GenesetPCA user manual

Setup

Necessary software and libraries

- Matlab
  - Statistics toolbox
  - isintersect.m (http://www.mathworks.com/matlabcentral/fileexchange/27673-2d-polygon-edges-intersection/content/polygonstuff/isintersect.m)
  - poly2poly.m (http://www.mathworks.com/matlabcentral/fileexchange/27673-2d-polygon-edges-intersection/content/polygonstuff/poly2poly.m)

Running the program

To run the program:

- Open Matlab
- Make sure all scripts are in directories which sit in your path ('set Path' button)

```
>> genesetpca_openingdialog.m
```

- You will now be asked to select a datafile and a geneset file (see figure 1)

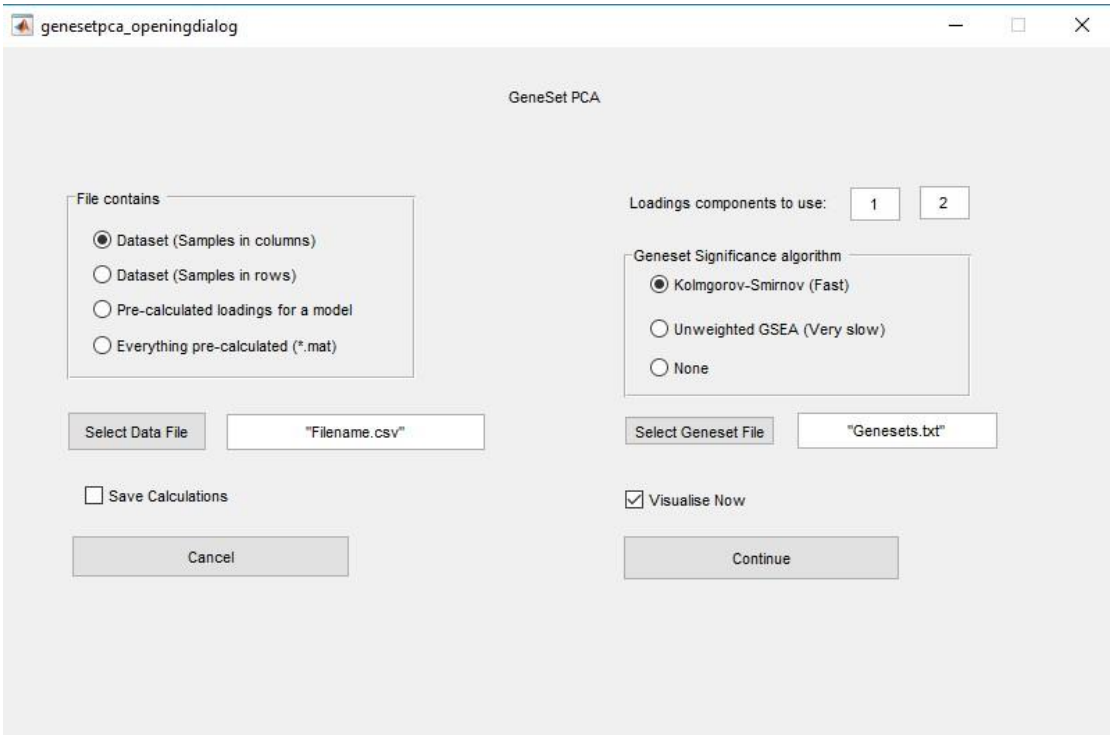

Figure 1 The opening dialog box for genesetpca.

The dataset file

The dataset file can be in 4 different formats:

- CSV (comma separated values) with the samples in columns. The first row contains sample names and the first column contains gene identifiers.
- CSV (comma separated values) with the samples in rows.
- Pre-calculated loadings for a model – this option should be chosen if you already have a PCA, or another type of multi-variate model which you have calculated loadings for (eg PLS-DA model or OPLS model). Again the file should be in CSV form, with the first row as labels for each column and each row should contain the model loadings for one gene. The first column must contain the gene identifiers.
- Everything Pre-calculated – this can quickly reload a file saved from a previous run of this program. **Hint** If you want to use this option, select it first before clicking the select file button, then the interface will automatically show the *.mat files rather than the *.csv files.

The geneset file

The geneset file should describes the sets of genes which are going to be used to analyse the loadings. A human geneset file with entrez-gene ids is provided with the code as an exemplar. The file format matches that output by Consensus Path Database (http://consensuspathdb.org/). Files for human, mouse and yeast can be downloaded from this site with a range of gene identifiers. Files downloaded from here should have their file extension changed from *.tab to *.txt in order to load without issues.

IMPORTANT – The gene identifier used in your geneset file, must match the gene identifier used in your dataset file.

For other organisms, the following file format can be used to construct appropriate files:

- Column 1 – A unique identifier for each geneset, plain text, no symbols.
- Column 4 – A comma separated list of gene identifiers for the genes in the gene set.
- Columns 2 and 3 can contain anything.
- Columns should be tab separated.
- Save as a text file (eg .txt).

Loadings components

By default the program will examine the first vs the second component of a PCA model. However, if you wish to look at other components this is possible using the boxes on the top right of the dialog box. This is particularly useful when you have already generated a multi-variate model and wish to examine the loadings of the components in which you see interesting behavior.

Geneset significance algorithm

Currently two geneset significance algorithms have been implemented:

- **Kolmogorov-Smirnov**
  This algorithm looks at the cumulative distribution of all the loadings given (in that component), and at the cumulative distribution of the loadings of genes in the geneset and looks for a significant difference.

- **UGSEA**
  Unweighted Gene Set Enrichment Analysis, is an implementation of the original GSEA algorithm (Subramanian et al., 2005). This looks for sets of genes which are grouped together (anywhere) in the list and evaluates significance through permutations.

A note on timings

Once the opening dialog has been completed, then the calculations associated with the model will be performed. Depending on the size of the dataset and the number of genesets in the geneset file and the algorithm selected this can take a while. Some example timings from a HP Elitebook are found below;

Explore significant genesets

One the calculations have been performed, if the 'Visualise now' checkbox was ticked then the explore_significant_pathways graphical user interface will be launched (see figure 2).

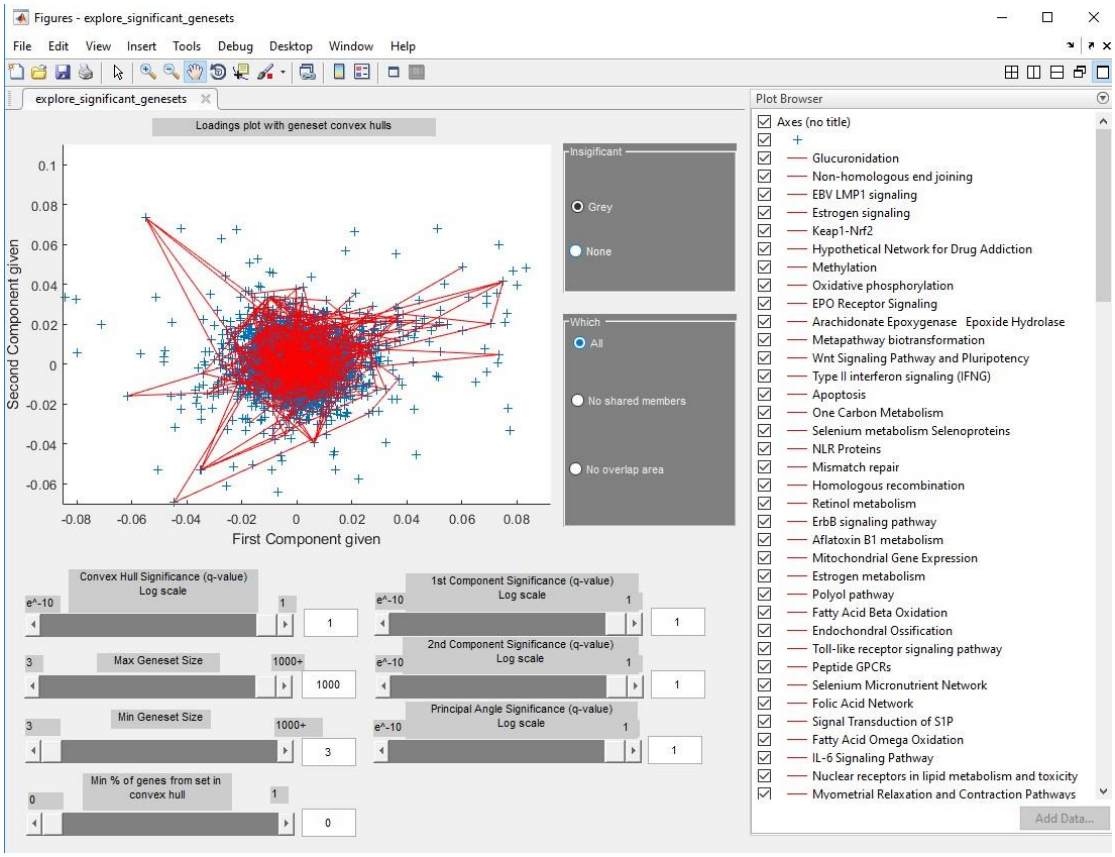

Figure 2 – Screenshot from explore significant genesets.

When using the GUI interface there are four tools which can be selected to navigate the interface, it is **crucial** to select the right tool for the task you wish to do. The four tools are; the pointer/arrow; the enlarging magnifying glass; the shrinking magnifying glass; the hand.

The tasks which can be performed in the GUI are grouped under the tool which **must be** selected to use them.

The hand (pan)

WARNING!!!
If an attempt is made to use the sliders or radio buttons when the hand isn't selected, the screen element will move around the screen – switch back to the hand before clicking near these elements.

The hand is the most important tool it is the default when the GUI is opened. When the hand is selected you can perform the following tasks;

- **Adjusting the sliders** – there are five sliders at the bottom of the GUI, these adjust the parameters which determine which sets are shown in the plot. All probability-based sliders are based on the Benjamini-Hochberg false discovery rate adjusted q-values and are adjustable according to a log scale. All sliders can also be adjusted using the text entry box found next to the slider, to allow additional precision.

The six sliders are;

- **Convex Hull significance** – this adjusts the significance of the polygon, less significant polygons cover more genes which are not in the set.

- **Max pathway size –** this adjusts the maximum size of the sets shown in the plot.

- **Min pathway size –** this adjusts the minimum size of the sets shown in the plot

- **PC1 significance –** this adjusts the significance of the enrichment of the set in the first principal component. Significantly enriched sets will tend to genes grouped together in the x-co-ordinate.

- **PC2 significance –** this does the same as above but for the second PC.

- **Loadings angle significance –** The loadings angle is the angle that a line from the origin to the point makes with the x-axis. These angles were calculated and then ranked. To avoid issues around $-\pi/2$ being in reality close to $\pi/2$ in the circle, a second angle was calculated as the angle from the same line to the negative x-axis. Significantly enriched sets will tend to have genes with similar angles.

- **Min % of genes from set in convex hull –** this allows you to filter out those sets whose genes are scattered across the plot, and therefore the optimal convex hull contains just a few genes from the set. (NB. The enrichment should filter out some of these genesets anyway, but this adds a second level of filtering which is sometimes useful.)

Example:

You see separation between your samples in the second component and you want to explore which genesets might be related to this – you should set the significance for the first PC to 1 (so all genesets pass this criteria) and explore stricter thresholds for the second PC. To keep from displaying polygons which cover too many genes, adjust the convex hull significance.

- **Changing how you view insignificant sets –** There are two options for those sets which don't meet the thresholds set through the sliders, firstly they can still be displayed, but in grey rather than red and secondly the display can be turned off. This is controlled through the radio buttons at the top right of the screen.

- **Cutting down on overlapping sets –** As the set databases often contain many very similar sets with different names this can cause difficulties with interpretation. We currently offer two simple options to cut down the number of overlapping displayed sets. Both of these options rely on greedy search, which means that the best set from those which meet the thresholds set in the sliders is selected, where best is defined as having the lowest value when the enrichment p-values for both principal components are multiplied. Then the second best set is examined to see if it overlaps with the best. It is only displayed if no overlap is found. In this way a set of non-overlapping sets is chosen for display. The two types of overlap considered are;

  - **Overlap of members** - This option says that displayed sets shouldn't contain the same members, the polygons may still overlap and even contain shared genes which are not members of both displayed sets.

  - **Overlap of area** – For this option the polygon is considered rather than the only the members of the set. This is the stricter option, however that does not necessarily mean that it will display less sets (for instance if the second set selected by the first method was very large, but it overlapped in area with the first set, then it wouldn't

be selected by this method and therefore more small sets could be selected in its place.)

WARNING!!! Sometimes the program can get lost following which sets should be displayed. In this case, choose to display all significant sets, and then reselect the option.

- **Moving the plot without zooming –** the hand, as is standard in Matlab, can also be used to drag the viewer to a different area of the plot, without applying the zoom.

 The arrow/pointer (edit plot)

The arrow/pointer is useful when the plot browser is open. The plot browser can be opened through the view menu, and selecting plot browser.

- **Selecting a set by polygon -** Click on any polygon when the arrow is highlighted and the name of the corresponding set will become highlighted in the plot browser.
- **Selecting a set by name -** By selecting the name in the plot browser the polygon showing that set in the plot will be highlighted.

 The magnifying glasses (zoom in/out)

The magnifying glasses can be used as usual in Matlab to zoom into/out of the plot. A single click will zoom a fixed amount, or an area can be selected through dragging, and then the plotted area will change to be the selected range.

Preparing data

When preparing data for use with this software we followed these steps;

- Downloaded the processed data from a repository (such as array express or GEO)
- Used g:convert (http://biit.cs.ut.ee/gprofiler/gconvert.cgi) to convert the gene ids given into entrez-gene ids (as our geneset file used entrez-gene ids – to use other identifiers select a different geneset file).
- Selected a single entrez-gene id (the first[1]) where a given id in the original dataset maps to more than one entrez-gene id.
- Summarised data to give a single value per entrez-gene id by taking the median value.
- Plotted a histogram of the log (base 2) of all the values, and selected a cutoff to remove unexpressed genes – remember most genes are unexpressed in most samples, so the large peak of genes at the low end of this histogram will be the unexpressed genes which should be discarded. A typical cutoff has been found to be 2^6 (64).
- Removed all genes whose maximum value in any sample was below the chose cutoff.
- Identify the samples using the sample relationship table downloaded from the repository, checking the order of the samples carefully before copying the labels across.

All these steps were generally performed with the data stored in Excel, and examples showing how this was done can be found in the example datasets.

- Finally we copied the final data table of the selected genes into Matlab, along with the labels and class columns.

[1] ***NB**. This strategy is not necessarily optimal and was chosen for convenience only. You may wish to check which genesets the alternative ids are in and select the one with the best coverage. Note – it is **not** a good idea to use all alternative ids, as when these appear in the same sets –as they frequently will, it will overestimate the significance of those sets.*

Frequently asked questions

- **Will Genesetpca analyse my PLS-DA model?**
    - Genesetpca can analyse other types of multi-variate models. To do this you need to first calculate the model yourself, and then upload a csv file containing the loadings of the model. From a raw dataset only PCA is analysed.
- **Can I save the output?**
    - Yes, you can save the output in several formats – either when generating the model and analysis through the initial dialog or through the save dialogs to save matlab figures as an interactive figure or a flat image.