AdFiT (Admixture Files Tool)

Version 1.5.0

A Program developped by Géraud GOURJON (PhD)
Post-doctoral researcher - University of Aix-Marseille II

With the participation of Anna DEGIOANNI
Maître de Conférences - University of Aix-Marseille II

UMR6578, Unité d'Anthropologie Bioculturelle
Aix-Marseille University / CNRS / EFS
Faculty of Medecine, Avenue Pierre Dramard, Marseille

ggourjon@hotmail.com

June 2009

This document describes the features and use of AdFiT, a computer program for a simplified creation of input files for admixture proportions estimation software.
In case of any trouble using AdFiT, please contact us.

Reference to quote:
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.

The Program

This easy-to-use software has been designed to help researchers in the creation of input files for several admixture estimation software programs. It provides input files for the following software:

Software	Author	Implemented methods	Method references	Download websites
ADMIX	Long	Weighted Least Squares	(Long 1991)	Available on request to its conceptor, 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)	Available on request to its conceptor, 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://www.rubic.rdg.ac.uk/~mab/software.html
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)	

Before using AdFiT and to fill correctly software specific parameters for each admixture estimation software (like sample size for ADMIX, mutation rate and time since admixture for Admix 2.0, or if genetic divergence has to be accounted for or not for LEADMIX), it is highly recommended to read the appropriate admixture estimation software documentation and the implemented method reference.

All information are given for the English version (refer to the documentation in French or in Spanish if you use a different language). You can select the language in the first window of AdFiT. All parameters and fields in AdFiT window is in the selected language.

AdFiT is an auto-extracted executable without any installation. By the way, when you launch it, it will create several dll files, which are necessary for its functionalities. These files are used by AdFiT and it is recommended to create a specific folder to put AdFiT.

Input File creation

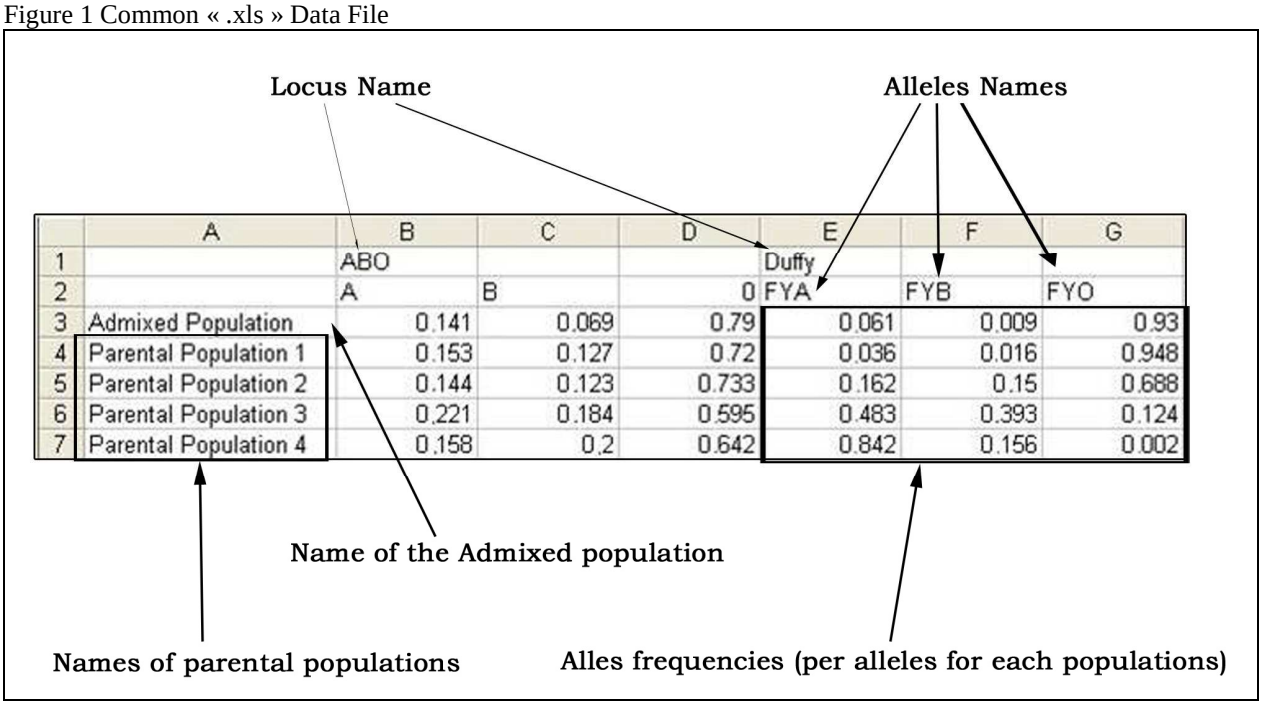
1. CREATION OF THE PRELIMINARY COMMON XLS FILE BEFORE TO RUN ADFIT

The first procedure consists to gather all data in a common spreadsheet file (figure 1). The actual used format is ".xls" (you can create it with Open Office Calc or with Microsoft© Office Excel).

1st line: Name of loci (starting to the 2nd column)
2nd line: Names of Alleles for each locus (starting to the 2nd column)
3rd line:
    1st column: Name of the admixed population
    Next columns: Alleles frequencies for each locus (total for each locus must be equal to 1)
Next lines:
    1st column: Name of the parental populations
    Next columns: Alleles frequencies for each locus (total for each locus must be equal to 1)

This file will be directly read through AdFiT.
It will only be necessary to select in AdFiT window, the admixture estimation software for which the input files is required and to fill specific parameters for this software (see below).

Figure 1 Common « .xls » Data File

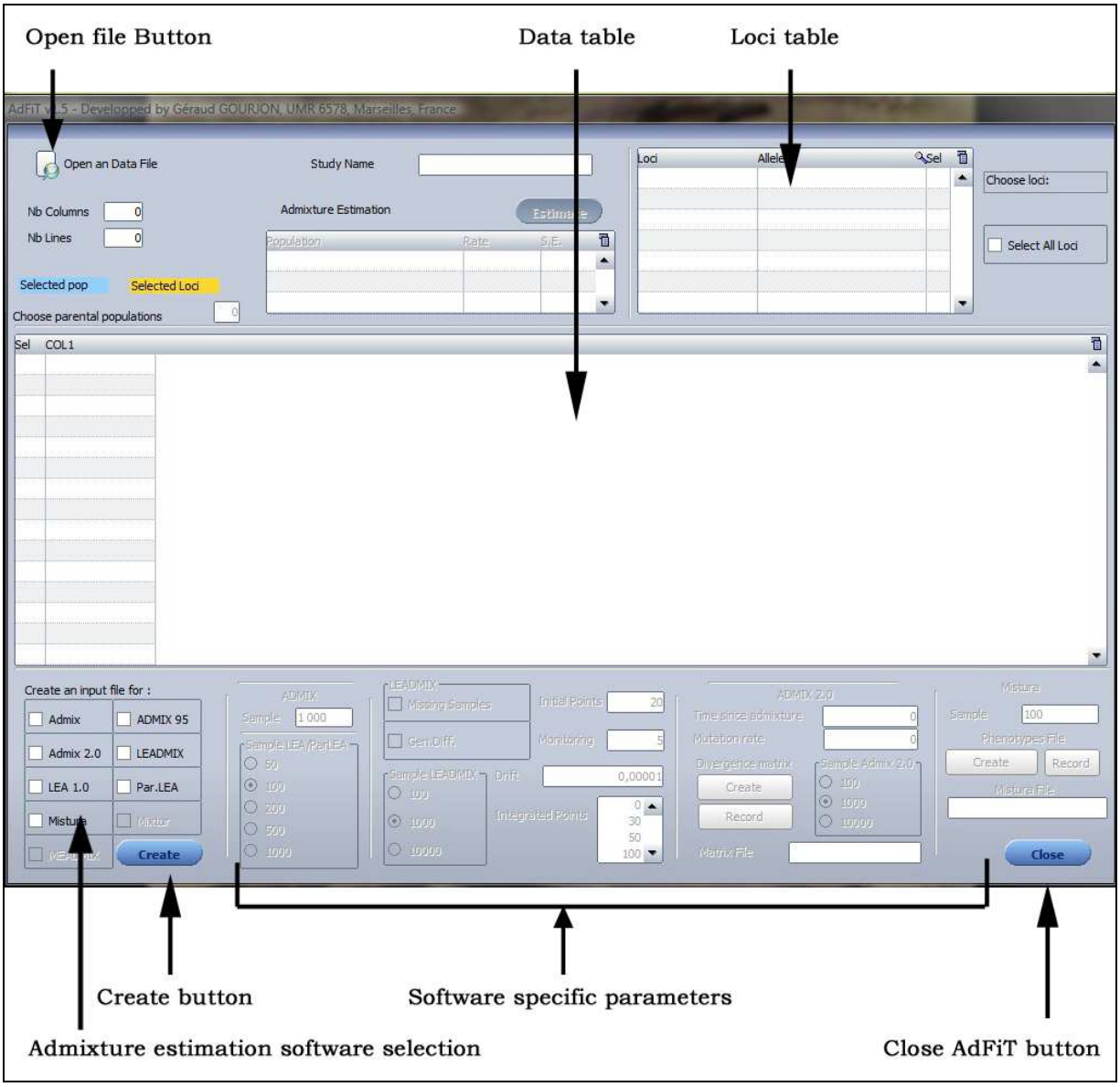


	A	B	C	D	E	F	G
1		ABO			Duffy		
2		A	B	0	FYA	FYB	FYO
3	Admixed Population	0.141	0.069	0.79	0.061	0.009	0.93
4	Parental Population 1	0.153	0.127	0.72	0.036	0.016	0.948
5	Parental Population 2	0.144	0.123	0.733	0.162	0.15	0.688
6	Parental Population 3	0.221	0.184	0.595	0.483	0.393	0.124
7	Parental Population 4	0.158	0.2	0.642	0.842	0.156	0.002

2. PROCEDURE FOR RUNNING ADFIT

Once this data file created, open it with the “Open a Data File” button. The Data Table will be filled automatically with these Data.
The first line in this table contains data for the admixed population which is all selected. The input files creation is in 3 steps:

1) Parental populations selection: in the Data Table, choose parental populations: tick in the “Sel” column in front of parental populations: selected parental populations are highlighted in blue
2) Loci selection: in the Loci table, tick in the “Sel” column to select Loci. Once can select all loci with the “Select all loci”option. Selected loci are highlighted in blue in Loci table and in orange in Data Table.
3) Select the admixture estimation software and eventually filled the specific parameters (refer to the documentation for each software below), then click on “Create” button. The input file is created on your desktop.



Input Files are named automatically. If you want to create several input files of the same type but with different markers (and for some software with different parental populations, see specific name for each software), remember to rename, to erase or to put it in another folder.

One priority was to design it as intuitive as possible. Most of the errors in the input files creation procedure have been envisaged and taken into account. Errors/Information messages have been added in AdFiT to explain encountered problems in every conceivable case. Moreover, these risks could be prevented in some cases. For example, LEA (Langella *et al.* 2001) software considers only 2 parental populations and it will be impossible to generate an input file with 3 or more parental populations (an error message will appear). Inversely, when, for a given allele, frequencies for all population are equal to 0, most of admixture estimation software won't work. It is nevertheless possible to create these input files with AdFiT but an explanatory message will warm users that the input files would cause disturbance and highlight the null alleles in the data sheet.

2. SPECIFIC PARAMETERS

ADMIX95 input file

No specific parameters for this software. Input files are generated immediately. It is named with the 2 first letters of selected parental populations (extension is .inp); max 8 characters.

ADMIX input file

The sample size must be filled (see ADMIX documentation). It impacts only the F_{st} value. It is named with the 2 first letters of selected parental populations (extension is .dat); max 8 characters.

LEA input file

The sample size must be filled (see LEA documentation). This could have influence in the estimation. To increase the sample size, increases results accuracy. However, computing time will be highly increased too.
The file is named "infile" (no extension). All files will be named similarly (LEA needs this format) so shift it in another folder before to create a new one.

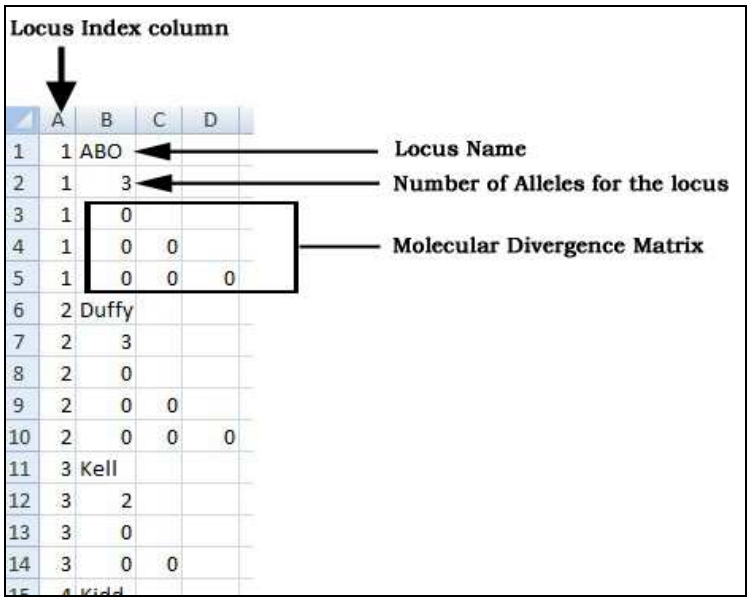
Parallel LEA input file

The sample size must be filled (see Parallel LEA documentation).
The file is named "infile" (no extension). All files will be named similarly so shift it in another folder before to create a new one.

Admix 2.0 input file

For the creation of the input file, a primary matrix ".xls" file, including molecular divergence between alleles for each locus, has to be generated. AdFiT creates an ".xls" file already filled with the correct number of loci and alleles For this purpose, once the Data file selected, click

on “select all button” (near the Loci table), select Admix 2.0 software and click on “Create” button in the Admix 2.0 specific parameters section under “Molecular divergence matrix”. This will generate on the desktop a “.xls” file including pre-formatted matrixes. Open this file with Calc (Open Office) or Excel (Microsoft Office) and enter the molecular divergence between alleles for each locus: substitute the “0” in matrixes by the molecular divergences values for all loci. Then, save it (“**save as**” in File Menu and not only “save”) in xls (Excel 97-2003 format on Microsoft Office 2007, and Excel 07/2000/XP on Open Office). Caution: do not save it in .txt or in .xlsx format nor .ods.
This file is required to create the input file. It has to be filled correctly if the molecular differentiation is taken into account. Otherwise, AdFiT fill it by values equal to 0 in the matrix (no differentiation between alleles). Refer to the Admix 2.0 documentation to fill it correctly. However, if you use Microsoft Office 2007 or Open Office Calc, it is **compulsory** to do this because of compatibility problems, even if you don’t use the molecular divergence (see the “**IMPORTANT NOTE**” below)



	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	

If this matrixes file has already been created, you can select it with the “Record” button in the Admix 2.0 parameters. Caution: select a Matrixes file filled with the correct loci in the same order as in Loci table. The name of selected matrix file appears under these buttons.
The selected matrixes file name appears in the box below these buttons.
Others parameters: (here again, refer to the Admix 2.0 documentation): time since admixture event (enter 0 if this time is unknown); mutation rate (must different from 0 if time since admixture is filled); and sample size.
When all parameters and matrix file are filled, select parental populations and loci and click on input files “create” button (under software selection).
It is named with the 2 first letters of selected parental populations (extension is .txt); max 8 characters.

IMPORTANT NOTE: some compatibility troubles rise when using Microsoft Office Excel 2007 and Open Office Calc. **To avoid any problems, on Microsoft Office 2007 remember to open the Matrixes file and to “save as” in Excel 97-2003 format (not “save”) before to create the input files**. Otherwise, Office 2007 considers it as a txt file instead of a xls file and AdFiT will not be able to generate the input files. **On Open Office, when you open it, a**

window with different conversion parameters will appear, click on "ok" then fill the file and save as "xls" Excel 07/2000/XP.

NOTE: When opening it the first time after the creation with AdFiT, an error message will occur in Microsoft Office, click "yes" ☺

LEADMIX input file

Refer to the LEADMIX documentation to fill these parameters correctly.

Specific parameters:

Missing samples: tick if data are missing for one parental population.

Genetic differentiation: tick if genetic differentiation has to be accounted for (mutation).

Sample size: select the sample size (will not modify results for point estimators)

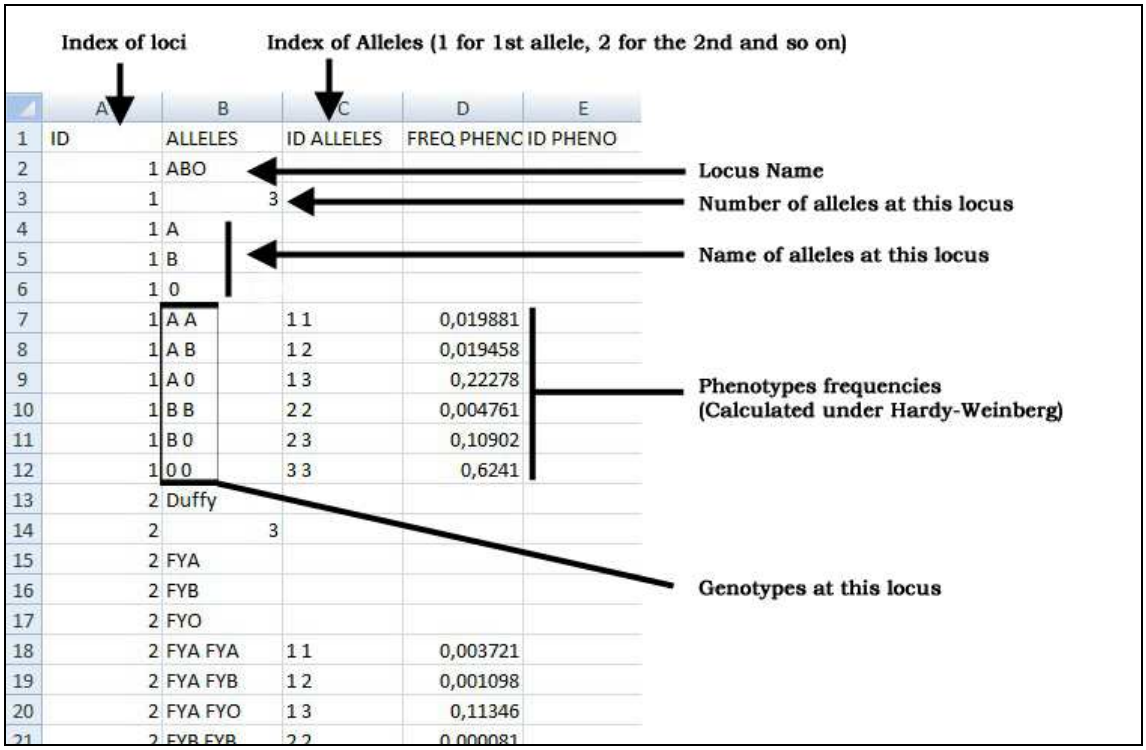
Drift: enter the drift rate.

Initial points and integrated points: Refer to the LEADMIX documentation to fill these parameters.

Monitoring: enter a digit (between one to five): this will not modify results but more details will be displayed during LEADMIX run.

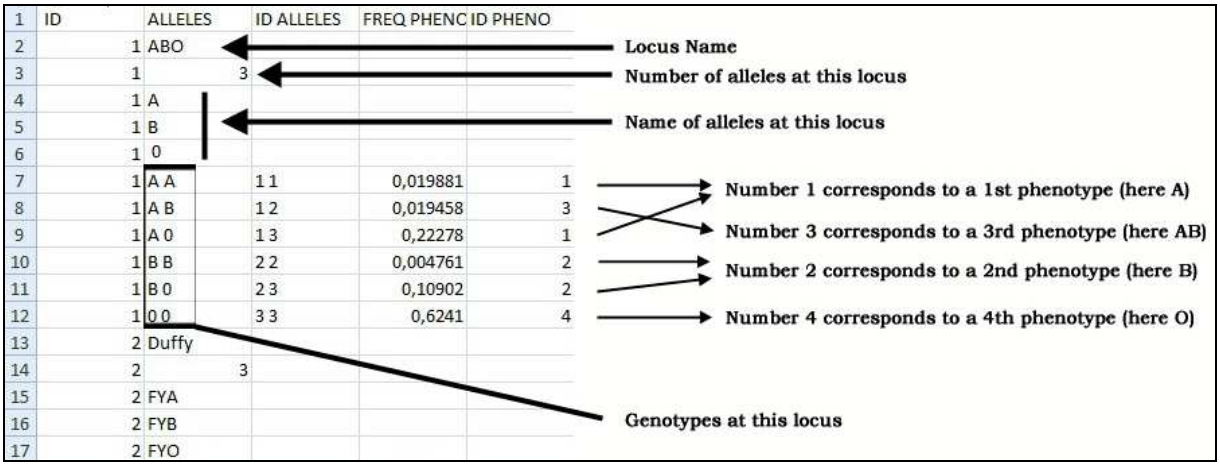
Mistura input file

For the creation of the input file, a primary "xls" file, establishing the correspondence between genotypes and phenotypes each locus, has to be generated. User need to specify wich genotypes correspond to which phenotype. AdFiT creates an "xls" file partially filled. For this purpose, once the Data file selected, click on "select all button" (near the Loci table), select Mistura software and click on "Create" button in the Mistura specific parameters section under "Phenotypes File". This will generate on the desktop an "xls" file including pre-formatted matrixes for phenotypes/genotypes (see below).



	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	

Open this file with Open Office Calc or Microsoft Office Excel. An error message will appear if Office 2007 is used, click "Yes". Similarly, a conversion window will appear on Open Office, click OK.
ID PHENO column must be entered as follow: Enter number 1 in front of each genetopype corresponding to the phenotype 1; enter number 2 in front of each genetopype corresponding to the phenotype 2; and so on. Numbers used for phenotypes are arbitrary but all phenotypes must be named with an integer with no gap: for example, for 4 different phenotypes, they will be noted as 1, 2, 3, 4 and not 1, 3, 4, 6. See the example with ABO blood group below.
Then, save it ("**save as**" in File Menu and not only "save") in Excel 97-2003 format (.xls). Caution: do not save it in .txt or in .xlsx format. Remember that this step is **compulsory.**
not 1, 3, 4, 6. Then, save it with "save as" (in File Menu) in xls (97-2003) format.



If this file has already been created, you can select it with the Record button. Caution: select a Phenotype file filled with the correct loci in the same order as in Loci table. The name of selected Phenotype file appears under these buttons.
When the sample size and phenotype file are filled, click on input files "create" button (under software selection).
It is named with "Mis" and the first letter of selected parental populations (extension is .txt). Mistura input file needs a study name, and it will be the name of the admixed population by default. You can change this name in the field "Study name".

IMPORTANT NOTE: **Remember to open the Phenotypes file and to "save as" in Excel 97-2003 format before to create the input files**. Otherwise, Office 2007 considers it as a txt file instead of a xls file and AdFiT will not be able to generate the input files. When opening it the first time after the creation with AdFiT, an error message will occur in Microsoft Office, click "yes" ☺. Similarly, **on Open Office, when you open it, a window with conversion parameters will appear, click on "ok" then fill the file and save as "xls" Excel 07/2000/XP.**

3. Other Parameters

Study Name

This field is automatically filled with the name of the hybrid population. You fill it with the required name, if you desire to change it.

Admixture Estimate

This function will estimate the admixture rates from the genetic distances, using Cavalli-Sforza formula (Cavalli-Sforza *et al.* 1994). Actually this function is disabled but will be functional in the v2.0.

4. Data Tables

All tables generated by AdFiT (Loci Table, Data Table, and in the v2.0, admixture rates Table) could be exported directly in an Excel or Word format, or be printed.

Requirements

Project name	AdFiT (Admixture Files Tool)
Version	1.5.0
Language	English, French, Spanish
Project home page	http://www.anthropologie-biologique.cnrs.fr/recherche/axe1equipe3.php
Operating system	Platform dependent (Windows 9X/XP/Vista)
License	Freely available for academic users (By quoting the reference)

References

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.