

PhD: a web database application for phenotype data management

J.-L. Li^{1,2,*}, M.-X. Li¹, H.-Y. Deng³, P.E. Duffy², and H.-W. Deng^{1,3,4}

1. Laboratory of Molecular and Statistical Genetics, College of Life Sciences, Hunan Normal University, Changsha, Hunan 410081, P. R. China

2. Seattle Biomedical Research Institute, 307 Westlake Avenue N, Suite 500, Seattle, WA 98109-5219, USA

3. Osteoporosis Research Center, Creighton University Medical Center, Omaha, NE 68131, USA

4. The Key Laboratory of Biomedical Information Engineering of Ministry of Education, and Institute of Molecular Genetics, School of Life Science and Technology, Xi'an Jiaotong University, Xi'an 710049, P.R.China.

Running title: Phenotype database management

* To whom Correspondence should be addressed at

Seattle Biomedical Research Institute

307 Westlake Avenue N, Suite 500, Seattle, WA 98109-5219, USA

Tel: +1 206 256 7483; Email: Jinlong.li@sbri.org

ABSTRACT

Summary: A database application has been developed for phenotype data management employing the Entity-Attribute-Value (EAV) model. By applying the EAV model, this application allows users to manage arbitrary phenotypes and customize data entry forms (DEFs); therefore it is suitable for different and multi-center projects.

Availability: <http://apps.sbri.org/gpdb> (Beta version).

Contacts: Jinlong.Li@sbri.org

Supplementary information: <http://apps.sbri.org/gpdb/supp.htm>

INTRODUCTION

In genetics and genomics studies of complex human diseases, large numbers of data points are commonly acquired. Efficient and effective management of the vast amount of data is then necessary. We developed a genotype database management system (DBMS), GenoDB, for large-scale management of genotype data derived from fluorescently labeled short tandem repeat microsatellite markers (Li et al. 2001). Recently, we developed another DBMS, SNPP, for efficient and convenient management of large-scale single nucleotide polymorphism (SNP) data (Zhao et al. 2004). However, both GenoDB and SNPP lack functions to manage phenotype data, which are essential in genetics and genomics studies for complex human diseases.

Phenotypes are heterogeneous for projects that have different aims. For instance, to search genes underlying osteoporosis, researches may focus on bone mass, while to search genes underlying obesity, phenotypes such as body mass index and percentage fat mass may be measured. In addition, data for some phenotype may often be sparse (not uniformly measured or even missing) for some subjects. One reason is that it is unnecessary to collect the phenotype data from all study subjects for some seemingly “non-essential” or unnecessary phenotypes. In addition, it may not be feasible to collect all of the phenotype data from all of the study subjects (missing data).

Instead of the conventional Entity-Relation (E-R) database model, the Entity-Attribute-Value (EAV) model, which conceptually involves tables with three columns [an entity identification, an attribute and the value for that attribute], is suitable for heterogeneous and sparse data. It has been successfully applied to building database applications to manage clinical data that are heterogeneous and sparse (Nadkarni et al. 1999). Although the domain of phenotype data intersects with that of clinical data, we should not simply adapt the clinical database application for phenotype data management. Because scientists often study phenotype, genotype and environmental data together for genetics and genomics research, it is necessary to develop a phenotype DBMS that interacts with genotype DBMS. Here, we present a phenotype DBMS (PhD), developed using the EAV database model, to complement our earlier developed genotype DBMS for microsatellite (Li et al. 2001) and SNP markers (Zhao et al. 2004).

IMPLEMENTATION**Architecture**

The server side of PhD includes a MySQL database server and Sun Java System Application Server. The software packages for installing the servers are freely available from <http://www.mysql.com/> and <http://java.sun.com/>.

Database Design (Figure 1)

In the EAV design, attributes with different data types (such as integer, real, string, etc) are separated into different tables. Therefore, each EAV table is a data type specific table. An “EAV” database may also use conventional tables when necessary and convenient (Nadkarni et al. 1999). Some tables in PhD are more suitable for E-R design; therefore, they follow E-R design rules.

Functionality

Defining, updating or deleting phenotypes

By employing EAV design, PhD allows users to manage phenotypes without predefining them in the database at the time of development. This function is particularly important for development of a general application for different projects and multi-center projects in which phenotypes may vary.

Importing phenotype data into PhD from other sources

Phenotype data can be imported into PhD from MS Excel or text delimited files. If traits in the file have all been defined in PhD, phenotype data can be directly imported into the database. Otherwise, new traits have to be defined prior to data importing.

Customizing data entry forms (DEFs) for different projects and users

One important function of PhD is that it lets users create customized DEFs and modify DEFs. This function makes PhD suitable for different and multi-center projects because DEFs are usually varied from project to project. It is also desirable even for a single project when the project evolves and the needs arise for modifying previous DEFs.

Querying, compiling and exporting phenotype data for analysis

PhD provides a function for users to query the database and export the query results for analysis. In addition, because scientists often study phenotype and genotype data together, PhD has been integrated with GenoDB (Li et al. 2001). In this way, phenotype data can be queried and compiled jointly with genotype and pedigree data for further data analysis or for feeding some popular linkage analysis software directly.

ACKNOWLEDGEMENT

The authors thank Tom Blackwell, Jake Masters and Jason Blue for their assistance in setting up the computing environment to host PhD at the Seattle Biomedical Research Institute. HWD and HYD were partially supported by grants from NIH.

REFERENCE

- Li, J.L. et al. (2001). Toward high-throughput genotyping: dynamic and automatic software for manipulating large-scale genotype data using fluorescently labeled dinucleotide markers. *Genome. Res.*, 11, 1304-1314.
- Nadkarni, P.M. et al. (1999) Organization of heterogeneous scientific data using the EAV/CR representation. *J. Am. Med. Inform. Assoc.*, 6: 478-493.
- Zhao L.J. et al. (2004) SNPP: automating large scale SNP genotype data management. *Bioinformatics*, 20, 1-3

FIGURE LEGEND

Figure 1: Database schema of PhD. The top section within the dash rectangle represents EAV design, while the bottom one represents ER design. An entity is represented as a solid rectangle with two sections. Field(s) in the top section of the rectangle denotes primary key(s). FK denotes foreign key.

Figure 1

