Gene Set Builder

Quick reference

Overview

INTRODUCTION

The Gene Set Builder (GSB) is a web-based application designed to help users create virtual “sets” of genes in an intuitive and user-friendly environment. The front-end of the application is driven by Hypertext Markup Language (HTML) with Cascading Style Sheets (CSS), Macromedia Flash, and JavaScript, while its backend is coded in Perl and employs a locally-installed MySQL database for data storage and retrieval.

The motive behind Gene Set Builder is a common issue: research in bioinformatics and genomics often involve the compilation and analysis of a set of genes, which require repetitive, manual work such as copying and pasting gene identifiers from a website onto a word processing document or spreadsheet. Also, handling extremely large sets that contain hundreds or thousands of genes can become a challenge when the data is kept in a “flat” format, such as a text file or a spreadsheet.

Gene Set Builder is designed to address this problem by helping researchers create, organize and store sets of genes in a powerful, readily-available, and useable format. The application saves time by handling many of the manual and repetitive aspects of building a gene set, such as hunting for transcripts or copying and pasting gene identifiers. With the help of a database and customizable user interface, the application can process and display vast amounts of information without overwhelming the user.

Other benefits include the option to attach comments and ratings to genes, as well as share sets with other users. Sets stored in the Gene Set Builder can be “exported” as a list of gene identifiers, a table, or sequences in FASTA format. Users can also instruct the program to create “mirror” homolog sets of an existing set, search Ensembl and GeneLynx using the built-in search engine, and obtain Ensembl transcripts using the software’s intuitive Transcript Manager. Of course, genes and sets that are no longer needed can be deleted to reduce clutter.

APPLICATION REQUIREMENTS

System

The Gene Set Builder requires a recent, graphical web browser running on a GUI platform. A screen resolution of at least 1024 pixels across (XGA) and a colour depth of 16 million are highly recommended. The optional, Perl-based API that can be used to retrieve gene and set information from the Gene Set Builder database requires a Perl interpreter and the DBI and DBD-mysql modules, as well as an Internet connection.

Browser

Gene Set Builder has been tested to work with recent versions of the Mozilla Firefox (www.mozilla.org) and Microsoft Internet Explorer (www.microsoft.com/windows/ie) web browsers. Other Mozilla-based browsers such as Netscape Navigator (www.netscape.com) should also be compatible with Gene Set Builder. Some KHTML browsers such as Konqueror (www.konqueror.org) have no problems rendering the Gene Set Builder, while older versions of Safari (www.apple.com/safari) cannot. Regardless of which browser to use, it is essential to enable JavaScript support. Due to the extensive use of tables, images, and inline frames, text browsing is not supported.

GSBM API

The GSBm API is a Perl module that enables other applications to retrieve gene and set objects from the GSB database. It can be downloaded from the Sign in page, http://www.cisreg.ca/gsb/.

Synopsis

subroutines	**functions**
get_genes_by_set	Retrieve the numeric id of the genes in a set. **In**: numeric id of the set (string) (e.g. 17) **Out**: numeric ids of the genes (array)
get_gene_ensembl get_gene_symbol get_gene_genelynx get_gene_entrez get_gene_refseq get_gene_uniprot	Retrieve the Ensembl, gene symbol, GeneLynx, Entrez Gene, RefSeq, or UniProt id of a gene. **In**: numeric id of the gene (string) (e.g. 1) **Out**: accession identifiers (string) (e.g. ENSG00000165029 or ABCA1, depending on which subroutine is used)
get_gene_cr get_gene_species get_gene_description get_gene_comment get_gene_comment_set	Other gene annotations: confidence rating, species, description, user comment, and set-specific user comment. **In**: numeric id of the gene (string) (e.g. 1) **Out**: number or text (string)
get_gene_ids get_gene_all_info	Retrieve all stored identifiers, or all stored identifiers + annotations, for a gene. **In**: numeric id of the gene (string) (e.g. 1) **Out**: key-value pairs (hash)

Keys for **get_gene_ids** and get_gene_all_info (**bolded** items apply for both subroutines; non-bolded items apply for only get_gene_all_info):

ensembl	=> Ensembl stable identifier
symbol	=> Gene symbol
genelynx	=> GeneLynx identifier
entrez	=> Entrez Gene accession
refseq	=> RefSeq accession
uniprot	=> UniProt accession
cr	=> Confidence rating
species	=> Species
description	=> Description
comment	=> User comment
comment_set	=> User comment (set)

get_set_info

Retrieve set annotations.

In: numeric id of the set (string) (e.g. 17)
Out: key-value pairs (hash)

Keys for **get_set_info**:

Name	=> Set name
description	=> Description
species	=> Species
author	=> Author
email	=> Email address
contact	=> Additional contact info
pubmed_id	=> PubMed ID

Usage example

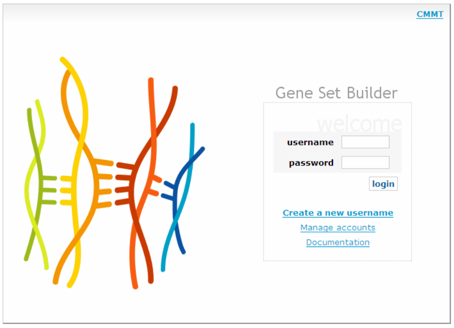
This script can be found in GSBm.zip, inside GSB_API_test.pl.

```
#!/usr/local/bin/perl

=for comment

        This PERL script demonstrates the basics of retrieving
        gene sets from the Gene Set Builder database.
        For questions or comments, please contact Dimas Yusuf,
        dyusuf@cmmt.ubc.ca.

=cut

#       In this script, we will retrieve information about set
#       number 17 and its genes. This set is a mouse set which
#       contains 4 Calcium Channel genes.

use GSBm::GSBm;
use strict;
```

```
#       To start things off, let's grab some set annotations.

my ($set_number, %set_data);

$set_number = "17";

%set_data = GSBm::GSBm::get_set_info($set_number);

#       The information is sent back as a hash, with the
#       following keys: name, description, species, author,
#       email, contact, and pubmed_id. Keep in mind that
#       sometimes, not all of these information are available.

my ($set_name, $set_description, $set_species, $set_pubmed_id);

$set_name        = $set_data{name};

$set_description = $set_data{description};

$set_species     = $set_data{species};

$set_pubmed_id   = $set_data{pubmed_id};

print "\n Some information about the set... \n";

print "Set name:    " . $set_name . "\n";
print "Species:     " . $set_species . "\n";
print "PubMed:      " . $set_pubmed_id . "\n";
print "Description: " . $set_description . "\n";

#       Now we will use the GSBm::GSBm::get_genes_by_set
#       subroutine to retrieve the entry numbers of the genes.

my @genes = GSBm::GSBm::get_genes_by_set($set_number);

print "\n These are the number ids of the genes in set \"" .
$set_name . "\", number " . $set_number . ". \n";

my $counter = 1;

foreach my $gene_db_number (@genes) {
        print "Gene " . $counter . ": " . $gene_db_number . "\n";
        $counter++;
}

#       Unfortunately, numbers are quite useless. For the next
#       step, we will retrieve the Ensembl ids of each gene via
#       the foreach statement.

$counter = 1;

print "\n These are the Ensembl ids of the genes in set \"" .
$set_name . "\"... \n";

foreach my $gene_number (@genes) {

        my @ensembl_id = GSBm::GSBm::get_gene_ensembl($gene_number);

        print "Gene " . $counter . ": " . $ensembl_id[0] . "\n";

        $counter++;

}

#       To get other information, the "ensembl" can be replaced
#       with "symbol" (forming "get_gene_symbol"), genelynx,
#       entrez, refseq, uniprot, cr, species, description,
#       comment, or comment_set. Cr is for confidence rating,
#       by the way.

#       Note how the information is sent as an array--this is
#       because a single gene can have multiple identifiers.
#       To get all gene identifiers, it is easier to use
#       get_gene_ids.

#       Remember to use a hash when receiving the data. Keys
#       include ensembl, symbol, genelynx, entrez, refseq,
#       and uniprot.

my $first_gene = $genes[0];

print "\n These are the various identifiers associated with the
first gene in the set, gene number " . $first_gene . "... \n";

my %gene_identifiers = GSBm::GSBm::get_gene_ids($first_gene);

print "Ensembl:      " . $gene_identifiers{ensembl} . "\n" ;
print "Gene symbol: " . $gene_identifiers{symbol} . "\n" ;
print "Genelynx:     " . $gene_identifiers{genelynx} . "\n" ;
print "Entrez Gene: " . $gene_identifiers{entrez} . "\n" ;
print "RefSeq:       " . $gene_identifiers{refseq} . "\n" ;
print "UniProt:      " . $gene_identifiers{uniprot} . "\n" ;

#       And finally, the subroutine get_gene_all_info will
#       retrieve all information associated with the gene.
#       Like get_gene_ids, it returns the data as a hash,
#       which utilizes the following keys:
#       ensembl => Ensembl stable identifier
#       symbol => Gene symbol
#       genelynx => GeneLynx identifier (human, mouse, or rat)
#       entrez => Entrez Gene accession
#       refseq => RefSeq accession
#       uniprot => UniProt accession
#       cr => Confidence rating
#       species => Species
#       description => Description
#       comment => User comment (general)
#       comment_set => User comment (set-specific)

print "\n More information... \n";

my %gene_all_info = GSBm::GSBm::get_gene_all_info($first_gene);

print "Confidence rating:  " . $gene_all_info{cr} . "\n" ;
print "Species:            " . $gene_all_info{species} . "\n" ;
print "Description:        " . $gene_all_info{description} . "\n" ;
print "Comment 1:          " . $gene_all_info{comment} . "\n" ;
print "Comment 2:          " . $gene_all_info{comment_set} . "\n" ;

print "\n Thank you, come again! \n";

#       End of script
```

Usage

STARTING UP

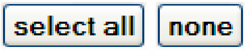


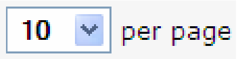
Gene Set Builder can be accessed via http://www.cisreg.ca/gsb/. Users can create a new username and password combination by clicking on the “**Create a new username**” link. Please remember your username and password as—for privacy and security—the software will not ask you for an email address or a password reminder question (incase you forgot your username or password.) Unused usernames can be deleted manually by the user by clicking on “**Manage accounts**”.

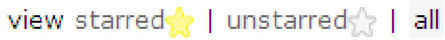
FUNCTIONS

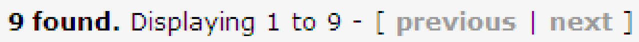
Common items

The following icons and functions are located throughout Gene Set Builder:

- ⟲ Click this icon to reload the page.
- 

These buttons allow you to select or deselect all checkboxes that are located on the page.
- 

You can use this handy drop down menu to specify how many genes, sets, or transcripts you would like displayed on a single page.
- 

This panel lets you display only starred, unstarred, or all items.
- 

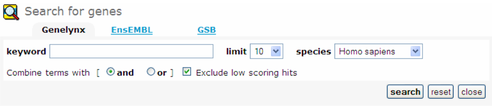
See how many genes or sets are available and crawl through multiple pages using this tool, located at the top right side of most areas.

Genes

Search for genes

Search GeneLynx and Ensembl for genes. GeneLynx is a consolidated resource centre for known human, mouse, and rat genes, while Ensembl stores known and predicted genes of over 15 different species. By clicking on the GSB tab, you can also search through your gene collection in the Gene Set Builder database.

Screenshot:



Search for genes

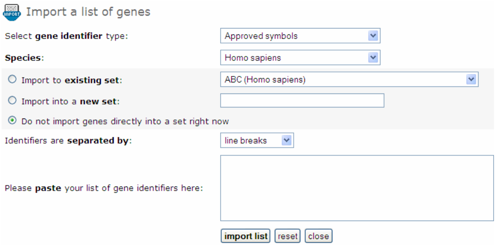
keyword:
limit: 10
species: Homo sapiens

Combine terms with [● and ○ or] ☑ Exclude low scoring hits

Import a list of genes

Users with a list of gene identifiers (Ensembl stable ids, GeneLynx, Entrez Gene, UniProt, RefSeq DNA, Affymetrix ids, or gene symbols) can use this tool to find more information about each gene and import the list into a set.

Screenshot:



Import a list of genes

Select **gene identifier** type: Approved symbols

Species: Homo sapiens

○ Import to **existing set**: ABC (Homo sapiens)

○ Import into a **new set**:

● Do not import genes directly into a set right now

Identifiers are **separated by**: line breaks

Please **paste** your list of gene identifiers here:

Add genes to set

Browse through all the genes that you have imported into the Gene Set Builder database. From here, you can delete genes, move them into sets, add comments, and view detailed information on each one.

There are some features implemented in this area that can help you view and organize your data. Genes of particular interest can be marked with yellow stars. Whenever you need to view these genes right away, you can click on the “view starred” link at the top of the page. Likewise, you can also click on “view … unstarred” and “view … all” to see genes that are not marked, and so on. The “select all” and “select none” buttons are useful whenever you want to sort or delete large amounts of genes. You can also sort the list by gene symbol, species, description, Ensembl stable id and GeneLynx id by clicking on the “Name”, “Species”, “Description”, “E” and “GL” icons.

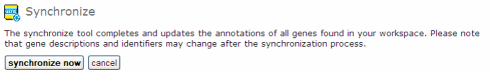
Functions that can be accessed in this area:

- **Clean up**: remove unused genes from the database.
- **View gene properties**
  See more information about the gene.
- **Add a comment**
  Attach a note to the gene.
- **Star or "un-star" a gene**
  Click on these icons to switch the yellow star on or off.

Synchronize

Obtain an expanded set of information for the genes in your collection, one which includes their RefSeq, UniProt, and Entrez Gene identifiers. Please be patient as the synchronization process requires several minutes.

Screenshot:



Synchronize

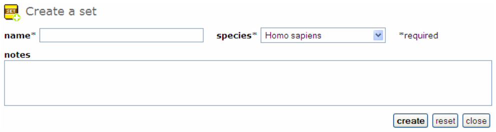
The synchronize tool completes and updates the annotations of all genes found in your workspace. Please note that gene descriptions and identifiers may change after the synchronization process.

Sets

Create a set

Sets can be made via this feature.

Screenshot:



Create a set

name*

species* Homo sapiens

*required

notes

Stored sets

You can browse through your sets by clicking on this icon.

Functions that can be accessed in this area:

- create a set Open the “Create a set” dialog

- delete set delete genes and set
  The “delete set” button deletes all selected sets, but not their genes (this feature is perhaps useful when you’d like to add the same genes to another set, but would not like to import them again.)

- GENE **Add genes to this set**
  Open the “Add genes to set” area, where you can browse through all imported genes and pick which ones you’d like to add to that particular set.

- SET **View set properties**
  View set annotations.

- SET **Export this set**
  Open the “Set exporter” dialog, where you can command the Gene Set Builder to export the set as a list of identifiers, a table, or a series of gene/cDNA sequences.

- SET **Share this set**
  This grayed out icon hints that the set is not shared. Clicking it will open the “Share this set” dialog.

- SET **Withdraw this set**
  This colour icon hints that the set is already shared. Click it to open the shared set manager, where you can alter annotations or “withdraw” the set’s shared status.

- ☆⭐ **Star or “un-star” a set**
  Click on these icons to switch the yellow star on or off.

When you have chosen a set, you can access the following functions:

- AT CG **View transcripts**
  View the Ensembl transcripts of the genes found in that particular set.

- GENE **View gene properties**

See more information about the gene.

- **Add a comment**
  Attach a set-specific note to the gene. This note will only accompany the gene in this set.

- **Confidence ratings**
  Click on the stars to change the gene's confidence rating:
  ★★★★★ Very high confidence (100%, or 5/5 stars)
  ★★★★☆ High confidence (80%, or 4/5 stars)
  ★★★☆☆ Moderate confidence (60%, or 3/5 stars)
  ★★☆☆☆ Low confidence (40%, or 2/5 stars)
  ★☆☆☆☆ Very low confidence (20%, or 1/5 stars)
  ☆☆☆☆☆ No confidence (0%, or 0/5 stars)

Generate homolog set

With the help of the Ensembl Ensmart Homology annotations, this tool takes one of your sets and converts it to another species.

Transcript manager

View available cDNA transcripts by set.

Functions that can be accessed in this area:

- **View set properties**
- **Obtain or delete transcripts**

When you have chosen a set, you can access the following functions:

- **Set as default transcript**
  These tiny blue icons are found beside each transcript. Click on it to set that transcript as the default transcript for the gene.
- **View set properties**
- **Export this set**

- **Share this set**
- **View gene properties**
- **Add a comment**

Browse shared sets

Display the sets of other users that are shared.

Functions that can be accessed in this area:

- **Browse all shared sets**
  Display all shared sets in the database.
- **Export this set**
- **Copy into your collection**
  Install a copy of the shared set in your account. Most set annotations will be preserved.

Set exporter

Generate a **List**, **Table**, **FASTA** or **API** output.

IMPORTANT: RefSeq, Entrez Gene, and UniProt gene identifiers are available for export only after the genes have been synchronized. Use the "Synchronize" tool to accomplish this task. When exporting a set as a FASTA-formatted list of transcripts, please ensure that you have already used the Transcript manager to obtain the necessary cDNA Ensembl transcripts for the genes in that set.

Set copier

Use this feature to create copies of existing sets.

API

Click here to view and delete your API-accessible gene sets.

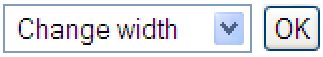
Tools

Documentation

Opens up this quick reference.

Feedback

View the contact page. Comments, criticisms, and suggestions can be forwarded to Wyeth Wasserman, email: wyeth@cmmt.ubc.ca. You can also email the developers: Dimas: dyusuf@cmmt.ubc.ca, and Jonathan: jlim@cmmt.ubc.ca.



Adjust the width of the user interface, up to 2560 pixels wide (QXGA). By default, a resolution of 1024 pixels (XGA) is selected.