AdFiT (Admixture Files Tool)

Version 1.5

Un programme créé par Géraud GOURJON
Chercheur associé - Université Aix-Marseille II

Avec la participation de Anna DEGIOANNI
Maître de Conférences - Université Aix-Marseille II

UMR6578, Unité d'Anthropologie Bioculturelle
Aix-Marseille Université / CNRS / EFS
Faculté de Médecine, Avenue Pierre Dramard, Marseille

ggourjon@hotmail.com

Juin 2009

Cette documentation décrit les caractéristiques et l'utilisation de AdFiT, un programme destiné à la création simplifiée des fichiers d'entrée pour les logiciels d'estimation du mélange génétique
Si vous rencontrez des problèmes ou des difficultés dans l'utilisation de ce logiciel, contactez nous par email.

Référence à citer

Gourjon G, Degioanni A (2009) AdFiT v1.7 (Admixture File Tool): input files creating tool for genetic admixture estimation software. *Bulletins et Mémoires de la Société d'Anthropologie de Paris* **21**, 223-229.

Le Programme

Ce logiciel simple d'utilisation a été conçu pour aider les chercheurs dans la creation des fichiers d'entrée pour plusieurs logiciels d'estimation du mélange génétique. Il fournit les fichiers d'entrée pour les logiciels suivants :

Software	Concepteur	Méthode	Publications (methode)	Site de téléchargement
ADMIX	Long	Weighted Least Squares	(Long 1991)	Disponible sur demande au Pr. S.J. Long (longjc@umich.edu)
ADMIX95	Bertoni	Gene Identity	(Chakraborty 1985)	http://www.genetica.fmed.edu.uy/software.htm
Mistura	Cabello Krieger	Maximum Likelihood	(Krieger *et al.* 1965) (Cabello& Krieger 1997)	Disponible sur demande au Pr. H. Krieger (hkrieger@icb.usp.br)
Admix 2.0	Bertorelle Dupanloup	Coalescence-based	(Bertorelle& Excoffier 1998) (Dupanloup& Bertorelle 2001)	http://web.unife.it/progetti/genetica/Isabelle/admix2_0.html
LEA	Beaumont Langella	Coalescent-based maximum Likelihood	(Chikhi *et al.* 2001) (Langella *et al.* 2001)	http://dm.unife.it/parlea
Parallel LEA	Giovannini et al.	Likelihoob-based approach	(Chikhi *et al.* 2001) (Giovannini et al. 2008)	http://dm.unife.it/parlea
LEADMIX	Wang	Maximum likelihood	(Wang 2003)	http://www.zoo.cam.ac.uk/ioz/people/wang.htm
		Least-Square	(Roberts& Hiorns 1965)	
		Weighted Least Squares	(Long 1991) (Chakraborty& Srinivasan 1992)	
		Coalescent-based	(Bertorelle& Excoffier 1998)	

Avant toute utilisation d'AdFiT et pour remplir correctement les paramètres spécifiques de chaque logiciel d'estimation du mélange (comme par exemple la taille de l'échantillon pour ADMIX, le taux de mutation et le temps depuis le mélange pour Admix 2.0 ou encore la prise en compte de la Divergence génétique pour LEADMIX), il est fortement recommandé de lire la documentation du logiciel correspondant et la publication de référence de la méthode.

Toutes les informations sont fournies pour la version Française, référez vous à la documentation en Anglais ou en Espagnol si vous utilisez un langage différent.
Lors de l'ouverture d'AdFiT, vous pouvez sélectionner le langage de votre choix. Tous les paramètres, champs et messages d'AdFiT seront dans la langue sélectionnée.

AdFiT est un exécutable sans installation. Par conséquent lors de la première utilisation du logiciel, de nombreux fichiers sont créés dans le même dossier où se trouve l'exécutable (dll nécessaires au fonctionnement du logiciel). Ces fichiers dll doivent être présents pour que le logiciel fonctionne. L'idéal est de placer l'exe dans un dossier spécialement créé pour cela. Il fonctionne sous Windows© XP et Visa, et la version 2.0 disponible prochainement fonctionnera sous Linux.

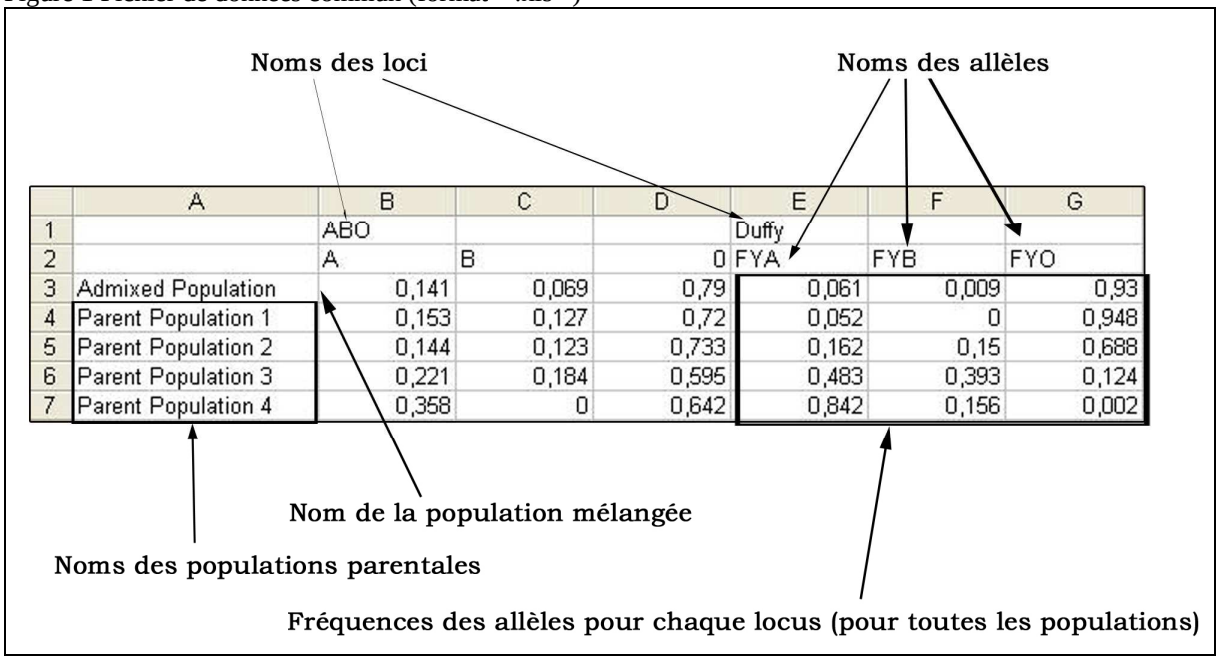
1. Creation du Fichier Commun « .xls » prealable a l'execution d'AdFiT

La première manipulation nécessite le regroupement toutes les données dans un fichier commun (figure1) dans le format « .xls » (il peut être créé avec Open Office Calc ou Microsoft Office Excel).

1ère ligne : Noms des Loci (en commençant à la deuxième colonne)
2ème ligne : Noms des Allèles pour chaque Locus (en commençant à la deuxième colonne)
3ème ligne :
1ère colonne : Nom de la population mélangée (population hybride, fille)
Colonnes suivantes : Fréquences alléliques pour chaque Locus (le total des fréquences alléliques pour chaque locus doit être égal à 1)

Ce fichier de données sera lu directement par AdFiT.
Il sera uniquement nécessaire de sélectionner sur AdFiT le logiciel d'estimation de mélange pour lequel le fichier d'entrée doit être créé et de remplir éventuellement les paramètres spécifiques du programme (voir ci dessous).

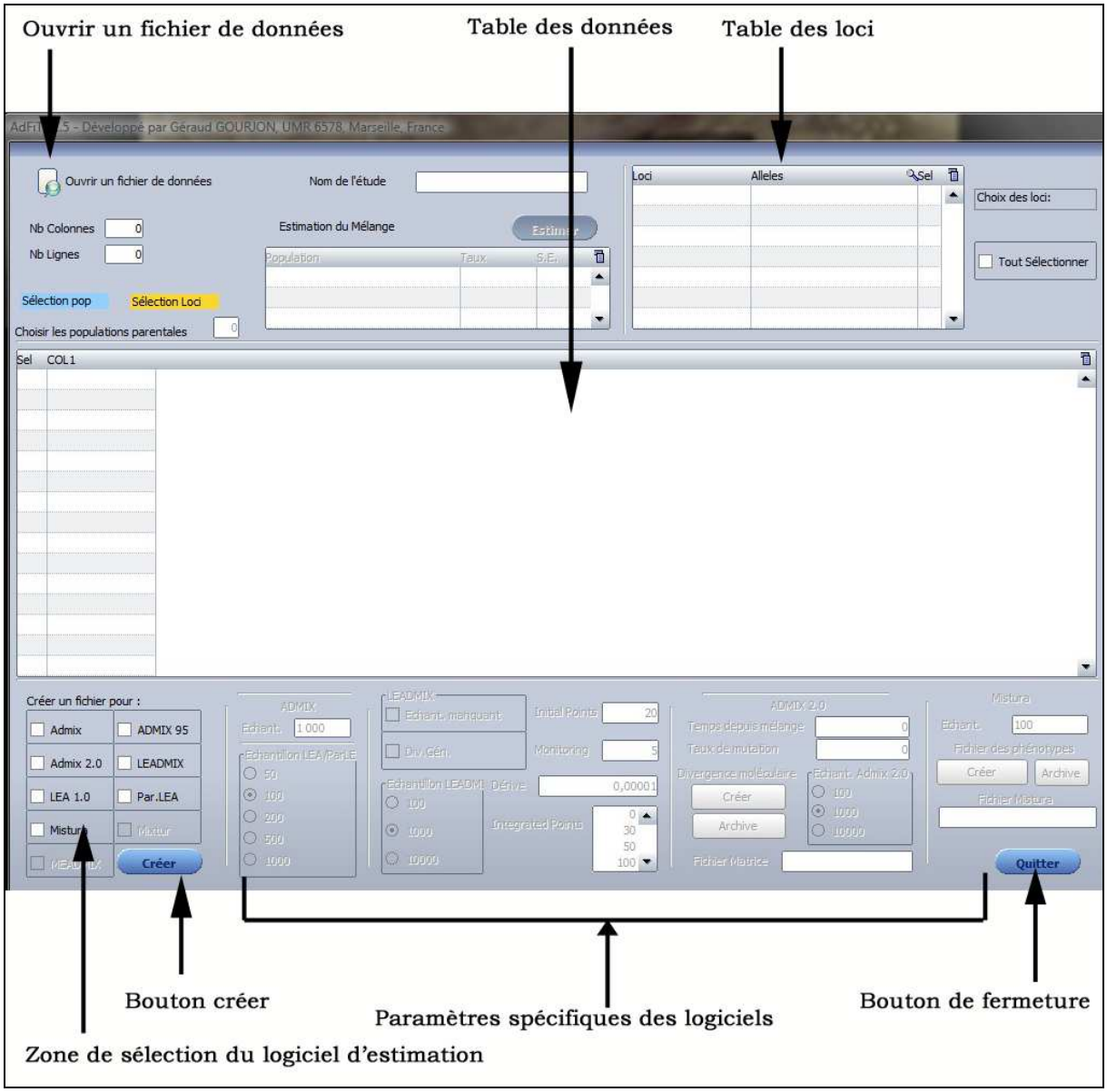
Figure 1 Fichier de données commun (format « .xls »)



	A	B	C	D	E	F	G
1		ABO			Duffy		
2		A	B	0	FYA	FYB	FYO
3	Admixed Population	0,141	0,069	0,79	0,061	0,009	0,93
4	Parent Population 1	0,153	0,127	0,72	0,052	0	0,948
5	Parent Population 2	0,144	0,123	0,733	0,162	0,15	0,688
6	Parent Population 3	0,221	0,184	0,595	0,483	0,393	0,124
7	Parent Population 4	0,358	0	0,642	0,842	0,156	0,002

2. PROCEDURE D'EXECUTION D'ADFIT

Une fois le fichier commun créé, il doit être ouvert en cliquant sur le bouton « Ouvrir un fichier de données» dans la fenêtre AdFiT. La table se rempli automatiquement dans AdFiT. La première ligne de la table des données concerne la population mélangée qui est toujours sélectionnée. La création du fichier d'entrée se fait ensuite en trois étapes :
1) choisir dans la table des données les populations parentales. Pour cela, il suffit de cocher dans la colonne « Sel » dans la table des Données : les populations parentales sélectionnées se surlignent en bleu ciel.
2) sélectionner les loci. On peut utiliser l'ensemble des infomations en cliquant sur « sélectionner tous les loci » ou bien choisir certains loci dans la dernière « Sél » de la table des loci : les loci sélectionnés se surlignent en bleu ciel dans la table des Loci et en orange dans la table des Données.
3) Sélectionner ensuite le logiciel d'estimation du mélange et remplir si nécessaire les paramètres spécifiques (référez vous ci dessous à la documentation de chaque logiciel d'estimation pour cela) puis cliquer sur le bouton « Créer ». Le fichier d'entrée pour le logiciel demandé est généré sur le bureau de votre ordinateur.



Les fichiers d'entrée sont nommés automatiquement. Si vous désirez créer plusieurs fichiers d'entrée du même type (avec des marqueurs différents, et selon le logiciel des populations parentales différentes, Cf le nom donné pour chaque logiciel) pensez à les renommer, les effacer ou bien les déplacer dans un autre dossier pour ne pas les perdre.

Une de nos priorités était de concevoir un logiciel aussi intuitif que possible. La majorité des erreurs envisageables lorsque le fichier est créé, a été anticipée et prise en compte. Des messages d'erreur ou d'information ont été introduits dans AdFiT pour expliquer les problèmes potentiels. De plus, ces risques peuvent être prévenus dans certains cas. Par exemple, LEA (Langella et al. 2001) considèrent un évènement de mélange impliquant uniquement deux populations parentales. Il sera donc impossible de générer un fichier d'entrée lorsque plus de deux populations parentales seront sélectionnées (un message d'erreur apparaît à l'écran). Inversement, quand pour un allèle donné, les fréquences pour toutes les populations sont égales à 0, la plupart des logiciels d'estimation ne fonctionnera pas. Il est tout de même possible de créer ces fichiers d'entrée avec AdFiT mais un message informatif apparaîtra pour signaler ce risque potentiel de rejet du fichier d'entrée par le logiciel d'estimation, et les allèles dont les fréquences sont égales à 0 pour toutes les populations seront surlignées en rouge dans la table des données.

2. PARAMÈTRES SPÉCIFIQUES

ADMIX95 input file

Aucun paramètre spécifique ne doit être renseigné pour ce logiciel. Le fichier d'entrée est créé automatiquement. Il est nommé par les 2 premières lettres des populations parentales sélectionnées (l'extension du fichier pour ADMIX95 est .inp)

ADMIX input file

La taille de l'échantillon doit être renseignée (voir la documentation d'ADMIX). Cela influence uniquement la valeur du F_{st}. Il est nommé par les 2 premières lettres des populations parentales sélectionnées (l'extension du fichier pour ADMIX est .dat)

LEA input file

La taille de l'échantillon doit être renseignée (voir la documentation de LEA). Ceci peut avoir un poids dans l'estimation du mélange par LEA. Augmenter la taille de l'échantillon augmente la précision des résultats mais le temps de simulation peut être fortement accru. Le fichier crée par Adfit est nommé « infile » (sans extension). Tous les fichiers sont nommés identiquement (format LEA) donc il est nécessaire de le déplacer dans un autre dossier avant de créer un nouveau fichier d'entrée (ne pas le renommer, LEA nécessite un fichier d'entrée nommé exclusivement « infile »).

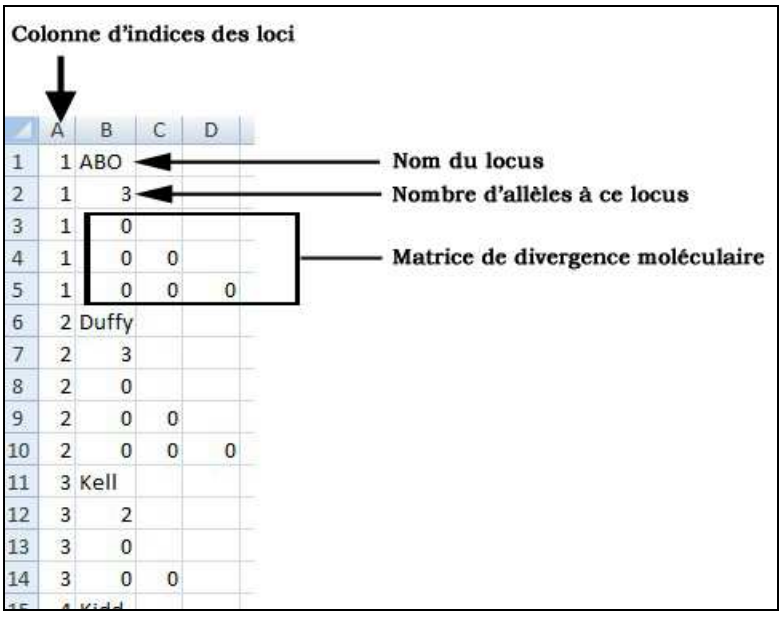
Parallel LEA input file

La taille de l'échantillon doit être renseignée (voir la documentation de Parallel LEA).
Le fichier crée par Adfit est nommé « infile » (sans extension). Tous les fichiers sont nommés identiquement (format LEA) donc il est nécessaire de le déplacer dans un autre dossier avant

de créer un nouveau fichier d'entrée (ne pas le renommer, LEA nécessite un fichier d'entrée nommé exclusivement « infile »).

Admix 2.0 input file

Pour la creation du fichier d'entrée, un fichier « .xls » intermédiaire contenant les matrices de divergences moléculaires entre les allèles pour chaque locus doit être créé. AdFiT génère ce fichier déjà partiellement prérempli avec les loci et les allèles. Pour cela, une fois le fichier de données sélectionné (et les données affichées dans la table des données), cliquez sur « Tout sélectionner » (à côté de la table des Loci), sélectionnez Admix 2.0 dans la zone de sélection du logiciel et cliquez sur « Créer » dans la zone des paramètres spécifiques d'Admix 2.0 en dessous de « Matrices de divergence moléculaire ». Ceci crée sur le Bureau le fichier « .xls » intermédiaire avec les matrices préformatées pour tous les loci. Ouvrez ce fichier avec votre tableur (Open Office Calc ou Microsoft Office Excel) et remplacez les « 0 » avec les valeurs de divergence moléculaires entre les allèles pour chaque locus et sauvegardez le fichier (utilisez « **enregistrer sous** » dans le menu et non pas « enregistrer ») en format « .xls » (format Excel 97-2003 avec Microsoft Office 2007, et Excel 07/2000/XP avec Open Office). Attention de ne pas l'enregistrer en format texte (txt), ou en format Open Office (.ods), ou en Excel 2007 (.xlsx).
Ce fichier est nécessaire pour créer le fichier d'entrée. Il doit être rempli correctement si la différenciation moléculaire est prise en compte. Dans le cas contraire, AdFiT rempli automatiquement les matrices pour tous les loci avec des 0 (considérant aucune différence moléculaire entre les allèles). Référez vous à la documentation du logiciel Admix 2.0 pour remplir correctement les matrices. Cependant, si vous utilisez Office 2007 ou Open Office Calc, il est **obligatoire** de procéder à une manipulation suite à des problèmes de compatibilité même si vous n'utilisez pas la différentiation, même si vous ne prenez pas en compte la divergence moléculaire (voir ci-dessous, la partie **IMPORTANT** plus bas)



	A	B	C	D
1	1	ABO		
2	1	3		
3	1	0		
4	1	0	0	
5	1	0	0	0
6	2	Duffy		
7	2	3		
8	2	0		
9	2	0	0	
10	2	0	0	0
11	3	Kell		
12	3	2		
13	3	0		
14	3	0	0	

Si un fichier de matrices a déjà été créé, vous pouvez le sélectionner avec le bouton « Archive » situé à côté du bouton de création des matrices. Attention : Sélectionnez un fichier contenant les matrices pour tous les loci du fichier de données, loci devant être dans le même ordre que dans la table des Loci.
Le nom du fichier sélectionné pour la création du fichier d'entrée apparaît sous ces deux boutons dans la case grisée.

Les autres paramètres doivent être également renseignés correctement (se référer à la documentation de Admix 2.0) : temps depuis le mélange (entrer 0 si ce temps est inconnu) ; taux de mutation (il doit être différent de 0 si le temps depuis le mélange a été spécifié) ; et la taille de l'échantillon.
Lorsque tous les paramètres et le fichier de matrices sont remplis, sélectionner les populations parentales et les loci pour le fichier d'entrée et cliquer sur le bouton "Création" (sous la zone de selection des logiciels d'estimation).
Il est nommé par les deux premières lettres des populations parentales sélectionnées (l'extension est .txt).

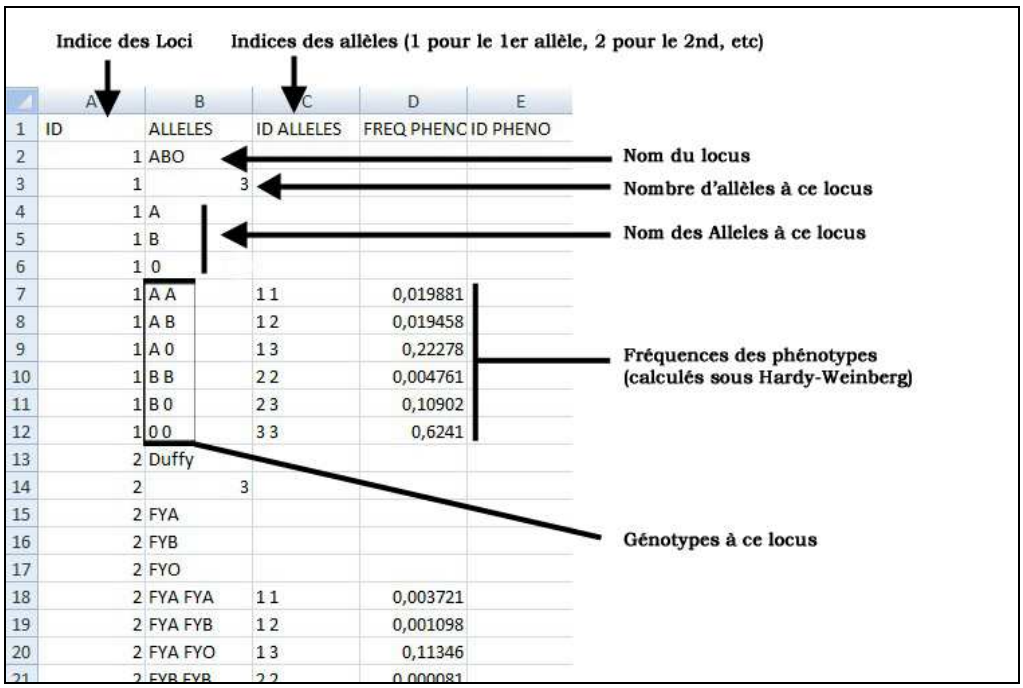
IMPORTANT: Des problèmes de compatibilité apparaissent avec Microsoft Office 2007 et Open Office Calc. Pour éviter ces problèmes, **sous Microsoft Office 2007**, pensez à ouvrir le fichier de Matrices et le sauvegarder avec « enregistrer sous » en format Excel 97-2003 avant de créer le fichier d'entrée. Si cette étape n'est pas réalisée, AdFiT ne sera pas en mesure de créer le fichier d'entrée. **Sous Open Office**, à la première ouverture du fichier, une fenêtre demandant une conversion de format apparait, cliquez simplement sur « ok » puis remplissez le fichier et sauvegardez le en « .xls ».
NOTE : Quand vous ouvrez le fichier Matrice pour la première fois avec Office 2007, un message d'erreur apparaît, cliquez sur « Oui ».

LEADMIX input file

Référez vous à la documentation de LEADMIX pour remplir sans erreurs ces paramètres.
Paramètres Spécifiques:
Echantillons manquants: Cocher cette case si les données sont manquantes pour une population parentale.
Div. Gén.: Cocher cette case si la differenciation génétique doit être prise en compte.
Taille de l'échantillon: Sélectionner une taille d'échantillon (cela ne modifie pas les resultants pour les estimateurs ponctuels).
Dérive: Entrer le taux de derive génétique.
Initial points et Int. points: Référez vous à la documentation de LEADMIX pour renseigner ces paramètres.
Monitoring: Entrer un nombre entier compris (entre 1 et 5). Ceci n'affecte pas les résultats mais permet d'afficher plus de détails lors de la simulation avec LEADMIX.

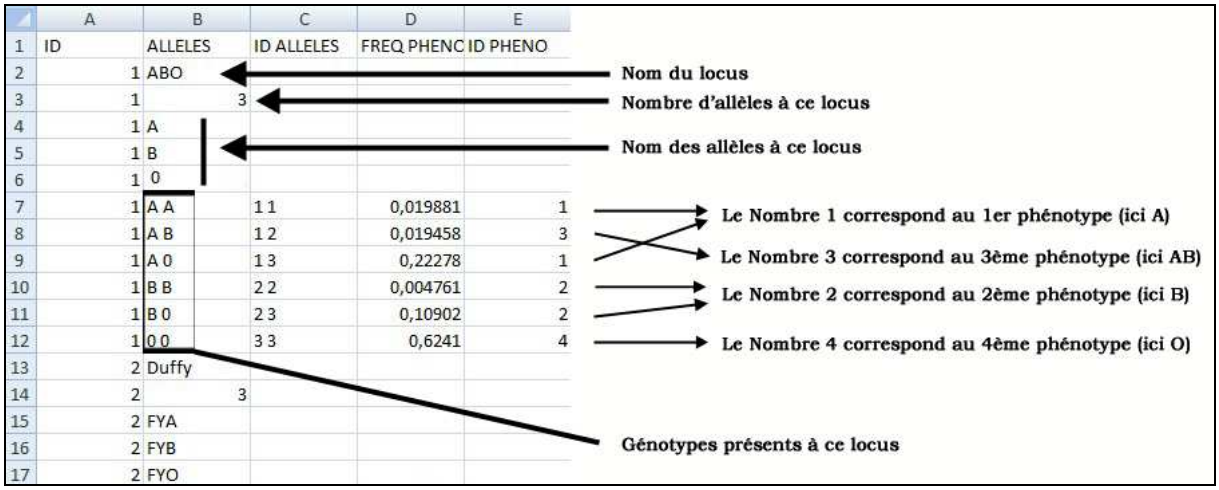
Mistura input file

Pour la réalisation du fichier d'entrée pour Mistura, il faut créer un fichier intermédiaire Il s'agit d'un fichier servant à établir la correspondance existant entre génotypes et phénotypes pour chaque locus. C'est l'utilisateur qui doit indiquer quels génotypes correspondent à chaque phénotype. Ce fichier intermédiaire est partiellement rempli. Pour cela, une fois le fichier de données sélectionné (et les données affichées dans la table des données), cliquer sur « Tout sélectionner » (à côté de la table des Loci), sélectionner Mistura dans la zone de sélection du logiciel et cliquer sur « Créer » dans la zone des paramètres spécifiques de Mistura, en dessous de « Fichier des Phénotypes ». Ceci crée sur le Bureau le fichier « .xls » intermédiaire avec les matrices préformatées pour tous les loci (voir ci dessous).



	A	B	C	D	E
1	ID	ALLELES	ID ALLELES	FREQ PHENC	ID PHENO
2	1	ABO			
3	1	3			
4	1	A			
5	1	B			
6	1	0			
7	1	A A	1 1	0,019881	
8	1	A B	1 2	0,019458	
9	1	A 0	1 3	0,22278	
10	1	B B	2 2	0,004761	
11	1	B 0	2 3	0,10902	
12	1	0 0	3 3	0,6241	
13	2	Duffy			
14	2	3			
15	2	FYA			
16	2	FYB			
17	2	FYO			
18	2	FYA FYA	1 1	0,003721	
19	2	FYA FYB	1 2	0,001098	
20	2	FYA FYO	1 3	0,11346	
21	2	FYB FYB	2 2	0,000081	

Ouvrir ce fichier avec Open Office Calc ou Microsoft Office Excel (sous Office 2007 un message d'erreur apparait, cliquer sur Oui ; sous Open Office un message de conversion de format apparait, cliquer sur OK).
La colonne ID PHENO doit être remplie comme suit : entrer le chiffre 1 en face de chaque Génotype correspondant au Phénotype 1 ; entrer le chiffre 2 en face de chaque Génotype correspondant au Phénotype 2, etc. Les chiffres utilisés doivent être des entiers et se suivre. Par exemple, pour 4 différents phénotypes, ils seront notés 1,2,3,4 et non 1,2,4,6. Voir l'exemple ci dessous pour le Groupe Sanguin ABO.
Ensuite, sauvegardez le fichier (utilisez « **enregistrer sous** » dans le menu et non pas « enregistrer ») en format « .xls » (format Excel 97-2003 avec Microsoft Office 2007, et Excel 07/2000/XP avec Open Office). Attention de ne pas l'enregistrer en format texte (txt), ou en format Open Office (.ods), ou en Excel 2007 (.xlsx). Cette manipulation est importante et obligatoire suite à des problèmes de compatibilité (voir ci-dessous, la partie **IMPORTANT**). Contrairement à lors de la création de fichier pour Admix2.0, cette manipulation est une étape de remplissage strictement obligatoire du fait de la nécessité de renseigner la colonne ID PHENO.



	A	B	C	D	E
1	ID	ALLELES	ID ALLELES	FREQ PHENC	ID PHENO
2	1	ABO			
3	1	3			
4	1	A			
5	1	B			
6	1	0			
7	1	A A	1 1	0,019881	1
8	1	A B	1 2	0,019458	3
9	1	A 0	1 3	0,22278	1
10	1	B B	2 2	0,004761	2
11	1	B 0	2 3	0,10902	2
12	1	0 0	3 3	0,6241	4
13	2	Duffy			
14	2	3			
15	2	FYA			
16	2	FYB			
17	2	FYO			

Si un fichier de Phénotypes a déjà été créé, vous pouvez le sélectionner avec le bouton « Archive » situé à côté du bouton de création. Attention : Sélectionnez un fichier contenant

les phénotypes pour tous les loci du fichier de données, loci devant être dans le même ordre que dans la table des Loci. Le nom du fichier sélectionné pour la création du fichier d'entrée apparaît sous ces deux boutons dans la case grisée.
Quand la taille de l'échantillon et le fichier de phénotypes sont remplis ; cliquer sur le bouton "Création" (sous la zone de selection des logiciels d'estimation).
Il est nommé par « Mis » + la première lettre des populations parentales sélectionnées (l'extension est .txt).
Mistura nécessite dans le fichier d'entrée le nom de l'étude. Ce nom correspond par défaut au nom de la population mélangée mais il peut être modifié dans le champ « Nom de l'étude » de la fenêtre d'AdFiT.

IMPORTANT: Des problèmes de compatibilité apparaissent avec Microsoft Office 2007 et Open Office Calc. Pour éviter ces problèmes, **sous Microsoft Office 2007**, pensez à ouvrir le fichier de Phénotypes et le sauvegarder avec « enregistrer sous » en format Excel 97-2003 avant de créer le fichier d'entrée. Si cette étape n'est pas réalisée, AdFiT ne sera pas en mesure de créer le fichier d'entrée. **Sous Open Office**, à la première ouverture du fichier, une fenêtre demandant une conversion de format apparait, cliquez simplement sur « ok » puis remplissez le fichier et sauvegardez le en « .xls ».
NOTE : Quand vous ouvrez le fichier Phénotype pour la première fois avec Office 2007, un message d'erreur apparaît, cliquez sur « Oui ».

3. AUTRES PARAMETRES

Nom de l'Etude

Ce champ est rempli automatiquement par AdFiT avec le nom de la population mélangée. Vous pouvez modifier ce nom directement dans le champ (évitez de préférence les noms trop complexes car ce nom est ensuite utilisé dans le fichier Mistura).

Estimation du Mélange

Cette function permettra d'estimer les taux de mélange selon la methode basée sur les distances génétiques de Cavalli-Sforza (Cavalli-Sforza *et al.* 1994). Actuellement cette fonction est en test et désactivée. Elle sera activée dans la version 2.0 qui sera disponible prochainement.

4. TABLES DE DONNÉES

Toutes les tables générées par AdFiT (Table des Loci, Table de Données et dans la v2.0, la Table des Taux de Mélange) peuvent être directement exportées dans un format Excel ou Word ou peuvent être imprimées avec un clic droit sur la table voulue.

Caractéristiques Techniques et prérequis

Nom du Projet	AdFiT (Admixture Files Tool)
Version	1.5
Langage	Français, Anglais, Espagnol
Site Web	http://www.anthropologie-biologique.cnrs.fr/recherche/axe1equipe3.php
Système d'Exploitation	Dépendant, nécessite Windows (9X/XP/Vista)
Licence	L'exécutable est libre d'accès pour un usage personnel (en citant la publication de reference)

Bibliographie

Bertorelle G, Excoffier L (1998) Inferring admixture proportions from molecular data. *Molecular Biology and Evolution* **15**, 1298-1311.

Cabello PH, Krieger H (1997) GENIOC: Sistema para análises de dados de genética. FIOCRUZ, Rio de Janeiro.

Cavalli-Sforza L, Menozzi P, Piazza A (1994) *The History and Geography of Human Genes* Princeton University Press, Princeton, New Jersey.

Chakraborty R (1985) *Gene identity in racial hybrids and estimation of admixture rates* Indian Anthropological Association, Delhi.

Chakraborty R, Srinivasan MR (1992) A modified best-maximum likelihood estimator of line regression with errors in both variables - An application for estimating genetic admixture. *Biometrical Journal* **34**, 567-576.

Chikhi L, Bruford MW, Beaumont MA (2001) Estimation of admixture proportions: A likelihood-based approach using Markov chain Monte Carlo. *Genetics* **158**, 1347-1362.

Dupanloup I, Bertorelle G (2001) Inferring admixture proportions from molecular data: Extension to any number of parental populations. *Molecular Biology and Evolution* **18**, 672-675.

Krieger H, Morton NE, Mi MP, *et al.* (1965) Racial admixture in North-Eastern Brazil. *Annals of Human Genetics* **29**, 113-125.

Langella O, Chikhi L, Beaumont MA (2001) LEA (likelihood-based estimation of admixture): a program to estimate simultaneously admixture and time since the admixture event. *Molecular Ecology Notes* **1**, 357-358.

Long JC (1991) The Genetic Structure of Admixed Populations. *Genetics* **127**, 417-428.

Roberts DF, Hiorns RW (1965) Methods of analysis of genetic composition of a hybrid population. *Human Biology* **37**, 38-&.

Wang JL (2003) Maximum-likelihood estimation of admixture proportions from genetic data. *Genetics* **164**, 747-765.