Journal of
Data Mining in Genomics & Proteomics

Müller and Pfaffl, J Data Mining Genomics Proteomics 2012, 3:2
DOI: 10.4172/2153-0602.1000117

Research Article Open Access

RNA Interference Off-target Screening using Principal Component Analysis

Jakob Müller* and Michael W. Pfaffl

Physiology Weihenstephan, Technische Universität München, Research Center for Nutrition and Food Science, Weihenstephaner Berg 3, 85350 Freising, Germany

Abstract

Off-target effects remain the major problem in any RNAi-knockdown application. Casting cell culture loss-of-function studies evaluated by heat map and principal component analysis (PCA) we realized that the PCA derived plots can clearly visualize off-target effects. Due to the inexistence of off-target effects in our cell culture model we created an *in silico* data model in order to demonstrate how PCA can be utilized therefore. With the presented *in silico* modulation it is possible to simulate the impact of various treatments on changing gene expression. Known effects caused by drug treatment or by inserted knockdowns could be clearly separated from unknown off-target effects. By creating various randomized gene expression data sets we demonstrate that PCA can assign more effective an off-target effect compared to a heat map gene regulation pattern.

Keywords: RNAi Off-target Screening; Principal Component Analysis; Heatmap; *In silico* Data Modelling; Virtual RT-qPCR

Abbreviations: Cq: Quantification Cycle; CV: Coefficient of Variation; EGCG: Epigallocatechin Gallate; HCA: Hierarchical Cluster Analysis; PCA: Principal Component Analysis; RNAi: RNA Interference; RT-qPCR: Reverse Transcription Quantitative Real-time Polymerase Chain Reaction; siRNA: Small Interfering RNA

Background

RNA interference (RNAi) is nowadays a common used technology for gene silencing. It is applied for therapeutic aims or identification of drug targets as well as for basic research on gene function. The still unsolved major problem with this technique is the potential to cause unspecific off-target gene regulations in the treated cell culture or organism. These off-target effects are well described in many studies [1-3]. Since its discovery [4] and utilisation [5,6] RNAi has become an easy tool to use in cell culture assays. Compared to knockout based loss-of-function studies RNAi-knockdowns can be applied easily as a combined application beside a drug treatment. The reaction of the cell culture model to the treatments can subsequently be assessed on RNA level. Conventionally expression proofing studies are performed by quantitative real-time RT-qPCR or hybridization arrays. To visualize the differential gene expression changes hierarchical cluster analysis (HCA) or heat maps are very common [7]. Effects of a single treatment, like a siRNA-sequence, on a very broad range of genes can be visualized that way. An alternative method of visualizing complex datasets is the principal component analysis (PCA). This statistical and visualization tool reduces the dimensionality of a dataset consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the dataset [8].

Using an adenoviral based RNAi knockdown-model, we casted a loss-of-function study (data not shown) identifying the influence of a plant secondary metabolite (EGCG) under various gene knockdowns in an immunological signalling pathway. The outcome of the experiments was among others analyzed by PCA. Doing this we observed by chance a synergetic side effect exclusively in those of our treatment groups in which one particular knockdown was combined with the drug appliance. These findings led to the idea that if we can separate the roots of an effect applying PCA we can use it vice versa and isolate RNAi originating gene regulations from drug effects and this means off-target screening.

For an approval of our assumption we tried to find an off-target gene regulation in our cell culture model caused by one of the viral induced siRNA-knockdowns we had in stock. But without access to genome wide transcriptional array results no off-target effect was found. Hence we decided to use *in silico* data modelling to demonstrate how PCA can be utilized beneficially for RNAi off-target screening.

Methods

Data modelling

The underlying dataset was justified to the gene expression data output of our standard cell culture assays. Thereby the extracted RNA initially undergoes a RT (Reverse Transcription) with subsequent qPCR (quantitative real-time polymerase chain reaction) gene expression analysis. Accordingly the layout of the data model was designed. It consists of eight treatment groups (a-h) each containing four replicates, leading to 32 samples (Table 1). The eight groups were divided into two sections, the drug-treated groups (b,d,f,h) and the media control treated groups (a,c,e,g). The sections in turn were sub-classified in four species each one characterized by a particular gene knockdown-combination. The four knockdown-combinations were

Knockdown-combination	kd-c	kd-I	kd-II	kd-I&II
Control (cell culture media)	a	c	e	g
Treatment (drug appliance)	b	d	f	h

Each treatment-variant (control, treatment) is combined with each knockdown-variant (kd-c, kd-I, kd-II, kd-I&II) leading to eight treatment groups (a - h), each n=4, leading to 32 samples

Table 1: Knockdown-assay layout.

***Corresponding author:** Jakob Müller, Physiology Weihenstephan, Technische Universität München, Research Center for Nutrition and Food Science, Weihenstephaner Berg 3, 85350 Freising, Germany, Tel: +49 8161 71 3867; Fax: +49 8161 71 4204; E-mail: jakob.mueller@wzw.tum.de

Received March 07, 2012; **Accepted** June 13, 2012; **Published** June 17, 2012

Citation: Müller J, Pfaffl MW (2012) RNA Interference Off-target Screening using Principal Component Analysis. J Data Mining Genomics Proteomics 3:116. doi:10.4172/2153-0602.1000116

composed from two target knockdowns (named "kd-I" or "kd-II") and one knockdown-control (named "kd-c"). The two treatment groups and the four knockdown-groups were combined pair wise (Table 1).

For each of the 32 samples a dataset of 21 genes was rendered using our self designed spreadsheet template (Microsoft Office Excel 2007, Additional file 1). The 21 genes where group-wise attributed to a supposed class of gene regulation. The classes differ in how genes respond, either to drug treatment or knockdown-appliance. Four of these classes where created: (class I) stabile expressed reference genes, (class II) target-genes knocked down by RNAi, (class III) genes regulated (up or down) by the drug treatment, (class IV) genes influenced by the drug treatment and additionally bearing a RNAi-off-target effect (Figures 1 and 2). The *in silico* expression data based on C_q-

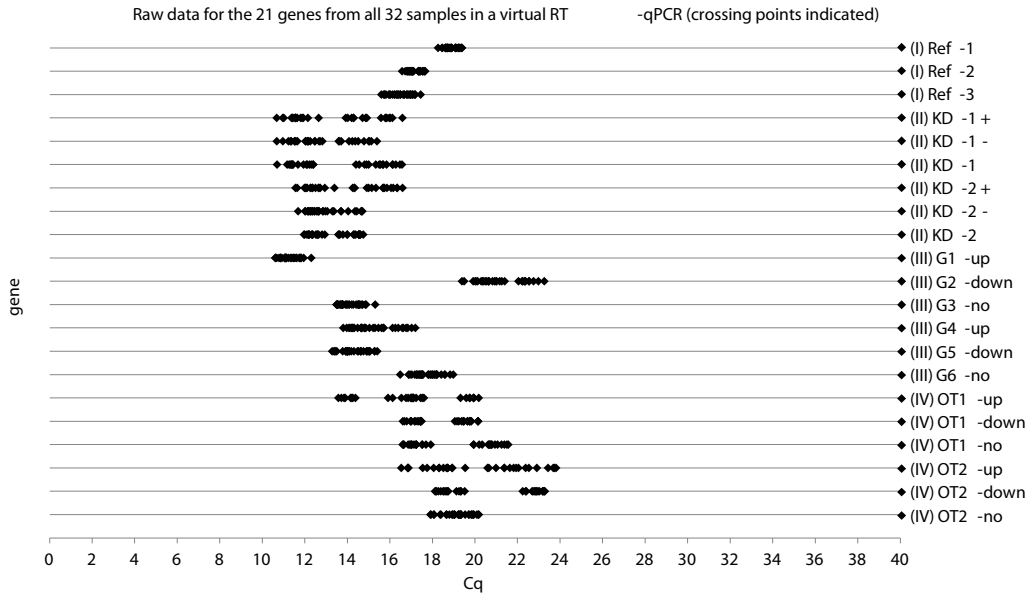

Figure 1: C_q-values plotted as virtual RT-qPCR. The one random output of our simulation which underlies all graphs in this work represents a raw dataset from a RT-qPCR gene expression analysis of 21 genes. The 32 C_q-values (originating from the eight different application combinations multiplied by the four replicates) are plotted gene wise on a line, like fluorescence-signals crossing a threshold in an *in vitro* RT-qPCR reaction. The random data distribution and within the emulated CV-value gets visible best in class I genes (unregulated reference-genes). Particular regulation effects become apparent by the separate clusters among the 32 data points of a gene.

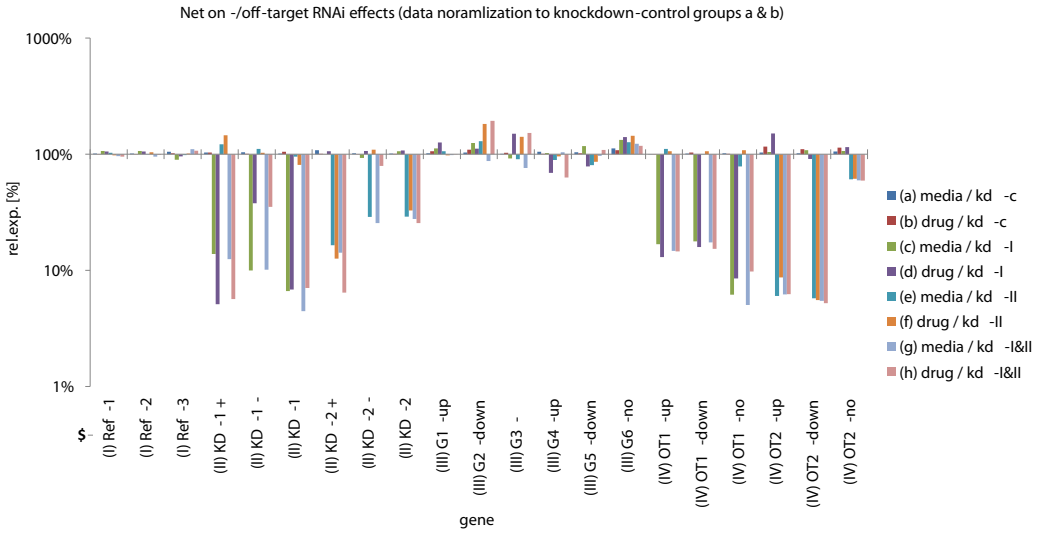

Figure 2: Knockdown-evaluation from the *in silico* data set. Relative expression levels of simulated genes after knockdown and/or drug appliance calculated by the $\Delta\Delta C_q$-method plotted against a logarithmic ordinate. Net knockdown-effects are exposed by normalizing data to knockdown control samples. The 21 genes are arranged in four classes: (class I) reference genes (Ref-1, Ref-2, Ref-3); (class II) knockdown-targets (KD-1+, KD-1-, KD-1, KD-2+, KD-2-, KD-2); (class III) drug targets (G1-up, G2-down, G3-no, G4-up, G5-down, G6-no); (class IV) drug targets with an additional off-target effect (OT1-up, OT1-down, OT1-no, OT2-up, OT2-down, OT2-no). The additional prefix specification "+" or "-" indicated at class II genes stands for an additional synergetic gene regulation effect which occurs when knockdown appliance is combined with drug appliance. The additional prefix specification "-up", "-down" or "-no" indicated at class III and IV genes stands for the direction of the predestinated gene regulation. For every gene eight expression levels are given corresponding to the eight treatment groups: a - h (Table 1).

values of every single sample for every individual gene was created by our random algorithm (Microsoft Office Excel 2007, Additional file 1). The C_q-value represents the number of cycles that are required by the fluorescence signal of a single sample in a qPCR-experiment to cross a defined threshold [9,10]. All factors which influence the C_q-value of a sample were randomly generated: the expression level of each gene, its variation coefficient, the range of up- or down-regulation, and the range of on-target or off-target knockdown. The data model delivers a dataset of 21 x 32 values (genes in columns x samples in rows). This *in silico* modelled dataset represents the raw data outcome of a corresponding RT-qPCR gene expression analysis obtained from our standard cell culture assays. For all plots presented the same raw dataset was used (Additional file 2). All variables influencing the data were completely randomized, but expression ranges, gene regulations and the inherent variability were chosen to be realistic based on the knowledge about gene expression we gained from our cell culture assays. The range of the gene regulation here was set from 1-10 folds, the knockdown-efficiency between 45% and 95% and the applied CV-value was justified to 35% of the emulated gene regulation. In order to evaluate the quality of the raw data the distribution of the 32 C_q -values (32 samples) corresponding to every gene is plotted in a virtual RT-qPCR (Figure 1).

Data processing

Gene expression analysis: The raw data was analyzed under the terms of the $\Delta\Delta C_q$-method [11], and then relative expression values (shown in percentage) were plotted gene-wise on a logarithmic ordinate.

Hierarchical cluster analysis: Heat maps of the modelled data were plotted in GenEx software (version 5.3.2.13, MultiD, Sweden) using the ΔC_q-values for each sample corresponding to the gene set [12].

Principal component analysis: The data processing using principal components analysis (PCA) was as well performed using GenEx software. Therefore the ΔC_q-values for each sample corresponding to the gene set were calculated and then pasted into the PCA-algorithm input box. According to the aim of the PCA (results and discussion) the data was mean centered to columns or transposed and then mean centered to rows [7].

Results and Discussion

For this work we designed a data emulator based on the layout of our cell culture assays. Every time used it delivers a unique dataset. Within this emulator the range of the RT-qPCR data constituting parameters can be set up. Subsequently the values are rendered randomly (Additional file 1). In order to show that this data is authentic compared to a cell culture experiment we plotted the raw data (C_q-values) in a virtual RT-qPCR (Figure 1). When this data is normalized to the control samples of the knockdown-treatments (group's a and b) the net knockdown-effects get visible (Figure 2). As a result this plot shows explicitly the knockdown-effect which weighs on a particular gene. The six genes in class II are all predestinated to be knockdown-targets by the simulation. Thus they show down-regulation after the normalization used in this plot. All other genes are no knockdown-targets. But nevertheless some of them show a significant down-regulation. These are the genes from class IV which bear an off-target effect. Slight regulations on the remaining genes (classes I and II) emerge from the random variation coefficient integrated in the data simulation. We inflicted additional gene regulations on some of the class II genes (Figure 2). The additional prefix specification "+" or "-" at the gene name indicates a synergetic

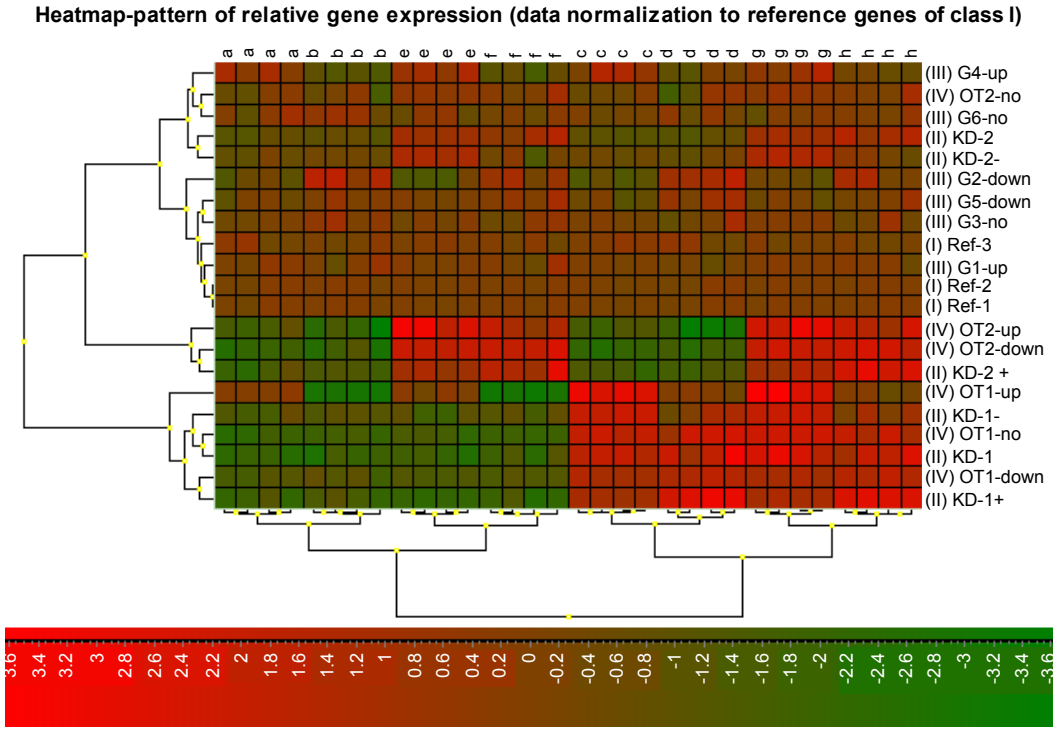

Figure 3: Heat map of relative expression data. ΔC_q-values (normalized against reference gene means) for all 32 samples corresponding to the 21 genes are plotted in GenEx software. Green indicates a relative up-regulation, red a relative down-regulation. The same dataset as in Figure 2 was applied.

effect on the gene regulation when the RNAi-knockdown and the drug appliance are inserted simultaneously. We include this special scenario, because it is a well known phenomenon in *in vitro* or *in vivo* models as we experienced it from our own cell culture assays (data not shown). The same gene expression data is plotted in a heat map in Figure 3. In this visualization mode the ΔC_q -values are directly used without any further normalization to control samples.

The randomly created dataset plotted in Figures 2 and 3 contains some special cases by chance. In the heatmap (Figure 3) we realize a separation between strong knockdown-effects evoked either by knockdown kd-I or knockdown kd-II. However only the genes influenced by knockdown kd-I can be globally separated clearly (KD-1, KD-1+, KD-1-, OT1, OT1-up, OT1-down). kd-II influenced genes fractionally merge undefined between red and green (KD-2, KD-2-, OT2-no). This happens due to the by chance relatively low knockdown-originated regulations on these genes. So, how can I adjust the undefined genes in order?

Therefore we additionally utilized PCA. A plot from a PCA will always have as much data points as the underlying dataset has rows. In every single data point you can include as much resources as you are interested in by integrating more or less columns of a dataset. In our case this means to consider more or less genes measured from one RNA-sample. Each gene will contribute its impact to the sample and influence its position in the PCA-plot. If a gene is strong regulated its impact is high. On the other hand a low regulated gene like a reference gene has poor impact on a sample. If such a gene (column) is included the position of a data point (sample) will not move in the PCA-plot. Figure 4a/b uses the same dataset as Figures 1,2 and 3. In the plots from Figure 4a/b the data was sorted as rows = samples. With the data arranged like this the PCA algorithm regards the impact which

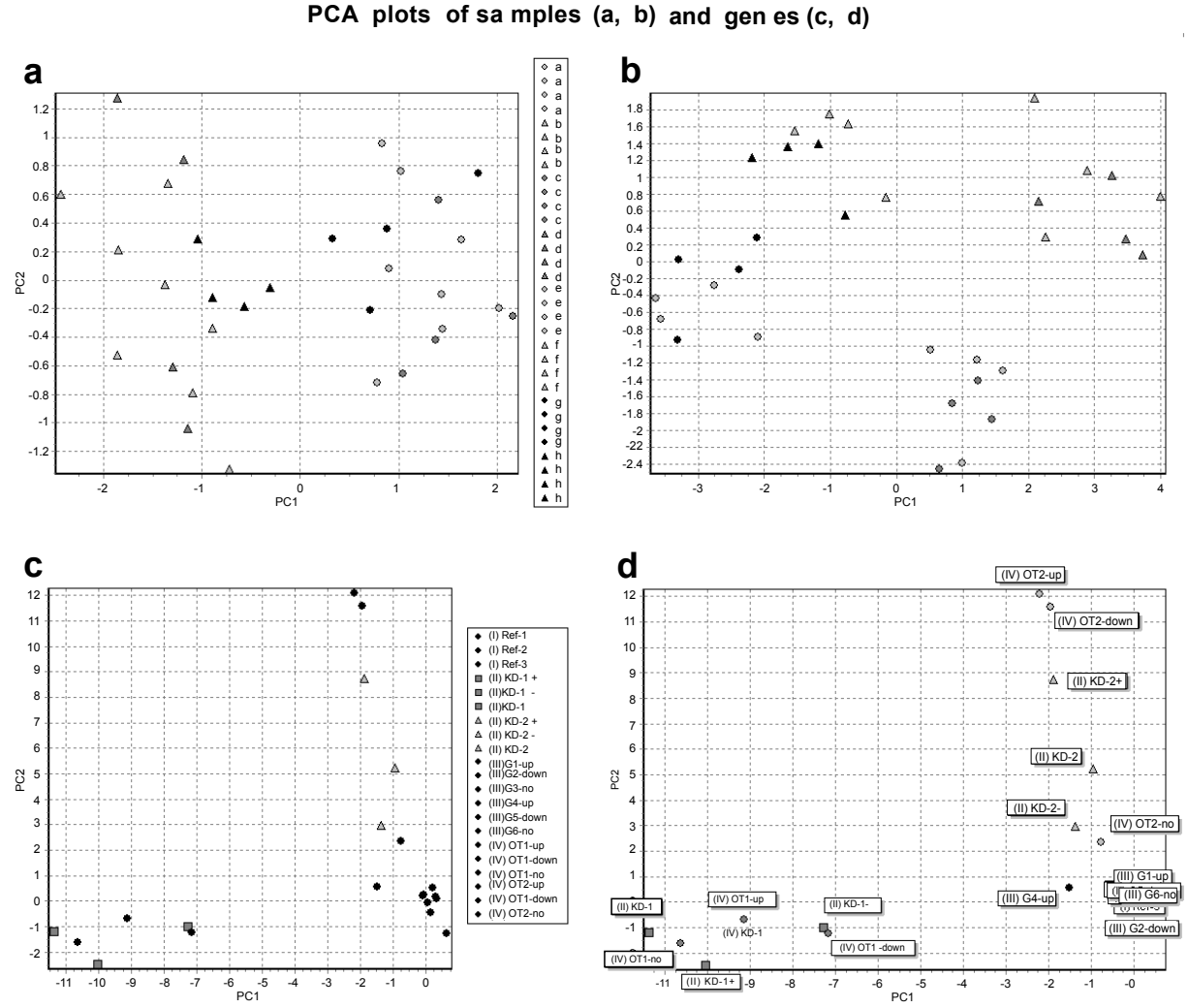

Figure 4: PCAs of the *in silico* data set from Figures 1,2 and 3. (a) PCA of all 32 samples (columns = genes, rows = samples) embracing exclusively genes of classes I and III. The samples cluster in two clouds according to media control (circles) or drug treatment (triangles). (b) PCA from the same data pool as in Figure 4a but including the class IV gene OT2-up. The incorporation of OT2-up divides the two clouds from Figure 4a by separating the samples which bear the off-target effect. (c) PCA with data transposed columns by rows including the complete dataset (32 samples x 21 genes). Genes of class II are indicated as squares (kd-I) or triangles (kd-II). Circles indicate genes from the three remaining classes. (d) Plot Identical with Figure 4c, but layout adjusted: gene names are given and class IV genes (with off-target effect) are now labeled in color according to the knockdown they originate from.

lies on each of the 32 samples. On the algorithm level that means the dimension of the collective gene regulation corresponding to a single sample is expressed as a vector. The resulting PCA-plot then simplifies the multi dimensional dataset and adjusts it to the biggest vectors. That way the major gene regulations are exposed in a two dimensional plot. The way one single gene can influence the plot from a PCA gets clear in Figures 4a/b. While in Figure 4a we can recognize two clouds, there are four clouds in Figure 4b. This is due to the impact that is load on the data points (samples) by integrating more or less resources (genes). In Figure 4a the PCA is cast only with genes that explicitly bear no knockdown-effect (classes I & III). As a result the data points cluster only in accordance with media (circles) or drug (triangles) appliance. In Figure 4b there is only one more column (only one class IV gene: OT2up) included into the dataset. This single load has enough impact to split the two clouds into four. The knockdown lying on OT2up represents a second vector of regulation. For those of the samples in Figure 4b which are treated with the knockdown kd-II the PCA algorithm exposes another vector because the provided effect (an off-target effect) is strong enough. For a screening that means if a gene set we want to analyze is not harassed by any off-target effects the PCA plot will remain in a treatment and a control cloud.

But we can go deeper into data analysis and retrieve more information about the gene regulation. In Figure 4c the PCA casted with the entire dataset is plotted. The major difference is that the data now is transposed columns by rows. In consequence every data point represents one gene including all 32 values from the corresponding samples. Three different kinds of symbols were attributed. Genes which are designated as knockdown-targets (class II) are labeled with squares (kd-I) or triangles (kd-II). All other genes (classes I, III & IV) are labeled as circles. The result is a pattern of the data branching into two directions (or vectors). In our model we have the advantage that we know which genes are additionally charged by an off-target effect (class IV genes). In an off-target screening this information would be the task. If we now label these genes with the color for the knockdown they bear we get a very clearly terminated evaluation (Figure 3d). Genes having an off-target effect from the knockdown kd-I cluster to the branch of kd-I target genes and accordingly behave genes with a kd-II originating off-target effect. This outcome is what we can use to screen for off-target effects using PCA. Genes clustering in a cloud containing a knockdown target are likely influenced by an off-target effect. They should be exclusively checked for off-target effects in an extra RT-qPCR experiment.

But where is the advantage compared to a heat map analysis? The trouble in any case is when weak regulated genes shall be assigned. At that point the vectors in a PCA plot provide us extra information. A very low regulated gene in our dataset is OT2-no. In addition to that the two knockdown targets KD-2 and KD-2- are only weakly down-regulated by the knockdown kd-II (Figures 2 and 3). If we look at these genes in the PCA-plots of Figure 4c/d we realize that this evaluation locates these genes in direction towards the cloud of genes affected by kd-II. Of course OT2-no merges slightly with the cloud of generally low regulated genes, but it settles at the border towards the kd-II containing cloud and explicitly not towards the kd-I influenced genes. In the heatmap these genes cannot be attributed to any treatment group and fade between other low regulated genes. At that point PCA can clearly provide more information about a gene regulation pattern than the HCA alone.

By using virtual RT-qPCR data modelling it is easy to create more similar datasets to demonstrate this evaluation method. For your concern we posted additional twelve PCA-plots online, even with extreme values for the regulative parameters or the CV-value (Additional file 2).

Conclusions

PCA in its innate manner takes that point of view on a dataset that visually maximizes the variation present in a dataset. As a result PCA cannot distinguish information which is not present in a dataset. But the task for the scientist is to pick the most vivid visualization mode for that what was measurable out of the processes which took place on biological level. Analyzing off-target effects using HCA and PCA gives us the advantage that a side effect can be easily attributed to its origin. In the PCA-plot the knockdown targets will cluster with genes regulated by the same cause. In a heat map only strong effects can be identified clearly. Weak regulations will fade between red and green. In addition the PCA assigns a direction to every source of gene regulation. Even when an effect is weak it will cluster towards that cloud of genes which are regulated by the same originator. Separated that way transcript expression patterns can efficiently be analyzed for siRNA specific gene regulations which means off-target screening in a very new mode.

Competing Interests

The authors declare that they have no competing interests.

Authors' Contributions

The idea of utilizing principal component analysis for RNAi-off-target screening was devised by JM. The data modelling spread sheet was designed by JM. The manuscript was drafted by JM and supervised by MP.

Acknowledgements

This study was supported by a grant from the Vereinigung zur Förderung der Milchwissenschaftlichen Forschung an der Technischen Universität München e.V. I vigorously thank Katrin Danowski for basic and Dr. Ales Tichopad for advanced guidance with the GenEx software.

References

1. Jackson AL, Bartz SR, Schelter J, Kobayashi SV, Burchard J, et al. (2003) Expression profiling reveals off-target gene regulation by RNAi. Nat Biotechnol 21: 635-637.
2. Jackson AL, Burchard J, Schelter J, Chau BN, Cleary M, et al. (2006) Widespread siRNA "off-target" transcript silencing mediated by seed region sequence complementarity. RNA 12: 1179-1187.
3. Lim LP, Lau NC, Garrett-Engele P, Grimson A, Schelter JM, et al. (2005) Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. Nature 433: 769-773.
4. Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, et al. (1998) Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature 391: 806-811.
5. Tuschl T, Zamore PD, Lehmann R, Bartel DP, Sharp PA (1999) Targeted mRNA degradation by double-stranded RNA in vitro. Genes Dev 13: 3191-3197.
6. Zamore PD, Tuschl T, Sharp PA, Bartel DP (2000) RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. Cell 101: 25-33.
7. Bergkvist A, Rusnakova V, Sindelka R, Garda JM, Sjogreen B, et al. (2010) Gene expression profiling-Clusters of possibilities. Methods 50: 323-335.
8. Jolliffe IT (2002) Introduction, Principal Component Analysis. Springer New York.
9. Bustin SA, Benes V, Garson JA, Hellemans J, Huggett J, et al. (2009) The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. Clin Chem 55: 611-622.
10. Freeman WM, Walker SJ, Vrana KE (1999) Quantitative RT-PCR: pitfalls and potential. Biotechniques 26: 112-122, 124-125.

11. Livak KJ, Schmittgen TD (2001) Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods 25: 402-408.

12. Pfaffl MW, Vandesompele J, Kubista M (2009) Data Analysis Software. Caister Academic Press, Norwich, UK.