InCroMAP: Integrated analysis of cross-platform microarray and pathway data

Clemens Wrzodek [1,*], Johannes Eichner [1] and Andreas Zell [1,*]

[1] Center for Bioinformatics Tuebingen (ZBIT), University of Tuebingen, 72076 Tübingen, Germany

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Summary: Microarrays are commonly used to detect changes in gene expression between different biological samples. Therefore, many analysis tools are available that offer visualizations, statistical analysis and more sophisticated analysis methods. Most of those tools are designed specifically for messenger RNA microarrays. But today, more and more different microarray platforms are available. Changes in DNA methylation, MicroRNA expression or even protein phosphorylation states can be detected with specialized arrays. For these microarray technologies, the amount of available tools is very little, compared to mRNA analysis tools. Especially, a joint analysis of different microarray platforms that have been employed on the same set of biological samples is hardly supported by most microarray analysis tools.

We here present InCroMAP, a high-level microarray data analysis tools that can be used to perform single and integrated microarray data analysis. Currently, InCroMAP supports mRNA, microRNA, DNA methylation and protein modification datasets. Several methods are offered that allow for an integrated analysis of those platforms. The available features of InCroMAP range from visualization of DNA methylation data, over annotation of microRNA targets and integrated gene set enrichment analysis to a joint visualization of all platforms in one pathway.

Availability: InCroMAP is freely available as Java$^{\text{TM}}$ application at www.cogsys.cs.uni-tuebingen.de/software/InCroMAP.

Contact: clemens.wrzodek@uni-tuebingen.de

1 INTRODUCTION

Typical workflows for the analysis of microarray data involve several steps. The procedures usually begin with the preparation of samples and arrays, followed by conduction of the experiments up to scanning the array in order to obtain the initial raw data in the computer. Depending on the actual array, several quality control and low-level data analysis steps are now performed *in silico*. Common workflows mostly include normalization, annotation of gene identifiers and calculation of statistical comparison values (such as p-values, fold changes or log ratios). Most of these tasks are often performed in R, a statistical programming language (www.r-project.org), or derived applications with a graphical user interface (e.g., MultiExperiment Viewer – Saeed *et al.*, 2006, or Mayday – Dietzsch *et al.*, 2006). These two mentioned tools also bridge the gap to more sophisticated array analysis and visualization methods. But typically, the processed datasets can now be used in various high-level data analysis tools for further evaluation and data mining. A popular example is the commercial Ingenuity pathways analysis software (www.ingenuity.com), which links processed microarray datasets with pathway analysis.

But most of these high-level analysis tools are specialized on single platform analysis. Only a few approaches and very little tools are available for an integrated analysis of high-throughput data from heterogenous platforms. Furthermore, not many software tools are freely available that offer suitable and easy-to-use analysis and visualization techniques for microarray platforms, other than mRNA expression.

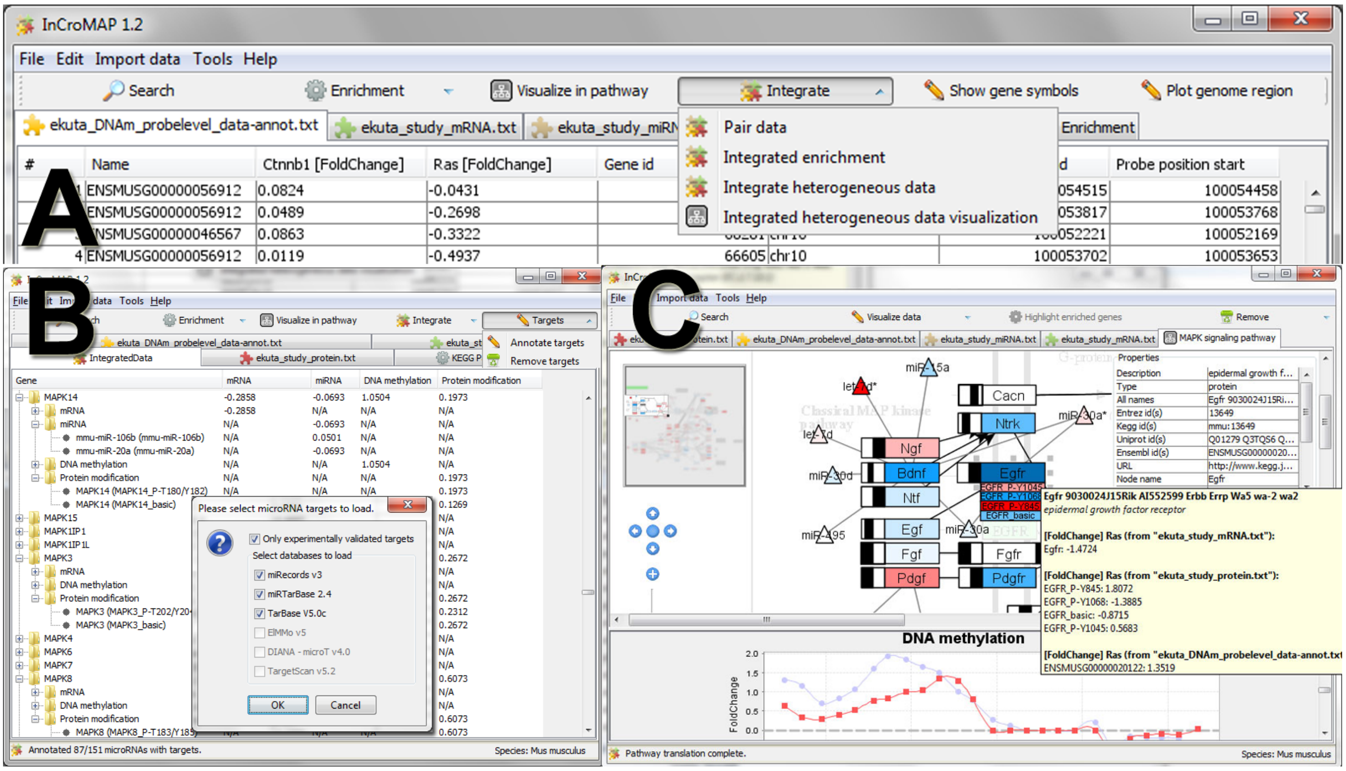
Therefore, we developed InCroMAP, an easy-to-use and interactive application with a graphical user interface that is specialized on an integrated analysis of cross-platform microarray and pathway data. InCroMAP supports DNA methylation, messenger RNA, microRNA and protein modification datasets. Besides these platforms, it is possible to import any platform that contains expression values that can somehow be assigned to genes. A special emphasis has been put on the usability of the application, hence all required files, e.g., for mapping gene identifiers to gene symbols, annotating mRNA targets to micoRNAs or pathways to visualize are either directly included in the application or downloaded dynamically in the background.

2 RESULTS

2.1 Integration of heterogeneous platforms

To integrated data from multiple platforms, a common denominator must be established. The vast majority of all data is somehow associated to genes hence, integration of multiple data types is performed by mapping each probe to a gene. This procedure is straightforward for protein or mRNA datasets. DNA methylation datasets are region-based and can be mapped on genes by defining a window upstream and downstream of each gene's transcription start site. InCroMAP proposes a window of $-2,000$ and $+500$ bps as default region, but users may change these values. Commonly, mRNA and protein data consists of protein coding genes and the relation between DNA methylation and mRNA is of regulatory nature. Therefore, integration of microRNA data is performed by annotating the genes of the mRNA targets to each microRNA. For this task, the user can choose between three microRNA target databases that contain experimentally verified targets and three databases with predicted targets (databases reviewed in Alexiou *et al.*, 2009).

[*] to whom correspondence should be addressed



#	Name	Ctnnb1 [FoldChange]	Ras [FoldChange]	Gene id			d	Probe position start
1	ENSMUSG00000056912	0.0824	-0.0431				054515	100054458
	ENSMUSG00000056912	0.0489	-0.2698				053817	100053768
	NSMUSG00000046567	0.0863	-0.3322				052221	100052169
4	ENSMUSG00000056912	0.0119	-0.4937		66605	chr10	100053702	100053653

Gene	mRNA	miRNA	DNA methylation	Protein modification
MAPK14	-0.2858	-0.0693	1.0504	0.1973
mRNA	-0.2858	N/A	N/A	N/A
miRNA	N/A	-0.0693	N/A	N/A
mmu-miR-106b (mmu-miR-106b)	N/A	0.0501	N/A	N/A
mmu-miR-20a (mmu-miR-20a)	N/A	-0.0693	N/A	N/A
DNA methylation	N/A	N/A	1.0504	N/A
Protein modification	N/A	N/A	N/A	0.1973
MAPK14 (MAPK14_P-T180/Y182)	N/A	N/A	N/A	0.1973
MAPK14 (MAPK14_basic)				0.1269
MAPK15				N/A
MAPK1IP1				N/A
MAPK1IP1L				N/A
MAPK3				0.2672
mRNA				N/A
DNA methylation				N/A
Protein modification				0.2672
MAPK3 (MAPK3_P-T202/Y20				0.2312
MAPK3 (MAPK3_basic)				0.2672
MAPK4				N/A
MAPK6				N/A
MAPK7				N/A
MAPK8				0.6073
mRNA				N/A
DNA methylation				N/A
Protein modification				0.6073
MAPK8 (MAPK8_P-T183/Y185)	N/A	N/A	N/A	0.6073

Please select microRNA targets to load.

☑ Only experimentally validated targets

Select databases to load

☑ miRecords v3
☑ miRTarBase 2.4
☑ TarBase V5.0c
☐ ElMMo v5
☐ DIANA - microT v4.0
☐ TargetScan v5.2

Annotated 87/151 microRNAs with targets.

Species: Mus musculus

Properties	
Description	epidermal growth f...
Type	protein
All names	Egfr 9030024J15Ri...
Entrez id(s)	13649
Kegg id(s)	mmu:13649
Uniprot id(s)	Q01279 Q3TQS6 Q...
Ensembl id(s)	ENSMUSG00000020...
URL	http://www.kegg.j...
Node name	Egfr

Egfr 9030024J15Rik AI552599 Erbb Errp Wa5 wa-2 wa2
epidermal growth factor receptor

[FoldChange] Ras (from "ekuta_study_mRNA.txt"):
Egfr: -1.4724

[FoldChange] Ras (from "ekuta_study_protein.txt"):
EGFR_P-Y845: 1.8072
EGFR_P-Y1068: -1.3885
EGFR_basic: -0.8715
EGFR_P-Y1045: 0.5683

[FoldChange] Ras (from "ekuta_DNAm_probelevel_data-annot.txt
ENSMUSG00000020122: 1.3519

Pathway translation complete.

Species: Mus musculus

Fig. 1. Different views of InCroMAP. A) the popup menu shows different methods that are provided for a joint analysis of heterogeneous microarray platforms. B) MicroRNA datasets can be annotated with three experimental and three predicted miRNA target databases directly from within the application. In the background the result of the 'integrate heterogeneous data' procedure is shown. C) Integrated pathway-based visualization of heterogenous microarray datasets allows to visualize up to four different platforms in a single pathway (excerpt from the 'MAPK signaling' pathway). Pathway nodes can be selected to get more detailed information, including various plots for all assigned expression values from every platform (DNA methylation in the promoter region of *Egfr* is depicted in the bottom part of C).

2.2 Integrated data analysis

A first approach to integratively investigate data from any two platforms is the 'data pairing' procedure. This procedure shows two datasets next to each other hence, simplifying common lookup task such as investigating the effect of a differentially regulated mRNA on protein level. Further, this view is especially suitable to inspect the effect of microRNA expression on target mRNAs. An arbitrary amount of platforms can be inspected, using the 'integrate heterogenous data' procedure. To keep the clarity, only the most relevant information, i.e., the expression values (as fold changes or p-values) are shown. Therefore, one row is created for each gene and one column for each platform. This procedure requires mostly summarization of multiple values, which user's can control by picking mean, median, maximum, etc. A special representation is provided that integrated a tree in a table to allow expanding the node of each gene to show all platforms with associated values. If these are expanded more detailed information, such as the single probes or microRNAs targeting this gene's mRNA is displayed. A popular method for a generic analysis of expression data is performing a gene set enrichment. We have extended this procedure to an integrated gene set enrichment that is able to perform enrichments across multiple platforms. The user can choose the datasets and thresholds for each dataset to calculate a p-value, using a hypergeometric test for each predefined gene set (see Backes *et al.*, 2007 for more information). InCroMAP supports gene sets from the KEGG PATHWAY database, Gene Ontology and any gene set from the molecular signatures database (www.broadinstitute.org/gsea/msigdb/).

Visualization of expression data in pathways is supported for each single platform. Furthermore, InCroMAP allows for integratively visualizing data from all platforms in a single pathway. Therefore, node color is changed according to mRNA expression and small boxes are added and colored according to each protein modification's expression value. MicroRNAs are added as small colored triangles to the graph and connected to their targets with novel edges. DNA methylation data is indicated with a black bar that shows the maximum differential peak in each gene's promoter (stretching from the middle to the left to indicate hypomethylation and to the right for hypermethylation). This is an interactive graph, therefore allowing users to modify the layout and selecting nodes to get more detailed information and plots about all associated expression values.

Besides those integrated analysis methods, InCroMAP allows plotting region-based DNA methylation data in a genome plot with boxes for gene bodies, which in turn can be colored, e.g., according to mRNA expression. Further, all enrichments can also be performed on any single dataset, which is straightforward for mRNA or protein datasets, but implementations that can also handle DNA methylation or microRNA data are less common.

ACKNOWLEDGEMENT

We gratefully acknowledge contributions from Andreas Dräger and Finja Büchel, as well as the whole MARCAR consortium.

Funding: The research leading to these results has received funding from the Innovative Medicine Initiative Joint Undertaking (IMI JU) under grant agreement nr. 115001 (MARCAR project).

Conflict of Interest: none declared.

REFERENCES

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.

Backes, C., Keller, A., Kuentzer, J., Kneissl, B., Comtesse, N., Elnakady, Y. A., Müller, R., Meese, E., and Lenhof, H.-P. (2007). GeneTrail–advanced gene set enrichment analysis. *Nucleic Acids Res*, **35**(Web Server issue), W186–W192.

Dietzsch, J., Gehlenborg, N., and Nieselt, K. (2006). Mayday–a microarray data analysis workbench. *Bioinformatics*, **22**(8), 1010–1012.

Saeed, A. I., Bhagabati, N. K., Braisted, J. C., Liang, W., Sharov, V., Howe, E. A., Li, J., Thiagarajan, M., White, J. A., and Quackenbush, J. (2006). TM4 microarray software suite. *Methods Enzymol*, **411**, 134–193.