ALFRED: An Allele Frequency Database for Anthropology

Michael V. Osier,[1] Kei-Hoi Cheung,[2] Judith R. Kidd,[1] Andrew J. Pakstis,[1] Perry L. Miller,[2] and Kenneth K. Kidd[1]*

[1] *Department of Genetics, Yale University School of Medicine, New Haven, Connecticut 06520-8005*
[2] *Center for Medical Informatics, Yale University School of Medicine, New Haven, Connecticut 06520-8009*

KEY WORDS human evolution; population genetics; gene frequency

ABSTRACT The deluge of data from the human genome project (HGP) presents new opportunities for molecular anthropologists to study human variation through the promise of vast numbers of new polymorphisms (e.g., single nucleotide polymorphisms or SNPs). Collecting the resulting data into a single, easily accessible resource will be important to facilitate this research. We created a prototype Web-accessible database named ALFRED (ALelle FREquency Database, http://alfred.med.yale.edu/alfred/) to store and make publicly available allele frequency data on diverse polymorphic sites for many populations. In constructing this database, we considered many different concerns relating to the types of information needed for anthropology, population genetics, molecular genetics, and statistics, as well as issues of data integrity and ease of access to data. We also developed links to other Web-based databases as well as procedures for others to make links to the data in ALFRED. Here we present an overview of the issues considered and provisional solutions, as well as an example of data already available. It is our hope that this database will be useful for research and teaching in a wide range of fields, and that colleagues from various fields will contribute to making ALFRED an important resource for many studies as yet unforeseeen. Am J Phys Anthropol 119:77–83, 2002.

© 2002 Wiley-Liss, Inc.

The information now available from the human genome project (HGP) provides the opportunity to study the genetic variation of the human species with more power and specificity than ever before. Major efforts are underway to identify large numbers of single nucleotide polymorphisms (SNPs) (e.g., Cargill et al., 1999; Goddard et al., 2000); >1,500,000 have already been identified and are being cataloged in HGBASE (Brookes et al., 2000), dbSNP (Sherry et al., 1999; Smigielski et al., 2000), and the SNP Consortium database TSC (URLs for Web sites cited are given in the Appendix). More than 2,000 short tandem repeat polymorphisms (STRPs) have been identified and used to construct linkage maps (Dib et al., 1996). Mere knowledge of the existence of specific variation in a DNA sequence can be a starting point for research, but most applications require the frequencies of the alleles at the varying site, both for planning research projects and for statistical analyses. An obvious question is, "What allele frequency should be used and how should it be estimated?" Human population data for classical genetic markers (e.g., blood groups) collected over the last 70+ years have shown that gene frequencies are population-specific and usually vary significantly around the world, as amply demonstrated, e.g., in Cavalli-Sforza et al. (1994). It is exactly that variation that is of interest to anthropologists, because it provides information on the evolutionary histories of populations through the effects of demography and migration, among other factors, on gene frequencies. The thousands of new genetic markers identified directly in DNA, of which SNPs are the largest class, are far less well-studied for gene frequency variation than the classical markers. However, the available data confirm the earlier studies that significant allele frequency variation among populations is the expectation (e.g., Kidd and Kidd, 1996; Calafell et al., 1998; Osier et al., 1999; Jorde et al., 2000); it will be a very rare SNP, STRP, or other DNA polymorphism that has nearly the same allele frequencies around the world.

The plethora of new genetic markers is responsible for related problems, i.e., the "empty matrix" in population comparisons, and ascertainment bias in polymorphism identification. Many polymorphic sites in the autosomes have been genotyped on no

Grant sponsor: NSF; Grant numbers: 9912028, 9632509, BCS0096588; Grant sponsor: NIGMS; Grant number: GM57672; Grant sponsor: NIAAA; Grant number: AA09379.

*Correspondence to: Kenneth K. Kidd, Department of Genetics, Yale University School of Medicine, 333 Cedar St., PO Box 208005, New Haven, CT 06520-8005. E-mail: kidd@biomed.med.yale.edu

Received 16 July 2001; accepted 9 December 2001.

DOI 10.1002/ajpa.10094
Published online in Wiley InterScience (www.interscience.wiley.com).

© 2002 WILEY-LISS, INC.

more than a few populations, often using samples of only a few individuals, commonly of European or unspecified ancestry. Estimation of population relationships/similarities, however, requires that each population studied be typed for almost all markers (ideally for all) in a study, and that data on multiple marker loci be used for the greatest accuracy. Simulation studies (Astolfi et al., 1979) based on a model of random genetic drift of diverging populations have shown that well over a dozen independent biallelic loci are required for accurate estimation of true population relationships. Because it has been rare for large numbers of populations to be studied for gene frequencies at large numbers of DNA-based loci, these requirements pose serious problems for global overviews, such as the most recent "interim" analysis presented on the Kidd Lab Web Site. The most recent of those interim analyses includes 29 populations and 17 loci, only a coarse overview of world populations and, in our opinion, only a minimally sufficient number of statistically independent alleles for accurate estimation of genetic relationships among closely related populations. Adequate data are rare for a truly robust global picture. Instead, what has generally developed is that different researchers study different genetic polymorphisms in different populations, leading to a matrix of "markers studied" by "populations studied" that is largely empty. There is little ability to compare populations studied by different groups when completely different loci are used. Not only is it impossible to compare studies; it is difficult to build on the findings of published studies. This empty matrix problem arises in large part because there are now thousands of polymorphic loci known, each of which is a potentially useful genetic marker for population studies, and there is no easy way to learn what loci have already been studied on what populations: the data are scattered through the medical, genetic, forensic, and anthropological literature. Moreover, different researchers are funded to study different specific sets of populations, preventing them from undertaking anthropologically comprehensive studies, or to study different specific genes or types of polymorphisms, preventing them from undertaking coordinated studies on their samples. Additionally, the ascertainment bias resulting from many of these polymorphic sites being ascertained in samples of European ancestry is a complicating factor in these same analyses. One hope has been that the Human Genome Diversity Project (http://www.stanford.edu/group/morrinst/hgdp.html) would have resulted in greater coordination, but that effort has been stalled for a variety of reasons. In the meantime, only a few studies involve multiple markers on a global sample of specific populations (Bowcock et al., 1994; Kidd and Kidd, 1996; Calafell et al., 1998; Chikhi et al., 1998; Deka et al., 1999; Jorde et al., 2000).

Databases play a key role in modern human genetic research. While individual laboratories need to develop their own (private) databases to manage their data in specific ways, only large public data repositories can make readily available the enormous quantity of data generated in many individual laboratories. We are not aware of any existing databases (private or public) that completely meet the research needs of the human population genetics and molecular anthropology communities. Although one can find gene frequency data in Web-accessible databases such as the CEPH (Centre d'Etude du Polymorphisme Humain, Paris) genotype database, GDB (Genome Data Base), dbSNP, and HGBASE, the primary foci of these databases are not gene frequency data. The CEPH genotype database focuses on samples of related individuals (families) that are almost all of entirely European ancestry, so even frequencies based on the biologically unrelated parents or grandparents are of limited generality. GDB stores mapping data on individual human chromosomes. Information related to populations (often inadequately identified) and gene frequencies, if available, is buried (as text) within a marker on a map, making systematic retrieval of useful gene frequency data extremely difficult. The data stored in dbSNP are centered around DNA sequence variants (mostly SNPs). While gene frequencies are sometimes included as part of the text of the SNP descriptions, one cannot readily generate gene-frequency reports on multiple populations and multiple SNPs. HGBASE is similarly focused on SNP definition, not on allele frequencies.

In view of these problems, we developed a database, ALFRED (ALlele FREquency Database), with a focus on flexible storage and retrieval of nuclear polymorphism allele frequency data, excluding the nonrecombining region of the Y chromosome. The prototype is accessible via the Kidd Lab Web Site or directly from the ALFRED Web server. Our efforts in developing this protoype were focused on making our own data available to others since 1) journal limitations on length prevent publication of comprehensive frequency tables (e.g., Tishkoff et al., 1996, 1998, 2000; Calafell et al., 1998; Kidd et al., 1998, 2000a), and 2) the scientific community needs access to published and supporting data, ideally through the Web. ALFRED also serves to update data on one or a few additional populations, following publication of a paper based on the then-available data. Small increments of data, while undoubtedly valuable, do not justify a separate publication, and usually would not be accepted by reputable journals. We are now starting to expand the database to include data from the literature and data submitted from other laboratories.

ALFRED is a work in progress. Since our initial descriptions of the database (Cheung et al., 2000a,b; Osier et al., 2001), the amounts and types of data have increased, and the database structure and interface have become more sophisticated. Detailed population and sample descriptions have been added for nearly all populations. Links have been made to the *Ethnologue* (Grimes, 1996) to enhance

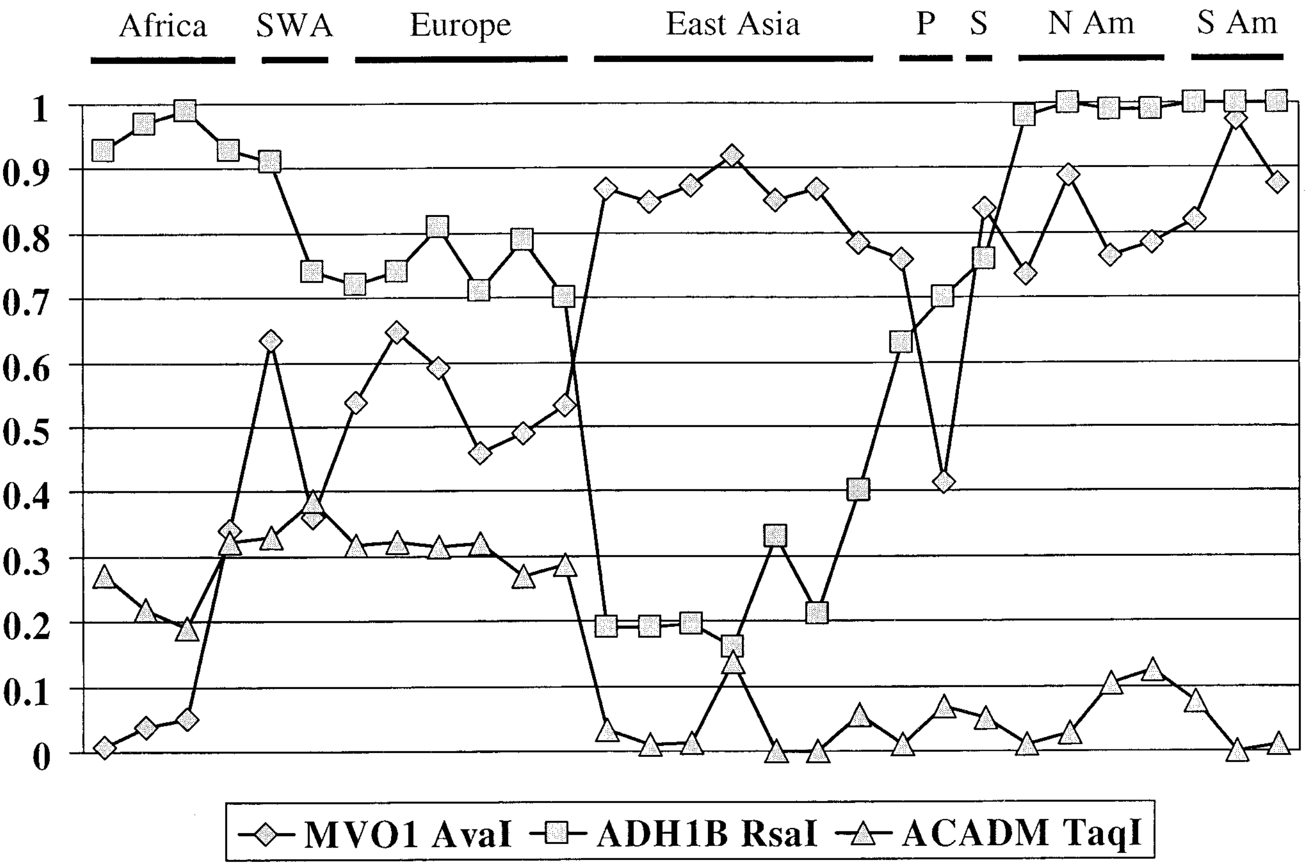

Fig. 1. Variation in allele frequencies for three biallelic sites. These data from ALFRED show frequency variation among populations within a geographic region, and variation between geographic regions, as well as a lack of global trends. The three sites are an *Ava*I site of unknown location (ALFRED UID SI000242I), an *Rsa*I site PCR-RFLP from ADH1B (SI000002C), and a *Taq*I site from ACADM (SI000162J); the *Rsa*I data are published (Osier et al., 1999), but the data on the other two sites are previously unpublished. From those site records in ALFRED, one can retrieve all the specific population and frequency data. For all three sites, the populations are in the same order, from left (Africa) to right (South America), with geographic regions marked by solid bars along the top.

these descriptions. Links are being made to dbSNP and HGBASE, and to TSC to enhance site descriptions by reference to the molecular sequences in these databases. A system of unique identifiers (UIDs) has been established so that publications can refer to specific records in ALFRED. A standard link format has been implemented so that others can make links over the Internet to the data in ALFRED. Finally, a tool to dynamically graph site frequencies and heterozygosities in an anthropological context has been added. We believe that this database can serve to connect many diverse fields interested in human populations and their histories with the highly relevant molecular data being developed. Therefore, we request community comment and suggestions so that ALFRED can evolve to more fully meet the needs of researchers in anthopology and human population genetics.

SCIENTIFIC CONCERNS

ALFRED is a sophisticated prototype focused on human gene frequency data for DNA-based polymorphisms. Several important elements relevant to such data are addressed in the database. A gene frequency is only meaningful in the context of the set of individuals who are tested. If the frequency is to be considered an estimate for the frequency in an ethnic group as a whole, the individuals tested must be carefully sampled to represent that ethnic group, and the sample must be well-documented. While gene frequencies in a geographic region tend to be similar, they can occasionally vary considerably across relatively short distances (Fig. 1). The differences across larger geographic distances are even less predictable. For these reasons, ALFRED stores information on the population name, primary language, and language family, plus a textual description of the population and a rectangle of the geographic coordinates within which that population resides. Similarly, each sample of a population is separately defined. Because different sampling regimes could result in significantly different allele frequency estimates for the same larger population, each sample of a population must be considered separately. When more than one sample has been studied, future researchers have the option of combining the data or treating them separately only if frequency estimates are given separately and referenced to a specific sample, and the samples are referenced to the larger population. Thus, an allele frequency is based on a specific sample of a population, and multiple different samples of a population are accommodated.

Another important consideration for scientists doing gene frequency studies is the definition of alleles as determined in the laboratory. While the underlying nucleotide difference at a SNP can be precisely defined, as in dbSNP and HGBASE, the laboratory assay yields a typing result (a phenotype) from which the underlying genotype is inferred. Different typing procedures are subject to different sorts of "errors." For example, any PCR-based method is subject to occasional systematic failure of amplification if there is a polymorphism in the population resulting in the occasional mismatching of a nucleotide present at the 3′ end of one of the primers. Such a mismatch variant could be common in some populations and, since it is molecularly close, it is likely to be in disequilibrium with one of the alleles at the polymorphism being typed. Such a variant could result in preferential failure to detect one of the alleles, with the consequence that some heterozygotes would falsely appear as homozygotes if a perfect codominant mode of inheritance is assumed, and would lead to a distortion in the allele frequency estimate. We detected one such "null" allele at D10S591, one of the standard linkage markers (Calafell et al., 1998). D10S591 is an STRP in the ABI linkage panel, with a common mismatch variant in Danes. Another was detected at the CD4 locus in complete disequilibrium with one of the STRP alleles in the Japanese (Watanabe et al., 1998). We detected yet another at COMT in the DNA flanking an intron 1 insertion/deletion polymorphism (unpublished). In all cases, these errors were population-specific, and a different primer would have produced a phenotype accurately reflecting a more perfect codominant system. Other typing methods (such as allele-specific oligonucleotides, TaqMan, oligonucleotide chips, and fluorescence polarization) are susceptible to their own types of "errors" in typing. Therefore, we consider it extremely important to associate the typing method used with a gene frequency. The polymorphic "sites" in ALFRED are primarily defined by their locus, alleles, and a brief description. In addition, and as a mechanism of handling the problem of the inherent errors in protocols, we are making efforts to link all sites to 1) a detailed typing protocol, presently available separately through the Kidd Lab Web Site, and 2) Web sites which provide relevant molecular definitions, such as dbSNP and HGBASE. For data from other laboratories, the structure is in place to include the full protocol within ALFRED. Links can additionally or alternatively be made to Web pages made available by those laboratories.

As we start to enter data from other laboratories into ALFRED, we plan to develop a system for processing data that should include several types of checks for completeness, inconsistencies, and errors. Association of the frequencies with the originator's (e.g., author's) identity will provide some information on quality, since laboratories have reputations. The requirement that a sufficient protocol accompany any data will allow tests of reproducibility as well as allow others to incorporate the same marker into their own research with relative ease. Of great importance will be input from the community about which criteria to implement.

RELATED DATABASES

From a social sciences perspective, one Web site that is especially relevant to the data in ALFRED is the *Ethnologue* (Grimes, 1996). The *Ethnologue* Web site has URLs that allow one to directly access either a country entry, which includes the list of all languages in that country, or 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 the *Ethnologue*. While it seems likely that tools could be written to assist curators in finding the correct URLs in the *Ethnologue* corresponding to entries in ALFRED and tying them into the ALFRED entries, at the moment we are adding these links "by hand" as personnel time allows. We expect that other Web-accessible databases relevant to the anthropologic focus of ALFRED will soon be available, or may already exist. We welcome suggestions.

As noted above, from a molecular genetic perspective, the HGBASE, dbSNP, and TSC databases are particularly important, since they define polymorphisms in terms of DNA sequence variation.

IMPLEMENTATION

The current ALFRED database is implemented using Microsoft Access, an SQL-compliant microcomputer-based database package. Access has features, including cross-tab queries, that allow users to produce summary reports involving aggregation. For example, we use cross-tab queries to generate a report that lists, for all population samples and polymorphic systems, how many individuals are typed within a population sample for a given polymorphic system.

The Web front end is based on Active Server Pages (ASP), with most of the user interface code written in Visual Basic Scripts (VBScript); the database access code is implemented using platform-independent standard Open Database Connectivity (ODBC). One advantage of using ASP is the ease with which data from databases may be accessed and published on the Web. While a small amount of client-side ASP code (using JavaScript) is used, nearly all of our code is run on the server side. We minimize client-side coding to avoid incompatibilities among different types and versions of Web browsers. We use Internet Information Server (IIS) running on Windows 2000 as our Web server (ASP is a part of IIS). The use of Access together with ASP enables us to achieve rapid prototyping, allowing user feedback during development. We currently plan to move our Access database to a more powerful database engine (Oracle) after community input into design and contents. Migration from Access to Oracle (or any relational database management system) should be straightforward because of the SQL standard. We chose Oracle because Yale University has a core Oracle service which will minimize administrative and fiscal costs and provide automatic backups of the data. The use of the Yale Oracle server will provide increased reliability and performance, providing the necessary speed and minimizing downtime.

STRUCTURE AND CONTENTS OF DATA

ALFRED is a relational database, with the data stored in tables that are cross-referenced in a variety of ways. Aspects of the structure and examples of some tables as they existed at the time are given in Cheung et al. (2000b). Detailed descriptions of these tables (i.e., their fields) and a pictorial representation of the table relationships as they currently exist can be found on the "About ALFRED" Web page available through the Kidd Lab Web Site or within ALFRED itself. Some relationships among records in the various tables are quite specific. For example, a population (represented in the Populations table) is represented by a specific sample (Samples table) used to determine frequencies (Frequencies table) of alleles (Alleles table) at a site (Sites table) within a locus (Loci table). All publication-related information is stored in a single table (Publications table), and intermediate tables are defined to link Publications to Frequencies, Samples, Sites, and Loci. Links to other Web pages are stored in a single table (URLs table) which is linked within ALFRED to the Loci, Sites, Populations, and Publications tables through intermediate tables. Contributor information is also stored, including contact information (Contributors table). The typing protocol used for a site can be stored directly in the database (Typing_ Protocol table) or linked to another Web page (through the intermediate Site_URL table to the URLs table).

Presently, the bulk of ALFRED's data has been generated in the Kidd Lab. The data are stored locally in PhenoDB (Cheung et al., 1996), a database system for storing and analyzing allele genotype/ phenotype data for individuals in populations as well as pedigrees, from which frequencies can be generated and uploaded into ALFRED. Other sources include data contributed by our collaborators, and haplotype data generated by our HAPLO program (Hawley and Kidd, 1995). As of September 2001, ALFRED contained 3,561 frequency tables (a single population sample typed for a single polymorphism or for a haplotype) for 231 sites (or haplotypes) in 150 loci. Of these sites at least 56 are SNPs, 25 are haplotype frequency estimates, and 135 are indels or STRPs. There are >90 population samples typed for at least one of these sites. Note that not all sites are typed for all populations, and vice versa.

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. From the first page of the interface, it is possible to obtain information about ALFRED, see some sample searches, and obtain some summaries and overviews of the data in ALFRED. Of the summaries retrievable, the Sites List is probably the most useful. It will retrieve a list of loci sorted alphabetically by locus name, with the polymorphic sites nested under each locus. The site names are active links to go directly to the record for that site. In addition, there are three pathways for retrieving data from ALFRED: a basic search, a search by locus, and a search by population. Ultimately, all pathways can retrieve tables of allele frequencies for subsets of populations and polymorphisms. Those frequency tables can be retrieved in semicolon-delimited format and readily copied onto a local workstation and uploaded into a spreadsheet for whatever displays and/or analyses the user desires.

The most flexible search page is the basic search. From this screen, one is able to go directly to any given record in the database if the UID for that record is known. For example, typing in any one of the site UIDs from Figure 1 and clicking on the search button will retrieve the specific record for that polymorphic site. This retrieval option is useful when publications or other databases give a unique ALFRED identifier as a reference. The remaining options are intended to allow exploration and/or retrieval without knowledge of the unique identifiers. For example, the user can search based on fields such as the locus name, locus symbol, and/or site name. With the exception of using wildcards (in this case, the "%" symbol), the searches require an exact match between the search text and the field contents.

We expect that the most common search pathways will be through the Loci and Populations options. These search pathways are organized similarly in a two-tiered heirarchical fashion. For loci, the first level is by chromosome, since the chromosomal location of a desired locus will generally be known. Clicking on the Loci option brings up a page with a list of chromosome numbers on the left. Clicking on a particular chromosome brings up the list of loci on that chromosome for which data exist in the database. The cursor can then be moved to a specific locus name, and a click brings up the page for that locus. The locus page lists all of the individual polymorphic sites (and haplotype systems) for which data exist. Clicking on one of those sites brings up the site description page from which one can retrieve the available frequency tables or graphs of allele frequencies or of estimated heterozygosities. A search starting from the Populations option follows logic similar to that for loci. The first page has a list of geographic regions on the left. Clicking on a region name brings up the list of populations in that region with data in the database. Clicking on a population name brings up descriptive information on that population and on the sample(s) for which allele frequencies have been calculated.

Two helpful features are the dynamic graphing of allele frequencies or estimated heterozygosities for a single site. These allow a quick overview of the variation among populations. From the detailed information page for the site of interest, the user can choose "Graph allele frequencies at this site" or "Graph estimated heterozygosities for this site" to dynamically produce the selected graph in a "stacked bar" histogram format, with populations sorted by rough geographic region. Currently the allele frequency output is limited to plots of the frequencies of the 10 most common alleles, on average, in the populations studied, plus a "residual" which is the sum of all other alleles in each population sample.

We are also working on improving the query interface. For instance, one path already implemented and available from the Basic Search page allows the user to choose from a list of populations and then pull up a list of all loci genotyped for any or all of those selected populations. This search query will then return a table of frequencies for that combination of populations and sites. The user might, however, desire the opposite search pathway: choose from a list of loci, and then pull up a list of all populations genotyped for any or all of those selected loci. This pathway has also been implemented. Other intuitive search pathways are being implemented to ease searching for data in ALFRED.

A RESOURCE FOR TEACHING

ALFRED is of value as a resource for student projects. Assembling large data sets of allele frequencies at multiple loci for multiple populations is difficult. Computer programs for analyses of such data are readily available (e.g., PHYLIP, Beerli and Felsenstein, 1999; ARLEQUIN, Schneider et al., 2000), but DNA-based marker data are not. ALFRED provides that source and provides it in an electronic form. Examples of possible projects such as phylogenetic trees and principal components analysis are given in the Teaching Aids section of the Kidd Lab Web Page (http://info.med.yale.edu/genetics/kkidd/teaching_aids.html).

CONCLUSIONS

Some issues that can be addressed using data from an allele frequency database are as follows: 1) Anthropologists can trace ancient migrations of human populations by comparing allele frequency data on contemporary populations across geographic regions. Clinal distributions in frequencies across long distances are indicative of large-scale population expansions and/or admixture; on the other hand, abrupt discontinuities in frequencies may be the result of more specific population movements from one location to a distant location. 2) Prehistoric demographic parameters of human populations may be estimated from a comparison of allele frequency data of contemporary human populations. Because allele frequencies change, in the absence of gene flow, as a function of time and effective population size (and, in some cases, selection), comparative data will allow relative estimates of population size, time since separation of various populations, and (perhaps) support for hypothesized selection at certain loci. 3) Linguistic similarities and genetic similarities between contemporary populations can be strong evidence for a recent common origin. Exceptions to such concordance are particularly noteworthy, and may be indicative of political rather than demic factors. 4) Theories developed from simulations or other mathematical formulations can be tested against real data. 5) As it becomes more feasible to type ancient DNA for variation at nuclear DNA loci, data on contemporary populations for the same loci will make inferences of ancestral relationships possible. 6) Population relationships based on genetic similarity can be inferred as a large amount of allele frequency data on a large number of populations becomes accessible. 7) The range of allele frequency patterns generated by population histories can be evaluated to provide a statistical basis for identifying the outlier loci that represent adaptive changes to selective pressures. 8) Researchers interested in a polymorphism in a particular gene will have a single source for an overview of variation at the locus. With just this brief enumeration of a few of the uses for an allele frequency database, two issues become clear: 1) an allele frequency database will provide an interface between disparate disciplines (e.g., linguistics, ethnography, population genetics, archaeology, and paleontology) and 2) to be maximally useful, such a database must contain clear descriptions of populations and loci, contain a large number of loci with data for many populations, and have links to other databases such as those for linguistics, polymorphisms, and the primary literature. To be maximally useful, the design and interface for a database such as ALFRED need the input of users. We plan to implement a feedback/comment section associated with the database to accumulate and publicize comments. Additionally, an external advisory board composed of members of the community will help guide development. ALFRED is designed to be a public resource. We invite readers to use it and send us their suggestions for improving its utility.

ACKNOWLEDGMENTS

Development of the prototype of ALFRED was provided by NSF grants 9912028 and 9632509 (to J.R.K.), NIGMS grant GM57672, and NIAAA grant AA09379. Funding for the continuing development of ALFRED is provided by NSF grant BCS0096588.

LITERATURE CITED

Astolfi P, Piazza A, Kidd KK. 1979. Testing of evolutionary independence in simulated phylogenetic trees. Syst Zool 27:391–400.

Beerli P, Felsenstein J. 1999. Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. Genetics 152:763–773.

Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR. 1994. High resolution of human evolutionary trees with polymorphic microsatellites. Nature 368:455– 457.

Brookes AJ, Lehvaslaiho H, Siegfried M, Boehm JG, Yuan YP, Sarkar CM, Bork P, Ortigao F. 2000. HGBASE: a database of SNPs and other variations in and around human genes. Nucleic Acids Res 28:356–360.

Calafell F, Shuster A, Speed WC, Kidd JR, Kidd KK. 1998. Short tandem repeat polymorphism evolution in humans. Eur J Hum Genet 6:38–49.

Cargill M, Altschuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, Lane CR, Lim EP, Kalyanaraman N, Nemesh J, Ziaugra L, Friedland L, Rolfe A, Warrington J, Lipshutz R, Daley GQ, Lander ES. 1999. Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat Genet 22:231–238.

Cavalli-Sforza LL, Menozzi P, Piazza A. 1994. The history and geography of populations. Princeton: Princeton University Press.

Cheung K-H, Nadkarni P, Silverstein S, Kidd JR, Pakstis AJ, Miller P, Kidd KK. 1996. Pheno DB: An integrated client/server database for linkage and population genetics. Comput Biomed Res 29:327–337.

Cheung KH, Osier MV, Kidd JR, Pakstis AJ, Miller PL, Kidd KK. 2000a. ALFRED: an allele frequency database for diverse populations and DNA polymorphisms. Nucleic Acids Res 28:361–363.

Cheung KH, Miller PL, Kidd JR, Kidd KK, Osier MV, Pakstis AJ. 2000b. ALFRED: a Web-accessible allele frequency database. In: Altman RB, Dunker AK, Hunter L, Lauderdale K, Klein TE, editors. Pacific Symposium on Biocomputing 2000 Proceedings. River Edge, NJ: World Scientific. p 639–650.

Chikhi L, Destro-Bisol G, Pascali V, Baravelli V, Dobosz M, Barbujani G. 1998. Clinal variation in the nuclear DNA of Europeans. Hum Biol 70:643–657.

Deka R, Guangyun S, Smelser D, Zhong Y, Kimmel M, Chakraborty R. 1999. Rate and directionality of mutations and effects of allele size constraints at anonymous, gene-associated, and disease-causing trinculeotide loci. Mol Biol Evol 16:1166–1177.

Dib C, Faure S, Fizames C, Samson D, Drouot N, Vignal A, Millasseau P, Marc S, Hazan J, Seboun E, Lathrop M, Gyapay G, Morissette J, Weissenbach J. 1996. A comprehensive genetic map of the human genome based on 5,264 microsatellites. Nature 380:152–154.

Goddard KA, Hopkins PJ, Hall JM, Witte JS. 2000 Linkage disequilibrium and allele-frequency distributions for 114 single-nucleotide polymorphisms in five populations. Am J Hum Genet 66:216–234.

Grimes BF. 1996. Ethnologue: languages of the world (13th edition). Dallas, TX: Summer Institute of Linguistics.

Hawley ME, Kidd KK. 1995. HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. J Hered 86:409–411.

Jorde LB, Watkins WS, Bamshad MJ, Dixon ME, Ricker CE, Seielstad MT, Batzer MA. 2000. The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data. Am J Hum Genet 66:979–988.

Kidd KK, Kidd JR. 1996. A nuclear perspective on human evolution In: Boyce AJ, Mascie-Taylor CGN, editors. Molecular biology and human diversity. Cambridge: Cambridge University Press. p 242–264.

Kidd KK, Morar B, Castiglione CM, Zhao H, Pakstis AJ, Speed WC, Bonne-Tamir B, Lu R-B, Goldman D, Lee C, Nam YS, Grandy DK, Jenkins T, Kidd JR. 1998. A global survey of haplotype frequencies and linkage disequilibrium at the DRD2 locus. Hum Genet 103:211–227.

Kidd JR, Pakstis AJ, Zhao H, Lu R-B, Okonofua FE, Odunsi A, Grigorenko E, Bonne-Tamir B, Friedlaender J, Schulz LO, Parnas J, Kidd KK. 2000a. Haplotypes and linkage disequilibrium at the phenylalanine hydroxylase locus, *PAH*, in a global representation of populations. Am J Hum Genet 66:1882–1899.

Osier MV, Cheung K-H, Kidd JR, Pakstis AJ, Miller PL, Kidd KK. 2001. ALFRED: an allele frequency database for diverse populations and DNA polymorphisms—an update. Nucleic Acids Res 29:317–319.

Osier MV, Pakstis J, Kidd JR, Lee J-F, Yin S-J, Ko H-C, Edenberg HJ, Lu R-B, Kidd KK. 1999. Linkage disequilibrium at the ADH2 and ADH3 loci and risk of alcoholism. Am J Hum Genet 64:1147–1157.

Price DH. 1989. Atlas of world cultures: a geographical guide to ethnographic literature. Newbury Park: Sage Publications.

Schneider S, Roessli D, Excoffier L. 2000. Arlequin version 2.000: a software for population genetics data analysis. Geneva: Genetics and Biometry Laboratory, University of Geneva, Switzerland.

Sherry ST, Ward M, Sirotkin K. 1999. dbSNP—database for single nucleotide polymorphisms and other classes of minor genetic variation. Genome Res 9:677–679.

Smigielski EM, Sirotkin K, Ward M, Sherry ST. 2000. dbSNP: a database of single nucleotide polymorphisms. Nucleic Acids Res 28:352–355.

Tishkoff SA, Dietzsch E, Speed W, Pakstis AJ, Cheung K, Kidd JR, Bonne-Tamir B, Santachiara-Benerecetti AS, Moral P, Watson E, Krings M, Pääbo S, Risch N, Jenkins T, Kidd KK. 1996. Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. Science 271:1380–1387.

Tishkoff SA, Goldman A, Calafell F, Speed WC, Deinard AS, Bonne-Tamir B, Kidd JR, Pakstis AJ, Jenkins T, Kidd KK. 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.

Tishkoff SA, Pakstis AJ, Stoneking M, Kidd JR, Destro-Bisol G, Sanjantila A, Lu RB, Deinard AS, Sirugo G, Jenkins T, Kidd KK, Clark AG. 2000. Short tandem-repeat polymorphism/alu haplotype variation at the PLAT locus: implications for modern human origins. Am J Hum Genet 67:901–925.

Watanabe G, Umetsu K, Yuasa I, Suzuki T. 1998. Simultaneous determination of STR polymorphism and a new nucleotide substitution in its flanking region at the CD4 locus. J Forensic Sci 43:733–737.

APPENDIX
URLs

ALFRED: http://alfred.med.yale.edu/alfred/
CEPH genotype database: http://cephb.fr/cephdb
GDB: http://www.gdb.org
TSC: http://snp.cshl.org/index.html
dbSNP: http://www.ncbi.nlm.nih.gov/SNP
HGBASE: http://hgbase.cgr.ki.se
Kidd Lab Web Site: http://info.med.yale.edu/genetics/kkidd
Ethnologue: http://www.sil.org/ethnologue