Documentation for InCroMAP

integrated analysis of microarray data from different platforms

Clemens Wrzodek

March 9, 2012

InCroMAP is a powerful, easy-to-use high-level cross-platform microarray dataset analysis tool. It provides several methods to analyze or visualize single datasets, as well as methods to perform integrated cross-platform analysis. Currently mRNA, miRNA (microRNA), DNA methylation and protein (modification) data are supported. InCroMAP can, for example, visualize all platforms at once in a pathway, provide detailed information, e.g., about promoter methylation or perform integrated mRNA and microRNA analysis. In general, InCroMAP is a tool for generic or pathway-based analysis and visualization of heterogeneous, cross-platform datasets.

Contents

1 Introduction

InCroMAP is an application that is able to import processed microarray data from different platforms and perform gene-centric integration of those. It holds many novel analysis methods that allow, e.g., integrated pathway-based visualization, gene set enrichment across multiple platforms or integrated tabular views. InCroMAP can handle directly the following four platforms:

1. messenger RNA (mRNA) expression data

2. micro RNA (miRNA) expression data

3. protein and protein modification expression data

4. DNA methylation data

InCroMAP is a high-level data analysis tool. This means that no low-level data processing methods (normalization, smoothing, etc.) are included in InCroMAP. The reader is referred to the R statistical computing language to perform low-level data processing for microarrays. But this also means that InCroMAP supports data from various microarray manufacturers. To import microarray data, InCroMAP just needs any gene identifier and some statistics, i.e. fold changes or p-values, for each probe. Therefore, it is also possible to use data from other platforms than the mentioned four. Please see the FAQ for more information on this.

Besides developing powerful analysis methods and visualization tools, we have put a huge amount of work to make InCroMAP as easy-to-use as possible. Therefore, InCroMAP already includes many pre-processed datasets and mapping files. This also includes, for example, integrated microRNA targets from different databases for multiple organisms or mapping files to convert different IDs to gene symbols.

The detailed list of features is very long and every feature is described in it's own section in more detail. The following is a list of the main features of InCroMAP:

- Getting started:
    - Read processed microarray data from tabular text files
    - Manually import gene lists
    - Show microRNA targets for human, rat or mouse, choosing one or multiples of the following databases:
        * Experimentally verified

- · miRecords
- · miRTarBase
- · TarBase
    - ∗ Predicted
        - · ElMMo
        - · DIANA microT
        - · TargetScan
- – Visualize KEGG Pathways

- Single dataset analysis:
    - Gene set enrichment analysis:
        - ∗ KEGG Pathway enrichment
        - ∗ Gene Ontology (GO) enrichment
        - ∗ Enrichments using any gene set from the Molecular Signatures Database (MSigDB)
    - Annotate targets of miRNAs
    - Pathway-based visualization of any microarray data
    - Locus-based visualization of DNA methylation data

- Integrated, cross-platform analysis methods (using data from heterogeneous platforms):
    - Data pairing (e.g., showing miRNA data together with mRNA targets)
    - Gene-centered and expandable integrated tabular view of multiple platforms
    - Integrated gene set enrichment analysis (enrichment across multiple platforms, same databases supported as mentioned before)
    - Pathway based visualization of microarray data from four different platforms in one pathway

- Pathway visualization features:
    - Visualization of various KEGG pathways
    - Highlighting enriched genes from a pathway enrichment analysis directly in a visualized KEGG pathway
    - Searching/Highlighting of genes/compounds, etc.
    - Visualization of data directly in a pathway:
        - ∗ Coloring pathway nodes according to an observation (e.g., mRNA fold change)
        - ∗ Adding miRNA nodes with relevant targets to the pathway

* Coloring miRNA nodes

* Extending pathway-nodes to visualize expression data for various protein modifications

* Extending pathway-nodes to show the amount of differential methylation in gene promoters

* Showing corresponding observations/ expression values directly in the tooltip of pathway nodes

– Inspecting locus-based DNA methylation regions for specific pathway genes

• Application features:

– Easy-to-use and user-friendly

– All features are included in one graphical user interface

– Automatic detection of file formats and contents

– Customization of various options with the "Preferences" dialog

– Help available directly in the application

– Cross-linked pathway view (double-click nodes to get more info)

– Save and export everything (to multiple formats)

– Everything included (targets, mappings,...) and non-included data is downloaded automatically

– Multiple FDR correction methods included

– Application is available for download and also runs as Java WebStart

2 Installation

InCroMAP comes as a Java JAR file. It can run out-of-the-box on all systems where a Java virtual machine is installed and does not require any further installations.

2.1 Requirements

2.1.1 Software

InCroMAP is entirely written in Java™ and runs on any operating system where a suitable Java Virtual Machine (JDK version 1.6 or newer) is installed. See, for example, the Java SE download page[1].

2.1.2 Hardware

With at least 1 GB main memory, you should be able to perform most tasks without any problem. For large datasets, you should have at least 2 GB of main memory.
An active internet connection is required for most operations.

2.2 Starting the application

If you downloaded a ZIP-file, you need to unzip it before starting the application. Depending on your operating system, you should use the provided shell scripts for starting the application. This is `start.sh` for Linux or `start.bat` for Windows. On MAC OS, you have to create your own shortcut. You can start the application on all operating systems by typing

```
java -jar -Xms128m -Xmx1024m InCroMAP.jar
```

on your command prompt. Please note that you might have to change `InCroMAP.jar` for the real name of the JAR-file, e.g., `InCroMAP1.2.0.jar`. In this example, a minimum of 128 MB and a maximum of 1024 MB of memory will be available for the program. In most cases, InCroMAP needs more than 128 MB memory, so it might be convenient to create a shortcut and start the application with as much memory as available. If you have 2 GB RAM, for example, you might want to start the application with the following command:

```
java -Xms128m -Xmx1400M -jar InCroMAP.jar
```

---

[1]`http://www.oracle.com/technetwork/java/javase/downloads/index.html`

For your convenience, we already created several start-scripts to run the application with as much memory as possible. How much memory you actually need strongly depends on the size of your input datasets.

It is strongly recommended to start the program with at least 1 GB memory (`-Xmx1G`).

3 How to get started

3.1 Prepare microarray data

First of all, make sure that your microarray data is in one of the required file formats. The application can read character separated value (CSV) files, which are mostly tab-separated tabular files with microarray data. To use EXCEL-data, you can simply open your EXCEL spreadsheet, click "File" and "Save as" and select "Tab-separated text file". For nearly all data formats, the most important columns are your observations, which can be p-values or fold-changes, and one column with any gene identifier. Supported gene identifiers are gene symbols, NCBI Entrez Gene IDs, Ensembl gene IDs, RefSeq IDs or KEGG IDs. The fold-changes should be on a log scale (i.e., log_2 data) – negative values represent downregulation, positive values upregulation and 0 is the base line for no expression change.

Note: InCroMAP can NOT process raw data. Please use only processed microarray data.

3.1.1 mRNA expression data

As always, your file must be formatted as character (preferably tab) separated text file. Necessary columns are only one for the gene identifier and at least one column for your fold-changes or p-values.

Example:

Listing 3.1: Input file example for messenger RNA expression data

```
Gene.Symbol  EntrezGene.ID  Ctnnb1_foldchange  Ras_foldchange
Copg         54161          0.17               0.64
Atp6v0d1     11972          0.04               -0.24
...          ...            ...                ...
```

3.1.2 miRNA expression data

MicroRNA datasets are required to be character separated text files. Besides columns with fold-changes or p-values, just one column, containing a systematic miRNA identifier, is required. The systematic name can refer to either the mature, or the pre-miRNA.

Example:

Listing 3.2: Input file example for micro RNA expression data

```
Systematic_name Ctnnb1_foldchange Ras_foldchange
mmu-miR-96      1.45              7.21
mmu-let-7e      -0.72             0.46
...             ...               ...
```

3.1.3 DNA methylation data

InCroMAP can read any DNA methylation data with references to chromosomal locations. This can lead to huge datasets with up to single base pair resolution. To avoid having too much data, it is strongly recommended to summarize the data to bins of, e.g., $50bps$. Most microarray based DNA methylation datasets are already in a perfect format, because the probe length is mostly $50 - 75bps$.

All probes must somehow be mapped to genes. Therefore, in addition to the fold-change/p-value columns, at least one of the following three columns are required:

1. An "Identifier" column that contains a direct (custom) mapping from every probe to a gene

2. One "Chromosome" column and a "Probe position" column

3. One column, containing chromosome and position as probe identifier (format: `CHRchromo-someFSposition`, e.g., `CHR12FS1234`). This is also used by some companies (e.g., Nimblegen) as probe identifier.

In the later two cases, a dialog will appear that let's you choose how to map the data to the genome. Probes can either be mapped to gene bodies or promoter regions. In the later case, you can specify the region up-/ and downstream of a TSS.
Note: it is always recommended to have chromosome and position information with your DNA methylation data. Without these information, creating genome-based plots is not possible!
Example:

Listing 3.3: Input file example for DNA methylation data

```
Gene                probe_start probe_end ProbeID           Ctnnb1_foldchange
ENSMUSG00000056912  100054458   100054515 CHR10FS100054458  0.08
ENSMUSG00000056912  100053768   100053817 CHR10FS100053768  0.04
ENSMUSG00000046567  100052169   100052221 CHR10FS100052169  0.08
...                 ...         ...       ...               ...
```

3.1.4 Protein modification expression data

Protein expression datasets, as all other datasets, require a gene-identifier and columns with fold-changes or p-values. Furthermore, it is strongly recommended having a column, indicating the

modification of the current protein (phosphorylation, cleavage, acetylation, methylation, etc). This is customized for, e.g., phosphoprotein datasets. Nevertheless, the modification column is optional. Thus, one can simply read protein expression datasets by just specifying an identifier and the observation columns.

Example:

Listing 3.4: Input file example for protein modification expression data

```
Analyte_ID   Gene_Symbol Modification EntrezGene Ctnnb1_foldchange
AKT1         AKT1        basic        11651      0.24
AKT1_P-S473  AKT1        P-S473       11651      -0.34
PRKAA1       PRKAA1      basic        105787     0.07
...          ...         ...          ...        ...
```

The modification column indicates, if and how the protein is modified and at which site. For example, "P-S473" indicates a phosphorylation of the serine 473 site. The term "basic" refers to the unmodified protein. These modifications prevent InCroMAP from mixing different values of the same protein, that originate from different protein modifications (example input data may originate from reverse-phase-protein-arrays, e.g., from Zeptosense). But for simple protein expression data (without modification expression information), one can simply omit the modification column.

3.2 Open microarray data in the application

In general, all data must be processed expression data in tabular text files (see Section 3.1 for descriptions and examples of the input file formats). Then, there are many possibilities to open datasets in InCroMAP. For example, you can simply drag&drop data into the application or select "File", "Open" to open a file. In the upcoming dialogs, you will need to define which type of microarray data is contained (mRNA, miRNA, DNA methylation or protein) and which columns contain what information.

Figure 3.1 shows an example of the file input dialog. First, the species must be selected (on top of the dialog). If the input file format has not been inferred automatically, one may click on the black "CSV Options" label to specify further options (like "column separator char" or if the file contains headers - see Figure 3.2).

In the table at the lower half of this dialog, one must specify the content of each column. This can be done by clicking on the combo box below the captions. In some cases (e.g., when selecting "Identifier"), another combo box below the first one will become available and the content must be further specified (e.g., when selecting "Identifier", this can be either a "Entrez Gene ID", "Gene Symbol", "Ensembl Gene ID", etc. - see Figure 3.3 for an example).

Figure 3.1:   Example of the data import dialog. On the top, the species must be selected. The "CSV Options" panel might be expanded to correct auto-detected input file properties. At the table in the lower half, the content of each column must be specified. By clicking on the "Edit observation names" button, on can rename "Observation 1" to "Ctnnb1", as we did to create this example picture.

Figure 3.2:   Expanded "CSV Options" allow to give details for the input file. These properties are auto-detected and only need to be changed if the auto-detection failed to correctly infer those properties.

3.2.1 Importing observations from expression data

Section 3.1 gives examples for input files and specifies, what columns are required for each data type. Besides the data type specific columns, each dataset must have columns, containing your expression data as processed observations. Currently, these can be fold-changes or p-values. To import the fold-changes or p-values, an observation must be assigned to a column by picking "Observation {Number}" from the content combo box. Then, another combo-box will become available,

Figure 3.3: Example for specifying column contents. In this example, Entrez Gene IDs are given as the content of one column.

which let's the user choose if the content contains fold-changes or p-values. This step is demonstrated in Figure 3.4 and must be performed for all observations one might want to import. The term "Observation" here refers to any column that contains experimental results as fold-changes or p-values.

Figure 3.4: Importing processed observations from expression data into the application.

3.2.2 Renaming observations

The name "Observation" might be very unspecific for many datasets. Therefore, above the table on the import dialog is a big button "Edit observation names". This button allows renaming all "Observations" to, e.g., "DMSO_vs_KO" or any other more meaningful name. If the input file contained column headers and all observations are already assigned to columns on the input file dialog, InCroMAP will automatically suggest taking the column headers as observation names, as soon as the button "Edit observation names" is pressed. Figure 3.5 demonstrates how to rename observations.
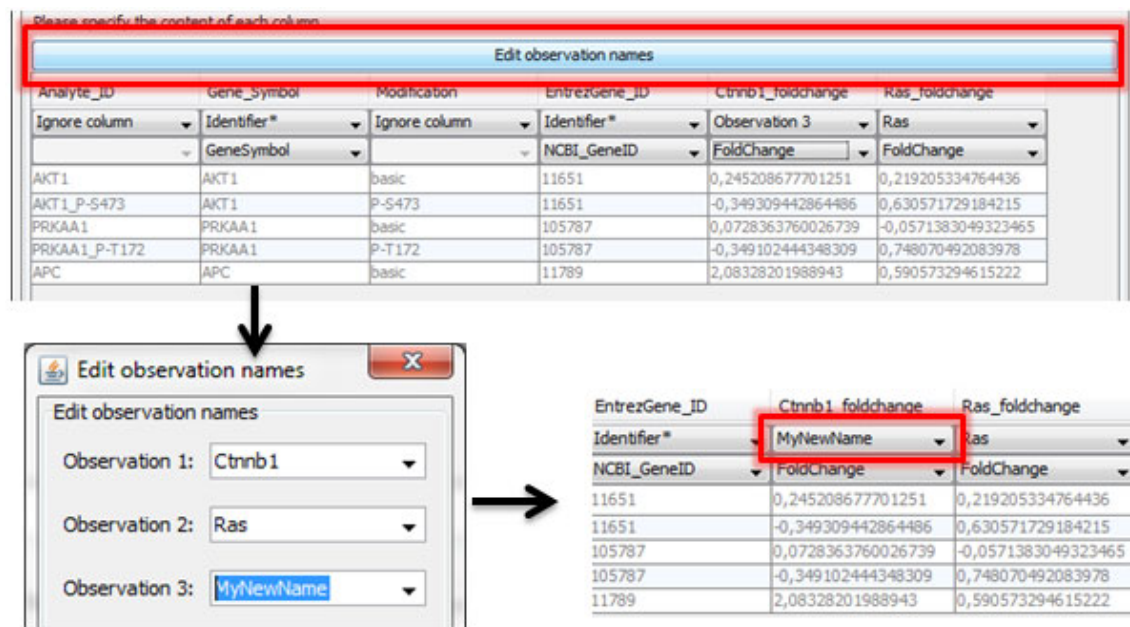
Figure 3.5: To rename observations, click the large "Edit observation names" button. In the upcoming dialog, you may enter new names for each observation. The new name will become effective immediately.

4 Single dataset analysis examples

This section describes some brief examples how to get started with single dataset analysis in In-CroMAP. As soon as you imported your data into the application, you'll see a table in a new tab, showing your data. On top of the tabs is a toolbar, showing some analysis functions. This toolbar always changes, depending on what is shown in the currently selected tab.

4.1 Enrichment analysis (every platform)

There are two ways to perform an enrichment analysis. One is to manually select lines in the table (corresponding to probes or genes) and right click on one of the selected lines. Then you can select an enrichment method (e.g., Gene Ontology or KEGG pathway) to perform this analysis.

The other way is to click the enrichment button in the Toolbar. Then, one needs to select an enrichment type to define a filter to select probes/genes for an enrichment. For example, one can pick a fold-change column, select "$|>=|$" (which is the same as "$\geq \pm$") and enter 1.0 in the last field. This would lead to an enrichment on all genes that are differentially expressed with a fold-change cutoff of 1.0 (i.e. "$fold - change \geq \pm 1.0$").

4.2 Pathway-based microarray data visualization (every platform)

InCroMAP offers multiple possibilities to visualize expression data in a pathway plot. One is, to select the tab containing the table with the input dataset and click on the "Visualize in pathway" button in the toolbar. In the upcoming dialog, one needs to pick the pathway of interest and an observation that contains the microarray data that should be visualized in the pathway. This will result in a picture, similar to Figure 4.1.

Another possibility is to somehow visualize a pathway (e.g., directly from a pathway enrichment analysis or by selecting "Import data" and "Show pathway" from the menu bar) and then click "Visualize data" and again "Visualize data" from the toolbar. See the top of Figure 4.1 for a screenshot of this option. Then, one needs to select the dataset that should be visualized in the pathway and the observation within this dataset. After clicking "Ok", the data will be visualized in the pathway.
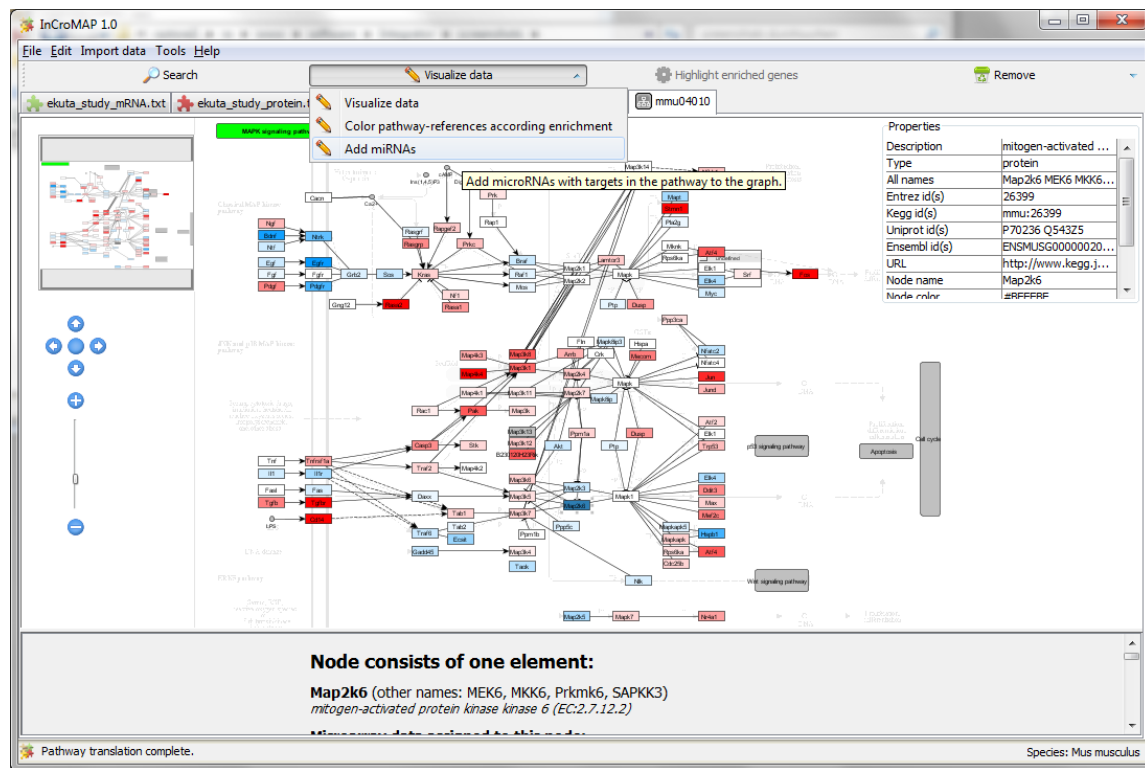
Figure 4.1:   Example of the "MAPK signaling pathway" with visualized mRNA fold-changes (encoded as node-colors). The "Map2k6" gene-node has been selected and the bottom shows a detail panel, containing detailed information about this gene and all associated expression data.

4.3 Show microRNA targets (miRNA only)

When a tab with microRNA data is currently visible, one can select "Targets" and "Annotate targets" from the toolbar. Then, in the upcoming dialog, one can choose which databases should be used to annotate micro RNA targets. There are currently three experimentally derived target databases (miRecords (see Xiao *et al.*, 2009), miRTarBase (see Hsu *et al.*, 2011) and TarBase (see Papadopoulos *et al.*, 2009)) and three databases with predictions (ElMMo (see Gaidatzis *et al.*, 2007), DIANA-microT (see Maragkakis *et al.*, 2011) and TargetScan (see Friedman *et al.*, 2009)) available.

Please note that for all predictions only targets that are above the suggested "high-confidence" threshold of the corresponding prediction database have been included in InCroMAP. Targets with low or medium prediction confidence have not been included at all.

Furthermore, many researchers follow the rule that "the more prediction methods suggest a target, the more reliable is this information". This rule has not been implemented into InCroMAP, because it is controversial. There are multiple reviews, suggesting that it is better to use one good

prediction algorithm, than combining multiples with an "AND". The three prediction methods offered by InCroMAP have been proven to be good prediction methods and also to perform good in combination. See Alexiou *et al.* (2009) for a verification of these statements and further information on micro RNA targets.

4.4 Visualize microRNA targets (every platform)

When a tab, containing a pathway plot (as shown, e.g., in Figure 4.1) is currently selected, users can click "Visualize data" and "Add miRNAs" to visualize micro RNAs as small rectangles with edges to their mRNA targets. This can be done, even if no miRNA expression data if available. For more information on the upcoming miRNA target database selection dialog, see Section 4.3.

4.5 Genome region plot (DNA methylation only)

Figure 4.2: XY-plot of DNA methylation data in the promoter region of the "Cyp2b10" gene.

DNA methylation data can be visualized as genome-region XY-plot. Please select a tab, containing DNA methylation data. There are now two possibilities to generate such a plot: The quick method is by selecting a probe (i.e., a line in the table), make a right click on it and select "Plot genome region". The application will then collect all probes that are associated with the same gene and create a genome region plot of them (see Figure 4.2 for an example).
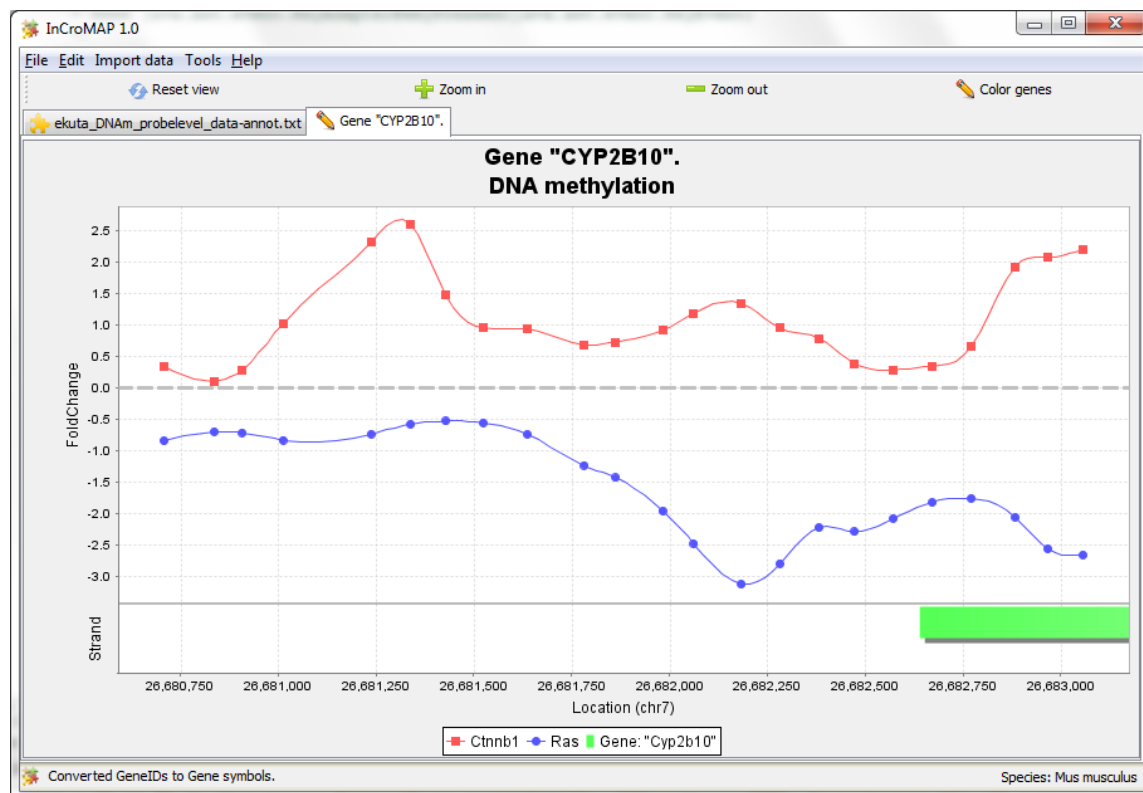
The other possibility is, to click the "Plot genome region" button in the toolbar. The upcoming dialog then let's you pick either a custom region by manually specifying chromosome, start and end coordinates. Or, after clicking "Plot a region, associated with a gene", you can select any gene and let InCroMAP plot all probes that are associated with this gene. At the bottom of this dialog, you can pick an observation from the current dataset that should be plotted. To include multiple observations in one plot (e.g., Figure 4.2 has two observations called "Ctnnb1" and "Ras"), one can pick "Include other observations with same signal type" (signal type refers to either p-values or fold-changes). If "Do not include other observations" is selected, only the selected observation will be plotted.

4.6 Metabolic overview (mRNA only)

Creating metabolic overviews is a good way to get started with novel mRNA expression datasets. This will create a plot that is similar to Figure 4.3. As this is an advanced analysis example, please make yourself familiar with the application before trying to follow the described steps.

From the application, select a tab that contains mRNA data. From this tab, select "Visualize in pathway". In the upcoming dialog pick "Metabolic pathways" from the pathway-selection combobox and any observation of interest from the combo-box below the pathway selection. After clicking "Ok", the application might need some time to fetch all required data from the KEGG-API. The result will now be a pathway, similar to Figure 4.3, in which all edges correspond to the mRNA expression level of enzymes. I.e., the colored edges express the mRNA fold-change (or p-value, depending on the selected observation) of the appropriate enzyme. You can click on a colored edge to get more information. Please note that in other pathways the mRNA expression levels are visualized as colored rectangles, whereas in the "Metabolic pathways"-pathway all mRNA expression levels are encoded in colored edges!

In this overview-pathway, it is also possible to visualize the importance of specific metabolic pathways. Please go back to your mRNA data tab and perform a KEGG pathway enrichment, as described, e.g., in Section 4.1. Now go back to the metabolic overview pathway again and select "Visualize data" and "Color pathway-references according enrichment" from the toolbar. Select the just performed KEGG pathway enrichment result in the appearing dialog and click "Ok". All nodes that are to references to other pathways (big rounded rectangles) now get colored according to the p-value in the selected pathway enrichment. I.e., the deeper blue they are, the more significant is the pathway (and the enrichment p-value). A white color is assigned to not significantly enriched pathways and grey means, that this pathway is not altered at all (did not appear in the enrichment). Figure 4.3 shows an example of the enrichment-based visualization of pathway-reference nodes.

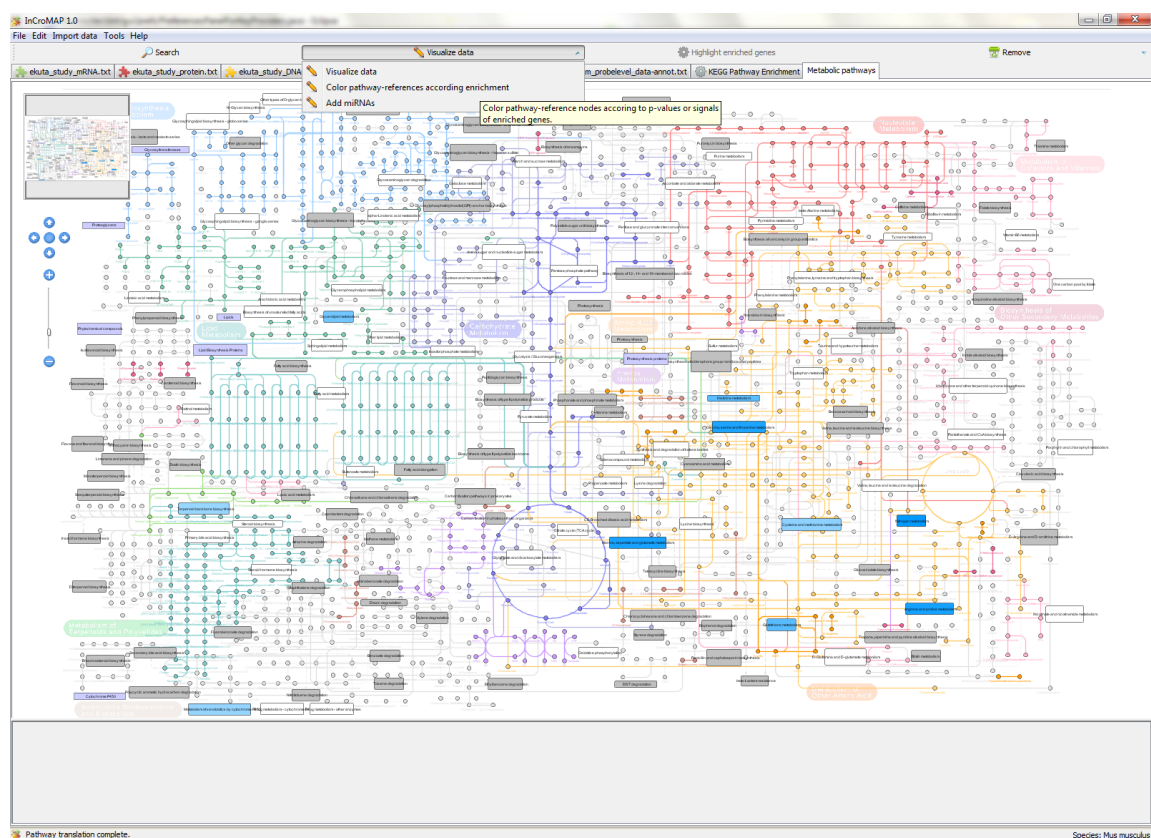Figure 4.3:   InCroMAP visualization of a metabolic overview pathway. All other referenced pathways (rectangle shaped nodes) are colored according to their enrichment p-value (i.e., how significant they are altered in input dataset). The darker-blue they are, the lower the p-value. White means $p - value \geq 0.05$ and grey means, that this pathway is not altered at all (not contained in pathway enrichment).

5 Integrated cross-platform microarray data analysis examples

This chapter describes analysis methods in the InCroMAP application that are available to users who have data from two or more different platforms for the same samples.

5.1 Integrated pathway-based visualization

Integrated pathway-based visualizations allow to visualize DNA methylation, mRNA expression, micro RNA expression and protein (modification) expression datasets in one pathway. See Figure 5.1 for an example. There are multiple ways to create such a visualization. You could just open any pathway and then visualize each of the datasets therein (see Section 4.2). Another easy way is to select "Tools" and "Integrated heterogeneous data visualization" from the menu bar. In the upcoming dialog, you first need to select the organism and the pathway of interest. For each data type is a checkbox in this dialog called, e.g., "Visualize mRNA data". Select the desired checkboxes and select a dataset and observation for each dataset (you need to have all datasets imported into InCroMAP, see Section 3.2). After clicking "Ok", a new tab with the pathway-based visualization of all your selected microarray datasets will show up.

5.2 Data pairing

Although this method makes especially sense for combinations with miRNA data, it can also be used for any other platform combination. Data pairing basically creates a huge table that puts one dataset in relation to another. We are going to explain this on some mRNA and miRNA example datasets. The result will be an integrated mRNA and miRNA analysis table, that contains microRNA fold-changes, mRNA target relations, and corresponding mRNA fold-changes (see Figure 5.2). Please select the tab, containing the tabular miRNA data. From the toolbar, select "Integrate" and "Pair data". In the upcoming dialog, you'll need to select the other dataset, which is in our example the mRNA dataset. Furthermore, you have the option to calculate a merged observation. This can be, e.g., the difference between miRNA and targeted mRNA expression fold-change or various other options. You can also unselect "Calculate a merged observation". After pressing "Ok", the paired table with two matched datasets will appear. If you haven't already annotated miRNA data with targets, the target database selection dialog (as described in Section 4.3) will show up before the final results. The data pairing with micro RNA datasets will
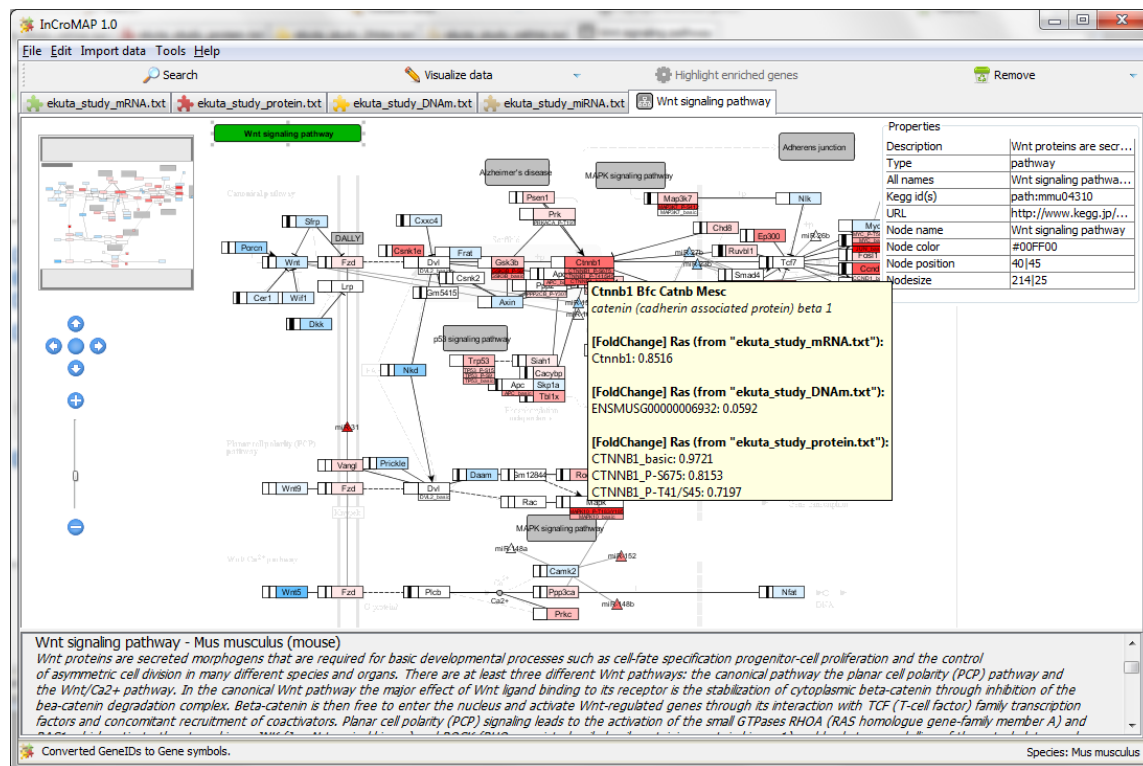
Figure 5.1: Pathway-based integrated cross-platform microarray visualization. This example shows the "Wnt signaling pathway". The rectangular nodes correspond to genes or gene families and their color indicates mRNA expression (blue means downregulated, red means upregulated). The box left of those nodes indicates promoter DNA methylation. The largest peak is shown and an empty box indicates no differential methylation. A bar to the left indicates hypomethylation and a bar, which goes from the middle of the box to the right indicates hypermethylation. Small rectangular boxes below the mRNA nodes indicate protein and protein modification expression. E.g., the "CTNNB1" gene has associated data from two different phosphorylation sites and the basic protein expression. Every of those boxes is red, which indicates upregulation. MicroRNAs are visualized as triangles with edges to their mRNA targets. MiRNA node color indicates miRNA expression.

lead to three additional columns. One relation column that simplifies the miRNA expression fold-change and target fold-change relation. E.g., "Up_Down" means that microRNA expression is upregulated and the target mRNA expression level is downregulated. The "Relation source" column gives the database(s), that contained or predicted this mRNA to be a target of the miRNA. In case of experimental databases, the PubMed identifier of the corresponding publication is given after the experimental database name. The third additional column is called "Score" and gives the score of the prediction method, if the target is a predicted target. Please refer to the correspond-

19

Figure 5.2: Example of the cross-platform "Data pairing" analysis method. Here: an integrated mRNA and microRNA analysis. On the left part of the figure, a micro RNA dataset is shown with expression p-values and fold-changes for each miRNA. In the middle, the target relationship for each miRNA is shown (source database, eventually prediction score, etc.). On the right side, the matching mRNA target is shown together with p-value and fold-change. The "Up_Down" column in the middle shows the relation of microRNA expression to mRNA expression (i.e., microRNA expression level is upregulated, mRNA expression level is downregulated).

ing miRNA target prediction databases itself for an interpretation of these scores. Please note that InCroMAP only shows targets that are above the recommended cutoff threshold for each miRNA target database. I.e., low confidence targets are not included in InCroMAP at all.

5.3 Integrated enrichment

Conventional enrichment analysis is just an interpretation of the current data type. For example, a pathway enrichment analysis on mRNA data will show pathways, that are altered in the mRNA dataset. Pathway enrichment analysis on DNA methylation data will show pathways that contain genes with significantly altered DNA methylation in their promoter regions.

With the integrated enrichment method in InCroMAP, the result of an enrichment analysis is not anymore just an interpretation of one platform - it's more close to what is really altered in the samples across multiple platforms. Therefore, the integrated enrichment is performed across multiple platforms and reflects alterations that occur on every (if datasets are connected with an AND) or any (if datasets are connected with an OR) platform.

To perform such an enrichment, please select "Tools" and "Integrated enrichment" from the menu bar. Now select your organism and all currently opened datasets from this organism should show up in the box below the organism selection. Pick all datasets that should be integrated into this enrichment (you can pick multiples by press and hold the CRTL-key of your keyboard) and select an enrichment type (Gene Ontology, KEGG pathway, etc.). At the bottom of this dialog, you can select if all the probes/ genes from the datasets should get connected with an "AND" or "OR" operator. For example, if you choose "AND", every gene must be altered in all selected datasets to get selected for the enrichment analysis. We recommend selecting "OR", because, e.g., an alteration in DNA methylation must not always also show an alteration in mRNA expression.

After confirming this dialog with "Ok", a threshold must be given for each dataset to define which genes should be picked for the enrichment analysis. Once all thresholds are defined, the result of the enrichment will show up.

5.4 Gene-based tabular integration of data from multiple platforms

Another method to get a cross-platform view of various samples is the expandable, gene-based tabular integration of heterogenous datasets. An example can be seen in Figure 5.3. At the root level, each row corresponds to one gene. In the columns, the associated values of various platforms are denoted. If one gene is expanded, second-level nodes show up that are named according to the platform (e.g., "mRNA"). If these nodes are expanded, the single probes associated with this gene appear. Thus, this integration method gives a possibility to quickly examine expression data, coming from various different platforms if one already knows some genes of interest.

To create such a table, select "Tools" and "Integrate heterogeneous data" from the menu bar. In the upcoming dialog, please select your organism and set a checkmark on every platform you want to integrate. Then, pick a dataset and an observation for each platform. Please note that this

Integrator 1.0

File   Edit   Import data   Tools

Search   Enrichment

Example_mRNA | Example_miRNA | ekuta_study_protein2.txt | mmu_Ctnnb1.plusEns.txt | IntegratedData

Gene	mRNA	miRNA	DNA methylation	Protein
EGFLAM	-0.11	N/A	N/A	N/A
EGFR	-0.944	0.3952	-0.1231	1.0616
mRNA	-0.944	N/A	N/A	N/A
Egfr (1424932_at)	-0.766	N/A	N/A	N/A
Egfr (1435888_at)	-0.944	N/A	N/A	N/A
Egfr (1451530_at)	-0.163	N/A	N/A	N/A
Egfr (1460420...)	-0.471	N/A	N/A	N/A
miRNA	N/A	0.3952	N/A	N/A
mmu-miR-30a (A_54_P2	N/A	0.3952	N/A	N/A
mmu-miR-30a (A_54_P2	N/A	0.1827	N/A	N/A
mmu-miR-30a (A_54_P4	N/A	0.1404	N/A	N/A
DNA methylation	N/A	N/A	-0.1231	N/A
Protein modification	N/A	N/A	N/A	1.0616
EGFR (EGFR)	N/A	N/A	N/A	0.3463
EGFR (EGFR_P-Y1045)	N/A	N/A	N/A	1.0616
EGFR (EGFR_P-Y1068)	N/A	N/A	N/A	0.1573
EGFR (EGFR_P-Y845)	N/A	N/A	N/A	-0.128
EGLN1	-0.215	N/A	-0.0948	N/A
mRNA	-0.215	N/A	N/A	N/A
DNA methylation	N/A	N/A	-0.0948	N/A
EGLN2	-0.114	N/A	-0.1102	N/A
EGLN3	0.678	N/A	-0.1308	N/A

Annotated 227/772 microRNAs with targets.   Species: Mus musculus

Figure 5.3: Example of an expandable, gene-based tabular view of microarray data from different platforms. Each top root node is one gene (e.g., the second row "EGFR") and each column shows the corresponding observation (in this example the fold-change) of each platform. If the node is expanded, novel nodes appear for each platform that contains data for this gene (e.g., the third row "mRNA"). If this is further expanded, the single values associated with this gene on the particular platform show up (in this example, all mRNA probes). For miRNA datasets, all miRNAs targeting this gene are shown.

is a gene-centric table. Thus, each dataset needs to be converted to contain gene-centric values. Therefore, you can select if you want to do this by calculating the mean, median, maximum, etc. in another combobox below the observation selection. You can pick separate options here for each platform. After confirming this dialog, the result table like shown, e.g., in Figure 5.3 will appear.

6 Help for different tabs and methods

6.1 Tabular tabs in general - searching, sorting and more

All input data from every platform is shown as a table. The first row is the header, which usually contains items like "#" (which is a simple number counting from one to infinity for every row) or "Name". You can click on any of those column headers to sort the table according to this column.

All tables in InCroMAP can be searched, by simply clicking on any row in the table (i.e., setting the focus on the table) and start typing. This will result in a text field to show up at the top left corner of the table. The table selection will automatically change to select an object that contains what you searched for. If you want to search for the next element, containing the same string, you need to press "F3" on your keyboard. Pressing "F3" once more will again search for the next hit, etc.

6.2 Enrichments and enrichment tabs

InCroMAP currently offers three different enrichments: KEGG pathway, Gene Ontology, and any enrichment using MSigDB files. For the first two, all required data will be obtained automatically. You just need to select genes for the enrichment (e.g., by setting a threshold for the fold-changes or p-values). The MSigDB enrichment will require you to specify a gene set file. You can download any gene set in GMT format from the molecular signatures database (MSigDB) at `http://www.broadinstitute.org/gsea/msigdb/` (see Subramanian *et al.* (2005)). On this homepage, please either download the "gene symbols" files or make sure that the "entrez genes ids" from the other files are valid for your current organism! For example, you can't perform an enrichment on a mouse dataset if the MSigDB file contains entrez gene ids for human. If you select "MSigDB enrichment" from within InCroMAP, the application will ask you to specify a path to any file, containing a gene set from the molecular signatures database as mentioned above.

6.2.1 How is the enrichment performed? And what are the columns?

InCroMAP first creates a gene-pool and then calculates p-values for every pre-defined gene set (e.g., pathway or GO-term) for this gene-pool. The results are presented to the user in a new table. The mentioned gene-pool is created by the user, typically by picking all significant genes ($p-value < 0.05$ or $abs(fold-change) > 1.5$). The gene sets are predefined by KEGG, Gene Ontology or the MSigDB. Now for every gene set, we investigate if it contains one or more genes from our gene-pool. The result is denoted in the "List ratio" column. This column denotes the
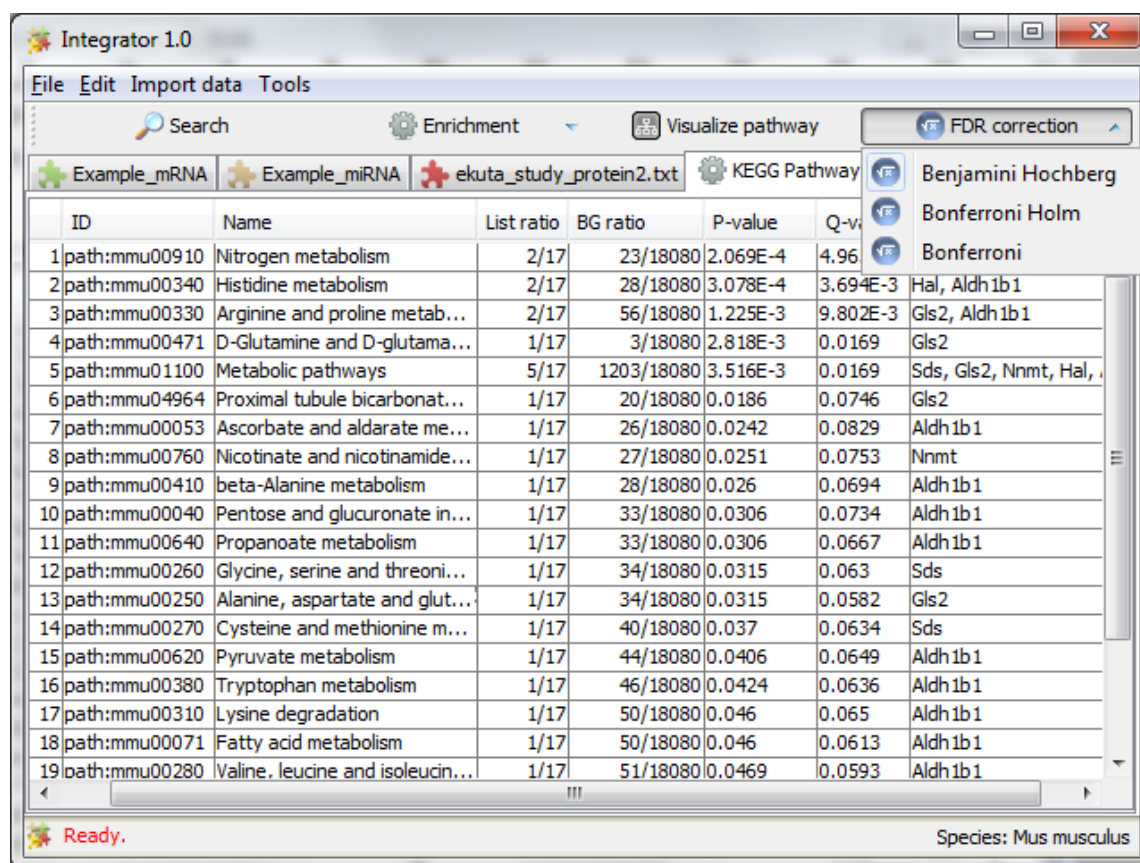
Figure 6.1: InCroMAP screenshot that shows the result of a KEGG pathway enrichment. The "FDR correction" menu shows different methods to face the problem of multiple testing.

number of genes from our gene-pool list that are in the gene set, described by the row, and the total gene-pool size. The "BG ratio" column denotes the total number of genes in the gene set and the total number of all genes in any gene set. From these values, the raw "p-value" is calculated using a hypergeometric test. The "q-value" column contains p-values, that have been corrected for multiple testing. By default, all p-values are being corrected using the "Benjamini & Hochberg" FDR correction method (see Benjamini and Hochberg (1995)). This can be changed to "Bonferroni" or "Bonferroni Holm" by clicking on the "FDR correction" button in the toolbar (see Figure 6.1). The last column denotes the actual genes from the gene-pool that are in the current gene set.

Example: we have created a KEGG pathway enrichment with some p-value cutoff on an mRNA dataset. After applying the threshold, we had a total of 17 genes in our gene-pool. The results of the enrichment are depicted in Figure 6.1. The third row now denotes an enriched pathway (i.e., gene set) called "Arginine and proline metabolism". Two of our 17 genes in the gene-pool are contained in this pathway. The pathway has a total of 56 genes and the total number of genes occurring in any

pathway is 18080. From these values, the p-value of 0.001225 is calculated with a hypergeometric test and, using Benjamini & Hochberg, the corrected p-value is 0.009802. The two genes from our gene-pool that are in the pathway are "Gls2" and "Aldh1b1", what is denoted in the last column.

For micro RNA datasets, the enrichment is a little bit more difficult. The gene sets behind these enrichment are mostly referring to protein coding genes. Thus, they are not directly usable for micro RNAs. Hence, the enrichment on miRNA datasets is performed on the targets of the corresponding miRNAs. So first, you need to annotate your miRNA dataset with targets, using the "Targets" and "Annotate targets" button on the toolbar. Then, as with every other dataset you can define a threshold for the p-values or fold-changes. This will result in a microRNA pool that meets the selected cutoff. This microRNA pool is now converted to a gene-pool by using all annotated targets from the microRNAs in the pool. The rest of the enrichment is the same procedure as described above. Except for the list ratio, which denotes the number of microRNAs that contain targets in this gene set (i.e., pathway) and the total number of microRNAs in the gene pool. In the preferences, you can change the list ratio when performing a microRNA enrichment to reflect the number of targets in this gene set, instead of the number of microRNAs that contain targets in the gene set.

6.3 Pathways and the pathway visualization tabs

The InCroMAP pathways are graphs that are layouted on top of an original KEGG pathway picture. In general, small circular nodes are small molecules, rectangular nodes with rounded corners describe referenced pathways (also called "pathway-reference nodes"), and rectangular-shaped nodes (smaller ones, without rounded corners) correspond to products of genes or gene families. Please note that KEGG sometimes puts multiple genes into a single node and if it's easier to visualize, a node can have multiple instances in a graph. Sometimes (e.g., in the "Metabolic pathways"-pathway), the proteins or enzymes are shown as big edges, connecting compounds. These are rare special cases that KEGG uses to enhance the quality of the overall picture. In these cases, only mRNA data can be visualized (resulting in a color-change of this big edge).

In every pathway, you can double click on a pathway-reference node to open the corresponding pathway in InCroMAP or you can click on any other node or edge to get further information about it. Furthermore, you can search for certain genes or other strings in a pathway, by selecting "Search" from the toolbar. This will change the font and border color of all nodes matching your search to red.

6.3.1 Visualize expression data in a pathway

When a pathway tab is currently visible, any opened expression dataset can be visualized in the pathway by selecting "Visualize data" and again "Visualize data" from the toolbar. Only one dataset per type (i.e., mRNA, miRNA, protein, or DNA methylation) can be visualized at a time, but all types can be visualized at once. In the following part we are referring to the default properties. Many options (like the color or threshold for upregulation) can be changed in the preferences!

Examples for pathway-based
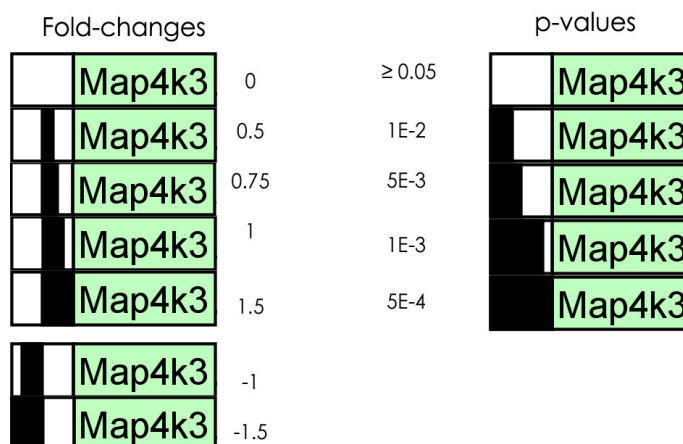## visualization of DNA methylation data

Figure 6.2: Examples of various fold-changes or p-values for DNA methylation data and their visualization in InCroMAP. These examples are only valid for the default settings. For example, the p-value required to fill the whole box can be changed in the preferences. Usually, the value visualized for a gene corresponds to the maximum peak in the gene promoter - but this can also be changed in the preferences.

For all data types, except DNA methylation, if fold-changes are visualized, the color red means upregulation and blue means downregulation. The darker the color, the higher/lower is the fold-change. Default requirement for a node to get the darkest red/blue color is a fold-change of ± 1.5. White color indicates no differential expression (by default, fold-changes between ± 0.5) and grey color means, that no value from the input dataset could be linked to this node (e.g., gene was not on microarray, removed by preprocessing, etc.). These properties are similar when visualizing p-values. A white color, by default, means that the p-value is > 0.05 and grey means, that no data is available. Significance is indicated by the darkness of the blue color on a log-scale, where 0.0005 is the threshold for the darkest blue color. All these properties (thresholds, colors, etc.) can be changed in the preferences.

To visualize data in a pathway, only one value for a node in the pathway can be visualized. These "nodes" are mostly genes, but sometimes also gene families (e.g., in the "MAPK signalling pathway", "MAPK1" and "MAPK3" are summarized in one node). Thus, InCroMAP sometimes needs to merge multiple probes and values to a single value. In the preferences, you might want to remove the checkmark on the "Remember gene center decision" box. If this is unchecked, InCroMAP will ask every time how the input data should be gene-centered. Therefore, several options are available: Mean, Median, Minimum, Maximum, Maximum distance to zero (i.e., max $|x|$),

NormalizedSumOfLog2Values, or Automatic. Automatic will take the most significantly probe (p-values) or the probe with maximum differential expression (fold-changes). Please see the chapter "Preferences" for more information on these options.
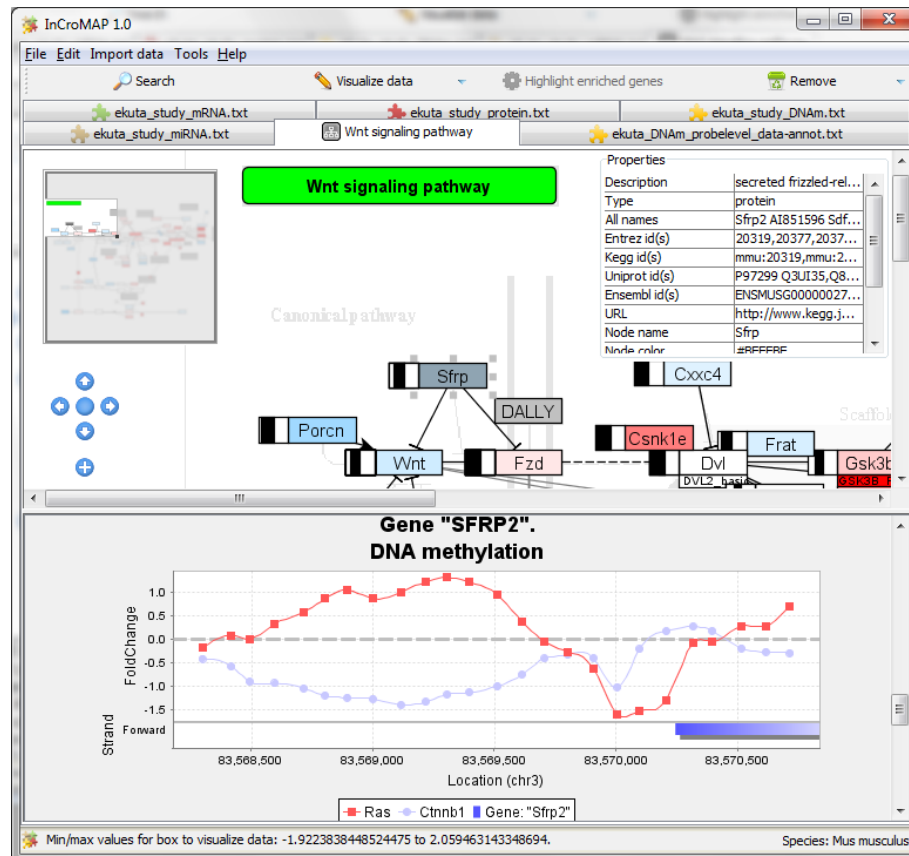

Figure 6.3: Example of an integrated pathway-based visualization. Here, three different data types (mRNA, protein and DNA methylation) have been visualized. The node "SFRP" is selected and the bottom half shows a detailed view of all expression data associated with this node. The DNA methylation data is plotted in detail as genome region plot of this gene.

Messenger RNA is visualized by changing the color of the rectangle-shaped nodes, according to mRNA expression. MicroRNA is visualized, by changing the color of the triangle-shaped miRNA nodes, according to miRNA expression. Protein data is visualized, by adding small boxes below the rectangle-shaped nodes. DNA methylation data is visualized, by putting black boxes left of the rectangle-shaped nodes. For DNA methylation, the same thresholds as for all other data types are valid. But the expression is not encoded in a color, but in a black bar that goes from the middle to the left, if the gene is hypomethylated and to the right, if the gene is hypermethylated. For p-values, a black bar goes from the left border of this box to the right one, depending on the significance. See

Figure 6.2 for examples and more details on the pathway-based visualization of DNA methylation data. With the default settings, the probe with maximum differential expression or maximum significance in the promoter determines the value for DNA methylation visualization of this gene. This is very similar to maximum peak detection approaches! But still, DNA methylation data is not very informative if it is summarized to a single value per gene and thus, InCroMAP includes a detail panel that contains a plot of the DNA methylation data in the genomic and promoter region of the selected gene. The black bar, added to the pathway, should show in a very simplified view which genes are interesting and important. For those genes then, the user can click on their nodes and in the lower half of the application a new frame is shown with detailed information about the nodes and all expression values associated with them. Please see Figure 6.3 for an example of the detailed DNA methylation visualization.

6.3.2 Other pathway-based visualization options

If you started by performing a KEGG pathway enrichment on your expression data and then selected a pathway from the enrichment and clicked "Visualize pathway", then the "Highlight enriched genes" option will become available in the toolbar. If this option is selected, each node in the pathway is colored grey, except for nodes that correspond to genes that were enriched in this pathway. These are highlighted in red (which is only a highlighting and does not express up- or downregulation). For example, Figure 6.1 shows the result of an enrichment. If we would select the third row ("Arginine and proline metabolism") and click on "Visualize pathway", then the "GLS2" and "ALDH1B1" genes would be highlighted in red.

Another visualization possibility (when a KEGG pathway enrichment has been performed) is coloring pathway-reference nodes according to the enrichment p-value. This can be done by selecting "Visualize data" and "Color pathway-references according enrichment" from the toolbar. This will change the color of all pathway-reference nodes (rectangular nodes with rounded corners) according to their p-value in the enrichment. The color scheme and requirements are the same as described in Section 6.3.1. An example of such colored reference nodes can be seen in Figure 4.3.

6.4 DNA methylation plots

InCroMAP can create genome-region, XY-plots of DNA methylation data. The DNA methylation data is plotted on the upper part of the figure and all genes in this genome region on the bottom. See Figure 4.2 for an example. Section 4.5 describes in detail how to create such a plot and how to select which observations should be contained in the plot.

Hint: To get the full name of a depicted gene, you can hover the gene with your mouse cursor and a tooltip with the full gene name will appear.

6.4.1 Integrated mRNA expression and DNA methylation plots

When a tab, containing only the XY-plot of DNA methylation data, is selected, InCroMAP can change the color of the genes at the bottom of this graph according to another expression dataset. Therefore, you can click on the "Color genes" button in the toolbar and select an expression dataset. This is very similar to changing the color of the pathway nodes. For example, you can create a DNA methylation plot of a custom genome region and select "Color genes" from the toolbar. Then, you have to select the dataset and observation you wish to use for coloring the gene (e.g., an mRNA dataset). InCroMAP will then change the color of all the genes, depicted at the bottom of this graph, according to the selected dataset. The colors and thresholds for this procedure are the same as for the pathway-visualization (see Section 6.3.1).

7 Preferences

There are many preferences that you can change to influence the analysis results and generic behavior of InCroMAP. Please note that any changes made to the preferences are only valid for new tabs/ analyses. E.g., if you add a navigation panel to the pathway tabs, it will not show up on the existing, but on all novel tabs.

InCroMAP options

These options influence generic InCroMAP properties, like how to merge multiple probes or how to visualize data in a pathway.

Gene center signals by For many operations, all signals (i.e., observations, fold-changes, etc.) that belong to one or multiple genes must be summarized to one value. For example, when visualizing data in a pathway, one value must be calculated that corresponds to one node. The following options are available for this calculation: Mean, Median, Minimum, Maximum, Maximum distance to zero (i.e., $\max |x|$), NormalizedSumOfLog2Values (i.e., $\frac{\sum_{i=1}^{n} \log_2 x}{n}$, useful for p-values), or Automatic. Automatic will take the minimum for p-values, "maximum distance to zero" for fold-changes, and mean for all others. This corresponds to taking the most significant probe (p-values) or the probe with maximum differential expression (fold-changes).

Remember gene center decision If this is unchecked, InCroMAP will ask every time how to merge multiple signals (i.e., the "Gene center signals by" option). If it is checked, the current value will be taken as default value without asking each time.

Count miRNA targets for list ratio InCroMAP requires protein coding genes for enrichments. Hence, for miRNA datasets, InCroMAP will take the miRNA targets of the selected miRNAs to perform the enrichment. The p-value of the enrichment depends on the list ratio (all terms explained in detail in Section 6.2.1). By default, InCroMAP will take the number of microRNAs that have targets in the current pathway and the total number of microRNAs in the gene-pool as list ratio. By selecting this option, InCroMAP will count the number of targets to calculate the list ratio, rather than the microRNAs themselves. I.e., number of targets of any microRNA from the gene-pool that occur in the current pathway and total number of all targets from all microRNAs in the gene-pool.

Fold change for maximum color This is the fold-change that is required for a gene to get the color that is defined some options below. All fold-changes closer to zero get brighter colors assigned.

Don't visualize fold changes between $\pm$ x All fold-changes between $\pm$ this value will get the "Color for no fold change".

P-value threshold for maximum color This is the p-value that is required to get the full "Color for minimum fold change" assigned. All greater p-values will get brighter colors.

Don't visualize p-values $>$ x All p-values greater than this value will get the "Color for no fold change".

Color for X Define various colors (the "Color for no value" is assigned to all nodes that can not be mapped on your input dataset).

DNA methylation maximum box width Define the width of the black boxes (in pixels) that appear when visualizing DNA methylation data in a pathway.

Protein modification box height Define the height of the boxes (in pixels) that appear below the nodes when visualizing protein data in a pathway.

KEGGtranslator options

These options influence the creation of the pathways from the source database.

Remove orphans If true, remove all nodes that have no edges before translating the pathway.

Label genes by For one KEGG object, multiple names are available. Choose how to assign one name to this object.

Show formula for compounds If true, shows the chemical formula for all compounds, instead of the name.

Remove white gene nodes If true, removes all white gene-nodes in the KEGG document. KEGG usually draws all enzymes that occur in homologous, but NOT the current organism in white.

Remove pathway references If true, removes all pathway-reference nodes.

Merge nodes with same edges If true, merges all nodes that have exactly the same relations (sources, targets and types).

Create edge labels If this is selected, describing labels for each edge in the graph are created.

Hide labels for compounds This option allows to hide labels for all compounds (=small molecules).

Draw grey arrows for reactions Some metabolic KEGG pathways contain only reactions and no relations (typically, only relations are visualized). If this option is selected, grey arrows for all reactions specified in a KEGG pathway are drawn in addition to the relations.

Graph options

These options influence the look and additional available panels for all pathway tabs.

Show logo in graph background If true, shows the InCroMAP logo in the background of each graph.

Show KEGG picture in graph background Allows to show the original KEGG picture in the background layer of a pathway.

Brighten KEGG background image Set a percentage for brightening the KEGG background image (see previous option).

Show navigation and overview panels Puts a navigation and overview panel on the left side of each pathway.

Show properties table Shows a properties table on the upper right corner of each pathway that contains additional information for each selected node or edge.

Layout edges If this is checked, an organic edge routing algorithm will be executed for every pathway.

8 FAQ / Troubleshooting

Where can I get help for a certain component/ option/ checkbox/ etc.?
Most elements in InCroMAP have tooltips. If you don't understand an option, you can get help in the first place by just pointing the mouse cursor over it and wait for the tooltip to show up ($\sim$ 3 seconds).

I'm getting a "java.lang.OutOfMemoryError: Java heap space"
Some operations need a lot of memory. If you simply start InCroMAP, without any JVM parameters, only 64 MB of memory are available. Please append the argument `-Xmx1024M` to start the application with 1 GB of main memory. See Section 2.2 for a more detailed description of how to start the application with additional memory. If possible, you should give the application 2 GB of main memory. A minimum of 1 GB main memory should be available to the application.

Is an internet connection required to run InCroMAP?
An internet connection is required for most operations. Many identifier mapping files and pathway-based visualizations require an active internet connection. However, if you import your data directly with NCBI Entrez Gene IDs and do not use the pathway-visualization or GO-enrichment, you should be able to run the application offline.

Which organisms are supported?
Currently mouse, human and rat are supported.

Can I import other microarray data types than the mentioned four?
Yes! Basically, everything that can be mapped to a gene can be imported. It is recommended to pick the mRNA data type (even if you don't have mRNA data), select any gene identifier and import your data as processed p-values or fold changes. With this method, you can perform integrated analysis, enrichment analysis and pathway-based visualization of basically every microarray type.

Where can I get the latest version?
Go to `http://www.cogsys.cs.uni-tuebingen.de/software/Integrator/`.

Is one node in a pathway always one gene?
No! KEGG often creates one node that groups multiple genes together. These can be either gene families (e.g. RRM1 and RRM2 in the pyrimidine metabolism) or functionally related genes. You should, in doubt, click on a node to see which genes are contained – this will also give you all

expression values, associated to all genes that are contained in the node.

How can I save or export a picture/ table/ etc.?
You can save almost every table as tab-separated text file and every picture as JPEG or PNG file. Simply select "File" and "Save as" from the menu bar.

Where can I get more information on the pathway translations?
Please see Wrzodek *et al.* (2011).

Error message while visualizing data in the pathway: no node matches the input data?
There are mainly two possible reasons for this error message. First, there really is no gene in the pathway contained in your source dataset. Second, it often occurs that people accidently mix organisms. If you show, for example, the "MAPK signalling pathway" for *Homo sapiens* and try to visualize a mouse dataset, you will also get this error. Please check on the pathway and on your source dataset that you picked the correct organism (it is shown for every tab at the bottom left corner of InCroMAP).

No targets could be annotated to my miRNA data?
This often occurs when you selected the wrong organism during import of your data. Please double-check that you selected the correct organism in the open-file dialog.

What does "Observation" or "Signal" mean?
The terms "Observation" or also "Signal" always refer to fold-changes or p-values. For example, if you have an experiment with some wildtype mouse *WT* that has a spontaneous tumor and a somehow treated mouse *TM* that also has a tumor. Then, in this example the mRNA expression can be measured for *WT healthy*, *WT tumor*, *MT healthy* and *MT tumor*. You could now make several observations from this, e.g., "*WT healthy vs. WT tumor*" or "*WT healthy vs. MT healthy*", etc. Those observations are usually expressed as p-values or fold-changes and they are meant when talking of observations in InCroMAP.

Bibliography

Alexiou, P., Maragkakis, M., Papadopoulos, G. L., Reczko, M., and Hatzigeorgiou, A. G. (2009). Lost in translation: an assessment and perspective for computational microrna target identification. *Bioinformatics*, **25**(23), 3049–3055.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**(1), 289–300.

Friedman, R. C., Farh, K. K.-H., Burge, C. B., and Bartel, D. P. (2009). Most mammalian mrnas are conserved targets of micrornas. *Genome Res*, **19**(1), 92–105.

Gaidatzis, D., van Nimwegen, E., Hausser, J., and Zavolan, M. (2007). Inference of mirna targets using evolutionary conservation and pathway analysis. *BMC Bioinformatics*, **8**, 69.

Hsu, S. D., Lin, F. M., Wu, W. Y., Liang, C., Huang, W. C., Chan, W. L., Tsai, W. T., Chen, G. Z., Lee, C. J., Chiu, C. M., Chien, C. H., Wu, M. C., Huang, C. Y., Tsou, A. P., and Huang, H. D. (2011). miRTarBase: a database curates experimentally validated microRNA-target interactions. *Nucleic Acids Res*, **39**(Database issue), D163–9. Hsu, Sheng-Da Lin, Feng-Mao Wu, Wei-Yun Liang, Chao Huang, Wei-Chih Chan, Wen-Ling Tsai, Wen-Ting Chen, Goun-Zhou Lee, Chia-Jung Chiu, Chih-Min Chien, Chia-Hung Wu, Ming-Chia Huang, Chi-Ying Tsou, Ann-Ping Huang, Hsien-Da England Nucleic Acids Res. 2011 Jan;39(Database issue):D163-9. Epub 2010 Nov 10.

Maragkakis, M., Vergoulis, T., Alexiou, P., Reczko, M., Plomaritou, K., Gousis, M., Kourtis, K., Koziris, N., Dalamagas, T., and Hatzigeorgiou, A. G. (2011). DIANA-microT Web server upgrade supports Fly and Worm miRNA target prediction and bibliographic miRNA to disease association. *Nucleic Acids Res*, **39**(Web Server issue), W145–8. Maragkakis, Manolis Vergoulis, Thanasis Alexiou, Panagiotis Reczko, Martin Plomaritou, Kyriaki Gousis, Mixail Kourtis, Kornilios Koziris, Nectarios Dalamagas, Theodore Hatzigeorgiou, Artemis G England Nucleic Acids Res. 2011 Jul;39(Web Server issue):W145-8. Epub 2011 May 6.

Papadopoulos, G. L., Reczko, M., Simossis, V. A., Sethupathy, P., and Hatzigeorgiou, A. G. (2009). The database of experimentally supported targets: a functional update of TarBase. *Nucleic Acids Res*, **37**(Database issue), D155–8. Papadopoulos, Giorgos L Reczko, Martin Simossis, Victor A Sethupathy, Praveen Hatzigeorgiou, Artemis G England Nucleic Acids Res. 2009 Jan;37(Database issue):D155-8. Epub 2008 Oct 27.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, **102**(43), 15545–15550.

Wrzodek, C., Drger, A., and Zell, A. (2011). Keggtranslator: visualizing and converting the kegg pathway database to various formats. *Bioinformatics*, **27**(16), 2314–2315.

Xiao, F., Zuo, Z., Cai, G., Kang, S., Gao, X., and Li, T. (2009). miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Res*, **37**(Database issue), D105–10. Xiao, Feifei Zuo, Zhixiang Cai, Guoshuai Kang, Shuli Gao, Xiaolian Li, Tongbin 1R21CA126209/CA/NCI NIH HHS/ R43 GM076941/GM/NIGMS NIH HHS/ England Nucleic Acids Res. 2009 Jan;37(Database issue):D105-10. Epub 2008 Nov 7.