BIOINFORMATICS

Vol. 00 no. 00 2012
Pages 1–5

Pathway-based visualization of cross-platform microarray datasets

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

Motivation: Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text.

Results: Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text

Contact: clemens.wrzodek@uni-tuebingen.de

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 (Golub *et al.*, 1999; Schena *et al.*, 1995). As these traditional gene-centered arrays were mostly limited to mRNA transcripts, the vast majority of visualization tools are still focused on mRNA datasets (e.g., Expression Profiler – Kapushesky *et al.*, 2004 or KegArray – Kanehisa *et al.*, 2006). To date, a plethora of different microarray platforms are readily available. These include gene-centered platforms which rely on current genome annotations as well as unbiased tiling arrays which 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 mRNA or miRNA transcripts, through proteins or protein modifications, to relevant functional elements such as exons, SNPs 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).

Several tools exist for the visual inspection of datasets from individual platforms (see Gehlenborg *et al.*, 2010, for some examples). However, the current inventory of publicly available tools, which are capable of integrating and jointly visualizing data from multiple microarray platforms, is still very limited. Here, we introduce a method for integrated pathway-centered visualization of datasets, generated from the same biological samples using different microarray platforms, which monitor complementary genomic and epigenomic features.

In contrast to commonly used region-based visualization methods (e.g., the UCSC genome browser – Kent *et al.*, 2002), we propose to visualize the microarray data in the context of specific signaling or metabolic pathways, which can in many cases be more easily related to the biological problem under study, than individual genes or genomic regions. In recent years, diverse tools were developed, which are specialized in pathway analysis (e.g., Ingenuity) or pathway visualization (e.g., Cytoscape – Cline *et al.*, 2007 or KEGG Atlas – Okuda *et al.*, 2008). Some of these tools offer visualizing experimental data in a pathway (e.g., KegArray – Kanehisa *et al.*, 2006, GenMAPP – Salomonis *et al.*, 2007 or MGV – Symons and Nieselt, 2011). For this purpose, the experimental data is typically mapped to a color gradient and displayed in the background color of the pathway nodes. GenMAPP or MGV even have capabilities to display multiple colors 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 which have no direct reference to the genes in a pathway (e.g., microRNAs 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 (1) 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 (3) the mapping of the probes to genes or genomic regions. After these preprocessing steps the microarray data has to be exported in tabular format. These tables have to contain two types of columns: (1) 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 (2) data columns containing either fold-changes and/or p-values resulting from basic statistical analysis of the microarray data.

*to whom correspondence should be addressed

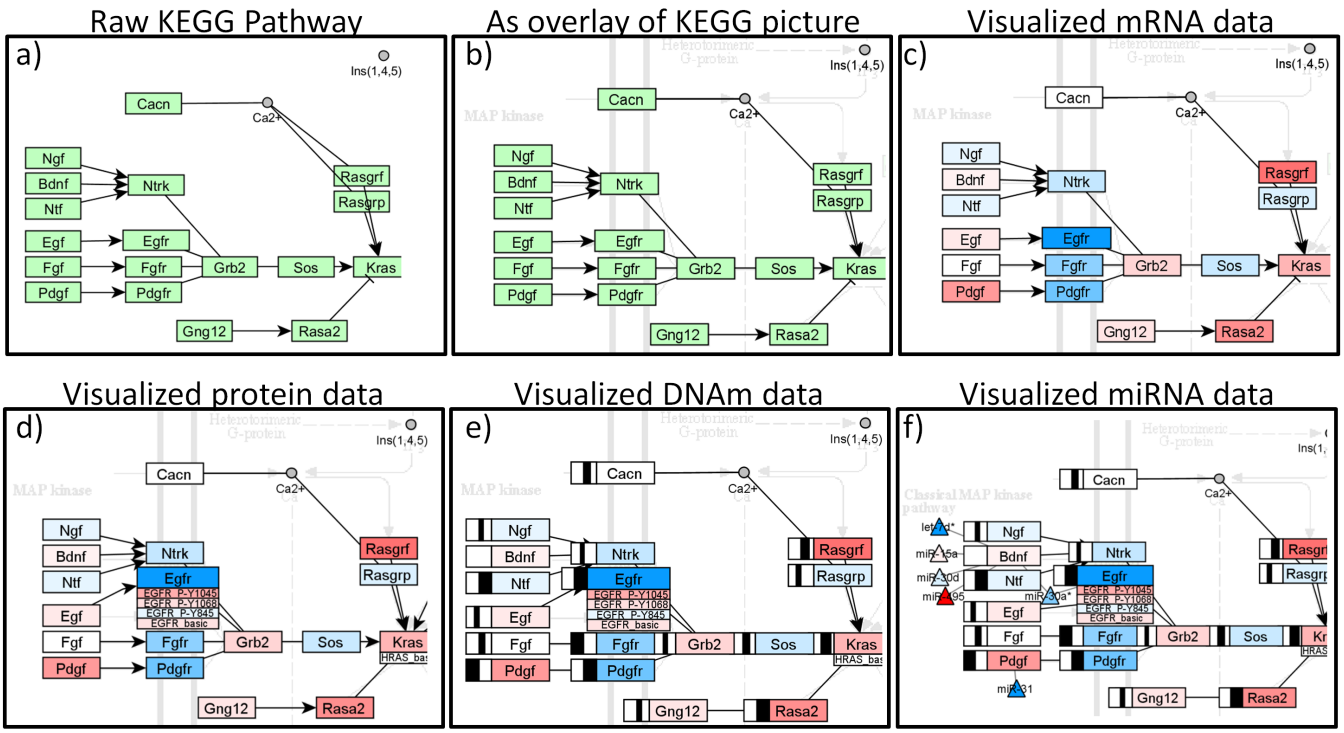

Fig. 1. These pictures show an excerpt of the KEGG MAPK signaling pathway. It is demonstrated how the pathway and each of the four supported platforms is visualized in the pathway. Picture a) shows the visualized KGML document as nodes and edges. b) demonstrates that the overall quality of KEGG pathway visualizations can be improved by underlaying the original KEGG pathway picture. c) presents the excerpt with visualized mRNA data, whereas red means upregulation, blue downregulation and lighter colors represent smaller fold-changes. d) shoes how protein modification datasets are visualized in the pathway. Each of the boxes below "Egfr" represent different modifications of this protein and the color of the box reflects the corresponding fold-change. e) illustrates DNA methylation peaks in promoter regions. A bar from the middle of the black box to the left represents hypomethylation and hypermethylation is expressed by a bar to the right. The size of the bar determines the size of the maximum DNA methylation peak. f) demonstrates the visualization of microRNA data by adding small triangles, representing microRNAs, and connecting them with grey edges to their mRNA targets. The color of the microRNA nodes determines the corresponding fold-change.

The pathway data is automatically imported from KEGG. In the KEGG PATHWAY database each pathway map is available for download as a KGML document which is internally converted into a GraphML document by InCroMAP. These GraphML files can be interactively visualized by InCroMAP using the yWorks graph viewer (TODO: Ref.). In order 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.

After uploading 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, DNA methylation 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 colored according to the expression changes measured in the underlying experiment. See Figure 1 for an illustration of all the above-mentioned 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 and to annotates all nodes with EntrezGene identifiers (Ref. zu NCBI EntrezGene). In short, KEGGtranslator converts all KGML entries to nodes and all relations to edges. Some basic errors are corrected automatically and appropriate shapes, colors and labels are inferred. Then, all nodes are annotated with diverse identifiers, 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 displayed 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 (see Figure 1b). Owing to this feature, additional information on cellular structure (e.g., schematic drawings of receptors involved in cell signaling) can be maintained. This feature also enhances the visualization of the global maps in KEGG PATHWAY, as additional

information on the organisation of subordinate pathway groups is provided by the pathway image.

2.2 Visualization of messenger-RNA expression data

As mRNA expression data is 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 color of the nodes.

As input our method requires preprocessed mRNA datasets with annotation columns (e.g., probeset and 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 corresponds to the color 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, the mean or median is calculated across probes, which were mapped to the same node, or the probe with the strongest or most significant signal (i.e., $\min p - value$ or $\max |fold - change|$) is adopted for coloring the node.

To visualize fold-changes or log-ratios in the context of pathways a color gradient ranging from blue to red is used to illustrate down- and upregulation, respectively. Non-differentially expressed genes are shown in white, and pathway nodes for which no mRNA data is available are displayed in grey. If desired *p*-values can be shown instead of fold-changes. For this purpose, we propose to map the negative logarithmized *p*-values to a color 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 color 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) (TODO: Refs), multiple measurements may correspond to the same gene. In this particular case, the expression change observed for each individual protein form is represented as a separate box below the corresponding node. Each of these boxes is then labeled according to the respective protein form and colored 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 (e.g., UniProt IDs) or genes (e.g. EntrezGene IDs), which allows us to perform a straightforward mapping to pathway nodes.

2.4 Visualization of DNA-methylation data

The DNA-methylation (DNAm) status observed in the proximal promoter of a gene is graphically represented by adding boxes to the left side of the pathway nodes. These boxes are drawn in a white box containing a black bar which stretches from the middle to the left in the case of hypomethylation and from the middle to the right to indicate hypermethylation.

The length of the black 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. Normally, mean log-ratios r are used as summary value, which can be computed from the probe-level fold-changes $x_1, ..., x_n$ using the formula $r = \frac{1}{n}\sum_{i=1}^{n}\log_2(x_i)$. Alternatively, the value of the maximum peak found in the promoter of a gene can be displayed. For peak detection Nimblegen recently proposed two algorithms which 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 the Nimblegen software SignalMap (http://www.nimblegen.com/products/lit/signalmap1.9usersguide.pdf).

The genomic location of individual peaks can be visualized in more detail in a DNAm profile plot (TODO: Auf entsprechende Figure verweisen). 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).

In order to relate the DNAm data to data from gene-centered microarrays (e.g., mRNA expression arrays), each probe has to be assigned 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 micro-RNA expression data

Visualizing microRNA (miRNA) datasets in the context of KEGG pathways is not straightforward, as these pathways do not contain miRNAs *a priori*. Therefore, in order to incorporate the data into a pathway, a connection must be established between the miRNAs and the protein-coding genes in the pathway. 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 either contain experimentally verified targets (e.g., miRecords (see Xiao *et al.*, 2009), miRTarBase (see Hsu *et al.*, 2011), TarBase (see Papadopoulos *et al.*, 2009)) or predicted miRNA targets (reviewed in Alexiou *et al.*, 2009) (TODO: Beispiele hinzufgen). We used a union of the three mentioned experimentally verified miRNA target databases for the example 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 colored 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.

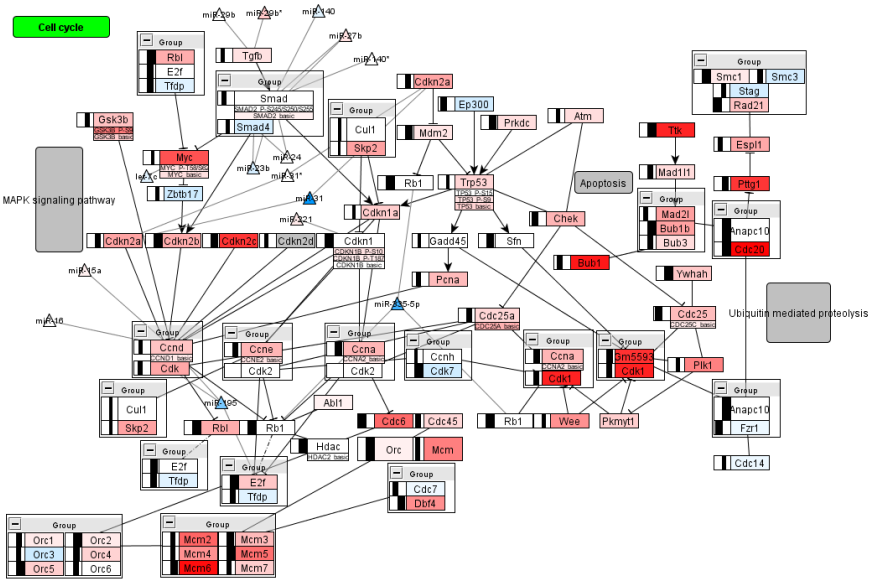

Fig. 2. Integrated visualization of datasets from four different platforms in the KEGG Cell Cycle pathway. The colors reflect the fold-changes, whereas red means upregulated and blue means downregulated. White corresponds to a fold-change of zero and darker colors represent stronger differential expression. The color of each node itself reflects the mRNA fold-change, e.g., "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 black boxes left of the nodes represent the maximum DNA methylation fold-change peak. A bar to the left represents hypomethylation and a bar to the right hypermethylation. In out example, "Ttk" shows a strong hypermethylation. MicroRNAs are added as small triangles to the pathway and connected with a grey edge to their mRNA targets. The color of the microRNA nodes reflects the fold-change, as described for the mRNA nodes.

3 RESULTS AND DISCUSSION

In this work, we present a novel methodology for the combined visualization of DNA methylation data, as well as 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.

Visualizing single datasets in pathways is a good first step that is also included in other methods and applications. Actually, any dataset with gene identifiers and expression values can be somehow visualized by changing node colors or other aspects of pathway nodes. Hence, it is not very hard to visualize mRNA datasets. But visualization of phosphoprotein, microRNA or DNA methylation datasets is not easy, because each datatype has its own characteristics.

For phosphoproteins, one destroys the data if expression values for differen protein modifications (i.e. phosphoforms) are merged. Hence, we add a separate box with its own visualized expression value for every protein and protein modification to the pathway graph.

For MicroRNAs, the general method with adding gene identifiers to datasets and mapping to the pathways does not work at all, because miRNAs are not contained in pathways. We visualize miRNAs together with an additional information: the miRNA target relation. By identifying protein coding genes, whose mRNAs are targets of miRNAs, one can add the miRNAs to the pathways and connect them to the protein coding genes by inserting edges to the corresponding targets.

DNA methylationmicroarray datasets are region-based and thus contain a huge number of probes and information. The connection to protein coding genes is usually performed by defining a region around the transcription start site (TSS) of a gene and assigning all probes in this region to the gene.

For example, with this procedure one can define that all probes within -2000 and +500 bps of a TSS are most important for the regulation of this gene's expression. But then, it is not recommended to visualize all this information in the pathway. For example, it would be possible to create small XY plots of the promoter region for every gene and add these pictures as small icons below each node. But this would clutter the picture and destroy the clarity of it. Thus, we concluded that it is most important to know if there are methylation changes in a promoter and maybe also if the gene 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 DNA methylation if one clicks on a pathway node). But the most important information in the first place is, if there are methylation changes at all. And this can be expressed with various summarization methods and a small black bar that grows with the amount of significant methylation changes in a gene promoter.

An example for a KEGG pathway with visualized mRNA, miRNA, protein and DNA methylation data can be seen in Figure 2. The visualization of multiple heterogenous datatypes can not be performed with either a loss of information or a loss of clarity. Since 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.

4 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 colors 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. Thus, today we have datasets that not only consist of mRNA data, but also DNA methylation, miRNA, protein or even more different levels. Analysis and especially visualization methods for such cross-platform datasets are very rare.

Here, we present a pathway-based cross-platform microarray visualization method that is perfectly suitable, e.g., as visualization method after pathway enrichment analyses. Especially the visualization of DNA methylation and miRNA datasets is a feature that can not be found in other visualization approaches. By integratively visualizing all datatypes, it helps researchers to discover potential relationships across multiple layers. For example, a miRNA can be upregulated and the corresponding mRNA target downregulated, whereupon a connected pathway protein could also be downregulated in protein expression. This effect, which involves information about pathway relations and miRNA targets, as well as miRNA, mRNA and protein expression, would be directly visible from the here presented visualization method.

ACKNOWLEDGEMENT

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

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.

Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**(2), 281–297.

Cline, M. S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., Workman, C., Christmas, R., Avila-Campilo, I., Creech, M., Gross, B., Hanspers, K., Isserlin, R., Kelley, R., Killcoyne, S., Lotia, S., Maere, S., Morris, J., Ono, K., Pavlovic, V., Pico, A. R., Vailaya, A., Wang, P.-L., Adler, A., Conklin, B. R., Hood, L., Kuiper, M., Sander, C., Schmulevich, I., Schwikowski, B., Warner, G. J., Ideker, T., and Bader, G. D. (2007). Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc*, **2**(10), 2366–2382.

Gehlenborg, N., O'Donoghue, S. I., Baliga, N. S., Goesmann, A., Hibbs, M. A., Kitano, H., Kohlbacher, O., Neuweger, H., Schneider, R., Tenenbaum, D., and Gavin, A.-C. (2010). Visualization of omics data for systems biology. *Nat Methods*, **7**(3 Suppl), S56–S68.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**(5439), 531–537.

Hoheisel, J. D. (2006). Microarray technology: beyond transcript profiling and genotype analysis. *Nat Rev Genet*, **7**(3), 200–210.

Hsu, S. D., Lin, F. M., Wu, W. Y., Liang, C., Huang, W. C., Chan, W. L., Tsai, W. T., Chen, G. Z., Lee, C. J., Chiu, C. M., Chien, C. H., Wu, M. C., Huang, C. Y., Tsou, A. P., and Huang, H. D. (2011). miRTarBase: a database curates experimentally validated microRNA-target interactions. *Nucleic Acids Res*, **39**(Database issue), D163–9.

Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K. F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., and Hirakawa, M. (2006). From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res*, **34**(Database issue), D354–D357.

Kapushesky, M., Kemmeren, P., Culhane, A. C., Durinck, S., Ihmels, J., Krner, C., Kull, M., Torrente, A., Sarkans, U., Vilo, J., and Brazma, A. (2004). Expression Profiler: next generation–an online platform for analysis of microarray data. *Nucleic Acids Res*, **32**(Web Server issue), W465–W470.

Kauffmann, A., Gentleman, R., and Huber, W. (2009). arrayQualityMetrics–a bioconductor package for quality assessment of microarray data. *Bioinformatics*, **25**(3), 415–416.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, D. (2002). The human genome browser at UCSC. *Genome Res*, **12**(6), 996–1006.

Lim, W. K., Wang, K., Lefebvre, C., and Califano, A. (2007). Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks. *Bioinformatics*, **23**(13), 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.

Okuda, S., Yamada, T., Hamajima, M., Itoh, M., Katayama, T., Bork, P., Goto, S., and Kanehisa, M. (2008). KEGG Atlas mapping for global analysis of metabolic pathways. *Nucleic Acids Res*, **36**(Web Server issue), W423–W426.

Papadopoulos, G. L., Reczko, M., Simossis, V. A., Sethupathy, P., and Hatzigeorgiou, A. G. (2009). The database of experimentally supported targets: a functional update of TarBase. *Nucleic Acids Res*, **37**(Database issue), D155–8.

Salomonis, N., Hanspers, K., Zambon, A. C., Vranizan, K., Lawlor, S. C., Dahlquist, K. D., Doniger, S. W., Stuart, J., Conklin, B. R., and Pico, A. R. (2007). GenMAPP 2: new features and resources for pathway analysis. *BMC Bioinformatics*, **8**, 217.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**(5235), 467–470.

Schumacher, A., Kapranov, P., Kaminsky, Z., Flanagan, J., Assadzadeh, A., Yau, P., Virtanen, C., Winegarden, N., Cheng, J., Gingeras, T., and Petronis, A. (2006). Microarray-based DNA methylation profiling: technology and applications. *Nucleic Acids Res*, **34**(2), 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**, Article3.

Symons, S. and Nieselt, K. (2011). MGV: a generic graph viewer for comparative omics data. *Bioinformatics*, **27**(16), 2248–2255.

Wrzodek, C., Dräger, A., and Zell, A. (2011). KEGGtranslator: visualizing and converting the KEGG PATHWAY database to various formats. *Bioinformatics*, **27**(16), 2314–2315.

Xiao, F., Zuo, Z., Cai, G., Kang, S., Gao, X., and Li, T. (2009). miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Res*, **37**(Database issue), D105–10.