BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 23 2012, pages 3021–3026
doi:10.1093/bioinformatics/bts583

Genome analysis

Advance Access publication October 9, 2012

Pathway-based visualization of cross-platform microarray datasets

Clemens Wrzodek*, Johannes Eichner and Andreas Zell*

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

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Traditionally, microarrays were almost exclusively used for the genome-wide analysis of differential gene expression. But nowadays, their scope of application has been extended to various genomic features, such as microRNAs (miRNAs), proteins and DNA methylation (DNAm). Most available methods for the visualization of these datasets are focused on individual platforms and are not capable of integratively visualizing multiple microarray datasets from cross-platform studies. Above all, there is a demand for methods that can visualize genomic features that are not directly linked to protein-coding genes, such as regulatory RNAs (e.g. miRNAs) and epigenetic alterations (e.g. DNAm), in a pathway-centred manner.

Results: We present a novel pathway-based visualization method that is especially suitable for the visualization of high-throughput datasets from multiple different microarray platforms that were used for the analysis of diverse genomic features in the same set of biological samples. The proposed methodology includes concepts for linking DNAm and miRNA expression datasets to canonical signalling and metabolic pathways. We further point out strategies for displaying data from multiple proteins and protein modifications corresponding to the same gene. Ultimately, we show how data from four distinct platform types (messenger RNA, miRNA, protein and DNAm arrays) can be integratively visualized in the context of canonical pathways.

Availability: The described method is implemented as part of the InCroMAP application that is freely available at www.cogsys.cs.uni-tuebingen.de/software/InCroMAP.

Contact: clemens.wrzodek@uni-tuebingen.de or andreas.zell@uni-tuebingen.de

Received on June 5, 2012; revised on September 5, 2012; accepted on September 21, 2012

1 INTRODUCTION

The first generation of microarray platforms was developed as a high-throughput technique for profiling the transcriptome of diverse biological systems (i.e. cells, organs or organisms) under various experimental conditions (Schena *et al*., 1995; Golub *et al*., 1999). To date, a plethora of different microarray platforms are readily available. These include gene-centred platforms that rely on current genome annotations and unbiased tiling arrays that interrogate large non-repetitive regions of the genome. Diverse types of platforms have been specifically designed for the interrogation of different genomic features, ranging from messenger RNA (mRNA) or microRNA (miRNA) transcripts, through proteins or protein modifications, to relevant functional elements, such as exons, single-nucleotide polymorphisms or promoters (Hoheisel, 2006). In addition to arrays serving for the quantification of global gene expression on the RNA or protein level, also epigenetic modifications, such as DNA methylation (DNAm), can be monitored on a genome-wide level using microarray technology (Schumacher *et al*., 2006).

Microarray datasets are typically not only inspected on the level of individual genes, but rather differential gene expression in the context of groups of functionally related genes. For instance, the overrepresentation of genes related to the same biological process or involved in the same pathway is investigated by using statistical tests (e.g. Fisher test or hypergeometric test) or gene set enrichment analysis (Subramanian *et al*., 2005). Traditional overrepresentation analysis methods require fixed sets of up- or downregulated genes that are usually defined with fold-change and/or *P*-value cut-offs. Modern gene set enrichment analysis-like methods, in contrast, do not require cut-offs to detect sets of functionally related genes showing a trend towards differential gene expression (Markowetz, 2010).

As a result of pathway enrichment, a list of significant pathways that contain genes with strong differential expression is returned. To obtain more details about this result (e.g. which genes are up- or downregulated within a pathway), suitable visualization methods are required.

Several tools exist for the visual inspection of datasets from individual platforms (see Gehlenborg *et al*., 2010, for some examples). As the first microarrays were mostly limited to mRNA transcripts, the majority of these visualization tools are still focused on mRNA datasets. However, the current inventory of publicly available tools, which are capable of integrating and jointly visualizing data from multiple microarray platforms, is still limited.

Although for genomic features, which are not directly linked to genes (such as DNAm or single-nucleotide polymorphism data), region-based visualization methods [e.g. the UCSC genome browser (Kent *et al*., 2002)] are commonly used, data from gene-centred genomic features (mRNAs, proteins, . . .) are often visualized using network-based visualization methods [e.g. Cytoscape (Cline *et al*., 2007) or KEGG Atlas (Okuda *et al*., 2008)]. Although the former approach may provide more detailed information about individual genes or genomic loci, conclusions about higher order mechanisms can often be more easily drawn from a pathway-based view of the data. Thus, we propose to first focus on pathways relevant to the studied phenomenon, which can be identified using enrichment analysis methods.

*To whom correspondence should be addressed.

© The Author 2012. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com

Subsequently, network hubs or key regulators can be inspected in more detail using complementary region-based visualization methods.

Here, we introduce a method for integrated pathway-centred visualization of datasets, generated from the same biological samples using different microarray platforms, which monitor complementary genomic and epigenomic features.

In recent years, diverse tools were developed, which are specialized in pathway analysis (e.g. Ingenuity) or pathway visualization. Some of these tools offer visualizing experimental data in a pathway [e.g. KegArray (Kanehisa *et al*., 2006), Pathline (Meyer *et al*., 2010), GenMAPP (Salomonis *et al*., 2007) or MGV (Symons and Nieselt, 2011)]. For this purpose, the experimental data are typically mapped to a colour gradient and displayed in the background colour of the pathway nodes. GenMAPP or MGV even have capabilities to display multiple colours in a single node (e.g. for the visualization of time-series experiments). MGV goes one step further and offers additional features to put profile plots or heatmaps inside nodes. However, all of these tools are not able to handle data from genomic features that have no direct reference to the genes in a pathway (e.g. miRNAs or genomic regions). Furthermore, none of these tools offers viable solutions that are tailored for the integration of multiple datasets obtained from heterogeneous microarray platforms.

2 METHODS

Before visualization, the microarray datasets of interest have to be preprocessed and annotated using platform-specific workflows (Smyth, 2004; López-Romero, 2011). These workflows usually involve (i) the quality control of the raw data (Kauffmann *et al*., 2009), (2) the data normalization to correct for background noise and experimental variation (Lim *et al*., 2007) and (iii) the mapping of probes to genes or genomic regions. After these preprocessing steps, the microarray data have to be exported in tabular format. These tables have to contain the following two types of columns: (i) annotation columns, containing probe or probeset IDs (e.g. Affymetrix IDs) and database IDs of the corresponding genes (e.g. Ensembl or Entrez IDs) and (ii) data columns containing either fold-changes or *P*-values resulting from basic statistical analysis of the microarray data.

The pathway data are automatically imported from the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa *et al*., 2006). In the KEGG PATHWAY database, each pathway map is available for download as a document in the KEGG Markup Language (KGML) which is internally converted into a graph structure by InCroMAP. To overcome limitations of the KGML format, one can create an overlay graph that shows the original KEGG pathway image in the background, which may provide the user with additional information about cellular structure and compartmentalization. Relevant pathways can be deduced by a gene set enrichment analysis or simply by manually choosing any pathway to visualize.

After importing the data, the KEGG pathway nodes, which correspond to genes or gene families, can be overlaid with fold-changes measured on mRNA or protein expression level. Furthermore, DNAm changes observed in the proximal promoter regions can be visualized. In the final step, additional nodes corresponding to miRNAs can be added to the pathway and coloured according to the expression changes measured in the underlying experiment. See Figure 1 for an illustration of all the aforementioned visualization steps.

2.1 Pathway visualization

The basic prerequisite for generating a pathway-based visualization is visualizing the pathway itself. For this purpose, we are using KEGGtranslator (see Wrzodek *et al*., 2011), which performs a basic conversion of the KEGG KGML documents to GraphML. In short, KEGGtranslator converts all KGML entries to nodes and all relations to edges. Compounds are depicted by small circular nodes, genes or proteins are rectangular-shaped nodes and group nodes are used to illustrate complexes or families. The original layout, given in the KGML document, is used to position all nodes. Then, all nodes are annotated with diverse identifiers [e.g. Entrez Gene identifiers (Maglott *et al*., 2005)], descriptions and further information. The resulting document provides the basis for the subsequently generated visualizations.

At least for some pathways, the KGML document available at KEGG does not contain all information that is depicted in the corresponding pathway map. Thus, an overlay graph can be generated which contains the original pathway map as a static transparent image in the background of the interactive graph plot (Fig. 1b). By using the given x and y coordinates together with height and width of each interactive node, it is possible to match each node exactly on the background pathway picture. However, to allow a clear distinction between the actual pathway in the foreground and the pathway background picture, the background picture should be brightened by at least 70%. Owing to this optional feature, additional information on cellular structure (e.g. schematic drawings of receptors involved in cell signalling) can be maintained.

2.2 Visualization of mRNA expression data

As mRNA expression data are typically available for the whole genome and thus also for the majority of nodes in a particular KEGG pathway, these data are displayed in the background colour of the nodes.

As input, our method requires preprocessed mRNA datasets with annotation columns (e.g. gene identifiers) and data columns, which are referred to as *observations*. In this context, observations can be any statistical significance (e.g. *P*-values) or comparative measure (e.g. fold-changes or log-ratios).

Next, these data have to be broken down to a single value for each pathway node, which then determines the colour of the node. As single nodes can represent multiple genes in KEGG, the intensities measured by probes, corresponding to the same node, have to be summarized. To this end, either the mean or median is calculated across these probes, or the probe with the strongest or most significant signal (i.e. min *P*-value or max $|fold\text{-}change|$) is decisive for colouring the node. Among others, the InCroMAP application provides all of these summarization methods.

To visualize fold-changes or log-ratios in the context of pathways, a colour gradient ranging from blue through white to red is used to illustrate down- and upregulation. Non-differentially expressed genes are shown in white, and pathway nodes for which no mRNA data are available are displayed in grey. If desired, *P*-values can be shown instead of fold-changes. For this purpose, we decided to map the negative logarithm of the *P*-values to a colour gradient, which leads to a more intuitive illustration of the observed significances. See Figure 1c for an example of visualized mRNA data.

2.3 Visualization of protein expression data

Visualization of protein datasets is performed by adding small boxes below pathway nodes and changing the colour of the boxes according to the corresponding protein expression data. As state-of-the-art experimental techniques (e.g. reverse-phase protein arrays, quantitative mass spectrometry) facilitate the distinction between different protein modifications (e.g. phosphorylated or acetylated forms of proteins), multiple measurements may correspond to the same gene (Pirnia *et al*., 2009; Yates *et al*., 2009). In this particular case, the expression change observed for each individual protein form is represented as a separate box below

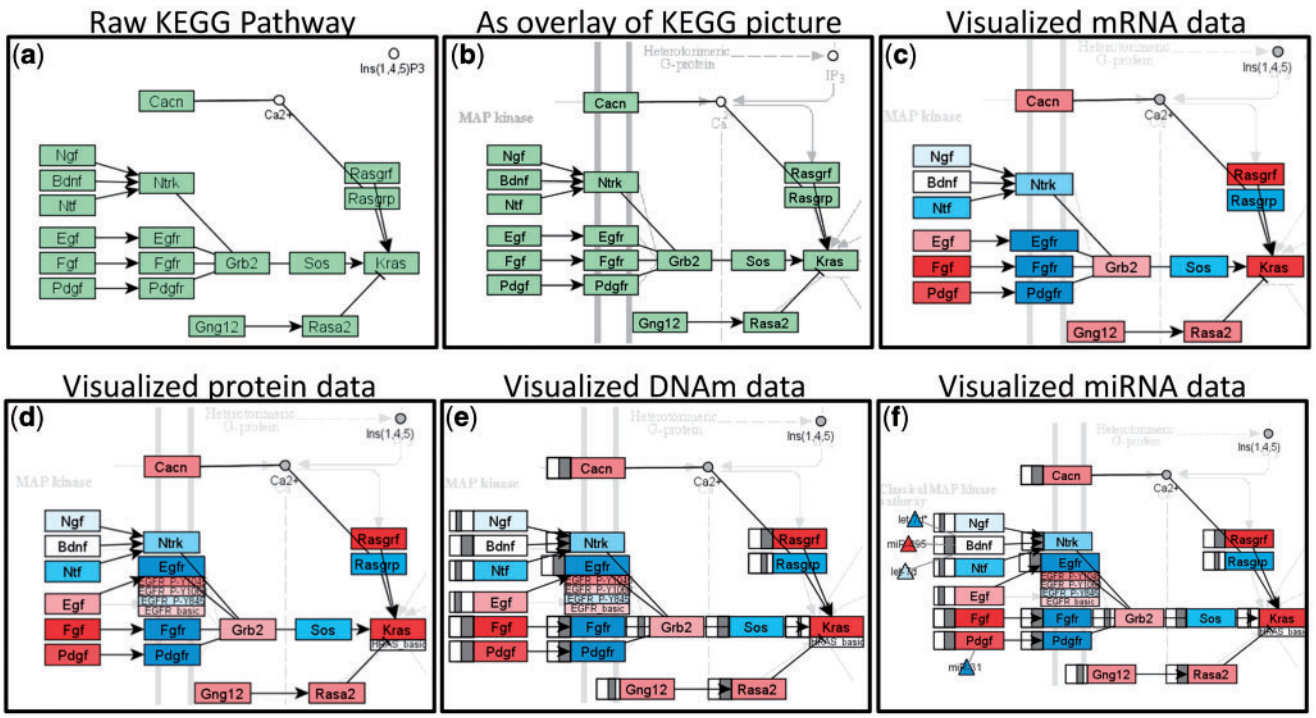

Fig. 1. These pictures show an excerpt of the KEGG 'MAPK signalling' pathway. It demonstrates how the pathway and each of the four supported platforms are visualized in the pathway. (**a**) The pathway from the KGML document visualized as a graph. (**b**) The information content of KEGG pathway visualizations can be improved by underlaying the original KEGG pathway picture. (**c**) The pathway overlaid with mRNA expression data, where red means upregulation, blue indicates downregulation and more saturated colours correspond to stronger differential expression. (**d**) Protein modification datasets in the pathway. Each box below the node *Egfr* represents a different modification of the corresponding protein, and the colour of the box reflects the corresponding fold-change. (**e**) DNAm peaks in the promoter regions. A bar from the middle of the grey box to the left represents hypomethylation, and hypermethylation is indicated by a bar to the right. The size of the bar corresponds to the size of the maximum DNAm peak. (**f**) The visualization of miRNA data by adding small triangles, representing miRNAs, and connecting them to their mRNA targets. The colour of the miRNA nodes refers to the corresponding fold-changes

the corresponding node. Each of these boxes is then labelled according to the respective protein form and coloured based on the underlying expression data, as described previously for mRNA datasets. We require protein datasets to be annotated with database identifiers referring to proteins or genes (e.g. Entrez Gene IDs), which allows us to perform a straightforward mapping to pathway nodes. For protein modification datasets, we further require a column, determining the actual modification for every protein.

2.4 Visualization of DNAm data

DNAm microarrays are similar to tiling arrays and contain many probes that can be assigned to a single gene. But visualizing the signal levels of all probes assigned to a certain gene in one pathway node would clutter the picture and reduce the clarity of it. For DNAm in gene promoters, it is most important to know whether there are methylation changes and whether the promoter is rather hyper- or hypomethylated. The methylation details can then be inspected manually later (e.g. the InCroMAP application provides a detailed XY plot of the DNAm if one clicks on a pathway node, see Fig. 2).

Therefore, the DNAm status, observed in the proximal promoter of a gene, is summarized and graphically represented by adding boxes to the left side of the pathway nodes. These boxes are drawn on a white background and contain a grey bar that stretches from the middle to the left, in the case of hypomethylation, and from the middle to the right to indicate hypermethylation. Instead of discriminating between hyper- and hypomethylation, one can simply visualize the amount of differential methylation. In this case, the grey bar would stretch from the left border of the box to the right border.

The length of the grey bar is proportional to a summary value reflecting the DNAm status of the promoter of a certain gene. This value is by default computed from the probes in a region ranging from −2000 bp upstream of the transcription start site to 500 bp downstream.

For mRNA expression arrays, the summarization of probes corresponding to the same transcript is usually achieved by the median polishing algorithm (see Wu and Irizarry, 2005), which effectively alleviates the influence of outlier probes. However, this algorithm is only applicable under the assumption that ideally the probe intensities are evenly distributed along transcripts. This assumption does not hold true for the DNAm of promoter regions, as the probe intensities are not expected to be evenly distributed along the region. Furthermore, outlier probes, corresponding to DNAm peaks, are mostly not because of probe-specific effects, but rather reflect a local change in the DNAm pattern, which may be biologically meaningful. A method to overcome this issue is the detection and visualization of maximum peaks within the DNAm data.

For peak detection, Nimblegen proposed two algorithms that are called 'windowed threshold detection' and 'second derivative peak detection'. A detailed description of these methods can be found in the user's guide of Nimblegen's SignalMap software (see NimbleGen Systems Inc., 2006). In addition to those sophisticated peak detection methods, one can also approximate peaks by using simple methods, such as picking the probe with maximum differential methylation (i.e. max |*fold-change*|). As an alternative to visualizing peaks derived from the probe-level fold-changes, one can also use a statistical approach to estimate the

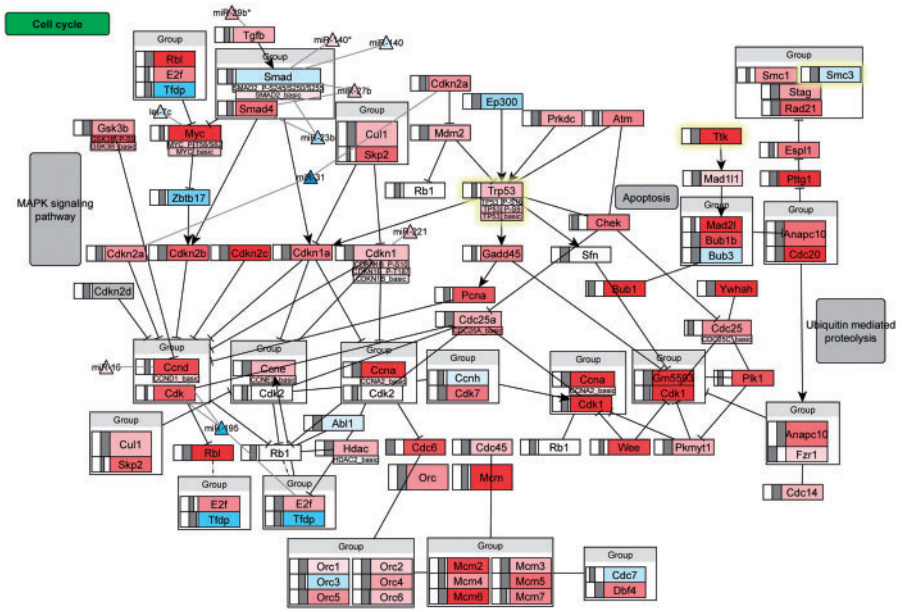

Fig. 2. Integrated visualization of datasets from four different platforms in the KEGG 'Cell Cycle' pathway. In general, colours reflect fold-changes, where red means upregulated and blue means downregulated. White indicates a fold-change of zero and darker colours correspond to stronger differential expression. The colour of each node itself reflects the mRNA fold-change, for example, *Ttk* shows a strong upregulation and *Smc3* is downregulated. Smaller boxes below nodes show the protein and protein modification expression. For example, we visualized three different forms of TRP53: a phosphorylation at the serine 15 (S15) site, which shows almost no expression change, a modification of the S9 site, which is upregulated, and the basic protein itself, which is also upregulated. The grey boxes on the left of the nodes represent the maximum DNAm fold-change peak. A bar to the left indicates hypomethylation and a bar to the right hypermethylation. In our example, *Ttk* shows a strong hypermethylation. All aforementioned genes are highlighted in yellow. miRNAs are added as small triangles to the pathway and are connected with a grey edge to their mRNA targets. The colour of the miRNA nodes reflects the fold-change, as described for the mRNA nodes

differential methylation in a promoter region. For this purpose, we propose to divide the promoter region r into segments $b_1, \ldots, b_n$, each containing a fixed number of probes. For each segment, a statistical test (e.g. ordinary or moderated t-test) can be used to detect differential methylation between two experimental groups. An overall significance value s can then be computed from the P-values $P(b_1, \ldots, P(b_n)$ using the formula $s = \frac{1}{n}\sum_{i=1}^{n} -\log(P(b_i))$. This binning approach has the advantage of being sensitive to local changes in DNAm, as promoter segments with significant changes in DNAm are highly weighted. Conversely, segments for which slight changes were observed only contribute marginally to the overall significance score.

InCroMAP, by default, approximates the peak by selecting the probe with maximum differential methylation. It further supports other options, such as the given formula s to process P-values and simple options that include mean, median, minimum and maximum. If multiple genes are associated to one node, all probes are pooled together and the summarization method is applied. By default, this will lead to a visual representation of the maximum peak contained in any gene that is represented by a node.

The purpose of this visualization technique is giving researchers a hint if and how many methylation changes are present in a gene promoter. For more details, the genomic location of individual peaks can be visualized in a DNAm profile plot (see Fig. 2). To this end, the fold-changes observed for the probes covering the promoter region of a certain gene are plotted along their genomic coordinates. This depiction is particularly useful for comparing the DNAm profiles and peaks found for different observations (i.e. sample groups).

To relate the DNAm data to data from gene-centred microarrays (e.g. mRNA expression arrays), each probe needs to be mapped to a gene. Thus, we require annotation columns containing the chromosome and the genomic position of each probe to facilitate the mapping from probes to genes.

2.5 Visualization of miRNA expression data

Visualizing miRNA datasets in the context of KEGG pathways is not straightforward, as these pathways do not contain miRNAs *a priori*. Therefore, to incorporate the data into a pathway, a connection must be established between the miRNAs and the protein-coding genes in the pathway.

A simple approach for putting miRNAs in context would be inspecting the gene locus of the miRNAs. This approach already led to some valuable results, for example, the detection of a strong co-regulation within the genomic DLK1 locus (Luk *et al.*, 2011). But this approach is not suitable for linking the miRNAs with genes in a pathway. First, genes from neighbouring loci are not generally co-regulated. Second, a pathway shows, for example, how a signal is processed within a certain organism. The genes in a pathway are usually functionally related or represent regulatory relationships. Hence, when adding new nodes to a pathway, the relation should also be made based on functional or regulatory relationships.

As the common mechanism of miRNAs involves the binding to complementary mRNA transcripts (Bartel, 2004), we propose to link miRNAs to their known target mRNAs. These target mRNAs can be obtained from diverse databases, which contain experimentally verified targets [e.g. miRecords (Xiao *et al.*, 2009), miRTarBase (Hsu *et al.*, 2011), TarBase (Papadopoulos *et al.*, 2009)] or predicted miRNA targets (reviewed in Alexiou *et al.*, 2009). We used a union of the three aforementioned experimentally verified miRNA target databases for the figures in this publication.

Based on a map containing all connections between miRNAs and their target mRNAs, the miRNAs monitored in a specific experiment can be added to a pathway of interest as small triangular nodes, which have outgoing edges to the pathway nodes corresponding to their target mRNAs. The triangular miRNA nodes are coloured according to their expression, as described previously for mRNA datasets. This leads to an integrated visualization of a pathway, overlaid with miRNA and mRNA expression data, and extended with putative miRNA–mRNA interactions. Figure 1f shows an example result of the described procedure.

3 RESULTS AND DISCUSSION

In this work, we present a novel methodology for the combined visualization of DNAm data and mRNA, (phospho-) protein and miRNA expression data in the context of canonical pathways. This methodology involves strategies for mapping the data from heterogeneous platform types to a common functional element, namely, a gene, embedded into a pathway which regulates higher order cellular functions or processes. Furthermore, we propose visualization techniques that are particularly suited for displaying quantitative data from diverse genomic features in an integrated graph plot, which represents a metabolic or signalling pathway of interest. However, generating a joint visualization of epigenomics, transcriptomics and proteomics data is challenging, as each data type has its own characteristics.

For proteomics data, we propose to map basic and modified proteins to their common source gene and draw each protein form as a separate box below the pathway node representing the source gene. Using this kind of visualization, one can, for instance, determine the activation state of a protein by comparing the colour (i.e. expression state) of the box corresponding to the basic form with the one that corresponds to the phosphorylated form.

In brief, for most mRNA expression datasets, it is sufficient to annotate the probesets with appropriate gene identifiers to link the data with the pathways of interest. For miRNA datasets, this linkage is complicated by the fact that miRNAs are normally not contained in pathways. miRNAs are known to have a regulatory function by binding to complementary mRNA transcripts (Bartel, 2004). These interactions are documented in public databases, which we use to extract interactions between miRNAs and their target mRNAs. Using this information, one can add the miRNAs to the pathways and connect them to the existing nodes by inserting edges to the corresponding targets.

The probe design of DNAm microarrays is typically not gene-centred, but is region based. Thus, in contrast to gene-centred arrays, the probes are not organized in probesets, which were designed specifically for the same gene. Instead, these arrays cover specific genomic regions of interest, for instance gene promoters. The connection to protein coding genes is usually established by defining a region around the transcription start site of a gene and assigning all probes in this region to the gene.

However, the visualization of every probe would vastly reduce the clarity of the visualization; therefore, a summary method is required. For DNAm changes, calculating a mean is often inappropriate because small peaks in a promoter region can already have a large influence on gene regulation. One of the summarization methods we consider advisable is the detection and visualization of maximum peaks, as it is sensitive to local changes in DNAm.

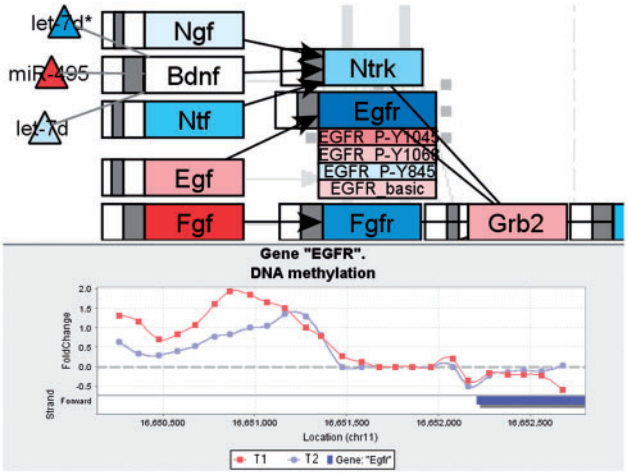

Fig. 3. Integrated visualization of the 'MAPK signalling' pathway with a detailed view of DNAm in the proximal promoter of *Egfr*. The grey DNAm bar next to the *Egfr* gene in the pathway plot stretches from the middle to the right of the box, which indicates strong hypermethylation. The length of the grey bar corresponds to the height of the maximum peak of the red curve. This curve displays the probe-level fold-changes plotted along the promoter region of *Egfr*. The light blue curve shows the DNAm profile of *Egfr*, observed in another sample group

Alternatively, if the data are properly normalized and smoothed (i.e. probe specific effects are alleviated), one can also take the probe with maximum differential methylation to approximate the peak. Nevertheless, every summarization method has its limitations. The maximum peak methods are definitely less robust against probe-specific effects, and the results may be inconclusive as a single positive peak may suggest hypermethylation of a region, while the majority of the promoter is clearly hypomethylated. On the other hand, if mean probe-level fold-changes are computed across larger regions, there is a higher chance that localized differential methylation is erased.

An example for a KEGG pathway with visualized mRNA, miRNA, protein and DNAm data can be seen in Figure 3. The visualization of multiple heterogenous datatypes cannot be performed with either a loss of information or a loss of clarity. As these integrated pathway visualizations should provide an overview, rather than a detailed listing of all possible information, we tried to keep the clarity and summarize information wherever possible.

At first glance, it always seems to be desirable putting as much information as possible into the pathway. For example, it would be possible to create small XY plots of promoter regions for every gene and add these pictures as small icons below each node. Other possibilities include the creation of heatmaps or visualization of multiple timepoints with various colours in a single node. But, especially if multiple datasets are visualized together in a pathway, it is important to summarize things and to not overload the whole picture. The main purpose of a pathway-based visualization is not to replace all other analysis and visualization methods, but rather giving a general overview that helps researchers to further analyse their data.

4 IMPLEMENTATION AND AVAILABILITY

The described methods are implemented in InCroMAP, a tool for integrated analysis of cross-platform microarray and Pathway data. InCroMAP is a Java™ application that provides an interactive, user-friendly and easy-to-use graphical user interface and is freely available under the LGPL version 3 license from www.cogsys.cs.uni-tuebingen.de/software/InCroMAP. The application can import all aforementioned data types, is able to automatically download and layout KEGG pathways and apply all described visualization methods on those pathways. The resulting graphs can be exported as JPG, GIF, TGF, GML or GraphML. Furthermore, many options are provided that control, for example, the mapping of expression values to a continuous colour gradient and allow for customization of the generated cross-platform pathway visualizations.

5 CONCLUSION

Pathway enrichment analysis is a common microarray data analysis tool to discover pathways, whose genes show significant expression changes. In most applications, these enrichments are endpoints in analysis workflows. There are some methods or applications that can show pictures of significantly altered pathways, and a few applications can even change node shapes or colours according to a single expression dataset. However, not only the amount of available microarray datasets is growing rapidly but also the number of available platforms. At present, microarray platforms exist for profiling diverse genomic features, as for instance mRNAs, miRNAs, proteins and promoter regions. However, the current repertoire of analysis and especially visualization methods that are specifically designed for the analysis of cross-platform datasets is still limited.

Here, we present a pathway-based cross-platform microarray visualization method that can, for example, be used to inspect relevant pathways detected by pathway enrichment analyses. Especially the visualization of DNAm and miRNA datasets is a feature that cannot be found in other visualization approaches. By integratively visualizing all datatypes, it helps researchers to discover potential relationships across multiple layers of gene regulation. For example, the hypomethylation of a promoter may cause the upregulation of a miRNA, which may in turn downregulate the corresponding target mRNA, whereupon a connected pathway protein could change its expression or activation state. Such complex cascades of effects, which involve multiple levels of gene regulation, can be deduced from the pathway images generated by the here presented visualization method.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge contributions from Andreas Dräger and Finja Büchel and 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 no 115001 (MARCAR project).

Conflict of Interest: none declared.

REFERENCES

Alexiou,P. *et al.* (2009) Lost in translation: an assessment and perspective for computational microRNA target identification. *Bioinformatics*, **25**, 3049–3055.

Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.

Cline,M.S. *et al.* (2007) Integration of biological networks and gene expression data using Cytoscape. *Nat. Protoc.*, **2**, 2366–2382.

Gehlenborg,N. *et al.* (2010) Visualization of omics data for systems biology. *Nat. Methods*, **7 (Suppl. 3)**, S56–S68.

Golub,T.R. *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Hoheisel,J.D. (2006) Microarray technology: beyond transcript profiling and genotype analysis. *Nat. Rev. Genet.*, **7**, 200–210.

Hsu,S.D. *et al.* (2011) miRTarBase: a database curates experimentally validated microRNA-target interactions. *Nucleic Acids Res.*, **39**, D163–D169.

Kanehisa,M. *et al.* (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, **34**, D354–D357.

Kauffmann,A. *et al.* (2009) arrayQualityMetrics—a bioconductor package for quality assessment of microarray data. *Bioinformatics*, **25**, 415–416.

Kent,W.J. *et al.* (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.

Lim,W.K. *et al.* (2007) Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks. *Bioinformatics*, **23**, i282–i288.

López-Romero,P. (2011) Pre-processing and differential expression analysis of Agilent microRNA arrays using the AgiMicroRna Bioconductor library. *BMC Genomics*, **12**, 64.

Luk,J.M. *et al.* (2011) DLK1-DIO3 genomic imprinted microRNA cluster at 14q32.2 defines a stemlike subtype of hepatocellular carcinoma associated with poor survival. *J. Biol. Chem.*, **286**, 30706–30713.

Maglott,D. *et al.* (2005) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **33**, D54–D58.

Markowetz,F. (2010) How to understand the cell by breaking it: network analysis of gene perturbation screens. *PLoS Comput. Biol.*, **6**, e1000655.

Meyer,M. *et al.* (2010) Pathline: a tool for comparative functional genomics. *Comput. Graph. Forum (Proc. EuroVis 2010)*, **29**, 1043–1052.

NimbleGen Systems Inc. (2006) *SignalMap User's Guide*. www.nimblegen.com/products/lit/signalmap1.9usersguide.pdf (22 March 2012, date last accessed).

Okuda,S. *et al.* (2008) KEGG Atlas mapping for global analysis of metabolic pathways. *Nucleic Acids Res.*, **36**, W423–W426.

Papadopoulos,G.L. *et al.* (2009) The database of experimentally supported targets: a functional update of TarBase. *Nucleic Acids Res.*, **37**, D155–D158.

Pirnia,F. *et al.* (2009) Novel functional profiling approach combining reverse phase protein microarrays and human 3-D ex vivo tissue cultures: expression of apoptosis-related proteins in human colon cancer. *Proteomics*, **9**, 3535–3548.

Salomonis,N. *et al.* (2007) GenMAPP 2: new features and resources for pathway analysis. *BMC Bioinformatics*, **8**, 217.

Schena,M. *et al.* (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.

Schumacher,A. *et al.* (2006) Microarray-based DNA methylation profiling: technology and applications. *Nucleic Acids Res.*, **34**, 528–542.

Smyth,G.K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**, Article 3.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.

Symons,S. and Nieselt,K. (2011) MGV: a generic graph viewer for comparative omics data. *Bioinformatics*, **27**, 2248–2255.

Wrzodek,C. *et al.* (2011) KEGGtranslator: visualizing and converting the KEGG PATHWAY database to various formats. *Bioinformatics*, **27**, 2314–2315.

Wu,Z. and Irizarry,R.A. (2005) Stochastic models inspired by hybridization theory for short oligonucleotide arrays. *J. Comput. Biol.*, **12**, 882–893.

Xiao,F. *et al.* (2009) miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Res.*, **37**, D105–D110.

Yates,J.R. *et al.* (2009) Proteomics by mass spectrometry: approaches, advances, and applications. *Annu. Rev. Biomed. Eng.*, **11**, 49–79.