THE FIRST MODULAR DATABASE OF INDONESIAN GENES-ASSOCIATED DISEASES INFORMATION

1,3,4 RIKA YULIWULANDARI, 2,3 NOVA EKA DIANA, 3 KINASIH PRAYUNI
1 Department of Pharmacology, Faculty of medicine, YARSI University, Jakarta, Indonesia
2 Department of Informatics, Faculty of Information Technology, YARSI University, Jakarta, Indonesia
3 Genomic Medicine Research Group, YARSI Research Institute, YARSI University, Jakarta, Indonesia
4 The Indonesian Pharmacogenomics Working Group
E-mail: rika.yuliwulandari@yarsi.ac.id, nova.diana@yarsi.ac.id

ABSTRACT

This paper aims to develop a modular schema database for gene-associated diseases incidence in Indonesia, which is called as Genomic Medicine Research Group Database (GMRGDB). A modular architecture is beneficial to facilitate any manipulation and addition of new genes-affiliated diseases in the future. Currently, GMRGDB contains two categories, N-acetyltransferase 2 (NAT2) and Human Leukocyte Antigen (HLA) databases. NAT2 database is a repository of NAT2 genes and anti-TB drugs association, while HLA database records diseases related to Human Leukocyte Antigen. GMRGDB consists of patients’ information from various ethnics in Indonesia, such as Javanese, Melayu, Betawi, Sundanese, and Batak. Users can view a number of samples for each database, detailed representation of Single Nucleotide Polymorphism (SNP), and also filter the data for specified ethnic through GMRGDB browser.

Keywords: Genes-Associated Disease, Modular Database, N-Acetyltransferase 2, Human Leukocyte Antigen, Indonesian Ethnics

1. INTRODUCTION

Several genetic mapping studies of some ethnics in Indonesia, especially Javanese and Sundanese population, have been conducted such as Human Leucocyte Antigen (HLA) and N-Acetyltransferase 2 (NAT2) studies [1, 2]. HLA is essential in identifying histocompatibility and disease susceptibility [3, 4]. The identification of HLA is also important for anthropological studies to elucidate genetic relationship among various ethnic groups in Indonesia [5]. Meanwhile, NAT2 has a major part to activate and deactivate arylamine and hydrazine drugs and carcinogen [6]. Polymorphisms in these genes are responsible for classifying the phenotype of NAT2 into a rapid, intermediate, and slow acetylator [6, 7].

The previous study of HLA in Javanese and Sundanese population showed the allele frequency of HLA-A, HLA-B, and HLA-DRB1 alleles; and also the broad relationship between these populations with other Asian inhabitants. Yuliwulandari et al reported that HLA allele types have an association with many diseases [1]. Previously, Yuliwulandari et al in their research also revealed that the slow acetylator of NAT2 variants were common in Indonesian populations, specifically within Javanese and Sundanese ethnics [2]. And in 2016, Yuliwulandari et al showed the association among those slow acetylators and susceptibility to Anti-Tuberculosis Drug Induced Liver Injury (AT-DILI) for Indonesian population [8].

There is yet a centralized database focusing on managing gene-associated diseases in Indonesia. Most of the data are still stationed in the disparate locations and there has not been an interconnected infrastructure to communicate the data. Hence, it is arduous for researchers to learn, analyze and find out the association of genes with the diseases. A centralized and comprehensive database for these data will definitely beneficial to researchers for cultivating more salient information and utilizing them on finding the solution of gene-associated diseases.

This paper aims to develop a modular database for representing gene-associated diseases in Indonesia. The organization of the article follows.
Section 1 describes the background of the research. Section 2 explains related works about methods that are used to develop this research. Next, Section 3 defines the methodologies for implementing the system. Section 4 contains the experiment results and its evaluations. Finally, Section 5 presents the conclusion and future works of our research.

2. RELATED WORKS

Many researchers in the world have adopting database techniques in information technology area to administer genes data. The National Center for Biotechnology Information (NCBI) is the most promising research institute on genes database area. This institute had conducted their researches since 2001 by creating dbSNP, a centralized database that integrated genome variations from various sources, such as GenBank, Pubmed, LocusLink and the Human Genome Project data [9]. In 2005, NCBI upgraded their LocusLink by adding Entrez Gene database which focused on the genomes that had been fully sequenced or were scheduled for intense sequence analysis. This database represented more detailed taxonomies of integrated data from NCBI’s Reference Sequence (RefSeq) and other databases available in NCBI website [10]. NCBI kept working on updating and completing their database by appending various resources such Entrez, BLAST, RefSeq, UniGene, HomoloGene, the Conserved Domain Database (CDD), Protein Clusters, etc. Detail explanation and references about their services were described in the NCBI documentation that was released on 2013 and 2015 [11, 12].

Japan, Singapore, and Malaysia are three of the most active countries that conducting genomes-database research in Asia. Some researchers in Japan Science and Technology (JST) and The University of Tokyo had started to collect Japanese genomes variation since 2000. They built JSNP, a repository of 150,000 Japanese Single Nucleotide Polymorphism (SNP) that aim was to identify the association between the polymorphism and common disease of the drug reactions [13]. Further, many institutions collaborated to develop more comprehensive Japanese genetic database.

Five institutions: Kyoto University, Tohoku University, National Research Institute for Child Health and Development, the University of Tokyo, and Yokohama University, worked together to develop The Human Genetic Variation Database (HGVD). This database stored Japanese genetic variation of 1208 individuals. Browser interface of HGVD provides many utilities for viewing allele and genotype frequencies, eQTL significance, and also the number of samples [14]. Another research, The Malaysian Node of the Human Variome Project (MyHVP) had started on collecting genotype and phenotype variation of Malaysian population since 2010. This project focused on managing ethnic-specific database; hence the clinicians and researchers can obtain up-to-date and accurate information on population-specific culture [15].

Complete and rigorous genomic database need a thorough process to identify entities and relationship among genes-variation and diseases. The ontology-based representation is beneficial to fully describe the association among entities. This approach had been widely to portray much medical and biological information. Lawrence Berkeley National Laboratory and Harvard University created Chado to represent the ontology-based relational database schema of genome-associated biological information [16]. Ontology was also utilized to identify and represent the incidence and spread of Malaria disease in Indonesia. A thorough process of attributes and relationship identification was crucial to extract important information from a huge number of data (malaria data warehouse) [17, 18]. Besides an impeccable entities and relationship representation, a well-developed database should provide the capability to adapt with the amendment and the transformation of the increasing genomes data. Modular-database architecture is perfect to facilitate any modification and data manipulation in the future [16].

3. MATERIAL AND METHODS

2.1 Data Source

This project aims to create a dynamic database for representing the gene-associated diseases information in Indonesia comprehensively, titled “Genomic Medicine Research Group Database (GMRGDB)”. Each disease has a number of features determine the condition of a patient, and it may share some common parameters or information with other diseases. The number of parameters and diseases might also increase in the future. Hence, the database should have the modularity features to accommodate any improvements and modifications.

In this research, currently we categorized the data into two categories based on the project type:

a. NAT2 Project contains the interconnection between Anti-Tuberculosis Drug-Induced
Liver Injury (AT-DILI) enzyme called N-acetyltransferase 2 (NAT2) and anti-TB drugs.

b. HLA Project consists of diseases data which are associated with Human Leukocyte Antigen (HLA).

We have successfully collected a total of 495 gene-associated data of various ethnics in Indonesia from 2015-2016. These data were classified into a control and patient data. A control data shows no negative effect of the designated drugs to the disease. While a patient data reveals information about the negative association of the designated drugs to the disease. Table I summarizes our research data. NAT2 database contains a total of 50 controls and 173 patients information for Javanese, Sundanese, Betawi, Melayu and Batak ethnics. Meanwhile, HLA project consists of a total of 99 and 173 data for controls and patient data, respectively.

2.2 Data Features Identification

As aforementioned in Table I, there are two groups of gene-associated diseases in our dataset. Those data were classified as control and patient data. Each group has a number of features determine the patient condition. Some features may be similar for different groups. Fig. 1 illustrates the identified parameters for NAT2 and HLA projects. Entities name with a Control-prefix depicts the control data for each project, while a Patient-prefix denotes a patient data. Patient NAT2 and Patient HLA data share the same SNP attributes: sgot0, sgpt0, bilirubin0, sgot1, sgpt1, bilirubin1, bilirubinDt, history, familyHistory, and therapy parameter.

Table 1: Genes-Associated Diseases Data In Indonesia (2015-2016)

<table>
<thead>
<tr>
<th>Project</th>
<th>Data Type</th>
<th>Ethnics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Control</td>
<td>Patient</td>
</tr>
<tr>
<td>NAT2</td>
<td>50</td>
<td>173</td>
</tr>
<tr>
<td></td>
<td>Javanese, Sundanese, Betawi, Melayu, Batak</td>
<td></td>
</tr>
<tr>
<td>HLA</td>
<td>99</td>
<td>173</td>
</tr>
<tr>
<td></td>
<td>Betawi, Javanese, Sundanese, Batak, Melayu</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1: Project Features Identification
Fig. 2 depicts the ontology of our current dataset that illustrates the interrelationships of the entities. The ‘n’ symbol in the figure describes the common features owned by two or more entities. For example, ‘blood’ and ‘dna’ attributes are associated with ControlHLA and PatientHLA entity. Another parameters, ‘history’, ‘familyHistory’, and ‘therapy’, are also shared by PatientNAT2 and PatientHLA entity. There are many ‘n’ symbols in the figure, so it means that the entities have many similar features for identifying the designated disease. In this figure, there are some entities that come together to create a virtual group, and there is one entity that standing alone and does not associate with other entity (ControlNAT2 entity). Smaller the distance between two entities, then those entities are closer and share more features compare to other entity. In contrary, the apparent non-zero direct straight line connection of one entity to another, then this entity has unique parameters that do not belong to others.

2.3 Database Construction and Implementation

In this research, some entities that represent gene-associated diseases have some similar features to diagnose the patient. Furthermore, there are some additional requirements to manage the gene-associated projects in our laboratory. Those requirements are:

a. The number of gene-associated projects may increase in the future

b. A project may have zero or more than one subproject

c. Each project and subproject have a set of attributes to identify the gene-associated disease

According to the aforementioned requirements, the proposed database should able to dynamically add a new project or subproject with a set of attributes. Fig. 3 illustrates the generic schema of our genes-associated disease projects. The main entities are PROJECTS, SUBPROJECTS, and USER. PROJECTS entity records a number of projects that are currently managed in our laboratory. PROJECT_DETAIL table saves the detail information of each project in PROJECTS table. A unique id and project name describe the project. Numerous attributes give a complete description of a project, which may be different for each. A project may have zero or more than one subprojects, in which each subproject have a collection of parameters that differ to their parent (a
List of all subprojects is summarized in SUBPROJECTS table. Each project is managed by a user that has been appointed as a person-in-charge (PIC) for this project. USER table stores more detail information for all users, including their role in using the system.

**Figure 3: Generic Modular Schema Of Main Database**

Fig. 4 shows the tree-tier architecture of our genes-associated disease database, which consists of the client, server, and database. We developed a web-client interface using Bootstrap framework which based on Hypertext Markup Language (HTML), Cascading Style Sheet (CSS), JavaScript, and Hypertext Processor (PHP) language programming. Apache Server was used to mediate the Structured Query Language (SQL) request from the client to our modular database, which was developed using MySQL.

4. RESULTS AND DISCUSSION

GMRGDB can be accessed by three types of user: database administrator, project-pic, and user portal. The project-pic is a user that the database administrator has assigned as a person-in-charge for managing the project. Once a user has been admitted as a project-pic, next he can add or manipulate the genes-associated diseases data. Those data can be accessed by all user portals, anyone who know the web application address and has an internet connection to access it.

The database is organized into two main categories:

a) **NAT2 database** contains patients’ information of NAT2 enzyme and anti-TB drug association. These data was divided into a control and patient categories, which consist of Java (N=65), Betawi (N=27), Sundanese (N=12), Melayu (N=4), Batak (N=9), and unknown (N=55) ethnics.

b) **HLA database** stores a total of 272 patients data related to HLA-disease information, with Melayu (N=133), Batak (N=10), Limo Panjang (N=2), and unknown (N=48) ethnics.
Search by a project name will display all patient information of each gene-associated disease belong to the project. Fig. 5 and 6 display all patient data in NAT2 and HLA database category, respectively. These figures display the detail information of patients’ Electronic Medical Record (EMR) such as identity, jobs, ethnic, and their medical histories. The display of search result operation is structured into two levels hence it is easier for the user to track the more detail medication history of patients. GMRGDB browser is going to disclose those data when users choose the given link under SNP or Pemeriksaan link.
GMRGDB application enables users to filter the data based on the specific ethnic. Fig. 7 and 8 illustrate the results of searching for Betawi ethnic on NAT2 database and Javanese ethnic on HLA database, respectively. Through this feature, the researcher might swiftly have a broad illustration about which Indonesian ethnic that is sensitive in reacting to NAT2 enzyme after consuming Tuberculosis drugs, e.g. Isoniazid. The user may also fast to discover the ethnic inhabitants with HLA alleles that are diagnosed with more diseases. Besides search on ethnic, the user can find the number of patients that have been treated with such medicine, such as Rifampicin as illustrated in Fig. 9. In this figure, the researchers may draw valuable information by associating the medicines treatment to their diagnostic and family history. Fig. 10 presents the SNPs variation of each individual in N-acetyltransferase 2 gene that reacts to anti-TB drugs. This display exhibits alleles and genotyping information of NAT2 enzymes after reacting to the given medicines. The user can access this detail by clicking the given link under SNP column in the search result tables.
Figure 8: Search result of Javanese ethnic on HLA database

Figure 9: Genomic Medicine Research Group (Gmrg) Database Application Architecture

Figure 10: Snps Information Of NAT2 Gene For One Individual
5. CONCLUSIONS

We developed a modular architecture of Genomic Medicine Research Group Database (GMRGDB) to manage gene-associated diseases in Indonesia. This database covers most Indonesian ethnics, such as Javanese, Betawi, Sundanese, Melayu, and Batak. Currently, GMRGDB consists two database categories, N-acetyltransferase 2 (NAT2) and Human Leukocyte Antigen (HLA) group. These databases record the interconnection among diseases, genes, and reaction to drugs. GMRGDB browser provides the functions to view a number of samples for each database, find SNPs information, filter the data based on specified ethnic, and connect to other genomic browsers, e.g. NCBI website.

GMRGDB database will be continuously enriched with other gene-associated diseases. Leprae data for Indonesian population will be appended to the GMRGDB in the coming work. GMRGDB browser should provide the feature to automatically investigate the relationships between SNP and the given drugs. The associations may be visually presented by generating an interactive and informative dashboard with a befitting graphs or charts.

ACKNOWLEDGMENT

We are fully thankful to The Indonesian Directorate General of Higher Education (DIKTI) of the Ministry of Higher Education, Research and Technology, and also YARSI Foundation for providing grants to execute this research.

REFERENCES:


