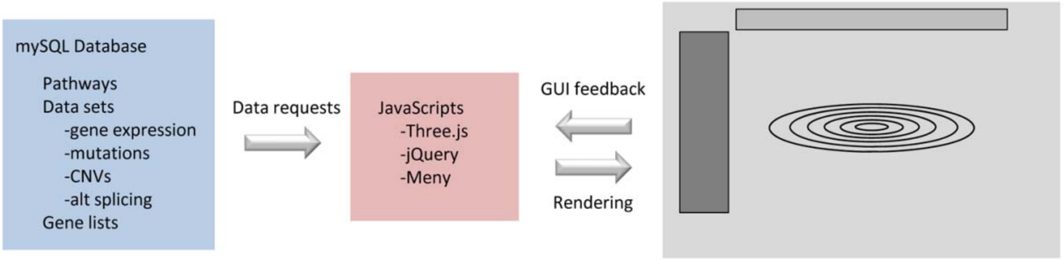
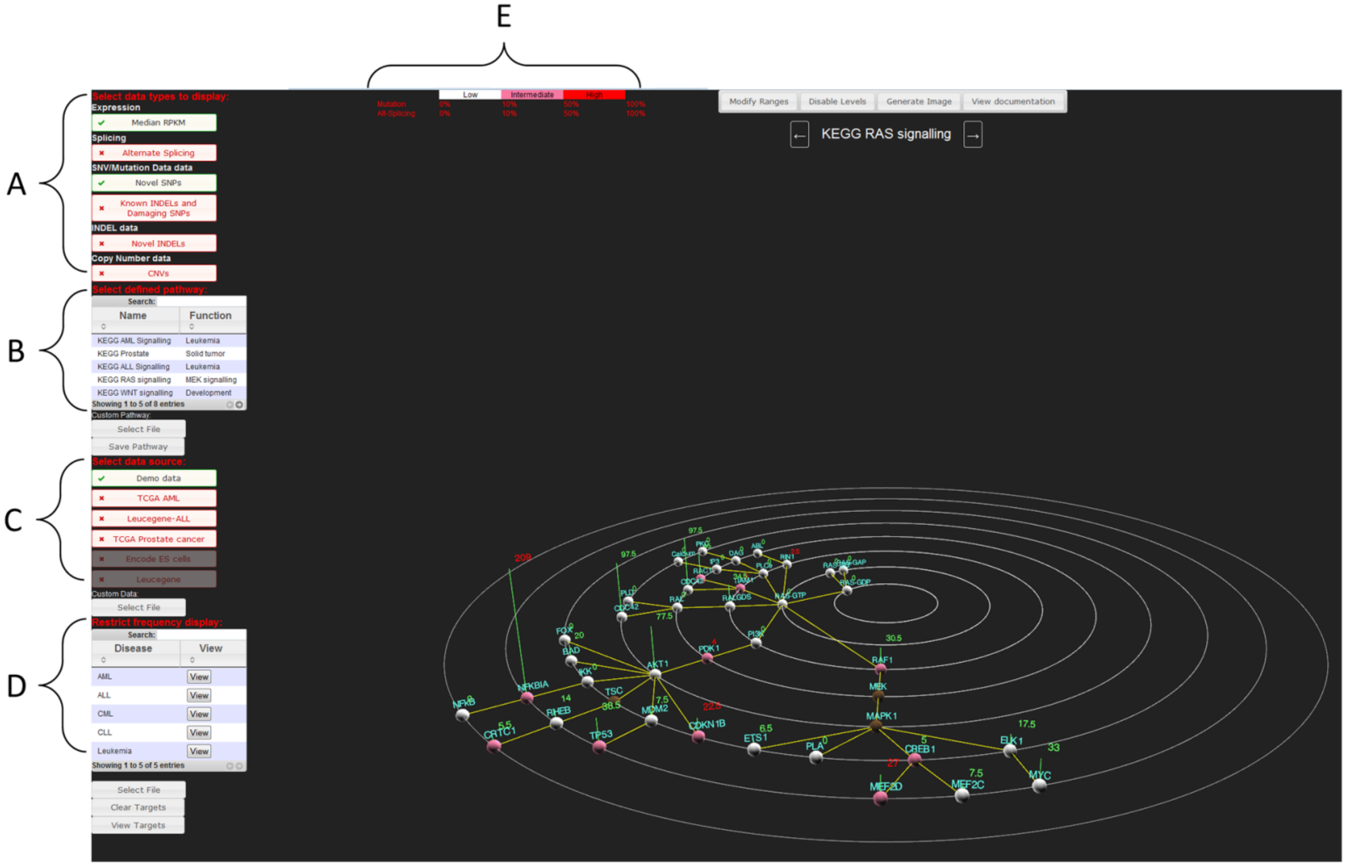
Cascade – A novel tool for exploring multidimensional RNA-seq data

Cascade is a software tool which allows the representation and the exploration of multidimensional RNA-seq data, and specifically cancer genomics data. The object of the software is to reduce the dimensionality of RNA-seq data into a single intuitive view that builds on existing biological knowledge. Cascade consists of a my SQL database used to store experimental data and a main interface written in JavaScript which uses the three.js JavaScript 3D library for rendering of known biological pathways.



Known biological pathways are displayed as a series of connected nodes, where each node represents a gene. The dimensionality of the RNA-seq is achieved by using node colour to represent mutation frequencies, vertical bars with values attached to each node to show average gene expression (RPKM) values and shape and Z-axis position changes to highlight alternative splicing and copy number variations respectively. The thresholds used for colour/shape changes for nodes are user tunable through a menu option on the main screen. Users interact with Cascade using a space saving menu on the left-hand side to select features of the RNA-seq data to be displayed (A), pre-defined or custom biological pathways to view (B), specific datasets to use for visualization of features selected (C). Additionally, predefined or custom disease gene lists (D) can be used to restrict the colouring thresholds defined by the used (E). The "modify ranges" button (top centre) allows users alter the thresholds required for node colour changes (mutations) or ring appearance (splicing).



	Low	Intermediate	High	
Mutation	0%	10%	50%	100%
Alt-Splicing	0%	10%	50%	100%

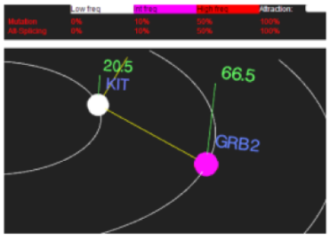
Name	Function
KEGG AML Signalling	Leukemia
KEGG Prostate	Solid tumor
KEGG ALL Signalling	Leukemia
KEGG RAS signalling	MEK signalling
KEGG WNT signalling	Development

Showing 1 to 5 of 8 entries

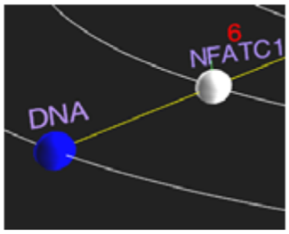
Disease	View
AML	View
ALL	View
CML	View
CLL	View
Leukemia	View

Showing 1 to 5 of 5 entries

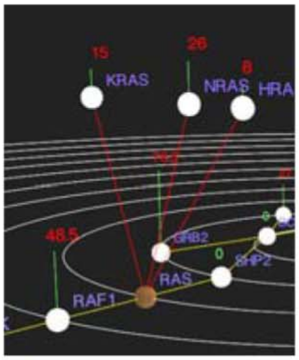
Examples of node colouring and elements in Cascade



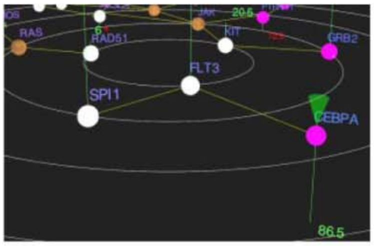
A gene (GRB2) with a mutation frequency between 10 and 50% in selected data set and an average RPKM value of 66.5.



Non-genic elements represented as blue nodes. Genes with outliers in dataset for gene expression have average RPKM values shown in red



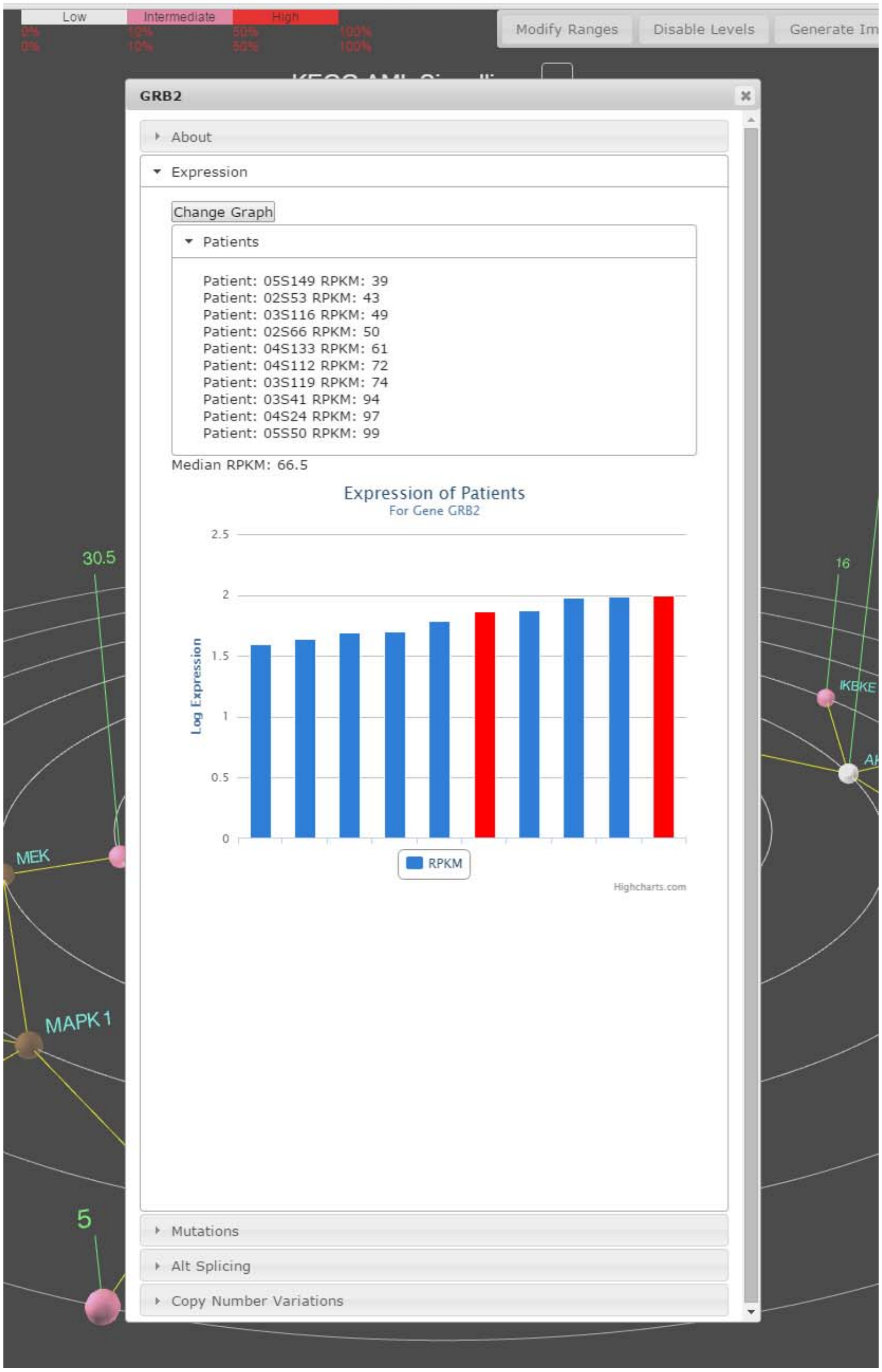
Families of genes represented as brown nodes; Members can be expanded by clicking on parent node



Copy number changes shown by green (loss) or red (gain) cones with node below or above plane respectively

Gene specific information

Clicking on specific nodes will bring up a vertically tabbed menu with information for: Generic description of gene retrieved dynamically from GeneCards, and sample specific information for gene expression (with mutated samples shown in red), mutations, alternative splicing and copy number variations. Samples can be sorted by column values within tabs.



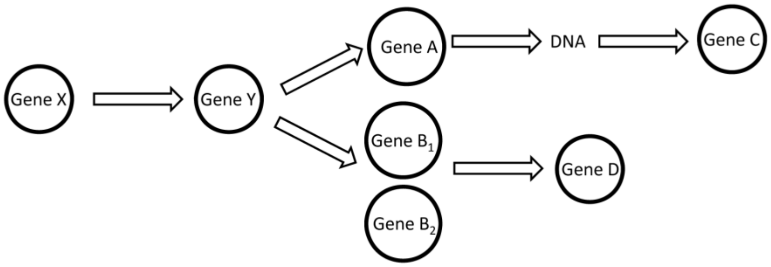
```
Patient: 05S149 RPKM: 39
Patient: 02S53 RPKM: 43
Patient: 03S116 RPKM: 49
Patient: 02S66 RPKM: 50
Patient: 04S133 RPKM: 61
Patient: 04S112 RPKM: 72
Patient: 03S119 RPKM: 74
Patient: 03S41 RPKM: 94
Patient: 04S24 RPKM: 97
Patient: 05S50 RPKM: 99
```

Median RPKM: 66.5

Creating pathways for Cascade

The biological pathways currently represented in Cascade are based on KEGG pathways that have been converted into a tab delimited format suitable for Cascade. There are two methods for adding a new pathway to Cascade: 1) Addition via the database 2) temporary addition via a file upload. For the database method, the information is stored in a MySQL table called "genes" which is queried when users select or change views in Cascade. Shown below is an example of a pathway and how the data can be encoded for use with Cascade in either format.

Pathway image



Format for database addition:

Gene_index	Pathway_ID	Gene	Children	Node_type
1	1	Gene X	Gene Y	gene
2	1	Gene Y	Gene A, GeneB_fam	gene
3	1	Gene A	DNA	gene
4	1	DNA	Gene C	endpoint
5	1	Gene C		endpoint
6	1	GeneB_fam:F(Gene B1|GeneB2)	Gene D	gene
7	1	Gene D		endpoint

File upload format:

Gene	Children	Node_type
Gene X	Gene Y	gene
Gene Y	Gene A, GeneB_fam	gene
Gene A	DNA	gene
DNA	Gene C	endpoint
Gene C		endpoint
GeneB_fam:F(Gene B1|GeneB2)	Gene D	gene
Gene D		endpoint

The only differences between the database addition compare to the file upload are the first two columns in the table, which represent a unique index and identifier for the pathway required only for the database. The genes in the pathway are organized into two columns representing "parent and child" nodes which allow connections to be drawn between them in Cascade. The 3rd column in the table contains a single HUGO gene name or unique identifier for a family of genes with HUGO gene names within brackets. The use of proper HUGO names is essential for the automated retrieval of gene information from NCBI. The 4th column lists all of the child nodes connected to the parent at the left, separated by commas. There is no specific limit to the number of child nodes that can be attached, although more than 8 (depending on where in the pathway they are) may result in sub-optimal displays. Families of genes can be defined by any name followed by a ":F" as a child node (e.g. RAS_Kinases:F) and whenever the family is defined in the gene column, the members are identified (using HUGO names) separated by pipes. The 5th column is a key word description for the parent node in the row used to select the shape and properties used to represent it. The choices are either "gene" (a coloured sphere) or "endpoint" (a blue square) and they can be used in any order with the pathways.

There are several important points to keep in mind for generating functional pathways:

1) All nodes within a pathway must be resolved. This means they must either connect to another node/end-point or they must have their own row specifying that they represent an endpoint.

2) The use of HUGO gene names is important for the NCBI gene information retrieval function.

3) Genes can be duplicated within a single view (i.e. appear within 2 alternative versions of the same pathway) as long as they are resolved within both pathways. Their behaviour is independent within the view (expanding a duplicated gene family will only affect the family node clicked, not both).

4) Cascade does not currently support converging pathways, but this feature is anticipated for future versions of the software.

Loading custom data for Cascade

As with custom pathways, custom patient data can be loaded either from the database or from user supplied files using the CSV format shown below:

```
id,patient,expression,snp,damaging,indel,splice,cnv
RTK,sample_1,35,1,0,0,0,0,
ILK,sample_1,87,0,0,0,0,0,
PI3K,sample_1,62,0,0,0,0,0,
PTEN,sample_1,50,0,0,0,0,0,
PDK1,sample_1,72,0,0.5,0,0,0,
AKT,sample_1,43,0,0,0,0,0,
MDM2,sample_1,48,0,0,0,0,0,
p53,sample_1,91,0,0,0,0,0,
```

The expression values entered need not be RPKM/FPKM data from RNA-seq experiments however if microarray data is used, it is likely that auto-scaling values in the main page (index_run.php) would need to be altered to provide a useful view of these data. Likewise, any of the data types could theoretically be co-opted to represent some other data type (e.g. Alternative splicing column could be used for DNA methylation) if the data can reasonably be binary encoded (e.g. 1= >50% DNA methylation, 0=<50%). For any data types not available to the user, zero values should be entered to maintain the format.

Cosmetic alterations for Cascade

Cascade uses a small file (config.php) to obtain various values for colours of text, background, links between nodes and font size. There are currently two colour palettes supplied that users can select by commenting out the alternative color palette through the use of # symbols for each line or multi line comment s (/* comment */). Likewise, the button names for data types (if co-opted as explained above) can be altered by any text editor using the "meny.php" file.