CaGE: Cardiac gene expression knowledgebase

Michael Bober1,2, Kevin Wiehe1,3, Christina Yung1, Tugba Onal Suzek1,3, Matthew Lin1, William Baumgartner Jr1 and Raimond Winslow1,*

1Whitaker Biomedical Engineering Institute, Center for Computational Medicine & Biology, The Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA, 2McKusick-Nathans Institute of Genetic Medicine, Department of Pediatrics, Johns Hopkins University School of Medicine, Baltimore, MD 21287, USA and 3Department of Computer Science, The Johns Hopkins University, Baltimore, MD 21218, USA

Received on September 4, 2001; revised on January 11, 2002; accepted on January 14, 2002

ABSTRACT

Summary: CaGE is a Cardiac Gene Expression knowledgebase we have developed to facilitate the analysis of genes important to human cardiac function. CaGE integrates the functionality of the LocusLink database with data from several human cardiac expression libraries, phenotypic data from OMIM™ and data from large-scale microarray gene expression studies to create a knowledgebase of gene expression in human cardiac tissue. The knowledgebase is fully searchable via the web using several intuitive query interfaces. Results can be displayed in several concise easy to navigate formats.

Availability: CaGE is located at http://www.cage.wbmei.jhu.edu

Contact: mbbober@pol.net; rwinslow@bme.jhu.edu

The acquisition and integration of genomic information from specific tissue types has the potential to yield clinically relevant insights in the medical setting. The extensive data already in existence, and which will continue to be acquired, demand that subject specific reference tools be created. With this in mind we have developed the Cardiac Gene Expression (CaGE) knowledgebase. CaGE is a gene-centric knowledgebase designed to serve as a web-based reference for physicians and scientists interested in genes expressed in human cardiac tissue.

In order for a gene to be designated as being expressed in human cardiac tissue it must be present in NCBI’s LocusLink (http://www.ncbi.nlm.nih.gov/LocusLink/; Pruitt and Maglott, 2001; Pruitt et al., 2000) and have evidence confirming its expression. The LocusLink identifier is used as a stable, central key field for CaGE. There are multiple sources used for the confirming evidence of expression. These include the designation of heart in the ‘express’ field of a Unigene cluster (http://www.ncbi.nlm.nih.gov/UniGene/; Boguski and Schuler, 1995) or the designation of heart in the ‘tissue_type’ field of a GenBank entry (http://www.ncbi.nlm.nih.gov/Genbank/index.html; Benson et al., 2000). Relationships can then be established between any sequence accession number, its assigned Unigene cluster and unique LocusLink. The two human cardiac gene expression libraries, which provide confirming evidence of expression in CaGE, are created by the Toronto Cardiac Gene Unit (http://www.tcgucardiacgene.ca/tcgu/Human/index.html; Hwang et al., 1997) and Body Maps (Okubo et al., 1992; http://bodymap.ims.u-tokyo.ac.jp). These libraries provide lists of RefSeq and GenBank sequence accession numbers that are mapped onto LocusLink via Unigene. The last source of evidence for expression comes from gene expression profiling experiments conducted on normal and dilated cardiomyopathic failing human hearts at the Johns Hopkins University School of Medicine using cDNA microarrays. The summary data from these experiments are included in CaGE http://www.cage.wbmei.jhu.edu/information.htm (Bober et al., 2001).

Given the rapid changes in the data sets discussed above, and the need for quality up-to-the-minute data, CaGE is completely rebuilt with new data daily. The cron function of Unix is used to automatically call PERL scripts, which then download and parse the approximately 550 Mb of relevant data on a nightly basis. All of the LocusLink, Unigene, GenBank and OMIM data is acquired using FTP, and then parsed. A table listing the URLs of the FTP and web source sites can be found on the CaGE information page (http://www.cage.wbmei.jhu.edu/information.htm). Toronto Cardiac Gene Unit and Body Maps data are acquired by repeated web hits and parsing. This is done on an IBM RS/6000 processor running the

*To whom correspondence should be addressed.
AIX 4.3.1 operating system. The PERL scripts use a module (LWP::Simple) to access web content and transfer it to the local server. Data is then parsed into tables that comprise CaGE. CaGE itself is being developed in IBM’s DB2 Universal Database version 7.0. The data is automatically uploaded into DB2 using a second set of PERL scripts.

The CaGE user interface consists of html pages that can be accessed from the URL: http://www.cage.wbmei.jhu.edu. Using the PERL::DBI (database interface) and a common gateway interface, the hyperlinks displayed dynamically interact with CaGE and generate html pages which display the results of the user’s query. Genes known to be expressed in human cardiac tissue can be browsed by the first letter or number of their official gene name. Basic searches can be performed to identify genes expressed in human cardiac tissue by searching either official or alias gene names, chromosome number or cytogenetic band, and official or alias gene symbols. Searches can also be done using a sequence accession number, Unigene cluster or LocusLink identifier. Lastly, clinical synopses of the cardiovascular system found within OMIM (OMIM, 2000) may be searched for genes known to be associated with a given human phenotype (e.g. cardiomyopathy or arrhythmia). The results of these queries generate html pages with a list of LocusLink identifiers and official gene names. The identifiers are hyperlinks that call a PERL script, which in turn generates what we call the gene ‘home page.’ The gene home page displays all data stored within CaGE for that given LocusLink. This will include official name, symbol, chromosomal location, alias names and symbols, the official gene product name, which data set or sets show the gene to be expressed and the cardiovascular clinical synopsis field of OMIM. Links to other references such as LocusLink, Unigene, RefSeq, GenBank, OMIM, Gene Cards, and KEGG are also displayed for each entry.

Currently, CaGE has 2 629 318 human GenBank accession numbers, 64 490 human Unigene Clusters and 46 320 human LocusLink identifiers represented in the knowledgebase. The current analysis shows that there are 6212 Unigene clusters expressed in human cardiac tissue. There are 658 human LocusLink identifiers in the BodyMaps Atria library, 695 in the BodyMaps Ventricle library and 2494 in the Toronto Cardiac Gene Unit library. An additional 1800 human LocusLink identifiers are from the Johns Hopkins gene expression data. Therefore, CaGE currently includes 7188 unique Human LocusLink identifiers expressed in human cardiac tissue. This is 15.5% of all Human LocusLink identifiers.

REFERENCES


OMIM™ (Online Mendelian Inheritance in Man) (2000) McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD).
