Surveillance intégrée pour les écosystèmes dulçaquicoles : Biomarqueurs et Omics chez le bivalve d'eau douce Dreissena polymorpha (BIOMICS)

Potentiel des approches omiques chez la dreissène pour le diagnostic environnemental

Rapport final

Convention ONEMA – Université de Lorraine
Laboratoire LIEC – UMR CNRS 7360

Auteurs
Romain Peden
Bénédicte Sohm, Pascal Poupin, Virginie Jouffret,
Simon Devin

Contexte et Objectifs

Les approches « -omiques » regroupent les techniques dites à haut débit et permettent d'étudier la complexité du vivant à différents échelons biologiques (transcriptomique pour l'ARN, protéomique pour les protéines...). Grâce à ces approches, il est possible de travailler sur les réponses physiologiques et moléculaires à la suite d'un stress environnemental (température, contaminations…). La puissance de ces techniques repose sur leur capacité à étudier un nombre considérable d'acteurs moléculaires simultanément et sans *a priori* (analyse globale). Elles ont bénéficié et bénéficient de l'essor technologique et numérique récent à la fois au niveau du traitement des échantillons (rapidité, accessibilité, coût...) mais aussi au niveau de l'analyse bio-informatique des données.

En écotoxicologie, l'utilisation des approches « -omiques » est relativement récent. Regroupées dans le terme « écotoxicogénomique », elles sont utilisées pour déterminer i) les modes d'action de contaminants, ii) les mécanismes d'adaptation des organismes vis à vis de ces stress chimiques ou iii) identifier des biomarqueurs d'exposition. Dans la littérature, les études utilisant une approche écotoxicogénomique sur des espèces aquatiques ont été multipliées par dix entre 2000 et 2013, parmi lesquelles trois sont majoritairement représentées : a) la ***transcriptomique*** avec l'utilisation de la qPCR[1], des puces à ADN et du séquençage à haut débit (RNAseq), b) la ***protéomique*** avec l'électrophorèse à deux dimension et/ou la spectrométrie de masse et enfin c) la **métabolomique**, nécessitant néanmoins des moyens d'analyses plus pointus (e.g. UPLC[2]) et l'utilisation de produits traceurs coûteux. Bien qu'il soit possible de s'affranchir de l'identification des transcrits ou protéines (e.g. Protein Expression Signature en protéomique), l'obstacle principal dans ces études est la difficulté à caractériser ces variants, notamment à cause de données génomique parcellaires. La transcriptomique et la protéomique ne peuvent donc être pleinement utilisées qu'à la condition de disposer d'un génome séquencé et annoté, qu'il soit de l'espèce étudiée ou phylogénétiquement proche, ou, dans une moindre mesure, d'un transcriptome de référence. C'est d'ailleurs l'objectif de la protéogénomique, discipline qui consiste à compléter et affiner l'annotation du génome d'organismes à partir données protéomiques sont générées à haut-débit.

L'identification des acteurs moléculaires induits face à un stress permet, comme évoqué précédemment, de proposer des biomarqueurs d'effet ou d'exposition à un stress. Ces derniers sont utilisés en biosurveillance pour permettre une évaluation rapide et fiable de l'état de santé d'un milieu. L'échelon moléculaire, du fait de la précocité et de la sensibilité de sa réponse à un stress, est souvent proposé comme cible favorite pour la découverte de ces biomarqueurs. Dans cette optique, l'utilisation

1 PCR quantitative (ou PCR en temps réel)
2 Chromatographie liquide ultra-performante

de la transcriptomique et de la protéomique est un atout étant donné leurs capacités à cribler un nombre considérable d'acteurs moléculaires – biomarqueurs – simultanément. Pour prendre en compte un plus grand nombre de réponses, une approche intégrative des biomarqueurs peut être menée. C'est ce qui a été réalisé avec l'Integrated Biomarker Index (IBR) où différents biomarqueurs sont sélectionnés et regroupés pour donner un index permettant ainsi d'être un indicateur du stress environnemental. A l'instar de l'IBR, cette approche intégrée a été adaptée récemment en protéomique sous le nom d'Integrated Biomarker Proteomic (IBP). Trente-quatre protéines ont été sélectionnées pour développer un index capable de qualifier et de quantifier la contamination. Les auteurs ont démontré qu'ils étaient ainsi capables de discriminer le type de contaminant (DTT ou cadmium) tout en décrivant leur effet dose-réponse.

La découverte de biomarqueurs passe également par la caractérisation précise de sa variation physiologique en dehors de tout stress chimique. Évaluer les variations naturelles d'un biomarqueur permet d'en établir les niveaux de base et d'ainsi interpréter avec plus de finesse les réponses mesurées. Par exemple, il a été démontré une variation journalière (circadienne) de transcrits et de protéines chez des espèces aquatiques des milieux marins et dulçaquicoles.

Le séquençage à haut débit des transcrits (*e.g.* RNAseq) permet de construire, pour des espèces non séquencées, des puces à ADN personnalisées. A l'inverse du RNAseq où tout les transcrits sont analysés sans *a priori*, les puces (ou microarray) peuvent être construites selon la ou les questions biologiques posées ou en sélectionnant des gènes d'intérêt. Des puces à ADN existent déjà chez *Dreissena polymorpha*. A terme, il devrait être possible d'obtenir une puce biomarqueur.

Stratégie mise en œuvre dans le cadre de la convention BIOMICS

En parallèle à l'approche de surveillance utilisant des biomarqueurs à une échelle (sub-) individuelle, nous proposons d'associer une approche de génomique environnementale (transcriptomique, protéomique). En effet, dans le cadre de l'évaluation des risques chimiques, l'approche omics est encore peu utilisée et si son intérêt est démontré dans cette étude de biosurveillance, elle permettra à terme d'améliorer les connaissances sur les mécanismes d'action toxique des contaminants chez les organismes aquatiques.

Démarche Méthodologique

1- Stratégie d'échantillonnage

Comme nous venons de le voir, l'objectif de ce volet du projet Biomics était d'évaluer le potentiel d'une approche omique comme au sein des méthodes de bioévaluation. Il était initialement prévu de séquencer le génome de *D. polymorpha*. Nous avons vite du renoncer à cette option, pour plusieurs raisons :

- le coût et le temps d'assemblage des données de séquençage, ainsi que les moyens humains à mobiliser pour réaliser ces tâches, était hors de portée de l'enveloppe financière disponible

- l'arrivée de *D. r. bugensis* dans les hydrosystèmes français concomitamment au démarrage de la convention nous aurait obligé, pour être pertinents, de réaliser également le séquençage du génome de cette deuxième espèce

- le séquençage du génôme est une brique indispensable pour comprendre les modes d'action des contaminants, pour explorer les patrons d'adaptation locale, mais ne nous permettait pas de faire des corrélations directes avec les niveaux de contamination relevés sur les différents sites d'étude ou avec les réponses biochimiques mesurées simultanément

Pour l'ensemble de ces raisons, nous nous sommes tournés vers le séquençage du transcriptome, qui reflète la quantité d'ARN disponible à un instant t dans la cellule, comme précurseur à la synthèse protéique, et en réponse à des modifications des conditions environnementales.

En se focalisant sur le transcriptome, nous pourrons alors évaluer le potentiel de cette échelle de réponse biologique à refléter à la fois un état de contamination du milieu et une modification des paramètres biologiques des organismes. Le fait de disposer, pour les mêmes dates, des réponses biochimiques, nous permettra en outre d'évaluer la cohérence des réponses entre 2 échelles d'intégration biologique différentes.

Pour chaque espèces, 7 populations ont été échantillonnées en avril/mai 2016. La stratégie d'échantillonnage visait à englober la plus grande variété possible de contextes hydromorphologiques et de profils de contamination (Figure 1). Notre objectif était aussi d'échantillonner des populations sympatriques afin de faciliter les comparaisons inter-espèces. Pour chaque population, 5 individus ont été séquencés indépendamment, pour pouvoir rendre compte de la variabilité aux échelles intra- et inter-populationnelles.

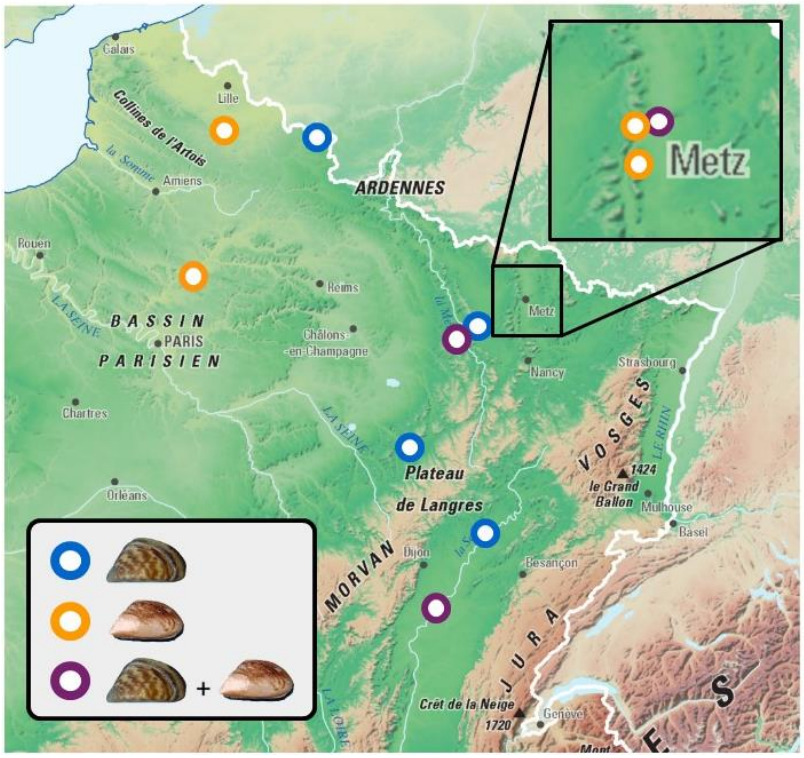


Figure 1 : localisation des sites échantillonnés pour les 2 espèces de dreissènidés. Bleu : <u>D. polymorpha</u> ; Orange : <u>D. r. bugensis</u> ; Violet : Populations sympatriques (la Meuse à Saint-Mihiel, le Canal de Jouy à Montigny-les-Metz et la Saone à Seurre), permettant de comparer les différences des transcriptomes entre les 2 espèces, indépendamment des paramètres environnementaux.

Les données de contamination des sédiments ont été analysées avec une Analyse en Composantes Principales (Figure 2). Le site de Condé apparait comme le plus contaminé, par les métaux et HAP, alors que Courcelles présente un profil de contamination par les HAP très marqué. Les sites de Scey, Saint-Mihiel, du Lac de Madine et de Pont-Sainte-Maxence peuvent être considérés comme des sites de référence pour cette étude. Les autres sites présentent des profils de contamination intermédiaires, principalement marqué par un gradient de concentration en HAP le long de l'axe F1.

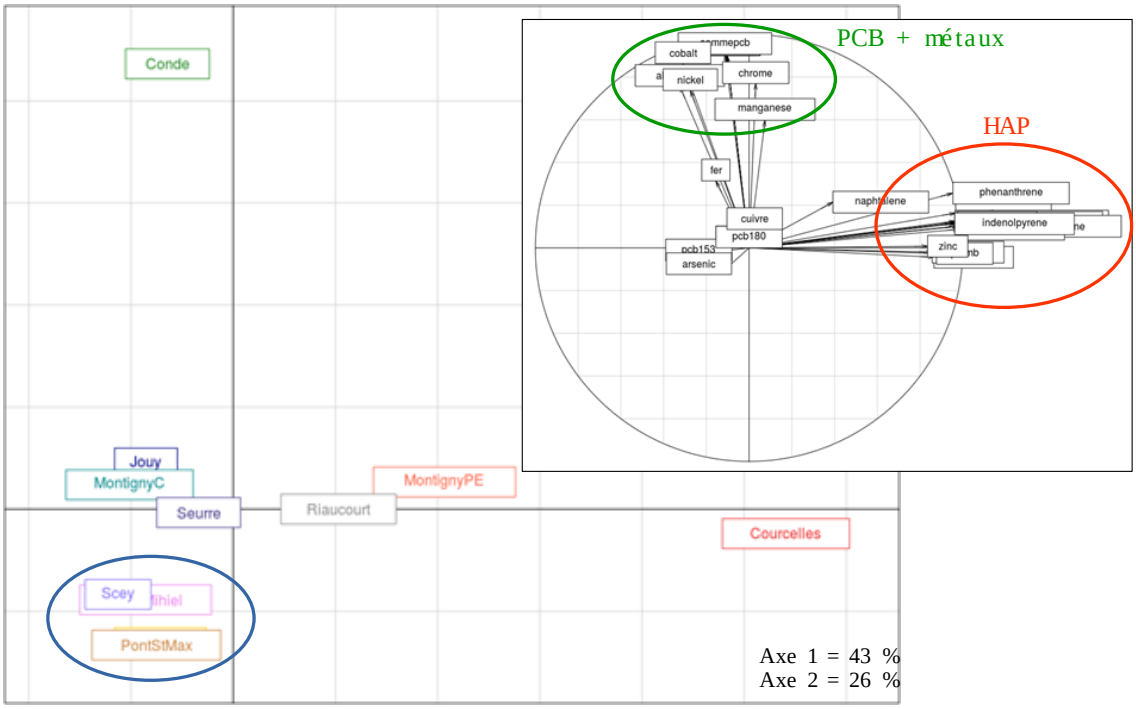


Figure 2 : Plan factoriel des stations et cercle des corrélations des contaminants du sédiment des 11

2- Séquençage RNASeq et Analyse Bioinformatique

2.1. Données générales du séquençage

L'extraction de l'ARN s'est déroulée au LIEC, en se focalisant sur la glande difestive. Les échantillons ont ensuite été expédiés au Pôle GenoToul (INRA Toulouse). Un séquençage par la méthode Illumina HiSeq3000, Reverse Forward 2*150pb a été réalisé. L'assemblage et l'alignement de novo a été réalisé avec runDRAP Oases. L'annotation a été faite avec Blastx contre les bases RefSeq et Swissprot, les données de séquençage du génome de Lottia gigantea, Lingula anatina et Crassostrea gigas. Plus de 50% des contigs séquencés ont pu être annotés (Figure 3).

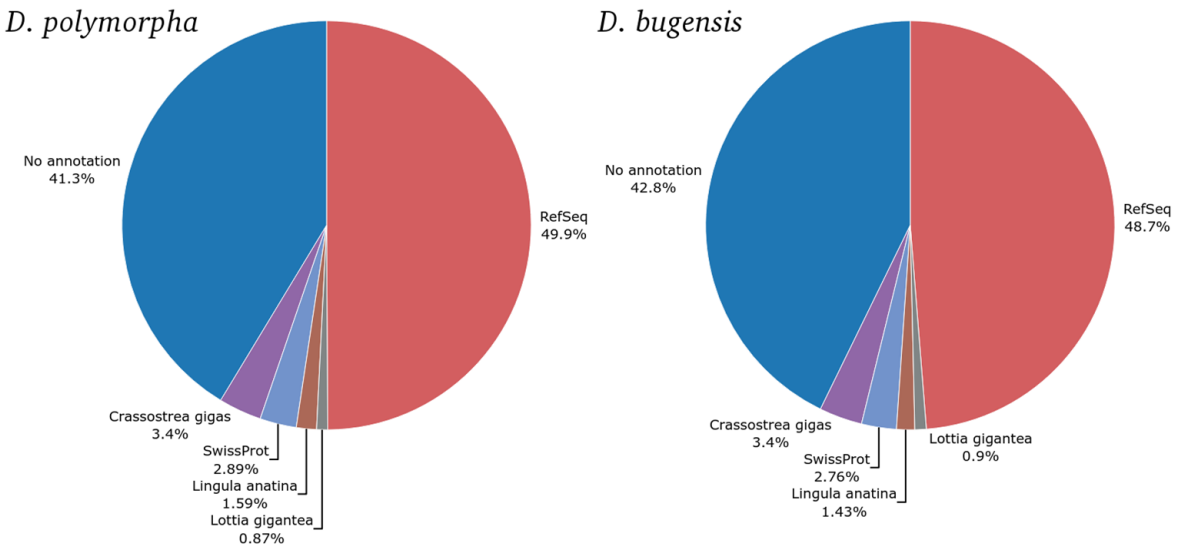


Figure 3: Annotation des contigs.

2.2. Dreissena polymorpha

Sept sites ont été validés : Condé, MontignyC (canal), Saint-Mihiel, Scey, Madine, Seurre et Riaucourt. Deux sites ont été exclus, Nantua et Pont-Sainte-Maxence, du fait de la présence d'un contig (CYPB, ADN mitochondrial) qui représente parfois plus de 20 % de reads (Figure 4). Sur ces deux sites, l'extraction a été réalisée sur des glandes digestives mis à -80°C directement versus RNAlater à température ambiante pour les autres échantillons. Il semblerait que le RNAlater pénètre mal les mitochondries d'où la sur-représentation de la CYPB dans les glandes conservées à -80°C.

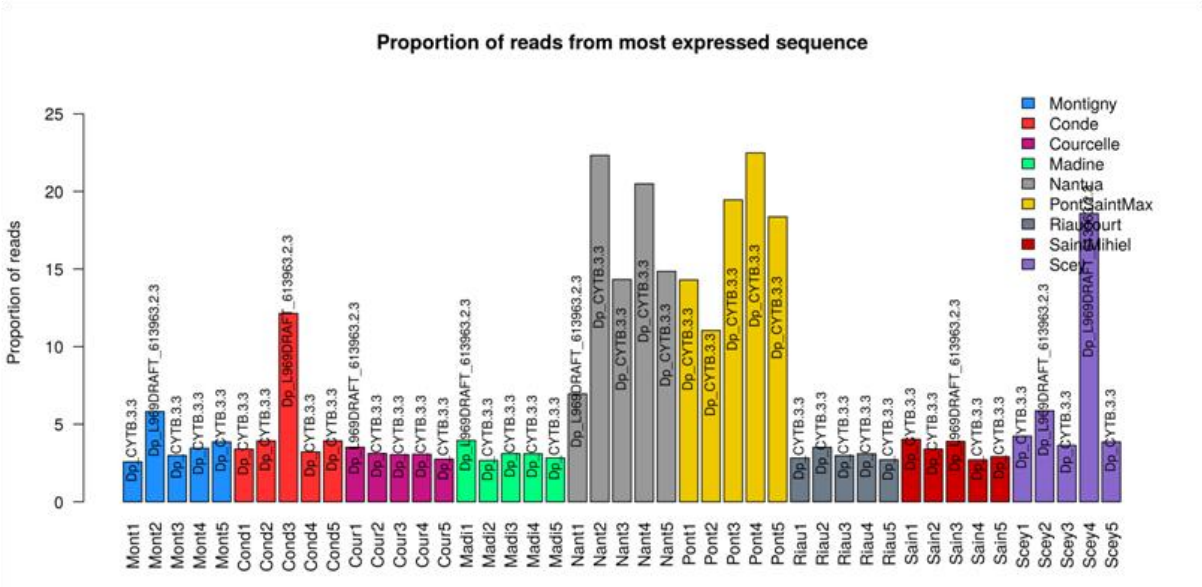


Figure 4 : Pourcentage des reads du contigs le plus exprimé par échantillon.

Séquence la plus exprimée (hors CYPB) : *Dp_L969DRAFT_613963.2.3* → Hypothetical protein L969DRAFT_613963, partial, *Mixia osmundae*. Deux échantillons concernés : *Cond3* et *Scey4*

- 44 538 contigs avec un N50 à 3 094 et un N90 à 1 244. Taille moyenne de librairies : 71 millions de séquences.
- 58,7 % des contigs annotés : contigs avec Goterms = 35 %, contigs avec KEGG = 22 %

2.3. Dreissena bugensis rostriformis

Sept sites ont été validés : Courcelles, Jouy aux Arches, Pont-Sainte-Maxence, MontignyPE (Moselle canalisée côté plan d'eau), MontignyC (canal), Seurre et Saint-Mihiel.

Séquence la plus exprimée : *Db_LOC106867323* → Predicted: Low Quality Protein: transcriptional enhancer factor TEF-1-like, *Octopus bimaculoides*. Présente à plus de 10 % dans 7 échantillons (Figure 5).

- 49 679 contigs avec un N50 à 2 674 et un N90 à 1 016. Taille moyenne de librairies : 74 millions de séquences.
- 57,2 % des contigs annotés : contigs avec Goterms = 34 % contigs avec KEGG = 21 %

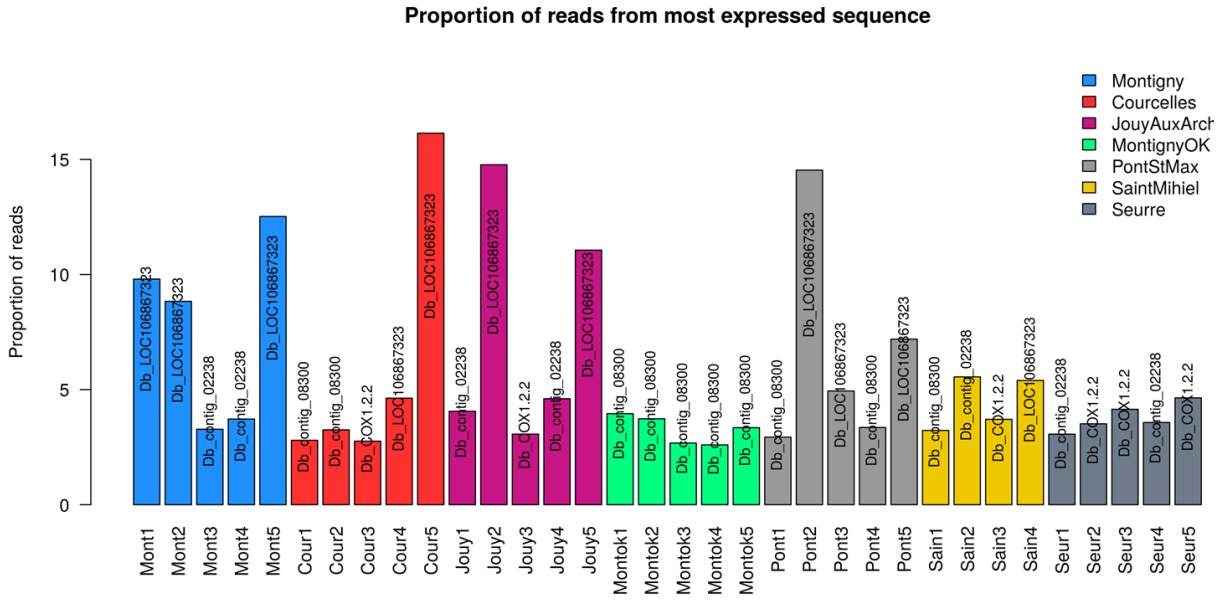


Figure 5: Pourcentage des reads du contigs le plus exprimé par échantillon.

Principaux Résultats

1. Analyses différentielles

Les données ont été analysées avec DESeq2 et edgeR via le package SARTools développé par l'équipe PF2 de l'Institut Pasteur. La normalisation sélectionnée pour edgeR est celle des *Trimmed Mean of M values* (TMM). Le seuil de *False Discovery Rate* (FDR) choisi est de 0.05 (Benjamini-Hochberg).
Vingt et une comparaisons ont été réalisées entre les sites pour chaque espèce, afin d'identifier le nombre de contigs différentiellement exprimés au moins une fois (Figure 6).

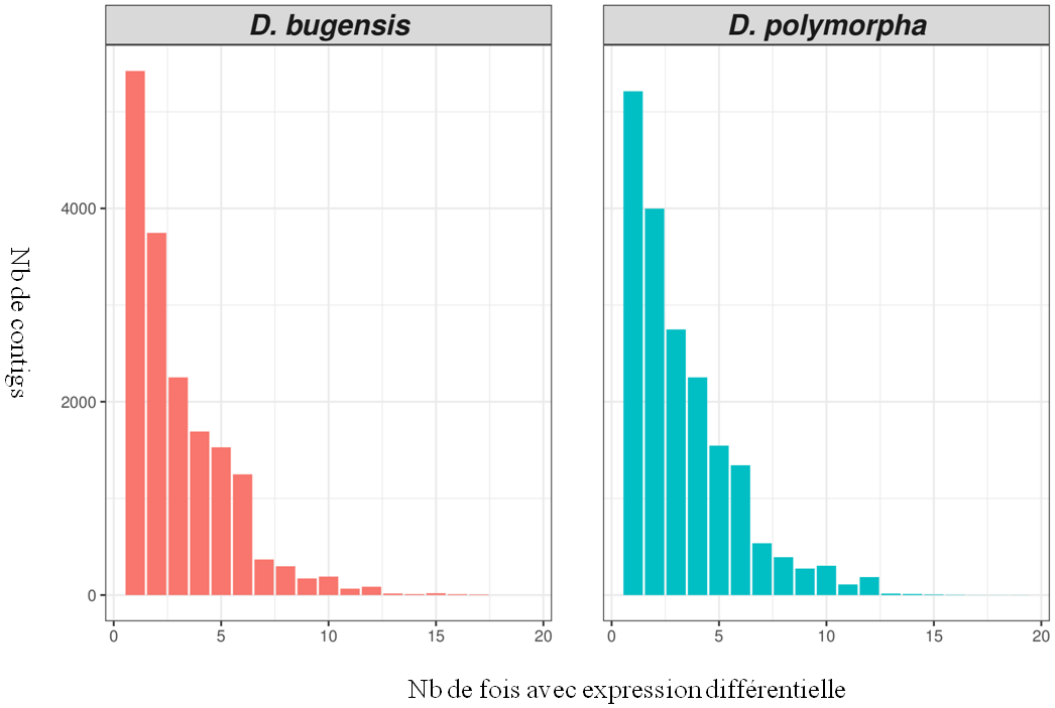


Figure 6 : histogramme de distribution des expressions différentielles

Les analyses suivantes ont été réalisées avec les 2 packages, en se focalisant dans cette première étape sur l'ensemble des contigs identifiés par l'expression différentielle.

1.1 Dreissena polymorpha

DESeq2 : Les analyses réalisées avec DESeq2 montrent que ~18 938 contigs sont différentiellement exprimés au moins dans une des 21 comparaisons. Ces contigs ont été sélectionnés afin de réaliser des analyses discriminantes. Ces ACP sont réalisées sur des données brutes (*raw.counts*) après transformation VST (*Variance Stabilizing Transformation*) sous DESeq2 (Figure 7).

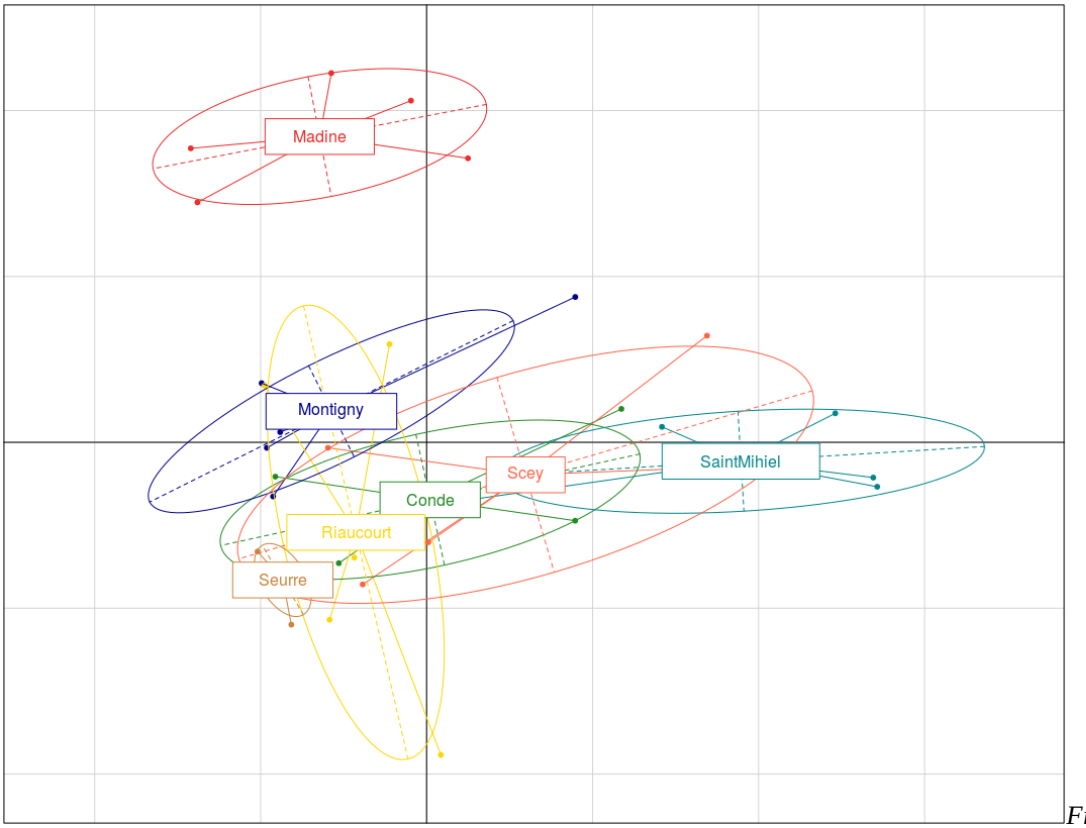
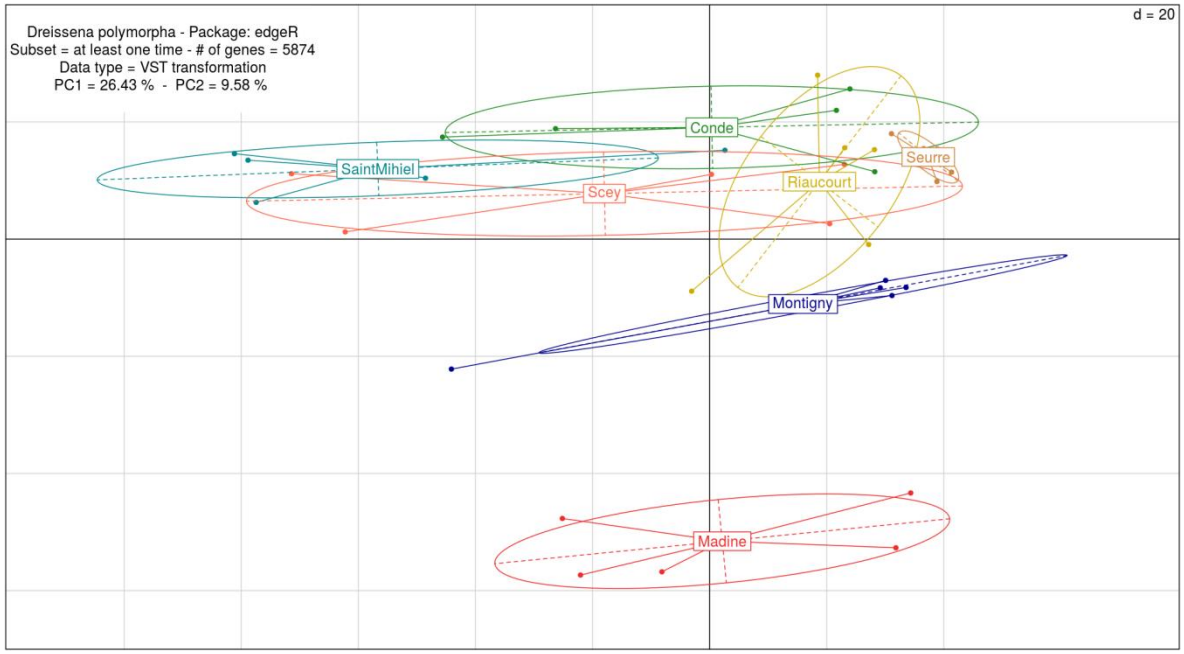


Figure 7: ACP des contigs exprimés au moins une fois sous DESeq2 chez D. polymorpha

edgeR: edgeR est plus conservatif que DESeq2 et applique avant les analyses statistiques une étape de filtration. Cette étape enlève ici 2 929 contigs qui ne satisfont pas la condition suivante : au moins 1 compte par million dans au moins 5 échantillons (low counts). Parmi les contigs restant, 5 874 sont sont différentiellement exprimés au moins une fois (Figure 8).



Dans les 2 cas, l'ACP

Figure 8: ACP des contigs exprimés au moins une fois sous edgeR chez D. polymorpha

ne permet de ne différencier que le site du Lac de Madine, qui est pourtant en tous points similaire à Scey et Saint-Mihiel sur le plan de la physico-chimie. Les autres sites, malgré les expressions différentielles nombreuses, ne sont pas bien différenciés sur le plan factoriel. Il est également surprenant de constater que la méthode EdgeR, qui élimine les évènements rares, produit un écrasement des données sur l'axe 2, l'ensemble de la variance semble être absorbée par l'axe 1.

1.2 Dreissena bugensis rostriformis

DESeq2 : Les analyses réalisées avec DESeq2 montrent que ~17 132 contigs sont différentiellement exprimés au moins dans une des 21 comparaisons (Figure 9).

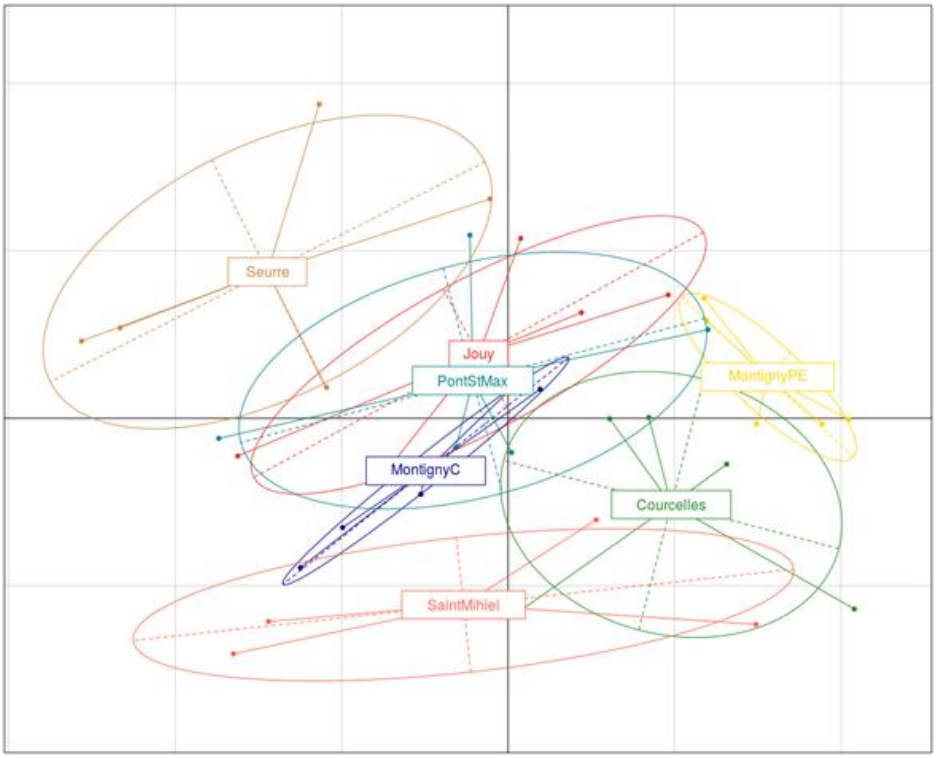


Figure 9: ACP des contigs exprimés au moins une fois sous DESeq2 chez D. bugensis

edgeR : 4 520 contigs ont été filtrés (low counts). 5 860 sont sont différentiellement exprimés au moins une fois (Figure 10).

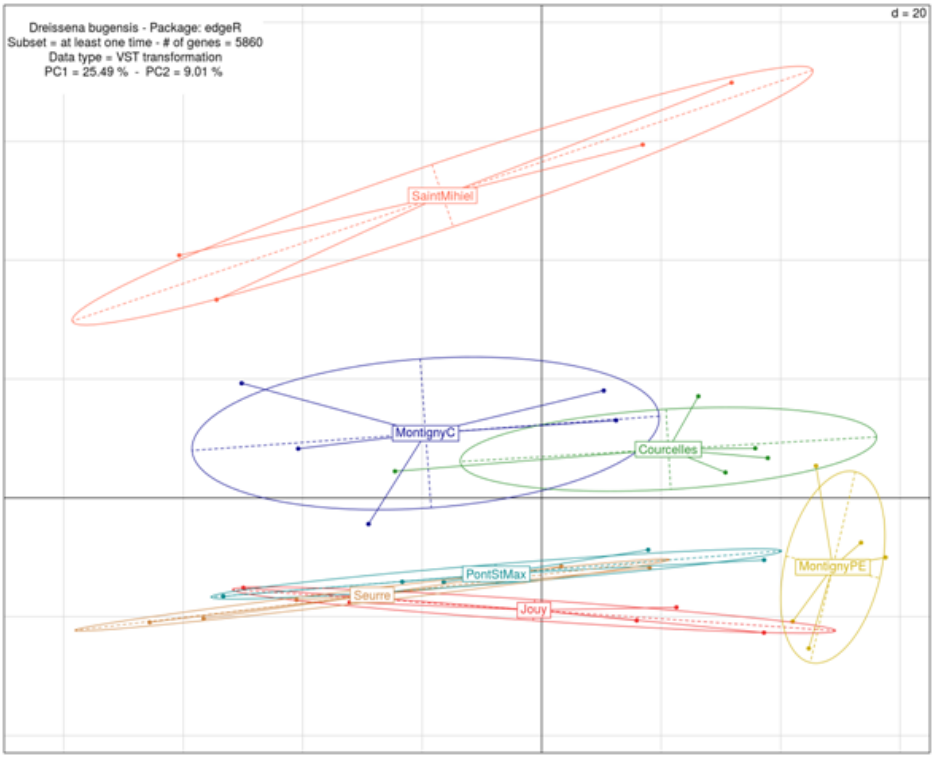


Figure 10: ACP des contigs exprimés au moins une fois sous DESeq2 chez D. bugensis

Pour *D. r. bugensis*, la méthode EdgeR produit de meilleurs résultats que DESeq2, et parvient à mieux séparer les sites en utilisant moins d'information.

2. Set maximisant la discrimination des sites

L'objectif de cette partie est d'obtenir un set de 1 000 gènes permettant une discrimination optimale des sites. Afin d'obtenir ce set, deux filtrations sont réalisées. Pour cette sélection, les données d'entrée sont les contigs filtrés par edgeR (41 610 pour *D. polymorpha* et 45 159 pour *D. r. bugensis*) pour lesquels les *cpm()* normalisés sont recalculés. Les étapes de la filtration :

1. Calcul du coefficient de variation (SD/mean*100) pour chaque contig pour chaque site (= pour chaque contig, un CV par site).
2. Application d'un seuil (*cutoff*) de façon à ce qu'uniquement les contigs ayant un CV inférieur à ce seuil dans x groupes soient sélectionnés.
3. Calcul du coefficient de variation sur les valeurs moyennées par groupe (= un CV par contig)
4. Les contigs sont triés par ordre décroissant de ce second CV.

Pour déterminer le CV à appliquer pour le premier *cutoff* (2), plusieurs seuils croissants de CV sont testés (entre 30 % et 41 % pour x = 7). Les 3000 premiers gènes sont utilisés successivement pour

faire un clustering hiérarchique (Figure 11). Le CV minimal permettant une clusterisation optimale des sites lorsque les 1000 gènes les plus variants sont utilisés est de 33 %.

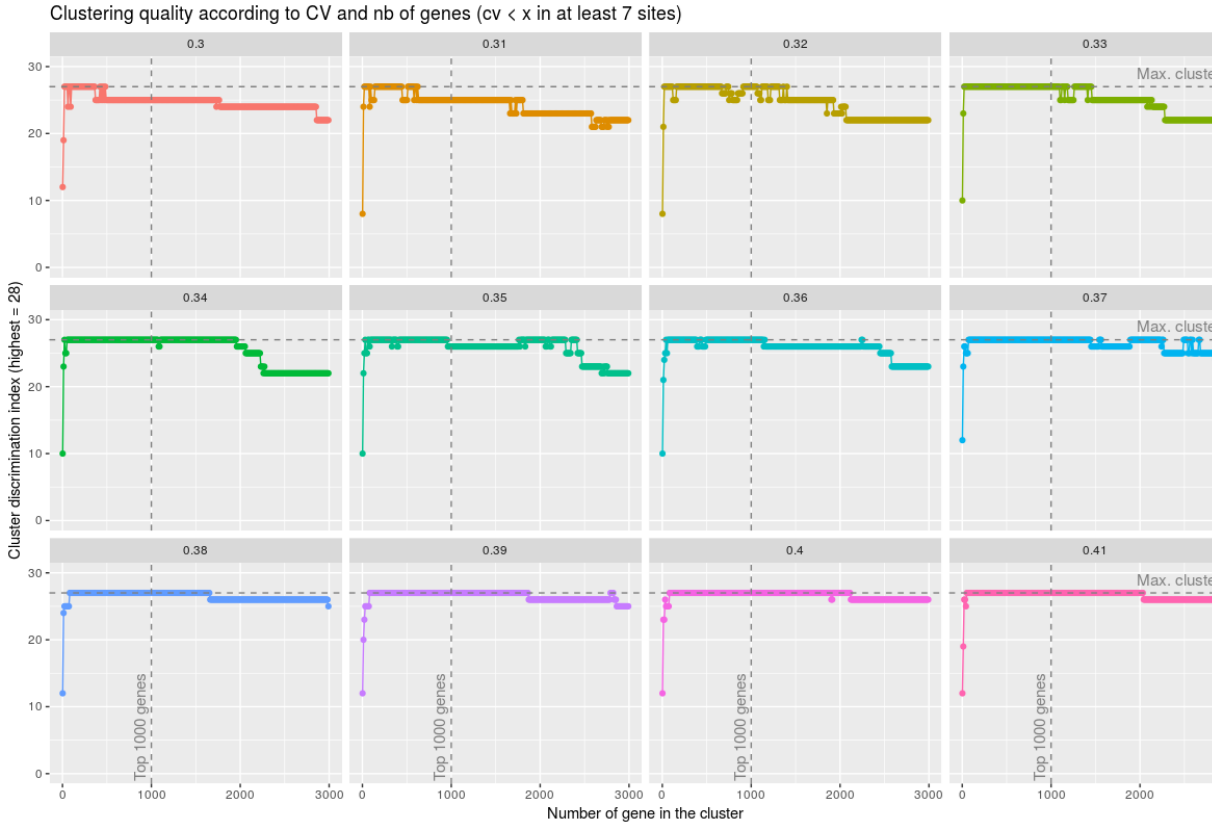


Figure 11: Qualité de la clusterisation hiérarchique (D. polymorpha). Pour chaque CV appliqué aux 7 sites,une clusterisation hiérarchique est réalisée sur les gènes classés du plus variant au moins variant. Le pas des gènes :10 par 10.

Néanmoins, il faut un CV de 37 % et au délà pour avoir une stabilité du *Max. clustering* entre 0 et 1000 gènes. La vérification de la stabilité du clustering en prenant les gènes par pas de 1 (au lieu de 10 par 10 dans la figure) montre que le clustering à un *cutoff* de 37 % et au delà est optimal. Lorsque la sélection des contig est moins stringente (x < 7), la qualité du clustering augmente à CV équivalent. Enfin, lorsque la même filtration est effectuée sur 5 sites (exclusion de 2 site au hasard), la qualité du clustering augmente également à CV équivalent.

Le meilleur *cutoff* à appliquer sur les données *D. polymorpha* doit donc être supérieur à 37 % (filtration la plus stringente). Arbitrairement, le *cutoff* sélectionné pour la suite est donc de 40 %.

Concernant *D. bugensis*, l'application des paramètres identiques (à *D. polymorpha*) ne permet pas d'obtenir une clusterisation maximale, même à un *cutoff* supérieur (> 60 %, Fig. 12).

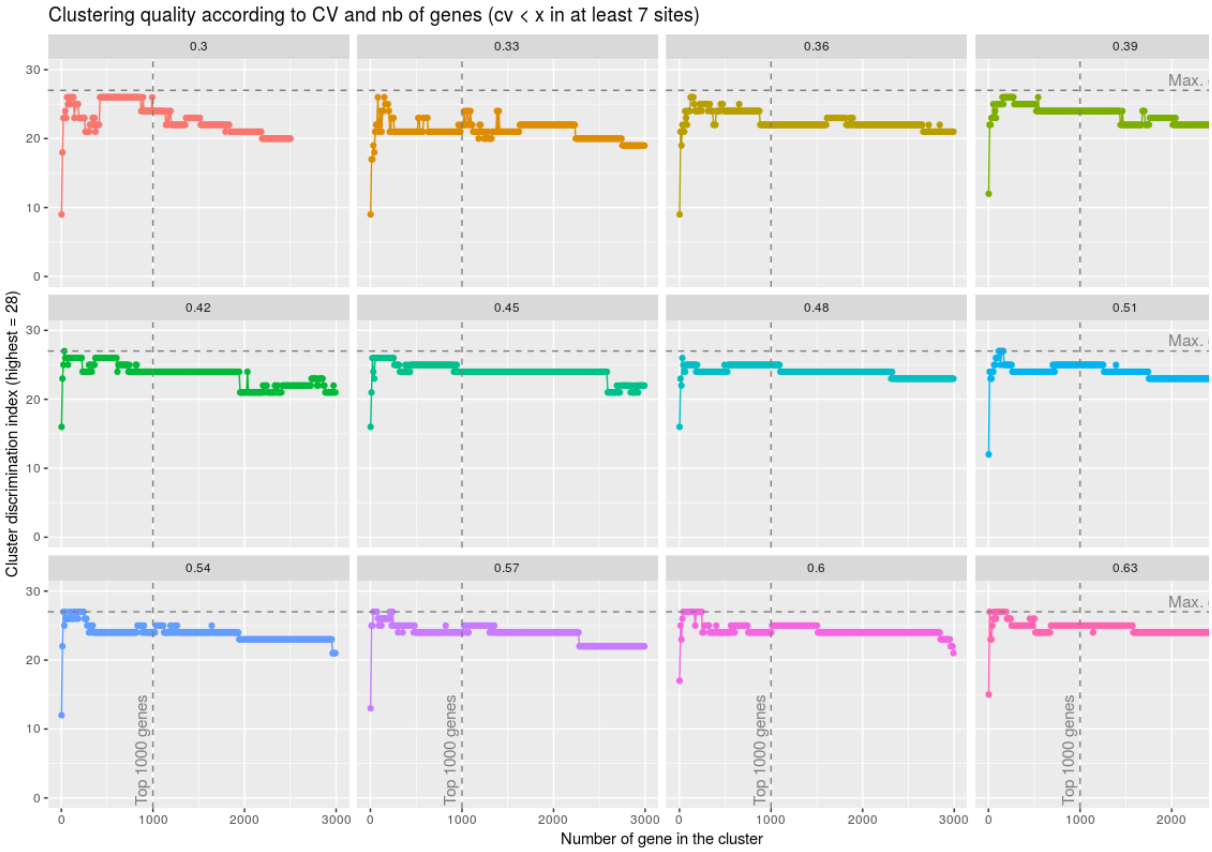


Figure 12: Qualité de la clusterisation hiérarchique (D. r. bugensis). Pour chaque CV appliqué aux 7 sites, uneclusterisation hiérarchique est réalisée sur les gènes classés du plus variant au moins variant. Le pas des gènes : 10par 10.

Pour obtenir une séparation optimale des sites, il faut diminuer à la fois le x (*i.e.* le nombre de groupe minimum où le *cutoff* est respecté) et le *cutoff* lui même. Pour un x de 6 et 5, les *cutoffs* permettant d'obtenir 1000 gènes discriminant sont respectivement 58 % et 45 % (Fig. 13).

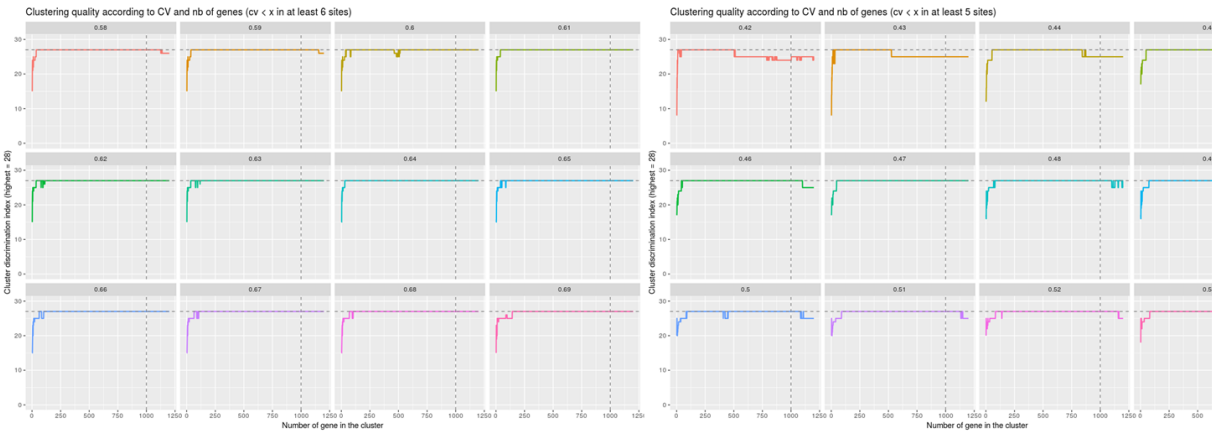


Figure 13: Qualité de la clusterisation hiérarchique (D. r. bugensis). Un x de 6 (gauche) et de 5 (droite) estappliqué. Le pas de gènes : 1 par 1. Attention, le nombre de gènes en abscisse est de 1200 et non 3000 comme surles autres graphiques.

3. Discrimination des sites sur la base des clusters

Les clusters de gènes construits ont ensuite été retravaillés, pour parvenir à discriminer les sites en utilisant des variables plus synthétiques.

Pour chaque espèce, une matrice de corrélation des niveaux d'expression entre l'ensemble des individus, tous sites confondus, a été produite, mettant en avant des coefficients de corrélation très élevés, et confirmant ainsi l'homogénéité des niveaux d'expression au sein de chaque cluster.

L'expression moyenne de l'ensemble des contigs par cluster pour chaque individu a ensuite été calculée, permettant de travailler, pour chaque espèce, sur une matrice réduite de 20 variables explicatives (la moyenne de chaque cluster) et 35 individus (5 individus x 7 sites).

La discrimination des sites a ensuite été réalisée par analyse PLS-DA (Partial Least Square Discriminant Analysis) (Figures 14 et 15).

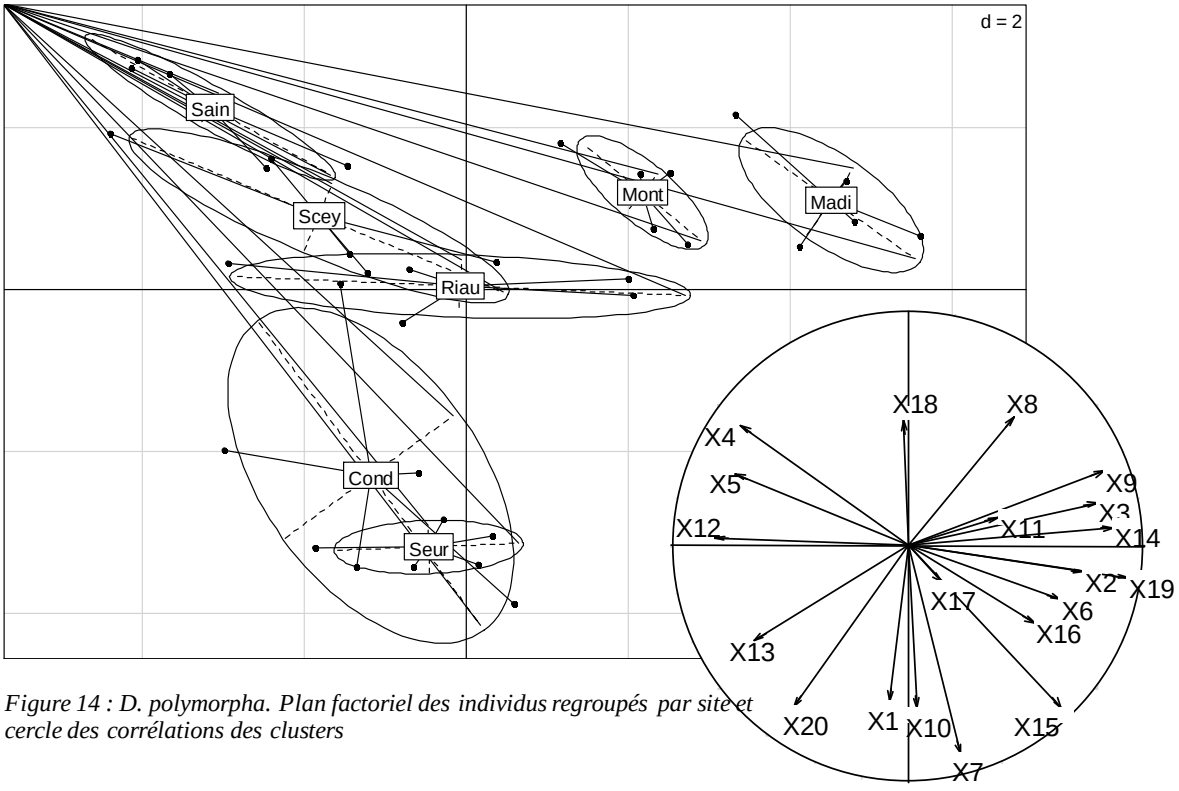


Figure 14 : D. polymorpha. Plan factoriel des individus regroupés par site et cercle des corrélations des clusters

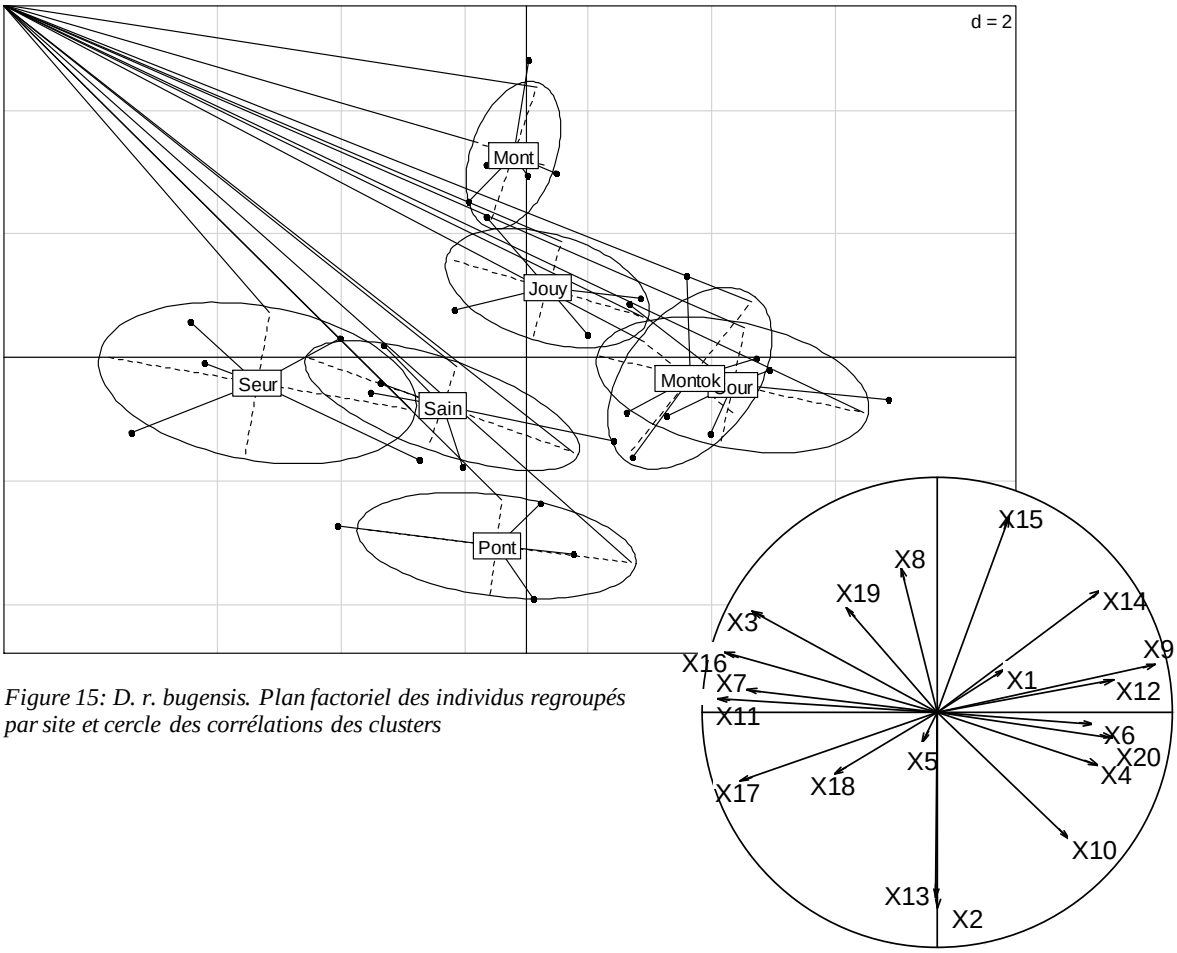


Figure 15: D. r. bugensis. Plan factoriel des individus regroupés par site et cercle des corrélations des clusters

Ces résultats mettent en évidence la très bonne capacité de cette variable synthétique à différencier les populations. La discrimination est plus efficace qu'en se basant sur les biomarqueurs biochimiques. Nous avons également testé la discrimination en ne se basant plus sur une variable synthétique (qui nécessite donc d'avoir au préalable quantifié l'expression de 1000 contigs), mais en opérant un choix aléatoire d'un contig par cluster. Les résultats sont aussi bons, c'est à dire que 20 contigs seulement ont la même performance de discrimination que nos 20 moyennes. Cependant, selon le tirage effectué, la distribution relative des populations sur le plan factoriel est modifiée.

Chaque combinaison de 20 contigs est donc susceptible de révéler un patron de réponse particulier, potentiellement spécifique d'une nature ou d'un niveau de stress. Le corollaire est que pour l'instant, nous n'avons pas encore pu identifier la combinaison qui permettrait de rendre compte de la contamination du milieu, ou de l'état de santé des organismes.

Conclusion et Perspectives

Une approche basée sur la transcriptomique pour réaliser de la biosurveillance apparait, au vu des résultats de cette approche exploratoire préliminaire, tout à fait envisageable.

Nous avons pu mettre en évidence la capacité d'un faible nombre de réponses, facilement mesurables à l'échelle individuelle, à discriminer les sites.

Les limites actuelles de l'approche sont :

- La difficulté à attribuer une signification biologique aux clusters → limite liée aux approches sur des organismes non séquencés

- La difficulté à établir un lien évident entre la physico-chimie (contamination) et les niveaux d'expressions

- Le faible nombre de populations étudié et l'absence de validation externe pour les sets de gènes regroupés dans les clusters

Les perspectives associées à ces limites sont :

- Identifier des gènes candidats par cluster, dont la fonction est identifiée, pour réaliser cette discrimination des sites et lui attribuer une signification biologique.

- La possibilité de développer des puces à ADN, étant donné le faible nombre de gènes nécessaires pour différencier les sites

- La nécessité de tester les gènes identifiés sur d'autres panels de sites, voire sur des stress standardisés au labo

Une autre interrogation est la valeur prédictive de ces approches moléculaires à l'échelle des populations. Par définition, nous avons réalisé un biomonitoring passif, donc uniquement sur des populations qui parviennent à se maintenir dans les sites étudiés. Ce sont donc systématiquement des populations pérennes, non menacées d'extinction (en tout cas nous n'avons noté aucun signe avant-coureur sur les 2 ans de suivi). Les modifications des niveaux d'expressions mesurées peuvent indiquer la mise en place de systèmes de défense, un coût énergétique, mais aucunement être indicatrice d'une menace pour le maintien des populations.