The Sri Lankan Genome Variation Database

Pubudu S. Samarakoon BSc, MSc
Tutor, Biomedical Informatics Course, Postgraduate Institute of Medicine, University of Colombo, Sri Lanka. Current address: Research Fellow, Department of Medical Genetics, University of Oslo, Norway.
E-mail address: saneth.samarakoon@gmail.com

Prof. Rohan W. Jayasekara MBBS, PhD, C.Biol, MSB (Lond)
Chair and Senior Professor of Anatomy; Director and Medical Geneticist, Human Genetics Unit, Faculty of Medicine, University of Colombo, Sri Lanka
E-mail address: rohanwj@hotmail.com

Prof. Vajira H. W. Dissanayake MBBS, PhD
Professor, Department of Anatomy; Medical Geneticist, Human Genetics Unit, Faculty of Medicine, University of Colombo, Sri Lanka
E-mail address: vajirahwd@hotmail.com

Sri Lanka Journal of Bio-Medical Informatics 2011;2(1):9-20
DOI: http://dx.doi.org/10.4038/sljbmi.v2i1.2861

Abstract

The Sri Lankan Genome Variation Database (SLGVD) is a database of single nucleotide polymorphisms found in Sinhalese, Sri Lankan Tamils and Moors - the three major ethnic groups in Sri Lanka. Studies of variations in genes among different groups of individuals in the Sri Lankan population have grown steadily during the past few years. These studies generate large amounts of genetic data that is important to study the occurrences of diseases that differ across ethnic groups. There is therefore a need for a central repository of this data. The SLGVD was created to fulfill this void. The SLGVD offers web based access to genetic variation information of Sri Lankan people. It would also be an important informatics tool for both research and clinical purposes. The database was designed conforming to guidelines issued by the Human Genome Variation Society (HGVS). In addition to variation data each variation is linked with the relevant entries of Online Mendelian Inheritance in Man (OMIM), dbSNP and GenBank databases at the National Centre of Biotechnology Information (NCBI), USA. Genotype and allele frequencies of each variation in different ethnic groups are represented in numerical and graphical format. The SLGVD can be freely accessed online at http://hgucolombo.org/SLGVD.aspx.

Keywords - Sri Lankan Population; Variation database; Sinhalese; Tamils; Moors

Introduction

Organising large datasets is becoming important with massive amount of data coming from increasing number of genome projects. Databases are an important tool for storing, organising and distributing biological data. Biological information that is extracted from such data can be distributed in many different forms and stored in many different databases. Such data and databases are unique because they are resources for both genomics and genetics. Genomics is concerned with the structure of genomes, while genetics is concerned with inheritance[(1)].

Genome databases can provide different types of views for complete genomes. The nucleotide sequences of the complete set of chromosomes of an organism can be annotated and presented in different forms such as cytogenetic maps, sequence maps and integrated genetic and physical maps.

Variation databases are useful for genetic inheritance and association studies. For example, the Human Gene Mutation Database (HGMD) is a collection of data of germ-line mutations

causing human inherited diseases. The database catalogs single base pair substitutions, deletions, duplications, insertions as well as other complex rearrangements implicated in disease [(2)].

Developments in genotyping technologies and bio-banking projects have enabled researchers to perform association studies more accurately and effectively so that the genetic basis of complex disorders can be explored rapidly. Information coming from these studies are organised and achieved in genetic association databases. The Genetics association database (GAD) is such a database. It allows the user to rapidly identify medically relevant polymorphisms from among a large volume of polymorphism and mutational data[(3)]. Human Genome Variation database of Genotype to Phenotype information (HGVbaseG2P) is another genetics association database that provides an easy way to access all the available association study data relevant to genes, genome regions, variations or diseases of interest[(4)].

Recent population genetic research has suggested that "race" represents a useful proxy for genetic studies in clinical practice and medical interventions[(5)]. Sample collection based on ethnicity is central to these studies. Individuals from "populations" (e.g. "Sinhalese", "Sri Lankan Tamils", "Moors" in the Sri Lankan population) are defined by cultural traits such as shared language, shared religion, or shared origins[(6)]. One problem with this approach may be that it is often not known whether these populations represent any relevant reproductive unit even a few generations back in time. An alternative approach is to sample humans according to geography without regarding cultural traits[(7)]. Understanding how genetic diversity is structured in human species holds great medical relevance[(8)]. For example, if major differences in allele frequencies exist between populations, individuals from different origins may be expected to respond differently to medical treatments[(9)]. Furthermore, understanding population structure is important in association studies where disease genes are identified by association with marker loci[(10)]. Prevalence of genetic variants in a population directly affects gene-disease associations and gene-gene and gene-environment interactions. Genetic epidemiological studies are employed, therefore, to examine the occurrence of genetic variants in a population prior to studying gene-disease associations and gene-gene and gene-environment interactions. Results of genetic epidemiological studies are important to ensure that such studies are conducted in adequately powered sample collections. Human Genome Epidemiology (HuGE) Published Literature database (HuGE Pub Lit) is a knowledge base on the World Wide Web that tracks the growing published literature on human genetic epidemiological studies[(11)].

The importance of cataloging variations, which effect genetic diseases in various ethnic communities, has been highlighted at a HUGO meeting on the Mutation Database Initiative[(12)]. The ultimate goal of this is to create a network of genetic databases that contains a complete record of all known mutations in the human genome. Ethnic and national mutation databases are specialised databases where networking of these with locus specific databases[(13)] will eventually lead to establishing a complete *Homo Sapiens* mutation database[(1)].

The Sri Lankan Genome Variation Database (SLGVD) is a centralised repository of Single Nucleotide polymorphisms (SNPs) found in Sinhalese, Sri Lankan Tamils and Moors, the three major ethnic groups in the Sri Lankan population. The SLGVD was an initiative of the Human Genetics Unit (HGU) of the Faculty of Medicine, University of Colombo, Sri Lanka. The variation data cataloged in SLGVD were derived from published and unpublished research performed on Sri Lankan populations, living in Sri Lanka and abroad.

Population background

The Sri Lankan population is multi-ethnic and multi-religious. The main ethnic groups are Sinhalese (73.9%), Tamils (17.8%) and Moors (7.4%), with smaller groups such as Malays and Burghers accounting for less than 1%. The Tamils in Sri Lanka are divided into two groups. They are Sri Lankan Tamils (12.6%) and Indian Tamils (5.2%). In terms of religious affiliation, the main groupings are Buddhist (69.3%), Hindus (15.5%), Muslims (7.6%) and Roman Catholics (6.9%). The largest communities within the Sri Lankan population are Sinhala Buddhists (approx. 69% of the population), Tamil Hindus (approx. 15%), and Moor Muslims (7.4%). [see http://www.statistics.gov.lk].

System design and implementation

The SLGVD was designed conforming to guidelines issued by the Human Genome Variation Society (HGVS)[(1)]. Unique identifiers of database objects separate databases from mere lists of information[(1)]. The SLGVD uses unique and stable identifier for each table and constraints in the database. The SLGVD also uses a unique identifier to identify each genetic variation. The Single Nucleotide Polymorphism Identity given in the National Institute of Biotechnology Information (NCBI) dbSNP database (dbSNP ID) is this unique identifier. Using it variation data stored in the database can be accessed independently.

The genomic context of all the variations in a database should be defined by specifying the proper species name from which the data was extracted[(1)]. The SLGVD is a repository of *Homo Sapiens Sapiens* (human) data, so the genomic context of the variations in the SLGVD is well defined.

The description of the variant is central to any variation database[(1)]. Each variation in the SLGVD is described in three levels. They are gene-centered description, variation-centered description and genotype-centered description. These levels are used to provide more clear and informative descriptions of the variations to the end user.

The information provided under the gene-centered description includes the official name and the official symbol of the gene assigned by the HUGO Gene Nomenclature Committee (HGNC), the cytogenetic location, the reference sequence, the Online Mendelian Inheritance in Man (OMIM) index of the gene with the variation, structural information, phylogenetic information, and disease association information. The structural information link provides access to information such as splice patterns, exon sequence and distribution, and gene regulatory sequences. The phylogeny link provides access to homologous sequences and the phylogenetic tree of the gene. The disease association link provides access to a list of all the diseases associated with the gene archived in the Genetic Association Database.

The information provided under the variation-centered description includes the SNP ID of the variation assigned in the NCBI dbSNP database, region where the variation is found in the gene and the allelic change. All the information is gathered from the NCBI dbSNP database. Since most of the Single Nucleotide Polymorphisms in the SLGVD are derived from disease association studies, for each variation stored in the database, in the variation-centered description, information on disease association studies performed using the variation as a marker is presented.

The information provided under the genotype-centered description includes submitter, population, phenotype, chromosome count, genotype frequencies and allele frequencies of the SNP observed in study populations, and references. SLGVD is a repository of variation data from different submitters, so that submitter is cited for each variation. Sources from which variation data was obtained are categorised into different panels based on the phenotype. Information about these phenotypes is presented with the variation data. Number of chromosomes counted from each population is specified under chromosome count. There are three genotypes and two alleles for each SNP. They are specified and their frequencies are provided both numerically and graphically. If the data is extracted from scientific journals or publications, the source is cited under references.

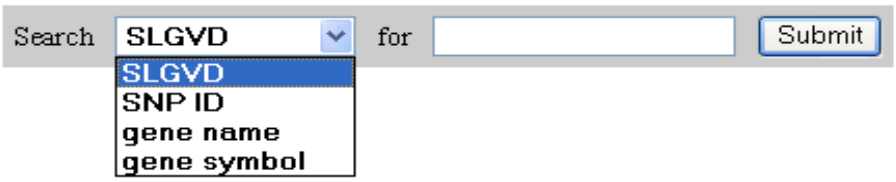
Database access

The SLGVD is accessible from http://hgucolombo.org/SLGVD.aspx. The database interface and data from the database can be viewed using all popular web browsers, e.g. Internet Explorer, Google Chrome, and Firefox. We used Hyper Text Markup Language (HTML) and PHP, Hypertext Preprocessor Language, to design the web interface to ensure that the users do not face any complication during the search and data retrieval processes.

Querying the database

The data in the SLGVD can be retrieved through four search schemes. These search schemes are presented as options in the dropdown menu of the SLGVD search tool bar. They are SLGVD, SNP ID, Gene name and Gene symbol (Figure 1).

Figure 1. SLGVD search tool bar



The SLGVD search option can be used to search the database using an unformatted query. An unformatted query may be a simple text that is part of a gene name or a gene symbol. It may even be an alpha numeric, e.g. a SNP ID. With this option the programme evaluates the query and search for all the SNP IDs, Gene names and Gene symbols in the database that have the search query within it.

The SNP ID option in the search toolbar is used to search the database using the SNP ID of the variation which is given by the NCBI SNP database. This option can be used to search for an exact match for the input query with the SNP IDs listed in the database.

The Gene name and Gene symbol options are used to search the database by the name and the symbol of the gene. These options can be used to search the database for gene names and gene symbols in the database that have search query as a part of the gene name or gene symbol listed in the database.
An important feature of the SLGVD search tool bar is that if a user search the database leaving the search cage blank, the users will be able retrieve a complete list of variations stored in the database.

Data submission

Submitters require a "Submitter ID" prior to submitting SNP data to the SLGVD. To request a "Submitter ID", submitters have to complete the submitter ID request form and send it to submitterid-request@hgucolombo.net. A confirmation e-mail will be sent to the contact e-mail address in the form. Each submitter will be asked to provide the information in Table 1 below before assigning a submitter ID.

Table 1. Submitter information needed for registration

Category	Description
Name	Complete name of the contact person
E-MAIL	E-mail of the contact person
FAX	Fax number of the contact person
TELEPHONE	Telephone number of the contact person
LAB NAME	Laboratory name
INSTITUTION	Name of Institution
ADDRESS	Complete mailing address

Submitters can suggest their own submitter IDs and if they have not been previously assigned in the SLGVD, the database administrators will assign those to submitters. If not, the database administrators will contact the submitter and assign similar IDs to the IDs suggested by submitters. Once a submitter receives a Submitter ID and is registered, the submitter can start submitting data via the online data submission pipeline or via e-mail.

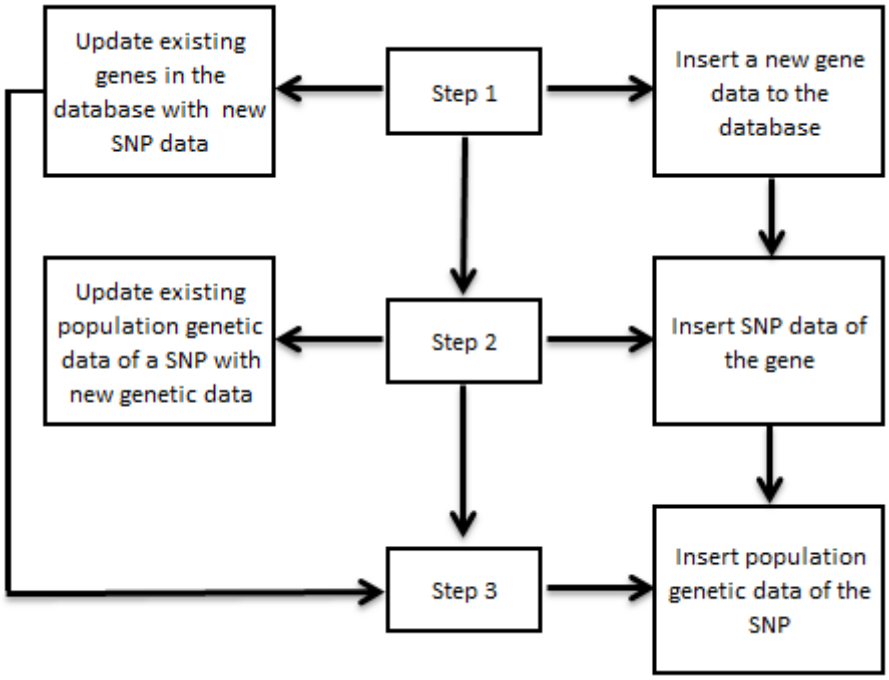
The online data submission pipeline is an interactive tool that displays archived genes, SNPs, genotypes, alleles in SLGVD and allows the user to add new data to the database. This tool is accessible from the data submission link of the SLGVD home page using the password given to the submitter at the time of registration. The password can be changed after the submitter's first login to the pipeline.

The online data submission pipeline of the SLGVD has three steps. Steps one and two, display data and allow user to interact and control the direction of the data submission process. Step one displays a list of genes stored in the database and step two displays a list of SNPs of a specific gene. Using these lists submitters can select the genes or the SNPs that needs to be updated. If the gene or the SNP is not listed, the submitter is given the option to insert a new gene or a new SNP into the database. In step three, the submitter can enter population genetic data to the database. The third step is accessed when a new SNP is inserted into the database via step two or when an existing gene is updated with new SNP data via step one. In each step of the pipeline, a catalog of data stored in the database is displayed. If the data stored during the submission process is incorrect, the user is given the option to delete the entire record set before moving to the next step. The online data submission pipeline is described diagrammatically in Figure 2.

Only the SLGVD curators and developers have access to the main database. Data from all the other submitters will be stored in a different database that stores genetic data to be evaluated. A SNP submission file is used to submit SNP data via e-mail. A template submission file is available at the SLGVD home page and the completed files can be mailed to slgvd-subs@hgucolombo.net. The SLGVD submission file format strictly follows the excel submission template SNPPOPUSE used by the NCBI dbSNP database

(www.ncbi.nlm.nih.gov/SNP) as the standard. Data submission via e-mail is useful in bulk SNP submissions. Data from the SNP submission files is also subjected to the evaluation process.

Figure 2. The online data submission pipeline



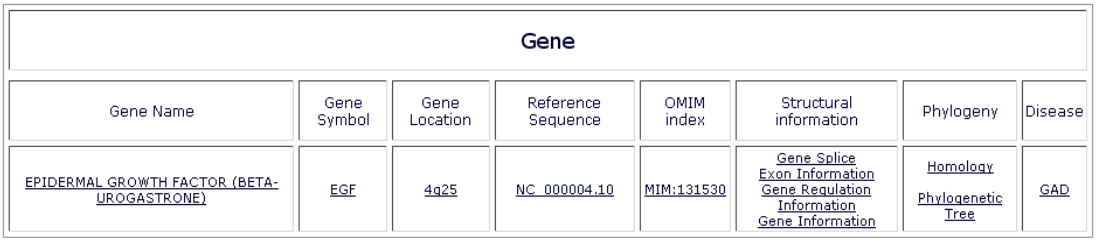
As database curators evaluate submissions, submitters will receive a submission report from the SLGVD with a list of updated database entries or a list of errors encountered during the data evaluation process. Submitters can resubmit the corrected file to slgvd-subs@hgucolombo.net in the latter case. Submitted data will not be publicly available until the committee that is responsible for quality control of submitted data has verified them.

Data display

Presenting search results in the clearest and simplest form to the end user is a key target of the SLGVD designers and developers. This goal is achieved by describing genetic data in three distinct levels as explained in the "system design and implementation" section. Data in each level (gene-centered, variation-centered and genotype-centered) is grouped into three separate tables. These tables are titled "Gene", "SNP" and "Genotype frequency". They are presented in the SLGVD search result web page (Figures 3, 4 and 5).

The gene table presents information from the "Gene-centered description" level of the SLGVD. Each entry of the "Gene" table is linked with relevant entries of the external databases listed in Table 2.

Figure 3. Gene Table



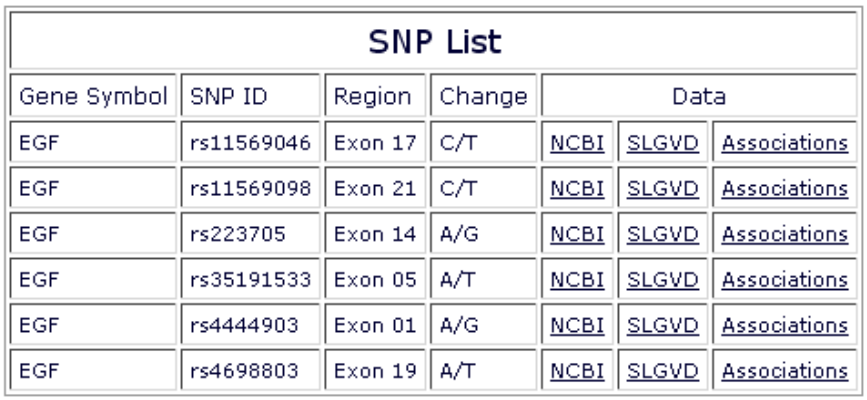
Gene							
Gene Name	Gene Symbol	Gene Location	Reference Sequence	OMIM index	Structural information	Phylogeny	Disease
EPIDERMAL GROWTH FACTOR (BETA-UROGASTRONE)	EGF	4q25	NC_000004.10	MIM:131530	Gene Splice Exon Information Gene Regulation Information Gene Information	Homology Phylogenetic Tree	GAD

Table 2. External databases linked to the Gene Table

Entry	Links to
Gene name	Relevant entry of the National Centre for Biotechnology Information (NCBI) gene database
Gene symbol	HUGO Gene Nomenclature Committee
Gene location	Relevant entry of the NCBI map viewer reference assembly
Reference sequence	Relevant entry of the NCBI Reference Sequences (RefSeq) of annotated reference Genomic assembly : Build 36.3
OMIM index	Relevant entry of the NCBI Online Mendelian Inheritance in Man
Structural Information Gene Splice Exon Gene regulation Gene	 Ensembl Gene Splice view Ensembl Gene Sequence view Ensembl Gene Regulation view Ensembl Gene Summary view
Phylogeny Homology Phylogenetic	 NCBI HomoloGene Database Ensembl Gene tree view
Disease	Genetic Association Database (GAD)

The SNP table presents information from the "variation-centered description" level of the database. In addition to the categories described in the variation centered description of the system design and implementation section, the data links of the SNP List table provides links to NCBI dbSNP data, SLGVD genotype frequency data and HGVBaseG2P data.

Figure 4. SNP Table



SNP List						
Gene Symbol	SNP ID	Region	Change	Data		
EGF	rs11569046	Exon 17	C/T	NCBI	SLGVD	Associations
EGF	rs11569098	Exon 21	C/T	NCBI	SLGVD	Associations
EGF	rs223705	Exon 14	A/G	NCBI	SLGVD	Associations
EGF	rs35191533	Exon 05	A/T	NCBI	SLGVD	Associations
EGF	rs4444903	Exon 01	A/G	NCBI	SLGVD	Associations
EGF	rs4698803	Exon 19	A/T	NCBI	SLGVD	Associations

The genotype frequency table presents genotype and allele frequency data observed in each study population in both text and graphical formats. Frequencies given in the table are represented graphically using colour coded lines. Three genotypes are represented in blue, red and green and the two alleles are represented in blue and red. The length of the coloured line represents the frequency. Submitter column of the table links to a web page that gives information about the submitter. If the SLGVD curators extracted the data from published literature and included it in the database, then the submitter is specified as SLGVD and a link to the abstract in the NCBI PubMed database is given under references. The phenotype column links to a webpage that describes the phenotype of the population and sampling criteria.

Figure 5. Genotype Frequency Table

Genotype frequencies of rs4444903

Submitter	Population	Phenotype	Chromosome Count	Genotype Frequency			Allele Frequency		References
				GG	GA	AA	G	A	
HGU	Sinhalese	Population	160	35	47.5	17.5	58.8	42.2	unpublished
HGU	Sri Lankan Tamil	Population	160	33.7	40	26.3	53.8	46.2	unpublished
HGU	Moor	Population	160	26.25	47.5	26.25	50	50	unpublished
HGU	Sinhalese	Preeclampsia	342	32.8	45.6	21.6	55.6	44.4	V.H.W.Dissanayake ,et.al (2007)

The result page will depend on the option selected from the drop-down menu of the SLGVD search toolbar. If the SLGVD was searched by gene name or gene symbol, then only the "Gene Table" and the "SNP Table" will appear in the result page. It will list all the genes that have the search text as a part of the gene name or gene symbol. If the search was performed leaving the search cage blank, the result tables will list all the genes and SNPs stored in the database. As the "genotype frequency table" is not provided in the SLGVD search result web page, using the SLGVD link in the data column of the SNP table, users can retrieve genotype frequency data. If the database is searched selecting the SNP ID option all three tables will appear in the result page. SNP list table with the list of all the SNPs stored in the database will be displayed if the search is performed selecting the SNP option and leaving the search cage blank. The result of the search using the SLGVD option provides a categorised view. Here, results are grouped into three tables. They are "SNP ID table", "Gene Name table" and "Gene Symbol table". These tables specify SNP ID searched by the user or lists gene names and gene symbols that have the search text as a part of the name or symbol.

Results and discussion

The SLGVD was developed and will be maintained by the Human Genetics Unit, Faculty of Medicine, University of Colombo, Sri Lanka. The database is a repository of genetic data from genetic research on Sri Lankan populations.

At present, data on 34 SNPs from 14 genes have been deposited into the SLGVD. Data in the database is manually examined by a panel of curators who are responsible for the quality of

the database. Almost all the SNPs currently in the database are reported to be associated with diseases. As such they have been the subject of attention in disease association studies. Table 3 provides a summarised description on each SNP archived in the database.

Table 3. Summarised description of SNP's in the database.

Gene Symbol	SNP ID	Region	Change	Disease
AGT	rs4762	Exon 02	T/C	CGEMS Prostate Cancer CGEMS Breast Cancer NINDS Parkinson's Disease FUSION Type 2 Diabetes Ischemic Stroke
AGT	rs699	Exon 02	C/T	CGEMS Prostate Cancer CGEMS Breast Cancer NINDS Parkinson's Disease AREDS Age-related Macular Degeneration (AMD) FUSION Type 2 Diabetes Ischemic Stroke
CYP2C9	rs1057910	Exon 07	C/A	CGEMS Prostate Cancer CGEMS Breast Cancer NINDS Parkinson's Disease FUSION Type 2 Diabetes Ischemic Stroke
CYP2C9	rs1799853	Exon 03	T/C	Preeclampsia
EGF	rs11569046	Exon 17	C/T	Preeclampsia
EGF	rs11569098	Exon 21	C/T	Preeclampsia
EGF	rs223705	Exon 14	A/G	Preeclampsia
EGF	rs35191533	Exon 05	A/T	Preeclampsia
EGF	rs4444903	Exon 01	A/G	Preeclampsia
EGF	rs4698803	Exon 19	A/T	Preeclampsia
F5	rs1800595	Exon 13	G/A	Preeclampsia
F5	rs6025	Exon 08	G/A	Preeclampsia
HBB	rs10768683	Intron 02	C/G	Preeclampsia
HBB	rs713040	Exon 01	T/C	Preeclampsia
HBB	rs7480526	Intron 02	A/C	Preeclampsia
HFE	rs1799945	Exon 02	C/G	Preeclampsia
HFE	rs1800562	Exon 04	G/A	CGEMS Prostate Cancer DGI Type 2 Diabetes CGEMS Breast Cancer NINDS Parkinson's Disease
LTA	rs4986978	Intron 01	A/G	Preeclampsia
MTHFR	rs1801131	Exon 08	C/A	Preeclampsia
MTHFR	rs1801133	Exon 05	T/C	CGEMS Prostate Cancer CGEMS Breast Cancer NINDS Parkinson's Disease AREDS Age-related Macular Degeneration (AMD)

Gene Symbol	SNP ID	Region	Change	Disease
				FUSION Type 2 Diabetes Ischemic Stroke
MTHFR	rs2274976	Exon 08	G/A	CGEMS Prostate Cancer CGEMS Breast Cancer NINDS Parkinson's Disease FUSION Type 2 Diabetes Ischemic Stroke
MTHFR	rs57431061	Exon 08	C/T	Preeclampsia
MTHFR	rs9651118	Exon 08	T/C	DGI Type 2 Diabetes WTCCC Bipolar Disorder WTCCC Crohn's Disease WTCCC Hypertension WTCCC Rheumatoid Arthritis WTCCC Coronary Artery Disease WTCCC Type 1 Diabetes WTCCC Type 2 Diabetes
TGFA	rs1058213	Exon 06	C/T	Preeclampsia
TGFA	rs11466285	Exon 06	T/C	Preeclampsia
TGFA	rs3771523	Exon 06	A/G	Preeclampsia
TNF	rs1800629	5' near gene	A/G	Preeclampsia
VKORC1	rs9923231		C/T	Preeclampsia
SLC11A1	rs2276631	Exon 03	C/T	Diabetes, type 1
SLC11A1	rs3731865	Intron 04	G/C	Diabetes, type 1
SERPINA1	rs17580		A/T	Cystic fibrosis
SERPINA1	rs28929474		A/G	Cystic fibrosis
HBEGF	rs41445351		C/T	Breast cancer (HGVST2) Parkinson's disease (HGVST6) Ischemic stroke (HGVST14) Type 2 Diabetes (HGVST5) Prostate cancer (HGVST1)
HBEGF	rs4150208		C/T	Breast cancer (HGVST2) Parkinson's disease (HGVST6) Ischemic stroke (HGVST14) Type 2 Diabetes (HGVST5) Prostate cancer (HGVST1)

The SLGVD is a much small database compared to other national and ethnic mutation databases, but the genotype and allele frequency data from Sri Lankan populations stored in the SLGVD is not in any other database. The size of the SLGVD is expected to grow rapidly when one considers the recent rapid development of genetic research in Sri Lanka.

The clear and simple data presentation formats of the SLGVD search results make it easy for users to locate genetic data without much hassle. As the SLGVD search result web page provides links to most of the major biological databases, this can be used as an interface to access information in these biological databases. Graphical and textual genetic data representation is a key feature of SLGVD. This format allows the SLGVD users to study

genotype and allele frequencies of different populations more easily and effectively. The SLGVD contains both published and unpublished genotype and allele frequency data. If the data is published, the reference column of the "genotype frequency table" provides links to publications, so that users can directly access the abstract/article for further information. The links enrich the data in the database and offer access to large amount of genetic information that researchers can use in their studies.

With the growth of the SLGVD in the future, it will be a starting point for genetic research in the Sri Lankan population. Researchers and clinicians can use the database as a starting point to access Sri Lankan information on SNPs and genes of their interest. Genotype and allele frequencies in the database will be useful for comparative population genetic studies among ethnic groups in Sri Lanka and elsewhere.

Conflicts of interest

The authors declare that they have no conflicts of interest.

Acknowledgement

The SLGVD was funded by funds from the Human Genetics Unit Development Fund, Faculty of Medicine, University of Colombo, Sri Lanka.

References

1. Scriver CR, Nowacki PM, Lehvaslaiho H. Guidelines and recommendations for content, structure, and deployment of mutation databases. *Hum Mutat* 1999; **13**:344-50. PMID: 10612816

2. Krawczak M, Cooper DN. The human gene mutation database. *Trends Genet* 1997; **13**:121-2. PMID: 9399854

3. Becker KG, Barnes KC, Bright TJ, Wang SA. The genetic association database. *Nat Genet* 2004; **36**:431-2. PMID: 15118671

4. Thorisson GA, Lancaster O, Free RC, Hastings RK, Sarmah P, *et al.* HGVbaseG2P: a central genetic association database. *Nucleic Acids Res* 2008; D797-802 http://dx.doi.org/10.1093/nar/gkn748

5. Risch N, Burchard E, Ziv E, Tang H. Categorization of humans in biomedical research: genes, race and disease. *Genome Biol* 2002; **3**:comment2007.1–2007.12, http://dx.doi.org/10.1186/gb-2002-3-7-comment2007

6. Bowcock A, Cavalli-Sforza L. The study of variation in the human genome. *Genomics* 1991; **11**:491-8. PMID: 1722770

7. King MC, Motulsky AG. Human genetics. Mapping human history. *Science* 2002; **298**:2342-3. PMID: 12493903

8. Serre D, Paabo S. Evidence for gradients of human genetic diversity within and among continents. *Genome Res* 2004; **14**:1679-85. http://dx.doi.org/10.1101/gr.2529604

9. Wilson JF, Weale ME, Smith AC, Gratrix F, Fletcher B, *et al*. Population genetic structure of variable drug response. *Nat Genet* 2001; **29**:265-9. http://dx.doi.org/10.1038/ng761

10. Lander ES, Schork NJ. Genetic dissection of complex traits. *Science* 1994; **265**:2037-48. http://dx.doi.org/10.1126/science.8091226

11. Lin BK, Clyne M, Walsh M, Gomez O, Yu W, *et al.* Tracking the epidemiology of human genes in the literature: the HuGE Published Literature database. *Am J Epidemiol* 2006; **164**:1-4. http://dx.doi.org/10.1093/aje/kwj175

12. Cotton RG. Progress of the HUGO mutation database initiative: a brief introduction to the human mutation MDI special issue. *Hum Mutat* 2000; **15**:4-6. http://dx.doi.org/10.1002/(SICI)1098-1004(200001)15:1<4

13. Claustres M, Horaitis O, Vanevski M, Cotton RG. Time for a unified system of mutation description and reporting: a review of locus-specific mutation databases. *Genome Res* 2002; **12**:680-8. http://dx.doi.org/10.1101/gr.217702