Gupta et al., Clin Med Biochem 2016, 2:1
DOI: 10.4172/2471-2663.1000111

Clinical & Medical Biochemistry

Research Article Open Access

RareDDB: An Integrated Catalog of Rare Disease Database

Hemant Gupta, Chandan Badapanda*, Arpita Ghosh and Surendra Chikara

Xcelris Labs Ltd., Bioinformatics Division, Ahmedabad, Gujarat, India

Abstract

Background: RareDDB repository for rare disease or orphan disease (http://rareddb.xcelrislabs.com/) is a freely accessible web-based user friendly database which provides detailed information for different types of rare diseases with their associated genes, SNPs along with functional annotations and drug's information.

Method: The RareDDB database was developed using information from various databases such as Orphanet, GHR, OMIM, RDI and dbSNP database using in-house perl script. RareDDB has been implemented using three-tier architecture.

Results: RareDDB contains 2,396 genes that are associated with 6,651 rare diseases and 379 drugs. RareDDB also contains 336,826 curated SNPs related to 1,553 rare diseases. Sequential BLAST homology of 2,396 genes resulted in total 5,900 Gene Ontology terms which includes 1,112 metabolic pathway terms. The orthologs analysis resulted in 849 common orthologs between mouse, yeast, zebra fish, Drosophila and worm using DIOPT server. RareDDB is also linked with databases such as PharmGKB, Drugbank, KEGG and Orphanet to provide comprehensive information about rare disease. In this study, we have also compared rare diseases and their genes between global populations and Indian sub-population which resulted in 521 and 431 common diseases and genes respectively.

Conclusion: RareDDB is a secondary database made by integrating primary data resources such as Orphanet, GHR, OMIM, RDI and dbSNP and also has linked databases for providing detailed information regarding rare diseases. This database comprises dedicated information on rare diseases, orphan drugs, SNPs, genes with their GO terms, Gene location on the chromosomes and orthologs. RareDDB database has a user friendly interface for searching and browsing information related to rare diseases.

Keywords: Rare diseases; Orphanet; dbSNP; RareDDB; Orthologs

Background

Rare diseases are identified by their low prevalence and their heterogeneity. A rare disease is that disease which occurs infrequently or rarely in the general population. In order to be classified as rare disease, each specific disease cannot affect more than a limited number of people out of the whole population. Moreover, rare diseases are not confined to the US or Europe, but effect people all over the globe and consequently represent a true global health issue. There are 5000 to 8000 rare diseases which, taken together, afflict up to 8% of the world's population. Approximately 7,000 rare diseases has been reported in the Orphanet database (www.orpha.net), about 80% of which are of genetic in origin and affect children at a very early age. Rare diseases may occur as a result of bacteria or viral infection, allergies and environmental causes or are degenerative and proliferative [1].

The common problem faced now a days is lack of scientific knowledge and databases equipped with quality information on the disease often results in a delay in diagnosis. As mentioned, due to the broad diversity of disorders and relatively common symptoms which can hide underlying rare diseases, initial misdiagnosis is very common. Also the symptoms differ not only from patient to patient but also from disease to disease and in the same disease, disease recognized as common can be considered as rare for another place, for example malaria is common disease for India but defined as rare in America [1].

In rare diseases limited research information is available. Major problems in rare diseases are due to the unavailability of integrative-functional annotations of rare disease genes at a particular place. Thus, there is growing and urgent need for web based repository of rare disease having information on SNPs, Drugs, Genes, etc. To address this issue, authors have developed a secondary database known as RareDDB (Rare Disease Database) which is accessible at http://rareddb.xcelrislabs.com//. This database is developed using data from primary databases such as Orphanet [2], OMIM [3], Ensemble [4], Drugbank [5], GHR [6], dbSNP [7] and RID [8]. RareDDB provides information of about 6,651 diseases and their associated genes with their functional annotations. RareDDB will be helpful to find rare diseases and their associated genes, along with their drug information which will be beneficial to clinicians and researchers working on rare diseases.

Materials and Methods

Database construction

In the current version, data has been curated from various primary databases such as Orphanet [2], OMIM [3], Ensembl [4], Drugbank [5], GHR [6], dbSNP [7] and Rare diseases India [8]. Also, Orthologous genes and GO terms were analyzed for different rare diseases, which are included in RareDDB. This study has made an attempt to gather the information from literatures, various databases. We have also analyzed the GO (Gene Ontology) terms and orthologous genes related to rare disease, which will contribute to the improvement of the diagnosis and

***Corresponding authors:** Chandan Badapanda, Xcelris Labs Ltd., Old Premchand Nagar Road, Opp. Satyagrah Chhavani, Bodakdev, Ahmedabad, Gujarat, India, Tel:+917966197777; E-mail: chandan.badapanda@xcelrislabs.com; chandan.bioinfo@gmail.com

Received December 28, 2015; **Accepted** January 14, 2016; **Published** January 24, 2016

Citation: Gupta H, Badapanda C, Ghosh A, Chikara S (2016) RareDDB: An Integrated Catalog of Rare Disease Database. Clin Med Biochem 2: 111. doi:10.4172/2471-2663.1000111

treatment of patients with rare diseases. The RareDDB database content has been extracted from Orphanet, GHR, OMIM, RDI and dbSNP database through the help of in-house perl script. RareDDB has been implemented using three-tier architecture. The web based application is created using Apache web server which is connected to the database using MYSQL through an application layer written in Perl-CGI as represented in Figure 1. The entire content and data collection is summarized in Figure 2. RareDDB is also linked with databases such as PharmGKB [9], Drugbank, KEGG [10] and Orphanet for providing comprehensive drug details related to rare diseases.

Blast homology searches and orthologus gene identification

Gene annotation was performed by BLASTX algorithm using non redundant protein database from National Centre for Biotechnology Information (NCBI). The functional annotation was carried based on gene ontology terms and enzyme classification codes (EC), using the BLAST2GO software suite v2.3.1 [11]. Sequences were searched against the NCBI non-redundant (nr) protein database using an E-value cut-off of 10-5. Total of 2,396 genes were classified into either 'Cellular Components' or 'Biological Process' or 'Molecular Function' in order to access the potential role of these proteins in human. DIOPT was used for rapid identification of orthologs [12]. DIOPT integrates existing approaches, facilitating rapid identification of orthologs among human, mouse, zebra fish, *C. elegans* (worm), Drosophila and *S. cerevisiae* (yeast).

Results and Discussion

Data collection and database content

RareDDB is a database which has integrated various databases and also has linked databases for providing detailed information regarding Rare diseases. We integrated 6,651 unique disease related to 2,396 genes from Orphanet database out of which only 3,834 diseases were found to be present in OMIM database. Similarly, we identified 428 diseases reported in Rare Disease India database, out of the total diseases 413 diseases were found to be common with Orphanet database and 15 diseases were found to be unique in Indian sub-population. We also observed out of 428, only 343 diseases were found to be with their OMIM annotation. Also, similar observation was seen at gene level.We obtained a total of 2,396 unique genes from Orphanet and 573 genes related to rare disease present in Indian Sub-population. Among all the genes, 521 rare disease genes were found to be common in both Orphanet and Rare Diseases India database, and only 52 genes were found to be exclusively present in Indian subpopulation. In total 6,651 diseases associated with 2,396 genes are present in RareDDB. A total of 379 drugs are incorporated in the RareDDB and also various databases are linked such as PharmGKB, Drugbank, KEGG and Orphanet.

Functional classification

Gene Ontology Consortium (GO), which provides a structured language that can be applied to the functions of genes and proteins in all organisms. GO mapping resulted in retrieving GO terms for annotated gene using different database. Total of 2,396 rare disease genes were annotated according to all the three GO sub vocabularies (i.e., 'cellular component' (CC); 'biological process' (BP); 'molecular function' (MF)). Genes associated with similar functions were assigned to same GO functional group. The rare disease genes under 'Biological Process' were mainly subdivided in to single-organism process (2265 terms), signalling (1016 terms), rhythmic process(75 terms), response to stimulus (1476 terms), reproductive process (312 terms), reproduction (213 terms), metabolic process (1820 terms), biological adhesion (239 terms) along with other biological process genes as represented in Figure 3A. The rare disease genes under 'Cellular Component' were mainly subdivided in synapse (134 terms), cell (2259 terms), symplast (1), organelle (1889 terms), nucleoid (11 terms), membrane-enclosed lumen (579 terms), membrane (1257 terms) along with other Cellular Component genes, as represented in Figure 3B. From the above analysis it was observed that majority of these rare disease genes were classified in cells followed by different organelles, membranes as well as in nervous system. Similarly, rare disease genes under 'Molecular Function' were mainly subdivided in binding proteins (2013 terms), catalytic activity (990 terms), channel regulatory activity (30 terms), chemo attractant activity (5 terms), electron carrier activity (41 terms), enzyme regulatory activity (137 terms), receptor activity (234 terms) along with other functional genes as represented in Figure 3C. A combined graph of rare disease genes considering all Biological Process, Cellular Components, and Molecular Functions terms are represented in Figure 3D.

SNP (Single Nucleotide Polymorphism)

To integrate the SNPs related to rare disease, we used the database of dbSNP built 137 from NCBI ftp site (ftp://ftp.ncbi.nlm.nih.gov/snp/) and human reference genome hg19 from UCSC browser [13]. Then these SNPs were mapped to the rare disease genes on the basis of the following criteria: a) SNP chromosome number equal to the gene chromosome number, b)SNP chromosome end greater than or equal to the gene chromosome start position, c) SNP chromosome end less than or equal to the gene chromosome end position. The above mentioned criteria's were achieved using "intersects" (BED tools) and in-house Perl script. By using these three criteria, we extracted 336,826 SNPs from dbSNP which were associated with 1,553 rare disease genes and are incorporated in RareDDB.

Orthologous genes identification

The identification of orthologous groups is useful for genome annotation, studies on gene/protein evolution, comparative genomics and the identification of taxonomically restricted sequences [14]. DIOPT was used for rapid identification of orthologs. DIOPT integrates existing approaches, facilitating rapid identification of orthologs among human, mouse, zebra fish, *C. elegans* (worm), Drosophila, and *S. cerevisiae* (yeast). The orthologs analysis was carried out between human rare disease genes along with Zebra fish, Drosophila, Mouse, Worm, and Yeast. We found that 2,379 human rare disease genes were having orthologous relationship with 2334, 1958, 2413, 1863 and 942 genes of Zebra fish, Drosophila, Mouse, Worm, and Yeast respectively. The comparative analysis of genes belonging to all five model organism which were showing orthologous relationship with rare disease genes are mentioned in Figure 4. This resulted in 849 orthologs which were common across all the five species, 848 orthologs were common between Zebra fish, Drosophila, Mouse and Worm, 235 orthologs were common between Zebra fish and Mouse, 52 orthologs were common between Zebra fish, Drosophila, Mouse and Yeast and so on as depicted in Figure 4 and also detailed information available in Supplementary File S1. The identified orthologous genes were associated with other databases such as MGI (Mouse Genome Informatics), Worm Base, Fly Base, SGD (Saccharomyces Genome Database) and ZFIN (The Zebra fish Model Organism Database) to facilitate the information about the function associated with those genes in other model organisms. This analysis provides important genes involved in the evolution of life and

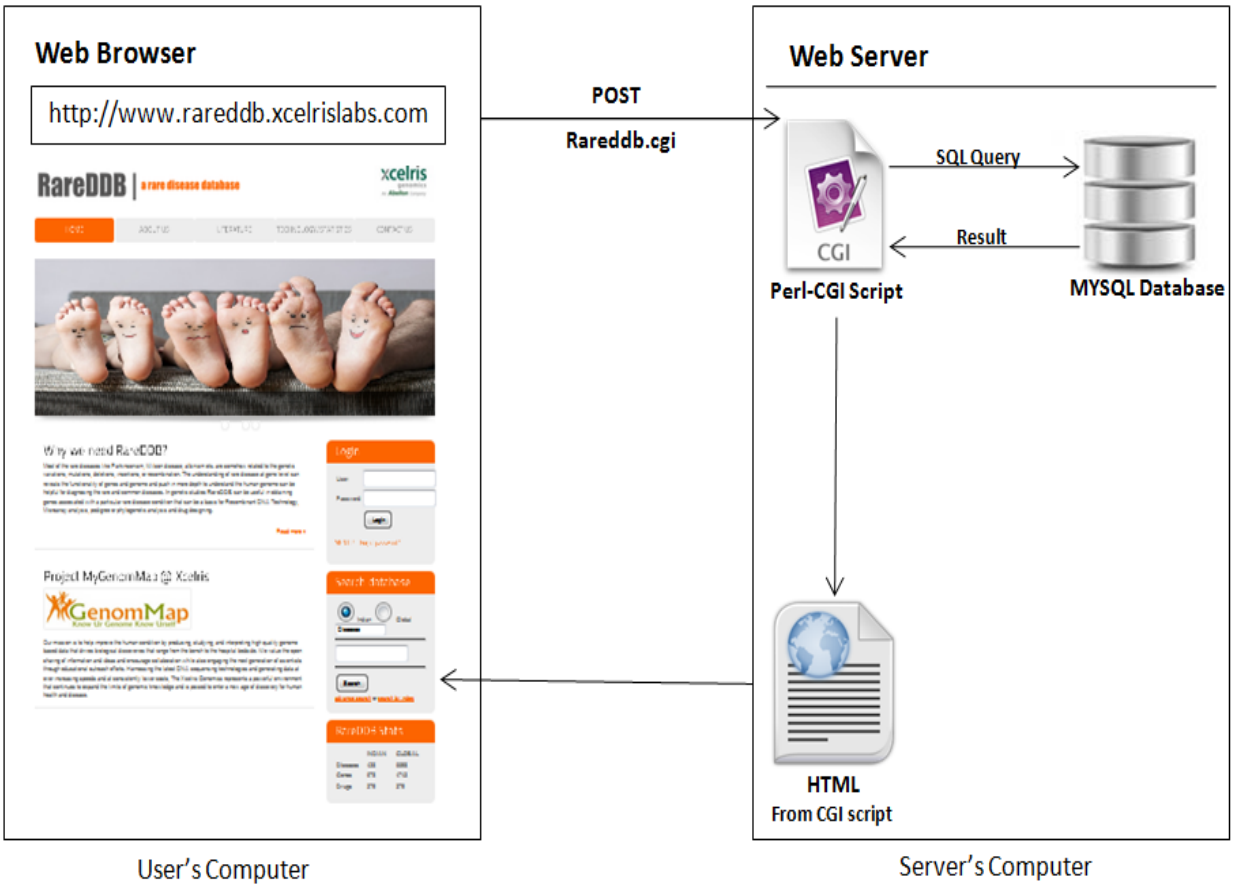

Figure 1: Three-tier architecture of RareDDB (Rare Disease Database).

will help in future to lead the way to a greater synthesis of evolutionary developmental biology and modern medicines.

Utility of RareDDB

RareDDB has web interface for the retrieval of rare disease information including description, clinical significance, rare disease related genes, genes chromosomal location, SNPs and drugs. A search can be initiated by just typing or pasting disease/drug/gene in the search box. User can select disease, gene or drug from the scroll down menu-bar above to the search box. A-Z index option is also available to list out the diseases, genes and drugs alphabetically. An example is provided in Figure 5 how a disease name known as Cystic fibrosis can be searched in details with its disease information, gene, SNP, Orthologs gene with the model organism and drug information.

Steps to use RareDDB:

1. Every page of RareDDB has a search database box at right side.
2. User can go with search box with a drop down menu bar (containing disease, gene and drug) and an A-Z index.
3. User can select one of the three option given in the drop down menu bar (default is diseases).
4. Write disease name/gene name (symbol)/drug name/drug substance in the search box.
5. After clicking on search, a new page will be loaded with list of results related to your query.
6. Click on appropriate one.
7. Another page will be loaded with complete information and hyperlinks of your interest.

User can find all the below mentioned information for rare diseases in RareDDB:

1. Disease description
2. SNPs with Clinical significance
3. Disease related genes
4. Chromosomal location of genes
5. Information regarding orthologous relationship with closely related model organism such as Zebra fish, Mouse, Fly, Worm and Yeast which will help in clinical trails
6. If any drug information is available for the particular disease and Drugbank entry for that drug will be displayed on the page

To get the above listed information, users need to browse several databases. Suppose if user wants to know all the above information for "Zellweger syndrome", user will have to visit Orphanet, to get detailed information about disease including its synonyms, description, health care resources, related genes and drugs. But, for other information user needs to visit OMIM for human genes and genetic phenotypes, GHR

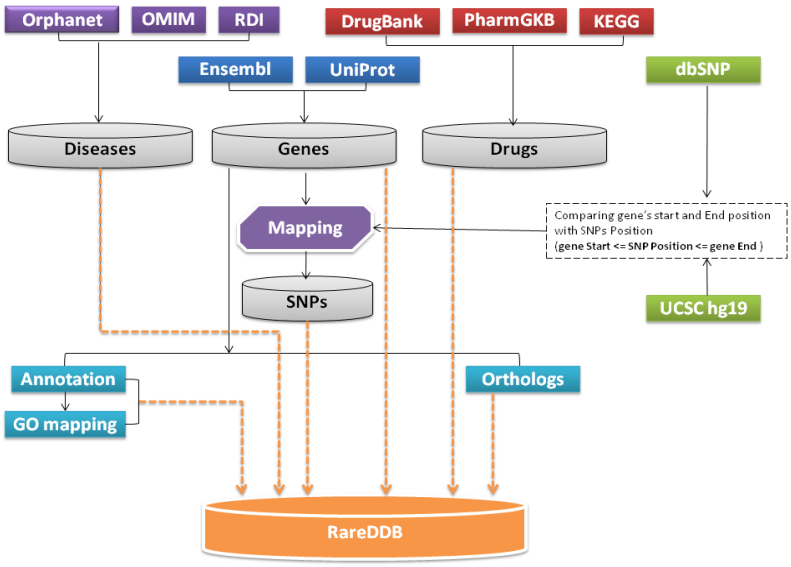

Figure 2: A schematic representation of RareDDB (Rare Disease Database). RareDDB is a secondary database developed by integrating data from various primary databases such as Orphanet, Ensembl, OMIM, Rare diseases India to extract the information on 6651 rare diseases that are linked to 2396 human genes. A total of 379 drug information is extracted from Drug bank, PharmKB and these are linked to KEGG database (Kyoto Encyclopedia of Genes and Genomes). A total of 336,826 SNPs were extracted from dbSNP which are associated with rare disease genes incorporated in RareDDB. Also, integrated functional annotation was performed to identify Gene Ontology terms, KEGG terms, and orthologous gene identification across different species.

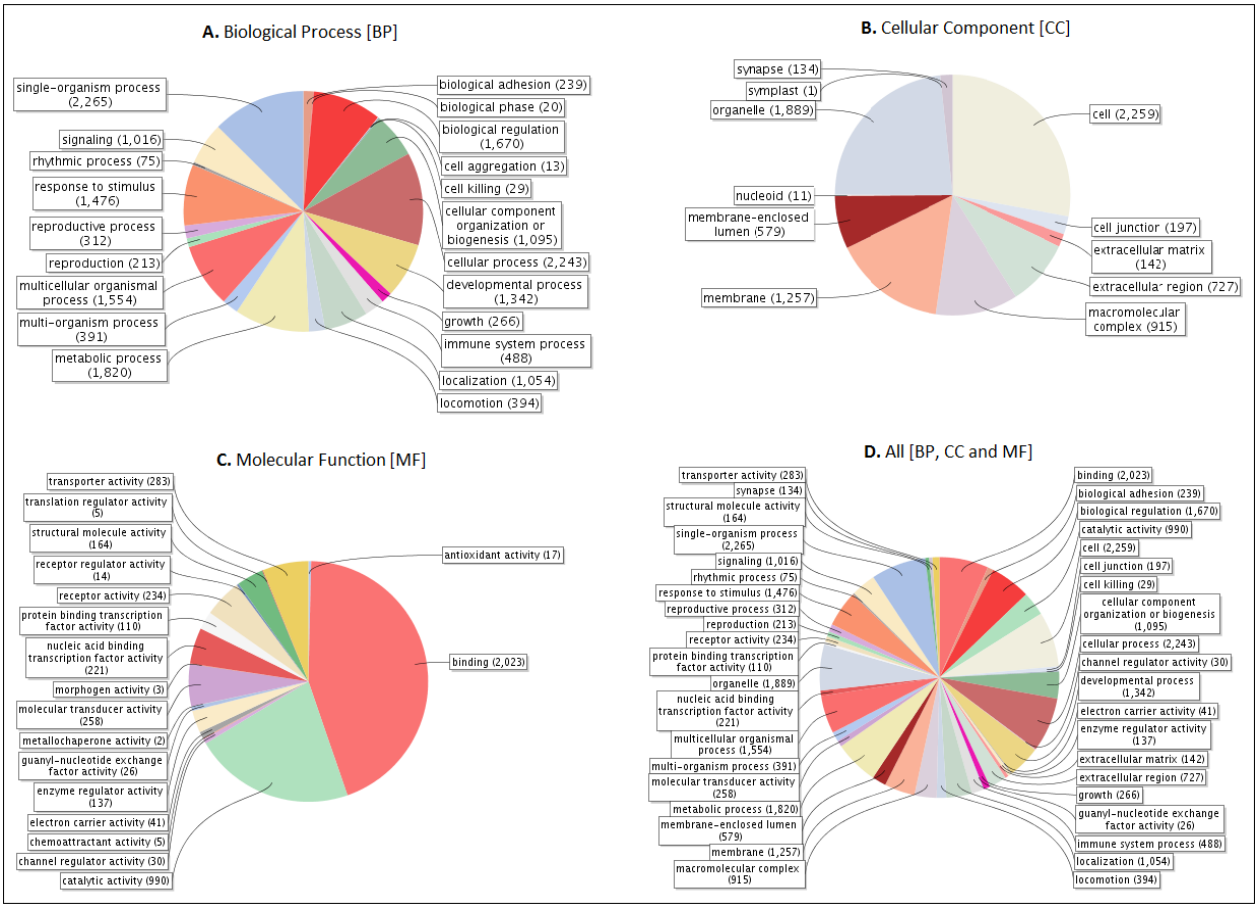

Figure 3: Gene ontology (GO) terms for rare disease genes. (A) Biological Process, (B) Cellular Component (C) Molecular Function, and (D) Combined graph representing Biological Process, Cellular Components, and Molecular Function.

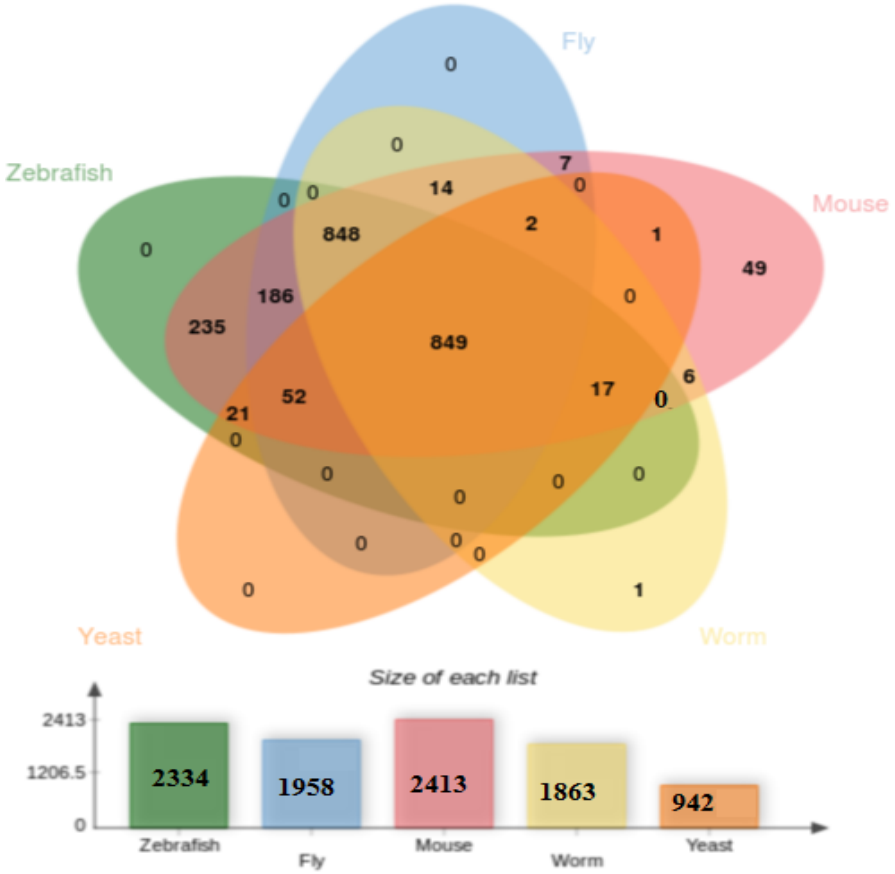

Figure 4: Representation of rare disease orthologs among mouse, zebra fish, *C. elegans*, Drosophila, and *S. cerevisia.*

Database Name	Disease information	Gene information	Chromosome location of the gene	Drug information	SNP details	Orthologus genes
Orphanet	Yes	Yes	Yes	Yes	No	No
GHR	Yes	Yes	Yes	No	No	No
OMIM	Yes	Yes	No	No	No	No
dbSNP	Yes	No	No	No	Yes	No
Drug Bank	No	No	No	Yes	No	No
RareDDB	Yes	Yes	Yes	Yes	Yes	Yes

Table 1: Comparison of RareDDB with other databases.

for disease description, related genes and its chromosomal location, DrugBank for further drug details (in the case of Zellweger syndrome, no drugs information available) and for SNP information related to disease user needs to visit dbSNP database. To sum up, for getting a detailed information of a particular disease user had to visit Orphanet, GHR, OMIM, DrugBank and dbSNP individually.

RareDDB is developed to address the solutions of above problems. It provides all this information at a single place, with external link to the primary data sources as shown in Table 1. RareDDB provides orthologous relationship of disease related genes with closely related model organisms. This knowledge of orthologous relationship of disease related genes among the closely related model organism facilitates further research prospective, for example to identify novel SNPs responsible for the particular disease condition. Hence, RareDDB is a comprehensive rare disease database containing all available information for the rare diseases.

Conclusion

Rare diseases are characterised by huge diversity of disorders and symptoms that affect few people in the population. The symptoms vary not only from disease to disease, but also within the same disease, so it requires detail information for treatment. The same type of disease in different people has very different clinical manifestations. Few of the diseases are very rare in the world but are common in some part of the world such as, Malaria is common in India. The treatment of malaria is also very much known in India. This information can help rest of the world for treatment which has malaria as rare disease. Majority of the rare diseases are identified to be of genetic origins, involving one or several genes or chromosomal abnormalities [1]. But still, till date we don't have a database with all the following information, disease description, SNP, genes, genes location on chromosomes, functional annotation, drug and orthologous genes across five model organisms.

RareDDB is an integrated catalog of rare diseases for global

Citation: Gupta H, Badapanda C, Ghosh A, Chikara S (2016) RareDDB: An Integrated Catalog of Rare Disease Database. Clin Med Biochem Open Access 2: 111. doi:10.4172/2471-2663.1000111

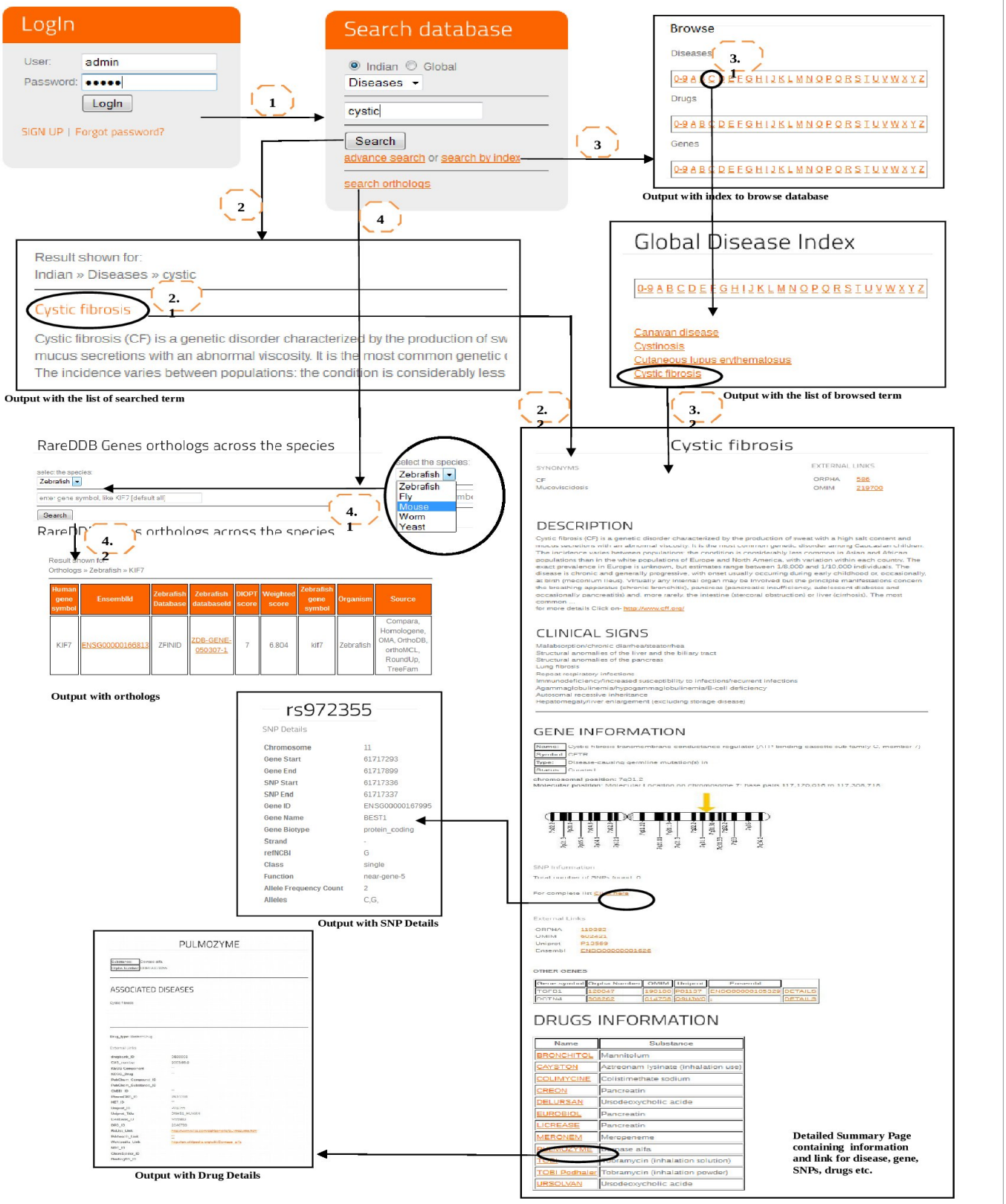

Figure 5: A snapshot of RareDDB database. The figure depicts that the search can be initiated by just typing or pasting disease/drug/gene name in the search box as represented in Step 2. User can also select disease, gene or drug from the scroll down menu-bar in the search box. A-Z index option is also available in RareDDB to list out the diseases, genes and drugs information alphabetically as represented in Step 3.1. After the search is completed a detailed page with disease, drug, gene, SNPs and orthologs information with their external links will be displayed as in Step 2-4.

population. RareDDB is built by gathering information from different databases which includes disease information, gene, chromosomal position, SNP information and drug information etc. Also, some unique features are added by functional annotation and Ortholog gene analysis to provide user all information under a single umbrella. The database may be beneficial to clinicians, researchers or anyone working on rare diseases or associated genes at molecular and clinical level.

The predicted orthologs across different species along with integrated bioinformatics approach should considerably enhance the interpretation of the rare disease. The smaller and genetically tractable models, such as *D. melanogaster*, *C. elegans* or *Danio rerio* (Zebra fish), can each provide critical information about genetic and cellular process underlying certain diseases in a more rapid and cost effective manner than traditional rodent-based or *in vitro* studied [15]. A total of 849 orthologs that are common across all the five species, indicates a conversed sub network of genes relevant to rare diseases. This will help in clinical trials and drug developments which will open ways for modern medicines. In future the orthologous analysis can be used to predict unique genes associated with diseases in the model organisms, which will allow us to understand aspects of more complex diseases. RareDDB provides comprehensive information about 6651 rare diseases that are associated with 2396 genes, 336,826 SNPs, and 379 drugs information.

Acknowledgements

We acknowledge management of Xcelris Labs Ltd., for support and encouragement to carry out this study.

References

1. Deliverska MY (2011) Rare diseases and genetic discrimination. Journal of IMAB - Annual Proceeding (Scientific Papers) 17: 116-119.
2. Weinreich SS, Mangon R, Sikkens JJ, Teeuw ME, Cornel MC (2008) Orphanet: a European database for rare diseases. Ned Tijdschr Geneeskd 152: 518-519.
3. Amberger J, Bocchini CA, Scott AF, Hamosh A (2009) McKusick's Online Mendelian Inheritance in Man (OMIM). Nucleic Acids Res 37: D793-796.
4. Flicek P, Amode MR, Barrell D, Beal K, Brent S, et al. (2012) Ensembl 2012. Nucleic Acids Res 40: D84-90.
5. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, et al. (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res 36: D901-906.
6. Genetics Home Reference (2016) Your guide to understand Genetic Condition.
7. Smigielski EM, Sirotkin K, Ward M, Sherry ST (2000) dbSNP: a database of single nucleotide polymorphisms. Nucleic Acids Res 28: 352-355.
8. Rare disease India (2015) Estimated rare diseases population in South Asian countries.
9. Hernandez-Boussard T, Whirl-Carrillo M, Hebert JM, Gong L, Owen R, et al. (2008) The pharmacogenetics and pharmacogenomics knowledge base: accentuating the knowledge. Nucleic Acids Res 36: D913-918.
10. Kanehisa M, Susumu G (2000) KEGG: Kyoto encyclopedia of genes and genomes. Nucleic acids research 28: 27-30.
11. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, et al. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21: 3674-3676.
12. Hu Y, Flockhart I, Vinayagam A, Bergwitz C, Berger B, et al. (2011) An integrative approach to ortholog prediction for disease-focused and other functional studies. BMC Bioinformatics 12: 357.
13. Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, et al. (2014) The UCSC Genome Browser database: 2014 update. Nucleic Acids Res 42: D764-770.
14. Li L, Stoeckert CJ Jr, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13: 2178-2189.
15. Pandey UB, Nichols CD (2011) Human disease models in Drosophila melanogaster and the role of the fly in therapeutic drug discovery. Pharmacol Rev 63: 411-436.

Citation: Gupta H, Badapanda C, Ghosh A, Chikara S (2016) RareDDB: An Integrated Catalog of Rare Disease Database. Clin Med Biochem Open Access 2: 111. doi:10.4172/2471-2663.1000111