Albrecht *et al. BMC Bioinformatics* (2017) 18:33
DOI 10.1186/s12859-016-1440-8

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

TTCA: an R package for the identification of differentially expressed genes in time course microarray data

Marco Albrecht[1,2*], Damian Stichel[1,3], Benedikt Müller[4], Ruth Merkle[5,6], Carsten Sticht[7], Norbert Gretz[7], Ursula Klingmüller[5,6], Kai Breuhahn[4] and Franziska Matthäus[1,8]

Abstract

Background: The analysis of microarray time series promises a deeper insight into the dynamics of the cellular response following stimulation. A common observation in this type of data is that some genes respond with quick, transient dynamics, while other genes change their expression slowly over time. The existing methods for detecting significant expression dynamics often fail when the expression dynamics show a large heterogeneity. Moreover, these methods often cannot cope with irregular and sparse measurements.

Results: The method proposed here is specifically designed for the analysis of perturbation responses. It combines different scores to capture fast and transient dynamics as well as slow expression changes, and performs well in the presence of low replicate numbers and irregular sampling times. The results are given in the form of tables including links to figures showing the expression dynamics of the respective transcript. These allow to quickly recognise the relevance of detection, to identify possible false positives and to discriminate early and late changes in gene expression. An extension of the method allows the analysis of the expression dynamics of functional groups of genes, providing a quick overview of the cellular response. The performance of this package was tested on microarray data derived from lung cancer cells stimulated with epidermal growth factor (EGF).

Conclusion: Here we describe a new, efficient method for the analysis of sparse and heterogeneous time course data with high detection sensitivity and transparency. It is implemented as R package TTCA (transcript time course analysis) and can be installed from the Comprehensive R Archive Network, CRAN. The source code is provided with the Additional file 1.

Keywords: Differential expression, Time series, EGF, Stimulation experiments, Gene ontology, Gene set analysis

Background

Time course microarray experiments are frequently conducted to study the dynamics of gene expression at several consecutive time points. Associated data sets often require own custom-made analysis strategies, and cannot been adequately exploited with standard methods which were established to compare groups. The variability of the dynamics, spanning from fast and transient to slower, long-lasting changes, is a challenge for the analysis of time series microarray data. In perturbation experiments, sampling frequency is often adapted to reflect the expected changes in gene expression. This kind of experimental design leads to irregularly sampled data sets. Irregular time sampling may also arise when time points are chosen to be omitted after quality control, for instance when the respective arrays represent outliers with respect to the global trajectory resulting from principal component analysis (PCA) as shown in Fig. 1. If replicates are considered, their number may also vary due to the experimental design or quality issues. Often time course-data provide only one replicate per time point.

*Correspondence: marco.albrecht@posteo.de
[1]Complex Biological Systems Group (BIOMS/IWR), Heidelberg, Im Neuenheimer Feld 294, 69120 Heidelberg, Germany
[2]Systems Biology Group, Université du Luxembourg, 7, avenue du Swing, L-4367 Belvaux, Luxembourg
Full list of author information is available at the end of the article

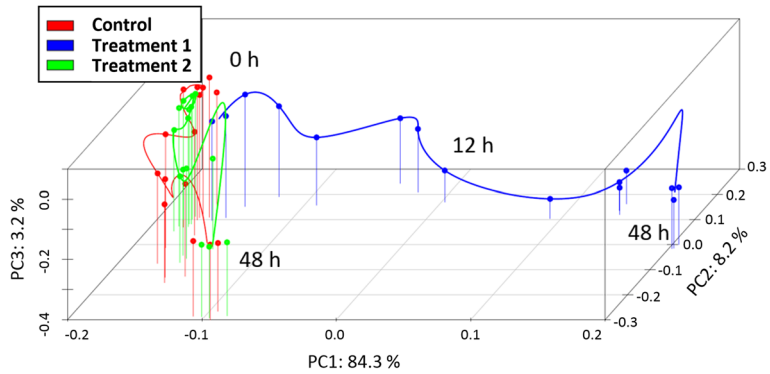

Fig. 1 Trajectory of the transcriptomes: Axes represent principal components explaining 95.7% of the variability in the data. Measurement points represent the entire transcriptome under three different stimulation experiments projected onto the first three principal components. For early time periods, all three transcriptomes correlate very well with each other. Over time, the transcriptomes develop stimulus dependent. Stimulus 1 leads to a strong change in the transcriptome, while stimulus 2 has a much smaller effect. Possible outliers are measurement points that show a large distance from the trajectory or from related replicates

The first methods applied on time course microarrays including SAM [1], ANOVA [2] and Limma [3] where extensions of methods for contrasts between states and do not include the order of time points into the analysis [4].

EDGE was one of the first methods taking the time sequence into account [5, 6]. EDGE involves a fit of natural cubic splines to gene expression profiles, and a bootstrap approach providing a reference distribution. MaSigPro (Microarray Significant Profiles) operates in a similar manner [7]. Sohn et al. have modified EDGE by using a permutation-approach and controlling the family wise error rate [8]. Later, they applied the FWER as a significance threshold and made the method more robust using quantile regression [9]. These methods have three drawbacks when used to analyse sparse data containing sharp transient expression changes. First, the information of the time course measurements is underestimated. Biologically meaningful peaks might be overlooked when the related measurement points are rejected as outliers. Second, the information of the permuted reference time course is overestimated. The permutation of the measurement points within the time sequence is often used to produce reference data of the same distribution, but without the original ordered pattern of dynamic changes. This estimation of the error rate can fail in sparse data sets when the expression dynamics exhibit a sharp peak. Here, permutation of the time points merely shifts but does not wipe out the peak. With this method, the signal-to-noise ratio of genes displaying fast variations in expression can be underestimated and related genes are erroneously removed from analysis. The third problem is that a large number of computationally expensive permutations is required, to avoid granularity in the resulting ranking [4]. Granularity refers in this case to hundreds of genes with exactly the same p-value. Repeated application of the method may shift a gene to another p-value cluster, which impedes reproducibility of the results.

An alternative method using multivariate empirical Bayes statistics and one-sample Hotelling T^2 statistics is implemented in the R package timecourse [10]. This package does not provide a significance threshold and requires a minimum number of replicates. Also, BETR (Bayesian Estimation of Temporal Regulation) [11], which uses random-effects models and considers co-expression, relies on time point replicates. Network-based methods combine cluster analysis with detection of differential expression and focus also on co-expression [12, 13]. But co-expression is a very strict assumption for the extraction of differentially expressed genes from time course data. In tightly regulated and dynamic gene regulatory networks, it seems to be very unlikely that cells do not regulate their genes at any of the sampled time points. Some of the target genes could have a negative feedback loop and could block their own expression, which could explain fast transient dynamic changes, while other target genes could have a positive feedback loop and therefore maintain gene expression longer. Additional regulation could happen after a longer time or very fast without protein translation, i.e. with functional large non-coding RNAs [14]. Longitudinal co-expression might overlook target genes that are affected by the stimulus, but which are additionally regulated by other dynamic mechanisms. Moreover, the longer the sampled time period is, the higher is the risk that initially unaffected genes show co-expression behaviour due to completely different mechanisms without relation to the stimulus. The risk is higher to detect false positive target genes.

Methods based on Gaussian processes select differentially expressed genes from one channel experiments [15] and from two channel experiments [16], implemented in

the R package gprege. However, the implemented Gaussian processes suffer from massive computational cost and the required time point replication. An alternative for two channel experiments is BATS (Bayesian Analysis of Time Series) [17, 18].

Another class of time course methods is based on principal component analysis (PCA) [19]. Inspired by a trend in the data analysis to fit the *true underlying functions* [20, 21], methods based on functional PCA (FPCA) were developed [22, 23]. The most recent method [23] can handle single replicated time course data, predict individual dynamics with PACE (Principal Component Analysis through Conditional Expectation) [24] and yields reasonable results for moderately slow expression dynamics. This method was successfully applied to clinical data derived from immune response studies [25]. For the data set considered in our study, involving perturbation experiments on cell cultures with fast expression changes, this method did not perform reliably. In particular, we observed counterintuitive differences between our original data and the original data being displayed by this method after preliminary transformation by PACE. First, the method transforms flat gene profiles into profiles exhibiting strong temporal changes, shown in Additional file 2: Figure S1A. Second, the transformed trajectories are too stiff to follow sharp peak behaviour like in Additional file 2: Figure S1B. This happens before the actual time course analysis method is applied.

Finally, even simple methods can yield good results for sparse data, for instance by computing distances or the area between curves [26, 27]. Also, a sliding window, capturing a small subset of consecutive measurement points, was discussed, but cannot be applied to non-equidistant measurements [4].

To sum up, most existing methods cannot reliably analyse sparse and irregularly sampled time course gene expression data sets. Further details and a method comparison are provided in the Additional files. A method overview is given in Additional file 2: Table S1.

Method TTCA

The method TTCA (transcript time course analysis) includes different scores to identify genes showing differential expression dynamics of various kinds.

The *dynamics score* $\mathcal{D}_i$ captures slow gene expression dynamics, the *peak score* $\mathcal{P}_i$ selects fast transient expression changes, the *integral score* $\mathcal{I}_i$ accounts for absolute changes in mRNA production level in different time periods, and a *relevance score* $\mathcal{R}_i$ provides information on existing references in the literature. A further option allows for gene ontology groups to be processed in a similar manner as individual genes. Additionally, the *minimum overlap score* Θ_i is computed to identify gene ontology groups with maximal separation of the group specific expression bandwidths between two conditions. Significance threshold and effect size are calculated for each score and the *consensus score* $\mathcal{C}_i$ combines the different scores for a final ranking.

For the detection of differential gene expression based on two channel microarray data, we recommend to create a constant gene expression profile as control profile. This control profile might start with the expression value of the first time point, or could be set to the average expression value of the experimentally derived gene expression profile.

The gene expression level is based on an assembled set of detected probes of 25 bp length. In this article, we focus on the expression dynamics of genes, however, those probe level signals can also be mapped to related transcripts or other longer oligonucleotides. These can equally be analysed with TTCA.

In the following section, preprocessing for microarray time series data is addressed. Next, all relevant scores and components of the proposed method are explained briefly.

Pre-processing of microarray time course data

Microarray data are usually afflicted by batch effects, i.e. unwanted variability in the samples arising from their experimental, technical and digital processing history. Batch effects can be introduced when samples are processed on different hybridisation batches (maximum 6-12 samples at once), or when a subset of the samples experienced slightly different experimental conditions (time of the day, new media, etc.). Many batch effects can be technically detected and can be removed if enough replicates are available. Microarray time course data sets are frequently sparse and the number of replicates per time point is low. In such data it is impossible to detect batch-effects [28]. Moreover, the frequently used quantile-normalisation, implemented in RMA [29], is based on the assumption that the majority of the genes shows a constant expression level. However, for time series experiments this might not be the case. Especially, cancer cells are known to have a high variability in their gene expression profiles [30]. Perturbation experiments might induce secondary gene responses that eventually result in considerable expression dynamics for a broad range of genes. It has been shown that thousands of genes can change their expression over time after stimulation [6]. Instead of using multi-array normalisation methods like RMA for time course analyses, we recommend to use within-array normalisation methods which process each array separately, independent of arrays taken at other time points. In particular, we recommend individual array standardisation with SCAN [31], which is robust against GC-content bias and some batch effects.

Dynamics score

We define a *dynamics score* in three steps based on the method EDGE [6] and its extension using quantile-regression [9].

The null hypothesis H_0 is that the stimulus does not significantly alter the expression level of gene i. Thus, the measurements of the respective conditions (i.e. treatment vs. control) are derived from the same expression pattern and can be combined for a single function fit. In Fig. 2a, the null hypothesis is represented by the fit to all measurement points without distinction between the conditions (dashed line). The alternative hypothesis H_1 is that the measurements are derived from different expression patterns, and that the two conditions have to be treated separately. Hence, the data is split into the two conditions, and each time course is fitted to an individual function (see the solid lines in Fig. 2a). The sum of the residuals of the two individual function fits should be smaller than the sum of the residuals of the single function fit to fulfil the alternative hypothesis H_1.

The fit is based on quantile regression [32]. The fitted function $g(t)$ and the residuals r_{ij} are obtained by minimising

$$\sum_{j=1}^{n} \rho_{0.5}(y_j - g(t_j)) - \underbrace{\lambda \int |g''(t)|dt}_{\text{smoothes the function}} . \quad (1)$$

The quantile regression algorithm is symbolised by $\rho_{0.5}$, and implemented in the R-package Quantreg [33] in function rqss(). The index 0.5 indicates the use of the median to provide the most robust curve fit. The continuous function g is fitted to the measurements $y_j, j \in \{1, \ldots, n\}$ taken at time points $t_j, j \in \{1, \ldots, n\}$ with n measurements in total. The first term of Eq. (1) represents the absolute, not the quadratic distance between the measurements y_j and the function $g(t_j)$. Microarrays are inflicted with a certain proportion of outliers [9]. If these outliers are weighted quadratically by least-square approaches, as most methods do, a Gaussian distributed error model is assumed. However, a Gaussian error model is not a good choice for the characteristics of frequent outliers, as this approach biases the fit stronger than the absolute distance. The second term of Eq. (1) penalises the absolute number of directional changes in the gene expression dynamics to avoid over-fitting. The penalisation term is weighted by

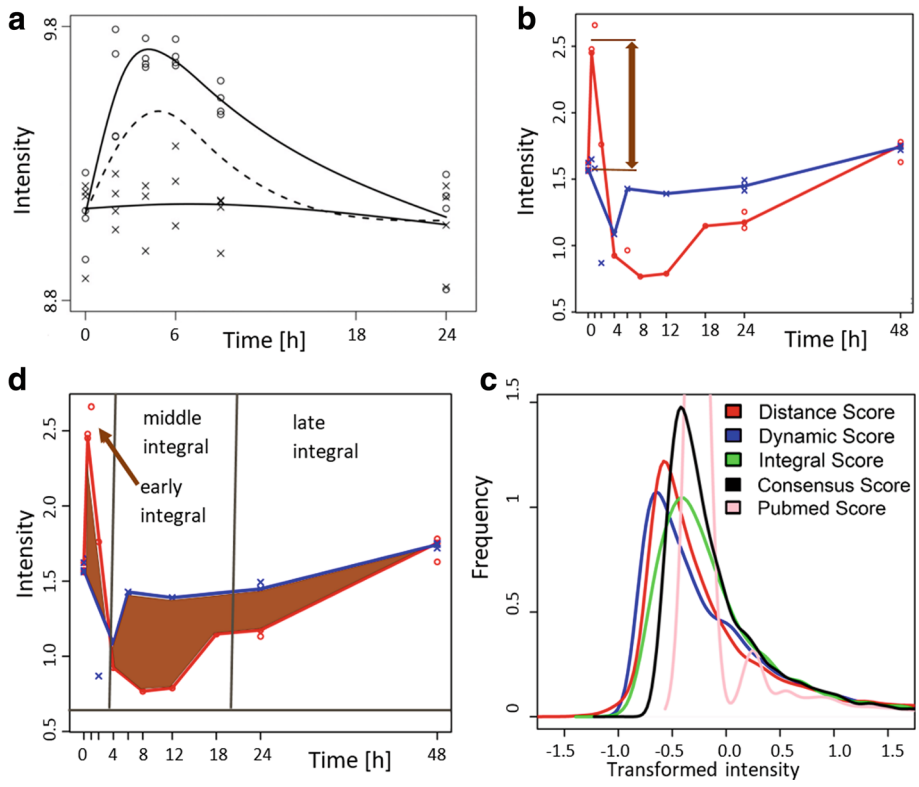

Fig. 2 Score characteristics: **a**) *Dynamics score*. The alternative hypothesis is represented by the *solid line*. The *dashed line* represent the null hypothesis (Picture source [6]). **b**) *Peak score*. Is based on the largest distance (*arrow*) between measurement points for two different stimuli. The *solid line* represents the fit achieved via quantile regression with Eq. (1). **c**) *Integral score*. The area between two dynamics indicates the absolute mRNA production change. This value can be computed for different time intervals. **d**) Different score distributions after z-transformation and the merged *consensus score* distribution

the scaling factor λ. We estimated $\lambda = 0.6$ for SCAN-processed data with the help of real-time PCR profiles from genes that are known to be differentially expressed after the stimulation. The obtained residual-vectors R_i are modified by weighting vectors Ω. These weights account for the uneven experimental design in the following way: First, each time point should have the same weight independent from the number of replicates. Second, more values in one condition than in the other result in higher residuals without a better fit. TTCA balances the uneven design. Third, to reduce the unwanted bias by this vector, the sum of all vector elements of the weighting vector is forced to the same value. The scalar product of the residual and weighting vector yields a scalar value for each gene.

The *dynamics score* D_i is then defined by

$$\mathcal{D}_i := \frac{\mathrm{H}_0}{\mathrm{H}_1} = \frac{< \Omega^{\mathrm{H}_0}, R_i^{\mathrm{H}_0} >}{< \Omega^{\mathrm{stim}}, R_i^{\mathrm{stim}} > + < \Omega^{\mathrm{ctrl}}, R_i^{\mathrm{ctrl}} >}.$$

The relation $\mathrm{H}_0/\mathrm{H}_1$ quantifies how much worse the null-hypothesis fits in comparison to the alternative hypothesis and is easy to interpret.

Peak score

Perturbation experiments may invoke fast and transient peak dynamics in a gene subset, where the peak might be captured by only a small number of measurements. In this case, peaks, although biologically meaningful, may be overlooked by microarray analysis methods. To account for this, we introduce the *peak score*. Let $T = \{t_1, \ldots, t_n\}$ denote the set of the measurement time points. For each time-point $t \in T$, we define $\mathrm{F}_{it}^{\mathrm{stim}}$ and $\mathrm{F}_{it}^{\mathrm{ctrl}}$ as the averages of all replicates for the stimulated and control conditions, respectively. The *peak score* is then given by

$$\mathcal{P}_i := \max_{t \in T} \left| \mathrm{F}_{it}^{\mathrm{stim}} - \mathrm{F}_{it}^{\mathrm{ctrl}} \right|.$$

The success of this approach has been pointed out by Di Camillo et al. [26]. To test whether differences between the expression profiles are significant, we use the robust 0.95 quantile of all available standard deviations, for a minimum of 1000 genes and multiple replicated measurement points as a noise-threshold. A gene i is considered as significant, if $\mathcal{P}_i$ is more than twice the noise-threshold (see Fig. 2b). To account for a possible correlation between the standard deviation and mean of gene expression, TTCA sorts the genes with respect to their mean values and divides them into a minimum of 8 groups, each containing at least 1000 genes. The noise-threshold is then computed separately for each group. TTCA can either use replicated time points to provide a noise threshold or the distribution of the score values to provide a significance threshold. Replicates are not required but can be used. If less than 4 measurement points are replicated, the program will provide only a ranking and the significance will be calculated as in the other scores as described below.

Instability score

Some genes, found highly significant in the previous scores, exhibit an extreme variance between replicates. If the median of the standard deviation of replicated measurements of gene i is two-fold larger than the gene group noise threshold, these genes are classified as unstable. The *instability score* is binary and appears in the results table together with a relative effect size, explained below. TRUE indicates instable genes that are likely false positives, and FALSE indicates genes with acceptable variance between replicates. For an example see the gene SNORA11 in Table 1 and Fig. 4.

Integral score

The *integral score* is intended to quantify the area between the expression profiles for control and treatment. To compute the integral between the two expression dynamics of each gene i we first linearly interpolate the missing values of the quantile regression at measured time points t and at time points where the curves intersect. We then estimate the area between the two dynamics D_i applying the trapezium rule. This integral

$$\mathcal{I}_i := \int_{t_1}^{t_2} \left| \mathrm{D}_i^{\mathrm{stim}}(t) - \mathrm{D}_i^{\mathrm{ctrl}}(t) \right| \mathrm{dt}$$

for each gene i serves as a measure for the difference in the mRNA production between the two conditions. Figure 2 C illustrates the *integral score*, which can be computed for different time intervals. Hereby, four separate scores are computed ($\mathcal{I}_i^{\mathrm{early}}, \mathcal{I}_i^{\mathrm{intermediate}}, \mathcal{I}_i^{\mathrm{late}}, \mathcal{I}_i^{\mathrm{complete}}$) to distinguish between the early response, the intermediate response, the late response, and the response over the whole period. The first three scores are defined for subsequent time-intervals, which can be defined by the user. These scores allow to distinguish between slowly and rapidly responding genes, and might also be used to distinguish a secondary response from the direct response to the stimulus. By using a z-score transformation and averaging of all three integral scores the *combined integral score* $\mathcal{I}_i^{\mathrm{comb}}$ is obtained. The *combined integral score* emphasises the largest changes in gene expression for each period stronger than the more outbalancing *complete integral score* $\mathcal{I}_i^{\mathrm{complete}}$.

Table 1 Compendious result table. The instability of SNORA11 is confirmed and the effect size is high, which indicates a false positive result. The plotted SNORA11 profile in Fig. 4 confirms this suspicion. The effect size of the peak score covers up to 26% of the detection range

Consensus rank	Gene name	Consensus score	Consensus score *p*-value	PubMed	Instability score	Effect size of peak score
1	CTGF	1.00	3.57E-05	73	0.009	0.26
2	EGR1	0.91	7.25E-05	101	0.006	0.23
3	SNORA11	0.62	8.18E-04	0	0.038	0.26
4	PTGS2	0.59	0.001	804	0.009	0.10
5	JUN	0.58	0.001	6789	0.005	0.11
6	GLIPR1	0.57	0.001	0	0.006	0.13
7	FOS	0.55	0.002	920	0.002	0.14
8	AREG	0.53	0.002	549	0.006	0.10
13	MIR4320	0.44	0.005	0	0.016	0.15
15	F3	0.44	0.006	65	0.011	0.10
19	IL8	0.41	0.007	43	0.018	0.13
20	EGR2	0.41	0.008	7	0.005	0.12
21	PCNA	0.40	0.009	583	0.003	0.03
29	DUSP5	0.37	0.013	4	0.012	0.10
36	MYC	0.34	0.017	984	0.002	0.06
37	ROS1	0.34	0.017	84	0.005	0.03
38	HIF1A	0.34	0.017	185	0.007	0.08
42	MIR554	0.34	0.018	0	0.004	0.15
45	IL24	0.32	0.022	0	0.003	0.06
49	TGFB2	0.31	0.025	121	0.008	0.04
51	TGFB1	0.30	0.027	887	0.004	0.03
52	JUNB	0.30	0.028	54	0.008	0.05

Relevance score

By using the R package RISmed [34] we query the PubMed database of publications for records that match both the gene name and the condition. For each gene i this yields a number of publications p_i. We use a log-transformation to normalise p_i between 0 and 1, and obtain the *relevance score*

$$\mathcal{R}_i := \log_{p_{\max}}(p_i),$$

where $p_{\max} := \max_i(p_i)$. This score indicates whether a gene is already well known to be associated with the condition or potentially a new target.

Consensus score

The *consensus score* is used for the final ranking of the genes and combines the four scores. By merging the *dynamics score* with the *peak score*, *combined integral score* and *relevance score*, and normalising the result to be between 0 and 1, we obtain

$$\mathcal{C}_i := \frac{\breve{\mathcal{D}}_i + \breve{\mathcal{P}}_i + \breve{\mathcal{I}}_i^{\text{comb}} + \breve{\mathcal{R}}_i}{4},$$

whereby score $\mathcal{S}$ is z-transformed $\breve{\mathcal{S}}$ before the average is computed. Figure 2d shows the z-transformed distributions of the score values. To better centre the relevance score distribution, only non-zero values are considered for the z-transformation.

Significance

Except for the *peak score* we did not define any significance threshold, yet. For the other scores a significance level can be computed by a one-sided, one-group hypothesis test. The program fits the Cauchy, Gamma, log-normal, logistic, normal, Poisson and Weibull distribution to the empirical distribution of score values using the function fitdistr() provided by the R package MASS [35]. The log-normal distribution is only defined for strictly positive values, however, by shifting the $x-$axis it can be fitted in the negative part as well. The obtained significance threshold is transformed back afterwards. The distribution function providing the best fit of the distribution of score values is automatically selected and plotted. To estimate the significance for a differentially expressed gene we provide the p-value as well as the effect size [36]. The effect size of the *peak score* is defined as the

distance between the expression dynamics, normalised by the maximum distance possible, i.e. the highest expression value within the data set minus the lowest expression value within the data set. The largest observed expression change in our data set covers 25.9% of the whole detection range and represents the effect size. The same normalisation is used for the *instability score* and also for the *integral score*, where the maximum area is given as the maximal distance multiplied by the time period. In the *consensus score*, a gene is considered to be significant, if it is considered significant in at least two scores.

Method extension for gene set analysis

To investigate the behaviour of functional groups, the genes are linked to gene ontology groups using the BiomaRt-package [37, 38]. Then the expression level at the initial time point is subtracted from the gene expression profile of each gene. Thus, all profiles are initially zero and only the expression change with respect to the first value is observed (Fig. 3a). Second, the average expression together with the upper sd_{u} and lower sd_{l} standard deviation of all genes within each ontology group are calculated for each time point. The upper standard deviation hereby accounts for all measurement points above the group mean and the lower standard deviation accounts for all measurement points below the average. Separation into upper and lower standard deviation helps to better recognise when the subset of the functional group shows increased (or decreased) expression. This would lead to enlarged upper (or lower) standard deviations, where the classical standard deviation does not allow such distinction. We then consider gene groups differentially expressed if their expression bandwidths are separated by the condition, i.e., that the variability between genes in the same ontology group are small in contrast to changes caused by different treatments. To test, whether the expression bandwidths are separated by condition, we distinguish two different cases, as shown in Fig. 3b. On the one hand, the band of the control can be higher than the band of the stimulus (case A), on the other hand, the situation can be reversed (case B). We search for the minimum overlap

$$\theta_{ij} = \frac{\max\left[\overbrace{(\mathrm{mean(C)} - \mathrm{sd_l(C)}) - (\mathrm{mean(S)} + \mathrm{sd_u(S)})}^{\text{Case A}}; \overbrace{(\mathrm{mean(S)} - \mathrm{sd_l(S)}) - (\mathrm{mean(C)} + \mathrm{sd_u(C)})}^{\text{Case B}}\right]}{\underbrace{\frac{1}{2}\left(\mathrm{sd_u(S)} + \mathrm{sd_l}(S) + \mathrm{sd_u(C)} + \mathrm{sd_l(C)}\right)}_{\text{Bandwith}}}$$

of both bandwidths for a combination of time points $j \in \{1, \ldots, n\}$ and genes i, where n indicates the total number of measurements per gene. We are only interested in the maximum distance between the bands or the minimum mutual overlap for the score $\Theta_i = \max\left(\theta_{i1}, \ldots, \theta_{ij}, \ldots, \theta_{in}\right)$ at each time point. Positive values indicate a separation of the bands and negative values indicate overlap. The average expression profile for each gene group and treatment is then used to calculate the other scores as described above. Hence, TTCA ranks functional groups high if they contain genes with similar expression pattern over time within a condition and if they clearly change the expression dynamics from one condition to the other. Although we did not compare the performance of the gene set module, the application on real data seems promising. Alternatively, the user can use the ranking of the individual genes to apply other methods for gene set analysis.

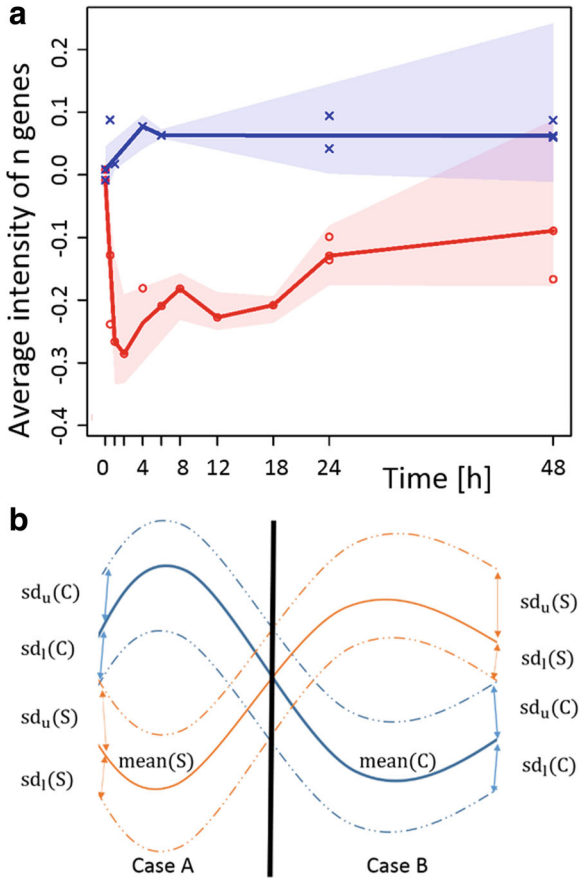

Fig. 3 Gene set analysis. **a**) Example analysis result with the average of *n* genes. **b**) Scheme for minimal overlap calculation. The *continuous lines* represent the average expression of the gene group at one time point for either the stimulated sample (S) or the control (C). The *dotted lines* represent the average with *upper* (Sd_u) or *lower* (Sd_l) standard deviation

Computation time and further packages

TTCA is computationally fast using about 1 h for one contrast. This includes the analysis of expression dynamics

and the generation of relevant figures on a standard laptop (i5 1.70 GHz, memory 12 GB). Furthermore, TTCA uses the R-package tcltk2 [39] for a progress bar and the R-package VennDiagram [40] to show automatically the overlap of significant genes across scores.

Methods for lung cancer data set

Cell seeding, growth factor stimulation and microarray processing

The cell line H1975 (NCI-H1975; ATCC: CRL-5908) were obtained from LGC Standards (Teddington, UK). The cell line was authenticated by STR-analysis (DSMZ, Braunschweig, Germany) and routinely checked for mycoplasma contamination. H1975 NSCLC cells were seeded in 6-well-plates with $1.33 \cdot 10^5$ cells per well. After incubation for 3 days, cells were washed 3 times and supplemented with DMEM without FCS for overnight starvation. On the following day, cells were stimulated with 50 ng/mL of EGF diluted in starvation medium. Samples were harvested after 0, 0.5, 1, 2, 4, 6, 8, 12, 24 and 48 hours. Subsequently RNA was extracted, as described below. Total cellular RNA was isolated with the NucleoSpin RNA II kit according to the manufacturers' instructions. RNA concentrations were determined by measuring the absorbance (230 - 400 nm) using a NanoDrop®ND-1000 spectrometer. The purity of the RNA was determined through the ratio of the absorbance at 260nm and 280nm. RNA with a ratio $\geq$ 1.8 was used for further analysis. After assessing RNA integrity using the Agilent Bioanalyzer, 100 ng in 3 μl per sample were handed over. After amplification, labelling with biotin and fragmentation of the RNA, hybridisation with GeneChip Human Gene 2.0 ST Array was performed for 16 h at 45 °C. Subsequently, washing and staining was performed using an Affymetrix Fluidics Station 450 and the microarray was scanned using an Affymetrix GeneArray Scanner 3000.

Microarray preprocessing

The method Single Channel Array Normalisation (SCAN) [31] was used for the preprocessing. For the mapping of probes to genes we used the Netaffix.v.34 annotation file which is available from the array manufacturer. For the transcript-level we used Brainarray-Ensembl-T-v.18.0.0 [41] for annotation. The quality was additionally assessed before and after preprocessing with the R-package ArrayQualityMetrics [42]. Four possible outliers were visible in the 3D-PCA-plot generated with pcaMethods [43]. They were investigated in contrast to other replicates or to the closest measurement points with Limma [44] and Piano [45] under use of BioMart [46] for GO-mapping. We assumed a problem with the magnesium concentration and excluded the affected arrays from the analysis.

Results and discussion

The approach presented here allows the identification of biologically relevant genes from noisy, sparse, and possibly incomplete time course gene expression data sets from perturbation experiments. In the case presented in our study, the administration of the potent mitogen EGF led to the identification of numerous known EGF/EGFR induced target genes as indicated by the *relevance score*, such as CTGF (Fig. 4), EGR1, PTGS2/COX2, and transcription factors of the AP1 family including JUN and FOS (Table 1).

The top-ranked genes represent key factors involved in the initiation and maintenance of a mitogenic response in tumour cells. Interestingly, many of the immediate EGF-dependent targets listed in Table 1 represented transcriptional regulators, for instance EGR1, EGR2, JUN, FOS, or MYC, and secreted chemokines like CTGF, IL8 (Fig. 4), or KITLG/SCF, illustrating that EGF is a central inducer of pro-proliferative gene expression and paracrine regulation in lung cancer. These results are confirmed by previous publications describing for example, that activation of the PI3K/AKT pathway, which typically stimulates the transcription factor AP1 consisting of JUN/FOS heterodimers, can stimulate IL8 production and secretion in NSCLC cells [47].

However, our approach not only confirmed findings from other studies. Even more important, we identified a long list of previously un-published downstream effectors (Additional file 3: Table S4; 18/79 (23%) significantly regulated genes have not been described in the context of EGF/EGFR signalling). For example, the target gene IL24 (Relevance Score: 0.32) has been shown to inhibit NSCLC cell migration suggesting that EGF-induced IL24 might shift tumour cells from a migratory to a mitotic phenotype [48] (Fig. 4). The high ranked gene GLIPR1 (Fig. 4) has recently been identified as tumour suppressor in lung cancer [49], however, the relationship between GLIPR1 and EGF was yet unknown. In addition, the significant regulation of the micro-RNAs miR-4320 (Relevance Score: 0.44; Fig. 4) and miR554 (Relevance Score: 0.34) suggests that EGF supports the oncogenic properties of NSCLC cells via miRNA-dependent mechanisms [50].

We compared TTCA with Limma, EDGE and MaSigPro (see Additional file 2). We assume, that the number of PubMed publications, linking EGF stimulation with individual genes, can be used to generate a ranking of expected target genes. Additional file 2: Table S2 shows the ranking of the top 100 expected genes, determined by TTCA, Limma, EDGE and MaSigPro. Additional file 3: Table S3 shows the top 100 gene names displayed by each method investigated. Additional file 2: Figures S2-S8 show the top ten expression profiles of each method investigated and a p-value distribution provided by EDGE. The code for the method comparison is in Additional file 2.

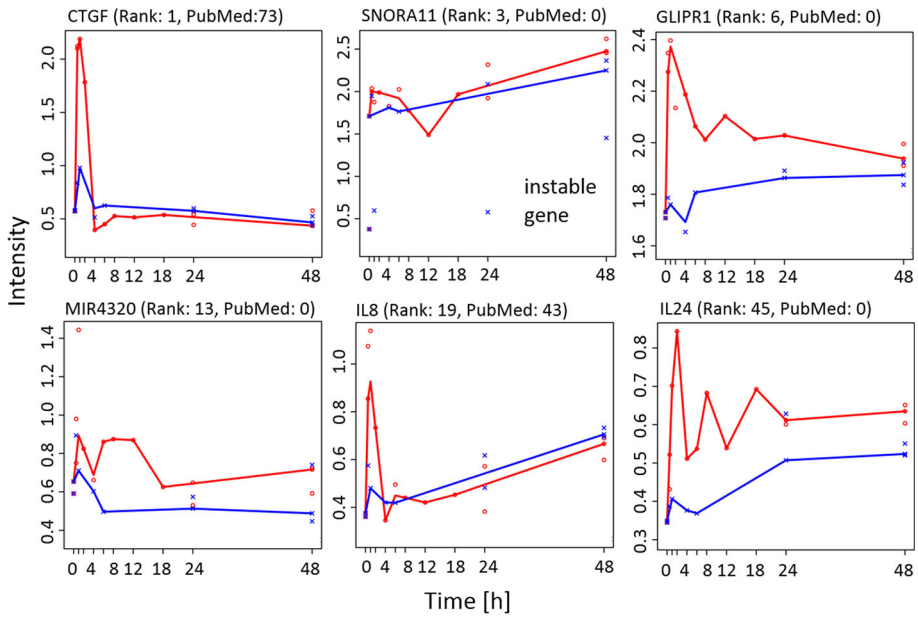

Fig. 4 Time course profiles of genes considered significant. *Red*: With EGF stimulation. *Blue*: Control. *Line*: Quantile regression. *Points*: Measurements. SNORA11 is ranked highly significant, but the instability score is high and identifies this finding as false positive

The source code of the TTCA method is in Additional file 1.

Conclusion

We have presented a new method for microarray time-series data analysis, specifically intended for difficult experimental designs with sparse measurements. Even when the experimental design involves a uniform data collection, experimental problems can lead to the exclusion of individual arrays, and thus to the loss of measurement points after quality control. Sufficient replicates are important for proper microarray data analysis [51] and remain important in more accurate next-generation sequencing [52]. However, even if such data are difficult, they nonetheless contain helpful hints for further investigations. TTCA is able to detect different characteristics of the changes in expression dynamics and always provides not only *p*-values but also effect sizes for an optimal significance interpretation [36].

Our method can also be applied for data sets with less complicated designs (regular sampling intervals, large number of replicates) and yield very good results, comparable with other tools. It should be noted, however, that the scores included in TTCA detect specifically expression patterns arising after perturbation or stimulation experiments. For detecting specific dynamical behaviours, e.g. oscillations, we recommend specialised methods like Lomb-Scargle periodograms [53], JTK-CYCLE [54] or GeneCycle [55].

We believe that the developed TTCA package is a valuable and efficient tool for the dissection of important information that is usually concealed by experimental and biological variations leading to data heterogeneity. The connection with the number of PubMed publications has to our knowledge never been included in other packages and supports the user in distinguishing between new and already known genes affected by the applied perturbation. Further new features (at least to our knowledge) are the automatic detection of the best density function, the approach to detect false positives (the instability score), or the distinction between early, middle and late response. Also, the outbalancing of the sampling design using weighting factors is an important new feature. Moreover, we provide a new gene set significance approach, which pools genes into gene ontology groups which expression bandwidths are separated (minimal overlap score). TTCA provides automatically quality checks and plots the gene expression profiles. Thus, the user can easily judge the performance of the package for any included data set. Strong advantages of TTCA are the high degree of transparency, the multitude of visual output for quality assessment, search flexibility and sensitivity also in cases where other methods cannot be applied.

Additional files

Additional file 1: R code of TTCA (EUPL). (TXT 816 kb)

Additional file 2: Method comparison. A table summarises the methods mentioned in the introduction, method shortcomings are further discussed and TTCA is compared with some applicable methods. Includes **Table S1**-**S3** and **Figures S1**-**S8**. (PDF 2165 kb)

Additional file 3: The complete result table is given in **Table S4**. (XLSX 816 kb)

Abbreviations

AKT: AKT serine/threonine kinase 1; ANOVA: Analysis of variance; AP1: Activator protein 1; AREG: Amphiregulin; BATS: Bayesian analysis of time series; BETR: Bayesian estimation of temporal regulation; COX2: Cytochrome C oxidase subunit II (see new name: PTGS2); CRAN: Comprehensive R archive network; CTGF: Connective tissue growth factor; DMEM: Dulbecco's modified eagle medium; DUSP5: Dual specificity phosphatase 5; EDGE: Extraction of differential gene expression; EGF: Epidermal growth factor; EGFR: Epidermal growth factor receptor; EGR1: Early growth response 1; EGR2: Early growth response 2; EUPL: European union public licence; F3: Coagulation factor III (thromboplastin, tissue factor); FCS: Fetal calf serum; FOS: FBJ murine osteosarcoma viral oncogene homolog; FPCA: Functional principle component analysis; FWER: Family wise error rate; GLIPR1: Glioma pathogenesis-related protein 1; GO: Gene ontology; HIF1A: Hypoxia-inducible factor 1-alpha; IL8: Interleukin 8; IL24: Interleukin 24; JUN: Jun proto-oncogene; JUNB: Transcription factor jun-B; KIT: V-Kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog; KITLG: KIT ligand; Limma: Linear models for microarray data; MaSigPro: Microarray significant profiles; MIR4320: MicroRNA 4320; MIR554: MicroRNA 554; MYC: V-myc avian myelocytomatosis viral oncogene homolog; NSCLC: Non-small cell lung cancer; PACE: Principle component analysis through conditional expectation; PCA: Principle component analysis; PCNA: Proliferating cell nuclear antigen; PCR: Polymerase chain reaction; PI3K: Phosphatidylinositol 3-kinase; Piano: Platform for integrative analysis of omics data; PTGS2: Prostaglandin-endoperoxide synthase 1; RMA: Robust multi-array average; RNA: Ribonucleic acid; ROS1: Proto-oncogene tyrosine-protein kinase ROS; SAM: Significance analysis of microarrays; SCAN: Single-channel array normalisation; SCF: Stem cell factor (see new name KITLG); SNORA11: Small nucleolar RNA, H/ACA box 11; TGFB1: Transforming growth factor beta 1; TGFB2: Transforming growth factor beta 2; TTCA: Transcript time course analysis

Acknowledgements

MA thanks Eric Koncina (Neuro Inflammation group, University of Luxembourg) for his support for translating the code to a user-friendly R-package at CRAN. MA thanks Sébastien de Landtsheer (Systems Biology group, University of Luxembourg) for proofreading the manuscript.

Funding

MA acknowledges currently the Horizon 2020 MSCA grant agreement, No 642295, www.melplex.eu. MA, DS and FM were supported by a grant from the Center for Modelling and Simulation in the Biosciences (BIOMS) of the Heidelberg University. KB was supported by a grant from the BMBF (LungSysII, FKZ 0316042B). RM and UK were supported by the German Center for Lung Research (DZL, 82DZL00404). The funding body was not involved in the design of the study and collection, analysis, and interpretation of data or in writing the manuscript.

Availability of data and materials

The program is freely distributed under European Union Public Licence (EUPL) and can directly be installed from CRAN [cran.rstudio.com/web/packages/TTCA], the official R package archive. The source code is provided in Additional file 1 and the current version is available upon request. Microarray data sets GSE84094 and GSE84095 have been uploaded to Gene Expression Omnibus (GEO) database at NCBI [ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE84094; ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE84095].

Authors' contributions

MA created the method concept, wrote the code and performed the statistical analyses. BM, RM, KB and UK performed the experimental design, conducted the experiments and KB interpreted the analysis results. NG and CS are responsible for microarray handling and the public availability of the data set. MA, DS, KB and FM wrote the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Author details

[1]Complex Biological Systems Group (BIOMS/IWR), Heidelberg, Im Neuenheimer Feld 294, 69120 Heidelberg, Germany. [2]Systems Biology Group, Université du Luxembourg, 7, avenue du Swing, L-4367 Belvaux, Luxembourg. [3]CCU Neuropathology Group, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 221, 69120 Heidelberg, Germany. [4]Institute of Pathology, Heidelberg University Hospital, Im Neuenheimer Feld 672, 69120 Heidelberg, Germany. [5]Systems Biology of Signal Transduction Group, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, 69120 Heidelberg, Germany. [6]Translational Lung Research Center (TLRC), Member of the German Center for Lung Research (DZL), Im Neuenheimer Feld 430, 69120 Heidelberg, Germany. [7]Medical Research Center, Medical Faculty Mannheim, University of Heidelberg, Theodor-Kutzer-Ufer 1-3, 68167 Mannheim, Germany. [8]Frankfurt Institute for Advanced Studies (FIAS), Goethe University Frankfurt, Ruth-Moufang-Straße 1, 60438 Frankfurt am Main, Germany.

Received: 8 July 2016 Accepted: 21 December 2016
Published online: 14 January 2017

References

1. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci. 2001;98(9):5116–21.
2. Kerr MK, Martin M, Churchill GA. Analysis of variance for gene expression microarray data. J Comput Biol. 2000;7(6):819–37.
3. S GK. Limma: linear models for microarray data In: Gentleman R, Carey V, Dudoit S, Irizarry R, Huber W, editors. Bioinformatics and Computational Biology Solutions Using R And Bioconductor. New York: Springer; 2005. p. 397–420.
4. Mutarelli M, Cicatiello L, Ferraro L, Grober OMV, Ravo M, Facchiano AM, Angelini C, Weisz A. Time-course analysis of genome-wide gene expression data from hormone-responsive human breast cancer cells. BMC Bioinforma. 2008;9(Suppl 2):12.
5. Leek JT, Monsen E, Dabney AR, Storey JD. EDGE: extraction and analysis of differential gene expression. Bioinformatics. 2006;22(4):507–8.
6. Storey JD, Xiao W, Leek JT, Tompkins RG, Davis RW. Significance analysis of time course microarray experiments. Proc Natl Acad Sci USA. 2005;102(36):12837–42.
7. Conesa A, Nueda MJ, Ferrer A, Talón M. maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments. Bioinformatics. 2006;22(9):1096–102.
8. Sohn I, Owzar K, George SL, Kim S, Jung SH. A permutation-based multiple testing method for time-course microarray experiments. BMC Bioinforma. 2009;10(1):336.
9. Sohn I, Owzar K, George SL, Kim S, Jung SH. A permutation-based multiple testing method for time-course microarray experiments. BMC Bioinforma. 2009;10(1):336.
10. Tai YC, Speed TP, et al. A multivariate empirical Bayes statistic for replicated microarray time course data. Ann Stat. 2006;34(5):2387–412.
11. Aryee MJ, Gutiérrez-Pabello JA, Kramnik I, Maiti T, Quackenbush J. An improved empirical bayes approach to estimating differential gene expression in microarray time-course data: BETR (Bayesian Estimation of Temporal Regulation). BMC Bioinforma. 2009;10(1):409.
12. Cheng C, Ma X, Yan X, Sun F, Li LM. MARD: a new method to detect differential gene expression in treatment-control time courses. Bioinformatics. 2006;22(21):2650–7.
13. Huang W, Cao X, Zhong S. Network-based comparison of temporal gene expression patterns. Bioinformatics. 2010;26(23):2944–51.
14. Moran VA, Perera RJ, Khalil AM. Emerging functional and mechanistic paradigms of mammalian long non-coding rnas. Nucleic Acids Res. 2012;40(14):6391–400.
15. Stegle O, Denby KJ, Cooke EJ, Wild DL, Ghahramani Z, Borgwardt KM. A robust Bayesian two-sample test for detecting intervals of differential

gene expression in microarray time series. J Comput Biol. 2010;17(3):355–67.
16. Kalaitzis AA, Lawrence ND. A simple approach to ranking differentially expressed gene expression time courses through gaussian process regression. BMC Bioinforma. 2011;12(1):180.
17. Angelini C, De Canditiis D, Mutarelli M, Pensky M. A Bayesian approach to estimation and testing in time-course microarray experiments. Stat Appl Genet Mol Biol. 2007;6(1).
18. Angelini C, Cutillo L, De Canditiis D, Mutarelli M, Pensky M. BATS: a Bayesian user-friendly software for analyzing time series microarray experiments. BMC Bioinforma. 2008;9(1):415.
19. Jonnalagadda S, Srinivasan R. Principal components analysis based methodology to identify differentially expressed genes in time-course microarray data. BMC Bioinforma. 2008;9(1):267.
20. Ramsay JO. Functional Data Analysis. Hoboken: John Wiley & Sons, Inc; 2006.
21. Coffey N, Hinde J. Analyzing time-course microarray data using functional data analysis-a review. Stat Appl Genet Mol Biol. 2011;10(1):1–32.
22. Liu X, Yang MCK. Identifying temporally differentially expressed genes through functional principal components analysis. Biostatistics. 2009;10(4):667–79.
23. Wu S, Wu H. More powerful significant testing for time course gene expression data using functional principal component analysis approaches. BMC Bioinforma. 2013;14(1):6.
24. Yao F, Müller HG, Wang JL. Functional data analysis for sparse longitudinal data. J Am Stat Assoc. 2005;100(470):577–90.
25. Henn AD, et al. High-resolution temporal response patterns to influenza vaccine reveal a distinct human plasma cell gene signature. Sci Rep. 2013;3(2327). doi:10.1038/srep02327.
26. Di Camillo B, Toffolo G, Nair SK, Greenlund LJ, Cobelli C. Significance analysis of microarray transcript levels in time series experiments. BMC Bioinforma. 2007;8(Suppl 1):10.
27. Minas C, Waddell SJ, Montana G. Distance-based differential analysis of gene curves. Bioinformatics. 2011;27(22):3135–41.
28. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet. 2010;11(10):733–9.
29. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003;4(2):249–64.
30. Stevens JB, Horne SD, Abdallah BY, Christine JY, Heng HH. Chromosomal instability and transcriptome dynamics in cancer. Cancer Metastasis Rev. 2013;32(3-4):391–402.
31. Piccolo SR, Sun Y, Campbell JD, Lenburg ME, Bild AH, Johnson WE. A single-sample microarray normalization method to facilitate personalized-medicine workflows. Genomics. 2012;100(6):337–44.
32. Koenker R, Vol. 38. Quantile Regression. New York: Cambridge University Press; 2005.
33. Koenker R, Portnoy S, Ng PT, Zeileis A, Grosjean P, Ripley BD. Quantreg: Quantile Regression. 2013. R package version 5.05. https://cran.r-project.org/web/packages/quantreg/index.html.
34. Kovalchik S. RISmed: Download Content from NCBI Databases. 2015. R package version 2.1.5. https://CRAN.R-project.org/package=RISmed.
35. Venables WN, Ripley BD. Modern Applied Statistics with S, 4th edn. New York: Springer; 2002. ISBN 0-387-95457-0. http://www.stats.ox.ac.uk/pub/MASS4.
36. Nuzzo R. Statistical errors: P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume. Nature. 2014;506.7487:150–153.
37. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, Huber W. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics. 2005;21(16):3439–40.
38. Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, Kasprzyk A. BioMart–biological queries made easy. BMC Genomics. 2009;10(1):22.
39. Grosjean P. SciViews-R: A GUI API for R. MONS, Belgium: UMONS; 2014. UMONS. https://cran.r-project.org/web/packages/tcltk2/index.html.
40. Chen H, Boutros PC. VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. BMC Bioinforma. 2011;12(1):35.
41. Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, Bunney WE, Myers RM, Speed TP, Akil H, et al. Evolving gene/transcript definitions significantly alter the interpretation of genechip data. Nucleic Acids Res. 2005;33(20):175–5.
42. Kauffmann A, Gentleman R, Huber W. arrayqualitymetrics-a bioconductor package for quality assessment of microarray data. Bioinformatics. 2009;25(3):415–6.
43. Stacklies W, Redestig H, Scholz M, Walther D, Selbig J. pcaMethods - a bioconductor package providing PCA methods for incomplete data. Bioinformatics. 2007;23(9):1164–7.
44. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol. 2004;3(1):1–25.
45. Väremo L, Nielsen J, Nookaew I. Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. Nucleic Acids Res. 2013;111.
46. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart. Nat Protoc. 2009;4(8):1184–91.
47. Zhang Y, Wang L, Zhang M, Jin M, Bai C, Wang X. Potential mechanism of interleukin-8 production from lung cancer cells: An involvement of egf–egfr–pi3k–akt–erk pathway. J Cell Physiol. 2012;227(1):35–43.
48. Panneerselvam J, Jin J, Shanker M, Lauderdale J, Bates J, Wang Q, Zhao YD, Archibald SJ, Hubin TJ, Ramesh R. Il-24 inhibits lung cancer cell migration and invasion by disrupting the sdf-1/cxcr4 signaling axis. PloS one. 2015;10(3):0122439.
49. Sheng X, Bowen N, Wang Z. GLI pathogenesis-related 1 functions as a tumor-suppressor in lung cancer. Mol Cancer. 2016;15(1):1.
50. Singh DK, Bose S, Kumar S. Role of microrna in regulating cell signaling pathways, cell cycle, and apoptosis in non-small cell lung cancer. Curr Mol Med. 2016;16(5):474–486.
51. Nguyen TT, Almon RR, DuBois DC, Jusko WJ, Androulakis IP. Importance of replication in analyzing time-series gene expression data: corticosteroid dynamics and circadian patterns in rat liver. BMC Bioinforma. 2010;11(1):279.
52. Hansen KD, Wu Z, Irizarry RA, Leek JT. Sequencing technology does not eliminate biological variability. Nat Biotechnol. 2011;29(7):572–3.
53. Glynn EF, Chen J, Mushegian AR. Detecting periodic patterns in unevenly spaced gene expression time series using lomb–scargle periodograms. Bioinformatics. 2006;22(3):310–6.
54. Hughes ME, Hogenesch JB, Kornacker K. Jtk_cycle: an efficient nonparametric algorithm for detecting rhythmic components in genome-scale data sets. J Biol Rhythm. 2010;25(5):372–80.
55. Ahdesmäki M, Lähdesmäki H, Gracey A, Yli-Harja O, et al. Robust regression for periodicity detection in non-uniformly sampled time-course gene expression data. BMC Bioinforma. 2007;8(1):233.