Integrated analysis and visualization of heterogeneous cross-platform microarray datasets

Clemens Wrzodek[1,*], Elif Unterberger[2], Johannes Eichner[1] Michael Schwarz[2], Andreas Zell[1]
1 Center for Bioinformatics Tuebingen (ZBIT), University of Tuebingen, Sand 1, 72076 Tübingen, Germany
2 Department of Toxicology, Institute of Experimental and Clinical Pharmacology and Toxicology, University of Tuebingen, Wilhelmstrasse 56, 72074 Tübingen, Germany
∗ **E-mail: clemens.wrzodek@uni-tuebingen.de**

Abstract

Konzept (technisch): Heute nicht nur mRNA sondern auch viele andere microarrays (miRNA, DNAm, etc.). Daten von den selben samples verfgbar. Methoden fr platformbergreifende datenanalyse und visualisierung sind sehr gefragt. Konzept (biologisch): Auf Ha-Ras / CTNNB1 eingehen, vor allem platformbergreifende zusammenhnge erlutern und RESULTS dass diese mit den genannten Methoden super analysier/ und visualisierbar sind.

Author Summary

Eine art subjektive zusammenfassung zu dem artikel schreiben...

Introduction

Vor Abgabe suchen und vereinheitlichen von

- **gene and genome nomenclature**
- ***p*-value**
- **fold-change**
- **Ctnnb1**
- **Ha-ras**
- **up-regulated and down-regulated**
- **BE vs AE (analyses vs. analyzes)**
- **irgendwo erklren dass unsere fold-changes immer log2 based sind!**

In the past, the microarray technology has primarily been used to detect changes in genome-wide gene expression of various biological systems. Usually, a set of biological samples is tested under various experimental and corresponding control conditions. This allows to calculate a relative change in gene expression between treatments and controls, but also between different treatments. Thus, for example, comparisons between differently mutated tumors or animal groups treated with different types of tumor promoters are possible. It has previously been shown that in mice, treatment with the anticonvulsant phenobarbital following a single ip injection of the mutagenic initiator N-diethylnitrosamine (DEN) predominantly leads to liver tumors harboring activating point mutations in the gene encoding the transcription factor β-Catenin, *Ctnnb1* (**Citation?**). In contrast, tumors isolated from animals treated with DEN only rather carry *Ha-ras*- or *B-raf*-mutations, both of which lead to constant activation of the MAP-kinase signaling

pathway [1]. In the study described by Stahl *et al.*, the microarray approach was used to explore the effect of phenobarbital-treatment on gene expression in mouse liver tumors as compared to non-tumor control tissue [2].

Today, the application of high-throughput techniques evolved from gene expression to a broad range of other genomic, proteomic and epigenomic features. MicroRNAs (miRNAs) are short non-coding RNAs that regulate gene-expression by potentially binding to complementary mRNA targets [3]. MiRNA expression can be measured very similarly to mRNA expression, by defining specific probesets for those RNAs. There are several microarray platforms available today to detect the expression of miRNAs [4].

When studying the proteome in context of signaling networks, it is very important to know the phosphorylation state of proteins. In many cases, only phosphorylated proteins become active and propagate a signal, e.g., in a pathway. The expression of basic and phosphorylated proteins can be measured with reverse-phase-protein arrays [5].

Epigenetic alterations are often detected by measuring DNA methylation of cytosines. It is known that DNA methylation is the only covalent modification of DNA and hypermethylation of cytosines in gene promoters leads to gene silencing of corresponding genes [6, 7]. The microarray technology has been extended to approaches that tile the promoter regions of defined genes to detect the degree of methylation in those regions [8].

Using all those microarray technologies (mRNA, miRNA, (phospho-)protein, and DNA methylation), on the same set of biological samples leads to high-dimensional, heterogeneous cross-platform datasets. Figure 1 For most of those individual platforms there are many microarray analysis methods available that go beyond the usual calculation and comparison of fold-changes or other statistical comparisons. But many interesting biological findings are not covered by single platform analysis. The biological system of a cell includes all levels of regulation and omics data expression and therefore, many conclusions can only be drawn when studying multiple genomic layers in parallel.

We here present methods that ease the integrated study of of multiple high-throughput datasets, in which several microarray platforms have been used to investigate various genomic features in the same set of biological samples. The usefulness of those methods is demonstrated by application to a cross-platform dataset of mRNA, miRNA, protein modification, and DNA methylation data, obtained from *Ha-ras*- or *Ctnnb1*-mutated liver tumors in *mus musculus*. These algorithms are going beyond single gene or protein expression analysis (i.e. complementary to locus-specific approaches) and thus into an area of computational biology with a very sparse research density. Interactions detected in heterogeneous datasets have the power to, e.g., reveal novel and very promising biomarker candidates.

Results and Discussion

Gene-based integration of heterogeneous datasets

All proposed methods are validated on a dataset of 19 mouse samples: 10 tumors (3 *Ha-ras*-mutated and not treated with phenobarbital (PB), and 7 *Ctnnb1*-mutated after PB-treatment) and 9 corresponding control tissue samples from the same animals (3 untreated, 6 PB-treated). Messenger RNA, microRNA, DNA methylation and protein modification expression datasets have been measured for all of those samples. We propose to integrate all datasets on gene-level. This is obvious for messenger RNA data, because mapping all probes on genes is a common task. For protein datasets, this is also straightforward, except that different protein modifications should not be mixed. Mapping DNA methylation data on genes requires to introduce a window around the transcription start site (TSS) of each gene. The Nimblegen chip-design provides probes for approximately 8,000 bps downstream and 3,000 bps upstream of each TSS. But including all probes within this window would lead to a huge amount of peaks for one gene, which is difficult to interpret. It has been shown that usually, the proximal promoter region has a stronger effect on a gene's regulation than more distant regions [9, 10]. Furthermore, many CpG islands

are overlapping the TSS, reaching into the 5' region of a gene [11]. And DNA methylation of CpG islands in promoter regions is crucial for the regulation of gene expression [12]. Putting all together, we decided to assign all probes to a gene that are within -2,000 bps and +500 bps of the TSS. MicroRNAs infer into gene-regulation by binding to complementary mRNA targets [13]. Mapping of microRNA data on genes can be realized by taking the locus of miRNA transcripts or by mapping miRNAs on their corresponding mRNA targets. Both approaches are eligible and have been confirmed by numerous publications [14,15]. However, the locus approach is difficult for integrated data analysis, because genes from neighboring loci are not necessarily co-regulated. And integration of different datasets is performed rather on a functional relationship than on a region-based relationship. Therefore, mapping from miRNA to genes is performed by mapping all probes from an miRNA to all genes, whose mRNAs are targets of the miRNA. For this purpose, numerous public databases are available that contain information about miRNAs and corresponding mRNA targets [15]. We worked with a union of three miRNA target databases that only contain experimentally validated tarkets: miRecords v3, miRTarBase v2.4 and TarBase v5.0c [16–18].

Methods for integrated data analysis

Having all platforms integrated, we propose numerous method for integrated representation, visualization and analysis of all platforms.

Data pairing

A first approach that is already suitable for any two datasets is "data pairing". Data pairing allows viewing two different datasets at once. In the first step, each value from dataset one is mapped on the matching value in dataset two. Then, both values are displayed next to each other in a tabular view. This procedure allows a first glance on integrated data that comes from two platforms. It is especially useful for showing microRNA and mRNA target interactions: the miRNA is shown, together with p-value and fold change on the left hand side, whereas the targeted mRNA is shown, together with p-value and fold change on the right hand side. In the middle between both datasets, the relation is displayed: This includes the source of this target mapping (miRNA target database or prediction algorithm), the confidence of this interaction and the inferred miRNA effect on the mRNA. The InCroMAP application includes all required algorithms to pair any two datasets - this includes the automated target annotation for miRNA data, the described gene-centering preprocessing steps and the actual creation and rendering of the paired table. An example can be seen in Figure 2.

Paired analysis of mRNA and miRNA expression data in *Ctnnb1*-mutated tumors revealed miR-495 to be the second most strongly up-regulated miRNA in this tumor type, whereas its expression remains unchanged in *Ha-ras*-mutated tumors. Annotation of experimentally validated targets showed that miR-495 regulates the gene expression of the *Onecut*-transcription factor HNF-6, which itself is down-regulated (see Figure 2 and Simion *et al.* [19]). HNF-6 is known to stimulate the expression of glucose-6-phosphatase (*G6pc*) by binding at its promoter region [20]. The proposed mode of action of miRNA-mediated expression regulation is that an miRNA binds to a complementary mRNA sequence, thus either leading to its degradation or inhibiting its translation at the ribosome. The assumption that miR-495 inhibits the translation of HNF-6 and thus the expression of *G6pc* is a possible explanation for the lack of *G6pc* on protein level in mouse liver tumors [21]. The fact that miR-495 is only up-regulated in *Ctnnb1*- but not in *Ha-ras*-mutated tumors whereas both are deficient in G6PC protein indicates that different regulatory mechanisms apply to the two tumor types.

Tabular integration of multiple heterogeneous datasets

Integration of multiple datasets into a tabular view is similar to data pairing, except that more than two datasets are aligned and displayed alongside. One important objective of multiple data integration is

showing the relevant information and hiding irrelevant information, as a table might get very big if all information, contained in a paired data table is shown. Thus, one row is created for each gene and one column for each platform. A single summary value of each gene is calculated for each platform and placed in the table. How the summary value is calculated largely depends on the platform and user-preference. There are arguments for taking the probe or peak with maximum differential expression, but still others prefer taking the mean, e.g., for mRNA datasets. The InCroMAP application, in which this integrated analysis method is implemented, allows for selecting the preferred method to calculate the summary value for each platform.

To provide researchers more information than the single summary value of each platform and gene, a special tabular representation is used that allows to collapse and expand rows in a table. Each gene can be expanded to see all platforms, providing data for this gene. Every platforms itself can be further expanded to show more detailed information. This includes, all probes and their expression values for mRNA data and the corresponding protein and protein modification expression for protein datasets. DNA methylation data is showing all peaks in the promoter region of this gene, when expanded and all single probes when the peaks are further expanded. Gene-based expansion of miRNA datasets will show all miRNAs that are targeting this gene, together with probe-level expression values if the miRNAs are further expanded. Figure 3 illustrates an example of this interactive tabular view.

The tabular integration of all platforms reflecting *Ctnnb1*-mutated tumors confirms a strong hypermethylation in the promoter region of *Egfr*, resulting in a mRNA decrease (see Figures 3 and 4) [22]. Down-regulation of *Egfr* is usually a cause for cancer [23]. Interestingly, neither the basic protein nor any phosphorylated variant of EGFR shows strong expression changes.

Integrated gene set enrichment

A common data analysis method, especially for mRNA data, is the gene set enrichment analysis. This analysis is performed by creating a gene list (typically by selecting the genes, showing huge differential expression in treatment versus control) and searching for significant enrichments therein. These enrichments can be pathways, gene ontology terms, or genes, regulated by certain transcription factors, or anything else. As a result, researchers can, for example, figure out which pathways are targeted by a certain treatment or compare targeted pathways of genotoxic or non-genotoxic compounds.

We propose a cross-platform extension of this procedure for multiple, integrated heterogeneous datasets. The method is depicted in Figure 5. As usually, platform-specific low-level processing and gene-centering procedures are performed for each platform. Each gene is then being assigned a p-value and fold change. Based on these values, a mixed gene pool of significant differentially expressed genes is created, by picking and merging the significantly differentially expressed genes from all platforms. On this mixed gene pool, an enrichment analysis is now performed. As a result, terms (e.g., pathways) are returned that are enriched across all platforms and not only effects of, for example, mRNA or miRNA data analysis. This provides a much broader insight into pathway changes in the organism, upon treatment with a specific compound, than enrichment analysis of a single platform.

Figure 6 shows a comparison between mRNA-only and an integrated gene set enrichment, using mRNA, miRNA, DNA methylation and protein modification data from *Ctnnb1*-mutated tumors. It is clearly visible that the integrated enrichment shows significant changes in many cancer related pathways, whereas the mRNA enrichment shows most significant changes in metabolic pathway maps. Having differentially cancer-related pathways as a results in tumor tissue shows the comprehensiveness of an integrated enrichment, in contrast to platform-specific results, showing in detail which pathways are differentially regulated in this particular platform.

Pathway-based visualization of integrated datasets

Still today, only few visualization techniques are available for integrated datasets. A wide-spread data visualization technique is, creating region-based tracks in a specific file format (common formats are BED or WIG) and using the UCSC genome browser [24]. This technique is good to get an insight into specific genome regions of interest, but this technique is not specialized on integrated data analysis and fails to give overall impressions. Especially if interesting genes lay on different chromosomes, it is not possible to visualize them together and if researchers do not already have genes of interest, the method fails completely to give a starting point in the visualized data.

Following up the common pathway enrichment analysis, integrated visualization of differentially regulated pathways and measured microarray data gives an overview of how compounds influence certain signaling networks. By changing node colors, shapes, adding new nodes to graphs or adding new labels, much information can be visualized directly in a pathway. Combinations of those possibilities allow visualizing mRNA fold change, miRNA fold change, Protein modification expression data and DNA methylation information directly in a pathway.

To get a comprehensive overview of the regulation of metabolic processes and affected pathways in a microarray dataset, we propose to modify KEGG's "Metabolic Pathways" map. This pathway visualizes compounds, enzymes and secondary pathways, which are involved in many important metabolic processes in an organism (see Figure 7 and Supplement **X**). To visualize specific microarray data and differentially regulated metabolic processes in this pathway, the color of various pathway elements is changed. All downregulated enzymes (depicted as edges between compounds) are colored blue and all upregulated enzymes are colored red. Stronger differential expression is indicated by more saturated colors. With this procedure, the expression of all contained enzymes is visualized, e.g., based on an mRNA dataset. Further, the metabolic overview pathway contains multiple rectangular nodes which are references to secondary metabolic pathways. The color of those referenced pathways can be changed to reflect the p-value of this pathway in an enrichment analyses. In other words, the color of referenced pathways can be changed to a more saturated color if the pathway is significantly differentially regulated in a microarray dataset and to a brighter color if it is less significantly differentially regulated. The resulting picture is an overview which metabolic processes and enzymes are up- or downregulated in any input microarray dataset. See Figure 7 for an example.

The described method can further be extended to visualize data from multiple platforms in any particular pathway. For most color-based visualizations, we define blue to indicate downregulation and red for upregulation. More saturated colors indicate stronger differential expressions and white is used to visualize no differential expression. Grey is used to show that no data is available from the input dataset. Messenger RNA expression is typically available for the majority of nodes in a pathway. Hence, the background color of every node in a pathway is changed to reflect the mRNA expression change. Protein expression is visualized by adding a small colored box below each node. If multiple measurements are available for differentially modified proteins (e.g., acetylated or phosphorylated isoforms), a separate box is added for each modification. Each box is labeled according to the proteins modification.

Integrating microRNA data into the pathway visualization is more difficult, as pathways usually consist of protein coding genes and compounds. MicroRNAs are not included in pathways and thus, a connection from each miRNA to the nodes in a pathway must be established. As already described, miRNAs are integrated with other platforms by querying miRNA target databases and put the miRNAs in relation to their corresponding targets. The same approach is used for the pathway-based visualization. Each miRNA that has a target within the current pathway is added as small rectangle, which is colored according to miRNA expression. The connection to the target is depicted by a line from the miRNA to the node, corresponding to the target mRNA.

DNA methylation data can be interpreted as a trajectory in a defined window for each gene. This information must be summarized, in order to crete a brief overview for each gene. Since most researchers want to know if a gene is rather hyper- or hypomethylated, we recommend to inspect and visualize the

DNA methylation peaks. Visualizing the mean or median is not very informative since small local peaks can already have a strong influence on gene expression. Hence, to get a single summary value for each gene, any peak detection algorithm can be applied to a DNA methylation dataset (see, e.g., the user's guide of Nimblegen's SignalMap software [25]) or the peak can be approximated by taking the probe with maximum differential expression on a normalized and smoothed DNA methylation dataset.

TODO: Folgende Bloecke ueberarbeiten und etwas results zusammentippen, auf glycolysis fig eingehen, zustzlich noch eine fig (wnt/map) mit alle 4 typen.

The summary value for DNA methylation data is visualized as a black bar that is drawn from the middle to the left side to indicate hypomethylation and from the middle to the right side to indicate hypermethylation. The total size of the bar is proportional to the summary value, i.e., the maximum DNA methylation peak. The aim of this visualization is giving a first hint if a gene promoter is differentially methylated. In the InCroMAP application, the gene can be selected to get a detailed plot of the actual DNA methylation trajectory in the corresponding promoter region (see Figure 4).

A global overview of metabolic pathways in *Ha-ras*- or *Ctnnb1*-mutated tumors on mRNA level provided a first insight into the profound metabolic changes taking place in the tumors. Integrated enrichment indicated individual pathways with the strongest differential regulation in the respective tumor tissue as compared to normal tissue. The visualizations clearly show characteristic perturbations in the metabolism of *Ctnnb1*-mutated tumors.

Major transcriptional changes take place in key pathways of energy metabolism, such as glycolysis and gluconeogenesis, the citric acid cycle or the urea cycle. In general, the key enzymes of gluconeogenesis are down-regulated whereas the expression of glucokinase, which catalyzes the first step of glycolysis, is up-regulated. Also, the rate-limiting enzymes of the citric acid cycle are up-regulated. This might indicate that this tumor type uses glucose as fuel rather than synthesizing it *de novo*. Furthermore, the key enzymes of the urea cycle as well as several enzymes involved in amino acid catabolism are characteristically down-regulated in *Ctnnb1*-mutated tumors which is consistent with previous findings [26].

All of the described visualization techniques allow for a joint visualization of mRNA, miRNA, DNA methylation and protein modification data (see Figure 8 for a detailed example of the described visualizations).

Materials and Methods

TODO: komplette section ueberarbeiten!!!

Samples

The samples used for the analyses presented here were taken from a previous experiment by Marx-Stoelting *et al.* [**?**]. In brief, male C3H mice received a single dose of DEN and were subsequently kept on a diet containing PB and a PB-free control diet, respectively. *Ctnnb1*-mutated tumors were isolated from the phenobarbital (PB)-treated mice whereas *Ha-ras*-mutated tumors were isolated from the animals which received a PB-free diet. Normal tissue samples were taken from the same livers.

Low-level data processing

In order to generate processed data from raw data, each data type needs its own low-level data processing procedure. It is important to define standardized procedures here, because the term "processed data" is not really defined. For example, datasets are sometimes being logarithmized and sometimes not. Another example is the method to calculate p-values or the statistical false discovery rate (FDR) correction afterwards.

As one can see, integrated data analysis does also cover the low-level data processing to generate suitable and comparable high-level data that can be used for the final integrated data analysis methods.

Furthermore, all data must be reduced to a common denominator. The most reasonable common denominator is a gene, thus, all data should be gene-centered after the defined low-level processing steps. The need for this step and our decision to gene-center all datasets is explained in more detail in Section 2.3.

Messenger RNA

EKUT-B received historical and novel mRNA expression data measured by Affymetrix microarrays from the partners EKUT-A, BSP, MUW and UCB in CEL file format. The respective CEL files containing the raw probe intensities were normalized using the Robust Multichip Average (RMA) method and the quality of the experiments was assessed using diverse plots and statistics implemented in the package "arrayQualityMetrics" for R/Bioconductor [2]. Figure 3 shows a selection of plots generated for Affymetrix gene expression data obtained from MUW. On the basis of extensive quality controls, arrays with limited quality were identified. The corresponding experiments are currently repeated by MUW.

Depending on the study design, a moderated F-statistic (time-series data) or a moderated t-statistic (single time points) was chosen to detect differentially expressed genes (implemention from "limma" package for R/Bioconductor) [3]. In order to correct for testing multiple genes and to ensure a false discovery rate less than 0.05 the Benjamini-Hochberg method was applied [4]. Additionally, fold change cutoffs of 2 and 0.5 were used to select upregulated and downregulated genes, respectively.

For each treatment and regulation state (i.e. up- or downregulated) a hypergeometric test was performed to identify enriched functional categories and pathways, respectively. Functional categories for characterizing genes with respect to biological process, molecular function and cellular compartment have been chosen according to the Gene Ontology Project (www.geneontology.org). Canonical Pathways were taken from the Biobase database TransPath (www.biobase-international.com), the Ingenuity Knowledge Base (www.ingenuity.com), and the KEGG database (www.genome.jp/kegg). Selected KEGG pathways were overlayed with fold changes, represented by a color gradient, and visualized in interactive graph plots (Figure 8B) to facilitate thorough visual inspection of drug-induced pathway alterations. Furthermore, venn diagrams were generated to illustrate the effects of NGC compounds, differing between cell types or depending on the duration of treatment (Figure 4).

The venn diagrams shown in Figure 4 A2 indicate a CPA-mediated activation of the cell cycle and DNA replication in rodent hepatocytes represented in the deregulated mRNAs. This result, which was exclusively found after in-vivo treatment followed by mRNA expression analysis of hepatocytes isolated from the CPA- and vehicle-treated animals, is consistent with the fact that considerable liver growth was observed for CPA-treated rats. As expected, an upregulation of CYP-dependent metabolism was found, which is most likely due to transcriptional activation by the CAR receptor. This CYP induction could be consistently observed throughout cell types (hepatocytes and mesenchymal cells) and experimental procedures (in-vivo and in-vitro treatment).

Micro RNA

Micro RNA data processing is very similar to messenger RNA data processing. Subsequently we only briefly describe the required steps here. The background-corrected miRNA expression data is logarithmized in the first step. Some miRNAs are not expressed in certain cells and thus, all probes that have a low signal/noise ratio throughout all chips (i.e. the signal is lower than the background noise) are removed. Afterwards, the probes are processed again similar to mRNA data by calculating means, fold changes, p-values and eventually applying statistical correction methods on the calculated p-values. Integration of micro RNA target databases

Micro RNAs influence the regulation of gene expression in two different ways: either by translational inhibition or target mRNA cleavage. Although there are some evidences for a transcriptional regulation by miRNAs,[5] the common way to interfere into protein biosynthesis is by binding to a messenger RNA and repressing translation or inducing cleavage [6].

Therefore, it is required to know the mRNA targets for a micro RNA to detect the impact of highly up- or downregulated miRNAs and to correlate miRNAs to other datasets. There are many publications, describing miRNA mRNA interactions[5, 7-16]. Flat data files, summarizing those interactions, can be downloaded from several miRNA target databases. These databases contain experimentally validated miRNA and mRNA target interactions and provide flat-file downloads in different file formats. But still today, these databases only contain subsets of all miRNA mRNA interactions and hence, many methods to predict mRNA targets of miRNAs have been published[7, 10, 13, 15]. These prediction methods can be used to complete the data available from experimental databases.

EKUT-B has created a comprehensive, integrated miRNA target database that can be used to annotate miRNA datasets with corresponding mRNA targets. This dataset has been created by integrating three experimental miRNA target databases (miRecords v3 [16], miRTarBase v2.4 [12], TarBase v5.0c [14]), removing duplicates and mapping all gene identifiers to common NCBI Gene IDs. This dataset of experimentally validated targets has been completed by adding high-confidence predictions of ElMMo v5 [15], DIANA microT v4.0 [13] and TargetScan v5.2 [10]. This has proven to be a good combination and integrating more prediction methods might even worsen the resulting miRNA target mapping [7].

Having this target database, gene-centering of miRNA data is performed as follows: first, each miRNA is assigned one value. This is done in the same manner as with mRNA data, e.g. by calculating the mean of all probes for a miRNA. Then, each gene is assigned one or multiple values of microRNAs, regulating the levels of the mRNA encoded by this gene or the translation of this genes mRNA.

DNA methylation

Within the MARCAR project, DNA methylation data is measured by Nimblegen promoter tiling arrays, which cover sequence regions of about -8kbps to +3kbps around the transcription start site (TSS) of each gene. This sequence region is tiled with hundreds of probes that are about 50bps in size. The raw probe intensities are processed, as for most microarray platforms, with loess or quantile normalization methods. The resulting probe intensities are smoothed, by incorporating neighboring probe intensities.

The usage of microarray chips to measure DNA methylation is not as easy to interpret as data from other platforms. DNA methylation in mammals affects nearly exclusively the C-5 of cytosines, followed immediately by a guanine. Consequently, probes with similar intensities, but a low CG-content should be treated different than probes with a high CG-content. Furthermore, these values are hard to interpret, because from those signal intensities one cannot read which CGs are methylated and which are not. Thus, methods have been developed that use a fully methylated dataset, in addition to the normal dataset [17]. The fully methylated dataset is obtained by methylation of all cytosines in a DNA and measuring the probe intensities of the same platforms as with the actual data. The MEDME procedure [17] can then be applied to these datasets. This procedure calculates relative methylation intensities, based on the fully methylated dataset and the CG content of all probes. These relative intensities are then used for further data analysis.

To perform an integrative analysis, it is now required to convert the region-based DNA methylation data to gene-based values. Because DNA methylation is a very region-specific effect, bringing DNA methylation data down to a single value for a gene cannot be done without loss of information. Simple procedures like taking the mean or median of all probes for a gene do not make sense, because local changes in DNA methylation that might have an impact on gene-regulation are not covered then. See for example Figure 5: simply taking the mean of both trajectories results in two very similar values, which do not reflect the actual DNA methylation pattern.

For DNA methylation analyzes, it is not only interesting if the whole promoter is hypo- or hypermethylated, but also if small genomic regions (like CpG islands) are hypo- or hypermethylated. To also catch these local changes, we established a procedure that calculates one value for one gene that tells the researcher if there are local or global methylation changes:

Step 1: Divide the covered sequence region (8kbp downstream and 3kbp upstream of the TSS for Nimblegen chips) into small bins of, e.g. 250bps.

Step 2: Compute p-value pgj based on moderated t-test for each gene g and each bin bj.

Step 3: Compute overall score pg for each gene as follows:

FORMEL FEHLT.

Result: The value pg is very small if the promoter does not show significant DNA methylation changes. A higher value tells the researcher that there are either local significant, or global medium significant methylation changes. A very high value for pg is assigned to promoters with significant global methylation changes.

This procedure does not differentiate between hyper- or hypomethylation and thus, again shows the need for single dataset analyses in addition to integrated data analysis. For an integrated data analysis, the most interesting question is, to detect novel markers that are only visible by having an integrated view on the data. For most of these procedures, a value telling if there are significant methylation changes in a gene-promoter is sufficient. For more detailed analysis, the researcher can then manually take a closer look at the data and see if it is a global or local effect and if the promoter is hypo- or hypermethylated.

Protein modification data

The raw fluorescence intensities measured by Zeptosens Reverse-Phase Protein Arrays (RPPM) were obtained from NMI as a spreadsheet file. At EKUT-B the analytes (i.e. proteins or protein modifications) quantified in diverse biological samples were annotated with UniProt IDs to facilitate the interpretation of the data in the context of signaling pathways. Next, the data was log-transformed and centered around the median intensity values observed for control samples to ease interpretability and obtain a symmetrical scale of protein or protein modification levels. Measurements, for which the background noise (i.e. signal of secondary antibody) was higher than the combined foreground and background signal (i.e. signal of primary and secondary antibody), were treated as missing values and imputed using the k-Nearest-Neighbor (kNN) algorithm.

In order to detect analytes, which are differentially expressed between two sample groups (e.g., treated vs. control samples), a moderated t-test (implementation from "limma" package for R/Bioconductor) was applied [3]. Subsequently, a correction for multiple hypothesis testing was done using the FDR control method proposed by Benjamini and Hochberg [4]. The results from this statistical analysis were visualized using volcano plots (Figure 6).

The results shown in Figure 6, which shows the volcano plots for NGC-treated rodent hepatocytes, are consistent with findings reported in the literature. For instance, the upregulation of diverse Cytochrome P450 enzymes is a known effect of phenobarbital (PB) treatment, which is mediated via the CAR/PXR receptor. The strong upregulation of Cyclin E2 indicates that PB treatment may also impact on cell cycle regulation.

The analytes were categorized into tissue- and treatment-specific regulation states, respectively, based on significance (p-values from moderated t-statistic) and fold changes. Specifically, regulation states were defined to characterize genes, with regard to the strength (weak, medium, high) and direction (up, down) of differential expression. Expression profiles, composed of the regulation states for each analyte, were clustered using a hierarchical average-linkage approach for the purpose of detecting mechanistic similarities among the treatments (Figure 7). To measure the pairwise similarity of expression profiles (corresponding to rows in Figure 7) an arbitrarily defined scoring matrix was built to score pairs of regulation states, which were in turn summarized across analytes and normalized using a geometric mean.

It becomes obvious from Figure 7 that GCs differ significantly from NGCs in their protein expression profiles. One of the most striking effects observed in Figure 8 is that the Ser235/Ser236-phosphorylated form of S6 Ribosomal protein shows strong upregulation for 100% (2/2) of the GCs, whereas strong downregulation was observed for 91% (10/11) of the NGCs, and no differential expression was detected for 100% (2/2) of the NCs. These results indicate that this protein, which is located in the 40S subunit of the ribosome, and which was reported to impact on cell growth and proliferation [18], is a promising marker for the early detection of nongenotoxic cancerogenicity. In combination with other candidate marker proteins showing class-specific, characteristic expressions patterns (e.g. STAT-3, c-Jun, BAX, etc.), protein signatures facilitating reliable detection of either GC or NGC compounds can be inferred from the data. For this purpose, further investigations are currently pursued by EKUT-B.

Furthermore, EKUT-B identified signaling networks, which are altered in tumor tissue (EKUT-A mouse study) or perturbed after treatment with NGC compounds (BSP rat study), by mapping the analytes to literature-based canonical pathways extracted from the Biobase Knowledge Library (BKL), the Ingenuity Knowledge Base (IKB) and the KEGG database. The impact on canonical pathways was visualized using bar plots showing the fraction of differentially regulated analytes for diverse pathways with potential relevance to the mechanism of NGC compounds (Figure 8).

Figure 8 shows that the differences in the expression profiles of GC and NGC illustrated in Figure 7 come along with considerable differences in the pathways addressed by representative compounds. One of the most obvious findings is that the Wnt pathway, which is known to play a crucial role in the mechanism of NGC compounds, such as PB [19], is exclusively perturbed by treatment with the NGC compounds DHEA and TAA. Among the differentially expressed analytes in the pathway are for instance phosphorylated ?-catenin, phosphorylated GSK3?, and p53 which are known to be key regulators of cell fate.

To obtain a more detailed view of the compound-specific pathway alterations, EKUT-B has implemented a flexible graph viewer which allows visualizing the regulation states of the analytes in the context of signaling networks extracted from KEGG. This is exemplified by Figure 9, which shows the effect of the compound TAA on the Wnt signaling pathway.

Implementation and availability

Acknowledgments

References

1. Aydinlik H, Nguyen TD, Moennikes O, Buchmann A, Schwarz M (2001) Selective pressure during tumor promotion by phenobarbital leads to clonal outgrowth of beta-catenin-mutated mouse liver tumors. Oncogene 20: 7812–7816.

2. Stahl S, Ittrich C, Marx-Stoelting P, Khle C, Ott T, et al. (2005) Effect of the tumor promoter phenobarbital on the pattern of global gene expression in liver of connexin32-wild-type and connexin32-deficient mice. Int J Cancer 115: 861–869.

3. He L, Hannon GJ (2004) MicroRNAs: small RNAs with a big role in gene regulation. Nat Rev Genet 5: 522–531.

4. Hoheisel JD (2006) Microarray technology: beyond transcript profiling and genotype analysis. Nat Rev Genet 7: 200–210.

5. Pirnia F, Pawlak M, Thallinger GG, Gierke B, Templin MF, 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.

6. Jones PA, Baylin SB (2007) The epigenomics of cancer. Cell 128: 683–692.

7. Bernstein BE, Meissner A, Lander ES (2007) The mammalian epigenome. Cell 128: 669–681.

8. Schumacher A, Kapranov P, Kaminsky Z, Flanagan J, Assadzadeh A, et al. (2006) Microarray-based DNA methylation profiling: technology and applications. Nucleic Acids Res 34: 528–542.

9. Cooper SJ, Trinklein ND, Anton ED, Nguyen L, Myers RM (2006) Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. Genome Res 16: 1–10.

10. Kim TH, Barrera LO, Zheng M, Qu C, Singer MA, et al. (2005) A high-resolution map of active promoters in the human genome. Nature 436: 876–880.

11. Gardiner-Garden M, Frommer M (1987) CpG islands in vertebrate genomes. J Mol Biol 196: 261–282.

12. Razin A, Cedar H (1991) DNA methylation and gene expression. Microbiol Rev 55: 451–458.

13. Bartel DP (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 116: 281–297.

14. Luk JM, Burchard J, Zhang C, Liu AM, Wong KF, 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.

15. Alexiou P, Maragkakis M, Papadopoulos GL, Reczko M, Hatzigeorgiou AG (2009) Lost in translation: an assessment and perspective for computational microRNA target identification. Bioinformatics 25: 3049–3055.

16. Xiao F, Zuo Z, Cai G, Kang S, Gao X, et al. (2009) miRecords: an integrated resource for microRNA-target interactions. Nucleic Acids Res 37: D105-10.

17. Hsu SD, Lin FM, Wu WY, Liang C, Huang WC, et al. (2011) miRTarBase: a database curates experimentally validated microRNA-target interactions. Nucleic Acids Res 39: D163-9.

18. Papadopoulos GL, Reczko M, Simossis VA, Sethupathy P, Hatzigeorgiou AG (2009) The database of experimentally supported targets: a functional update of TarBase. Nucleic Acids Res 37: D155-8.

19. Simion A, Laudadio I, Prévot PP, Raynaud P, Lemaigre FP, et al. (2010) MiR-495 and miR-218 regulate the expression of the Onecut transcription factors HNF-6 and OC-2. Biochem Biophys Res Commun 391: 293–298.

20. Beaudry JB, Pierreux CE, Hayhurst GP, Plumb-Rudewiez N, Weiss MC, et al. (2006) Threshold levels of hepatocyte nuclear factor 6 (HNF-6) acting in synergy with HNF-4 and PGC-1alpha are required for time-specific gene expression during liver development. Mol Cell Biol 26: 6037–6046.

21. Weber G, Cantero A (1955) Glucose-6-phosphatase activity in normal, pre-cancerous, and neoplastic tissues. Cancer Res 15: 105–108.

22. Montero AJ, Díaz-Montero CM, Mao L, Youssef EM, Estecio M, et al. (2006) Epigenetic inactivation of EGFR by CpG island hypermethylation in cancer. Cancer Biol Ther 5: 1494–1501.

23. Zhang H, Berezov A, Wang Q, Zhang G, Drebin J, et al. (2007) ErbB receptors: from oncogenes to targeted cancer therapies. J Clin Invest 117: 2051–2058.

24. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, et al. (2002) The human genome browser at UCSC. Genome Res 12: 996–1006.

25. NimbleGen Systems Inc (2006) SignalMap User's Guide .

26. Stahl S, Ittrich C, Marx-Stoelting P, Khle C, Altug-Teber O, et al. (2005) Genotype-phenotype relationships in hepatocellular tumors from mice and man. Hepatology 42: 353–361.

Figure Legends

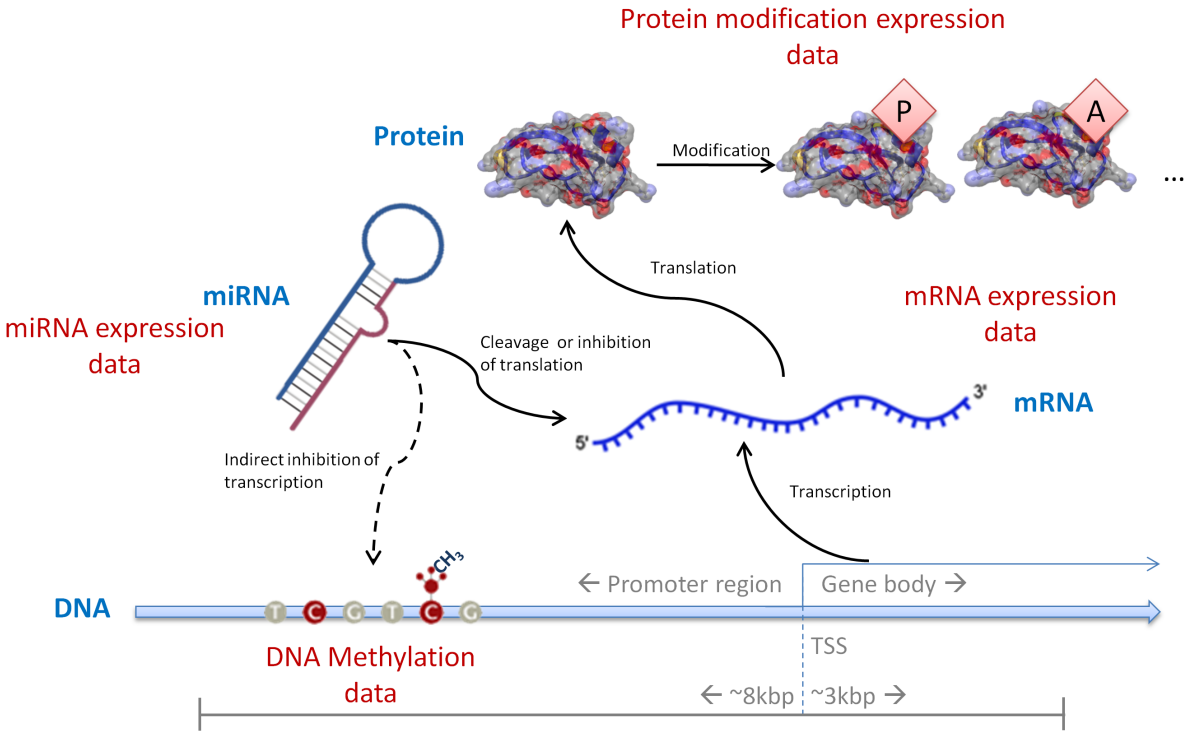
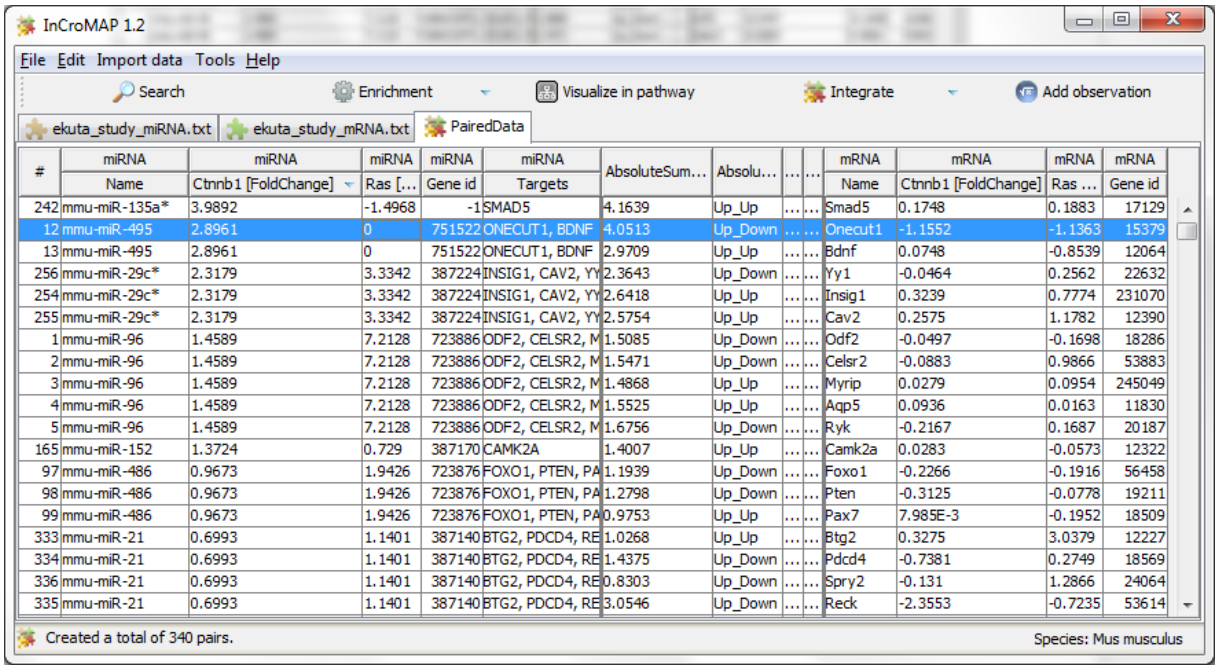


Figure 1. Visualization of diverse microarray platforms that have been employed to perform an integrated data analysis. 1 Caption berarbeiten, 2 nach Mglichkeit referenz auf farben entfernen, 3 Referenz in Introduction auf Figure einbauen. Red fonts describe the actual data types and corresponding platforms. This figure is restricted to genomic interactions relevant for this deliverable. At the bottom of the figure, a DNA sequence is given, for which methylation data is available. This DNA is transcribed to an mRNA. The transcription might be regulated by methylated regions on the gene promoter. Furthermore, miRNAs might inhibit the translation from mRNA to a protein. Both, mRNA and miRNA expression is measured with Affymetrix and Agilent microarrays. In the end, translated proteins might get modified, e.g., by phosphorylation or acetylation. The expression of some basic isoforms and specific modifications is determined, using Zeptosens arrays.



#	miRNA Name	miRNA Ctnnb1 [FoldChange]	miRNA Ras [...]	miRNA Gene id	miRNA Targets	AbsoluteSum...	Absolu...	...	...	mRNA Name	mRNA Ctnnb1 [FoldChange]	mRNA Ras ...	mRNA Gene id
242	mmu-miR-135a*	3.9892	-1.4968	-1	SMAD5	4.1639	Up_Up	...	...	Smad5	0.1748	0.1883	17129
12	mmu-miR-495	2.8961	0	751522	ONECUT1, BDNF	4.0513	Up_Down	...	...	Onecut1	-1.1552	-1.1363	15379
13	mmu-miR-495	2.8961	0	751522	ONECUT1, BDNF	2.9709	Up_Up	...	...	Bdnf	0.0748	-0.8539	12064
256	mmu-miR-29c*	2.3179	3.3342	387224	INSIG1, CAV2, YY	2.3643	Up_Down	...	...	Yy1	-0.0464	0.2562	22632
254	mmu-miR-29c*	2.3179	3.3342	387224	INSIG1, CAV2, YY	2.6418	Up_Up	...	...	Insig1	0.3239	0.7774	231070
255	mmu-miR-29c*	2.3179	3.3342	387224	INSIG1, CAV2, YY	2.5754	Up_Up	...	...	Cav2	0.2575	1.1782	12390
1	mmu-miR-96	1.4589	7.2128	723886	ODF2, CELSR2, M	1.5085	Up_Down	...	...	Odf2	-0.0497	-0.1698	18286
2	mmu-miR-96	1.4589	7.2128	723886	ODF2, CELSR2, M	1.5471	Up_Down	...	...	Celsr2	-0.0883	0.9866	53883
3	mmu-miR-96	1.4589	7.2128	723886	ODF2, CELSR2, M	1.4868	Up_Up	...	...	Myrip	0.0279	0.0954	245049
4	mmu-miR-96	1.4589	7.2128	723886	ODF2, CELSR2, M	1.5525	Up_Up	...	...	Aqp5	0.0936	0.0163	11830
5	mmu-miR-96	1.4589	7.2128	723886	ODF2, CELSR2, M	1.6756	Up_Down	...	...	Ryk	-0.2167	0.1687	20187
165	mmu-miR-152	1.3724	0.729	387170	CAMK2A	1.4007	Up_Up	...	...	Camk2a	0.0283	-0.0573	12322
97	mmu-miR-486	0.9673	1.9426	723876	FOXO1, PTEN, PA	1.1939	Up_Down	...	...	Foxo1	-0.2266	-0.1916	56458
98	mmu-miR-486	0.9673	1.9426	723876	FOXO1, PTEN, PA	1.2798	Up_Down	...	...	Pten	-0.3125	-0.0778	19211
99	mmu-miR-486	0.9673	1.9426	723876	FOXO1, PTEN, PA	0.9753	Up_Up	...	...	Pax7	7.985E-3	-0.1952	18509
333	mmu-miR-21	0.6993	1.1401	387140	BTG2, PDCD4, RE	1.0268	Up_Up	...	...	Btg2	0.3275	3.0379	12227
334	mmu-miR-21	0.6993	1.1401	387140	BTG2, PDCD4, RE	1.4375	Up_Down	...	...	Pdcd4	-0.7381	0.2749	18569
336	mmu-miR-21	0.6993	1.1401	387140	BTG2, PDCD4, RE	0.8303	Up_Down	...	...	Spry2	-0.131	1.2866	24064
335	mmu-miR-21	0.6993	1.1401	387140	BTG2, PDCD4, RE	3.0546	Up_Down	...	...	Reck	-2.3553	-0.7235	53614

Figure 2. Pairing of miRNA (left side) and mRNA (right side) datasets reveals miR-495 as a potential regulator of the HNF-6 coding gene *Onecut1*. **1 Caption berarbeiten, 2 Ueberlegen ob Ras Expressionswerte ausgeblendet werden sollten und wenn nciht, "in Ctnnb1" in titel.**

Tables

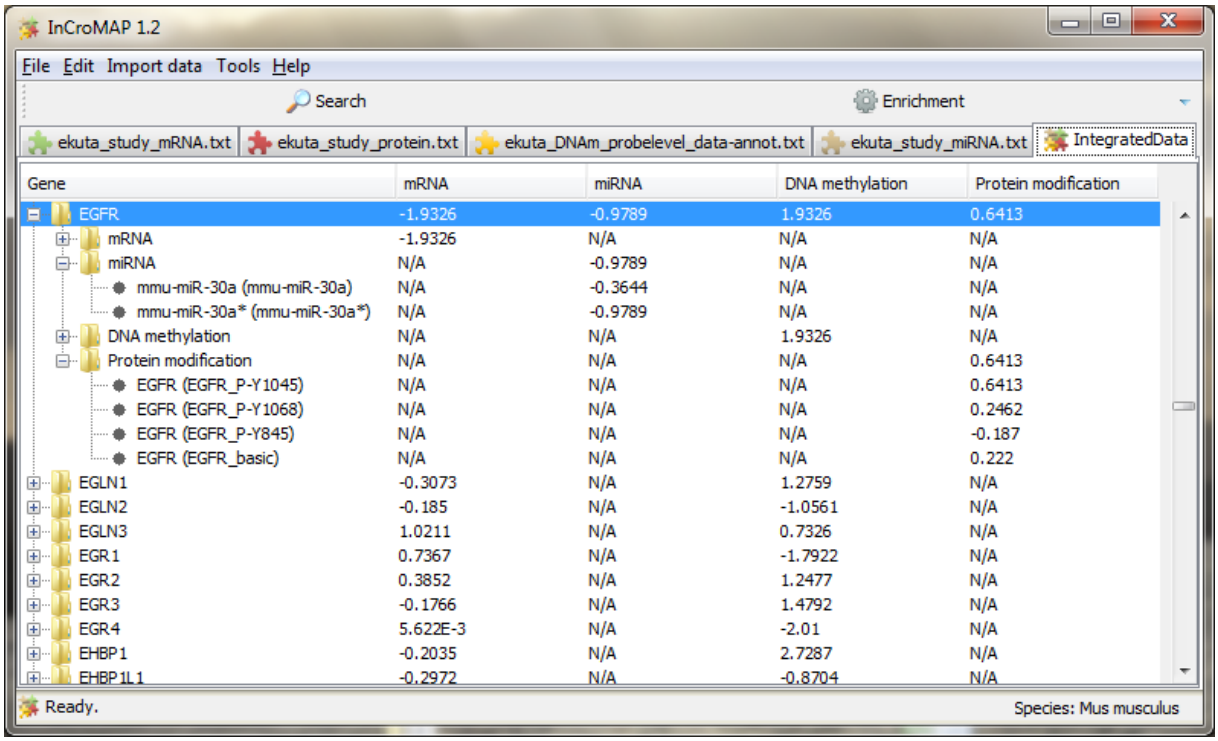


Figure 3. Multiple integration of four different platforms from *Ctnnb1* mutated tumors shows an mRNA decrease of *Egfr* as a potential effect of DNA methylation increase in the promoter region. Bildunterschrift

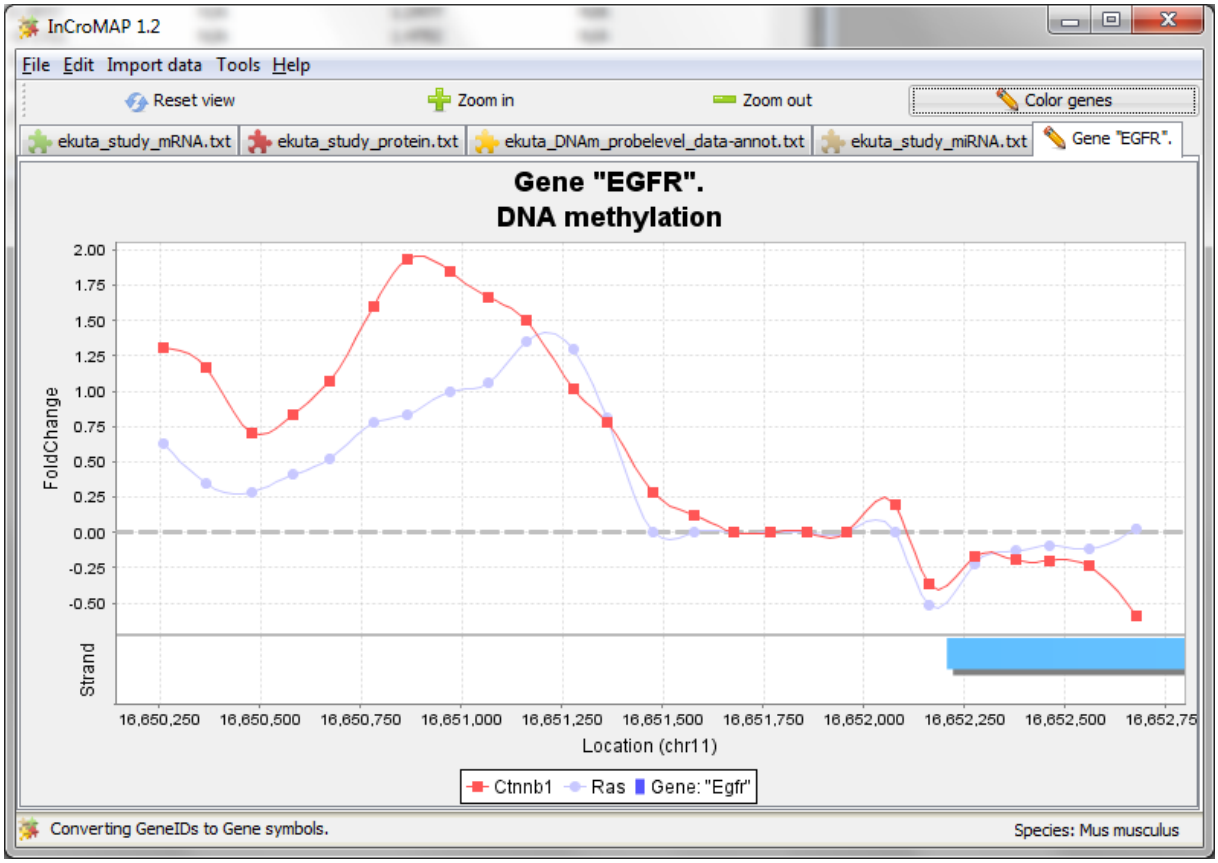


Figure 4. DNA methylation in the promoter region of *Egfr* in *Ctnnb1*-mutated tumors. **Bildunterschrift, 2 sagen dass dies auch ein paired ist, da mRNA entsprechend eingefrbt wurde**

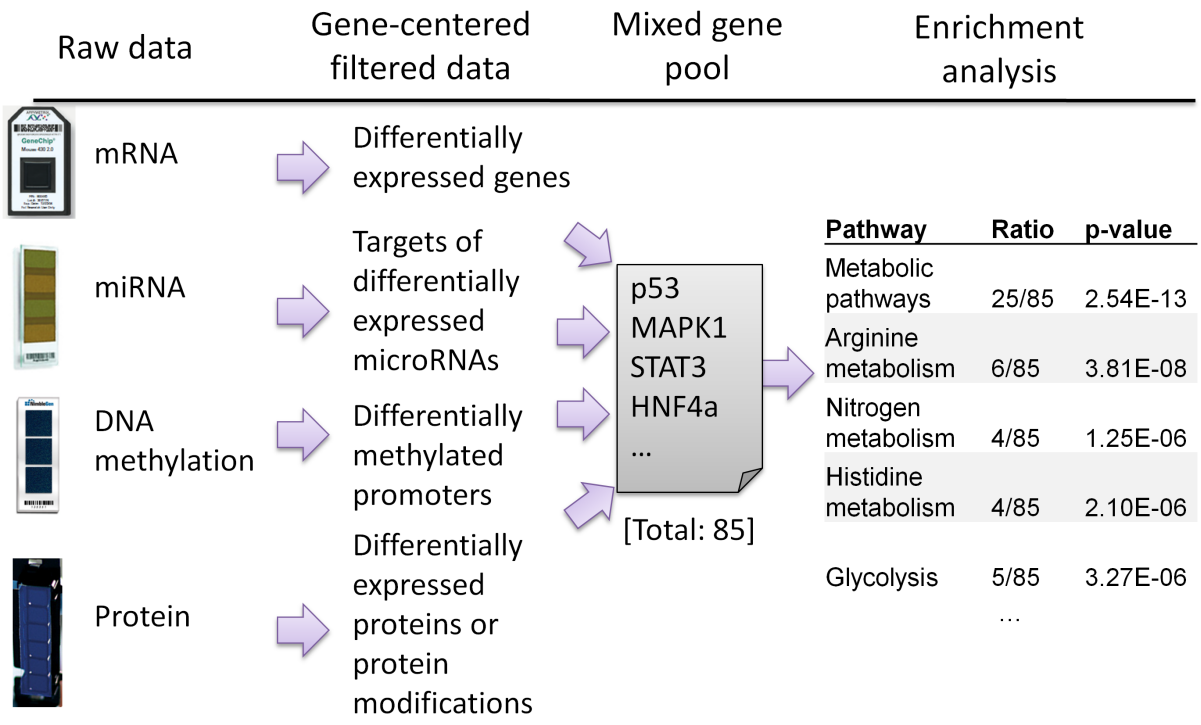


Figure 5. Proposed procedure for a cross-platform gene-set enrichment analysis. Bildunterschrift, 2 sagen dass 85 usw. examples sind.

#	ID	Name	List ratio	BG ratio	P-value	Q-value
1	path:mmu05215	Prostate cancer	31/3209	90/18080	6.543E-5	0.0145
2	path:mmu05200	Pathways in cancer	80/3209	326/18080	4.21E-4	0.0467
3	path:mmu04350	TGF-beta signaling pathway	27/3209	85/18080	6.877E-4	0.0509
4	path:mmu00910	Nitrogen metabolism	11/3209	23/18080	7.084E-4	0.0393
5	path:mmu05216	Thyroid cancer	13/3209	30/18080	7.428E-4	0.033
6	path:mmu05221	Acute myeloid leukemia	20/3209	57/18080	8.312E-4	0.0308
7	path:mmu04110	Cell cycle	36/3209	128/18080	1.179E-3	0.0374
8	path:mmu05213	Endometrial cancer	18/3209	52/18080	1.678E-3	0.0466
9	path:mmu05210	Colorectal cancer	21/3209	65/18080	1.888E-3	0.0466
10	path:mmu04540	Gap junction	26/3209	88/18080	2.37E-3	0.0526
11	path:mmu00250	Alanine, aspartate and glutamate metabolism	13/3209	34/18080	2.64E-3	0.0533
12	path:mmu05218	Melanoma	22/3209	72/18080	3.064E-3	0.0567
13	path:mmu05222	Small cell lung cancer	25/3209	87/18080	3.957E-3	0.0676
14	path:mmu05214	Glioma	20/3209	66/18080	4.838E-3	0.0767
15	path:mmu04976	Bile secretion	21/3209	71/18080	5.289E-3	0.0783
16	path:mmu04510	Focal adhesion	48/3209	200/18080	5.459E-3	0.0757
17	path:mmu05219	Bladder cancer	14/3209	43/18080	8.303E-3	0.1084
18	path:mmu00561	Glycerolipid metabolism	16/3209	52/18080	8.801E-3	0.1085

Integrated enrichment (mRNA, miRNA, DNAm, protein)

#	ID	Name	List ratio	BG ratio	P-value	Q-value
1	path:mmu00910	Nitrogen metabolism	10/937	23/18080	7.702E-8	1.502E-5
2	path:mmu04976	Bile secretion	15/937	71/18080	2.275E-6	2.218E-4
3	path:mmu01100	Metabolic pathways	95/937	1203/18080	9.317E-6	6.056E-4
4	path:mmu00330	Arginine and proline metabolism	12/937	56/18080	1.936E-5	9.437E-4
5	path:mmu04110	Cell cycle	19/937	128/18080	2.318E-5	9.041E-4
6	path:mmu00260	Glycine, serine and threonine metabolism	9/937	34/18080	3.647E-5	1.185E-3
7	path:mmu00250	Alanine, aspartate and glutamate metabolism	9/937	34/18080	3.647E-5	1.016E-3
8	path:mmu03320	PPAR signaling pathway	13/937	82/18080	2.127E-4	5.185E-3
9	path:mmu00480	Glutathione metabolism	10/937	54/18080	3.14E-4	6.804E-3
10	path:mmu00340	Histidine metabolism	7/937	28/18080	3.835E-4	7.479E-3
11	path:mmu05146	Amoebiasis	15/937	116/18080	6.413E-4	0.0114
12	path:mmu04350	TGF-beta signaling pathway	12/937	85/18080	9.885E-4	0.0161
13	path:mmu00561	Glycerolipid metabolism	9/937	52/18080	9.887E-4	0.0148
14	path:mmu04512	ECM-receptor interaction	12/937	86/18080	1.09E-3	0.0152
15	path:mmu04540	Gap junction	12/937	88/18080	1.317E-3	0.0171
16	path:mmu04964	Proximal tubule bicarbonate reclamation	5/937	20/18080	2.593E-3	0.0316
17	path:mmu05200	Pathways in cancer	28/937	326/18080	2.905E-3	0.0333
18	path:mmu05216	Thyroid cancer	6/937	30/18080	3.182E-3	0.0345

mRNA enrichment

Figure 6. Comparison of an integrated and mRNA enrichment, based on data from Ctnnb1-mutated tumors. Bildunterschrift, Metabolic vs. Cancer

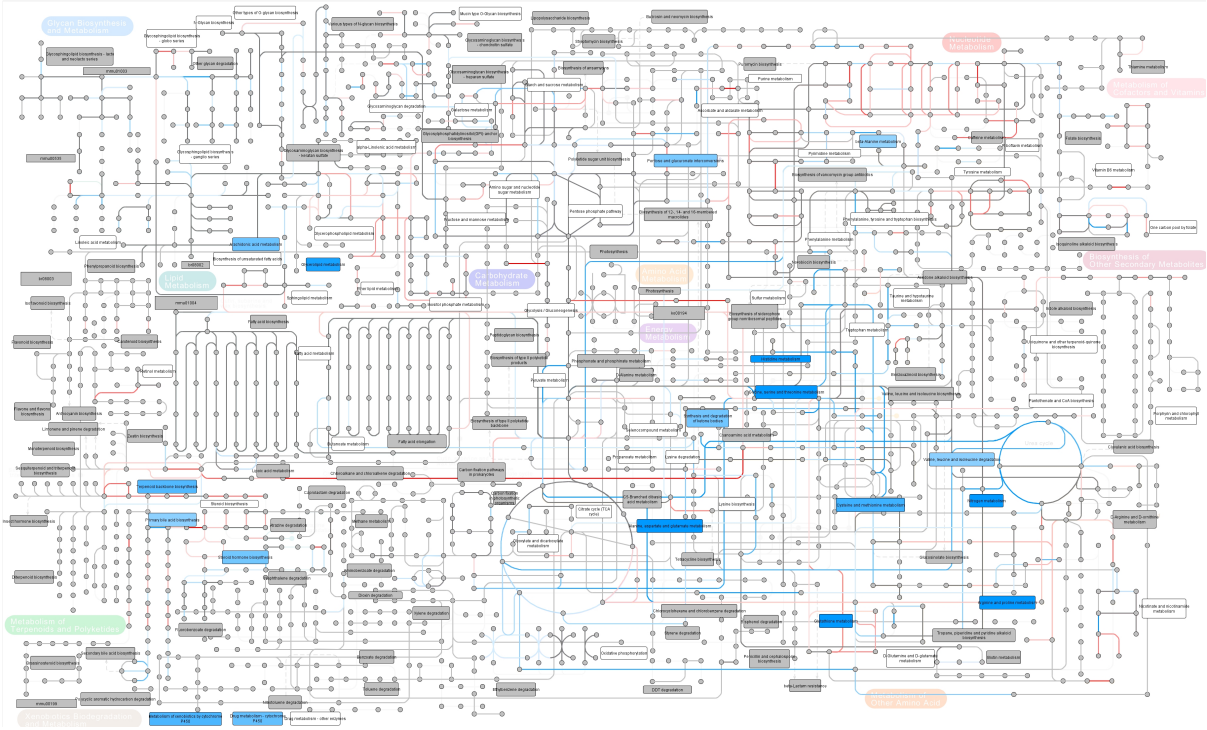


Figure 7. Comparison of an integrated and mRNA enrichment, based on data from Ctnnb1-mutated tumors. Bildunterschrift, 2 Integrates a) pathway b) enrichment c) (on mRNA). 3 Can be extended to line-color mRNA 4 Volles Bild + mRNA kanten im supplement
Was sollten wir hier genau zeigen?
– Gesamtes Bild oder bestimmter Ausschnitt? Wenn Ausschnitt, welcher?
– Kanten nach mRNA einfaerben oder nicht?
– Compound Beschriftungen verstecken oder zeigen? (Derzeit sind sie nicht beschriftet!)

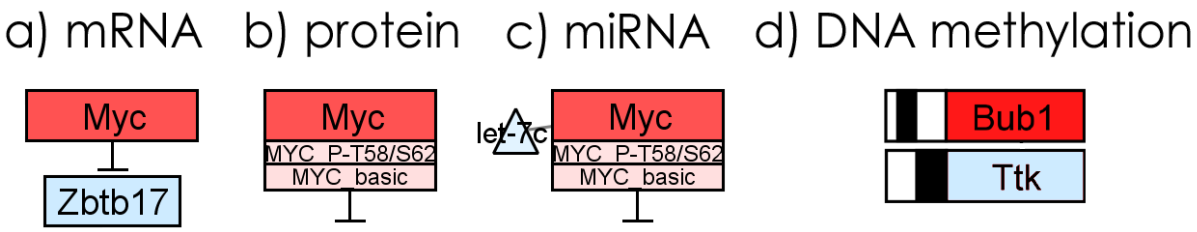


Figure 8. Pathway-based visualization of different platforms. Bildunterschrift

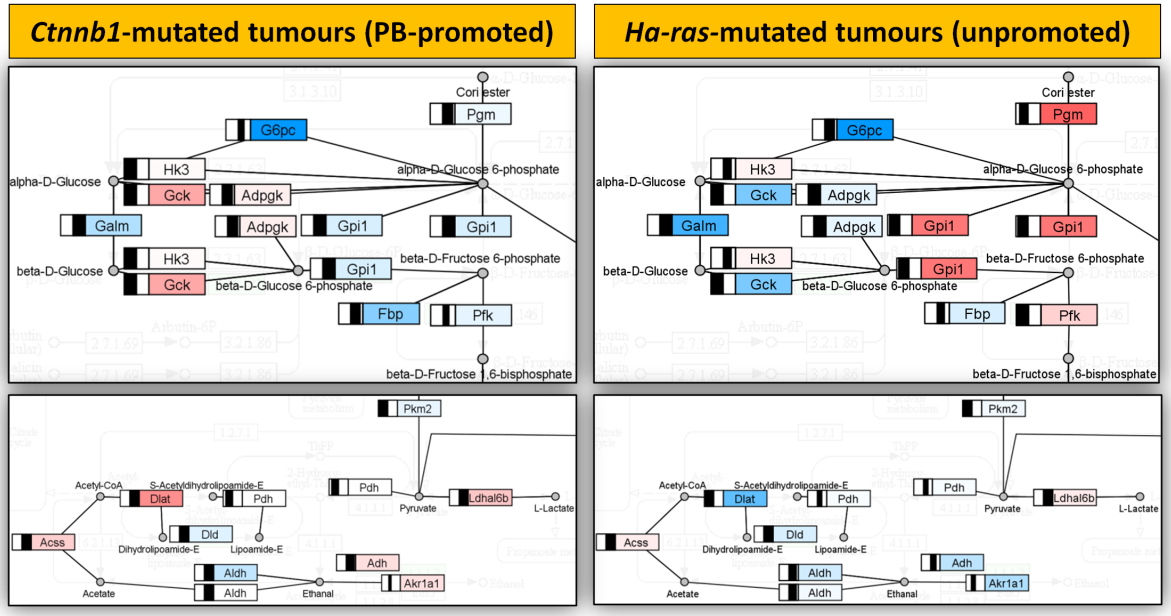


Figure 9. Transcriptional and epigenetic changes in Glycolysis/Gluconeogenesis during profiling of *Ha-Ras* and *Ctnnb1* mutated tumors. **Bildunterschrift, Legende (nur DNAm, mRNA)** Glucokinase (Gck) – katalysiert den 1. Schritt des Glucose-Abbaus in der Glykolyse: Genexpression hochreguliert, Promotorregion hypomethyliert. Glucose-6-phosphatase (G6PC) – katalysiert den der Gck entgegengesetzten Schritt in der Glukoneogenese, also dem Synthese von Glukose. Dieses Gen ist in beiden Tumortypen runterreguliert, dazu passt, dass die Promotorregion bei beiden weitestgehend hypermethyliert ist. Fructose-1,6-bisphosphatase (FBP) – auch ein Enzym der Gluconeogenese, in beiden Tumortypen runterreguliert mit hypermethylierter Promotorregion. Bei Pgm, Aldo und anderen eingefrbten Enzymen wre ich persnlich jetzt noch vorsichtig mit irgendwelchen Aussagen, weil da oft nur sehr wenige Sonden gemessen wurden oder die Sonden widersprchliche Werte zeigen. Korrelation von DNAm und mRNA ist teilweise recht interresant!