LSHGD: A database for human leprosy susceptible genes

C. George Priya Doss a,⁎, N. Nagasundaram b, Jain Srajan b, Chakraborty Chiranjib b

a Medical Biotechnology Division, School of Biosciences and Technology, Center for Nanobiotechnology, VIT University, Vellore 632014, Tamil Nadu, India
b School of Biosciences and Technology, VIT University, Vellore 632014, Tamil Nadu, India

1. Introduction

The field of human genetics of infectious diseases aims to define the genetic variations accounting for inter-individual variability in the course of human infections [1]. Leprosy (Hansen’s disease) is a chronic granulomatous infectious disease of human whose etiological agent, Mycobacterium leprae, was identified by G. H. A. Hansen in the 19th century. In 2011, there were new cases reported, predominantly in Asia, Africa, and Latin America [2]. According to official reports received from WHO during 2011 from 130 countries and territories, the global registered prevalence of leprosy at the beginning of 2011 stood at 192,246 cases, while the number of new cases detected during 2010 was 228,474 (excluding the small number of cases in Europe). Leprosy is an unusual disease in many aspects, not the least of which is that, despite effective multi-drug therapy, the steady state number of leprosy cases is about equal to the annual number of new cases. These numbers are evidence of a high transmission rate. Until the modes and sources of transmission are well understood, it is unlikely that we will be able to interfere with the transmission or be able to eradicate this disease. Leprosy is primarily a disease of the skin and peripheral nervous system. There were also studies that involved the eyes, bone, lymph nodes, nasal structures, and tests [3]. The Ridley-Jopling classification uses histopathological and clinical features and the bacteriological index as well. It classifies leprosy in tuberculoid (TT), borderline tuberculoid (BT), borderline (BB), borderline-lepromatous (BL), and lepromatous (LL) categories. In another classification by WHO, multibacillary leprosy includes the lepromatous (LL), borderline lepromatous (BL), and borderline (BB) forms, and paucibacillary leprosy encompasses the tuberculoid (TT) and borderline tuberculoid (BT) forms [4]. During the past years, few studies aiming to explore the host susceptibility to leprosy have been explored and published. The results of epidemiological studies (twin studies and complex segregation analyses) [5–7] and genome-wide analysis (linkage and association) [8] in leprosy infection clearly indicated the involvement of host genetic factors to determine susceptibility to develop the disease and the individual’s response to the infection with M. leprae. Host genetic factors may therefore largely determine which exposed individuals develop disease. Several studies indicate that leprosy pathogenesis is a three-step process: (i) group of genes controls susceptibility to infection, (ii) different genes control the clinical manifestation of disease, and (iii) genes influencing the development, in a proportion of affected individuals, of leprosy reversal reaction type 1 (RR1). A recent human genome-wide linkage or association studies have observed a significant association with or linked to leprosy in various genes C3, C4B, CCDC122, CCL3, CCL5, CFB, COL3A1, CR1, CTLA4, DEFB1, FCN2, HLA-DQA1, HLA-DQB1, HLA-DRB1, HSPA1A, ICAM1, IFNG, IL10, IL12B, IL12B1, IL12RB1, LACC1, LAMA2, LRRK2, LTA, LTA4H, MBL2, MICA, MICB, MRCI, NOD2, P450, PARK2, PGL1/SDHD, RIPK2, SLC11A1, TAP1, TIRAP, TLR1, TLR2, TLR4, TNF, and TNFSF15, [5–9–11]. Recent advances in high-throughput genotyping, next generation sequencing, and very high-density microarrays have generated a tremendous amount of human genetic variation data associated with human traits and diseases especially Single Nucleotide Polymorphisms (SNPs), Copy Number Variations (CNVs), and Insertions and deletions (Indels). There were quite a lot of studies that have reported the direct

⁎ Corresponding author at: VIT University, Vellore 632014, Tamil Nadu, India.
E-mail addresses: georgecp77@yahoo.co.in, georgepriyadoss@vit.ac.in (C. George Priya Doss).

© 2012 Elsevier Inc. All rights reserved.
involvement of SNPs in host genetics susceptibility to leprosy. SNPs provide an opportunity to understand the association between genotype and phenotype in a much broader manner. These mounting studies on SNPs insist their role in better understanding the resultant phenotypic variations in individuals with an endeavor towards new drug design and development. Mapping of SNPs to protein domains and analyzing at the structural level will reveal the full extent to which they can alter the interactivity of proteins. To understand the key structural changes induced by SNPs, we have included three dimensional (3D) structures of the corresponding proteins in this database. Although human associated studies related to leprosy have received great attention from experimental researchers; there is no single comprehensive database to address human leprosy associated gene information. The amount of genetic data generating will also increase drastically in the coming years. To address this issue the need for a comprehensive and well organized collection of genetic data from multiple published studies is urgently needed to provide information in one cluster. Currently there are few databases available related to leprosy such as The Leprosy VNTR Database [12], the Leprosy—Mycobacterium Leprae Database (DDTRP) [14] stating the information related to M. leprae and its drug resistance. In this present study, we describe a new web based interface called LSHGD which provides detailed information for researchers who are interested in analyzing the involvement and association of human genome in leprosy. Our main urge is to provide a convenient information platform for molecular epidemiologists and clinicians working on leprosy. The information about the 45 leprosy susceptible human genes [15–56] is shown in Table 1. A primary function of the database is to provide a way of identifying associated SNPs that are likely to have an impact on molecular function. As the first database, freely accessible LSHGD aims to provide comprehensive sets of genetic information both at the sequence and structural levels for extending functional analysis.

2. Data collection

The main goal of our presented work is to develop an interface that provides a direct link between leprosy and human associated genes. The gene information regarding human leprosy susceptible genes [15–56] are provided in Table 1. The data contents can scale up in two ways. (i) The literature information on the impact of human genes and its association with leprosy were compiled from published literatures according to Pub Med (http://www.ncbi.nlm.nih.gov/PubMed/), OMIM (http://www.ncbi.nlm.nih.gov/omim/), and UniProtKB (http://www.uniprot.org/) [57]. OMIM contains textual information, pictures, and reference information. It also contains copious links to NCBI’s Entrez database of MEDLINE articles and sequence information. UniProt was used to validate the involvement of gene in mutation; where this gene is implicated: diseases, a selected bibliography with hyperlinks to MEDLINE abstracts. (ii) The genomic information types include genes, chromosomal regions, normal variants (SNPs), associated proteins and their pathways, etc. The SNP information of associated genes was retrieved from the NCBI dbSNP (http://www.ncbi.nlm.nih.gov/snp/). The annotated sequence is required to determine the functional class (nSNP, cds-synon, UTR, and intronic) of the SNP. Nucleotide coordinates (numerical position of a nucleotide in a gene sequence) and genotypic variation information were also obtained from this sequence. We retrieved related information from Public databases and gene annotation includes NCBI [58], OMIM [57], UniProt [59], Ensembl [60], PDB [61], HGNC [62] and KEGG [63].

2.1. Database design

We provide an intuitive, well organized and user friendly web interface that allows users to explore the detailed information of leprosy susceptible human genes. It includes gene name, gene ID, chromosome number, chromosomal position, chromosomal orientation, gene function, taxonomic identifier, RefSeq_protein, RefSeq_mRNA, OMIM ID, HGNC ID, Pub Med ID, genotypic variation, residue change, function class, SNP reference, associated protein, associated PDB ID, UniProt ID, and associated pathways. Majorly, sequence information was incorporated from NCBI (DNA), protein related information from UniProt and protein structures from PDB. Clicking links leads to more detailed information.

2.2. LSHGD organization and system design

Upon accessing LSHGD (Link: http://www.vit.ac.in/leprosy/leprosy.htm) a click on the first interactive button named “Home” introduces the user about the database. A click on the second interactive button...
named “Gene Database” allows the user to specifically choose the type of information from the check box menu that includes gene information, chromosomal information, associated protein information, and natural variants information from the database. A click on the third interactive button named “Classification of Gene”, allows the user to visualize the classification of human genes susceptible to leprosy based on associated protein mechanisms. A click on the last interactive button named “Associated protein study” allows the user to find several other protein associated studies in different stages of leprosy. A click on the last interactive button named “Link to Other Database” allows user to link other related databases Leprosy—Mycobacterium Leprae Database [13], and Database of Drug Targets for Resistant Pathogens (DDTRP). The following interfaces are displayed in Fig. 1. LSHGD consists of three major software components: an Apache HTTP server, a MySQL database and an Aspx script. The back end data analysis programs were written in asp script using the Microsoft Visual Studio 2008 tool.

### 2.3. Database implementation

A maximum of 45 genes are displayed in the interface. In gene information the interface provides related information such as Gene Symbol, Gene Name, Gene Function, Entrez Gene ID, RefSeq_Protein, RefSeq_mRNA, OMIM_ID, PubMed_ID, HGNC_ID, Primary Reference related to leprosy susceptibility, population studied, pathway Information and other cross references. Pathway information for each gene allows the user to identify disease related and associated pathways. Ethnicity will provide information about the predisposition of leprosy genes in different populations. Similarly, Gene_Name, Gene_ID, Chromosome_Number, Chromosomal_Position and Orientation in chromosomal information have been provided. These information will definitely provide the user the information about susceptibility of genes in a specific chromosomal region. Notably, gene information and chromosome information will serve as genomic platform for database users. To explore the possible relationships between genes and diseases we have incorporated information about natural variants. However, only a small subset of SNP information has been documented. It is easy to determine a preliminary profile of the relationship between genes and diseases using the information such as Gene Symbol, Entrez Gene ID, Associated SNP, Genotypic Variation, Residue Change, Functional Annotation, Consequence of Transcript and SNP Associated with other Disease. Information such as functional annotation and consequence of transcript encoded have been provided for the significant SNPs by GWAS and pathway information for each gene in the natural variants and gene information interface of the revised database. Similarly, Gene_Name, Associated_protein, Chromosomal_position, UniProt_ID, Associated_PDB_ID and Associated_Pathway in protein information exist. From this natural variants and protein information result table, the user can interlink SNP associated information at the structural level. These interfaces created in this database will satisfy the demands of the scientific community by providing in depth knowledge about leprosy association with human variants.

### 2.4. Designed instruction for users

The freely accessible LSHGD available at http://www.vit.ac.in/leprosy/leprosy.htm aims to provide a comprehensive set of genetic information both at the sequence and structural levels for extending functional analysis. The database is composed of five interactive buttons in the main display (Fig. 2) and explained their functions by considering TLR1 as example.

(i) **Main Page:** the database is composed of five interactive buttons in the main display namely Home, Gene Database, Classification of genes, Associated protein study and Link to other databases. Home is the main page of the database. By clicking, the user can access the summarized introduction about the database (Fig. 2A).

(ii) **Gene Database:** This represents the main characteristics of 45 leprosy susceptible human genes. Search for genes in LSHGD by using the drop down menu provided in the interface. From this page the user has the option of retrieving information of their choice such as Gene Information (Fig. 2B), Chromosomal Information (Fig. 2C), Natural Variants Information.
(Fig. 2D), and Protein Information (Fig. 2E) by using four radio buttons. These information will provide in depth knowledge about the gene function, population studied, SNP information, annotation, consequence of transcript, primary reference associated with the disease and other cross references.

(iii) Classification of genes: This page displays the classification of 45 genes namely Innate Immune Receptors Genes (6), Cytokines Genes (1), Transport Genes (1), Tissue-Specific Markers Genes (1), Innate Immune Effectors Molecules and Serum Proteins Genes (3), and Other Candidate Genes (31).

(iv) Associated protein studies: This page (Fig. 2F) represents other associated proteins in different stages of leprosy namely LL, BL, BB, TT, BT, Reversal reactions (RRs) and Erythema nodosum leprosum (ENL).

(v) Link to other databases: This page displays two interlinked database links which will be helpful in understanding multi-drug resistance in leprosy.

3. Future improvement

As the amount of genomic information generating will be increasing drastically in the coming years, we will continuously collect the latest gene and SNP related data disease data sets to keep up-to-date. In the future we would like to improvise our database in three main ways. One is to further provide link gene symbol to their corresponding accession numbers in NCBI (NP for protein and NM for mRNA). Next is to provide link to PDB ID information. Lastly, the incorporation of in silico tool prediction results into database will be able to classify SNPs pathogenic from neutral ones. We assure this information will definitely will make a way for genotype-phenotype researches as well as pharmacogenetics studies.

Conflict of interest

The authors declare that we don’t have a conflict of interest.

Acknowledgment

The authors thank the management of VIT University for providing the facilities to carry out this work.

References
