ÉTUDE ET DÉCONTAMINATION DU TRANSCRIPTOME *DE NOVO* DU NÉMATODE DORÉ *GLOBODERA ROSTOCHIENSIS*

par

Joël Lafond Lapalme

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

FACULTÉ DES SCIENCES
UNIVERSITÉ DE SHERBROOKE

Sherbrooke, Québec, Canada, juin 2016

Le 10 juin 2016

Le jury a accepté le mémoire de Monsieur Joël Lafond-Lapalme dans sa version finale.

Membres du jury

Docteur Benjamin Mimee
Directeur de recherche
Agriculture et Agroalimentaire Canada
Centre de recherche de St-Jean-sur-Richelieu

Professeur Peter Moffett
Codirecteur de recherche
Département de biologie

Professeur Sébastien Rodrigue
Évaluateur interne
Département de biologie

Professeur Pierre-Étienne Jacques
Président-rapporteur
Département de biologie

SOMMAIRE

Le nématode doré, *Globodera rostochiensis*, est un nématode phytoparasite qui peut infecter des plantes agricoles telles la pomme de terre, la tomate et l'aubergine. En raison des pertes de rendement considérables associées à cet organisme, il est justifiable de quarantaine dans plusieurs pays, dont le Canada. Les kystes du nématode doré protègent les œufs qu'ils contiennent, leur permettant de survivre (en état de dormance) jusqu'à 20 ans dans le sol. L'éclosion des œufs n'aura lieu qu'en présence d'exsudats racinaires d'une plante hôte compatible à proximité. Malheureusement, très peu de connaissances sont disponibles sur les mécanismes moléculaires liés à cette étape-clé du cycle vital du nématode doré.

Dans cet ouvrage, nous avons utilisé la technique *RNA-seq* pour séquencer tous les ARNm d'un échantillon de kystes du nématode doré afin d'assembler un transcriptome *de novo* (sans référence) et d'identifier des gènes jouant un rôle dans les mécanismes de survie et d'éclosion. Cette méthode nous a permis de constater que les processus d'éclosion et de parasitisme sont étroitement reliés. Plusieurs effecteurs impliqués dans le mouvement vers la plante hôte et la pénétration de la racine sont induits dès que le kyste est hydraté (avant même le déclenchement de l'éclosion).

Avec l'aide du génome de référence du nématode doré, nous avons pu constater que la majorité des transcrits du transcriptome ne provenaient pas du nématode doré. En effet, les kystes échantillonnés au champ peuvent contenir des contaminants (bactéries, champignons, etc.) sur leur paroi et même à l'intérieur du kyste. Ces contaminants seront donc séquencés et assemblés avec le transcriptome *de novo.*

Ces transcrits augmentent la taille du transcriptome et induisent des erreurs lors des analyses post-assemblages. Les méthodes de décontamination actuelles utilisent des alignements sur des bases de données d'organismes connus pour identifier ces séquences provenant de contaminants. Ces méthodes sont efficaces lorsque le ou les contaminants sont connus (possède un génome de référence) comme la contamination humaine. Par contre, lorsque le ou les contaminants sont inconnus, ces méthodes deviennent insuffisantes pour produire un transcriptome décontaminé de qualité.

Nous avons donc conçu une méthode qui utilise un algorithme de regroupement hiérarchique des séquences. Cette méthode produit, de façon récursive, des sous-groupes de séquences homogènes en fonction des patrons fréquents présents dans les séquences. Une fois les groupes créés, ils sont étiquetés comme contaminants ou non en fonction des résultats d'alignements du sous-groupe. Les séquences ambiguës ayant aucun ou plusieurs alignements différents sont donc facilement classées en fonction de l'étiquette de leur groupe. Notre méthode a été efficace pour décontaminer le transcriptome du nématode doré ainsi que d'autres cas de contamination. Cette méthode fonctionne pour décontaminer un transcriptome, mais nous avons aussi démontré qu'elle a le potentiel de décontaminer de courtes séquences brutes. Décontaminer directement les séquences brutes serait la méthode de décontamination optimale, car elle minimiserait les erreurs d'assemblage.

Mots-clés : Nématode doré, *Globodera rostochiensis*, éclosion, transcriptome, assemblage *de novo*, gènes différentiellement exprimés, décontamination, Model-based Categorical Sequence Clustering, MCSC

REMERCIEMENTS

J'aimerais remercier mon directeur Dr Benjamin Mimee, il a su me motiver et me guider pour améliorer mes aptitudes en recherche durant toute ma maîtrise. Merci à Marc-Olivier Duceppe qui a été mon mentor durant les deux dernières années ainsi qu'à tous les employés et étudiants de l'équipe de nématologie du CRDH. Merci également à mon codirecteur Peter Moffett et à mes deux conseillers Sébastien Rodrigue et Pierre-Étienne Jacques pour leurs précieux conseils.

Un merci spécial à mes parents Jeannine Lapalme et Pierre Lafond qui m'ont soutenu moralement (et financièrement) durant mes 5 années à l'Université de Sherbrooke. Finalement, j'aimerais remercier ma copine Julie qui m'a soutenu et qui m'a aidé à corriger plusieurs de mes textes ainsi que tous mes amis (TC).

TABLE DES MATIÈRES

LISTE DES ABRÉVIATIONS

NKPT : Nématode à kyste de la pomme de terre

PCN : *Potato cyst nematode*

J2 : Nématode juvénile de 2^e stade larvaire

GDE : Gène différentiellement exprimé

DEG : *Differentially expressed gene*

PRD : *Potato root diffusate*

TRD : *Tomato root diffusate*

MCSC : *Model-based categorical sequence clustering*

ADN : Acide désoxyribonucléique

ADNc : Acide désoxyribonucléique complémentaire

ARN : Acide ribonucléique

ARNm : Acide ribonucléique messager

RNA-seq : Séquençage à haut débit de l'ARN

qRT-PCR : *quantitative Real-Time polymerase chain reaction*

LISTE DES TABLEAUX

LISTE DES FIGURES

CHAPITRE 1
INTRODUCTION

1.1. Nématode à kyste

Les nématodes sont présents dans tous les types d'écosystèmes, que ce soit dans l'eau, dans le sol, sous un climat nordique ou tropical. Certains nématodes sont saprophytes (libres) alors que d'autres peuvent parasiter les animaux, les insectes ou les plantes. On rapporte plus de 4100 espèces de nématodes parasites des plantes (Decraemer and Hunt 2006) et les pertes économiques annuelles causées par ces nématodes sont estimées à 80 milliards $ US (Nicol et al. 2011). À ce niveau, les deux groupes les plus dommageables sont les nématodes à galles et les nématodes à kyste, tous membres de la famille des *Heteroderidae*. Les nématodes à kyste comptent 115 espèces (Turner & Subbotin 2013) parmi ceux-ci, les plus importants sont *Heterodera schachtii*, le nématode à kyste de la betterave, *Heterodera glycines*, le nématode à kyste du soya et les nématodes à kyste de la pomme de terre (NKPT), *Globodera pallida* (nématode à kyste pâle) et *G. rostochiensis* (nématode doré). Les NKPT sont des organismes de quarantaine qui causent des pertes de rendements estimées à 9 % de la production mondiale de pomme de terre (Jones et al. 2013). Ils s'attaquent à la famille des *Solonaceae* comprenant des plantes agricoles telles la pomme de terre, la tomate et l'aubergine (Bélair 2005). Le nématode doré, *G. rostochiensis,* est originaire d'Amérique du Sud (Evans et al. 1975), il est maintenant présent sur tous les continents. Il a été détecté dans 75 pays dont le Canada où il a été détecté sur l'île de Vancouver en 1965, à Terre-Neuve en 1962 et plus récemment à St-Amable en 2006 (Sun et al. 2007). Le nématode doré parasite les racines des pommes de terre pour se nourrir et se reproduire. Cela provoque un stress et réduit l'apport de nutriments disponibles pour la plante. Les symptômes visibles (réduction de croissance, jaunissement du feuillage), lorsque présents, ne sont pas spécifiques au NKPT et peuvent facilement être confondus avec d'autres problèmes. Par contre, dans le sol, la présence de kystes sur les racines confirmera l'atteinte par un NKPT. Le développement des tubercules sera aussi réduit. Tout cela causera d'importantes pertes de rendement, pouvant aller jusqu'à 90 % dans certaines conditions (Nicol et al. 2011).

1.1.1. Cycle de vie

Le kyste du nématode doré de la pomme de terre est formé du corps durci de la femelle. Cette protection permet aux œufs qu'il contient de survivre, en état de dormance, plus de 20 ans dans le sol en attendant des conditions favorables pour son développement. Chaque kyste contient entre 200 et 500 œufs chez *G. rostochiensis* (Evans et Stone 1977). L'éclosion des œufs est induite par le contact avec l'exsudat racinaire d'une plante hôte. La survie du kyste en état de dormance ainsi que son mécanisme d'éclosion spécifique sont des méthodes de survie à long terme qui font du nématode doré une espèce difficile à éradiquer d'un champ. Suite à l'éclosion, le nématode du deuxième stade juvénile (J2) migre vers les racines de la plante hôte, il est guidé par des stimuli provenant de la plante comme l'exsudat racinaire. Le nématode J2 ne peut se nourrir tant qu'il n'a pas pénétré la racine et pris le contrôle d'un groupe de cellules de la plante. Afin de se nourrir des nutriments de l'hôte, il formera un site de nutrition en 6 à 11 jours (Robinson et al. 1987). Pour établir son site de développement, le J2 doit d'abord pénétrer la racine et parcourir les cellules à l'aide de son stylet et de la sécrétion d'enzymes (effecteurs) spécifiques au stade J2 (Tytgat et al. 2002). Son site de nutrition appelé syncytium est composé de centaines de cellules (Bohlmann and Sobczak 2014). Une fois son site établi, le J3 devient sédentaire et se développe jusqu'au moment de la reproduction. Le mâle, qui a une forme beaucoup plus allongée, se déplace alors hors de la racine pour aller féconder la femelle dont le corps sailli à l'extérieur. Lorsque la femelle atteint sa maturité, elle meurt et son corps durci forme le kyste. Pour le nématode doré, il n'y a généralement qu'une seule génération par année en climat tempéré (Whitehead 1997).

1.1.2. Moyens de contrôle

La relation intime du nématode doré avec la plante, de même que sa capacité à survivre pour de longues périodes dans le sol en font un parasite efficace et difficile à éradiquer. Les méthodes de lutte actuelles comprennent les nématicides et les fumigants, néfastes pour l'environnement et maintenant interdits dans plusieurs pays; les rotations de culture, qui doivent être réalisées sur plusieurs années (6 et plus) pour être efficaces et l'utilisation de cultivars résistants. Ces derniers reposent sur la capacité de certaines lignées de pomme de terre de reconnaître le ravageur. Cette détection se fera par l'entremise du produit d'un gène de résistance de la plante qui reconnaîtra le produit d'un gène d'avirulence exprimé par le nématode. Cette reconnaissance entraînera une réponse hypersensible causant la mort des cellules végétales infestées. Ces gènes de résistance ne seront donc efficaces que contre les populations de nématodes exprimant le gène d'avirulence correspondant. Pour le nématode doré, cinq pathotypes (Ro1 – Ro5) ont été identifiés en fonction de leur développement sur différents génotypes de pomme de terre (Kort et al. 1977). Cependant, l'assignation de certaines populations à un pathotype peut être ambigüe (Nijboer and Parlevliet 1990), c'est pourquoi la compréhension des interactions entre le nématode et son hôte est essentielle afin d'améliorer les méthodes d'identification et de pouvoir développer des cultivars résistants efficaces.

1.1.3. Éclosion

L'éclosion est une étape clé chez les NKPT. Les larves dans les œufs se développent jusqu'au stade J2, ensuite elles stoppent leur développement pour tomber en état de dormance jusqu'à l'éclosion. Le mécanisme de survie des NKPT permet aux larves dans le kyste de survivre pendant plusieurs années en attendant des conditions optimales pour éclore. Cette survie à long terme est rendue possible par la synthèse de tréhalose. Ce sucre contenu à l'intérieur de l'œuf permettra à la larve de survivre à une déshydratation presque totale lors de la diapause (Atkinson et al. 1987). L'éclosion peut parfois se produire spontanément, mais généralement, elle est provoquée par le contact du kyste avec l'exsudat racinaire d'une plante (Perry et al. 2002). Perry and Beane (1982) ont montré que seulement cinq minutes d'exposition à l'exsudat étaient nécessaires pour provoquer l'éclosion. La larve reste donc en diapause jusqu'au signal d'éclosion qui provoque une réduction rapide des concentrations de tréhalose. Cette réaction modifie également la perméabilité de la membrane de l'œuf pour permettre à la larve de s'hydrater et de s'activer. À l'aide de son stylet, le nématode perce l'œuf et le kyste pour en sortir et se déplacer vers les racines de la plante hôte. En plus du stylet, le nématode utilise des enzymes sécrétés comme la chitinase (Cotton et al. 2014) pour dégrader la coquille et faciliter ses déplacements.

Le contrôle de l'éclosion du nématode est une des voies possibles pour le développement de nouvelles méthodes de lutte. La stimulation de l'éclosion par l'utilisation de plantes non-hôtes produisant des exsudats compatibles, comme la morelle de Balbis, *Solanum sisymbriifolium,* est déjà utilisée (Timmermans et al. 2007). Par contre, cette méthode ne génère aucun produit vendable et n'est donc pas économiquement viable. Certains auteurs ont également tenté de déclencher l'éclosion à l'aide d'exsudats racinaires produits en laboratoire. Blaauw et al. (2001)

et Snyder (2011) ont obtenu une certaine efficacité, mais l'exsudat de pomme de terre est très complexe (Byrne et al. 1998). Il contient un mélange de stimulateurs et d'inhibiteurs qui le rendent difficile à reproduire à partir d'une plante ou par synthèse. D'autres auteurs ont suggéré l'induction de l'éclosion via son contrôle génétique et ont entrepris d'étudier l'expression des gènes en cause (Jones et al. (1997). Par contre, la sensibilité et la puissance des techniques utilisées à l'époque n'ont pas permis d'identifier de tels gènes.

1.2. Études transcriptomiques

1.2.1. *RNA-seq*

Les études transcriptomiques permettent d'étudier les éléments fonctionnels du génome. Plusieurs méthodes ont été développées pour faire l'analyse de l'expression des transcrits d'un organisme. Toutes ces méthodes utilisent l'ADN complémentaire (ADNc) qui est un ADN synthétisé à partir d'un brin d'ARN. Les *microarrays* (puce d'ADN), par exemple, utilisent l'hybridation d'ADN complémentaire (ADNc), pour évaluer l'expression de plusieurs transcrits (DeRisi et al. 1997). Cette méthode est limitée à l'analyse des gènes qui sont déjà connus dans la littérature. Une autre approche consiste à séquencer directement l'ADN complémentaire avec la technologie de séquençage Sanger (Velculescu et al. 1995). Par contre, cette technologie est contraignante, car elle est très coûteuse pour évaluer une grande quantité de gènes. Avec l'émergence du séquençage de nouvelle génération, le séquençage à haut débit d'ARN (*RNA-seq)* est maintenant un incontournable pour les études transcriptomiques. Cette nouvelle méthode de séquençage permet d'étudier le profil d'expression de l'ensemble du transcriptome de façon quantitative.

Elle consiste à extraire l'ARNm puis à le convertir en une librairie de fragments d'ADN complémentaires avec des adaptateurs de séquençages au bout de chaque fragment. Ces fragments sont ensuite séquencés sur une plateforme de séquençage à haut débit (Wang et al. 2009). Les millions de séquences résultantes peuvent être assemblées pour former un transcriptome ou bien simplement alignées sur un transcriptome de référence. L'alignement des séquences servira ensuite à quantifier l'expression de chaque transcrit. De plus, un des avantages de cette méthode est qu'elle ne requiert pas de génome de référence. Un assemblage *de novo* du transcriptome peut être fait directement avec les séquences issues du *RNA-seq*.

Il existe une multitude d'analyses qui peuvent être réalisées à partir de données *RNA-seq*, par exemple : l'étude de l'épissage alternatif dans différentes conditions, l'identification de nouveaux gènes/transcrits/isoformes, l'identification de marqueurs moléculaires dans les régions transcrites et l'identification de gènes différentiellement exprimés (GDEs). L'étude des GDEs à partir de données *RNA-seq* est très populaire. Les GDEs sont des gènes dont l'expression est induite ou réprimée entre deux ou plusieurs traitements. Plusieurs méthodes statistiques ont été développées pour détecter des GDEs. Les plus populaires sont DESeq (Anders and Huber 2010, Love et al. 2014), edgeR (Robinson et al. 2010) et Cuffdiff (Trapnell et al. 2013) qui utilisent une distribution binomiale négative pour modéliser la variation d'expression. D'autres méthodes alternatives existent comme NOISeq (Tarazona et al. 2011) qui utilise une approche non paramétrique ainsi que BaySeq (Hardcastle and Kelly 2010) et BitSeq (Glaus et al. 2012) qui utilisent une approche bayésienne. Pour chaque expérimentation, les résultats entre les méthodes varient, il n'y a pas de consensus sur la meilleure méthode à utiliser, mais les méthodes les plus stables semblent être celles qui utilisent la distribution binomiale négative (Guo et al. 2013). Un test statistique selon l'approche choisi produit une *p-value* qui est la probabilité qu'un gène soit différentiellement exprimé lors du test alors qu'en réalité il ne l'est pas. La *p-value* est ensuite ajustée en fonctions du nombre de tests (un par gène) effectués.

Cet ajustement a pour effet d'augmenter la *p-value* pour contrer les résultats faux-positifs dû au hasard (Anders and Huber 2010).

1.2.2. Assemblage *de novo*

Tel qu'indiqué précédemment, l'assemblage *de novo* permet de reconstruire un transcriptome à partir des données *RNA-seq*, et ce, sans référence. La reconstruction des transcrits à partir de courtes séquences est une tâche très complexe qui requiert beaucoup de temps de calcul et de mémoire vive. Il existe plusieurs logiciels libres pour faire l'assemblage *de novo* de transcriptome : Trinity (Grabherr et al. 2011, Haas et al. 2013), Oases (Schulz et al. 2012), EBARDenovo (Chu et al. 2013) et SOAPdenovo-trans (Xie et al. 2014). Il y a aussi des logiciels commerciaux comme CLC Genomics Workbench (Qiagen) et SeqMan NGen (DNASTAR) capable de faire des assemblages *de novo*. Tous ces logiciels utilisent un algorithme basé sur les graphes de de Bruijn (Compeau et al. 2011) pour reconstruire les transcrits à partir des séquences de *RNA-seq*. L'assemblage du transcriptome est une étape importante, car il est la base de plusieurs analyses de *RNA-seq*, en particulier, l'identification de GDEs (Davidson and Oshlack 2014). Plusieurs facteurs peuvent influencer la qualité d'un transcriptome *de novo* (Baker 2012). Par exemple, peu importe le choix de l'assembleur, il y a toujours présence de chimères dans l'assemblage (Yang and Smith 2013). Les chimères sont des transcrits résultant de la fusion incorrecte de plusieurs gènes entre eux. Ces transcrits produisent des erreurs lors des analyses et augmentent le nombre de transcrits dans le transcriptome. Généralement, dans un transcriptome *de novo* il y a plusieurs transcrits pour représenter un gène (Davidson and Oshlack 2014). Ces transcrits redondants peuvent être de vrais isoformes biologiques, mais peuvent aussi provenir d'erreurs d'assemblage provoquées par des séquences répétées ou des variations génétiques à l'intérieur de l’échantillon. Un nombre trop élevé de transcrits *de novo* diminue la puissance statistique lors de l’analyse de GDE car la *p-value* est corrigée pour les tests multiples. De plus, il est difficile de correctement quantifier l’expression d’un gène lorsqu’il est représenté par plusieurs transcrits et les résultats d’analyses sont difficiles à interpréter (Davidson and Oshlack 2014). Il est préférable de corriger ces erreurs d’assemblage afin de maximiser la puissance statistique et de réduire le nombre de transcrits non informatifs. Pour corriger les chimères, il existe quelques méthodes comme EBARDenovo (Chu et al. 2013) qui détecte les chimères directement lors de l’assemblage et mRNAmarkup (Brendel and Standage) qui consiste à aligner les transcrits sur une base de données de protéines connues afin de séparer les chimères codant pour plus d’une protéine. Pour regrouper les transcrits semblables, les assembleurs Trinity et Oases possèdent une méthode qui regroupe les transcrits selon le graphe obtenu lors de l’assemblage. Davidson and Oshlack (2014) ont développé une nouvelle méthode qui regroupe les transcrits selon les séquences partagées. Cette méthode robuste utilise un test statistique en comparant les données d’expression de chaque transcrit afin de regrouper les isoformes, mais aussi les transcrits redondants en un *cluster* qui représente le gène.

1.3. Décontamination de sequences

1.3.1. Contamination de sequences

Sans génome de référence, il est difficile de valider si les séquences produites par un séquenceur appartiennent bel et bien à l’organisme à l’étude. La contamination est un autre facteur qui peut réduire la qualité d’un assemblage *de novo*. Lors d’une expérience de *RNA-seq*, un contaminant est un organisme, autre que l’organisme à l'étude, qui se retrouve dans l'échantillon lors de l'extraction d'ARN. Ces ARN provenant de contaminants seront séquencés et assemblés. Les transcrits et erreurs d'assemblage générés par les contaminants augmenteront la taille du transcriptome ce qui, comme mentionné ci-dessus, diminue la puissance statistique et rend l'interprétation des résultats plus difficile. En plus de la contamination par de mauvaises manipulations en laboratoire, il existe plusieurs cas ou la contamination d'un échantillon est inévitable. Par exemple, lors de l'étude d'un organisme qui est infecté par un autre organisme ou lorsque l'organisme à l'étude provient d'un échantillon de sol comme pour le nématode doré. Il est un candidat à la contamination car les kystes sont récoltés à partir d'échantillons de sol où une panoplie de contaminants (champignons, protozoaires, etc.) peuvent se retrouver, sur la paroi ou même à l'intérieur du kyste. Le *RNA-seq* est utilisé sur de plus en plus d'organismes non modèles. La décontamination du transcriptome *de novo* est donc une étape clé afin d'avoir des résultats valides et faciles à interpréter.

1.3.2. Méthodes de décontamination existantes

Les méthodes de décontamination existantes sont toutes basées sur un même procédé : l'alignement des transcrits sur des bases de données. Ces méthodes utilisent les alignements sur des séquences d'organismes connus pour prédire si un transcrit provient d'un contaminant ou non. Deconseq (Schmieder and Edwards 2011) utilise deux bases de données : une *white list* qui est un groupe de séquences d'organismes génétiquement proches de l'organisme à l'étude et la *black list,* qui est une base de données de séquences de contaminants potentiels ou d'organismes proches de ces contaminants. Les alignements sur ces bases de données sont faits avec BWA (Li and Durbin 2009, Li and Durbin 2010) puis en fonction des résultats, il classe les transcrits. Une autre méthode consiste à faire un alignement BLAST sur la *white list* (ou la *black list*) et de sélectionner les transcrits ayant le meilleur résultat d'alignement (Willner et al. 2009a). Ces méthodes basées sur des alignements sont efficaces, mais seulement dans certains cas. Par exemple, Schmieder and Edwards (2011) ont montré que DeconSeq pouvait décontaminer différents types de données, mais la contamination était toujours de source connue (humaine dans ce cas). L'efficacité de ces méthodes est dépendante de la qualité des bases de données, ce qui rend la décontamination complexe lorsqu'il y a plusieurs contaminants, lorsque les contaminants ne sont pas identifiés, ou lorsque les contaminants ne possèdent pas de génome de référence. Dans ces situations, certains transcrits n'auront aucun alignement et d'autres auront plusieurs alignements, ces transcrits ambiguës sont difficiles à classer. La qualité d'un alignement peut varier selon plusieurs facteurs comme la longueur des séquences et la complexité de la base de données choisie. D'autres facteurs comme le choix du type d'alignement et ses différents paramètres peuvent influencer les résultats (Xiong et al. 2014). L'ensemble de ces facteurs rend les méthodes de décontamination difficiles d'utilisation pour plusieurs utilisateurs, entre autres pour les cas de contamination provenant d'échantillon de sol. Un échantillon de sol contient en effet plusieurs organismes de tous genres. Cette variété de contaminants potentiels, dont la plupart ne possèdent pas de génome de référence, rend leur identification difficile.

1.3.3. Regroupement de sequences

Les méthodes de décontamination existantes sont efficaces seulement dans certains cas comme la contamination humaine, mais mal adaptées lorsque peu d'informations est disponible sur les contaminants. Peu de recherches ont été faites pour développer des méthodes de décontamination qui n'utilisent pas de bases de données. Willner et al. (2009b) ont montré que des analyses en composantes principales de fréquences de dinucléotides peuvent mettre en évidence les différences entre plusieurs génomes. Par contre, les résultats n'ont pas été générés dans un contexte de décontamination de transcrits ou de séquences individuelles. Cependant, l'existence de patrons (dinucléotides, trinucléotides, etc.) distinctifs pour un génome démontre qu'il est possible de catégoriser les séquences selon ces patrons. La difficulté est d'identifier les patrons et d'utiliser des séquences suffisamment longues pour qu'elles soient représentatives du génome. Les algorithmes de regroupement (*clustering*) de séquences sont conçus pour regrouper des séquences ayant des patrons similaires ensemble. Pour faire ces regroupements, les algorithmes ont besoin d'une mesure de similarité entre les séquences qui est calculée selon des alignements. Dans certains cas, les alignements de séquences sont impossibles ou très complexes (Van Walle et al. 2004). De plus, l'alignement multiple de plusieurs milliers de séquence varie selon l'aligneur utilisé et requiert beaucoup de temps de calcul. Le MCSC (Xiong et al. 2014) pour « Model-based Categorical Sequence Clustering » est un algorithme de division hiérarchique pour les séquences catégoriques. Cet algorithme a la particularité de pouvoir regrouper les séquences similaires sans l'utilisation d'alignements. Il s'est montré efficace dans différents domaines comme la détection de faillite, la reconnaissance vocale et le regroupement de séquences de protéines (Xiong et al. 2011). Plus récemment, Glouzon et al. (2014) ont fait l'analyse d'ARN d'un viroïde. Les résultats obtenus à l'aide du MCSC ont permis de mettre en évidence les mutations clés de l'évolution de la population de viroïdes. Le MCSC utilise un modèle de probabilité conditionnelle pondéré pour diviser l'ensemble des séquences en deux groupes (*clusters*) homogènes en fonction des patrons fréquents dans les séquences. Ces groupes permettent ensuite de calculer un indice de similarité entre une séquence et son groupe. Après chaque division suit une étape d'optimisation pour que la distance entre chaque séquence et son groupe soit minimale. Ces deux étapes s'appliquent récursivement jusqu'au niveau de division désiré afin d'obtenir des groupes plus petits et plus homogènes.

1.4. Objectifs

Le premier objectif de cet ouvrage est l'étude des gènes du nématode doré, *Globodera rostochiensis,* impliqués dans le processus de survie et d'éclosion. Cet objectif inclut : assembler le transcriptome du nématode doré ainsi qu'identifier les gènes réprimés ou induits lors de différents stades à l'aide du séquençage à haut-débit d'ARN. Le deuxième objectif est de développer une méthode de décontamination de transcriptome *de novo* applicable pour tous les organismes non modèles ne possédant pas de génome de référence.

Le chapitre 2 présente les méthodes et résultats du premier objectif. Nous avons utilisé plusieurs outils bio-informatiques afin d'assembler et d'identifier les gènes différentillement exprimés du nématode doré. De plus, nous avons comparé ces résultats à une étude similaire réalisée sur *Globodera pallida*, une espèce proche du nématode doré qui s'attaque aussi à la pomme de terre et dont les mécanismes de survie et d'éclosion sont similaires. Par la suite, dans le chapitre 3, à l'aide d'un algorithme de regroupement, nous avons développé une méthode de décontamination qui permet d'éliminer les contaminants d'un transcriptome *de novo*. Nous avons évalué cette méthode entre autres sur le transcriptome assemblé au chapitre 2 mais aussi sur différents jeux de données. L'objectif étant de développer une méthode de décontamination générale.

CHAPITRE 2
ANALYSIS OF POTATO CYST NEMATODES, *GLOBODERA ROSTOCHIENSIS* AND *G. PALLIDA*, TRANSCRIPTOMES EVOLUTION DURING DIAPAUSE AND HATCHING

2.1. Mise en contexte

Le nématode doré, *Globodera rostochiensis*, est un nématode phytoparasite qui peut infecter des plantes agricoles telles la pomme de terre, la tomate et l'aubergine. En raison des pertes de rendements considérables associées à cet organisme, il est justifiable de quarantaine dans plusieurs pays incluant le Canada. Cette problématique est directement reliée à son cycle de vie particulier. Ses œufs sont protégés à l'intérieur d'un kyste où ils peuvent survivre plus de 20 ans en état de dormance. Également, l'éclosion est synchronisée avec la présence d'un hôte compatible à proximité, induite par la détection d'exsudats racinaires. Une connaissance approfondie des processus de dormance et d'éclosion permettrait de développer de nouvelles approches pour lutter contre le nématode doré. Nous avons utilisé le séquençage à haut débit d'ARN (*RNA-seq*) afin d'étudier l'expression des gènes lors de ces deux événements. Nos travaux ont montré que des centaines de gènes sont induits dans les kystes dormants et les kystes exposés à l'exsudat racinaire.

Dans cet article, soumis à *Molecular Plant Pathology* le 30 mai 2016 nous présentons les résultats d'une étude transcriptomique sur les deux espèces de nématode à kyste de la pomme de terre *Globodera rostochiensis* et *G. pallida.* Nous avons étudié les gènes induits et réprimés lors du processus d'éclosion du nématode à l'aide du transcriptome *de novo* ainsi que du transcriptome de référence des deux espèces.

Les auteurs de cet article sont : Marc-Olivier Duceppe, Joël Lafond-Lapalme, Juan Emilio Palomares-Rius, Michaël Sabeh, Vivian Blok, Peter Moffett et Benjamin Mimee. Leurs contributions ont été les suivantes : Marc-Olivier Duceppe a effectué les manipulations au laboratoire en lien avec *G. rostochiensis*, préparé les librairies de séquençage, contribué à l'assemblage et aux analyses statistiques et participé à la rédaction du manuscrit. La contribution de Joël Lafond-Lapalme est égale à celle du premier auteur, il a assemblé et annoté les transcriptomes *de novo* des deux espèces, exécuté et interprété les analyses statistiques sur l'expression des gènes et a participé significativement à la rédaction du manuscrit. Juan Emilio Palomares-Rius a préparé les librairies de séquençage pour *G. pallida* et contribué à la rédaction de l'article. Michaël Sabeh a réalisé les essais de qRT-PCR servant à valider l'expression du gène NEP1. Vivian Block et Peter Moffett ont participé à l'analyse critique des résultats et à l'écriture du manuscrit. Benjamin Mimee a obtenu le financement, conceptualisé l'étude, supervisé les travaux, participé à l'analyse des résultats et à la rédaction du manuscrit.

Le matériel supplémentaire de cet article est en annexe (A-H).

ANALYSIS OF POTATO CYST NEMATODES, *GLOBODERA ROSTOCHIENSIS* AND *G. PALLIDA*, TRANSCRIPTOMES EVOLUTION DURING DIAPAUSE AND HATCHING

Marc-Olivier Duceppe[1,†,§], Joël Lafond-Lapalme[1,2,§], Juan Emilio Palomares-Rius[3], Michaël Sabeh[1], Vivian Blok[3], Peter Moffett[2] and Benjamin Mimee[1*]

[1]Agriculture and Agri-Food Canada, 430, boulevard Gouin Saint-Jean-sur-Richelieu (Québec) J3B 3E6, Canada.
[2] Département de Biologie, Université de Sherbrooke, Sherbrooke, J1K 2R1, Canada.
[3]Cell and Molecular Sciences, James Hutton Institute, Invergowrie, Dundee, DD2 5DA, United Kingdom.
[§] These authors contributed equally to this work.
†Marc-Olivier Duceppe's current address: Canadian Food Inspection Agency, Ottawa Laboratory Fallowfield (OLF), 3851 Fallowfield Road, Ottawa (Ontario) K2H 8P9, Canada
*Correspondance: Email: benjamin,mimee@agr.gc.ca, Fax: 450-346-7740;

Keywords: *Globodera pallida*, *Globodera rostochiensis*, hatching, survival, cyst nematode, *RNA-seq*, transcriptome.

2.2. Summary

Potato cyst nematodes (PCNs), *Globodera rostochiensis* and *G. pallida*, cause important economic losses in potato crop. They are hard to manage because of their ability to remain dormant in soil for many years. Although general knowledge about these plant parasitic nematodes has considerably increased over the past decades, very little is known about molecular events involved in cyst dormancy and hatching, key steps in PCN management. Here, we have studied the evolution of PCN transcriptomes from dry cysts to hatched juveniles using RNA-Seq. Several genes related to cell detoxification were up-regulated in the dry cyst, the dormant stage of PCN. Important changes in gene expression were also highlighted during hydration in both species. Most of the up-regulated genes during this stage were involved in increasing cell membrane permeability to calcium and water. Exposure of hydrated cysts to root exudates resulted in significantly different transcriptional response between *G. pallida* and *G. rostochiensis*. After 48h of exposure, no genes were consistently modulated for *G. pallida* while significant changes were observed after only 8h for *G. rostochiensis* and 278 differentially expressed genes were identified after 48h.The first gene to be up-regulated after soaking *G. rostochiensis* in root exudate was *nep-1*, coding for the transmembrane metalloprotease neprilysin. This enzyme is able to activate/inactivate peptide hormones and could be involved in a cascade of events leading to hatching. Chitinase and several known effector genes were also up-regulated during the hatching process.

2.3. Introduction

Potato cyst nematodes (PCNs), *Globodera rostochiensis* and *G. pallida*, are major plant-parasitic nematodes of potato and are found infesting fields alone or a as mixtures of both species (Pylypenko et al. 2005). They are present in the major world potato production areas and are quarantine organisms in many countries (Nicol et al. 2011, Yu et al. 2010). Yield losses are usually proportional to initial soil contamination (Greco et al. 1982, Seinhorst 1982) and are estimated at 2 t/ha of potatoes for every 20 eggs/g of soil (Brown 1969). Yield losses of potato in excess of 50% due to PCN are reported in the literature (Trudgill 1986). PCNs can also attack other crops (tomato, eggplant) and several Solanaceous weeds such as nightshades, which can serve as reservoirs (Sullivan 2007, Mimee et al. 2014). These nematodes belong to the family *Heteroderidae* and originated in South America. They were probably introduced to Europe along with potato breeding material around 1850 (Turner 1998).

Like other specialized parasites, PCN life cycle is synchronized with their hosts to optimize the chances of successful invasion (Perry 1989a). This synchrony is possible because PCN unhatched juveniles have the ability to remain dormant until a stimulus from the host is perceived, indicating favourable conditions for hatching. PCN eggs are trapped inside the dead female body, forming the cyst structure, and can survive in soil for over 20 years (Evans and Stone 1977). Hatching occurs in response to root diffusate from a suitable host plant growing nearby. However, some eggs will only hatch on restimulation, a strategy to increase population persistence throughout growing seasons and to lower competition between hatched juveniles (Perry 1989a). Variable spontaneous hatching also occurs, depending on field conditions (Turner 1996).

The PCN hatching process can be divided in three major stages: (i) changes in eggshell permeability; (ii) activation of the larva; and (iii) eclosion (Perry and Moens 2011). Trehalose inside the eggs is associated with hatching and survival. It provides an osmotic stress on the unhatched larva inducing quiescence, where locomotion and utilization of energy reserves are inhibited, thus providing protection against environmental stresses (Atkinson et al. 1987). The hatching process starts with a permeability change of the eggshell lipid layer involving Ca^{2+} (Clarke and Perry 1985), and subsequent leakage of trehalose in response to host root diffusates (Clarke et al. 1978). With the loss of osmotic pressure, juveniles become hydrated and increasingly active, leading to cutting of the eggshell and hatching. Changes in the lipid content and fatty acid composition of the larvae also occur in the egg after exposition to potato root diffusate (Holz et al. 1998). A number of external environmental factors, including host plant root diffusates, soil temperature and moisture, soil oxygen, soil microorganisms, minerals and organic substances, can serve as hatch inducers or can influence hatching (Pridannikov et al. 2007). Natural compound (solanoeclepin A), synthetic analogues (Benningshof et al. 2002) and other chemicals as picrolonic acid, sodium thiocyanate, alpha-solanine, and alpha-chaconine partially stimulate the hatching process, with greater hatching levels for *G. rostochiensis* than for *G. pallida* (Byrne et al. 2001). Using potato root diffusate (PRD), Perry and Beane (1982) showed that a single 5-min exposure to potato root diffusate (PRD) was enough to induce hatching of *G. rostochiensis* eggs while weekly 5-min exposures to PRD induced hatching of *G. pallida* eggs (Forrest and Perry 1980).

The series of physiological and behavioral events associated with hatching suggest that gene expression may be involved. However, very little is known about which genes are expressed during PCN hatching and which play a key roles. Jones et al. (1997), using differential display as analytical technique, did not find any changes in gene expression linked to exposure to PRD in *G. rostochiensis*. On the other hand, they found a few differentially expressed genes (DEGs) associated with cyst survival, but none of them showed significant homology to known sequences. Similarly, Qin et al. (2000) highlighted a few coding sequences by cDNA-AFLP related to *G. rostochiensis* cyst survival. Other studies have showed indirect observations of increased transcriptional activity during hatching of PCNs. Perry (1989b) as well as Atkinson et al. (1987) found an accumulation of secretory granules and an increase of nucleolus size of the dorsal oesophageal glands of *G. rostochiensis* within a few hours of exposure to PRD. Likewise, Blair (1999) found an increase in staining of a nucleic acid specific dye in unhatched second stage juveniles of *G. rostochiensis* after three days of exposure to tomato root diffusate (TRD). Recent transcriptome analysis of *G. pallida* has shown that 526 genes were up-regulated at the transition from encysted eggs (containing dormant J2) and hatched J2 nematodes (Cotton et al. 2014). This large-scale activation of transcription illustrates the metabolic changes and the need for upregulation of genes involved in root-penetration and other secreted proteins interacting with plant defense mechanisms. However, this study was not designed to capture early gene activation during hatching or to analyze genes involved in survival.

Here, we combine two experiments that use a high throughput/high-resolution technique, RNA-Seq, to study the evolution of the transcriptome of *G. rostochiensis* and *G. pallida* during diapause and throughout the hatching process, from dry cyst to hatched J2

2.4. Results

2.4.1. *De novo* transcriptome assemblies

To study the survival and hatching process, *G. rostochiensis* dry cysts were exposed to potato root diffusate until hatching. RNA was extracted at nine moments (dry cysts, hydrated cysts, after 6 different lengths of exposure to PRD and hatched J2). For *G. pallida*, five treatments were sampled from dry cysts to 48h of exposure to TRD. The sequencing of RNA-Seq libraries generated 511M reads for *G. rostochiensis* and 213M reads for *G. pallida*. The number of Trinity components obtained from these reads for *G. rostochiensis* was very high at 239k (assembly statistics are summarized in Table S1). This high number is attributable to the presence of sequences from contaminants. To reduce the number of contigs not belonging to *G. rostochiensis* transcriptome, we used a decontamination method called contaminant contigs removal by counts (CCRbC). The decontamination algorithm kept only 39% of the 239,134 original contigs produced by Trinity. The final *G. rostochiensis* Trinity transcriptome had 93,089 contigs, about three times as many contigs as *G. pallida*, which had 31,346. The *G. rostochiensis* reference transcriptome obtained from Augustus (Stanke et al. 2004) gene prediction on the reference genome (Sanger Institute, unpublished data) had 13,650 contigs and *G. pallida* reference transcriptome (Cotton et al. 2014) had 16,417 contigs. The comparison of *de novo* transcriptomes with these references showed that only 19.1% of the *G. rostochiensis de novo* contigs had a BLAST hit on the reference transcriptome compared to 70% for *G. pallida de novo* contigs. However, those contigs covered 96.9% and 82.7% of the *G. rostochiensis* and *G. pallida* reference transcriptome respectively.

2.4.2. DEGs analyses

Differentially expressed genes (DEGs) for *de novo* transcriptomes and computed using *G. rostochiensis* and *G. pallida* reference transcriptomes are summarized in Figure 1 and detailed in Tables S2 to S5. There were more DEGs in the *de novo* Trinity transcriptomes than when using the reference transcriptomes for both species. The ratio of up and down-regulated genes in each treatment was similar between Trinity and reference analysis for both species. However, *G. pallida* had more up-regulated than down-regulated genes in the dry cyst and the opposite at all hatching time-points (5h, 24h and 48h of exposure to TRD). On the other hand, *G. rostochiensis* had more down-regulated than up-regulated genes in the dry cyst and the opposite during hatching (24h, 48h and 7 days of exposure to PRD).

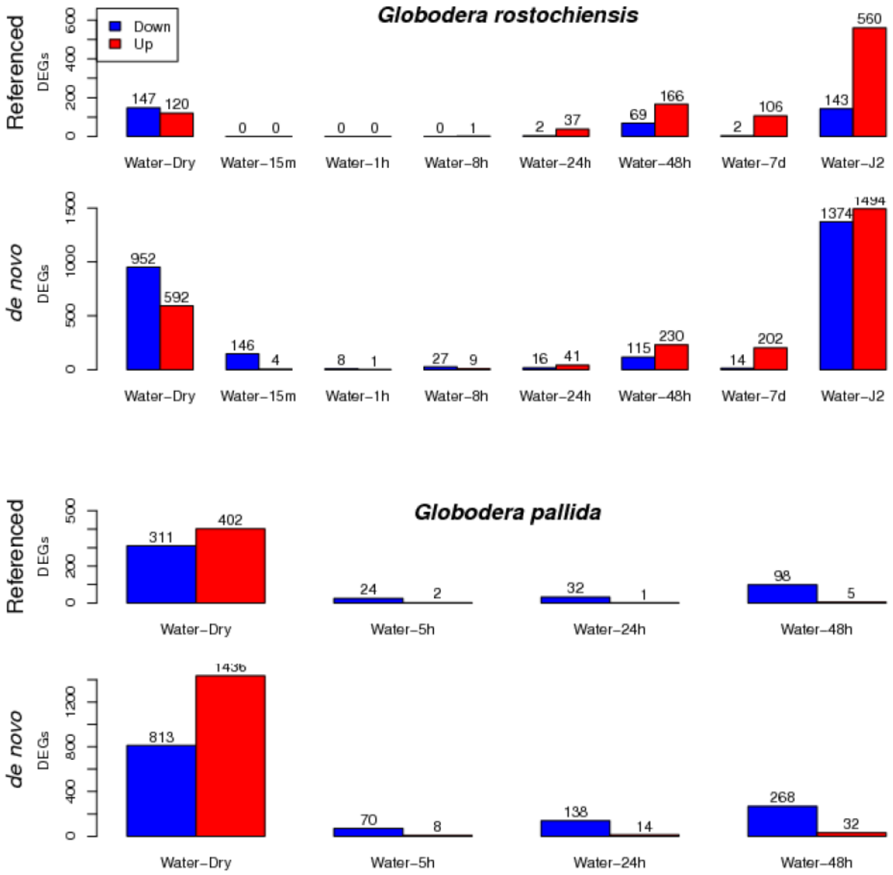


Figure 1. Pairwise counts of differentially expressed genes (DEGs) in each treatment for each transcriptome. The control treatment "water" represents hydrated cyst and time represents the soaking time in PRD/TRD after hydration.

2.4.3. Survival

Many contigs were differentially expressed in the dry cysts. Using the *de novo* transcriptomes, we found 592 contigs for *G. rostochiensis* and 1436 for *G. pallida* that were up-regulated in dry cyst when compared to hydrated cysts and considered involved in cyst survival. In the same manner, there were 952 down-regulated contigs for *G. rostochiensis* and 813 for *G. pallida* that were actually contigs that were up-regulated in response to hydration (Figure 1). The BLAST results for all these contigs can be found in supplemental tables S2-S5. We found seven relevant DEGs that had homologs in all four transcriptomes that could be involved in survival of PCN (Table 1).

Table 1. Differentially expressed genes (DEGs) up-regulated in dry cysts that were common to *G. rostochiensis* and *G. pallida* in both the *de novo* and reference-based transcriptomes.

Transcript name		
Trinity G. *rostochiensis*		
Reference *G. rostochiensis*		
Trinity *G. pallida*		
Reference *G. pallida*	BLAST results	DEG fold change
comp209610_c0,	dihydrodiol dehydrogenase	4.3
G14.T1,		4.6
comp35682_c0_seq2		1.5
GPLIN_000780500		1.9
comp238116_c0,	thiazole biosynthetic enzyme	3.0
G9808.T1,		3.7
comp29709_c0_seq5		1.9
GPLIN_000109200		2.1
comp248107_c0	protein ttr	3.2
G6983.T1		3.7
comp35561_c0_seq4		1.7
GPLIN_000178900		1.7
comp217753_c0	dorsal gland cell-specific expression protein	1.6
G1821.T1		3.0
comp36097_c0_seq6		1.7
GPLIN_000717000		2.8
comp233714_c0	adipocyte plasma membrane-associated protein	2.0
G1457.T1		3.0

comp31390_c0_seq10		1.5
GPLIN_001294200		1.7
comp252050_c0	superoxide dismutase	2.3
G3110.T1		3.2
comp30335_c0_seq3		1.7
GPLIN_000288300		1.9
comp212223_c0	an1-type zinc finger protein	2.1
G9913.T1		2.6
comp28302_c0_seq5		1.4
GPLIN_000417800		1.7

2.4.4. Hatching

The first *G. rostochiensis* transcript to be significantly up-regulated in both the Trinity and reference-based transcriptomes, after 8h exposure to PRD, encodes for a protein similar to neprilysin NEP-1 (comp140896_c0, Figure 1, Table S2 and S3). The expression of this gene was confirmed by RT-PCR (Figure S1). We also found 39 common BLAST results when comparing the up-regulated genes in hatching treatments (8h, 24h and 48h cysts soak in PRD) from the Trinity and the reference transcriptome of *G. rostochiensis* (Table 2). No DEGs were found simultaneously in both *G. pallida* transcriptomes in hatching treatments (up to 48h).

Table 2. Differentially expressed genes (DEGs) up-regulated after 8h, 24h or 48h exposure to potato root diffusate that were common to *G. rostochiensis de novo* and reference-based transcriptomes.

Contig *de novo*		*de novo* FC	Up-regulated treatments *de novo*
Contig reference	Blast results	reference FC	Up-regulated treatments reference
comp140896_c0	protein nep-1	6.4	8h
G11130.T1		3.2	8h, 24h, 48h

comp79822_c0 G3826.T1	pectate lyase 2	9.8 8.0	24h, 48h 24h,48h
comp223900_c0 G9188.T1	fatty acid elongation protein 3	8.5 8.5	24h, 48h 24h, 48h
comp233971_c0 G4478.T1	transport and golgi organization-like	7.4 4.2	24h, 48h 24h, 48h
comp197008_c0 G5298.T1	extracellular solute-binding protein family 1	6.4 4.5	24h, 48h 24h, 48h
comp239365_c0 G6254.T1	cre-mig-17 protein	6.9 4.2	24h, 48h 48h
comp250236_c1 G3528.T1	acid phosphatase-1	7.4 6.0	24h, 48h 24h, 48h
comp252640_c1 G7269.T1	arabinogalactan endo-beta-galactosidase	6.9 4.9	24h, 48h 24h, 48h
comp241201_c2 G8616.T1	histidine acid phosphatase family protein	5.6 3.7	24h, 48h 24h, 48h
comp146670_c0 G7095.T1	pectate lyase 1	6.0 4.9	24h, 48h 24h, 48h
comp258474_c0 G11848.T1	protein cht-2	4.9 5.6	24h, 48h 24h, 48h
comp253737_c1 G10850.T1	sodium bicarbonate transporter-like protein 11	5.2 3.7	24h, 48h 24h, 48h
comp205597_c0 G4741.T1	alpha-carbonic anhydrase	5.2 4.9	24h, 48h 24h, 48h
comp249939_c0 G9520.T1	expansin partial	4.9 4.9	24h, 48h 24h, 48h
comp82167_c0 G4316.T1	phosphoglycerate mutase	5.6 6.9	24h, 48h 24h, 48h
comp242049_c0	glutamine synthetase	3.4	24h, 48h

G3175.T1		1.8	48h
comp212021_c0 G6661.T1	peptidase c13 family protein	8.0 5.6	48h 48h
comp254346_c0 G926.T1	lysosomal protective	8.0 4.0	48h
comp256008_c0 G307.T1	protein mlt-7	7.4 3.2	48h 48h
comp257944_c0 G13014.T1	tartrate-resistant acid phosphatase type 5-like	6.4 6.0	48h 48h
comp231807_c0 G5991.T1	c52 protein	5.6 8.0	48h 48h
comp250073_c0 G9300.T1	hypothetical protein Aave_2802	5.6 5.6	48h 48h
comp184777_c0 G5119.T1	protein fat- isoform a	4.9 3.4	48h 48h
comp252939_c0 G11187.T1	protein del- isoform a	4.5 3.0	48h 48h
comp242752_c0 G7081.T1	beta-endoglucanase	4.2 6.0	48h 24h, 48h
comp258555_c1 G6471.T1	ghf5 endo- -beta-glucanase precursor	4.2 4.5	48h 48h
comp220907_c0 G12000.T1	hypothetical protein LOAG_17131	8.0 2.1	8h 48h
comp171900_c0 G7218.T1	hydroxyacyl-coenzyme a mitochondrial precursor	4.0 2.6	48h 48h
comp219369_c0 G10382.T1	beta- levanase invertase	4.0 3.7	48h 48h
comp234017_c0 G1410.T1	transmembrane amino acid transporter	4.0 1.7	48h 48h
comp208748_c0	cathepsin z precursor	3.7	48h

G8230.T1		2.8	24h, 48h
comp249497_c3	rbp-1 protein	3.7	48h
G11341.T1		5.2	24h, 48h
comp235317_c0	protein ugt-49	3.7	48h
G7585.T1		4.2	48h
comp250308_c0	n-acetylated-alpha-linked acidic dipeptidase 2	3.7	48h
G6374.T1		3.2	48h
comp204787_c0	Protein C36E8.1	3.4	48h
G10574.T1		1.4	48h
comp248143_c0	protein nep- isoform a	3.2	48h
G5673.T1		9.8	24h, 48h

2.4.5. DEGs clustering

We used a clustering algorithm to group the general expression pattern of each DEG. This analysis was used to narrow our search for DEGs belonging to important clusters and to explore genes that have expression patterns similar to known genes. We performed a hierarchical clustering of the expression pattern of the 4,094 unique DEGs of the Trinity transcriptome of *G. rostochiensis*. This method built 195 clusters. Because trehalose is known to be involved in survival, we identified a cluster with an expression pattern similar to *trehalose 6-phosphate synthase* (Figure 2A). This cluster contained 31 DEGs that were up-regulated in dry cysts (Table S6). We also empirically selected the cluster showing the best expression pattern for survival: a high expression level in dry cysts followed by a decrease in expression in all other treatments. This cluster (Figure 2B) contained 10 DEGs listed in Table S7.

For hatching, we selected the cluster containing NEP-1, which was found to be up-regulated in both transcriptomes in the hatching treatments of *G. rostochiensis*. This cluster (Figure 2C) contained 11 genes (Table S8). The cluster with the best expression pattern for hatching: low expression in dry and hydrated cysts followed by an increase in expression in early contact to PRD then a plateau and finally a decrease in expression in the larval stages was also studied (Figure 2D). The 13 DEGs from this cluster were up-regulated in at least one hatching treatment (Table S9).

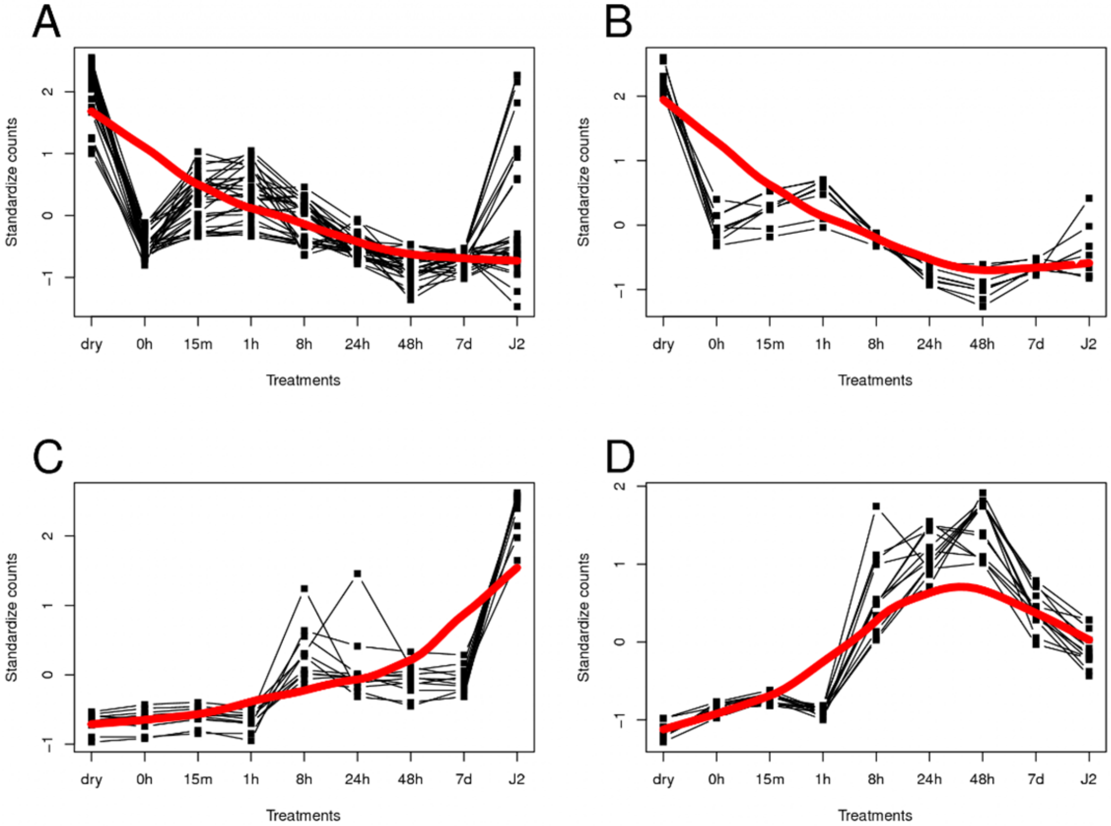


Figure 2. Clusters of expression in *G. rostochiensis* trinity transcriptome. A) Cluster containing *trehalose 6-phosphate synthase* gene. B) Cluster with a specific pattern for cyst survival. C) Cluster containing the *nep-1* gene. D) Cluster with a specific pattern for hatching.

2.5. Discussion

Throughout their evolution, cyst nematodes have developed remarkable abilities to ensure reproduction success and species persistence. One of the most impressive strategies is the ability of potato cyst nematodes, *Globodera rostochiensis* and *G. pallida*, to synchronize their hatching with the presence of a suitable host and to survive in soil for several years (Evans 1977). Very little was known about the genetic control behind long-term dormancy and hatching. In this work, we highlighted important genetic pathways that are activated during these key life stages using RNA-Seq. Sequence contamination from soil/cyst microorganisms was found to be a big challenge. For *G. rostochiensis*, more than 60% of the transcripts obtained were contaminant sequences. A simple decontamination algorithm (CCRbC) was developed and successfully removed most of these contaminating sequences without losing important information. Indeed, a horizontal coverage of 96.9% was obtained when aligning the remaining transcripts to the reference.

During dormancy, cysts nematodes stay in an anhydrobiotic state, surviving almost complete desiccation (Ellenby 1968). These organisms are protected by physical structures such as cyst and eggshell that slow the rate of water loss during desiccation, which is thought to be very important for cryptobiosis survival (Womersley et al. 1998). However, additional adaptations are needed for long-term survival. One of these mechanisms is the accumulation of trehalose inside the juvenile body. Trehalose may replace bound water by attaching to polar side groups on proteins and phospholipids, thus maintaining the balance between hydrophilic and hydrophobic forces acting on the molecules and preventing their collapse (Perry and Moens 2011). In this study, we have found that *trehalose 6-phosphate synthase* was up-regulated in dry cysts in comparison with hydrated eggs in *G. rostochiensis* (Table

S8). The enzyme with the opposite biochemical function, *trehalase*, which catalyzes the conversion of trehalose to glucose, was found to be up-regulated in dry cysts of *G. pallida*. This is not surprising however as trehalose is mostly synthesized during the early phases of cryptobiosis. Afterwards, trehalose will serve as an energy reserve and overexpression of a trehalase in *G. pallida* could reflect its use. Trehalose is thus very important, but not sufficient to ensure survival during desiccation and other adaptations at the cellular and subcellular levels are required (Perry and Wright 1998).

One important stress that dormant cysts have to cope with is the accumulation of reactive oxygen species (ROS) including superoxide ($O_2^{\bullet -}$), hydroxyl ($^{\bullet}OH$) radicals and peroxide (H_2O_2). These molecules are highly reactive and can damage nucleic acids, proteins and lipids. Desiccation will affect the control mechanisms that maintain low levels of ROS in cells. The resulting increase in ROS, if not controlled, can lead to deteriorative processes such as ageing and eventually death (Beckman and Ames 1998). Thus, some organisms have developed mechanisms to detoxify the cells and to prevent damages to macromolecules and lipid peroxidation. Antioxidants are the main molecules capable of balancing ROS levels. In the present study, we have found that several enzymatic antioxidant pathways were up-regulated in dry cysts. One of the most common superoxide radical scavengers, *superoxide dismutase* (comp252050_c0), as well as a *dehydrogenase* (comp209610_c0) were up-regulated in dry cysts versus hydrated cysts of both species, in all four transcriptomes (Table 1). Another very interesting finding is the overexpression of a thiazole biosynthetic enzyme (comp238116_c0) in dry cysts in all transcriptomes. This transcript is similar to the *thi4* gene, a key component in the biosynthesis of thiamin (vitamin B_1). Most animal does not have the machinery to synthesized B vitamins as they can easily find it through their diet. Thus, it was a surprise to find the genes for the biosynthesis of vitamin B_6 in the genome of the soybean cyst nematode, *Heterodera glycines* (Craig et al. 2008). Additional genes for vitamin B_1, B_5 and B_7 with evidences of horizontal gene transfer from bacteria were found soon after (Craig et al. 2009). These genes were also recently identified in the genome of *G. pallida* (Cotton et al. 2014) and *G. rostochiensis* (Eves-van den Akker et al., 2016). Others have previously discussed the necessity of these genes in nutrition and hypothesized on a possible limitation of B vitamins at feeding sites as general plant defense mechanism (Craig et al. 2008). Craig et al. (2009) also proposed, among other roles, that the antioxidant properties of Vitamin B_6 could be used to protect the nematode from reactive oxygen species. Vitamin B_1 is also known to have strong antioxidant properties and our results indicate that this molecule could play an important role in detoxifying ROS under anhydrobiotic conditions. This role of Vitamin B_1 in the protection of cells against oxidative damage during drought has been proposed in plants and is well documented (Tunc-Ozdemir et al. 2009).

Another transcript (comp24112_c0; Table S7) coding for a selenoprotein (thioredoxin) was overexpressed in dry cysts. The protein encoded by this gene also has antioxidant properties and was found to play an important role in aging and longevity in different organisms (Pu et al. 2015, Yoshida et al. 2005, Martin-Romero et al. 2001). Selenoproteins contains selenocysteine which is a rare amino acid using codon UGA (usually coding for termination of translation) combined with a special mRNA structure called the *selenocysteine insertion sequence* (Zinoni et al. 1990). Interestingly, *thioredoxin reductase* is the only selenoprotein reported in nematodes (Taskov et al. 2005) and one of the only enzymes with peroxidase activity known in nematodes. This gene, which showed the highest fold change in our study, was reported to be essential for life in many organisms and is currently a promising target for the development of antiparasitic drugs against nematodes in humans (Salinas et al. 2011). Several other genes, implicated in post-transcriptional regulation and coding for RNA-binding proteins and histone deacetylases were also up-regulated in dry cysts.

Both nematodes showed important changes in gene expression when they became hydrated in comparison to the dehydrated eggs. The most notable up-regulated genes during the hydration process are shown in Table 3. They include protein *gcy-9* (comp242611_c0), a guanylyl cyclase that is part of a signaling cascade activated by low intracellular calcium that leads to the synthesis of cyclic guanosine monophosphate (cGMP), which in turn allows the entry of calcium into the cell. This is consistent with the findings of Atkinson et al. (1987) who showed that the levels of cAMP and cGMP influence hatching of *G. rostochiensis*. This elevation in hydrated cysts could prepare the cells for a better reactivity to hatching factors which act in a calcium-mediated way (Atkinson and Ballantyne 1979). In the same manner, the expression of the transmembrane protein *four domain-type voltage-gated ion channel alpha-1 subunit* (comp257544_c2; table 3) will restore the permeability of cell membrane to calcium, as well as the expression of the cation channel protein *del* (comp252939_c1; table 3). The gene *mua-3* (comp258240_c1; table 3), which is predicted to have a calcium ion-binding activity, was also up-regulated during hydration. Another gene that was up-regulated during cyst hydration encodes for a *beta-endoglucanase,* which is an important effector for host root infection. Goellner et al. (2000) also found expression of this beta-endoglucanase encoding gene prior to hatching in *Globodera tabacum* eggs.

Table 3. Differentially expressed genes (DEGs) up-regulated in hydrated cysts that were common to *G. rostochiensis* and *G. pallida* in both the *de novo* and reference-based transcriptomes.

Transcript name		
Trinity *G. rostochiensis*		
Reference *G. rostochiensis*		
Trinity *G. pallida*		
Reference *G. pallida*	BLAST results	DEG fold change

comp252939_c1	protein del- isoform a	5.3
G11187.T1		3.2
comp32699_c0_seq2		2.3
GPLIN_000940400		1.9
comp258240_c1	transmembrane cell adhesion receptor mua-3	6.5
G2584.T1		3.2
comp37877_c0_seq2		2.6
GPLIN_000889800		2.0
comp89713_c0	lipase family protein	4.9
G9703.T1		3.2
comp26677_c0_seq1		2.5
GPLIN_000757300		1.7
comp34138_c0	beta- -endoglucanase	32.0
G9434.T1		17.1
comp32250_c0_seq8		3.0
GPLIN_000552400		2.0
comp233971_c0	transport and golgi organization-like protein	8
G4478.T1		4.6
comp31082_c0_seq4		3.0
GPLIN_000347500		3.0
comp199786_c0	protein unc- isoform b	11.3
G494.T1		4.9
comp36789_c0_seq1		1.9
GPLIN_000299900		1.4
comp257544_c2	four domain-type voltage-gated ion channel alpha-1 subunit	11.3
G4366.T1		5.7
comp37850_c0_seq16		2.0
GPLIN_000712300		2.3
comp242611_c0	protein gcy-9	18.4
G8365.T1		8.0
comp51248_co_seq1		6.5
GPLIN_00139600		7.0

Gene expression analysis during hatching for *G. pallida* was not possible in this study because of the limited time points available. *G. pallida* eggs take longer to hatch (Turner and Rowe 2006), and few up-regulated genes were found in the first 48 hours following exposure to root exudates. On the other hand, 278 differentially expressed genes were identified during the same period for *G. rostochiensis*. The first gene to be significantly differentially expressed in the two *G. rostochiensis* transcriptomes was *nep-1* (comp140896_c0; Table 2), coding for a neprilysin protein and up-regulated eight hours after exposure to PRD. Neprilysins (NEPs) are transmembrane zinc-metalloproteases that are well conserved throughout the animal kingdom. They were first identified in nematodes by Sajid and Isaac (Sajid and Isaac 1995). NEPs are able to hydrolyse peptide bonds at the N terminus of hydrophobic amino acids of a variety of substrates (e.g. enkephalins, tachykinins, neurotensins) thereby not only allowing the degradation of peptides, but also the post-transcriptional modification of inactive precursor peptides (Spanier et al. 2005). In *Caenorhabditis elegans*, NEP-1 is involved in locomotion and pharyngeal pumping and is highly expressed prior hatching (Spanier et al. 2005). More than 20 putative neprilysin genes were identified in *C. elegans* (Coates et al. 2000). Here, we found 11 different transcripts for NEPs. Other Zn^{2+}-metalloproteases could also play a significant role in hatching, such as a novel matrix metalloproteinase in *Heterodera glycines* (Hg-MMP) identified by Kovaleva et al. (2004).

Another interesting gene, up-regulated at 24h and 48h following exposure to PRD, is *cht-2* (comp258474_c0; Table 2) coding for a chitinase. This enzyme catabolizes chitin, a polysaccharide made of β-1,4-N-acetyl-D-glucosamine, a compound that is not present in host plants and found only in the eggshell in plant-parasitic nematodes. Endochitinases were also identified in the soybean cyst nematode, *H. glycines* (Schwekendiek et al. 1999) and in preparasitic *Meloidogyne incognita* (Dautova et al. 2001). Several other polysaccharide-degrading enzymes genes were also up-regulated during hatching (Table 2). Most of them are essential for plant colonization and prepare the nematode for its infective stage. *Beta-endoglucanases* (comp242752_c0 & comp258555_c1), *beta-levanase invertase* (comp219369_c0), and *arabinogalactan endo-beta-galactosidase* (comp252640_c1) are all involved in the degradation of plant cell walls. This last enzyme hydrolyses arabinogalactans found in dicot cell walls and may be specific to cyst nematodes as it is present in *G. pallida* and *H. schachtii* but absent from *M. incognita* and *M. hapla* (Cotton et al., 2014). Several phosphatases were also up-regulated, including *histidine acid phosphatase* (comp241201_c2), encoding a phytase that catalyses the hydrolysis of phytate (inositol hexakisphosphate), an important storage of phosphorus in many plants. A recent study has shown that specific down-regulation of the gene encoding myo-inositol phosphate synthase in plants reduces its susceptibility to cyst nematodes (Jain et al. 2015). Several genes coding for peptidases were also up-regulated during hatching in both *G. rostochiensis* transcriptomes. It has been proposed that secreted peptidases could play a role in parasitism in phytonematodes (Shinya et al. 2013) These enzymes are known to contribute to host specificity, host range and virulence in animal parasite nematodes (Williamson et al. 2006). In this study, several peptidases (comp239365_c0, comp250308_c0, comp208748_c0 & comp212021_c0) other than NEPs were found to be overexpressed during the pre-parasitic stage. Some of these were also present in *G. rostochiensis* secretions (Robertson et al. 1999) and involved in the hatching process in different nematodes (Hishida et al. 1996, Perry et al. 1992).

Finally, other known effector genes were also up-regulated in hatching treatments in both the Trinity and reference transcriptomes of *G. rostochiensis* (Table 2). These include *expansin* (comp249939_c0), *pectate lyases* (comp146670_c0 & comp79822_c0) and *rbp-1* (comp249497_c3). Pectate lyases are essential for breaking down the pectin component of plant cell walls, these enzymes were believed to be absent from animals before they were described in *G. rostochiensis* (Popeijus et al. 2000). Both pectate lyases and expansin proteins of *G. rostochiensis* induce strong phenotypes when expressed in planta suggesting virulence function (Ali et al., 2015) RBP-1 is a homologue of Ran binding proteins to microtubules (*ranbpm)* and was identified in *G. pallida* by Blanchard et al. (2005). The protein contains a SPRY domain and a signal peptide and was strongly suspected to be involved in parasitism. This protein was later identified as the avirulence factor recognised by the potato resistance protein Gpa2 (Sacco et al. 2009). This, combined with the high polymorphism of this gene (Carpentier et al. 2012) suggests that this gene family may be under strong selection pressure to evade recognition by the host. In the present study, 66 different transcripts with RBP-1 BLAST results were identified. This confirms the high genetic diversity, probable alternative splicing and high potential for adaptation in this gene (Jones et al. 2009).

In conclusion, the dormant state of potato cyst nematodes is not quiescent in term of gene expression. We have shown that a great number of genes, most being involved in cell detoxification, are specifically up-regulated during that period in both species. On the other hand, hatching seems to be triggered by only a few pathways. Cell permeability, calcium and cGMP levels were already modulated by hydration and exposure to root diffusate seems to only affect a small number of genes. Several transmembrane metalloprotease, including NEP-1, were activated early in the process and seem to be responsible for the activation of a cascade of events leading to hatching.

2.6. Experimental procedures

2.6.1. Root diffusates

For *Globodera rostochiensis*, potato plants cv. Snowden were grown in perlite, in 2L containers, until they reached about 15 cm-high. At this point, potato root diffusate (PRD) was harvested once a week, for six consecutive weeks, by the method of Fenwick (1949). Briefly, soil was drenched with tap water until saturation. An extra 50 mL of tap water was then added to the pot and the flowing liquid was collected. The collected liquid was used to repeat this procedure two more times. The final collected liquid was filtered (KenAG, D-547) to obtain PRD. PRD samples were kept at 4°C in dark plastic bottles until the last one was harvested. Then, all six weekly-sampled PRDs were pooled, freeze-dried and stored at -20°C. Final volume was recorded prior lyophilization, as well as final weight after lyophilization, for proper PRD reconstitution. PRD was reconstituted from powder with nanopure water at a final concentration of 0.5 X and passed through a 0.2 µm filter prior use.

For *G. pallida*, tomato plants (cv. MoneyMaker) were grown in 6-inch pots containing Levington Bio-Multicompost (a mixture of sand, soil and peat). When plants reached 4-weeks old, roots were removed carefully from compost, washed and placed in 250 ml flasks with distilled water. After an incubation period of 4 h, roots were removed and the remaining diffusate was filtered using Whatman no. 1 filter paper. Filtered tomato root diffusate (TRD) samples were kept at 4°C and used within 1 week.

2.6.2. Sample description

G. rostochiensis cysts were recovered by flotation (Fenwick 1940) from soil samples collected in the fall 2011 in Saint-Amable (Quebec, Canada). Cysts were stored dried for at least one year in the dark at room temperature prior to hatching experiments. A time course experiment was set up to study the evolution of the transcriptome of *G. rostochiensis* during diapause and hatching. The following physiological stages (treatments) were studied: dry cyst, cyst soaked in water for one week (hydration), hydrated cysts soaked in PRD for 15 min, 1 h, 8 h, 24 h, 48 h and 7 d and hatched J2 larvae. Each cyst sample contained 1000 cysts placed in a mesh bag (Ankom, F57). Cysts were soaked in 30 mL of filtered (0.2 µm) tap water or 0.5X PRD, in a petri dish. Water and PRD were changed every day. No hatching occurred during the hydration period. Hatched J2s were harvested daily for a two-week period and pooled for further analysis. Experiment was repeated two times.

G. pallida (population Lindley from the James Hutton Institute collection) was maintained in glasshouse conditions on the susceptible potato cultivar Desirée. Plants were inoculated with 5,000 eggs and maintained in a growth chamber adjusted to 20±1°C, 60 to 90% relative humidity, and a 14-h photoperiod of fluorescent light of 360±25 µE $m^{-2}s^{-1}$ in a mixture of 2:1 of sand:loam in root-trainers (Ronaash, Kelso, UK). After plants had died, cysts were extracted from the soil by thoroughly mixing infested soil with water in a plastic bucket and settling for 15 seconds. The supernatant was poured through a 750 µm-pore sieve nested over a 250 µm-pore sieve. Cysts were collected from the finer mesh sieve and kept at 4°C until used for egg hatching and gene expression experiments. Cysts were soaked in water for 4 days and subsequently transferred to TRD for 5, 24 and 48 h. Eggs were released from cysts using a tissue homogenizer with a clearance of 0.46-0.54 mm between the glass pestle and the homogenizer tube. Cyst walls were removed from eggs by pouring the solution through a 100 µm-pore sieve nested over a 5 µm-pore sieve. Eggs were concentrated by decantation and centrifugation. Experiment was repeated two times.

2.6.3. Total RNA extraction, library preparation and sequencing

For *G. rostochiensis*, cysts soaked in PRD were washed thoroughly with distilled water prior to RNA extraction to remove as much potential contaminants as possible. Samples were homogenized in 700 µL of RTL plus buffer with one 6 mm zirconium bead and ~150 µL of 1 mm zirconium beads using the PowerLyzer 24 homogenizer (MO BIO, Carlsbad, CA, USA) and stored at -80°C until RNA purification. Total RNA was extracted using the RNeasy Plus mini kit (Qiagen, Mississauga, Canada) according to manufacturer's instructions. Total RNA samples were store at -80°C prior RNA-Seq library preparation. RNAs were quantified with the NanoDrop 2000 (Thermo Scientific). RNA integrity was assessed with the Bioanlalyzer 2100 (Agilent Technologies) using the RNA 6000 Nano kit. All RNA samples had a RIN value higher than 7 and a 260/230 ratio value over 2.

Library preparation and sequencing were performed at McGill University and Génome Québec Innovation Centre (Montreal, Canada) using the TruSeq RNA sample prep kit v2 (Illumina) and a HiSeq 2000 sequencer (Illumina). For each replicate, all nine samples were multiplexed and sequenced in one lane for 100 bp paired-end reads.

For *G. pallida*, total RNA was extracted using RNeasy Plus Micro Kit (Qiagen, Hilden, Germany) following the manufacturer's instructions. DNA digestion was conducted on column during RNA extraction using RNase-Free DNase set (Qiagen, Hilden, Germany), as recommended. All RNA samples had a RIN value higher than 7 and a

260/230 ratio value over 2. Total RNA was quantified using a 2100 Bioanalyzer (Agilent Technologies) and the Small RNA kit (Agilent Technologies) following the manufacturer's instructions. Libraries and sequencing were produced and sequenced in Sanger Institute facilities. Illumina transcriptome libraries were produced using polyadenylated mRNA purified from total RNA using methods previously described (Choi et al., 2011) except size selection, which was done using the Caliper LabChip XT.

2.6.4. Sequence processing and differential expression analysis

Reads were trimmed from the 3' end with a minimal phred score of 30 using the Trimmomatic software. Illumina sequencing adapters were removed. Trimmed reads shorter than 32 bp were discarded and orphan reads were kept for the assembly.

The Trinity assembler version 20131110 with minimum contig length set to 300 and other parameters set to default (Haas et al. 2013, Grabherr et al. 2011) was used on normalized trimmed reads (30X coverage) to create a *de novo* transciptome for both *G. rostochiensis* and *G. pallida* using default parameters. Minimum contig length was set to 300. A custom script was applied to the Trinity transcriptomes to keep only the longest isoform of each component. Then, we apply the contaminant contig removal by counts (CCRbC) to remove contaminants sequences in the transcriptome. The CCRbC is a transcriptome decontamination method for RNA-Seq data. It uses as input the counts matrix (n x m) produces by RSEM version 1.2.8 (Li and Dewey 2011) where the n contigs are represented by n rows and the r replicates of t treatments are represented by r*t = m columns. The first step is to sum all treatments together for each replicate and for each contigs. This will result in a n by r matrix. Non-contaminant contigs are those who have at least one count for every replicates. Contaminant contigs are removed by cutting rows that contains at least one zero in the n by r matrix. Differential expression (DE; *P*<0.05 FDR-corrected) was measured using the DESeq2 Bioconductor package version 1.6.3 with the parametric wald test (Love et al. 2014) in R statistical software 3.1.2 . We also uses the RSEM software to count gene expression on *G. pallida* reference transcriptome (Cotton et al. 2014) and *G. rostochiensis* reference transcriptome obtained with augustus (Stanke et al. 2004) gene prediction from the reference genome (eves-van den akker et al., 2016). Differential expression using DESeq2 was also measured on both reference gene expression matrix produces by RSEM. Contig identification was performed using BLASTx (e-value < 1e-10) against NCBI nr database. Gene ontology (GO) and InterproScan annotations were done using Blast2GO version 3.2.7 (Conesa et al. 2005). To compare the *de novo* transcriptome to the reference transcriptome of both species, we did a BLASTn (P-value < 1e-5) of each *de novo* transcriptome on their respective reference transcriptome as BLAST databases.

2.6.5. Clustering

DEGs were clustered using the *hclust* function (*cluster* package) and the *cutreeDynamicTree* function (*dynamicTreeCut* package) in R. A matrix containing the fold changes of all DEGs compared in chronological order (e.g. dry-0h, 0h-15m, 15m-1h, etc.) was used as clustering input. Expression patterns across treatments, as well as presence of candidate genes, were also used to identify the most interesting clusters.

2.6.6. RNA extraction and cDNA synthesis for RT-PCR

Each *G. rostochiensis* sample (7 day soaked in water, 1h, 15 min, 8h, 24h, 48h and 7 day soaked in root diffusate, dry cyst and J2 larvae) contained 1000 cysts. Each sample was homogenized in 650 µl buffer RLT Plus (Qiagen) with one 6 mm zirconium grinding bead and 200 µL of 1 mm zirconium beads using the PowerLyzer® 24 Homogenizer (MO BIO) before RNA extraction. Total RNA was extracted using RNeasy Mini Kit (Qiagen) according to the manufacturer's instruction. All samples were treated with DNase (DNase I, New England Biolabs). The 2100 Bioanalyzer (Agilent Technologies) was used to analyze RNA concentration and purity. First strand complementary DNA (cDNA) was synthesized with SuperScript II reverse transcriptase (Invitrogen) according to the manufacturer's instruction, with 0.5 µg of total RNA and using oligo(dT)$_{18}$.

2.6.7. RT-PCR

Primers were designed using PrimerQuest (Integrated DNA Technologies, Inc.) (Table S10). The amplification efficiencies were calculated using the web-based Real-Time PCR Miner algorithm (ver. 4.0) (Zhao and Fernald 2005). Reactions were performed using TaqMan Universal PCR Master Mix (Applied Biosystems) on Mx3000P qPCR System (Agilent Technologie). qPCRs were performed in a 20 µl reaction volume with 1X SYBR green (SYBR Green I Nucleic Acid Gel Stain, Invitrogen), 250 nM of reverse and forward primers and 1 µl of cDNA template. The cycle details were as follow: initial denaturation 95 °C for 10 min, 40 cycles of 95 °C for 20 s and 60 °C for 60 s. A melting curve analysis followed the amplification cycles to examine the specificity of the reaction. Relative expression analysis of the *nep-1* gene was calculated using the 2-ΔΔCT method (Livak and Schmittgen 2001). Three genes (GR, PMP-3 and aaRS) reported as stables in all *Globodera* spp. life stages (Sabeh et al. in preparation) were used as reference for normalization. Hydrated cysts was the treatment used as calibrator to calculate the fold change in the other treatments. Two repetitions of the experiment were carried out.

2.7. Data availability

Globodera rostochiensis Illumina 100bp paired-end reads are available through NCBI under the bioproject accession number PRJNA274143. *Globodera pallida* Illumina sequence reads are available through the European Nucleotide Archive (http://www.ebi.ac.uk/ena) under the accession numbers ERR202482-ERR202486 (first repetition) and ERR202488-ERR202492 (second repetition). *G. pallida* reference transcriptome is available through the Sanger Institute (ftp.sanger.ac.uk/pub/project/pathogens/Globodera/pallida/)

2.8. Acknowledgments

The authors declare that they have no conflict of interest. This work was supported by the Agriculture and Agri-Food Canada Agriflex program (B. M.), the National Institute for Food and Agriculture from the United States Department of Agriculture (award #2015-69004-23634) (B. M. & V. B.), by the Education Spanish Ministry under the “Ayudas para la movilidad postdoctoral en centros extranjeros” scheme (J. E. P.-R.) and by the Scottish Government through the James Hutton Institute (V. B.).

2.9. References

Atkinson, H. and Ballantyne, A. (1979) Evidence for the involvement of calcium in the hatching of *Globodera rostochiensis*. Annals of applied Biology, 93, 191-198.

Atkinson, H. J., Taylor, J. D. and Fowler, M. (1987) Changes in the second stage juveniles of *Globodera rostochiensis* prior to hatching in response to potato root diffusate. Annals of Applied Biology, 110, 105-114.

Beckman, K. B. and Ames, B. N. (1998) The free radical theory of aging matures. Physiological reviews, 78, 547-581.

Benningshof, J. C. J., IJsselstijn, M., Wallner, S.R., Koster, A.L., Blaauw, R.H., van Ginkel, A.E., Briere, J.F., van Maarseveen, J.H., Rutjes, F.P.J.T. and Hiemstra, H. (2002) Studies towards the total synthesis of solanoeclepin A: synthesis and potato cyst nematode hatching activity of analogues containing the tetracyclic left-hand substructure. Journal of the Chemical Society-Perkin Transactions, 1, 14.

Blair, L., Perry, R. N., Oparka, K. and Jones, J. T. (1999) Activation of transcription during the hatching process of the potato cyst nematode *Globodera rostochiensis.* Nematology, 1, 103-111.

Blanchard, A., Esquibet, M., Fouville, D. and Grenier, E. (2005) Ranbpm homologue genes characterised in the cyst nematodes *Globodera pallida* and *Globodera 'mexicana'*. Physiological and molecular plant pathology, 67, 15-22.

Bolger, A. M., Lohse, M. and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 30, 2114-2120

Brown, E. B. (1969) Assessment of the damage caused to potatoes by potato cyst eelworm, *Heterodera rostochiensis* Woll. Annals of Applied Biology, 63, 493-502.

Byrne, J. T., Maher, N.J. and Jones, P.W. (2001) Comparative responses of *Globodera rostochiensis* and *G. pallida* to hatching chemicals. Journal of Nematology, 33, 195-202.

Carpentier, J., Esquibet, M., Fouville, D., Manzanares-Dauleux, M. J., Kerlan, M. C. and Grenier, E. (2012) The evolution of the Gp-Rbp-1 gene in *Globodera pallida* includes multiple selective replacements. Molecular plant pathology, 13, 546-555.

Clarke, A. and Perry, R. (1985) Egg-shell calcium and the hatching of *Globodera rostochiensis*. International journal for parasitology, 15, 511-516.

Clarke, A., Perry, R. and Hennessy, J. (1978) Osmotic stress and the hatching of *Globodera rostochiensis*. Nematologica, 24, 384-392.

Coates, D., Siviter, R. and Isaac, R. (2000) Exploring the *Caenorhabditis elegans* and Drosophila melanogaster genomes to understand neuropeptide and peptidase function. Biochemical Society Transactions, 28, 464-469.

Conesa, A., Gotz, S., Garcia-Gomez, J. M., Terol, J., Talon, M. and Robles, M. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics, 21, 3674-3676.

Cotton, J. A., Lilley, C. J., Jones, L. M., Kikuchi, T., Reid, A. J., Thorpe, P., et al. (2014) The genome and life-stage specific transcriptomes of *Globodera pallida* elucidate key aspects of plant parasitism by a cyst nematode. Genome Biol, 15, R43.

Craig, J. P., Bekal, S., Hudson, M., Domier, L., Niblack, T. and Lambert, K. N. (2008) Analysis of a horizontally transferred pathway involved in vitamin B6 biosynthesis from the soybean cyst nematode *Heterodera glycines*. Molecular biology and evolution, 25, 2085-2098.

Craig, J. P., Bekal, S., Niblack, T., Domier, L. and Lambert, K. N. (2009) Evidence for horizontally transferred genes involved in the biosynthesis of vitamin B1, B5, and B7 in *Heterodera glycines*. Journal of nematology, 41, 281.

Dautova, M., Rosso, M.-N., Abad, P., Gommers, F. J., Bakker, J. and Smant, G. (2001) Single pass cDNA sequencing-a powerful tool to analyse gene expression in preparasitic juveniles of the southern root-knot nematode Meloidogyne incognita. Nematology, 3, 129-139.

Ellenby, C. (1968) Dessication survival in the plant parasitic nematodes, *Heterodera rostochiensis* Wollenweber and Ditylenchus dipsaci (Khun) Filipjew. Proc. Roy. Soc. B., 169, 203-213.

Evans, K. a. S., A. R. (1977) A review of the distribution and biology of the potato cyst-nematodes *Globodera rostochiensis* and *G. pallida*. Pans, 23, 178-189.

Eves-van den Akker, S., Laetsch, D. R., Thorpe, P., Lilley, C. J., Danchin, E. G. J., Da Rocha, M., et al. (2016) The genome of the yellow potato cyst nematode, *Globodera rostochiensis*, reveals insights into the basis of parasitism and virulence. Genome Biology., 17, 1-23.

Fenwick, D. W. (1940) Methods for the recovery and counting of cysts of Heterodera schachtii from soil. Journal of Helminthology, 18, 155-172.

Fenwick, D. W. (1949) Investigations on the emergence of larvae from cysts of the potato-root eelworm *Heterodera rostochiensis*. I. Technique and variability. Journal of Helminthology, 23, 157-170.

Forrest, J. and Perry, R. (1980) Hatching of *Globodera pallida* eggs after brief exposures to potato root diffusate. Nematologica, 26, 130-132.

Goellner, M., Smant, G., De Boer, J., Baum, T. and Davis, E. L. (2000) Isolation of beta-1, 4-endoglucanase genes from Globodera tabacum and their expression during parasitism. Journal of Nematology, 32, 154.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature biotechnology, 29, 644-652.

Greco, N., Di Vito, M., Brandonisio, A., Giordano, I. and De Marinis, G. (1982) The effect of *Globodera pallida* and *G. rostochiensis* on potato yield. Nematologica, 28, 379-386.

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

Hishida, R., Ishihara, T., Kondo, K. and Katsura, I. (1996) hch-1, a gene required for normal hatching and normal migration of a neuroblast in *C. elegans*, encodes a protein related to TOLLOID and BMP-1. The EMBO journal, 15, 4111.

Holz, R. A., Wright, D. J. and Perry, R. N. (1998) Changes in the lipid content and fatty acid composition of 2nd-stage juveniles of *Globodera rostochiensis* after rehydration, exposure to the hatching stimulus and hatch. Parasitology Today, 116, 183-190.

Jain, R., Lilley, C. J. and Urwin, P. E. (2015) Reduction of phytate by down-regulation of Arabidopsis thaliana MIPS and IPK1 genes alters susceptibility to beet cyst nematodes. Nematology, 17, 401-407.

Jones, J. T., Kumar, A., Pylypenko, L. A., Thirugnanasambandam, A., Castelli, L., Chapman, S., et al. (2009) Identification and functional characterization of effectors in expressed sequence tags from various life cycle stages of the potato cyst nematode *Globodera pallida*. Mol Plant Pathol, 10, 815-828.

Jones, J. T., Robertson, L., Perry, R. N. and Robertson, W. M. (1997) Changes in gene expression during stimulation and hatching of the potato cyst nematode *Globodera rostochiensis*. Parasitology, 114, 309-315.

Kovaleva, E. S., Masler, E. P., Skantar, A. M. and Chitwood, D. J. (2004) Novel matrix metalloproteinase from the cyst nematodes *Heterodera glycines* and *Globodera rostochiensis*. Molecular and biochemical parasitology, 136, 109-112.

Li, B. and Dewey, C. N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics, 12, 323.

Livak, K. J. and Schmittgen, T. D. (2001) Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods, 25, 402-408.

Love, M. I., Huber, W. and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2.Genome biology, 15, 1-21.

Martin-Romero, F. J., Kryukov, G. V., Lobanov, A. V., Carlson, B. A., Lee, B. J., Gladyshev, V. N., et al. (2001) Selenium metabolism in Drosophila: selenoproteins, selenoprotein mRNA expression, fertility, and mortality. The Journal of biological chemistry, 276, 29798-29804.

Mimee, B., Andersen, R., Bélair, G., Vanasse, A. and Rott, M. (2014) Impact of quarantine procedures on weed biodiversity and abundance: Implications for the management of the golden potato cyst nematode, *Globodera rostochiensis*. Crop Protection, 55, 21-27.

Nicol, J. M., Turner, S. J., Coyne, D., Den Nijs, L., Hockland, S. and Maafi, Z. T. (2011) Current nematode threats to world agriculture. In: Genomics and molecular genetics of plant-nematode interactions. Springer, pp. 21-43.

Perry, R. and Beane, J. (1982) The effects of brief exposures to potato root diffusate on the hatching of *Globodera rostochiensis*. Revue de Nématologie, 5, 221-224.

Perry, R., Knox, D. and Beane, J. (1992) Enzymes released during hatching of *Globodera rostochiensis* and *Meloidogyne incognita*. Fundamental and applied nematology, 15, 283-288.

Perry, R. N. (1989a) Dormancy and hatching of nematode eggs. Parasitology Today, 5, 377-383.

Perry, R. N. and Moens, M. (2011) Survival of plant-parasitic nematodes outside the host. In: Molecular and physiological basis of nematode survival. (Perry, R. N., Wharton, D.A., ed.). Wallingford, UK: CAB International, pp. 1-26.

Perry, R. N. and Wright, D. J. (1998) The physiology and biochemistry of free-living and plant-parasitic nematodes. CAB INTERNATIONAL.

Perry, R. N., Zunke, U. and Wyss, U. (1989b) Observations on the response of the dorsal and subventral oesophageal glands of *Globodera rostochiensis* to hatching stimulation. Revue de Nematologie, 12, 91-96.

Perry, R. N. a. M., M. (2011) Survival of parasitic nematodes outside the host. In: Molecular and physiological basis of nematode survival. (Perry, R. N. a. W., D.A., ed.). Wallingford, UK: CAB International.

Popeijus, H., Overmars, H., Jones, J., Blok, V., Goverse, A., Helder, J., et al. (2000) Enzymology: degradation of plant cell walls by a nematode. Nature, 406, 36-37.

Pridannikov, M. V., Petelina, G.G., Palchuk, M.V., Masler, E.P. and Dzhavakhiya, V.G. (2007) Influence of components of *Globodera rostochiensis* cysts on the *in vitro* hatch of second-stage juveniles. Nematology, 9, 837-844.

Pu, M., Ni, Z., Wang, M., Wang, X., Wood, J. G., Helfand, S. L., et al. (2015) Trimethylation of Lys36 on H3 restricts gene expression change during aging and impacts life span. Genes & development, 29, 718-731.

Pylypenko, L. A., Uehara, T., Phillips, M. S., Sigareva, D. D. and Blok, V. C. (2005) Identification of *Globodera rostochiensis* and *G. pallida* in the Ukraine by PCR. European Journal of Plant Pathology, 111, 39-46.

Qin, L., Overmars, H., Helder, J., Popeijus, H., van der Voort, J. R., Groenink, W., et al. (2000) An efficient cDNA-AFLP-based strategy for the identification of putative pathogenicity factors from the potato cyst nematode *Globodera rostochiensis.* Molecular Plant-Microbe Interactions, 13, 830-836.

Robertson, L., Robertson, W. and Jones, J. (1999) Direct analysis of the secretions of the potato cyst nematode *Globodera rostochiensis.* Parasitology, 119, 167-176.

Sabeh, M., Duceppe, M.-O., St-Arnaud, M. and Mimee, B. (in preparation) Transcriptome-wide selection of a reliable set of reference genes for *Globodera rostochiensis* using RNA-seq and RT-qPCR data.

Sacco, M. A., Koropacka, K., Grenier, E., Jaubert, M. J., Blanchard, A., Goverse, A., et al. (2009) The cyst nematode SPRYSEC protein RBP-1 elicits Gpa2- and RanGAP2-dependent plant cell death. PLoS Pathog, 5, e1000564.

Sajid, M. and Isaac, R. (1995) Identification and properties of a neuropeptide-degrading endopeptidase (neprilysin) of Ascaris suum muscle. Parasitology, 111, 599-608.

Salinas, G., Bonilla, M., Otero, L., Lobanov, A. V. and Gladyshev, V. N. (2011) Selenoproteins in parasites. In: Selenium. Springer, pp. 471-479.

Schwekendiek, A., Maier, T., Womack, C., Byrne, D., De Boer, J., Davis, E., et al. (1999) Initial characterisation of endochitinase genes of the plant-parasitic nematode *Heterodera glycines*. J Nematol, 31, 568.

Seinhorst, J. W. (1982) The relationship in field experiments between population density of *Globodera rostochiensis* before planting potatoes and yield of potato tubers. Nematologica, 28, 277-284.

Shinya, R., Morisaka, H., Kikuchi, T., Takeuchi, Y., Ueda, M. and Futai, K. (2013) Secretome analysis of the pine wood nematode Bursaphelenchus xylophilus reveals the tangled roots of parasitism and its potential for molecular mimicry. PloS one, 8, e67377.

Spanier, B., Sturzenbaum, S. R., Holden-Dye, L. M. and Baumeister, R. (2005) *Caenorhabditis elegans* neprilysin NEP-1: an effector of locomotion and pharyngeal pumping. Journal of molecular biology, 352, 429-437.

Stanke, M., Steinkamp, R., Waack, S. and Morgenstern, B. (2004) AUGUSTUS: A web server for gene finding in eukaryotes. Nucleic acids research, 32, W309-W312.

Sullivan, M. J., Inserra, R. N., Franco, J., Moreno-Leheude, I. and Greco, N. (2007) Potato cyst nematodes: plant host status and their regulatory impact. Nematropica, 37, 193-201.

Taskov, K., Chapple, C., Kryukov, G. V., Castellano, S., Lobanov, A. V., Korotkov, K. V., et al. (2005) Nematode selenoproteome: the use of the selenocysteine insertion system to decode one codon in an animal genome? Nucleic acids research, 33, 2227-2238.

Trudgill, D. L. (1986) Yield losses caused by potato cyst nematodes - a review of the current position in Britain and prospects for improvements. Ann Appl Biol 108, 181-198.

Tunc-Ozdemir, M., Miller, G., Song, L., Kim, J., Sodek, A., Koussevitzky, S., et al. (2009) Thiamin confers enhanced tolerance to oxidative stress in Arabidopsis. Plant physiology, 151, 421-432.

Turner, S. J. (1996) Population decline of potato cyst nematodes (*Globodera rostochiensis, G. pallida*) in field soils in Northen Ireland. Annals of Applied Biology, 129, 315-322.

Turner, S. J. and Rowe, J. A. (2006) Cyst nematodes. In: Plant Nematology. (Perry, R. N. a. M., M. , ed.). Wallingford, UK: CAB International, pp. 91-122.

Turner, S. J. a. E., K. (1998) The origins, global distribution and biology of potato cyst nematodes (*Globodera rostochiensis* (woll.) and *Globodera pallida* Stone). In: Potato Cyst Nematodes. (Marks, R. J., and Brodie, B.B., , ed.). Wallingford: CABI Publishing.

Williamson, A. L., Lustigman, S., Oksov, Y., Deumic, V., Plieskatt, J., Mendez, S., et al. (2006) Ancylostoma caninum MTP-1, an astacin-like metalloprotease secreted by infective hookworm larvae, is involved in tissue migration. Infection and immunity, 74, 961-967.

Womersley, C., Wharton, D. and Higa, L. (1998) Survival biology. The physiology and biochemistry of free-living and plant-parasitic nematodes, 271-302.

Yoshida, T., Nakamura, H., Masutani, H. and Yodoi, J. (2005) The Involvement of Thioredoxin and Thioredoxin Binding Protein-2 on Cellular Proliferation and Aging Process. Annals of the New York Academy of Sciences, 1055, 1-12.

Yu, Q., Ye, W., Sun, F. and Miller, S. (2010) Characterization of*Globodera rostochiensis*(*Tylenchida*: *Heteroderidae*) associated with potato in Quebec, Canada. Canadian Journal of Plant Pathology, 32, 264-271.

Zhao, S. and Fernald, R. D. (2005) Comprehensive algorithm for quantitative real-time polymerase chain reaction. Journal of computational biology : a journal of computational molecular cell biology, 12, 1047-1064.

Zinoni, F., Heider, J. and Böck, A. (1990) Features of the formate dehydrogenase mRNA necessary for decoding of the UGA codon as selenocysteine. Proceedings of the National Academy of Sciences, 87, 4660-4664.

CHAPITRE 3
A NEW METHOD FOR DECONTAMINATION OF DE NOVO TRANSCRIPTOME USING A HIERARCHICAL CLUSTERING ALGORITHM

3.1. Mise en contexte

Nous avons vu au chapitre 2 que le séquençage de nouvelle génération permet d’obtenir une grande quantité d’information génétique rapidement et que cette méthode était très utile, notamment pour les études transcriptomiques. Par contre, nous avons aussi réalisé que dans certains cas, comme pour l’analyse de *G. rostochiensis* en provenance du sol, les assemblages et analyses sont compliqués par la présence de contaminants. Ce problème est encore pire pour les espèces sans génome de référence car il devient difficile de différencier les séquences de l’organisme à l’étude de celles de contaminants. Malheureusement, la plupart des méthodes de décontamination existantes sont basées sur l’alignement des séquences sur des bases de données. Ces méthodes requièrent donc une connaissance préalable des contaminants et sont peu efficaces lorsque ceux-ci sont inconnus. Le MCSC (*Model-based Categorical Sequence Clustering*) est un algorithme qui convertit un groupe de séquences en un modèle mathématique et les classe ensuite en sous-groupes. Nous démontrons dans cet article, soumis à *Bioinformatics* le 9 mars 2016, qu’il est possible de décontaminer les séquences d’un transcriptome en utilisant la méthode de décontamination du MCSC que nous avons développée sans aucune connaissance *a priori* des contaminants présents, ni besoin de génome de référence.

Les auteurs de cette étude sont : Joël Lafond-Lapalme, Marc-Olivier Duceppe, Shengrui Wang, Peter Moffett et Benjamin Mimee. Leurs contributions ont été les suivantes : Joël Lafond-Lapalme a développé, testé et comparé les méthodes de décontamination. Il a aussi contribué significativement à la rédaction du manuscrit. Marc-Olivier Duceppe a réalisé les manipulations en laboratoire, contribué aux analyses et à la rédaction du manuscrit. Shengrui Wang a contribué à l'intégration de l'algorithme à la méthode, à l'interprétation des résultats ainsi qu'à la rédaction du manuscrit. Peter Moffett a contribué à l'analyse critique des résultats et à la rédaction du manuscrit. Benjamin Mimee a obtenu le financement, supervisé les travaux, participé à l'analyse des résultats et à la rédaction du manuscrit.

Le matériel supplémentaire de cet article est en annexe (I-P).

A NEW METHOD FOR DECONTAMINATION OF DE NOVO TRANSCRIPTOME USING A HIERARCHICAL CLUSTERING ALGORITHM

Joël Lafond-Lapalme [1,2], Marc-Olivier Duceppe [1], Shengrui Wang [3], Peter Moffett [2] , Benjamin Mimee [1,*]

[1]Agriculture and Agri-Food Canada, 430, boulevard Gouin, Saint-Jean-sur-Richelieu, Canada, J3B 3E6, [2] Département de Biologie, Université de Sherbrooke, Sherbrooke, J1K 2R1, Canada, [3]Département d'Informatique, Université de Sherbrooke, Sherbrooke, J1K 2R1, Canada

3.2. Abstract

Motivation: Identification of contaminating sequences in a *de novo* assembly is challenging due to the absence of information on target species. For sample types where the target organism is impossible to isolate from its matrix, such as endoparasites, endosymbionts and soil-harvested samples, contamination is unavoidable. A few post-assembly decontamination methods are currently available. However, these are all based on alignments to databases, which can lead to poor decontamination.

Results: Here, we present a new database-free decontamination method based on a hierarchical clustering algorithm called MCSC. This method uses frequent patterns found in sequences to create clusters. These clusters are then linked to the target species or tagged as contaminants using classic alignment tools. The main advantage of this decontamination method is that it allows misaligned, ambiguous and unknown sequences to be tagged correctly and it does not depend on databases.

Availability: Scripts and documentation about the MCSC decontamination method are available at https://github.com/Lafond-LapalmeJ/MCSC_Decontamination

Supplementary information: Supplementary data are available at *Bioinformatics* online.

3.3. Introduction

Transcriptomics is essential to gain knowledge about molecular functions in complex organisms. Next generation sequencing (NGS) allows molecular biology to be scaled at the whole-genome level in a cost-effective way. RNA-seq is a NGS method that has remarkably improved gene expression quantification over the past decade (Wang et al. 2009). One of the main advantages of RNA-seq is that it does not require prior information on the genome or transcriptome of the target organism to monitor gene expression, which is very useful for non-model organisms lacking a well-annotated reference genome (Robertson et al. 2010). *De novo* transcriptome assembly builds a list of transcripts that gives a good representation of the real transcriptome, using only the RNA-seq reads as input. It is a popular, fast and cost-effective way to improve transcriptomics analyses. However, *de novo* assembly is a challenging task that requires caution to avoid errors and thereby obtain the best estimation of the “real” transcriptome (Yang and Smith 2013). Currently, mainstream sequencing technology produces very high quality reads of about 100-300 bp. Powerful algorithms are needed to reconstruct the original transcripts by overlapping short reads. Those algorithms can manage hundreds of millions of reads, but they are time consuming and require a great deal of memory.

For most organisms, no reference genome or transcriptome is available. As such, it is difficult to evaluate the quality of a *de novo* assembly (Ghangal et al. 2013). In recent years, many new *de novo* algorithms have been developed to assemble short reads based on the de Bruijn graph concept (Compeau et al. 2011) and most of these are open source. The most popular algorithms include Trinity (Haas et al. 2013, Grabherr et al. 2011), Oases (Schulz et al. 2012) and SOAPdenovo-trans (Xie et al. 2014).

Commercial software, including CLC Genomics Workbench (Qiagen) and SeqMan NGen (DNASTAR), are also able to perform such assemblies.

Post-assembly RNA-seq analyses, like variant discovery and identification of differentially expressed genes (DEGs), are dependent on assembly quality (Davidson and Oshlack 2014). Many factors can influence the quality of a *de novo* assembled transcriptome (reviewed in Baker 2012). For example, assembly artifacts like chimeras and redundant transcripts are always present (Yang and Smith 2013). These increase the total number of transcripts, which decreases the statistical power of post-assembly analyses (Baker 2012). In a perfect transcriptome, every existing transcript is represented by a unique contig. In practice, this is never the case because the clustering algorithms of *de novo* assemblers are implemented in such a way to find a compromise between sensitivity and accuracy (Compeau et al. 2011). Until an affordable technology capable of producing long reads of very high quality is readily available, post-assembly chimera and isoform removal will continue to be required.

Biological contaminants in samples can also be a major factor affecting assembly quality. In the absence of a reference genome, it is difficult to differentiate contaminating transcripts with certainty. Sometimes, it is simply impossible to avoid the presence of contaminant RNA in samples. For example, cyst nematodes are harvested from soil where a plethora of organisms are present. Bacteria, fungi, pollen, plant matter, etc. can be found on the surface or inside the cysts. Consequently, contaminant mRNAs will be sequenced together with the mRNAs of the target organism. Because of their short length, those contaminant reads cannot be discarded prior to the *de novo* assembly process, creating much more complex de Bruijn graphs and increasing the number of transcripts and assembly artifacts. These additions created by the contaminants reads inflate the transcriptome size (Yang and

Smith 2013). As long as the dominant sequencing technologies produce short length reads, a decontamination step will continue to be required for *de novo* transcriptomes assembled from contaminated samples. Existing decontamination methods are unavailable to satisfactorily improve *de novo* transcriptome assemblies containing multiple or unknown contaminants.

Several transcriptome decontamination methods have been proposed. Among them, DeconSeq (Schmieder and Edwards 2011) aligns the reads or transcripts to a white and a black list. The white list is a database containing sequences from the organism of interest, if available, or closely related species. The black list contains sequences from known contaminating organisms. Since the latter can be hard to identify, for example many soil-born organisms are yet to be discovered, one method to get an overview of the type of organisms present in the sample consists of running a BLAST analysis of the raw transcriptome assembly and listing the organisms present. The main drawback of database-dependent decontamination methods like DeconSeq is that their efficiency relies on data availability and quality. A good decontamination method will maximize contaminant contig removal while minimizing the removal of valid contigs. mRNAmarkup (Brendel and Standage; To be submitted) is another database-dependent tool for post-assembly transcriptome decontamination. Its main difference from DeconSeq is that it uses multiple database types to identify contaminant transcripts (white list, conserved domains and full length cDNAs).

Database-dependent methods are usually restricted to decontamination of well-known model organisms like humans (Schmieder and Edwards 2011). In other cases, database-dependent methods are less efficient because many transcripts align either on none or both of the white and the black list. Such ambiguous situations can lead to errors in contaminant identification and removal.

The methods outlined above are not suited to decontaminate for unknown or multiple contaminants. We explored new approaches using algorithms that do not depend on databases. Recently, a new transcriptome decontamination method based on the low probability of finding a same contaminant in every replicate was developed in our lab (Duceppe, publication in preparation). Transcripts that are not present in every replicate, all treatments together, are eliminated. This method is purely numerical and thus independent of alignments or database quality. It is only based on gene expression counts. On the other hand, it is highly dependent of the experimental design and the quality of the RNA-seq data. The algorithm is more efficient when many treatments and replicates are available and when the risk of having all the same contaminating organisms in all the samples is minimized.

Recently, Xiong et al. (2014) developed a divisive hierarchical clustering algorithm for categorical sequences named MCSC, standing for Model-based Categorical Sequence Clustering. This alignment-free algorithm has previously shown good performance for protein sequence clustering (Xiong et al. 2011), as well as for viroid RNA sequences obtained using 454 technology (Glouzon et al. 2014).

In this paper, we develop a new decontamination pipeline for assemblies based on the MCSC algorithm. Our method can effectively clean *de novo* assembled transcriptomes from two different types of samples: 1) golden nematode cysts highly contaminated with unknown soil-born microorganisms and 2) carrot weevils infected with a parasitic nematode. The method has been assessed by mixing in silico the published transcriptomes of *Loa loa* and *Fusarium oxysporum*. The feasibility of applying the decontamination pipeline to raw reads has also been evaluated.

3.4. Methods

3.4.1. Datasets

Four different contaminated datasets were used to test our new decontamination pipeline. The first was a *de novo* transcriptome from *G. rostochiensis* contaminated with a plethora of soil-borne microorganisms. RNA-seq details can be found in Supplemental information S1. The second dataset was a *de novo* transcriptome from the carrot weevil *Listronotus oregonensis* contaminated with the nematode *Bradynema listronoti,* a known parasite of the carrot weevil (Zeng et al., 2007). See Supplemental information S2 for RNA-seq details. The third dataset was generated in silico using a mixture of *Loa loa* transcriptome (PRJNA37757), a human parasitic nematode reference transcriptome, and the fungal plant pathogen *Fusarium oxysporum* reference transcriptome (PRJNA67069). Finally, to test if the MCSC algorithm could identify contaminants in short reads before any assembly steps, we used 366,956 raw reads from Ion Torrent sequencing of a DNA sample from *Globodera rostochiensis* cyst (bioproject accession number PRJNA314586).

3.4.2. *De novo* assembly

Reads obtained from RNA-seq analyses on *G. rostochiensis* and *L. oregonensis* were trimmed from the 3’ end with a minimal phred score of 30 using Trimmomatic 0.30 (Bolger et al. 2014). Illumina sequencing adapters were removed. Trimmed reads shorter than 32 bp were discarded and orphan reads were kept for the assembly. Normalization of trimmed reads (coverage 30x) was performed using Trinity normalization utility (Haas et al. 2013, Grabherr et al. 2011) to reduce memory requirement and decrease assembly runtime. The *de novo* transcriptome assembly was performed with trinity 20131110 using the 30x-normalized reads as input with default parameters, except for the minimum contig lengths, which was set to 300.

3.4.3. Gene clustering and chimera removal

Transcriptomes were submitted to Corset version 1.04 (Davidson and Oshlack 2014) with default parameters. Corset clusters transcripts to regroup isoforms, remove clusters with less than 10 supporting reads and produces a gene level expression table. From each fasta file build by Corset, chimeras were removed using the recursive chimera detection script of the mRNAmarkup version 1.0 (Brendel and Standage) pipeline with NEMBASE4 (Elsworth et al. 2011) as a reference database. Details of this database are described below. The reduced transcriptomes were used to test the decontamination methods.

3.4.4. Decontamination methods

We compared the MCSC method against three other methods. The CCRbC decontamination method is described in supplemental material S3 and was implemented in R. DeconSeq version 0.4.3 (Schmieder and Edwards 2011) using default parameters with white and black lists described in the databases section below. We also used a BLAST (Altschul et al. 1997) transcriptome decontamination method that consists of keeping only transcripts that have a BLAST hit on the nematode database NEMBASE4 described in the databases section. This method was used to evaluate a simple BLAST method with only one database versus more complex methods.

3.4.5. Databases

Many decontamination methods, such as DeconSeq and mRNAMarkup, are database-dependent. They usually require a white list or a black list. The MCSC-based decontamination method also uses two databases to identify contaminant clusters (groups of sequences). NEMBASE4 (Elsworth et al. 2011), a database containing clustered EST datasets from 63 different nematode species, was used as white list for *G. rostochiensis* transcriptome. The black list for *G. rostochiensis* transcriptome included the transcriptomes of the top hit non-nematode species identified by a BLAST of the raw Trinity transcriptome against the non-redundant database of NCBI (Figure 3).

For the *L. oregonesis* transcriptome, the white list consisted of a combination of the transcriptomes of three related species: *Acyrthosiphon pisum* (PRJNA13657)*,* *Tribolium castaneum* (PRJNA12540) and *Dendroctonus ponderosae* (PRJNA162621). The black list consisted of all sequences available in the NEMBASE4 database (Elsworth et al. 2011). All BLAST databases were built with makeblastdb command from BLAST+ 2.2.29+ and all DeconSeq databases were built as described by Schmieder and Edwards (2011).

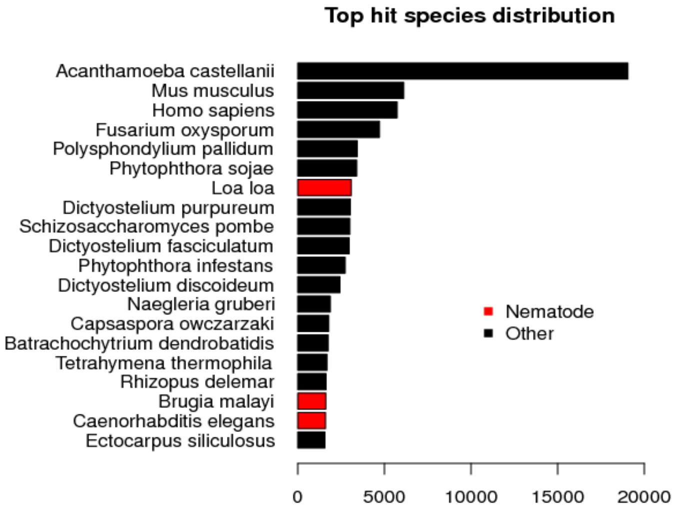


Figure 3. Species distribution of the Trinity transcriptome. Distribution of the best BLAST hit of each transcripts in the trinity transcriptome. Nematode species are in red. Only the top 20 species hits are shown.

3.4.6. MCSC decontamination method

The MCSC algorithm clusters sequences based on a weighted conditional probability distribution (WCPD) model. This statistical model allows the algorithm to build an effective representation of each cluster by using a probabilistic suffix tree and compute sequence-cluster similarities instead of typical sequence-sequence similarities done by other database-dependent algorithms. The WCPD model is a high-order Markov model with variable memory lengths. This particularity allows it to model each cluster of sequences by making optimum use of frequent patterns found within sequences without having to fix in advance the length of the actual patterns and without having to explore all the patterns of specific lengths. The MCSC algorithm works as follow. At the beginning, there is only one cluster that contains all the sequences. This cluster is divided in two preliminary clusters (Figure 4) by a fuzzy multiple correspondence analysis (F-MCA) (Xiong et al. 2014) on the vector representation of the sequences. It is equivalent to representing each sequence by the coefficients of a first-order Markov model and performing the division based on a singular value decomposition approach. Next, a statistical center is calculated for each cluster and the Chi-square similarity of each sequence is computed with respect to each cluster. A sequence is reassigned to the other cluster if it is more similar to that cluster. When these reassignments based on the first-order Markov model are over, a WCPD model is built for each cluster and some reassignments are performed for final improvement. Finally, the worst cluster is identified for further division using the same procedure. This division process is run recursively until it reaches the desired clustering level in terms of the number of clusters.

To identify contaminants, we first used the MCSC to cluster sequences. The algorithm was used as described by (Xiong et al. 2014), except that the original stopping criteria was replaced by a fixed number of clusters. Multiple clustering levels were tested, from 2 to 32 clusters (one to five iterations). After the clustering, a BLAST of all sequences against the white and black lists, described above, was used to compute a white list ratio (WR) for each cluster.

$$WR = \frac{\#hit_{white} * \ \mu_{white}}{\#hit_{all} * \ \mu_{all}}$$

WR represents the number of BLAST hits on the white list ($\#hit_{white}$) weighted by the average BLAST score on the white list (μ_{white}), divided by the total number of hits ($\#hit_{all}$) weighted by the average BLAST score (μ_{all}). Clusters with a WR ratio lower than 0.8 were labelled as contaminants.

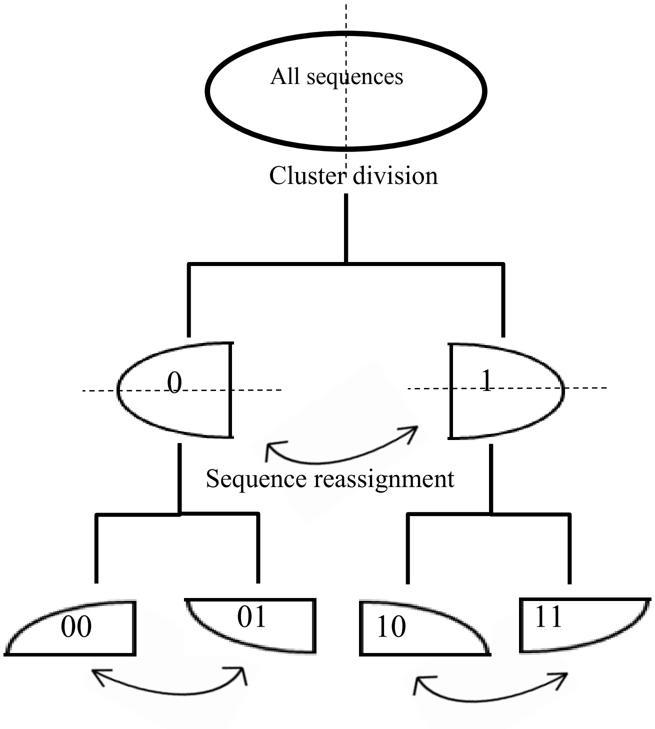


Figure 4. Cluster division by the MCSC. The first step splits, according to their relative similarities, all sequences into two clusters named "0" and "1". The second step reassigns sequence to the other cluster if the distance between the sequence and its cluster is not minimal. Reassignments stop when both clusters are stable and a final WCPD model is built for each cluster. These steps are repeated recursively on each cluster.

3.4.7. Comparison of decontamination methods

Two strategies were used to evaluate decontamination efficiency. First, the transcriptomes obtained with the four decontamination methods described above were compared with the reduced *de novo* transcriptome (non-decontaminated) and the reference transcriptome obtained from the Augustus (Stanke et al. 2004) predicted genes of the reference genome (Eves-van den Akker et al., 2016). All of these transcriptomes were BLASTed against the NCBI non-redundant protein database (Pruitt et al. 2007). From these results, we computed the species distribution of the best hit of each transcript using Blast2GO version 3.2.7 (Conesa et al. 2005b). For *G. rostochiensis*, we have done a BLAST analysis of every *de novo* transcriptomes on the reference transcriptome to compute how many genes of the reference were covered by at least one hit (*E*-value < 1e-50). Secondly, the lists of differentially expressed genes (DEG; P<0.05 FDR-corrected) before and after decontamination were compared. DEGs were identified with the DESeq2 R package version 1.6.3 with the parametric wald test (Anders and Huber 2010, Love et al. 2014) using the count tables produced by Corset (Davidson and Oshlack 2014) as input and by doing pairwise comparisons between hydrated cysts (reference) against all other conditions. To evaluate the gain in statistical power, we computed the mean p-value and adjusted p-value of all common DEGs between transcriptomes. We also compared the DEGs found in *de novo* transcriptomes to those obtained using the reference transcriptome. By using the BLASTn command, we computed the number of common DEGs (*E*-value < 1e-10) and the percentage of DEGs that had a BLAST hit on the *G. rostochiensis* genome.

3.4.8. Simulated contamination

To evaluate precisely its efficacy, the MCSC decontamination method was used to decontaminate the transcriptome of the nematode *Loa loa* from the fungus *Fusarium oxysporum*. The 15,440 genes of the nematode and the 24,828 genes of the fungus were randomly shuffled to create a contaminated transcriptome. The resulting mixture of sequences was then processed by the MCSC method and after one division (2 clusters) the number of genes of each organism in each cluster was numerated.

3.4.9. Decontamination of raw reads

We evaluated the raw read decontamination ability of the MCSC on 366,956 reads from *G. rostochiensis* contaminated by exogenous DNA by comparing the species distribution of the raw and the MCSC-decontaminated reads. We also performed a BLAST of all the reads against the reference genome of *G. rostochiensis* to compute the number of hits for the raw and decontaminated reads.

3.5. Results

Sequencing of *G. rostochiensis* and L. *oregonensis* RNA-seq libraries generated about 511M and 151M 100 bp paired-end reads, respectively. Raw transcriptome assembly with Trinity produced 679,382 transcripts for *G. rostochiensis* and 293,441 transcripts for *L*. *oregonensis*. The reduced transcriptomes, after chimera and redundant isoform removal, contained 122,553 and 70,507 transcripts, respectively. These reduced transcriptomes were used to test the decontamination methods.

3.5.1. *G. rostochiensis* transcriptome decontamination

The *G. rostochiensis* reduced transcriptome decontaminated using the CCRbC method had 92,426 transcripts that were expressed in all replicates of all treatments with a minimal count of 10. The *G. rostochiensis* reduced transcriptome decontaminated with DeconSeq included 91,234 transcripts, and the dataset decontaminated by the BLAST decontamination method produced a transcriptome of 61,663 transcripts. For the MCSC decontamination method, the WR scores of clusters obtained with MCSC run at different levels of clustering (2, 4, 8, 16, 32) are presented in Figure 5. The optimal number of clusters was determined empirically. As the clustering level increased, the number of sequences per cluster decreased, thus reducing the accuracy of cluster assignment. Therefore, the number of iterations that yields the lowest number of clusters capturing the maximum number of target sequences has to be determined. For our *G. rostochiensis* transcriptome, at n=2 ($2^{1\ iteration}$), both clusters had a WR value lower than the 0.8 threshold. n=4 and n=8 were similar, but the extra iteration generated an extra cluster with a WR > 0.8. Going from n=8 to n=16 only split the good clusters in half (six good clusters instead of three), suggesting that the optimal clustering level was n=8. From the eight clusters selected, three had a WR of 79% or more. These three clusters contained 33,806 transcripts.

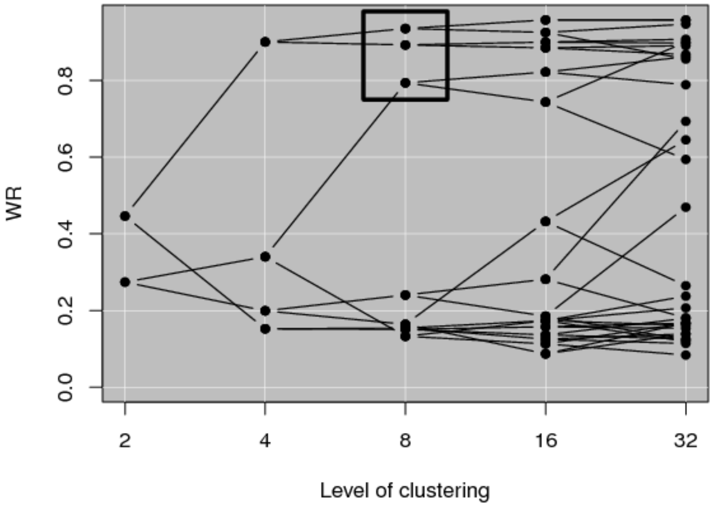


Figure 5. Clustering evaluation of the of *G. rostochiensis* transcriptome. White-Ratio (WR) of all clusters at different clustering levels. The dots inside the black rectangle are the three clusters selected and representing the decontaminated transcriptome.

All *G. rostochiensis* transcriptomes were BLASTed and their top hit species distributions were compared. The reduced transcriptome, which was used as input to the decontamination methods, had only seven nematodes species among its top hit species, and the best nematode species (*Loa loa*) was in third position (Figure 6A). The DeconSeq transcriptome had almost the same representation of nematodes in its top 20 species (Figure S1A). The CCRbC (Figure S1B) and the BLAST decontamination (Figure S1C) methods also only had seven nematodes species in the top 20, but *Loa loa* was in second position, behind the protozoa *Acanthamoeba castellanii,* which ranked first in all previous transcriptomes. The MCSC was the most efficient decontamination method, with 14 nematodes species out of the 20 top species and an all-nematode top five (Figure 6B). Finally, the reference transcriptome (predicted genes from the genome sequence) had 15 nematodes species in the top 20 and an all-nematode top five as well (Figure 4C).

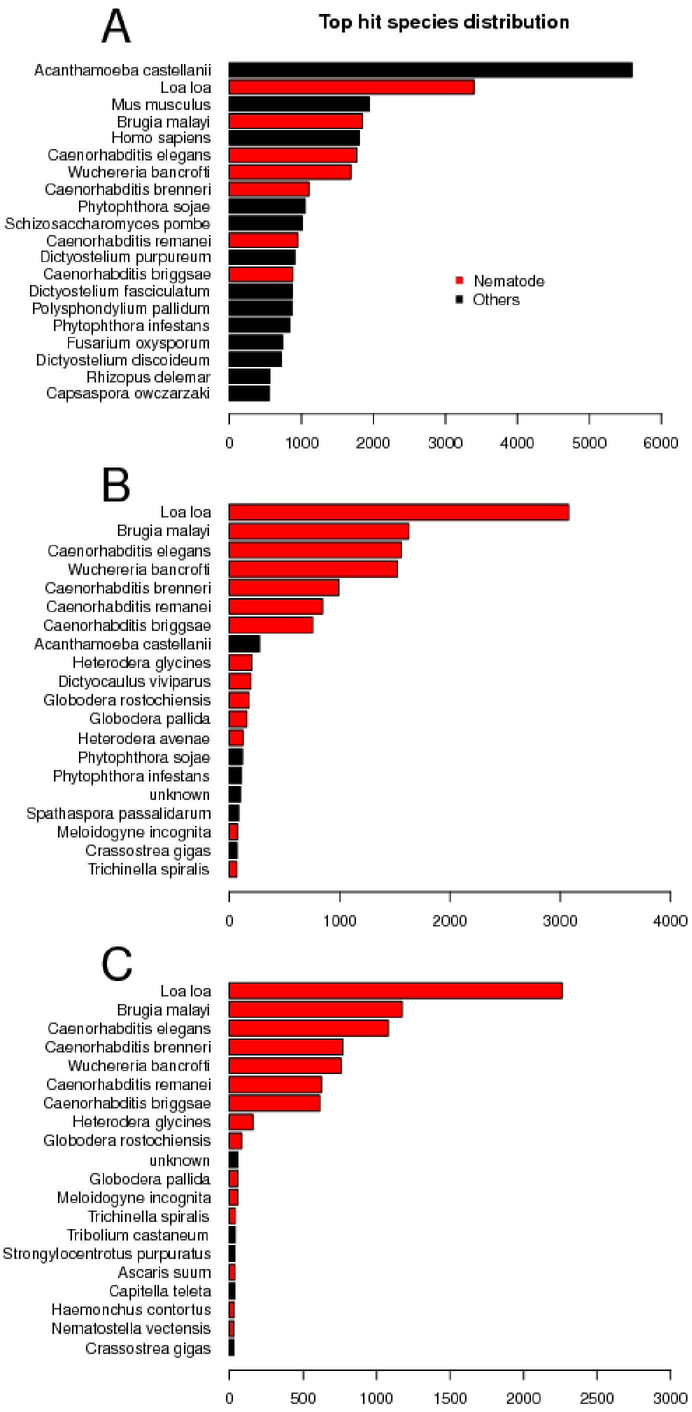


Figure 6. Species distribution of *G. rostochiensis* transcriptome. Distribution of the best BLAST hit of each transcript in the **A)** Reduced transcriptome, **B)** MCSC transcriptome and **C)** Reference transcriptome. Nematode species are in red. Only the top 20 is shown.

We evaluated the coverage percentage of each *de novo* transcriptome on the reference transcriptome. Among the 122,553 transcripts of the reduced transcriptome, 24,354 (5.03%) had a BLAST hit (*E*-value < 1e-50) on the reference transcriptome and those hits covered 11,928 genes, which represents a rate of coverage of 83.36%. In comparison, 20,730 (61.3%) of the 33,806 transcripts retained by the MCSC method had a good BLAST hit and covered 11,036 genes (77.1% coverage). All BLAST results on the reference transcriptome are summarized in Table S1.

To quantify the efficacy of the decontamination process, we compared the number and the similarity of the DEGs from the reduced transcriptome and the four decontaminated transcriptomes to the reference genome (Table 4). The MCSC transcriptome had a total of 1,733 DEGs compared to 3,190 and 2,508 DEGs for the CCRbC and the reduced transcriptomes, respectively. Although the MCSC transcriptome had the lowest number of DEGs, 95% of them had a BLAST hit on the *G. rostochiensis* genome. In comparison, only 55% and 60% of the DEGs had a hit on the genome for CCRbC and reduced transcriptome respectively, suggesting a greater number of residual contaminant transcripts.

The statistical power was also increased with the MCSC decontamination. Mean p-values were computed between the 1,313 common DEGs found in the reduced, CCRbC and MCSC transcriptomes. Reduced Trinity DEGs had a mean FDR adjusted p-value of 1.38% against 0.998% (significantly lower; t-test 99%) for the CCRbC transcriptome and 0.761% (significantly higher; t-test 99%) for the MCSC. Results were similar for the non-adjusted p-value (Table S2).

Table 4. DEG analysis of *G. rostochiensis* transcriptomes.

Transciptome	Nb. of DEGs	Common DEGs with reference transcriptome[1]	DEGs found in the reference genome[2]
Reduced	2508	778 (31%)	1517 (61%)
CCRbC	3190	887 (25%)	1760 (55%)
MCSC	1733	799 (45%)	1654 (95%)

[1] Number of DEGs with a common DEG from the reference transcriptome (1e-10).

[2] Number of DEGs with a BLAST hit (1e-10) on the reference genome.

3.5.2. *L. oregonensis* transcriptome decontamination

L. oregonensis reduced transcriptome decontamination with the MCSC algorithm required 3 iterations (n=8 or n=2^3) for optimal results (Figure 7), yielding five clusters with a WR over 90% and including 53,328 transcripts. Top species distribution of the

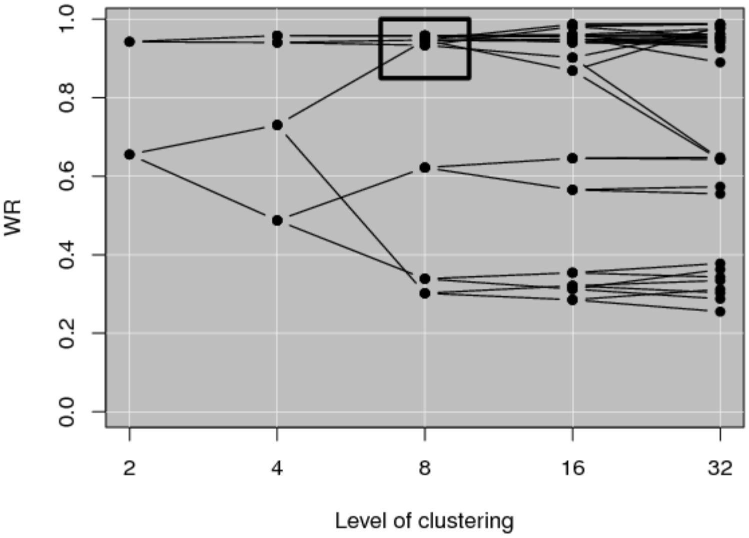


Figure 7. Clustering of the transcriptome of *L. oregonensis* by the MCSC algorithm. White-Ratio (WR) of all clusters at different clustering level. The black rectangle represents the optimal clustering level.

reduced transcriptome and the MCSC-decontaminated transcriptome is shown in Figure S2. Among the six most abundant species, only two were arthropods in the reduced transcriptome (pre-decontamination; Figure S2A). In comparison, five out of the six most frequent species were arthropods after using the MCSC method (Figure S2B).

3.5.3. Decontamination of a simulated sample

To further evaluate the efficiency of the MCSC decontamination method, the sequences of two published transcriptomes, namely *Loa loa* (15,440 transcripts) and *Fusarium oxysporum* (24,828 transcripts)*,* were randomly shuffled. The mixed transcriptome (40,268 transcripts) was submitted to the MCSC clustering algorithm. The optimal number of iterations was established at one (2^1 clusters). The cluster with a WR value of ~0.8 had 14,665 transcripts, from which 14,473 (98.66%) belonged to *Loa loa*. The second cluster had 25,603 sequences and 24,631 (96.2%) of them belonged to *Fusarium oxysporum*.

3.5.4. Raw read decontamination

For this experiment, a single iteration yielded the optimal clustering level (Figure 8). The cluster with a WR of ~0.8, containing 97,806 reads, was further analysed to evaluate the efficiency of the MCSC algorithm in decontaminating raw reads. Top species distribution of BLAST hits of the non-decontaminated reads were mostly proteobacteria species and no nematode species were present. In contrast, top hit species of MCSC decontaminated reads had seven nematode species among the top 11 (Figure S3B). Blasting the raw reads on the draft genome of *G. rostochiensis* resulted in 49,818 hits out of 366,956 (14%). When the same test was done with the MCSC decontaminated reads, we obtained 42,203 hits out of 97,806 (43%). From the 49,818 reads that align on the genome, 42,203 were kept by the MCSC method and only 7,615 were lost out of 269,150 eliminated reads.

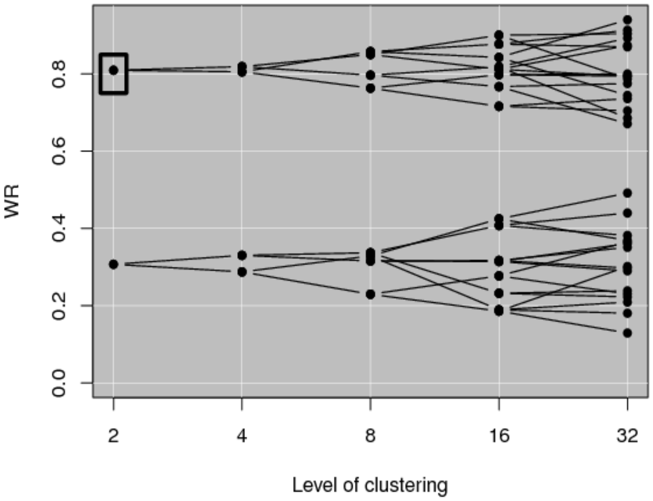


Figure 8. Clustering evaluation of contaminated *G. rostochiensis* DNA reads. White-Ratio (WR) of all clusters at different clustering levels. The black dot inside the rectangle represents the decontaminated reads.

3.6. Discussion

Sample contamination can be a serious issue for transcriptomics studies involving non-model species. Without a reference genome or transcriptome available, contaminant sequences are difficult to identify and remove. Their presence increases the total number of transcripts of *de novo* assembled transcriptomes and increases the occurrence of assembly errors such as chimeras. As a result, the overall quality of the assembly is decreased, as well as post-assembly analysis statistical power (Davidson and Oshlack 2014).

Most of the current methods for raw reads or transcriptome assembly decontamination are based on sequence alignments (Schmieder and Edwards 2011). Because they are database-dependent, they are usually more successfully applied to model organisms. In that case, transcriptome decontamination can be done by using only a white list. For non-model organisms, when no reference sequences are available, populating the white or the black list can be a challenging task. BLASTing all of the contigs from the contaminated transcriptome can provide insights about the contaminating organisms. Such information can then be used to help create a black list, but this process remains tedious, time consuming and lacks accuracy.

In order to overcome these limitations, a purely statistical decontamination algorithm named CCRbC was developed by our group (Duceppe et al.; publication in progress). Although this method is very fast and simple to implement, it can only be reliable if the contaminants are not present in all replicates. For that reason, the CCRbC method did not work well with the carrot weevil RNA-seq dataset because the contaminant *B. listronoti* was present in all replicates.

To overcome these issues, we have developed a new decontamination method based on sequence clustering which does not rely on alignments or databases. Database alignments are only used post-clustering to identify which clusters are contaminants. Consequently, there is no need to have extensive knowledge about the target and the contaminating organisms to run the MCSC decontamination method.

This characteristic makes it well suited to clustering unknown transcripts, which is of main interest for *de novo* transcriptome decontamination.

Although the MCSC, as well as its earlier version called DHCS (Xiong et al. 2011), is a general clustering algorithm, it is quite interesting to see how well it performs on biological sequences (Xiong et al. 2011, Xiong et al. 2014). Glouzon et al. (2014) successfully used the MCSC algorithm to cluster RNA viroid 454 sequences and identify new mutation patterns. Here, we have integrated the MCSC algorithm to create a method that was successfully applied to decontaminate multiple *de novo* transcriptome assemblies. We have shown that the method was efficient to decontaminate a *de novo* transcriptome assembled from *G. rostochiensis* field-harvested samples containing a myriad of other soil organisms (Figure 4). The method was also successful at decontaminating a *L. oregonensis de novo* transcriptome which was contaminated by the parasite nematode *B. listronoti* (Figure. S2). For both transcriptomes, the MCSC clustering showed a clear separation between target and contaminating clusters (Figure 5-7).

The MCSC-decontaminated transcriptomes showed a better top species distribution than their non-decontaminated counterparts, with less overall contaminants and loss of fewer target species sequences. The top species distribution of MCSC-decontaminated *G. rostochiensis* transcriptome (Figure 6B) was very similar to the reference transcriptome distribution (Figure 6C). These results suggest that the MCSC algorithm removed the contaminant sequences while leaving the sequences from the organism of interest. The MCSC algorithm eliminated 88,747 transcripts from the reduced Trinity transcriptome, from which only 3,624 (4.1%) had a BLAST hit on the reference transcriptome (Table S1). The proportion of good transcripts discarded is probably overestimated due to assembly artefacts and highly conserved transcripts between *G. rostochiensis* and contaminant transcriptomes. MCSC decontamination also increased the statistical power to identify DEGs (Table 4). The average p-value and adjusted p-value of the 1,313 common DEGs was almost cut in half between the reduced trinity and the MCSC decontaminated transcriptome. The method also removed many DEGs that were actually contaminants (Table 4). Those improvements will help in obtaining better results from *de novo* DEGs analysis.

Although post-assembly decontamination is of high interest, it would be ideal if a decontamination algorithm could be applied at the read level. Eliminating the contaminating reads prior to assembly would greatly improve the assembly process *per se*. The presence of contaminating reads results in larger *de novo* transcriptomes containing more assembly errors and artefacts (Schmieder and Edwards 2011). Today's mainstream sequencing technology produces short length reads, for which most decontamination algorithms are not well suited. Indeed, many genes are well conserved between species or even kingdoms (Kaul et al. 2000). The short read length increases the likelihood that a given read sequence will harbor a high homology between the target species and one or more contaminating organisms. It is then challenging to accurately link that one specific read to the white or the black list. Our MCSC-based method was applied to raw reads DNA sample from *G. rostochiensis* cyst. Although the WR values showed a clear separation between clusters (Figure 8), the top species distribution of target species cluster appeared to have many contaminant sequences remaining (Figure S3). This drop in performance compared to the post-assembly decontamination was expected due to the short read length and the high level of contamination. The MCSC still performed well with only 2% of the removed reads having a BLAST hit on the *G. rostochiensis* (target) reference genome. These results suggest that the MCSC algorithm should perform well with longer raw reads. The current implementation of the MCSC algorithm uses only one processor. Future work should involve multiprocessing the algorithm to exploit the availability of high performance, multicore workstations and decontamination of large raw reads datasets on large computing servers.

The MCSC algorithm can efficiently cluster different types of data (Glouzon et al. 2014, Xiong et al. 2014). We showed that it can be a powerful tool to identify and remove contaminating sequences from various *de novo* transcriptome assemblies. To be effective, the user must select the optimal level of clustering and select which clusters will form the decontaminated dataset. We suggest using two databases, a black and a white list, to compute WRs, which are used to identify contaminant clusters. This labeling method can be customized for each dataset, depending on the type of sequences or contaminants. To improve this method, it would be ideal to develop an algorithm that would identify the optimal number of clusters, label the clusters and gather only the sequences of the organism of interest automatically.

In the coming years, the availability of NGS at lower cost will stimulate exotic organism whole genome/transcriptome sequencing for which no reference sequence is available. The MCSC-based method presented here provides an efficient way to decontaminate assemblies from non-model organisms by using the information contained in the sequences themselves. Using our methods, we have achieved decontamination levels and accuracy similar to what can be obtained when a reference genome is available. This tool should be used, or at least tried, in all projects in which sample contamination is unavoidable.

3.7. Acknowledgements

The authors want to thank Pierre-Yves Véronneau for technical assistance in DNA extraction and sequencing library preparation. We also thank Jean-Pierre Sehi Glouzon and Tengke Xiong for the MCSC algorithm and source code.

3.8. Funding

This work was supported by Agriculture and Agri-Food Canada (AAFC) and the United States Department of Agriculture (USDA). This research is part of a Coordinated Agricultural Project (CAP) “Risk Assessment and eradication of *Globodera* spp. in US production of potato” supported by award #2015-69004-23634 from the National Institute for Food and Agriculture, (https://www.globodera.org/).

3.9. References

Altschul, S.F.*, et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* 1997;25(17):3389-3402.

Anders, S. and Huber, W. Differential expression analysis for sequence count data. *Genome biology* 2010;11(10).

Baker, M. De novo genome assembly: what every biologist should know. *Nature Methods* 2012;9(4):333-337.

Bolger, A.M., Lohse, M. and Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;30(15):2114-2120.

Brendel, V.P. and Standage, D.S. mRNAmarkup: quality control and annotation of de novo transcriptome assemblies.

Compeau, P.E., Pevzner, P.A. and Tesler, G. How to apply de Bruijn graphs to genome assembly. *Nature biotechnology* 2011;29(11):987-991.

Conesa, A.*, et al.* Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 2005;21(18):3674-3676.

Cotton, J.A.*, et al.* The genome and life-stage specific transcriptomes of *Globodera pallida* elucidate key aspects of plant parasitism by a cyst nematode. *Genome Biol* 2014;15:R43.

Davidson, N. and Oshlack, A. Corset: enabling differential gene expression analysis for de novo assembled transcriptomes. *Genome Biology* 2014;15(7):410.

Elsworth, B., Wasmuth, J. and Blaxter, M. NEMBASE4: The nematode transcriptome resource. *International Journal for Parasitology* 2011;41(8):881-894.

Eves-van den Akker, S., *et al.* (2016) The genome of the yellow potato cyst nematode, *Globodera rostochiensis*, reveals insights into the basis of parasitism and virulence. Genome Biology., 17, 1-23.

Glouzon, J.P.*, et al.* Deep-sequencing of the peach latent mosaic viroid reveals new aspects of population heterogeneity. *PLoS One* 2014;9(1):e87297.

Grabherr, M.G.*, et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology* 2011;29(7):644-652.

Haas, B.J.*, et al.* De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocol* 2013;8(8):1494-1512.

Kaul, S.*, et al.* Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* 2000;408(6814):796-815.

Love, M.I., Huber, W. and Anders, S. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. Genome biology 2014;15:1-21.

Pruitt, K.D., Tatusova, T. and Maglott, D.R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research* 2007;35(suppl 1):D61-D65.

Robertson, G.*, et al.* De novo assembly and analysis of RNA-seq data. *Nature Methods* 2010;7(11):909-912.

Schmieder, R. and Edwards, R. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLoS One* 2011;6(3):e17288.

Schulz, M.H.*, et al.* Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* 2012;28(8):1086-1092.

Stanke, M.*, et al.* AUGUSTUS: A web server for gene finding in eukaryotes. *Nucleic acids research* 2004;32(WEB SERVER ISS.):W309-W312.

Wang, Z., Gerstein, M. and Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009;10(1):57-63.

Xie, Y.*, et al.* SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics* 2014;30(12):1660-1666.

Xiong, T.*, et al.* A New Markov Model for Clustering Categorical Sequences. In, *Data Mining (ICDM), 2011 IEEE 11th International Conference on*. 2011. p. 854-863.

Xiong, T.*, et al.* A novel variable-order Markov model for clustering categorical sequences. *Knowledge and Data Engineering, IEEE Transactions on* 2014;26(10):2339-2353.

Yang, Y. and Smith, S. Optimizing de novo assembly of short-read RNA-seq data for phylogenomics. *BMC Genomics* 2013;14(1):328.

CHAPITRE 4
DISCUSSION GÉNÉRALE ET CONCLUSION

L'objectif initial de ce projet était l'étude transcriptomique du nématode doré lors du processus d'éclosion. Cette étude nous a permis d'identifier plusieurs gènes induits durant la dormance du kyste et lors de l'éclosion. Nous avons entre autre démontré l'importance des mécanismes de détoxification cellulaire lors de la dormance et confirmé la modulation de l'expression des gènes liés au tréhalose. La surexpression de plusieurs enzymes de détoxification connues comme la superoxide dismutase a été mise en évidence. De plus, l'étude a permis d'émettre de nouvelles hypothèses sur cette étape du cycle de vie. Par exemple, il était admis que la plupart des espèces du règne animal ne synthétisaient pas les vitamines B puisque les besoins en celles-ci sont facilement comblés dans l'alimentation. Or, les gènes pour leur synthèse ont récemment été retrouvés chez les nématodes phytoparasites et les auteurs ont avancé des besoins nutritifs particuliers des stades parasitaires. Notre étude a plutôt démontré que la synthèse de la vitamine B1 était activée durant la dormance alors que les besoins nutritifs sont nuls. La vitamine agirait plutôt comme antioxydant afin de préserver l'intégrité des macromolécules durant cette phase. En ce qui concerne l'éclosion, nos travaux ont permis de mettre en évidence la surexpression de plusieurs gènes d'endopeptidases dont le gène *nep-1* codant pour une métalloprotéase matricielle, la néprilysine. Cette enzyme est bien connue chez l'humain pour son rôle dans le développement de plusieurs cancers et de la maladie d'Alzheimer. Son activité permet d'activer ou d'inactiver plusieurs hormones peptidiques. Elle pourrait donc être à l'origine d'une cascade biochimique menant à l'éclosion des larves et est dès lors une cible très intéressante.

Par contre, la réalisation de cette étude nous a permis de constater que plusieurs transcrits présents dans le transcriptome *de novo* provenaient d'organismes contaminants. Ces séquences provenant de contaminants sont difficiles à identifier parmi les millions de courtes séquences contenues dans un fichier de séquençage de nouvelle génération, particulièrement lors d'une expérience *de novo* (sans génome de référence). C'est ce qui nous a menés à notre deuxième objectif : développer une méthode de décontamination qui s'adapte mieux au contexte d'un transcriptome *de novo* où l'on a peu d'information sur l'organisme à l'étude et ses contaminants. Notre méthode catégorise les séquences selon les patrons fréquents trouvés dans l'ensemble des séquences. Les groupes créés permettent de facilement identifier les contaminants. La méthode de décontamination du MCSC a l'avantage de pouvoir s'appliquer à n'importe quel contexte de contamination, contrairement aux méthodes existantes qui sont seulement efficaces lorsque le contaminant possède un génome de référence.

Les études à l'aide d'un transcriptome *de novo* sont un moyen d'analyser rapidement et à faible coût l'ensemble des gènes impliqués dans différents processus pour des organismes qui ne possèdent pas de génome de référence. Nous avons développé une méthode qui donne un transcriptome fiable et semblable à celui d'un transcriptome de référence malgré des échantillons hautement contaminés. Selon nos résultats, cette méthode a aussi le potentiel de pouvoir décontaminer directement les séquences brutes produites par le séquençage de nouvelle génération. Cette solution produirait un transcriptome encore plus propre, avec moins d'erreurs d'assemblage. Cette méthode est très prometteuse et devrait très rapidement être adoptée dans une multitude de situations.

L'utilisation d'un algorithme d'apprentissage pourrait être une voie possible pour obtenir une méthode « optimal » de décontamination de séquences. Par exemple, des algorithmes de *deep learning* peuvent, à l’aide de millions d’images préalablement analysées, identifier des animaux sur une image quelconque. Un algorithme du même genre pourrait, en fonction des séquences connues et identifiées dans la littérature, identifier les séquences appartenant à un organisme quelconque. Par ailleurs, nous avons utilisé l’algorithme du MCSC à des fins de décontamination, mais cet algorithme est conçu pour former des groupes de séquences homogènes, peu importe le contexte. Il serait intéressant de le tester sur des données non contaminées afin d’évaluer son potentiel à reconnaître des groupes de séquences distinctes à l’intérieur d’un même organisme. Cela pourrait par exemple mettre en évidence des séquences provenant de transferts de gènes horizontaux ou des mutations importantes dans l’évolution d’une espèce.

ANNEXES

5.1. Appendix A

Figure S1. Expression of neprilysin gene *nep-1* by RT-PCR

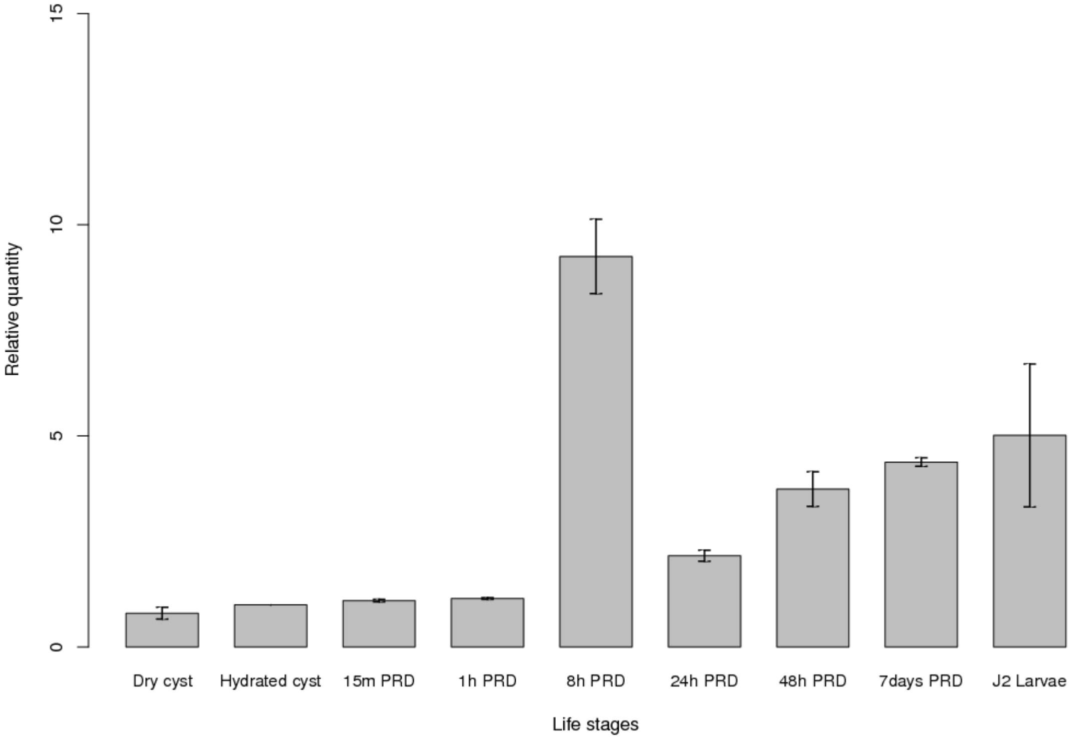


Figure S1: Expression of neprilysin gene *nep-1* by RT-PCR in dry cysts, hydrated eggs exposed to PRD for 15 min, 1h, 8h, 24h, 48h, 7 days and hatched J2. Hydrated cyst expression level was use as calibrator.

5.2. Appendix B

Table S1. *Globodera rostochiensis* and *G. pallida* Trinity transcriptome assembly statistics.

Metric	*Globodera rostochiensis*	*Globodera pallida*
Paired-end reads (100 bp)	511,208,631	213,156,265
Normalized pairs (30 X)	41,045,675	7,903,581
Trinity transcripts	1,075,007	300,796
Trinity components	239,134	31,346
Final transcriptome	**93,089**	**31,346**
Longest contig (bp)	18,280	15,960
N50 (bp)	1,293	2,527
GC (%)	47	47

5.3. Appendix C

Table S2-S5. DEGs during hatching in *Globodera* transcriptomes

S2 *G. rostochiensis* Trinity transcriptome

S3 *G. rostochiensis* reference transcriptome

S4 *G. pallida* Trinity transcriptome

S5 *G. pallida* reference transcriptome

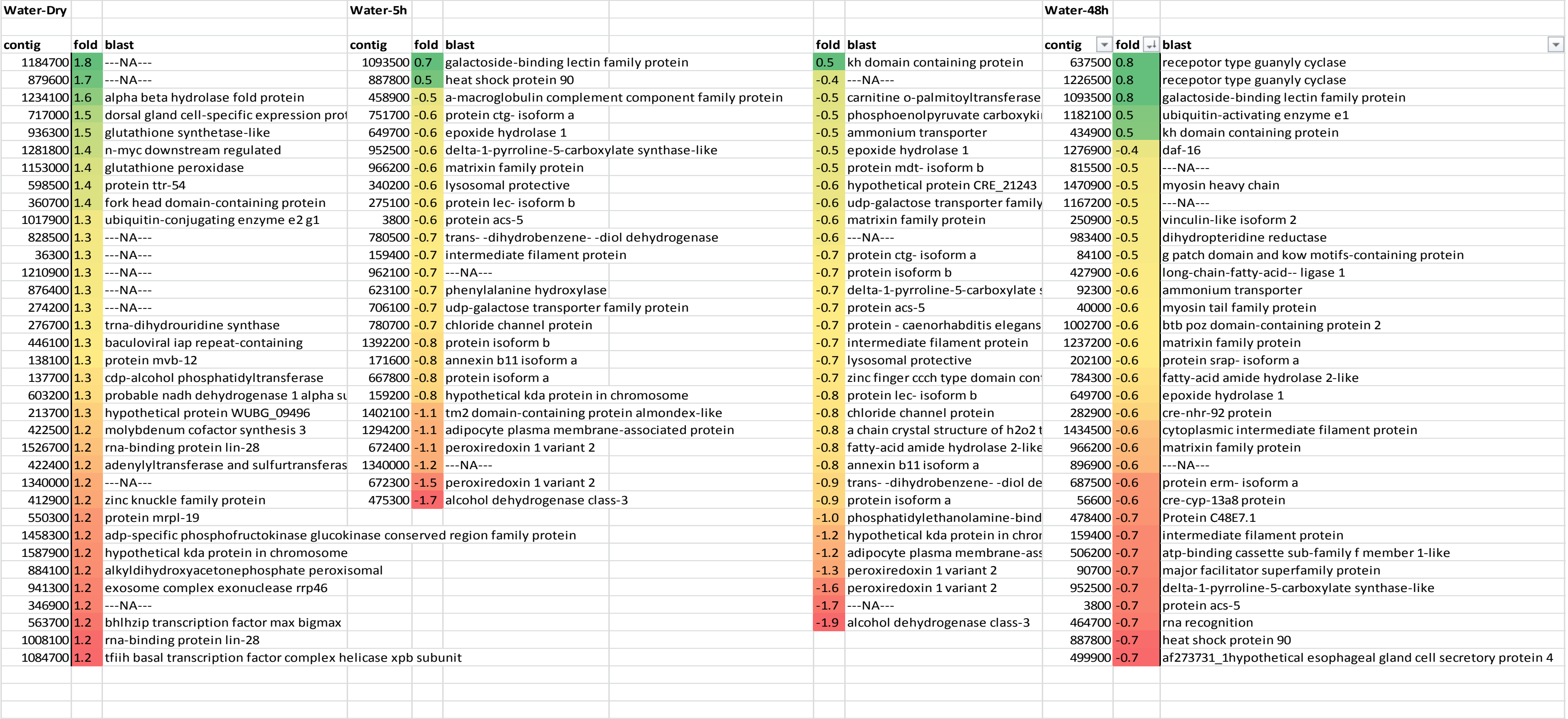
Seulement la première page de chaque tableau est affichée. Pour les tableaux complets, voir le matériel supplémentaire de l'article en ligne.

Dry-Water			Water-15m			Water-1h			Water-8h			Water-24h			Water-48h			Water-7d			Water-J2		
contig	fold	blast	contig	fold	blast	contig	fold	blast	contig	fold	blast	contig	fold	blast	contig	fold	blast	contig	fold	blast	contig	fold	blast
255433_c0	7.5	structural polypro	257847_c0	5.9	---NA---	257215_c0	5.3	---NA---	258542_c0	5.0	fibro-slime fan	220545_c0	6.0	---NA---	253529_c3	5.1	enolase	254788_c1	6.1	cellulose synt	152329_c0	7.7	---NA---
225922_c1	5.7	---NA---	247163_c0	5.5	---NA---	255324_c1	-3.9	actin	151504_c0	4.8	alcohol dehydr	143625_c0	5.6	---NA---	250115_c0	4.9	cog1938 domai	255723_c0	6.0	cysteine prote	253435_c0	7.3	rna-binding domai
240085_c0	5.5	glucose-repressib	250669_c1	5.3	protein dd3-3-li	249858_c0	-4.2	elongation facto	213450_c1	4.6	nucleoside dip	231234_c0	5.3	---NA---	182827_c0	4.7	protein	256324_c2	6.0	protein	263948_c0	7.3	---NA---
258085_c0	5.4	aspartic protease	256870_c0	4.5	---NA---	257332_c0	-5.0	---NA---	178353_c0	3.1	galactosylgalac	247163_c0	5.2	---NA---	231486_c0	4.5	protein nep- is	256749_c2	6.0	elongation fac	243049_c0	7.3	---NA---
220915_c0	5.4	conserved hypoth	252270_c0	-1.2	gcn20-type atp-	255728_c0	-5.0	oxidoreductase	220907_c0	3.0	hypothetical p	254050_c0	5.1	hydroxylamine	178293_c0	4.3	---NA---	233666_c1	5.9	60s ribosomal	258155_c1	7.1	von willebrand fac
254784_c0	5.4	lea domain-contai	240072_c1	-2.0	---NA---	252798_c0	-5.1	40s ribosomal pr	148687_c0	3.0	---NA---	231486_c0	3.5	protein nep- isc	178240_c0	4.0	---NA---	243603_c1	5.9	fructose-bisph	245392_c0	7.0	---NA---
233577_c0	5.4	cell surface	240072_c0	-2.0	---NA---	248622_c1	-5.1	cysteine protein	140896_c0	2.7	protein nep-1	178293_c0	3.5	---NA---	207374_c1	4.0	---NA---	226440_c2	5.9	60s ribosomal	251449_c0	7.0	dual isoform b
241242_c0	5.3	cell surface	253102_c1	-2.2	glutathione s-tr	257705_c1	-5.5	major vault prot	239479_c0	2.0	protein isoforr	178240_c0	3.2	---NA---	209592_c0	3.9	protein	253524_c0	5.8	---NA---	235438_c0	6.9	---NA---
245836_c1	5.3	ferritin	249288_c1	-2.2	60s ribosomal p	258737_c0	-6.2	fibro-slime fami	231155_c0	2.0	protein isoforr	232270_c0	3.1	protein	152861_c0	3.8	---NA---	251864_c1	5.8	s-adenosylme	245392_c1	6.8	---NA---
175045_c0	5.2	conserved hypoth	254422_c2	-2.2	14-3-3 protein				258830_c0	-1.6	---NA---	79822_c0	3.1	pectate lyase 2	253814_c1	3.8	---NA---	250146_c0	5.6	40s ribosomal	253435_c1	6.7	nucleolin protein
217009_c0	5.2	upf0591 membran	245057_c0	-2.2	ras gtpase				218370_c0	-1.7	profilin ii	223900_c0	2.9	fatty acid elong	85047_c0	3.8	protein isoforr	241666_c0	5.5	---NA---	249147_c1	6.7	aaa family atpase
240195_c0	5.1	mismatched base	248808_c0	-2.3	elongation factor partial				252998_c0	-2.0	major vault prc	233971_c0	2.9	transport and gc	258752_c1	3.7	---NA---	243464_c2	5.4	60s ribosomal	256874_c0	6.7	---NA---
242367_c0	5.1	glucose-repressibl	258731_c0	-2.4	myosin ii heavy chain				255351_c0	-2.1	ethyl tert-buty	197008_c0	2.8	extracellular so	246807_c1	3.6	---NA---	256702_c0	5.3	---NA---	249857_c1	6.7	---NA---
245410_c0	5.1	defective mitochc	254823_c0	-2.4	cathepsin l2				258785_c0	-2.1	polyketide syn	231227_c0	2.8	fad-linked oxid	259655_c0	3.6	glycosyltransfe	252268_c1	4.9	60s ribosomal	212079_c0	6.6	gland protein g20e
71217_c0	5.0	---NA---	255132_c2	-2.5	alpha-tubulin				258212_c0	-2.1	n-acyl-d-glucos	152061_c0	2.8	pectate lyase 2	178494_c0	3.6	glycosyltransfe	254378_c1	4.7	dd3_dicdi ame	258576_c1	6.6	tldc domain-conta
222797_c0	5.0	chaperone heat sh	253520_c0	-2.5	40s ribosomal protein s1				240072_c1	-2.1	---NA---	186241_c0	2.7	protein isoform	79822_c0	3.5	pectate lyase 2	241087_c0	4.7	60s ribosomal	256292_c4	6.5	heat shock proteir
256253_c2	5.0	cre-asp-6 protein	241964_c0	-2.5	onent of the counting factor lex				254823_c0	-2.2	cathepsin l2	179924_c0	2.7	extracellular so	251015_c1	3.5	---NA---	142227_c0	4.4	lim domain co	251557_c3	6.5	elongation factor
245001_c0	5.0	---NA---	258181_c2	-2.6	ethyl tert-butyl ether degradation				258590_c1	-2.2	n-acyl-d-glucos	239365_c0	2.7	cre-mig-17 prot	232690_c0	3.5	protein isoforr	257647_c2	4.3	elongation fac	255959_c1	6.4	---NA---
234746_c0	5.0	msc1 protein	257786_c1	-2.6	clathrin heavy chain 1-like				240072_c0	-2.2	---NA---	254948_c0	2.5	---NA---	228898_c0	3.4	---NA---	254540_c1	4.3	cathepsin l2	248387_c0	6.4	---NA---
222909_c0	4.9	---NA---	248578_c0	-2.6	60s ribosomal protein l4				206723_c0	-2.5	hypothetical p	250236_c1	2.5	acid phosphatas	204519_c0	3.4	---NA---	258405_c0	4.3	protein dd3-3-	27762_c0	6.4	60s ribosomal prot
269472_c0	4.9	tubulin gamma ch	254314_c1	-2.6	activation domain containing protein				179442_c0	-2.9	acto_acaca ame	252640_c1	2.5	arabinogalactar	254320_c1	3.4	---NA---	257326_c0	4.2	protein dd3-3-	254100_c0	6.3	40s ribosomal prot
253509_c1	4.9	40s ribosomal prot	213523_c0	-2.6	14-3-3-like partial				257705_c1	-3.3	major vault prc	252640_c0	2.5	arabinogalactar	231227_c0	3.4	fad-linked oxi	252121_c1	4.2	adenosylhom	255620_c0	6.3	actin binding prote
255040_c1	4.8	---NA---	255685_c1	-2.6	60s ribosomal protein				257272_c0	-3.3	---NA---	212359_c0	2.5	acid sphingomy	232987_c0	3.3	protein isoforr	257434_c0	4.1	extracellular r	251486_c0	6.3	hypothetical prote
229594_c1	4.8	hypothetical prote	252058_c0	-2.7	altered inheritance rate of mitochondria protein				250134_c0	-3.3	elongation fact	241201_c2	2.4	histidine acid pl	213467_c0	3.3	---NA---	253841_c1	4.1	---NA---	246769_c0	6.3	40s ribosomal prot
29022_c0	4.7	conidiation-specif	254810_c1	-2.7	protein				32107_c0	-3.3	acto_acaca ame	220435_c0	2.4	protein isoform	253814_c2	3.3	---NA---	253769_c0	4.0	40s ribosomal	184405_c0	6.2	---NA---
238954_c0	4.7	chaperone heat sh	250781_c2	-2.7	calreticulin precursor				236065_c0	-3.5	atp-dependent	236889_c0	2.4	protein isoform	251834_c0	3.3	---NA---	251708_c1	4.0	gelsolin-like p	252218_c0	6.2	---NA---
209115_c0	4.7	---NA---	243250_c0	-2.8	---NA---				255243_c0	-3.6	plasma membr	254948_c1	2.3	---NA---	152061_c0	3.3	pectate lyase 2	249857_c1	3.8	---NA---	251164_c1	6.2	sequestosome-1 i
205196_c0	4.7	sulfotransferase	218370_c0	-2.8	profilin ii				257979_c1	-3.7	fibro-slime fan	146670_c0	2.3	pectate lyase 1	30126_c0	3.3	---NA---	254194_c0	3.8	hypothetical p	246503_c0	6.2	---NA---
231561_c1	4.6	msc1 protein	239295_c1	-2.8	60s ribosomal protein l20				249931_c0	-4.2	glycogen synth	258474_c0	2.3	protein cht-2	231619_c0	3.3	sodium- and ch	253814_c1	3.7	---NA---	50554_c0	6.2	phospholipase a2
247705_c1	4.6	---NA---	143492_c0	-2.8	actin binding protein				252058_c0	-4.3	altered inherit	253737_c0	2.2	sodium bicarbo	211868_c0	3.3	---NA---	257848_c0	3.7	alpha- sarcom	258183_c1	6.1	heat shock proteir
203920_c1	4.6	polyadenylate bin	252998_c0	-2.8	major vault protein				252339_c0	-4.3	transcriptional	205597_c0	2.2	alpha-carbonic a	256802_c0	3.2	---NA---	254794_c0	3.6	partial	141609_c0	6.1	---NA---
213135_c0	4.5	#NAME?	255219_c1	-2.8	ribosomal protein l7				243416_c0	-4.4	glucose-repres	258665_c0	2.2	---NA---	252171_c0	3.2	---NA---	178293_c0	3.6	---NA---	81081_c0	6.0	---NA---
86818_c0	4.5	protein	256496_c1	-2.9	spherulation-specific family 4				249967_c0	-4.4	carnitine acety	251015_c2	2.1	---NA---	222746_c0	3.2	---NA---	231486_c0	3.6	protein nep- i	248122_c0	6.0	protofilament ribb
238301_c0	4.5	cytochrome oxida	258830_c0	-2.9	---NA---				252568_c0	-4.5	hypothetical p	258466_c3	2.0	pa2l_caeel ame	250236_c1	3.2	acid phosphata	258796_c0	3.6	type a von wil	147397_c0	6.0	pao retrotranspos
245486_c0	4.5	hypothetical prote	228343_c0	-2.9	elongation factor 1-beta				254529_c0	-5.1	c2h2 transcript	258752_c0	2.0	---NA---	255621_c1	3.2	---NA---	257204_c2	3.5	elongation fac	254127_c0	6.0	protein
247367_c1	4.5	cell surface	256078_c1	-2.9	tryptophan halogenase family protein				253559_c0	-5.3	26s proteasom	249939_c0	2.0	expansin partia	181322_c0	3.2	protein	246807_c1	3.4	---NA---	249857_c0	6.0	---NA---
251009_c0	4.5	universal stress pr	256890_c0	-2.9	actin binding protein							82167_c0	1.9	phosphoglycera	253218_c1	3.2	---NA---	254320_c1	3.4	---NA---	254559_c0	6.0	ribosomal p0
30177_c0	4.4	upf0591 membran	140999_c0	-2.9	60s ribosomal protein l36-2-like							258765_c1	1.9	---NA---	214364_c0	3.2	sodium- and ch	254320_c0	3.4	---NA---	239213_c0	5.9	---NA---
216169_c0	4.4	---NA---	252998_c1	-2.9	major vault protein							253643_c1	1.9	expansin partia	254320_c0	3.2	---NA---	239213_c0	3.4	---NA---	257921_c1	5.9	rna-binding regior
223300_c0	4.3	---NA---	228919_c0	-3.0	duf614 family protein							242049_c0	1.7	glutamine synth	252092_c1	3.2	---NA---	251015_c1	3.4	---NA---	251145_c1	5.9	14-3-3 protein eps
226058_c0	4.3	hypothetical prote	249288_c2	-3.0	60s ribosomal protein l9-b							256719_c2	1.7	heat shock prot	223900_c0	3.2	fatty acid elon	178240_c0	3.4	---NA---	254293_c0	5.9	tldc domain-conta
223253_c0	4.3	elongation factor	257788_c0	-3.0	d-xylulose 5-phosphate d-fructose 6-phosphate phosphoketolase							225458_c0	-1.2	---NA---	253218_c0	3.2	---NA---	141609_c0	3.4	---NA---	246807_c1	5.9	---NA---
245242_c1	4.3	hypothetical prote	258646_c0	-3.0	elongation factor 2							254814_c0	-1.3	pyrroline-5-carb	254948_c0	3.1	---NA---	239114_c1	3.4	---NA---	257804_c0	5.9	chaperone dnak
143452_c0	4.3	mismatched base	239295_c0	-3.0	60s ribosomal protein l20							219089_c0	-1.4	---NA---	209014_c0	3.1	---NA---	258466_c3	3.3	pa2l_caeel am	253814_c1	5.8	---NA---
229809_c0	4.3	---NA---	255217_c0	-3.0	alpha- sarcomeric-like isoform 1							230944_c0	-1.6	---NA---	216758_c0	3.1	---NA---	253218_c1	3.3	---NA---	199123_c0	5.8	zinc finger bed doi
187991_c0	4.2	---NA---	258785_c0	-3.0	polyketide synthase							230649_c0	-1.6	---NA---	248708_c0	3.1	predicted prot	30126_c0	3.3	---NA---	31401_c0	5.8	bestrophin family
146311_c0	4.2	fic domain-contair	179442_c0	-3.0	acto_acaca ame: full=actobindin							206806_c0	-1.6	---NA---	237663_c0	3.1	zinc transporte	256088_c1	3.3	pa2l_caeel am	256094_c2	5.8	luminal-binding p
250219_c0	4.2	---NA---	258488_c0	-3.0	prophenoloxidase							268553_c0	-1.7	cre-hsp- protein	253841_c1	3.1	---NA---	253814_c2	3.3	---NA---	258554_c0	5.8	heat shock
254705_c0	4.2	---NA---	200680_c0	-3.0	ribosomal protein s9							261782_c0	-1.7	---NA---	251523_c1	3.1	---NA---	232048_c0	3.3	phosphoglyce	212477_c0	5.8	guanine deaminas
238779_c0	4.2	glucose-repressib	222943_c0	-3.0	---NA---							143948_c0	-1.8	cre-hsp- protein	251015_c2	3.1	---NA---	93165_c1	3.2	---NA---	246769_c1	5.8	40s ribosomal prot
233918_c1	4.1	hypothetical prote	256810_c0	-3.0	chaperonin 60							144283_c0	-1.8	---NA---	242393_c0	3.1	---NA---	256802_c0	3.2	---NA---	254031_c5	5.8	adenosylhomocys

Dry-Water			Water-15m			Water-1h			Water-8h			Water-24h			Water-48h			Water-7d			Water-J2		
contig	fold	blast	contig	fold	blast	contig	fold	blast	contig	fold	blast	contig	fold	blast	contig	fold	blast	contig	fc	blast	contig	fold	blast
G5558	4.1	---NA---							G11130	1.8	protein nep-1	G10190	3.9	---NA---	G12558	3.9	---NA---	G9591	3.0	---NA---	G4594	7.0	rna recognition motif domain containing prot
G4328	4.1	sulfotransferase										G5673	3.7	protein nep- isoform a	G11539	3.7	---NA---	G7024	3.0	rbp-1 protein	G11538	7.0	---NA---
G4109	3.5	---NA---										G9591	3.7	---NA---	G1518	3.7	---NA---	G10190	3.0	---NA---	G8588	6.9	---NA---
G3188	3.4	peroxiredoxin 1 variant 2										G5674	3.3	protein nep- isoform a	G7024	3.6	rbp-1 protein	G5674	2.9	protein nep- isof	G8583	6.5	---NA---
G9820	3.4	protein										G3826	3.3	pectate lyase 2	G10190	3.5	---NA---	G7938	2.9	---NA---	G11811	6.4	---NA---
G6929	3.3	protein - caenorhabditis elegans										G11848	3.0	protein cht-2	G3030	3.2	---NA---	G4316	2.7	phosphoglycerat	G9818	6.4	---NA---
G5557	3.0	cuticlin 1										G3528	3.0	acid phosphatase-1	G8539	3.1	---NA---	G5673	2.6	protein nep- isof	G7024	6.1	rbp-1 protein
G7617	3.0	phosphoglycerate mutase family protein										G10138	3.0	---NA---	G9582	3.1	---NA---	G5283	2.6	isoform a	G2906	6.1	gland protein g19b10
G2118	3.0	solute carrier family facilitated glucose transporter member 1-like										G6370	2.9	protein ugt-54	G5283	3.1	isoform a	G7028	2.5	---NA---	G11539	6.0	---NA---
G1115	3.0	---NA---										G4741	2.8	alpha-carbonic anhydra	G7026	3.1	rbp-1 protein	G1518	2.4	---NA---	G8539	6.0	---NA---
G10683	2.8	cre-gst-9 protein										G9626	2.7	---NA---	G10861	3.1	rbp-1 protein	G9887	2.4	fad binding doma	G8902	6.0	---NA---
G5719	2.7	me1 protein										G7081	2.7	beta- -endoglucanase	G9591	3.0	---NA---	G1458	2.3	---NA---	G12359	5.9	xylitol oxidase
G2900	2.5	---NA---										G9887	2.7	fad binding domain-cor	G5991	3.0	c52 protein	G3826	2.3	pectate lyase 2	G9969	5.9	bestrophin family protein
G6462	2.5	redox-regulatory protein fam213a										G7095	2.7	pectate lyase 1	G1458	3.0	---NA---	G10062	2.3	---NA---	G12558	5.9	---NA---
G7499	2.4	---NA---										G1553	2.7	---NA---	G7938	3.0	---NA---	G2083	2.3	lysosomal protec	G10749	5.9	rbp-1 protein
G4565	2.4	dna ligase d										G4316	2.7	phosphoglycerate muta	G7930	3.0	---NA---	G12558	2.3	---NA---	G1518	5.8	---NA---
G4949	2.4	mitogen-activated protein kinase kinase kinase mlt-like										G11341	2.7	rbp-1 protein	G13038	3.0	---NA---	G10138	2.3	---NA---	G7388	5.8	---NA---
G3523	2.4	btb poz domain containing protein										G7269	2.7	arabinogalactan endo- -	G12348	3.0	---NA---	G12772	2.3	rbp-1 protein	G8338	5.7	---NA---
G1468	2.4	mannosyl oligosaccharide glucosidase										G13137	2.6	protein chil-4	G7027	3.0	---NA---	G7026	2.3	rbp-1 protein	G7031	5.7	---NA---
G976	2.3	galectin-4 isoform 1										G9188	2.6	fatty acid elongation pr	G5673	2.9	protein nep- isofo	G8539	2.2	---NA---	G10861	5.7	rbp-1 protein
G9280	2.3	hypothetical protein CRE_22550										G9520	2.6	expansin partial	G10050	2.9	---NA---	G10050	2.2	---NA---	G7710	5.7	---NA---
G7884	2.3	---NA---										G4478	2.5	transport and golgi orga	G12961	2.9	---NA---	G5991	2.2	c52 protein	G3391	5.6	transmembrane protein 135-like
G6289	2.3	Protein Y47D3B.1										G5298	2.5	extracellular solute-bin	G10864	2.9	rbp-1 protein	G6575	2.2	beta- -endogluca	G10462	5.5	---NA---
G1374	2.3	---NA---										G8616	2.3	histidine acid phosphat	G7025	2.9	---NA---	G12961	2.2	---NA---	G1455	5.5	---NA---
G5857	2.2	f-box only protein 25 isoform 1										G8878	2.3	expansin partial	G4316	2.9	phosphoglycerate	G5016	2.2	ubiquitin c	G4303	5.5	---NA---
G7215	2.2	af344865_1 esophageal gland cell protein hgg-20										G10850	2.2	sodium bicarbonate tra	G5016	2.8	ubiquitin c	G8108	2.2	tartrate-resistant	G9297	5.4	rbp-1 protein
G14	2.2	trans- -dihydrobenzene- -diol dehydrogenase										G926	2.2	lysosomal protective	G7028	2.8	---NA---	G9582	2.1	---NA---	G6703	5.4	---NA---
G6489	2.2	hydroxyethylthiazole kinase										G2362	2.1	protein	G7451	2.8	---NA---	G4740	2.1	protein isoform a	G7677	5.3	---NA---
G8038	2.1	embryonic fatty acid-binding protein bm-fab-1										G12887	2.0	conserved hypothetical	G12265	2.8	---NA---	G8362	2.1	---NA---	G7027	5.3	---NA---
G1063	2.1	epoxide hydrolase 1-like										G11425	2.0	---NA---	G4740	2.8	protein isoform a	G9626	2.1	---NA---	G1311	5.3	---NA---
G11882	2.1	domain containing protein										G4076	2.0	protein ugt-47	G12772	2.8	rbp-1 protein	G11539	2.1	---NA---	G4045	5.2	---NA---
G5897	2.1	small heat shock protein										G8230	1.9	cathepsin z precursor	G5674	2.8	protein nep- isofo	G3528	2.1	acid phosphatase	G13533	5.2	---NA---
G13143	2.1	cre-ugt-64 protein										G11353	1.8	protein	G12455	2.7	---NA---	G7301	2.1	---NA---	G13070	5.2	rbp-1 protein
G8058	2.1	protein ztf-2										G6359	1.8	acid sphingomyelinase-	G10138	2.7	---NA---	G1134	2.1	protein isoform a	G5290	5.2	---NA---
G4156	2.1	---NA---										G11130	1.6	protein nep-1	G3826	2.7	pectate lyase 2	G9300	2.1	hypothetical prot	G10156	5.2	beta- -endoglucanase
G3180	2.1	transthyretin-like family protein										G2576	1.3	fructose-bisphosphate a	G13147	2.7	phosphatidylinosi	G10462	2.1	---NA---	G1136	5.1	guanine deaminase
G7759	2.1	---NA---										G4065	1.3	protein isoform c	G6961	2.7	---NA---	G12265	2.1	---NA---	G10864	5.1	rbp-1 protein
G7542	2.0	---NA---										G9819	-2.1	rna-binding protein lin-	G7038	2.7	---NA---	G7038	2.1	---NA---	G13509	5.1	---NA---
G3019	2.0	protein dj-1-like										G4996	-2.1	---NA---	G13030	2.6	---NA---	G13147	2.1	phosphatidylinos	G3821	5.1	---NA---
G1583	2.0	u3 small nucleolar rna interacting protein 2													G13033	2.6	---NA---	G1553	2.1	---NA---	G12022	5.1	---NA---
G9047	2.0	acyl- desaturase													G7820	2.6	---NA---	G10864	2.1	rbp-1 protein	G12626	5.1	---NA---
G4326	2.0	protein xbx- isoform b													G2083	2.6	lysosomal protecti	G12348	2.1	---NA---	G6623	5.0	---NA---
G4406	2.0	cre-rle-1 protein													G10201	2.6	---NA---	G6172	2.1	aspartic protease	G11544	5.0	hypothetical protein CAEBREN_08140
G11320	2.0	high mobility group protein													G6370	2.6	protein ugt-54	G7081	2.0	beta- -endogluca	G10609	5.0	glycosyl hydrolase family 31 protein
G7106	2.0	high mobility group protein													G13014	2.6	tartrate-resistant a	G7031	2.0	---NA---	G12796	5.0	---NA---
G8793	2.0	heat shock protein beta-1-like isoform 1													G7301	2.6	---NA---	G6370	2.0	protein ugt-54	G7930	5.0	---NA---
G1292	2.0	molybdenum cofactor synthesis protein 3													G8108	2.6	tartrate-resistant a	G7025	2.0	---NA---	G9582	5.0	---NA---
G3476	2.0	dual specificity catalytic domain containing protein													G11851	2.5	---NA---	G3391	2.0	transmembrane p	G10201	5.0	---NA---
G8839	1.9	protein ugt-64													G3528	2.5	acid phosphatase-	G6621	2.0	---NA---	G12852	5.0	---NA---
G4029	1.9	---NA---													G5290	2.5	---NA---	G11848	2.0	protein cht-2	G7962	4.9	---NA---
G3472	1.9	---NA---													G7081	2.5	beta- -endoglucan	G7027	1.9	---NA---	G11026	4.9	---NA---
G9808	1.9	thiazole biosynthetic enzyme													G1553	2.5	---NA---	G7930	1.9	---NA---	G6188	4.9	protein sams- isoform a
G9645	1.9	protein isoform b													G6661	2.5	peptidase c13 fam	G10861	1.9	rbp-1 protein	G13038	4.9	---NA---
G8601	1.9	---NA---													G1134	2.5	protein isoform a	G13014	1.9	tartrate-resistant	G7025	4.9	---NA---
G6983	1.9	protein ttr- isoform a													G11891	2.5	---NA---	G13137	1.9	protein chil-4	G7026	4.9	rbp-1 protein
G3665	1.9	methylthioribulose-1-phosphate dehydratase													G9300	2.5	hypothetical prote	G11851	1.9	---NA---	G4333	4.9	protein isoform a
G12858	1.9	aaa atpase domain containing protein													G9887	2.5	fad binding domai	G4225	1.9	major facilitator s	G12171	4.9	rbp-1 protein
G13175	1.9	intermediate filament protein													G6621	2.5	---NA---	G13533	1.9	---NA---	G6644	4.9	---NA---
G2063	1.9	protein best-8													G11649	2.5	---NA---	G12170	1.9	---NA---	G11368	4.9	aspartic protease sp-2

2
3

Water-Dry			Water-5h			Water-24h			Water-48h		
contig	**fold**	**blast**	**contig**	**fold**	**blast**	**contig**	**fold**	**blast**	**contig**	**fold**	**blast**
22290_c0_seq1	2.2	---NA---	38133_c0_seq1	1.5	rna-dependent rna polymerase	38133_c0_seq1	1.4	rna-dependent rna polymerase	27594_c0_seq4	3.2	conserved protein
16627_c0_seq1	2.0	conserved protein	27443_c0_seq3	1.5	---NA---	33540_c0_seq10	1.1	---NA---	29583_c0_seq1	1.7	---NA---
23660_c0_seq1	1.7	---NA---	35102_c0_seq2	1.3	---NA---	17717_c0_seq1	1.0	polyprotein	27513_c0_seq2	1.4	melibiase family protein
30213_c0_seq1	1.7	---NA---	17717_c0_seq1	1.0	polyprotein	27443_c0_seq3	1.0	---NA---	28057_c0_seq1	1.4	melibiase family protein
25583_c0_seq1	1.7	recepotor type guanyly cyclase	23052_c0_seq2	0.9	protein lec- isoform b	28234_c0_seq14	0.9	---NA---	38133_c0_seq1	1.3	rna-dependent rna polymerase
28910_c0_seq2	1.7	---NA---	24760_c0_seq1	0.9	---NA---	35102_c0_seq2	0.8	---NA---	31775_c0_seq89	1.3	recepotor type guanyly cyclase
35432_c1_seq1	1.6	---NA---	19531_c0_seq2	0.8	nadh-ubiquinone oxidoreductase subun	22334_c0_seq10	0.7	---NA---	36550_c0_seq8	1.2	recepotor type guanyly cyclase
3149_c0_seq1	1.6	---NA---	28749_c0_seq6	0.7	heat shock protein 90	24165_c0_seq1	0.7	protein lron-10	36229_c0_seq7	1.2	recepotor type guanyly cyclase
25217_c0_seq1	1.5	---NA---	25840_c0_seq2	-0.4	protein ttr- isoform a	37938_c0_seq47	0.6	zinc knuckle family protein	31937_c0_seq8	1.2	protein gcy-9
25727_c0_seq1	1.5	rrna intron-encoded homing endonuclea	21405_c0_seq5	-0.4	galectin like protein	23490_c0_seq3	0.6	defective mitochondrial respiration	30331_c0_seq14	1.2	protein gcy-9
25931_c0_seq8	1.5	cold-shock dna-binding domain-containi	19783_c0_seq4	-0.4	---NA---	28405_c0_seq1	0.6	kh domain containing protein	23052_c0_seq2	1.1	protein lec- isoform b
15211_c0_seq2	1.5	---NA---	32830_c0_seq5	-0.4	serine carboxypeptidase	31727_c0_seq2	0.5	ubiquitin-activating enzyme e1	24760_c0_seq1	1.1	---NA---
24140_c0_seq2	1.5	---NA---	32375_c0_seq1	-0.4	matrixin family protein	29151_c0_seq3	0.5	kh domain containing protein	31150_c0_seq1	1.1	amine flavin-containing
22161_c0_seq3	1.4	cysteine-rich pdz-binding	27085_c0_seq1	-0.5	udp-galactose transporter family protei	31311_c0_seq8	0.5	heat shock protein 23-like	37639_c0_seq1	1.1	aaa atpase domain containing protein
24928_c0_seq3	1.4	alpha beta hydrolase fold protein	32589_c0_seq4	-0.5	long-chain-fatty-acid-- ligase 1 isoform 1	27085_c0_seq1	-0.4	udp-galactose transporter family pr	26477_c0_seq5	1.1	---NA---
25333_c0_seq1	1.4	---NA---	32661_c0_seq3	-0.5	long-chain-fatty-acid-- ligase 1 isoform 1	24493_c0_seq1	-0.4	troponin skeletal muscle-like	37791_c0_seq1	1.0	aaa atpase domain containing protein
38133_c0_seq1	1.4	rna-dependent rna polymerase	31242_c0_seq1	-0.5	phenylalanine hydroxylase	26137_c0_seq32	-0.5	fimbrin plastin	32463_c0_seq3	0.9	tartrate-resistant acid phosphatase type 5
23682_c0_seq1	1.4	---NA---	31820_c0_seq1	-0.5	malate dehydrogenase	28211_c0_seq4	-0.5	---NA---	26963_c0_seq5	0.9	hypothetical protein WUBG_09559
23024_c0_seq1	1.4	hypothetical protein TSTA_040370	26257_c0_seq1	-0.5	udp-galactose transporter family protei	31890_c0_seq13	-0.5	hypothetical protein CRE_13221	17717_c0_seq1	0.9	polyprotein
21754_c0_seq1	1.4	---NA---	37684_c0_seq1	-0.5	bifunctional polynucleotide phosphatas	28182_c0_seq5	-0.5	fimbrin plastin	31727_c0_seq2	0.9	ubiquitin-activating enzyme e1
25686_c0_seq24	1.4	alpha beta hydrolase fold protein	26125_c0_seq5	-0.6	---NA---	26566_c0_seq1	-0.5	nadph--cytochrome p450 reductase	26938_c0_seq14	0.9	---NA---
22652_c0_seq1	1.3	cysteine-rich pdz-binding	26984_c0_seq7	-0.6	---NA---	36111_c0_seq1	-0.5	protein tyr-6	23490_c0_seq3	0.9	defective mitochondrial respiration famil
25073_c0_seq6	1.3	glutathione peroxidase	29831_c0_seq1	-0.6	---NA---	19825_c0_seq1	-0.5	39s ribosomal protein mitochondria	28405_c0_seq1	0.9	kh domain containing protein
15839_c0_seq1	1.3	---NA---	32433_c0_seq9	-0.6	cre-hsp- protein	30596_c0_seq17	-0.5	hypothetical kda protein in chromos	29259_c0_seq2	0.8	ubiquitin-activating enzyme e1
21286_c0_seq1	1.3	---NA---	31913_c0_seq1	-0.6	trans- -dihydrobenzene- -diol dehydrog	28386_c0_seq3	-0.5	small heat shock protein	33544_c0_seq2	0.8	galectin 4
17717_c0_seq1	1.3	polyprotein	30765_c0_seq7	-0.6	cre-hsp- protein	29597_c0_seq18	-0.5	protein isoform a	17555_c0_seq2	0.8	nucleoside diphosphate kinase
37316_c1_seq1	1.3	fork head domain-containing protein	35682_c0_seq2	-0.6	trans- -dihydrobenzene- -diol dehydrog	38021_c1_seq5	-0.5	serine threonine-protein phosphata	29151_c0_seq3	0.8	kh domain containing protein
30431_c0_seq2	1.3	retinoic acid receptor gamma-a isoform 2	29696_c0_seq8	-0.6	---NA---	31018_c0_seq4	-0.5	beta -mannosyltransferase egh	31617_c0_seq2	0.8	pre-mrna-processing-splicing factor 8
35794_c0_seq1	1.3	partial	22488_c0_seq1	-0.6	high mobility group protein	34002_c0_seq19	-0.6	galectin like protein	33467_c0_seq1	0.8	tartrate-resistant acid phosphatase type 5
22649_c0_seq2	1.3	---NA---	30081_c0_seq2	-0.6	annexin b11 isoform a	23414_c0_seq5	-0.6	small heat shock protein	22040_c0_seq3	0.8	atp synthase subunit mitochondrial-like
20752_c0_seq4	1.3	---NA---	29014_c0_seq9	-0.6	---NA---	24992_c0_seq60	-0.6	protein tyr-5	32295_c0_seq2	0.8	pre-mrna-processing-splicing factor 8
27267_c0_seq1	1.3	---NA---	29905_c1_seq40	-0.6	hypothetical kda protein in chromosome	26393_c0_seq38	-0.6	protein tyr-5	26369_c0_seq1	0.7	galectin-9 isoform 2
24314_c0_seq4	1.2	hypothetical protein LOAG_06975	35256_c0_seq28	-0.6	selenium binding protein 1	37330_c0_seq27	-0.6	c-1-tetrahydrofolate cytoplasmic	36134_c0_seq4	-0.2	---NA---
17957_c0_seq2	1.2	polyprotein 1	22257_c0_seq1	-0.6	high mobility group protein	32917_c0_seq3	-0.6	lysosomal protective	29052_c0_seq6	-0.2	atp-binding cassette sub-family f membe
14822_c0_seq1	1.2	protein	35761_c0_seq9	-0.6	Hypothetical protein CBG16435	29696_c0_seq8	-0.6	---NA---	37580_c0_seq1	-0.2	myosin heavy chain



Water-Dry			Water-5h						Water-48h		
contig	fold	blast	contig	fold	blast		fold	blast	contig	fold	blast
1184700	1.8	---NA---	1093500	0.7	galactoside-binding lectin family protein		0.5	kh domain containing protein	637500	0.8	recepotor type guanyly cyclase
879600	1.7	---NA---	887800	0.5	heat shock protein 90		-0.4	---NA---	1226500	0.8	recepotor type guanyly cyclase
1234100	1.6	alpha beta hydrolase fold protein	458900	-0.5	a-macroglobulin complement component family protein		-0.5	carnitine o-palmitoyltransferase	1093500	0.8	galactoside-binding lectin family protein
717000	1.5	dorsal gland cell-specific expression prot	751700	-0.6	protein ctg- isoform a		-0.5	phosphoenolpyruvate carboxyki	1182100	0.5	ubiquitin-activating enzyme e1
936300	1.5	glutathione synthetase-like	649700	-0.6	epoxide hydrolase 1		-0.5	ammonium transporter	434900	0.5	kh domain containing protein
1281800	1.4	n-myc downstream regulated	952500	-0.6	delta-1-pyrroline-5-carboxylate synthase-like		-0.5	epoxide hydrolase 1	1276900	-0.4	daf-16
1153000	1.4	glutathione peroxidase	966200	-0.6	matrixin family protein		-0.5	protein mdt- isoform b	815500	-0.5	---NA---
598500	1.4	protein ttr-54	340200	-0.6	lysosomal protective		-0.6	hypothetical protein CRE_21243	1470900	-0.5	myosin heavy chain
360700	1.4	fork head domain-containing protein	275100	-0.6	protein lec- isoform b		-0.6	udp-galactose transporter family	1167200	-0.5	---NA---
1017900	1.3	ubiquitin-conjugating enzyme e2 g1	3800	-0.6	protein acs-5		-0.6	matrixin family protein	250900	-0.5	vinculin-like isoform 2
828500	1.3	---NA---	780500	-0.7	trans- -dihydrobenzene- -diol dehydrogenase		-0.6	---NA---	983400	-0.5	dihydropteridine reductase
36300	1.3	---NA---	159400	-0.7	intermediate filament protein		-0.7	protein ctg- isoform a	84100	-0.5	g patch domain and kow motifs-containing protein
1210900	1.3	---NA---	962100	-0.7	---NA---		-0.7	protein isoform b	427900	-0.6	long-chain-fatty-acid-- ligase 1
876400	1.3	---NA---	623100	-0.7	phenylalanine hydroxylase		-0.7	delta-1-pyrroline-5-carboxylate s	92300	-0.6	ammonium transporter
274200	1.3	---NA---	706100	-0.7	udp-galactose transporter family protein		-0.7	protein acs-5	40000	-0.6	myosin tail family protein
276700	1.3	trna-dihydrouridine synthase	780700	-0.7	chloride channel protein		-0.7	protein - caenorhabditis elegans	1002700	-0.6	btb poz domain-containing protein 2
446100	1.3	baculoviral iap repeat-containing	1392200	-0.8	protein isoform b		-0.7	intermediate filament protein	1237200	-0.6	matrixin family protein
138100	1.3	protein mvb-12	171600	-0.8	annexin b11 isoform a		-0.7	lysosomal protective	202100	-0.6	protein srap- isoform a
137700	1.3	cdp-alcohol phosphatidyltransferase	667800	-0.8	protein isoform a		-0.7	zinc finger ccch type domain con	784300	-0.6	fatty-acid amide hydrolase 2-like
603200	1.3	probable nadh dehydrogenase 1 alpha su	159200	-0.8	hypothetical kda protein in chromosome		-0.8	protein lec- isoform b	649700	-0.6	epoxide hydrolase 1
213700	1.3	hypothetical protein WUBG_09496	1402100	-1.1	tm2 domain-containing protein almondex-like		-0.8	chloride channel protein	282900	-0.6	cre-nhr-92 protein
422500	1.2	molybdenum cofactor synthesis 3	1294200	-1.1	adipocyte plasma membrane-associated protein		-0.8	a chain crystal structure of h2o2 t	1434500	-0.6	cytoplasmic intermediate filament protein
1526700	1.2	rna-binding protein lin-28	672400	-1.1	peroxiredoxin 1 variant 2		-0.8	fatty-acid amide hydrolase 2-like	966200	-0.6	matrixin family protein
422400	1.2	adenylyltransferase and sulfurtransferas	1340000	-1.2	---NA---		-0.8	annexin b11 isoform a	896900	-0.6	---NA---
1340000	1.2	---NA---	672300	-1.5	peroxiredoxin 1 variant 2		-0.9	trans- -dihydrobenzene- -diol de	687500	-0.6	protein erm- isoform a
412900	1.2	zinc knuckle family protein	475300	-1.7	alcohol dehydrogenase class-3		-0.9	protein isoform a	56600	-0.6	cre-cyp-13a8 protein
550300	1.2	protein mrpl-19					-1.0	phosphatidylethanolamine-bind	478400	-0.7	Protein C48E7.1
1458300	1.2	adp-specific phosphofructokinase glucokinase conserved region family protein					-1.2	hypothetical kda protein in chror	159400	-0.7	intermediate filament protein
1587900	1.2	hypothetical kda protein in chromosome					-1.2	adipocyte plasma membrane-ass	506200	-0.7	atp-binding cassette sub-family f member 1-like
884100	1.2	alkyldihydroxyacetonephosphate peroxisomal					-1.3	peroxiredoxin 1 variant 2	90700	-0.7	major facilitator superfamily protein
941300	1.2	exosome complex exonuclease rrp46					-1.6	peroxiredoxin 1 variant 2	952500	-0.7	delta-1-pyrroline-5-carboxylate synthase-like
346900	1.2	---NA---					-1.7	---NA---	3800	-0.7	protein acs-5
563700	1.2	bhlhzip transcription factor max bigmax					-1.9	alcohol dehydrogenase class-3	464700	-0.7	rna recognition
1008100	1.2	rna-binding protein lin-28							887800	-0.7	heat shock protein 90
1084700	1.2	tfiih basal transcription factor complex helicase xpb subunit							499900	-0.7	af273731_1hypothetical esophageal gland cell secretory protein 4

6
7

5.4. Appendix D

Table S6: Cluster of transcripts from *G. rostochiensis* Trinity transcriptome with expression similar to *trehalose 6-phosphate synthase.*

Contig	up-regulated Treatment	BLAST	Fold Change hydrated cyst vs dry cyst
comp24112_c0	Dry	histone h3	1.9
comp224595_c0	Dry	mgc83793 protein	1.6
comp246692_c0	Dry	ring finger and ccch-type zinc finger domain-containing protein 2	4.0
comp251106_c0	Dry	tubulin polyglutamylase ttll11-like	2.0
comp254088_c1	Dry	serine palmitoyltransferase 2	1.7
comp256950_c0	Dry	ribosomal protein s7p s5e containing protein	1.3
comp266765_c0	Dry	NA	3.7
comp188635_c0	Dry	histone	1.9
comp226839_c0	Dry	protein cogc-8	1.6
comp248755_c0	Dry	NA	2.0
comp251314_c3	Dry	adenylate kinase 2	2.5
comp254759_c0	Dry	probable trans-2-enoyl- mitochondrial-like	2.0
comp257619_c0	Dry	transcription initiation factor iib	1.6
comp208184_c0	Dry	hypothetical protein Bm1_35525	4.3
comp236180_c0	Dry	cre-rle-1 protein	3.7

comp249325_c0	Dry	g10 protein	1.9
comp252340_c0	Dry	cytoplasmic intermediate filament protein	2.1
comp254763_c0	Dry	NA	1.9
comp257819_c0	Dry	probable dolichyl pyrophosphate glc1man9 c2 alpha-glucosyltransferase	2.8
comp220998_c0	Dry	cop9 signalosome complex subunit	1.9
comp238646_c0	Dry	NA	4.0
comp250328_c1	Dry	NA	1.9
comp253394_c0	Dry	g10 protein	1.7
comp256639_c0	Dry	NA	2.1
comp258202_c0	Dry	islet cell autoantigen 1	2.3
comp222505_c0	Dry	cytoplasmic intermediate filament protein	2.7
comp246505_c0	Dry	NA	1.6
comp250328_c0	Dry	NA	3.5
comp253978_c0	Dry	polypyrimidine tract binding protein	1.8
comp256753_c0	Dry	trehalose 6-phosphate synthase	3.2
comp258460_c2	Dry	NA	2.2

5.5. Appendix E

Table S7: Cluster of transcripts from *G. rostochiensis* Trinity transcriptome with putative expression patterns of genes involved in cyst survival.

Contig	up-regulated Treatment	BLAST	Fold change dry cyst vs hydrated cyst
comp89407_c0	Dry	NA	3.7
comp140507_c0	Dry	NA	3.0
comp204557_c0	Dry	Selenoprotein	6.1
comp232167_c0	Dry	hypothetical protein CRE_16869	2.0
comp248011_c0	Dry	NA	2.0
comp250515_c1	Dry	Annexin	2.3
comp255917_c0	Dry	lysine histidine transporter	2.1
comp256424_c1	Dry	Zinc finger protein	1.7
comp258276_c1	Dry	briggsae cbr-pap-1 protein	2.8
comp258560_c0	Dry	NA	2.8

5.6. Appendix F

Table S8: Cluster of transcripts from *G. rostochiensis* Trinity transcriptome with expression similar to *nep-1.*

Contig	up-regulated Treatment	BLAST	Mean fold-change in up-regulated hatching treatments
comp91041_c0	48h-7d-J2	gland protein g20e03	4.9
comp140896_c0	8h-J2	protein nep-1	6.5
comp213450_c1	8h-J2	nucleoside diphosphate kinase	24.3
comp220622_c0	J2	Patched familly protein	NA
comp226361_c0	J2	NA	NA
comp235689_c0	J2	heat shock protein hsp20	NA
comp241314_c0	J2	beta-endoglucanase-1 precursor	NA
comp242450_c0	7d-J2	major facilitator superfamily domain-containing protein 8	NA
comp249021_c0	48h-J2	NA	5.3
comp251094_c0	J2	patched family protein	NA
comp263359_c0	J2	NA	NA

5.7. Appendix G

Table S9: Cluster of transcripts from *G. rostochiensis* Trinity transcriptome with putative expression pattern of genes involved in hatching.

Contig	Up-regulated treatment	BLAST	mean fold change between up-regulated hatching treatment
comp152061_c0	24h-48h	pectate lyase 2	8.3
comp178494_c0	48h	galactosylgalactosylxylosylprotein 3-beta-glucuronosyltransferase 2	12.1
comp252092_c1	48h	NA	9.2
comp79822_c0	24h-48h	pectate lyase 2	9.8
comp186241_c0	24h-48h	Peptidase M14, carboxypeptidase A domain and Proteinase inhibitor	6.1
comp220435_c0	24h-48h	Peptidase M14, carboxypeptidase A domain and Proteinase inhibitor	4.9
comp143987_c0	24h-48h	fatty acid elongation protein 3	7.0
comp205597_c0	24h-48h	carbonate dehydratase	5.3
comp179924_c0	24h-48h	ABC transporter substrate-binding protein	6.5
comp204845_c0	48h	lysosomal acid phosphatase	5.3
comp146670_c0	24h-48h	pectate lyase 1	5.9
comp236889_c0	24h-48h	general substrate transporter and Major facilitator superfamily domain-containing	6.1

		protein	
comp197008_c0	24h-48h	ABC transporter substrate-binding protein	6.7

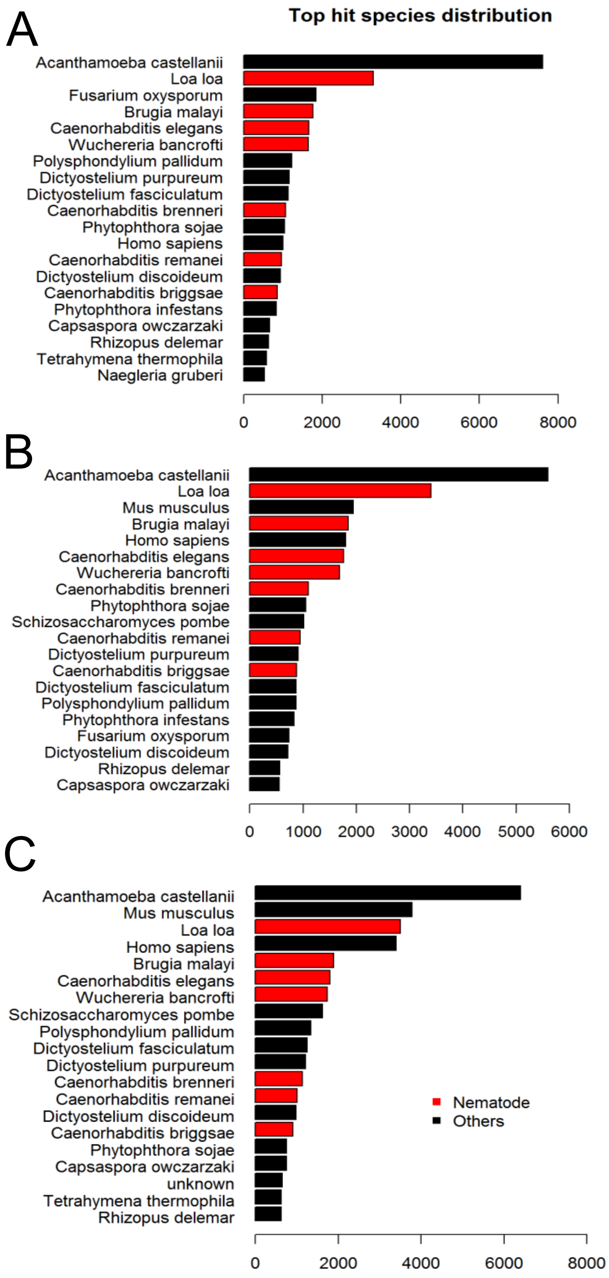
5.8. Appendix H

Table S10: Information and primer sequences used in this study

Gene symbol	Gene description	Forward primer	Reverse primer	Amplicon length (bp)	Amplification efficiency (Log)
Pmp-3	Putative membrane transporter	CTGGTTGCTGAGCAGGATAA	GATGAAGCCCGATTGGTAGAA	102	1.92
AaRS	Aminoacyl tRNA synthetase	CGGATTTACGGACCTTGTCTAC	GGGAATCCGTCACGCTTAAT	84	1.98
GR	Glutathione reductase	TTGAGAGACCATGCCGATTAC	GAGTTGAGACGCCGAATGT	102	1.90
Nep-1	Neprilysin	GCTGAAATGGTGGAGAAAGTG	TTTGACGCCCGAGTAGAAG	457	1.91

5.9. Appendix I

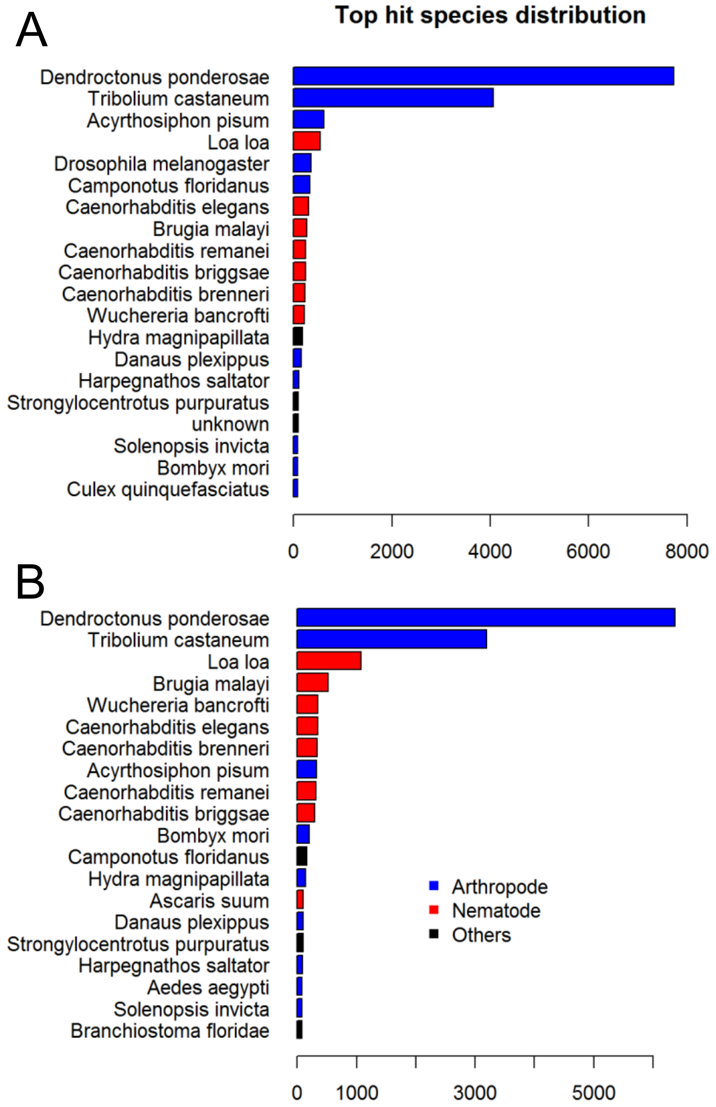
Figure S1. Species distribution of *G. rostochiensis* transcriptome.



Species distribution of *G. rostochiensis* transcriptome. Distribution of the best BLAST hit of each transcript in the **A)** DeconSeq transcriptome, **B)** CCRbC transcriptome and **C)** BLAST transcriptome. Nematode species are in red. Only the top 20 is shown.

5.10. Appendix J

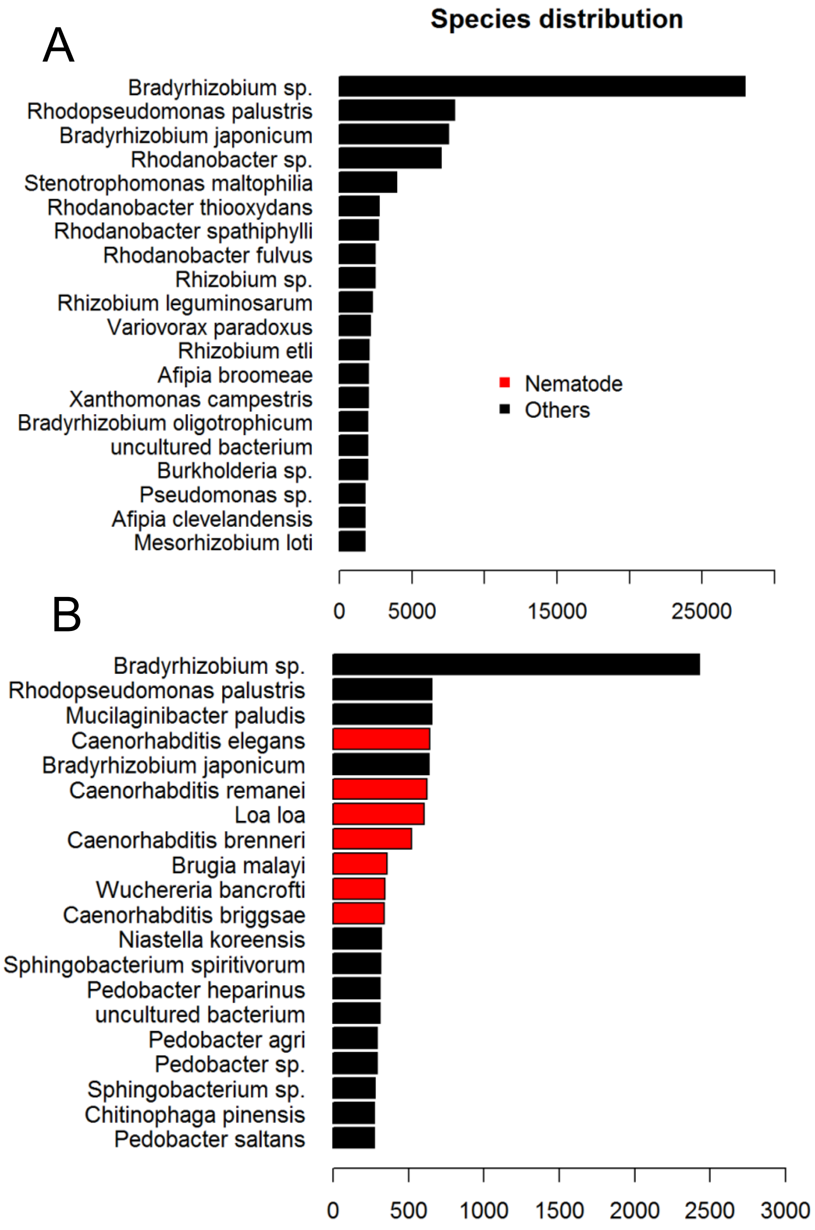
Figure S2. Species distribution of *L. oregonensis* transcriptome.



Species distribution of *G. rostochiensis* transcritpome. Distribution of the best BLAST hit of each transcript in the **A)** Trinity transcriptome, **B)** MCSC transcriptome. Only the top 20 is shown.

5.11. Appendix K

Figure S3. Species distribution of *G. rostochiensis*



Species distribution of *G. rostochiensis* reads. Distribution of the best BLAST hit of each reads in the **A)** Raw reads, **B)** MCSC-decontaminated reads. Only the top 20 is shown.

5.12. Appendix L

Supplemental information S1. *G. rostochiensis* RNA-seq librairie preparation.

Data availability

Globodera rostochiensis Illumina 100bp paired-end reads are available through NCBI under the bioproject accession number PRJNA274143.

Root diffusates

For *Globodera rostochiensis*, potato plants cv. Snowden were grown in perlite, in 2L containers, until they reached about 15 cm-high. At this point, potato root diffusate (PRD) was harvested once a week, for six consecutive weeks, by the method of Fenwick (1949). Briefly, soil was drenched with tap water until saturation. An extra 50 mL of tap water was then added to the pot and the flowing liquid was collected. The collected liquid was used to repeat this procedure two more times. The final collected liquid was filtered (KenAG, D-547) to obtain PRD. PRD samples were kept at 4°C in dark plastic bottles until the last one was harvested. Then, all six weekly-sampled PRDs were pooled, freeze-dried and stored at -20°C. Final volume was recorded prior lyophilization, as well as final weight after lyophilization, for proper PRD reconstitution. PRD was reconstituted from powder with nanopure water at a final concentration of 0.5 X and passed through a 0.2 µm filter prior use.

Sample description

G. rostochiensis cysts were recovered by flotation (Fenwick 1940) from soil samples collected in the fall 2011 in Saint-Amable (Quebec, Canada). Cysts were stored dried for at least one year in the dark at room temperature prior hatching experiments. A time course experiment was setup to study the evolution of the transcriptome of *G. rostochiensis* during hatching. The following physiological stages (treatments) were studied: dry cyst, cyst soaked in water for one week (hydratation), hydrated cysts soaked in PRD for 15 min, 1 h, 8 h, 24 h, 48 h and 7 d and hatched J2 larvae. Each cyst sample contained 1000 cysts placed in a mesh bag (Ankom, F57). Cysts were soaked in 30 mL of filtered (0.2 µm) tap water or 0.5X PRD, in a petri dish. Water and PRD were changed every day. No hatching occurred during the hydration period. Hatched J2s were harvested daily for a two-week period and pooled for further analysis. Experiment was repeated two times.

Total RNA extraction, library preparation and sequencing

For *G. rostochiensis*, cysts soaked in PRD were washed thoroughly with distilled water prior to RNA extraction to remove as much potential contaminants as possible. Samples were homogenized in 700 µL of RTL plus buffer with one 6 mm zirconium bead and ~150 µL of 1 mm zirconium beads using the PowerLyzer 24 homogenizer (MO BIO, Carlsbad, CA, USA) and stored at -80°C until RNA purification. Total RNA was extracted using the RNeasy Plus mini kit (Qiagen, Mississauga, Canada) according to manufacturer's instructions. Total RNA samples were store at -80°C prior to RNA-seq library preparation. RNAs were quantified with the NanoDrop 2000 (Thermo Scientific). RNA integrity was assessed with the Bioanlalyzer 2100 (Agilent Technologies) using the RNA 6000 Nano kit. All RNA samples had a RIN value higher than 7 and a 260/230 ratio value over 2.

Library preparation and sequencing were performed at McGill University and Génome Québec Innovation Centre (Montreal, Canada) using the TruSeq RNA sample prep kit v2 (Illumina) and a HiSeq 2000 sequencer (Illumina). For each replicate, all nine samples were multiplexed and sequenced in one lane for 100 bp paired-end reads.

For *G. pallida*, total RNA was extracted using RNeasy Plus Micro Kit (Qiagen, Hilden, Germany) following the manufacturer's instructions. DNA digestion was conducted on column during RNA extraction using RNase-Free DNase set (Qiagen, Hilden, Germany), as recommended. All RNA samples had a RIN value higher than 7 and a 260/230 ratio value over 2. Total RNA was quantified using a 2100 Bioanalyzer

(Agilent Technologies) and the Small RNA kit (Agilent Technologies) following the manufacturer’s instructions. Libraries and sequencing were produced and sequenced in Sanger Institute facilities. Illumina transcriptome libraries were produced using polyadenylated mRNA purified from total RNA using methods previously described (Choi et al., 2011) except size selection, which was done using the Caliper LabChip XT.

Transcriptome libraries were denatured with 0.1 M sodium hydroxide and diluted to 6 pM in a hybridisation buffer to allow the template strands to hybridise to adapters attached to the flowcell surface. Cluster amplification was performed on the Illumina cluster station or cBOT using the V4 cluster generation kit following the manufacturer’s protocol and then a SYBRGreen QC was performed to measure cluster density and determine whether to pass or fail the flowcell for sequencing, followed by linearization, blocking and hybridization of the R1 sequencing primer. The hybridized flow cells were loaded onto the Illumina Genome Analyser IIX for 76 or 100 cycles of sequencing-by-synthesis using the V4 or V5 SBS sequencing kit then, in situ, the linearization, blocking and hybridization step was repeated to regenerate clusters, release the second strand for sequencing and to hybridise the R2 sequencing primer followed by another 76 or 100 cycles of sequencing to produce paired-end reads. These steps were performed using proprietary reagents according to the manufacturer's recommended protocol (https://icom.illumina.com/). Data were analysed from the Illumina Genome Analyser IIx or HiSeq sequencing machines using the RTA1.6 or RTA1.8 analysis pipelines.

5.13. Appendix M

Supplemental information S2. *L. oregonensis* RNA-seq library preparation.

Data availability

L. oregonensis Illumina 100bp paired-end reads are available through NCBI under the bioproject accession number PRJNA313413.

A comparison of RNA transcripts from carrot weevils infected or not with B. listronoti was realised using next-generation sequencing (RNA-Seq) in order to evaluate the impact of the parasite on its host gene expression. Samples consisted of female adult carrot weevils infected with B. listronoti or healthy females spiked with the B. listronoti content of parasitized individuals obtained by dissection. All the weevils were multiplied in vitro in controlled environment and each treatment was repeated three times. Samples were homogenized in 700 µL of RTL plus buffer (Qiagen, Mississauga, Canada), with one 6 mm zirconium bead and 200 µL of 1 mm zirconium beads, using the PowerLyzer 24 Homogenizer (3 x 45 s, at 2500 rpm; MO BIO, Carlsbad, CA, USA), and stored at -80°C until RNA purification. Total RNA was extracted using the RNeasy Plus mini kit (Qiagen) according to manufacturer's instructions and stored at -80°C prior to library preparation. RNA samples were quantified with the NanoDrop 2000 (ThermoFisher Scientific, Mississauga, Canada) and their integrity was assessed with the Bioanalyzer 2100 (Agilent Technologies) using the RNA 6000 Nano kit. Library preparation and sequencing were performed at the Institute for Research in Immunology and Cancer (IRIC; Université de Montréal, Montreal, QC, Canada) using the TruSeq RNA sample prep kit v2 (Illumina, San Diego, CA, USA) and the HiSeq 2000 sequencer (Illumina). All the samples were multiplexed in a single sequencing lane.

5.14. Appendix N

Supplemental information S3. Contaminant contigs removal by counts (CCRbC).

The contaminant contigs removal by counts (CCRbC) is a transcriptome decontamination method for RNA-seq data. It use as input the counts matrix (*n x m)* produces by Corset (Davidson and Oshlack 2014) where the *n* contigs are represented by *n* rows and the *r* replicates of *t* treatments are represented by *r*t = m* columns. The first step is to sum all treatments together for each replicate and for each contigs. This will results in a *n* by *r* matrix. Non-contaminant contigs are those who have at least one count for every replicates. Contaminant contigs are remove by cutting rows that contains at least one zero in the *n* by *r* matrix.

5.15. Appendix O

Table S1: BLAST of decontaminated transcriptomes on the reference transcriptome.

	Nb transcripts with BLAST hit on ref transcriptome	**Nb reference genes covered by the *de novo* transcriptome**	**% genes covered**
Trinity reduce	24 354	11 928	83.3
DeconSeq	24 142	11 870	82.9
CCRbC	23 063	11 907	83.2
MCSC	20 730	11 036	77.1

Blast of the reduce transcriptome and three decontaminated transcriptome on the reference transcriptome who contains 14,309 genes.

5.16. Appendix P

Table S2: Average P-value of the 1,313 common genes.

	P-value	P-value FDR adjusted
Transcriptome reduce	0.045%	1.38%
Transcritome CCRbC	0.034%	0.998%
Transcriptome DHCS	0.021%	0.761%

BIBLIOGRAPHIE

Anders, S. and Huber, W. (2010) Differential expression analysis for sequence count data. *Genome Biol.,* **11**.

Atkinson, H. J., Taylor, J. D. and Fowler, M. (1987) Changes in the second stage juveniles of *Globodera rostochiensis* prior to hatching in response to potato root diffusate. *Ann. Appl. Biol.,* **110,** 105-114.

Baker, M. (2012) De novo genome assembly: what every biologist should know. *Nature Methods,* **9,** 333-337.

Bélair, G. (2005) Nematodes, these roundworms that harm plants... by their roots. *Phytoprotection,* **86,** 65-69.

Blaauw, R. H., Brière, J.-F., de Jong, R., Benningshof, J. C., van Ginkel, A. E., Fraanje, J.*, et al.* (2001) Intramolecular Photochemical Dioxenone-Alkene [2+ 2] Cycloadditions as an Approach to the Bicyclo [2.1. 1] hexane Moiety of Solanoeclepin A. *The Journal of organic chemistry,* **66,** 233-242.

Bohlmann, H. and Sobczak, M. (2014) The plant cell wall in the feeding sites of cyst nematodes. *Front. Plant Sci.,* **5**.

Brendel, V. P. and Standage, D. S. mRNAmarkup: quality control and annotation of de novo transcriptome assemblies.

Byrne, J., Twomey, U., Maher, N., Devine, K. J. and Jones, P. W. (1998) Detection of hatching inhibitors and hatching factor stimulants for golden potato cyst nematode, *Globodera rostochiensis*, in potato root leachate. *Ann. Appl. Biol.,* **132,** 463-472.

Chu, H. T., Hsiao, W. W. L., Chen, J. C., Yeh, T. J., Tsai, M. H., Lin, H.*, et al.* (2013) EBARDenovo: Highly accurate de novo assembly of RNA-Seq with efficient chimera-detection. *Bioinformatics,* **29,** 1004-1010.

Compeau, P. E., Pevzner, P. A. and Tesler, G. (2011) How to apply de Bruijn graphs to genome assembly. *Nat Biotechnol,* **29,** 987-991.

Cotton, J. A., Lilley, C. J., Jones, L. M., Kikuchi, T., Reid, A. J., Thorpe, P.*, et al.* (2014) The genome and life-stage specific transcriptomes of *Globodera pallida* elucidate key aspects of plant parasitism by a cyst nematode. *Genome Biol,* **15,** R43.

Davidson, N. and Oshlack, A. (2014) Corset: enabling differential gene expression analysis for de novo assembled transcriptomes. *Genome Biol.,* **15,** 410.

Decraemer, W. and Hunt, D. J. (2006) Structure and classification. In: *Plant Nematology.* pp. 3-32.

DeRisi, J. L., Iyer, V. R. and Brown, P. O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science,* **278,** 680-686.

Evans, K., Franco, J. and De scurrah, M. M. (1975) Distribution of species of potato cyst-nematodes in South America. *Nematologica,* **21,** 365-369.

Glaus, P., Honkela, A. and Rattray, M. (2012) Identifying differentially expressed transcripts from RNA-seq data with biological variation. *Bioinformatics,* **28,** 1721-1728.

Glouzon, J. P., Bolduc, F., Wang, S., Najmanovich, R. J. and Perreault, J. P. (2014) Deep-sequencing of the peach latent mosaic viroid reveals new aspects of population heterogeneity. *PLoS One,* **9,** e87297.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I.*, et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol,* **29,** 644-652.

Guo, Y., Li, C. I., Ye, F. and Shyr, Y. (2013) Evaluation of read count based RNAseq analysis methods. *BMC Genomics,* **14**.

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

Hardcastle, T. J. and Kelly, K. A. (2010) BaySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinf.,* **11**.

Jones, J. T., Haegeman, A., Danchin, E. G., Gaur, H. S., Helder, J., Jones, M. G.*, et al.* (2013) Top 10 plant-parasitic nematodes in molecular plant pathology. *Mol. Plant Pathol.,* **14,** 946-961.

Jones, J. T., Robertson, L., Perry, R. N. and Robertson, W. M. (1997) Changes in gene expression during stimulation and hatching of the potato cyst nematode *Globodera rostochiensis*. *Parasitology,* **114,** 309-315.

Kort, J., Ross, H., Rumpenhorst, H. J. and Stone, A. R. (1977) An international scheme for identifying and classifying pathotypes of potato cyst-nematodes *Globodera rostochiensis* and *G. pallida*. *Nematologica,* **23,** 333-339.

Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics,* **25,** 1754-1760.

Li, H. and Durbin, R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics,* **26,** 589-595.

Love, M. I., Huber, W. and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *bioRxiv*.

Nicol, J. M., Turner, S. J., Coyne, D., Den Nijs, L., Hockland, S. and Maafi, Z. T. (2011) Current nematode threats to world agriculture. In: *Genomics and molecular genetics of plant-nematode interactions.* Springer, pp. 21-43.

Nijboer, H. and Parlevliet, J. E. (1990) Pathotype-specificity in potato cyst nematodes, a reconsideration. *Euphytica,* **49,** 39-47.

Perry, R. and Beane, J. (1982) The effects of brief exposures to potato root diffusate on the hatching of *Globodera rostochiensis*. *Revue de Nématologie,* **5,** 221-224.

Perry, R. N., Beane, J., Marett, C. C. and Tylka, G. L. (2002) Comparison of the rate of embryogenic development of *Globodera rostochiensis* and *G. pallida* using flow cytometric analysis. *Nematology,* **4,** 553-555.

Robinson, M., Atkinson, H. and Perry, R. (1987) The influence of soil moisture and storage time on the motility, infectivity and lipid utilization of second stage juveniles of the potato cyst nematodes *Globodera rostochiensis* and *G. pallida*. *Revue de Nématologie,* **10,** 343-348.

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

Schmieder, R. and Edwards, R. (2011) Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLoS One,* **6,** e17288.

Schulz, M. H., Zerbino, D. R., Vingron, M. and Birney, E. (2012) Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics,* **28,** 1086-1092.

Snyder, S. A. (2011) Natural product synthesis: Making nematodes nervous. *Nature Chemistry,* **3,** 422-423.

Sun, F., Miller, S., Wood, S. and Côté, M. J. (2007) Occurrence of potato cyst nematode, *Globodera rostochiensis*, on potato in the Saint-Amable region, Quebec, Canada. *Plant Dis.,* **91,** 908.

Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A. and Conesa, A. (2011) Differential expression in RNA-seq: A matter of depth. *Genome Res.,* **21,** 2213-2223.

Timmermans, B., Vos, J., Stomph, T., Van Nieuwburg, J. and Van der Putten, P. (2007) Field performance of Solanum sisymbriifolium, a trap crop for potato cyst nematodes. II. Root characteristics. *Ann. Appl. Biol.,* **150,** 99-106.

Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L. and Pachter, L. (2013) Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol,* **31,** 46-53.

Tytgat, T., De Meutter, J., Vanholme, B., Claeys, M., Verreijdt, L., Gheysen, G.*, et al.* (2002) Development and pharyngeal gland activities of Heterodera schachtii infecting Arabidopsis thaliana roots. *Nematology,* **4,** 899-908.

Van Walle, I., Lasters, I. and Wyns, L. (2004) Align-m - A new algorithm for multiple alignment of highly divergent sequences. *Bioinformatics,* **20,** 1428-1435.

Velculescu, V. E., Zhang, L., Vogelstein, B. and Kinzler, K. W. (1995) Serial analysis of gene expression. *Science,* **270,** 484-487.

Wang, Z., Gerstein, M. and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet,* **10,** 57-63.

Whitehead, A. G. (1997) *Plant nematode control*. CAB international.

Willner, D., Furlan, M., Haynes, M., Schmieder, R., Angly, F. E., Silva, J.*, et al.* (2009a) Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals. *PLoS One,* **4**.

Willner, D., Thurber, R. V. and Rohwer, F. (2009b) Metagenomic signatures of 86 microbial and viral metagenomes. *Environ. Microbiol.,* **11,** 1752-1766.

Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., *et al.* (2014) SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics,* **30,** 1660-1666.

Xiong, T., Wang, S., Jiang, Q. and Huang, J. Z. (2011) A New Markov Model for Clustering Categorical Sequences. In: *Data Mining (ICDM), 2011 IEEE 11th International Conference on.* pp. 854-863.

Xiong, T., Wang, S., Jiang, Q. and Huang, J. Z. (2014) A novel variable-order Markov model for clustering categorical sequences. *Knowledge and Data Engineering, IEEE Transactions on,* **26,** 2339-2353.

Yang, Y. and Smith, S. (2013) Optimizing de novo assembly of short-read RNA-seq data for phylogenomics. *BMC Genomics,* **14,** 328.