Les codes-barres ADN des lépidoptères. Partie 2 : protocoles expérimentaux

Rodolphe Rougerie

Résumé : Ce second article est une présentation de la mise en pratique de l'utilisation des codes-barres ADN, depuis le prélèvement d'un échantillon de tissu jusqu'à la production des séquences ADN.

Mots-clés : Lépidoptères, codes-barres ADN, identification spécifique, taxonomie, protocoles.

Summary: This second paper presents the practical application of DNA barcoding, from tissue sampling to sequence production.

Keywords: Lepidoptera, DNA barcodes, species identification, taxonomy, protocols.

Après la présentation lors d'un précédent article de ce que sont ces fameux « codes-barres ADN » (Rougerie, 2011), ce second volet développe les détails de leur mise en pratique, depuis les spécimens jusqu'à la génération des séquences. Ce qui suit s'adresse aux lectrices et lecteurs qui envisagent peut-être de participer à des campagnes de « barcoding », ou à celles et ceux qui souhaitent tout simplement mieux comprendre l'arrière-scène qui n'est pas toujours détaillée dans les publications utilisant cet outil génétique. Cette présentation s'inscrit dans le contexte du projet iBOL (international Barcode of Life) et fait essentiellement référence aux protocoles mis en place par le laboratoire canadien de l'université de Guelph, le CCDB (Canadian Centre for DNA Barcoding), pionnier et leader dans le domaine. De même, il est fait référence à la base de données en ligne BOLD (Barcode of Life Datasystems, www.boldsystems.org) qui centralise la très grande majorité des codes-barres ADN produits mondialement, tous groupes taxonomiques confondus. Il n'en demeure pas moins que les méthodes, les protocoles et les précautions détaillées ici s'appliquent plus généralement à tout projet similaire réalisé dans un autre laboratoire ou utilisant une autre base de données.

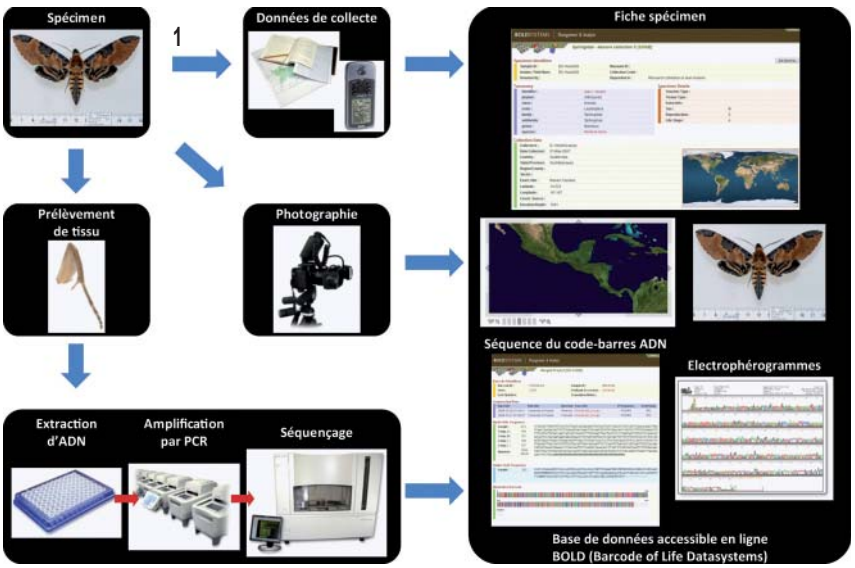


Figure 1. Déroulement des opérations menant à la production d'une fiche spécimen dans la base de données en ligne BOLD. Cette fiche intègre les données propres au spécimen et à la séquence du code-barres ADN qui a été généré en laboratoire à partir d'un prélèvement de tissu. Au 12 avril 2012, BOLD contient 746 855 fiches spécimens de ce type pour les lépidoptères, représentant 81 915 espèces (655 835 spécimens de 72 848 espèces sont séquencés).

► DU SPÉCIMEN À LA SÉQUENCE – MODE D'EMPLOI (fig. 1)

Étape 1 : sur le terrain ou dans la collection

(a) La sélection des spécimens. C'est bien sûr là que tout commence, par ce choix qui n'est pas du tout trivial et qui doit prendre en compte à la fois des critères techniques et pratiques, mais également les exigences propres au contexte de la démarche de chacun :

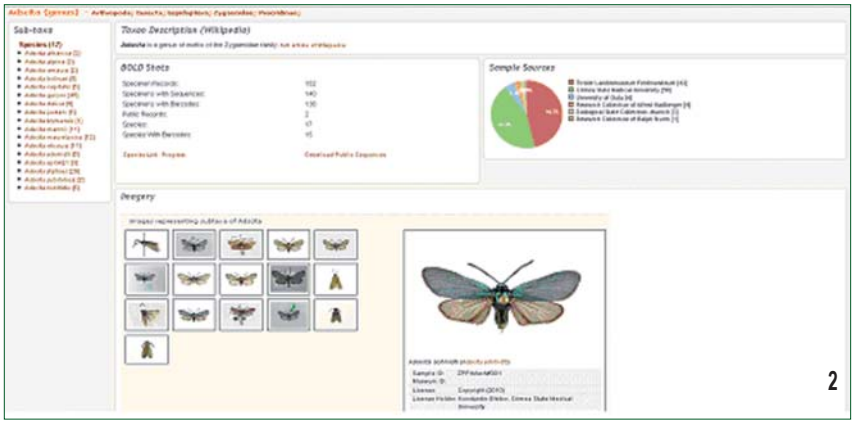
- le choix des espèces : lorsqu'il s'agit de contribuer à construire une librairie de référence ou de résoudre des problèmes d'ordre taxonomique ou de nomenclature, il est important de synchroniser l'échantillonnage avec celui déjà effectué. Il est conseillé de privilégier les espèces absentes de la librairie ou des spécimens d'origine géographique intéressante. Ces précautions sont cruciales pour pouvoir s'intégrer dans les programmes existant tels iBOL, car leur respect conditionne la prise en charge financière des travaux en laboratoire. La meilleure façon de savoir ce qui a déjà été échantillonné et ce qui peut l'être est de consulter le « taxonomy browser » de BOLD (fig. 2) qui liste le nombre de spécimens séquencés pour chaque espèce et où l'on peut lancer une recherche par nom ou bien naviguer jusqu'au taxon recherché via une classification hiérarchique. En revanche, si la démarche consiste à utiliser les codes-barres ADN pour identifier des spécimens inconnus, le problème du choix des espèces ne se pose pas, et il suffit de trouver un laboratoire qui acceptera de prendre en charge ces échantillons.
- le choix des spécimens : si l'échantillonnage de papillons est relativement simple, il demande cependant un travail substantiel qui doit conduire à un choix raisonné des spécimens pour optimiser le taux de succès. Il est très frustrant en effet de n'obtenir de séquences que pour la moitié ou moins des spécimens que l'on a passé des jours à échantillonner, sans parler du gaspillage causé par les efforts vains réalisés au laboratoire. Au premier rang des critères de choix, l'âge des spécimens : la molécule d'ADN se dégrade progressivement avec le temps, et plus la date de collecte du spécimen est ancienne, plus les chances d'obtenir son code-barres s'amenuisent (fig. 3). En règle générale, pour des spécimens conservés à sec après capture et préparation, le succès est de l'ordre de 95-100% pour des spécimens frais (1 à 2 ans), 80-90% pour des spécimens de 2 à 10 ans. Au-delà, le taux de succès devient moins satisfaisant et décroit rapidement avec l'ancienneté de la collecte. Il peut cependant être relevé à des valeurs tout à fait convenables (au-delà de 80%) pour des spécimens capturés depuis 20 à 30 ans grâce à l'utilisation de protocoles de laboratoire dédiés à cet ADN dégradé. En fait, des protocoles ont même été développés avec succès pour générer les codes-barres de spécimens vieux de plus d'un siècle (en particulier pour les spécimens types, voir par ex. Hausmann *et al.*, 2009). Pour le coup, on sort ici du barcoding en « routine » et ces protocoles sont plus coûteux et prennent beaucoup plus de temps ; il est donc recommandé de toujours sélectionner les spécimens les plus récemment collectés. Le taux de succès peut être également affecté si les spécimens sont passés par le ramolissoir, car la dégradation de l'ADN est accélérée dans des conditions humides, ou – pire – si

des champignons viennent à se développer. Cette étape classique de préparation des papillons n'est cependant absolument pas rédhibitoire. A l'inverse, la congélation stoppe ou ralentit considérablement la dégradation de l'ADN et des spécimens congelés depuis des années donnent des résultats proches de ceux obtenus avec des spécimens frais. Enfin, certains agents chimiques utilisés pour tuer les spécimens ou les préserver en collection sont aussi dommageables, dans une certaine mesure, pour l'ADN ; c'est le cas de l'éther acétique, du thymol, du benzène et du dichlorvos. Les spécimens tués dans un flacon à cyanure ou par injection d'ammoniaque semblent se prêter parfaitement à la génération de séquences d'ADN, de même que l'utilisation de paradichlorobenzène ou de naphtaline (produits cancérigènes dont l'utilisation est désormais réglemntée) pour la préservation des collections.

- le nombre de spécimens échantillonnés par espèce peut varier ; il est de 5 à 10 en moyenne dans les librairies existantes. Si l'objectif est de créer ou compléter une librairie de référence, on veillera à choisir des spécimens typiques d'une part (par leur habitus et/ou leur localité), mais aussi des individus représentant la variation morphologique individuelle observée au sein de l'espèce et de provenances géographiques couvrant au maximum son aire de répartition. Si une nouvelle espèce est en cours de description, il est tout particulièrement intéressant de veiller à ce que le futur holotype soit échantillonné ; en effet, si l'avenir réserve la découverte d'une espèce cryptique dans ce même groupe, il sera alors précieux de pouvoir prendre en compte le code-barres du type pour déterminer sans ambigüité qui est qui.

(b) Le prélèvement de tissu. Une fois le spécimen choisi, la première chose à faire est de lui attribuer un code unique qui permettra d'associer le papillon (ou la larve si c'est une chenille en alcool) avec le prélèvement qui va être fait. Dans la base de données BOLD, ce code est baptisé « SampleID » ; il doit être inscrit sur une étiquette épinglé sous le spécimen. Son format est libre, mais il est très souvent formé des initiales identifiant la collection ou le récolteur, suivies d'un numéro incrémenté pour chaque nouveau spécimen. Faire précéder ce code d'un suffixe renvoyant à sa signification, par exemple « BC » pour « barcode », permettra de le distinguer au premier coup d'œil d'autres codes renvoyant à des préparations de genitalia par exemple. Une simple patte convient généralement parfaitement, sauf pour les très petits microlépidoptères pour lesquels il peut être préférable d'en placer deux par tube, surtout si la collecte du spécimen date de plusieurs années. Pour les très gros papillons (Sphingidae, Saturniidae, Lasiocampidae, Cossidae, etc.) une patte entière représente beaucoup trop de tissu et il faut en couper un fragment car sinon l'excès de tissu peut inhiber certaines des réactions qui seront réalisées par la suite en laboratoire. Ces recommandations peuvent varier selon les protocoles utilisés dans les différents laboratoires, mais elles s'appliquent généralement lorsque les codes-barres sont produits en routine et que les échantillons sont assemblés dans une plaque de 96 « puits » (Fig. 4). Dans ce type de plaque, on prendra soin de déposer une goutte d'éthanol (95%) dans le fond de chaque puits afin d'éviter des problèmes liés à l'électricité statique qui pourrait faire « sauter » les pattes au moment de l'ouverture.

Un simple fichier Excel (il existe des fichiers spécifiques prêts-à-l'emploi) permet de faire le lien entre le spécimen et le tissu prélevé en associant le code SampleID avec la position dans la plaque dont les puits sont numérotés de A1 à H12. Pour diminuer les risques de contamination, c'est-à-dire le transfert accidentel d'ADN d'un puits à son voisin (via des écailles par exemple), on évitera d'échantillonner dans une même plaque et surtout à proximité les uns des autres des spécimens frais et des spécimens beaucoup plus vieux. Aussi, il est important de nettoyer après chaque prélèvement les pinces utilisées ; on peut pour cela les passer à la flamme d'un briquet pendant quelques secondes (après les avoir essuyées si beaucoup d'écailles s'y trouvent collées) ou bien les tremper dans un flacon d'alcool et les essuyer avec un mouchoir en papier.



2

Figure 2. Capture d'écran de la page du « taxonomy browser » de BOLD détaillant l'avancement de l'échantillonnage pour le genre Adscita (Zygaenidae). Ce site permet de connaître le nombre de spécimens déjà échantillonnés pour chaque taxon, leurs provenances géographiques et les sources (institutions, collections privées) de ce matériel.

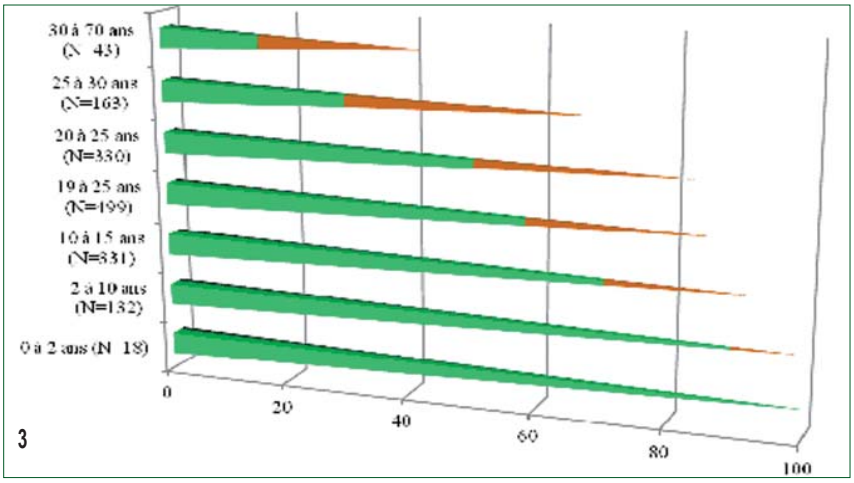
(c) Collecte des données. Un fichier Excel prêt à l'emploi permet là encore la saisie de ces données en vue de leur compilation dans BOLD. Ces données sont connectées au spécimen via le code SampleID et se divisent en quatre grands groupes : (1) les informations sur le spécimen (code SampleID, lieu de dépôt, personne ayant échantillonné le tissu) ; (2) la taxonomie (classification, identification spécifique, identificateur) ; (3) les détails propres au spécimen (sexe, stade de développement, mode de reproduction, informations supplémentaires, notes) ; (4) les données de récolte (collecteur, date, pays, province, région, site exact, coordonnées GPS, altitude). La base de données BOLD contient elle-même davantage de champs (p.ex. précision et source des coordonnées GPS, de l'altitude) qui ne sont pas intégrés cependant dans les fiches de saisie standard.

(d) Photographie. Une photographie doit accompagner chaque spécimen échantillonné. Il s'agira du spécimen étalé de préférence, pour en faciliter l'identification, mais cela peut aussi être un individu non préparé si l'on souhaite prélever la patte avant de ramollir l'insecte pour accroître le succès du séquençage. BOLD peut associer plusieurs images à un spécimen et il sera possible d'ajouter une image supplémentaire plus tard, lorsque le spécimen aura été étalé. Une photo de genitalia peut aussi être ajoutée le cas échéant, ou même une photographie du spécimen dans la nature avant capture par exemple. Là encore, un fichier Excel dédié à la prise en charge des photographies permettra le transfert des fichiers images dans BOLD via le code SampleID et le nom des fichiers, tout en y associant un copyright au passage pour le photographe. Comme la saisie de données et les multiples actions décrites ci-dessus sont autant de sources d'erreurs potentielles, une astuce consiste à photographier le spécimen en faisant en sorte que l'étiquette portant le code SampleID soit visible. Cela permet de détecter un éventuel décalage entre les données et les images et de résoudre plus facilement d'éventuelles erreurs passées inaperçues.

Une fois ces tâches effectuées, la plaque de pattes ayant été remplie et les fichiers et images étant prêts, le tout peut être envoyé au laboratoire qui traitera les échantillons. Les spécimens sources ont une importance cruciale désormais comme référence pour les séquences d'ADN qui seront produites. Il est donc fondamental d'en assurer une bonne conservation et un classement qui permet d'y accéder facilement. Idéalement, ces spécimens doivent être déposés, ou du moins destinés à l'être, dans une collection d'accès public comme un Muséum ou une université.

Étape 2 : au laboratoire

Bien que différents laboratoires utilisent du matériel, des réactifs et des protocoles différents, le traitement des échantillons suit invariablement les même phases : (1) extraction de l'ADN, (2) amplification par PCR (Polymerase Chain Reaction) et (3)



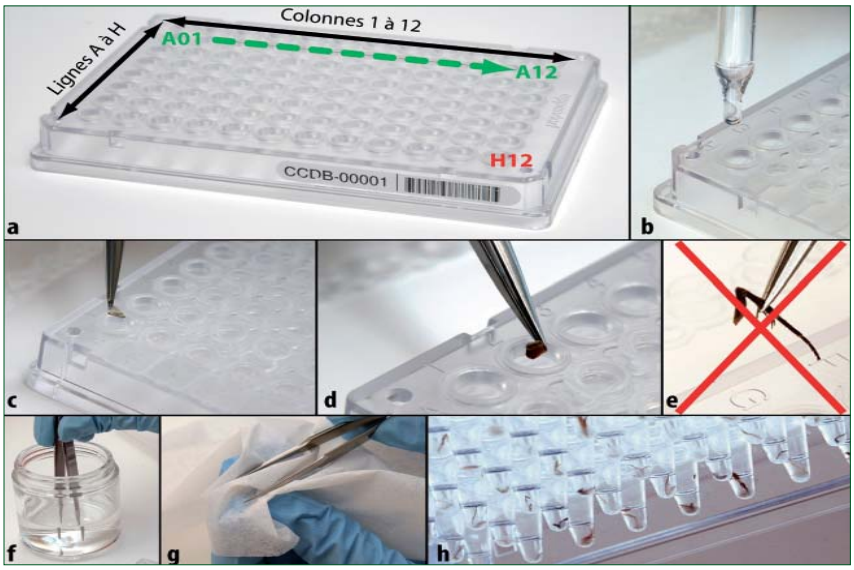


Figure 3. Taux de succès du séquençage en fonction de l'ancienneté de la collecte des spécimens. Ce diagramme présente les résultats obtenus pour près de 1500 papillons échantillonnés dans la collection de lépidoptères de Haute-Normandie de Bernard Dardenne. Une distinction est apportée entre les codes-barres complets (plus de 500 bp en longueur, en vert) et partiels (moins de 500 bp, en orange). La qualité de ces derniers est moins bonne, mais ils offrent néanmoins souvent une excellente discrimination des espèces.

Figure 4. Matériel d'échantillonnage utilisé pour collecter les tissus en vue de l'extraction de l'ADN. Le modèle de plaque le plus couramment utilisé possède 96 puits (a) dans lesquels on déposera une goutte (b) ou environ 30µL d'éthanol à 95%. L'échantillon peut être une patte pour les espèces de petite taille (c, h), ou un fragment (d) pour de gros papillons. Il faut éviter d'échantillonner trop de tissu (e) et nettoyer les pinces entre chaque prélèvement (f-g).

séquençage du gène visé (fig. 1). Dans un laboratoire dédié comme le CCDB dont la capacité de production dépasse 300 000 séquences par an (près de 60 plaques traitées par semaine !) ces étapes sont largement automatisées et gérées avec précision à l'aide d'un logiciel assurant la traçabilité de chaque échantillon dans le laboratoire. Chaque échantillon entrant dans le laboratoire se voit attribué un second code unique, le « ProcessID », auquel toutes les données générées au laboratoire seront associées et elles-mêmes liées dans BOLD au code SampleID. En pratique toutes ces étapes peuvent être réalisées dans la même journée, mais les contraintes des laboratoires font souvent durer le processus plus longtemps. Au CCDB par exemple, il faut 2 à 4 semaines après l'entrée de la plaque au laboratoire pour que les séquences soient générées et transférées dans BOLD (incluant la phase d'édition manuelle, voir plus loin). En routine, les échantillons de tissus ne sont pas conservés après extraction de l'ADN, mais le protocole utilisé n'est cependant pas destructif et il demeure possible de récupérer les tissus pour les spécimens les plus précieux. On préfèrera ainsi parfois utiliser l'abdomen entier de microlépidoptères pour l'extraction de l'ADN et celui-ci pourra être utilisé par la suite pour étude morphologique.

L'amplification du fragment de gène visé par PCR est une étape clé du travail en laboratoire. Il s'agit de produire expérimentalement des milliers de copies d'un fragment du génome grâce à l'utilisation d'une enzyme de réplication associée à de courts fragments d'ADN synthétisés spécifiquement – les amorces, qui définissent le début et la fin de la séquence ciblée. Pour du matériel frais, on peut généralement amplifier le code-barres ADN (un fragment de 658 paires de bases (bp) du gène mitochondrial COI) en une seule fois et il existe des amorces dites « universelles » que l'on pourra utiliser avec succès chez tous les Lépidoptères. En revanche si le spécimen source est plus ancien, la dégradation de l'ADN aura causé la fragmentation de la molécule dans les cellules et il est alors nécessaire de viser par PCR plusieurs fragments plus courts pour finalement reconstruire un code-barres complet. Cette approche alternative donne d'excellents résultats en visant deux fragments de 300-400 bp pour des spécimens de 10 à 30 ans. Elle peut s'appliquer également à des spécimens beaucoup plus anciens en amplifiant et en assemblant six fragments de 80 à 120 bp.

Étape 3 : dans l'arrière salle, au laboratoire

Lorsque toutes les étapes expérimentales ont été effectuées, le produit final du séquençage se présente sous la forme de fichiers informatiques appelés électrophérogrammes. Ces résultats bruts consistent en une série de courbes sinusoïdales identifiant la séquence des nucléotides du gène amplifié. Ils sont traduits automatiquement par un logiciel, mais il est néanmoins nécessaire d'opérer un contrôle visuel et d'éditer manuellement les séquences pour s'assurer d'une qualité optimale des résultats. Après cette étape, les séquences peuvent être transférées dans la base de données. Sur BOLD, chaque code-barres est associé à ses électrophérogrammes qui peuvent être consultés pour vérifier d'éventuelles erreurs dans l'édition manuelle des séquences.

C'est à ce stade que seront filtrées les séquences de mauvaise qualité et les éventuelles contaminations. Celles-ci peuvent être évidentes si le contaminant est taxonomiquement très éloigné, mais pourront parfois demander un examen beaucoup plus approfondi des résultats pour contrôler si une incohérence des résultats est causée par une contamination par l'ADN d'un puits voisin (ou une inversion des échantillons dans la plaque, etc.) ou plus simplement par une erreur d'identification.

C'est après ces vérifications que l'on va pouvoir analyser les résultats et utiliser ces codes-barres ADN comme référence permettant d'identifier des spécimens inconnus ou encore comme un jeu de données complémentaire en taxonomie pour caractériser des espèces nouvelles, confirmer des mises en synonymies, etc. Nous aborderons les méthodes d'analyse couramment utilisées dans une prochaine livraison d'*oreina*.

► REMERCIEMENTS

L'auteur remercie Eric Drouet pour sa relecture du manuscrit ainsi que le CCDB pour la permission d'utiliser leurs photographies (droits réservés au Biodiversity Institute of Ontario. ■

BIBLIOGRAPHIE

Hausmann (A.), Hebert (P.D.N.), Mitchell (A.), Rougerie (R.), Sommerer (M.), Edwards (T.) et Young (C.J.), 2009. – Revision of the Australian *Oenochroma vinaria* Guenée, 1858 species-complex (Lepidoptera: Geometridae, Oenochrominae): DNA barcoding reveals cryptic diversity and assesses status of type specimen without dissection. *Zootaxa*, **2239**, 1-21.

Rougerie (R.), 2011. – Les codes-barres ADN des Lépidoptères. Partie 1 : un outil pour l'identification des espèces. *Oreina*, **15**, 7-9.

Laboratoire ECODIV / Université de Rouen, Bâtiment IRESE A Place Émile Blondel F-76821 Mont-Saint-Aignan Cedex
rrougeri@gmail.com