N°d'ordre NNT : 2022LYO1006

THESE de DOCTORAT DE L'UNIVERSITE DE LYON
opérée au sein de
l'Université Claude Bernard Lyon 1

Ecole Doctorale N° 341
Evolution Ecosystèmes Microbiologie Modélisation (E2M2)

Spécialité de doctorat : Biomathématiques, bioinformatique, génomique évolutive
Discipline : Bioinformatique et écotoxicologie

Soutenue publiquement le 07/09/2022 par :
Natacha Koenig

Approches bioinformatiques pour l'exploration sans *a priori* des voies moléculaires chez les espèces non modèles *:* le cas du métabolisme lipidique chez *Gammarus fossarum*

Devant le jury composé de :

VIEIRA Cristina	Professeure des Universités, Université Lyon 1	Présidente
COUTELLEC Marie-Agnès	Chargée de Recherche, INRAE Rennes	Rapporteure
GONZALEZ Patrice	Chargé de Recherche, CNRS Bordeaux	Rapporteur
BROCHIER-ARMANET Céline	Professeure des Universités, Université Lyon 1	Examinatrice
PRUD'HOMME Sophie	Maître de Conférences, Université de Lorraine	Examinatrice
GEFFARD Olivier	Directeur de recherche, INRAE Villeurbanne	Directeur de thèse
CALEVRO Federica	Directrice de Recherche, INRAE Villeurbanne	Invitée
DEGLI ESPOSTI Davide	Charge de recherche, INRAE Villeurbanne	Co-directeur de thèse invité

RESUME

L'utilisation des espèces modèles en science environnementale se confronte à plusieurs verrous, tels que leur distribution géographique et leur représentativité vis à vis de la diversité des espèces dans les milieux aquatiques. Les connaissances liées aux organismes modèles limitent l'extrapolation et la prédiction des réponses pour les organismes des milieux. Les avancées des nouvelles technologies de séquençage et de spectrométrie de masse (MS) permettent d'étendre l'acquisition de données moléculaires aux organismes de pertinence environnementale, comme l'amphipode *Gammarus fossarum*.

L'objectif de cette thèse consiste au **développement d'approches bioinformatiques sans *a priori* et multiomique pour exploiter les données omiques de plus en plus disponibles chez les espèces non modèles et contourner les limitations de l'annotation fonctionnelle à partir des bases de données existantes.**

Premièrement, une analyse de réseaux de coexpression a été réalisée à l'aide de la méthode WGCNA, dans le but d'identifier les acteurs moléculaires (*i.e.* protéines) régulant la physiologie de la reproduction et la réponse des testicules aux contaminants. Pour ceci, deux jeux de données précédemment publiés ont été étudiés. Le premier a combiné les profils protéomiques des gonades mâles et femelles à différents stades de maturation et ceux des embryons au cours de leur développement. Le second provenait d'une précédente étude et était composé de données protéomiques issues de testicules de gammares exposés à trois contaminants : le pyriproxyfène (Pyr), le cadmium (Cd) et le méthoxyfénozide (Met). Ces travaux ont permis d'identifier des modules de protéines coexprimées et spécifiques des différents stades de développement et de reproduction, et de l'exposition aux différents contaminants. Les analyses d'enrichissement de ces modules ont révélé de nouvelles hypothèses pour comprendre les processus biologiques sous-jacents. Notamment, un module lié à l'embryogenèse a montré l'importance des voies

moléculaires comme l'épissage de l'ARN et a confirmé la diversité des protéines de la famille Large Lipid Transport Proteins impliquées dans la vitellogenèse. Une régulation fine entre la production d'énergie et les évènements dépendants de l'actine-myosine dans la spermatogenèse a été mise en évidence. L'enrichissement des modules liés à l'exposition aux contaminants a identifié des protéines : (i) impliquées dans l'organisation du cytosquelette et la réponse au stress oxydatif, en lien avec le Cd, (ii) associées au Pyr et en lien avec la réponse au stress du réticulum endoplasmique, et (iii) associées au Met mais restreintes aux amphipodes dont les fonctions ne sont pas encore annotées. Ce travail a permis d'identifier des protéines clés que l'analyse différentielle précédemment utilisée n'avait pas permis de révéler.

Deuxièmement, une stratégie multiomique a été développée pour décrire les voies métaboliques impliquées dans le métabolisme lipidique chez *G. fossarum*. L'outil d'annotation génomique (CycADS), qui permet de reconstruire des voies métaboliques, a été adapté pour exploiter les données transcriptomiques disponibles chez le mâle et la femelle, en réduisant la redondance des isoformes des transcrits. La reconstruction des voies métaboliques a permis d'identifier plus de 70 voies impliquées dans le métabolisme lipidique (ML) pour le mâle et la femelle. Les données de MS ont permis de valider la détection d'une centaine d'enzymes catalysant les réactions constituant les voies identifiées dans les différents organes.

Ce travail de thèse constitue une base solide pour l'utilisation des données omiques d'organismes non modèles dans une stratégie d'exploration sans *a priori* ou multiomique. Les résultats ont mis en évidence des protéines potentiellement impliquées dans les processus biologiques liés aux stades de développement, à la reproduction et au ML chez *G. fossarum*, ainsi que dans la toxicité testiculaire. Ces résultats soulignent également l'intérêt de l'application d'approches omiques au niveau des organes pour identifier les possibles différences de MoA de contaminants en fonction de l'organe cible.

ABSTRACT

The use of model species in environmental science is confronted with several scientific limitations, such as their geographical distribution and their representativeness of the diversity of species in aquatic environments. Indeed, the knowledge related to model organisms does not allow the extrapolation and the prediction of the response(s) of organisms present in the environment. Advances in new sequencing and mass spectrometry (MS) technologies allow the extension of large-scale molecular data acquisition to organisms of environmental relevance, such as the amphipod *Gammarus fossarum*.

The overall objective of this thesis is **to develop unbiased and multiomics bioinformatics approaches to exploit the increasingly available omics data in non-model species and to overcome the limitations of functional annotation from existing databases**.

First, a coexpression network analysis was performed using the WGCNA method, with the aim of identifying the molecular actors (*i.e.* proteins) regulating reproductive physiology and testicular response to contaminants. For this, two datasets were studied. The first one combined the proteomic profiles of male and female gonads at different stages of maturation and those of embryos during their development. The second was from a previous study that consisted of proteomic data from gammarid testes exposed to three contaminants: pyriproxyfen (Pyr), cadmium (Cd) and methoxyfenozide (Met). This work allowed the identification of co-expressed protein modules specific to different developmental and reproductive stages, and to exposure to different contaminants. Enrichment analyses of these modules supported new hypotheses for understanding the underlying biological processes. A module related to embryogenesis showed the importance of molecular pathways such as RNA splicing and confirmed the diversity of the Large Lipid Transport Protein family. A fine regulation between energy production and actin-myosin dependent events in spermatogenesis was demonstrated. The enrichment of modules

related to contaminant exposure identified proteins: (i) involved in cytoskeletal organization and oxidative stress response, related to Cd, (ii) associated with Pyr and related to the endoplasmic reticulum stress response, and (iii) related to Met but restricted to amphipods whose functions are not yet annotated. This work identified key proteins that were not revealed by the previously used differential analysis.

Second, a multiomics strategy was developed to describe the metabolic pathways involved in ML in *G. fossarum*. The genomic annotation tool (CycADS), which allows the reconstruction of metabolic pathways, was adapted to exploit the available transcriptomic data in both males and females, reducing the redundancy of transcript isoforms. Metabolic pathway reconstruction identified over 70 pathways involved in ML in both males and females. The MS data validated the detection of about 100 enzymes catalyzing the reactions constituting the identified pathways in the different organs.

This thesis work provides a solid foundation for the use of omics data from non-model organisms in without *a priori* or multiomics exploration strategy. The results identified proteins potentially involved in biological processes related to developmental stages, reproduction and ML in *G. fossarum*, as well as testicular toxicity. These results also highlight the value of applying omics approaches at the organ level to identify possible differences in contaminant MoA depending on the target organ.

REMERCIEMENTS

Je remercie toutes les personnes qui ont participé à l'élaboration de ma thèse et en premier lieu mon directeur de thèse Davide Degli Esposti, pour son implication, son soutien sans faille et sa grande disponibilité tout au long de ces années. Je tiens aussi à remercier Olivier Geffard pour sa participation dans la direction de ma thèse et ses nombreux conseils durant la rédaction de ce manuscrit. Je vous suis reconnaissante de m'avoir fait confiance tout en ayant pris le temps de m'aiguiller au mieux. Merci pour tous vos encouragements et votre bienveillance. J'ai beaucoup appris à vos côtés et travailler avec vous a été un réel plaisir.

Je voudrais ensuite remercier tous les membres du jury d'avoir accepté d'évaluer ce travail et particulièrement Marie-Agnès Coutellec et Patrice Gonzalez, rapporteurs de cette thèse. Je remercie Céline Brochier-Armanet et Sophie Prud'homme pour avoir été mes examinatrices, ainsi que Cristina Vieira pour avoir présidé ce jury de thèse.

Je remercie INRAE de m'avoir financé ainsi que l'ANR grâce aux projets « JCJC PLAN-TOX », « PROTEOGAM », et « APPROve ». Je tiens à remercier Jean Armengaud et Christine Almunia du laboratoire LI2D au CEA de Marcoule, et Sophie Ayciriex de l'ISA de Villeurbanne pour leur participation scientifique ainsi que le temps consacré à ma recherche. Je souhaite aussi remercier Federica Calevro et Hubert Charles de l'UMR Biologie Fonctionnelle, insectes et interactions (BF2i) pour leur accueil et notamment Patrice Baa-Puyoulet pour sa disponibilité et sa pédagogie pour l'enseignement des rouages de CycADS.

Je tiens à remercier tous les membres du laboratoire d'écotoxicologie de INRAE (et ceux qui ont été de passage) et les camarades doctorants pour m'avoir accueilli chaleureusement et m'avoir permis de passer un séjour mémorable au sein de l'équipe, que ce soit grâce à l'ambiance, aux apéros du vendredi, aux pauses au coin café ou aux fameux repas de Noël.

Je remercie enfin tout particulièrement mes ami(e)s qui ne m'ont pas beaucoup vu ces derniers mois. Vos attentions, vos encouragements et tous les bons moments passés ensemble m'ont accompagnée tout au long de ces années, et m'ont permis de garder le cap.

Enfin à mon mari, Nicolas, qui a été à mes côtés au quotidien dans les bons et les mauvais moments et surtout pendant cette phase de rédaction qui a été un vrai challenge. Alors, merci pour ton soutien infaillible, d'avoir su m'épauler et d'avoir pris soin de notre fils quand j'ai été moins présente. Vous avoir tous les deux à mes côtés a été une source de motivation considérable.

Pour finir je remercie ma famille, mon frère Dimitri et ma sœur Mélissa pour m'avoir écouté me plaindre et pour votre soutien inconditionnel. Je mesure la chance que j'ai de vous avoir à mes côtés. Je souhaite enfin dédier cette thèse à mes parents et particulièrement à mon père disparu juste avant que je ne commence ce projet. Vous avez toujours été un exemple de force et détermination. Merci de m'avoir inculqué ces valeurs de résilience et de courage. J'espère vous rendre fiers et être à la hauteur de vos nombreux sacrifices pour me permettre d'être celle que je suis aujourd'hui.

TABLE DES MATIERES

LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DES ABREVIATIONS

ACP	Analyse en Composante Principale
ADNc	ADN complémentaire
AG	Acide Gras
BLAST	Basic Local Alignment Search Tool
BUSCO	Benchmarking Universal Single-Copy Orthologues
CBZ	Carbamazépine
CDS	Coding DNA Sequence
CycADS	Cyc Annotation Database System
DEG	Differentially Expressed Genes
DBG	De Bruijn Graph
EC	Enzyme Commission
ERA	Environmental Risk Assessment
EST	Expressed Sequence Tag
FC	Fold Change
FDR	False Discovery Rate
GFF3	General Feature Format 3
G. fossarum	*Gammarus fossarum*
GO	Gene Ontology
HAP	Hydrocarbures Aromatiques Polycycliques
HMG-CoA	Hydroxyméthylglutaryl – CoA
IsoPct	Isoforme Percentage
JH	Juvenile Hormone
KAAS	KEGG Automatic Annotation Server
KEGG	Kyoto Encyclopedia of Genes and Genomes
KNN	K Nearest Neighbours
KO	KEGG Orthology
LC	Liquid Chromatography
LLTP	Large Lipid Transfer Protein
ME	Module Eigengene

MDS	Multi Dimensional Scaling
ML	Métabolisme Lipidique
MoA	Mécanisme d'Action
MRM	Multiple Reactions Monitoring
MS	Mass Spectrometry ou Spectrométrie de Masse
MSEA	Metabolite Set Enrichment Analysis
MSI	Imagerie par Spectrométrie de Masse
NGS	Next Generation Sequencing ou Séquençage de Nouvelle Génération
ORF	Open Reading Frame
PE	Perturbateurs Endocriniens
PGDB	Pathway/Genome Database
PUFA	Polyunsaturated Fatty Acid
RE	Réticulum Endoplasmique
RMN	Résonanace Magnétique Nucléaire
RNA-Seq	Séquençage ARN
RSEM	RNA-Seq by Expectation Maximisation
RT	Temps de Rétention
SC	Spectral Count
TAG	Triacylglycérides
TMM	Trimmed Mean of M-values
WGCNA	Weighted Gene Coexpression Network Analysis

LISTE DES VALORISATIONS ECRITES ET ORALES

ORALES

Koenig N., Almunia C., Armengaud J., Chaumot A., Geffard O., Degli Esposti D. **Co-expression network analysis identifies distinct protein modules and pathways associated with cadmium and pyriproxyfen testicular toxicity in *Gammarus fossarum***. Society of Environmental Toxicology and Chemistry (SETAC) Europe Annual Meeeting, 3-6 mai 2021 (en visioconférence).

Koenig N., Almunia C., Chaumot A., Armengaud J., Geffard O., Degli Esposti D. **Réseaux de coexpression pour l'analyse de données de protéomique pour la compréhension des mécanismes d'action de contaminants chez une espèce non modèle, *Gammarus fossarum***. Colloque Société d'Écotoxicologie Fondamentale et Appliquée (SEFA), Lyon (France), 24-25 juin 2019.

POSTERS

Koenig N., Baa-Puyoulet P., Leprêtre M., Gaillard J-C., Armengaud J., Charles H., Calevro F., Chaumot A., Geffard O., Degli Esposti D. **Intégration de données multi-omiques pour l'identification d'enzymes clés du métabolisme lipidique chez l'espèce sentinelle *Gammarus fossarum*.** GDR Ecotoxicologie, Rennes, 23-24 novembre 2021.

Koenig N., Almunia C., Chaumot A., Armengaud J., Geffard O., Degli Esposti D. **Réseaux de co-expression pour l'analyse de données de protéomique pour la compréhension des mécanismes d'action de contaminants chez une espèce non-modèle, Gammarus fossarum**. Journées Ouvertes en Biologie, Informatique et Mathématiques (JOBIM), Nantes, 2-5 juillet 2019.

ARTICLES

Koenig, N., Almunia, C., Bonnal-Conduzorgues, A., Armengaud, J., Chaumot, A., Geffard, O., Degli Esposti, D., 2021. **Co-expression network analysis identifies novel molecular pathways associated with cadmium and pyriproxyfen testicular toxicity in Gammarus fossarum**. Aquatic Toxicology 235, 105816. https://doi.org/10.1016/j.aquatox.2021.105816

Degli Esposti, D., Almunia, C., Guery, M.-A., Koenig, N., Armengaud, J., Chaumot, A., Geffard, O., 2019. **Co-expression network analysis identifies gonad- and embryo-associated protein modules in the sentinel species Gammarus fossarum.** Scientific Reports 9, 7862. https://doi.org/10.1038/s41598-019-44203-5

AVANT-PROPOS

Ce projet de thèse s'articule à l'interface de l'écotoxicologie aquatique, la biologie moléculaire à haut contenu d'information (les omiques, notamment transcriptomique, protéomique et lipidomique), la bioinformatique et les biostatistiques. Ce projet a été initié grâce à la collaboration de trois laboratoires : le laboratoire d'écotoxicologie d'INRAE Villeurbanne, le groupe ANABIO-MS de l'Institut des Sciences Analytiques (ISA) à Villeurbanne et le Laboratoire Innovations Technologiques pour la Détection et le Diagnostic (LI2D) au CEA de Marcoule. Les travaux ont été en partie financés par le projet ANR JCJC PLAN-TOX (ANR-18-CE34-0008), le projet ANR Proteogam (ANR-14-CE21-0006) et le projet ANR APPROve (ANR-18-CE34-0013). Au cours de cette thèse une nouvelle collaboration a été mise en place avec l'équipe SYMT (génomique fonctionnelle des interactions trophiques dans la symbiose) de l'unité mixte de recherche (INSA-INRAE) Biologie Fonctionnelle, insectes et interactions (BF2i).

Les avancées des nouvelles technologies de séquençage et de spectrométrie de masse permettent d'étendre l'acquisition des données moléculaires aux organismes étudiés en écotoxicologie comme sentinelles de la qualité environnementale, telles que l'amphipode *Gammarus fossarum*. Avec la croissante acquisition de données omiques chez ces espèces, il est aujourd'hui nécessaire de développer des outils et approches bioinformatiques qui permettent d'exploiter au mieux les données disponibles. Ces outils permettraient dans le cadre de l'écotoxicologie d'améliorer la prédiction des effets toxiques des contaminants sur les communautés aquatiques en acquérant des connaissances sur les voies métaboliques impliquées dans les processus clés pour le maintien des populations. C'est dans ce cadre, que ce projet s'intéresse à la **caractérisation des voies métaboliques impliquées dans la biosynthèse, le transport et la dégradation des lipides via des approches bioinformatiques sans *a priori* et multiomiques chez le crustacé *G. fossarum*.**

Ce manuscrit est organisé en 4 chapitres. Le chapitre I fait un état des lieux des connaissances sur (i) les approches omiques, (ii) les approches bioinformatiques pour la compréhension des mécanismes d'action et des voies moléculaires en écotoxicologie, (iii) le métabolisme lipidique et (iv) l'espèce non modèle *G. fossarum.* Le chapitre II présente les résultats des travaux réalisés à partir d'une analyse de réseaux de coexpression sur des données d'études précédentes, dans le cadre de l'étude des stades de reproduction et de développement des embryons de gammare. Ce chapitre s'appuie sur deux articles publiés, le premier dans Scientific Reports (Degli Esposti et al., 2019) et le deuxième dans Aquatic Toxicology (Koenig et al., 2021). Le chapitre III présente les résultats de travaux d'une approche multiomique (transcriptomique et protéomique) pour la caractérisation du métabolisme lipidique, et s'appuie sur un article en préparation. Le chapitre IV présente une discussion de l'ensemble des résultats obtenus ainsi que les perspectives envisagées.

CHAPITRE I - ÉTAT DE L'ART

1. LES APPROCHES OMIQUES

1.1 Historique et définitions

Les avancées des technologies expérimentales à haut débit et bioinformatiques ont permis le développement des approches dites omiques en biologie moléculaire depuis les années 1990, permettant l'exploration des systèmes biologiques de façon exhaustive et sans *a priori* (Fischer, 2005). L'apparition de la technologie de séquençage Sanger (dite aujourd'hui de première génération) dans les années 70, a révolutionné la biologie en permettant de déchiffrer les génomes (Sanger et al., 1977) (Figure 1). Les méthodes de « séquençage de nouvelle génération » dites NGS sont ensuite apparues dans les années 2000, tel que le séquençage par amplification en pont (implémenté sur les séquenceurs de type Illumina) (Heather and Chain, 2016) (Figure 1). Ces NGS sont plus puissantes (*e.g.* pour la méthode Illumina, un débit de 1300 Mb par run, une vitesse de 4 jours par run, un coût d'environ 9000 USD par run (Mardis, 2008)) que le séquençage Sanger et permettent à de nombreux laboratoires d'accéder au séquençage. À ce jour, les techniques de séquençage de troisième génération ou dit « long read » ont été mises au point et permettent une amélioration de la qualité des génomes (Figure 1).

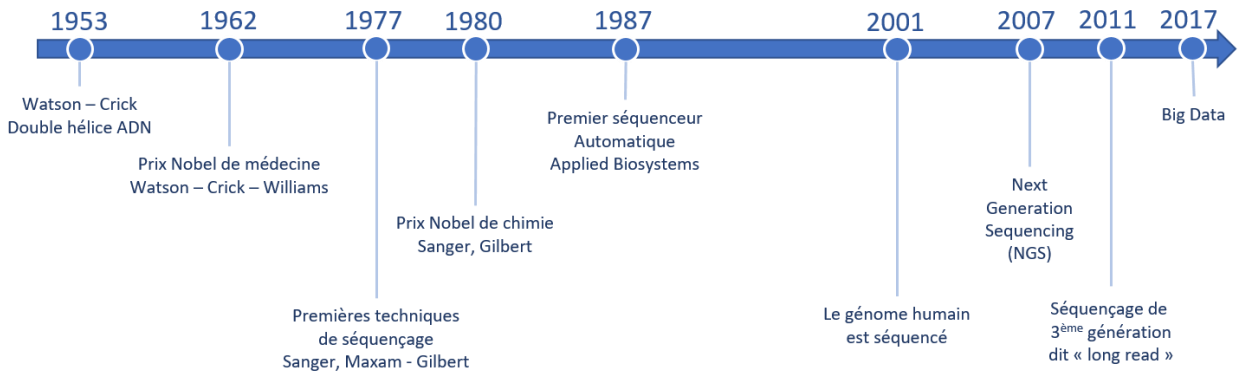


Figure 1 : Chronologie de l'évolution du séquençage des acides nucléiques (inspirée de (Fournier, 2022)).

Les approches omiques permettent des analyses globales à plusieurs niveaux moléculaires : le génome (ADN), le transcriptome (ARN), le protéome (protéines), le métabolome (métabolites cellulaires), le lipidome (lipides). Ces notions seront détaillées ci-après.

1.2 La génomique

Le génome correspond à l'ensemble du matériel génétique codé dans l'ADN. Il est composé de séquences codantes et non codantes. Les séquences codantes sont transcrites en ARN messagers (ARNm) puis traduites en protéines. Les séquences non codantes peuvent ne pas être transcrites, ou transcrites en ARN mais non traduites par la suite (ARN ribosomique (ARNr), ARN de transfert (ARNt), long ARN non codant et micro ARNs) (Figure 2). Une partie de l'ADN non codant est composée de séquences régulatrices (*i.e.* promoteurs, amplificateurs dits « enhancers ») qui jouent un rôle sur la transcription des gènes (*i.e.* l'expression des gènes) (Figure 2).

La génomique structurale vise à déterminer la structure physique de l'ADN jusqu'aux protéines en identifiant les séquences informatives (*i.e.* gènes avec ou sans introns, les éléments transposables, séquences régulatrices, séquences répétées, etc.) (Figure 2). Dans la suite de ce chapitre et dans ce manuscrit, nous nous intéresserons principalement aux gènes eucaryotes codants pour des protéines.

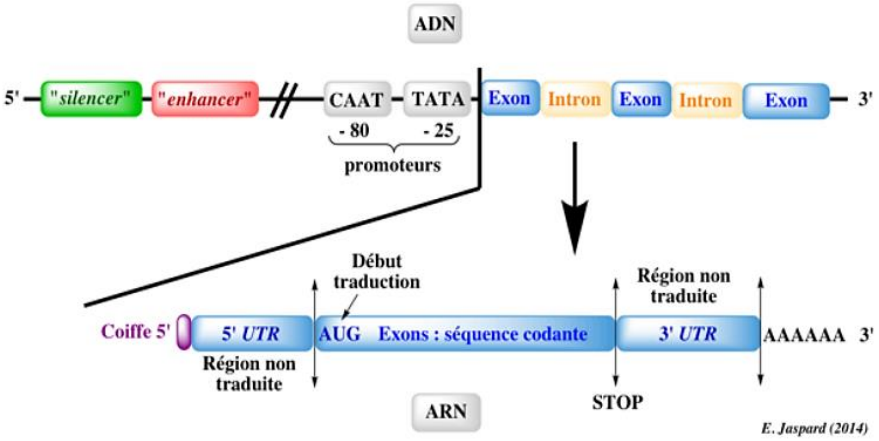


Figure 2 : Structure des séquences génomiques : de l'ADN à l'ARN (E. Jaspard 2014).

Ainsi, pour permettre l'analyse des données issues du séquençage d'ADN, des stratégies d'annotation du génome ont été développées. L'annotation du génome consiste à analyser la séquence nucléotidique en ayant 3 objectifs majeurs (Figure 3) :

(i) **Localiser les gènes et les régions codantes correspondantes** : **la génomique structurale**. À partir du code génétique (*i.e.* la syntaxe du génome) il est possible de déduire les séquences des ARN et des protéines, et de prédire la structure tridimensionnelle des protéines (Jumper et al., 2021). Il est aussi éventuellement possible de prédire et d'inférer les fonctions des protéines par des approches de phylogénie par exemple (Sahraeian et al., 2015), mais cette approche ne permet pas d'évaluer leur taux d'expression, les modifications de transcription ou encore les modifications post-traductionnelles des protéines.

(ii) **Identifier ou prédire leur fonction biologique** : **la génomique fonctionnelle** permet d'étudier la manière dont les gènes et les régions intergéniques du génome participent aux différents processus biologiques et produisent un phénotype particulier. La génomique fonctionnelle se concentre sur l'expression dynamique des produits des gènes dans un contexte spécifique (*e.g.* stade de développement donné, exposition à un contaminant). Les connaissances en génomique fonctionnelle tentent de relier le génotype au phénotype.

(iii) **Comprendre les interactions entre ces entités biologiques** : **l'annotation relationnelle** consiste à tenter d'identifier les interactions (ou relations) entre les gènes et leurs produits impliqués dans les voies métaboliques (Mercier, 2017).

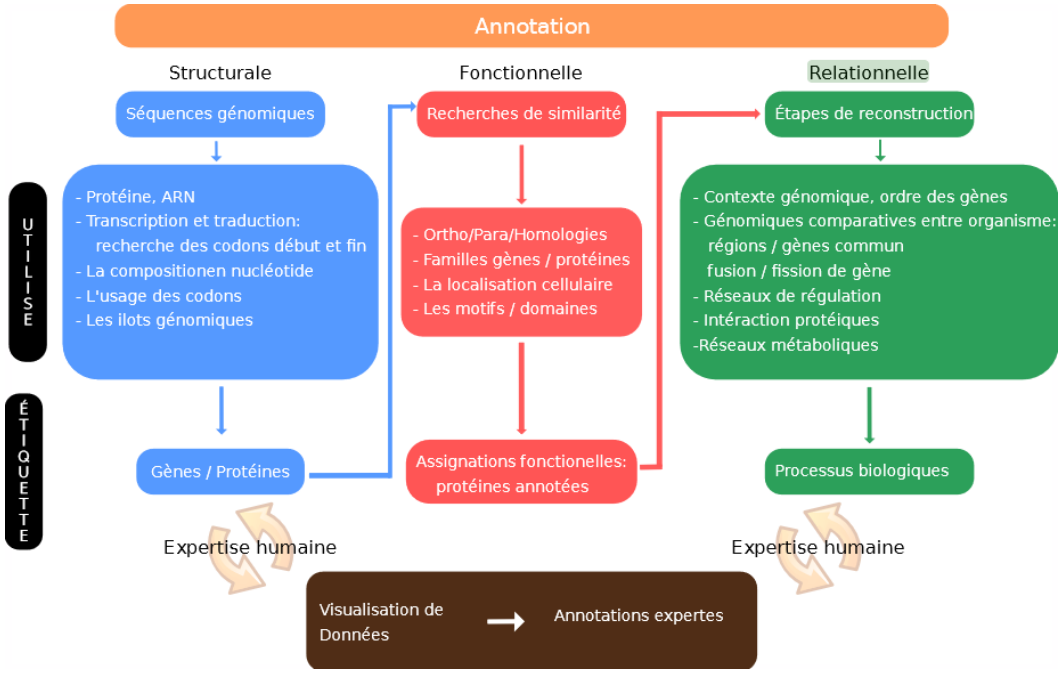


Figure 3 : Schéma des trois niveaux d'annotations des génomes : L'annotation structurale, l'annotation fonctionnelle et l'annotation relationnelle (Mercier, 2017).

Pour accéder à l'information contenue dans les gènes codants pour les protéines, la méthode du séquençage shotgun de l'ADN a été développée. Pour cela, l'ADN est fragmenté et ces fragments sont lus par le séquenceur en petits fragments aléatoires, appelés lectures ou reads. Ces fragments peuvent être en lectures simples (single reads) ou en lectures couplées (paired-end) sur le brin sens et anti-sens de l'ADN (Figure 4). Les lectures doivent ensuite être assemblées pour retrouver la séquence et la structure du génome. L'assemblage du génome consiste à aligner (si un génome de référence est disponible) ou fusionner (si aucun génome de référence n'existe) les lectures pour former des séquences plus longues pour reconstituer la séquence du fragment d'ADN original (Figure 4).

Toutefois, avant de procéder à l'assemblage des lectures en contigs, il est nécessaire d'évaluer la qualité des données. Cela est possible grâce à plusieurs outils existants. L'outil FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) (Andrews, 2010) permet de générer des statistiques relatives à des critères (*i.e.* qualité des lectures, teneur globale en bases G et C,

contamination par les adaptateurs ou encore la proportion de lectures dupliquées) pour évaluer la qualité des données. L'outil Trimmomatic (Bolger et al., 2014) permet de supprimer les lectures ayant une accumulation d'erreurs aux extrémités générées par le séquençage Illumina (Fuller et al., 2009). L'outil cutadapt (Martin, 2011) ou le programme Trimmomatic (Bolger et al., 2014) suppriment les lectures présentant des séquences d'amorces non éliminées. Des outils comme BWA (Li and Durbin, 2009) ou Kraken 2 (Wood et al., 2019) identifient et suppriment les séquences potentielles de contaminants provenant d'autres organismes entrainants des contigs chimériques ou des erreurs d'assemblages.

Après avoir évalué leur qualité, les lectures doivent être assemblées par des algorithmes informatiques pour former des fragments contigus, appelés contigs (Figure 4). Pour que l'assemblage soit le plus correct possible, il est nécessaire d'avoir un chevauchement suffisant entre les lectures, nécessitant une couverture de séquence élevée (*i.e.* profondeur de lecture) (Ekblom and Wolf, 2014). Néanmoins, certains contigs peuvent ne pas se chevaucher, laissant ainsi un gap qui est représenté par des « N » et correspond à n'importe lequel des 4 nucléotides possibles (A, T, G ou C).

Ces contigs peuvent ensuite être assemblés en scaffolds (parfois appelés supercontigs) qui correspondent à l'ensemble des contigs orientés et ordonnés. Il est possible que des gaps de longueur connue soient encore présents et qu'ils soient représentés par des « N » dans la séquence (Ellegren, 2014) (Figure 4). À ce stade-là, le génome reste à l'état d'ébauche (dit draft) et n'est pas dans sa forme finale, c'est-à-dire mappée (ou alignée) sur les chromosomes.

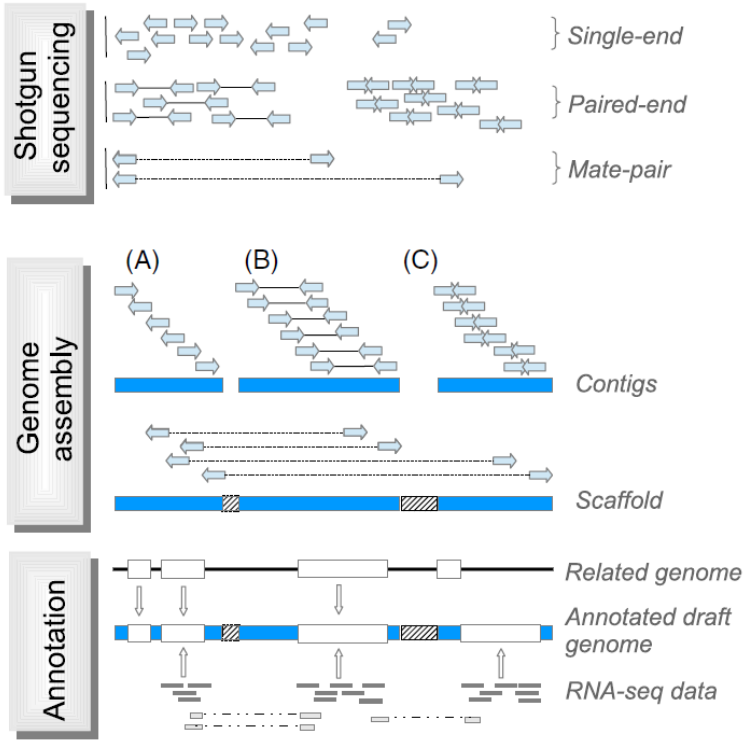


Figure 4 : Schéma simplifié et termes du séquençage, de l'assemblage et de l'annotation du génome (Ekblom and Wolf, 2014).

À ce jour, la grande majorité des génomes sont publiés à l'état d'ébauche (Lewin et al., 2022, 2018). Les scaffolds (même assemblés *de novo*) sont souvent assez longs pour contenir et permettre la découverte de gènes. Le mapping des scaffolds sur les chromosomes nécessite un effort important de génotypage (Romanov et al., 2009). Pour la plupart des organismes vivants ces cartes génétiques n'existent pas, et cela explique que le nombre de génomes complets reste limité et ne représente que 1% des génomes dans les bases de données (Figure 5). La majorité des assemblages de génome se trouve sous la forme de scaffolds (58%) et contigs (27%) (Figure 5). La plupart des analyses post-assemblage et les applications (*e.g.* annotation, étude comparative, phylogénétique) ne nécessitent pas forcément le génome complet, mais peuvent être menées à partir de contigs ou scaffolds de bonne qualité.

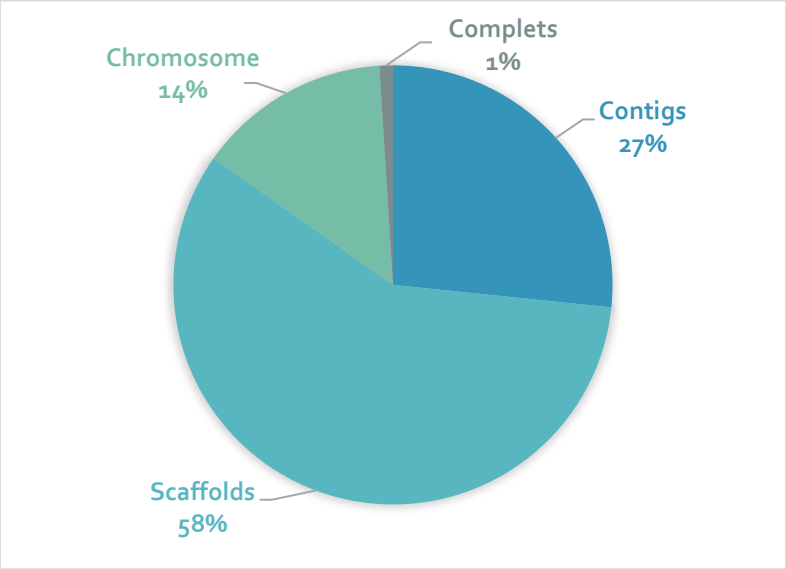


Figure 5 : Statut de l'assemblage des génomes, basé sur tous les génomes eucaryotes présents (21899) classés dans la base de données NCBI (NCBI, 2022), (inspirée de Bradman, 2015).

Dans le cas de séquençage d'ADN chez des espèces n'ayant pas de génome de référence, les données génomiques doivent être assemblées de manière *de novo* (*i.e.* absence d'un autre génome de référence comme guide). Pour cela, il existe plusieurs outils qui se différencient par leur performance (*i.e.* vitesse, qualité et précision de la séquence du génome, évolutivité) (Bradnam et al., 2013; Di Genova et al., 2021; Earl et al., 2011; Miller et al., 2010; Narzisi and Mishra, 2011). La plupart des logiciels sont créés pour les données long reads comme le séquençage Sanger, mais sont aujourd'hui utilisés et adaptés pour les données de lectures courtes. Il existe également des algorithmes d'assemblages hybrides qui permettent de combiner des données de lectures courtes et longues (Di Genova et al., 2021). On peut différencier ces différents assembleurs en trois catégories selon les paradigmes d'assemblages utilisés (Ekblom and Wolf, 2014; Miller et al., 2010; Nagarajan and Pop, 2013) (Figure 6) :

( ) Ceux qui utilisent les algorithmes de type « **overlap-layout-consensus** » (**OLC**). L'assembleur commence par identifier toutes les paires de lectures suffisamment chevauchantes et les organise ensuite en un graphe. Ces algorithmes sont souvent considérés comme très gourmands en termes de

temps de calcul pour les données Illumina, mais certains assembleurs utilisent tout de même ces algorithmes pour les données de courtes lectures, tels que Edena (Hernandez et al., 2008), SGA (Simpson and Durbin, 2012) ou encore FERMI (Li, 2012).

(i) Ceux qui utilisent les algorithmes de type **graphes de Bruijn** (**DBG**). Ces assembleurs divisent les lectures brutes en sous-chaînes de caractères de longueur k (k-mer), et modélisent la relation entre ces sous-chaînes en un graphe. Ce paradigme utilisant des correspondances exactes, les erreurs de séquençage peuvent dégrader la qualité de l'assemblage et être plus difficile à utiliser avec des lectures de plus en plus longues et imprécises. Les programmes comme Euler (Pevzner et al., 2001), Velvet (Zerbino and Birney, 2008), AllPaths (Butler et al., 2008), ABySS (Simpson et al., 2009) ou SOAPdenovo (Luo et al., 2012) utilisent des DBG.

(ii) Ceux qui utilisent les deux types d'algorithmes (OLC-DBG) dits « **greedy** » ou « **glouton** ». Ces assembleurs font un compromis entre le meilleur chevauchement des lectures et l'assemblage qui est déjà construit. Les choix effectués par l'assembleur sont intrinsèquement locaux et ne prennent pas en compte la relation globale entre les lectures. Ce paradigme n'est pas très répandu, étant donné le manque d'utilisation de la globalité de l'information, mais on retrouve par exemple l'outil VCAKE (Jeck et al., 2007).

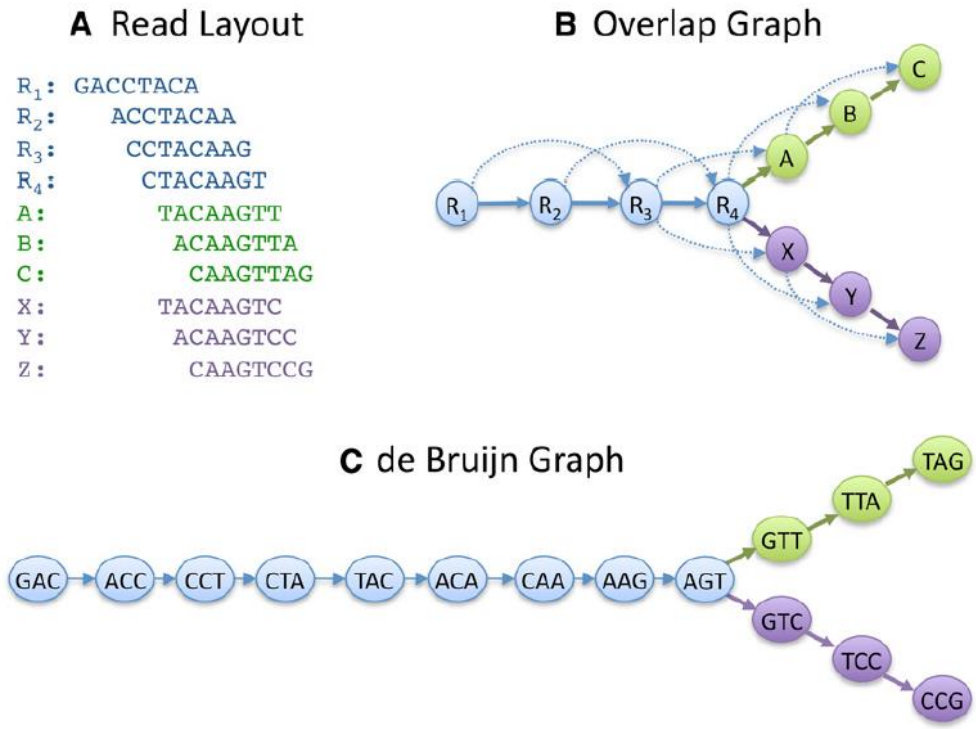


Figure 6 : Différences entre un graphe overlap-layout-consensus (OLC) et un graphe de Bruijn (DBG) pour l'assemblage *de novo* des génomes. (A) 10 lectures de 8 paires de bases (bp), (B) un nœud = une lecture, une arête = un overlap de plus de 5 bp, (C) un noeud = un 3-mer pour chaque k-mer des lectures, une arête = overlap de k-1 bases (Schatz et al., 2010).

Il est difficile de savoir à l'avance quel outil sera le plus pertinent selon le projet d'assemblage, la structure des données, la composition en bases, les séquences répétées, etc. (Ekblom and Wolf, 2014). Certains assembleurs favorisent la minimisation des défauts d'assemblage, alors que d'autres préfèrent améliorer la contiguïté du génome (Ekblom and Wolf, 2014). Plusieurs études de benchmark existent pour comparer les différents assembleurs (Miller et al., 2010; Nagarajan and Pop, 2013) à partir de jeux de données dits « gold-standards » (*i.e.* de très haute qualité comme des génomes de référence complets ou presque complets) (Haiminen et al., 2011; Lin et al., 2011; Salzberg et al., 2012; Zhang et al., 2011) ou des données simulées (Barthelson et al., 2011; Earl et al., 2011; Haiminen et al., 2011; Lin et al., 2011). De manière générale, d'après ces comparaisons d'outils, il est conseillé de tester plusieurs méthodes d'assemblages et d'évaluer la qualité qui convient le mieux aux données disponibles (Ekblom and Wolf, 2014).

Après l'assemblage en contigs, l'étape de scaffolding est généralement incluse dans la plupart des programmes. Il existe des programmes qui permettent de faire cette étape séparément pour plus de contrôle et pour pouvoir l'adapter à des génomes complexes, tels que SSPACE (Boetzer et al., 2011) ou BESST (Sahlin et al., 2014). Les gaps peuvent être supprimés après le scaffolding en se basant sur les données des paires de lectures. Pour ce faire, on retrouve les logiciels iMAGE (Tsai et al., 2010), GapCloser (Li et al., 2010) ou GapFiller (Boetzer and Pirovano, 2012). Si l'on a à disposition des données de séquençage de lectures longues, ces dernières permettent également de déduire les régions des gaps dans les scaffolds (English et al., 2012). Par exemple, le logiciel PBJelly permet d'aligner les longues lectures sur les ébauches d'assemblage (English et al., 2014).

Après avoir obtenu l'assemblage, l'étape suivante consiste à évaluer sa contiguïté (*i.e.* l'absence de gaps) et sa complétude. L'évaluation peut se faire à partir des caractéristiques intrinsèques de l'assemblage ou par des données externes à celui-ci (Bradnam et al., 2013; Ekblom and Wolf, 2014) :

- **Qualité de l'assemblage :**
    - Une de métrique la plus répandue est la **N50**, qui mesure la fragmentation de l'assemblage du génome. Cette métrique représente la longueur du contig la plus courte après avoir « parcouru » 50% de la longueur totale du génome (Figure 7). Généralement, plus la N50 est élevée, moins l'assemblage du génome est fragmenté, plus le génome est considéré comme bien assemblé. Toutefois, cette métrique mesure uniquement la contiguïté de l'assemblage et non la précision, car elle ne peut pas identifier la présence potentielle de biais de fusion des contigs (Yandell and Ence, 2012).

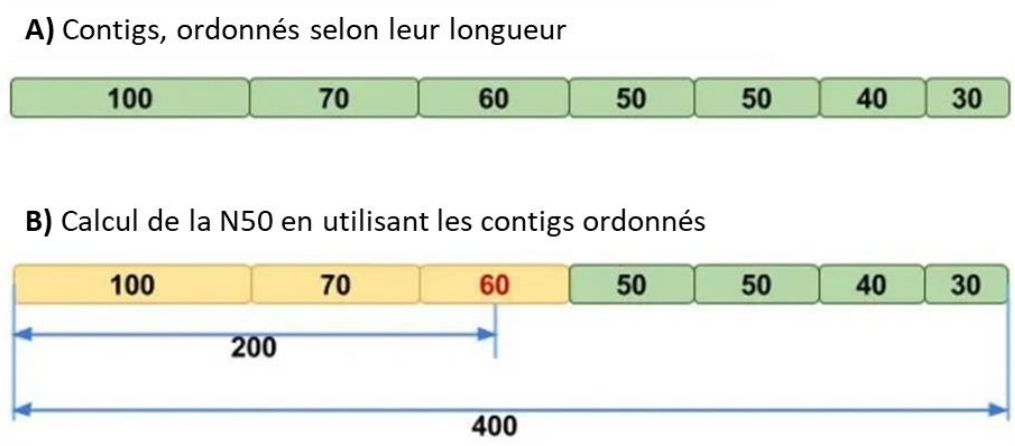


Figure 7 : Exemple de calcul de N50 à partir de 7 contigs. Ici, N50 = 60 kbp (adaptée de Videvall, 2017).

- Le **taux de couverture** peut être utilisé pour le génome et les gènes. La couverture du génome fait référence au pourcentage du génome contenu dans l'assemblage en fonction de l'estimation de la taille du génome. Un assemblage avec une couverture du génome de 90 à 95% est considéré comme bon (Yandell and Ence, 2012). La couverture des gènes est le pourcentage de gènes contenus dans l'assemblage du génome. La couverture des gènes est souvent nettement supérieure au pourcentage de couverture génomique, car les régions répétitives difficiles à assembler sont généralement pauvres en gènes (Yandell and Ence, 2012).

- **Erreurs de séquençage :**
  - Les **assemblages** sont généralement fragmentés et contiennent des **erreurs** telles que la détection de régions avec une profondeur de couverture inhabituelle qui est soit trop élevée (indiquant peut-être la présence d'une séquence répétée) soit trop faible (indiquant peut-être une jointure incorrecte sur la séquence d'ADN) ; un grand nombre de discordances entre la séquence assemblée et les lectures de séquençage (souvent trouvées dans les régions répétées ou les erreurs de jointure); et un appariement incohérent de lectures couplées (Nagarajan and Pop, 2013). Il

est possible de s'aider de **données externes à l'assemblage**, telles que des données de cartographie (Myers et al., 2000; Nagarajan et al., 2008), le transcriptome (Zimin et al., 2009) ou encore le génome d'espèces proches phylogénétiquement (Gnerre et al., 2009; Meader et al., 2010).

- L'**ADN d'autres organismes** est susceptible de contaminer les échantillons génomiques à divers stades (*i.e.* à l'échantillonnage et au laboratoire) et sera présent dans les données de séquençage. Il peut s'agir de contamination non voulue (*e.g.* par l'Homme qui manipule les échantillons), mais la **contamination** peut aussi être liée à la présence de parasites ou micro-organismes de l'espèce (Ekblom and Wolf, 2014). Pour détecter ce genre de contamination, des outils basés sur l'alignement local des séquences (BLAST, Kraken 2) peuvent être utilisés (Wood et al., 2019). Cette étape peut aussi être réalisée en amont de l'assemblage.

- **La complétude :**

En complément de la mesure de la N50, Simão et al. (2015) ont proposé une mesure pour l'évaluation quantitative de l'assemblage du génome (**redondance**) et de l'exhaustivité (**complétude**) de l'annotation en prenant en compte l'évolution du contenu des gènes (Manni et al., 2021; Simão et al., 2015). Il s'agit de l'outil Benchmarking Universal Single-Copy Orthologues (**BUSCO**), mettant à disposition un ensemble de gènes orthologues présent en copie unique. La base de données OrthoDB de gènes orthologues (www.orthdoDB.com) a été utilisée pour définir les ensembles BUSCO pour 6 clades phylogénétiques majeurs. Ces gènes orthologues ont été sélectionnés à partir de l'échantillonnage de centaines de génomes. À l'intérieur de ces génomes échantillonnés, des groupes de gènes orthologues présents en copie unique à plus de 90% ont été sélectionnés (Simão

et al., 2015). Parmi l'ensemble des gènes orthologues sélectionnés pour BUSCO, on dénombre : 3354 gènes pour les vertébrés, 1013 pour les arthropodes, 954 pour les métazoaires, 758 pour les champignons et 255 pour les eucaryotes (Manni et al., 2021). Les clades majeurs couvrent un nombre important de phylums. En alignant les gènes d'un génome donné aux gènes présents dans la base de données de gènes orthologues BUSCO spécifique à l'espèce étudiée, on obtient des correspondances (Figure 8) :

(i) Considérées comme « complètes » si les gènes BUSCO sont retrouvés complets et en copie unique (C pour Complete) (*i.e.* une seule fois dans les gènes du génome donné)

(ii) Considérées comme comme « dupliquées » (D pour Duplicated) si les correspondances complètes sont trouvées plusieurs fois dans les gènes du génome donné

(iii) Partielles (*i.e.* moins de 90% de la longueur moyenne des gènes) qui sont considérées comme « fragmentées » (F pour Fragmented)

(iv) Les gènes sans correspondance sont considérés comme « manquants » (M pour Missing).

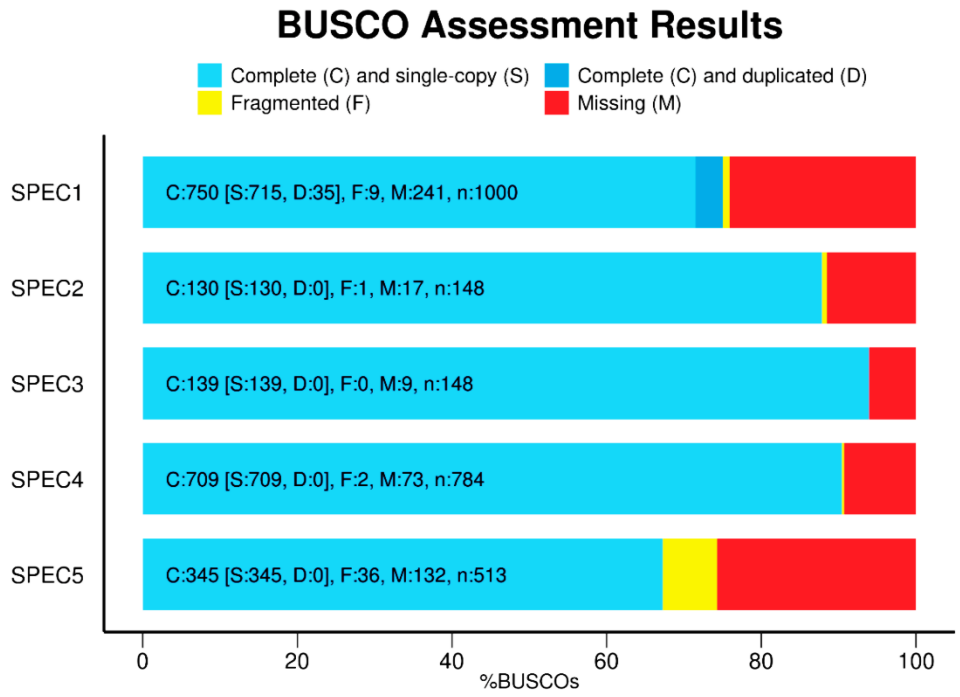


Figure 8 : Exemple de résultats BUSCO ("User guide BUSCO v5.2.2," 2019)

Une méta-analyse a mis en avant le fait que les assemblages avec des valeurs de contigs N50 et de scaffolds N50 élevés avaient aussi des scores BUSCO élevés (Jauhal and Newcomb, 2021). Cependant un score BUSCO élevé pouvait être aussi obtenu à partir d'un assemblage avec un faible N50. Cela confirme que la N50 n'est pas forcément pertinente pour évaluer le contenu d'information biologique d'un génome assemblé. Un score BUSCO élevé et un faible N50 pourraient aussi être liés à une quantité importante d'introns dans les gènes (Jauhal and Newcomb, 2021).

Après l'obtention de l'assemblage du génome de qualité satisfaisante, il est possible de localiser les gènes grâce aux méthodes bioinformatiques. Pour ceci, il est nécessaire de repérer et de décrire les signaux caractéristiques présents dans les séquences d'ADN (Figure 9) (Médigue et al., 2002). Ces signaux correspondent à des « motifs consensus » nécessaires à l'expression des gènes tels que les sites d'épissage, les sites de fixation des ribosomes, les régions promotrices, etc. Un motif est une séquence courte et conservée que l'on peut associer à un rôle fonctionnel ou structural. On peut prendre en exemple les séquences CAAT et TATA (boîte TATA) bien connues,

comme séquences promotrices pour la transcription de certains gènes eucaryotes (Smale and Kadonaga, 2003; Yang et al., 2007) (Figure 9). Plusieurs outils ont été développés pour rechercher des motifs dans les séquences génomiques tels les sites d'épissage (Brendel et al., 1998; Tolstrup et al., 1997), les régions promotrices de la transcription (Prestridge, 1995) ou encore les sites d'initiation de la traduction (Pedersen and Nielsen, 1997). Cependant, il ne suffit pas de connaître le début et la fin d'une région codante en identifiant les « Open Reading Frame » (ORFs) ou « cadres de lectures ouverts » situés entre un codon START et un codon STOP pour caractériser les gènes (Figure 9). En effet, cette simple stratégie ne permet pas toujours d'identifier le codon d'initiation de la traduction et par ailleurs les séquences codantes des exons sont parfois courtes. La taille importante de certains génomes et les exons courts bordés par de grands introns rendent le processus de prédiction des gènes difficile pour les eucaryotes (Goel et al., 2013). De plus, les séquences codantes des eucaryotes sont soumises à l'épissage alternatif, c'est-à-dire un processus d'excision des introns qui permet la jonction des exons de différentes manières lors de l'épissage de l'ARN (Schellenberg et al., 2008) (Figure 9). Par exemple, chez *Homo sapiens* on estime que plus de 95 % des gènes humains présentent au moins un site d'épissage alternatif (Sleator, 2010).

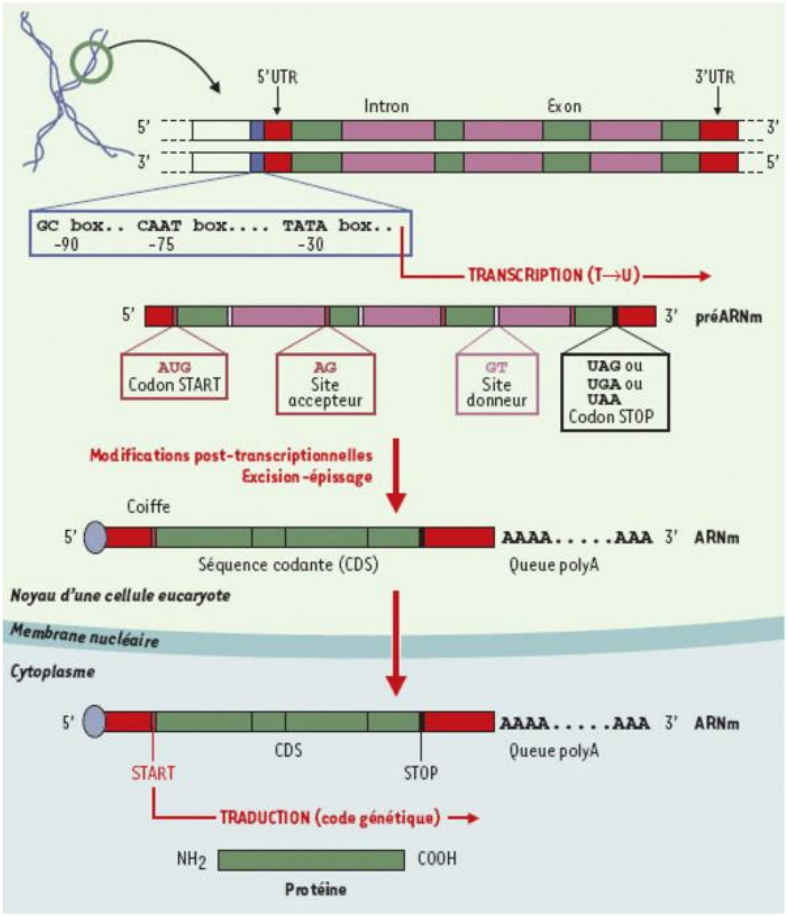


Figure 9 : Structure des gènes chez les organismes eucaryotes (Médigue et al., 2002)

Les méthodes actuelles permettant la prédiction des gènes *in silico* peuvent être classées en deux typologies selon l'information utilisée par les programmes et les algorithmes employés (Figure 10) :

(i) Les marqueurs basés sur le **contenu de l'ADN** ou dits **empiriques** permettent de discriminer les régions codantes et non codantes (*i.e.* introns, régions non traduites (UTR), régions intergéniques). Parmi ces méthodes, il existe deux approches : **intrinsèques** et **extrinsèques** (Figure 10):

a. Les indicateurs de prédiction de gènes **intrinsèques** : la fréquence des hexamères (*i.e.* courtes séquences d'ADN composées de 6 nucléotides) (Fickett and Tung, 1992; Mathé et al., 2002), en sachant qu'il existe 4096 hexamères possibles, et

seulement 40 se retrouvent dans les régions codantes. Il existe également d'autres capteurs intrinsèques tels que la composition en nucléotide (*i.e.* contenu en bases G et C) ou l'utilisation des codons (Sleator, 2010).

b. En se basant sur l'hypothèse que les séquences codantes sont plus conservées que les séquences non codantes (Mathé et al., 2002), les méthodes **extrinsèques** identifient les exons très conservés à partir de l'homologie (inter ou intra génomique) à l'aide de méthodes d'alignement local comme FASTA (Pearson and Lipman, 1988) ou BLAST (Altschul et al., 1990) dans les bases de données d'ADN génomique, d'ADN complémentaires (ADNc) ou de protéine (Figure 10).

(ii) Les marqueurs basés sur **les signaux** caractéristiques (transcription, traduction, épissage) permettent de détecter la présence de sites fonctionnels spécifiques des gènes. Ces méthodes sont dites ***ab initio*** (Figure 10). Il existe une multitude de modèles statistiques qui définissent les régions codantes tels que, les arbres de décision (Salzberg et al., 1998), les modèles de Markov (Korf, 2004; Majoros et al., 2004; Stanke and Waack, 2003; Ter-Hovhannisyan et al., 2008), les méthodes de discrimination comme les réseaux neuronaux (Snyder and Stormo, 1995; Xu et al., 1996), les champs aléatoires conditionnels (DeCaprio et al., 2007), les machines à vecteur de support (SVM) (Schweikert et al., 2009). Des programmes comme GENSCAN (Burge and Karlin, 1997), TWISCAN (Flicek et al., 2003) utilisent ces modèles pour prédire les gènes.

En outre, les méthodes empiriques peuvent aussi être combinées aux méthodes *ab initio* pour améliorer et préciser la structure et l'organisation des gènes (Allen et al., 2004; Sleator, 2010), par exemple avec l'outil GenomeScan (extension de GENSCAN) (Yeh et al., 2001), JIGSAW (Allen and Salzberg, 2005) ou encore MAKER (Cantarel et al., 2008).

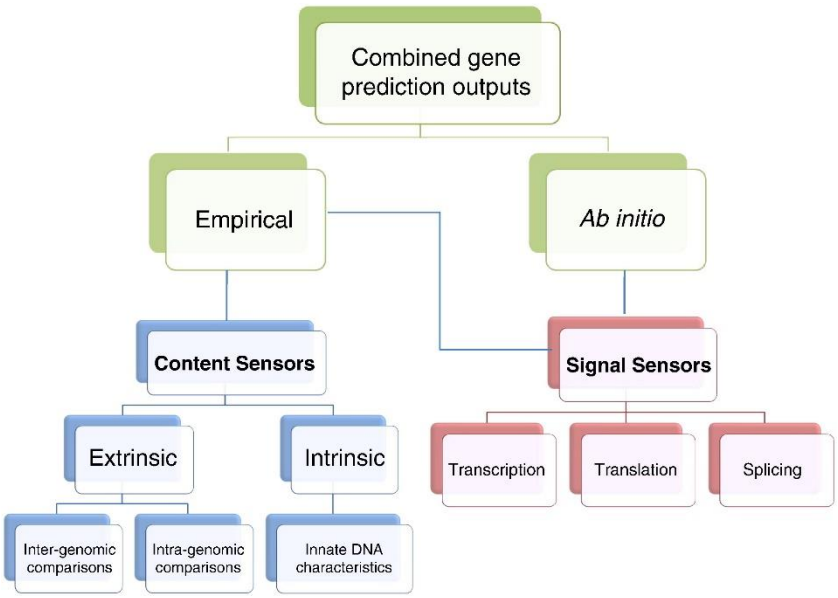


Figure 10 : Classement des différents algorithmes de prédiction des gènes eucaryotes (Sleator, 2010).

En complément des méthodes *ab initio* et empiriques, l'utilisation des **données expérimentales d'expression des gènes** (*i.e.* transcriptome, protéome) peut apporter un soutien important dans l'annotation des génomes. En effet, la transcription inverse de l'ARNm permet d'obtenir les ADNc des gènes (Médigue et al., 2002). Il existe aussi les « expressed sequence tags » (marqueurs de séquences exprimés), dits **EST**, qui proviennent du séquençage des extrémités d'ADNc qui peuvent permettre de mettre en évidence des régions exoniques de l'ADN. Ces EST permettent alors d'identifier de manière plus précise les jonctions intron-exon grâce à des programmes de recherche de similarité entre séquences à partir de banque d'ADNc (Florea et al., 1998; Stanke et al., 2008) ou d'EST (Jiang and Jacob, 1998; Stanke et al., 2006) comme AUGUSTUS (Stanke et al., 2004). Cette méthode a cependant des limites concernant la qualité moindre des EST comparé à l'ADN génomique et la taille relativement courte de ces tags. De ce fait, l'acquisition de données transcriptomiques par séquençage est souvent incluse dans les projets génomiques aujourd'hui (Kao et al., 2016; Poynton et al., 2018).

Le **protéome** peut également être utilisé pour améliorer la prédiction des gènes. En effet, il est possible à partir de base de données de protéines comme UniProtKB/Swiss-Prot (The UniProt Consortium, 2009), de comparer les ORFs (*i.e.* traduction dans les 6 cadres de lectures des séquences génomiques) à ces protéines annotées dans les bases de données. La meilleure conservation des séquences protéiques (*i.e.* dégénérescence du code génétique), ainsi que la taille plus importante de leur alphabet permettent de rendre la comparaison protéines-protéines plus sensible que la comparaison ADN-ADN.

1.3 La transcriptomique

La transcriptomique est l'étude de tous les transcrits d'ARN codants et non codants produits à partir de l'ADN (Nguyen et al., 2019) (Figure 11). La transcriptomique vise à décrire les profils globaux d'ARN à l'échelle des cellules, organes ou organismes entiers à un moment donné ou dans une condition physiologique, pathologique ou d'exposition déterminée.

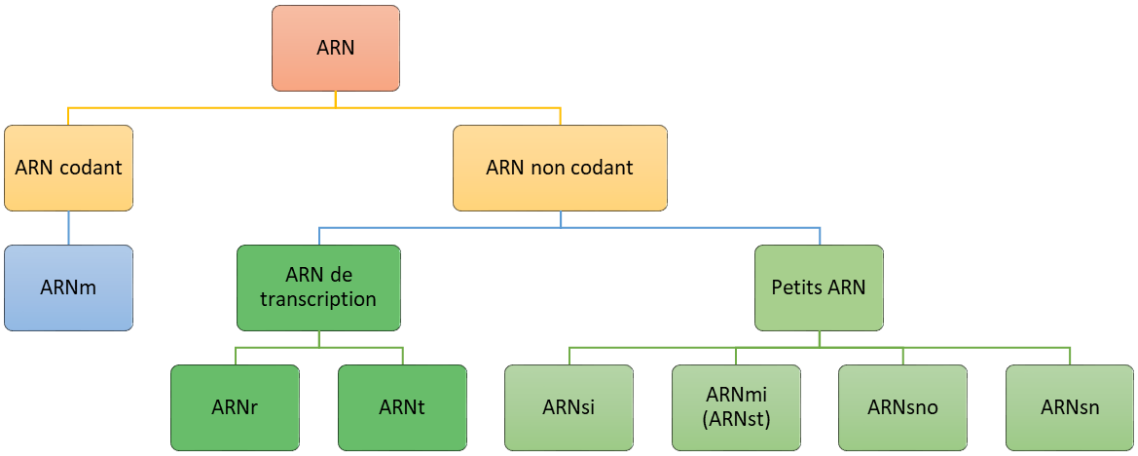


Figure 11 : Les différents types d'ARN. ARNm : ARN messager, ARNr : ARN ribosomique, ARNt : ARN de transfert, ARNsi : petit ARN interférent ou small interfering ARN, ARNmi : micro ARN, ARNsno : petit ARN nucléolaire ou small nucleolar ARN, ARNsn : petit ARN nucléaire ou small nuclear ARN (adaptée de Buckingham, 2003).

Pour associer un gène à une fonction biologique (*i.e.* une protéine), comprendre les évènements clés (quand et comment) de la régulation de la traduction (passage d'un ARNm à une protéine) est crucial. L'objectif de la génomique fonctionnelle est d'étudier le **transcriptome** (ARN)

et le **protéome** (protéines traduites à partir des ARNm) pour associer les fonctions des produits de la transcription (ARN) et les protéines à ces gènes (Figure 12).

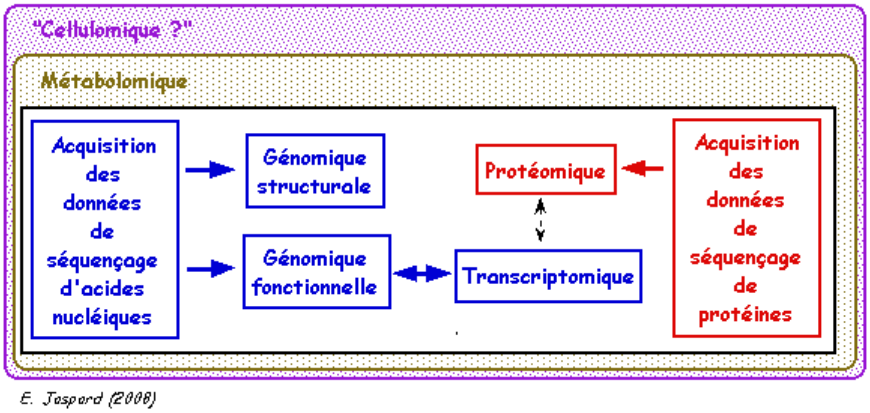


Figure 12 : Les volets complémentaires de la génomique (Jaspard, 2008).

Plusieurs méthodes moléculaires, comprenant l'analyse en série/bouchon de l'expression génique (SAGE/CAGE), l'hybridation soustractive avec suppression (SSH), l'EST, les puces à ADN (microarray), le séquençage d'ARN (RNA-Seq) et la PCR quantitative (qPCR) ont été développées pour la caractérisation de la transcription génique globale (Duran et al., 2016). Certaines de ces techniques, telles que SAGE, CAGE et EST (basées sur le séquençage Sanger de l'ADNc ou EST) ne sont plus utilisées pour les analyses transcriptomiques (Nguyen and Alfaro, 2020). Les techniques de puces à ADN sont disponibles pour certains organismes modèles tels que *Daphnia magna* ou le poisson *Danio rerio* (Giraudo et al., 2015). Les techniques de séquençage de l'ARN (dit RNA-Seq) permettent d'étendre la transcriptomique à des organismes non modèles d'intérêt (Trapp et al., 2016a) (Figure 13). Le RNA-Seq est un outil révolutionnaire qui utilise des technologies de séquençage haut débit combinées à des méthodes informatiques pour cartographier et quantifier les transcriptomes (Lowe et al., 2017; Wang et al., 2009). Il présente plusieurs avantages par rapport aux puces à ADN (*i.e.* une meilleure sensibilité et spécificité, connaissance préalable du génome non nécessaire) et est donc devenu la technologie dominante (Duran et al., 2016; Nguyen and Alfaro, 2020).

Contrairement au génome, le transcriptome est variable dans le temps et selon les types cellulaires et les environnements (Lowe et al., 2017). Cela nécessite de multiplier les réplicats biologiques étant donné la nature de la variance biologique des populations (Trapp et al., 2016a). Les millions de lectures de séquences courtes générées par le RNA-Seq nécessitent le développement de méthodes de calcul pour la caractérisation du transcriptome. Les étapes d'analyses varient selon la question biologique et la disponibilité ou non de données génomiques ou transcriptomiques de référence (Duran et al., 2016). La Figure 14 donne un aperçu des différentes étapes de l'analyse du transcriptome pour l'annotation et l'enrichissement. L'analyse comprend généralement le contrôle de la qualité et le prétraitement des données de séquençage, l'assemblage du transcriptome (*de novo* ou avec transcriptome de référence), la quantification de l'expression, l'analyse statistique de l'expression et l'annotation fonctionnelle. Les différentes étapes de l'analyse peuvent être effectuées à l'aide différents outils automatisés ou de pipelines existants et mis à disposition, comme GENE-Counter (Cumbie et al., 2011), Tuxedo (Trapnell et al., 2012), PRADA (Torres-García et al., 2014), MAP-R-SEQ (Kalari et al., 2014) et celui de Yalamanchili et al. (2017).

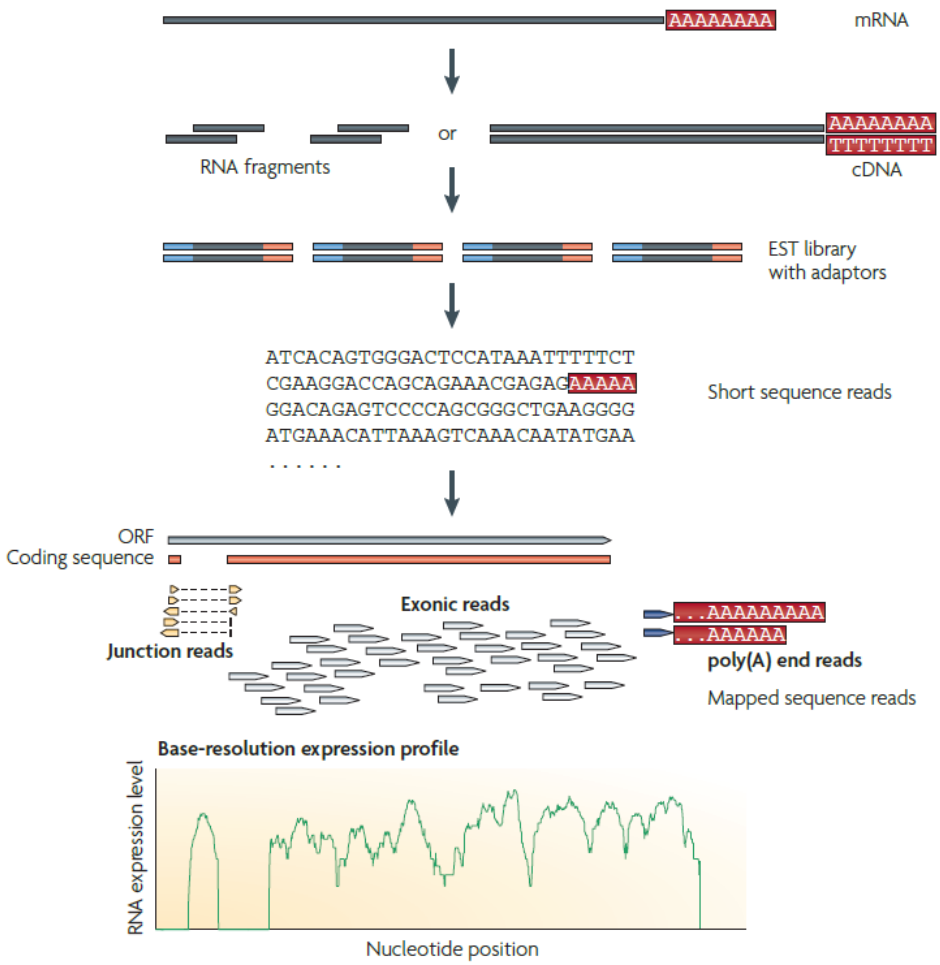


Figure 13 : Principe du RNA-Seq : Les ARNm sont convertis en une bibliothèque de fragments d'ADNc par fragmentation d'ARN. Des adaptateurs de séquençage (bleus) sont ajoutés à chaque fragment d'ADNc et une courte séquence est obtenue à partir de chaque ADNc grâce au séquençage. Les lectures de séquence résultantes sont alignées avec le génome ou le transcriptome de référence et classées en trois types : lectures exoniques, lectures de jonction et lectures poly(A) terminale. Ces trois types sont utilisés pour générer un profil d'expression de résolution de base pour chaque gène, comme illustré en bas ; une ORF de levure avec un intron est montrée. (Wang et al., 2009).

Une application importante de la transcriptomique consiste à comparer les différences d'expression génique entre différents groupes d'individus afin d'accéder aux variations de phénotypes entre les groupes (Nachtomy et al., 2007). Le RNA-Seq permet aussi d'obtenir des valeurs d'expression relatives de milliers de gènes par échantillon, ce qui permet une comparaison quantitative de l'expression des gènes entre deux conditions biologiques (Prat and Degli-Esposti, 2019). Au-delà de la quantification de l'expression génique, les données générées par RNA-Seq facilitent la découverte de nouveaux transcrits, l'identification de gènes épissés alternativement et la détection de l'expression spécifique d'un allèle (Kukurba and Montgomery, 2015).

La première étape de l'analyse des données est le contrôle de la qualité des lectures, de manière semblable aux données génomiques (voir partie 1.2) (Figure 14). La qualité des lectures est primordiale pour éviter les erreurs d'alignements, d'assemblages ou encore de quantification de l'expression. Par exemple, le « Phred quality score Q » (ou scores Phred) permet de mesurer la qualité de l'identification des bases de nucléotides (A, T, G ou C) générées par le séquençage. Ce score indique la probabilité que la base d'un nucléotide soit appelée de manière incorrecte par le séquenceur. Ainsi, un score phred de 30 (*i.e.* Q30) pour une base correspond à 1 chance sur 1000 que la base soit appelée incorrectement (*i.e.* taux d'erreur), et à une précision de 99,9%. Généralement, un score phred de 20 (*i.e.* 99% de précision d'appel de la base) est considéré comme un seuil minimum pour la qualité des lectures.

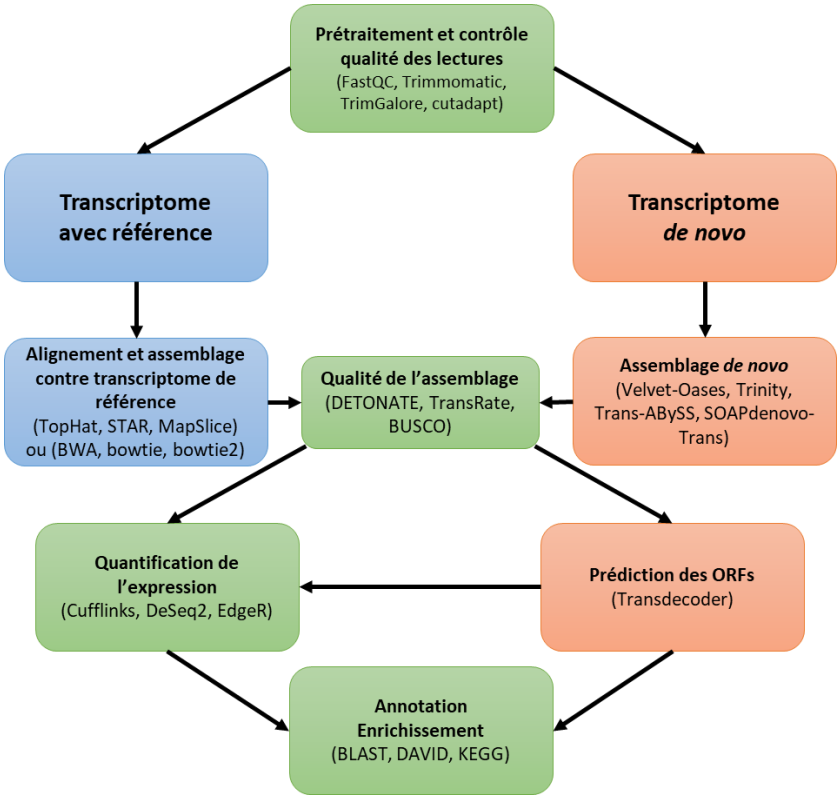


Figure 14 : Étapes d'analyse RNA-Seq *de novo* ou avec transcriptome de référence. Entre parenthèses : exemples d'outils. En bleu : spécificité de l'assemblage avec référence. En orange : spécificité de l'assemblage de novo. En vert : commun aux deux stratégies (adaptée de Anamika et al., 2016).

Pour évaluer la qualité des transcriptomes, les métriques utilisées en génomique sont à prendre avec précaution. Il a été montré que des métriques comme la N50 ne reflétait pas forcément la qualité des assemblages, car les transcriptomes ne cherchent pas à obtenir les contigs les plus longs (*i.e.* un N50 élevé) mais plutôt un contig par transcrit (O'Neil and Emrich, 2013). De plus, les contigs les plus fortement exprimés ne sont pas nécessairement les plus longs, et la plupart des transcrits auront des niveaux d'expression relativement faible (Senatore et al., 2015). Les métriques ou les scores basés sur l'annotation des transcrits sont de meilleurs évaluateurs, comme DENOTATE (Li et al., 2014), TransRate (Smith-Unna et al., 2016) ou rnaQUAST (Bushmanova et al., 2016). Par exemple, Transrate permet de calculer le taux de lectures alignées sur les transcrits après assemblage et le traduit en un score (Smith-Unna et al., 2016). DETONATE propose, le score RSEM-EVAL (Li et al., 2014) qui est basé sur un modèle probabiliste qui dépend uniquement d'un assemblage et des lectures utilisées pour le construire. La complétude est aussi évaluable avec l'outil BUSCO (Simão et al., 2015).

Après le prétraitement et le contrôle qualité, les lectures sont prêtes pour l'assemblage et l'analyse du profil transcriptomique. Il existe deux approches : (i) l'assemblage *de novo* quand le génome de référence n'est pas disponible, et (ii) l'alignement à partir d'un génome de référence quand il est disponible pour l'espèce étudiée. Dans l'approche *de novo*, les lectures de séquençage sont alignées directement à partir de leur chevauchement respectif (Anamika et al., 2016). Pour la deuxième approche avec le génome de référence, les lectures sont alignées sur celui-ci et les transcrits sont ensuite assemblés (Tableau 1). Les différences techniques et les objectifs de ces deux approches sont listés dans le Tableau 1, et présentés ci-après.

Tableau 1 : Différence entre l'assemblage à partir d'un génome de référence et l'assemblage *de novo* (adaptée de Anamika et al., 2016).

Assemblage par génome de référence	**Assemblage *de novo***
Génome de référence nécessaire pour assembler le transcriptome	Transcriptome assemblé *de novo*
Moins intense en calcul	Plus intense en calcul
Les contaminants et les artefacts de séquençage ne sont pas une préoccupation majeure	Les contaminants et les artefacts de séquençage peuvent entraîner une mauvaise qualité du transcriptome assemblé
La qualité de l'alignement des transcrits dépend des aligneurs d'épissage	L'alignement n'est pas nécessaire
Possibilité d'assembler des transcrits de faible abondance	Difficulté à assembler les transcrits de faible abondance, à moins que la profondeur de séquençage ne soit élevée
Fonctionne bien avec des données de faible profondeur de séquençage (~ 10X)	Fonctionne bien avec des données de profondeur de séquençage élevées (~ 30X)
Moins efficace pour identifier de nouvelles isoformes et SNP	Efficace pour identifier de nouvelles isoformes et SNP
L'exhaustivité et la contiguïté du transcriptome sont relativement plus élevées	L'exhaustivité et la contiguïté du transcriptome sont relativement plus faibles, en particulier pour les données de faible profondeur de séquençage

Dans le cas de l'assemblage *de novo*, les assembleurs utilisent les mêmes principes que les assembleurs pour les séquences génomiques (voir la partie 1.2). Les chevauchements entre lectures permettent l'assemblage *de novo*. L'algorithme le plus couramment utilisé est celui des graphes de Bruijn (voir page 24). La liste de différents assembleurs et leurs caractéristiques sont présentées dans le Tableau 2. Ainsi, à la suite de la reconstruction du graphique, les transcrits sont extraits avec les outils tels que Velvet-Oases (Schulz et al., 2012), Trans-ABySS (Robertson et al., 2010), Trinity

(Grabherr et al., 2011), SOAPdenovo-Trans (Xie et al., 2014), ou encore récemment rnaSPADES (Bushmanova et al., 2019).

Tableau 2 : Liste des différents assembleurs de transcriptomes *de novo* (adaptée de Anamika et al., 2016).

Nom de l'outil	Algorithme	Type de lectures	Références
Mira	OLC	Lectures simples ou appariées	(Chevreux et al., 2004)
Velvet-Oases	Graphe de Bruijn	Lectures simples ou appariées	(Schulz et al., 2012; Zerbino and Birney, 2008)
Trans-ABySS	Graphe de Bruijn	Lectures simples ou appariées	(Robertson et al., 2010)
Trinity	Graphe de Bruijn	Lectures simples ou appariées	(Grabherr et al., 2011)
IDBA-Tran	Graphe de Bruijn	Lectures appariées	(Peng et al., 2013)
SOAPdenovo-Trans	Graphe de Bruijn	Lectures simples ou appariées	(Xie et al., 2014)
Bayesembler	Modèle bayésien	Lectures appariées	(Maretty et al., 2014)
rnaSPADES	Graphe de Bruijn	Lectures simples ou appariées	(Bushmanova et al., 2019)

L'alignement des lectures contre un génome de référence permet de (i) faire un comptage des lectures de séquençage et de quantifier l'expression (*i.e.* abondance) des transcrits, et (ii) d'identifier de nouveaux sites d'épissage et de nouvelles isoformes. Afin de déterminer l'abondance des transcrits à partir des lectures pour quantifier l'expression, il est nécessaire d'aligner les lectures sur un génome ou un transcriptome de référence pour déterminer l'origine de la lecture. Cette approche est plus courante chez les espèces modèles, car il augmente les informations potentielles

qui peuvent être obtenues (*e.g.* l'identification de nouveaux transcrits et gènes) et parce que de nombreux transcriptomes sont incomplets (Duran et al., 2016). L'alignement est une tâche difficile, car les lectures de RNA-Seq sont relativement courtes et peuvent correspondre à des régions non contiguës du génome en raison de jonctions d'épissage. De plus, les outils d'alignement doivent faire face aux erreurs d'appariements et aux indels (*i.e.* insertion ou délétion dans une séquence) causés par la variation génomique et les erreurs de séquençage. De nombreux outils d'alignement ont été développés. Une liste complète de ces outils et de leurs propriétés a été initialement publiée par Fonseca et al. (2012) et est tenu à jour sur le site de bio.tools (https://bio.tools/) (Ison et al., 2016). Il existe deux cas de figure, le premier correspond à un alignement à partir d'une annotation existante. En effet, s'il existe déjà une annotation des exons – gènes – transcrits avec leur position, plusieurs outils permettent d'aligner les courtes lectures comme BWA (Li and Durbin, 2009), bowtie (Langmead et al., 2009) ou bowtie2 (Langmead and Salzberg, 2012). Dans le deuxième cas, en absence d'annotations existantes ou si l'objectif est d'identifier des nouveaux exons ou isoformes, il est possible de reconstruire la structure du transcriptome à partir des lectures et de l'alignement. Il existe des outils permettant l'alignement sans s'appuyer sur des sites d'épissages déjà connus, tels que TopHat (Trapnell et al., 2009), MapSplice (Wang et al., 2010) et STAR (Dobin et al., 2013). À la suite de l'alignement les lectures, l'un des outils les plus couramment utilisés est Cufflinks (Trapnell et al., 2010), qui permet de connecter les lectures à partir de l'emplacement de leur alignement épissé.

Pour quantifier l'expression des gènes, il est nécessaire d'aligner les lectures brutes du jeu de données de départ contre le transcriptome nouvellement assemblé. Cette quantification peut s'effectuer à différents niveaux : exons, transcrits ou gènes. L'expression d'un gène peut être exprimée comme la somme de l'expression de toutes ses isoformes, en comptant les lectures par

gène selon la référence utilisée pour l'alignement (Duran et al., 2016). Le comptage des lectures doit être normalisé en raison de la variabilité induite par le biais de la longueur des lectures (Oshlack and Wakefield, 2009) et de la profondeur du séquençage par échantillon (Duran et al., 2016; Mortazavi et al., 2008) (Tableau 3). Les outils expriment cette quantification généralement en nombres de lectures brutes, en lectures par kilobase de transcrits par million de lectures alignées (RPKM), ou en fragments par kilobase de transcrits par million de lectures alignées (FPKM) (Tableau 3). Il existe donc différentes unités qui permettent de traduire le niveau d'expression des transcrits comme défini dans le Tableau 3 (Anamika et al., 2016) :

Tableau 3 : Les différentes mesures de quantification de l'expression.

Unité	Définition	Remarques
Nombre de lectures	Nombre de lectures chevauchant une unité génomique (transcrit ou gène)	
CPM : nombre de lectures mappées par million	Nombres de lectures calibrées par le nombre de fragments séquencés multiplié par un million.	Utilisée dans les analyses différentielles réalisées à l'aide du package R EdegR (Robinson et al., 2010).
RPKM : lectures par kilobase de transcrits par million	Calculé en divisant le nombre de lectures par la longueur et le total des lectures séquencées puis en multipliant par un million.	Seulement pour lectures simples (single –end)
FPKM : fragments par kilobase de transcrits par million	Similaire au RPKM mais prend en compte un seul fragment (pas les lectures). Pour les données paired-end : deux lectures peuvent correspondre à un seul fragment du génome. Dans ce cas, un seul compte sera ajouté.	Lectures par pairs (paired-end)
TPM : transcrits par million	Calculé en divisant le ratio entre le nombre de lectures sur la longueur du transcrit donné, par la somme des ratios de tous les transcrits, en multipliant le tout par un million.	Pour estimer l'abondance des transcrits

Des algorithmes ont été développés pour estimer l'expression au niveau des transcrits pour appréhender le problème des transcrits apparentés partageant plusieurs lectures (Conesa et al., 2016). Cufflinks (Trapnell et al., 2010) estime l'expression des transcrits à partir d'un alignement sur le génome, obtenu à partir d'aligneurs tels que TopHat (Trapnell et al., 2009) en utilisant une approche probabiliste qui estime les abondances des transcrits. Cette approche prend en compte des biais tels que la distribution non uniforme des lectures sur la longueur du gène. Les algorithmes qui quantifient l'expression à partir des alignements du transcriptome comprennent notamment RSEM (RNA-Seq by Expectation Maximization) (Li and Dewey, 2011), eXpress (Roberts and Pachter, 2013) ou kallisto (Bray et al., 2016). Ces méthodes répartissent les lectures s'alignant à plusieurs emplacements entre les transcrits (*i.e.* multi-mapping reads) et produisent des valeurs normalisées, corrigées des biais de séquençage (Roberts et al., 2011).

Suite à la quantification de l'expression des gènes, il est possible de procéder à l'analyse de l'expression différentielle des transcrits. Cette analyse exige que les valeurs d'expression soient comparées entre les échantillons. Une fois que les comptages de la quantification de l'expression ont été réalisés pour chaque transcrit, l'expression différentielle des gènes peut être mesurée en normalisant, en modélisant et en testant statistiquement les données entre les différentes conditions.

Comme la quantification RNA-seq est basée sur des comptes de lecture qui sont attribués de manière absolue ou probabiliste aux transcrits, les premières approches pour calculer l'expression différentielle utilisaient des distributions de probabilité discrètes, telles que la distribution de Poisson ou la distribution binomiale négative (Anders and Huber, 2010; Robinson and Smyth, 2007). Cependant, en prenant en compte la variance de l'échantillonnage des petits comptes de lecture (notamment avec peu de réplicats), et une normalisation (comme la normalisation TMM ou

la suppression des effets batch), les données peuvent perdre leur nature discrète et s'apparenter à une distribution continue (Conesa et al., 2016). Certaines méthodes, comme edgeR (Robinson et al., 2010), prennent en données d'entrée des comptages de lecture bruts et corrigent des sources de biais possibles dans le modèle statistique pour effectuer une normalisation intégrée ainsi qu'une analyse d'expression différentielle (Tableau 4). Dans d'autres méthodes, l'expression différentielle nécessite que les données soient préalablement normalisées pour éliminer tous les biais possibles. DESeq2 (Love et al., 2014), comme edgeR (Robinson et al., 2010), utilise la binomiale négative comme distribution de référence et fournit sa propre approche de normalisation (Anders and Huber, 2010; Love et al., 2014) (Tableau 4). Les méthodes baySeq (Hardcastle and Kelly, 2010) et EBSeq (Leng et al., 2013) sont des approches bayésiennes, également basées sur le modèle binomial négatif, qui définissent une collection de modèles pour décrire les différences entre les groupes expérimentaux, pour calculer la probabilité postérieure de chacun d'entre eux pour chaque gène (Tableau 4).

Compte tenu de la baisse du prix du séquençage, il est recommandé que les expériences RNA-Seq comportent un minimum de trois réplicats biologiques lorsque la disponibilité des échantillons n'est pas limitée afin de permettre à toutes les méthodes d'expression différentielle de bénéficier de la reproductibilité entre les réplicats (Conesa et al., 2016). Les résultats de ces analyses sont des listes de gènes avec des tests par paires associés (dits pairwise tests) pour l'expression différentielle entre les échantillons selon les conditions, ainsi que les estimations de probabilité de ces différences (Lowe et al., 2017).

Tableau 4 : Liste non exhaustive des différents logiciels réalisant une analyse différentielle (adapté de Lowe et al., 2017)

Logiciel	Langage/Environnement	Spécialisation	Référence
Cuffdiff2	C++	Analyse des transcrits au niveau des isoformes	(Trapnell et al., 2013)
EdgeR	R/Bioconductor	Toute donnée de comptages en génomique	(Robinson et al., 2010)
DESeq2	R/Bioconductor	Types de données flexibles, réplication faible	(Love et al., 2014)
Limma/Voom	R/Bioconductor	Puces à ADN ou données RNA-Seq, analyse des isoformes, design expérimental flexible	(Ritchie et al., 2015)
BaySeq	R/Bioconductor	Approche bayésienne	(Hardcastle and Kelly, 2010)
EBseq	R/Bioconductor	Approche bayésienne	(Leng et al., 2013)

La visualisation et la synthèse des résultats d'expression différentielle sont importantes pour les interprétations en aval. Par exemple, une heatmap (Figure 15) est une technique de visualisation des données couramment utilisée pour trouver des groupes (dits clusters) de gènes différentiellement exprimés. Les heatmaps permettent de représenter les gènes ou transcrits en fonction des échantillons et des conditions (Figure 15). .

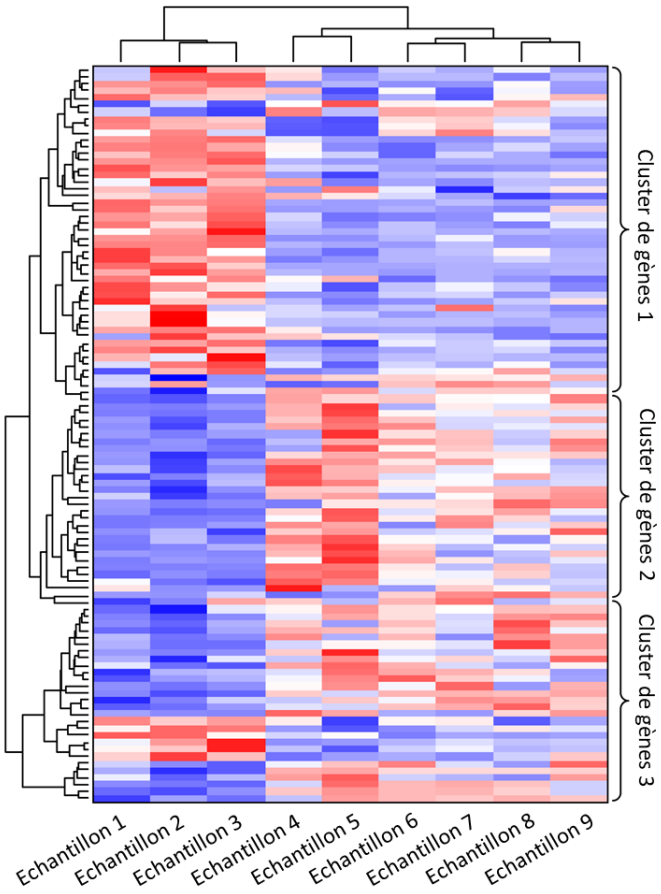


Figure 15 : Heatmap permettant d'identifier les modèles de profils d'expression des gènes ou transcrits à travers différents échantillons et conditions. Chaque colonne contient les variations des mesures d'expression des gènes/transcrits. Une surexpression est représentée par la couleur rouge, une expression moyenne (ou absence de modulation) par la couleur blanche et une sous-expression par la couleur bleue. Les gènes ayant des profils d'expression similaires sont regroupés grâce aux clustering hiérarchiques (arbres) représentés en haut et à gauche de la heatmap (adaptée de Lowe et al., 2017).

En parallèle, l'annotation des séquences transcriptomiques vise à associer les séquences transcriptomiques à une information ou fonction biologique exploitable. L'annotation est restreinte aux transcrits possédant des ORFs et donc à la séquence codant pour les protéines (CDS) (Sieber et al., 2018). Les régions codantes dans les transcrits assemblés *de novo* peuvent être recherchées à l'aide d'outils de prédiction des ORF tels que Transdecoder (http://transdecoder.Github.io), getorf (du package EMBOSS http://emboss.sourceforge.net) (Rice et al., 2000) ou encore plus récemment orfipy (Singh and Wurtele, 2021).

Pour procéder à **l'annotation fonctionnelle** des transcrits, il est possible d'utiliser des programmes de recherches d'homologie de séquence, qui associent la fonction de la protéine connue à celle prédite. Il est aussi possible de reconnaître des domaines fonctionnels qui correspondent à des motifs d'acides aminés caractéristiques qui appartiennent à des classes de protéines telles que les lipases, les protéases, les kinases, etc. Des bases de données comme InterPro (Blum et al., 2021) regroupent 13 bases de données de signatures protéiques en une seule ressource centrale. L'interface InterProScan permet à partir de la base de données InterPro, la reconnaissance des signatures protéiques et de retrouver la fonction potentielle des transcrits assemblés (Zdobnov and Apweiler, 2001). Cette analyse reste néanmoins coûteuse en temps de calcul (Duarte et al., 2021). Alternativement, il existe des outils comme Trinotate (http://trinotate.github.io) (Bryant et al., 2017) qui peuvent intégrer des recherches BLAST quand une espèce parente proche est bien annotée (Duarte et al., 2021). En complément de l'annotation automatique, il est possible d'annoter manuellement en validant ou en invalidant les fonctions prédites grâce aux résultats expérimentaux. En effet, l'annotation manuelle est indispensable pour limiter la propagation des erreurs qui peuvent rapidement se répandre dans les bases de données.

Afin d'interpréter biologiquement la liste de transcrits différentiellement exprimés, il est pertinent d'identifier les catégories fonctionnelles des gènes et les voies biologiques qui se trouvent enrichies. Pour ce faire, **l'analyse d'enrichissement** (ou de surreprésentation) recherche des ensembles de gènes qui sont significativement surreprésentés dans une liste de gènes donnée, par rapport à un ensemble de gènes de fond. Ces ensembles de gènes sont généralement des gènes qui fonctionnent ensemble dans une voie biologique connue. En pratique, ils sont compilés à partir de bases de données d'annotation de gènes et de voies telles que GO (Ashburner et al., 2000), KEGG (Kanehisa and Goto, 2000), Reactome (Croft et al., 2011), Wikipathways (Slenter et al.,

2018), BioCyc (Karp et al., 2005) ou autres (Simillion et al., 2017). La liste de gènes différentiellement exprimés est testée statistiquement pour l'enrichissement, en quantifiant ou en comparant les fréquences des termes GO parmi différents groupes de séquences originales. L'approche la plus simple consiste à utiliser un test exact unilatéral de Fisher, également appelé test hypergéométrique, pour déterminer l'importance de la surreprésentation d'un ensemble de gènes dans la liste d'entrée. L'inconvénient de cette méthode est qu'elle nécessite une frontière nette entre les gènes inclus et exclus. D'autres types d'analyses, comme la plupart des expériences de transcriptomique, renvoient une liste de valeurs p (*i.e.* p-value) associées à chaque gène. Ces p-value expriment l'importance de l'expression différentielle d'un gène entre différentes conditions. La définition d'une frontière entre les gènes exprimés de manière différentielle (DEG) et les non-DEG repose alors sur l'application d'un seuil arbitraire de la p-value, qui influence le résultat d'une analyse d'enrichissement (Pan et al., 2005). Ces méthodes restent simples d'utilisation et permettent d'appliquer des analyses d'enrichissement pour les organismes non modèles qui ne possèdent pas d'annotation fonctionnelle définie. Il existe également d'autres méthodes alternatives que Tarca et al (2013) ont examiné et évalué. Dans leur article, les auteurs font la distinction entre les méthodes Functional Class Scoring (FCS) qui calculent un score basé sur une valeur statistique, telle que la p-value ou le rang, pour tous les gènes qui appartiennent à un ensemble de gènes donné, et les méthodes Single-Sample (SS) où pour chaque ensemble de gènes, un score par échantillon est calculé. Malgré ces développements, la méthode hypergéométrique est encore largement utilisée, principalement en raison de sa simplicité et parce qu'elle peut être appliquée à d'autres questions que l'enrichissement (Simillion et al., 2017). Plusieurs plateformes accomplissent ce type d'analyse, incluant : Blast2GO (Conesa et al., 2005), DAVID (Dennis et al., 2003), GoMiner (Zeeberg et al., 2005), GFINDer (Masseroli et al., 2005), GSEA (Subramanian et al., 2005), GOrilla (Eden et al., 2009) ou encore Enrichr (E. Y. Chen et al., 2013; Kuleshov et al., 2016)

(liste non exhaustive). Il existe également des outils de visualisation des analyses d'enrichissement comme BiNGO (Maere et al., 2005) ou ClueGO (Bindea et al., 2009), deux plug-in de Cytoscape (Shannon et al., 2003), qui permet de représenter les catégories GO dans des réseaux biologiques plutôt que comme de longues listes ou d'arbres hiérarchiques. Bien que chaque outil ait des caractéristiques et des atouts distincts, comme l'ont examiné Khatri et al (2005), ils adoptent tous une stratégie de base commune pour mapper systématiquement un grand nombre de gènes d'intérêt dans une liste avec l'annotation biologique associée (*e.g.* les termes d'ontologie des gènes). Ils mettent ensuite en évidence statistiquement l'annotation biologique la plus surreprésentée (enrichie) parmi des milliers de termes et de contenus liés. Les analyses d'enrichissement sont aussi réalisables en se basant sur d'autres types de bases de données telles que les voies métaboliques de KEGG (Kanehisa and Goto, 2000), de MetaCyc (Caspi et al., 2008) , de BioCarta (Nishimura, 2001) ou encore les voies Reactome (Croft et al., 2011; Fabregat et al., 2018).

L'analyse d'enrichissement est une stratégie prometteuse qui augmente la probabilité pour les chercheurs d'identifier les processus biologiques les plus pertinents (Huang et al., 2009).

1.4 La protéomique et la protéogénomique

1.4.1 Définition et objectifs

La protéomique est l'étude à grande échelle du protéome, qui correspond à l'ensemble des protéines exprimées dans les cellules, les tissus, les organes ou les organismes entiers dans des conditions et à un moment précis (Graves and Haystead, 2002). L'étude des protéines s'étend à leur expression, leur fonction ou encore leur structure. De même qu'en transcriptomique, la dynamique du protéome varie selon le type de cellules, dans le temps, et est modifiable selon les conditions environnementales (Fliser et al., 2007). De plus, les approches protéomiques permettent de donner des informations sur les modifications post-traductionnelles qu'elles peuvent subir (Nguyen et al.,

2019; Rodrigues et al., 2012) (Figure 16). En effet, comme la Figure 16 le présente, la formation d'ARNm n'est que la première étape de plusieurs évènements pouvant aboutir à une protéine : l'épissage de l'ARN pour former l'ARNm peut générer de multiples isoformes de protéines ; la stabilité et l'efficacité de la traduction peuvent être régulées par l'ARNm ; les protéines peuvent être régulées par des mécanismes supplémentaires tels que la modification post-traductionnelle, la protéolyse ou la compartimentation (Graves and Haystead, 2002) (Figure 16).

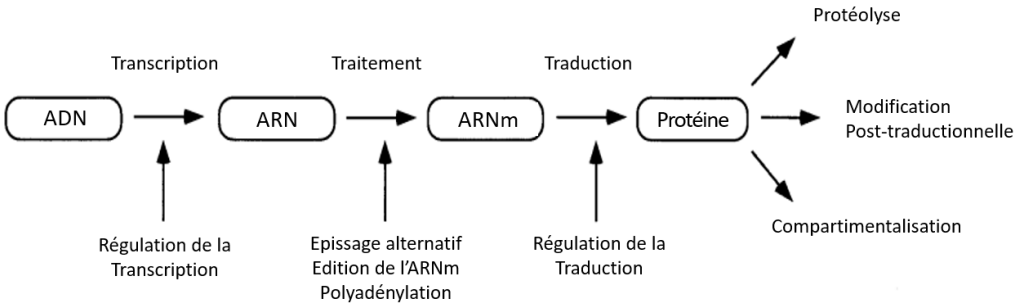


Figure 16 : Du gène aux produits du gène (Graves and Haystead, 2002)

Les outils utilisés dans l'analyse protéomique ont été développés à partir de l'électrophorèse sur gel bidimensionnelle traditionnelle (2D-GE). Il s'agit d'une méthodologie protéomique simple, fiable et économiquement abordable (Rodrigues et al., 2017). La limite de cette méthode traditionnelle est sa faible sensibilité, en permettant l'analyse des protéines solubles les plus abondantes. Par conséquent, une approche sans gel basée sur la spectrométrie de masse (MS) couplée à la chromatographie liquide (LC) a récemment émergée et représente l'approche la plus courante aujourd'hui. En raison de son haut débit, les approches basées sur la LC-MS sont capables d'analyser des milliers de peptides dans un seul échantillon, ce qui la rend très efficace pour identifier les protéines et analyser les modifications post-traductionnelles (Nguyen and Alfaro, 2020; Rodrigues et al., 2017).

Avec les progrès des omiques et des technologies analytiques, la protéomique est de plus en plus appliquée à diverses disciplines (Debnath et al., 2010). Dans le domaine de l'épidémiologie et

de l'immunologie, elle est principalement utilisée pour identifier des biomarqueurs potentiels pour le diagnostic de maladies (Debnath et al., 2010; Wheelock et al., 2013).

1.4.2 Outils et stratégies en protéomique

La protéomique shotgun fait référence à l'utilisation de techniques de protéomique ascendante (bottom-up) pour identifier les protéines dans des mélanges complexes et à grande échelle. Elle utilise une combinaison de chromatographie liquide à haute performance à la spectrométrie de masse (LC-MS/MS) (Aebersold and Mann, 2003; Lin et al., 2003; Nesvizhskii, 2007; Zhang et al., 2006). La spectrométrie de masse permet de détecter et mesurer l'abondance des molécules analysées. Dans cette partie, nous allons traiter des aspects quantitatifs de la protéomique shotgun.

La démarche générale de l'analyse protéomique comprend principalement quatre étapes : (i) la préparation des échantillons, (ii) la séparation des protéines ou des peptides digérés, (iii) l'interprétation des spectres, les mesures d'abondance et l'identification des protéines par MS (top-down ou bottom-up), et (iv) le data mining (*i.e.* exploration des données) de la très grande quantité d'informations générées (*e.g.* annotation fonctionnelle ou analyse comparative) (Figure 17). Les trois premières étapes sont étroitement liées aux techniques analytiques (Hu et al., 2007). La dernière étape relève de l'analyse bioinformatique des données (Figure 17). Ainsi, l'utilisation de statistiques d'échantillonnage semble être une approche simple et pratique pour mesurer l'abondance relative des protéines (Florens et al., 2002; Liu et al., 2004). Par exemple, le spectral count ou le nombre total de spectres MS/MS acquis sur les peptides d'une protéine donnée est corrélé linéairement avec l'abondance de la protéine (Liu et al., 2004).

Pour permettre l'identification des spectres, les bases de données protéiques universelles issues des données de la littérature telles que celle du NCBI

(https://www.ncbi.nlm.nih.gov/protein), UniprotKB/Swiss-Prot (The UniProt Consortium, 2009) ou encore neXtProt (Lane et al., 2012) sont utilisées (Dumas, 2020). Cette approche consiste à faire correspondre les spectres analysés à des spectres théoriques identifiés, dans une bibliothèque de peptides associés aux protéines (Granholm, 2014). Swiss-Prot est une base de données où les informations sont vérifiées manuellement et leur fiabilité est renforcée par les données expérimentales de la communauté scientifique (Dumas, 2020). Swiss-Prot a pour objectif de mettre à disposition des séquences protéiques (i) en limitant la redondance, (ii) en fournissant des annotations comme la fonction de la protéine, les modifications post-traductionnelles associées, les protéines similaires, les domaines fonctionnels, la structure, les voies dans lesquelles elles sont impliquées ou encore leur localisation cellulaire, (iii) et ayant des liens avec d'autres banques de données.

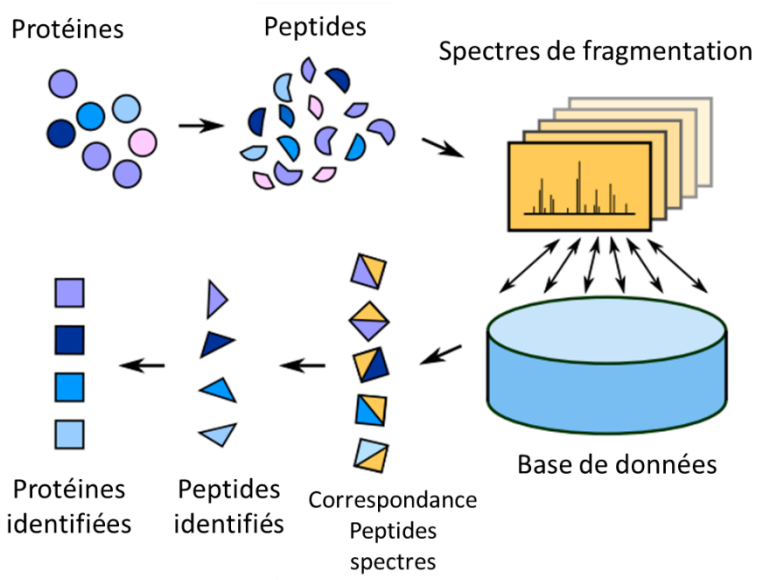


Figure 17 : Schéma du principe de la protéomique shotgun. À partir d'un échantillon de protéines, celles-ci sont digérées en peptides grâce à une protéase et les spectres des fragments sont obtenus après l'analyse LC-MS/MS. Du côté bioinformatique, les spectres sont appariés avec des peptides dans une base de données protéiques. Les peptides identifiés à l'aide des spectres permettent d'identifier les peptides et les protéines de l'échantillon de départ. (adaptée de (Granholm, 2014).

La Figure 18 présente les principales étapes de l'analyse biostatistique et bioinformatique de données de protéomique shotgun et sont détaillées ci-après (Sinitcyn et al., 2018; Tyanova et al.,

2016). Une fois que les protéines ont été identifiées et quantifiées sur de nombreux échantillons, une matrice de valeurs representant les abondances de protéines ou des rapports d'abondance de protéines (ou des groupes de protéines) pour chaque échantillon est construite pour permettre les analyses quantitatives (Figure 18-1) (Sinitcyn et al., 2018). Certaines étapes préparatoires précèdent la plupart des analyses, telles que la normalisation des intensités ou des ratios, le filtrage des données et/ou l'imputation potentielle des valeurs manquantes (Figure 18-2). Une tâche courante dans la protéomique de découverte consiste à identifier les protéines modulées (*e.g.* par l'exposition à des contaminants) et à les distinguer du reste du protéome, notamment grâce à l'analyse différentielle (Figure 18-3A). L'analyse différentielle est possible à l'aide des outils statistiques et de visualisation utilisés en transcriptomique (voir partie 1.3) (Figure 18-3B, Figure 18-4A). L'analyse en composantes principales (ACP) est une méthode alternative de visualisation des effets principaux selon les conditions expérimentales et de la relation inter et intra échantillons (Figure 18-3B).

Lorsqu'un groupe intéressant de protéines a été identifié (*e.g.* par des tests statistiques, une classification hiérarchique ou une ACP), une analyse d'enrichissement peut être effectuée pour trouver des processus biologiques, des complexes ou des voies communes à ces protéines, ce qui est semblable à l'analyse transcriptomique (Figure 18-4B) (voir partie 1.3). Dans ce but, des sources d'annotation telles que l'ontologie des gènes (Gene Ontology, 2015), les appartenances à des voies avec des bases de données comme Reactome (Croft et al., 2011; Fabregat et al., 2018) ou des complexes protéiques (Ruepp et al., 2007) sont nécessaires. Les processus biologiques étudiés présentent souvent des changements temporels, avec des protéines qui suivent un pattern attendu, par exemple des changements périodiques dans le cycle cellulaire (Figure 18-4C) (Sinitcyn et al., 2018). D'autres études impliquent de mesurer une réponse aux changements de dose de

stimuli. Dans ces situations, des méthodes peuvent être appliquées pour détecter les changements de concentrations suivant un modèle donné. Pour ce cas de changements temporels périodiques, l'analyse attribuera une amplitude de variation et un temps de pointe pour chaque protéine (Robles et al., 2014).

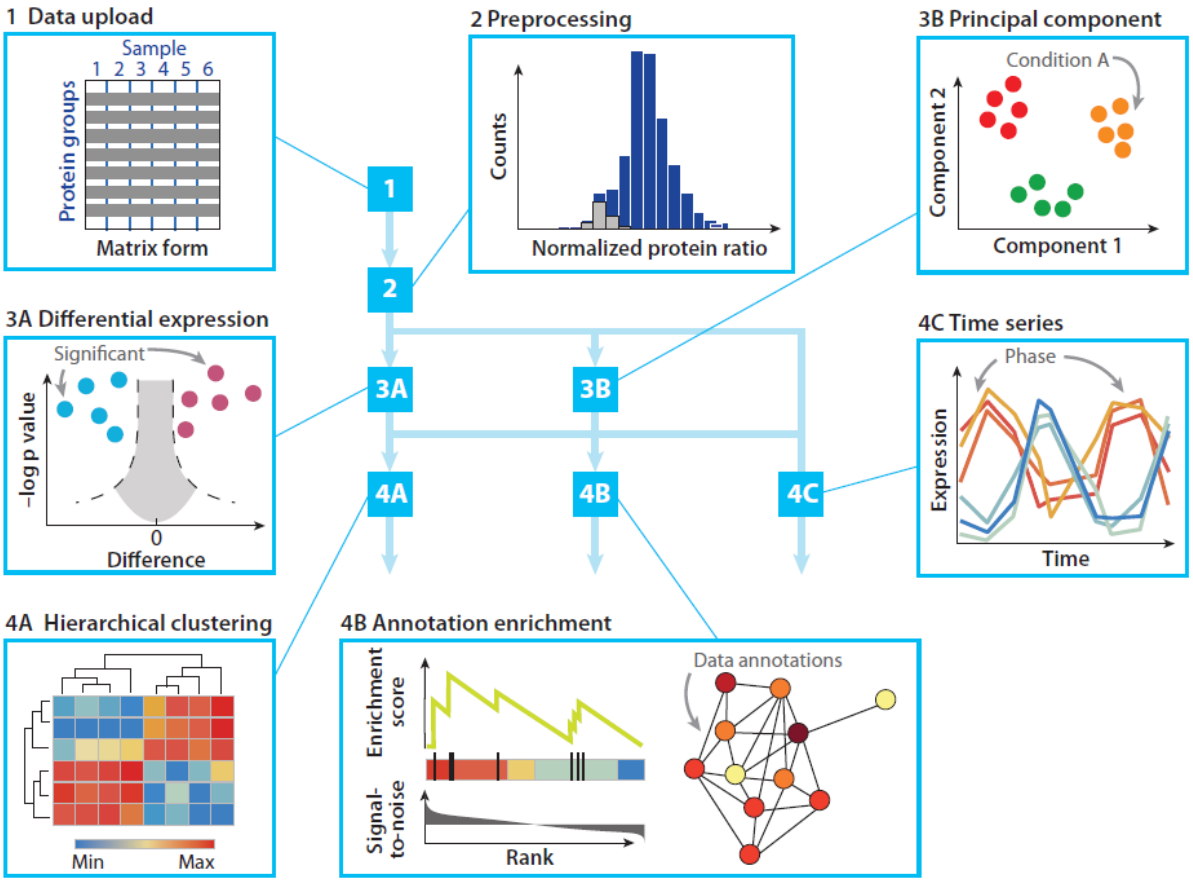


Figure 18 : Principe de l'analyse de données en protéomique shotgun. 1- Téléchargement des données. 2- prétraitement des données. 3A- Expression différentielle. 3B- Analyse en composantes principales (ACP). 4A- Classification hiérarchique. 4B- Analyse d'enrichissement des annotations. 4C- Analyses des séries temporelles. (adaptée de Sinitcyn et al., 2018; Tyanova et al., 2016)

La protéogénomique est définie comme l'amélioration des annotations structurales des génomes grâce au couplage des méthodes d'annotation des génomes (top-down) et/ou la transcriptomique aux méthodes de protéomique shotgun (bottom-up) (Armengaud et al., 2014a; Jaffe et al., 2004; Nesvizhskii, 2014). En effet, la protéogénomique permet (i) de valider la prédiction des gènes, (ii) d'identifier des gènes non prédits, (iii) de corriger des sites d'initiation et

de terminaison de la transcription ainsi que des ORFs, (iv) et d' identifier les sites de modifications post-traductionnelles (Ansong et al., 2008) (Figure 19). Il existe des outils capables de localiser les peptides identifiés sur le génome et ensuite de modifier cette annotation pour permettre la correction par la protéogénomique, tels que 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). Il existe néanmoins une limitation technique dans la correction des annotations, qui est la différence de format des données de protéomiques et de transcriptomiques (Cogne, 2019). Deux nouveaux formats ont donc été proposés, le format proSAM et le format proBAM (Menschaert and Fenyö, 2017; X. Wang et al., 2016). Ces formats uniformisent les sorties classiques (SAM ou BAM) des logiciels de transcriptomiques pour mieux les intégrer aux sorties de protéogénomiques pour l'annotation (Cogne, 2019).

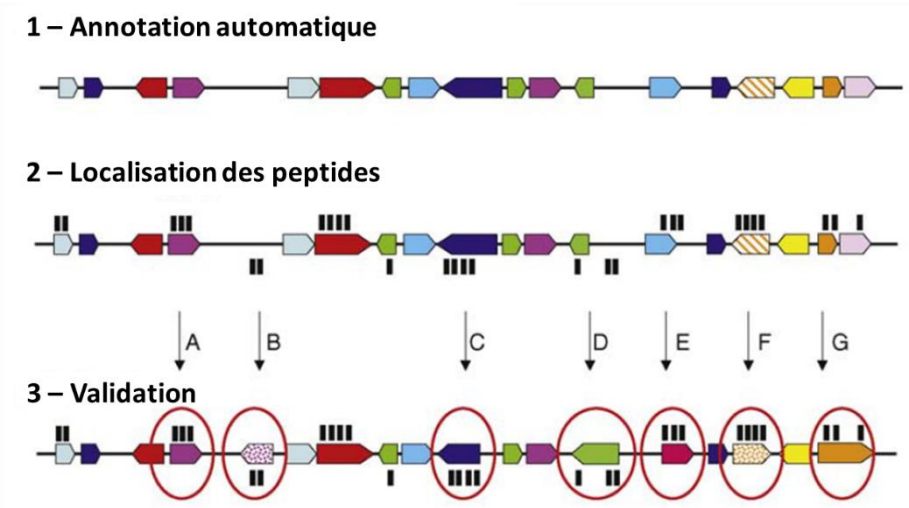


Figure 19 : Validation et correction de l'annotation grâce aux données de MS. La stratégie comporte les étapes suivantes : 1 – annotation automatique du génome, 2 – localisation des peptides sur le génome, 3 – validation manuelle des erreurs d'annotation. Les différents peptides détectés par spectrométrie de masse appariés à la séquence d'acide nucléique sont indiqués par des rectangles noirs soit en haut (cadre de lecture direct) soit en bas (cadre de lecture inverse) de la séquence. A) L'existence d'un gène prédit est validée par la présence de trois peptides différents. B) L'existence d'un gène non annoté est confirmée. Le codon d'initiation du gène est situé en aval (C) ou en amont (D) du site précédemment annoté. E) Trois peptides détectés indiquent que le gène doit être annoté sur un autre cadre de lecture. F) Quatre peptides détectés montrent que le gène doit être réorienté comme le gène est transcrit à partir de l'autre brin. G) Plusieurs peptides provenant de la même protéine révèlent la présence d'un décalage du cadre de lecture ou d'un codon stop mal placé. (adaptée de Armengaud, 2009).

La protéogénomique est une alternative efficace et rapide pour obtenir un large catalogue de protéines chez des espèces pour lesquelles le séquençage du génome n'est pas encore accessible. Effectivement, pour les organismes non modèles, les bases de données protéomiques universelles sont composées de peptides homologues qui proviennent d'organismes apparentés et ne permettent donc pas l'identification des protéines spécifiques de l'espèce étudiée (Dumas, 2020). Pour pallier cette limite, la protéogénomique utilise le séquençage des ARNm de l'espèce d'intérêt pour créer une base de données protéiques théorique générée à partir de la traduction *in silico* dans les 6 cadres de lectures des séquences d'ARNm (transcrits) : les ORFs (Figure 20). Cette base de données théorique et spécifique de l'organisme étudié est ensuite utilisée pour l'identification des spectres MS/MS de la protéomique shotgun, qui ne seraient pas présents dans les bases de données protéomiques universelles (Figure 20). Il existe de nombreux outils bioinformatiques adaptés à l'analyse protéogénomique, de la création de bases de données jusqu'à l'interprétation des peptides (Menschaert and Fenyö, 2017). Ces différents outils ont été recensés et classés dans la revue de Menschaert et Fenyö (2017), tels que PROTEOFORMER (Crappé et al., 2015) pour la génération de base de données, ENOSI (Woo et al., 2014) ou Peppy (Risk et al., 2013) pour la recherche automatique, ou encore le workflow Peptimapper (Guillot et al., 2019) pour la correction de l'annotation.

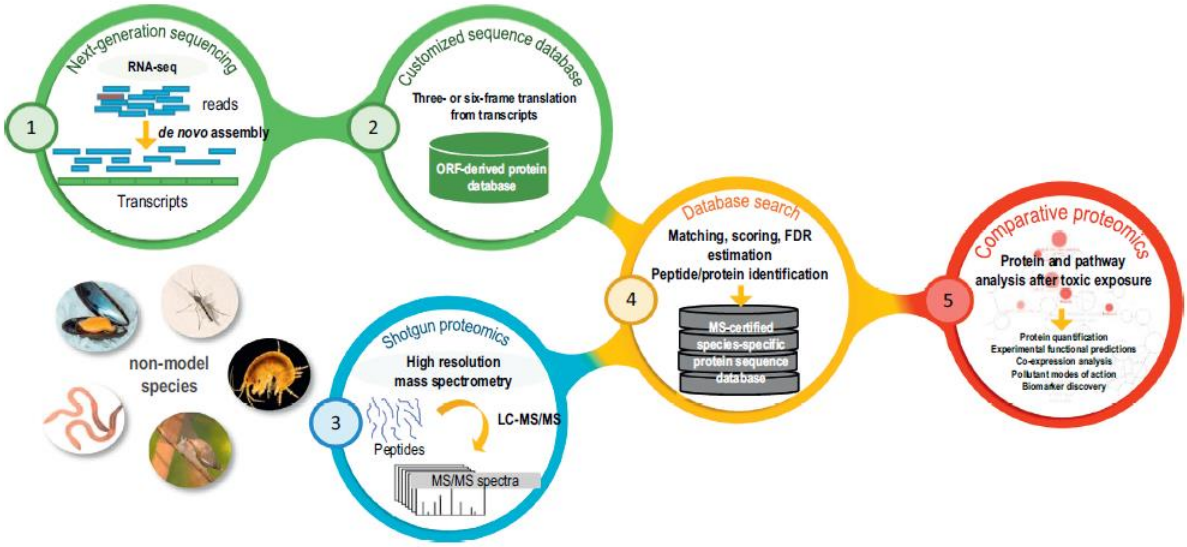


Figure 20 : Stratégie de protéogénomique chez les espèces non modèles (Gouveia et al., 2019a)

La protéogénomique est donc devenue une méthode de référence pour l'identification des protéines chez les organismes non modèles (Armengaud et al., 2014a) et a été appliquée avec succès à plusieurs espèces (Gouveia et al., 2019b). Cette approche a permis de découvrir, par exemple, des séquences protéiques spécifiques au nématode *Heligosomoides polygyrus* (Moreno et al., 2011), la tomate domestiquée *Solanum lycopersicum* (Lopez-Casado et al., 2012), l'amphipode crustacé *Gammarus fossarum* (Trapp et al., 2014b), le bivalve *Mytilus edulis* (Campos et al., 2016), ou la fougère apogame *Dryopteirs affinis* (Grossmann et al., 2017). Toutefois, l'attribution des fonctions reste difficile (Trapp et al., 2014a). Les approches d'annotations fonctionnelles actuelles sont basées sur la similarité des séquences ou des corrélations fonctionnelles. Ces annotations sont majoritairement réalisées à partir de la physiologie ou de la biologie des espèces pour lesquelles il existe des données disponibles (Duarte Gouveia et al., 2017; Gouveia et al., 2019b; Trapp et al., 2014b). Pour étudier le lien entre une protéine et sa fonction physiologique ou son rôle comme biomarqueur d'une exposition ou d'une maladie, il est nécessaire que son abondance soit mesurée. En effet, la mesure de l'expression des protéines (*i.e.* abondance) permet l'étude de leur dynamique dans des conditions données, telles que l'exposition à des polluants ou des variations physiologiques (Lepretre, 2019). L'estimation de l'abondance à grande échelle de protéomes

entiers est la technique de choix en protéomique clinique pour la phase initiale de découverte de biomarqueurs, car elle permet l'identification robuste des profils d'expression de protéines modulées en réponse à un stimulus (Gouveia et al., 2019b). En sciences environnementales, l'analyse shotgun a été appliquée pour obtenir une mesure de l'abondance relative des protéines exprimées de manière différentielle après des expositions à des substances toxiques (Campos et al., 2016; Ralston-Hooper et al., 2013; Trapp et al., 2015). Ces informations peuvent être utilisées pour élucider les modes d'action moléculaires (MoA) des contaminants et/ou pour développer des méthodes sensibles pour la quantification des biomarqueurs (Armengaud et al., 2014a).

1.5 Les omiques pour le métabolisme : la métabolomique et la lipidomique

1.5.1 Métabolomique

La métabolomique est une omique apparue depuis la fin des années 1990, qui a connu une croissance rapide au cours de la dernière décennie (Patti et al., 2012). Elle vise à étudier l'ensemble des métabolites dans les échantillons biologiques (*e.g.* les cellules, les tissus, les fluides corporels, les organismes entiers) que l'on appelle le métabolome (Wishart, 2007), correspondant à des composés de faible poids moléculaire (inférieur à 2000 Dalton (Da)) (Laudicella et al., 2020). Chez la levure *Saccharomyces cerevisiae*, le métabolome dénombre environ 600 métabolites (Förster et al., 2003). Chez l'Homme, on compte jusqu'à 40 000 métabolites (Wishart et al., 2013). Pour finir, chez les végétaux on compte jusqu'à 200 000 métabolites (Fiehn, 2002). L'étude du métabolome d'une espèce reste donc complexe, car un grand nombre de métabolites, ainsi que leurs voies de métabolisation et leurs rôles biochimiques restent spécifiques d'une espèce (Dumas, 2020).

Le métabolome englobe principalement « l'endo-métabolome » qui correspond aux produits naturels du métabolisme, et le « xéno-métabolome » (ou appelé exposome) qui correspond aux substances chimiques exogènes (*i.e.* xénobiotiques) qui peuvent être retrouvées et

biotransformées dans l'organisme (Dumas, 2020). Deux grandes classes de métabolites composent l'endo-métabolome : ceux provenant du métabolisme primaire, communs à plusieurs espèces et essentiels au développement, et ceux du métabolisme secondaire, plus spécifiques de l'espèce donnée. Le métabolome est un groupe de molécules ayant une grande diversité, tels que les lipides, les acides aminés, les peptides, les acides nucléiques, les acides organiques, les vitamines et les thiols (Shi, 2017), présentant des propriétés physico-chimiques, des concentrations et poids moléculaires très hétérogènes (Dumas, 2020; Dumas et al., 2022; Dunn and Ellis, 2005). Étant donné que les métabolites sont le produit final de l'expression génique et de l'activité enzymatique et cellulaire, ils sont très sensibles aux changements environnementaux. Ces différences de métabolites peuvent être utilisées pour identifier des biomarqueurs et des voies métaboliques particulières indicatives de chaque condition spécifique d'exposition.

La métabolomique utilise diverses technologies à haut débit, telles que la spectroscopie infrarouge (IR), la spectroscopie Raman (RAMAN), la résonance magnétique nucléaire (RMN). Elle utilise également de nombreuses techniques de spectrométrie de masse (MS), y compris la spectrométrie de masse à infusion directe (DI-MS), la matrice- spectrométrie de masse assistée par désorption/ionisation laser (MALDI-IMS), l'électrophorèse capillaire-spectrométrie de masse (CE-MS), la chromatographie en phase gazeuse spectrométrie de masse (GC-MS) et chromatographie liquide-spectrométrie de masse (LC-MS). Avec des capacités de débit et de résolution suffisamment élevées, la RMN et la MS sont les outils analytiques les plus largement appliqués (Alfaro and Young, 2018; Nguyen et al., 2019).

Stratégies d'analyses

La Figure 21 décrit les grandes étapes nécessaires à une étude métabolomique, de la mise en place du design expérimental à l'interprétation des données en passant par l'analyse statistique.

Les étapes pour traduire des données métabolomiques en informations biologiquement significatives impliquent : le prétraitement des données (traitement des spectres, normalisation des données), analyses statistiques en aval, l'identification des métabolites, et enfin la visualisation et l'interprétation des données (Figure 21). Dans cette partie, nous nous focaliserons et synthétiserons uniquement les étapes à partir du prétraitement des données (Figure 21). La métabolomique non ciblée fournit une vue holistique des petites molécules dans l'échantillon biologique et a un grand potentiel pour des études de découverte pour générer de nouvelles hypothèses. C'est une analyse globale de « tous » les métabolites mesurables dans un échantillon biologique, y compris des métabolites inconnus dans le but de générer des profils ou empreintes (Gorrochategui et al., 2016). En effet, les empreintes obtenues peuvent être comparées pour permettre de souligner les différences ou similarités entre des groupes d'échantillons (témoins versus tests) (Dumas, 2020; Nicholson et al., 1999). La métabolomique non ciblée se traduit souvent par une quantité massive de données. L'identification des métabolites est indispensable pour conférer une signification biologique (Alonso et al., 2015; Dunn et al., 2013; Gorrochategui et al., 2016), nécessitant des outils chimio-informatiques (*i.e.* biostatistiques et informatiques) et bioinformatiques sophistiqués de l'acquisition des données, aux analyses statistiques en passant par le traitement des données (Gorrochategui et al., 2016; Schrimpe-Rutledge et al., 2016).

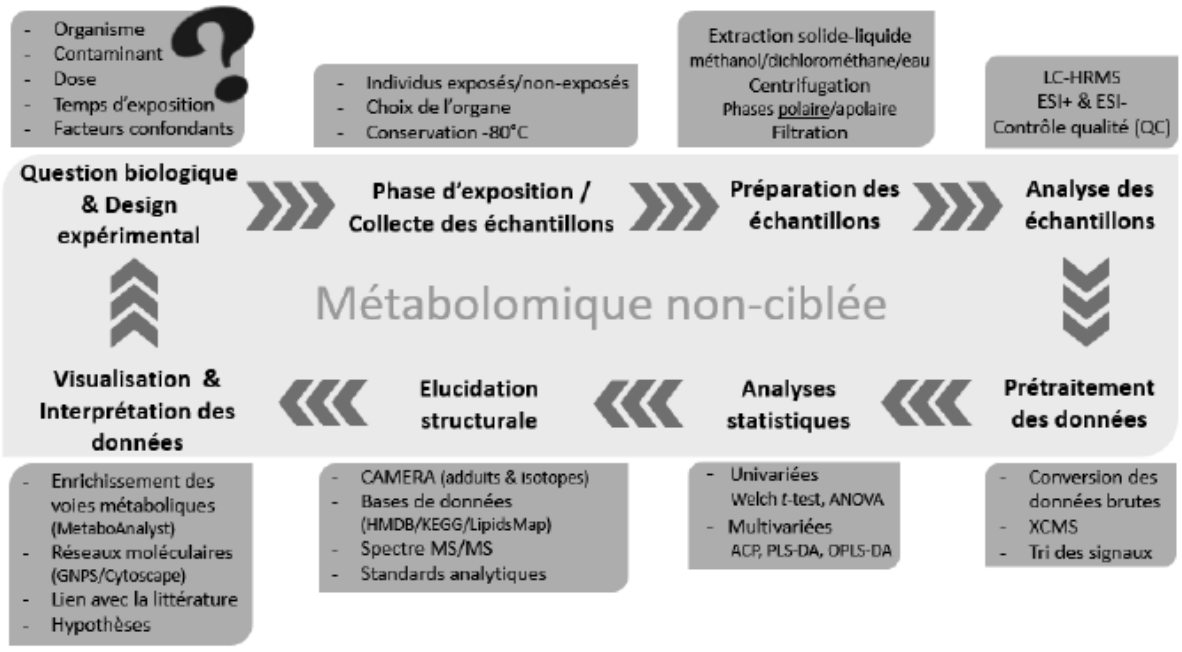


Figure 21 : Démarche de l'approche de la métabolomique non ciblée (Dumas, 2020).

Les méthodes de métabolomique non ciblée par spectrométrie de masse génèrent des données caractérisées par trois dimensions : le temps de rétention (RT), le rapport masse sur charge (m/z) du composé et son intensité (représentant son abondance). Il est par la suite nécessaire de transformer les données brutes à la sortie des instruments en une matrice contenant les métabolites pour chaque échantillon (Figure 21). Ces données transformées pourront être exploitées pour révéler l'information d'intérêt biologique grâce à des analyses statistiques (*i.e.* approches univariées et multivariées). Il existe des méthodes statistiques non supervisées (*e.g.* l'ACP) pour obtenir une vision sans *a priori* globale. Il est recommandé de combiner les approches univariées et multivariées pour maximiser l'extraction d'informations pertinentes provenant des données (Vinaixa et al., 2012).

Il est ensuite nécessaire de procéder à l'élucidation structurale chimique des signaux pour identifier les métabolites (Figure 21). Cette étape reste un des plus grands défis pour la métabolomique non ciblée (Dumas, 2020), et a pour but d'associer un rapport m/z et un RT à une structure chimique précise et donc à un métabolite. De manière générale, on estime à moins de

20% le nombre de signaux d'intérêt annotés (Blaženović et al., 2018). L'identification de nouveaux métabolites nécessite la combinaison de méthodes analytiques telles que la RMN, la MS/MS ou encore le marquage isotopique, ainsi que d'outils de prédiction et de calculs pour les spectres d'ions fragmentés (Rathahao-Paris et al., 2015). Il existe également la construction de réseaux moléculaires basés sur les similarités spectrales (issus de MS/MS). Par exemple, Global Natural Product Social Molecular Networking (GNPS) est une plateforme en ligne qui permet le partage de données de spectres pour reconstruire ces réseaux moléculaires (M. Wang et al., 2016). Récemment, de nouveaux pipelines intégrant des outils bioinformatiques pour les tâches d'identification de métabolites ont vu le jour et permettent une identification fiable dans des mélanges (Benton et al., 2015; Bingol et al., 2016; Dührkop et al., 2015; Kouassi Nzoughet et al., 2017; van der Hooft et al., 2016).

Le contexte biologique d'un métabolite est la voie métabolique, c'est-à-dire une succession de réactions chimiques dépendantes les unes des autres. Chaque réaction est catalysée par une enzyme spécifique et forme une étape de synthétisation ou dégradation d'une molécule. Le maintien de l'homéostasie cellulaire et les besoins fonctionnels de l'organisme sont assurés par la régulation des voies métaboliques (Dumas, 2020). Replacer les métabolites dans leur contexte biologique et dans les voies métaboliques est une approche qui permet de déterminer les voies métaboliques les plus impactées selon un stress.

Au cours des dernières années, les connaissances biologiques disponibles pour les études de métabolomique n'ont cessé d'augmenter. Les bases de données biologiques telles que KEGG (Kanehisa and Goto, 2000), sur les voies des petites molécules ou « small molecule pathway database » (SMPDB ; (Jewison et al., 2014)), EHMN (Ma et al., 2007), WikiPathways (Slenter et al., 2018)) et MetaCyc (Caspi et al., 2008) fournissent de nouvelles informations de plus en plus précises

sur un grand nombre de voies métaboliques (Tableau 5). La disponibilité de ces données permet donc l'utilisation de méthodes d'enrichissement des ensembles de métabolites (*i.e.* MSEA pour Metabolite Set Enrichment Analysis), basées sur les mêmes principes que les méthodes pour les données d'expression génique, comme détaillée précédemment dans ce manuscrit (voir partie 1.3) (Khatri et al., 2012). Les chercheurs disposent actuellement d'une grande variété de logiciels pour analyser les voies métabolomiques. Des outils comme, Impala (Kamburov et al., 2011), MetScape2 (Karnovsky et al., 2012) ou Metaboanalyst (Chong et al., 2019; Xia et al., 2012) sont des logiciels qui implémentent également des méthodes MSEA spécifiques. En complément, des applications de visualisation, telles que Paintomics (García-Alcalde et al., 2011), Vanted (Rohn et al., 2012) et Cytoscape (Smoot et al., 2011) fournissent différents outils de visualisation des voies. À l'aide de ces outils, les métabolites sont cartographiés sur des voies métaboliques prédéfinies, et permettent un haut niveau d'interaction avec les données.

Tableau 5 : Base de données de métabolites et de voies métaboliques (adapté de (Alonso et al., 2015)).

Base de données	Description	Site Web	Référence
Kyoto Encyclopedia of Genes and Genomes (KEGG)	551 voies, 18919 métabolites et 11774 réactions biochimiques	http://www.genome.jp/kegg	(Kanehisa et al., 2021)
MetaCyc	2749 voies provenant de 3045 organismes différents	http://metacyc.org	(Caspi et al., 2020)
The small molecule pathway database (SMPDB)	1594 métabolites mappant 727 voies de petites molécules trouvées chez les humains	http://www.smpdb.ca	(Jewison et al., 2014)
WikiPathways	2857 voies	http://wikipathways.org	(Martens et al., 2021)
Plant metabolic network (PMN/PlantCyc)	Base de données de voies multiespèces pour la métabolomique des plantes	http://www.plantcyc.org	(Hawkins et al., 2021)

Limites des approches de métabolomiques

Les bases de données de métabolites restent encore limitées, rendant leur identification complexe, ce qui représente un obstacle pour l'interprétation biologique. De plus, le concept de voie métabolique présente actuellement certaines limites (Barupal et al., 2018; Dumas, 2020) :

(i) La définition d'une voie métabolique peut varier d'une base de données à l'autre, celle-ci étant faite manuellement.

(ii) Un métabolite peut se retrouver impliqué dans différentes voies métaboliques. Cette redondance peut parfois biaiser l'interprétation et les analyses d'enrichissement des voies.

Une solution proposée est l'organisation des métabolites en classe chimique en fonction de leur similarité structurale (*i.e.* Chemical Ontology) (Barupal et al., 2018; Moreno et al., 2015; Tsuyuzaki et al., 2015). Cette approche a été développée sous le nom de ChemRICH (Chemical Similarity Enrichment Analysis, http://chemrich.fiehnlab.ucdavis.edu/) (Barupal and Fiehn, 2017) et propose :

(i) S'il existe des métabolites qui ne sont pas toujours associés à une voie métabolique ou une réaction enzymatique, ils pourraient tout de même être associés à une classe chimique, grâce à l'approche ChemRICH.

(ii) Pour limiter la redondance des métabolites et ainsi les biais d'enrichissement, il serait possible d'affecter une seule classe chimique à un seul métabolite (plutôt que plusieurs voies métaboliques). Les analyses d'enrichissement permettraient d'indiquer les classes de métabolites les plus impactées par l'effet donné, plutôt que les voies métaboliques.

Pour conclure, la métabolomique reste relativement abordable (environ 30 à 40 USD par échantillon) par rapport à la transcriptomique ou la protéomique (environ 200 à 500 USD par échantillon). La métabolomique est donc une omique intéressante, car elle permet des études sans protéome ou protéogénome, et ce malgré la comparaison des voies métaboliques entre différentes espèces qui reste problématique, et l'incertitude sur le concept de voie métabolique.

1.5.2 Lipidomique

Les lipides constituent un sous-ensemble du métabolome (Figure 22) (Ayciriex, 2010). On observe une augmentation de la littérature concernant l'analyse des lipides en raison de leur rôle central dans le métabolisme énergétique, la structure cellulaire, la signalisation et la fonction des protéines ainsi que les progrès technologiques (Kyle et al., 2021; O'Donnell et al., 2020). Les termes « lipidome » et « lipidomique » ont été introduits pour la première fois par Kishimoto et al. (2001), puis défini par Han et Gross (2003). La lipidomique est l'étude à grande échelle de la structure et de la fonction de tous les lipides dans un échantillon, appelée lipidome (Han and Gross, 2003; Laudicella et al., 2020; Navas-Iglesias et al., 2009).

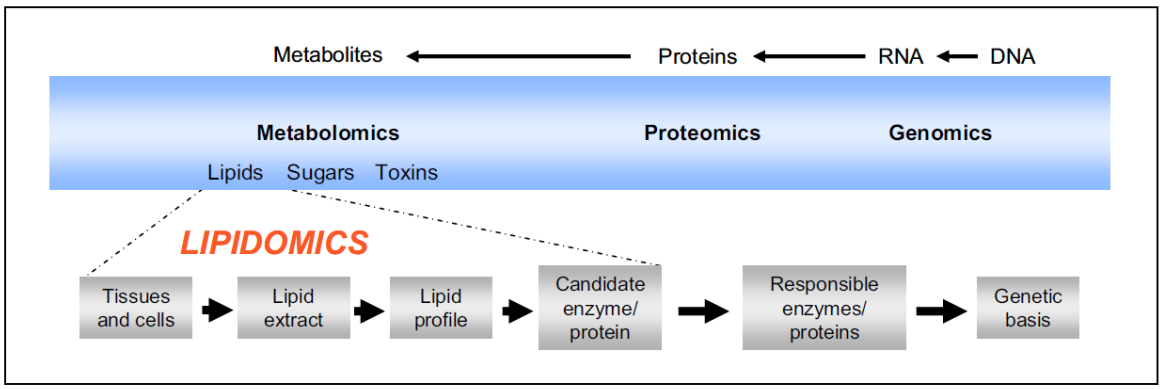


Figure 22 : Échelles métabolomiques incluant la génomique, la protéomique et la lipidomique (Ayciriex, 2010; Wenk, 2010)

Les études utilisant la lipidomique ont doublé au cours des cinq dernières années, les progrès récents en spectrométrie de masse (MS) ainsi que les développements bioinformatiques permettent aux chercheurs d'identifier en routine des centaines de lipides uniques (Aimo et al.,

2015; Fahy et al., 2019, 2009; Kyle et al., 2021; Ni et al., 2017a, 2017b). Cependant, comprendre la pertinence biologique des lipides et leurs effets sur la fonction d'autres biomolécules (*e.g.* les protéines) est un grand défi dans le domaine de la lipidomique, compte tenu de sa diversité et de sa complexité (Harayama and Riezman, 2018).

Les lipides sont classés en différentes catégories, chacune avec sa propre hiérarchie de classes et de sous-classes basée sur une structure de squelette commune, et peuvent comprendre une variété de chaînes acyles. Leurs propriétés physico-chimiques dépendent de ces paramètres, qui déterminent également le rôle qu'ils jouent dans les systèmes biologiques. La nomenclature des lipides a évolué au cours des dernières années pour devenir plus simple et plus représentative de leur nature chimique. The International Lipid Classification and National Nomenclature Committee appelé LIPID MAPS (http://www.lipidmaps.org) (Fahy et al., 2009) classe les lipides en huit catégories. Ces classes sont ensuite divisées par leurs propriétés structurelles et chimiques plus spécifiques. Il existe actuellement plus de 40 000 structures lipidiques dans cette base de données. La complexité du lipidome est souvent sous-décrite par les pipelines lipidomiques actuels, car la plupart des lipides ne peuvent pas être entièrement décris structurellement et représentent des espèces isobares agrégées (*i.e.* ayant une même masse moléculaire, mais une structure différente). Par exemple, l'incorporation de données quantitatives pour les glycérolipides complexes individuels ou les phospholipides est impossible dans la plupart des étapes d'analyse, car seules les valeurs moyennes de l'ensemble de la sous-classe de lipides peuvent être représentées, et non les valeurs pour les lipides individuels. Actuellement, les analyses basées sur les classes peuvent identifier en toute confiance quelques centaines de lipides par rapport aux milliers potentiellement présents (Kyle et al., 2021).

Une revue des logiciels d'identification lipidomiques actuels est disponible (Kyle et al., 2021; Misra and Mohapatra, 2019; O'Shea and Misra, 2020), y compris sur le site web de LIPID MAPS (https://www.lipidmaps.org/resources/tools). Cependant, l'analyse automatisée puis l'identification des données MS et MS/MS restent souvent un goulot d'étranglement en lipidomique. Ces dernières années, plusieurs approches ont permis l'amélioration de l'identification des lipides, comme par exemple le logiciel LipidHunter (https://github.com/SysMedOs/lipidhunter) (Ni et al., 2017a, 2017b). En utilisant une approche ascendante (bottum-up), LipidHunter ne nécessite pas de bases de données structurelles ou spectrales et peut être facilement ajusté pour s'adapter aux conditions analytiques spécifiques à l'utilisateur. Avec l'augmentation de la sensibilité et de la résolution des plateformes MS, la complexité et la diversité des lipidomes naturels peuvent désormais être étudiées de manière beaucoup plus détaillée. Comme pour l'ADN (épigénétique) et les protéines (modifications post-traductionnelles, PTM), il existe des modifications des lipides par des réactions enzymatiques et non enzymatiques, notamment l'oxydation, la nitration, la sulfatation et l'halogénation. Ces modifications sont nécessaires pour réguler une fonction biologique complexe, créant ainsi un nouveau niveau de complexité du lipidome (*i.e.* l'épilipidome) (Ni et al., 2019), pour lequel il existe des outils spécialement développés pour identifier les modifications des lipides tels que le logiciel LPPtiger (Ni et al., 2017a, 2017b).

Les bases de données consultables dans le domaine de la lipidomique sont des ressources relativement rares. En 2003, le comité LIPID MAPS a été formé, ainsi qu'une base de données associée et organisée sur la structure lipidique des espèces moléculaires individuelles. SwissLipids (www.swisslipids.org) est une ressource de connaissances dédiée aux lipides et à leur biologie (Aimo et al., 2015). Cette base de connaissances comprend une bibliothèque avec plus de 590 000 structures lipidiques connues et théoriques appartenant à plus de 500 classes de lipides, chacune

enrichie d'informations sur les composants lipidiques, les réactions et les enzymes, avec des liens vers la littérature. Tous les lipides sont cartographiés dans la base de données d'ontologie chimique ChEBI (www.ebi.ac.uk/chebi/) (Hastings et al., 2016), qui est utilisée pour décrire le métabolisme des lipides dans la base de données Rhea sur les réactions biochimiques (www.rhea-db.org) (Morgat et al., 2020), et liées aux enzymes correspondantes présentes dans UniProtKB (www.uniprot.org) (The UniProt Consortium, 2021). Actuellement, LIPID MAPS et SwissLipids, utilisent la même nomenclature simplifiée pour décrire les lipides. Dans la majorité des cas, les noms des lipides contiennent à eux seuls les détails essentiels pour déterminer la nature chimique des lipides (*e.g.* "nom commun" dans LIPID MAPS et "abréviation" dans SwissLipids). En utilisant les informations fournies par les noms communs des lipides, deux outils d'enrichissement lipidomique, Lipid Mini-On (Clair et al., 2019) et LION/web (Molenaar et al., 2019), ont récemment été développés pour aider les utilisateurs dans l'interprétation biologique des altérations du lipidome.

A ce jour, l'intégration des données lipidomiques progresse encore lentement, principalement dû aux difficultés d'identification de la structure exacte des lipides ainsi qu'au manque de connaissances biologiques dans les bases de données. Plusieurs approches de calcul ont toutefois été développées pour relier le mapping des voies lipidomiques et l'intégration d'autres données omiques (Lam et al., 2021). Une de ces approches est la base de données WikiPathways (O'Donnell et al., 2019; Slenter et al., 2018), qui à partir de neuf voies hautement conservées chez la souris a permis de retrouver les homologues chez l'Homme pour alimenter et mettre à jour la base de données avec des fonctions jusque-là inconnues (Dennis et al., 2010; Dennis and Norris, 2015; Kyle et al., 2021). Les données WikiPathways pour les lipides sont accessibles pour l'analyse des voies et

l'intégration omique, par l'interface graphique et l'automatisation dans PathVisio (Ellis et al., 2018; Kutmon et al., 2014) et Cytoscape (Shannon et al., 2003).

2. LES APPROCHES BIOINFORMATIQUES POUR LA COMPREHENSION DES MECANISMES D'ACTION ET DES VOIES MOLECULAIRES EN ECOTOXICOLOGIE

2.1 L'apport des omiques en écotoxicologie

2.1.1 Définition de l'écotoxicologie

L'écotoxicologie est une discipline à l'interface entre l'écologie et la toxicologie qui s'est consolidée à partir des années 1970, à la suite des pollutions survenues après la Seconde Guerre mondiale (Jouany, 1971; Truhaut, 1977, (Butler, 1978; Kendall et al., 2001; Moriarty, 1983; Newman and Zhao, 2008). Ces événements ont permis la prise de conscience de l'impact négatif des substances chimiques et des rejets toxiques sur l'environnement et sur l'Homme. La toxicologie a pour objectif d'évaluer le danger et le risque liés à l'exposition à un agent chimique, physique ou biologique. L'écotoxicologie étudie le comportement et les effets des polluants sur les écosystèmes, qu'il s'agisse d'agents d'origine artificielle ou naturelle dont l'Homme modifie la répartition et/ou les cycles dans les différents compartiments de la biosphère, ainsi que le devenir de ces polluants. En effet, la toxicologie limite ses études aux organismes alors que l'écotoxicologie étudie l'impact des contaminants sur les individus ainsi que sur les populations, les écosystèmes et la dynamique associée (Figure 23).

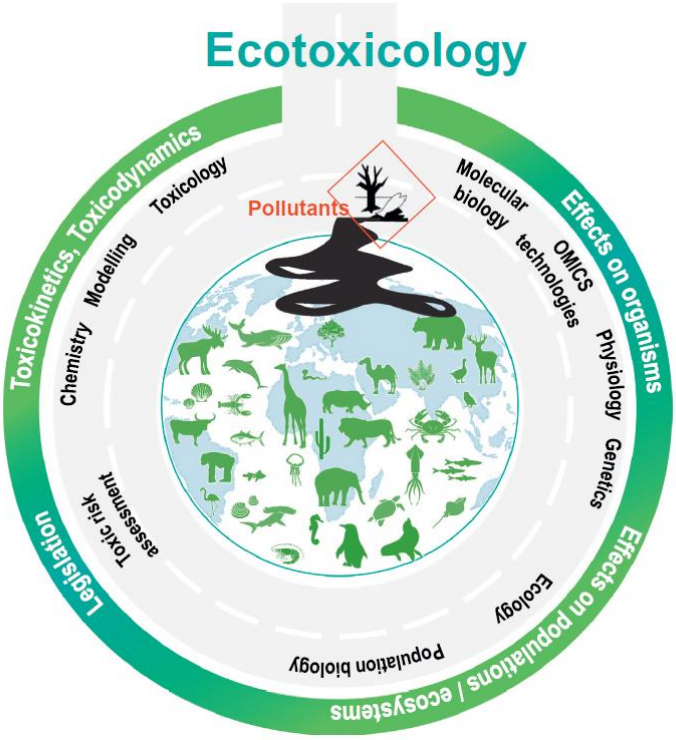


Figure 23 : L'écotoxicologie ou l'étude des effets des produits chimiques toxiques sur les organismes biologiques et les écosystèmes, à la croisée de plusieurs disciplines (Gouveia et al., 2019b).

2.1.2 *L'ère des omiques en écotoxicologie*

Le changement d'échelle induit par les technologies émergentes, plus particulièrement le séquençage de nouvelle génération (NGS), la spectrométrie de masse haute résolution et la bioinformatique, a révolutionné l'exploration des systèmes biologiques. Ces technologies offrent une perspective descendante (top-down) relativement non biaisée et ont changé la façon dont les chercheurs abordent les études écotoxicologiques (Garcia-Reyero and Perkins, 2011). Elles ouvrent des perspectives de recherche pour approfondir nos connaissances sur les systèmes vivants, notamment chez les espèces non modèles, mais avec une pertinence environnementale. Les approches génomiques, protéomiques et métabolomiques proposent de nouveaux outils pour répondre aux défis environnementaux tels qu'évaluer et prédire l'impact du réchauffement climatique et de la pollution environnementale sur les écosystèmes (Faure and Joly, 2015; Prat and

Degli-Esposti, 2019). Dans ce contexte, le terme « écotoxicogénomique » a été proposé par Snape et al. (2004) pour décrire la réponse d'un système biologique à une exposition à une substance toxique aux niveaux transcriptomique, protéomique et métabolomique (Figure 24). L'écotoxicogénomique a comme objectif la détermination des liens de cause à effet entre l'exposition d'un organisme à un contaminant et la modification de l'expression de ses biomolécules. Cette discipline vise ainsi à identifier des biomarqueurs d'effets ou d'exposition aux substances toxiques.

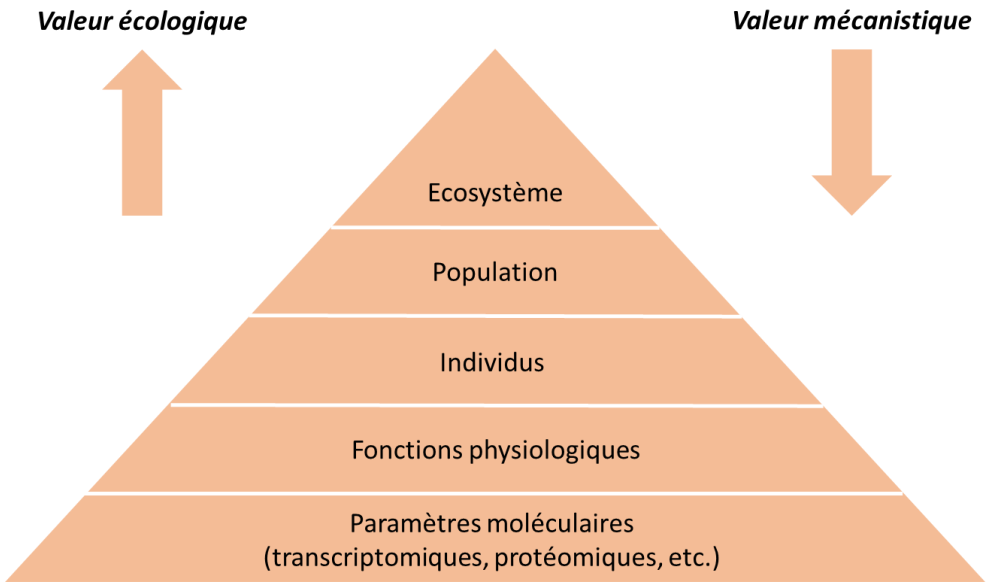


Figure 24 : Cadre conceptuel pour l'écotoxicogénomique (d'après Snape et al., 2004)

Une revue bibliométrique traitant de l'application des omiques en écotoxicologie, indique que sur 648 études, la transcriptomique est la méthode la plus fréquemment appliquée (43 %), suivie de la protéomique (30 %), de la métabolomique (13 %) et enfin de la multiomique (13 %) (Figure 25-A) (Ebner, 2021). Une augmentation de l'utilisation de la protéomique a été observée pour les études de 2017 à 2019, avec un nombre constant d'études utilisant la transcriptomique (Figure 25-A) (Ebner, 2021). Une tendance à combiner les omiques pour étudier l'impact des facteurs de stress sur les organismes a été observée, les études multiomiques constituant la majorité (44 %) de la

littérature en 2020 (Figure 25-A). Toutes années confondues, les études multiomiques ont le plus souvent utilisé une combinaison de transcriptomique et de protéomique (38 %), suivie de la transcriptomique et de la métabolomique (33 %) et de la protéomique et de la métabolomique (21 %). Cependant, les ensembles de données multiomiques au-delà de deux couches restent encore rares (Ebner, 2021).

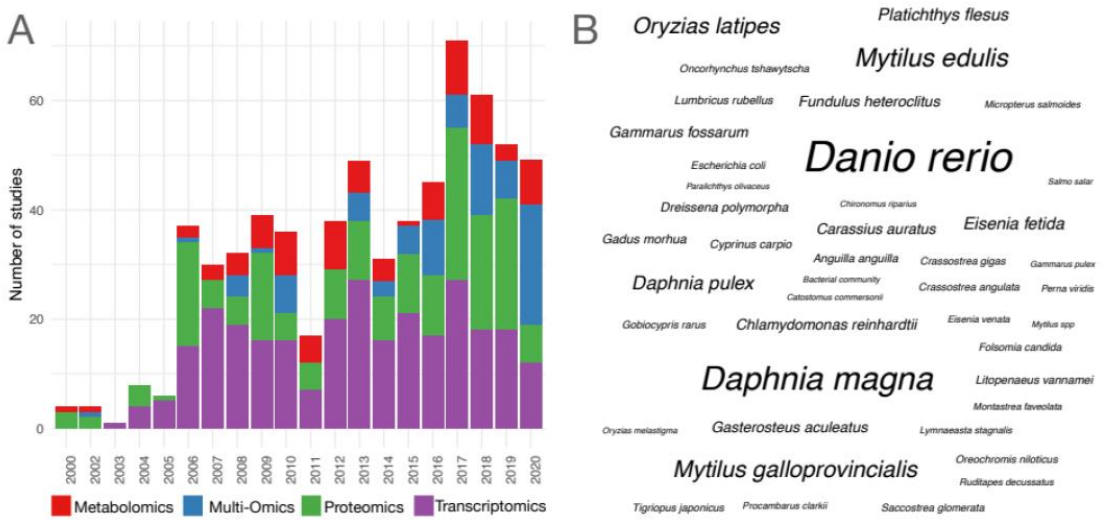


Figure 25 : (A) Études écotoxicologiques et de stress écologiques entre 2000 et 2020 qui ont utilisé une ou plusieurs méthodes omiques pour étudier les changements moléculaires suite à une exposition à des facteurs de stress (environnementaux et chimiques). (B) Nuage de mots montrant les espèces modèles et non modèles représentatives étudiées au cours de toutes les années (seules les espèces avec n > 2 études sont présentées). La taille des mots est proportionnelle au nombre d'études (Ebner, 2021).

2.1.3 Des espèces modèles aux espèces non modèles

Bien que les espèces modèles soient essentielles pour la découverte et l'acquisition de connaissances en biologie moléculaire et en physiologie moléculaire, leur utilisation en science environnementale présente des limites. De par la grande diversité des milieux et la répartition géographique des espèces, les études réalisées sur les organismes modèles en conditions de laboratoire ne permettent pas toujours d'extrapoler et de prédire les réponses pour les organismes présents dans les milieux naturels (Banks and Stark, 1998; Calow et al., 1997; Van Straalen, 2003).

Parce qu'il n'est pas possible de tester les effets de tous les produits chimiques sur toutes les espèces et tous les scénarios d'exposition, les chercheurs ont comme défi de pouvoir projeter les effets mesurés chez les espèces modèles aux espèces des écosystèmes (Ebner, 2021), ou de développer des approches généralisables à toutes espèces (Benson and Giulio, 2006; Ebner, 2021; Kunin et al., 2005; Luo et al., 2015; Siepel et al., 2005). Par exemple, la connaissance des réponses à une exposition peut être utilisée pour prédire les modes d'action d'agents similaires dans d'autres groupes (Waters and Fostel, 2004). Ces effets dits « de pont » accélèrent considérablement la capacité à évaluer l'impact environnemental sur un ensemble diversifié d'espèces au fil du temps (Martyniuk, 2018).

À ce jour, de nombreux organismes sont utilisés comme organismes sentinelles en écotoxicologie. Les espèces sentinelles renseignent sur les modifications de l'écosystème par des changements au niveau moléculaire, cellulaire, physiologique ou comportemental, qui révèlent leur exposition à des substances polluantes. Par exemple, *Gammarus pulex* ou *Gammarus fossarum* (Bertin et al., 2016; Kunz et al., 2010) et *Mytilus edulis* sont fréquemment utilisées pour les espèces aquatiques (Beyer et al., 2017). Chez les espèces terrestres, les vers de terre comme *Eisenia spp*. (Gong and Perkins, 2016) et *Lumbricus spp*. (Calisi et al., 2019) sont considérés comme d'excellentes espèces sentinelles compte tenu de leur contact étroit avec l'environnement et de leurs rôles essentiels dans la pédogenèse, la structure, la fertilité et la chaîne alimentaire terrestre (Prat and Degli-Esposti, 2019).

2.2 Approche de biologie des systèmes

2.2.1 Définition du concept de réseau biologique

La biologie des systèmes est l'étude intégrée des systèmes biologiques (*e.g.* cellules, tissus, organes ou organismes entiers) au niveau moléculaire. Elle propose une approche descendante (top-down) et intégrative qui complète la biologie réductionniste (*i.e.* l'approche ascendante

(bottom-up) traditionnelle) (Garcia-Reyero and Perkins, 2011; Wanjek, 2011) (Figure 26). Les approches descendantes permettent la découverte de composants nouveaux et essentiels aux processus biologiques sans *a priori* (Figure 26) (Garcia-Reyero and Perkins, 2011). Par analogie, la toxicologie des systèmes consiste à étudier la perturbation d'une cellule, d'un organite, d'un tissu, d'un organe ou d'un organisme dans les conditions spécifiées d'exposition à des substances biologiques ou chimiques combinés ou pas à d'autres facteurs de stress. L'analyse des modifications moléculaires, de l'expression des transcrits, des protéines et des métabolites et leur intégration sous forme d'interactions entre tous les éléments d'un système permettent une meilleure description et compréhension des mécanismes de toxicité (Prat and Degli-Esposti, 2019).

De puissants outils d'analyse existent (*e.g.* Ingenuity Pathway Analysis (IPA) de Qiagen® ; http://www.ingenuity.com) pour intégrer et interpréter les données de différentes expériences omiques (RNA-Seq, profils d'expression transcriptomique, métabolomique ou protéomique). Cytoscape (Shannon et al., 2003) est un outil permettant l'intégration de réseaux d'interaction biomoléculaire, avec des données d'expression à haut débit (*e.g.* transcriptomique ou protéomique) et d'autres états moléculaires, dans un cadre conceptuel unifié. BioNetBuilder est un plug-in pour Cytoscape, qui intègre les interactions moléculaires et d'autres types de données à haut débit provenant de plusieurs bases de données publiques, pour construire des réseaux biologiques pour toutes les espèces pour lesquelles de telles données peuvent être trouvées (Avila-Campillo et al., 2007). Ces analyses peuvent révéler les mécanismes moléculaires impliqués et identifier de nouvelles cibles ou biomarqueurs potentiels dans le cadre des systèmes biologiques étudiés. Les données omiques combinées à ces analyses de fouille permettent de générer des hypothèses mécanistes qui pourront ensuite être vérifiées avec des tests plus fonctionnels, ciblés

et spécifiques (*e.g.* expression de gènes, modification post-traductionnelles, gene silencing, CRISPR/Cas9, etc.) (Prat and Degli-Esposti, 2019).

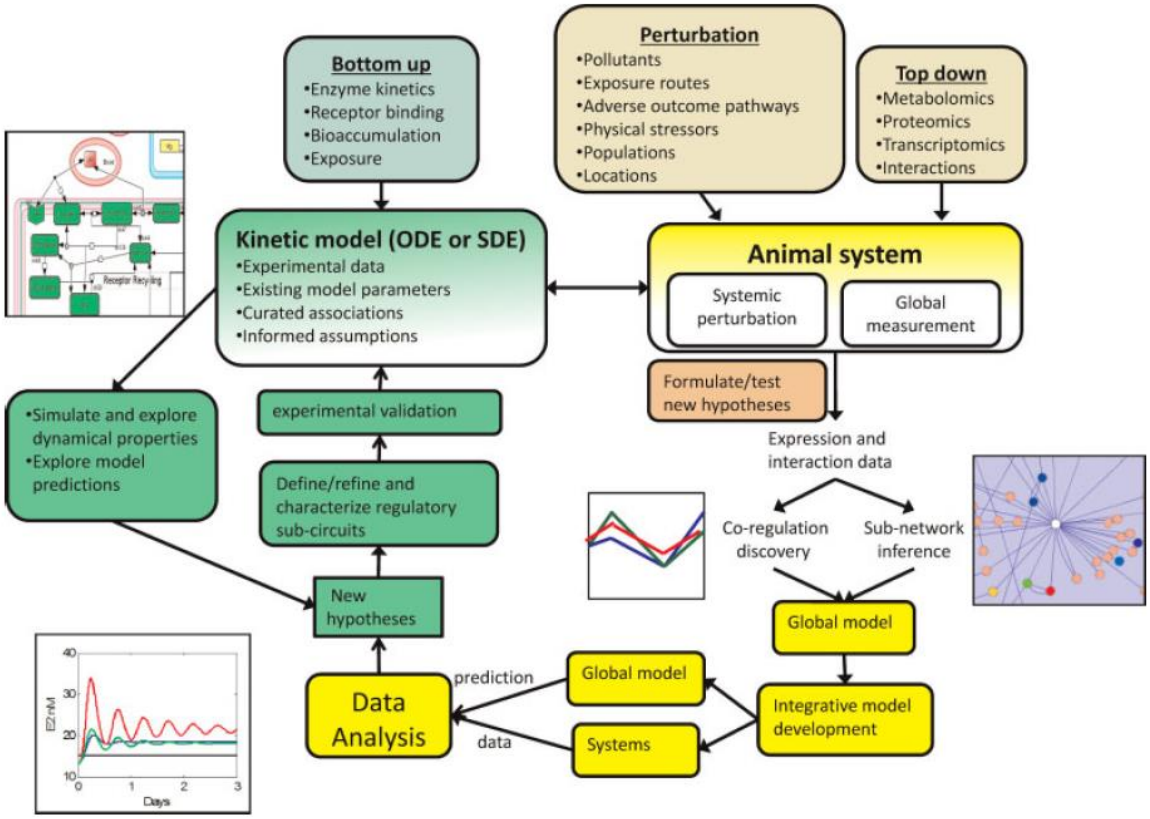


Figure 26 : L'approche de la biologie des systèmes commence par une hypothèse testée en perturbant le système. Les changements moléculaires sont mesurés à plusieurs niveaux à l'aide de technologies à haut débit. Les ensembles de données obtenus conduiront au développement d'algorithmes pour déduire des modèles prédictifs. Les simulations permettront d'élucider les révisions du système pour produire les résultats souhaités. Une approche « descendante et ascendante » sera essentielle pour manipuler avec précision un circuit biologique et prédire ses résultats au niveau du système. ODE : équation différentielle ordinaire ; SDE : équation différentielle stochastique (Garcia-Reyero and Perkins, 2011).

2.2.2 *Biologie des systèmes en écotoxicologie*

Bien que les sciences « omiques » aient été divisées en de nombreuses sous-disciplines (*e.g.* la génomique, la transcriptomique, la protéomique, la métabolomique, la lipidomique, l'épigénomique, la fonctionomique, l'immunogénomique, l'immunoprotéomique, l'interactomique et la pathomique), elles relèvent toutes du domaine de la biologie des systèmes (Leung, 2018). La biologie des systèmes vise à intégrer les réponses globales au sein d'un organisme du génotype au phénotype et est appelée l'étude intégrée des disciplines omiques. D'autre part, le terme « toxicologie systémique » a été utilisé pour décrire l'intégration des approches de biologie

systémique avec la toxicologie traditionnelle. Avec l'avènement des NGS, ainsi que les applications de la spectrométrie de masse aux systèmes biologiques, les chercheurs peuvent désormais examiner l'image globale du système, plutôt que d'examiner individuellement des gènes, des protéines ou des métabolites (Caputo, 2020; Simmons et al., 2015). Cette vision holistique apportée par la biologie de systèmes a récemment attiré l'attention de la communauté écotoxicologique pour son potentiel d'identification de liens cause-effet entre expositions environnementales et effets indésirables (Ankley et al., 2009, 2006; Garcia-Reyero and Perkins, 2011).

L'évaluation des risques écologiques (ERA : Ecological Risk Assessment) s'est longtemps appuyée sur les données de survie, de croissance ou développement et de reproduction. Deux facteurs clés ont limité l'utilisation des données omiques et de la biologie des systèmes dans l'ERA : le manque de liens établis entre les réponses des sous-organismes et les effets au niveau de la population, en plus du manque d'outils d'extrapolation appropriés (Ralston-Hooper et al., 2013; Van Aggelen et al., 2010). De nombreux efforts pour appliquer les omiques à l'écotoxicologie et à l'ERA sont en cours. Ankley et al. (2006) ont analysé le rôle de la toxicogénomique dans le domaine de l'écotoxicologie réglementaire, en explorant l'utilisation de la génomique pour détecter les mécanismes d'action de différents produits chimiques. Un autre exemple est celui décrit par Ankley et al. (2009) sur les effets de composés perturbateurs endocriniens chez deux modèles de poissons, le tête-de-boule et le poisson zèbre. Le projet utilise une approche systémique (*i.e.* combinaison d'approches omiques, bioinformatiques et de modélisation) pour déterminer les voies de toxicité de 12 composés modèles perturbateurs endocriniens avec différents mécanismes d'action. Les auteurs ont développé des indicateurs d'exposition et des modèles prédictifs des effets basés sur les mécanismes d'action. Dans une autre étude, DeWit et al. (2008) ont utilisé une approche

transcriptomique et protéomique combinée pour évaluer les effets moléculaires du retardateur de flamme, le tétrabromobisphénol-A (TBBPA), sur le foie des poissons zèbres. Cette combinaison a permis aux auteurs de détecter de façon originale une interférence entre l'homéostasie de la thyroïde et la vitamine A chez les poissons exposés, ainsi que des réponses au stress général et oxydatif. Il existe quelques exemples récents de transcriptomique combinée à la métabolomique. Williams et al. (2009) ont utilisé une puce d'ADNc (*i.e.* micro-array) d'épinoches à trois épines (*Gasterosteus aculeatus*) et une approche RMN pour élucider les réponses des poissons à l'exposition aux hydrocarbures aromatiques polycycliques (HAPs). Ils ont montré des changements dans l'expression des gènes liés à la biosynthèse des acides biliaires, au métabolisme des stéroïdes et à la fonction endocrinienne, ainsi que des changements dans les concentrations de taurine, malonate, glutamate et alanine. Santos et al. (2010) ont étendu cette approche afin d'étudier l'impact du cuivre sur l'épinoche à trois épines, montrant des modulations de gènes et des changements métaboliques liés au mécanisme de toxicité du cuivre connu et conservé chez les mammifères.

Récemment, une approche écotoxicologique intégrative sur l'effet chronique de l'exposition des poissons médaka adultes à diverses souches de cyanobactéries (productrices de microcystines connues pour leur forte hépatotoxicité) a été réalisée à l'aide d'analyses combinées d'anatomopathologie, de protéomique et de métabolomique (Le Manach et al., 2018). Pour explorer la relation entre les différents métabolites et les protéines enrichies, une analyse de réseau a été réalisée grâce à l'outil IPA (*i.e.* Ingenuity Pathway Analysis) pour illustrer la réponse moléculaire fonctionnelle impliquée, à partir de ce qui est connu chez l'Homme (Marie, 2020).

2.2.3 L'analyse des réseaux de coexpression

Les systèmes biologiques sont dynamiques et tendent vers un équilibre quantitatif entre les composants qui détermine souvent le résultat biologique (*e.g.* traits physiologiques, réponse à un contaminant). Ainsi, l'étude des propriétés structurelles des réseaux moléculaires sous-jacents peut aider à comprendre le comportement cellulaire du point de vue systémique (Eisenberg et al., 2000). Il existe plusieurs types de réseaux : les réseaux d'interactions protéiques (Habibi et al., 2014) ; les réseaux de régulation génique (Davidson and Levin, 2005) ; réseaux métaboliques (Palsson, 2006). Ici, nous nous focaliserons sur les réseaux de coexpression comme méthode pour l'analyse des données transcriptomiques, protéomiques ou métabolomiques (Guarascio et al., 2019).

La traduction de l'information moléculaire en une meilleure compréhension biologique par l'analyse conventionnelle de l'expression différentielle reste un défi majeur (Ruffalo et al., 2015). Par exemple, il est difficile d'évaluer le rôle et l'implication des protéines ou métabolites avec une faible abondance ou associés à un faible fold change (*i.e.* facteur de changement entre deux mesures) identifiés par l'analyse de l'expression différentielle (Pei et al., 2014; Ruffalo et al., 2015).

Les réseaux de coexpression de gènes sont des approches utilisées pour analyser les interactions entre gènes au niveau du système biologique. Ces approches construisent des réseaux de relations basés sur des mesures d'association par paires d'éléments (*e.g.* gènes, protéines). Ils se sont révélés être des outils puissants pour déchiffrer les réponses cellulaires ou identifier les voies critiques pertinentes pour les traits ou conditions clés, telles que le lien entre les gènes et une maladie, ou les gènes et le poids (Lehtinen et al., 2013; Zhang and Horvath, 2005; Zhang et al., 2016). Un réseau de coexpression génique est constitué de profils d'expression génique représentés sous forme de nœuds et de connexions géniques, qui se produisent si deux gènes sont significativement coexprimés (déterminés par des corrélations d'expression génique par paires)

(Zhang and Horvath, 2005) (Figure 27). L'analyse pondérée du réseau de coexpression de gènes (WGCNA pour Weighted Gene Coexpression Network Analysis) a été développée pour décrire la relation de corrélation entre (i) les gènes à travers des échantillons analysés par puces à ADN ou de RNA-Seq, (ii) et la relation de corrélation entre des clusters de gènes fortement corrélés (appelés modules) et des conditions externes ou des traits biologiques (Langfelder and Horvath, 2008). Les analyses des réseaux de coexpression génique des données transcriptomiques ont fourni des informations utiles pour comprendre les processus biologiques fondamentaux chez de nombreuses espèces d'arthropodes. Par exemple, WGCNA a fourni des preuves de voies conservées à travers 16 espèces différentes de fourmis impliquées dans la division du travail de la reproduction (Morandin et al., 2016). De plus, une analyse de réseau a mis en évidence le rôle des gènes métaboliques et de signalisations cellulaires dans la réponse au stress chez le crustacé modèle *Daphnia magna*, lorsque la simple analyse d'expression différentielle n'a pas permis de trouver des réponses biologiques significatives (Orsini et al., 2018).

L'un des principaux avantages de ce modèle statistique est d'être axé sur les données et de ne pas nécessiter d'hypothèse. Il représente donc une grande opportunité pour les organismes non modèles pour lesquels les séquences génomiques, les annotations de gènes ou les voies métaboliques et de signalisation ne sont pas encore connues. Malgré les avantages cités ci-dessus, l'application de la méthode WGCNA a été relativement peu utilisée avec d'autres types de données omiques, tels que les données protéomiques et métabolomiques ces dernières années (Kunowska et al., 2015; Pei et al., 2014; Su et al., 2014; Yu et al., 2015). Dans cette thèse, nous présentons les travaux de deux analyses de réseaux de coexpression appliquées à des jeux de données de protéomique shotgun chez *Gammarus fossarum* (voir page 115).

Un réseau de coexpression peut être représenté sous plusieurs formes. L'une est un graphe non orienté, où un nœud représente un gène (ou une protéine ou un métabolite) et une paire de nœuds est relié par une arête (non orienté) s'il existe une relation de coexpression significative (Stuart et al., 2003) (Figure 27). Cela peut être aussi sous la forme d'une matrice de similarité gène-gène (Figure 28-1).

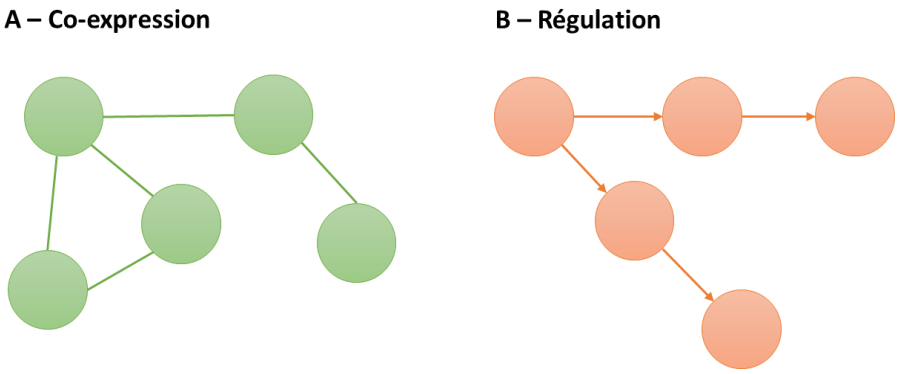


Figure 27 : Différence entre les réseaux de coexpression (A) et les réseaux de régulation de gènes (B).

La transformation appropriée des ensembles de données en protéomique et en métabolomique est une condition préalable à la capture de la véritable corrélation dans les ensembles de données (Pei et al., 2017). Par exemple, le nombre total de signaux détectés peut être significativement différent en raison de la taille, du poids ou du volume des échantillons (Wu and Li, 2016). Par conséquent, sans effectuer une transformation appropriée pour normaliser les données d'abondance des protéines et des métabolites, la corrélation de coexpression ne peut pas être correctement construite. Dans le cas de données d'abondance protéomique et métabolomique semi-quantitatives, une transformation appropriée peut être encore plus importante pour stabiliser les variances et capturer la véritable corrélation (Nie et al., 2007). En ce qui concerne les données de protéomique et de métabolomique, les jeux de données se retrouvent souvent incomplets en raison de l'identification imparfaite des séquences codantes au sein d'un génome et de la sensibilité des technologies de détection (Lazar et al., 2016). Une grande proportion de valeurs manquantes est souvent rencontrée (Albrecht et al., 2010; Huan and Li,

2015). Pour pallier cela, plusieurs approches, telles que la méthode des k plus proches voisins (dit kNN), la méthode des moindres carrés et la méthode des moindres carrés locaux, ont été proposées (Nie et al., 2007; Xia et al., 2015). Chacune des méthodes d'imputation (*i.e.* remplacement des données manquantes) est consacrée à un type spécifique de mécanismes d'absence (*e.g.* erreurs analytiques comme la détection d'une espèce chimique, erreurs bioinformatiques comme la mauvaise identification d'un peptide) (Lazar et al., 2016). Cela nécessite une compréhension et un contrôle précis de l'imputation de chaque valeur manquante (Pei et al., 2017). Enfin, une étape supplémentaire pour préparer le jeu de données consiste à prendre en compte et corriger l'effet batch (ou de lot) possible dans les échantillons. Une fois ces étapes essentielles de prétraitement des données effectuées, la construction et les analyses du réseau de coexpression se déroulent en trois grandes étapes (Van Dam, 2017) (Figure 28) :

1. Dans la première étape de **calcul de la coexpression**, les relations individuelles entre les gènes sont généralement définies sur la base de mesures de corrélation (*e.g.* corrélations de Pearson ou de Spearman) (Ala et al., 2008; Guttman et al., 2011) entre chaque paire de gènes (Figure 28-1). Ces relations décrivent la similarité entre les profils d'expression de chaque paire de gènes possible dans tous les échantillons.
2. Dans la deuxième étape, les associations de coexpression sont utilisées pour **construire un réseau** de nœuds (gènes) et d'arêtes (présence et force de la coexpression) (Figure 28-2). Un réseau de coexpression peut être pondéré ou non pondéré. Dans un réseau pondéré, tous les nœuds sont connectés les uns aux autres. Ces connexions ont des valeurs de

poids continues entre 0 et 1 qui indiquent la force de la co-régulation entre les gènes. Dans un réseau non pondéré, l'interaction entre les paires de gènes est binaire, c'est-à-dire 0 ou 1, indiquant que les gènes sont soit connectés, soit non connectés. Un réseau non pondéré peut être créé à partir d'un réseau pondéré, par exemple en considérant que tous les gènes dont la corrélation est supérieure à un certain seuil sont connectés et tous les autres non connectés.

3. Dans la troisième étape, **les modules** (ou cluster de gènes fortement coexprimés) **sont identifiés** en utilisant l'une des nombreuses techniques de clustering disponibles (*e.g.* k-means, clustering hiérarchique) (D'haeseleer, 2005) (Figure 28-3). Le clustering, dans les analyses de coexpression, est une méthode qui peut être utilisée pour identifier des groupes de gènes qui ont un profil d'expression similaire à travers plusieurs échantillons, pour produire des groupes de gènes co-exprimés plutôt que seulement des paires.

Après avoir défini les modules de gènes coexprimés, les eigengènes de module (dits ME pour module eigengenes) sont calculés pour chaque module. Les ME sont des vecteurs représentatifs des profils d'expression des protéines dans un module. Ils représentent la composante principale de chaque module et permettent d'explorer la corrélation entre les modules et les traits phénotypiques, tels que les conditions d'exposition. Les modules sont ensuite interprétés par des analyses d'enrichissement fonctionnel, une méthode qui peut être utilisée pour identifier et classer les catégories fonctionnelles surreprésentées dans une liste de gènes (Chen et al., 2009; de Magalhães et al., 2010; Gupta et al., 2014). Il est aussi possible de se focaliser sur les éléments (*i.e.* gènes, protéines ou métabolites) « hub » de chaque module, qui présentent le plus haut degré de

connectivité au sein d'un module et qui jouent donc potentiellement un rôle clé dans les voies biologiques (Horvath and Dong, 2008). Ces éléments hub sont identifiés par leurs propriétés au sein du réseau, telles que la connectivité totale ou intramodulaire (*i.e.* à l'intérieur d'un module) qui représente la façon dont les éléments sont liés aux autres éléments du réseau entier ou au sein du même module, respectivement. Après identification des éléments hub correspondants, il est possible d'exporter les réseaux vers des logiciels de visualisation, tels que VisANT (Hu et al., 2008) ou Cytoscape (Shannon et al., 2003). La visualisation aide à comprendre et interpréter les topologies du réseau et des modules.

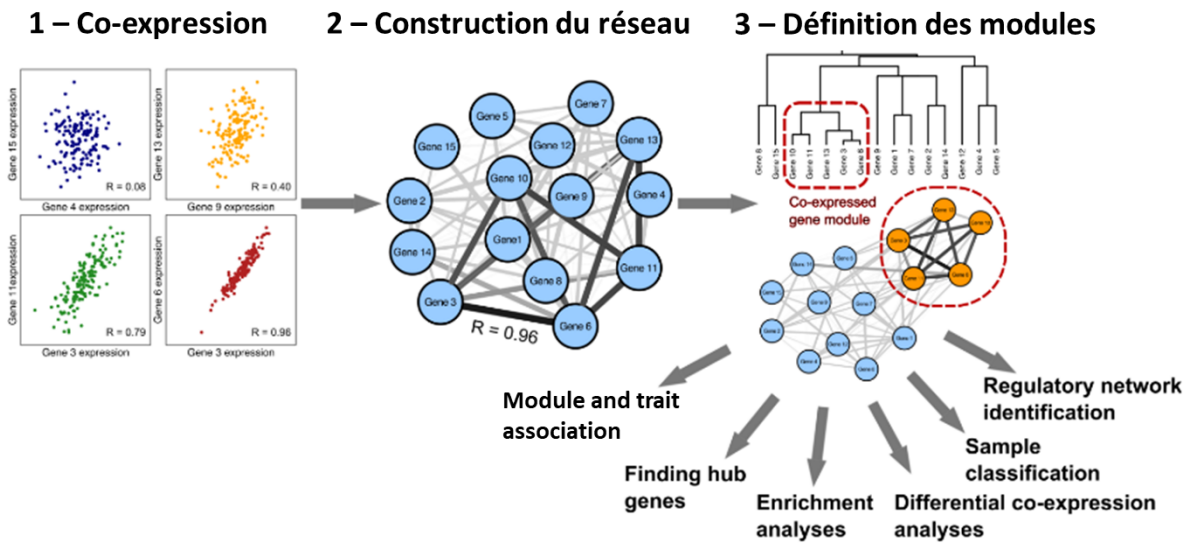


Figure 28 : Schéma d'analyse de réseau de coexpression (adaptée de Van Dam, 2017).

2.3 Les approches multiomiques pour l'annotation

2.3.1 Définition et intérêt de la multiomique

Actuellement, la plupart des études toxicologiques et écotoxicologiques utilisant des approches omiques impliquent un seul type de plate-forme, par exemple la transcriptomique ou la protéomique. Chaque approche ne détecte que les biomolécules d'un seul type et peut conduire à l'identification de biomarqueurs, mais ne permet pas de fournir une compréhension systémique

des voies de toxicité (Caputo, 2020). La principale motivation pour l'intégration de plusieurs jeux de données omiques est d'aller vers une vision holistique des réponses à un toxique (Canzler et al., 2020) et ainsi de résoudre plusieurs verrous, notamment l'analyse des mécanismes moléculaires, l'identification des biomarqueurs ou encore l'annotation des gènes et protéines (Tini, 2018). Le dernier point sera traité plus en détail dans le chapitre III (voir page 141).

Pour mieux comprendre les mécanismes moléculaires qui caractérisent les traits complexes (*e.g.* la réponse ou les mécanismes de tolérance aux contaminants), on recherche les interactions entre les biomolécules de différentes plates-formes et les voies biologiques, notamment au moyen d'analyse de réseaux (Chari et al., 2010; Glass et al., 2013; Kuo et al., 2013; Piwowar and Jurkowski, 2015; L. Wang et al., 2014). Des réseaux ont été construits par l'outil 3Omics (Piwowar and Jurkowski, 2015) pour décrire les connexions entre les données transcriptomiques, protéomiques et métabolomiques, à l'aide d'analyse de corrélation et de l'exploration de la littérature (Tini, 2018).

Un autre objectif important de l'intégration des données multiomiques est l'identification des biomolécules qui caractérisent un phénotype. L'intégration des données omiques doit permettre une compréhension plus fine des interactions entre les biomolécules provenant de différentes données omiques par rapport à l'analyse individuelle de chaque jeu de données omique (Bonnet et al., 2015; Y. Chen et al., 2013; Kim et al., 2016; Vaske et al., 2010; Wahl et al., 2015; Wang et al., 2015)

2.3.2 Les informations des différentes couches moléculaires

Pour faciliter la détection des réponses des voies moléculaires à l'exposition par des produits chimiques, les couches omiques doivent être choisies pour interroger de manière optimale les voies d'intérêt, si elles sont connues. Par exemple, l'abondance de certains métabolites (*e.g.* hormones) peut être mesurée par la métabolomique, alors que les niveaux des enzymes sont

préférentiellement mesurés en protéomique (Canzler et al., 2020). Or, les effets de certaines modifications dans l'abondance des métabolites peuvent provenir de mécanismes de régulation des voies métaboliques ou à l'expression des protéines et des activités enzymatiques.

Dans le cas des espèces non modèles, Canzler et al. (2020) proposent d'inclure plusieurs omiques telles que la transcriptomique, la protéomique, la métabolomique et/ou la phosphoprotéomique pour améliorer la confiance dans l'identification d'une voie moléculaire de réponse au contaminant (Canzler et al., 2020). La transcriptomique comprend des informations sur les ARN régulateurs et permet d'avoir des informations sur une grande partie du génome (et de fonctions biologiques associées) transcrit par rapport à la protéomique. La protéomique, quant à elle, est beaucoup plus proche du phénotype et englobe plusieurs mécanismes de régulation qui conduisent à des changements au niveau des protéines, sans changements remarquables au niveau du transcriptome. De plus, même des altérations radicales au niveau du transcriptome n'entraînent pas nécessairement des changements détectables au niveau des protéines correspondantes, du fait des effets de la traduction et des effets post-traductionnels (Canzler et al., 2020). Enfin, la métabolomique diffère de la protéomique et de la transcriptomique du fait qu'elle (i) détecte la réponse à des niveaux très différents d'une voie métabolique et (ii) qu'elle saisit simultanément les molécules d'une voie qui répondent à des échelles de temps extrêmement différentes (*e.g.* les messagers secondaires formés presque instantanément à l'inverse de composants de la membrane cellulaire) (Canzler et al., 2020).

2.3.3 Challenges de l'intégration de données multiomiques

Les principaux défis dans l'intégration des données multiomiques sont l'hétérogénéité intrinsèque des données et la complexité biologique des interactions inter/intraomiques. L'hétérogénéité des données fait référence à des mesures qui ne sont pas prises à la même échelle,

avec des distributions variées et provenant de différentes plateformes omiques. Les méthodes statistiques et bioinformatiques doivent donc garantir que les résultats ne sont pas biaisés en faveur des omiques de plus grande dimension ou de plus grande variance (Tini, 2018). Autrement dit, les résultats biologiquement significatifs peuvent être à la fois soutenus par des signaux plus faibles impliquant différentes données omiques ou fortement induits par un seul type de données. Ce problème peut être résolu en mettant les données à l'échelle ou en réduisant la dimension des données en sélectionnant uniquement les plus informatives (*e.g.* réduire la redondance) (Hira and Gillies, 2015). Bien que la combinaison d'un plus grand nombre de niveaux biologiques permet d'obtenir une image plus complète du système biologique étudié (Bersanelli et al., 2016), elle augmente également la quantité de bruit ajoutée au modèle, intensifiant ainsi la découverte de faux positifs et les difficultés d'interprétation (Tini, 2018). Pour ceci, de nombreux outils ont été optimisés pour l'intégration multiomique. Parmi ceux-ci, deux grands types d'intégration multiomique se distinguent : l'intégration conceptuelle et l'intégration statistique (Dumas, 2020; Misra et al., 2019; Pinu et al., 2019).

Dans le cas de **l'intégration conceptuelle**, il existe deux types de stratégies, utilisant les connaissances antérieures relatives aux voies et aux réseaux métaboliques (*e.g.* interactions biochimiques, génétique) provenant de base de données ou de précédentes études (Dumas, 2020) :

(i) **L'intégration s'appuyant sur les ontologies ou les voies métaboliques**. Des analyses d'enrichissement sont applicables en se basant sur les voies métaboliques et en prenant en compte les protéines intervenant dans les voies de synthèse ou de dégradation des métabolites (Feng et al., 2018). Ces analyses sont aussi applicables au termes GO qui appartiennent à une nomenclature commune aux gènes, protéines ou métabolites qui permettent d'établir des liens entre les données omiques.

(ii) **L'intégration s'appuyant sur les réseaux biologiques**. Les réseaux biologiques ont l'avantage de prendre en compte les relations ou interactions existantes entre les entités biologiques. La construction de réseaux biologiques est possible à partir des bases de données ou des études antérieures. Les réseaux protéine-protéine et les réseaux métaboliques sont les deux approches les plus utilisées. De nombreuses bases de données servent de source de connaissances pour leur construction, telles que IntAct (Orchard et al., 2014), MINT (Licata et al., 2012), InnateDB (Lynn et al., 2008) pour les réseaux protéine-protéine ; ou KEGG (Kanehisa and Goto, 2000), MetaCyc (Caspi et al., 2008), WikiPathways (Slenter et al., 2018), Recon3D (Brunk et al., 2018) pour les réseaux métaboliques (Zhou et al., 2020).

Quant à **l'intégration statistique**, se base uniquement sur les données à disposition grâce à des corrélations telles que (Dumas, 2020) :

(i) **La corrélation matricielle** qui vise à expliquer au mieux le lien entre la variabilité intra et inter omiques, en identifiant les corrélations linéaires correspondantes. On retrouve ce type de stratégie dans le package R mixOmics (Rohart et al., 2017).

(ii) **La corrélation univariée** (*e.g.* corrélation de Spearman, corrélation de Pearson) cherche à caractériser les liens entre des molécules individuelles d'un niveau omique et celles d'un autre niveau omique (Chong and Xia, 2017).

(iii) **L'analyse multibloc** est une stratégie multivariée, qui considère les différentes omiques comme des blocs distincts dans le modèle pour ensuite les combiner (Cavill et al., 2016).

2.3.4 La multiomique en écotoxicologie

Dans une étude récente, Lv et al. (2022) ont combiné des approches transcriptomiques et protéomiques pour comprendre les mécanismes sous-jacents liés aux effets de la toxicité du cadmium sur la fonction de reproduction de l'araignée *Pardosa pseudoannulata*. En combinant les gènes et protéines différentiellement exprimés de l'ovaire, ils ont observé une inhibition des gènes et des protéines impliqués dans le traitement des protéines (*i.e.* « protein processing ») dans le réticulum endoplasmique, dans la vitellogenèse, et dans le système enzymatique antioxydant.

Pour identifier les mécanismes d'actions d'un contaminant et mieux comprendre les origines biochimiques de la modulation des métabolites, déterminer les liens directs entre métabolites et protéines se révèle être pertinent. À ce jour, un faible nombre d'études intègrent les données protéomiques et métabolomiques (Chen et al., 2018; Dumas et al., 2022; Ji et al., 2016; Sun et al., 2019; Xiang et al., 2021). La majorité des études combinant ces approches opte pour l'intégration conceptuelle en s'intéressant aux protéines et métabolites différentiellement impliqués dans les voies métaboliques et leurs interactions. Concernant le reste des études, une intégration en tant que telle n'est pas forcément proposée, et les données omiques sont traitées séparément (Dumas, 2020). Dans l'étude de Sun et al. (2019), les approches protéomiques et métabolomiques ont été combinées pour élucider les réponses moléculaires au niveau du cerveau de *Danio rerio* après avoir été exposé à l'oxyde de graphène. Les protéines et les métabolites impliqués ont été identifiés grâce à l'outil KEGG pour l'analyse des voies et KEGG PATHWAY (Moriya et al., 2007) pour comprendre les liens entre protéines et métabolites. Les auteurs ont observé des modifications du métabolisme des acides aminés, du cycle de Krebs ou encore de la synthèse des acides gras. Ces modulations ont mis en évidence une potentielle perturbation du métabolisme énergétique par l'oxyde de graphène. Très récemment, l'étude de Dumas et al. (2022) a exploré une approche métabolomique et protéogénomique intégrée. Cette étude inclut une stratégie statistique de

fusion de données et d'analyse multibloc pour comprendre les événements moléculaires et les processus cellulaires déclenchés par la carbamazépine (CBZ) chez la moule *Mytilus galloprovincialis* (Dumas et al., 2022). Les résultats restent préliminaires, mais indiquent que le mécanisme d'action de la CBZ serait probablement lié à l'induction de l'autophagie, soit par la déplétion en inositol, soit par le stress du réticulum endoplasmique. Néanmoins, ces études n'intègrent pas l'ensemble des interactions et corrélations potentielles entre protéines et métabolites, mais utilisent uniquement les voies métaboliques et les annotations fonctionnelles. Pour pallier ces limites, les méthodes utilisant les réseaux moléculaires ou les analyses statistiques pourraient fournir de plus amples informations. Toutefois, dans le cas d'espèces non modèles, les connaissances liées aux interactions entre gènes, protéines et métabolites proviennent généralement de l'extrapolation des annotations d'autres espèces présentes dans les bases de données (Wanichthanarak et al., 2015). Il est donc nécessaire de déverrouiller l'annotation des espèces non modèles, grâce à des stratégies et outils qui permettent de révéler des gènes, protéines ou métabolites propres à l'espèce étudiée. La protéogénomique précédemment présentée (voir la partie 1.4) qui couple les méthodes de RNA-Seq et de protéomique shotgun, permet d'améliorer l'annotation des génomes (Nesvizhskii, 2014). De même, l'outil de reconstruction et d'annotation fonctionnelle CycADS (Vellozo et al., 2011) couplé aux données de protéomique shotgun permet de reconstruire des voies métaboliques sans *a priori* (voir page 141). Ces méthodes permettent de préciser et affiner les annotations des biomolécules, et de planter les bases pour de futures études d'intégration multiomiques à proprement parler (*e.g.* intégration statistique).

3. LE METABOLISME LIPIDIQUE

3.1 Connaissances disponibles

En tant que macronutriments, les lipides sont une source majeure d'énergie et d'acides gras essentiels (AGE). Ce sont des composants structurels importants des biomembranes, qui agissent en tant que porteurs de vitamines liposolubles (*i.e.* soluble dans les graisses), et fonctionnent comme précurseurs d'eicosanoïdes (*i.e.* dérivés de l'oxydation des AG comme l'acide arachidonique), d'hormones et de cofacteurs enzymatiques pour les animaux aquatiques, y compris les crustacés (National Research Council, 2011). Le métabolisme des lipides comprend l'absorption, le transport, la biosynthèse et la dégradation des lipides. Il implique également plusieurs voies biochimiques avec des enzymes clés et des facteurs de transcription (Chen et al., 2015; Dutta and Sinha, n.d.; Obeid et al., 1993; Sun et al., 2020; van Meer et al., 2008). Les lipides constituent un groupe de molécules structurellement diverses qui peuvent être classées en plusieurs catégories, notamment les acyles gras, les glycérolipides, les glycérophospholipides, les sphingolipides, les stérolipides, les prénolipides, les saccharolipides et les polycétides, selon le système de classification des lipides, LIPIDMAPS (Fahy et al., 2019, 2009). Ils représentent une importante réserve organique chez la plupart des espèces de crustacés (Santos et al., 1997). Les phospholipides, et particulièrement la phosphatidyl choline et la phosphatidyl ethanolamine, sont les principaux lipides de l'hémolymphe, faisant environ 65% du total des lipides dans ce tissu, tandis que les triglycérides représentent les lipides les plus stockés (Allen, 1972; Gilbert and O'Connor, 1970).

De nombreux polluants s'avèrent être des perturbateurs endocriniens qui pourraient provoquer des perturbations de l'homéostasie lipidique chez les espèces aquatiques (Fu et al., 2021) qui est cruciale pour leur développement, leur maintien et leur reproduction (de Mendoza and Pilon, 2019; Klose et al., 2012; Zhang and Rock, 2008). Plusieurs études montrent que certains de

ces polluants, les obésogènes, pourraient interférer avec l'homéostasie des lipides et provoquer des effets toxiques sur un certain nombre d'espèces animales aquatiques (Capitão et al., 2017; Fuertes et al., 2020)

3.2 Effets des facteurs biologiques

Plusieurs études ont montré chez les amphipodes, que le genre, le cycle de reproduction ou encore le stade de développement des individus modulent les profils et quantités en lipides (acides gras polyinsaturés (PUFAs), lipides totaux) (Fu et al., 2021) et en énergie (glycogène) (Correia et al., 2003; Fu et al., 2021; Gismondi et al., 2012; Rosa and Nunes, 2002; Sroda and Cossu-Leguille, 2011). Les différences de réserves énergétiques (lipides et protéines) entre les mâles et les femelles pourraient s'expliquer par l'oogenèse (Buikema Jr and Benfield, 1979) et l'entretien des œufs (Sroda and Cossu-Leguille, 2011). Plaistow et al. (2003) ont observé le même schéma chez *G. pulex* avec un contenu lipidique plus élevé chez les femelles que chez les mâles (Plaistow et al., 2003). Les œufs sont composés de vitellogénines, et chez les amphipodes, d'autres protéines appartenant à la famille des protéines de transfert des grands lipides (« large lipid transfer proteins » ou LLTP) constituant le vitellus (Meusy, 1980, Trapp et al 2016, Degli Esposti et al 2019). Dans une étude comparant le mâle et la femelle de *G. roeseli*, des différences ont été observés dans les capacités de défense antioxydantes et les réserves énergétiques ont été observées. En effet, des niveaux plus élevés de lipides et de protéines antioxydantes ont été mesurés chez les femelles, ce qui suggère que les femelles pourraient avoir de meilleures capacités de défense contre le stress oxydatif que les mâles (Sroda and Cossu-Leguille, 2011). Pour deux espèces de céphalopodes marins, *Eledone cirrhosa* et *Eledone moschata*, la teneur en lipides et en protéines augmente dans la gonade pendant l'ovogenèse chez les femelles (Rosa et al., 2004).

Le métabolisme des lipides a également été étudié dans des organes tels que l'hépatopancréas d'espèces de crabes. D'un point de vue anatomo-physiologique, l'hépatopancréas est un organe clé pour l'absorption et le stockage des nutriments chez divers crustacés, et il joue un rôle important dans le métabolisme des lipides, le statut nutritionnel et le stockage de l'énergie (Wang et al., 2008; W. Wang et al., 2014). L'hépatopancréas stocke une grande quantité de substances énergétiques, en particulier des lipides, comme réserve d'énergie à utiliser pour l'organisme pendant la mue, la privation alimentaire et la reproduction. L'hépatopancréas intervient aussi dans la régénération des membres dans la phase de croissance des crustacés par la β-oxydation (une voie majeure du catabolisme des acides gras) qui a lieu dans les mitochondries et le peroxysome des cellules (Huang et al., 2015; Tocher, 2003). L'hépatopancréas est un organe de choix pour étudier le métabolisme des lipides et le métabolisme énergétique (Yuan et al., 2019). Plusieurs études ont signalé la cooccurrence d'une diminution de la teneur en lipides dans l'hépatopancréas et d'une augmentation de la teneur en lipides dans l'ovaire pendant la maturation ovarienne chez les crustacés (Alava et al., 2007; Castille and Lawrence, 1989; Spaargaren and Haefner, 1994).

Aujourd'hui, il est possible d'aborder l'étude de la composition lipidique et sa dynamique grâce aux méthodes de hautes résolutions de spectrométrie de masse (MS) (lipidomique non ciblée ou shotgun) qui ont permis notamment d'augmenter considérablement l'identification d'espèces de lipides (Züllig and Köfeler, 2021). La lipidomique shotgun a été appliquée avec succès pour décrire le lipidome d'une variété d'espèces dont *Drosophila* (Carvalho et al., 2012; Knittelfelder et al., 2020; Palm et al., 2012) ou encore du crustacé d'eau douce *Daphnia magna* (Taylor et al., 2017). Pour aller plus loin dans la compréhension des modes d'action, la caractérisation des lipides et de leur distribution spatiale dans les tissus est devenue cruciale. Récemment, une étude par

lipidomique shotgun et MSI (Imagerie MS : Imagerie par spectrométrie de masse) a permis de caractériser et de cartographier aux niveaux des tissus, le lipidome de *G. fossarum*, ceci en lien avec le genre et les stades reproductifs des femelles (Fu et al., 2021). Les auteurs ont découvert des lipides à base de sulfate dans l'hépatopancréas et leur accumulation dans les ovocytes matures. Cela suggère que l'hépatopancréas fonctionne également comme un organe de stockage des lipides chez les amphipodes, et qu'il pourrait libérer les lipides nécessaires aux ovocytes pour faciliter leur maturation (Fu et al., 2021).

3.3 Effets des contaminants

Les contaminants chimiques peuvent cibler les voies du métabolisme lipidique chez les crustacés. Il a été montré que certains perturbateurs endocriniens, connus comme obésogènes (*e.g.* tributylétain), interférents avec le métabolisme lipidique des vertébrés et affectent la distribution et la synthèse des lipides chez le crustacé modèle *Daphnia magna* (Fuertes et al., 2019; Jordão et al., 2016a, 2016b; Jordão Rita et al., 2015). Des résultats récents ont montré que l'exposition aux juvénoïdes, au tributylétain et au bisphénol A favorisait l'accumulation de lipides de stockage tels que les triacylglycérols et le cholestérol chez *Daphnia magna*, montrant que la perturbation lipidique observée chez les vertébrés peut se retrouver aussi chez les invertébrés (Fuertes et al., 2018; Jordão et al., 2016a, 2016b; Jordão Rita et al., 2015). Dans l'étude de Fuertes et al. (2019), les modifications transcriptomiques consécutives à l'accumulation des lipides induit par le pyriproxyfène (*i.e.* pesticide juvénoïde), le bisphénol A et le tributylétain ont été analysées (Fuertes et al., 2019). L'annotation fonctionnelle a révélé que les gènes régulés à la hausse étaient impliqués dans les voies métaboliques des acides gras, des glycérophospholipides et des

glycérolipides, les constituants membranaires et les voies métaboliques de la chitine et de la cuticule (Fuertes et al., 2019).

D'autres médicaments, comme les statines se retrouvent rejetées dans les cours d'eau à des concentrations allant jusqu'à plusieurs nano grammes par litre (Jelic et al., 2011). La pravastatine est un médicament très utilisé chez l'Homme pour traiter les dyslipidémies (*i.e.* augmentation du cholestérol dans le sang). Chez l'Homme, la pravastatine inhibe l'hydroxyméthyl-glutaryl coenzyme A (HMG-CoA) réductase (HMGR) qui catalyse la conversion de l'HMG-CoA en mévalonate, une étape clé dans la synthèse du cholestérol chez les vertébrés (Figure 29). L'HMGR est une enzyme très conservée chez les eucaryotes et peut donc être une cible commune aux statines . La voie de synthèse du mévalonate est impliquée chez les arthropodes pour la synthèse des hormones juvéniles (JHs) qui jouent un rôle sur le développement embryonnaire, la métamorphose, la synthèse de la vitellogénine et la production de phéromones (Bellés et al., 2005; Nijhout, 1994). Cette régulation intervient par le biais de l'hormone juvénile III (chez les insectes) ou par le méthyl farnésoate (MF) (chez les crustacés), qui est produit à partir du mévalonate. Des effets négatifs de la simvastatine ont été observés sur la reproduction de *G. locusta* (Neuparth et al., 2014).

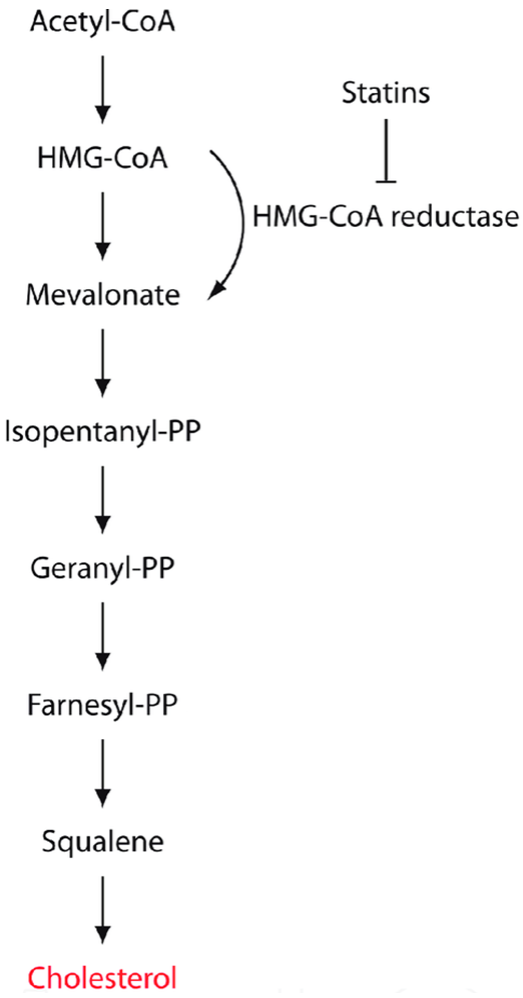


Figure 29 : Schéma simplifié de la voie du mévalonate (adaptée de Hissa and Pontes, 2018).

L'étude du lipidome s'est aujourd'hui élargie aux espèces non modèles. Chez *G. fossarum*, des effets endocriniens (*e.g.* une maturation accélérée des ovocytes, des ovocytes vitellogènes plus petits et une diminution de la production de spermatozoïdes) ont été observés après une exposition à des contaminants tels que le cadmium, les insecticides juvenoïdes ou les agonistes au recepteur à l'écdysone. (Schirling et al., 2005; Trapp et al., 2015). Il est donc très important de caractériser le lipidome de cet organisme, de comprendre le mécanisme moléculaire sous-jacent de ces effets endocriniens et de développer des biomarqueurs pour une évaluation précoce du risque de contamination. Une première évaluation de la perturbation du lipidome chez *G. fossarum* exposé à un perturbateur endocrinien a été récemment rapportée (Arambourou et al., 2018), cependant,

seul un nombre limité de classes de lipides et d'espèces moléculaires sont connus et caractérisés (Arambourou et al., 2018; Fu et al., 2020; Kolanowski et al., 2007). Récemment, un lipidome profond a été caractérisé, décrit et cartographié chez cette espèce d'intérêt, *G. fossarum* (Fu et al., 2021).

4. LE GAMMARE COMME ESPECE SENTINELLE EN ECOTOXICOLOGIE

4.1 Description de l'espèce

Gammarus est un genre de crustacé amphipode appartenant au sous-ordre des Gammaridae, représenté par plus de 200 espèces (Martin and Davis, 2001; Väinölä et al., 2008) (Tableau 6).

Tableau 6 : Systématique de *G. fossarum* (Martin and Davis, 2001)

EMBRANCHEMENT	Arthropode
SUPER-CLASSE	Crustacé
CLASSE	Malacostracé
SOUS-CLASSE	Eumalacostracé
SUPER-ORDRE	Péracaride
ORDRE	Amphipode
SOUS-ORDRE	Gammaridea
FAMILLE	Gammaridae
GENRE	*Gammarus*
ESPÈCE	*fossarum* (Koch, 1835)

Pour ces travaux de thèse, nous avons choisi de travailler sur l'espèce *Gammarus fossarum* (Koch, 1835) (Figure 30). Cette espèce a longtemps été considérée comme une sous-espèce de *Gammarus pulex*, et était nommée *Gammarus pulex fossarum* (Wautier and Roux, 1959).

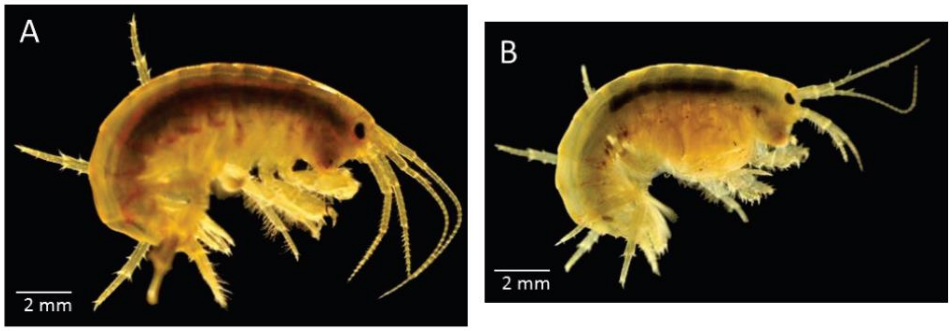


Figure 30 : *Gammarus fossarum* mâle (A) et femelle (B) (Source : Hervé Quéau, INRAE Lyon) (Gouveia, 2017).

Cette espèce de crustacé d'eau douce est largement répandue en Europe et en Asie Mineure . D'après Piscart et Bollache (2012), la présence de *G. fossarum* s'étendrait sur tout le territoire français sauf la Corse, la Bretagne et une partie de la Normandie (Figure 31). Il s'agirait de l'espèce de gammare la plus répandue après *G. pulex* (Piscart and Bollache, 2012).

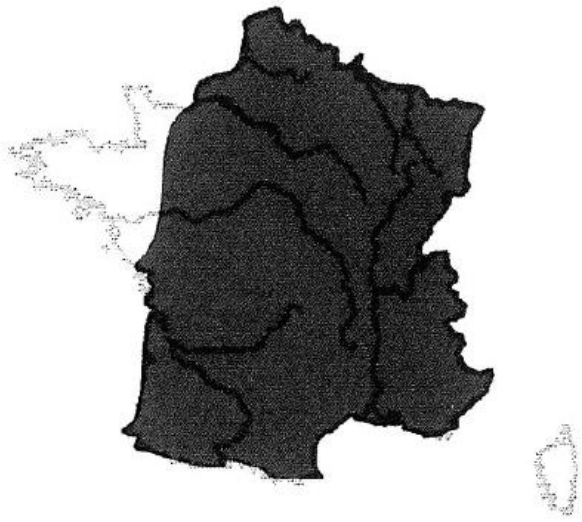


Figure 31 : Distribution de *Gammarus fossarum* en France (Piscart and Bollache, 2012).

Les facteurs abiotiques tels que la salinité, l'oxygène, l'acidité et la pollution jouent un rôle important dans leur distribution (Coulaud et al., 2014; Macneil et al., 1999; Maltby, 1995; Peeters et al., 1998; Verberk et al., 2018). Leur habitat optimal correspond aux eaux alcalines, peu profondes, à forts courants, pauvres en nutriments et riches en oxygène (Peeters et al., 1998). Les

gammares sont des espèces clés dans les systèmes aquatiques de par leur présence en forte densité et leur rôle majeur dans le processus de dégradation des litières. Ils constituent aussi une réserve de nourriture pour les macroinvertébrés, les oiseaux, les poissons, les amphibiens et d'autres crustacés (Gouveia, 2017; Macneil et al., 1999).

Le corps des amphipodes est caractérisé par un aplatissement latéral (Tachet et al., 2000) et est divisé en 4 parties : le prosome (la tête), le mésosome (thorax) et le métasome et l'urosome formant l'abdomen (Piscart and Bollache, 2012) (Figure 32). Le prosome contient le céphalon ainsi que les deux yeux. Dans le mésosome, deux paires de gnathopodes permettent au gammare de se fixer sur un substrat, et les 4 paires de péréiopodes ont une fonction de locomotion. Dans le métasome, les pléopodes sont continuellement agités pour ventiler les cavités branchiales, et jouent un rôle important pour la locomotion.

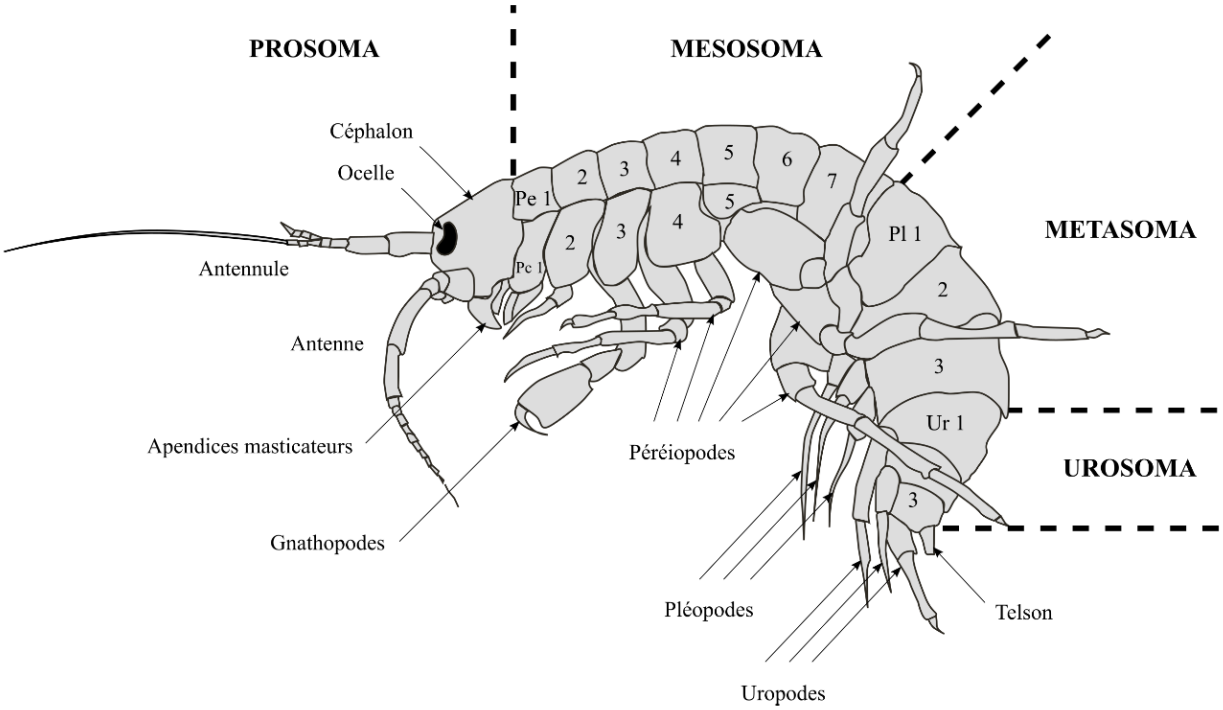


Figure 32 : Morphologie des gammaridés (d'après Cribiu, 2020) ; modifiée d'après (Xuereb, 2009).

La durée d'un cycle de vie de *G. fossarum* peut aller jusqu'à 2 ans (Tachet et al., 2000). Ils ont une croissance discontinue avec des mues successives. *G. fossarum* est une espèce itéropare, dont les femelles se reproduisent plusieurs fois dans l'année (Coulaud et al., 2014; Felten, 2003). Le cycle de mue est parfaitement synchronisé avec le cycle de reproduction chez les femelles (Chaumot et al., 2020; Ratier, 2019). Il se divise en 5 stades : AB pour après la mue, C1 et C2 pour l'inter-mue, D1 et D2 pour la pré-mue (Geffard et al., 2010). Le développement des embryons est aussi divisé en 5 stades (S1-S5) (Figure 33) (Geffard et al., 2010). La température joue un rôle clé sur la durée le cycle de reproduction et de mue, et est corrélée négativement à la durée du cycle. Le cycle de mue dure 30 jours à 12°C, alors qu'il dure environ deux mois à 7°C (Coulaud, 2012; Geffard et al., 2010).

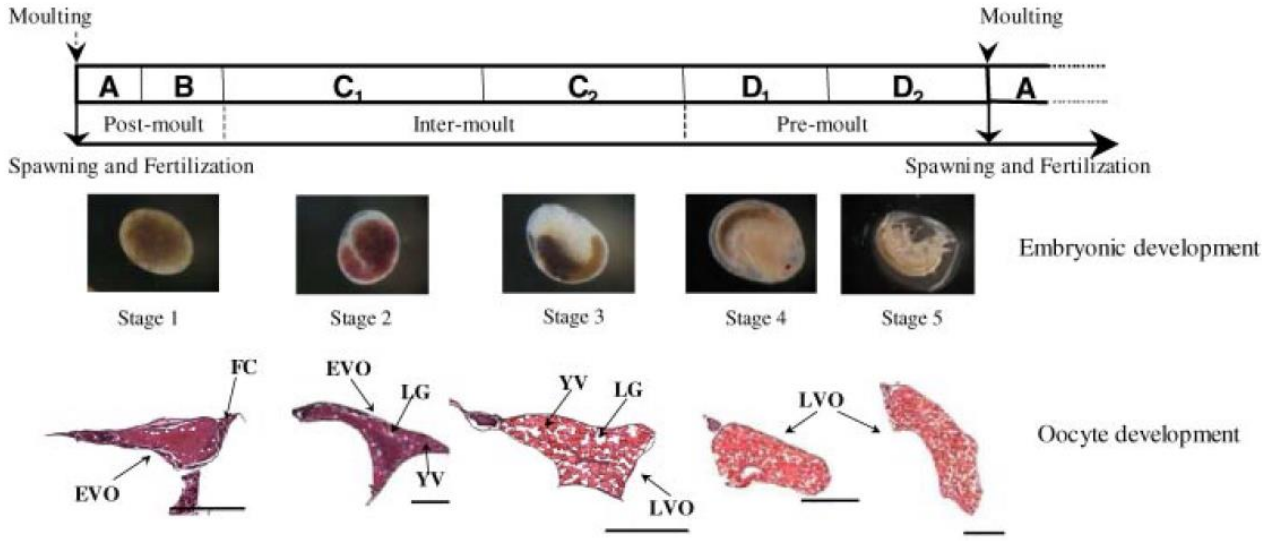


Figure 33 : Développement embryonnaire et phases de mue pour *Gammarus fossarum*. FC : cellules folliculaires, EVO : ovocytes vitellogènes précoces, LVO : ovocytes vitellogènes tardifs, YV : vésicule du vitellus, LG : globule lipidique (Geffard et al., 2010).

4.2 Le gammare en écotoxicologie

4.2.1 Modèle biologique

Les gammares sont largement utilisés comme espèces modèles en écotoxicologie, ils sont (i) très répandus et présents en abondance ce qui permet des comparaisons inter-sites, (ii) leur prélèvement et leur maintien en élevage sont relativement faciles grâce aux connaissances sur leur biologie et écologie, (iii) les observations et expositions peuvent être faites *in situ*, (iv) ils présentent un dimorphisme sexuel permettant la différenciation des genres, (v) leur cycle de reproduction est relativement court, bien connu et documenté (Geffard et al., 2010), (vi) ils passent l'ensemble de leur vie dans l'eau, et (vii) ils sont aussi sensibles à de nombreux stress et accumulent une large gamme de contaminants (Kunz et al., 2010).

4.2.2 Espèce sentinelle

L'étude des organismes sentinelles (au niveau de l'organisme entier, tissulaire et moléculaire) et de leurs biomarqueurs, permettrait de détecter au plus tôt les perturbations observées dans

l'environnement pour en définir la source et les solutions nécessaires (Amiard and Amiard-Triquiet, 2008; Rivière, 1993).

L'utilisation de *Gammarus spp*. pour les études en écotoxicologie a commencé dans les années 1970 (Kunz et al., 2010). Étant sensibles à une large gamme de polluants, elles représentent donc des espèces de choix pour une utilisation en laboratoire et sur le terrain (*in situ*) (Besse et al., 2013; Coulaud et al., 2014; Fialkowski et al., 2003; Geffard, 2014; Gerhardt, 2011; Kunz et al., 2010). Plusieurs biomarqueurs individuels ou subindividuels ont été développés ces dernières décennies chez le gammare (Chaumot et al., 2015; Kunz et al., 2010). On retrouve notamment :

(i) Des biomarqueurs **individuels** de toxicité générale mesurant les impacts des polluants sur les traits d'histoire de vie telle que la reproduction avec des mesures biométriques de la fertilité ou le cycle de mue (Geffard et al., 2010), le taux d'alimentation en mesurant la consommation des feuilles (Coulaud et al., 2011) ;

(ii) Des biomarqueurs **subindividuels** liés à la perturbation endocrinienne avec la quantification de la vitellogénine chez les femelles de *G. fossarum* (Jubeaux et al., 2012; Simon et al., 2010), ou la désynchronisation de processus physiologiques tels que la mue, l'ovogenèse (croissance ovocytaire) et le développement embryonnaire (Geffard et al., 2010) ;

(iii) Des biomarqueurs **subindividuels** liés spécifiquement à un mode d'action tel que la génotoxicité (Lacaze et al., 2011), la neurotoxicité avec l'activité acétylcholine estérase (AChE) (Xuereb et al., 2009) et les changements dans l'activité digestive des enzymes (amylase, trypsine, cellulase, etc.) (Charron et al., 2013; Dedourge-Geffard et al., 2013).

(iv) Ou encore des approches de modélisation de la dynamique des populations comme biomarqueurs **populationnels** (Coulaud et al., 2014). Ces méthodes permettent de prédire des éventuels impacts populationnels d'effet mesurés aux niveaux subindidivuels et individuels (Vigneron, 2015).

4.3 Données omiques disponibles chez *Gammarus spp.*

Malgré l'utilisation de *Gammarus spp.* dans de nombreuses études en écotoxicologie, le génome des espèces les plus utilisées, tel que *G. fossarum* et *G. pulex*, n'est pas disponible et ainsi ils sont considérés comme des organismes non modèles en écotoxicologie moléculaire. Toutefois, l'utilisation des approches omiques chez *Gammarus spp*. est de plus en plus répandue (Armengaud et al., 2014a). Dans la suite du document nous détaillerons les études et données omiques disponibles chez *Gammarus sp.* et un tableau recensant les études omiques disponibles à ce jour se trouve à la fin de ce paragraphe (*Tableau* 7).

4.3.1 Génomes

Aujourd'hui, un seul draft de génome est disponible pour l'espèce *Gammarus lacustris* (Jin et al., 2019) (*Tableau* 7). La taille estimée de ce draft était de 5,07 gigabases (Gb), et couvrait 37,55% du génome estimé (*i.e.* 13,5 Gb). Il contenait 443304 scaffolds avec un N50 de 2578 paires de bases. En comparaison avec les génomes publiés d'autres espèces aquatiques, la qualité du génome préliminaire de *G. lacustris* est relativement faible. La grande taille du génome de *G. lacustris* et sa complexité semble être difficile pour le séquençage de nouvelle génération, ainsi que l'extraction d'ADN de bonne qualité chez cette espèce (Jin et al., 2019).

4.3.2 Transcriptomes

Grâce à la généralisation des nouvelles technologies de séquençage, il est devenu plus accessible de générer les données transcriptomiques d'espèces non modèles, comme *Gammarus*

spp. Plusieurs études transcriptomiques sont disponibles à ce jour pour une dizaine d'espèces de gammaridés (*Tableau 7*). Ces études ont globalement été réalisées afin de fournir des ressources moléculaires à partir desquelles les effets des contaminants peuvent être étudiés.

4.3.3 Métabolomes

Une étude de métabolique ciblée a permis la quantification de 29 métabolites de *G. pulex* par LC-MS/MS (Gómez-Canela et al., 2016) (*Tableau 7*). Il s'agit de la première étude, montrant des altérations des voies liées à la synthèse des protéines, au stress oxydatif et aux cascades de signalisation. Une autre étude a été récemment réalisée chez *G. fossarum* exposé à un mélange de deux médicaments (*i.e.* oxazépam, carbamazépine) en laboratoire (Bonnefoy et al., 2019) (*Tableau 7*). Il a été observé un effet significatif de l'exposition aux contaminants et du genre des individus sur les métabolites mesurés chez *G. fossarum.*

4.3.4 Lipidomes

Un nombre limité de classes de lipides et d'espèces moléculaires ont été décrites chez *G. fossarum* (Arambourou et al., 2018; Kolanowski et al., 2007; Konschak et al., 2021) (Tableau 7). A ce jour, les progrès en MS ont permis de cartographier de manière exhaustive le lipidome de *G. fossarum*, en lien avec le genre et les stades de reproduction des femelles (Fu et al., 2021) (Tableau 7).

4.3.5 Protéomes

Les premiers travaux d'écotoxicoprotéomique étaient basés sur des approches de gel électrophorèse bidimensionnelle (2D) couplées à la spectrométrie de masse en tandem. L'étude de Leroy et al. (2010) s'était intéressée à caractériser l'impact des contaminants PCB sur le protéome de *G. pulex*. En utilisant une approche par 2D-PAGE comme outil de séparation, 560 spots de protéines ont été détectés, avec 21 protéines présentant des différences d'expression significatives entre exposés et contrôles. Bien que cette technique permet d'identifier les modifications post-

traductionnelles, elle permet uniquement l'analyse des protéines connues (*i.e.* annotées) solubles et les plus abondantes (Gouveia et al., 2019b).

D'autres études en protéomique shotgun se sont aussi intéressées à la caractérisation des protéomes de *Gammarus spp*. (Gismondi et al., 2017, 2015; Trapp et al., 2014b). Gismondi et al. ont réalisé une approche protéomique sans gel pour étudier les différences entre les sexes dans les réponses de *G. pulex* exposé au BDE-47 (Gismondi et al., 2015) et la réponse de *G. fossarum* aux expositions chroniques à trois métaux lourds (*i.e.* cuivre, cadmium et plomb) (Gismondi et al., 2017) (Tableau 7). Ces travaux ont identifié 45 et 35 protéines, respectivement, dont 25 et 23 étaient exprimées de manière significativement différente entre les conditions. Dans la première étude, Gismondi et al. (2015) ont mis en évidence des différences dans les réponses entre mâles et femelles, soulignant que le genre pourrait être un facteur confondant dans l'évaluation écotoxicologique. Dans l'étude de Gismondi et al. (2017), la reproduction semble être fortement affectée chez les gammares exposés chroniquement aux métaux lourds. Les niveaux d'expression de protéines impliquées dans le transfert d'énergie et le métabolisme ont montré des remaniements énergétiques pour faire face aux expositions de métaux. Ces résultats soutiennent le fait que les pressions métalliques induisent un coût fonctionnel et énergétique pour les individus de *G. fossarum*.

Une autre étude a également analysé le potentiel de la protéomique comparative comme approche multimarqueurs de la contamination métallique chez *G. pulex* (Vellinger et al., 2016). Bien qu'ayant identifié 264 protéines (dans toutes les conditions étudiées), la majorité était impliquée dans les fonctions de ménage ou le métabolisme énergétique, et n'a pas permis l'identification de protéine spécifique de l'exposition au cadmium.

Le point commun de ces travaux est le nombre limité de protéines identifiées par la spectrométrie de masse (Gouveia, 2017). En effet, ce biais provient généralement de limitations bioinformatiques telles que la recherche d'homologues sur des bases de données existantes avec des espèces éloignées de celle étudiée, ce qui engendre une faible attribution de spectres. Cette limitation technique a été dépassée avec l'application de l'approche de la protéogénomique en 2014, qui couple des méthodes de RNA-Seq et de protéomique shotgun pour la découverte de protéines spécifiques chez des espèces non modèles (Trapp et al., 2014b, p. 2014). La spectrométrie de masse en tandem de nouvelle génération permet de vastes études de protéomes et également des analyses comparatives, où un très grand nombre de protéines est comparé entre diverses conditions, mais à ce jour, elle n'a été utilisée que dans un nombre limité d'études concernant seulement une espèce d'amphipode, *G. fossarum* (Armengaud et al., 2014a; Trapp et al., 2014b, 2018). En effet, afin d'explorer plus en profondeur les mécanismes moléculaires impliqués dans la reproduction de *G. fossarum*, Trapp et al. (2014) ont effectué une première étude protéogénomique permettant d'identifier près de 1873 protéines en spectrométrie de masse, dont 218 spécifiques de l'espèce (Trapp et al., 2014b). Ensuite, une étude protéomique de la réponse des gammares mâles aux perturbateurs endocriniens a conduit à de fortes modulations de 14 protéines « inédites » (ou dites orphelines) spécifiques aux mâles dans les testicules, sur un total de 871 protéines analysées (Trapp et al., 2015). Une autre étude parue en 2018 a évalué la modulation des protéines chez *G. fossarum* exposé au pyriproxyfène, un insecticide analogue de l'hormone juvénile, à différentes doses (Trapp et al., 2018, p. 2018). La comparaison des abondances des protéines a permis d'identifier 32 et 21 protéines modulées entre les conditions et contrôles. Ces études ont permis d'alimenter le catalogue de potentiels candidats de biomarqueurs et de leur application en biosurveillance.

Une autre étude fonctionnelle s'intéressant à l'identification des protéines vitellines présentes dans le protéome femelle de crustacé amphipode a été réalisée (Trapp et al., 2016b). Par similitude avec la génomique comparative, les concepts de core-protéome et de pan-protéome ont été proposés (Trapp et al., 2016b). Le core-protéome correspond aux protéines qui seraient conservées dans toutes les espèces d'une branche d'un arbre phylogénétique donné et produites pour une condition donnée. Le pan-protéome correspond à toutes les protéines qui sont présentes dans une condition pour toutes les espèces d'une branche d'un arbre phylogénétique donné (Trapp et al., 2016b). Cette étude visait l'obtention des informations sur le core protéome des gonades des femelles pour les amphipodes de cinq espèces différentes, toutes du sous-ordre *Senticaudata : G. fossarum, G. pulex, G. roeseli, P hawaiensis* et *H. azteca* (Trapp et al., 2016b).

Chez le mâle et la femelle de *G. pulex,* une première exploration du core-protéome a été réalisée par Cogne et al. (2019a). Les individus provenaient de deux sites différents, permettant ainsi la caractérisation de la divergence potentielle du protéome induit dans un site par une contamination au cadmium biodisponible naturellement. Les protéines impliquées dans la lipidation des protéines, le métabolisme des glucides, la protéolyse, l'immunité innée, la réponse au stress oxydatif et le transport des lipides se sont révélées plus abondantes chez les animaux sur le site exposé au cadmium, tandis que les hémocyanines étaient moins abondantes.

À partir des données expérimentales de protéomique shogun, il est possible de développer une méthode quantitative. La quantification ciblée par spectrométrie de masse en mode *Selected Reaction Monitoring* (SRM), également appelé *Multiple Selected Reaction Monitoring* (MRM), a récemment été proposée comme outil de diagnostic pour mesurer des biomarqueurs et évaluer l'état de santé des organismes sentinelles. Par exemple, chez le crustacé *G. fossarum*, des méthodes de quantification ont été développées et utilisées afin de tester la pertinence d'une

trentaine de protéines comme biomarqueurs spécifiques de l'espèce dans le cadre de programme de biosurveillance des rivières (Charnot et al., 2018, 2017; D. Gouveia et al., 2017; Duarte Gouveia et al., 2017).

Tableau 7 : Synthèse des données omiques disponibles chez les amphipodes.

Omique	Organe	Stress	Espèce	Référence
Génomique	Muscle	Pas de stress	*G. lacustris*	(Jin et al., 2019)
Transcriptomique	Muscle, tête, hépatopancréas, et les gonades (Mâles, femelles, juvéniles)	Parasite intersexuel	*Echinogammarus marinus*	(Short et al., 2014)
Transcriptomique	Céphalons, caeca, oocytes, testicules (Mâles et femelles)	Pas de stress	*G. fossarum*	(Trapp et al., 2014b)
Transcriptomique	Hépatopancréas (Mâles et femelles)	PCB (benzo(a)pyrene)	*G. pulex*	(Gismondi and Thomé, 2016)
Transcriptomique	Embryons	Zone estuarienne	*G. chevreuxi*	(Truebano et al., 2016)
Transcriptomique	Organisme entier (Mâles et femelles)	Habitats (grottes et en surface)	*G. minus*	(Carlini and Fong, 2017)
Transcriptomique	Organisme entier (Mâles)	Hypoxie chronique	*G. chevreuxi*	(Collins et al., 2017)
Transcriptomique	Organisme entier (Femelles si possible)	Pas de stress	Espèces du Baikal	(Naumenko et al., 2017)
Transcriptomique	Muscle	Pression hydrostatique et température	*Eogammarus possjeticus*	(Chen et al., 2019)
Transcriptomique	Organisme entier (Mâles et femelles)	Pas de stress	*Gammarus fossarum A, G. fossarum B, G. fossarum C, G. wautieri, G. pulex, Echinogammarus berilloni, Echinogammarus marinus*	(Cogne et al., 2019c)

Transcriptomique	Organisme entier	Pas de stress	*Eulimnogammarus verrucosus, Eulimnogammarus cyaneus, G. lacustris*	(Drozdova et al., 2019)
Transcriptomique	Tissus internes (Mâles et femelles)	Pas de stress	*G. fossarum*	(Caputo et al., 2020)
Transcriptomique	Organisme entier (Femelles)	Simvastatine	*G. locusta*	(Neuparth et al., 2020)
Transcriptomique	Organisme entier	Phénanthrène (HAP) et acétone (solvant)	*Eulimnogammarus verrucosus, Eulimnogammarus cyaneus, G. lacustris*	(Shatilina et al., 2020)
Métabolomique	Organisme entier	Triclosan, Nimésulide, Propanolol	*G. pulex*	(Gómez-Canela et al., 2016)
Métabolomique	Organisme entier (Mâles et femelles)	Oxazépam, Carbamazépine	*G. fossarum*	(Bonnefoy et al., 2019)
Métabolomique	Organisme entier	Triclosan, Nimésulide, Propanolol	*G. pulex*	(Sheikholeslami et al., 2020)
Lipidomique (composition en AGs)	Organisme entier	Pas de stress	*G. fossarum, G. pulex, G. roeseli, Pontogammarus robustoides, Dikerogammarus haemobaphes*	(Kolanowski et al., 2007)
Lipidomique (composition en AGs)	Organisme entier (Mâles et femelles)	Fenoxycarbe	*G. fossarum*	(Arambourou et al., 2018)
Lipidomique (composition en AGs)	Organisme entier	Régime alimentaire	*Gammarus sp.*	(Kühmayer et al., 2020)
Lipidomique (composition en AGs)	Organisme entier (Mâles et femelles)	Azoxystrobine (strobilurine)	*G. fossarum*	(Konschak et al., 2021)
Lipidomique (composition en AGs)	Organisme entier	Déchets aquaculture	*G. insensibilis*	(Jiménez-Prada et al., 2021)
Lipidomique (shotgun)	Organisme entier (Femelles)	Pas de stress	*G. fossarum*	(Fu et al., 2020)

Lipidomique (shotgun)	Organisme entier (Mâles et femelles)	Pas de stress	*G. fossarum*	(Fu et al., 2021)
Protéomique	Organisme entier	PCB	*G. pulex*	(Leroy et al., 2010)
Protéomique	Organisme entier (Mâle, femelles et embryons)	Hormones de crustacés (20-hydroxyecdysone, methylfarnesoate) et insecticide (methoxyfénoside), fongicide (propiconasole) et produits pharmaceutiques (benzophenone, carbamazépine, cyproterone, R-propanolol)	*G. fossarum*	(Jubeaux et al., 2012)
Protéomique	Organisme entier (Mâle et femelles)	BDE-47	*G. pulex*	(Gismondi et al., 2015)
Protéomique	Organisme entier (Mâles)	Cadmium, Arsenate	*G. pulex*	(Vellinger et al., 2016)
Protéomique	Organisme entier (Mâles et femelles)	Cadmium, Cuivre, Plomb	*G. fossarum*	(Gismondi et al., 2017)
Protéomique (shotgun)	Gonades, céphalons, caeca (Mâles et femelles)	Pas de stress	*G. fossarum*	(Trapp et al., 2014b)
Protéomique (shotgun)	Testicules (Mâles)	Cadmium, Pyriproxyfène, Méthoxyfénoside	*G. fossarum*	(Trapp et al., 2015)
Protéomique (shotgun)	Ovaires (Femelles)	Pas de stress	*G. fossarum, G. pulex, G. roeseli, P hawaiensis et H. azteca*	(Trapp et al., 2016a)
Protéomique (shotgun)	Embryons et ovaires (Femelles)	Pas de stress	*G. fossarum*	(Trapp et al., 2016b)
Protéomique (shotgun)	Testicules (Mâles)	Pyriproxyfen	*G. fossarum*	(Trapp et al., 2018)
Protéomique (shotgun)	Organismes entier (Mâles et femelles)	Cadmium biodisponible	*G. pulex*	(Cogne et al., 2019a)

Protéomique (ciblée)	Organisme entier (Mâles et femelles)	Sites contaminés	*G. fossarum*	(Duarte Gouveia et al., 2017)
Protéomique (ciblée)	Organisme entier (Mâles et femelles)	Cadmium, Plomb	*G. fossarum*	(D. Gouveia et al., 2017)
Protéomique (ciblée)	Organisme entier (Mâles et femelles)	Pas de stress	*G. fossarum*	(Charnot et al., 2017)
Protéomique (ciblée)	Organisme entier	Pas de stress	*G. fossarum*	(Charnot et al., 2018)
Protéomique (ciblée)	Organisme entier (Mâles et femelles)	Pas de stress	*G. fossarum*	(Faugere et al., 2020)

5. OBJECTIFS ET DEMARCHE EXPERIMENTALE

Cette thèse s'inscrit dans la continuité des développements de connaissances moléculaires autour de l'espèce non modèle *Gammarus fossarum*, portés par le laboratoire d'écotoxicologie d'INRAE Villeurbanne.

L'objectif général du projet consiste en l'utilisation d'approches bioinformatiques pour l'analyse de données multiomiques (notamment transcriptomique et protéomique) chez les espèces non modèles et au contournement des limitations de l'annotation fonctionnelle dues aux bases de données existantes qui ne représentent que des espèces éloignées. Dans le cadre de cette thèse, nous mettons à l'œuvre une stratégie pour décrire et caractériser les voies métaboliques impliquées dans la biosynthèse, le transport et la dégradation des lipides chez l'espèce sentinelle *G. fossarum*.

5.1 La biologie des systèmes pour l'analyse de données de protéogénomique pour la compréhension de la toxicité moléculaire liée à l'activité testiculaire de *G. fossarum*

Les analyses des réseaux de coexpression se sont montrés utiles pour la compréhension de l'information biologique contenue dans les données omiques. Dans ce contexte, nous cherchons à

démontrer que les approches d'analyse des réseaux sans *a priori* (dites « data driven ») fournissent une méthodologie pertinente pour étudier les ensembles de données protéomiques de *G. fossarum* déjà existants. C'est dans ce cadre que l'analyse des réseaux de coexpression (weighted gene coexpression network analysis, WGCNA) proposée par Langfelder et Horvath (2008) a été adaptée aux jeux de données de protéomique shotgun. L'analyse de réseaux de coexpression est une méthode d'exploration de données qui va être testée pour identifier et mieux comprendre le rôle des protéines dans la reproduction et dans la réponse aux contaminants perturbateurs endocriniens. De plus, nous avons testé la pertinence de l'analyse de coexpression pour proposer des mécanismes d'action des différents contaminants en cas d'exposition au laboratoire.

5.2 Exploitation de données transcriptomiques et protéomiques de mâle et femelle de *G. fossarum* pour la caractérisation du métabolisme lipidique

L'objectif de cette deuxième partie est la caractérisation du métabolisme lipidique en utilisant des ressources multiomiques. Pour cela, nous avons dans un premier temps cherché à démontrer la faisabilité de l'adaptation de l'outil CycADS, originalement destiné à l'annotation des voies métaboliques à partir des données génomiques (Vellozo et al., 2011), à nos données transcriptomiques à disposition pour *G. fossarum* (Cogne et al., 2019c). Dans un second temps, nous avons intégré les données issues de la protéogénomique d'organes de *G. fossarum* pour valider les enzymes identifiées par spectrométrie de masse avec les données d'annotation fonctionnelle des transcrits. Les connaissances sur les voies impliquées dans le métabolisme lipidique chez cette espèce non modèle sont encore limitées. Nous nous sommes donc placés dans ce cadre pour identifier sans hypothèse *a priori*, les acteurs clés du métabolisme lipidique de *G. fossarum.* La reconstruction des voies métaboliques impliquées dans l'homéostasie des lipides et

des espèces lipidiques présentes chez les espèces sentinelles pourrait faciliter la compréhension de leurs implications dans la réponse à des contaminants.

CHAPITRE II – L'ANALYSE DE RESEAUX DE COEXPRESSION EN PROTEOMIQUE POUR L'ECOTOXICOLOGIE

Ce chapitre regroupe les résultats qui illustrent l'intérêt d'utiliser des approches de biologie des systèmes avec des organismes sentinelles non modèles, ne disposant pas de génome annoté, pour identifier les acteurs moléculaires (*i.e.* protéines) régulant la physiologie et/ou la réponse aux contaminants. Plus précisément, les deux publications montrent les résultats et l'originalité de l'utilisation d'une analyse de réseau de coexpression sur des données de protéomique shotgun déjà disponibles chez le crustacé d'eau douce *G. fossarum*. La première publication (voir partie 1) s'inscrit dans le cadre de la modulation physiologique de la reproduction de cette espèce d'amphipode en analysant les profils protéomiques des gonades mâles et femelles à différents stades de maturation et les embryons au cours de leur développement. Le deuxième article (voir partie 2) traite de la toxicité testiculaire liée à trois contaminants (*i.e.* cadmium, pyriproxyfène, méthoxyfénoside) à partir de données provenant d'une précédente étude de Trapp et al. (2015). Cette étude antérieure n'avait pas réussi à identifier de protéines spécifiques des mécanismes d'actions (MoA) des contaminants en utilisant une approche comparative (*i.e.* analyse différentielle) et nous avons testé la pertinence de l'analyse de coexpression pour l'identification des MoA specifiques de differents perturbateurs endocriniens potentiels.

1. DECOUVERTE DES PROCESSUS BIOLOGIQUES CLES LIES A LA REPRODUCTION CHEZ LE GAMMARE

1.1 Synthèse

Les technologies de séquençage et de spectrométrie de masse de nouvelles générations ont récemment élargi la disponibilité des transcriptomes et des protéomes. En effet la protéogénomique permet de coupler les données de RNA-Seq et de protéomique shotgun, une alternative pour les espèces dépourvues de génome annoté. Une base de données protéiques théorique (ORFs) est générée à partir de la traduction *in silico* dans les 6 cadres de lectures des séquences d'ARNm. Cette base de données théorique et spécifique de l'organisme étudié est ensuite utilisée pour l'identification des spectres MS/MS de la protéomique shotgun, qui ne seraient pas présents dans les bases de données protéomiques disponibles. Cependant, la nature partielle des données transcriptomiques et protéomiques (*i.e.* qui ne couvrent pas la totalité des acteurs moléculaires) et les incertitudes concernant les annotations fonctionnelles notamment chez les organismes non modèles réduisent la capacité d'interprétation biologique des données. Les approches basées sur les réseaux peuvent aider à extraire des informations biologiques importantes des jeux de données omiques. En effet, l'analyse de réseaux de coexpression permet de regrouper les protéines dans des modules de coexpression et d'identifier les protéines centrales (dites « hub ») de ces modules. Il est ensuite possible d'identifier les relations entre les modules et les traits d'histoires de vie ou d'exposition aux contaminants.

Dans cette partie, nous avons utilisé pour la première fois chez *G. fossarum*, **une analyse de réseau de coexpression pour étudier l'organisation des protéines dans les organes reproducteurs et les embryons à partir de données de protéomique shotgun disponibles.** Nous avons illustré l'utilisation de l'outil « Weighted Gene Coexpression Network Analysis », (*i.e.* WGCNA) pour l'identification des protéines régulant la maturation des gonades et le

développement embryonnaire chez le crustacé d'eau douce *Gammarus fossarum*. Les jeux de données protéomiques comprenaient des échantillons de testicules, d'ovaires ou d'embryons à différents stades de maturation ou de développement, et en cinq réplicats par condition. L'abondance des protéines a été mesurée en spectral count, ainsi la table de comptage comprenait au départ 1199 protéines.

Nous avons identifié un module de protéines coexprimées, associé aux embryons et corrélé avec les stades de développement correspondant à l'organogenèse chez les embryons, enrichi en protéines impliquées dans l'édition et l'épissage de l'ARN. Un module associé aux ovaires était enrichi en protéines de type vitellogénine et en protéines coagulables, confirmant la diversité des protéines de transfert des grands lipides impliqués dans la maturation des ovocytes chez cet amphipode d'eau douce. De plus, nos résultats ont mis en évidence une régulation fine entre la production d'énergie par la glycolyse et les événements dépendants de l'actine-myosine dans la spermatogenèse de *G. fossarum*.

Cette étude illustre l'intérêt d'appliquer des approches de biologie des systèmes à des espèces non modèles afin d'améliorer la compréhension des mécanismes moléculaires régulant des événements physiologiques importants ayant une pertinence écologique. Elle illustre également une approche efficace pour exploiter et retirer un maximum d'informations disponibles dans les grands jeux de données, dont l'exploitation un à un limite l'acquisition de connaissances.

1.2 Article n°1 : Coexpression network analysis identifies gonad- and embryo-associated protein modules in the sentinel species *Gammarus fossarum*

SCIENTIFIC REPORTS

OPEN

Co-expression network analysis identifies gonad- and embryo-associated protein modules in the sentinel species *Gammarus fossarum*

Received: 4 October 2018
Accepted: 10 May 2019
Published online: 27 May 2019

Davide Degli Esposti[1], Christine Almunia[2], Marc-Antoine Guery[1], Natacha Koenig[1], Jean Armengaud[2], Arnaud Chaumot[1] & Olivier Geffard[1]

Next generation sequencing and mass spectrometry technologies have recently expanded the availability of whole transcriptomes and proteomes beyond classical model organisms in molecular biology, even in absence of an annotated genome. However, the fragmented nature of transcriptomic and proteomic data reduces the ability to interpret the data, notably in non-model organisms. Network-based approaches may help extracting important biological information from -omics datasets. The reproductive cycle of the freshwater crustacean *Gammarus fossarum*.provides an excellent case study to test the relevance of a network analysis in non-model organisms. Here, we illustrated how the use of a co-expression network analysis (based on Weighted Gene Co-expression Network Analysis algorithm, WGCNA) allowed identifying protein modules whose expression profiles described germ cell maturation and embryonic development in the freshwater crustacean *Gammarus fossarum*. Proteome datasets included testes, ovaries or embryos samples at different maturation or developmental stages, respectively. We identified an embryonic module correlated with mid-developmental stages corresponding to the organogenesis and it was characterized by enrichment in proteins involved in RNA editing and splicing. An ovarian module was enriched in vitellogenin-like proteins and clottable proteins, confirming the diversity of proteins belonging to the large lipid transfer family involved in oocytes maturations in this freshwater amphipod. Moreover, our results found evidence of a fine-tuned regulation between energy production by glycolysis and actin-myosin-dependent events in *G. fossarum* spermatogenesis. This study illustrates the importance of applying systems biology approaches to emergent animal models to improve the understanding of the molecular mechanisms regulating important physiological events with ecological relevance.

Advances in nucleic acid sequencing technologies and mass spectrometry have recently increased the availability of whole transcriptomes and proteomes beyond classical model organisms, even in absence of an annotated genome[1]. The use of proteogenomics approaches, that couple species-specific RNA-sequencing followed by the acquisition of shotgun proteomic data by high-resolution mass spectrometry for a straightforward interpretation of peptide spectra, has opened the way to get insights into the molecular mechanisms involved in the physiology and the response to environmental stress in many species of ecological relevance[1,2].

However, the fragmented nature of proteomic and transcriptomic data without a reference genome usually prevents extensive interpretation of the data. This is particularly true in the context of the molecular physiology of non-model organisms whose genomes are evolutionarily distant from well-annotated genomes, such as those of the fly *D. melanogaster* or the human genome. Network-based approaches provide an excellent methodological

[1]Irstea, UR RiverLy, Ecotoxicology Team. Centre de Lyon-Villeurbanne, 5 rue de la Doua CS 20244, 69625, Villeurbanne, France. [2]Laboratoire Innovations technologiques pour la Détection et le Diagnostic (Li2D), Service de Pharmacologie et Immunoanalyse (SPI), CEA, INRA, F-30207, Bagnols sur Cèze, France. Correspondence and requests for materials should be addressed to D.D. (email: davide.degli-esposti@irstea.fr)

framework to investigate the rich content of information from omics datasets[3,4]. Weighted Gene Co-expression Network Analysis (WGCNA) is a systems biology method originally conceived for describing correlation patterns among genes across microarrays data[4]. One of the main advantages of this statistical model is to be data driven and it represents a great opportunity in non-model organisms for which genome sequences, gene annotations, or functional pathways are not still available. WGCNA has been successfully used also for count-based (i.e. RNA-Seq based) gene expression data[5,6]. However, shotgun proteomics data have been poorly explored using WGCNA to date, despite the possibility to apply this approach to the spectral count metric largely used in this phenotypic methodology[7]. Concurrently, gene co-expression network analyses of transcriptomic data have provided useful insights in understanding fundamental biological processes in many arthropod species. For instance, WGCNA provided evidence of conserved pathways across 16 different species of ants involved in reproductive division of labor[8]. Moreover, a network analysis highlighted the role of metabolic and cell signaling genes in the stress response in the model crustacean *Daphnia magna*, when simple differential expression analysis failed to find meaningful biological answers[9].

Our group has provided the first proteomic-based description of the reproductive systems of the freshwater amphipod *Gammarus fossarum*[2,10]. *G. fossarum* is an established and ecologically relevant test species and it has been extensively used in ecotoxicology to monitor the quality of freshwater bodies[11–13].

Morphological and biochemical alterations in the gammarid female or male gonads are currently used as endpoint in ecotoxicology[14,15]. Testes have been reported to be particularly sensitive to genotoxic insult and minimal sperm DNA damage induced an increased abnormality rate in embryos, with potential consequences for population dynamics of this species[16]. Morphological abnormalities in embryonic development were also observed after parental exposure to heavy metals (Cd), or various organic compounds (such as nonylphenol or phenoxycarb)[14,17]. However, we do not have any detailed knowledge on the metabolic and signaling pathways underlying the full reproductive process in *G. fossarum*. The aim of this work is to develop systems biology approaches for identifying fundamental pathways and key hub proteins involved in the reproductive process and that could provide predictive biomarkers of physiological impairment following contaminant exposure. In particular, we aimed to use a co-expression network analysis through a pooled proteomics dataset including testes, ovaries or embryos at different maturation or developmental stages to identify protein pathways regulating germ cell maturation and embryonic development of *G. fossarum*. This provides an excellent case study to test the relevance of a network analysis in non-model organisms, since the structure and the molecular physiology of sperms, oocytes, and embryos are highly differentiated.

Materials and Methods

Proteomics datasets. We retrieved protein abundance data measured as spectral counts (SC) from two published shotgun proteomics datasets obtained by high-resolution tandem mass spectrometry using a LTQ-Orbitrap XL instrument. Only proteins validated with at least two different peptides were retained[2,10]. The datasets were merged in a matrix $n \times m$, with n indicating the total number of proteins and m the total number of samples. Finally, the matrix consisted in 1,199 protein abundance data expressed as SC for 85 samples (Supplementary Data 1). The samples included *G. fossarum* ovaries (individual whole paired organs) and embryos (6 embryos per female) proteomes[10] at different stages of oocyte maturation and embryo development[10,14]. Five biological replicates for each stage (AB, C1, C2, D1, D2 for ovaries and S1 to S5 for embryos) were included in this dataset. Testes were individually sampled from male organisms at 6 stages of spermatogenesis, namely at pre-copula stage (mature male in amplexus at the end of the spermatogenesis), and days after copulations (J0, J1, J2, J3, J4, J7)[2,15]. At J7 (7 days after copulation), the spermatogenesis cycle is completed. Five biological replicates for each spermatogenesis stage were included in this dataset. The annotation of the 1,199 proteins was updated compared with the original studies, by using a semi-automatic pipeline using DIAMOND[18]. Briefly, protein sequences identified by peptides that matched an ORF entry of the GFOSS RNA-seq database[2] were blasted against the NCBInr and Swissprot databases. The closest predicted protein (indicated with NCBInr ID or SwissProt ID) and its E-value were associated with the corresponding GFOSS contig. The Gene Ontology (GO) annotation of proteins homolog found in Swissprot database was performed using data provided by the GO Consortium through the GO website and the web application Amigo (http://amigo.geneontology.org). Two levels (1 and 2) of GO annotation were extracted.

Data pre-processing and normalization. In order to identify possible outliers, sample clustering was performed using the hierarchical clustering function implemented in the lumi R package. One sample (AB-n1) was identified as an outlier and thus excluded from the final dataset. In particular, this sample was characterized by the lowest number of total measured SC. Principal Component Analysis (PCA) was used to analyze and identify the variables explaining the maximum variance associated to the proteomic data of the different developmental stages of *G. fossarum* gonads and its embryos. In order to avoid the noise and poor reproducibility associated with low abundant proteins, proteins with less than three SC in the most abundant condition were filtered out[19]. Finally, a dataset comprising 84 samples and 375 proteins was retained for further analyses. Before network modeling, counts need to be normalized across samples[4,7]. Normalization was performed using the *calcNormFactors* function from the R package edgeR[20]. While this function has been originally implemented for RNA-Seq count data, it applies equally well to spectral count data[19]. The *calcNormFactors* function automatically adjusts for the total number of spectral counts per sample (also known as original library size in RNA-Seq experiments) and normalizes for protein composition, taking into account the fact that highly abundant proteins tend to contribute more peptide/spectra than those lowly abundant[20]. *calcNormFactors* calculate a normalization factor that was applied to normalize SC values of each protein in each sample.

Network analysis. The network properties of the proteome of the reproductive system of *G. fossarum* were analyzed using the R package Weighted Gene Coexpression Network Analysis (WGCNA)[3,4]. Shotgun proteomics gives access to a count-based quantification of peptides that can be aggregated to a contig/protein level when a species-specific RNA-Seq database is available, as it is the case for the datasets analyzed in this study. Thus, we used the WGCNA package to investigate the count-based proteomic database of gonadic and embryonal tissues obtained from the freshwater amphipod *G. fossarum*. WGCNA was used to construct a protein co-expression network, identify protein modules in the network and correlate the identified modules to external information, such as tissue and stage. In particular, network dendrogram was built using the *blockwiseModules* function, choosing an "unsigned" network type in order to keep relationships of negatively correlated proteins and a "signed" topological overlap matrix (TOM) in order to subtract connections affected by noise[3]. TOM values are a measure of proximity (i.e. interconnection) between expressed proteins. They range from 0 to 1, with 1 indicating maximal proximity, i.e. very high interconnection. We set 20 the minimum number of proteins forming a module. Proteins outside of any modules (indicating low co-expression) were combined together in a grey module. Different clustering methods (static, dynamic, hybrid and dynamic-hybrid methods) were used to delimit the network into distinct modules[21]. Module eigengenes were calculated using the *moduleEigengenes* function. Module eigengenes are defined as the principal component of each module and they represent the protein expression profile in a module. Module eigengenes allow to explore how related the modules are and the correlation between modules and phenotypic traits, such as tissue type or maturation/developmental stages. Network properties, such as total (k_{total}) or intramodular (k_{within}) connectivity or module membership (ME) have been used to identify proteins that show the high degree of connectivity within a module (hub proteins). Due to their central position in the network, hub proteins are expected to play important biological role within their module.

Reproducibility. An R script combining the whole analysis pipeline is available as Supplementary Document 1. Spectral count data and the associated metadata are available as Supplementary Data 1 and Supplementary Data 2, respectively.

Results

Global proteomics profiles distinguish testes, ovaries and embryos in *Gammarus fossarum*. In order to have a systemic view of the molecular processes in the reproductive organs and in the embryonic development of the sentinel species *G. fossarum*, we analyzed proteomics data obtained from individually sampled testes and ovaries at different maturation status and from individually sampled embryos obtained at different developmental stages (see Material and Methods section). Sample clustering and Principal Component Analysis (PCA) based on raw data did not change after filtering and normalization (Fig. 1A,C for raw data, Fig. 1B,D for normalized samples). The proteomes of the testes and ovaries were clearly separated and the difference across tissues accounted for up to 72% of the variance in the data (Fig. 1C,D). However, testes appeared to cluster tightly, independently from the maturation stage, while ovaries showed a much higher variation in their proteomic profiles during the maturation process. The protein profiles of embryos showed also a high level of variability, essentially explained by between-stage variability distributed along the embryonic development. Early stage embryo profiles (S1 and S2) clustered close to the vitellogenic ovary profiles (C1-D2 stage) while middle stages (S3 and S4) clustered independently and the last embryonic stage (S5) clustered even further (Fig. 1C,D).

Distinct modules of co-expressed proteins are identified in the proteome of the reproductive system of *Gammarus fossarum*. We used four different clustering methods to explore the protein network of the gonads and embryos. Figure 2 shows the differences in terms of protein clustering between the four methods. The hybrid method (indicated as *blockwise Colors)*, the dynamic method and the mixed dynamic-hybrid method identified four distinct modules with high overlap among them. In order to minimize spurious associations induced by forcing proteins into proper modules but at the same time to maximize the sensitivity in detecting biologically relevant modules, the clustering performed by the dynamic method was retained for further analyses. On this basis, the four distinct modules were identified as a yellow module (31 proteins), a turquoise module (112 proteins), a blue module (74 proteins) and a brown module (32 proteins) (Fig. 2). The values of topological overlap among the proteins of the network showed that the yellow and the blue modules clustered distinctly each other and from the other modules (Fig. 3A). In order to quantify the similarity between modules and to see how the different tissues fit the eigengene network, we used the module eigengenes. Figure 3B shows the yellow module co-clustered exclusively with the embryonic tissues while the blue module clustered only with the ovaries. Interestingly, the two other modules (turquoise and brown) were close each other and co-clustered with testes (Fig. 3A,B). In this context, it was possible to associate each tissue or organ to one or two distinct co-expressed modules, independently from their maturation stage, identifying a previously unappreciated organization of the protein expression profiles in the reproductive tissues and embryonic development of *G. fossarum*[2,10].

Co-expressed modules mirrors gonad maturation and embryos developmental stages. By correlating the module eigengenes with each stage, we observed that two embryonic mid-developmental stages (S3 and S4) were the main drivers of the association between the yellow module and the embryonic traits, with positive correlation values of 0.59 and 0.66, respectively (p-values = $3*10^{-9}$ and $6*10^{-12}$ respectively) (Fig. 4). On the contrary, the earliest stage of embryonic development (S1) showed no correlation with the yellow module (−0.13, p-value = 0.2), while it was weakly but significantly correlated (0.27, p-value = 0.01) with the ovarian-associated module (blue), reflecting its cytoplasmic inheritance from the egg. Concerning the ovaries, all stages were positively correlated with the blue module, except the earliest stage (AB). All stages from C2 to D2 are mainly involved in vitellogenesis, indicating that the blue module contains mainly proteins involved in this physiological pathway. Moreover, the blue module was negatively correlated to (i.e. repressed in) testes (−0.73, p-value = $3*10^{-15}$),

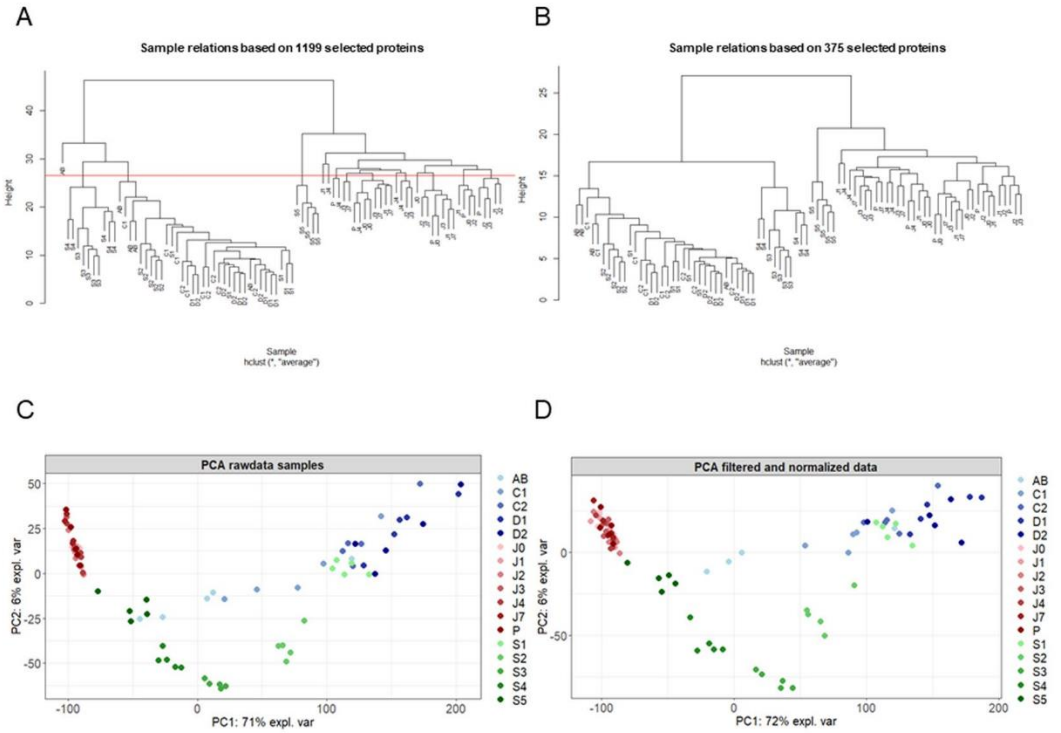


Figure 1. Proteomic profiles clearly distinguish testes, ovaries and embryos of *G. fossarum*. Blue dots indicate the different molt stages in females (AB, C1, C2, D1 and D2). Red dots indicate the testes at different maturation stages: P indicates testes in the pre-copula stage. J0, J1, J2, J3, J4, J7 indicate the days after copulation. Green dots indicate the different embryonic development stages (S1, S2, S3, S4, and S5). (**A**) Hierarchical clustering of raw data. (**B**) Hierarchical clustering after filtering and normalization. (**C**) Principal component analysis of raw data. (**D**) Principal component analysis after filtering and normalization.

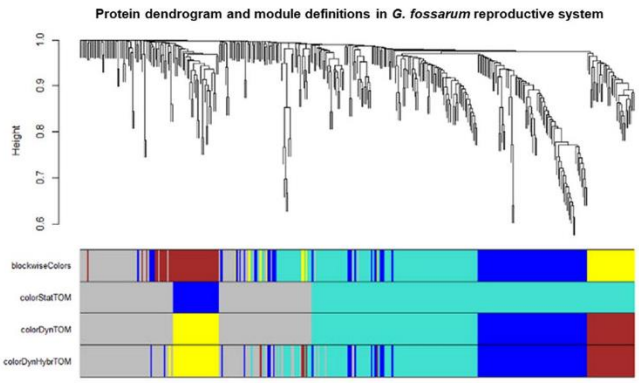


Figure 2. Distinct protein modules can be identified using co-expression network analysis. Protein dendrogram and module delimitation based on different clustering methods (see Materials & Method section). Modules group the proteins that have a high level of co-expression, based on pair-wise correlations between protein abundance. Four protein modules (indicated by colors: yellow, turquoise, blue or brown) were identified by 3 different clustering methods.

indicating that this pathway is strictly controlled in male gonads. Similarly, the brown and turquoise modules were highly correlated with the testes and significantly repressed in both embryos and ovaries (Fig. 4). Finally, even proteins that are not strongly co-expressed (grey module) were found to have an expression profile similar to the turquoise module, with high positive correlation to the testes and negative correlation to the ovaries. This is likely due to the choice of a conservative clustering in defining the co-expression modules. Indeed, a few proteins showed also a relatively high value of module membership (ME > 0.60) for the turquoise module (Supplementary Document 2) and could be probably involved in the same biological pathway activated in the testes and concomitantly repressed in the ovaries.

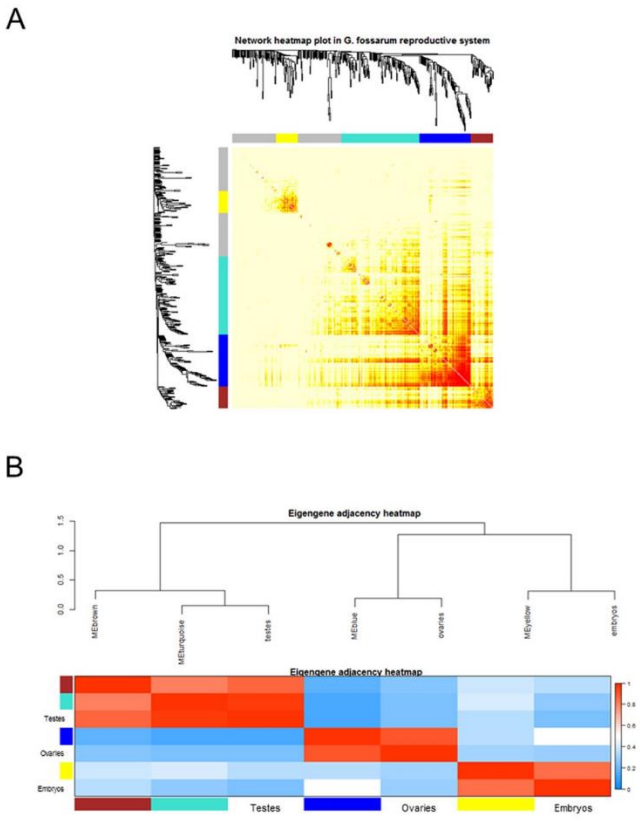


Figure 3. Distinct modules interact differently each other. (**A**) The heatmap shows the Topological Overlap Matrix (TOM) values among the proteins of the network delimited in modules by the dynamic method. Light yellow color represents low overlap (*i.e.* low interconnection), while dark red color represents high overlap (*i.e.* high interconnection). (**B**) The module eigengene adjacency showed by hierarchical clustering and heatmap. A module eigengene summarizes the protein expression profile of each module. Testes cluster together with the turquoise and brown module eigengenes, embryos with the yellow one and ovaries with the blue one.

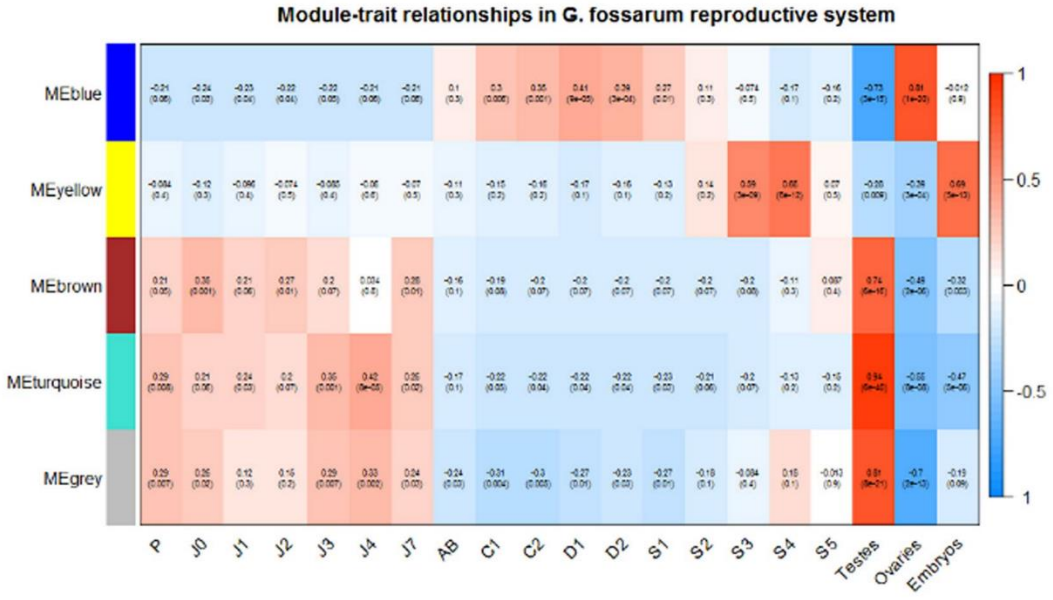


Figure 4. Testes, ovaries and embryos of *G. fossarum* are distinctively associated with different protein modules. Each row corresponds to a module eigengene, each column to a different developmental (for the embryos) and maturity (for the gonads) stages (on the left side) or different organ/tissue (on the right side). In each cell, the correlation value and the p-value (in parenthesis) for each module-stage association are shown.

Modules	Large Lipid Transport Proteins		Cell Division	Actin-Myosin	Energy metabolism	RNA processing†	Protein synthesis
	Vitellogenin-like	Clottable protein-like					
Blue	**14 (19)****	8 (11)‡	4 (5)	3 (4)	4 (5)	0 (0)	2 (3)
Yellow	1 (3)	2 (6)	3 (10)	1 (3)	1 (3)	**7 (23)****	**11 (35)****
Brown	0 (0)	0 (0)	1 (3)	**27 (84)****	0 (0)	0 (0)	0 (0)
Turquoise	3 (3)	3 (3)	10 (9)	15 (13)	**28 (25)****	0 (0)	15 (13)

Table 1. Number and percentage (in parentheses) of proteins belonging to the main molecular functions identified in the protein network. Comparison of relative abundance was made among the 4 co-expression modules. P-values were calculated with Fisher exact test. (**p-value < 0.01). †We included in this category proteins involved in transcriptional and splicing regulation or mRNA transport. ‡p-value = 0.05025.

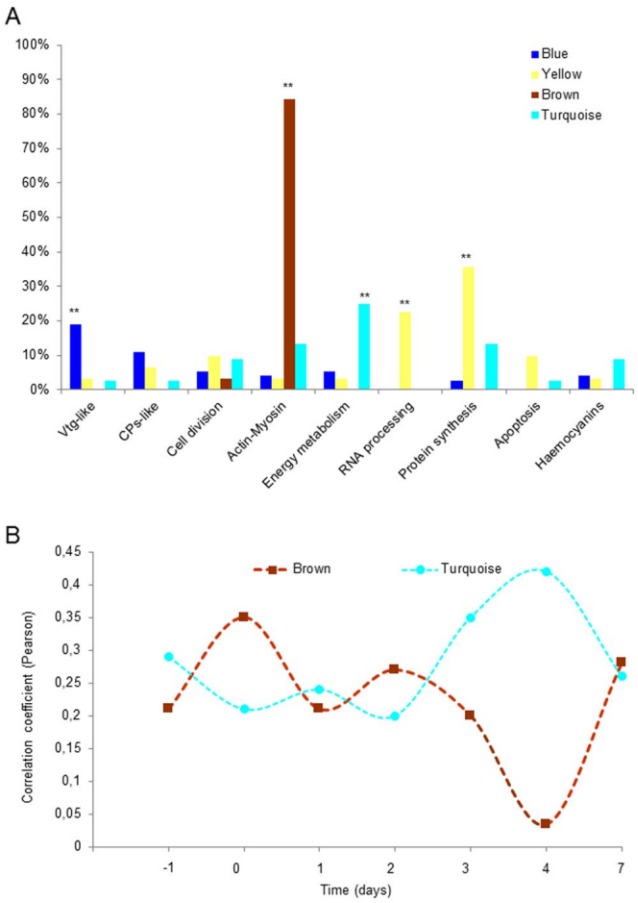


Figure 5. Different protein modules are differentially enriched in distinct molecular functions. (**A**) Percentage of modules' proteins for each molecular function identified. **p-value < 0.01, Fisher's exact-test. (**B**) Temporal variation of the correlation coefficient between brown module and turquoise module eigengenes and the phenotypic trait "testes". The results show that the two modules are highly interconnected and their expression profiles regulated in a tightly coordinate manner.

Module protein composition and hub proteins provide new insights in the reproductive biology and embryonic developmental processes in *Gammarus fossarum*. We looked at the predicted biological role of the different proteins that clustered into the same co-expression module in order to investigate the biological significance of the identified co-expressed modules. We used the output of the NCBInr and Swissprot BLAST results to associate each protein to one category of the following molecular functions or protein families: the Large Lipid Transport Protein superfamily which included both Vitellogenin-like proteins (Vtg-like) and clottable protein-like (CPs-like), cell cycle and division, actin-myosin families, energy metabolism, RNA processing (including proteins regulating the transcriptional process, RNA splicing and mRNA transport to cytoplasm), protein synthesis (including riboproteins and factors involved in protein translation), apoptosis (including both potential activators and inhibitors), and haemocyanin-like proteins. The module correlated with the embryos (yellow) had a significant higher proportion of proteins involved in RNA processing (23%) and protein synthesis (35%) compared with the other modules (Table 1 and Fig. 5A, p-value < 0.01, Fisher's exact test), suggesting an

Contig N	Module	Swissprot ACC	NCBInr ACC	E-value†	Organisms††	k_{total}	k_{within}	Module Membership
199971_fr3	Blue	Q94637	XP_018007759.1	1.40E-28	*H. atzeca*	15.340	13.757	0,97
276_fr4	Blue	O94518	XP_018025391.1	1.00E-10	*H. atzeca*	14.672	13.362	0.96
30649_fr2	Blue	Q868N5	AHK05984.1	9.00E-115	*G. marinus*	13.918	12.428	0.95
39606_fr3	Blue	—	XP_018007759.1	2.80E-96	*H. atzeca*	19.291	13.048	0.95
67263_fr2	Blue	Q17RH7	XP_017986259.1	0.049	*E. sinecaudum*	13.666	12.186	0.94
121249_fr2	Yellow	P51400	XP_018009377.1	2.60E-42	*H. atzeca*	2.409	2.259	0.89
203833_fr2	Yellow	P48810	XP_018010441.1	3.70E-97	*H. atzeca*	2.359	2.090	0.88
201834_fr4	Yellow	—	XP_018025391.1	2,00	*H. atzeca*	1.968	1.858	0.87
122312_fr4	Yellow	P49041	XP_018016935.1	5.40E-26	*H. atzeca*	2.079	1.912	0.87
153160_fr6	Yellow	P21895	XP_018020084.1	3.80E-45	*H. atzeca*	2.060	1.671	0.86
20975_fr4	Brown	Q24756	AFP95338.1	2,20E-36	*P. clarkii*	4.582	3.378	0.91
194796_fr2	Brown	P05661	BAK61430.1	1,40E-78	*M. japonicus*	6.384	3.384	0.90
37276_fr6	Brown	P05661	XP_018022402.1	1,70E-70	*H. atzeca*	9.565	3.522	0.90
182086_fr1	Brown	P05661	XP_018022403.1	7,60E-83	*H. atzeca*	3.602	3.019	0.89
48693_fr3	Brown	P30163	XP_020899190.1	1,50E-52	*E. pallida*	10.374	3.416	0.89
191065_fr3	Turquoise	P11979	XP_018006716.1	1,70E-32	*H. atzeca*	11.387	7.798	0.94
141775_fr4	Turquoise	Q5R2J2	CAQ60115.1	4,40E-118	*G. locusta*	10.926	6.974	0.89
34845_fr5	Turquoise	Q8JZW4	XP_018022776.1	4,80E-168	*H. atzeca*	10.154	6.587	0.89
2278_fr6	Turquoise	P07764	XP_018027036.1	5,00E-132	*H. atzeca*	12.427	6.770	0.89
16972_fr2	Turquoise	P35381	XP_018018775.1	3,00E-146	*H. atzeca*	8.411	5.042	0.88

Table 2. Top hub proteins in the four co-expressed modules of the reproductive organs and embryos of *G. fossarum*. †E-value < 0.001 were considered significant. Proteins with E-values > 0.001 should be considered without homologs. ††*H.: Hyalella; G.: Gammarus; E. sinecaudum: Eremothecium sinecaudum; P.: Procambarus; M.: Marsupenaeus; E. pallida: Exaiptasia pallida*.

active cellular remodeling. Moreover, among the top five hub proteins (those showing the highest module membership value), contigs 121249_fr2 and 203833_fr2 were predicted to have functions involved in RNA editing and splicing regulation while 122312_fr4 is predicted to be involved in protein translation (Table 2, Supplementary Document 2). These results suggest the functional importance of RNA processing, gene isoform switch and protein translation in *G. fossarum* embryogenesis, particularly during the organogenesis (S3 and S4 stages).

The blue module correlated with the ovaries and in particular with the stages involved in secondary vitellogenesis. It presented a significant higher proportion of proteins annotated as vitellogenins or their precursors (19%) (Table 1, Fig. 5A, p-value < 0.01, Fisher's exact test). Moreover, clottable protein-like proteins, another group of proteins belonging to the large lipid transport proteins (LLTP), showed a slight enrichment in this module compared with the others (11%, p-value $= 0.05025$, Fisher's exact test). The functional importance of lipid transport in the ovaries was also confirmed using network statistics that showed the top four hub proteins belonged to either Vtg-like or clottable protein-like subfamilies (Table 2, Supplementary Document 2). It merits noting that LLTP proteins may arise from an ancient duplication event leading to paralogs of Vtg sequences and that the "clottable-like" annotation derives from neofunctionalization processes of proteins belonging to the group of metazoan Vtg[22].

Interestingly, we found two taxonomically-restricted proteins in the yellow module and in the blue module (coded by the contigs 201834_fr4 and 67263_fr2) that showed high intramodular connectivity and module membership, characteristics of a hub role in their respective modules. These proteins do not show any significant homology (E-value > 0.001) with proteins currently available in the NCBInr and Swissprot databases. This result suggests that protein with key roles in the organogenesis and oocyte maturation may have evolved divergently in amphipods and it shows the strength of co-expression network analysis to get new molecular insights in non-model organisms.

Finally, the two modules correlated with the amphipod testes showed two distinct functional profiles. The brown module was highly enriched in proteins annotated as members of the actin or myosin families (84%, Table 1 and Fig. 5, Supplementary Document 1, p-value < 0.01, Fisher's exact test). In particular, among the top five hub proteins, we found the top four to be annotated as myosin light (contig 20975_fr4) or heavy (194796_fr2, 37276_fr6, 182086_fr1) chains and the fifth to be an actin isoform (48693_fr3) (Table 2, Supplementary Document 2). The turquoise module presented instead a significant higher proportion of proteins involved in the energy metabolism, in particular in the glycolysis pathway and ATP synthesis (25%, Table 1 and Fig. 5, Supplementary Document 2, p-value < 0.01, Fisher's exact test). The contig 191065_fr3 that codes for the rate limiting enzyme pyruvate kinase was the top hub protein in the module. Other two enzymes involved in the glycolytic pathway (a putative GAPDH and a predicted fructose aldolase, coded by the contigs 141775_fr4 and 2278_fr6, respectively) were also found among the top five hub proteins (Table 2, Supplementary Document 2). Finally, we observed that the values of correlation coefficients between the eigengenes of the turquoise and brown modules and the post-copula times showed opposite trends, with the maximum difference at day 4 (Fig. 5B). It is also of interest noting that the peak in the brown module is at day 0 of copulation, suggesting a contribution of

the muscle contraction of the gonad during fertilization. These results suggest that the two pathways interact in a coordinated manner during spermatogenesis in *G. fossarum* testes.

Discussion

In this study, we used for the first time a co-expression network analysis to investigate the protein organization of the reproductive organs and embryos in the freshwater amphipod *G. fossarum* using shotgun proteomics data. We showed that WGCNA was well suited to extract biological information using shotgun proteomics data from a non-model organism. In our case study that considered contrasted biological samples at different maturation or developmental stages, this methodology allowed us to identify different modules of co-expressed proteins correlated with the spermatogenesis cycle, and some specific physiological stage of ovary maturation or embryonic development, shedding new light on the molecular physiology of the reproductive system in *G. fossarum*. We found one module (named as yellow) correlated with embryos, notably with the mid-developmental stages S3 and S4. A second module (named as blue) correlated with the ovaries, notably with stages from C1 to D2; and finally the two other modules (namely the brown and the turquoise) both correlated with the testes, with the trends in the correlation coefficients suggesting these two pathways are mutually controlled during the amphipod spermatogenesis. Compared with our previous studies[2,10], we were able to identify key proteins involved in physiological pathways such as oocyte maturation, spermatogenesis and embryonic development without making use of *a priori* protein function predicted by standard sequence similarity search, but by using hierarchical clustering to construct modules from protein abundance data obtained by label-free proteomics. We showed for the first time that embryogenesis in a non-model species, namely *G. fossarum*, is characterized by the activation of factors involved in RNA processing, such as RNA editing or RNA splicing control. Interestingly, mechanisms involved in the regulation of alternative splicing have been identified in evolutionarily distant metazoans, such as the fruit fly and the mouse which are well characterized models[23,24]. Our data-driven approach enabled us to identify two proteins with hub properties, 201834_fr4 in the yellow module and 67263_fr2 in the blue module, that had no significant homology among the NCBInr and Swissprot databases by sequence similarity search (E-value < 0.001). This result highlights the potential of co-expression network analyses in identifying taxon-restricted proteins with key roles in molecular physiology processes. Moreover, the observation that one of these proteins is involved in organogenesis is in line with the recent proposal, based on the comparison of the developmental transcriptomes of 10 different species, of a divergent mid-development transition that uses species-specific functions in embryo development for defining the phyletic body plan[25]. This result is coherent with the observation that S1 stage corresponds to a 2 cell-stage embryos and most of their proteomic profiles are inherited by the maternal egg, while stage S3 and S4 are mainly involved in the organogenesis[10,14].

Our network analysis found evidence of two different pathways involved in testicular processes, namely the glycolytic pathways and the actin-myosin system in *G. fossarum*. While testicular metabolism is scarcely investigated in arthropods, an early biochemical study reported that glycolysis played an important energetic role in both early and late spermatogenesis in *Drosophila hydei*[26]. However, glycolysis has been reported as a key pathway in the spermatogenesis of different vertebrate species. Notably, glycolytic enzymes have been found in the carp seminal plasma using mass spectrometry[27]. A gene expression screening across human, mouse, and rat testes found a functional enrichment in genes involved in glycolysis and pyruvate metabolism[28]. Moreover, a study suggested that mouse spermatogonial stem cells might be primed for conditions that favor glycolytic activity, a bioenergetics state that helps to maintain their functional integrity[29]. Interestingly, we previously reported the contig 40028_fr4 (annotated as a glycogen phosphorylase) among the proteins altered in the testes of male gammarids exposed to pyriproxyfen, suggesting that this pathway might be sensitive to chemical exposures in this amphipod species[30].

Actins and myosins play also an important role in spermatogenesis[31]. Myosins belong to a large superfamily of molecular motors and they are involved in different processes during spermatogenesis, such as acrosomal formation or spermatid individualization[32]. Actin is responsible for the formation of specific sub-cellular structures in arthropods[31]. In *Drosophila*, actin forms apical bundles required for spermatid individualization. Moreover, a new cytoskeletal structure composed of actin and microtubules, namely the acroframosome, has been recently described in the crustacean *Macrobrachium nipponense*[33]. Together, these studies suggest the importance of actin and myosin interactions and the glycolytic pathway as the energy fuel during spermatogenesis in metazoans, including crustaceans, reinforcing our results that showed that these interactions are distinctly activated in *G. fossarum* testes.

Finally, the co-expression analysis performed in this study confirmed and expanded our previous results that showed a high diversity of proteins belonging to the LLTP superfamily involved in yolk formation[10]. We found strong co-expression of 22 contigs annotated as Vtg-like or clottable protein like proteins in the ovaries, expanding the previous set of 8 contigs. While multiple copies of Vtg genes were reported in many arthropods[34,35], it might be possible that some of the contigs represent a fragmented reconstruction of the original mRNA, giving an overestimation of the total number of vitellogenins or clottable proteins in *G. fossarum*. Our co-expression correlations among different Vtg-like contigs might help annotation efforts and improve *de novo* transcriptome assemblies, likely providing a better estimation of the number of LLTP proteins in this freshwater amphipod.

In conclusion, in this case study we performed the first co-expression network analysis on the proteome of male, female and embryos of *G. fossarum*, an emergent model species in molecular ecotoxicology. We found evidence of the importance of unappreciated molecular pathways involved in the amphipod embryogenesis, notably RNA splicing, and confirmed the diversity of proteins belonging to the large lipid transfer family. Moreover, we were able to identify proteins with a hub role in embryogenesis and vitellogenesis that do not have any close homologs in sequenced animal genomes. This result shows the strength of co-expression network methods in generating working hypothesis on specific proteins that standard homology comparisons or differential expression analysis would have failed to identify. Finally, our results found evidence of a fine-tuned regulation between

energy production and myosin-dependent events in *G. fossarum* spermatogenesis. This study illustrates the relevance of applying systems biology approaches to emergent animal models to improve the understanding of the molecular mechanisms of physiological events with high ecological relevance.

Network analyses are promising but still not fully exploited approaches for the understanding of the molecular physiology of sentinel organisms in ecotoxicology. These tools will help to link adverse outcomes to gene or protein modules, informing the Adverse Outcome Pathway framework with the underlying molecular mechanisms of toxicity.

Data Availability

Original mass spectrometry data are available via the PRIDE repository with the dataset identifier PXD000576 and PXD001002. Spectral count data are directly available in Supplementary Data 1.

References

1. Armengaud, J. *et al*. Non-model organisms, a species endangered by proteogenomics. *J Proteomics* **105**, 5–18 (2014).
2. Trapp, J. *et al*. Proteogenomics of Gammarus fossarum to document the reproductive system of amphipods. *Mol. Cell Proteomics* **13**, 3612–3625 (2014).
3. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol* **4**, Article17 (2005).
4. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
5. Degli Esposti, D. *et al*. Identification of novel long non-coding RNAs deregulated in hepatocellular carcinoma using RNA-sequencing. *Oncotarget* **7**, 31862–31877 (2016).
6. Iancu, O. D. *et al*. Utilizing RNA-Seq data for *de novo* coexpression network inference. *Bioinformatics* **28**, 1592–1597 (2012).
7. Pei, G., Chen, L. & Zhang, W. WGCNA Application to Proteomic and Metabolomic Data Analysis. *Meth. Enzymol.* **585**, 135–158 (2017).
8. Morandin, C. *et al*. Comparative transcriptomics reveals the conserved building blocks involved in parallel evolution of diverse phenotypic traits in ants. *Genome Biol.* **17**, 43 (2016).
9. Orsini, L. *et al*. Early transcriptional response pathways in Daphnia magna are coordinated in networks of crustacean-specific genes. *Mol. Ecol.* **27**, 886–897 (2018).
10. 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).
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. 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).
13. 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).
14. Geffard, O. *et al*. Ovarian cycle and embryonic development in Gammarus fossarum: application for reproductive toxicity assessment. *Environ. Toxicol. Chem.* **29**, 2249–2259 (2010).
15. Lacaze, E. *et al*. DNA damage in caged Gammarus fossarum amphipods: a tool for freshwater genotoxicity assessment. *Environ. Pollut.* **159**, 1682–1691 (2011).
16. Lacaze, E., Geffard, O., Goyet, D., Bony, S. & Devaux, A. Linking genotoxic responses in Gammarus fossarum germ cells with reproduction impairment, using the Comet assay. *Environ. Res.* **111**, 626–634 (2011).
17. Arambourou, H. *et al*. Phenotypic defects in newborn Gammarus fossarum (Amphipoda) following embryonic exposure to fenoxycarb. *Ecotoxicol. Environ. Saf.* **144**, 193–199 (2017).
18. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
19. Gregori, J., Sanchez, A. & Villanueva, J. msmsTests: LC-MS/MS Differential Expression Tests. R package version 1.18.0. (2013).
20. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).
21. Langfelder, P., Zhang, B. & Horvath, S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* **24**, 719–720 (2008).
22. Avarre, J.-C., Lubzens, E. & Babin, P. J. Apolipocrustacein, formerly vitellogenin, is the major egg yolk precursor protein in decapod crustaceans and is homologous to insect apolipophorin II/I and vertebrate apolipoprotein B. *BMC Evol. Biol.* **7**, 3 (2007).
23. Guilgur, L. G. *et al*. Requirement for highly efficient pre-mRNA splicing during Drosophila early embryonic development. *Elife* **3**, e02181 (2014).
24. Revil, T., Gaffney, D., Dias, C., Majewski, J. & Jerome-Majewska, L. A. Alternative splicing is frequent during early embryonic development in mouse. *BMC Genomics* **11**, 399 (2010).
25. Levin, M. *et al*. The mid-developmental transition and the evolution of animal body plans. *Nature* **531**, 637–641 (2016).
26. Geer, B. W., Martensen, D. V., Downing, B. C. & Muzyka, G. S. Metabolic changes during spermatogenesis and thoracic tissue maturation in Drosophila hydei. *Dev. Biol.* **28**, 390–406 (1972).
27. Dietrich, M. A. *et al*. Characterization of carp seminal plasma proteome in relation to blood plasma. *J Proteomics* **98**, 218–232 (2014).
28. Liu, F. *et al*. Comparative and functional analysis of testis-specific genes. *Biol. Pharm. Bull.* **34**, 28–35 (2011).
29. Helsel, A. R., Oatley, M. J. & Oatley, J. M. Glycolysis-Optimized Conditions Enhance Maintenance of Regenerative Integrity in Mouse Spermatogonial Stem Cells during Long-Term Culture. *Stem Cell Reports* **8**, 1430–1441 (2017).
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. Sun, X., Kovacs, T., Hu, Y.-J. & Yang, W.-X. The role of actin and myosin during spermatogenesis. *Molecular Biology Reports* **38**, 3993–4001 (2011).
32. Li, Y.-R. & Yang, W.-X. Myosin superfamily: The multi-functional and irreplaceable factors in spermatogenesis and testicular tumors. *Gene* **576**, 195–207 (2016).
33. Li, Z., Pan, C.-Y., Zheng, B.-H., Xiang, L. & Yang, W.-X. Immunocytochemical studies on the acroframosome during spermiogenesis of the caridean shrimp Macrobrachium nipponense (Crustacea, Natantia). *Invertebrate Reproduction & Development* **54**, 121–131 (2010).
34. Provost-Javier, K. N., Chen, S. & Rasgon, J. L. Vitellogenin gene expression in autogenous Culex tarsalis. *Insect Mol. Biol.* **19**, 423–429 (2010).
35. Wurm, Y. *et al*. The genome of the fire ant Solenopsis invicta. *Proc. Natl. Acad. Sci. USA* **108**, 5679–5684 (2011).

Acknowledgements

The authors thank the Institut Carnot "Risque" initiative and the ANR program PROTEOGAM (ANR-14-CE21-0006-02) for their financial support.

Author Contributions

D.D.E. designed research. D.D.E., C.A., M.A.G. and N.K. analyzed data. D.D.E., C.A., A.C., A.J., O.G. wrote the paper.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-44203-5.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

2. DECOUVERTE DES PROCESSUS BIOLOGIQUES CLES LIES A LA TOXICITE CHEZ LE GAMMARE

2.1 Synthèse

Dans une étude précédente, Trapp et al. (2015) ont montré que l'exposition en laboratoire de mâles *G. fossarum* au pyriproxyfène (Pyr) et au cadmium (Cd) induisait une diminution dose-dépendante des spermatozoïdes. Les auteurs ont cherché à identifier, à partir de données d'abondance de protéomique shotgun, des biomarqueurs protéiques afin d'améliorer la surveillance de l'exposition aux PE et de la fonction reproductive chez les gammares mâles. En utilisant une analyse différentielle de l'abondance des protéines, les résultats et l'interprétation des mécanismes à l'origine des effets néfastes restaient très limités (Trapp et al., 2015). Afin d'aller plus loin dans la compréhension des mécanismes moléculaires impliqués dans la toxicité testiculaire chez *G. fossarum*, une stratégie de biologie des systèmes sur les mesures d'abondance de protéomique shotgun a été effectuée.

Un des objectifs de cette partie est d'utiliser l'analyse de réseaux de coexpression pour **réexploiter le maximum d'informations sur des données issues de la protéomique shotgun déjà disponibles chez un organisme sentinel**. Cette approche a également pour but de **tester la capacité de l'analyse de réseaux de coexpression d'identifier les acteurs protéiques possiblement modulés après exposition à trois contaminants (Pyr, Cd et Met) dans les testicules de *G. fossarum*.** Pour cela, la même approche de biologie des systèmes que pour l'article n°1 (voir partie 1 ci-dessus), WGCNA, a été utilisée sur des données provenant de l'étude citée (Trapp et al., 2015). Le jeu de données protéomiques de départ comportait 871 protéines dont l'abondance (en spectral count) était reportée dans une table de comptage. Chaque condition (*i.e.* les 3 contaminants et deux contrôles) était composée de 5 réplicats.

L'analyse de réseau de coexpression a permis d'identifier dix modules de protéines coexprimées, dont quatre étaient significativement corrélés à l'exposition aux contaminants. L'analyse d'enrichissement des protéines a identifié deux modules associés à l'exposition au Cd, l'un impliqué dans l'organisation du cytosquelette et l'autre dans la réponse au stress oxydatif. En particulier, le premier module était fortement enrichi en protéines appartenant principalement à la famille des myosines, qui sont impliquées dans les processus d'organisation du squelette au cours de la morphogenèse du sperme mature chez les métazoaires (Li and Yang, 2016; Sun et al., 2011). Le module associé à l'exposition au Pyr était enrichi en protéines liées au stress du réticulum endoplasmique, notamment en protéines de choc thermique et en calréticuline, toutes ayant une place centrale (*i.e.* protéines hub) dans le module. Le module corrélé à l'exposition au Met était caractérisé par une proportion importante de protéines spécifiques des amphipodes dont les fonctions ne sont pas encore caractérisées. Ceci montre l'indépendance de la méthode face aux annotations. Ces résultats soulignent la capacité d'une analyse de réseau de coexpression à identifier des protéines taxonomiquement restreintes ayant un intérêt physiologique potentiel dans des organismes modèles émergents, et qui permet d'approfondir l'exploitation de données existantes. Nos résultats montrent ainsi que les réseaux de coexpression sont des outils efficaces et adaptés pour identifier de nouveaux modes d'action (MoA) potentiels à partir d'espèces sentinelles environnementales, telles que *G. fossarum*, en utilisant une approche de protéogénomique.

2.2 Article n°2 : Coexpression network analysis identifies novel molecular pathways associated with cadmium and pyriproxyfen testicular toxicity in *Gammarus fossarum*

Aquatic Toxicology 235 (2021) 105816

Contents lists available at ScienceDirect

Aquatic Toxicology

journal homepage: www.elsevier.com/locate/aqtox

Co-expression network analysis identifies novel molecular pathways associated with cadmium and pyriproxyfen testicular toxicity in *Gammarus fossarum*

Natacha Koenig [a], Christine Almunia [b], Aurore Bonnal-Conduzorgues [a], Jean Armengaud [b], Arnaud Chaumot [a], Olivier Geffard [a], Davide Degli Esposti [a,*]

[a] INRAE, UR RiverLy, Ecotoxicology Team. Centre de Lyon-Grenoble Auvergne Rhône-Alpes, 5 rue de la Doua CS 20244, 69625 Villeurbanne, France
[b] Université Paris-Saclay, Département Médicaments et Technologies pour la Santé (DMTS), CEA, INRAE, SPI-Li2D, F-30207 Bagnols-sur-Cèze, France

ARTICLE INFO

Keywords:
Co-expression networks
Gammarus fossarum
Mode of action
Insecticides
Heavy metals
Proteogenomics
Systems biology
Endocrine disruptors

ABSTRACT

Omics approaches are continuously providing new clues on the mechanisms of action of contaminants in species of environmental relevance, contributing to the emergence of molecular ecotoxicology. Co-expression network approaches represent a suitable methodological framework for studying the rich content of omics datasets. This study aimed to find evidence of key pathways and proteins related to the testicular toxicity in the sentinel crustacean species *Gammarus fossarum* exposed to endocrine disruptors using a weighted protein co-expression network analysis. From a shotgun proteomics dataset of male gonads of *G. fossarum* organisms exposed to cadmium (Cd), pyriproxyfen (Pyr) and methoxyfenozide (Met) in laboratory conditions, four distinct modules were identified as significantly correlated to contaminants' exposure. Protein set enrichment analysis identified modules involved in cytoskeleton organization and oxidative stress response associated with the Cd exposure. The module associated with Pyr exposure was associated with endoplasmic reticulum stress (ER) response, and the module correlated with Met exposure was characterized by a significant proportion of amphipod-restricted proteins whose functions are still not characterized. Our results show that co-expression networks are efficient and adapted tools to identify new potential mode of actions from environmental sentinel species, such as *G. fossarum*, using a proteogenomic approach, even without an annotated genome.

1. Introduction

Anthropogenic activities have left a chemical pollution legacy affecting the aquatic ecosystems in the last decades. Among these contaminants, endocrine disruptors (EDs) are chemicals that have the potential to interfere with the endocrine system, affecting development and reproduction, among other physiological processes (Åke et al., 2013). Some insecticides are conceived to be toxic against a limited range of pest species, specifically insects, with a mode of action (MoA) targeting arthropod endocrine pathways. Juvenile hormone (JH) agonists, such as pyriproxyfen (Pyr), bind to the JH receptor methoprene-tolerant (MET). By inducing its heterodimerization with a species-specific partner protein, these agonists deregulate the transcription of JH responsive genes, inhibiting metamorphosis and embryogenesis in several insects (Jindra and Bittova, 2020; Dhadialla et al., 1998). Another class of insecticides, the nonsteroidal ecdysone (Ec) agonists, such as methoxyfenozide (Met), binds to the Ec receptor with high affinity. The resulting agonist-receptor complex induces developmental arrest by repressing late-responsive genes necessary for cuticle synthesis and ecdysis (Retnakaran et al., 2003). These two pathways coordinate post-embryonic development and molt cycle, and both insecticide families have morphostatic effects leading to increased larval lethality (Retnakaran et al., 2003; Moura and Souza-Santos, 2020; Nakagawa, 2005). Heavy metals, such as cadmium (Cd) is also

Abbreviations: ANOVA, Analysis of variance; Cd, Cadmium; Ec, Ecdyson; ED, Endocrine disruptor; ER, Endoplasmic reticular; FC, Fold change; FDR, False discovery rate; GO, Gene ontology; JH, Juvenile hormone; ME, Module eigenprotein; Met, Methoxyfenozide; MET, Methoprene-tolerant; MM, module membership; MoA, Mode of action; ORF, Open reading frame; PCA, Principal component analysis; Pyr, Pyriproxyfen; SC, Spectral count; TOM, Topological overlap matrix; WGCNA, Weighted gene co-expression network analysis.

* Corresponding author
E-mail address: davide.degli-esposti@inrae.fr (D.D. Esposti).

https://doi.org/10.1016/j.aquatox.2021.105816
Received 6 January 2021; Received in revised form 12 March 2021; Accepted 19 March 2021
Available online 26 March 2021

frequently detected in continental freshwaters (Alric et al., 2019; Ciliberti et al., 2017) and known as toxicant for both vertebrates and invertebrates, with reproductive effects in many freshwater crustaceans (Geffard et al., 2010; Sadeq and Beckerman, 2019; Cribiu et al., 2020; Jaegers and Gismondi, 2020).

The emergence of high-throughput molecular data-collection techniques, collectively known as –omics (*e.g.,* transcriptomics, proteomics, metabolomics), has increasingly allowed to simultaneously interrogate the status of organ, tissue or cell components and to determine how and when these molecules interact with each other. These technologies are contributing to changing the experimental approaches to investigate impacts of EDs by the possibility to consider the multiplicity of endogenous targets of single contaminants (Cuvillier-Hot and Lenoir, 2020). However, most of the research work available today on EDs' MoA is based on vertebrate model organisms because of their genomic annotation availability, thus limiting the extrapolation of the results on a large scale of diverse taxa, including invertebrates (Cuvillier-Hot and Lenoir, 2020).

The advances in genome sequencing and mass spectrometry has expanded the possibility of the use of omics approaches in so-called non-model organisms of ecological relevance, such as the freshwater crustacean *Gammarus fossarum* (Gouveia et al., 2019). The molecular physiology of these organisms and the molecular mechanisms in response to environmental contaminants are still challenging to assess due to potential taxon-specific pathways and uncertainties in gene and protein interactions. In this context, data-driven co-expression network analyses provide a powerful methodology to investigate –omics dataset to explore the system-level functionality of genes. We have recently adapted the weighted gene co-expression network analysis (WGCNA) (Zhang and Horvath, 2005; Langfelder and Horvath, 2008) to proteomics datasets in *G. fossarum* reproductive physiology and identified protein modules specific to vitellogenesis in female ovaries and to organogenesis in embryos (Degli Esposti et al., 2019).

In a previous study, Trapp et al. showed that laboratory exposure to pyriproxyfen and cadmium in males *G. fossarum* induced a dose-related decrease in spermatozoa number, while methoxyfenozide, did not exert any toxic effect on sperm maturation (Trapp et al., 2015). The authors aimed to identify protein biomarkers for improved monitoring of EDs exposure and reproductive function in male gammarids. Using differential abundance analysis, they identified a core of 44 proteins commonly altered by both insecticides (Met and Pyr) and Cd. With few proteins (from 2 for Cd to 7 for Pyr) specific for each contaminant, the interpretation of the mechanisms behind the adverse outcome associated with Cd and Pyr was limited (Trapp et al., 2015). In this study, we show the ability of the co-expression network analysis to allow a broader analysis of the biological pathways underlying an adverse outcome in *G. fossarum* testes and to identify protein modules that can inform about the MoA of contaminants that have an impact on the male amphipod reproductive function. This analysis has implemented the clustering of proteins in co-expression modules and the identification of hub proteins. A correlation matrix has been computed to identify the relationships among modules and the exposure traits. Finally, Gene Ontologies (GO) annotations and network properties of the proteins have been used to improve the functional interpretation of the data.

2. Materials & methods

2.1. Proteomics datasets

We retrieved protein abundance data measured as spectral counts (SC) obtained by high-resolution tandem mass spectrometry using a LTQ-Orbitrap XL instrument from Trapp et al. (Trapp et al., 2015). Original mass spectrometry data are available at the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE repository with the data set identifier PXD001267. Only proteins validated with at least two different peptides were retained (Trapp et al., 2015). The dataset consisted in a matrix of 871 protein abundance data expressed as SC for 40 male gonads (Table S1). Five biological replicates for each condition (control, control with solvent (*i.e.,* 0.005% acetone), cadmium at 0.3 µg/l or 3 µg/l, pyriproxyfen at 0.5 or 50 µg/l, methoxyfenozide at 0.001 or 0.1 µg/l) were included in this dataset. The highest concentrations were chosen because of the observed effects at 50 µg/l (decreased number of spermatozoa for Cd and Pyr), while the lowest concentrations represent a no observed effect concentration (NOEC) that could hide early proteome profile modifications in response to a stress (Trapp et al., 2015). Detailed experimental conditions were described in Trapp et al. (Trapp et al., 2015). Briefly, males were exposed for two spermatogenesis cycles (15 days) and testes were individually sampled from each organism. The annotation of the 871 proteins was updated compared with the original study, by a semiautomatic pipeline using DIAMOND (Buchfink et al., 2015). Protein sequences identified by peptides that matched an ORF entry of the GFOSS RNA-seq database (Trapp et al., 2014) were blasted against the NCBInr and Swissprot databases. The closest predicted protein (indicated with NCBInr ID or SwissProt ID) and its E-value were associated with the corresponding GFOSS contig. The Gene Ontology (GO) annotation of protein homologs found in Swissprot database was performed using data provided by the GO Consortium through the GO website and the web application Amigo (http://amigo.geneontology.org). GO and GO slim levels were extracted.

2.2. Data pre-processing and normalization

In order to identify possible outliers, sample clustering was performed using the hierarchical clustering function, *plotSampleRelation*, implemented in the lumi R package (Du et al., 2008). Despite high inter-individual variability, no extreme outliers were identified and all samples were retained for further analysis (Fig. S1-A). Principal Component Analysis (PCA) was used to analyze and identify the variables explaining the maximum variance associated to the proteomic data of the conditions. No clear-cut treatment driven clustering of the gonads was observed, indicating inter-individual variability explain most of the variance in this dataset. To avoid the noise and poor reproducibility associated with low abundant proteins, proteins with less than three SC in the most abundant condition were filtered out (Gregori et al., 2013). Finally, a dataset comprising 40 samples and 312 proteins was retained for further analyses. Before network modeling, counts need to be normalized across samples (Langfelder and Horvath, 2008; Pei et al., 2017). Spectral counts were normalized using the *calcNormFactors* function from the R package edgeR (Robinson and Oshlack, 2010), as described in Degli Esposti et al. 2019 (Degli Esposti et al., 2019). Hierarchical clustering and PCA did not change significantly after normalization (Fig. S1-B, D).

2.3. Network analysis

The R package Weighted Gene Co-expression Network Analysis (WGCNA) (Zhang and Horvath, 2005; Langfelder and Horvath, 2008) was used to investigate the count-based proteomic database of the male gonads exposed or not to the three contaminants. To increase the statistical power of the analysis, the different doses were combined into a single chemical exposure for Cd, Pyr and Met. WGCNA was performed as described in Degli Esposti et al. (Degli Esposti et al., 2019). In order to build the network, the *networkType* argument was set as "unsigned" so that negatively correlated proteins are considered as connected. The *TOMType* argument was set as "signed" to consider possible anti-reinforcing connection strengths (Zhang and Horvath, 2005). The minimum number of proteins in a module was set at ten.

Module eigenproteins (ME) were calculated for each module. ME are vectors representative of the protein expression profiles in a module. ME represents the principal component of each module and it allows exploring how related the modules are and the correlation between

Table 1
Fundamental protein network properties observed in the *Gammarus fossarum* testes.

Network properties	Values
Density	0.01460815
Centralization	0.02794005
Heterogeneity	0.50169786
Mean Cluster Coefficient	0.04337556
Mean Connectivity	4.54313467

modules and phenotypic traits, such as exposure conditions.

The hub proteins are molecules with the highest degree of connectivity within a module. These hub proteins are expected to play a key role in biological pathways. These hubs are identified by their network properties, such as total (k_{total}) or intramodular (k_{within}) connectivity or module membership (MM). The k_{total} or intramodular connectivity k_{within} are defined as the way the proteins are related to other proteins of the entire network or within the same module, respectively. The MM is representative of the degree of correlation between proteins and modules. If MM approaches 0, the protein may not be part of the module, if the MM approaches 1 or −1, the protein is highly correlated to the module. After identification of the corresponding hub genes, Cytoscape v3.7.2 (Shannon et al., 2003) was also used to visualize and interpret the network and modules' topologies. The network was exported from WGCNA to Cytoscape with the dedicated function *exportNetworkToCytoscape*. The protein-protein interactions are priorly selected with a TOM value threshold higher than 0.1.

2.4. Protein set enrichment analysis

For proteins included in the network analysis, 165 Gene Ontology (GO) terms were retained having at least one annotated protein. The functional enrichment was measured in terms of fold change in the percentage of proteins observed in a module compared with the network for a given GO term (Eq. (1)).

$$\frac{\%\,{}^{\text{modi}}_{\text{GOj}}\text{P}}{\%\,{}^{\text{net}}_{\text{GOj}}\text{P}} \quad (1)$$

where i correspond to a given module and j to a given GO term. A Fisher exact test was performed to evaluate statistical significance of the enrichment followed by a Benjamini-Hochberg correction (False Discovery Rate, FDR) for multiple comparison. The GO terms were considered as significantly enriched with protein overrepresentation greater than two (fold change, FC>2) compared with the rest of the network and/or having an FDR ≤0.1.

2.5. Statistical analysis

One-way ANOVA ($P < 0.05$) followed by Tukey's HSD tests were used to test for differences in expression levels of various proteins for each module using R. When variances were not homogeneous, the Kruskal-Wallis non parametric analysis of variance ($P < 0.05$) followed by pairwise Wilcoxon tests were performed. All statistical analyses were performed using R (v 3.5.0) and R-studio (v 1.2).

2.6. Reproducibility

An R script combining the whole analysis is available on Github (https://github.com/NatachaKoenig/CdPyrMet-Network-Analysis). Spectral count data and the associated metadata are available as Table S1 and Table S2, respectively.

3. Results and discussion

3.1. Molecular functional characterization of testicular protein networks under chemical stress in Gammarus fossarum

We first explored the network properties of the proteomes obtained in the 40 male gonads exposed or not to Cd, acetone, Pyr or Met, performing a weighted protein co-expression analysis based on the abundance level of 312 proteins. Our data fitted (R^2=0.75) a scale-free topology by selecting a power of three (Fig. S2). The main characteristics of the testicular protein network of *G. fossarum* testes are reported in Table 1. The density and centralization (centralization = 0.028; density = 0.015) suggest that the protein network is composed of distinct groups of co-expressed proteins (modules) (Horvath and Dong, 2008). The mean network connectivity for each protein is 4.5 (Table 1, Fig. S2) with protein connectivity (*i.e.,* variability of protein degree) ranging from 1.16 to 13.17. After the definition of the protein network, we applied a dynamic clustering method implemented in WGCNA to identify modules of highly correlated proteins (Fig. 1A). We identified ten modules and each module was given a color-coded name (black, blue, brown, green, magenta, pink, purple, red, turquoise and yellow) using the standard WGCNA nomenclature. Proteins with low co-expression are outside of any modules and were combined together in a gray module. The number of proteins in each module ranged from 12 (purple module) to 45 (turquoise module). The values of topological overlap among the

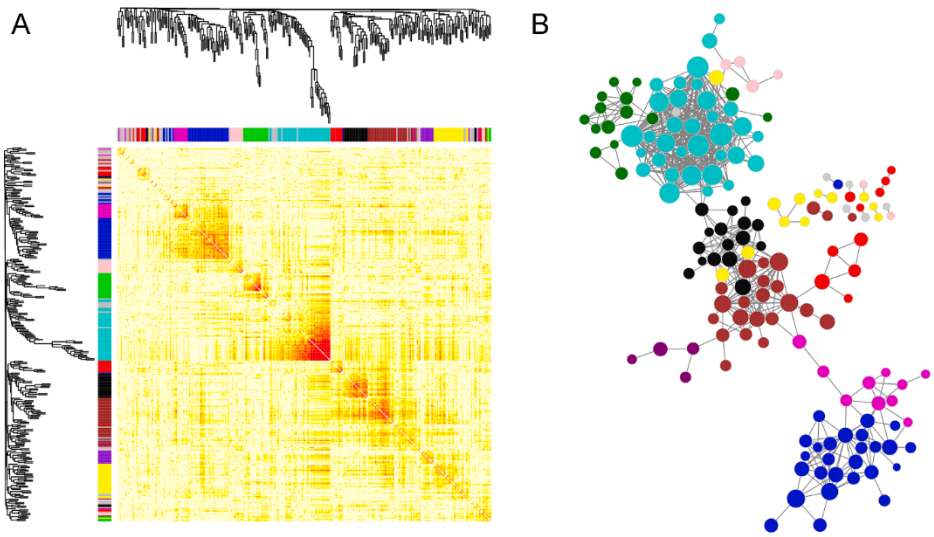


Fig. 1. Protein-protein interaction network of the male gonad proteome in the amphipod *Gammarus fossarum*. (A) The heatmap plot of the topological overlap in the network. Each cluster branch represents a protein. Each dot in the heatmap represents protein-protein topological overlap values (*i.e.*, protein-protein connectivity). The gradient, from light yellow to dark red, shows the overlap intensity between the proteins within a module. The color bar on the top and the left shows the module assignment and the protein dendrogram associated. (B) The network of the most highly connected proteins. Proteins with a topological overlap above a threshold of 0.1 are shown. Each node represents a protein, and each edge represents connection between two proteins. The size of the node is related to the total connectivity (k_{total}). The bigger the node, the more the protein is related to other proteins of the whole network.

Table 2
The most significant GO term enrichment by module.

Module	GO term	% Proteins[a]	Fold Change[b]	p-value	FDR
Black	Uncharacterized proteins[c]	34.8	3.50	0.003	0.1
Blue	Supramolecular fiber_CC	88.1	2.67	7.7E-12	1.6E-09
	Cellular component assembly_BP	90.5	2.45	2.4E-12	2.2E-09
	Cytoskeleton_CC	85.7	2.68	3.2E-11	2.2E-09
	Cytoskeleton organization_BP	85.7	2.60	7.3E-11	3.0E-09
	Anatomical structure formation involved in morphogenesis_BP	85.7	2.60	7.3E-11	3.0E-09
Brown	Biosynthetic process_BP	66.67	2.08	3.4E-05	0.007
	Unfolded protein binding_MF	23.08	5.77	1.8E-04	0.018
	Protein folding_BP	28.20	3.53	6.3E-04	0.043
	Cell death_BP	35.90	2.39	2.5E-03	0.102
	Cytoplasm_CC	82.05	1.44	2.1E-03	0.102
Green	Synapse_CC	29.17	5.83	4.4E-04	0.030
	Plasma membrane organization_BP	25.00	12.50	2.3E-04	0.030
	Extracellular matrix organization_BP	29.17	5.83	4.4E-04	0.030
	Structural molecule activity_MF	45.83	2.29	6.9E-03	0.223
	Locomotion_BP	54.17	2.01	6.9E-03	0.223
Purple	Mitochondrion_CC	100.0	5.56	3.9E-09	8.0E-07
	Small molecule metabolic process_BP	100.0	4.00	1.6E-07	1.6E-05
	Lipid metabolic process_BP	50.00	5.56	7.1E-04	0.036
	Cellular amino acid metabolic process_BP	33.33	11.11	1.6E-03	0.064
	Sulfur compound metabolic process_BP	25.00	25.00	2.0E-03	0.067
Red	DNA binding_MF	30.43	4.35	2.5E-03	0.257
	Nucleus_CC	65.22	2.04	1.9E-03	0.257
	Nuclear chromosome_CC	26.09	4.35	6.1E-03	0.413
Turquoise	Hydrolase activity, acting on glycosyl bonds_MF	26.67	6.67	6.0E-06	0.001
	Immune system process_BP	44.44	2.12	1.5E-03	0.100
	Catabolic process_BP	40.00	2.11	2.9E-03	0.119
	Molecular transducer activity_MF	22.22	3.17	3.8E-03	0.130
	Carbohydrate metabolic process_BP	28.89	2.41	4.9E-03	0.144
Yellow	Antioxidant activity_MF	13.33	13.33	2.7E-03	0.149
	Detoxification_BP	13.33	13.33	2.7E-03	0.149
	Peroxisome_CC	16.67	8.33	2.9E-03	0.149

[a] It represents the percentage of proteins in the module belonging to the corresponding GO term.
[b] It is calculated as detailed in the Material and Methods section.
[c] The black module was significantly enriched in proteins of unknown function and annotated as uncharacterized, which is not a GO term. However this suggests the module could be taxon specific.

proteins of the network showed high levels of interconnection between some identified modules, such as the blue with the magenta and the brown, with red and the black modules (Fig. 1B). On the other hand, the yellow module is mostly peripheral to the rest of the network, with proteins that show lower intra-modular connectivity values than those shown by the proteins in the other modules (Table S3).

Functional annotation using the Gene Ontology databases showed that 90% of the 312 proteins belonging to the network were annotated under at least one GO term. In particular, 89% of the protein could be associated with a molecular function, 79% with a cellular components and 82% with a biological process, while 10% (31 proteins) of the proteins remained GO orphan and were indeed annotated as possible homologs of uncharacterized proteins mostly found encoded in the *Hyallela atzeca* genome (Poynton et al., 2018). Using the GO annotation, we performed a protein set enrichment analysis (see Materials and Methods section) to investigate the biological functions of the protein modules identified. Following these criteria, 7 modules out of 10 presented a significant enrichment of at least 3 GOs. The most significant GOs per module are shown in Table 2.

Interestingly, each module shows distinct GOs enrichments, supporting the hypothesis that the co-expression analysis could identify distinct biological processes embedded in the testicular protein network. For instance, two modules account for the energy metabolism pathways identified in this protein network, namely the turquoise and the purple modules. The turquoise module present an enrichment in proteins involved in the carbohydrate metabolism, an important pathway in crustacean testicles (Degli Esposti et al., 2019; Liu et al., 2019) while the purple module is enriched in enzymes involved in the lipid catabolism via the beta-oxidation pathway.

Three other modules, the blue, the red and the green ones, seems to play an important role in spermatogenesis. In particular, the blue module is highly enriched in proteins mainly belonging to the myosin family, that are involved in the skeleton organization processes during the mature sperm morphogenesis in metazoans (Li and Yang, 2016; Sun et al., 2011). Notably, this module presents a similar protein composition to a protein module previously identified in gammarids testes (Degli Esposti et al., 2019), showing the consistency of co-expression network analysis in identifying preserved and reproducible modules of biological significance (Langfelder et al., 2011). The green module is likely to be involved in spermatogenesis since it is mainly enriched in protein members of the spectrin family, all included in the most-enriched GO terms (synapses and plasma membrane organization). Indeed, spectrins are important components of spermatozoid nuclei in metazoan (Ghosh-Roy et al., 2004; Ocampo et al., 2005) and members of this family were previously identified as important regulators of cell contractility in the myoepithelial cells of the *C. elegans* spermatheca (Wirshing and Cram, 2018). The red module presents instead an enrichment in nuclear proteins and histone proteins. Histones are essential in chromatin organization and they are almost completely displaced by protamines during the post-meiotic stages of the spermatogenesis (Hundertmark et al., 2018).

The brown and the yellow modules also showed significant GO term enrichment, both in functions related with stress response pathways. The brown module was enriched in protein related to endoplasmic reticulum (ER) stress, notably heat shock proteins and calreticulin. The yellow module was enriched in proteins related to oxidative stress response, such as catalase and peroxiredoxin. Finally, the black module was not enriched in any known GO term; however, it harbored a significant proportion (8 proteins corresponding to 34% of the module proteins) of proteins whose functions are still not characterized. These

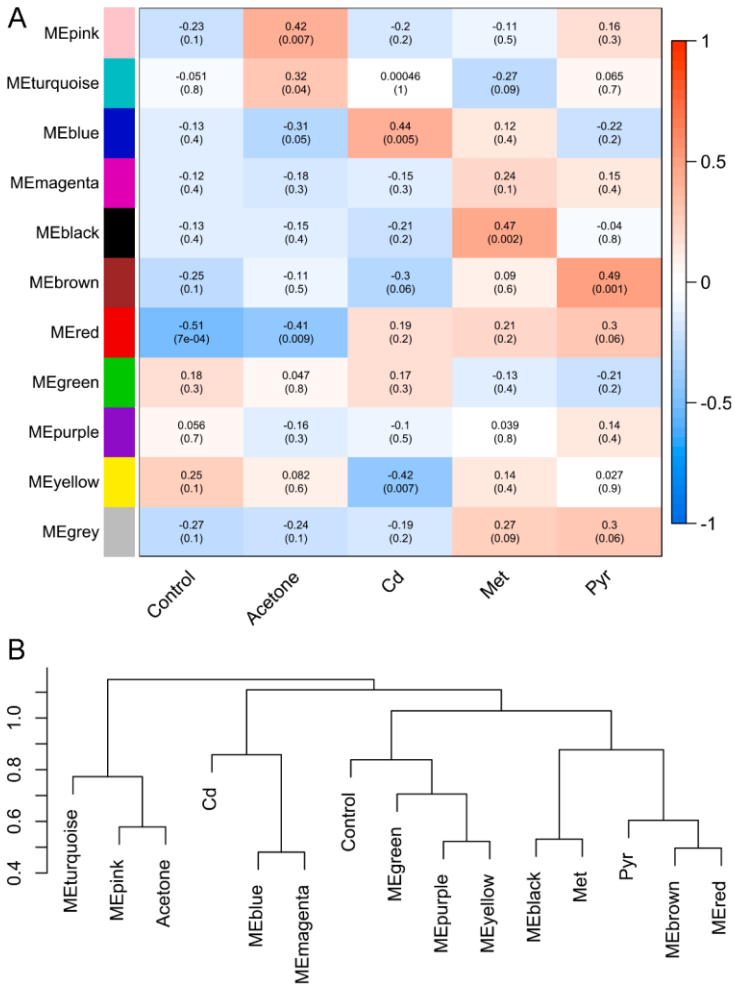


Fig. 2. Associations between the modules and the chemical exposures. (A) The modules-trait relationships heatmap. Module eigenproteins (ME) (labeled by color), which represents the protein expression profile of each module, are in rows, and exposure traits are in columns. The gradient represents the strength of the correlation between the module eigengene and the exposure trait (controls or contaminants) with the Pearson correlation coefficient (ρ) and the associated p-value between parentheses. In the heatmap, red color represents the positive correlations (ρ>0), while the blue represents the negative correlation (ρ<0). (B) The eigenprotein adjacency obtained by hierarchical clustering shows the similarity between the module eigenproteins (labeled by color name) and the exposure trait.

results show the potential of data-driven network analysis in identifying yet unknown biological functions involved in gammarids or amphipod reproduction that are likely to be taxon specific, as reported for other rapid evolving proteins involved in gamete recognition in invertebrates (Vacquier and Swanson, 2011; Lobov et al., 2019).

In conclusion, the network approach here presented has identified distinct modules of co-expressed proteins involved in specific metabolic and biological pathways directly or indirectly involved in the spermatogenesis process in *Gammarus fossarum* testicles.

3.2. *Module-exposure correlations indicate chemical-specific clusters of proteins and suggest new insecticides' modes of action*

In order to explore the mode of action of the different chemical exposures, we correlated the module eigenproteins with each contaminant exposure (Fig. 2). Module-exposure correlations show indeed that each chemical exposure had a different module profile (Fig. 2). Interestingly, acetone (used as solvent to solubilize the insecticides Met and Pyr (Trapp et al., 2015) and not displaying any toxic effect at the tested levels (Trapp et al., 2015) exert a biological effect by affecting the expression of proteins belonging to the turquoise and pink modules (Fig. 2). These two modules are topologically related (Fig. 1) and while the pink module did not show any significant GO term enrichment, the turquoise module is significantly enriched in enzymes involved in carbohydrate metabolism, with glycosidase enzymes exhibiting the highest connectivity values (Table 2, Fig. S3).

Then, we started to examine the modules correlated with the exposures that were reported to induce a dose-related decrease in sperm count, namely Cd and Pyr (Trapp et al., 2015), by integrating the network properties of each single protein in the modules with the results of the GO term enrichment analysis (Table 2).

Based on the GO term annotation, our results show that the global expression profile (eigenprotein) of oxidative stress response in the yellow module was negatively correlated to Cd exposure (Pearson ρ=−0.42, p-value=0.007), while the eigenprotein of the blue module, enriched in proteins involved in the morphogenesis of sperm cells, was positively correlated to Cd exposure (Pearson ρ=0.44, *p*-value=0.005). Indeed, a *G. fossarum* catalase (contig 110912_fr3) was the second top hub protein in the yellow module (Table 3, Fig. 3A), while one myosin heavy chain (contig 14_fr3) was concomitantly among the top hub protein in the blue module (Fig. 3C). It merits noting that neither of these proteins showed a significant expression change in the testicles of Cd treated organisms compared to the control group (Fig. 3B,D). This result highlights the power of this data-driven method looking at expression relationship between proteins. Co-expression network analyses help identify potentially affected pathways when changes in the abundances of single proteins are too weak to be detected, either due to limited statistical power (*e.g.,* too few biological replicates) or technological limits (*e.g.*, sensitivity of mass spectrometers). Interestingly, Cd was previously shown to induce spermatozoa structural anomalies in other arthropod models, such as the lepidopteran *Bombyx mori* (Yuan et al., 2016) and the coleopteran *Blaps polycresta* (the ground beetle) (Shonouda and Osman, 2018). Moreover, the main mechanism responsible for Cd toxicity in various organs, included testes, is the induction of

Table 3
Top hub proteins for each identified modules in the testicular protein network of *G. fossarum*.

Contig	Module	Name	kTotal	kWithin
461_fr1	black	Uncharacterized protein LOC108668644	5,94	3,21
23928_fr5	black	Uncharacterized protein LOC108668644	6,11	3,21
7908_fr4	black	Uncharacterized protein LOC108668644	5,14	3,03
439_fr3	black	Neurogenic locus notch homolog protein 1-like	7,12	3,00
192788_fr6	black	Hypothetical protein LR48_Vigan02g018500	7,76	2,95
1836_fr6	blue	Titin-like isoform X3	7,53	5,26
14_fr3	blue	Myosin heavy chain, muscle-like isoform X7	8,97	5,01
155642_fr5	blue	Alpha-actinin, sarcomeric-like isoform X1	7,62	4,67
422_fr4	blue	Myosin heavy chain, muscle-like isoform X17	9,63	4,21
40293_fr4	blue	Paramyosin, long form-like isoform X1	7,74	4,11
21529_fr6	brown	Spliceosome RNA helicase DDX39B-like	9,06	4,86
5310_fr1	brown	Calreticulin-like	7,72	4,60
12402_fr1	brown	Heat shock protein HSP 90-alpha-like	9,60	4,20
48901_fr2	brown	Heat shock protein 70	9,81	4,19
191476_fr2	brown	Tubulin alpha-1 chain isoform X11	8,24	3,78
156_fr4	green	Spectrin alpha chain-like isoform X1	5,72	3,69
4456_fr5	green	Spectrin alpha chain-like isoform X2	5,90	3,54
470_fr4	green	Spectrin beta chain-like	6,22	3,04
7983_fr6	green	Myosin heavy chain, non-muscle-like isoform X1	4,62	2,81
166827_fr1	green	Hemocyanin	5,39	2,01
141642_fr4	magenta	Myosin regulatory light chain 2-like	5,80	2,64
20975_fr4	magenta	Myosin light chain 1	6,63	2,23
23125_fr3	magenta	Uncharacterized protein LOC108676008	5,01	1,54
199608_fr3	magenta	Tropomyosin isoform X23	3,88	1,53
48693_fr3	magenta	Actin-1	4,57	1,40
110918_fr2	pink	Hemolymph clottable protein-like isoform X1	5,20	1,90
48868_fr5	pink	Hemocyanin subunit 1	4,25	1,66
189695_fr2	pink	Alpha-2-macroglobulin-like isoform X6	4,46	1,39
144152_fr3	pink	Alpha-tectorin-like isoform X2	3,94	1,27
20335_fr3	pink	Hemolymph clottable protein-like isoform X1	3,01	1,17
7953_fr6	purple	Propionyl-CoA carboxylase alpha chain, mitochondrial-like	6,77	1,55
153121_fr6	purple	Dehydrogenase/reductase SDR family member 4-like	4,47	1,19
7625_fr2	purple	3-oxoacyl-[acyl-carrier-protein] synthase 2-like	5,32	1,06
1344_fr4	purple	Dihydrolipoyllysine-residue acetyltransferase component of pyruvate dehydrogenase complex, mitochondrial-like	5,54	1,03
51377_fr2	purple	Aspartate aminotransferase, mitochondrial-like	5,56	1,02
65834_fr4	red	Histone H3-like	5,60	2,52
2101_fr3	red	Histone H2A.V isoform X2	7,62	2,27
32861_fr4	red	Uncharacterized protein LOC100759437	4,34	2,02
13250_fr6	red	Uncharacterized protein LOC111062361	6,40	1,90
36242_fr3	red	Hemocyanin B chain-like	5,90	1,08
Contig	Module	Name	kTotal	kWithin
38283_fr1	turquoise	Sucrase-isomaltase, intestinal-like	13,18	9,98
9667_fr6	turquoise	Beta-1,3-glucan-binding protein-like	12,55	9,89
189210_fr4	turquoise	Alpha glucosidase	12,21	9,84
10149_fr3	turquoise	Putative metallocarboxypeptidase ECM14	12,20	9,82
4757_fr2	turquoise	Amylase	11,33	8,99
209438_fr3	yellow	Na+/K+ ATPase alpha subunit	5,65	2,62
110912_fr3	yellow	Catalase	6,67	2,06
210555_fr5	yellow	Transitional endoplasmic reticulum ATPase TER94 isoform X2	5,78	1,93
9002_fr1	yellow	Tubulin alpha-1A	4,12	1,81
10171_fr5	yellow	Tubulin beta-1 chain-like	4,16	1,81

oxidative stress (Bhardwaj et al., 2020). While our study design ignores possible dynamic variation in the expression and efficiency of the oxidative stress response, our results suggest that chronic or sub chronic Cd exposures may affect the antioxidant defense pathway in gammarid male gonads. Interestingly, Cd exposure did not induce genotoxicity in *G. fossarum* male gonads at concentrations similar to those used in the original study (Trapp et al., 2015; Lacaze et al., 19), and this network analysis support these results, since the modules associated with Cd exposure did not show any enrichment or hub proteins directly involved in DNA repair or other genotoxic damage response. Finally, it is worth of note that the catalase (a top hub protein belonging to the oxidative stress response in the yellow module) is highly connected with two tubulins, proteins that are involved in sperm maturation, together with the myosin and the actin proteins that are enriched in the other protein module associated with Cd exposure (blue module). Thus, our results open the hypothesis of a cross-talk between the molecular actors involved in sperm morphology and oxidative stress regulation and that Cd might affect both processes.

Pyr exposure correlated with the brown module that is enriched in proteins involved in the response to ER stress (Table 2). A calreticulin-like protein (contig 5310_fr1), a key protein in keeping intracellular Ca^{2+} homeostasis, and two heat shock proteins (contigs 12402_fr1 and 48901_fr2), all involved in protein folding and ER stress response, were among the top hub proteins in the module (Table 3, Fig. 4A). Similarly to Cd-related hub proteins, these proteins did not present any significant abundance difference compared with their paired-control (acetone exposed organisms), neither with the other exposures (Fig. 4B-D). The ER stress has been recently identified as a novel mode of action involved in testicular toxicity *in vivo* in mice (Li et al., 2020) and *in vitro* in rodent cells (Ham et al., 2020). In particular, the synthetic phenolic antioxidant butylated hydroxyanisole (BHA) suppresses cell viability and induces cell cycle arrest by dysregulating calcium homeostasis and inducing ER stress in mouse testis cell lines (Li et al., 2020). Interestingly, Pyr was shown to alter Ca^{2+} homeostasis *in vitro* and *in vivo* in *Danio rerio* testes (Staldoni de Oliveira et al., 2021). Intracellular Ca^{2+} alterations were associated with increased lipid peroxidation and decreased antioxidant

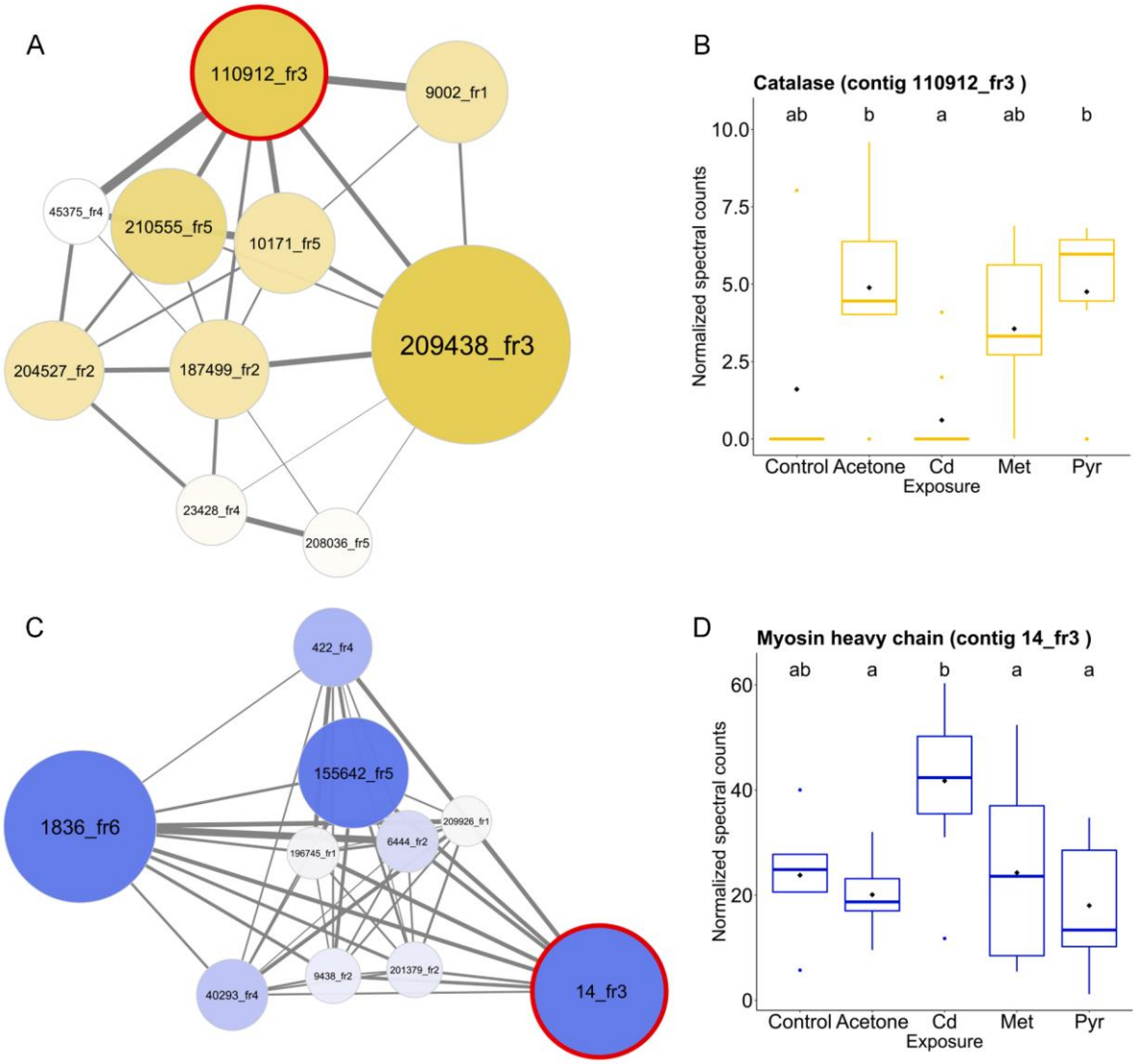


Fig. 3. Network protein interactions in modules correlated with Cd exposure. For each module the node size and the color node intensity are proportional to the k_{within} (intra-modular connectivity). The edge width represents the correlation strength between two proteins (TOM value). The red circles indicate the top hub protein belonging to the most enriched GO term in the module. In the boxplots, the mean of each box plot is represented by the black diamond. Comparison of the protein abundance was performed by ANOVA and the Tukey test. If two conditions don't share any letter in common, the protein abundance is significantly different ($p<0.05$). (A) Visualization of the yellow module proteins, negatively associated with the Cd exposure. (B) Protein abundance levels for the catalase (contig 110912_fr3) depending on the exposure. (C) Visualization of the blue module proteins, positively associated with the Cd exposure. (D) Protein abundance levels for a myosin heavy chain (contig 14_fr3) depending on the exposure.

capacity, leading to altered spermatogenesis (Staldoni de Oliveira et al., 2021). Our network analysis identify the *G. fossarum* calreticulin as a hub protein in the module associated with Pyr exposure, suggesting that Pyr may induce ER stress and heat shock response via an alteration of intracellular Ca^{2+} homeostasis. Our enrichment analysis of the protein module correlated with Pyr exposure showed that the same proteins involved in the ER stress response also belongs to the "cell death" GO term annotation (Table 2). In fact, it is known that where ER homeostasis cannot be restored, ER stress-induced cellular dysfunction may lead to cell death (Sano and Reed, 2013). Indeed, our analysis provides the first evidence of a potential implication of heat shock response and ER stress in the testicular toxicity of an invertebrate model widely used in ecotoxicology, suggesting a novel mode of action of insect growth regulators targeting the arthropods' juvenile hormone pathway.

While Met-did not induce any effect in sperm count (Trapp et al., 2015), our network analysis identified a module enriched in proteins showing a high homology with *H. atzeca* uncharacterized proteins (black module, Table 2). Three of these proteins also showed the highest level of intra-module connectivity, suggesting they have a key role in structuring protein-protein interactions in this module (Table 3). These results highlight the ability of a data-driven co-expression network analysis in identifying taxonomically restricted proteins with potential physiological interest in emerging model organisms.

In conclusion, our results show that applying co-expression analysis in ecotoxicoproteomics may help identify contaminant MoA in environmentally relevant model organisms with limited genomic knowledge, providing new mechanistic hypotheses, although further experimental investigation is required to confirm these hypotheses. Network methods applied to omics data in ecotoxicology can provide useful information to build Adverse Outcome Pathways in multiple species of environmental relevance.

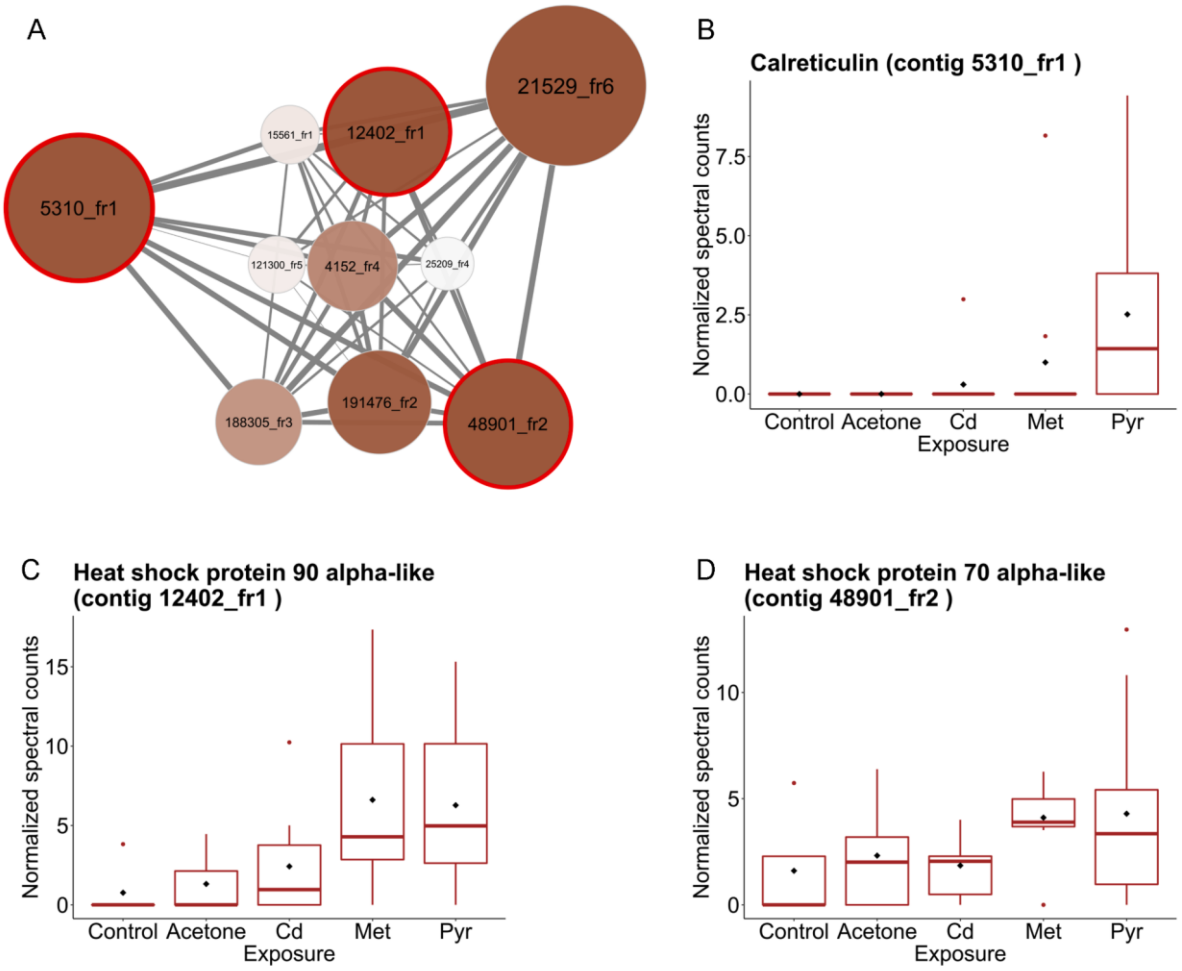


Fig. 4. Network protein interactions in the module correlated with Pyr exposure. (A) Visualization of the brown module proteins, positively associated with the Pyr exposure. The node size, the color node intensity and edge widths are represented as in Fig. 3. The red circles indicate top hub proteins in unfolded protein binding and protein folding GO terms. (B) The protein abundance levels of calreticulin (contig 5310_fr1). (C) Protein abundance levels of the heat shock protein 90 alpha-like (contig 12402_fr1). (D) Protein abundance levels of heat shock protein 70 (contig 48901_fr1). The mean of each box plot is represented by the black diamond. Comparison of the protein abundance was performed by ANOVA and the Tukey test. If two conditions don't share any letter in common, the protein abundance is significantly different ($p<0.05$).

Author agreement

All authors have seen and approved the final version of the manuscript being submitted. They warrant that the article is the authors' original work, hasn't received prior publication and isn't under consideration for publication elsewhere.

Davide Degli Esposti, PhD

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

N.K. is supported by a PhD fellowship from INRAE. This work was partially supported by the French National Research Program for Environmental and Occupational Health of Anses (2019/1/112). The funding agency had no involvement in study design, analysis or interpretation of data.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.aquatox.2021.105816.

References

Åke, Bergman B., Heindel Jerrold, J., Tim, Kasten K., Kidd Karen, A., Susan, Jobling J., Maria, Neira N., Thomas, Zoeller Z.R., Georg, Becher B., Poul, Bjerregaard B., Riana, Bornman B., Ingvar, Brandt B., Andreas, Kortenkamp K., Derek, Muir M., Brune, Drisse D.MarieM.-Noël N., Roseline, Ochieng O., Skakkebaek, Niels N.E., Sundén, Byléhn B.Agneta A., Taisen, Iguchi I., Jorma, Toppari T., Woodruff, Tracey T.J., 2013. The impact of endocrine disruption: a consensus statement on the state of the science. Environ. Health Perspect. 121 (4), a104–a106. https://doi.org/10.1289/ehp.1205448.

Jindra, M., Bittova, L., 2020. The juvenile hormone receptor as a target of juvenoid "insect growth regulators. Arch. Insect Biochem. Physiol. 103 (3), e21615. https://doi.org/10.1002/arch.21615.

Dhadialla, T.S., Carlson, G.R., Le, D.P., 1998. New insecticides with ecdysteroidal and juvenile hormone activity. Annu. Rev. Entomol. 43, 545–569. https://doi.org/10.1146/annurev.ento.43.1.545.

Retnakaran, A., Krell, P., Feng, Q., Arif, B., 2003. Ecdysone agonists: mechanism and importance in controlling insect pests of agriculture and forestry. Arch. Insect Biochem. Physiol. 54 (4), 187–199. https://doi.org/10.1002/arch.10116.

Moura, J.A.S., Souza-Santos, L.P., 2020. Environmental risk assessment (ERA) of pyriproxyfen in non-target aquatic organisms. Aquat. Toxicol. Amst. Neth. 222, 105448 https://doi.org/10.1016/j.aquatox.2020.105448.

Nakagawa, Y., 2005. Nonsteroidal ecdysone agonists. Vitam. Horm. 73, 131–173. https://doi.org/10.1016/S0083-6729(05)73005-3.
Alric, B., Geffard, O., Chandesris, A., Ferréol, M., François, A., Perceval, O., Piffady, J., Villeneuve, B., Chaumot, A., 2019. Multisubstance indicators based on caged Gammarus bioaccumulation reveal the influence of chemical contamination on stream macroinvertebrate abundances across France. Environ. Sci. Technol. 53 (10), 5906–5915.
Ciliberti, A., Chaumot, A., Recoura-Massaquant, R., Chandesris, A., François, A., Coquery, M., Ferréol, M., Geffard, O., 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. https://doi.org/10.1016/j.watres.2017.04.031.
Geffard, O., Xuereb, B., Chaumot, A., Geffard, A., Biagianti, S., Noël, C., Abbaci, K., Garric, J., Charmantier, G., Charmantier-Daures, M., 2010. Ovarian cycle and embryonic development in Gammarus fossarum: application for reproductive toxicity assessment. Environ. Toxicol. Chem. 29 (10), 2249–2259. https://doi.org/10.1002/etc.268.
Sadeq, S.A., Beckerman, A.P., 2019. The chronic effects of copper and cadmium on life history traits across cladocera species: a meta-analysis. Arch. Environ. Contam. Toxicol. 76 (1), 1–16. https://doi.org/10.1007/s00244-018-0555-5.
Cribiu, P., Devaux, A., Garnero, L., Abbaci, K., Bastide, T., Delorme, N., Quéau, H., Degli Esposti, D., Ravanat, J.-.L., Geffard, O., Bony, S., Chaumot, A.A., 2020. Population dynamics" perspective on the delayed life-history effects of environmental contaminations: an illustration with a preliminary study of cadmium transgenerational effects over three generations in the crustacean Gammarus. Int. J. Mol. Sci. 21 (13), 4704. https://doi.org/10.3390/ijms21134704.
Jaegers, J., Gismondi, E., 2020. Gammarid exposure to pyriproxyfen and/or cadmium: what effects on the methylfarnesoate signalling pathway? Environ. Sci. Pollut. Res. 27 (25), 31330–31338. https://doi.org/10.1007/s11356-020-09419-3.
Cuvillier-Hot, V., Lenoir, A., 2020. Invertebrates facing environmental contamination by endocrine disruptors: novel evidences and recent insights. Mol. Cell. Endocrinol. 504, 110712 https://doi.org/10.1016/j.mce.2020.110712.
Gouveia, D., Almunia, C., Cogne, Y., Pible, O., Degli-Esposti, D., Salvador, A., Cristobal, S., Sheehan, D., Chaumot, A., Geffard, O., Armengaud, J., 2019. Ecotoxicoproteomics: a decade of progress in our understanding of anthropogenic impact on the environment. J. Proteomics 198, 66–77. https://doi.org/10.1016/j.jprot.2018.12.001.
Zhang, B., Horvath, S., 2005. A general framework for weighted gene co-expression network analysis. Stat. Appl. Genet. Mol. Biol. 4 (1) https://doi.org/10.2202/1544-6115.1128.
Langfelder, P., Horvath, S., 2008. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9 (1), 559. https://doi.org/10.1186/1471-2105-9-559.
Degli Esposti, D., Almunia, C., Guery, M.-.A., Koenig, N., Armengaud, J., Chaumot, A., Geffard, O., 2019. Co-expression network analysis identifies gonad- and embryo-associated protein modules in the sentinel species Gammarus fossarum. Sci. Rep. 9 (1), 7862. https://doi.org/10.1038/s41598-019-44203-5.
Trapp, J., Armengaud, J., Pible, O., Gaillard, J.-.C., Abbaci, K., Habtoul, Y., Chaumot, A., Geffard, O., 2015. Proteomic investigation of male Gammarus fossarum, a freshwater crustacean, in response to endocrine disruptors. J. Proteome Res. 14 (1), 292–303. https://doi.org/10.1021/pr500984z.
Buchfink, B., Xie, C., Huson, D.H., 2015. Fast and sensitive protein alignment using DIAMOND. Nat. Methods 12 (1), 59–60. https://doi.org/10.1038/nmeth.3176.
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. https://doi.org/10.1074/mcp.M114.038851 mcp.M114.038851.
Du, P., Kibbe, W.A., Lin, S.M.LumiL., 2008. A pipeline for processing illumina microarray. Bioinformatics 24 (13), 1547–1548.
Gregori, J., Sánchez, A., Villanueva, J., 2013. MsmsTests: LC-MS/MS Differential Expression Tests. R Package Version 1.14, 0.
Pei, G., Chen, L., Zhang, W., 2017. WGCNA Application to proteomic and metabolomic data analysis. Methods Enzymol. 585, 135–158. https://doi.org/10.1016/bs.mie.2016.09.016.
Robinson, M.D., Oshlack, A., 2010. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 11 (3), R25. https://doi.org/10.1186/gb-2010-11-3-r25.
Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T., 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13 (11), 2498–2504.
Horvath, S., Dong, J., 2008. Geometric interpretation of gene coexpression network analysis. PLoS Comput. Biol. 4 (8), e1000117 https://doi.org/10.1371/journal.pcbi.1000117.
Poynton, H.C., Hasenbein, S., Benoit, J.B., Sepulveda, M.S., Poelchau, M.F., Hughes, D.S.T., Murali, S.C., Chen, S., Glastad, K.M., Goodisman, M.A.D., Werren, J.H., Vineis, J.H., Bowen, J.L., Friedrich, M., Jones, J., Robertson, H.M., Feyereisen, R., Mechler-Hickson, A., Mathers, N., Lee, C.E., Colbourne, J.K., Biales, A., Johnston, J.S., Wellborn, G.A., Rosendale, A.J., Cridge, A.G., Munoz-Torres, M.C., Bain, P.A., Manny, A.R., Major, K.M., Lambert, F.N., Vulpe, C.D., Tuck, P., Blalock, B.J., Lin, Y.-.Y., Smith, M.E., Ochoa-Acuña, H., Chen, M.-.J.M., Childers, C.P., Qu, J., Dugan, S., Lee, S.L., Chao, H., Dinh, H., Han, Y., Doddapaneni, H., Worley, K.C., Muzny, D.M., Gibbs, R.A., Richards, S., 2018. The toxicogenome of hyalella azteca: a model for sediment ecotoxicology and evolutionary toxicology. Environ. Sci. Technol. 52 (10), 6009–6022. https://doi.org/10.1021/acs.est.8b00837.
Liu, X., Luo, B.-.Y., Feng, J.-.B., Zhou, L.-.X., Ma, K.-.Y., Qiu, G.-.F., 2019. Identification and profiling of microRNAs during gonadal development in the giant freshwater prawn Macrobrachium rosenbergii. Sci. Rep. 9 (1), 2406. https://doi.org/10.1038/s41598-019-38648-x.
Li, Y.-.R., Yang, W.-.X., 2016. Myosin superfamily: the multi-functional and irreplaceable factors in spermatogenesis and testicular tumors. Gene 576 (1), 195–207. https://doi.org/10.1016/j.gene.2015.10.022. Pt 2.
Sun, X., Kovacs, T., Hu, Y.-.J., Yang, W.-.X., 2011. The role of actin and myosin during spermatogenesis. Mol. Biol. Rep. 38 (6), 3993–4001.
Langfelder, P., Luo, R., Oldham, M.C., Horvath, S., 2011. Is my network module preserved and reproducible? PLoS Comput. Biol. 7 (1), e1001057 https://doi.org/10.1371/journal.pcbi.1001057.
Ghosh-Roy, A., Kulkarni, M., Kumar, V., Shirolikar, S., Ray, K., 2004. Cytoplasmic dynein-dynactin complex is required for spermatid growth but not axoneme assembly in drosophila. Mol. Biol. Cell 15 (5), 2470–2483. https://doi.org/10.1091/mbc.e03-11-0848.
Ocampo, J., Mondragón, R., Roa-Espitia, A.L., Chiquete-Félix, N., Salgado, Z.O., Mújica, A., A, 2005. ctin, myosin, cytokeratins and spectrin are components of the guinea pig sperm nuclear matrix. Tissue Cell 37 (4), 293–308. https://doi.org/10.1016/j.tice.2005.03.003.
Wirshing, A.C.E., Cram, E.J., 2018. Spectrin regulates cell contractility through production and maintenance of actin bundles in the Caenorhabditis elegans spermatheca. Mol. Biol. Cell 29 (20), 2433–2449. https://doi.org/10.1091/mbc.E18-06-0347.
Hundertmark, T., Gärtner, S.M.K., Rathke, C., Renkawitz-Pohl, R., 2018. Nejire/DCBP-mediated histone H3 acetylation during spermatogenesis is essential for male fertility in Drosophila melanogaster. PLoS ONE 13 (9), e0203622. https://doi.org/10.1371/journal.pone.0203622.
Vacquier, V.D., Swanson, W.J., 2011. Selection in the rapid evolution of gamete recognition proteins in marine invertebrates. Cold Spring Harb. Perspect. Biol. 3 (11), a002931 https://doi.org/10.1101/cshperspect.a002931.
Lobov, A.A., Maltseva, A.L., Mikhailova, N.A., Granovitch, A.I., 2019. The molecular mechanisms of gametic incompatibility in invertebrates. Acta Naturae 11 (3), 4–15. https://doi.org/10.32607/20758251-2019-11-3-4-15.
Yuan, H., Qin, F., Guo, W., Gu, H., Shao, A., 2016. Oxidative stress and spermatogenesis suppression in the testis of cadmium-treated Bombyx Mori Larvae. Environ. Sci. Pollut. Res. Int. 23 (6), 5763–5770. https://doi.org/10.1007/s11356-015-5818-3.
Shonouda, M., Osman, W., 2018. Ultrastructural alterations in sperm formation of the beetle, Blaps Polycresta (Coleoptera: tenebrionidae) as a biomonitor of heavy metal soil pollution. Environ. Sci. Pollut. Res. Int. 25 (8), 7896–7906. https://doi.org/10.1007/s11356-017-1172-y.
Bhardwaj, J.K., Panchal, H., Saraf, P., 2020. Cadmium as a testicular toxicant: a review. J. Appl. Toxicol. JAT. https://doi.org/10.1002/jat.4055.
Lacaze, E., Geffard, O., Bony, S., Devaux, A., 2010 Jul 19. Genotoxicity assessment in the amphipod Gammarus fossarum by use of the alkaline Comet assay. Mutat Res. 700 (1–2), 32–38. https://doi.org/10.1016/j.mrgentox.2010.04.025.
Li, J., Gao, L., Zhu, B.-.B., Lin, Z.-.J., Chen, J., Lu, X., Wang, H., Zhang, C., Chen, Y.-.H., Xu, D.-.X., 2020. Long-term 1-nitropyrene exposure induces endoplasmic reticulum stress and inhibits steroidogenesis in mice testes. Chemosphere 251, 126336. https://doi.org/10.1016/j.chemosphere.2020.126336.
Ham, J., Lim, W., You, S., Song, G., 2020. Butylated hydroxyanisole induces testicular dysfunction in mouse testis cells by dysregulating calcium homeostasis and stimulating endoplasmic reticulum stress. Sci. Total Environ. 702, 134775 https://doi.org/10.1016/j.scitotenv.2019.134775.
Staldoni de Oliveira, V., Gomes Castro, A.J., Marins, K., Bittencourt Mendes, A.K., Araújo Leite, G.A., Zamoner, A., Van Der Kraak, G., Mena Barreto Silva, F.R., 2021. Pyriproxyfen induces intracellular calcium overload and alters antioxidant defenses in Danio rerio testis that may influence ongoing spermatogenesis. Environ Pollut 270, 116055. https://doi.org/10.1016/j.envpol.2020.116055.
Sano, R., Reed, J.C., 2013 Dec. ER stress-induced cell death mechanisms. Biochim Biophys Acta 1833 (12), 3460–3470. https://doi.org/10.1016/j.bbamcr.2013.06.028.

CHAPITRE III – ANNOTATION MULTIOMIQUE POUR LA CARACTERISATION SANS *A PRIORI* DU METABOLISME LIPIDIQUE CHEZ *G. FOSSARUM*

1. RESUME

Les nouvelles technologies de séquençage (NGS) et de spectrométrie de masse (MS) permettent l'exploration sans *a priori* pour des espèces non modèles. Le métabolisme lipidique (ML) est central chez les animaux, car il est impliqué dans la constitution des membranes cellulaires, dans la production d'énergie par le catabolisme des acides gras (AGs) et dans la synthèse d'hormones, telles que les hormones stéroïdiennes. Chez les crustacés, le ML intervient dans de nombreux processus physiologiques tels que la mue, la reproduction, ou encore l'homéostasie énergétique. De plus, le ML peut être la cible potentielle de certains contaminants présents dans l'eau. Dans cette étude, nous cherchons à développer une approche multiomique sans *a priori* pour exploiter les données moléculaires disponibles chez de nombreuses espèces non modèles d'intérêt, afin de décrire le métabolisme lipidique chez l'espèce sentinelle *Gammarus fossarum.*

Les travaux ont été réalisés à partir de transcriptomes ainsi que de données de protéomique shotgun d'organes, chez le mâle et la femelle de *G. fossarum*. Dans ce cadre, un des verrous à résoudre est le problème de la redondance d'informations contenue dans les assemblages disponibles. Pour cela, nous avons utilisé une stratégie qui permet de sélectionner et conserver uniquement les isoformes les plus exprimées. Ensuite, nous avons adapté un outil génomique de reconstruction des voies métaboliques aux données transcriptomiques. La reconstruction des voies métaboliques a permis d'identifier 78 et 76 voies impliquées dans le ML pour le mâle et la femelle, respectivement. Les données de MS ont permis de valider la détection et l'existence d'une centaine d'enzymes catalysant les réactions constituant les voies identifiées dans les différents organes.

L'analyse multiomique à partir de l'intégration de données de réseaux métaboliques a permis l'identification des potentiels acteurs clés du ML qui n'avaient pas été décrits jusqu'alors. Cette étude est une première description des acteurs du ML dans les organes de *G. fossarum*. Ceci ouvre de nouvelles pistes de travail en protéomique ciblée pour identifier des biomarqueurs et pour la compréhension du rôle du ML dans le cadre de l'exposition aux contaminants.

Mots clés : amphipode, transcriptomique, protéomique, *Gammarus fossarum*, métabolisme lipidique, lipidome, intégration multiomique.

2. ABBRÉVIATIONS

AG	Acide Gras
BLAST	Basic Local Alignment Search Tool
BUSCO	Benchmarking Universal Single-Copy Orthologues
CycADS	Cyc Annotation Database System
EC	Enzyme Commission
FC	Fold Change
FDR	False Discovery Rate
GFF3	General Feature Format 3
G. fossarum	*Gammarus fossarum*
GO	Gene Ontology
HMG-CoA	Hydroxyméthylglutaryl – CoA
IsoPct	Isoforme Percentage
KAAS	KEGG Automatic Annotation Server
KEGG	Kyoto Encyclopedia of Genes and Genomes
KO	KEGG Orthology
MDS	Multi Dimensional Scaling
ML	Métabolisme Lipidique
MS	Mass Spectrometry ou Spectrométrie de Masse
NGS	Next Generation Sequencing ou Séquençage de Nouvelle Génération
ORF	Open Reading Frame
PGDB	Pathway/Genome Database
PUFA	Polyunsaturated Fatty Acid
RE	Réticulum Endoplasmique
RNA-Seq	Séquençage ARN
RSEM	RNA-Seq by Expectation Maximisation
RT	Temps de Rétention
SC	Spectral Count
TAG	Triacylglycérides

3. INTRODUCTION

Le métabolisme lipidique chez les arthropodes et les crustacés en particulier, intervient dans différents processus physiologiques tels que la mue, la reproduction, la croissance ou encore la synthèse d'hormones, et assure l'homéostasie énergétique des organismes (Tessier et al., 1983; Zeng et al., 2018). Chez les crustacés, plusieurs études s'intéressant à la composition lipidique ont montré que le genre, le cycle de reproduction ou encore le stade de développement des individus gouvernent les profils et quantités en lipides (acides gras polyinsaturés (PUFAs), lipides totaux) (Fu et al., 2021) et en énergie (glycogène) (Correia et al., 2003; Gismondi et al., 2012; Rosa and Nunes, 2002; Sroda and Cossu-Leguille, 2011). Des études ont montré une plus grande quantité de lipides chez la femelle de *Gammarus pulex* plutôt que chez le mâle (Plaistow et al., 2003), ceci pouvant être expliqué par l'oogenèse (Buikema Jr and Benfield, 1979) et l'accumulation des lipoprotéines constituant les œufs, telles que les vitellogénines et les protéines clottable-like (Meusy, 1980; Sroda and Cossu-Leguille, 2011; Trapp et al., 2016a).

Les facteurs environnementaux tels que la température peuvent aussi moduler les voies du métabolisme lipidique et les réserves énergétiques (Sroda and Cossu-Leguille, 2011). Chez *G. roeseli*, il a été observé que les niveaux de lipides étaient plus importants en hiver qu'en été, probablement dus au stade de reproduction et aux comportements alimentaires régulés ou contrôlés par les saisons (température) chez cette espèce (Sroda and Cossu-Leguille, 2011). En automne, la chute des feuilles constitue notamment un apport important de nourritures dans les rivières et s'accompagne d'une augmentation des quantités de lipides, permettant un stockage des éléments constitutifs des organismes. En été, la diminution des lipides pourrait être liée au coût énergétique de la reproduction provenant d'un nombre de cycles de reproduction plus important lié aux températures plus élevées (Sutcliffe, 1993).

Les contaminants chimiques peuvent aussi impacter les voies du métabolisme lipidique. Il a été montré que certains perturbateurs endocriniens interférents avec le métabolisme lipidique des vertébrés. Les obésogènes (*e.g.* tributylétain) sont connus pour affecter la distribution et la synthèse des lipides chez le crustacé *Daphnia magna* (Fuertes et al., 2018; Jordão et al., 2016a, 2016b; Jordão Rita et al., 2015). L'exposition au tributylétain favorise l'accumulation de lipides de stockage tels que les triacylglycérols et le cholestérol dans des gouttelettes lipidiques montrant que la perturbation lipidique observée chez les vertébrés peut se retrouver aussi chez les invertébrés comme la daphnie (Fuertes et al., 2018; Jordão et al., 2016a, 2016b; Jordão Rita et al., 2015).

Certains médicaments utilisés chez l'homme pour traiter les dyslipidémies, comme les statines, agissent en inhibant l'enzyme responsable de l'étape limitante la synthèse du mévalonate, l'hydroxyméthylgluratyl-CoA (HMG-CoA) réductase, une protéine très conservée chez les eucaryotes (Barros et al., 2020; Istvan and Deisenhofer, 2001; Neuparth et al., 2020, 2014; Santos et al., 2016). La voie de synthèse du mévalonate est également impliquée chez les arthropodes pour la synthèse des hormones juvéniles (JHs) qui jouent un rôle sur le développement embryonnaire, la métamorphose, la synthèse de la vitellogénine et la production de phéromones (Bellés et al., 2005; Nijhout, 1994). Ces composés se retrouvent dans les milieux aquatiques, à des concentrations de l'ordre du ng/L (Jelic et al., 2011) et représentent ainsi une problématique environnementale, notamment comme potentiels perturbateurs endocriniens chez des espèces non-cibles. La régulation moléculaire de ces fonctions biologiques fondamentales (*e.g.* les voies liées à la synthèse d'hormones chez *Daphnia magna*) est décrite chez des espèces dites modèles, pour lesquelles le génome est parfaitement annoté. Toutefois, les génomes de trois crustacés sont disponible à ce jour :

Daphnia magna (Colbourne et al., 2011), *Hyalella azteca* (Poynton et al., 2018), et *Parhyale hawaiensis* (Kao et al., 2016).

Si les espèces modèles jouent un rôle clé dans la découverte et l'acquisition de connaissances en biologie et physiologie moléculaires, en revanche leur utilisation en sciences environnementales reste limitée et se confronte à plusieurs verrous scientifiques, comme leur distribution géographique et la diversité des milieux aquatiques. Les connaissances et observations faites sur les organismes modèles en conditions de laboratoire ne permettent pas d'extrapoler et de prédire les réponses des organismes présents dans les milieux (Banks and Stark, 1998; Calow et al., 1997; Van Straalen, 2003). Il existe une quinzaine d'outils recensés (Mendoza et al., 2019) pour explorer les données génomiques des espèces modèles et reconstruire les réseaux métaboliques, comme AureMe (Aite et al., 2018), MetaDraft (Hanemaaijer et al., 2017), CarveMe (Machado et al., 2018) ou encore Pathway Tools (Karp et al., 2021). Les bases de données BioCyc offrent un cadre pour l'analyse intégrée des réseaux métaboliques chez les espèces modèles comme *Drosophila melanogaster*. Ces bases de données sont construites automatiquement à l'aide de Pathway Tools (Karp et al., 2021). Elles comprennent des données sur les voies métaboliques liées aux informations sur le génome. Néanmoins, pour construire ces bases de données, il est nécessaire de mettre en commun toutes les annotations relatives au génome donné (*i.e.* Gene Ontology, Enzyme Commission) dans un fichier de synthèse spécifique de Pathway Tools. Ainsi, le système de gestion Cyc Annotation Database System (CycADS) a été développé pour faciliter la création et la mise à jour de ces bases de données en automatisant le processus d'intégration des annotations (Baa-Puyoulet et al., 2016; Vellozo et al., 2011).

De nos jours, même si les crustacés forment un des plus grands groupes taxonomiques, il reste néanmoins peu connu au niveau moléculaire. Toutefois les avancées des nouvelles technologies de séquençage et de spectrométrie de masse permettent d'étendre l'acquisition des données moléculaires à large échelle (*e.g.* par transcriptomique, protéomique ou métabolomique) aux organismes de pertinence environnementale, comme l'amphipode d'eau douce *Gammarus fossarum* (Cogne et al., 2019c; Fu et al., 2021; Trapp et al., 2016b). C'est une espèce largement répartie au niveau européen, qui joue un rôle clé dans le fonctionnement des écosystèmes aquatiques, notamment comme source de nourriture pour de nombreuses espèces et qui est de plus en plus utilisée en écotoxicologie, reconnue comme sensible à une large gamme de contaminants (Chaumot et al., 2015; Kunz et al., 2010). Dans le cas du gammare, les connaissances moléculaires se concentrent principalement au niveau du transcriptome et du protéome. Des études métabolomiques voient aussi le jour sur l'étude des altérations dues aux médicaments (Bonnefoy et al., 2019; Gómez-Canela et al., 2016; Sheikholeslami et al., 2020). L'approche couplée de séquençage d'ARNm et spectrométrie de masse à haute résolution, dite la protéogénomique (Armengaud et al., 2014b; Gouveia et al., 2019a) permet d'élargir l'acquisition de connaissances moléculaires chez les espèces non modèles en augmentant considérablement le nombre de protéines identifiées (1873 dont 218 spécifiques de l'espèce) par spectrométrie de masse (Trapp et al., 2014b). Plus récemment, des études en lipidomique ont été menées chez *Gammarus spp.* décrivant la composition en acides gras (Jiménez-Prada et al., 2021; Kolanowski et al., 2007; Kühmayer et al., 2020), les variations en lien avec l'exposition à des contaminants (Arambourou et al., 2018; Konschak et al., 2021) ou la distribution dans les tissus grâce à l'imagerie par spectrométrie de masse (MSI) (Fu et al., 2021). Néanmoins, à ce

jour, chez le gammare, comme chez de nombreuses espèces d'intérêt environnemental, les voies moléculaires contrôlant et régulant le métabolisme lipidique restent non décrites.

L'amélioration de la connaissance des réseaux métaboliques chez ces espèces est primordiale afin d'améliorer la compréhension des mécanismes moléculaires de toxicité pour les contaminants environnementaux. Pour ceci, il est nécessaire de développer des stratégies bioinformatiques qui permettent d'exploiter sans *a priori* et d'intégrer l'ensemble ou la majeure partie des données omiques existantes. L'objectif de cette étude est la caractérisation des voies métaboliques des lipides dans les organes de l'amphipode d'eau douce *Gammarus fossarum* par une approche multiomique qui intègre le transcriptome et le protéome (Figure 34). Nous nous sommes intéressés au ML pour sa sensibilité aux perturbations du milieu et le peu d'étude le documentant de manière systémique chez les crustacés malgré sa centralité dans le métabolisme.

La démarche expérimentale est illustrée dans la Figure 34. Le premier aspect du travail consiste à rendre les données transcriptomiques déjà disponibles chez *G. fossarum* (Cogne et al., 2019c) utilisables par un outil bioinformatique développé initialement pour des données de génomique, CycADS. Par la suite, et dans le but d'obtenir une première description du ML qui intègre les niveaux d'expression des ARNm et des protéines, des données de protéomique provenant d'organes de mâles et femelles de *G. fossarum* ont été combinées. Cette intégration permet de valider les enzymes qui sont aussi identifiées en MS et d'identifier des profils d'expression spécifiques de différents organes.

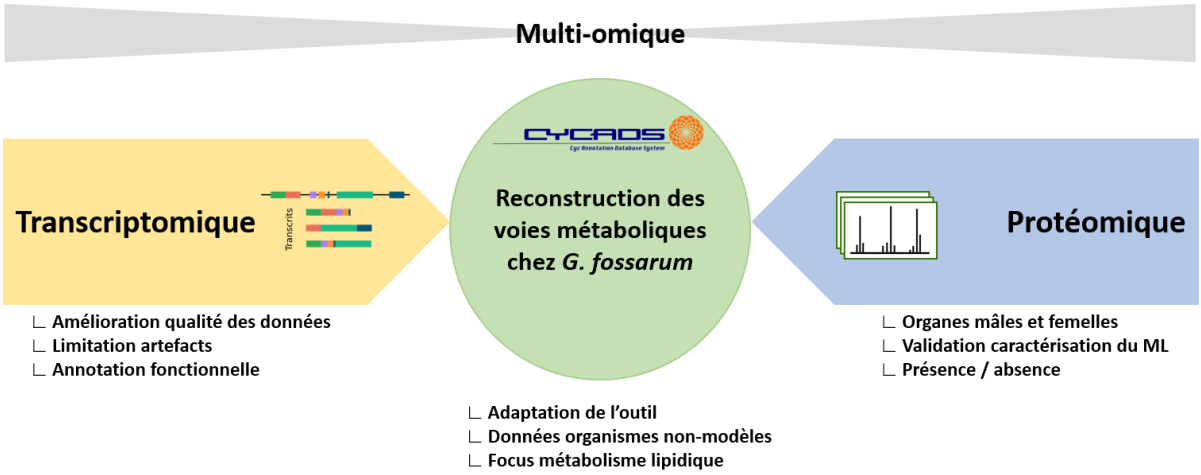


Figure 34 : Démarche méthodologique pour la caractérisation du métabolisme lipidique chez *Gammarus fossarum*.

4. MATERIELS ET METHODES

4.1 Ressources transcriptomiques

Les transcriptomes du mâle et de la femelle de *Gammarus fossarum B* proviennent des travaux de Cogne et al. (2019). Les séquences des lectures originales sont disponibles dans l'archive « NCBI Sequence Reads Archive » sous le numéro d'accession SRR8089722 ("RNAseq GFBM," 2018) (BioProject PRJNA497972 and BioSample SAMN10259937) pour le mâle et sous le numéro d'accession SRR8089729 ("RNAseq GFBF," 2018) (BioProject PRJNA497972 and BioSample SAMN10259936) pour la femelle. Les assemblages des transcriptomes d'origine sont disponibles dans la base de données « NCBI Transcriptome Shotgun Assembly Sequence Database » avec l'identifiant GenBank GHDA01000000 ("Assemblage GFBM," 2018) (BioProject PRJNA497972 and BioSample SAMN10259937) pour la mâle et avec l'identifiant GenBank GHCZ01000000 ("Assemblage GFBF," 2018) (BioProject PRJNA497972 and BioSample SAMN10259937) pour la femelle. Les transcriptomes ont ensuite été analysés séparément.

4.2 Ressources protéomiques

Les données de protéomique ont été acquises pour cette étude en particulier, à partir de trois mâles et trois femelles de l'espèce *Gammarus fossarum* qui ont été échantillonnés. Six organes ou régions anatomiques (céphalon, branchies, caeca, intestin, gonades mâles, gonades femelles et le reste du corps) ont été prélevés pour chaque organisme et congelés en azote liquide.

Extraction des protéines

Pour la protéomique shotgun, chaque organe à analyser a été directement dissous dans 40 µL de tampon d'échantillon LDS (Invitrogen), à l'exception du céphalon et du reste du corps qui ont été préalablement broyés dans le LDS en ajoutant une bille d'acier de 4 mm puis en utilisant un homogénéisateur de tissus. Les échantillons ont été soumis à 1 min de sonication (sonicateur transonic 780H) et ont été portés pendant 5 min à 95 °C. Les broyats d'organes ont été totalement dissous dans le tampon d'échantillon LDS, puis 35 µL de chaque échantillon ont été soumis à une SDS-PAGE sur un NuPAGE 10 puits gradient 4-12 % (Invitrogen) pendant 10 min à 150 V avec du tampon MES. Les gels ont été colorés avec le colorant bleu de Coomassie Safe (Invitrogen) et décolorés pendant la nuit avec de l'eau. La totalité des protéines de chaque puits a été extraite sous forme d'une seule bande de polyacrylamide et traitée pour une nouvelle décoloration et un traitement à l'iodoacétamide. Les protéines ont été protéolysées avec de la trypsine de qualité séquençage (Roche) en utilisant 0,01 % de tensioactif Protease-MAX (Promega).

Spectrométrie de masse

Les mélanges de peptides résultants ont été analysés en mode d'acquisition dépendant des données avec un spectromètre de masse haute résolution Orbitrap Exploris

(ThermoFisher) couplé à un système LC UltiMate 3000 (Dionex-LC Packings), exploité comme décrit précédemment (Trapp et al., 2015).

Identification des protéines et quantification par comptage spectral

Les listes de pics ont été générées avec le logiciel Mascot Daemon (version 2.3.2 ; Matrix Science) en utilisant le filtre d'importation de données *extract_msn.exe* (Thermo). Les options du filtre d'importation de données ont été définies sur 400 (masse minimale), 5000 (masse maximale), 0 (tolérance de regroupement), 0 (balayages intermédiaires) et 1000 (seuil), comme décrites précédemment (Christie-Oleza et al., 2012). Les spectres MS/MS ont été attribués à des séquences peptidiques avec le moteur de recherche Mascot Daemon 2.3.2 (Matrix Science) par rapport à la base de données personnalisée dérivée du transcriptome assemblé.

Deux tables ont été obtenues pour chaque individu à partir de la spectrométrie de masse. La première table (Table S1 dans Supplementary Data 1, Supplementary Data 2) correspond aux peptides (32925 séquences de peptides uniques mâles, 39018 séquences de peptides uniques femelles) détectés en spectrométrie de masse et leurs caractéristiques associées (identifiant de la protéine associée, séquence, longueur, rapport M/Z, modifications, etc.). La deuxième table (Table S2 dans Supplementary Data 1, Supplementary Data 2) contient les mesures d'abondance des protéines pour chaque échantillon exprimé en « spectral count » (SC). En effet, les spectral count sont une estimation de l'abondance d'une protéine dans les échantillons à travers le nombre de spectres MS/MS qui sont associés (Liu et al., 2004). Plus un peptide est abondant, plus il a de chance d'être fragmenté dans le spectromètre de masse. La table S2 du mâle

(Supplementary Data 1) est une matrice de 5073 protéines pour 18 échantillons. La Table S2 de la femelle (Supplementary Data 2) est une matrice de 5678 protéines pour 18 échantillons.

4.3 Assemblage des transcriptomes

Les assemblages disponibles représentent la totalité des contigs reconstruits par Trinity. Dans le but de réduire la redondance des transcrits alternatifs présents dans les assemblages d'origine, nous avons choisi de filtrer les possibles artéfacts de transcrits et les transcrits avec un niveau d'expression faible comme recommandé par Trinity (Haas et al., 2013). Pour cela, nous avons évalué le pourcentage du niveau d'expression (abondance) pour un transcrit donné en comparaison à tous les transcrits à l'intérieur d'un cluster d'isoformes (Haas et al., 2013), dit pourcentage d'isoforme (IsoPct) (Figure 35, étape n°1). Pour estimer l'abondance des transcrits sans le génome de référence, nous avons utilisé une méthode basée sur un alignement des lectures de séquençage au transcriptome assemblé *de novo*. Nous avons utilisé les scripts en Perl (*align_and_estimate_abundance.pl, abundance_estimates_to_matrix.pl, filter_low_expr_transcripts.pl*) de Trinity v2.9.1 (Grabherr et al., 2011) qui fournissent un support direct pour réaliser la méthode d'alignement avec bowtie2 v2.4.2 (Langmead and Salzberg, 2012) et la quantification avec RSEM v1.3.3 (Li and Dewey, 2011). Une fois la mesure d'abondance réalisée pour chaque transcrit, nous avons filtré les transcrits en utilisant l'option *highest_iso_only*. Cette option permet de conserver les transcrits alternatifs avec le plus haut niveau d'abondance, traduit par l'IsoPct. Il en résulte deux fichiers fasta des transcriptomes ainsi nettoyés pour le mâle et la femelle de *G. fossarum* (Figure 35, étape n°2). En parallèle, les fichiers des ORFs (Open Reading Frame) qui correspondent à la traduction des régions de séquences codantes (CDS) étaient disponibles pour chaque transcriptome dans la publication de Cogne et al. (2019). Les deux fichiers des ORFs correspondant au mâle et à la femelle ont aussi été filtrés à l'aide des identifiants des

transcrits que nous avons conservés après le nettoyage des transcriptomes. Ces deux nouveaux transcriptomes et fichiers d'ORFs serviront à la suite du pipeline d'annotation (Figure 35, étape n°3).

Pour évaluer la qualité des assemblages ainsi obtenus, une mesure de complétude des transcriptomes, basée sur la présence d'orthologues, a été effectuée à l'aide de l'outil « Benchmarking Universal Single-Copy Orthologues » (BUSCO) (Simão et al., 2015) version 4.1.2. Cet outil a été utilisé avec la base de données « arthropoda_odb10 » de 1013 gènes orthologues en copie unique des arthropodes (Kriventseva et al., 2019). Cette base de données est la plus proche taxonomiquement des gammaridés. BUSCO a été utilisé en mode transcriptome avec les options par défaut.

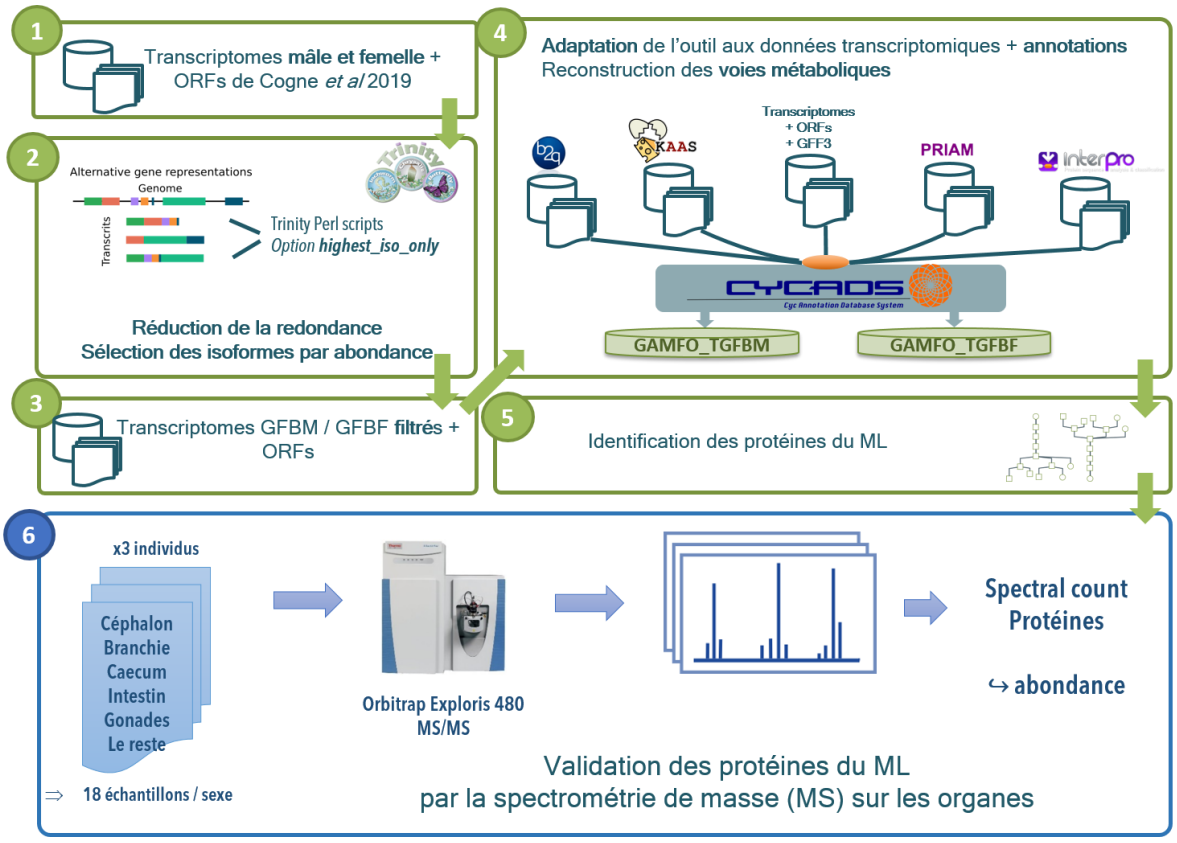


Figure 35 : Schéma global du pipeline d'annotation des transcriptomes et de la validation par la spectrométrie de masse chez le mâle et la femelle de *G. fossarum*.

4.4 Annotation fonctionnelle des assemblages par CycADS

CycADS est un système qui permet la standardisation de l'annotation d'un génome. Dans cette étude, le pipeline a été adapté pour les données transcriptomiques de *G. fossarum*. Pour y parvenir, nous avons créé un fichier d'annotation structurelle artificielle (GFF3) en concaténant tous les contigs. Après cela, CycADS a été utilisé comme déjà fait par Vellozo et al. (2011) (Figure 35, étape n° 3).

En particulier, nous avons réalisé une annotation fonctionnelle sur nos transcriptomes mâle et femelle nettoyés avec les pipelines sur machine locale comme Blast2GO (Conesa et al., 2005; Conesa and Götz, 2008), Priam (Claudel-Renard et al., 2003) et InterproScan (Jones et al., 2014), et le pipeline KAAS-KEGG en ligne (Moriya et al., 2007). Les informations fonctionnelles ont été collectées (KEGG Orthologie (KO), numéro Enzyme Commission (EC) et Gène Ontologie (GO)) dans la base de données à l'aide du module collecteur d'annotations de CycADS (Figure 35, étape n° 3). Toutes les annotations ont été extraites et collectées dans un fichier « Pathological » qui a été utilisé dans le compartiment « Pathway Tools » (Karp et al., 2010) pour générer le BioCyc Pathway Genome Database (PGDB) correspondant et réaliser la reconstruction du métabolisme (Figure 35, étape n° 3). Les paramètres par défaut ont été utilisés pour toutes les configurations logicielles. Les alignements BLAST ont été effectués par rapport à la base de données de référence des séquences de protéines UniProtKB/Swiss-Prot (The UniProt Consortium, 2019, 2009).

4.5 Reconstruction du métabolisme

Les bases de données métaboliques de BioCyc du mâle et de la femelle peuvent être visualisées sur une plateforme interactive (BioCyc) qui permet d'explorer le métabolisme dans sa globalité et les liens entre les différentes voies sont accessibles. Pour caractériser les voies du ML, nous avons extrait de la base de données toutes les voies se trouvant dans les

catégories fonctionnelles relatives à la biosynthèse (« Biosynthesis > Fatty Acid and Lipid Biosynthesis ») et la dégradation des lipides (« Degradation/Utilization/Assimilation > Fatty Acid and Lipid Degradation »). La voie de biosynthèse du mévalonate a été ajoutée à la sélection, car c'est le métabolite clé pour la synthèse des terpénoïdes, lipides ramifiés parmi lesquels on retrouve les hormones juvéniles des arthropodes. Cette sous-sélection a été effectuée grâce à la fonction « Browse Pathway Ontology » et au système des SmartTables qui permet le croisement de différentes caractéristiques. Cette stratégie permet d'être plus facilement exhaustif et reproductible d'un organisme à l'autre, car le recensement des voies et des enzymes annotées s'effectue à partir des grandes catégories fonctionnelles et non des voies individuelles qui peuvent varier ou ne pas être retrouvées selon l'espèce.

Pour analyser la complétude des voies métaboliques, nous avons utilisé la plateforme en ligne MetExplore (Cottret et al., 2018). Cette plateforme permet l'exploration des voies métaboliques et l'analyse de données omiques de façon interactive. Les identifiants des voies du ML extraites de nos bases de données GAMFO_TGFBM et GAMFO_TGFBF ont été mappés sur le métabolisme entier pour récupérer les réactions enzymatiques correspondantes. Après le mapping de nos voies sur le métabolisme global, nous avons obtenu les données de la complétude des annotations des voies du métabolisme lipidique.

Les deux bases de données GAMFO_TGFBM et GAMFO_TGFBF sont accessibles sur demande.

4.6 Intégration des données protéomiques

Les protéines identifiées dans chaque organe ont été dénombrées à partir des peptides retrouvés dans les échantillons (Table S1 dans Supplementary Data 1, Supplementary Data 2). Pour pouvoir identifier les enzymes dont l'expression peut être faible et spécifique à un

organe, seules les protéines identifiées par un spectral count et présentes dans au moins 3 échantillons ont été retenues (Table S2 dans Supplementary Data 1, Supplementary Data 2). À partir de cette table filtrée, l'analyse différentielle de l'abondance des protéines a été réalisée à l'aide du package R EdgeR (Robinson et al., 2010) version 3.32.1. Bien que cette fonction ait été initialement conçue pour les données de comptage de RNA-Seq, elle s'applique également aux données de comptage spectral en protéomique (Gregori et al., 2013). La sélection des protéines différentiellement exprimées a été basée sur un seuil de FDR (False Discovery Rate : taux de fausse découverte) de 0.05 et d'un changement absolu d'expression de 2 (|Fold Change| > 2).

5. RESULTATS ET DISCUSSION

5.1 Amélioration des transcriptomes

Réduction de la redondance des isoformes

Nous avons utilisé les transcriptomes mâle (GFBM) et femelle (GFBF) provenant de Cogne et al. (2019), car ils étaient les plus récents et chaque individu avait été génotypé. Ils sont composés de 344409 et 325379 transcrits et correspondent à 177384 et 154449 ORFs (Open Reading Frame) respectivement (Tableau 8). Les deux assemblages originaux ont été assemblés par Cogne et al. (2019) à l'aide de Trinity v2.4 (Grabherr et al., 2011) pour constituer une base de données la plus exhaustive possible et pouvoir assigner un maximum de spectre de masse lors de l'acquisition des protéomes. Cette stratégie implique la présence de transcrit et isoformes redondantes et de potentiels artefacts de transcrits issus d'erreurs d'assemblage. Ainsi, nous avons choisi d'exclure les possibles artefacts et de réduire la redondance, en prenant en compte le niveau d'expression des contigs et en ne gardant que le transcrit le plus fortement exprimé (Haas et al., 2013). Une fois filtrés, les deux nouveaux transcriptomes GFBM et GFBF contiennent 77% et 67% des transcrits originaux,

correspondant à 69% et 67% des ORFs originaux respectivement (Tableau 8). L'annotation GO a permis de conserver un peu plus de la moitié (53% et 51%) des annotations, de même pour les annotations EC (58% et 54%), pour le mâle et la femelle respectivement. La collecte des annotations fonctionnelles a permis la reconstruction des voies métaboliques, et de mapper 8258 et 6748 enzymes pour le mâle et la femelle sur ces voies. La réduction de la complexité de nos assemblages a permis de conserver l'information essentielle pour la reconstruction des voies métaboliques, en conservant une grande partie des transcriptomes de départ et des annotations associées. Dans la base de données de *Drosophila melanogaster* (v 1.0.1), on compte 4965 enzymes annotées (Baa-Puyoulet et al., 2016; McQuilton et al., 2012). Dans celle de *Daphnia pulex* (v 1.0) on compte 3672 enzymes (Baa-Puyoulet et al., 2016; Nordberg et al., 2014). Ces deux espèces ayant des génomes annotés, le nombre important d'enzymes retrouvées chez *G. fossarum* à partir des transcriptomes peut être expliqué par la présence de plusieurs transcrits annotés pour la même enzyme, et ce malgré la réduction de la redondance de nos assemblages. Notre méthode de filtrage ne permet pas d'identifier des artefacts d'annotation. D'un point de vue biologique, si deux isoformes sont annotées par la même réaction enzymatique, l'outil Pathway Tools conserve les deux annotations. Dans ce cas, la présence d'isozymes est également envisageable.

Tableau 8 : Statistiques des transcriptomes GFBF et GFBM avant et après réduction de la redondance.

Individu	Transcrits		ORFs		Annotation GO (Interproscan)		Annotations EC (Blast2GO, KASS-KEGG, Priam)	
	Avant filtrage	Après filtrage	Avant filtrage	Après filtrage	Avant filtrage	Après filtrage	Avant filtrage	Après filtrage
GFBM	325379	213220	154449	85939	210333	106396	38697	21015
GFBF	344409	230507	177384	105108	236602	126154	42127	24803

Qualité des transcriptomes

Les résultats BUSCO montrent que plus de 90% des 1013 gènes orthologues sont retrouvés dans les 4 transcriptomes assemblés, cela indique que les assemblages des transcriptomes *G. fossarum* contiennent l'information biologique attendue par rapport à leur appartenance phylogénétique (Figure 36). Environ 40% des gènes complets sont retrouvés en copie unique dans les transcriptomes originaux, contre environ 60% dans les transcriptomes filtrés (Figure 36). On constate donc que le filtrage de nos transcriptomes a permis l'augmentation de la présence de gènes en copie unique, avec une légère augmentation (d'environ 2%) des gènes manquants. La part de gènes dupliqués pourrait venir de différents haplotypes, ou isoformes encore présents et donc d'un résidu de redondance des assemblages que la méthode choisie n'aurait pas permis d'éliminer. On peut aussi noter l'absence du sous-embranchement des crustacés dans la base de données de gènes orthologues. En effet, on note uniquement la présence de la classe des insectes et celle des arachnides. Or, la complétude des transcriptomes est corrélée positivement à la proportion de gènes BUSCO complets, mais cette évaluation peut être biaisée par le nombre d'espèces proches de l'espèce d'intérêt (Amil-Ruiz et al., 2021; Seppey et al., 2019).

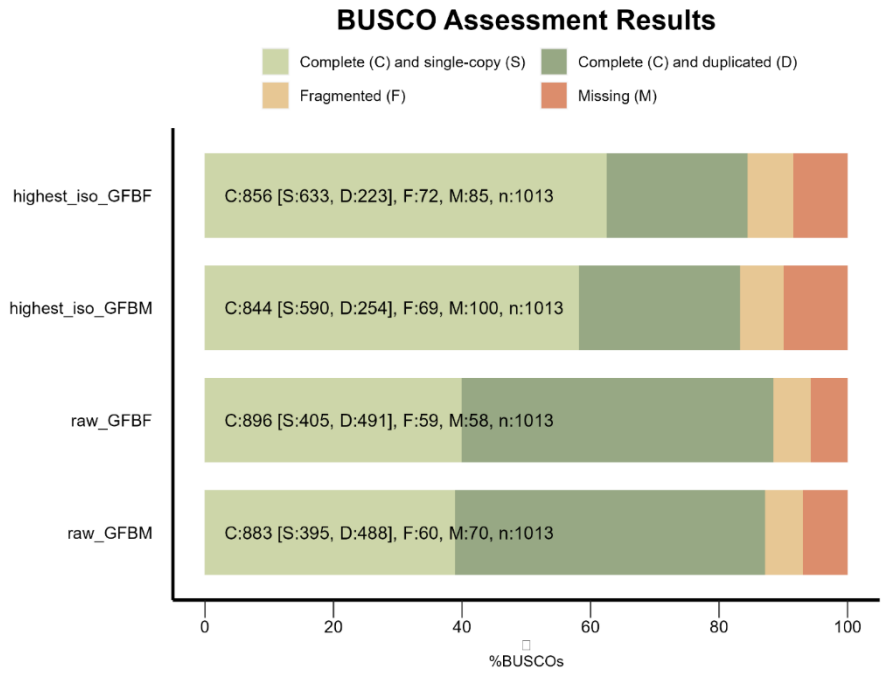


Figure 36 : Analyse de la complétude des transcriptomes originaux (raw) et filtrés (highest_iso) du mâle et de la femelle de *G. fossarum* via BUSCO des arthropodes.

5.2 Création des bases de données du métabolisme de *Gammarus fossarum*

Les transcriptomes filtrés ont ensuite été annotés par le pipeline CycADS (Vellozo et al., 2011). Deux bases de données métaboliques de *G. fossarum*, appelées GAMFO_TGFBF pour le mâle et GAMFO_TGFBM pour la femelle, ont été obtenues à partir de l'analyse de 77967 et 94499 séquences de polypeptides (Tableau S1). Ces chiffres diffèrent du nombre de séquences d'ORFs de départ (85939 et 105108) étant donné que pour un transcrit il est possible d'avoir plusieurs ORFs donc plusieurs protéines/polypeptides, certains ont donc été éliminés, car redondants.

Ces deux bases de données contiennent au total 358 et 366 voies annotées respectivement, dont 349 voies communes aux deux genres (Tableau S2), 9 voies identifiées uniquement chez la femelle et 17 uniquement chez le mâle (Tableau S3). Ces voies métaboliques sont composées de 2764 et 2850 réactions enzymatiques, pour le mâle et la

femelle respectivement (Tableau S1). La grande majorité des voies étant communes, l'annotation et la reconstruction des voies métaboliques globales à partir des données transcriptomiques des deux individus semblent cohérentes. Sur la totalité des voies, on observe 323 voies (82%) chez le mâle et 307 voies (80%) chez la femelle, contenant plus de 75% de réactions annotées avec des enzymes. Seules 3 voies (1%) chez le mâle et 4 (1%) chez la femelle contiennent moins de 25% de réactions annotées (Figure 37). Les 3 voies retrouvées chez le mâle sont équivalentes chez la femelle. Celles-ci sont impliquées dans la biosynthèse d'eumélanine, la biosynthèse des clusters fer-soufre et la nitrosylation /denitrosylation des protéines. Les voies de biosynthèse d'eumélanine sont possiblement mal annotées, car il a été montré récemment dans une étude que la mélanine cuticulaire des insectes était différente de celle des mammifères (Barek et al., 2018). La quatrième voie peu annotée chez la femelle n'est pas du tout retrouvée chez le mâle, et est impliquée dans la biosynthèse de phéomélanine.

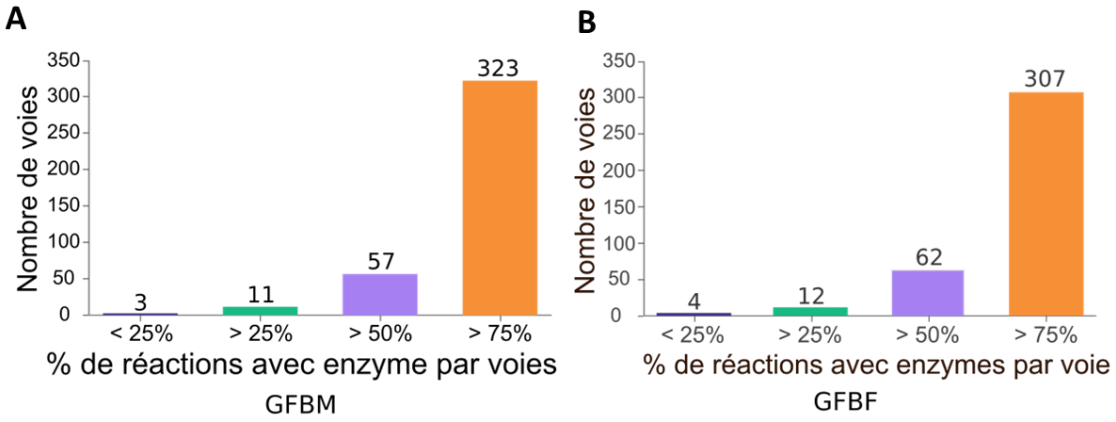


Figure 37 : Complétude de l'annotation des voies métaboliques via MetExplore, (A) chez le mâle GFBM et (B) la femelle GFBF.

BioCyc contient d'autres bases de données métaboliques d'arthropodes, notamment d'organismes modèles tels que l'insecte diptère, *Drosophila melanogaster* (Baa-Puyoulet et

al., 2019). Nous avons donc comparé nos résultats avec les données de cette espèce. Le métabolisme global reconstruit de *D. melanogaster* contient 2689 réactions enzymatiques pour 247 voies, ainsi que 3175 enzymes mappées sur l'ensemble des voies métaboliques (Figure S1). En ce qui concerne la complétude de l'annotation de la globalité des voies, 171 voies (69%) contiennent plus de 75% de réactions avec des enzymes annotées et 5 voies (2%) contiennent moins de 25% de réactions avec des enzymes annotées (Figure S1).

Les résultats obtenus pour le mâle et la femelle de *G. fossarum*, sont qualitativement similaires à ceux d'un organisme modèle pour lequel la reconstruction des voies métaboliques globale a été réalisée à l'aide d'un génome. Ces travaux montrent que l'exploitation de données transcriptomiques chez une espèce non modèle est donc faisable en adaptant l'outil CycADS à ce type de données, et ouvre de nouvelles perspectives pour d'autres espèces non modèles dont les données sont déjà disponibles.

5.3 Intégration des données protéomiques pour l'annotation métabolique

Après avoir obtenu une description du ML à partir de l'annotation des transcriptomes, nous avons combiné les données protéomiques dans le but de valider la présence d'enzymes vues en spectrométrie de masse. L'annotation fonctionnelle et les approches omiques permettent d'accélérer l'identification de gènes et de leurs produits, et de décrire les relations fonctionnelles entre ceux-ci (Ge et al., 2003). Cependant, les données provenant d'une source omique unique peuvent contenir des biais tels que des faux négatifs ou des faux positifs. C'est pourquoi il est important d'interpréter les résultats émanant d'une source de données omique unique avec précaution (Ge et al., 2003). Pour pallier ces limitations, l'intégration de plusieurs sources de données omiques est recommandée (Ge et al., 2003; Reeves et al., 2009).

Le métabolisme global

Les profils protéomiques globaux sont composés de 5062 protéines chez le mâle et 5663 chez la femelle (Table S2 dans Supplementary Data 1, Supplementary Data 2). Chez le mâle, la présence de 1189 protéines annotées sur les voies métaboliques a été validée en spectrométrie de masse (MS), cela représente 14,4% des enzymes mappées (8258) et 23,5% des protéines identifiées par MS. Chez la femelle, 1320 protéines annotées sur les voies métaboliques sont validées en MS, cela représente 19,6% des enzymes mappées (6748) et 23,3 % des protéines identifiées par MS. Nous retrouvons plus de protéines dont la présence est validée en MS chez la femelle que chez le mâle. La femelle possède plus de protéines vues en MS, même si le nombre de protéines annotées est supérieur chez le mâle. Nous ne pouvons pas attribuer cette différence à un réel effet biologique, mais plutôt à un effet base de données étant donné la présence de deux transcriptomes différents comme référence.

Le métabolisme lipidique

Nous nous sommes focalisés sur l'identification des voies métaboliques impliquées dans le ML de *G. fossarum* en combinant les annotations transcriptomiques et les données de protéomique. Nous observons chez le mâle 78 voies dont 447 réactions et 1108 enzymes uniques (*i.e.* une enzyme peut être impliquée dans différentes réactions enzymatiques) annotées comme appartenant au ML. Chez la femelle, nous comptons 76 voies, dont 436 réactions et 802 enzymes uniques (Tableau 9). Les 76 voies sont communes aux deux genres. Seules deux voies ont été annotées spécifiquement chez le mâle : « fatty acid β-oxidation II (plant peroxysome) » (95 enzymes annotées) et « superpathway of geranylgeranyldiposphate biosynthesis » (68 enzymes annotées) (Figure S2*).* Parmis ces deux voies retrouvées chez le mâle, seule la voie de la biosynthèse du géranyldiphosphate est

retrouvée chez la drosophile et la daphnie. Cette voie appartient à la classe des supervoies de la biosynthèse des diterpénoïdes.

Tableau 9 : Caractéristiques de la reconstruction du métabolisme lipidique chez *Gammarus fossarum* par CycADS pour le mâle (GAMFO_TGFBM) et la femelle (GAMFO_TGFBF).

PGDB	Voies métaboliques	Réactions enzymatiques	Enzymes
GAMFO_TGFBM	78	447	1108
GAMFO_TGFBF	76	436	802

Sur la totalité des voies du métabolisme lipidique, nous observons 63 voies (81%) composées de plus de 75% de réactions annotées avec des enzymes, et aucune voie n'en contient moins de 25% (Figure 38). Nous observons la même tendance chez la femelle avec 60 voies (79%) composées de plus de 75% de réactions annotées avec des enzymes (Figure 38).

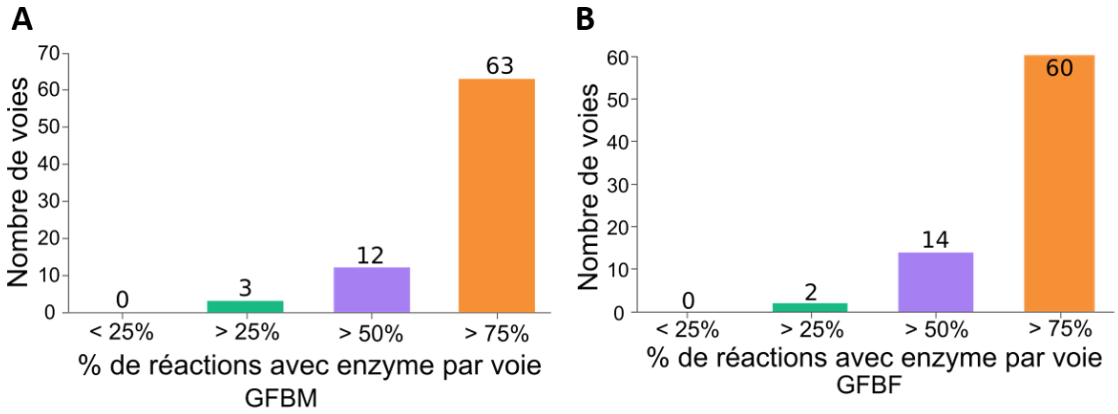


Figure 38 : Complétude de la reconstruction des voies du métabolisme lipidique pour le mâle (GFBM) et la femelle (GFBF) de *Gammarus fossarum*.

À titre de comparaison avec des organismes modèles, nous avons recensé le nombre de voies impliquées dans le métabolisme lipidique au même titre que notre sélection des catégories GO d'intérêt. On observe 62 voies du métabolisme lipidique chez *Daphnia pulex*,

un crustacé d'eau douce (Baa-Puyoulet et al., 2015; Nordberg et al., 2014) et 61 voies du ML chez *Drosophila melanogaster*, un insecte diptère (Baa-Puyoulet et al., 2019; McQuilton et al., 2012). Ainsi, on retrouve plus de voies chez le mâle et la femelle de *G. fossarum* que chez ces deux espèces modèles. La plupart des voies sont communes aux quatre organismes, entre 1 et 13% des voies de la drosophile et de la daphnie manquent chez GFBM ou GFBF. Ces comparaisons montrent la pertinence de notre stratégie pour la reconstruction des voies métaboliques d'un organisme non modèle à partir du transcriptome, et ce même si de potentiels effets bases de données sont à prendre en compte.

Chez le mâle, la présence de 155 protéines annotées comme impliquées dans les voies du métabolisme lipidique a été validée en MS. Cela représente 14% des enzymes mappées sur les voies métaboliques (1108) (Figure 39). Chez la femelle, 161 protéines putativement annotées sont validées en MS, elles représentent 20% des enzymes mappées sur les voies métaboliques (802) (Figure 39).

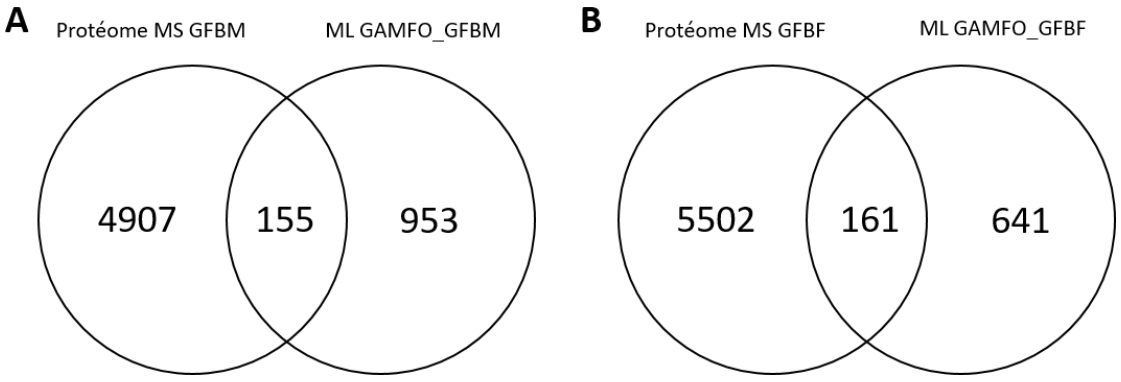


Figure 39 : Diagramme de Venn des protéines du métabolisme lipidique mappées dans GamfoCyc et les protéines présentes en spectrométrie de masse (A) pour le mâle et (B) la femelle. « ML GAMFO_GFBM et ML GAMFO_GFBF » représentent les bases de données du métabolisme lipidique. « Protéome MS GFBM/GFBF » représente les protéines retrouvées en MS dans les différents organes de *G. fossarum*, tous métabolismes confondus.

5.4 Les profils d'expression dans les organes

La composition des profils protéomiques

Les données protéomiques étant collectées à partir d'organes de *G. fossarum*, il est intéressant de décrire la composition des profils protéomiques et la présence de voies particulières du métabolisme lipidique dans les différents organes. On s'intéresse tout d'abord au nombre de protéines présentes dans les différents échantillons et organes. La Figure 40 présente pour le mâle et la femelle, les profils d'abondance de chaque protéine annotée comme faisant part d'au moins une voie du métabolisme lipidique de *G. fossarum* pour chaque réplicat d'organe.

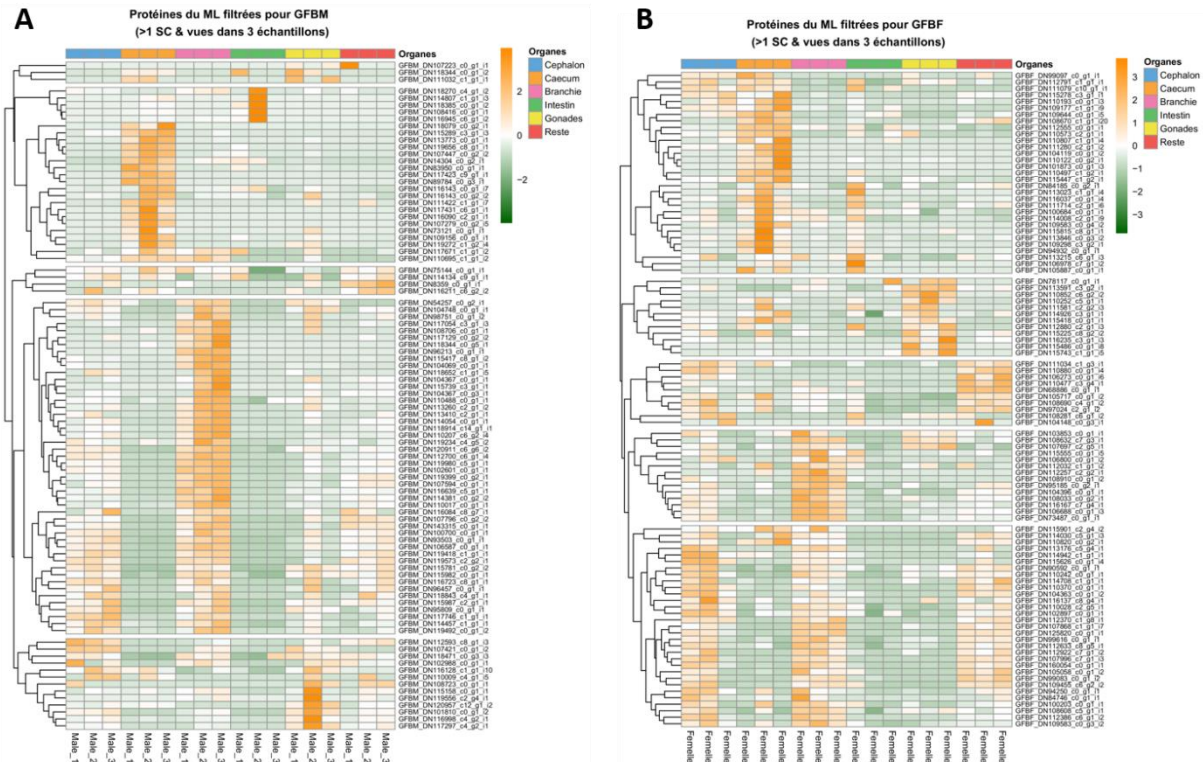


Figure 40 : Répartition des protéines vues en spectrométrie de masse selon organes et échantillons. (A) GFBM, (B) GFBF.

Les protéines de départ du ML (155 pour GFBM et 161 pour GFBM) ont été filtrées par le seuil de spectral count (≥ 1) et le nombre d'échantillons minimum dans lequel les protéines sont présentes (≥3). Nous avons choisi un filtrage relativement large des spectral count pour

une première approche de découverte et ceux-ci ont été normalisés. Il reste donc 93 protéines pour GFBM et 98 protéines pour GFBF pour le reste de l'analyse.

Le nombre de protéines du ML retrouvées dans les organes est du même ordre de grandeur entre le mâle et la femelle (Figure S2). La Figure 40 permet d'apprécier la répartition de l'abondance des protéines entre réplicats, individus et organes. La Figure 41 présente les tendances que l'on peut observer sur la Figure 40. Les distances entre échantillons ont été représentées à l'aide des diagrammes de positionnement multidimensionnel (MDS ou « Multi Dimensional Scaling »). Tous les échantillons et tous les organes contiennent des protéines du ML validées en MS chez le mâle et la femelle. Nous pouvons observer que les 3 réplicats de chaque organe sont assez homogènes, sauf pour les intestins (Figure 40, Figure S2). Cette observation est associée à une faible présence de protéines extraites des intestins, qui se reflète par un déficit en spectral count par rapport aux autres organes chez le mâle (2 voire 3 fois moins de protéines présentes), et chez la femelle (Figure 40, Figure S2). Les réplicats des autres organes tels que les branchies, les céphalons, les restes ou encore les testicules sont assez proches chez le mâle, ce qui suggère une faible variabilité de la composition en protéines du métabolisme lipidique entre ces organes. Chez la femelle, nous observons une meilleure séparation des organes (variabilité inter organes importante) (Figure 40B, Figure 41B).

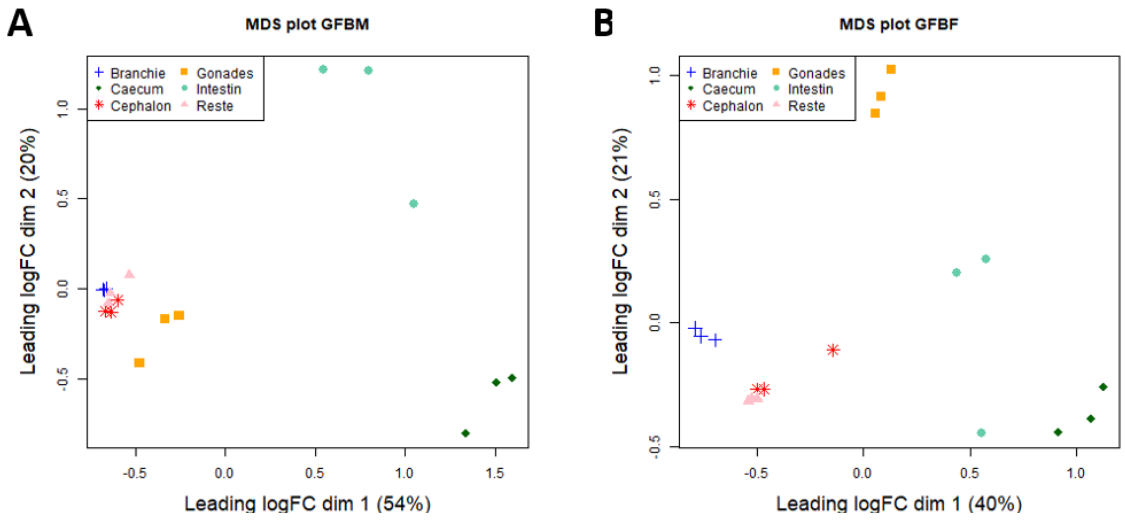


Figure 41 : Diagramme multidimensional scaling (MDS) (ou positionnement multidimensionnel) des échantillons de protéines pour le mâle (A) et la femelle (B) de *G. fossarum*. Clustering non supervisé des échantillons de protéines.

On observe chez le mâle et la femelle, 3 profils d'expression de protéines qui distinguent les branchies, les caeca et les gonades (testicules et ovaires) (Figure 40). La Figure 40 et la Figure 41 montrent donc un organotropisme marqué des différentes enzymes du ML dans les branchies, les caeca et les gonades (testicules et ovaires) chez le mâle et la femelle.

Analyse différentielle chez le mâle et la femelle de G. fossarum

Dans un premier temps nous avons commencé par analyser les différences entre les branchies et le reste des organes (Figure 42, Figure S3) (Supplementary Data 3). Chez le mâle, nous avons trouvé 39 protéines plus présentes (FDR<0.05, FC>2) dans les branchies (Figure 42A, Figure S3A). Chez la femelle, nous avons trouvé 26 protéines plus abondantes dans les branchies (FDR<0.05 et FC>2) (Figure 42B, Figure S3B).

Nous avons ensuite testé l'expression différentielle des protéines dans les gonades versus le reste des organes chez le mâle et la femelle (Figure 43, Figure S4) (Supplementary Data 3). Chez le mâle, 31 protéines sont plus abondantes dans les testicules (FDR<0.05 et

FC>2) (Figure 43A, Figure S4A). Chez la femelle, 21 protéines DE sont surreprésentées dans les ovaires (FDR<0.05 et FC>2) (Figure 43B, Figure S4B).

Parmi les protéines les plus abondantes dans les branchies et les gonades du mâle et de la femelle, on retrouve des protéines annotées comme intervenant dans les voies de β-oxydation des AGs, dans la synthèse des PUFAs ou encore dans la dégradation des plasmalogènes. La β-oxydation des acides gras (AGs) permet de fournir de l'énergie à l'organe en brulant les acides gras, et les PUFAs constituent majoritairement les membranes et sont retrouvés chez les mammifères, notamment dans les testicules de rat (Salem et al., 2015). Quant aux plasmalogènes, ils ont été décrits dans la littérature comme les plus présents dans le céphalon, les nerfs, les testicules et les branchies et peuvent jouer un rôle dans la perméabilité des membranes (Chapelle, 1987; Nagan and Zoeller, 2001; Rapport and Alonzo, 1960). Ils appartiennent à une catégorie de glycérophospholipides membranaires qui semblent jouer un rôle important dans la lutte contre les oxydants puisque les cellules qui en sont dépourvues par mutation génétique montrent une hypersensibilité vis-à-vis des oxydants (Aboshi et al., 2012; Nagan and Zoeller, 2001). Les plasmalogènes sont connus aussi pour inhiber l'oxydation induite par les ions métalliques (Aboshi et al., 2012). Une étude récente de Gestin et al. (2021), a montré que la branchie bioaccumule le cadmium sans le dépurer. Ces lipides pourraient jouer un rôle dans la protection contre l'exposition aux métaux lourds et au stress oxydant dans un organe constitutif exposé aux variations environnementales.

On note aussi plus spécifiquement la présence dans les gonades du mâle et de la femelle de voies impliquées dans la synthèse de l'anandamide (*i.e.* composé cannabinoïde endogène), présent principalement dans le cerveau, le rein et les testicules, notamment chez

les invertébrés, les vertébrés et les mammifères (Battista et al., 2012). Une étude a permis de découvrir des récepteurs à cannabinoïde chez un crustacé et d'autres espèces invertébrés (McPartland et al., 2006). Il a été montré que l'activité testiculaire pouvait être régulée par l'anandamide chez la grenouille *Pelophylax esculentus* par le système de la kisspeptine (Ciaramella et al., 2016). On retrouve également dans les gonades du mâle et de la femelle des protéines intervenant dans le métabolisme de la sphingosine. Les céramides (produits de la dégradation des sphingomyélines) sont métabolisés par la céramidase pour générer la sphingosine. La sphingosine est produite au cours des premiers stades de l'apoptose et permet sa régulation (Cuvillier, 2002; Pettus et al., 2002). Cette enzyme pourrait être un facteur maternel impliqué dans le développement embryonnaire comme la prolifération, la morphogenèse ou l'apoptose (Sinha Hikim and Swerdloff, 1999).

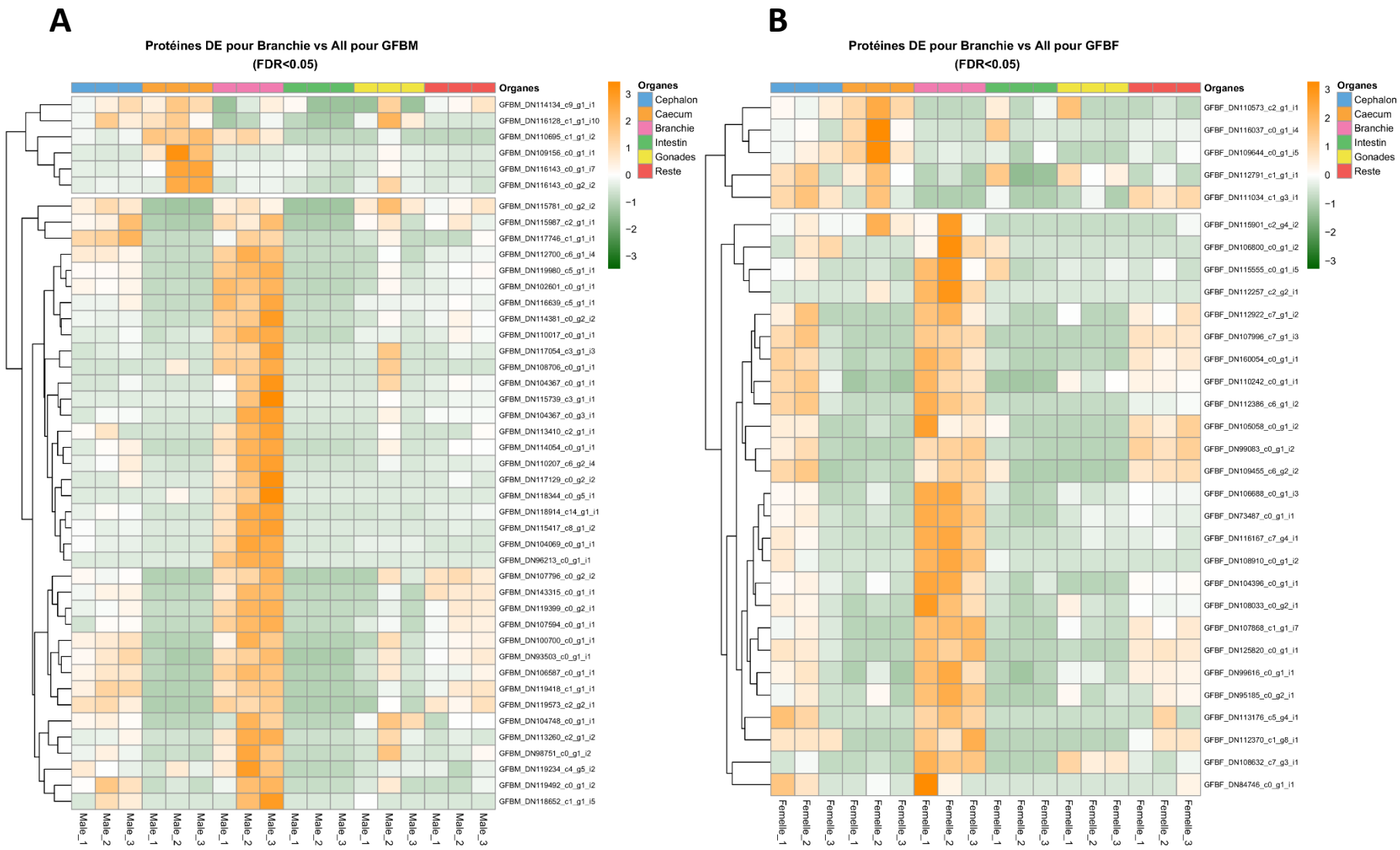


Figure 42 : Abondances différentielles des protéines pour les branchies versus le reste des organes (FDR < 0.05, LFC>1). (A) GFBM, (B) GFBF.

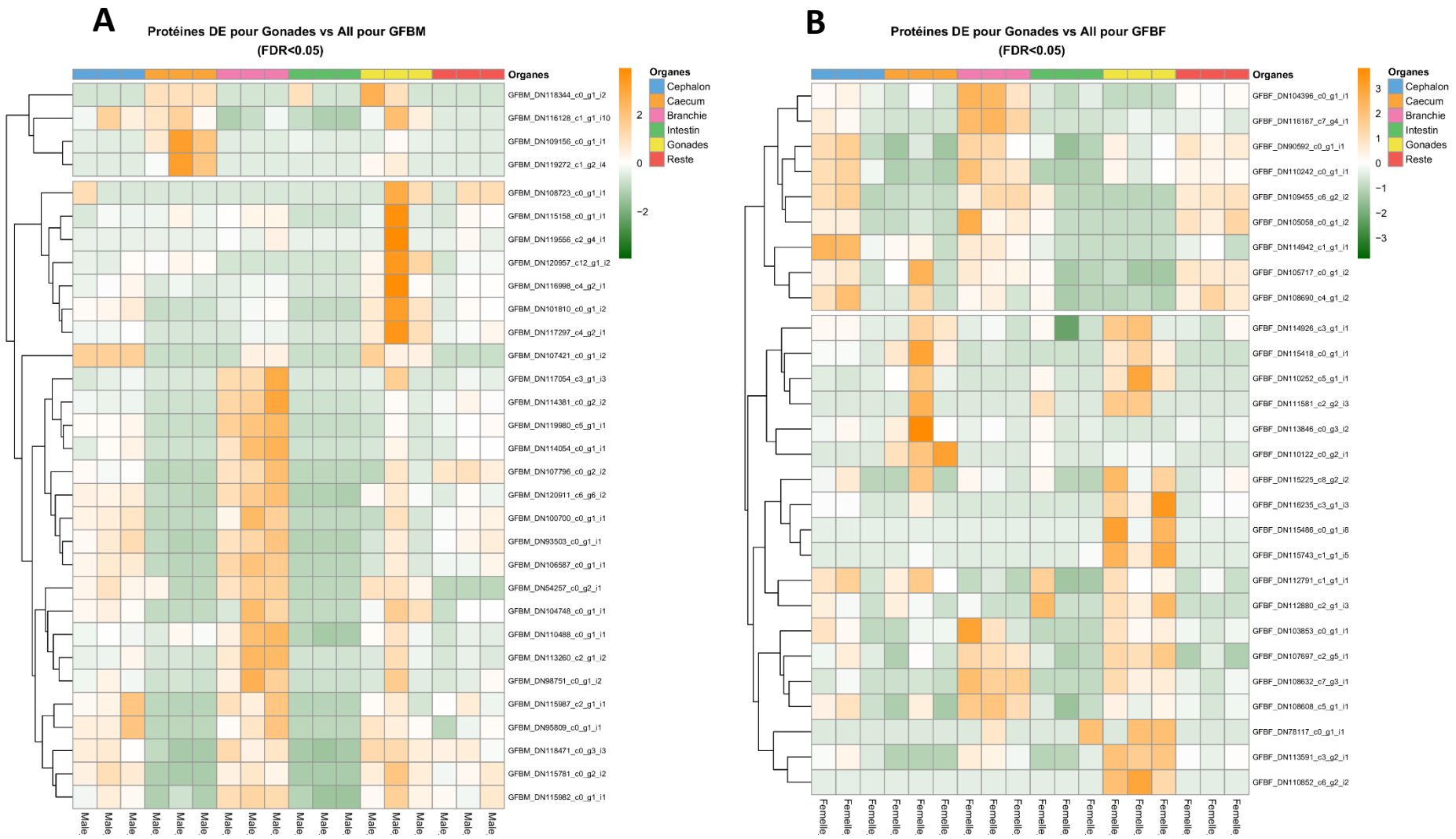


Figure 43 : Abondances différentielles des protéines pour les gonades versus le reste des organes (FDR < 0.05,LFC > 1). (A) GFBM, (B) GFBF.

Notre analyse sans *a priori* des profils des enzymes lipidiques dans différents organes montre un organotropisme spécifique des voies du métabolisme lipidique chez *G. fossarum*. Notamment, les branchies sont caractérisées par une dégradation des plasmalogènes, lipides impliqués dans la défense antioxydante et les gonades par une capacité à synthétiser l'anandamide, un endocannabinoïde récemment identifié comme médiateur lipidique impliqué dans différents aspects de la reproduction et conservés chez des taxons divergents. Ces résultats soulignent l'intérêt de l'application d'approches omiques au niveau des organes des espèces sentinelles pour identifier et évaluer les possibles différences de MoA de contaminants en fonction de l'organe ciblé.

6. REFERENCES

Aboshi, T., Nishida, R., Mori, N., 2012. Identification of plasmalogen in the gut of silkworm (Bombyx mori). Insect Biochem Mol Biol 42, 596–601. https://doi.org/10.1016/j.ibmb.2012.04.006

Aite, M., Chevallier, M., Frioux, C., Trottier, C., Got, J., Cortés, M.P., Mendoza, S.N., Carrier, G., Dameron, O., Guillaudeux, N., Latorre, M., Loira, N., Markov, G.V., Maass, A., Siegel, A., 2018. Traceability, reproducibility and wiki-exploration for "à-la-carte" reconstructions of genome-scale metabolic models. PLOS Computational Biology 14, e1006146. https://doi.org/10.1371/journal.pcbi.1006146

Amil-Ruiz, F., Maria Herruzo-Ruiz, A., Fuentes-Almagro, C., Baena-Angulo, C., Manuel Jimenez-Pastor, J., Blasco, J., Alhama, J., Michan, C., 2021. Constructing a de novo transcriptome and a reference proteome for the bivalve Scrobicularia plana: Comparative analysis of different assembly strategies and proteomic analysis. Genomics 113, 1543–1553. https://doi.org/10.1016/j.ygeno.2021.03.025

Arambourou, H., Fuertes, I., Vulliet, E., Daniele, G., Noury, P., Delorme, N., Abbaci, K., Barata, C., 2018. Fenoxycarb exposure disrupted the reproductive success of the amphipod Gammarus fossarum with limited effects on the lipid profile. PLOS ONE 13, e0196461. https://doi.org/10.1371/journal.pone.0196461

Armengaud, J., Trapp, J., Pible, O., Geffard, O., Chaumot, A., Hartmann, E.M., 2014. Non-model organisms, a species endangered by proteogenomics. Journal of Proteomics, Special Issue: Proteomics of non-model organisms 105, 5–18. https://doi.org/10.1016/j.jprot.2014.01.007

Baa-Puyoulet, P., Parisot, N., Charles, H., Calevro, F., 2019. ArthropodaCyc - Drosophila melanogaster [WWW Document]. Summary of Drosophila melanogaster, version 1.0.1. URL http://arthropodacyc.cycadsys.org/organism-summary?object=DROME (accessed 12.18.21).

Baa-Puyoulet, P., Parisot, N., Colella, S., 2015. ArthropodaCyc - Daphnia pulex [WWW Document]. Summary of Daphnia pulex, version 1.0. URL http://arthropodacyc.cycadsys.org/organism-summary?object=DAPPU (accessed 12.18.21).

Baa-Puyoulet, P., Parisot, N., Febvay, G., Huerta-Cepas, J., Vellozo, A.F., Gabaldón, T., Calevro, F., Charles, H., Colella, S., 2016. ArthropodaCyc: a CycADS powered collection of BioCyc databases to analyse and compare metabolism of arthropods. Database (Oxford) 2016. https://doi.org/10.1093/database/baw081

Banks, J.E., Stark, J.D., 1998. What is ecotoxicology? An ad-hoc grab bag or an interdisciplinary science? Integrative Biology: Issues, News, and Reviews: Published in Association with The Society for Integrative and Comparative Biology 1, 195–204.

Barek, H., Sugumaran, M., Ito, S., Wakamatsu, K., 2018. Insect cuticular melanins are distinctly different from those of mammalian epidermal melanins. Pigment Cell Melanoma Res 31, 384–392. https://doi.org/10.1111/pcmr.12672

Barros, S., Coimbra, A.M., Alves, N., Pinheiro, M., Quintana, J.B., Santos, M.M., Neuparth, T., 2020. Chronic exposure to environmentally relevant levels of simvastatin disrupts zebrafish brain gene signaling involved in energy metabolism. J Toxicol Environ Health A 83, 113–125. https://doi.org/10.1080/15287394.2020.1733722

Battista, N., Meccariello, R., Cobellis, G., Fasano, S., Di Tommaso, M., Pirazzi, V., Konje, J.C., Pierantoni, R., Maccarrone, M., 2012. The role of endocannabinoids in gonadal function and fertility along the evolutionary axis. Mol Cell Endocrinol 355, 1–14. https://doi.org/10.1016/j.mce.2012.01.014

Bellés, X., Martín, D., Piulachs, M.-D., 2005. The Mevalonate Pathway and the Synthesis of Juvenile Hormone in Insects. Annual Review of Entomology 50, 181–199. https://doi.org/10.1146/annurev.ento.50.071803.130356

Bonnefoy, C., Fildier, A., Buleté, A., Bordes, C., Garric, J., Vulliet, E., 2019. Untargeted analysis of nanoLC-HRMS data by ANOVA-PCA to highlight metabolites in Gammarus fossarum after in vivo exposure to pharmaceuticals. Talanta 202, 221–229. https://doi.org/10.1016/j.talanta.2019.04.028

Buikema Jr, A.L., Benfield, E.F., 1979. Use of macroinvertebrate life history information in toxicity tests. Journal of the Fisheries Board of Canada 36, 321–328.

Calow, P., Sibly, R.M., Forbes, V., 1997. Risk assessment on the basis of simplified life-history scenarios. Environmental Toxicology and Chemistry: An International Journal 16, 1983–1989.

Caspi, R., Billington, R., Keseler, I.M., Kothari, A., Krummenacker, M., Midford, P.E., Ong, W.K., Paley, S., Subhraveti, P., Karp, P.D., 2020. The MetaCyc database of metabolic pathways and enzymes - a 2019 update. Nucleic Acids Research 48, D445–D453. https://doi.org/10.1093/nar/gkz862

Caspi, R., Foerster, H., Fulcher, C.A., Kaipa, P., Krummenacker, M., Latendresse, M., Paley, S., Rhee, S.Y., Shearer, A.G., Tissier, C., Walk, T.C., Zhang, P., Karp, P.D., 2008. The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res 36, D623-631. https://doi.org/10.1093/nar/gkm900

Chapelle, S., 1987. Plasmalogens and O-alkylglycerophospholipids in aquatic animals. Comparative Biochemistry and Physiology Part B: Comparative Biochemistry 88, 1–6. https://doi.org/10.1016/0305-0491(87)90068-X

Chaumot, A., Geffard, O., Armengaud, J., Maltby, L., 2015. Gammarids as reference species for freshwater monitoring, in: Aquatic Ecotoxicology. Elsevier, pp. 253–280.

Christie-Oleza, J.A., Miotello, G., Armengaud, J., 2012. High-throughput proteogenomics of Ruegeria pomeroyi: seeding a better genomic annotation for the whole marine Roseobacter clade. BMC Genomics 13, 73. https://doi.org/10.1186/1471-2164-13-73

Ciaramella, V., Meccariello, R., Chioccarelli, T., Sirleto, M., Fasano, S., Pierantoni, R., Chianese, R., 2016. Anandamide acts via kisspeptin in the regulation of testicular activity of the frog, Pelophylax esculentus. Mol Cell Endocrinol 420, 75–84. https://doi.org/10.1016/j.mce.2015.11.011

Claudel-Renard, C., Chevalet, C., Faraut, T., Kahn, D., 2003. Enzyme-specific profiles for genome annotation: PRIAM. Nucleic acids research 31, 6633–6639.

Cogne, Y., Degli Esposti, D., Pible, O., Gouveia, D., Geffard, O., Chaumot, A., 2019a. YC02. figshare. https://doi.org/10.6084/m9.figshare.c.4568087.v1

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., 2019b. De novo transcriptomes of 14 gammarid individuals for proteogenomic analysis of seven taxonomic groups. Sci Data 6, 1–7. https://doi.org/10.1038/s41597-019-0192-5

Colbourne, J.K., Pfrender, M.E., Gilbert, D., Thomas, W.K., Tucker, A., Oakley, T.H., Tokishita, S., Aerts, A., Arnold, G.J., Basu, M.K., Bauer, D.J., Cáceres, C.E., Carmel, L., Casola, C., Choi, J.-H., Detter, J.C., Dong, Q., Dusheyko, S., Eads, B.D., Fröhlich, T., Geiler-Samerotte, K.A., Gerlach, D., Hatcher, P., Jogdeo, S., Krijgsveld, J., Kriventseva, E.V., Kültz, D., Laforsch, C., Lindquist, E., Lopez, J., Manak, J.R., Muller, J., Pangilinan, J., Patwardhan, R.P., Pitluck, S., Pritham, E.J., Rechtsteiner, A., Rho, M., Rogozin, I.B., Sakarya, O., Salamov, A., Schaack, S., Shapiro, H., Shiga, Y., Skalitzky, C., Smith, Z., Souvorov, A., Sung, W., Tang, Z., Tsuchiya, D., Tu, H., Vos, H., Wang, M., Wolf, Y.I., Yamagata, H., Yamada, T., Ye, Y., Shaw, J.R., Andrews, J., Crease, T.J., Tang, H., Lucas, S.M., Robertson, H.M., Bork, P., Koonin, E.V., Zdobnov, E.M., Grigoriev, I.V., Lynch, M., Boore, J.L., 2011. The Ecoresponsive Genome of Daphnia pulex. Science 331, 555–561. https://doi.org/10.1126/science.1197761

Conesa, A., Götz, S., 2008. Blast2GO: A Comprehensive Suite for Functional Analysis in Plant Genomics. International Journal of Plant Genomics 2008, 1–12. https://doi.org/10.1155/2008/619832

Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M., Robles, M., 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21, 3674–3676.

Correia, A.D., Costa, M.H., Luis, O.J., Livingstone, D.R., 2003. Age-related changes in antioxidant enzyme activities, fatty acid composition and lipid peroxidation in whole body Gammarus locusta (Crustacea: Amphipoda). Journal of Experimental Marine Biology and Ecology 289, 83–101. https://doi.org/10.1016/S0022-0981(03)00040-6

Cottret, L., Frainay, C., Chazalviel, M., Cabanettes, F., Gloaguen, Y., Camenen, E., Merlet, B., Heux, S., Portais, J.-C., Poupin, N., Vinson, F., Jourdan, F., 2018. MetExplore: collaborative edition and exploration of metabolic networks. Nucleic Acids Res 46, W495–W502. https://doi.org/10.1093/nar/gky301

Cuvillier, O., 2002. Sphingosine in apoptosis signaling. Biochim Biophys Acta 1585, 153–162. https://doi.org/10.1016/s1388-1981(02)00336-0

Fu, T., Knittelfelder, O., Geffard, O., Clement, Y., Testet, E., Elie, N., Touboul, D., Abbaci, K., Shevchenko, A., Lemoine, J., Chaumot, A., Salvador, A., Degli-Esposti, D., Ayciriex, S., 2021. Shotgun lipidomics and mass spectrometry imaging unveil diversity and dynamics in Gammarus fossarum lipid composition. iScience 24, 102115. https://doi.org/10.1016/j.isci.2021.102115

Fuertes, I., Jordão, R., Casas, J., Barata, C., 2018. Allocation of glycerolipids and glycerophospholipids from adults to eggs in Daphnia magna: Perturbations by compounds that enhance lipid droplet accumulation. Environ Pollut 242, 1702–1710. https://doi.org/10.1016/j.envpol.2018.07.102

Ge, H., Walhout, A.J.M., Vidal, M., 2003. Integrating "omic" information: a bridge between genomics and systems biology. Trends Genet 19, 551–560. https://doi.org/10.1016/j.tig.2003.08.009

Gestin, O., Lacoue-Labarthe, T., Coquery, M., Delorme, N., Garnero, L., Dherret, L., Ciccia, T., Geffard, O., Lopes, C., 2021. One and multi-compartments toxico-kinetic modeling to understand metals' organotropism and fate in Gammarus fossarum. Environment International 156, 106625.

GHCZ00000000.1 Gammarus fossarum female :: NCBI [WWW Document], 2018. URL https://www.ncbi.nlm.nih.gov/Traces/wgs/?val=GHCZ01 (accessed 12.16.21).

GHDA00000000.1 Gammarus fossarum male :: NCBI [WWW Document], 2018. URL https://www.ncbi.nlm.nih.gov/Traces/wgs/?val=GHDA01 (accessed 12.16.21).

Gismondi, E., Beisel, J.-N., Cossu-Leguille, C., 2012. Influence of gender and season on reduced glutathione concentration and energy reserves of Gammarus roeseli. Environmental Research 118, 47–52. https://doi.org/10.1016/j.envres.2012.06.004

Gómez-Canela, C., Miller, T.H., Bury, N.R., Tauler, R., Barron, L.P., 2016. Targeted metabolomics of Gammarus pulex following controlled exposures to selected pharmaceuticals in water. Science of The Total Environment 562, 777–788. https://doi.org/10.1016/j.scitotenv.2016.03.181

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

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., 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29, 644–652. https://doi.org/10.1038/nbt.1883

Gregori, J., Sánchez, A., Villanueva, J., 2013. msmsTests: LC-MS/MS Differential Expression Tests. R package version 1.14. 0.

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., MacManes, M.D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C.N., Henschel, R., LeDuc, R.D., Friedman, N., Regev, A., 2013. De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. Nat Protoc 8, 10.1038/nprot.2013.084. https://doi.org/10.1038/nprot.2013.084

Hanemaaijer, M., Olivier, B.G., Röling, W.F.M., Bruggeman, F.J., Teusink, B., 2017. Model-based quantification of metabolic interactions from dynamic microbial-community data. PLOS ONE 12, e0173183. https://doi.org/10.1371/journal.pone.0173183

Istvan, E.S., Deisenhofer, J., 2001. Structural Mechanism for Statin Inhibition of HMG-CoA Reductase. Science 292, 1160–1164. https://doi.org/10.1126/science.1059344

Jelic, A., Gros, M., Ginebreda, A., Cespedes-Sánchez, R., Ventura, F., Petrovic, M., Barcelo, D., 2011. Occurrence, partition and removal of pharmaceuticals in sewage water and sludge during wastewater treatment. Water Research 45, 1165–1176. https://doi.org/10.1016/j.watres.2010.11.010

Jiménez-Prada, P., Hachero-Cruzado, I., Guerra-García, J.M., 2021. Aquaculture waste as food for amphipods: the case of Gammarus insensibilis in marsh ponds from southern Spain. Aquacult Int 29, 139–153. https://doi.org/10.1007/s10499-020-00615-z

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics 30, 1236–1240.

Jordão, R., Campos, B., Piña, B., Tauler, R., Soares, A.M.V.M., Barata, C., 2016a. Mechanisms of Action of Compounds That Enhance Storage Lipid Accumulation in Daphnia magna. Environmental Science & Technology 50, 13565–13573. https://doi.org/10.1021/acs.est.6b04768

Jordão, R., Garreta, E., Campos, B., Lemos, M.F.L., Soares, A.M.V.M., Tauler, R., Barata, C., 2016b. Compounds altering fat storage in Daphnia magna. Sci Total Environ 545–546, 127–136. https://doi.org/10.1016/j.scitotenv.2015.12.097

Jordão Rita, Casas Josefina, Fabrias Gemma, Campos Bruno, Piña Benjamín, Lemos Marco F.L., Soares Amadeu M.V.M., Tauler Romà, Barata Carlos, 2015. Obesogens beyond Vertebrates: Lipid Perturbation by Tributyltin in the Crustacean Daphnia magna. Environmental Health Perspectives 123, 813–819. https://doi.org/10.1289/ehp.1409163

Kao, D., Lai, A.G., Stamataki, E., Rosic, S., Konstantinides, N., Jarvis, E., Di Donfrancesco, A., Pouchkina-Stancheva, N., Semon, M., Grillo, M., 2016. The genome of the crustacean Parhyale hawaiensis, a model for animal development, regeneration, immunity and lignocellulose digestion. elife 5, e20062.

Karp, P.D., Midford, P.E., Billington, R., Kothari, A., Krummenacker, M., Latendresse, M., Ong, W.K., Subhraveti, P., Caspi, R., Fulcher, C., Keseler, I.M., Paley, S.M., 2021. Pathway Tools version 23.0 update: software for pathway/genome informatics and systems biology. Briefings in Bioinformatics 22, 109–126. https://doi.org/10.1093/bib/bbz104

Karp, P.D., Paley, S.M., Krummenacker, M., Latendresse, M., Dale, J.M., Lee, T.J., Kaipa, P., Gilham, F., Spaulding, A., Popescu, L., 2010. Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. Briefings in bioinformatics 11, 40–79.

Kolanowski, W., Stolyhwo, A., Grabowski, M., 2007. Fatty Acid Composition of Selected Fresh Water Gammarids (Amphipoda, Crustacea): A Potentially Innovative Source of Omega-3 LC PUFA. Journal of the American Oil Chemists' Society 84, 827–833. https://doi.org/10.1007/s11746-007-1116-7

Konschak, M., Zubrod, J.P., Baudy, P., Fink, P., Kenngott, K.G.J., Englert, D., Röder, N., Ogbeide, C., Schulz, R., Bundschuh, M., 2021. Chronic effects of the strobilurin fungicide azoxystrobin in the leaf shredder Gammarus fossarum (Crustacea; Amphipoda) via two effect pathways. Ecotoxicology and Environmental Safety 209, 111848. https://doi.org/10.1016/j.ecoenv.2020.111848

Kriventseva, E.V., Kuznetsov, D., Tegenfeldt, F., Manni, M., Dias, R., Simão, F.A., Zdobnov, E.M., 2019. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. Nucleic Acids Research 47, D807–D811. https://doi.org/10.1093/nar/gky1053

Kühmayer, T., Guo, F., Ebm, N., Battin, T.J., Brett, M.T., Bunn, S.E., Fry, B., Kainz, M.J., 2020. Preferential retention of algal carbon in benthic invertebrates: Stable isotope and fatty acid evidence from an outdoor flume experiment. Freshwater Biology 65, 1200–1209. https://doi.org/10.1111/fwb.13492

Kunz, P., Kienle, C., Gerhardt, A., 2010. Gammarus spp. in Aquatic Ecotoxicology and Water Quality Assessment: Toward Integrated Multilevel Tests. Reviews of environmental contamination and toxicology 205, 1–76. https://doi.org/10.1007/978-1-4419-5623-1_1

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9, 357–359. https://doi.org/10.1038/nmeth.1923

Li, B., Dewey, C.N., 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12, 323. https://doi.org/10.1186/1471-2105-12-323

Liu, H., Sadygov, R.G., Yates, J.R., 2004. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. Anal Chem 76, 4193–4201. https://doi.org/10.1021/ac0498563

Machado, D., Andrejev, S., Tramontano, M., Patil, K.R., 2018. Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. Nucleic Acids Research 46, 7542–7553. https://doi.org/10.1093/nar/gky537

McPartland, J.M., Agraval, J., Gleeson, D., Heasman, K., Glass, M., 2006. Cannabinoid receptors in invertebrates. Journal of Evolutionary Biology 19, 366–373. https://doi.org/10.1111/j.1420-9101.2005.01028.x

McQuilton, P., St. Pierre, S.E., Thurmond, J., the FlyBase Consortium, 2012. FlyBase 101 – the bas ics of navigating FlyBase. Nucleic Acids Research 40, D706–D714. https://doi.org/10.1093/nar/gkr1030

Mendoza, S.N., Olivier, B.G., Molenaar, D., Teusink, B., 2019. A systematic assessment of current genome-scale metabolic reconstruction tools. Genome Biology 20, 158. https://doi.org/10.1186/s13059-019-1769-1

Meusy, J.J., 1980. Vitellogenin, the extraovarian precursor of the protein yolk in Crustacea: a review. Reproduction Nutrition Développement 20, 1–21.

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A.C., Kanehisa, M., 2007. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic acids research 35, W182–W185.

Nagan, N., Zoeller, R.A., 2001. Plasmalogens: biosynthesis and functions. Progress in Lipid Research 40, 199–229. https://doi.org/10.1016/S0163-7827(01)00003-0

Neuparth, T., Machado, A.M., Montes, R., Rodil, R., Barros, S., Alves, N., Ruivo, R., Castro, L.F.C., Quintana, J.B., Santos, M.M., 2020. Transgenerational inheritance of chemical-induced signature: A case study with simvastatin. Environment International 144, 106020. https://doi.org/10.1016/j.envint.2020.106020

Neuparth, T., Martins, C., Santos, C.B. de los, Costa, M.H., Martins, I., Costa, P.M., Santos, M.M., 2014. Hypocholesterolaemic pharmaceutical simvastatin disrupts reproduction and population growth of the amphipod Gammarus locusta at the ng/L range. Aquatic Toxicology 155, 337–347. https://doi.org/10.1016/j.aquatox.2014.07.009

Nijhout, H., 1994. Insect Hormones (Princeton, NJ: Princeton Uni- todes: themes and variations. Trends Genet 17, 206–213.

Nordberg, H., Cantor, M., Dusheyko, S., Hua, S., Poliakov, A., Shabalov, I., Smirnova, T., Grigoriev, I.V., Dubchak, I., 2014. The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. Nucleic Acids Research 42, D26–D31. https://doi.org/10.1093/nar/gkt1069

Pettus, B.J., Chalfant, C.E., Hannun, Y.A., 2002. Ceramide in apoptosis: an overview and current perspectives. Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids 1585, 114–125. https://doi.org/10.1016/S1388-1981(02)00331-1

Plaistow, S.J., Bollache, L., Cézilly, F., 2003. Energetically costly precopulatory mate guarding in the amphipod Gammarus pulex: causes and consequences. Animal behaviour 65, 683–691.

Poynton, H.C., Hasenbein, S., Benoit, J.B., Sepulveda, M.S., Poelchau, M.F., Hughes, D.S.T., Murali, S.C., Chen, S., Glastad, K.M., Goodisman, M.A.D., Werren, J.H., Vineis, J.H., Bowen, J.L., Friedrich, M., Jones, J., Robertson, H.M., Feyereisen, R., Mechler-Hickson, A., Mathers, N., Lee, C.E., Colbourne, J.K., Biales, A., Johnston, J.S., Wellborn, G.A., Rosendale, A.J., Cridge, A.G., Munoz-Torres, M.C., Bain, P.A., Manny, A.R., Major, K.M., Lambert, F.N., Vulpe, C.D., Tuck, P., Blalock, B.J., Lin, Y.-Y., Smith, M.E., Ochoa-Acuña, H., Chen, M.-J.M., Childers, C.P., Qu, J., Dugan, S., Lee, S.L., Chao, H., Dinh, H., Han, Y., Doddapaneni, H., Worley, K.C., Muzny, D.M., Gibbs, R.A., Richards, S., 2018. The Toxicogenome of Hyalella azteca: A Model for Sediment Ecotoxicology and Evolutionary Toxicology. Environmental Science & Technology 52, 6009–6022. https://doi.org/10.1021/acs.est.8b00837

Rapport, M.M., Alonzo, N.F., 1960. The Structure of Plasmalogens: V. LIPIDS OF MARINE INVERTEBRATES. Journal of Biological Chemistry 235, 1953–1956. https://doi.org/10.1016/S0021-9258(18)69342-1

Reeves, G.A., Talavera, D., Thornton, J.M., 2009. Genome and proteome annotation: organization, interpretation and integration. J R Soc Interface 6, 129–147. https://doi.org/10.1098/rsif.2008.0341

Rna-seq transcriptome Gammarus Fossarum B female [WWW Document], 2018. . NCBI Sequence Read Archive. URL https://www.ncbi.nlm.nih.gov/sra/SRR8089729 (accessed 12.16.21).

Rna-seq transcriptome Gammarus Fossarum B male [WWW Document], 2018. . NCBI Sequence Read Archive. URL https://www.ncbi.nlm.nih.gov/sra/SRR8089722 (accessed 12.16.21).

Robinson, M.D., McCarthy, D.J., Smyth, G.K., 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139–140.

Rosa, R., Nunes, M.L., 2002. Changes in Organ Indices and Lipid Dynamics during the Reproductive Cycle of Aristeus antennatus, Parapenaeus longirostris, and Nephrops norvegicus (Decapoda) from the Portuguese South Coast. Crustaceana 75, 1095–1105.

Salem, N.M., Lin, Y.H., Moriguchi, T., Lim, S.Y., Salem, N., Hibbeln, J.R., 2015. Distribution of omega-6 and omega-3 polyunsaturated fatty acids in the whole rat body and 25 compartments. Prostaglandins Leukot Essent Fatty Acids 100, 13–20. https://doi.org/10.1016/j.plefa.2015.06.002

Santos, M.M., Ruivo, R., Lopes-Marques, M., Torres, T., de los Santos, C.B., Castro, L.F.C., Neuparth, T., 2016. Statins: An undesirable class of aquatic contaminants? Aquat Toxicol 174, 1–9. https://doi.org/10.1016/j.aquatox.2016.02.001

Seppey, M., Manni, M., Zdobnov, E.M., 2019. BUSCO: Assessing Genome Assembly and Annotation Completeness, in: Kollmar, M. (Ed.), Gene Prediction: Methods and Protocols, Methods in Molecular Biology. Springer, New York, NY, pp. 227–245. https://doi.org/10.1007/978-1-4939-9173-0_14

Sheikholeslami, M.N., Gómez-Canela, C., Barron, L.P., Barata, C., Vosough, M., Tauler, R., 2020. Untargeted metabolomics changes on Gammarus pulex induced by propranolol, triclosan, and nimesulide pharmaceutical drugs. Chemosphere 260, 127479. https://doi.org/10.1016/j.chemosphere.2020.127479

Simão, 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, 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Sinha Hikim, A.P., Swerdloff, R.S., 1999. Hormonal and genetic control of germ cell apoptosis in the testis. Rev Reprod 4, 38–47. https://doi.org/10.1530/ror.0.0040038

Sroda, S., Cossu-Leguille, C., 2011. Seasonal variability of antioxidant biomarkers and energy reserves in the freshwater gammarid Gammarus roeseli. Chemosphere 83, 538–544. https://doi.org/10.1016/j.chemosphere.2010.12.023

Sutcliffe, D.W., 1993. Reproduction in Gammarus (Crustacea, Amphipoda): female strategies, in: Freshwater Forum. pp. 26–64.

Tessier, A.J., Henry, L.L., Goulden, C.E., Durand, M.W., 1983. Starvation in Daphnia: Energy reserves and reproductive allocation1. Limnology and Oceanography 28, 667–676. https://doi.org/10.4319/lo.1983.28.4.0667

The UniProt Consortium, 2019. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Research 47, D506–D515. https://doi.org/10.1093/nar/gky1049

The UniProt Consortium, 2009. The Universal Protein Resource (UniProt) 2009. Nucleic Acids Research 37, D169–D174. https://doi.org/10.1093/nar/gkn664

Trapp, J., Almunia, C., Gaillard, J.-C., Pible, O., Chaumot, A., Geffard, O., Armengaud, J., 2016a. Proteogenomic insights into the core-proteome of female reproductive tissues from crustacean amphipods. Journal of proteomics 135, 51–61.

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

Trapp, J., Armengaud, J., Pible, O., Gaillard, J.-C., Abbaci, K., Habtoul, Y., Chaumot, A., Geffard, O., 2015. Proteomic investigation of male Gammarus fossarum, a freshwater crustacean, in response to endocrine disruptors. Journal of Proteome Research 14, 292–303. https://doi.org/10.1021/pr5009842

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. Molecular & Cellular Proteomics mcp.M114.038851. https://doi.org/10.1074/mcp.M114.038851

Van Straalen, N.M., 2003. Peer reviewed: ecotoxicology becomes stress ecology. Environmental science & technology 37, 324A-330A.

Vellozo, A.F., Véron, A.S., Baa-Puyoulet, P., Huerta-Cepas, J., Cottret, L., Febvay, G., Calevro, F., Rahbé, Y., Douglas, A.E., Gabaldón, T., Sagot, M.-F., Charles, H., Colella, S., 2011. CycADS: an annotation database system to ease the development and update of BioCyc databases. Database (Oxford) 2011, baro08. https://doi.org/10.1093/database/baro08

Zeng, Y., Ren, K., Zhu, X., Zheng, Z., Yi, G., 2018. Chapter One - Long Noncoding RNAs: Advances in Lipid Metabolism, in: Makowski, G.S. (Ed.), Advances in Clinical Chemistry, Advances in Clinical Chemistry. Elsevier, pp. 1–36. https://doi.org/10.1016/bs.acc.2018.07.001

7. ANNEXES

7.1 Supplementary data

Supplementary Data 1

Les données supplémentaires associées à cet article peuvent être trouvées sur figshare avec le lien suivant : https://figshare.com/s/280186a4783a0acc8db5. Les données sont enregistrées sous forme de deux dossiers contenant la Table S1-GFBM et la Table S2-GFBM.

Supplementary Data 2

Les données supplémentaires associées à cet article peuvent être trouvées sur figshare avec le lien suivant : https://figshare.com/s/3244d922163fec420ce5. Les données sont enregistrées sous forme de deux dossiers contenant la Table S1-GFBF et la Table S2-GFBF.

Supplementary Data 3

Les données supplémentaires associées à cet article peuvent être trouvées sur figshare avec le lien suivant : https://figshare.com/s/9af8b0e9a5e0848533e6. Les données sont enregistrées sous forme de deux fichiers contenant les résultats de l'analye différentielle pour les organes chez GFBM et GFBF.

7.2 Supplementary figures

Tableau S1 : Tableau résumé des statistiques des bases de données GamfoCyc (GAMFO_TGFBF et GAMFO_TGFBM).

PGDB	Polypeptides	Enzymatic Reactions	Pathways	Unique pathways
GAMFO_TGFBF	77967	2764	385	9
GAMFO_TGFBM	94499	2850	394	17

Tableau S2 : Comparaison des voies par classe de voies.

Pathway Class	*G. fossarum TGFBF*	*G. fossarum TGFBM*
Biosynthesis	234	238
Amine and Polyamine Biosynthesis	7	7
Amino Acid Biosynthesis	27	26
Aminoacyl-tRNA Charging	3	3
Aromatic Compound Biosynthesis	3	3
Carbohydrate Biosynthesis	29	32
Cell Structure Biosynthesis	0	1
Cofactor, Carrier, and Vitamin Biosynthesis	38	41
Fatty Acid and Lipid Biosynthesis	58	58
Metabolic Regulator Biosynthesis	5	5
Nucleoside and Nucleotide Biosynthesis	23	23
Other Biosynthesis	4	2
Polyprenyl Biosynthesis	5	5
Secondary Metabolite Biosynthesis	16	15
Storage Compound Biosynthesis	0	0
Tetrapyrrole Biosynthesis	3	3
Generation of Precursor Metabolites and Energy	19	19
Metabolic Clusters	14	14
Bioluminescence	0	0
Detoxification	9	9
Transport	0	0
Macromolecule Modification	23	24
Activation/Inactivation/Interconversion	5	5
Degradation/Utilization/Assimilation	116	122
Alcohol Degradation	4	5
Aldehyde Degradation	1	1
Amine and Polyamine Degradation	8	9
Amino Acid Degradation	28	29
Aromatic Compound Degradation	3	3
C1 Compound Utilization and Assimilation	3	2
Carbohydrate Degradation	13	13
Carboxylate Degradation	8	9
Chlorinated Compound Degradation	1	1
Cofactor, Prosthetic Group, Electron Carrier Degradation	1	1
Degradation/Utilization/Assimilation - Other	1	1
Fatty Acid and Lipid Degradation	15	16
Hormone Degradation	2	2
Inorganic Nutrient Metabolism	9	10
Nucleoside and Nucleotide Degradation	13	14
Polymeric Compound Degradation	4	4
Protein Degradation	0	0
Secondary Metabolite Degradation	4	5
Glycan Pathways	17	19
Signal transduction pathways	0	0
Total	411	423

Tableau S3 : Table des voies métaboliques uniques par organisme.

Unique Pathways in Organism	*G. fossarum TGFBF*	*G. fossarum TGFBM*
Unique Pathways	D-gluconate degradation	2-aminoethylphosphonate biosynthesis
	ethene biosynthesis III (microbes)	2-methylcitrate cycle I
	formaldehyde oxidation VII (THF pathway)	ABH and Lewis epitopes biosynthesis from type 2 precursor disaccharide
	L-citrulline biosynthesis	cellulose biosynthesis
	L-citrulline degradation	coenzyme A biosynthesis I (prokaryotic)
	L-dopa and L-dopachrome biosynthesis	coenzyme A biosynthesis II (eukaryotic)
	NAD salvage pathway II (PNC IV cycle)	fatty acid β-oxidation II (plant peroxisome)
	pheomelanin biosynthesis	glycerol degradation V
	UMP biosynthesis III	L-histidine degradation III
		L-tryptophan degradation X (mammalian, via tryptamine)
		oxalate degradation VI
		pyridoxal 5'-phosphate salvage I
		pyrimidine ribonucleosides degradation
		spermine and spermidine degradation II
		the visual cycle I (vertebrates)
		UDP-α-D-glucuronate biosynthesis (from *myo*-inositol)
		UMP biosynthesis I

Tabealu S4 : Liste voies communes du ML chez GFBM et GFBF.

76 common pathways in GFBM and GFBF
(4Z,7Z,10Z,13Z,16Z)-docosapentaenoate biosynthesis (6-desaturase)
2-methyl-branched fatty acid β-oxidation
3-phosphoinositide biosynthesis
3-phosphoinositide degradation
acyl-CoA hydrolysis
acyl-[acyl-carrier protein] thioesterase pathway
anandamide biosynthesis I
anandamide biosynthesis II
anandamide degradation
anandamide lipoxygenation
arachidonate biosynthesis III (6-desaturase, mammals)
arachidonate biosynthesis IV (8-detaturase, lower eukaryotes)
biotin-carboxyl carrier protein assembly
cardiolipin biosynthesis II
CDP-diacylglycerol biosynthesis I
ceramide *de novo* biosynthesis
ceramide degradation (generic)
cis-vaccenate biosynthesis
D-myo-inositol-5-phosphate metabolism
diacylglycerol and triacylglycerol biosynthesis
docosahexaenoate biosynthesis III (6-desaturase, mammals)
dolichol and dolichyl phosphate biosynthesis
fatty acid β-oxidation I (generic)
fatty acid β-oxidation III (unsaturated, odd number)
fatty acid β-oxidation V (unsaturated, odd number, di-isomerase-dependent)
fatty acid biosynthesis initiation (mitochondria)
fatty acid biosynthesis initiation (type I)
fatty acid elongation -- saturated
gala-series glycosphingolipids biosynthesis
γ-linolenate biosynthesis II (animals)
gondoate biosynthesis (anaerobic)
icosapentaenoate biosynthesis II (6-desaturase, mammals)
icosapentaenoate biosynthesis VI (fungi)
juniperonate biosynthesis
lacto-series glycosphingolipids biosynthesis

long-chain fatty acid activation
mitochondrial L-carnitine shuttle
monoacylglycerol metabolism (yeast)
neolacto-series glycosphingolipids biosynthesis
octanoyl-[acyl-carrier protein] biosynthesis (mitochondria, yeast)
oleate biosynthesis II (animals and fungi)
palmitate biosynthesis (type I fatty acid synthase)
palmitate biosynthesis (type II fatty acid synthase)
palmitoleate biosynthesis
palmitoleate biosynthesis I (from (5Z)-dodec-5-enoate)
palmitoleate biosynthesis II (plants and bacteria)
phosphatidate biosynthesis (yeast)
phosphatidylcholine acyl editing
phosphatidylcholine biosynthesis I
phosphatidylcholine biosynthesis II
phosphatidylcholine resynthesis via glycerophosphocholine
phosphatidylethanolamine biosynthesis II
phosphatidylethanolamine biosynthesis III
phosphatidylglycerol biosynthesis II (non-plastidic)
phosphatidylinositol biosynthesis II (eukaryotes)
phospholipases
phospholipid remodeling (phosphatidate, yeast)
phospholipid remodeling (phosphatidylcholine, yeast)
phospholipid remodeling (phosphatidylethanolamine, yeast)
plasmalogen biosynthesis
plasmalogen degradation
sciadonate biosynthesis
sphingolipid biosynthesis (mammals)
sphingomyelin metabolism
sphingosine and sphingosine-1-phosphate metabolism
stearate biosynthesis I (animals)
stearate biosynthesis II (bacteria and plants)
stearate biosynthesis III (fungi)
sterol:steryl ester interconversion (yeast)
tetradecanoate biosynthesis (mitochondria)
triacylglycerol degradation
ultra-long-chain fatty acid biosynthesis
valproate β-oxidation

very long chain fatty acid biosynthesis I
very long chain fatty acid biosynthesis II
zymosterol biosynthesis

Only in GFBM
fatty acid β-oxidation II (plant peroxisome)
superpathway of geranylgeranyldiphosphate biosynthesis I (via mevalonate)

Figure S1 : Complétude des voies globales de *Drosophila melanogaster* par MetExplore.

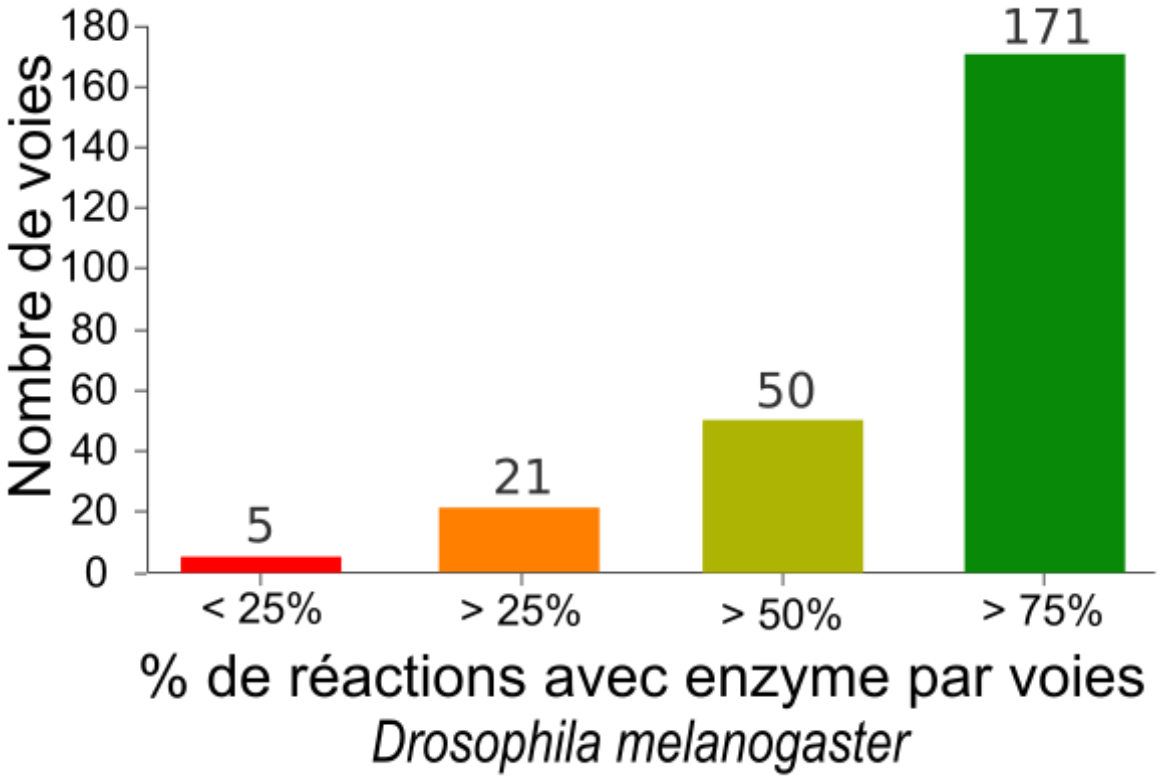


Figure S2 : Effectif en nombre de protéines dans les organes et les échantillons de GFBM et de GFBF.

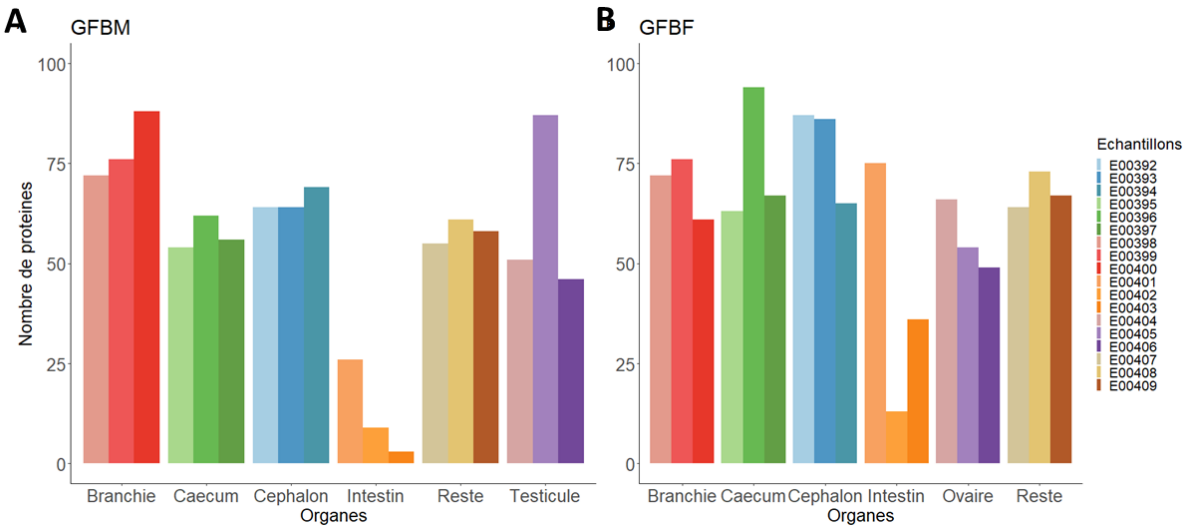


Figure S3 : MA plot de l'analyse différentielle des branchies versus le reste des organes. Selon la définition des contrastes de l'analyse différentielle, les protéines indiquées comme « down » en bleue, correspondent aux protéines plus abondantes dans la branchie pour le mâle (A) et pour la femelle (B), et inversement.

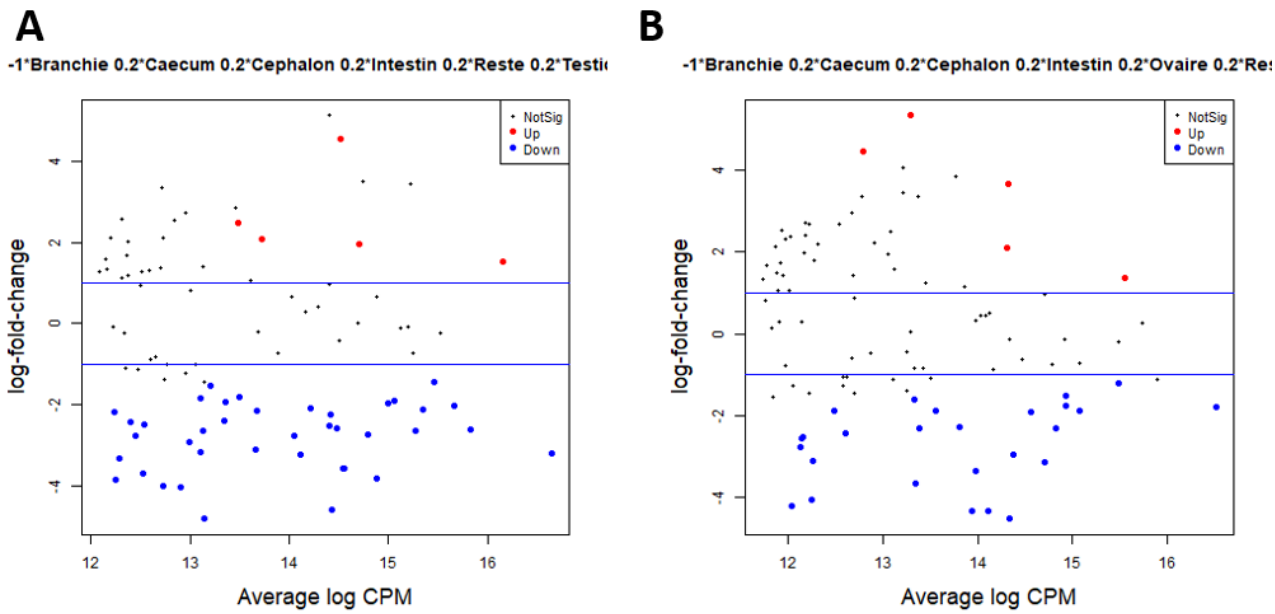
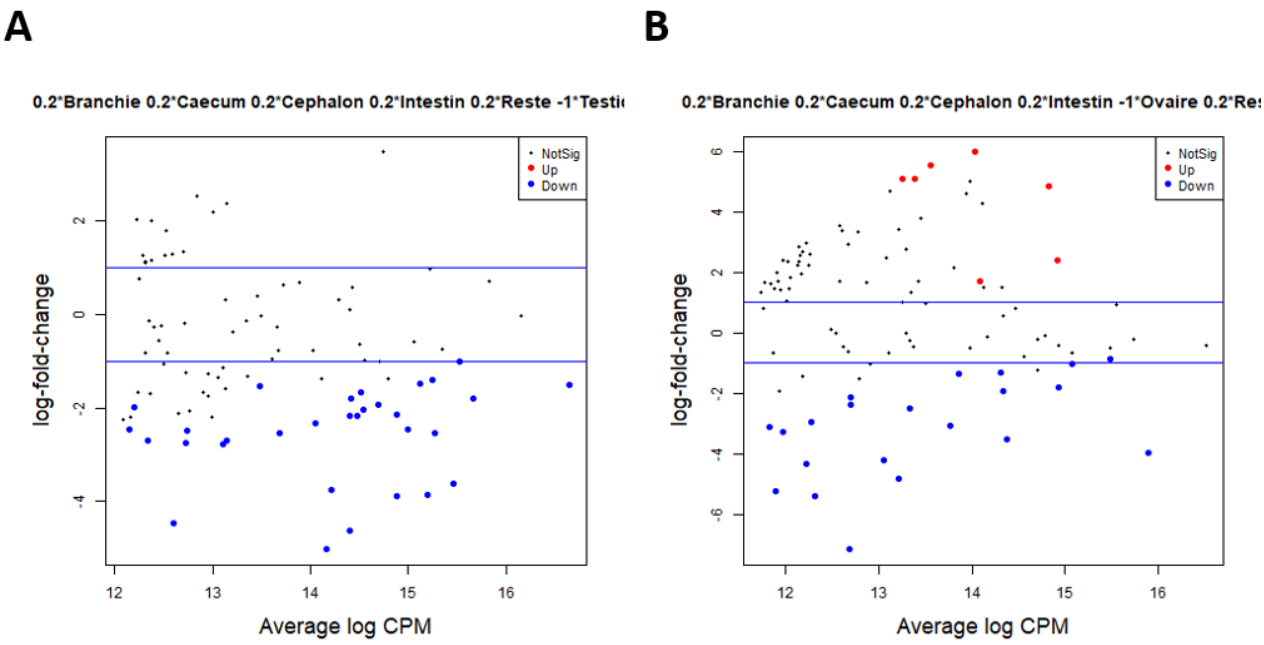


Figure S4 : MA plot de l'analyse différentielle des Gonades vs le reste des organes. Selon la définition des contrastes de l'analyse différentielle, les protéines indiquées comme « down » en bleue, correspondent aux protéines plus abondantes dans les gonades pour le mâle (A) et pour la femelle (B), et inversement.



Chapitre IV – Discussion et perspectives

En écotoxicologie, de nombreux biomarqueurs reposent sur la mesure de traits d'histoires de vie et la mesure d'activité enzymatique dont la spécificité chez l'espèce d'intérêt n'est pas toujours démontrée et validée. Pour mieux comprendre les modes d'action de contaminants, le niveau moléculaire doit être mieux pris en compte, toutefois l'absence de génome chez les espèces non modèles d'intérêt environnemental constitue une limite forte. *Gammarus fossarum* a été identifié comme une espèce sentinelle non modèle pour évaluer les risques écotoxicologiques des contaminants uniques ou en mélange, au laboratoire ou sur le terrain (*in situ*) (Adam et al., 2010 ; Besse et al., 2013 ; Gouveia et al., 2017 ; Kunz et al., 2010 ; Trapp et al., 2015 ; Wigh et al., 2017). Les avancées des technologies NGS et de spectrométrie de masse (MS) permettent aujourd'hui d'étendre l'acquisition de données omiques (*e.g.* par transcriptomique, protéomique ou métabolomique) aux organismes non modèles de pertinence environnementale, comme *G. fossarum*.

Le travail de cette thèse s'est focalisé sur le développement d'une approche bioinformatique multiomique pour exploiter les données omiques déjà disponibles (*i.e.* transcriptomiques et protéomiques) chez *Gammarus fossarum*, ceci afin de contourner l'absence d'un génome de référence pour l'annotation fonctionnelle et la caractérisation du métabolisme lipidique. Dans un premier temps, nous aborderons les principaux verrous rencontrés : (i) la normalisation des données transcriptomiques et protéomiques pour l'analyse de données et (ii) l'adaptation de l'outil CycADS à nos données transcriptomiques, initialement conçu pour des données génomiques. Dans ce cadre, nous nous sommes intéressés à la description et la caractérisation sans *a priori* des voies moléculaires du métabolisme lipidique chez *G. fossarum*. Dans un second temps, nous discuterons les principaux résultats biologiques et leur intérêt en

écotoxicologie. Et enfin, nous proposerons des éléments de réflexions à propos des études multiomiques en écotoxicologie pour les espèces non modèles de pertinence environnementale.

1. ADAPTATION DES DONNEES OMIQUES D'ORGANISMES NON MODELES POUR LES OUTILS BIOINFORMATIQUES

Une multitude de méthodes et algorithmes ont été initialement développés pour l'analyse des données omiques. Cependant, l'explosion des nouvelles méthodes d'acquisition de données à haut débit, telles que la spectrométrie de masse à haute résolution ou le séquençage short reads a conduit à une grande diversité de jeux de données, et par conséquent le besoin en stratégies bioinformatiques adaptées. Pour ces travaux, nous avions à disposition des données de protéomique shotgun (Trapp et al., 2015) et transcriptomiques chez *G. fossarum* (Cogne et al., 2019c). Dans cette partie, nous discuterons des différents verrous méthodologiques que nous avons dû lever afin d'utiliser des méthodes disponibles dans la littérature sur ces jeux de données.

1.1 Apport bioinformatique de l'analyse de réseaux de coexpression pour exploiter les données issues de protéogénomique

L'analyse de réseaux de coexpression par l'outil WGCNA (Langfelder and Horvath, 2008), est une méthode conçue initialement pour les données de puces à ADN et qui a été ensuite étendue aux données RNA-Seq, or nous avions à disposition des données de protéomique shotgun. L'imputation (*i.e.* remplacement) des données manquantes et la normalisation des données sont essentielles (Karpievitch et al., 2012; Rahmatbakhsh et al., 2021), en effet, contrairement aux données transcriptomiques, les données protéomiques sont souvent incomplètes dues aux limites de sensibilité, de détection des peptides ou d'identification de séquences codantes. Ces données manquantes sont un obstacle ou un vecteur de biais pour l'analyse de réseaux de coexpression, car elles peuvent être associées à une abondance faible ou nulle pour des protéines non détectées. Pour pallier cette limite, des algorithmes, initialement conçus pour des données de microarray, ont

été proposés pour être appliqués aux données protéomiques, telles que les k plus proches voisins (KNN : K Nearest Neighbor) (Ling and Dong-Mei, 2009), la méthode des moindres carrés ou encore la méthode des moindres carrés locaux (Nie et al., 2007; Pei et al., 2017; Xia et al., 2015). Ces méthodes permettent d'imputer statistiquement les valeurs manquantes dans les échantillons, en se basant sur les valeurs expérimentales existantes des réplicats ou en les comparant au niveau d'abondance relative de protéines apparentées. Ces méthodes statistiques et les différents mécanismes associés ont été résumé dans Lazar et al. (2016) et Salgado et al. (2016). Néanmoins, il n'existe pas de consensus dans le choix de la stratégie d'imputation à appliquer (Pei et al., 2017) et un trop grand nombre de données manquantes peut toujours poser un problème pour la précision de l'imputation.

Nous avions 84% de valeurs manquantes dans les données de l'article n°1 (voir page 118), et 70% dans les données de l'article n°2 (voir page 130). Nous avons donc réalisé plusieurs tests pour tenter d'imputer le reste des valeurs manquantes encore présentes dans nos données, notamment avec le package R DAPAR (Wieczorek et al., 2017) conçu pour l'analyse statistique des données protéomiques. Nous avons utilisé la méthode KNN pour imputer les valeurs manquantes de notre jeu de données de l'article n°1 (voir chapitre II-1). Après avoir appliqué cet algorithme sur nos données, aucune relation de coexpression entre les protéines n'a pu être détectée sur le nouveau jeu de données (*i.e.* avec imputation des valeurs manquantes), ce qui était biologiquement très improbable. En effet, les méthodes sont performantes lorsqu'il y a moins de valeurs manquantes (<10 %) et deviennent imprécises lorsque la proportion de valeurs manquantes augmente (>20 %) (Gao et al., 2015 ; Lazar et al., 2016 ; Webb-Robertson et al., 2015), il a donc été préférable de supprimer les peptides présentant une grande fraction d'entrées manquantes (Pei et al., 2017). Dans notre étude, pour éviter le bruit de fond associé aux protéines peu abondantes et présentant

des valeurs manquantes, nous avons choisi d'exclure de notre analyse les protéines ayant moins de trois spectral count (SC) dans les échantillons (Gregori et al., 2013). Dans la littérature, on retrouve que la proportion de protéines qui présentent au moins une valeur manquante est très importante dans les jeux de données protéomiques, pouvant varier entre 70% et 90% de protéines (Albrecht et al., 2010). Pour le jeu de données de l'article n°1 (voir page 118), on passait donc de 1199 protéines au départ à 375 protéines, donc une perte de 69% des protéines initiales. Pour le jeu de données de l'article n°2 (voir page 130), on passait de 871 protéines au départ à 312 protéines, donc une perte de 64% des protéines initiales.

Une normalisation adéquate des données est nécessaire lorsque l'on cherche à déterminer des corrélations ou des différences entre entités dans nos jeux de données omiques. La normalisation vise à éliminer les biais systématiques pour permettre les inférences statistiques (Karpievitch et al., 2012). En effet, les échantillons peuvent différer en termes de taille d'échantillons, de poids ou encore de quantité de protéines extraites dans chaque échantillon, et ces paramètres doivent être pris en compte pour pouvoir comparer deux échantillons sans altérer ou biaiser cette détection (Chen et al., 2020). Il est donc important de normaliser les données d'abondance des protéines des échantillons pour construire les corrélations de coexpression dans nos études (Langfelder and Horvath, 2008; Pei et al., 2017). Pour la construction de nos réseaux de coexpression, la normalisation des spectral count a été effectuée à l'aide de la fonction *calcNormFactors* du package R edgeR (Robinson and Oshlack, 2010). Derrière cette fonction est implémentée l'approche TMM (*i.e.* weighted trimmed mean of M-values ou log-intensity ratios). La normalisation TMM se base sur l'hypothèse que la plupart des caractéristiques (*e.g.* protéines) ne sont pas différentiellement exprimées (Robinson and Oshlack, 2010). Selon cette hypothèse, cette approche sélectionne un échantillon comme référence et calcule un facteur de normalisation TMM pour les autres

échantillons (Robinson and Oshlack, 2010). Cette méthode de normalisation prend en compte le fait que les protéines très abondantes ont tendance à se voir attribuer davantage de peptides (spectres) que les moins abondantes (Gregori et al., 2013; Robinson and Oshlack, 2010). Les spectral count doivent ensuite être transformés en log pour normaliser les distributions. Cette transformation de données est directement implémentée dans le package edgeR et n'est donc pas nécessaire avant la normalisation TMM par EdgeR (Rahmatbakhsh et al., 2021). Cette méthode de normalisation comporte néanmoins certaines limites, telles que l'absence de prise en compte de la taille des protéines et l'hypothèse que la longueur de la protéine sera constante entre les échantillons. Selon Dillies et al (2013) et Branson et Freitas (2016), la méthode TMM implémentée dans le package edgeR est une des méthodes les plus performantes pour éliminer les biais dans les données protéomiques en comparaison avec les méthodes de normalisation par : comptage médian (*i.e.* median count), le quantile supérieur, la moyenne du comptage total ou encore la normalisation par quantile. Sur la base de toutes les connaissances de la littérature, nous avons donc utilisé l'approche de normalisation TMM. L'approche optimale de normalisation dépend de l'environnement expérimental des différentes études protéomiques et doit être évaluée individuellement avec des outils de visualisation tels que les MA plot, le clustering hiérarchique ou encore les heatmaps qui aident à déterminer l'efficacité des étapes de prétraitement des données (Chen et al., 2020).

1.2 Apport bioinformatique de l'analyse multiomique et reconstruction des réseaux métaboliques

Dans un second temps, nous avons exploité l'outil de reconstruction et d'annotation fonctionnelle CycADS (Vellozo et al., 2011) dans une démarche multiomique pour exploiter les données transcriptomiques et de protéomique shotgun pour reconstruire le métabolisme lipidique chez *G. fossarum*. Dans le contexte de cette thèse, nous avons utilisé la transcriptomique pour

obtenir les informations sur le contenu des gènes codants pour les protéines. Nous avons donc travaillé à partir des transcriptomes mâles et femelles de *Gammarus fossarum B* provenant de l'étude de Cogne et al. (2019). Ces transcriptomes étaient déjà assemblés *de novo* à l'aide de l'assembleur Trinity (Grabherr et al., 2011).

L'outil CycADS est en premier lieu destiné aux données génomiques, qui nécessitent en entrée un génome et un fichier d'annotation structurelle (*i.e.* GFF3) correspondant. Dans l'optique d'exploiter les données omiques de plus en plus disponibles pour un grand nombre d'espèces dont le génome n'est pas encore séquencé et annoté, comme le gammare, nous avons choisi d'adapter nos données transcriptomiques à cet outil et de montrer la faisabilité de cette démarche.

Les assembleurs de transcriptomes *de novo* produisent généralement beaucoup plus de transcrits que ce à quoi on pourrait s'attendre sur la base du nombre de gènes dans le génome (Raghavan et al., 2022). La performance et la précision de l'assemblage *de novo* du transcriptome dépendent largement de la complexité du génome (*e.g.* la taille du génome, le nombre de paralogues, le niveau de ploïdie), de la couverture de lecture différentielle des données séquencées et des erreurs de séquençage. Par exemple, Bryant et al. (2017) rapportent avoir assemblé plus de 1,5 million de séquences pour un transcriptome de l'amphibien axolotl (*Ambystoma mexicanum*), alors que son génome compte environ 23 000 gènes (Nowoshilow et al., 2018). S'ajoute à cela, le fait que la transcription est un processus biologique générant du bruit de fond. Par exemple, Dunham et al. (2012) affirment que plus de 80 % du génome d'*Homo sapiens* est transcrit même si moins de 3% (Hangauer et al., 2013) des produits transcrits codent pour des protéines. L'assemblage du transcriptome est complexe et différent de l'assemblage du génome dans lequel la couverture de lecture est plus uniforme. Ainsi, les contigs assemblés *de novo* comprennent des artefacts de transcription, des pré-ARNm et des ARNnc en plus des transcrits codant pour des

protéines (Freedman et al., 2021). Une autre source de séquences supplémentaires est l'épissage alternatif (McManus and Graveley, 2011; Zhang et al., 2021; Zhao et al., 2011) qui se manifeste par des isoformes de transcription. Ces séquences peuvent être des artéfacts menant à des erreurs d'annotations qui peuvent se répandre dans les bases de données et dans les analyses fonctionnelles en aval. La réduction de la complexité de l'assemblage peut donc être une étape importante pour obtenir un ensemble de séquences biologiquement pertinentes (Raghavan et al., 2022). La réduction de nos jeux de données permet également de réduire considérablement le temps de calcul du pipeline de CycADS que nous utilisons. En effet, si on prend l'exemple des espèces *Daphnia pulex* ou *Drosophila melanogaster*, toutes deux présentent dans la base de données ArthropodaCyc, leur proportion de séquences codantes pour des protéines est 5 à 6 fois inférieure à nos transcriptomes bruts (Baa-Puyoulet et al., 2016 ; Ye et al., 2017). Il est important de noter que la réduction de la redondance de l'assemblage ne doit être effectuée que si cela est nécessaire. La provenance des contigs assemblés *de novo* est inconnue, et ils peuvent donc tous porter des informations biologiques pertinentes. Le filtrage des isoformes est une tâche intrinsèquement heuristique. Il est tout à fait possible, par exemple, de régler les paramètres de telle sorte que des paralogues étroitement liés soient regroupés. Dans un tel cas, des séquences qui devraient être représentées dans l'assemblage seront perdues. De plus, dans le cas des isoformes, il est souvent impossible d'identifier l'isoforme la plus biologiquement pertinente (Haas et al., 2013). Par exemple, l'isoforme la plus longue n'est pas nécessairement la plus exprimée (et vice versa). Il est possible que l'isoforme la plus longue ou la plus exprimée ne soit pas celle qui est réellement représentative du gène et de la protéine concomitante. L'isoforme la plus longue peut être le résultat d'une erreur de l'assembleur dans l'extension du contig biologiquement pertinent, ou le résultat de la rétention d'un intron dans le transcrit. Par la suite, la protéine correspondante peut ne pas être la plus longue de la cohorte, ou peut même être absente en raison de l'aberration

de l'ORF correspondant. Il faut donc faire preuve d'une extrême prudence lors de la réduction de la redondance des assemblages et de la réduction des redondances, car un nettoyage non adapté peut entraîner la perte de séquences autrement informatives pour les analyses en aval (Raghavan et al., 2022).

Dans le but de réduire la redondance en choisissant l'isoforme représentative, il existe différentes méthodes ou critères : comme l'isoforme la plus longue, l'isoforme ayant la plus longue séquence codante (CDS) (Wang et al., 2020), l'isoforme ayant la plus longue séquence d'acides aminés traduite (ORFs) ou l'isoforme ayant le plus grand nombre de lectures (abondance) (Gonzàlez-Porta et al., 2013). Si l'on se base sur l'hypothèse que les transcrits ayant une faible abondance, sont ceux qui plus de chance d'avoir un impact phénotypique moindre, les transcrits avec un faible niveau d'expression peuvent être éliminés de l'assemblage (Raghavan et al., 2022). Ainsi, nous avons opté pour cette hypothèse de travail qui nous permet de conserver le plus d'information biologique, en conservant uniquement l'isoforme la plus abondante parmi les transcrits d'un même cluster d'isoformes (*i.e.* dérivant potentiellement d'un locus) (Haas et al., 2013). Nous avons retenu les isoformes avec la mesure d'abondance du pourcentage d'isoformes (IsoPct) le plus élevé pour chaque locus, avec l'option *highest_iso_only* (Yang and Smith, 2013). Cette option nous a donc permis de garder au moins un transcrit pour chaque locus, en conservant environ 77% et 67% des transcrits originaux, pour les transcriptomes mâle et femelle respectivement. Ces transcrits correspondent à 69% et 67% des ORFs originaux du mâle et de la femelle, respectivement.

Il existe d'autres mesures d'abondance disponibles et calculées par Trinity, comme les FPKM ou les TPM qui consistent à exclure les transcrits qui peuvent être considérés comme faiblement exprimés sur la base de mesures d'abondance. Les contigs dont le support de lecture est inférieur

à un seuil (*e.g.* TPM < 1) peuvent être écartés de l'assemblage (Raghavan et al., 2022). En faisant varier les seuils de filtrage de nos isoformes, nous avons remarqué qu'en nous basant sur TPM < 1 nous perdions une très grande quantité de transcrits (plus de 80% des transcrits de départ chez le mâle et la femelle) et des locus entiers ou les transcrits étaient trop faiblement exprimés. Il s'avère que l'option de filtrage *highest_iso_only* était équivalente au seuil de TPM < 0.1 dans nos données (après avoir réalisé des tests de seuils de TPM), en termes de nombre de transcrits restant dans nos transcriptomes mâle et femelle (environ 64% des transcrits de départ). N'ayant pas de consensus établi à ce jour, toutes ces approches peuvent être aussi efficaces les unes que les autres, et sont susceptibles de dépendre de l'ensemble des données (Raghavan et al., 2022).

Pour évaluer qualitativement et quantitativement la réduction de la redondance des isoformes et l'impact sur la structure de nos assemblages, nous avons utilisé la mesure de complétude BUSCO (« Benchmarking Universal Single-Copy Orthologues ») (Simão et al., 2015) sur nos transcriptomes. Les résultats BUSCO ont montré que plus de 83% des 1013 gènes orthologues sont retrouvés complets en copie unique ou dupliqués dans les deux transcriptomes filtrés en ne gardant que l'isoforme la plus exprimée par locus (*highest_iso_GFBF* et *highest_iso_GFBM*) (Figure 36). Environ 40% des gènes complets sont retrouvés en copie unique dans les transcriptomes originaux, contre environ 60% dans les transcriptomes filtrés (Figure 36). On constate donc que le filtrage de nos transcriptomes a permis l'augmentation de la proportion de gènes en copie unique. Il n'existe pas de score BUSCO absolu pour définir ce qui constitue un « bon » score BUSCO, mais ce score peut être comparé de façon relative à d'autres transcriptomes assemblés si la base de données des gènes orthologues est identique. En effet, Caputo et al. (2020) observent un score de 96% de gènes complets, mais à partir d'une base de données de métazoaires comprenant 978 gènes BUSCO (metazoa_odb9) (Simão et al., 2015), tandis que la base de données de référence utilisée dans notre

étude est celle des arthropodes (arthopoda_odb10) avec 1013 gènes BUSCO (Kriventseva et al., 2019). Tandis que dans l'étude de Lipaeva et al. (2021), la redondance des isoformes des transcriptomes de trois espèces d'amphipodes du lac Baïkal a été réduite. Les scores BUSCO ont été calculés sur la même version de base de données de gènes orthologues d'arthropodes que pour notre étude et les auteurs ont observé un effet positif du filtrage des isoformes sur la proportion de gènes complets présents dans les transcriptomes, avec une augmentation du nombre de gènes en copie unique (Lipaeva et al., 2021). Les scores BUSCO pour les gènes complets en copie unique ou dupliqués allaient de 72% à 82,9% (Lipaeva et al., 2021).

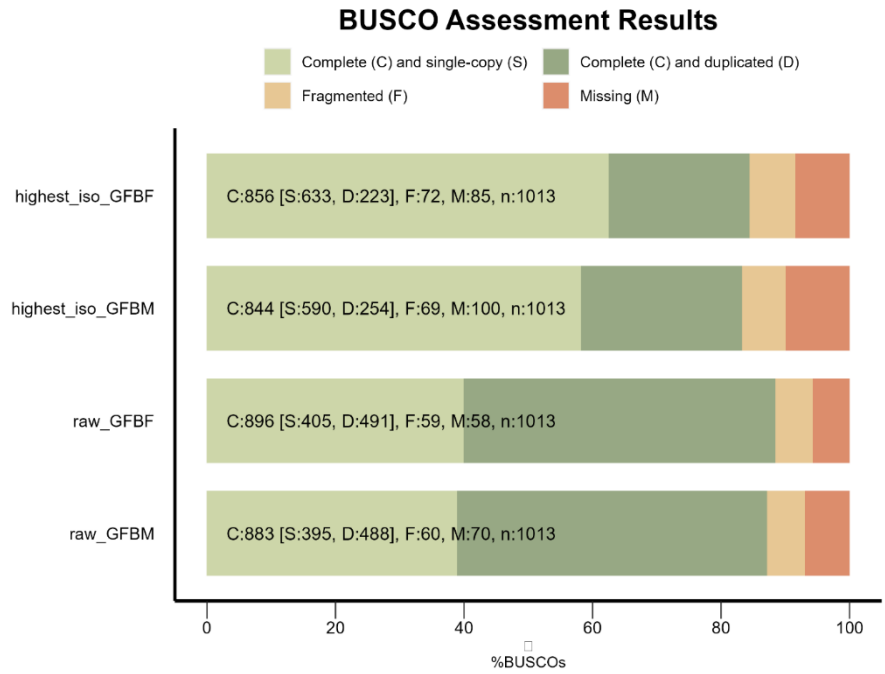


Figure 44 : Analyse de la complétude des transcriptomes originaux (raw) et filtrés (highest_iso) du mâle et de la femelle de G. fossarum via BUSCO des arthropodes.

Cependant, la part de gènes dupliqués restant dans la plupart des études pourrait venir de différents haplotypes, ou isoformes encore présentes et donc d'un résidu de redondance des assemblages que la méthode choisie n'aurait pas permis d'éliminer. Cela est cohérent avec la méthode de filtrage, qui a été choisie volontairement pour perdre le moins d'informations biologiques (*i.e.* locus) en restant le plus exhaustif possible pour maximiser les chances

d'identification des enzymes d'intérêt au cours de l'annotation. On peut aussi noter l'absence du sous-embranchement des crustacés dans la base de données de gènes orthologues, mais uniquement la présence de la classe des insectes et celle des arachnides. La complétude des transcriptomes est corrélée positivement à la proportion de gènes BUSCO complets, mais cette évaluation peut être biaisée par la sous-représentation d'espèces proches de l'espèce d'intérêt (Amil-Ruiz et al., 2021 ; Seppey et al., 2019).

2. APPORTS BIOLOGIQUES DES APPROCHES SANS *A PRIORI* CHEZ *GAMMARUS FOSSARUM*

Les études permettant d'acquérir des données omiques et de découvrir de nombreux gènes, protéines et métabolites ne cessent de croître chez les espèces non modèles, mais celles-ci ne sont pas exploitées dans leur globalité. Les approches actuelles d'annotation fonctionnelle sont basées sur la similarité des séquences ou des corrélations fonctionnelles à partir de la physiologie chez les espèces qui possèdent davantage de données physiologiques ou biologiques dans les bases de données (Duarte Gouveia et al., 2017; Gouveia et al., 2019b; Trapp et al., 2014b). Ainsi, il est nécessaire de développer des stratégies bioinformatiques pour exploiter sans *a priori* et l'ensemble ou la majeure partie des données omiques existantes chez les espèces d'intérêt environnemental pour lequel on manque de génomes de référence. Un des verrous principaux est l'attribution des fonctions pour les protéines identifiées (Trapp et al., 2014a). Il est donc essentiel de déverrouiller l'annotation des espèces non modèles grâce à des stratégies et outils qui permettent de révéler la fonction des gènes et protéines propres à l'espèce étudiée. L'objectif de cette partie du travail est de montrer d'une part que le domaine de la biologie des systèmes permet de répondre à ce besoin grâce à des approches sans *a priori* pour les données omiques et d'autre part, nous avons cherché à montrer qu'une approche multiomique permet de combiner des données de protéomique shotgun et de transcriptomique d'une espèce non modèle pour l'annotation des voies du métabolisme

lipidique. Dans un premier temps, des données de protéomique shotgun d'organes reproducteurs et d'embryons de gammares ont été utilisées pour réaliser deux analyses de réseaux de coexpression afin de mieux comprendre les acteurs moléculaires de la reproduction et ceux liés à la toxicité testiculaire des contaminants. Dans un second temps, des données transcriptomiques couplées à des données de protéomique shotgun ont été exploitées par l'outil CycADS pour décrire le réseau des voies métaboliques lipidiques chez *Gammarus fossarum* (Vellozo et al., 2011). A notre connaissance, c'est la première étude ayant utilisée cet outil avec les données transcriptomiques couplées à la protéomique shotgun chez une espèce non modèle.

2.1 Les réseaux de coexpression pour l'annotation

La traduction de l'information en une meilleure compréhension biologique par l'analyse conventionnelle de l'expression différentielle reste un défi majeur (Ruffalo et al., 2015; Trapp et al., 2015). Dans le cadre des analyses d'expression différentielle, il est difficile d'identifier les protéines avec une faible abondance ou associées un faible fold change, qui jouent un rôle régulateur important dans les réponses biologiques contre les perturbations environnementales externes (Pei et al., 2014 ; Ruffalo et al., 2015). Pour pallier ces limites, les réseaux de coexpression peuvent fournir un aperçu des processus cellulaires sous-jacents grâce aux modules de gènes coexprimés (Eisen et al., 1998). Dans le chapitre II (voir page 115) nous avons utilisé pour la première fois chez *G. fossarum* une analyse de réseau de coexpression pour étudier l'organisation des protéines, d'une part dans les organes reproducteurs et les embryons, et d'autre part dans les testicules suite à une exposition à des contaminants. Nous avons montré que l'outil WGCNA (« Weighted Gene Coexpression Network Analysis ») était pertinent pour extraire les informations biologiques à partir de données de protéomique shotgun. En effet, il permet de regrouper les protéines dans des modules de coexpression, d'identifier les protéines centrales (dites hub) de ces modules et

d'identifier les relations entre les modules et les traits d'histoires de vie ou d'exposition aux contaminants.

Dans l'article n°1 (voir page 118), qui se place dans le contexte de la reproduction, nous avons cherché à identifier des voies protéiques régulant la maturation des gonades et le développement embryonnaire chez le crustacé d'eau douce *Gammarus fossarum*. Cette méthodologie a permis d'identifier quatre modules de protéines coexprimés en corrélation avec les stades de développement embryonnaire, aux ovaires et aux testicules. Un premier module était corrélé aux stades intermédiaires du développement embryonnaire (S3 et S4), correspondant à l'organogenèse. Ce module était enrichi en protéines impliquées dans l'édition et l'épissage de l'ARN et la synthèse de protéines. Un deuxième module associé aux ovaires était enrichi en protéines de type vitellogénine et en protéines de la coagulation, montrant la diversité des protéines appartenant à la grande famille de transfert lipidique impliquée dans la maturation des ovocytes chez cet amphipode d'eau douce. L'analyse de coexpression réalisée dans cette étude a confirmé et étendu les résultats précédents qui montraient une grande diversité de protéines appartenant à la superfamille LLTP (*i.e.* « Large Lipid Transport Protein » pour protéines de transport des grands lipides), impliquée dans la formation du vitellus (Trapp et al., 2016b). Nous avons trouvé une forte coexpression de 22 contigs annotés comme des protéines de type Vtg (*i.e.*, vitellogénine) ou de type protéine clottable dans les ovaires. Alors que des copies multiples des gènes Vtg ont été rapportées chez de nombreux arthropodes (Provost-Javier et al., 2010 ; Wurm et al., 2011), il est possible que certains des contigs représentent une reconstruction fragmentée de l'ARNm original, donnant une surestimation du nombre total de vitellogénines ou de protéines clottables chez *G. fossarum*. Nos corrélations de coexpression entre les différents contigs de type Vtg pourraient aider les efforts d'annotation et améliorer les assemblages *de novo* de

transcriptome, fournissant probablement une meilleure estimation du nombre de protéines LLTP chez cet amphipode. Enfin, nous avons identifié deux autres modules associés aux testicules, l'un enrichi en protéines de la glycolyse et le second en protéines de type actines-myosines.

Par rapport aux études précédentes de l'équipe (Trapp et al., 2016b, 2014b), nous avons pu identifier des protéines clés impliquées dans des voies physiologiques telles que la maturation des ovocytes, la spermatogenèse et le développement embryonnaire sans faire appel à une fonction protéique *a priori* prédite par une recherche standard de similarité de séquence, mais en utilisant le clustering hiérarchique pour construire des modules à partir des données d'abondance des protéines obtenues par protéomique shotgun.

Dans l'article n°2 (voir page 130) nous avons étudié les voies et les protéines clés liées à la toxicité testiculaire chez *G. fossarum*. Notre jeu de données provenait de l'étude de Trapp et al. (2015). Il consistait en des données de protéomique shotgun de gonades mâles de *G. fossarum*, exposés au cadmium (Cd), au pyriproxyfène (Pyr) et au méthoxyfénozide (Met) dans des conditions de laboratoire. L'analyse de réseau de coexpression a permis d'identifier dix modules de protéines coexprimées, dont quatre modules distincts étaient significativement corrélés à l'exposition aux contaminants. L'analyse d'enrichissement de l'annotation des protéines a identifié des modules impliqués dans l'organisation du cytosquelette et la réponse au stress oxydatif associée à l'exposition au Cd. Ce module était en particulier fortement enrichi en protéines appartenant à la famille des myosines, impliquées dans les processus d'organisation du squelette au cours de la morphogenèse du sperme mature chez les métazoaires (Li and Yang, 2016; Sun et al., 2011). Ce module présente une composition protéique similaire à un module protéique précédemment identifié dans les testicules de gammares dans l'article n°1 (voir page 118), montrant la cohérence de l'analyse des réseaux de coexpression dans l'identification de modules préservés avec une

signification biologique (Langfelder et al., 2011). Il est intéressant de noter que le Cd a déjà été montré comme induisant des anomalies structurelles des spermatozoïdes chez d'autres modèles d'arthropodes, tels que le lépidoptère *Bombyx mori* (Yuan et al., 2016) et le coléoptère *Blaps polycresta* (le carabinier) (Shonouda and Osman, 2018). De plus, le principal mécanisme responsable de la toxicité du Cd dans divers organes, dont les testicules, est l'induction du stress oxydatif (Bhardwaj et al., 2020).

Le module associé à l'exposition au Pyr était associé à la réponse au stress du réticulum endoplasmique (RE). Ce module était enrichi en protéines liées au stress du réticulum endoplasmique, notamment en protéines de choc thermique et en calréticuline, toutes ayant une place centrale (*i.e.* protéines hub) dans le module. Les protéines centrales (dites hub) des modules sont des protéines ayant une plus forte connectivité avec les autres protéines à l'intérieur du module. On fait l'hypothèse qu'elles jouent un rôle clé dans les voies biologiques. Le stress du réticulum endoplasmique a été récemment identifié comme un nouveau mode d'action impliqué dans la toxicité testiculaire *in vivo* chez la souris (Li et al., 2020) et *in vitro* dans les cellules de rongeurs (Ham et al., 2020). Il est intéressant de noter que le Pyr modifie également l'homéostasie du Ca^{2+} *in vitro* et *in vivo* dans les testicules de *Danio rerio* (Staldoni de Oliveira et al., 2021). Les altérations du Ca2+ intracellulaire ont été associées à une augmentation de la peroxydation lipidique et à une diminution de la capacité antioxydante, entraînant une altération de la spermatogenèse (Staldoni de Oliveira et al., 2021). Notre analyse de réseau identifie la calréticuline de *G. fossarum* comme une protéine centrale (hub) dans le module associé à l'exposition au Pyr, suggérant que le Pyr peut induire un stress du réticulum endoplasmique et une réponse de choc thermique par une altération de l'homéostasie du Ca2+ intracellulaire. Notre analyse d'enrichissement du module protéique corrélé à l'exposition au Pyr a montré que les mêmes

protéines impliquées dans la réponse au stress du réticulum endoplasmique appartiennent également à l'annotation GO terme "mort cellulaire". En effet, il est connu que lorsque l'homéostasie du RE ne peut être restaurée, le dysfonctionnement cellulaire induit par le stress du RE peut conduire à la mort cellulaire (Sano and Reed, 2013). Notre analyse fournit la première preuve d'une implication potentielle de la réponse au choc thermique et du stress du RE dans la toxicité testiculaire d'un modèle d'invertébré largement utilisé en écotoxicologie, suggérant un nouveau mode d'action des régulateurs de croissance des insectes ciblant la voie de l'hormone juvénile des arthropodes.

Le module corrélé à l'exposition au Met était caractérisé par une proportion importante de protéines spécifiques des amphipodes dont les fonctions ne sont pas encore caractérisées. Alors que le Met n'induit aucun effet sur le nombre de spermatozoïdes (Trapp et al., 2015), notre analyse de réseau a identifié un module enrichi en protéines présentant une homologie élevée avec les protéines non caractérisées chez *Hyalella azteca*. Trois de ces protéines ont également montré le plus haut niveau de connectivité intramodule, ce qui suggère qu'elles ont un rôle clé dans la structuration des interactions protéine-protéine dans ce module. Ces résultats soulignent la capacité d'une analyse de réseau de coexpression à identifier des protéines taxonomiquement restreintes ayant un intérêt physiologique potentiel dans des organismes modèles émergents. Ce constat montre également l'indépendance de l'analyse de réseau de coexpression face aux annotations moins présentes dans les bases de données pour les espèces non modèles notamment.

Ces résultats montrent le potentiel de l'analyse de réseau sans *a priori* pour identifier et décrire des fonctions biologiques encore inconnues impliquées dans la reproduction ou la toxicité moléculaire chez une espèce sentinelle, qui sont susceptibles d'être spécifiques à un taxon, comme cela a été rapporté pour d'autres protéines à évolution rapide impliquées dans la reconnaissance

des gamètes chez les invertébrés (Lobov et al., 2019; Vacquier and Swanson, 2011). L'analyse de réseau de coexpression permet de chercher des relations fonctionnelles entre protéines non annotées pour les associer avec d'autres protéines dont on connait les fonctions. On peut ensuite, en se basant sur l'hypothèse que ces protéines sont coexprimées, les associer à ces mêmes fonctions. En conclusion, nos résultats montrent que l'application de l'analyse de coexpression à l'écotoxicoprotéomique peut aider à identifier les MoA de contaminants chez des organismes non modèles et pertinents pour l'environnement, avec des connaissances génomiques limitées, fournissant de nouvelles hypothèses mécanistiques. De plus, les analyses de réseaux sont des approches relativement peu exploitées pour la compréhension de la physiologie moléculaire des organismes sentinelles en écotoxicologie. Ces outils permettront d'établir un lien entre les effets indésirables et les modules de gènes ou de protéines, ce qui éclairera le cadre de la voie des effets indésirables sur les mécanismes moléculaires sous-jacents de la toxicité.

2.2 Reconstruction des voies du métabolisme lipidique

Après avoir adapté nos données transcriptomiques en réduisant la redondance des isoformes, comme expliqué ci-dessus (voir la partie 1.2), nous avons pu reconstruire les voies métaboliques du mâle et de la femelle de *Gammarus fossarum*. Les deux bases de données utilisées contiennent 358 et 366 voies métaboliques, et 2764 et 2850 réactions enzymatiques, chez le mâle et la femelle respectivement. Pour les deux genres, plus de 80% des voies contiennent plus de 75% de réactions annotées, et seulement 1% en contiennent moins de 25%. En comparant ces résultats à d'autres bases de données d'organismes modèles disponibles sur BioCyc (Caspi et al., 2008), telles que *Drosophila melanogaster* (Baa-Puyoulet et al., 2019 ; McQuilton et al., 2012), les résultats de complétude d'annotation qualitative sont similaires, validant l'adaptation de la méthode CycADS à nos données transcriptomiques. Ce résultat permet d'envisager de mettre à profit les données déjà disponibles pour d'autres espèces non modèles (Cogne et al., 2019c) et ouvre de nouvelles

perspectives pour identifier les voies moléculaires impliquées dans les organes d'organismes sentinelles. Comme décrit dans la partie précédente (voir partie 1.2), les données transcriptomiques, du fait de la complexité liée à l'assemblage, peuvent contenir des erreurs ou des artefacts. Il est donc recommandé d'intégrer ou du moins multiplier les sources de données omiques pour optimiser l'annotation (Ge et al., 2003 ; Reeves et al., 2009). C'est pourquoi l'apport des données protéomiques permet de confirmer l'existence des enzymes impliquées dans les différentes voies métaboliques chez le gammare. Globalement, les profils protéomiques sont composés de 5062 protéines chez le mâle, et 5663 chez la femelle. En couplant les profils protéomiques aux annotations des voies métaboliques avec CycADS, nous retrouvons 1189 protéines validées par la protéomique chez le mâle, et 1320 chez la femelle.

Pour montrer l'intérêt des approches multiomiques, nous nous sommes placés dans le cadre du métabolisme lipidique. En effet, le métabolisme lipidique peut être la cible d'un grand nombre de contaminants chez de nombreuses espèces (Barros et al., 2020 ; Jord et al., 2015 ; Jordão et al., 2016 ; Neuparth et al., 2014). Chez les crustacés, les lipides jouent un rôle majeur dans de nombreux processus physiologiques. Pour extraire et identifier ces voies impliquées dans le métabolisme lipidique, nous nous sommes basés sur les catégories fonctionnelles Gene Ontology (GO). Nous avons observé 78 voies (dont 1108 enzymes uniques) impliquées dans le métabolisme lipidique chez le mâle et 76 voies (dont 802 enzymes uniques) chez la femelle. Chez le mâle et la femelle, on observe environ 80% des voies avec plus de 75% de réactions annotées. En comparaison avec deux espèces modèles comme *D. melanogaster* (Baa-Puyoulet et al., 2019 ; McQuilton et al., 2012) et *Daphnia pulex* (Baa-Puyoulet et al., 2015 ; Nordberg et al., 2014), nous retrouvons plus de voies annotées comme appartenant au métabolisme lipidique chez notre espèce *G. fossarum*. Ces comparaisons montrent la pertinence de notre stratégie pour la reconstruction des voies

métaboliques d'un organisme non modèle à partir du transcriptome, et ce même si de potentiels effets bases de données sont à prendre en compte. Les profils protéomiques ont permis de valider la présence de 155 protéines annotées comme impliquées dans les voies du métabolisme lipidique chez le mâle et 161 chez la femelle.

Pour évaluer s'il existe des voies du métabolisme lipidique spécifique à certains organes, nous nous sommes intéressés aux profils d'expression des protéines du métabolisme lipidique dans les branchies, le caecum, le céphalon, l'intestin, les gonades et le reste du corps du gammare à partir de nos données protéomiques. Après un prétraitement des données pour supprimer les protéines trop faiblement abondantes, il restait 93 protéines pour le mâle et 98 pour la femelle. Nous avons validé par la protéomique, la présence de protéines du ML dans tous les organes du mâle et de la femelle (Figure 45). Pour les deux genres, 3 profils d'expression de protéines spécifiques des branchies, des caeca et des gonades (testicules et ovaires), ont été identifiés, montrant un organotropisme marqué des différentes enzymes dans ces organes (Figure 45).

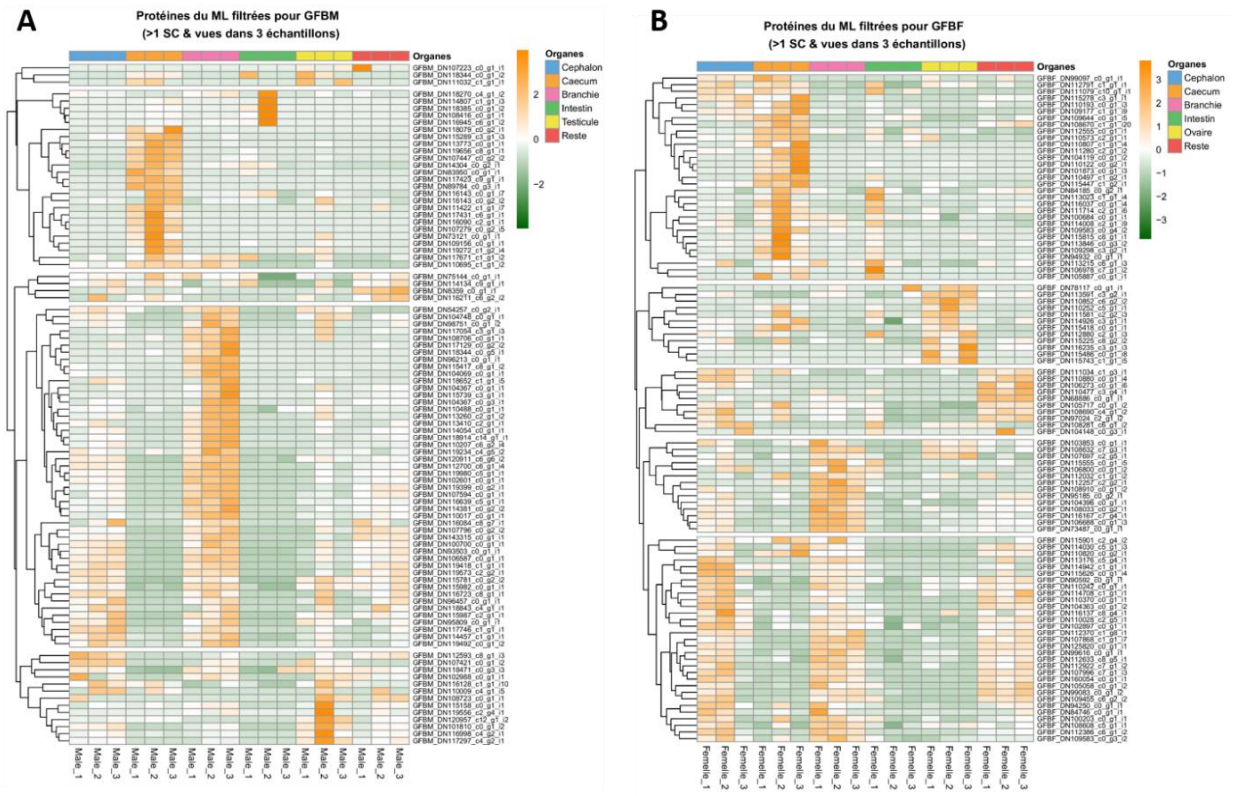


Figure 45 : Répartition des protéines vues en spectrométrie de masse selon les organes et les échantillons. (A) GFBM (B) GFBF.

Une analyse différentielle des spectral count a été menée pour identifier les voies et les enzymes du ML qui caractérisent cet organotropisme chez le mâle et la femelle de *G. fossarum.* Parmi les protéines abondantes dans les branchies, on retrouve des protéines annotées comme intervenant dans les voies de β-oxydation des acides gras (AGs), dans la synthèse des PUFAs ou encore dans la dégradation des plasmalogènes pour les mâles et les femelles. La β-oxydation des AGs permet de fournir de l'énergie à l'organe en brulant les AGs. Les plasmalogènes appartiennent à une catégorie de glycérophospholipides membranaires qui semblent jouer un rôle important dans la lutte contre les oxydants puisque les cellules qui en sont dépourvues par mutation génétique montrent une hypersensibilité vis-à-vis des oxydants (Aboshi et al., 2012; Nagan and Zoeller, 2001). On observe aussi dans les branchies du mâle, des protéines associées à la voie de la géranyldiphosphate par la voie du mévalonate. La voie du mévalonate est une voie qui permet la

synthèse d'hormones juvéniles (Bellés et al., 2005). C'est une étape clé dans la synthèse du méthyl-farnésoate qui joue un rôle important dans la reproduction des crustacés (Neuparth et al., 2014 ; Santos et al., 2016).

Les protéines différentiellement exprimées dans les testicules et les ovaires sont principalement des protéines associées à des voies de β-oxydation des AGs et de synthèse des PUFAs. Les PUFAs sont des constituants majoritairement des membranes et sont retrouvés chez les mammifères comme le rat et notamment dans les testicules (Salem et al., 2015). On note aussi la présence de voies impliquées dans la synthèse de l'anandamide. L'anandamide (composé cannabinoïde endogène) est présent principalement dans le cerveau, le rein et les testicules, notamment chez les invertébrés, vertébrés et mammifères (Battista et al., 2012). Il a été montré que l'activité testiculaire pouvait être régulée par l'anandamide chez la grenouille *Pelophylax esculentus* par le système de la kisspeptine (Ciaramella et al., 2016). Une étude a confirmé la présence des récepteurs à cannabinoïde chez le crustacé *Jasus edwardi* et d'autres espèces invertébrées (McPartland et al., 2006).

On retrouve aussi dans les gonades du mâle et de la femelle des protéines intervenant dans le métabolisme de la sphingosine. Les céramides (produits de la dégradation des sphingomyélines) sont métabolisées par la céramidase pour générer la sphingosine. La sphingosine est produite au cours des premiers stades de l'apoptose et permet sa régulation (Cuvillier, 2002 ; Pettus et al., 2002). Cette enzyme pourrait être un facteur maternel impliqué dans le développement embryonnaire comme la prolifération, la morphogenèse ou l'apoptose (Sinha Hikim and Swerdloff, 1999). Parmi les protéines en abondance différentielle dans les ovaires et les testicules, on retrouve des protéines impliquées dans la biosynthèse des plasmalogènes. Les plasmalogènes appartiennent à la famille des phospholipides et sont plus présents dans le céphalon, les nerfs, les

testicules et les branchies et peuvent jouer un rôle dans la perméabilité des membranes (Chapelle, 1987; Nagan and Zoeller, 2001; Rapport and Alonzo, 1960).

Ces travaux soulignent l'intérêt de l'intégration multiomiques pour permettre l'identification des voies métaboliques d'intérêt qu'on ne peut pas toujours obtenir avec l'annotation classique. Notre analyse sans *a priori* des profils d'enzymes lipidiques montre un organotropisme des différentes voies du métabolisme lipidique chez *G. fossarum*. Bien que ces résultats permettent de confirmer certaines hypothèses de la littérature, toutes les voies mises en avant ne sont pas strictement spécifiques aux organes cités. Nous avons pu retrouver la présence de certaines voies dans plusieurs organes, mais en abondance plus faible. L'organotropisme ainsi observé ne nous permet pas de conclure sur le fait que les différences d'abondance reflètent une réalité métabolique. Pour ceci, il serait intéressant de combiner des données métabolomiques et lipidomique pour valider ce type d'hypothèse.

3. PERSPECTIVES : LES ETUDES MULTIOMIQUES CHEZ LES ESPECES NON MODELES

Les études multiomiques commencent à voir le jour en écotoxicologie pour élucider les MoA des contaminants, mieux comprendre les mécanismes sous-jacents des contaminants et déterminer les liens entre les différentes molécules (Chen et al., 2018 ; Dumas et al., 2022 ; Ji et al., 2016 ; Lv et al., 2022 ; Sun et al., 2019 ; Xiang et al., 2021). La principale motivation pour l'intégration de données provenant de différents niveaux (*e.g.* le transcriptome, le protéome, le métabolome) est d'améliorer la compréhension du système biologique et d'approfondir les connaissances mécanistiques qui ne sont que partiellement interprétables en mesurant un seul niveau du système biologique (*e.g.* la transcriptomique seule) (Canzler et al., 2020). Cependant, l'analyse intégrative de données omiques reste difficile, en raison d'un certain degré d'hétérogénéité entre ces données et de la complexité intrinsèque des données générées (Wu et

al., 2019). Le défi consiste donc à extraire les signaux biologiques contenus dans les différentes sources omiques, masqués par les grandes variations (*i.e.* hétérogénéité, complexité) existantes dans les données. Différentes stratégies de fusion peuvent être envisagées en fonction de la structure et de la nature des données associées (Steinmetz et al., 1999). Nous avons proposé de combiner deux jeux de données omiques différents (transcriptomique et protéomique) pour comprendre et décrire le métabolisme lipidique. Cette stratégie nous a permis d'identifier les voies du ML chez l'espèce sentinelle non modèle *G. fossarum* en adaptant un outil destiné à des données de génomique chez les espèces modèles. Nous avons validé et précisé de manière exhaustive les annotations et la reconstruction des voies métaboliques obtenues à partir de la transcriptomique seule. L'apport des données protéomique a permis de confirmer la présence de protéines attribuées au ML par une approche sans *a priori*.

Du point de vue de l'analyse des données, les études multiomiques permettent une meilleure interprétation des résultats quand les échantillons sont appariés, c'est-à-dire sur des échantillons où toutes les couches omiques par réplicats sont générées à partir d'un seul individu (Canzler et al., 2020). Dans le cas d'échantillons non appariés, il est nécessaire de calculer les corrélations à partir des statistiques de « groupe » (*i.e.* agrégats) comme l'expression moyenne ou les fold-change par groupe de traitement, et non sur les échantillons indivuels. Cette agrégation limite l'utilisation de certaines méthodes d'intégration comme la réduction de dimensionnalité utilisée dans l'étude de Dumas et al (2022). Pour tendre vers ce type d'analyse multiomique, les échantillons de tissus d'un même individu devraient être divisés pour obtenir les différentes couches omiques (*i.e.* coextraction), alors que classiquement les données omiques proviennent de différents réplicats provenant de différents individus ou tissus (Cavill et al., 2016). À ce jour, pour les organismes de petite taille, des méthodes de co-extraction voient le jour pour permettre l'acquisition de plusieurs

données omiques de différente nature à partir d'un seul organisme (Faugere, 2022). Toutefois, pour réaliser des analyses organe-centrées, il sera probablement nécessaire de procéder à des pools (*i.e.* mutualisation) d'organes de différents organismes issus des mêmes conditions d'exposition.

Dans notre cas de figure, notre méthode s'approche d'une fusion de données de haut niveau (*i.e.* fusion après avoir traité les différents jeux données séparément, s'opposant au bas niveau en fusionnant à partir des données brutes) (Boccard and Rudaz, 2014; Smolinska et al., 2019). Ce type de fusion de données permet d'étudier individuellement chaque bloc de données (transcriptome et protéome) pour ensuite rassembler les résultats et offrir une image globale des données. Notre objectif était dans un premier temps descriptif et de comprendre le métabolisme lipidique chez *G. fossarum*. En raison de la complexité et du grand volume de données associées aux études multiomiques, il est nécessaire d'adapter l'analyse statistique à la question scientifique posée (Graw et al., 2021 ; Hasin et al., 2017). L'intégration et l'analyse statistique d'une étude multiomique dépendent de la sélection des plateformes omiques et des types de données qui leur sont associés (*e.g.* valeurs de comptage, pourcentages) (Figure 46). Cependant, un chercheur doit se poser plusieurs questions pour décider quel outil ou quelle méthode seront les plus appropriés (Graw et al., 2021), telles que : la collecte et la préparation des échantillons, la profondeur et/ou la qualité du séquençage, la compatibilité des données pour l'intégration multiomique, ou encore la quantité de signal qui peut potentiellement être perdue après normalisation et/ou filtrage (Misra et al., 2019). Comme indiqué ci-dessus, la question biologique est le moteur du type de méthode d'analyse choisi et des facteurs tels que l'échantillonnage, le type de plateforme et la qualité des données.

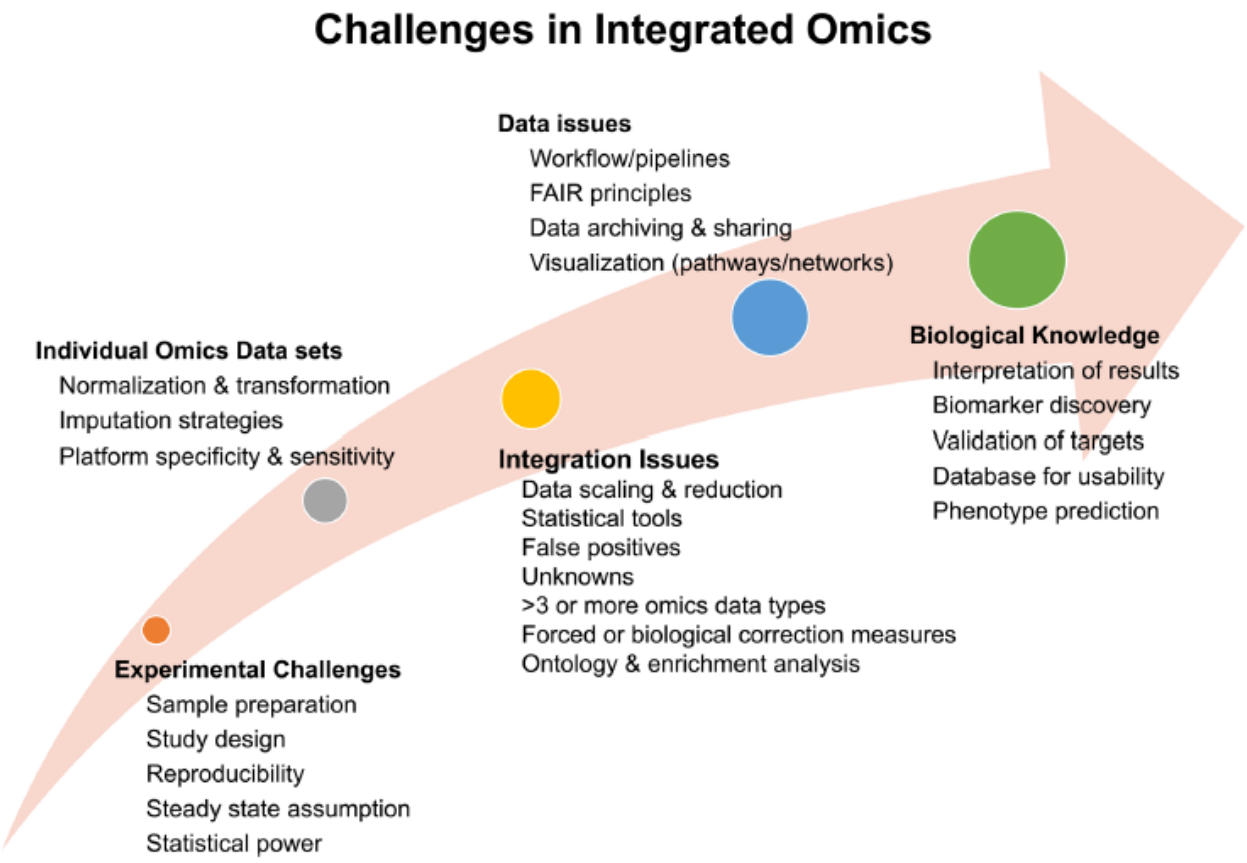


Figure 46 : Challenges pour l'intégration multiomique qui englobent (A) le design expérimental (B) les jeux de données omiques individuels (C) les problèmes d'intégration (D) les problèmes de données et (E) les connaissances biologiques (Misra et al., 2019).

De même, le choix des plates-formes omiques dépend fortement de la question de recherche. On aurait pu dans notre étude se contenter de la reconstruction des voies métaboliques à partir des données transcriptomiques uniquement. L'outil CycADS ne nécessite pas l'apport de données de protéomique, toutefois, nous n'aurions pas pu observer les différences de profils enzymatiques qui nous permet d'aller vers une vision intégrative multi-organes du métabolisme lipidique chez *Gammarus fossarum*. En général, la transcriptomique est souvent le premier choix pour les enquêtes écotoxicologiques en raison de son faible coût, de ses pipelines de calcul bien établis et de la plus grande quantité de données qu'elle permet d'obtenir par rapport à la protéomique ou à la métabolomique (Liang et al., 2020). Toutefois, les coûts de la protéomique shotgun sont

aujourd'hui comparables à ceux de la transcriptomique et permettent d'allier et d'étendre l'accessibilité de ce type d'omique pour les études écotoxicologiques.

En écotoxicologie, nous nous intéressons principalement aux réponses phénotypiques des contaminants environnementaux (Liang et al., 2020). Les protéines sont plus pertinentes que les transcrits, car elles sont les médiatrices directes du phénotype résultant (Liang et al., 2020), elles dirigent tous les niveaux du phénotype : les protéines structurelles dictent la forme physique, les enzymes catalysent les réactions biochimiques et les protéines agissent comme des acteurs de signalisation, des anticorps, des transporteurs, des pompes ioniques et des facteurs de transcription pour contrôler l'expression génétique (Liang et al., 2020). Ces effets traductionnels et post-traductionnels ne se reflètent pas au niveau du transcriptome. Une autre application clé de la protéomique en écotoxicologie consiste à prédire les effets écotoxicologiques des contaminants dans l'environnement. Associées à d'autres omiques (e.g. la transcriptomique et la métabolomique), les approches protéomiques non ciblées fournissent des informations moléculaires à grande échelle et peuvent mettre en évidence les biomarqueurs moléculaires de grande importance (Gouveia et al., 2019b).

Enfin, nous pourrions mettre en regard des données de métabolique, comme la lipidomique non ciblée (*i.e.* shotgun) sur des organes de *G. fossarum* pour compléter la caractérisation du métabolisme lipidique en validant l'existence de métabolites intervenant dans les voies du ML et proposer d'éventuelles hypothèses de biomarqueurs. Les lipides sont en effet les produits finaux de voies que l'on a identifiées et ils peuvent être très sensibles aux changements environnementaux et fournissent une mesure la plus directe du phénotype. À ce jour, il existe par exemple des données de lipidomique shotgun associées à des données de spectrométrie par imagerie chez *G. fossarum*,

pour explorer la composition et la distribution spatiale des lipides pour différents stades de reproduction et de développement (Fu et al., 2021).

3.1 Protéomique ciblée pour la modulation du métabolisme lipidique

L'intégration multiomique nous a permis d'identifier les acteurs clés du métabolisme lipidique chez une espèce sentinelle. À partir de ces résultats et des protéines annotées comme appartenant aux voies du métabolisme lipidique, une méthode d'analyse par spectrométrie de masse ciblée pourrait être développée pour tester l'intérêt de ces enzymes comme cibles des contaminants. En effet, cette approche permettrait de mesurer l'impact de la perturbation de processus physiologiques liés au métabolisme énergétique ou à l'exposition aux contaminants chez *G. fossarum*, sur ces protéines d'intérêt (Faugere et al., 2020 ; Lepretre et al., 2020). Ainsi, une première étape consisterait à mettre au point la méthode analytique pour la détection des protéines du ML en protéomique ciblée. Ici, l'apport de la bioinformatique serait de possiblement réduire le temps de mise au point de la méthode *in silico* en faisant un « pré-screening » des peptides (provenant de la protéomique shotgun) permettant d'orienter les peptides potentiels. En effet, un des verrous existants est la difficulté à prédire les peptides les plus facilement identifiables et détectables (peptides rapporteurs) en spectrométrie de masse.

À partir des séquences en acides aminés des protéines annotées du ML, on pourra retrouver dans les fichiers de protéomique shotgun les peptides associés et leur nombre de spectral count. Après validation de la liste finale de peptides à chercher en protéomique ciblée, les échantillons pourront être utilisés avec une méthode telle que la Scout-MRM (Faugere et al., 2020). Il sera possible avec les résultats de réaliser une analyse différentielle pour observer les variations dans l'expression des protéines liées au métabolisme lipidique et de pouvoir proposer des indicateurs de cette modulation. À partir des peptides rapporteurs ainsi identifiés, il serait possible de se focaliser

sur l'étude de la modulation de métabolisme lipidique, par exemple (i) en lien avec la physiologie et les mécanismes énergétiques, en modifiant le régime alimentaire de *G. fossarum*, (ii) ou pour l'identifier et caractériser des changements suite à une exposition aux contaminants.

3.2 Obtenir un transcriptome de référence pour *G. fossarum*

Pour les organismes pour lesquels un génome de référence n'est pas disponible, l'assemblage *de novo* du transcriptome permet d'avoir un aperçu des transcrits tissu-spécifique de l'espèce d'intérêt. Pour ce faire, nous avions à disposition deux transcriptomes distincts, un pour le mâle et un pour la femelle. Ces ressources ne nous ont pas permis de comparer directement les différences potentielles de compositions en protéines des deux genres. Pour pallier cette limite, nous proposons d'obtenir un transcriptome de « référence » pour l'espèce *G. fossarum* en fusionnant les deux transcriptomes de départ. Toutefois, si une approche idéale pour obtenir le transcriptome de référence n'est pas clairement définie dans la littérature, dans notre cas, il serait intéressant de combiner les transcriptomes des deux genres pour obtenir le transcriptome le plus représentatif en termes de diversité de répertoire génique et permettre la comparaison directe entre genres pour les analyses futures. La combinaison d'assemblages générés par des assembleurs *de novo* simples peut conduire à la construction d'un assemblage de transcriptome plus complet (Cerveau and Jackson, 2016; Huerlimann et al., 2018; Sadat-Hosseini et al., 2020). Pour cela, il existe deux types de stratégies, les méthodes préassemblage et les méthodes post-assemblage. Les méthodes préassemblage utilisent les lectures brutes des transcriptomes, puis les concatènent dans un seul fichier global de lectures et procèdent à l'assemblage de manière classique. De cette façon, même les transcrits les plus faiblement exprimés pourraient être capturés dans l'assemblage. Dans notre cas, il serait pertinent de commencer par une approche de fusion préassemblage à l'aide de l'assembleur Trinity. D'une part, Trinity est probablement l'algorithme le plus populaire pour l'analyse transcriptomique *de novo*, car il présente une bonne sensibilité avec un temps de calcul

raisonnable (Amil-Ruiz et al., 2021). Il récupère la plupart des transcrits exprimés sous forme de séquences complètes, mais il peut produire de grandes quantités de chimères trans-self (*i.e.* séquences assemblées dans des directions opposées) et de redondances (Yang and Smith, 2013). Il est donc nécessaire d'optimiser cette étape de nettoyage pour obtenir un transcriptome biologiquement pertinent et qui permette de réaliser les analyses en aval sans être parasité par des artefacts présents en trop grand nombre. D'autre part, nos deux transcriptomes originaux ont été assemblés à l'aide de Trinity, ce qui peut permettre de comparer les anciennes versions avec la version fusionnée plus facilement.

En résumé, il serait intéressant et pertinent d'établir une feuille de route consensus ou une plateforme ressource avec les avantages et inconvénients d'une stratégie selon la problématique écotoxicologique d'intérêt. Des revues de la littérature et des benchmarks permettent déjà entre autres de tendre vers cela en présentant des vues d'ensemble des processus et outils pour l'assemblage et l'annotation *de novo* des données RNA-Seq (Anamika et al., 2016 ; Lowe et al., 2017 ; Raghavan et al., 2022).

CONCLUSION GENERALE

Dans cette thèse, nous avons montré que les analyses de réseaux moléculaires sont des approches prometteuses, mais encore peu exploitées pour la compréhension de la physiologie moléculaire des organismes sentinelles en écotoxicologie. Ces outils permettent d'établir un lien entre les effets indésirables et les modules de gènes ou de protéines, en informant sur les voies et les mécanismes moléculaires sous-jacents de la toxicité. Ils montrent également la force et l'originalité des méthodes de réseaux de coexpression pour générer des hypothèses de travail sur des protéines spécifiques que les comparaisons d'homologie standard ou l'analyse d'expression différentielle n'auraient pas réussi à identifier. Finalement, nos résultats montrent que l'application de l'analyse de coexpression en écotoxicoprotéomique peut aider à identifier les mécanismes d'actions (MoA) des contaminants dans des organismes modèles pertinents pour l'environnement avec des connaissances génomiques limitées, fournissant de nouvelles hypothèses mécanistiques, bien que d'autres recherches expérimentales soient nécessaires pour confirmer ces hypothèses.

Notre approche de reconstruction des voies métaboliques par une intégration multiomique (*i.e.* transcriptomique et protéomique) nous a permis de faire la preuve de concept de l'adaptation d'un outil bioinformatique initialement destiné aux données d'organismes modèles, à nos données transcriptomiques dans un premier temps. Et dans un second temps, le couplage des données des profils protéomiques des différents organes de *G. fossarum* à nos données transcriptomiques, montre un organotropisme spécifique des différentes voies du métabolisme lipidique chez *G. fossarum*.

Dans l'ensemble, ce travail de thèse constitue une base solide pour l'utilisation des données omiques d'organismes non modèles déjà existantes dans une stratégie d'exploration sans *a priori* ou d'intégration multiomique. Les résultats ont permis de mettre en évidence des protéines potentiellement impliquées dans les processus biologiques liés aux stades de développement et de

reproduction de *G. fossarum*, ainsi que dans la toxicité testiculaire. Ces résultats soulignent également l'intérêt de l'application d'approches omiques au niveau des organes des espèces sentinelles pour identifier et évaluer les possibles différences de MoA de contaminants en fonction de l'organe cible. Ces deux méthodologies seraient applicables à diverses espèces non modèles ayant à disposition des données transcriptomiques et protéomiques, sans de nouvelles acquisitions de données.

REFERENCES BIBLIOGRAPHIQUES

Aboshi, T., Nishida, R., Mori, N., 2012. Identification of plasmalogen in the gut of silkworm (Bombyx mori). Insect Biochem Mol Biol 42, 596–601. https://doi.org/10.1016/j.ibmb.2012.04.006

Adam, O., Degiorgi, F., Crini, G., Badot, P.-M., 2010. High sensitivity of Gammarus sp. juveniles to deltamethrin: Outcomes for risk assessment. Ecotoxicology and Environmental Safety 73, 1402–1407. https://doi.org/10.1016/j.ecoenv.2010.02.011

Aebersold, R., Mann, M., 2003. Mass spectrometry-based proteomics. Nature 422, 198–207. https://doi.org/10.1038/nature01511

Aimo, L., Liechti, R., Hyka-Nouspikel, N., Niknejad, A., Gleizes, A., Götz, L., Kuznetsov, D., David, F.P.A., van der Goot, F.G., Riezman, H., Bougueleret, L., Xenarios, I., Bridge, A., 2015. The SwissLipids knowledgebase for lipid biology. Bioinformatics 31, 2860–2866. https://doi.org/10.1093/bioinformatics/btv285

Aite, M., Chevallier, M., Frioux, C., Trottier, C., Got, J., Cortés, M.P., Mendoza, S.N., Carrier, G., Dameron, O., Guillaudeux, N., Latorre, M., Loira, N., Markov, G.V., Maass, A., Siegel, A., 2018. Traceability, reproducibility and wiki-exploration for "à-la-carte" reconstructions of genome-scale metabolic models. PLOS Computational Biology 14, e1006146. https://doi.org/10.1371/journal.pcbi.1006146

Ala, U., Piro, R.M., Grassi, E., Damasco, C., Silengo, L., Oti, M., Provero, P., Cunto, F.D., 2008. Prediction of Human Disease Genes by Human-Mouse Conserved Coexpression Analysis. PLOS Computational Biology 4, e1000043. https://doi.org/10.1371/journal.pcbi.1000043

Alava, V.R., Quinitio, E.T., De Pedro, J.B., Priolo, F.M.P., Orozco, Z.G.A., Wille, M., 2007. Lipids and fatty acids in wild and pond-reared mud crab Scylla serrata (Forsskål) during ovarian maturation and spawning. Aquaculture Research 38, 1468–1477. https://doi.org/10.1111/j.1365-2109.2007.01793.x

Albrecht, D., Kniemeyer, O., Brakhage, A.A., Guthke, R., 2010. Missing values in gel-based proteomics. PROTEOMICS 10, 1202–1211. https://doi.org/10.1002/pmic.200800576

Alfaro, A.C., Young, T., 2018. Showcasing metabolomic applications in aquaculture: a review. Reviews in Aquaculture 10, 135–152.

Allen, J.E., Pertea, M., Salzberg, S.L., 2004. Computational Gene Prediction Using Multiple Sources of Evidence. Genome Res. 14, 142–148. https://doi.org/10.1101/gr.1562804

Allen, J.E., Salzberg, S.L., 2005. JIGSAW: integration of multiple sources of evidence for gene prediction. Bioinformatics 21, 3596–3603. https://doi.org/10.1093/bioinformatics/bti609

Allen, W.V., 1972. Lipid transport in the dungeness crab Cancer magister dana. Comparative Biochemistry and Physiology Part B: Comparative Biochemistry 43, 193-IN8. https://doi.org/10.1016/0305-0491(72)90216-7

Alonso, A., Marsal, S., Julià, A., 2015. Analytical Methods in Untargeted Metabolomics: State of the Art in 2015. Frontiers in Bioengineering and Biotechnology 3.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. Journal of molecular biology 215, 403–410.

Amiard, J.-C., Amiard-Triquiet, C., 2008. Les biomarqueurs dans l'évaluation de l'état écologique des milieux aquatiques. Lavoisier.

Amil-Ruiz, F., Maria Herruzo-Ruiz, A., Fuentes-Almagro, C., Baena-Angulo, C., Manuel Jimenez-Pastor, J., Blasco, J., Alhama, J., Michan, C., 2021. Constructing a de novo transcriptome and a reference proteome for the bivalve Scrobicularia plana: Comparative analysis of different assembly strategies and proteomic analysis. Genomics 113, 1543–1553. https://doi.org/10.1016/j.ygeno.2021.03.025

Anamika, K., Verma, S., Jere, A., Desai, A., 2016. Transcriptomic profiling using next generation sequencing-advances, advantages, and challenges. Next generation sequencing-advances, applications and challenges 9, 7355–7365.

Anders, S., Huber, W., 2010. Differential expression analysis for sequence count data. Nat Prec 1–1. https://doi.org/10.1038/npre.2010.4282.1

Andrews, S., 2010. FastQC: a quality control tool for high throughput sequence data.

Ankley, G.T., Bencic, D.C., Breen, M.S., Collette, T.W., Conolly, R.B., Denslow, N.D., Edwards, S.W., Ekman, D.R., Garcia-Reyero, N., Jensen, K.M., Lazorchak, J.M., Martinović, D., Miller, D.H., Perkins, E.J., Orlando, E.F., Villeneuve, D.L., Wang, R.-L., Watanabe, K.H., 2009. Endocrine disrupting chemicals in fish: Developing exposure indicators and predictive models of effects based on mechanism of action. Aquatic Toxicology 92, 168–178. https://doi.org/10.1016/j.aquatox.2009.01.013

Ankley, G.T., Daston, G.P., Degitz, S.J., Denslow, N.D., Hoke, R.A., Kennedy, S.W., Miracle, A.L., Perkins, E.J., Snape, J., Tillitt, D.E., 2006. Toxicogenomics in regulatory ecotoxicology.

Ansong, C., Purvine, S.O., Adkins, J.N., Lipton, M.S., Smith, R.D., 2008. Proteogenomics: needs and roles to be filled by proteomics in genome annotation. Briefings in Functional Genomics 7, 50–62. https://doi.org/10.1093/bfgp/elm010

Arambourou, H., Fuertes, I., Vulliet, E., Daniele, G., Noury, P., Delorme, N., Abbaci, K., Barata, C., 2018. Fenoxycarb exposure disrupted the reproductive success of the amphipod Gammarus fossarum with limited effects on the lipid profile. PLOS ONE 13, e0196461. https://doi.org/10.1371/journal.pone.0196461

Armengaud, J., 2009. A perfect genome annotation is within reach with the proteomics and genomics alliance. Current Opinion in Microbiology, Ecology and Industrial Microbiology • Techniques 12, 292–300. https://doi.org/10.1016/j.mib.2009.03.005

Armengaud, J., Trapp, J., Pible, O., Geffard, O., Chaumot, A., Hartmann, E.M., 2014a. Non-model organisms, a species endangered by proteogenomics. Journal of Proteomics, Special Issue: Proteomics of non-model organisms 105, 5–18. https://doi.org/10.1016/j.jprot.2014.01.007

Armengaud, J., Trapp, J., Pible, O., Geffard, O., Chaumot, A., Hartmann, E.M., 2014b. Non-model organisms, a species endangered by proteogenomics. Journal of Proteomics, Special Issue: Proteomics of non-model organisms 105, 5–18. https://doi.org/10.1016/j.jprot.2014.01.007

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G., 2000. Gene Ontology: tool for the unification of biology. Nat Genet 25, 25–29. https://doi.org/10.1038/75556

Avila-Campillo, I., Drew, K., Lin, J., Reiss, D.J., Bonneau, R., 2007. BioNetBuilder: automatic integration of biological networks. Bioinformatics 23, 392–393. https://doi.org/10.1093/bioinformatics/btl604

Ayciriex, S., 2010. Caractérisation de lysolipide acyltransférases chez S. cerevisiae - Apport de la Spectrométrie de Masse (These de doctorat). Bordeaux 2.

Baa-Puyoulet, P., Parisot, N., Charles, H., Calevro, F., 2019. ArthropodaCyc - Drosophila melanogaster [WWW Document]. Summary of Drosophila melanogaster, version 1.0.1. URL http://arthropodacyc.cycadsys.org/organism-summary?object=DROME (accessed 12.18.21).

Baa-Puyoulet, P., Parisot, N., Colella, S., 2015. ArthropodaCyc - Daphnia pulex [WWW Document]. Summary of Daphnia pulex, version 1.0. URL http://arthropodacyc.cycadsys.org/organism-summary?object=DAPPU (accessed 12.18.21).

Baa-Puyoulet, P., Parisot, N., Febvay, G., Huerta-Cepas, J., Vellozo, A.F., Gabaldón, T., Calevro, F., Charles, H., Colella, S., 2016. ArthropodaCyc: a CycADS powered collection of BioCyc databases to analyse and compare metabolism of arthropods. Database (Oxford) 2016. https://doi.org/10.1093/database/baw081

Banks, J.E., Stark, J.D., 1998. What is ecotoxicology? An ad-hoc grab bag or an interdisciplinary science? Integrative Biology: Issues, News, and Reviews: Published in Association with The Society for Integrative and Comparative Biology 1, 195–204.

Barek, H., Sugumaran, M., Ito, S., Wakamatsu, K., 2018. Insect cuticular melanins are distinctly different from those of mammalian epidermal melanins. Pigment Cell Melanoma Res 31, 384–392. https://doi.org/10.1111/pcmr.12672

Barnard, J.L., Barnard, C.M., 1983. Freshwater Amphipoda of the world, I Evolutionary patterns, II Handbook and bibliography. Hayfield Associates, Mt. Vernon, Virginia.

Barros, S., Coimbra, A.M., Alves, N., Pinheiro, M., Quintana, J.B., Santos, M.M., Neuparth, T., 2020. Chronic exposure to environmentally relevant levels of simvastatin disrupts zebrafish brain gene signaling involved in energy metabolism. J Toxicol Environ Health A 83, 113–125. https://doi.org/10.1080/15287394.2020.1733722

Barthelson, R., McFarlin, A.J., Rounsley, S.D., Young, S., 2011. Plantagora: modeling whole genome sequencing and assembly of plant genomes. PLoS One 6, e28436. https://doi.org/10.1371/journal.pone.0028436

Barupal, D.K., Fan, S., Fiehn, O., 2018. Integrating bioinformatics approaches for a comprehensive interpretation of metabolomics datasets. Current Opinion in Biotechnology, Analytical Biotechnology 54, 1–9. https://doi.org/10.1016/j.copbio.2018.01.010

Barupal, D.K., Fiehn, O., 2017. Chemical Similarity Enrichment Analysis (ChemRICH) as alternative to biochemical pathway mapping for metabolomic datasets. Sci Rep 7, 14567. https://doi.org/10.1038/s41598-017-15231-w

Battista, N., Meccariello, R., Cobellis, G., Fasano, S., Di Tommaso, M., Pirazzi, V., Konje, J.C., Pierantoni, R., Maccarrone, M., 2012. The role of endocannabinoids in gonadal function and fertility along the evolutionary axis. Mol Cell Endocrinol 355, 1–14. https://doi.org/10.1016/j.mce.2012.01.014

Bellés, X., Martín, D., Piulachs, M.-D., 2005. The Mevalonate Pathway and the Synthesis of Juvenile Hormone in Insects. Annual Review of Entomology 50, 181–199. https://doi.org/10.1146/annurev.ento.50.071803.130356

Benson, W.H., Giulio, R.T.D. (Eds.), 2006. Genomic Approaches for Cross-Species Extrapolation in Toxicology. CRC Press, Boca Raton. https://doi.org/10.1201/9781420043648

Benton, H.P., Ivanisevic, J., Mahieu, N.G., Kurczy, M.E., Johnson, C.H., Franco, L., Rinehart, D., Valentine, E., Gowda, H., Ubhi, B.K., Tautenhahn, R., Gieschen, A., Fields, M.W., Patti, G.J., Siuzdak, G., 2015. Autonomous Metabolomics for Rapid Metabolite Identification in Global Profiling. Anal. Chem. 87, 884–891. https://doi.org/10.1021/ac5025649

Bersanelli, M., Mosca, E., Remondini, D., Giampieri, E., Sala, C., Castellani, G., Milanesi, L., 2016. Methods for the integration of multi-omics data: mathematical aspects. BMC Bioinformatics 17 Suppl 2, 15. https://doi.org/10.1186/s12859-015-0857-9

Bertin, D., Labadie, P., Ferrari, B.J.D., Sapin, A., Garric, J., Geffard, O., Budzinski, H., Babut, M., 2016. Potential exposure routes and accumulation kinetics for poly- and perfluorinated alkyl compounds for a freshwater amphipod: Gammarus spp. (Crustacea). Chemosphere 155, 380–387. https://doi.org/10.1016/j.chemosphere.2016.04.006

Besse, J.-P., Coquery, M., Lopes, C., Chaumot, A., Budzinski, H., Labadie, P., 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 Research 47, 650–660. https://doi.org/10.1016/j.watres.2012.10.024

Beyer, J., Green, N.W., Brooks, S., Allan, I.J., Ruus, A., Gomes, T., Bråte, I.L.N., Schøyen, M., 2017. Blue mussels (Mytilus edulis spp.) as sentinel organisms in coastal pollution monitoring: A review. Marine Environmental Research 130, 338–365. https://doi.org/10.1016/j.marenvres.2017.07.024

Bhardwaj, J.K., Panchal, H., Saraf, P., 2020. Cadmium as a testicular toxicant: A Review. J Appl Toxicol. https://doi.org/10.1002/jat.4055

Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., Fridman, W.-H., Pagès, F., Trajanoski, Z., Galon, J., 2009. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics 25, 1091–1093. https://doi.org/10.1093/bioinformatics/btp101

Bingol, K., Bruschweiler-Li, L., Li, D., Zhang, B., Xie, M., Brüschweiler, R., 2016. Emerging new strategies for successful metabolite identification in metabolomics. Bioanalysis 8, 557–573. https://doi.org/10.4155/bio-2015-0004

Blaženović, I., Kind, T., Ji, J., Fiehn, O., 2018. Software Tools and Approaches for Compound Identification of LC-MS/MS Data in Metabolomics. Metabolites 8, 31. https://doi.org/10.3390/metabo8020031

Blum, M., Chang, H.-Y., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., Nuka, G., Paysan-Lafosse, T., Qureshi, M., Raj, S., Richardson, L., Salazar, G.A., Williams, L., Bork, P., Bridge, A., Gough, J., Haft, D.H., Letunic, I., Marchler-Bauer, A., Mi, H., Natale, D.A., Necci, M., Orengo, C.A., Pandurangan, A.P., Rivoire, C., Sigrist, C.J.A., Sillitoe, I., Thanki, N., Thomas, P.D., Tosatto, S.C.E., Wu, C.H., Bateman, A., Finn, R.D., 2021. The InterPro protein families and domains database: 20 years on. Nucleic Acids Research 49, D344–D354. https://doi.org/10.1093/nar/gkaa977

Boccard, J., Rudaz, S., 2014. Harnessing the complexity of metabolomic data with chemometrics. Journal of Chemometrics 28, 1–9. https://doi.org/10.1002/cem.2567

Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D., Pirovano, W., 2011. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics 27, 578–579. https://doi.org/10.1093/bioinformatics/btq683

Boetzer, M., Pirovano, W., 2012. Toward almost closed genomes with GapFiller. Genome Biol 13, R56. https://doi.org/10.1186/gb-2012-13-6-r56

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bonnefoy, C., Fildier, A., Buleté, A., Bordes, C., Garric, J., Vulliet, E., 2019. Untargeted analysis of nanoLC-HRMS data by ANOVA-PCA to highlight metabolites in Gammarus fossarum after in vivo exposure to pharmaceuticals. Talanta 202, 221–229. https://doi.org/10.1016/j.talanta.2019.04.028

Bonnet, E., Calzone, L., Michoel, T., 2015. Integrative Multi-omics Module Network Inference with Lemon-Tree. PLOS Computational Biology 11, e1003983. https://doi.org/10.1371/journal.pcbi.1003983

Bradman, K., 2015. Tales of drafty genomes: part 3 – all genomes are complete…except for those that aren't [WWW Document]. ACGT. URL http://www.acgt.me/blog/2015/3/3/tales-of-drafty-genomes-part-3-all-genomes-are-completeexcept-for-those-that-arent (accessed 3.23.22).

Bradnam, K.R., Fass, J.N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., Boisvert, S., Chapman, J.A., Chapuis, G., Chikhi, R., Chitsaz, H., Chou, W.-C., Corbeil, J., Del Fabbro, C., Docking, T.R., Durbin, R., Earl, D., Emrich, S., Fedotov, P., Fonseca, N.A., Ganapathy, G., Gibbs, R.A., Gnerre, S., Godzaridis, E., Goldstein, S., Haimel, M., Hall, G., Haussler, D., Hiatt, J.B., Ho, I.Y., Howard, J., Hunt, M., Jackman, S.D., Jaffe, D.B., Jarvis, E.D., Jiang, H., Kazakov, S., Kersey, P.J., Kitzman, J.O., Knight, J.R., Koren, S., Lam, T.-W., Lavenier, D., Laviolette, F., Li, Y., Li, Z., Liu, B., Liu, Y., Luo, R., Maccallum, I., Macmanes, M.D., Maillet, N., Melnikov, S., Naquin, D., Ning, Z., Otto, T.D., Paten, B., Paulo, O.S., Phillippy, A.M., Pina-Martins, F., Place, M., Przybylski, D., Qin, X., Qu, C., Ribeiro, F.J., Richards, S., Rokhsar, D.S., Ruby, J.G., Scalabrin, S., Schatz, M.C., Schwartz, D.C., Sergushichev, A., Sharpe, T., Shaw, T.I., Shendure, J., Shi, Y., Simpson, J.T., Song, H., Tsarev, F., Vezzi, F., Vicedomini, R., Vieira, B.M., Wang, J., Worley, K.C., Yin, S., Yiu, S.-M., Yuan, J., Zhang, G., Zhang, H., Zhou, S., Korf, I.F., 2013. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. Gigascience 2, 10. https://doi.org/10.1186/2047-217X-2-10

Branson, O.E., Freitas, M.A., 2016. A multi-model statistical approach for proteomic spectral count quantitation. J Proteomics 144, 23–32. https://doi.org/10.1016/j.jprot.2016.05.032

Bray, N.L., Pimentel, H., Melsted, P., Pachter, L., 2016. Near-optimal probabilistic RNA-seq quantification. Nature Biotechnology 34, 525–527. https://doi.org/10.1038/nbt.3519

Brendel, V., Kleffe, J., Carle-Urioste, J.C., Walbot, V., 1998. Prediction of splice sites in plant pre-mRNA from sequence properties. Journal of molecular biology 276, 85–104.

Brunk, E., Sahoo, S., Zielinski, D.C., Altunkaya, A., Dräger, A., Mih, N., Gatto, F., Nilsson, A., Preciat Gonzalez, G.A., Aurich, M.K., Prlić, A., Sastry, A., Danielsdottir, A.D., Heinken, A., Noronha, A., Rose, P.W., Burley, S.K., Fleming, R.M.T., Nielsen, J., Thiele, I., Palsson, B.O., 2018. Recon3D enables a three-dimensional view of gene variation in human metabolism. Nat Biotechnol 36, 272–281. https://doi.org/10.1038/nbt.4072

Bryant, D.M., Johnson, K., DiTommaso, T., Tickle, T., Couger, M.B., Payzin-Dogru, D., Lee, T.J., Leigh, N.D., Kuo, T.-H., Davis, F.G., Bateman, J., Bryant, S., Guzikowski, A.R., Tsai, S.L., Coyne, S., Ye, W., Freeman, R.M., Peshkin, L., Tabin, C.J., Regev, A., Haas, B.J., Whited, J.L., 2017. A tissue-mapped axolotl de novo transcriptome enables identification of limb regeneration factors. Cell Rep 18, 762–776. https://doi.org/10.1016/j.celrep.2016.12.063

Buckingham, S., 2003. The major world of microRNAs. Nature, 2nd Symposium. Understanding the RNAissance.

Buikema Jr, A.L., Benfield, E.F., 1979. Use of macroinvertebrate life history information in toxicity tests. Journal of the Fisheries Board of Canada 36, 321–328.

Burge, C., Karlin, S., 1997. Prediction of complete gene structures in human genomic DNA. Journal of molecular biology 268, 78–94.

Bushmanova, E., Antipov, D., Lapidus, A., Prjibelski, A.D., 2019. rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. GigaScience 8, giz100. https://doi.org/10.1093/gigascience/giz100

Bushmanova, E., Antipov, D., Lapidus, A., Suvorov, V., Prjibelski, A.D., 2016. rnaQUAST: a quality assessment tool for de novo transcriptome assemblies. Bioinformatics 32, 2210–2212. https://doi.org/10.1093/bioinformatics/btw218

Butler, G.C., 1978. Principles of ecotoxicology. Wiley New York.

Butler, J., MacCallum, I., Kleber, M., Shlyakhter, I.A., Belmonte, M.K., Lander, E.S., Nusbaum, C., Jaffe, D.B., 2008. ALLPATHS: de novo assembly of whole-genome shotgun microreads. Genome Res 18, 810–820. https://doi.org/10.1101/gr.7337908

Calisi, A., Latino, M.E., Corallo, A., Grimaldi, A., Ferronato, C., Antisari, L.V., Dondero, F., 2019. Biomarkers in Soil Organisms: Their Potential use in the Assessment of Soil Pollution and Remediation, in: Bioremediation of Agricultural Soils. CRC Press.

Calow, P., Sibly, R.M., Forbes, V., 1997. Risk assessment on the basis of simplified life-history scenarios. Environmental Toxicology and Chemistry: An International Journal 16, 1983–1989.

Campos, A., Danielsson, G., Farinha, A.P., Kuruvilla, J., Warholm, P., Cristobal, S., 2016. Shotgun proteomics to unravel marine mussel (Mytilus edulis) response to long-term exposure to low salinity and propranolol in a Baltic Sea microcosm. Journal of Proteomics, Environment and (Prote)-OMICS 137, 97–106. https://doi.org/10.1016/j.jprot.2016.01.010

Cantarel, B.L., Korf, I., Robb, S.M.C., Parra, G., Ross, E., Moore, B., Holt, C., Sánchez Alvarado, A., Yandell, M., 2008. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res 18, 188–196. https://doi.org/10.1101/gr.6743907

Canzler, S., Schor, J., Busch, W., Schubert, K., Rolle-Kampczyk, U.E., Seitz, H., Kamp, H., von Bergen, M., Buesen, R., Hackermüller, J., 2020. Prospects and challenges of multi-omics data integration in toxicology. Arch Toxicol 94, 371–388. https://doi.org/10.1007/s00204-020-02656-y

Capitão, A., Lyssimachou, A., Castro, L.F.C., Santos, M.M., 2017. Obesogens in the aquatic environment: an evolutionary and toxicological perspective. Environment International 106, 153–169. https://doi.org/10.1016/j.envint.2017.06.003

Caputo, D.R., 2020. The Effects of a Wastewater Effluent on Population, Transcriptomic and Metabolomic Markers in the Freshwater Amphipod Gammarus Fossarum. University of Portsmouth.

Caputo, D.R., Robson, S.C., Werner, I., Ford, A.T., 2020. Complete transcriptome assembly and annotation of a critically important amphipod species in freshwater ecotoxicological risk assessment: Gammarus fossarum. Environment International 137, 105319. https://doi.org/10.1016/j.envint.2019.105319

Carlini, D.B., Fong, D.W., 2017. 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. https://doi.org/10.1371/journal.pone.0186173

Carvalho, M., Sampaio, J.L., Palm, W., Brankatschk, M., Eaton, S., 2012. Effects of diet and development on the Drosophila lipidome. Molecular Systems Biology 8, 600. https://doi.org/10.1038/msb.2012.29

Caspi, R., Billington, R., Keseler, I.M., Kothari, A., Krummenacker, M., Midford, P.E., Ong, W.K., Paley, S., Subhraveti, P., Karp, P.D., 2020. The MetaCyc database of metabolic pathways and enzymes - a 2019 update. Nucleic Acids Research 48, D445–D453. https://doi.org/10.1093/nar/gkz862

Caspi, R., Foerster, H., Fulcher, C.A., Kaipa, P., Krummenacker, M., Latendresse, M., Paley, S., Rhee, S.Y., Shearer, A.G., Tissier, C., Walk, T.C., Zhang, P., Karp, P.D., 2008. The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res 36, D623-631. https://doi.org/10.1093/nar/gkm900

Castille, F.L., Lawrence, A.L., 1989. Relationship Between Maturation and Biochemical Composition of the Gonads and Digestive Glands of the Shrimps Penaeus Aztecus Ives and Penaeus Setiferus (L.). Journal of Crustacean Biology 9, 202–211. https://doi.org/10.1163/193724089X00025

Cavill, R., Jennen, D., Kleinjans, J., Briedé, J.J., 2016. Transcriptomic and metabolomic data integration. Briefings in Bioinformatics 17, 891–901. https://doi.org/10.1093/bib/bbv090

Cerveau, N., Jackson, D.J., 2016. Combining independent de novo assemblies optimizes the coding transcriptome for nonconventional model eukaryotic organisms. BMC Bioinformatics 17, 525. https://doi.org/10.1186/s12859-016-1406-x

Chapelle, S., 1987. Plasmalogens and O-alkylglycerophospholipids in aquatic animals. Comparative Biochemistry and Physiology Part B: Comparative Biochemistry 88, 1–6. https://doi.org/10.1016/0305-0491(87)90068-X

Chari, R., Coe, B.P., Vucic, E.A., Lockwood, W.W., Lam, W.L., 2010. An integrative multi-dimensional genetic and epigenetic strategy to identify aberrant genes and pathways in cancer. BMC Systems Biology 4, 67. https://doi.org/10.1186/1752-0509-4-67

Charnot, A., Gouveia, D., Armengaud, J., Almunia, C., Chaumot, A., Lemoine, J., Geffard, O., 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, 3969–3991. https://doi.org/10.1007/s00216-017-0348-0

Charnot, A., Gouveia, D., Ayciriex, S., Lemoine, J., Armengaud, J., Almunia, C., Chaumot, A., Geffard, O., Salvador, A., 2018. On-line solid phase extraction liquid chromatography-mass spectrometry method for multiplexed proteins quantitation in an ecotoxicology test specie: gammarus fossarum. Journal of Applied Bioanalysis 4, 81–101.

Charron, L., Geffard, O., Chaumot, A., Coulaud, R., Queau, H., Geffard, A., Dedourge-Geffard, O., 2013. Effect of water quality and confounding factors on digestive enzyme activities in Gammarus fossarum. Environ Sci Pollut Res 20, 9044–9056. https://doi.org/10.1007/s11356-013-1921-5

Chaumot, A., Coulaud, R., Adam, O., Quéau, H., Lopes, C., Geffard, O., 2020. In Situ Reproductive Bioassay with Caged Gammarus fossarum (Crustacea): Part 1—Gauging the Confounding Influence of Temperature and Water Hardness. Environmental Toxicology and Chemistry 39, 667–677. https://doi.org/10.1002/etc.4655

Chaumot, A., Geffard, O., Armengaud, J., Maltby, L., 2015. Gammarids as reference species for freshwater monitoring, in: Aquatic Ecotoxicology. Elsevier, pp. 253–280.

Chen, C., Hou, J., Tanner, J.J., Cheng, J., 2020. Bioinformatics Methods for Mass Spectrometry-Based Proteomics Data Analysis. Int J Mol Sci 21, 2873. https://doi.org/10.3390/ijms21082873

Chen, E.Y., Tan, C.M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G.V., Clark, N.R., Ma'ayan, A., 2013. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics 14, 128. https://doi.org/10.1186/1471-2105-14-128

Chen, H., Diao, X., Wang, H., Zhou, H., 2018. An integrated metabolomic and proteomic study of toxic effects of Benzo[a]pyrene on gills of the pearl oyster Pinctada martensii. Ecotoxicology and Environmental Safety 156, 330–336. https://doi.org/10.1016/j.ecoenv.2018.03.040

Chen, J., Bardes, E.E., Aronow, B.J., Jegga, A.G., 2009. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Research 37, W305–W311. https://doi.org/10.1093/nar/gkp427

Chen, J., Liu, H., Cai, S., Zhang, H., 2019. Comparative transcriptome analysis of Eogammarus possjeticus at different hydrostatic pressure and temperature exposures. Sci Rep 9, 3456. https://doi.org/10.1038/s41598-019-39716-y

Chen, Q.-L., Luo, Z., Huang, C., Zheng, J.-L., Pan, Y.-X., Song, Y.-F., Hu, W., 2015. Molecular cloning and tissue mRNA levels of 15 genes involved in lipid metabolism in Synechogobius hasta. European Journal of Lipid Science and Technology 117, 471–482. https://doi.org/10.1002/ejlt.201400164

Chen, Y., Wu, X., Jiang, R., 2013. Integrating human omics data to prioritize candidate genes. BMC Med Genomics 6, 57. https://doi.org/10.1186/1755-8794-6-57

Chevreux, B., Pfisterer, T., Drescher, B., Driesel, A.J., Müller, W.E.G., Wetter, T., Suhai, S., 2004. Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. Genome Res 14, 1147–1159. https://doi.org/10.1101/gr.1917404

Chong, J., Wishart, D.S., Xia, J., 2019. Using MetaboAnalyst 4.0 for Comprehensive and Integrative Metabolomics Data Analysis. Current Protocols in Bioinformatics 68, e86. https://doi.org/10.1002/cpbi.86

Chong, J., Xia, J., 2017. Computational Approaches for Integrative Analysis of the Metabolome and Microbiome. Metabolites 7, 62. https://doi.org/10.3390/metabo7040062

Christie-Oleza, J.A., Miotello, G., Armengaud, J., 2012. High-throughput proteogenomics of Ruegeria pomeroyi: seeding a better genomic annotation for the whole marine Roseobacter clade. BMC Genomics 13, 73. https://doi.org/10.1186/1471-2164-13-73

Ciaramella, V., Meccariello, R., Chioccarelli, T., Sirleto, M., Fasano, S., Pierantoni, R., Chianese, R., 2016. Anandamide acts via kisspeptin in the regulation of testicular activity of the frog, Pelophylax esculentus. Mol Cell Endocrinol 420, 75–84. https://doi.org/10.1016/j.mce.2015.11.011

Clair, G., Reehl, S., Stratton, K.G., Monroe, M.E., Tfaily, M.M., Ansong, C., Kyle, J.E., 2019. Lipid Mini-On: mining and ontology tool for enrichment analysis of lipidomic data. Bioinformatics 35, 4507–4508. https://doi.org/10.1093/bioinformatics/btz250

Claudel-Renard, C., Chevalet, C., Faraut, T., Kahn, D., 2003. Enzyme-specific profiles for genome annotation: PRIAM. Nucleic acids research 31, 6633–6639.

Cogne, Y., 2019. Bioinformatique pour l'exploration de la diversité inter-espèces et inter-populations: hétérogénéité & données multi-omiques 121.

Cogne, Y., Almunia, C., Gouveia, D., Pible, O., François, A., Degli-Esposti, D., Geffard, O., Armengaud, J., Chaumot, A., 2019a. Comparative proteomics in the wild: Accounting for intrapopulation variability improves describing proteome response in a Gammarus pulex field population exposed to cadmium. Aquatic Toxicology 214, 105244. https://doi.org/10.1016/j.aquatox.2019.105244

Cogne, Y., Degli Esposti, D., Pible, O., Gouveia, D., Geffard, O., Chaumot, A., 2019b. YC02. figshare. https://doi.org/10.6084/m9.figshare.c.4568087.v1

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., 2019c. De novo transcriptomes of 14 gammarid individuals for proteogenomic analysis of seven taxonomic groups. Sci Data 6, 1–7. https://doi.org/10.1038/s41597-019-0192-5

Colbourne, J.K., Pfrender, M.E., Gilbert, D., Thomas, W.K., Tucker, A., Oakley, T.H., Tokishita, S., Aerts, A., Arnold, G.J., Basu, M.K., Bauer, D.J., Cáceres, C.E., Carmel, L., Casola, C., Choi, J.-H., Detter, J.C., Dong, Q., Dusheyko, S., Eads, B.D., Fröhlich, T., Geiler-Samerotte, K.A., Gerlach, D., Hatcher, P., Jogdeo, S., Krijgsveld, J., Kriventseva, E.V., Kültz, D., Laforsch, C., Lindquist, E., Lopez, J., Manak, J.R., Muller, J., Pangilinan, J., Patwardhan, R.P., Pitluck, S., Pritham, E.J., Rechtsteiner, A., Rho, M., Rogozin, I.B., Sakarya, O., Salamov, A., Schaack, S., Shapiro, H., Shiga, Y., Skalitzky, C., Smith, Z., Souvorov, A., Sung, W., Tang, Z., Tsuchiya, D., Tu, H., Vos, H., Wang, M., Wolf, Y.I., Yamagata, H., Yamada, T., Ye, Y., Shaw, J.R., Andrews, J., Crease, T.J., Tang, H., Lucas, S.M., Robertson, H.M., Bork, P., Koonin, E.V., Zdobnov, E.M., Grigoriev, I.V., Lynch, M., Boore, J.L., 2011. The Ecoresponsive Genome of Daphnia pulex. Science 331, 555–561. https://doi.org/10.1126/science.1197761

Collins, M., Tills, O., Spicer, J.I., Truebano, M., 2017. De novo transcriptome assembly of the amphipod Gammarus chevreuxi exposed to chronic hypoxia. Marine genomics 33, 17–19.

Conesa, A., Götz, S., 2008. Blast2GO: A Comprehensive Suite for Functional Analysis in Plant Genomics. International Journal of Plant Genomics 2008, 1–12. https://doi.org/10.1155/2008/619832

Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M., Robles, M., 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21, 3674–3676.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M.W., Gaffney, D.J., Elo, L.L., Zhang, X., Mortazavi, A., 2016. A survey of best practices for RNA-seq data analysis. Genome Biology 17, 13. https://doi.org/10.1186/s13059-016-0881-8

Correia, A.D., Costa, M.H., Luis, O.J., Livingstone, D.R., 2003. Age-related changes in antioxidant enzyme activities, fatty acid composition and lipid peroxidation in whole body Gammarus locusta (Crustacea: Amphipoda). Journal of Experimental Marine Biology and Ecology 289, 83–101. https://doi.org/10.1016/S0022-0981(03)00040-6

Cottret, L., Frainay, C., Chazalviel, M., Cabanettes, F., Gloaguen, Y., Camenen, E., Merlet, B., Heux, S., Portais, J.-C., Poupin, N., Vinson, F., Jourdan, F., 2018. MetExplore: collaborative edition and exploration of metabolic networks. Nucleic Acids Res 46, W495–W502. https://doi.org/10.1093/nar/gky301

Coulaud, R., 2012. Modélisation et changements d'échelles pour l'évaluation écotoxicologique : application à deux macroinvertébrés aquatiques, Gammarus fossarum (crustacé amphipode) et potamopyrgus antipodarum (mollusque gastéropode) (phdthesis). Université Claude Bernard - Lyon I.

Coulaud, R., Geffard, O., Coquillat, A., Quéau, H., Charles, S., Chaumot, A., 2014. Ecological Modeling for the Extrapolation of Ecotoxicological Effects Measured during in Situ Assays in Gammarus. Environmental Science & Technology 48, 6428–6436. https://doi.org/10.1021/es501126g

Coulaud, R., Geffard, O., Xuereb, B., Lacaze, E., Quéau, H., Garric, J., Charles, S., Chaumot, A., 2011. In situ feeding assay with Gammarus fossarum (Crustacea): Modelling the influence of confounding factors to improve water quality biomonitoring. water research 45, 6417–6429.

Crappé, J., Ndah, E., Koch, A., Steyaert, S., Gawron, D., De Keulenaer, S., De Meester, E., De Meyer, T., Van Criekinge, W., Van Damme, P., Menschaert, G., 2015. PROTEOFORMER: deep proteome coverage through ribosome profiling and MS integration. Nucleic Acids Research 43, e29. https://doi.org/10.1093/nar/gku1283

Cribiu, P., 2020. Étude des effets inter et transgénérationnels de l'exposition parentale au stress chimique chez le crustacé amphipode Gammarus fossarum (phdthesis). Université de Lyon.

Croft, D., O'Kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., Caudy, M., Garapati, P., Gopinath, G., Jassal, B., Jupe, S., Kalatskaya, I., Mahajan, S., May, B., Ndegwa, N., Schmidt, E., Shamovsky, V., Yung, C., Birney, E., Hermjakob, H., D'Eustachio, P., Stein, L., 2011. Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res 39, D691-697. https://doi.org/10.1093/nar/gkq1018

Cumbie, J.S., Kimbrel, J.A., Di, Y., Schafer, D.W., Wilhelm, L.J., Fox, S.E., Sullivan, C.M., Curzon, A.D., Carrington, J.C., Mockler, T.C., Chang, J.H., 2011. GENE-Counter: A Computational Pipeline for the Analysis of RNA-Seq Data for Gene Expression Differences. PLOS ONE 6, e25279. https://doi.org/10.1371/journal.pone.0025279

Cuvillier, O., 2002. Sphingosine in apoptosis signaling. Biochim Biophys Acta 1585, 153–162. https://doi.org/10.1016/s1388-1981(02)00336-0

Davidson, E., Levin, M., 2005. Gene regulatory networks. Proceedings of the National Academy of Sciences 102, 4935–4935. https://doi.org/10.1073/pnas.0502024102

de Magalhães, J.P., Finch, C.E., Janssens, G., 2010. Next-generation sequencing in aging research: Emerging applications, problems, pitfalls and possible solutions. Ageing Research Reviews 9, 315–323. https://doi.org/10.1016/j.arr.2009.10.006

de Mendoza, D., Pilon, M., 2019. Control of membrane lipid homeostasis by lipid-bilayer associated sensors: A mechanism conserved from bacteria to humans. Progress in Lipid Research 76, 100996. https://doi.org/10.1016/j.plipres.2019.100996

De Wit, M., Keil, D., Remmerie, N., Ven, K. van der, Brandhof, E.-J. van den, Knapen, D., Witters, E., Coen, W.D., 2008. Molecular targets of TBBPA in zebrafish analysed through integration of genomic and proteomic approaches. Chemosphere 74, 96–105. https://doi.org/10.1016/j.chemosphere.2008.09.030

Debnath, M., Prasad, G.B.K.S., Bisen, P.S., 2010. Molecular Diagnostics: Promises and Possibilities. Springer Science & Business Media.

DeCaprio, D., Vinson, J.P., Pearson, M.D., Montgomery, P., Doherty, M., Galagan, J.E., 2007. Conrad: gene prediction using conditional random fields. Genome Res 17, 1389–1398. https://doi.org/10.1101/gr.6558107

Dedourge-Geffard, O., Charron, L., Hofbauer, C., Gaillet, V., Palais, F., Lacaze, E., Geffard, A., Geffard, O., 2013. Temporal patterns of digestive enzyme activities and feeding rate in gammarids (Gammarus fossarum) exposed to inland polluted waters. Ecotoxicology and Environmental Safety 97, 139–146. https://doi.org/10.1016/j.ecoenv.2013.07.016

Degli Esposti, D., Almunia, C., Guery, M.-A., Koenig, N., Armengaud, J., Chaumot, A., Geffard, O., 2019. Co-expression network analysis identifies gonad- and embryo-associated protein modules in the sentinel species Gammarus fossarum. Scientific Reports 9, 7862. https://doi.org/10.1038/s41598-019-44203-5

Dennis, E.A., Deems, R.A., Harkewicz, R., Quehenberger, O., Brown, H.A., Milne, S.B., Myers, D.S., Glass, C.K., Hardiman, G., Reichart, D., Merrill, A.H., Sullards, M.C., Wang, E., Murphy, R.C., Raetz, C.R.H., Garrett, T.A., Guan, Z., Ryan, A.C., Russell, D.W., McDonald, J.G., Thompson, B.M., Shaw, W.A., Sud, M., Zhao, Y., Gupta, S., Maurya, M.R., Fahy, E., Subramaniam, S., 2010. A Mouse Macrophage Lipidome*♦. Journal of Biological Chemistry 285, 39976–39985. https://doi.org/10.1074/jbc.M110.182915

Dennis, E.A., Norris, P.C., 2015. Eicosanoid storm in infection and inflammation. Nat Rev Immunol 15, 511–523. https://doi.org/10.1038/nri3859

Dennis, G., Sherman, B.T., Hosack, D.A., Yang, J., Gao, W., Lane, H.C., Lempicki, R.A., 2003. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biology 4, P3. https://doi.org/10.1186/gb-2003-4-5-p3

D'haeseleer, P., 2005. How does gene expression clustering work? Nat Biotechnol 23, 1499–1501. https://doi.org/10.1038/nbt1205-1499

Di Genova, A., Buena-Atienza, E., Ossowski, S., Sagot, M.-F., 2021. Efficient hybrid de novo assembly of human genomes with WENGAN. Nat Biotechnol 39, 422–430. https://doi.org/10.1038/s41587-020-00747-w

Dillies, M.-A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, G., Castel, D., Estelle, J., Guernec, G., Jagla, B., Jouneau, L., Laloë, D., Le Gall, C., Schaëffer, B., Le Crom, S., Guedj, M., Jaffrézic, F., 2013. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Briefings in Bioinformatics 14, 671–683. https://doi.org/10.1093/bib/bbs046

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R., 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21. https://doi.org/10.1093/bioinformatics/bts635

Drozdova, P., Rivarola-Duarte, L., Bedulina, D., Axenov-Gribanov, D., Schreiber, S., Gurkov, A., Shatilina, Z., Vereshchagina, K., Lubyaga, Y., Madyarova, E., Otto, C., Jühling, F., Busch, W., Jakob, L., Lucassen, M., Sartoris, F.J., Hackermüller, J., Hoffmann, S., Pörtner, H.-O., Luckenbach, T., Timofeyev, M., Stadler, P.F., 2019. Comparison between transcriptomic responses to short-term stress exposures of a common Holarctic and endemic Lake Baikal amphipods. BMC Genomics 20, 712. https://doi.org/10.1186/s12864-019-6024-3

Duarte, G., Yu., P., Geras'kin, S., 2021. A Pipeline for Non-model Organisms for de novo Transcriptome Assembly, Annotation, and Gene Ontology Analysis Using Open Tools: Case Study with Scots Pine. BIO-PROTOCOL 11. https://doi.org/10.21769/BioProtoc.3912

Dührkop, K., Shen, H., Meusel, M., Rousu, J., Böcker, S., 2015. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. Proceedings of the National Academy of Sciences 112, 12580–12585. https://doi.org/10.1073/pnas.1509788112

Dumas, T., 2020. Les approches –omiques, métabolomique et protéomique, pour l'étude de la relation de cause à effet entre contaminants émergents, produits pharmaceutiques et organismes marins, Mytilus galloprovincialis (phdthesis). Université Montpellier.

Dumas, T., Courant, F., Almunia, C., Boccard, J., Rosain, D., Duporté, G., Armengaud, J., Fenet, H., Gomez, E., 2022. An integrated metabolomics and proteogenomics approach reveals molecular alterations following carbamazepine exposure in the male mussel Mytilus galloprovincialis. Chemosphere 286, 131793. https://doi.org/10.1016/j.chemosphere.2021.131793

Dunn, W.B., Ellis, David.I., 2005. Metabolomics: Current analytical platforms and methodologies. TrAC Trends in Analytical Chemistry 24, 285–294. https://doi.org/10.1016/j.trac.2004.11.021

Dunn, W.B., Erban, A., Weber, R.J.M., Creek, D.J., Brown, M., Breitling, R., Hankemeier, T., Goodacre, R., Neumann, S., Kopka, J., Viant, M.R., 2013. Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics. Metabolomics 9, 44–66. https://doi.org/10.1007/s11306-012-0434-4

Duran, R.C.D., Menon, S., Wu, J., 2016. The Analyses of Global Gene Expression and Transcription Factor Regulation, in: Wu, J. (Ed.), Transcriptomics and Gene Regulation, Translational Bioinformatics. Springer Netherlands, Dordrecht, pp. 1–35. https://doi.org/10.1007/978-94-017-7450-5_1

Dutta, A., Sinha, D.K., n.d. Zebrafish lipid droplets regulate embryonic ATP homeostasis to power early development. Open Biology 7, 170063. https://doi.org/10.1098/rsob.170063

Earl, D., Bradnam, K., St John, J., Darling, A., Lin, D., Fass, J., Yu, H.O.K., Buffalo, V., Zerbino, D.R., Diekhans, M., Nguyen, N., Ariyaratne, P.N., Sung, W.-K., Ning, Z., Haimel, M., Simpson, J.T., Fonseca, N.A., Birol, İ., Docking, T.R., Ho, I.Y., Rokhsar, D.S., Chikhi, R., Lavenier, D., Chapuis, G., Naquin, D., Maillet, N., Schatz, M.C., Kelley, D.R., Phillippy, A.M., Koren, S., Yang, S.-P., Wu, W., Chou, W.-C., Srivastava, A., Shaw, T.I., Ruby, J.G., Skewes-Cox, P., Betegon, M., Dimon, M.T., Solovyev, V., Seledtsov, I., Kosarev, P., Vorobyev, D., Ramirez-Gonzalez, R., Leggett, R., MacLean, D., Xia, F., Luo, R., Li, Z., Xie, Y., Liu, B., Gnerre, S., MacCallum, I., Przybylski, D., Ribeiro, F.J., Yin, S., Sharpe, T., Hall, G., Kersey, P.J., Durbin, R., Jackman, S.D., Chapman, J.A., Huang, X., DeRisi, J.L., Caccamo, M., Li, Y., Jaffe, D.B., Green, R.E., Haussler, D., Korf, I., Paten, B., 2011. Assemblathon 1: a competitive assessment of de novo short read assembly methods. Genome Res 21, 2224–2241. https://doi.org/10.1101/gr.126599.111

Ebner, J.N., 2021. Trends in the Application of "Omics" to Ecotoxicology and Stress Ecology. Genes 12, 1481.

Eden, E., Navon, R., Steinfeld, I., Lipson, D., Yakhini, Z., 2009. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics 10, 48. https://doi.org/10.1186/1471-2105-10-48

Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D., 1998. Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences 95, 14863–14868. https://doi.org/10.1073/pnas.95.25.14863

Eisenberg, D., Marcotte, E.M., Xenarios, I., Yeates, T.O., 2000. Protein function in the post-genomic era. Nature 405, 823–826. https://doi.org/10.1038/35015694

Ekblom, R., Wolf, J.B.W., 2014. A field guide to whole-genome sequencing, assembly and annotation. Evolutionary Applications 7, 1026–1042. https://doi.org/10.1111/eva.12178

Ellegren, H., 2014. Genome sequencing and population genomics in non-model organisms. Trends Ecol Evol 29, 51–63. https://doi.org/10.1016/j.tree.2013.09.008

Ellis, S.R., Paine, M.R.L., Eijkel, G.B., Pauling, J.K., Husen, P., Jervelund, M.W., Hermansson, M., Ejsing, C.S., Heeren, R.M.A., 2018. Automated, parallel mass spectrometry imaging and structural identification of lipids. Nat Methods 15, 515–518. https://doi.org/10.1038/s41592-018-0010-6

English, A.C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., Qin, X., Muzny, D.M., Reid, J.G., Worley, K.C., Gibbs, R.A., 2012. Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. PLOS ONE 7, e47768. https://doi.org/10.1371/journal.pone.0047768

English, A.C., Salerno, W.J., Reid, J.G., 2014. PBHoney: identifying genomic variants via long-read discordance and interrupted mapping. BMC Bioinformatics 15, 180. https://doi.org/10.1186/1471-2105-15-180

Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B., Milacic, M., Roca, C.D., Rothfels, K., Sevilla, C., Shamovsky, V., Shorser, S., Varusai, T., Viteri, G., Weiser, J., Wu, G., Stein, L., Hermjakob, H., D'Eustachio, P., 2018. The Reactome Pathway Knowledgebase. Nucleic Acids Research 46, D649–D655. https://doi.org/10.1093/nar/gkx1132

Fahy, E., Alvarez-Jarreta, J., Brasher, C.J., Nguyen, A., Hawksworth, J.I., Rodrigues, P., Meckelmann, S., Allen, S.M., O'Donnell, V.B., 2019. LipidFinder on LIPID MAPS: peak filtering, MS searching and statistical analysis for lipidomics. Bioinformatics 35, 685–687. https://doi.org/10.1093/bioinformatics/bty679

Fahy, E., Subramaniam, S., Murphy, R.C., Nishijima, M., Raetz, C.R.H., Shimizu, T., Spener, F., Meer, G. van, Wakelam, M.J.O., Dennis, E.A., 2009. Update of the LIPID MAPS comprehensive classification system for lipids1. Journal of Lipid Research 50, S9–S14. https://doi.org/10.1194/jlr.R800095-JLR200

Faugere, J., 2022. Développements analytiques pour l'analyse multi-omique en chromatographie liquide couplée à la spectrométrie de masse (These en préparation). Lyon.

Faugere, J., Gouveia, D., Ayciriex, S., Chaumot, A., Almunia, C., François, A., Armengaud, J., Lemoine, J., Geffard, O., Degli-Esposti, D., Salvador, A., 2020. High-multiplexed monitoring of protein biomarkers in the sentinel Gammarus fossarum by targeted scout-MRM assay, a new vision for ecotoxicoproteomics. Journal of Proteomics 226, 103901. https://doi.org/10.1016/j.jprot.2020.103901

Faure, D., Joly, D., 2015. Next-generation sequencing as a powerful motor for advances in the biological and environmental sciences. Genetica 143, 129–132.

Felten, V., 2003. Effets de l'acidification des ruisseaux vosgiens sur la biologie, l'écologie et l'écophysiologie de Gammarus fossarum Koch, 1835 (Crustacea Amphipoda): Approche intégrée à différents niveaux d'organisation. (PhD Thesis). Metz.

Feng, J., Zhang, Q., Zhou, Y., Yu, S., Hong, L., Zhao, S., Yang, J., Wan, H., Xu, G., Zhang, Y., Li, C., 2018. Integration of Proteomics and Metabolomics Revealed Metabolite–Protein Networks in ACTH-Secreting Pituitary Adenoma. Frontiers in Endocrinology 9.

Fialkowski, W., Fialkowska, E., Smith, B.D., Rainbow, P.S., 2003. Biomonitoring Survey of Trace Metal Pollution in Streams of a Catchment Draining a Zinc and Lead Mining Area of Upper Silesia, Poland Using the Amphipod Gammarus fossarum. International Review of Hydrobiology 88, 187–200. https://doi.org/10.1002/iroh.200390014

Fickett, J.W., Tung, C.-S., 1992. Assessment of protein coding measures. Nucleic acids research 20, 6441–6450.

Fiehn, O., 2002. Metabolomics – the link between genotypes and phenotypes. Plant Mol Biol 48, 155–171. https://doi.org/10.1023/A:1013713905833

Fischer, H.P., 2005. Towards quantitative biology: Integration of biological information to elucidate disease pathways and to guide drug discovery, in: Biotechnology Annual Review. Elsevier, pp. 1–68. https://doi.org/10.1016/S1387-2656(05)11001-1

Flicek, P., Keibler, E., Hu, P., Korf, I., Brent, M.R., 2003. Leveraging the Mouse Genome for Gene Prediction in Human: From Whole-Genome Shotgun Reads to a Global Synteny Map. Genome Res. 13, 46–54. https://doi.org/10.1101/gr.830003

Fliser, D., Novak, J., Thongboonkerd, V., Argiles, A., Jankowski, V., Girolami, M.A., Jankowski, J., Mischak, H., 2007. Advances in urinary proteome analysis and biomarker discovery. Journal of the American Society of Nephrology 18, 1057–1071. https://doi.org/10.1681/ASN.2006090956

Florea, L., Hartzell, G., Zhang, Z., Rubin, G.M., Miller, W., 1998. A computer program for aligning a cDNA sequence with a genomic DNA sequence. Genome research 8, 967–974.

Florens, L., Washburn, M.P., Raine, J.D., Anthony, R.M., Grainger, M., Haynes, J.D., Moch, J.K., Muster, N., Sacci, J.B., Tabb, D.L., Witney, A.A., Wolters, D., Wu, Y., Gardner, M.J., Holder, A.A., Sinden, R.E., Yates, J.R., Carucci, D.J., 2002. A proteomic view of the Plasmodium falciparum life cycle. Nature 419, 520–526. https://doi.org/10.1038/nature01107

Fonseca, N.A., Rung, J., Brazma, A., Marioni, J.C., 2012. Tools for mapping high-throughput sequencing data. Bioinformatics 28, 3169–3177. https://doi.org/10.1093/bioinformatics/bts605

Förster, J., Famili, I., Fu, P., Palsson, B.Ø., Nielsen, J., 2003. Genome-Scale Reconstruction of the Saccharomyces cerevisiae Metabolic Network. Genome Res. 13, 244–253. https://doi.org/10.1101/gr.234503

Fournier, T., 2022. Les apports du séquençage long read en génomique 29.

Freedman, A.H., Clamp, M., Sackton, T.B., 2021. Error, noise and bias in de novo transcriptome assemblies. Molecular Ecology Resources 21, 18–29. https://doi.org/10.1111/1755-0998.13156

Fu, T., Knittelfelder, O., Geffard, O., Clement, Y., Testet, E., Elie, N., Touboul, D., Abbaci, K., Shevchenko, A., Lemoine, J., Chaumot, A., Salvador, A., Degli-Esposti, D., Ayciriex, S., 2021. Shotgun lipidomics and mass spectrometry imaging unveil diversity and dynamics in Gammarus fossarum lipid composition. iScience 24, 102115. https://doi.org/10.1016/j.isci.2021.102115

Fu, T., Oetjen, J., Chapelle, M., Verdu, A., Szesny, M., Chaumot, A., Degli-Esposti, D., Geffard, O., Clément, Y., Salvador, A., Ayciriex, S., 2020. In situ isobaric lipid mapping by MALDI–ion mobility separation–mass spectrometry imaging. Journal of Mass Spectrometry 55, e4531. https://doi.org/10.1002/jms.4531

Fuertes, I., Jordão, R., Casas, J., Barata, C., 2018. Allocation of glycerolipids and glycerophospholipids from adults to eggs in Daphnia magna: Perturbations by compounds that enhance lipid droplet accumulation. Environ Pollut 242, 1702–1710. https://doi.org/10.1016/j.envpol.2018.07.102

Fuertes, I., Jordão, R., Piña, B., Barata, C., 2019. Time-dependent transcriptomic responses of Daphnia magna exposed to metabolic disruptors that enhanced storage lipid accumulation. Environmental Pollution 249, 99–108. https://doi.org/10.1016/j.envpol.2019.02.102

Fuertes, I., Piña, B., Barata, C., 2020. Changes in lipid profiles in Daphnia magna individuals exposed to low environmental levels of neuroactive pharmaceuticals. Science of The Total Environment 733, 139029. https://doi.org/10.1016/j.scitotenv.2020.139029

Fuller, C.W., Middendorf, L.R., Benner, S.A., Church, G.M., Harris, T., Huang, X., Jovanovich, S.B., Nelson, J.R., Schloss, J.A., Schwartz, D.C., Vezenov, D.V., 2009. The challenges of sequencing by synthesis. Nat Biotechnol 27, 1013–1023. https://doi.org/10.1038/nbt.1585

Gao, L., Pei, G., Chen, L., Zhang, W., 2015. A global network-based protocol for functional inference of hypothetical proteins in Synechocystis sp. PCC 6803. Journal of Microbiological Methods 116, 44–52. https://doi.org/10.1016/j.mimet.2015.06.013

García-Alcalde, F., García-López, F., Dopazo, J., Conesa, A., 2011. Paintomics: a web based tool for the joint visualization of transcriptomics and metabolomics data. Bioinformatics 27, 137–139. https://doi.org/10.1093/bioinformatics/btq594

Garcia-Reyero, N., Perkins, E.J., 2011. Systems biology: Leading the revolution in ecotoxicology. Environmental Toxicology and Chemistry 30, 265–273. https://doi.org/10.1002/etc.401

Ge, H., Walhout, A.J.M., Vidal, M., 2003. Integrating "omic" information: a bridge between genomics and systems biology. Trends Genet 19, 551–560. https://doi.org/10.1016/j.tig.2003.08.009

Geffard, O., 2014. Proposition d'une espèce bio-indicatrice, Gammarus fossarum : Démarches, méthodes et outils pour diagnostiquer et comprendre la contamination chimique et la toxicité des milieux aquatiques (thesis). Habilitation à Diriger des Recherches, Université Lyon I.

Geffard, O., Xuereb, B., Chaumot, A., Geffard, A., Biagianti, S., Noël, C., Abbaci, K., Garric, J., Charmantier, G., Charmantier-Daures, M., 2010. Ovarian cycle and embryonic development in Gammarus fossarum: Application for reproductive toxicity assessment. Environmental Toxicology and Chemistry 29, 2249–2259. https://doi.org/10.1002/etc.268

Gene Ontology, C., 2015. Gene ontology consortium: going forward. vol. 43, pp. D1049-56. Nucleic Acids Res.

Gerhardt, A., 2011. GamTox: A Low-Cost Multimetric Ecotoxicity Test with Gammarus spp. for In and Ex Situ Application. International Journal of Zoology 2011, e574536. https://doi.org/10.1155/2011/574536

Gestin, O., Lacoue-Labarthe, T., Coquery, M., Delorme, N., Garnero, L., Dherret, L., Ciccia, T., Geffard, O., Lopes, C., 2021. One and multi-compartments toxico-kinetic modeling to understand metals' organotropism and fate in Gammarus fossarum. Environment International 156, 106625.

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, 2731–2741. https://doi.org/10.1002/pmic.201400265

GHCZ00000000.1 Gammarus fossarum female :: NCBI [WWW Document], 2018. URL https://www.ncbi.nlm.nih.gov/Traces/wgs/?val=GHCZ01 (accessed 12.16.21).

GHDA00000000.1 Gammarus fossarum male :: NCBI [WWW Document], 2018. URL https://www.ncbi.nlm.nih.gov/Traces/wgs/?val=GHDA01 (accessed 12.16.21).

Gilbert, L.I., O'Connor, J.D., 1970. CHAPTER 8 - Lipid Metabolism and Transport in Arthropods11Work of the authors and their colleagues cited in this article was supported by grant AM-02818 from the National Institutes of Health, in: Florkin, M.,

Scheer, B.T. (Eds.), Chemical Zoology. Academic Press, pp. 229–253. https://doi.org/10.1016/B978-0-12-395538-8.50035-5

Giraudo, M., Douville, M., Houde, M., 2015. Chronic toxicity evaluation of the flame retardant tris (2-butoxyethyl) phosphate (TBOEP) using Daphnia magna transcriptomic response. Chemosphere 132, 159–165. https://doi.org/10.1016/j.chemosphere.2015.03.028

Gismondi, E., Beisel, J.-N., Cossu-Leguille, C., 2012. Influence of gender and season on reduced glutathione concentration and energy reserves of Gammarus roeseli. Environmental Research 118, 47–52. https://doi.org/10.1016/j.envres.2012.06.004

Gismondi, E., Mazzucchelli, G., De Pauw, E., Joaquim-Justo, C., Thomé, J.P., 2015. Gender differences in responses in Gammarus pulex exposed to BDE-47: A gel-free proteomic approach. Ecotoxicology and Environmental Safety 122, 205–213. https://doi.org/10.1016/j.ecoenv.2015.07.038

Gismondi, E., Thomé, J.P., 2016. Transcriptome of the freshwater amphipod Gammarus pulex hepatopancreas. Genomics Data 8, 91–92. https://doi.org/10.1016/j.gdata.2016.04.002

Gismondi, E., Thomé, J.-P., Urien, N., Uher, E., Baiwir, D., Mazzucchelli, G., De Pauw, E., Fechner, L.C., Lebrun, J.D., 2017. Ecotoxicoproteomic assessment of the functional alterations caused by chronic metallic exposures in gammarids. Environmental Pollution 225, 428–438. https://doi.org/10.1016/j.envpol.2017.03.006

Glass, K., Huttenhower, C., Quackenbush, J., Yuan, G.-C., 2013. Passing Messages between Biological Networks to Refine Predicted Interactions. PLOS ONE 8, e64832. https://doi.org/10.1371/journal.pone.0064832

Gnerre, S., Lander, E.S., Lindblad-Toh, K., Jaffe, D.B., 2009. Assisted assembly: how to improve a de novo genome assembly by using related species. Genome Biol 10, R88. https://doi.org/10.1186/gb-2009-10-8-r88

Goel, N., Singh, S., Aseri, T.C., 2013. A Review of Soft Computing Techniques for Gene Prediction. ISRN Genomics 2013, e191206. https://doi.org/10.1155/2013/191206

Gómez-Canela, C., Miller, T.H., Bury, N.R., Tauler, R., Barron, L.P., 2016. Targeted metabolomics of Gammarus pulex following controlled exposures to selected pharmaceuticals in water. Science of The Total Environment 562, 777–788. https://doi.org/10.1016/j.scitotenv.2016.03.181

Gong, P., Perkins, E.J., 2016. Earthworm toxicogenomics: A renewed genome-wide quest for novel biomarkers and mechanistic insights. Applied Soil Ecology, ISEE-10: The 10th International Symposium on Earthworm Ecology, 22-27 June 2014, Athens, Georgia, USA 104, 12–24. https://doi.org/10.1016/j.apsoil.2015.11.005

Gonzàlez-Porta, M., Frankish, A., Rung, J., Harrow, J., Brazma, A., 2013. Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. Genome Biol 14, R70. https://doi.org/10.1186/gb-2013-14-7-r70

Gorrochategui, E., Jaumot, J., Lacorte, S., Tauler, R., 2016. Data analysis strategies for targeted and untargeted LC-MS metabolomic studies: Overview and workflow. TrAC Trends in Analytical Chemistry 82, 425–442. https://doi.org/10.1016/j.trac.2016.07.004

Gouveia, D., 2017. Approches moléculaires pour la découverte, le développement et l'application de biomarqueurs de toxicité chez les gammaridéscité spécifiques applicables au sein de la diversité des gammaridés (These de doctorat). Lyon.

Gouveia, D., Almunia, C., Cogne, Y., Pible, O., Degli-Esposti, D., Salvador, A., Cristobal, S., Sheehan, D., Chaumot, A., Geffard, O., 2019a. Ecotoxicoproteomics: A decade of progress in our understanding of anthropogenic impact on the environment. Journal of proteomics 198, 66–77.

Gouveia, D., Almunia, C., Cogne, Y., Pible, O., Degli-Esposti, D., Salvador, A., Cristobal, S., Sheehan, D., Chaumot, A., Geffard, O., Armengaud, J., 2019b. Ecotoxicoproteomics: A decade of progress in our understanding of anthropogenic impact on the environment. Journal of Proteomics, 10 Year Anniversary of Proteomics 198, 66–77. https://doi.org/hh

Gouveia, Duarte, Chaumot, A., Charnot, A., Almunia, C., François, A., Navarro, L., Armengaud, J., Salvador, A., 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, 13417–13426. https://doi.org/10.1021/acs.est.7b03736

Gouveia, D., Chaumot, A., Charnot, A., Queau, H., Armengaud, J., Almunia, C., Salvador, A., 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. Aquatic Toxicology 190, 199–209. https://doi.org/10.1016/j.aquatox.2017.07.007

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., 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29, 644–652. https://doi.org/10.1038/nbt.1883

Granholm, V., 2014. The accuracy of statistical confidence estimates in shotgun proteomics.

Graves, P.R., Haystead, T.A.J., 2002. Molecular biologist's guide to proteomics. Microbiology and Molecular Biology Reviews 66, 39–63. https://doi.org/10.1128/MMBR.66.1.39-63.2002

Graw, S., Chappell, K., Washam, C.L., Gies, A., Bird, J., Robeson, M.S., Byrum, S.D., 2021. Multi-omics data integration considerations and study design for biological systems and disease. Mol. Omics 17, 170–185. https://doi.org/10.1039/D0MO00041H

Gregori, J., Sánchez, A., Villanueva, J., 2013. msmsTests: LC-MS/MS Differential Expression Tests. R package version 1.14. 0.

Grossmann, J., Fernández, H., Chaubey, P.M., Valdés, A.E., Gagliardini, V., Cañal, M.J., Russo, G., Grossniklaus, U., 2017. Proteogenomic Analysis Greatly Expands the Identification of Proteins Related to Reproduction in the Apogamous Fern Dryopteris affinis ssp. affinis. Frontiers in Plant Science 8.

Guarascio, M., Manco, G., Ritacco, E., 2019. Network Models, in: Ranganathan, S., Gribskov, M., Nakai, K., Schönbach, C. (Eds.), Encyclopedia of Bioinformatics and Computational Biology. Academic Press, Oxford, pp. 968–977. https://doi.org/10.1016/B978-0-12-809633-8.20427-5

Guillot, L., Delage, L., Viari, A., Vandenbrouck, Y., Com, E., Ritter, A., Lavigne, R., Marie, D., Peterlongo, P., Potin, P., Pineau, C., 2019. Peptimapper: proteogenomics workflow for the expert annotation of eukaryotic genomes. BMC Genomics 20, 56. https://doi.org/10.1186/s12864-019-5431-9

Gupta, S., Ellis, S.E., Ashar, F.N., Moes, A., Bader, J.S., Zhan, J., West, A.B., Arking, D.E., 2014. Transcriptome analysis reveals dysregulation of innate immune response genes and neuronal activity-dependent genes in autism. Nat Commun 5, 5748. https://doi.org/10.1038/ncomms6748

Guttman, M., Donaghey, J., Carey, B.W., Garber, M., Grenier, J.K., Munson, G., Young, G., Lucas, A.B., Ach, R., Bruhn, L., Yang, X., Amit, I., Meissner, A., Regev, A., Rinn, J.L., Root, D.E., Lander, E.S., 2011. lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature 477, 295–300. https://doi.org/10.1038/nature10398

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., MacManes, M.D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C.N., Henschel, R., LeDuc, R.D., Friedman, N., Regev, A., 2013. De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. Nat Protoc 8, 10.1038/nprot.2013.084. https://doi.org/10.1038/nprot.2013.084

Habibi, I., Emamian, E.S., Abdi, A., 2014. Quantitative analysis of intracellular communication and signaling errors in signaling networks. BMC Syst Biol 8, 89. https://doi.org/10.1186/s12918-014-0089-z

Haiminen, N., Kuhn, D.N., Parida, L., Rigoutsos, I., 2011. Evaluation of methods for de novo genome assembly from high-throughput sequencing reads reveals dependencies that affect the quality of the results. PLoS One 6, e24182. https://doi.org/10.1371/journal.pone.0024182

Ham, J., Lim, W., You, S., Song, G., 2020. Butylated hydroxyanisole induces testicular dysfunction in mouse testis cells by dysregulating calcium homeostasis and stimulating endoplasmic reticulum stress. Sci Total Environ 702, 134775. https://doi.org/10.1016/j.scitotenv.2019.134775

Han, X., Gross, R.W., 2003. Global analyses of cellular lipidomes directly from crude extracts of biological samples by ESI mass spectrometry: a bridge to lipidomics. Journal of Lipid Research 44, 1071–1079. https://doi.org/10.1194/jlr.R300004-JLR200

Hanemaaijer, M., Olivier, B.G., Röling, W.F.M., Bruggeman, F.J., Teusink, B., 2017. Model-based quantification of metabolic interactions from dynamic microbial-community data. PLOS ONE 12, e0173183. https://doi.org/10.1371/journal.pone.0173183

Hangauer, M.J., Vaughn, I.W., McManus, M.T., 2013. Pervasive Transcription of the Human Genome Produces Thousands of Previously Unidentified Long Intergenic Noncoding RNAs. PLOS Genetics 9, e1003569. https://doi.org/10.1371/journal.pgen.1003569

Harayama, T., Riezman, H., 2018. Understanding the diversity of membrane lipid composition. Nat Rev Mol Cell Biol 19, 281–296. https://doi.org/10.1038/nrm.2017.138

Hardcastle, T.J., Kelly, K.A., 2010. baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics 11, 422. https://doi.org/10.1186/1471-2105-11-422

Hasin, Y., Seldin, M., Lusis, A., 2017. Multi-omics approaches to disease. Genome Biol 18, 83. https://doi.org/10.1186/s13059-017-1215-1

Hastings, J., Owen, G., Dekker, A., Ennis, M., Kale, N., Muthukrishnan, V., Turner, S., Swainston, N., Mendes, P., Steinbeck, C., 2016. ChEBI in 2016: Improved services and an expanding collection of metabolites. Nucleic Acids Research 44, D1214–D1219. https://doi.org/10.1093/nar/gkv1031

Hawkins, C., Ginzburg, D., Zhao, K., Dwyer, W., Xue, B., Xu, A., Rice, S., Cole, B., Paley, S., Karp, P., Rhee, S.Y., 2021. Plant Metabolic Network 15: A resource of genome-wide metabolism databases for 126 plants and algae. Journal of Integrative Plant Biology 63, 1888–1905. https://doi.org/10.1111/jipb.13163

Heather, J.M., Chain, B., 2016. The sequence of sequencers: The history of sequencing DNA. Genomics 107, 1–8. https://doi.org/10.1016/j.ygeno.2015.11.003

Hernandez, D., François, P., Farinelli, L., Osterås, M., Schrenzel, J., 2008. De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. Genome Res 18, 802–809. https://doi.org/10.1101/gr.072033.107

Hira, Z.M., Gillies, D.F., 2015. A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data. Adv Bioinformatics 2015, 198363. https://doi.org/10.1155/2015/198363

Hissa, B., Pontes, B., 2018. Role of Membrane Cholesterol in Modulating Actin Architecture and Cellular Contractility, Cholesterol - Good, Bad and the Heart. IntechOpen. https://doi.org/10.5772/intechopen.76532

Horvath, S., Dong, J., 2008. Geometric interpretation of gene coexpression network analysis. PLoS Comput Biol 4, e1000117. https://doi.org/10.1371/journal.pcbi.1000117

Hu, L., Ye, M., Jiang, X., Feng, S., Zou, H., 2007. Advances in hyphenated analytical techniques for shotgun proteome and peptidome analysis—A review. Analytica Chimica Acta 598, 193–204. https://doi.org/10.1016/j.aca.2007.07.046

Hu, Z., Snitkin, E.S., DeLisi, C., 2008. VisANT: an integrative framework for networks in systems biology. Briefings in Bioinformatics 9, 317–325. https://doi.org/10.1093/bib/bbn020

Huan, T., Li, L., 2015. Counting Missing Values in a Metabolite-Intensity Data Set for Measuring the Analytical Performance of a Metabolomics Platform. Anal. Chem. 87, 1306–1313. https://doi.org/10.1021/ac5039994

Huang, D.W., Sherman, B.T., Lempicki, R.A., 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4, 44–57. https://doi.org/10.1038/nprot.2008.211

Huang, S., Wang, J., Yue, W., Chen, J., Gaughan, S., Lu, W., Lu, G., Wang, C., 2015. Transcriptomic variation of hepatopancreas reveals the energy metabolism and biological processes associated with molting in Chinese mitten crab, Eriocheir sinensis. Sci Rep 5, 14015. https://doi.org/10.1038/srep14015

Huerlimann, R., Wade, N.M., Gordon, L., Montenegro, J.D., Goodall, J., McWilliam, S., Tinning, M., Siemering, K., Giardina, E., Donovan, D., Sellars, M.J., Cowley, J.A., Condon, K., Coman, G.J., Khatkar, M.S., Raadsma, H.W., Maes, G.E., Zenger, K.R., Jerry, D.R., 2018. De novo assembly, characterization, functional annotation and expression patterns of the black tiger shrimp (Penaeus monodon) transcriptome. Sci Rep 8, 13553. https://doi.org/10.1038/s41598-018-31148-4

Ison, J., Rapacki, K., Ménager, H., Kalaš, M., Rydza, E., Chmura, P., Anthon, C., Beard, N., Berka, K., Bolser, D., Booth, T., Bretaudeau, A., Brezovsky, J., Casadio, R., Cesareni, G., Coppens, F., Cornell, M., Cuccuru, G., Davidsen, K., Vedova, G.D., Dogan, T., Doppelt-Azeroual, O., Emery, L., Gasteiger, E., Gatter, T., Goldberg, T., Grosjean, M., Grüning, B., Helmer-Citterich, M., Ienasescu, H., Ioannidis, V., Jespersen, M.C., Jimenez, R., Juty, N., Juvan, P., Koch, M., Laibe, C., Li, J.-W., Licata, L., Mareuil, F., Mičetić, I., Friborg, R.M., Moretti, S., Morris, C., Möller, S., Nenadic, A., Peterson, H., Profiti, G., Rice, P., Romano, P., Roncaglia, P., Saidi, R., Schafferhans, A., Schwämmle, V., Smith, C., Sperotto, M.M., Stockinger, H., Vařeková, R.S., Tosatto, S.C.E., de la Torre, V., Uva, P., Via, A., Yachdav, G., Zambelli, F., Vriend, G., Rost, B., Parkinson, H., Løngreen, P., Brunak, S., 2016. Tools and data services registry: a community effort to document bioinformatics resources. Nucleic Acids Research 44, D38–D47. https://doi.org/10.1093/nar/gkv1116

Istvan, E.S., Deisenhofer, J., 2001. Structural Mechanism for Statin Inhibition of HMG-CoA Reductase. Science 292, 1160–1164. https://doi.org/10.1126/science.1059344

Jaffe, J.D., Berg, H.C., Church, G.M., 2004. Proteogenomic mapping as a complementary method to perform genome annotation. Proteomics 4, 59–77.

Jaspard, E., 2008. Génomique fonctionnelle et protéomique : introduction. URL http://biochimej.univ-angers.fr/Page2/COURS/9ModulGenFoncVeg/3Introduction/1IntroGenFonc.htm

Jauhal, A.A., Newcomb, R.D., 2021. Assessing genome assembly quality prior to downstream analysis: N50 versus BUSCO. Molecular Ecology Resources 21, 1416–1421. https://doi.org/10.1111/1755-0998.13364

Jeck, W.R., Reinhardt, J.A., Baltrus, D.A., Hickenbotham, M.T., Magrini, V., Mardis, E.R., Dangl, J.L., Jones, C.D., 2007. Extending assembly of short DNA sequences to handle error. Bioinformatics 23, 2942–2944. https://doi.org/10.1093/bioinformatics/btm451

Jelic, A., Gros, M., Ginebreda, A., Cespedes-Sánchez, R., Ventura, F., Petrovic, M., Barcelo, D., 2011. Occurrence, partition and removal of pharmaceuticals in sewage water and sludge during wastewater treatment. Water Research 45, 1165–1176. https://doi.org/10.1016/j.watres.2010.11.010

Jewison, T., Su, Y., Disfany, F.M., Liang, Y., Knox, C., Maciejewski, A., Poelzer, J., Huynh, J., Zhou, Y., Arndt, D., Djoumbou, Y., Liu, Y., Deng, L., Guo, A.C., Han, B., Pon, A., Wilson, M., Rafatnia, S., Liu, P., Wishart, D.S., 2014. SMPDB 2.0: Big Improvements to the Small Molecule Pathway Database. Nucleic Acids Research 42, D478–D484. https://doi.org/10.1093/nar/gkt1067

Ji, C., Li, F., Wang, Q., Zhao, J., Sun, Z., Wu, H., 2016. An integrated proteomic and metabolomic study on the gender-specific responses of mussels Mytilus galloprovincialis to tetrabromobisphenol A (TBBPA). Chemosphere 144, 527–539. https://doi.org/10.1016/j.chemosphere.2015.08.052

Jiang, J., Jacob, H.J., 1998. EbEST: an automated tool using expressed sequence tags to delineate gene structure. Genome research 8, 268–275.

Jiménez-Prada, P., Hachero-Cruzado, I., Guerra-García, J.M., 2021. Aquaculture waste as food for amphipods: the case of Gammarus insensibilis in marsh ponds from southern Spain. Aquacult Int 29, 139–153. https://doi.org/10.1007/s10499-020-00615-z

Jin, S., Bian, C., Jiang, S., Sun, S., Xu, L., Xiong, Y., Qiao, H., Zhang, W., You, X., Li, J., Gong, Y., Ma, B., Shi, Q., Fu, H., 2019. Identification of Candidate Genes for the Plateau Adaptation of a Tibetan Amphipod, Gammarus lacustris, Through Integration of Genome and Transcriptome Sequencing. Frontiers in Genetics 10. https://doi.org/10.3389/fgene.2019.00053

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics 30, 1236–1240.

Jord, ão R., Casas, J., Fabrias, G., Campos, B., Pi, ña B., Lemos, M.F.L., Soares, A.M.V.M., Tauler, R., Barata, C., 2015. Obesogens beyond Vertebrates: Lipid Perturbation by Tributyltin in the Crustacean Daphnia magna. Environmental Health Perspectives 123, 813–819. https://doi.org/10.1289/ehp.1409163

Jordão, R., Campos, B., Piña, B., Tauler, R., Soares, A.M.V.M., Barata, C., 2016a. Mechanisms of Action of Compounds That Enhance Storage Lipid Accumulation in Daphnia magna. Environmental Science & Technology 50, 13565–13573. https://doi.org/10.1021/acs.est.6b04768

Jordão, R., Garreta, E., Campos, B., Lemos, M.F.L., Soares, A.M.V.M., Tauler, R., Barata, C., 2016b. Compounds altering fat storage in Daphnia magna. Sci Total Environ 545–546, 127–136. https://doi.org/10.1016/j.scitotenv.2015.12.097

Jordão Rita, Casas Josefina, Fabrias Gemma, Campos Bruno, Piña Benjamín, Lemos Marco F.L., Soares Amadeu M.V.M., Tauler Romà, Barata Carlos, 2015. Obesogens beyond Vertebrates: Lipid Perturbation by Tributyltin in the Crustacean Daphnia magna. Environmental Health Perspectives 123, 813–819. https://doi.org/10.1289/ehp.1409163

Jouany, J.M., 1971. Nuisances et ecologie. Actualités Pharmaceutiques 69, 11–22.

Jubeaux, G., Simon, R., Salvador, A., Quéau, H., Chaumot, A., Geffard, O., 2012. 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. Aquatic Toxicology 112–113, 72–82. https://doi.org/10.1016/j.aquatox.2012.01.011

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.A.A., Ballard, A.J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A.W., Kavukcuoglu, K., Kohli, P., Hassabis, D., 2021. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kalari, K.R., Nair, A.A., Bhavsar, J.D., O'Brien, D.R., Davila, J.I., Bockol, M.A., Nie, J., Tang, X., Baheti, S., Doughty, J.B., Middha, S., Sicotte, H., Thompson, A.E., Asmann, Y.W., Kocher, J.-P.A., 2014. MAP-RSeq: Mayo Analysis Pipeline for RNA sequencing. BMC Bioinformatics 15, 224. https://doi.org/10.1186/1471-2105-15-224

Kamburov, A., Cavill, R., Ebbels, T.M.D., Herwig, R., Keun, H.C., 2011. Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA. Bioinformatics 27, 2917–2918. https://doi.org/10.1093/bioinformatics/btr499

Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M., Tanabe, M., 2021. KEGG: integrating viruses and cellular organisms. Nucleic Acids Research 49, D545–D551. https://doi.org/10.1093/nar/gkaa970

Kanehisa, M., Goto, S., 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 28, 27–30. https://doi.org/10.1093/nar/28.1.27

Kao, D., Lai, A.G., Stamataki, E., Rosic, S., Konstantinides, N., Jarvis, E., Di Donfrancesco, A., Pouchkina-Stancheva, N., Semon, M., Grillo, M., 2016. The genome of the crustacean Parhyale hawaiensis, a model for animal development, regeneration, immunity and lignocellulose digestion. elife 5, e20062.

Karaman, G., Pinkster, S., 1977. Freshwater Gammarus Species from Europe, North Africa and Adjacent Regions of Asia (Crustacea-Amphipoda) in: Bijdragen tot de Dierkunde Volume 47 Issue 2 (1977). https://doi.org/10.1163/26660644-04702003

Karnovsky, A., Weymouth, T., Hull, T., Tarcea, V.G., Scardoni, G., Laudanna, C., Sartor, M.A., Stringer, K.A., Jagadish, H.V., Burant, C., Athey, B., Omenn, G.S., 2012. Metscape 2 bioinformatics tool for the analysis and visualization of metabolomics and gene expression data. Bioinformatics 28, 373–380. https://doi.org/10.1093/bioinformatics/btr661

Karp, P.D., Midford, P.E., Billington, R., Kothari, A., Krummenacker, M., Latendresse, M., Ong, W.K., Subhraveti, P., Caspi, R., Fulcher, C., Keseler, I.M., Paley, S.M., 2021. Pathway Tools version 23.0 update: software for pathway/genome informatics and systems biology. Briefings in Bioinformatics 22, 109–126. https://doi.org/10.1093/bib/bbz104

Karp, P.D., Ouzounis, C.A., Moore-Kochlacs, C., Goldovsky, L., Kaipa, P., Ahrén, D., Tsoka, S., Darzentas, N., Kunin, V., López-Bigas, N., 2005. Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. Nucleic Acids Research 33, 6083–6089. https://doi.org/10.1093/nar/gki892

Karp, P.D., Paley, S.M., Krummenacker, M., Latendresse, M., Dale, J.M., Lee, T.J., Kaipa, P., Gilham, F., Spaulding, A., Popescu, L., 2010. Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. Briefings in bioinformatics 11, 40–79.

Karpievitch, Y.V., Dabney, A.R., Smith, R.D., 2012. Normalization and missing value imputation for label-free LC-MS analysis. BMC Bioinformatics 13 Suppl 16, S5. https://doi.org/10.1186/1471-2105-13-S16-S5

Kendall, R.J., Anderson, T.A., Baker, R.J., Bens, C.M., Carr, J.A., Chiodo, L.A., Cobb III, G.P., Dickerson, R.L., Dixon, K.R., Frame, L.T., 2001. Ecotoxicology. USDA National Wildlife Research Center-Staff Publications 516.

Khatri, P., Drăghici, S., 2005. Ontological analysis of gene expression data: current tools, limitations, and open problems. Bioinformatics 21, 3587–3595. https://doi.org/10.1093/bioinformatics/bti565

Khatri, P., Sirota, M., Butte, A.J., 2012. Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. PLOS Computational Biology 8, e1002375. https://doi.org/10.1371/journal.pcbi.1002375

Kim, M., Rai, N., Zorraquino, V., Tagkopoulos, I., 2016. Multi-omics integration accurately predicts cellular state in unexplored conditions for Escherichia coli. Nat Commun 7, 13090. https://doi.org/10.1038/ncomms13090

Kishimoto, K., Urade, R., Ogawa, T., Moriyama, T., 2001. Nondestructive Quantification of Neutral Lipids by Thin-Layer Chromatography and Laser-Fluorescent Scanning: Suitable Methods for "Lipidome" Analysis. Biochemical and Biophysical Research Communications 281, 657–662. https://doi.org/10.1006/bbrc.2001.4404

Klose, C., Surma, M.A., Gerl, M.J., Meyenhofer, F., Shevchenko, A., Simons, K., 2012. Flexibility of a Eukaryotic Lipidome – Insights from Yeast Lipidomics. PLOS ONE 7, e35063. https://doi.org/10.1371/journal.pone.0035063

Knittelfelder, O., Prince, E., Sales, S., Fritzsche, E., Woehner, T., Brankatschk, M., Shevchenko, A., 2020. Sterols as dietary markers for Drosophila melanogaster. Biochim. Biophys. Acta Mol. Cell Biol. Lipids 1865, 158683. https://doi.org/10.1016/j.bbalip.2020.158683

Koenig, N., Almunia, C., Bonnal-Conduzorgues, A., Armengaud, J., Chaumot, A., Geffard, O., Degli Esposti, D., 2021. Co-expression network analysis identifies novel molecular pathways associated with cadmium and pyriproxyfen testicular toxicity in Gammarus fossarum. Aquatic Toxicology 235, 105816. https://doi.org/10.1016/j.aquatox.2021.105816

Kolanowski, W., Stolyhwo, A., Grabowski, M., 2007. Fatty Acid Composition of Selected Fresh Water Gammarids (Amphipoda, Crustacea): A Potentially Innovative Source of Omega-3 LC PUFA. Journal of the American Oil Chemists' Society 84, 827–833. https://doi.org/10.1007/s11746-007-1116-7

Konschak, M., Zubrod, J.P., Baudy, P., Fink, P., Kenngott, K.G.J., Englert, D., Röder, N., Ogbeide, C., Schulz, R., Bundschuh, M., 2021. Chronic effects of the strobilurin fungicide azoxystrobin in the leaf shredder Gammarus fossarum (Crustacea; Amphipoda) via two effect pathways. Ecotoxicology and Environmental Safety 209, 111848. https://doi.org/10.1016/j.ecoenv.2020.111848

Korf, I., 2004. Gene finding in novel genomes. BMC Bioinformatics 5, 59. https://doi.org/10.1186/1471-2105-5-59

Kouassi Nzoughet, J., Bocca, C., Simard, G., Prunier-Mirebeau, D., Chao de la Barca, J.M., Bonneau, D., Procaccio, V., Prunier, F., Lenaers, G., Reynier, P., 2017. A Nontargeted UHPLC-HRMS Metabolomics Pipeline for Metabolite Identification: Application to Cardiac Remote Ischemic Preconditioning. Anal. Chem. 89, 2138–2146. https://doi.org/10.1021/acs.analchem.6b04912

Kriventseva, E.V., Kuznetsov, D., Tegenfeldt, F., Manni, M., Dias, R., Simão, F.A., Zdobnov, E.M., 2019. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. Nucleic Acids Research 47, D807–D811. https://doi.org/10.1093/nar/gky1053

Kühmayer, T., Guo, F., Ebm, N., Battin, T.J., Brett, M.T., Bunn, S.E., Fry, B., Kainz, M.J., 2020. Preferential retention of algal carbon in benthic invertebrates: Stable isotope and fatty acid evidence from an outdoor flume experiment. Freshwater Biology 65, 1200–1209. https://doi.org/10.1111/fwb.13492

Kukurba, K.R., Montgomery, S.B., 2015. RNA Sequencing and Analysis. Cold Spring Harb Protoc 2015, 951–969. https://doi.org/10.1101/pdb.top084970

Kuleshov, M.V., Jones, M.R., Rouillard, A.D., Fernandez, N.F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S.L., Jagodnik, K.M., Lachmann, A., McDermott, M.G., Monteiro, C.D., Gundersen, G.W., Ma'ayan, A., 2016. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res 44, W90–W97. https://doi.org/10.1093/nar/gkw377

Kunin, V., Ahren, D., Goldovsky, L., Janssen, P., Ouzounis, C.A., 2005. Measuring genome conservation across taxa: divided strains and united kingdoms. Nucleic Acids Res 33, 616–621. https://doi.org/10.1093/nar/gki181

Kunowska, N., Rotival, M., Yu, L., Choudhary, J., Dillon, N., 2015. Identification of protein complexes that bind to histone H3 combinatorial modifications using super-SILAC and weighted correlation network analysis. Nucleic Acids Research 43, 1418–1432. https://doi.org/10.1093/nar/gku1350

Kunz, P., Kienle, C., Gerhardt, A., 2010. Gammarus spp. in Aquatic Ecotoxicology and Water Quality Assessment: Toward Integrated Multilevel Tests. Reviews of environmental contamination and toxicology 205, 1–76. https://doi.org/10.1007/978-1-4419-5623-1_1

Kuo, T.-C., Tian, T.-F., Tseng, Y.J., 2013. 3Omics: a web-based systems biology tool for analysis, integration and visualization of human transcriptomic, proteomic and metabolomic data. BMC Syst Biol 7, 64. https://doi.org/10.1186/1752-0509-7-64

Kutmon, M., Evelo, C.T., Coort, S.L., 2014. A network biology workflow to study transcriptomics data of the diabetic liver. BMC Genomics 15, 971. https://doi.org/10.1186/1471-2164-15-971

Kyle, J.E., Aimo, L., Bridge, A.J., Clair, G., Fedorova, M., Helms, J.B., Molenaar, M.R., Ni, Z., Orešič, M., Slenter, D., Willighagen, E., Webb-Robertson, B.-J.M., 2021. Interpreting the lipidome: bioinformatic approaches to embrace the complexity. Metabolomics 17, 55. https://doi.org/10.1007/s11306-021-01802-6

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. Environmental Research 111, 626–634. https://doi.org/10.1016/j.envres.2011.03.012

Lam, S.M., Wang, Z., Li, B., Shui, G., 2021. High-coverage lipidomics for functional lipid and pathway analyses. Analytica Chimica Acta 1147, 199–210. https://doi.org/10.1016/j.aca.2020.11.024

Lane, L., Argoud-Puy, G., Britan, A., Cusin, I., Duek, P.D., Evalet, O., Gateau, A., Gaudet, P., Gleizes, A., Masselot, A., Zwahlen, C., Bairoch, A., 2012. neXtProt: a knowledge platform for human proteins. Nucleic Acids Research 40, D76–D83. https://doi.org/10.1093/nar/gkr1179

Langfelder, P., Horvath, S., 2008. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9, 559. https://doi.org/10.1186/1471-2105-9-559

Langfelder, P., Luo, R., Oldham, M.C., Horvath, S., 2011. Is my network module preserved and reproducible? PLoS Comput Biol 7, e1001057. https://doi.org/10.1371/journal.pcbi.1001057

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9, 357–359. https://doi.org/10.1038/nmeth.1923

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

Laudicella, V.A., Whitfield, P.D., Carboni, S., Doherty, M.K., Hughes, A.D., 2020. Application of lipidomics in bivalve aquaculture, a review. Reviews in Aquaculture 12, 678–702. https://doi.org/10.1111/raq.12346

Lazar, C., Gatto, L., Ferro, M., Bruley, C., Burger, T., 2016. Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies. https://doi.org/10.1021/acs.jproteome.5b00981

Le Manach, S., Sotton, B., Huet, H., Duval, C., Paris, A., Marie, A., Yépremian, C., Catherine, A., Mathéron, L., Vinh, J., Edery, M., Marie, B., 2018. Physiological effects caused by microcystin-producing and non-microcystin producing Microcystis aeruginosa on medaka fish: A proteomic and metabolomic study on liver. Environmental Pollution 234, 523–537. https://doi.org/10.1016/j.envpol.2017.11.011

Lehtinen, S., Marsellach, F.X., Codlin, S., Schmidt, A., Clément-Ziza, M., Beyer, A., Bähler, J., Orengo, C., Pancaldi, V., 2013. Stress induces remodelling of yeast interaction and co-expression networks. Mol. BioSyst. 9, 1697–1707. https://doi.org/10.1039/C3MB25548D

Leng, N., Dawson, J.A., Thomson, J.A., Ruotti, V., Rissman, A.I., Smits, B.M.G., Haag, J.D., Gould, M.N., Stewart, R.M., Kendziorski, C., 2013. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. Bioinformatics 29, 1035–1043. https://doi.org/10.1093/bioinformatics/btt087

Lepretre, M., 2019. Caractérisation protéomique par spectrométrie de masse du compartiment hémolymphatique de la moule zébrée Dreissena polymorpha. (These de doctorat). Reims.

Lepretre, M., Palos-Ladeiro, M., Faugere, J., Almunia, C., Lemoine, J., Armengaud, J., Geffard, A., Salvador, A., 2020. From shotgun to targeted proteomics: rapid Scout-MRM assay development for monitoring potential immunomarkers in Dreissena polymorpha. Anal. Bioanal. Chem. 412, 7333–7347. https://doi.org/10.1007/s00216-020-02868-2

Leroy, D., Haubruge, E., De Pauw, E., Thomé, J.-P., Francis, F., 2010. Development of ecotoxicoproteomics on the freshwater amphipod Gammarus pulex: identification of PCB biomarkers in glycolysis and glutamate pathways. Ecotoxicology and environmental safety 73, 343–352.

Leung, K.M., 2018. Joining the dots between omics and environmental management. Integrated Environmental Assessment and Management 14, 169–173. https://doi.org/10.1002/ieam.2007

Lewin, H.A., Richards, S., Lieberman Aiden, E., Allende, M.L., Archibald, J.M., Bálint, M., Barker, K.B., Baumgartner, B., Belov, K., Bertorelle, G., Blaxter, M.L., Cai, J., Caperello, N.D., Carlson, K., Castilla-Rubio, J.C., Chaw, S.-M., Chen, L., Childers, A.K., Coddington, J.A., Conde, D.A., Corominas, M., Crandall, K.A., Crawford, A.J., DiPalma, F., Durbin, R., Ebenezer, T.E., Edwards, S.V., Fedrigo, O., Flicek, P., Formenti, G., Gibbs, R.A., Gilbert, M.T.P., Goldstein, M.M., Graves, J.M., Greely, H.T., Grigoriev, I.V., Hackett, K.J., Hall, N., Haussler, D., Helgen, K.M., Hogg, C.J., Isobe, S., Jakobsen, K.S., Janke, A., Jarvis, E.D., Johnson, W.E., Jones, S.J.M., Karlsson, E.K., Kersey, P.J., Kim, J.-H., Kress, W.J., Kuraku, S., Lawniczak, M.K.N., Leebens-Mack, J.H., Li, X., Lindblad-Toh, K., Liu, X., Lopez, J.V., Marques-Bonet, T., Mazard, S., Mazet, J.A.K., Mazzoni, C.J., Myers, E.W., O'Neill, R.J., Paez, S., Park, H., Robinson, G.E., Roquet, C., Ryder, O.A., Sabir, J.S.M., Shaffer, H.B., Shank, T.M., Sherkow, J.S., Soltis, P.S., Tang, B., Tedersoo, L., Uliano-Silva, M., Wang, K., Wei, X., Wetzer, R., Wilson, J.L., Xu, X., Yang, H., Yoder, A.D., Zhang, G., 2022. The Earth BioGenome Project 2020: Starting the clock. Proceedings of the National Academy of Sciences 119, e2115635118. https://doi.org/10.1073/pnas.2115635118

Lewin, H.A., Robinson, G.E., Kress, W.J., Baker, W.J., Coddington, J., Crandall, K.A., Durbin, R., Edwards, S.V., Forest, F., Gilbert, M.T.P., Goldstein, M.M., Grigoriev, I.V., Hackett, K.J., Haussler, D., Jarvis, E.D., Johnson, W.E., Patrinos, A., Richards, S., Castilla-Rubio, J.C., van Sluys, M.-A., Soltis, P.S., Xu, X., Yang, H., Zhang, G., 2018. Earth BioGenome Project: Sequencing life for the future of life. Proceedings of the National Academy of Sciences 115, 4325–4333. https://doi.org/10.1073/pnas.1720115115

Li, B., Dewey, C.N., 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12, 323. https://doi.org/10.1186/1471-2105-12-323

Li, B., Fillmore, N., Bai, Y., Collins, M., Thomson, J.A., Stewart, R., Dewey, C.N., 2014. Evaluation of de novo transcriptome assemblies from RNA-Seq data. Genome Biology 15, 553. https://doi.org/10.1186/s13059-014-0553-5

Li, H., 2012. Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly. Bioinformatics 28, 1838–1844. https://doi.org/10.1093/bioinformatics/bts280

Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, J., Gao, L., Zhu, B.-B., Lin, Z.-J., Chen, J., Lu, X., Wang, H., Zhang, C., Chen, Y.-H., Xu, D.-X., 2020. Long-term 1-nitropyrene exposure induces endoplasmic reticulum stress and inhibits steroidogenesis in mice testes. Chemosphere 251, 126336. https://doi.org/10.1016/j.chemosphere.2020.126336

Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., Huang, Q., Cai, Q., Li, B., Bai, Y., Zhang, Zhihe, Zhang, Y., Wang, W., Li, Jun, Wei, F., Li, H., Jian, M., Li, Jianwen, Zhang, Zhaolei, Nielsen, R., Li, Dawei, Gu, W., Yang, Z., Xuan, Z., Ryder, O.A., Leung, F.C.-C., Zhou, Y., Cao, J., Sun, X., Fu, Y., Fang, X., Guo, X., Wang, B., Hou, R., Shen, F., Mu, B., Ni, P., Lin, R., Qian, W., Wang, G., Yu, C., Nie, W., Wang, Jinhuan, Wu, Z., Liang, H., Min, J., Wu, Q., Cheng, S., Ruan, J., Wang, M., Shi, Z., Wen, M., Liu, B., Ren, X., Zheng, Huisong, Dong, D., Cook, K., Shan, G., Zhang, Hao, Kosiol, C., Xie, X., Lu, Z., Zheng, Hancheng, Li, Y., Steiner, C.C., Lam, T.T.-Y., Lin, S., Zhang, Q., Li, G., Tian, J., Gong, T., Liu, H., Zhang, D., Fang, L., Ye, C., Zhang, J., Hu, W., Xu, A., Ren, Y., Zhang, G., Bruford, M.W., Li, Q., Ma, L., Guo, Y., An, N., Hu, Y., Zheng, Y., Shi, Y., Li, Z., Liu, Q., Chen, Y., Zhao, J., Qu, N., Zhao, S., Tian, F., Wang, Xiaoling, Wang, H., Xu, L., Liu, X., Vinar, T., Wang, Y., Lam, T.-W., Yiu, S.-M., Liu, S., Zhang, Hemin, Li, Desheng, Huang, Y., Wang, Xia, Yang, G., Jiang, Z., Wang, Junyi, Qin, N., Li, L., Li, Jingxiang, Bolund, L., Kristiansen, K., Wong, G.K.-S., Olson, M., Zhang, X., Li, S., Yang, H., Wang, Jian, Wang, Jun, 2010. The sequence and de novo assembly of the giant panda genome. Nature 463, 311–317. https://doi.org/10.1038/nature08696

Li, Y.-R., Yang, W.-X., 2016. Myosin superfamily: The multi-functional and irreplaceable factors in spermatogenesis and testicular tumors. Gene 576, 195–207. https://doi.org/10.1016/j.gene.2015.10.022

Liang, X., Martyniuk, C.J., Simmons, D.B.D., 2020. Are we forgetting the "proteomics" in multi-omics ecotoxicology? Comparative Biochemistry and Physiology Part D: Genomics and Proteomics 36, 100751. https://doi.org/10.1016/j.cbd.2020.100751

Licata, L., Briganti, L., Peluso, D., Perfetto, L., Iannuccelli, M., Galeota, E., Sacco, F., Palma, A., Nardozza, A.P., Santonico, E., Castagnoli, L., Cesareni, G., 2012. MINT, the molecular interaction database: 2012 update. Nucleic Acids Research 40, D857–D861. https://doi.org/10.1093/nar/gkr930

Lin, D., Tabb, D.L., Yates, J.R., 2003. Large-scale protein identification using mass spectrometry. Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics 1646, 1–10. https://doi.org/10.1016/S1570-9639(02)00546-0

Lin, Y., Li, J., Shen, H., Zhang, L., Papasian, C.J., Deng, H.-W., 2011. Comparative studies of de novo assembly tools for next-generation sequencing technologies. Bioinformatics 27, 2031–2037. https://doi.org/10.1093/bioinformatics/btr319

Ling, W., Dong-Mei, F., 2009. Estimation of Missing Values Using a Weighted K-Nearest Neighbors Algorithm, in: 2009 International Conference on Environmental Science and Information Application Technology. Presented at the 2009 International Conference on Environmental Science and Information Application Technology, pp. 660–663. https://doi.org/10.1109/ESIAT.2009.206

Lipaeva, P., Vereshchagina, K., Drozdova, P., Jakob, L., Kondrateva, E., Lucassen, M., Bedulina, D., Timofeyev, M., Stadler, P., Luckenbach, T., 2021. Different ways to play it cool: Transcriptomic analysis sheds light on different activity patterns of three amphipod species under long-term cold exposure. Molecular Ecology 30, 5735–5751. https://doi.org/10.1111/mec.16164

Liu, H., Sadygov, R.G., Yates, J.R., 2004. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. Anal Chem 76, 4193–4201. https://doi.org/10.1021/ac0498563

Lobov, A.A., Maltseva, A.L., Mikhailova, N.A., Granovitch, A.I., 2019. The Molecular Mechanisms of Gametic Incompatibility in Invertebrates. Acta Naturae 11, 4–15. https://doi.org/10.32607/20758251-2019-11-3-4-15

Lopez-Casado, G., Covey, P.A., Bedinger, P.A., Mueller, L.A., Thannhauser, T.W., Zhang, S., Fei, Z., Giovannoni, J.J., Rose, J.K.C., 2012. Enabling proteomic studies with RNA-Seq: The proteome of tomato pollen as a test case. PROTEOMICS 12, 761–774. https://doi.org/10.1002/pmic.201100164

Love, M.I., Huber, W., Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology 15, 550. https://doi.org/10.1186/s13059-014-0550-8

Lowe, R., Shirley, N., Bleackley, M., Dolan, S., Shafee, T., 2017. Transcriptomics technologies. PLoS Comput Biol 13, e1005457. https://doi.org/10.1371/journal.pcbi.1005457

Luo, H., Gao, F., Lin, Y., 2015. Evolutionary conservation analysis between the essential and nonessential genes in bacterial genomes. Sci Rep 5, 13210. https://doi.org/10.1038/srep13210

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Yunjie, Tang, J., Wu, G., Zhang, H., Shi, Y., Liu, Yong, Yu, C., Wang, B., Lu, Y., Han, C., Cheung, D.W., Yiu, S.-M., Peng, S., Xiaoqian, Z., Liu, G., Liao, X., Li, Y., Yang, H., Wang, Jian, Lam, T.-W., Wang, Jun, 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience 1, 18. https://doi.org/10.1186/2047-217X-1-18

Lv, B., Peng, Yong, Peng, Yuan-de, Wang, Z., Song, Q., 2022. Integrated transcriptomics and proteomics provide new insights into the cadmium-induced ovarian toxicity on Pardosa pseudoannulata. Chemosphere 297, 134255. https://doi.org/10.1016/j.chemosphere.2022.134255

Lynn, D.J., Winsor, G.L., Chan, C., Richard, N., Laird, M.R., Barsky, A., Gardy, J.L., Roche, F.M., Chan, T.H., Shah, N., 2008. InnateDB: facilitating systems-level analyses of the mammalian innate immune response. Molecular systems biology 4, 218.

Ma, H., Sorokin, A., Mazein, A., Selkov, A., Selkov, E., Demin, O., Goryanin, I., 2007. The Edinburgh human metabolic network reconstruction and its functional analysis. Molecular systems biology 3, 135.

Machado, D., Andrejev, S., Tramontano, M., Patil, K.R., 2018. Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. Nucleic Acids Research 46, 7542–7553. https://doi.org/10.1093/nar/gky537

Macneil, C., Dick, J.T.A., Elwood, R.W., 1999. The dynamics of predation on Gammarus spp. (Crustacea: Amphipoda). Biological Reviews 74, 375–395. https://doi.org/10.1017/S0006323199005368

Maere, S., Heymans, K., Kuiper, M., 2005. BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. Bioinformatics 21, 3448–3449. https://doi.org/10.1093/bioinformatics/bti551

Majoros, W.H., Pertea, M., Salzberg, S.L., 2004. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics 20, 2878–2879. https://doi.org/10.1093/bioinformatics/bth315

Maltby, L., 1995. Sensitivity of the crustaceans Gammarus pulex (L.) and Asellus aquaticus (L.) to short-term exposure to hypoxia and unionized ammonia: Observations and possible mechanisms. Water Research 29, 781–787. https://doi.org/10.1016/0043-1354(94)00231-U

Manni, M., Berkeley, M.R., Seppey, M., Simão, F.A., Zdobnov, E.M., 2021. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Molecular Biology and Evolution 38, 4647–4654. https://doi.org/10.1093/molbev/msab199

Mardis, E.R., 2008. The impact of next-generation sequencing technology on genetics. Trends in Genetics 24, 133–141. https://doi.org/10.1016/j.tig.2007.12.007

Maretty, L., Sibbesen, J.A., Krogh, A., 2014. Bayesian transcriptome assembly. Genome Biology 15, 501. https://doi.org/10.1186/s13059-014-0501-4

Marie, B., 2020. Disentangling of the ecotoxicological signal using "omics" analyses, a lesson from the survey of the impact of cyanobacterial proliferations on fishes. Science of The Total Environment 736, 139701. https://doi.org/10.1016/j.scitotenv.2020.139701

Martens, M., Ammar, A., Riutta, A., Waagmeester, A., Slenter, D.N., Hanspers, K., A. Miller, R., Digles, D., Lopes, E.N., Ehrhart, F., Dupuis, L.J., Winckers, L.A., Coort, S.L., Willighagen, E.L., Evelo, C.T., Pico, A.R., Kutmon, M., 2021. WikiPathways: connecting communities. Nucleic Acids Research 49, D613–D621. https://doi.org/10.1093/nar/gkaa1024

Martin, J.W., Davis, G.E., 2001. An updated classification of the recent Crustacea. Natural History Museum of Los Angeles County Los Angeles.

Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17, 10–12. https://doi.org/10.14806/ej.17.1.200

Martyniuk, C.J., 2018. Are we closer to the vision? A proposed framework for incorporating omics into environmental assessments. Environ Toxicol Pharmacol 59, 87–93. https://doi.org/10.1016/j.etap.2018.03.005

Masseroli, M., Galati, O., Pinciroli, F., 2005. GFINDer: genetic disease and phenotype location statistical analysis and mining of dynamically annotated gene lists. Nucleic Acids Research 33, W717–W723. https://doi.org/10.1093/nar/gki454

Mathé, C., Sagot, M., Schiex, T., Rouzé, P., 2002. Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Research 30, 4103–4117. https://doi.org/10.1093/nar/gkf543

McManus, C.J., Graveley, B.R., 2011. RNA structure and the mechanisms of alternative splicing. Current Opinion in Genetics & Development, Differentiation and gene regulation 21, 373–379. https://doi.org/10.1016/j.gde.2011.04.001

McPartland, J.M., Agraval, J., Gleeson, D., Heasman, K., Glass, M., 2006. Cannabinoid receptors in invertebrates. Journal of Evolutionary Biology 19, 366–373. https://doi.org/10.1111/j.1420-9101.2005.01028.x

McQuilton, P., St. Pierre, S.E., Thurmond, J., the FlyBase Consortium, 2012. FlyBase 101 – the basics of navigating FlyBase. Nucleic Acids Research 40, D706–D714. https://doi.org/10.1093/nar/gkr1030

Meader, S., Hillier, L.W., Locke, D., Ponting, C.P., Lunter, G., 2010. Genome assembly quality: assessment and improvement using the neutral indel model. Genome Res 20, 675–684. https://doi.org/10.1101/gr.096966.109

Médigue, C., Bocs, S., Labarre, L., Mathé, C., Vallenet, D., 2002. L'annotation in silico des séquences génomiques - Bio-informatique (1). Med Sci (Paris) 18, 237–250. https://doi.org/10.1051/medsci/2002182237

Mendoza, S.N., Olivier, B.G., Molenaar, D., Teusink, B., 2019. A systematic assessment of current genome-scale metabolic reconstruction tools. Genome Biology 20, 158. https://doi.org/10.1186/s13059-019-1769-1

Menschaert, G., Fenyö, D., 2017. Proteogenomics from a bioinformatics angle: A growing field. Mass Spectrometry Reviews 36, 584–599. https://doi.org/10.1002/mas.21483

Mercier, J., 2017. Logique paracohérente pour l'annotation fonctionnelle des génomes au travers de réseaux biologiques.

Meusy, J.J., 1980. Vitellogenin, the extraovarian precursor of the protein yolk in Crustacea: a review. Reproduction Nutrition Développement 20, 1–21.

Miller, J.R., Koren, S., Sutton, G., 2010. Assembly algorithms for next-generation sequencing data. Genomics 95, 315–327. https://doi.org/10.1016/j.ygeno.2010.03.001

Misra, B.B., Langefeld, C., Olivier, M., Cox, L.A., 2019. Integrated omics: tools, advances and future approaches. Journal of Molecular Endocrinology 62, R21–R45. https://doi.org/10.1530/JME-18-0055

Misra, B.B., Mohapatra, S., 2019. Tools and resources for metabolomics research community: A 2017–2018 update. ELECTROPHORESIS 40, 227–246. https://doi.org/10.1002/elps.201800428

Molenaar, M.R., Jeucken, A., Wassenaar, T.A., van de Lest, C.H.A., Brouwers, J.F., Helms, J.B., 2019. LION/web: a web-based ontology enrichment tool for lipidomic data analysis. GigaScience 8, giz061. https://doi.org/10.1093/gigascience/giz061

Morandin, C., Tin, M.M.Y., Abril, S., Gómez, C., Pontieri, L., Schiøtt, M., Sundström, L., Tsuji, K., Pedersen, J.S., Helanterä, H., Mikheyev, A.S., 2016. Comparative transcriptomics reveals the conserved building blocks involved in parallel evolution of diverse phenotypic traits in ants. Genome Biology 17, 43. https://doi.org/10.1186/s13059-016-0902-7

Moreno, P., Beisken, S., Harsha, B., Muthukrishnan, V., Tudose, I., Dekker, A., Dornfeldt, S., Taruttis, F., Grosse, I., Hastings, J., Neumann, S., Steinbeck, C., 2015. BiNChE: A web tool and library for chemical enrichment analysis based on the ChEBI ontology. BMC Bioinformatics 16, 56. https://doi.org/10.1186/s12859-015-0486-3

Moreno, Y., Gros, P.-P., Tam, M., Segura, M., Valanparambil, R., Geary, T.G., Stevenson, M.M., 2011. Proteomic Analysis of Excretory-Secretory Products of Heligmosomoides polygyrus Assessed with Next-Generation Sequencing Transcriptomic Information. PLOS Neglected Tropical Diseases 5, e1370. https://doi.org/10.1371/journal.pntd.0001370

Morgat, A., Lombardot, T., Coudert, E., Axelsen, K., Neto, T.B., Gehant, S., Bansal, P., Bolleman, J., Gasteiger, E., de Castro, E., Baratin, D., Pozzato, M., Xenarios, I., Poux, S., Redaschi, N., Bridge, A., The UniProt Consortium, 2020. Enzyme annotation in UniProtKB using Rhea. Bioinformatics 36, 1896–1901. https://doi.org/10.1093/bioinformatics/btz817

Moriarty, F., 1983. The study of pollutants in ecosystems. Ecotoxicolgy. Academic Pres 289.

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A.C., Kanehisa, M., 2007. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic acids research 35, W182–W185.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., Wold, B., 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5, 621–628. https://doi.org/10.1038/nmeth.1226

Myers, E.W., Sutton, G.G., Delcher, A.L., Dew, I.M., Fasulo, D.P., Flanigan, M.J., Kravitz, S.A., Mobarry, C.M., Reinert, K.H.J., Remington, K.A., Anson, E.L., Bolanos, R.A., Chou, H.-H., Jordan, C.M., Halpern, A.L., Lonardi, S., Beasley, E.M., Brandon, R.C., Chen, L., Dunn, P.J., Lai, Z., Liang, Y., Nusskern, D.R., Zhan, M., Zhang, Q., Zheng, X., Rubin, G.M., Adams, M.D., Venter, J.C., 2000. A Whole-Genome Assembly of Drosophila. Science 287, 2196–2204. https://doi.org/10.1126/science.287.5461.2196

Nachtomy, O., Shavit, A., Yakhini, Z., 2007. Gene expression and the concept of the phenotype. Stud Hist Philos Biol Biomed Sci 38, 238–254. https://doi.org/10.1016/j.shpsc.2006.12.014

Nagan, N., Zoeller, R.A., 2001. Plasmalogens: biosynthesis and functions. Progress in Lipid Research 40, 199–229. https://doi.org/10.1016/S0163-7827(01)00003-0

Nagarajan, N., Pop, M., 2013. Sequence assembly demystified. Nat Rev Genet 14, 157–167. https://doi.org/ekb

Nagarajan, N., Read, T.D., Pop, M., 2008. Scaffolding and validation of bacterial genome assemblies using optical restriction maps. Bioinformatics 24, 1229–1235. https://doi.org/10.1093/bioinformatics/btn102

Narzisi, G., Mishra, B., 2011. Comparing de novo genome assembly: the long and short of it. PLoS One 6, e19175. https://doi.org/10.1371/journal.pone.0019175

National Research Council, 2011. Nutrient requirements of fish and shrimp. National academies press.

Naumenko, S.A., Logacheva, M.D., Popova, N.V., Klepikova, A.V., Penin, A.A., Bazykin, G.A., Etingova, A.E., Mugue, N.S., Kondrashov, A.S., Yampolsky, L.Y., 2017. Transcriptome-based phylogeny of endemic Lake Baikal amphipod species flock: fast speciation accompanied by frequent episodes of positive selection. Molecular Ecology 26, 536–553. https://doi.org/10.1111/mec.13927

Navas-Iglesias, N., Carrasco-Pancorbo, A., Cuadros-Rodríguez, L., 2009. From lipids analysis towards lipidomics, a new challenge for the analytical chemistry of the 21st century. Part II: Analytical lipidomics. TrAC Trends in Analytical Chemistry 28, 393–403. https://doi.org/10.1016/j.trac.2008.12.004

NCBI, 2022. Genome List [WWW Document]. URL https://www.ncbi.nlm.nih.gov/genome/browse/#!/eukaryotes/ (accessed 3.23.22).

Nesvizhskii, A.I., 2014. Proteogenomics: concepts, applications and computational strategies. Nat Methods 11, 1114–1125. https://doi.org/10.1038/nmeth.3144

Nesvizhskii, A.I., 2007. Protein identification by tandem mass spectrometry and sequence database searching. Methods Mol Biol 367, 87–119. https://doi.org/10.1385/1-59745-275-0:87

Neuparth, T., Machado, A.M., Montes, R., Rodil, R., Barros, S., Alves, N., Ruivo, R., Castro, L.F.C., Quintana, J.B., Santos, M.M., 2020. Transgenerational inheritance of chemical-induced signature: A case study with simvastatin. Environment International 144, 106020. https://doi.org/10.1016/j.envint.2020.106020

Neuparth, T., Martins, C., Santos, C.B. de los, Costa, M.H., Martins, I., Costa, P.M., Santos, M.M., 2014. Hypocholesterolaemic pharmaceutical simvastatin disrupts reproduction and population growth of the amphipod Gammarus locusta at the ng/L range. Aquatic Toxicology 155, 337–347. https://doi.org/10.1016/j.aquatox.2014.07.009

Newman, M.C., Zhao, Y., 2008. Ecotoxicology Nomenclature: LC, LD, LOC, LOEC, MAC.

Nguyen, T.V., Alfaro, A.C., 2020. Applications of omics to investigate responses of bivalve haemocytes to pathogen infections and environmental stress. Aquaculture 518, 734488. https://doi.org/10.1016/j.aquaculture.2019.734488

Nguyen, T.V., Alfaro, A.C., Merien, F., 2019. Omics approaches to investigate host–pathogen interactions in mass mortality outbreaks of Crassostrea gigas. Reviews in Aquaculture 11, 1308–1324. https://doi.org/10.1111/raq.12294

Ni, Z., Angelidou, G., Hoffmann, R., Fedorova, M., 2017a. LPPtiger software for lipidome-specific prediction and identification of oxidized phospholipids from LC-MS datasets. Sci Rep 7, 15138. https://doi.org/10.1038/s41598-017-15363-z

Ni, Z., Angelidou, G., Lange, M., Hoffmann, R., Fedorova, M., 2017b. LipidHunter Identifies Phospholipids by High-Throughput Processing of LC-MS and Shotgun Lipidomics Datasets. Anal Chem 89, 8800–8807. https://doi.org/10.1021/acs.analchem.7b01126

Ni, Z., Goracci, L., Cruciani, G., Fedorova, M., 2019. Computational solutions in redox lipidomics – Current strategies and future perspectives. Free Radical Biology and Medicine, Redox lipidomics and adductomics Advanced analytical strategies to study oxidzed lipids and lipid-protein adducts 144, 110–123. https://doi.org/10.1016/j.freeradbiomed.2019.04.027

Nicholson, J.K., Lindon, J.C., Holmes, E., 1999. "Metabonomics": understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. Xenobiotica 29, 1181–1189. https://doi.org/10.1080/004982599238047

Nie, L., Wu, G., Culley, D.E., Scholten, J.C.M., Zhang, W., 2007. Integrative Analysis of Transcriptomic and Proteomic Data: Challenges, Solutions and Applications. Critical Reviews in Biotechnology 27, 63–75. https://doi.org/10.1080/07388550701334212

Nijhout, H., 1994. Insect Hormones (Princeton, NJ: Princeton Uni- todes: themes and variations. Trends Genet 17, 206–213.

Nishimura, D., 2001. BioCarta. Biotech Software & Internet Report 2, 117–120. https://doi.org/10.1089/152791601750294344

Nordberg, H., Cantor, M., Dusheyko, S., Hua, S., Poliakov, A., Shabalov, I., Smirnova, T., Grigoriev, I.V., Dubchak, I., 2014. The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. Nucleic Acids Research 42, D26–D31. https://doi.org/10.1093/nar/gkt1069

Nowoshilow, S., Schloissnig, S., Fei, J.-F., Dahl, A., Pang, A.W.C., Pippel, M., Winkler, S., Hastie, A.R., Young, G., Roscito, J.G., Falcon, F., Knapp, D., Powell, S., Cruz, A., Cao, H., Habermann, B., Hiller, M., Tanaka, E.M., Myers, E.W., 2018. The axolotl genome and the evolution of key tissue formation regulators. Nature 554, 50–55. https://doi.org/10.1038/nature25458

Obeid, L.M., Linardic, C.M., Karolak, L.A., Hannun, Y.A., 1993. Programmed Cell Death Induced by Ceramide. Science 259, 1769–1771. https://doi.org/10.1126/science.8456305

O'Donnell, V.B., Dennis, E.A., Wakelam, M.J.O., Subramaniam, S., 2019. LIPID MAPS: Serving the next generation of lipid researchers with tools, resources, data, and training. Science Signaling 12, eaaw2964. https://doi.org/10.1126/scisignal.aaw2964

O'Donnell, V.B., Ekroos, K., Liebisch, G., Wakelam, M., 2020. Lipidomics: Current state of the art in a fast moving field. WIREs Systems Biology and Medicine 12, e1466. https://doi.org/10.1002/wsbm.1466

O'Neil, S.T., Emrich, S.J., 2013. Assessing De Novo transcriptome assembly metrics for consistency and utility. BMC Genomics 14, 465. https://doi.org/10.1186/1471-2164-14-465

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N.H., Chavali, G., Chen, C., del-Toro, N., Duesbury, M., Dumousseau, M., Galeota, E., Hinz, U., Iannuccelli, M., Jagannathan, S., Jimenez, R., Khadake, J., Lagreid, A., Licata, L., Lovering, R.C., Meldal, B., Melidoni, A.N., Milagros, M., Peluso, D., Perfetto, L., Porras, P., Raghunath, A., Ricard-Blum, S., Roechert, B., Stutz, A., Tognolli, M., van Roey, K., Cesareni, G., Hermjakob, H., 2014. The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res 42, D358-363. https://doi.org/10.1093/nar/gkt1115

Orsini, L., Brown, J.B., Shams Solari, O., Li, D., He, S., Podicheti, R., Stoiber, M.H., Spanier, K.I., Gilbert, D., Jansen, M., Rusch, D.B., Pfrender, M.E., Colbourne, J.K., Frilander, M.J., Kvist, J., Decaestecker, E., De Schamphelaere, K.A.C., De Meester, L., 2018. Early transcriptional response pathways in Daphnia magna are coordinated in networks of crustacean-specific genes. Molecular Ecology 27, 886–897. https://doi.org/10.1111/mec.14261

O'Shea, K., Misra, B.B., 2020. Software tools, databases and resources in metabolomics: updates from 2018 to 2019. Metabolomics 16, 36. https://doi.org/10.1007/s11306-020-01657-3

Oshlack, A., Wakefield, M.J., 2009. Transcript length bias in RNA-seq data confounds systems biology. Biol Direct 4, 14. https://doi.org/10.1186/1745-6150-4-14

Palm, W., Sampaio, J.L., Brankatschk, M., Carvalho, M., Mahmoud, A., Shevchenko, A., Eaton, S., 2012. Lipoproteins in Drosophila melanogaster—Assembly, Function, and Influence on Tissue Lipid Composition. PLOS Genetics 8, e1002828. https://doi.org/10.1371/journal.pgen.1002828

Palsson, B.Ø., 2006. Systems Biology: Properties of Reconstructed Networks. Cambridge University Press.

Pan, K.-H., Lih, C.-J., Cohen, S.N., 2005. Effects of threshold choice on biological conclusions reached during analysis of gene expression by DNA microarrays. Proceedings of the National Academy of Sciences 102, 8961–8965. https://doi.org/10.1073/pnas.0502674102

Patti, G.J., Yanes, O., Siuzdak, G., 2012. Metabolomics: the apogee of the omic triology. Nat Rev Mol Cell Biol 13, 263–269. https://doi.org/10.1038/nrm3314

Pearson, W.R., Lipman, D.J., 1988. Improved tools for biological sequence comparison. Proceedings of the National Academy of Sciences 85, 2444–2448.

Pedersen, A.G., Nielsen, H., 1997. Neural network prediction of translation initiation sites in eukaryotes: perspectives for EST and genome analysis., in: Ismb. Citeseer, pp. 226–233.

Peeters, E.T.H.M., Gardeniers, P, J.J., 1998. Logistic regression as a tool for defining habitat requirements of two common gammarids. Freshwater Biology 39, 605–615. https://doi.org/10.1046/j.1365-2427.1998.00304.x

Pei, G., Chen, L., Wang, J., Qiao, J., Zhang, W., 2014. Protein network signatures associated with exogenous biofuels treatments in cyanobacterium Synechocystis sp. PCC 6803. Frontiers in Bioengineering and Biotechnology 2. https://doi.org/10.3389/fbioe.2014.00048

Pei, G., Chen, L., Zhang, W., 2017. WGCNA Application to Proteomic and Metabolomic Data Analysis. Meth. Enzymol. 585, 135–158. https://doi.org/10.1016/bs.mie.2016.09.016

Peng, Y., Leung, H.C.M., Yiu, S.-M., Lv, M.-J., Zhu, X.-G., Chin, F.Y.L., 2013. IDBA-tran: a more robust de novo de Bruijn graph assembler for transcriptomes with uneven expression levels. Bioinformatics 29, i326–i334. https://doi.org/10.1093/bioinformatics/btt219

Pettus, B.J., Chalfant, C.E., Hannun, Y.A., 2002. Ceramide in apoptosis: an overview and current perspectives. Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids 1585, 114–125. https://doi.org/10.1016/S1388-1981(02)00331-1

Pevzner, P.A., Tang, H., Waterman, M.S., 2001. An Eulerian path approach to DNA fragment assembly. Proc Natl Acad Sci U S A 98, 9748–9753. https://doi.org/10.1073/pnas.171285098

Pinu, F.R., Beale, D.J., Paten, A.M., Kouremenos, K., Swarup, S., Schirra, H.J., Wishart, D., 2019. Systems Biology and Multi-Omics Integration: Viewpoints from the Metabolomics Research Community. Metabolites 9, 76. https://doi.org/10.3390/metabo9040076

Piscart, C., Bollache, L., 2012. Crustacés amphipodes de surface (Gammares d'eau douce), Introduction Pratique à la Systématique des Organismes des Eaux Continentales Françaises. Association française de limnologie.

Piwowar, M., Jurkowski, W., 2015. ONION: Functional Approach for Integration of Lipidomics and Transcriptomics Data. PLOS ONE 10, e0128854. https://doi.org/10.1371/journal.pone.0128854

Plaistow, S.J., Bollache, L., Cézilly, F., 2003. Energetically costly precopulatory mate guarding in the amphipod Gammarus pulex: causes and consequences. Animal behaviour 65, 683–691.

Poynton, H.C., Hasenbein, S., Benoit, J.B., Sepulveda, M.S., Poelchau, M.F., Hughes, D.S.T., Murali, S.C., Chen, S., Glastad, K.M., Goodisman, M.A.D., Werren, J.H., Vineis, J.H., Bowen, J.L., Friedrich, M., Jones, J., Robertson, H.M., Feyereisen, R., Mechler-Hickson, A., Mathers, N., Lee, C.E., Colbourne, J.K., Biales, A., Johnston, J.S., Wellborn, G.A., Rosendale, A.J., Cridge, A.G., Munoz-Torres, M.C., Bain, P.A., Manny, A.R., Major, K.M., Lambert, F.N., Vulpe, C.D., Tuck, P., Blalock, B.J., Lin, Y.-Y., Smith, M.E., Ochoa-Acuña, H., Chen, M.-J.M., Childers, C.P., Qu, J., Dugan, S., Lee, S.L., Chao, H., Dinh, H., Han, Y., Doddapaneni, H., Worley, K.C., Muzny, D.M., Gibbs, R.A., Richards, S., 2018. The Toxicogenome of Hyalella azteca: A Model for Sediment Ecotoxicology and Evolutionary Toxicology. Environmental Science & Technology 52, 6009–6022. https://doi.org/10.1021/acs.est.8b00837

Prat, O., Degli-Esposti, D., 2019. 6 - New Challenges: Omics Technologies in Ecotoxicology, in: Gross, E., Garric, J. (Eds.), Ecotoxicology. Elsevier, pp. 181–208. https://doi.org/10.1016/B978-1-78548-314-1.50006-7

Prestridge, D.S., 1995. Predicting Pol II promoter sequences using transcription factor binding sites. Journal of molecular biology 249, 923–932.

Provost-Javier, K.N., Chen, S., Rasgon, J.L., 2010. Vitellogenin gene expression in autogenous Culex tarsalis. Insect Molecular Biology 19, 423–429. https://doi.org/10.1111/j.1365-2583.2010.00999.x

Raghavan, V., Kraft, L., Mesny, F., Rigerte, L., 2022. A simple guide to de novo transcriptome assembly and annotation. Briefings in Bioinformatics 23, bbab563. https://doi.org/10.1093/bib/bbab563

Rahmatbakhsh, M., Gagarinova, A., Babu, M., 2021. Bioinformatic Analysis of Temporal and Spatial Proteome Alternations During Infections. Frontiers in Genetics 12.

Ralston-Hooper, K.J., Turner, M.E., Soderblom, E.J., Villeneuve, D., Ankley, G.T., Moseley, M.A., Hoke, R.A., Ferguson, P.L., 2013. Application of a Label-free, Gel-free Quantitative Proteomics Method for Ecotoxicological Studies of Small Fish Species. Environ. Sci. Technol. 47, 1091–1100. https://doi.org/10.1021/es303170u

Rapport, M.M., Alonzo, N.F., 1960. The Structure of Plasmalogens: V. LIPIDS OF MARINE INVERTEBRATES. Journal of Biological Chemistry 235, 1953–1956. https://doi.org/10.1016/S0021-9258(18)69342-1

Rathahao-Paris, E., Alves, S., Junot, C., Tabet, J.-C., 2015. High resolution mass spectrometry for structural identification of metabolites in metabolomics. Metabolomics 12, 10. https://doi.org/10.1007/s11306-015-0882-8

Ratier, A., 2019. Modélisation toxico-cinétique de la bioaccumulation de composés organiques persistants par des invertébrés benthiques d'eau douce (These de doctorat). Lyon.

Reeves, G.A., Talavera, D., Thornton, J.M., 2009. Genome and proteome annotation: organization, interpretation and integration. J R Soc Interface 6, 129–147. https://doi.org/10.1098/rsif.2008.0341

Rice, P., Longden, I., Bleasby, A., 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16, 276–277. https://doi.org/10.1016/s0168-9525(00)02024-2

Risk, B.A., Spitzer, W.J., Giddings, M.C., 2013. Peppy: Proteogenomic Search Software. J. Proteome Res. 12, 3019–3025. https://doi.org/10.1021/pr400208w

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., Smyth, G.K., 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research 43, e47. https://doi.org/10.1093/nar/gkv007

Rivière, J.-L., 1993. Les animaux sentinelles. Le Courrier de l'environnement de l'INRA 20, 59–68.

Rna-seq transcriptome Gammarus Fossarum B female [WWW Document], 2018. . NCBI Sequence Read Archive. URL https://www.ncbi.nlm.nih.gov/sra/SRR8089729 (accessed 12.16.21).

Rna-seq transcriptome Gammarus Fossarum B male [WWW Document], 2018. . NCBI Sequence Read Archive. URL https://www.ncbi.nlm.nih.gov/sra/SRR8089722 (accessed 12.16.21).

Roberts, A., Pachter, L., 2013. Streaming fragment assignment for real-time analysis of sequencing experiments. Nat Methods 10, 71–73. https://doi.org/10.1038/nmeth.2251

Roberts, A., Trapnell, C., Donaghey, J., Rinn, J.L., Pachter, L., 2011. Improving RNA-Seq expression estimates by correcting for fragment bias. Genome Biology 12, R22. https://doi.org/10.1186/gb-2011-12-3-r22

Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S.D., Mungall, K., Lee, S., Okada, H.M., Qian, J.Q., Griffith, M., Raymond, A., Thiessen, N., Cezard, T., Butterfield, Y.S., Newsome, R., Chan, S.K., She, R., Varhol, R., Kamoh, B., Prabhu, A.-L., Tam, A., Zhao, Y., Moore, R.A., Hirst, M., Marra, M.A., Jones, S.J.M., Hoodless, P.A., Birol, I., 2010. De novo assembly and analysis of RNA-seq data. Nat Methods 7, 909–912. https://doi.org/10.1038/nmeth.1517

Robinson, M.D., McCarthy, D.J., Smyth, G.K., 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139–140.

Robinson, M.D., Oshlack, A., 2010. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology 11, R25. https://doi.org/10.1186/gb-2010-11-3-r25

Robinson, M.D., Smyth, G.K., 2007. Moderated statistical tests for assessing differences in tag abundance. Bioinformatics 23, 2881–2887. https://doi.org/10.1093/bioinformatics/btm453

Robles, M.S., Cox, J., Mann, M., 2014. In-Vivo Quantitative Proteomics Reveals a Key Contribution of Post-Transcriptional Mechanisms to the Circadian Regulation of Liver Metabolism. PLOS Genetics 10, e1004047. https://doi.org/10.1371/journal.pgen.1004047

Rodrigues, P.M., Campos, A., Kuruvilla, J., Schrama, D., Cristobal, S., 2017. Proteomics in Aquaculture: Quality and Safety, in: Proteomics in Food Science: From Farm to Fork. pp. 279–295. https://doi.org/10.1016/B978-0-12-804007-2.00017-5

Rodrigues, P.M., Silva, T.S., Dias, J., Jessen, F., 2012. PROTEOMICS in aquaculture: Applications and trends. Journal of Proteomics 75, 4325–4345. https://doi.org/10.1016/j.jprot.2012.03.042

Rohart, F., Gautier, B., Singh, A., Cao, K.-A.L., 2017. mixOmics: An R package for 'omics feature selection and multiple data integration. PLOS Computational Biology 13, e1005752. https://doi.org/10.1371/journal.pcbi.1005752

Rohn, H., Junker, A., Hartmann, A., Grafahrend-Belau, E., Treutler, H., Klapperstück, M., Czauderna, T., Klukas, C., Schreiber, F., 2012. VANTED v2: a framework for systems biology applications. BMC Systems Biology 6, 139. https://doi.org/10.1186/1752-0509-6-139

Romanov, M.N., Tuttle, E.M., Houck, M.L., Modi, W.S., Chemnick, L.G., Korody, M.L., Mork, E.M.S., Otten, C.A., Renner, T., Jones, K.C., Dandekar, S., Papp, J.C., Da, Y., Green, E.D., Magrini, V., Hickenbotham, M.T., Glasscock, J., McGrath, S., Mardis, E.R., Ryder, O.A., 2009. The value of avian genomics to the conservation of wildlife. BMC Genomics 10, S10. https://doi.org/10.1186/1471-2164-10-S2-S10

Rosa, R., Costa, P.R., Pereira, J., Nunes, M.L., 2004. Biochemical dynamics of spermatogenesis and oogenesis in Eledone cirrhosa and Eledone moschata (Cephalopoda: Octopoda). Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology 139, 299–310. https://doi.org/10.1016/j.cbpc.2004.08.002

Rosa, R., Nunes, M.L., 2002. Changes in Organ Indices and Lipid Dynamics during the Reproductive Cycle of Aristeus antennatus, Parapenaeus longirostris, and Nephrops norvegicus (Decapoda) from the Portuguese South Coast. Crustaceana 75, 1095–1105.

Ruepp, A., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Stransky, M., Waegele, B., Schmidt, T., Doudieu, O.N., Stümpflen, V., 2007. CORUM: the comprehensive resource of mammalian protein complexes. Nucleic acids research 36, D646–D650.

Ruffalo, M., Koyutürk, M., Sharan, R., 2015. Network-Based Integration of Disparate Omic Data To Identify "Silent Players" in Cancer. PLOS Computational Biology 11, e1004595. https://doi.org/10.1371/journal.pcbi.1004595

Sadat-Hosseini, M., Bakhtiarizadeh, M.R., Boroomand, N., Tohidfar, M., Vahdati, K., 2020. Combining independent de novo assemblies to optimize leaf transcriptome of Persian walnut. PLOS ONE 15, e0232005. https://doi.org/10.1371/journal.pone.0232005

Sahlin, K., Vezzi, F., Nystedt, B., Lundeberg, J., Arvestad, L., 2014. BESST - Efficient scaffolding of large fragmented assemblies. BMC Bioinformatics 15, 281. https://doi.org/10.1186/1471-2105-15-281

Sahraeian, S.M., Luo, K.R., Brenner, S.E., 2015. SIFTER search: a web server for accurate phylogeny-based protein function prediction. Nucleic Acids Research 43, W141–W147. https://doi.org/10.1093/nar/gkv461

Salem, N.M., Lin, Y.H., Moriguchi, T., Lim, S.Y., Salem, N., Hibbeln, J.R., 2015. Distribution of omega-6 and omega-3 polyunsaturated fatty acids in the whole rat body and 25 compartments. Prostaglandins Leukot Essent Fatty Acids 100, 13–20. https://doi.org/10.1016/j.plefa.2015.06.002

Salgado, C.M., Azevedo, C., Proença, H., Vieira, S.M., 2016. Missing Data, in: MIT Critical Data (Ed.), Secondary Analysis of Electronic Health Records. Springer International Publishing, Cham, pp. 143–162. https://doi.org/10.1007/978-3-319-43742-2_13

Salzberg, S., Delcher, A.L., Fasman, K.H., Henderson, J., 1998. A decision tree system for finding genes in DNA. Journal of Computational Biology 5, 667–680.

Salzberg, S.L., Phillippy, A.M., Zimin, A., Puiu, D., Magoc, T., Koren, S., Treangen, T.J., Schatz, M.C., Delcher, A.L., Roberts, M., Marçais, G., Pop, M., Yorke, J.A., 2012. GAGE: A critical evaluation of genome assemblies and assembly algorithms. Genome Res 22, 557–567. https://doi.org/10.1101/gr.131383.111

Sanders, W.S., Wang, N., Bridges, S.M., Malone, B.M., Dandass, Y.S., McCarthy, F.M., Nanduri, B., Lawrence, M.L., Burgess, S.C., 2011. The Proteogenomic Mapping Tool. BMC Bioinformatics 12, 115. https://doi.org/10.1186/1471-2105-12-115

Sanger, F., Nicklen, S., Coulson, A.R., 1977. DNA sequencing with chain-terminating inhibitors. PNAS 74, 5463–5467. https://doi.org/10.1073/pnas.74.12.5463

Sano, R., Reed, J.C., 2013. ER stress-induced cell death mechanisms. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research 1833, 3460–3470. https://doi.org/10.1016/j.bbamcr.2013.06.028

Santos, E.A., Nery, L.E., Keller, R., Gonçalves, A.A., 1997. Evidence for the involvement of the crustacean hyperglycemic hormone in the regulation of lipid metabolism. Physiol Zool 70, 415–420. https://doi.org/10.1086/515846

Santos, E.M., Ball, J.S., Williams, T.D., Wu, H., Ortega, F., van Aerle, R., Katsiadaki, I., Falciani, F., Viant, M.R., Chipman, J.K., Tyler, C.R., 2010. Identifying Health Impacts of Exposure to Copper Using Transcriptomics and Metabolomics in a Fish Model. Environ. Sci. Technol. 44, 820–826. https://doi.org/10.1021/es902558k

Santos, M.M., Ruivo, R., Lopes-Marques, M., Torres, T., de los Santos, C.B., Castro, L.F.C., Neuparth, T., 2016. Statins: An undesirable class of aquatic contaminants? Aquat Toxicol 174, 1–9. https://doi.org/10.1016/j.aquatox.2016.02.001

Schatz, M.C., Delcher, A.L., Salzberg, S.L., 2010. Assembly of large genomes using second-generation sequencing. Genome Res. 20, 1165–1173. https://doi.org/10.1101/gr.101360.109

Schellenberg, M.J., Ritchie, D.B., MacMillan, A.M., 2008. Pre-mRNA splicing: a complex picture in higher definition. Trends in biochemical sciences 33, 243–246.

Schirling, M., Jungmann, D., Ladewig, V., Nagel, R., Triebskorn, R., Köhler, H.-R., 2005. Endocrine Effects in Gammarus fossarum (Amphipoda): Influence of Wastewater Effluents, Temporal Variability, and Spatial Aspects on Natural Populations. Arch Environ Contam Toxicol 49, 53–61. https://doi.org/10.1007/s00244-004-0153-6

Schrimpe-Rutledge, A.C., Codreanu, S.G., Sherrod, S.D., McLean, J.A., 2016. Untargeted Metabolomics Strategies—Challenges and Emerging Directions. J. Am. Soc. Mass Spectrom. 27, 1897–1905. https://doi.org/10.1007/s13361-016-1469-y

Schulz, M.H., Zerbino, D.R., Vingron, M., Birney, E., 2012. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics 28, 1086–1092. https://doi.org/10.1093/bioinformatics/bts094

Schweikert, G., Zien, A., Zeller, G., Behr, J., Dieterich, C., Ong, C.S., Philips, P., De Bona, F., Hartmann, L., Bohlen, A., Krüger, N., Sonnenburg, S., Rätsch, G., 2009. mGene: accurate SVM-based gene finding with an application to nematode genomes. Genome Res 19, 2133–2143. https://doi.org/10.1101/gr.090597.108

Senatore, A., Edirisinghe, N., Katz, P.S., 2015. Deep mRNA Sequencing of the Tritonia diomedea Brain Transcriptome Provides Access to Gene Homologues for Neuronal Excitability, Synaptic Transmission and Peptidergic Signalling. PLOS ONE 10, e0118321. https://doi.org/10.1371/journal.pone.0118321

Seppey, M., Manni, M., Zdobnov, E.M., 2019. BUSCO: Assessing Genome Assembly and Annotation Completeness, in: Kollmar, M. (Ed.), Gene Prediction: Methods and Protocols, Methods in Molecular Biology. Springer, New York, NY, pp. 227–245. https://doi.org/10.1007/978-1-4939-9173-0_14

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T., 2003. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res. 13, 2498–2504. https://doi.org/10.1101/gr.1239303

Shatilina, Z., Drozdova, P., Bedulina, D., Rivarola-Duarte, L., Schreiber, S., Otto, C., Jühling, F., Aulhorn, S., Busch, W., Lubyaga, Y., 2020. Transcriptome-level effects of the model organic pollutant phenanthrene and its solvent acetone in three amphipod species. Comparative Biochemistry and Physiology Part D: Genomics and Proteomics 33, 100630.

Sheikholeslami, M.N., Gómez-Canela, C., Barron, L.P., Barata, C., Vosough, M., Tauler, R., 2020. Untargeted metabolomics changes on Gammarus pulex induced by propranolol, triclosan, and nimesulide pharmaceutical drugs. Chemosphere 260, 127479. https://doi.org/10.1016/j.chemosphere.2020.127479

Shi, L., 2017. Untargeted Metabolomics and Novel Data Analysis Strategies to Identify Biomarkers of Diet and Type 2 Diabetes.

Shonouda, M., Osman, W., 2018. Ultrastructural alterations in sperm formation of the beetle, Blaps polycresta (Coleoptera: Tenebrionidae) as a biomonitor of heavy metal soil pollution. Environ Sci Pollut Res Int 25, 7896–7906. https://doi.org/10.1007/s11356-017-1172-y

Short, S., Yang, G., Guler, Y., Green Etxabe, A., Kille, P., Ford, A.T., 2014. Crustacean intersexuality is feminization without demasculinization: implications for environmental toxicology. Environmental science & technology 48, 13520–13529.

Sieber, P., Platzer, M., Schuster, S., 2018. The Definition of Open Reading Frame Revisited. Trends Genet 34, 167–170. https://doi.org/10.1016/j.tig.2017.12.009

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., Weinstock, G.M., Wilson, R.K., Gibbs, R.A., Kent, W.J., Miller, W., Haussler, D., 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 15, 1034–1050. https://doi.org/10.1101/gr.3715005

Simão, 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, 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Simillion, C., Liechti, R., Lischer, H.E.L., Ioannidis, V., Bruggmann, R., 2017. Avoiding the pitfalls of gene set enrichment analysis with SetRank. BMC Bioinformatics 18, 151. https://doi.org/10.1186/s12859-017-1571-6

Simmons, D.B.D., Benskin, J.P., Cosgrove, J.R., Duncker, B.P., Ekman, D.R., Martyniuk, C.J., Sherry, J.P., 2015. Omics for aquatic ecotoxicology: Control of extraneous variability to enhance the analysis of environmental effects. Environmental Toxicology and Chemistry 34, 1693–1704. https://doi.org/10.1002/etc.3002

Simon, R., Jubeaux, G., Chaumot, A., Lemoine, J., Geffard, O., Salvador, A., 2010. 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). Journal of Chromatography A 1217, 5109–5115. https://doi.org/10.1016/j.chroma.2010.06.015

Simpson, J.T., Durbin, R., 2012. Efficient de novo assembly of large genomes using compressed data structures. Genome Res 22, 549–556. https://doi.org/10.1101/gr.126953.111

Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J., Birol, I., 2009. ABySS: a parallel assembler for short read sequence data. Genome research 19, 1117–1123.

Singh, U., Wurtele, E.S., 2021. orfipy: a fast and flexible tool for extracting ORFs. Bioinformatics 37, 3019–3020. https://doi.org/10.1093/bioinformatics/btab090

Sinha Hikim, A.P., Swerdloff, R.S., 1999. Hormonal and genetic control of germ cell apoptosis in the testis. Rev Reprod 4, 38–47. https://doi.org/10.1530/ror.0.0040038

Sinitcyn, P., Rudolph, J.D., Cox, J., 2018. Computational Methods for Understanding Mass Spectrometry–Based Shotgun Proteomics Data. Annual Review of Biomedical Data Science 1, 207–234. https://doi.org/10.1146/annurev-biodatasci-080917-013516

Sleator, R.D., 2010. An overview of the current status of eukaryote gene prediction strategies. Gene 461, 1–4. https://doi.org/10.1016/j.gene.2010.04.008

Slenter, D.N., Kutmon, M., Hanspers, K., Riutta, A., Windsor, J., Nunes, N., Mélius, J., Cirillo, E., Coort, S.L., Digles, D., Ehrhart, F., Giesbertz, P., Kalafati, M., Martens, M., Miller, R., Nishida, K., Rieswijk, L., Waagmeester, A., Eijssen, L.M.T., Evelo, C.T., Pico, A.R., Willighagen, E.L., 2018. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. Nucleic Acids Res 46, D661–D667. https://doi.org/10.1093/nar/gkx1064

Smale, S.T., Kadonaga, J.T., 2003. The RNA Polymerase II Core Promoter. Annual Review of Biochemistry 72, 449–479. https://doi.org/10.1146/annurev.biochem.72.121801.161520

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, 1134–1144. https://doi.org/10.1101/gr.196469.115

Smolinska, A., Engel, J., Szymanska, E., Buydens, L., Blanchet, L., 2019. General Framing of Low-, Mid-, and High-Level Data Fusion With Examples in the Life Sciences, in: Data Handling in Science and Technology. Elsevier, pp. 51–79. https://doi.org/10.1016/B978-0-444-63984-4.00003-X

Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.-L., Ideker, T., 2011. Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics 27, 431–432. https://doi.org/10.1093/bioinformatics/btq675

Snape, J.R., Maund, S.J., Pickford, D.B., Hutchinson, T.H., 2004. Ecotoxicogenomics: the challenge of integrating genomics into aquatic and terrestrial ecotoxicology. Aquatic Toxicology 67, 143–154. https://doi.org/10.1016/j.aquatox.2003.11.011

Snyder, E.E., Stormo, G.D., 1995. Identification of protein coding regions in genomic DNA. Journal of molecular biology 248, 1–18.

Spaargaren, D.H., Haefner, P.A., Jr., 1994. Interactions of Ovary and Hepatopancreas During the Reproductive Cycle of Crangon Crangon (L.) . II. Biochemical Relationships. Journal of Crustacean Biology 14, 6–19. https://doi.org/10.1163/193724094X00425

Specht, M., Stanke, M., Terashima, M., Naumann-Busch, B., Janßen, I., Höhner, R., Hom, E.F.Y., Liang, C., 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, 1814–1823. https://doi.org/10.1002/pmic.201000621

Sroda, S., Cossu-Leguille, C., 2011. Seasonal variability of antioxidant biomarkers and energy reserves in the freshwater gammarid Gammarus roeseli. Chemosphere 83, 538–544. https://doi.org/10.1016/j.chemosphere.2010.12.023

Staldoni de Oliveira, V., Gomes Castro, A.J., Marins, K., Bittencourt Mendes, A.K., Araújo Leite, G.A., Zamoner, A., Van Der Kraak, G., Mena Barreto Silva, F.R., 2021. Pyriproxyfen induces intracellular calcium overload and alters antioxidant defenses in Danio rerio testis that may influence ongoing spermatogenesis. Environmental Pollution 270, 116055. https://doi.org/10.1016/j.envpol.2020.116055

Stanke, M., Diekhans, M., Baertsch, R., Haussler, D., 2008. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics 24, 637–644. https://doi.org/10.1093/bioinformatics/btn013

Stanke, M., Steinkamp, R., Waack, S., Morgenstern, B., 2004. AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Research 32, W309–W312. https://doi.org/10.1093/nar/gkh379

Stanke, M., Tzvetkova, A., Morgenstern, B., 2006. AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. Genome Biology 7, S11. https://doi.org/10.1186/gb-2006-7-s1-s11

Stanke, M., Waack, S., 2003. Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics 19, ii215–ii225. https://doi.org/10.1093/bioinformatics/btg1080

Steinmetz, V., Sévila, F., Bellon-Maurel, V., 1999. A Methodology for Sensor Fusion Design: Application to Fruit Quality Assessment. Journal of Agricultural Engineering Research 74, 21–31. https://doi.org/10.1006/jaer.1999.0428

Stuart, J.M., Segal, E., Koller, D., Kim, S.K., 2003. A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. Science 302, 249–255. https://doi.org/10.1126/science.1087447

Su, Y., Wang, J., Shi, M., Niu, X., Yu, X., Gao, L., Zhang, X., Chen, L., Zhang, W., 2014. Metabolomic and network analysis of astaxanthin-producing Haematococcus pluvialis under various stress conditions. Bioresource Technology 170, 522–529. https://doi.org/10.1016/j.biortech.2014.08.018

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P., 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. PNAS 102, 15545–15550. https://doi.org/10.1073/pnas.0506580102

Sun, J., Zhou, Q., Hu, X., 2019. Integrating multi-omics and regular analyses identifies the molecular responses of zebrafish brains to graphene oxide: Perspectives in environmental criteria. Ecotoxicology and Environmental Safety 180, 269–279. https://doi.org/10.1016/j.ecoenv.2019.05.011

Sun, P., Jin, M., Jiao, L., Monroig, Ó., Navarro, J.C., Tocher, D.R., Betancor, M.B., Wang, X., Yuan, Y., Zhou, Q., 2020. Effects of dietary lipid level on growth, fatty acid profiles, antioxidant capacity and expression of genes involved in lipid metabolism in juvenile swimming crab, *Portunus trituberculatus*. Br J Nutr 123, 149–160. https://doi.org/10.1017/S0007114519002563

Sun, X., Kovacs, T., Hu, Y.-J., Yang, W.-X., 2011. The role of actin and myosin during spermatogenesis. Molecular biology reports 38, 3993–4001.

Sutcliffe, D.W., 1993. Reproduction in Gammarus (Crustacea, Amphipoda): female strategies, in: Freshwater Forum. pp. 26–64.

Tachet, H., Richoux, P.H., Bournaud, M., Usseglio-Polatera, P., 2000. Invertébrés d'eau douce : systématique, biologie, écologie. CNRS éditions, Paris 588.

Tarca, A.L., Bhatti, G., Romero, R., 2013. A Comparison of Gene Set Analysis Methods in Terms of Sensitivity, Prioritization and Specificity. PLOS ONE 8, e79217. https://doi.org/10.1371/journal.pone.0079217

Taylor, N.S., White, T.A., Viant, M.R., 2017. Defining the Baseline and Oxidant Perturbed Lipidomic Profiles of Daphnia magna. Metabolites 7, 11. https://doi.org/10.3390/metabo7010011

Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y.O., Borodovsky, M., 2008. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. Genome Res 18, 1979–1990. https://doi.org/10.1101/gr.081612.108

Tessier, A.J., Henry, L.L., Goulden, C.E., Durand, M.W., 1983. Starvation in Daphnia: Energy reserves and reproductive allocation1. Limnology and Oceanography 28, 667–676. https://doi.org/10.4319/lo.1983.28.4.0667

The ENCODE Project Consortium, 2012. An Integrated Encyclopedia of DNA Elements in the Human Genome. Nature 489, 57–74. https://doi.org/10.1038/nature11247

The UniProt Consortium, 2021. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Research 49, D480–D489. https://doi.org/10.1093/nar/gkaa1100

The UniProt Consortium, 2019. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Research 47, D506–D515. https://doi.org/10.1093/nar/gky1049

The UniProt Consortium, 2009. The Universal Protein Resource (UniProt) 2009. Nucleic Acids Research 37, D169–D174. https://doi.org/10.1093/nar/gkn664

Tini, G., 2018. The influence of the inclusion of biologicalknowledge in statistical methods to integratemulti-omics data.

Tocher, D.R., 2003. Metabolism and Functions of Lipids and Fatty Acids in Teleost Fish. Reviews in Fisheries Science 11, 107–184. https://doi.org/10.1080/713610925

Tolstrup, N., Rouzé, P., Brunak, S., 1997. A branch point consensus from Arabidopsis found by non-circular analysis allows for better prediction of acceptor sites. Nucleic Acids Research 25, 3159–3163.

Torres-García, W., Zheng, S., Sivachenko, A., Vegesna, R., Wang, Q., Yao, R., Berger, M.F., Weinstein, J.N., Getz, G., Verhaak, R.G.W., 2014. PRADA: pipeline for RNA sequencing data analysis. Bioinformatics 30, 2224–2226. https://doi.org/10.1093/bioinformatics/btu169

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, 1469–1470. https://doi.org/10.1093/bioinformatics/btu051

Trapnell, C., Hendrickson, D.G., Sauvageau, M., Goff, L., Rinn, J.L., Pachter, L., 2013. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol 31, 46–53. https://doi.org/10.1038/nbt.2450

Trapnell, C., Pachter, L., Salzberg, S.L., 2009. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25, 1105–1111. https://doi.org/10.1093/bioinformatics/btp120

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., Pachter, L., 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7, 562–578. https://doi.org/10.1038/nprot.2012.016

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

Trapp, J., Almunia, C., Gaillard, J.-C., Pible, O., Chaumot, A., Geffard, O., Armengaud, J., 2016a. Proteogenomic insights into the core-proteome of female reproductive tissues from crustacean amphipods. Journal of proteomics 135, 51–61.

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

Trapp, J., Armengaud, J., Pible, O., Gaillard, J.-C., Abbaci, K., Habtoul, Y., Chaumot, A., Geffard, O., 2015. Proteomic investigation of male Gammarus fossarum, a freshwater crustacean, in response to endocrine disruptors. Journal of Proteome Research 14, 292–303. https://doi.org/10.1021/pr500984z

Trapp, J., Armengaud, J., Salvador, A., Chaumot, A., Geffard, O., 2014a. Next-Generation Proteomics: Toward Customized Biomarkers for Environmental Biomonitoring. Environ. Sci. Technol. 48, 13560–13572. https://doi.org/10.1021/es501673s

Trapp, J., Geffard, O., Imbert, G., Gaillard, J.-C., Davin, A.-H., Chaumot, A., Armengaud, J., 2014b. Proteogenomics of Gammarus fossarum to document the reproductive system of amphipods. Molecular & Cellular Proteomics mcp.M114.038851. https://doi.org/10.1074/mcp.M114.038851

Trapp, J., Gouveia, D., Almunia, C., Pible, O., Degli Esposti, D., Gaillard, J.-C., Chaumot, A., Geffard, O., Armengaud, J., 2018. 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. Front. Environ. Sci. 6. https://doi.org/10.3389/fenvs.2018.00054

Truebano, M., Tills, O., Spicer, J.I., 2016. Embryonic transcriptome of the brackishwater amphipod Gammarus chevreuxi. Marine Genomics 28, 5–6. https://doi.org/10.1016/j.margen.2016.02.002

Truhaut, R., 1977. Ecotoxicology: objectives, principles and perspectives. Ecotoxicology and environmental safety 1, 151–173.

Tsai, I.J., Otto, T.D., Berriman, M., 2010. Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. Genome Biol 11, R41. https://doi.org/10.1186/gb-2010-11-4-r41

Tsuyuzaki, K., Morota, G., Ishii, M., Nakazato, T., Miyazaki, S., Nikaido, I., 2015. MeSH ORA framework: R/Bioconductor packages to support MeSH over-representation analysis. BMC Bioinformatics 16, 45. https://doi.org/10.1186/s12859-015-0453-z

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

User guide BUSCO v5.2.2 [WWW Document], 2019. URL https://busco.ezlab.org/busco_userguide.html (accessed 2.12.22).

Vacquier, V.D., Swanson, W.J., 2011. Selection in the rapid evolution of gamete recognition proteins in marine invertebrates. Cold Spring Harb Perspect Biol 3, a002931. https://doi.org/10.1101/cshperspect.a002931

Väinölä, R., Witt, J.D.S., Grabowski, M., Bradbury, J.H., Jazdzewski, K., Sket, B., 2008. Global diversity of amphipods (Amphipoda; Crustacea) in freshwater, in: Balian, E.V., Lévêque, C., Segers, H., Martens, K. (Eds.), Freshwater Animal Diversity Assessment, Developments in Hydrobiology. Springer Netherlands, Dordrecht, pp. 241–255. https://doi.org/10.1007/978-1-4020-8259-7_27

Van Aggelen, G., Ankley, G.T., Baldwin, W.S., Bearden, D.W., Benson, W.H., Chipman, J.K., Collette, T.W., Craft, J.A., Denslow, N.D., Embry, M.R., Falciani, F., George, S.G., Helbing, C.C., Hoekstra, P.F., Iguchi, T., Kagami, Y., Katsiadaki, I., Kille, P., Liu, L., Lord, P.G., McIntyre, T., O, 'Neill Anne, Osachoff, H., Perkins, E.J., Santos, E.M., Skirrow, R.C., Snape, J.R., Tyler, C.R., Versteeg, D., Viant, M.R., Volz, D.C., Williams, T.D., Yu, L., 2010. Integrating Omic Technologies into Aquatic Ecological Risk Assessment and Environmental Monitoring: Hurdles, Achievements, and Future Outlook. Environmental Health Perspectives 118, 1–5. https://doi.org/10.1289/ehp.0900985

Van Dam, S., 2017. Development and exploitation of GeneFriends: An online database for gene and transcript co-expression analysis (phd). University of Liverpool.

van der Hooft, J.J.J., Wandy, J., Barrett, M.P., Burgess, K.E.V., Rogers, S., 2016. Topic modeling for untargeted substructure exploration in metabolomics. Proceedings of the National Academy of Sciences 113, 13738–13743. https://doi.org/10.1073/pnas.1608041113

van Meer, G., Voelker, D.R., Feigenson, G.W., 2008. Membrane lipids: where they are and how they behave. Nat Rev Mol Cell Biol 9, 112–124. https://doi.org/10.1038/nrm2330

Van Straalen, N.M., 2003. Peer reviewed: ecotoxicology becomes stress ecology. Environmental science & technology 37, 324A-330A.

Vaske, C.J., Benz, S.C., Sanborn, J.Z., Earl, D., Szeto, C., Zhu, J., Haussler, D., Stuart, J.M., 2010. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics 26, i237–i245. https://doi.org/10.1093/bioinformatics/btq182

Vellinger, C., Sohm, B., Parant, M., Immel, F., Usseglio-Polatera, P., 2016. Investigating the emerging role of comparative proteomics in the search for new biomarkers of metal contamination under varying abiotic conditions. Science of The Total Environment 562, 974–986.

Vellozo, A.F., Véron, A.S., Baa-Puyoulet, P., Huerta-Cepas, J., Cottret, L., Febvay, G., Calevro, F., Rahbé, Y., Douglas, A.E., Gabaldón, T., Sagot, M.-F., Charles, H., Colella, S., 2011. CycADS: an annotation database system to ease the development and update of BioCyc databases. Database (Oxford) 2011, baro08. https://doi.org/10.1093/database/baro08

Verberk, W.C.E.P., Leuven, R.S.E.W., van der Velde, G., Gabel, F., 2018. Thermal limits in native and alien freshwater peracarid Crustacea: The role of habitat use and oxygen limitation. Functional Ecology 32, 926–936. https://doi.org/10.1111/1365-2435.13050

Videvall, E., 2017. What's N50? The Molecular Ecologist. URL https://www.molecularecologist.com/2017/03/29/whats-n50/ (accessed 2.8.22).

Vigneron, A., 2015. Capacités d'adaptation des populations naturelles à la contamination des milieux aquatiques : cas d'étude du cadmium chez le crustacé Gammarus fossarum (These de doctorat). Lyon 1.

Vinaixa, M., Samino, S., Saez, I., Duran, J., Guinovart, J.J., Yanes, O., 2012. A guideline to univariate statistical analysis for LC/MS-based untargeted metabolomics-derived data. Metabolites 2, 775–795.

Wahl, S., Vogt, S., Stückler, F., Krumsiek, J., Bartel, J., Kacprowski, T., Schramm, K., Carstensen, M., Rathmann, W., Roden, M., Jourdan, C., Kangas, A.J., Soininen, P., Ala-Korpela, M., Nöthlings, U., Boeing, H., Theis, F.J., Meisinger, C., Waldenberger, M., Suhre, K., Homuth, G., Gieger, C., Kastenmüller, G., Illig, T., Linseisen, J., Peters, A., Prokisch, H., Herder, C., Thorand, B., Grallert, H., 2015. Multi-omic signature of body weight change: results from a population-based cohort study. BMC Medicine 13, 48. https://doi.org/10.1186/s12916-015-0282-y

Wang, J., Zuo, Y., Man, Y., Avital, I., Stojadinovic, A., Liu, M., Yang, X., Varghese, R.S., Tadesse, M.G., Ressom, H.W., 2015. Pathway and Network Approaches for Identification of Cancer Signature Markers from Omics Data. J Cancer 6, 54–65. https://doi.org/10.7150/jca.10631

Wang, K., Singh, D., Zeng, Z., Coleman, S.J., Huang, Y., Savich, G.L., He, X., Mieczkowski, P., Grimm, S.A., Perou, C.M., MacLeod, J.N., Chiang, D.Y., Prins, J.F., Liu, J., 2010. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Res 38, e178. https://doi.org/10.1093/nar/gkq622

Wang, L., Xiao, Y., Ping, Y., Li, J., Zhao, H., Li, F., Hu, J., Zhang, H., Deng, Y., Tian, J., Li, X., 2014. Integrating Multi-Omics for Uncovering the Architecture of Cross-Talking Pathways in Breast Cancer. PLOS ONE 9, e104282. https://doi.org/10.1371/journal.pone.0104282

Wang, L., Yan, B., Liu, N., Li, Y., Wang, Q., 2008. Effects of cadmium on glutathione synthesis in hepatopancreas of freshwater crab, Sinopotamon yangtsekiense. Chemosphere 74, 51–56. https://doi.org/10.1016/j.chemosphere.2008.09.025

Wang, M., Carver, J.J., Phelan, V.V., Sanchez, L.M., Garg, N., Peng, Y., Nguyen, D.D., Watrous, J., Kapono, C.A., Luzzatto-Knaan, T., Porto, C., Bouslimani, A., Melnik, A.V., Meehan, M.J., Liu, W.-T., Crüsemann, M., Boudreau, P.D., Esquenazi, E., Sandoval-Calderón, M., Kersten, R.D., Pace, L.A., Quinn, R.A., Duncan, K.R., Hsu, C.-C., Floros, D.J., Gavilan, R.G., Kleigrewe, K., Northen, T., Dutton, R.J., Parrot, D., Carlson, E.E., Aigle, B., Michelsen, C.F., Jelsbak, L., Sohlenkamp, C., Pevzner, P., Edlund, A., McLean, J., Piel, J., Murphy, B.T., Gerwick, L., Liaw, C.-C., Yang, Y.-L., Humpf, H.-U., Maansson, M., Keyzers, R.A., Sims, A.C., Johnson, A.R., Sidebottom, A.M., Sedio, B.E., Klitgaard, A., Larson, C.B., Boya P, C.A., Torres-Mendoza, D., Gonzalez, D.J., Silva, D.B., Marques, L.M., Demarque, D.P., Pociute, E., O'Neill, E.C., Briand, E., Helfrich, E.J.N., Granatosky, E.A., Glukhov, E., Ryffel, F., Houson, H., Mohimani, H., Kharbush, J.J., Zeng, Y., Vorholt, J.A., Kurita, K.L., Charusanti, P., McPhail, K.L., Nielsen, K.F., Vuong, L., Elfeki, M., Traxler, M.F., Engene, N., Koyama, N., Vining, O.B., Baric, R., Silva, R.R., Mascuch, S.J., Tomasi, S., Jenkins, S., Macherla, V., Hoffman, T., Agarwal, V., Williams, P.G., Dai, J., Neupane, R., Gurr, J., Rodríguez, A.M.C., Lamsa, A., Zhang, C., Dorrestein, K., Duggan, B.M., Almaliti, J., Allard, P.-M., Phapale, P., Nothias, L.-F., Alexandrov, T., Litaudon, M., Wolfender, J.-L., Kyle, J.E., Metz, T.O., Peryea, T., Nguyen, D.-T., VanLeer, D., Shinn, P., Jadhav, A., Müller, R., Waters, K.M., Shi, W., Liu, X., Zhang, L., Knight, R., Jensen, P.R., Palsson, B.Ø., Pogliano, K., Linington, R.G., Gutiérrez, M., Lopes, N.P., Gerwick, W.H., Moore, B.S., Dorrestein, P.C., Bandeira, N., 2016. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nat Biotechnol 34, 828–837. https://doi.org/10.1038/nbt.3597

Wang, W., Wu, X., Liu, Z., Zheng, H., Cheng, Y., 2014. Insights into hepatopancreatic functions for nutrition metabolism and ovarian development in the crab Portunus trituberculatus: gene discovery in the comparative transcriptome of different hepatopancreas stages. PLoS One 9, e84921. https://doi.org/10.1371/journal.pone.0084921

Wang, X., Slebos, R.J.C., 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 *. Molecular & Cellular Proteomics 15, 1164–1175. https://doi.org/10.1074/mcp.M115.052860

Wang, Z., Gerstein, M., Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10, 57–63. https://doi.org/10.1038/nrg2484

Wang, Z.-Y., Leushkin, E., Liechti, A., Ovchinnikova, S., Mößinger, K., Brüning, T., Rummel, C., Grützner, F., Cardoso-Moreira, M., Janich, P., Gatfield, D., Diagouraga, B., de Massy, B., Gill, M.E., Peters, A.H.F.M., Anders, S., Kaessmann, H., 2020. Transcriptome and translatome co-evolution in mammals. Nature 588, 642–647. https://doi.org/10.1038/s41586-020-2899-z

Wanichthanarak, K., Fahrmann, J.F., Grapov, D., 2015. Genomic, Proteomic, and Metabolomic Data Integration Strategies. Biomark Insights 10s4, BMI.S29511. https://doi.org/10.4137/BMI.S29511

Wanjek, C., 2011. Systems Biology as Defined by NIH [WWW Document]. NIH Intramural Research Program. URL https://irp.nih.gov/catalyst/v19i6/systems-biology-as-defined-by-nih (accessed 6.18.22).

Waters, M.D., Fostel, J.M., 2004. Toxicogenomics and systems toxicology: aims and prospects. Nat Rev Genet 5, 936–948. https://doi.org/10.1038/nrg1493

Wautier, J., Roux, A., 1959. Note sur les Gammares du groupe Pulex dans la région lyonnaise. Publications de la Société Linnéenne de Lyon 28, 76–83. https://doi.org/10.3406/linly.1959.8046

Webb-Robertson, B.-J.M., Wiberg, H.K., Matzke, M.M., Brown, J.N., Wang, J., McDermott, J.E., Smith, R.D., Rodland, K.D., Metz, T.O., Pounds, J.G., Waters, K.M., 2015. Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics. J Proteome Res 14, 1993–2001. https://doi.org/10.1021/pr501138h

Wenk, M.R., 2010. Lipidomics: New Tools and Applications. Cell 143, 888–895. https://doi.org/10.1016/j.cell.2010.11.033

Wheelock, C.E., Goss, V.M., Balgoma, D., Nicholas, B., Brandsma, J., Skipp, P.J., Snowden, S., Burg, D., D'Amico, A., Horvath, I., Chaiboonchoe, A., Ahmed, H., Ballereau, S., Rossios, C., Chung, K.F., Montuschi, P., Fowler, S.J., Adcock, I.M., Postle, A.D., Dahlén, S.-E., Rowe, A., Sterk, P.J., Auffray, C., Djukanovic, R., U-BIOPRED Study Group, 2013. Application of 'omics technologies to biomarker discovery in inflammatory lung diseases. Eur Respir J 42, 802–825. https://doi.org/10.1183/09031936.00078812

Wieczorek, S., Combes, F., Lazar, C., Giai Gianetto, Q., Gatto, L., Dorffer, A., Hesse, A.-M., Couté, Y., Ferro, M., Bruley, C., Burger, T., 2017. DAPAR & ProStaR: software to perform statistical analyses in quantitative discovery proteomics. Bioinformatics 33, 135–136. https://doi.org/10.1093/bioinformatics/btw580

Wigh, A., Geffard, O., Abbaci, K., Francois, A., Noury, P., Bergé, A., Vulliet, E., Domenjoud, B., Gonzalez-Ospina, A., Bony, S., Devaux, A., 2017. Gammarus fossarum as a sensitive tool to reveal residual toxicity of treated wastewater effluents. Science of The Total Environment 584–585, 1012–1021. https://doi.org/10.1016/j.scitotenv.2017.01.154

Williams, T.D., Wu, H., Santos, E.M., Ball, J., Katsiadaki, I., Brown, M.M., Baker, P., Ortega, F., Falciani, F., Craft, J.A., Tyler, C.R., Chipman, J.K., Viant, M.R., 2009. Hepatic Transcriptomic and Metabolomic Responses in the Stickleback (Gasterosteus aculeatus) Exposed to Environmentally Relevant Concentrations of Dibenzanthracene. Environ. Sci. Technol. 43, 6341–6348. https://doi.org/10.1021/es9008689

Wishart, D.S., 2007. Current Progress in computational metabolomics. Briefings in Bioinformatics 8, 279–293. https://doi.org/10.1093/bib/bbm030

Wishart, D.S., Jewison, T., Guo, A.C., Wilson, M., Knox, C., Liu, Y., Djoumbou, Y., Mandal, R., Aziat, F., Dong, E., Bouatra, S., Sinelnikov, I., Arndt, D., Xia, J., Liu, P., Yallou, F., Bjorndahl, T., Perez-Pineiro, R., Eisner, R., Allen, F., Neveu, V., Greiner, R., Scalbert, A., 2013. HMDB 3.0—The Human Metabolome Database in 2013. Nucleic Acids Research 41, D801–D807. https://doi.org/10.1093/nar/gks1065

Woo, S., Cha, S.W., Merrihew, G., He, Y., Castellana, N., Guest, C., MacCoss, M., Bafna, V., 2014. Proteogenomic Database Construction Driven from Large Scale RNA-seq Data. J. Proteome Res. 13, 21–28. https://doi.org/10.1021/pr400294c

Wood, D.E., Lu, J., Langmead, B., 2019. Improved metagenomic analysis with Kraken 2. Genome Biology 20, 257. https://doi.org/10.1186/s13059-019-1891-0

Wu, C., Zhou, F., Ren, J., Li, X., Jiang, Y., Ma, S., 2019. A Selective Review of Multi-Level Omics Data Integration Using Variable Selection. High-Throughput 8, 4. https://doi.org/10.3390/ht8010004

Wu, Y., Li, L., 2016. Sample normalization methods in quantitative metabolomics. Journal of Chromatography A, Editors' Choice X 1430, 80–95. https://doi.org/10.1016/j.chroma.2015.12.007

Wurm, Y., Wang, J., Riba-Grognuz, O., Corona, M., Nygaard, S., Hunt, B.G., Ingram, K.K., Falquet, L., Nipitwattanaphon, M., Gotzek, D., Dijkstra, M.B., Oettler, J., Comtesse, F., Shih, C.-J., Wu, W.-J., Yang, C.-C., Thomas, J., Beaudoing, E., Pradervand, S., Flegel, V., Cook, E.D., Fabbretti, R., Stockinger, H., Long, L., Farmerie, W.G., Oakey, J., Boomsma, J.J., Pamilo, P., Yi, S.V., Heinze, J., Goodisman, M.A.D., Farinelli, L., Harshman, K., Hulo, N., Cerutti, L., Xenarios, I., Shoemaker, D., Keller, L., 2011. The genome of the fire ant Solenopsis invicta. Proceedings of the National Academy of Sciences 108, 5679–5684. https://doi.org/10.1073/pnas.1009690108

Xia, J., Mandal, R., Sinelnikov, I.V., Broadhurst, D., Wishart, D.S., 2012. MetaboAnalyst 2.0—a comprehensive server for metabolomic data analysis. Nucleic Acids Research 40, W127–W133. https://doi.org/10.1093/nar/gks374

Xia, J., Sinelnikov, I.V., Han, B., Wishart, D.S., 2015. MetaboAnalyst 3.0—making metabolomics more meaningful. Nucleic Acids Research 43, W251–W257. https://doi.org/10.1093/nar/gkv380

Xiang, Q.-Q., Yan, H., Luo, X.-W., Kang, Y.-H., Hu, J.-M., Chen, L.-Q., 2021. Integration of transcriptomics and metabolomics reveals damage and recovery mechanisms of fish gills in response to nanosilver exposure. Aquat. Toxicol. 237, 105895. https://doi.org/10.1016/j.aquatox.2021.105895

Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., Huang, W., He, G., Gu, S., Li, S., Zhou, X., Lam, T.-W., Li, Y., Xu, X., Wong, G.K.-S., Wang, J., 2014. SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. Bioinformatics 30, 1660–1666. https://doi.org/10.1093/bioinformatics/btu077

Xu, Y., Mural, R.J., Einstein, J.R., Shah, M.B., Uberbacher, E.C., 1996. GRAIL: a multi-agent neural network system for gene identification. Proceedings of the IEEE 84, 1544–1552. https://doi.org/10.1109/5.537117

Xuereb, B., 2009. Développement de marqueurs de neurotoxicité et de perturbations endocrines chez l'amphipode d'eau douce Gammarus fossarum (PhD Thesis). Doctorat, spécialité:" Toxicologie de l'environnement", Université de Metz.

Xuereb, B., Lefèvre, E., Garric, J., Geffard, O., 2009. Acetylcholinesterase activity in Gammarus fossarum (Crustacea Amphipoda): Linking AChE inhibition and behavioural alteration. Aquatic Toxicology 94, 114–122. https://doi.org/10.1016/j.aquatox.2009.06.010

Yalamanchili, H.K., Wan, Y.-W., Liu, Z., 2017. Data Analysis Pipeline for RNA-seq Experiments: From Differential Expression to Cryptic Splicing. Current Protocols in Bioinformatics 59, 11.15.1-11.15.21. https://doi.org/10.1002/cpbi.33

Yandell, M., Ence, D., 2012. A beginner's guide to eukaryotic genome annotation. Nat Rev Genet 13, 329–342. https://doi.org/10.1038/nrg3174

Yang, C., Bolotin, E., Jiang, T., Sladek, F.M., Martinez, E., 2007. Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. Gene 389, 52–65. https://doi.org/10.1016/j.gene.2006.09.029

Yang, Y., Smith, S.A., 2013. Optimizing de novo assembly of short-read RNA-seq data for phylogenomics. BMC Genomics 14, 328. https://doi.org/10.1186/1471-2164-14-328

Ye, Z., Xu, S., Spitze, K., Asselman, J., Jiang, X., Ackerman, M.S., Lopez, J., Harker, B., Raborn, R.T., Thomas, W.K., Ramsdell, J., Pfrender, M.E., Lynch, M., 2017. A New Reference Genome Assembly for the Microcrustacean Daphnia pulex. G3 Genes|Genomes|Genetics 7, 1405–1416. https://doi.org/10.1534/g3.116.038638

Yeh, R.-F., Lim, L.P., Burge, C.B., 2001. Computational Inference of Homologous Gene Structures in the Human Genome. Genome Res. 11, 803–816. https://doi.org/10.1101/gr.175701

Yu, X., Niu, X., Zhang, X., Pei, G., Liu, J., Chen, L., Zhang, W., 2015. Identification and mechanism analysis of chemical modulators enhancing astaxanthin accumulation in Haematococcus pluvialis. Algal Research 11, 284–293. https://doi.org/10.1016/j.algal.2015.07.006

Yuan, H., Qin, F., Guo, W., Gu, H., Shao, A., 2016. Oxidative stress and spermatogenesis suppression in the testis of cadmium-treated Bombyx mori larvae. Environ Sci Pollut Res Int 23, 5763–5770. https://doi.org/10.1007/s11356-015-5818-3

Yuan, Y., Sun, P., Jin, M., Wang, X., Zhou, Q., 2019. Regulation of Dietary Lipid Sources on Tissue Lipid Classes and Mitochondrial Energy Metabolism of Juvenile Swimming Crab, Portunus trituberculatus. Front Physiol 10, 454. https://doi.org/10.3389/fphys.2019.00454

Zdobnov, E.M., Apweiler, R., 2001. InterProScan--an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17, 847–848. https://doi.org/10.1093/bioinformatics/17.9.847

Zeeberg, B.R., Qin, H., Narasimhan, S., Sunshine, M., Cao, H., Kane, D.W., Reimers, M., Stephens, R.M., Bryant, D., Burt, S.K., Elnekave, E., Hari, D.M., Wynn, T.A., Cunningham-Rundles, C., Stewart, D.M., Nelson, D., Weinstein, J.N., 2005. High-Throughput GoMiner, an "industrial-strength" integrative gene ontology tool for interpretation of multiple-microarray experiments, with application to studies of Common Variable Immune Deficiency (CVID). BMC Bioinformatics 6, 168. https://doi.org/10.1186/1471-2105-6-168

Zeng, Y., Ren, K., Zhu, X., Zheng, Z., Yi, G., 2018. Chapter One - Long Noncoding RNAs: Advances in Lipid Metabolism, in: Makowski, G.S. (Ed.), Advances in Clinical Chemistry, Advances in Clinical Chemistry. Elsevier, pp. 1–36. https://doi.org/10.1016/bs.acc.2018.07.001

Zerbino, D.R., Birney, E., 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18, 821–829. https://doi.org/10.1101/gr.074492.107

Zhang, B., Horvath, S., 2005. A General Framework for Weighted Gene Co-Expression Network Analysis. Statistical Applications in Genetics and Molecular Biology 4. https://doi.org/10.2202/1544-6115.1128

Zhang, B., VerBerkmoes, N.C., Langston, M.A., Uberbacher, E., Hettich, R.L., Samatova, N.F., 2006. Detecting Differential and Correlated Protein Expression in Label-Free Shotgun Proteomics. J. Proteome Res. 5, 2909–2918. https://doi.org/10.1021/pr0600273

Zhang, Q., Li, J., Xue, H., Kong, L., Wang, Y., 2016. Network-based methods for identifying critical pathways of complex diseases: a survey. Mol. BioSyst. 12, 1082–1089. https://doi.org/10.1039/C5MB00815H

Zhang, W., Chen, J., Yang, Y., Tang, Y., Shang, J., Shen, B., 2011. A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies. PLoS One 6, e17915. https://doi.org/10.1371/journal.pone.0017915

Zhang, Y., Qian, J., Gu, C., Yang, Y., 2021. Alternative splicing and cancer: a systematic review. Sig Transduct Target Ther 6, 1–14. https://doi.org/10.1038/s41392-021-00486-7

Zhang, Y.-M., Rock, C.O., 2008. Membrane lipid homeostasis in bacteria. Nat Rev Microbiol 6, 222–233. https://doi.org/10.1038/nrmicro1839

Zhao, Q.-Y., Wang, Y., Kong, Y.-M., Luo, D., Li, X., Hao, P., 2011. Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study. BMC Bioinformatics 12, S2. https://doi.org/10.1186/1471-2105-12-S14-S2

Zhou, G., Li, S., Xia, J., 2020. Network-Based Approaches for Multi-omics Integration, in: Li, S. (Ed.), Computational Methods and Data Analysis for Metabolomics, Methods in Molecular Biology. Springer US, New York, NY, pp. 469–487. https://doi.org/10.1007/978-1-0716-0239-3_23

Zhu, Y., Hultin-Rosenberg, L., Forshed, J., Branca, R.M.M., Orre, L.M., Lehtiö, J., 2014. SpliceVista, a Tool for Splice Variant Identification and Visualization in Shotgun Proteomics Data *. Molecular & Cellular Proteomics 13, 1552–1562. https://doi.org/10.1074/mcp.M113.031203

Zimin, A.V., Delcher, A.L., Florea, L., Kelley, D.R., Schatz, M.C., Puiu, D., Hanrahan, F., Pertea, G., Van Tassell, C.P., Sonstegard, T.S., Marçais, G., Roberts, M., Subramanian, P., Yorke, J.A., Salzberg, S.L., 2009. A whole-genome assembly of the domestic cow, Bos taurus. Genome Biol 10, R42. https://doi.org/10.1186/gb-2009-10-4-r42

Živić, I., Marković, Z., 2007. Distribution of the Species Gammarus balcanicus and Gammarus fossarum on the Territory of Serbia (Central Part of the Balkan Peninsula). Crustaceana 80, 67–76.

Züllig, T., Köfeler, H.C., 2021. High Resolution Mass Spectrometry in Lipidomics. Mass Spectrometry Reviews 40, 162–176. https://doi.org/10.1002/mas.21627