Building your own gene sets with WhichGenes, a new webtool for creating biological hypothesis to test in gene set based methods.

Daniel Glez-Peña[1], Florentino Fdez-Riverola[1,2], David G. Pisano[3], Gonzalo Gómez-López[3].

[1]Higher Technical School of Computer Engineering, University of Vigo, Ourense, Spain
[2]Informatics Department, University of Vigo, Vigo, Pontevedra, Spain
[3] Bioinformatics Unit (UBio), Spanish National Cancer Research Centre (CNIO), Madrid, Spain

INTRODUCTION

During the past several years, bioinformatics enrichment tools have played a very important and successful role contributing to the gene functional analysis of large gene for various high-throughput methods. From the large amount of tools that are currently available in the community, two widely used approaches can be identified: (i) individual gene analysis (IGA), which evaluates the significance of individual genes between two groups of samples compared, and (ii) gene set analysis (GSA), free from the problems of the 'cutoff-based' methods.

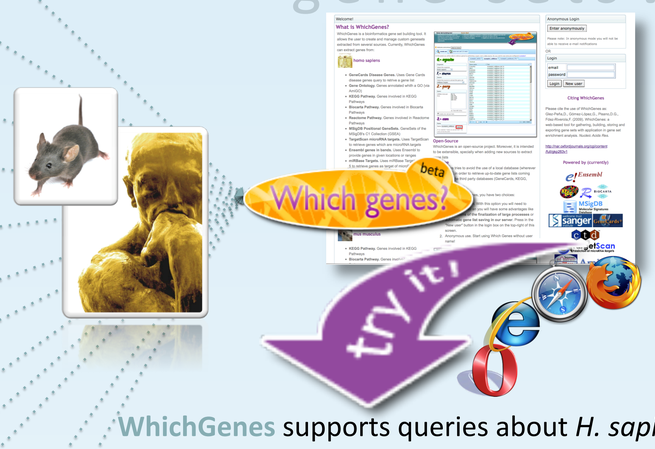
GSA approach has received a great deal of attention because, from a biological perspective, functionally related genes often display a coordinated expression to accomplish their roles in the cell.

In this direction, GSA methods enable the understanding of cellular processes as an intricate network of functionally related components. However, while tremendous effort has been invested during last years in developing new GSA methods, minimal effort has been put in developing tools that can help researches gather, store and manage gene sets containing large 'interesting' gene lists.

As a result, the process of routinely compiling and using sets of genes often involves tedious and time consuming steps that involucrate very different (and sometimes uncompleted) sources of information.

WhichGenes is an open source, database-free, web-based tool for easily gathering, building, storing and exporting gene sets coming from multiple data sources. It allows researchers to elaborate custom hypotheses in the form of lists of genes in order to further use them as input in existing GSA tools.

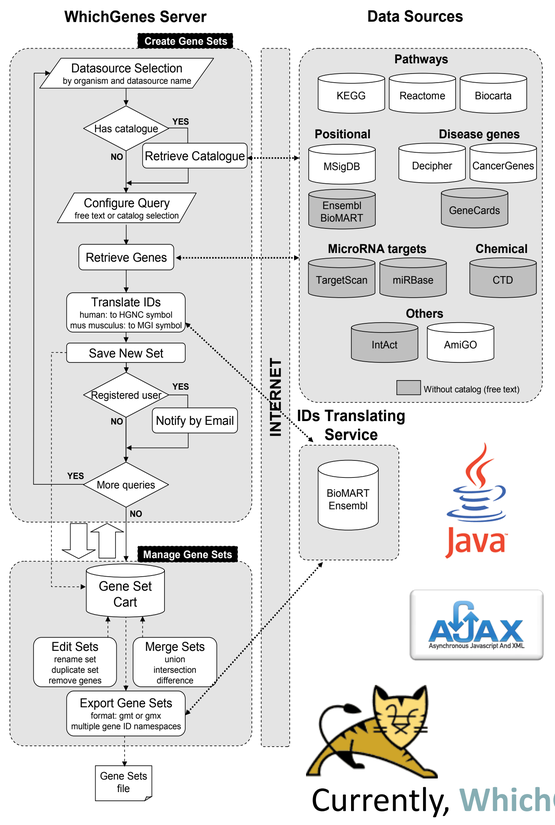
looking for gene sets?



WhichGenes supports queries about *H. sapiens* and *Mus musculus* by retrieving up-to-date gene lists directly coming from multiple databases including Ensembl, MSigDB, KEGG, Biocarta, Reactome, GeneCards, CancerGenes, Decipher, Diseases CTD, TargetScan, miRBase, IntAct, Chemical CTD and AmiGO.

Generated gene sets can be easily combined and exported using a wide variety of supported gene identifiers (i.e. Affy, Agilent, HUGO, Ensembl, mgi, refseq,etc.).

IMPLEMENTATION



WEB SERVER

WhichGenes is implemented as an AJAX-enabled web application programmed in J2SE 1.5 Java language. The ZK development framework (http://www.zkoss.org) was used to construct the user interface as well as different technologies to access the source databases including the BioMart XML-based query service for Ensembl and Reactome, the KEGG API and a custom developed library to download, parse, process and extract information from HTML pages, since many databases can only be accessed via their web sites. Caching techniques have been also implemented in order to reduce the number of accesses to the external data sources, especially when retrieving catalogues.

API

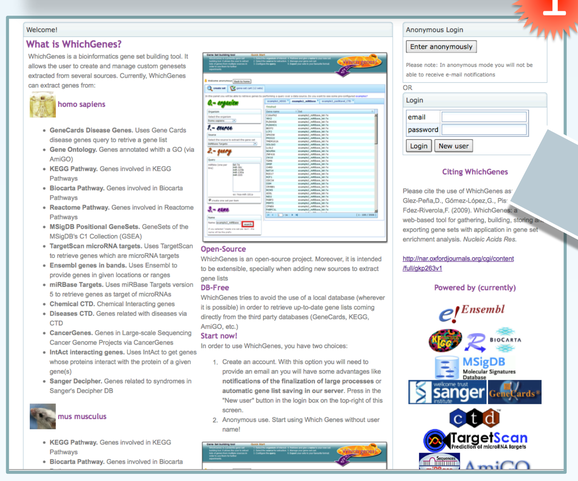
WhichGenes also implements a REST-based programming interface allowing developers to retrieve gene sets directly from our server and using them within their own algorithms (http://www.whichgenes.org/api/).

The system runs on a standard Tomcat 5.5 Web application server without any local database. An anonymous login is available for those who are interested in testing the system (without e-mail notifications) as well as preloaded searches and demo gene sets.

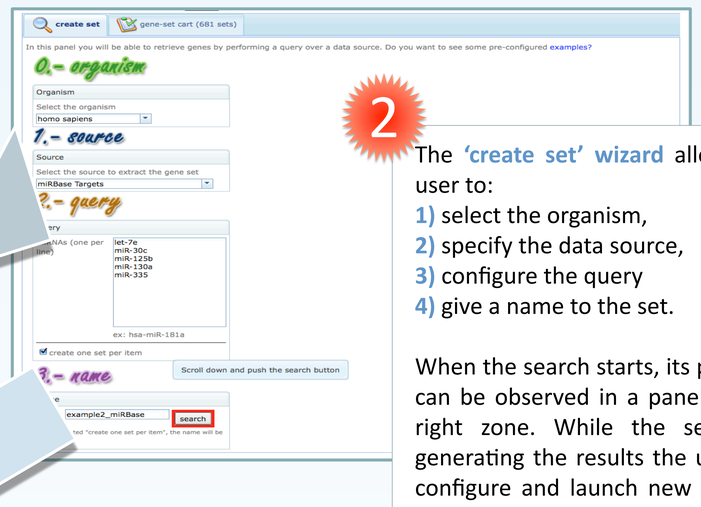
Currently, WhichGenes runs in Internet Explorer 7, Firefox 3, Opera 9.62 and Safari 3.

USING WHICHGENES ...always updated...

WhichGenes presents a web interface based on a simple to use 4-step wizard to create groups of interesting genes plus an intuitive control panel to manage and export them. **Users can login either anonymously or using their own account**.



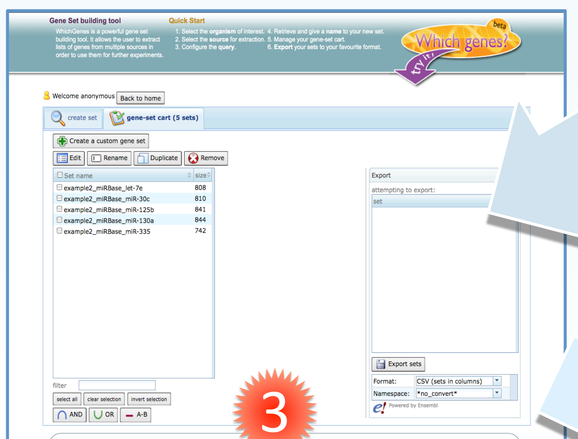
1



2

The **'create set' wizard** allows the user to:
1) select the organism,
2) specify the data source,
3) configure the query
4) give a name to the set.

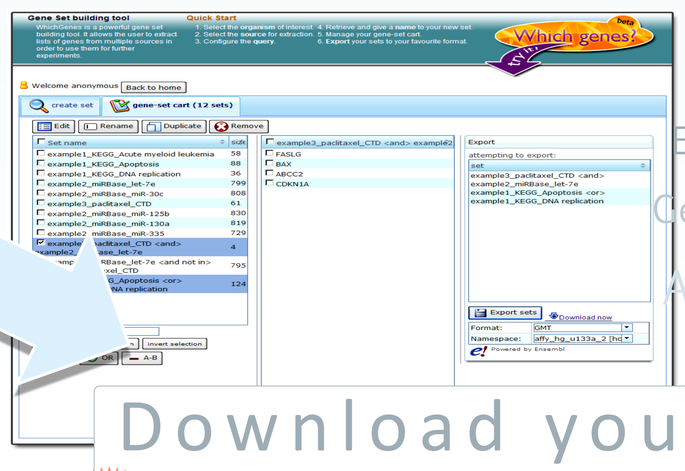
When the search starts, its progress can be observed in a panel on the right zone. While the search is generating the results the user can configure and launch new queries, which will run simultaneously.



3

Every time a new set is generated by WhichGenes, it is automatically added to the private **'gene-set cart' where the user can manage his own groups.**

The functionalities of this control panel allow users to obtain common genes, intersections and/or differences amongst chosen gene sets.



Ensembl ID
Entrez Gene
Gene Symbol
Mgi symbol
Affymetrix
RefSeq
Agilent
etc

4

Download your gene sets!!

CONCLUSIONS

✓ WhichGenes offers final-user facilities as well as a programmatic API for intuitively extracting lists of genes from multiple sources in order to extent the scope of existing gene set based analysis methods.

✓ The system is able to access and integrate information coming from a continuously growing plethora of information repositories including pathways, diseases, chemicals, GO, microRNAs, etc.

✓ The 'database-free' nature of WhichGenes implies that external data sources are accessed just in time, retrieving up-to-date information without using obsolete local mirrors of existing data sources.

✓ The user can not only retrieve and integrate 'interesting' gene lists from multiple repositories automatically, but also combine these sets (with the '*and*', '*or*' and '*difference*' operators) to build more complex hypotheses.

✓ WhichGenes can export gene sets in commonly used text formats (.csv, etc.) where the final gene namespace can be changed to a more appropriate one, including multiple popular database gene identifiers and several microarray probe set IDs.

www.whichgenes.org

Please cite the use of **WhichGenes** as:
Glez-Peña,D., Gómez-López,G., Pisano,D.G., Fdez-Riverola,F. (2009). WhichGenes: a web-based tool for gathering, building, storing and exporting gene sets with application in gene set enrichment analysis. Nucleic Acids Res. Web server Issue, 2009.