Original Research Article

International Journal of Bioassays
ISSN: 2278-778X
www.ijbio.com
OPEN ACCESS

DSD: A DATABASE OF SNP ASSOCIATED DISEASES

Sameer Chaudhary*, Sapana Mehendale, Shweta Metkar, Tanvee Pardeshi, Akhil Jobby, Vaibhav Kandale, Shama Mujawar and Apurva Patange
RASA Life Science Informatics, 3rd Floor, Dhanashree Apartment, Opposite Chittaranjan Vatika, Model Colony, Shivaji Nagar, Pune-411016, India.
www.rasalsi.com

Received for publication: July 04, 2014; **Revised**: July 21, 2014; **Accepted**: August 11, 2014

Abstract: Single nucleotide polymorphisms, frequently called SNPs, are the most common type of genetic variation among people. They occur normally throughout a person's DNA and studies show the serious association between SNPs and disease. Substantial data regarding SNPs is available but specific and detailed information regarding SNPs and genes related to particular diseases is either unavailable or has to be gathered from several databases. Unlike other existing SNP databases which fail to produce SNP-disease association or doesn't show related PubMed references or doesn't give details regarding the SNPs etc., DSD is a manually curated database containing information regarding SNPs known to cause many diseases. It is the one stop reference for information related to SNPs where disease can be browsed from a list. Each of the record contains details of SNPs, its related gene name, gene id, gene symbol, mRNA allele and protein residue change, UniProt ID, reference Pubmed articles and NCBI assay ID. This database also contains Gene view, variation view and SNP view to provide user a single platform to retrieve the information related to an SNP associated with that disease. DSD also gives population details. It will be a unique resource to study SNPs related to many diseases and help the scientific community to carry out the fundamental research. In the current release, 371 unique SNP entries are included in which data regarding thirty four types of cancer and five polygenic diseases. The future release of DSD will feature many more diseases such as genetically predisposed, polygenic and autoimmune. The database can be accessed at www.rasadbsnp.com

Key Words: SNPs, Disease, Manually curated database, SNP specific database

INTRODUCTION

Single nucleotide polymorphisms (SNPs) are the most common type of genetic variations in humans, accounting for approximately 90% of sequence differences at an overall frequency of about one per 1000 bases. New techniques for identification of the SNPs in the human population are resulting in an exponential expansion in finding the known SNPs. The NIHSNP database (http://www.NCBI.nlm.nih.gov/SNP) contains approximately 2.8 million cases [2]. It is predicted that knowing of an individual's SNP genotype will provide a basis for assessing susceptibility to disease and the optimal choice of therapies. However, a major challenge has been to understand how and when the variants cause disease. SNPs in coding regions (cSNPs) and regulatory regions are most likely to affect gene function. The studies on SNPs show that each gene contains about four cSNPs. Of them, half cause missense mutation in the respective proteins whereas the other half is silent [6] [4].

The Human Gene Database states that missense mutations are responsible for almost half of DNA mutations which are known to cause genetic disease. These mutations are single causative factors of rare monogenic inherited disorders. It is believed that frequent missense mutations arising from cSNPs are associated with common polygenic diseases such as heart disease and hypertension [2].

The understanding of the complex disease biology and their pathogenesis has advanced, thanks to the insights provided by all the genetic experiments and their genome wide association studies (GWAS). It is really important to understand the clinical significance associated with the diseases rather than studying its pathogenesis and prognosis [6]. Variability in prognosis is observed in most complex diseases and has an impact upon patient well-being to a great extent. In spite of this, prognosis in most diseases remains poorly understood and largely unaddressed by genetic technologies that have provided insights into disease development. Several studies carried out show that diseases caused by genetic and environmental factors leading to SNPs, are the biggest reason for occurrence of diseases such as Cancer, Alzheimer's, Autism etc. and also influence many polygenic diseases. As a result, several scientists via small experiments and research projects using high throughput techniques of microarray expression profiling compared normal cells and diseased cells. This comparison led to enormous data generation and revealed that thousands of genes are expressed very differently. This revelation has given scientists the

***Corresponding Author:**
Dr. Sameer Chaudhary,
RASA Life Science Informatics,
3rd Floor, Dhanashree Apartment,
Opposite Chittaranjan Vatika, Model Colony,
Shivaji Nagar, Pune-411016, India.
www.rasalsi.com

resources to explore molecular mechanism and identify the specific genes and SNPs related to a particular disease for example cancer [3][6][8].

MicroRNAs (miRNA) are recently identified gene regulators which at abnormal levels can implicate in all cancer subtypes studied and also many diseases. MiRNAs tend to bind to the 3' untranslated regions (UTR) of their target genes, regions which are evolutionarily highly conserved. As miRNAs control functioning of many mRNAs, chances of cellular transformation and mutation is high, resulting in single nucleotide polymorphism [7]. Hence the role of miRNA SNP in diseases like cancer is being identified. For example miRNA125a, shown to be altered in breast cancer, has a variant allele at a SNP in the mature miRNA sequence that decreases expression of this gene which causes breast cancer. Recent studies have shown that SNPs disrupting miRNA regulation of genes can affect the disease predisposition. Hence, there is a need to identify SNPs which are associated with cancer subtypes and other diseases [5][9][11].

DSD is a manually curated database which provides information regarding not all but some SNPs responsible for causing diseases. Information regarding the particular SNP, its mRNA sequence and the change happening, the change in function and the residue can be searched in this database. As far as our knowledge, this database is one of its kind which gives information regarding disease specific genes and their SNPs which has a clinical significance, provides population details and PubMed articles establishing the relation. The other SNP databases such as dbSNP (NCBI), miRdSNP, SCAN, mtSNP database, CaSNP, DisGeNet, GeneSNPs, SNPedia, Genetic Association Database ,Japanese SNP Database, Gene Cards, VnD, DACS-DB, SNP control database gives details of SNPs for a particular kind of disease or any SNPs which may or may not have direct involvement in disease [2][7][10]. Like in DACS- DB (Disease associated cytokine SNPs database) (http://www.iupui.edu/~cytosnp), association is established between cytokine related SNPs and diseases and not all diseases are covered. In short this database covers a narrow range as opposed to DSD which provides a wider platform for disease associated SNPs.

Hence DSD can be useful in collecting information regarding various types of cancers, their genes and the SNPs and its information from the same page. In the current release of the database, we have collected information related to thirty four types of cancers and five polygenic diseases which integrates information regarding the 372 SNPs and related reference Pubmed articles.

Data Curation

The data regarding SNPs in the database was extensively searched and manually curated from online databases and literature. Presently DSD consists of 372 SNP entries in 39 diseases. The SNP data was retrieved by giving MeSH search for the disease which successively gave the list of genes involved in the disease. The gene data was annotated from NCBI Gene database. Using the Gene details, related SNPs for the disease were retrieved by literature mining and manual curation from Pubmed articles. The SNP data was then fetched from the NCBI dbSNP database. This process was repeated for each disease which is included in DSD. Data was cleaned in order to remove duplicity and redundancy using Perl scripts. From the data obtained it was clear that many SNPs were involved in the pathogenesis of diseases.

Each entry in the database contains details like Disease, SNP, alleles responsible for the alteration, allele origin, chromosome position, NCBI SNP Assay Id, NCBI Reference Sequence Id, Clinical Significance, Validation Status, Pubmed Id and article name, Population, P Value of the SNP. Gene information like Gene Symbol, NCBI Gene Id, Gene Name, Chromosome number, Gene Position, Function, mRNA Position, mRNA allele Change, Protein position, Protein Residue change, mRNA Accession number, Protein Accession number and Uniprot ID. The information regarding the type of study i.e. "In vitro", "In silico" and "Suspected" (the SNPs suspected to cause that disease) were duly noted in tabular format along with their respective Pubmed IDs. DSD also contains gene view, variation view and SNP view for the respective SNP and gene entry.

Database Implementation

The Disease SNP Database is a java based database. It uses JSP as front end and MySQL as back end, a relational database management system. All data in database is inserted from excel sheet using SQL queries. The public user interface read from the database and there is no intermediate database for links provided in interface. All external links are synchronized with the DSD database entries. It also gives the external link for references along with its paper which will be displayed in reference tab.

As mentioned earlier the current version of DSD contains data about 173 genes, 372 SNPs, 39 Diseases and 607 references. Following two graphs describes the exact distribution of the available data in the database.

Table 1: The summary of data fields covered in each entry of DSD

SNP Details		**mRNA and Protein Details**
Disease name	NCBI Assay ID	mRNA Accession Number
NCBI SNP ID	ChromosomePosition	mRNA Position
Population	Allele Origin	Allele Change
Gene Symbol	Validation Status	Function
Chromosome Number	P Value	Protein Accession Number
Allele		Protein Position
Clinical Significance		Residue Change
Study		Uniprot Id

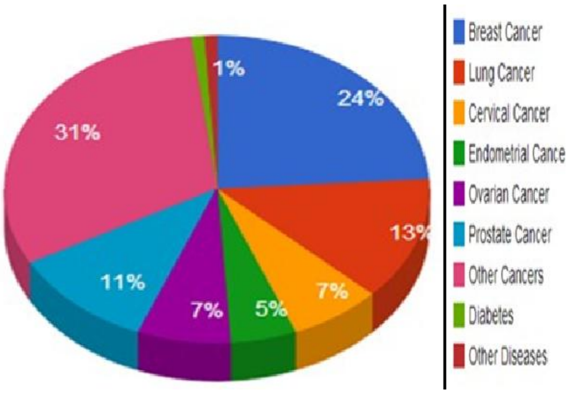


Figure 1: This graphs demonstrates the SNP distribution in the diseases included in the DSD

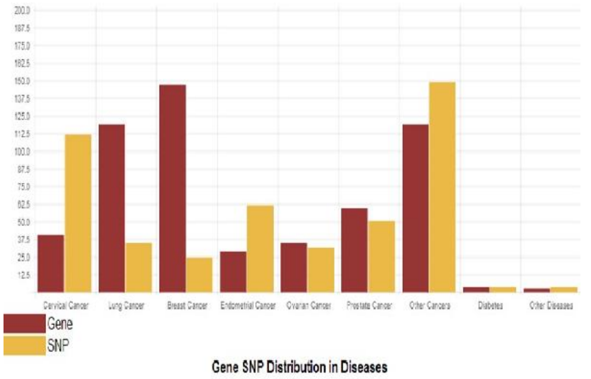


Figure 2: This graph represents the Gene and SNP distribution in diseases included in DSD

Accessing the Data

DSD database gives the users two options to query the database **(i)** Simple search **(ii)** Advanced search.

In simple search the user can search his query by Gene symbol, NCBI SNP ID, DSD ID and Disease name. The result page will display information regarding the unique DSD IDs given to the SNP entries. Also, disease name NCBI SNP ID, gene symbol, Population, P-value, SNP allele and the related reference Pubmed ID are displayed. Clicking the DSD ID link will open further details about the SNP i.e. gene name, gene id, chromosome number and position, SNP ID, related NCBI assay ID and UniProt ID. It will also display information regarding mRNA accession ID, its position, allele change and protein accession ID, its position, residue change. Other information regarding function of the allele change and orientation of SNP to chromosome will also be exhibited depending on the user search query. If the user wishes to know more regarding how a particular gene is associated with disease then, he can click on the reference Pubmed ID shown on the result page. Along with the information displayed related to query, user can also visualize the Ensemble Gene view, NCBI variation and SNP views on the same page.

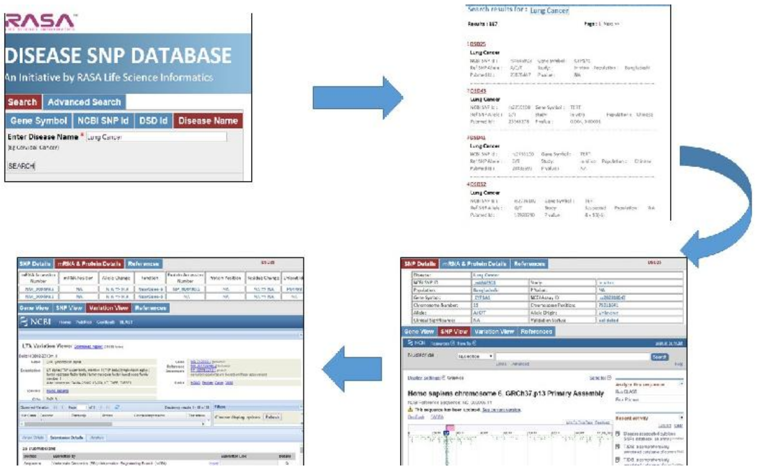


Figure 3: Picture montage of simple search performed in DSD using the disease name "Lung Cancer". This image shows the expected output generated by DSD and how it is represented.

In the advance search, the user can search using very specific criteria. This criterion asks the user to specify the query. The fields include disease name, chromosome number, gene symbol and population. The user will have the choice of including or excluding these four fields by choosing the options "and" and "not" thus customizing the search and the expected result. See figure 4 for search query flowchart.

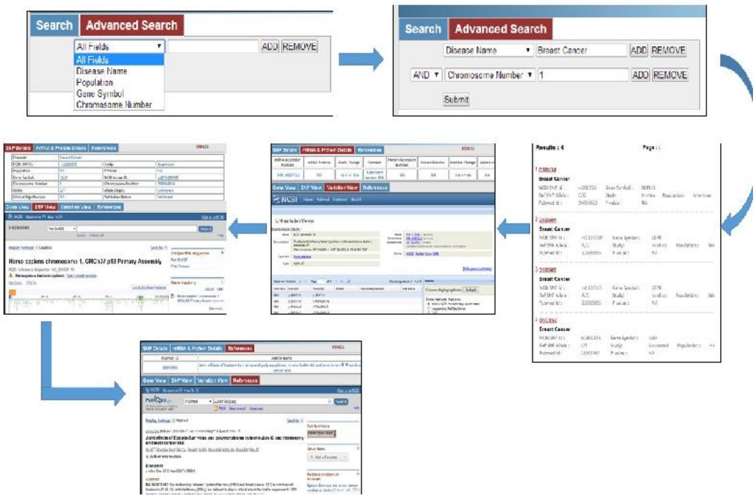
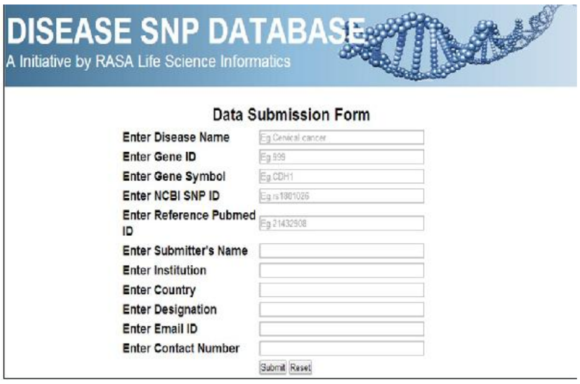


Figure 4: The picture montage of Advance search performed in DSD by using the field's Disease name "Breast Cancer" and Chromosome number "1". This image also displays the expected output and the search flow.

Furthermore, an online submission facility has been provided for the user to add SNP entries which are related to diseases like cancer, diabetes etc. This is possible by filling up an online data submission form whose link can be seen on right hand upper corner on the home screen. The form asks for basic details such as disease name, Gene ID, Gene symbol, NCBI SNP ID and Pubmed ID along with the user's personal information. After the user adds required information with specified fields (the fields with star are mandatory), the database would be updated after validating the uploaded data.



Data Submission Form

Enter Disease Name	Eg Cervical cancer
Enter Gene ID	Eg 999
Enter Gene Symbol	Eg CDH1
Enter NCBI SNP ID	Eg rs1801026
Enter Reference Pubmed ID	Eg 21432908
Enter Submitter's Name	
Enter Institution	
Enter Country	
Enter Designation	
Enter Email ID	
Enter Contact Number	

Figure 4: This is the image of the data submission form available on DSD

DISCUSSION

Although there is a lot of information acquired the SNPs and the mutations causing diseases, which has been documented in a great detail in scientific literature and textbooks and even online resources, but there is no one database which establishes a clear relationship between a SNP and a disease. Large numbers of databases are available online which are cancer-specific and disease specific like Human Lung Cancer Database, Flybase, SNP4Disease, CCDB etc. but they are specific to a particular disease and its genetic association. SNP details are not very elaborate if at all its mentioned [1][3][5][8][10]. DSD comprises information related to diseases which is curated manually after thorough screening of the genes and SNPs available in the scientific literature. This gives DSD an upper hand amongst other databases

The data in the database is represented in a very systematic format. The user has various search criteria to make query searching fast, efficient and easier. The GUI is very user friendly and hence the data retrieval is easy. By creating DSD, our effort was to create a database which will provide the user insight into pathogenesis of many diseases be it cancer, polygenic or genetically predisposed diseases in quick and user-friendly way. The DSD is a SNP related database not a disease database hence the information obtained will focus more on the molecular part and will help the user to get detailed information. DSD creates a big platform for many diseases like several types of cancer, diabetes etc wherein all possible information is available and would contribute largely to the scientific community in carrying out fundamental research. The majority of the diseases in DSD are cancer and its various subtypes (current version). With the help of the information present in DSD, cancer pathogenesis can be studied in a better way.

In conclusion, Disease SNP Database (DSD) is a unique resource for studying and understanding involvement of SNPs in various diseases. We believe that, our database can contribute largely to the field of life science research by providing quick and accurate information all in one place.

Future Developments

The current version of DSD contains 372 unique SNPs associated to many diseases supported by Pubmed records. The current release has data regarding thirty four types of cancer and five polygenic diseases. The next release of DSD will have information regarding Obesity, Inflammatory bowel disease (IBD), Cleft palate, Multiple sclerosis, Cystic fibrosis, Porphyria, Beta Thalassemia, Phenylketonuria, Canavan disease, Neuroblastoma, Alzheimer's disease Asthama, Lupus, Autoimmune Thyroiditis, their SNPs, genes and the general information regarding them. We estimate that in approximately one year we plan to update DSD. Until that time, the current data will be updated on a regular basis. We propose to add new tools to make information retrieval and analysis of the obtained data easier. Our aim is to increase the quality of this database by providing veracious data and sophisticated tools.

REFERENCES

1. Amoreira C, Hindermann W and Grunau C. An improved version of the DNA methylation database (MethDB). Nucleic Acids Res. 2003, 31, 75–77.

2. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Federhen S, Feolo M, Fingerman IM, Geer LY, Helmberg W, Kapustin Y, Landsman D, Lipman DJ, Lu Z, Madden TL, Madej T, Maglott DR, Marchler-Bauer A, Miller V, Mizrachi I, Ostell J, Panchenko A, Phan L, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Shumway M, Sirotkin K, Slotta D, Souvorov A, Starchenko G, Tatusova TA, Wagner L, Wang Y, Wilbur WJ, Yaschenko E, Ye J. Database resources of the National Center for Biotechnology Information. Nucleic Acids Research, 2011, 39 (1), D38-D51.

3. He X, Chang s, Zhang J, Zhao Q, Xiang H, Kusonmano K, Yang L, Sun ZS, Yang H and Wang J. Methy Cancer: the database of human DNA methylation and cancer. Nucleic Acids Research, 2008, 36, D836–D841.

4. Lee JC, Espéli M, Anderson CA, Linterman MA, Pocock JM, Williams NJ, Roberts R, Viatte S, Fu B, Peshu N, Hien TT, Phu NH, Wesley E, Edwards C, Ahmad T, Mansfield JC, Gearry R, Dunstan S, Williams TN, Barton A, Vinuesa CG; UK IBD Genetics Consortium, Parkes M, Lyons PA, Smith KG..Human SNP Links Differential Outcomes in Inflammatory and Infectious Disease to a FOXO3-Regulated Pathway. Cell, 2013, 155(1), 1-13.

5. Jiang Q, Wang Y, Hao Y, Juan L, Teng M, Zhang X, Li M, Wang G and Liu, Y. miR2Disease: a manually curated database for microRNA deregulation in human disease. Nucleic Acids Research, 2009, 37, D98–D104.

6. Chin LJ, Ratner E, Leng S, Zhai R, Nallur S, Babar I, Muller RU, Straka E, Su L, Burki EA, Crowell RE, Patel R, Kulkarni T, Homer R, Zelterman D, Kidd KK, Zhu Y, Christiani DC, Belinsky SA, Slack FJ, Weidhaas JB. Cancer Risk Small Cell Lung– Untranslated Region Increases Non'3 KRAS microRNA Complementary Site in the let-7A SNP in a. Cancer Research, 2008, 68, 8535-8540.

7. Wang L, Xiong Y, Sun Y, Fang Z, Li L, Ji H, Shi T. HLungDB: an integrated database of human lung cancer research. Nucleic Acids Research, 2009, 38, D665-D669.

8. Ongenaert M, Van NL, De MT, Menschaert G, Bekaert S and Van C. W. PubMeth: a cancer methylation database combining text-mining and expert annotation. Nucleic Acids Research, 2008, 36, 842-846.

9. Lu M, Zhang Q, Deng M, Miao J, Guo Y, Gao W, Cui Q. An Analysis of Human MicroRNA and Disease Associations. PLOSOne, 2008, 3 (10), 1-5.

10. Agarwal SM, Raghav D, Singh H and Raghava G.P.S. CCDB: a curated database of genes involved in cervix cancer. Nucleic Acids Research, 2010, 39, 975-979.

11. Wang Z and Moult J SNPs, Protein Structure, and Disease. Human Mutation, 2001, 17, 263-270.

Source of support: Nil
Conflict of interest: None Declared