Nucleic Acids Research, 2001, Vol. 29, No. 1 317–319

ALFRED: an allele frequency database for diverse populations and DNA polymorphisms—an update

Michael V. Osier, Kei-Hoi Cheung[1], Judith R. Kidd, Andrew J. Pakstis, Perry L. Miller[1] and Kenneth K. Kidd*

Department of Genetics, Yale University School of Medicine, 333 Cedar Street, PO Box 208005, New Haven, CT 06520-8005, USA and [1]Center for Medical Informatics, Yale University School of Medicine, 333 Cedar Street, PO Box 208009, New Haven, CT 06520-8009, USA

Received September 21, 2000; Accepted September 27, 2000

ABSTRACT

ALFRED (the ALelle FREquency Database) is designed to store and disseminate frequencies of alleles at human polymorphic sites for multiple populations, primarily for the population genetics and molecular anthropology communities. Currently ALFRED has information on over 180 polymorphic sites for more than 70 populations. Since our initial release of the database we have focussed on increasing the quantity and quality of data, making reciprocal links between ALFRED and other related databases, and providing useful tools to make the data more comprehensible to the end user. ALFRED is accessible from the Kidd Lab home page (http://info.med.yale.edu/genetics/kkidd/) or from ALFRED directly (http://alfred.med.yale.edu/alfred/index.asp).

INTRODUCTION

Major efforts are now underway to identify large numbers of single nucleotide polymorphisms (SNPs) (for examples see 1,2); >100 000 have already been identified and are being cataloged in publicly accessible databases including HGBASE (3) and dbSNP (4). Mere knowledge of the existence of specific variation in DNA sequence can be useful, but most applications require the frequencies of the alleles at the varying site both in planning research projects and in statistical analyses. The databases cataloging sequence variations are not focussed on allele frequencies and do not completely meet the research needs of the human population genetics and molecular anthropology communities.

We have prototyped a web-accessible database, 'ALFRED' (ALlele FREquency Database), with a focus on flexible storage and retrieval of gene frequency data. The prototype is accessible via the Kidd Lab home page or directly from the ALFRED Web site. We developed ALFRED initially to make our own data available to others since journal limitations on length prevented publication of all of the actual frequency tables (for examples see 5–9) and the published and supporting data need to be available to the scientific community, ideally through the Web. We also needed methods for making available updates to those studies: small increments of data that do not justify publication.

Since our initial description of the database (10,11), additional data have been added, including some data from other researchers, but the emphasis has been on completeness of polymorphism, population and sample descriptions for existing frequency entries and on structural and functional enhancements. As of September 27, 2000, ALFRED contains 2865 allele frequency tables, each giving frequencies of all alleles at one site in one population sample. Structurally, the database and interface have become more sophisticated. Detailed population and sample descriptions have been added for nearly all populations. Links to the online version of Ethnologue have been added for many populations to specify the language spoken. Links have been made to dbSNP and HGBASE to add molecular specificity to site descriptions. A standard link format has been implemented so that others can make links to the data in ALFRED. Finally, a tool to graph site frequencies dynamically in a somewhat anthropological context has been added. We believe that this database can help to connect the many diverse fields interested in human populations and their histories with the highly relevant molecular data being developed. During the coming year the amount of data in ALFRED (number of polymorphisms and populations) will be significantly increased.

NEW DATA

Much information has been added to fill in the previously described structure for polymorphisms and populations with frequencies already in the database. For many population and sample records, detailed descriptive information has been added. In the POPULATIONS table, this information includes the primary spoken language and language family with links to the Ethnologue Web site as described below, longitude and latitude ranges defining a geographic rectangle bounding the rough location of the population, and general information about the population. Additionally, all populations are given a decimal-based identifier for the purpose of sorting in the dynamic graphing tool described below. For the specific samples analyzed, the additional information includes the geographical location of individuals sampled and the names of those researchers involved in the collection and preservation of

*To whom correspondence should be addressed: Tel: +1 203 785 2654; Fax: +1 203 785 6568; Email: kidd@biomed.med.yale.edu

the sample. Details have also been placed in the SITES table. In some cases this added detail is simply a more accurate description of the polymorphic site, sometimes including molecular data such as the nature of the variation at the nucleotide level or changes to the protein and their effect. Where known, the ancestral allele states and how they were determined are included.

ENHANCEMENTS TO LINKING

Links to other databases

Precise specification of population, sample, polymorphism and typing procedure are all important if future researchers are to compare their new frequency data with frequencies in ALFRED and meaningfully interpret similarities and differences. The first two are being addressed by textual descriptions, citations of publications using the specific samples and links to other databases, most notably Ethnologue. Specifications of the polymorphism are provided by inclusion of the primers and typing protocols used. But links to other databases can provide additional information outside the scope of ALFRED.

At the moment, the most obviously relevant molecular databases for ALFRED to link to are the large and rapidly growing ones such as dbSNP (4) and HGBASE (3), which define polymorphisms at the molecular level. Most of the work will fall to the ALFRED curators to establish the logical connections between polymorphism data and molecular definitions. In some cases, such as the DRD2 CA dinucleotide repeat (ALFRED UID SI000141G), we have already established pointers in ALFRED to the correct records in other databases. Other relevant molecular databases include the sequence database, GenBank, and some of the other interconnected NCBI databases (PubMed, OMIM, LocusLink, Map Viewer), GDB (Genome Data Base), GeneCard and the mtDNA databases MitoMap and MOUSE. The structures in ALFRED now allow for links to those as well, but they have not been implemented.

Non-molecular databases are also important. As for links to molecular databases, the database structures are in place and only the curatorial effort is needed to identify the links and put them into the database. We have identified several different databases that could/should have links established. ALFRED uses the official gene names and symbols but official locus names are not used consistently in the literature. Links to GDB and the Human Gene Nomenclature Committee Web sites for the specific loci will help remove any ambiguity about which loci have data in ALFRED. Other locus information that may be relevant to users can be found in the NCBI databases. The obvious link into those databases is through LocusLink, allowing the user to search Online Mendelian Inheritance in Man (OMIM) for clinically related information, to GenBank for detailed DNA sequence information, to PubMed for literature citations and to dbSNP for information about all documented variation (polymorphisms) at the locus. In the literature table in ALFRED we already have the URLs for the PubMed entries for those papers; this allows the reader to view the abstract (in most cases) and occasionally view or download the whole paper.

From the social sciences perspective we have identified one Web site that is especially relevant: Ethnologue (12; online version http://www.sil.org/ethnologue/). The Ethnologue Web site has URLs that allow one to go directly to a country entry which includes the list of all languages in that country or directly to the text for a specific language in that list. Alternatively, if the language code is used, a URL will retrieve all links to that language. For language family, a URL can go to the Language Family Index in Ethnologue. It seems likely that tools could be written to help future curators find the correct URLs in Ethnologue corresponding to entries in ALFRED and tie them into the ALFRED entries. At the moment we are adding these links manually as personnel time allows.

Links into ALFRED

Clearly, stable URLs and UIDs for external links into ALFRED are also needed. Since our original descriptions of ALFRED (10,11) we have added a UID (unique identifier) field to each major table. Instead of ALFRED's interface searching for an item by name/symbol we have used this UID field to identify the item uniquely. Using these UIDs also allows external databases to make Web links to data in ALFRED. We have just developed a standard URL format for such links that should remain stable but administrators of other databases have not yet tested links into ALFRED. A detailed description of the structure of the stable URLs is listed in the Supplementary Material.

DYNAMIC GRAPHING TOOL

While the main focus of ALFRED has been to make the frequency data accessible, we have also added a tool to help the end user visualize the data. From the detailed information page for any polymorphism, one can click on 'Graph allele frequencies at this site' to view a dynamically generated graph of allele frequencies at that site. The existing tool is satisfactory for SNPs and other polymorphisms with few alleles, but is not useful for highly polymorphic multiallelic loci. Currently, the tool selects and graphs the first ten non-zero allele frequencies encountered in the search results from the database. All other allele frequencies are summed and graphed as a 'residual' bar. We are exploring algorithms to determine which ten alleles are 'most informative' in the set of search results to avoid meaningless graphs of rare alleles. This simple tool, however, currently makes much of the data in ALFRED far more comprehensible for the end user.

CONCLUSIONS

One activity in the past year has been to fill in preexisting fields in the database, but the primary focus has been to enhance interconnection among databases. Different databases being developed provide different foci and information which when interconnected will assist the researchers on the client side of the Internet. While the contents of ALFRED are still primarily those from one laboratory, several researchers have indicated a desire/willingness to add their data, both published and unpublished. The database will continue to publish frequencies in well-defined population samples in contrast to most of the samples that have been used in the large SNP discovery projects. A primary focus will continue to be on exploring new means (e.g. extensible markup language or XML) of interconnecting data already available on the Web.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

This work is supported in part by NIH grant 1-P01-GM57672 (K.K.K., J.R.K.), National Library of Medicine grant G08 LM05583 (P.L.M.), NSF grant BCS9912028 (J.R.K.), and NIAAA grant 2-R01-AA09379-06A1.

REFERENCES

1. Cargill,M., Altschuler,D., Ireland,J., Sklar,P. Ardlie,K., Patil,N., Shaw,N., Lane,C.R., Lim,E.P., Kalyanaraman,N. *et al.* (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genet.*, **22**, 231–238.
2. Goddard,K.A., Hopkins,P.J., Hall,J.M. and Witte,J.S. (2000) Linkage disequilibrium and allele-frequency distributions for 114 single-nucleotide polymorphisms in five populations. *Am. J. Hum. Genet.*, **66**, 216–234.
3. Brookes,A.J., Lehvaslaiho,H., Siegfried,M., Boehm,J.G., Yuan,Y.P., Sarkar,C.M., Bork,P. and Ortigao,F. (2000) HGBASE: a database of SNPs and other variations in and around human genes. *Nucleic Acids Res.*, **28**, 356–360.
4. Smigielski,E.M., Sirotkin,K., Ward,M. and Sherry,S.T. (2000) dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Res.*, **28**, 352–355. Updated article in this issue: *Nucleic Acids Res.* (2001), **29**, 308–311.
5. Tishkoff,S.A., Dietzsch,E., Speed,W., Pakstis,A.J., Cheung,K., Kidd,J.R., Bonne-Tamir,B., Santachiara-Benerecetti,A.S., Moral,P., Watson,E., Krings,M., Pääbo,S., Risch,N., Jenkins,T. and Kidd,K.K. (1996) Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science*, **271**, 1380–1387.
6. Tishkoff,S.A., Goldman,A., Calafell,F., Speed,W.C., Deinard,A.S., Bonne-Tamir,B., Kidd,J.R., Pakstis,A.J., Jenkins,T. and Kidd,K.K. (1998) A global haplotype analysis of the myotonic dystrophy locus: implications for the evolution of modern humans and for the origin of myotonic dystrophy mutations. *Am. J. Hum. Genet.*, **62**, 1389–1402.
7. Tishkoff,S.A., Pakstis,A.J., Stoneking,M., Kidd,J.R., Destro-Bisol,G., Sanjantila,A., Lu,R.-B., Deinard,A.S., Sirugo,G., Jenkins,T., Kidd,K.K. and Clark,A.G. (2000) Short Tandem-Repeat Polymorphism/*Alu* Haplotype Variation at the PLAT Locus: Implications for Modern Human Origins. *Am. J. Hum. Genet.*, **67**, 901–925.
8. Calafell,F., Shuster,A., Speed,W.C., Kidd,J.R. and Kidd,K.K. (1998) Short tandem repeat polymorphism evolution in humans. *Eur. J. Hum. Genet.*, **6**, 38–49.
9. Kidd,J.R., Pakstis,A.J., Zhao,H., Lu,R.-B., Okonofua,F.E., Odunsi,A., Grigorenko,E., Bonne-Tamir,B., Friedlaender,J., Schulz,L.O., Parnas,J. and Kidd,K.K. (2000) Haplotypes and Linkage Disequilibrium at the Phenylalanine Hydroxylase Locus, *PAH*, in a Global Representation of Populations. *Am. J. Hum. Genet.*, **66**, 1882–1899.
10. Cheung,K.H., Osier,M.V., Kidd,J.R., Pakstis,A.J., Miller,P.L. and Kidd,K.K. (2000) ALFRED: an allele frequency database for diverse populations and DNA polymorphisms. *Nucleic Acids Res.*, **28**, 361–363.
11. Cheung,K.H., Miller,P.L., Kidd,J.R., Kidd,K.K., Osier,M.V. and Pakstis,A.J. (2000) ALFRED: A Web-Accessible Allele Frequency Database. *Pacific Symposium on Biocomputing 2000 Proceedings*, World Scientific, New Jersey, pp. 639–650.
12. Grimes,B.F. (1996) *Ethnologue: Languages of the World*, 13th Edn. Summer Institute of Linguistics, Dallas, Texas, pp. 966.