270–271 Nucleic Acids Research, 2003, Vol. 31, No. 1
DOI: 10.1093/nar/gkg043

ALFRED: the ALelle FREquency Database. Update

H. Rajeevan, M. V. Osier, K.-H. Cheung[1], H. Deng, L. Druskin[1], R. Heinzen, J. R. Kidd, S. Stein, A. J. Pakstis, N. P. Tosches[1], C.-C. Yeh, P. L. Miller[1] and K. K. Kidd*

Department of Genetics and [1]Center for Medical Informatics, Yale University School of Medicine, New Haven, CT 06520-8005, USA

Received September 18, 2002; Accepted September 27, 2002

ABSTRACT

Elaboration of ALFRED (http://alfred.med.yale.edu) is being continued in two directions. One of which is developing tools for efficiently annotating the entries and checking the integrity of the data already in the database while the other is to increase the quantity and accessibility of data. Information contained in ALFRED such as, polymorphic sites, number of populations and frequency tables (one sample typed for one site) has significantly increased.

DATABASE DESCRIPTION

The ALelle FREquency Database (ALFRED) is designed to store and disseminate frequencies of alleles at human autosomal polymorphic sites for multiple defined population samples, primarily for the population genetics and molecular anthropology communities. The focus is on allele frequencies of normal, common DNA variants, i.e., polymorphisms, in samples of anthropologically defined populations. Links are provided to molecular databases for precise definitions and locations of the polymorphisms and to anthropologic databases for linguistic, ethnographic and demographic information on the populations sampled. References to publications are associated with the frequencies and linked to PubMed, whenever possible. Many polymorphisms have linkes to low-tech protocols suitable for small laboratories engaged in anthropologic research. ALFRED has information on 672 polymorphic sites typed on at least one population sample and 288 populations typed for at least one polymorphism. ALFRED is accessible from http://alfred.med.yale.edu.

RECENT DEVELOPMENTS

Effort on the past year has been concentrated in three areas: (1) development of curatorial tools, (2) implementation of a more robust and sustainable Oracle database, and (3) increasing the quantity and quality of data.

Curatorial tools

Unseen by the user, these software tools provide integrity checks and allow the curators to more efficiently annotate entries and add web links to appropriate entries in other database. When entering data into ALFRED, our curators are using a controlled vocabulary including official locus names and symbols as a way to achieve data quality. We have implemented automatic checks to enforce this. Integrity checks of the data already in ALFRED are run periodically to ensure data accuracy and consistency.

Conversion to Oracle

We have converted the entire system from its current Access database implementation to Oracle to allow for the considerable expansion of the data in the coming year. The Oracle version is currently being tested and the queries optimized. It should be the active system by early 2003.

Data expansion and development

With new unpublished allele frequency data from Kidd Lab and frequency data from published articles, the number of frequency tables (one sample typed for one site) increased from 3561 (September, 2001) to 6301 (September, 2002). The staffs are systematically extracting gene frequency data on DNA sequence variants from recent issues of major human genetics and physical anthropology journals. To ensure the high quality of our data, we perform data curation in an iterative and systematic way before importing data into ALFRED. Descriptions of a large number of loci, polymorphisms, alleles, populations and samples from the literature are nearing completion and will be loaded with the associated allele frequency tables. Additional web links to literature and public databases such as GenBank, PubMed, GDB, OMIM, LocusLink, and Ethnologue have been added for the existing entries.

Data accessibility

A Document Type Definition (DTD) has been developed for importing and exporting ALFRED data in XML format. All information in ALFRED can now be put into a single compressed ‘data dump’ file in the declared XML format. The data dump can include either all relevant information

*To whom correspondence should be addressed. Tel: +1 203 785 2654; Email: kidd@biomed.med.yale.edu

(including descriptions) or only the data relevant to statistical analyses. These files are available on request by email to http://alfred.med.yale.edu/alfred/feedback.asp.

Web viewing enhancement

New graphical overviews of the database contents have been implemented to direct users to the more extensive 'comparative' aspects of the database. A 'sites per population' web page (http://alfred.med.yale.edu/alfred/sitesperpop_graph.asp) shows graphically (and numerically) the number of allele frequency tables for each population. Currently, the maximum is 385 for the Han. A 'populations per site' web page (http://alfred.med.yale.edu/alfred/ popspersite_graph.asp) similarly represents the number of allele frequency tables for each polymorphic site. Currently, the maximum is 49 for the CD4 pentanucleotide repeat polymorphism.

ACKNOWLEDGEMENTS

Ongoing funding of ALFRED is provided by NSF grant BCS0096588. Initial funding for ALFRED was provided by NSF grant SBR-9632509 and USPHS grants P01GM57672, R01AA09379 and T15LM07056.

REFERENCE

1. Osier,M.V., Cheung,K.H., Kid,J.R., Pakstis,A.J., Miller,P.L. and Kidd,K.K. (2002) ALFRED: An Allele Frequency Database for Anthropology. *Am. J. Phys. Anthropol.*, **119**, 77–83.