



**HAL**  
open science

# Apports de la Génomique Environnementale dans la Caractérisation des Réseaux d'Interactions Ecologiques au Sein des Communautés d'Arthropodes pour une Gestion Durable des Agroécosystèmes

Jean-François Martin

► **To cite this version:**

Jean-François Martin. Apports de la Génomique Environnementale dans la Caractérisation des Réseaux d'Interactions Ecologiques au Sein des Communautés d'Arthropodes pour une Gestion Durable des Agroécosystèmes. Sciences du Vivant [q-bio]. Université de Montpellier - Institut Agro, 2021. tel-04170254

**HAL Id: tel-04170254**

**<https://institut-agro-montpellier.hal.science/tel-04170254>**

Submitted on 25 Jul 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License



Université de Montpellier

Ecole Doctoral GAIA

Dossier de candidature

Habilitation à Diriger des recherches

Apports de la Génomique Environnementale dans  
la Caractérisation des Réseaux d'Interactions  
Ecologiques au Sein des Communautés  
d'Arthropodes pour une Gestion Durable des  
Agroécosystèmes

**JEAN-FRANÇOIS MARTIN**

UMR Centre de Biologie pour la gestion des Populations

Institut Agro | Montpellier SupAgro, INRAE, CIRAD, IRD



## REMERCIEMENTS

Je voudrais ici remercier toutes celles et ceux qui ont permis d'aboutir à ce temps de réflexion sur mon parcours depuis... 20 ans déjà ! à commencer par l'ensemble de ceux qui m'ont fait confiance en m'ouvrant les opportunités qui ont façonné ce parcours, Henri Descimon, Pierre Taberlet, Le CSIRO, l'EBCL et bien entendu Montpellier SupAgro.

Rien de tout ce que j'ai vécu n'aurait eu la même saveur sans les collaborateurs et étudiants nombreux avec qui j'ai eu le plaisir de travailler. Chacune de ces rencontres ont été riches d'enseignements pour ma perception de mon rôle d'encadrant et j'espère leur avoir en retour permis d'avancer dans leur cheminement. Parmi eux j'ai une pensée et des remerciements particuliers évidemment pour celles et ceux avec qui les collaborations ont été les plus intenses, Boris, Marie-Claude, Thibault, Vincent, Manu et Mélodie.

Dans ces remerciements je mets à part mon complice de toujours, André, avec qui c'est toujours le même plaisir de travailler après 25 ans ! Merci d'avoir accompagné, stimulé et relancé mon parcours quand j'en avais besoin, maintenant c'est à toi d'écrire !

Dans le projet de recherche que je mène, rien ne serait possible sans Johannes, sorte de superwoman qui permet de rendre possible l'improbable une pipette à la main, qui ne recule jamais face à la difficulté quand elle est persuadée que c'est la voie la meilleure, qui a pris complètement en charge les développements moléculaires, mais aussi la tâche ingrate de la production des données en environnement exigeant (moi) tout en gardant un niveau d'excellence au quotidien qui m'était inconnu. J'espère que nous avancerons dans ce projet ensemble encore longtemps et je te remercie pour ton engagement sans faille et ta bonne humeur constante qui m'ont gardé sur les rails même au milieu des tunnels.

Je remercie enfin toutes celles et ceux qui m'ont encouragé à soumettre ce document quand je le repoussais sans cesse, que je ne considérais pas que le moment était le mieux choisi. Sans ces encouragements ce document et la réflexion qu'il m'a permis de coucher sur le papier ne seraient pas terminés aujourd'hui. Merci en particulier à Marie-Stéphane, Alex, Vincent et Laure pour leur soutien quotidien dans cette étape.

Pour finir je remercie Laure qui, non contente de me supporter au quotidien (choisis le sens que tu préfères) a littéralement fait en sorte que je puisse aller au bout du projet, en m'encourageant chaque jour de notre vie mais aussi en sacrifiant beaucoup d'énergie et de nuits pour s'occuper de nos chers bambins. Je sais les sacrifices que tu as consentis ces derniers temps en particulier pour que je puisse créer la bulle de quiétude qui m'a permis d'aboutir. Rien n'aurait été possible sans toi.

## TABLE DES MATIERES

Remerciements .....	1
Table des Matieres .....	0
1. Curriculum Vitae .....	1
1.1 Etat civil.....	1
1.2 Cursus professionnel, fonctions exercées, mobilité .....	1
1.3 Titres et formations .....	2
1.4 Compétences acquises par formation.....	3
1.5 Encadrement d'activites de recherche .....	3
1.6 Enseignement .....	5
1.7 Jurys et comités de thèse .....	6
1.8 animation de la recherche et Expertise .....	6
1.9 Obtention de contrats de recherche.....	7
1.10 Dépôt de brevet ou de licence d'exploitation.....	8
1.11 Liste de publications (56) .....	8
1.12 Chapitre d'ouvrage (2) .....	14
1.13 Communications orales et posters.....	14
2. Parcours et Activités de recherche hors projet Principal.....	16
2.1 reconstruire l'histoire évolutive de populations naturelles en déséquilibre (1999-2006).....	16
2.2 Le temps de l'adaptation (2007-2009).....	26
2.3 De la variation génétique à la génomique – une transition methodologique charniere (2009-2013).....	28
3. Le Projet de recherche : La génomique environnementale pour analyser les interactions dans les communautés .....	33
3.1 Le cadre structurel .....	33
3.2 Orientation générale.....	33
3.3 apports des reseaux d'interactions .....	37
3.4 la demarche et les problematiques associées .....	46
4 - Réflexion sur les activités et perspectives .....	56
Bibliographie.....	59

# 1. CURRICULUM VITAE


## 1.1 ETAT CIVIL

### Jean-François Martin

- Maître de conférences Hors Classe à Institut Agro | Montpellier SupAgro depuis le 1<sup>er</sup> janvier 2003 - 18 ans d'ancienneté
- Emploi 02-154, CNECA 2 au Ministère de l'Agriculture et de l'Alimentation (équivalent CNU section 67)
- Né le 19 mars 1973 à Paray-le-Monial (Saône et Loire)
- Nationalité Française
- Situation familiale : Pacs

### Etablissement Actuel :

Institut Agro | Montpellier SupAgro – Département Biologie et Ecologie – UMR Centre de Biologie pour la Gestion des Populations (CBGP) - Campus International de Baillarguet, 34980 Montferrier/Lez

 : +33 768 950 612

 : [jean-francois.martin@supagro.fr](mailto:jean-francois.martin@supagro.fr)

 : [@JFMartinSupagro](https://twitter.com/JFMartinSupagro)

## 1.2 CURSUS PROFESSIONNEL, FONCTIONS EXERCEES, MOBILITE

**2012-2013** **Mise en délégation à Aix-Marseille Université** pendant une année universitaire pour effectuer un changement thématique (voir mémoire).

**2003-actuel** **Maître de Conférences** : Montpellier SupAgro – UMR Centre de Biologie pour la Gestion des Populations (CBGP) - Montpellier

**2002** **Postdoctorat** : EBCL : European Biological Control Laboratory, Campus de Baillarguet Montferrier-sur-Lez : 9 mois (analyse de la diversité en lutte biologique).

**2000-2002** **Postdoctorat** : CSIRO : Commonwealth Scientific and Industrial Research Organisation European Laboratory, Campus de Baillarguet-Montferrier-sur-Lez : 18 mois (mise en place et exploitation d'un laboratoire de biologie moléculaire dans une unité de lutte biologique)

**1999-2000** **Postdoctorat** : Laboratoire de Biologie des Populations d'Altitude (LBPA), UMR CNRS 5553 Université Joseph Fourier- Grenoble – France 10 mois (Phylogéographie des perdrix grises et Phylogénie du genre Capra)

**1999**      **Attaché Temporaire d'Enseignement et de Recherche (ATER)**  
6 mois temps plein Aix-Marseille Université, Marseille

**1997-1999**    **Thèse de doctorat** dans l'EA 2022 Biodiversité (Aix-Marseille  
Université & Stanford University, CA, USA) - 2 ans

---

## ACTIVITES ACTUELLES

**Activités de recherche** au sein de l'UMR Centre de Biologie pour la Gestion des Populations (CBGP) : Génomique environnementale, Analyse moléculaire de réseaux trophiques, Bioinformatique

**Activités d'enseignement** dans le Département Biologie et Ecologie (BE) : Génétique des populations, Evolution biologique et Ecologie animale appliquées à la Gestion des populations, compétences pour la Recherche Scientifique.

## 1.3 TITRES ET FORMATIONS

**1999 : Thèse de doctorat** soutenue le 25 octobre 1999 à Marseille, Formation doctorale Biosciences de l'Environnement et Santé. Titre : « Phylogénies moléculaires : exemples d'applications de l'espèce à la phylogénie des grands taxa ». Aix Marseille Université. Mention très honorable avec félicitations du jury.

Composition du jury :

H. Descimon	Professeur, Université Aix-Marseille I, France	Directeur de thèse
P. Taberlet	DR1, CNRS U5553, Grenoble, France	Rapporteur
W.B. Watt	Professeur, Stanford University, CA, Etats-Unis	Rapporteur
G. Brun	MCF, Université Aix-Marseille I, France	Examineur
E. Faure	Professeur, Université Aix-Marseille I, France	Examineur
P. Pontarotti	DR1, INSERM UI19, Marseille, France	Examineur

**1997 : D.E.A.** « Ecosystèmes Aquatiques, Méditerranéens et Montagnards » Université des Sciences et Techniques de St Jérôme (Aix-Marseille III) Mention Bien.

**1996 : Maîtrise de Sciences Naturelles** Université de Provence (Aix-Marseille I) Mention Assez Bien.

**1995 : Licence de Sciences Naturelles** Université de Provence (Aix-Marseille I) Mention Assez Bien.

## 1.4 COMPETENCES ACQUISES PAR FORMATION

- **Langues** : Ecriture Anglais Scientifique (Lu, écrit, parlé)
- **Organisation et management** : management d'équipe, concepts et outils de productivité, gestion de projet Agile, gestion du temps, lecture rapide.
- **Biologie moléculaire** : Construction de librairie HTS pour des applications diversifiées (metabarcoding, transcriptomique, GBS, ddRAD, capture), design expérimental haut débit, qPCR pour génotypae et quantification relative et absolue.
- **Bioinformatique** : administration système Linux (Ubuntu), notions de programmation (Python, R), versioning (Git), reporting (markdown), containerisation (virtualisation Singularity) en environnement HPC et bioinformatique (metabarcoding, transcriptomique, SNPs).
- **Analyse de données** : Manipulation et visualisation de données (readr, dplyr, tidyr, ggplot2), reporting (Rmarkdown, Jupyter Lab), versioning (Git), analyse de données de réseau écologique.
- **Pédagogie** : Pédagogie en enseignement supérieur, apprentissage par problème, approche par compétences, apprentissage incluant du distanciel.
- **Intégrité scientifique et Science Ouverte** : recherche reproductible, traitement des manquements à l'intégrité scientifique, open access, open data, Gestion des données de la recherche.

## 1.5 ENCADREMENT D'ACTIVITES DE RECHERCHE

### POST-DOCTORANTS :

Les post-doctorants ci-après étaient sous ma responsabilité ou co-responsabilité directe.

- Vincent Lesieur<sup>1,2</sup> : lutte biologique contre une espèce envahissante en Australie, *Sonchus oleraceus* (Asteraceae). CBGP 4 ans de 2017-2020, Aujourd'hui Research Scientist au CSIRO
- Emmanuel Guivier<sup>6,20,29</sup> : Microbiota Diversity Within and Between the Tissues of Two Wild Interbreeding Species. Aix-Marseille Université 3 ans de 2017-2019
- Grégory Mollot : System approach for the TRAnSition to bio-DIVersified agroecosystems. CBGP 2 ans de 2016-2017
- Scott McCairns<sup>7</sup> : Approche pan-génomique de la phylogeographie et de l'adaptation de *Pseudorasbora parva*. CBGP 2 ans de 2015-2016. Aujourd'hui Chercheur à INRAE
- Vincent Lesieur<sup>5,14,15</sup> : Phylogéographie et routes de colonisation d'insectes ravageurs des cultures aux Etats-Unis. CBGP 1 an en 2015. Aujourd'hui Research Scientist au CSIRO



- Emmanuel Corse<sup>8,32</sup> : Evolution du régime alimentaire de poissons cyprinidae. Deux séjours de 1 mois en 2014. Université Aix-Marseille
- Jamie Winternitz: Evolution du CMH chez les rongeurs. 3 mois en 2014. Académie des sciences de République Tchèque, Studenec.
- Andrea Simkova<sup>43,45</sup> : Adaptation du CMH. Plusieurs séjours de 3 mois de 2003 à 2008. Université de Brno, République Tchèque.

---

## DOCTORANTS :

Dans la liste présentée, Mélodie Ollivier correspond à un co-encadrement au sens formel du terme. Pour les autres encadrements il s'agit de participation à leur encadrement dans un aspect de leurs travaux.

- **Mélodie Ollivier**<sup>1,2</sup> (co-direction formelle) : Réseaux d'interaction biologique et implications pour la lutte biologique contre une espèce envahissante en Australie, *Sonchus oleraceus* (Asteraceae). 2017-2020. Montpellier SupAgro, Montpellier. Aujourd'hui enseignante-chercheuse contractuelle à PURPAN
- David Fletcher : Rôle de la plasticité dans l'adaptation génomique de *Pseudorasbora parva*. Encadrement des aspects de Génomique, thèse en cours. 2014-2017. Université de Bournemouth, Grande Bretagne.
- Martina Vyskočilová<sup>40,43,45</sup> : Phylogénie du CMH chez les poissons. Plusieurs séjours de 1 mois pendant sa thèse. 2009-2011. Université de Brno, République Tchèque. Aujourd'hui en post-doctorat à l'université de Prague.
- Adam Konecky : génotypage de rongeurs et évolution du CMH. Plusieurs séjours de 6 mois pendant sa thèse. 2007-2009. Université de Brno, République Tchèque. Aujourd'hui en post-doctorat à Turin, Italie
- Eva Ottova<sup>48</sup> : Phylogénie du CMH chez les poissons. Plusieurs séjours de 6 mois pendant sa thèse. 2005-2007. Université de Brno, République Tchèque. Actuellement Responsable d'un centre de conservation en république tchèque.
- Thibaut Malausa<sup>23,24,25,27,28,30,33,36,41,C1</sup> : Structure génétique des populations de pyrale du maïs. Thèse au CBGP. 2004-2006. Montpellier. Actuellement Chercheur à l'INRAE
- Grégory Molot<sup>17</sup> : Impact des pratiques agricoles sur le réseau trophique dans les bananeraies en Martinique, barcoding environnemental. 2009-2011. Actuellement Auto-entrepreneur.

---

## ETUDIANTS EN MASTER :

- Fanny Bénetière : master 2 : Caractérisation des relations trophiques au sein du cortège d'Arthropodes associés à *Sonchus oleraceus* et implication pour la lutte biologique contre cette adventice invasive en Australie. Partie Europe. (mars-août 2018)
- Maeva Labouyrie : master 1 : Caractérisation des relations trophiques au sein du cortège d'arthropodes associés à *Sonchus oleraceus* et implication pour la lutte biologique contre cette adventice invasive en Australie. Partie Australie. (septembre-décembre 2018) – aujourd'hui doctorante
- Maxime Corbin<sup>2</sup> : master 1 : Caractérisation de la variabilité fonctionnelle et morphologique entre différentes populations de *Sonchus oleraceus*, via une expérimentation en jardin commun. (mars-juillet 2018) – aujourd'hui doctorant.
- Camille Rouvière : master 2 : Mise au point d'une méthode de génotypage haut-débit par l'utilisation de la courbe de fusion haute résolution pour établir la structure de la diversité génétique de *C. assimilis*.
- Boris Fumana<sup>5,48,50</sup> : master 2 : Caractérisation de la structure de la diversité génétique de *Ceuthorrhynchus assimilis* et implications pour le contrôle biologique de *Lepidium draba* – aujourd'hui enseignant-chercheur à l'université de Clermont Auvergne
- Céline Jolivet : master 2 : Phylogeny of the genus *Raphanus* and biological control against Wild radish – aujourd'hui chercheuse au Thünen Institute (Allemagne)

## 1.6 ENSEIGNEMENT

**Presentation.** La majeure partie de mes activités d'enseignant se déroule dans le cadre de la formation initiale des deux cursus d'ingénieurs et du master 3A de l'Institut Agro | Montpellier SupAgro. Ces enseignements se retrouvent dans les trois années de formation jusqu'en option. Le fil conducteur qui structure mon engagement est, partout où c'est possible, de mettre l'accent sur la double entrée Ecologie et Evolution pour appréhender l'organisation et le fonctionnement de la biodiversité à toutes les échelles dans le contexte de transitions vers la durabilité. A travers mes interventions ou mes responsabilités d'organisation, mon objectif général est d'optimiser la cohérence des disciplines liées à l'écologie et l'évolution dans nos parcours et de favoriser la réflexion des élèves sur les interactions biologiques qui régissent non seulement les agrosystèmes mais de façon plus générale les écosystèmes en intégrant divers niveaux d'organisation du vivant et en privilégiant des approches intégratives de l'écologie. Les notions que je souhaite transmettre au cours du cursus des élèves vont de la prise de conscience de la complexité des systèmes biologiques

en première année à l'analyse des interactions biotiques dans les divers modules auxquels je participe et/ou anime par la suite. **Perspectives.** Je pense avoir trouvé ma place dans l'ensemble des collectifs institutionnels et pédagogiques. Ma participation à de nombreux groupes de travail transversaux et mon investissement renouvelé me permettent d'avoir une vision assez claire de ma mission d'enseignement, d'autant qu'elle est cohérente avec mon activité de recherche. L'animation du collectif partout où je suis en responsabilité est une de mes priorités pour optimiser la pertinence du projet pédagogique quand c'est possible, et insuffler à ma modeste échelle la transition pédagogique vers des approches actives de l'apprentissage.

## 1.7 JURYS ET COMITES DE THESE

J'ai participé à différents jurys de recrutement d'ingénieurs en Biologie Moléculaire pour l'INRA, l'université d'Aix-Marseille et l'IRD (respectivement en 2011, 2012 et 2003) ainsi que pour le recrutement d'un administrateur de parc informatique pour l'unité (TR, INRA) en 2012 (voir annexe H). J'ai été membre de la commission de spécialistes de l'université Aix-Marseille entre 2004 et 2007.

J'ai participé à trois comités de thèse pour Pascaline Dumas (2009-2013), Grégory Mollot (2008-2012) et David Fletcher (2014-2018).

## 1.8 ANIMATION DE LA RECHERCHE ET EXPERTISE

---

### REFERENT A L'INTEGRITE SCIENTIFIQUE

A ce titre j'assure les missions définies par le *vademecum* édité par l'OFIS (Office Français d'Intégrité Scientifique) pour la mise en place de la circulaire du MESRI relative à l'Intégrité Scientifique. Je prépare la signature des chartes nationales et européennes, de déontologie scientifique. Je participe à la conférence des signataires et au réseau national des Référents. J'assure la promotion de l'intégrité scientifique dans la collectivité scientifique de l'établissement et je traite les allégations de manquement à l'intégrité scientifique quand elles concernent un membre de cette collectivité. Je rends compte de ces traitements à la direction de l'établissement en indépendance hiérarchique.

Ma lettre de mission pour ces activités : <https://tinyurl.com/HDRJFMmissions>

---

### DELEGUE A LA POLITIQUE SCIENCE OUVERTE

A ce titre, j'ai pour mission d'apporter mon soutien à la Direction déléguée aux formations et à la politique Scientifique de l'établissement dans ce domaine en étant responsable de la veille réglementaire, la coordination de mes actions avec mes homologues au sein des opérateurs de recherche partenaires et j'informe la collectivité scientifique de l'établissement des formations disponibles. J'assure la promotion du développement de bonnes pratiques en matière de recherche et de publication en libre

accès à travers la plateforme HAL de l'établissement. Je participe aux actions de communication sur la diffusion des résultats de la recherche.

Ma lettre de mission pour ces activités : <https://tinyurl.com/HDRJFMmissions>

---

## MEMBRE DU COLLEGE DONNEES DU COMITE NATIONAL POUR LA SCIENCE OUVERTE

[Le Collège Données du comité pour la Science Ouverte](#) a pour missions de faire naître des actions à partir des enjeux identifiés par les communautés de recherche ou les spécialistes, portés par les membres du collège ou issus de veilles diverses ; construire ou faire construire des actions en application des orientations/décisions politiques relayées par le CoSO ; Coordonner les activités des divers groupes et procéder au suivi des groupes projets qui relèvent du Collège. Dans ce cadre je participe aux travaux pléniérs du Collège et je coordonne actuellement le Groupe de Travail sur la diffusion des données de Recherche liées aux publications.

Ma lettre de mission pour ces activités : <https://tinyurl.com/HDRJFMcoso>

---

## EVALUATEUR AGENCE DE FINANCEMENT ET REVIEWING

J'ai été évaluateur pour le Programme-cadre pour la recherche et le développement technologique PCRD7 (2007-2013) et pour l'académie des Sciences de République Tchèque (2005-2015).

Je suis relecteur pour plusieurs publications internationales (6-10 par an) ayant des thématiques de publication autour de la génomique, les approches NGS et la bioinformatique pour résoudre des questions scientifiques dans ma thématique (Mol Ecol Res., BMC bioinformatics, BMC genomics, Plos One, Nucleic Acid Res, Nature Methods...).

### 1.9 OBTENTION DE CONTRATS DE RECHERCHE

Ci-après figure la liste des contrats de recherche obtenus ces dix dernières années ainsi que la somme allouée à mes activités et les collaborations impliquées hors UMR.

**2020- 2023 ANR Cultiver et Protéger Autrement** – MoBiDiv : Mobiliser et Sélectionner la diversité cultivée intra et inter spécifique pour un changement systémique vers une agriculture zéro-pesticide – collaborateur – 35k€ - *UMR AGAP, Montpellier*

**2017-2018 Agropolis Fondation - CAPTURE:** Long Sequence DNA CAPTURE for agrobiodiversity studies – 12k€ - UMR DIADE, Montpellier

**2016-2020 Collaborative Research Agreement CSIRO:** new biocontrol solutions for sustainable management of weed impacts to agricultural profitability) – co-leader – 435k€ - *CSIRO European Laboratory*

**2014-2018 Agropolis Fondation** - Projet Etendard STRADIV: System approach for the transition to biodiversified agroecosystems – leader WP – 232k€ - *UMR GECO, Montpellier*

**2013-2017 ANR BIOADAPT** – GENESIS: the role of GENetic diversity and phenotypic plasticity in adaptations to changing Environments: a genomic analySIS of a biological invasion – leader WP – 138k€ - *UMR IMBE, Marseille & Bournemouth University, UK*

## 1.10 DEPOT DE BREVET OU DE LICENCE D'EXPLOITATION

Jean-François **Martin** (Montpellier SupAgro), Thibaut Malausa (INRA), André Gilles (Aix-Marseille Université), Stéphanie Ferreira (Genoscreen S.A.) : protocole universel de construction de banques enrichies en microsatellites par séquençage haut débit 454 GS-FLX Titanium pyrosequencing.

## 1.11 LISTE DE PUBLICATIONS (56)

En termes de production scientifique, j'ai participé à la rédaction de 56 articles de rang A au total, avec un équilibre sur la position en tant qu'auteur selon les critères d'évaluation du Ministère de l'Agriculture et de l'Alimentation (16 en premier ou second auteur, 19 en dernier ou avant-dernier auteur, 20 autres positions).

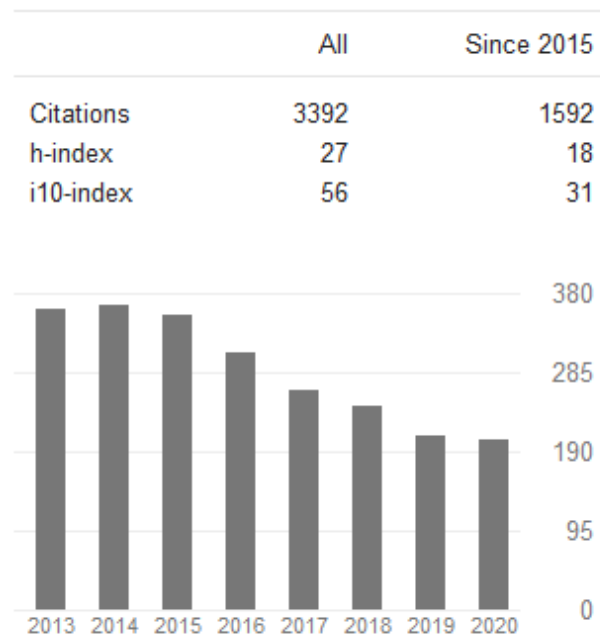


Figure 1 : histogramme du nombre de citations annuelles de mes publications. Extrait à partir des données de [Google Scholar](https://scholar.google.com/) au 17/11/2020).

- Melodie Ollivier, Vincent Lesieur, Sathyamurthy Raghu, Jean-François **Martin**. Characterizing ecological interaction networks to support risk assessment in classical biological control of weeds. 2020. *Current Opinions in Insect Science* 38 : 40-47. Doi : 10.1016/j.cois.2019.12.002

2. Melodie Ollivier, Elena Kazakou, Maxime Corbin, Kevin Sartori, Ben Gooden, Vincent Lesieur, Thierry Thomann, Jean-François **Martin**, Marie Stéphane Tixier. Trait differentiation between native and introduced populations of the invasive plant *Sonchus oleraceus* L. (Asteraceae). 2020. *Neobiota* 55 : 85-115. doi : 10.3897/neobiota.55.49158
3. Kevin Bethune, Cédric Mariac, Marie Couderc, Nora Scarcelli, Sylvain Santoni, Morgane Ardisson, Jean-François **Martin**, Rommel Montúfar, Valentin Klein, François Sabot, Yves Vigouroux, Thomas L. P. Couvreur : Long-fragment targeted capture for long-read sequencing of plastomes. *Applications in Plant Sciences*, 2019 7(5) : e1243, doi:10.1002/aps3.1243
4. Jakub Kreisinger, Lucie Schmiedová, Adéla Petrželková, Oldřich Tomášek, Marie Adámková, Romana Michálková, Jean-François **Martin**, Tomáš Albrecht: *Fecal microbiota associated with phytohaemagglutinin-induced immune response in nestlings of a passerine bird*. *Ecology and Evolution* 09/2018; 8(19)., DOI:10.1002/ece3.4454
5. Vincent Lesieur, Jean-François **Martin**, Harriet L. Hinz, Boris Fumanal, Rouhollah Sobhian, Marie-Claude Bon: *Implications of a phylogeographic approach for the selection of Ceutorhynchus assimilis as a potential biological control agent for Lepidium draba*. *Biological Control* 05/2018; 123., DOI:10.1016/j.biocontrol.2018.05.001
6. Emmanuel Guivier, Jean-François **Martin**, Nicolas Pech, Arnaud Ungaro, Rémi Chappaz, André Gilles: *Microbiota Diversity Within and Between the Tissues of Two Wild Interbreeding Species*. *Microbial Ecology* 09/2017;., DOI:10.1007/s00248-017-1077-9
7. Arnaud Ungaro, Nicolas Pech, Jean-François **Martin**, R. J. Scott McCairns, Jean-Philippe Mévy, Rémi Chappaz, André Gilles: *Challenges and advances for transcriptome assembly in non-model species*. *PLoS ONE* 09/2017; 12(9):e0185020., DOI:10.1371/journal.pone.0185020
8. Emmanuel Corse, Emese Megléc, Gaït Archambaud, Morgane Ardisson, Jean-François **Martin**, Christelle Tougard, Rémi Chappaz, Vincent Dubut: *A from-benchtop-to-desktop workflow for validating HTS data and for taxonomic identification in diet metabarcoding studies*. *Molecular Ecology Resources* 08/2017; 17(6)., DOI:10.1111/1755-0998.12703
9. Sylvain Piry, Catherine Wipf-Scheibel, Jean-François **Martin**, Maxime Galan, Berthier: *High throughput amplicon sequencing to assess within- and between-host genetic diversity in plant viruses*. DOI:10.1101/168773
10. Lucie Kropáčková, Hana Pechmanová, Michal Vinkler, Jana Svobodová, Hana Velová, Martin Těšický, Jean-François **Martin**, Jakub Kreisinger: *Variation between the oral and faecal microbiota in a free-living passerine bird, the great tit (Parus major)*. *PLoS ONE* 06/2017; 12(6):e0179945., DOI:10.1371/journal.pone.0179945
11. Lucie Kropáčková, Martin Těšický, Tomáš Albrecht, Jan Kubovčíak, Dagmar Čížková, Oldřich Tomášek, Jean-François **Martin**, Lukáš Bobek, Tereza Králová, Petr Procházka, Jakub Kreisinger: *Co-diversification of gastrointestinal microbiota and phylogeny in passerines is not explained by ecological divergence*. *Molecular Ecology* 04/2017; 26(19)., DOI:10.1111/mec.14144
12. Jakub Kreisinger, Lucie Kropáčková, Adéla Petrželková, Marie Adámková, Oldřich Tomášek, Jean-François **Martin**, Romana Michálková, Tomáš Albrecht: *Temporal Stability and the Effect of Transgenerational Transfer on Fecal*

- Microbiota Structure in a Long Distance Migratory Bird*. *Frontiers in Microbiology* 02/2017; 8(e0123933)., DOI:10.3389/fmicb.2017.00050
13. M. E. Maggia, Yves Vigouroux, J. F. Renno, Fabrice Duponchelle, E. Desmarais, J. Nunez, C. García-Dávila, F. M. Carvajal-Vallejos, Emmanuel Paradis, J. F. **Martin**, Cédric Mariac: *DNA Metabarcoding of Amazonian Ichthyoplankton Swarms*. *PLoS ONE* 01/2017; 12(1):e0170009., DOI:10.1371/journal.pone.0170009
  14. Vincent Lesieur, Jean-François **Martin**, David K. Weaver, Kim A. Hoelmer, David R. Smith, Wendell L. Morrill, Nassera Kadiri, Frank B. Peairs, Darren M. Cockrell, Terri L. Randolph, Debra K. Waters, Marie-Claude Bon: *Phylogeography of the wheat stem sawfly, Cephus cinctus Norton (Hymenoptera: Cephidae): implications for pest management*. *PLoS ONE* 12/2016; 11(12):e0168370., DOI:10.1371/journal.pone.0168370
  15. V. Lesieur, M. Jeanneau, J. F. **Martin**, M. C. Bon: *Development and characterization of 11 microsatellite markers in the root-gall-forming weevil, Ceutorhynchus assimilis (Coleoptera: Curculionidae)*. *Applied Entomology and Zoology* 04/2016; 51(3)., DOI:10.1007/s13355-016-0414-7
  16. Antoine Fraimout, Anne Loiseau, Donald K. Price, Anne Xuéreb, Jean-François **Martin**, Renaud Vitalis, Simon Fellous, Vincent Debat, Arnaud Estoup: *New set of microsatellite markers for the spotted-wing Drosophila suzukii (Diptera: Drosophilidae): A promising molecular tool for inferring the invasion history of this major insect pest*. *European Journal of Entomology* 07/2015; 112(4)., DOI:10.14411/eje.2015.079
  17. Gregory Mollot, Pierre-François Duyck, Pierre Lefeuvre, Françoise Lescourret, Jean-François **Martin**, Sylvain Piry, Elsa Canard, Philippe Tixier: *Cover Cropping Alters the Diet of Arthropods in a Banana Plantation: A Metabarcoding Approach*. *PLoS ONE* 04/2014; 9(4):e93740., DOI:10.1371/journal.pone.0093740
  18. Emese Megléc, Nicolas Pech, André Gilles, Vincent Dubut, Pascal Hingamp, Aurélie Trilles, Rémi Grenier, Jean-François **Martin**: *QDD version 3.1: A user-friendly computer program for microsatellite selection and primer design revisited: Experimental validation of variables determining genotyping success rate*. *Molecular Ecology Resources* 04/2014; 14(6)., DOI:10.1111/1755-0998.12271
  19. François Michel, Emese Megléc, Jean-François **Martin**, Henri Descimon: *Erebia serotina Descimon & de Lesse 1953 (Lepidoptera: Nymphalidae), a recurrent hybrid between two distantly related species*. *Annales- Societe Entomologique de France* 03/2013; 49(1):100-116., DOI:10.1080/00379271.2013.774741
  20. Sylvain Piry, Emmanuel Guivier, A Realini, J-F **Martin**: *|SE|S|AM|E| Barcode: NGS-oriented software for amplicon characterization - application to species and environmental barcoding*. *Molecular Ecology Resources* 07/2012; 12(6):1151-7., DOI:10.1111/j.1755-0998.2012.03171.x
  21. Emese Megléc, Nicolas Pech, André Gilles, Jean-François **Martin**, Michael G Gardner: *A shot in the genome: How accurately do shotgun 454 sequences represent a genome?*. *BMC Research Notes* 05/2012; 5(1):259., DOI:10.1186/1756-0500-5-259
  22. Malé Pierre-Jean G, **Martin** Jean-François, Galan Maxime, Deffontaine Valérie, Bryja Josef, Cosson Jean-François, Michaux Johan, Charbonnel Nathalie: *Discongruence of Mhc and cytochrome b phylogeographical patterns in Myodes*



- glareolus* (Rodentia: Cricetidae). Biological Journal of the Linnean Society 04/2012; 105(4):881-899., DOI:10.1111/j.1095-8312.2011.01799.x
23. Malausa Thibaut, Gilles André, Meglécz Emese, Blanquart Hélène, Duthoy Stéphanie, Costedoat Caroline, Dubut Vincent, Pech Nicolas, Castagnone-Sereno Philippe, Délye Christophe, Feau Nicolas, Frey Pascal, Gauthier Philippe, Guillemaud Thomas, Hazard Laurent, Le Corre Valérie, Lung-Escarmant Brigitte, Malé Pierre-Jean G, Ferreira Stéphanie, **Martin** Jean-François: *High-throughput microsatellite isolation through 454 GS-FLX Titanium pyrosequencing of enriched DNA libraries*. Molecular Ecology Resources 07/2011; 11(4):638-44., DOI:10.1111/j.1755-0998.2011.02992.x
  24. André Gilles, Emese Meglécz, Nicolas Pech, Stéphanie Ferreira, Thibaut Malausa, Jean-François **Martin**: *Accuracy and Quality Assessment of 454 GS-FLX Titanium Pyrosequencing*. BMC Genomics 05/2011; 12(1):245., DOI:10.1186/1471-2164-12-245
  25. Melthide Sinama, Vincent Dubut, Caroline Costedoat, André Gilles, Marius Junker, Thibaut Malausa, Jean-François **Martin**, Gabriel Nève, Nicolas Pech, Thomas Schmitt, Marie Zimmermann, Emese Meglécz: *Challenges of microsatellite development in Lepidoptera: Euphydryas aurinia (Nymphalidae) as a case study*. European Journal of Entomology 04/2011; 108(2):261–266., DOI:10.14411/eje.2011.035
  26. Y Le Conte, C Alaux, J-F **Martin**, J R Harbo, J W Harris, C Dantec, D Séverac, S Cros-Arteil, M Navajas: *Social immunity in honeybees (Apis mellifera): Transcriptome analysis of varroa-hygienic behaviour*. Insect Molecular Biology 03/2011; 20(3):399-408., DOI:10.1111/j.1365-2583.2011.01074.x
  27. Vincent Dubut, Rémi Grenier, Emese Meglécz, Rémi Chappaz, Caroline Costedoat, Delphine Danancher, Stéphane Descloux, Thibaut Malausa, Jean-François **Martin**, Nicolas Pech, André Gilles: *Development of 55 novel polymorphic microsatellite loci for the critically endangered Zingel asper L. (Actinopterygii: Perciformes: Percidae) and cross-species amplification in five other percids*. European Journal of Wildlife Research 12/2010; 56(6):931-938., DOI:10.1007/s10344-010-0421-x
  28. Malvina Andris, Gudbjorg I Aradottir, G Arnau, Asta Audzijonyte, Emilie C Bess, Francesco Bonadonna, G Bourdel, Joël Bried, Gregory J Bugbee, P A Burger, H Chair, P C Charruau, A Y Ciampi, L Costet, Paul J Debarro, H Delatte, Marie-Pierre Dubois, Mark D B Eldridge, Phillip R England, D Enkhbileg, B Fartek, Michael G Gardner, Karen-Ann Gray, Rasanthi M Gunasekera, Steven J Hanley, Nathan Havil, James P Hereward, Shotaro Hirase, Yan Hong, Philippe Jarne, Qi Jianfei, Rebecca N Johnson, Manami Kanno, Akihiro Kijima, Hyun C Kim, Kwan S Kim, Woo-Jin Kim, Elizabeth Larue, Jang W Lee, Jeong-Ho Lee, Chunhong Li, Minghui Liao, Nathan Lo, Andrew J Lowe, Thibaut Malausa, Pierre-Jean G Malé, Michelle D Marko, Jean-François **Martin**, Russell Messing, Karen J Miller, Byeong-Wha Min, Jeong-In Myeong, S Nibouche, Ann E Noack, Jae K Noh, Jérôme Orivel, Choul-Ji Park, D Petro, Kittipath Prapayotin-Riveros, Angélique Quilichini, B Reynaud, Cynthia Riginos, A M Risterucci, Harley A Rose, I Sampaio, K Silbermayr, M B Silva, N Tero, Ryan A Thum, C C Vinson, Adam Vorsino, Charles R Vossbrinck, C Walzer, Jason C White, Ania Wiczorek, Mark Wright: *Permanent Genetic Resources added to Molecular Ecology Resources Database 1 June 2010 - 31 July 2010*. Molecular Ecology Resources 11/2010; 10(6):1106-1108., DOI:10.1111/j.1755-0998.2010.02916.x



29. Emese Megléc, Sylvain Piry, Erick Desmarais, Maxime Galan, André Gilles, Emmanuel Guivier, Nicolas Pech, Jean-François **Martin**: *SESAME (SEquence Sorter & Amplicon Explorer): Genotyping based on high-throughput multiplex amplicon sequencing*. *Bioinformatics* 11/2010; 27(2):277-8., DOI:10.1093/bioinformatics/btq641
30. Jean-François **Martin**, Nicolas Pech, Emese Megléc, Stéphanie Ferreira, Caroline Costedoat, Vincent Dubut, Thibaut Malausa, André Gilles: *Representativeness of microsatellite distributions in genomes, as revealed by 454 GS-FLX Titanium pyrosequencing*. *BMC Genomics* 10/2010; 11(1):560., DOI:10.1186/1471-2164-11-560
31. Vincent Dubut, Melthide Sinama, Jean-François **Martin**, Emese Megléc, Juliette Fernandez, Rémi Chappaz, André Gilles, Caroline Costedoat: *Cross-species amplification of 41 microsatellites in European cyprinids: A tool for evolutionary, population genetics and hybridization studies*. *BMC Research Notes* 05/2010; 3(1):135., DOI:10.1186/1756-0500-3-135
32. Emmanuel Corse, Caroline Costedoat, Rémi Chappaz, Nicolas Pech, Jean-François **Martin**, André Gilles: *A PCR-based method for diet analysis in freshwater organisms using 18S rDNA barcoding on faeces*. *Molecular Ecology Resources* 01/2010; 10(1):96-108., DOI:10.1111/j.1755-0998.2009.02795.x
33. Emese Megléc, Caroline Costedoat, Vincent Dubut, André Gilles, Thibaut Malausa, Nicolas Pech, Jean-François **Martin**: *QDD: A user-friendly program to select microsatellite markers and design primers from large sequencing projects*. *Bioinformatics* 12/2009; 26(3):403-4., DOI:10.1093/bioinformatics/btp670
34. Vincent Dubut, Jean-François **Martin**, André Gilles, Jeroen VAN Houdt, Rémi Chappaz, Caroline Costedoat: *Isolation and characterization of polymorphic microsatellite loci for the dace complex: *Leuciscus leuciscus* (Teleostei: Cyprinidae)*. *Molecular Ecology Resources* 07/2009; 9(4):1179-83., DOI:10.1111/j.1755-0998.2009.02594.x
35. Vincent Dubut, Jean-François **Martin**, Caroline Costedoat, Rémi Chappaz, André Gilles: *Isolation and characterization of polymorphic microsatellite loci in the freshwater fishes *Telestes souffia* and *Telestes muticellus* (Teleostei: Cyprinidae)*. *Molecular Ecology Resources* 05/2009; 9(3):1001-5., DOI:10.1111/j.1755-0998.2009.02539.x
36. Anne Loiseau, Thibaut Malausa, Eric Lombaert, Jean-François **Martin**, Arnaud Estoup: *Isolation and characterization of microsatellites in the harlequin ladybird, *Harmonia axyridis* (Coleoptera, Coccinellidae), and cross-species amplification within the family Coccinellidae*. *Molecular Ecology Resources* 05/2009; 9(3):934-7., DOI:10.1111/j.1755-0998.2009.02517.x
37. H Hürner, J F **Martin**, A Ribas, A Arrizabalaga, J R Michaux: *Isolation, characterization and PCR multiplexing of polymorphic microsatellite markers in the edible dormouse, *Glis glis**. *Molecular Ecology Resources* 05/2009; 9(3):885-7., DOI:10.1111/j.1755-0998.2008.02365.x
38. Michel Yvon, Baptiste Monsion, Jean-François **Martin**, Serafín Gutiérrez, Stéphane Blanc: *PCR-based amplification and analysis of specific viral sequences from individual plant cells*. *Journal of virological methods* 05/2009; 159(2):303-7., DOI:10.1016/j.jviromet.2009.04.016
39. Marie-Pierre Chapuis, Julie-Anne Popple, Stephen J Simpson, Arnaud Estoup, Jean-François **Martin**, **Martin** Steinbauer, Laurence McCulloch, Gregory A Sword: *Eight polymorphic microsatellite loci for the Australian plague locust,*

- Chortoicetes terminifera*. Molecular Ecology Resources 11/2008; 8(6):1414-6., DOI:10.1111/j.1755-0998.2008.02204.x
40. Radka Poláková, Martina Vyskočilová, Jean-François **Martin**, Herman L. Mays Jr, Geoffrey E. Hill, Josef Bryja, Tomáš Albrecht, Herman L. Mays: *A multiplex set of microsatellite markers for the scarlet rosefinch (Carpodacus erythrinus)*. Molecular Ecology Notes 11/2007; 7(6):1375 - 1378., DOI:10.1111/j.1471-8286.2007.01892.x
  41. Thibaut Malausa, Laurianne Leniaud, Jean-François **Martin**, Philippe Audiot, Denis Bourguet, Sergine Ponsard, Siu-Fai Lee, Richard G Harrison, Erik Dopman: *Molecular Differentiation at Nuclear Loci in French Host Races of the European Corn Borer (Ostrinia nubilalis)*. Genetics 09/2007; 176(4):2343-55., DOI:10.1534/genetics.107.072108
  42. Polakova R, Vyskocilova M, **Martin** JF, Mays HL, Bryja J, Hill GE, Albrecht T: *A multiplex of microsatellite markers for scarlet rosefinch Carpodacus erythrinus..* Molecular Ecology Notes 07/2007; 7:1375-1378.
  43. Martina Vyskočilová, Andrea Šimková, Jean-François **Martin**: *Isolation and characterization of microsatellites in Leuciscus cephalus (Cypriniformes, Cyprinidae) and cross-species amplification within the family Cyprinidae*. Molecular Ecology Notes 05/2007; 7(6):1150 - 1154., DOI:10.1111/j.1471-8286.2007.01813.x
  44. Coeur d'Acier, E. Jousselin, J-F. **Martin**, J.-Y. Rasplus: *Phylogeny of the Genus Aphis Linnaeus, 1758 (Homoptera: Aphididae) inferred from mitochondrial DNA sequences*. Molecular Phylogenetics and Evolution 04/2007; 42:598-611., DOI:10.1016/j.ympev.2006.10.006
  45. Melanie L. Haines, Jean-François **Martin**, Rowan M. Emberson, Pauline Syrett, Toni M. Withers, Sue P. Worner: *Can sibling species explain the broadening of the host range of the broom seed beetle, Bruchidius villosus (F.) (Coleoptera: Chrysomelidae) in New Zealand?*. New Zealand Entomologist 02/2007; 30(1)., DOI:10.1080/00779962.2007.9722146
  46. Martina Vyskočilová, Markéta Ondračková, Andrea Šimková, Jean-François **Martin**: *Isolation and characterization of microsatellites in Neogobius kessleri (Perciformes, Gobiidae) and cross-species amplification within the family Gobiidae*. Molecular Ecology Notes 01/2007; 7(4):701 - 704., DOI:10.1111/j.1471-8286.2007.01682.x
  47. Chiraz Jridi, Jean-François **Martin**, Véronique Marie-Jeanne, Gérard Labonne, Stéphane Blanc: *Distinct Viral Populations Differentiate and Evolve Independently in a Single Perennial Host Plant*. Journal of Virology 04/2006; 80(5):2349-57., DOI:10.1128/JVI.80.5.2349-2357.2006
  48. B Fumanal, J-F **Martin**, M.C. Bon: *High through-put characterization of insect morphocryptic entities by a non-invasive method using direct-PCR of fecal DNA*. Journal of Biotechnology 10/2005; 119(1):15-9., DOI:10.1016/j.jbiotec.2005.04.011
  49. Eva Ottová, Andrea Simková, Jean-François **Martin**, Joëlle Goüy de Bellocq, Milan Gelnar, Jean-François Allienne, Serge Morand: *Evolution and trans-species polymorphism Of MHC class IIb genes in cyprinid fish*. Fish & Shellfish Immunology 04/2005; 18(3):199-222., DOI:10.1016/j.fsi.2004.07.004
  50. Boris Fumanal, J F **Martin**, Rouhollah Sobhian, Arnaud Blanchet, M.C. Bon: *Host range of Ceutorhynchus assimilis (Coleoptera: Curculionidae), a candidate for biological control of Lepidium draba (Brassicaceae) in the USA*. Biological Control 07/2004; 30(3)., DOI:10.1016/j.biocontrol.2004.03.001

51. M.D. Salducci, J.-F. **Martin**, N. Pech, R. Chappaz, C. Costedoat, A. Gilles: *Deciphering the evolutionary biology of freshwater fish using multiple approaches - Insights for the biological conservation of the Vairone (Leuciscus souffia souffia)*. Conservation Genetics 01/2004; 5(1):63-77., DOI:10.1023/B:COGE.0000014054.57397.3f
52. W. B. Watt, C. W. Wheat, E. H. Meyer, J.-F. Martin : Adaptation at specific loci. VII. Natural selection, dispersal and the diversity of molecular–functional variation patterns among butterfly species complexes (Colias: Lepidoptera, Pieridae). Molecular Ecology 2003 : 12,5 :1265 :1275. DOI : 10.1046/j.1365-294X.2003.01804.x
53. Jean-Francois **Martin**, André Gilles, Matthias Lörtscher, Henri Descimon: *Phylogenetics and differentiation among the western taxa of the Erebia tyndarus group (Lepidoptera: Nymphalidae)*. Biological Journal of the Linnean Society 03/2002; 75(3):319 - 332., DOI:10.1046/j.1095-8312.2002.00022.x
54. J.-F. **Martin**, A Gilles, H Descimon: *Molecular Phylogeny and Evolutionary Patterns of the European Satyrids (Lepidoptera: Satyridae) as Revealed by Mitochondrial Gene Sequences*. Molecular Phylogenetics and Evolution 05/2000; 15(1):70-82., DOI:10.1006/mpev.2000.0757
55. B. Barascud, J. F. **Martin**, Michel Baguette, H. Descimon: *Genetic consequences of an introduction-colonization process in an endangered butterfly species*. Journal of Evolutionary Biology 07/1999; 12(4-4):697-709.
56. Rémi Chappaz, André Gilles, Anne Miquelis, Laurent Cavalli, Jean-François **Martin**: *Différenciation génétique et hybridation chez le cyprin Leuciscus leuciscus*. Comptes Rendus de l'Académie des Sciences - Series III - Sciences de la Vie 11/1998; 321(11):933-940., DOI:10.1016/S0764-4469(99)80008-0

## 1.12 CHAPITRE D'OUVRAGE (2)

Thibaut Malausa, Laurianne Leniaud, **Jean-François Martin**, Philippe Audiot, Denis Bourguet, Sergine Ponsard, Siu-Fai Lee, Richard G. Harrison: *the European Corn Borer (Ostrinia nubilalis)*. 2006

**Jean-François Martin**, André Gilles, and Henri Descimon : Species Concepts and Sibling Species: The Case of *Leptidea sinapis* and *Leptidea reali*. Butterflies: Ecology and Evolution Taking Flight. 2005

## 1.13 COMMUNICATIONS ORALES ET POSTERS

Determining ecological interaction networks to assess risks and potential changes induced by importation biological control of weeds. September 2019. 4th symposium on ecological network (Paris, FRANCE) M. Ollivier, V. Lesieur, A. Sheppard, **J-F. Martin** et M-S. Tixier. Communication

How the study of interaction networks in native range helps the selection of biocontrol agents for the control of *Sonchus oleraceus* in Australia? June 2019. 19ème colloque de biologie de l'insecte (Albi, FRANCE) M. Ollivier, V. Lesieur, M. Jourdan, T. Thomann, S. Raghu, L. Morin, A. Sheppard, **J-F. Martin** et M-S. Tixier. Communication

Molecular analysis of ecological interactions for optimizing biocontrol of the invasive weed *Sonchus oleraceus* L. (Asteraceae) in Australia. September 2018. 21st

Australian Weed Conference (Sydney, AUSTRALIA) M. Ollivier, V. Lesieur, M. Jourdan, T. Thomann, S. Raghu, L. Morin, A. Sheppard, **J-F. Martin** et M-S. Tixier. Communication

Molecular analysis of ecological interactions for optimizing biocontrol of the invasive weed *Sonchus oleraceus* L. (Asteraceae) in Australia. August 2018. 15th International Symposium on Biological Control of Weeds (Engelberg, SUISSE) M. Ollivier, V. Lesieur, M. Jourdan, T. Thomann, S. Raghu, L. Morin, A. Sheppard, **J-F. Martin** et M-S. Tixier. Communication

Comparison of prey consumption between outdoor and farm cats in a rural free-ranging population of domestic cats (*Felis s. catus*). July 2014. 1st Feline Science Forum. Marie-Amélie Forin-Wiart, Maxime Galan, Gérald Umhang, Sylvain Piry, Vanessa Hormaz Bastid, Pauline Hubert, **J-F. Martin**, Frank Boué, Jean-François Cosson, Claire Larose, Marie-Lazarine Poulle. Communication

Food web study, a community approach for biological control of the weed *Sonchus oleraceus* L. October 2017. Conference: 5th International Symposium: Weeds and Invasive Plants (Chios, GRECE). ) M. Ollivier, V. Lesieur, M. Jourdan, T. Thomann, S. Raghu, L. Morin, A. Sheppard, **J-F. Martin** et M-S. Tixier. Poster

Meaningful application of the new 454 large scale pyrosequencing technology (Roche GS-FLX 454) to the identification of microsatellites for small-scale research projects. January 2011. M Jeanneau, M.C. Bon, **J-F. Martin**. Poster

## 2. PARCOURS ET ACTIVITES DE RECHERCHE HORS PROJET PRINCIPAL

Pour structurer les activités de recherche qui m'ont conduit au programme actuellement au cœur de ma recherche, j'ai choisi d'adopter dans cette section du mémoire une structure qui épouse globalement des grandes périodes de ma carrière post-thèse et relate l'évolution des thématiques que j'ai abordées ainsi que les compétences que j'ai développées dans ce processus. Les périodes ne sont données qu'à titre indicatif et ne correspondent pas nécessairement à des changements thématiques brusques mais plutôt à une typologie du cœur de thématique majoritaire à cette période. Ce cheminement me paraissait être le plus éclairant sur mon parcours et à même de l'expliquer au mieux.

### 2.1 RECONSTRUIRE L'HISTOIRE EVOLUTIVE DE POPULATIONS NATURELLES EN DESEQUILIBRE (1999-2006)

Les articles associés à cette thématique : #5, 14, 27, 28, 37, 39-43, 45-48, 50, 51, 54, 55

A l'issue de mon doctorat en octobre 1999, j'avais acquis une certaine expertise en matière de macro évolution et de reconstruction phylogénétique, majoritairement dans le domaine de la phylogénie de taxa (species tree) mais aussi, par des collaborations ponctuelles, dans celui de l'évolution de gènes candidats sans objectif de transposition des résultats aux espèces porteuses du polymorphisme analysé (gene trees). Dans ce dernier cas les enjeux étaient tout autant de l'ordre de la reconstruction de l'histoire évolutive du gène étudié que parfois de détecter des signatures de sélection à l'échelle macro-évolutive.

A la suite de cette première expérience, pour répondre à ma décision de compléter ma formation par des approches d'évolution au niveau intraspécifique, mais aussi par ma volonté de maîtriser l'acquisition des données moléculaires que j'avais peu développée durant mon travail de thèse, j'ai fait un premier séjour post-doctoral de 10 mois dans l'UMR Laboratoire d'Ecologie Alpine (UJF-CNRS à Grenoble, aujourd'hui LECA) sous la direction de Pierre Taberlet. Pendant cette période, j'ai eu en charge un projet de recherche ponctuel sur la phylogéographie de la Perdrix grise. Ce projet de génétique de la conservation (populations en crash démographique et espèce en danger de disparition) a été pour moi une opportunité d'apprentissage intense, tant sur le plan de l'acquisition des marqueurs moléculaires que de l'analyse de la diversité génétique au niveau intraspécifique, champ qui m'était alors étranger.

Cette expérience m'a permis d'être sélectionné par le CSIRO ((Commonwealth Scientific and Industrial Research Organisation) dans son laboratoire européen à Montpellier pour mettre sur pied un laboratoire de biologie moléculaire commun avec l'USDA (European Biological Control Laboratory, EBCL), deux opérateurs de recherche présents sur le même campus international. Cette tâche, effectuée en étroite collaboration avec ma collègue Marie-Claude Bon (USDA, EBCL), m'a permis de mettre à profit l'expérience acquise précédemment et de la transposer au contexte

de la phylogéographie des ravageurs de cultures ou des adventices envahissantes en retraçant leurs routes de colonisation sur la base de l'analyse de leur diversité intraspécifique. Cette expérience a également été une occasion d'appréhender les enjeux sociétaux liés à ces systèmes. Ce fut aussi le moment des premiers encadrements d'étudiants de master en mon nom propre. Derrière une diversité d'organismes cibles des programmes de contrôle biologiques du CSIRO et de l'USDA, cette période (2000-2002) a correspondu à une poursuite de ma montée en connaissances quant aux concepts évolutifs associés au polymorphisme intraspécifique. D'autre part elle a également été très riche d'enseignements sur les aspects touchant au design expérimental, au choix des marqueurs moléculaires pertinents et à leur mise en œuvre par l'analyse fréquentiste ou bayésienne permettant d'aborder des tests de scénarii démographiques sous un angle nouveau à cette époque.

Les modèles d'étude ont été divers, plus précisément de deux types :

- Des plantes envahissantes (le passereau drave (*Lepidium draba*) et le radis sauvage (*Raphanus raphanistrum*)) pour lesquelles nous cherchions à clarifier la structure de la diversité génétique ainsi que l'origine ;
- Des Arthropodes, agents de lutte biologique potentiels contre ces adventices (le charançon *Ceutorhynchus assimilis* pour lutter contre *L. draba* par exemple) ou des ravageurs des cultures comme la cochenille *Maconellicoccus hirsutus* ou la tenthrède du blé *Cephus cinctus*

Le projet le plus aboutit dans ce cadre est sans aucun doute celui mené sur le charançon *Ceutorhynchus assimilis* (articles # 5,15,48,50) qui a concerné un étudiant de master (B. Fumanal) et plus récemment un post-doctorant (V. Lesieur). Partant d'une problématique de lutte biologique classique s'intéressant à la recherche d'un agent de contrôle contre *L. draba*, ce projet nous a amené à dérouler un sujet complexe et multi-facettes qui a impliqué successivement l'analyse de la structure de la diversité génétique de ce charançon (Figure 2), à mettre en évidence des entités génétiques cryptiques, à tester la présence de traces d'hybridation entre ces entités, et enfin à en établir le spectre de plantes hôtes *in Natura* mais aussi en expérimentation. Les résultats de ces recherches ont éclairé le processus de sélection de cette espèce comme agent de contrôle biologique pour lutter contre *L. draba*.

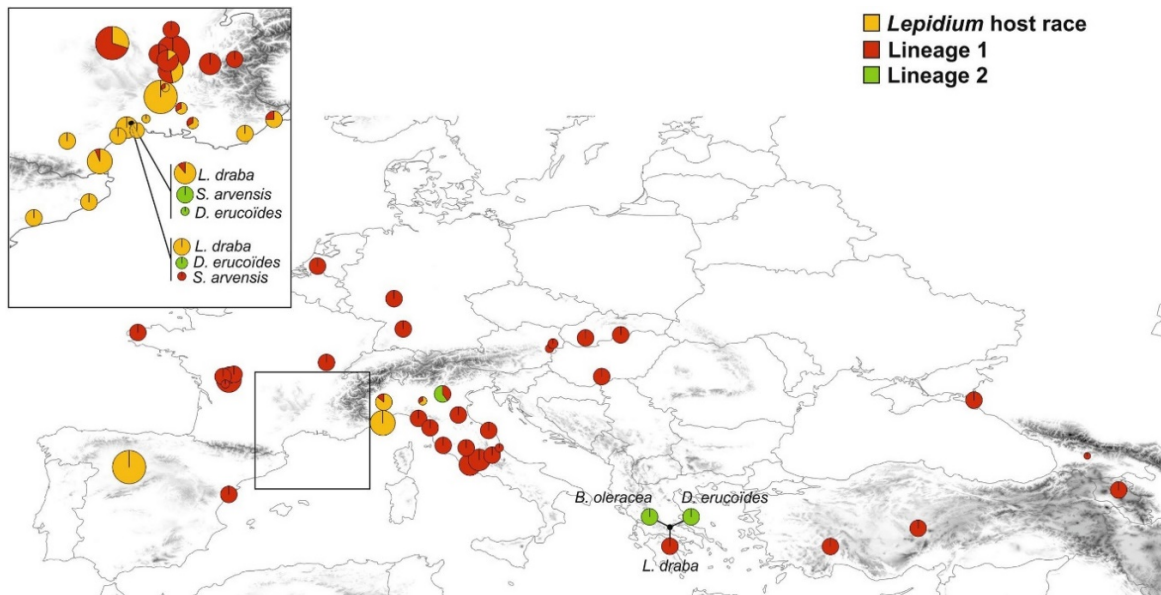


Fig. 2. Distribution géographique des différentes lignées de *Ceutorhynchus assimilis*. La taille des diagrammes circulaires est proportionnelle à la taille de l'échantillon. (Figure extraite de l'article #5)

Un second programme dans ce thème s'est intéressé à la tenthrède du blé, *Cephus cinctus* Norton (Hymenoptera : Cephidae), ravageur clé du blé dans le nord des grandes plaines d'Amérique du Nord et récemment étendu vers le sud. Dans ce projet nous avons examiné à la fois la divergence interspécifique entre les échantillons collectés en Amérique du Nord et en Asie du Nord-Est, l'aire de répartition supposée de *C. cinctus*, mais nous avons également caractérisé la structure de la diversité génétique dans les principales régions productrices de blé en Amérique du Nord en utilisant une combinaison de marqueur ADN mitochondrial et de microsatellites dans des échantillons collectés à la fois dans des champs de blé et dans des plantes sauvages. Les résultats obtenus (Figure 3) tendent à exclure l'hypothèse d'un déplacement récent des populations de tenthrèdes du blé nuisibles de la zone nord. Le déplacement de l'utilisation de la plante hôte par les populations humaines locales semble être le scénario le plus probable. Comme dans le cas de *C. assimilis*, ces résultats ont été évalués dans le cadre de la lutte biologique et ont fourni une base évolutive solide pour définir des stratégies de gestion de ce ravageur.







## **L'analyse de la différenciation moléculaire des races d'hôtes françaises de la pyrale du maïs (*Ostrinia nubilalis*).**

Les populations françaises de la pyrale du maïs sont constituées de deux races d'hôtes sympatriques et génétiquement différenciées. Ce système biologique est un bon candidat pour étudier les processus qui pourraient être impliqués dans la spéciation sympatrique, mais les conditions initiales de la divergence entre les races d'hôtes doivent être élucidées. Les généalogies génétiques ont permis de fournir un aperçu des processus impliqués dans la spéciation. Nous avons utilisé les séquences d'ADN de quatre gènes nucléaires pour (i) documenter la structure génétique des deux races d'hôtes françaises préalablement délimitées par des marqueurs allozymatiques, (ii) trouver les gènes directement ou indirectement impliqués dans l'isolement reproductif entre les races hôtes, et (iii) estimer le temps écoulé depuis la divergence des deux taxons et valider si cette estimation est compatible avec l'hypothèse que cette divergence soit correspond plutôt au résultat d'un changement d'hôte vers le maïs après son introduction en Europe il y a environ 500 ans. Les généalogies obtenues (Figure 4) ont révélé un polymorphisme partagé omniprésent, mais ont confirmé la différenciation entre les deux races d'hôtes. Des écarts importants par rapport aux prévisions des modèles neutres d'évolution moléculaire ont été détectés sur trois loci, mais ceux-ci ils n'avaient *a priori* aucun lien direct avec la mise en place d'isolement reproducteur entre les races d'hôtes. De plus, les estimations du temps écoulé depuis la divergence entre les races d'hôtes françaises variaient de 75 000 à 150 000 ans, ce qui suggère que les deux taxons ont divergé bien avant l'introduction du maïs en Europe.

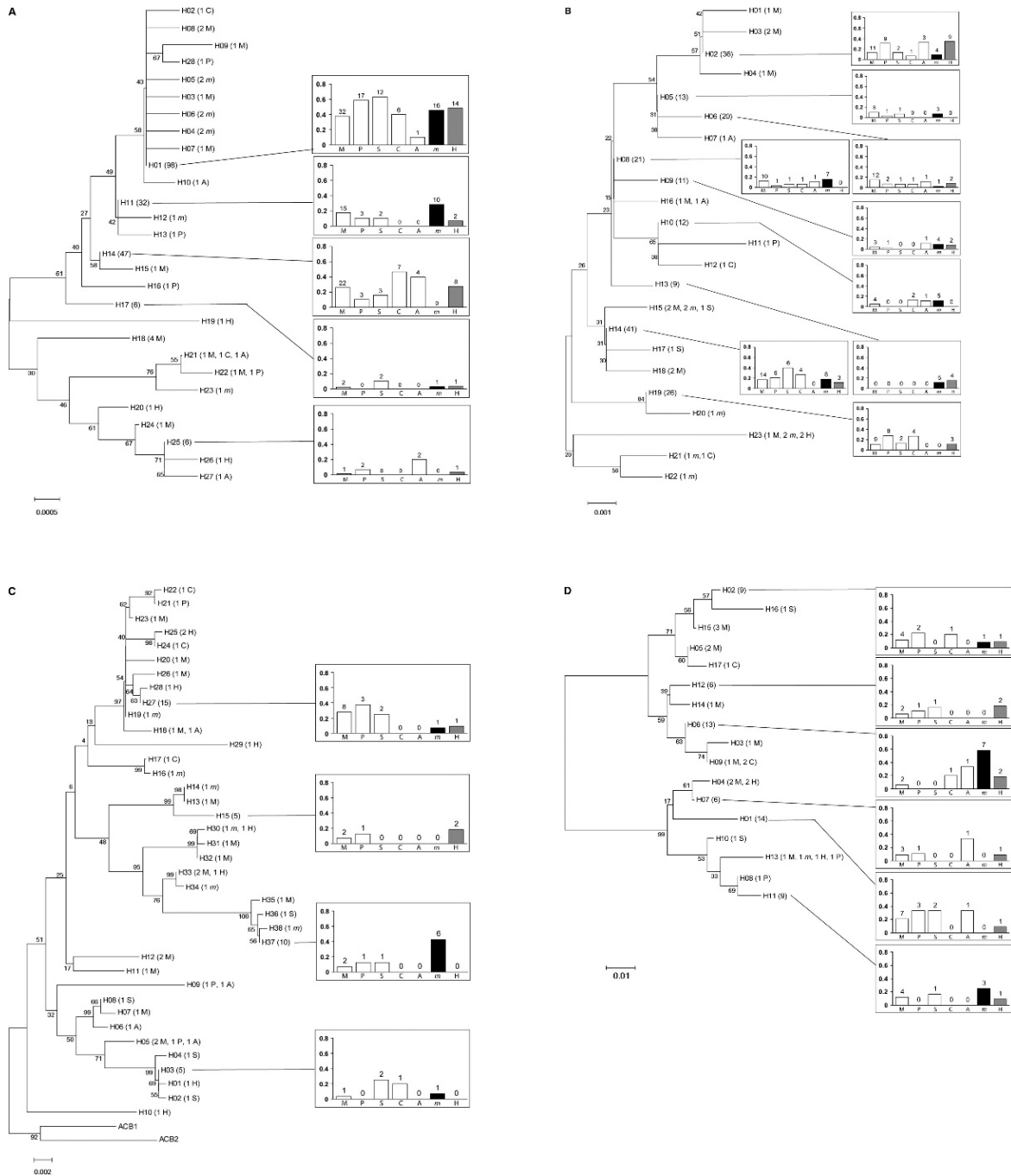


Figure 4 : arbres NJ pour les quatre loci : (A) Tpi, (B) Ket, (C) Pbp, (D) Mpi. H01-H30: numéro de l'haplotype, suivi du nombre d'individus porteurs de chaque haplotype (entre parenthèses) et de la ou des plantes hôtes sur lesquelles ces individus ont été prélevés (les noms sont codés comme suit : Les plantes hôtes : M, maïs ; m, armoise ; P, poivron ; H, houblon ; C, lampourde ; S, sorgho ; A, amarante. Pour les haplotypes partagés par plus de quatre individus prélevés sur plus de quatre plantes différentes, un diagramme (à droite) représente (i) parmi tous les individus prélevés sur une plante donnée, la proportion partageant l'haplotype (barres) et (ii) le nombre correspondant d'individus de cette plante (au-dessus des barres). Lorsque des séquences étaient disponibles, *Ostrinia furnacalis* (la pyrale du maïs asiatique, ACB) a été utilisée en tant que groupe de référence. (Figure extraite de l'article #41)

## **L'analyse des populations virales distinctes qui se différencient et évoluent de manière indépendante dans une seule plante hôte pérenne**

A cette époque (2006) la structure complexe des populations de virus faisait l'objet d'études intensives sur les bactéries, les animaux et les plantes depuis plus de dix ans. S'il était clair qu'une énorme diversité génétique était rapidement générée lors de la réplication virale, la répartition de cette diversité au sein d'un même hôte restait un domaine obscur. Parmi les virus animaux, seules les populations du virus de l'immunodéficience humaine et du virus de l'hépatite C avaient été étudiées de manière approfondie au niveau intra-hôte, où elles sont structurées comme des métapopulations, ce qui démontre que l'hôte ne peut pas être considéré simplement comme un "sac" contenant un essaim homogène ou non structuré de génomes viraux mutants. Chez les plantes, quelques rapports avaient suggéré une possible distribution hétérogène des variantes de virus à différents endroits de l'hôte, mais n'avaient fourni aucun indice sur la façon dont cette hétérogénéité est structurée. Nous avons présenté l'étude la plus exhaustive de la structure et de l'évolution d'une population virale jamais signalée au niveau intra-hôte à cette époque, grâce à l'analyse d'un Prunus infecté par le virus de la sharka du prunier pendant plus de 13 ans à la suite d'une seule inoculation et à l'utilisation de l'analyse de la variance moléculaire à différents niveaux hiérarchiques combinée à l'analyse des clades emboîtées (Figure 5). Nous avons démontré que, suite à l'invasion systémique de l'hôte, la population du virus se différencie en plusieurs populations distinctes qui sont isolées dans différentes branches, où elles évoluent indépendamment grâce à l'expansion de leur aire de répartition contiguë tout en colonisant des organes nouvellement formés. De plus, ces résultats nous ont permis de discuter des preuves que l'arbre abrite une énorme "banque" de clones viraux, chacun isolé dans une des myriades de feuilles.

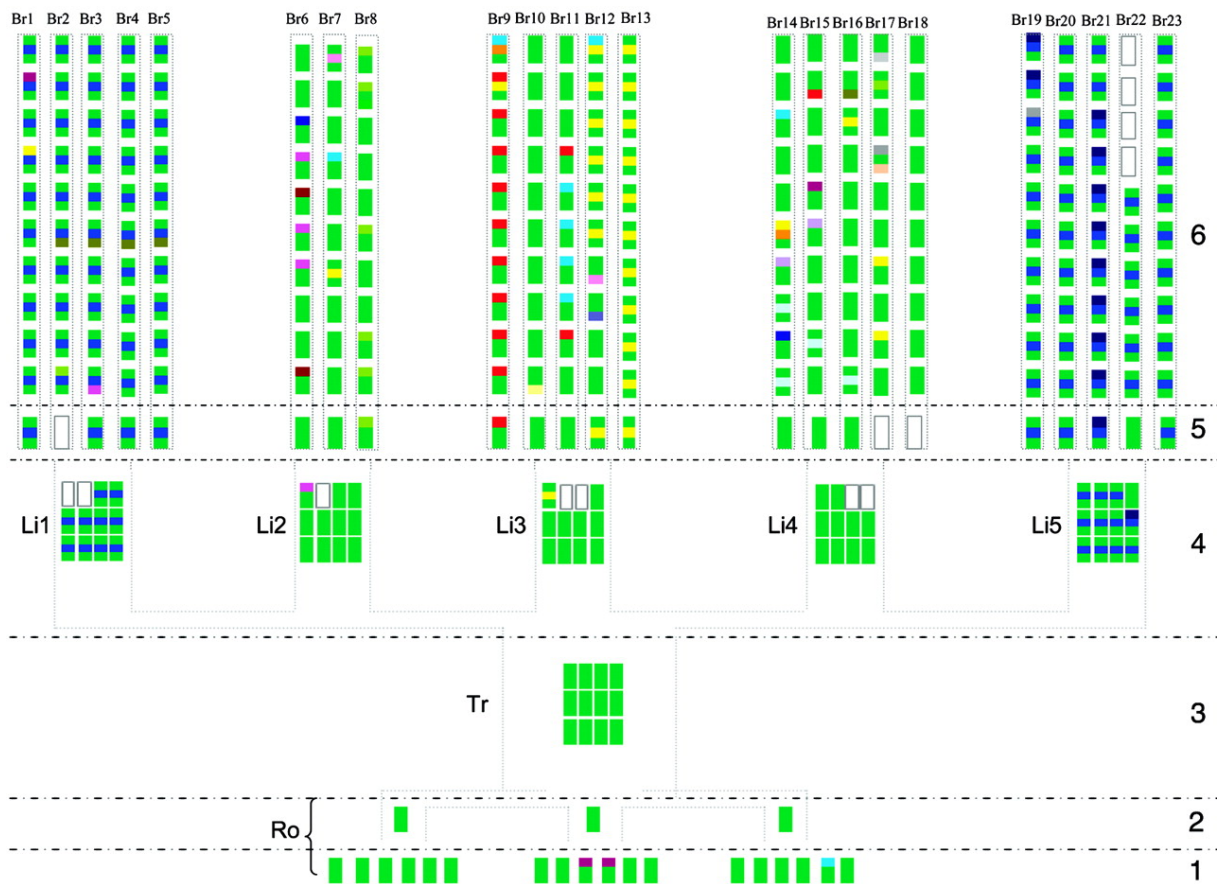


Fig. 5 : Distribution des haplotypes superposée à l'architecture schématique de l'arbre. Chaque haplotype est représenté par un mince rectangle composé de trois compartiments distincts illustrant les différentes séquences des régions A, B et C. Pour faciliter la représentation schématique des haplotypes, le premier haplotype détecté, H1, pour les régions A, B et C est représenté en vert. Lorsque la séquence d'une région génomique donnée varie, la couleur du compartiment correspondant change. Les haplotypes vierges correspondent aux échantillons dont l'immunocapture RT-PCR a échoué à plusieurs reprises. (1) haplotypes détectés dans les 18 jeunes échantillons de racines terminales. (2) haplotypes détectés dans les trois principaux échantillons de racines. (3) haplotypes détectés dans les 12 échantillons d'écorce du tronc. (4) haplotypes détectés dans les 60 échantillons d'écorce des membres constitutifs. (5) haplotypes détectés dans les 23 échantillons d'écorce des branches d'un an. (6) haplotypes détectés dans les 230 échantillons de feuilles. (Figure extraite de l'article #47)

En parallèle de ces activités, je conservais ma collaboration de longue haleine (datant de 1996) avec l'équipe d'André Gilles (Aix-Marseille Université) sur la structure génétique et l'histoire évolutive des poissons cyprinidés ; recherches dans lesquelles ces mêmes compétences étaient utiles malgré un grand écart quant au modèle biologique. J'illustre ici cette collaboration à travers un des travaux clef.

### **Décrypter la biologie évolutive des poissons d'eau douce à l'aide de multiples approches - perspectives pour la conservation biologique de la Vairone (*Leuciscus souffia souffia*)**

L'organisation de la variabilité génétique est de première importance dans la conception des stratégies de conservation. Dans ce contexte, les unités de conservation sont souvent définies en utilisant le concept d'unités significatives évolutives (USE) ce qui pourrait apparaître comme un guide utile mais qui était très rarement utilisé à cette époque. Une autre difficulté survient lorsque les espèces s'hybrident comme c'est le cas pour les poissons cyprinidés pour lesquels le débat rejoint les discussions sur les définitions des espèces. Par exemple, le vairone, *Leuciscus souffia* (Teleostei : Cyprinidae) est une espèce protégée (UICN, Convention de Berne, Habitat), mais la législation ne tient pas compte de son statut taxonomique ambigu, ni de son caractère la variabilité ni ses préférences écologiques. Dans l'analyse de sa biologie évolutive, nous avons examiné la structure génétique et la phylogéographie de la sous-espèce *Leuciscus souffia souffia* dans son aire de répartition (France) avec une combinaison de morphologie, d'allozymes et de séquences d'ADN mitochondrial en utilisant des analyses principalement basées sur l'AMOVA et l'analyse des clades emboîtées (Figure 6). Nous avons ensuite déchiffré la biologie évolutive de ce poisson par l'analyse combinée de la morphologie, des marqueurs moléculaires d'origine nucléaire et de l'ADN mitochondrial. Nous avons montré un découplage entre l'homogénéité morphologique et une différenciation génétique modérée, ce qui donnait des indications pour définir les unités de gestion (selon la définition de Moritz). Nous avons conclu cette étude en proposant des orientations tant pour la protection des populations menacées que pour la conservation du potentiel d'évolution approximé sur la base des dynamiques évolutives trouvées pour les populations examinées et à la lumière des définitions d'USE les plus souvent utilisées. Ces conclusions sont toujours valides aujourd'hui.



## 2.2 LE TEMPS DE L'ADAPTATION (2007-2009)

Les articles associés à cette thématique : #26, 52

Durant ces premières années d'activité au CBGP, j'avais en parallèle continué de travailler avec l'équipe du Biological Department de Stanford (W.B. Watt) dans laquelle j'avais pris ma place pendant ma thèse (en parallèle du thème principal de celle-ci). Dans ce cadre, je travaillais en partenariat avec cette équipe sur l'interaction Génome-Environnement par une approche Gène candidat sous contrainte sélective. Le cadre conceptuel était lié à l'impact de la sélection sur la thermo-stabilité du Lépidoptère ciblé par le programme (*Colias eurytheme*) à travers l'analyse du polymorphisme génétique à un gène impliqué dans le métabolisme (PGI). Le projet faisait un lien étroit entre écophysiologie, écologie et biologie moléculaire<sup>52</sup>. Les développements européens de ces activités étaient financés par un Appel à Proposition de l'Institut Français de la Biodiversité de 2003 à 2005 (38k€). Malheureusement la canicule de l'été 2003 a stoppé net le programme car elle a éliminé les populations du dit Lépidoptère, ce qui a conduit à l'annulation anticipée du projet.

A partir de 2007, j'ai cherché à réconcilier mes activités avec le domaine « interaction génome/environnement » que j'avais dû interrompre à regret quelques années plus tôt et donc à faire passer au second plan les projets relatifs à la génétique des populations neutre. Cela m'a amené à développer un axe d'étude de l'adaptation des abeilles aux pressions pathogène de l'acarien *Varroa* en partenariat avec Maria Navajas (DR INRA au CBGP). L'intégration à ce programme était un challenge personnel important car il ne faisait pas appel à la majeure partie de mes compétences et imposait d'en acquérir un grand nombre de nouvelles dans les domaines de l'analyse des puces à ADN, du transcriptome, de l'expression différentielle et de la génomique en général liée au passage à l'échelle de l'approche gène candidat vers la variation à l'échelle du génome.

Le cadre global de cette étude était le développement par les abeilles d'une immunité sociale consistant en la coopération des individus pour contrôler la maladie dans la ruche. Nous avons identifié un ensemble de gènes impliqués dans cette immunité sociale en analysant le transcriptome cérébral d'abeilles très « hygiéniques », qui détectent et éliminent efficacement le couvain infecté par l'acarien. La fonction de ces gènes candidats détectés ne semblait pas favoriser une plus grande sensibilité olfactive chez les abeilles hygiéniques, comme on l'avait jusqu'alors supposé dans la littérature. Cependant, la comparaison de leur profil génomique avec ceux d'autres comportements suggéraient un lien avec les soins du couvain et les abeilles africanisées très hygiéniques à l'égard des varroas. Ces résultats ont représenté un premier pas vers l'identification des gènes impliqués dans l'immunité sociale (Figure 7) et ont donné ainsi un premier aperçu de son évolution<sup>26</sup>.

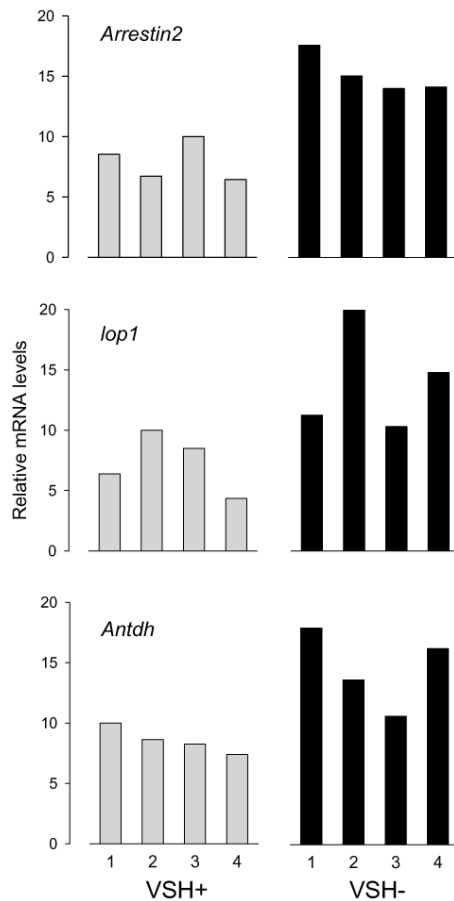


Fig. 7 : validation des résultats des puces à ADN avec la PCR quantitative en temps réel. Niveaux d'expression cérébrale de trois gènes (*Arrestin 2*, *lop1* et *Antdh*) identifiés par l'étude par microarray comme étant exprimés différemment entre les abeilles ayant un taux d'hygiène faible (VSH-) et élevé (VSH+). Les individus des colonies VSH- et VSH+ utilisées pour les puces ont été testés. Chaque barre représente un échantillon de colonie (pool de 10 abeilles). Des différences significatives ont été constatées pour chaque gène à l'aide d'un test U de Mann-Whitney ( $P < 0,05$ ). (Figure extraite de l'article #26)

Courant 2009, la stratégie nationale de l'INRA a abouti au recentrage de toutes les activités liées aux abeilles à Avignon ce qui a mis fin de fait à ces recherches au CBGP pour M. Navajas et par extension pour les miennes. S'en est alors suivie une période de transition de quatre ans que je résume dans le paragraphe suivant.



## 2.3 DE LA VARIATION GENETIQUE A LA GENOMIQUE – UNE TRANSITION METHODOLOGIQUE CHARNIERE (2009-2013)

Les articles associés ou découlant de cette thématique : #3, 7, 15, 16, 18, 21, 23-25, 27-31, 33-36

Suite à la réorientation forcée de mes activités en 2009, celles-ci ont pris une dimension méthodologique qui était dans la logique de l'approfondissement du potentiel de la génomique dans la résolution de questions de biologie et d'écologie évolutive. L'émergence à cette époque des technologies de séquençage haut débit était une opportunité technologique révolutionnant l'approche de la structure de la diversité génétique à l'échelle du génome au-delà des organismes modèles. Grâce à ces approches de séquençage massif, l'information à l'échelle du génome n'était plus la panacée de quelques organismes modèles comme l'abeille mais accessibles pour toutes les espèces. C'est donc la combinaison de la génétique des populations et des approches de génomique développées sur l'abeille par les microarrays qui m'ont amené à développer les méthodes d'acquisition de données moléculaires pour faire de la génétique des populations à l'échelle du génome, alliant développement technique en labo et développements informatiques dans le traitement des données obtenues. Cette approche intégrée de l'acquisition et du traitement des données a structuré mon activité avec le bénéfice attendu de produire des données de marqueurs moléculaires de qualité pour tester les hypothèses évolutives ou démographiques qui agitaient nos équipes.

Au-delà de l'élargissement de mes compétences personnelles, cette activité a permis à de nombreux collègues partenaires d'introduire l'utilisation de marqueurs moléculaires à haut débit dans leur quotidien ce qui a été un fort motif de satisfaction. En effet, à cette époque, les microsatellites étaient le marqueur de choix pour analyser la structure génétique et les flux de gènes entre populations. Leur isolement et caractérisation était fastidieuse et nécessitait des constructions de banques enrichies qui nécessitaient du clonage bactérien. L'arrivée des technologies de séquençage haut débit produisant de longs fragments (400pb !) était une révolution contournant le clonage et massifiant la disponibilité de marqueurs. Avec mes collègues Thibaut Malausa (CR INRAE Sophia-Antipolis) et André Gilles (MCF Aix-Marseille Université) nous avons imaginé un ensemble cohérent de procédures d'acquisition, de traitement et d'analyse des données moléculaires pour produire des marqueurs microsatellites en haut débit. Cette méthode était adaptable et trouvait une application quelle que soit l'espèce d'intérêt. Ce projet a été mené à bien de 2009 à 2011 sous le nom de Consortium ECOMICRO et a irrigué plus d'une centaine d'équipes utilisant la génétique des populations en France, une cinquantaine de plus en Europe. Il a donné lieu à un dépôt de licence exploitée par un partenaire industriel privé et à la publication d'un peu plus de 200 articles au dernier comptage en 2015. Trois éléments fondamentaux ont assuré le succès de ce projet méthodologique : i) une méthode d'isolement de microsatellites à haut débit couplée au pyroséquençage 454 GS-FLX de bibliothèques d'ADN enrichies<sup>23</sup>, ii) la validation de la représentativité des banques

produites<sup>21,30</sup> et l'évaluation de la qualité des séquences obtenues<sup>24</sup>, iii) des outils de traitement des séquences obtenues pour en extraire les marqueurs microsatellites<sup>18,33</sup>.

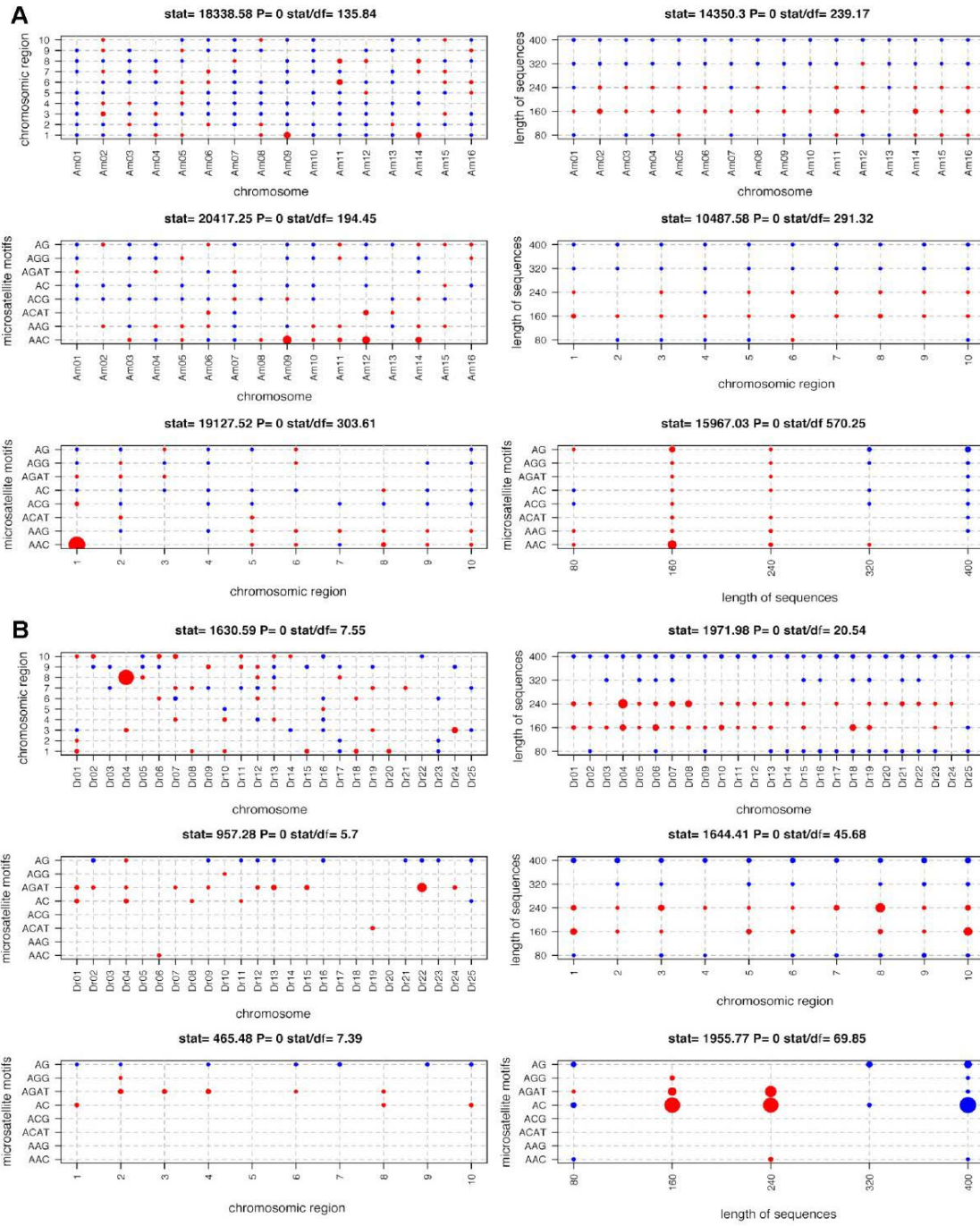
### **Isolement de microsatellites à haut débit grâce au pyroséquençage de bibliothèques d'ADN enrichies au titane 454 GS-FLX**

L'utilisation des marqueurs microsatellites était limitée par les difficultés liées à leur isolement *de novo* des espèces pour lesquelles aucune ressource génomique n'était disponible. Nous avons décrit une méthode à haut débit pour isoler les marqueurs microsatellites basée sur le couplage de l'enrichissement multiplex par sondes biotinylées des microsatellites et du séquençage haut débit sur des plateformes 454 GS-FLX Titanium. La procédure a été calibrée sur une espèce modèle (*Apis mellifera*) et validée sur 13 autres espèces de divers groupes taxonomiques (animaux, plantes et champignons), y compris des taxons pour lesquels de graves difficultés étaient auparavant rencontrées avec les méthodes traditionnelles. Nous avons obtenu de 11 497 à 34 483 séquences (c'était considérable à l'époque !). Selon les espèces et le nombre de loci microsatellites détectés allait de 199 à 5791. Nous avons ainsi démontré que cette procédure pouvait être facilement et avec succès appliquée à une grande variété de groupes taxonomiques, à un coût bien moindre que celui qui aurait été possible avec les protocoles traditionnels. Cette méthode a permis d'accélérer l'acquisition de marqueurs génétiques de haute qualité pour les organismes non modèles.

### **Représentativité des distributions des microsatellites dans les génomes, telle que révélée par le pyroséquençage du titane 454 GS-FLX**

L'approche développée n'était réellement efficace que dans la mesure où les bibliothèques enrichies de microsatellites étaient représentatives du génome à partir duquel elles avaient été isolées. Nous avons analysé les divergences potentielles de représentativité dans la distribution des marqueurs microsatellites obtenus pour deux génomes de référence (*Apis mellifera* et *Danio rerio*), sélectionnés sur la base de leur extrême hétérogénéité en termes de proportions et de distributions des microsatellites sur les chromosomes. Le génome d'*A. mellifera*, en particulier, s'est avéré très hétérogène, en raison de taux de recombinaison extrêmement élevés, mais le seul biais introduit systématiquement dans les bibliothèques enrichies en multiplex concernait la longueur des séquences, avec une surreprésentation des séquences de 160 à 320 pb de longueur (Figure 8). D'autres écarts par rapport aux proportions ou distributions attendues des motifs sur les chromosomes ont été observés, mais la signification et l'intensité de ces écarts étaient pour la plupart limités. En outre, aucune concurrence négative constante entre les sondes multiplexées n'a été observée pendant la phase d'enrichissement des motifs.

Nous avons pu en conclure que la méthode était fiable et permettait d'améliorer le développement des microsatellites, car elle n'introduisait pas de biais majeur en termes de proportions et de distribution des microsatellites.



**Fig. 8 : Comparaison des données 454 de séquences microsatellites (pour les 8 sondes) pour *Apis mellifera* (A) et *Danio rerio* (B), pour chacun des six tableaux bidirectionnels considérés : chromosome x région chromosomique, chromosome x longueur des séquences, chromosome x motifs microsatellites, région chromosomique x longueur des séquences, région chromosomique x motifs microsatellites et longueur des séquences x motifs microsatellites. "Stat" fait référence aux statistiques globales  $\chi^2$  pour le test d'indépendance entre les deux variables définissant chaque tableau. Pour chaque cellule du tableau, un point rouge correspond à un nombre observé significativement plus élevé que prévu dans l'hypothèse d'indépendance entre les deux variables définissant le tableau, un point bleu correspond à un nombre observé significativement plus faible que prévu dans l'hypothèse d'indépendance entre les deux variables définissant le tableau ; aucun point indique un écart non significatif. La taille du point est inversement proportionnelle à la valeur P, l'échelle étant telle que les comparaisons de tailles entre les tableaux sont significatives. (Figure extraite de l'article #30)**

Une fois la généralité de l'approche assurée, la qualité et la représentativité des motifs vérifiées, nous avons développé des outils de traitement bio-informatiques pour exploiter ces banques enrichies en motifs microsatellites pour que le non spécialiste puisse s'en emparer en autonomie. C'est dans cette optique que nous avons conçu

**QDD : un logiciel convivial permettant de sélectionner des marqueurs microsatellites et de concevoir des amorces à partir de grands projets de séquençage**

QDD (Quick and Dirty for Dummies) a été conçu comme un programme en libre accès qui fournit un outil convivial pour la détection des microsatellites et la conception d'amorces à partir de grands ensembles de séquences d'ADN. Le programme permet de conduire toutes les étapes du traitement des séquences brutes obtenues par pyroséquençage de banques d'ADN enrichies, mais il est également applicable aux données obtenues par d'autres méthodes de séquençage, en utilisant les fichiers FASTA en entrée. Les tâches suivantes sont accomplies par QDD (Figure 9) : tri des barcodes moléculaires, retrait des adaptateurs de séquençage/vecteurs, élimination des séquences redondantes, détection d'éventuelles multicopies génomiques (loci dupliqués ou éléments transposables), sélection rigoureuse des microsatellites cibles et conception d'amorces personnalisables fournies dans un tableau de synthèse.

Ce logiciel a été utilisé dans l'analyse de plus de 700 banques enrichies en motifs microsatellites dans le cadre du consortium ECOMICRO et au-delà par la communauté dans les années qui ont suivi.

A la fin de ce projet, qui était ponctuel par construction, nous avons fait le constat que l'expertise acquise était transférable à des approches davantage centrées sur l'ADN environnemental qui commençait alors à prendre son essor et à montrer son potentiel dans l'utilisation de données moléculaires pour résoudre des questions centrées sur l'écologie des communautés et les interactions (travaux du LECA en particulier). Après réflexion, et dans la continuité d'une démarche orientée « méthodes », j'ai obtenu en 2012-2013 un an de délégation à l'université Aix-Marseille, avec qui j'étais en collaboration continue depuis 15 ans, pour approfondir les aspects méthodologiques de la génomique environnementale. Pendant cette année, j'ai pu pousser cette logique jusqu'au bout avec un investissement à 100% de mon temps sur les activités de développement puisque déchargé des enseignements. Cette période a été riche d'enseignements et de réflexion sur le champ des possibles dans le croisement de la génomique et de l'écologie des interactions et a été à la source du positionnement thématique de mon projet scientifique présenté dans les pages suivantes.

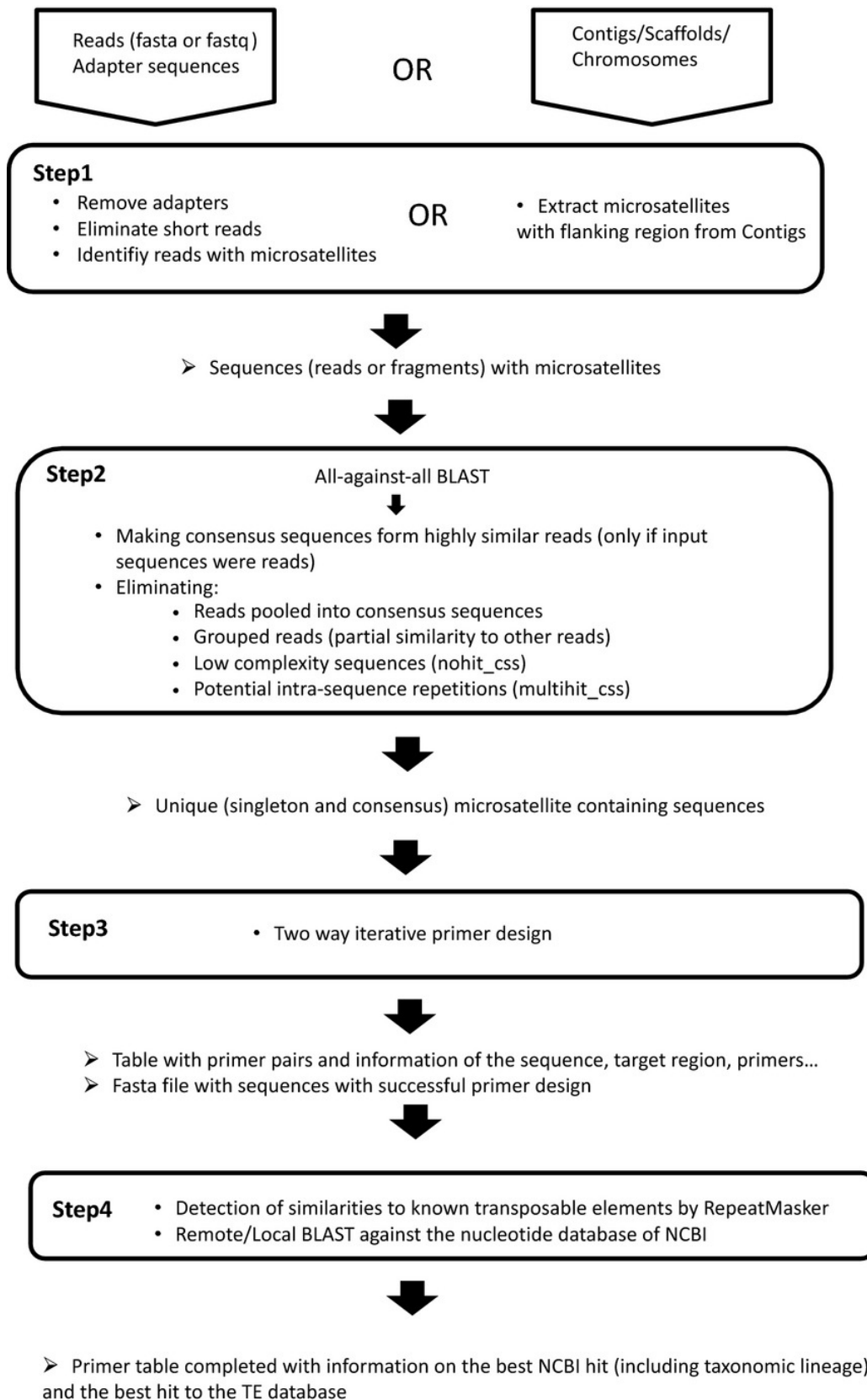


Fig. 9 : Organigramme du pipeline QDD détaillant les différentes étapes du traitement des données issues du séquençage haut débit des banques enrichies en motifs microsatellites. (Figure extraite de l'article #18).

## 3. LE PROJET DE RECHERCHE : LA GENOMIQUE ENVIRONNEMENTALE POUR ANALYSER LES INTERACTIONS DANS LES COMMUNAUTES

### 3.1 LE CADRE STRUCTUREL

Mon positionnement recherche est dans le cadre du **département Biologie et Ecologie de l'Institut Agro | Montpellier SupAgro**. Ce département se situe à l'interface de deux axes scientifiques : transitions agroécologique, ingénierie prédictive du vivant. Le **Centre de Biologie pour la Gestion des Populations (CBGP)** est une UMR composée de quatre tutelles (INRA, IRD, CIRAD et Montpellier SupAgro). Elle a pour vocation de comprendre les mécanismes qui régissent les populations d'organismes importants pour l'agriculture, l'environnement et la santé humaine. L'objectif finalisé du CBGP est de contribuer à l'amélioration de stratégies de lutte contre les espèces nuisibles et à l'identification de stratégies de conservation pour des populations naturelles menacées. L'unité a une structure atypique sans équipe et animée autour de groupes de réflexion thématiques. Ce fonctionnement a, de mon point de vue, des aspects positifs tels que le non cloisonnement dans une équipe, mais aussi des aspects négatifs comme l'isolement et le manque de stratégie et réflexion collective visible. Ce sentiment est accentué par les appels d'offre nationaux et internationaux qui s'appuient sur des réseaux à grande échelle. Paradoxalement il est bien plus simple de monter des projets avec des collègues hors unité qu'en son sein. L'ensemble de l'unité représente 86 personnels permanents, pour deux tiers des chercheurs. Nous sommes trois enseignants-chercheurs dans l'unité. Depuis quatre ans, le nombre croissant de contrats de recherche dans lequel je suis impliqué a entraîné la mise à disposition par l'UMR de J. Tavoillot, assistante ingénieur (en CDI à l'IRD) spécialisée dans l'acquisition de données moléculaires. D'autre part, cette croissance de mes activités de recherche sur contrat m'a permis de recruter une doctorante, Mélodie Ollivier (aujourd'hui Maître de conférence contractuelle à PURPAN), trois postdoctorants (l'un étant aujourd'hui Research Scientist au CSIRO, un second CR à l'INRA) une assistante ingénieur (dont le contrat vient de se terminer) et plusieurs étudiants de Master (2018). Ces étudiants et personnels (CDI ou CDD) sont sous ma responsabilité directe ou en coresponsabilité avec Marie-Stéphane Tixier.

### 3.2 ORIENTATION GENERALE

Les articles associés ou découlant de cette thématique : #1, 2, 4, 6, 8-13, 17, 20, 32, 48

Il y a six ans, à la faveur de ma mise en délégation d'un an, j'ai engagé une réflexion approfondie sur mon positionnement au sein du département. Ces discussions/réflexions avaient pour objectif de choisir entre deux options (parmi d'autres) à la fois conceptuelles, thématiques et méthodologiques distinctes :

- Me diriger vers la thématique « Amélioration des plantes » du département (cette thématique s'attache à comprendre les mécanismes aboutissant à la diversité des plantes cultivées afin de la modifier pour répondre aux enjeux en termes de transition agroécologique dans un contexte de changements globaux et d'économie / préservation des ressources). Cela aurait pu être dans la logique de mes activités autour de phylogéographie et la génétique des populations.
- Me diriger vers la thématique « Protection des plantes » du département. Les recherches dans ce thème ont pour objectif de comprendre les interactions biologiques entre les plantes, leurs bio-agresseurs et les ennemis naturels de ces derniers pour proposer des stratégies de protection des plantes permettant de réduire les pertes de production en garantissant des produits sains, en limitant les impacts environnementaux et en contribuant au travers de systèmes de production durables aux équilibres des écosystèmes, qu'ils soient fortement anthropisés ou non. Cette thématique faisait appel à la génomique environnementale que j'étais en train de développer à Marseille.

C'est cette seconde option que j'ai choisie suite à la prise en compte de plusieurs éléments résumés ci-après.

Tout d'abord je suis arrivé à un stade de mon parcours auquel je voulais (re)donner du sens sociétal à mes activités de recherche. Dans les deux cas, le cadre des transitions agroécologiques est parfaitement aligné avec cet objectif ainsi qu'avec l'ensemble de mes missions d'enseignement, transfert et recherche.

Dans ce champ très large, c'est l'étude des interactions écologiques à laquelle j'ai souhaité me consacrer car j'y vois un point d'entrée naturel vers la compréhension du fonctionnement des écosystèmes et par là-même vers l'optimisation des services écosystémiques, par exemple le contrôle biologique sur lequel je me focalise aujourd'hui. Les outils de la génomique environnementale sont puissants dans la mise au jour des interactions écologiques, trophiques en particulier car elle permet de détecter des interactions difficilement observables, soit parce qu'elles sont rares, dans des conditions d'observation complexes (dans le sol, chez les acariens par exemple), ou qu'elles sont indirectes. La génomique, appliquée à l'ADN environnemental ou dans mon cas au metabarcoding des régimes alimentaires des Arthropodes, vient donc apporter un éclairage objectif sur les interactions. Elle permet un niveau d'intégration des régimes alimentaires couvrant quelques heures tout en limitant les contraintes d'échantillonnage et/ou d'observation. En choisissant cette thématique, je pouvais donc mettre à profit le fond méthodologique que j'avais acquis en génomique pour pouvoir me focaliser rapidement sur les résultats biologiques et écologiques des résultats des analyses et aller à terme vers la compréhension du fonctionnement de écosystèmes.

Cet angle d'approche du fonctionnement des écosystèmes par l'analyse des réseaux d'interaction permet d'adresser un grand nombre de questions posées dans le champ



de l'écologie en général, de celui du contrôle biologique en particulier. Je me suis focalisé sur le fonctionnement des communautés d'Arthropodes, principalement dans les agrosystèmes. A cette échelle d'organisation de la biodiversité, on peut analyser finement la structure des interactions et le fonctionnement du réseau reconstruit lui-même [1, 2]. L'objectif de cette approche est aujourd'hui de déterminer la structure des réseaux d'interaction plantes/Arthropodes et Arthropodes/Arthropodes, de comparer les propriétés de ces réseaux dans des situations contrastées (invasion/natif, lutte chimique/lutte biologique) et d'en tirer des indications sur le fonctionnement de ces communautés [3] et sur l'optimisation des services écosystémiques associés (notamment le contrôle biologique, mais cela serait également transférable par exemple à la production végétale ou la pollinisation). Pour ce faire nous utilisons des approches de metabarcoding couplées à de l'analyse de réseau pour identifier les interactions trophiques, principalement sur la base de l'analyse des régimes alimentaires des Arthropodes. Cette thématique s'appuie sur la collaboration avec des entomologistes, des botanistes et des écologues. Elle nécessite également la maîtrise combinée des aspects moléculaires, de la bio-informatique et de l'analyse de réseau qui sont au cœur des développements de la phase actuelle de mon projet.

Les interactions trophiques sur lesquelles je me focalise permettent d'ouvrir une fenêtre d'exploration des processus écologiques qui produisent les services écosystémiques.

On pourrait sans doute trouver réducteur de décliner mon orientation dans le cadre des services écosystémiques qui ne représentent que des aspects anthropocentrés de la fonction écologique pour améliorer le bien-être humain. Au-delà de l'enjeu appliqué direct qui m'importe comme je l'ai explicité, je pense que cela peut aussi être justifié par le fait que les mécanismes écologiques qui sous-tendent les services écosystémiques englobent la complexité des interactions possibles entre les espèces. De plus, les relations entre la biodiversité et les mécanismes qui sous-tendent la plupart des services écosystémiques sont désormais bien établis et reconnus, tant empiriquement [4] que théoriquement [5]. Cependant la relation entre les pratiques agricoles et les services est mal comprise. L'une des raisons en est que les mécanismes écologiques qui sous-tendent les services écosystémiques englobent des interactions complexes entre les espèces, entre les espèces et les pratiques culturelles mais aussi les politiques des parties prenantes. On manque encore d'outils de modélisation mécaniste pour analyser et explorer l'effet des options de gestion sur la fourniture de multiples services écosystémiques. De tels outils sont nécessaires pour concevoir des systèmes de culture innovants à l'échelle des champs et des paysages, ainsi qu'à plus grande échelle pour planifier les futures politiques de gestion territoriale. Ces modèles basés sur des processus ou des mécanismes doivent également tenir compte de la complexité des interactions adressées par les réseaux écologiques qui sont une pierre nécessaire à la construction de cette connaissance. J'y reviendrai dans les perspectives.



J'illustrerai mes activités dans ce contexte par deux programmes de recherche actuellement en cours.

### **Effets de la bio-diversification sur la gestion des processus de lutte contre les ravageurs des cultures (au sein du programme STRADIV)**

Comprendre comment la diversité végétale et sa gestion modifient les processus biophysiques et écologiques est une condition préalable à la conception de systèmes biodiversifiés efficaces. L'intensité de la lutte contre les ravageurs dépend dans une large mesure de la structure communautaire dans laquelle les ravageurs et leurs ennemis naturels sont intégrés. La structure de la communauté végétale affecte d'autres communautés, avec des effets directs bottom-up (quantité et qualité des ressources primaires soutenant les réseaux alimentaires) et avec des effets indirects, notamment par la modification des habitats. Il n'est pas facile de démêler comment la diversité végétale altère le fonctionnement des communautés et la lutte contre les ravageurs, en raison de la difficulté de mesurer les interactions réelles sur le terrain. Nous utilisons dans ce programme le metabarcoding de marqueurs moléculaires pour identifier le lien trophique entre les consommateurs et les ressources (y compris les organismes nuisibles pour les plantes cultivées cibles du programme) au sein du réseau d'interactions trophiques de l'agroécosystème. Cette méthode conduit à une reconstruction complète de la structure du réseau alimentaire et permet une analyse complète des processus de régulation des ravageurs. L'utilisation de la méthode dans un ensemble de situations contrastées est une initiative ambitieuse en matière d'écologie des communautés appliquée aux agroécosystèmes. Comme la méthode est strictement normalisée en utilisant les mêmes protocoles sur tous les sites, notre ensemble de données permettra, nous l'espérons une compréhension générique de la relation entre la diversité végétale, la structure des interactions trophiques et la lutte contre les ravageurs. Nous nous concentrons sur les nuisibles les plus importants dans chaque type d'agroécosystème : les vers blancs du riz à Madagascar, le charançon du bananier en Martinique, le scolyte du caféier au Costa Rica. En parallèle, dans chaque situation étudiée, l'abondance des Arthropodes nuisibles est mesurée afin de quantifier le service écosystémique de contrôle assuré par la communauté. La structure des réseaux trophiques et le potentiel de régulation des ravageurs sont analysés en contrastant les options de gestion.

**Structure des réseaux d'interactions trophiques et implications pour la lutte biologique contre une espèce envahissante en Australie, *Sonchus oleraceus* (Asteraceae) dans le cadre de l'accord de collaboration avec le CSIRO** (ayant pour objectif de bâtir une approche intégrative du processus d'invasion et contribuer à la sélection des agents de lutte via l'analyse des réseaux écologiques).

Dans ce programme, nous tâchons de retracer les interactions écologiques pour décrire la composition et la structure de la communauté d'herbivores en interaction avec la plante cible du programme (*S. oleraceus*). L'analyse de réseaux sur la base de données moléculaires du régime alimentaire des Arthropodes améliore notre

compréhension du fonctionnement des écosystèmes grâce à l'échantillonnage direct rendu possible par cette approche et tire parti de la puissance des liens individuels sans les inconvénients de l'intégration dans le temps qu'impliquent les études fondées sur les isotopes. Elle précise également quelle est la spécificité de ces herbivores pour la plante cible et s'ils sont contrôlés par des ennemis naturels *in natura*. Enfin, la comparaison de la structure des réseaux trophiques locaux dans l'aire de distribution d'origine de la plante renseigne sur la présence de tendances générales au-delà des effets locaux liés aux conditions environnementales. De même, la comparaison des réseaux entre les aires indigènes et les aires envahissantes offre de nouvelles possibilités d'appréhender le processus invasif, et en particulier répondre à des questions telles que :

- la plante envahissante cible est-elle mieux maîtrisée par les herbivores dans la zone indigène que dans la zone envahissante ?
- Les réseaux écologiques des aires de distributions indigènes et envahissantes présentent-ils des guildes ou des familles comparables ?
- Quelles sont les interactions au niveau tritrophique (adventices/herbivores/ennemis naturels tels que prédateurs et/ou parasitoïdes) et quel est le rôle des agents potentiels de lutte biologique dans le réseau ?
- Enfin, l'analyse des réseaux et l'écologie des communautés visent à plus long terme à anticiper la perturbation engendrée par l'introduction de plantes envahissantes et d'agents potentiels dans un agroécosystème.

L'ensemble de ces défis et toutes ces questions utilisent le metabarcoding des herbivores comme une source de données pour décrire les réseaux d'interactions trophiques.

### 3.3 APPORTS DES RESEAUX D'INTERACTIONS

Comme introduit précédemment, l'analyse des réseaux d'interactions écologiques est une approche prometteuse pour comprendre l'assemblage des communautés basées sur des niches, reflétant la complexité des interactions entre les espèces et les processus écosystémiques sous-jacents [6]. De telles analyses peuvent renforcer notre compréhension des facteurs fondamentaux de l'assemblage des communautés [7, 8], des processus co-évolutifs [9], de la réponse des écosystèmes aux invasions biologiques et au changement global [10, 11], et de la gestion des services écosystémiques [12, 13].

L'analyse des réseaux d'interactions écologiques pourrait donc bénéficier à la régulation des espèces envahissantes, une discipline qui vise à réassocier une espèce envahissant un nouvel environnement et ses ennemis naturels spécialisés (c'est-à-dire les agents de contrôle biologique) pour la réguler. Bien que la compréhension des interactions entre les espèces soit préconisée depuis plus de 20 ans [14-16], l'évaluation des risques de la lutte biologique classique repose encore principalement

sur des tests expérimentaux. L'étude des réseaux d'interactions pourrait apporter un éclairage innovant à ces programmes de recherche en abordant les questions pratiques inhérentes au contrôle biologique (Figure 10 illustrant le cas du contrôle d'une plante envahissante). Cette section passe en revue les approches et les méthodes qui ont été utilisées pour répondre à ces questions et souligne le potentiel de l'analyse des réseaux écologiques dans le contexte d'espèces nuisibles et/ou envahissantes. Elle est illustrée dans le cadre de la régulation d'une espèce d'adventice envahissante mais la plupart des points relevés sont compatibles avec le contexte de ravageurs Arthropodes.

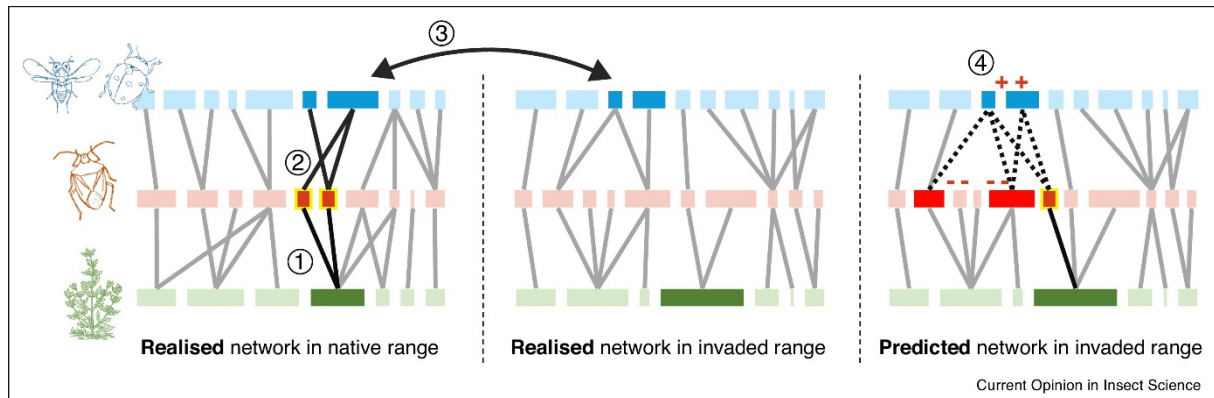


Fig. 10 : réseaux tripartites illustrant les interactions entre les communautés de plantes (en vert), d'herbivores (en orange) et d'ennemis naturels (en bleu) (composés de prédateurs et de parasitoïdes). De gauche à droite, le premier volet se trouve dans l'aire de distribution indigène, le second dans l'aire de distribution envahie. Le troisième est un réseau prédictif supposé dans l'aire de répartition envahie. Comme dans la représentation conventionnelle des réseaux tripartites, chaque espèce est représentée par un rectangle, dont la largeur reflète son abondance relative dans la communauté. L'analyse des réseaux vise à améliorer la sélection d'un agent de contrôle biologique (ACB) avec un risque minimal d'attaques non ciblées et d'effets indirects sur la communauté réceptrice. Le processus est divisé en plusieurs étapes. 1 : Rechercher les herbivores spécifiques à la plante cible (rectangles vert foncé) sur le terrain et déterminer les ACB potentielles (rectangles orange foncé). Déterminer les associations sur le terrain peut fournir des informations plus réalistes sur les interactions entre les espèces, que de se fier uniquement à des tests dans des conditions contrôlées. 2 : Identifier les ennemis naturels potentiels de ces ACB présumées. Les ennemis naturels pourraient i) menacer l'efficacité des ACB et ii) être la source d'effets indirects sur la communauté réceptrice par le biais d'interactions indirectes. 3 : Comparer les réseaux écologiques réalisés dans les aires de répartition indigènes et envahies en se basant sur i) les propriétés structurales et architecturales et ii) la taxonomie et les guildes. On s'attend à ce que les réseaux associés aux espèces cibles diffèrent entre les aires de répartition indigènes et envahies en termes de richesse en espèces, de guildes trophiques et de complexité. En outre, si des espèces taxonomiquement proches d'ennemis naturels sont trouvées entre les aires de répartition indigènes et les aires de répartition envahies (rectangles bleu foncé), un ACB introduit est plus susceptible d'être attaqué par ces nouveaux ennemis naturels. 4 : Prédire les associations d'espèces possibles suite au lâcher et à l'établissement d'un ACB. Le troisième réseau présente les effets indirects possibles (ligne pointillée) de l'ACB introduit via des parasitoïdes partagés avec des herbivores indigènes (rectangles rouges). Cette interaction indirecte (compétition apparente) est susceptible d'avoir des effets négatifs sur les herbivores indigènes et pourrait se répercuter sur d'autres niveaux trophiques. (Figure extraite de l'article #1)

## **Définir la spécificité des potentiels agents de contrôle biologique et leur gamme d'hôtes pour réduire les risques d'attaques non intentionnelles**

La lutte biologique classique contre les espèces d'Arthropodes nuisibles ou les plantes envahissantes utilise des ennemis naturels spécialisés de ces cibles pour réduire sélectivement la dynamique de leur population en dessous d'un seuil économique acceptable. Une première étape essentielle de ce processus est la compilation d'inventaires des ennemis naturels associés à la cible dans son aire de distribution naturelle. La spécificité d'un potentiel agent de Contrôle Biologique (ACB) est ensuite étudiée pour réduire le risque d'effets non intentionnels et indésirables sur les espèces non ciblées [17-20]. Ces tests sont généralement conçus selon la méthode phylogénétique centrifuge [21, 22] et réalisés dans des environnements standardisés, dans des conditions de choix et de non-choix. Cette approche prudente peut conduire à des faux positifs car la gamme d'hôtes réalisée sur le terrain est potentiellement plus restreinte que la gamme d'hôtes fondamentale explorée par l'expérimentation [23]. L'évaluation des risques uniquement dans des conditions expérimentales a toujours été considérée comme trop simpliste et l'accent est mis de plus en plus sur les évaluations de la gamme d'hôtes sur le terrain dans la zone d'origine [18]. Cela implique de caractériser les interactions dans des communautés diversifiées et d'être capable de décrire le spectre d'hôtes sur le terrain grâce à la construction de réseaux bipartite (par exemple la Figure 11).

**Illustration par le cas de *Sonchus oleraceus***, une plante envahissante des agrosystèmes en Australie, à travers la caractérisation du réseau trophique reconstruit dans la zone d'origine de la plante (Europe)

Au total, la reconstruction du réseau trophique basé sur la combinaison de l'analyse moléculaire du régime alimentaire et l'observation directe sur le terrain a conduit à identifier 47 taxons herbivores sur *S. oleraceus*, dont 37 identifiés au niveau de l'espèce (Figure 11).

Ces herbivores appartenait à cinq ordres différents : Hemiptera (c'est-à-dire pucerons, véritables insectes et cicadelles, 45 %), Diptères (25 %), Coléoptères (19 %), Lépidoptères (0,06 %) et Hyménoptères (0,04 %). Leur spectre d'hôtes sur le terrain (c'est-à-dire le nombre de plantes ressources distinctes par herbivore), telle que définie par notre échantillonnage, est représentée dans les Figures 11 et 12. Quinze taxons ont été prélevés exclusivement sur *S. oleraceus*, et deux autres taxons sur *S. oleraceus* et *Sonchus asper*. Ces taxons sont des agents candidats potentiels (spectre d'hôtes apparemment limitée au genre *Sonchus*, sous-groupe *Sonchinae*). Six autres espèces ont été détectées uniquement sur des membres de la tribu des *Chicorieae* (*Aphis craccivora* Koch, *Ophiomyia cunctata* Hendel, *Phytomyza lateralis* Fallén, *Campiglosa producta* Loew, *L. punctiventris* et *T. formosa*). Nous avons identifié 38 autres espèces de plantes comme ressources végétales complémentaires pour les espèces herbivores collectées chez *S. oleraceus*. La généralité (c'est-à-dire le nombre de ressources végétales) de ces espèces herbivores allait de 1 à 18, *Philaenus spumarius* L. étant la plus polyphage des 47 espèces herbivores trouvées sur *S. oleraceus* (Figure 12).

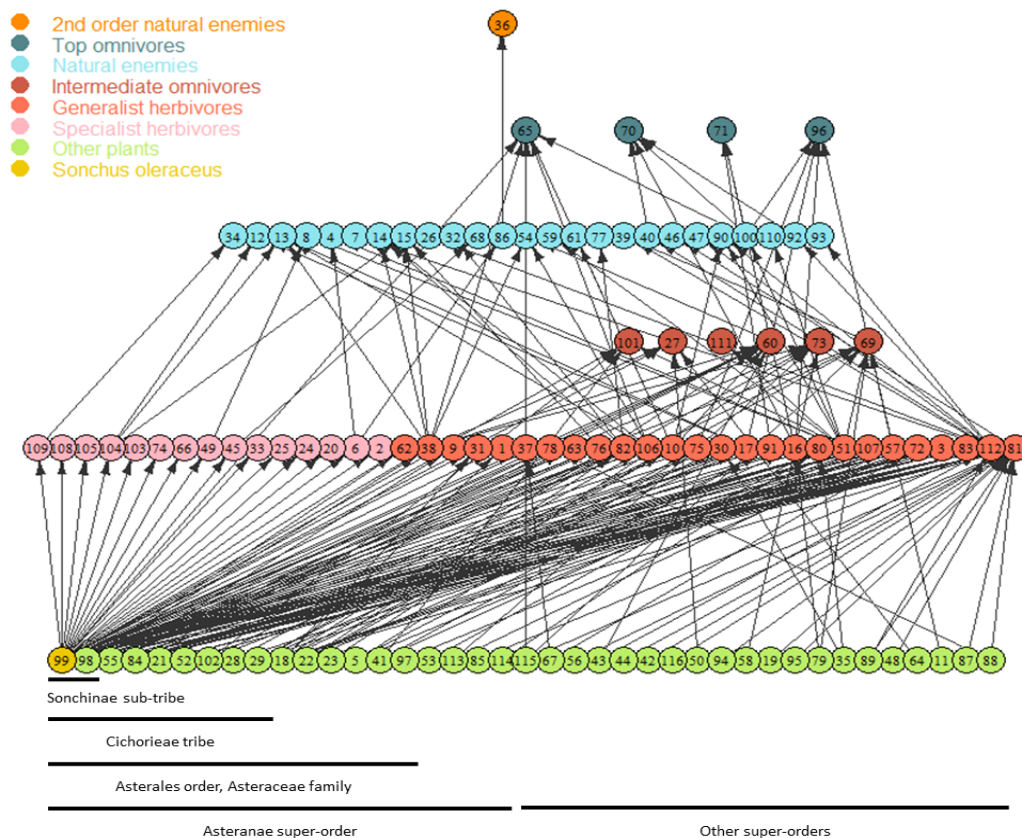


Fig. 11 : Réseau multitrophique reconstitué à partir de *S. oleraceus* (nœud jaune) et de 38 autres plantes (nœuds vert) utilisées par les herbivores de *S. oleraceus* (les nœuds rose représentent les herbivores spécialistes présumés, tandis que les nœuds orange représentent les herbivores non spécifiques de *S. oleraceus*). Les plantes sont classées selon leur parenté phylogénétique avec *S. oleraceus*. Les ennemis naturels des espèces herbivores sont représentés par des nœuds bleu clair. Les nœuds aux niveaux intermédiaires sont des espèces omnivores révélées par des analyses moléculaires. (Figure extraite d'un article en cours de révision)

Dans la lutte biologique contre les adventices, la spécialisation écologique des herbivores est un élément clé qui conditionne leur sélection en tant que potentiel ACB. Les préférences d'interaction réalisées se reflètent dans les modèles de réseau [24] car les contraintes écologiques et évolutives tendent à façonner la structure modulaire des réseaux (Modularité : groupes d'espèces, par exemple des modules, fortement associés à un ensemble particulier d'espèces végétales).

Le calcul des indices de spécialisation est couramment effectué pour les réseaux de pollinisation [25, 26] et prend en compte différents aspects de l'architecture, tant pour l'ensemble du réseau que pour les espèces individuelles. La spécialisation peut être mesurée en comptant le nombre de ressources par espèce (Généralité, voir Figure 11 par exemple) ou en quantifiant la dépendance d'une espèce à une ressource donnée (Force d'interaction), (mais voir [25] pour d'autres indices et leurs corrélations).

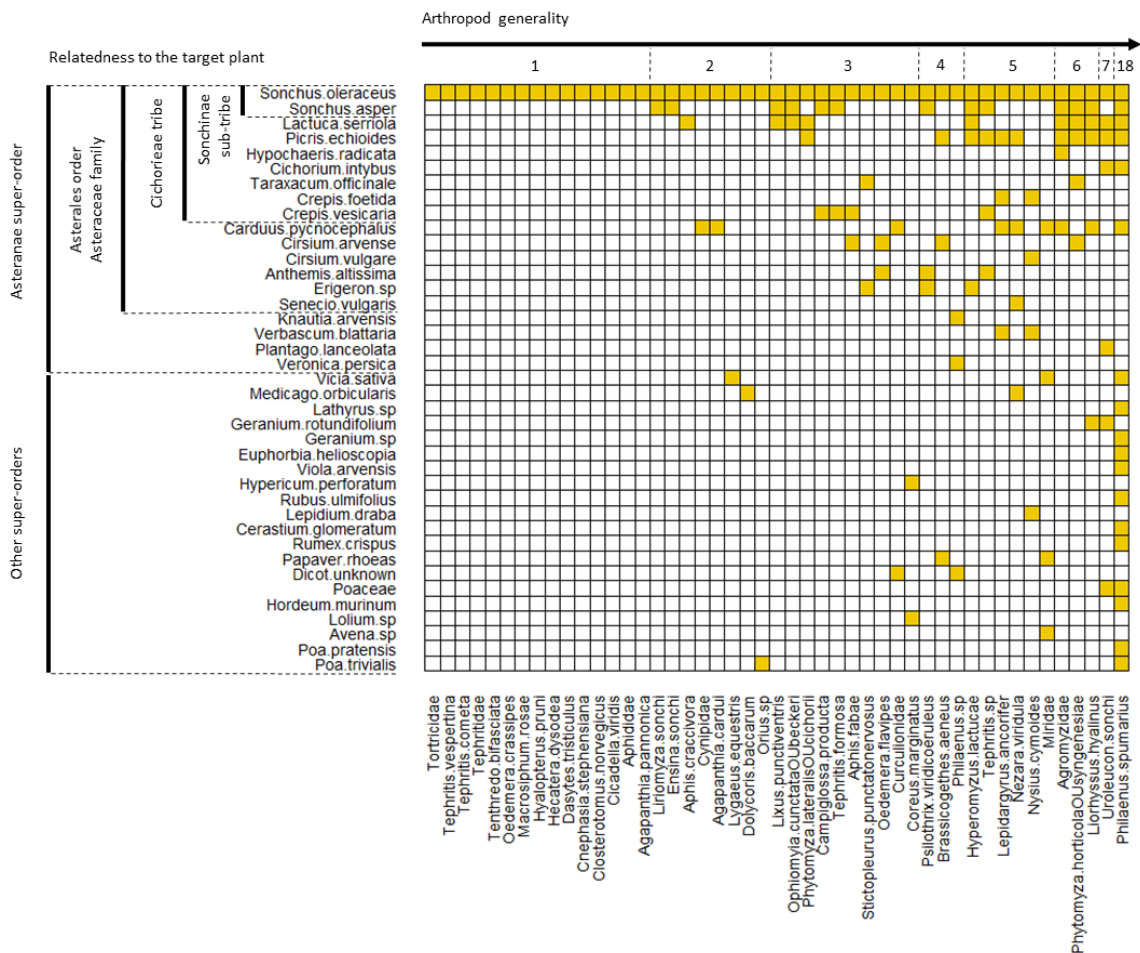


Fig. 12 : matrice d'interaction entre les herbivores échantillonnés sur *Sonchus oleraceus* et leurs plantes ressources utilisées, indiquant la gamme d'hôtes écologiques des herbivores telle que définie par l'échantillonnage intensif sur le terrain dans le sud de la France, au cours du printemps 2018. Les carrés jaunes représentent l'utilisation d'une plante par un Arthropode. Les plantes sont classées en fonction de leur parenté phylogénétique avec *S. oleraceus* et les arthropodes sont classés en valeurs de généralité croissante (c'est-à-dire le nombre de ressources par espèce). (Figure extraite d'un article en cours de révision)



Les schémas de spécialisation observés peuvent être avérés, mais aussi dus à une faible complétude de l'échantillonnage (nous y reviendrons dans la section suivante), ou à des différences intrinsèques dans l'attrait ou l'abondance des ressources. L'indice de spécialisation Distance-based Specialization Index (DSI), récemment développé [27], une extension de l'indice de spécificité des espèces (SSI), tient compte de la similarité phylogénétique et de l'abondance des espèces végétales hôtes (voir l'application des deux indices dans [28]). Cet indice prometteur tient également compte des différences d'abondance et de l'effort d'échantillonnage des consommateurs, ce qui permet des comparaisons solides entre les guildes d'herbivores.

L'amélioration des techniques moléculaires associée aux analyses de réseaux trophiques offre de multiples possibilités d'acquérir des connaissances sur les interactions qui se produisent dans les environnements naturels, et peut donc aider à caractériser le spectre d'hôtes *in natura* des candidats agents de contrôle biologique, et compléter ou orienter les tests expérimentaux de spécificité des hôtes.

### **Recherche des ennemis naturels des potentiels Agents de Contrôle Biologique pour améliorer les prévisions d'effets indirects**

L'ajout de parasitoïdes et de prédateurs pour ajouter une dimension aux réseaux bipartites est utile pour analyser l'influence de ce niveau trophique sur l'efficacité de l'ACB et pour détecter la probabilité d'effets indirects sur la dynamique de la communauté. Les connaissances sur les parasitoïdes des ACB sont généralement obtenues dans le cadre de l'élevage des candidats identifiés dans les enquêtes dans les aires de distribution indigènes. La caractérisation des prédateurs des ACB est plus difficile car l'observation directe est nécessaire. Le metabarcoding permet de détecter les proies dans le contenu digestif des Arthropodes, mais aussi les parasitoïdes à un stade précoce chez leurs hôtes. Dans la lutte biologique contre les insectes, l'utilisation de technologies moléculaires avancées pour la construction de réseaux écologiques a été récemment développée [29, 30] et serait directement transférable à la lutte biologique contre les adventices. L'exploitation d'organismes nouvellement introduits par des parasitoïdes de la communauté bénéficiaire est une association nouvelle qui a été prouvée dans le cadre d'invasions biologiques [31, 32]. De même, la prédation de l'ACB par des ennemis naturels indigènes est un phénomène qui a également été observé [14, 33, 34]. Ces découvertes confirment la capacité des organismes introduits à modifier la structure du réseau trophique. Bien qu'il ait été suggéré d'utiliser les analyses de réseau pour évaluer la sécurité des ACB après leur lâcher [14, 16], elles pourraient également être utiles pour les évaluations des effets indirects avant la lâcher.

**Illustration pour *Sonchus oleraceus*** de l'utilisation des réseaux trophiques reconstruits dans la zone d'origine de la plante (Europe) pour déterminer le cortège d'ennemis naturels des ACB *in natura*.

Dans le sous-réseau généré par les 39 taxons végétaux partageant au moins une espèce d'herbivore avec *S. oleraceus*, les herbivores associés et leurs ennemis

naturels, on a dénombré 116 nœuds et 213 interactions (Figure 11). Une analyse de ce sous-réseau a indiqué que les herbivores collectés sur *S. oleraceus* constituaient également une ressource pour divers ennemis naturels. En particulier, 19 des 47 taxons herbivores collectés ont été attaqués par plusieurs espèces de parasitoïdes (12 espèces de la famille des Braconidae, 1 de Figitidae et 1 d'Ichneumonidae) et de prédateurs (6 espèces d'Arachnida, 1 de Cantharidae, 2 de Coccinellidae, 3 de Syrphidae et 1 d'Orthoptera). Les analyses moléculaires ont révélé des modèles particuliers d'omnivorie impliquant plusieurs espèces d'Heteroptera. Nous avons distingué les omnivores intermédiaires (espèces se nourrissant à la fois de plantes et d'herbivores, comme les membres des Tephritidae et des Aphididae), et les omnivores supérieurs (espèces se nourrissant d'herbivores et d'ennemis naturels, comme les membres des Syrphidae). Ces ennemis naturels seront à comparer avec les guildes présentes en Australie (détectées par la même approche) pour estimer les risques d'effets indirects sur les communautés envahies.

Outre la compétition apparente, d'autres effets indirects sont identifiables grâce à l'analyse de réseau qui peut être utilisée pour rechercher des motifs particuliers impliquant des ACB. Malgré leur valeur potentielle, l'usage de ces analyses dans l'évaluation préalable au lâcher dans les programmes de régulation des nuisibles ou de adventices garde une grande marge de progression [35].

### **Comparer les réseaux écologiques entre des situations contrastées**

Comparer des situations contrastées, par exemple liées à des différences de pratique culturales ou bien sûr aux facteurs abiotiques ou encore liées aux contingences évolutives est un défi majeur de l'écologie des interactions car elle permet de répondre à des questions centrales en écologie dont les retombées sont directement applicables. Par d'autres contextes, l'introduction d'organismes dans des communautés établies ouvre la perspective de nouvelles associations créées dans la communauté bénéficiaire. L'évaluation de l'ampleur des modifications causées par ces perturbations nécessite une comparaison des différentes situations étudiées. Par exemple, les réseaux trophiques associés aux espèces végétales envahissantes sont supposés 1) être composés d'espèces plus généralistes et 2) être moins diversifiés (au niveau des herbivores et des niveaux trophiques supérieurs) que la structure de la communauté indigène [36, 37].

Comparer des réseaux d'interactions n'est pas une question triviale qui peut se régler sur la base de l'observation directe des réseaux, même dans les cas simples (Figure 13 pour s'en convaincre)



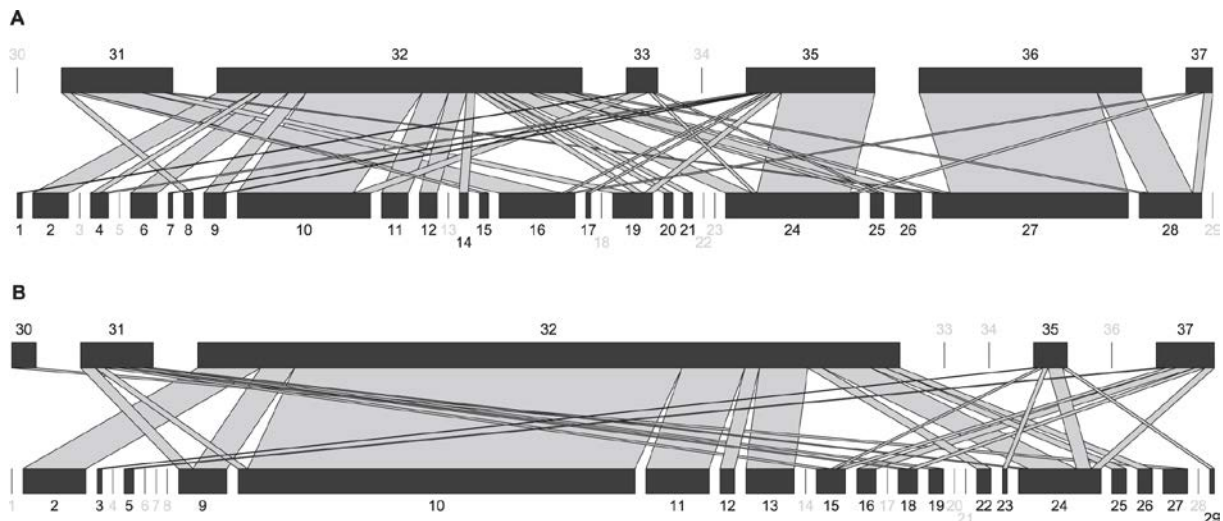


Fig. 13 : Réseaux trophiques bipartites des interactions prédateur-proie sur (A) sol nu, et (B) bananeraies cultivées en couverture. (Figure extraite de l'article #17)

Un outil classique de comparaison de communautés se base sur les différents indices de diversité spécifiques et d'équitabilité. La comparaison des compositions de taxons utilise l'indice de similarité Bray-Curtis de manière très courante car il permet d'évaluer la différence de composition des espèces entre deux échantillons en tenant compte des données d'abondance [38-40]. Cependant cet indice ne prend pas en compte par exemple la phylogénie des taxa impliqués ou les éléments de diversité fonctionnelle tout aussi pertinents dans ce cadre que la diversité taxonomique.

D'autre part, les taxa sont parfois différents d'une situation à l'autre, ce qui crée une difficulté d'interprétation des différences. On s'intéresse alors à la représentativité comparée des guildes et des « espèces trophiques » dans le but de focaliser les analyses sur les interactions elles-mêmes et sur leurs significations écologiques.

Les comparaisons structurelles des réseaux trophiques reposent quant à elles sur l'utilisation d'indices descripteurs des réseaux pour extraire des informations sur les propriétés des espèces (par exemple, le ratio proies/consommateurs, la proportion d'espèces par niveau trophique), les propriétés des liens (par exemple, la densité des liens, la connectivité) et les asymétries entre les proies et les consommateurs (par exemple, la généralité, la vulnérabilité) [41]. Ces mesures peuvent permettre une meilleure compréhension des interactions écologiques potentielles des situations comparées mais restent le plus souvent centrées sur des indicateurs globaux, au mieux sur des niveaux trophiques particuliers.

Enfin, depuis plus récemment, les réseaux d'interactions écologiques sont comparés sous l'angle des motifs qui les composent, marqueurs directs de types d'interactions précises. La combinaison de la fréquence de chaque motif dans chaque réseau et l'agglomération des distances ainsi calculées entre les réseaux comparés permet la hiérarchisation des similarités entre réseaux sur la base de la distribution des motifs,

ce qui semble donner plus de support aux fonctions des interactions dans le processus de comparaison des situations (Figure 14).

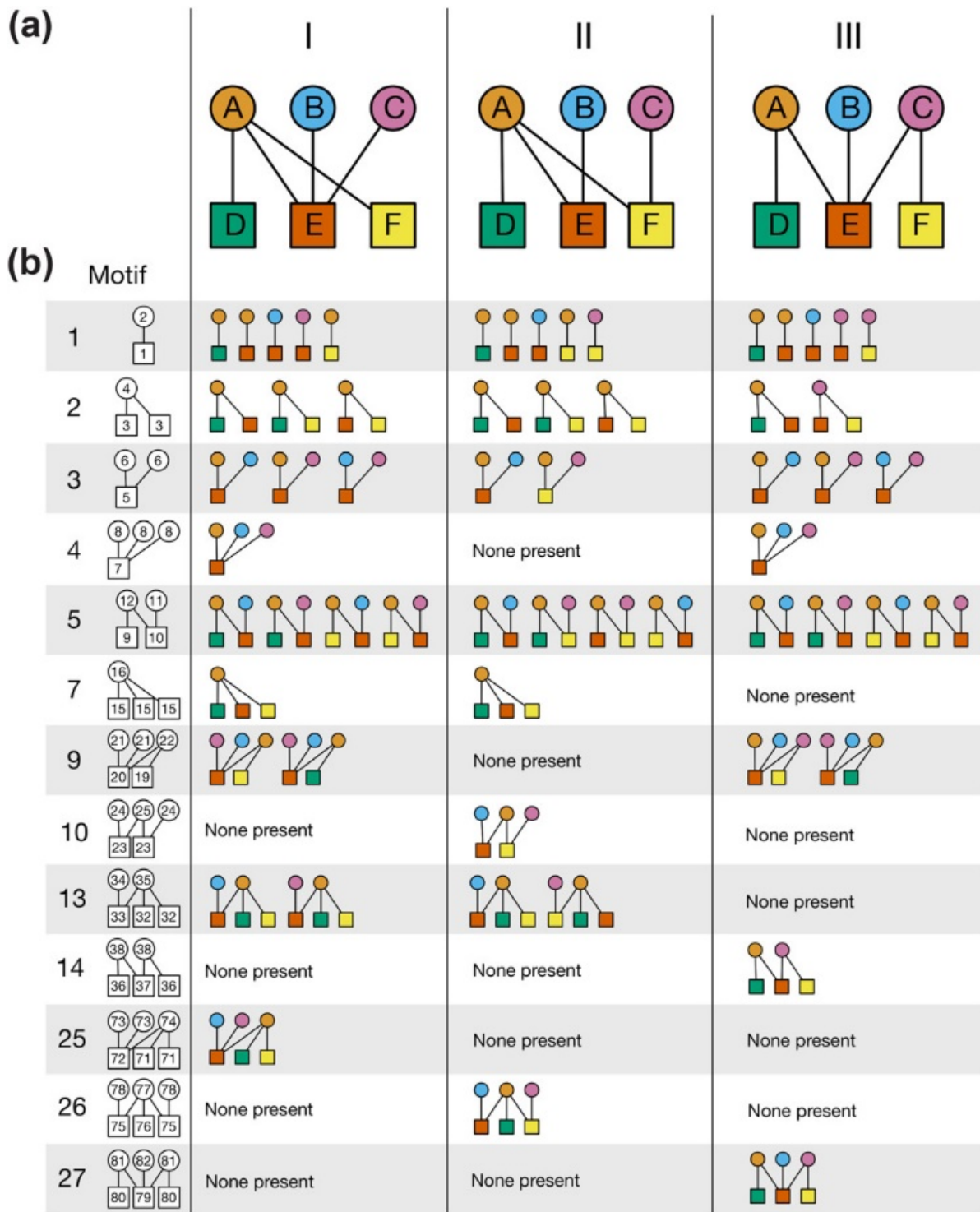


Fig. 14 : Décomposition de trois exemples de réseaux en leurs motifs constitutifs de deux à six nœuds. (a) Trois exemples de réseaux. (b) Tableau indiquant les motifs constitutifs de chaque réseau. La première colonne indique le motif compté : le grand nombre renvoie à l'ID du motif ; le petit nombre à l'intérieur de chaque nœud renvoie aux positions uniques que les espèces peuvent occuper à l'intérieur de chaque motif. La deuxième, troisième et quatrième colonne montre les occurrences de chaque motif dans les réseaux I, II et III respectivement. Les couleurs des nœuds se réfèrent aux espèces impliquées

dans chaque motif. Pour des raisons de visualisation, sont exclus les motifs qui ne sont présents dans aucun réseau. (Figure reproduite de [42])

### **Prédire les interactions pour évaluer les risques par la modélisation**

Des modèles prédictifs des réseaux trophiques sont également bienvenu dans le domaine de la régulation des ravageurs ou des adventices [43]. La combinaison d'une structure de réseau trophique avec des modèles de population dynamiques génèrerait des réseaux trophiques dynamiques qui pourraient améliorer notre compréhension et notre capacité à prédire les changements dus à l'introduction d'espèces [43-45]. Dans une étude récente [46], un modèle de réseau a été proposé pour évaluer les effets directs et indirects de différentes méthodes de régulation sur la dynamique des efflorescences algales. Les nœuds clés du réseau ont été identifiés comme particulièrement efficaces pour contrôler les efflorescences algales, et de fortes influences indirectes ont été observées entre les groupes fonctionnels. Cette méthodologie pourrait être adaptée et le développement sophistiqué de modèles dans des domaines de recherche étroitement liés à la biologie des invasions [47] et à la gestion des écosystèmes [48, 49] serait également transférable au domaine de la régulation des adventices ou des ravageurs et participeraient de ce fait de l'amélioration de la compréhension du fonctionnement des écosystèmes.

### **3.4 LA DEMARCHE ET LES PROBLEMATIQUES ASSOCIEES**

La pertinence des réseaux écologiques dépend de la fiabilité des données et des méthodes utilisées pour les construire. Dans ce processus allant de l'échantillonnage à l'interprétation des réseaux, la qualité de chaque étape est cruciale dans la fiabilité et la représentativité du résultat final. L'image de la rupture de la chaîne du froid est assez juste pour comparer l'impact d'une approximation dans l'une des étapes qui aboutit à la reconstruction d'un réseau à partir de données récoltées sur le terrain. Dans cette section, je vais m'attacher à résumer les difficultés rencontrées et la démarche que nous avons utilisé pour les résoudre autant que faire se peut et tenter de maximiser la fiabilité des réseaux d'interactions reconstruits.

Ces difficultés ont une conséquence immédiate : comme l'échantillonnage incomplet et la capacité générale à révéler avec précision les interactions entre les espèces peuvent introduire un biais dans la majorité des descripteurs de réseau [24], les analyses et les comparaisons des réseaux résultants exigent des praticiens qu'ils soient pleinement conscients des pièges et des possibilités qu'offre la méthode choisie. La construction de réseaux écologiques fiables nécessite donc de connaître les avantages et les limites inhérents à chaque méthode afin de choisir la méthodologie adaptée au système étudié et aux questions de recherche abordées. La révélation des liens trophiques réalisés repose traditionnellement sur des techniques à forte intensité de main-d'œuvre basées sur des observations directes sur le terrain[50], l'élevage ou

des dissections microscopiques du contenu intestinal et des fèces [51]. Tout en fournissant des informations comportementales significatives, ces approches présentent des limites majeures lorsqu'on travaille sur des espèces souterraines ou nocturnes et empêchent l'étude alimentaire des insectes qui se nourrissent de sève [52]. Les approches basées sur les empreintes d'alcane des plantes, l'électrophorèse des protéines du contenu intestinal, l'analyse des isotopes stables, la détection des protéines des proies à l'aide d'anticorps polyclonaux et monoclonaux (ELISA) et les méthodes basées sur l'ADN peuvent aider à surmonter les obstacles de l'identification visuelle [15, 53, 54]. Cependant, les performances de ces techniques dépendent du contexte [52]. Les empreintes d'alcane des plantes et l'électrophorèse des protéines ne sont pas adaptées pour refléter l'étendue du régime alimentaire des espèces généralistes (fournissant des motifs de bandes chevauchantes non interprétables) [55]. Les études d'enrichissement isotopique ont l'avantage de fournir des informations sur des échelles temporelles plus longues, en intégrant les flux énergétiques passés plutôt que le repas le plus récent. Cependant, les signatures isotopiques sont sujettes à des variations entre les espèces qui peuvent conduire à interprétations de liens trophiques incohérents et peu clairs [56]. Après les techniques basées sur l'ADN, l'approche des anticorps monoclonaux est la deuxième méthode la plus utilisée pour l'évaluation des réseaux trophiques en agriculture [57]. Les antigènes de proies présentent l'avantage d'être détectables plus longtemps après leur consommation [58], par rapport à la dégradation rapide de l'ADN des proies dans l'intestin des consommateurs. Bien que les antigènes soient de bons marqueurs pour détecter la consommation d'une proie spécifique par une série de prédateurs, ils ne sont pas adaptés aux analyses complexes des réseaux trophiques car leur développement serait coûteux et long [59].

Les méthodes basées sur l'ADN, qui sont au cœur de notre méthodologie pour acquérir les données d'interactions sur le terrain, sont de plus en plus utilisées dans l'élucidation de réseaux trophiques en agriculture [51, 57, 60]. Le plus souvent, le metabarcoding [61] associé aux technologies de séquençage haut débit [62] nous offre la possibilité de traiter efficacement un grand nombre d'échantillons. Par exemple, à partir d'un échantillon de contenu digestif d'Arthropodes prédateurs prélevé sur le terrain, il est possible de reconstruire son régime alimentaire et de révéler des interactions difficiles à observer, telles qu'hôte-parasitoïde, quel que soit le stade de vie de l'insecte [63, 64]. Cependant, ces méthodes sont également sujettes à des sources d'erreurs potentielles [65] qui sont l'objet de cette section.

**Le dispositif d'échantillonnage et le stockage** peuvent produire des interactions faussement positives (par le biais de contaminations externes, de prédation secondaire ou de fouille des cadavres) [29, 66, 67]. Le piégeage de masse a été éliminé de nos processus dès 2014 car sujet à la prédation secondaire et totalement artificielle dans les pièges. Il est préférable de prélever les insectes individuellement, à l'aide d'un aspirateur à embouts uniques ou directement à la main avec des pinces stériles [67]. Cette méthode à forte intensité de temps peut être adaptée en limitant le

temps de collecte à des périodes standard sur chaque site de collecte, normalisant ainsi l'effort d'échantillonnage pour des comparaisons ultérieures entre les sites. Cependant quel que soit le temps passé, les limites du raisonnable s'opposent bien souvent à une **complétude satisfaisante de l'échantillonnage** pour qu'il soit représentatif des interactions réelles. Soit parce que les périodes d'échantillonnage sont réduites (à l'échelle de la journée, de la saison et de la dynamique de la communauté), et/ou parce qu'il reflète une activité observable à un moment  $t$  qui ne reflète pas nécessairement l'ensemble des interactions. La dimension spatiale est souvent incomplètement traitée également pour les mêmes raisons de coût humain dans l'acquisition des données de terrain. Pour essayer de mesurer cette complétude, nous utilisons une adaptation des courbes d'accumulation en espèces (bien connues dans la caractérisation des échantillonnages d'espèces), modifiée pour visualiser l'accumulation des interactions par paires d'espèces [68] (Figure 15). Ces courbes sont une indication précieuse sur le niveau de complétude de l'échantillonnage directement au niveau des interactions et permettent soit de tenter de compléter celui-ci en fonction de l'objectif, soit de relativiser la représentativité des résultats.

### Accumulation analysis - Interactions

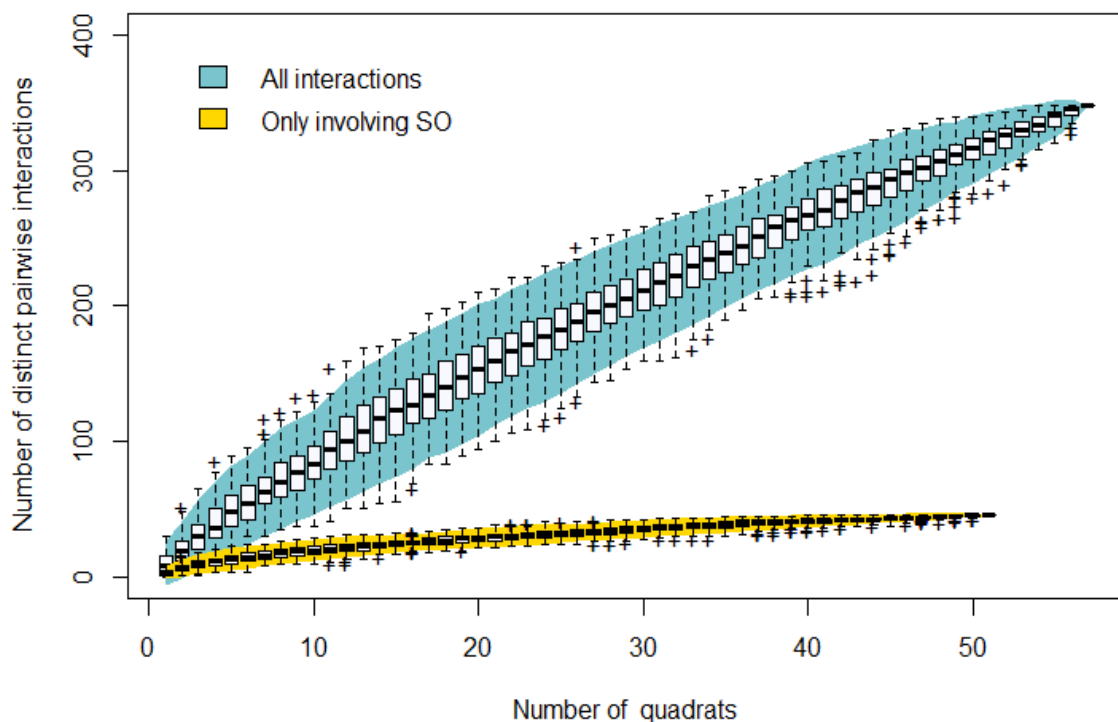


Fig. 15 : Sous-réseau généré par les 39 taxons végétaux mentionnés dans la section précédente, leurs herbivores associés et leurs ennemis naturels, était composé de 116 nœuds et 213 interactions (Figure 11). Une analyse de ce sous-réseau a indiqué que les herbivores collectés sur *S. oleraceus* constituaient également une ressource pour divers ennemis naturels. La courbe d'accumulation impliquant *S. oleraceus* atteint un plateau (visible en normalisant les échelles) alors que celle des interactions globales n'atteint pas ce plateau et relativise donc la représentativité des interactions en dehors de celles impliquant *S. oleraceus*. (Figure extraite d'un article en cours de révision)





Figure 16 : Représentation de la dissimilarité des communautés (distances de Bray-Curtis) en deux dimensions par une mise à l'échelle multidimensionnelle non métrique (NMDS). Pour les communautés végétales (a et c), la valeur du stress est de 0,14 (avec k=3) et pour les communautés d'arthropodes (b et d), la valeur du stress est de 0,17 (avec k=2). Les différences entre les régions sont indiquées à gauche : la région semi-océanique en rouge, la région continentale en bleu et la région méditerranéenne en vert. À droite, les différences entre les trois sessions d'échantillonnage sont indiquées : la date 1 (avril) en bleu clair, la date 2 (mai) en gris clair et la date 3 (juin) en gris foncé. (Figure extraite d'un article en cours de révision)

Outre le caractère incomplet de l'échantillonnage [24], **la stabilité et la détectabilité de l'ADN** sont également une source de faux négatifs. Des tests de sensibilité peuvent être utilisés pour évaluer combien de temps après l'ingestion une proie ou une plante l'ADN peut être détecté dans l'intestin du consommateur [80]. Par ailleurs, pour maximiser la conservation des échantillons avant l'extraction de l'ADN, nous avons choisi de les stocker dès leur capture sur le terrain directement dans le tampon de lyse utilisé dans le protocole d'extraction d'ADN total et stocké à 4°C jusqu'à l'extraction.

Le protocole d'extraction d'ADN lui-même a été repensé à zéro avec plusieurs objectifs en tête :

- Augmenter le taux d'extractions d'ADN réussies (au sens qu'elles permettent l'amplification du ou des marqueurs cibles) au-delà des 80% que nous obtenions avec les kits commerciaux (Qiagen)
- Permettre une automatisation de l'ensemble de la chaîne de traitement physico-chimique des échantillons

Nous avons opté, en collaboration avec l'UMR AGAP et Sylvain Santoni en particulier (Montpellier), pour une remise à plat du protocole sur les bases fondamentales de biochimie, avec une phase de lyse comprenant une protéase et une cellulase (adaptée aux cellules végétales), une capture de l'ADN sur billes magnétiques à surface recouverte de silice *via* des processus robotisables, et des purifications robotisables également sur les mêmes matériels [69, 70]. Cet investissement dans la méthode a permis à une personne (Johanne Tavoillot) d'extraire en quelques semaines plus de 5 000 échantillons avec un taux de réussite à la PCR de 97% en moyenne selon le marqueur.

Une fois l'ADN des échantillons extrait, **la détection du régime alimentaire** est un défi en soit. La diversité des consommateurs disqualifie l'approche par bloqueur qui pourrait être utile pour empêcher l'amplification du marqueur utilisé pour le metabarcoding chez un consommateur particulier. Il a donc fallu miser sur la profondeur de séquençage pour extraire les séquences des proies ingérées de l'ensemble des séquences de l'échantillon. L'arrivée récente des plateformes Illumina Novaseq a permis en cinq ans de diviser le prix des séquences par 15, ce qui permet d'accéder à une profondeur très importante pour un très grand nombre d'échantillons simultanément.

Avant le séquençage toutefois, il reste deux verrous. Le premier est **la sélection du marqueur moléculaire** (metabarcode) ou le design d'un couple d'amorces pour amplifier le marqueur d'intérêt choisi pour sa variabilité et la disponibilité des banques de référence taxonomiques. Même si le gène mitochondrial *Cytochrome Oxydase I* (COI) est généralement reconnu pour ses propriétés dans la délimitation des espèces d'arthropodes [71], il n'existe pas de paire d'amorces universelles permettant d'amplifier l'ensemble des Arthropodes. Nous avons donc utilisé la combinaison de trois jeux d'amorces pour amplifier le gène COI avec une large couverture taxonomique [67, 72-75] (Figure 17). Une étape technique en cours de développement est le multiplexage de ces amplifications pour réduire les coûts.

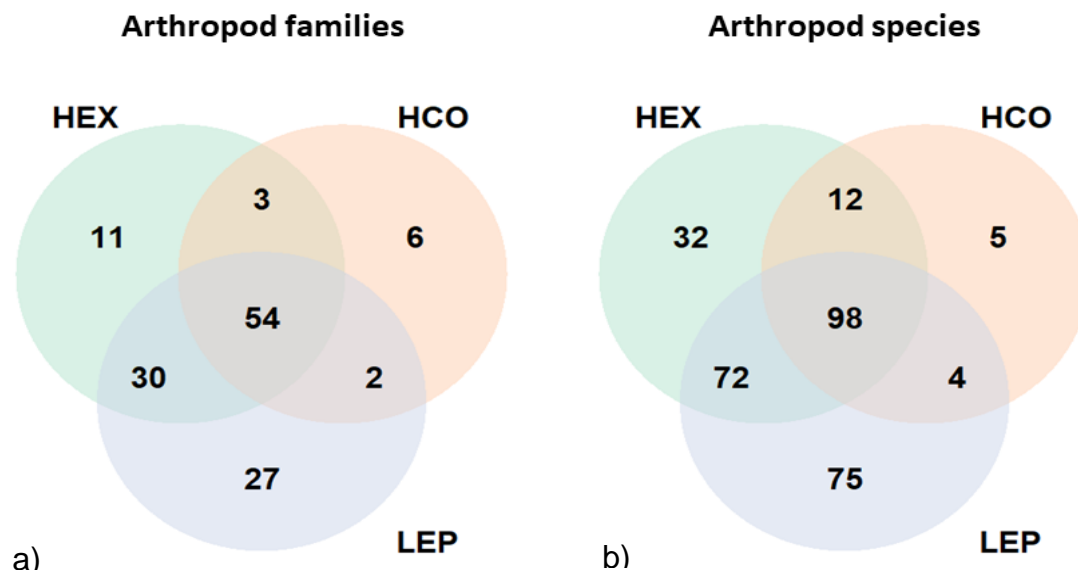


Fig. 17 : Diagramme de Venn de la couverture taxonomique obtenue avec les trois marqueurs COI utilisés dans l'analyse des communautés d'Arthropodes en interaction avec *S. oleraceus*, représentée par a) la diversité des familles d'arthropodes et b) la diversité des espèces d'arthropodes.

Pour les plantes, il est relativement consensuel de considérer que de multiples marqueurs végétaux sont nécessaires pour déterminer les espèces végétales [76]. Nous sommes partis dans cette direction avec deux marqueurs classiques (rbcL et matK) qui n'ont pas donné de résultats satisfaisant en termes de polymorphisme utilisable à l'échelle spécifique ou même générique. Ce défi reste à résoudre aujourd'hui (il y a un potentiel pour les approches utilisant la capture d'exons en multiplex même si les quantités d'ADN sont généralement limitantes pour assurer la reproductibilité de la capture). Cela nous a contraint à ne prendre en compte que les interactions Arthropodes/Plantes réellement observées dans les parties internes des plantes disséquées individuellement, procédure extrêmement chronophage (trois fois plus longue que l'échantillonnage sur le terrain lui-même) mais qui, par sa démarche conservative, permet d'éviter les faux positifs.



**Le multiplexage massif d'échantillons** pour rentabiliser les opérations de séquençage nous a demandé d'innover dans le domaine du « tagging moléculaire » utilisé pour identifier chaque échantillon après séquençage. Nous avons défini des étiquettes de 9 bases [77] maximisant leur diversité et dissimilarité (distance de Hamming) ce qui nous a permis de réduire le taux de chimères artificielles de 7,3% (admises dans la littérature et corroborée par notre expérience [78]) à 0.41%. Cet effort nous a permis d'éliminer virtuellement tout risque d'interaction artificielle liée à une erreur de lecture des étiquettes ou à la création de recombinants pendant les phases de construction des bibliothèques Illumina. Enfin, nous avons repris complètement le processus de traitement des données de séquence depuis la sortie du séquenceur jusqu'à l'assignation des variants identifiés afin de fiabiliser le processus, l'automatiser entièrement et permettre sa parallélisation et ainsi assurer la scalabilité du traitement quel que soit le nombre de séquences [79, 80] (Figure 18 pour un schéma résumé).

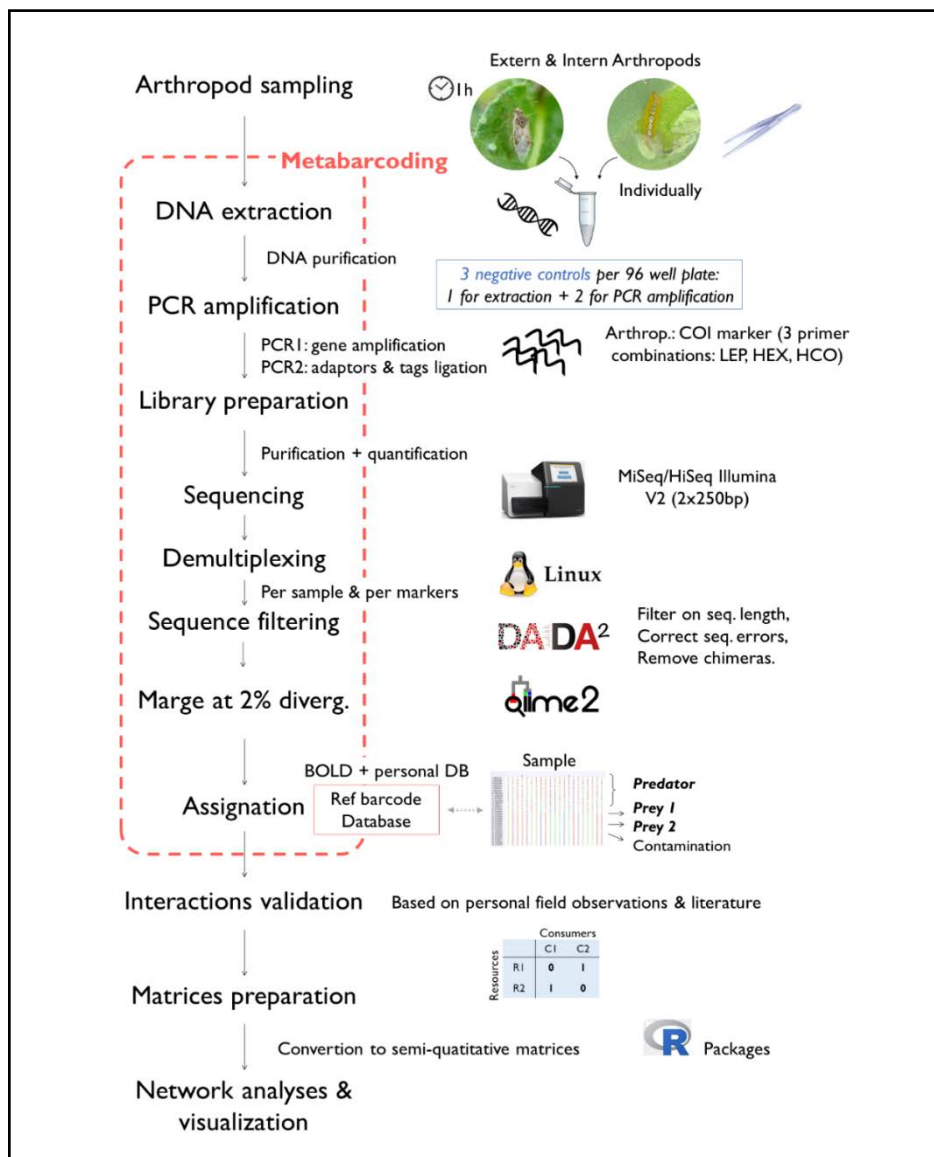


Fig. 18 : Aperçu de l'ensemble du processus de reconstruction des interactions entre les plantes, les herbivores et les ennemis naturels. (Figure extraite d'un article en révision)

La phase la plus sensible du traitement des données est sans aucun doute **l'identification précise des variants** (assignation). Elle dépend, au-delà de l'algorithme utilisé, principalement de la qualité et de l'exhaustivité de la base de données de référence interrogée [81, 82]. Malheureusement, les banques de référence publiques (BOLD, Genbank) ont un taux d'erreur taxonomique ne permettant pas de leur faire une confiance aveugle. C'est d'ailleurs le cas aussi de la propre banque de référence du laboratoire (plusieurs milliers d'espèces d'Arthropodes présentes en Europe) qui n'est pas exempte de contradictions malgré des efforts de qualification considérables. Pour tenter de résoudre ce problème, nous avons choisi d'adopter une démarche d'assignation taxonomique hybride utilisant à la fois une banque de référence créée pour chaque programme, la banque de référence du CBGP et BOLD. Nous comparons les assignations dans les trois banques de références et chaque ambiguïté est levée manuellement par une démarche experte accompagnée par les entomologistes de l'unité.

Finalement, **la transformation des données assignées en matrice de contiguïté**, qui permet la reconstruction des réseaux et leur analyse, nécessite à ce jour une étape de validation manuelle et faisant appel à l'écologie, les données recueillies sur le terrain, l'entomologie et la botanique. Elle s'appuie également sur la littérature disponible donnant un cadre aux interactions plausibles [83]. Ce processus est couteux en temps et son objectivation résumée dans la Figure 19 a été rendue nécessaire pour éviter au maximum des effets liés à l'analyste. Finalement, **l'utopie de quantifier les proportions relatives des proies dans le régime alimentaire** d'un individu doit être abandonnée car de multiples raisons (dégradation différentielle et différences de taille pour ne citer que les deux plus évidentes) expliquent l'absence de lien entre le nombre de séquences obtenues pour un taxon et le nombre d'individus correspondants ou leur proportion dans le régime alimentaire. A ce stade, seule la fréquence d'occurrence d'une proie a du sens au niveau des espèces et c'est bien cette information que nous conservons dans les analyses qui utilisent des approches quantitatives pour pondérer les interactions.

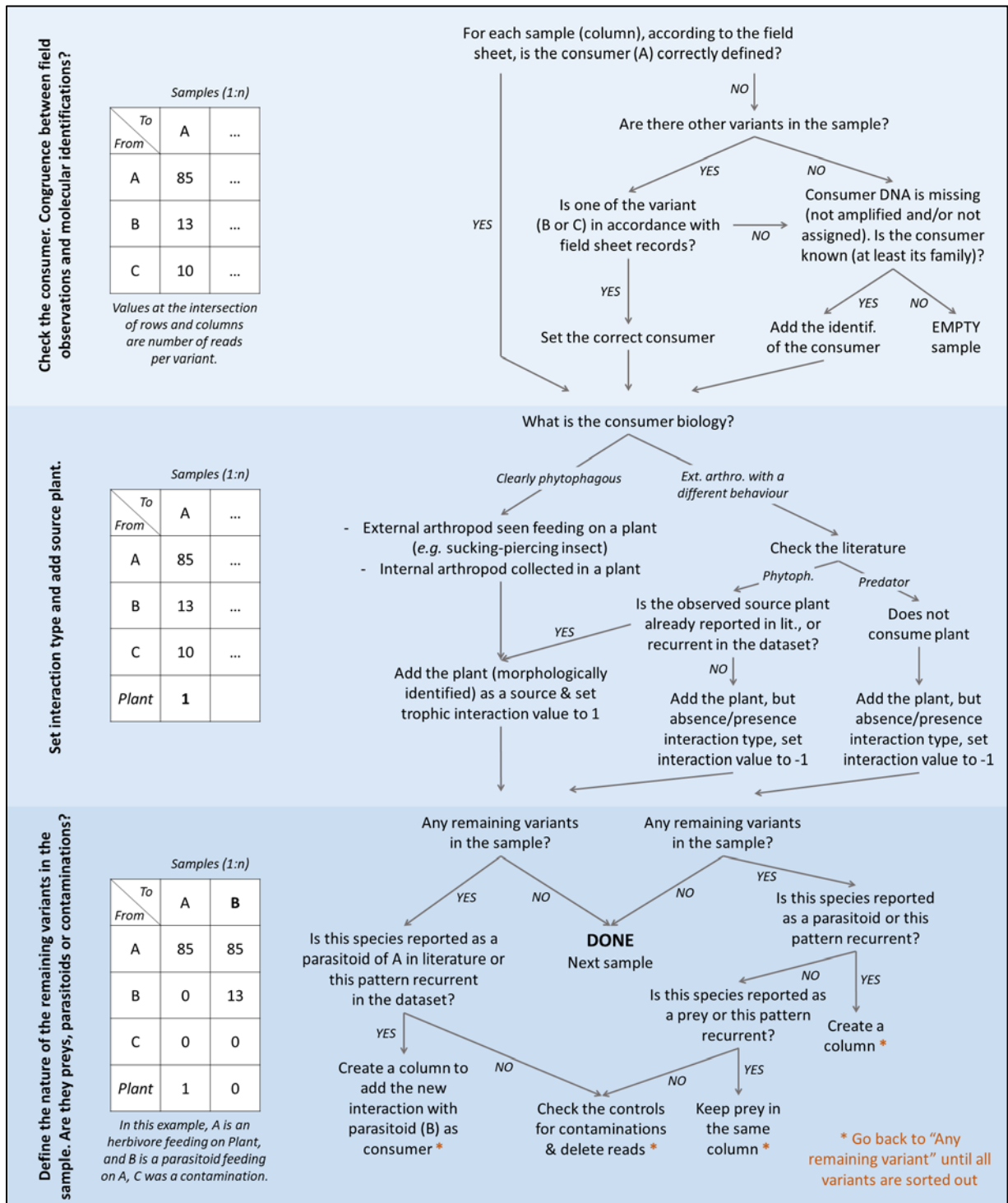


Fig. 19 : Schéma présentant le processus suivi pour la transformation des données brutes du séquençage en une matrice de contiguïté.

La représentativité et la fiabilité des données utilisées pour reconstruire et analyser les réseaux est un enjeu majeur qui nous a demandé des efforts et un investissement considérable, tant humain que financier. Ces efforts ont abouti à la faveur de la thèse de Mélodie Ollivier dont les premiers réseaux ont permis de faire des progrès significatifs vis-à-vis des dernières difficultés liées la conversion des assignations des variants en matrices de contiguïté. Aujourd'hui je pense qu'on peut dire que nous avons l'expertise dans l'acquisition et le traitement des données pour construire les réseaux d'interactions écologiques pour les relations Arthropodes/Arthropodes et dans une certaine mesure Arthropodes/Plantes. Dans ce processus nous avons appris l'importance de l'hybridation des approches moléculaires et plus traditionnelles (observation directe sur le terrain, dissection au labo, mise en élevage de larves par exemple) qui sont précieuses dans la validation des assignations et la construction de la matrice de contiguïté.

Cet aspect de développement méthodologique est contrasté car il témoigne de la nécessité d'avoir investi dans la refonte du processus d'acquisition mais il a mobilisé toute notre attention ce qui a freiné notre capacité à pleinement utiliser puis développer cette approche dans l'analyse du fonctionnement des communautés en interaction. Je veux donc dans la partie suivante, pour conclure, introduire en perspectives quelles sont les pistes d'évolution du projet de recherche à différents termes et tracer les dimensions potentielles qu'il pourrait explorer.

## 4 - REFLEXION SUR LES ACTIVITES ET PERSPECTIVES

Le bilan des six dernières années est très positif sur le plan qualitatif avec le développement d'une thématique claire, innovante et utile tant sur les plans académiques que plus appliqués. J'ai pu trouver les subventions adéquates pour soutenir cette activité de reconstruction et d'analyse des réseaux d'interactions écologiques et construire assez rapidement un groupe et des collaborations solides et efficaces dont le travail commence à porter ses fruits en termes de valorisation de la recherche (12 de mes 15 dernières publications depuis 2014 sont centrées sur cette thématique). Notre courbe collective (j'inclue l'ensemble du groupe : ITA, doctorante et postdoctorants) de montée en compétence dans l'analyse des interactions écologiques a une belle dynamique en perspective et les premiers résultats sont très encourageants et reconnus par la communauté dans les meetings (pris de la meilleure présentation dans le meeting ICE à Sydney en septembre dernier pour Mélodie Ollivier par exemple). Tous les voyants sont au vert là aussi donc pour valoriser l'expérience dans une phase maintenant tournée vers l'analyse et le fonctionnement des écosystèmes.

---

### PREMIERE ETAPE A COURT TERME : LE TEMPS DE L'IMMERSION DANS LES « NETWORK SCIENCES »

Maintenant que nous avons obtenu des données dont la fiabilité est grandement améliorée, je peux me consacrer pleinement à l'analyse de réseaux et viser la maîtrise analytique dans le contexte des programmes en cours, en particulier celui concernant la régulation de *S. oleraceus*, mais aussi pour le programme STRADIV dont les données sont acquises. De même, nous avons développé un programme sur les interactions chez les communautés d'acariens prédateurs/proies dont les données sont produites également. Ce que j'entends par « maîtriser », c'est connaître et savoir appliquer le panel de méthodes liées à l'analyse des réseaux pour répondre efficacement aux questions biologiques posées par les programmes qui les utilisent comme une approche du fonctionnement des communautés. Cet objectif est assez personnel car il assoit le socle de compétences de bases nécessaire au développement des perspectives à plus long terme. Les réseaux ont des caractéristiques universelles qui ont été approfondies et étudiées dans de très nombreux champs disciplinaires. Mon objectif est bien entendu d'approfondir mes compétences dans l'analyse des réseaux d'interactions écologiques. Cependant, la généricité des mécanismes sous-tendus par les réseaux en fait un outil dont la portée va bien au-delà de l'écologie des interactions et des communautés. J'irai donc explorer les champs très divergents utilisant ces réseaux, y compris hors Sciences de la Vie, pour en retirer les usages et les enseignements remobilisables dans le contexte de ce projet. L'objectif pragmatique est d'être indépendant et je l'espère pertinent dans l'analyse d'ici 2022 et à partir de là de progresser régulièrement dans l'expertise, sur

le modèle de ce que nous faisons dans le domaine de l'acquisition des données, moléculaires en particulier où les premières étapes ont été franchies collectivement.

---

## UN SECOND TEMPS : EXPRIMER PLEINEMENT LE POTENTIEL DE L'APPROCHE RESEAUX D'INTERACTIONS DANS LE CADRE DU LIEN BIODIVERSITE-FONCTIONNEMENT DES ECOSYSTEMES

La littérature est riche dans le domaine du rôle que l'analyse des réseaux d'interactions écologiques ont dans l'étude des relations Biodiversité-Fonctionnement des écosystèmes [84-88]. En effet, les réseaux trophiques en particulier, forment une représentation intégrée des relations entre espèces constituant les communautés en intégrant leurs dynamiques des populations, l'évolution des traits d'histoire de vie, la diversité taxonomique, phylogénétique et fonctionnelle, et l'expression des filtres écologiques dans l'assemblage des communautés. Ces réseaux sont des ponts entre l'écologie et la biologie évolutive qui permettent d'accéder à des fractions de la compréhension du fonctionnement des écosystèmes. Cette richesse conceptuelle s'étend dans de multiples domaines de recherche de l'écologie qui sont associés à autant de dimensions évolutives et cette matrice du champ des possibles impose de faire des choix de positionnement à moyen terme. Il me semble que pour ma part, une évolution du projet à moyen terme peut avoir deux directions qui me semblent cohérentes avec mon parcours et objectifs généraux, ces directions n'étant pas incompatibles bien entendu.

La première serait de générer, combiner et analyser des réseaux d'interactions écologiques déclinant l'ensemble des grandes interactions présentes dans les communautés (en restant centré sur les agroécosystèmes). Allier les informations des réseaux construits sur les relations Arthropodes/Arthropodes, Arthropodes/Plantes, Plantes/communautés bactériennes pour ne citer que celles-ci, produirait une vision en « méta-réseau » qui permettrait l'étude des processus de régulation à cette échelle alors qu'ils sont encore dans la majorité des cas fragmentés dans des équipes disciplinaires spécialisées dans un organisme d'étude spécifique ou focalisé sur un service écosystémique particulier. De façon complémentaire, on peut également imaginer d'incorporer les aspects de diversité génétique intraspécifique, par exemple des plantes cultivées comme nous nous apprêtons à le faire dans le cadre du programme ANR PPR Mobidiv (Collaboration avec l'UMR AGAP pour cet aspect).

En contrepoint de cette approche empirique, la seconde option serait d'utiliser les outils de la modélisation pour étudier la relation Diversité / fonctionnement de l'écosystème au-delà des indices de diversité globaux qui ne prennent pas en compte l'identité et la fonction des espèces dans le réseau. Le besoin de prédiction de l'impact de perturbations sur la structure et la dynamique des réseaux d'interactions fait également sens dans la démarche d'utilisation des réseaux pour optimiser les services écosystémiques et, par exemple, minimiser le risque lié au lâcher d'agents de contrôle

biologiques exotiques dans des communautés locales. D'un point de vue théorique, ces développements permettraient de participer de la compréhension des processus de régulation des communautés, enrichie des possibilités conjointe de manipulation expérimentale des acteurs du réseau dans les systèmes relativement simples caractéristiques des agrosystèmes.

Dans les deux *scenarii* d'évolution à long terme de ce projet de recherche (une multitude d'autres options sont susceptibles d'émerger d'ici là !), la constante est bien la compréhension des mécanismes régissant le fonctionnement des écosystèmes. Celle-ci impliquera la multiplicité et la complémentarité des compétences nécessaires pour mener à bien le projet. Cette complémentarité des acteurs s'exprime dans le spectre taxonomique comme disciplinaire et implique un réseau de collaborations efficaces sur le long terme, et partageant au moins partiellement l'objectif. La communauté scientifique de Montpellier, et au-delà sur ces thématiques au niveau global se prête particulièrement bien à ces deux options étant donné le panel de compétences existantes et la réelle volonté dans les discussions au quotidien d'intégrer ces compétences dans un collectif au service de la compréhension du fonctionnement de ces communautés au sens large.

Dans ce contexte, les compétences à développer sont non seulement celles d'un expert dans une des disciplines (pour ma part il est acquis que j'ai débuté un glissement de focale depuis l'écologie moléculaire vers l'écologie des interactions), mais aussi celles d'un facilitateur d'émergence et de réalisation de programmes de recherche dans lesquels pourront s'exprimer ces complémentarités, qu'elles soient disciplinaires ou relatives à l'objet d'étude. Mon profil universitaire issu d'une formation en Sciences Naturelles (qui n'existe plus, ayant laissé la place à des disciplines plus spécifiques) a probablement un rôle dans ma perception de l'importance de ce positionnement.

En termes de formation et de direction des recherches des futurs chercheurs et scientifiques en général, j'ai à cœur de promouvoir et de rendre tangible cette capacité de développement permanent de son expertise dans son domaine disciplinaire. De même, il est crucial que les futurs scientifiques puissent, par leur curiosité, se nourrir des apports de champs qui ne sont pas dans son cœur de compétences. Cette impulsion vers l'ouverture supportée par un socle disciplinaire solide est à mon sens centrale dans mon rôle de direction de recherche. Evidemment, de par mes fonctions dans mon établissement, je ne saurais conclure sans préciser que parmi les valeurs que j'estime nécessaires à la conduite de la Recherche, l'intégrité scientifique et la transparence sont des éléments centraux de ce que je vise à promouvoir et développer chez les collaborateurs, étudiants ou non avec qui j'ai la chance de travailler.



## BIBLIOGRAPHIE

1. Bartomeus, I., et al., *A common framework for identifying linkage rules across different types of interactions*. *Functional Ecology*, 2016. **30**(12): p. 1894-1903.
2. Thébault, E. and C. Fontaine, *Stability of Ecological Communities and the Architecture of Mutualistic and Trophic Networks*. *Science*, 2010. **329**(5993): p. 853-856.
3. Lewinsohn, T.M., et al., *Structure in plant–animal interaction assemblages*. *Oikos*, 2006. **113**(1): p. 174-184.
4. Letourneau, D.K., et al., *Does plant diversity benefit agroecosystems? A synthetic review*. *Ecological Applications*, 2011. **21**(1): p. 9-21.
5. Isbell, F., et al., *High plant diversity is needed to maintain ecosystem services*. *Nature*, 2011. **477**(7363): p. 199-U96.
6. Borrett, S.R., J. Moody, and A. Edelman, *The rise of Network Ecology: Maps of the topic diversity and scientific collaboration*. *Ecological Modelling*, 2014. **293**: p. 111-127.
7. Pimm, S.L., J.H. Lawton, and J.E. Cohen, *Food web patterns and their consequences*. *Nature*, 1991. **350**(6320): p. 669-674.
8. Barraclough, T.G., *How Do Species Interactions Affect Evolutionary Dynamics Across Whole Communities?* *Annual Review of Ecology, Evolution, and Systematics*, 2015. **46**(1): p. 25-48.
9. Bascompte, J. and P. Jordano, *Plant-Animal Mutualistic Networks: The Architecture of Biodiversity*. *Annual Review of Ecology, Evolution, and Systematics*, 2007. **38**(1): p. 567-593.
10. Traveset, A. and D.M. Richardson, *Mutualistic Interactions and Biological Invasions*. *Annual Review of Ecology, Evolution, and Systematics*, 2014. **45**(1): p. 89-113.
11. Seibold, S., et al., *The Necessity of Multitrophic Approaches in Community Ecology*. *Trends in Ecology & Evolution*, 2018. **33**(10): p. 754-764.
12. Dee, L.E., et al., *Operationalizing Network Theory for Ecosystem Service Assessments*. *Trends in Ecology & Evolution*, 2017. **32**(2): p. 118-130.
13. McDonald-Madden, E., et al., *Using food-web theory to conserve ecosystems*. *Nature Communications*, 2016. **7**(1): p. 10245.
14. Memmott, J., *Food Webs as a Tool for Studying Nontarget Effects in Biological Control*, in *Nontarget Effects of Biological Control*. 2000: Boston, MA. p. 147-163.
15. Sheppard, S.K. and J.D. Harwood, *Advances in molecular ecology: Tracking trophic links through predator-prey food-webs*. *Functional Ecology*, 2005. **19**(5): p. 751-762.
16. Willis, A.J. and J. Memmott, *The potential for indirect effects between a weed, one of its biocontrol agents and native herbivores: A food web approach*. *Biological Control*, 2005. **35**(3): p. 299-306.
17. Suckling, D.M. and R.F.H. Sforza, *What Magnitude Are Observed Non-Target Impacts from Weed Biocontrol?* *PLoS ONE*, 2014. **9**(1): p. e84847.



18. Hinz, H.L., R.L. Winston, and M. Schwarzländer, *How Safe Is Weed Biological Control? A Global Review of Direct Nontarget Attack*. The Quarterly Review of Biology, 2019. **94**(1): p. 1-27.
19. Louda, S.M., et al., *Ecological Effects of an Insect Introduced for the Biological Control of Weeds*. Science, 1997. **277**(5329): p. 1088-1090.
20. Louda, S.M., et al., *Nontarget effects--the Achilles' heel of biological control? Retrospective analyses to reduce risk associated with biocontrol introductions*. Annual Review of Entomology, 2003. **48**: p. 365-396.
21. Corcket, E., B. Giffard, and R.F.H. Sforza, *Food Webs and Multiple Biotic Interactions in Plant–Herbivore Models*. Advances in Botanical Research, 2017. **81**: p. 111-137.
22. Briese, D.T., *Translating host-specificity test results into the real world: The need to harmonize the yin and yang of current testing procedures*. Biological Control, 2005. **35**(3): p. 208-214.
23. Fowler, V.S., et al., *How can ecologists help practitioners minimize non-target effects in weed biocontrol?* Journal of Applied Ecology, 2012. **49**(2): p. 307-310.
24. Dormann, C.F., J. Fründ, and H.M. Schaefer, *Identifying Causes of Patterns in Ecological Networks: Opportunities and Limitations*. Annual Review of Ecology, Evolution, and Systematics, 2017. **48**(1): p. 559-584.
25. Dormann, C.F., *How to be a specialist? Quantifying specialisation in pollination networks*. 2011.
26. Memmott, J., *The structure of a plant-pollinator food web*. Ecology Letters, 1999. **2**(5): p. 276-280.
27. Jorge, L.R., et al., *Phylogenetic trophic specialization: a robust comparison of herbivorous guilds*. Oecologia, 2017. **185**(4): p. 551-559.
28. Redmond, C.M., et al., *High specialization and limited structural change in plant-herbivore networks along a successional chronosequence in tropical montane forest*. Ecography, 2019. **42**(1): p. 162-172.
29. Derocles, S.A.P., et al., *Biomonitoring for the 21st Century: Integrating Next-Generation Sequencing Into Ecological Network Analysis*. Advances in Ecological Research, 2018. **58**: p. 1-62.
30. Mollot, G., et al., *Cover Cropping Alters the Diet of Arthropods in a Banana Plantation: A Metabarcoding Approach*. PLoS ONE, 2014. **9**(4): p. e93740.
31. Herron-Sweet, C.R., et al., *Native parasitoids associated with the biological control agents of *Centaurea stoebe* in Montana, USA*. Biological Control, 2015. **86**: p. 20-27.
32. Murillo Pacheco, H., et al., *Food web associations and effect of trophic resources and environmental factors on parasitoids expanding their host range into non-native hosts*. Entomologia Experimentalis et Applicata, 2018. **166**(4): p. 277-288.
33. Pearson, D.E. and R.M. Callaway, *Indirect effects of host-specific biological control agents*. Trends in Ecology & Evolution, 2003. **18**(9): p. 456-461.

34. Pearson, D.E. and R.M. Callaway, *Indirect nontarget effects of host-specific biological control agents: Implications for biological control*. *Biological Control*, 2005. **35**(3): p. 288-298.
35. Carneiro, L.G., et al., *Apparent competition can compromise the safety of highly specific biocontrol agents*. *Ecology Letters*, 2008. **11**(7): p. 690-700.
36. Memmott, J., et al., *The invertebrate fauna on broom, Cytisus scoparius, in two native and two exotic habitats*. *Acta Oecologica*, 2000. **21**(3): p. 213-222.
37. Cornell, V.H. and B.A. Hawkins, *Accumulation of native parasitoid species on introduced herbivores: a comparison of hosts as natives and hosts as invaders*. *The American naturalist*, 1993. **141**(6): p. 847-65.
38. Delmas, E., et al., *Analysing ecological networks of species interactions*. *Biological Reviews*, 2019.
39. Eitzinger, B., et al., *Assessing changes in arthropod predator–prey interactions through DNA-based gut content analysis—variable environment, stable diet*. *Molecular Ecology*, 2019.
40. Ximenes Pinho, B., W. Dáttilo, and I.R. Leal, *Structural breakdown of specialized plant-herbivore interaction networks in tropical forest edges*. *Global Ecology and Conservation*, 2017. **12**: p. 1-8.
41. Bersier, L.-F., C. Banašek-Richter, and M.-F. Cattin, *Quantitative descriptors of Food-web matrices*. *Ecology*, 2002. **83**(9): p. 2394-2407.
42. Simmons, B.I., et al., *bmotif: a package for motif analyses of bipartite networks*. *bioRxiv*, 2018: p. 302356.
43. Barratt, B.I.P., et al., *Progress in risk assessment for classical biological control*. *Biological Control*, 2010. **52**(3): p. 245-254.
44. Tylianakis, J.M. and A. Binzer, *Effects of global environmental changes on parasitoid–host food webs and biological control*. *Biological Control*, 2014. **75**: p. 77-86.
45. López-Núñez, F.A., et al., *Four-trophic level food webs reveal the cascading impacts of an invasive plant targeted for biocontrol*. *Ecology*, 2017. **98**(3): p. 782-793.
46. Mao, X., et al., *An ecological-network-analysis based perspective on the biological control of algal blooms in Ulansuhai Lake, China*. *Ecological Modelling*, 2018. **386**: p. 11-19.
47. Romanuk, T.N., et al., *Predicting invasion success in complex ecological networks*. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 2009. **364**(1524): p. 1743-54.
48. Pires, M.M., *Rewilding ecological communities and rewiring ecological networks*. *Perspectives in Ecology and Conservation*, 2017. **15**(4): p. 257-265.
49. Baker, C.M., et al., *A novel approach to assessing the ecosystem-wide impacts of reintroductions*. *Ecological Applications*, 2019. **29**(1): p. e01811.
50. Sander, E.L., J.T. Wootton, and S. Allesina, *Ecological Network Inference From Long-Term Presence-Absence Data*. *Scientific reports*, 2017. **7**(1): p. 7154.

51. Evans, D.M., et al., *Merging DNA metabarcoding and ecological network analysis to understand and build resilient terrestrial ecosystems*. Functional Ecology, 2016. **30**(12): p. 1904-1916.
52. Pompanon, F., et al., *Who is eating what: Diet assessment using next generation sequencing*. Molecular Ecology, 2012. **21**(8): p. 1931-1950.
53. De Barba, M., et al., *DNA metabarcoding multiplexing and validation of data accuracy for diet assessment: application to omnivorous diet*. Molecular Ecology Resources, 2014. **14**(2): p. 306-323.
54. Wirta, H.K., et al., *Complementary molecular information changes our perception of food web structure*. Proceedings of the National Academy of Sciences, 2014. **111**(5): p. 1885-1890.
55. Symondson, W.O.C., *Molecular identification of prey in predator diets*. Molecular Ecology, 2002. **11**(4): p. 627-641.
56. Traugott, M., et al., *Evaluating <sup>15</sup>N/<sup>14</sup>N and <sup>13</sup>C/<sup>12</sup>C isotope ratio analysis to investigate trophic relationships of elaterid larvae (Coleoptera: Elateridae)*. Soil Biology and Biochemistry, 2007. **39**(5): p. 1023-1030.
57. González-Chang, M. and M.-C. Lefort, *Food webs and biological control: A review of molecular tools used to reveal trophic interactions in agricultural systems*. Food Webs, 2016. **9**: p. 4-11.
58. Schenk, D. and S. Bacher, *Detection of shield beetle remains in predators using a monoclonal antibody*. Journal of Applied Entomology, 2004. **128**(4): p. 273-278.
59. Chen, Y., et al., *Identifying key cereal aphid predators by molecular gut analysis*. Molecular ecology, 2000. **9**(11): p. 1887-98.
60. Roslin, T. and S. Majaneva, *The use of DNA barcodes in food web construction—terrestrial and aquatic ecologists unite!* Genome, 2016. **59**(9): p. 603-628.
61. Taberlet, P., et al., *Towards next-generation biodiversity assessment using DNA metabarcoding*. Molecular Ecology, 2012. **21**(8): p. 2045-2050.
62. Shendure, J. and H. Ji, *Next-generation DNA sequencing*. Nature Biotechnology, 2008. **26**(10): p. 1135-1145.
63. Garipey, T.D., T. Haye, and J. Zhang, *A molecular diagnostic tool for the preliminary assessment of host-parasitoid associations in biological control programmes for a new invasive pest*. Molecular Ecology, 2014. **23**(15): p. 3912-3924.
64. Hrček, J. and H.C.J. Godfray, *What do molecular methods bring to host–parasitoid food webs?* Trends in Parasitology, 2015. **31**(1): p. 30-35.
65. Alberdi, A., et al., *Scrutinizing key steps for reliable metabarcoding of environmental samples*. Methods in Ecology and Evolution, 2018. **9**(1): p. 134-147.
66. King, R.A., et al., *Suction sampling as a significant source of error in molecular analysis of predator diets*. Bulletin of Entomological Research, 2012. **102**(03): p. 261-266.

67. King, R.A., et al., *Molecular analysis of predation: A review of best practice for DNA-based approaches*. *Molecular Ecology*, 2008. **17**(4): p. 947-963.
68. Jordano, P., *Sampling networks of ecological interactions*. *Functional Ecology*, 2016. **30**(12): p. 1883-1893.
69. Brady, J.A., et al., *High-throughput DNA isolation method for detection of Xylella fastidiosa in plant and insect samples*. *J Microbiol Methods*, 2011. **86**(3): p. 310-2.
70. Cruaud, A., et al., *Using insects to detect, monitor and predict the distribution of Xylella fastidiosa: a case study in Corsica*. *Sci Rep*, 2018. **8**(1): p. 15628.
71. Hebert, P.D.N., et al., *Biological identifications through DNA barcodes*. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 2003. **270**(1512): p. 313-321.
72. Brandon-Mong, G.J., et al., *DNA metabarcoding of insects and allies: an evaluation of primers and pipelines*. *Bulletin of Entomological Research*, 2015. **105**(6): p. 717-727.
73. Folmer, et al., *DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates*. *Molecular Marine Biology and Biotechnology*, 1994. **3**(5): p. 294-299.
74. Leray, M., et al., *A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents*. *Frontiers in Zoology*, 2013. **10**(1): p. 34.
75. Marquina, D., A.F. Andersson, and F. Ronquist, *New mitochondrial primers for metabarcoding of insects, designed and evaluated using in silico methods*. *Molecular Ecology Resources*, 2019. **19**(1): p. 90-104.
76. Zhu, C., D. Gravel, and F. He, *Seeing is believing? Comparing plant–herbivore networks constructed by field co-occurrence and DNA barcoding methods for gaining insights into network structures*. *Ecology and Evolution*, 2019.
77. Martin, J.-F., *error-proof\_indexes v1.0*. 2019.
78. Galan, M., et al., *Metabarcoding for the parallel identification of several hundred predators and their prey: Application to bat species diet analysis*. *Mol Ecol Resour*, 2018. **18**(3): p. 474-489.
79. Bolyen, E., et al., *QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science*. 2018.
80. Callahan, B.J., et al., *DADA2: High resolution sample inference from Illumina amplicon data*. *Nature methods*, 2016. **13**(7): p. 581-583.
81. Vilgalys, R., *Taxonomic misidentification in public DNA databases*. *New Phytologist*, 2003. **160**(1): p. 4-5.
82. Creedy, T.J., W.S. Ng, and A.P. Vogler, *Toward accurate species-level metabarcoding of arthropod communities from the tropical forest canopy*. *Ecology and Evolution*, 2019. **9**(6): p. 3105-3116.
83. Bladmineerders Online, d., *Plant Parasites of Europe – leafminers, galls and fungi*. 2020.

84. D'Alelio, D., et al., *Ecological-network models link diversity, structure and function in the plankton food-web*. Scientific Reports, 2016. **6**(1): p. 21806.
85. Kirwan, L., et al., *Diversity–interaction modeling: estimating contributions of species identities and interactions to ecosystem function*. Ecology, 2009. **90**(8): p. 2032-2038.
86. Reiss, J., et al., *Emerging horizons in biodiversity and ecosystem functioning research*. Trends in Ecology & Evolution, 2009. **24**(9): p. 505-514.
87. Saint-Béat, B., et al., *Trophic networks: How do theories link ecosystem structure and functioning to stability properties? A review*. Ecological Indicators, 2015. **52**: p. 458-471.
88. Ulanowicz, R.E., R.D. Holt, and M. Barfield, *Limits on ecosystem trophic complexity: insights from ecological network analysis*. 2014. **17**(2): p. 127-136.



ELSEVIER



# Characterizing ecological interaction networks to support risk assessment in classical biological control of weeds

Melodie Ollivier<sup>1</sup>, Vincent Lesieur<sup>1,2</sup>, Sathyamurthy Raghu<sup>3</sup> and Jean-François Martin<sup>1</sup>

A key element in weed biological control is the selection of a biological control agent that minimizes the risks of non-target attack and indirect effects on the recipient community. Network ecology is a promising approach that could help decipher tritrophic interactions in both the native and the invaded ranges, to complement quarantine-based host-specificity tests and gain insights on potential interactions of biological control agents. This review highlights practical questions addressed by networks, including 1) biological control agent selection, based on specialization indices, 2) risk assessment of biological control agent release into a novel environment, *via* particular patterns of association such as apparent competition between agent(s) and native herbivore(s), 3) network comparisons through structural metrics, 4) potential of network modelling and 5) limits of network construction methods.

## Addresses

<sup>1</sup> CBGP, Montpellier SupAgro, INRAE, CIRAD, IRD, Univ Montpellier, Montpellier, France

<sup>2</sup> CSIRO Health and Biosecurity, European Laboratory, Montferrier sur Lez, 34980, France

<sup>3</sup> CSIRO Health & Biosecurity, GPO Box 2583, Brisbane, Qld 4001, Australia

Corresponding author: Ollivier, Melodie ([melodie.ollivier@supagro.fr](mailto:melodie.ollivier@supagro.fr))

Current Opinion in Insect Science 2020, 38:40–47

This review comes from a themed issue on **Parasites/parasitoids/biological control**

Edited by **Heinz Müller-Schärer** and **Urs Schaffner**

For a complete overview see the [Issue](#) and the [Editorial](#)

Available online 30th January 2020

<https://doi.org/10.1016/j.cois.2019.12.002>

2214-5745/Crown Copyright © 2020 Published by Elsevier Inc. All rights reserved.

## Introduction

Ecological networks (popularized as ‘food-cycles’ by Elton in 1927 [1]) describe flows of matter and energy within a community. For example, trophic networks (food-webs) help to understand antagonistic interactions, e.g. such as predation, parasitism and herbivory [2]. Deciphering such networks is a promising approach to

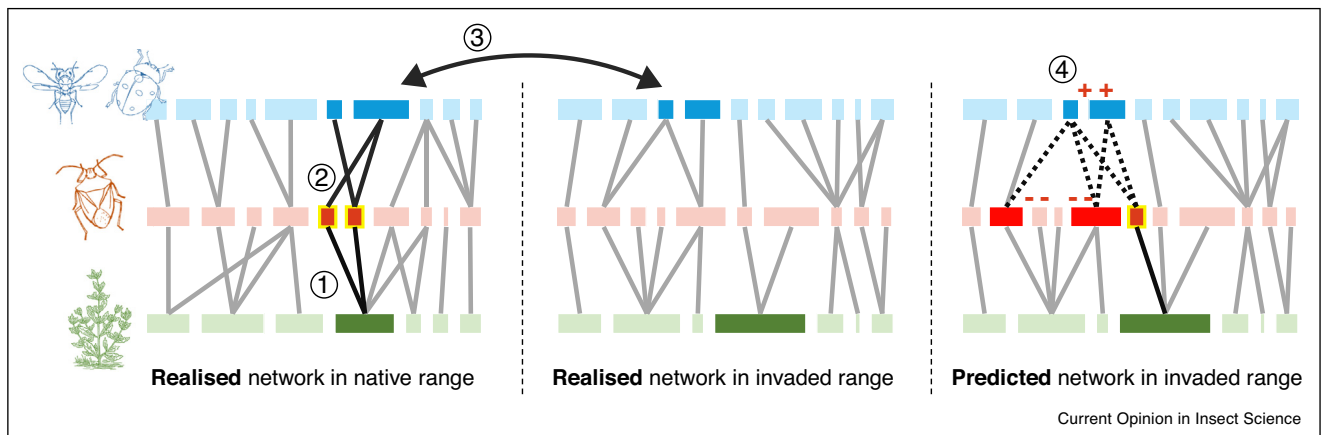
gain insight into niche-based community assembly, reflecting the complexity of species interactions and underlying ecosystem processes [3]. Such analyses can strengthen our understanding of fundamental drivers of community assembly [4,5], co-evolutionary processes [6], ecosystem response to biological invasions and global change [7,8], and ecosystem services management [9,10].

Network ecology could therefore benefit weed biological control, a discipline that aims to re-associate a plant species invading a novel environment with its specialist natural enemies (i.e. biological control agents). Although understanding species interactions has been advocated for more than 20 years [11–13], assessing risks still mostly rely on experimental tests. Network ecology could enhance such research programs through addressing practical questions inherent to weed biological control (Figure 1). This article reviews the approaches and methods that have been used to answer these questions and highlights the potential of ecological network analysis in the context of weed biological control. This review also provides a brief overview of the benefits and pitfalls of main network construction methods.

## Defining the community of herbivores and their host range to improve prediction of non-target attacks

Classical biological control of weeds uses specialist natural enemies of the target plants to selectively reduce their population dynamics under an acceptable economic threshold. A vital first step in this process is the compilation of inventories of natural enemies associated with the target weed in its native range. The specificity of a candidate biological control agent (BCA) is subsequently explored to reduce adverse effects on non-target plants [14,15]. Such tests are generally designed according to the centrifugal phylogenetic method [16] and performed in standardized environments under choice and no-choice conditions. This conservative approach can lead to false positive interactions as the realized field host range is potentially more restricted than the fundamental host range [17]. Risk evaluation solely under experimental conditions has always been known to be simplistic and increasing emphasis is being placed on field host-range assessments in the native range [15]. This implies characterizing interactions in diversified plant communities and being able to describe the realized field host range of

Figure 1



Illustrative tripartite networks showing interactions among communities of plants (green), herbivores (orange), and natural enemies (blue) (composed of predators and parasitoids). The first occur in native range, the second in invaded range. The third is a putative predictive network in the invaded range. As in conventional representation of tripartite networks, each species is represented by a rectangle, whose width reflects its relative abundance in the community. Analysis of networks is intended to enhance the selection of a biological control agent (BCA) with minimal risk of non-target attacks and indirect effects on the recipient community. The process is divided into several steps. 1: Look for herbivores specific to the target plant (dark green rectangles) in the field and determine potential BCAs (dark orange rectangles). Determining field associations may provide more realistic information about species interactions, than relying solely on tests under controlled conditions. 2: Identify potential natural enemies of these putative BCAs. Natural enemies could *i*) threaten BCA efficiency and, *ii*) be source of indirect effects on recipient community via indirect interactions. 3: Compare realized ecological networks in native versus invaded ranges based on *i*) structural and architectural properties and, *ii*) taxonomy. Networks associated with target species are expected to differ between native and invaded ranges in terms of species richness, trophic guilds and complexity. Moreover, if taxonomically closely related species of natural enemies are found between native and invaded ranges (dark blue rectangles), an introduced BCA is more likely to be attacked by these new natural enemies. 4: Predict possible species associations following the release and establishment of a BCA. The third network presents possible indirect effects (dotted line) of the introduced BCA *via* shared parasitoids with native herbivores (red rectangles). This indirect interaction (apparent competition) is likely to have adverse effects on native herbivores and could cascade across other trophic levels.

arthropods through the construction of bipartite networks (BPNs).

Two recent studies [18,19\*\*] characterized the diet of insects analyzing their gut content and reconstructed BPNs based on metabarcoding (molecular identification of the diet through high throughput amplicon sequencing). Zhu *et al.* [19\*\*] in particular identified host plant species of 239 Lepidoptera species, sampled in subtropical forest in China. By comparing traditional observations of plant–herbivore interactions and morphological identifications versus molecular analyses of gut-content and DNA identifications, this study revealed 46 plant species exclusively detected by molecular methods as well as an overall higher species resolution of ecological interactions than originally thought with traditional observation. On a community-wide scale, environmental DNA from wild flowers also proved useful to discover cryptic and unknown plant–arthropod interactions of diverse ecological groups, for example, pollinators, gall inducers, and herbivore species [20].

In weed biological control, ecological specialization of herbivores is a key requisite in their selection as a BCA. Realized interaction preferences are reflected in network

patterns [21\*] as ecological and evolutionary constraints tend to shape the modular structure of networks (*Modularity*: groups of species, e.g. modules, strongly associated with a particular set of plant species).

The computation of specialization indices is commonly performed for pollination networks [22,23] and capture different aspect of the architecture both for the whole network and individual species. Specialization can be measured by counting the number of resources per species (*Generality*) or by quantifying the dependence of a species upon a given resource (*Interaction strength*), (but see Ref. [22] for more indices and their correlations). Specialization patterns observed may be real, but also due to low sampling completeness, or intrinsic differences in resource attractiveness or abundance. Null models allow correcting for such possible artefacts [24]. Tools like *econullnet* [25\*] have been developed to look for resource preferences of a consumer by comparing observed and expected link strengths for every resource of a given consumer species. Novotny *et al.* [26] investigated the specificity pattern among feeding guilds of herbivorous arthropods by estimating their *effective specialization*, an index defined as the proportion of herbivore species feeding on a particular host plant and being unique to



this plant (detailed in Ref. [27]). A more recent study [28\*] compared the average herbivore specialization between the interior of a tropical forest and edges, that are supposed more disturbed and dominated by generalists. Here, the *specialization index* ( $d'$ ) translates to how the observed interactions of a species differed from randomly sampled interactions with other partners. The recently developed *distance-based specialization index* (*DSI*) [29], an extension of the *species specificity index* (*SSI*), accounts for phylogenetic similarity and abundance of hosts plant species (see application of both indices in Ref. [30]). This promising index also accounts for differences in abundance and sampling effort of consumers, which enables robust comparisons among herbivore guilds.

The improvement of molecular techniques coupled with contemporary analyses of BPNs offers multiple opportunities for acquiring insights on interactions occurring in natural environments, and may thus help to characterize field host range of biological control candidates, and complement host-specificity tests.

### Looking for predators and parasitoids of herbivores to improve predictions of indirect effects

Adding parasitoids and predators to convert bipartite networks (BPNs) into tripartite networks (TPNs) can also assist weed biological control. Describing such networks in native and invaded ranges could help analyze the influence of the third trophic level on BCA efficiency and detect the likelihood of indirect effects on the community dynamics.

Knowledge about parasitoids of BCAs is usually obtained as part of rearing BCAs identified in native range surveys. Characterizing predators of herbivores is more challenging as direct observation is required. Metabarcoding enables the detection of prey in arthropod gut-content but also early stage parasitoids in their hosts. In insect biological control, the use of advanced molecular technologies for constructing ecological networks has been recently developed [31\*\*,32] and could be directly transferable to weed biological control. The exploitation of newly introduced organisms by parasitoids of the recipient community is a novel association that has been repeatedly found in the context of biological invasions [33,34\*]. Likewise, predation of the BCA by native natural enemies is a pattern that has also been observed [11,35,36]. These discoveries confirm the ability of introduced organisms to modify food web structure. In a more recent study [37], a post-release food web was constructed involving the two BCAs of the weed *Melaleuca quinquenervia*. The results showed that generalists predators impact the population dynamics of the two released BCAs, although not significantly to diminish biological control efficacy. The community-wide effects of BCAs introduced to Hawaii were also explored *via* the

construction of a TPN and non-target effects have been identified on native communities [38]. Although suggestions have been made to use network analyses in assessing post-release safety of BCAs [11,13], they could also have value in pre-release assessments of indirect effects.

To our knowledge, in classical biological control of weeds, the only study attempting to quantify risk of indirect impacts before the release of the BCA has been in Portugal on *Trichilogaster acaciaelongifoliae*, a gall insect on *Acacia longifolia* [39\*\*]. The authors focused on apparent competition between the BCA and native herbivores due to a shared natural enemy, which in the worst case can lead to the extinction of the native species [40]. From a plant-gall insect-parasitoid TPN, they calculated the proportion of shared parasitoids between the BCA and a native galler. They estimated the potential for the galling BCA to affect the community according to two scenarios: 1) the BCA interacts only with similar species as those currently known to be in interaction with in its area of origin; 2) the galler interacts with all species belonging to the same family as the parasitoids currently known to attack it in its area of origin. In doing so, they predicted the potential for the BCA to interact with native parasitoids and resulting in highly significant indirect effects on the native gall insect.

In addition to apparent competition, other indirect effects could be monitored through network analyses (Figure 2). Network motifs capture the meso-scale structure of a particular species assemblage. Tools like *bmotif* [41] can help count motifs, and species occurrence within motifs, of a BPN and could be employed to look for particular motifs involving BCAs. Despite their potential value, there has been limited use of ecological networks analyses in pre-release assessment in weed biological control programs.

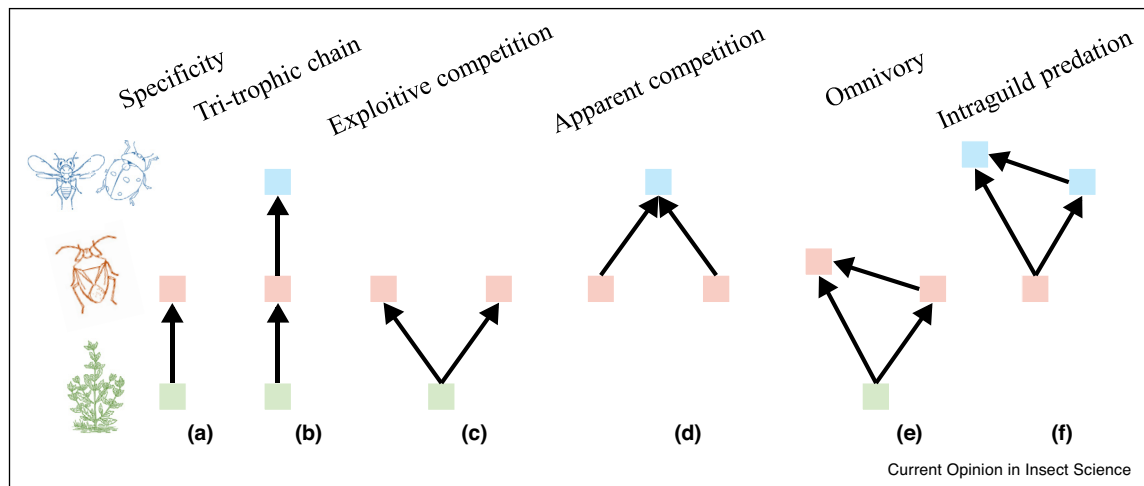
### Comparing ecological networks between native and invaded ranges to describe novel interactions

The introduction of organisms into established communities raises the prospect of novel associations created in the recipient community. Assessing the extent of modification caused by either invasive alien plants (IAPs) or BCAs to recipient communities requires comparison with a reference, that is, communities from the native range, for network structure and species composition. Trophic networks associated with IAP species are hypothesized to 1) be composed of more generalist species and 2) be less diversified (at the herbivore and higher trophic levels) than the native community structure [42,43].

Memmott *et al.* [42] compared, between native and invaded habitats, the arthropod fauna on the IAP Scotch broom, *Cystisus scoparius*, before the BCA release. They



Figure 2



Common motifs studied in ecological networks to explore community assembly. (a) 2-node motif that can be encountered in bipartite and tripartite networks (a) specific relationship between a plant and a herbivore species). Motifs (b)–(d) are 3-node motifs present in bipartite and tripartite structures. (b) Tri-trophic chain (a) plant consumed by a herbivore, which is then preyed upon by its natural enemy). (c) Exploitive competition (a) similar resource shared by two consumers). (d) Apparent competition (a) shared consumer between two resource species). Motifs (e) and (f) cannot be represented in bipartite or tripartite structure, since they represent species interacting within the same community. However, these kinds of interactions occur frequently in natural ecosystems and can be visualized and studied in more complex graphs displaying intermediate trophic levels. (e) Omnivory (a) consumer feeding on diversified food sources, including plants and arthropods, for example, carabid beetles feeding of crop pests and weed seeds). (f) Intraguild predation (predation among a group of natural enemies also sharing a same resource, for example, among natural enemies of aphids, mirids can feed on syrphid eggs).

confirmed that the generalist species were more abundant in the exotic range, while specialist species were dominant in the native range. By analyzing TPNs, authors observed higher herbivore richness in the native range, divided into seven feeding guilds, whereas some guilds (seed and flower feeders) were absent in the invaded range. The increased biomass and abundance of herbivores in the native range coincided with higher natural enemy biomass and abundance. Similar observations have been made by comparing the structure of parasitoid complexes associated with herbivores in their native and invaded range [43] that also pointed out a correlation between the abundance of parasitoids attacking a host in its native versus invaded ranges. In a rare study that investigated realized interactions through network comparison after release of a BCA [44], food webs constructed from the two galling BCAs of the IAP *A. longifolia*, revealed similar taxonomic patterns at the family and super family levels and guild compositions. This study indicates the predictive power of food webs.

When comparing taxa compositions, the *Bray–Curtis similarity index* is the most commonly used. It allows assessing the difference in species composition between two samples considering abundance data [28,45,46]. Structural comparisons of trophic networks rely on the use of networks descriptors to extract information on species properties (e.g. *Ratio of prey to consumers*, *Proportion of species per*

*trophic level*), link properties (e.g. *Link density*, *Connectance*), and consumer-prey asymmetries (e.g. *Generality*, *Vulnerability*) [47]. These metrics, that can be elucidated using various analytical packages (e.g. *bipartite* [48], *cheddar* [49], *foodweb* [50], and *enaR* [51]), can enable a richer understanding of potential ecological interactions of candidate BCAs in the native versus invaded ranges.

### Predicting interactions to assess risks also means modelling

Predictive models of food webs are an additional important and helpful tool in biological control of weeds [52]. By combining the description of a static food web structure with dynamic population models, dynamic food webs could further our understanding and ability to predict changes due to species introductions [52,53]. In a recent study [54], a network model was proposed, based on phosphorus flows, to assess the direct and indirect effects of different biological control methods on the dynamics of algal blooms. Key nodes were identified in the network as particularly efficient to control algal blooms, and strong indirect influences were observed between functional groups. This methodology could be adapted to classical biological control. Sophisticated development of models in closely related research fields of invasion biology [55] and ecosystem management [56,57] would also be transferable to weed biological control.

## Selecting the best methods for reconstructing reliable ecological networks

The relevance of ecological networks for weed biological control depends on the reliability of the data and the methods used to build them. According to the method, ecological networks summarize different kind of species interactions [45\*\*]. Field collection provides networks representing realized interactions, but may be subject to false negative inference due to insufficient sampling effort [21\*,58]. Networks based on literature or database surveys [59,60] for supplementing field observations lead to likely interactions. In addition, models and machine-learning algorithms may be used on data such as presence-absence [61], body size [62,63] or species traits [64,65], to generate predicted interaction networks.

Constructing reliable ecological networks requires knowledge about the benefits and limits inherent to each method in order to choose the methodology suited to the studied system and the research questions addressed. Revealing realized trophic links traditionally rely on labor-intensive techniques based on direct field observations, rearing, or microscopic dissections of gut content and faeces [66]. While providing meaningful behavioral information, these approaches present major limitations when working on below-ground or nocturnal species and prevent the dietary study of sap feeders insects [67]. Approaches relying on plant alkane fingerprints, protein electrophoresis of gut content, stable isotope analysis, detection of prey proteins based on polyclonal and monoclonal antibodies (ELISA) and DNA-based methods can help overcome barriers of visual identification [12,68]. However, the performances of these techniques are context-dependent [67]. Plant alkane fingerprints and protein electrophoresis are not suited to reflect diet breadth of generalist species (providing uninterpretable overlapping banding patterns) [69]. Isotopic enrichment studies have the advantage of providing information over longer temporal scales, integrating past energy flows rather than just the most recent meal. However, isotopic signatures are subject to variations among species that can lead to inconsistent and unclear trophic links [70]. After DNA-based techniques, the monoclonal antibody approach is the second-most used method for the evaluation of food webs in agriculture [71]. Preys antigens offer the benefit of being detectable for a longer period following their consumption [72], compared with rapid degradation of prey DNA in consumer gut content. Although antigens are good markers for screening the consumption of a specific prey by a range of predators, they are not suitable for complex food web analyses as their development would be expensive and time-intensive [73].

DNA-based methods are increasingly used in contemporary food web elucidation in agriculture [66,71,74]. Most commonly, DNA metabarcoding [75] associated with Next Generation Sequencing (NGS) technologies [76]

offers us the possibility to efficiently process large number of samples in the context of biological control. For example, from a field-collected sample of arthropod gut contents, food-range can be tracked and difficult-to-observe interactions, such as host-parasitoid interactions, can be revealed, regardless of insect life stage [77,78]. However, these methods are also prone to potential sources of errors. Sampling device and storage can infer false positive interactions (through external contaminations, secondary predation or scavenging) [31\*\*,79,80]. Insects would be best collected individually, using an aspirator or by hand directly with sterile forceps [80]. This time-intensive method can be adapted by limiting collection time to standard periods at each collection site, normalizing the sampling effort for between site comparisons. Besides sampling incompleteness [21\*], DNA stability and detectability are also a source of false negative interactions. Sensitivity tests may be used to assess how long after ingesting a prey or plant DNA can be detected in consumer gut [80]. Multiple primer set combinations are also recommended to amplify DNA with a large taxonomic coverage [81]. While the mitochondrial gene COI is generally recognized for its properties in arthropod species delineation [82], multiple plant markers are needed for determining plant species [19\*\*]. Lastly, the accurate identification of DNA fragments will fully depend on the quality and completeness of the reference database queried [83,84].

Since sampling incompleteness and the general ability to accurately reveal species interactions may introduce bias to a majority of network descriptors [21\*], the analyses and comparisons of resulting networks require practitioners to be fully aware of the pitfalls and potential that a chosen method offers.

## Conclusion

Characterizing and analyzing ecological interaction networks structure in both, native and invaded ranges generates insights on the processes underpinning effective biological control. It also enables projections of the direct and indirect effects that a biological control agent would have and assist choosing a species that would: 1) be specific to the plant based on natural interactions recorded, 2) possess few natural enemies or natural enemies that would belong to different taxonomic groups as those encountered in the range of introduction. Network analyses, supplemented by advanced molecular methods, could enhance the development of safe biological control strategies and also improve the confidence in biological control among regulators and the general public

## Conflict of interest statement

Nothing declared.

## Acknowledgements

This project was supported by funding from the Australian Government Department of Agriculture, as part of its Rural R&D for Profit programme, through AgriFutures Australia (Rural Industries Research and Development Corporation) (PRJ—010527).

## References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- of special interest
- of outstanding interest

1. Elton CS: *Animal Ecology*. New York: Macmillan Co.; 1927.
2. Memmott J, Martinez ND, Cohen JE: **Predators, parasitoids and pathogens: species richness, trophic generality and body sizes in a natural food web**. *J Anim Ecol* 2000, **69**:1-15.
3. Borrett SR, Moody J, Edelman A: **The rise of network ecology: maps of the topic diversity and scientific collaboration**. *Ecol Modell* 2014, **293**:111-127.
4. Pimm SL, Lawton JH, Cohen JE: **Food web patterns and their consequences**. *Nature* 1991, **350**:669-674.
5. Barraclough TG: **How do species interactions affect evolutionary dynamics across whole communities?** *Annu Rev Ecol Evol Syst* 2015, **46**:25-48.
6. Bascompte J, Jordano P: **Plant-animal mutualistic networks: the architecture of biodiversity**. *Annu Rev Ecol Evol Syst* 2007, **38**:567-593.
7. Traveset A, Richardson DM: **Mutualistic interactions and biological invasions**. *Annu Rev Ecol Evol Syst* 2014, **45**:89-113.
8. Seibold S, Cadotte MW, Maclvor JS, Thorn S, Müller J: **The necessity of multitrophic approaches in community ecology**. *Trends Ecol Evol* 2018, **33**:754-764.
9. Dee LE, Allesina S, Bonn A, Eklöf A, Gaines SD, Hines J, Jacob U, McDonald-Madden E, Possingham H, Schröter M *et al.*: **Operationalizing network theory for ecosystem service assessments**. *Trends Ecol Evol* 2017, **32**:118-130.
10. McDonald-Madden E, Sabbadin R, Game ET, Baxter PWJ, Chadès I, Possingham HP: **Using food-web theory to conserve ecosystems**. *Nat Commun* 2016, **7**:10245.
11. Memmott J: **Food webs as a tool for studying nontarget effects in biological control**. *Nontarget Effects of Biological Control*. US: Springer; 2000, 147-163.
12. Sheppard SK, Harwood JD: **Advances in molecular ecology: tracking trophic links through predator-prey food-webs**. *Funct Ecol* 2005, **19**:751-762.
13. Willis AJ, Memmott J: **The potential for indirect effects between a weed, one of its biocontrol agents and native herbivores: a food web approach**. *Biol Control* 2005, **35**:299-306.
14. Suckling DM, Sforza RFH: **What magnitude are observed nontarget impacts from weed biocontrol?** *PLoS One* 2014, **9**:e84847.
15. Hinz HL, Winston RL, Schwarzländer M: **How safe is weed biological control? A global review of direct nontarget attack**. *Q Rev Biol* 2019, **94**:1-27.
16. Corcket E, Giffard B, Sforza RFH: **Food webs and multiple biotic interactions in plant-herbivore models**. *Adv Bot Res* 2017, **81**:111-137.
17. Fowler SV, Paynter Q, Dodd S, Groenteman R: **How can ecologists help practitioners minimize non-target effects in weed biocontrol?** *J Appl Ecol* 2012, **49**:307-310.
18. Frei B, Guenay Y, Bohan DA, Traugott M, Wallinger C: **Molecular analysis indicates high levels of carabid weed seed consumption in cereal fields across Central Europe**. *J Pest Sci* 2019, **92**:935-942 <http://dx.doi.org/10.1007/s10340-019-01109-5>.
19. Zhu C, Gravel D, He F: **Seeing is believing? Comparing plant-herbivore networks constructed by field co-occurrence and DNA barcoding methods for gaining insights into network structures**. *Ecol Evol* 2019, **9**:1764-1776 <http://dx.doi.org/10.1002/ece3.4860>.  
The study compares the efficiency of two methods to reconstruct ecological network: traditional field observations and DNA-based method. The DNA-based method detected plants species that could not be identified with observational method. Molecular methods provided a higher resolved ecological network.
20. Thomsen PF, Sigsgaard EE: **Environmental DNA metabarcoding of wild flowers reveals diverse communities of terrestrial arthropods**. *Ecol Evol* 2019, **9**:1665-1679.
21. Dormann CF, Fründ J, Schaefer HM: **Identifying causes of patterns in ecological networks: opportunities and limitations**. *Annu Rev Ecol Evol Syst* 2017, **48**:559-584.  
This review highlights how sampling design can affect resulting interaction matrices, especially species abundances, specialization patterns and network descriptors. The review also deals with the utility of null models to correct such effects and suggests that additional information based on species traits would lead to satisfactory resolved networks to address co-evolutionary processes.
22. Dormann CF: **How to be a specialist? Quantifying specialisation in pollination networks**. *Netw Biol* 2011, **1**:1-20.
23. Memmott J: **The structure of a plant-pollinator food web**. *Ecol Lett* 1999, **2**:276-280.
24. Blüthgen N, Fründ J, Vázquez DP, Menzel F: **What do interaction network metrics tell us about specialization and biological traits**. *Ecology* 2008, **89**:3387-3399.
25. Vaughan IP, Gotelli NJ, Memmott J, Pearson CE, Woodward G, Symondson WOC: **Econullnet: an R package using null models to analyse the structure of ecological networks and identify resource selection**. *Methods Ecol Evol* 2018, **9**:728-733.  
Recent R package that have been developed to assess whether observed interactions between species are randomly distributed or reflect particular species association having biological significations, such as preference pattern for a particular resource. It relies on null models computation to test the significance of resource-consumers interactions, considering resources availability. It would be particularly rewarding to look for specificity patterns between plants and potential biocontrol agents in classical biological control.
26. Novotny V, Miller SE, Baje L, Balagawi S, Basset Y, Cizek L, Craft KJ, Dem F, Drew RAI, Hulcr J *et al.*: **Guild-specific patterns of species richness and host specialization in plant-herbivore food webs from a tropical forest**. *J Anim Ecol* 2010, **79**:1193-1203.
27. May RM: **How many species?** *Philos Trans R Soc London Ser B Biol Sci* 1990, **330**:293-304.
28. Ximenes Pinho B, Dáttilo W, Leal IR: **Structural breakdown of specialized plant-herbivore interaction networks in tropical forest edges**. *Glob Ecol Conserv* 2017, **12**:1-8.  
Study that compares the level of specialization of plant-herbivore interactions. It employs a set of tools to test for particular association patterns such as a standardized specialization index, null model randomizations and a modularity index. These approaches would be valuable to transfer in the context of biocontrol agent selection.
29. Jorge LR, Novotny V, Segar ST, Weiblen GD, Miller SE, Basset Y, Lewinsohn TM: **Phylogenetic trophic specialization: a robust comparison of herbivorous guilds**. *Oecologia* 2017, **185**:551-559.
30. Redmond CM, Auga J, Gewa B, Segar ST, Miller SE, Molem K, Weiblen GD, Butterill PT, Maiyah G, Hood ASC *et al.*: **High specialization and limited structural change in plant-herbivore networks along a successional chronosequence in tropical montane forest**. *Ecography (Cop)* 2019, **42**:162-172.
31. Derocles SAP, Kitson JJJ, Massol F, Pauvert C, Plantegenest M, Vacher C, Evans DM: **Biomonitoring for the 21st century: integrating next-generation sequencing into ecological network analysis**. *Adv Ecol Res* 2018, **58**:1-62.  
The understanding of ecosystem functioning being crucial to enhance ecosystem services, this paper reviews the potential of Next-Generation Sequencing to construct highly resolved multilayers ecological networks.

It also lists the shortcomings inherent to this cutting-edge technology and discusses possible solutions.

32. Mollot G, Duyck P-F, Lefeuvre P, Lescouret F, Martin J-F, Piry S, Canard E, Tixier P: **Cover cropping alters the diet of arthropods in a banana plantation: a metabarcoding approach.** *PLoS One* 2014, **9**:e93740.
33. Herron-Sweet CR, Littlefield JL, Lehnhoff EA, Burkle LA, Mangold JM: **Native parasitoids associated with the biological control agents of *Centaurea stoebe* in Montana, USA.** *Biol Control* 2015, **86**:20-27.
34. Murillo Pacheco H, Vanlaerhoven SL, Marcos García MÁ,
  - Hunt DW: **Food web associations and effect of trophic resources and environmental factors on parasitoids expanding their host range into non-native hosts.** *Entomol Exp Appl* 2018, **166**:277-288.

This study highlights the ability of introduced species to create novel interactions with the recipient community. Results show the native parasitoids extended their host range to attack the invasive alien butterfly, changing the structure of the food web.
35. Pearson DE, Callaway RM: **Indirect effects of host-specific biological control agents.** *Trends Ecol Evol* 2003, **18**:456-461.
36. Pearson DE, Callaway RM: **Indirect nontarget effects of host-specific biological control agents: implications for biological control.** *Biol Control* 2005, **35**:288-298.
37. Tipping PW, Martin MR, Nimmo KR, Smart MD, Wear EW: **Food web associations among generalist predators and biological control agents of *Melaleuca quinquevnia*.** *Biol Control* 2016, **101**:52-58.
38. Henneman ML, Memmott J: **Infiltration of a Hawaiian community by introduced biological control agents.** *Science* (80-) 2001, **293**:1314-1316.
39. López-Núñez FA, Heleno RH, Ribeiro S, Marchante H,
  - Marchante E: **Four-trophic level food webs reveal the cascading impacts of an invasive plant targeted for biocontrol.** *Ecology* 2017, **98**:782-793.

This pioneering study uses ecological network theory to assess the risks associated with classical biological control. This study attempts to quantify the risks of indirect impacts on the recipient community, through apparent competition between a biocontrol agent and native species that share the same parasitoids. When considering comparable taxonomical specimens at family level, results predict that indirect effects of the biocontrol agent on native galler will be highly significant.
40. Carvalheiro LG, Buckley YM, Ventim R, Fowler SV, Memmott J: **Apparent competition can compromise the safety of highly specific biocontrol agents.** *Ecol Lett* 2008, **11**:690-700.
41. Simmons BI, Sweering MJM, Schillinger M, Dicks LV, Sutherland WJ, Di Clemente R: **bmotif: a package for motif analyses of bipartite networks.** *Methods Ecol Evol* 2019, **10**:695-701 <http://dx.doi.org/10.1111/2041-210X.13149>.
42. Memmott J, Fowler SV, Paynter Q, Sheppard AW, Syrett P: **The invertebrate fauna on broom, *Cytisus scoparius*, in two native and two exotic habitats.** *Acta Oecologica* 2000, **21**:213-222.
43. Cornell HV, Hawkins BA: **Accumulation of native parasitoid species on introduced herbivores: a comparison of hosts as natives and hosts as invaders.** *Am Nat* 1993, **141**:847-865.
44. Veldtman R, Lado TF, Botes A, Procheş Ş, Timm AE, Geertsema H, Chown SL: **Creating novel food webs on introduced Australian acacias: indirect effects of galling biological control agents.** *Divers Distrib* 2011, **17**:958-967.
45. Delmas E, Besson M, Brice MH, Burkle LA, Dalla Riva GV,
  - Fortin MJ, Gravel D, Guimarães PR, Hemby DH, Newman EA et al.: **Analysing ecological networks of species interactions.** *Biol Rev* 2019, **94**:16-36 <http://dx.doi.org/10.1111/brv.12433>.

This paper reviews the tools available in network analysis to address ecological questions regarding species interactions. It highlights their methodological development, the appropriate metrics to analyses ecological networks, and the potential and limitations of these approaches. Furthermore, strategies to test ecological hypotheses through the comparisons of community structure are presented.
46. Eitzinger B, Abrego N, Gravel D, Huotari T, Vesterinen EJ, Roslin T: **Assessing changes in arthropod predator-prey interactions through DNA-based gut content analysis—variable environment, stable diet.** *Mol Ecol* 2019, **28**:266-280 <http://dx.doi.org/10.1111/mec.14872>.
47. Bersier L-F, Banašek-Richter C, Cattin M-F: **Quantitative descriptors of food-web matrices.** *Ecology* 2002, **83**:2394-2407.
48. Dormann CF, Frund J, Bluthgen N, Gruber B: **Indices, graphs and null models: analyzing bipartite ecological networks.** *Open Ecol J* 2009, **2**:7-24.
49. Hudson LN, Emerson R, Jenkins GB, Layer K, Ledger ME, Pichler DE, Thompson MSA, O'Gorman EJ, Woodward G, Reuman DC: **Cheddar: analysis and visualisation of ecological communities in R.** *Methods Ecol Evol* 2013, **4**:99-104.
50. Perdomo G, Sunnucks P, Thompson RM: **foodweb-package: Visualisation and analysis of food web networks in foodweb: visualisation and analysis of food web networks.** [date unknown].
51. Borrett SR, Lau MK: **enaR: an R package for ecosystem network analysis.** *Methods Ecol Evol* 2014, **5**:1206-1213.
52. Barratt BIP, Howarth FG, Withers TM, Kean JM, Ridley GS: **Progress in risk assessment for classical biological control.** *Biol Control* 2010, **52**:245-254.
53. Tyliranakis JM, Binzer A: **Effects of global environmental changes on parasitoid-host food webs and biological control.** *Biol Control* 2014, **75**:77-86.
54. Mao X, Wei X, Yuan D, Jin Y, Jin X: **An ecological-network-analysis based perspective on the biological control of algal blooms in Ulansuhai Lake, China.** *Ecol Modell* 2018, **386**:11-19.
55. Romanuk TN, Zhou Y, Brose U, Berlow EL, Williams RJ, Martinez ND: **Predicting invasion success in complex ecological networks.** *Philos Trans R Soc Lond B Biol Sci* 2009, **364**:1743-1754.
56. Pires MM: **Rewilding ecological communities and rewiring ecological networks.** *Perspect Ecol Conserv* 2017, **15**:257-265.
57. Baker CM, Bode M, Dexter N, Lindenmayer DB, Foster C, MacGregor C, Plein M, McDonald-Madden E: **A novel approach to assessing the ecosystem-wide impacts of reintroductions.** *Ecol Appl* 2019, **29**:e01811.
58. Jordano P: **Sampling networks of ecological interactions.** *Funct Ecol* 2016, **30**:1883-1893.
59. Poisot T, Gravel D, Leroux S, Wood SA, Fortin M-J, Baiser B, Cirtwill AR, Araújo MB, Stouffer DB: **Synthetic datasets and community tools for the rapid testing of ecological hypotheses.** *Ecography (Cop)* 2016, **39**:402-408.
60. Beas-Luna R, Novak M, Carr MH, Tinker MT, Black A, Caselle JE, Hoban M, Malone D, Iles A: **An online database for informing ecological network models.** *PLoS One* 2014, **9**:e109356 <http://kelpforest.ucsc.edu>.
61. Sander EL, Wootton JT, Allesina S: **Ecological network inference from long-term presence-absence data.** *Sci Rep* 2017, **7**:7154.
62. Bohan DA, Caron-Lormier G, Muggleton S, Raybould A, Tamaddoni-Nezhad A: **Automated discovery of food webs from ecological data using logic-based machine learning.** *PLoS One* 2011, **6**:e29028.
63. Gravel D, Poisot T, Albouy C, Velez L, Mouillot D: **Inferring food web structure from predator-prey body size relationships.** *Methods Ecol Evol* 2013, **4**:1083-1090.
64. Crea C, Ali RA, Rader R: **A new model for ecological networks using species-level traits.** *Methods Ecol Evol* 2016, **7**:232-241.
65. Bartomeus I, Gravel D, Tyliranakis JM, Aizen MA, Dickie IA, Bernard-Verdier M: **A common framework for identifying linkage rules across different types of interactions.** *Funct Ecol* 2016, **30**:1894-1903.
66. Evans DM, Kitson JJJ, Lunt DH, Straw NA, Pocock MJO: **Merging DNA metabarcoding and ecological network analysis to understand and build resilient terrestrial ecosystems.** *Funct Ecol* 2016, **30**:1904-1916.



67. Pompanon F, Deagle BE, Symondson WOC, Brown DS, Jarman SN, Taberlet P: **Who is eating what: diet assessment using next generation sequencing.** *Mol Ecol* 2012, **21**:1931-1950.
68. De Barba M, Miquel C, Boyer F, Mercier C, Rioux D, Coissac E, Taberlet P: **DNA metabarcoding multiplexing and validation of data accuracy for diet assessment: application to omnivorous diet.** *Mol Ecol Resour* 2014, **14**:306-323.
69. Symondson WOC: **Molecular identification of prey in predator diets.** *Mol Ecol* 2002, **11**:627-641.
70. Traugott M, Pázmándi C, Kaufmann R, Juen A: **Evaluating  $^{15}\text{N}/^{14}\text{N}$  and  $^{13}\text{C}/^{12}\text{C}$  isotope ratio analysis to investigate trophic relationships of elaterid larvae (Coleoptera: Elateridae).** *Soil Biol Biochem* 2007, **39**:1023-1030.
71. González-Chang M, Lefort M-C: **Food webs and biological control: a review of molecular tools used to reveal trophic interactions in agricultural systems.** *Food Webs* 2016, **9**:4-11.
72. Schenk D, Bacher S: **Detection of shield beetle remains in predators using a monoclonal antibody.** *J Appl Entomol* 2004, **128**:273-278.
73. Chen Y, Giles KL, Payton ME, Greenstone MH: **Identifying key cereal aphid predators by molecular gut analysis.** *Mol Ecol* 2000, **9**:1887-1898.
74. Roslin T, Majaneva S: **The use of DNA barcodes in food web construction—terrestrial and aquatic ecologists unite!** *Genome* 2016, **59**:603-628.
75. Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E: **Towards next-generation biodiversity assessment using DNA metabarcoding.** *Mol Ecol* 2012, **21**:2045-2050.
76. Shendure J, Ji H: **Next-generation DNA sequencing.** *Nat Biotechnol* 2008, **26**:1135-1145.
77. Garipey TD, Haye T, Zhang J: **A molecular diagnostic tool for the preliminary assessment of host-parasitoid associations in biological control programmes for a new invasive pest.** *Mol Ecol* 2014, **23**:3912-3924.
78. Hrčák J, Godfray HCJ: **What do molecular methods bring to host-parasitoid food webs?** *Trends Parasitol* 2015, **31**:30-35.
79. King RA, Davey JS, Bell JR, Read DS, Bohan DA, Symondson WOC: **Suction sampling as a significant source of error in molecular analysis of predator diets.** *Bull Entomol Res* 2012, **102**:261-266.
80. King RA, Read DS, Traugott M, Symondson WOC: **Molecular analysis of predation: a review of best practice for DNA-based approaches.** *Mol Ecol* 2008, **17**:947-963.
81. Alberdi A, Aizpurua O, Gilbert MTP, Bohmann K: **Scrutinizing key steps for reliable metabarcoding of environmental samples.** *Methods Ecol Evol* 2018, **9**:134-147.
82. Hebert PDN, Cywinska A, Ball SL, deWaard JR: **Biological identifications through DNA barcodes.** *Proc R Soc London Ser B Biol Sci* 2003, **270**:313-321.
83. Vilgalys R: **Taxonomic misidentification in public DNA databases.** *New Phytol* 2003, **160**:4-5.
84. Creedy TJ, Ng WS, Vogler AP: **Toward accurate species-level metabarcoding of arthropod communities from the tropical forest canopy.** *Ecol Evol* 2019, **9**:3105-3116.



# Cover Cropping Alters the Diet of Arthropods in a Banana Plantation: A Metabarcoding Approach

Gregory Mollot<sup>1,2</sup>, Pierre-François Duyck<sup>1,3\*</sup>, Pierre Lefeuvre<sup>3</sup>, Françoise Lescouret<sup>2</sup>, Jean-François Martin<sup>4</sup>, Sylvain Piry<sup>5</sup>, Elsa Canard<sup>1,6</sup>, Philippe Tixier<sup>1,7</sup>

**1** CIRAD, UR 26 Systèmes de culture à base de bananiers, plantains et ananas, PRAM, Le Lamentin, Martinique, France, **2** INRA, UR-1115 Plantes et Systèmes de culture Horticoles, Avignon, France, **3** CIRAD, UMR PVBMT, CIRAD/Université de La Réunion, Pôle de Protection des Plantes, Saint-Pierre, La Réunion, France, **4** Montpellier-SupAgro, UMR CBGP, Montferrier-sur-Lez, France, **5** INRA, UMR1062 CBGP, Montferrier-sur-Lez, France, **6** CNRS-IRD, UMR 2724 MIVEGEC, Montpellier, France, **7** CIRAD – CATIE, Departamento de Agricultura y Agroforestería, CATIE, Turrialba, Costa Rica

## Abstract

Plant diversification using cover crops may promote natural regulation of agricultural pests by supporting alternative prey that enable the increase of arthropod predator densities. However, the changes in the specific composition of predator diet induced by cover cropping are poorly understood. Here, we hypothesized that the cover crop can significantly alter the diet of predators in agroecosystems. The cover crop *Brachiaria decumbens* is increasingly used in banana plantations to control weeds and improve physical soil properties. In this paper, we used a DNA metabarcoding approach for the molecular analysis of the gut contents of predators (based on mini-COI) to identify 1) the DNA sequences of their prey, 2) the predators of *Cosmopolites sordidus* (a major pest of banana crops), and 3) the difference in the specific composition of predator diets between a bare soil plot (BSP) and a cover cropped plot (CCP) in a banana plantation. The earwig *Euborellia carabea*, the carpenter ant *Camponotus sexguttatus*, and the fire ant *Solenopsis geminata* were found to contain *C. sordidus* DNA at frequencies ranging from 1 to 7%. While the frequencies of predators positive for *C. sordidus* DNA did not significantly differ between BSP and CCP, the frequency at which *E. carabea* was positive for Diptera was 26% in BSP and 80% in CCP; the frequency at which *C. sexguttatus* was positive for *Jalysus spinosus* was 14% in BSP and 0% in CCP; and the frequency at which *S. geminata* was positive for *Polytus mellerbergi* was 21% in BSP and 3% in CCP. *E. carabea*, *C. sexguttatus* and *S. geminata* were identified as possible biological agents for the regulation of *C. sordidus*. The detection of the diet changes of these predators when a cover crop is planted indicates the possible negative effects on pest regulation if predators switch to forage on alternative prey.

**Citation:** Mollot G, Duyck P-F, Lefeuvre P, Lescouret F, Martin J-F, et al. (2014) Cover Cropping Alters the Diet of Arthropods in a Banana Plantation: A Metabarcoding Approach. PLoS ONE 9(4): e93740. doi:10.1371/journal.pone.0093740

**Editor:** Dawn Sywassink Luthe, Pennsylvania State University, United States of America

**Received:** December 5, 2013; **Accepted:** March 6, 2014; **Published:** April 2, 2014

**Copyright:** © 2014 Mollot 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:** This research is a part of a PhD funded by the CIRAD and funded by the project “sustainable cropping systems design” from E.U. FEDER (grant PRESAGE n°30411). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

\* E-mail: duyck@cirad.fr

## Introduction

Agriculture faces the challenges of providing more food and energy while adapting to climate change and mitigating environmental impacts. One of the most promising approaches to meet these challenges is the design of new agroecosystems based on the management of ecological processes rather than on application of fertilizers and pesticides [1]. For instance, the regulation of crop pests through top-down and bottom-up effects remains a potential alternative to reduce the ecological imprints of agroecosystems while maintaining production [2]. This regulation could rely on the management of primary resources, such as the addition of cover crops [3]. The underlying hypothesis is that cover crops enable the development of alternative preys leading to higher densities and diversities of generalist arthropod predators. The larger the densities of predators, the higher the consumption of herbivore pests - provided that the pest remains a favourite prey [4,5]. It follows that designing environmentally friendly cropping systems requires a clear understanding of food web functions.

In spite of the substantial research conducted in the last decade, food web ecology suffers from a lack of efficient and comprehen-

sive methods to measure trophic links *in natura* with accuracy. To date, trophic links are often inferred by abundance measurements of predators and prey [6–10], stable isotope analyses [11–17], and protein-based approaches [18,19]. Although recent molecular approaches, like multiplex-PCR, have enabled identification of specific prey in the gut contents and faeces of a wide range of predators [20], these methods are not suited for the detection of unexpected prey [21]. Recently, Next Generation Sequencing (NGS) technology has been used to examine the specific composition of the diet of a given predator or herbivore and to describe two-level food webs [for a review, see 22]. The development of DNA metabarcoding now enables researchers to measure trophic links without *a priori* knowledge of the consumed species and to determine the diet of each individual [22]. The metabarcoding approach is based on the pyrosequencing of a DNA barcode (amplified with universal primers) that can discriminate and identify species in a DNA mixture. This method could be used to analyse gut contents of arthropods, because it has the potential to identify the complex diet of generalist predators.

Plant diversification has strong bottom-up effects on multi-trophic interaction networks, especially on lower trophic levels [23], but it remains unclear whether the addition of a cover crop in agroecosystems actually leads to enhanced pest control by predators [24]. In banana agroecosystems in Martinique, the cover crop *Brachiaria decumbens* Stapf, a tropical C4 grass, is increasingly used to control weeds and to improve physical soil properties. In these agroecosystems, Mollet *et al.* [25] showed that *Solenopsis geminata* (F.), a generalist predator feeding on the banana weevil *Cosmopolites sordidus* (Germar), the major banana insect pest, was more abundant in cover cropped plots (CCP) than in bare soil plots (BSP). Along with this increase in densities, *S. geminata* exhibited a change in isotopic signature, indicating that it fed on the C4 pathway provided by the new resource. In the same study, the monitoring of eggs of *C. sordidus* artificially deposited in plots showed that predation on eggs was always higher in CCP than in BSP. However, the specific changes in the diets of predators affected by the cover crop are unclear. Identifying the prey consumed and determining the rate at which they are consumed by the major generalist predators would help us understand the effects of the cover crop on predator diets and thus on the regulation of *C. sordidus*.

Using a metabarcoding approach based on the COI barcode, we assessed the diet of eight ground-dwelling predators commonly found in banana plantations in Martinique: wolf spiders from the Lycosidae family, the earwig *Euborellia caraibea* Hebard, the carpenter ant *Camponotus sexguttatus* (F.), the trap jaw ant *Odontomachus baurii* Emery, the fire ant *S. geminata*, the little fire ant *Wasmannia auropunctata* (Roger), rove beetles from the Staphilinidae family, and centipedes from the Scolopendridae family. We amplified a shortened fragment of COI (mitochondrial cytochrome *c* oxidase I) from the gut contents of predators to identify (1) DNA sequences of their prey, (2) predators of *C. sordidus*, and (3) the difference in predator diet between CCP and BSP. Based on these results, we make suggestions about how to use and manage a *B. decumbens* cover crop to control the populations of *C. sordidus* in banana plantations. Finally, we discuss the technical implications of the use of the metabarcoding approach to assess the diet of ground-dwelling predators.

## Materials and Methods

### Ethics Statement

All of the authors declare that the experiments performed in the present study comply with the current laws of France. No specific permits were required for the described field study, which involved sampling of invertebrates and plant species. No specific permits were required to perform the described study in this location, which is an experimental farm owned by CIRAD. All of the authors confirm that the location is not privately owned or protected in any way and that the field studies did not involve endangered or protected species.

### Study Sites

Sampling was conducted in Martinique (French West Indies) between January and June 2011. Samples were collected from an experimental farm in Rivière Lézarde (14°39'45.04"N; 60°59'59.08"W) in two adjacent plots: a bare soil plot (BSP) of 300 m<sup>2</sup> and a *B. decumbens* cover cropped plot (CCP) of 368 m<sup>2</sup>. Both plots were in the sixth year of banana production without insecticide application; plants were unsynchronized and harvested throughout the year.

## Sampling

The first step of the procedure was the construction of a reference bank of DNA sequences that included every possible arthropod prey taxa from the studied sites. To construct this bank of sequences, we designed a sampling scheme to capture most of the arthropod diversity in banana agroecosystems. We collected two to four samples (one sample corresponds to one individual of a given taxon) belonging to each of 15 taxa commonly found in banana fields (taxa and trapping methods are listed in **Table S1**). Soil-surface arthropod samples were collected with dry pitfall traps and pseudostem traps (one-half of a section of fresh banana pseudostem, 50 cm long), whereas flying arthropod samples (*Gryllus*, Cicadellidae, and Pentatomidae) were collected by 15-s sessions with a suction sampler (D-vac, Rincon-Vitova Insectaries, Inc., Ventura, California, USA). We also directly collected samples with clean forceps to obtain arthropods that were not trapped by pitfall traps, pseudostem traps, or vacuum sampling.

The samples for the diet analyses were obtained by collecting individual samples from the most common ground-dwelling predators ( $n = 572$ ) (**Table S2**). The recovery of DNA from the gut contents of predators was optimized by placing samples in a portable cooler (4°C) in the field so as to decrease enzymatic activity and prevent DNA degradation. Samples were collected every 12 h from dry pitfall traps and pseudostem traps, which were distributed at 4-m intervals over the plots, and by direct capture. As indicated by King *et al.* [20], predation events occurring in traps remain a substantial concern for diet analyses. To reduce this possible source of error, we focused on direct capture, frequently collected predators in the traps, and excluded samples from traps that contained fragments of herbivores or other predators. Samples were placed in separate tubes in 96% ethanol; the tubes were temporarily kept in a portable cooler until they were transported to the laboratory and stored at  $-20^{\circ}\text{C}$ .

## Feeding Trials and Positive Controls

To determine whether it was possible to detect *C. sordidus* DNA in the gut contents of predators, we collected 10 additional samples of *O. baurii* in the field by direct capture and placed them individually in tubes with moistened cotton and without food for 96 h. At this stage, seven samples were killed and then placed individually in clean tubes containing 96% ethanol and stored at  $-20^{\circ}\text{C}$ ; among them, three samples were analysed alone, and four samples were analysed after the addition of one *C. sordidus* egg before DNA extraction. The remaining three samples, which had been kept alive, were placed in new tubes with moistened cotton and one *C. sordidus* egg (one predator and one egg per tube); after 12 h, each individual of *O. baurii* had fed on the provided egg and was placed in a clean tube in 96% ethanol and stored at  $-20^{\circ}\text{C}$ . In a "positive control" experiment, we also analysed the quantity of sequences recovered as a function of the number of *C. sordidus* eggs in a sample without predators; this experiment used 1 egg ( $n = 9$ ), two eggs ( $n = 3$ ), and three eggs ( $n = 3$ ).

## Construction of the Mini-COI Bank of Sequences by SANGER Sequencing for Taxa Assignment

Predation was studied by amplifying and sequencing the mitochondrial *cytochrome c oxidase I* (COI) gene, which is widely used for species-level identification of animals [26]. Legs of two to four frozen samples of each taxon collected in the banana plantations (taxa and trapping methods are listed in **Table S1**) were used for the construction of the COI bank of sequences. Total DNA was extracted from legs with the DNeasy Blood and Tissue kit (Qiagen, Germany) following the manufacturer's

protocol. The long fragment of COI was amplified with the universal primers LCOI490 and HCO2198 [27] in a 20- $\mu$ l volume containing 0.5 U of HotStarTaq plus DNA polymerase (Qiagen), 3 mM MgCl<sub>2</sub>, 400  $\mu$ M of each dNTP, 10  $\mu$ M of each primer, and 8  $\mu$ l of arthropod DNA extract. After an initial activation of the DNA polymerase for 5 min at 95°C, the amplification was performed with 5 cycles of 1 min at 95°C, 1 min at 45°C, and 1.5 min at 72°C; followed by 30 cycles of 1 min at 95°C, 1 min at 48°C, and 1.5 min at 72°C; and a final extension of 5 min at 72°C. Amplicons were sequenced using the Sanger method on both strands and for each sample with the ABI3730XL analyser (Applied Biosystems) by the MacroGen sequencing service (Seoul, South Korea). Sequences were assembled and aligned with Geneious Pro 5.5.3 (Biomatters, New Zealand) before the mini-COI barcode was extracted. We deposited 15 sequences of COI in GenBank, and these included the sequences of two species previously not recorded in GenBank (see **Table S1**). The final mini-COI bank of sequences included the 15 sequences obtained from samples of the banana plantations and 20 additional COI sequences obtained from GenBank after the BLAST of the raw 454 sequences (sequences recovered with GenBank in **Table S3**).

#### 454 Pyrosequencing of Mini-COI

We used a shortened fragment of COI, the mini-COI fragment (127 bp), which was amplified with primers Uni-MinibarF1 and Uni-MinibarR1 designed by Meusnier *et al.* [28]. Total DNA was extracted from the dissected gut contents or from the whole body (when body size was <1 cm) of the ground-dwelling predators with the DNeasy Blood and Tissue kit following the manufacturer's protocol. To enable deconvolution of pooled 454 sequencing runs such that individual sequences could be traced back to a particular sample, we tagged the 5' end of the PCR primers with different combinations of seven nucleotides (the tags are listed in **Table S4**). A total of 30 different tags enabled us to process the 572 samples of ground-dwelling predators for diet analyses and the 59 samples for positive controls (taxa, trapping methods, and positive controls are listed in **Table S2**). Deconvolution of the pooled sequences was performed with an exact search of the tag sequences. These tags differed in at least three nucleotides, which reduced the risk of incorrect assignment of sequences to sample ID in case of a sequencing error.

Amplification of mini-COI was performed in a 20- $\mu$ l volume containing 0.5 U of HotStarTaq plus DNA polymerase (Qiagen), 3 mM MgCl<sub>2</sub>, 400  $\mu$ M of each dNTP, 10  $\mu$ M of each primer, and 8  $\mu$ l of arthropod DNA extract. After an initial activation of the DNA polymerase for 5 min at 95°C, the mini-COI was amplified with 5 cycles of 60 s at 95°C, 60 s at 46°C, and 30 s at 72°C; 35 cycles of 60 s at 95°C, 60 s at 53°C, and 30 s at 72°C; and a final extension of 5 min at 72°C. Although blocking probes are often used to reduce the sequencing of predator DNA [29], they also make the procedure more cumbersome, especially when numerous taxa are analysed. In the current study, the pyrosequencing generated enough sequences so that prey could be identified even with over-representation of predator DNA sequences (see Results). An analysis of the migration of the PCR products obtained with tagged mini-COI primers enabled us to sort PCR products as a function of their signal intensity (null, low, medium, strong). PCR products from gut contents of ground-dwelling predators were concentrated in an oven at 35°C for 12 h. Then, we standardized the DNA concentration in all samples before pooling them in an equimolar solution. Pooled PCR products were purified with the QIAquick Gel Extraction Kit (Qiagen, Germany) and sequenced using the 454 GS FLX Titanium platform (Roche, Basel, Switzerland) of Beckman Coulter Genomics (Danvers, MA, USA).

#### Bioinformatics Processing of Raw Sequences and Taxonomic Assignment

Pyrosequencing outputs were analysed using |SE|S|AM|E| BARCODE, a software designed to process the large amount of DNA sequences generated by pyrosequencing [for more details on the bioinformatics pipeline used, see 30,31]. The software loads a table in which the sequence of each tag is recorded for each sample and the FASTA file obtained from the pyrosequencing step. Sequences of interest are recovered after the BLAST of each raw sequence to a consensus sequence of the marker (threshold E-value =  $e^{-2}$ , which is low to recover a maximum of sequences). This marker consensus sequence (127 bp) was obtained after the alignment [MUSCLE algorithm, 32] of all sequences of taxa found in the banana plantations (**Table S1**).

Mini-COI barcodes were assigned to taxa from the bank of sequences using the BLAST+ (E-value =  $e^{-20}$ ) and FASTA algorithm (85% similarity threshold) available in |SE|S|AM|E| BARCODE. Final barcoding identification was performed by assigning sequences with a Nearest Neighbour algorithm. Decision rules applied during the processing of sequences lead to an assignment to a higher taxonomic rank when the percentages of similarity were strictly equal between two or more species. Unique sequences were removed so that a minimum of two sequences was used for the barcoding identification. Sequences with fewer than 120 nucleotides (without primers) were discarded.

#### Comparison of the Diet of the Ground-dwelling Predators in BSP vs. CCP

Although the number of sequences found for each prey was determined for each predator sample, this quantitative information could not be processed directly as indicated by Pompanon *et al.* [22]. The number of sequences obtained for a given prey was converted into binary information (presence/absence), and the trophic links were quantified by the number of predator samples among the population that were positive for each prey taxon. It is important to note that sequences assigned to a taxonomic rank higher than order were discarded from the analyses. Finally, we calculated the differences of frequency of consumption for each identified prey taxon between BSP and CCP and determined whether the differences were statistically significant using a Fisher's exact test implemented within the statistical program R [33].

## Results

#### Control Experiments

We first assessed our ability to amplify and detect *C. sordidus* mini-COI from 1, 2, or 3 *C. sordidus* eggs in a sample that lacked predator tissue or predator DNA. We correctly assigned the samples to *C. sordidus* for 14 of the 15 samples (see first three rows in **Table 1**). Nevertheless, no significant correlation was obtained between the number of eggs analysed and the number of sequences. We also performed trials in which each of three *O. baurii* was either fed with one *C. sordidus* egg and then analysed or in which each of four *O. baurii* was mixed with one *C. sordidus* egg and then analysed; although the detection rate was <100%, *C. sordidus* was detected in both cases (see rows 4–6 in **Table 1**). In negative controls (pure water), DNA of *Myospora lauta* Stein was detected (two sequences) (see rows 7 and 8 in **Table 1**) and was consequently excluded from further analyses because we suspected a possible cross-contamination (a total of 36 sequences of *M. lauta* were detected in four samples of *C. sexguttatus*).



**Table 1.** Positive controls, feeding trials, and negative controls.

Sample	n	Barcoding identification	Rank	Frequency (%)	Number of sequences		
					Total	Mean	SE
1 egg of <i>C. sordidus</i>	9	<i>Cosmopolites sordidus</i>	Species	89	14	4	
2 eggs of <i>C. sordidus</i>	3	<i>Cosmopolites sordidus</i>	Species	100	39	19	
3 eggs of <i>C. sordidus</i>	3	<i>Cosmopolites sordidus</i>	Species	100	55	21	
<i>O. baurii</i> kept without food for 96 h	3	<i>Cosmopolites sordidus</i>	Species	0	0	0	
<i>O. baurii</i> fed with 1 egg of <i>C. sordidus</i>	3	<i>Cosmopolites sordidus</i>	Species	33	1	1	
<i>O. baurii</i> kept without food for 96 h + 1 egg of <i>C. sordidus</i>	4	<i>Cosmopolites sordidus</i>	Species	50	15	16	
Water (negative control)	10	Neoptera	Subclass	10	0.2	0.2	
		<i>Myospila lauta</i>	Species	10	0.2	0.2	

Positive controls (rows 1–3 in the table) were designed to estimate the number of sequences recovered as a function of the quantity of *C. sordidus* material assayed, which was varied by processing different numbers of *C. sordidus* eggs. The feeding trial (rows 4–6 in the table) was designed to determine whether *C. sordidus* DNA could be detected in the gut contents of a predator (*O. baurii* in this case) that had consumed one *C. sordidus* egg (in related treatments, the predator had not consumed a *C. sordidus* egg or was processed with one egg that it had not consumed). Negative controls (rows 7 and 8 in the table) consisted of ultra-pure HPLC grade water. doi:10.1371/journal.pone.0093740.t001

## Pyrosequencing Outputs and Resolution of the Mini-COI Barcode

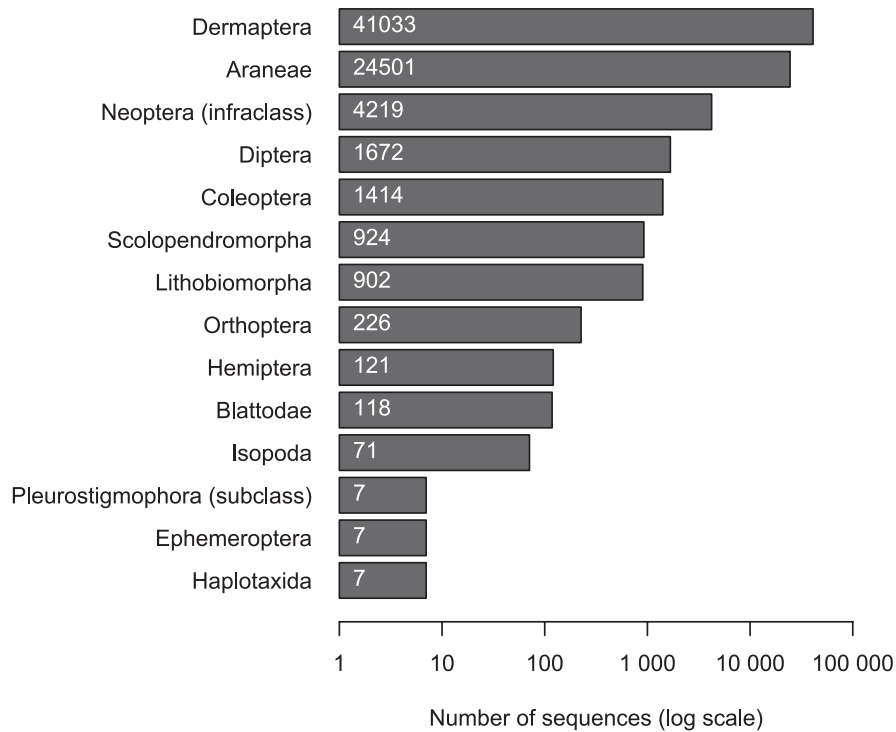
Among the 177,186 raw sequences obtained from the gut contents of predators, 118,180 sequences (67%) were identified as mini-COI marker, of which 75,313 sequences (43%) were successfully assigned to a taxon. A total of 56 distinct taxa belonging to 14 orders were identified (predator diet analyses and positive controls are indicated in **Figure 1**); 71.4% were identified to species, 80.3% to genus, and 91.0% to family. Among these taxa, nine were identified from sequences obtained in this study (**Table S2**) while the 47 remaining taxa were recovered from COI sequences recorded in GenBank (**Table S3**). Sequences of taxa belonging to the Lycosidae and to *E. caraibea* were very abundant (87% of the assigned sequences, **Table 2**). It is highly probable that these taxa were so efficiently amplified during the PCR step that they were later overrepresented in the samples. Consequently, we considered these taxa as a possible source of false positive prey detection, and we removed them from the diet analyses of predators.

## Identification of Prey from the Gut Contents of Ground-dwelling Predators

A total of 29 prey taxa were identified from the gut contents of the eight ground-dwelling predators taxa found in the banana plantations (**Figure 2**). DNA sequences of the banana weevil *C. sordidus* were recovered from samples of three species, with relatively low frequencies of consumption (frequency of consumption refers to the percentage of samples that were positive for DNA of the prey in question): the earwig *E. caraibea* (frequencies: BSP = 3%, CCP = 7%), the carpenter ant *C. sexguttatus* (CCP = 3%), and the fire ant *S. geminata* (BSP = 1%). The BLAST of sequences against the COI sequences recorded in GenBank enabled identification of dipteran arthropods that were not sampled in the plots during the study. We did not identify prey consumed by samples from the Scolopendridae family in either plot. Because of the absence of prey sequences in one of the two plots, the diet change could not be assessed for samples from the Lycosidae family, *O. baurii*, or the Staphilinidae family (frequencies of consumption are listed in **Table S5**).

## Difference in the Diets of Ground-dwelling Predators between BSP and CCP

Twenty-nine prey taxa were identified from the gut contents of ground-dwelling predators. Among these, 22 prey taxa were identified in BSP while 19 were identified in CCP; 12 prey taxa were detected in both plots. Frequencies of prey detected from gut contents significantly differed in the two plots for some of the ground-dwelling predators (**Figure 3**). Interestingly, whereas *Jalysus spinosus* (Say) was detected as the main prey of the carpenter ant *C. sexguttatus* in BSP (14% positive), no sample of *C. sexguttatus* was positive for this prey in CCP (Fisher's exact test,  $p$ -value = 0.0043). The frequency at which the earwig *E. caraibea* was positive for dipteran DNA was 26% in the BSP and 80% in the CCP (Fisher's exact test,  $p$ -value < 0.0001). The frequency at which the fire ant *S. geminata* was positive for DNA of the little banana weevil *Polytus mellerborgi* Heller was 21% in the BSP and 3% in the CCP (Fisher's exact test,  $p$ -value = 0.0006). Samples of *C. sexguttatus* that were positive for banana weevil DNA were found only in CCP, while samples of *S. geminata* that were positive for banana weevil DNA were found only in BSP. *Cosmopolites sordidus* was detected in the gut contents from two samples of *E. caraibea* in BSP ( $n = 53$ ) and from two other samples in CCP ( $n = 30$ ).



**Figure 1. Number of sequences assigned to each order after mini-COI barcode sequencing.** Data were derived from 454 pyrosequencing run, and sequences were assigned to taxa with SE[S|AM|E] BARCODE (minimum of two sequences, BLAST+ E-value =  $e^{-20}$ , FASTA 85% similarity threshold and Nearest Neighbour algorithm for the final identification). Bar charts indicate the total number of assigned sequences obtained for all taxa belonging to each taxonomic group. 42,915 sequences remained unassigned. doi:10.1371/journal.pone.0093740.g001

## Discussion and Conclusions

We used a metabarcoding approach to analyse the gut contents of eight ground-dwelling predators in banana plantations. We demonstrated that the addition of a new primary resource in the agroecosystem modified the diet for some of the predators. This is one of few studies that has used metabarcoding to investigate the gut contents of the arthropod food web. The results suggest that this is a feasible approach for ecological studies and also revealed some issues that will require methodological adjustments.

The efficiency of the mini-COI PCR was highly variable depending on the predator taxa examined, and this remains a major problem in achieving comprehensive identification of the prey ingested by a predator. While some of the predator taxa (e.g., the Lycosidae) were over-represented in the amplicon set, others were not amplified at all (e.g., the Hymenoptera). Regarding the latter, Yu *et al.* [34] also recently reported difficulties with COI sequencing of hymenopteran samples with 454 technology. This problem indicates that optimization of PCR conditions and primers is essential. The combination of several barcodes [35] could be used to achieve a complete description of the predator diets. The differences in PCR efficiencies also resulted in our inability to include some taxa in the food web and confirmed that metabarcoding data cannot be interpreted in a quantitative manner [22]. Relating the number of DNA sequences detected to the amount of material ingested appears to be an intractable problem. The quantification of trophic interactions would be enhanced by coupling the metabarcoding approach with stable isotope analysis [36] that enables assessment of the quantity of material used for the construction of an organism [14]. A second problem was the large number of the sequences that were unassigned despite our effort to sample comprehensively and thus

to include arthropod diversity from the studied sites in the bank of sequences. In addition, more than 4,000 sequences were assigned to the Neoptera infra-class level. Such uninformative identification may result from the incompleteness of the bank of sequences or from the incorrect identification of COI sequences recorded in GenBank [37]. The non-assignment of sequences should be prevented by an exhaustive and rigorous sampling in the studied ecosystem.

Despite the methodological shortcomings mentioned above (i.e., it must be borne in mind that the food webs described here are incomplete), metabarcoding is an excellent approach for inferring the food web from the natural environment and for detecting differences in its topology among treatments. These advantages were particularly appealing for our study, in which we aimed to detect differences in the predation of a major pest between two banana plantation management systems.

In the current study, the banana weevil *C. sordidus* was identified in the gut contents of three ground-dwelling predators: the earwig *E. caribeana*, the fire ant *S. geminata*, and the carpenter ant *C. sexguttatus*. These three species have previously been shown to feed on the banana weevil. In laboratory trials, earwigs of the genus *Euborellia* in Kenya [38] and other dermapterans in Indonesia [39] attacked *C. sordidus* eggs and larvae. The fire ant *S. geminata* is often described as an important generalist predator [40–42] and was observed to feed on *C. sordidus* in banana agroecosystems in Martinique [25]. Here, we showed that *S. geminata* workers that were directly sampled in the field were positive for *C. sordidus* DNA; however, the role of this species in the consumption of the pest may be underestimated in our study because ant workers usually carry prey to their nest to feed the colony [43]. Carpenter ants in the genus *Camponotus* were found in pseudostem leaf sheaths

**Table 2.** Frequency of taxa identified corresponding to each taxon analysed with the mini-COI barcodes.

Sample (n)	Barcoding identification	Rank	Frequency (n)	Number of sequences
Lycosidae (20)	<i>Trochosa</i>	Genus	100 (20)	18088
	<i>Pardosa</i>	Genus	65 (13)	2884
	<i>Pardosa milvina</i>	Species	15 (9)	9
	Lycosidae	Family	10 (2)	29
	<i>Pardosa amentata</i>	Species	5 (1)	11
	<i>Pardosa giebeli</i>	Species	5 (1)	2
	<i>Pardosa paludicola</i>	Species	5 (1)	3
	<i>Varacosa avara</i>	Species	5 (1)	4
<i>E. carai-bea</i> (97)	<i>E. carai-bea</i>	Species	100 (97)	39114
Scolopendra (6)	<i>Scolopendra</i>	Genus	100 (6)	544
	<i>Scolopendra mutilans</i>	Species	67 (4)	323
	Henicopidae	Family	50 (3)	864
	<i>Lamyctes hellyeri</i>	Species	50 (3)	8
	<i>Anopsobius giribeti</i>	Species	17 (1)	28
	<i>Scolopendra subspinipes</i>	Species	17 (1)	10
	<i>Scolopendra multidentis</i>	Species	17 (1)	8
	Pleurostigmophora	Subclass	17 (1)	7
	<i>Otostigmus aculeatus</i>	Species	17 (1)	4
Cicadellidae (3)	Cicadellidae	Family	67 (2)	29
<i>C. sordidus</i> (15)	<i>Cosmopolites sordidus</i>	Species	93 (14)	407
Oniscidae (2)	Oniscidae	Family	50 (1)	69
Blattodea (2)	<i>Blatella germanica</i>	Species	100 (2)	101
<i>P. mellerborgi</i> (2)	<i>Polytus mellerborgi</i>	Species	100 (2)	365
<i>Gryllus</i> (1)	<i>Gryllus</i>	Genus	100 (1)	136
	<i>Orocharis saltator</i>	Species	100 (1)	59
	Gryllidae	Family	100 (1)	16
<i>A. castelnaui</i> (2)	<i>Tribolium castaneum</i>	Species	100 (2)	16
Lumbricidae (2)	(not assigned)	–	–	–
Rhinocricidae (2)	(not assigned)	–	–	–
Paradoxosomatidae (2)	(not assigned)	–	–	–
<i>O. baurii</i> (96)	(not assigned)	–	–	–
<i>C. sexguttatus</i> (103)	(not assigned)	–	–	–
<i>S. geminata</i> (155)	(not assigned)	–	–	–
<i>W. auropunctata</i> (108)	(not assigned)	–	–	–
Staphilinidae (10)	(not assigned)	–	–	–

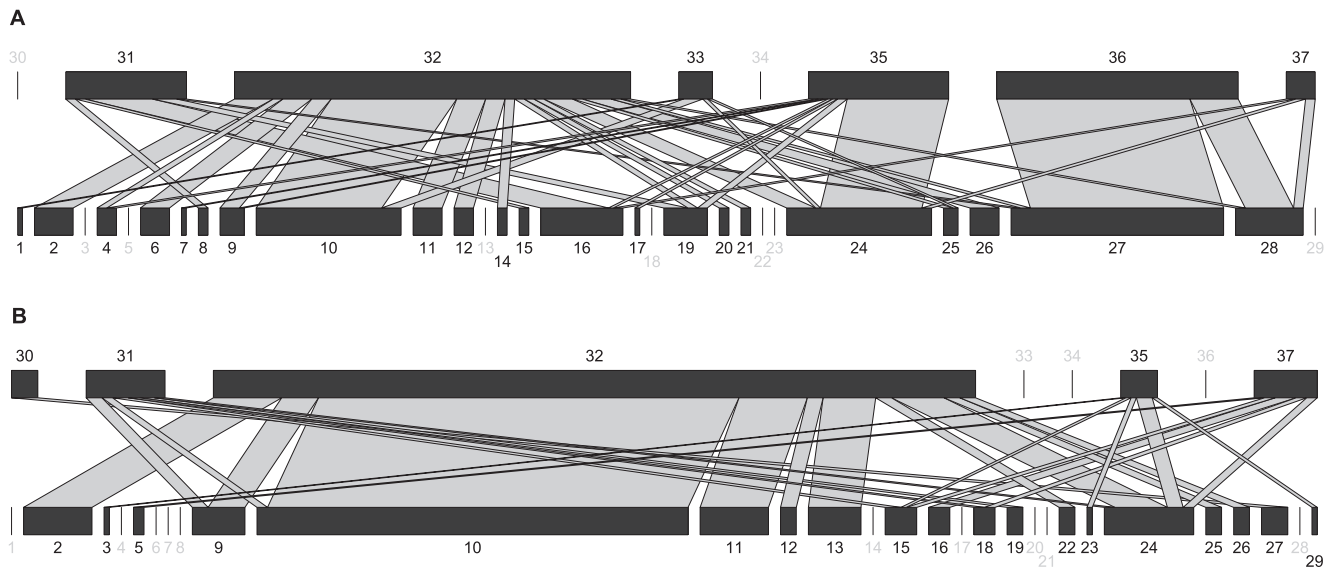
Generated by 454 pyrosequencing, the table displays the taxonomic rank, the frequencies of samples and the corresponding sample size in brackets, and the number of sequences corresponding to taxa identified by barcoding and belonging to the same order of the taxa analysed. A frequency of 100% means that all samples of the taxa analysed had at least two sequences of the taxa identified by barcoding (BLAST+ with E-value =  $10^{-20}$ , FASTA with 85% similarity threshold, and Nearest Neighbour algorithm for the final identification).

doi:10.1371/journal.pone.0093740.t002

and leaf trash in Indonesia and were predicted to forage in the banana plant [39], which is the habitat of immature stages of *C. sordidus*. Here, the diet analysis of *C. sexguttatus* revealed the consumption of *C. sordidus* by ant workers that were trapped in the field. In contrast, although ants in the genus *Odontomachus* are thought to be predators of *C. sordidus* [44] and although *O. baurii* consumed *C. sordidus* eggs in a laboratory assay of the current study, we did not detect *C. sordidus* DNA in *O. baurii* trapped in the field. From these results, we identified three species of predators that could be considered for the control of *C. sordidus* populations in banana plantations. Two of the three species are ants, which are

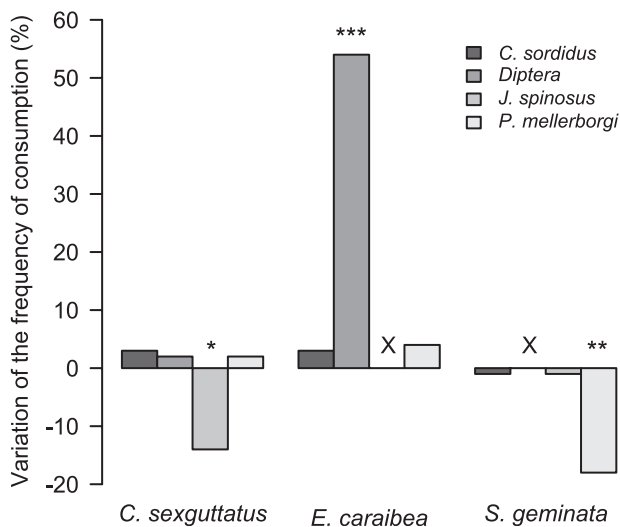
often assumed to play a key role in the regulation of *C. sordidus* in banana plantations [44–46].

In this study, we described the changes in the diets of generalist predators induced by plant diversification of a cropping system. We demonstrated that the use of a *B. decumbens* cover crop in banana plantations altered the arthropod food web, with significant changes in the frequency of consumption of some of the prey. Duyck *et al.* [47] found similar results based on stable isotope analyses, i.e., the trophic positions of generalist predators were changed by cover cropping. In the current study, the percentage of samples of the earwig *E. carai-bea* that were positive for dipteran DNA was higher in CCP (80%) than in BSP (26%).



**Figure 2. Bipartite food webs of predator-prey interactions on (A) bare soil, and (B) cover cropped banana plantation.** For each web, lower bars represent relative abundance of consumed prey, and upper bars represent relative abundance of positive ground-dwelling predators, each drawn at different scale. The width of links between ground-dwelling predators and prey represents the frequency of consumption. Numbers in grey indicate unlinked taxa. Visualization was performed with the R package "bipartite" [50]. 1: *Anopheles claviger*. 2: *Anopheles nimbus*. 3: *Baetis rhodani*. 4: *Blatella germanica*. 5: *Calliphora vomitoria*. 6: *Carabidae* spp. 7: *Codophila varia*. 8: *Coridius chinensis*. 9: *Cosmopolites sordidus*. 10: Diptera. 11: *Drosophila anceps*. 12: *Drosophila melanica*. 13: *Drosophila montana*. 14: *Gryllus*. 15: Hemiptera. 16: *Jalysus spinosus*. 17: *Nebria chinensis*. 18: *Neoneides muticus*. 19: *Nezara viridula*. 20: Oniscidae. 21: *Ophyra spinigera*. 22: *Periplaneta americana*. 23: *Podisus seriveventris*. 24: *Polytus mellerborgi*. 25: *Resseliella yagoi*. 26: *Sarcophila*. 27: *Scolopendra*. 28: *Scolopendra mutilans*. 29: *Stephensiella sterrei*. 30: Lycosidae. 31: *Camponotus sexguttatus*. 32: *Euborellia carai-bea*. 33: *Odontomachus baurii*. 34: Scolopendridae. 35: *Solenopsis geminata*. 36: Staphilinidae. 37: *Wasmannia auropunctata*. doi:10.1371/journal.pone.0093740.g002

This diet change suggests the *B. decumbens* cover crop probably increases the abundance of dipterans and thereby increased their consumption by *E. carai-bea*. In sugarcane fields in Hawaii, weeds favour dipterans by providing food, shade, and resting areas [48].



**Figure 3. Diet changes of ground-dwelling predators between bare soil plot and cover cropped plot.** The bar charts display the difference of frequencies of consumption for each prey calculated between the two plots, with the bare soil plot as a reference. The significance of the difference of frequencies of consumption observed between plots was assessed with a Fisher's exact test. (\*\*\*:  $p$ -value < 0.0001; \*\*:  $p$ -value < 0.001; \*:  $p$ -value < 0.01). Black crosses above bar charts indicate that the prey was not detected in both treatments. doi:10.1371/journal.pone.0093740.g003

Management of the *B. decumbens* cover crop by mowing is required to maintain the trade-off between the increase of predator densities and pest control. However, the increase in alternative prey (i.e., prey other than the target pest) in the diet of generalist predators exemplifies the processes that can dampen the positive effects of cover crops on pest regulation. In other words, the predators may increase consumption of non-pests without increasing consumption of pests.

In conclusion, it is essential to disentangle trophic interactions in order to achieve a better understanding of ecosystem resilience and persistence following disturbances [49], such as plant diversification [3]. DNA metabarcoding allows direct inference of trophic interactions and enables the assessment of arthropod diet. Although the method has limitations, including the inability to discriminate between direct predation, secondary predation, and scavenging [20], it has the potential to be very useful for describing arthropod food webs. Here, we identified new and unexpected trophic interactions in the predator-prey system in banana plantations. The accurate determination of trophic networks will challenge current models of trophic interactions and will contribute to food web theory and ecosystem management. In addition to its application to individual food webs, DNA metabarcoding could be used to link different food webs, such as those that describe micro-organisms, plants, arthropods, and larger animals.

## Supporting Information

**Table S1 List of taxa collected in banana plantations in order to sequence CO1 with SANGER method.** This sampling was performed to build the bank of sequence from individuals of the studying site. Sequences were aligned (MUSCLE algorithm) to build the consensus sequence used during the

bioinformatics processing of raw sequences (Consensus sequence of mini-CO1:5'-TTTATATTTTATTTTGGARCTTGAG-CAGGAATAGTAGGAACCTTCATTAAGAATAHTTATTC-GAGCAGAAATTAGGAMAACCCGGATCATTAAATTGGT-GATGATCAAATTTATAATGTTATTGTTACA-3').  
(DOCX)

**Table S2 List of taxa collected in banana plantations for the 454 pyrosequencing.** Taxa were collected for diet analyses (ground dwelling predators, n=572 samples), and for positive controls of the 454 pyrosequencing run (n=59 samples). Positive controls are designed to check the efficiency of the pyrosequencing run.  
(DOCX)

**Table S3 List of species identified with GenBank.** These taxa were recovered by blasting raw sequences derived from the 454 pyrosequencing run from the gut contents of ground-dwelling predators to GenBank database. Samples of these prey species were not collected during the sampling campaign designed in order to construct the bank of sequences. Identification to higher taxonomic rank results from an equal score calculated between two or more sequences of species recorded in GenBank, and the table displays taxa only for sequences identified to species rank.  
(DOCX)

**Table S4 List of tags used for forward and reverse primers.** Each sample had a specific combination of tagged primers allowing an assignment of sequence to its respective sample ID. Each primer was added with a 7 nucleotide sequence (tag) at the 5'-end.

## References

- Tilman D, Cassman KG, Matson PA, Naylor R, Polasky S (2002) Agricultural sustainability and intensive production practices. *Nature* 418: 671–677.
- Letourneau DK, Armbrrecht I, Riviera BS, Lerma JM, Carmona EJ, et al. (2011) Does plant diversity benefit agroecosystems? A synthetic review. *Ecol Appl* 21: 9–21.
- Simon S, Bouvier J-C, Debras J-F, Sauphanor B (2009) Biodiversity and pest management in orchard systems. A review. *Agron Sustain Dev* 30: 139–152.
- Barberì P, Burgio G, Dinelli G, Moonen AC, Otto S, et al. (2010) Functional biodiversity in the agricultural landscape: relationships between weeds and arthropod fauna. *Weed Research* 50: 388–401.
- Wise DH, Moldenhauer DM, Halaj J (2006) Using stable isotopes to reveal shifts in prey consumption by generalist predators. *Ecol Appl* 16: 865–876.
- Blanchard JL, Jennings S, Law R, Castle MD, McCloghrie P, et al. (2009) How does abundance scale with body size in coupled size-structured food webs? *J Anim Ecol* 78: 270–280.
- Cohen JE, Jonsson T, Carpenter SR (2003) Ecological community description using the food web, species abundance, and body size. *Proc Natl Acad Sci U S A* 100: 1781–1786.
- Jennings S, Mackinson S (2003) Abundance-body mass relationships in size-structured food webs. *Ecol Lett* 6: 971–974.
- Jonsson T, Cohen JE, Carpenter SR (2005) Food Webs, Body Size, and Species Abundance in Ecological Community Description. *Adv Ecol Res*. 1–84.
- van Veen FJF, Müller CB, Pell JK, Godfray HCJ (2008) Food web structure of three guilds of natural enemies: predators, parasitoids and pathogens of aphids. 0021–8790 2008: 191–200.
- Cabana G, Rasmussen JB (1996) Comparison of aquatic food chains using nitrogen isotopes. *Proc Natl Acad Sci U S A* 93: 10844–10847.
- Fry B (1988) Food web structure on Georges Bank from stable C, N, and S isotopic compositions. *Limnol Oceanogr* 33: 1182–1190.
- Peterson BJ, Howarth RW, Garritt RH (1985) Multiple stable isotopes used to trace the flow of organic matter in estuarine food webs. *Science* 227: 1361–1363.
- Ponsard S, Arditi R (2000) What can stable isotopes ( $\delta^{15}\text{N}$  and  $\delta^{13}\text{C}$ ) tell about the food web of soil macro-invertebrates? *Ecology* 81: 852–864.
- Post DM (2002) Using stable isotopes to estimate trophic position: Models, methods, and assumptions. *Ecology* 83: 703–718.
- Vander Zanden MJ, Rasmussen JB (1999) Primary consumer delta C-13 and delta N-15 and the trophic position of aquatic consumers. *Ecology* 80: 1395–1404.
- Vander Zanden MJ, Rasmussen JB (2001) Variation in delta N-15 and delta C-13 trophic fractionation: Implications for aquatic food web studies. *Limnol Oceanogr* 46: 2061–2066.
- Harwood JD, Bostrom MR, Hladilek EE, Wise DH, Obrycki JJ (2007) An order-specific monoclonal antibody to *Diptera* reveals the impact of alternative prey on spider feeding behavior in a complex food web. *Biol Control* 41: 397–407.
- Symondson WOC (2002) Molecular identification of prey in predator diets. *Mol Ecol* 11: 627–641.
- King RA, Read DS, Traugott M, Symondson WOC (2008) Molecular analysis of predation: a review of best practice for DNA-based approaches. *Mol Ecol* 17: 947–963.
- Harper GL, King RA, Dodd CS, Harwood JD, Glen DM, et al. (2005) Rapid screening of invertebrate predators for multiple prey DNA targets. *Mol Ecol* 14: 819–827.
- Pompanon F, Deagle BE, Symondson WOC, Brown DS, Jarman SN, et al. (2012) Who is eating what: diet assessment using next generation sequencing. *Mol Ecol* 21: 1931–1950.
- Scherber C, Eisenhauer N, Weisser WW, Schmid B, Voigt W, et al. (2010) Bottom-up effects of plant diversity on multitrophic interactions in a biodiversity experiment. *Nature* 468: 553–556.
- Bugg RL, Waddington C (1994) Using cover crops to manage arthropod pests of orchards - A review. *Agric Ecosyst Environ* 50: 11–28.
- Mollot G, Tixier P, Lescouret F, Quilici S, Duyck PF (2012) New primary resource increases predation on a pest in a banana agroecosystem. *Agricultural and Forest Entomology*.
- Hebert PDN, Cywinska A, Ball SL, DeWaard JR (2003) Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London, Series B: Biological Sciences* 270: 313–321.
- 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. *Molecular Marine Biology and Biotechnology* 3: 294–297.
- Meusnier I, Singer GAC, Landry JF, Hickey DA, Hebert PDN, et al. (2008) A universal DNA mini-barcode for biodiversity analysis. *BMC Genomics* 9.
- Leray M, Agudelo N, Mills SC, Meyer CP (2013) Effectiveness of annealing blocking primers versus restriction enzymes for characterization of generalist diets: unexpected prey revealed in the gut contents of two coral reef fish species. *Plos One* 8.
- Megléc E, Piry S, Desmarais E, Galan M, Gilles A, et al. (2010) Sesame (SEquence Sorter & AMplicon Explorer): Genotyping based on high-throughput multiplex amplicon sequencing. *Bioinformatics* 27: 277–278.
- Piry S, Guivier E, Realini A, Martin JF (2012) |SE|S|AM|E| Barcode: NGS-oriented software for amplicon characterization - application to species and environmental barcoding. *Molecular Ecology Resources* 12: 1151–1157.
- Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32: 1792–1797.

33. Team RDC (2012) R: A language and environment for statistical computing. Vienna, Austria.
34. Yu DW, Ji YQ, Emerson BC, Wang XY, Ye CX, et al. (2012) Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *MEE* 3: 613–623.
35. Valentini A, Pompanon F, Taberlet P (2009) DNA barcoding for ecologists. *Trends Ecol Evol* 24: 110–117.
36. Carreon-Martinez L, Heath DD (2010) Revolution in food web analysis and trophic ecology: Diet analysis by DNA and stable isotope analysis. *Mol Ecol* 19: 25–27.
37. Harris DJ (2003) Can you bank on GenBank? *Trends Ecol Evol* 18: 317–319.
38. Koppenhofer AM (1993) Egg predators of the banana weevil, *Cosmopolites sordidus* (Germar) (Coleoptera, Curculionidae) in western Kenya. *J Appl Entomol* 116: 352–357.
39. Abera-Kalibata AM, Hasyim A, Gold CS, Van Driesche R (2006) Field surveys in Indonesia for natural enemies of the banana weevil, *Cosmopolites sordidus* (Germar). *Biol Control* 37: 16–24.
40. Nickerson JC, Kay CAR, Buschman LL, Whitcomb WH (1977) The presence of *Spissistilus festinus* as a factor affecting egg predation by ants in soybeans. *Fla Entomol* 60: 193–199.
41. Risch SJ (1981) Ants as important predators of rootworm eggs in the neotropics. *J Econ Entomol* 74: 88–90.
42. Vandenberg H, Bagus A, Hassan K, Muhammad A, Zega S (1995) Predation and parasitism on eggs of two pod-sucking bugs, *Nezara viridula* and *Pizodorus hybneri*, in soybean. *Int J Pest Manag* 41: 134–142.
43. Dornhaus A, Holley JA, Pook VG, Worswick G, Franks NR (2008) Why do not all workers work? Colony size and workload during emigrations in the ant *Temnothorax albipennis*. *Behav Ecol Sociobiol* 63: 43–51.
44. Abera-Kalibata AM, Gold CS, Van Driesche R (2008) Experimental evaluation of the impacts of two ant species on banana weevil in Uganda. *Biological Control* 46: 147–157.
45. Gold CS, Pena JE, Karamura EB (2001) Biology and integrated pest management for the banana weevil *Cosmopolites sordidus* (Germar) (Coleoptera: Curculionidae). *Integrated Pest Management Reviews* 6: 79–155.
46. Perfecto I, Castañeras A (1998) Deployment of the predaceous ants and their conservation in agroecosystems. In: Barbosa P, editor. *Conservation Biological Control*. San Diego: Academic Press. 269–290.
47. Duyck PF, Lavigne A, Vinatier F, Achard R, Okolle JN, et al. (2011) Addition of a new resource in agroecosystems: Do cover crops alter the trophic positions of generalist predators? *Basic Appl Ecol* 12: 47–55.
48. Topham M, Beardsley JWJ (1975) Influence of nectar source plants on the New Guinea sugarcane weevil parasite, *Lixophaga sphenophori* (Villeneuve). *Proc Hawaii Entomol Soc* 12: 145–154.
49. Berlow EL, Neutel AM, Cohen JE, de Ruiter PC, Ebenman B, et al. (2004) Interaction strengths in food webs: issues and opportunities. *J Anim Ecol* 73: 585–598.
50. Dormann CF, Fruend J, Bluethgen N, B G (2009) Indices, graphs and null models: analyzing bipartite ecological networks. *The Open Ecology Journal* 2: 7–24.



# |SE|S|AM|E| Barcode: NGS-oriented software for amplicon characterization – application to species and environmental barcoding

S. PIRY,\* E. GUIVIER,\* A. REALINI† and J.-F. MARTIN†

\*INRA (UMR CBGP Centre de Biologie Pour la Gestion des Populations), Campus international de Baillarguet, CS 30016, F-34988 Montpellier-sur-Lez Cedex, France, †Montpellier SupAgro (UMR CBGP Centre de Biologie Pour la Gestion des Populations), Campus international de Baillarguet, CS 30016, F-34988 Montpellier-sur-Lez Cedex, France

## Abstract

Progress in NGS technologies has opened up new opportunities for characterizing biodiversity, both for individual specimen identification and for environmental barcoding. Although the amount of data available to biologist is increasing, user-friendly tools to facilitate data analysis have yet to be developed. Our aim, with |SE|S|AM|E| BARCODE, is to provide such support through a unified platform. The sequences are analysed through a pipeline that (i) processes NGS amplicon runs, filtering markers and samples, (ii) builds reference libraries and finally (iii) identifies (barcodes) the sequences in each amplicon from the reference library. We use a simulated data set for specimen identification and a recently published data set for environmental barcoding to validate the method. The results obtained are consistent with the expected characterizations (*in silico* and previously published, respectively). |SE|S|AM|E| BARCODE and its documentation are freely available under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported Licence for Windows and Linux from <http://www1.montpellier.inra.fr/CBGP/NGS/>.

**Keywords:** biodiversity, database, diet analysis, metagenomics

Received 22 February 2012; revision received 30 May 2012; accepted 31 May 2012

## Introduction

Major breakthroughs are being made in the characterization of biodiversity on the basis of molecular markers. In particular, international initiatives for taxon identification based on DNA barcoding have been boosted by continuing advances in sequencing technology (Hebert *et al.* 2003). This important development is directly related to the generalization of next-generation sequencing (NGS), which delivers massive DNA sequences at a reasonable cost and level of effort. This has led directly to the accumulation of sequences in databases ranging from Barcode of Life and GenBank/EMBL repositories to in-house projects. Much effort has been focused so far on the normalization of database systems (barcode data standard) and the links between the major databases. There is still an urgent need to provide analytical tools for mining the huge amount of data available to end-users, who have diverse backgrounds and interests in the characterization of biodiversity.

In this context, we have created |SE|S|AM|E| BARCODE, a user-friendly web application for barcoding

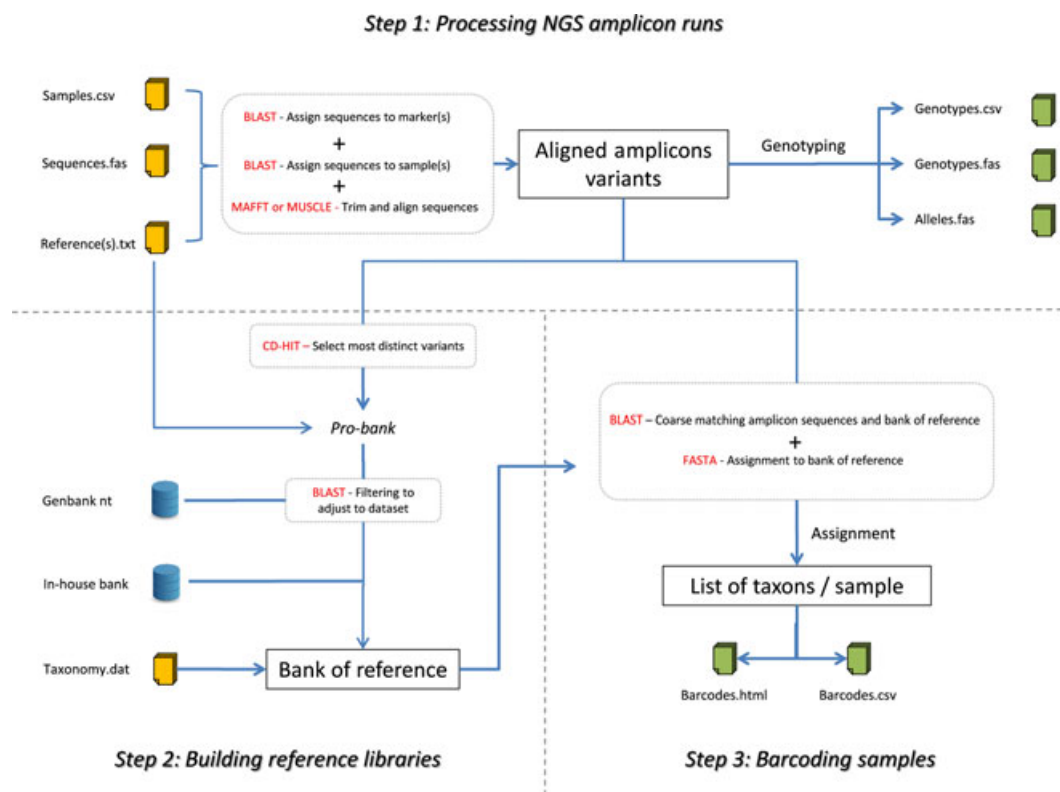
amplicon sequences obtained through NGS. The software automatically processes DNA sequence data sets, sorting for multiplexed loci and samples identified by oligonucleotide tags, and allows for the assignment of sequences to a preloaded reference library (from GenBank to in-house scale). The results are displayed both as sample-based synthetic tables and graphical representations. The output of |SE|S|AM|E| BARCODE is designed with multiple applications in mind, including specimen assignment to species referenced in DNA libraries, molecular characterization of trophic networks and, more generally, metagenomics and environmental barcoding studies (Soininen *et al.* 2009; Valentini *et al.* 2009).

## Methods

### Step 1: Processing NGS amplicon runs

DNA reads from a sequencing run are imported as a single FASTA file, making the application independent of the acquisition technology. The simultaneous analysis of several markers sequenced together in a single run is supported. Each DNA sequence is BLASTed (Camacho *et al.* 2009) against the marker reference sequence or sequences provided by the user for the run (Fig. 1).

Correspondence: Jean-François Martin, Fax: +33-499-623-345; E-mail: martinjf@supagro.inra.fr



**Fig. 1** Overview/workflow of ISE|S|AM|E| BARCODE for barcoding NGS amplicon runs. The three main steps of the workflow are represented and separated by dashed lines. Rectangular boxes depict the main outcomes of the three steps of the pipeline: Aligned amplicon variants for the processing of NGS amplicon runs, reference library for the building of reference libraries and the list of taxa/samples for the sample barcoding step. Special attention was paid to the input and output files, differentiating the plain text files from the databases used in the workflow. We focused on the conceptual processes, but the external software used is indicated for each process.

This allows the assignment of sequences to a specific marker and the determination of read orientation. Sequences can be assigned to samples through the use of unique oligonucleotide tags, as previously described (Galan *et al.* 2010). Any combination, from no tag (only one amplicon in the data set) to the use of both forward and reverse tags, is possible. The information about each amplicon is read from a comma-separated values (CSV) file that contains sample names, marker names, ploidy levels, primer and tag sequences, population and species names when relevant and taxon ID. For each amplicon, the sequences and the reference marker are aligned by MAFFT (Kato *et al.* 2005) or MUSCLE (Edgar 2004), according to the administrator's choice. Tags and primers are trimmed off from the sequences, based on the alignment. Finally, singleton sequences or incomplete (too short) sequences can be filtered out from the analysis.

The manual exploration of the alignments of sequences for each individual amplicon is integrated. This includes summary statistics, such as the number of sequences assigned to the amplicon and the number of user-validated alleles, if relevant. Selecting an amplicon

displays the alignment of all sequences from this amplicon. Identical sequences are pooled into variants and their number and frequencies in the amplicon are provided. Each variant is characterized by its length, total occurrence in the whole run and the number of amplicons in which it is found. Users can validate selected variants as alleles by a simple click. A unique allele identifier is attributed in a given project. This allows comparisons between multiple runs within a project. After allele validation, the genotype table, or the FASTA file of alleles can be produced and exported. Filtering options are available to display this information for selected data sets only (e.g. run/populations/species).

### Step 2: Building reference libraries

The assignment of variants to references libraries requires complementary information: a library to which amplicon sequences will be assigned, and taxonomic information relating to the sequences included in the database. We decided to leave keep this process highly flexible, using open formats so that the user can fully



control both aspects of the library, and pre-installed GenBank features. Sequence libraries can be imported as an *ad hoc* tabular file by the user or automatically built by |SE|S|AM|E| BARCODE. This automatic building is based on two steps. The first is the building of a pro-bank data set for the extraction of sequences from the GenBank database as relevant references for the barcoding of the sequencing run (Fig. 1).

The pro-bank contains at least one reference sequence (used also as a reference marker). The variants from the sequencing run can also be added to the pro-bank, to maximize representation of the taxonomic coverage of the sequencing run. We minimized computation time and maximized representativeness using CD-HIT (Li & Godzik 2006) to build a hierarchical distance framework describing the distance between the variants in the sequencing run and to select the most divergent sequences at a given threshold for subsequent addition to the pro-bank data set. Finally, sequences can be freely added or deleted to customize the pro-bank data set according to the user's interest, thus leaving as many options as possible open to the user. In the second step, each sequence from the pro-bank is BLASTed against GenBank and all matching sequences at a given e-value threshold are imported into the reference library. We prefer this method over searching for annotation fields, which are often incomplete or ambiguous (for example, Cytochrome *c* oxidase subunit 1 may be listed alternatively as COI, *co1* or *cox1* in GenBank). The matching sequences are appended to the bank reference and checked for redundancy. This automatic building ensures optimal coverage of the bank reference, given the sequencing run and the GenBank database. The taxonomic information is provided as an *ad hoc* tabular file that contains a list of taxon names, their rank, their ID and the ID of their parent taxon, in a recursive scheme. The format is that used in GenBank, although the information can be customized to fit the needs of any in-house project (see Supporting Information).

### Step 3: Barcoding samples

The assignment of the variants to the library is based on the combination of BLAST and FASTA (Pearson & Lipman 1988). In a first step, BLAST is used to restrict the assignment within the library to related sequences (based on the e-value threshold retained). FASTA is then performed on these related sequences from the library, to estimate similarity with the query variant. The number of sequences from the library used for assignment can be customized according to computing limitations. Subsequently, for the defined subset of related sequences from the library, assignment is based on the percentage difference between the variant considered and the

sequences from the library. Assignment is based on a similarity threshold (expressed as a percentage difference) and all sequences compatible with the threshold are considered equally potential assignments (Fig. 1). Rather than assigning a variant only to the most closely related sequence in the library (although that is possible), we chose to consider all sequences above the similarity threshold as equally likely to be assigned. This assignment fuzziness is chosen to overcome the issues caused by the relatively high error rate of recent sequencing technologies (Gilles *et al.* 2011), as well as taxonomic uncertainties. This potential uncertainty is taken into account by assigning the variant to the nearest taxon common to the retained sequences. The results are produced as a table of assigned taxa for each amplicon in the sequencing run and a graphical representation of the selected samples (potentially multiple amplicons) displaying the taxonomic content, both as qualitative and as quantitative displays.

### Implementation

|SE|S|AM|E| BARCODE resulted from ongoing efforts to generate a user-friendly tool for biologists and ecologists, enabling them to manage and barcode DNA sequence data sets easily. The original processing of the sequencing runs and the manual exploration phases (step 1) were upgraded from SESAME (Meglecz *et al.* 2011), as the tools and interfaces developed in this software were particularly well adapted to the barcoding application. The barcoding application could have been developed independently and connected to SESAME, but we wanted to make the user experience as easy and straightforward as possible and we decided to publish the barcoding software presented here as a single, independent package.

### Installing |SE|S|AM|E| BARCODE

With this constant concern to facilitate the user's experience, we provide both the installation files required to set up a traditional server (see Supporting Information) and a virtual version of the server already installed, available as a single OVA appliance file usable on any computer with virtualization software installed (tested with Virtualbox (<https://www.virtualbox.org/>)). In both cases, the installation guide (see Supporting Information) helps the user to install and deploy the application step by step. The minimal requirements for the machine (virtualized or not) are as follows: 32 bits, 4 Gb (native) or 8 Gb (virtual) Ram, 20 Gb HDD free space, 1 cpu (better with 2 cores). Recommended requirements for optimal performance are the following: 64 bits, 12 Gb Ram (native) or 16 Gb or more (virtual), 500 Gb HDD, 8 cores.

### Using the application

The user interacts with |SE|S|AM|E| BARCODE through a database-driven web application. The administration tab helps to define the configuration of paths to data (locally downloaded nt database and GenBank related links) and external programs installed, such as BLAST, FASTA, CD-HIT, MUSCLE and MAFFT. The administrator also defines the privileges of users for current projects (collection of sequencing runs). Use of the application is based on the workflow detailed in the Methods section, each major step corresponding to one or multiple tabs.

*Step 1: Processing sequencing runs.* The definition of the reference markers (tab 'Markers') is a prerequisite for the processing of sequencing runs. Optional parameters can be used to adapt processing to a range of different projects, including DNA barcoding. The current project hub, (tab 'Runs') compiles the information from all the sequencing runs in a specific project and allows for the processing of a new sequencing run through the assistant 'Uploads and analyses'. This assistant guides the user from the raw FASTA file and sample information (CSV file) to processed amplicon-based alignments ready for genotyping or barcoding analyses. This involves multiple steps of filtering and sorting the data to reflect the sequence diversity for each amplicon in the sequencing run. Optional triggers and details are provided in the user guide (see Supporting Information). The user can explore and edit the samples (tab 'Samples') on a project/run basis, with multiple filter options and exportation options available. Finally, the amplicon explorer (tab 'Genotyping → Sequences Analysis') helps the user to assess the sequence content of each amplicon as a multiple sequence alignment. |SE|S|AM|E| BARCODE has been designed to allow multiple export formats and to provide a direct interface for further external analyses (tab 'Genotyping → Results').

*Step 2: Building reference libraries.* Taxonomy is the core element matched to the reference sequences. We strongly encourage the use of the GenBank Taxonomy file, for which direct implementation is possible *via* the tab 'Barcoding → Taxonomy' although in-house taxonomy files can also be used. Some tools for exploring and filtering the loaded taxonomy are available in the same tab. The tab 'Barcoding → Bank of Sequences builder' allows reference libraries to be built in multiple ways, as detailed in the Methods section. The main window displays the sequences currently used in the pro-bank to build the reference library. This pro-bank can be filled manually (one by one) or automatically derived from the sequencing runs themselves. Finally, the bank of sequences used for assignment can be explored, filtered and manipulated in

the tab 'Barcoding → Bank of sequences'. The management of the library, through the filtering and removal of misannotated sequences, is an important step, greatly increasing the accuracy of barcoding processing results. Moreover, it may be of particular interest to export the library in its entirety or in part for different projects using the same set of sequence data.

*Step 3: Barcoding samples.* Amplicons are barcoded from the tab 'Barcoding → Processing'. All the parameters for ensuring the best trade-off between accuracy and computing time are can be managed through this interface. In particular, the restriction of assignments to a taxon can greatly reduce computing times, but may lead to the user missing unrelated species in the amplicon (human contamination for example). In the same way, the manipulation of BLAST e-value and/or FASTA similarity threshold allows maximum flexibility (documented in detail in the user guide, see Supporting Information). Once processed, the barcoded amplicons can be analysed and visualized in the tab 'Barcoding → Results'. Each amplicon/pool of amplicons from the Project/Run can be detailed for its assigned taxons, in a qualitative or quantitative manner. The assignment method (nearest neighbour or user-defined similarity threshold) can be modified as required, to explore the sensitivity of the method with particular settings. The Samples by Taxonomy grid allows the user to obtain detailed information on the distribution of taxons across the amplicons. Double-clicking on a taxon in the grid displays an assignment table with details of the similarity between sequences from the amplicon and individual sequences from the reference library. The Synthetic taxonomic image panel displays nested taxonomic boxes representing the taxonomic diversity and structure present in the amplicon of interest. All these qualitative and/or quantitative results can be exported for external use, in multivariate analyses of sample diversity, for example.

### Data sets analysed

The algorithms used in |SE|S|AM|E| BARCODE were validated with two data sets. The first highlights the power and limitation of barcoding specimens to species with GenBank nt and is based on simulated data, whereas the second data set was taken from the literature and demonstrates the usefulness of the application in an environmental barcoding context.

#### *Barcoding aphid species with simulated NGS data*

We selected 15 COI sequences of 658 bases from 15 different and closely related species of aphids (Hemiptera: Aphidoidea). Their similarity (1 – percentage divergence)

ranged from 91.1% to 99.6%. For depiction of the error rates of NGS sequencing technologies, such as 454 sequencing, we simulated an error model (normally distributed random error, available on request) and simulated 1000 derived sequences for each species. This provided a complete data set of 15 simulated amplicons, each corresponding to one original sequence. We added primers and oligonucleotide tags *in silico* and generated a FASTA file and a corresponding CSV file describing the 15 simulated amplicons. The reference sequence was chosen randomly from the original data set (for testing sensitivity during pro-bank construction).

*Step 1: Processing of the sequencing run.* The 15 000 sequences representing the 15 amplicons (1000 sequences each) were analysed through the run assistant, using default parameters, yielding a total of 8887 variants. This large number of different variants results from the noising algorithm used to simulate the sequencing error rate.

*Step 2: Building reference libraries.* The taxonomy was loaded directly from GenBank. The pro-bank was built from the variants from the sequencing run, with default parameters. Only two distinct sequences (the reference marker sequence and a variant from the sequencing run) were kept in the pro-bank, based on hierarchical clustering and the default similarity threshold. They were used to compute the reference library for the project. This library contained 556 368 sequences. Within this library, 391 443 records did not seem to be assigned to a species rank. We decided to remove these sequences from the analysis, as the only taxonomic information available for these records was 'Hemiptera sp.', which is of no use for species identification, at least in the case study. The final library thus contained 164 925 sequences, against which the assignment process was performed.

*Step 3: Barcoding samples.* The amplicons were barcoded with the default 95% similarity threshold. Most amplicons were assigned to the Hemiptera or Aphididae. This is not surprising given the close similarity of the selected sequences and it provides support for our assignment result based on the closest common taxon. The simulated error rate here did not differ considerably from that observed for real data (Gilles *et al.* 2011). Caution is therefore required when choosing a barcoding marker, to ensure that it has sufficient polymorphism to overcome the sequencing error. We modified the similarity threshold to 99.7%, to correct for sequence similarities and obtained a correct assignment of all amplicons to the expected species (Fig. 2). The whole analysis took 2 h 25 min with our computer system, from the importation of the sequences to the final barcode results.

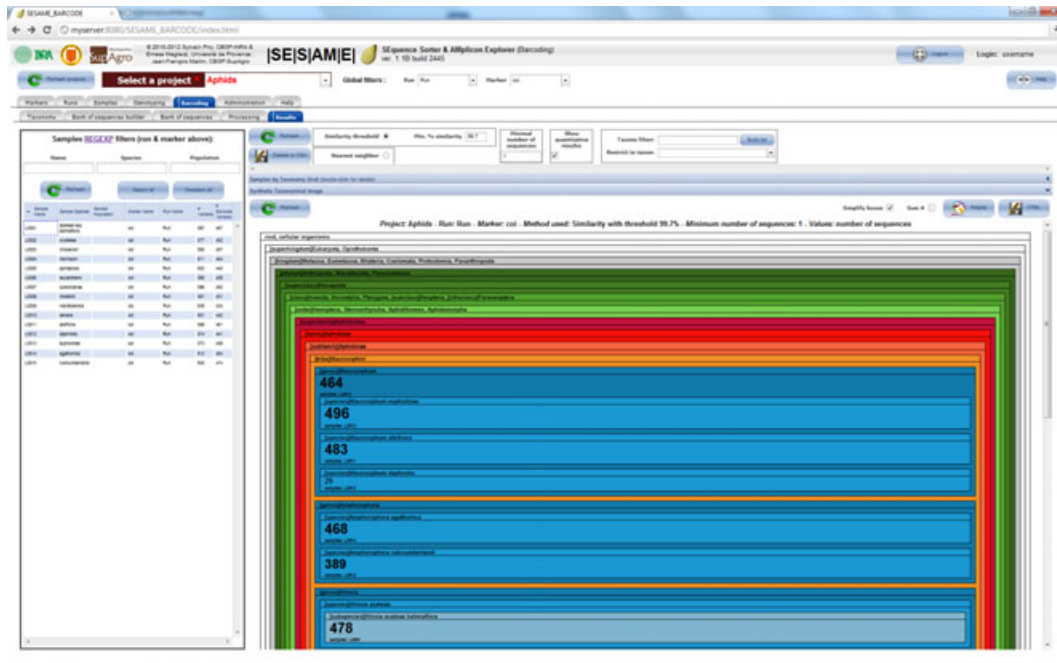
### *Pyrosequencing of prey DNA in reptile faeces (Molecular Ecology Resources 12–2, 2012)*

The validation of |SE|S|AM|E| BARCODE is also particularly important for environmental barcoding data. We checked the validity of our developed method with an already published work from Mol. Ecol. Resources on reptile diets (Brown *et al.* 2012). The aim of the original article was to determine which earthworm species are exploited by slow worms (the legless lizard *Anguis fragilis*) and was based on 454 pyrosequencing data from faecal samples. It provided a quantitative analysis of diet for four different habitats. We reanalysed these data and checked whether our method gave similar results. The data set consists of four FASTA files corresponding to the different habitats analysed. We added the published primers and arbitrary oligonucleotide tags to mimic a single-sequencing run.

*Step 1: Processing of the sequencing run.* In total, 1896 sequences provided in the original paper were analysed for the four habitats, yielding a total of 538 different variants.

*Step 2: Building reference libraries.* The taxonomy was directly loaded from GenBank. The pro-bank was built from the variants from the sequencing run, with default parameters. Three distinct sequences (the reference marker sequence and two variants from the sequencing run) were kept in the pro-bank and used to compute the reference library for the project. This reference library contained 636 sequences. This smaller number of sequences than for the previous project reflects the lower level of taxonomic coverage in GenBank, requiring the addition of an in-house library that was used in the original study.

*Step 3: Barcoding samples.* The amplicons were barcoded with the default 95% similarity threshold for BLAST+FASTA assignment. We ensured that our results were comparable with those of the original article, using the nearest neighbour assignment method. The results are provided in Table 1. Most results from |SE|S|AM|E| BARCODE were consistent with the original study. The two recurrent differences (whatever the parameters used) were (i) assignment to *Dendrodrilus rubidus*, only recently added to GenBank and probably absent when the data set was first analysed and (ii) the absence of *Aporrectodea caliginosa* from the sample from the 'Caerphilly' site and of *Lumbricus festivus* from the sample from the 'East Cowes' site. This discrepancy may result from the use of an in-house bank in addition to GenBank in the original study (Brown *et al.* 2012). The whole analysis took 40 min with our computer system, from the importation of the sequences to the final barcode results provided in Table 1.



**Fig. 2** Sample barcoding results. This screenshot of the web application corresponds to the Synthetic taxonomic image. The 15 simulated amplicons were barcoded at a 99.5% similarity threshold. Customizable colours (see online version) correspond to the taxon ranks from root to subspecies. As the ‘Show quantitative results’ box has been ticked, the values reported on the graphic representation correspond to the assigned sequences for a particular taxon.

**Table 1** Barcoding samples results. This table corresponds to the Sample x taxonomic information exported from the software. The four habitats were barcoded using the nearest neighbor algorithm after a 95% similarity threshold search. Numbers correspond to the DNA sequences assigned to a specific taxon for each region (as defined in Brown et al., 2012)

Taxonomy (nearest Neighbor)	Caerphilly	Ringwood	East Cowes	Flat Holm	Total
[genus] Lumbricus, Lumbricus rubellus complex, [species] <i>Lumbricus rubellus</i>	488	374	55	38	955
[genus] Lumbricus, [species] <i>Lumbricus castaneus</i>	19	17	12	2	50
[genus] Lumbricus, Lumbricus terrestris complex, [species] <i>Lumbricus terrestris</i>	10	3	4	112	129
[genus] Aporectodea, [species] <i>Aporrectodea longa</i>	33	20	100	0	153
[genus] Aporectodea, [species] <i>Aporrectodea caliginosa</i>	0	1	7	73	81
[genus] Aporectodea, [species] <i>Aporrectodea tuberculata</i>	0	4	7	0	11
[genus] Allolobophora, [species] <i>Allolobophora chlorotica</i>	10	0	1	14	25
[genus] Dendrodrilus, [species] <i>Dendrodrilus rubidus</i>	15	1	0	0	16
Unidentified	87	6	58	18	169

## Conclusion

Obtaining high number of sequences for amplicons from specimens or environmental samples has become a reality for molecular ecologists. The generation of such large amounts of data presents new challenges in the representation, understanding and analysis of the structure of the massive number of sequences that can be obtained. In this respect, bioinformatics tools are required to cope with the identification of amplicons sequenced in this high-throughput environment. ISE|S|AM|E| BARCODE fills this gap and brings the routine sequencing and DNA

barcoding of a large number of amplicons into the reach of molecular ecologists. This software was developed specifically for biologists, with a user-friendly interfacing and easy installation. Furthermore, most functions to support analyses are unique and provide a strong and productive framework for high-throughput amplicon barcoding.

## Acknowledgements

We thank André Gilles and Gregory Mollot for their useful comments in the design of ISE|S|AM|E| BARCODE and, in



particular, for their help in defining the appropriate output formats useful for molecular ecology studies. We thank Armelle Coeur-d'Acier, who participated in the design of the simulated data and provided the sequences. We thank Laurent Soldati for major improvements to the application design. We thank Julie Sappa from Alex Edelman and associates for major English improvements of the manuscript. We thank Maxime Galan for useful comments on the manuscript. Funding: This work was supported by a grant from the Scientific Council of Montpellier SupAgro.

## References

- Brown DS, Jarman SN, Symondson WOC (2012) Pyrosequencing of prey DNA in reptile faeces: analysis of earthworm consumption by slow worms. *Molecular Ecology Resources*, **12**, 259–266.
- Camacho C, Coulouris G, Avagyan V *et al.* (2009) BLAST plus: architecture and applications. *BMC Bioinformatics*, **10**, doi:10.1186/1471-2105-10-421.
- Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, **32**, 1792–1797.
- Galan M, Guivier E, Caraux G, Charbonnel N, Cosson JF (2010) A 454 multiplex sequencing method for rapid and reliable genotyping of highly polymorphic genes in large-scale studies. *BMC Genomics*, **11**, doi:10.1186/1471-2164-11-296.
- Gilles A, Meglecz E, Pech N *et al.* (2011) Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. *BMC Genomics*, **12**, doi:10.1186/1471-2164-12-245.
- Hebert PDN, Ratnasingham S, deWaard JR (2003) Barcoding animal life: cytochrome *c* oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society of London Series B-Biological Sciences*, **270**, S96–S99.
- Katoh K, Kuma K, Toh H, Miyata T (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Research*, **33**, 511–518.
- Li WZ, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.
- Meglecz E, Piry S, Desmarais E *et al.* (2011) SESAME (SEquence Sorter & AMplicon Explorer): genotyping based on high-throughput multiplex amplicon sequencing. *Bioinformatics*, **27**, 277–278.
- Pearson WR, Lipman DJ (1988) Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the United States of America*, **85**, 2444–2448.
- Soininen EM, Valentini A, Coissac E *et al.* (2009) Analysing diet of small herbivores: the efficiency of DNA barcoding coupled with high-throughput pyrosequencing for deciphering the composition of complex plant mixtures. *Frontiers in Zoology*, **6**, 16.
- Valentini A, Pompanon F, Taberlet P (2009) DNA barcoding for ecologists. *Trends in Ecology & Evolution*, **24**, 110–117.

---

S.P. conceived the idea, developed the software and participated in writing the manuscript, E.G. participated in designing the software, managed the quality assurance and wrote the manuals as well as participated in writing the manuscript. A.R. developed the software. J.-F.M. conceived the idea, tested the software, wrote the manuscript and managed the project.

---

## Data Accessibility

|SE|S|AM|E| BARCODE and its documentation are freely available under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported Licence for Windows and Linux from <http://www1.montpellier.inra.fr/CBGP/NGS/>.

## Supporting Information

Additional supporting information may be found in the online version of this article.

**Appendix S1** Installation of |SE|S|AM|E| BARCODE for LINUX (Ubuntu/Debian style).

**Appendix S2** Installation of |SE|S|AM|E| BARCODE virtual machine distribution (Windows/Mac/Linux).

**Appendix S3** |SE|S|AM|E| BARCODE installation for WINDOWS.

**Appendix S4** Genotyping and barcoding based on high-throughput multiplex amplicon sequencing. User guide – February 2012.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

# High-throughput microsatellite isolation through 454 GS-FLX Titanium pyrosequencing of enriched DNA libraries

THIBAUT MALAUSA,\* ANDRÉ GILLES,+ EMESE MEGLÉCZ,+ HÉLÈNE BLANQUART,‡  
STÉPHANIE DUTHOY,‡ CAROLINE COSTEDOAT,+ VINCENT DUBUT,+ NICOLAS PECH,+  
PHILIPPE CASTAGNONE-SERENO,\* CHRISTOPHE DÉLYE,§ NICOLAS FEAU,¶ PASCAL FREY,\*\*  
PHILIPPE GAUTHIER,++ THOMAS GUILLEMAUD,\* LAURENT HAZARD,‡‡ VALÉRIE LE CORRE,§  
BRIGITTE LUNG-ESCHARMANT,¶ PIERRE-JEAN G. MALÉ,§§ STÉPHANIE FERREIRA‡  
and JEAN-FRANÇOIS MARTIN++

\*INRA, UMR 1301 IBSV INRA/UNSA/CNRS, 400 Route des Chappes, BP 167, 06903 Sophia-Antipolis Cedex, France, †Aix-Marseille Université, CNRS, IRD, UMR 6116 – IMEP, Equipe Evolution Génome Environnement, Centre Saint-Charles, Case 36, 3 Place Victor Hugo, 13331 Marseille Cedex 3, France, ‡Genoscreen, Genomic Platform and R&D, Campus de l'Institut Pasteur, 1 rue du Professeur Calmette, Bâtiment Guérin, 59000 Lille, France, §INRA, UMR 1210 Biologie et Gestion des Adventices, 17 rue Sully, 21000 Dijon, France, ¶INRA, UMR 1202 BIOGECO, Equipe de Pathologie Forestière, Domaine de Pierroton, 69 route d'Arcachon, 33612 Cestas Cedex, France, \*\*INRA, Nancy-Université, UMR 1136, Interactions Arbres – Microorganismes, IFR 110, 54280 Champenoux, France, ++UMR CBGP (INRA/IRD/Cirad/Montpellier SupAgro), Campus International de Baillarguet, CS 30016, 34988 Montferrier-sur-Lez Cedex, France, ‡‡INRA – UMR 1248 AGIR, BP 52627, 31326 Castanet-Tolosan Cedex, France, §§UMR Evolution et Diversité Biologique (Université Toulouse III; CNRS), 118 Route de Narbonne, 31062 Toulouse, France

## Abstract

Microsatellites (or SSRs: simple sequence repeats) are among the most frequently used DNA markers in many areas of research. The use of microsatellite markers is limited by the difficulties involved in their *de novo* isolation from species for which no genomic resources are available. We describe here a high-throughput method for isolating microsatellite markers based on coupling multiplex microsatellite enrichment and next-generation sequencing on 454 GS-FLX Titanium platforms. The procedure was calibrated on a model species (*Apis mellifera*) and validated on 13 other species from various taxonomic groups (animals, plants and fungi), including taxa for which severe difficulties were previously encountered using traditional methods. We obtained from 11 497 to 34 483 sequences depending on the species and the number of detected microsatellite loci ranged from 199 to 5791. We thus demonstrated that this procedure can be readily and successfully applied to a large variety of taxonomic groups, at much lower cost than would have been possible with traditional protocols. This method is expected to speed up the acquisition of high-quality genetic markers for nonmodel organisms.

**Keywords:** enriched library, genetic marker, genotyping, microsatellite isolation, next-generation sequencing, primer design, pyrosequencing

Received 17 May 2010; revision received 27 December 2010; accepted 5 January 2011

## Introduction

Microsatellites (or SSRs: simple sequence repeats) are among the most frequently used DNA markers in many areas of research (Sunnucks 2000). However, their availability and quality are limited by the difficulties of *de novo* development in species for which no genomic information is available. The most commonly used procedure

(enrichment of genomic DNA in microsatellite motifs, cloning and sequencing of the enriched DNA library by the Sanger method) is difficult, time-consuming and costly. Enrichment methods generally use a few specific repeated motifs, generally selected without prior knowledge of their abundance in the genome (Castoe *et al.* 2010), hence introducing potential bias in genome representativeness. Furthermore, only a few hundred sequences are generally obtained because of the high cost of Sanger sequencing. Next-generation sequencing through 454 GS-FLX technology (Roche Applied Science)

Correspondence: Thibaut Malausa, Fax: +33 492386401;  
E-mail: thibaut.malausa@sophia.inra.fr

has opened up new opportunities for microsatellite isolation. First, large shotgun genomic libraries have proved sufficient to isolate a satisfactory number of markers in a few studies (Abdelkrim *et al.* 2009; Allentoft *et al.* 2009; Castoe *et al.* 2010). Second, 454 GS-FLX sequencing can be used to sequence enriched library, thus offering a higher cost-efficiency (Santana *et al.* 2009). However, the use of pyrosequencing applied to microsatellite isolation has remained rare. This situation could evolve quickly if new procedures taking profit of pyrosequencing technology updates (sequencing longer and more numerous DNA fragments) and easily accessible for research teams could be set up. We present here a new method for high-throughput microsatellite isolation combining DNA enrichment procedures with the use of multiplexed microsatellite probes and the update Titanium of the 454 GS-FLX technology. This method was initially developed in the model species *Apis mellifera* (Linnaeus, 1758) [Insecta: Hymenoptera: Apidae], a species with a genome particularly rich in microsatellites (Solignac *et al.* 2007). Its efficiency was subsequently assessed and validated with 13 other species from various taxonomic groups, including fungi, plants and animals. This procedure is now available on our platform (Lille, France) for any research team interested in rapid and low-cost development of wide SSR libraries.

## Methods

### *Construction and sequencing of multiplex-enriched libraries*

An optimization of classical biotin-enrichment methods (Kijas *et al.* 1994) was used to prepare the enriched libraries. Genomic DNA was extracted from various tissues (depending on the species), with the DNeasy Tissue Kit (QIAGEN) and the DNeasy Plant Mini Kit (QIAGEN). Enrichment was carried out at Genoscreen (Lille, France), according to the following procedure. Genomic DNA (1 µg) was sonicated or digested with *RsaI* (FERMENTAS) for 1 h at 37 °C, according to the manufacturer's recommendations, and was then ligated to standard adapters (Adap-F: GTTAAAGGCCTAGCTAGCAGAATC and Adap-R: GATTCTGCTAGCTAGGCCTT). This step was repeated until the average length of DNA fragment was <1500 bp. Samples were then purified on a Nucleofast PCR plate (MACHEREY-NAGEL). Eight biotin-labelled oligonucleotides, corresponding to eight targeted microsatellite motifs, were hybridized to the ligated DNA at 56 °C for 20 min, after initial denaturation of the ligated DNA. The choice of the targeted microsatellite motifs was based on the screening of thirteen published genome sequences or whole-genome

shotgun (WGS) sequences (insects: *A. mellifera*, *Anopheles gambiae*, *Drosophila melanogaster*, *D. yakuba*, *D. simulans*, *Bombyx mori*, *Tribolium castaneum*; Vertebrates: *Takifugu rubripes*, *Danio rerio*, *Gallus gallus*, *Bos taurus*, *Mus musculus* and *Rattus norvegicus*). All perfect microsatellites with at least five repetitions for all di-hexa motifs were extracted. We then identified the 12 most frequent motifs for each genome (Table S1, Supporting information). From this pool of 30 motifs, we selected eight. This selection was based on (i) motif frequencies in different genomes and (ii) melting temperature compatibility (56 °C). At the same time, we avoided using motifs likely to produce hairpin structures, even if highly frequent (e.g. AT, CG and AAT). The following eight probes were designed to enrich total DNA in these motifs: (AG)<sub>10</sub>, (AC)<sub>10</sub>, (AAC)<sub>8</sub>, (AGG)<sub>8</sub>, (ACG)<sub>8</sub>, (AAG)<sub>8</sub>, (ACAT)<sub>6</sub> and (ATCT)<sub>6</sub>.

The enrichment step was completed with Dynabeads (INVITROGEN). The resulting enriched DNA was amplified with primers corresponding to the library adapters, over 25 cycles (20 s at 95 °C, 20 s at 60 °C and 90 s at 72 °C) and a final extension step of 30 min at 72 °C. The PCR products were purified with a QIAquick PCR purification kit (QIAGEN).

The sample concentration of purified PCR products was determined by quantifying Picogreen fluorescence (Invitrogen), and the fragment size distribution was determined by running 1 µL of each sample on an Agilent Bioanalyzer 2100, using a DNA 7500 chip (Agilent Technologies). The following manufacturer's protocols were carried out at Genoscreen (Lille, France): fragment end polishing, adaptor ligation, during which specific multiplex identifiers (MIDs) were added, library immobilization, fill-in reaction and single-stranded DNA library isolation. The small fragment removal step was not included to avoid the loss of any genetic information. The single-strand DNA profile and quantification were determined by running 1 µL of each sample on an Agilent Bioanalyzer 2100 with a RNA Pico 6000 chip. The concentration (pg/µL) obtained was then used to calculate the number of molecules of the final product/µL: [single-strand DNA (pg/µL)]/[MW of nucleotide (325) × base pair length of DNA strand] × [6.02 × 10<sup>23</sup>]. The single-strand templates were subsequently diluted to a normalized concentration of 1 × 10<sup>8</sup> molecules/µL, and multiplexing by equimolar mixture was performed for the analysis of four samples on a 1/8 GsFLX PTP or eight samples on a 1/4 GsFLX PTP. In each GsFLX PTP region, samples were distinguished thanks to their MIDs. Each multiplex library was previously titrated to accurately determine the number of DNA copies per bead required for maximum sequencing quality. Emulsion PCR and sequencing were then carried out according to the GS-FLX protocol, with no modification.



### Calibration test

We set up a calibration test, using DNA samples of *A. mellifera*, to assess the minimum number of 454 loading beads required to obtain sequences of satisfactory quality in sufficient amounts. We loaded calibrated quantities of beads (125 000; 75 000; 50 000 and 25 000 beads) onto four regions of a GsFLX plate delimited by a 16-region gasket.

### Enriched vs. shotgun library

We evaluated the benefits of DNA enrichment in microsatellite motifs, by comparing the data obtained after sequencing DNA libraries of *A. mellifera* with and without enrichment. This comparison was made with 125 000 loading beads as it provided the maximum number of sequences in a given 1/16 GsFLX plate.

### Validation on 13 taxa

Enriched libraries, generated as described earlier, were constructed for 13 additional taxa from various taxonomic groups (Table 1) to assess the general value of the method. The loading of 75 000 beads was performed for each species (one species on a 1/16 GsFLX plate, four species on 1/8 plate or eight species on 1/4 plate—using Roche MIDs) as calibration tests revealed that this num-

ber provided highest quality and cost-efficiency (see Results).

### Data analysis and automated primer design

The QDD pipeline (Megléczy *et al.* 2010) was used to analyse the 454 sequences and design primers for amplification of the detected microsatellite motifs. Sequences were sorted according to their MID (when used), and the MID sequence was subsequently removed. Enrichment adaptors (Adap-F and Adap-R) were then removed from sequences, and sequences with no detected adapter were discarded. Sequences shorter than 80 bp and sequences containing microsatellite motifs shorter than five repeats were discarded. Sequence similarities were detected through an 'all against all' BLAST analysis. Sequences with significant BLAST hits (e-value = 1E-40, microsatellites being soft-masked) but with flanking region identity levels below 90% were discarded to avoid potential intragenomic multicopy sequences. Using BLAST allowed identification of flanking regions with low, but significant similarities. This is a conservative step that aims to eliminate repetitive sequences (e.g. minisatellites and transposable elements), which are unlikely to provide a clear amplification pattern. Sequences displaying only BLAST hits for which pairwise similarity between the complete

**Table 1** Name and systematic position of the species used to set and test the procedure; number and length of sequences generated, Accession no. in the NCBI Short Read Archives. Libraries were also submitted to the Dryad Database, doi:10.5061/dryad.8297 (<http://dx.doi.org/10.5061/dryad.8297>)

Name	Division	Class	Order	Number of sequences	Length		SRA accession
					Mean	Maximum	
<i>Apis mellifera</i> (shotgun)	Arthropoda	Insecta	Hymenoptera	37 870	275	765	SRS150264.1
<i>A. mellifera</i> (enriched-125K)	Arthropoda	Insecta	Hymenoptera	39 473	251	766	SRS150263.1
<i>A. mellifera</i> (enriched-75K)	Arthropoda	Insecta	Hymenoptera	26 428	258	681	SRS150262.1
<i>A. mellifera</i> (enriched-50K)	Arthropoda	Insecta	Hymenoptera	30 041	259	610	SRS150261.1
<i>A. mellifera</i> (enriched-25K)	Arthropoda	Insecta	Hymenoptera	11 571	259	639	SRS150260.1
<i>Venturia canescens</i>	Arthropoda	Insecta	Hymenoptera	21 716	189	539	SRS140293.2
<i>Euphydryas aurinia</i>	Arthropoda	Insecta	Lepidoptera	11 497	184	562	SRS150273.1
<i>Pseudococcus viburni</i>	Arthropoda	Insecta	Hemiptera	29 528	237	611	SRS150265.1
<i>Diabrotica virgifera</i>	Arthropoda	Insecta	Coleoptera	15 207	259	595	SRS140297.2
<i>Bursaphelenchus xylophilus</i>	Nematoda	Secernentea	Aphelenchida	12 286	240	615	SRS140294.2
<i>Barbus meridionalis</i>	Chordata	Actinopterygii	Cypriniformes	13 010	150	477	SRS150271.1
<i>Danio rerio</i>	Chordata	Actinopterygii	Cypriniformes	15 833	193	543	SRS150272.1
<i>Gerbillus nigeriae</i>	Chordata	Mammalia	Rodentia	21 740	153	884	SRS150274.1
<i>Armillaria ostoyae</i>	Basidiomycota	Agaricomycetes	Agaricales	32 488	179	809	SRS150270.1
<i>Phytophthora alni</i> subsp. <i>uniformis</i>	Heterokontophyta	Oomycetes	Peronosporales	34 483	209	997	SRS150269.1
<i>Festuca eskia</i>	Magnoliophyta	Liliopsida	Poales	25 577	246	598	SRS150267.1
<i>Hirtella physophora</i>	Magnoliophyta	Eudicotyledones	Malpighiales	34 316	197	849	SRS150266.1
<i>Papaver rhoeas</i>	Magnoliophyta	Eudicotyledones	Ranunculales	13 825	240	570	SRS150268.1

overlapping part of the flanking regions was over 90% were grouped into contigs aligned by ClustalW, and a 2/3 majority rule was used to build a consensus sequence. These consensus sequences substituted the corresponding single reads. Unique sequences (with no BLAST hit according to our criteria) and consensus sequences were used to form a validated set of sequences that was used for further analyses. Primers were designed automatically using the Primer3 algorithm (Rozen & Skaletsky 2000) implemented within QDD. PCR primers were designed only if (i) the target microsatellite had at least five repeats, (ii) the resulting PCR product was between 80 and 500 bp long, (iii) the flanking region contained, at most, a five-base mononucleotide stretch or two repeats of any di-hexa base-pair motif, (iv) the annealing temperature of primers was between 50 and 64 °C, and the difference in annealing temperature between the forward and the reverse primer was <4 °C and (v) the self-complementarities of primers and the complementarities between primers matched the quality criteria used as default parameters in Primer3. We deliberately chose stringent criteria, as the number of microsatellite loci identified through pyrosequencing is large and the most time-consuming and extensive step is the subsequent validation of the designed primers.

## Results

### *Calibration test*

The calibration tests revealed strong positive correlation between the numbers of beads and sequences ( $R^2=0.851$ ,  $P = 0.015$  through permutation test), and although the quality (as inferred by the percentage of sequences that passed the Roche 454 GsFLX Titanium default quality filters) was not homogeneous among the numbers of loading beads ( $\chi^2 = 142.29$ ,  $P < 0.0001$ ), there was no significant correlation between beads number and quality ( $R^2=0.053$ ,  $P > 0.05$ ). Based on this analysis, the loading of 75 000 beads was retained for the final protocol because this number of beads (i) provided sufficient number of sequences (26 428 sequences) with a high-quality index (63.35% of the sequences passed the 454 GsFLX Titanium quality filters) and (ii) allows the sequencing of up to four enriched libraries onto a 1/8 GsFLX plate or eight libraries onto a 1/4 GsFLX plate, using MIDs.

### *Enriched vs. shotgun library*

We obtained 37 870 and 39 473 sequences for the *A. mellifera* libraries without and with enrichment (hereafter referred to as the 'shotgun' and 'enriched' libraries), respectively (Table 1). Short sequences and sequences without microsatellite motifs were discarded, and the

remaining sequences were filtered for redundancy and checked for multiple copies in the data set with our bioinformatics pipeline QDD (Megléczy *et al.* 2010). In total, 5157 and 6230 loci containing microsatellites matching the quality criteria implemented in QDD were detected in the shotgun and enriched libraries, respectively. This high number of loci detected in the shotgun library is not surprising, given the high quantities of microsatellites contained in the genome of this organism (Solignac *et al.* 2007). With the shotgun method, 97% of the validated loci corresponded to sequences observed once in the raw data set, whereas the remaining 3% of the validated loci corresponded to consensus sequences (Table 2). With the enrichment method, 22% of the validated loci corresponded to consensus sequences, taking into account variation between sequences (caused by intraspecific polymorphism and polymerase or 454 sequencing errors). QDD was then used to design primer pairs for each validated locus. The primer design was successful in 2045 loci from the shotgun library and in 2200 loci from the enriched library. The properties of the markers designed from the two libraries differed considerably: the percentage of markers displaying microsatellite motifs consisting exclusively of A/T nucleotides was 39% with the shotgun library and only 8% with the enriched library (Table 2). Interestingly, enrichment allowed isolation of around 3 times more primer pairs (543 vs. 186) designed around microsatellite motifs of more than eight perfect repeats (excluding AT-motifs), i.e. optimal microsatellite markers.

### *Validation on 13 taxa*

The data sets obtained after the validation tests each contained 11 497–34 483 sequences depending on the species (Table 1). Mean sequence length was between 150 and 275 bp (Table 1), and maximum sequence length was between 477 and 997 bp (Table 1). The number of validated loci containing microsatellites ranged from 199 to 5791 (Table 2). We found that 3–28% of the validated loci corresponded to consensus sequences (Table 2), providing information about the polymorphism of sequences. Despite the heavy constraints imposed on primer design, primers were successfully designed for 94–1162 loci, even for species for which severe difficulties were previously encountered using traditional methods (Megléczy *et al.* 2004; Péténian *et al.* 2005; Dutech *et al.* 2007). The primers designed targeted various microsatellite motifs in each species, with the respective proportions of each motif differing among species (Table 2). Both the contrasted numbers and proportions of motifs found in the different organisms are not surprising with regard to the available literature (Lagercrantz *et al.* 1993; Toth *et al.* 2000; Megléczy *et al.* 2007; Richard *et al.* 2008).

**Table 2** Description of the microsatellite libraries produced for the 14 tested species: number of microsatellite loci validated by the QDD program; number of perfect (only one repeated motif) and compound (2 or more repeated motifs or interrupted microsatellites) microsatellite motifs among the loci for which PCR primers were designed by QDD; frequencies of the various types of perfect microsatellites. Data are obtained from 454 runs in which 75 000 beads were loaded for each species, except *Apis mellifera* shotgun, enriched-125K, enriched-75K, enriched-50K and enriched-25K, for which 125 000, 125 000, 75 000, 50 000 and 25 000 beads were loaded, respectively

Name	Validated loci	% loci identified from several sequences	Design of primers		Motif type of perfect ms											Others (%)
			Perfect ms	Compound	AC (%)	AG (%)	AAC (%)	AAG (%)	ACG (%)	AGG (%)	ACAT (%)	ATCT (%)	AT-based (%)			
<i>A. mellifera</i> (shotgun)	5157	3	1451	594	3	35	1	5	2	2	6	0	0	0	39	10
<i>A. mellifera</i> (enriched-125K)	6230	22	1516	684	5	58	3	12	1	1	7	0	0	0	8	6
<i>A. mellifera</i> (enriched-75K)	4674	21	1207	563	5	54	2	13	1	1	9	0	0	10	5	5
<i>A. mellifera</i> (enriched-50K)	5025	21	1203	540	5	56	3	12	1	1	8	0	0	9	5	5
<i>A. mellifera</i> (enriched-25K)	2519	18	656	290	5	55	3	12	1	1	8	0	0	11	5	5
<i>Venturia canescens</i>	2404	23	483	192	15	62	5	9	2	2	2	0	0	2	3	3
<i>Euphydryas aurinia</i>	1627	10	252	96	41	17	7	11	0	0	0	6	4	6	8	8
<i>Pseudococcus viburni</i>	1311	23	315	136	16	15	14	7	6	6	18	2	1	3	19	19
<i>Diabrotica virgifera</i>	1731	5	319	109	14	11	1	60	1	3	3	2	2	2	5	5
<i>Bursaphelenchus xylophilus</i>	199	28	71	23	20	45	11	13	0	0	3	0	0	0	8	8
<i>Barbus meridionalis</i>	2562	11	572	229	80	15	0	2	0	0	0	0	1	1	1	1
<i>Danio rerio</i>	5791	8	770	322	73	8	5	1	0	0	1	0	1	1	11	11
<i>Gerbillus nigeriae</i>	2048	3	163	78	36	19	10	1	0	0	3	3	9	2	17	17
<i>Armilaria ostoyae</i>	4015	6	235	69	20	10	22	10	4	6	6	0	1	6	20	20
<i>Phytolthora alni</i>	550	16	186	53	25	23	10	10	4	8	8	1	1	3	17	17
<i>Festuca eskia</i>	1474	8	475	151	31	35	8	10	1	6	6	0	1	4	4	4
<i>Hirtella physophora</i>	3552	24	809	353	16	24	9	18	2	11	11	1	2	6	10	10
<i>Papaver rhoas</i>	1072	4	334	120	11	24	13	43	0	4	4	0	0	1	4	4

## Discussion

Even with the stringent selection of loci imposed by the analysis parameters chosen, the numbers of microsatellite markers isolated with this new procedure were satisfactory for all the tested species. When compared to traditional isolation techniques, these numbers are much larger and were obtained with a much lower budget and within a shorter length of time. Indeed, whereas the isolation of microsatellites from an enriched library typically involves the screening and Sanger sequencing of a couple of hundred clones over a period of 5 weeks at a cost of more than US\$ 5000, the process tested here allows the simultaneous screening of thousands of DNA fragments, over a 2-week period, at a cost of less than US\$ 1500 (including expenses related to consumables and staff). The sequences produced are shorter than those obtained by Sanger sequencing, but this is not a problem for microsatellite marker development because the PCR products used for genotyping are usually 100–400 bp in size. Besides, this read-length constraint should be overcome with subsequent updates of 454 platforms. Comparing our results to the first studies using pyrosequencing to isolate microsatellites is not straightforward because (i) the way microsatellites were defined and the analysis methods used differ substantially, (ii) the organisms used to test the protocols are different and (iii) the total cost to generate a given amount of exploitable microsatellite sequences was generally not provided. However, the efficiency as evaluated from our results (between 1% and 8% of amplifiable markers in the obtained sequences; see methods for the definition of amplifiable markers) appears higher than in Abdelkrim *et al.* (2009) who reported around 0.1% of amplifiable microsatellite markers in their shotgun library and of the same order of magnitude as in Castoe *et al.* 2010 (~3.5% of amplifiable microsatellites in one species) and Santana *et al.* 2009 (between ~2% and ~5%).

Our tests revealed that the use of the multiplex DNA enrichment of our procedure appears highly advisable to optimize cost-efficiency of microsatellite isolation. Comparisons between enriched and shotgun *A. mellifera* libraries revealed two major advantages of enrichment. First, enrichment slightly increased the overall number of microsatellite loci isolated and reduced the proportion of unwanted motifs such as AT-based motifs (39–8%) that are likely to generate difficult amplification during genotyping. Counting the numbers of amplifiable optimal microsatellite markers ( $\geq 8$  repetitions of perfect and not AT-rich motifs, i.e. the most valuable markers for the end-users) revealed that enrichment improved marker

isolation efficiency by almost 300% (543 vs. 186 markers isolated in *A. mellifera* with and without enrichment) for an additional cost of around 10%. Moreover, the benefits of enrichment are likely much underestimated here because *A. mellifera* genome is rich in microsatellites (Solignac *et al.* 2007). Second, it increases the number of multiple reads obtained for a given microsatellite locus, which enables to design PCR primers targeting nonpolymorphic sequences flanking the microsatellite motif. Such an approach should decrease the probability of designing markers with a high percentage of null alleles because of mismatches between primers and polymorphic nucleotides in flanking regions that can occur in some individuals or populations. The analysis of the large data sets obtained, using programs like QDD, also enables to sort and discard loci that are likely to be found at multiple sites in the genome.

The validation of our method on 13 taxa did not reveal negative effects of enrichment on microsatellite isolation efficiency. Comparison of motif frequencies in the libraries generated here and in published genome sequences of closely related taxa (when available) shows that enrichment successfully favoured the most common motifs (Tables 2 and S1, Supporting information). Second, results do not indicate that lower isolation efficiency could be because of enrichment failure: we found no negative correlation between the percentage of unwanted motifs and the number of detected microsatellite loci ( $r = 0.16$ ;  $P > 0.05$ ). Lower yields in some species may rather be caused by technical issues (e.g. random manipulation effects or lower DNA quality of available samples) or lower abundances of microsatellites in genomes.

In conclusion, we demonstrated that our procedure coupling multiplex enrichment, pyrosequencing and sequence selection can be readily and successfully applied to a large variety of taxonomic groups (Table 2). The procedure is particularly cost-efficient as only a small part of a 454 GsFLX plate is needed to isolate high numbers of microsatellite markers. It is expected to speed up the acquisition of high-quality genetic markers for nonmodel organisms.

## Acknowledgements

We thank M. Galan for useful comments on previous versions of the manuscript and S. Nielsen and J. Sappa for major improvements to English grammar throughout the text. This work was supported by the AIP BioRessources 'EcoMicro' grant from the French *Institut National de la Recherche Agronomique* (INRA), permanent institutional support from Montpellier SupAgro, University Aix-Marseille I, INRA and the R&D budget of Genoscreen (Lille, France).

## References

- Abdelkrim J, Robertson BC, Stanton JAL, Gemmell NJ (2009) Fast, cost-effective development of species-specific microsatellite markers by genomic sequencing. *BioTechniques*, **46**, 185–191.
- Allentoft ME, Schuster SC, Holdaway RN *et al.* (2009) Identification of microsatellites from an extinct moa species using high-throughput (454) sequence data. *BioTechniques*, **46**, 195–200.
- Castoe TA, Poole AW, Gu W *et al.* (2010) Rapid identification of thousands of copperhead snake (*Agkistrodon contortrix*) microsatellite loci from modest amounts of 454 shotgun genome sequence. *Molecular Ecology Resources*, **10**, 341–347.
- Dutech C, Enjalbert J, Fournier E *et al.* (2007) Challenges of microsatellite isolation in fungi. *Fungal Genetics and Biology*, **44**, 933–949.
- Kijas JMH, Fowler JCS, Garbett CA, Thomas MR (1994) Enrichment of microsatellites from the *Citrus* genome using biotinylated oligonucleotide sequences bound to streptavidin-coated magnetic particles. *BioTechniques*, **16**, 656–662.
- Lagercrantz U, Ellegren H, Andersson L (1993) The abundance of various polymorphic microsatellite motifs differs between plants and vertebrates. *Nucleic Acids Research*, **21**, 1111–1115.
- Megléc E, Petenian F, Danchin E *et al.* (2004) High similarity between flanking regions of different microsatellites detected within each of two species of Lepidoptera: *Parnassius apollo* and *Euphydryas aurinia*. *Molecular Ecology*, **13**, 1693–1700.
- Megléc E, Anderson SJ, Bourguet D *et al.* (2007) Microsatellite flanking region similarities among different loci within insect species. *Insect Molecular Biology*, **16**, 175–185.
- Megléc E, Costedoat C, Dubut V *et al.* (2010) QDD: a user-friendly program to select microsatellite markers and design primers from large sequencing projects. *Bioinformatics*, **26**, 403–404.
- Péténian F, Megléc E, Genson G, Rasplus JY, Faure E (2005) Isolation and characterization of polymorphic microsatellites in *Parnassius apollo* and *Euphydryas aurinia* (Lepidoptera). *Molecular Ecology Notes*, **5**, 243–245.
- Richard GF, Kerrest A, Dujon B (2008) Comparative genomics and molecular dynamics of DNA repeats in Eukaryotes. *Microbiology and Molecular Biology Reviews*, **72**, 686–727.
- Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods in Molecular Biology*, **132**, 365–386.
- Santana QC, Coetzee MPA, Steenkamp ET *et al.* (2009) Microsatellite discovery by deep sequencing of enriched genomic libraries. *BioTechniques*, **46**, 217–223.
- Solignac M, Mougel F, Vautrin D, Monnerot M, Cornuet J-M (2007) A third-generation microsatellite-based linkage map of the honey bee, *Apis mellifera*, and its comparison with the sequence-based physical map. *Genome Biology*, **8**, R66.
- Sunnucks P (2000) Efficient genetic markers for population biology. *Trends in Ecology and Evolution*, **15**, 199–203.
- Toth G, Gaspari Z, Jurka J (2000) Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Research*, **10**, 967–981.

## Supporting Information

Additional supporting information may be found in the online version of this article.

**Table S1** Proportions of microsatellite motifs in selected genomes.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.



RESEARCH ARTICLE

Open Access

# Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing

André Gilles<sup>1†</sup>, Emese Megléc<sup>1†</sup>, Nicolas Pech<sup>1†</sup>, Stéphanie Ferreira<sup>2</sup>, Thibaut Malausa<sup>3</sup> and Jean-François Martin<sup>4\*</sup>

## Abstract

**Background:** The rapid evolution of 454 GS-FLX sequencing technology has not been accompanied by a reassessment of the quality and accuracy of the sequences obtained. Current strategies for decision-making and error-correction are based on an initial analysis by Huse *et al.* in 2007, for the older GS20 system based on experimental sequences. We analyze here the quality of 454 sequencing data and identify factors playing a role in sequencing error, through the use of an extensive dataset for Roche control DNA fragments.

**Results:** We obtained a mean error rate for 454 sequences of 1.07%. More importantly, the error rate is not randomly distributed; it occasionally rose to more than 50% in certain positions, and its distribution was linked to several experimental variables. The main factors related to error are the presence of homopolymers, position in the sequence, size of the sequence and spatial localization in PT plates for insertion and deletion errors. These factors can be described by considering seven variables. No single variable can account for the error rate distribution, but most of the variation is explained by the combination of all seven variables.

**Conclusions:** The pattern identified here calls for the use of internal controls and error-correcting base callers, to correct for errors, when available (e.g. when sequencing amplicons). For shotgun libraries, the use of both sequencing primers and deep coverage, combined with the use of random sequencing primer sites should partly compensate for even high error rates, although it may prove more difficult than previous thought to distinguish between low-frequency alleles and errors.

## Background

Scientific strategies and approaches based on next-generation sequencing (NGS) have been revolutionizing genetics over the last few years. Many aspects of basic, applied and clinical research now rely on the generation of enormous amounts of sequence data from various sample sources, to assess polymorphism (mostly SNPs), or expression data (RNA-Seq) at the genome level [1,2]. This shift in the scale of sequence acquisition has been achieved by simultaneous progress in bioinformatics, the availability of genome assemblies and key technical findings in the domains of biochemistry and sequencing device physics [3]. In this context, the 454 GS-FLX (Roche Diagnostics Corporation), Illumina<sup>®</sup> technology (Illumina, Inc.) and SOLiDTM systems (Applied

Biosystems<sup>TM</sup>) offer a number of complementary solutions for specific requirements (see Metzker [4] for a review). 454 GS-FLX Titanium technology provides around 1,000,000 sequences in a single 10-hour run. These sequences, with an average read length equal to 330 bp, may be up to 500 bp in shotgun libraries conditions, much longer than can be obtained with the other available approaches. This makes mapping easier, particularly for repetitive regions, and facilitates *de novo* genome sequencing, exome capture, metagenomics and amplicon sequencing [4].

One of the basic questions arising from this spectacular increase in sequence volume concerns the possible detrimental effects of this shift in quantity on the quality of the obtained data. In other words, is there a tradeoff between the quantity and quality of information? It is widely accepted that next-generation sequencing approaches generate such large amounts of sequence data that even if overall accuracy (derived from error rate) or quality (percentage of error-free sequences) is

\* Correspondence: martinjf@supagro.inra.fr

† Contributed equally

<sup>4</sup>UMR CBGP (INRA/IRD/Cirad/Montpellier SupAgro), Campus international de Baillarguet, CS 30016, F-34988 Montferrier-sur-Lez cedex, France  
Full list of author information is available at the end of the article

suboptimal it is still possible to reconstruct polymorphism rigorously by comparing redundant sequences that cover the same genomic region multiple times (i.e. depth of coverage provides accuracy, not the individual read) [5-8]. This is the typical “quick and dirty” view of NGS. This approach may sound reasonable, but it is based on assumptions such as low error rate and error randomness for the unambiguous detection of polymorphism. If this last assumption is challenged, even a low error rate has a huge impact on sequence analysis, as in cases of related allele detection, paralogous sequences or pseudogene identification. In these cases, the “quick and dirty” approach is inadequate, because consensus sequence calculation is accurate only if these three sources of sequence diversity are distinguishable from the error due to background noise [9].

In 2007, S. Huse and collaborators raised the question of the accuracy and quality of massively parallel pyrosequencing GS20 systems, performing an empirical analysis of the per-base error rate [10]. This was needed as “the quality score of a position is not a measure of a confidence that the correct base is called at that position, as with a traditional PHRED score. Instead, the GS20 quality score is a measure of confidence that the homopolymer length at a position is correct” [10,11]. They used V6 hypervariable region sequences from cloned microbial ribosomal DNA for this purpose. They concluded that the accuracy rate was 99.51%, on average, and that 82% of the sequences contained no error. They also demonstrated that 39% of the errors corresponded to homopolymer effects [10]. Finally, they detected no significant correlation between error and distance from the 5' end of the sequences for 101 positions. Surprisingly, despite changes in this technology over the last four years, the accuracy and quality of 454-based sequences has not been reevaluated and this previous study remains the basic reference used by the scientific community to account for error rate in 454 GS-FLX systems (181 articles citing this study at the time of writing). Over the same period, chemistry, acquisition devices (CCD cameras in particular) and quality filtering algorithms have evolved. A new analysis is therefore required, and this was the main goal of this work.

Furthermore, in addition to estimating the per-base error rate, we aimed to identify the potential causes of sequencing errors and possible solutions for improving both the accuracy and quality of pyrosequences. We selected several variables likely to affect sequencing errors directly or indirectly: (i) the position of the nucleotide base within the sequence (the beginning of the sequence may be more accurate than the end), (ii) the primary structure of the sequence, including, in particular, the presence of homopolymers, (iii) the length of the sequence generated (a sequence may be short due to

quality filtering, resulting from an accumulation of errors or the stochastic ending of polymerization), and (iv) the position of the bead carrying the sequence both within and between the regions on a PT plate (PicoTiterPlate) (edge effect), and between multiple PT plates. Our analyses are based on Roche test fragments. These are sequences used for GS-FLX Titanium diagnostics that are included in all runs, but not subjected to PCR amplification before sequencing. Thus with these fragments we estimate the sequencing error due to pyrosequencing. Huse et al. [10] found that the experimental sequences they used display error rate five times higher than the GS20 Roche test fragments (0.1% vs. 0.49%). Since almost all of their results are based on experimental sequences, we cannot directly compare our results to theirs. However, we do not intend to focus on a general error rate, but rather assess the effect of several variables on error generation.

## Results and Discussion

### Accuracy and quality of sequences

We assessed the quality of the sequences obtained by 454 GS-FLX Titanium sequencing, using the control DNA fragment Type I sequences (provided with 454 sequencing kits) as reference templates (see Materials and Methods for details). As these internal controls are added to the pyrosequencing process during the sequencing step, they are modified only by sequencing errors and are not related to any previous step. The quality of these control sequences is not influenced by the samples themselves, particularly with Titanium technology, in which loading beads are isolated from each other and there should therefore be no interference from adjacent beads. We analyze here the 86,237 sequences that passed the quality filters, representing the six control DNA fragments from three 454 GS-FLX runs. These results revealed several general trends in the sequencing error generated by 454 GS-FLX Titanium technology (Table 1). It also provided detailed information about the different types of error: insertion, deletion, mismatches and ambiguous base calls. We first analyzed the error on the first 101 sequenced positions from the 5' end (with reference to the sequencing primer) of the control DNA fragments. We compared the sequences obtained with those for the GS20 system and then extended the error analysis to full-length sequences (500 to 592 bases, depending on the reference sequence analyzed, see Materials and Methods for details).

The error rate for the first 101 sequenced positions (corresponding to 8,596,016 examined bases) displayed a mean = 0.534% (95% CI: [0.529, 0.539]) (45,895 erroneous bases) for 454 GS-FLX Titanium data. This global error rate is five times higher than the error rate obtained by the analyses of GS20 test fragments and is



**Table 1 Comparative analysis of the accuracy and quality of sequences**

	# of sequences	% of error-free sequences	# of positions	Insertions	Deletions	Mismatch	Ambiguous	Total % of error
GS20 (101)	34015	82.00%	32801429	0.18%	0.13%	0.08%	0.10%	0.49%
Ref 1 (101)	16052	87.12%	1605640	0.15%	0.05%	0.01%	0.01%	0.22%
Ref 2 (101)	16466	60.01%	1600327	0.42%	0.23%	0.04%	0.01%	0.70%
Ref 3 (101)	12215	72.96%	1228804	0.17%	0.19%	0.01%	0.01%	0.38%
Ref 4 (101)	9908	56.43%	984452	0.30%	0.37%	0.03%	0.00%	0.70%
Ref 5 (101)	15880	50.93%	1595718	0.34%	0.48%	0.05%	0.01%	0.88%
Ref 6 (101)	15716	75.17%	1581075	0.25%	0.10%	0.00%	0.01%	0.36%
Total	86237	67.57%	8596016	0.27%	0.23%	0.02%	0.01%	0.53%
Ref 1 (572)	16052	6.75%	5359696	0.52%	0.46%	0.10%	0.12%	1.20%
Ref 2 (552)	16466	9.75%	4789285	0.89%	0.28%	0.10%	0.08%	1.35%
Ref 3 (500)	12215	18.75%	4180478	0.30%	0.35%	0.07%	0.12%	0.84%
Ref 4 (532)	9908	6.88%	2572843	0.56%	0.71%	0.19%	0.11%	1.57%
Ref 5 (592)	15880	7.46%	6171098	0.38%	0.38%	0.06%	0.07%	0.89%
Ref 6 (516)	15716	11.81%	6027338	0.60%	0.17%	0.07%	0.04%	0.88%
Total	86237	10.09%	29100738	0.54%	0.36%	0.09%	0.09%	1.07%

The different types of error are detailed for each reference sequence for 454 sequencing. Errors are classified according to the nomenclature used by Huse *et al.* (2007): insertions, deletions, mismatches and ambiguous base calls (see materials and methods). Error rates are given for two length categories (first 101 bases vs. full length).

similar to that obtained from for GS20 experimental sequences. Indeed, 0.49% of the positions were erroneous for a comparable dataset relating to 101 positions (Table 1). If we break down the global error rate for all reference sequences according to the type of error, insertions are found to be the most common errors (mean = 0.273% [0.269, 0.276]; mode  $q_{1/2}$  = 0.215), followed by deletions (0.232% [0.229, 0.235];  $q_{1/2}$  = 0.170), mismatches (0.022% [0.021, 0.023];  $q_{1/2}$  = 0.010), and ambiguous base calls (0.007% [0.006, 0.007];  $q_{1/2}$  = 0.010). This pattern is entirely consistent with that described by Huse *et al.* [10]. This pattern is in agreement with the study of 454 GS-FLX [12] but markedly different from IlluminaTM sequencing, in which insertions and deletions of single bases occur less frequently than mismatches [13,14]. In total, 58,269 sequences (67.57% [67.26, 67.88]) of this length were found to be free from error. This trend is similar to that reported for GS20 experimental sequences, for which 82% of sequences matched the corresponding reference sequence perfectly. Unfortunately the data are not available for GS20 test fragments.

If we restricted the analysis to full-length sequences (500 to 592 positions), we found for the 86,237 sequences that passed the 454 quality filters (29,100,738 bases) that 312,351 bases were erroneous (1.073% [1.069, 1.077]). The pattern observed for the first 101 positions was confirmed for the full-length sequence data, with insertions (0.541% [0.538, 0.543];  $q_{1/2}$  = 0.465) and deletions (0.359% [0.357, 0.362];  $q_{1/2}$  = 0.350) being the most common types of error and mismatches (0.088% [0.087, 0.089];  $q_{1/2}$  = 0.085) and ambiguous base calls (0.085% [0.084, 0.086];  $q_{1/2}$  = 0.090) making a

smaller contribution to global error rate. Only 8,702 of the 86,237 full-length sequences (10.09% [9.89, 10.29]) had no error with respect to the corresponding reference sequence. This result strongly contrasts with the higher proportion of error-free sequences for the first 101 bases.

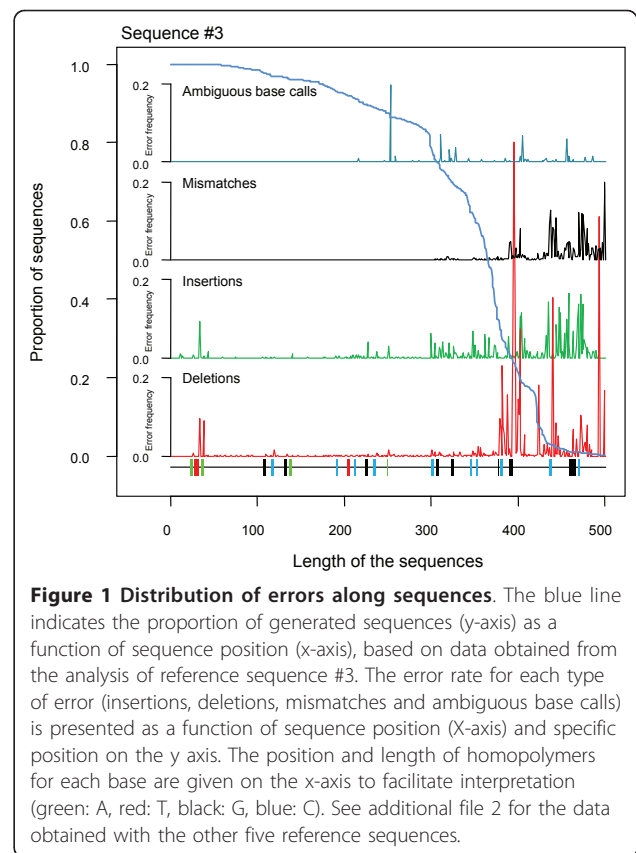
The comparison of error rates between sequences of different lengths (first 101 positions vs full-length sequences) highlighted two key developments in addition to the doubling of the global error rate for full-length sequences as found elsewhere [15]. This length-associated overall increase in error rates did not reflect a common mechanism for all types of error, as insertion and deletion rates increased only slightly (by factors of 2 and 1.5, respectively), whereas mismatch and ambiguous base call rates increased to a much greater extent (by factors of 5 and 9, respectively). This decoupling of the changes in rate for different types of errors modified the contribution to global error rate of the various types of errors. Thus, mismatch and ambiguous base call errors made a greater contribution to global error rate for longer sequences, although their effects remained moderate. Thus, overall error rates and the rates of different types of error are not uniform for the sequences obtained by 454 GS-FLX Titanium sequencing. Consequently, the conclusions drawn for short sequences should not be directly extrapolated to longer sequences, as sequence length affects error rates. Another key result in this in-depth analysis of error was the finding that error rate (1.07%) should be seen in the light of the large number of erroneous sequences (89.91%) in the dataset. This combination of a low error rate and a large number of erroneous sequences results from the

occurrence of only very small numbers of errors in individual sequences, on average. These findings conflict with those reported for GS20 sequencing and suggest that the removal of erroneous sequences may not be useful, to increase the overall quality.

However, the consequences of this may be relatively minor even if most sequences display errors (89.91 [89.71, 90.11]), as the overall error rate is low, with only 1.07% of bases being problematic. It is widely believed that deep sequencing coverage (multiple independent sequences for the same locus) should make it possible to correct for errors in this context [16]. Like other types of high-throughput sequencing, 454 pyrosequencing is thought to be suitable for use in this context. Indeed, for some applications, such as SNP discovery in whole-genome sequences [17,18] or amplicon sequencing [7,19], an almost unlimited number of sequences may be obtained. We need to consider the number of sequences required to correct an erroneous position appropriate, at a given probability, for an error rate of 1.07%. As detailed in additional file 1, at least five sequences would be required to correct for random error at low error rates (<10% error rate). However, an analysis of error along the length of the sequence (comparing the first 101 bases with the full-length sequence) indicated that error rate was heterogeneous along the length of the sequence. Longer sequences therefore would be subject to higher error rates at their 3' ends. The distribution of error, as illustrated in Figure 1, does not fit a stochastic model, for any error type. Most of the positions are correct, but a few have high error rates, even exceeding 50% in some cases. There is no clear way to resolve the issue, particularly when this pattern (error hot spots) is repeatable between runs [9].

This pattern is particularly problematic for 454 data, as the number of sequences significantly decreases after 300 bases (see Figure 1 and additional file 2 for illustration) whatever the reference sequence considered (for a total length ranging from 500 to 592). In summary, for the longest sequences (>300 positions), the combination of higher error rates along the length of the sequence, combined with the decrease in the number of sequences available, may make it difficult to correct errors. This difficulty results from a deficit in the number of sequences required to decrease the probability of erroneous assignment for a given sequence position, under a reasonable coverage threshold (i.e. minimum number of reads per bp required, see additional file 1).

This issue is further complicated by the heterogeneous distribution of the error types among the six different control DNA reference sequences, within and between gasket regions for a PT GS-FLX Titanium plate and also between PT plates, as initially estimated from the large standard errors (derived from table 1) in the error



**Figure 1 Distribution of errors along sequences.** The blue line indicates the proportion of generated sequences (y-axis) as a function of sequence position (x-axis), based on data obtained from the analysis of reference sequence #3. The error rate for each type of error (insertions, deletions, mismatches and ambiguous base calls) is presented as a function of sequence position (X-axis) and specific position on the y axis. The position and length of homopolymers for each base are given on the x-axis to facilitate interpretation (green: A, red: T, black: G, blue: C). See additional file 2 for the data obtained with the other five reference sequences.

estimate. This overall variability of error distribution makes it difficult to draw any clear conclusions ruling out particular parameters that might potentially influence error rates or to identify a single mechanism accounting for the observed errors in the dataset. This pattern requires an in-depth analysis of the interaction and explanatory power of various factors before we can assess the degree of sequencing error and identify solutions for preventing artifacts.

#### Interactions between variables and error characterization

The evolution of 454 technology combines progress in chemistry, acquisition devices, such as CCD cameras and PT plates handling equipment, and improvements in quality filters and base-calling algorithms. All these modifications are potential sources of variation in the amount, length and quality of sequences. In this work, we analyzed the interaction of seven variables identified as potential sources of sequencing error. We characterized sequencing error as a function of information about position in the sequence (*Position* and *Seq.length*), the presence of homopolymers (*Homopolymer*) and reference sequence type (*Seq.type*), all considered being sequence-specific information. Location on the PT plate was also taken into account through the region of origin

(*Region*), the distance of beads to the region center (*Dist.region*) or the plate center (*Dist.plate*, see Materials and Methods for details) as both the flow of chemicals through the plate and the central position of the CCD camera may play a role in the error generation. Before this analysis, we tested the hypothesis of homogeneous error rates on the three PT plates. This hypothesis was rejected ( $\chi^2 = 2613.3$ ,  $df = 2$ ,  $P < 2 \times 10^{-16}$ ). The significant result obtained in this test is mostly due to the high power of detection associated with the large number of samples available, but this heterogeneity requires the specification of individual parameter values for the logistic model describing each PT plate. The three runs were therefore analyzed separately. This approach did not prevent us from extracting the common trends influencing error rate and distribution. The models (for each plate and for each type of error) explained between 14.32% and 37.38% of the error distribution and were highly significant ( $P < 2 \times 10^{-16}$ ).

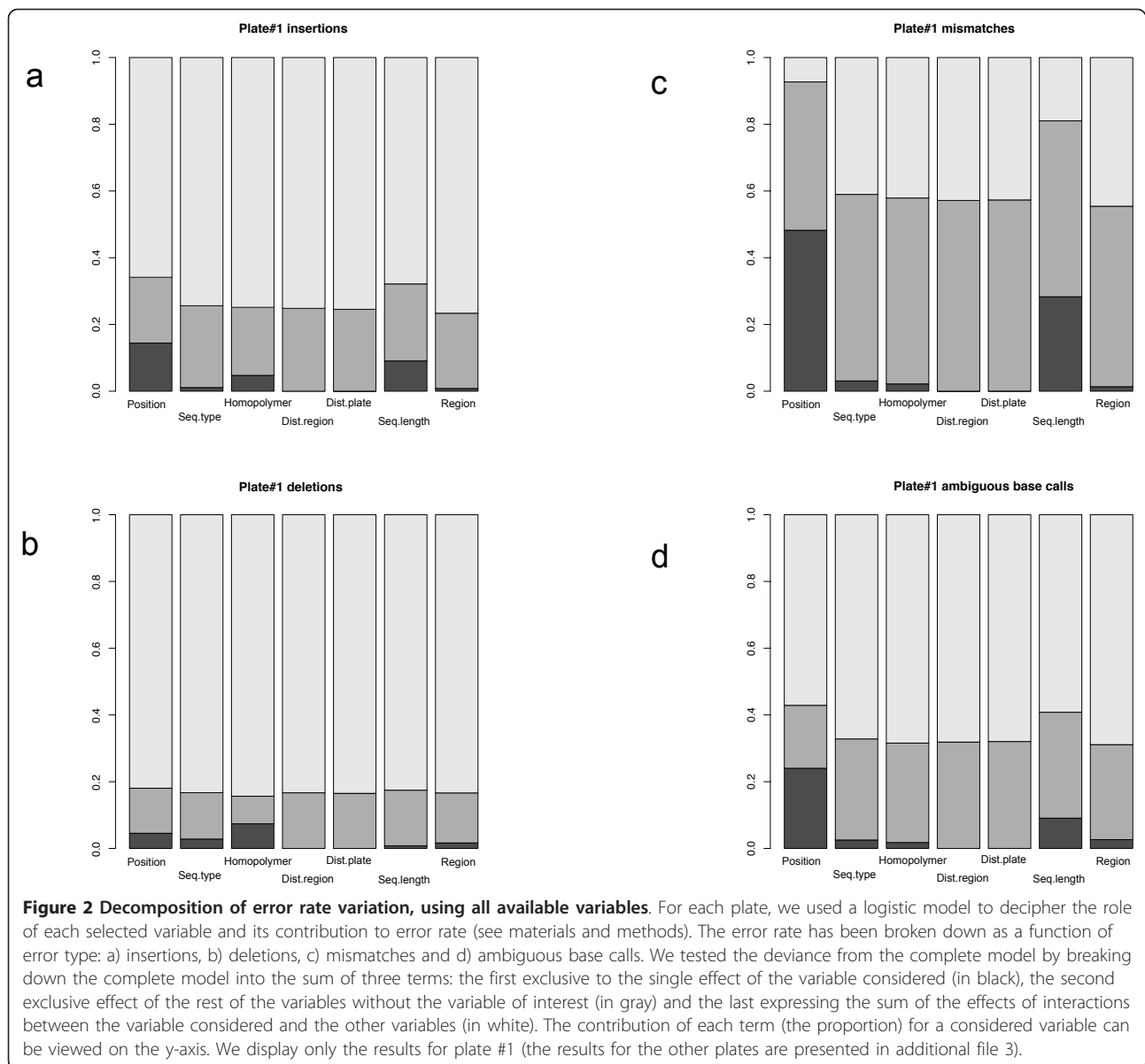
The nullity of  $r$  (Bravais-Pearson correlation coefficient) between pairs of the seven variables was tested independently for each run. As the usual assumptions required to infer the distribution of the test statistics were not met, we used permutations to approximate the distribution of the test statistic under  $H_0$ . We used a type I error rate of 0.05 and Benjamini-Hochberg correction [20] to take multiple testing into account. Most of the pairs of variables (74.29%) were significantly correlated, using a threshold of  $\alpha = 0.05$  in a permutation test for multiple testing. However 41.85% of the pairs of variables correlated with  $0.005 < r < 0.05$ . The pair of variables displaying the strongest correlation was the position of the error in the reference sequence (*position*) and sequence length (*Seq.length*), with  $0.40 < r < 0.50$ , depending on the PT plate considered. The second strongest correlation was that between distance to the region center (*Dist.region*) and distance to the PT plate center (*Dist.plate*), with  $0.38 < r < 0.62$ .

The nature and significance of a correlation between two variables does not provide any information about the ability of this combination of variables to explain a third variable [21]. For each plate and each kind of error, we considered a logistic model [22] (see materials and methods for the detailed procedure) accounting for the binary (error) variable in terms of the seven variables considered. For the separation of the effect of a given explicative variable from the combined effect of the other variables, we propose (see materials and methods) breaking down each explanatory variable into three additive terms: the effect of the variable itself, the combined effect of the other variables and the rest. The combined effect of the variables ranged from 20% to 80% of the total variation in error rate (Figure 2 and additional file 3). More specifically, for individual error

types, the combined effect accounted for  $38.00\% \pm 13.05$  of the total information for mismatch errors,  $64.10\% \pm 4.54$  for ambiguous base call errors,  $75.83\% \pm 3.78$  for insertion errors and  $79.95\% \pm 3.08$  for deletion errors. The remaining information results from the specific effects of each variable. These high percentages of shared information highlight the high degree to which the error can be explained by combinations of variables. This may be due to partial redundancy of the information contained in each variable or the combined contribution to the total amount of error explained [21]. In the first case, a variable may substitute for the effect of others, whereas, in the second, only the combined information provided by each variable can account for the observed pattern. The results of correlation analysis, indicating that most regression coefficients were low, ruled out redundancy as the primary cause of the observed pattern, as most variables were independent. There is therefore no single variable consistently accounting for the distribution of sequencing error, as detailed in Figure 2. We investigated the main trends highlighted by the logistic model, by focusing on the distribution of sequencing error at sequence level. We then characterized the variables most strongly influencing error in terms of the location of the bead carrying the sequence, in a given region of a PT plate.

At DNA sequence level, we detailed the variables individually accounting for the highest proportion of the error rate for each error type. It was essential to bear in mind, during this analysis, the fact that most of the explanatory power of these variables was obtained with combinations of variables. We analyzed each type of error independently.

For insertion errors (Figure 2), the variable *Homopolymer* accounted for  $5.97\% \pm 1.33$  of the variation in error on its own, and was concurrent to the error rate. This finding is consistent with available published empirical observations linking errors to homopolymers [9]. The variable *Position* accounted for  $11.94\% \pm 2.22$  of the variation and was also concurrent to error. In other words, the error rate due to insertions increased along the sequence. Finally, the variable *Seq.length* accounted for  $5.48\% \pm 3.13$  of the variation. Insertion rates were lower for longer sequences and higher for shorter sequences. These last two results may appear paradoxical, but the combined information for these variables indicates that the distribution of insertion errors along sequences is not random, with more insertions in 3' end, whatever the length of the sequence considered. This is fully explained if we considered that i) the number of sequences decreases with length (Figure 1), hence changing the number of sequences for which error rates are computed with respect to the reference and ii) the quality filtering process (v2.3) implemented in the GS-FLX system involves the trimming of reads



with many off-peak signal intensities by the software. In particular, for insertions, the TrimBack Valley Filter trims sequences from the 3' end until the number of valley flows (intermediate signal intensity, i.e., a signal intensity occurring in the valley between the peaks for 1-mer and 2-mer incorporations, or 2-mer and 3-mer, etc.) is  $< 1.25\%$  [23]. This implies that short sequences are not short because the strand synthesis stops prematurely, but due to a rapid decrease in the quality of the flowgram (raw sequence) resulting from early out-of-phase synthesis. Trimming eliminates the 3' end with above-threshold ambiguous base calls, but the remaining sequence still contains errors.

For deletion errors, *Seq.type* accounted for  $2.36\% \pm 1.39$  of the variation, reflecting substantial heterogeneity

between the reference sequences. The variables *Homopolymer* (accounting for  $6.89\% \pm 0.89$  of the variation) and *Position* (accounting for  $8.93\% \pm 5.21$  of the variation) were both concurrent to the deletion rate. Deletion errors tend to occur more frequently in homopolymers and their rates are higher towards the 3' end of sequences.

Finally, mismatch and ambiguous base call error rates were both found to be linked to *Position* ( $45.24\% \pm 4.04$  and  $25.85\% \pm 1.71$ , respectively) and *Seq.length* ( $25.00\% \pm 9.61$  and  $7.66\% \pm 2.11$ , respectively), with higher error rates found in 3' positions within sequences and longer sequences tending to have lower error rates.

Given this pattern, the next step in the integration of information is characterizing the effect of bead



localization on error rate. In particular, it is useful to consider whether position in a particular region or on the PT plate is linked to error rate. Heterogeneity in error rate as a function of bead location was found for insertions and deletions, whatever the PT plate analyzed. Heterogeneity was observed at both the region and plate scales. More precisely, error rate variation was mostly accounted for by the combination of several variables but, when the distribution of insertion errors fitted a gradient following the Y-axis in each region (Figure 3 and additional file 4), it was not accounted for by the variable *Dist.region* alone. However, the proportion of the model accounted for by the remaining variables is small ( $23.01\% \pm 2.62$ ). Adding the *Dist.region* to the model increases explanatory power to  $76.99\% \pm 2.62$ . The situation was similar for extraction of the signal at plate level, with *Dist.plate* increasing the explanatory power to  $77.39\% \pm 2.12$ . In summary, all regions had heterogeneous insertion and deletion error rates, but there were conserved gradients along both the x and y axes. Inverse physical gradients were observed for insertions and deletions. The covariation of these error types and sequence length indicates that they are influenced by a single latent variable (Figure 3).

## Conclusions

### From statistical inference to technical causes and perspectives

As detailed in the results and discussion section, error rate variability is mostly accounted for by the combination of the seven variables analyzed. However, the heterogeneous physical pattern may be partially driven by the combined influence of the central CCD camera (edge effect) with chemical flow direction (Y-axis). This explanation is, however, insufficient in itself to account for the observed pattern, and other variables clearly influence error rate. The negative relationship between insertion and deletion errors is probably related to physical acquisition issues, but chemistry-related artifacts probably also have an effect (through the related statistical variables analyzed), including the CAFIE effect (carry forward and incomplete extension) in particular. Carry forward occurs when a trace amount of nucleotide remains in a well after the apyrase wash, perpetuating premature nucleotide incorporations for specific sequence combinations during the next base flow and contributing to signal 'noise'. Incomplete extension occurs when some DNA strands on a bead fail to incorporate during the appropriate base flow. The strands that fail to incorporate must await another flow cycle for sequencing to continue and are thus incorporated out-of-phase with the rest of the strands [23].

This study clearly demonstrates that sequencing error rate, as deciphered here, is a heterogeneous feature in

454 GS-FLX Titanium pyrosequencing. We cannot extrapolate the results obtained for other technologies, such as the GS20 system, to this system, nor is the use of a single global error rate inappropriate. Our results provide information about the number of sequences required to correct for a specific erroneous position, when detected, but this procedure requires the error rate to be computed from within the 454 PT plate regions in which the physical distribution of error rate is heterogeneous. Internal DNA controls should therefore be used when appropriate [7,19,24] (readily available for amplicon sequencing), together with an error-corrected base caller [25], and routine procedures taking error data into account should be defined. When error rate is not estimated, a large number of potential false-positive polymorphisms would be expected and only post-sequencing validation can account for these artifacts [26,27]. For the resolution of this issue, the use of both sequencing primers and deep coverage, combined with the use of random sequencing priming sites, should partially compensate for error – even for high error rates – although it may be more difficult to distinguish between low-frequency alleles and errors than previously anticipated.

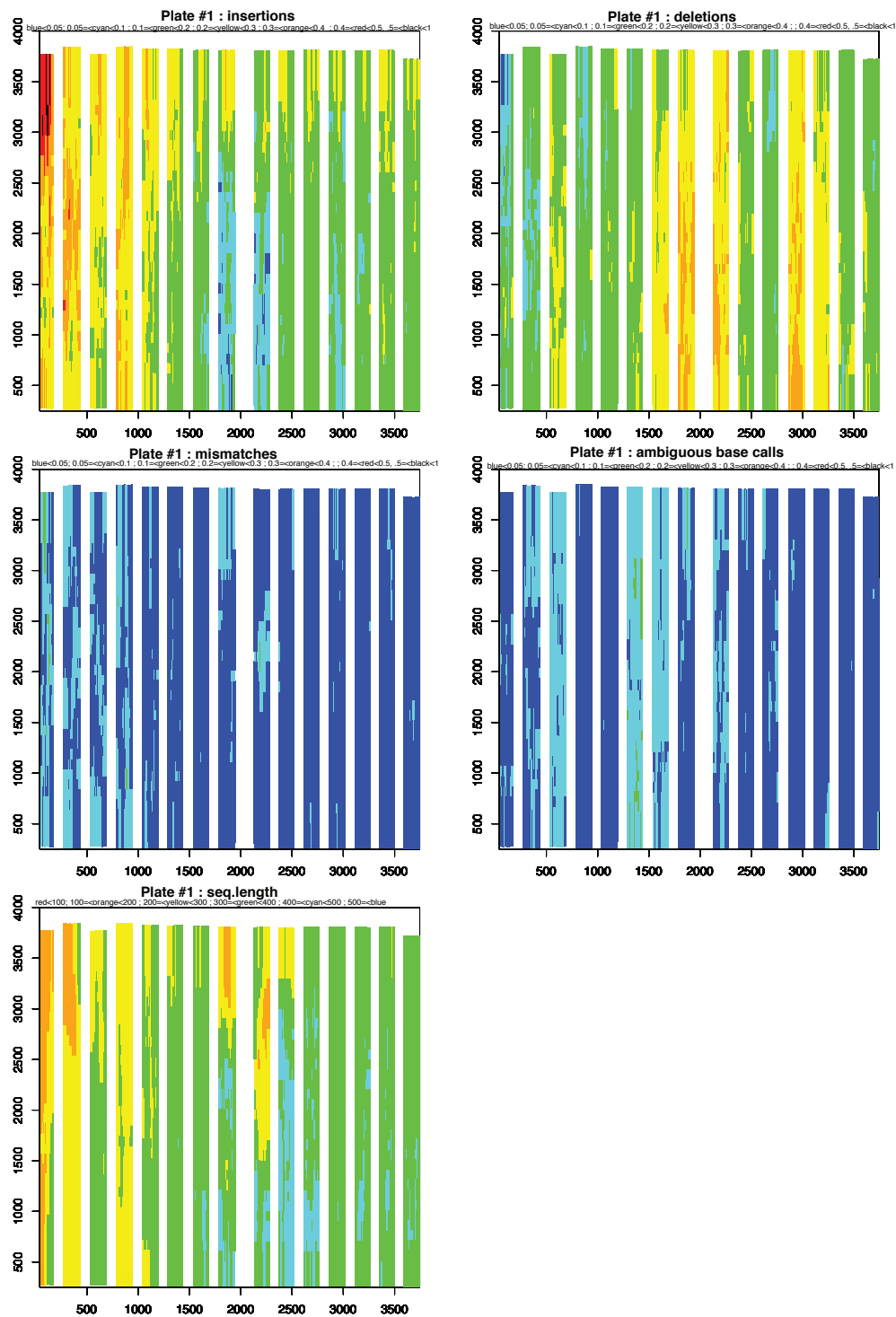
## Methods

### Experimental design and reference sequences

We used the six control DNA fragment Type I sequences (as provided in Roche 454 protocols) as reference sequences. This made it possible to use a large number of strictly identical templates to characterize the sequencing error rate of this technology. The sequences generated constituted a set of three replicates from three different runs, making it possible to assess the quality and accuracy of the 454 GS-FLX Titanium method. Six references were used, with lengths ranging from 500 to 592 bp and GC contents from 52.75% to 65.85%; each of these reference sequences contained a large number of homopolymers (20 to 34), defined as a succession of three or more identical bases. Homopolymer positions are shown on Figure 1 and in additional file 2. The reference sequences are provided in additional file 5.

All reference sequence positions were classified according to the presence and length of a homopolymer: (i) the first and last bases of a homopolymer and those within two bases on either side of a homopolymer were coded "1". All the other positions within the homopolymer were coded "3" to "6" (the length of the homopolymer). All positions outside these zones (not influenced by the homopolymer) were coded "0".

The dataset consisted of 86,237 sequences, corresponding to 29,100,738 positions. Sequencing was



**Figure 3 Spatial distribution of error rate variation.** For each error type and sequence length, the x-axis represents the spatial location of 454 reads and the y-axis represents the y-coordinates on the PT plate. The results presented in this figure correspond to plate #1. Data for the other two runs is presented in additional file 4. The 15 strips represent the 15 regions. We display separately the four types of error (insertions, deletions, mismatches and ambiguous base calls) and the length of the sequences generated. Colors indicate the ranges of error rates, from 0 to 1 (or the length of the sequences, from 0 to 500), using a sliding window (see materials and methods).

carried out at Genoscreen, France. We aimed to identify factors linked to error rate. For a tractable analysis, we analyzed a dataset corresponding to all the positions at which an error was detected, plus a similar number of error-free positions randomly selected from the whole original dataset.

### Sequencing error analysis

Reads (see additional file 6) were sorted according to their reference sequences, by BLASTn [28]. Each read was aligned to its reference sequence, to identify the positions and the number of sequencing errors. For optimization of the pairwise alignment parameters, the total number of errors was counted in a test dataset of 500 kb for a series of gap opening and gap extension penalties. The final analyses were carried out with ClustalW [29], using "1" as the gap opening penalty, and "10" as the gap extension penalty.

In the analyses, the observation unit was the position on the 454-generated sequences. These positions were transformed into the position on the reference sequence. Insertions are reported with respect to the position of the base preceding the gaps. For each position, a binary variable was defined indicating the presence or absence of a sequencing error. An error is defined here as discordance between two homologous positions: the first in the reference sequence and the second in the generated sequence. Discordance may refer to an insertion, a deletion, a nucleotide mismatch or an ambiguous base call (N) with respect to a non-available nucleotide determination on the replicate sequence (according to Huse et al. [10]). We investigated the pattern of 454-error type, focusing on the following seven factors: (i) **Position**, position in the sequence expressed as a proportion of the total length of the reference sequence (treated as a quantitative variable); (ii) **Seq.type**, the different reference sequences (qualitative variable with 6 settings); (iii) **Homopolymer**, type of homopolymer linked to the position as defined above; (iv) **Dist.region**, Euclidean distance between the generated sequence (bead) and the center of the region on the plate; (v) **Dist.plate**, Euclidean distance between the generated sequence and the center of the plate; (vi) **Seq.length**, length of the considered generated sequence (the observed sequence length results from the GS-FLX quality filtering process); (vii) **Region**, region of the plate in which the replicate was observed, region of the considered replicate.

The R package was used for all statistical tests [30]. The significance of regression coefficients was assessed by a permutation test with Benjamini-Hochberg correction, with  $\alpha = 0.05$ . As we studied both qualitative and quantitative variables, we decided to transform the

qualitative variables. The various possible settings of each qualitative variable were therefore replaced by a binary variable (dummy variable).

Let us define as  $\pi_i$  the sequencing error rate for the position  $i$ . As this value is supposed to vary as a function of the factors defined above, we have  $\pi_i = P(Y_i = 1/x_i^*) = E(Y_i/x_i^*) = \pi(x_i^*)$ .  $Y_i$  is the binary variable equal to 1 if an error is present and 0 otherwise.  $x_i^*$  is the vector  $(x_{1i}, x_{2i}, \dots, x_{7i})$  of the explanatory variables. We chose to model the error rate  $\pi(x_i^*)$  with a logistic model [22]:

$$\pi(x_i^*) = \frac{e^{(\sum_{i=1}^7 \beta \times x_i) + \beta_0}}{1 + e^{(\sum_{i=1}^7 \beta \times x_i) + \beta_0}}$$

Maximum likelihood estimators were considered to estimate the parameters of the model. Tests of significance of the parameters were then carried out with Student's t test. A model was generated for each of the three plates and for each of the error types (insertion, deletion, mismatch and N). All the analyses were performed with R (version 2.6.0).

The contribution of a given explanatory variable  $xi$  is assessed as follows. Let us denote by *comp.mod* the logistic model including all the variables considered, and *dev(comp.mod)*, its deviance. Let us define *dev(sub.model)* as the deviance associated with the model including all the variables other than the considered  $xi$ . Then,  $part(xi) = (dev(sub.model) - dev(comp.mod)) / dev(comp.mod)$  expresses the contribution of  $xi$  in addition to the other variables. We can symmetrically define the participation of all the variables other than  $xi$ :  $part(whole \setminus xi) = (dev(xi) - dev(comp.mod)) / (dev(comp.mod))$ . Hence the deviance of the complete model may be broken down into the sum of three terms: the first exclusive to  $xi$ , the second exclusive to the rest of the variables and the last expressing the explanation common to  $xi$  and the other variables:  $1 = part(xi) + part(whole \setminus xi) + (1 - part(xi) - part(whole \setminus xi))$ .

### Additional material

#### Additional file 1: Number of sequences to correct erroneous positions.

1a: this file illustrates the number of sequences necessary to obtain a majority of correct sequences. The x-axis shows the error rate and the y-axis shows the number of sequences needed, according to three possible probabilities: 0.001 0.01 and 0.05. 1b the x-axis shows the error rate for a given position (ranging from 0 to 0.5); the y-axis shows the cumulative proportion of erroneous sequences sampled (ranging from 0 to 0.5) in the total sample. Sample size varies from 10 to 100, 500 and 1,000 sequences. For a given error rate and a cumulative proportion of erroneous sequences in the sample of size N, the probability of observing this combination is indicated in color: green: 1 to 0.95, blue: 0.95 to 0.8, yellow: 0.8 to 0.6, orange: 0.6 to 0.5, red: 0.5 to 0.4, gray: 0.4 to 0.2 and white: below 0.2. For example, if the error rate is 0.2, the probability of observing a cumulative proportion of erroneous sequences in the sample of between 0 and 0.2 ranges between 0.4 and 0.5 (red envelope). In this case, the probability of there being 20% erroneous sequences in the sample is between 0.4 and 0.5. If we consider the same



error rate (0.2) with 40% erroneous sequences, then the probability ranges from 0.8 to 0.95 (blue envelope). If N increases, the variance of the probability envelopes decreases.

**Additional file 2: Distribution of errors along the reference sequences.** The blue line represents the proportion of sequences generated (y-axis) according to the sequence position (x-axis), using data obtained from the analysis of reference 5 reference sequences (excluding reference #3, which is displayed in Figure 1). The error rate for each type of error (insertions, deletions, mismatches and ambiguous base calls) is presented as a function of the sequence position (x-axis) and specific position on the y-axis. The position and length of homopolymers for each base is given on the x-axis to facilitate interpretation (green: A, red: T, black: G, blue: C).

**Additional file 3: Breakdown of error rate variation using all available variables.** For each plate, we used a logistic model to decipher the role of each selected variable in explaining the variation of error rate (see materials and methods). The figure is broken down by error type: a) insertions, b) deletions, c) mismatches and d) ambiguous base calls. We tested the deviance from the complete model by breaking down the model into the sum of three terms: the first exclusive to the single effect of the variable considered (in black), the second exclusive effect of the rest of the variables without the variable of interest (in gray) and the last expressing the sum of the effects of interactions between the variable considered and the other variables (in white). The contribution of each term (the proportion) for a considered variable can be viewed on the y-axis. Additional file 3 displays the results for plates #2 and #3 (results from the plate #1 are presented as Figure 2).

**Additional file 4: Spatial localization of error rate variation.** For each error type and the sequence length, the x-axis represents the spatial localization of 454 reads as x-coordinates and the y-axis represents the y-coordinates on the PT plate. The results presented in this additional data file 4 correspond to plates #2 and #3. The strips represent the regions. We display separately the four types of error (insertions, deletions, mismatches and ambiguous base calls) and the length of the generated sequences. Colors represent the ranges of error rates from 0 to 1 (or the length of the sequences from 0 to 500), using a sliding window (see materials and methods).

**Additional file 5: FASTA file of the 6 reference sequences.** The six reference DNA sequences used in this analysis are found in the corresponding FASTA file. They correspond to the control DNA fragments of type I provided with 454 GS-FLX Titanium sequencing kits. As such, the polymorphism displayed by the sequences corresponds purely to sequencing errors.

**Additional file 6: Raw data sequences from 454 GS-FLX Titanium sequencing.** This file contains three archives, including the raw FASTA files for each sequencing run.

#### Acknowledgements

We thank G. Nève for assistance with figure design. We thank M. Galan for useful comments on previous versions of the manuscript and S. Nielsen and J. Sappa (Alex Edelman) for major improvements of English grammar throughout the text. This work was supported by the AIP BioRessources "EcoMicro" grant from the French *Institut National de la Recherche Agronomique* (INRA), permanent institutional support from Montpellier SupAgro, University Aix-Marseille I, and the R&D budget of Genoscreen (Lille, France).

#### Author details

<sup>1</sup>Aix-Marseille Université, CNRS, IRD, UMR 6116 - IMEP, Equipe Evolution Génome Environnement, Centre Saint-Charles, Case 36, 3 place Victor Hugo, 13331 Marseille Cedex 3, France. <sup>2</sup>Genoscreen, Genomic Platform and R&D, Campus de l'Institut Pasteur, 1 rue du Professeur Calmette, Bâtiment Guérin, 4ème étage, 59000 Lille, France. <sup>3</sup>Institut National de la Recherche Agronomique, UMR 1301, Equipe BPI, 400 route des Chappes, BP 167, 06900 Sophia-Antipolis Cedex, France. <sup>4</sup>UMR CBGP (INRA/IRD/Cirad/Montpellier SupAgro), Campus international de Baillarguet, CS 30016, F-34988 Montpellier-sur-Lez cedex, France.

#### Authors' contributions

AG conceived the study and wrote the manuscript. EM participated in the design of the study, performed the bioinformatics analysis and helped to write the manuscript. NP participated in the design of the study, performed the statistical analysis and helped to write the manuscript. SF participated in the design and performed the molecular biology. TM helped to write the manuscript. JFM conceived the study and wrote the manuscript. All authors have read and approved the final manuscript.

Received: 20 July 2010 Accepted: 19 May 2011 Published: 19 May 2011

#### References

1. Zhou XG, Ren LF, Li YT, Zhang M, Yu YD, Yu J: **The next-generation sequencing technology: A technology review and future perspective.** *Science China-Life Sciences* 53(1):44-57.
2. Reis-Filho JS: **Next-generation sequencing.** *Breast Cancer Research* 2009, **11**.
3. Lundin S, Stranneheim H, Pettersson E, Klevebring D, Lundeberg J: **Increased Throughput by Parallelization of Library Preparation for Massive Sequencing.** *Plos One* 5(3).
4. Metzker ML: **Applications of next-generation sequencing. Sequencing technologies - the next generation.** *Nature Reviews Genetics* 11(1):31-46.
5. Hahn DA, Ragland GJ, Shoemaker DD, Denlinger DL: **Gene discovery using massively parallel pyrosequencing to develop ESTs for the flesh fly *Sarcophaga crassipalpis*.** *Bmc Genomics* 2009, **10**.
6. Aury JM, Cruaud C, Barbe V, Rogier O, Mangenot S, Samson G, Poulain J, Anthouard V, Scarpelli C, Artiguenave F, Wincker P: **High quality draft sequences for prokaryotic genomes using a mix of new sequencing technologies.** *Bmc Genomics* 2008, **9**.
7. Babik W, Taberlet P, Ejsmond MJ, Radwan J: **New generation sequencers as a tool for genotyping of highly polymorphic multilocus MHC system.** *Molecular Ecology Resources* 2009, **9**(3):713-719.
8. Galan M, Guivier E, Caraux G, Charbonnel N, Cosson JF: **A 454 multiplex sequencing method for rapid and reliable genotyping of highly polymorphic genes in large-scale studies.** *Bmc Genomics* 11(1):296.
9. Kunin V, Engelbrekton A, Ochman H, Hugenholtz P: **Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates.** *Environ Microbiol* 2010, **12**(1):118-123.
10. Huse SM, Huber JA, Morrison HG, Sogin ML, Mark Welch D: **Accuracy and quality of massively parallel DNA pyrosequencing.** *Genome Biology* 2007, **8**(7).
11. Margulies M, Egholm M, Altman W, Attiya S, Bader J, Bemben L, Berka J, Braverman M, Chen Y, Chen Z: **Genome sequencing in microfabricated high-density picolitre reactors.** *Nature* 2005, **437**:376-380.
12. Huse SM, Welch DM, Morrison HG, Sogin ML: **Ironing out the wrinkles in the rare biosphere through improved OTU clustering.** *Environ Microbiol* 2010, **12**(7):1889-1898.
13. Dohm JC, Lottaz C, Borodina T, Himmelbauer H: **Substantial biases in ultra-short read data sets from high-throughput DNA sequencing.** *Nucleic Acids Res* 2008, **36**(16).
14. Hoffmann S, Otto C, Kurtz S, Sharma CM, Khaitovich P, Vogel J, Stadler PF, Hackermuller J: **Fast Mapping of Short Sequences with Mismatches, Insertions and Deletions Using Index Structures.** *Plos Computational Biology* 2009, **5**(9).
15. Campbell PJ, Pleasance ED, Stephens PJ, Dicks E, Rance R, Goodhead I, Follows GA, Green AR, Futreal PA, Stratton MR: **Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing.** *Proc Natl Acad Sci USA* 2008, **105**(35):13081-13086.
16. Wheat CW: **Rapidly developing functional genomics in ecological model systems via 454 transcriptome sequencing.** *Genetica* 138(4):433-451.
17. Saunders IW, Brohede J, Hannan GN: **Estimating genotyping error rates from Mendelian errors in SNP array genotypes and their impact on inference.** *Genomics* 2007, **90**(3):291-296.
18. Quinlan AR, Stewart DA, Stromberg MP, Marth GT: **Pyrobayes: an improved base caller for SNP discovery in pyrosequences.** *Nat Meth* 2008, **5**(2):179-181.
19. Wegner KM: **Massive parallel MHC genotyping: titanium that shines.** *Molecular Ecology* 2009, **18**(9):1818-1820.
20. Benjamini Y, Hochberg Y: **Controlling the false discovery rate - a practical and powerful approach to multiple testing.** *Journal of the Royal Statistical Society Series B-Methodological* 1995, **57**(1):289-300.

21. Hamilton D: **Sometimes R2 is greater than RYX1(2) + RYX2(2) - correlated variables are not always redundant.** *American Statistician* 1987, **41(2)**:129-132.
22. McCullagh P, Nelder JA: **Generalized Linear Models.** London: Chapman and Hall;; 2 1989.
23. **Genome Sequencer FLX System Software Manual, version 2.3.** Branford, CT 06405, USA: 454 Life Sciences Corp., A Roche Company; 2009.
24. Hoff KJ: **The effect of sequencing errors on metagenomic gene prediction.** *Bmc Genomics* 2009, **10**.
25. Saeed F, Khokhar A, Zagordi O, Beerenwinkel N: **Multiple Sequence Alignment System for Pyrosequencing Reads.** In *Bioinformatics and Computational Biology, Proceedings* Edited by: Rajasekaran S 2009, 362-375, vol. 5462.
26. Lynch M: **Estimation of Allele Frequencies From High-Coverage Genome-Sequencing Projects.** *Genetics* 2009, **182(1)**:295-301.
27. Lynch M: **Estimation of Nucleotide Diversity, Disequilibrium Coefficients, and Mutation Rates from High-Coverage Genome-Sequencing Projects.** *Molecular Biology and Evolution* 2008, **25(11)**:2409-2419.
28. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucl Acids Res* 1997, **25(17)**:3389-3402.
29. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG: **Clustal W and clustal X version 2.0.** *Bioinformatics* 2007, **23(21)**:2947-2948.
30. R Development Core Team: **R: A Language and Environment for Statistical Computing.** R Foundation for Statistical Computing, Vienna, Austria; 2010 [<http://www.R-project.org>], ISBN 3-900051-07-0.

doi:10.1186/1471-2164-12-245

**Cite this article as:** Gilles *et al.*: Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. *BMC Genomics* 2011 **12**:245.

**Submit your next manuscript to BioMed Central  
and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)

