Pacific Symposium on Biocomputing 5:636-647 (2000)

ALFRED: A WEB-ACCESSIBLE ALLELE FREQUENCY DATABASE

KEI-HOI CHEUNG, PERRY L. MILLER
Center for Medical Informatics, Yale University School of Medicine, 333 Cedar Street, P.O. Box 208009 New Haven, CT 06520-8009, USA
kei.cheung, perry.miller@yale.edu

JUDITH R. KIDD, KENNETH K. KIDD, MICHAEL V. OSIER, ANDREW J. PAKSTIS
Department of Genetics, Yale University School of Medicine, 333 Cedar Street, P.O. Box 208005 New Haven, CT 06520-8005, USA
kidd@biomed.med.yale.edu
andrew.pakstis, michael.osier@yale.edu

Abstract

We present a Web-accessible database (ALFRED) that allows public access to gene frequency data for a diverse set of population samples and genetic systems. The data in ALFRED are modeled based on the experience and needs of a single laboratory, but with the expectation that the database will meet the needs of a much broader scientific community that needs population-specific gene frequency estimates. Our database currently contains data on more than 40 populations representing most major regions of the world and data on more than 150 genetic systems including SNPs, STRPs, and insertion-deletion polymorphisms. While data are not available for all population-genetic system combinations, over 2000 allele frequency tables already exist. In this paper, we enumerate the broad needs in the scientific domain, describe their significance, and describe how we have designed the database to meet those needs. We compare our database with dbSNP, the NCBI database that has a broader but overlapping purpose.

Introduction

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. Mere knowledge of 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. But what allele frequency should be used and how should it be estimated? Human population data collected over the last 70+ years for classical genetic markers (blood groups, etc.) have amply demonstrated that there is no such thing as "the general population" when it comes to gene frequencies – gene frequencies are population specific and usually vary significantly around the world (1). The recent advent of thousands of new genetic markers identified directly in DNA, of which Single Nucleotide Polymorphisms (SNPs) are the largest class, are far less well studied for gene frequency variation, but confirm the earlier studies that significant allele frequency variation among populations is the expectation; it will be a very rare SNP that has nearly the same allele frequencies around the world.

Databases play a key role in the modern human genetic research we discuss here. While individual laboratories need to develop their own (private) databases to manage their data in specific ways, large public data repositories are needed to

distribute data generated by individual laboratories in support of a broad dissemination of scientific knowledge. We are not aware of any existing databases (private or public) that completely meet our research needs. Although one can find gene frequency data in databases such as the CEPH genotype database (http://www.cephb.fr/cephdb/), GDB (http://www.gdb.org/), and dbSNP (http://www.ncbi.nlm.nih.gov/SNP/), the primary focus of these databases is not on gene frequencies. The CEPH genotype database focuses on samples of related individuals (families), while gene frequencies should be determined based on unrelated individuals. Moreover, the families 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 and gene frequencies, if available, is buried (as text) within a marker on a map, making retrieval of useful gene frequency data almost impossible. The data stored in dbSNP are centered around populations and genes (mostly SNPs). However, gene frequencies are included as part of the SNP descriptions. Therefore, one cannot readily generate gene frequency reports on multiple populations and multiple SNPs. In view of these problems, we have developed a database (Allele Frequency Database, or "ALFRED") with a focus on flexible storage and retrieval of gene frequency data (accessible via http://info.med.yale.edu/genetics/kkidd/).

In developing this database we pose several questions: How are the frequency data to be assembled and made available? What are the scientific requirements for a meaningful frequency estimate? How are the data to be managed? What are the database requirements? Here we present some guidelines and examples based on our experience with many SNPs as well as other DNA polymorphisms (STRPs, VNTRs, indels, RFLPs) in more than 40 populations representing all major regions of the world. To illustrate how these questions are not fully addressed by existing databases, we compare our database with dbSNP.

The Scientific Domain

Our database is focused on human gene frequency data for DNA-based polymorphisms. There are several important elements relevant to such data. 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 well documented. Additionally, it is not generally possible to predict gene frequencies in one population from those in another population. Figure 1 shows the gene frequency variation for four independent bi-allelic polymorphisms. While gene frequencies in a geographic region tend to be similar, they can occasionally vary considerably across relatively short distances. The differences across larger geographic distances are unpredictable. Consequently, a mixed-ethnic sample is essentially worthless allowing no inference about any of the component ethnic groups. For example, the SNP screening panel assembled by NHGRI of individuals from many different

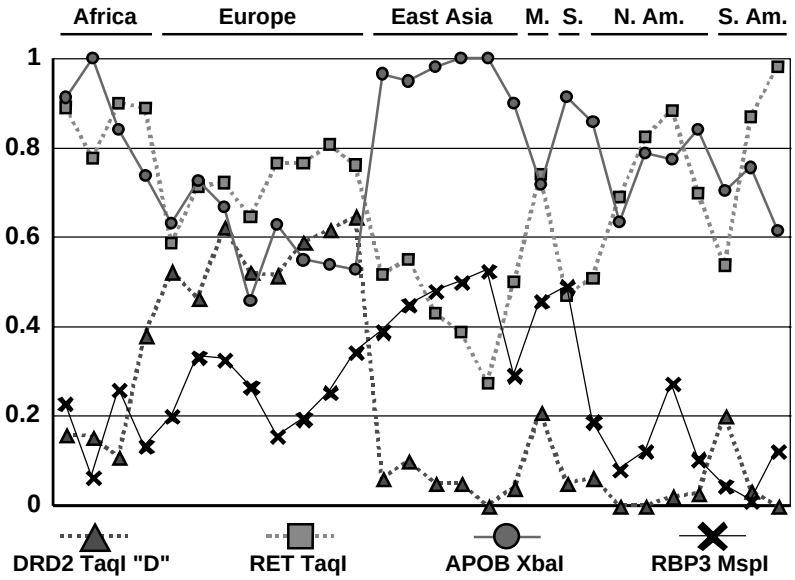


Figure 1. Allele frequency variation for four SNPs in 28 populations from Africa (4); Europe and the Middle East (8); East Asia (6); Melanesia (1); Siberia (1); North America (5); and South America (3). These largely unpublished data collected in the Kidd Lab represent some of the data already in ALFRED where specific population samples are identified

ethnic groups but with ethnic origins of individual samples deliberately expunged and unknowable will generate scientifically worthless gene frequencies.

Another important consideration is the definition of alleles as determined in the laboratory. While the underlying nucleotide difference at a SNP can be precisely defined, 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 a failure of amplification if there is a different nucleotide present at the 3'end of one of the primers. Such a variant could be common in some populations and, since it is molecularly close, likely in disequilibrium with one of the alleles at the SNP to be typed. This would result in preferential failure to detect one of the SNP alleles with the consequence that many heterozygotes would falsely appear as homozygotes since only one chromosome would allow amplification. We detected one such "null" allele at one of the standard linkage markers (2). Another has been detected at the CD4 locus in complete disequilibrium with one of the STRP alleles in Japanese (3). In both cases a different primer would have accurately genotyped the sample. Other typing methods – allele specific oligonucleotides, TaqMan, oligonucleotide chips, etc – are susceptible to their own types of "errors" in typing.

Therefore, it is extremely important to associate a gene frequency with the typing method used.

A problem in the field is determining whether the same sample of a population has been used in different studies, especially when studies of different loci by different investigators are interpreted differently. Clear documentation of a sample is important but we also want to point from a population sample to other studies done on the same sample.

Several types of questions can be approached from the types of data being assembled: (1) genetic similarities of human populations and the recent evolutionary histories of these populations; (2) inferences on general evolutionary processes responsible for the genetic similarities/differences among populations; (3) the evolutionary histories of individual genes; and others. The first of these requires reasonably complete data – the empty matrix problem – and as many loci as possible for the set of populations being studied. The third requires data on as many populations as possible as the history of the locus is entwined with the histories of the populations.

System Overview

The architectural components of our system are depicted in Fig. 2. As shown in the figure, multiple data sources can be used and referenced but currently the vast majority of our gene frequency data is obtained from our phenotype database (4) representing data produced in our laboratory. To automate the process, we have written a program to extract publication quality data from PhenoDB and load the new data into ALFRED. Currently, the program is executed manually. Periodically, we receive or request data from our collaborators who are interested in exchanging data with us. We also estimate haplotype frequencies from the raw site data using the program HAPLO (5). Both sources of data are used in publications, and therefore entered into ALFRED. The data stored in ALFRED are made accessible to the public via a Web interface that is cross-browser compatible

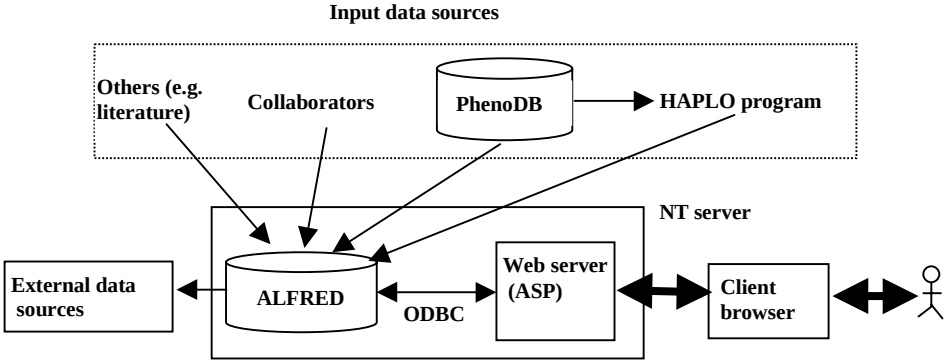


Figure 2 System overview.

(e.g., Netscape and Microsoft Internet Explorer).

To broaden the utility of our data, we are creating links to other Web sites. For example, we have established links to NCBI's PubMed (http://www.ncbi.nlm.nih.gov/PubMed/) to connect site and sample descriptions and published frequencies with the original literature. We also plan to link our data to dbSNP at NCBI. Linking ALFRED to dbSNP is important because, as discussed in the next section, the two databases store related but not identical data and ALFRED's frequency data can be submitted to dbSNP. Links between ALFRED and dbSNP will be established when we are able to identify which entries correspond (tables that store such correspondences can then be created).

ALFRED vs. dbSNP

Despite the recent recognition of the importance of providing public access to gene frequency data, inadequate efforts have been made to address the following questions:

- What kind of data should be stored?
- How should the data be organized?
- How do we ease data navigation as well as data retrieval?

Our approach is not a universal solution but is one focused on the needs of a specific domain. In the following, we compare our database with dbSNP (with the understanding that both databases are still evolving). Ours is a small-scale database designed to be used by a small number of collaborating laboratories. In contrast, dbSNP represents a large data repository designed to serve the global needs of the genome community. One of the reasons why we use dbSNP for our comparison is that it is similar to ALFRED. In addition, it is the only large public database to which we can submit our data including SNPs, populations, and gene frequencies. The emphasis of this comparison is not on promoting one database over the other (we consider the databases to be complimentary to each other). Instead we intend to use the comparison to shed some insights into the above questions.

Data Contents

A. Population Samples

At the time of writing, dbSNP contains 36 population samples and ALFRED has more than 40 populations. Most samples in ALFRED represent a distinct geographic region with a distinct ethnic background and fairly detailed descriptions of each are being entered. On the other hand, some samples in dbSNP (e.g., Cau and European) represent a mixture of individuals who come from regions that have different ethnic backgrounds. In dbSNP, some samples submitted by one laboratory seem to overlap and many samples appear from the often sketchy descriptions to be a mixture of very diverse ethnic origins. Such samples have little value for population genetics.

B. Genetic systems

ALFRED is a curated database – frequencies are only included when data reach "publication quality" in completeness and quality of typing results. dbSNP is not so curated – the individual researchers decide what is "good enough" in terms of

frequency, etc. What qualifies as "publication quality data"? Certainly not all data are scientifically meaningful. We set a minimal typed sample size of 20 individuals (40 chromosomes) to minimize standard errors. In smaller samples, standard errors are often larger than the frequency of the least common variant(s). Data in ALFRED are also either from a reasonably large number of populations typed for a single site, or a large number of sites typed on a single population.

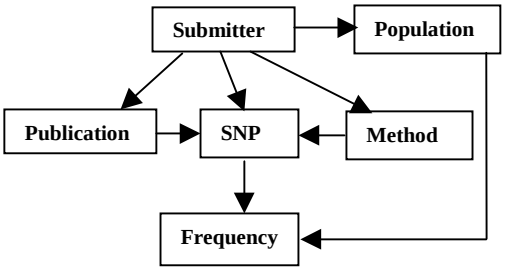
The ideal goal is to have frequency data for all sites for all populations, though this is not practical. Although we set no formal minimal number of populations typed for a site, in practice most SNP sites chosen to be placed in the database are typed on at least 15 population samples. All samples typed for less than 15 sites are part of a larger collection of samples for those sites the sample is typed for. For example, the only site in the database typed on the Woloff is the CD4 locus. However, this site is typed on a larger collection of 49 samples. This informal qualification for a minimal number population-genetic system combinations represents an attempt to minimize missing cells in the matrix. The minimization of missing data is useful for a broad range of scientific applications such as the creation of genetic distance trees in molecular anthropology. As new data become available to fill in these empty cells, they will be added into the database. An exception exists for some STRP loci used in standard linkage panels – many such markers are available for only a few populations which serve as guides to frequency variation in non-European populations.

Data Organization

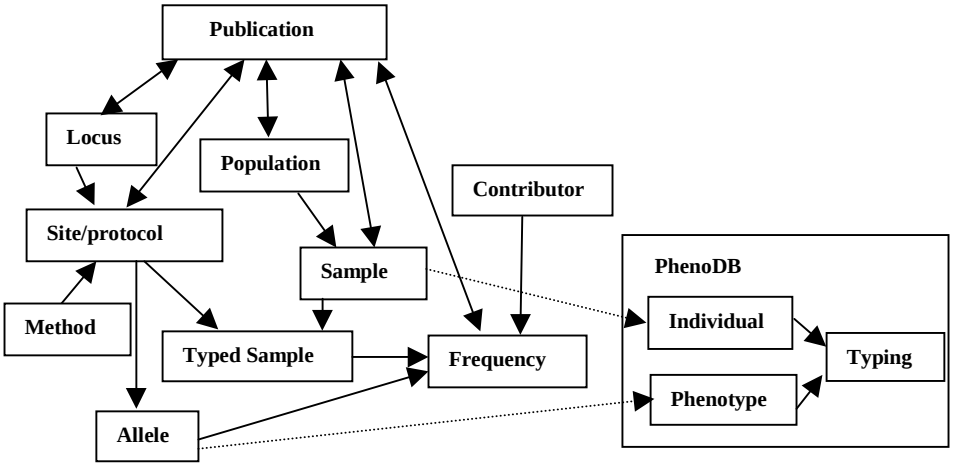
Figs. 3 (a) and (b) show the categories of data components and their relationships for dbSNP and ALFRED, respectively. As indicated in the figures, both databases store similar kinds of information in different structures. In dbSNP, there are six categories of data, namely, **submitter**, **publication**, **SNP**, **method** (which describes how an SNP can be assayed), **population**, and **frequency**. Each submitter ID typically identifies the director of the laboratory/institution submitting the data. Each submitter can submit data on multiple publications, methods, populations, and SNPs. In addition, each submitter ID can be associated with multiple batches of SNP submissions, each of which is identified by a batch ID (which usually identifies the person who actually submits the data). Each publication is linked to a list of SNPs. Multiple SNPs may be assayed using the same method. There is a many-to-many relationship between population and SNP with frequency as the bridge.

ALFRED has more data categories than dbSNP mainly because additional information is captured in modeling genetic systems and population samples. **Locus** is used to represent a DNA segment on a chromosome. Each chromosomal segment may contain multiple polymorphic sites, defined in the database as a **site/protocol**. Each site can be also characterized by multiple experimental **methods** such as PCR-RFLP or PCR-ASO or STRP. As discussed in **Scientific Domain**, different typing protocols may be vulnerable to different "errors." Therefore, ALFRED provides the

ability to store and report the full typing protocol used on a site. Each site is associated with multiple **alleles** representing the polymorphic variants.



(a)



(b)

Figure 3 (a) data components and relationships of dbSNP and (b) data components and relationships of ALFRED and a portion of PhenoDB. Such structural information was up-to-date as of July 15, 1999. This can be considered a variation of the UML notation with the rectangle boxes representing "classes" and connecting lines representing "associations". Different styles of arrows are used to indicate multiplicity of relationships. A single-headed arrow represents a one-to-many relationship; a double-headed arrow represents a many-to-many relationship.

Another difference between dbSNP and ALFRED is that the latter differentiates between samples and populations. Multiple samples of individuals can

be collected for the same population. For example, sometimes one laboratory has a sample from a single population from one village, and another laboratory has a sample from the same population from a different region. These two samples may have slight genetic differences. An additional category **typed sample** is used to specify a many-to-many relationship between samples and sites so that information such as the number of individuals within a sample typed on a site can be recorded. **Frequency** then represents a many-to-many link between **allele and typed sample**. Unlike dbSNP in which publication is only linked to a list of SNPs, publication can be linked to different data categories including sites, loci, populations, samples, and frequencies.

User Interface

The Web interface of dbSNP and that of ALFRED share many similarities. For example, both of them allow users to see a full list of populations. In addition, both systems allow genetic systems (SNPs in dbSNP) to be listed by chromosomes. In the following, we point out some of the differences between our system and dbSNP. Also, we illustrate the features that are present in our system but currently absent in dbSNP.

A. Visibility of Object Identifier

Both dbSNP and ALFRED incorporate an object-oriented notion, namely, object identity into their database designs. That is, each database record is uniquely identified by an object identifier (OID), which is computer-generated. However, different decisions are made on the visibility of OIDs. In dbSNP, OIDs such as submitter handles and population IDs are always displayed to the user, but they are not in ALFRED. We decide to hide OIDs whenever possible because the user may find such identifiers difficult to interpret (especially when there are no explanations next to them), while standardized names (e.g., locus and population names) are intuitive for the human user.

B. Frequency data retrieval

The current interface of dbSNP seems to be designed with a primary focus on SNPs. In other words, information such as gene frequency is contained as part of the information of a SNP. As an example of showing how gene frequency information can be buried deeply in dbSNP, we describe below the steps that are required to retrieve gene frequency data on a population/locus combination.

- list all populations
- click on a population ID
- click on the SNP list link to obtain the list of SNPs that are typed on the population
- click on a SNP ID for its detailed information, which then includes the gene frequency information.

In contrast to dbSNP, the primary focus of ALFRED is on allele frequency data. Because of this, we allow the user to access allele frequency more readily. Fig. 4 llustrates how a user can retrieve gene frequency information for two populations (Ami and Cheyenne) and all the typed systems on chromosome 5. Note that

multiple population names (separated by commas) can be entered in the population field. Fig. 5 shows the corresponding gene frequency results. As shown in the figure, there is only one locus (SLC6A3) on chromosome 5 that is typed on the two populations.

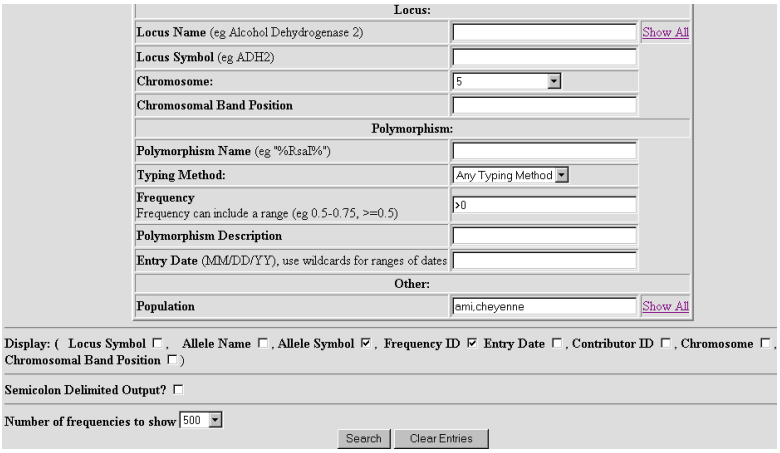


Figure 4 A search form illustrating how a query for retrieving gene frequency data can be constructed.

Population Name	Locus Name	Polymorphism Name	Sample Size (2N)	Sample ID	Allele Symbol	Frequency
Ami	SLC6A3	3-prime-VNTR	76	2	10	0.868
				2	9	0.132
Cheyenne	SLC6A3	3-prime-VNTR	112	23	10	0.982
				23	9	0.018

Figure 5 Results of the example query shown in Figure 4.

When a list of populations or loci is displayed, dbSNP allows the user to click on only one item for detailed information. On the other hand, ALFRED allows the user to select multiple items by making checkbox selection on each row. This provides flexibility for the user to make associations that otherwise might not have been made. It also allows the user to customize which information to export when requesting semicolon-delimited results for their own analyses. There is no easy export mechanism in dbSNP.

C. Summary Report on Frequency Data

Since the emphasis of ALFRED is on gene frequency data, it is useful to report on what populations are typed on what loci (currently, this feature is not available in dbSNP). Figure 6 shows such a report in a table. In the figure, we see that each

column label (except for the first one) represents a site and each row label (except for the first one) represents a population. The number in each table cell represents the number of chromosomes that are typed for the population/locus pair. Dividing the number by two gives the actual number of individuals typed since each individual has two chromosomes. If the number is zero, it means that the corresponding sample is not typed on the corresponding locus.

Population	3-prime-VNTR	ACADM/TAQ	APOB EcoRI	APOB XbaI	CD4 Alu deletion	CD4 CTTTT pentanucleotide repeat	CFTR CA Repeat	D10S189	D10S190	D10S191
Adygei	108	104	108	108	108	100	104	0	0	0
Ami	76	78	80	80	78	70	0	0	0	0
Arizona Pima	90	102	74	90	0	0	0	0	0	0
Arizona Pima	0	0	0	0	84	74	0	0	0	0
Ashkenazi Jews	0	0	0	0	102	102	0	0	0	0
Asiatic Indians	0	0	0	0	108	108	0	0	0	0
Atayal	84	84	84	84	84	80	0	0	0	0
Basque	0	0	0	0	128	124	212	0	0	0
Biaka Pygmies	130	136	136	136	134	122	126	132	0	134
Cambodians	50	0	44	50	50	48	0	0	0	0

Figure 6 A summary table on population/loci typings.

Implementation

We adopt a PC-based technology to implement our system. Our general belief is that a PC-based system is cheaper and easier to maintain. We have established experience of using such a technology to build a number of client-server database systems (e.g., NeuronDB [6] and ACT/DB [7]).

Our backend database was implemented using Microsoft Access running on an NT server (version 4.0 with service pack 5). The main reasons for using Access include its ease of use and flexibility. For example, it allows queries (views) including cross-tab queries to be defined graphically based on existing tables and queries. With Access, we can rapidly prototype our database for user feedback. In addition, Access is SQL-standard. Therefore, if performance becomes an issue because of large data size, one can port an Access database to a more powerful database management system (e.g., Oracle) without too much effort.

Our NT server comes with the IIS (Internet Information Server) Web server. Our Web interface was implemented using ASP (Active Server Page), which is a part of the Web server. ASP allows VBScript or JavaScript to be embedded in HTML documents and executed on either the client or server side. We chose to use only server-side scripting because of browser compatibility. The ASP server-side scripting approach is more flexible and efficient (due to the shared memory model)

than the traditional CGI approach. In coding database access, we use ODBC (Open Database Connectivity) scripting mechanism, which allows the same code to be written to access different types of database systems including relational systems such as Oracle, Sybase, and Access as well as object-oriented systems such as Gemstone.

To ease interaction with other databases, the interface scripting generally utilizes the "post" method of passing form data to the receiving ASP script in a predictable format. When using "post", all the information passed to the receiving script is visible in the URL of the receiving script. This allows creators of other WWW pages to easily create a link to ALFRED without needing to use a complicated form construct on their page. Instead, they can simply copy and paste a URL from the report generated by ALFRED in their browser into their HTML document. The "post" method is used by NCBI's resources, such as PubMed, and is what made addition of links to PubMed citations easy for our interface construction.

Discussion and Future Directions

In the construction of SNP databases, there are many subtle complexities which must be planned for long before construction of the database can begin. In this paper, we have focused on data stringency, data management, and implementation. Data stringency is qualified by sufficiently large typed samples, and minimization of untyped population-genetic systems combinations. The implementation is focused on speed, ease of construction, and ease of access and integration with other WWW projects. The data management principles maximize the quantity of represented information and ease of data entry and retrieval.

Data stringency requires basic statistical soundness, such as minimal sample sizes, as well as meeting the needs of population genetic analyses. Minimizing missing data creates a better framework to test scientific questions. For instance, a genetic distance tree based upon 15 samples provides more information than a tree based upon 5 samples. Also, biotech companies are starting to focus on "designer drugs." A therapy might be useful for those individuals with one allele, but not the other. That allele may be common in one population, but not in others, meaning the drug is useful only in a specific portion of the world. A reasonably complete and statistically robust data set is very important to answer many scientifically interesting questions.

The WWW interface not only makes the data readily accessible to the majority of the scientific community and WWW developers, but can also be utilized in other settings such as the classroom. While browsing through the data, additional representations of the information can be included in the reports. For example, when looking at the information on the DRD2 TaqI "A" site, a histogram of allele frequencies in a wide range of populations is easily appended to the page, providing a more comprehensible view of the data. Citations to the original literature not only provide vital links for the scientific community, but also encourage classroom viewers to investigate the information they are examining. Detailed population and sample descriptions not only inform the researcher who is viewing the data of

potential ascertainment bias, but can spur the imagination of a future geneticist or anthropologist. Readily accessible data is beneficial for a wide range of classroom uses.

Data management is centered around maximizing information represented in the database while easing data entry. Currently, data can be generated faster than is practical for manual entry. Methods of automating data entry are critical. Additionally, logical links between types of data, such as a polymorphic site and its typing protocol, locus, and citation, enhance the value of the data. A format of data representation that makes navigation easy and logical for the end user is just as important as the rapid entry of data.

In the future, we feel developing the database along the guidelines above will increase its usefulness. As with any database, it is important to focus on ease of data entry and ease of accessing data. Through evolving a single pipeline into the database that can draw on multiple sources of data, entry of new data can be more efficiently automated. Through providing more intuitive means to access the data, such as searching for loci, site, and sample information directly instead of through the results of a frequency search, accessing the data will become easier and more useful for the end user. Emphasizing the concepts of automated data entry and an intuitive user interface for the end user improves the usefulness of any database of human genetic variation.

Acknowlegments

This work is supported in part by NIH grants T15 LM07056 (PLM), AA09379 (KKK,JRK), and GM57672 (KKK,JRK), National Library of Medicine grant G08 LM05583 (PLM), and NSF grant SBR9632509 (JRK).

References

1. L.L. Cavalli-Sforza P. Menozzi, A. Piazza, (Princeton University Press, Princeton, New Jersey, 1994).

2. F. Calafell, A. Shuster, W.C. Speed, J.R. Kidd, K.K. Kidd, Eur. J. Hum. Genet. 6:38-49 (1998)

3. G. Wantanabe, K. Umetsu, I. Yuasa, and T. Suzuki, J. of Foren. Sci. 43:733-737(1998)

4. K.-H. Cheung, P. Nadkarni, S. Silverstein, J.R. Kidd, A.J. Pakstis, P. Miller, and K.K. Kidd, Computers and Biomedical Research 29:327-337 (1996)

5. M.E. Hawley and K.K. Kidd, J. of Hered. 86:409-11 (1995)

6. L. Marenco, P. Nadkarni, E. Skoufos, G. Shepherd, D. Phil, and P. Miller, AMIA, In press (1999)

7. P. Nadkarni, C. Brandt, S. Frawley, F. Sayward, R. Einbinder, D. Zelterman, L. Schacter, and P. Miller, J. of the Amer. Med. Informatics Assoc. 5, 139-151 (1998)