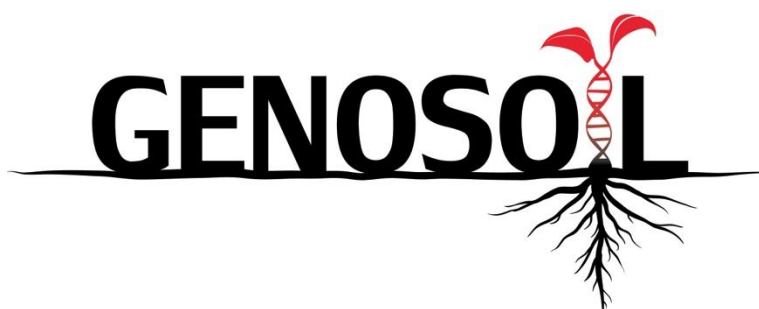




APPROCHE METAGENOMIQUE POUR L'ETUDE DE LA BIODIVERSITE TOTALE DU SOL

Thibaud Decaëns
Laboratoire d'Ecologie (EA 1293) – Université de Rouen
thibaud.Decaëns@univ-rouen.fr

Rapport final
05 mars 2014

Numéro de contrat Ministère/ADEME : 1060C0092

REMERCIEMENTS :

Les participants au projet tiennent à remercier ici le Lycée Agricole d'Yvetot, le Conservatoire des Sites de haute Normandie, l'ONF, ainsi que Thierry Lecomte (Conservateur de la RN des Courtils de Bouquelon) pour les autorisations d'échantillonner sur les sites d'étude. Nous remercions aussi Paul Hebert qui au travers du projet IBOL (International Barcode Of Life) a permis la réalisation des bibliothèques de référence. Enfin nous remercions Pierre Taberlet qui a accepté d'appliquer le protocole issu de l'ANR MetaBar à certains de nos échantillons.

L'ADEME EN BREF

L'Agence de l'Environnement et de la Maîtrise de l'Energie (ADEME) participe à la mise en œuvre des politiques publiques dans les domaines de l'environnement, de l'énergie et du développement durable. Afin de leur permettre de progresser dans leur démarche environnementale, l'agence met à disposition des entreprises, des collectivités locales, des pouvoirs publics et du grand public, ses capacités d'expertise et de conseil. Elle aide en outre au financement de projets, de la recherche à la mise en œuvre et ce, dans les domaines suivants : la gestion des déchets, la préservation des sols, l'efficacité énergétique et les énergies renouvelables, la qualité de l'air et la lutte contre le bruit.

L'ADEME est un établissement public sous la tutelle du ministère de l'écologie, du développement durable et de l'énergie et du ministère de l'enseignement supérieur et de la recherche. www.ademe.fr

Copyright

Toute représentation ou reproduction intégrale ou partielle faite sans le consentement de l'auteur ou de ses ayants droit ou ayants cause est illicite selon le Code de la propriété intellectuelle (art. L 122-4) et constitue une contrefaçon réprimée par le Code pénal. Seules sont autorisées (art. 122-5) les copies ou reproductions strictement réservées à l'usage privé de copiste et non destinées à une utilisation collective, ainsi que les analyses et courtes citations justifiées par la caractère critique, pédagogique ou d'information de l'œuvre à laquelle elles sont incorporées, sous réserve, toutefois, du respect des dispositions des articles L 122-10 à L 122-12 du même Code, relatives à la reproduction par reprographie.

Résumé

Le projet GENOSOIL a pour objectif principal de développer une méthode de quantification de la biodiversité du sol se basant sur une combinaison d'approches de taxonomie moléculaire mettant en œuvre le barcode ADN et la métagénomique (barcode environnemental à partir d'échantillons totaux d'invertébrés et/ou d'échantillons de sol). La démarche s'appuie sur (1) la constitution de bibliothèques de référence de barcodes ADN ainsi que (2) l'utilisation des outils de séquençage de nouvelle génération (Roche 454, Illumina MiSeq) pour caractériser les communautés d'invertébrés présentes dans des échantillons de sol ou issus d'extraction de faune. L'élaboration et la calibration du protocole ont été faites à l'aide d'échantillons d'invertébrés et de sol prélevés dans différents gradients environnementaux caractéristiques des écosystèmes du Nord-Ouest de la France.

Les bibliothèques de référence de barcodes ADN comportent au total, 2338 invertébrés qui ont été analysés et conservés dans des collections de référence, représentant quelques 500 espèces ou unités taxonomiques opérationnelles moléculaires (MOTUs) réparties dans 21 groupes taxonomiques. La couverture taxonomique, bien qu'inégale en fonction des groupes taxonomiques, est très bonne pour les groupes clefs tels que les collemboles, les myriapodes ou les araignées, pour lesquels une proportion importante des espèces connues de la région Haute Normandie a d'ores et déjà été intégrée dans les bases de données ; cette proportion est de 100 % pour le groupe des vers de terre.

La campagne de prélèvement visant à fournir les échantillons nécessaires au volet métagénomique a été réalisée selon deux gradients environnementaux échantillonnés pour un total de 54 points de prélèvement. Trois points provenant de trois types d'habitat échantillonnés ont été traités avec le protocole mis au point à base de séquençage massif parallèle sur une plateforme Roche 454. Il a été possible de caractériser ces trois types d'habitats par la composition de leurs communautés d'invertébrés du sol au niveau spécifique, générique familial et ordinal.

De plus une autre méthode de recouvrement de la diversité des sols à niveau taxonomique inférieur (entre Embranchement et Ordre) a été expérimentée avec la recherche de traces d'ADN extracellulaires dans 24 échantillons de sols secs. L'application du protocole issu de l'ANR MetaBar a permis de recouvrer une partie significative de la diversité des sols. Les rapports de dominance dans les proportions relatives de séquences produites indiquent bien une correspondance avec les proportions attendues entre grands groupes trophiques.

Mots clés : barcode ADN, biodiversités des sols, invertébrés, metabarcoding, ADN extracellulaire, communautés

Sommaire :

Remerciements :	2
1. Biodiversité des sols et approche moléculaire –Présentation générale du projet.....	6
1.1 Objectif	6
1.2 Sites d'étude.....	6
1.3 Analyse paysagère.....	6
1.4 Echantillonnage	8
1.4.1 Echantillons d'invertébrés pour les bibliothèques de barcodes ADN.....	8
1.4.2 Echantillons de sol et d'invertébrés pour l'approche environnementale.....	8
2. Résultats des différentes analyses moléculaires.....	9
2.1 Bibliothèque de référence générale.....	9
2.1.1 Analyses moléculaires.....	9
2.1.2 Résultats.....	9
2.1.3 Discussion.....	11
2.2 Le cas des vers de terre – une bibliothèque de référence complète pour la Haute Normandie....	12
2.2.1 Sites d'étude.....	13
2.2.2 Collecte des vers de terre	14
2.2.3 Analyses de données.....	14
2.2.4 Résultats.....	14
2.2.5 Discussion.....	16
2.2.6 Conclusion	18
2.3 Séquençage environnemental de l'ADN extracellulaire – Protocole MetaBar	19
2.3.1 Site d'étude	19
2.3.2 Echantillonnage sol	19
2.3.3 Analyse moléculaire	19
2.3.4 Identification des séquences	20
2.3.5 Résultats/ discussion.....	20
2.3.6 Conclusion	24
2.4 Séquençage environnemental de l'ADN intracellulaire	24
2.4.1 Sites d'étude.....	24
2.4.2 Analyse moléculaires.....	24
2.4.3 Analyse de données.....	25
2.4.4 Résultats/discussion.....	25
2.4.5 Conclusion	30
3. Conclusions générales / perspectives	30
4. Appui potentiel aux politiques publiques.....	31
5. Productions scientifiques	31
5.1 Bases de données et collections de référence.....	31
5.2 Publications réalisées ou en cours de préparation dans des journaux internationaux	31
5.3 Communications.....	32
5.4 Rapports de stage.....	33
6. Bibliographie.....	33

INFORMATIONS GENERALES CONCERNANT LE PROJET

Date d'engagement : 17 DECEMBRE 2010

Durée : 36 mois

Financement : 645 344.20 €

Montant MEDDTL/ADEME: 150 384.76 €

Co-financement(s):

Région Haute Normandie	55 410.00 €
Autofinancement Univ Rouen	244 852.63 €
Autofinancement CNRS	74 185.22 €
Autofinancement INRA	68 463.32 €
Autofinancement Univ Guelph	52 048.30 €

• *Participants au projet :*

Decaëns T.	ECODIV, Univ de Rouen
Chauvat M.	ECODIV, Univ de Rouen
Akpa-Vinceslas M.	ECODIV, Univ de Rouen
Aubert M.	ECODIV, Univ de Rouen
Forey E.	ECODIV, Univ de Rouen
Langlois E.	ECODIV, Univ de Rouen
Huste A.	ECODIV, Univ de Rouen
Porco D.	ECODIV, Univ de Rouen
Rougerie R.	ECODIV, Univ de Rouen
Bodilis J.	M2C, Univ de Rouen CNRS
Sebag D.	M2C, Univ de Rouen CNRS
Dubs F.	BIOEMCO, IRD
Dupont L.	BIOEMCO, Uni Paris XII
Lavelle P.	BIOEMCO, IRD
Roy V.	BIOEMCO, Univ Paris XII
Rossi J.P.	CBGP, IRNA – Montpellier
Deharveng L.	OSEB, MNHN
Judson M.	OSEB, MNHN
Oszwald J.	COSTEL, Univ Rennes 2
Hajibabaei M.	Univ de Guelph
Shokralla S.	Univ de Guelph
Smith M.A.	Univ de Guelph
Dolédec S.	BEL-LEHF, Univ Lyon
Hedde M.	PESSAC, INRA

1. BIODIVERSITE DES SOLS ET APPROCHE MOLECULAIRE –PRESENTATION GENERALE DU PROJET

1.1 Objectif

Le projet GENOSOIL a pour objectif principal de développer une méthode de caractérisation de la biodiversité du sol en s'appuyant sur une combinaison d'approches de taxonomie moléculaire basées sur le barcode ADN et le métabarcoding (barcode environnemental à partir d'échantillons totaux d'invertébrés et/ou d'échantillons de sol).

La démarche s'appuie sur (1) la constitution de bibliothèques de référence de barcode ADN dans le cadre des campagnes internationales de barcode (projet iBOL- <http://ibol.org/>) pour une sélection de groupes d'invertébrés clefs ; (2) l'élaboration de protocoles d'analyse de l'ADN environnemental du sol à l'aide des outils de séquençage de nouvelle génération (Roche 454 et Illumina MiSeq). L'élaboration du protocole a été effectuée à l'aide d'échantillons d'invertébrés et de sol prélevés dans différents gradients environnementaux caractéristiques des écosystèmes du Nord-Ouest de la France.

1.2 Sites d'étude

Conformément aux recommandations du conseil scientifique, le choix des sites a porté sur des zones ayant déjà fait l'objet d'études dans le cadre de projets financés par l'ADEME : St Adrien et Eawy. Ces deux sites ont été antérieurement étudiés dans le cadre de deux projets financés respectivement lors des appels d'offres « Bioindicateurs » (coord. P. Lavelle) et « GESSOL 3 » (coord. F. Dubs). Pour la première phase de l'échantillonnage de GENOSOIL, les efforts ont été concentrés sur un gradient successional composé de 3 stades de végétation calcicole (pelouse ouverte, pelouse en cours d'enfrichement et forêt), ainsi que sur un cycle d'exploitation forestière constitué de 3 stades de régénération forestière sur sol limoneux acide (hêtraie de 20, 65 et 130 ans). Ce choix permet ainsi d'envisager une valorisation des résultats génomiques au travers d'une mise en relation entre la biodiversité édaphique, les principaux types de sol et la dynamique de la végétation.

En complément de ce plan d'échantillonnage et des méthodes prévues, une expérimentation de recouvrement de l'ADN extracellulaire (ADN issu de la décomposition des organismes) à partir d'échantillons secs de sol a été effectuée sur du matériel prélevé dans une prairie pâturée du lycée agricole d'Yvetot lors du projet ANR EDISP (coord. J. Mathieu).

1.3 Analyse paysagère

La première action a été d'établir la cartographie des gradients paysagers. Afin d'effectuer cette tâche, il a été indispensable de cartographier les types d'occupation des sols végétaux. Pour ce faire, un corpus de données de photographies aériennes et d'images satellites a été rassemblé pour les sites d'Eawy (conservatoire des sites naturels de Haute-Normandie) et de St Adrien (images provenant de la banque de données IGN et de SPOT 5). Ce corpus de données cartographiques ainsi constitué a permis de suivre la dynamique de l'occupation des sols végétaux entre 1947 et 2009. Pour établir ce suivi, des cartes de synthèse des dynamiques végétales ont été mises en place pour 1947, 1955, 1963, 1973, 1994, 2000, 2003 et 2009. Ces cartes tiennent compte, par exemple, des états de surface typiques des zones de prairies, de forêts, de zones humides ou d'espaces végétaux anthropisés comme les vergers (Figure 1).

Les cartes produites ont servi de support à la définition de l'emplacement spatial des points d'échantillonnage et à l'élaboration des analyses des gradients paysagers sur lesquels ils se positionnent. Pour la campagne de 2011, nous avons ainsi identifié et décrit le gradient de succession sur substrat calcaire (site de St Adrien), ainsi que le gradient de maturation d'un peuplement forestier (site d'Eawy). D'autres gradients potentiellement utilisables pour la prochaine campagne d'échantillonnage ont également été identifiés : un gradient d'intensification agricole entre 1949 et 2009, un gradient de gestion forestière (diversification du peuplement et longueur de la rotation

forestière), et un gradient de fragmentation d'habitat avec une étude du lien entre chaque patch et ses voisins et la dynamique de leur relation dans le temps.

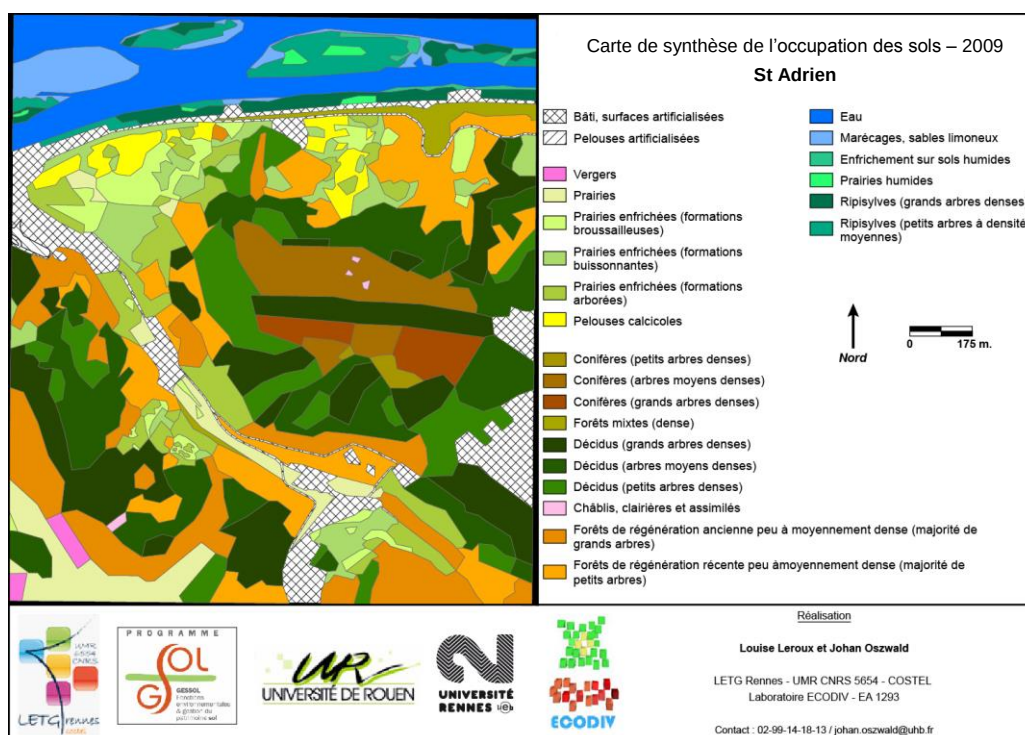


Fig. 1 Carte de synthèse des dynamiques de l'occupation des sols en 2009 sur le site de St Adrien.

Pour chaque stade de gradient, trois parcelles indépendante (réplicats) ont été identifiées, et 3 points d'échantillonnage ont été positionnés dans chacune d'entre-elles (pseudoréplicats, cf Fig 2, plan d'échantillonnage).

Description formelle des sites d'étude :

- St Adrien est situé dans une zone de points chauds de biodiversité à l'échelle régionale. Elle se caractérise par une dynamique de déprise agricole après abandon de l'élevage ovin et des cultures (vigne en particulier) auquel ils étaient consacrés jusqu'aux années 50. Les sols sont des rendzines carbonatées. La végétation est une mosaïque de stades successionnels comprenant des pelouses sèches, des stades pré-forestiers et forestiers. Le site est caractérisé par trois ensembles typologiques majeurs d'occupation des sols : les occupations du sol d'origine agricoles, forestières ou de zones humides. Les espaces agricoles sont particulièrement marqués par les systèmes prairiaux à différents stades de recolonisation par les formations broussailleuses, buissonnantes ou arborées. Il est caractérisé par la présence de pelouses calcicoles et plus exceptionnellement de vergers. Les espaces forestiers sont majoritairement constitués de végétation arborée décidue à différents stades d'âge. Au centre de la forêt de St Adrien, on observe la présence de vastes parcelles de conifères constituées de petits arbres et d'arbres de taille moyenne. De nombreuses parcelles de forêt de régénération ancienne ou récente, peu à moyennement dense sont observables sur le site d'anciennes prairies sur les coteaux. Enfin, les zones humides sont nettement concentrées autour du réseau fluvial.

- Yvetot et Eawy sont deux sites localisés sur les sols limoneux des plateaux de la région rouennaise. Le premier site est localisée dans une zone mixte comprenant des prairies permanentes et temporaires (élevage laitier et de viande), des cultures annuelles céréalières conventionnelles ou sans labour. Le site d'Eawy est principalement constitué de forêts décidues denses de grands arbres ou d'anciennes coupes arborées datant de 1947. Certaines parcelles forestières sont marquées par des éclaircissements progressifs des massifs, mais cela reste cantonné à quelques secteurs marginaux. Sur ce site géré par l'ONF, plusieurs types de gestion sont appliqués : futaie régulière et taillis sous futaie de hêtre, chêne et/ou épicéa.

1.4 Echantillonnage

1.4.1 Echantillons d'invertébrés pour les bibliothèques de barcodes ADN

Les bibliothèques de référence ont été assemblées grâce à un échantillonnage massif combinant deux approches :

1 – Un échantillonnage *quantitatif*, consistant à extraire les invertébrés à partir de volumes ou de surfaces de sol connus : par berlèse sur des carottes de 6cm de diamètre et 10cm de profondeur pour la mésofaune ; par tri manuel de blocs de sol de 25cm de côté et de 10cm de profondeur pour la macrofaune ; par application de 10 litres de formol à 0.4 % sur une surface de 1 m² pour les vers de terre.

2 – Un échantillonnage *qualitatif* consistant à prospecter systématiquement les microhabitats présents dans la zone d'étude. Les récoltes ont été faites à vue sur le terrain mais aussi par extraction de la faune au moyen d'appareils de Berlèse. D'autres techniques (fauchage dans la végétation, pièges malaises) ont également été utilisées pour échantillonner les adultes de certains groupes d'insectes dont les larves se retrouvent fréquemment dans les sols (diptères et coléoptères notamment).

Ces approches sont complémentaires et permettent d'obtenir la couverture nécessaire à l'établissement d'un référentiel exhaustif de la zone d'étude. Ce protocole a été appliqué sur 54 points dont 27 en milieux prairial et 27 en milieux forestier. Au total 2338 spécimens appartenant à 21 groupes taxonomiques ont ainsi été échantillonnés et barcodés (Tableau 1).

1.4.2 Echantillons de sol et d'invertébrés pour l'approche environnementale

Le plan d'échantillonnage suivi a été le même que celui proposé dans le projet (Figure 2) : pour chaque type d'écosystème (prairie ou forêt), trois répétitions indépendantes ont été choisies, et trois pseudoréplicats distants de 20m les uns des autres ont été positionnés. Sur chacun de ces points, une série d'échantillons a été prélevée de façon à permettre la collecte des vers de terre, de la macrofaune et de la mésofaune. Une carotte de sol supplémentaire a été prélevée en prévision de l'analyse environnementale sur la matrice de sol.

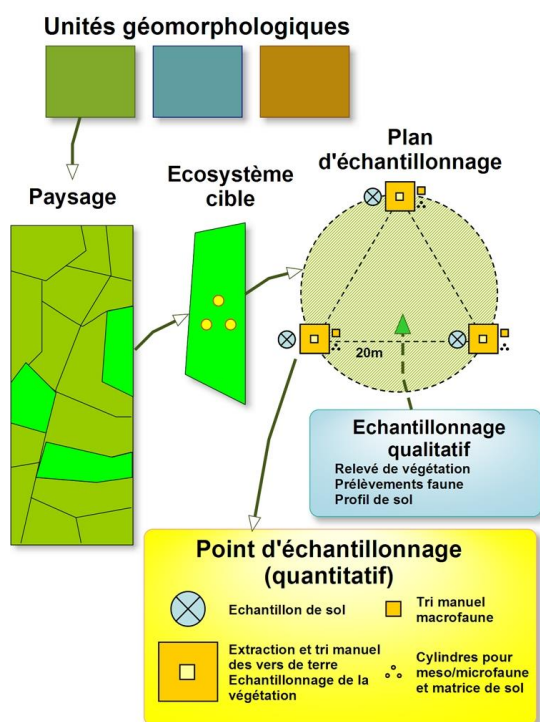


Fig. 2 Plan d'échantillonnage des organismes du sol, de la végétation et du sol. La description géomorphologique et paysagère a été utilisée pour localiser les écosystèmes cibles (répliqués trois fois chacun). Les organismes du sol, la végétation et le sol sont échantillonnés sur trois points

d'échantillonnage dans chaque répétition (échantillonnage quantitatif), et sur la surface environnante (échantillonnage qualitatif).

L'échantillonnage de la phase 1 du projet a porté sur (1) un gradient successional sur substrat calcaire comprenant trois stades (pelouse ouverte, pelouse en cours d'enfrichement et forêt de 70 ans) et (2) un cycle de régénération forestière comprenant des hêtraies d'âge croissant sur sol limoneux de plateau (parcelles de 20, 65 et 130 ans). La méthodologie appliquée est la même que celle suivie pour la phase préliminaire (Figure 2). Au total, 24 parcelles et 72 points d'échantillonnages ont ainsi été prélevés.

2. RESULTATS DES DIFFERENTES ANALYSES MOLECULAIRES

2.1 Bibliothèque de référence générale

Ce chapitre présente les résultats concernant le séquençage individuel de spécimens qui ont été identifiés morphologiquement. Ceci a permis la constitution de bibliothèques de références qui établissent un lien entre la dénomination nomenclaturale taxonomique et la séquence du fragment de COI désigné comme barcode universel pour le règne animal. Ce lien permettra d'établir dans les parties suivantes du projet une correspondance entre la quantité massive de séquences produite par les méthodes de métagénomique et le référentiel taxonomique auquel se rattache tout le corpus de connaissances biologiques accumulé jusqu'à nos jours.

2.1.1 Analyses moléculaires

Pour chaque spécimen d'une taille supérieure à 6mm (macrofaune) l'ADN a été extrait à partir d'un fragment de muscle ou de patte équivalent à 1mm² de tissus. Pour les animaux dont la taille était inférieure à 6mm, la totalité du spécimen a été utilisée pour une extraction d'ADN non-destructive, qui a ensuite permis la récupération de leur enveloppe tégumentaire et donc leur examen morphologique (Porco *et al.*, 2010). L'extraction d'ADN a été effectuée dans 40ul de tampon de lyse additionné de protéinase K à 56°C pendant toute une nuit. Les phases suivantes sont celles de l'extraction automatisée d'ADN avec utilisation de plaques filtres de 96 puits publiées par (Ivanova *et al.*, 2006). La région barcode standard (5' du gène mitochondrial COI) a été amplifiée en utilisant les amorces universelles LCO1490 - HCO2198 (Folmer *et al.*, 1994). Les extraits n'ayant pas pu être amplifiés avec cette première paire d'amorces, l'ont été avec deux autres paires incluant des amorces internes (LepF1-MLepR1 - MLepF1-LepR1) (Hajibabaei *et al.*, 2006). Le protocole standard du CCDB (Canadian Center for DNA barcoding) (http://www.ccdb.ca/docs/CCDB_Amplification.pdf) a été utilisé pour l'amplification, et les produits de cette amplification ont été testés sur E-gel 96 Agarose 2 % (Invitrogen). Les amplicons non-purifiés ont été séquencés dans les deux directions en utilisant les amorces additionnées d'une queue M13 (M. Hajibabaei *et al.*, 2005). Les produits de séquençage ont ensuite été purifiés en utilisant le protocole Agencourt CleanSEQ et analysés sur séquenceur ABI 3730 (Applied Biosystems). Les séquences ont été assemblées et éditées avec Sequencher 4.5 (GeneCode Corporation) et alignées avec BIOEDIT version 7.0.5.3 (Hall, 1999). Les séquences sont publiquement disponibles sur BOLD (Barcode Of Life Data system) dans le jeu de données 'Bibliothèques de référence GENOSOIL -Invertébrés du sol' [DS-BGENO].

2.1.2 Résultats

L'obtention de séquences de COI (région barcode) pour un grand nombre d'espèces identifiées a permis de constituer des bibliothèques de référence pour les principaux groupes cibles retrouvés dans les sols. Certains groupes épigées ont été également échantillonnés (Lépidoptères, Diptères, Coléoptères) car leur représentation dans les sols est souvent relative aux stades larvaires,

qui sont par ailleurs difficiles voire impossibles à identifier morphologiquement. Incidemment cette étude nous a donc permis de caractériser les habitus larvaires de certaines espèces connues principalement par la morphologie de l'adulte.

Groupe taxonomique	Effectif	# MOTUs
Diplura	1	1
Lepidoptera	4	1
Orthoptera	5	3
Thysanoptera	5	3
Psocoptera	6	1
Pseudoscorpiones	9	3
Homoptera	16	5
Enchytraeidae	18	12
Opiliones	19	4
Gastropoda	26	10
Hymenoptera	43	8
Hemiptera	44	29
Isopoda	68	4
Plectidae	68	1
Myriapoda	92	35
Coleoptera	153	24
Diptera	178	49
Araneae	221	100
Acarina	225	85
Collembola	401	80
Lumbricidae	736	36

Table 1 Liste des principaux groupes taxonomiques échantillonnés, effectifs des spécimens barcodés pour alimenter les bibliothèques de référence, et nombre d'unités taxonomiques opérationnelles moléculaires (MOTUs) pour chacun d'entre eux.

Les 21 groupes ont été échantillonnés à la fois dans les milieux prairiaux et forestiers pour un total de 2338 spécimens représentant environ 500 espèces ou unités taxonomiques opérationnelles moléculaires (MOTUs) définis par un seuil de variabilité génétique variable d'un groupe taxonomique à l'autre (Table 1). L'ensemble des séquences de référence de COI obtenues a été placé dans des projets spécifiques sur la plateforme bioinformatique BOLD (jeu de données 'Bibliothèques de référence GENOSOIL -Invertébrés du sol' [DS-BGENO], Figure 3). Ces bibliothèques qui constituent la première étape pour une documentation complète de la biodiversité du sol par le barcode ADN, sont en accès libre.

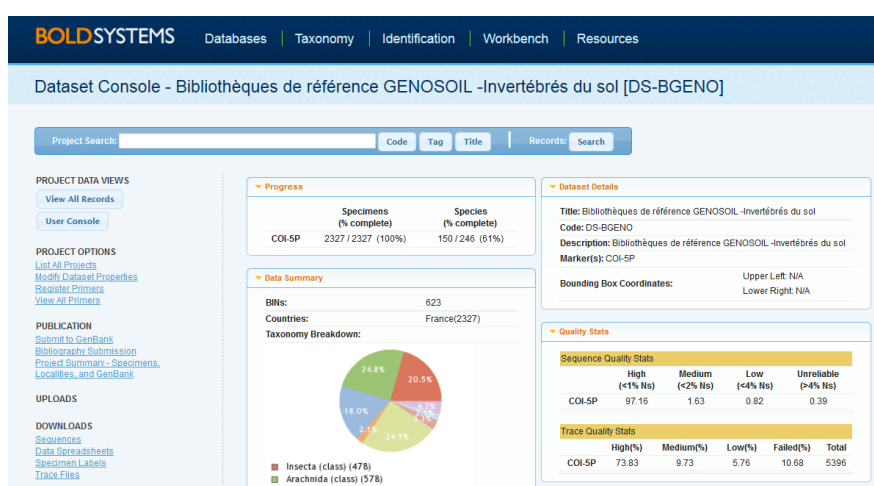


Fig. 3 Copie d'écran du projet-container [ESSNO] sur la plateforme BOLD. Ce projet contient les séquences de COI obtenues pour les invertébrés autres que les vers de terre, groupe qui dispose de son propre container.

2.1.3 Discussion

La Figure 4 représente la façon dont les séquences sont traitées dans un premier temps (arbre des distances calculées à l'aide de l'indice K2P) pour identifier des unités taxonomiques opérationnelles moléculaires (MOTUs) ou encore clusters de barcodes, qui peuvent le cas échéant correspondre à des espèces morphologiquement bien identifiées, ou à des espèces cryptiques indissociables par leurs habitus.

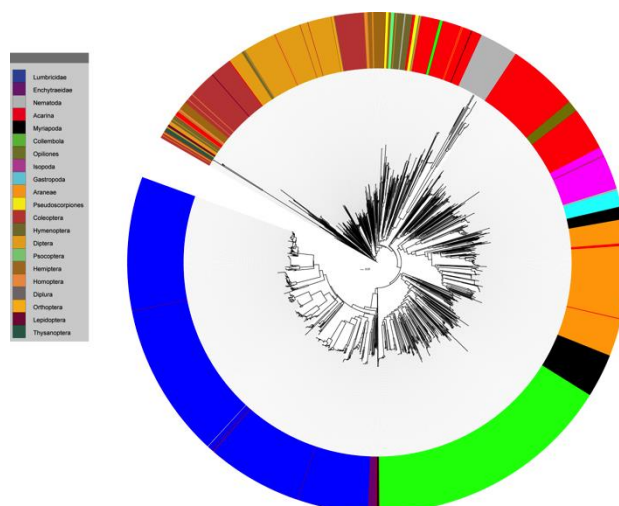


Fig. 4 Arbre des distances au plus proche voisin (distances du K2P) réalisé sur l'ensemble des séquences obtenues pour les bibliothèques de référence. Les couleurs visualisent la position des principaux groupes taxonomiques sur l'arbre.

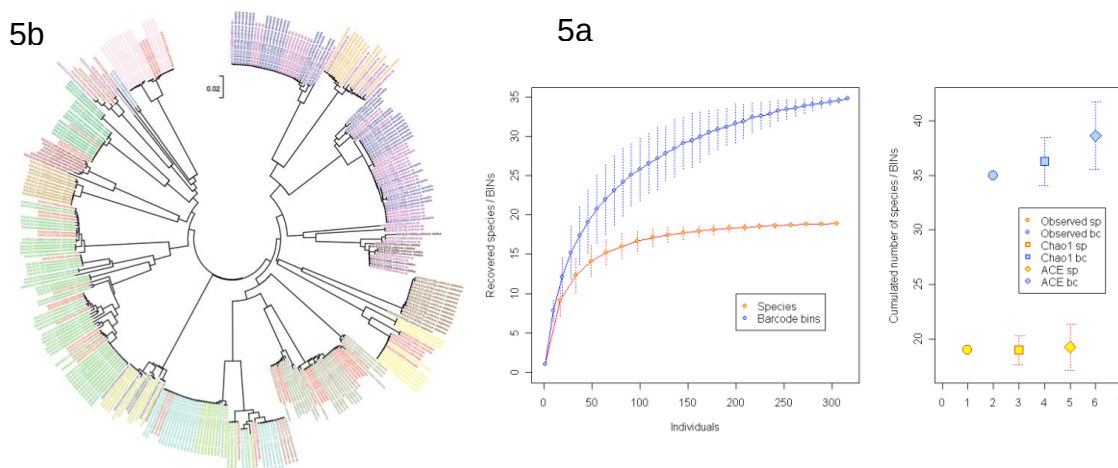


Fig. 5 (a) Arbre des distances au plus proche voisin (distances du K2P) réalisé sur les vers de terre, et (b) estimations des richesses régionales à partir d'identifications morphologiques versus moléculaires (Decaëns et al., 2013). Chao1 et ACE sont deux estimateurs de richesse basés sur la proportion de singletons dans les échantillons. BC = barcode clusters, équivalent à MOTU dans le texte.

Le nombre de MOTUs observé est ensuite utilisé comme un proxy permettant de réaliser des estimations de richesse spécifique pour les différents habitats échantillonnés. L'avantage de cette approche par rapport aux estimations classiques basées sur des identifications morphologiques est qu'elle permet l'identification, et donc la prise en compte, des juvéniles (Richard *et al.*, 2010) mais aussi des spécimens fragmentés au moment de l'échantillonnage (les juvéniles pouvant représenter la majorité de l'échantillon dans certaines situations). Elle permet également de différencier les espèces cryptiques pour lesquelles les critères morphologiques sont inapplicables (Figure 5).

[illegible][illegible]

Il faut noter que ces espèces cryptiques sont susceptibles de présenter des caractéristiques et écologiques différentes, ce qui peut impacter directement les disciplines zoologie et l'écologie qui utilisent certaines de ces espèces comme modèles et bioindicateurs. Un exemple frappant est donné par le vers de terre de référence *Lumbricus terrestris* qui a été révélé être un ensemble de deux espèces cryptiques (*L. terrestris* et *L. lineus*) ayant des spécificités écologiques sensiblement différentes (James *et al.*, 2010) ainsi que la collembole *Parisotoma notabilis* (Porco *et al.*, 2012) qui est l'une des espèces les plus communes dans les relevés de sols Européens. L'étude systématique effectuée au cours de ce stage vise à mettre en lumière d'autres cas de ce type dans divers groupes de la faune du sol afin de leur donner une valeur de référence et de valorisation [cf. paragraphe suivant sur les vers de terre].

2.2 Le cas des vers de terre – une bibliothèque de référence complète pour la Haute Normandie

12

(Eisenhauer *et al.*, 2009), ainsi que leur impact au niveau des communautés végétales (Forey *et al.*, 2011).

Les Lumbricidae Européens sont sans conteste les vers de terre qui ont été les plus étudiées depuis les prémices de la systématique linnéennes, en outre, ils ont fait partie des premiers organismes du sol décrits par Linné en 1758. Si la plupart de ces espèces ont été décrites depuis longtemps, et utilisées en tant qu'hypothèses scientifiques fiables, certaines d'entre elles se sont avérées être des complexes d'espèces (*A. chlorotica* (King *et al.*, 2008); *L. terrestris* (James *et al.*, 2010), *A. caliginosa* (Perez-Losada *et al.*, 2005)). Ces cas de diversité cryptique ont été révélés par l'usage en routine des méthodes de taxonomie moléculaire. Il a parfois été suggéré que cette occurrence d'entités de niveau spécifique dans les délimitations morphologiques d'une espèce pourrait être reliée à des différences en termes de préférences écologiques (James *et al.*, 2010, Porco *et al.*, 2013).

Ces résultats suggèrent la présence de nombreuses d'« espèces chimériques », soit de noms d'espèces uniques qui représentent en fait des groupes entiers d'espèces distinctes (Porco *et al.*, 2012, Decaëns *et al.*, 2013). En conséquence, on peut suspecter que de nombreuses données de la littérature ont été faussement accumulées sous un même nom. Cette situation souligne le besoin d'établir de façon systématique des bibliothèques de référence moléculaires pour la détection précoce de tels cas, de façon à corriger les erreurs accumulées dans la littérature jusqu'à présent.

Nous présentons ici l'assemblage d'une telle bibliothèque de référence pour une région Européenne connue et échantillonnée depuis longtemps pour laquelle 21 espèces ont été recensées jusqu'à présent (Bouché, 1972, Decaëns *et al.*, 2008). Cette réalisation, outre l'intérêt scientifique général décrit ci-dessus, va permettre dans le cadre de l'emploi de méthodes de metabarcoding d'établir une sorte de calibrage de la méthode ainsi qu'une comparaison précise de résultats obtenus avec les différentes méthodes d'identifications qu'elles soient morphologiques ou moléculaires.

2.2.1 Sites d'étude

Neuf sites ont été échantillonnés, parmi lesquels on retrouve des habitats forestiers et prairiaux, dont ceux qui sont ciblés dans les gradients de l'étude Genosoil : (1) St Adrien et Henouville qui sont des coteaux calcaires typiques de la vallée de Seine, traditionnellement utilisées et entretenues dans le cadre du pâturage ovin, la forêt de coteau de St Adrien a aussi été échantillonnée ; (2) le Marais Vernier avec à la fois un habitat forestier et prairial de la plaine inondable de la Seine ; (3) la forêt de Eawy qui est la plus importante hêtraie de Haute Normandie (gérée par l'ONF) ; (4) le Campus de Mont St Aignan pour les habitats prairie et forêt ayant subis un certain niveau d'anthropisation; (5) Lycée agricole d'Yvetot, une pâture intensive typique ; (6) Radepont, habitat prairial anthropisé (Table 2).

Localité	Type d'habitat	N
Henouville	Coteau calcaire	47
St Adrien	Forêt de plateau	18
St Adrien	Coteau calcaire	91
Marais Vernier, réserve des Manneville	Forêt alluviale	10
Marais Vernier, réserve des Manneville	Prairie alluviale	41
Forêt d'Eawy	Forêt de plaine	123
Mont Saint Aignan, Campus universitaire	Pelouse	129
Mont Saint Aignan, Campus universitaire	Forêt	91
Lycée agricole Yvetot	Pâturage	89
Radepont	Prairie	2

Table 2 Effectif échantillonné pour les différentes localités et types d'habitat.

2.2.2 Collecte des vers de terre

Comme décrits dans la partie 1-4-1, l'échantillonnage des spécimens s'est fait aussi bien de façon quantitative (aspersion d'un mètre carré avec une solution de formol à 0.4 %, assorti du tri manuel d'un monolithe de 25cm de côté), que de façon qualitative (fouille de sol et de tronc ponctuelle). Tous les échantillons ont été préservés au congélateur en éthanol 100 % jusqu'au moment du prélèvement de tissu pour extraction d'ADN, ils ont ensuite été mis en collection à température ambiante. Les identifications morphologiques ont été effectuées avec la clef de Bouché (1972) avec une rectification nomenclaturale (Sims & Gerard, 1985). Au total 561 vers ont été sélectionnés pour être barcodés. Les spécimens sont conservés dans les collections du laboratoire ECODIV.

2.2.3 Analyses de données

2.2.3.1 Distances moléculaires et délimitation des MOTUs

Les analyses de distances ont été effectuées via le logiciel MEGA5 (Tamura *et al.*, 2011), en utilisant l'algorithme de Neighbor-Joining (Saitou & Nei, 1987) avec le modèle utilisant le 'paramètre Kimura-2' (Kimura, 1980) pour estimer les distances génétiques. La robustesse des nœuds a été évaluée par des analyses de bootstrap avec 1000 pseudoréplicats. L'arbre obtenu a été représenté avec l'utilitaire en ligne ITOL (Letunic & Bork, 2007). Les Unités Taxonomiques Moléculaires Opérationnelles (MOTUs) ont été définies avec le logiciel 'mothur' en utilisant la commande 'Hcluster' avec l'option 'Furthest neighbor' (Schloss *et al.*, 2009).

2.2.3.2 Courbes de raréfaction et estimateurs de diversité

Les courbes de raréfactions concernant la diversité spécifique ont été générées avec EcoSim 7.71 (Gotelli & Entsminger, 2006) avec un intervalle de confiance de 95 %, puis ont été tracées sous R 2.15.0 (R Development Core Team, 2012) en utilisant le paquet 'Plotrix' (Lemon, 2006). Elles ont été calculées séparément pour les MOTUs dans leur ensemble et pour les espèces identifiées morphologiquement, en fonction de l'effort d'échantillonnage. De plus, la richesse spécifique maximale théorique a été calculée pour chaque catégorie avec les estimateurs de diversité Chao1 et ACE via le paquet 'Vegan' (Oksanen *et al.*, 2012) dans R 2.15.0.

2.2.4 Résultats

Vingt-deux espèces ont pu être morphologiquement identifiées parmi les 561 spécimens analysés. La mise en relation du nombre de MOTUs définies dans le jeu de données avec différentes valeurs seuil de distances génétiques utilisées pour les délimiter, montre une forme caractéristique en plateau qui représente la stabilisation du nombre de MOTUs autour d'une valeur optimale qui correspondrait à un niveau de divergence génétique correspondant à l'espèce biologique (Plaisance *et al.*, 2011) (Figure 8). Ce plateau s'étend des valeurs seuils allant de 8 % à 14 %, délimitant respectivement 36 à 33 MOTUs. Le calcul des fréquences de distances k2p prises deux à deux ont permis de déterminer quelle partie du plateau correspond au 'barcode gap' (c'est à dire la séparation entre la variation intraspécifique et interspécifique). Cette zone est située entre 8 % et 9 % (Fig. 8). Nous avons sélectionné la première valeur de divergence trouvée après la borne supérieure du gap dans la distribution des fréquences : 9 %. Appliqué à la totalité du jeu de données, cette valeur seuil a permis de recouvrir 36 MOTUs incluant de nombreuses lignées cryptiques pour neuf des espèces morphologiques cibles : *D. octaedra*, *E. tetraedra*, *A. ictérica*, *A. chlorotica*, *A. rosea*, *A. caliginosa*, *L. rubellus*, *L. castaneus* et *O. tyrtaeum* (Figure 9).

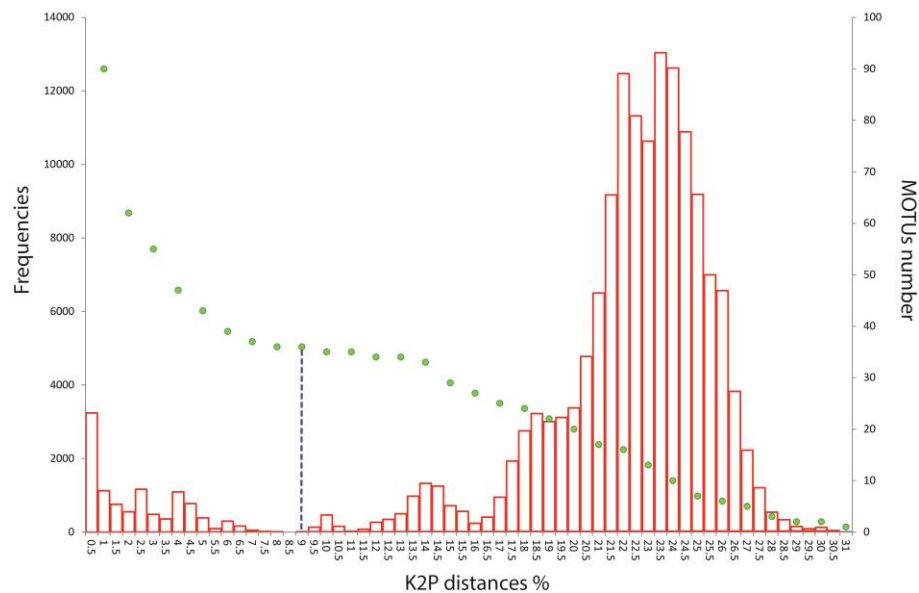


Fig. 8 Désignation de la valeur seuil (Trait pointillé bleu). En rouge l'histogramme des fréquences des distances génétiques deux à deux (Axe des ordonnées à gauche). En vert le nombre de MOTUS délimitées en fonction des différentes valeurs seuils (Axe des ordonnées à droite).

Treize de ces MOTUS correspondent bien à des espèces morphologiques. Pour celles-ci, la distance intraspécifique maximale varie entre 0 % et 4.6 %, et les valeurs minimales intraspécifiques entre 9.2 % et 19.7 % (Table x). Les 23 autres MOTUS sont des lignées discrètes qui apparaissent au sein d'espèces nominales pour lesquelles on observe donc des maximums de divergence intraspécifique élevés (de 9.2 % à 23.5 %, Table x). Les distances à l'intérieur de ces MOTUS varient entre 0 % et 7.65 % alors que les distances entre ces MOTUS vont de 7.7 % jusqu'à 23,5 %. La plus faible divergence observée entre MOTUS est celle qui sépare les deux MOTUS pour *E. tetraedra*.

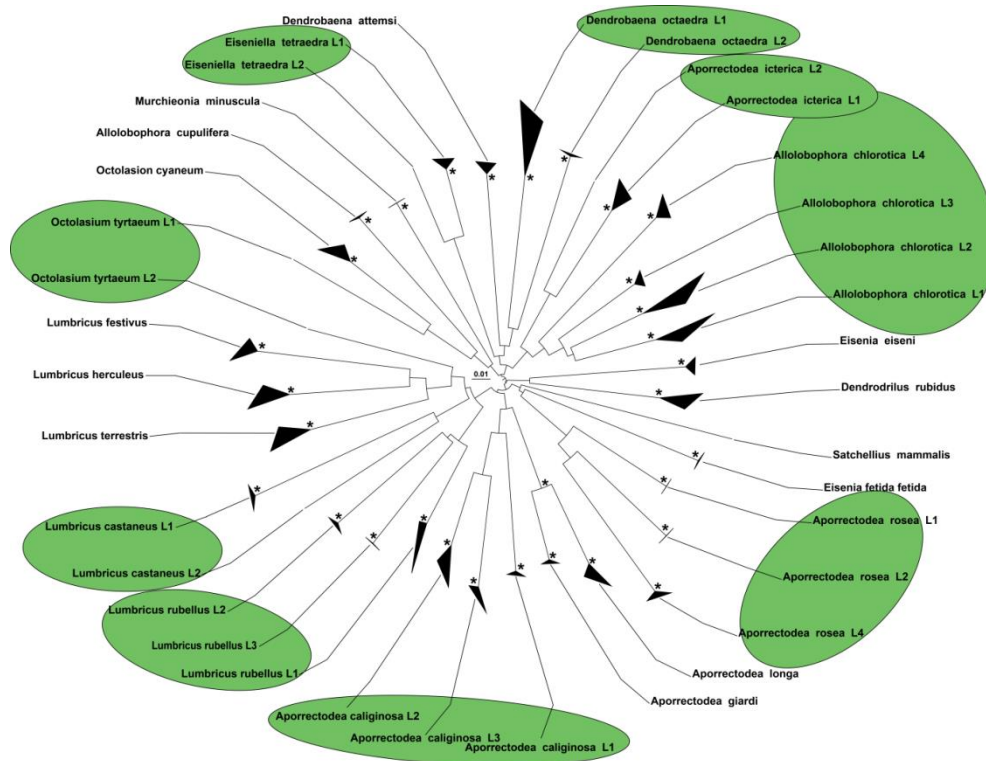


Fig. 9 Arbre de neighbor joining de distances génétiques k2p avec les 36 MOTUS détectées. Les nœuds soutenus par des valeurs de bootstrap ≥ 99 % sont marqués par un astérisque. Les MOTUS cryptiques au sein d'une même espèce morphologique sont entourées d'une ellipse verte.

Le gain permis par l'utilisation du barcode ADN concernant la diversité du groupe est de 38.8 %. Si on prend en compte les estimateurs de richesse (ACE et Chao1), ce gain atteint les 45 % avec une diversité prédite à 40 MOTUs (Figure 10).

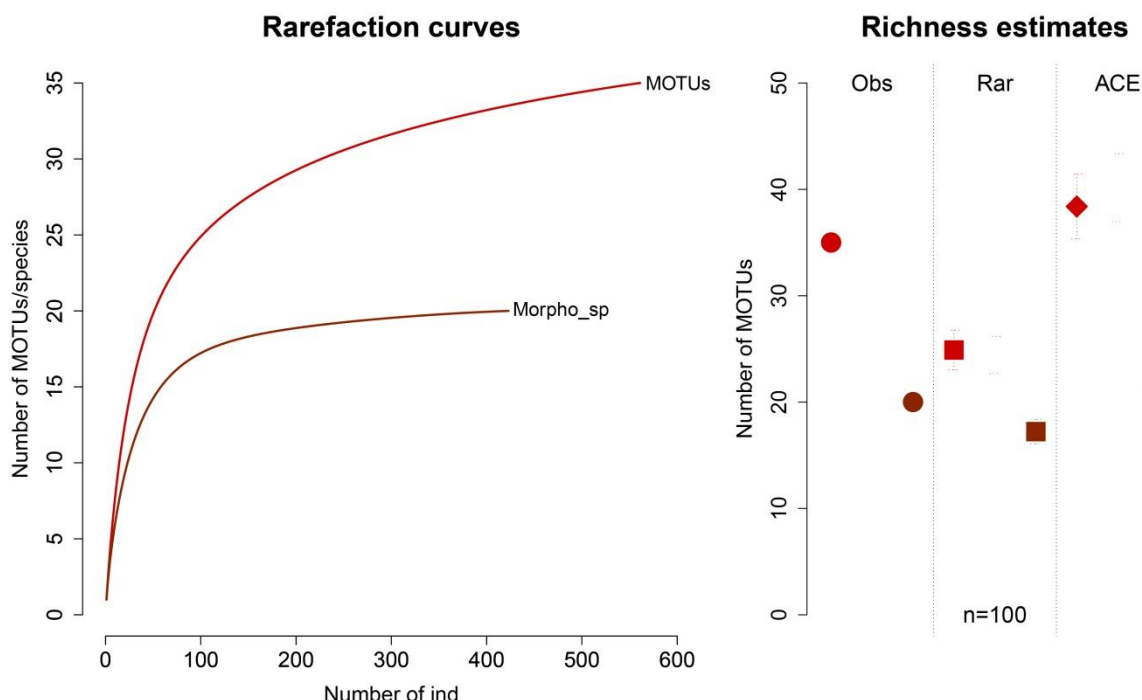


Fig. 10 Courbes d'accumulation et estimateurs de diversité pour les espèces morphologiques (Morpho_sp) et les unités taxonomiques opérationnelles moléculaires (MOTUs).

2.2.5 Discussion

2.2.5.1 Barcode ADN – outil d'identification et apport au niveau biodécouverte

La constitution de cette bibliothèque de référence a montré que l'utilisation du barcode ADN permet non seulement l'identification précise d'espèces pour les vers de terre, mais aussi d'accéder à un niveau d'identification et de connaissance de la diversité supérieur à celui atteint par des moyens d'identification morphologique classique (qui aurait laissé de côté 38.8 % de la diversité du groupe). Les distances génétiques mesurées entre et au sein des MOTUs cryptiques détectées pour les espèces morphologiques cibles sont comparables à celles que l'on trouve en intra et interspécifique pour des espèces morphologiques bien définies. Ceci suggère fortement que ces MOTUs représentent en fait des entités de niveau spécifique qui pourront être ultérieurement décrites. Certains de ces cas sont déjà documentés (*A. chlorotica* - King et al. 2008; *L. terrestris/L. herculeus* - James et al. 2010, *A. caliginosa* - Peres-Losada et al. 2005, *A. rosea*, *L. rubellus* - Porco et al. 2013), mais d'autres ont été nouvellement détectés par cette étude : *A. icterica*, *D. octaedra*, *E. tetraedra*, *L. castaneus* et *O. tyrtaeum*. Sur les 22 espèces échantillonnées en Haute Normandie, 9 présentent des cas de diversité cryptique. Une extension géographique et/ou une augmentation de l'effort d'échantillonnage pourrait permettre de recouvrer plus d'éléments de cette diversité jusque-là ignorée au sein d'espèces européennes communes. L'apport en termes de biodécouverte est substantiel pour une zone d'étude très connue, en climat tempéré et pour un groupe extrêmement visible et étudié. Des découvertes d'une telle amplitude sont d'habitude réservées au milieu tropical.

Si on utilise le ratio de 38.8 % trouvé dans cette étude entre la diversité attendue d'après les espèces décrites d'une part, et la diversité inférée à partir du barcode ADN d'autre part, afin d'établir la diversité du groupe à d'autres échelles géographiques, l'augmentation peut être spectaculaire. La faune mondiale des vers de terre passerait de 5282 à 7331 espèces. Mais on est très certainement en dessous d'une estimation réaliste, car de nombreuses régions du globe sont encore très insuffisamment étudiées à l'instar des tropiques. L'estimation serait plus juste pour l'Europe, qui est

une zone ayant globalement subi le même effort d'échantillonnage, on passerait pour celle-ci de 432 espèces de Lumbricidae à 599. Pour la France, dont l'estimation serait encore plus fiable, on passerait de 127 à 176 espèces sur un territoire que les taxonomistes classiques ont déjà échantillonné avec soin (cf. échantillonnage Bouché 1972 qui a couvert toute la France avec un maillage assez régulier).

Un certain nombre des espèces touchées par cette diversité cryptique avaient été décrites comme 'polytypiques' (comportant un certain nombre de sous-espèces et de variétés) par Bouché en 1972. Nous avons donc examiné les spécimens issus des différentes MOTUs de façon à pouvoir éventuellement les rattacher à certaines de ces variétés ou sous-espèces. Aucun des caractères morphologiques examinés ne nous ont permis un tel rattachement à la nomenclature infraspécifique de Bouché (1972), confirmant bien le caractère cryptique de cette variation. En revanche, cet examen a permis d'observer des tendances assez marquées au niveau de caractères morphométriques comme la taille, le poids ou le nombre de segments, comme c'est le cas entre les deux espèces *L. terrestris* et *L. herculeus* anciennement regroupées sous le nom *L. terrestris* (James et al. 2010).

Même si ces tendances comportent des recouvrements importants, elles appuient le degré de différenciation génétique détecté. C'est le cas au niveau du poids pour les deux MOTUs *L. castaneus* L1 et *L. castaneus* L2 ainsi que les trois MOTUs appartenant à *A. calliginosa*. La MOTU *A. calliginosa* L1 se détache aussi très nettement des 2 autres lignées en ce qui concerne le nombre de segments ainsi que la longueur. C'est également le cas de la MOTU *A. chlorotica* L4 dont les spécimens sont en moyenne plus courts (Figure 11).

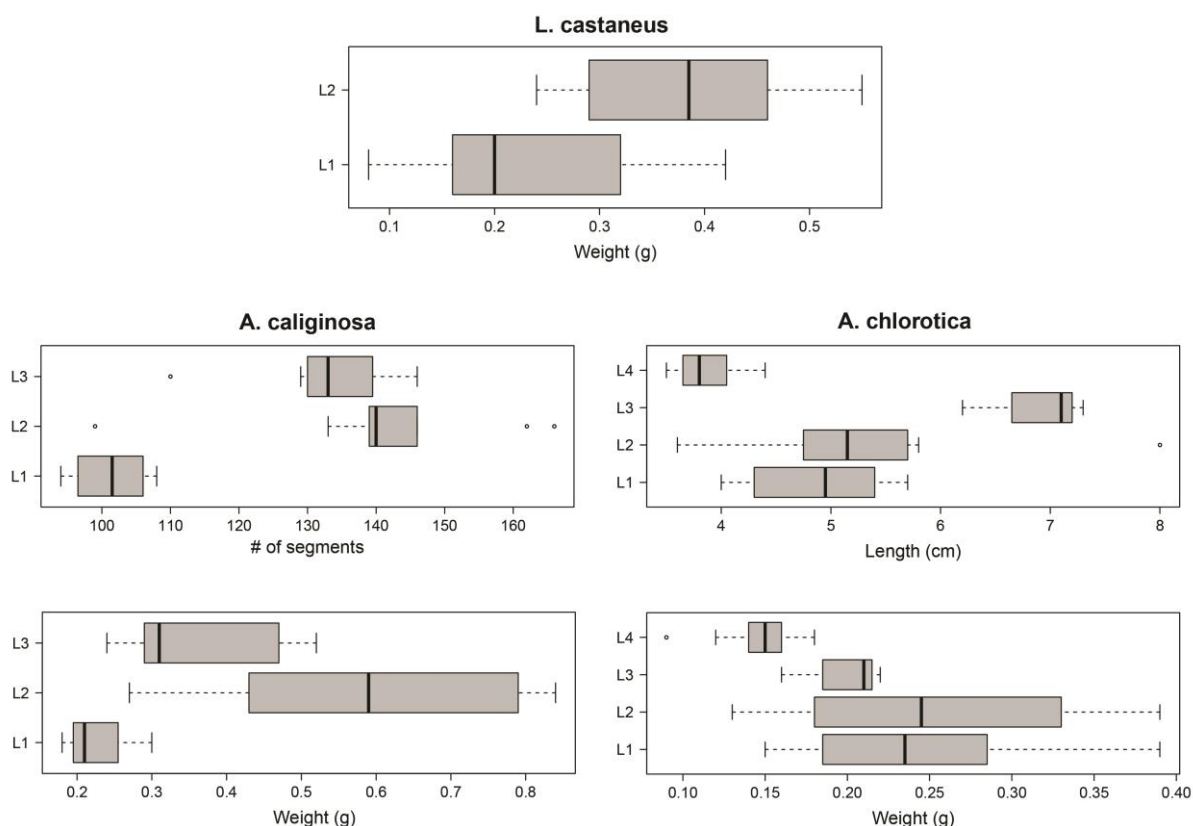


Fig. 11 Mesures morphométriques concernant le poids, la taille et le nombre de segments. Sont représentées ici les mesures qui permettent de définir des tendances au sein des groupes de MOTUs.

Ces différences morphométriques, à confirmer sur un échantillonnage plus large, pourraient suggérer un type de spéciation par polyploïdisation, un phénomène déjà avéré pour plusieurs des espèces chez lesquelles des cas de diversité cryptique ont été détectés (Vlasenko et al., 2011) (Vsevolodova-Perel & Bulatova, 2008) et qui pourrait entraîner l'apparition de classes morphométriques discrètes de cet ordre.

D'autre part, ce travail a été l'occasion de régler le problème de l'espèce *Aporrectodea tuberculata* réputée proche de *A. caliginosa* : celle-ci s'est avérée n'être qu'un morphotype présent dans les populations d'une des MOTUs de *A. caliginosa* (*A. caliginosa* L2) au sein de laquelle elle est mêlée avec le morphotype *A. caliginosa* typique *paratypica* (Fig x).

2.2.5.2 Différences écologiques potentielles

Cette précision accrue apportée au niveau de l'identification, va modifier significativement notre vision des communautés d'espèces présentes dans les différents habitats, mais va aussi possiblement impacter notre connaissance du fonctionnement de ces mêmes communautés d'invertébrés dans la matrice des sols.

Ces MOTUs cryptiques de niveau spécifique, présentent comme on l'a vu un haut niveau d'individualisation génétique et parfois même d'importantes variations phénotypiques, et se doivent donc d'être annotées de façon systématique dans la littérature. Cela permettra d'accumuler des données sur les propriétés écologiques respectives de ces MOTUs qui resteraient sinon agglomérées sous les noms d'« espèces chimères » utilisés jusqu'à présent (Porco et al. 2012, Decaëns et al. 2013), et pour lesquelles les résultats des nombreux travaux déjà publiés sont donc sujet à caution (*L. terrestris* / *L. herculeus* : 1153 ; *A. chlorotica* : 37 ; *A. rosea* : 31 ; *A. caliginosa* : 125 ; *A. icterica* : 8 ; *D. octaedra* : 70 ; *E. tetraedra* : 31 ; *L. castaneus* : 11 ; *L. rubellus* : 203 ; *O. tyrtaeum* : 10 – Source : Web of knowledge Novembre 2013). De manière plus active, une telle annotation systématique de ces lignées cryptiques dans les différentes disciplines utilisant ces espèces permettra à des expérimentations dont ce sera spécifiquement le but de caractériser les différences biologiques et écologiques de ces entités.

Au niveau des préférences d'habitat, on a déjà détecté des tendances bien marquées chez *L. terrestris*/*L. herculeus* avec *L. herculeus* que l'on retrouve plus en milieu boisé alors que *L. terrestris* est majoritairement présent en prairie (James et al. 2010). Mais les préférences d'habitats ne sont pas les seuls facteurs indirects à prendre en compte pour mettre en lumière les différences écologiques potentielles de ces MOTUs cryptiques. Dans bon nombre de localités, on a retrouvé plusieurs de ces lignées cryptiques en sympatrie. Cela renforce l'hypothèse du niveau spécifique de ces entités et suggère un évitement compétitif au niveau de l'exploitation de la ressource, même si leur similarité morphologique va à rebours des hypothèses émises jusque-là (Decaëns et al., 2011). Cette potentielle spécificité trophique met aussi en exergue le besoin urgent de caractérisation et de prise en compte de ces MOTUs cryptiques pour la compréhension du fonctionnement des sols.

2.2.5.3 Compilation et homogénéisation avec les jeux de données existants

L'agglomération de cette bibliothèque avec les données issues de la littérature permet déjà d'établir une correspondance nomenclaturale entre les études passées et ce travail, et ainsi de mettre en évidence d'éventuelles erreurs d'identification. C'est le cas par exemple avec la séquence JN850542.1 (numéro d'accension Genbank) tirée du travail par Peres-Losada et al. en 2012 et identifiée comme *Aporrectodea longa*, mais qui tombe dans la MOTU correspondant à *A. giardi* dans notre étude. Ce conflit est certainement dû soit à une erreur d'identification ou à une contamination croisée pour ce spécimen. Cet outil permettra donc incidemment la rectification de jeux de données moléculaires accumulés jusque-là.

2.2.6 Conclusion

La constitution de bibliothèques de référence exhaustives, qui est un prérequis nécessaire à l'application de méthode de métabarcoding, donne un bon exemple de ce que peut concrètement apporter cette première phase à elle seule. Elle permet évidemment l'identification des spécimens quel que soit leur stade de développement (juvéniles ou cocons) ou leur état, mais elle permet également de mettre au jour une portion non négligeable de la diversité d'un groupe qu'on pensait jusque-là extrêmement bien connue pour la zone d'étude. En plus du bon en avant effectué au niveau de la connaissance de la biodiversité pour des groupes pourtant réputés connus, cet outil, utilisé en routine, en préalable à toute expérimentation dans différentes disciplines, permettra de recommencer la tâche

d'accumulation des données biologiques et écologiques sur chacune entité cryptique qui constituent le groupe d'espèces. On retrouvera alors l'utilité initiale du référencement transversal que constituent les noms d'espèces dans la littérature scientifique, et incidemment une meilleure compréhension du compartiment sol des écosystèmes.

2.3 Séquençage environnemental de l'ADN extracellulaire – Protocole MetaBar

Ce protocole permet de séquencer des fragments d'ADN dit extracellulaires : il s'agit d'ADN fragmenté qui n'est pas extrait des tissus des organismes, mais qui est celui directement disséminé dans l'environnement après la mort et la décomposition de ceux-ci. Ces ADN extracellulaires sont plus fragmentés que ceux que l'on peut trouver dans une cellule encore intacte, c'est pourquoi la taille du fragment cible amplifié est peu importante (100pb). Ainsi, ce protocole permet de détecter la présence d'un organisme sur un point d'échantillonnage même si aucun de ces représentants vivant n'est présent dans l'aliquot de sol prélevé.

2.3.1 Site d'étude

Le site ciblé est une des prairies pâturée de 5 ans d'âge du Lycée agricole d'Yvetot incluse en tant que site d'étude dans le projet EDISP (ANR-08-JCJC-0023). La mise en place a consisté en un labour profond du sol qui a ensuite été additionné, pour les couches superficielles, de 50 tonnes à l'hectare de fumier de vache. La parcelle a ensuite été ensemencée d'ivraie, de trèfle blanc, de fétuque élevée et de fléole des prés. Pâturée sur 6 mois par deux à cinq vaches laitières à l'hectare, elle reçoit une fertilisation annuelle de 180 kg d'azote par hectare.

2.3.2 Echantillonnage sol

Vingt-quatre carottes de sols ont été prélevées sur la parcelle selon un maillage régulier de 20m de côté environ (Figure 12). Ces carottes de sol ont été tamisées, homogénéisées et séchées, puis conservées 3 ans à sec avant d'être aliquotées pour l'analyse. L'échantillonnage représente un aliquot de 5g de sol pour chacune des 24 carottes.



Fig. 12 Réseau des 24 points d'échantillonnage pris sur une des prairies du lycée agricole d'Yvetot.

2.3.3 Analyse moléculaire

Les détails du protocole ne sont pas disponibles car ils font partie d'un autre projet encore non publié (MetaBar – Taberlet et al. en préparation). Nous pouvons néanmoins en décrire le principe dans les grandes lignes : l'extraction d'ADN total s'effectue à partir d'aliquots de sol de 1g sur les 5g déjà échantillonnés, suit une amplification d'un fragment cible de 100pb du gène nucléaire 18S et un séquençage sur une plateforme Illumina MiSeq.

2.3.4 Identification des séquences

Après filtrage des séquences trop courtes et chimères, les 100 haplotypes les plus abondants ont été comparés aux séquences de 18S correspondantes à la portion amplifiée par le protocole MetaBar présentes sur Genbank. L'algorithme utilisé est Megablast (Altschul *et al.*, 1990 ; BLAST+ v2.2.26), optimisé pour reconnaître les séquences hautement similaires. Seuls les niveaux de similarité compris entre 99 et 100 % ont été pris en compte pour l'identification. Cela a permis d'établir une correspondance au niveau des grands embranchements, classes et ordres.

2.3.5 Résultats/ discussion

La variation du gène nucléaire 18S est moins importante que celle observée pour les gènes mitochondriaux utilisés en routine pour le barcode. De ce fait, la portion amplifiée de 100pb ne permet pas de capturer une variation correspondant au niveau spécifique. On obtient donc une série d'haplotypes pour ce fragment de gène nucléaire. Chacun de ces haplotypes est susceptible d'appartenir à un large panel d'espèces. Ces données permettent néanmoins dans un premier temps d'obtenir d'importantes informations sur la présence, et à un certain niveau, sur l'abondance des grands groupes présents dans les sols (Figure 13, Table 3).

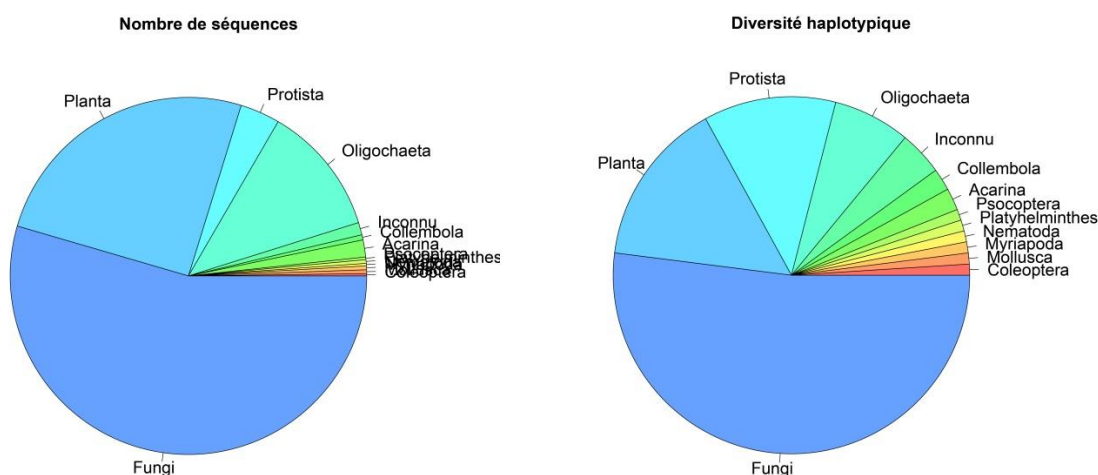


Fig. 13 Proportion des différents groupes taxonomiques détectés en nombre de séquences et d'haplotypes produits.

Groupes taxonomiques	Nb haplotypes	Nb séquences
Coleoptera	1	1324
Mollusca	1	2241
Myriapoda	1	2247
Nematoda	1	1421
Platyhelminthes	1	2126
Psocoptera	1	1542
Acari	2	9927
Collembola	2	3003
Inconnu	4	7232
Oligochaeta	7	75781
Protista	12	23233
Planta	15	162585
Fungi	52	350643

Table 3 Nombre de séquences et d'haplotypes obtenus par l'analyse pour les différents groupes taxonomiques dont les correspondances ont été établies avec les séquences disponibles sur Genbank.

Les haplotypes les plus amplifiés sur les 24 échantillons appartiennent aux trois groupes attendus comme dominants en termes de biomasse dans les sols hors bactéries : les champignons, les plantes et les oligochètes avec respectivement 350 643, 162 585 et 75 781 séquences produites. Les champignons et les plantes (représentées par la masse racinaire) sont majoritaires, aussi bien en nombre de séquences produites qu'en nombre d'haplotypes, qui à ce niveau-là traduit une représentation phylogénétique des groupes assez large si on les considère comme autant de groupes d'espèces. Viennent ensuite les protistes et les oligochètes, avec une inversion attendue entre nombre d'haplotypes et le nombre de séquences produites qui provient du différentiel de taille et de biomasse entre les deux types d'organismes. Puis on retrouve les organismes de la mésofaune et de la microfaune qui ont des densités et des biomasses moindres dans les sols comme les Collembolles, les acariens et les nématodes. On observe un nombre d'haplotypes similaire pour les acariens et les collembolles, mais un nombre de séquences plus important pour les acariens (9927) que pour les collembolles (3003), ce qui correspond mieux aux diversités/densité de populations relatives connues pour les deux groupes. La faible représentation des prédateurs de la mésofaune et macrofaune (Myriapodes Chilopodes et Coléoptères Staphylinidae) est aussi un résultat attendu au niveau de l'écologie des réseaux trophiques. L'application d'une technique employant un gène mitochondrial permettrait de recouvrer une image plus précise de la diversité de ces groupes en atteignant le niveau spécifique.

Il est intéressant de noter dans tous les échantillons, mais à des intensités différentes en nombre de séquences, la présence d'un platyhelminthe qui correspond à 99 % à une séquence de Turbellariés. Ce groupe n'est habituellement présent que dans les sols tropicaux, mais plusieurs espèces invasives de planaires prédatrices ont été détectées dans de nombreuses régions du globe : *Platydemus manokwari* prédateur d'escargots suspecté d'avoir mis en danger la faune endémique de nombreuses îles (Iwai *et al.*, 2010), *Bipalium adventitium* prédateur sur les vers *Lumbricus rubellus* et *Lumbricus terrestris* (Fiore *et al.*, 2004), et de nombreux autres cas revus dans Fiore *et al.* 2004. Dernièrement, une attention particulière a été portée à un planaire invasif provenant de Nouvelle Zélande, *Arthurdendyus triangulatus*, qui avait pourtant été détecté en Europe il y a une cinquantaine d'années (Royaume unis, Irlande et îles Faeroe). Ce planaire s'est révélé être un prédateur des vers de terre, en particulier de l'anécique *Lumbricus terrestris*, dont il peut significativement réduire la biomasse (20 %) (Murchie & Gordon, 2013). La comparaison avec une séquence de 18S déposée sur Genbank pour cette espèce a permis d'écarter qu'il s'agissait d'elle, Néanmoins, lors de la comparaison de séquences, la similarité à 99 % avec un haplotype de turbellariés pouvant recouvrir plusieurs espèces, signale la très possible présence d'au moins une espèce de planaire exotique en Haute Normandie qui pourrait potentiellement exercer une pression de prédation sur des espèces de vers ou de gastéropodes natifs. Ce résultat donnera lieu à des recherches ciblées qui permettront de confirmer ce signalement.

De plus, même si il existe des biais d'amplification pour les techniques PCR appliquées sur les mélanges de plusieurs ADN qui pourraient venir fausser les proportions relatives exprimées pour un point (Medinger *et al.*, 2010), ces résultats permettent d'avoir une information sur la composition relative des communautés d'invertébrés du sol pour les principaux groupes dominants en chacun des points d'échantillonnage (Figure 14). En faveur des résultats obtenus, il est à noter que certaines proportions en nombre de séquences correspondent aux quelques ratio connus entre biomasse moyenne de certains groupes. C'est le cas entre Protistes et champignons, le ratio rapporté avec une dominance de ces derniers atteint un facteur 14 (Gobat *et al.*, 2010); on retrouve au niveau du nombre de séquences produites ici ce ratio de même orientation et d'ordre de grandeur similaire avec un facteur 15.

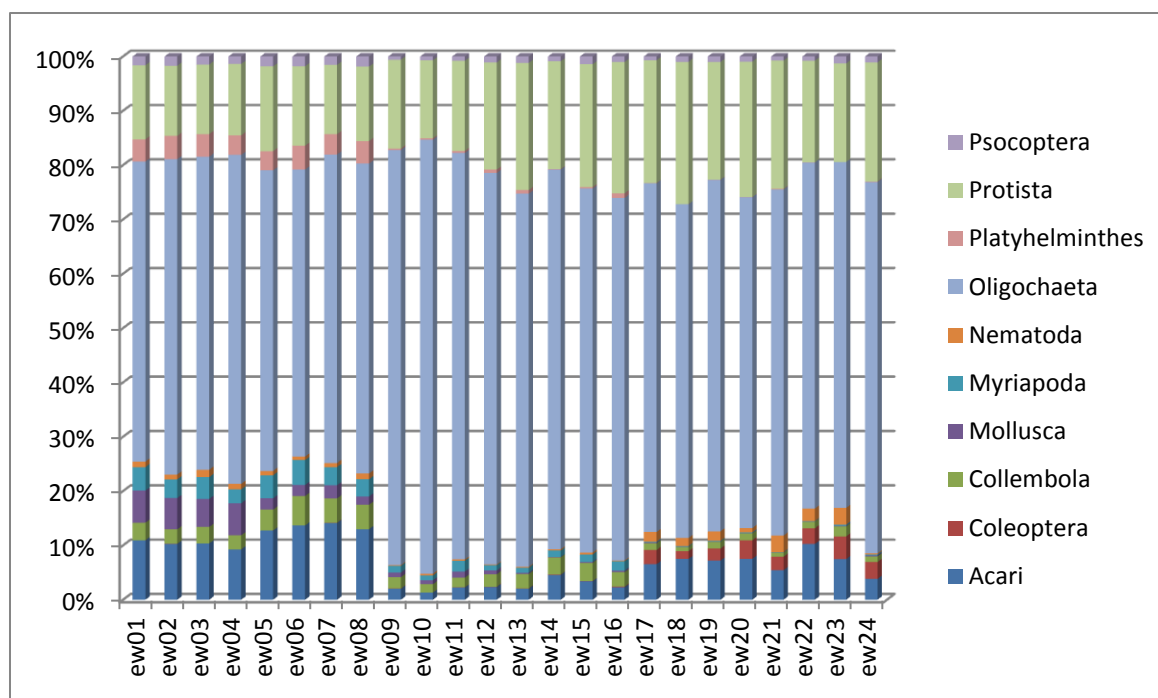


Fig. 14 Composition relative des communautés pour les 24 points d'échantillonnage.

Ces données permettent la visualisation des proportions de séquences entre les différents points d'échantillonnage de la parcelle, et ainsi de cartographier la présence des groupes cibles (Figure 15). Ceci permet, dans une certaine mesure, d'inférer leur niveau de densité de population, avec encore une fois une réserve quant aux éventuels biais méthodologiques induits par l'amplification d'un mélange de plusieurs ADN. On observe, pour les données obtenues, une certaine hétérogénéité spatiale selon les groupes, ce qui suggère, pour certains taxa, leurs probables relations trophiques. On a par exemple, une correspondance entre la distribution des Platyhelminthes et des mollusques qui pourrait être relative, comme évoqué dans le paragraphe précédent, à la prédation de planaires invasifs sur la faune locale de gastéropodes. Il faudrait bien sûr pour étayer plus avant cette hypothèse un échantillonnage plus large ainsi qu'un niveau d'indentification plus élevé, mais le potentiel de l'outil, applicable à des sols secs, est fortement soutenu par ces premiers résultats. Dans le même ordre d'idée, on remarquera la distribution extrêmement polarisée de l'haplotype correspondant au groupe de Coléoptères de la famille prédatrice des Staphylinidae qui pourrait être possiblement liée à la dynamique des populations de proies. On pourrait penser à des dispositifs expérimentaux qui mettraient en évidence la dynamique spatio-temporelle de structuration des communautés d'invertébrés du sol à une échelle courte, saisonnière, ou bien longue, pluriannuelle, permettant de mesurer, entre autres, l'impact de changements climatiques ou d'usage des sols sur l'occupation spatiale par différents groupes.

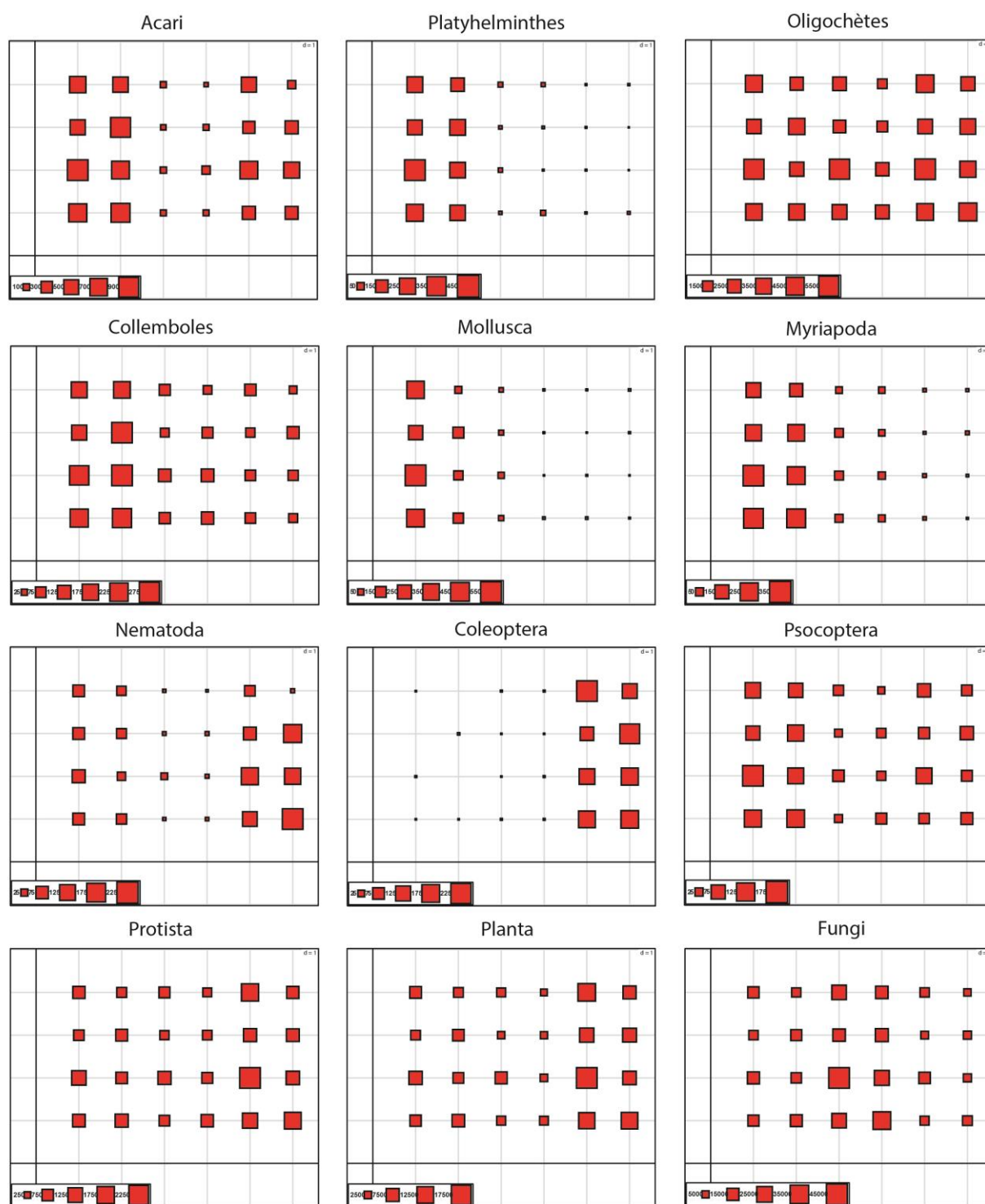


Fig. 15 Distribution spatiale du nombre de séquences selon la grille d'échantillonnage des sols pour les groupes taxonomiques détectés.

2.3.6 Conclusion

Les résultats obtenus avec ce dispositif expérimental montrent que l'on peut potentiellement utiliser des sols secs stockés depuis plusieurs années pour recouvrer des traces d'ADN, permettant elles-mêmes de remonter à la composition des communautés. Ces informations permettent de spatialiser la diversité des invertébrés, et on peut facilement imaginer un croisement de telles données avec des mesures de propriétés physico-chimiques des sols échantillonnés. La confrontation systématique des données de communautés avec de nombreux éléments (biotiques et abiotiques) à une échelle étendue, permettrait d'émettre de solides hypothèses quant aux facteurs clefs qui déterminent leur structuration. Ce type d'approche est particulièrement prometteur si on se place dans une optique d'utilisation des échantillons prélevés dans le cadre du Réseau de Mesures de la Qualité des Sols.

Ici, l'essai a été effectué avec un fragment de gène nucléaire dont la variabilité ne permet qu'une discrimination au niveau taxonomique de l'embranchement à l'ordre, mais le succès de cette démarche permet d'envisager l'application de techniques impliquant des gènes mitochondriaux tel que COI (ou 16S – gène utilisé dans le protocole MetaBar en ce qui concerne les vers – Bienert et al. 2012) pour lesquels nous avons de nombreuses séquences de référence, ou en tout cas les extraits d'ADN qui permettraient de les constituer rapidement pour un autre marqueur. L'utilisation de marqueurs mitochondriaux rendra possible l'établissement de listes d'espèces, ou du moins l'obtention, avec le dénombrement des MOTUS, d'un proxy de la diversité spécifiques dans différents groupe cibles dont les bibliothèques de référence n'auront pas été constituées. Le gain pour l'étude de l'écologie des communautés d'invertébrés du sol sera alors substantiel.

2.4 Séquençage environnemental de l'ADN intracellulaire

Cette technique met en œuvre l'amplification de fragments du barcode standard COI de taille importante (250 et 320 pb). Avec une telle taille de marqueur, on se positionne ici non plus sur les fragments d'ADN à l'état libre (extracellulaire) comme dans le protocole précédent, mais sur de l'ADN d'animaux pris vivants ou entrés récemment en stade de décomposition dans l'échantillon prélevé. L'un des éléments importants à explorer pour ce type d'approche est donc de trouver un volume d'échantillonnage suffisant pour une capture représentative de la diversité. L'avantage de cette technique est qu'elle repose sur la comparaison avec des bibliothèques de référence déjà constituées pour les sites d'étude, et permet donc d'accéder à un niveau d'identification spécifique pour une caractérisation précise des communautés d'invertébrés du sol.

2.4.1 Sites d'étude

Un point a été échantillonné par habitat en forêt d'Eawy, ainsi qu'en prairie et en forêt de coteau calcaire de St Adrien (description des sites cf. paragraphe 1.3) selon le protocole décrit plus haut avec trois pseudoréplicats distant de 20m sur chaque point (cf. paragraphe 1.4.2). Chaque pseudoréplicat est constitué d'une carotte de sol de 5cm de diamètre sur 10cm de longueur.

2.4.2 Analyse moléculaires

Un sous-échantillon de chaque carotte de sol homogénéisée a été traité avec un kit Power Soil DNA isolation (Mobio, USA). L'ADN total de la suspension homogène ainsi obtenue a été extrait avec un kit Nucleospin tissue (Macherey-Nagel Inc.) et élué dans 70ul d'eau ultrapure. Deux fragments de la région barcode de COI ont été amplifiés avec des réactions PCR en parallèle en utilisant deux paires d'amorces ((a) LCO1490_F (Folmer et al 1994) et 250_R (Shokralla et al. 2014 in prep.) ; (b) Univ_B_F and Univ_E_R (Mehrdad Hajibabaei *et al.*, 2012)). Un second cycle PCR a ensuite été réalisé avec les mêmes amorces additionnées d'amorces de fusion hybrides 454 ainsi que des séquences-étiquettes permettant l'identification individuelle d'échantillons. Chaque réaction PCR comprenait 2ul d'ADN, 17,5ul d'eau ultrapure, 2,5ul de tampon 10x, 1ul de MgCl₂ (50 mM), 0,5 ul de

dNTP (10 mM), 0,5ul de chaque amorce forward et reverse (10 mM) et 0,5 ul de polymérase Taq Platinum Invitrogen (5 U/ul) dans un volume total de 25 ul. Programme de température PCR : 95°C - 5 min.; 15 cycles de 94°C - 40 sec., 46°C - 1 min., et 72°C - 30 sec.; 72°C - 5 min. Les amplicons ont été purifiés sur colonnes de purification MiniElute PCR Qiagen avec un volume d'élution de 50 ul d'eau ultrapure. Les amplicons purifiés de la première PCR ont été utilisés comme matrice de la seconde PCR de 30 cycle d'amplification. Les PCR ont été effectuées avec un thermocycleur Eppendorf Mastercycler ep gradient S. Chaque stade expérimental comprenait des contrôles négatifs.

Tous les amplicons ont été purifiés et quantifiés fluorométriquement. Une quantité équimolaire d'amplicons tagués ont été séquencés en bidirectionnel avec un séquenceur 454 Genome Sequencer FLX System (Roche Diagnostics GmbH) avec le kit GS Titanium.

2.4.3 Analyse de données

Toutes les séquences ont été coupées en 3', et une limite minimale de score Phred de 20 a été appliquée (taille de fenêtre=10, pas d'incrément=5) avec PRINSEQ (Schmieder & Edwards 2011). Les séquences-étiquettes ont été utilisées pour séparer les séquences issues des différents jeux d'amorces. Après que les séquences-étiquettes et celle des amorces ait été enlevées, les répliques de séquences ont été supprimées avec PRINSEQ. Le bruit de fond a été supprimé en agglomérant les séquences à 99 % de similarité avec USEARCH (Edgar, 2010). Les séquences chimères ont été éliminées par une recherche de novo avec UCHIME (Edgar *et al.*, 2011). Les séquences de moins de 200pb ainsi que les singletons (séquences uniques) ont été exclues.

Les séquences ont été comparées avec la base de données nucléotides de Genbank (version Novembre 2013) et la bibliothèque de référence GENOSOIL grâce à l'algorithme Megablast (Altschul *et al.* 1990 ; BLAST+ v2.2.26). Les 100 premières réponses avec des valeurs de 'e' supérieures à 1e-20 ont été gardées. L'assignation taxonomique a été générée en analysant ces 100 réponses avec MEGAN 4.70.4 (Huson *et al.*, 2011). Les paramètres de l'ancêtre commun le plus basal (LCA) ont été ajustés pour chaque échantillon et les assignements à l'embranchement, l'ordre, la famille, le genre et l'espèce ont été assignés pour chaque séquence. La richesse spécifique et l'indice de Shannon ont été calculés avec EstimateS 9.1.0 (Colwell R.K., 2013). Les analyses multivariées ont été conduites sous PRIMER v6 (Clarke & Gorley, 2006).

2.4.4 Résultats/discussion

2.4.4.1 Caractérisation des communautés d'invertébrés du sol – Optimisation du protocole

La double comparaison avec la bibliothèque de référence GENOSOIL et les séquences présentes pour COI sur Genbank ont permis de détecter une partie de la diversité présente dans les carottes de sol prélevées pour les principaux groupes cibles du projet au niveau spécifique (Table 4). On observe une dominance marquée du nombre de séquences produites pour les Enchytraeidae dans les pseudoréplicats du point réalisé dans la forêt d'Eawy. Dominance retrouvée dans l'un des pseudoréplicats de la forêt et de la prairie des deux points échantillonnés sur le coteau calcaire de St Adrien pour ces deux habitats. En revanche, peu de séquences de Lumbricidae ont été recouvrées sur les différents échantillons : on obtient un signal seulement sur deux des pseudoréplicats de la forêt de coteau calcaire avec un seul cas de dominance en abondance de séquence. Ce faible résultat tient probablement à la fois (1) au fait que l'on amplifie des fragments d'ADN de taille importante, donc issus de tissu relativement frais et (2) à la méthode d'échantillonnage, soit une carotte de sol de 5cm de diamètre sur 10cm de profondeur avec laquelle la probabilité d'échantillonner des animaux de taille importante est faible. On peut appliquer cette remarque également à de nombreux groupes de la macrofaune dont la densité est faible mais qui vont être ponctuellement surreprésentés en terme d'abondance de séquences du fait de leur taille et donc de la quantité relative de tissus qu'ils vont représenter dans le mélange.

Si on ne considère seulement les membres de la mésofaune, principalement enchytraeides, collembolés et acariens, pour lesquels la méthode d'échantillonnage est plus adaptée, on a une dominance globale des enchytraeides dans la majorité des échantillons. Pour les deux groupes d'arthropodes principaux dans les sols, on constate une dominance nette de l'abondance des

séquences de collemboles sur les acariens dans tous les échantillons, ce qui est l'inverse de ce que l'on connaît des communautés de la mésofaune dans ces habitats. Cette réponse discordante s'explique lorsqu'on analyse les résultats obtenus pour des catégories taxonomiques inférieures avec des listes élargies au niveau familles et ordres. Cet élargissement procède de la récupération des unités taxonomiques opérationnelles moléculaires (MOTUs) qui ont une identification en deçà de l'espèce (Table 5 et 6). On observe alors un renversement de la tendance vue pour les espèces de collemboles et d'acarien. Ce renversement est d'autant plus net que le niveau taxonomique est bas avec une dominance marquée des acariens rétablie dans la majorité des échantillons, ce qui est plus conforme à nos connaissances. Une telle discordance s'explique par l'incomplétude des bibliothèques de référence pour le niveau espèce, en particulier pour les acariens dont la majorité n'est pas identifiée à l'espèce, que ce soit sur la bibliothèque GENOSOIL ou Genbank. Les acariens étant l'un des groupes les plus diversifiés des invertébrés de la faune du sol avec environ 45 000 espèces au niveau mondial (contre 8000 pour les Collemboles à titre de comparaison), ils demanderaient un effort d'échantillonnage beaucoup plus important au niveau des bibliothèques de référence pour avoir une réponse spécifique plus fiable lors de l'analyse des données de séquençage massif parallèle. De ce fait, de nombreuses séquences produites, n'étant pas identifiables pour le moment, n'apparaissent pas dans les bilans, l'importance de cette catégorie 'inconnue' décroissant évidemment avec le niveau taxonomique auquel on analyse les données. De façon générale, un effort supplémentaire à ce qui a été possible de réaliser dans le cadre de ce projet, permettrait, en ciblant spécifiquement quelques groupes importants tels que les acariens, les collemboles, les enchytraeides et les diptères, de réaliser des bilans de communautés plus précis, et cela simplement en réanalysant les résultats déjà obtenus.

Ces trois groupes, hors enchytraeides, ressortent de l'étude lorsque l'on analyse l'abondance des séquences, avec, dans tous les échantillons, une dominance des collemboles, acariens et diptères (présents sous formes larvaires). Ce sont les groupes pour lesquels le type d'échantillonnage choisi ici convient le mieux.

					EA20			SA70			SABO		
					1	2	3	1	2	3	1	2	3
Annelida	Haplotaxida	Enchytraeidae	<i>Cognettia</i>	<i>cognetti</i>	0	0	0	0	0	0	0	65	0
			<i>Enchytraeus</i>	<i>bulbosus</i>	0	0	0	0	0	0	2	2	0
				<i>norvegicus</i>	24	0	65	24	0	0	0	0	0
			<i>Fridericia</i>	<i>galba</i>	0	0	4	0	0	0	0	0	0
				<i>isseli</i>	25	0	0	25	26	0	0	0	0
				<i>magna</i>	0	30	0	0	0	0	0	0	0
			<i>Henlea</i>	<i>nasuta</i>	0	0	0	0	3	0	0	0	0
				<i>perpusilla</i>	0	0	0	0	0	0	225	0	0
			<i>Marionina</i>	<i>clavata</i>	32	0	0	32	0	0	0	0	0
			<i>Oconnorella</i>	<i>cambrensis</i>	1200	2034	1000	1200	0	6	0	370	0
			<i>Apporectodea</i>	<i>longa</i>	0	0	0	0	61	8	0	0	0
			<i>Satchellius</i>	<i>mammalis</i>	0	0	0	0	875	0	0	0	0
					2	2	4	2	0	0	0	0	0
					0	0	0	0	5	0	0	0	0
Arthropoda	Collembola	Entomobryidae	<i>Entomobrya</i>	<i>multifasciata</i>	0	0	0	0	0	0	0	0	0
			<i>Lepidocyrtus</i>	<i>lanuginosus</i>	0	0	0	0	0	0	0	0	0
			<i>Orchesella</i>	<i>cincta</i>	0	0	18	0	3	0	0	0	0
		Hypogastruridae	<i>Ceratophysella</i>	<i>denticulata</i>	2	0	0	2	0	0	0	0	0
			<i>Willemia</i>	<i>denisi</i>	0	2	0	0	0	0	0	0	0
		Isotomidae	<i>Folsomia</i>	<i>quadriloculata</i>	2	0	0	2	0	0	0	0	0
			<i>Isotomiella</i>	<i>minor</i>	12	2	0	12	0	0	0	0	0
			<i>Parisotoma</i>	<i>notabilis</i>	35	75	0	35	119	41	120	0	0
		Araneae	<i>Tegenaria</i>	<i>agrestis</i>	0	0	2	0	0	0	0	0	0
			<i>Dysdera</i>	<i>alegranzaensis</i>	0	0	0	0	0	0	2	0	0
			<i>Linyphiidae</i>	<i>Megalephyphantes</i>	0	0	0	0	0	0	0	0	2
	Chilopoda	<i>Lithobius</i>	<i>forficatus</i>	0	85	0	0	0	0	0	0	0	0
		<i>Elateridae</i>	<i>Agrypnus</i>	<i>murinus</i>	0	0	0	0	0	0	0	37	0
		<i>Elateridae</i>	<i>Athous</i>	<i>puncticollis</i>	0	0	0	0	0	0	0	28	0
	Diptera	<i>Nitidulidae</i>	<i>Meligethes</i>	<i>erysimicola</i>	0	0	2	0	0	0	0	0	0
		<i>Chironomidae</i>	<i>Ablabesmyia</i>	<i>monilis</i>	0	0	0	0	0	0	2	0	0
		<i>Culicidae</i>	<i>Anopheles</i>	<i>pristinus</i>	2	0	0	2	0	0	0	0	0
	Geophilomorpha	<i>Tephritidae</i>	<i>Urophora</i>	<i>cardui</i>	0	3	0	0	0	0	0	0	0
		<i>Schendylidae</i>	<i>Schendyla</i>	<i>nemorensis</i>	0	0	0	0	106	0	0	0	0
		<i>Acanthosomatidae</i>	<i>Acanthosoma</i>	<i>haemorrhoidale</i>	0	0	0	0	0	0	0	0	3
	Hemiptera	<i>Miridae</i>	<i>Apolygus</i>	<i>subpulchellus</i>	0	2	0	0	0	0	0	0	0
		<i>Braconidae</i>	<i>Charmon</i>	<i>extensor</i>	0	0	0	0	0	0	2	0	0
		<i>Vespidae</i>	<i>Ancistrocerus</i>	<i>parietum</i>	0	0	0	0	0	0	17	0	0
	Lepidoptera	<i>Tortricidae</i>	<i>Pandemis</i>	<i>cerasana</i>	0	13	0	0	0	0	0	0	0
		<i>Acrididae</i>	<i>Locusta</i>	<i>migratoria</i>	0	0	0	0	0	0	5	0	2
		<i>Orthoptera</i>	<i>Achipteria</i>	<i>coleoprata</i>	0	0	2	0	0	0	0	0	0
Sarcoptiformes		<i>Camisiidae</i>	<i>Platynothrus</i>	<i>peltifer</i>	0	6	3	0	20	0	0	0	0
		<i>Eremaeidae</i>	<i>Euermaeus</i>	<i>oblongus</i>	2	0	0	2	0	0	0	0	0
		<i>Hypochothoniidae</i>	<i>Hypochothonius</i>	<i>rufulus</i>	0	0	0	0	0	0	18	0	0
		<i>Nothridae</i>	<i>Nothrus</i>	<i>silvestris</i>	6	0	0	6	0	0	0	3	0
		<i>Scutoverticidae</i>	<i>Scutovertex</i>	<i>sculptus</i>	5	0	0	5	0	0	0	0	0
					0	0	0	0	0	0	0	0	0
					0	0	0	0	0	0	0	0	0

Table 4 Liste des espèces détectées dans les différents pseudoréplicats des trois habitats échantillonnés. Le nombre de séquences est figuré avec un gradient de couleur en fonction de l'abondance recouvrée pour chaque espèce.

			EA20			SA70			SABO		
			1	2	3	1	2	3	1	2	3
Annelida	Haplotaenia	Enchytraeidae	1838	3374	6926	0	545	2561	1117	456	902
		Lumbricidae	2	3	16	91	1005	8	2	465	122
Arthropoda	Araneae	Agelenidae	0	0	2	0	0	0	0	0	0
		Araenidae	3	0	3	0	0	0	0	0	0
		Cybaeidae	0	0	0	0	0	0	5	0	0
		Dysderidae	0	0	0	0	0	0	2	0	0
		Linyphiidae	0	0	0	0	0	0	2	0	2
		Lycosidae	0	0	0	0	0	0	0	0	8
		Pholcidae	0	0	3	0	0	0	0	0	0
		Salicidae	0	0	4	0	0	0	0	0	0
		Geophilidae	0	0	29	0	0	0	0	0	0
		Lithobiidae	0	87	0	0	0	0	0	0	0
Coleoptera	Coleoptera	Carabidae	0	0	0	0	0	0	8	0	8
		Cerambycidae	0	0	47	0	0	0	0	0	0
		Chrysomelidae	7	0	0	0	0	0	0	0	0
		Coccinellidae	0	0	0	0	0	0	2	0	0
		Curculionidae	0	0	324	0	0	0	0	0	0
		Elateridae	0	0	0	0	0	0	0	65	0
		Hydraenidae	0	0	0	0	0	2	0	0	0
		Leiodidae	0	9	0	0	0	0	0	0	0
		Nitidulidae	0	0	2	0	0	0	0	0	0
		Scarabaeidae	0	0	5	0	0	0	0	0	0
Collembola	Collembola	Entomobryidae	2	12	22	0	8	0	0	0	6
		Hypogastruridae	2	4	0	0	7	0	4	0	0
		Isotomidae	62	118	26	3	137	61	120	2	18
		Neaturidae	16	0	0	0	0	0	0	0	0
		Neelidae	0	7	14	0	0	5	52	3	0
		Onychiuridae	14	0	2	0	3	5	4	15	0
		Sminthuridae	0	0	53	0	0	0	0	0	0
		Tomoceridae	0	29	0	0	12	0	0	0	2
		Tullbergidae	0	158	0	0	0	0	4	0	119
		Cecidomyiidae	0	3	0	0	0	0	0	0	0
Diptera	Diptera	Chironomidae	23	10	0	13	15	79	104	4	77
		Chloropidae	0	0	0	0	0	0	0	0	0
		Culicidae	2	0	0	0	0	2	11	5	0
		Drosophilidae	67	0	22	0	0	0	2	0	0
		Ephyridae	0	0	0	0	0	0	10	3	0
		Phoridae	0	0	0	0	0	0	0	0	0
		Psychodidae	0	2	0	0	0	0	0	0	0
		Rhagionidae	55	0	0	0	0	0	0	0	0
		Sciaridae	0	0	16	0	0	187	0	0	0
		Simuliidae	0	4	0	0	0	0	0	0	0
Geophilomorpha	Geophilomorpha	Tachinidae	0	2	4	0	0	0	0	2	0
		Tephritidae	0	3	0	0	0	0	0	0	0
		Linotaenidae	20	0	0	0	0	0	0	0	0
		Schendylidae	0	0	0	0	106	0	0	0	0
		Acanthosomatidae	0	0	0	0	0	0	0	0	3
		Alydidae	227	0	0	0	0	0	0	0	0
		Cicadellidae	0	6	4	0	0	0	0	0	0
		Issidae	0	0	0	0	5	0	3	0	0
		Largidae	0	14	0	0	0	0	0	0	0
		Lygaeidae	0	0	0	0	0	31	0	0	0
Hymenoptera	Hymenoptera	Miridae	0	4	0	0	0	3	0	0	0
		Pentatomidae	0	2	0	0	0	0	0	0	0
		Reduviidae	0	0	0	0	0	0	0	3	0
		Braconidae	0	0	2	0	0	7	7	4	0
		Colletidae	0	4	0	0	0	0	0	0	0
		Formicidae	0	0	0	0	0	0	9	0	0
		Ichneumonidae	0	0	0	0	0	9	0	0	0
		Vespidae	0	0	0	0	0	0	19	0	0
		Asellidae	0	3	0	0	2	8	6	0	2
		Ligidae	0	0	0	0	2	2	0	0	0
Isopoda	Isopoda	Philosciidae	0	0	0	0	0	0	0	0	0
		Stenasellidae	4	0	0	0	0	0	0	0	0
		Trichoniscidae	0	0	0	0	0	0	0	0	0
		Ixodidae	0	0	7	0	0	0	0	0	0
		Geometridae	0	2	2	0	6	0	0	0	0
		Noctuidae	0	10	0	0	0	0	0	0	0
		Prodoxidae	0	0	0	0	0	0	3	0	0
		Saturniidae	0	0	0	0	0	6	0	0	3
		Tortricidae	0	13	0	0	0	0	0	0	0
		Parasitidae	0	0	0	0	0	6	0	0	0
Mesostigmata	Mesostigmata	Phalangidae	0	0	0	0	0	0	0	0	0
		Acrididae	16	22	12	0	4	0	5	65	112
		Gryllidae	0	0	0	0	3	20	0	0	0
		Pamphagidae	0	0	0	0	0	0	0	0	2
		Chthonidae	0	21	0	0	0	0	0	0	0
		Achipteridae	16	18	8	0	4	0	0	16	0
		Brachychthoniidae	7	2	0	2	4	13	0	17	272
		Camisiidae	0	6	0	0	2	13	7	0	0
		Cepheidae	0	0	0	0	0	0	0	9	0
		Ceratopidae	20	0	0	0	0	0	0	0	0
Opiliones	Opiliones	Ceratopidae	0	0	0	0	0	0	0	3	0
		Damaeidae	0	47	0	0	0	0	0	0	0
		Eniochthonidae	0	0	0	3	0	2	0	0	0
		Eremaeidae	48	0	0	0	0	0	0	0	0
		Hypochthoniidae	0	0	0	0	0	0	18	0	0
		Nothridae	6	0	0	0	0	0	0	3	0
		Opidae	40	5	38	0	4	164	0	83	0
		Scutoverticidae	10	0	0	0	0	0	0	0	0
		Tectocephidae	0	0	0	0	3	0	0	0	0
		Cryptopidae	0	0	0	0	0	0	0	0	0
Scolopendromorpha	Scolopendromorpha	Scolopendridae	0	0	0	0	0	0	11	0	0
		Scutigerellidae	0	0	0	2	0	0	0	4	9
		Eupodidae	43	0	15	0	2	506	141	105	587
		Pentaleidae	0	0	0	0	0	0	0	304	0
		Rhagidae	0	0	2	0	0	0	0	0	0
		Tydeidae	0	0	0	0	0	29	0	2	0
		Diplomatidae	0	0	0	0	0	21	0	0	15
		Chondrinidae	0	0	0	0	0	0	9	0	0
		Clausidae	0	0	0	0	0	0	0	0	2
		Helicidae	0	0	2	0	0	0	0	0	0

Table 5 Liste des familles détectées dans les différents pseudoréplicats des trois habitats échantillonnés. Le nombre de séquences est figuré avec un gradient de couleur en fonction de l'abondance recouvrée pour chaque famille.

	EA20			SA70			SABO		
	1	2	3	1	2	3	1	2	3
Araneae	3	20	12	8	6	12	11	4	89
Architaenioglossa	4	0	0	0	0	21	0	0	15
Chilopoda	0	87	29	0	0	0	0	0	0
Coleoptera	7	9	559	0	0	2	14	67	8
Collembola	101	392	151	5	207	89	207	22	172
Diptera	151	30	53	13	15	333	148	16	77
Geophilomorpha	20	0	0	0	106	2	0	0	0
Haplotaxida	3420	5464	9447	186	1681	2860	1268	2808	1154
Hemiptera	246	28	4	0	5	34	3	5	8
Hymenoptera	0	4	2	0	0	21	30	4	3
Isopoda	4	3	0	0	4	12	6	0	2
Ixodida	0	0	7	0	0	0	0	0	0
Lepidoptera	0	44	2	2	8	9	10	18	11
Mesostigmata	0	0	2	0	3	6	0	0	0
Opiliones	0	6	0	0	0	0	0	0	0
Orthoptera	16	25	12	0	11	20	5	65	119
Pseudoscorpiones	0	21	0	0	0	0	0	0	0
Sarcoptiformes	418	116	78	5	24	206	25	150	272
Scolopendromorpha	0	0	0	0	0	0	11	0	0
Stylommatophora	0	0	2	0	0	0	11	0	2
Symphyla	0	0	0	2	0	0	0	4	9
Trombidiformes	128	32	95	0	20	586	224	759	596

Table 6 Liste des ordres détectées dans les différents pseudoréplicats des trois habitats échantillonnés. Le nombre de séquences est figuré avec un gradient de couleur en fonction de l'abondance recouvrée pour chaque ordre.

Une liste d'espèces, de familles et d'ordres a ainsi été obtenue pour chaque type d'habitat, avec les réserves dues à l'incomplétude des bibliothèques de référence principalement pour le niveau spécifique. Néanmoins, on remarque, pour ces trois niveaux, une forte hétérogénéité dans la mesure des communautés entre les pseudoréplicats de chaque habitat, avec en moyenne 78 % des espèces, 66 % des familles et 24 % des ordres ayant été retrouvés dans un seul des pseudoréplicats pour un habitat donné (Table 7). Ceci montre la nécessité absolue d'effectuer plusieurs pseudoréplicats pour chaque point échantillonné, et du même coup met en exergue le besoin d'un calibrage du nombre de pseudoréplicats nécessaires pour optimiser la capture de la diversité en invertébrés du sol présente au niveau d'un point d'échantillonnage. Ce niveau de calibrage variera bien sûr en fonction de la taille et la densité des taxa ciblés ; il faudra par exemple beaucoup plus de pseudoréplicats pour rendre compte de la diversité en chilopodes qu'en collembolles du fait de la densité beaucoup plus faible du groupe de myriapode prédateur. Le cas extrême étant la diversité des vers de terre qui pourra difficilement être capturée avec ce type d'échantillonnage, spécialement lorsqu'il est lié à un protocole utilisant des tailles de marqueurs requérant la présence de tissus frais ou dégradé depuis peu dans l'échantillon. Pour ce type de groupe il faudra s'orienter soit vers un autre protocole d'échantillonnage plus spécifique, soit vers l'utilisation d'un marqueur plus petit ciblant l'ADN extracellulaire comme dans le chapitre précédent.

	Espèces									Familles									Ordres								
	EA20			SA70			SABO			EA20			SA70			SABO			EA20			SA70			SABO		
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
Nb de taxa détectés par pseudoréplicat	13	11	9	13	9	3	9	6	4	25	32	29	6	21	25	28	23	20	12	15	15	7	12	15	14	12	15
Nb taxa communs 3 pseudoréplicats	2			1			0			7			4			6			9			6					11
Nb taxa communs 2 pseudoréplicats	4			2			2			10			9			11			2			6					3
Nb taxa communs 1 pseudoréplicats	19			7			15			44			22			31			7			4					2
Nb Total de taxa pour le point	25			10			17			61			35			48			20			16					16
Proportion de taxa trouvées sur un seul pseudoréplicat	76%			70%			88%			72%			63%			64%			35%			25%					12%

Table 7 Résumé des résultats obtenus pour chaque pseudoréplicat et chaque point d'échantillonnage de chaque habitat en termes de nombre de taxa (espèces, familles et ordres).

Ces résultats montrent, pour les cortèges d'espèces et de familles, une diversité supérieure en forêt de plaine (Eawy), intermédiaire en prairie de coteau calcaire et plus faible en forêt de coteau

calcaire. Des résultats qui pourront être confirmés par l'analyse d'autres échantillons de ces mêmes habitats et possiblement couplés à la trajectoire historique d'évolution de l'occupation des sols sur ces mêmes points.

2.4.4.2 Comparaison entre habitats

La figuration des communautés par projection sur trois axes, permet de visualiser les similitudes à différents niveaux taxonomiques - espèces, genres et familles -, et cela pour quatre grands ensembles, les annélides, les collemboles, les gastéropodes et le reste de la communauté d'arthropodes (comprenant principalement les ordres d'insectes et les acariens) (Figure 16). On voit sur ce graphe, que très peu d'espèces sont partagées entre les trois habitats cibles. Au niveau générique, on a des similitudes marquées pour les trois habitats en ce qui concerne les annélides, en revanche, on observe une différenciation notable du cortège des arthropodes hors collemboles qui caractérise largement la prairie de coteau calcaire (SABO) avec le plus grand nombre de genres caractéristiques. Il est à noter également qu'une spécificité au niveau des gastéropodes rapproche la prairie et la forêt de St Adrien par rapport au site d'Eawy. Que ce soit au niveau des annélides ou des collemboles, on a une grande proximité des deux points de St Adrien et la forêt d'Eawy (SA70 et EA20) au niveau des communautés, avec cependant des genres caractéristiques pour cette dernière dans les deux groupes. Les familles de collemboles et d'annélides semblent très peu caractériser les trois types d'habitats, la différenciation venant des familles d'arthropodes hors Collemboles qui sont dominés par les insectes et les acariens.

On a donc une bonne caractérisation au niveau des communautés d'espèces recouvrées pour les trois types d'habitat, caractérisation qui s'érode naturellement avec la dégradation du niveau taxonomique, mais qui reste néanmoins très importante pour les familles et les genres d'insectes, de myriapodes et d'acariens (groupés sous 'Autres Arthropodes').

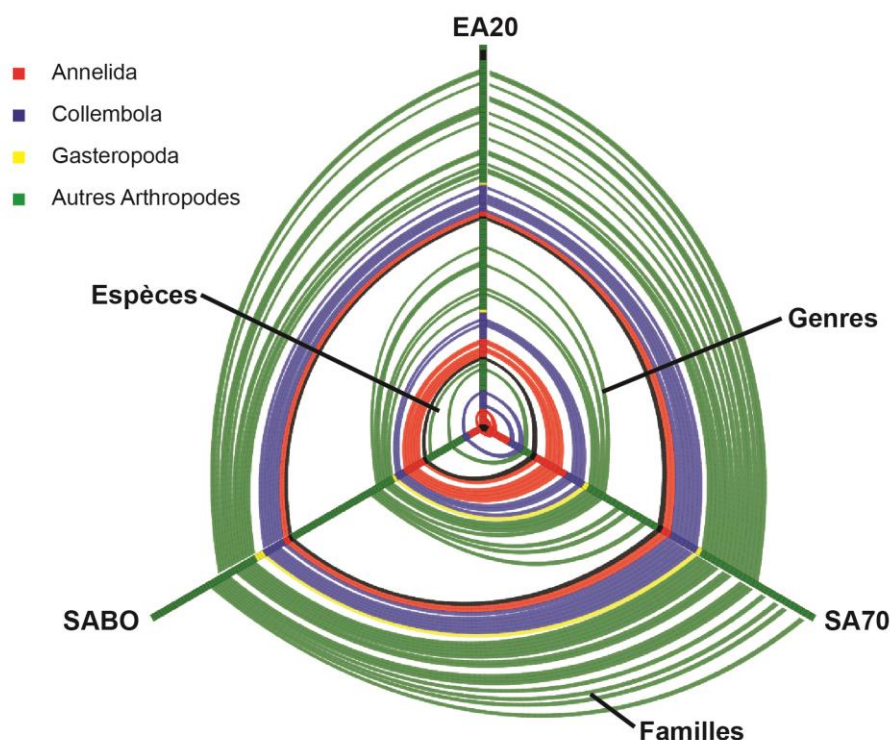


Fig. 16 Projection sur les 3 axes des communautés d'invertébrés recouvrées dans les 3 types d'habitat pour les niveaux taxonomiques espèces, genres et familles. Les arcs de cercles représentent les taxa communs entre les types d'habitat.

2.4.5 Conclusion

L'application de ce protocole a permis de recouvrer les communautés d'invertébrés du sol dans trois types d'habitats caractéristiques de Haute Normandie avec la possibilité d'en comparer les diversités à différents niveaux taxonomiques. Il ressort de ces premiers résultats que la technique d'échantillonnage peut être encore être optimisée suivant les groupes et que la complétude des bibliothèques de références doit être augmentée de façon à obtenir une meilleure caractérisation des communautés. Il est cependant à noter que le niveau de diversité spécifique de certains groupes pourra être inféré de manière plus approfondie dans un second jeu d'analyses avec un seuil de divergence génétique fixé par rapport à ce qui a été observé précédemment dans ces mêmes groupes. On obtiendrait alors un proxy de la diversité spécifique en nombre de MOTUS. Cette technique permettra ainsi de caractériser les niveaux de diversité spécifique de nombreux groupes et de comparer la composition relative de leurs communautés sans avoir recours à une bibliothèque de référence détaillée. De la même façon, le séquençage en parallèle pour un même échantillon d'autres gènes employés spécifiquement comme barcode dans d'autres groupes essentiels dans le fonctionnement des sols enrichirait le spectre de l'analyse : 18S pour les nématodes, 16S pour les bactéries et ITS pour les champignons.

3. CONCLUSIONS GENERALES / PERSPECTIVES

Cette étude a permis d'une part d'initier l'assemblage systématique de bibliothèques de référence pour de nombreux groupes invertébrés du sol en Haute Normandie, et d'autre part d'expérimenter deux techniques novatrices de recouvrement de la composition des communautés par des approches de métabarcoding.

L'une d'elle, MetaBar, utilisant un marqueur nucléaire de petite taille dont les cibles principales sont les fragments d'ADN extracellulaires issus de la dégradation de tissus des différents organismes présents dans le compartiment sol des écosystèmes. Cette méthode, par l'utilisation d'un marqueur à variabilité modérée par rapport à un marqueur mitochondrial, permet de caractériser, dans un premier temps, l'occurrence de grands embranchements, ordres ou familles. L'utilisation complémentaire d'un marqueur mitochondrial, comme pour la caractérisation des communautés de vers de terre (Bienert *et al.*, 2012), pourrait permettre d'accéder à un niveau spécifique de caractérisation des communautés. L'un des avantages majeur de cette technique est que l'on peut la mettre en œuvre sur des échantillons de sol secs et donc potentiellement valoriser les pédothèques constituées jusque-là.

L'autre méthode utilise deux fragments plus longs du marqueur mitochondrial COI correspondant aux bibliothèques de références assemblées pour les invertébrés du sol de Haute Normandie. Son application a permis de caractériser les communautés de trois types d'habitats à un niveau spécifique, générique, familial et ordinal grâce à une comparaison avec ces bibliothèques. Des bibliothèques de référence plus complètes permettront d'établir un diagnostic plus précis de ces communautés. Cependant, ces données, permettront, par l'application de différents niveaux de seuil selon les groupes, d'obtenir un proxy du nombre d'espèce.

Pour les deux méthodes, des expérimentations, quant au volume nécessaire d'échantillon et d'échantillonnage, pour obtenir un niveau de représentativité optimum des communautés d'invertébrés du sol, sont encore à effectuer de façon à parfaire la standardisation des protocoles. En outre, on pourrait imaginer le couplage des deux protocoles, ce qui permettrait par exemple de palier avec l'approche ADN extracellulaire les manquements de l'autre protocole au niveau de l'échantillonnage de la grande faune de type vers de terre.

Au final, le projet GENOSOIL a permis d'expérimenter et d'initier le développement d'une véritable boîte à outils pour la connaissance des communautés d'invertébrés du sol. A la suite des différentes mises au point de protocole nécessaire, on aura à disposition un panel d'approches qui permettront de répondre à la fois à des questions sur la structuration des communautés de façon transversale au niveau spatial (paysages) ou au niveau temporel (changements de pratiques de gestion), mais aussi relatives à la dynamique des réseaux trophiques, à des échelles, et pour une largeur de champ taxonomique jusqu'ici hors d'atteinte par les méthodes classiques d'étude.

4. APPUI POTENTIEL AUX POLITIQUES PUBLIQUES

De plus en plus d'intérêt est porté au sol dans les politiques publiques suite à une prise de conscience sociétale de la fonction essentielle de ce compartiment des écosystèmes du fait de la délivrance de biens et de services écosystémiques tels, qu'entre autres, la production de ressources alimentaires et énergétiques, la régulation des flux hydriques, ou encore celle des cycles biogéochimiques. Nombreux sont ces biens et services dont la production est liée à l'activité biologique des communautés d'invertébrés du sol qui restent, paradoxalement, très mal connues. Cette méconnaissance rend les prises de décisions rationnelles en termes de politiques locales, nationales ou européennes, au mieux difficile. L'implémentation systématique des types d'approches génétiques qui ont été développés au cours de ce projet permettra d'avoir une vision qualitative et quantitative de ces communautés pour de large portion de territoire. Celle-ci pourra être liée de façon précise à la quantification des facteurs biotiques et abiotiques critiques pour la production de services et biens d'intérêt. Les impacts des aménagements du paysage, avec les changements d'usage des terres que ceux-ci impliquent, pourront alors être mesurés finement au niveau de marqueurs d'intérêt associés à certains éléments des communautés, et donneront le potentiel de prédiction nécessaire à une prise de décision basée sur des indicateurs objectifs.

Concrètement, cela permettra d'établir des indicateurs de la santé des sols pour mesurer objectivement les effets de politiques de restauration, de réaménagement, de changement d'utilisation des terres ou de l'installation d'infrastructures. On pourra alors avoir une vision en termes de conservation et de bioindication liée à un compartiment des écosystèmes pour lequel de telles approches restaient difficiles voire impraticables.

L'autre application importante qui peut découler de l'utilisation de ce panel de techniques est la détection d'éléments invasifs dans les communautés des sols. L'utilisation en routine sur des sites sentinelles permettrait d'établir une veille quant à l'apparition et la progression de tels éléments. En outre, cela produirait des informations précieuses concernant l'impact de ces espèces invasives sur les communautés d'invertébrés édaphiques, que ce soit via la compétition pour la ressource, la prédation directe ou encore la transformation physico-chimique de la matrice des sols ; différents facteurs qui pourront être mesurés et corrélés à la pression des propagules d'invasifs. L'application de veille sanitaire peut aussi être envisagée de la même façon pour des ravageurs agricoles et sylvicoles (par exemple pour les nématodes phytoparasites). Cette démarche fournirait des informations de premier choix aux services régionaux de protection des végétaux et permettrait d'élaborer des stratégies de contrôle appropriées.

On voit donc que le développement de ces approches, et la standardisation des protocoles attendant, en permettant leur utilisation en routine, produiraient une masse d'informations importantes, quasiment en temps réel, qui éclairerait, à l'aide d'indicateurs objectifs, la prise de décisions pour une gestion plus efficace de l'usage des sols et des ressources attenantes.

5. PRODUCTIONS SCIENTIFIQUES

5.1 Bases de données et collections de référence

Comme mentionné précédemment, l'ensemble des séquences de référence de COI obtenues a été placé dans des projets spécifiques sur la plateforme bioinformatique BOLD (jeu de données 'Bibliothèques de référence GENOSOIL -Invertébrés du sol' [DS-BGENO]). L'ensemble de ces données est en accès libre avant la fin du projet GENOSOIL, de façon à servir de base de référence pour l'utilisation des approches métagénomiques développées dans le projet GENOSOIL ou dans d'autres comme par exemple le projet MetaBar coordonné par P. Taberlet.

5.2 Publications réalisées ou en cours de préparation dans des journaux internationaux

Porco D., Dupont L., Bodilis J., Chang C. H., Dolédec S., Dubs F., Hajibabaei M., James S., Lavelle P., Oszwald J., Richard B., Rossi J.-P., Rougerie R., Roy V., Sebaq D., Shokralla S., Decaëns T. A

reference library of DNA barcodes for the earthworm of upper Normandy: Biodiversity assessment and cryptic diversity. En préparation.

Decaëns T., Porco D., Rougerie R., Brown G.G., James S.W. (2013) Potential of DNA barcoding for earthworm research in taxonomy and ecology. *Applied Soil Ecology*, 65, 35-42.

Porco D., Decaëns T., Deharveng L., James S.W., Skarżyński D., Erséus C., Butt K., Richard B., Hebert P. (2013) Biological invasions in soil: DNA barcoding as a monitoring tool in a multiple taxa survey targeting European earthworms and collembolans in North America. *Biological Invasion*, 15, 899-910.

Dupont L., Lazrek F., Porco D., King R.A., Rougerie R., Symondson W.O.C., Livet A., Richard B., Decaëns T., Butt K.R., Mathieu J. (2011) New insight into the genetic structure of the *Allolobophora chlorotica* aggregate in Europe using microsatellite and mitochondrial data. *Pedobiologia*, 54(4), 217-224.

Rossi J.-P. (2011) rich: an R package to analyse species richness. *Diversity* 3, 112-120.

Antérieures à 2011 (phase préliminaire de GENOSOIL)

James S., Porco D., Decaëns T., Richard B., Rougerie R., Erséus C. (2010) DNA Barcoding Reveals Cryptic Diversity in *Lumbricus terrestris* L., 1758 (Clitellata): Resurrection of *L. herculeus* (Savigny, 1826). *PLoS One*, 5, e15629.

Richard B., Decaëns T., Rougerie R., James S., Porco D., Hebert P. (2010) Re-integrating earthworm juveniles into soil biodiversity studies: species identification through DNA barcoding. *Molecular Ecology Resources*, 10, 606–614.

5.3 Communications

Decaëns T., Porco D., James S., Smith M.A., Rougerie R., Hajibabaei M. (2012) Barcoding the underground – recent advances and new prospects for the study of soil animal biodiversity. Conférence invitée (Keynote Talk) au « XVth International Colloquium on Soil Zoology », Coimbra, Portugal, 6-10 août 2012.

Decaëns T., Porco D., Sebag D., Oszwald J., Rougerie R., Husté A., Dupont L., Dolédec S., Rossi J.-P., Hajibabaei M. (2012) The GENOSOIL-Fr project: insights from metagenomics into the study of total biodiversity of soils. Poster au « XVth International Colloquium on Soil Zoology », Coimbra, Portugal, 6-10 août 2012.

Decaëns T., Rougerie R., James S., Porco D., Hebert P. 2012 DNA barcodes as a tool for the study of invertebrate biodiversity. Conférence invitée à l'Universidade do Estado de Santa Catarina, Lages, Brésil, 03 mars 2012.

Decaëns T., Porco D., Sebag D., Oswald J., Rougerie R., Husté A., Dupont L., Doledec S., Rossi J.-P., Hajibabaei M. (2011) The GENOSOIL-Fr project : insights from metagenomics into the study of total biodiversity of soils. Poster et lightning au « Fourth International Barcode of Life Conference », Adelaide, Australie, 30 novembre – 3 décembre 2011.

Antérieures à 2011 (phase préliminaire de GENOSOIL)

Decaëns T., Porco D., James S.W. (2010) El potencial de los barcodes de DNA en los estudios taxonómicos y ecológicos con lombrices de tierra. Conférence invitée au « Encuentro Latino Americano de Ecología y Taxonomía de Oligochetos », Curitiba, Brésil, 12-14 octobre 2010.

Decaëns T., Rougerie R., Deharveng L., Porco P., James S.W., Chang C.H., Richard B., Hebert P.D.N. (2010) DNA barcodes for soil animal taxonomy: transcending the final frontier. Présentation

orale, colloque « ECBOL2 - 2nd Conference of the European Consortium for the Barcode of Life », Braga, Portugal, 2-4 June 2010.

James S.W., Decaëns T., Porco D., Richard B., Erseus C., Meusnier I., Rougerie R., Hebert P.D.N. (2010) Chimerical biology: a plea for good taxonomy as exemplified by the common nightcrawler (*Lumbricus terrestris* L.). Poster au colloque « ECBOL2 - 2nd Conference of the European Consortium for the Barcode of Life », Braga, Portugal, 2-4 June 2010.

Porco D., Decaëns T., Deharveng L., James S., Butt K., Mark S., Rougerie R., Richard B., Hebert P. (2010) Barcoding invasive: a new tool for invasion monitoring in soil. Présentation orale, colloque « ECBOL2 - 2nd Conference of the European Consortium for the Barcode of Life », Braga, Portugal, 2-4 June 2010.

Richard B., Decaëns T., James S., Rougerie R., Porco D., Hebert P.D.N. (2010) Re-integrating earthworm juveniles into soil biodiversity studies: species identification through DNA barcoding. Poster, colloque « ECBOL2 - 2nd Conference of the European Consortium for the Barcode of Life », Braga, Portugal, 2-4 June 2010.

5.4 Rapports de stage

Coupiereau D. (2012) *Evaluation de la diversité des vers de terre dans deux gradients environnementaux à l'aide du barcoding ADN*. Stage de L3 EBO (Univ Rouen), réalisé au lab ECODIV (encadrant: T. Decaëns).

Salim H. (2012) *Utilisation du barcoding ADN pour l'estimation de la richesse spécifique des invertébrés du sol dans différents écosystèmes de Normandie*. Stage de M2 Environnement Continental et Hydrosociété, parcours Sol (AgroParisTech/UPMC), réalisé au lab ECODIV (encadrants: T. Decaëns & D. Porco).

James S.W. (2011) *Utilisation du barcoding ADN dans l'étude des vers de terre : mise en évidence de la diversité cryptique et amélioration de nos connaissances écologiques*. Rapport de fin de recherche, bourse « Marie Curie France Régions », laboratoire ECODIV (resp. scientifique : T. Decaëns).

Lebarque M. (2011) *Caractérisation différentielle en prairie et forêt des communautés de la mésofaune et de la macrofaune*. Stage de L3 EBO (Univ Rouen), réalisé au lab ECODIV (Encadrants : T. Decaëns & D. Porco).

Lallemant H. (2011) *Caractérisation différentielle en prairie et forêt des communautés de la mésofaune et de la macrofaune*. Stage de L3 EBO, réalisé au Lab ECODIV (Encadrants : T. Decaëns & D. Porco).

Leroux L. (2011) *Caractérisation des dynamiques paysagères entre 1947 et 2010 dans la zone de St Adrien (Normandie)*. Stage de Master 1 TASE (Université Rennes 2) réalisé au laboratoire COSTEL (encadrant : J. Oszwald).

Poirier E. (2011) *Caractérisation des dynamiques paysagères entre 1947 et 2010 dans la zone de Eawy (Normandie)*. Stage de Master 1 TASE (Université Rennes 2) réalisé au laboratoire COSTEL (encadrant : J. Oszwald).

6. BIBLIOGRAPHIE

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215, 403-410.

Bienert, F., De Danieli, S., Miquel, C., Coissac, E., Poillot, C., Brun, J. J., et al. (2012). Tracking earthworm communities from soil DNA. *Molecular Ecology*, 21, 2017-2030.

Bouché, M. B. (1972). *Lombriciens de France*. Ecologie et systématique. Institut natn Rech agron Publ, 1-671.

Clarke, K. R., & Gorley, R. N. (2006). Primer v6: User manual/tutorial. Plymouth.

Colwell R.K. 2013. Estimates: Statistical estimation of species richness and shared species from samples. Version 9: <http://purl.oclc.org/estimates>.

- Decaëns, T., Margerie, P., Aubert, M., Hedde, M., & Bureau, F. (2008). Assembly rules within earthworm communities in north-western France - a regional analysis. *Applied Soil Ecology*, 39, 321-335.
- Decaëns, T., Margerie, P., Renault, J., Bureau, F., Aubert, M., & Hedde, M. (2011). Niche overlap and species assemblage dynamics in an ageing pasture gradient in north-western France. *Acta Oecologica-International Journal of Ecology*, 37, 212-219.
- Decaëns, T., Porco, D., Rougerie, R., Brown, G. G., & James, S. W. (2013). Potential of DNA barcoding for earthworm research in taxonomy and ecology. *Applied Soil Ecology*, 65, 35-42.
- Edgar, R. C. (2010). Search and clustering orders of magnitude faster than blast. *Bioinformatics*, 26, 2460-2461.
- Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., & Knight, R. (2011). Uchime improves sensitivity and speed of chimera detection. *Bioinformatics*, 27, 2194-2200.
- Eisenhauer, N., Milcu, A., Nitschke, N., Sabais, A. C. W., Scherber, C., & Scheu, S. (2009). Earthworm and belowground competition effects on plant productivity in a plant diversity gradient. *Oecologia*, 161, 291-301.
- Fiore, C., Tull, J. L., Zehner, S., & Ducey, P. K. (2004). Tracking and predation on earthworms by the invasive terrestrial planarian *Bipalium adventitium* (tricladida, platyhelminthes). *Behavioural Processes*, 67, 327-334.
- Folmer, O., Black, M., Hoeh, W., Lutz, R., & Vrijenhoek, R. (1994). DNA primers for amplification of mitochondrial cytochrome c oxidase subunit i from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology*, 3, 294-299.
- Forey, E., Barot, S., Decaëns, T., Langlois, E., Laossi, K. R., Margerie, P., et al. (2011). Importance of earthworm-seed interactions for the composition and structure of plant communities: A review. *Acta Oecologica-International Journal of Ecology*, 37, 594-603.
- Gobat, J. M., Aragno, M., & Matthey, W. (2010). *Le sol vivant: Bases de pédologie, biologie des sols* (3e édition): Presses polytechniques et universitaires romandes.
- Gotelli N. J. & Entsminger G. L., 2006. *Ecosim: Null models software for ecology*. Ver. 7. Jericho, Vermont, USA: Acquired Intelligence and Kesity-Bear.
- Hajibabaei, M., DeWaard, J. R., Ivanova, N. V., Ratnasingham, S., Dooh, R. T., Kirk, S. L., et al. (2005). Critical factors for assembling a high volume of DNA barcodes. *Philosophical Transactions of the Royal Society B-Biological Sciences*, 360, 1959-1967.
- Hajibabaei, M., Smith, M. A., Janzen, D. H., Rodriguez, J. J., Whitfield, J. B., & Hebert, P. D. N. (2006). A minimalist barcode can identify a specimen whose DNA is degraded. *Molecular Ecology Notes*, 6, 959-964.
- Hajibabaei, M., Spall, J. L., Shokralla, S., & van Konynenburg, S. (2012). Assessing biodiversity of a freshwater benthic macroinvertebrate community through non-destructive environmental barcoding of DNA from preservative ethanol. *Bmc Ecology*, 12.
- Hall, T. A. (1999). Bioedit: A user-friendly biological sequence alignment editor and analysis program for windows 95/98/nt. *Nucleic Acids Symposium Series*, 41, 95-98.
- Huson, D. H., Mitra, S., Ruscheweyh, H. J., Weber, N., & Schuster, S. C. (2011). Integrative analysis of environmental sequences using megan4. *Genome Research*, 21, 1552-1560.
- Ivanova, N. V., Dewaard, J. R., & Hebert, P. D. N. (2006). An inexpensive, automation-friendly protocol for recovering high-quality DNA. *Molecular Ecology Notes*, 6, 998-1002.
- Iwai, N., Sugiura, S., & Chiba, S. (2010). Prey-tracking behavior in the invasive terrestrial planarian *platydemus manokwari* (platyhelminthes, tricladida). *Naturwissenschaften*, 97, 997-1002.
- James, S. W., Porco, D., Decaëns, T., Richard, B., Rougerie, R., & Erséus, C. (2010). DNA barcoding reveals cryptic diversity in *Lumbricus terrestris* L., 1758 (Clitellata): Resurrection of *L. Hercules* (Savigny, 1826). *PLoS ONE*, 5, e15629.
- Jones, C. G., Lawton, J. H., & Shachak, M. (1994). Organisms as ecosystem engineers. *Oikos*, 69, 373-386.
- Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide-sequences. *Journal of Molecular Evolution*, 16, 111-120.
- King, R. A., Tibble, A. L., & Symondson, W. O. C. (2008). Opening a can of worms: Unprecedented sympatric cryptic diversity within british Lumbricid earthworms. *Molecular Ecology*, 17, 4684-4698.
- Lavelle, P., Decaëns, T., Aubert, M., Barot, S., Blouin, M., Bureau, F., et al. (2006). Soil invertebrates and ecosystem services. *European Journal of Soil Biology*, 42, S3-S15.
- Lavelle, P., & Spain, A. V. (2001). *Soil ecology*: Kluwer Academic Publishers.
- Lemon, J. (2006). Plotrix: A package in the red light district of r. *R-News*, 6, 8-12.
- Letunic, I., & Bork, P. (2007). Interactive tree of life (itol): An online tool for phylogenetic tree display and annotation. *Bioinformatics*, 23, 127-128.

- Makeschin, F. (1997). Earthworms (Lumbricidae: Oligochaeta): Important promoters of soil development and soil fertility: Marcel Dekker, Inc.
- Medinger, R., Nolte, V., Pandey, R. V., Jost, S., Ottenwalder, B., Schlotterer, C., et al. (2010). Diversity in a hidden world: Potential and limitation of next-generation sequencing for surveys of molecular diversity of eukaryotic microorganisms. *Molecular Ecology*, 19, 32-40.
- Murchie, A. K., & Gordon, A. W. (2013). The impact of the 'new zealand flatworm', *Arthurdendyus triangulatus*, on earthworm populations in the field. *Biological Invasions*, 15, 569-586.
- Author 2012. Vegan: Community ecology package. R package version 2.0-3.
- Perez-Losada, M., Bloch, R., Breinholt, J. W., Pfenninger, M., & Dominguez, J. (2012). Taxonomic assessment of Lumbricidae (Oligochaeta) earthworm genera using DNA barcodes. *European Journal of Soil Biology*, 48, 41-47.
- Perez-Losada, M., Eiroa, J., Mato, S., & Dominguez, J. (2005). Phylogenetic species delimitation of the earthworms *Eisenia fetida* (Savigny, 1826) and *Eisenia andrei* Bouché, 1972 (Oligochaeta, Lumbricidae) based on mitochondrial and nuclear DNA sequences. *Pedobiologia*, 49, 317-324.
- Plaisance, L., Caley, M. J., Brainard, R. E., & Knowlton, N. (2011). The diversity of coral reefs: What are we missing? *PLoS ONE*, 6.
- Porco, D., Bedos, A., Greenslade, P., Janion, C., Skarżyński, D., Stevens, M. I., et al. (2012). Challenging species delimitation in collembola: Cryptic diversity among common springtails unveiled by DNA barcoding. *Invertebrate Systematics*, 26, 470-477.
- Porco, D., Decaëns, T., Deharveng, L., James, S., Skarżyński, D., Erséus, C., et al. (2013). Biological invasions in soil: DNA barcoding as a monitoring tool in a multiple taxa survey targeting European earthworms and springtails in North America. *Biological Invasions*, 15, 899-910.
- Porco, D., Potapov, M., Bedos, A., Busmachiu, G., Weiner, W. M., Hamra-Kroua, S., et al. (2012). Cryptic diversity in the ubiquitous species *Parisotoma notabilis* (Collembola, Isotomidae): A long-used chimeric species? *PLoS ONE*, 7, e46056.
- Porco, D., Rougerie, R., Deharveng, L., & Hebert, P. (2010). Coupling non-destructive DNA extraction and voucher retrieval for small soft-bodied arthropods in a high-throughput context: The example of Collembola. *Molecular Ecology Resources*, 10, 942-945.
- Author 2012. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.
- Richard, B., Decaëns, T., Rougerie, R., James, S. W., Porco, D., & Hebert, P. D. N. (2010). Re-integrating earthworm juveniles into soil biodiversity studies: Species identification through DNA barcoding. *Molecular Ecology Resources*, 10, 606-614.
- Saitou, N., & Nei, M. (1987). The neighbor-joining method - a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4, 406-425.
- Scheu, S. (1987). The role of substrate feeding earthworms (Lumbricidae) for bioturbation in a beechwood soil. *Oecologia*, 72, 192-196.
- Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., et al. (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, 75, 7537-7541.
- Schmieder, R., & Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27, 863-864.
- Sims, R. W., & Gerard, B. M. (1985). Earthworms: Keys and notes for the identification and study of the species: Linnean Society of London and the Estuarine and Brackish-Water Sciences Association.
- Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., & Kumar, S. (2011). Mega5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution*, 28, 2731-2739.
- Vlasenko, R. P., Mezherin, S. V., Garbar, A. V., & Kotsuba, Y. (2011). Polyploid races, genetic structure and morphological features of earthworm *Aporrectodea rosea* (Savigny, 1826) (Oligochaeta, Lumbricidae) in Ukraine. *Comparative Cytogenetics*, 5, 91-103.
- Vsevolodova-Perel, T. S., & Bulatova, N. S. (2008). Polyploid races of earthworms (Lumbricidae, Oligochaeta) in the east European plain and Siberia. *Biology Bulletin*, 35, 385-388.

ANNEXE : TEXTES DES PUBLICATIONS

DNA Barcoding Reveals Cryptic Diversity in *Lumbricus terrestris* L., 1758 (Clitellata): Resurrection of *L. herculeus* (Savigny, 1826)

Samuel W. James^{1,3*}, David Porco^{2,3}, Thibaud Decaëns³, Benoit Richard³, Rodolphe Rougerie^{2,3}, Christer Erséus⁴

1 Biodiversity Institute, Kansas University, Lawrence, Kansas, United States of America, **2** Canadian Centre for DNA Barcoding, Biodiversity Institute of Ontario, Guelph, Canada, **3** Laboratoire d'Ecologie, UPRES-EA 1293 ECODIV, FED SCALE, Bâtiment IRESE A, UFR Sciences et Techniques, Université de Rouen, Mont Saint Aignan, France, **4** Department of Zoology, University of Gothenburg, Göteborg, Sweden

Abstract

The widely studied and invasive earthworm, *Lumbricus terrestris* L., 1758 has been the subject of nomenclatural debate for many years. However these disputes were not based on suspicions of heterogeneity, but rather on the descriptions and nomenclatural acts associated with the species name. Large numbers of DNA barcode sequences of the cytochrome oxidase I obtained for nominal *L. terrestris* and six congeneric species reveal that there are two distinct lineages within nominal *L. terrestris*. One of those lineages contains the Swedish population from which the name-bearing specimen of *L. terrestris* was obtained. The other contains the population from which the syntype series of *Eterion herculeum* Savigny, 1826 was collected. In both cases modern and old representatives yielded barcode sequences allowing us to clearly establish that these are two distinct species, as different from one another as any other pair of congeners in our data set. The two are morphologically indistinguishable, except by overlapping size-related characters. We have designated a new neotype for *L. terrestris*. The newly designated neotype and a syntype of *L. herculeus* yielded DNA adequate for sequencing part of the cytochrome oxidase I gene (COI). The sequence data make possible the objective determination of the identities of earthworms morphologically identical to *L. terrestris* and *L. herculeus*, regardless of body size and segment number. Past work on nominal *L. terrestris* could have been on either or both species, although *L. herculeus* has yet to be found outside of Europe.

Citation: James SW, Porco D, Decaëns T, Richard B, Rougerie R, et al. (2010) DNA Barcoding Reveals Cryptic Diversity in *Lumbricus terrestris* L., 1758 (Clitellata): Resurrection of *L. herculeus* (Savigny, 1826). PLoS ONE 5(12): e15629. doi:10.1371/journal.pone.0015629

Editor: Dirk Steinke, Biodiversity Institute of Ontario, University of Guelph, Canada

Received: August 26, 2010; **Accepted:** November 15, 2010; **Published:** December 29, 2010

Copyright: © 2010 James et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: Most of this research was supported through grants from Natural Science and Engineering Research Council of Canada (NSERC) and Genome Canada to Paul Hebert (Biodiversity Institute of Ontario). S. W. James was supported by United States National Science Foundation Grant DEB 0516439 and by a Marie Curie France Regions grant from the Région Haute Normandie. R. Rougerie and D. Porco were supported by post-doctoral fellowship grant from the Conseil Régional de Haute Normandie through the Grand Réseau de Recherche Sciences de l'Environnement et Gestion des Risques. T. Decaëns, B. Richard, R. Rougerie, D. Porco, and S. James were supported by the SClences Appliquées à l'Environnement research federation through the Functions and Determinants of Biodiversity program. C. Erséus was supported by ArtDatabanken (Swedish Taxonomy Initiative) and The Royal Society of Arts and Sciences in Göteborg. None of these funding agencies had any role in the conducting of the study or in the preparation of the manuscript. It is fair to say that until this work is published, most of them remain unaware of its existence. Paul Hebert, as director of the Canadian Centre for DNA Barcoding at the Biodiversity Institute of Ontario, made it a priority to direct the necessary resources toward the study, but the resources were ultimately derived from grants by NSERC and Genome Canada. Dr. Hebert cannot be cast in the role of a funding agency, but rather in the role of a laboratory director making decisions about how to use existing support.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: sjames@ku.edu

Introduction

Lumbricus terrestris L., 1758 occupies an important place in the nomenclature of earthworms, having been the first earthworm named [1], and has an important place in biological science and science education. It has been used in many studies of earthworm anatomy, behaviour, physiology and ecology, achieving the status of model organism long before this term came into common use in the last few decades. Virtually every student of biology in secondary or higher education systems of the Western world, as well as many places influenced by textbooks produced therein, has been presented the example of "*Lumbricus terrestris*" as an object of study. In some cases it may be doubted that *L. terrestris* was in fact on the dissection tray, but no matter; the name was used and biologists everywhere recall this as "the earthworm." Darwin [2] referred loosely to earthworms in this way, though it is likely that

one of the species whose activities he observed was *L. terrestris*. The species came into further prominence as an economic resource through the fish bait trade, notably in North America, where astonishing numbers are gathered from Canadian golf courses for domestic bait use and export to the United States of America [3]. Finally, *L. terrestris* is now considered an invasive species and has a prominent role in transforming soils and organic matter accumulations where it has invaded ecosystems previously devoid of earthworms, or replaced species with a comparable ecology [4,5]. It routinely reaches population densities capable of consuming the entire annual leaf fall of north temperate deciduous forests, which is far more than above-ground herbivores normally do except in massive pest outbreaks [4,6,7].

With all the scientific, educational and popular attention devoted to *L. terrestris*, it is rather surprising to find that we do not really know what it is. Some of the problems stemmed from

the brief original description and lack of a type specimen, but this was rectified in a detailed consideration of the nomenclatural history and identities of various earthworms cited as *L. terrestris*, with the designation of a neotype from the probable type locality in Uppsala, Sweden [8]. That neotype is now missing (The Natural History Museum, in litt.). Savigny [9], in describing *Enterion herculeum* Savigny, 1826 deposited a series of specimens which exists to this day. However, *E. herculeum* was later placed in the synonymy of *L. terrestris* [8,10]. Richard et al. [11] detected two genetic clusters within nominal *L. terrestris*, reopening the debate. In this paper we revisit the question of the identities of the earthworms known as *L. terrestris* and *L. herculeus*, with the application of molecular and morphological data. We designate a new neotype for *L. terrestris* and provide a DNA barcode record for a syntype of *L. herculeus*. The latter is from Savigny's specimens, which were automatically syntypes because no holotype was fixed by the author. This barcoded specimen is here designated as the lectotype of *L. herculeus*.

Materials and Methods

We examined 200 specimens of "*L. terrestris*" recently collected in Europe and North America, four specimens topotypic to the former neotype [8] of *L. terrestris* collected in Uppsala, Sweden in 1972; a syntype of *L. herculeus* and several specimens of five congeners: *L. castaneus* (Savigny, 1826); *L. centralis* Bouché, 1972; *L. festinus* Savigny, 1826; *L. friendi* Cognetti, 1904; and *L. rubellus* Hoffmeister, 1843 (Table 1). Morphological examinations were confined to nominal *L. terrestris* (including *L. herculeus*), including four specimens topotypic to the 1973 neotype of *L. terrestris*, six specimens from Parc du Chateau Brunoy, and five specimens collected in 2008 from Parc du Gally on the grounds of the Versailles Palace, the location of Savigny's material from the environs of Paris (M.B. Bouché, pers. comm., based on notes of Savigny). Tissue samples were obtained from three of the topotypic Uppsala *L. terrestris* collected in 1972, a syntype of *L. herculeus*, the five congeners, and the 198 recent specimens of "*L. terrestris*" one of which is the new neotype (GenBank HM388349; BOLD EW-ECO-0533) and topotypic to the former neotype of *L. terrestris*.

In total, 230 specimens from 6 species of *Lumbricus* were used for genetic examination of the divergence within the genus *Lumbricus* (Table 1). All these worms were processed for the campaign 'Barcoding Earthworms' (BCEW) at two different laboratories.

Samples processed at Canadian Centre for DNA Barcoding

Lysis of the tissues was carried out in 50µl volume of lysis buffer and proteinase K incubated at 56°C overnight. DNA extraction followed a standard automated protocol on 96-well glass fiber plates (Ivanova et al. 2006). The 5' region of COI used as a standard DNA barcode was amplified using M13 tailed primers LCO1490 and HCO2198 [12]. Failed samples from this first pass were amplified with a pair of internal primers combined with full length ones LepF1-MLepR1 and MLepF1-LepR1 [13]. A standard PCR reaction protocol was used for PCR amplifications and products were checked on a 2% E-gel 96Agarose (Invitrogen). Unpurified PCR amplicons were sequenced in both directions using M13 tails as primers. The sequencing reactions followed standard protocols of the Canadian Centre for DNA Barcoding (CCDB) [14], with products subsequently purified using Agencourt CleanSEQ protocol (Agencourt) and processed using BigDye version 3.1 on an ABI 3730 DNA Analyzer (Applied Biosystems).

The specimen from the type series of *Lumbricus herculeus* from the Savigny 1821 collection and the 1972 topotypic *L. terrestris*

Table 1. Specimens included in the study.

Species	Country	Region	N
<i>Lumbricus castaneus</i>	France	Seine Maritime	9
	Andorra	Santa Julia	1
			10
<i>Lumbricus centralis</i>	France	Provence-Alpes-Cote d'Azur	1
<i>Lumbricus festinus</i>	France	Seine Maritime	8
	France	Ile de France	1
			9
<i>Lumbricus friendi</i>	France	Midi-Pyrenees	1
<i>Lumbricus rubellus</i>	France	Haute Normandie	9
<i>Lumbricus terrestris</i>	Canada	Ontario	63
	Denmark	Jutland, Arhus	1
	France	Languedoc-Roussillon	1
	France	Ile de France	7
	France	Haute Normandie	41
	France	Bretagne	2
	Norway	Nordland	1
	Norway	Hordaland	1
	Sweden	Jämtland	1
	Sweden	Scania	2
	Sweden	Småland	2
	Sweden	Uppland	2
	Sweden	Värmland	1
	Sweden	Västerbotten	1
	Sweden	Västergötland	2
	United States	Ohio	7
	United States	Iowa	8
			144
<i>Lumbricus herculeus</i>	Denmark	Jutland, Arhus	1
	France	Seine Maritime	24
	France	Ile de France	22
	Sweden	Scania	9
			56

The text references to lineages L1 and L2 correspond to *L. terrestris* and *L. herculeus*, respectively.
doi:10.1371/journal.pone.0015629.t001

specimens were sampled for DNA (hereafter referred to as the museum specimens). The age and preservation of the specimens from which these tissues were sampled, however, demanded a different approach for extraction and amplification. Extraction was done manually with the Nucleospin tissue extraction Kits and PCR amplification was done with 6 pairs of primers in order to amplify overlapping fragments of about 160bp (Table 2). The same primers and the standard protocol of the CCDB [14] were used for the sequencing of those fragments.

Specimens processed at University of Gothenburg

Twenty-four specimens morphologically identified as *L. terrestris* were collected in Scandinavia (21 in Sweden, two in Denmark, and two in Norway), in 2008–2009. DNA was extracted from a tissue sample of each worm with the QIAGEN DNeasy® Blood & Tissue Kit, after which PCR reactions were performed using the COI primers LCO1490 (forward) and HCO2198 (reverse) [12],

Table 2. Primers used to obtain short overlapping barcode sequence fragments.

5' - 3'	
1st pair	
LCO1490_t1	TGTAAACGACGCCAGTGGTCAACAAATCATAAGATATTGG
EWL1-1R	CGCCAATRAAGACTGGTATYAC
2nd pair	
EWL1-2F	TTATACAATACAATCGTTACTGC
EWL1-2R	GAAACTARGAGAATAAGRAGAGGG
3rd pair	
EWL1-3F	CATAAGATTTGACTCTCTCC
EWL1-3R	AGRATAGAGGAYGCACCTGC
4th pair	
EWL1-4F	CTTGCCAGRAATCTCGCCCA
EWL1-4R	ACAAAYAGAGGGATTCGYCTAG
5th pair	
EWL1-5F	TCCCTCCATTTRGCAGGKGC
EWL1-5R	GTTARGAGTATTGTGATTGCYCK
6th pair	
EWL1-6F	AATYACAGTAGTYCTCTCTCTCT
HCO2198_t1	CAGGAACAGCTATGACTAACTTCAGGGTGACCAAAAAATCA

doi:10.1371/journal.pone.0015629.t002

the reverse primer sometimes replaced by COI?E [15]; all following standard protocols. PCR products were purified using an Omega E.Z.N.A. cycle-pure kit, and sent to Macrogen, South Korea, for sequencing.

Sequence analysis

Sequences were assembled with Sequencer 4.5 (GeneCode Corporation, Ann Arbor, MI, USA) and aligned by eye using BIOEDIT version 7.0.5.3 [16]; we observed no indels in this coding region of the mitochondrial genome and therefore all base positions were aligned with confidence in positional homology. Distance analyses were conducted with MEGA4 [17] using a Neighbor-Joining [18] algorithm and distances corrected with the Kimura-2 parameter model [19]. The robustness of nodes was evaluated through bootstrap re-analysis of 1000 pseudoreplicates.

Results

Barcode data

230 full length barcodes were obtained ranging between 508 and 658 bp. From the syntype of *L. herculeus* we obtained 5 of the 6 fragments with 4 consecutive overlapping ones producing a continuous sequence of 480 bp. Only one of the 1972 Uppsala museum specimens of *L. terrestris* yielded a sequence, a 144 bp fragment. A complete COI barcode was obtained from the specimen collected in Uppsala in 2009 (the replacement neotype, GenBank HM388349; BOLD EW-ECO-0533). Sequences are publicly available on BOLD [20]; <http://www.barcodinglife.org>) within the project LTERH and in GenBank (Table S1).

The mean intraspecific and interspecific variations for COI in the genus *Lumbricus* are 1.24% and 19.81%, respectively, except for nominal *L. terrestris* which exhibits the highest intraspecific value in the dataset at 8.93% and a range of 0% to 19%. These extreme values are due to the presence of two highly divergent

Table 3. Kimura 2-parameter mean genetic distances (%) between and within *Lumbricus* spp.

<i>Lumbricus castaneus</i>	0.45					
<i>Lumbricus centralis</i>	22.99	-				
<i>Lumbricus festivus</i>	23.46	18.55	0.36			
<i>Lumbricus friendi</i>	22.54	14.80	17.54	-		
<i>Lumbricus rubellus</i>	21.04	21.06	21.01	17.72	0.57	
<i>Lumbricus terrestris</i>	23.74	18.23	19.82	19.90	18.45	3.37
<i>Lumbricus herculeus</i>	21.62	18.60	17.85	19.43	20.29	17.50

The principal diagonal has intraspecific distances; all others are interspecific.
doi:10.1371/journal.pone.0015629.t003

groups of individuals within the nominal species. Separating the two lineages, divergence within *L. terrestris* s.s is 3.37%, and within *L. herculeus* it is 1.54% (Table 3). Comparing the distribution of the *L. terrestris* intraspecific divergences to what is exhibited among the other species in the genus (Figure 1) we see clearly that the divergence between the two groups found in nominal *L. terrestris* is comparable to the distances among other species of the genus. The mean interspecific divergence between the two *L. terrestris* lineages is 17.5%.

An unrooted Neighbor-Joining (NJ) tree of *Lumbricus* barcode sequences placed all nominal *L. terrestris* in two well-supported and

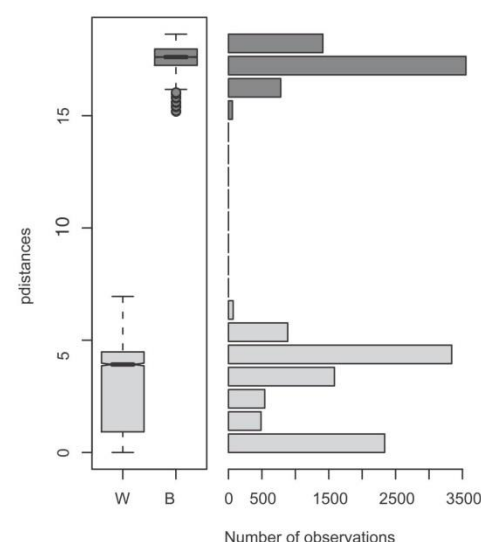


Figure 1. Boxplots of intraspecific (W; gray bars) and interspecific (B; black bars) genetic distances (K2P) for *Lumbricus* species with corrected taxonomy separating *L. terrestris* from *L. herculeus*. Each boxplot represents: the discarded outliers (external dots), the smallest and largest observations (external bars), the lower and upper quartiles (limits of the box) and the median (within-box black line). The boxplots are notched and indicate that medians differ if the notches do not overlap.
doi:10.1371/journal.pone.0015629.g001

divergent clusters (Figure 2A). All other species represented by more than one individual also fell into well-supported clusters.

In a second step the short barcode sequences from the 1972 museum specimen of *L. terrestris* and from the syntype of *L. herculeus* were introduced in the dataset. Although both the sequences were short, they allowed an accurate assignment of each specimen to one of the *L. terrestris* lineages (Figure 2B). The discriminating power of mini-barcodes is established [13] and here we used these short sequences in favorable conditions as the divergence between the two lineages of *L. terrestris* is very high (17.5%). Thus each of the type-related specimens was successfully assigned to a lineage in the NJ analysis, one to the *L. terrestris* cluster and one to the *L. herculeus* cluster (Figure 2B). From this point forward in the results and discussion, we use the two species names in the restricted sense supported by these data, unless enclosed in quotation marks.

In a separate NJ analysis (tree not shown) the cytochrome oxidase I gene barcode region of the complete “*L. terrestris*” mitochondrial genome sequence [21; GenBank NC001673.1] fell within the *L. terrestris* cluster.

Morphology

Morphological examination of the fresh specimens, each registered in such manner that a COI barcode sequence can be matched to an individual worm, indicates that there are differences in segment number (125 vs 143, *L. herculeus* and *L. terrestris* respectively), body mass (1.7g vs 3.2g), and body length (107mm vs. 148 mm) between the two groups (Figures 3A–C). These differences are not as clear-cut as the genetic differences, there being overlap in the distributions of the three measurements. Put simply, small *L. terrestris* can be smaller than large *L. herculeus*, but they have strongly divergent COI sequences. The two species are illustrated in Figure 4.

The specimen of *L. terrestris* in the vial labeled as neotype (Natural History Museum, London; Register No. 1973.1.1) is shorter by 12 mm and has 6 fewer segments than the neotype described by Sims [8]. The other specimens of the same series from Uppsala are all either longer or shorter, or have more or fewer segments than the specimen described in Sims (1973) (E. Sherlock, *in litt.*), so the Sims [8] neotype is presumed lost.

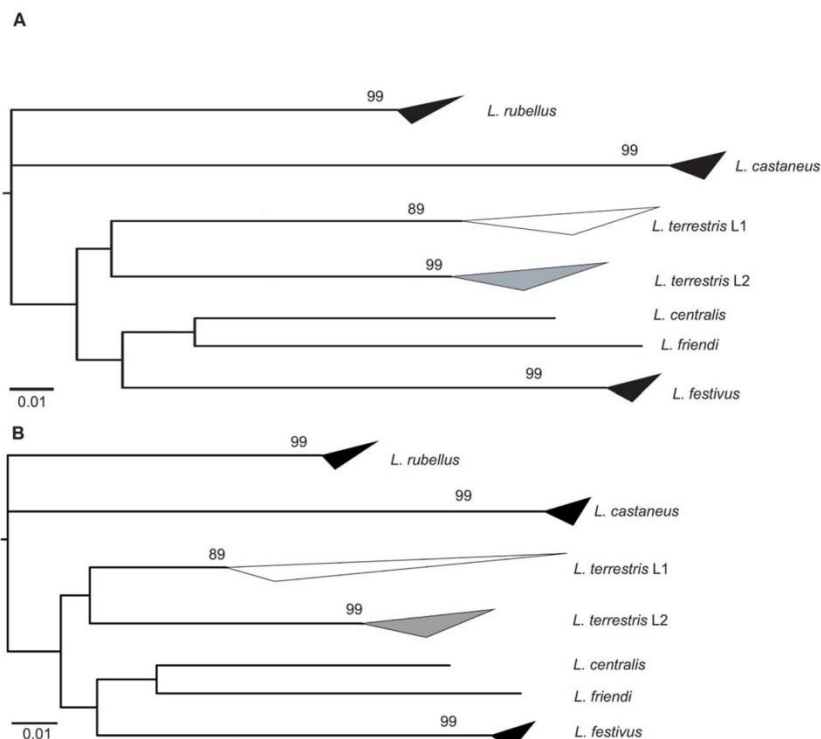


Figure 2. Neighbor joining trees (K2P) for 6 species of the genus *Lumbricus*, based on the COI 5' 'barcoding fragment'; bootstrap support values for each cluster shown on its subtending branch. The upper and lower sides of each triangle represent respectively the maximum and minimum genetic distances within a species. A. Without type or museum material of the two *L. terrestris* lineages. B. Reconstruction with type specimens and museum material. Higher genetic variation of *L. terrestris* L1 in B. is due to the short sequence (144bp) of the 1972 museum specimen.
doi:10.1371/journal.pone.0015629.g002

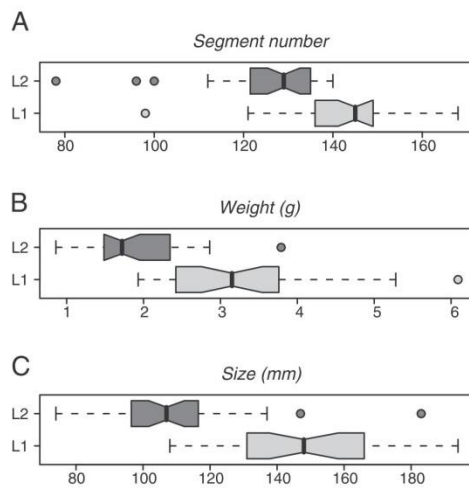


Figure 3. Boxplots of morphological features for *L. terrestris* (L1; N=30) and *L. herculeus* (L2; N=36): A. Segment number; B. Body weight; C. Body length. Each boxplot represents: the discarded outliers (external dots), the smallest and largest observations (external bars), the lower and upper quartiles (limits of the box) and the median (within-box black line). The boxplots are notched and indicate that medians differ if the notches do not overlap.
doi:10.1371/journal.pone.0015629.g003

Geographical distributions

In the Scandinavian material, which covers latitudes from 55 to 68°N, *L. herculeus* was found at seven localities all concentrated in the western part of Sweden's southernmost province, Scania, and at one Danish locality (at Århus in Jutland), the northernmost site being near Båstad, Sweden (56°23'N), while *L. terrestris* was found scattered over Sweden (to Västerbotten in the north, at about 64°N) and also above the Polar Circle on the Norwegian West coast (Nordland, at 68.7°N). All our North American records belonged to *L. terrestris*, while French records were mixed. Two



Figure 4. *Lumbricus herculeus* (left) and *Lumbricus terrestris* (right); specimens depicted are respectively smaller and larger than average for their respective species.
doi:10.1371/journal.pone.0015629.g004

Swedish sites, the Danish site and three French sites had both species, while others had only one.

Neotypes

We turn now to the means of defining these two very similar species. As we have indicated, size is the only morphological difference, and it is not reliable in the overlapping sections of the size distributions. Above we noted that the neotype of *L. terrestris* collected at Uppsala in 1972 is no longer in the Natural History Museum (London). Neotypes can be replaced "...when no name-bearing type specimen (i.e. holotype, lectotype, syntype or prior neotype) is believed to be extant and an author considers that a name-bearing type is necessary to define the nominal taxon objectively," [22, Art. 75.1].

If the designation of a new neotype would only replace one morphologically undefinitive specimen with another, then there is little or no justification for the designation. However, we have successfully isolated and sequenced some DNA from a contemporary member of the same population (GenBank HQ024541) as the missing neotype. We also obtained sequence data from specimen CE6377M collected at the same location (GenBank HM388349; BOLD EW-ECO-0533). Then we unequivocally clustered these resulting sequences with those of numerous other individuals which by size are generally identifiable as *L. terrestris*. We also demonstrated a substantial genetic difference between the *L. terrestris* cluster and the related and cryptic congener *L. herculeus*.

Designation of a neotype must demonstrate exceptional need. We believe this to be the case. All of the points raised by Sims [8] regarding stability of nomenclature are still valid today. As Sims [8] indicated in his arguments for the designation of a neotype, *L. terrestris* occupies not only an important historical position in the nomenclature of earthworms as the first earthworm described and type species of *Lumbricus* and the Lumbricidae, it has also been a model organism for research and education in Biology. It is of considerable importance for the stability of nomenclature that it be possible to determine the identity of an earthworm matching the physical characters of *L. terrestris* and *L. herculeus*. Our results indicate that morphological examinations are not sufficient for the identification of these species, but that DNA sequences are. We have established the utility of the barcode fragment of the COI gene [23,24] for this purpose but do not confine the method to this gene. The proposed neotype now has a COI sequence tag which can unequivocally be used to characterize and recognize the taxon *L. terrestris* in a way hitherto impossible. To date there are no known *Lumbricus* species with a sufficiently similar sequence to cause any confusion in DNA-based identification of species within this genus, let alone in the discrimination of *L. terrestris* and the morphologically virtually identical *L. herculeus*.

To satisfy the provisions of ICZN [22] Art. 75.3.1-7, the qualifying conditions for validly designating a neotype, we offer the following points:

1. Our designation of a neotype is necessary to clarify the taxonomic status of *L. terrestris*, in order that it can be distinguished from *L. herculeus*.
2. The characters differentiating *L. terrestris* from *L. herculeus* are differences in aligned, positionally homologous COI gene DNA sequence bases.
3. The partial COI sequence derived from the neotype (GenBank HM388349; BOLD EW-ECO-0533), and the physical description in Sims [8], which is identical in all but measurements to the designated neotype, and the measurements given here below are sufficient to ensure correct recognition of the specimen designated.

4. The prior neotype is missing from the collection of the Natural History Museum (London) and the staff made a thorough search of the premises, but failed to locate the specimen.
5. Other than the particular body size and segment number measurements, the designated specimen is anatomically consistent with the prior neotype.
6. One of the collectors of the prior neotype (Tryggve Persson) was consulted on the collection event location of 13 October 1972, and the new neotype was taken from the same locality in the Botanical Gardens at Uppsala, Sweden. The location was chosen because that was the location satisfying this condition (proximity to original collection site of Linnaeus) for the prior neotype. We follow this established precedent, which is in any case the valid type locality following the designation of the prior neotype.
7. The new neotype has been deposited in the Swedish Museum of Zoology, with catalogue number given below.

Our choice of neotype specimen is not one of the specimens collected at the same time as the prior neotype, even though we were able to obtain a short (144 bp) sequence of the COI gene from one of three attempted. Therefore regarding "Recommendation 75A. Choice of neotypes," [22] we designate a new specimen preserved in a manner that allows extraction of high-quality DNA. Thus the definition and delimitation of the taxon need not be based only on the short "mini-barcode" obtained from the 1972 specimen, because future researchers will be able to use small samples of the newly designated neotype for further genetic data. In short, we maintain that for purposes of molecular definition and delimitation of the taxon, the 1972 material is in poor condition.

The following synonymy is modified from: <http://earthworms.elte.hu/Hungary/lumbricus.htm> by removal of the references to *L. herculeus*, which was included by Cs. Csuzdi in the synonymy of *L. terrestris*. In all cases where the authors in the following synonymy did not make any distinction between *L. terrestris* and *L. herculeus*, we indicate that the *L. terrestris* referred to may be partly attributable to another species. However we make no claims about the identities of any of the other junior synonyms.

***Lumbricus terrestris* Linnaeus, 1758**

- Lumbricus terrestris*: Linnaeus, 1758 Systema Naturae, 10: 647.
Lumbricus agricola Hoffmeister, 1842 Verm. Lumbric., p. 24.
Lumbricus infelix Kinberg, 1867 Öfv. Akad. Förh., 23: 98.
Lumbricus americanus Perrier, 1872 N. Arch. Mus. Paris, 8: 44.
Lumbricus terrestris (part.): Orley 1885 Érték. term. tud. köréből, 15: 30.
Lumbricus studei Ribaucourt, 1896 Rev. suisse Zool., 4: 5.
Lumbricus terrestris (part.): Michaelsen 1900 Das Tierreich, 10: 511.
Lumbricus terrestris (part.): Szűts 1909 Állattani Közlemények, 8: 142.
Lumbricus terrestris (part.): Zicsi 1959 Acta zool. hung., 5: 433.
Lumbricus terrestris (part.): Zicsi 1968 Opusc. Zool. Budapest, 8: 130.
Lumbricus terrestris (part.): Sims 1973. Bull. Zool. Nom. 30:32.
Lumbricus terrestris (part.): Zicsi 1982 Acta zool. hung., 28: 443.
Lumbricus terrestris (part.): Easton 1983 Earthworm Ecology, p. 482.
Lumbricus terrestris (part.): Zicsi 1991 Opusc. Zool. Budapest, 24: 173.
Lumbricus terrestris (part.): Mršić 1991 Acad. Sci. Art. Slov. (Hist. Nat.), 31: 481.
Lumbricus terrestris terrestris (part.): Qju & Bouché 2000 Doc. pedozool. integrol, 4: 192.

Neotype. Clitellate specimen. Sweden, Uppland, Uppsala, Uppsala Botanical Garden, walkway adjacent to lawn, 59°51.07'N, 017°37.65'E. 04 June 2009. Collector: Christer Erséus. Swedish Museum of Natural History (Stockholm) catalogue number SMNH Type- 8035.

Other material: 4 clitellates, Sweden, Uppland, Uppsala, Uppsala Botanical Garden., 13 October 1972. B. Axelsson, U. Lohm, T. Persson collectors; 1 clitellate, France, Parc du Château de Versailles, Ile de France 48° 43'43.52" N, 2° 06'56.74" 1 November 2008, M. Hedde collector; 2 clitellates, France, Essone, Ile de France, Parc du Château de Brunoy, 48° 48'41.61" N, 2° 29'38.25" 11 November 2008, M. Hedde and T. Decaëns collectors.

Description of neotype and other material: The neotype is in two fragments, the anterior consisting of 59 segments and the posterior of 95 segments, for a total of 154 segments, with total length (strongly contracted) 89 mm. There are herniations on the right side at 15/16 and 28/29, and slight abrasions to the left side of segments 46–52. This damage was present on the specimen at the time of collection from a walkway. The clitellum is at 32–37, the tubercula pubertatis at 33–36, and there are genital markings surrounding enlarged AB genital setae on segments 31–37 and right side of segment 38. The first muscular septum is always 19/20, which is displaced about a half-segment length posteriorly to lie close to septum 20/21.

Examining the 1972 and the French material, the typhlosolar convolutions are very distinctive. From the beginning of the typhlosole in XXII it has lateral flaps oriented vertically. The ventral edges of each flap bifurcate and fuse with the split sections of the flaps anterior and posterior to the flap in question. The fused parts form a short bar extending across the center ventral face of the typhlosole to meet the lateral flaps of the other side, which are also split and fused as just described. The short bars take the appearance of the rungs of a ladder whose lengthwise components are made of the fusion points of the lateral flaps. This pattern originates in segments 23–24 and gradually fades out over two or three segments between 47 and 52, after which the typhlosole has a smooth surface and a circular to oval cross-section. The typhlosole ends abruptly over one or two segments anywhere from 99 to 117, though most commonly in 100–108.

Remarks: The neotype and the 1972 specimens agreed in all other particulars with the description in Sims [8], except for the muscularity of septa behind the gizzard. It is possible that Sims [8] mistook the two closely-spaced septa for one and misjudged the count because of the displacement. To be certain, we counted anteriorly from segment 25, where septum 24/25 is in line with its external segment boundary. No morphological differences from the nominal *L. terrestris* as traditionally defined were detected. The sequences given (Genbank HQ024541, HM388349) are the only present means of objective determination of the species, in relation to the slightly smaller *L. herculeus*.

***Lumbricus herculeus* Savigny, 1826 (synonymy modified from <http://earthworms.elte.hu/Hungary/lumbricus.htm>)**

- Enterion herculeum*: Savigny, 1826 Mem. Acad. Sci. Inst. Fr., 5: 180.
Lumbricus terrestris (part.): Orley 1885 Érték. term. tud. köréből, 15: 30.
Lumbricus terrestris (part.): Michaelsen 1900 Das Tierreich, 10: 511.
Lumbricus terrestris (part.): Szűts 1909 Állattani Közlemények, 8: 142.
Lumbricus terrestris (part.): Pop 1943 Ann. Hist.-Nat. Mus. Hung., 36: 19.
Lumbricus terrestris (part.): Zicsi 1959 Acta zool. hung., 5: 433.

Lumbricus terrestris (part.): Bouché 1972 Inst. Nat. Rech. Agron. p. 352.

Lumbricus terrestris (part.): Sims 1973. Bull. Zool. Nom. 30:32.

Lumbricus terrestris (part.): Zicsi 1968 Opusc. Zool. Budapest, 8: 130.

Lumbricus terrestris (part.): Zicsi 1982 Acta zool. hung., 28: 443.

Lumbricus terrestris (part.): Easton 1983 Earthworm Ecology, p. 482.

Lumbricus terrestris (part.): Zicsi 1991 Opusc. Zool. Budapest, 24: 173.

Lumbricus terrestris (part.): Mršić 1991 Acad. Sci. Art. Slov. (Hist. Nat.), 31: 481.

Lumbricus terrestris terrestris (part.): Qiu & Bouché 2000 Doc. pedozool. integrol, 4: 192.

Lectotype. Clitellate specimen. In the general area of Paris France. 1821. Collector J.C. Savigny. Paris, Musée Nationale d'histoire Naturelle. Label data: *Enterion herculeum* Savigny, Paris 1821.

Other material: 4 clitellates, France, Parc du Château de Versailles, Ile de France 48° 43'43.52" N, 2° 06'56.74" 1 November 2008, M. Hedde collector; 4 clitellates, France, Essone, Ile de France, Parc du Château de Brunoy, 48° 48'41.61" N, 2° 29'38.25" 11 November 2008, M. Hedde and T. Decaëns collectors.

Lectotype: Three fragments in one vial, consisting of the first 22 segments, segments 23–50, and segments 51 to 145. There is a partial cut in the 5th segment and a small knotted thread is inserted in the 9th segment, in the manner of those used for tagging larger animal specimens. Total length, 114 mm. Genital markings are present on AB of 9–11, left 26, 31–38, and left 39. The clitellum is saddle-shaped on 32–37, with tubercula pubertatis on 33–36. No pigmentation is visible. No dissection was performed on this delicate specimen. The 480 bp DNA barcode sequence from this specimen is GenBank HQ024540.

The other material examined had no differences from the *L. terrestris* specimens other than measurements (Figs 3A–C., which include additional specimens to those examined in detail). Septal muscularity, and typhlosome morphology and termination were all indistinguishable other than a slight tendency of *L. herculeus* specimens to have fewer atyphlosolate segments than *L. terrestris*. However these numbers overlapped, like the other quantitative measures.

Savigny did not designate any type specimen(s), so here we defined this species in the most simple and direct manner possible at this time. The somewhat softened clitellate specimen was not subjected to further examination, and is now designated as the lectotype of *L. herculeus*. The description of *L. terrestris* by Sims [8] serves as a source of morphological details. Our Figures 3A–C, the observations on the specimens examined for the above descriptions, and the descriptive data of Bouché and Beugnot [25] give the extent of morphological differences between this species and *L. terrestris*.

Discussion

These results indicate that “*L. terrestris*” as traditionally identified is composed of two species that have not been discriminated in the literature. Morphological examination can only make reliable distinctions between average or larger *L. terrestris* and average or smaller *L. herculeus*. Bouché and Beugnot [25] reached the same conclusion regarding what they considered as two sympatric populations of “*L. herculeus*” with nomenclature following Bouché’s 1970 advocacy of that name over “*L. terrestris*.” Our collection includes 18 individuals from the Parc du Chateau de Brunoy location sampled by Bouché and Beugnot (1972), of which 16 fall in *L. herculeus* and the other 2 in *L. terrestris*. The segment number

and size variations Bouché and Beugnot [25] reported are the same as we observed between *L. herculeus* and *L. terrestris*. The two species are best distinguished by molecular data, which will work on all sizes and life stages of the individuals, from egg capsules to adults. The sequences from 1972 and 2009 specimens topotypic to the missing Sims [8] *L. terrestris* neotype were in the *L. terrestris* cluster.

In a similar situation, that of earthworms questionably separable by size, body coloration, and some genital papillae, Chang et al. [26] found that molecular data strongly supported separation of two species from nominal *Amyntas wulinensis* Tsai, Shen and Tsai, 2001. In the *A. wulinensis* case, the size, color, and papillae characters giving the initial indications of lineage diversity are traditionally not considered reliable in Asian earthworm taxonomy. *Eisenia fetida* (Savigny, 1826) and *E. andrei* Bouché, 1972 are only sometimes separable by color, but are two genetically distinct and isolated species [27].

Savigny [9] did not attribute an author to *Enterion herculeum* and did not indicate that it was a new species, which was not required in his time. Nor did he expressly indicate that *Enterion herculeum* is a replacement name for some other nominal species group [25, Art. 72.7]. Therefore the two names are not objective synonyms, and *L. herculeus* is a junior subjective synonym of *L. terrestris*. In any case the brief description by Savigny is a valid indication [25, Art. 12.2.] of the identity of the worm and intent of the author. The effect of our work is to restore a junior subjective synonym (*L. herculeus*) to species status. In consideration of the molecular data and nomenclatural procedure, we remove *L. herculeus* from the synonymy of *L. terrestris*, and thereby restrict *L. terrestris* to the cluster whose members are larger.

Apparently, Savigny was describing a new species, named *Enterion herculeum*, among 21 other names in his document. Had his specimens been *L. terrestris* s.s., the name *herculeus* would definitely pass into synonymy. By happenstance, he collected and applied a name to what is now defined as a separate species. Where Linnaeus worked in Sweden he saw earthworms on the surface at night [1: p. 648: “*ascendit noctu*”]. Sweden has both species, with *L. herculeus* only found in Scania so far, but regardless of which species Linnaeus saw, *L. terrestris* was defined by the Sims [8] neotype and is now defined by its replacement. On the other side of the Atlantic, the brisk trade in fish bait and classroom specimens is so far known to consist only of *L. terrestris*. North American investigators may rely on the lack of records of *L. herculeus*, but do so at their peril. We would not be surprised to find *L. herculeus* in North America or other continents where “*L. terrestris*” is known to occur.

This revision introduces doubt about the true identity of the species involved in any publication on “*L. terrestris*”, even if vouchers were deposited. Larger *L. terrestris* can be fairly certainly identified, as can smaller *L. herculeus*. Otherwise, old vouchers fixed in un-buffered formaldehyde solutions may not yield usable DNA, and therefore may not be identifiable. However, the worm used in Boore and Brown [21] for a complete mitochondrial genome sequence (GenBank NC001673.1) was apparently *L. terrestris*.

At this point it is an open question whether or not research on these two highly similar species has been tainted by the taxonomic confusion of the last 200 years. Are there conflicting results from similar studies, which could be resolved by establishing the true identities of the earthworms involved? Here we do not speak of the instances where careless study of “the earthworm, *Lumbricus terrestris*” actually referred to some other species, but to those in which a perfectly honest error was made, because taxonomists had no access to the types of data necessary to make the distinctions we are making here, and are now easily obtained. The two species

seldom co-occur in northern France, which could be due to competitive exclusion in various habitats that favor one or the other species. Alternatively they could be different enough to have distinct habitat preferences, and seldom come into competition. All northern European populations have been established by human-aided and natural dispersal since the retreat of the last European ice sheet. Thus we are not speaking of natural allopatric distributions but of a combination of accidents of arrival and competitive exclusion [28].

The obvious consequence of the revision is that any future identification of *Lumbricus* species closely resembling *L. terrestris* and *L. herculeus* should be accomplished in part by comparing DNA sequences including the COI barcode region. The genetic “gap” between the two is large and there are no known intermediate populations, so the results should be very clear.

Supporting Information

Table S1
(DOC)

References

1. Linnaeus C (1758) *Systema Naturae* (10th edition). 1:1–824. Salvi: Holmiae. 824 p.
2. Darwin C (1881) *The Formation of Vegetable Mould through the Action of Worms, with Observations on their Habits*. London: Murray. 326 p.
3. Tomlin AD (1983) The earthworm bait market in North America. In: Satchell JE, ed. *Earthworm Ecology*. London: Chapman and Hall. pp 331–338.
4. Hale CM, Frelich LE, Reich PB, Pastor J (2005) Effects of European earthworm invasion on soil characteristics in Northern Hardwood Forests of Minnesota, USA. *Ecosyst* 8: 911–927.
5. Hendrix PF, Callahan MA, Jr., Drake J, Huang C-Y, James SW, et al. (2008) Pandora's box contained bait: the global problem of introduced earthworms. *Ann Rev Ecol Syst* 39: 593–613.
6. Landsber J, Ohmart C (1989) Levels of insect defoliation in forests: Patterns and concepts. *Trends Ecol Evol* 4: 96–100.
7. Muzika RM, Liebhold AM (2001) Effects of gypsy moth defoliation in oak-pine forests in the Northeastern United States. In: Liebhold AM, McManus ML, Otvos IS, Fosbroke SLC, eds. *Proceedings: Integrated management and dynamics of forest defoliating insects*. Gen. Tech. Rep. NE-277. Newtown Square, PA: U.S. Department of Agriculture, Forest Service, Northeastern Research Station. pp 117–123.
8. Sims RW (1973) *Lumbricus terrestris* Linnaeus 1758 (Annelida, Oligochaeta): designation of a neotype in accordance with accustomed usage. Problems rising from the misidentification of the species by Savigny (1822 & 1826). *Bull Zool Nom* 30: 27–33.
9. Savigny JC In: Cuvier J, ed (1826) *Analyses de travaux de l'Académie Royale des Sciences pendant l'année 1821, partie physique*. Zoologie. *Mém Acad Sci Inst France (Hist)* 5: 176–184.
10. Michaelsen W (1900) Oligochaeta. In: Spengler JW, ed. *Das Tierreich*. Vol. 10. Berlin: Friedländer.
11. Richard B, Decaens T, Rougerie R, James SW, Porco D, et al. (2009) Re-integrating earthworm juveniles into soil biodiversity studies: species identification through DNA barcoding. *Mol Ecol Res* 10: 606–614.
12. Folmer O, Black M, Hoch W, Lutz R, Vrijenhoek R (1994) DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Mol Mar Biol Biotechnol* 3: 294–299.
13. Hajibabaei M, Smith MA, Janzen FH, Rodriguez JJ, Whitfield JB, et al. (2006) A minimalist barcode can identify a specimen whose DNA is degraded. *Mol Ecol Notes* 6: 959–964.

Acknowledgments

Paul Hebert of the Biodiversity Institute of Ontario and the Canadian Centre for DNA Barcoding allocated the resources necessary to complete the majority of the research. Isabelle Meunier helped design the mini-barcode primers. The following people assisted in field or laboratory work related to this paper: Anna Ansebo, Thomas Cedhagen, Thomas Dahlgren, Pierre De Wit, Arne Larsson, Emelie Lindquist, Maria Lindström, and Anna-Stina Ögren. Jimmy Cassone and Emma Sherlock provided loans from the Paris and London museums, respectively. We dedicate this paper to our friend and colleague, Marcel Bouché, un homme avec un bon nez pour les vers (a man with a good nose for worms).

Author Contributions

Conceived and designed the experiments: SWJ DP RR TD BR CE. Performed the experiments: SWJ DP RR TD BR CE. Analyzed the data: SWJ DP TD BR CE. Contributed reagents/materials/analysis tools: SWJ DP BR CE. Wrote the paper: SWJ TD DP.

14. Hajibabaei M, deWaard JR, Ivanova NV, Ratnasingham S, Dooh RT, et al. (2005) Critical Factors for assembling a high volume of DNA barcodes. *Phil Trans R Soc B* 360: 1959–1967.
15. Bely AE, Wray GA (2004) Molecular phylogeny of naiddid worms (Annelida: Clitellata) based on cytochrome c oxidase I. *Mol Phylogenet Evol* 30: 50–63.
16. Hall TA (1999) BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucl Acids Symp Ser* 41: 95–98.
17. Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol* 24: 1596–1599.
18. Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4: 406–425.
19. Kimura M (1980) A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* 16: 111–120.
20. Ratnasingham S, Hebert PDN (2007) BOLD: The Barcode of Life Data System (www.barcodinglife.org). *Mol Ecol Notes* 7: 355–364.
21. Boore JL, Brown WM (1995) Complete sequence of the mitochondrial DNA of the annelid worm *Lumbricus terrestris*. *Genetics* 141: 305–319.
22. International Commission on Zoological Nomenclature (1999) *International Code of Zoological Nomenclature*. London: The International Trust for Zoological Nomenclature. 306 p.
23. Hebert PDN, Cywinski A, Ball SL, deWaard JR (2003) Biological identifications through DNA bar-codes. *Proc R Soc B* 270: 313–321.
24. Hebert PDN, Ratnasingham S, deWaard JR (2003) Barcoding animal life: cytochrome c oxidase subunit I divergences among closely related species. *Proc R Soc B* 270(Suppl): S96–S99.
25. Bouché MB, Beugnot M (1972) La complexité taxonomique de *Lumbricus herculeus* illustrée par les caractéristiques de populations de stations de la R.C. P. 40. *Rev Ecol Biol Sol* 9: 697–704.
26. Chang C-H, Lin Y-H, Chen I-H, Chuang S-C, Chen J-H (2007) Taxonomic reevaluation of the Taiwanese montane earthworm *Amyntas uulimensis* Tsai, Shen and Tsai, 2001 (Oligochaeta: Megascotocidae): polytypic species or species complex? *Org Divers Evol* 7: 231–240.
27. Pérez-Losada M, Eiroa J, Mato S, Domínguez J (2005) Phylogenetic species delimitation of the earthworms *Eisenia fetida* (Savigny, 1826) and *Eisenia andrei* Bouché, 1972 (Oligochaeta, Lumbricidae) based on mitochondrial and nuclear DNA sequences. *Pedobiologia* 49: 317–324.
28. Decaens T, Margerie P, Aubert M, Hedde M, Bureau F (2008) Assembly rules within earthworm communities in North-Western France- a regional analysis. *Appl Soil Ecol* 39: 321–335.

**MOLECULAR ECOLOGY
RESOURCES**

Molecular Ecology Resources (2009)

doi: 10.1111/j.1755-0998.2009.02822.x

DNA BARCODING

Re-integrating earthworm juveniles into soil biodiversity studies: species identification through DNA barcoding

B. RICHARD,* T. DECAËNS,* R. ROUGERIE,† S. W. JAMES,‡ D. PORCO† and P. D. N. HEBERT†

*Laboratoire d'Ecologie, EA 1293 ECODEV, FED SCALE, Bâtiment IRESE A, UFR Sciences et Techniques, Université de Rouen, F-76821 Mont Saint Aignan cedex, France, †Canadian Centre for DNA Barcoding, Biodiversity Institute of Ontario, 50 Stone Road East, Guelph, ON, N1G 2W1 Canada, ‡Natural History Museum and Biodiversity Institute, Kansas University, 1345 Jayhawk Blvd., Lawrence, KS 66045, USA

Abstract

Species identification of earthworms is usually achieved by careful observation of morphological features, often sexual characters only present in adult specimens. Consequently, juveniles or cocoons are often impossible to identify, creating a possible bias in studies that aim to document species richness and abundance. DNA barcoding, the use of a short standardized DNA fragment for species identification, is a promising approach for species discrimination. When a reference library is available, DNA-based identification is possible for all life stages. In this study, we show that DNA barcoding is an unrivaled tool for high volume identification of juvenile earthworms. To illustrate this advance, we generated DNA barcodes for specimens of *Lumbricus* collected from three temperate grasslands in western France. The analysis of genetic distances between individuals shows that juvenile sequences unequivocally match DNA barcode clusters of previously identified adult specimens, demonstrating the potential of DNA barcoding to provide exhaustive specimen identification for soil ecological research.

Keywords: DNA barcoding, earthworms, juveniles, species identifications

Received 13 September 2009; revision received 15 November 2009; accepted 23 November 2009

Introduction

Studies on the biology and ecology of earthworms often depend on species diagnoses. Species identifications in this group require minute examinations of external and/or internal morphology of adults, necessitating the involvement of expert taxonomists. In many cases, identifications are complicated by the lack of stable, easily scored diagnostic characters or by environmentally induced variability in morphological features. In addition, many characters important in species diagnosis involve the position and structure of the clitellum and the associated tubercular pubertatis (Bouché 1972; Sims & Gerard 1999), characters, which are only observable in sexually mature specimens. As a result of their lack of diagnostic characters, the identification of juveniles of closely related species (e.g. members of the genus *Lumbricus*) is impossible in most cases. As a result, taxonomists can only

provide generic identifications for juveniles, hampering soil studies for species richness evaluation. Attempts to characterize earthworm species based on electrophoresis analysis – as developed for instance by Bogh (1992) – revealed higher resolution for species identification in samples containing many immature stages or incomplete body fragments, but the method remained largely ignored by soil biologists and has never been developed as a routine, which could be used to process the high number of individuals that an ecological study often comprises.

More recently, DNA barcoding has emerged as a promising standardized approach for rapid species identifications in taxonomically complex groups. It uses a 658 bp fragment of the mitochondrial gene cytochrome-c oxidase I (COI) as a standard DNA tag for species discrimination and identification in the animal kingdom (Hebert *et al.* 2003). The effectiveness of DNA barcodes in species identification has been shown in varied taxonomic groups (Hebert *et al.* 2004a,b; Ward *et al.* 2005; Smith *et al.* 2007; Borisenko *et al.* 2008) and they are increasingly being used in species descriptions as well (Decaëns & Rougerie 2008; Martinez *et al.* 2008; Vaglia *et al.* 2008;

Correspondence: Thibaud Decaëns, Tel: 33 + (0)2 3276 9442; Fax: 33 + (0)2 3514 6655; E-mail: thibaud.decaens@univ-rouen.fr

2 DNA BARCODING

Gibbs 2009). Their potential use in ecological studies has been discussed by Valentini *et al.* (2009). In the domain of soil biodiversity studies, DNA barcoding has been proposed as a promising approach to resolve the strong taxonomic impediment (Decaëns *et al.* 2006, 2008; Rougerie *et al.* 2009). Confirming the early results of DNA data for earthworm taxonomy (Perez-Losada *et al.* 2005), recent applications of DNA barcoding in earthworms are very promising (Huang *et al.* 2007; Chang *et al.* 2009) and the assembly of a comprehensive reference library for this group has been recently undertaken.

The genus *Lumbricus* is a temperate genus represented in France by nine species (Bouché 1972). All of them are morphologically very similar and the key distinctive features are the adult size and the exact position of the clitellum. These two characters are not applicable to juvenile individuals, which are consequently almost impossible to identify at the species level (Bouché 1972). In this study, we demonstrate and emphasize the unique potential of DNA barcoding as a mean to reliably identify juvenile worms of the genus *Lumbricus*.

Material and methods

The study was carried out in Haute-Normandie, France between February and March 2008. Earthworms were sampled from seven grassland and six forest sites located

within the three main landscape units of the region: the plateau area north of the Seine River, the chalky slopes of the valley, and the floodplain of the Seine River (Table 1 and Fig. 1). Depending on the sampling site, soils were NEOLUVISOLS and LUVISOLS with acidic pH (4.0–6.0) developed on loess material; RENDOSOLS with neutral pH (7.0–8.0), aggregated structure and fast organic matter turnover; and HISTOSOLS with a neutral pH (6.0–8.0) underlain by peat layers and hydromorphic horizons whose depth depends on topography (INRA 1998) (Table 1).

In each sampling site, three points located 15 m apart from each other were sampled by combining formalin extraction and hand sorting: (i) 10 L of 0.4% formaldehyde were applied to a 1 m² area and (ii) a soil section of 25 × 25 × 25 cm was dug out in the middle of the area and hand sorted 15 min after formaldehyde application. This method is recommended in temperate soils to allow the best estimations of earthworm populations (Baker & Lee 1993). The efficiency of formaldehyde extraction is in fact known to vary significantly depending on the species, and hand sorting is thus carried out to recover those specimens that remain in the soil. All collected earthworms were stored alive and killed in 50% ethanol in the laboratory.

Specimens of the genus *Lumbricus* were separated from others and adults were morphologically identified to species using Sims & Gerard (1999). Barcoding was

Table 1 Details of collecting sites. Soil classification according to INRA (1998), landscape units after Decaëns *et al.* (2008)

Exact site	GPS coordinates		Sampling date	Soil	Landscape unit	Habitat	Labels
	Latitude	Longitude					
Mont Saint Aignan, campus of the University of Rouen	N49°27'28.1"	E001°04'14.5"	5 April 2007	RENDOSOL	Plateau	Grassland	PGr1
Mont Saint Aignan, campus of the University of Rouen	N49°27'32.2"	E001°04'37.0"	11 & 14 February 2008	RENDOSOL	Plateau	Grassland	PGr2
Yvetot, 'Lycée Agricole'	N49°36'33.4"	E000°44'29.6"	31 March 2008	NEOLUVISOL	Plateau	Grassland	PGr3
Mont Saint Aignan, campus of the University of Rouen	N49°27'26.8"	E001°04'10.0"	6 June 2007	RENDOSOL	Plateau	Forest	PFo1
Saint Adrien	N49°22'26.0"	E001°07'46.5"	21 February 2008	RENDOSOL	Plateau	Forest	PFo2
Eawy forest	N49°41'25.8"	E001°16'35.4"	26 February 2008	LUVISOL	Plateau	Forest	PFo3
Eawy forest	N49°41'43.5"	E001°17'43.2"	26 March 2008	LUVISOL	Plateau	Forest	PFo4
Saint Adrien	N49°22'17.4"	E001°07'51.3"	20 February 2008	RENDOSOL	Chalky slope	Grassland	CGr1
Henouville	N49°29'01.3"	E000°55'49.8"	17 March 2008	RENDOSOL	Chalky slope	Grassland	CGr2
Mont Saint Aignan, campus of the University of Rouen	N49°27'28.1"	E001°04'14.5"	8 June 2007	RENDOSOL	Chalky slope	Forest	CFo1
Marais Vernier, 'Réserve des Manneville'	N49°26'01.4"	E000°30'36.6"	5 March 2008	HISTOSOL	Floodplain	Grassland	FGr1
Marais Vernier, 'Réserve des Manneville'	N49°25'42.9"	E000°31'09.8"	2 April 2008	HISTOSOL	Floodplain	Grassland	FGr2
Marais Vernier, 'Réserve des Manneville'	N49°25'29.5"	E000°31'05.2"	2 April 2008	HISTOSOL	Floodplain	Forest	FFo1

Sequences were assembled with Sequencer 4.5 (Gene Code Corporation, Ann Arbor, MI, USA) and aligned by eye using BIOEDIT version 7.0.5.3 (Hall 1999); we observed no indels in this coding region of the mitochondrial genome and therefore all base positions were aligned with confidence in positional homology. Sequences are publicly available on BOLD (Ratnasingham & Hebert 2007; <http://www.barcodinglife.org>) within the project EWNOR (Earthworms of Normandie) and in GenBank (accession numbers in Appendix). Distance analyses were conducted with MEGA4 (Tamura *et al.* 2007) using a

All pairwise distances involving at least one unidentified juvenile specimen show the same distribution within the bimodal pattern as those involving only adult specimens (Fig. 2), and the analysis of genetic similarity combining adults and juveniles clearly associates both age classes within the same genetic clusters (Fig. 3). Thus, the barcoding approach permits reliable and straightforward species-level identifications for all juveniles that would otherwise be impossible to identify. This results in the identification of morphologically undetected species in two sampling sites out of thirteen (Table 2), and in three

4 DNA BARCODING

Table 2 Number of specimens per species estimated with and without identification of juveniles. The numbers between brackets refer to the relative proportion (%) of each species in the total of specimens that were identified to species level. The numbers between brackets in the juvenile column refer to the proportion of juveniles (%) in the total number of specimens of each sampling site. Label sites refer to individual codes for each collecting site as given in Table 1; Species abbreviations as follow: Lter, *L. terrestris*; Lter1 & Lter2, first and second clusters of *L. terrestris* as identified in Fig. 3; Lfes, *L. festivus*; Lcas, *L. castaneus*; Lrub, *L. rubellus*; Juv, juvenile specimens; bold italic numbers highlight those species, which were only detected through the DNA-barcoding approach

Label Sites/ Species	Without DNA barcode identification					With DNA barcode identification				
	Lter	Lfes	Lcas	Lrub	Juv	Lter1	Lter2	Lfes	Lcas	Lrub
PGr1	5 (62.5)	2 (25)	1 (12.5)	0 (0.0)	9 (52.9)	8 (50.0)	3 (18.7)	4 (25.0)	1 (6.2)	0 (0.0)
PGr2	9 (69.2)	3 (23.1)	1 (7.7)	0 (0.0)	8 (38.1)	4 (19.1)	12 (57.1)	3 (14.3)	2 (9.5)	0 (0.0)
PGr3	2 (50.0)	1 (25)	0 (0.0)	1 (25)	4 (50.0)	0 (0.0)	2 (25.0)	5 (62.5)	0 (0.0)	1 (12.5)
PFo1	1 (100)	0 (0.0)	0 (0.0)	0 (0.0)	2 (66.7)	0 (0.0)	3 (100)	0 (0.0)	0 (0.0)	0 (0.0)
PFo2	0 (0.0)	0 (0.0)	1 (100)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	1 (100)	0 (0.0)
PFo3	0 (0.0)	0 (0.0)	3 (100)	0 (0.0)	1 (25.0)	0 (0.0)	0 (0.0)	0 (0.0)	3 (75.0)	1 (25.0)
PFo4	0 (0.0)	0 (0.0)	8 (66.7)	4 (33.3)	3 (20)	0 (0.0)	0 (0.0)	0 (0.0)	8 (53.3)	7 (46.7)
CGr1	8 (100)	0 (0.0)	0 (0.0)	0 (0.0)	9 (52.9)	2 (11.8)	15 (88.2)	0 (0.0)	0 (0.0)	0 (0.0)
CGr2	9 (100)	0 (0.0)	0 (0.0)	0 (0.0)	9 (50.0)	18 (100)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
CFo1	7 (100)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	7 (100)	0 (0.0)	0 (0.0)	0 (0.0)
FGGr1	0 (0.0)	5 (50)	0 (0.0)	5 (50)	5 (33.3)	0 (0.0)	0 (0.0)	5 (33.3)	0 (0.0)	10 (66.7)
FGGr2	0 (0.0)	0 (0.0)	0 (0.0)	2 (100)	2 (50.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	4 (100.0)
VFo1	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	2 (100)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	2 (100)

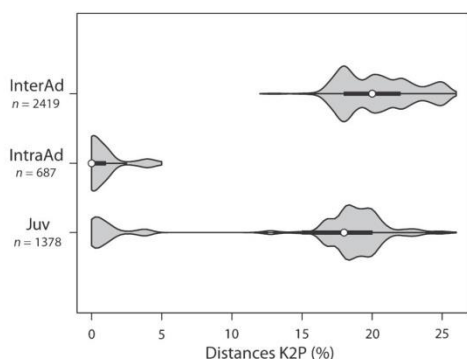


Fig. 2 Violin plots of intraspecific (IntraAd), interspecific (InterAd) pairwise distances for adult specimens and for juvenile specimens (Juv). The number of pairwise comparisons is indicated. The medians of data (white points), black bars indicating the interquartile range, and spikes extending to the upper- and lower-adjacent values are included.

other cases, the barcoding approach permits detection of both clusters of *L. terrestris*, which also increases the estimated species richness. Finally, molecular identifications also affected species relative abundance in almost all cases where more than one species was present. As a consequence, the global assessment of earthworm assemblages is significantly improved when extended to juvenile specimens through the use of DNA barcodes for species identification.

We thus emphasize that DNA barcoding, by enabling such identifications and by revealing cryptic species, represents a very important breakthrough for the integration of earthworm diversity data into soil studies. Indeed, earthworm biological and ecological research depends much on reliable species lists and has long been impeded by the impossibility to properly consider juveniles. So far, authors have either considered juveniles specimens as a distinct category (Margerie *et al.* 2001; Blackshaw *et al.* 2007; Pelosi *et al.* 2008) or have simply excluded them from analysis (Smetak *et al.* 2007). The collecting locality and/or the composition of local earthworm community is also occasionally used to infer species identity of juvenile specimens, but the method is error-prone, especially when considering the prevalence of sympatric cryptic species within common earthworm taxa (King *et al.* 2008). As exemplified by our results, these strategies unquestionably affect density/abundance measurements and can potentially lead to underestimates of species richness.

The value of DNA barcodes in extending species identification to juvenile specimens is not unique to the genus *Lumbricus* nor to earthworms; it can be profitable to the study of any group for which DNA barcodes are diagnostic and identification of some development stages is challenging. The reduction of sequencing cost, the multiplication of sequencing facilities, and more specifically the development of an international DNA barcoding initiative (iBOL; <http://www.ibolproject.org>) are rapidly facilitating access to this approach, making it a potential standard procedure in soil biodiversity assessment. Because earthworms are key organisms of soil

DNA BARCODING 5

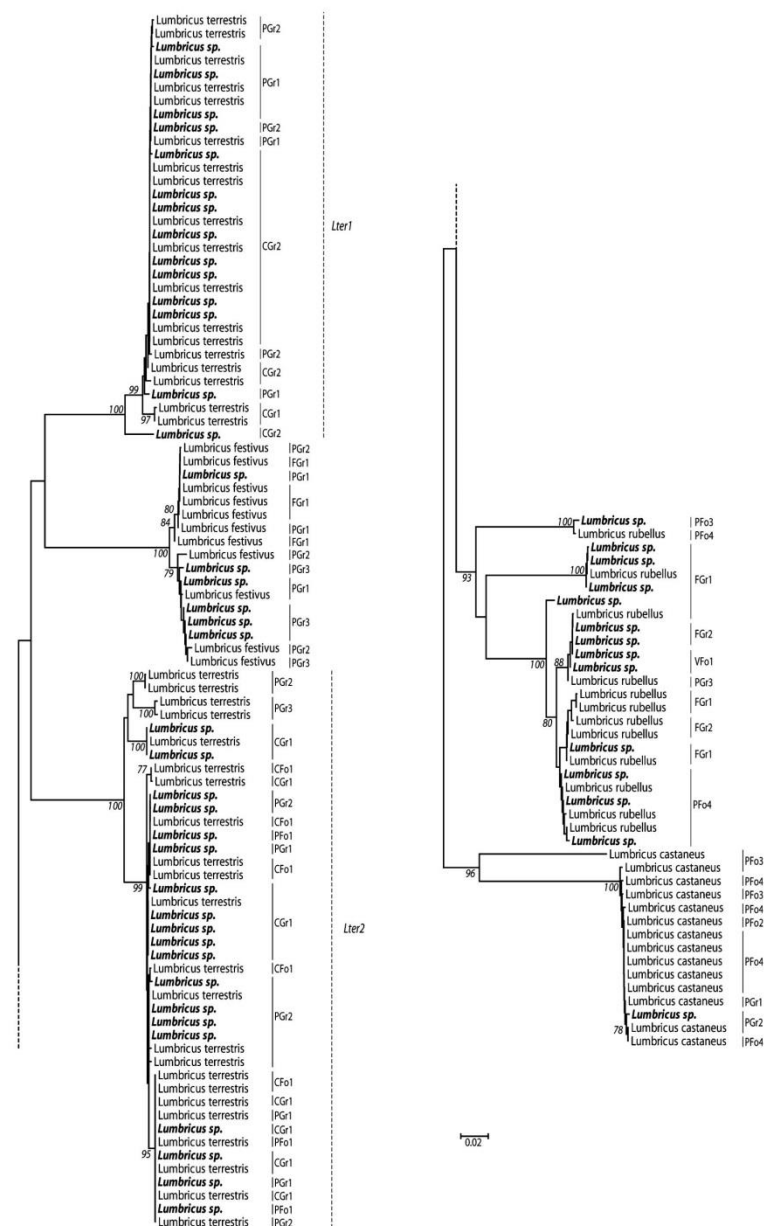


Fig. 3 Neighbor-joining tree (K2P distances) of DNA barcodes for 131 *Lumbricus* specimens. Juvenile specimens are in bold italics. Bootstrap values greater than 75% are shown above branches. Lter1 and Lter2 refer to the two distinct genetic clusters of *L. terrestris*. Locality codes are figured on the tree and refer to codes used in Fig. 1 and in Table 1.

6 DNA BARCODING

ecosystems and are important indicators of soil health and quality (Paoletti 1999; Lavelle *et al.* 2006; Suthar 2009), it is crucial to improve the resolution and the reliability of species identification (Nahmani *et al.* 2006) and to enable identifications for all life stages. DNA barcoding can meet both these needs and should rapidly be integrated into soil studies addressing earthworm diversity or based on diversity data.

Acknowledgements

We thank NSERC and Genome Canada through the Ontario Genomics Institute for supporting barcode analysis, and the Lycée Agricole d'Yvetot, the Conservatoire de Sites de Haute Normandie, the Office National des Forêts and the Réserve Naturelle des Manneville for allowing earthworm sampling in their soils.

References

- Baker G, Lee KE (1993) Earthworms. In: *Field Samplings and Methods of Analysis* (ed. Carter MR), pp. 359–371. Lewis Publishers, Boca Raton.
- Blackshaw RP, Donovan SE, Hazarika S, Bol R, Dixon ER (2007) Earthworm responses to long term agricultural management practices: spatial relationships with soil properties. *European Journal of Soil Biology*, **43**, S171–S175.
- Bogh PS (1992) Identification of earthworms (Lumbricidae): choice of methods and distinction criteria. *Megadrilogica*, **4**, 163–174.
- Borisenko AV, Lim BK, Ivanova NV, Hanner RH, Hebert PDN (2008) DNA barcoding in surveys of small mammal communities: a field study in Suriname. *Molecular Ecology Resources*, **8**, 471–479.
- Bouché M (1972) *Lombriciens de France*. Ecologie et Systématique. Institut National de Recherches Agronomiques, Paris.
- Chang C-H, Rougerie R, Chen J-H (2009) Identifying earthworms through DNA barcodes: Pitfalls and promise. *Pedobiologia*, **52**, 171–180.
- Decaëns T, Rougerie R (2008) Descriptions of two new species of Hemileucinae (Lepidoptera: Saturniidae) from the region of Muzo in Colombia- Evidence from morphology and DNA barcodes. *Zootaxa*, **1944**, 34–52.
- Decaëns T, Jiménez JJ, Gioia C, Measey J, Lavelle P (2006) The values of soil animals for conservation biology. *European Journal of Soil Biology*, **42**, S23–S38.
- Decaëns T, Lavelle P, Jimenez JJ (2008) Priorities for conservation of soil animals. *CAB reviews: Perspectives in Agriculture, Veterinary Science, Nutrition and Natural Resources*, **3**, 14.
- Folmer O, Black M, Hoeh W, Lutz RRV (1994) DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology*, **3**, 294–299.
- Gibbs J (2009) Integrative taxonomy identifies new (and old) species in the *Lasioglossum* (*Dialictus*) *tegulare* (Robertson) species group (Hymenoptera, Halictidae). *Zootaxa*, **2032**, 1–38.
- Hajibabaei M, DeWaard JR, Ivanova NV *et al.* (2005) Critical factors for assembling a high volume of DNA barcodes. *Philosophical Transactions of the Royal Society B-Biological Sciences*, **360**, 1959–1967.
- Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series*, **41**, 95–98.
- Hebert PDN, Alina C, Shelley LB, Jeremy RD (2003) Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences*, **270**, 313–321.
- Hebert PDN, Penton EH, Burns JM, Janzen DH, Hallwachs W (2004a) Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 14812–14817.
- Hebert PDN, Stoeckle MY, Zemlak TS, Francis CM (2004b) Identification of birds through DNA barcodes. *Plos Biology*, **2**, 1657–1663.
- Huang J, Xu Q, Sun ZJ, Tang GL, Su ZY (2007) Identifying earthworms through DNA barcodes. *Pedobiologia*, **51**, 301–309.
- INRA (1998) *A Sound Reference Base for Soils - "Référentiel Pédologique"*. INRA, Paris.
- Ivanova NV, Dewaard JR, Hebert PDN (2006) An inexpensive, automation-friendly protocol for recovering high-quality DNA. *Molecular Ecology Notes*, **6**, 998–1002.
- Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide-sequences. *Journal of Molecular Evolution*, **16**, 111–120.
- King RA, Tibble AL, Symondson WOC (2008) Opening a can of worms: unprecedented sympatric cryptic diversity within British lumbricid earthworms. *Molecular Ecology*, **17**, 4684–4698.
- Lavelle P, Decaëns T, Aubert A *et al.* (2006) Soil invertebrates and ecosystem services. *European Journal of Soil Biology*, **42**, S3–S15.
- Margerie P, Decaens T, Bureau F, Alard D (2001) Spatial distribution of earthworm species assemblages in a chalky slope of the Seine Valley (Normandy, France). *European Journal of Soil Biology*, **37**, 291–296.
- Martinez JJ, Zaldivar-Riveron A, Saez AG (2008) Reclassification of *Bracon mendocinus*, a gall-associated doryctine wasp, and description of a new closely related species of *Allorhogas* (Hymenoptera: Braconidae). *Journal of Natural History*, **42**, 2689–2701.
- Nahmani J, Lavelle P, Rossi JP (2006) Does changing the taxonomic resolution alter the value of soil macroinvertebrates as bioindicators of metal pollution? *Soil Biology and Biochemistry*, **38**, 385–396.
- Paoletti MG (1999) The role of earthworms for assessment of sustainability and as bioindicators. *Agriculture Ecosystems & Environment*, **74**, 137–155.
- Pelosi C, Bertrand M, Capowiez Y, Boizard H, Roger-Estrade J (2008) Earthworm collection from agricultural fields: comparisons of selected expellants in presence/absence of hand-sorting. *European Journal of Soil Biology*, **45**, 176–183.
- Perez-Losada M, Eiroa J, Mato S, Dominguez J (2005) Phylogenetic species delimitation of the earthworms *Eisenia fetida* (Savigny, 1826) and *Eisenia andrei* Bouche, 1972 (Oligochaeta, Lumbricidae) based on mitochondrial and nuclear DNA sequences. *Pedobiologia*, **49**, 317–324.
- Ratnasingham S, Hebert PDN (2007) BOLD: The Barcode of Life Data System (<http://www.barcodinglife.org>). *Molecular Ecology Notes*, **7**, 355–364.

DNA BARCODING 7

- Rougerie R, Decaëns T, Deharveng L *et al.* (2009) DNA barcodes for soil animal taxonomy: transcending the final frontier. *Pesquisa Agropecuária Brasileira*, **44**, 789–801.
- Saitou N, Nei M (1987) The Neighbor-Joining Method - a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, **4**, 406–425.
- Sims RW, Gerard BM (1999) *Earthworms*. FSC Publications, London.
- Smetak KM, Johnsn-Maynard JL, Lloyd JE (2007) Earthworm population density and diversity in different-aged urban systems. *Applied Soil Ecology*, **37**, 161–168.
- Smith MA, Wood DM, Janzen DH, Hallwachs W, Hebert PDN (2007) DNA barcodes affirm that 16 species of apparently generalist tropical parasitoid flies (Diptera, Tachinidae) are not all generalists. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 4967–4972.
- Suthar S (2009) Earthworm communities a bioindicator of arable land management practices: a case study in semiarid region of India. *Ecological Indicators*, **9**, 588–594.
- Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. *Molecular Biology and Evolution*, **24**, 1596–1599.
- Vaglia T, Haxaire J, Kitching IJ, Meusnier I, Rougerie R (2008) Morphology and DNA barcoding reveal three cryptic species within the *Xylophanes neoptolemus* and *loelia* species-groups (Lepidoptera: Sphingidae). *Zootaxa*, **1923**, 18–36.
- Valentini A, Pompanon F, Taberlet P (2009) DNA barcoding for ecologists. *Trends in Ecology & Evolution*, **24**, 110–117.
- Ward RD, Zemlak TS, Innes BH, Last PR, Hebert PDN (2005) DNA barcoding Australia's fish species. *Philosophical Transactions of the Royal Society B-Biological Sciences*, **360**, 1847–1857.

Appendix

Accession numbers of the COI sequences in BOLD and GenBank

BOLD Sample ID	BOLD Process ID	Species names	GenBank accession numbers	COI-5P length (Ambiguous base-calls)	Site labels
EW-ECO-0012	EWNOR012-07	<i>Lumbricus terrestris</i>	FJ937319	658 (0n)	PGr1
EW-ECO-0013	EWNOR013-07	<i>Lumbricus terrestris</i>	FJ937315	658 (0n)	PGr1
EW-ECO-0014	EWNOR014-07	<i>Lumbricus terrestris</i>	FJ937316	658 (0n)	PGr1
EW-ECO-0015	EWNOR015-07	<i>Lumbricus terrestris</i>	FJ937317	658 (0n)	PGr1
EW-ECO-0016	EWNOR016-07	<i>Lumbricus terrestris</i>	FJ937318	658 (0n)	PGr1
EW-ECO-0017	EWNOR017-07	<i>Lumbricus festivus</i>	FJ937287	618 (0n)	PGr1
EW-ECO-0018	EWNOR018-07	<i>Lumbricus festivus</i>	FJ937289	658 (0n)	PGr1
EW-ECO-0019	EWNOR019-07	<i>Lumbricus</i> sp.	FJ937295	658 (0n)	PGr1
EW-ECO-0020	EWNOR020-07	<i>Lumbricus</i> sp.	GU206212	248 (5n)	PGr1
EW-ECO-0021	EWNOR021-07	<i>Lumbricus</i> sp.	FJ937296	658 (0n)	PGr1
EW-ECO-0022	EWNOR022-07	<i>Lumbricus</i> sp.	FJ937297	658 (0n)	PGr1
EW-ECO-0023	EWNOR023-07	<i>Lumbricus</i> sp.	FJ937298	658 (0n)	PGr1
EW-ECO-0024	EWNOR024-07	<i>Lumbricus</i> sp.	FJ937299	658 (0n)	PGr1
EW-ECO-0025	EWNOR025-07	<i>Lumbricus</i> sp.	FJ937300	658 (0n)	PGr1
EW-ECO-0026	EWNOR026-07	<i>Lumbricus</i> sp.	FJ937301	465 (0n)	PGr1
EW-ECO-0027	EWNOR027-07	<i>Lumbricus castaneus</i>	FJ937285	658 (0n)	PGr1
EW-ECO-0029	EWNOR029-07	<i>Lumbricus</i> sp.	FJ937308	658 (0n)	PGr1
EW-ECO-0059	EWNOR059-07	<i>Lumbricus terrestris</i>	GU014224	669 (0n)	PFo1
EW-ECO-0060	EWNOR060-07	<i>Lumbricus</i> sp.	GU014230	669 (0n)	PFo1
EW-ECO-0065	EWNOR065-07	<i>Lumbricus</i> sp.	GU014231	669 (0n)	PFo1
EW-ECO-0077	EWNOR077-07	<i>Lumbricus terrestris</i>	GU014223	669 (0n)	CFo1
EW-ECO-0089	EWNOR089-07	<i>Lumbricus terrestris</i>	GU014225	669 (0n)	CFo1
EW-ECO-0090	EWNOR090-07	<i>Lumbricus terrestris</i>	GU014226	669 (0n)	CFo1
EW-ECO-0091	EWNOR091-07	<i>Lumbricus terrestris</i>	GU014227	669 (0n)	CFo1
EW-ECO-0092	EWNOR092-07	<i>Lumbricus terrestris</i>	GU014228	669 (0n)	CFo1
EW-ECO-0093	EWNOR093-07	<i>Lumbricus terrestris</i>	GU014229	669 (0n)	CFo1
EW-ECO-0094	EWNOR094-07	<i>Lumbricus terrestris</i>	GU206226	589 (0n)	CFo1
EW-ECO-0096	EWNOR095-08	<i>Lumbricus</i> sp.	FJ937304	658 (0n)	PGr2
EW-ECO-0098	EWNOR097-08	<i>Lumbricus</i> sp.	GU206211	633 (0n)	PGr2
EW-ECO-0099	EWNOR098-08	<i>Lumbricus terrestris</i>	FJ937312	644 (0n)	PGr2
EW-ECO-0100	EWNOR099-08	<i>Lumbricus terrestris</i>	FJ937313	632 (0n)	PGr2
EW-ECO-0101	EWNOR100-08	<i>Lumbricus terrestris</i>	FJ937314	658 (0n)	PGr2
EW-ECO-0108	EWNOR107-08	<i>Lumbricus</i> sp.	FJ937305	644 (0n)	PGr2
EW-ECO-0109	EWNOR108-08	<i>Lumbricus</i> sp.	FJ937306	634 (0n)	PGr2
EW-ECO-0110	EWNOR109-08	<i>Lumbricus</i> sp.	FJ937307	644 (0n)	PGr2

8 DNA BARCODING

Appendix Continued

BOLD Sample ID	BOLD Process ID	Species names	GenBank accession numbers	COI-5P length (Ambiguous base-calls)	Site labels
EW-ECO-0111	EWNOR110-08	<i>Lumbricus terrestris</i>	FJ937321	634 (0n)	PGr2
EW-ECO-0112	EWNOR111-08	<i>Lumbricus terrestris</i>	FJ937322	644 (0n)	PGr2
EW-ECO-0113	EWNOR112-08	<i>Lumbricus terrestris</i>	FJ937323	644 (0n)	PGr2
EW-ECO-0120	EWNOR119-08	<i>Lumbricus terrestris</i>	FJ937324	645 (0n)	PGr2
EW-ECO-0121	EWNOR120-08	<i>Lumbricus terrestris</i>	FJ937327	644 (0n)	PGr2
EW-ECO-0122	EWNOR121-08	<i>Lumbricus terrestris</i>	FJ937320	657 (0n)	PGr2
EW-ECO-0129	EWNOR128-08	<i>Lumbricus</i> sp.	FJ937309	634 (0n)	PGr2
EW-ECO-0130	EWNOR129-08	<i>Lumbricus</i> sp.	FJ937310	519 (0n)	PGr2
EW-ECO-0131	EWNOR130-08	<i>Lumbricus</i> sp.	FJ937311	572 (1n)	PGr2
EW-ECO-0132	EWNOR131-08	<i>Lumbricus festivus</i>	FJ937290	655 (0n)	PGr2
EW-ECO-0133	EWNOR132-08	<i>Lumbricus festivus</i>	FJ937291	644 (0n)	PGr2
EW-ECO-0134	EWNOR133-08	<i>Lumbricus festivus</i>	FJ937286	570 (1n)	PGr2
EW-ECO-0135	EWNOR134-08	<i>Lumbricus castaneus</i>	FJ937284	655 (0n)	PGr2
EW-ECO-0156	EWNOR155-08	<i>Lumbricus terrestris</i>	GU206239	643 (1n)	CGr1
EW-ECO-0157	EWNOR156-08	<i>Lumbricus terrestris</i>	GU206225	658 (0n)	CGr1
EW-ECO-0158	EWNOR157-08	<i>Lumbricus terrestris</i>	GU206238	508 (0n)	CGr1
EW-ECO-0166	EWNOR165-08	<i>Lumbricus</i> sp.	GU206224	650 (1n)	CGr1
EW-ECO-0167	EWNOR166-08	<i>Lumbricus</i> sp.	GU206223	658 (0n)	CGr1
EW-ECO-0168	EWNOR167-08	<i>Lumbricus</i> sp.	GU206222	632 (0n)	CGr1
EW-ECO-0175	EWNOR174-08	<i>Lumbricus terrestris</i>	GU206237	580 (0n)	CGr1
EW-ECO-0176	EWNOR175-08	<i>Lumbricus terrestris</i>	GU206221	658 (0n)	CGr1
EW-ECO-0177	EWNOR176-08	<i>Lumbricus</i> sp.	GU206220	658 (0n)	CGr1
EW-ECO-0178	EWNOR177-08	<i>Lumbricus</i> sp.	GU206219	658 (0n)	CGr1
EW-ECO-0179	EWNOR178-08	<i>Lumbricus</i> sp.	GU206218	658 (0n)	CGr1
EW-ECO-0194	EWNOR193-08	<i>Lumbricus terrestris</i>	GU206217	658 (0n)	CGr1
EW-ECO-0195	EWNOR194-08	<i>Lumbricus terrestris</i>	GU206216	658 (0n)	CGr1
EW-ECO-0196	EWNOR195-08	<i>Lumbricus terrestris</i>	GU206215	658 (0n)	CGr1
EW-ECO-0197	EWNOR196-08	<i>Lumbricus</i> sp.	GU206214	658 (0n)	CGr1
EW-ECO-0198	EWNOR197-08	<i>Lumbricus</i> sp.	GU206213	658 (0n)	CGr1
EW-ECO-0199	EWNOR198-08	<i>Lumbricus</i> sp.	GU206210	658 (0n)	CGr1
EW-ECO-0215	EWNOR214-08	<i>Lumbricus castaneus</i>	GU206161	658 (0n)	PFo2
EW-ECO-0227	EWNOR226-08	<i>Lumbricus castaneus</i>	GU206160	658 (0n)	PFo3
EW-ECO-0228	EWNOR227-08	<i>Lumbricus castaneus</i>	GU206159	658 (0n)	PFo3
EW-ECO-0230	EWNOR229-08	<i>Lumbricus</i> sp.	GU206170	658 (0n)	PFo3
EW-ECO-0254	EWNOR253-08	<i>Lumbricus castaneus</i>	GU206163	658 (0n)	PFo3
EW-ECO-0264	EWNOR263-08	<i>Lumbricus festivus</i>	GU206168	658 (0n)	FGr1
EW-ECO-0265	EWNOR264-08	<i>Lumbricus festivus</i>	GU206167	658 (0n)	FGr1
EW-ECO-0266	EWNOR265-08	<i>Lumbricus festivus</i>	GU206166	658 (0n)	FGr1
EW-ECO-0267	EWNOR266-08	<i>Lumbricus rubellus</i>	GU206191	603 (0n)	FGr1
EW-ECO-0268	EWNOR267-08	<i>Lumbricus rubellus</i>	GU206190	636 (0n)	FGr1
EW-ECO-0269	EWNOR268-08	<i>Lumbricus rubellus</i>	GU206189	593 (17n)	FGr1
EW-ECO-0278	EWNOR277-08	<i>Lumbricus</i> sp.	GU206209	630 (0n)	FGr1
EW-ECO-0279	EWNOR278-08	<i>Lumbricus</i> sp.	GU206208	646 (0n)	FGr1
EW-ECO-0280	EWNOR279-08	<i>Lumbricus</i> sp.	GU206183	292 (1n)	FGr1
EW-ECO-0284	EWNOR283-08	<i>Lumbricus festivus</i>	GU206165	618 (1n)	FGr1
EW-ECO-0285	EWNOR284-08	<i>Lumbricus</i> sp.	GU206207	658 (0n)	FGr1
EW-ECO-0293	EWNOR292-08	<i>Lumbricus rubellus</i>	GU206187	458 (4n)	FGr1
EW-ECO-0294	EWNOR293-08	<i>Lumbricus rubellus</i>	GU206186	653 (0n)	FGr1
EW-ECO-0295	EWNOR294-08	<i>Lumbricus rubellus</i>	GU206185	554 (11n)	FGr1
EW-ECO-0296	EWNOR295-08	<i>Lumbricus festivus</i>	GU206164	648 (0n)	FGr1
EW-ECO-0297	EWNOR296-08	<i>Lumbricus</i> sp.	GU206184	511 (0n)	FGr1
EW-ECO-0298	EWNOR297-08	<i>Lumbricus</i> sp.	GU206206	518 (9n)	FGr1
EW-ECO-0306	EWNOR305-08	<i>Lumbricus terrestris</i>	GU206236	648 (0n)	CGr2
EW-ECO-0307	EWNOR306-08	<i>Lumbricus terrestris</i>	GU206235	648 (0n)	CGr2
EW-ECO-0308	EWNOR307-08	<i>Lumbricus terrestris</i>	GU206234	640 (0n)	CGr2

DNA BARCODING 9

Appendix Continued

BOLD Sample ID	BOLD Process ID	Species names	GenBank accession numbers	COI-5P length (Ambiguous base-calls)	Site labels
EW-ECO-0309	EWNOR308-08	<i>Lumbricus</i> sp.	GU206233	648 (0n)	CGr2
EW-ECO-0310	EWNOR309-08	<i>Lumbricus</i> sp.	GU206205	648 (0n)	CGr2
EW-ECO-0311	EWNOR310-08	<i>Lumbricus</i> sp.	GU206204	648 (0n)	CGr2
EW-ECO-0320	EWNOR319-08	<i>Lumbricus terrestris</i>	GU206232	648 (0n)	CGr2
EW-ECO-0321	EWNOR320-08	<i>Lumbricus terrestris</i>	GU206231	648 (0n)	CGr2
EW-ECO-0322	EWNOR321-08	<i>Lumbricus terrestris</i>	GU206230	648 (0n)	CGr2
EW-ECO-0325	EWNOR324-08	<i>Lumbricus</i> sp.	GU206203	648 (0n)	CGr2
EW-ECO-0326	EWNOR325-08	<i>Lumbricus</i> sp.	GU206202	648 (0n)	CGr2
EW-ECO-0327	EWNOR326-08	<i>Lumbricus</i> sp.	GU206201	648 (0n)	CGr2
EW-ECO-0334	EWNOR333-08	<i>Lumbricus terrestris</i>	GU206229	648 (0n)	CGr2
EW-ECO-0335	EWNOR334-08	<i>Lumbricus terrestris</i>	GU206228	648 (0n)	CGr2
EW-ECO-0336	EWNOR335-08	<i>Lumbricus terrestris</i>	GU206227	648 (0n)	CGr2
EW-ECO-0337	EWNOR336-08	<i>Lumbricus</i> sp.	GU206200	648 (0n)	CGr2
EW-ECO-0338	EWNOR337-08	<i>Lumbricus</i> sp.	GU206199	648 (0n)	CGr2
EW-ECO-0339	EWNOR338-08	<i>Lumbricus</i> sp.	GU206198	648 (0n)	CGr2
<i>EW-ECO-0346</i>	<i>EWNOR345-08</i>	<i>Lumbricus rubellus</i>	<i>GU206183</i>	<i>292 (0n)</i>	<i>PFo4</i>
EW-ECO-0347	EWNOR346-08	<i>Lumbricus rubellus</i>	GU206182	511 (0n)	PFo4
EW-ECO-0348	EWNOR347-08	<i>Lumbricus rubellus</i>	GU206181	635 (0n)	PFo4
EW-ECO-0349	EWNOR348-08	<i>Lumbricus</i> sp.	GU206180	648 (0n)	PFo4
EW-ECO-0350	EWNOR349-08	<i>Lumbricus castaneus</i>	GU206158	658 (0n)	PFo4
EW-ECO-0351	EWNOR350-08	<i>Lumbricus castaneus</i>	GU206157	658 (0n)	PFo4
EW-ECO-0352	EWNOR351-08	<i>Lumbricus castaneus</i>	GU206156	658 (0n)	PFo4
EW-ECO-0370	EWNOR369-08	<i>Lumbricus castaneus</i>	GU206162	648 (0n)	PFo4
EW-ECO-0371	EWNOR370-08	<i>Lumbricus</i> sp.	GU206197	635 (2n)	PFo4
<i>EW-ECO-0372</i>	<i>EWNOR371-08</i>	<i>Lumbricus</i> sp.	<i>GU206196</i>	<i>565 (34n)</i>	<i>PFo4</i>
EW-ECO-0373	EWNOR372-08	<i>Lumbricus</i> sp.	GU206179	635 (0n)	PFo4
EW-ECO-0376	EWNOR375-08	<i>Lumbricus castaneus</i>	GU206155	658 (0n)	PFo4
EW-ECO-0377	EWNOR376-08	<i>Lumbricus castaneus</i>	GU206154	658 (0n)	PFo4
EW-ECO-0378	EWNOR377-08	<i>Lumbricus castaneus</i>	GU206153	658 (0n)	PFo4
EW-ECO-0379	EWNOR378-08	<i>Lumbricus castaneus</i>	GU206152	658 (0n)	PFo4
EW-ECO-0380	EWNOR379-08	<i>Lumbricus rubellus</i>	GU206178	644 (0n)	PFo4
EW-ECO-0381	EWNOR380-08	<i>Lumbricus rubellus</i>	GU206169	658 (0n)	PFo4
EW-ECO-0386	EWNOR385-08	<i>Lumbricus</i> sp.	FJ937293	658 (0n)	PGr3
EW-ECO-0387	EWNOR386-08	<i>Lumbricus</i> sp.	FJ937294	658 (0n)	PGr3
EW-ECO-0388	EWNOR387-08	<i>Lumbricus festivus</i>	FJ937288	584 (0n)	PGr3
EW-ECO-0398	EWNOR397-08	<i>Lumbricus</i> sp.	FJ937302	644 (0n)	PGr3
EW-ECO-0399	EWNOR398-08	<i>Lumbricus</i> sp.	FJ937303	658 (0n)	PGr3
EW-ECO-0408	EWNOR407-08	<i>Lumbricus terrestris</i>	FJ937325	658 (0n)	PGr3
EW-ECO-0409	EWNOR408-08	<i>Lumbricus terrestris</i>	FJ937326	658 (0n)	PGr3
EW-ECO-0410	EWNOR409-08	<i>Lumbricus rubellus</i>	FJ937292	545 (0n)	PGr3
<i>EW-ECO-0416</i>	<i>EWNOR415-08</i>	<i>Lumbricus rubellus</i>	<i>GU206177</i>	<i>470 (51n)</i>	<i>FGr2</i>
EW-ECO-0418	EWNOR417-08	<i>Lumbricus rubellus</i>	GU206175	613 (0n)	FGr2
EW-ECO-0421	EWNOR420-08	<i>Lumbricus rubellus</i>	GU206173	579 (0n)	FGr2
EW-ECO-0422	EWNOR421-08	<i>Lumbricus</i> sp.	GU206195	658 (0n)	FGr2
<i>EW-ECO-0423</i>	<i>EWNOR422-08</i>	<i>Lumbricus</i> sp.	<i>GU206194</i>	<i>541 (59n)</i>	<i>FGr2</i>
EW-ECO-0424	EWNOR423-08	<i>Lumbricus</i> sp.	GU206172	658 (0n)	FGr2
<i>EW-ECO-0425</i>	<i>EWNOR424-08</i>	<i>Lumbricus</i> sp.	<i>GU206193</i>	<i>426 (38n)</i>	<i>FGr2</i>
<i>EW-ECO-0426</i>	<i>EWNOR425-08</i>	<i>Lumbricus</i> sp.	<i>GU206192</i>	<i>380 (39n)</i>	<i>FGr2</i>
EW-ECO-0430	EWNOR429-08	<i>Lumbricus</i> sp.	GU206171	620 (0n)	VFo1
EW-ECO-0432	EWNOR431-08	<i>Lumbricus</i> sp.	GU206151	658 (0n)	VFo1
<i>EW-ECO-0434</i>	<i>EWNOR433-08</i>	<i>Lumbricus</i> sp.	<i>GU206150</i>	<i>269 (2n)</i>	<i>VFo1</i>

For each sequence, the length and the number of ambiguous base-calls are specified. Site labels refer to Table 1. Lines in italics signify that the sequence was not used in the data analyses (size < 400 bp and/or number of ambiguous base-calls >10)



ELSEVIER

Contents lists available at ScienceDirect

Pedobiologia - International Journal of Soil Biology

journal homepage: www.elsevier.de/pedobi



New insight into the genetic structure of the *Allolobophora chlorotica* aggregate in Europe using microsatellite and mitochondrial data

L. Dupont^{a,*}, F. Lazrek^a, D. Porco^{b,d}, R.A. King^{c,g}, R. Rougerie^{b,d}, W.O.C. Symondson^c, A. Livet^a, B. Richard^{d,h}, T. Decaëns^d, K.R. Butt^e, J. Mathieu^f

^a UMR Bioemco, équipe IBIOS, Faculté des Sciences et Technologie, Université Paris-Est Créteil, 61 Avenue du Général de Gaulle, 94010 Créteil Cedex, France

^b Biodiversity Institute of Ontario, University of Guelph, 50 Gordon Street, Guelph, Ontario N1G 2W1, Canada

^c Cardiff School of Biosciences, Biomedical Sciences Building, Cardiff University, Museum Avenue, Cardiff CF10 3AX, UK

^d Laboratoire d'Ecologie, EA 1293 ECODIV, UFR Sciences et Techniques, Université de Rouen, 76821 Mont Saint Aignan Cedex, France

^e School of Built and Natural Environment, University of Central Lancashire, Preston PR1 2HE, UK

^f UPMC Université Paris 06, UMR Bioemco, 46 rue d'Ulm, 75230 Paris Cedex 05, France

^g College of Life and Environmental Sciences, Geoffrey Pope Building, Stocker Road, Exeter EX4 4QD, UK

^h Laboratoire Biosol, Esitpa – Ecole d'Ingénieurs en Agriculture, 3 rue du Tronquet BP 40118 76134 Mont-Saint-Aignan Cedex, France

ARTICLE INFO

Article history:

Received 2 December 2010

Received in revised form 1 March 2011

Accepted 1 March 2011

Keywords:

Cryptic diversity
DNA barcoding
Earthworms
Enriched genomic library
Microsatellites
Population sub-structure

ABSTRACT

The taxonomic status of *Allolobophora chlorotica* is still the subject of considerable discussion. After the recent experimental demonstration that the two regularly observed colour morphs (pink and green) were not fully interfertile and likely represent two distinct species, molecular analyses revealed a more complex picture with the occurrence of five mitochondrial lineages (named L1 to L5) in *A. chlorotica*, two within the green morph and three within the pink morph. Although nuclear markers (AFLPs) confirmed that the pink morph might consist of three different taxa, AFLPs data suggested that the green morph may be a single taxon. In order to further characterize the population genetic structure within *A. chlorotica* in addition to test for reproductive isolation between lineages, eight polymorphic microsatellite loci were developed using a microsatellite-enriched genomic library from an individual belonging to lineage L2. The number of alleles per locus varied from 9 to 29 in a set of individuals belonging to L2. Considerable sub-structure was observed within L2 populations suggesting a low dispersal capability in this species and a distribution of the individuals in patches that could function as panmictic units. These markers were transferable to the other lineages but only five loci could be amplified consistently in four of the lineages (L1, L2, L3 and L4). Microsatellite data confirmed that the green morph represents a single taxon. Although the integrity of the three previously documented lineages within the pink morph was also generally supported by the data, microsatellites provided evidence for hybridization between lineages and between morphs in the field. Moreover, mitochondrial data revealed the existence of two additional mitochondrial lineages within the pink morph. The taxonomic status of the pink morph remains thus unclear, requiring a thorough and comprehensive study.

© 2011 Elsevier GmbH. All rights reserved.

Introduction

Allolobophora chlorotica (Savigny, 1826) is a widespread temperate earthworm species with a native distribution over much of Europe and Anatolia (Bouché 1972). It is also known as an invasive species in North America (Bouché 1972), South America, New Zealand (Sims and Gerard 1999) and North Africa (Rouabah and Descamps 2001). This endogeic (soil-dwelling) species is an obligatory bi-parental species (Lowe and Butt 2007, 2008) usually located in the rhizosphere not more than 60 mm from the soil surface (Sims and Gerard 1999).

Two colour forms, one green and one pink, have been described in *A. chlorotica*; they are known to be caused by the presence or absence of a bilin pigment (Satchell 1967), respectively. The two morphs have contrasting ecological preferences linked to soil moisture. Field observations have shown that the green morph dominates in wet soils while the pink morph prevails in dry soils (Satchell 1967). Furthermore, laboratory experiments confirmed that growth and maturation of the green morph were significantly reduced in dry-soil conditions while growth of the pink morph was not significantly different in wet or dry conditions (Lowe and Butt 2007).

Although mixed populations are present in grasslands (Satchell 1967; King et al. 2008), Lowe and Butt (2007) suggested that: "soil moisture content may act to isolate the two morphs, providing, in extremes, a barrier to inter-morphic mating". Breeding exper-

* Corresponding author. Tel.: +33 0145171664.

E-mail addresses: lise.dupont@u-pec.fr, lise.dupont@univ-paris12.fr (L. Dupont).

iments have shown that the viability of cocoons produced by the green morph in pink $\times$ green pairings was severely restricted (Lowe and Butt 2008). Moreover, back-crossing of hybrids with pure bred morphs has indicated that hybrids are male sterile. Thus, these results suggest postzygotic isolating mechanisms and reinforce the idea that the two morphs may represent distinct species.

A recent molecular analysis of *A. chlorotica* using mitochondrial cytochrome oxidase I (COI) and ribosomal 16S genes has shown the presence of five highly divergent lineages in this species (King et al. 2008). Interestingly, the two colour morphs did not share any COI haplotypes. In addition, amplified fragment length polymorphism (AFLP) markers confirmed the existence of three genetically isolated pink morph lineages but showed that the two green morph lineages may be interbreeding and therefore represent a single taxon.

In this work, we revisit the complex genetic status of *A. chlorotica* by extending the existing dataset for COI mitochondrial lineages (King et al. 2008) and by combining it to the analysis of microsatellite markers specifically designed for that study and whose transferability is tested across the different lineages. Because they are co-dominant and locus-specific, microsatellites are indeed the markers of choice to test for reproductive isolation between cryptic lineages (Dupont 2009; Lazrek et al. 2011). The relationships between lineages are discussed in the light of these new data, looking for evidence of inter-lineages hybridization in the field and for possible phylogeographic differentiation of the lineages.

Materials and methods

Specimens and DNA isolation

Samples of *A. chlorotica* were mainly collected from two populations in England (Croston) and in France (Yvetot), as well as from eight other localities in France (Fig. 1, Appendix A). The total genomic DNA was extracted using either the DNeasy Tissue Kit (Qiagen) or a routine silica-based 96-well extraction automation protocol (Ivanova et al. 2006; Table 1). In addition, DNA samples from King et al. (2008) study were added to the dataset (Fig. 1).

COI sequencing and analyses

To determine which mitochondrial lineages each *A. chlorotica* individual belonged to, we targeted the same portion of the COI gene used by King et al. (2008); this gene fragment is homologous to the DNA region proposed as a standard DNA barcode for animals (Hebert et al. 2003). All samples, except samples from King et al. (2008) study, were processed at the Canadian Centre for DNA Barcoding (CCDB) in the context of the global earthworm barcoding campaign (EarthwormBOL, Rougerie et al. 2009) part of the International Barcode of Life initiative (iBOL). PCR amplifications and DNA sequencing were performed according to the standard protocols used in CCDB (Hajibabaei et al. 2005) and available at: <http://www.ccbd.ca/pa/ge/research/protocols>. The sequences produced are publically accessible on BOLD (Barcode of Life Data Systems – <http://www.boldsystems.org/>) in the project 'ACGSL' and on Genbank (accessions numbers in Appendix A).

In order to establish the correspondence of the generated sequences to the five lineages (named L1 to L5) of *A. chlorotica* uncovered in the study by King et al. (2008), a local BLAST as implemented in the software BioEdit (Hall 1999) was carried out using a nucleotide database formed by all the haplotypes of the King et al. (2008) study with their respective assignments to lineages L1 to L5. The match with the highest *E*-value was used to assign a lineage to the query sequence.

Relationships among haplotypes were examined using a haplotypic network constructed using a reduced-median algorithm (Bandelt et al. 1995) as implemented in the software NETWORK 4.1.1.1 (www.fluxus-engineering.com/sharenet.htm). A maximum parsimony algorithm was applied to infer the most parsimonious branch connections between the haplotypes detected among all the populations sampled.

Microsatellite cloning and screening

An enriched microsatellite library from tandem TC and TG repeats was constructed following Billote et al. (1999). Extracted DNA of a single individual from St Opportune la Mare (Normandie, France) and belonging to lineage L2 (haplotype h25, see Appendix A) was digested by *Rsa*I, purified using spin columns (QIA quick nucleotide removal, Qiagen) and ligated to *Rsa*I linkers (*Rsa*21: 5'-CTCTTGCTTACGCGTGGACTA-3' and *Rsa*25: 5'-GAGAACGAATGCGCACCTGAT-3'). Ligated genomic DNA was amplified by polymerase chain reaction (PCR) using *Rsa*21 primer before hybridization to 5'-biotinylated-(TC)₁₀ and -(TG)₁₀ probes. Target sequences were captured with streptavidin-coated magnetic particles (MagneSphere paramagnetic particles, Promega). After amplification, DNA fragments were ligated into pCR⁴-TOPO[®] plasmid vector (Invitrogen) and transformed into JM109 competent *Escherichia coli* (Promega). Five hundred and fifty-two positive clones were screened for microsatellites by PCR using combinations of (TC)₁₀ and (TG)₁₀ primers and the standard T3 and T7 primers. DNA inserts from 238 positive clones were amplified using T3 and T7 primers and sequenced using the BigDye version 1.1 dye-terminator system (Applied Biosystems) and visualized using an ABI 310 DNA Analyzer (Applied Biosystems). Primers were designed for 40 loci using the Primer3 software (<http://frodo.wi.mit.edu/primer3/>). A preliminary PCR amplification test was performed and 19 primer pairs were selected for screening on a ABI 310 DNA Analyzer. The forward strand primer of each microsatellite was labelled with one of the four fluorescent dyes (FAM, PET, NED or VIC).

Molecular analyses of microsatellite data

All loci were amplified with the same touchdown PCR procedure using a PTC200 thermocycler (MJ research): 15 μ l final volume, 3 μ l of DNA template diluted to 1:50, 1 $\times$ Fermentas PCR buffer, 0.125 mM of each dNTP, 1.5 mM MgCl₂, 0.5 U of Taq polymerase (Fermentas), 0.25 μ M fluorescent-labelled forward primer and 0.25 μ M non-labelled reverse primer. Touchdown PCR included initial denaturation for 3 min at 94 °C, then 35 s at 94 °C, 45 s at initial temperature (*T_i*, Table 2) decreasing by 1 °C per cycle for 10 cycles, 1 min at 72 °C, 25–30 cycles of 35 s at 94 °C, 45 s at *T_i* – 10 °C, 1 min at 72 °C, with a final elongation of 10 min at 72 °C. The PCR products were analyzed on an ABI 310 sequencer using the LIZ500 size standard (Applied Biosystems). Alleles were scored using GenScan software (Applied Biosystems) and were confirmed manually. Of the 19 primer pairs, 11 were discarded because of ambiguous banding pattern or monomorphism. All studied individuals were genotyped at the remaining eight loci. Individuals not belonging to lineage L2 were genotyped twice at each locus in order to limit genotyping errors.

Statistical analyses of microsatellite data

For two populations of the green morph (Yvetot and Croston), the total number of alleles (*N_{all}*), the average number of alleles per locus (*N_{loc}*), the observed heterozygosity (*H_o*), and the gene diversity (*H_e*) were estimated using the program GENETIX V. 4.05 (Belkhir et al. 2004). The null hypothesis of independence

Please cite this article in press as: Dupont, L., et al., New insight into the genetic structure of the *Allolobophora chlorotica* aggregate in Europe using microsatellite and mitochondrial data. Pedobiologia - Int. J. Soil Biol. (2011), doi:10.1016/j.pedobi.2011.03.004

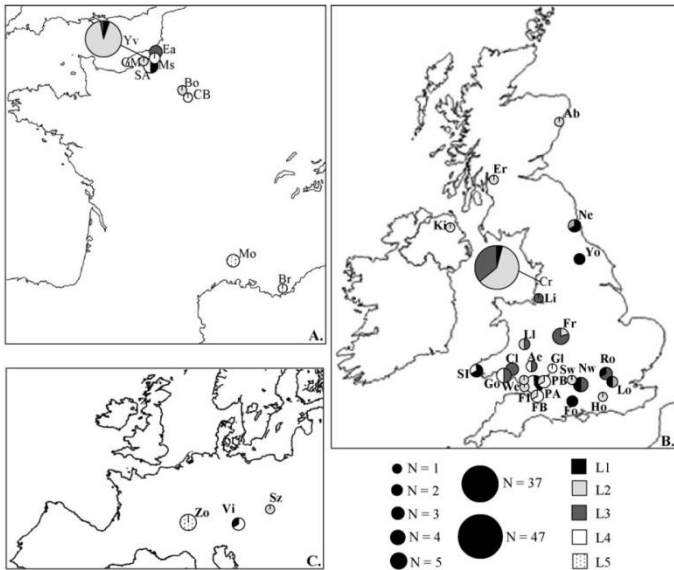


Fig. 1. Location of the study individuals in France (A), United Kingdom (B) and elsewhere in Europe (C). Pie charts display the relative frequencies for a given locality of each of the five mitochondrial (COI) lineages (L1–L5) as reported by King et al. (2008) (location codes are given in Appendix A, and samples from King et al. (2008) are indicated in bold).

Table 1
Characterization of 8 microsatellite loci in *Allolobophora chlorotica*.

Locus	Repeat array	Primer 5' → 3'	T _i	Clone size (bp)	N _{Tall}
Ac127	(CA) ₄ TA(CA) ₁₈ N ₁₄ (CA) ₄	F: VIC-CCGAGCAGTAAGTGAAGCC R: AAACCTACGAGGACGCAACG	61	306	45
Ac170	(TG) ₆ T ₂ (TG) ₄ CGTGCG(TG) ₇ (TCTG) ₄	F: NED-TGGGTTATAAGCATCTGAA R: ATACGATAGGGACCGGTATT	61	179	38
Ac418	(TG) ₄ TA(TG) ₄ N ₆ (TG) ₃	F: FAM-TCTTGTCTGTAATGAAGTGGGT R: TTCACAAGCCTTGCCGAGTA	61	300	25
Ac419	(CA) ₄ CG(CA) ₂ N ₈ (CA) ₂ N ₆ (CA) ₃ N ₄ (CA) ₅	F: PET-GCAGTCACGCTTGTGATAA R: ATACAGACGGCAAGGGAAC	61	257	34
Ac476	(CA) ₆ TA(CA) ₂ CC(CA) ₅	F: PET-AGTAACGATGAAGTAGAGTAG R: AACTTACCATTGTGCTGAC	50	140	16
Ac527	(GA) ₁₀	F: VIC-TACCAGAGGAAAGCACAT R: CACAAATAATCGAGCAAG	56	213	16
Ac528	(AC) ₇	F: VIC-AAGTCTGTGCTGCTCTG R: CAAATGCTCCATCAATAA	61	176	14
Ac529	(AC) ₆ TA(AC) ₂	F: NED-ACACTAACGCACTAACACTCA R: ACAAGAATAGACTGTGGC	56	287	23

Locus name, repeat array, primer sequences, initial temperature (T_i) in touchdown PCR programme, allele size of the sequenced clones and total number of alleles (N_{Tall}) over the whole dataset, are indicated. Clone sequences are available from GenBank under the accession numbers HQ660947–54.

Table 2
Variability parameters for each population and locus: number of alleles (N_{all}), observed heterozygosity (H_o), gene diversity (H_e), estimator of the fixation index F_{IS}($\hat{f}$), result of exact test for deviation from Hardy–Weinberg expectations (*p < 0.05; **p < 0.01; ***p < 0.005). Both populations are only composed of individuals belonging to lineage L2. In bold, single locus values that remained significant after FDR correction.

	Yvetot (N = 34)				Croston (N = 28)			
	N _{all}	H _o	H _e	$\hat{f}$	N _{all}	H _o	H _e	$\hat{f}$
Ac127	13	0.706	0.793	0.112	9	0.786	0.813	0.034**
Ac170	14	0.618	0.802	0.233**	9	0.571	0.769	0.261*
Ac418	12	0.882	0.877	−0.005	8	0.571	0.769	0.261**
Ac419	6	0.471	0.764	0.388**	7	0.429	0.758	0.439***
Ac476	5	0.588	0.660	0.111	5	0.571	0.560	−0.020
Ac527	8	0.588	0.740	0.208*	4	0.464	0.701	0.342
Ac528	9	0.647	0.780	0.173**	7	0.750	0.807	0.072***
Ac529	10	0.647	0.773	0.165	4	0.407	0.621	0.349*
Overall	9.63	0.643	0.774	0.171***	6.63	0.569	0.725	0.218***

Please cite this article in press as: Dupont, L., et al., New insight into the genetic structure of the *Allolobophora chlorotica* aggregate in Europe using microsatellite and mitochondrial data. Pedobiologia – Int. J. Soil Biol. (2011), doi:10.1016/j.pedobi.2011.03.004

between loci was tested from statistical genotypic disequilibrium analysis using GENEPOP V 3.4 (Raymond and Rousset 1995). To adjust for multiple comparisons, the FDR method (Benjamini and Hochberg 1995) as implemented in the software SGoF (<http://webs.uvigo.es/acraaj/SGoF.htm>) was applied.

Exact tests for deviations from Hardy–Weinberg equilibrium (HWE) were computed for each locus using GENEPOP V 3.4. Departures from Hardy–Weinberg expectations within populations were quantified by calculating the Weir and Cockerham $\hat{f}$ estimator of the fixation index F_{IS} (Weir and Cockerham 1984) using GENEPOP V 3.4. Null alleles are a common problem with microsatellite loci and can lead to artefactual heterozygote deficiencies. The method of van Oosterhout et al. (2006) as implemented in the software MICRO-CHECKER (<http://www.microchecker.hull.ac.uk/>) was used to estimate the potential frequency of null alleles (r) and to adjust the observed allele and genotype frequencies appropriately. Note that this method, in addition to all other methods for detecting null alleles, assumes that deviations from HWE do not result from other causes (such as Wahlund effect i.e. deficit of heterozygosity attributable to subpopulation structure).

In order to determine if deviation from HWE might be due to inbreeding, we used a method that estimates selfing rate (i.e. the fraction of the population that result from self-fertilization) from the multi-locus correlation structure of a population sample (David et al. 2007) as implemented in the software RMES (<http://ftp.cefe.cnrs.fr>). Partial selfing not only creates heterozygote deficiencies but also generates identity disequilibria, i.e. correlations in heterozygosity among different loci, yielding an excess of multi-heterozygous genotypes (as well as of multi-homozygous genotypes) compared with the random assortment of genotypes across loci under full outcrossing.

To investigate the genetic structure among populations, exact tests of allelic differentiation were carried out between populations using GENEPOP v.3.4 software. Moreover, Weir and Cockerham's (1984) estimator of the fixation index $F_{ST}(\theta)$ was calculated with GENEPOP v.3.4.

To discriminate the lineage patterning within the species, two graphical methods were chosen. An unrooted neighbour-joining tree was constructed with the Populations software Ver. 1.2.28 (<http://bioinformatics.org/~tryphon/populations/>) using the Cavalli-Sforza and Edward's (1967) chord distance estimated between individuals belonging to the 5 mitochondrial lineages. As a complementary approach, a 2D-factorial correspondence analysis was performed using GENETIX. This method allows visual examination of the distribution of the different multilocus genotypes in a two-dimensional space whose principal axes are linear combinations of different alleles at different loci.

Results

DNA barcoding

The same 582 bp fragment of the COI gene used in King et al. (2008) was compared among 153 individuals (89 individuals from this study for which the sequence was long enough and 64 individuals from King et al. (2008) study, see the Appendix A for details). In total, 54 haplotypes were identified (Fig. 1). Five groups of haplotypes corresponding to lineages L1–L5, as defined by King et al. (2008), were created on the basis of the first match in the local BLAST: 17, 14, 7, 10 and 6 haplotypes were obtained respectively.

The relationships among mitochondrial haplotypes were inferred using a haplotype network (Fig. 2). The haplotypes of the lineage L1 occupied a central position in the network. These haplotypes were linked by unresolved torsos and several mutational steps have been inferred, indicating an insufficient sampling cov-

erage. In lineage L2, some haplotypes were closely related while four haplotypes were highly divergent (H27, H28, H53, H52). Three of these haplotypes were observed in the Yvetot population. Similarly, in L4, two haplotypes (H43, H36) were separated by more than 20 mutational steps from the main group of haplotypes. Finally, we report three new haplotypes closer to lineage L5 than to any other lineage (as revealed by the local blast and the haplotypes network analysis), but extremely divergent from any previously obtained haplotypes. Distant by more than 45 mutational steps from the closest known haplotypes, these sequences might represent two new distinct lineages (H48 to H50, and H51; see Fig. 2).

Microsatellite characteristics and within-population polymorphism

Microsatellite marker development in *A. chlorotica* was difficult with only 8 useful markers out of 238 clones (Table 1). The sequenced fragments included imperfect microsatellites with interrupted di-nucleotide repeats. Flanking sequences were scattered with short stretches (from 2 to 3 di-nucleotides) of microsatellite motifs, which made the design of specific primers difficult.

Overall, within the 5 lineages ($N = 166$) the total number of alleles per locus (N_{Tall}) ranged from 14 to 45 (Table 1). The levels of polymorphism of the loci were tested in two groups of individuals belonging to the lineage L2 of *A. chlorotica*. The number of alleles per locus, the observed heterozygosity (H_o) and gene diversity (H_e) for each population and each locus are detailed in Table 2. The mean number of alleles per locus was lower in the Croston population ($N_{Tall} = 6.63$) than that in Yvetot ($N_{Tall} = 9.63$). Gene diversity (H_e) was similar in both populations, ranging from 0.560 to 0.877 (Table 2).

Significant genotypic disequilibria after FDR correction were observed in 4 pairs of loci in the Croston population: Ac127–Ac528, Ac528–Ac529, Ac419–Ac528 and Ac127–Ac476 ($p < 0.001$).

Heterozygote deficiencies, null allele frequency and multi-locus heterozygosity

Heterozygote deficiency, indicated by significant deviation from HWE, was revealed for 3 loci in Yvetot and for 6 loci in Croston, after FDR correction (Table 2). Null alleles are likely to occur at loci Ac170 and Ac419 in both populations ($0.124 < r < 0.133$ and $0.181 < r < 0.193$, respectively) and at loci Ac418 ($r = 0.125$), Ac527 ($r = 0.161$) and Ac529 ($r = 0.159$) in the Croston population. After correction of the data for null alleles, significant departures from HWE were still observed for the locus Ac170 in the Yvetot population and for the loci Ac418 and Ac419 in the Croston population, as well as for the loci Ac527 in Yvetot and Ac127, Ac528 in Croston for which no null allele was detected. This result may partly be explained by an incomplete recovery of null alleles due to size homoplasy (Chapuis and Estoup 2007). Another explanation is that departure from HWE is due to more than null alleles. Apart from null alleles, heterozygote deficiencies at neutral loci can result from inbreeding or Wahlund effect. Heterozygote deficiencies are not congruent among loci, indicating that selfing is unlikely. Moreover, no identity disequilibrium was detected. Thus, even partial selfing seems improbable. Spatial or temporal Wahlund effect may partly explain heterozygote deficiencies. Wahlund effect is expected when F_{ST} values are higher than F_{IS} values (David et al. 2007). The F_{ST} value, estimated between both populations using the uncorrected dataset ($F_{ST} = 0.112$, $p = 0.000$) was lower than F_{IS} values. However, the F_{ST} value, estimated using the corrected dataset ($F_{ST} = 0.098$, $p = 0.000$) was higher than the corrected F_{IS} estimated in Croston ($F_{IS} = 0.046$, $p = 0.000$) and close to the corrected F_{IS} estimated in Yvetot ($F_{IS} = 0.116$, $p = 0.000$). Thus, two factors (null

Please cite this article in press as: Dupont, L., et al., New insight into the genetic structure of the *Allolobophora chlorotica* aggregate in Europe using microsatellite and mitochondrial data. Pedobiologia - Int. J. Soil Biol. (2011), doi:10.1016/j.pedobi.2011.03.004

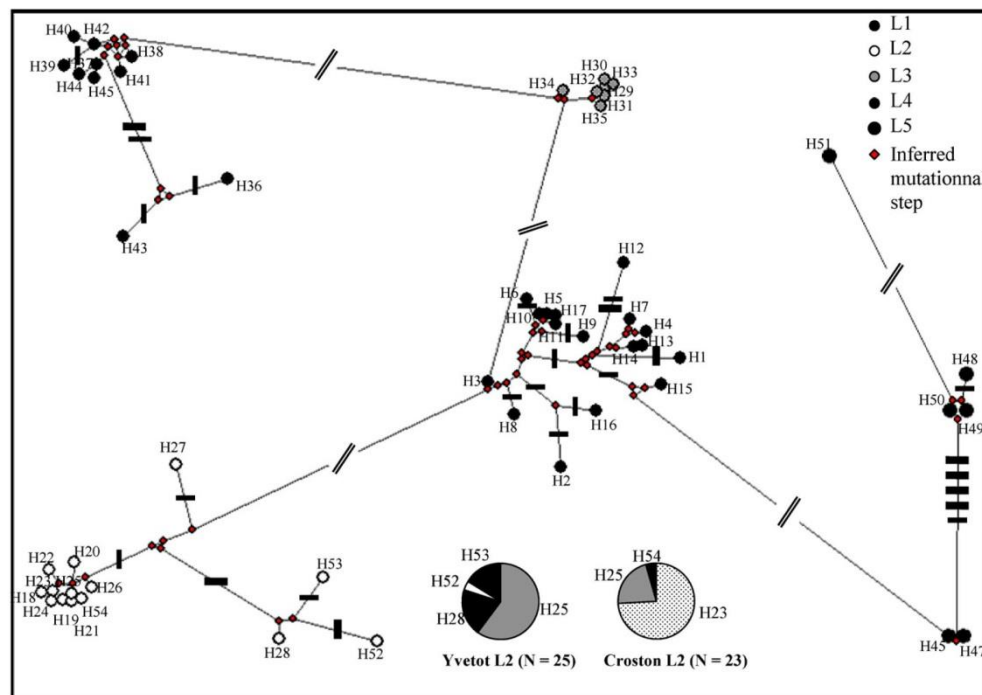


Fig. 2. Haplotype network showing the relationships between haplotypes (H) belonging to the five lineages (L) as reported by King et al. (2008). Unlabelled diamond indicates inferred steps not found in the dataset. Connecting lines show mutational pathways among haplotypes. Grey bars represent 5–9 mutational steps and black bars represent 10 mutational steps. More than 50 mutational steps are indicated by parallel bars. No sign on the line indicates that less than 5 mutational steps are necessary to connect two haplotypes. Pie charts display the relative frequencies of lineage L2 haplotypes in both study populations at Yvetot and Croston. L5 is actually composed of several groups of highly divergent haplotypes that may be named L5 (H46 and H47), L6 (H48, H49 and H50) and L7 (H51).

alleles and Wahlund effect) are likely responsible for the departures from HWE expectations.

individuals are grouped given their sampling locality and not their lineage.

Transferability of microsatellite markers among cryptic lineages

Discussion

Characteristics of newly developed microsatellites

Although all markers amplified in each lineage, percentages of successful amplifications were not equivalent (Table 3). An extremely low level of successful amplification was observed in lineage L1 for locus Ac527 (24%), in lineage L4 for loci Ac419 (10%) and Ac529 (10%) and in lineage L5 for the Ac528 (33%). In lineage L2 and 3, reliable percentage of successful amplification was observed for all loci, except Ac529 in L3 (81%). Thus, comparison between individuals belonging to different lineage was achieved using data from the five microsatellites that amplified well in four of the lineages (L1, L2, L3 and L4): Ac127, Ac170, Ac418, Ac476, Ac528.

The eight microsatellite markers developed in this study were polymorphic in the two populations of *A. chlorotica* grouping L2 individuals but the number of alleles was lower in the Croston population (Table 2). In the Yvetot population, the mean number of alleles per locus ($N_{all} = 9.63$) was comparable to that found in *Lumbricus rubellus* ($N_{all} = 9.75$, Harper et al. 2006) but lower than in the two other earthworm species for which microsatellites have been developed: N_{all} was 12.8 and 12.5 in populations of *Lumbricus terrestris* (Velavan et al. 2007), and *Hormogaster elisae* (Novo et al. 2008), respectively. The gene diversity was also lower in the Yvetot population ($H_e = 0.77$) than in populations of *L. terrestris* ($H_e = 0.85$, Velavan et al. 2007), and *H. elisae* ($H_e = 0.82$, Novo et al. 2008).

The NJ tree and the FCA clearly showed one cluster of individuals belonging to lineages L2 and L3, one cluster of L4 individuals, one cluster of L1 individuals and, to a lesser extent, one cluster of L5 individuals (Figs. 3 and 4A). However, some exceptions were observed. In particular, three individuals of the L3 (L3.18–20) formed a separated cluster. Moreover, some L2 and L4 individuals were observed in the L1 cluster (Fig. 3) and some L5 individuals were observed in the L4 cluster.

Depending on the population screened, null alleles might exist in 2 (25%) or 5 (63%) loci. Occurrence of null alleles was also suggested to explain heterozygote deficiencies in the other earthworm species. Thus, 25%, 40% and 60% of the loci might present null alleles in *L. rubellus* (Harper et al. 2006), *L. terrestris* (Velavan et al. 2007), and *H. elisae* (Novo et al. 2008), respectively. The presence of null alleles has been frequently reported in population genetic studies for a wide range of taxa (Dakin and Avise 2004; DeWoody et al.

Please cite this article in press as: Dupont, L., et al., New insight into the genetic structure of the *Allolobophora chlorotica* aggregate in Europe using microsatellite and mitochondrial data. Pedobiologia - Int. J. Soil Biol. (2011), doi:10.1016/j.pedobi.2011.03.004

Table 3
Percentage of successful amplifications of the eight microsatellite loci in the five mitochondrial lineages defined by King et al. (2008). Total number of allele plus allele size range in base pairs between parentheses, are indicated for each locus. Lowest levels of successful amplifications are indicated in bold.

Loci	Lineage L1 N=21	Lineage L2 N=80	Lineage L3 N=37	Lineage L4 N=20	Lineage L5 N=9
Ac127	100%18 (251–299)	100%18 (266–322)	100%15 (269–334)	100%15 (247–350)	100%4 (247–288)
Ac170	71% 8 (176–206)	100%29 (160–217)	95% 17 (160–217)	100%14 (168–227)	78% 7 (165–227)
Ac418	100%13 (282–308)	100%15 (285–306)	97% 12 (282–306)	95% 11 (291–310)	44% 3 (278–289)
Ac419	95% 18 (259–334)	100%12 (233–290)	100%10 (245–266)	10% 4 (270–313)	89% 8 (257–282)
Ac476	90% 8 (117–165)	99% 7 (117–175)	100% 5 (122–175)	90% 6 (120–150)	89% 3 (114–128)
Ac527	24% 4 (207–211)	100% 9 (197–213)	100%10 (197–217)	50% 5 (195–217)	78% 7 (193–211)
Ac528	100% 6 (170–191)	100%12 (162–191)	97% 8 (162–178)	100% 6 (170–191)	89% 7 (162–191)
Ac529	95% 11 (283–364)	98% 13 (272–356)	81% 9 (274–297)	10% 2 (356–358)	33% 5 (285–354)

2006). Some taxa, such as insects and molluscs, have a particularly high frequency of null alleles (review in Chapuis and Estoup 2007). The association between the presence of null alleles and highly variable flanking regions has also been demonstrated repeatedly (Chapuis and Estoup 2007). In *A. chlorotica*, flanking sequences were scattered with short stretches of microsatellite motif, which may be particularly instable regions and may explain the occurrence of null alleles as well as the difficulty to amplify some loci in other lineages. For instance, the percentage of successful amplification was low for locus Ac527 in lineages L1 (24%) and L4 (50%). Only five loci could be amplified consistently in four of the lineages (L1, L2, L3 and L4). The percentage of successful amplification was only of 44% for one of these loci (Ac418) in L5.

Relationships between *A. chlorotica* lineages

Microsatellite data corroborate the AFLP results of King et al. (2008) showing that the two green morph lineages (L2 and L3) are interbreeding and thus represent a single taxon (Figs. 3 and 4).

Moreover, microsatellite data confirm that pink lineages L1 and L4 are distinct genetic units. In the group of mitochondrial haplotypes pooled within lineage L5 (note that the method used for assigning sequences to lineages forced the assignment to one of the five lineages initially known, independently of the genetic distance observed) haplotypes of individuals sampled in distant localities (Zolikoffen, Montfaucon and Bormes) were highly divergent with 45 to more than 50 mutational steps (Fig. 2). These are thought to represent two additional mitochondrial lineages within *A. chlorotica* and we propose to name them as lineages L6 for haplotypes H48, H49 and H50 from Montfaucon (Fig. 1; Mo) and L7 (haplotype H51 from Bormes (Fig. 1; Br). Microsatellite data confirm the important genetic difference between L5 samples from Zolikoffen (Switzerland) and samples of L6 from Montfaucon (France, Gard). These results stress out the need for more samples coming from different geographical localities to clarify the taxonomic status of the pink morph lineage.

Microsatellite genotypes suggest that hybridization may be possible between pink morph lineages: some L4 individuals were

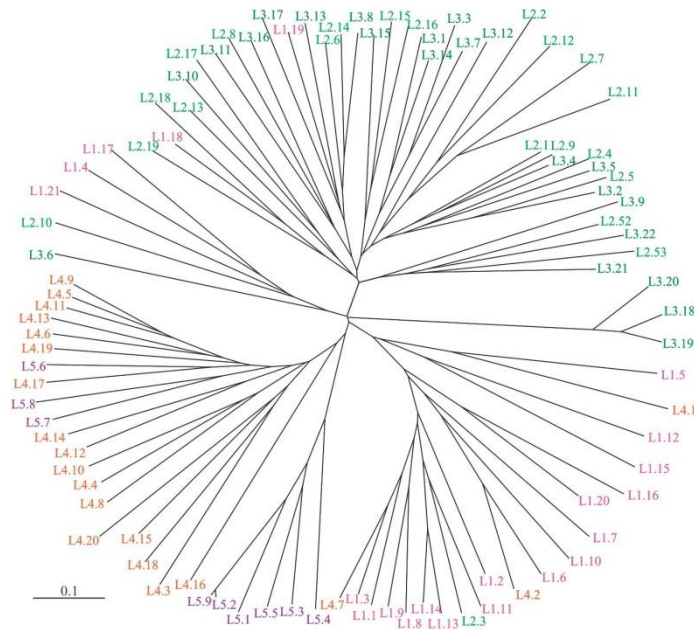


Fig. 3. Unrooted neighbour-joining tree based on the Cavalli-Sforza and Edwards chord distance (D_c) inferred from variation at five microsatellite loci for 93 individuals belonging to the five mitochondrial lineages of *Allolobophora chlorotica*. In the Croston population, only two individuals from lineage L2 and lineage L3 were used for the analysis. Similarly in Yvetot, only 2 individuals of lineage L2 were used.

Please cite this article in press as: Dupont, L., et al., New insight into the genetic structure of the *Allolobophora chlorotica* aggregate in Europe using microsatellite and mitochondrial data. Pedobiologia - Int. J. Soil Biol. (2011), doi:10.1016/j.pedobi.2011.03.004

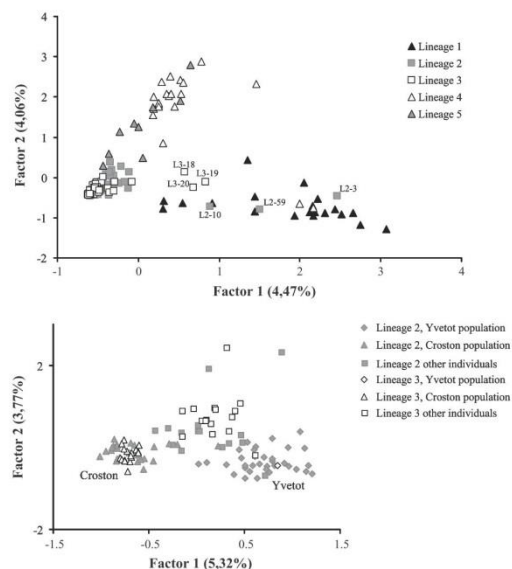


Fig. 4. Plot of the first two axes (factors) of a factorial correspondence analysis (AFC) based on allelic variation (A) at five microsatellite loci for 166 *A. chlorotica* belonging to the five mitochondrial lineages and (B) at eight microsatellite loci for 110 *A. chlorotica* belonging to the green morph (lineages L2 and L3 without the individuals L3.18–20, L2–3, L2–10 and L2–59, label codes are given in appendix A).

observed in the cluster formed by L1 individuals and some L5 individuals were genetically close to L4 individuals (Figs. 3 and 4). Surprisingly, our results also suggest that hybridization between morphs may happen in the field. Indeed, some L2 individuals were genetically close to L1 individuals. Lowe and Butt (2008) demonstrated in the laboratory that hybridization between morphs was possible. Indeed, in pink $\times$ green pairings, both pink and green individuals produced cocoons but cocoons viability was low, in particular for cocoons produced by individuals of the green morph. Moreover, back-crossing of hybrids with pure bred morphs has indicated that hybrids are male sterile, a result suggesting that introgression between morphs is limited. An extensive genetic study of a group of individuals belonging to both colour morphs from the same locality is necessary to confirm that introgression is restricted in nature.

Sub-structure of *A. chlorotica* populations

An important genetic structure was revealed within lineage L2 populations using both mitochondrial and nuclear markers. Several divergent mitochondrial haplotypes were observed for the same lineage, within the same site. For instance, in the L2 individuals from Yvetot, the main haplotype H25 was separated by more than 15 mutational steps to the other haplotypes found in this population (H28, H52 and H53, Fig. 2). Moreover, microsatellite data revealed a Wahlund effect within the Yvetot and Croston populations. Such a result may be explained by the sampling of individuals belonging to different sub-populations that were genetically differentiated. This may be as a result of recent mixing of different populations due to human activities and/or an important genetic structure at small scales due to limited gene flow between sub-populations. In the collembolan *Orchesella cincta*, microsatellite data revealed population genetic substructuring on a scale

of 50 m (van der Wurff et al. 2003). In earthworms, such sub-populations might correspond to patches of individuals (10–50 m wide) revealed in earthworms by geostatistics studies (Rossi et al. 1997; Cannavacciuolo et al. 1998; Margerie et al. 2001). The maintenance of genetic structuring at this fine scale suggests that dispersal capability is limited in *A. chlorotica*.

At a larger scale, such a lack of gene flow results in the genetic isolation of the populations. Thus, the three L3 individuals sampled in Eawy, a large beech forest in Normandie, France, belonged to a distinct cluster on the basis of the microsatellite genotypes. They shared the same COI haplotype (H34), which was not observed in other localities and which was different to the main group of haplotypes observed in this lineage (Fig. 1). This result suggests genetic isolation of this population.

Conclusion

Eight polymorphic microsatellite markers are now available for studying the population genetic structure of the green morph of *A. chlorotica*, and five of them can be used more broadly across the two colour morphs. Our results confirm that the two mitochondrial lineages of the green morph are interbreeding and seem genetically isolated from the pink morph although a few cases of hybridization between morphs have been noticed. With the congruent support from breeding experiments (Lowe and Butt 2008), ecological (Lowe and Butt 2007) and molecular data (this study, King et al. 2008) we consider that the green morph of *A. chlorotica* should be treated as a distinct species. Our work completes King et al. (2008) hypotheses that the pink morph is formed of several distinct lineages whose taxonomic status remains to be elucidated. Our results also show that within a lineage, populations are genetically structured, a result suggesting low vagility and limited gene flow between patches of earthworms. As a consequence, further studies aiming at delimiting taxonomic units within the *A. chlorotica* aggregate need denser geographical sampling, in particular in areas of sympatry between lineages, to properly balance the effect of inherited reproductive isolation and geographic structure caused by the patchy distribution of the earthworm populations and a limited level of gene flow favouring local genetic divergence.

Acknowledgements

This work was funded by the French National Research Agency (ANR) as a part of the project EDISP No. ANR-08-JCJC-0023. The authors are grateful to the "Lycée Agricole" of Yvetot for sampling facilities. We thank S. James for collecting and sending specimens as well as S. Barot, M. Berthou, M. Blouin, G. Caro, T. Desjardins, F. Dubs, A. Ferreira, A. Husté, S. Latchoumaya, C. Lecarpentier, N. Mylinh and V. Roy for their contribution to the first EDISP sampling trip. The generation of DNA barcodes at the Canadian Centre for DNA Barcoding was supported by grants from Genome Canada through the Ontario Genomics Institute.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.pedobi.2011.03.004.

References

- Bandelt, H.J., Forster, P., Sykes, B.C., Richards, M.B., 1995. Mitochondrial portraits of human-populations using Median Networks. *Genetics* 141, 743–753.
- Belkhir, K., Borsa, P., Goudet, J., Chikhi, L., Bonhomme, F., 2004. GENETIX 4.05, logiciel sous Windows pour la génétique des populations. Laboratoire Génome, Population, Interactions, CNRS UMR 5000, Université Montpellier II, Montpellier, France.

Please cite this article in press as: Dupont, L., et al., New insight into the genetic structure of the *Allolobophora chlorotica* aggregate in Europe using microsatellite and mitochondrial data. *Pedobiologia - Int. J. Soil Biol.* (2011), doi:10.1016/j.pedobi.2011.03.004

- Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B* 57, 289–300.
- Billote, N., Lagoda, P., Risterucci, A., Baurens, F., 1999. Microsatellite enriched libraries: applied methodology for the development of ISSR markers in tropical crops. *Fruits* 54, 277–288.
- Bouché, M.B., 1972. Lombriciens de France. Ecologie et systématique INRA, Paris.
- Cannavacciuolo, M., Bellido, A., Cluzeau, D., Gascuel, C., Trehen, P., 1998. A geostatistical approach to the study of earthworm distribution in grassland. *Appl. Soil Ecol.* 9, 345–349.
- Cavalli-Sforza, L.L., Edwards, A.W.F., 1967. Phylogenetic analysis. Models and estimation procedures. *Am. J. Hum. Genet.* 19, 233–257.
- Chapuis, M.P., Estoup, A., 2007. Microsatellite null alleles and estimation of population differentiation. *Mol. Biol. Evol.* 24, 621–631.
- Dakin, E.E., Avise, J.C., 2004. Microsatellite null alleles in parentage analysis. *Heredity* 93, 504–509.
- David, P., Pujol, B., Viard, F., Castella, V., Goudet, J., 2007. Reliable selfing rate estimates from imperfect population genetic data. *Mol. Ecol.* 16, 2474–2487.
- DeWoody, J.A., Nason, J.D., Hipkins, V.D., 2006. Mitigating scoring errors in microsatellite data from wild populations. *Mol. Ecol. Notes* 6, 951–957.
- Dupont, L., 2009. Perspectives on the application of molecular genetics to earthworm ecology. *Pedobiologia* 52, 191–205.
- Hajibabaei, M., DeWaard, J.R., Ivanova, N.V., Ratnasingham, S., Dooh, R.T., Kirk, S.L., Mackie, P.M., Hebert, P.D.N., 2005. Critical factors for assembling a high volume of DNA barcodes. *Philos. Trans. R. Soc. B* 360, 1959–1967.
- Hall, T.A., 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp. Ser.* 41, 95–98.
- Harper, G.L., Cesarini, S., Casey, S.P., Morgan, A.J., Kille, P., Bruford, M.W., 2006. Microsatellite markers for the earthworm *Lumbricus rubellus*. *Mol. Ecol. Notes* 6, 325–327.
- Hebert, P.D.N., Cywinska, A., Ball, S.L., Dewaard, J.R., 2003. Biological identifications through DNA barcodes. *Philos. R. Soc. B. Biol. Sci.* 270, 313–321.
- Ivanova, N.V., Dewaard, J.R., Hebert, P.D.N., 2006. An inexpensive, automation-friendly protocol for recovering high-quality DNA. *Mol. Ecol. Notes* 6, 998–1002.
- King, R.A., Tibble, A.L., Symondson, O.C., 2008. Opening a can of worms: unprecedented sympatric cryptic diversity within British lumbricid earthworms. *Mol. Ecol.* 17, 4694–4698.
- Lazrek, F., Velavan, T.P., Mathieu, J., Dupont, L., 2011. Application of molecular genetics to earthworm ecology: current research and promising future directions. In: Karaca, A. (Ed.), *Biology of Earthworms*. Elsevier, Berlin, pp. 279–298.
- Lowe, C.N., Butt, K.R., 2007. Life-cycle traits of the dimorphic earthworm species *Allolobophora chlorotica* (Savigny, 1826) under controlled laboratory conditions. *Biol. Fertil. Soils* 43, 495–499.
- Lowe, C.N., Butt, K.R., 2008. *Allolobophora chlorotica* (Savigny, 1826): evidence for classification as two separate species. *Pedobiologia* 52, 81–84.
- Margerie, P., Decaens, T., Bureau, F., Alard, D., 2001. Spatial distribution of earthworm species assemblages in a chalky slope of the Seine Valley (Normandy, France). *Eur. J. Soil Biol.* 37, 291–296.
- Novo, M., Velavan, T.P., Almodóvar, A., Schulenburg, H., Díaz-Cosín, D.J., Michiels, N.K., 2008. Microsatellite markers for the drought-resistant earthworm *Hormogaster elisae*. *Mol. Ecol. Notes* 8, 901–903.
- Raymond, M., Rousset, F., 1995. GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *J. Hered.* 86, 248–249.
- Rossi, J.-P., Lavelle, P., Albrecht, A., 1997. Relationships between spatial pattern of the endogeic earthworm *Polypheretima elongata* and soil heterogeneity. *Soil Biol. Biochem.* 29, 485–488.
- Rouabah, L., Descamps, M., 2001. Biology of *Lumbricus terrestris*, *Allolobophora chlorotica* and *Dendrobaena pygmaea* (Annelida, Oligochaeta) in eastern Algeria. *Bull. Soc. Zool. Fr.* 126, 49–58.
- Rougerie, R., Decaens, T., Deharveng, L., Porco, D., James, S., Chang, C., Richard, B., Potapov, M., Suhardjono, Y., Hebert, P.D.N., 2009. DNA barcodes for soil animal taxonomy. *Pesqui. Agropec. Bras.* 44, 789–802.
- Satchell, J.E., 1967. Colour dimorphism in *Allolobophora chlorotica* Sav. (Lumbricidae). *J. Anim. Ecol.* 36, 623–630.
- Sims, R.W., Gerard, B.M., 1999. Synopsis of the British Fauna (31)–Earthworms. The Linnean Society of London and the Estuarine and Brackish-Water Sciences Association, London, pp. 1–169.
- van der Wurff, A.W.G., Isaaks, J.A., Ernsting, G., Van Straalen, N.M., 2003. Population substructures in the soil invertebrate *Orchesella cincta*, as revealed by microsatellite and TE-AFLP markers. *Mol. Ecol.* 12, 1349–1359.
- van Oosterhout, C., Weetman, D., Hutchinson, W.F., 2006. Estimation and adjustment of microsatellite null alleles in nonequilibrium populations. *Mol. Ecol. Notes* 6, 255–256.
- Velavan, T.P., Schulenburg, H., Michiels, N., 2007. Development and characterization of novel microsatellite markers for the common earthworm (*Lumbricus terrestris* L.). *Mol. Ecol. Notes* 7, 1060–1062.
- Weir, B.S., Cockerham, C.C., 1984. Estimating F-statistics for the analysis of population structure. *Evaluation* 38, 1358–1370.

Please cite this article in press as: Dupont, L., et al., New insight into the genetic structure of the *Allolobophora chlorotica* aggregate in Europe using microsatellite and mitochondrial data. *Pedobiologia - Int. J. Soil Biol.* (2011). doi:10.1016/j.pedobi.2011.03.004

Diversity **2011**, 3, 112–120; doi:10.3390/d3010112

OPEN ACCESS

diversity

ISSN 1424-2818

www.mdpi.com/journal/diversity

Article

rich: An R Package to Analyse Species Richness

Jean-Pierre Rossi

INRA, UMR CBGP (INRA/IRD/Cirad/Montpellier SupAgro), Campus International de Baillarguet, CS 30016, F-34988 Montferrier-sur-Lez cedex, France; E-Mail: rossi@supagro.inra.fr;
Tel.: +33-4-30-63-04-30; Fax: +33-4-99-62-33-45

Received: 18 January 2011 / Accepted: 10 February 2011 / Published: 16 February 2011

Abstract: The paper describes *rich*, a new R package to perform species richness estimation and comparison. Species richness is the simplest surrogate for the more complex concept of species biodiversity. It is relatively easy to assess although estimations strongly depend on sampling intensity with the consequence that richness estimations should be standardized to perform valid comparisons. The R package *rich* allows such corrections as well as the computation of various statistics and implements different randomization tests to compare cumulative and average species richness of two communities. These tests are useful for ranking sites or communities which is a classical goal in restoration ecology and conservation biology.

Keywords: species richness; rarefaction; bootstrap; randomization test; biodiversity; community ecology

1. Introduction

Species richness is a basic surrogate for the more complex concept of ecological diversity [1]. It is broadly used as a measure of biodiversity with various objectives such as monitoring biodiversity in order to prioritize management or conservation actions [2–4] or design ecological indicators [5,6]. There are many other mathematical indices intended to measure species diversity many of which incorporate species abundance.

However, general agreement among researchers about which index should be used is currently lacking, and this is a major reason for the continued study of species richness in local communities [7,8].

Yet, problems with simple species richness estimation exist [7]. The observed species richness is dependent on sample size and individuals density [9] and this leads to difficulties when comparing communities with, for example, the aim of ranking sites after applying different management options or performing restoration actions. The problem of sample size dependency is broadly acknowledged but the effect of individual density appears perhaps less often considered although it has been very clearly stressed by some authors [9]. The dependence of species richness upon density of sampled individuals is two-fold. First, it requires the use of rarefaction procedures in order to estimate the expected richness of the community with highest density if it was sampled at a density similar to that of the community with lower density [9]. Second, specific statistical tests need to be elaborated in order to compare the former richness estimates while controlling for differences in community densities. The present article focuses on these questions through presenting a new R package entitled *rich* dedicated to species richness estimation and comparison among communities differing—or not—in terms individuals density.

Comparing measures of species richness is an important task and it is not always straightforward. If the cumulative value (*i.e.*, the overall richness of a set of sampling units) is used to describe a community, only one value is available with no estimate of its dispersion. On the contrary, if one use the average value of the set of sampling units, the resulting mean and standard deviation can be computed. In the latter case, comparing communities can be achieved using standard tests like t-tests or anova tests provided the corresponding assumptions are met. On the other hand, these classical tests cannot be used when the cumulative species richness is considered and other approaches must be used. Evaluating the respective merits of cumulative and average species richness as measures of biodiversity is beyond the scope of this paper. I will here solely mention the statistical tools available in *rich* to process this type of data.

R is a free statistical software [10] allowing a vast array of data processing and analysis. Various packages are developed by users and constitute a efficient way to spread custom methods and functions. The aim in writing *rich* was to provide a simple tool for statistical analysis of species richness data. The package provides functions to perform rarefaction [9], bootstrap re-estimation of richness [11] and richness statistical comparison by means of a randomization test [12] with, or without, controlling for differences in individuals density.

2. General Overview of *rich*

rich comprises different functions that process data in the form of a matrix with species as columns and sampling units as rows. Some of these functions necessitate the R packages *boot* [13] and *vegan* [14]. The package also contains different data sets that can be used to illustrate its capacities. For instance the data sets *ef* and *ea* describe the abundance of soil macrofauna species in a set of soil samples collected in a secondary forest and a cultivated plot in French Guiana, respectively [15]. These data will be used in the examples examined below.

Rarefaction curves traditionally report the average values of randomized species richness derived from resampling without replacement. A consequence of this is that the variance of the species richness estimation among randomizations decreases with sample size and equals 0 at the right-hand of the curve. As a consequence, such estimations cannot be used to compare different data sets. Sampling with replacement, on the contrary, provides meaningful variance of average species richness and thus

allows comparison between different data sets. This corresponds to performing bootstrap on species richness and allows useful additional computations (bias, confidence intervals) [12].

3. Species Richness Computations

The function `rich` processes a species $\times$ sample data matrix and returns various statistics: the bootstrap estimates of species average and cumulative richnesses, their associated bias ([12], p. 36) and confidence interval.

```
> data(ef)
> test<-rich(matrix=ef, nrandom=499, verbose=TRUE)
> test$scr # observed cumulative species richness
[1] 121
> test$mr # observed mean value of species richness over the n samples
[1] 10.4
```

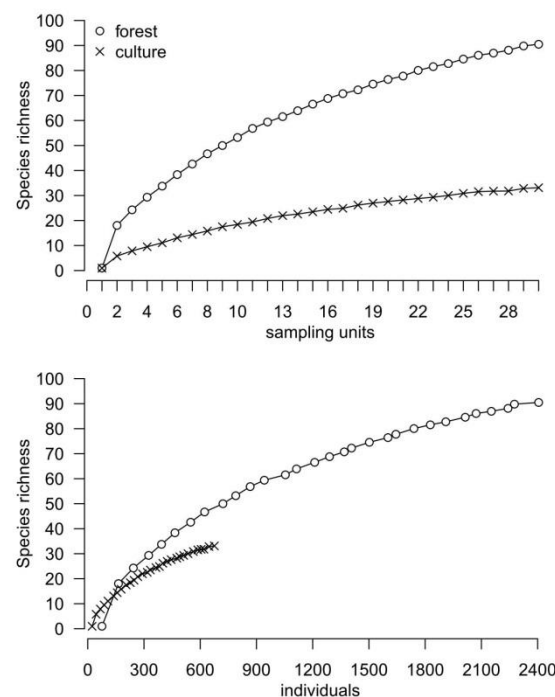
`rich` returns useful information about rare species: the number of singletons and doubletons (species with at most one or two individual(s), respectively) and the number of uniques and duplicates (species encountered in only one or two sample(s), respectively). `rich` also returns the total number of zeroes in the raw data table *i.e.*, empty cells, which value is directly related to the difference between the average and the cumulative species richness.

4. Rarefaction

4.1. Rarefaction Curves

Rarefaction curves are important diagnostic tools that consist of the plot of randomized richness against the sampling intensity. `rich` operates using resampling with replacement so that the variance among randomizations remains meaningful for large number of sampling units or individuals (*i.e.*, right-hand of rarefaction curves) and thus can be used to compare richnesses. The sampling intensity may be represented by the number of sampling units and may be rescaled to individuals [9]. The function `rarc` provides such rarefaction data in the form of a data frame which can be used to draw rarefaction curves. Figure 1 illustrates the rarefaction curves based on the number of sampling units (upper panel) and individuals density (lower panel) for species richness of soil macrofauna in two contrasted land-uses, a forest and a cultivated plot in French Guiana (details in [15]). It appears that species richness is higher in the forest. However, because richness depends on the density of the individuals, rarefaction curves may be rescaled to density such as the lower panel in Figure 1. Interestingly, it can be seen that for low densities the cultivated plot might have been considered more diverse than the forest.

Figure 1. Rarefaction curves for species richness of soil macrofauna in two study plots in French Guiana. Upper panel: sample-based rarefaction curves. Lower panel: individual-based rarefaction curve. In this example, the individual density is larger in the forest plot (open circles) which also hosts the higher species richness than the cultivated plot (crosses) (data from [15]). These curves are based on the output of the R function `rarc` from the package `rich`. `rarc` performs rarefaction using resampling with replacement.



4.2. Rarefying Species Richness

The comparison of species richness of 2 communities may be biased if the sampling effort strongly differs. For that reason, some authors (e.g., [9]) recommend that species richness of the community surveyed using the largest number of sampling units (or characterized by the largest individuals density) be rescaled to be comparable to the data set describing the second community. Another strategy implies extrapolation rather than rarefaction [8]. `rich` comes with a function called `raref` that interpolates the species richness for a given density on the basis of the rarefaction curve considered above (again, resampling with replacement is used). This is illustrated by analysis the data sets `ef` and `ea` to estimate the species richness of soil macrofauna if a forest plot (`ef`) for a sampling intensity corresponding to the maximum individuals density recorded in a nearby cultivated plot (`ea`) (i.e., 670 individuals). The

corresponding richness is estimated as 48 species whereas an overall forest richness of 121 was recorded (see above).

```
> data(ea) # culture plot
> forest<-raref(matrix=ef, dens=sum(ea), nrandom=500)
> forest$Sinterp[2]
[1] 48.09114
```

A second function, `raref2` allows rarefying species richness using bootstrap. For a given density threshold D and a given tolerance t , a subset of the rows (*i.e.*, sampling units) of the community matrix is randomly selected under the constraint that the resulting density is comprised between $D - t \times D$ and $D + t \times D$. Using `raref2` with the `ef` data set and a density threshold of $d = 670$ led to a rarefied richness of 51 species. This method additionally provides a bootstrap standard deviation of the mean richness (10.8 in the example).

```
> raref2(matrix=ef, dens=sum(ea), tolerance=0.01, nrandom=999)
$mean.boot
[1] 51.13614
$sd.boot
[1] 10.76379
```

5. Comparing Cumulative Species Richnesses

Comparing cumulative species richnesses is not straightforward because each community is described by only one value and consequently usual statistical tests do not apply. For that reason, published papers often consider average richnesses and this point is discussed below. The test implies evaluating the null hypothesis “there is no difference between observed richnesses but sampling fluctuations” against an alternate hypothesis “richnesses are different”. The nature of the alternate hypothesis will, as usual, determine whether the test is uni or bilateral. `rich` offers two functions to perform such hypothesis testing.

`c2cv` is dedicated to simple cases where communities are considered as comparable without rarefaction. Let S_1 and S_2 be the richness of community 1 and 2. The difference between these values, $d = S_1 - S_2$ is computed and compared to n similar differences d_{rand} obtained after randomizing samples between communities. The principle of this test is simple and documented in ([12], p. 7). The function `c2cv` reports a variety of results including the randomized values of d . This test rests on the observed value of d as compared to the quantiles of the corresponding randomized values of a user-fixed probability level. Using the function `c2cv` with $n = 999$ randomizations to compare the forest and the cultivated sites examined above leads to an observed difference of 76 species deemed highly significant:

```
> data(efea)
> c2cv(com1=efea$ef, com2=efea$ea, nrandom=999, verbose=FALSE)
$res
cv1          121.000
cv2           45.000
```



```
cv1-cv2          76.000
p                0.001
quantile 0.025   -32.025
quantile 0.975    32.050
```

The function `c2rcv` operates in a very similar way but incorporates a procedure allowing to control for the effect of differences in density of community 1 and 2. The richness associated to the data matrix with the higher density is first rarefied so as to get an estimate of its richness for a number of individuals comparable to that of the community with lower density. A randomization test similar to the one above is performed on the difference between the observed richness of the community with lower density and the rarefied richness of the community with higher density. Using this function with the example of the forest and the cultivated plots examined above leads to similar conclusion that the forest hosts a higher diversity:

```
data(efea)
ex<-c2rcv(com1=efea$ef,com2=efea$ea,nrandom=999,tolerance=0.01)
$dmean
[1] 10.35035
$q1 # quantile for p=0.025
[1] -68.05
$q2 #quantile for p=0.975
[1] -4.95
```

The package `rich` includes a function `c2m` which enables the statistical comparison of average species richness over 2 sets of samples. Comparing average species richness may be problematic in cases where standard statistical tests cannot be used safely because some assumptions are violated. This is the case, e.g., for Student t-test when samples have different sample variances. A practical solution is to use randomization tests [12]. The function `c2m` allows to perform such randomization tests to compare the two sets of sampling units describing the species richness of two communities. If we come back to the example of the species richness in the forest and the cultivated plot, the former appears to have a higher average number of species per sampling unit than the latter:

```
> data(efea)
> x<-rich(efea$ef,nrandom=50,verbose=TRUE)
> y<-rich(efea$ea,nrandom=50,verbose=TRUE)
> c2m(pop1=x$sumrow,pop2=y$sumrow,
+ nrandom=99,verbose=FALSE)
$res
mv1          10.40000000
mv2           3.46666667
mv1-mv2       6.93333333
p             0.01000000
quantile 0.025 -2.06666667
quantile 0.975  2.27666667
```

6. Shared Species

The function `shared` computes the absolute and relative number of species shared by a set of communities. Each community must be described by a matrix as a component of a R `list`. `shared` returns a data frame where the observed community richness is given on the diagonal, the shared species is given above the diagonal and the total richness of the pooled sites is given below the diagonal. This leads to a synthetic table such as the one presented below and e.g., in [11,15].

```
> data(efeb)
> shared(efeb)
      ef eb
ef 121  9
eb 134 22
```

7. Discussion and Perspectives

7.1. Rarefaction

There are many indices designed to describe biodiversity but the question of comparing these measures has not received the attention it deserves. The aim in writing `rich` was to provide a tool to perform simple richness comparisons with randomization procedures. The rarefaction operations were implemented because they are often considered useful when comparing communities sampled with very different intensities. The package provides a function (`raref2`) based on a bootstrap procedure to estimate the species richness. Its main interest is that bootstrap additionally provides an estimate of species richness statistical dispersion (*i.e.*, bootstrap standard deviation).

In some cases however, the difference in individuals density is directly reflecting the differences in community structure rather than effects of sampling regimes. In the example of the soil macrofauna richness in the secondary forest and the cultivated plot, strong differences in densities as well as in richness were observed. The first effect of cutting down the forest and settling a cultivated plot is the destruction of many micro-habitats which leads to the decrease of the density of most populations as well as the extinction of numerous species [15]. In this example, it is interesting to consider that for very low densities, the rarefied richness of the forest would appear to be lower than the former which is obviously a wrong conclusion.

Such differences in the rarefaction curves may be explained by very marked differences in species spatial patterning *i.e.*, aggregation and environmental heterogeneity leading to more species being sampled at low sampling intensity in the cultivated plot. The point here is that such different systems as an old secondary tropical forest and a cultivated plot differ by many ecological aspects amongst which the overall organism density and diversity. Controlling for possible confounding density effects may not always be adequate and might even be a confounding operation in itself.

7.2. Average Richness versus Cumulative Richness

`rich` offers two different ways of comparing the richness associated to two sets of sampling units. These comparisons can be made either on cumulative or on average richnesses. This point is important

because differences between communities may depend to a large extent on the the distribution of species among sampling units. In other words, phenomena such as species spatial aggregation or the proportion of rare species may lead to marked differences of our appreciation of the differences in richness of two communities according to the measure considered. The R package `rich` provides functions to perform such tests and allows to perform rarefaction when using cumulative richness. Future versions of the package may extend the procedure to more than 2 samples comparison and to genetic data so as to allow analyses of allelic richness.

8. Availability

Stable version is available from CRAN: <http://cran.r-project.org/mirrors.html>. Development version is available from R-Forge <https://r-forge.r-project.org/projects/rich/>. Both versions can be installed directly from R. `rich` is distributed under the GNU General Public Licence.

Acknowledgements

The author is grateful to R-Forge for hosting `rich` and to three anonymous reviewers who provided helpful comments on earlier versions of this note.

References

1. Magurran, A. *Measuring Biological Diversity*; Blackwell Science: Oxford, UK, 2004.
2. Kerr, J. Species richness, endemism, and the choice of areas for conservation. *Conserv. Biol.* **1997**, *11*, 1094-1100.
3. Caro, T.; O'Doherty, G. On the use of surrogate species in conservation biology. *Conserv. Biol.* **1999**, *13*, 805-814.
4. Mathieu, J.; Rossi, J.P.; Mora, P.; Lavelle, P.; Martins, P.F.D.S.; Rouland, C.; Grimaldi, M. Recovery of soil macrofauna communities after forest clearance in Eastern Amazonia, Brazil. *Conserv. Biol.* **2005**, *19*, 1598-1605.
5. Rossi, J.P.; van Halder, I. Towards indicators of butterfly biodiversity based on a multiscale landscape description. *Ecol. Indic.* **2010**, *10*, 452-458.
6. Rossi, J.P. Extrapolation and biodiversity indicators: handle with caution! *Ecol. Indic.* **2011**, doi: 10.1016/j.ecolind.2010.09.002.
7. Hellmann, J.; Fowler, G. Bias, precision, and accuracy of four measures of species richness. *Ecol. Appl.* **1999**, *9*, 824-834.
8. Colwell, R.K.; Coddington, J.A. Estimating Terrestrial Biodiversity through Extrapolation. *Phil. Trans.R. Soc. Lond. B* **1994**, *345*, 101-118.
9. Gotelli, N.; Colwell, R. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecol. Lett.* **2001**, *4*, 379-391.
10. R Development Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2009.
11. Rossi, J.P.; Mathieu, J.; Cooper, M.; Grimaldi, M. Soil macrofaunal biodiversity in Amazonian pastures: Matching sampling with patterns. *Soil Biol. Biochem.* **2006**, *38*, 2178-2187.

Diversity **2011**, *3***120**

12. Manly, B. *Randomization and Monte Carlo Methods in Biology*; Chapman & Hall: London, UK, 1997.
13. Canty, A.; Ripley, B.D. boot: Bootstrap R (S-Plus) Functions, 2009; R package version 1.2-41.
14. Oksanen, J.; Blanchet, F.G.; Kindt, R.; Legendre, P.; O'Hara, R.B.; Simpson, G.L.; Solymos, P.; Stevens, M.H.H.; Wagner, H. vegan: Community Ecology Package, 2010; R package version 1.17-3.
15. Rossi, J.P.; Celini, L.; Mora, P.; Mathieu, J.; Lapied, E.; Nahmani, J.; Ponge, J.F.; Lavelle, P. Decreasing fallow duration in tropical slash-and-burn agriculture alters soil macro-invertebrate diversity: A case study in southern French Guiana. *Agric. Ecosyst. Environ.* **2010**, *135*, 148-154.

© 2011 by the author; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>.)

Biol Invasions (2013) 15:899–910
DOI 10.1007/s10530-012-0338-2

ORIGINAL PAPER

Biological invasions in soil: DNA barcoding as a monitoring tool in a multiple taxa survey targeting European earthworms and springtails in North America

David Porco · Thibaud Decaëns · Louis Deharveng · Samuel W. James · Dariusz Skarżyński · Christer Erséus · Kevin R. Butt · Benoit Richard · Paul D. N. Hebert

Received: 8 October 2011 / Accepted: 31 August 2012 / Published online: 12 September 2012
© Springer Science+Business Media B.V. 2012

Abstract Biological invasions are increasingly recognized as a potent force altering native ecosystems worldwide. Many of the best documented cases involve the massive invasions of North America by plant and animal taxa native to Europe. In this study, we use DNA barcoding to survey the occurrence and genetic structure of two major groups of soil invertebrates in both their native and introduced ranges:

Collembola and earthworms. Populations of ten species of earthworms and five species of Collembola were barcoded from both continents. Most of these species exhibited a similar genetic structure of large and stable populations in North America and Europe, a result supporting a scenario of multiple invasions. This was expected for earthworm species involved in human economic activities, but not foreseen for Collembola species de facto unintentionally introduced. This study also establishes that invasive species surveys employing DNA barcoding gain additional resolution over those based on morphology as they

Electronic supplementary material The online version of this article (doi:10.1007/s10530-012-0338-2) contains supplementary material, which is available to authorized users.

D. Porco (✉) · T. Decaëns · S. W. James · B. Richard
Laboratoire d'Ecologie, EA 1293 ECODIV, FED
SCALE, Bâtiment IRESE A, UFR Sciences et
Techniques, Université de Rouen, Place Emile Blondel,
76821 Mont Saint Aignan, Cedex, France
e-mail: david.porco.gm@gmail.com

D. Porco · P. D. N. Hebert
Canadian Centre for DNA Barcoding, University
of Guelph, 50 Stone Road East, Guelph,
ON N1G 2W1, Canada

L. Deharveng
Museum National d'Histoire Naturelle,
UMR7205 «Origine, Structure et Evolution de la
Biodiversité», 45 rue Buffon, CP50, 75005 Paris, France

Present Address:

S. W. James
Department of Biology, University of Iowa,
Iowa City, IA 52242, USA

D. Skarżyński
Department of Evolutionary Biology and Ecology,
Wrocław University, Przybyszewskiego 63/77,
51-148 Wrocław, Poland

C. Erséus
Department of Biology and Environmental Sciences,
University of Gothenburg, Box 463, 405 30 Göteborg,
Sweden

K. R. Butt
School of Built and Natural Environment, University
of Central Lancashire, Preston PR1 2HE, UK

B. Richard
Laboratoire BioSol, Esitpa-Ecole d'Ingénieurs en
Agriculture, 3 rue du Tronquet, 76134 Mont-Saint-
Aignan, Cedex, France

allow evaluation of cryptic lineages exhibiting different invasion histories.

Keywords Invasion survey · Underground · Molecular taxonomy · Springtails · Lumbricidae · Anthropogenic dispersion

Introduction

Both the pace of non-indigenous species European colonization (Addison 2009), and the success of biological invasions have dramatically increased with the rise of international trade. These biotic exchanges are leading to an increasing homogenization of communities (Olden et al. 2004; Winter et al. 2010), and represent one of the main causes of biodiversity decline (Primack 2000; Wilson 2002). Accordingly, scientific interest in this topic has expanded rapidly over the last 30 years. While soil invasions have generally received little attention, some taxa have recently been studied, e.g. earthworms (Hendrix et al. 2008). This neglect is largely due to the 'taxonomic impediment' (Rodman and Cody 2003): in most soil groups, species identification is a difficult task and specialist expertise is often lacking (Decaëns et al. 2006). In the past, earthworms, among other soil invertebrates, received more attention because they represent the most conspicuous element of the soil biological invasions due to their large body size and their role as soil engineers. However, it is undoubtedly the case that smaller-bodied invaders, having no such perceivable activity, also occur in soil. In this study, the invasive patterns in two model taxa of different size are compared: earthworms and Collembola from their native (European) and invasive (North American) ranges.

Over past decades, there has been increasing documentation of the occurrence and consequences of the invasions of European earthworms in North America (Addison 2009). After Pleistocene glaciations, the extinction of native North American earthworms, except in southern refuges, left many regions of North America devoid of native species at the time of European colonization (Addison 2009), and allowed non-native species spreading (Reynolds 1973). Various human activities have led to recurrent invasions (Cameron et al. 2007; Cameron et al. 2008). The effects on soil ecosystem properties and functioning include major impacts on litter decomposition,

horizon mixing, the cycling of carbon, nitrogen (Burtelow et al. 1998; Groffman et al. 2004; Suárez et al. 2006; Wironen and Moore 2006), phosphorous (Suárez et al. 2004) and calcium (Holdsworth et al. 2008). These changes have major impacts on the plant (Hale et al. 2008), microbial, mesofaunal (Eisenhauer et al. 2007; McLean et al. 2006) and fungal communities (McLean and Parkinson 2000).

By contrast, biological invasions by Collembola and their impact on habitats have been much less documented in spite of their ecological significance (Hopkins 1997). To date, Collembola invasion has only been documented in a few sub-Antarctic Islands (Greenslade and Wise 1984; Gabriel et al. 2001; Greenslade 2002a; Myburgh et al. 2007; Greenslade 2008; Greenslade and Convey 2012), in Australia (Womersley 1939; King et al. 1985; Greenslade et al. 2002; Greenslade 2002b; Oliver et al. 2005; Greenslade 2008) and in New Zealand (Salmon 1941). Shown to competitively exclude certain native species (Convey et al. 1999), invasive Collembola display a greater ability to adapt to rising temperature and so could be favored by climate changes (Slabber et al. 2007; Janion et al. 2009). Disturbance, land use changes (Yeates 1991; Yeates et al. 2000), the use of fertilizers, and high grazing intensity (Oliver et al. 2005) also favors invasive species.

The lack of taxonomic expertise and data on existing invasive taxa in soil biota, emphasize the need for new tools to rapidly assess both local and global situations. In this context, molecular systematic, especially DNA barcoding (Hebert et al. 2003), is valuable for detecting or identifying invasive species of plants (Van de Wiel et al. 2009) and animals (Darling and Blum 2007). Successfully tested in earthworms and Collembola (Rougerie et al. 2009; Decaëns et al. 2012), this approach also allows the identification of juvenile specimens (Richard et al. 2010; Chown et al. 2008), preventing an underestimation of the propagule pressure for invasive species.

This study represents the first multi-taxa application of DNA barcoding to the investigation of biological invasions in soil. The effectiveness of this tool for detection and confirmation of invasive status in these two target groups is tested. In addition, we aim to use the DNA barcoding data to compare genetic structures between native and invaded areas as a proxy of invasion history (number of introductions, propagule pressure).

Materials and methods

Sampling

In total, 941 specimens belonging to 10 species of Lumbricidae and 275 specimens from 5 species of Collembola were collected from several localities in Europe and North America (Supplementary Material Table 1). In addition, 94 sequences of Lumbricidae (AM774234–AM774293, EU035477–EU035492, FJ214209–FJ214213, FJ214215, FJ214218, FJ214221, FJ214222, FJ214227, FJ214230–FJ214232, FJ214234, FJ214235, FJ214237, FJ374776, FJ374777) and one sequence of Collembola (EU869805) were downloaded from GenBank. All the other specimens were processed in the context of the International Barcode of life project (iBOL) for the campaigns ‘Barcoding Collembola’ and ‘Barcoding Earthworms’ (<http://www.earthwormbol.org>). Specimen data are available in the project ‘DATASET-INVSOIL1’ on the Barcode of Life Data System (www.boldsystems.org). Each sequence in these libraries is linked to a specimen identified by taxonomic specialists and deposited as a voucher in a public institution.

Preliminary remarks concerning the species origin

For Lumbricidae the direction of circulation between North America and Europe is clear as it is well established that the invasive species present in North America originated from Europe (Gates 1966). The situation is more difficult to assess for Collembola. The five species that were chosen as models for this study have all been flagged as introduced European species in the literature (Maynard 1951; Stach 1966; Christiansen and Bellinger 1998), but without reference to a broader biogeographical context. Here we first evaluated this hypothesis with all the elements available for the five species of Collembola studied: *Neanura muscorum*, *Orchesella villosa*, *Orchesella cincta*, *Parisotoma notabilis* and *Sminthurinus elegans*.

For *Neanura muscorum*, the hypothesis of a European origin is strongly supported by the fact that all other species of the genus are restricted to Eastern Europe (Bellinger et al. 2011). The situation is more complex for the four other species, as they belong to genera widely distributed across the Holarctic, making the biogeographical arguments less clear.

Eastern Russia is the main center of diversification for *Parisotoma* with the highest number of species (Potapov 2001). Moreover *P. notabilis* was not found in the most oriental part of the genus’ distribution in Eastern Asia (Potapov 2001; Yosii 1977), while it is very abundant in Western Europe. From these chorological data, it can be assumed that *P. notabilis* was introduced from Europe to North America; a phylogenetic framework would nevertheless be useful to investigate this hypothesis further.

Orchesella villosa and *O. cincta* were already claimed to be introduced species in North America (Maynard 1951; Christiansen and Bellinger 1998). Like *Parisotoma*, *Orchesella* exhibits its highest diversity in the old world, with a majority of species found in Europe (Bellinger et al. 2011), by contrast, it is only represented in the western regions. So, these two species are also likely to originate from Europe.

Sminthurinus could be another old world genus with its highest species diversity in Europe (Bellinger et al. 2011). This pattern suggests that the species *S. elegans* used in this study is likely to be an introduced species in North America.

So the European origin is a reliable working hypothesis for the species of Lumbricidae and Collembola examined for this study.

Molecular analysis

DNA extraction was achieved in 50 µL of lysis buffer with proteinase K incubated at 56 °C overnight. It was performed following a standard automated protocol using 96-well glass fiber plates (Ivanova et al. 2006). For Collembola, the entire animal was lysed, and specimens were recovered after DNA extraction using a specially designed work flow (Porco et al. 2010a). The 5’ region of COI used as a standard DNA barcode was amplified using M13 tailed primers LCO1490 and HCO2198 (Folmer et al. 1994). Samples that failed to generate an amplicon were subsequently amplified with a pair of internal primers combined with full length ones LepF1-MLepR1 and MLepF1-LepR1 (Hajibabaei et al. 2006). A standard PCR reaction protocol was used for amplifications, and products were checked on a 2 % E-gel 96Agarose (Invitrogen). Unpurified PCR amplicons were sequenced in both directions using M13 tailed primers (Hajibabaei et al. 2005), with products subsequently purified using Agencourt CleanSEQ protocol and processed using

BigDye version 3.1 on an ABI 3730 DNA Analyzer (Applied Biosystems). Sequences were assembled with Sequencher 4.5 (GeneCode Corporation, Ann Arbor, MI, USA) and aligned by eye using BIOEDIT version 7.0.5.3 (Hall 1999). As we observed no indels in the COI sequences, sequence alignment was unambiguous. Sequences are publicly available on BOLD in the project DATASET-INVSOIL1 and on GenBank (accession numbers AM774234–AM774293, EU035477–EU035492, EU869805, FJ214209–FJ214213, FJ214215, FJ214218, FJ214221, FJ214222, FJ214227, FJ214230–FJ214232, FJ214234, FJ214235, FJ214237, FJ374776, FJ374777, FJ937284, FJ937285, FJ937292, FJ937295–FJ937297, FJ937299, FJ937300, FJ937304–FJ937325, GQ373501–GQ373508, GQ373667–GQ373670, GQ373867–GQ373872, GU013792–GU013814, GU013817, GU013819–GU013825, GU013830, GU013831, GU013834–GU013839, GU014221–GU014232, GU206152–GU206159, GU206161, GU206162, GU206169–GU206173, GU206175, GU206178–GU206182, GU206184, GU206187, GU206190, GU206191, GU206213–GU206239, GU656217, GU656242, GU656408–GU656411, GU656418–GU656423, GU656430–GU656433, GU656446, GU656448–GU656451, GU656502, GU656503, GU656507, GU656612, GU656614–GU656619, GU656621, GU656625, GU656627–GU656629, GU656705, GU656706, GU656709–GU656711, GU656778, GU656886, HM388349–HM388353, HM397729, HM397730, HM397758–HM397761, HM398181, HM398303–HM398307, HM398316–HM398319, HM398981–HM398989, HQ024539, HQ024542–HQ024671, HQ682449–HQ682497, HQ942779, JQ908634–JQ935043, JQ935046–JQ935066, JQ935069–JQ935072, JQ935074–JQ935076, JQ935090, JQ935092, JQ935101–JQ935110, JQ935112, JQ935113, JQ935115, JQ935116, JQ935118, JQ935121, JQ935123, JQ935125, JQ935127, JQ935130–JQ935132, JQ935134, JQ935135, JQ935137, JQ935138, JQ935140–JQ935147, JQ935151, JQ935152, JQ935155, JQ935156, JQ935160–JQ935163, JQ935165–JQ935172, JQ935181–JQ935203, JX008093–JX008098).

Calculations

Distance analyses were conducted with MEGA4 (Tamura et al. 2007) using a Neighbor-Joining algorithm (Saitou and Nei 1987). Distances were estimated using the Kimura-2 parameter model (Kimura 1980).

The robustness of nodes was evaluated through bootstrap re-analysis of 1000 pseudoreplicates. The trees were replotted using the online utility iTOL (Letunic and Bork 2007).

For the five species of Collembola and five species of earthworms for which the sampling effort was comparable in terms of population number on each continent (Supplementary Material Table 1), genetic diversity (haplotype diversity (h), nucleotide diversity (π)) and gene flow/population exchange between the two continents (Nm , F_{st}) were analysed along with descriptive statistics, mismatch distributions (population growth-decline model, Raggedness, R^2 values were estimated), and neutrality tests (Tajima's D , Fu's F_s). All these values were calculated with DnaSP v5.10 (Librado and Rozas 2009).

Results

Some well characterized species among those included in the present study consist of several distinct genetic lineages already found in previous studies (Chahartaghi 2007; King et al. 2008; Pérez-Losada et al. 2009; James et al. 2010), but a few additional lineages were detected here (Table 1). For the remaining species, the COI barcode clusters obtained corresponded to the morphological identifications of the specimens. The neighbor joining trees for both groups, showed North American matches for the different European species and their lineages (Figs. 1 and 2). Although all 15 nominal species of Collembola and Lumbricidae were detected in North America, some of their COI lineages were not (*Parisotoma notabilis* L3; *Allolobophora chlorotica* L1, L3, L4, L5; *Aporrectodea rosea* L3; *Aporrectodea caliginosa* L1; *Lumbricus rubellus* L1; *Lumbricus herculeus*—throughout the text, we will consider *L. terrestris* and *L. herculeus* as lineages to allow one to keep track of the fact that we address entities that were until recently considered as cryptic lineages within *L. terrestris*). A more exhaustive sampling of the target species might reveal more haplotypes than found in this study and possibly more lineages. A clear gap between the interspecific and intraspecific variation was found in both datasets. For Collembola, the mean intraspecific variation was 2.32 % versus 26.15 % interspecific divergence and 21.43 % for interlineage distances in *P. notabilis* (Supplementary Material Table 2; Supplementary

Table 1 COI lineages within nominal species

Species	Number of COI lineages	Lineages names	Source
Earthworms			
<i>Allolobophora chlorotica</i>	5	L1, L2, L3, L4, L5	King et al. (2008)
<i>Aporrectodea caliginosa</i>	3	L1, L2, L3	Pérez-Losada et al. (2009)
<i>Aporrectodea longa</i>	1		
<i>Aporrectodea rosea</i>	4	L1, L2, L3, L4	King et al. (2008)
<i>Dendrobaena octaedra</i>	1		
<i>Dendrodrilus rubidus</i>	1		
<i>Lumbricus castaneus</i>	1		
<i>Lumbricus rubellus</i>	2	L1, L2	King et al. (2008)
<i>Lumbricus terrestris</i>	2	<i>L. terrestris</i> / <i>L. herculeus</i>	James et al. (2010)
<i>Octolasion cyaneum</i>	1		
Springtails			
<i>Neanura muscorum</i>	1		
<i>Orchesella cincta</i>	1		
<i>Orchesella villosa</i>	1		
<i>Parisotoma notabilis</i>	3	L1, L2, L3	Porco et al. (2012)
<i>Sminthurinus elegans</i>	1		

Material Fig. 1). The mean intraspecific variation was 1.47 % in earthworms versus 22.09 % for interspecific divergence and 15.88 % for interlineage distances (Supplementary Material Table 3; Supplementary Material Fig. 2).

Table 2 displays the results of population genetics analyses for the species and lineages having a comparable sampling effort in Europe and North America. For most species, both haplotype and nucleotide diversity are high and equivalent between European and North American populations. There were a few exceptions in genetic diversity evenness between the two continents. The Collembola species *S. elegans*

exhibited lower nucleotide diversity in North America. All Collembola species except *O. villosa* showed a high value of Nm and a low Fst. The earthworm lineages *Ap. rosea* L4 and *Ap. caliginosa* L2 had similar haplotype diversity on both continents but lower nucleotide diversity values in North America. Among those species only *Ap. caliginosa* L2 exhibited a low value of Nm and relatively high Fst. All species with an evenness of both haplotype and nucleotide diversity exhibited a high Nm value and low Fst. For the two groups, comparable mismatch distributions were found (multimodal or unimodal) in both areas (Supplementary Material Figs. 3, 4). Fu's Fs, Tajima's D, raggedness and R2 values were not significant (Table 2), allowing one to reject the hypothesis of recent demographic expansion, thus confirming that these populations are well settled and stable.

Discussion

Barcode efficiency and its relevance for monitoring soil invasions

Our results showed that COI sequence data provide an appropriate mean of detecting biological invasions in two major groups of soil invertebrates where such studies would normally be constrained by the taxonomic impediment. Such DNA barcode analyses will enable large-scale studies requiring substantial numbers of routine identifications. The ability of DNA barcodes to distinguish closely related species in earthworms and Collembola has been established in several prior studies (Hogg and Hebert 2004; Chang et al. 2009; Richard et al. 2010; Porco et al. 2010b). However, this approach also brings a finer resolution than morphology by revealing potential cases of cryptic diversity in well studied species (Table 1). Our study revealed a number of cases in which COI divergences between lineages of a single nominal 'species' were as great as typical interspecific distances (Supplementary Material Figs. 1 and 2), suggesting the likely presence of cryptic species. The specific status of many of these lineages has, in fact, already been confirmed through supplemental molecular or morphological analysis (King et al. 2008; Pérez-Losada et al. 2009; James et al. 2010; Porco et al. 2012). This high discrimination power has sometimes been considered problematic (Armstrong

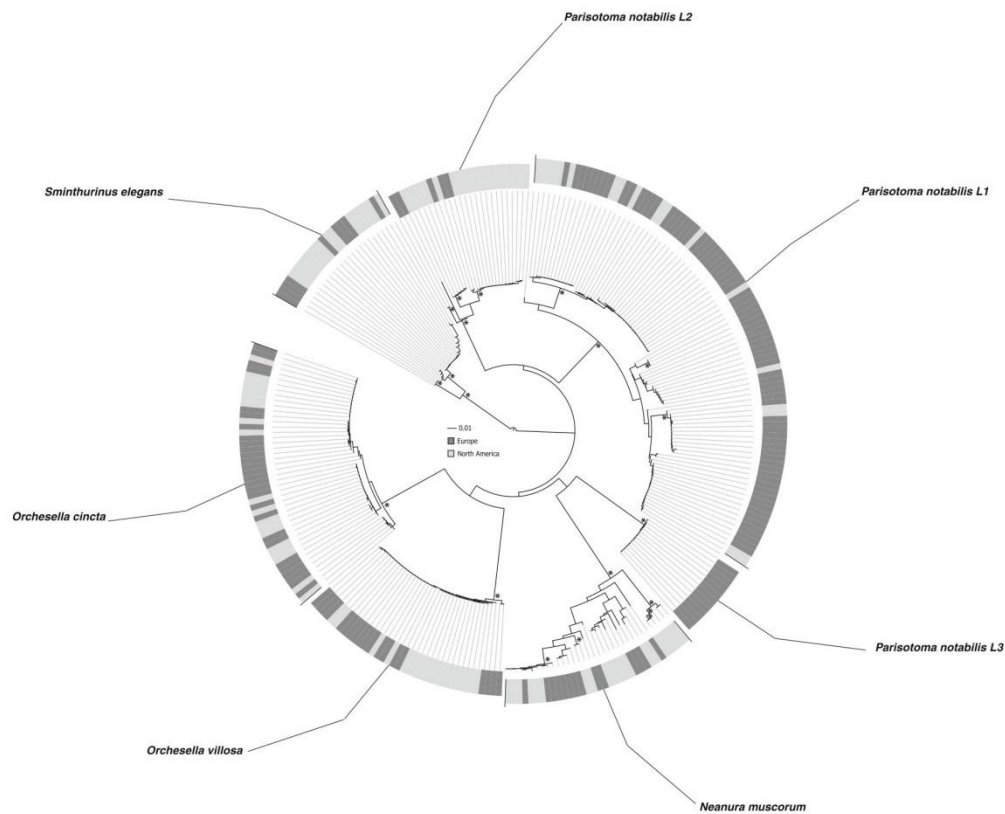


Fig. 1 Neighbor joining K2P tree obtained from 275 specimens among 5 species of Collembola. Bootstrap values $\geq 99\%$ are marked by an asterisk over the branch

and Ball 2005), but this conclusion ignores the growing evidence that most cases of high intraspecific divergence reflect overlooked species or incorrect identifications. By contrast, the importance of considering these cryptic lineages is crucial to gain a proper understanding of biological invasions. Indeed, the COI lineages detected within a single species can exhibit different invasive patterns, occurring sympatrically in some of the localities (Supplementary material Table S1). Consideration of these lineages makes it possible to discern biogeographic trends of invasions that would be overlooked through morphological analysis alone. For example, some lineages of our target species were not detected in North America (*L. herculeus*, *L. rubellus* L1, and *P. notabilis* L3; see Supplementary material Table S1), suggesting potential ecological or biological differences, further

supporting their status as distinct species. It is important to note that the use of provisional names enables the assembly and comparison of distributional data on each of these distinct genetic entities long before the description of any new taxa or the connection of described taxa to their valid name. It was for this reason that we adopted the provisional names for five lineages of *Al. chlorotica* (King et al. 2008) and introduced a similar number system for the lineages in other species.

A second criticism concerning the utility of DNA barcodes for the detection of invasive species involved the current limitations in taxonomic coverage (Darling and Blum 2007). However, barcode coverage is expanding rapidly for numerous taxonomic groups (International Barcode of Life project—<http://ibol.org/>) overcoming this disadvantage. This ongoing effort will

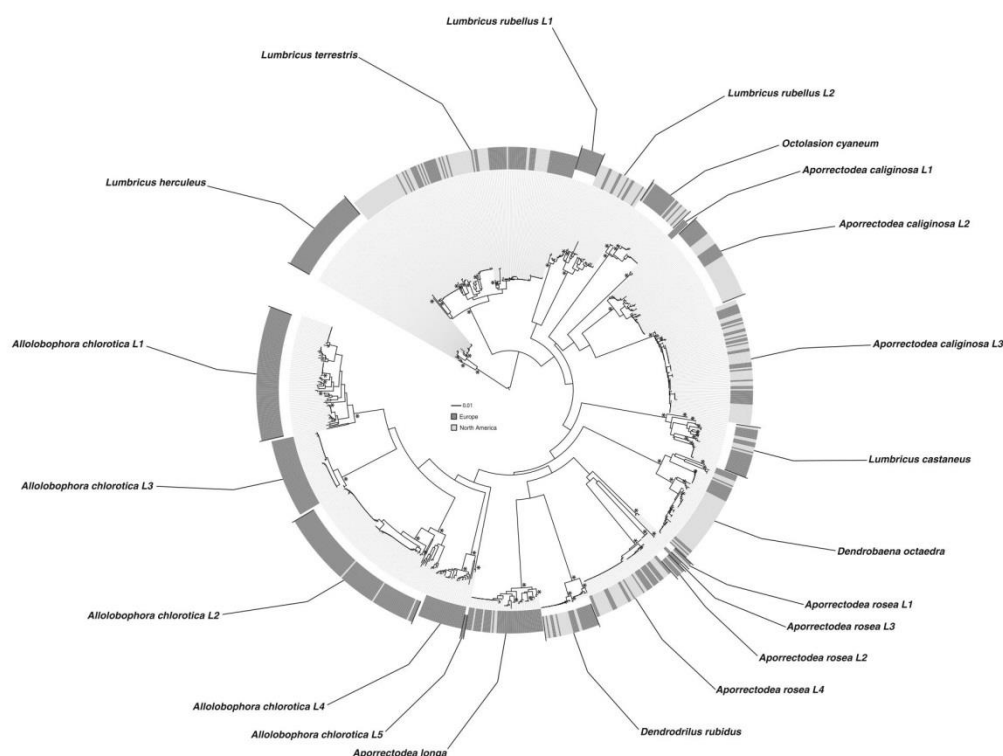


Fig. 2 Neighbor joining K2P tree obtained from 941 specimens among 10 species of Lumbricidae. Bootstrap values $\geq 99\%$ are marked by an asterisk over the branch

undoubtedly increase the range of this approach both geographically and taxonomically, leading to the discovery of many cases of overlooked biological introductions or invasions in all biota. The utility of DNA barcoding in this regard has already been revealed in Lepidoptera (Chown et al. 2008; deWaard et al. 2009), the insect order with the best parameterized barcode library.

Comparing the genetic structure of native and invasive populations

COI data generated for the detection of invasive species can also allow the comparison of their genetic structure with that in their native distribution. Such analyses reveal information on the genetic diversity and demography in the different species for both areas,

providing an indirect assessment of the propagule pressure in the invaded one.

In this study, all species showed similar patterns of mismatch distributions in native and invaded areas. If the mismatch distributions in the European native populations reflect historical events, the similar profiles for introduced populations suggest recurrent introduction events involving large numbers of individuals from across European source populations. Certainly, mutation and drift are unlikely to be responsible for the presence of so much variation given the brief interval since the introduction in North America. We detected erratic mismatch distributions in six species (*D. octaedra*, *L. terrestris*, *Ap. caliginosa* L2, *N. muscorum*, *O. cincta*, *P. notabilis* L1) indicating stable rather than expanding populations over a substantial period of time, a conclusion supported by the high average number of pairwise

Table 2 Genetic diversity indices and neutrality test statistics (N = number of individuals; H = number of haplotypes; N pop. = number of populations; π = nucleotide diversity; h = haplotype diversity; SE = Standard error; Nm = gene flow estimate; Fst = genetic differentiation; D = Tajima's D; Fs = Fu's Fs; r = Ruggedness; R2 = Ramos-Onsins and Rozas R2)

	N	H	N	N	h	(SE)	π	(SE)	Nm	Fst	D*	Fs*	r*	R2*	Mean Pairwise Differences
<i>Lumbricus terrestris</i> Europe	69	23	21		0.893	(0.022)	0.02316	(0.00227)	2.07	0.19421	-1.0598	8.8145	0.0216	0.0758	10.723
<i>Lumbricus terrestris</i> North America	82	14	9		0.836	(0.026)	0.0337	(0.00142)			-0.1632	-0.276	0.1048	0.1021	13.177
<i>Dendrobaena octaedra</i> Europe	34	20	4		0.900	(0.041)	0.04294	(0.00322)	2.56	0.16346	-1.5007	-1.4168	0.0277	0.1371	26.111
<i>Dendrobaena octaedra</i> North America	44	22	5		0.932	(0.023)	0.01985	(0.00443)			0.175	1.6269	0.0292	0.0772	12.071
<i>Aporrectodea caliginosa</i> L3 Europe	33	11	9		0.877	(0.034)	0.00699	(0.00057)	6.8	0.06852	-0.9934	-0.793	0.055	0.0832	4.045
<i>Aporrectodea caliginosa</i> L3 North America	49	8	7		0.832	(0.024)	0.00529	(0.00048)			-0.46247	1.5977	0.0741	0.0949	3.446
<i>Aporrectodea caliginosa</i> L2 Europe	21	10	5		0.862	(0.057)	0.01376	(0.00125)	0.53	0.48655	1.05499	0.327	0.0388	0.1702	6.352
<i>Aporrectodea caliginosa</i> L2 North America	35	8	5		0.709	(0.073)	0.00672	(0.00120)			-0.0802	1.0022	0.0805	0.1221	3.19
<i>Aporrectodea rosea</i> L4 Europe	22	3	7		0.567	(0.051)	0.00379	(0.00025)	1.48	0.2524	2.5208	3.5366	0.5667	0.2603	2.082
<i>Aporrectodea rosea</i> L4 North America	33	3	4		0.322	(0.097)	0.00195	(0.00057)			0.1329	1.8106	0.5024	0.1297	1.038
<i>Parisotoma notabilis</i> L1 Europe	76	14	5		0.874	(0.019)	0.03196	(0.00099)	21.52	0.0227	2.42353	4.215	0.0735	0.1817	9.108
<i>Parisotoma notabilis</i> L1 North America	17	8	7		0.853	(0.066)	0.03121	(0.00572)			1.084	3.052	0.0735	1817	11.235
<i>Neanura muscorum</i> Europe	14	8	3		0.868	(0.068)	0.03301	(0.00640)	3.33	0.13056	-0.5549	3.3187	0.1005	0.1221	16.769
<i>Neanura muscorum</i> North America	20	9	3		0.895	(0.041)	0.03969	(0.00390)			1.2518	3.444	0.0462	0.1943	10.533
<i>Orchesella cincta</i> Europe	28	10	5		0.791	(0.059)	0.00842	(0.00079)	16.35	0.03154	0.82505	0.5175	0.1309	0.152	5.407
<i>Orchesella cincta</i> North America	19	4	5		0.702	(0.065)	0.01105	(0.00094)			1.65115	6.644	0.3378	0.2062	6.187
<i>Orchesella villosa</i> Europe	19	3	4		0.374	(0.130)	0.00114	(0.00046)	0.39	0.55992	-0.8295	-0.085	0.2264	0.0994	5.96
<i>Orchesella villosa</i> North America	18	3	2		0.451	(0.117)	0.00073	(0.00021)			-0.43794	-0.377	0.1362	0.0998	0.477
<i>Sminthurinus elegans</i> Europe	9	3	3		0.6667	(0.105)	0.03519	(0.00534)	0.96	0.34278	2.3682	9.0513	0.4568	0.2674	14.111
<i>Sminthurinus elegans</i> North America	16	2	3		0.500	(0.074)	0.00284	(0.00042)			1.6871	2.7081	0.75	0.25	1

* None of the values of D, Fs, r and R2 were found significant

differences and the results of the neutrality tests. By contrast, we found unimodal mismatch distributions in four species (*Ap. caliginosa* L3, *Ap. rosea* L4, *O. villosa*, *S. elegans*) with low numbers of pairwise differences, suggesting a bottleneck in their source origin (Europe) followed by recent expansion (Supplementary Material Fig. 3 and 4).

A similar level of genetic diversity was found in European and North American populations of most species (Table 2) indicating no bottleneck effects.

Accordingly, the combined native and invasive range genetic structure (Fst) and gene flow estimates (Nm) suggests massive and recurrent introductions of each species, supporting the loss of the “genetic paradox” in soil invasions as already emphasized for aquatic organisms (Roman and Darling 2007).

However, a few species in both groups had less genetic diversity in North America than in Europe (Table 2). For example, *D. octaedra* and *S. elegans* exhibited lower nucleotide diversity in North America suggesting their rapid expansion from a small number of individuals and the loss of variation following colonization due to low effective population numbers. *Ap. rosea* L4 had less nucleotide diversity in North America, but comparable haplotype diversity, suggesting rapid expansion after the introduction of individuals from a few genetically distinct populations. However, even these three species showed low Fst and high population exchange (Nm) between the two continents.

The low nucleotide diversity in North American populations of *Ap. caliginosa* L2 was associated with high Fst values suggesting genetic drift or that the native source of invading populations was not sampled in this study. Native populations of the collembolan *O. villosa* seem to have experienced a bottleneck event as both nucleotide and haplotype diversity values were low. In fact, these values were similar in populations from both Europe and North America, but Fst was high. This divergence likely also reflects the origin of North American populations from parts of Europe that were not sampled in this study.

Our genetic results, indicating either multiple and massive introduction events or expansion patterns can be diversely interpreted for the two target taxa of this study. Human activity likely explains high propagule pressure for some earthworm species that we analyzed. For example, *L. terrestris* and *D. octaedra*,

which are widely used as fishing bait, showed high genetic similarity between European and North American populations, a result consistent with multiple, large-scale introductions implied by such a trade. On the other hand, a species like *Ap. rosea*, not directly useful for any human activity, exhibited a genetic pattern suggestive of the introduction of a small number of individuals. Our data also confirmed the rarity of *Al. chlorotica* in North America (Whalen 2004), a fact that may reflect its low resistance to freezing (Tiunov et al. 2006) or low tolerance to soil acidity (Pearce 1972). However, this rarity may also reflect the absence of economic interest in this species, meaning that introduction events have been accidental, keeping the propagule pressure low. The fact that some individuals have survived suggests that this species could, in time, gain a broader invasive distribution.

The effect of human activity on collembolans is undoubtedly more homogenous as all introductions have been accidental. These unintentional introductions have generally resulted in similar genetic structure on both continents (*P. notabilis*, *N. muscorum*, *O. cincta* and *O. villosa*), suggesting massive and multiple introductions. However, this dispersal almost certainly reflects massive unidirectional exchanges of specimens from Europe to North America in soil used as ship ballast (Vazquez and Simberloff 2001).

Massive parallel sequencing technologies: future application?

The rapid expansion in species coverage for DNA barcode libraries will soon make it possible to employ rapidly-developing environmental sequencing approaches. This transition will enable inexpensive, rapid and accurate surveys permitting extensive screening of invasive species in various target groups. Massively parallelized sequencing will certainly be a useful tool to evaluate the impact of invasive species on the biological diversity of native fauna (Darling and Blum 2007). Because this technology cannot yet produce accurate quantitative data (Amend et al. 2010), propagule pressure cannot be assessed directly. However, the data generated by these techniques for population genetic analysis will allow a rapid indirect appraisal of propagule pressure on broader geographical and taxonomical scales.

Conclusion

With the taxonomic impediment affecting most soil taxa, molecular taxonomy tools such as DNA barcoding, appear to be an appropriate solution for the early, effective and accurate detection of biological invasions, especially for small and understudied taxa such as Collembola. This method also provides an opportunity to consider cryptic diversity in common and well-studied species, a benefit which was here proven critical as the various genetic entities exhibited diverse patterns of invasiveness possibly related to different ecological traits.

In most cases, the genetic structure of North American and European populations in both earthworms and Collembola were similar, fitting consistently with multiple and massive introduction events and thus a high propagule pressure. If this was anticipated in earthworm species involved in human trade, it was less expected in Collembola. These results indicates that passive introductions over a period of time can be as successful as intentional ones, suggesting this could be a general trend in other groups of the soil fauna, especially small body size ones which are easily overlooked.

Facilitation phenomena among invasive taxa, through the alterations of ecosystems, underlie a wider impact on a large scale and should be further investigated. Accordingly, this first study investigating biological invasions conjointly in two soil groups, advocates the necessity of further broad surveys accounting for several taxonomic groups of soil communities at the same time. DNA barcoding has the potential to make a significant contribution with extensive reference libraries available and a rapidly building capacity for those still lacking. Its use can be particularly critical to uncover and quantify 'cryptic invasions' such as that of the Collembola on a large scale. Moreover, this could also be the opportunity to monitor the many taxa rarely sampled in standard surveys.

Acknowledgments This work was supported by grants to PDNH from NSERC and from the government of Canada through Genome Canada and the Ontario Genomics Institute, (SCALE) research federation through the "Functions and Determinants of Biodiversity" (BIODIV) program. D. Porco was supported by post-doctoral fellowships grant from the Conseil Régional de Haute Normandie and from NSERC. C. Erséus was supported by the Swedish Taxonomy Initiative

(ArtDatabanken), and the Adlerbert Research Foundation. S. James was supported by a Marie Curie France Regions fellowship award to the Laboratoire EA 1293 ECODIV at the University of Rouen.

References

- Addison J (2009) Distribution and impacts of invasive earthworms in Canadian forest ecosystems. *Biol Invasions* 11: 59–79
- Amend AS, Seifert KA, Bruns TD (2010) Quantifying microbial communities with 454 pyrosequencing: does read abundance count? *Mol Ecol* 19:5555–5565
- Armstrong KF, Ball SL (2005) DNA barcodes for biosecurity: invasive species identification. *Phil Trans R Soc B-Biol Sci* 360:1813–1823
- Bellinger PF, Christiansen KA, Janssens F (2011) Checklist of the Collembola of the World. In: <http://www.collembola.org>
- Burlow AE, Bohlen PJ, Groffman PM (1998) Influence of exotic earthworm invasion on soil organic matter, microbial biomass and denitrification potential in forest soils of the northeastern United States. *Appl Soil Ecol* 9:197–202
- Cameron EK, Bayne EM, Clapperton MJ (2007) Human-facilitated invasion of exotic earthworms into northern boreal forests. *Ecoscience* 14:482–490
- Cameron EK, Bayne EM, Colman DW (2008) Genetic structure of invasive earthworms *Dendrobaena octaedra* in the boreal forest of Alberta: insights into introduction mechanisms. *Mol Ecol* 17:1189–1197
- Chahartaghi M (2007) Trophic niche differentiation, sex ratio and phylogeography of European Collembola. Ph.D. Thesis
- Chang CH, Rougerie R, Chen JH (2009) Identifying earthworms through DNA barcodes: pitfalls and promise. *Pedobiologia* 52:171–180
- Chown SL, Sinclair BJ, van Vuuren BJ (2008) DNA barcoding and the documentation of alien species establishment on sub-Antarctic Marion Island. *Polar Biol* 31:651–655
- Christiansen K, Bellinger P (1998) The collembola of North America North of the Rio Grande. A taxonomic analysis. Grinnell College, Grinnell, pp 1–1520
- Convey P, Greenslade P, Arnold RJ et al (1999) Collembola of sub-Antarctic South Georgia. *Polar Biol* 22:1–6
- Darling JA, Blum MJ (2007) DNA-based methods for monitoring invasive species: a review and prospectus. *Biol Invasions* 9:751–765
- Decaens T, Jiménez JJ, Gioia C et al (2006) The values of soil animals for conservation biology. *Eur J Soil Biol* 42:S23–S38
- Decaens T, Porco D, Rougerie R et al. (2012) Potential of DNA barcoding for earthworm research in taxonomy and ecology. *Appl Soil Ecol* (in press)
- deWaard JR, Landry JF, Schmidt BC et al (2009) In the dark in a large urban park: DNA barcodes illuminate cryptic and introduced moth species. *Biodivers Conserv* 18:3825–3839
- Eisenhauer N, Partsch S, Parkinson D et al (2007) Invasion of a deciduous forest by earthworms: changes in soil chemistry,

- microflora, microarthropods and vegetation. *Soil Biol Biochem* 39:1099–1110
- Folmer O, Black M, Hoeh W et al (1994) DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Mol Mar Biol Biotech* 3:294–299
- Gabriel AGA, Chown SL, Barendse J et al (2001) Biological invasions of Southern Ocean islands: the Collembola of Marion Island as a test of generalities. *Ecography* 24: 421–430
- Gates GE (1966) Requiem for megadrile utopias. A contribution toward the understanding of the earthworm fauna of North America. *Proc Biol Soc Wash* 79:239–254
- Greenslade P (2002a) Assessing the risk of exotic Collembola invading subantarctic islands: prioritising quarantine management. *Pedobiologia* 46:338–344
- Greenslade P (2002b) Systematic composition and distribution of Australian cave collembolan faunas with notes on exotic taxa. *Helictite* 38:11–16
- Greenslade P (2008) Has survey effort of Australia's islands reflected conservation and biogeographical significance? An assessment using Collembola. *Eur J Soil Biol* 44:458–462
- Greenslade P, Convey P (2012) Exotic Collembola on subantarctic islands: pathways, origins and biology. *Biol Invasions* 14:405–417
- Greenslade P, Wise KAJ (1984) Additions to the Collembolan fauna of the Antarctic. *Trans R Soc S Aust* 108:203–206
- Greenslade P, Simpson JA, Grgurinovic CA (2002) Collembola associated with fungal fruit-bodies in Australia. *Pedobiologia* 46:345–352
- Groffman PM, Bohlen PJ, Fisk MC et al (2004) Exotic earthworm invasion and microbial biomass in temperate forest soils. *Ecosystems* 7:45–54
- Hajibabaei M, DeWaard JR, Ivanova NV et al (2005) Critical factors for assembling a high volume of DNA barcodes. *Phil Trans R Soc B-Biol Sci* 360:1959–1967
- Hajibabaei M, Janzen DH, Burns JM et al (2006) DNA barcodes distinguish species of tropical Lepidoptera. *Proc Natl Acad Sci USA* 103:968–971
- Hale CM, Frelich LE, Reich PB et al (2008) Exotic earthworm effects on hardwood forest floor, nutrient availability and native plants: a mesocosm study. *Oecologia* 155:509–518
- Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp Ser* 41:95–98
- Hebert PDN, Cywinska A, Ball SL et al (2003) Biological identifications through DNA barcodes. *Proc R Soc Lond Ser B-Biol Sci* 270:313–321
- Hendrix PF, Callahan MA, Drake JM et al (2008) Pandora's Box contained Bait: the global problem of introduced earthworms. *Annu Rev Ecol Evol Syst* 39:593–613
- Hogg ID, Hebert PDN (2004) Biological identification of springtails (Hexapoda : Collembola) from the Canadian Arctic, using mitochondrial DNA barcodes. *Can. J. Zool. Rev. Can Zool* 82:749–754
- Holdsworth AR, Frelich LE, Reich PB (2008) Litter decomposition in earthworm-invaded northern hardwood forests: role of invasion degree and litter chemistry. *Ecoscience* 15:536–544
- Hopkin SP (1997) Biology of the springtails (Insecta: Collembola). Oxford University Press, New York, Tokyo
- Ivanova NV, Dewaard JR, Hebert PDN (2006) An inexpensive, automation-friendly protocol for recovering high-quality DNA. *Mol Ecol Notes* 6:998–1002
- James SW, Porco D, Decaëns T et al (2010) DNA barcoding reveals cryptic diversity in *Lumbricus terrestris* L., 1758 (Clitellata): resurrection of *L. herculeus* (Savigny, 1826). *PLoS ONE* 5:e15629
- Janion C, Worland MR, Chown SL (2009) Assemblage level variation in springtail lower lethal temperature: the role of invasive species on sub-Antarctic Marion Island. *Physiol Entomol* 34:284–291
- Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide-sequences. *J Mol Evol* 16:111–120
- King KL, Greenslade P, Hutchinson KJ (1985) Collembolan associations in natural versus improved pastures of the New-England tableland, NSW: distribution of native and introduced species. *Aust J Ecol* 10:421–427
- King RA, Tibble AL, Symondson WOC (2008) Opening a can of worms: unprecedented sympatric cryptic diversity within British lumbricid earthworms. *Mol Ecol* 17:4684–4698
- Letunic I, Bork P (2007) Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* 23:127–128
- Librado P, Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25:1451–1452
- Maynard EA (1951) A monograph of the Collembola or springtail insects of New York state. Comstock Publishing Company Inc, Ithaca
- McLean MA, Parkinson D (2000) Field evidence of the effects of the epigeic earthworm *Dendrobaena octaedra* the microfungus community in pine forest floor. *Soil Biol Biochem* 32:351–360
- McLean MA, Migge-Kleian S, Parkinson D (2006) Earthworm invasions of ecosystems devoid of earthworms: effects on soil microbes. *Biol Invasions* 8:1257–1273
- Myburgh M, Chown SL, Daniels SR et al (2007) Population structure, propagule pressure, and conservation biogeography in the sub-Antarctic: lessons from indigenous and invasive springtails. *Divers Distrib* 13:143–154
- Olden JD, Poff NL, Douglas MR et al (2004) Ecological and evolutionary consequences of biotic homogenization. *Trends Ecol Evol* 19:18–24
- Oliver I, Garden D, Greenslade PJ et al (2005) Effects of fertiliser and grazing on the arthropod communities of a native grassland in South-Eastern Australia. *Agric Ecosyst Environ* 109:323–334
- Pérez-Losada M, Ricoy M, Marshall JC et al (2009) Phylogenetic assessment of the earthworm *Aporrectodea caliginosa* species complex (Oligochaeta: Lumbricidae) based on mitochondrial and nuclear DNA sequences. *Mol Phylogenet Evol* 52:293–302
- Pearce TG (1972) Acid intolerant and ubiquitous Lumbricidae in selected habitats in North Wales. *J Anim Ecol* 41:397
- Porco D, Bedos A, Deharveng L (2010a) Description and DNA barcoding assessment of the new species *Deutonura gibbosa* (Collembola: Neanuridae: Neanurinae), a common springtail of Alps and Jura. *Zootaxa* 2639:59–68
- Porco D, Rougerie R, Deharveng L et al (2010b) Coupling non-destructive DNA extraction and voucher retrieval for small

- soft-bodied Arthropods in a high-throughput context: the example of Collembola. *Mol Ecol Resour* 10:942–945
- Porco D, Potapov M, Bedos A et al (2012) Cryptic diversity in the ubiquitous species *Parisotoma notabilis* (Collembola, Isotomidae): a long used chimeric species? *PLoS One* (in press)
- Potapov M (2001) Isotomidae. Staatliches Museum für Naturkunde Görlitz
- Primack RB (2000) A primer of conservation biology, 2nd edn. Sinauer Associates, Sunderland
- Reynolds JW (1973) The earthworms of Connecticut (Oligochaeta: Lumbricidae, Megascolecidae and Sparganophilidae). *Megadrilogica* 1:1–4
- Richard B, Decaens T, Rougerie R et al (2010) Re-integrating earthworm juveniles into soil biodiversity studies: species identification through DNA barcoding. *Mol Ecol Resour* 10:606–614
- Rodman JE, Cody JH (2003) The taxonomic impediment overcome: NSF's partnerships for enhancing expertise in taxonomy (PEET) as a model. *Syst Biol* 52:428–435
- Roman J, Darling JA (2007) Paradox lost: genetic diversity and the success of aquatic invasions. *Trends Ecol Evol* 22:454–464
- Rougerie R, Decaens T, Deharveng L et al (2009) DNA barcodes for soil animal taxonomy. *Pesqui Agropecu Bras* 44:789–802
- Saitou N, Nei M (1987) The neighbor-joining method—a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4:406–425
- Salmon JT (1941) The Collembolan Fauna of New Zealand, including a discussion of its distribution and affinities. *Trans R Soc N Z* 70:282–431
- Slabber S, Worland MR, Leinaas HP et al (2007) Acclimation effects on thermal tolerances of springtails from sub-Antarctic Marion Island: indigenous and invasive species. *J Insect Physiol* 53:113–125
- Stach J (1966) On the Collembola of Newfoundland and Nova Scotia. *Acta Zool Crac* 11:211–222
- Suárez ER, Pelletier DM, Fahey TJ et al (2004) Effects of exotic earthworms on soil phosphorus cycling in two broadleaf temperate forests. *Ecosystems* 7:28–44
- Suárez ER, Fahey TJ, Yavitt JB et al (2006) Patterns of litter disappearance in a northern hardwood forest invaded by exotic earthworms. *Ecol Appl* 16:154–165
- Tamura K, Dudley J, Nei M et al (2007) MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. *Mol Biol Evol* 24:1596–1599
- Tiunov AV, Hale CM, Holdsworth AR et al (2006) Invasion patterns of Lumbricidae into the previously earthworm-free areas of northeastern Europe and the western Great Lakes region of North America. *Biol Invasions* 8:1223–1234
- Van De Wiel CCM, Van Der Schoot J, Van Valkenburg J et al (2009) DNA barcoding discriminates the noxious invasive plant species, floating pennywort (*Hydrocotyle ranunculoides* L.f.), from non-invasive relatives. *Mol Ecol Resour* 9:1086–1091
- Vazquez DP, Simberloff D (2001) Taxonomic selectivity in surviving introduced insects in the United States. Kluwer Academic/Plenum Publ, New York
- Whalen JK (2004) Spatial and temporal distribution of earthworm patches in corn field, hayfield and forest systems of southwestern Quebec, Canada. *Appl Soil Ecol* 27:143–151
- Wilson EO (2002) The future of life. Vintage Books, New York
- Winter M, Kuhn I, La Sorte FA et al (2010) The role of non-native plants and vertebrates in defining patterns of compositional dissimilarity within and across continents. *Glob Ecol Biogeogr* 19:332–342
- Wironen M, Moore TR (2006) Exotic earthworm invasion increases soil carbon and nitrogen in an old-growth forest in southern Quebec. *Can J For Res—Rev Can Rech For* 36:845–854
- Womersley H (1939) Primitive insects of South Australia. Silverfish Springtails and their allies, Adelaide
- Yeates GW (1991) Impact of historical changes in land-use on the soil fauna. *N Z J Ecol* 15:99–106
- Yeates GW, Hawke MF, Rijkse WC (2000) Changes in soil fauna and soil conditions under *Pinus radiata* agroforestry regimes during a 25-year tree rotation. *Biol Fertil Soils* 31:391–406
- Yosii R (1977) Critical checklist of the Japanese species of Collembola. *Contrib Biol Lab Kyoto Univ* 25:141–170

