F-30200 Bagnols-sur-Cèze, France; jean.armengaud@cea.fr; Tel: +00 33 4 66 79 68 02; Fax: +00 33 4 66 79 19 05.

Word count: **3,982; Characters count: 22,561; Running title**: **Proteogenomics-guided evaluation**

ABSTRACT

Proteogenomics gains momentum as genomics, transcriptomics and proteomics can be today easily performed on any new species. This approach allows defining key molecular players when comparing contrasted conditions. For animals and plants, RNAseq-informed proteomics is the most popular approach but requires the most performant *de novo* RNAseq assembly and optimized translation strategy. Here, several options for pre-treating RNAseq Illumina reads before assembly and for translating the resulting contigs into useful polypeptide sequences have been explored. Based on transcriptomics and proteomics experimental datasets acquired on individual animals belonging to the freshwater crustacean *Gammarus fossarum*, the ratio of MS/MS spectra assigned to peptide sequences helps to define the most relevant procedure. Removing reads with mean quality score below 17 which represents a single probable nucleotide error on reads with length of 150 bp prior assembly increases the proteomics outcome. The best translation option using Transdecoder is with a minimal ORF length of 50 amino acids and systematic selection of ORFs above 300 amino acids. Transcriptome assembly and translation informed by proteomics pave the way to further improvement in proteogenomics.

Keywords: assembly, *de novo*, non-model organisms, proteogenomics, proteomics, quality metrics, RNAseq, transcriptomics.

INTRODUCTION

Despite their high relevance in ecology and ecotoxicology, many animals from the environment cannot be cultivated and bred in laboratory conditions. No genetic information or tool to engineer them are available, thus limiting scientific opportunities. In the case of arthropods which include insects, arachnids, myriapods, and crustaceans, and account for a large majority of known living animal species, only a small proportion of species have been genome sequenced [1, 2].

Transcriptomics and proteomics are two interesting tools to describe the dynamics of the molecular players of any organism even if a reference genome sequence and annotation are missing. These omics approaches are highly complementary, as shown by the so-called "proteomics informed by transcriptomics" strategy. Indeed, proteogenomics gains momentum, especially for non-model organisms [3]. Transcript reads generated by next-generation sequencing are assembled by *de novo* approaches and used to create a protein sequence database that is crucial for interpreting MS/MS spectra acquired by tandem mass spectrometry on proteins. This alliance was successful for characterizing various animals, such as *Apis mellifera* [4], *G. fossarum* [5], and *Plutella xylostella* [6]. This proteogenomics strategy is useful for improving knowledge in molecular physiology and discovering biomarkers in the field of ecotoxicology [7, 8]. In such context, the success regarding peptide assignation and protein identification resides in the quality of the transcriptome assembly and creation of the most appropriate protein sequence database.

RNA-seq data generated by Illumina should be filtered and treated before any assembly to avoid bias due to parasite sequences or low-quality sequences. For example, oligonucleotide adapter sequences inserted by ligation with RNA fragments for library preparation may be found in several sequencing reads due to variable insert sizes. The lower confidence of nucleotide calling at the 3' end of reads can be circumvented by removing these extremities to keep only high quality nucleotide sequences. This trimming has been shown to be highly beneficial if done after adaptor removal and its parameters impact the quality of subsequent analysis [9-12]. Another main treatment consists in removing rRNA reads which can be retrieve despite the use of polyA selection of RNAs or other rRNA depletion methods prior library construction [13]. Guidelines have been proposed for performing these pre-treatments of raw Illumina data, but only few studies have been reported for evaluating the impact of these treatments mostly on model organism [11, 12, 14] and the discriminative parameters used to benchmark them may be subjected to discussion.

When dealing with organisms for which no sequenced genome is available, transcriptome assembly should be performed by *de novo* approach. Numerous tools for such *de novo* assembly have been proposed [15]. For transcriptome assembly, several studies benchmarked RNA-Seq assemblers [15-20]. Surprisingly, no specific tool delivered the best results for all datasets when comparing datasets from different kingdoms of life [15]. For this comparison, standard metrics such as the size of assembled contigs or the ratio and size of conserved orthologous genes were used [21, 22]. Interestingly, a new metric was recently proposed to check the extent of correct transcriptome coverage by means of the proteogenomics concept. Proteomics data in the form of tandem mass spectrometry spectra acquired on peptides from the same organism can be measured and then interpreted with the translated *de novo* assembled transcriptome. The ratio of MS/MS spectra assigned to peptide sequences has been shown an interesting criterion to benchmark transcriptome assembly tools. With such approach, the *de novo* assemblers

MAPS [23] and Trinity [24] were favourably evaluated.

For optimal interpretation in proteogenomics, the size of the 3 frame-translated transcriptome databases should not become too abnormally large as the attribution rate is strongly dependent on the size of the database [25, 26]. Better evaluation of false positives with such inflated databases is to be considered [27, 28] and cascade searches can be implemented to focus on the most probable polypeptide sequences [29]. Another interesting alternative is to reduce the database size by selecting the most probable polypeptide sequences using ORF length as criterion [30] or based on learning machine search engine such as Transdecoder [26, 31]. However, to the best of our knowledge, no comparative study has been reported on the impact of such database size reduction and the correlated loss of information in proteomics informed by transcriptomics for non-model organisms [32].

In the framework of a large proteogenomics project devoted at defining biomarkers from a sentinel animal useful in ecotoxicology, the freshwater *G. fossarum* [5, 33], we obtained both high quality RNAseq and shotgun proteomics datasets on individual organisms. No reference genome sequence is yet available for this organism used for active biomonitoring of aquatic environments. Here, we explored with these datasets how filtering performed before *de novo* assembly with Trinity, a popular *de novo* transcriptome assembly tool based on *de Bruijn* graphs, could trigger a potential loss of useful information. For this, we combined standard and proteogenomics-derived metrics for evaluating the quality of *de novo* transcriptome assembly following multiple pre-treatments of Illumina raw data. Optimal parameters for translating the improved transcriptome were evaluated for maximizing the proteomics output.

Materials & Methods

Sequencing files, SRA accession

RNAseq on *G. fossarum* female was previously performed and raw files are available from NBCI with the accession SRR8089727 [34].

Read pre-treatments

Removing adaptators and/or trimming were done with TrimGalore! v0.4.2 (Babraham Bioinformatics) using the Galaxy wrapper from toolshed bgruening repository. All parameters were set as default except the quality threshold (Phred score: Q) for trimming which was set at 5, 10, 15 or 20. Removal of rRNA sequences was done with Bowtie2 v2.2.6.2 [35] using the Galaxy wrapper from toolshed devteam repository and SILVA (SILVA Release 132, [36]) or NCBI rRNA (2018-03 download) as database with default parameters. Sequencing reads with N or with mean quality score below 16.9897 or 21.7609 were filtered with homemade script available at https://github.com/YannickCogne/Qfiltering. In all cases unpaired reads were deleted. Systematic removal of 12 nucleotides at the start of each read was done using the trim function on Galaxy with default parameters considering Fastq files option.

Transcriptome assembly

After pre-treatment, reads from the two different lanes produced by sequencing were merged for each pair-end part. Trinity v2.4 [31] was used to assemble reads considering pair-end and strand orientation while other parameters were set at default values with a k value fixed at 25 and a minimum contig length of 200 bp.

Translation of transcriptome for building proteogenomic databases

First, a systematic 3 frame translation of each contig was done from stop to stop codons with a homemade script (https://github.com/YannickCogne/3Frame_Translate). Only polypeptide sequences with length above 7 amino acids were kept. For the best assembly (GFAF07) as defined in our study a translation was performed using an in-house modified version of Transdecoder v3.0.1 [31] allowing start of an ORF without ATG initiator codon (Stop to Stop ORF). Parameters for the minimal length of the ORF to consider for the Hidden Markov Model recognition of the Transdecoder algorithm after selection of the top 500 ORFs as training dataset, and the minimal length of ORF to systematically take into account were tested while all the other parameters were set as default. For evaluating false positives with a decoy search, each protein database was merged with its own decoy built by decoyPyrat [37] with default parameters.

Sampling, protein extraction and mass spectrometry

Three female individuals from a population of *G. fossarum* were sampled from the Seebach river in north-eastern France as previously described [34]. Back to the laboratory, developing embryos were removed from the marsupial brood pouch just before females were quickly frozen in liquid nitrogen. Proteins from each animal were extracted and analysed by shotgun proteomics in conditions similar as those previously described [38]. Briefly, proteins were extracted by bead-beating in LDS sample buffer (Invitrogen) and subjected to SDS-PAGE for a short electrophoretic migration. The whole-protein content from each well was extracted as a single polyacrylamide band, processed as described for trypsin proteolysis [39]. The resulting peptides were analysed in data-dependent mode with a Q-Exactive HF mass spectrometer (Thermo) operated as described [40].

MASCOT searches

MS/MS spectra were assigned to peptide sequences with the MASCOT Daemon 2.3.2 search engine (Matrix Science). Peptide mass tolerance was set at 5 ppm; fragment mass tolerance was set at 0.02 Da, and missed cleavages were allowed until a maximum of 2. Carbamidomethylation of cysteines was searched as a fixed modification. Deamination of Asparagine or Glutamine and oxidation of methionine were set as variable modifications. Peptide-to-spectrum matches were extracted with the Perl MASCOT script fdr_stats.pl with specific p value threshold defined for a given FDR threshold value with the corresponding decoyPyrat [37] search.

Evaluation of search strategy yields

Quality assessment of transcriptomes was done with Transrate v1.0.1 [22] which generates standard metrics. No reference protein sequences were used for the evaluation with Transrate. The validation of the assembly quality of detectable orthologs in all the assemblies was done with BUSCO v2.0 [21]. The database used for BUSCO analyses was Arthropoda_odb9 which contains 1066 orthologous gene at the nearest taxon level available for *Gammarus*. Three criteria were used for proteomics evaluation of databases. The first one was the number of different peptides identified for a given sample. The second one was the number of spectra identified for a given sample. The last one was the ratio R* between the sum of the number of new peptides brought by an assembly by comparing it to each of the other assemblies and the sum of the number of peptides lost by this assembly relatively to each of the other assemblies. For a given assembly *i* and all assemblies numerated from 1 for GFAF01 to 13 for GFAF13:

$$R_i^* = \frac{\sum_{j=13}^{j=1} new\ Peptides\ in\ assembly\ i\ compared\ to\ assembly\ j}{\sum_{j=13}^{j=1} new\ Peptides\ in\ assembly\ j\ compared\ to\ assembly\ i}$$

This R*ratio was calculated for each of the thirteen assemblies and was used as criterion to rank the thirteen assemblies. The higher R* indicates the assembly consensus with the higher number of new peptides and the lower number of lost peptides.

Mass spectrometry and proteomics data

The mass spectrometry and proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE [41] partner repository with the dataset identifiers PXD014166 and DOI 10.6019/PXD014166. [The dataset is available with the Username: reviewer26888@ebi.ac.uk and Password: K6HHdssc].

RESULTS AND DISCUSSION

Experimental datasets acquired on *Gammarus fossarum* individuals

Figure 1 shows the general workflow for a classical proteogenomics study. Here, four *G. fossarum* females were sampled from the Seebach river in north-eastern France. One animal (S0) was previously genotyped by amplifying and sequencing its Cytochrome c oxidase subunit I (COI) gene from DNA extracted from a single leg [34]. Total mRNAs were extracted from this animal, polyA mRNAs were then purified and Illumina sequenced using HIseq3000, resulting in 40,979,915 reads of 150 bp for each element of the pair. After assembly, the transcriptome is translated to obtain a theoretical protein database. As shown in Figure 1, proteins were extracted from the three other animals and each individual batch was subjected to shotgun proteome analysis. For each animal, the proteomics data consisted in 43,457 (animal S1), 40,432 (animal S2) and 41,099 (animal S3) MS/MS high-resolution spectra. These MS/MS spectra were interpreted into peptide sequences against the theoretical protein database derived from the transcriptome, resulting in a list of mass spectrometry-certified proteins and their abundances estimated by spectral counts.

RNAseq read preprocessing and assembling

Several strategies and parameters for selecting RNAseq reads prior assembling them were compared. **Table 1** lists the different options chosen according to classical practice for circumventing several well-known types of error detected in Illumina generated data. A first route consisted in keeping all the available reads whatever their quality. This control assembly was named GFAF01. A second assembly procedure follows the Harvard protocol (https://informatics.fas.harvard.edu/best-practices-for-de-novo-transcriptome-assembly-with-trinity.html) recommending removal of Illumina sequencing adapters, trimming the low-quality reads at Q score lower than 5, and filtering potential rRNA gene reads deposited in the SILVA database. In this treatment named GFA02 we noted that no reads were trimmed at such low threshold. Therefore, we also performed several assembly procedures relying on adapter removal and trimming at different Q values for exploring this parameter. GFAF03, GFAF04, GFAF05, and GFAF06 were done for thresholds of low-quality base of reads fixed at Q below 5, 10, 15, and 20, respectively. As trimming at Q below 5 (GFAF03) has no effect on the data, this strategy is equivalent to a procedure where only adapters are removed. The procedures GFAF07 and GFAF08 were done after removing reads with mean Q score below 16.9897 and 21.7609, respectively, and filtering missing unpaired reads. These last two quality thresholds were chosen according a mean Q score representing 1 or 3 probable nucleotide

error on reads with length of 150 bp. Assembly GFAF09 is done after removal of reads containing at least one N base and filtering missing unpaired reads after this treatment. Assembly GFAF10 consisted in systematic removal of 12 nucleotides at the start of each read. Assemblies GFAF11 and GFAF12 were done removing mapped reads on SILVA rRNA database, and SILVA rRNA and NCBI rRNA databases, respectively. GFAF13 was done combining GFAF07 and GFAF09 treatments. Proportions of reads and nucleotides kept after each treatment are shown in **Table 1**. Removing the 12 first nucleotides of each read (GFAF10) induced the highest loss of information with 8% of nucleotides removed. Trimming at Q score of 20 (GFAF06) or removing reads with mean Q score below 21.7609 (GFAF08) are also quite stringent options that resulted in important information losses with 2.3% and 3.2%, respectively. Other thresholds induced a loss of less than 1% of information.

Table 1. Strategies of read pre-treatment and resulting databases.

Database	Pre-treatment	Reads (%)	Nucleotides (%)
GFAF01	**Nothing**	**100.00**	**100.00**
GFAF02	**Remove adaptateur+ Trimming (Q>5) + Remove ARNr (SILVA)**	**99.98**	**99.17**
GFAF03	**Remove adaptateur+ Trimming (Q<5)**	**100.00**	**99.19**
GFAF04	**Remove adaptateur+ Trimming (Q<10)**	**100.00**	**99.15**
GFAF05	**Remove adaptateur+ Trimming (Q<15)**	**99.99**	**98.89**
GFAF06	**Remove adaptateur+ Trimming (Q<20)**	**99.35**	**97.71**
GFAF07	**Remove reads Qmoyen < 16.9897**	**99.49**	**99.49**
GFAF08	**Remove reads Qmoyen < 21.7609**	**96.76**	**96.76**
GFAF09	**Remove reads with at least one N**	**99.88**	**99.88**
GFAF10	**Remove 12 nucleotides at the start of reads**	**100.00**	**92.00**
GFAF11	**Remove ARNr (SILVA)**	**99.98**	**99.98**
GFAF12	**Remove ARNr (SILVA) & Remove ARNr (NCBI)**	**99.85**	**99.85**
GFAF13	**Remove reads Qmoyen < 16.9897 & remove reads with at least one N**	**99.38**	**99.38**

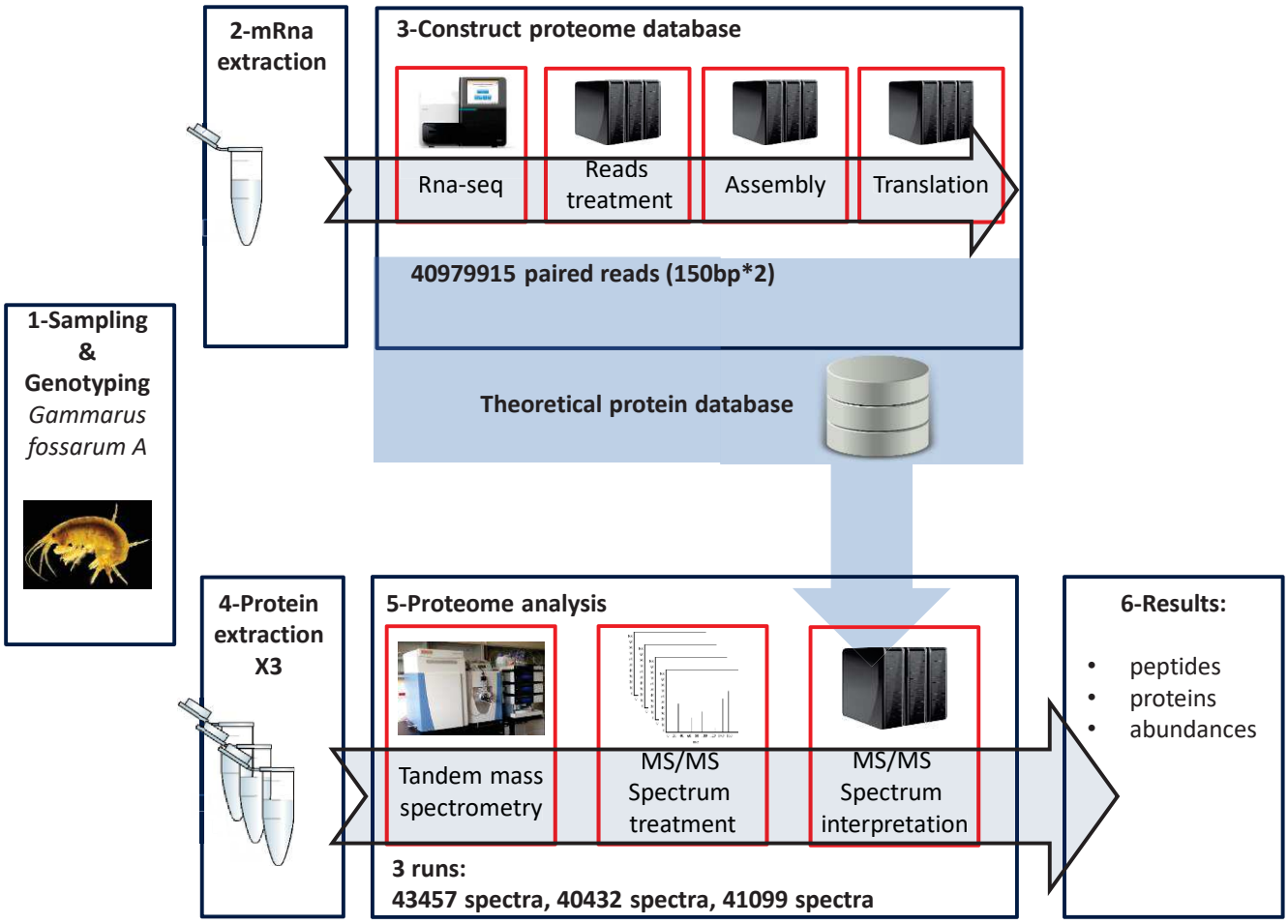


Figure 1: **Strategy for obtaining high quality RNAseq and shotgun proteomics data on genotyped *G. fossarum*.** The different steps are numbered from 1 to 6. A theoretical protein sequence database was derived from RNAseq obtained on one individual. This database was used to assign peptide sequences to three proteomic datasets acquired on three individuals. As explained in the main text, several databases were constructed to assess different parameters of RNAseq pre-treatments before assembly and translation.

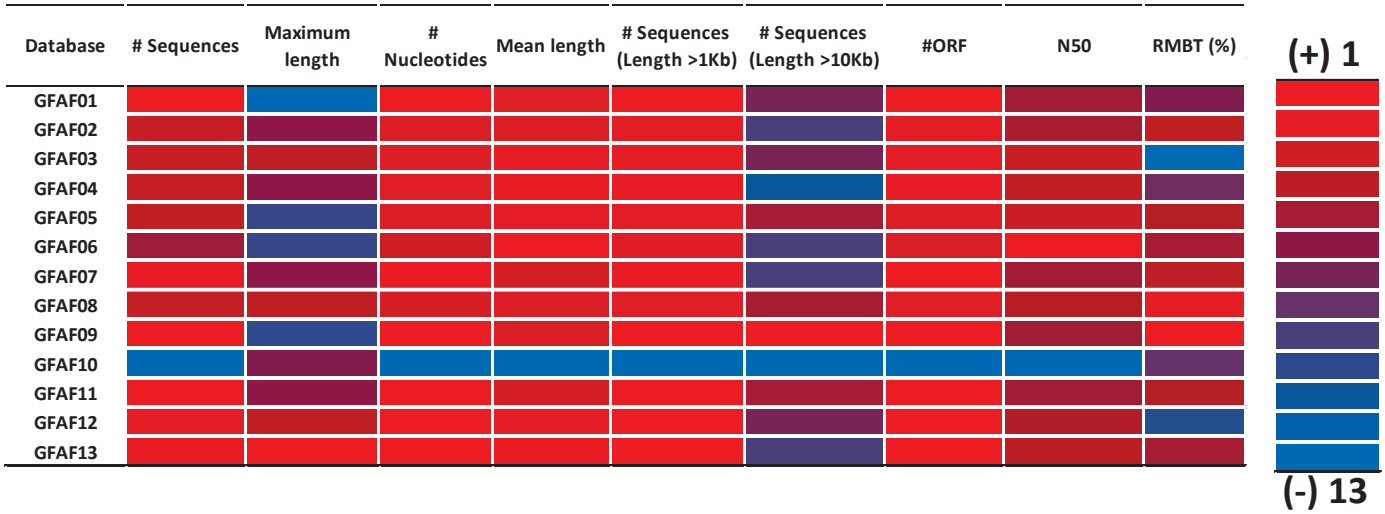
Evaluation of read treatment by classical transcriptomic criteria

Classical transcriptomics criteria for evaluating the read treatment after assembly are: the number of contig sequences, the global size of the assembled contigs, the mean length of the contigs, the number of sequences above a given size, the number of contigs containing open reading frame, the number of contigs that covers a given ratio of the total dataset, and the remapping percentage of initials raw data on assembled transcript. The higher the values in each category, the better the treatment is considered. The thirteen assemblies were ranked based on each of these criteria from the best (as 1) to the worst (as 13). The resulting scores for the 9 criteria (**Supplementary Table S1**) are visualized as a heatmap (**Figure 2, panel A**). Assembly comparison using these criteria do not result in any consensual optimum, the best assembly changing upon a single criterion. The assembly GFAF10 which consists in the systematic removal of the first 12 nucleotides of each read appears as the worst assembly in most cases because of the amount of lost information. Therefore, this treatment should clearly be avoided. Another interesting parameter frequently used to monitor the quality of an assembly and the completeness of a transcriptome is relying on the number of universal single-copy orthologs found in the dataset. **Figure 2 (Panel B)** shows the results for a BUSCO analysis at Arthropods level. For this, orthologous sequences defined at Arthropods level were searched in each transcriptome and classified as complete retrieval, partial retrieval or missing. Here, no optimum assembly was found when comparing the 13 pre-treatment strategies before the assembly based on the number of sequences found in each category**.** In most cases the number of complete orthologs was in the range 857-861 and the missing orthologous sequences were in the range of 67-70 sequences. In agreement with previous results, the assembly GFAF10 has significantly more missing and fragmented sequences compared to the others, and once again appears to be the worst assembly. Obviously, any of the classical transcriptomics criteria could distinguish the other assembly pre-treatment strategies for the non-model organism *G. fossarum* for which no reference genome is currently available.

Evaluation of read treatment by means of proteogenomic criteria

The proteomics data acquired on samples S1, S2 and S3 were interpreted querying the recorded MS/MS spectra against a database consisting of the translated assembled contigs of each of the thirteen strategies. The number of assigned MS/MS spectra or the number of proteins detected with the thirteen databases can be compared for highlighting the best pre-treatment strategy. This interpretation was performed at different FDR values (from 1% to 15% with increments of 1%) considering the results of a decoy database search. At each FDR and for each proteomic dataset, a p-value threshold is established. **Figure 3 (panel A)** shows the mean coefficient of variation of these threshold p-values when considering the 13 different assemblies along the FDR considered. Interestingly, we noticed that for each of the three proteomic samples the mean coefficient of variation is higher at low FDR thresholds (e.g. 15% at FDR 1%) and significantly decreases at high FDR thresholds (**Figure 3, panel A).** Furthermore, a higher variation is found at low FDR thresholds. This is somehow expected as a stringent FDR threshold lowers both the numbers of true positive results and hits on the decoy database, resulting in discontinuous, sporadic values, and thus increasing variations. To minimize these fluctuations, we compared the different strategies at a common FDR value of 5%. For each of the three proteomics datasets, we interpreted the MS/MS spectra against the

A-



B-

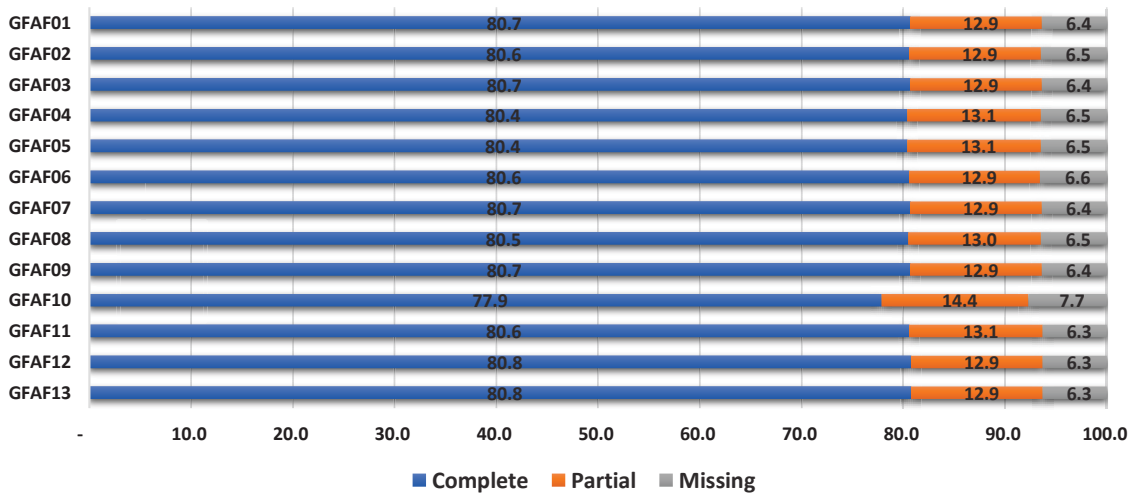


Figure 2: **Classical assessment of RNAseq assemblies. A.** The thirteen assays numbered from GFAF01 to GFAF13 were ranked from the best to the worse in terms of number of sequences, maximum length of contigs, number of nucleotides, mean length, number of sequences with length above 1kb, number of sequences with length above ten kb, number of ORFs, N50, and percentage of reads that could be mapped back to transcripts (RMBT). **B.** BUSCO analysis showing for the thirteen options the percentage of complete, partial and missing *Arthropoda* orthologs.

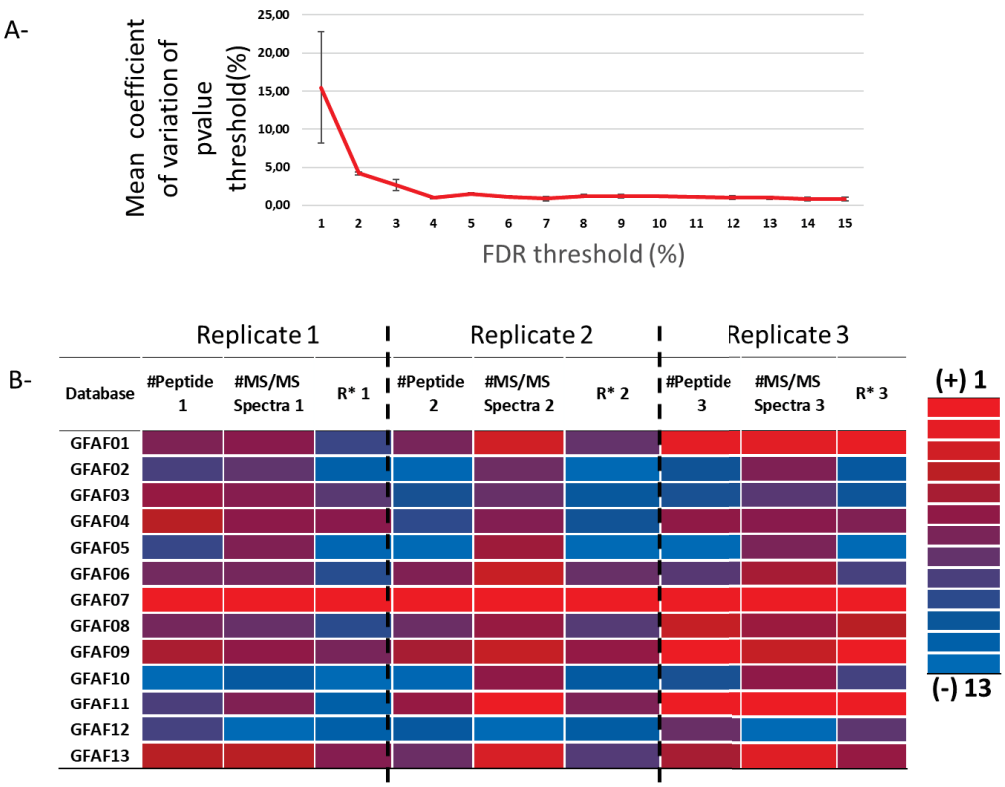


Figure 3: **Proteogenomics-based assessment of RNAseq assemblies. A.** Mean of variation coefficient of pvalue thresholds for the thirteen assays depending on the FDR threshold. **B.** The thirteen assays (GFAF01 to GFAF13) were ranked from the best to the worse in terms of number of identified peptides, number of assigned MS/MS spectra, and R* coefficient for each of the three MS/MS datasets (1, 2, and 3).

thirteen databases and established the number of peptide sequences identified and the number of MS/MS spectra assigned to peptide sequences. Furthermore, when merging the thirteen results, we established the total number of peptide sequences detected by tandem mass spectrometry. We also calculated the R* parameter, representing the ratio of additional detected peptides versus lost detected peptides. **Figure 3 (Panel B)** shows the heatmap of the 13 pre-treatments applied on the three proteomic datasets with each of these three criteria (**Supplementary Table S2**) and ranked from the best (1) to the worse (13). This heatmap indicates that some assembly options gave better proteomic results than the others. For example, the assembly pre-treatment GFAF07 is giving the highest number of peptides and number of assigned MS/MS spectra for the first two datasets. On this basis, the assembly pre-treatment GFAF09 can be ranked as the second-best strategy. This observation resulted in the creation of the assembly pre-treatment GFAF13 where the quality filters applied in GFAF07 and GFAF09 were merged, *i.e.* reads with at least one N and reads with mean Q score below 17 were removed. However, this assembly did not result better than the assembly pre-treatment GFAF07 (**Figure 3, Panel B**). Remarkably, the assembly pre-treatment GFAF01 which consists in no filtering the data can also be considered as a well performing route. In line with the transcriptomic criteria described earlier, the assembly pre-treatment GFAF10 shows the worse results when proteogenomic criteria were applied with a loss of an average of 63 detected peptides and 173 assigned MS/MS spectra compared to the best assembly (GFAF07). Two other assembly pre-treatments, GFAF05 and GFAF02, are also of low performances.

Optimization of the assembly translation assessed by proteogenomics

A first protein sequence database (DB00) was created by systematic translation from STOP-to-STOP of the whole transcriptome obtained by the GFAF07 pre-treatment strategy before *de novo* assembly. This database comprises 3,767,382 sequences totalling 105,436,491 amino acids. In order to reduce this database to the most probable content a selection of the relevant CDS can be performed with the Transdecoder tool. Two translation parameters using Transdecoder were explored resulting in the creation of various hypothetical protein sequence databases. These parameters were: i) the minimal length of the ORF to consider for the validation by Transdecoder algorithm after selection of the top 500 ORFs as training dataset, and ii) the minimal length of ORF to systematically take into account. **Figure 4 (panel A)** shows the 20 assays that were performed together with the selected values for the two parameters taking the GFAF07 optimized assembly as entry. The size of these databases was comprised in the range 10,099,231 – 62,638,091 amino acids. In terms of size, they compare favourably with the STOP-to-STOP systematic translation which size is 105,436,491 residues. The 20 resulting databases were used to interpret the three experimental MS/MS datasets and the results were compared, including the reference initial database, *i.e.* GFAF 3 frame. **Figure 4 (panel B)** presents the mean of the coefficient of variation of the threshold p-values of the 21 searches performed at different FDR. In this case, we noted a higher variation at low and high FDR thresholds, with a minimum at 3% FDR. **Figure 4 (panel A)** shows the results of the search performed at FDR 3% for the 21 databases in terms of number of PSMs and peptide sequences. A total of 1,250 additional MS/MS spectra could be interpreted when using the translation strategy 19 compared to the reference strategy (GFAF 3 frame). The number of peptide sequences detected resulted also higher with 369 additional sequences while the size of the database decreased drastically (-88%), thus speeding the calculation process. Transdecoder allows to

A-

Database	length ORF considerated (AA)	length conserved ORF (nucleotide)	Database size	# Peptides	#MS/MS spectra	R*	Size reduction (%)	Recall (%)
GFAF 3 frame	-	-	105436491	5192	14432	0.4	-	100
DB01	7	25	62638091	5332	14830	0.6	40.6	98.2
DB02	7	50	58809926	5360	14917	0.6	44.2	98.1
DB03	7	75	52366810	5381	15005	0.7	50.3	97.8
DB04	7	100	44525077	5379	15043	0.7	57.8	97.5
DB05	7	150	33214179	5443	15239	0.9	68.5	97.2
DB06	7	200	25917109	5489	15395	1.2	75.4	96.9
DB07	30	200	25159217	5489	15404	1.2	76.1	96.9
DB08	50	200	24267101	5474	15385	1.1	77	96.7
DB09	7	300	19303626	5535	15542	1.6	81.7	96.6
DB10	30	300	18491077	5516	15503	1.4	82.5	96.5
DB11	50	300	17538973	5522	15537	1.5	83.4	96.4
DB12	100	300	14920818	5430	15354	0.9	85.8	94
DB13	7	600	15719252	5546	15598	1.7	85.1	96.2
DB14	30	600	14874281	5543	15603	1.7	85.9	96.2
DB15	50	600	13888175	5555	15661	1.8	86.8	96
DB16	100	600	11214438	5479	15518	1.1	89.4	93.7
DB17	7	900	14634772	5537	15589	1.6	86.1	96
DB18	30	900	13779424	5538	15609	1.6	86.9	96
DB19	50	900	12783322	5561	15682	1.8	87.9	95.8
DB20	100	900	10099231	5466	15504	1	90.4	93.5

B-

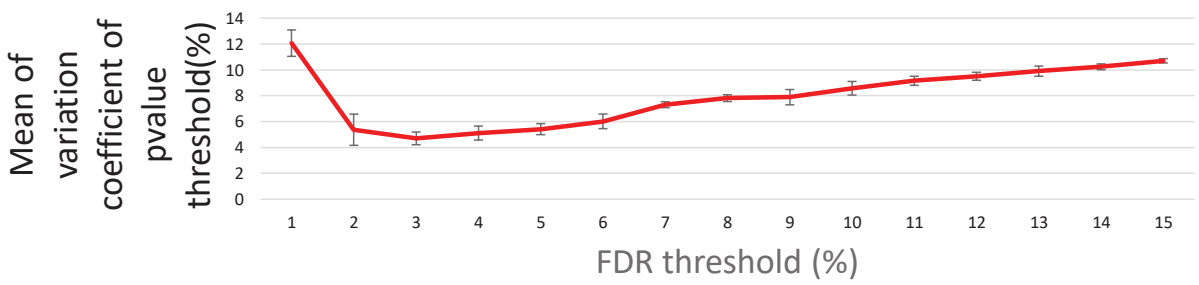


Figure 4: **Proteogenomics-based assessment of RNAseq data translation strategies. A.** The thirteen assays (GFAF01 to GFAF13) were ranked from the best to the worse in terms of number of identified peptides, number of assigned MS/MS spectra, and R* coefficient for the three merged MS/MS datasets. The resulting size reduction for each database is indicated in percentage as well as the recall percentage. **B.** Mean of variation coefficient of pvalue thresholds for the thirteen procedures depending on the FDR threshold.

decrease the database size while the number of peptides detected in the STOP-to-STOP comprehensive database are highly recall in all translation done with transdecoder (93% to 98% of peptides retrieved). In order to compare the 20 new databases, we calculated a R* factor that emphasizes for each database search the ratio of peptide sequences found consensual with the other searches. On this basis, the best database is DB19 obtained using Transdecoder with a minimal ORF length of 50 amino acids and systematic selection of ORFs above 300 amino acids. With this database a total of 61,126 peptide-to-spectrum were assigned to peptides, revealing a set of 5619 specific peptides and a total of 859 proteins validated with at least two specific peptides when merging the three whole-body proteome samples (**Supplementary Table S3**).

Concluding remarks

Deep RNAseq on non-model organisms can be performed to establish an exhaustive protein sequence database useful for proteomics interpretation. Here, we recorded RNAseq and shotgun proteomics datasets on individual *G. fossarum* animals. The different strategies for pretreating the RNAseq reads are not equal in terms of MS/MS spectra assigned to peptide sequences. Therefore, proteogenomics is an interesting tool to define the most interesting RNAseq assembly. We have shown that removing reads with mean Q score equivalent to a single probable nucleotide error on reads with length of 150 bp is prior assembly perform better than all other read pre-treatments. Gains in terms of MS/MS interpretation can be obtained by optimizing the translation strategy. Here, we have shown that using Transdecoder for translating the *de novo* assembled contigs a minimal ORF length of 50 amino acids and systematic selection of ORFs above 300 amino acids is optimal for this sequencing dataset and transcriptome assembling.

ACKNOWLEDGEMENTS

We thank the Institut National de Recherche en Sciences et Technologies pour l'Environnement et l'Agriculture (France), the Commissariat à l'Energie Atomique et aux Energies Alternatives (France) through the transversal toxicology program (PPTOX), and the Agence Nationale de la Recherche program "ProteoGam" (ANR-14-CE21-0006-02) for financial support.

REFERENCES

1. Evans, J. D., The i5K Initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. *J Hered* **2013,** *104*, (5), 595-600.

2. Thomas, G. W. C.; Dohmen, E.; Hughes, D. S. T.; Murali, S. C.; Poelchau, M.; Glastad, K.; Anstead, C. A.; Ayoub, N. A.; Batterham, P.; Bellair, M.; Binford, G. J.; Chao, H.; Chen, Y. H.; Childers, C.; Dinh, H.; Doddapaneni, H.; Duan, J. J.; Dugan, S.; Esposito, L. A.; Friedrich, M.; Garb, J.; Gasser, R. B.; Goodisman, M. A. D.; Gundersen-Rindal, D. E.; Han, Y.; Handler, A. M.; Hatakeyama, M.; Hering, L.; Hunter, W. B.; Ioannidis, P.; Jayaseelan, J. C.; Kalra, D.; Khila, A.; Korhonen, P. K.; Lee, C. E.; Lee, S. L.; Li, Y.; Lindsey, A. R. I.; Mayer, G.; McGregor, A. P.; McKenna, D. D.; Misof, B.; Munidasa, M.; Munoz-Torres, M.; Muzny, D. M.; Niehuis, O.; Osuji-Lacy, N.; Palli, S. R.; Panfilio, K. A.; Pechmann, M.; Perry, T.; Peters, R. S.; Poynton, H. C.; Prpic, N.-M.; Qu, J.; Rotenberg, D.; Schal, C.; Schoville, S. D.; Scully, E. D.; Skinner, E.; Sloan, D. B.; Stouthamer, R.; Strand, M. R.; Szucsich, N. U.; Wijeratne, A.; Young, N. D.; Zattara, E. E.; Benoit, J. B.; Zdobnov, E. M.; Pfrender, M. E.; Hackett, K. J.; Werren, J. H.; Worley, K. C.; Gibbs, R. A.; Chipman, A. D.; Waterhouse, R. M.; Bornberg-Bauer, E.; Hahn, M. W.; Richards, S., The Genomic Basis of Arthropod Diversity. *bioRxiv* **2018**, 382945.

3. Armengaud, J.; Trapp, J.; Pible, O.; Geffard, O.; Chaumot, A.; Hartmann, E. M., Non-model organisms, a species endangered by proteogenomics. *J Proteomics* **2014,** *105*, 5-18.

4. McAfee, A.; Harpur, B. A.; Michaud, S.; Beavis, R. C.; Kent, C. F.; Zayed, A.; Foster, L. J., Toward an Upgraded Honey Bee (Apis mellifera L.) Genome Annotation Using Proteogenomics. *J Proteome Res* **2016,** *15*, (2), 411-21.

5. Trapp, J.; Geffard, O.; Imbert, G.; Gaillard, J. C.; Davin, A. H.; Chaumot, A.; Armengaud, J., Proteogenomics of Gammarus fossarum to document the reproductive system of amphipods. *Mol Cell Proteomics* **2014,** *13*, (12), 3612-25.

6. Zhu, X.; Xie, S.; Armengaud, J.; Xie, W.; Guo, Z.; Kang, S.; Wu, Q.; Wang, S.; Xia, J.; He, R.; Zhang, Y., Tissue-specific Proteogenomic Analysis of Plutella xylostella Larval Midgut Using

a Multialgorithm Pipeline. *Mol Cell Proteomics* **2016,** *15*, (6), 1791-807.

7. Gouveia, D.; Almunia, C.; Cogne, Y.; Pible, O.; Degli-Esposti, D.; Salvador, A.; Cristobal, S.; Sheehan, D.; Chaumot, A.; Geffard, O.; Armengaud, J., Ecotoxicoproteomics: A decade of progress in our understanding of anthropogenic impact on the environment. *J Proteomics* **2019,** *198*, 66-77.

8. Trapp, J.; Armengaud, J.; Salvador, A.; Chaumot, A.; Geffard, O., Next-generation proteomics: toward customized biomarkers for environmental biomonitoring. *Environ Sci Technol* **2014,** *48*, (23), 13560-72.

9. Del Fabbro, C.; Scalabrin, S.; Morgante, M.; Giorgi, F. M., An extensive evaluation of read trimming effects on Illumina NGS data analysis. *PLoS One* **2013,** *8*, (12), e85024.

10. Macmanes, M. D., On the optimal trimming of high-throughput mRNA sequence data. *Front Genet* **2014,** *5*, 13.

11. Mbandi, S. K.; Hesse, U.; Rees, D. J.; Christoffels, A., A glance at quality score: implication for de novo transcriptome reconstruction of Illumina reads. *Front Genet* **2014,** *5*, 17.

12. Williams, C. R.; Baccarella, A.; Parrish, J. Z.; Kim, C. C., Trimming of sequence reads alters RNA-Seq gene expression estimates. *BMC Bioinformatics* **2016,** *17*, 103.

13. Lahens, N. F.; Kavakli, I. H.; Zhang, R.; Hayer, K.; Black, M. B.; Dueck, H.; Pizarro, A.; Kim, J.; Irizarry, R.; Thomas, R. S.; Grant, G. R.; Hogenesch, J. B., IVT-seq reveals extreme bias in RNA sequencing. *Genome Biol* **2014,** *15*, (6), R86.

14. MacManes, M. D., Establishing evidenced-based best practice for the de novo assembly and evaluation of transcriptomes from non-model organisms. *bioRxiv* **2016**, 035642.

15. Holzer, M.; Marz, M., De novo transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-Seq assemblers. *Gigascience* **2019,** *8*, (5).

16. He, B.; Zhao, S.; Chen, Y.; Cao, Q.; Wei, C.; Cheng, X.; Zhang, Y., Optimal assembly strategies of transcriptome related to ploidies of eukaryotic organisms. *BMC Genomics* **2015,** *16*, 65.

17. Huang, X.; Chen, X. G.; Armbruster, P. A., Comparative performance of transcriptome assembly methods for non-model organisms. *BMC Genomics* **2016,** *17*, 523.

18. Rana, S. B.; Zadlock, F. J. t.; Zhang, Z.; Murphy, W. R.; Bentivegna, C. S., Comparison of De Novo Transcriptome Assemblers and k-mer Strategies Using the Killifish, Fundulus heteroclitus. *PLoS One* **2016,** *11*, (4), e0153104.

19. Wang, S.; Gribskov, M., Comprehensive evaluation of de novo transcriptome assembly programs and their effects on differential gene expression analysis. *Bioinformatics* **2017,** *33*, (3), 327-333.

20. Zhao, Q. Y.; Wang, Y.; Kong, Y. M.; Luo, D.; Li, X.; Hao, P., Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study. *BMC Bioinformatics* **2011,** *12 Suppl 14*, S2.

21. Simao, F. A.; Waterhouse, R. M.; Ioannidis, P.; Kriventseva, E. V.; Zdobnov, E. M., BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **2015,** *31*, (19), 3210-2.

22. Smith-Unna, R.; Boursnell, C.; Patro, R.; Hibberd, J. M.; Kelly, S., TransRate: reference-free quality assessment of de novo transcriptome assemblies. *Genome Res* **2016,** *26*, (8), 1134-44.

23. Ma, J.; Saghatelian, A.; Shokhirev, M. N., The influence of transcript assembly on the proteogenomics discovery of microproteins. *PLoS One* **2018,** *13*, (3), e0194518.

24. Luge, T.; Fischer, C.; Sauer, S., Efficient Application of De Novo RNA Assemblers for Proteomics Informed by Transcriptomics. *J Proteome Res* **2016,** *15*, (10), 3938-3943.

25. Otte, K. A.; Schlotterer, C., Polymorphism-aware protein databases - a prerequisite for an unbiased proteomic analysis of natural populations. *Mol Ecol Resour* **2017,** *17*, (6), 1148-1155.

26. Sheynkman, G. M.; Shortreed, M. R.; Cesnik, A. J.; Smith, L. M., Proteogenomics: Integrating Next-Generation Sequencing and Mass Spectrometry to Characterize Human Proteomic Variation. *Annu Rev Anal Chem (Palo Alto Calif)* **2016,** *9*, (1), 521-45.

27. Kall, L.; Storey, J. D.; MacCoss, M. J.; Noble, W. S., Posterior error probabilities and false discovery rates: two sides of the same coin. *J Proteome Res* **2008,** *7*, (1), 40-4.

28. Kim, S.; Gupta, N.; Pevzner, P. A., Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. *J Proteome Res* **2008,** *7*, (8), 3354-63.

29. Jagtap, P.; Goslinga, J.; Kooren, J. A.; McGowan, T.; Wroblewski, M. S.; Seymour, S. L.; Griffin, T. J., A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies. *Proteomics* **2013,** *13*, (8), 1352-7.

30. de Groot, A.; Dulermo, R.; Ortet, P.; Blanchard, L.; Guerin, P.; Fernandez, B.; Vacherie, B.; Dossat, C.; Jolivet, E.; Siguier, P.; Chandler, M.; Barakat, M.; Dedieu, A.; Barbe, V.; Heulin, T.; Sommer, S.; Achouak, W.; Armengaud, J., Alliance of proteomics and genomics to unravel the specificities of Sahara bacterium Deinococcus deserti. *PLoS Genet* **2009,** *5*, (3), e1000434.

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

32. Li, H.; Joh, Y. S.; Kim, H.; Paek, E.; Lee, S. W.; Hwang, K. B., Evaluating the effect of database inflation in proteogenomic search on sensitive and reliable peptide identification. *BMC Genomics* **2016,** *17*, (Suppl 13), 1031.

33. Trapp, J.; Armengaud, J.; Gaillard, J. C.; Pible, O.; Chaumot, A.; Geffard, O., High-throughput proteome dynamics for discovery of key proteins in sentinel species: Unsuspected vitellogenins diversity in the crustacean Gammarus fossarum. *J Proteomics* **2016,** *146*, 207-14.

34. Cogne, Y.; Degli-Esposti, D.; Pible, O.; Gouveia, D.; François, A.; Bouchez, O.; Eché, C.; Ford, A.; Geffard, O.; Armengaud, J.; Chaumot, A.; Almunia, C., De novo transcriptomes of 14 gammarid individuals for proteogenomic

analysis of 7 different taxonomical groups. *Nature Scientific Data* **2019**, (In revision).

35. Langmead, B.; Salzberg, S. L., Fast gapped-read alignment with Bowtie 2. *Nat Methods* **2012,** *9*, (4), 357-9.

36. Quast, C.; Pruesse, E.; Yilmaz, P.; Gerken, J.; Schweer, T.; Yarza, P.; Peplies, J.; Glockner, F. O., The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* **2013,** *41*, (Database issue), D590-6.

37. Wright, J. C.; Choudhary, J. S., DecoyPyrat: Fast Non-redundant Hybrid Decoy Sequence Generation for Large Scale Proteomics. *J Proteomics Bioinform* **2016,** *9*, (6), 176-180.

38. Cogne, Y.; Almunia, C.; Gouveia, D.; Pible, O.; François, A.; Degli-Esposti, D.; Geffard, O.; Armengaud, J.; Chaumot, A., Comparative proteomics in the wild: accounting for intrapopulation variability improves describing proteome response in a Gammarus pulex field population exposed to cadmium. *Aquatic toxicology* **2019**, (In press).

39. Hartmann, E. M.; Allain, F.; Gaillard, J. C.; Pible, O.; Armengaud, J., Taking the shortcut for high-throughput shotgun proteomic analysis of bacteria. *Methods Mol Biol* **2014,** *1197*, 275-85.

40. Klein, G.; Mathe, C.; Biola-Clier, M.; Devineau, S.; Drouineau, E.; Hatem, E.; Marichal, L.; Alonso, B.; Gaillard, J. C.; Lagniel, G.; Armengaud, J.; Carriere, M.; Chedin, S.; Boulard, Y.; Pin, S.; Renault, J. P.; Aude, J. C.; Labarre, J., RNA-binding proteins are a major target of silica nanoparticles in cell extracts. *Nanotoxicology* **2016,** *10*, (10), 1555-1564.

41. Perez-Riverol, Y.; Csordas, A.; Bai, J.; Bernal-Llinares, M.; Hewapathirana, S.; Kundu, D. J.; Inuganti, A.; Griss, J.; Mayer, G.; Eisenacher, M.; Perez, E.; Uszkoreit, J.; Pfeuffer, J.; Sachsenberg, T.; Yilmaz, S.; Tiwary, S.; Cox, J.; Audain, E.; Walzer, M.; Jarnuczak, A. F.; Ternent, T.; Brazma, A.; Vizcaino, J. A., The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res* **2019,** *47*, (D1), D442-D450.

List of supplementary material

Table S1. Classical transcriptomic evaluation of the 13 assembly strategies.

Table S2. Proteogenomics evaluation of the 13 assembly strategies.

Table S3. List of proteins identified in the best assembly and translation workflow (GFAF07 assembly and DB19 translated.

Chapitre 2: Réalisation des assemblages des transcriptomes des 7 espèces sélectionnées.

Afin d'alimenter les connaissances moléculaires sur les Gammares pour l'interprétation de données protéomiques ou pour l'exploitation en termes de séquences de transcrits, il est essentiel de réaliser des transcriptomes de référence de plusieurs espèces de Gammaridés. Pour cela, le séquençage Illumina profond d'individus, plus précisément d'un mâle et d'une femelle, de 7 espèces de Gammaridés les plus répandus en Europe a été réalisé.

Les travaux réalisés dans le chapitre précédent nous ont permis d'établir la méthodologie de traitement des données transcriptomiques optimale pour l'assemblage des 14 transcriptomes. Ces assemblages ont été soumis à un contrôle de qualité puis ils ont été annotés fonctionnellement. Enfin l'ensemble des données initiales et interprétées de ces assemblages a été soumis à la base de données NCBI, et est présenté dans un manuscrit publié dans Nature Scientific Data. Ce chapitre présente le bilan de l'ensemble de ces étapes réalisées sur les 14 assemblages dorénavant disponibles pour l'ensemble de la communauté scientifique.

De novo transcriptomes of 14 gammarid individuals for proteogenomic analysis of 7 different taxonomical groups

Yannick Cogne[1], Davide Degli-Esposti[2], Olivier Pible[1], Duarte Gouveia[1], Adeline François[2], Olivier Bouchez[3], Camille Eché[3], Alex Ford[4], Olivier Geffard[2], Jean Armengaud[1#], Arnaud Chaumot[2], Christine Almunia[1]

[1]Laboratoire Innovations technologiques pour la Détection et le Diagnostic (Li2D), Service de Pharmacologie et Immunoanalyse (SPI), CEA, INRA, F-30207 Bagnols-sur-Cèze, France.

[2]Irstea, UR MALY Laboratoire d'écotoxicologie, centre de Lyon-Villeurbanne, F-69625 Villeurbanne, France.

[3]GeT-PlaGe, Genotoul, INRA Auzeville, F-31320 Castanet-Tolosan, France.

[4]School of Biological Sciences, Institute of Marine Sciences Laboratories, P04 9LY Portsmouth, United Kingdom.

[#]**Corresponding author:** Jean Armengaud, CEA-Marcoule, DRF-Li2D, Laboratory "Innovative technologies for Detection and Diagnostics", BP 17171, F-30200 Bagnols-sur-Cèze, France; jean.armengaud@cea.fr; Tel: +00 33 4 66 79 68 02; Fax: +00 33 4 66 79 19 05.

ABSTRACT

Gammarids are amphipods with a worldwide distribution, inhabiting fresh and marine waters. They play an important role in aquatic ecosystems and are well established sentinel species in ecotoxicology. We sequenced the transcriptomes of a male individual and a female individual for seven different taxonomical groups belonging to the two genders *Gammarus* and *Echinogammarus*: *Gammarus fossarum A, G. fossarum B, G. fossarum C*, *Gammarus wautieri*, *Gammarus pulex*, *Echinogammarus berilloni* and *Echinogammarus marinus*. These taxa were chosen to explore the molecular diversity of transcribed genes of genotyped individuals from these groups. Transcriptomes were *de novo* assembled and annotated. High-quality assembly was confirmed by BUSCO comparison against the Arthropod dataset. The fourteen RNA-seq derived protein sequence databases proposed here will be an important resource for proteogenomics on these non-model organisms of ecotoxicological relevance. These transcriptomes will serve as reliable reference sequences for whole transcriptome and proteome studies on other gammarids, in primer design for cloning specific genes or monitoring their specific expression, and in analyses of molecular differences between gammarid species.

Keywords: Reference transcriptomes, proteogenomic databases, sentinel species, water quality biomonitoring, ecotoxicology.

Background & summary

Gammarid amphipods are typically a few millimeters long animals and present in a wide range of aquatic habitats[1]. In freshwater ecosystems, they are key animal species because of their central role within food webs and their biomass abundance as often they are the most dominant macro-invertebrates. They are a prey for many species, but also predators for many invertebrate species. They are also scavengers and shredders, and detritivores involved in leaf litter breakdown, playing a central role in the decomposition of organic matter in general. Thus, they modulate the composition of freshwater communities of invertebrates[2]. They are the subject of many recent studies regarding their sensitivity to pollutants[3-7].

Marine and freshwater resources are of the utmost importance for Life. Human-made chemical contaminants released into aquatic environments compromise the quality of water bodies, threatening the resident biodiversity and the services supplied by such ecosystems. Their quality should be evaluated not only by measuring the concentrations of pollutants present, but also by monitoring how life is impacted by the bioavailable pollutants and their synergistic/antagonist effects[8]. For this, biomonitoring with encaged representative sentinel species has proved to be a valuable tool for efficient ecotoxicological studies[9-13]. Specific traits such as molt delay, growth impairment, or reproduction defects can be monitored on sensitive animals subjected to toxic environments. These data can be then integrated into a quantitative water quality index that can be used by stakeholders in charge of aquatic ecosystem and water resource management[14]. Because of their ecological representativeness, invertebrates are commonly employed as test organisms in marine and ecotoxicological assessment. Specifically, gammarids have been successfully used as sentinel species of freshwater ecosystems after important investments in analysing their physiological responses to toxicants[15-23] and biomonitoring in caging systems[9,12]. Specific biomarkers were proposed and monitored by innovative methods such as tandem mass spectrometry[19,24-26]. Next-generation proteomics contributed to improve the knowledge on the molecular responses of gammarids to toxicants and led to the proposal of a large panel of appropriate biomarkers[27-30]. This approach was successful after deriving a protein sequence database from an RNAseq transcriptome translated in all the possible reading frames. This proteogenomics concept allowed establishing a large catalogue of protein sequences comprising 1,873 mass-spectrometry-certified proteins, thus representing an important amphipod proteomic resource[29].

Molecular resources regarding gammarids are still scarce[31]. No gammarid whole genome sequence was available till very recently. A first draft genome of *Gammarus lacustris* was released comprising 443,304 scaffoldds[32]. The genomes of two related amphipods, *Parhyale hawaiensis*[33] and *Hyalella azteca*[34] have been sequenced. RNAseq datasets were made available for *P. hawaiensis*[35-37], *Echinogammarus marinus*[38], *Eogammarus possjeticus*[39], *Gammarus fossarum*[29], *Gammarus chevreuxi*[40], *Gammarus pulex*[41], and *Gammarus minus*[42]. However, these datasets are not of equal quality in terms of mRNA sequence coverage, a crucial parameter for proteogenomics interpretation[43], assembled from mRNAs extracted from a pool of various animals or from specific tissues, and sometimes even not any more accessible as it is the case for *E. marinus* because the repository used[44] no longer exists.

The present data consist of assembled transcriptome sequences of fourteen different gammarids, seven males and seven females, namely *Gammarus fossarum A (Müller type A)*, *G. fossarum B (Müller type B)*, *G. fossarum C (Müller type C)*, *Gammarus wautieri*, *Gammarus pulex*, *Echinogammarus berilloni* and *Echinogammarus marinus*. These

transcriptomes were performed in order to obtain the same sequencing depth and quality (full length mRNAs from whole organism), as well as to apply the same assembly and translation pipelines. They were done on single animals to avoid sequence heterogeneity. The transcriptomes have been annotated for serving as reference protein sequence databases for proteogenomics of these sentinel animals that will be soon conducted to gain more fundamental knowledge for better aquatic environmental risk assessment. For this, an interesting strategy could be to interpret MS/MS shotgun data first on the most appropriate specific single-organism database, and then a follow-up search on a multi-organism database. The transcriptomes presented here will also serve in comparative analysis for better defining the molecular diversity amongst gammarids and will be a valuable sequence resource for future ecotoxicological studies.

Methods

Experimental design

Freshwater gammarids were collected in four geographically-distant French rivers (Table 1). One population of *Gammarus fossarum* was sampled in north-eastern France (Seebach river), previously identified as sheltering the cryptic subspecies type A from the three types defined in Müller et al.[45], Westram et al.[46], and Weiss et al.[47]. The second river (Pollon river) situated in mid-eastern area France corresponding to a sympatric situation supplied organisms belonging to *Gammarus fossarum* type B, type C and *Gammarus pulex* species. *Gammarus wautieri* were collected in the Galaveyson river in the Dauphiné region, and *Echinogammarus berilloni* organisms from a fourth river in south-western France (Saucats river). These freshwater gammarids were all sampled using a hand net by kick sampling and transported to the laboratory. After one-week maintenance in buckets containing water sampled from their respective origin site, kept at 12°C with a constant aeration, under a 16/8-h light/dark photoperiod, and with conditioned alder leaves as food supply, couples in amplexus were isolated for species determination before RNA extraction. Couples whose females had well developed ovaries were selected. Embryos were removed from the marsupial pouch of females for five of these couples. Based on the description of the reproductive cycle in *Gammarus fossarum*[48], we were able to select one couple per species in the last stage of the reproductive cycle (pre-molting stage for the female) by retaining pairs whose females were carrying embryos at the end of their embryonic development stage (stage 4 or 5) for RNA extraction. For the marine species, *E. marinus* were collected from beneath seaweed in the intertidal zone of Portsmouth, southern England, in the same population as a previous study[38]. After one-month lab maintenance in buckets with filtrated natural seawater at 10°C under a 12 h light/12 h dark photoperiod and fed with fucoid seaweed, organisms were live transported in damp seaweed from United-Kingdom to France (one-day travel). Then they were maintained for few hours in aquaria with reconstituted seawater (salinity 30 ‰) before the selection of organisms. For this species, it was not possible to recover couples in amplexus. One free-swimming male and one free-swimming female were isolated from the batch of sampled organisms. Stage 1 embryos were recovered from the female marsupium, indicating that this female was in a post-molting stage.

Species determination was first conducted according to morphological criteria[49]. To distinguish between the three cryptic lineages A, B, C within the *G. fossarum* species, a molecular species assignment was carried out by amplifying the 5' part of the mtDNA cytochrome c oxidase subunit I (COI) using universal primers (LCO1490

Table 1: Sampling information and number of reads for each sample before and after filtering by mean quality for the 14 transcriptomes.

Species	Code Name	Sex	River	City	Country	GPS		Number of raw reads	Number of reads after filtering
Echinogammarus berilloni	EGSF	Female	Saucats	Saucats	France	44°39'34"N	0°34'25"W	80 482 966	80 277 434
Echinogammarus berilloni	EGSM	Male	Saucats	Saucats	France	44°39'34"N	0°34'25"W	90 372 154	90 118 242
Echinogammarus marinus	EGUF	Female	sea coast	Portsmouth	UK	50°47'41"N	1°01'50"W	85 032 246	84 652 454
Echinogammarus marinus	EGUM	Male	sea coast	Portsmouth	UK	50°47'41"N	1°01'50"W	70 768 994	70 540 528
*Gammarus fossarum A**	GFAF	Female	Seebach	Fellering	France	47°53'31"N	6°58'53"E	81 959 830	81 543 116
*Gammarus fossarum A**	GFAM	Male	Seebach	Fellering	France	47°53'31"N	6°58'53"E	95 167 986	94 695 372
*Gammarus fossarum B**	GFBF	Female	Pollon	Saint-Maurice-de-Rémens	France	45°57'21"N	5°15'44"E	96 361 300	96 093 396
*Gammarus fossarum B**	GFBM	Male	Pollon	Saint-Maurice-de-Rémens	France	45°57'21"N	5°15'44"E	85 125 996	84 758 816
*Gammarus fossarum C**	GFCF	Female	Pollon	Saint-Maurice-de-Rémens	France	45°57'21"N	5°15'44"E	78 459 708	77 977 148
*Gammarus fossarum C**	GFCM	Male	Pollon	Saint-Maurice-de-Rémens	France	45°57'21"N	5°15'44"E	75 598 166	75 407 534
Gammarus pulex	GPCF	Female	Pollon	Saint-Maurice-de-Rémens	France	45°57'21"N	5°15'44"E	84 202 086	83 965 920
Gammarus pulex	GPCM	Male	Pollon	Saint-Maurice-de-Rémens	France	45°57'21"N	5°15'44"E	89 235 492	89 025 410
Gammarus wautieri	GWF	Female	Galaveyson	Le Grand Serre	France	45°16'27''N	5°07'08"E	80 192 262	79 695 588
Gammarus wautieri	GWM	Male	Galaveyson	Le Grand Serre	France	45°16'27''N	5°07'08"E	63 959 618	63 638 482

*Müller type.

Table 2: Assembly quality metrics.

	EGSF	EGSM	EGUF	EGUM	GFAF	GFAM	GFBF	GFBM	GFCF	GFCM	GPCF
n_seqs	166,100	211,358	162,914	133,658	182,439	383,876	325,379	344,409	280,883	324,661	245,224
largest	21,406	28,082	25,426	29,815	11,828	22,574	26,858	21,757	29,633	25,029	17,350
n_bases	178,852,651	228,738,512	168,030,154	142,457,935	118,459,292	283,956,781	259,691,927	263,406,154	226,877,323	236,552,608	198,832,95
mean_len	1076.8	1082.2	1031.4	1065.8	649.3	739.7	798.1	764.8	807.7	728.6	810.8
n_over_1k	42,496	54,408	44,211	38,307	31,373	76,176	66,143	67,014	57,497	58,066	50,528
n_over_10k	498	827	348	324	5	156	345	308	303	202	232
n_with_orf	35,470	58,284	38,503	30,621	32,784	78,940	62,829	65,479	53,123	56,151	40,810
mean_orf (%)	41.0	47.9	46.0	43.3	51.9	54.2	50.5	50.7	49.3	50.2	45.3
n90	355	361	340	357	270	282	285	284	289	273	283
n50	2,646	2,594	2,278	2,299	963	1,240	1,518	1,354	1,555	1,290	1,622
n10	7,494	7,850	6,812	6,736	2,978	4,256	5,442	5,071	5,593	4,958	5,522
gc(%)	42.7	42.5	43.6	43.4	42.6	41.8	43.6	43.0	43.8	43.4	43.5
RMBT(%)*	91.7	94.4	89.6	91.9	88.2	83.9	90.1	82.7	87.7	86.1	81.9
G-RMBT(%)*	80.7	86.8	75.2	73.5	76.5	65.1	82.0	61.9	75.8	70.0	63.9
Score#	0.16	0.16	0.11	0.11	0.18	0.12	0.13	0.10	0.12	0.11	0.10

*RMBT means Reads Mapping Back on the Transcriptome; G-RMBT means Good Reads Mapping Back on the Transcriptome

Score calculated by Transrate.

Table 3: Accessions for the 14 transcriptomes.

Code Name	Transcriptome accession	Read accession	BioProject	BioSample
EGSF	GHCT01000000	*SRR8089732*	PRJNA497972	SAMN10259946
EGSM	GHCU01000000	*SRR8089733*	PRJNA497972	SAMN10259947
EGUF	GHCW01000000	*SRR8089734*	PRJNA497972	SAMN10259948
EGUM	GHCV01000000	*SRR8089735*	PRJNA497972	SAMN10259949
GFAF	GHCX01000000	*SRR8089727*	PRJNA497972	SAMN10259934
GFAM	GHCY01000000	*SRR8089728*	PRJNA497972	SAMN10259935
GFBF	GHCZ01000000	*SRR8089729*	PRJNA497972	SAMN10259936
GFBM	GHDA01000000	*SRR8089722*	PRJNA497972	SAMN10259937
GFCF	GHDC01000000	*SRR8089723*	PRJNA497972	SAMN10259938
GFCM	GHDB01000000	*SRR8089724*	PRJNA497972	SAMN10259939
GPCF	GHCP01000000	*SRR8089725*	PRJNA497972	SAMN10259940
GPCM	GHCQ01000000	*SRR8089720*	PRJNA497972	SAMN10259941
GWF	GHCR01000000	*SRR8089730*	PRJNA497972	SAMN10259944
GWM	GHCN01000000	*SRR8089731*	PRJNA497972	SAMN10259945

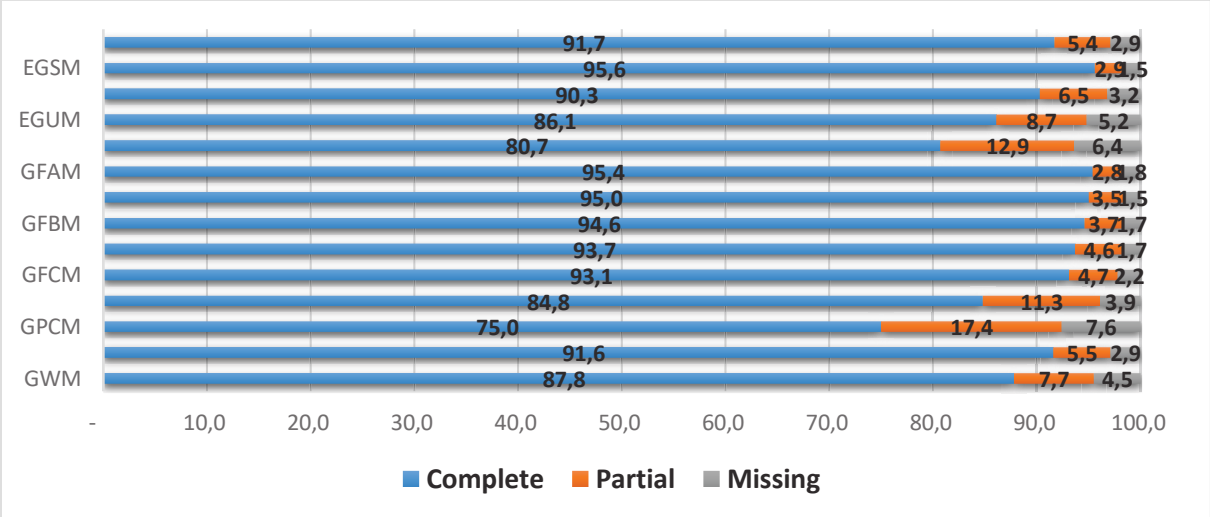


Figure 1. BUSCO assessment results for the 14 assembled transcriptomes.

and HCO2198)[50]. For this, DNA was extracted from one or two pereopods (depending on individual size) cut from organisms before their conditioning for RNA extraction. DNA extraction was performed using the Nucleospin tissue XS kit (Macherey-Nagel) and 10 ng of DNA for each organism were amplified using the mtDNA cytochrome c oxidase subunit I (COI) using universal primers LCO1490 (GGT CAA CAA ATC ATA AAG ATA TTG G) and HCO2198 (TAA ACT TCA GGG TGA CCA AAA AAT CA). The PCR conditions (45 cycles) consisted in the denaturation at 95°C for 30 sec, annealing at 50°C for 30 sec and elongation at 72°C for 1 min for each cycle. PCR products were purified by ultrafiltration using the Nucleofast kit (Macherey-Nagel). Purified amplicons were prepared for sequencing using the BigDye Terminator v3.1 kit (ThermoFisher) and then sequenced on a DNA analyser ABI 3730XL (ThermoFisher). Sequencing data were analyzed using the Sequencher 5.4.6 program (Genecodes). COI sequences (freely available from figshare, YC02_COI sequences and phylogenetic tree[51]) were aligned for building phylogenetic tree encompassing reference sequences from Weiss et al.[47] and Lagrue et al.[52]. The phylogenetic tree (freely available from figshare, YC02_COI sequences and phylogenetic tree[51]) positions the COI sequences of the *Gammarus* organisms selected for RNA sequencing in relation to reference sequences taken from the literature (SeaView software[53]; BioNJ method based on J-C distance). The robustness of the different groupings has been evaluated by bootstrap procedure (100 iterations). COI sequences were obtained for all *Gammarus* individuals except for the female *G. fossarum C*, but this individual was in precopulatory amplexus with the male COI-genotyped as *G. fossarum C*. However, in the same location (Pollon River), we were also able to obtain the COI genotype of 15 additional couples, all of which were found to be non-heterospecific (4 *G. fossarum* B, 3 *G. fossarum* C, 8 *G. pulex*). Westraam et al (2011) made the same observation in the Glovelier river which shelters *G. fossarum* A and B, with only one heterospecific pair for a total of 64 genotyped couples. Lagrue et al (2014) also observed that mixed couples are rare in the field for *Gammarus* lineages with COI distance superior to 4%. Considering that the divergence between the COI-genotyped *G. fossarum* B and C specimens is about 17% in the Pollon river, it is very unlikely that this female does not belong to the *G. fossarum* C species.

Dataset generation

Gammarids were put in RNAlater (Sigma) and kept at 4°C overnight. Then RNAlater was removed and organisms were snap frozen in liquid nitrogen and kept at -80°C until the RNA extraction was performed. Organisms were first homogenized in the lysis buffer using a bead homogenizer and then RNAs were extracted using the Qiagen fibrous kit (Qiagen). RNA quantity, quality and integrity were evaluated using Nanodrop (Thermo Fisher) and Bioanalyzer (Agilent). RNA-Seq libraries were generated using the TruSeq stranded mRNA Sample Prep kit (Illumina). mRNA were purified using poly-(T) beads from 2 µg of each total RNA sample, then cleaved in segments of 155 bp on average (120-210 bp range). Then, cleaved RNA fragments were primed with random hexamers and reverse transcribed into first strand cDNA. A second strand of cDNA was consecutively synthetized, and double-stranded cDNA was purified using beads. The 3' ends of the blunt fragments then were adenylated. Indexed adapters were ligated to the cDNA fragments enriched by PCR (11 cycles). Libraries then were purified and quality-assessed using a Fragment Analyzer (Advanced Analytical Technologies). The 16 libraries were quantified by qPCR using the Kapa Library Quantification Kit (Roche). They were normalized, multiplexed on 1 pool. Libraries were then sequenced on two lanes

of Hiseq3000 (Illumina) using a paired-end read length of 2x150 pb with the HiSeq Reagent Kits (Illumina). The 2 HiSeq lanes produced an average number of 40.0 ± 8 millions of read pairs per library. Control quality of reads was performed by FastQC version V0.11.2 (Babraham Bioinformatics). Detailed results are freely available from figshare (YC02_QC data[51]). The data records are stored as fourteen folders containing each four folders per transcriptome.

De novo assembly

For each sample reads from two different lanes were merged for each pair-end part. Filtering at mean Qphred score with a threshold set at 16.99 was done and unpaired reads resulting were removed using a homemade script. The numbers of reads for each sample before and after filtering by mean quality are presented in Table 1. Trinity v2.4[54] was used to assemble reads for each sample considering pair-end, strand orientation (-SS_lib type RF) and all other Trinity parameters were set at default values with k set at the value 25 and minimum contig length at 200 base pair.

Assessment of assembly quality

Quality assessment of transcriptomes was done with Transrate v1.0.1[55] which generates standard metrics and remapping statistics. No reference protein sequences were used for the evaluation with Transrate. The main metrics are shown in Table 2. For the validation of the quality of all the assemblies BUSCO v2.0[56] was used. The database used for BUSCO analyses was Arthropoda_odb9 which contains 1066 orthologous gene at the nearest taxon level (*i.e.* Arthropods) available for *Gammarus*. Results are shown in Figure 1. A high level of single copy orthologs retrieval is noted for the fourteen assemblies with at least 75% ratio. Furthermore, less than 8% of orthologs are missing in the worst case, and less than 5% are missing in 11 transcriptomes.

Annotation

For each sample the transcripts were annotated using the Trinotate v3.1.1 annotation pipeline[54]. Swissprot database was used as main database and Amphipods proteins referenced on Uniref were used as custom database. Blastx and Blastp were used for the similarity search step with an e-value cutoff set at 1e-2. Blast results were then used for generate the annotation report with the same e-value cutoff.

Code availability

Filtering step before assembly was done with homemade Pythonv2.7 script freely available (https://github.com/YannickCogne/Qfiltering) and automatized with bash script for each sample.

DATA RECORDS

Reads

Read sequences for each sample were deposited in the NCBI Sequence Reads Archive. These data have accessions SRR8089720[57], SRR8089722-SRR8089725[58-61], and SRR8089727-SRR8089735[62-70], as indicated in Table 3 with the corresponding Bioproject and Biosample codes. The FastQC results for the fourteen samples are freely available from figshare (YC02 _QC data[51]). The data records are stored as fourteen folders containing each four folders per transcriptome.

Transcriptomes

Transcriptome assembly for each sample were deposited in the NCBI Transcriptome Shotgun Assembly Sequence Database. These data have been deposited in GenBank, and have accessions GHCN01000000[71], GHCP01000000-GHCR01000000[72-74], GHCT01000000-GHCZ01000000[75-81], GHDA01000000-GHDC01000000[82-84], as indicated in Table 3 with the corresponding Bioproject and Biosample codes.

Proteogenomic databases

Translations of coding sequence regions were produced for each transcriptome by Transdecoder v3.0.1[54] from stop to stop codons, analysing only top strand with 500 top longest ORFs used for training, and retaining 600 base pair ORFs and only proteins with a minimum length of 50 amino acids. The fourteen translations are freely available as FASTA files from figshare (YC02_Transcriptome translated ORFs[51]).

Annotation

Annotation for each assembly is freely available as Excel file from figshare (YC02_Transcript annotations[51]). The folder contains fourteen Excel files.

TECHNICAL VALIDATION

Transrate

Transrate analyses show good remapping back results with more than 80% of reads remapped and almost assemblies with more than classed 70% as good mapping. Raw results from Transrate is freely available from figshare (YC02_ Transrate results[51]).

Busco

Busco analyzes show high level of single copy orthologs retrieval in all assemblies with all assemblies having more than 75% and almost assemblies with 90%.

ACKNOWLEDGEMENTS

We thank the Institut National de Recherche en Sciences et Technologies pour l'Environnement et l'Agriculture (France), the Commissariat à l'Energie Atomique et aux Energies Alternatives (France), and the Agence Nationale de la Recherche "ProteoGam" program (ANR-14-CE21-0006-02) for financial support.

AUTHOR CONTRIBUTIONS

YC, AF, OG, JA, AC and CA conceived the study. DDE, DG, AF, OB, CE, AC, and CA performed the experimental work. YC, DDE, OP, JA, AC, and CA analysed the data. YC and JA wrote the manuscript with help from all the co-authors.

REFERENCES

1 MacNeil, C., Dick, J. T. A. & Elwood, R. W. The trophic ecology of freshwater Gammarus Spp. (crustacea:amphipoda): problems and perspectives concerning the functional feeding group concept. *Biological Reviews* **72**, 349–364 (1997).

2 Kelly, D. W., Dick, J. T. A. & Montgomery, W. I. The functional role of Gammarus (Crustacea, Amphipoda): shredders, predators, or both? *Hydrobiologia* **485**, 199–203 (2002).

3 Arce-Funck, J. *et al.* High stoichiometric food quality increases moulting organism vulnerability to pollutant impacts: An experimental test with Gammarus fossarum (Crustacea: Amphipoda). *Sci Total Environ* **645**, 1484-1495 (2018).

4 Ganser, B. *et al.* Wastewater alters feeding rate but not vitellogenin level of Gammarus fossarum (Amphipoda). *Sci Total Environ* **657**, 1246-1252 (2019).

5 Konemann, S. *et al.* Combination of In Situ Feeding Rate Experiments and Chemical Body Burden Analysis to Assess the Influence of Micropollutants in Wastewater on Gammarus pulex. *Int J Environ Res Public Health* **16**, in press (2019).

6 Munz, N. A., Fu, Q., Stamm, C. & Hollender, J. Internal Concentrations in Gammarids Reveal Increased Risk of Organic Micropollutants in Wastewater-Impacted Streams. *Environ Sci Technol* **52**, 10347-10358 (2018).

7 von Fumetti, S. & Blaurock, K. Effects of the herbicide Roundup(R) on the metabolic activity of Gammarus fossarum Koch, 1836 (Crustacea; Amphipoda). *Ecotoxicology* **27**, 1249-1260 (2018).

8 Gouveia, D. *et al.* Ecotoxicoproteomics: A decade of progress in our understanding of anthropogenic impact on the environment. *J Proteomics* **198**, 66-77 (2018).

9 Besse, J. P. *et al.* Caged Gammarus fossarum (Crustacea) as a robust tool for the characterization of bioavailable contamination levels in continental waters: towards the determination of threshold values. *Water Res* **47**, 650-660 (2013).

10 Chaumot, A., Geffard, O., Armengaud, J. & Maltby, L. in *Aquatic Ecotoxicology - Advancing tools for dealing with emerging risks* (eds C. Amiard-Triquet, J.-C. Amiard, & C. Mouneyrac) 253-280 (Academic Press, London, 2015).

11 Ciliberti, A. *et al.* Caged Gammarus as biomonitors identifying thresholds of toxic metal bioavailability that affect gammarid densities at the French national scale. *Water Res* **118**, 131-140 (2017).

12 Lacaze, E. *et al.* DNA damage in caged Gammarus fossarum amphipods: a tool for freshwater genotoxicity assessment. *Environ Pollut* **159**, 1682-1691 (2011).

13 Trapp, J., Armengaud, J., Salvador, A., Chaumot, A. & Geffard, O. Next-generation proteomics: toward customized biomarkers for environmental biomonitoring. *Environ Sci Technol* **48**, 13560-13572 (2014).

14 Coulaud, R. *et al.* In situ feeding assay with Gammarus fossarum (Crustacea): Modelling the influence of confounding factors to improve water quality biomonitoring. *Water Res* **45**, 6417-6429 (2011).

15 Barros, S. *et al.* Chronic effects of triclocarban in the amphipod Gammarus locusta: Behavioural and biochemical impairment. *Ecotoxicol Environ Saf* **135**, 276-283 (2017).
16 Chaumot, A., Gos, P., Garric, J. & Geffard, O. Additive vs non-additive genetic components in lethal cadmium tolerance of Gammarus (Crustacea): novel light on the assessment of the potential for adaptation to contamination. *Aquat Toxicol* **94**, 294-299 (2009).
17 Correia, A. D., Lima, G., Costa, M. H. & Livingstone, D. R. Studies on biomarkers of copper exposure and toxicity in the marine amphipod Gammarus locusta (Crustacea): I. Induction of metallothionein and lipid peroxidation. *Biomarkers* **7**, 422-437 (2002).
18 Felten, V. *et al.* Physiological and behavioural responses of Gammarus pulex (Crustacea: Amphipoda) exposed to cadmium. *Aquat Toxicol* **86**, 413-425 (2008).
19 Jubeaux, G. *et al.* Vitellogenin-like proteins in the freshwater amphipod Gammarus fossarum (Koch, 1835): functional characterization throughout reproductive process, potential for use as an indicator of oocyte quality and endocrine disruption biomarker in males. *Aquat Toxicol* **112-113**, 72-82 (2012).
20 Kohler, S. A., Parker, M. O. & Ford, A. T. Species-specific behaviours in amphipods highlight the need for understanding baseline behaviours in ecotoxicology. *Aquat Toxicol* **202**, 173-180 (2018).
21 Maltby, L. & Crane, M. Responses of Gammarus pulex (Amphipoda, Crustacea) to metalliferous effluents: identification of toxic components and the importance of interpopulation variation. *Environ Pollut* **84**, 45-52 (1994).
22 Xuereb, B., Chaumot, A., Mons, R., Garric, J. & Geffard, O. Acetylcholinesterase activity in Gammarus fossarum (Crustacea Amphipoda) Intrinsic variability, reference levels, and a reliable tool for field surveys. *Aquat Toxicol* **93**, 225-233 (2009).
23 Xuereb, B., Noury, P., Felten, V., Garric, J. & Geffard, O. Cholinesterase activity in Gammarus pulex (Crustacea Amphipoda): characterization and effects of chlorpyrifos. *Toxicology* **236**, 178-189 (2007).
24 Gouveia, D. *et al.* 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 (2017).
25 Gouveia, D. *et al.* Assessing the relevance of a multiplexed methodology for proteomic biomarker measurement in the invertebrate species Gammarus fossarum: A physiological and ecotoxicological study. *Aquat Toxicol* **190**, 199-209 (2017).
26 Simon, R. *et al.* Mass spectrometry assay as an alternative to the enzyme-linked immunosorbent assay test for biomarker quantitation in ecotoxicology: application to vitellogenin in Crustacea (Gammarus fossarum). *J Chromatogr A* **1217**, 5109-5115 (2010).
27 Trapp, J. *et al.* High-throughput proteome dynamics for discovery of key proteins in sentinel species: Unsuspected vitellogenins diversity in the crustacean Gammarus fossarum. *J Proteomics* **146**, 207-214 (2016).
28 Trapp, J. *et al.* Proteomic investigation of male Gammarus fossarum, a freshwater crustacean, in response to endocrine disruptors. *J Proteome Res* **14**, 292-303 (2015).
29 Trapp, J. *et al.* Proteogenomics of Gammarus fossarum to document the reproductive system of amphipods. *Mol Cell Proteomics* **13**, 3612-3625 (2014).
30 Trapp, J. *et al.* Digging Deeper Into the Pyriproxyfen-Response of the Amphipod Gammarus fossarum With a Next-Generation Ultra-High-Field Orbitrap Analyser: New Perspectives for Environmental Toxicoproteomics. *Frontiers in Environmental Science* **6** (2018).
31 Armengaud, J. *et al.* Non-model organisms, a species endangered by proteogenomics. *J Proteomics* **105**, 5-18 (2014).
32 Jin, S. *et al.* Identification of Candidate Genes for the Plateau Adaptation of a Tibetan Amphipod, Gammarus lacustris, Through Integration of Genome and Transcriptome Sequencing. *Front Genet* **10**, 53 (2019).
33 Kao, D. *et al.* The genome of the crustacean Parhyale hawaiensis, a model for animal development, regeneration, immunity and lignocellulose digestion. *Elife* **5** (2016).
34 Poynton, H. C. *et al.* The Toxicogenome of Hyalella azteca: A Model for Sediment Ecotoxicology and Evolutionary Toxicology. *Environ Sci Technol* **52**, 6009-6022 (2018).
35 Blythe, M. J. *et al.* High through-put sequencing of the Parhyale hawaiensis mRNAs and microRNAs to aid comparative developmental studies. *PLoS One* **7**, e33784(2012).
36 Nestorov, P., Battke, F., Levesque, M. P. & Gerberding, M. The maternal transcriptome of the crustacean Parhyale hawaiensis is inherited asymmetrically to invariant cell lineages of the ectoderm and mesoderm. *PLoS One* **8**, e56049 (2013).
37 Zeng, V. *et al.* De novo assembly and characterization of a maternal and developmental transcriptome for the emerging model crustacean Parhyale hawaiensis. *BMC Genomics* **12**, 581 (2011).
38 Short, S. *et al.* Crustacean intersexuality is feminization without demasculinization: implications for environmental toxicology. *Environ Sci Technol* **48**, 13520-13529 (2014).
39 Chen, J., Liu, H., Cai, S. & Zhang, H. Comparative transcriptome analysis of Eogammarus possjeticus at different hydrostatic pressure and temperature exposures. *Sci Rep* **9**, 3456 (2019).
40 Truebano, M., Tills, O. & Spicer, J. I. Embryonic transcriptome of the brackishwater amphipod Gammarus chevreuxi. *Mar Genomics* **28**, 5-6 (2016).
41 Gismondi, E. & Thome, J. P. Transcriptome of the freshwater amphipod Gammarus pulex hepatopancreas. *Genom Data* **8**, 91-92 (2016).
42 Carlini, D. B. & Fong, D. W. The transcriptomes of cave and surface populations of Gammarus minus (Crustacea: Amphipoda) provide evidence for positive selection on cave downregulated transcripts. *PLoS One* **12**, e0186173 (2017).
43 Trapp, J. *et al.* Proteogenomic insights into the core-proteome of female reproductive tissues from crustacean amphipods. *J Proteomics* **135**, 51-61 (2016).
44 Jones, M. & Blaxter, M. afterParty: turning raw transcriptomes into permanent resources. *BMC Bioinformatics* **14**, 301 (2013).
45 Muller, J. Mitochondrial DNA variation and the evolutionary history of cryptic Gammarus fossarum types. *Mol Phylogenet Evol* **15**, 260-268 (2000).
46 Westram, A. M., Jokela, J., Baumgartner, C. & Keller, I. Spatial distribution of cryptic species diversity in european freshwater amphipods (Gammarus fossarum) as revealed by pyrosequencing. *PLoS One* **6**, e23879 (2011).

47 Weiss, M., Macher, J. N., Seefeldt, M. A. & Leese, F. Molecular evidence for further overlooked species within the Gammarus fossarum complex (Crustacea: Amphipoda). *Hydrobiologia* **721**, 165-184 (2014).

48 Geffard, O. *et al.* Ovarian cycle and embryonic development in Gammarus fossarum: application for reproductive toxicity assessment. *Environ Toxicol Chem* **29**, 2249-2259 (2010).

49 Piscart, C. & Bollache, L. *Crustacés amphipodes de surface : gammares d'eau douce*. (Association Française de Limnologie, 2012).

50 Folmer, O., Black, M., Hoeh, W., Lutz, R. & Vrijenhoek, R. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Mol Mar Biol Biotechnol* **3**, 294-299 (1994).

51 Cogne, Y. *et al.* YC02. *Figshare* https://doi.org/10.6084/m9.figshare.c.4568087.v1 (2019).

52 Lagrue, C. *et al.* Confrontation of cryptic diversity and mate discrimination within Gammarus pulex and Gammarus fossarum species complexes. *Freshwater Biology* **59**, 2555-2570 (2014).

53 Gouy, M., Guindon, S. & Gascuel, O. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol* **27**, 221-224 (2010).

54 Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**, 644-652 (2011).

55 Smith-Unna, R., Boursnell, C., Patro, R., Hibberd, J. M. & Kelly, S. TransRate: reference-free quality assessment of de novo transcriptome assemblies. *Genome Res* **26**, 1134-1144 (2016).

56 Seppey, M., Manni, M. & Zdobnov, E. M. BUSCO: Assessing Genome Assembly and Annotation Completeness. *Methods Mol Biol* **1962**, 227-245 (2019).

57 NCBI Sequence Read Archive, http://identifiers.org/insdc.sra:SRR8089720 (2019).

58 NCBI Sequence Read Archive, http://identifiers.org/insdc.sra:SRR8089722 (2019).

59 NCBI Sequence Read Archive, http://identifiers.org/insdc.sra:SRR8089723 (2019).

60 NCBI Sequence Read Archive, http://identifiers.org/insdc.sra:SRR8089724 (2019).

61 NCBI Sequence Read Archive, http://identifiers.org/insdc.sra:SRR8089725 (2019).

62 NCBI Sequence Read Archive, http://identifiers.org/insdc.sra:SRR8089727 (2019).

63 NCBI Sequence Read Archive, http://identifiers.org/insdc.sra:SRR8089728 (2019).

64 NCBI Sequence Read Archive, http://identifiers.org/insdc.sra:SRR8089729 (2019).

65 NCBI Sequence Read Archive, http://identifiers.org/insdc.sra:SRR8089730 (2019).

66 NCBI Sequence Read Archive, http://identifiers.org/insdc.sra:SRR8089731 (2019).

67 NCBI Sequence Read Archive, http://identifiers.org/insdc.sra:SRR8089732 (2019).

68 NCBI Sequence Read Archive, http://identifiers.org/insdc.sra:SRR8089733 (2019).

69 NCBI Sequence Read Archive, http://identifiers.org/insdc.sra:SRR8089734 (2019).

70 NCBI Sequence Read Archive, http://identifiers.org/insdc.sra:SRR8089735 (2019).

71 GenBank, https://identifiers.org/ncbi/insdc:GHCN01000000 (2019).

72 GenBank, https://identifiers.org/ncbi/insdc:GHCP01000000 (2019).

73 GenBank, https://identifiers.org/ncbi/insdc:GHCQ01000000 (2019).

74 GenBank, https://identifiers.org/ncbi/insdc:GHCR01000000 (2019).

75 GenBank, https://identifiers.org/ncbi/insdc:GHCT01000000 (2019).

76 GenBank, https://identifiers.org/ncbi/insdc:GHCU01000000 (2019).

77 GenBank, https://identifiers.org/ncbi/insdc:GHCV01000000 (2019).

78 GenBank, https://identifiers.org/ncbi/insdc:GHCW01000000 (2019).

79 GenBank, https://identifiers.org/ncbi/insdc:GHCX01000000 (2019).

80 GenBank, https://identifiers.org/ncbi/insdc:GHCY01000000 (2019).

81 GenBank, https://identifiers.org/ncbi/insdc:GHCZ01000000 (2019).

82 GenBank, https://identifiers.org/ncbi/insdc:GHDA01000000 (2019).

83 GenBank, https://identifiers.org/ncbi/insdc:GHDB01000000 (2019).

84 GenBank, https://identifiers.org/ncbi/insdc:GHDC01000000 (2019)

Chapitre 3: Mise au point d'une méthodologie d'étude de la variabilité inter-espèce au sein des Gammares orientée biomarqueurs.

Les biomarqueurs établis et validés dans les travaux des équipes de l'IRSTEA et du CEA pourraient être appliqués sur des individus prélevés *in natura* afin d'établir un profil de santé des environnements aquatiques en Europe, en utilisant les Gammares comme référence. Pour cela, il est essentiel de valider la présence des séquences de biomarqueurs dans l'ensemble des espèces répandues en Europe. A cet effet, l'ensemble des assemblages réalisés au cours de cette thèse a été exploité afin de rechercher si les séquences de biomarqueurs sont conservées. Une méthodologie de recherche de ces séquences au niveau des assemblages transcriptomiques issus du RNAseq a dû être mise au point afin de trouver les séquences codant pour les protéines orthologues à celles contenant les biomarqueurs puis d'y rechercher la séquence des peptides à doser par protéomique ciblée.

Dans le cas où ces peptides ne peuvent être retrouvés, il est nécessaire de proposer la séquence du peptide biomarqueur variant présent dans le groupe taxonomique visé pour permettre de futures études écotoxicologiques sur ce groupe spécifique. La méthodologie mise au point dans ce chapitre se propose de fournir une première liste de peptides pour des validations expérimentales ultérieures. De plus, la détection de séquences peptidiques spécifiques peut permettre une distinction taxonomique des organismes sans nécessité de génotypage lors de l'analyse du protéome. La limite de cet exercice réside dans le nombre limité de données RNAseq, limite qui pourrait être levée par une analyse systématique des protéomes de chaque groupe taxonomique. Ce manuscrit va être soumis au journal Environmental Science & Technology.

Defining molecular biomarkers across a large panel of Gammarids for environmental biomonitoring

Yannick Cogne[1], Duarte Gouveia[1], Jean-Charles Gaillard[1], Olivier Pible[1], Davide Degli-Esposti[2], Olivier Geffard[2], Arnaud Chaumot[2], Jean Armengaud[1], Christine Almunia[1#]

[1]Laboratoire Innovations technologiques pour la Détection et le Diagnostic (Li2D), Service de Pharmacologie et Immunoanalyse (SPI), CEA, INRA, F-30207 Bagnols-sur-Cèze, France. [2]Irstea, UR MALY Laboratoire d'écotoxicologie, centre de Lyon-Villeurbanne, F-69625 Villeurbanne, France.

[#]**Corresponding author:** Christine Almunia, CEA-Marcoule, DRF-Li2D, Laboratory "Innovative technologies for Detection and Diagnostics", BP 17171, F-30200 Bagnols-sur-Cèze, France; christine.almunia@cea.fr; Tel: +00 33 4 66 79 13 01

Word count: 5,106; Characters count: 29,355; Running title: *Gammaridae* molecular biomarkers

I. Abstract:

Freshwater biomonitoring based on caging of *Gammarus fossarum* as sensitive animal sentinel represents an interesting tool for assessing water quality. In this context, relevant biomarker peptides from *Gammarus fossarum B* have been selected to estimate molt delay and other key physiological traits. A script was developed to automatically search for the presence of these biomarker peptides encoded in the transcriptomes from 14 individual animals belonging to four *Gammarus* species and one *Echinogammarus* species of the *Gammaridae* family. The search results indicated the relevance of these biomarkers in each given species since they were found in their respective translated transcriptome. In the case of absence of an identical sequence, a peptide substitute is proposed that can be monitored by tandem mass spectrometry for reliable quantitation of the biomarker. Interestingly, some peptide sequences derived from proteins involved in osmoregulation were shown to be well conserved across the different species highlighting the strong constraint on the corresponding molecular mechanisms.

II. Introduction:

Active biomonitoring is increasingly used for surveying European freshwaters. For this, the use of encaged amphipod *Gammarus fossarum* offers the advantage to be sensitive to a large range of pollutants. Up to date, diagnosis of the quality of aquatic environments is based on the analysis of biometric measurement including fertility, molting cycle, oocyte growth to reveal the presence of endocrine disruptors and leaf consumption and acetylcholinesterase (AchE) level for evaluation of neurotoxic contaminants. Disadvantages of using such markers of life traits are the difficulties in terms of standardization and the sensitivity of the measurement. Consequently, efforts were made to identify and monitor specific molecular protein biomarkers by mass spectrometry. The general strategy included a discovery phase and a validation phase. First, whole organism and specific organ proteomes of *Gammarus fossarum* B exposed or not exposed to contaminants in laboratory conditions were analyzed by shotgun mass spectrometry. These experimental data were interpreted by differential proteogenomics to select the potential biomarkers and the most adapted reporter-peptides[1, 2]. In the second phase, these peptide-biomarkers were evaluated by targeted mass spectrometry performed on proteins extracted from bred

Gammarids treated and non-treated in laboratory conditions or maintained for several days *in natura*[3, 4]. This two-stage procedure enabled to define a list of thirty-eight reporter peptides for twenty-six proteins which are representatives of key biological processes, such as molting, immunity, hormonal regulation, osmolarity and detoxification according to their functional annotation. Based on a multiplex assay of these reporter-peptides and their temporal follow-up, the risk assessment of water pollution can be reliably evaluated on standardized organisms from the same subspecies[5]. Moreover, it can help to characterize the type of contamination according to the biological process imbalanced in response to contaminants. For instance, in the case of significant modulation of the abundance of vitellogenins, it could be suspected that endocrine disruptors would be the predominant pollutants[6]. However, before using this novel approach to monitor freshwater, further work has to be done. Firstly, the range of cutoff concentration values between clean and polluted sites of each protein reporter-peptide has to be settled. Secondly, the application of these peptides has to be challenged among the *Gammarus* species biodiversity, in view of the molecular divergence between species due to a higher diversification rates and a longer evolutionary times[7-10].

To examine the transferability of peptides to different *Gammarus* species, individual transcriptomes of a male and a female from seven taxonomical groups of *Gammarids* were sequenced (Cogne et al. 2019). *Gammarus fossarum* A, *Gammarus fossarum* B, *Gammarus fossarum* C, *Gammarus wautieri*, *Gammarus pulex*, *Echinogammarus berilloni* and *Gammarus marinus* were chosen to include both far and closely-taxonomically related species. In the present work, we systematically verified the presence of peptides representative of a set of key biomarkers in order to establish their universal nature amongst Gammaridae. We developed a script that carry out this verification automatically and propose a peptide-sequence substitute in case of absence in a given taxonomical group.

III. Methods and materials

A. Databases accessions

Database accessions for the 14 transcriptomes used in this study are indicated in Table 1. CDSs and proteins are available from Cogne et al. (Cogne et al. 2019).

B. Search for biomarkers in assemblies

The strategy for the selection of universal peptide-biomarkers among the *Gammaridae* family is based on the use of the peptides previously identified with *Gammarus fossarum B* as the reference sequence. The home-made python script that was developed to verify the presence of biomarkers in each species is available at https://github.com/YannickCogne/BAITS. The script takes as input a table including the description of the biomarkers (protein, identifier, peptide sequence) and a fasta file which contains the reference gene sequences encoding the peptide biomarkers. The script needs also as input for each species the fasta files listing all the CDSs and their corresponding protein sequences. No parameters are available for modification without coding step. As output, the script produce the following tables in TSV format: liste_pept.tab (list of all peptides found by the script), Table_pept.tab (table indicating the presence of the initial biomarkers and the potential substitute), Table_pept_ident.tab (table with the identity score between pairwise alignment area and biomarkers), Table_pept_tryp.tab (table for checking tryptic sites), Table_prot.tab (table with proteins corresponding to the initial peptides or their substitutes).

C. Proteome extraction and trypsin proteolysis

Three male organisms from four different taxonomical groups (*G. fossarum B*, *G. pulex*, *G. marinus*, and *E. berilloni*) were analyzed individually by shotgun proteomics. Protein extraction was done as previously described[11]. Briefly, each organism was mechanically homogenized by bead-beating (one 3.2 mm steel bead per tube) in 20 µl of LDS sample buffer (Invitrogen) per milligram of organism. The homogenates were centrifuged at 10,000g

for 3 min in order to pellet cellular debris, and the resulting supernatant collected into a new tube. Samples were then incubated for 5 min at 99˚C. A 20 µL aliquot of each sample was then subjected to SDS-PAGE for a short electrophoretic migration. The whole-protein content from each well was extracted as a single polyacrylamide band, which was then submitted to the proteolysis process with trypsin (Roche) using 0.01% ProteaseMAX surfactant (Promega) as previously described [12].

D. Tandem mass spectrometry analysis

The peptide mixtures were then analyzed in data-dependent mode with a Q-Exactive HF mass spectrometer (Thermofisher) coupled with an UltiMate 3000 LC system (Dionex-LC Packings) operated essentially as described[13]. Peptides were desalted and then resolved onto a nanoscale C18 PepMapTM 100 capillary column (LC Packings) with a 90-min gradient of CH3CN, 0.1% formic acid, at a flow rate of 0.2 µL/min. Following the Top20 method, full scan mass spectra were acquired from *m/z* 350 to 1800 at a resolution of 60,000 and MS/MS spectra from *m/z* 200 to 2000 at a resolution of 15,000. Ion selection for MS/MS fragmentation and measurement was performed applying a dynamic exclusion window of 10 sec.

E. MS/MS spectra interpretation

MS/MS spectra were assigned to peptide sequences by the MASCOT Daemon 2.3.2 search engine (Matrix Science) searching against the corresponding translated CDS database of *Echinogammrus berilloni*, *Gammarus marinus*, *Gammarus pulex* and *Gammarus fossarum* B. This assignation was done with the following parameters: full-trypsin specificity, maximum of two missed cleavages, mass tolerances of 5 ppm on the parent ion and 0.02 Da on the MS/MS, carboxyamidomethylated cysteine (+57.0215) as a fixed modification, oxidized methionine (+15.9949) and deamidation of asparagine and glutamine (+0.9848) as variable modifications. All peptide matches presenting a MASCOT peptide score with a p-value of less than 0.05 were filtered and assigned to a protein without parsimony to avoid any bias between isoforms. Complete results are in the supplementary data file 3.

IV. Results and discussion

A. Species selection for the study of biomarker transversality

Among the 304 species of the *Gammaridae* family[14], seven representative taxonomical groups were chosen on the basis of their occurrence in European aquatic environment: *G. fossarum* A (GFA), *G. fossarum* B (GFB), *G. fossarum* C (GFC), *G. wautieri (GW)* , *G. pulex (GPC)*, *E. berilloni (EGS)* and *G. marinus (EGU)*. Male and female transcriptomes have been assembled separately and are distinguished by the last letter (M and F, respectively). Four groups of biomarkers have been defined previously[4]: osmoregulation (with 3 isoforms of Na+/K+ ATPase, molt-related proteins (5 polypeptides), general biomarkers (11 proteins), and female specific yolk proteins (7 polypeptides). The 38 reporter-peptides for monitoring the abundance of each of these proteins have been selected from *Gammarus fossarum* B peptidome[4]. We checked whether these peptides were encoded in the RNAseq transcriptome assembly of the seven taxonomical groups.

B. Peptide biomarker search strategy

Figure 1 shows the nine steps for searching peptide biomarkers in the *de novo* assembled transcriptomes.

Step 1: Preparation of the nucleotide bait: CDS containing the biomarker sequence, considered as the reference sequence, was extracted from the reference database (GFOSS) to build a BLASTn database.

Step 2: Fishing in the ocean of species-assemblies: CDS homologs of the reference sequence were searched in each of the 14 assemblies, using the reference database and the BLASTn tool with default parameters.

Step 3: Selection of the best fishes: All the CDS homologs found were in silico translated and to select the best CDS homologs for a given

assembly, a pairwise global alignment was performed with no identity matrix (an identity scores for 1) and a high penalty for opening gaps (set as 15). The sequences with the higher alignment score were selected for each assembly and then in silico translated into proteins.

Step 4: Validation of the compliance between the best fishes and the biomarker: The biomarker sequence was searched in the results of the step 3. If the peptide-biomarker is found in one assembly, the search process stops for this assembly. When the peptide is not found in an assembly, the process went on with the step 5.

Step 5: Preparation of the protein bait: Translated CDS, defined as best in step 3, were extracted from each assembly to build a BLASTp database, which will be used as the new reference database.

Step 6: Re-fishing for uncompliant fishes: Translated CDS homologs were searched in each of the remaining assemblies from the step 4, using the new reference database build from the step 5 and the BLASTp tool with default parameters.

Step 7: Selection of potential new fishes: the translated CDS homologs, with the higher identity score in a given assembly were selected using a pairwise global alignment as described in step 3.

Step 8: Validation of the compliance between the new fishes and the biomarker: The biomarker sequence was searched in the results of the step 7. If the peptide-biomarker is found in one assembly, the search process stops for this assembly, if not, the script enables to go further for finding substitute biomarkers.

Step 9: Suggestion of a substitute biomarker. The substitute biomarkers were extracted from the region selected from the pairwise alignment between the biomarker and the translated CDS having the higher identity score (see step 7) and framed with trypsic sites. The strategy of this search is presented in Figure 1. We first established the most relevant similar sequences at the contig level in each taxonomical groups by sequence similarity search taking the nucleotidic sequence from *Gammarus fossarum* B encompassing the peptide information as query. The resulting nucleotidic sequence hits

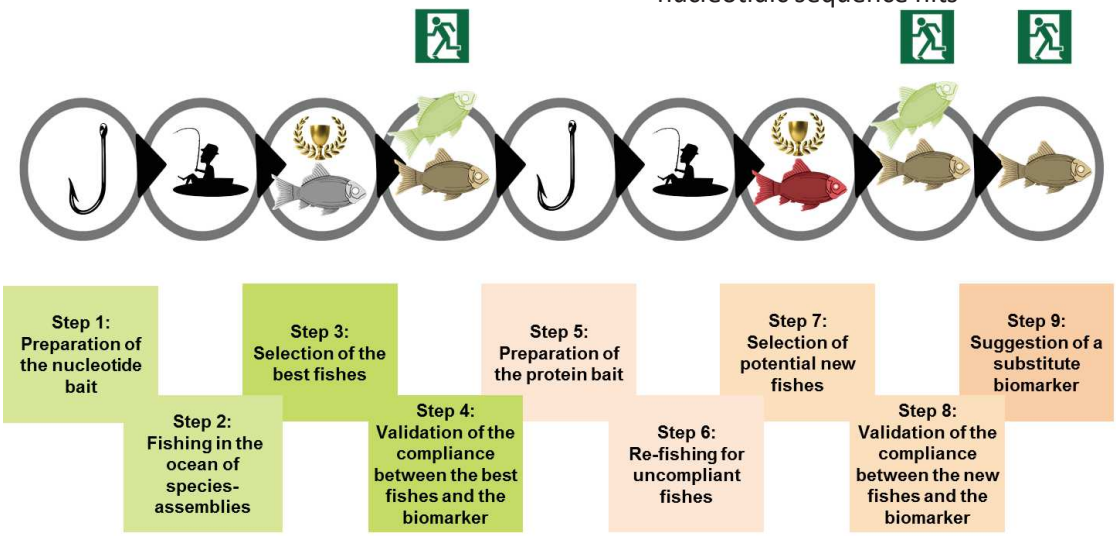


Figure 1: **BAITS for biomarker search**. Description of the steps for searching biomarkers or substitute in transcriptome CDS and their sequence translation. The Exit icon represents each step where the search is stopped after the given biomarker was found. The green squares indicate the step that use nucleotide sequences from the GFOSS reference database and the salmon squares, the steps that use the sequences obtained after the first search in all the transcriptome for a given biomarker. Green fishes are the proteins of each species which contain the reference peptide sequence and the salmon colored fishes are the proteins containing the peptide substitute of each species.

C. Results on biomarkers

The search results are summarized in **Figure 2** where the presence or absence of every peptide-biomarker in each species-specific protein is indicated with different colors. Further data are available in the Supplementary Data File 1 such as the name of the contigs encompassing the peptide coding information. The potential of the methodology is first illustrated with the catalase biomarker as example. As shown in **Figure 2**, four reporter peptides have been selected for this protein of interest: ADPALGQAIQER, LADNIAGHVINTQEFIR, NLPADQAAALASSDPDYAIR, and LGSNFLQIPVNCPYR, which initially originated in the reference database GFOSS, from two polypeptides, protein ID-110912 and protein ID-45375[15], previously shown to be down-regulated after *G. fossarum* was exposed to cadmium or lead, in laboratory conditions[3]. Here, because the quality of assembly is better when using a single organism (Cogne et al., 2019) compared to a transcriptome done on a pool of organisms (Trapp et al., 2014), only one single catalase protein is mentioned per organism. The reporter peptides of catalases, selected on the basis on the proteome exploration of *G. fossarum* B, do not suit their final use for *G. pulex*, *G. marinus* and the *E. berilloni* as they were not found in their respective proteome. Furthermore, the search against the 14 individual transcriptomes allowed us to select 20 protein sequences that slightly differ. Only one isoform was detected in *G. fossarum B, G. fossarum C, G. wautieri, E. berilloni* and *G. marinus*, but two were observed in *G. pulex*, and more than two were detected in *G. fossarum A*. **Figure 3** shows a partial view of the multi-alignment of all these isoforms highlighting the four conserved peptide sequences in the second half of the protein [amino acid 244 to amino acid 496]. This multi-alignment highlights the possibilities to select isoform using all peptides except ADPALGQAIQER for *G. fossarum A* and *G. pulex*. In most case, substitute peptides are conserved in terms of tryptic cleavage site except for *E. berilloni* peptide ADPALGQAIQER which has the last amino acid R replaced by Q amino acid and a tryptic site found 3 amino acids further downstream. Another exception found is for the peptide NLPADQAAALASSDPDYAIR which has a substitute with another tryptic site in *E. berilloni* and *G. pulex* isoforms. The full length catalases alignment is described with the supplementary figure 1.

1. Osmoregulation

In contrast to the reporter-peptides of female-yolk-proteins- showing a high diversity between species, almost all the other selected reporter-peptides derived from proteins involved in the osmoregulation process were conserved among the different species. There was only one exception with the peptide LQTNPDTGLSTAEAR, from the GFOSS proteinID-209438, which was not found in *E. berilloni*. This result shows that the selected peptides are conserved among the studied species, allowing their universal use for mass spectrometry-based biomonitoring.

2. Molt process

In the same way, among the nine peptides selected to be the reporter of molt-related proteins process, only one peptide sequence, AFWGSLPLR, derived from the protein ID-144144 was found to be universal among the studied *Gammarids* and six are common among the five *Gammarus* species, except *E. berilloni*. This peptide is very important for growth monitoring since the juvenile hormone carboxylesterase, from which it is derived, is involved in the regulation network of a number of physiological processes such as molting, metamorphosis, and reproduction. It was previously demonstrated that it was down modulated in contaminated sites with the same amplitude as the second reporter peptide of the JHE-like carboxylesterase[4]. On the other hand, the two other peptides selected for the molt monitoring, ILTTMWADFAR reporter of the

protein ID-144144 annotated as JHE-like Carboxylesterase, and LVLGTATYGR reporter of the protein ID-181833 annotated as a chitinase, were not found after their search in *E. berilloni* and *G. marinus* proteomes. Unexpectedly, the peptide LVLGTATYGR was

not found in the proteome of the *G. pulex* male whereas it was detected in the female proteome. The hypothesis is that the amino acid sequence of this peptide could vary between individuals of the same species. Further study with a high number of organisms is required to confirm this hypothesis. The same result was obtained with the search of the peptide GIDIIGDAFEADR extracted from the protein ID-2562 annotated as prophenoloxidase. This peptide was not found in male proteome of *G. fossarum* C and in male and female proteomes of *G. fossarum* A. The other peptides from the protein ID-2562, APILEGYFSK, GIDFGTTQSVR and ATQPSYTVAQLELPGVNITR, are suitable to monitor the protein level in all the species of *Gammarus*, their response to contaminants being demonstrated to evolve in correlation[5].

1. Other general functions

Transglutaminases are well known to be also involved in the innate immune response to pathogens and in blood coagulation[16]. Thus, complemented by the measurement of reporter-peptide from prophenoloxidase, transglutaminases help to distinguish viral and bacterial diseases from contaminant effects. Two peptides, VLAVDILAK and GTLAVIPVQNR, were derived from protein ID-7169 and protein ID-1917, respectively, both annotated as transglutaminases. According to the Smith-Waterman algorithm, these two proteins showed 49.7% of sequence identity. However, these peptides are not present in the *E. berilloni* and *G. marinus*. On the other hand, in the case of chitinase-like protein, there is no other perspective to monitor them with every studied species, except by choosing other peptide signatures specific for *E. berilloni* and *G. marinus*. Chitinase is crucial for the initiation of the molting process, thus this enzyme is considered as a relevant biomarker for detecting a molt delay or acceleration. Endocrine disruptor, like pesticides, are not the only contaminants able to influence the molting cycle. Metals such as Cadmium, lead, zinc, mercury and selenium were demonstrated to delay the molting process[17].

For making distinction between endocrine disruptors and metal effects on molting cycle, it is required to observe the oxidative stress level. Among the proteins annotated as protein effectors of the oxidative stress such as catalase and glutathion-S-transferase (GST), a set of six reporter peptides were selected. Moreover, among the selected proteins of anti-oxidative defense, the two reporter-peptides of GST could not be considered as universal. They are not applicable to E. berilloni, *G. marinus* and *G. pulex*. To summarize, monitoring the oxidative stress through the peptides selected will be possible for only part of the *Gammarus* genus, including the species *G. fossarum and G. wautieri* but excluding *G. marinus* while these last two species are phylogenetically very close18. For the other species, the search of peptide substitutes must be performed. Here, the peptide, reporters of the protein ID—100125, annotated as a cytochrome P450, distinguishes *G. fossarum* and *G. wautieri*. Indeed, the peptide ILEDFVDVFNR was found in the *G. fossarum* and *G. pulex* proteomes whereas it was absent in those of *G. wautieri*.

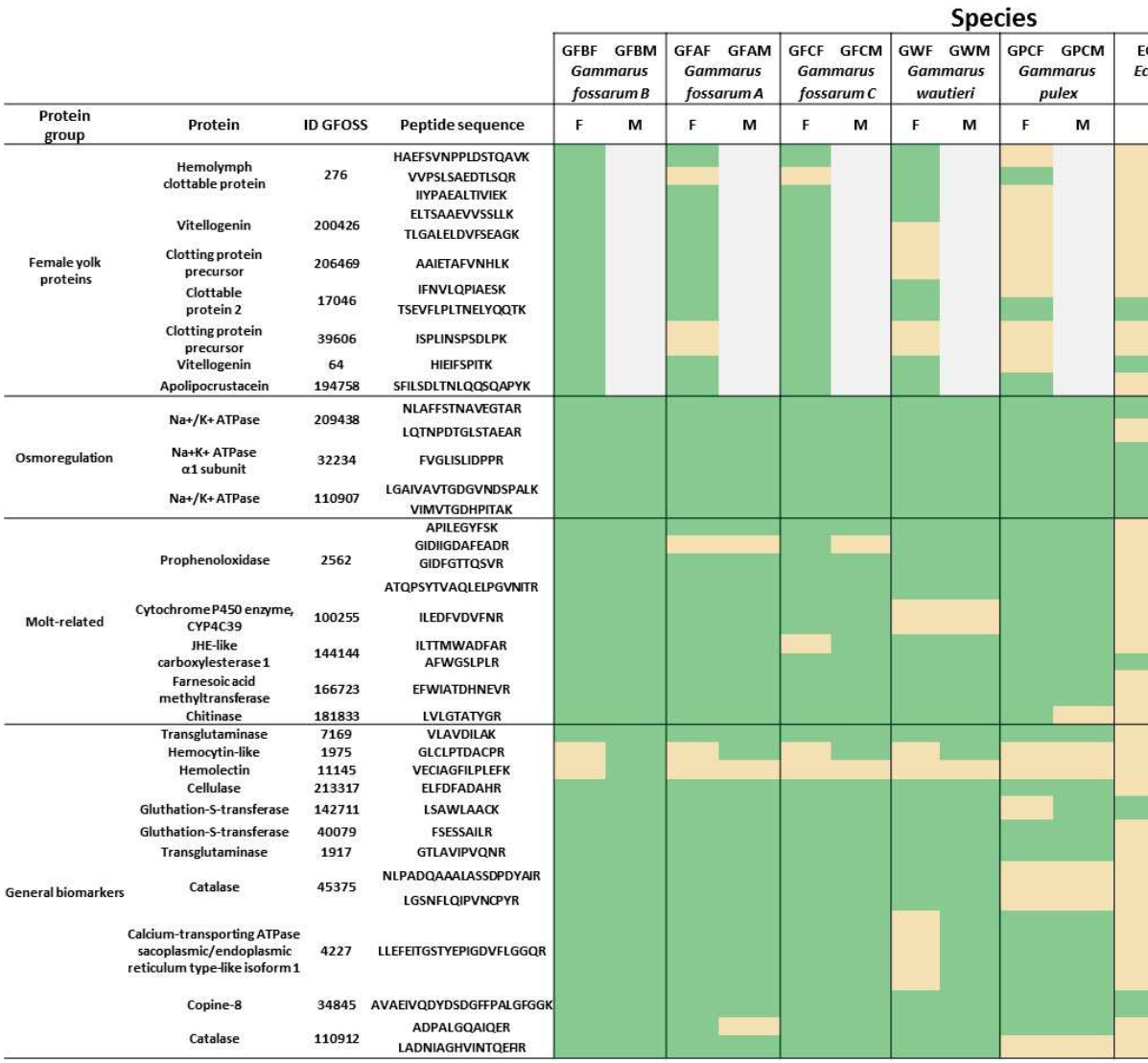

Species
GFBF GFBM
Gammarus fossarum B
GFAF GFAM
Gammarus fossarum A
GFCF GFCM
Gammarus fossarum C
GWF GWM
Gammarus wautieri
GPCF GPCM
Gammarus pulex
F M F M F M F M F M
Protein group
Protein
ID GFOSS
Peptide sequence
Female yolk proteins
Hemolymph clottable protein
276
HAEFSVNPPLDSTQAVK
VVPSLSAEDTLSQR
Vitellogenin
200426
IIYPAEALTIVIEK
ELTSAAEVVSSLLK
TLGALELDVFSEAGK
Clotting protein precursor
206469
AAIETAFVNHLK
Clottable protein 2
17046
IFNVLQPIAESK
TSEVFLPLTNELYQQTK
Clotting protein precursor
39606
ISPLINSPSDLPK
Vitellogenin
64
HIEIFSPITK
Apolipocrustacein
194758
SFILSDLTNLQQSQAPYK
Osmoregulation
Na+/K+ ATPase
209438
NLAFFSTNAVEGTAR
LQTNPDTGLSTAEAR
Na+K+ ATPase α1 subunit
32234
FVGLISLIDPPR
Na+/K+ ATPase
110907
LGAIVAVTGDGVNDSPALK
VIMVTGDHPITAK
Molt-related
Prophenoloxidase
2562
APILEGYFSK
GIDIIGDAFEADR
GIDFGTTQSVR
ATQPSYTVAQLELPGVNITR
Cytochrome P450 enzyme, CYP4C39
100255
ILEDFVDVFNR
JHE-like carboxylesterase 1
144144
ILTTMWADFAR
AFWGSLPLR
Farnesoic acid methyltransferase
166723
EFWIATDHNEVR
Chitinase
181833
LVLGTATYGR
General biomarkers
Transglutaminase
7169
VLAVDILAK
Hemocytin-like
1975
GLCLPTDACPR
Hemolectin
11145
VECIAGFILPLEFK
Cellulase
213317
ELFDFADAHR
Gluthation-S-transferase
142711
LSAWLAACK
Gluthation-S-transferase
40079
FSESSAILR
Transglutaminase
1917
GTLAVIPVQNR
Catalase
45375
NLPADQAAALASSDPDYAIR
LGSNFLQIPVNCPYR
Calcium-transporting ATPase sacoplasmic/endoplasmic reticulum type-like isoform 1
4227
LLEFEITGSTYEPIGDVFLGGQR
Copine-8
34845
AVAEIVQDYDSDGFFPALGFGGK
Catalase
110912
ADPALGQAIQER
LADNIAGHVINTQEFIR

Figure 2: **Biomarkers conservation amongst taxonomical groups**. The BAITS results are summarized by the P assembly of the reference biomarkers previously established from experiments performed with *Gammarus Fossarum E* control. The green color indicates the conservation of the biomarker sequence and the salmon color indicates the abs given transcriptome assembly. In the latter case, substitute were purposed. The grey color indicates that the peptide are very low expressed in male. The male *Gammarus fossarum* A was referenced as GFAM and female GFAF, the m GFBF, *G. fossarum* C was shortened as GFCM and GFCF for male and female respectively, *Gammarus wautieri*, as *Gammarus pulex*, GPM and GPF for male and female respectively. *Echinogammarus berilloni* male was named EGSM *marinus*, the male was referenced as EGUM and the female EGUF.

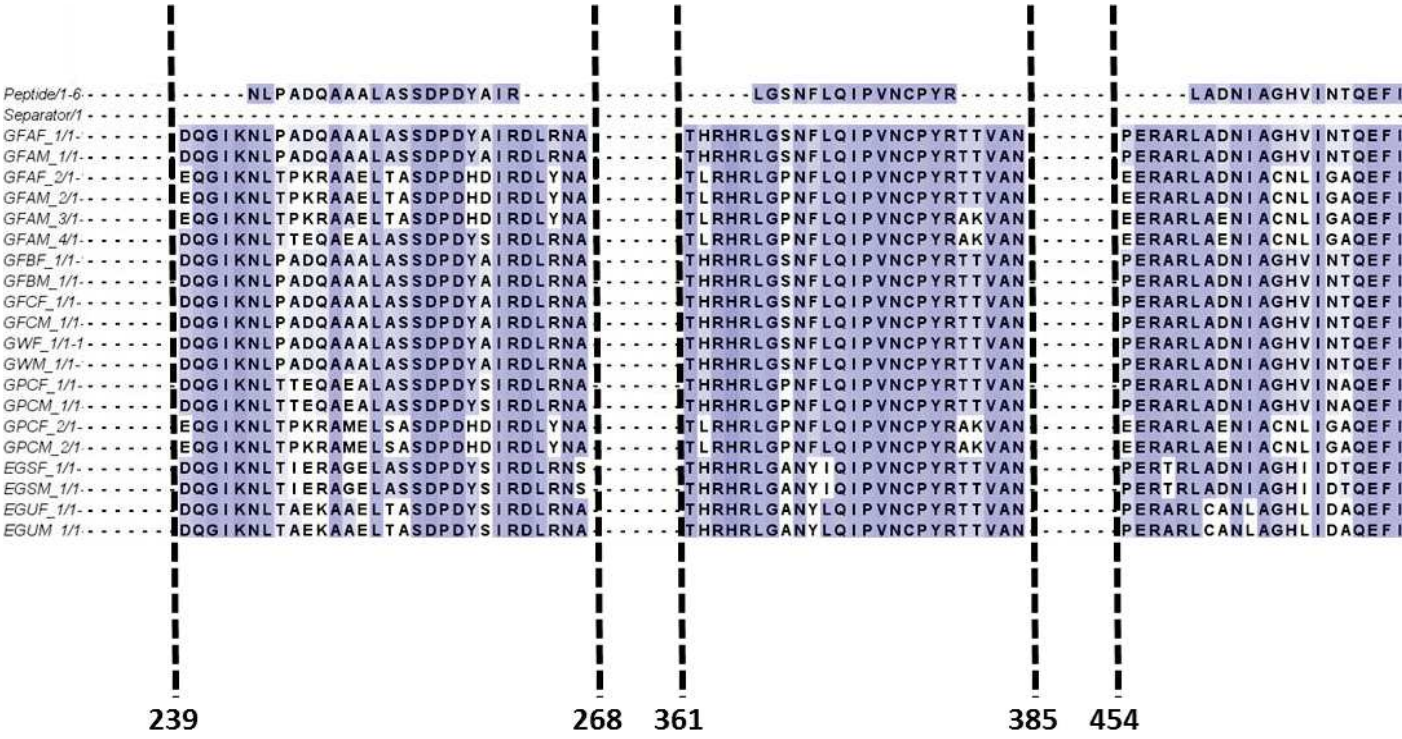


Figure 3: **Alignment of the catalase sequences from all the species and biomarker location**. Conserved amin
gradation indicating the amino acid variation takes into account the biochemical properties of each amino acid.

2. Vitellogenins

The selected peptide-biomarkers from the vitellogenin-like-proteins were only detected in the female transcriptome, as these proteins being naturally low expressed in male [19]. In the context of ecotoxicology, these proteins were identified as relevant biomarkers of endocrine disruption. Among the eleven peptide-biomarkers selected for this protein family, none is shared between every *Gammaridae* species. The peptide TSEVFLPLTNELYQQTK derived from the GFOSS vitellogenin-like protein ID-17046 annotated as clottable protein 2 is the only one usable as a universal biomarker among the female freshwater species. On the contrary, the peptide was not found in the proteome of the species *G. marinus*, which is a marine species. This peptide is a key peptide as it was demonstrated as a reliable signature of the presence of contaminant[4] and a reporter peptide of the most abundant yolk protein in female proteomes. This peptide was shown sensitive to pollutants and indicative of each reproductive cycle stage[20]. In the female yolk-protein group, no peptide is specific of a given species. Among the eleven reporter peptides selected, nine could be used to monitor freshwater with the *G. fossarum* species without subtype distinction. On the contrary, the peptide VVPSLSAEDTLSQR, derived from the protein ID-276, was found in the proteome of the species subtype B of *G. fossarum*, *G. wautieri* and *G. pulex*. Remarkably, this peptide discriminates *G. fossarum* subtypes.

D. Substitute peptides and proteomics validation

In the case where the peptide biomarker is not conserved in one species, substitute biomarkers were searched and selected in the same region from the most similar sequences. The substitute is proposed taking care of the possible change of trypsin cleavage but its detectability in tandem mass spectrometry should be verified experimentally. The automatized script provides the peptide sequences derived from the similar protein sequence at the same locus but respecting the trypsin proteolysis pattern. Such substitute peptides were found for all the proteins as shown in Supplementary Data 1.

As shown in Table 1, peptide substitutes were proposed for monitoring the cellulase biomarker in *E. berilloni* and *G. marinus*. The reporter peptide ELFDFADAHR was replaced by the peptide ELFDFADSYR for the former and the peptide ELFDFADNYR for the later organism. These two substitutes each differ from the reference sequence from *G. fossarum B* by two amino acids. The localization of the peptide biomarker does not change in the protein of each species (see the corresponding alignment of cellulase homologue sequences in the supplementary file 2). We checked whether these two peptides could be detected in animal experimental proteomes. For this, we recorded the proteome by label-free shotgun approach on three animals treated individually for 4 different species. An average of 38634 MS/MS spectra were recorded per animal and among these 30.8% were peptide-to-spectrum matches. The abundances of the specific peptide biomarker were assessed by their spectral counts and are reported in Table 2. The peptide substitutes could be detected in *E. berilloni* and *G. marinus*, but their level of detection by mass spectrometry slightly changed according to species, indicating a lower expression of the cellulase in *G. marinus*, a marine species. Here no cross-detection of the peptide substitute was found for each species.

V. Conclusion

As expected, conservation of biomarker peptides is decreasing along the phylogenetic distance between the considered species and the *G. fossarum B* reference. Interestingly, some peptides have a strong divergence, such as those from the vitellogenin-like-protein family, enabling a clear distinction between species. Remarkably, two peptides are able to distinguish among the three subtypes of G. *fossarum*. Thus, we defined here a set of specific peptides to simultaneously distinguish *Gammarids* species and quantify several key physiological traits. As an example, the peptides specific of osmoregulation and molting cycle can be analyzed together to monitor the moult stages as molting cycle and osmoregulation are tightly

linked[21]. In conclusion, the method developed here for the analysis of the transferability of biomarkers among species, selected on the basis of one species, enabled to verify the relevance of peptide biomarker for a given species and to find peptide substitutes for their biomonitoring by tandem mass spectrometry. Moreover, we observed that proteins involved in osmoregulation are well conserved among the Gammarid species whereas a high diversity among those belonging to the vitellogenin family was observed. The homemade developed script is applicable for any transcriptomic datasets and is available at https://github.com/YannickCogne/BAITS, associated to its documentation.

Table 1: Transcriptome accession and code name for each individual assembly.

	Female		Male	
Species	Code Name	Transcriptome accession	Code Name	Transcriptome accession
Echinogammarus berilloni	EGSF	GHCT01000000	EGSM	GHCU01000000
Echinogammarus marinus	EGUF	GHCW01000000	EGUM	GHCV01000000
Gammarus fossarum A	GFAF	GHCX01000000	GFAM	GHCY01000000
Gammarus fossarum B	GFBF	GHCZ01000000	GFBM	GHDA01000000
Gammarus fossarum C	GFCF	GHDC01000000	GFCM	GHDB01000000
Gammarus pulex	GPCF	GHCP01000000	GPCM	GHCQ01000000
Gammarus wautieri	GWF	GHCR01000000	GWM	GHCN01000000

Table 2: Detection of the cellulose peptide substitutes by shotgun tandem mass spectrometry.

Species	Peptide sequence	Peptide mass	Spectral counts per animal		
			#1	#2	#3
Gammarus fossarum B	ELFDFADAHR	1220,569	9	8	6
Gammarus pulex	ELFDFADAHR	1220,569	8	8	7
Echinogammarus berilloni	ELFDFADSYR	1262,568	2	4	3
Gammarus marinus	ELFDFADNYR	1289,579	0	1	1

VI. Acknowledgements

We thank the Institut National de Recherche en Sciences et Technologies pour l'Environnement et l'Agriculture (France), the Commissariat à l'Energie Atomique et aux Energies Alternatives (France), and the Agence Nationale de la Recherche program "ProteoGam" (ANR-14-CE21-0006-02) for financial support.

VII. References

1. Trapp, J.; Armengaud, J.; Pible, O.; Gaillard, J. C.; Abbaci, K.; Habtoul, Y.; Chaumot, A.; Geffard, O., Proteomic investigation of male Gammarus fossarum, a freshwater crustacean, in response to endocrine disruptors. *J Proteome Res* **2015,** *14*, (1), 292-303.

2. Trapp, J.; Gouveia, D.; Almunia, C.; Pible, O.; Esposti, D. D.; Gaillard, J. C.; Chaumot, A.; Geffard, O.; Armengaud, J., Digging deeper into the pyriproxyfen-response of the amphipod Gammarus fossarum with a next-generation ultra-high-field orbitrap analyser: New perspectives for environmental toxicoproteomics. *Frontiers in Environmental Science* **2018,** *6*, (JUN).

3. Gouveia, D.; Chaumot, A.; Charnot, A.; Queau, H.; Armengaud, J.; Almunia, C.; Salvador, A.; Geffard, O., Assessing the relevance of a multiplexed methodology for proteomic biomarker measurement in the invertebrate species Gammarus fossarum: A physiological and ecotoxicological study. *Aquat Toxicol* **2017,** *190*, 199-209.

4. Gouveia, D.; Chaumot, A.; Charnot, A.; Almunia, C.; Francois, A.; Navarro, L.; Armengaud, J.; Salvador, A.; Geffard, O., Ecotoxico-Proteomics for Aquatic Environmental Monitoring: First in Situ Application of a New Proteomics-Based Multibiomarker Assay Using Caged Amphipods. *Environ Sci Technol* **2017,** *51*, (22), 13417-13426.

5. Charnot, A.; Gouveia, D.; Armengaud, J.; Almunia, C.; Chaumot, A.; Lemoine, J.; Geffard, O.; Salvador, A., Multiplexed assay for protein quantitation in the invertebrate Gammarus fossarum by liquid chromatography coupled to tandem mass spectrometry. *Anal Bioanal Chem* **2017,** *409*, (16), 3969-3991.

6. Gouveia, D.; Almunia, C.; Cogne, Y.; Pible, O.; Degli-Esposti, D.; Salvador, A.; Cristobal, S.; Sheehan, D.; Chaumot, A.; Geffard, O.; Armengaud, J., Ecotoxicoproteomics: A decade of progress in our understanding of anthropogenic impact on the environment. *Journal of Proteomics* **2018**.

7. Douzery, E. J.; Snell, E. A.; Bapteste, E.; Delsuc, F.; Philippe, H., The timing of eukaryotic evolution: does a relaxed molecular clock reconcile proteins and fossils? *Proc Natl Acad Sci U S A* **2004,** *101*, (43), 15386-91.

8. Dunn, C. W.; Hejnol, A.; Matus, D. Q.; Pang, K.; Browne, W. E.; Smith, S. A.; Seaver, E.; Rouse, G. W.; Obst, M.; Edgecombe, G. D.; Sorensen, M. V.; Haddock, S. H.; Schmidt-Rhaesa, A.; Okusu, A.; Kristensen, R. M.; Wheeler, W. C.; Martindale, M. Q.; Giribet, G., Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* **2008,** *452*, (7188), 745-9.

9. Peterson, K. J.; Lyons, J. B.; Nowak, K. S.; Takacs, C. M.; Wargo, M. J.; McPeek, M. A., Estimating metazoan divergence times with a molecular clock. *Proc Natl Acad Sci U S A* **2004,** *101*, (17), 6536-41.

10. Trapp, J.; Almunia, C.; Gaillard, J. C.; Pible, O.; Chaumot, A.; Geffard, O.; Armengaud, J., Proteogenomic insights into the core-proteome of female reproductive tissues from crustacean amphipods. *J Proteomics* **2016,** *135*, 51-61.

11. Cogne, Y.; Almunia, C.; Gouveia, D.; Pible, O.; François, A.; Degli-Esposti, D.; Geffard, O.; Armengaud, J.; Chaumot, A., Comparative proteomics in the wild: accounting for intrapopulation variability improves describing proteome response in a Gammarus pulex field population exposed to cadmium. *Aquatic Toxicology* **2019**, 105244.

12. Hartmann, E. M.; Allain, F.; Gaillard, J. C.; Pible, O.; Armengaud, J., Taking the shortcut for high-throughput shotgun proteomic analysis of bacteria. *Methods Mol Biol* **2014,** *1197*, 275-85.

13. Klein, G.; Mathe, C.; Biola-Clier, M.; Devineau, S.; Drouineau, E.; Hatem, E.; Marichal, L.; Alonso, B.; Gaillard, J. C.; Lagniel, G.; Armengaud, J.; Carriere, M.; Chedin, S.; Boulard, Y.; Pin, S.; Renault, J. P.; Aude, J. C.; Labarre, J., RNA-binding proteins are a major target of silica nanoparticles in cell extracts. *Nanotoxicology* **2016,** *10*, (10), 1555-1564.

14. Väinölä, R.; Witt, J. D. S.; Grabowski, M.; Bradbury, J. H.; Jazdzewski, K.; Sket, B., Global diversity of amphipods (Amphipoda; Crustacea) in freshwater. *Hydrobiologia* **2008,** *595*, (1), 241-255.

15. Trapp, J.; Geffard, O.; Imbert, G.; Gaillard, J. C.; Davin, A. H.; Chaumot, A.; Armengaud, J., Proteogenomics of Gammarus fossarum to document the reproductive system of amphipods. *Mol Cell Proteomics* **2014,** *13*, (12), 3612-25.

16. Wang, Z.; Wilhelmsson, C.; Hyrsl, P.; Loof, T. G.; Dobes, P.; Klupp, M.; Loseva, O.; Morgelin, M.; Ikle, J.; Cripps, R. M.; Herwald, H.; Theopold, U., Pathogen entrapment by transglutaminase--a conserved early innate immune mechanism. *PLoS Pathog* **2010,** *6*, (2), e1000763.

17. Hosamani N, R. S., Reddy RP, Crustacean Molting: Regulation and Effects of Environmental Toxicants. . *Journal of Marine Science: Research & Development* **2017,** *7*, (5).

18. Karaman, G. S.; Pinkster, S., Freshwater Gammarus Species from Europe, North-Africa and Adjacent Regions of Asia (Crustacea-Amphipoda) .1. Gammarus Pulex-Group and Related Species. *Bijdr Dierkd* **1977,** *47*, (1), 1-97.

19. Xuereb, B.; Bezin, L.; Chaumot, A.; Budzinski, H.; Augagneur, S.; Tutundjian, R.; Garric, J.; Geffard, O., Vitellogenin-like gene expression in freshwater amphipod Gammarus fossarum (Koch, 1835): functional characterization in females and potential for use as an endocrine disruption biomarker in males. *Ecotoxicology* **2011,** *20*, (6), 1286-99.

20. Trapp, J.; Armengaud, J.; Gaillard, J. C.; Pible, O.; Chaumot, A.; Geffard, O., High-throughput proteome dynamics for discovery of key proteins in sentinel species: Unsuspected vitellogenins diversity in the crustacean Gammarus fossarum. *J Proteomics* **2016,** *146*, 207-14.

21. Van Mai, H.; Fotedar, R., Haemolymph constituents and osmolality as functions of moult stage, body weight, and feeding status in marron, Cherax cainii () and yabbies, Cherax destructor (Clark, 1936). *Saudi J Biol Sci* **2018,** *25*, (4), 64.

Chapitre 4: Observation de la variabilité intra-population au sein d'une analyse *in situ* de l'impact environnemental

La baisse du coût des analyses en spectrométrie de masse permet désormais de l'appliquer à un grand nombre d'échantillons. L'intérêt est d'augmenter la sensibilité de détection dans des analyses de comparaison protéomique, afin de mieux cerner les différents mécanismes moléculaires mis en jeu au sein des organismes vivants. Différents mécanismes peuvent se mettre en place en réponse à des polluants. Pour pouvoir observer et augmenter la sensibilité de détection, nous avons travaillé sur l'exploitation de méthodes statistiques permettant de regrouper et différencier les échantillons de façon plus détaillées.

Dans le cas de cette étude nous avons sélectionné des *Gammarus pulex* provenant de deux rivières différentes, Pollon et Brameloup. Le site de Pollon est un site de référence considéré comme non pollué (site contrôle) alors que le site de Brameloup est un site soumis à une pollution au cadmium. Le protéome complet de 10 mâles et 10 femelles *Gammarus pulex* pour chacun des sites a été analysé, soit 40 animaux traités individuellement.

Les travaux de ce chapitre ont permis de mettre au point une méthodologie applicable à un plus grand nombre d'individus afin de faire des études de protéomique dites de population. Ces études à large échelle peuvent être considérées désormais comme un enjeu majeur en écotoxicoprotéomique. Ce manuscrit a été soumis et accepté dans le journal Aquatic ecotoxicology.

Comparative proteomics in the wild: accounting for intrapopulation variability improves describing proteome response in a *Gammarus pulex* field population exposed to cadmium

Yannick Cogne[1], Christine Almunia[1], Duarte Gouveia[1], Olivier Pible[1], Adeline François[2], Davide Degli-Esposti[2], Olivier Geffard[2], Jean Armengaud[1#], Arnaud Chaumot[2]

[1]Laboratoire Innovations technologiques pour la Détection et le Diagnostic (Li2D), Service de Pharmacologie et Immunoanalyse (SPI), CEA, INRA, F-30207 Bagnols-sur-Cèze, France.

[2]Irstea, UR RiverLy, Laboratoire d'écotoxicologie, centre de Lyon-Villeurbanne, F-69625 Villeurbanne, France.

[#]**Corresponding author :** Jean Armengaud, CEA-Marcoule, DRF-Li2D, Laboratory "Innovative technologies for Detection and Diagnostics", BP 17171, F-30200 Bagnols-sur-Cèze, France; jean.armengaud@cea.fr; Tel: +00 33 4 66 79 68 02; Fax: +00 33 4 66 79 19 05.

Word count: 5,143; Characters count: 30,611; Running title: Gammarid intrapopulation proteomic variability

ABSTRACT

High-throughput proteomics can be performed on animal sentinels for discovering key molecular biomarkers signing the physiological response and adaptation of organisms. Ecotoxicoproteomics is today amenable by means of proteogenomics to small arthropods such as Gammarids which are well known sentinels of aquatic environments. Here, we analysed two regional *Gammarus pulex* populations to characterize the potential proteome divergence induced in one site by natural bioavailable mono-metallic contamination (cadmium) compared to a non-contaminated site. Two RNAseq-derived protein sequence databases were established previously on male and female individuals sampled from the reference site. Here, individual proteomes were acquired on 10 male and 10 female paired organisms sampled from each site. Proteins involved in protein lipidation, carbohydrate metabolism, proteolysis, innate immunity, oxidative stress response and lipid transport were found more abundant in animals exposed to cadmium, while hemocyanins were found in lower abundance. The intrapopulation proteome variability of long-term exposed *G. pulex* was inflated relatively to the non-contaminated population. These results show that, while remaining a challenge for such organisms with not yet sequenced genomes, taking into account intrapopulation variability is important to better define the molecular players induced by toxic stress in a comparative field proteomics approach.

Keywords: proteogenomics, sentinel species, intraspecies variability, ecotoxicology.

INTRODUCTION

The assessment of toxic effects of chemicals on aquatic animal species and the advancement of knowledge about the underlying mechanisms acquired in predictive ecotoxicology have so far been based mainly on the use of exposure laboratory bioassays under standardized conditions. These tests have been developed to gain reproducibility, but standardization results detrimental to their representativeness with regard to species and the variability of their populations within ecosystems (Calow, 1996; Breitholtz et al., 2006). Indeed, laboratory tests consider strains of organisms maintained specifically under controlled breeding conditions, with an objective of high homogeneity between organisms with low genetic diversity, making clonal models such as *Daphnia* successful in aquatic animal ecotoxicology. Similarly, the environmental realism of experimental designs achievable in the laboratory, which in most cases are limited to periods covering a unique life stage of the animal lifespan, sometimes one generation and rarely a few generations in animal species, questions their representativeness in relation to long-term exposures of natural populations (Chapman, 2002; Coutellec and Barata, 2013). Over the past decade, one of the advantages of biological reductionism inherent in this approach allowed to benefit from the latest major advances in molecular biology. Indeed, global omics methodologies were applied for a large variety of laboratory model organisms offering a deep gain of knowledge on the toxic mechanisms of many substances. Examples include works on zebrafish (Sukardi et al., 2010), fathead minnow (Larkin et al., 2007), or daphnia (Colbourne et al., 2011). Among omics methodologies, proteomics shows great potential to advance in the understanding of the molecular responses to exposure to toxic substances by offering the advantage of directly targeting the effector molecules of the physiological processes affected: the proteins (Karr, 2008; Silvestre et al., 2012). Comparative proteomics that now benefits from the unprecedented progress of high-throughput label-free mass spectrometry measurements allows establishing mechanistic links between molecular disturbances induced by the action of contaminants and impacts on the physiology of organisms (e.g., (Pillai et al., 2014)).

While laboratory approaches remain undeniably important in ecotoxicology for screening the potential toxicity of chemicals and gaining insights into mechanisms of environmental toxicity, the current advances in molecular technologies make possible the implementation of omics analyses for environmental species of ecological importance (Martyniuk and Simmons, 2016). Proteogenomics is an integrative approach that allows acquisition of large-scale proteomes of sentinel species in aquatic environments, challenging the boundary between model and non-model species (Armengaud et al., 2014; Gouveia et al., 2019a). As an example, a large proteome catalogue was established for the amphipod *Gammarus fossarum* after interpreting tandem mass spectrometry data acquired on proteins based on a newly sequenced and *de novo*-assembled transcriptome (Trapp et al., 2014). The proteome modulation of this freshwater crustacean in response to laboratory exposure to two insecticides and cadmium (Cd) known as endocrine disruptors was also straightforwardly monitored by differential proteomics (Trapp et al., 2015). This approach based on the combination of transcriptomic and proteomic data can be applied to any targeted species (Armengaud et al., 2014). Comparative proteomics should be amenable for the study of field populations living in rivers with contrasting levels of contamination. However, the ability to discriminate against the effects of contaminants, considering field populations, may be affected by the inter-individual variability of responses within each population (Simmons et al., 2015). In wild populations, natural differences in life and exposure history between classes of individuals, along with the occurrence of physiological acclimatization processes relying on either individual phenotypic flexibility or developmental plasticity (Piersma and Drent, 2003; Uller, 2008), may result in

introducing strong inter-individual variability into the proteomic patterns in exposed populations (Karr, 2008). In this vein, some authors have put forward the concept of proteomic reaction norm in understanding the effects of contaminants in natural populations (Silvestre et al., 2012). The outcomes on phenotypic inter-individual variability of long-term evolutionary processes in exposed populations are also important but still poorly documented in ecotoxicology. As reviewed for Cd sensitivities in invertebrates (Dallinger and Hockner, 2013), both the possibility of a tendency towards homogenization or, on the contrary, phenotypic diversification within populations could be observed. Modification of effective sizes and migratory flows in relation to the demographic depression of exposed populations also constitute potential sources of within-population phenotypic variability (Dallinger and Hockner, 2013). Hence, unlike comparisons to laboratory control, within-population variability could compromise our ability to statistically detect the sub-proteome differentially expressed in the field when comparing historically exposed populations with reference field populations (Simmons et al., 2015; Bahamonde et al., 2016).

Using the example of two regional *Gammarus pulex* populations, the purpose of the present methodological study is to illustrate the importance of taking into account the population proteome variation possibly induced by contamination in the differential analysis of proteomes between field populations. Similarly to the study of Vigneron et al (Vigneron et al., 2015), one of these two populations is naturally exposed to bioavailable mono-metallic contamination (Cd) due to geological influence in a context of a crystalline bedrock headwater basin. Taking advantage of the rare opportunity of this *in natura* laboratory of a historical mono-metallic exposure, the intrapopulation proteome variability of long-term exposed *G. pulex* was assessed relatively to a non-contaminated population. Thus, 40 individual proteomes were established by proteogenomics based on one male and one female transcriptomes specifically and previously obtained in this *G. pulex* lineage (Cogne et al., 2019). This large set of individual data then allowed a comparative proteomics analysis that explicitly took into account individual variability. We report here on the data analysis strategy adopted to identify the proteomic signal and its variability which discriminate the two populations and potentially Cd long term exposure effects. The objective here is to highlight and quantify the gain provided by a stratified data analysis strategy that takes into account intrapopulation variability in a comparative field proteomics approach. The fine functional description of proteins differentially detected between populations and among groups of individuals from the contaminated population is discussed here. This work remains a challenge in itself in such species which belongs to taxonomic groups where the annotation of sequenced genomes is still at its early stage.

MATERIAL AND METHODS

Gammarid sampling

G. pulex were collected in two rivers located in mid-eastern France. In previous studies addressing the use of caged *Gammarus* as biomonitors of aquatic contaminations, we characterized the level of bioavailable metallic contamination in rivers from different areas in this region (Besse et al., 2013; Urien et al., 2016). In particular, four neighbouring stations along the Doux and Cance rivers were identified in 2009 as presenting high levels of Cd bioavailable contamination. These localized high levels of gammarid Cd-contamination were explained by the influence of natural geochemical sources on a crystalline bedrock conjugated with extremely low levels of water mineralization (particularly calcium) which enhance the bioaccumulation of this metallic trace element in *Gammarus* (Pellet et al., 2009). Based on this observation, we investigated creeks among the upstream tributaries of the Doux and Cance rivers, where

we identified *G. pulex* populations inhabiting these localized Cd-contaminated river networks. Among these populations, we selected the Brameloup population (45°07'51"N; 4°25'00"E) in which Cd bioavailable contamination was measured in 2015 as two times higher than the national threshold of bioavailable contamination as defined in the national study of Ciliberti et al (Ciliberti et al., 2017). In contrast, we selected the Pollon river (45°57'21"N; 5°15'44"E) in the lowland plain of the Ain river, where no Cd contamination was recorded (Besse et al 2013). The Pollon river shelters a *G. pulex* population in sympatry with *G. fossarum* organisms. Cd levels recorded in native gammarids in 2015 confirm the Cd-free status in this station (internal Cd concentrations four times lower than the threshold of bioavailable contamination; Ciliberti et al 2017). Of note, the two sites showed contrasting physico-chemical characteristics, especially in terms of water hardness (5 mg $CaCO3.L^{-1}$ in Brameloup *vs* 250 mg $CaCO3.L^{-1}$ in Pollon) due to either crystalline or limestone areas, respectively.

Organisms were collected in August 2017 by kick sampling using a net. Adults were selected using a 2-2.5mm sieve, and quickly transported to the laboratory in plastic buckets containing freshwater from the station. In the laboratory, organisms were kept only few days in water sampled in their respective station of origin. Sexually mature male and female organisms were sampled in amplexus (paired male and female). One or two pereopods were cut from each male in order to ascertain by molecular barcoding (COI) that the selected individuals belong to the *G. pulex* species. Embryos were quickly removed from the marsupium of females (about 30 sec). Males and females were then directly frozen in liquid nitrogen and stored at −80 °C until proteomic analysis.

Protein extraction

Each animal was analysed individually by shotgun proteomics in standard conditions. For this, each animal was mechanically homogenized by bead-beating (one 3.2 mm steel bead per tube) in 20 µl of LDS sample buffer (Invitrogen) per milligram of organism. The homogenates were centrifuged at 10,000g for 3 minutes in order to pellet cellular debris, and the resulting supernatant collected to a new tube. Samples were then incubated for 5 min at 99 ˚C. A 20 µL aliquot of each sample was then subjected to SDS-PAGE for a short electrophoretic migration, as described previously (Trapp et al., 2015). The whole-protein content from each well was extracted as a single polyacrylamide band, processed as described (Hartmann et al., 2014), and submitted to proteolysis with trypsin (Roche) using 0.01% ProteaseMAX surfactant (Promega).

Tandem mass spectrometry

The peptide mixtures were analysed in data-dependent mode with a Q Exactive HF mass spectrometer (Thermo) coupled with an UltiMate 3000 LC system (Dionex-LC Packings). This instrument was operated essentially as described (Klein et al., 2016). Briefly, peptides were desalted and then resolved onto a nanoscale C18 PepMapTM 100 capillary column (LC Packings) with a 90-min gradient of CH3CN, 0.1% formic acid, at a flow rate of 0.2 µL/min. Following a Top20 method, peptides were analysed with scan cycles initiated by a full scan of peptide ions in the Orbitrap analyser, followed by high-energy collisional dissociation and MS/MS scans on the 20 most abundant precursor ions. Full scan mass spectra were acquired from *m/z* 350 to 1800 at a resolution of 60,000. Ion selection for MS/MS fragmentation and measurement was performed applying a dynamic exclusion window of 10 sec.

Peptide assignation and proteomics data analysis

MS/MS spectra were assigned to peptide sequences by the MASCOT Daemon 2.3.2 search engine (Matrix Science) searching against a gender-specific RNAseq-derived database obtained by *de novo* assembly with Trinity version 2.4 (Haas et al., 2013) followed by ORF search with Transdecoder as described

in Cogne et al. (2019, submitted). The male database contains 111,751 putative protein sequences totalling 17,598,514 amino acids while the female database comprises 121,147 putative protein sequences totalling 22,904,430 amino acids. For MS/MS spectra assignment, the parameters were: full-trypsin specificity, maximum of one missed cleavages, mass tolerances of 5 ppm on the parent ion and 0.02 Da on the MS/MS, carboxyamidomethylated cysteine (+57.0215) as a fixed modification, and oxidised methionine (+15.9949) and deamidation of asparagine and glutamine (+0.9848) as variable modifications. All peptide matches presenting a MASCOT peptide score with a *p*-value of less than 0.055, corresponding to an FDR of 2% as evaluated with the DecoyPyrat procedure (Wright and Choudhary, 2016), were filtered and assigned to a protein without parsimony to avoid any bias between isoforms. Annotation of detected protein was done using Swissprot and NCBInr databases with the blast tool. PSMs were counted for each protein group defined on the basis of their similar annotation without parsimony. Functional annotation of proteins and gene ontology classification were done as previously described (Trapp et al., 2018). Due to the lack of an annotated genome, we used the output of the GO terms to associate each protein to one category of the following molecular functions or protein families: hemocyanins, bioluminescence, innate immunity, actins, myosins, protein lipidation, carbohydrate metabolisms, proteolysis, oxidative stress response, mitochondrial proteins, proteins of unknown or uncharacterized function and proteins belonging to other functions.

Mass spectrometry and proteomics data

The mass spectrometry and proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE (Perez-Riverol et al., 2019) partner repository with the dataset identifiers PXD013656 and 10.6019/PXD013656 for the *G. pulex* males, and PXD013712 and 10.6019/PXD013712 for the *G. pulex* females. [The former dataset is available with the Username: reviewer49791@ebi.ac.uk and Password: DDNFCqmj, while the later dataset is available with the Username: reviewer51278@ebi.ac.uk and Password: Pnpvyt8r].

Differential detection and molecular diversity analysis

A matrix with abundances of all MS/MS-detected proteins based on cumulated spectral counting of peptides present at least once in each population has been established for all male individuals on one hand, and for all female individuals on the other. For the differential comparison of protein abundances, spectral counts were normalized as described (Liu et al., 2004). Statistical evaluation of differential detection was carried out using the Tfold method which is based on a Student test considering sample with unequal variance and bilateral repartition of differential abundances of proteins (Carvalho et al., 2008). Bonferroni correction was applied on p-values using the number of proteins detected with at least two different peptides for each sex. When comparing Brameloup and Pollon sites, proteins with a corrected p-value below 0.05 and a Tfold absolute value higher than 1.5 were considered as differentially expressed between the two rivers. Distance calculation was performed using Spearman correlation from the matrix of abundances of each protein previously generated independently for each sex and river. Protein abundances of individuals from Brameloup were clustered using pvclust version 2.0 R package (Suzuki and Shimodaira, 2006) using spearman distance calculation, average method clustering and 1000 bootstrap. Principal component analysis was done using spectral counts of peptides shared between male and female using the PCA function from R version 3.4.2.

RESULTS AND DISCUSSION

Shotgun proteomics of two *G. pulex* populations

Figure 1 shows the strategy for proteomic data acquisition and interpretation. Ten males and ten females from two *G. pulex* populations, namely Pollon and Brameloup sites, were sampled and their soluble proteomes were individually established by high-throughput shotgun proteomics. In average, a total of 39,794 (+/- 2,978) MS/MS spectra were recorded per sample. The whole dataset for the males (765,425 MS/MS spectra) was interpreted against a protein sequence database derived from an RNAseq experiment done on a *G. pulex* male individual sampled from the Pollon site. In total, the ratio of peptide-to-spectrum matches (PSMs) per MS/MS spectra is 28.5%, indicating a rather good quality of the protein sequence database. This result is in line with those already reported for other *Gammarids* analyzed with a similar strategy and pipeline (Trapp et al., 2016). A total of 9,347 different peptides were confidently listed for the *G. pulex* animals. They validated the presence of 1,385 proteins identified with at least two different peptides (**See supplementary Table S1 in Gouveia et al., 2019b**). A core detected proteome comprising 606 proteins could be delineated on the basis of the polypeptides detected systematically in all the 20 analyzed males. The cumulative spectral count of these proteins represents 92.8% of the total. The same strategy was performed for the female dataset (826,340 MS/MS spectra). For females, the ratio of PSMs per MS/MS represents 29.5%. A total of 9,869 different peptides were obtained, validating the presence of 1,444 proteins identified with at least two different peptides (**See supplementary Table S2 in Gouveia et al., 2019b**). In this case, the core detected proteome comprises 609 proteins and their spectral count represents 92.8%. Thus, a rather high homogeneity of the proteomics results is obtained for the two sexes.

PSMs were counted for each protein group defined on the basis of their annotation. In both sexes, the most abundant protein group is hemocyanin which appears as diverse with numerous detected proteoforms. Hemocyanin is the respiratory pigment of crustaceans that facilitates oxygen transport. The sequence variability of the three subunits was already highlighted in *Penaeid* shrimps with several isoforms established by RNAseq (Johnson et al., 2016) and in other arthropods (Costa-Paiva et al., 2018; Scherbaum et al., 2018). Here, this protein group represents based on the cumulated PSMs 21.3% for the male dataset and 18.1% for the female dataset. For the male dataset, the other most abundant protein groups are: endoglucanase (3.3%), an uncharacterized protein (contig GPCF_DN113133) (2.5%), twitchin kinase (1.7%), and glucosylceramidase (1.7%). The other most abundant protein groups for the female dataset are: hemolymph clottable protein 5.3% of PSMs, and the same uncharacterized protein as in males (2.9%), endoglucanase (2.7%) and glucosylceramidase (1.5%).

Global comparative proteomics of the two *G. pulex* populations

The male individuals sampled in Brameloup were compared in terms of protein abundance assessed by spectral counts to those sampled in Pollon (**See Figure 1 in Gouveia et al., 2019b**). In this case, only trypsic peptides detected in both populations were taken into account, resulting in a list restricted to 1,206 polypeptides. This allows avoiding bias due to a possible genome divergence between both populations that could occur and generate some variants on accessory proteins (Lespinet et al., 2002; Aravind et al., 2006). The abundance of proteins were evaluated by spectral count, a label free methodology well adapted for comparing organisms with different genetic backgrounds and without genome sequence reference. A total of 84 proteins were found with significant differentiating abundances with stringent thresholds (TFold change set at 1.5 and Bonferroni corrected p-value at 0.05), thus representing ~7% of the analyzed proteome. Among these, 61 were more abundant in Brameloup, while 23 resulted more abundant in Pollon. The five most differentially abundant proteins in Brameloup males compared to Pollon males are: non-catalytic subunit of oplophorus-luciferin 2-monooxygenase (x4.3), lipopolysaccharide and beta-1,3-glucan binding

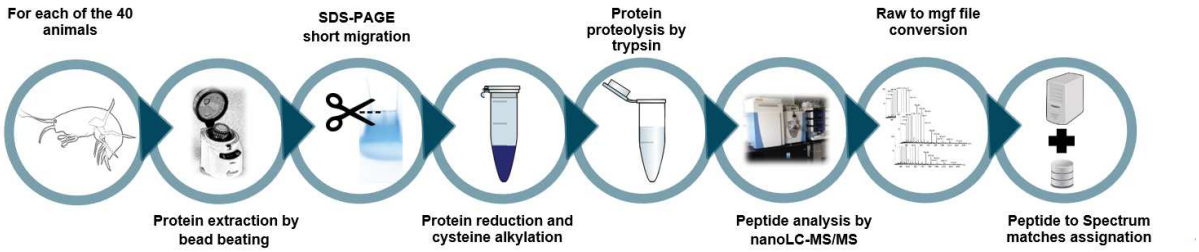


Figure 1. Experimental flowchart for generating proteomics data on individual animals. For the 40 animals, proteins were extracted by bead beating, purified by denaturing electrophoresis on a polyacrylamide gel, and in-gel proteolyzed with trypsin. Each of the 40 peptide fractions were resolved by reverse phase chromatography and analysed by high-resolution tandem mass spectrometry. The MS/MS data were interpreted against a RNAseq derived database.

protein (x3.7), hydroxypyruvate isomerase (x3.6), endoglucanase A-like (x3.3), and another beta-1,3-glucan-binding protein isoform (x3.1). The five most differentially accumulated proteins in Pollon males compared to Brameloup males are: fatty acid-binding protein (:4.8), ornithine aminotransferase (:3.7), and three uncharacterized proteins (:4.2, :3.4, and :3.1). Then, we aggregated the different proteins into biological functional groups and compared the different functional profiles in the two populations (**Figure 2, Panel A**). Interestingly, the Brameloup gammarids had a significant higher abundance of proteins involved in protein lipidation, carbohydrate metabolisms (notably cellulose degradation), proteolysis and innate immunity. Moreover, a significant increase in the abundance of proteins involved in oxidative stress response and lipid transport was also observed, although the global abundance of these proteins in the Brameloup population was relatively low (25-33 spectral counts). On the other side, hemocyanins were significantly less detected in the Brameloup population compared with the Pollon population. Down regulation of hemocyanin expression was already observed in the white shrimp *Litopenaeus vannamei* exposed to different concentrations of Cd (Bautista-Covarrubias et al., 2014). Indeed, Cd can replace copper in the oxygen binding site of hemocyanins and modulate the oxygen affinity as demonstrated with the protein from the blue crab *Callinectes sapidus* (Brouwer et al., 1982). Based on laboratory toxicity assays on a freshwater crab with high Cd concentrations (Wang et al., 2013), exposure of *G. pulex* to Cd in the Brameloup river might induce a chronic Cd-induced oxidative stress that elevates the amount of production for specific proteins such as the monooxygenase enzymes.

For the female comparison, the list of polypeptides is restricted to 1,286 items when only peptides detected in both populations were taken into account. For females 153 proteins are defined as differentially detected representing ~10% of the analyzed proteome **(See Figure 2 in Gouveia et al., 2019b).** A group of 70 polypeptides were found less abundant in Brameloup compared to Pollon female animals, while 83 were found more abundant. The five more abundant protein groups are: cellobiohydrolase (x5.7), the non-catalytic subunit of oplophorus-luciferin 2-monooxygenase (x4.6), mannan endo-1,4-beta-mannosidase-like (x3.9), hydroxypyruvate isomerase (x3.8), and lipopolysaccharide and beta-1,3-glucan binding protein (x3.8). The five less abundant proteins are: ornithine aminotransferase (:5.95), three uncharacterized proteins (:4.5, :4.0, and :3.5) and fatty acid-binding protein (:3.7). Thus a rather striking similarity is observed in male and female animals in terms of the most differentially detected proteins. As for males, the most differentially detected proteins between the two study sites are implicated in oxidative stress, digestion, and host defense against invading microorganisms. Similarly, when aggregated by biological functions (**Figure 2, Panel B**), we observed a functional profile very similar to that observed in males. In particular, the Brameloup female gammarids showed significant higher abundance in proteins involved in proteolysis, protein lipidation, carbohydrate metabolisms, innate immunity, oxidative stress response and bioluminescence. On the other side, proteins of the myosin and actin families and mitochondrial proteins were less abundant in Brameloup females compared with Pollon females, while hemocyanins were not significantly different. The difference mentioned for actin also apply for males. The differences in animal manipulations between males and females should not influence their proteome contents as this process was quick compared to general protein turnover. Finally, our strategy may have favored interpretation of Pollon dataset compared to Brameloup dataset as we used a Pollon RNAseq-derived database. To avoid any bias in the interpretation only common peptides between the two populations were taken into account, thus restricting the comparison to a core proteome of 1,206 conserved proteins. The main difference between the Brameloup and Pollon sites is the cadmium exposure in the former. However, the adaptation of the animals to habitats with different physico-chemical parameters (*e.g.* water hardness) or trophic conditions, interspecific interactions, in

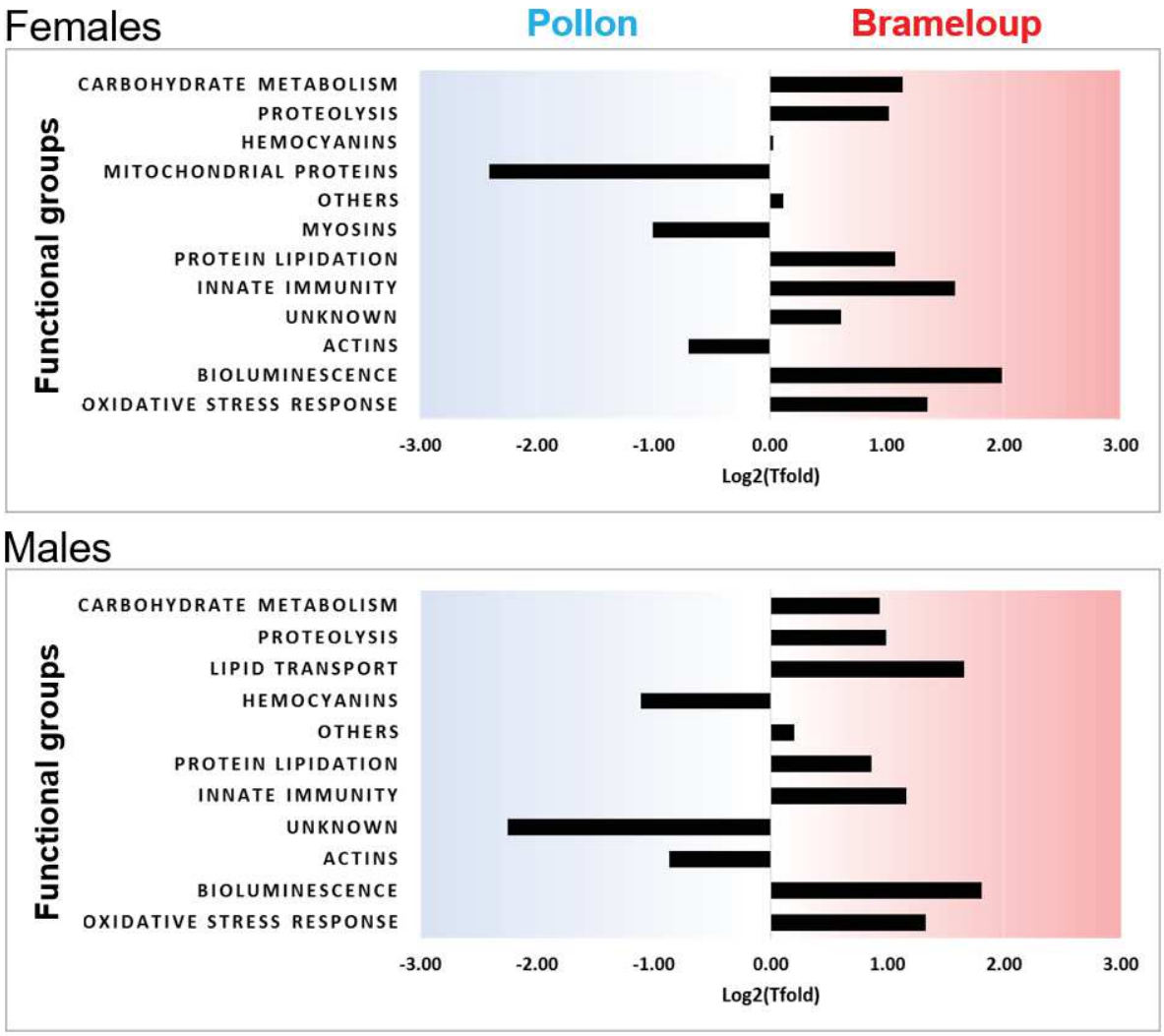


Figure 2. Functional groups highlighted by global comparative proteomics and their fold changes. Functional groups were established in males (Panel A) and females (Panel B) by comparing protein abundances in Brameloup versus Pollon *G. pulex* populations (n=10 per condition).

addition to changes in response to cadmium exposure, could also explain at least in part some of the observed proteomic changes.

Evaluation of *G. pulex* intrapopulation variability at the individual scale by means of proteomics

We analyzed the individual detection of the 1,206 proteins identified in the male populations in one hand, and the 1,286 proteins identified in the female populations on the other. Taking into account their detection and their abundances as estimated by their spectral counts in the 10 animals of each population and sex, a distance matrix was constructed based on Spearman correlation of their distributions. **Figure 3** shows the violin plot of the distribution of these distances, representative of the protein spectral count uniformity within the four groups of animals. While the mean distance is relatively similar for the four populations, the ranges of variability are quite differing. The Pollon male population is clearly the most homogeneous group. The male and female Brameloup populations exhibit a higher heterogeneity compared to their Pollon counterparts (**Figure 3**). In order to explain this difference, we performed a hierarchical clustering of the individual proteomes of each of the four groups of animals. **Figure 4** shows the results of this clustering for the males (Panel A) and females (Panel C) from Brameloup. The Brameloup male BM3 individual is out-grouped while two groups can be delineated but at a rather low relationship distance (below 0.13) in males with BM1, BM4, BM6, BM8 and BM10 on one hand, and BM9, BM2, BM5 and BM7 on the other. For the females there are two groups which are clearly distinguished, named FBC1 and comprising two subgroups (BF2, BF4, and BF7 animals on one hand, and BF6, BF8, BF9, and BF10 individuals on the other) that can be delineated at lower relationship distance (below 0.13), and FBC2 which comprises three other individuals (BF1, BF3, and BF5). The same hierarchical clustering applied on Pollon animals did not show the delineation of groups with a relationship distance above 0.13 for Male (panel B). However, hierarchical clustering for Pollon females resulted into the definition of two clusters at the same threshold, named FPC1 (PF4, PF5, PF8, PF2 and PF6) and FPC2 (PF1, PF3, PF7, PF9 and PF10). These clusters highlight natural heterogeneity for females from the reference population, but slightly lower than for animals from the Brameloup population (proteomic relationship distances of ~0.13 vs ~0.14, respectively). Moreover, distances between individuals from the FBC1 cluster are relatively low. **Figure 5** shows the results of a principal component analysis of the proteome content from the 40 samples. The proteomes of males and females from Pollon are clustered while for Brameloup, the BM3 animal is clearly outlier and two clusters are defined: FBC2 on the first hand, and MBC1/FBC1 on the other, thus confirming the previous results. These clustering results clearly explain the heterogeneity observed in the violin plot (**Figure 3**), as one single individual is clearly distant from the others in Brameloup males on the first hand, and higher distances are observed for Brameloup females compared to Pollon females on the other. Of note, the existence of different proteome states within the two populations is not likely due to the existence of cryptic genetic sub-populations, considering that paired males and females (same code number on **Figure 3**) are not clustered following the same delineation. Figure 3 shows the violin plot of the Brameloup males calculated without the BM3 outlier, showing still more variability than the Pollon males.

Comparative proteomics of the two *G. pulex* populations taking into account population heterogeneity

Once the proteome heterogeneity of the four populations established, it was interesting to reveal whether the proteomic comparison is impacted or not. For the proteomic comparison of males from Brameloup versus males from Pollon, we removed the BM3 individual outlier, thus considering the MBC1 cluster as a homogeneous animal dataset (**See supplementary Table S1 in Gouveia et al., 2019b**). In this case, a total of 133 proteins were defined as differentially detected, now

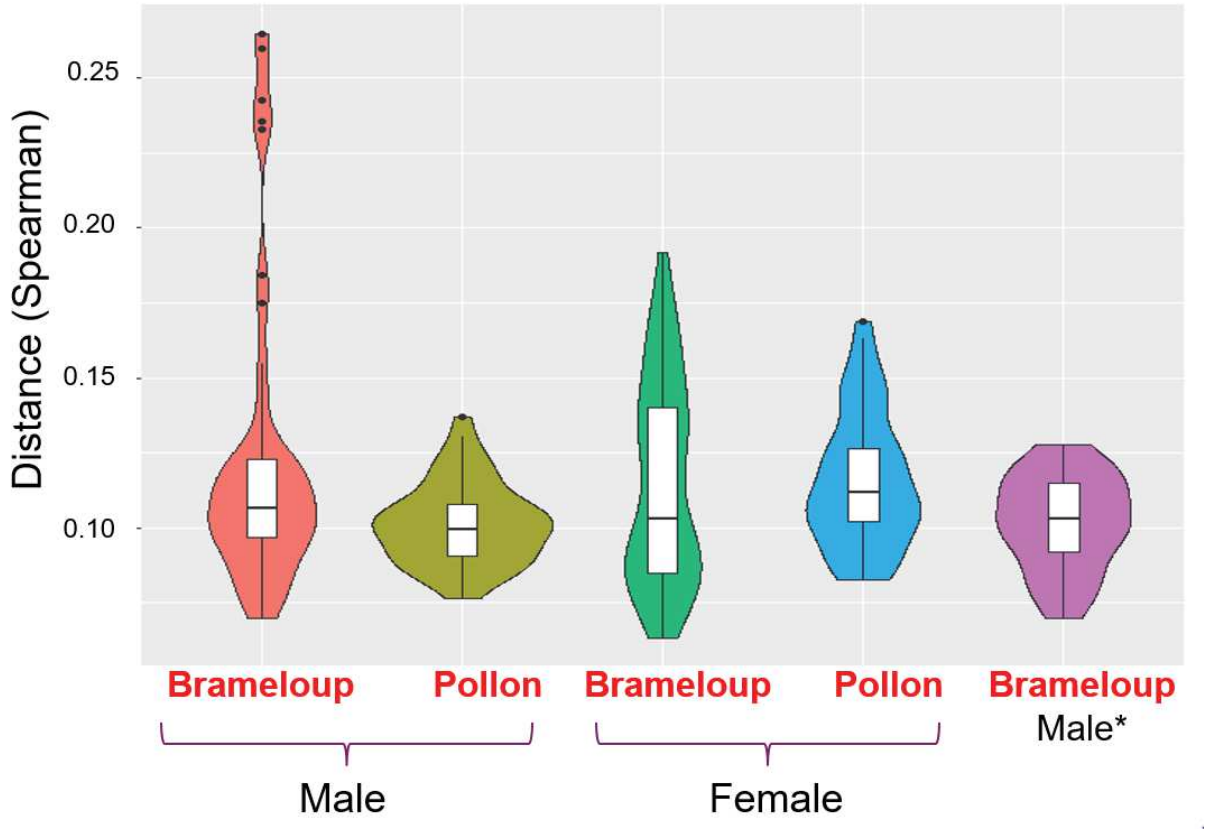


Figure 3. Violin plot representing the spearman correlation of the distributions of protein abundances in individual animals. Boxplot showing mean Spearman distance among all individuals is indicated for each population (n=10 per condition). Brameloup Male* was calculated without the BM3 outlier (n=9).

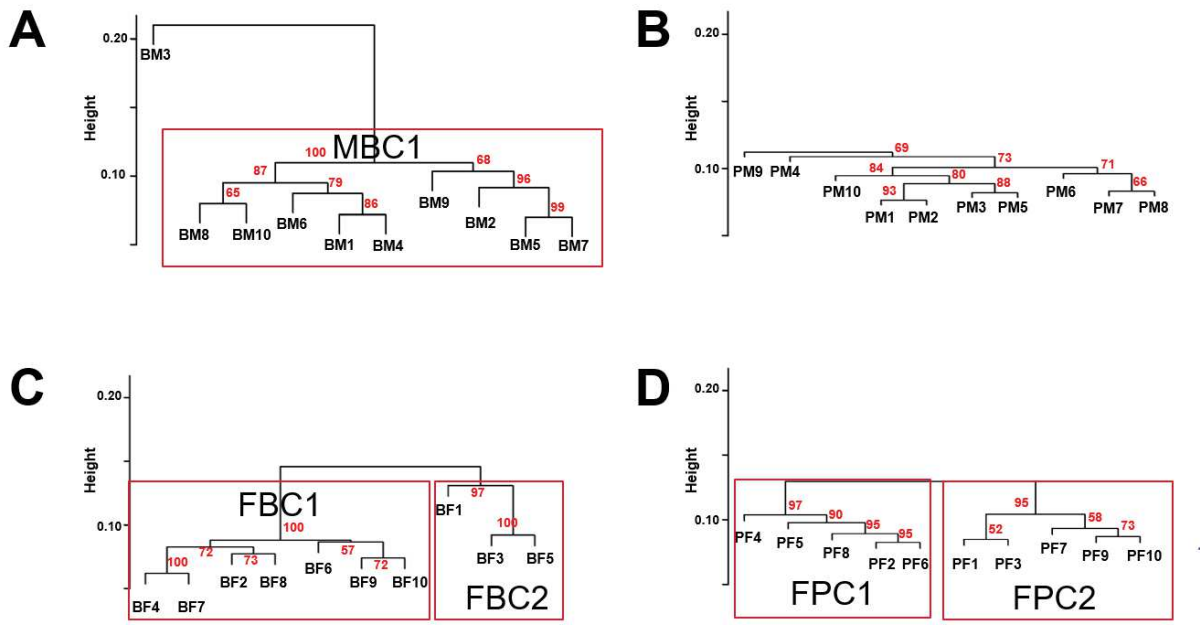


Figure 4. Dendogram representation of hierarchical clustering of individuals for the four animal groups. The hierarchical clustering was done using spearman distance matrix of protein abundance for each population (n=10 per condition): male Brameloup (Panel A), male Pollon (Panel B), female Brameloup (Panel C), and female Pollon (Panel D). Red rectangles indicate clusters defined with a distance threshold above 0.13.

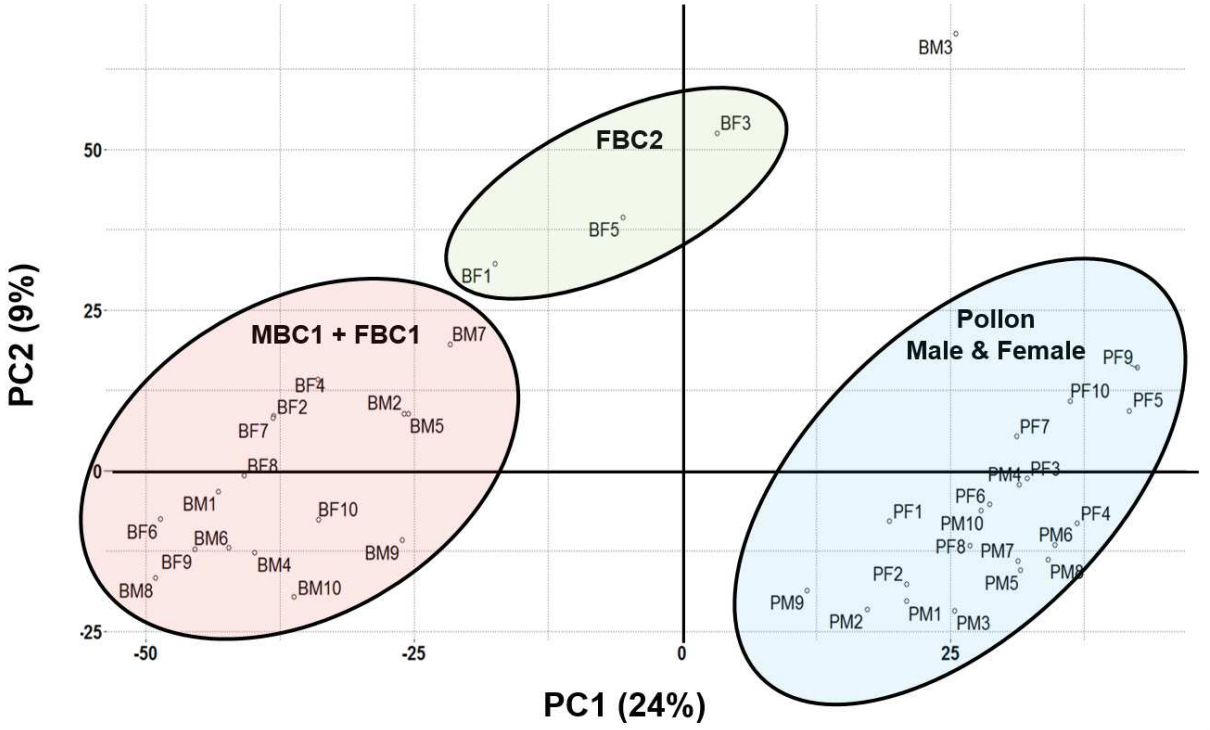


Figure 5. Principal component analysis of the 40 individual proteomes. Individuals grouping allows detection of 3 main groups and an outlier with the first two dimensions (n=40). MBC1 and FBC1 clusters defined in the dendograms from Figure 4 were found grouped, as well as all Pollon male and female proteomes. These groups are indicated with coloured ellipses.

representing ~10% of the total detected proteome**.** When compared to the proteins identified as differentially detected using the total population of Brameloup males, 61 new proteins were highlighted as differentially detected with the MBC1 cluster and 12 proteins previously identified were not retained at the same fold change and p-value thresholds as indicated in the Venn diagram shown in **Figure 6**. From the newly identified proteins, the proteins with the highest fold change considering proteins more differentially detected in Brameloup population is an uncharacterized protein (x1.8) and the less detected is the ribosomal protein L19e (:2). Because Brameloup females were delineated into two distinct clusters, two proteomic comparisons were done. First, the protein abundances from the FBC1 cluster (7 animals) from Brameloup females were compared to the Pollon dataset, resulting in 223 proteins defined as differentially detected which represented ~15% of the total proteome **(See supplementary Table S2 in Gouveia et al., 2019b).** Then, the protein abundances from the FBC2 cluster (3 animals) were compared to the Pollon dataset. In this case, a total of 28 proteins were defined as differentially detected, representing ~2% of the total validated proteins**.** Interestingly, the FBC1 comparison highlighted the family 7 cellobiohydrolase (x5.9), whose catalytic activity is closely linked to nutrition, and the 14-3-3 protein (:6.6), a conserved regulatory protein. The FBC2 comparison pinpointed at the oplophorus-luciferin 2-monooxygenase

non-catalytic subunit (x4.0) and an uncharacterized protein (:10.9) as the two proteins exhibiting the most abundance differences. Moreover mannan endo-1,4-beta-mannosidase-like protein is the highest newly overexpressed protein (x3.2) in Brameloup compared to the comparisons of all animals. When merging the three comparisons, a total of 273 proteins can be considered as differentially detected as reported in **Figure 6**. Regarding at the differences between the three comparisons, 105 new proteins were highlighted with the FBC1 comparison while 15 were highlighted with the FBC2 comparison. Importantly, almost all the proteins found when the ten individuals were compared from each group can be found with the FBC1-restricted comparison (**Figure 6**). Globally, it appears that taking into account intrapopulation heterogeneity allows disclosing approximately 50% more proteins as differentially expressed between the two populations. Given the one population *vs* one population design of this first study, all of these modulated proteins could nevertheless be unrelated with the long term Cd exposure of the Brameloup population. Their modulation could potentially be explained by other environmental sources of variability between these two first investigated sites. The modulation of each of these candidate proteins will now have to be tested on additional reference and cadmium contaminated populations, covering a wider range of environmental habitat conditions, in order to confirm their involvement in the specific response to long-term exposure to this metallic contaminant.

Conclusive remarks

Currently, most comparative proteomics studies are usually restricted to a rather small number of biological replicates, typically three to five, due to the costs of such experiments and the number of conditions to explore (Pisani et al., 2017; Gallois et al., 2018). Pooling animals into artificially designed biological replicates could lead to averaging the measured signal and is generally not recommended (Gouveia et al., 2019a). Here, we have performed a deep proteomic analysis of 40 individual animals sampled in the wild leading to an invaluable dataset on two populations of *G. pulex* sampled from a natural bioavailable Cd contamination site and a reference site. While comparing the genetic landscape of animals became a sound molecular tool for assessing the heterogeneity of populations (Gutierrez-Rodriguez et al., 2017), to our knowledge the present study is the first to exploit high-throughput proteomics data for assessing animal heterogeneity in several populations. This promises easier functional insights into the molecular heterogeneity of populations, from which advances in the understanding of the effects of long term exposure to chemical contamination in wild populations could benefit. Indeed, we have shown that a quick analysis can be performed based on the abundance of all the proteins detected in individual animals giving a picture of the heterogeneity of the population under scrutiny.

Based on our results, it appears now obvious that taking into account this individual heterogeneity is essential but requires the analysis of numerous individuals for avoiding clearly defined outliers or eventually considering sub-grouped individuals for a more accurate comparative proteomics.

ACKNOWLEDGEMENTS

We thank the Institut National de Recherche en Sciences et Technologies pour l'Environnement et l'Agriculture (France), the Commissariat à l'Energie Atomique et aux Energies Alternatives (France), and the Agence Nationale de la Recherche program "ProteoGam" (ANR-14-CE21-0006-02) for financial support. We gratefully acknowledge Jean-Charles Gaillard for expert assistance with mass spectrometry.

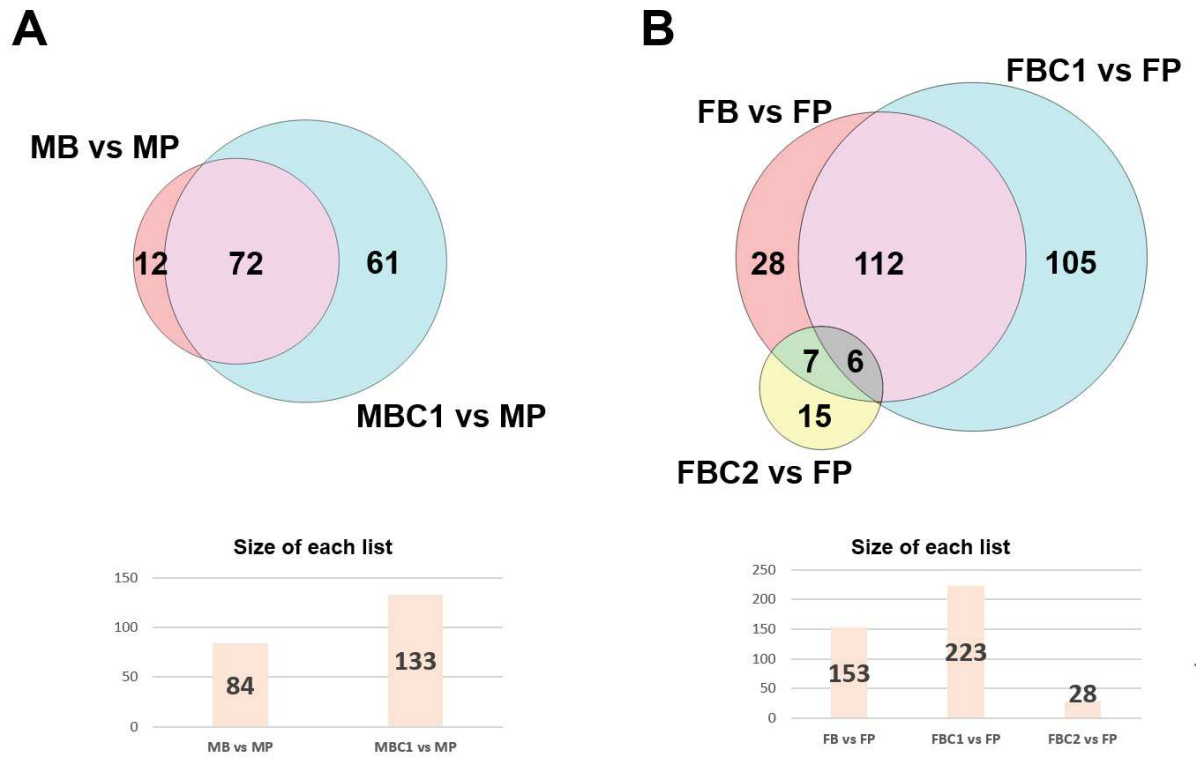


Figure 6. Venn diagram indicating the proteins whose abundance is differentially modulated when considering Brameloup animal clusters. Proteins were defined as differentially detected when comparing Brameloup cluster and Pollon population on the basis on an absolute fold change above 1.5 and a Bonferroni corrected p value below 0.05. Panel A represents results for males (n=20) and Panel B shows the results for females (n=20).

REFERENCES

Aravind, L., Iyer, L.M., and Koonin, E.V. (2006) Comparative genomics and structural biology of the molecular innovations of eukaryotes. *Curr Opin Struct Biol* **16**: 409-419.

Armengaud, J., Trapp, J., Pible, O., Geffard, O., Chaumot, A., and Hartmann, E.M. (2014) Non-model organisms, a species endangered by proteogenomics. *J Proteomics* **105**: 5-18.

Bahamonde, P.A., Feswick, A., Isaacs, M.A., Munkittrick, K.R., and Martyniuk, C.J. (2016) Defining the role of omics in assessing ecosystem health: Perspectives from the Canadian environmental monitoring program. *Environ Toxicol Chem* **35**: 20-35.

Bautista-Covarrubias, J.C., Velarde-Montes, G.J., Voltolina, D., Garcia-de la Parra, L.M., Soto-Jimenez, M.F., and Frias-Espericueta, M.G. (2014) Humoral and haemocytic responses of Litopenaeus vannamei to Cd exposure. *ScientificWorldJournal* **2014**: 903452.

Besse, J.P., Coquery, M., Lopes, C., Chaumot, A., Budzinski, H., Labadie, P., and Geffard, O. (2013) Caged Gammarus fossarum (Crustacea) as a robust tool for the characterization of bioavailable contamination levels in continental waters: towards the determination of threshold values. *Water Res* **47**: 650-660.

Breitholtz, M., Ruden, C., Hansson, S.O., and Bengtsson, B.E. (2006) Ten challenges for improved ecotoxicological testing in environmental risk assessment. *Ecotoxicol Environ Saf* **63**: 324-335.

Brouwer, M., Bonaventura, C., and Bonaventura, J. (1982) Heavy metal ion interactions with Callinectes sapidus hemocyanin: structural and functional changes induced by a variety of heavy metal ions. *Biochemistry* **21**: 2529-2538.

Calow, P. (1996) Variability: noise or information in ecotoxicology? *Environ Toxicol Pharmacol* **2**: 121-123.

Carvalho, P.C., Fischer, J.S., Chen, E.I., Yates, J.R., 3rd, and Barbosa, V.C. (2008) PatternLab for proteomics: a tool for differential shotgun proteomics. *BMC Bioinformatics* **9**: 316.

Chapman, P.M. (2002) Integrating toxicology and ecology: putting the "eco" into ecotoxicology. *Mar Pollut Bull* **44**: 7-15.

Ciliberti, A., Chaumot, A., Recoura-Massaquant, R., Chandesris, A., Francois, A., Coquery, M. et al. (2017) Caged Gammarus as biomonitors identifying thresholds of toxic metal bioavailability that affect gammarid densities at the French national scale. *Water Res* **118**: 131-140.

Cogne, Y., Degli-Esposti, D., Pible, O., Gouveia, D., François, A., Bouchez, O. et al. (2019) De novo transcriptomes of 14 gammarid individuals for proteogenomic analysis of 7 different taxonomical groups. *Nature Scientific Data*: In revision.

Colbourne, J.K., Pfrender, M.E., Gilbert, D., Thomas, W.K., Tucker, A., Oakley, T.H. et al. (2011) The ecoresponsive genome of Daphnia pulex. *Science* **331**: 555-561.

Costa-Paiva, E.M., Schrago, C.G., Coates, C.J., and Halanych, K.M. (2018) Discovery of Novel Hemocyanin-Like Genes in Metazoans. *Biol Bull* **235**: 134-151.

Coutellec, M.A., and Barata, C. (2013) Special issue on long-term ecotoxicological effects: an introduction. *Ecotoxicology* **22**: 763-766.

Dallinger, R., and Hockner, M. (2013) Evolutionary concepts in ecotoxicology: tracing the genetic background of differential cadmium sensitivities in invertebrate lineages. *Ecotoxicology* **22**: 767-778.

Gallois, N., Alpha-Bazin, B., Ortet, P., Barakat, M., Piette, L., Long, J. et al. (2018) Proteogenomic insights into uranium tolerance of a Chernobyl's Microbacterium bacterial isolate. *J Proteomics* **177**: 148-157.

Gouveia, D., Almunia, C., Cogne, Y., Pible, O., Degli-Esposti, D., Salvador, A. et al. (2019a) Ecotoxicoproteomics: A decade of progress in our understanding of anthropogenic impact on the environment. *J Proteomics*. **198**: 66-77.

Gouveia, D., Cogne, Y, Gaillard, J.-C., Almunia, C., Pible, O., François, A., Degli-Esposti, D., Geffard, O., Chaumot, A., Armengaud, J. (2019b) Shotgun proteomics datasets acquired on Gammarus pulex animals sampled from the wild. *Data in Brief*. Submitted.

Gutierrez-Rodriguez, J., Goncalves, J., Civantos, E., and Martinez-Solano, I. (2017) Comparative landscape genetics of pond-breeding amphibians in Mediterranean temporal wetlands: The positive role of structural heterogeneity in promoting gene flow. *Mol Ecol* **26**: 5407-5420.

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J. et al. (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* **8**: 1494-1512.

Hartmann, E.M., Allain, F., Gaillard, J.C., Pible, O., and Armengaud, J. (2014) Taking the shortcut for high-throughput shotgun proteomic analysis of bacteria. *Methods Mol Biol* **1197**: 275-285.

Johnson, J.G., Burnett, L.E., and Burnett, K.G. (2016) Uncovering Hemocyanin Subunit Heterogeneity in Penaeid Shrimp using RNA-Seq. *Integr Comp Biol* **56**: 1080-1091.

Karr, T.L. (2008) Application of proteomics to ecology and population biology. *Heredity (Edinb)* **100**: 200-206.

Klein, G., Mathe, C., Biola-Clier, M., Devineau, S., Drouineau, E., Hatem, E. et al. (2016) RNA-binding proteins are a major target of silica nanoparticles in cell extracts. *Nanotoxicology* **10**: 1555-1564.

Larkin, P., Villeneuve, D.L., Knoebl, I., Miracle, A.L., Carter, B.J., Liu, L. et al. (2007) Development and validation of a 2,000-gene microarray for the fathead minnow (Pimephales promelas). *Environ Toxicol Chem* **26**: 1497-1506.

Lespinet, O., Wolf, Y.I., Koonin, E.V., and Aravind, L. (2002) The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Res* **12**: 1048-1059.

Liu, H., Sadygov, R.G., and Yates, J.R., 3rd (2004) A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal Chem* **76**: 4193-4201.

Martyniuk, C.J., and Simmons, D.B. (2016) Spotlight on environmental omics and toxicology: a long way in a short time. *Comp Biochem Physiol Part D Genomics Proteomics* **19**: 97-101.

Pellet, B., Geffard, O., Lacour, C., Kermoal, T., Gourlay-france, C., and Tusseau-vuillemin, M.H. (2009) A model predicting waterborne cadmium bioaccumulation in Gammarus pulex: the effects of dissolved organic ligands, calcium, and temperature. *Environ Toxicol Chem* **28**: 2434-2442.

Perez-Riverol, Y., Csordas, A., Bai, J., Bernal-Llinares, M., Hewapathirana, S., Kundu, D.J. et al. (2019) The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res* **47**: D442-D450.

Piersma, T., and Drent, J. (2003) Phenotypic flexibilty and the evolution of organismal design. *Trends in Ecology and Evolution* **18**: 228-233.

Pillai, S., Behra, R., Nestler, H., Suter, M.J., Sigg, L., and Schirmer, K. (2014) Linking toxicity and adaptive responses across the transcriptome, proteome, and phenotype of Chlamydomonas reinhardtii exposed to silver. *Proc Natl Acad Sci U S A* **111**: 3490-3495.

Pisani, C., Gaillard, J.C., Odorico, M., Nyalosaso, J.L., Charnay, C., Guari, Y. et al. (2017) The timeline of corona formation around silica nanocarriers highlights the role of the protein interactome. *Nanoscale* **9**: 1840-1851.

Scherbaum, S., Hellmann, N., Fernandez, R., Pick, C., and Burmester, T. (2018) Diversity, evolution, and function of myriapod hemocyanins. *BMC Evol Biol* **18**: 107.

Silvestre, F., Gillardin, V., and Dorts, J. (2012) Proteomics to assess the role of phenotypic plasticity in aquatic organisms exposed to pollution and global warming. *Integr Comp Biol* **52**: 681-694.

Simmons, D.B., Benskin, J.P., Cosgrove, J.R., Duncker, B.P., Ekman, D.R., Martyniuk, C.J., and Sherry, J.P. (2015) Omics for aquatic ecotoxicology: control of extraneous variability to enhance the analysis of environmental effects. *Environ Toxicol Chem* **34**: 1693-1704.

Sukardi, H., Ung, C.Y., Gong, Z., and Lam, S.H. (2010) Incorporating zebrafish omics into chemical biology and toxicology. *Zebrafish* **7**: 41-52.

Suzuki, R., and Shimodaira, H. (2006) Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* **22**: 1540-1542.

Trapp, J., Geffard, O., Imbert, G., Gaillard, J.C., Davin, A.H., Chaumot, A., and Armengaud, J. (2014) Proteogenomics of Gammarus fossarum to document the reproductive system of amphipods. *Mol Cell Proteomics* **13**: 3612-3625.

Trapp, J., Almunia, C., Gaillard, J.C., Pible, O., Chaumot, A., Geffard, O., and Armengaud, J. (2016) Proteogenomic insights into the core-proteome of female reproductive tissues from crustacean amphipods. *J Proteomics* **135**: 51-61.

Trapp, J., Armengaud, J., Pible, O., Gaillard, J.C., Abbaci, K., Habtoul, Y. et al. (2015) Proteomic investigation of male Gammarus fossarum, a freshwater crustacean, in response to endocrine disruptors. *J Proteome Res* **14**: 292-303.

Trapp, J., Gouveia, D., Almunia, C., Pible, O., Degli-Esposti, D., Gaillard, J.-C. et al. (2018) Digging deepper into the pyriproxyfen-response of the amphipod Gammarus fossarum with a next-generation ultra-high-field orbitrap analyser: new perspectives for environmental toxicoproteomics. *Frontiers in Environmental Sciences* **6**: 1-12.

Uller, T. (2008) Developmental plasticity and the evolution of parental effects. *Trends Ecol Evol* **23**: 432-438.

Urien, N., Lebrun, J.D., Fechner, L.C., Uher, E., Francois, A., Queau, H. et al. (2016) Environmental relevance of laboratory-derived kinetic models to predict trace metal bioaccumulation in gammarids: Field experimentation at a large spatial scale (France). *Water Res* **95**: 330-339.

Vigneron, A., Geffard, O., Coquery, M., Francois, A., Queau, H., and Chaumot, A. (2015) Evolution of cadmium tolerance and associated costs in a Gammarus fossarum population inhabiting a low-level contaminated stream. *Ecotoxicology* **24**: 1239-1249.

Wang, J., Zhang, P., Shen, Q., Wang, Q., Liu, D., Li, J., and Wang, L. (2013) The effects of cadmium exposure on the oxidative state and cell death in the gill of freshwater crab Sinopotamon henanense. *PLoS One* **8**: e64020.

Wright, J.C., and Choudhary, J.S. (2016) DecoyPyrat: Fast Non-redundant Hybrid Decoy Sequence Generation for Large Scale Proteomics. *J Proteomics Bioinform* **9**: 176-180.

Discussions et perspectives:

I-La protéogénomique, un nouvel évaluateur de la qualité des séquences et assemblages

La qualité de l'interprétation des données de protéomique dépend de la qualité des analyses par spectrométrie de masse, certes, mais surtout de la qualité et de la quantité d'informations contenues dans les bases de données de requêtes. Ces bases de données sont donc le point clé d'une analyse exhaustive des spectres. Dans le cadre de projets d'études où les données du transcriptome sont essentielles pour construire la base de données protéique, l'assemblage des séquences en transcrits est une étape déterminante de la qualité des données. Or, aujourd'hui, les méthodes utilisées pour l'évaluation de la qualité des assemblages sont basées essentiellement sur des paramètres métriques comme la longueur des séquences (N50) et leur distribution ou la recherche d'orthologues dans une base de données restreinte aux séquences connues (BUSCO). Les résultats sont seulement des indicateurs de la qualité des assemblages, ils ne sont pas suffisamment sensibles pour distinguer les assemblages les mieux adaptés à la construction de bases de données protéiques. De plus, comme vu précédemment, ces méthodes peuvent ne pas être suffisamment discriminantes pour mettre en place une auto-paramétrisation des outils de traitement de données de séquençage. A cet effet, l'utilisation de la protéomique comme un outil évaluateur de la qualité des assemblages s'avère avantageuse. En effet, cette méthodologie peut s'appliquer aussi bien à un organisme modèle qu'à un organisme ne possédant pas de génome de référence. Elle est tout à fait adaptée à l'optimisation des paramètres d'assemblages de transcriptome dont la finalité est de construire des bases de données protéiques pour l'interprétation des données issues d'analyses de protéome par spectrométrie de masse. Comme il a été démontré dans les travaux de thèse, la sensibilité accrue de cette méthode par rapport aux méthodologies transcriptomiques permet d'effectuer des choix sur les étapes de prétraitement des assemblages aussi bien que sur les logiciels d'assemblage comme l'ont mis en évidence Ma et al. 2018 et Luge et al. 2016.

Nous avons pu voir l'impact de l'utilisation de méthodologies d'évaluation protéogénomique dans le gain de l'interprétation des spectres générés en protéomique. Même si le gain en nombre de spectres ou de peptides reste faible, il est d'intérêt de collecter un maximum d'information dans lesquelles le moindre peptide pourrait permettre d'identifier une nouvelle protéine faiblement exprimée et complexe à détecter. En effet, la majeure partie des spectres attribués correspond à des protéines majoritaires du protéome qui représentent, pour la plupart, les transcrits majoritaires, plus simples à assembler. Afin de découvrir des nouvelles fonctions biologiques d'intérêt, ou de mettre en évidence la régulation fine des processus biologiques fondamentaux, il est important d'améliorer l'ensemble des méthodologies permettant d'approfondir la détection des protéines plus faiblement exprimées.

A partir de ce principe, nous pouvons proposer de mettre en œuvre une méthodologie d'évaluation automatique et de paramétrisation des assemblages. Dans un premier temps, une base de données peut être établie en regroupant les analyses de spectrométrie de masse obtenues à partir d'organismes sélectionnés afin d'être représentative des différents niveaux taxonomiques a l'instar de BUSCO. Cette base de données aurait pour objectif de fournir une première évaluation protéomique des assemblages d'organismes non-modèles. La méthodologie serait basée sur la sélection des organismes les plus proches au niveau taxonomique de l'organisme non-modèle étudié. Il est cependant à noter que l'attribution sera tout de même réduite lors de l'utilisation d'organismes taxonomiquement distants. Aussi, pour évaluer l'impact de ces distances, les données protéomiques

générées au sein du projet Protéogam pourraient servir d'étalonnage afin de valider l'écart taxonomique maximal acceptable des organismes sélectionnés. Les données protéomiques de référence d'organismes modèles peuvent être extraites de la base de données PRIDE. L'utilisateur pourrait alors choisir d'utiliser soit les données de l'organisme le plus proche soit ses propres données, spécifiques de l'organisme étudié, pour évaluer plus précisément l'assemblage. En parallèle, de multiples assemblages peuvent être effectués pour examiner, par exemple, toutes les combinaisons possibles de prétraitement, logiciels et paramètres d'assemblage. Cette étape étant limitante en temps et non-parallélisable, un ou plusieurs sous-ensembles des lectures initiales pourraient être utilisés afin d'obtenir un score pour chacune des combinaisons. La méthodologie serait donc essentiellement dépendante du nombre de lectures considérées dans chaque méthode ainsi que du nombre de réplicas à effectuer pour obtenir des résultats concordant avec la totalité des données. Afin de mettre au point ce workflow d'évaluation, les données issues des assemblages réalisés (> 200) dans la publication de Hölzer & Marz 2019 évaluant l'utilisation de 10 outils d'assemblage par les méthodologies actuelles peuvent être exploités. Ces assemblages, réalisés sur des organismes modèles tels que l'homme et la souris ou encore *Escherichia coli* peuvent mettre en avant des discordances ou concordances entre la méthodologie protéomique et les méthodologies transcriptomiques sur les performances des outils d'assemblage. De plus, l'accessibilité de l'ensemble des données peut permettre d'évaluer les paramètres de ce workflow tels que le nombre de lectures à sélectionner et le nombre de réplicas à effectuer pour approcher le résultat théorique final.

De nos jours, un enjeu majeur en terme de séquençage NGS est la génération de longues lectures. Cependant, les technologies actuelles fournissent des données à fort taux d'erreur comparées aux technologies courtes lectures. Dans le cas de la technologie MinIon (Oxford nanopore), un des soucis majeurs provient de la traduction du signal électrique en base nucléotidique. Il existe ainsi un nombre important de bases mal détectées et mal séquencées, dû à la forte complexité de l'interprétation de ce signal électrique. Les longues lectures contenant suffisamment d'informations pour représenter plusieurs peptides peuvent donc être corrigées par protéogénomique. Les bases mal détectées sont généralement identifiées par des nucléotides N et lorsque les séquences sont traduites *in silico*, les codons contenant un N sont traduits par un X. Ainsi, par l'association de données de protéomiques avec les données de transcriptomiques, les séquences contenant des X, identifiées à l'aide du moteur de recherche Mascot seront corrigées. Deux méthodes sont applicables. La première méthode utilise la capacité de Mascot à proposer l'acide aminé le plus proche des données de protéomique, la deuxième est de remplacer les X par l'ensemble des 4 nucléotides et de générer ensuite les peptides correspondants. La détection de ces derniers permettra de déterminer la bonne séquence nucléotidique et de corriger ces longues lectures.

Enfin, il est mis en évidence l'impact, somme toute faible, d'effectuer les prétraitements avant assemblage pour l'interprétation protéogénomique. Ces étapes de prétraitement représentent une copie quasi-totale des données de séquençage qui représentent plusieurs Gigaoctets d'espace disque. Il est donc important de prendre conscience que cette duplication représente, à terme, une quantité importante de données dupliquées. En effet, ces sauvegardes représentent de nos jours un enjeu majeur de par la limitation à terme de l'espace informatique global disponible pour une équipe. Il faut aussi considérer que le stockage et le temps utilisé pour ces prétraitements représentent une dépense énergétique non-négligeable. Il est donc crucial d'étudier l'intérêt et l'impact sur les analyses finales de ces traitements.

Enfin la protéogénomique est aujourd'hui une méthodologie essentielle dans la découverte de protéines provenant d'organismes non modèles, n'ayant pas à ce jour de génome séquencé. Cependant, l'ensemble des données et méthodologies mises au point à ce jour en protéogénomique

permettront aussi de mieux annoter les génomes qui pourraient être proposés dans les prochaines années. Il y a donc une nécessité de conserver et partager l'ensemble des données protéomiques d'organismes non modèles afin de préparer le futur en exploitant l'ensemble de ces connaissances.

II-Les données de gammares, une avancée pour l'écotoxicologie aquatique

L'acquisition des nombreuses données transcriptomiques et protéomiques, réparties sur 7 espèces différentes de gammares est une avancée essentielle pour l'utilisation de biomarqueurs toxicologiques *in natura*. En effet, ces données, exploitées pour analyser l'applicabilité des biomarqueurs identifiés à partir de *Gammarus fossarum* à toutes les espèces de gammares *in natura*, ont montré la conservation d'une portion non-négligeable de leurs séquences. Dans le cas de leur potentielle utilisation, cette observation doit cependant être complétée par une étude toxicologique pour vérifier que la réponse des différents organismes est similaire. En effet, si la séquence peptidique du biomarqueur est bien retrouvée au sein de différentes espèces de gammares, il est cependant encore impossible de confirmer que son implication dans la réponse aux différents polluants est la même que pour *Gammarus fossarum*. Néanmoins, ces données mettent en évidence une relation entre la conservation des séquences et les fonctions des protéines auxquels les peptides biomarqueurs sélectionnés appartiennent. En effet, on observe que certaines protéines, comme celles de l'osmorégulation, contiennent des peptides biomarqueurs transversaux. Par contre, les protéines spécifiques des femelles telles que les vitellogénines sont beaucoup plus variables au sein des gammares et ainsi peuvent permettre une discrimination des espèces sans nécessité de génotypage préalable. Enfin, ces données ont permis de montrer de nouvelles pistes pour le choix de biomarqueurs basées sur l'utilisation de séquences conservées ou de proposer d'éventuels substituts dont l'implication dans la réponse aux polluants doit encore être définie par protéomique ciblée.

Le séquençage individuel des gammares a été réalisé pour se concentrer sur l'assemblage le plus simple possible sans inclure la variabilité génétique. D'une part, ce choix permet de ne pas avoir l'impact de la diversité génétique au sein d'une espèce donnée par rapport à une autre dans la qualité de l'assemblage. D'autre part, ce choix, combiné au génotypage préalable de chaque échantillon, permet d'établir des bases de données sans mélange d'espèces. En effet, les gammares utilisés au sein de ces études possèdent peu de caractères morphologiques distinctifs et ne peuvent pas être sélectionnés uniquement sur ces critères. Par exemple, le précédent séquençage GFOSS n'ayant pas inclus cette distinction possède des séquences de *Gammarus fossarum* et *pulex* au sein de la base de données ne pouvant pas permettre les études inter-espèces effectuées durant notre projet.

Les études entre les protéomes des *Gammarus pulex* ont permis de mettre en évidence la modulation de protéines entre un site pollué (Brameloup) et un site de référence (Pollon). Lors du séquençage, une femelle et un mâle de chaque site ont aussi été séquencés. Afin de simplifier l'étude, le transcriptome du site de Pollon a été utilisé. Cependant, l'existence des données transcriptomiques pour les deux sites peut permettre des études supplémentaires. D'une part, une comparaison de l'expression des transcrits pourrait être réalisée entre les deux sites. Cependant, il n'existe pas de réplicas biologiques permettant d'établir statistiquement des différences. A défaut de pouvoir réaliser ce type d'étude, la corrélation des transcrits et protéines modulées entre les deux sites pourrait être mise en évidence. En revanche, l'analyse de la conservation des séquences de transcrits entre les deux sites peut être réalisée. En effet, dans le cas des organismes *Gammarus pulex* provenant de Brameloup une diversité génomique et donc transcriptomique plus restreinte est attendue dans le cas où les polluants exercent une pression négative. En protéomique, la diversité au sein des populations ne suit

pas cette règle et il se trouve que la diversité en utilisant la variation d'expression est plus grande dans la population Brameloup (site contaminé). L'utilisation de séquences orthologues à une branche externe d'un organisme modèle (*Daphnia pulex*) pourrait permettre de mettre en évidence des pressions sélectives, plus fortes sur les gènes en étudiant le taux de mutation synonyme telle qu'a pu le faire l'étude d'Ofria et al. 2003.

De nombreuses données protéomiques générées peuvent encore être exploitées suite à cette thèse. Deux projet principaux sont prévus. Possédant les protéomes individuels globaux de 20 individus (10 mâles, 10 femelles) pour *Gammarus fossarum B*, *Gammarus Pulex*, *Gammarus marinus* et *Echinogammarus berilloni* une première étude de la biodiversité inter-espèces pourrait être réalisée. Au sein de cette étude la variabilité intra-population protéomique de chacune de ces espèces doit être évaluée avant d'analyser la stabilité des protéomes de chacune des espèces. Avec la création d'une base de données, fusionnée pour l'étude, l'identification de toutes ces espèces pourrait être faite sans génotypage. Pour cela d'une part, la base de données de core-peptides devra être créée, permettant une étude basée uniquement sur les peptides communs entre les différentes espèces sans se soucier de l'origine taxonomique de celles-ci. D'autre part, la base de données de pan-peptides servirait pour évaluer l'origine taxonomique d'un individu non génotypé et, en fonction du nombre de peptides spécifiques à un taxon donné, établir son origine. Pour cela, le protéome de 10 individus (5 mâles et 5 femelles) pour chaque espèce pourrait être utilisé pour définir les bases de données et 10 autres pour valider la méthodologie.

En ce qui concerne le transcriptome des 14 espèces de gammare, une exploitation taxonomique est envisageable. En effet, Rõdelsperger et al. 2018 utilisent les transcriptomes assemblés, issus de plusieurs lignées de nématodes afin d'en définir la taxonomie. L'avantage de la définition taxonomique ainsi réalisée repose sur la prise en compte de l'histoire évolutive d'un grand ensemble de gènes. En effet, la taxonomie réalisée habituellement s'applique à l'étude des séquences d'un ou quelques gènes pour une grande proportion d'individus. En réalité, cette taxonomie a pour défaut de représenter uniquement l'histoire évolutive de ces quelques gènes qui peuvent diverger par rapport à l'histoire évolutive de l'espèce. En ce qui concerne l'approche phylotranscriptomique, l'utilisation d'un grand ensemble de gènes orthologues retrouvés à travers les assemblages des différentes espèces permet d'établir une taxonomie plus proche de celle de l'espèce et non pas réduite à quelques gènes. En revanche, le coût du séquençage ne permet cependant pas d'obtenir autant d'individus que dans le cas de la taxonomie habituelle pour établir leur histoire évolutive. Il serait donc intéressant de comparer une taxonomie établie par phylotranscriptomique des 7 espèces avec la taxonomie actuelle basée sur la séquence de la sous-unité Cytochrome Oxidase I. A partir de ces résultats, les distances, établies sur les différents gènes orthologues, peuvent être utilisées afin de définir des gènes ayant une histoire évolutive stable. Ces gènes, une fois définis, peuvent être des candidats potentiels pour la définition de biomarqueurs transversaux supplémentaires au sein des gammares.

En parallèle, les données transcriptomiques générées permettent des analyses supplémentaires. Par exemple l'exploitation conjointe des données de séquençage issues des transcriptomes de mâles et de femelles permettrai l'assemblage de nouveaux transcrits. Le taux de spectres attribués issus de cette analyse pourrait être comparé pour chaque transcriptome indépendamment de la fusion de ceux-ci (mâles et femelles assemblées séparément puis associés) ou alors d'un assemblage commun. D'autre part, les assemblages réalisés sur les espèces cryptiques de *Gammarus fossarum* (A, B, C) peuvent être exploités avec des approches phylotranscriptomiques afin de définir les distances taxonomiques et déterminer les principales différences au sein du patrimoine génétique de ces espèces.

III-Protéogénomique des populations, perspectives

L'étude de l'expression différentielle a permis de mettre en évidence une hétérogénéité dans la diversité des réponses biologiques en fonction des sites de prélèvement des organismes. La variation de cette diversité, plus élevée dans le site Brameloup, pourrait être due à la présence de polluants (cadmium). Cette étude permet de mettre en évidence la nécessité de vérifier la variabilité des protéomes au sein de différentes populations afin de pouvoir adapter le nombre d'individus à sélectionner pour des études de protéomique différentielle. En effet, si cette diversité augmente, il est intéressant d'utiliser un plus grand nombre d'individus afin de mettre en évidence les différents mécanismes de réaction établis au sein d'une même population dans la zone polluée. Pour bien établir la diversité standard d'une population, il est nécessaire d'obtenir des données protéomiques inter-populations d'une espèce. Dans le cas de *Gammarus fossarum B* qui se trouve être l'espèce utilisée pour la définition des biomarqueurs, 20 (10 mâles, 10 femelles) protéomes sont disponibles pour 3 sites différents. Connaissant les différentes caractéristiques physico-chimiques des sites, une étude d'expression différentielle entre les 3 sites serait réalisable. De plus, en intégrant la variabilité protéomique de chacun des 3 sites, cette étude permettrait l'analyse de la variabilité intra espèces de notre organisme sentinelle. Cette information aiderait à mieux estimer le risque et le nombre de peptides biomarqueurs à utiliser afin de définir la présence de polluants *in natura* en utilisant cette espèce.

Les études sur un grand nombre d'individus ont mis en évidence les limites actuelles de l'analyse individualisée des échantillons. En effet, les bases de données générées par les protocoles d'assemblages RNA-seq possèdent de nombreux variants d'épissage de séquences. Ce phénomène induit, après traduction, la présence de nombreuses protéines similaires qui possèdent la quasi-totalité de leurs peptides en commun. Or en protéomique, lors de l'analyse de l'attribution des peptides aux protéines un système de classification est utilisé. En effet, parmi les protéines d'un groupe qui partagent les mêmes peptides, c'est la protéine qui possède le plus de peptides différents détectés qui est identifiée. Les autres protéines, très souvent isoformes de la protéine leader, sont considérées comme des sous-ensembles de celle-ci et ne sont pas sélectionnées. Cependant, lors de l'exploitation d'un grand nombre d'échantillons protéomiques avec des bases de données possédant des isoformes, cette protéine leader peut ne pas être la même dans tous les échantillons en fonction des peptides détectés. Il faut donc rétablir les attributions des peptides aux protéines et de leur nombre de spectres avant toute étude d'expression différentielle. Un logiciel nommé Proteocount a été développé au cours de cette thèse afin de traiter ce point. Cet outil permet de recompter le nombre de spectres attribués à travers plusieurs échantillons en réattribuant les spectres et les peptides à toutes les protéines validées sans appliquer de parcimonie. Cependant, il est important de prendre en compte que les spectres sont comptés de multiples fois à travers les échantillons au niveau des comptages de spectres par protéines dans les résultats. Afin de corriger cela, une parcimonie pour l'attribution des peptides communs est applicable. La parcimonie repose principalement sur l'utilisation des peptides discriminants des protéines et de leur nombre de spectres attribués pour établir un ratio de présence des protéines et ainsi distribuer les spectres en fonction. Le souci ici est la disparité de la détection des différents peptides en spectrométrie de masse. En effet, les peptides d'une même protéine ne sont pas toujours visibles à un même niveau. Une autre solution est de ne compter que les spectres attribués à des peptides différenciants. Cette solution induit cependant une perte importante en informations. Enfin, cette analyse peut soit être appliquée échantillon par échantillon, soit à l'ensemble des échantillons. Dans ce dernier cas, le jeu de données est complet, mais les isoformes peuvent être différentes si les conditions expérimentales changent. Ainsi, le choix de la stratégie n'est pas évident.

Afin de traiter le problème de la répartition des spectres une autre solution proposée est celle de considérer le nombre de spectres par fonction biologique. En effet des protéines isoformes ou possédant un grand nombre de peptides communs ont une fonction biologique similaire. De plus, la question biologique repose essentiellement sur la compréhension des mécanismes biologiques impactés par la présence de polluants dans un premier temps. Le souci de ce type d'étude est l'absence actuellement d'outil permettant une analyse intégrative de ce type. En effet, il serait intéressant de mettre au point un outil permettant d'effectuer automatiquement l'annotation puis la redistribution des spectres en fonction des fonctions biologiques des protéines et l'étude statistique des répartitions entre les différentes conditions expérimentales incluant la représentation graphique. Cependant, il est à noter que ce type d'approche entraîne une perte de sensibilité de détection dans le cas où certaines protéines définies comme ayant une même fonction peuvent avoir des variations différentielles antagonistes entre deux conditions. De plus, dans le cas d'étude des organismes non-modèles, un nombre important de protéines peuvent ne pas avoir d'annotation fonctionnelle, et donc entraîner une perte d'information.

Enfin ces derniers travaux ont permis d'entrevoir l'intérêt d'une protéomique réalisée sur des échantillons individualisés. Dans notre cas, cette étude a permis de mettre en évidence des mécanismes de réponses aux stress environnementaux différents au sein d'une même population. Un échantillonnage plus grand permettra d'approfondir les connaissances de ces différents mécanismes et d'intégrer un plus grand nombre d'échantillons jusqu'alors considérés comme des cas particuliers. En effet, ces derniers ont, en réalité, développé un mécanisme adaptatif différent pour lequel nous n'avions pas suffisamment échantillonné pour détecter les différences par rapport à la référence. Il serait donc intéressant d'observer la répartition des distances obtenues en utilisant une centaine d'échantillons et donc utiliser une approche basée sur l'analyse de population. A cet effet l'ensemble de ce projet de thèse permet d'imaginer le traitement rapide d'un grand nombre d'échantillons grâce aux optimisations réalisées sur la base de données. De plus, il existe des outils statistiques adaptés à l'analyse de larges populations. C'est le cas des outils disponibles dans le package mixOmics avec par exemple l'analyse discriminante par régression des moindres carrés partiels (PLS-DA). Cette approche assure la classification des organismes plus sensible afin de comprendre l'ensemble des mécanismes de réponses aux stress environnementaux d'une population.

Conclusion

Ce projet de thèse qui s'est déroulé au sein du projet ANR Protéogam a permis d'approfondir les connaissances précédemment acquises lors des projets de Judith Trapp et de Duarte Gouveia. L'orientation bioinformatique de ce projet a permis d'étudier plus en profondeur la diversité moléculaire des Gammares dans l'objectif d'exploiter l'information issue d'échantillonnages *in natura*. A cet effet, dans un premier temps, les données protéomiques de 164 individus représentant 7 espèces de Gammares ont été générées. Afin d'interpréter ces données, les transcriptomes d'un mâle et d'une femelle de chacune de ces espèces ont été séquencés. Pour interpréter l'ensemble de ces données avec une reproductibilité fiable j'ai dû installer, paramétrer et mettre au point une plateforme Galaxy/Docker en tenant compte des consignes de sécurité de l'organisme d'accueil.

Afin de maximiser l'interprétation des données protéomiques en utilisant les bases de données produites à partir des données transcriptomiques, nous avons mis au point une méthodologie d'évaluation de la qualité d'assemblage et de la génération des bases de données. Cette méthodologie repose sur l'exploitation du taux d'attribution protéomique pour optimiser les étapes liées à la construction des bases de données. L'apport principal de cette méthodologie repose sur une sensibilité accrue dans la détection de la qualité de l'assemblage par rapport aux méthodes usuelles.

Dans un second temps l'exploitation et la définition des optima pour le traitement et la génération de la base de données ont permis de construire les 14 transcriptomes ainsi que les bases de données de requêtes protéogénomiques dédiées. Ces transcriptomes et les bases de données ont pu être mis à disposition sur NCBI. L'ensemble des données brutes est aussi accessible sur ce même site afin de permettre d'éventuels projets complémentaires.

Une fois les bases de données protéogénomiques de chaque espèce obtenue, nous avons validé la présence des séquences de biomarqueurs définis pour *Gammarus fossarum B*. Pour cela, nous avons mis au point une méthodologie d'étude de la variabilité inter-espèce au sein des Gammares avec pour cible les séquences peptidiques biomarqueurs potentiels précédemment définies pour *Gammarus fossarum B*. Nous avons de plus proposé des séquences, substituts potentiels afin de permettre de nouvelles pistes pour le développement de nouvelles références de dosage dans le cas de biomarqueurs non conservés.

Enfin, l'analyse protéomique individuelle de l'ensemble des échantillons nous a permis d'explorer une nouvelle piste d'analyse des protéomes avec un point de vue axé sur l'analyse de population. Ces analyses nous ont permis de mettre en évidence la variabilité intra-population notamment en réponse face à des stress environnementaux. Ces méthodologies mettent en évidence l'importance d'un échantillonnage individuel et la valeur ajoutée d'une étude à plus large échelle afin de comprendre au mieux les différents mécanismes d'adaptation mis en place en réponse aux stress.

Références

Afgan, E., Baker, D., Batut, B., van den Beek, M., Bouvier, D., Cech, M., . . . Blankenberg, D. (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res, 46(W1), W537-W544. doi: 10.1093/nar/gky379

Alexeev, D., Kostrjukova, E., Aliper, A., Popenko, A., Bazaleev, N., Tyakht, A., . . . Govorun, V. (2012). Application of Spiroplasma melliferum proteogenomic profiling for the discovery of virulence factors and pathogenicity mechanisms in host-associated spiroplasmas. J Proteome Res, 11(1), 224-236. doi: 10.1021/pr2008626

Alfaro, J. A., Sinha, A., Kislinger, T., & Boutros, P. C. (2014). Onco-proteogenomics: cancer proteomics joins forces with genomics. Nat Methods, 11(11), 1107-1113. doi: 10.1038/nmeth.3138

Armengaud, J. (2009). A perfect genome annotation is within reach with the proteomics and genomics alliance. Curr Opin Microbiol, 12(3), 292-300. doi: 10.1016/j.mib.2009.03.005

Armengaud, J., Trapp, J., Pible, O., Geffard, O., Chaumot, A., & Hartmann, E. M. (2014). Non-model organisms, a species endangered by proteogenomics. J Proteomics, 105, 5-18. doi: 10.1016/j.jprot.2014.01.007

Baerenfaller, K., Grossmann, J., Grobei, M. A., Hull, R., Hirsch-Hoffmann, M., Yalovsky, S., . . . Baginsky, S. (2008). Genome-scale proteomics reveals Arabidopsis thaliana gene models and proteome dynamics. Science, 320(5878), 938-941. doi: 10.1126/science.1157956

Bambino, K., & Chu, J. (2017). Zebrafish in Toxicology and Environmental Health. Curr Top Dev Biol, 124, 331-367. doi: 10.1016/bs.ctdb.2016.10.007

Baudet, M., Ortet, P., Gaillard, J. C., Fernandez, B., Guerin, P., Enjalbal, C., . . . Armengaud, J. (2010). Proteomics-based refinement of Deinococcus deserti genome annotation reveals an unwonted use of non-canonical translation initiation codons. Mol Cell Proteomics, 9(2), 415-426. doi: 10.1074/mcp.M900359-MCP200

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., . . . Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. Nature, 456(7218), 53-59. doi: 10.1038/nature07517

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 30(15), 2114-2120. doi: 10.1093/bioinformatics/btu170

Brunner, E., Ahrens, C. H., Mohanty, S., Baetschmann, H., Loevenich, S., Potthast, F., . . . Aebersold, R. (2007). A high-quality catalog of the Drosophila melanogaster proteome. Nat Biotechnol, 25(5), 576-583. doi: 10.1038/nbt1300

Carvalho, P. C., Fischer, J. S., Chen, E. I., Yates, J. R., 3rd, & Barbosa, V. C. (2008). PatternLab for proteomics: a tool for differential shotgun proteomics. BMC Bioinformatics, 9, 316. doi: 10.1186/1471-2105-9-316

Castellana, N., & Bafna, V. (2010). Proteogenomics to discover the full coding content of genomes: a computational perspective. J Proteomics, 73(11), 2124-2135. doi: 10.1016/j.jprot.2010.06.007

Castellana, N. E., Payne, S. H., Shen, Z., Stanke, M., Bafna, V., & Briggs, S. P. (2008). Discovery and revision of Arabidopsis genes by proteogenomics. Proc Natl Acad Sci U S A, 105(52), 21034-21038. doi: 10.1073/pnas.0811066106

Castellana, N. E., Shen, Z., He, Y., Walley, J. W., Cassidy, C. J., Briggs, S. P., & Bafna, V. (2014). An automated proteogenomic method uses mass spectrometry to reveal novel genes in Zea mays. Mol Cell Proteomics, 13(1), 157-167. doi: 10.1074/mcp.M113.031260

Charnot, A., Gouveia, D., Armengaud, J., Almunia, C., Chaumot, A., Lemoine, J., . . . Salvador, A. (2017). Multiplexed assay for protein quantitation in the invertebrate Gammarus fossarum by liquid chromatography coupled to tandem mass spectrometry. Anal Bioanal Chem, 409(16), 3969-3991. doi: 10.1007/s00216-017-0348-0

Chen, M., Hu, Y., Liu, J., Wu, Q., Zhang, C., Yu, J., . . . Wu, J. (2015). Improvement of genome assembly completeness and identification of novel full-length protein-coding genes by RNA-seq in the giant panda genome. Sci Rep, 5, 18019. doi: 10.1038/srep18019

Christie-Oleza, J. A., Miotello, G., & Armengaud, J. (2013). Proteogenomic definition of biomarkers for the large Roseobacter clade and application for a quick screening of new environmental isolates. J Proteome Res, 12(11), 5331-5339. doi: 10.1021/pr400554e

Connon, R. E., Geist, J., & Werner, I. (2012). Effect-based tools for monitoring and predicting the ecotoxicological effects of chemicals in the aquatic environment. Sensors (Basel), 12(9), 12741-12771. doi: 10.3390/s120912741

Cox, J., & Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol, 26(12), 1367-1372. doi: 10.1038/nbt.1511

Crappe, J., Ndah, E., Koch, A., Steyaert, S., Gawron, D., De Keulenaer, S., . . . Menschaert, G. (2015). PROTEOFORMER: deep proteome coverage through ribosome profiling and MS integration. Nucleic Acids Res, 43(5), e29. doi: 10.1093/nar/gku1283

da Veiga Leprevost, F., Gruning, B. A., Alves Aflitos, S., Rost, H. L., Uszkoreit, J., Barsnes, H., . . . Perez-Riverol, Y. (2017). BioContainers: an open-source and community-driven framework for software standardization. Bioinformatics, 33(16), 2580-2582. doi: 10.1093/bioinformatics/btx192

Dai, Y. J., Jia, Y. F., Chen, N., Bian, W. P., Li, Q. K., Ma, Y. B., . . . Pei, D. S. (2014). Zebrafish as a model system to study toxicology. Environ Toxicol Chem, 33(1), 11-17. doi: 10.1002/etc.2406

Dalzon, B., Bons, J., Diemer, H., Collin-Faure, V., Marie-Desvergne, C., Dubosson, M., . . . Rabilloud, T. (2019). A Proteomic View of Cellular Responses to Anticancer Quinoline-Copper Complexes. Proteomes, 7(2). doi: 10.3390/proteomes7020026

de Groot, A., Dulermo, R., Ortet, P., Blanchard, L., Guerin, P., Fernandez, B., . . . Armengaud, J. (2009). Alliance of proteomics and genomics to unravel the specificities of Sahara bacterium Deinococcus deserti. PLoS Genet, 5(3), e1000434. doi: 10.1371/journal.pgen.1000434

Dedourge-Geffard, O., Palais, F., Biagianti-Risbourg, S., Geffard, O., & Geffard, A. (2009). Effects of metals on feeding rate and digestive enzymes in Gammarus fossarum: an in situ experiment. Chemosphere, 77(11), 1569-1576. doi: 10.1016/j.chemosphere.2009.09.042

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., . . . Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics, 29(1), 15-21. doi: 10.1093/bioinformatics/bts635

Elias, J. E., & Gygi, S. P. (2010). Target-decoy search strategy for mass spectrometry-based proteomics. Methods Mol Biol, 604, 55-71. doi: 10.1007/978-1-60761-444-9_5

Eng, J. K., McCormack, A. L., & Yates, J. R. (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J Am Soc Mass Spectrom, 5(11), 976-989. doi: 10.1016/1044-0305(94)80016-2

Evans, V. C., Barker, G., Heesom, K. J., Fan, J., Bessant, C., & Matthews, D. A. (2012). De novo derivation of proteomes from transcriptomes for transcript and protein identification. Nat Methods, 9(12), 1207-1211. doi: 10.1038/nmeth.2227

Fu, S., Liu, X., Luo, M., Xie, K., Nice, E. C., Zhang, H., & Huang, C. (2017). Proteogenomic studies on cancer drug resistance: towards biomarker discovery and target identification. Expert Rev Proteomics, 14(4), 351-362. doi: 10.1080/14789450.2017.1299006

Gallien, S., Perrodou, E., Carapito, C., Deshayes, C., Reyrat, J. M., Van Dorsselaer, A., . . . Lecompte, O. (2009). Ortho-proteogenomics: multiple proteomes investigation through orthology and a new MS-based protocol. Genome Res, 19(1), 128-135. doi: 10.1101/gr.081901.108

Gatto, L., & Christoforou, A. (2014). Using R and Bioconductor for proteomics data analysis. Biochim Biophys Acta, 1844(1 Pt A), 42-51. doi: 10.1016/j.bbapap.2013.04.032

Ghali, F., Krishna, R., Perkins, S., Collins, A., Xia, D., Wastling, J., & Jones, A. R. (2014). ProteoAnnotator--open source proteogenomics annotation software supporting PSI standards. Proteomics, 14(23-24), 2731-2741. doi: 10.1002/pmic.201400265

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet, 17(6), 333-351. doi: 10.1038/nrg.2016.49

Gouveia, D., Almunia, C., Cogne, Y., Pible, O., Degli-Esposti, D., Salvador, A., . . . Armengaud, J. (2019). Ecotoxicoproteomics: A decade of progress in our understanding of anthropogenic impact on the environment. J Proteomics, 198, 66-77. doi: 10.1016/j.jprot.2018.12.001

Gouveia, D., Chaumot, A., Charnot, A., Almunia, C., Francois, A., Navarro, L., . . . Geffard, O. (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(22), 13417-13426. doi: 10.1021/acs.est.7b03736

Gouveia, D., Chaumot, A., Charnot, A., Queau, H., Armengaud, J., Almunia, C., . . . Geffard, O. (2017). Assessing the relevance of a multiplexed methodology for proteomic biomarker measurement in the invertebrate species Gammarus fossarum: A physiological and ecotoxicological study. Aquat Toxicol, 190, 199-209. doi: 10.1016/j.aquatox.2017.07.007

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., . . . Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol, 29(7), 644-652. doi: 10.1038/nbt.1883

Grimplet, J., Gaspar, J. W., Gancel, A. L., Sauvage, F. X., & Romieu, C. (2005). Including mutations from conceptually translated expressed sequence tags into orthologous proteins improves the preliminary assignment of peptide mass fingerprints on non-model genomes. Proteomics, 5(11), 2769-2777. doi: 10.1002/pmic.200401177

Gruning, B., Dale, R., Sjodin, A., Chapman, B. A., Rowe, J., Tomkins-Tinch, C. H., . . . Bioconda, T. (2018). Bioconda: sustainable and comprehensive software distribution for the life sciences. Nat Methods, 15(7), 475-476. doi: 10.1038/s41592-018-0046-7

Guillot, L., Delage, L., Viari, A., Vandenbrouck, Y., Com, E., Ritter, A., . . . Pineau, C. (2019). Peptimapper: proteogenomics workflow for the expert annotation of eukaryotic genomes. BMC Genomics, 20(1), 56. doi: 10.1186/s12864-019-5431-9

Halvey, P. J., Wang, X., Wang, J., Bhat, A. A., Dhawan, P., Li, M., . . . Slebos, R. J. (2014). Proteogenomic analysis reveals unanticipated adaptations of colorectal tumor cells to deficiencies in DNA mismatch repair. Cancer Res, 74(1), 387-397. doi: 10.1158/0008-5472.CAN-13-2488

Hansen, K. D., Brenner, S. E., & Dudoit, S. (2010). Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Res, 38(12), e131. doi: 10.1093/nar/gkq224

Helmy, M., Sugiyama, N., Tomita, M., & Ishihama, Y. (2010). Onco-proteogenomics: a novel approach to identify cancer-specific mutations combining proteomics and transcriptome deep sequencing. Genome Biol, 11. doi: Artn P17

10.1186/Gb-2010-11-S1-P17

Holzer, M., & Marz, M. (2019). De novo transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-Seq assemblers. Gigascience, 8(5). doi: 10.1093/gigascience/giz039

Jaffe, J. D., Berg, H. C., & Church, G. M. (2004). Proteogenomic mapping as a complementary method to perform genome annotation. Proteomics, 4(1), 59-77. doi: 10.1002/pmic.200300511

Jagtap, P., Goslinga, J., Kooren, J. A., McGowan, T., Wroblewski, M. S., Seymour, S. L., & Griffin, T. J. (2013). A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies. Proteomics, 13(8), 1352-1357. doi: 10.1002/pmic.201200352

Johansson, H. J., Socciarelli, F., Vacanti, N. M., Haugen, M. H., Zhu, Y., Siavelis, I., . . . Lehtio, J. (2019). Breast cancer quantitative proteome and proteogenomic landscape. Nat Commun, 10(1), 1600. doi: 10.1038/s41467-019-09018-y

Kalume, D. E., Peri, S., Reddy, R., Zhong, J., Okulate, M., Kumar, N., & Pandey, A. (2005). Genome annotation of Anopheles gambiae using mass spectrometry-derived data. BMC Genomics, 6, 128. doi: 10.1186/1471-2164-6-128

Kim, M. S., Pinto, S. M., Getnet, D., Nirujogi, R. S., Manda, S. S., Chaerkady, R., . . . Pandey, A. (2014). A draft map of the human proteome. Nature, 509(7502), 575-581. doi: 10.1038/nature13302

Kopylova, E., Noe, L., & Touzet, H. (2012). SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics, 28(24), 3211-3217. doi: 10.1093/bioinformatics/bts611

Koster, J., & Rahmann, S. (2012). Snakemake--a scalable bioinformatics workflow engine. Bioinformatics, 28(19), 2520-2522. doi: 10.1093/bioinformatics/bts480

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., . . . Marra, M. A. (2009). Circos: an information aesthetic for comparative genomics. Genome Res, 19(9), 1639-1645. doi: 10.1101/gr.092759.109

Kurtzer, G. M., Sochat, V., & Bauer, M. W. (2017). Singularity: Scientific containers for mobility of compute. PLoS One, 12(5), e0177459. doi: 10.1371/journal.pone.0177459

Lacaze, E., Geffard, O., Goyet, D., Bony, S., & Devaux, A. (2011). Linking genotoxic responses in Gammarus fossarum germ cells with reproduction impairment, using the Comet assay. Environ Res, 111(5), 626-634. doi: 10.1016/j.envres.2011.03.012

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol, 10(3), R25. doi: 10.1186/gb-2009-10-3-r25

LeDuc, R. D., Schwammle, V., Shortreed, M. R., Cesnik, A. J., Solntsev, S. K., Shaw, J. B., . . . Tsybin, Y. O. (2018). ProForma: A Standard Proteoform Notation. J Proteome Res, 17(3), 1321-1325. doi: 10.1021/acs.jproteome.7b00851

Lepretre, M., Almunia, C., Armengaud, J., Salvador, A., Geffard, A., & Palos-Ladeiro, M. (2019). The immune system of the freshwater zebra mussel, Dreissena polymorpha, decrypted by proteogenomics of hemocytes and plasma compartments. J Proteomics, 202, 103366. doi: 10.1016/j.jprot.2019.04.016

Levitsky, L. I., Klein, J. A., Ivanov, M. V., & Gorshkov, M. V. (2019). Pyteomics 4.0: Five Years of Development of a Python Proteomics Framework. J Proteome Res, 18(2), 709-714. doi: 10.1021/acs.jproteome.8b00717

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25(14), 1754-1760. doi: 10.1093/bioinformatics/btp324

Lopez-Barea, J. (1995). Biomarkers in ecotoxicology: an overview. Arch Toxicol Suppl, 17, 57-79.

Low, S. C., & Berry, M. J. (1996). Knowing when not to stop: selenocysteine incorporation in eukaryotes. Trends Biochem Sci, 21(6), 203-208.

Ma, J., Saghatelian, A., & Shokhirev, M. N. (2018). The influence of transcript assembly on the proteogenomics discovery of microproteins. PLoS One, 13(3), e0194518. doi: 10.1371/journal.pone.0194518

Macmanes, M. D. (2014). On the optimal trimming of high-throughput mRNA sequence data. Front Genet, 5, 13. doi: 10.3389/fgene.2014.00013

Marques, A. T., Anjo, S. I., Bhide, M., Varela Coelho, A., Manadas, B., Lecchi, C., . . . Ceciliani, F. (2019). Changes in the intestinal mucosal proteome of turkeys (Meleagris gallopavo) infected with haemorrhagic enteritis virus. Vet Immunol Immunopathol, 213, 109880. doi: 10.1016/j.vetimm.2019.06.001

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., . . . DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res, 20(9), 1297-1303. doi: 10.1101/gr.107524.110

Menon, R., & Omenn, G. S. (2010). Proteomic characterization of novel alternative splice variant proteins in human epidermal growth factor receptor 2/neu-induced breast cancers. Cancer Res, 70(9), 3440-3449. doi: 10.1158/0008-5472.CAN-09-2631

Menschaert, G., & Fenyo, D. (2017). Proteogenomics from a bioinformatics angle: A growing field. Mass Spectrom Rev, 36(5), 584-599. doi: 10.1002/mas.21483

Mo, F., Hong, X., Gao, F., Du, L., Wang, J., Omenn, G. S., & Lin, B. (2008). A compatible exon-exon junction database for the identification of exon skipping events using tandem mass spectrum data. BMC Bioinformatics, 9, 537. doi: 10.1186/1471-2105-9-537

Muth, T., & Renard, B. Y. (2018). Evaluating de novo sequencing in proteomics: already an accurate alternative to database-driven peptide identification? Brief Bioinform, 19(5), 954-970. doi: 10.1093/bib/bbx033

Nagaraj, S. H., Waddell, N., Madugundu, A. K., Wood, S., Jones, A., Mandyam, R. A., . . . Grimmond, S. M. (2015). PGTools: A Software Suite for Proteogenomic Data Analysis and Visualization. J Proteome Res, 14(5), 2255-2266. doi: 10.1021/acs.jproteome.5b00029

Nanduri, B., Lawrence, M. L., Vanguri, S., Pechan, T., & Burgess, S. C. (2005). Proteomic analysis using an unfinished bacterial genome: the effects of subminimum inhibitory concentrations of antibiotics on Mannheimia haemolytica virulence factor expression. Proteomics, 5(18), 4852-4863. doi: 10.1002/pmic.200500112

Nesvizhskii, A. I. (2014). Proteogenomics: concepts, applications and computational strategies. Nat Methods, 11(11), 1114-1125. doi: 10.1038/nmeth.3144

Nielsen, P., & Krogh, A. (2005). Large-scale prokaryotic gene prediction and comparison to genome annotation. Bioinformatics, 21(24), 4322-4329. doi: 10.1093/bioinformatics/bti701

Ofria, C., Adami, C., & Collier, T. C. (2003). Selective pressures on genomes in molecular evolution. J Theor Biol, 222(4), 477-483.

Peterson, E. S., McCue, L. A., Schrimpe-Rutledge, A. C., Jensen, J. L., Walker, H., Kobold, M. A., . . . Webb-Robertson, B. J. (2012). VESPA: software to facilitate genomic annotation of prokaryotic organisms through integration of proteomic and transcriptomic data. BMC Genomics, 13, 131. doi: 10.1186/1471-2164-13-131

Risk, B. A., Spitzer, W. J., & Giddings, M. C. (2013). Peppy: proteogenomic search software. J Proteome Res, 12(6), 3019-3025. doi: 10.1021/pr400208w

Rison, S. C., Mattow, J., Jungblut, P. R., & Stoker, N. G. (2007). Experimental determination of translational starts using peptide mass mapping and tandem mass spectrometry within the proteome of Mycobacterium tuberculosis. Microbiology, 153(Pt 2), 521-528. doi: 10.1099/mic.0.2006/001537-0

Robinson, J. T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. Nat Biotechnol, 29(1), 24-26. doi: 10.1038/nbt.1754

Rodelsperger, C., Roseler, W., Prabh, N., Yoshida, K., Weiler, C., Herrmann, M., & Sommer, R. J. (2018). Phylotranscriptomics of Pristionchus Nematodes Reveals Parallel Gene Loss in Six Hermaphroditic Lineages. Curr Biol, 28(19), 3123-3127 e3125. doi: 10.1016/j.cub.2018.07.041

Rubiano-Labrador, C., Bland, C., Miotello, G., Guerin, P., Pible, O., Baena, S., & Armengaud, J. (2014). Proteogenomic insights into salt tolerance by a halotolerant alpha-proteobacterium isolated from an Andean saline spring. J Proteomics, 97, 36-47. doi: 10.1016/j.jprot.2013.05.020

Ruggles, K. V., Krug, K., Wang, X., Clauser, K. R., Wang, J., Payne, S. H., . . . Mani, D. R. (2017). Methods, Tools and Current Perspectives in Proteogenomics. Mol Cell Proteomics, 16(6), 959-981. doi: 10.1074/mcp.MR117.000024

Sanders, W. S., Wang, N., Bridges, S. M., Malone, B. M., Dandass, Y. S., McCarthy, F. M., . . . Burgess, S. C. (2011). The proteogenomic mapping tool. BMC Bioinformatics, 12, 115. doi: 10.1186/1471-2105-12-115

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., . . . Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res, 13(11), 2498-2504. doi: 10.1101/gr.1239303

Sheynkman, G. M., Johnson, J. E., Jagtap, P. D., Shortreed, M. R., Onsongo, G., Frey, B. L., . . . Smith, L. M. (2014). Using Galaxy-P to leverage RNA-Seq for the discovery of novel protein variations. BMC Genomics, 15, 703. doi: 10.1186/1471-2164-15-703

Sheynkman, G. M., Shortreed, M. R., Cesnik, A. J., & Smith, L. M. (2016). Proteogenomics: Integrating Next-Generation Sequencing and Mass Spectrometry to Characterize Human Proteomic Variation. Annu Rev Anal Chem (Palo Alto Calif), 9(1), 521-545. doi: 10.1146/annurev-anchem-071015-041722

Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics, 31(19), 3210-3212. doi: 10.1093/bioinformatics/btv351

Smith-Unna, R., Boursnell, C., Patro, R., Hibberd, J. M., & Kelly, S. (2016). TransRate: reference-free quality assessment of de novo transcriptome assemblies. Genome Res, 26(8), 1134-1144. doi: 10.1101/gr.196469.115

Smith, L. M., Sanders, J. Z., Kaiser, R. J., Hughes, P., Dodd, C., Connell, C. R., . . . Hood, L. E. (1986). Fluorescence detection in automated DNA sequence analysis. Nature, 321(6071), 674-679. doi: 10.1038/321674a0

Soares, R., Franco, C., Pires, E., Ventosa, M., Palhinhas, R., Koci, K., . . . Varela Coelho, A. (2012). Mass spectrometry and animal science: protein identification strategies and particularities of farm animal species. J Proteomics, 75(14), 4190-4206. doi: 10.1016/j.jprot.2012.04.009

Song, L., & Florea, L. (2015). Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads. Gigascience, 4, 48. doi: 10.1186/s13742-015-0089-y

Specht, M., Stanke, M., Terashima, M., Naumann-Busch, B., Janssen, I., Hohner, R., . . . Hippler, M. (2011). Concerted action of the new Genomic Peptide Finder and AUGUSTUS allows for automated proteogenomic annotation of the Chlamydomonas reinhardtii genome. Proteomics, 11(9), 1814-1823. doi: 10.1002/pmic.201000621

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., & Morgenstern, B. (2006). AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res, 34(Web Server issue), W435-439. doi: 10.1093/nar/gkl200

Subbannayya, Y., Pinto, S. M., Gowda, H., & Prasad, T. S. (2016). Proteogenomics for understanding oncology: recent advances and future prospects. Expert Rev Proteomics, 13(3), 297-308. doi: 10.1586/14789450.2016.1136217

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., . . . Korbel, J. O. (2015). An integrated map of structural variation in 2,504 human genomes. Nature, 526(7571), 75-81. doi: 10.1038/nature15394

Sun, H., Xing, X., Li, J., Zhou, F., Chen, Y., He, Y., . . . Xie, L. (2013). Identification of gene fusions from human lung cancer mass spectrometry data. BMC Genomics, 14 Suppl 8, S5. doi: 10.1186/1471-2164-14-S8-S5

Szklarczyk, D., Santos, A., von Mering, C., Jensen, L. J., Bork, P., & Kuhn, M. (2016). STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data. Nucleic Acids Res, 44(D1), D380-384. doi: 10.1093/nar/gkv1277

The, M., MacCoss, M. J., Noble, W. S., & Kall, L. (2016). Fast and Accurate Protein False Discovery Rates on Large-Scale Proteomics Data Sets with Percolator 3.0. J Am Soc Mass Spectrom, 27(11), 1719-1727. doi: 10.1007/s13361-016-1460-7

Thompson, D. G., Wojtaszek, B. F., Staznik, B., Chartrand, D. T., & Stephenson, G. R. (2004). Chemical and biomonitoring to assess potential acute effects of Vision herbicide on native amphibian larvae in forest wetlands. Environ Toxicol Chem, 23(4), 843-849.

Tovchigrechko, A., Venepally, P., & Payne, S. H. (2014). PGP: parallel prokaryotic proteogenomics pipeline for MPI clusters, high-throughput batch clusters and multicore workstations. Bioinformatics, 30(10), 1469-1470. doi: 10.1093/bioinformatics/btu051

Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics, 25(9), 1105-1111. doi: 10.1093/bioinformatics/btp120

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., . . . Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol, 28(5), 511-515. doi: 10.1038/nbt.1621

Trapp, J., Almunia, C., Gaillard, J. C., Pible, O., Chaumot, A., Geffard, O., & Armengaud, J. (2015). Data for comparative proteomics of ovaries from five non-model, crustacean amphipods. Data Brief, 5, 1-6. doi: 10.1016/j.dib.2015.07.037

Trapp, J., Almunia, C., Gaillard, J. C., Pible, O., Chaumot, A., Geffard, O., & Armengaud, J. (2016). Proteogenomic insights into the core-proteome of female reproductive tissues from crustacean amphipods. J Proteomics, 135, 51-61. doi: 10.1016/j.jprot.2015.06.017

Trapp, J., Armengaud, J., Gaillard, J. C., Pible, O., Chaumot, A., & Geffard, O. (2016). High-throughput proteome dynamics for discovery of key proteins in sentinel species: Unsuspected vitellogenins diversity in the crustacean Gammarus fossarum. J Proteomics, 146, 207-214. doi: 10.1016/j.jprot.2016.07.007

Trapp, J., Armengaud, J., Salvador, A., Chaumot, A., & Geffard, O. (2014). Next-generation proteomics: toward customized biomarkers for environmental biomonitoring. Environ Sci Technol, 48(23), 13560-13572. doi: 10.1021/es501673s

Trapp, J., Geffard, O., Imbert, G., Gaillard, J. C., Davin, A. H., Chaumot, A., & Armengaud, J. (2014). Proteogenomics of Gammarus fossarum to document the reproductive system of amphipods. Mol Cell Proteomics, 13(12), 3612-3625. doi: 10.1074/mcp.M114.038851

Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M. Y., Geiger, T., . . . Cox, J. (2016). The Perseus computational platform for comprehensive analysis of (prote)omics data. Nat Methods, 13(9), 731-740. doi: 10.1038/nmeth.3901

Ubaida Mohien, C., Hartler, J., Breitwieser, F., Rix, U., Remsing Rix, L., Winter, G. E., . . . Colinge, J. (2010). MASPECTRAS 2: An integration and analysis platform for proteomic data. Proteomics, 10(14), 2719-2722. doi: 10.1002/pmic.201000075

Venter, E., Smith, R. D., & Payne, S. H. (2011). Proteogenomic analysis of bacteria and archaea: a 46 organism case study. PLoS One, 6(11), e27587. doi: 10.1371/journal.pone.0027587

Wang, X., Slebos, R. J., Chambers, M. C., Tabb, D. L., Liebler, D. C., & Zhang, B. (2016). proBAMsuite, a Bioinformatics Framework for Genome-Based Representation and Analysis of Proteomics Data. Mol Cell Proteomics, 15(3), 1164-1175. doi: 10.1074/mcp.M115.052860

Wang, X., Slebos, R. J., Wang, D., Halvey, P. J., Tabb, D. L., Liebler, D. C., & Zhang, B. (2012). Protein identification using customized protein sequence databases derived from RNA-Seq data. J Proteome Res, 11(2), 1009-1017. doi: 10.1021/pr200766z

Wang, X., & Zhang, B. (2013). customProDB: an R package to generate customized protein databases from RNA-Seq data for proteomics search. Bioinformatics, 29(24), 3235-3237. doi: 10.1093/bioinformatics/btt543

Warde-Farley, D., Donaldson, S. L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., . . . Morris, Q. (2010). The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic Acids Res, 38(Web Server issue), W214-220. doi: 10.1093/nar/gkq537

Wilmes, P., Andersson, A. F., Lefsrud, M. G., Wexler, M., Shah, M., Zhang, B., . . . Banfield, J. F. (2008). Community proteogenomics highlights microbial strain-variant protein expression within activated sludge performing enhanced biological phosphorus removal. ISME J, 2(8), 853-864. doi: 10.1038/ismej.2008.38

Woo, S., Cha, S. W., Merrihew, G., He, Y., Castellana, N., Guest, C., . . . Bafna, V. (2014). Proteogenomic database construction driven from large scale RNA-seq data. J Proteome Res, 13(1), 21-28. doi: 10.1021/pr400294c

Xuereb, B., Chaumot, A., Mons, R., Garric, J., & Geffard, O. (2009). Acetylcholinesterase activity in Gammarus fossarum (Crustacea Amphipoda) Intrinsic variability, reference levels, and a reliable tool for field surveys. Aquat Toxicol, 93(4), 225-233. doi: 10.1016/j.aquatox.2009.05.006

Yates, J. R., 3rd, Eng, J. K., & McCormack, A. L. (1995). Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. Anal Chem, 67(18), 3202-3210.

Zeng, Z. B., Kao, C. H., & Basten, C. J. (1999). Estimating the genetic architecture of quantitative traits. Genet Res, 74(3), 279-289.

Zerbino, D. R., & Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res, 18(5), 821-829. doi: 10.1101/gr.074492.107

Zhang, X., Xia, P., Wang, P., Yang, J., & Baird, D. J. (2018). Omics Advances in Ecotoxicology. Environ Sci Technol, 52(7), 3842-3851. doi: 10.1021/acs.est.7b06494

Zhu, Y., Hultin-Rosenberg, L., Forshed, J., Branca, R. M., Orre, L. M., & Lehtio, J. (2014). SpliceVista, a tool for splice variant identification and visualization in shotgun proteomics data. Mol Cell Proteomics, 13(6), 1552-1562. doi: 10.1074/mcp.M113.031203