

Outils de gestion des ressources biologiques

Frédéric Sanchez^{1*}, Tou Cheu Xiong¹

Résumé. Les recherches en biologie génèrent un grand nombre d'échantillons biologiques. Ces ressources biologiques se présentent ainsi sous diverses formes : des collections de graines de différentes variétés végétales, des plasmides bactériens, des clones bactériens, des amorces oligonucléotidiques, etc. De plus, le nombre de ces échantillons augmente rapidement avec chaque projet de recherche. Dès lors, les méthodes de clonage se sont diversifiées et sont devenues plus performantes, multipliant le nombre de clones à sauvegarder. La nécessité d'une gestion de ces ressources est donc primordiale pour une bonne traçabilité et un bon suivi de l'avancée des projets de recherche. En effet, la gestion de ces ressources par des outils informatiques a permis de centraliser toutes les ressources de plusieurs projets, autorisant non seulement une traçabilité de chaque échantillon mais également la recherche et la disponibilité des ressources déjà réalisées pour la communauté scientifique. La génération de bases de données réalisées à l'aide d'outils informatiques a donné lieu à une gestion plus efficace de ces ressources. Cet article présente un outil mis en place pour la gestion des ressources avec le logiciel FileMaker Pro, au sein de l'UMR de Biochimie et Physiologie moléculaire des Plantes (B&PMP) à l'INRA de Montpellier.

Mots clés : bases de données, ressources biologiques, FileMarker Pro, traçabilité

Introduction

La quantité de données biologiques disponible augmente de façon considérable depuis des décennies. Des bases de données ont ainsi été conçues afin de créer des collections cohérentes et structurées. Jusqu'à la fin des années 1960, elles avaient une organisation classique en fichiers (ref.1). Des logiciels ont émergé ensuite, facilitant leur utilisation. Ils ont été nommés Système de Gestion de Bases de Données (SGBD) et ont pour but d'assurer la structuration, la mise à jour et la consultation des informations d'une base de données. Il en existe plusieurs comme par exemple MySQL, Oracle database, FileMaker Pro (ref. 2). Aujourd'hui, les bases de données sont omniprésentes dans le domaine de la recherche. La plus célèbre est la base de données bibliographiques scientifiques Medline. On la consulte en ligne à l'aide du moteur de recherche PubMed (<http://www.ncbi.nlm.nih.gov/pubmed>).

Les laboratoires de recherche sont parmi les plus importants pourvoyeurs d'échantillons. Au sein de l'Unité Biochimie et Physiologie Moléculaire des Plantes (B&PMP), les échantillons produits ou achetés sont des graines, de l'ADN bactérien (plasmide), des clones bactériens et des oligonucléotides servant d'amorces pour des amplifications d'ADN ou d'ARN. Grâce aux nouvelles méthodes en biologie moléculaire, plusieurs milliers d'échantillons peuvent être générés pour un projet de recherche (**Tableau 1**). De plus ces échantillons sont essentiellement des organismes génétiquement modifiés (OGM), utilisés pour la recherche. Par conséquent, une traçabilité et un suivi de ces échantillons sont indispensables pour l'obtention d'agréments et pour la démarche d'assurance qualité recherche (AQR). Pour toutes ces raisons, les informations cruciales de chaque échantillon doivent être répertoriées telles que la provenance du laboratoire, la méthodologie employée, etc. Pour centraliser ces informations, la plupart des équipes utilisent des logiciels non dédiés comme Excel par exemple. Ces logiciels ont l'avantage d'être simple d'utilisation mais ne sont pas adéquats pour gérer des grandes quantités de

¹ INRA, UMR 386 Biochimie et Physiologie Moléculaire des Plantes, Institut Claude Grignon Bat 7 F-34060 Montpellier, France
frederic.sanchez@supagro.inra.fr

Frédéric Sanchez, Tou Cheu Xiong

données de nature différente. Le besoin d'un outil de gestion des ressources adapté et plus efficace s'est vite avéré nécessaire. Dans l'Unité B&PMP, la mutualisation des graines utilisées dans les différents programmes de recherche a nécessité la gestion et la création d'une base de données. Le logiciel FileMaker Pro a été introduit dans ce but. La base de données ainsi générée était constituée de toutes les informations associées pour chaque lignée. Cette base de données a rapidement présenté l'inconvénient de demander trop de rubriques à renseigner pour chaque échantillon, ce qui a rendu fastidieux sa mise en place. Néanmoins cette expérience fut très utile dans le travail de réflexion, pour la conception de nouvelles bases de données non seulement pour les graines mais aussi pour les clones bactériens, plasmides et amorces. Le choix a été de développer des nouvelles bases de données plus simples et spécifiques pour chaque type de ressources (graines, plasmides, amorces, bactéries). Le choix du logiciel s'est porté sur FileMaker Pro car il a l'avantage d'être modulable et accessible à des non-informaticiens. De plus les chercheurs, ingénieurs et techniciens de l'Unité ont été déjà familiarisés à ce logiciel. Cet article présente une version aboutie des nouvelles bases de données générées avec Filemarker pro qui ont été utilisées sur une période de 4-5 années dans l'Unité. Ces bases sont destinées à la gestion des échantillons par des biologistes ayant peu ou pas d'expérience en programmation.

Tableau 1. Exemple de projet de recherche pour lequel 1024 échantillons ont été implantés dans les différentes bases de données

Projet	Graines	Plasmides	Amorces	Bactéries	Total échantillons
Etude de la régulation de canaux ioniques chez <i>Arabidopsis thaliana</i>	140	154	379	351	1024

Présentation des bases FileMaker pro

FileMaker Pro est un logiciel de gestion de bases de données développé par *FileMaker Inc*, filiale d'Apple. Il a été conçu en 1985 et fonctionne sous Mac OS X, Windows, en réseau poste à poste ou en mode client-serveur. Il ne nécessite aucune commande SQL (structured query language) pour faire des recherches ou créer des rubriques, mais permet de créer une interface graphique basée sur des boutons déclenchant des programmes (scripts). FileMaker Pro est utilisé par exemple pour la gestion de stocks dans le domaine du marketing (dossiers clients) et de la santé (dossiers patients). Jusqu'à 250 utilisateurs peuvent utiliser simultanément en réseau une base de données. Différents niveaux de droit d'accès sont possibles (ref.3). Des codes d'accès permettent ainsi d'avoir des droits plus ou moins restreints pour l'utilisation des bases (**Figure 1**). L'administrateur possède tous les droits. De ce fait, seul l'administrateur peut modifier et mettre à jour les bases sans restriction. Les utilisateurs n'ont qu'un accès consultatif (**Figure 1**).

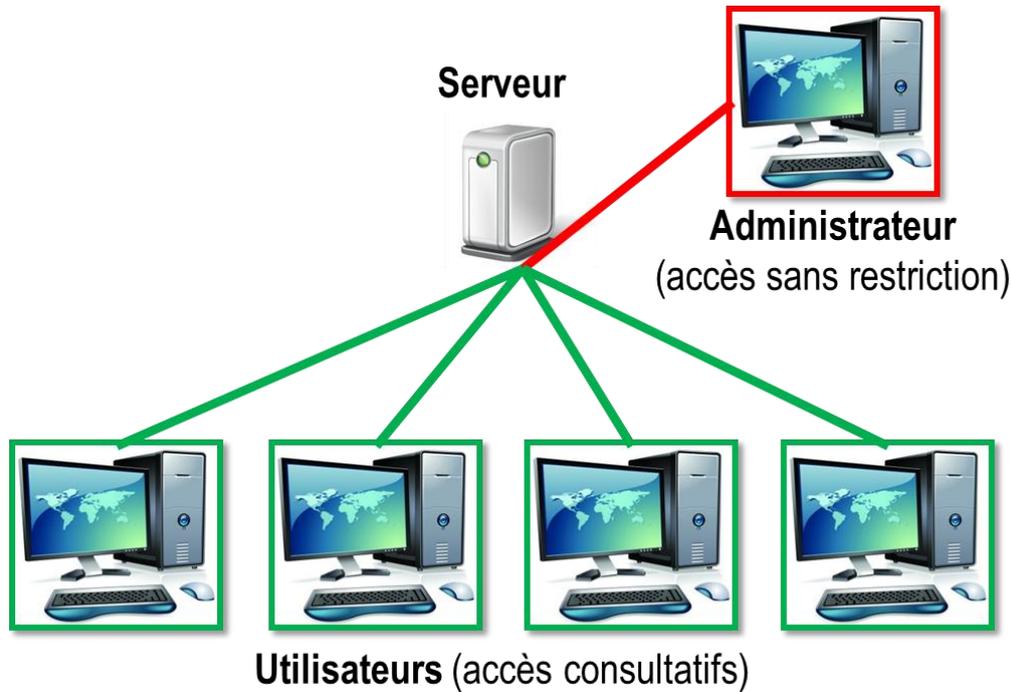


Figure 1. Partage de la base de données en réseau. Le réseau est composé d'un Administrateur (en rouge) qui rentre des échantillons dans la base de données et peut modifier la structure de l'interface. Les utilisateurs (en vert) sur d'autres postes informatiques peuvent consulter les bases pour rechercher des échantillons. Les bases sont stockées sur un serveur en réseau.

La création d'une base de données avec Filemaker pro est très aisée et ce fait grâce à des outils graphiques. Ces bases sont très facilement modulables. Cela permet par exemple de créer un code couleur personnalisable ou d'ajouter des boutons d'action (**Figure 2A**). L'ajout ou la suppression de rubriques à renseigner est aussi très fluide. La version de Filemaker pro 7.0v3 a été utilisée pour créer les bases présentées dans cet article.

Exemple de bases de données pour des collections de graines, plasmides bactériens, clones bactériens et d'amorces oligonucléotidiques

Pour les différents types de ressources présentes dans le laboratoire (graines, plasmides bactériens, clones bactériens et amorces oligonucléotidiques), nous avons choisi de créer une base de données spécifique pour chacune d'elles. Quatre bases de données ont été conçues. Chacune de ces bases partage la même structure, mais a été dissociée afin de simplifier sa conception (rubriques à renseigner différentes) et son utilisation. Seuls les boutons d'action (rechercher, supprimer, annuler, sommaire...) sont identiques pour toutes les bases (**Figure 2A**). Pour chaque base de données, le type et le niveau d'information sont différents. Chaque base possède des rubriques propres (**Figure 2-4**). L'exportation de l'ensemble de la base sous fichier texte (.csv) peut être réalisée pour créer un tableau récapitulatif. L'importation de données dans la base peut être également accomplie avec un fichier texte. Ces bases disposent de plusieurs modes de vue : détaillé, sommaire, vue boîte (**Figure 2**). Le passage d'un mode à l'autre se fait par l'utilisation de boutons d'action.

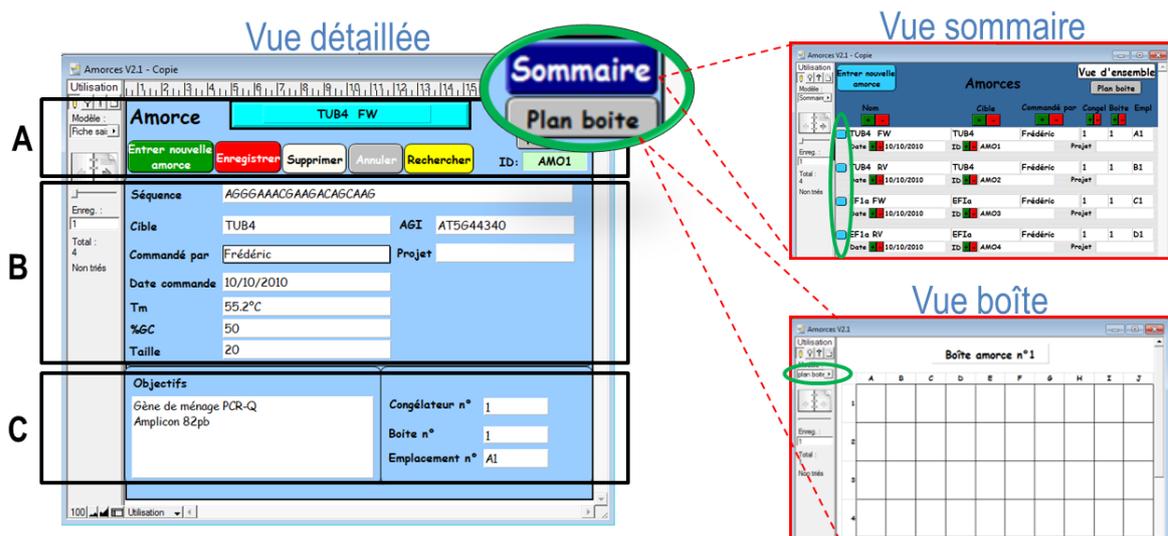


Figure 2. Présentation générale d'une base de données créée avec Filemaker Pro.

Ci-dessus différentes vues de la base de données pour les amorces oligonucléotidiques. La vue détaillée (à gauche) est constituée de : (A) les boutons d'action, (B) les rubriques informations, (C) les objectifs de fabrication et la localisation de l'échantillon. La vue sommaire (en haut à droite) présente l'ensemble des échantillons enregistrés sous forme de liste. La vue boîte (en bas à droite) correspond au plan de localisation des échantillons dans leur boîte de stockage. Le passage entre les trois vues se fait à l'aide de boutons d'action (entourés en vert).

La saisie des informations concernant l'échantillon s'opère en mode « vue détaillée ». Il suffit ensuite de cliquer sur les champs et de les remplir. Un code d'identification (ID) est incrémenté automatiquement pour chaque échantillon (ici AMO1, AMO2, AMO3...). L'ID est modifiable au départ, et il est différent pour chaque base. Toutes les bases de données ont des rubriques adaptées et spécifiques à remplir. Pour la base de données des amorces, les informations essentielles telles que la séquence, le gène cible (nom et AGI), la température d'hybridation (Tm), le pourcentage en GC et la taille de l'amorce peuvent être renseignées pour chaque amorce (**Figure 2B**). Dans la rubrique « objectifs » (**Figure 2C**), il est possible d'ajouter des commentaires supplémentaires. La localisation physique de l'échantillon doit également être renseignée (**Figure 2C**) pour permettre de trouver rapidement les échantillons (voir exemple de recherche **Figure 5**).

La « vue sommaire » est utilisée pour voir de manière synthétique la liste exhaustive des échantillons présents dans la base. Seules les informations minimum (nom, gène cible, nom du propriétaire et localisation de l'échantillon) sont visibles (**Figure 2**). Le passage entre les deux vues « détaillée » ou « sommaire » est aisé grâce aux boutons d'action (**Figure 2**). La « vue boîte » correspond au plan de boîte qui contient les échantillons stockés. Ce plan de boîte est un tableau Excel (.xls) incorporé dans la base. Ce tableau n'est pas connecté aux rubriques et doit être rempli indépendamment. Cette option permet d'imprimer le plan de boîte et de le coller sur le couvercle des boîtes de stockage dans le but de localiser plus rapidement l'échantillon dans son lieu de stockage.

Le remplissage des rubriques est facilité par des menus déroulants. Par exemple, dans la base de données des clones bactériens les menus déroulants des rubriques concernant les antibiotiques et les souches de bactéries sont disponibles (Figure 3).

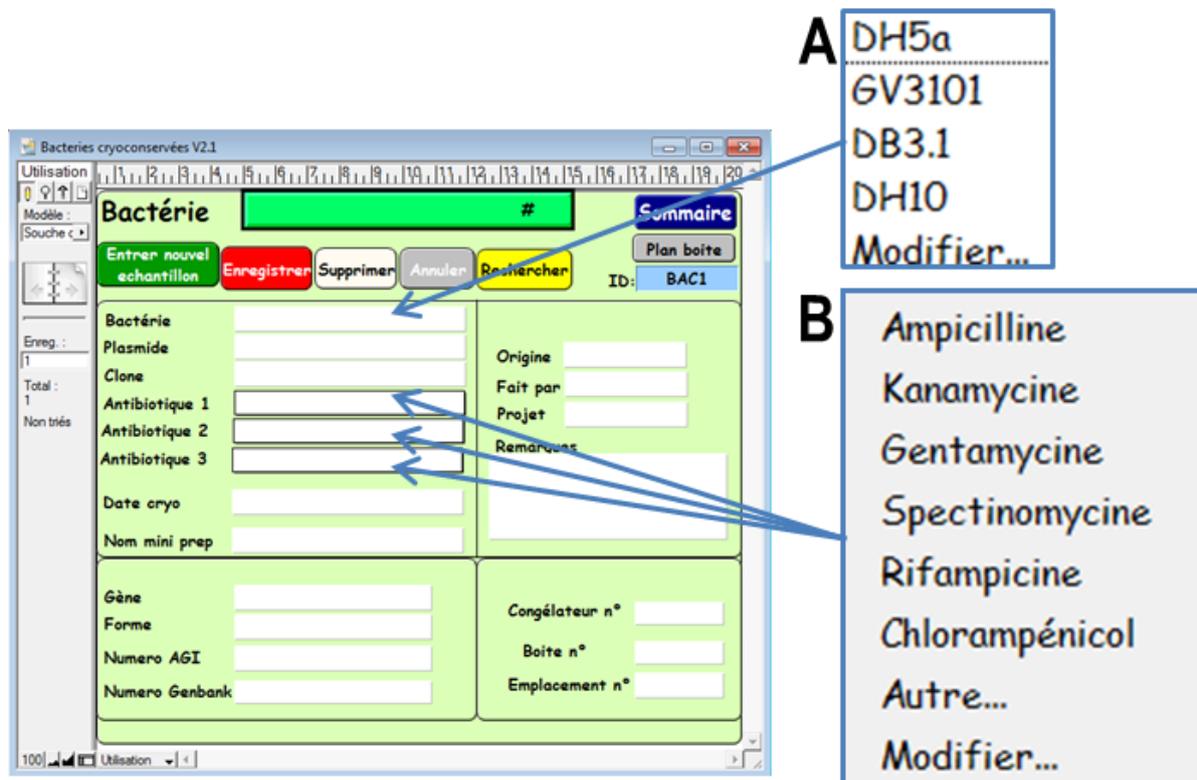


Figure 3. Détails du contenu des menus déroulants des rubriques de la base pour les bactéries.

A. Rubrique contenant les souches bactériennes les plus utilisées. B. Rubriques contenant la liste des antibiotiques auxquels les bactéries sont résistantes.

Des types d'informations supplémentaires peuvent être incorporés dans chaque base de données. Des rubriques peuvent contenir des fichiers tels que des cartes de plasmides ou des séquences nucléotidiques. Il suffit de faire glisser dans la rubrique le fichier contenant la carte ou la séquence. Les cartes et séquences peuvent être sous différents formats tels que : image (.jpg), document (.pdf) ou .xdna (format de fichier du logiciel Serial cloner freeware http://serialbasics.free.fr/Serial_Cloner.html) (Figure 4). L'incorporation de ces fichiers permet d'avoir accès rapidement à des informations supplémentaires spécifiques à l'échantillon.

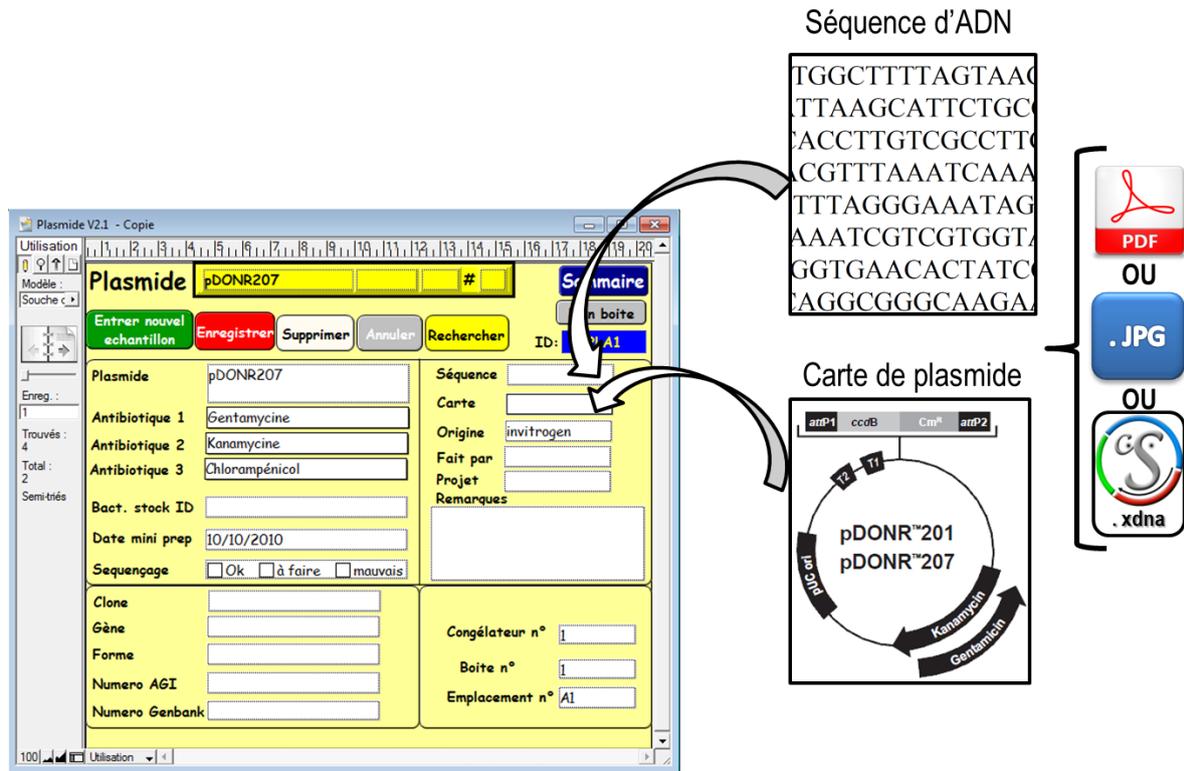
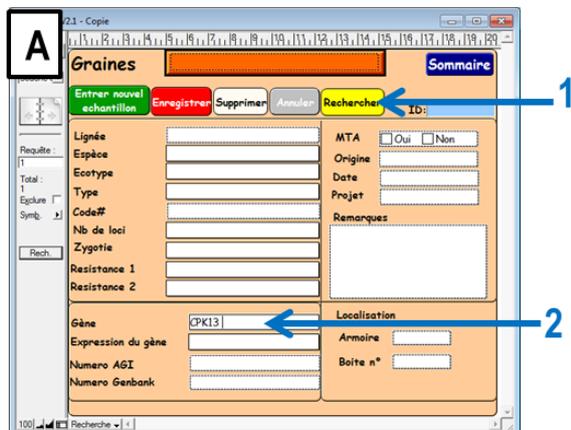


Figure 4. Incorporation de données supplémentaires dans la base.

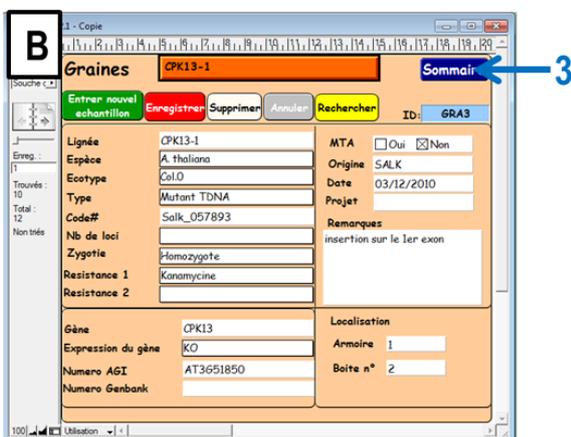
Les séquences d'ADN ou les cartes de plasmides aux formats jpg, pdf ou xdna peuvent être incorporées dans les rubriques « séquence » et « carte » en faisant glisser-déposer.

Exemple de recherche d'échantillons dans la base de données pour les graines

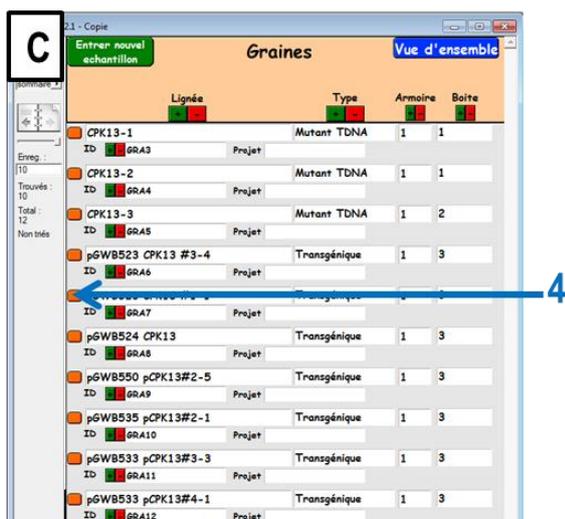
Le principal avantage d'une base de données est la recherche aisée d'un échantillon. Grâce aux renseignements fournis pour chaque échantillon dans la base, une recherche selon les différentes rubriques (nom, souche, propriétaire, etc.) est possible. De plus, les informations sur la localisation physique de l'échantillon sont rapidement accessibles. La procédure de recherche de lignées d'*Arabidopsis thaliana*, à partir de la rubrique gène, a été résumée ci-dessous pour illustrer un exemple de recherche (**Figure 5**).



A. (1) Cliquez sur le bouton d'action « Rechercher », (2) rentrez le nom du gène recherché dans la rubrique gène, validez par entrée.



B. Affichage d'un des résultats de la recherche en vue détaillée, (3) Cliquez sur « Sommaire » pour avoir une vue d'ensemble des lignées disponibles comportant le gène recherché visible en (C).



C. Visualisez les échantillons grâce à cette vue sommaire, localisez ces échantillons avec le numéro ID GRA et le numéro de boîte, (4) pour avoir plus de détails sur un échantillon cliquez sur le bouton orange foncé correspondant à la lignée souhaitée.

Figure 5. Procédure de recherche de lignées d'Arabidopsis thaliana dans la base de données graines.

Discussion

L'activité de recherche d'un laboratoire génère un grand nombre d'échantillons. Une organisation rigoureuse sous forme de bases de données en permet une gestion efficiente. Sans elles, il peut arriver que les échantillons soient inexploitable, voire non localisables à cause de la perte des informations associées. Cela peut présenter un important préjudice au niveau scientifique et un coup financier conséquent. Les bases de données permettent une centralisation efficace des informations relatives aux ressources biologiques, qui sont de plus accessibles à tous *via* un réseau. FileMaker Pro permet de créer facilement des bases de données modulables sans l'utilisation de programmation. Son principal inconvénient est d'être sous licence propriétaire (payant). Des alternatives *open sources* (gratuites) à FileMaker Pro existent tel que MySQL, cependant elles nécessitent des compétences en langage de programmation, ce qui est un frein à leur utilisation, surtout pour des biologistes.

Conclusion

Les bases de données développées avec FileMaker Pro ont considérablement amélioré l'organisation, la traçabilité et la centralisation des données au sein de l'Unité. Cela représente maintenant plusieurs milliers d'échantillons dans les bases de données. Les interrogations de la base pour rechercher des informations importantes concernant des échantillons (localisation, origine...) sont très aisées. Par ailleurs, au-delà de la gestion des ressources, cet outil a permis de fournir rapidement les informations essentielles nécessaires pour la demande d'agrément OGM (organisme génétiquement modifié). Il est aussi très utile dans le cadre de l'AQR (assurance qualité recherche) car il améliore grandement la traçabilité des échantillons. Pour ces raisons, ces bases sont maintenant utilisées par beaucoup d'équipes de recherche de l'UMR pour leurs gestions des échantillons.

Ces outils constituent donc une solution pertinente pour des équipes ou des Unités de recherche en biologie, voire même dans d'autres domaines, qui n'ont pas encore d'outil de gestion de leurs ressources.

Les bases de données vierges sont disponibles sur simple demande.

Références bibliographiques

1. <http://eric.univ-lyon2.fr/~jdarmont/docs/old/sise-bd.pdf>
2. http://fr.wikipedia.org/wiki/Base_de_donn%C3%A9es
3. http://fr.wikipedia.org/wiki/FileMaker_Pro