RESEARCH

Analysis of genes regulated by isoforms of the epidermal growth factor receptor in a glioblastoma cell line

Claus Weinholdt[1*], Henri Wichmann[2], Johanna Kotrba[2], David H. Ardell[4], Matthias Kappler[2], Alexander W. Eckert[2], Dirk Vordermark[2] and Ivo Grosse[1,3]

Correspondence:
laus.weinholdt@informatik.uni-alle.de
Institute of Computer Science, Martin Luther University Halle–Wittenberg , Germany
ull list of author information is vailable at the end of the article

Abstract

Background: The epidermal growth factor receptor (EGFR) is a major regulator of proliferation in tumor cells. Elevated expression levels of EGFR are associated with prognosis and clinical outcomes of patients in a variety of tumor types. There are at least four mRNA splice variants of the mRNA encoding four protein isoforms of EGFR in humans, named I through IV. The full-length (protein) isoform of EGFR is encoded by (mRNA) splice variant I, while (mRNA) splice variants II-IV encode for the shorter (protein) isoforms II-IV, but all EGFR (protein) isoforms are capable of binding the epidermal growth factor (EGF). EGFR is an important target of established tumor therapeutics like antibodies or small molecules which are successfully used over years. But, the functions and the potential as a biomarker of EGFR isoforms II-IV are unclear and motivating a deeper analysis of the soluble forms of this receptor. Here, we analyze transcriptome data of glioblastoma cell line SF767 to identify putative target genes regulated by the EGFR protein isoforms translated from mRNA splice variants II-IV and not by the full-length EGFR protein isoform translated from mRNA splice variant I nor by other receptors such as *c-ERBB2/HER2*, *c-ERBB3/HER3*, or *c-ERBB4/HER4* [1].

Results: To accomplish this goal, we performed two RNAi experiments, one to selectively knockdown EGFR mRNA splice variant I alone, one to knockdown all four EGFR mRNA splice variants I - IV, and one negative control with no RNAi and compared genome-wide expression levels using DNA microarrays of potential target genes in a glioblastoma cells line both with and without stimulation by EGF. Due to the nested experimental designs, the associated effects of a knock-down of EGFR splice variants (II-IV) can be calculated by subtracting the effects found after the knockdown of (I-IV) minus the effects found after the knockdown of only EGFR splice variants I. For this type of nested experimental design we developed a two-step bioinformatics approach, named Bayesian gene selection criterion (BGSC) approach, for the identification of putative target genes of EGFR isoforms II – IV based on the six expression levels observed for each potential target gene, using a novel algorithm based on the Bayesian Information Criterion. Finally, we experimentally validated a set of six putative target genes specifically regulated by EGFR isoforms II-IV and not by other receptors, and we found that qPCR validations confirmed the predictions in all cases.

Conclusions: By performing RNAi experiments for three poorly investigated EGFR isoforms in a glioblastoma cell line, we were able to successfully classify $1,140$ putative target genes specifically regulated by EGFR isoforms II – IV using our BGSC approach. This approach is suited to the analysis of nested experimental designs, because of nested mRNA splice variants and resulting in just directly measurable gene regulations. We provide a modular software implementation that allows users to easily apply the developed approach to data of other receptors like splice variants of the receptor *c-ERBB2/HER2* with a similar experimental design as well as to easily adapt this approach to other experimental designs.

Keywords: EGFR; RNAi; Bayesian Information Criteria (BIC); splice variants

1 Background

NOTE: isoform is used for protein and splice variants for mRNA !!! e.g EGFR isoforms II-IV produced from splice variants II-IV

Glioblastoma is the most malignant and most frequent primary cerebral tumor in adults and is responsible for 65% of all brain tumors [2]. One potential molecular target amplified in 36% of glioblastoma patients is the epidermal growth factor receptor (EGFR), and the expression of EGFR is associated with prognosis in cancer [3]. EGFR is known to affect growth and survival signals and to play a key role in the regulation of cell proliferation, differentiation, and migration of various tumor entities [4]. Therefore, EGFR is well known as a prognostic tumor marker and therapeutic target in different tumor entities.

The full-length transmembrane glycoprotein isoform of EGFR consists of three functional domains of which the extracellular domain is capable of binding at least seven different ligands such as EGF, AREG, or TGF-α [5]. However, there are different truncated mRNA splice variants (II, III, and IV) of the EGFR. Up to now only the full-length isoform of EGFR synthesized from splice variant I is well-investigated but comparatively little is known about the biological significance of EGFR splice variants II-IV [6]. These isoforms are capable of binding EGF [7] despite lacking the intra-cellular tyrosine-kinase domain [8], but information is very limited regarding their functional relevance. Different tumor therapies targeting EGFR (via antibodies or small molecules) often do not have the expected success. The EGFR isoforms II-IV may be responsible for that therapy failure because on one hand small molecules are targeting the tyrosine-kinase domain which is not present in the protein structure of EGFR isoforms II-IV. On the other hand, therapeutic antibody often bind the extracellular N-terminus of EGFR which is present in the structure of EGFR isoforms II-IV, too. However, such a EGFR specific antibody therapy required a interaction of the EGFR-bound therapeutic antibody and the presenting cell. The EGFR isoforms II-IV are solved protein forms do not identify the expressing cell itself but diffuse in the extracellular space and probably bind to surrounding non-tumor cell and may mislead the immune system.

This motivated our present work of perturbing the profile of EGFR mRNA splice variants using alternate small interfering RNAs (siRNAs) that differentially target these splice variants and of measuring the resulting expression responses using traditional microarrays. Specifically, we performed RNA interference (RNAi) experiments specific to EGFR splice variants [9] with cell line SF767 due to its strong EGFR-expression and good response on EGFR-target therapy in biological assays such as proliferation, SRB-assay, and cell-cycle analysis. Our studies (data not shown) leads to the assumption that this cell lines SF767 dispose of an intact EGFR-pathway. The study of such a pathway is useful only in a cell line with an inducible (e.g. by EGF-stimulation) and intact pathway. In a recent study, we found that a therapeutic antibody against EGFR (cetuximab) to not have an impact on cellular growth rate, cell cycle distribution, cell migration, clonogenic survival, and radio-sensitivity because of the missing immune system in that in vitro experiments for glioblastoma cell lines. The targeting of EGFR splice variants via siRNA reduced the growth rate of such cell in vitro [10].

However, it is impossible to knockdown by RNAi only EGFR splice variants II-IV and not EGFR splice variants I, because there does not exit an EGFR splice

variants II-IV specific region and this leads to the nested experimental design (Table 2). Due to this situation, the associated effects of a knock-down of EGFR splice variants II-IV can only be computed indirectly by subtracting the effects found after the knockdown of EGFR splice variants I-IV minus the effects found after the knockdown of only EGFR splice variants I. Furthermore, the problem of just indirectly measurable gene regulation or receptor effects because of nested splice variants is common in nearly all species. Therefore, we develop a two-step bioinformatics approach for the identification of putative target genes, which we test based on quantitative reverse transcription polymerase chain reaction (qPCR) validation experiments. This rest of the paper is structured as follows: In section 2 we introduce the two-step Bayesian gene selection criterion (BGSC) approach for identifying putative target genes regulated by EGF via EGFR isoforms II-IV and not by the full-length isoform or other receptors. In section 2.1 corresponding to the first step of the BGSC approach, we define groups with distinct expression patterns considering all eight possible modes of EGF-triggered transcriptional gene regulation via two sets of EGFR isoforms as well as non-EGFR receptors. In section 2.2 corresponding to the second step of the BGSC approach, we classify each gene into one of the four groups based on the Bayesian information criterion, and thereby obtain a prediction of putative target genes regulated by EGF specifically via EGFR isoforms II-IV. In section 2.3 we discuss the identified set of putative target genes, describe the qPCR experiments, and discuss the qPCR which validated our results. In section 3 we conclude our findings and in section 4 we provide additional information about our method.

2 Results and discussion

2.1 First step of the BGSC approach - grouping of genes

As a starting point, we consider all possible combinations of EGF-stimulated gene regulation under three possible signaling pathways, namely via the splice variant I isoform of EGFR, via the splice variants II-IV isoforms of EGFR, or via other non-EGFR receptors. Hence, each gene can be grouped into exactly one of the following eight gene groups `A` - `H`, which are graphically represented by Figure 1:

- Group `A` contains genes not regulated by EGF.
- Group `B` contains genes regulated by EGF not via EGFR isoforms I-IV, but via other receptors.
- Group `C` contains genes regulated by EGF via isoforms II-IV of EGFR and not via isoform I of EGFR and not via other receptors.
- Group `D` contains genes regulated by EGF via isoform I isoform of EGFR and not via isoforms II-IV of EGFR and not via other receptors.
- Group `E` contains genes regulated by EGF via isoforms II-IV of EGFR and via other receptors and not via isoform I of EGFR.
- Group `F` contains genes regulated by EGF via isoform I of EGFR and via isoforms II-IV of EGFR and via other receptors.
- Group `G` contains genes regulated by EGF via isoform I of EGFR and via other receptors and not via isoforms II-IV of EGFR.
- Group `H` contains genes regulated by EGF via isoform I of EGFR and via isoforms II-IV of EGFR and not via other receptors.

Next, we consider for each RNAi treatment if the genes of each group would be differentially regulated after EGF-stimulation. To conceptually analyze the gene expression of each group we denote by 1 a theoretical regulation (up or down) of the group after addition of EGF and denote by 0 no regulation, corresponding to showing no change in gene expression without and with EGF addition. Further, we define groups as regulated after EGF-stimulation if there was at least one incoming arrow to the group, and we define groups with no incoming arrow as unregulated. We consider three experimental manipulations with RNAi: a negative control without RNA interference, RNAi with siRNA against EGFR splice variant I, henceforth called siRNA$_I$, RNAi with siRNA against all EGFR splice variants, henceforth called siRNA$_{ALL}$ (Figure 2).

First, we consider the negative control without RNA interference. Here, none of the EGFR splice variants are down-regulated by a siRNA, so all target genes of EGFR isoforms and target genes of other receptors of EGF can be induced by EGF. Hence, we expect differential expression under EGF stimulation of genes belonging to groups `B - H` and no differential expression of genes belonging to group `A` (Figure 2a).

Second, we consider RNAi treatment with siRNA$_I$. Here, only splice variant I of EGFR is down-regulated by siRNA$_I$, so only target genes of EGFR isoforms II-IV and target genes of other receptors of EGF can be induced by EGF. Hence, we expect differential expression by EGF treatment of genes belonging to groups `B, C, E - H` and no differential expression of genes belonging to groups `A and D` (Figure 2b).

Third, we consider RNAi treatment with siRNA$_{ALL}$. Here, all four splice variants of EGFR are down-regulated by siRNA$_{ALL}$, so only target genes of other receptors of EGF can be induced by EGF. Hence, we expect differential expression by EGF treatment of genes belonging to groups `B, E - G` and no differential expression of genes belonging to groups `A, C, D and H` (Figure 2c).

Figure 3 summarizes the different expression patterns of Figure 2. We find that the eight gene groups show only four different expression patterns, so we reduce the eight gene groups `A - H` to four simplified gene groups a - d, where group `A` becomes group a, the union of groups `B and E - G` becomes group b, the union of groups `C` and `H` becomes group c, and group `D` becomes group d.

These simplified groups can be easily interpreted as follows: genes of the group a are never regulated by EGF, whereas genes of groups $b - d$ are regulated by EGF. genes of the group b are regulated by EGF only through other receptors besides EGFR isoforms. genes of the group c are regulated by EGFR isoforms II-IV and not by other receptors. And genes of the group d are regulated by EGFR isoform I and not other receptors or EGFR isoforms II-IV. Based on this reduction, we can now formulate the goal of this work as the identification of putative target genes regulated by EGFR isoforms II-IV and not by other receptors or more crisply as the goal of identifying genes of group c.

2.2 Second step of the BGSC approach - classification of genes

In step 2 we classify each potential target gene into one of four groups $z \in \{a, b, c, d\}$ introduced in section 2.1 based on the Bayesian Information Criterion, and thereby

identify target genes regulated by EGF via EGFR isoforms II-IV as those classified into group c.

In this second step we apply the oversimplified, but commonly accepted, assumption that the log-transformed expression of each gene is normally distributed [11] with a gene-specific and treatment-specific mean and variance.

For each gene, we additionally assume heteroscedasticity, or equality of the six variances, of the six log-normally distributed expression values under each of the six experimental conditions, an assumption very commonly made in t-tests, ANOVA, or other statistical tests. We further assume that the six means of these six Gaussian distributions are sample-specific according to group membership as in Table 4.

Specifically, we assume genes of the group a (not regulated by EGF) to show no differential expression at all under each of the six experimental treatments, as manifested by equality of the six means of the six Gaussian distributions (Figure 3, yellow column).

Second, we assume genes of the group b (regulated by EGF through other receptors besides any EGFR isoform) to show differential expression under EGF-stimulation, irrespective of RNAi treatment targeting any EGFR isoform (Figure 3, blue column). Thus we assume genes of the group b to have two different mean logarithmic expression levels, one in samples 1, 3, and 5, and another potentially different one in samples 2, 4, and 6. We denote these two mean logarithmic expression levels by μ_{b0} (Table 4b red) and μ_{b1} (Table 4b blue) respectively.

Third, we assume genes of the group c (regulated by EGFR isoform II-IV and not by other receptors) to show differential expression between the negative control and siRNA$_{ALL}$ treatments (Figure 3, red column) under EGF-stimulation. Thus we assume genes of the group c to have two different mean logarithmic expression levels, one in samples 1, 3, 5 and 6, and another potentially different one in samples 2 and 4. We denote these two mean logarithmic expression levels by μ_{c0} (Table 4c red) and μ_{c1} (Table 4c blue) respectively.

Fourth, we assume genes of the group d (regulated by EGFR isoform I only) to show differential expression between the negative control and siRNA$_I$ treatment (Figure 3, green column) under EGF-stimulation.

Thus we assume genes of the group c to have two different mean logarithmic expression levels, one in samples 1, 3, 4, 5 and 6, and another potentially different one in sample 2. We denote these two mean logarithmic expression levels by μ_{d0} (Table 4d red) and μ_{d1} (Table 4d blue) respectively.

For genes of the group a we denote the two model parameters μ_a and σ_a of the six Gaussian distributions by $\theta_a = (\mu_a, \sigma_a)$, and for each of the three groups $z \in \{b, c, d\}$ we denote the three model parameters μ_{z0}, μ_{z1}, and σ_z of the six Gaussian distributions by $\theta_z = (\mu_{z0}, \mu_{z1}, \sigma_z)$.

Assuming conditional independence of the six logarithmic expression levels given group z and model parameters θ_z, we can write the likelihood $p(x|z, \theta_z)$ of data x given group z and model parameters θ_z as a product of six univariate Gaussian distributions with the corresponding mean μ_a, or means μ_{z0} and μ_{z1}, and the corresponding variance σ_z^2 (equation 1 and equation 2). Using the maximum likelihood principle, we obtain the estimates of model parameters θ_a by equations 8a and 8b and model parameters θ_z for $z \in \{b, c, d\}$ by equations 8c - 8e.

To illustrate this approach, in Figure 4 we show for gene 'tumor potentiating region' (TPR) the six measured logarithmic expression levels together with the univariate Gaussian probability density estimated for group a and the three pairs of univariate Gaussian probability densities estimated for each of the three groups $z \in \{b, c, d\}$. Visually, it is easy to see that the model of group c best fits the expression profile of this gene, as it yields the best separation between the two estimated means and the smallest estimated pooled variance. Consistent with this visual observation, the four corresponding likelihoods of the six measured logarithmic expression levels are $p(x|a, \theta_a) = 0.004$, $p(x|b, \theta_b) = 0.035$, $p(x|c, \theta_c) = 4.22$, and $p(x|d, \theta_d) = 0.012$, i.e., the likelihood of the six measured logarithmic expression levels of gene TPR is highest for group c.

However, performing classification through model selection based on maximum likelihood is problematic when the number of free model parameters is not identical among all models under comparison. In the present work, model a has two free model parameters, while models b, c, and d have three. Hence, a naive classification based on a maximum likelihood criterion would give a spurious advantage to models b, c, and d with three free model parameters over model a with only two parameters. In order to eliminate that spurious advantage, we compute marginal likelihoods $p(x|z)$ using the approximation of Schwarz et al. [12] commonly referred to as Bayesian Information Criterion (Section 4.5). Applying this approximation to gene TPR we obtain the four marginal likelihoods of the six measured logarithmic expression levels $p(x|a, \theta_a) = 0.001$, $p(x|b, \theta_b) = 0.002$, $p(x|c, \theta_c) = 0.287$, and $p(x|d, \theta_d) = 0.001$. Consistent with the visual observation of Figure 4, we find that the marginal likelihood for group c is highest.

To obtain the approximate posterior probability $p(z|x)$, we now simply use Bayes' formula $p(z|x) = \frac{p(x|z) \cdot p(z)}{p(x)}$ for group $z \in \{a, b, c, d\}$, where $p(z)$ is the prior probability of group z, and the denominator $p(x)$ is the sum of the four numerators $p(x|z) \cdot p(z)$ for $z \in \{a, b, c, d\}$. We assume that 70% of all genes are not regulated by EGF, so we define the prior probability for group a by $p(a) = 0.70$, and we further assume that the remaining 30% of the genes fall equally in groups with EGF-regulation, so we define the prior probabilities for groups b, c, and d by $p(b) = p(c) = p(d) = 0.1$. Using these prior probabilities, we obtain for gene TPR the four approximate posterior probabilities $p(a|x) = 0.016$, $p(b|x) = 0.008$, $p(c|x) = 0.973$, and $p(d|x) = 0.003$. We find that the approximate posterior probability for group c is highest, so we finally assign gene TPR to group c.

By applying this approach of computing the four approximate posterior probabilities for each gene and assigning each gene to that group z with the highest approximate posterior probability, we classify to group a $8,449$ genes, to group b $3,822$ genes, to group c $3,143$ genes, and to group d $1,328$ genes.

2.3 Identification of genes belonging to group c

We define the subset of the $1,140$ genes with an approximate posterior probability for group c exceeding 0.75 as putative target genes regulated by EGFR isoforms II-IV and not by other receptors, and we scrutinize six of these genes in the sequel. Three of these are up-regulated genes with an approximate posterior $p(c|x) > 0.75$ and with a $\log_2$-fold change $\hat{\mu}_{c1} - \hat{\mu}_{c0} > 0.5$: CKAP2L, ROCK1, and TPR, and

three are down-regulated genes with an approximate posterior $p(c|x) > 0.75$ and with a $\log_2$-fold change $\hat{\mu}_{c1} - \hat{\mu}_{c0} < -0.5$: ALDH4A1, CLCA2, and GALNS.

In order to validate the six logarithmic expression levels $x_1, \ldots, x_6$ of each of the six genes CKAP2L, ROCK1, TPR, ALDH4A1, CLCA2, and GALNS, we perform 108 qPCR experiments comprising three biological replicates for each gene and each treatment. Figure 5 shows the six $\log_2$-fold changes $\hat{\mu}_{c1} - \hat{\mu}_{c0}$ of the microarray experiments and of the qPCR experiments. We find that the six $\log_2$-fold changes of the microarray experiments and those of the qPCR experiments are not identical, but in good agreement, yielding a Pearson correlation coefficient of 0.99. Moreover, the error bars, computed as the square root of the pooled standard deviation of group c (equation 8e), of all six genes overlap between microarray experiments and qPCR experiments.

Furthermore, the good correlation of microarray and qPCR experiment suggests that our microarray data is valid and next we investigate the biological context of the six genes in group c.

Recently, cytoskeleton-associated protein2 (CKAP2) was reported as a possible new prognostic breast cancer marker. It was shown that the chromatin CKAP2-positive cell count (CPCC) is a significant prognostic indicator for disease-free survival in the HER2-negative luminal subtype of breast cancer patients [13].

The Rho-associated kinase 1 (ROCK1) is known to play an important role in the EGF-induced formation of stress fibers in keratinocyte [14] and to be involved in the cofilin pathway in breast cancer [15]. In addition, ROCK1 has been found to promote migration, metastasis, and invasion of tumor cells and also to facilitate morphological cell shape transformations through modifications of the actinomyosin cytoskeleton [16].

It was found that depletion of the tumor potentiating region (TPR) by RNAi triggers G0-G1 arrest and that TPR depletion plays a role in controlling cellular senescence [17]. Moreover, TPR regulates the nuclear export of unspliced RNA and participates in processing and degradation of aberrant mRNA transcripts [18], a mechanism that is very important for the regulation of genes and their deregulation in cancer cells.

Furthermore, it was demonstrated that the aldehyde dehydrogenase 4 family member A1 (ALDH4A1) is induced in a tumor cell line in response to DNA damage in a p53-dependent manner. ALDH4 contains a potential p53 binding sequence in intron 1, and p53 is often mutated in tumor cells [19].

A second gene that is transcriptionally regulated by DNA damage in a p53-dependent manner is the calcium-activated chloride channel regulator 2 (CLCA2). Inhibition of CLCA2 stimulates cancer cell migration and invasion [20]. Moreover, hCLCA2 could be a marker of epithelial differentiation (EMT), and knockdown of hCLCA2 causes cell overgrowth as well as enhanced migration and invasion. These changes are accompanied by down-regulation of E-cadherin and up-regulation of vimentin, and loss of hCLCA2 may promote metastasis [20]. In addition, loss of breast epithelial marker hCLCA2 has been reported to promote an epithelial-to-mesenchymal transition and to indicate a higher risk of metastasis [21].

With respect to the galactosamine sulfatase (GALNS), an effect of 17β-estradiol on the expression of GALNS could be detected by qPCR experiments in a breast cancer cell line, which is a hint to a tumor association of GALNS [22].

Our analysis proves that EGFR II-IV isoforms may trigger the up-regulation of ROCK1 and the down-regulation of CLCA2. Both deregulations are positively associated with processes of migration, metastasis and invasion of tumor cells. The up-regulation of CKAP2L [23] by EGFR II-IV isoforms identified a link to processes of cell-cycle progression of stem or progenitor cells. Overall, the interpretation of the impact of EGFR II-IV isoforms on three of six validated gene transcript seems likely that these isoforms are involved in processes of migration and metastasis of clonogenic (stem) cell. This is strongly associated with a more aggressive tumor and a worse prognosis of tumor disease. We could show that our Bayesian gene selection criterion (BGSC) approach worked well for this nested experimental design and we were able to detected genes putatively regulated by EGFR isoforms II-IV and not by other receptors. Therefore, we suppose that our approach can possibly also used in similar cases of nested mRNA splice variants of receptors leading to just indirect measurable regulation of genes.

3 Conclusions

We have performed RNAi experiments to analyze the expression of three poorly investigated isoforms II-IV of the epidermal growth factor receptor in a glioblastoma cell line, and we have developed the Bayesian gene selection criterion approach for the identification of putative target genes of these splice variant isoforms. We have identified $3,143$ putative target genes, out of which $1,140$ genes have an approximate posterior greater than 0.75, and we have tested six of these genes by 108 qPCR experiments. These six genes include ROCK1, which is known to be associated with EGFR regulation, as well as CKAP2L, TPR, ALDH4A1, CLCA2, and GALNS. We have found that the six $\log_2$-fold changes of the microarray expression levels and those of the qPCR expression levels show a Pearson correlation coefficient of 0.99 (p-value = 0.00002), suggesting that the set of $1,140$ genes might possibly contain some further putative target genes of isoforms II-IV of the epidermal growth factor receptor in tumor cells.

The analysis of isoform-specific effects in combination with RNAi treatments are an elegant way to directly down-regulate specific mRNA splice variants but that often leads to a nested experimental design for which often no standard procedure exists. Our two-step BGSC procedure of first defining easily interpretable conceptual groups of genes associated with isoforms and subsequently classifying genes based on the approximated posterior to these groups seems to be a promising approach in such a situation. This approach is easily adaptable to other and more complex experimental designs. The microarray data and the R package *BGSC* (https://github.com/GrosseLab/BGSC) including a script for reproducing the results and plots are available as supplementary materials to this work.

4 Methods

4.1 Glioblastoma cell line SF767

We obtained glioblastoma cell line SF767 from Cynthia Cowdrey (Neurosurgery Tissue Bank, University of California, San Francisco, USA). We cultured cell line SF767 in RPMI1640 medium (Lonza, Walkersville, USA) containing 10% (Vol/Vol) fetal bovine serum, 1% (Vol/Vol) sodium pyruvate, 185 U/ml penicillin, and 185

$\mu g/ml$ ampicillin and maintain it at 37°C in a humidified atmosphere containing 3% (Vol/Vol) CO2.

4.2 RNAi

The design and application of siRNA specific for EGFR mRNA were performed by a program provided by MWG (Eurofins Genomics, Ebersberg, Germany). The sequences of the double-stranded EGFR-specific siRNAs correspond to a 21-bp sequences of the EGFR-cDNA (NCBI-ref NM_005228.3) for the first siRNA (siRNA_I) at positions 4,094 - 4,116 and for the second siRNA (siRNA_{ALL}) at positions 1,258 - 1,278 (Table 1). To ensure that the EGFR-specific siRNA do not interact with other transcripts, we used the sequences of siRNA_I and siRNA_{ALL} to perform a BLAST search with Nucleotide BLAST against the human genome database (http://www.ncbi.nlm.nih.gov/) and the siRNA-Check of SpliceCenter suite [24]. To prevent off-target effects of siRNA-treatment, we transfected cells with 50 nM targeting siRNA (siRNA_I and siRNA_{ALL}) in RPMI complete medium. For transfecting we use the reagent INTERFERin$^{\text{TM}}$ according to the manufacturer's instructions (Polyplus Transfection, Illkirch, France). We verified the efficiency of knockdown by siRNA treatment by analyzing EGFR-mRNA by qPCR as well as EGFR-protein level via Western blot technique [10].

4.3 Illumina BeadChip Microarray

Before microarray analysis RNA integrity and concentration was examined on an Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA, USA) using the RNA 6.000 LabChip Kit (Agilent Technologies) according to the manufacturer's instructions. Illumina BeadChip analysis was conducted at the microarray core facility of the Interdisciplinary Center for Clinical Research (IZKF) Leipzig (Faculty of Medicine, University of Leipzig). 250 ng RNA per sample were ethanol precipitated with GlycoBlue (Invitrogen) as carrier and dissolved at a concentration of 100-150 $ng/\mu l$ prior to probe synthesis using the TargetAmp$^{\text{TM}}$- Nano Labeling Kit for Illumina Expression BeadChip (Epicentre Biotechnologies, Madison, WI, USA). 750 Fng of cRNA were hybridized to Illumina HT-12 v4 Expression BeadChips (Illumina, San Diego, CA, USA) and scanned on the Illumina HiScan instrument according to the manufacturer's specifications.

The *read.ilmn* function of the *limma* package [25] was used to read the 47317 microarray probes into *R*. The *neqc* function of *limma* was used to perform a background correction followed by quantile normalization, using negative control probes for background correction and both negative and positive controls for normalization. The $16,742$ Array probes corresponding to $14,389$ genes which displayed a significant hybridisation signal (Illumina signal detection statistic at $P < 0.05$) in all probes are used for the further analysis.

4.4 Experimental design

For investigating which genes are activated by the four EGFR isoforms I - IV in glioblastoma cell line SF767, we use RNAi, as described in section 4.2, for a selective down-regulation of EGFR splice variants (Table 3 rows) with and without EGF treatment (Table 3 columns). Specifically, we applied the three different RNAi

treatments – (a) control without RNAi, (b) RNAi with siRNA$_I$, and (c) RNAi with siRNA$_{ALL}$– to glioblastoma cell line SF767. We did not need to perform an RNAi double treatment with siRNA$_{ALL}$ and siRNA$_I$, because the effect of siRNA$_I$ is nested in the effect of siRNA$_{ALL}$ and so the double treatment would have the same effect as RNAi with siRNA$_{ALL}$ (Table 2).

In case (a), we performed a control experiment without RNAi treatment. In this case, EGFR is not down-regulated by a siRNA, so target genes of all EGFR splice variants and of other receptors of EGF should be differentially expressed in columns 1 and 2, i.e., should have different logarithmic expression levels x_1 and x_2.

In case (b), we performed an RNAi with siRNA$_I$, which can bind only to the full-length splice variant I of EGFR. Hence, siRNA$_I$ down-regulates splice variant I, but not the other splice variants II-IV, and in this case target genes of EGFR splice variants II-IV and of other receptors of EGF should be differentially expressed in columns 1 and 2, i.e., should have different logarithmic expression levels x_3 and x_4.

In case (c), we performed an RNAi with siRNA$_{ALL}$, which can bind to all four splice variants of EGFR, and which subsequently down-regulates them. In this case, only target genes of other receptors of EGF should be differentially expressed in columns 1 and 2, i.e., should have different logarithmic expression levels x_5 and x_6.

4.5 Probabilistic modeling of gene expression

We proposed a probabilistic model for the logarithmic expression pattern $x = (x_1, \ldots, x_6)$ for each of the four groups $z \in \{a, b, c, d\}$.

First, we assumed that the three logarithmic expression levels x_1, x_3, and x_5 corresponding to no EGF treatment are similar to each other, which corresponds to the assumption that the RNAi treatment should have no effect in case of no EGF treatment. Second, we assumed that the three logarithmic expression levels x_2, x_4, and x_6 follow the expression patterns described in Section 2.1 and summarized in Figure 3.

In order to mathematically formulate the model assumptions, we introduced six indicator variables $g_1, \ldots, g_6$ for the groups $\tilde{z} \in \{b, c, d\}$ that indicate if the six logarithmic expression levels $x_1, \ldots, x_6$ are expected to be different from x_1. Specifically, we defined $g_n = 1$ if x_n is expected to be different from x_1 for $n = 1, \ldots, 6$ and $g_n = 0$ otherwise. genes of the group a are defined as showing no effect on the EGF treatment and therefore g_n is by definition 0.

By definition, we obtained that $g_1 = 0$ for each of the three groups $\tilde{z}$. By the first model assumption, we obtained that $g_1 = g_3 = g_5 = 0$ for each of the three groups $\tilde{z}$. By the second model assumption, we obtained that (g_2, g_4, g_6) is equal to the corresponding column of Figure 3 for each of the three groups $\tilde{z}$. Table 4 summarizes the values of the indicator variables $g_1, \ldots, g_6$ for each of the three groups $b - d$.

Third, we assumed that the logarithmic expression levels $x_1, \ldots, x_6$ are statistically independent and normally distributed. By combining all three model assump-

tions, we obtained the likelihood

$$p(x|a,\theta_a) = \prod_{n=1}^{6} \mathcal{N}(x_n|\mu_a,\sigma_a) \tag{1}$$

$$p(x|\tilde{z},\theta_{\tilde{z}}) = \prod_{n=1}^{6} \mathcal{N}(x_n|\mu_{\tilde{z}g_n},\sigma_{\tilde{z}}) \tag{2}$$

for each of the four gene groups $z \in \{a,b,c,d\}$, where

$$\mathcal{N}(x_n|\mu_a,\sigma_a) = \frac{1}{\sqrt{2\pi}\sigma_a} \times e^{-\frac{(x_n-\mu_a)^2}{2\sigma_a^2}} \tag{3}$$

$$\tag{4}$$

denotes the density of the normal distribution, $\theta_a = (\mu_a,\sigma_a)$ denotes the parameter of model a and

$$\mathcal{N}(x_n|\mu_{\tilde{z}g_n},\sigma_{\tilde{z}}) = \frac{1}{\sqrt{2\pi}\sigma_{\tilde{z}}} \times e^{-\frac{(x_n-\mu_{\tilde{z}g_n})^2}{2\sigma_{\tilde{z}}^2}} \tag{5}$$

denotes the density of the normal distribution, $\theta_{\tilde{z}} = (\mu_{\tilde{z}0},\mu_{\tilde{z}1},\sigma_{\tilde{z}})$ denotes the parameter of model $\tilde{z}$, and g_n are the indicator variables from Table 4.

4.6 Posterior approximation by the Bayesian Information Criterion

Next, we seek the approximate posterior

$$p(z|x) = \frac{p(x|z)\cdot p(z)}{p(x)} \tag{6}$$

for each $z \in \{a,b,c,d\}$ and each gene, where $p(z)$ is the prior probability of group z.

For the four models of Section 4.5, the approximations of the marginal likelihoods based on the Bayesian Information Criterion are

$$p(x|z) \propto \frac{p(x|z,\hat{\theta}_z)}{\sqrt{6}^{|\theta_z|}}, \tag{7}$$

where 6 is the number of data points and $|\theta_z|$ is the number of free parameters of model z, which is 2 for group a and 3 for groups $b-d$, and where the maximum-likelihood estimators $\hat{\theta}_z$ are

$$\hat{\mu}_a = \frac{\sum_{n=1}^{6} x_n}{6} \tag{8a}$$

$$\hat{\sigma}_a^2 = \frac{1}{5}\left(\sum_{n=1}^{6}(x_n - \hat{\mu}_a)^2\right) \tag{8b}$$

$$\hat{\mu}_{\tilde{z}0} = \frac{\sum_{n=1}^{6} x_n \cdot (1 - g_{\tilde{z}n})}{\sum_{n=1}^{6}(1 - g_{\tilde{z}n})} \tag{8c}$$

$$\hat{\mu}_{\tilde{z}1} = \frac{\sum_{n=1}^{6} x_n \cdot g_{\tilde{z}n}}{\sum_{n=1}^{6} g_{\tilde{z}n}} \tag{8d}$$

$$\hat{\sigma}_{\tilde{z}}^2 = \frac{1}{4}\left(\sum_{n=1}^{6}(x_n - \hat{\mu}_{\tilde{z}0})^2 \cdot (1 - g_{\tilde{z}n}) + \sum_{n=1}^{6}(x_n - \hat{\mu}_{\tilde{z}1})^2 \cdot g_{\tilde{z}n}\right) \tag{8e}$$

for $\tilde{z} \in \{b, c, d\}$, and where $g_{\tilde{z}n}$ denotes the indicator variable g_n of group $\tilde{z}$. Based on these approximations, we computed $p(z|x)$ and then performed Bayesian model selection by assigning each gene to that group z with the maximum approximate posterior $p(z|x)$.

Competing interests
The authors declare that they have no competing interests.

Author's contributions
HW, JK, MK, and AWE designed and performed the biological experiments. CW and IG designed the algorithm, CW implemented the algorithm, and CW and IG performed the data analysis. CW, HW, MK, DA and IG wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements
We thank Ralf Eggeling, Ioana Lemnian, and Teemu Roos for valuable discussions and the Microarray Core Facility of the Interdisciplinary Center of Clinical Research (IZKF) at Leipzig for performing the microarray experiments.

Author details
[1]Institute of Computer Science, Martin Luther University Halle–Wittenberg , Germany. [2]Department of Oral and Maxillofacial Plastic Surgery, Martin Luther University Halle–Wittenberg, Germany. [3]German Center of Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Germany. [4]Program in Quantitative and Systems Biology, School of Natural Sciences, University of California, Merced, CA, USA.

References

1. Sridhar, S.S., Seymour, L., Shepherd, F.A.: Inhibitors of epidermal-growth-factor receptors: a review of clinical research with a focus on non-small-cell lung cancer. The lancet oncology **4**(7), 397–406 (2003)
2. Ohgaki, H., Kleihues, P.: Epidemiology and etiology of gliomas. Acta neuropathologica **109**(1), 93–108 (2005)
3. Ohgaki, H., Kleihues, P.: Genetic pathways to primary and secondary glioblastoma. The American journal of pathology **170**(5), 1445–1453 (2007)
4. Yarden, Y.: The EGFR family and its ligands in human cancer: signalling mechanisms and therapeutic opportunities. European journal of cancer **37**, 3–8 (2001)
5. Citri, A., Yarden, Y.: EGF–ERBB signalling: towards the systems level. Nature reviews Molecular cell biology **7**(7), 505–516 (2006)
6. Maramotti, S., Paci, M., Manzotti, G., Rapicetta, C., Gugnoni, M., Galeone, C., Cesario, A., Lococo, F.: Soluble Epidermal Growth Factor Receptors (sEGFRs) in Cancer: Biological Aspects and Clinical Relevance. International journal of molecular sciences **17**(4), 593 (2016)
7. Wilken, J.A., Perez-Torres, M., Nieves-Alicea, R., Cora, E.M., Christensen, T.A., Baron, A.T., Maihle, N.J.: Shedding of soluble epidermal growth factor receptor (sEGFR) is mediated by a metalloprotease/ fibronectin/ integrin axis and inhibited by cetuximab. Biochemistry **52**(26), 4531–4540 (2013)
8. Reiter, J.L., Maihle, N.J.: Characterization and expression of novel 60-kda and 110-kda EGFR isoforms in human placenta. Annals of the New York Academy of Sciences **995**(1), 39–47 (2003)
9. Izquierdo, M.: Short interfering RNAs as a tool for cancer gene therapy. Cancer gene therapy **12**(3), 217–227 (2004)
10. Wichmann, H., Güttler, A., Bache, M., Taubert, H., Rot, S., Kessler, J., Eckert, A.W., Kappler, M., Vordermark, D.: Targeting of EGFR and HER2 with therapeutic antibodies and siRNA. Strahlentherapie und Onkologie, 1–12 (2014)
11. Long, A.D., Mangalam, H.J., Chan, B.Y.P., Tolleri, L., Hatfield, G.W., Baldi, P.: Improved statistical inference from DNA microarray data using analysis of variance and A Bayesian statistical framework. Analysis of global

gene expression in Escherichia coli K12. Journal of Biological Chemistry **276**(23), 19937–19944 (2001). doi:10.1074/jbc.M010192200
12. Schwarz, G.: Estimating the dimension of a model. Ann. Statist. **6**(2), 461–464 (1978)
13. Sim, S.H., Bae, C.-D., Kwon, Y., Hwang, H.-L., Poojan, S., Hong, H.-I., Kim, K., Kang, S.-H., Kim, H.-S., Um, T.-H., Park, I.H., Lee, K.S., Jung, S.-Y., Lee, S., Kang, H.-S., Lee, E.S., Kim, M.-K., Hong, K.-M., Ro, J.: CKAP2 (cytoskeleton-associated protein2) is a new prognostic marker in HER2-negative luminal type breast cancer. PLOS ONE **12**(8), 0182107 (2017). doi:10.1371/journal.pone.0182107
14. Ohuchi, H.: Wakayama symposium: Epithelial-mesenchymal interactions in eyelid development. The ocular surface **10**(4), 212–216 (2012)
15. Wang, W., Eddy, R., Condeelis, J.: The cofilin pathway in breast cancer invasion and metastasis. Nature Reviews Cancer **7**(6), 429–440 (2007)
16. Rath, N., Olson, M.F.: Rho-associated kinases in tumorigenesis: reconsidering ROCK inhibition for cancer therapy. EMBO reports **13**(10), 900–908 (2012)
17. David-Watine, B.: Silencing nuclear pore protein Tpr elicits a senescent-like phenotype in cancer cells. PloS one **6**(7), 22423 (2011)
18. Rajanala, K., Nandicoori, V.: Localization of nucleoporin Tpr to the nuclear pore complex is essential for Tpr mediated regulation of the export of unspliced RNA. PloS one **7**(1), 29921 (2012)
19. Yoon, K., Nakamura, Y., Arakawa, H.: Identification of ALDH4 as a p53-inducible gene and its protective role in cellular stresses. Journal of human genetics **49**(3), 134–140 (2004)
20. Sasaki, Y., Koyama, R., Maruyama, R., Hirano, T., Tamura, M., Sugisaka, J., Suzuki, H., Idogawa, M., Shinomura, Y., Tokino, T.: CLCA2, a target of the p53 family, negatively regulates cancer cell migration and invasion. Cancer biology & therapy **13**(14), 1512–1521 (2012)
21. Walia, V., Yu, Y., Cao, D., Sun, M., McLean, J., Hollier, B., Cheng, J., Mani, S., Rao, K., Premkumar, L., Elble, R.: Loss of breast epithelial marker hCLCA2 promotes epithelial-to-mesenchymal transition and indicates higher risk of metastasis. Oncogene **31**(17), 2237–2246 (2011)
22. Garcia, S., Nagai, M.: Transcriptional regulation of bidirectional gene pairs by 17-β-estradiol in MCF-7 breast cancer cells. Brazilian Journal of Medical and Biological Research **44**(2), 112–122 (2011)
23. Yumoto, T., Nakadate, K., Nakamura, Y., Sugitani, Y., Sugitani-Yoshida, R., Ueda, S., Sakakibara, S.-i.: Radmis, a novel mitotic spindle protein that functions in cell division of neural progenitors. PloS one **8**(11), 79895 (2013)
24. Ryan, M.C., Zeeberg, B.R., Caplen, N.J., Cleland, J.A., Kahn, A.B., Liu, H., Weinstein, J.N.: Splicecenter: a suite of web-based bioinformatic applications for evaluating the impact of alternative splicing on RT-PCR, RNAi, microarray, and peptide-based studies. BMC bioinformatics **9**(1), 313 (2008)
25. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., Smyth, G.K.: limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic acids research, 007 (2015)

Figures

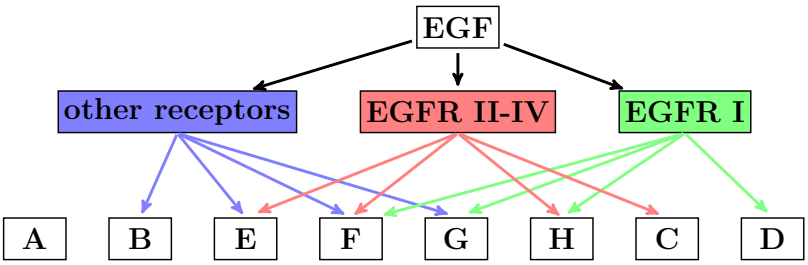
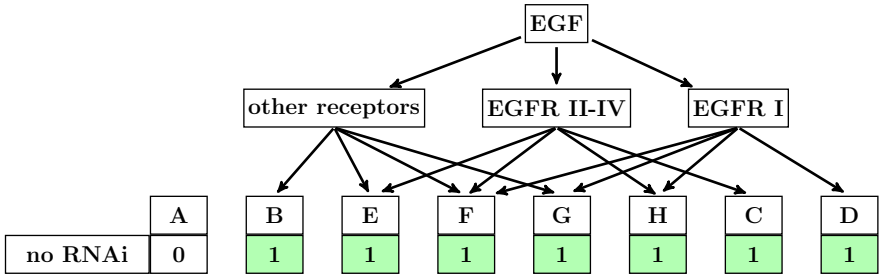
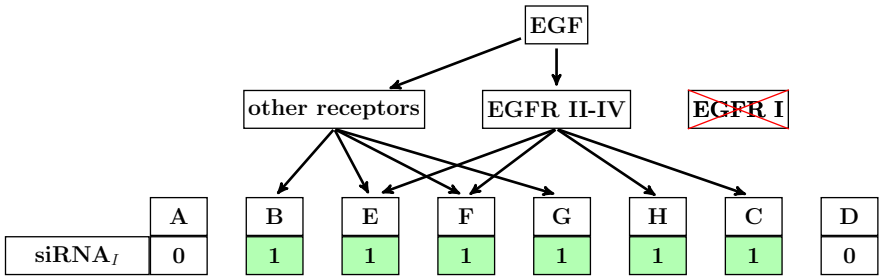


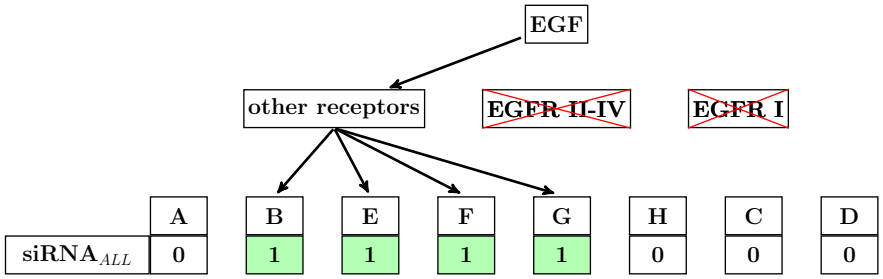
Figure 1: **Graphical representation of eight different modes of transcriptional regulation** by EGFR or other receptors of EGF. Each gene can be regulated by splice by some combination of splice variant I of EGFR (green arrows), splice variants II-IV of EGFR (red arrows), and other receptors of EGF (blue arrows), resulting in eight gene groups `A` - `H`.



(a) Control treatment without RNAi.



(b) RNAi with siRNA$_I$.



(c) RNAi with siRNA$_{ALL}$.

Figure 2: **Graphical representation of the groups** of genes differentially expressed by EGF treatment for each of the three cases of RNAi treatment. These eight gene groups (`A` - `H`) contains all theoretical possible models of regulation of a gene, after EGF addition. Sub-figure (a) corresponds to the control experiment without RNAi treatment, sub-figure (b) corresponds to RNAi treatment with siRNA$_I$, and sub-figure (c) corresponds to RNAi treatment with siRNA$_{ALL}$. Red crosses indicate the down-regulation of EGFR by RNAi treatment with siRNA$_I$ (b) or siRNA$_{ALL}$ (c). The change of gene expression (up or down) by EGF treatment is indicated by 1 and no change by 0, i.e., all genes except those of gene group `A` should be differentially expressed in the control experiment (a), all genes except those of gene groups `A` and `D` should be differentially expressed in experiment (b), and all genes except those of gene groups `A`, `C`, `D`, and `H` should be differentially expressed in experiment (c).

	Differential expression by EGF							
group	**A**	**B**	**E**	**F**	**G**	**H**	**C**	**D**
no RNAi	**0**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
siRNA$_I$	**0**	**1**	**1**	**1**	**1**	**1**	**1**	**0**
siRNA$_{ALL}$	**0**	**1**	**1**	**1**	**1**	**0**	**0**	**0**
simplified group	*a*	*b*				*c*		*d*

Figure 3: **Reduction of the conceptual gene groups.** genes of the group `A` are never differentially expressed by EGF treatment. genes of the group `B,E,F, and G` are always differentially expressed by EGF treatment. genes of the group `C and H` are differentially expressed by EGF treatment in case of control treatment (no RNAi) or simultaneously treatment with siRNA$_I$, whereas not differentially expressed by EGF treatment in case of simultaneous treatment with siRNA$_{ALL}$. genes of the group `D` are differentially expressed by EGF treatment in case of control treatment (no RNAi), whereas not differentially expressed by EGF treatment in case of simultaneously treatment with siRNA$_I$ or siRNA$_{ALL}$.

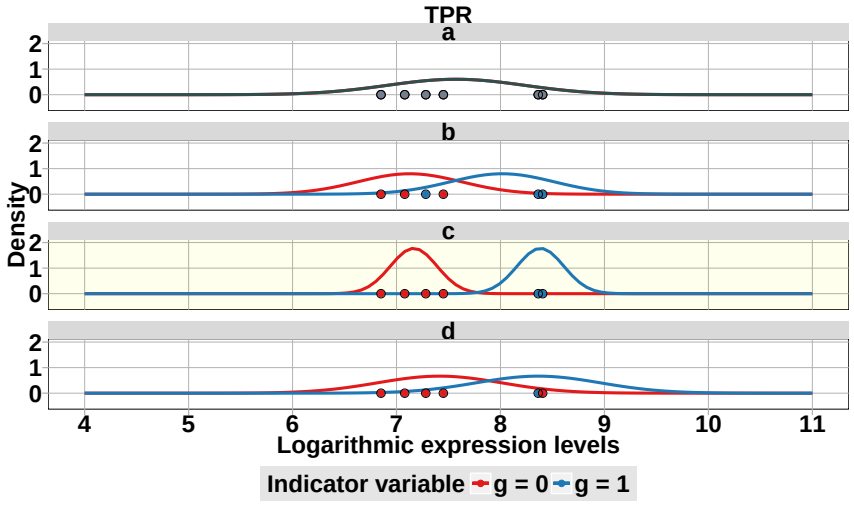


Figure 4: **Probability density plot of the Gaussian distributions of TRP** For group a we plot as black curve the probability density of the Gaussian distribution and mark the $\log_2$-expression values $x_1, \ldots, x_6$ of TPR with black circles which are colored according to Table 4a. For groups $b - d$ we plot with red and blue curves the probability densities of the Gaussian distributions and mark the $\log_2$-expression values $x_1, \ldots, x_6$ of TPR with circles which are colored according to indicator variables from Table 4 $g = 0$ in red and $g = 1$ in blue. For group a we assume that all six expression levels stem from the same Gaussian. In this case, the mean μ and standard deviation σ of this Gaussian (black) are equal to μ and σ of the six expression levels. For the groups b - d we assume that all six expression levels stem from a mixture of two Gaussian distributions with independent means μ_0 and μ_1, and one pooled standard deviation σ. For groups b - d we assume that the expression levels x_1, x_3, and x_5, x_1, x_3, x_4, and x_5, and x_1, x_3, x_4, x_5, and x_6 stem from the Gaussian based on μ_0 (red), respectively. For groups b - d we assume that the expression levels x_2, x_4, and x_6, x_2 and x_4, and x_2 stem from the Gaussian based on μ_1 (blue), respectively.

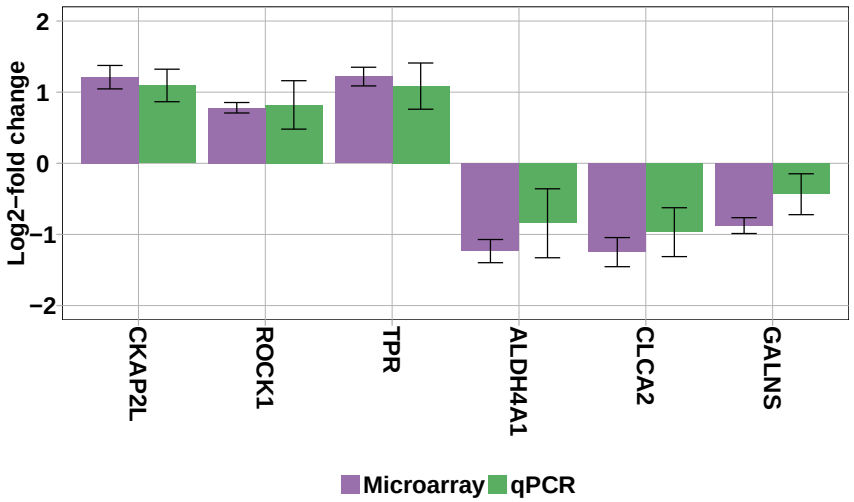


Figure 5: **Comparison of microarray and qPCR $\log_2$-fold changes.** Based on the microarray expression data described in Sections 2, 3, 4, and 5, we obtain an up-regulation for genes CKAP2L, ROCK1, and TPR and a down-regulation for genes ALDH4A1, CLCA2, and GALNS. The error bars of group c are the pooled standard error using the Satterthwaite approximation. Based on the qPCR data, we obtain qualitatively and quantitatively similar results with overlapping error bars, yielding a Person correlation coefficient of the $\log_2$-fold changes of the microarray experiments and those of the qPCR experiments of 0.99.

Tables

Table 1: Design of siRNA$_{ALL}$ and siRNA$_I$

siRNA	sequence $5' \rightarrow 3'$	localization	target-mRNA	corresponding mRNA
I	AACGCAUCCAGCAAGAAUA	4098-4116	EGFR I	NM_005228.3
ALL	CGGAAUAGGUAUUGGUGAA	1260-1278	EGFR I	NM_005228.3
			EGFR II	NM_201282.1
			EGFR III	NM_201283.1
			EGFR IV	NM_201284.1

Table 2: Experimental design where the columns present the RNAi treatment – without RNAi, RNAi with siRNA$_I$, and RNAi with siRNA$_{ALL}$– and the row present if the treatment is applied (+ applied , - not applied). The six corresponding logarithmic expression values per gene are denoted by $x_1, \ldots, x_6$. The nested experiments marked with $*$ (asterisk) would lead to equal expressions and therefore these experiments are not needed.

EGF	siRNA$_I$	siRNA$_{ALL}$	expression value
−	−	−	x_1
−	+	−	x_3
−	−	+	x_5
−	+	+	${x_5}^*$
+	−	−	x_2
+	+	−	x_4
+	−	+	x_6
+	+	+	${x_6}^*$

Table 3: Experimental design where the rows present the RNAi treatment – without RNAi, RNAi with siRNA$_I$, and RNAi with siRNA$_{ALL}$– and the columns present the EGF treatment. The six corresponding logarithmic expression values per gene are denoted by $x_1, \ldots, x_6$.

	no EGF	EGF
no RNAi	x_1	x_2
RNAi by siRNA$_I$	x_3	x_4
RNAi by siRNA$_{ALL}$	x_5	x_6

Table 4: **Schematic Expression pattern** For the gene groups $b - d$ the indicator variables g_n are equal to 0 if the logarithmic expression levels x_n are expected to be similar to x_1 and 1 otherwise. The four no-EGF columns are equal to 0 by model assumption 1, and the four EGF columns are equal to the corresponding columns of Figure 3 by model assumption 2. For group a the indicator variables g_n are by definition equal to 0.

group a	**no EGF**	**EGF**
no RNAi	$g_1 = 0$	$g_2 = 0$
siRNA$_I$	$g_3 = 0$	$g_4 = 0$
siRNA$_{ALL}$	$g_5 = 0$	$g_6 = 0$

(a) Expression pattern for group a.

group b	**no EGF**	**EGF**
no RNAi	$g_1 = 0$	$g_2 = 1$
siRNA$_I$	$g_3 = 0$	$g_4 = 1$
siRNA$_{ALL}$	$g_5 = 0$	$g_6 = 1$

(b) Expression pattern for group b.

group c	**no EGF**	**EGF**
no RNAi	$g_1 = 0$	$g_2 = 1$
siRNA$_I$	$g_3 = 0$	$g_4 = 1$
siRNA$_{ALL}$	$g_5 = 0$	$g_6 = 0$

(c) Expression pattern for group c.

group d	**no EGF**	**EGF**
no RNAi	$g_1 = 0$	$g_2 = 1$
siRNA$_I$	$g_3 = 0$	$g_4 = 0$
siRNA$_{ALL}$	$g_5 = 0$	$g_6 = 0$

(d) Expression pattern for group d.