Excel® : importer, mettre en forme et manipuler des données de génotypage microsatellites produites par le logiciel GeneMapper® V4.1. avec l'aide de macros codées en langage Visual Basic for Application

Jérôme OLIVARES[1]

Résumé. Cet article présente les fonctionnalités d'un fichier Excel® permettant par le biais de plusieurs Macros VBA d'importer, mettre en forme et manipuler des données de génotypage issues de l'analyse faite par le logiciel GeneMapper® en Version 4.1. Les différentes fonctionnalités et outils seront exposés. Le fichier est en téléchargement libre.

Mots clés : macro VBA, Visual Basic Application, Microsoft® Office Excel®, ABI3730, GeneMapper®, génotypage microsatellites, SSR.

Introduction

L'utilisation de séquenceurs capillaires pour le génotypage, de par leur capacité de multiplexage des marqueurs génétiques analysés, permet l'acquisition d'un volume important de données. Dans le cas présent nous utilisons un séquenceur à capillaires ABI3730xl de la société AppliedBiosystems® pour le génotypage de marqueurs microsatellites, dont les données sont interprétées par le logiciel d'analyse GeneMapper® V4.1. Cette analyse se fait par le biais de fenêtres de lecture ou "BIN" au sein desquelles on considère toutes les lectures comme identiques.
Cependant les rapports d'analyses générés sont au format texte/tabulation et présentent les données sous forme d'une liste brute dont le peu de lisibilité rend difficile l'analyse qualitative des résultats. Or, lors du déroulement d'un projet, le contrôle de la qualité des données produites est indispensable afin d'ajuster les conditions expérimentales, de sélectionner les échantillons à refaire ou encore de choisir de conserver ou supprimer un marqueur génétique. Il faut donc passer par une étape de filtrer-copier-coller afin de transposer les données sous forme de tableau bien plus lisible. De plus, lors de l'analyse de locus microsatellites il est fréquent d'obtenir un pic positionné entre deux BIN, nous le rattachons alors au BIN précédent ou suivant. Pour garder une trace de cette réattribution dans nos fichiers de génotypages, nous notons en vert les valeurs rattachées au BIN inférieur et en bleu les valeurs rattachées au BIN supérieur. Les lectures douteuses sont notées en rouge et un court descriptif est inséré en commentaire. Toute cette mise en forme rajoute encore à la lourdeur du traitement des données. Une fois les résultats mis en forme et validés il faut encore les exporter dans un format compatible avec les outils statistiques nécessaires à l'exploitation et la valorisation des résultats.
Lorsqu'elles sont réalisées manuellement ces étapes sont particulièrement longues, fastidieuses et les sources d'erreurs sont multiples.

Le fichier présent a pour but d'automatiser l'import de données, la transposition sous forme de tableau, de fournir des outils pour analyser qualitativement les résultats, aider au choix de ce qu'il faut garder, jeter ou refaire et formater les données pour les rendre utilisables par le pluggins GenAlex (Peakall et Smouse 2006 et 2012) ou le logiciel Genepop (Raymond et Rousset, 1995 ; Rousset, 2008). Il est compatible avec les rapports générés par la version 4.1 de GeneMapper®.

[1] INRA, UR 1115 Plantes et Systèmes de culture horticoles, 84914 Avignon, France
jerome.olivares@avignon.inra.fr

Jérôme Olivares

Pré-requis et limitations

Le fichier, dans sa dernière version est en téléchargement libre au bas de ma page de présentation : http://www6.paca.inra.fr/psh/Equipes-de-recherche/Ecologie-de-la-Production-Integree/Les-Personnes/Jerome-Olivares
Microsoft® Office Excel® 2007 ou supérieur est indispensable à son utilisation
Il n'est pas utilisable sous LibreOffice, OpenOffice ou sous environnement émulé de type "Wine" sous Ubuntu.
Les macros utilisées sont parfaitement dynamiques : elles détectent elles mêmes les dimensions des données à gérer (**Annexe 1**).
Il convient donc d'éviter de supprimer ou ajouter des cellules manuellement dans les colonnes A, B, ainsi que les lignes N°1 et 2 afin de ne pas nuire à la détection des plages de travail ou au fonctionnement des options.
Le nom des feuilles de calculs ne doit pas être modifié.

Préambule : format des données et export depuis GeneMapper® V4.1

Une des fonctionnalités particulièrement intéressante du logiciel de lecture GeneMapper® est la possibilité de fournir des résultats sous forme de valeurs numériques et/ou de texte. Nous utilisons cette fonctionnalité pour intégrer, lorsque cela est nécessaire, nos informations de couleurs et de commentaires directement aux lectures pour qu'elles soient traitées ensuite par les macros.

L'option "Custom" du menu d'attribution des tailles d'allèles permet cette intégration (**Figure 1**): l'information "couleur" est la première à noter suivi d'un espace puis de la taille à proprement dite. Les couleurs programmées pour être reconnues sont rouge, vert, et bleu, si rien n'est précisé (la majorité des cas) c'est la couleur standard noire qui est appliquée. Si on veut adjoindre un commentaire, il suffit de rajouter un espace et "com:" suivi du texte souhaité. Une des macros se chargera plus tard de changer la couleur de la police de caractère des lectures en fonction de la couleur que nous lui avons attribuée et intègrera tout le texte placé à droite de "com:" dans un commentaire de cellule. Enfin nous identifions les échantillons contaminés avec le mot "contamination" dans le commentaire.

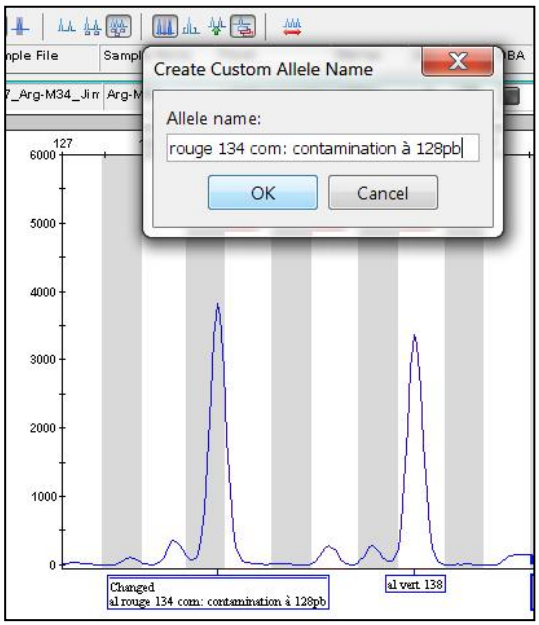

Figure 1. *L'option "Create Custom Allele Name" permet d'intégrer différentes informations que l'on peut récupérer par la suite à l'aide d'une macro VBA adaptée.*

Notes : si un décalage est fréquent pour une taille et qu'une attribution a été décidé on peut créer un BIN supplémentaire entre 2 BINs standards et lui donner la valeur "*vert 138*" par exemple, pour automatiser l'annotation et alléger le processus de lecture.

Limitations : pour être correctement mis en forme il faut respecter les espaces séparant les éléments "couleur", "lecture" et "com:".

Impératif : une fois les lectures terminées, le rapport doit être exporté au format "Texte/tabulation" depuis GeneMapper® avec les tous les champs suivants et en respectant impérativement l'ordre suivant (**Annexe 2**):

Marker / Sample File / Allele 1 / Allele 2 / Sample Name / Comments / UD2 / UD3 / Peak 1 / Allele 1-ht / Peak 2 / Allele 2-ht / Run name

Description des fonctionnalités

Importation du rapport de génotypage

Cette importation se fait depuis la feuille *"données brutes"* qui est destinée à accueillir les données brutes exportées; un bouton "Menu" donne accès à deux options.

L'option "*Importer un rapport au format GeneMapper®*" vide complètement la feuille et ouvre l'explorateur de fichiers afin d'aller chercher le rapport de génotypage préalablement exporté depuis GeneMapper®. Une fois sélectionné, après vérification, si tous les champs requis sont bien présent et dans le bon ordre, toutes les données sont copiées sur la feuille "*données brutes*". Il s'effectue alors sur les données brutes une mise en forme sommaire qui consiste à remplacer toutes les valeurs de lecture vide par un zéro, remplacer tous les "points textes" des colonnes I,J,K,et L par un "point numérique" pour considérer les valeurs comme des nombres et pas du texte et enfin trier la feuille par Marker croissant.

A l'issu de l'import une boîte de dialogue propose de créer un rapport de génotypage brut. Dans l'affirmative la macro transfère les données sous forme de tableau sur la feuille "*Genotypage_brut*", dans la négative elle s'arrête et l'utilisateur reste sur la feuille "*données brutes*" pour importer un rapport supplémentaire par exemple.
L'option "*Importer un rapport supplémentaire*", répète l'opération précédente mais ne vide pas la feuille.
Les données sont ajoutées à la suite de celles déjà importées. Cette fonction permet d'importer différents rapports issus de différents manipulateurs, projets ou de différentes périodes par exemple.

Notes : le champ "*Comments*", est réservé aux informations de type "population", il sera renommé "POP" par la suite, les champs "*UD2*" *et* "*UD3*" sont des informations supplémentaires liées à l'échantillon dont le titre et le contenu sont à la convenance de l'utilisateur en fonction du projet traité (sexe, poids…), les titres des autres champs ne doivent pas être modifiés.

Limitations : les données importées doivent être uniques, en aucun cas il ne faut importer un même RUN plusieurs fois. Si deux lectures d'un même RUN (même "Run Name") sont importées, elles ne seront pas fusionnées, seule la première des deux importées sera considérée.

Localisation des 0, des contaminations et des lectures "rouges" sur un plan de plaque

La feuille "*map*" permet d'obtenir la position sur un plan de plaque pour chaque locus microsatellite ou Marker des échantillons dont la lecture est égale à 0, jugée douteuse et précédée par un "rouge", ou identifiée comme contaminée par le mot "*contamination*" dans un commentaire.

La finalité est d'identifier les éventuelles erreurs de manipulations ou problèmes techniques qui conduisent à des schémas d'échecs suspects comme une ligne ou une colonne entière d'échantillons contaminés ou non amplifiés, chose assez difficile à détecter lors de la lecture ou du traitement manuel des données.
Une macro liste les Run Names et les Markers présent sur la feuille "*données brutes*", trace un plan de plaque pour chaque. Puis pour chaque échantillon, si la lecture est égale à "0", applique un fond jaune à la position de l'échantillon sur le plan de plaque correspondant. Si le mot "contamination" est dans la lecture un fond mauve sera appliqué, et si le mot "rouge" est présent c'est un fond rouge qui est appliqué. Dans les autres cas de lecture, la couleur appliquée est le blanc. Les positions grisées (#NA) correspondent aux cas où le traçage est Non Applicable, dans le cas où le Marker n'est pas présent à la position (**Figure 2**). Ces cas de figures se présentent lors d'une migration d'une demi-plaque, si on a plusieurs Multiplex sur une même plaque ou si des données sont manquantes car non analysées ou non exportées.

Impératif : pour être fonctionnelle la position de l'échantillon doit être notée au début des "Sample Files", "*C07_xxxxxx.fsa*", "*B12_xxxxxx.fsa*", etc. Ce paramètre se règle depuis le "Results Group Editor" (**Annexe 3**).

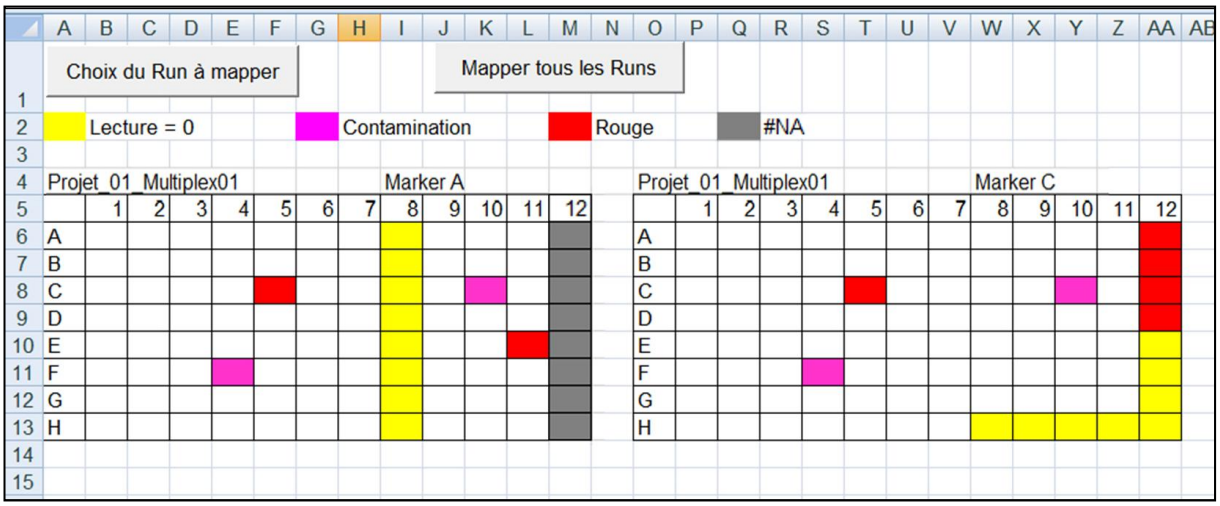

Figure 2. Les échantillons en position F04 et C10 de la plaque "Projet-01_Multiplex01" ont été identifiés contaminés pour les Markers A et C. La disposition des "0" laisse supposer un problème de pipetage (Marker A) ou d'évaporation sur un coin de la plaque (Marker C). Aucune lecture pour le Marker A n'est disponible pour les échantillons de la colonne 12.

Deux boutons sont présents sur la feuille: "*Choix du Run à mapper*" ouvre une liste déroulante qui répertorie tous les "Run Names" contenus dans la feuille "*données brutes*" et permet d'analyser tous les Markers du Run sélectionné. Le bouton "*Mapper tous les Runs*" analyse tous les Markers de tous les Runs.

Notes : le nom de Run est, dans notre cas, composé du nom de plaque (variable) suivi d'une partie fixe de 29 caractères (date-opérateur...). Cette partie fixe est supprimée par la macro pour ne restituer que le nom de plaque. Si les noms de Run ne sont pas formatés de la même manière ils peuvent être tronqués. (**Annexe 3**).

Limitations : ne considère pas les lectures "0" qui ont un commentaire du type " 0 com : …. ".

Représentation graphique des tailles

La feuille "*Graphique*" permet d'obtenir un graphique de la taille de toutes les lectures pour un même Marker. Le but étant d'avoir une représentation de toutes les tailles réelles des lectures indépendamment du BIN. Cela peut permettre d'aider à la prise de décision concernant la lecture des allèles hors BIN ou la définition même de la position des BINs. Si on trie les données brutes par date de Run on peut détecter les éventuelles déviations expérimentales au cours du temps (**Figure 3**).

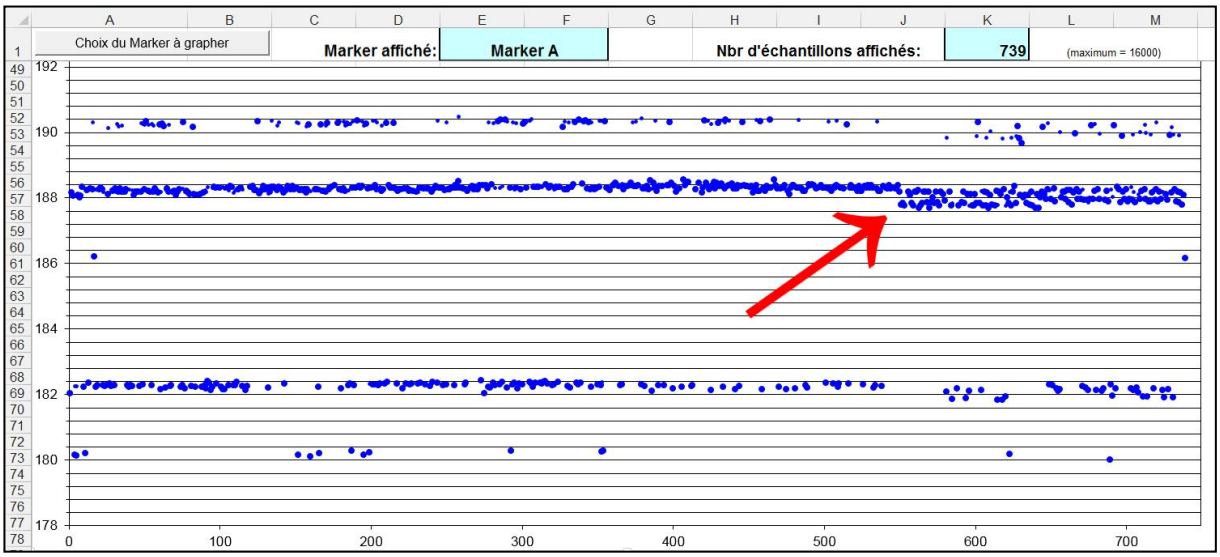

Figure 3. La flèche rouge montre un décrochage très net dans l'estimation des tailles aux alentours de 188 pb pour les 180 derniers individus analysés.

Un bouton "*Choix du locus à grapher*" permet de lister tous les Markers présent dans la feuille "*données brutes*". Lorsque l'on choisit un Marker la macro filtre les échantillons qui ont une valeur positive dans la colonne "*Peak 1*" et "*Peak 2*" pour le Marker sélectionné et alimente le graphique avec. Si on sélectionne un autre Marker, le filtre s'actualise et le graphique se met à jour. Un bouton "tri par taille" est aussi présent, et permet de trier tous les allèles du Marker sélectionné par taille croissante, et ainsi de visualiser différemment les groupes d'allèles identiques.

Note : les tailles affichées sur l'axe Y sont les tailles réelles Peak 1 et Peak 2, ce ne sont pas les valeurs de BIN. L'axe X correspond au nombre d'échantillons intégrés dans le graphique.

Limitations : le graphique n'affiche au maximum que les 16000 premiers échantillons du Marker choisi.

Rapport de génotypage brut

La feuille "*Génotypage_brut*" est destinée à recevoir les données brutes sous forme de tableau. C'est plutôt une feuille de travail permettant d'analyser rapidement les données au fur et à mesure de leur acquisition et de choisir des ajustements à faire ou des échantillons à ré-analyser avec l'aide des outils décrits plus loin.

Les données y sont intégrées de deux manières : soit à la fin de l'étape d'importation des données depuis la feuille *"données brutes"* soit par l'option "*Créer un nouveau rapport de génotypage*" disponible en cliquant sur le bouton menu. Pour éviter tout risque d'erreur, le travail s'effectue sur le "Sample File" qui est spécifique d'un échantillon situé à une position dans la plaque pour un Run et un multiplex de Markers donné.
Un ensemble de macros liste et colle sur la feuille tous les Markers puis tous les "Sample files" avec les "Sample Name", "Comments" (renommé POP), "UD2" et "UD3" qui y sont liés, délimitant ainsi un tableau. Une fois le tableau créé, toutes les lectures sont transférées aux coordonnées "Sample File/Marker" correspondantes. Chaque échantillon est caractérisé par autant de "Sample File" qu'il y a de Multiplex ou de reprises et donc autant de lignes dans le tableau, on obtient alors cet aspect décalé dans le remplissage des données (**Figure 4**).

	A	B	C	D	E	T	U	V	W	X	Y	Z	AA
1	Menu					Marker A		Marker B		Marker C		Marker D	
2	Sample Name	Sample File	POP	UD2	UD3	Allèle 1	Allèle 2	Allèle 1	Allèle 2	Allèle 1	Allèle 2	Allèle 1	Allèle 2
3	Echantillon 001	F05_Echantillon 001_Multi1.fsa	POP A	M	Gros			228	232			304	304
4	Echantillon 002	F06_Echantillon 002_Multi1.fsa	POP A	M	Gros			0	0			304	304
5	Echantillon 003	F07_Echantillon 003_Multi1.fsa	POP A	F	Petit			0	0			304	304
6	Echantillon 004	F08_Echantillon 004_Multi1.fsa	POP A	F	Petit			222	226			304	304
7	Echantillon 005	F09_Echantillon 005_Multi1.fsa	POP A	M	Gros			222	242			304	304
8	Echantillon 006	F10_Echantillon 006_Multi1.fsa	POP A	M	Gros			238	238			304	304
9	Echantillon 001	F05_Echantillon 001_Multi2.fsa	POP A	M	Gros	134	138			297	297		
10	Echantillon 002	F06_Echantillon 002_Multi2.fsa	POP A	M	Gros	132	134	contamination à 128pb		291	291		
11	Echantillon 003	F07_Echantillon 003_Multi2.fsa	POP A	F	Petit	134	134			289	291		
12	Echantillon 004	F08_Echantillon 004_Multi2.fsa	POP A	F	Petit	134	138			291	291		
13	Echantillon 005	F09_Echantillon 005_Multi2.fsa	POP A	M	Gros	136	136			289	291		
14	Echantillon 006	F10_Echantillon 006_Multi2.fsa	POP A	M	Gros	134	136			291	291		

Figure 4. *Les Markers A et C appartiennent à un même Multiplex, les Markers B et D à un autre. Les données des échantillons 1 à 6 sont réparties par multiplex. Une fois la mise en forme appliquée les valeurs prennent la couleur choisie et les commentaires sont insérés.*

Un bouton "Menu" est présent sur la feuille et offre un certain nombre d'outils :
"*Créer un nouveau rapport de génotypage*" : vide la feuille et transfère les données depuis la feuille "données brutes", pour mettre à jour les données, des corrections ou avant un transfert vers la feuille "*Génotypage final*".
"*Mise en forme des données*": applique sur chaque valeur du tableau la mise en forme choisie lors de la lecture en changeant la couleur de police des valeurs en vert, bleu ou rouge et en insérant le cas échéant les commentaires correspondants.
"*Tri par ordre croissant…*": permet de trier tout le tableau en fonction des colonnes A, C, D ou E au choix.
"*Filtrer et compter …* ": filtre toutes les lignes contenant une valeur d'allèle égale à zéro, rouge ou les deux, au choix, et donne en bout de ligne le nombre de valeurs concernées. Permet de compter le nombre d'échantillons

en échec pour chaque Marker et de calculer un taux d'échec qui est reporté en bas de tableau. Permet enfin, si on utilise un filtre manuel, sur un allèle donné par exemple, d'obtenir le nombre de lignes affichées.
"*Annuler tous les filtres*": permet de supprimer tous les filtres en cours mais aussi d'afficher toutes les lignes ou colonnes que l'on a volontairement masquées.
"*Trouver les valeurs non numériques*": les valeurs de génotypage étant de type numérique, si la mise en forme s'est mal déroulée ou si des valeurs "?" correspondant à des lectures non valides sont présentes, ce bouton permet de les repérer.
"*Exporter vers Génotypage final*": permet de transférer et fusionner les données par "*Sample Name*" sur la feuille "*Genotypage_final*".

Notes : la fonction "Calculer les taux d'échecs par Marker" intègre une estimation du taux d'erreur de génotypage qui ne peut pas être estimé sur cette feuille mais qui est fonctionnel sur la feuille "*Genotypage_final*" pour des raisons décrites plus loin.

Limitations : la fonction "*Exporter vers Génotypage final*" ne peut pas s'effectuer sur des "*Sample Names*" purement numériques (1,10, 256, etc.), les noms d'échantillons doivent donc contenir au minimum une lettre. De plus, seules les valeurs sont transférées, si la mise en forme couleur/commentaire a été appliquée, celle-ci est perdue. Pour la conserver il faut effectuer le transfert sur des données brutes non mises en formes en les réimportant avant le transfert par exemple.

Rapport de génotypage fusionné final

La feuille "*Genotypage_final*" est aussi destinée à recevoir les données brutes sous forme de tableau mais cette fois-ci le transfert se fait depuis le tableau de la feuille "*Genotypage_brut*" en fonction du nom de l'échantillon et non plus du "*Sample File*". Le but est de récupérer sur une seule ligne toutes les lectures des différents Multiplex et des ré-analyses. Pour cela tous les Marker et tous les "*Samples Names*" sont listés et copiés sans doublons pour délimiter un nouveau tableau. Les échantillons listés plusieurs fois car présent dans plusieurs Multiplex ou ayant fait l'objet de reprises sont notés en "gras". Puis toutes les valeurs "*POP*", "*UD2*", "*UD3*" et toutes les lectures sont copiées aux coordonnées "*Sample name/Marker*". Si pour un échantillon donné la valeur copiée est unique elle est simplement collée; s'il existe plusieurs valeurs identiques (si il y a des reprises par exemple) la valeur est mise en "gras"; s'il existe plusieurs valeurs mais différentes, elles sont toutes collées dans la cellule et séparées les unes des autres par un "#". Si une des lectures était un "0" et pas la deuxième, le "0" est remplacée par la deuxième lecture. Pour finir la mise en forme couleurs/commentaires est à nouveau appliquée et si besoin, un fond jaune vient surligner les cellules contenant un "#" (**Figure 5**).

	A	B	C	D	E	F	G	H	I	J	K	L
1	Menu				Marker A		Marker B		Marker C		Marker D	
2	Sample Name	POP	UD2	UD3	Allèle 1	Allèle 2	Allèle 1	Allèle 2	Allèle 1	Allèle 2	Allèle 1	Allèle 2
3	**ECHANTILLON 001**	**POP A**	**M**	**Gros**	134	138	228	232	297	297	304	304
4	**ECHANTILLON 002**	**POP A**	**M**	**Gros**	**132**	**134**	222	238	**291**	**291**	**304**	**304**
5	**ECHANTILLON 003**	**POP A**	**F**	**Petit**	**134**	**134**	228#230	238#232	**289**	**291**	**304**	**304**
6	**ECHANTILLON 004**	**POP A**	**F**	**Petit**	134	138	222	226	291	291	304	304
7	**ECHANTILLON 005**	**POP A**	**M**	**Gros**	136	136	222	242	289	291	304	304
8	**ECHANTILLON 006**	**POP A**	**M**	**Gros**	134	136	238	238	291	291	304	304
9	**ECHANTILLON 007**	**POP A**	**M**	**Petit**	128	130	222	230	291	291	304	304
10	**ECHANTILLON 008**	**POP A**	**M**	**Petit**	128	128	222	222	289	297	298	304
11	**ECHANTILLON 009**	**POP A**	**M**	**Gros**	134	136	236	238	291	297	304	304
12	**ECHANTILLON 010**	**POP A**	**M**	**Petit**	132	136	222	228	289	291	301	304
13	**ECHANTILLON 011**	**POP A**	**F**	**Petit**	132	132	226	232	289	291	301	304
14	**ECHANTILLON 012**	**POP A**	**F**	**Gros**	138	140	222	230	291	297	301	304

***Figure 5**. Les valeurs des Markers A, C et D pour les échantillons 002 et 003 ont été confirmées par une deuxième analyse et sont notées en gras. La deuxième analyse du Marker B de l'échantillon 003 donne une lecture différente de la première: les deux valeurs sont collées dans la cellule séparées par un "#" et surlignées en jaune.*

Cette fonction est particulièrement utile pour les échantillons analysés plusieurs fois pour un même Multiplex, elle permet d'identifier en gras les lectures concordantes d'un Run à l'autre et les lectures différentes par un "#".
A charge pour le manipulateur d'identifier les raisons de cette différence d'interprétation.

La fonction "Calculer les taux d'échecs par Marker" évoquée précédemment, calcule le ratio Nbr de "#" sur Nbr d'échantillons repris ("#" + "gras") pour estimer un taux d'erreur de génotypage pour chaque Marker. Plus le nombre d'échantillons repris sera grand plus ce taux sera significatif. Cette fusion des génotypes n'étant pas réalisée sur la feuille "*Génotypage_brut*", ce taux n'est pas calculé sur cette feuille.

Un bouton "Menu" est présent sur la feuille et offre les mêmes fonctions de tris, calculs, et filtres que sur la feuille "*Génotypage_brut*" avec, cependant, deux boutons supplémentaires :

"Afficher les différences de lectures" : permet de filtrer les lignes du tableau contenant un "#".
"Exporter le génotypage" : permet de transférer, concaténer et mettre en forme les valeurs de génotypage au format GenAlEx ou Genepop au choix. Chaque export vide complètement la feuille de destination et transfère les données sur la feuille correspondante "*Export GenAlex*" ou "*Export Genepop*".

Notes : le pluggins GenAlEx et toute sa documentation peut être téléchargé à l'adresse suivante : http://biology.anu.edu.au/GenAlEx/Download.html. Une fois installé et intégré à Excel® il est destiné à travailler exclusivement sur les données présentes sur la feuille "*Export GenAlex*".

Limitations : la fusion des données se faisant lignes par lignes et colonnes par colonnes, celle-ci peut s'avérer assez longue et très dépendante de la puissance de calcul de l'ordinateur utilisé. Pour 1000 échantillons et 15 Markers cela va de 3 min pour une machine récente à plus de 20 min pour les plus anciennes.

Analyse des données par Genepop

A la fin de l'export lancé depuis la feuille "*Génotypage_final*" l'affichage bascule sur la feuille "*Export Genepop*", une boîte de dialogue propose de sauvegarder le rapport, si ce choix est validé, la sauvegarde s'effectue alors au format texte avec un nom et un emplacement défini par l'utilisateur. La macro propose alors d'analyser le fichier sauvegardé avec le logiciel Genepop, dans ce cas l'utilisateur sélectionne le répertoire d'installation de Genepop par le biais de l'explorateur de fichier et le logiciel se lance dans une fenêtre MSDOS en ligne de commande et travaillera sur le fichier précédemment sauvegardé (**Figure 6**).

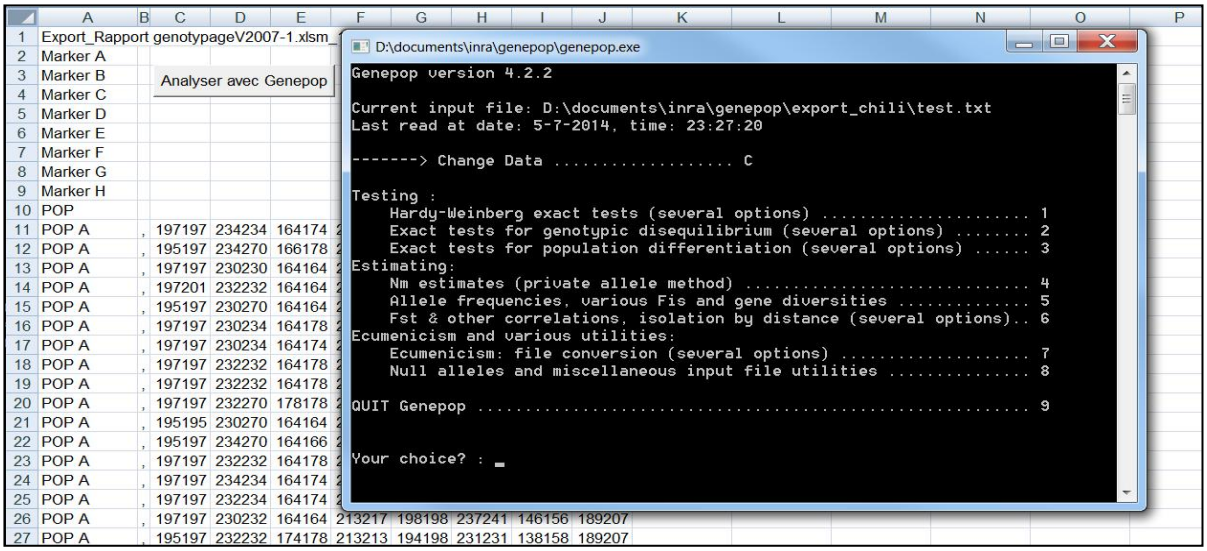

Figure 6. Le logiciel Genepop se lance en ligne de commande dans une fenêtre MSDOS à la fin de l'exportation ou en cliquant le bouton "Analyser avec Genepop" présent sur la feuille. Son utilisation reste traditionnelle.

Dès lors on peut utiliser les différentes fonctionnalités de Genepop à loisir, les fichiers d'analyses générés seront enregistrés dans le même répertoire que la sauvegarde sur laquelle s'effectuent les analyses, Excel® reste en attente pendant ces opérations.

Une fois les analyses réalisées il faut quitter Genepop pour retourner sous Excel®; pour finir il est proposé d'importer tous les fichiers d'analyse un par un sur une feuille de calcul.
Lors de cet import une mise en forme est appliquée qui peut prendre quelques minutes sur certains fichiers très volumineux. Un bouton "Analyser avec Genepop" est présent sur la feuille et permet de répéter toute la succession des opérations précédentes.
Les fichiers de sauvegardes ou d'analyses générés restent disponibles dans le répertoire choisi pour être utilisés ultérieurement.

Notes : le logiciel Genepop et toute sa documentation peuvent être téléchargés à l'adresse suivante :
http://kimura.univ-montp2.fr/~rousset/Genepop.htm
Cette fonctionnalité a été développée à partir de Genepop version 4.2.2, des différences de mise en forme peuvent être éventuellement observées sur des versions antérieures ou à venir.

Limitations : les options 2,3 et 4 du menu n°8 "*Null alleles and miscellaneous input file utilities*" ne peuvent pas être lancées selon ce procédé; il faut procéder manuellement à "l'ancienne".
L'option "*Only create genotypic contingency tables*" du menu N°2 génère un fichier contenant un très grand nombre de ligne dont la mise en forme peut s'avérer assez longue.

Conclusion

L'utilisation des macros sous Excel® a profondément changé notre manière de gérer et manipuler nos données de génotypage microsatellites. Les temps de mise en forme ont été considérablement réduits, la sécurisation des transferts indéniable. Les outils d'analyses qualitatifs nous permettent d'avoir une grande quantité d'informations et une réactivité sans précédent dans le choix de ce qu'il faut garder ou refaire. Ils facilitent certaines étapes jusque-là assez longues et rébarbatives de contrôle des erreurs de manipulation et de formatage des données pour les analyses statistiques. Les possibilités du VBA nous permettent de répondre à tous nos besoins et ont même généré une nouvelle manière de réaliser, surveiller et analyser nos expérimentations en matière de génotypage. Ce fichier dans sa forme actuelle est pensé pour être le plus dynamique possible afin d'être utilisable quelque soit le projet, le nombre de marqueurs génétiques ou le nombre d'échantillons pour peu que l'on fasse du génotypage microsatellites. Pour les autres applications de type AFLP, RFLP ou autre il faudra envisager d'adapter le code et les fonctionnalités aux nouveaux besoins. Pour les logiciels de génotypage autres que GeneMapper® il faudra aussi adapter le code pour rendre les rapports compatibles avec le mode de fonctionnement de ce fichier. Dans cette optique, le code est libre d'accès sous la responsabilité des utilisateurs.
Toute modification du code, ou utilisation d'une version antérieure ou postérieure à la version 4.1 du logiciel GeneMapper® peuvent éventuellement générer des bugs.

Enfin, ce fichier reste un outil et doit être considéré comme tel : il ne faut pas hésiter à contrôler et valider les résultats qu'il fourni et s'assurer régulièrement qu'il n'y a pas eu de déviation ou de perte entre la lecture du chromatogramme initial et le rendu final du résultat.

Remerciements

Mes remerciements à Anne Roig (URFM-Unité Ecologie des Forêts méditerranéennes) et Pierre Franck (PSH-Plantes et Systèmes de cultures horticoles) pour la pertinence de leurs remarques et corrections.

Références bibliographiques

Peakall R, Smouse PE (2012) GenAlEx 6.5 : Genetic analysis in Excel. Population genetic software for teaching and research-an update. *Bioinformatics* **28**, 2537-2539.
Freely available as an Open access article from here: http://bioinformatics.oxfordjournals.org/content/28/19/2537

Peakall R, Smouse PE (2006) GENALEX 6 : Genetic analysis in Excel. Population genetic software for teaching and research. *Mol Ecol Notes* **6**, 288-295.

Raymond M, Rousset F (1995) GENEPOP (version 1.2): Population genetics software for exact tests and ecumenicism. J. *Heredity* **86** : 248-249.

Rousset F (2008) Genepop'007 : A complete reimplementation of the Genepop software for Windows and Linux. *Mol Ecol Resources* **8** : 103-106.

Annexes

Annexe 1

Le volume des données à traiter est déterminé par détection du nombre de lignes (échantillons) et du nombre de colonnes (Marker).
La dernière ligne à gérer est définie par la dernière cellule contenant une valeur dans la colonne A (ou B), que l'on trouve grâce à l'expression: *Range("A1048576").End(xlUp).Row*
La dernière colonne à gérer est définie par la dernière cellule de la ligne N°2 contenant une valeur et que l'on trouve grâce à l'expression: *Range("XFD2").End(xlToLeft).Column*.
Cette détection des plages de travail assure le caractère dynamique du fichier et permet d'adapter automatiquement les calculs et les opérations au nombre d'échantillons ou de locus analysés.
Il est donc impératif de ne pas supprimer ou ajouter des valeurs manuellement en bas des colonnes A et B ainsi qu'en bout de la ligne N°2 pour ne pas fausser la détection des plages de travail.

Annexe 2

Le format du rapport de lecture peut se modifier dans la rubrique "Report setting editor" sous GeneMapper®.
Pour être exploitable par ce fichier le rapport doit contenir les champs de la colonne "Display Name" ordonnés comme il suit (**Figure 7**).
Lors de l'importation du rapport par le fichier, un contrôle s'effectue pour vérifier que ce format est respecté.

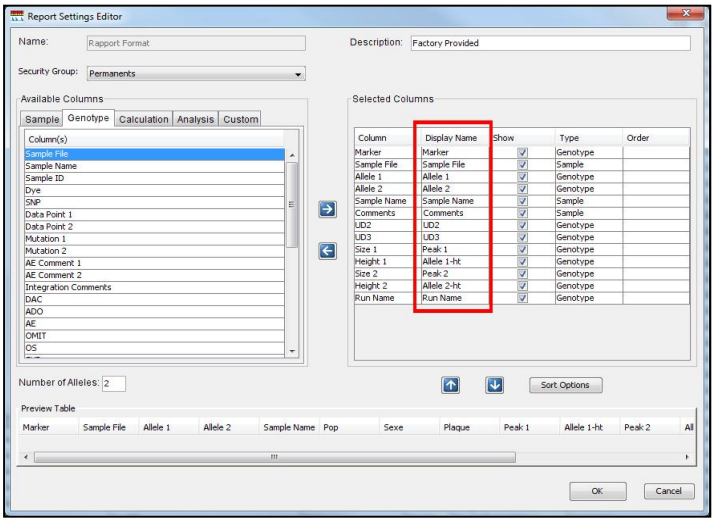

Figure 7. *Le Report Settings Editor permet de choisir et ordonner les champs qui constitueront le rapport de génotypage.*

Marker : nom du locus microsatellite.
Sample File: nom du fichier contenant le chromatogramme issu du séquenceur.
Allele 1 : valeur de l'allèle bas (BIN).
Allele 2 : valeur de l'allèle haut (BIN).
Sample Name: nom de l'échantillon.
Comments/ UD2/ UD3 : commentaires personnel ajouté lors de la création de la feuille de route du séquenceur.
Peak 1: taille réelle de l'allèle bas.
Allele 1-ht : hauteur du pic de fluorescence de l'allèle bas exprimé en "Rfu".

Peak 2 : taille réelle de l'allèle haut.
Allele 1-ht : hauteur du pic de fluorescence de l'allèle haut exprimé en "Rfu".
Run Name : nom de la plaque contenant l'échantillon, additionné souvent de la date et de l'heure de migration.

Annexe 3

Les outils graphiques de la feuille "map" ont besoin de récupérer les informations de position de chaque échantillon dans la plaque d'injection sur séquenceur.
Cette information doit être contenue dans le nom du "Sample File" et peut être paramétrée dans le "Results Group Editor" situé dans le logiciel (Data Base Collection V3.1.1) de pilotage du séquenceur (**Figure 8**).

L'onglet "Naming" permet de choisir quels types d'informations composeront le "Sample File". Dans le cas présent la position de l'échantillon est donnée par l'information "Well Position" et se doit d'être positionné en premier dans la partie gauche du nom du "Sample File" pour pouvoir être extraite par la macro.
De la même manière, le nom des Runs est un assemblage, dans le cas présent, du nom de la plaque choisi par l'opérateur (variable) et d'une partie fixe composée par la date de Run et l'identifiant de plaque. Cette partie fixe de 29 caractères est supprimée par la macro lors du traçage des plans de plaques par souci de lisibilité.

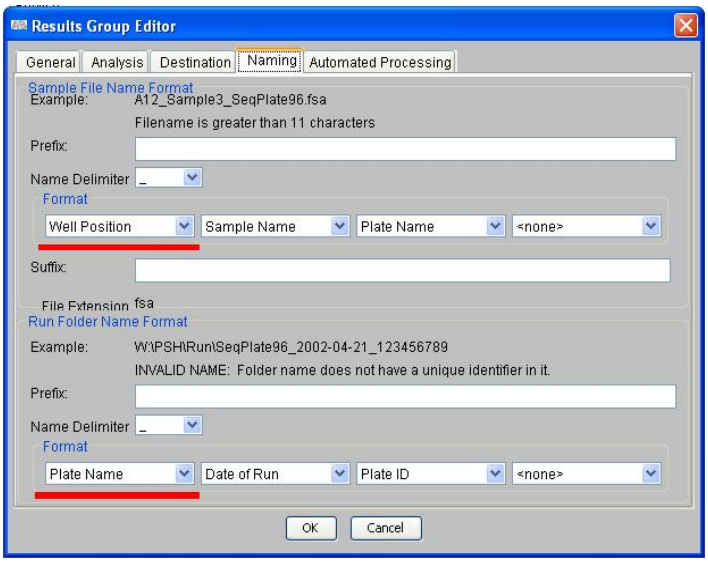

Figure 8. *Le Results Group Editor permet, entre autre, de formater le nom des "Sample File" et des "Run Names".*

Cependant si cette partie fixe fait plus ou moins de 29 caractères, le nom affiché sera tronqué.
On peut éventuellement modifier cette valeur à l'aide de l'éditeur de macros en modifiant le module "Mappage":
-à la ligne N°70 changer la valeur 29 dans la chaîne *Replace(KeyRun, Right(KeyRun, **29**), "")* par la valeur propre à chaque utilisateur ;
-répéter ce changement à la ligne N°165 pour la chaîne *Replace(SingleRun, Right(SingleRun, **29**), "")*
Sauvegarder et fermer l'éditeur de macros.