Documentation for Integrator

integrated analysis of microarray data from different platforms

Clemens Wrzodek

March 7, 2012

TODO: Please write the abstract.

ModuleMaster is a novel application for finding *cis*-regulatory modules (CRMs) in sets of co-expressed genes. The application comes with a newly developed method, which does not only consider transcription factor binding information but also multivariate functional relationships between regulators and target genes to improve the detection of CRMs. Given only the results of a microarray and subsequent clustering experiment, the program includes all necessary data and algorithms to perform every step to the final CRMs. This workbench possesses an easy-to-use graphical user interface, together with job processing and a plethora of command line options, making ModuleMaster a fully-featured program for large scale batch processing. The detected CRMs can be visualized and evaluated in various ways: i.e., generating GraphML- and R-based whole regulatory network visualizations or generating SBML files for subsequent analytical processing and dynamic modeling.

Contents

1 Introduction

Integrator 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. Integrator 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

Integrator is a high-level data analysis tool. This means, that no low-level data processing methods (normalization, smoothing, etc.) are included in Integrator. The reader is referred to the R statistical computing language to perform low-level data processing for microarrays. But this also means that Integrator supports data from various microarray manufacturers. To import microarray data, Integrator 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 Integrator as easy-to-use as possible. Therefore, Integrator 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 Integrator:

- 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:
    - Experimenttally 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 analyis (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 accorings 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 content
  - Customization of various options via preferences
  - 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

Integrator 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

Integrator 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 Integrator.jar
```

on your command prompt. 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, Integrator 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 Integrator.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 1. Prepare microarray data

First of all, make sure that your microarray data is in one of the required 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 ID, Ensembl gene ID, RefSeq ID or KEGG ID.

Note: Integrator 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

Integrator 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: ”CHRchromosomeFSposition”, 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
...          ...          ...           ...         ...
```

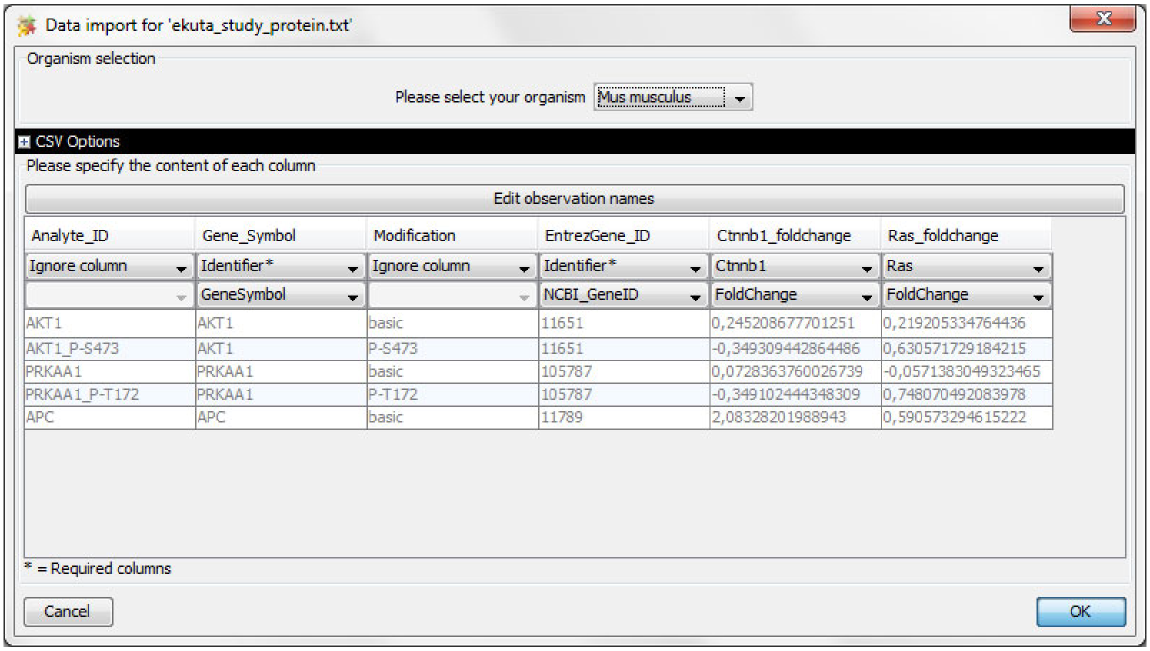
The modification column indicates, if and how the protein is modified and at which site. For example, ”P-S473” indicates a phosphorylation of the serine 473 site. The term ”basic” refers to the unmodified protein. These modifications prevent InCroMAP from mixing different values of the same protein, but coming from different protein modifications (example input data may come 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 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 example 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).



Data import for 'ekuta_study_protein.txt'

Organism selection

Please select your organism: Mus musculus

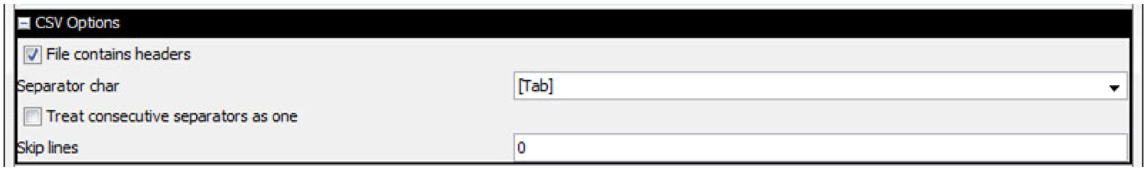
CSV Options

Please specify the content of each column

Analyte_ID	Gene_Symbol	Modification	EntrezGene_ID	Ctnnb1_foldchange	Ras_foldchange
Ignore column	Identifier*	Ignore column	Identifier*	Ctnnb1	Ras
	GeneSymbol		NCBI_GeneID	FoldChange	FoldChange
AKT1	AKT1	basic	11651	0,245208677701251	0,219205334764436
AKT1_P-S473	AKT1	P-S473	11651	-0,349309442864486	0,630571729184215
PRKAA1	PRKAA1	basic	105787	0,0728363760026739	-0,0571383049323465
PRKAA1_P-T172	PRKAA1	P-T172	105787	-0,349102444348309	0,748070492083978
APC	APC	basic	11789	2,08328201988943	0,590573294615222

* = Required columns

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.



CSV Options

☑ File contains headers

Separator char: [Tab]

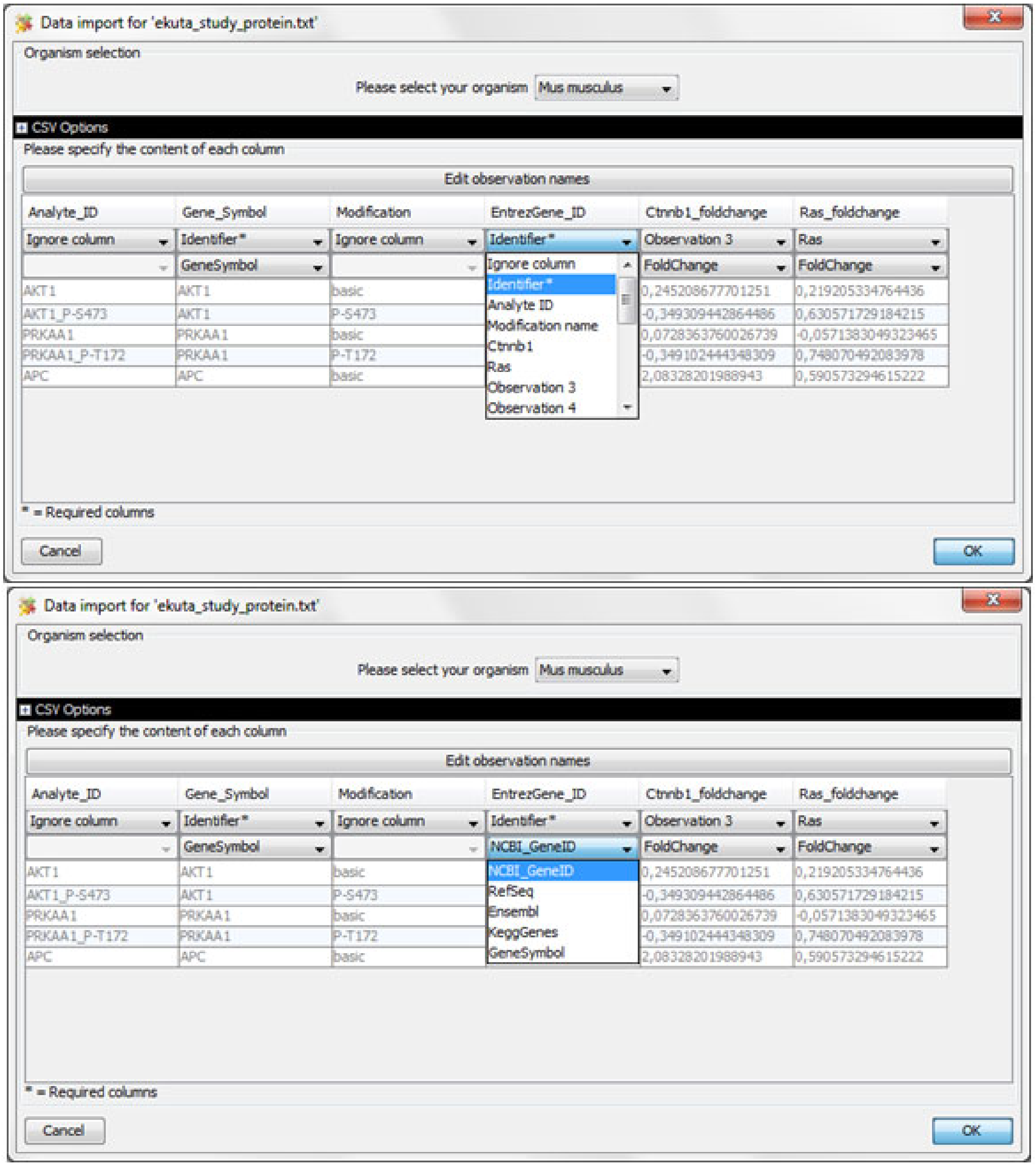
☐ Treat consecutive separators as one

Skip lines: 0

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, 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.



Data import for 'ekuta_study_protein.txt'

Organism selection

Please select your organism: Mus musculus

CSV Options

Please specify the content of each column

Edit observation names

Analyte_ID	Gene_Symbol	Modification	EntrezGene_ID	Ctnnb1_foldchange	Ras_foldchange
Ignore column	Identifier*	Ignore column	Identifier*	Observation 3	Ras
	GeneSymbol			FoldChange	FoldChange
AKT1	AKT1	basic		0,245208677701251	0,219205334764436
AKT1_P-S473	AKT1	P-S473		-0,349309442864486	0,630571729184215
PRKAA1	PRKAA1	basic		0,0728363760026739	-0,0571383049323465
PRKAA1_P-T172	PRKAA1	P-T172		-0,349102444348309	0,748070492083978
APC	APC	basic		2,08328201988943	0,590573294615222

* = Required columns

Cancel

OK

Data import for 'ekuta_study_protein.txt'

Organism selection

Please select your organism: Mus musculus

CSV Options

Please specify the content of each column

Edit observation names

Analyte_ID	Gene_Symbol	Modification	EntrezGene_ID	Ctnnb1_foldchange	Ras_foldchange
Ignore column	Identifier*	Ignore column	Identifier*	Observation 3	Ras
	GeneSymbol		NCBI_GeneID	FoldChange	FoldChange
AKT1	AKT1	basic		0,245208677701251	0,219205334764436
AKT1_P-S473	AKT1	P-S473		-0,349309442864486	0,630571729184215
PRKAA1	PRKAA1	basic		0,0728363760026739	-0,0571383049323465
PRKAA1_P-T172	PRKAA1	P-T172		-0,349102444348309	0,748070492083978
APC	APC	basic		2,08328201988943	0,590573294615222

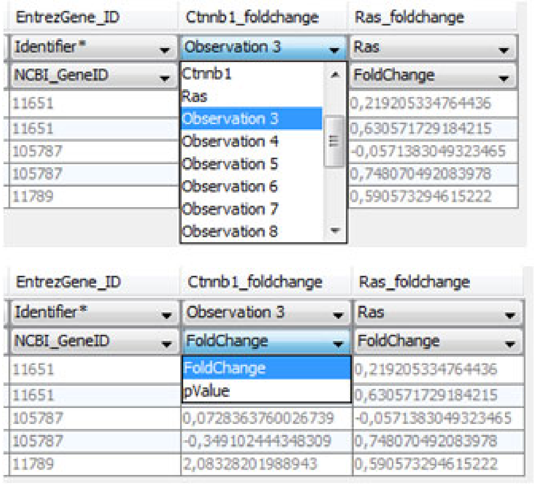
* = Required columns

Cancel

OK

Figure 3.3: Example for specifying column contents. In this example, entrez gene ids are given as the content of one column.

The term ”Observation” refers to any column content, for which fold-changes or p-values can be calculated.



EntrezGene_ID	Ctnnb1_foldchange	Ras_foldchange
Identifier*	Observation 3	Ras
NCBI_GeneID		FoldChange
11651		0,219205334764436
11651		0,630571729184215
105787		-0,0571383049323465
105787		0,748070492083978
11789		0,590573294615222

Import expression data by assigning an observation to this column

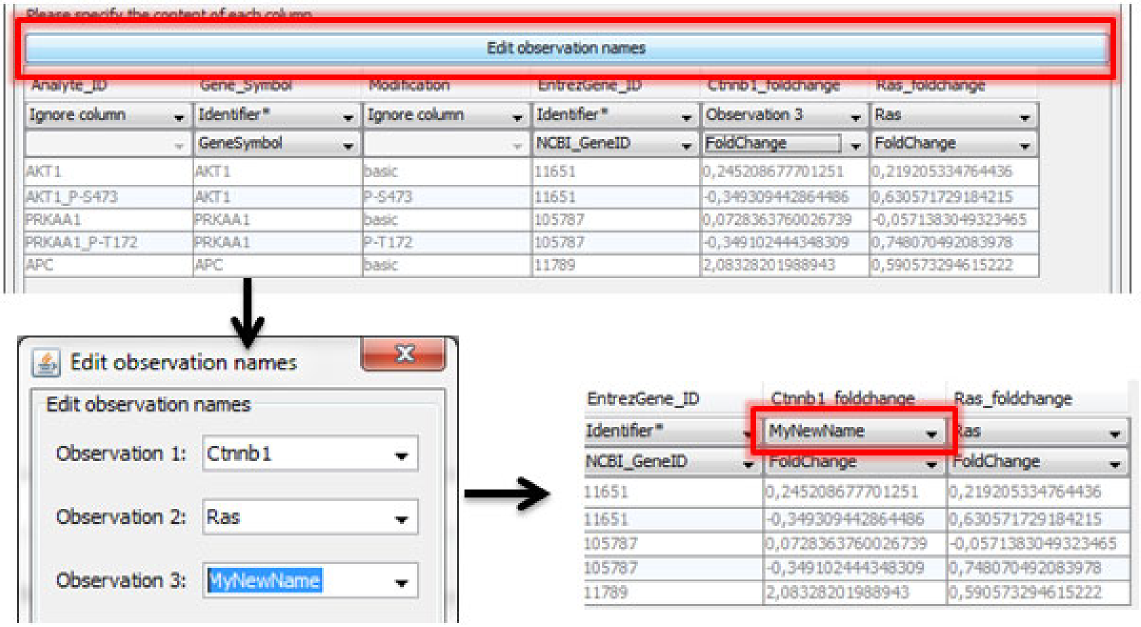
EntrezGene_ID	Ctnnb1_foldchange	Ras_foldchange
Identifier*	Observation 3	Ras
NCBI_GeneID	FoldChange	FoldChange
11651		0,219205334764436
11651		0,630571729184215
105787	0,0728363760026739	-0,0571383049323465
105787	-0,349102444348309	0,748070492083978
11789	2,08328201988943	0,590573294615222

...and picking the observation type in the combo box below.

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 more meaningful names. 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 observations names, as soon as the button ”Edit observation names” is pressed. Figure 3.5 demonstrates how to rename observations.



Please specify the content of each column

Analyte_ID	Gene_Symbol	Modification	EntrezGene_ID	Ctnnb1_foldchange	Ras_foldchange
Ignore column	Identifier*	Ignore column	Identifier*	Observation 3	Ras
	GeneSymbol		NCBI_GeneID	FoldChange	FoldChange
AKT1	AKT1	basic	11651	0,245208677701251	0,219205334764436
AKT1_P-S473	AKT1	P-S473	11651	-0,349309442864486	0,630571729184215
PRKAA1	PRKAA1	basic	105787	0,0728363760026739	-0,0571383049323465
PRKAA1_P-T172	PRKAA1	P-T172	105787	-0,349102444348309	0,748070492083978
APC	APC	basic	11789	2,08328201988943	0,590573294615222

Edit observation names

Edit observation names

Observation 1: Ctnnb1

Observation 2: Ras

Observation 3: MyNewName

EntrezGene_ID	Ctnnb1_foldchange	Ras_foldchange
Identifier*	MyNewName	Ras
NCBI_GeneID	FoldChange	FoldChange
11651	0,245208677701251	0,219205334764436
11651	-0,349309442864486	0,630571729184215
105787	0,0728363760026739	-0,0571383049323465
105787	-0,349102444348309	0,748070492083978
11789	2,08328201988943	0,590573294615222

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 InCroMAP. 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)

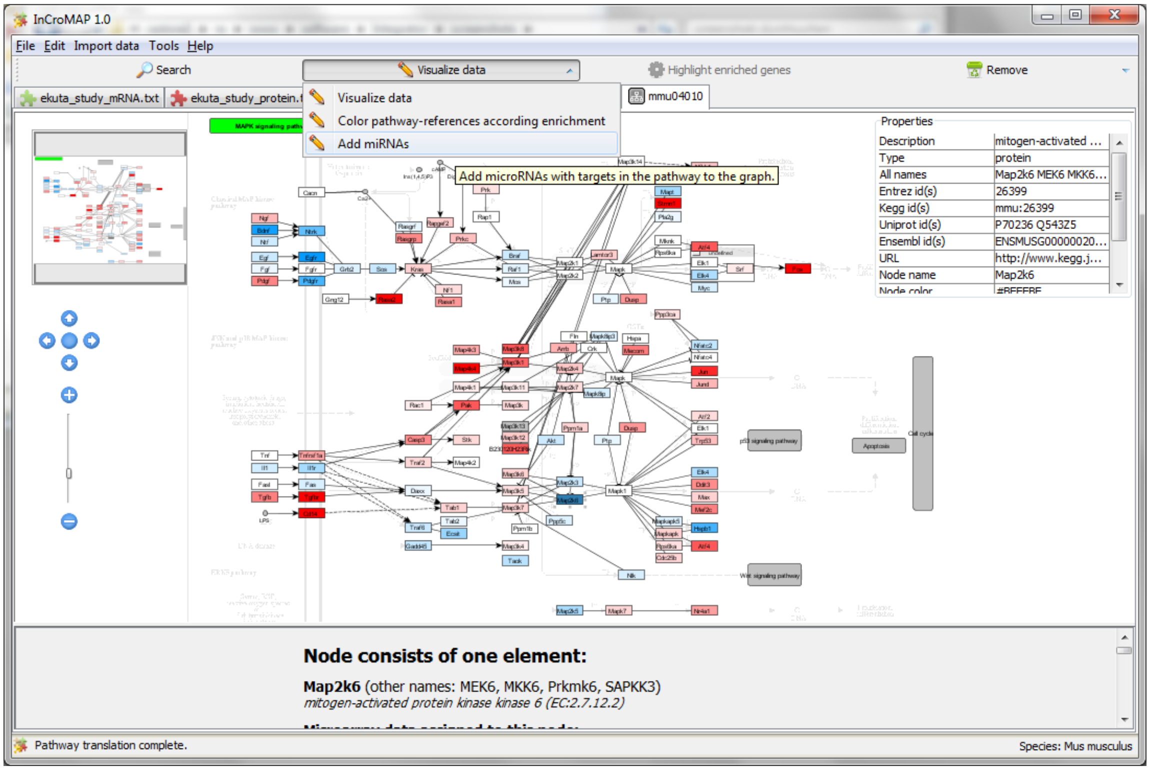
There are two ways of performing an enrichment analysis. On is, to manually select lines in the table (corresponding to probes or genes) and right click on one of the selected lines. Then, you can select an enrichment method (e.g., Gene Ontology or KEGG pathway) to perform this analysis.

The other way is to click the enrichment button in the Toolbar. Then, on needs to select an enrichment type and one will need to define a filter to select probes/genes for an enrichment. For example, in the upcoming dialog one can pick a fold-change column, select "$|>=|$" (which is the same as "$\geq \pm$") and enter 1.0 in the last field. This example 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)

There are multiple possibilities to visualize expression data in a pathway-plot, using InCroMAP. On is, for example, 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.



Properties

Description	mitogen-activated ...
Type	protein
All names	Map2k6 MEK6 MKK6...
Entrez id(s)	26399
Kegg id(s)	mmu:26399
Uniprot id(s)	P70236 Q543Z5
Ensembl id(s)	ENSMUSG00000020...
URL	http://www.kegg.j...
Node name	Map2k6
Node color	#BEEFBE

Node consists of one element:

Map2k6 (other names: MEK6, MKK6, Prkmk6, SAPKK3)
mitogen-activated protein kinase kinase 6 (EC:2.7.12.2)

Microarray data assigned to this node:

Pathway translation complete.

Species: Mus musculus

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 and three databases with predictions available.

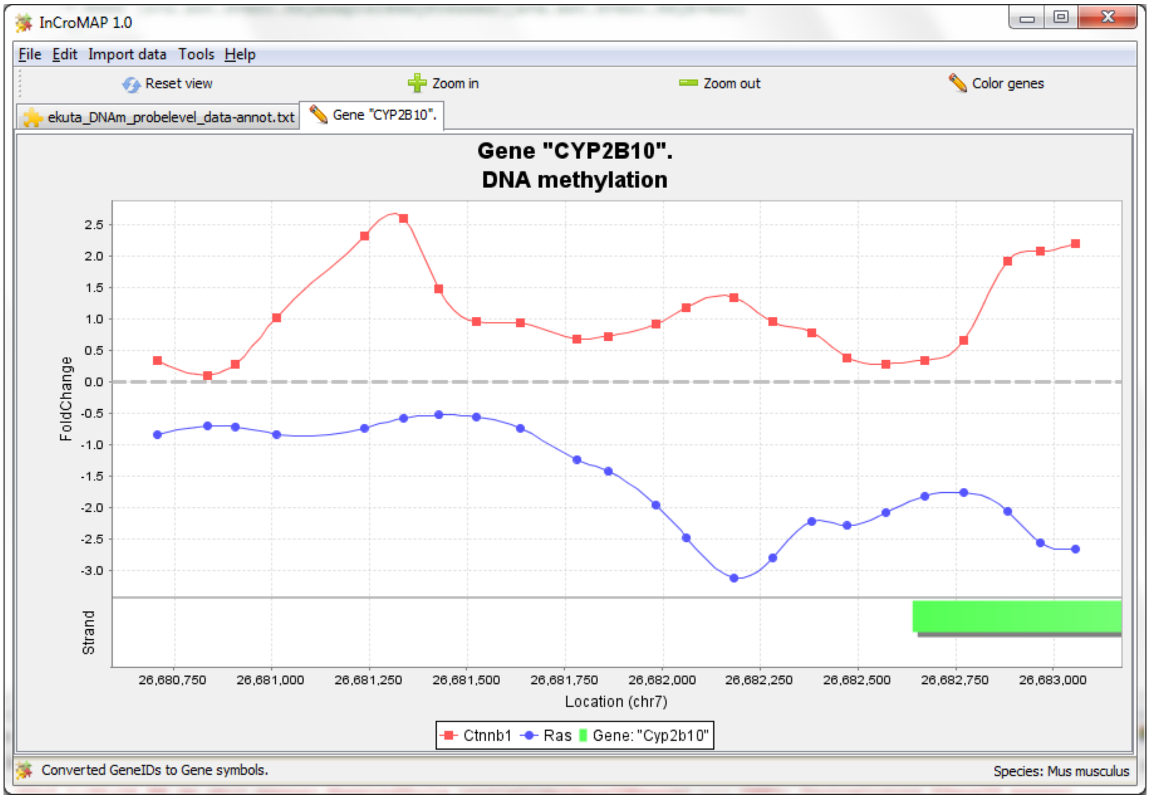
Please note that for all predictions, only targets that are aboved the suggested ”High-confidence” threshold of the cooresponding prediction database have been included in InCroMAP. Targets with low or medium prediction confidence have not been included at all.

Furthermore, many researchers follow the rule that ”the more prediction methods suggest a target, the more reliable this information is”. This rule has not been implemented into InCroMAP, because this conclusion is wrong. There are multiple reviews, suggesting that it is better to use one good prediction algorithm, than combining multiples with an ”OR”. 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)



Converted GeneIDs to Gene symbols.

Species: Mus musculus

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 possibilites to generate such a plot: The quick method is by selecting and probe (line in the table), make a right click on it and select ”Plot genome region”. The application will then collect all probes that are assigned to the selected gene and create a genome plot of them (see Figure 4.2 for an example).

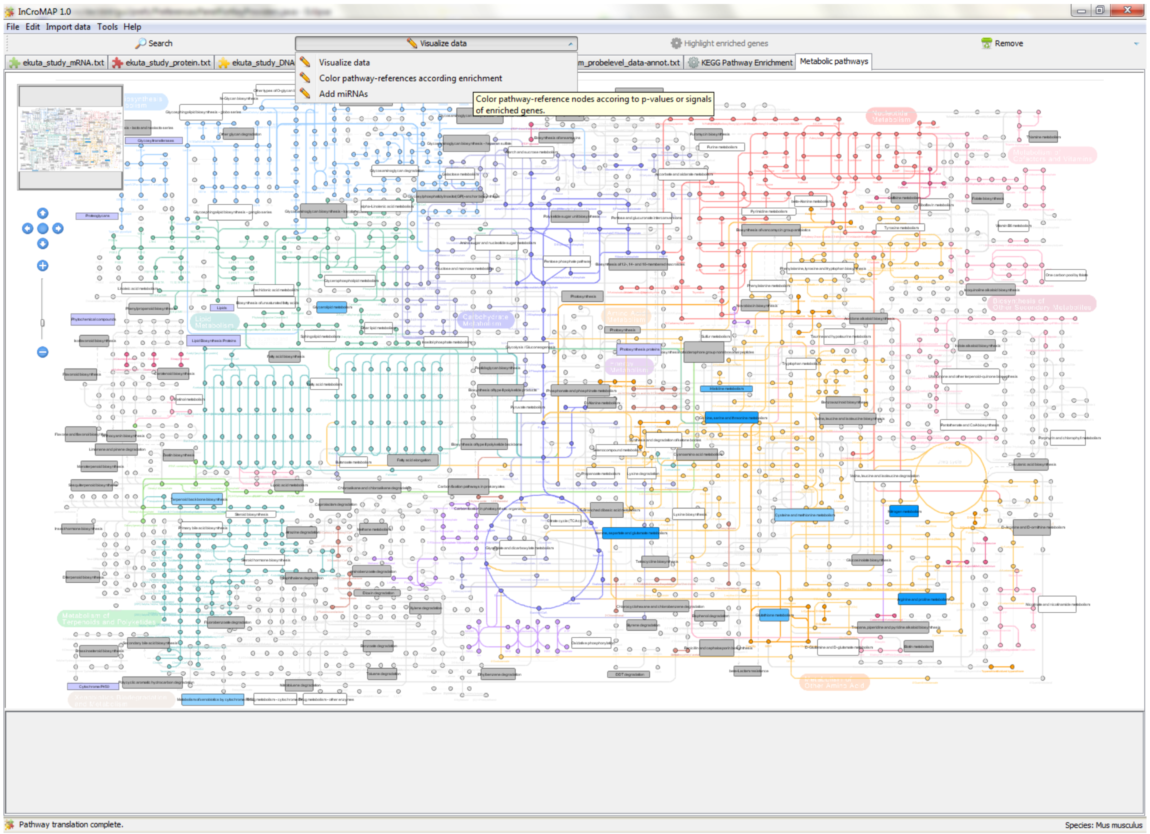
The other possibility is, to click the ”Plot genome region” button in the toolbar. The upcoming dialog then let’s you pick either a custom region by manually specifying chromosome, start and end coordinates. Or, after clicking ”Plot a region, associated to a gene”, you can select any gene and let InCroMAP plot all probes that are associated to 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)

A good way of getting started with novel mRNA expression datasets is, creating metabolic overviews. This will create a plot, 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 pathway, 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 correspond to references to other pathways (big rectangles with rounded edges) 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 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.



Color pathway-reference nodes accoring to p-values or signals of enriched genes.

Pathway translation complete.

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

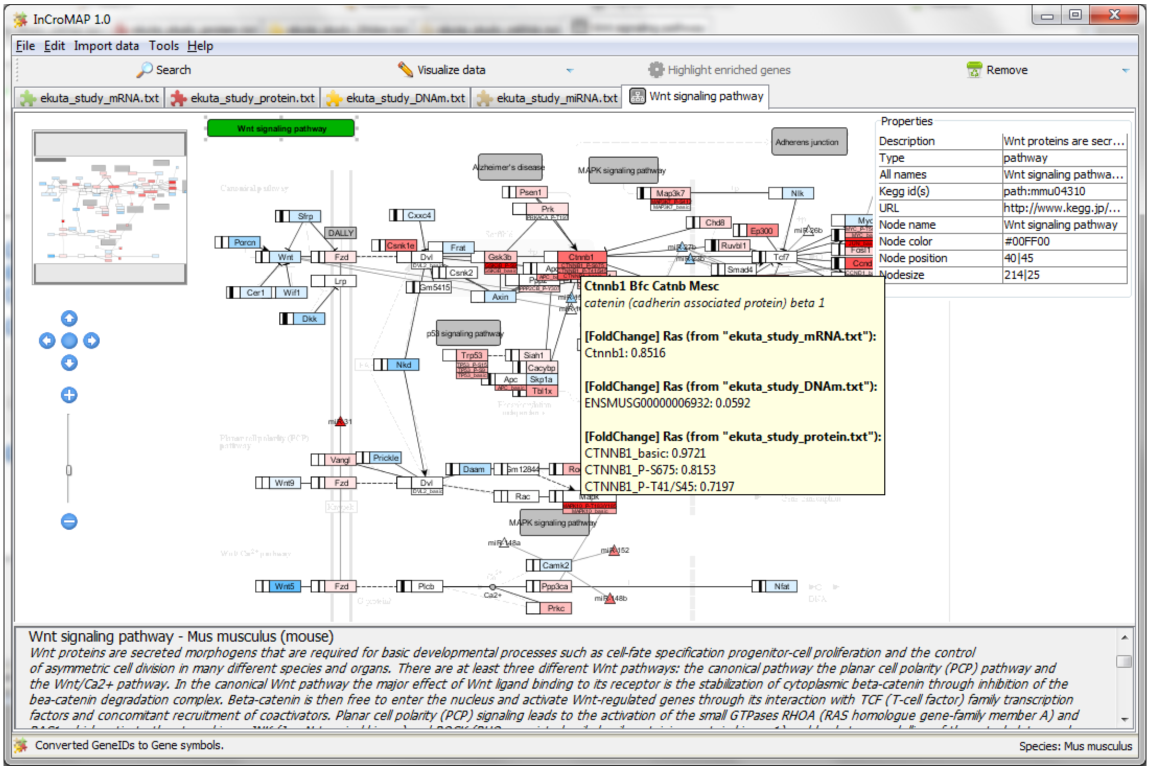
This chapter describes analysis methods in the InCroMAP application, that are available to users who have data from two or more different platforms for the same samples.

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 with any other platform combinations. It 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 targetted 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 lead to three additional columns. One



Properties

Description	Wnt proteins are secr…
Type	pathway
All names	Wnt signaling pathwa…
Kegg id(s)	path:mmu04310
URL	http://www.kegg.jp/…
Node name	Wnt signaling pathway
Node color	#00FF00
Node position	40|45
Nodesize	214|25

Ctnnb1 Bfc Catnb Mesc
catenin (cadherin associated protein) beta 1

[FoldChange] Ras (from "ekuta_study_mRNA.txt"):
Ctnnb1: 0.8516

[FoldChange] Ras (from "ekuta_study_DNAm.txt"):
ENSMUSG00000006932: 0.0592

[FoldChange] Ras (from "ekuta_study_protein.txt"):
CTNNB1_basic: 0.9721
CTNNB1_P-S675: 0.8153
CTNNB1_P-T41/S45: 0.7197

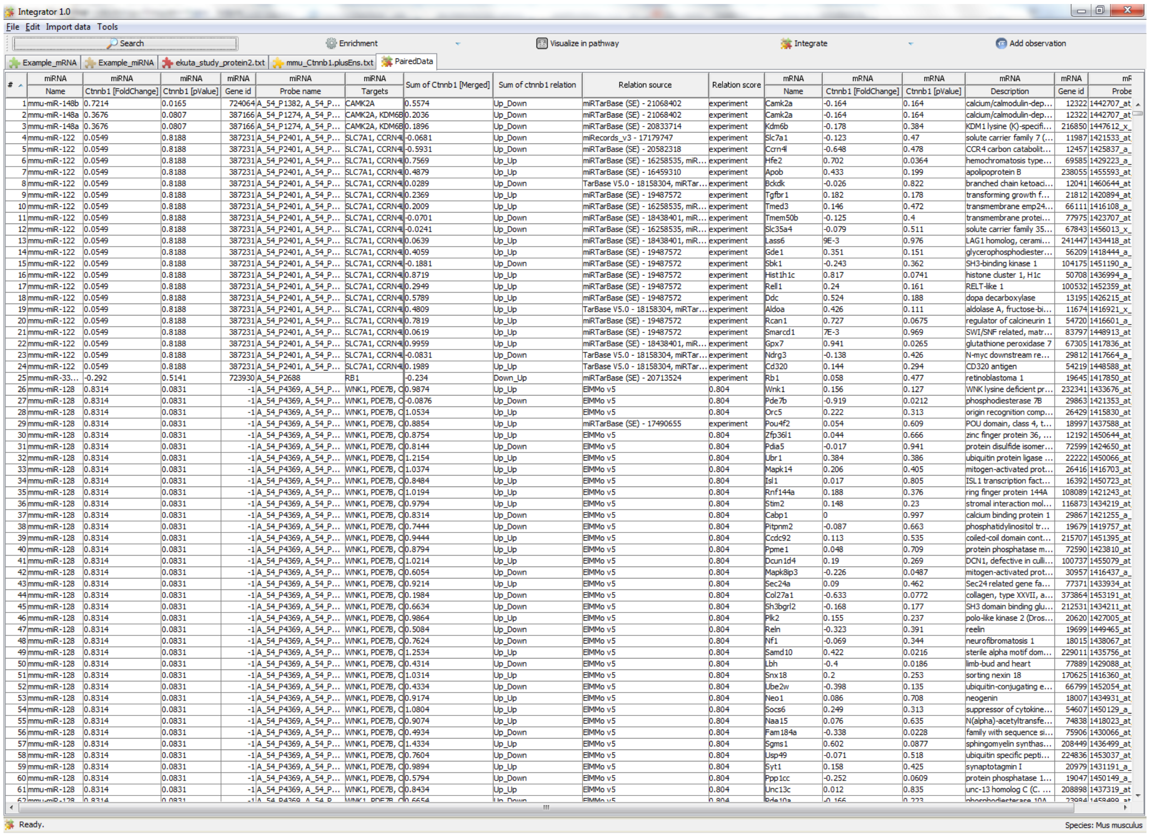
Wnt signaling pathway - Mus musculus (mouse)
Wnt proteins are secreted morphogens that are required for basic developmental processes such as cell-fate specification progenitor-cell proliferation and the control of asymmetric cell division in many different species and organs. There are at least three different Wnt pathways: the canonical pathway the planar cell polarity (PCP) pathway and the Wnt/Ca2+ pathway. In the canonical Wnt pathway the major effect of Wnt ligand binding to its receptor is the stabilization of cytoplasmic beta-catenin through inhibition of the bea-catenin degradation complex. Beta-catenin is then free to enter the nucleus and activate Wnt-regulated genes through its interaction with TCF (T-cell factor) family transcription factors and concomitant recruitment of coactivators. Planar cell polarity (PCP) signaling leads to the activation of the small GTPases RHOA (RAS homologue gene-family member A) and

Converted GeneIDs to Gene symbols.

Species: Mus musculus

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 indificates 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 proein 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.

relation column will simplify the miRNA expression fold-change and target fold-change relation. E.g., "Up_Down" means that microRNA expression is upregulated and the target (e.g., 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



#	miRNA	miRNA	miRNA	miRNA	miRNA	miRNA	Sum of Ctnnb1 [Merged]	Sum of ctnnb1 relation	Relation source	Relation score	mRNA	mRNA	mRNA	mRNA	mRNA	mR
	Name	Ctnnb1 [FoldChange]	Ctnnb1 [pValue]	Gene id	Probe name	Targets					Name	Ctnnb1 [FoldChange]	Ctnnb1 [pValue]	Description	Gene id	Probe
1	mmu-miR-148b	0.7214	0.0165	724064	A_54_P1382, A_54_P...	CAMK2A	0.5574	Up_Down	miRTarBase (SE) - 21068402	experiment	Camk2a	-0.164	0.164	calcium/calmodulin-dep...	12322	1442707_at
2	mmu-miR-148a	0.3676	0.0807	387166	A_54_P1274, A_54_P...	CAMK2A, KDM6B	0.2036	Up_Down	miRTarBase (SE) - 21068402	experiment	Camk2a	-0.164	0.164	calcium/calmodulin-dep...	12322	1442707_at
3	mmu-miR-148a	0.3676	0.0807	387166	A_54_P1274, A_54_P...	CAMK2A, KDM6B	0.1896	Up_Down	miRTarBase (SE) - 20833714	experiment	Kdm6b	-0.178	0.384	KDM1 lysine (K)-specifi...	216850	1447612_x_
4	mmu-miR-122	0.0549	0.8188	387231	A_54_P2401, A_54_P...	SLC7A1, CCRN4L	-0.0681	Up_Down	miRecords_v3 - 17179747	experiment	Slc7a1	-0.123	0.47	solute carrier family 7 (...	11987	1421533_at
5	mmu-miR-122	0.0549	0.8188	387231	A_54_P2401, A_54_P...	SLC7A1, CCRN4L	-0.5931	Up_Down	miRTarBase (SE) - 20582318	experiment	Ccrn4l	-0.648	0.478	CCR4 carbon catabolit...	12457	1425837_a_
6	mmu-miR-122	0.0549	0.8188	387231	A_54_P2401, A_54_P...	SLC7A1, CCRN4L	0.7569	Up_Up	miRTarBase (SE) - 16258535, miR...	experiment	Hfe2	0.702	0.0364	hemochromatosis type...	69585	1429223_a_
7	mmu-miR-122	0.0549	0.8188	387231	A_54_P2401, A_54_P...	SLC7A1, CCRN4L	0.4879	Up_Up	miRTarBase (SE) - 16459310	experiment	Apob	0.433	0.199	apolipoprotein B	238055	1455593_at
8	mmu-miR-122	0.0549	0.8188	387231	A_54_P2401, A_54_P...	SLC7A1, CCRN4L	0.0289	Up_Down	TarBase V5.0 - 18158304, miRTar...	experiment	Bckdk	-0.026	0.822	branched chain ketoaci...	12041	1460644_at
9	mmu-miR-122	0.0549	0.8188	387231	A_54_P2401, A_54_P...	SLC7A1, CCRN4L	0.2369	Up_Up	miRTarBase (SE) - 19487572	experiment	Tgfbr1	0.182	0.178	transforming growth f...	21812	1420894_at
10	mmu-miR-122	0.0549	0.8188	387231	A_54_P2401, A_54_P...	SLC7A1, CCRN4L	0.2009	Up_Up	miRTarBase (SE) - 16258535, miR...	experiment	Tmed3	0.146	0.472	transmembrane emp24...	66111	1416108_a_
11	mmu-miR-122	0.0549	0.8188	387231	A_54_P2401, A_54_P...	SLC7A1, CCRN4L	-0.0701	Up_Down	miRTarBase (SE) - 18438401, miR...	experiment	Tmem50b	-0.125	0.4	transmembrane protei...	77975	1423707_at
12	mmu-miR-122	0.0549	0.8188	387231	A_54_P2401, A_54_P...	SLC7A1, CCRN4L	-0.0241	Up_Down	miRTarBase (SE) - 16258535, miR...	experiment	Slc35a4	-0.079	0.511	solute carrier family 35...	67843	1456013_x_
13	mmu-miR-122	0.0549	0.8188	387231	A_54_P2401, A_54_P...	SLC7A1, CCRN4L	0.0639	Up_Up	miRTarBase (SE) - 18438401, miR...	experiment	Lass6	9E-3	0.976	LAG1 homolog, cerami...	241447	1434418_at
14	mmu-miR-122	0.0549	0.8188	387231	A_54_P2401, A_54_P...	SLC7A1, CCRN4L	0.4059	Up_Up	miRTarBase (SE) - 19487572	experiment	Gde1	0.351	0.151	glycerophosphodiester...	56209	1418444_a_
15	mmu-miR-122	0.0549	0.8188	387231	A_54_P2401, A_54_P...	SLC7A1, CCRN4L	-0.1881	Up_Down	miRTarBase (SE) - 19487572	experiment	Sbk1	-0.243	0.362	SH3-binding kinase 1	104175	1451190_a_
16	mmu-miR-122	0.0549	0.8188	387231	A_54_P2401, A_54_P...	SLC7A1, CCRN4L	0.8719	Up_Up	miRTarBase (SE) - 19487572	experiment	Hist1h1c	0.817	0.0741	histone cluster 1, H1c	50708	1436994_a_
17	mmu-miR-122	0.0549	0.8188	387231	A_54_P2401, A_54_P...	SLC7A1, CCRN4L	0.2949	Up_Up	miRTarBase (SE) - 19487572	experiment	Rell1	0.24	0.161	RELT-like 1	100532	1452359_at
18	mmu-miR-122	0.0549	0.8188	387231	A_54_P2401, A_54_P...	SLC7A1, CCRN4L	0.5789	Up_Up	miRTarBase (SE) - 19487572	experiment	Ddc	0.524	0.188	dopa decarboxylase	13195	1426215_at
19	mmu-miR-122	0.0549	0.8188	387231	A_54_P2401, A_54_P...	SLC7A1, CCRN4L	0.4809	Up_Up	TarBase V5.0 - 18158304, miRTar...	experiment	Aldoa	0.426	0.111	aldolase A, fructose-bi...	11674	1416921_x_
20	mmu-miR-122	0.0549	0.8188	387231	A_54_P2401, A_54_P...	SLC7A1, CCRN4L	0.7819	Up_Up	miRTarBase (SE) - 19487572	experiment	Rcan1	0.727	0.0675	regulator of calcineurin 1	54720	1416601_a_
21	mmu-miR-122	0.0549	0.8188	387231	A_54_P2401, A_54_P...	SLC7A1, CCRN4L	0.0619	Up_Up	miRTarBase (SE) - 19487572	experiment	Smarcd1	7E-3	0.969	SWI/SNF related, matr...	83797	1448913_at
22	mmu-miR-122	0.0549	0.8188	387231	A_54_P2401, A_54_P...	SLC7A1, CCRN4L	0.9959	Up_Up	miRTarBase (SE) - 18438401, miR...	experiment	Gpx7	0.941	0.0265	glutathione peroxidase 7	67305	1417836_at
23	mmu-miR-122	0.0549	0.8188	387231	A_54_P2401, A_54_P...	SLC7A1, CCRN4L	-0.0831	Up_Down	TarBase V5.0 - 18158304, miRTar...	experiment	Ndrg3	-0.138	0.426	N-myc downstream re...	29812	1417664_a_
24	mmu-miR-122	0.0549	0.8188	387231	A_54_P2401, A_54_P...	SLC7A1, CCRN4L	0.1989	Up_Up	TarBase V5.0 - 18158304, miRTar...	experiment	Cd320	0.144	0.294	CD320 antigen	54219	1448588_at
25	mmu-miR-33...	-0.292	0.5141	723930	A_54_P2688	RB1	-0.234	Down_Up	miRTarBase (SE) - 20713524	experiment	Rb1	0.058	0.477	retinoblastoma 1	19645	1417850_at
26	mmu-miR-128	0.8314	0.0831	-1	A_54_P4369, A_54_P...	WNK1, PDE7B, C	0.9874	Up_Up	ElMMo v5	0.804	Wnk1	0.156	0.127	WNK lysine deficient pr...	232341	1433676_at
27	mmu-miR-128	0.8314	0.0831	-1	A_54_P4369, A_54_P...	WNK1, PDE7B, C	-0.0876	Up_Down	ElMMo v5	0.804	Pde7b	-0.919	0.0212	phosphodiesterase 7B	29863	1421353_at
28	mmu-miR-128	0.8314	0.0831	-1	A_54_P4369, A_54_P...	WNK1, PDE7B, C	1.0534	Up_Up	ElMMo v5	0.804	Orc5	0.222	0.313	origin recognition comp...	26429	1415830_at
29	mmu-miR-128	0.8314	0.0831	-1	A_54_P4369, A_54_P...	WNK1, PDE7B, C	0.8854	Up_Up	miRTarBase (SE) - 17490655	experiment	Pou4f2	0.054	0.609	POU domain, class 4, t...	18997	1437588_at
30	mmu-miR-128	0.8314	0.0831	-1	A_54_P4369, A_54_P...	WNK1, PDE7B, C	0.8754	Up_Up	ElMMo v5	0.804	Zfp36l1	0.044	0.666	zinc finger protein 36, ...	12192	1450644_at
31	mmu-miR-128	0.8314	0.0831	-1	A_54_P4369, A_54_P...	WNK1, PDE7B, C	0.8144	Up_Down	ElMMo v5	0.804	Pdia5	-0.017	0.941	protein disulfide isomer...	72599	1424650_at
32	mmu-miR-128	0.8314	0.0831	-1	A_54_P4369, A_54_P...	WNK1, PDE7B, C	1.2154	Up_Up	ElMMo v5	0.804	Ubr1	0.384	0.386	ubiquitin protein ligase ...	22222	1450066_at
33	mmu-miR-128	0.8314	0.0831	-1	A_54_P4369, A_54_P...	WNK1, PDE7B, C	1.0374	Up_Up	ElMMo v5	0.804	Mapk14	0.206	0.405	mitogen-activated prot...	26416	1416703_at
34	mmu-miR-128	0.8314	0.0831	-1	A_54_P4369, A_54_P...	WNK1, PDE7B, C	0.8484	Up_Up	ElMMo v5	0.804	Isl1	0.017	0.805	ISL1 transcription fact...	16392	1450723_at
35	mmu-miR-128	0.8314	0.0831	-1	A_54_P4369, A_54_P...	WNK1, PDE7B, C	1.0194	Up_Up	ElMMo v5	0.804	Rnf144a	0.188	0.376	ring finger protein 144A	108089	1421243_at
36	mmu-miR-128	0.8314	0.0831	-1	A_54_P4369, A_54_P...	WNK1, PDE7B, C	0.9794	Up_Up	ElMMo v5	0.804	Stim2	0.148	0.23	stromal interaction mol...	116873	1434219_at
37	mmu-miR-128	0.8314	0.0831	-1	A_54_P4369, A_54_P...	WNK1, PDE7B, C	0.8314	Up_Down	ElMMo v5	0.804	Cabp1	0	0.997	calcium binding protein 1	29867	1421255_a_
38	mmu-miR-128	0.8314	0.0831	-1	A_54_P4369, A_54_P...	WNK1, PDE7B, C	0.7444	Up_Down	ElMMo v5	0.804	Pitpnm2	-0.087	0.663	phosphatidylinositol tr...	19679	1419757_at
39	mmu-miR-128	0.8314	0.0831	-1	A_54_P4369, A_54_P...	WNK1, PDE7B, C	0.9444	Up_Up	ElMMo v5	0.804	Ccdc92	0.113	0.535	coiled-coil domain cont...	215707	1451395_at
40	mmu-miR-128	0.8314	0.0831	-1	A_54_P4369, A_54_P...	WNK1, PDE7B, C	0.8794	Up_Up	ElMMo v5	0.804	Ppme1	0.048	0.709	protein phosphatase m...	72590	1423810_at
41	mmu-miR-128	0.8314	0.0831	-1	A_54_P4369, A_54_P...	WNK1, PDE7B, C	1.0214	Up_Up	ElMMo v5	0.804	Dcun1d4	0.19	0.269	DCN1, defective in culli...	100737	1455079_at
42	mmu-miR-128	0.8314	0.0831	-1	A_54_P4369, A_54_P...	WNK1, PDE7B, C	0.6054	Up_Down	ElMMo v5	0.804	Mapk8ip3	-0.226	0.0487	mitogen-activated prot...	30957	1416437_a_
43	mmu-miR-128	0.8314	0.0831	-1	A_54_P4369, A_54_P...	WNK1, PDE7B, C	0.9214	Up_Up	ElMMo v5	0.804	Sec24a	0.09	0.462	Sec24 related gene fa...	77371	1433934_at
44	mmu-miR-128	0.8314	0.0831	-1	A_54_P4369, A_54_P...	WNK1, PDE7B, C	0.1984	Up_Down	ElMMo v5	0.804	Col27a1	-0.633	0.0772	collagen, type XXVII, a...	373864	1453191_at
45	mmu-miR-128	0.8314	0.0831	-1	A_54_P4369, A_54_P...	WNK1, PDE7B, C	0.6634	Up_Down	ElMMo v5	0.804	Sh3bgrl2	-0.168	0.177	SH3 domain binding glu...	212531	1434211_at
46	mmu-miR-128	0.8314	0.0831	-1	A_54_P4369, A_54_P...	WNK1, PDE7B, C	0.9864	Up_Up	ElMMo v5	0.804	Plk2	0.155	0.237	polo-like kinase 2 (Dros...	20620	1427005_at
47	mmu-miR-128	0.8314	0.0831	-1	A_54_P4369, A_54_P...	WNK1, PDE7B, C	0.5084	Up_Down	ElMMo v5	0.804	Reln	-0.323	0.391	reelin	19699	1449465_at
48	mmu-miR-128	0.8314	0.0831	-1	A_54_P4369, A_54_P...	WNK1, PDE7B, C	0.7624	Up_Down	ElMMo v5	0.804	Nf1	-0.069	0.344	neurofibromatosis 1	18015	1438067_at
49	mmu-miR-128	0.8314	0.0831	-1	A_54_P4369, A_54_P...	WNK1, PDE7B, C	1.2534	Up_Up	ElMMo v5	0.804	Samd10	0.422	0.0216	sterile alpha motif dom...	229011	1435756_at
50	mmu-miR-128	0.8314	0.0831	-1	A_54_P4369, A_54_P...	WNK1, PDE7B, C	0.4314	Up_Down	ElMMo v5	0.804	Lbh	-0.4	0.0186	limb-bud and heart	77889	1429088_at
51	mmu-miR-128	0.8314	0.0831	-1	A_54_P4369, A_54_P...	WNK1, PDE7B, C	1.0314	Up_Up	ElMMo v5	0.804	Snx18	0.2	0.253	sorting nexin 18	170625	1416360_at
52	mmu-miR-128	0.8314	0.0831	-1	A_54_P4369, A_54_P...	WNK1, PDE7B, C	0.4334	Up_Down	ElMMo v5	0.804	Ube2w	-0.398	0.135	ubiquitin-conjugating e...	66799	1452054_at
53	mmu-miR-128	0.8314	0.0831	-1	A_54_P4369, A_54_P...	WNK1, PDE7B, C	0.9174	Up_Up	ElMMo v5	0.804	Neo1	0.086	0.708	neogenin	18007	1434931_at
54	mmu-miR-128	0.8314	0.0831	-1	A_54_P4369, A_54_P...	WNK1, PDE7B, C	1.0804	Up_Up	ElMMo v5	0.804	Socs6	0.249	0.313	suppressor of cytokine...	54607	1450129_a_
55	mmu-miR-128	0.8314	0.0831	-1	A_54_P4369, A_54_P...	WNK1, PDE7B, C	0.9074	Up_Up	ElMMo v5	0.804	Naa15	0.076	0.635	N(alpha)-acetyltransfe...	74838	1418023_at
56	mmu-miR-128	0.8314	0.0831	-1	A_54_P4369, A_54_P...	WNK1, PDE7B, C	0.4934	Up_Down	ElMMo v5	0.804	Fam184a	-0.338	0.0228	family with sequence si...	75906	1430066_at
57	mmu-miR-128	0.8314	0.0831	-1	A_54_P4369, A_54_P...	WNK1, PDE7B, C	1.4334	Up_Up	ElMMo v5	0.804	Sgms1	0.602	0.0877	sphingomyelin synthas...	208449	1436499_at
58	mmu-miR-128	0.8314	0.0831	-1	A_54_P4369, A_54_P...	WNK1, PDE7B, C	0.7604	Up_Down	ElMMo v5	0.804	Usp49	-0.071	0.518	ubiquitin specific pepti...	224836	1453037_at
59	mmu-miR-128	0.8314	0.0831	-1	A_54_P4369, A_54_P...	WNK1, PDE7B, C	0.9894	Up_Up	ElMMo v5	0.804	Syt1	0.158	0.425	synaptotagmin I	20979	1431191_a_
60	mmu-miR-128	0.8314	0.0831	-1	A_54_P4369, A_54_P...	WNK1, PDE7B, C	0.5794	Up_Down	ElMMo v5	0.804	Ppp1cc	-0.252	0.0609	protein phosphatase 1...	19047	1450149_a_
61	mmu-miR-128	0.8314	0.0831	-1	A_54_P4369, A_54_P...	WNK1, PDE7B, C	0.8434	Up_Up	ElMMo v5	0.804	Unc13c	0.012	0.835	unc-13 homolog C (C. ...	208898	1437319_at
62	mmu-miR-128	0.8314	0.0831	-1	A_54_P4369, A_54_P...	WNK1, PDE7B, C	0.6654	Up_Down	ElMMo v5	0.804	Pde10a	-0.166	0.223	phosphodiesterase 10A	23984	1458499_at

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).

target is a predicted target. Please refer to the corresponding 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- or medium confidence targets are not included in InCroMAP at all.

5.3 Integrated enrichment

Normal enrichment analyzes are 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.

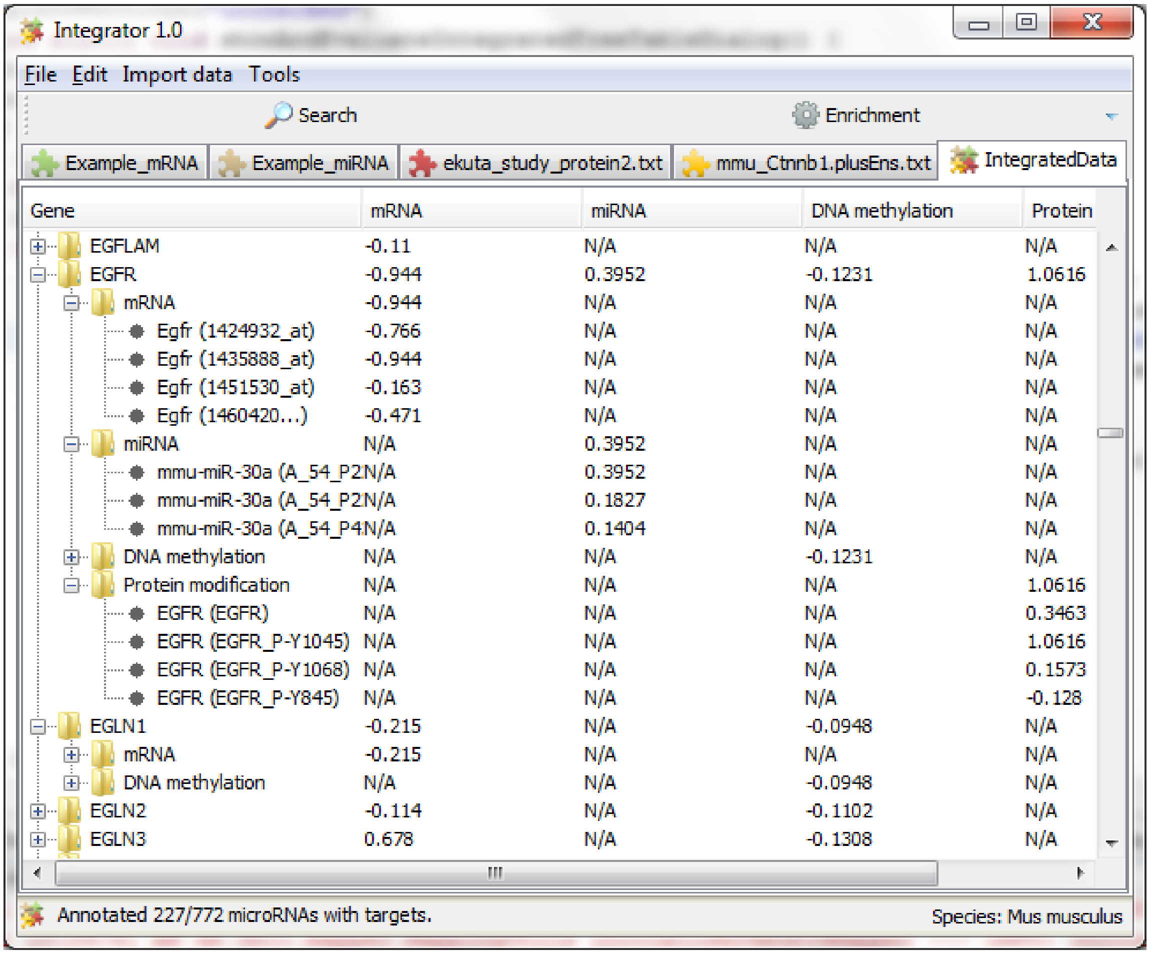
To perform such an enrichment, please select ”Tools” and ”Integrated enrichment” from the menu bar. Now select your organism and all currently opened datasets from this organism should show up in the box below the organism selection. Pick all datasets that should be integrated into this enrichment (you 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. If 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 to this gene appear. Thus, this integration methods gives a possibility to quickly examine expression data, coming from various different platforms if one already knows the 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



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 columns 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 to this gene on the particular platform show up (in this example, all mRNA probes). For miRNA datasets, all miRNAs targetting this gene are shown.

want to integrate. Then, pick a dataset and an observation for each platform. Please note that this 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 show up.

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.

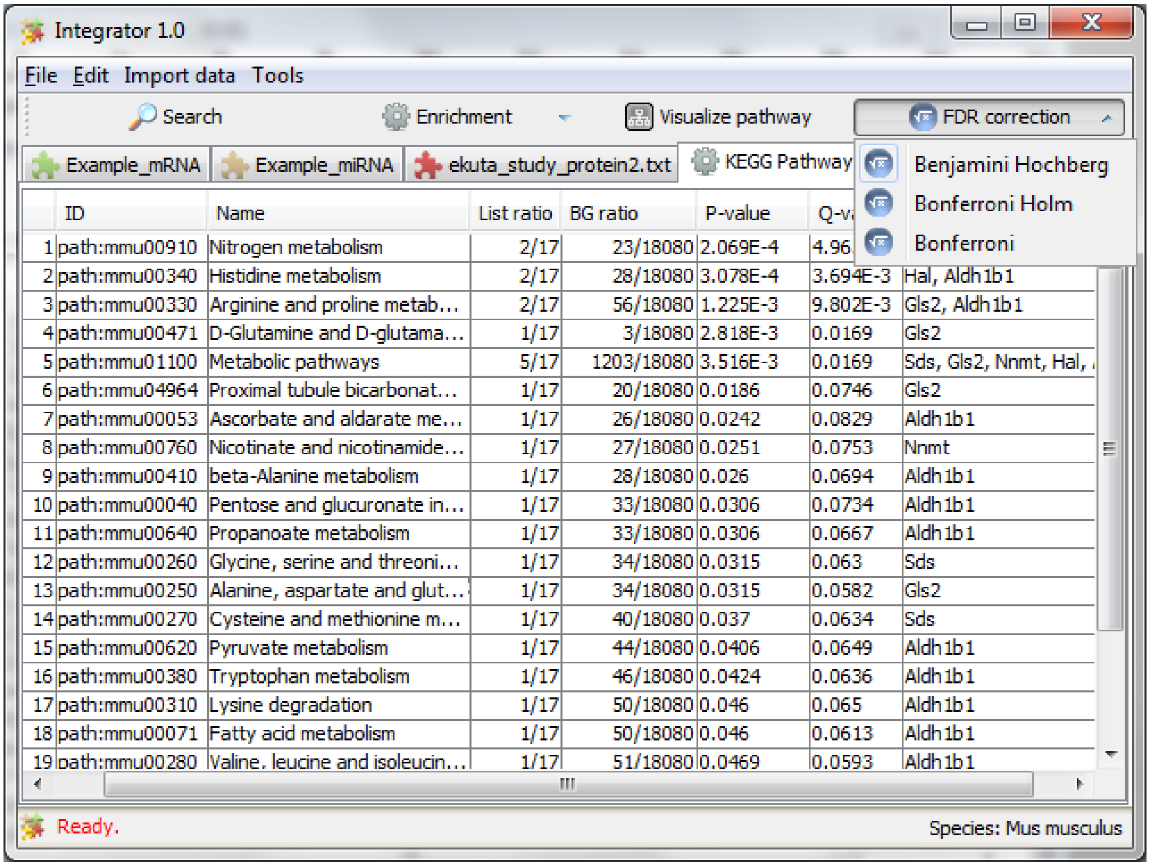
All tables in InCroMAP can be searched, by simply clicking on any row in the table (i.e. setting the focus on the table) and start typing. This will result in a textfield to show up at the top left edge of the table. The table selection will automatically change to select an object that contains what 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-change or p-values). The MSigDB enrichment will require you to specify your enrichment files. You can download any gene set in GTM 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 from 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 number of genes from our gene-pool list that are in the gene-set, described by the row, and the total



	ID	Name	List ratio	BG ratio	P-value	Q-va	
1	path:mmu00910	Nitrogen metabolism	2/17	23/18080	2.069E-4	4.96	
2	path:mmu00340	Histidine metabolism	2/17	28/18080	3.078E-4	3.694E-3	Hal, Aldh1b1
3	path:mmu00330	Arginine and proline metab...	2/17	56/18080	1.225E-3	9.802E-3	Gls2, Aldh1b1
4	path:mmu00471	D-Glutamine and D-glutama...	1/17	3/18080	2.818E-3	0.0169	Gls2
5	path:mmu01100	Metabolic pathways	5/17	1203/18080	3.516E-3	0.0169	Sds, Gls2, Nnmt, Hal,
6	path:mmu04964	Proximal tubule bicarbonat...	1/17	20/18080	0.0186	0.0746	Gls2
7	path:mmu00053	Ascorbate and aldarate me...	1/17	26/18080	0.0242	0.0829	Aldh1b1
8	path:mmu00760	Nicotinate and nicotinamide...	1/17	27/18080	0.0251	0.0753	Nnmt
9	path:mmu00410	beta-Alanine metabolism	1/17	28/18080	0.026	0.0694	Aldh1b1
10	path:mmu00040	Pentose and glucuronate in...	1/17	33/18080	0.0306	0.0734	Aldh1b1
11	path:mmu00640	Propanoate metabolism	1/17	33/18080	0.0306	0.0667	Aldh1b1
12	path:mmu00260	Glycine, serine and threoni...	1/17	34/18080	0.0315	0.063	Sds
13	path:mmu00250	Alanine, aspartate and glut...	1/17	34/18080	0.0315	0.0582	Gls2
14	path:mmu00270	Cysteine and methionine m...	1/17	40/18080	0.037	0.0634	Sds
15	path:mmu00620	Pyruvate metabolism	1/17	44/18080	0.0406	0.0649	Aldh1b1
16	path:mmu00380	Tryptophan metabolism	1/17	46/18080	0.0424	0.0636	Aldh1b1
17	path:mmu00310	Lysine degradation	1/17	50/18080	0.046	0.065	Aldh1b1
18	path:mmu00071	Fatty acid metabolism	1/17	50/18080	0.046	0.0613	Aldh1b1
19	path:mmu00280	Valine, leucine and isoleucin...	1/17	51/18080	0.0469	0.0593	Aldh1b1

Ready.

Species: Mus musculus

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.

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 where 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. Thus, 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 of 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 preferences, you can change the list ratio when performing a microRNA enrichment to be 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 round corners describe referenced pathways (also called ”pathway-reference nodes”), and rectangular-shaped nodes (smaller ones, without round 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.

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 and DNA methylation) can be visualized at a time, but all types can be visualized in the same pathway. 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!

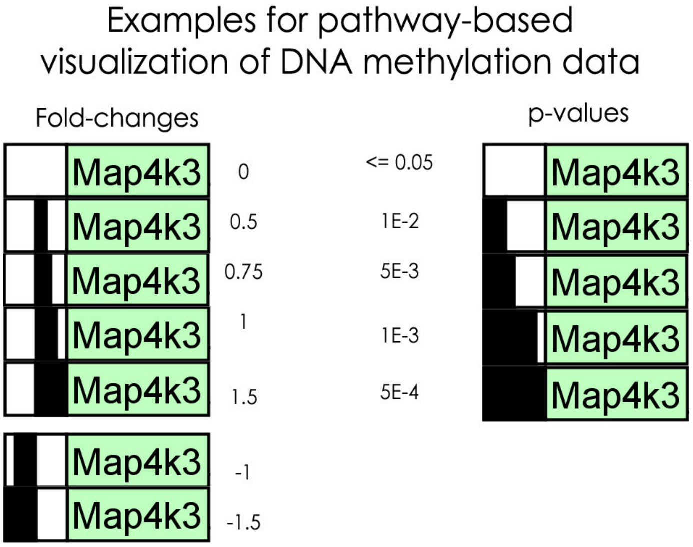
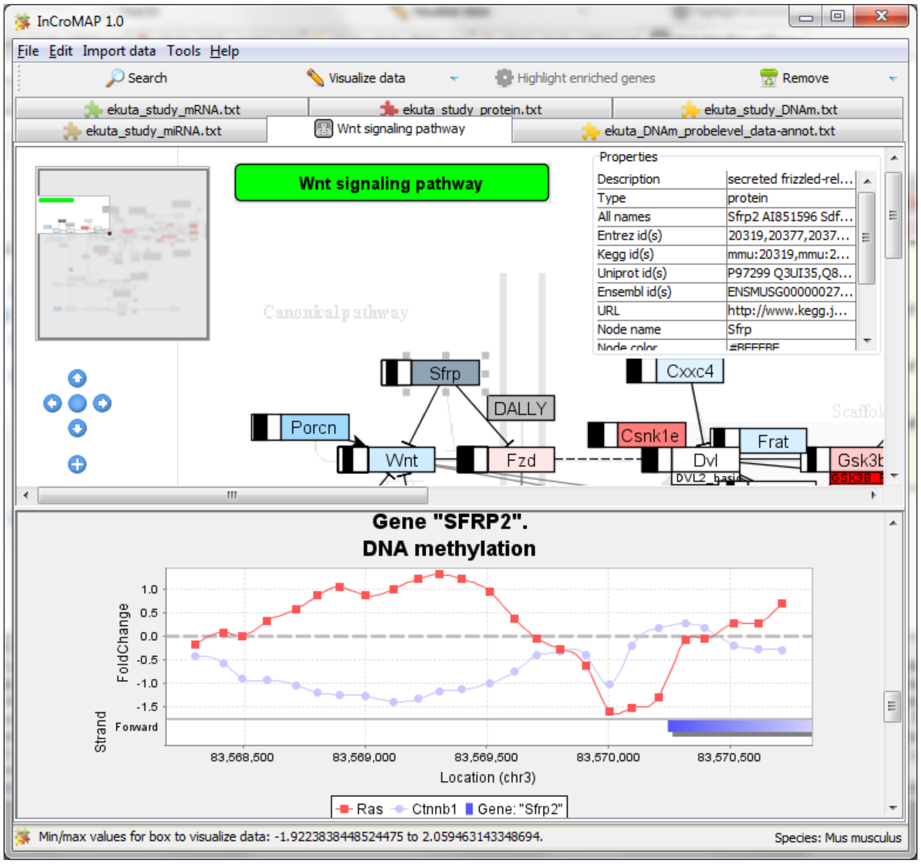


Figure 6.2: Examples of various fold-changes or p-values for DNA methylation data and their visualization in InCroMAP. These examples are only valid for the default settings. For example, the p-value required to fill the whole box can be changed in the preferences. Usually, the value visualized for a 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 associated to this node (e.g., gene was not on microarray, or 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 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 properties.

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 $|x|$), NormalizesSumOfLog2Values (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 significantly probe (p-values) or the probe with maximum differential expression (fold-changes).



Description	secreted frizzled-rel...
Type	protein
All names	Sfrp2 AI851596 Sdf...
Entrez id(s)	20319,20377,2037...
Kegg id(s)	mmu:20319,mmu:2...
Uniprot id(s)	P97299 Q3UI35,Q8...
Ensembl id(s)	ENSMUSG00000027...
URL	http://www.kegg.j...
Node name	Sfrp
Node color	#BFFFBF

Min/max values for box to visualize data: -1.9223838448524475 to 2.059463143348694.

Species: Mus musculus

Figure 6.3: Examples of an integrated pathway-based visualization. Here, three different data types (mRNA, protein and DNA methylation) have been visualized. The node ”SFRP” has been 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 boxed 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, we added a detail panel that contains a plot of the DNA methylation 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 boxes and in the lower half of the application, a new frame is shown with detailed information about the node and all expression values associated with it. Please see Figure 6.3 for an example of the detailed DNA methylation visualization.

6.3.2 Other visualization options

If you started by performing a KEGG pathway enrichment on your expression data and then selected and 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 round 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

6.4.1 Integrated mRNA expression and DNA methylation plots

PTPRT gen ist interessant

TODO: Fold-change muss logarithmiert sein. Sonst wird er bei einfrbung logarithmiert (base: 2)! (Bei data input erwhnen).

DONE? TODO: Tabellen funktion, auf suche, DNAm plot & enrichment eingehen und pw visualization einleiten DONE? enrichment, pw visualization separat. DONE? Integrated analysis methods.

DONE? - F3 Knopf bei suche

Please note: in this section we refer to the default values. The colores for upregulation

- Tutorial fr ein downloadbares sample file, von import bis zur enrichment und data in pathway ansicht

2. Open (webstart or download, simply double-click) 3. Analyze

Important Notes (in FAQ?) erklrend das kegg merhere gene in 1 knoten zusammenfasst

Jeden Screen noch mal erklren, inkl. aller columns (list ratio), MSigbDB enrichment In Analysis examples auf erklrungen verlinkgen Preferences erklren (formeln aus Vis.Expr.Data.In.PW hierher verschieben) DONE? Bei enrichment sagen wie es mit miRNA, DNAm usw. funktioniert.

TODO PAPER: Abgrenzung zu GenMAPP und Ingenuity Application ist auch ein ”KEGG Pathway viewer”

TODO: Gliederung:

mRNA data analysis miRNA data DNAm protein

Enrichment Tab

Pathway-tab

Integrated analysis

7 FAQ / Troubleshooting

Please note: There are separate FAQs for steps 1 (see Section **??**), 2 (see Section **??**) and 3 (see Section **??**).

I'm getting a "java.lang.OutOfMemoryError: Java heap space"
Some operations need a lot of memory. If you simply start Integrator, 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 or more of memory. Especially reading DNA methylation datasets takes a huge amount of memory.

Is an internet connection required to run Integrator?
An internet connection is required for most operations. Many identifier mapping files and pathway-based visualization 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 than 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/`.

TODO: Can I save XY (pathway plot picture, etc). Erklrung: File and Save

TODO: include items from KEGGtranslator FAQ (missing reactions and such).

TODO: Warum keine miRNA annotiert? (o.,) =¿ Falscher organismus gewhlt.

TODO: Signal / Observation generell erklren was damit gemeint ist

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.

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.