International Journal of Cardiology xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

International Journal of Cardiology

journal homepage: www.elsevier.com/locate/ijcard

HGDB: A web retrieving cardiovascular-associated gene data

Faezeh Noorabad-Ghahroodi [c], Samaneh Abdi [b], Amir Hossein Zand [a], Mohammad Najafi [a,*]

[a] Cellular and Molecular Research Center, Biochemistry Department, Iran University of Medical Sciences, Iran
[b] West Tehran Islamic Azad University, School of Engineering, Iran
[c] Biochemistry Department, School of Medicine, Shiraz University of Medical Sciences, Iran

ARTICLE INFO

Article history:
Received 6 August 2016
Received in revised form 25 November 2016
Accepted 4 January 2017
Available online xxxx

Keywords:
Gene
Cardiovascular disease
HGDB
Database

ABSTRACT

Background: The use of data obtained from high throughput techniques in genetics studies is an essential subject in biology. The system approaches of networking and enriching may improve the data management. Here, we annotated the molecular features for cardiovascular-associated genes and presented the HGDB search-based database (www.hgdb.ir).
Methods: The initial seed data was primarily used from Gene Ontology and was automatically enriched with other molecular features. The data was managed in a SQL popular and open source.
Results: The search tabs on the HGDB homepage were applied for ID/Name Gene, chromosome, cell organelle and all gene options. The search results were presented on the gene text-based and source link-based descriptions.
Conclusions: The HGDB is a friendly website to present gene data in the cardiovascular field.

1. Introduction

Coronary artery disease (CAD) is one of the most important causes of death in the world. Our knowledge of how it affects some people may help to screen the main factors involved in the disease [1]. Although classic risk factors including age, sex, smoking, race and lipidemia are known for at least 30–40% of mortality and morbidity, it is in large part inherited so that more comprehensive evaluations are needed for lifestyle and genetic risk factors [2,3]. In recent years, global techniques have been applied to understand the molecular mechanisms involved in multi-factorial diseases. Genome-wide association studies (GWAS) have suggested the role of several loci in the pathogenesis of coronary heart diseases [4]. However, the complementary studies based on the management, improvement and enrichment of these data can be organized in the field of computational biology.

In global approaches, the disease-associated gene data obtained primarily from OMICS and text-mining tools may be clustered and scored on subcellular features, and other gene and protein characteristics. Here, we introduced a web retrieving cardiovascular-associated gene database enriched automatically to molecular annotations (www.hgdb.ir).

2. Methods

2.1. Data collection

Initial seed data (Fig. 1, A) was obtained from Gene Ontology (Cardiovascular Annotation Gene Initiative). The data were improved through several stages, including i, adding gene data obtained from searching PubMed with the keywords of coronary artery disease (CAD) and gene (Fig. 1, D); ii, adding the interacted genes with the improved initial seed data (Fig. 1, F) prepared from the Protein-Protein Interaction (PPI) servers (IntAct, String, HPRD); and iii, the data flow was followed by annotating the molecular features into the gene list (Fig. 1, F^{*E}). The genes were annotated and enriched with the gene ID, protein ID, gene ontology (Term, GOSlim), binary interaction (Reactome), domain and motif (Interpro, SMART, PRINTS, Pfam and PIRFSF) via transferring data automatically through the Biomart database and Cytoscape software (ver. 2.8).

2.2. Web contents

The data retrieval allows users to access the annotation information including the gene text-based (36 items) and source link-based (16 items) data. The text-based results contain Gene Name, GO Term Definition, MIM, Gene Description, HGNC symbol, UniProt Gene Name, Chromosome, Banding, Inter1 (Interaction), Inter2 (Interaction), Technique (based on the interaction methods), GOSLIM GOA description and UniProt/TrEMBL accession. Moreover, some well ID-defined items were linked to primary databases such as Entrez Gene ID, CCDS ID, EMBL ID, Ensembl Gene ID, Unigene ID, HGNC ID, UniProt, SwissProt ID, GO ontology, PDB ID and UCSC ID as well as to secondary databases consisting of PIRSF SuperFamily ID, Interpro ID, PFAM ID, PRINTS ID, SMART ID and TIGRFam ID.

2.3. Web management

The data was managed in SQL, which is a popular and open source database management system, widely used in bioinformatics and biomedical databases. Furthermore, the HGDB data were organized using visual studio space and also ASP.NET, C.NET and HTML programming languages (www.hgdb.ir).

* Corresponding author at: Cellular and Molecular Research Center, Biochemistry Department, Iran University of Medical Sciences, Tehran, Iran.
E-mail address: nbsmmsbn@iums.ac.ir (M. Najafi).

http://dx.doi.org/10.1016/j.ijcard.2017.01.043
0167-5273/

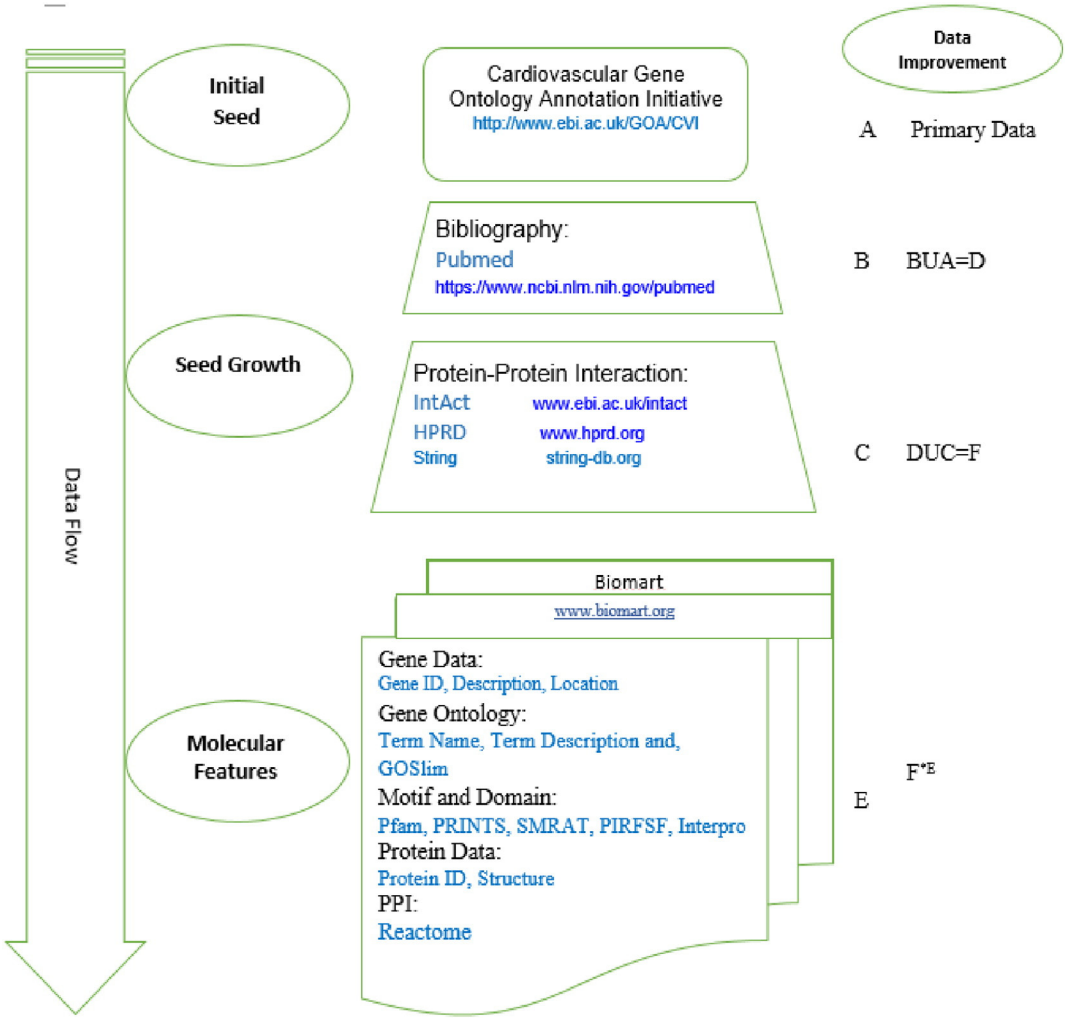

Fig. 1. Data flow diagram. The initial data (A) improved by adding the search results of PubMed and Protein-Protein Interaction databases (D and F). Then, the gene data is annotated and enriched with molecular features (F^{*E}).

2.4. Statistics analysis

Data analysis was statistically performed using a statistical software package (Ver. 18.0, Chicago, SPSS Inc.). The gene distributions between chromosomes and cell organelles were evaluated with the chi-square test.

3. Results

The tabs on the HGDB homepage are considered for searching a gene, chromosome and cellular organelle.

3.1. Search by gene

A user may search a gene using several options. They are defined on the 'Search gene' page and contain Gene Name, HGNC symbol, UniProt/SwissProt ID and Entrez Gene ID (Fig. 2, 1). If a result is found (including gene name, UniProt/SwissProt ID, HGNC symbol, Entrez Gene ID and Link), it is suggested that the gene has been previously reported or related to coronary artery disease (CAD). However, a negative result suggests that the gene is not reported in primary data sources. Clicking on the "More" tab in the page of the gene report opens a new page. It contains two data sets, descriptive and link, presenting not only function, gene location and cellular compartmentation but also their links with primary and secondary databases.

3.2. Search by chromosome

On opening the "Chromosome search" page, there are the tabs for all somatic, sex and mitochondrial (MT) chromosomes (Fig. 2, 2). By clicking on each tab, a gene list located on each chromosome is presented in the page of the gene report. The analysis of the gene list may be used for the gene linkage associations. Similar to the descriptions of the previous section, more data can be considered when clicking the "More" tabs. Table 1 shows the gene distributions on chromosomes in the HGDB database ($P < 0.0001$).

3.3. Search by cell organelle

By clicking on each organelle in the "Cell organelle" page, a gene list related to CAD appears in the page of the gene report. Based on the organelle locations, the genes can present for ribosome, mitochondrion, extracellular, nuclear, vacuolar transport, endosome, plasma membrane, cytoplasm, ER and Golgi body, lysosome, endosome, cytoskeleton and peroxisome organelles. Furthermore, other bioinformatics and descriptive data can be considered when clicking the "More" tabs (Fig. 2, 3).

The "All gene" tab on the HGDB homepage is the gene browser of the database. By clicking on the tab, the total gene list is considered on the page. Furthermore, the gene descriptions are shown by clicking on the "More" tabs.

Please cite this article as: F. Noorabad-Ghahroodi, et al., HGDB: A web retrieving cardiovascular-associated gene data, Int J Cardiol (2017), http://dx.doi.org/10.1016/j.ijcard.2017.01.043