



## Documentation for Integrator

**integrated analysis of microarray data from different platforms**

Clemens Wrzodek

March 5, 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

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Installation</b>	<b>4</b>
2.1	Requirements . . . . .	4
2.2	Starting the application . . . . .	4
<b>3</b>	<b>How to get started</b>	<b>6</b>
3.1	1. Prepare microarray data . . . . .	6
3.2	2. Open microarray data in the application . . . . .	8
<b>4</b>	<b>Single dataset analysis examples</b>	<b>13</b>
4.1	Enrichment analysis (every platform) . . . . .	13
4.2	Pathway-based microarray data visualization (every platform) . . . . .	13
4.3	Show microRNA targets (miRNA only) . . . . .	14
4.4	Visualize microRNA targets (every platform) . . . . .	15
4.5	Genome region plot (DNA methylation only) . . . . .	15
4.6	Metabolic overview (mRNA only) . . . . .	16
<b>5</b>	<b>Integrated cross-platform microarray data analysis examples</b>	<b>18</b>
5.1	Integrated pathway-based visualization . . . . .	18
5.2	Data pairing . . . . .	18
5.3	Integrated enrichment . . . . .	20
5.4	Gene-based tabular integration of data from multiple platforms . . . . .	20
<b>6</b>	<b>FAQ / Troubleshooting</b>	<b>21</b>
	<b>Bibliography</b>	<b>22</b>

# 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 its 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:
    - \* Experimentally verified

- miRecords
  - miRTarBase
  - TarBase
  - \* Predicted
    - EIMMo
    - 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 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<sup>1</sup>.

#### 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
```

---

<sup>1</sup><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.

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). Moreover, 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.

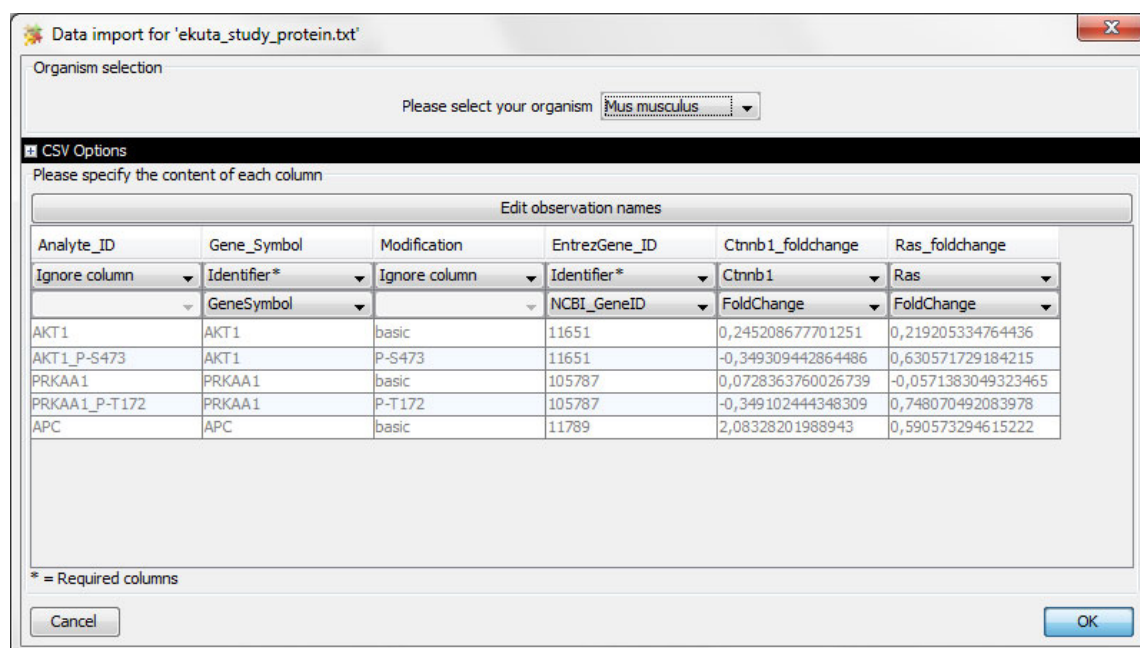
Figure 3.1 shows an example of the file input dialog. First, the species must be selected (on top of the dialog). If the input file format has not been inferred automatically, one may click on the black "CSV Options" label to specify further options (like "column separator char" or if the file contains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

### 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

## 3.2 2. Open microarray data in the application



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*	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

Cancel OK

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.

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

### 3 How to get started

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		Identifier*	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	NCBI_GeneID	0,245208677701251	0,219205334764436
AKT1_P-S473	AKT1	P-S473	RefSeq	-0,349309442864486	0,630571729184215
PRKAA1	PRKAA1	basic	Ensembl	0,0728363760026739	-0,0571383049323465
PRKAA1_P-T172	PRKAA1	P-T172	KeggGenes	-0,349102444348309	0,748070492083978
APC	APC	basic	GeneSymbol	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.

EntrezGene_ID	Ctnnb1_foldchange	Ras_foldchange
Identifier*	Observation 3	Ras
NCBI_GeneID	Ctnnb1	FoldChange
11651	Ras	0,219205334764436
11651	Observation 3	0,630571729184215
105787	Observation 4	-0,0571383049323465
105787	Observation 5	0,748070492083978
11789	Observation 6	0,590573294615222
	Observation 7	
	Observation 8	

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

Import expression data by assigning an observation to this column

...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.

### 3 How to get started

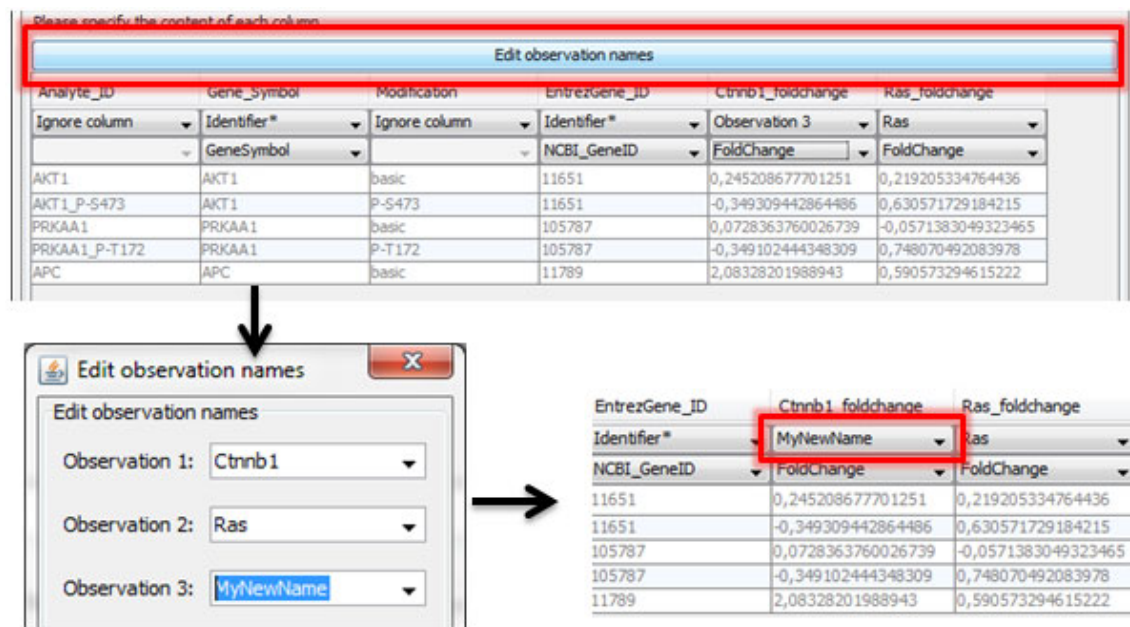


Figure 3.5: To rename observations, click the large "Edit observation names" button. In the upcoming dialog, you may enter new names for each observation. The new name will become effective immediately.



## 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. 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 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 " $\geq$ " (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. One 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.

## 4 Single dataset analysis examples

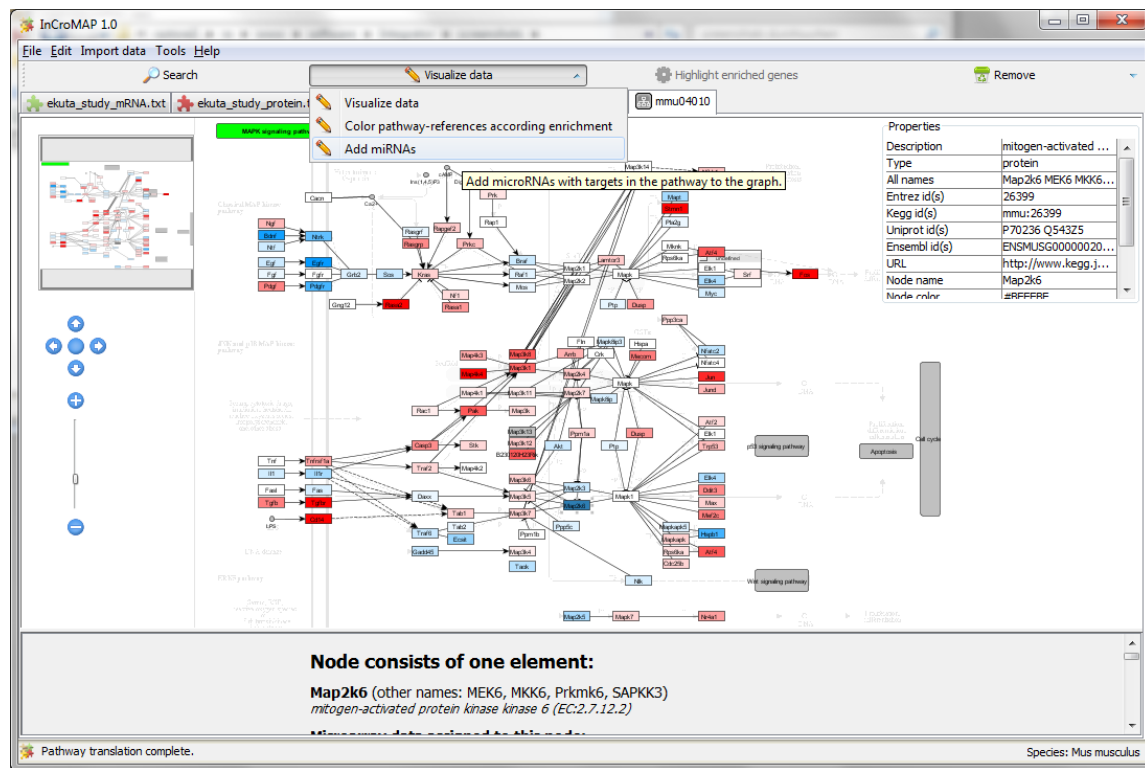


Figure 4.1: Example of the "MAPK signaling pathway" with visualized mRNA fold-changes (encoded as node-colors). The "Map2k6" gene-node has been selected and the bottom shows a detail panel, containing detailed information about this gene and all associated expression data.

### 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 above the suggested "High-confidence" threshold of the coresponding 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 is available. For more information on the upcoming miRNA target database selection dialog, see Section 4.3.

## 4.5 Genome region plot (DNA methylation only)

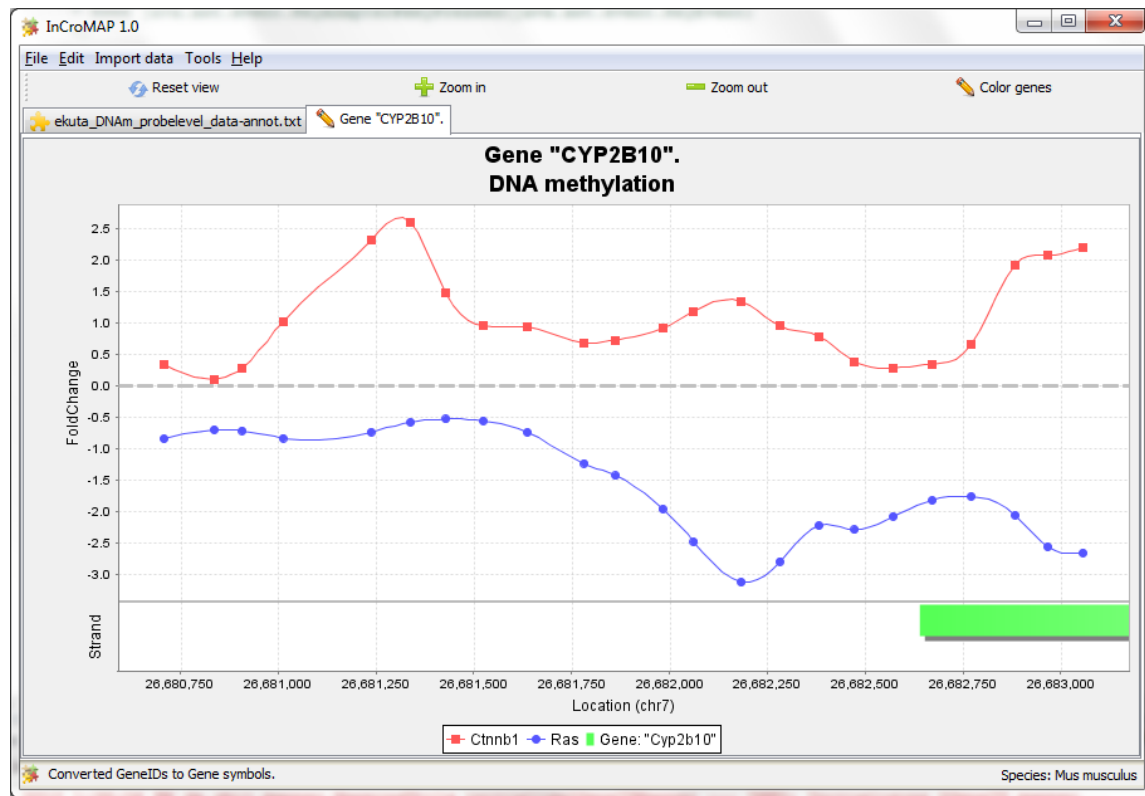


Figure 4.2: XY-plot of DNA methylation data in the promoter region of the Cyp2b10 gene.

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 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 lets 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 combo-box 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.

## 4.6 Metabolic overview (mRNA only)

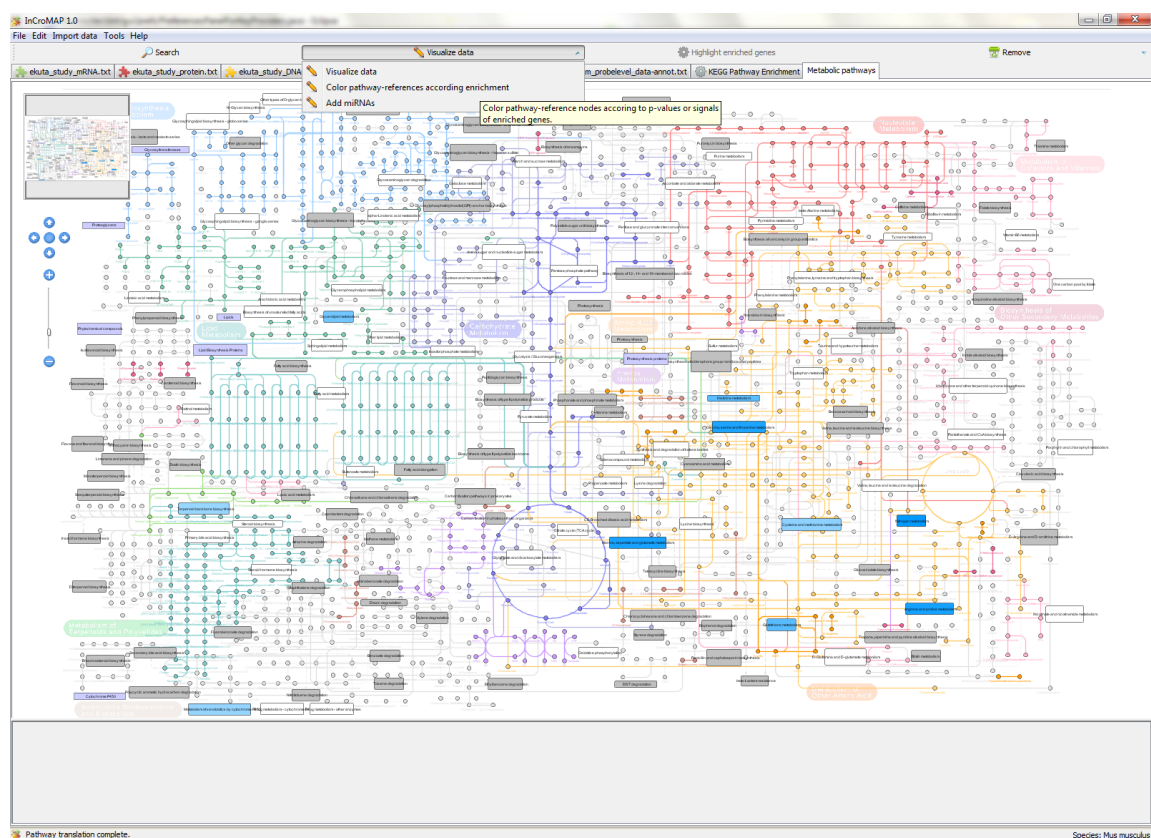


Figure 4.3: InCroMAP visualization of a metabolic overview pathway. All other referenced pathways (rectangle shaped nodes) are colored according to their enrichment p-value (i.e., how significant they are altered in input dataset). The darker-blue they are, the lower the p-value. White means  $p - value \geq 0.05$  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

TODO: TEXT SCHREIBEN

### 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 5.1). 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 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 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

## 5.2 Data pairing

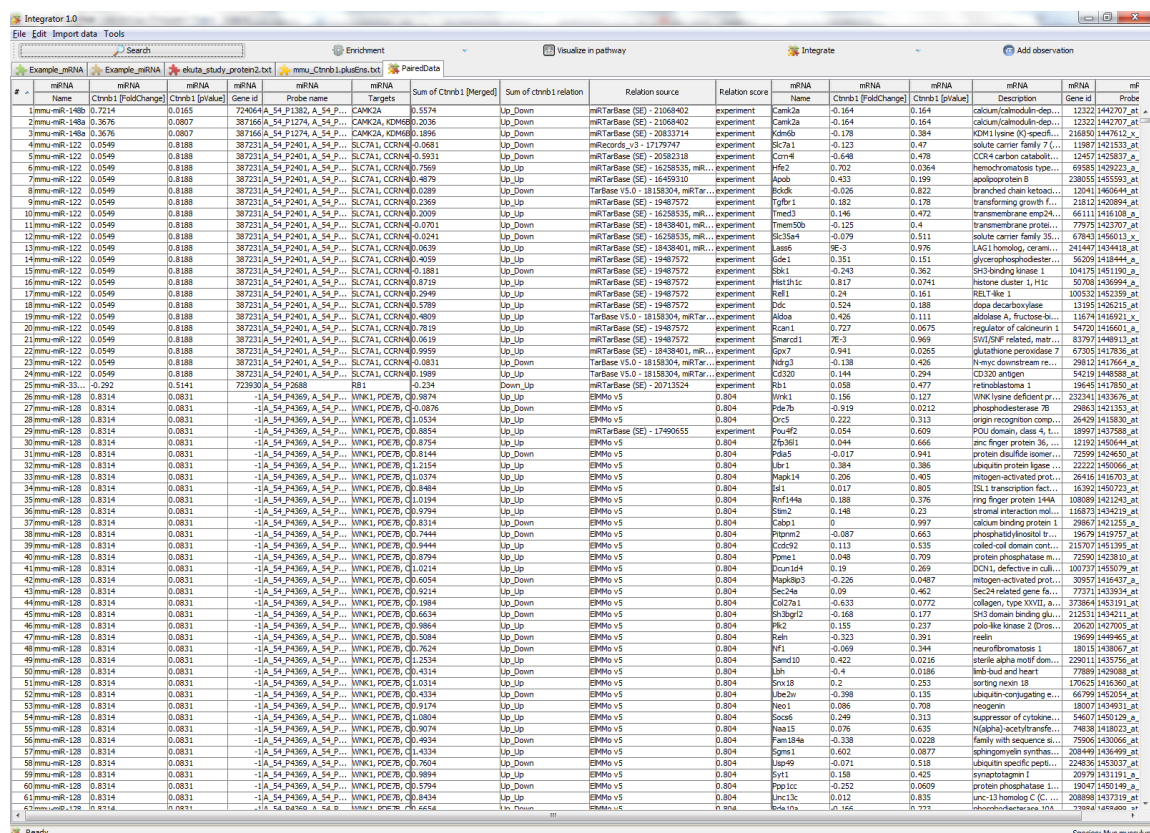


Figure 5.1: Example of the cross-platform "Data pairing" analysis method. Here: an integrated mRNA and microRNA analysis. On the left part of the figure, a micro RNA dataset is shown with expression p-values and fold-changes for each miRNA. In the middle, the target relationship for each miRNA is shown (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).



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

### 5.4 Gene-based tabular integration of data from multiple platforms

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

- F3 Knopf bei suche - 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

TODO: Create a ZIP for download, including start scripts.

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



## 6 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 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/>.

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

TODO: Warum keine miRNA annotiert? (o.) =  $\neq$  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.