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

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

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

Received September 1, 1999; Revised and Accepted October 25, 1999

ABSTRACT

We have developed a publicly accessible database (ALFRED, the ALlele FREquency Database) that catalogues allele frequency data for a wide range of population samples and DNA polymorphisms. This database is web-accessible through our laboratory (Kidd Lab) Web site: http://info.med.yale.edu/genetics/kkidd . ALFRED currently contains data on 60 populations and 156 genetic systems including single nucleotide polymorphisms (SNPs), short tandem repeat polymorphisms (STRPs), variable number of tandem repeats (VNTRs) and insertion–deletion polymorphisms. While data are not available for all population–DNA polymorphism combinations, over 2000 allele frequency tables have been entered. Our database is designed (i) to address our specific research requirements as well as broader scientific objectives; (ii) to allow researchers and interested educators to easily navigate and retrieve data of interest to them; and (iii) to integrate links to other related public databases such as dbSNP, GenBank and PubMed.

INTRODUCTION

With the approaching completion of the primary goal of the Human Genome Project, producing a consensus reference human DNA sequence, one important next step is to study the variation in that DNA sequence both among individuals within the same population and among different human populations. Individuals may have different combinations of the variants that exist and populations will have different frequencies of those variants. These levels of variation in DNA sequence have implications for the individual practice of medicine, for public health aspects of medicine, and for research design and data interpretation. At the individual level, differences in DNA sequences may determine differences in physiologic process or differences in binding of a drug to its target molecule. At the population level the different frequencies of the variants would lead to different frequencies of individuals with a disease susceptibility or resistance to a drug. At the research level, one of our studies (1) has demonstrated how failure to consider gene frequency differences for the dopamine transporter gene across Europe probably led other researchers to conclude falsely that this gene affected smoking behavior. In addition to these medically relevant aspects of DNA sequence variation, patterns of the variation among populations can be used to trace population histories and determine how modern humans spread around the world (2–4).

Thus, the biomedical community needs to know not only about the existence of DNA sequence variants but also the frequencies of those variants in different populations. The fundamental information for defining the population variation is the table of gene (or allele) frequencies for a specific DNA population in well defined (geographically and/or ethnically) population samples. As the need for and the amount of such data grow, we need to meticulously catalogue the data and make them easily accessible to the public. Allele frequency data currently appear piecemeal and in diverse journals, making them difficult to locate. Moreover, as the data become voluminous, journals will not allocate space for the raw allele frequency tables. Yet those basic data must be available to the broad research communities. Indeed, there is not only a need to know the existence of such variants, but also a need to know the frequencies of the alleles in different populations. To achieve these goals, we have created ALFRED (the ALlele FREquency Database). At the time of writing, ALFRED is still evolving rapidly, in terms of the structure and contents of the data as well as the user interface.

Although the database requirements for supporting the study of human genetic variation has been well recognized (5,6), little has been described in the literature about the design and development of such databases. To the best of our knowledge, dbSNP (developed at NCBI) is the only existing (public) database that is similar to ALFRED. It was designed, however, for a different purpose and does not serve our specific needs well. For example, some population samples stored in dbSNP consist of an ethnically heterogeneous mixture of individuals while we stress the importance of using population samples that are ethnically homogeneous with attendant descriptions that are sufficiently detailed to allow the sample to be replicated. Another difference is that dbSNP does not, as yet, provide

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

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors flexible retrieval of allele frequencies. Allele frequencies are only presented within the context of the SNP description. In other words, one cannot readily generate gene frequency reports on multiple populations and multiple SNPs from dbSNP.

SCIENTIFIC CONCERNS

To provide useful allele frequency information of good quality, sound population genetic principles must be applied. Samples must be carefully selected so as to be representative of the population. To assure researchers accessing the information that samples are carefully selected, the samples must be carefully documented. To ensure that the resulting allele frequency estimates are accurate, the typing protocols must be precisely documented, since different protocols are susceptible to different sources of error. For example, if a primer designed to amplify around a variable restriction site in a PCR–RFLP system is itself on a variable site, some alleles of the PCR–RFLP will be missed in the typing because of the failure of the PCR to amplify DNA from some chromosomes. Consequently, documentation of allele frequency data should include the specific typing protocol used as well as the population sample, etc. To encourage users, especially upcoming scientists still in the educational system, to read the original literature, we also include in ALFRED both relevant literature citations and links to PubMed records where possible, allowing for rapid online retrieval of at least the article abstract, if not the full article.

IMPLEMENTATION

Instead of mingling user interface design with database design, we separate front end development (of Web user interface) from back end development (of the underlying database). The backend database for ALFRED is currently in Microsoft Access. We chose Access because of its flexibility and ease-of-use, allowing us to achieve rapid prototyping. The ability to prototype rapidly is key to facilitating user-developer interaction. In addition, Access provides a convenient GUI (Graphical User Interface), which greatly simplifies creation/modification of database objects including table definitions, table relationships and queries (views). Access is SQL compliant, with a few proprietary features such as cross-tab queries that are used to produce summary reports involving aggregation. As our data grow and performance becomes an issue, our backend database can be ported to a more powerful relational database management system (e.g., Oracle or Microsoft SQL Server) without too much effort.

The user interface of ALFRED was implemented using ASP (Active Server Page) technology, which is an integral part of the Microsoft IIS (Internet Information Server) Web server running on Windows NT. The ASP environment allows server-side code written in VBScript (Visual Basic Script) and Javascript to be embedded within HTML documents. In our case, we used mainly VBScript for server-side scripting with small amounts of Javascript for both client-side and server-side scripting. We minimized client-side scripting to avoid the problem of incompatibilities among different types and versions of Web browsers. To make our user interface code compatible with different database platforms, we used ODBC (Open Database Connectivity) to implement database access.

STRUCTURE AND CONTENTS OF DATA

The core of our database consists of tables representing information on loci and populations. The common link between these categories of data is captured in the frequency table, which contains our frequency data. Populations (represented by the Populations table) are sampled (Samples table) to determine frequencies (Frequencies table) of alleles (Alleles table) at a site (Site table) within a locus (Loci table). All publication-related information is stored in a single table (Publications). Intermediate tables are defined to link Publications to Frequencies, Samples, Sites and Loci. Other categories of information include contributors (Contributors table) and the typing protocol used for a site (Typing_Protocol table). Detailed descriptions of these tables (i.e., their fields) and a pictorial representation of the table relationships can be found in the Supplementary material accompanying this manuscript at NAR Online.

In addition to textual data, graphs and images are also included where relevant and available. For instance, histograms exist for allele frequencies for 30 populations for 16 sites on the Kidd Lab Histogram Page. These are currently linked into the database table Sites by including the HTML tag for the relevant histogram. These histograms are thereby displayed automatically as part of the detailed site information pages (see Supplementary material for amplification on detailed site pages).

The bulk of ALFRED's data comes from data generated in the Kidd Lab and stored in PhenoDB (7), a database system for storing and analyzing genotype/phenotype data for individuals in populations as well as pedigrees. Other sources include data contributed by our collaborators, and haplotype data generated by our HAPLO program (8). As of late summer 1999, ALFRED contained 19 605 frequency records (a single frequency of a single allele for a single population) for 156 sites in 137 loci. There are >60 population samples typed for these sites. Note that not all sites are typed for all populations and vice versa. Typing protocols are currently being incorporated into the database. While the exact modeling of typing protocols within the database is being worked out, we are establishing links to specific protocols within the Kidd Lab Protocols Page (which is rapidly evolving itself) using the same HTML tag approach described above for the histograms.

USER INTERFACE

We implemented a Web interface because the Internet provides the most widely available means of access. The primary design goal of our Web interface is ease of use. Multiple 'pathways' into the data allow users to access the information they want. For example, the user may choose to retrieve information on a locus, information on a population, or a table of frequencies based on one or more search specifications. All these search results provide entryways for further searching. As an example, after searching for locus information, a list of sites typed for that locus is included. By clicking on the site of choice, a page of information about that site comes up. From there, the user may choose to perform a frequency search, bringing up a table of frequencies for that site for all typed populations. We also provide the user with an option to return semicolon-delimited text for their search results. This text can be readily imported by the user into other programs such as a spreadsheet program for further analysis. Finally, our system provides a summary table that indicates the number of individuals typed in each sample for each site. Such a report is useful for project management because it reveals the status of population/loci typings.

CONCLUSIONS

We describe an allele frequency database that records gene frequency data on a diverse set of population samples and DNA polymorphisms. In constructing the database, the following were emphasized:

- Data quality. To make our data useful for meaningful scientific study, we try to ensure that each population sample analyzed is well documented. We also make an effort to describe, in a structured fashion, the details of a polymorphic site, including its typing method and the protocol used.
- User-friendly interface. We adopted the Web technology to allow our data to be broadly disseminated to the scientific community. Hypertext links are used to connect our data to other Web sites such as PubMed. As demonstrated in the search examples included in the Supplementary material accompanying this paper, our Web interface allows flexible retrieval of frequency data on any combination of populations and loci without too much learning effort. We also provide useful summary reports including one that indicates what populations are typed on what loci as well as the number of individuals typed.

By addressing these issues, the resulting Web site becomes informative and easy to use. We expect that the main users of our Web site will include medical genetic researchers and anthropologists who need to seek high-quality population-specific genetic variation data. We also envision, however, that our Web site can be accessed by other types of users including educators and students who represent the future of genetics-based research. We are currently in the process of incorporating more illustrative materials such as gene frequency histograms (in color) into the Kidd Lab Web site to help achieve this broader goal.

At the moment ALFRED serves primarily to make our data readily available to others and as a testbed for database design. If it proves generally useful, ALFRED could be expanded to contain a much broader scope of data. Any large expansion would require developing additional procedures to ensure integrity for information extracted from the literature or submitted by other researchers. At a minimum, we expect that what we and the community learn from our experience with ALFRED will help in the development of better databases to meet the future needs in this domain, whether designed by us or others.

ACKNOWLEDGEMENTS

This work is supported in part by NIH grants T15 LM07056 (PLM), LM05583 (PLM), AA09379 (KKK, JRK) and GM57672 (KKK, JRK), and NSF grant SBR9632509 (JRK).

SUPPLEMENTARY MATERIAL

Supplementary material consisting of:

- ALFRED data structure: detailed table contents;
- graphical representations of data table relationships; and
- example searches

is available at NAR online.

REFERENCES

1. Kang,A.M., Palmetier,M.A. and Kidd,K.K. (1999) *Biol. Psychiatry*, **46**, 151–160.
2. Cavalli-Sforza,L.L., Menozzi,P. and Piazza,A. (1994) Princeton University Press, Princeton, NJ.
3. Tishkoff,S.A., Dietzsch,E., Speed,W., Pakstis,A.J., Kidd,J.R., Cheung,K., Bonne-Tamir,B., Santachiara-Benerecetti,A.S., Moral,P., Krings,M., Paabo,S., Watson,E., Risch,N., Jenkins,T. and Kidd,K.K. (1996) *Science*, **271**, 1380–1387.
4. 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) *Am. J. Hum. Genet.*, **62**, 1389–1402.
5. Collins,F.S., Brooks,L.D. and Chakravarti,A. (1998) *Genome Res.*, **8**, 1229–1231.
6. Collins,F.S., Guyer,M.S. and Chakravarti,A. (1997) *Science*, **278**, 1580–1581.
7. Cheung,K.-H., Nadkarni,P., Silverstein,S., Kidd,J.R., Pakstis,A.J., Miller,P. and Kidd,K.K. (1996) *Comput. Biomed. Res.*, **29**, 327–337.
8. Hawley,M.E. and Kidd,K.K. (1995) *J. Hered.*, **86**, 409–411.