Bulletins et Mémoires de la Société d'Anthropologie de Paris, n.s., t. 21, 2009, 3-4, p. 223-229

ADFIT V1.7 (ADMIXTURE FILES TOOL): INPUT FILES CREATING TOOL FOR POPULATION GENETIC ADMIXTURE ESTIMATION SOFTWARE

ADFIT V1.7 (ADMIXTURE FILES TOOL) : OUTIL DE CRÉATION DES FICHIERS D'ENTRÉE POUR LES LOGICIELS D'ESTIMATION DU MÉLANGE GÉNÉTIQUE ENTRE POPULATIONS

Géraud Gourjon[1], Anna Degioanni[1]

Abstract

The six most commonly used software programs in the literature for population genetic admixture estimation since 20 years, are ADMIX, ADMIX95, Mistura, Admix 2.0, LEA, and LEADMIX, and each one has its own specific file format and filename extensions. We present a specific tool that can help in the file creation process: *AdFiT* (Admixture Files Tool) *version 1.7*, a multi-language software (English, French and Spanish). It allows, from a common file containing data, to instantaneously create input files for these six software programs and two others, Parallel LEA (ParLEA) and Mixtur. The use of several software programs and all the available genetic markers appears to be mandatory to estimate efficiently the admixture rates. This requires the creation of multiple input files, for each software file format and leads to a wide data handling, increasing the risk of keyboarding errors. *AdFit* can do it quickly, simply (easily) and without errors. *AdFiT* is freely available online for academic users.

Keywords: genetic admixture, human population, admixture rates, software tool.

Résumé

Les logiciels d'estimation du mélange génétique entre populations les plus utilisés dans la littérature scientifique au cours de ces vingt dernières années sont : ADMIX, ADMIX95, Mistura, Admix 2.0, LEA et LEADMIX. Chacun de ces logiciels nécessite un fichier d'entrée avec un format et une extension spécifiques. Nous présentons ici un outil d'aide dans le processus de création des fichiers d'entrée : AdFiT *(Admixture Files Tool), version 1.7, un logiciel multilingue (anglais, français et espagnol). À partir d'un fichier commun de données, il permet de créer instantanément les fichiers d'entrée pour ces 6 logiciels, ainsi que pour deux autres, Parallel LEA (ParLEA) et Mixtur. L'utilisation de plusieurs logiciels et de tous les marqueurs génétiques disponibles est nécessaire pour estimer efficacement les taux de mélange. Cela exige la création de fichiers d'entrée multiples, avec le format propre à chaque logiciel et conduit à une grande manipulation de données, augmentant les risques d'erreurs de saisie.* AdFit *permet de faire tout cela rapidement, facilement et sans erreur.* AdFiT *est disponible gratuitement en ligne pour une utilisation non commerciale.*

Mots-clés : *mélange génétique, population humaine, taux de mélange, logiciels d'aide.*

1. UMR 6578, Unité d'Anthropologie Bioculturelle, Université Aix-Marseille 2, 13916 Marseille CEDEX, France, e-mail: geraud.gourjon@etumel.univmed.fr

Manuscrit reçu le 06 octobre 2009, accepté le 21 décembre 2009

INTRODUCTION

Since the last five decades, many estimators have been proposed for determining the contribution of parental populations to hybrid population gene pool (Glass, Li 1953; Krieger *et al*. 1965; Roberts, Hiorns 1965; Elston 1971; Thompson 1973; Chakraborty 1975; Szathmary, Reed 1978; Wijsman 1984; Chakraborty 1985; Chakraborty 1986; Wijsman, Neves 1986; Long 1991; Chakraborty, Srinivasan 1992; Mitchell *et al*. 1993; Cavalli-Sforza *et al*. 1994; Bertorelle, Excoffier 1998; Estoup *et al*. 1999; Helgason *et al*. 2000; Di Benedetto *et al*. 2001; Dupanloup, Bertorelle 2001; Beaumont *et al*. 2002; Wang 2003; Excoffier *et al*. 2005; Maca-Meyer *et al*. 2005; Wang 2006; Giovannini *et al*. 2009). Some of these estimators have been implemented in software programs for genetic admixture estimation. The six most commonly used software programs for genetic admixture estimation in the literature since 20 years, are ADMIX, ADMIX95, Admix 2.0, Mistura, LEA, and LEADMIX, and each one has its own specific file format and filename extensions.

In the goal to test them extensively with a set of genetic markers including blood groups, mtDNA, Y-Chromosome SNP, and Kir genes, and among multiple parental populations (having historical links or not) (G. Gourjon *et al*. unpublished study), we discovered: firstly, a varying level of performance, as well as frequent false positive results, being source of misinterpretations of admixture rates; and secondly, an increased risk of keyboarding errors in handling data creation in the different input files. Software programs implement various methods with their own flaws and advantages, and their different subjacent admixture models. Since these methods do not take same parameters into account (*i.e*. drift, mutation, sampling error, etc.), it is also clear that the simultaneous use of several software programs is necessary to estimate efficiently admixture rates and to evaluate influence of these parameters. Moreover, it seems crucial to use allele frequencies from all available genetic markers (both classical and molecular), especially when a sexual bias is suspected in the admixture dynamics.

In order to solve this problem, we have created a specific tool that can help in this file creation process: *AdFiT* (Admixture Files Tool) *version 1.7* is a multi-language software (English, French and Spanish) which allows, starting from a common file containing data for all populations (parental and admixed ones), to instantaneously create input files for Mixtur (Wijsman 1984), ADMIX (Long 1991), ADMIX95 (Chakraborty 1985), Mistura (Krieger *et al*. 1965; Cabello, Krieger 1997), Admix2.0 (Bertorelle, Excoffier 1998; Dupanloup, Bertorelle 2001), LEA (Chikhi *et al*. 2001; Langella *et al*. 2001), LEADMIX (Wang 2003), and Parallel LEA (Giovannini *et al*. 2009). *AdFiT* is freely available (http://www.anthropologie-biologique.cnrs.fr/recherche/axe1equipe3.php) for academic users only.

Version 2.0 is expected to be released soon, it will allow the creation of input files for other software programs such as MEADMIX (Wang 2006), and will provide admixture rate preview, estimated from genetic distances (Cavalli-Sforza *et al*. 1994). This new version will allow getting a better view of genetic admixture in the studied hybrid population.

PRINCIPLE AND PROCEDURES

Input files for the eight software programs (table I) can be easily generated with *AdFiT*. To be the most compatible and user-friendly as possible, user has only to create a generic spreadsheet file (*fig. 1*) containing all required data for all populations (allele frequencies, names of alleles and loci). The file must have ".xls" format (created from Open Office Calc or Microsoft© Office Excel). We chose this format because it offers the opportunity to concatenate different kinds of raw data in a simplified format (the most commonly used and friendly spreadsheet) and allows "juggling" between original and published data (spreadsheet, text, database), generating all wanted combinations of loci, parental populations and admixture estimation software programs. This requires only once "cut paste" of used data.

The version 2.0 (in preparation) will have an increased accessibility by allowing an Open Office ".ods" format and by running on Linux Operating Systems (Open Office and Linux OS are under GNU Lesser General Public Licence).

AdFiT directly reads the data file: user just needs to select the admixture estimation software and to complete, if needed, specific program parameters (such as sample size for ADMIX).

For two of them, intermediate files are required and are partially generated by *AdFit*. No software allows staying clear from this step.

—For Admix 2.0 (Bertorelle, Excoffier 1998; Dupanloup, Bertorelle 2001), a matrix including molecular divergence between alleles for each locus

has to be generated. *AdFiT* creates an ".xls" file that automatically includes the correct number of loci and alleles: users can fill it easily and correctly. This file is required if the molecular differentiation has to be taken into account.

—For Mistura (Cabello, Krieger 1997), a genotypes/phenotypes file allows *AdFiT* to fill properly the input files.

Complete description of the functions and guidelines for use of *AdFit* are available online (pdf documentations are in the three languages, English, French, and Spanish).

Admixed population name — Locus name — Alleles name

	A	B	C	D	E	F	G	H	I
1		Locus1			Locus2			Locus3	
2		Allele1	Allele2	Allele3	Allele1	Allele2	Allele3	Allele1	Allele2
3	Admixed Population	0.141	0.069	0.79	0.061	0.009	0.93	0.012	0.988
4	Parental Population 1	0.153	0.127	0.72	0.036	0.016	0.948	0.011	0.989
5	Parental Population 2	0.144	0.123	0.733	0.162	0.15	0.688	0.072	0.928
6	Parental Population 3	0.221	0.184	0.595	0.483	0.393	0.124	0.029	0.971
7	Parental Population 4	0.158	0.2	0.642	0.842	0.156	0.002	0.001	0.999
8	Parental Population 5	0.168	0.138	0.694	0.255	0.15	0.595	0.007	0.993

Parental population name — Alleles frequencies

Fig. 1—Generic AdFiT *input file.*

*Fig. 1 - Fichier d'entrée commun d'*AdFiT.

ADVANTAGES

We have evaluated users' perception of *AdFiT* about its effectiveness (how the software fits for purpose) and its efficiency (time required for its use). Usability appears to be very good and *AdFiT* answers correctly to the purpose. Time for creating input files for the different software programs, including a very important number of combinations of parental populations/markers/alleles is very short. No errors have been found in these generated files. Gain of time is really substantial, representing the greatest advantage of this software. This allows researchers to multiply the simulation numbers and consequently to increase the accuracy of the admixture rates determination.

Our second priority was to make the use of the software as intuitive as possible, with a design allowing its possible use nearly without reading the instructions. The main errors occurring during the input file procedure, when done manually, have been identified and taken into account. Information or error messages have been added to explain encountered problems in every conceivable situation, with explanation if risks could be prevented or not. For example, LEA software (Langella *et al.* 2001) considers only 2 parental populations: it is therefore impossible to generate an input file with *AdFiT* having more than 2 parental populations (explanatory message appears). Moreover, when, for a given allele, frequencies for all population are equal to 0, most of admixture estimation software programs simply do not run. It is nevertheless possible to create these input files with *AdFiT* but an informative message warns user that the input files would cause disturbance and the null alleles are highlighted in the data sheet. In addition, *AdFiT* checks for each locus and each population if the sum of allele frequencies is equal to 1 and gives the list of involved alleles/loci/populations.

Software	Author	Year	Implemented methods	Method references	Download websites
Mixtur	Wijsman	1984	Least-Square	Wijsman 1984	Available on request to its conceptor, E. Wijsman *wijsman@u.washington.edu*
ADMIX	Long	1991	Weighted Least Squares	Long 1991	Available on request to its conceptor, J. Long *longjc@umich.edu*
ADMIX 95	Bertoni	1995	Gene Identity	Chakraborty 1985	*http://www.genetica.fmed.edu.uy/software.htm*
Mistura	Cabello, Krieger	1997	Maximum Likelihood	Krieger *et al.* 1965 Cabello, Krieger 1997	Available on request to its conceptor, H. Krieger *hkrieger@icb.usp.br*
Admix 2.0	Bertorelle, Dupanloup	1998 2001	Coalescence-based	Bertorelle, Excoffier 1998 Dupanloup, Bertorelle 2001	*http://web.unife.it/progetti/genetica/Isabelle/admix2_0.html*
LEA	Beaumont, Langella	2001	Coalescent-based maximum Likelihood	Chikhi *et al.* 2001 Langella *et al.* 2001	*http://dm.unife.it/parlea*
LEADMIX	Wang	2003	Maximum likelihood	Wang 2003	*http://www.zsl.org/science/research/software/leadmix,1153.AR.html*
			Least-Square	Roberts, Hiorns 1965	
			Weighted Least Squares	Long 1991 Chakraborty, Srinivasan 1992	
			Coalescent-based	Bertorelle, Excoffier 1998	
Parallel LEA	Giovannini *et al.*	2008	Likelihood-based approach	Chikhi *et al.* 2001 Giovannini *et al.* 2009	*http://dm.unife.it/parlea*

Table I—Admixture estimation software programs taken into account with AdFiT v1.7.

Tabl. I - Logiciel d'estimation du mélange pris en compte par AdFiT v1.7.

CONCLUSION

In this paper, a new program (*AdFiT*, *table II*), which allows creating instantaneously input files for the most commonly used admixture estimation software, is presented. It offers the advantages to be fast, accurate, and easy-to-use. We hope that this software will be useful for our community

However, some limitations concerning the restricted number of software taken into account and the Operating System (Windows© only) could not be overlooked. The 2.0 version will partially solve these issues and other software will be added and *AdFiT* will run on Linux.

Project name	AdFiT (Admixture Files Tool)
Version	1.7
Language	English, French, Spanish
Project home page	http://www.anthropologie-biologique.cnrs.fr/recherche/axe1equipe3.php
Operating system	Windows XP/Vista
Programming language	WinDev
License	Freely available for academic users (quote this reference)

Table II—Technical characteristics.

Tabl. II - Caractéristiques techniques.

BIBLIOGRAPHY

Beaumont (M.), Zhang (W.), Balding (D.) 2002, Approximate Bayesian computation in population genetics, *Genetics* 162, 4: 2025-2035.

Bertorelle (G.), Excoffier (L.) 1998, Inferring admixture proportions from molecular data, *Molecular Biology and Evolution* 15, 10: 1298-1311.

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

Cavalli-Sforza (L.L.), Menozzi (P.), Piazza (A.) 1994, *The History and Geography of Human Genes*, Princeton University Press, Princeton, 1088 p.

Chakraborty (R.) 1975, Estimation of race admixture—New method, *American Journal of Physical Anthropology* 42, 3: 507-511.

Chakraborty (R.) 1985, Gene identity in racial hybrids and estimation of admixture rates, *in* Y. Ahuja, J.V. Neel (eds), *Genetics Microdifferentiation in Human and Other Animal Populations*, Indian Anthropol Association, Delhi University Antropology Department, Delhi, India, p. 171-180.

Chakraborty (R.) 1986, Gene admixture in human-populations—Models and predictions, *Yearbook of Physical Anthropology* 29: 1-43.

Chakraborty (R.), Srinivasan (M.R.) 1992, A modified best-maximum likelihood estimator of line regression with errors in both variables—An application for estimating genetic admixture, *Biometrical Journal* 34, 5: 567-576.

Chikhi (L.), Bruford (M.W.), Beaumont (M.A.) 2001, Estimation of admixture proportions: A likelihood-based approach using Markov chain Monte Carlo, *Genetics* 158, 3: 1347-1362.

Di Benedetto (G.), Ergüven (A.), Stenico (M.), Castrì (L.), Bertorelle (G.), Togan (I.), Barbujani (G.) 2001, DNA diversity and population admixture in Anatolia, *American Journal of Physical Anthropology* 115, 2: 144-156.

Dupanloup (I.), Bertorelle (G.) 2001, Inferring admixture proportions from molecular data: Extension to any number of parental populations, *Molecular Biology and Evolution* 18, 4: 672-675.

Elston (R.C.) 1971, The estimation of admixture in racial hybrids, *Annals of Human Genetics* 35, 1: 9-17.

Estoup (A.), Cornuet (J-M.), Rousset (F.), Guyomard (R.) 1999, Juxtaposed Microsatellite Systems as Diagnostic Markers for Admixture: Theoretical Aspects, *Molecular Biology and Evolution* 16, 7: 898-908.

Excoffier (L.), Estoup (A.), Cornuet (J.) 2005, Bayesian analysis of an admixture model with mutations and arbitrarily linked markers, *Genetics* 169, 3: 1727-1738.

Giovannini (A.), Zanghirati (G.), Beaumont (M.), Chikhi (L.), Barbujani (G.) 2009, A novel parallel approach to the likelihood-based estimation of admixture in population genetics, *Bioinformatics* 25, 11: 1440-1441.

Glass (B.), Li (C.C.) 1953, The dynamics of racial intermixture—An analysis based on the american negro, *American Journal of Human Genetics* 5, 1: 1-20.

Helgason (A.), Sigurðardóttir (S.), Nicholson (J.), Sykes (B.), Hill (E.W.), Bradley (D.G.), Bosnes (V.), Gulcher (J.R.), Ward (R.), Stefansson (K.) 2000, Estimating Scandinavian and Gaelic ancestry in the male settlers of Iceland, *American Journal of Human Genetics* 67, 3: 697-717.

Krieger (H.), Morton (N.), Mi (M.), Azevêdo (E.), Freire-Maia (A.), Yasuda (N.) 1965, Racial admixture in north-eastern Brazil, *Annals of Human Genetics* 29, 2: 113-125.

Langella (O.), Chikhi (L.), Beaumont (M.A.) 2001, LEA (likelihood-based estimation of admixture): a program to estimate simultaneously admixture and time since the admixture event, *Molecular Ecology Notes* 1, 4: 357-358.

Long (J.C.) 1991, The Genetic Structure of Admixed Populations, *Genetics* 127, 2: 417-428.

Maca-Meyer (N.), Cabrera (V.), Arnay (M.), Flores (C.), Fregel (R.), González (A.), Larruga (J.) 2005, Mitochondrial DNA diversity in 17th-18th century remains

from Tenerife (Canary Islands), *American Journal of Physical Anthropology* 127, 4: 418-426.

MITCHELL (B.), WILLIAMS-BLANGERO (S.), CHAKRABORTY (R.), VALDEZ (R.), HAZUDA (H.), HAFFNER (S.), STERN (M.) 1993, A comparison of three methods for assessing Amerindian admixture in Mexican Americans, *Ethnicity and Disease* 3, 1: 22-31.

ROBERTS (D.F.), HIORNS (R.W.) 1965, Methods of analysis of genetic composition of a hybrid population, *Human Biology* 37, 1: 38-43.

SZATHMARY (E.), REED (T.) 1978, Calculation of the maximum amount of gene admixture in a hybrid population, *American Journal of Physical Anthropology* 48, 1: 29-33.

THOMPSON (E.) 1973, The Icelandic admixture problem, *Annals of Human Genetics* 37, 1: 69-80.

WANG (J.L.) 2003, Maximum-likelihood estimation of admixture proportions from genetic data, *Genetics* 164, 2: 747-765.

WANG (J.L.) 2006, A coalescent-based estimator of admixture from DNA sequences, *Genetics* 173, 3: 1679-1692.

WIJSMAN (E.) 1984, Techniques for estimating genetic admixture and applications to the problem of the origin of the Icelanders and the Ashkenazi Jews, *Human Genetics* 67, 4: 441-448.

WIJSMAN (E.), NEVES (W.) 1986, The use of nonmetric variation in estimating human population admixture: a test case with Brazilian blacks, whites, and mulattos, *American Journal of Physical Anthropology* 70, 3: 395-405.