BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 1 2013, pages 77–83
doi:10.1093/bioinformatics/bts631

Gene expression

Advance Access publication October 26, 2012

RIP-chip enrichment analysis

Florian Erhard[1,*], Lars Dölken[2] and Ralf Zimmer[1]

[1]Institut für Informatik, Ludwig-Maximilians-Universität München, 80333 München, Germany and [2]Department of Medicine, University of Cambridge, Addenbrookes Hospital, Hills Road, CB20QQ Cambridge, UK

Associate Editor: Trey Ideker

ABSTRACT

Motivation: RIP-chip is a high-throughput method to identify mRNAs that are targeted by RNA-binding proteins. The protein of interest is immunoprecipitated, and the identity and relative amount of mRNA associated with it is measured on microarrays. Even if a variety of methods is available to analyse microarray data, e.g. to detect differentially regulated genes, the additional experimental steps in RIP-chip require specialized methods. Here, we focus on two aspects of RIP-chip data: First, the efficiency of the immunoprecipitation step performed in the RIP-chip protocol varies in between different experiments introducing bias not existing in standard microarray experiments. This requires an additional normalization step to compare different samples and even technical replicates. Second, in contrast to standard differential gene expression experiments, the distribution of measurements is not normal. We exploit this fact to define a set of biologically relevant genes in a statistically meaningful way.

Results: Here, we propose two methods to analyse RIP-chip data: We model the measurement distribution as a gaussian mixture distribution, which allows us to compute false discovery rates (FDRs) for any cut-off. Thus, cut-offs can be chosen for any desired FDR. Furthermore, we use principal component analysis to determine the normalization factors necessary to remove immunoprecipitation bias. Both methods are evaluated on a large RIP-chip dataset measuring targets of Ago2, the major component of the microRNA guided RNA-induced silencing complex (RISC). Using published HITS-CLIP experiments performed with the same cell line as used for RIP-chip, we show that the mixture modelling approach is a necessary step to remove background, which computed FDRs are valid, and that the additional normalization is a necessary step to make experiments comparable.

Availability: An R implementation of REA is available on the project website (http://www.bio.ifi.lmu.de/REA) and as supplementary data file.

Contact: florian.erhard@bio.ifi.lmu.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on July 3, 2012; revised on September 20, 2012; accepted on October 17, 2012

1 INTRODUCTION

Gene expression is a highly complex process that is controlled on multiple levels by various proteins and RNAs. Various experimental protocols have been established to measure expression levels of mRNAs or proteins, targets of transcription factors or post-transcriptional regulators and many other parameters of gene expression in a genome-wide manner. Each step of such a high-throughput experiment may introduce systematic errors (bias) or random variation (noise) into the generated data, and specialized methods are necessary to deal with particular kind of bias and noise and to answer specific questions using high-throughput data.

The most widely used high-throughput experiments are based on microarrays or next-generation sequencing (NGS) and are designed to measure the amount of all mRNAs in one or multiple conditions (Malone and Oliver, 2011). Based on the raw intensities from a microarray experiment or the sequencing reads from an NGS experiment, several analytical steps are taken, including normalization, summarization and statistical evaluation (Gentleman, 2005). There is a vast amount of literature describing various methods fulfilling these steps to identify differentially regulated genes (Fundel *et al.*, 2008; Irizarry *et al.*, 2006; Marioni *et al.*, 2008; Park *et al.*, 2003; Wang *et al.*, 2009).

Chromatin immunoprecipitation followed by microarray analysis or NGS (ChIP-chip/ChIP-seq) can determine the targets of DNA-binding proteins and has successfully been applied to a wide range of transcription factors and cell types (Birney *et al.*, 2007; Johnson *et al.*, 2007; Ren *et al.*, 2000). In addition to the above-mentioned analysis methods necessary for microarray and NGS data, it has been recognized that additional methods are necessary to successfully determine target sites on the genome, and thus a variety of methods is described in the literature (Ho *et al.*, 2011; Park, 2009; Zhu *et al.*, 2010).

In recent years, it has become apparent that transcriptional regulation is only one part of the machinery carrying out gene regulation. RNA-binding proteins (RBPs) and RNA-binding ribonucleoproteins (RNPs) play important roles and are responsible for splicing, RNA editing, regulation of translation and RNA degradation (Bartel, 2009; Nishikura, 2006; Witten and Ule, 2011). These processes are highly regulated by sequence-specific binding of RBPs or RNPs to the mRNA. MicroRNAs are small 20–24 nt long RNA molecules, which have emerged in recent years as important post-transcriptional regulators involved in all known multicellular organisms. They play important roles in development, tumorigenesis and viral infection. They act by guiding the RNA-induced silencing complex (RISC) to mRNAs by binding to their 3′ untranslated region (3′-UTR) in a sequence-specific manner, which leads to inhibition of translation or RNA degradation (Bartel, 2009).

A powerful experimental high-throughput technique to detect targets of RBPs or ribonucleoproteins, such as RISC, is based on immunoprecipitation (IP) of the RBP or RNP with associated

*To whom correspondence should be addressed.

mRNAs followed by microarray or NGS measurement (RIP-chip/RIP-seq) (Hendrickson *et al.*, 2008; Landthaler *et al.*, 2008; Karginov *et al.*, 2007; Mukherjee *et al.*, 2009; Stoecklin *et al.*, 2008). Targets of the RBP/RNP are enriched in the RIP experiment in comparison with a control measurement using an unspecific antibody or total RNA. Novel techniques including HITS-CLIP, iCLIP and PAR-CLIP also include cross-linking of the protein to the mRNA followed by digestion of the unprotected mRNA to determine the precise location of the target site (Chi *et al.*, 2009; Hafner *et al.*, 2010; König *et al.*, 2010).

The main question in a RIP-chip experiment is to determine the set of target genes of the immunoprecipitated protein. A basic answer is a sorted list of *enrichment values* that can be computed for each gene by dividing the intensity value in the IP fraction microarray by the intensity in the control microarray. This is very similar to standard differential gene expression (DE) experiments: Here, differentially regulated genes can be determined by a sorted list of *fold changes* computed for each gene by dividing the intensity in condition A by the intensity in condition B. Consequently, RIP-chip data are often analysed using standard methods borrowed from the DE setup such as fold changes (Stoecklin *et al.*, 2008), *t* statistics (Mukherjee *et al.*, 2009) or moderated t statistics (Hendrickson *et al.*, 2008).

However, as indicated above, additional experimental steps may introduce additional bias: In contrast to log fold change distributions of DE experiments, log enrichment distributions of RIP-chip experiments are not normal but typically have heavier right tails (Dölken *et al.*, 2010; Mukherjee *et al.*, 2009, see also Fig. 1). This is an indication that RIP-chip is able to separate true targets from the background very efficiently. Here, we exploit these skewed distributions to estimate the biological significance of genes. This is different from the statistical significance usually computed for DE experiments, where *P*-values are related to the reproducibility of the measurements and not to biological relevance.

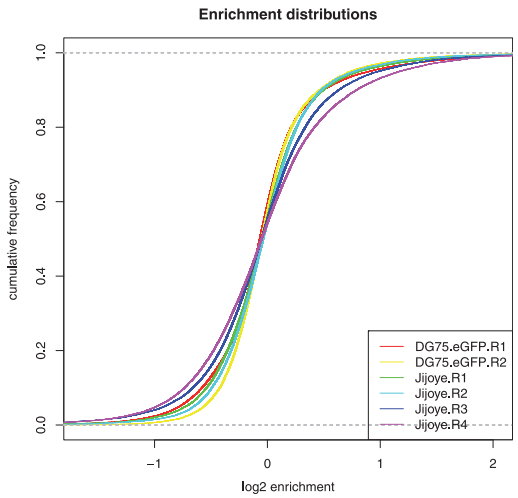

Fig. 1. Measurement distributions for our Ago2 RIP-chip experiment. The distributions of the enrichment values are shown. Although the intensity distributions of the various microarrays are properly normalized using RMA (see Supplementary Fig. S1), the enrichment distributions are significantly different from each other. This is a consequence of differing IP efficiencies and must be accounted for when analysing the respective data

The above-mentioned question about the set of target genes in a RIP-chip experiment only considers a single condition, in contrast to a DE experiment. However, especially for RISC-IP experiments, an additional question is to determine differential microRNA targets between two or several conditions. For instance, if these conditions are *control* and *transfected microRNA* (Hendrickson *et al.*, 2008), differential targets would be targets of the transfected microRNA, and if there are *uninfected* and *virus-infected* cells, differential targets would include targets of viral microRNAs (Dölken *et al.*, 2010). The answer to this question can be given by genes that are more enriched in condition A than in condition B, either by choosing two cut-offs on the corresponding enrichment values (i.e. at least *x* fold enriched in A and at most *y* fold enriched in B) (Dölken *et al.*, 2010), by computing *differential enrichment values* as the ratio of the two enrichment values (Hendrickson *et al.*, 2008) or by a mixture of both approaches (Karginov *et al.*, 2007).

All answers to this question necessarily have to compare enrichment values (i.e. ratios of intensities) of the conditions. However, IP efficiencies may vary between independent experiments, and it is important to account for this bias when comparing enrichment values. Obviously, the same problem exists for the summarization of replicate measurements (see Figs 1 and 4).

Here, we develop a suite of methods to properly analyse RIP-chip datasets, which take care of the unique properties of such data introduced by the IP: First, we use a Gaussian mixture model approach to find statistically meaningful cut-offs for enrichment values. We show that this approach can be used to filter unexpressed genes, that it allows to compute false discovery rates (FDRs) for sets of biological significant genes and that it is in fact a necessary step to make experiments comparable with each other. We also address the problem of differing IP efficiencies by introducing a principal component analysis (PCA)-based method to normalize enrichment distributions in a data-dependent manner. We use publicly available HITS-CLIP data measured for the same cell lines (Riley *et al.*, 2012) as standard-of-truth for evaluation and show that the proposed methods provide significant improvements for the analysis of RIP-chip data.

2 METHODS

2.1 Data processing

The RIP-chip data for this article has been taken from our study of herpes viral RISC-IP experiments (Dölken *et al.*, 2010). After the publication of the original study, additional replicates have been measured, and all chips including the new ones have been processed as described (Dölken *et al.*, 2010). Briefly, RNA from Ago2-IPs and either BrdU-IPs or total RNA have been measured on Affymetrix GeneST arrays, and all raw data have been normalized using robust multi-array average (RMA) (Irizarry *et al.*, 2003), $\log_2$ enrichment values have been computed by subtracting the control-IP/total RNA log intensity from the Ago2-IP log intensity for each probeset and each replicate experiment. Then, probesets have been mapped to Ensembl genes by using the annotation derived from Biomart. HITS-CLIP clusters (i.e. high confidence microRNA target sites) used to evaluate the mixture model approach for the cell line Jijoye has been downloaded from the supplementary

data of Riley *et al.* (2012). We also repeated the same analysis using PAR-CLIP data for Jijoye that has been measured and analysed as described in Hafner *et al.* (2010) in the laboratory of Markus Landthaler at the MDC Berlin (will be published elsewhere).

2.2 Mixture model fitting

Gaussian mixture models for sets of log enrichment values are fitted using the Mclust package in R (Fraley and Raftery, 2002). *Z*-scores for each gene can then be computed using the background distribution:

$$zscore(g) = \frac{e(g) - \mu_{bg}}{\sigma_{bg}} \tag{1}$$

Here, $e(g)$ is the $\log_2$ enrichment of gene g, μ_{bg} and σ_{bg} are the mean and standard deviation of the background component of the gaussian mixture model (which we always take as the component with the smaller mean). The FDR for a cut-off c is defined as the expected fraction of background genes g_b with $e(g_b) > c$ among all genes g with $e(g) > c$:

$$FDR(c) = \frac{1 - cdf_{bg}(c)}{|\{g|e(g) > c\}|} . |BG| \tag{2}$$

cdf_{bg} is the cumulative distribution function of the background component of the mixture model and $|BG|$ the expected number of background genes (estimated by the mixture model). This is mathematically equivalent the Benjamini-Hochberg multiple testing correction (Benjamini and Hochberg, 1995) for the one-sided *P*-values derived from the *z*-scores in Equation (1) multiplied with the expected fraction of background genes.

For the running window approach (see Section 3.1), we first selected a window of w genes with the smallest Ago2-IP intensities and fitted the mixture model (first window). Then we removed the s genes with smallest Ago2-IP intensities and added the next s smallest still unselected genes and again fitted a mixture model. This step was repeated until the window reached the top Ago2-IP intensities. For the analyses, we chose $w = 1000$ and $s = 20$.

We use two metrics to evaluate the fit of background and target distributions:

$$d(bg, t) = \frac{\mu_t - \mu_{bg}}{\sigma_{bg}} \tag{3}$$

$$skew(E) = -\log_{10}(ksp(E, -E)) \tag{4}$$

bg and t are the background and target components of the mixture model, respectively, E is the set of $\log_2$ enrichment values used to fit the mixture model and $ksp(E, -E)$is the *P*-value of the Kolmogorov–Smirnov test comparing the distribution of E with the distribution of negated enrichment values—E. Thus, the distance score $d(bg, t)$ measures the distance of the background and target distributions with respect to the width of the background distribution, whereas the asymetry score $skew(E)$ measures the skewness of the distribution without the need to fit a mixture model.

2.3 PCA

PCA is performed using the function prcomp in R. When there are k experiments/replicates and, therefore, k $\log_2$ enrichment values per gene, PCA is applied to the k-dimensional space of genes. The first principal component is the direction of the greatest variance, and is used to compute the summary enrichment value $\hat{e}(g)$ of gene g by taking the dot product of the replicate measurement *z*-scores $\langle z_1(g), \ldots, z_k(g)\rangle$ and the direction of the first principal component $PC1$:

$$\hat{e}(g) = \langle z_1(g), \ldots, z_k(g)\rangle . PC1 \tag{5}$$

The geometrical interpretation of this weighted average of the replicate enrichment values is that the dot product does an orthogonal projection of the k dimensional point g onto the first principal component and measures the distance to the origin.

It is not necessary to centre the points before PCA, as we perform PCA on the *z*-scores derived from the mixture modelling approach [see Equation (1)], the point cloud is naturally centred at the means of the background distributions, and centring at the overall mean may not be appropriate. Also, we perform PCA only on targets (as defined by an FDR of 1%). This is necessary because if the number of background genes is much higher than the number of target genes, stochasticity in the background could mask the effects in the target genes to some extent.

Differential targets will deviate from this vector in a specific direction: e.g. if we have two replicates of two conditions A and B and, therefore, an enrichment vector $\langle z_{a_1}, z_{a_2}, z_{b_1}, z_{b_2}\rangle$, any gene that is target specifically in A has greater enrichment in A than in B: z_{a_1} and z_{a_2} is greater than z_{b_1} and z_{b_2}. Thus, if there are enough differential targets, the second principal component will point into the direction of the deviations of their enrichment vectors. Therefore, the summarized differential enrichment value $\hat{e}_d(g)$ can be computed similarly to the overall summary enrichment value in Equation 5 by taking the dot product of the *z*-score vector $\langle z_1(g), \ldots, z_k(g)\rangle$ and the direction of the second principal component $PC2$:

$$\hat{e}_d(g) = \langle z_1(g), \ldots, z_k(g)\rangle . PC2 \tag{6}$$

Both the enrichment value $\hat{e}(g)$ and the differential enrichment value $\hat{e}_d(g)$ incorporate a linear normalization that removes bias owing to differing IP efficiencies.

3 RESULTS

3.1 Select relevant genes

The first step in our analysis of RIP-chip data is the filtering of unexpressed genes. On modern microarrays, such as the Affymetrix GeneST arrays, used for our data, probesets against all known human genes are available. Even if virtually all probesets have non-zero intensities, we can expect that only a fraction of all genes is expressed in a specific condition. We noticed that the asymmetry of the log enrichment distribution is not observable over the whole range of IP intensities (see Fig. 2a and Supplementary Fig. S2). For low intensity genes, the distribution indeed looks normal, which is expected for a set of genes that is not or almost not expressed. Therefore, we used a running window approach for fitting the mixture model (see Section 2) and evaluated each window with respect to the distance of the two components of the fitted model and the extent of asymmetry (see Fig. 2b). At intensity values ≈5, a significant increase in both scores was observable. We chose to use all genes above an intensity cut-off where $d(bg,t) > 1$ and $skew(E) > 2$, i.e. where the means of the two mixture components are at least one standard deviation away from each other, and where the asymmetry becomes significant with *P*-value 0.01.

We performed this running window approach for all previously published RIP-chip experiments from Dölken *et al.* (2010) and for two additional replicates of the control cell line DG75-eGFP and the Epstein-Barr Virus (EBV)-infected cell line Jijoye, respectively. For the additional DG75-eGFP replicates, the microarrays showed poor RNA quality and applying our running window approach to these bad quality experiments indeed did not yield a mixture model (see Fig. 2c). Thus, we can apply our method also for filtering poor quality experiments

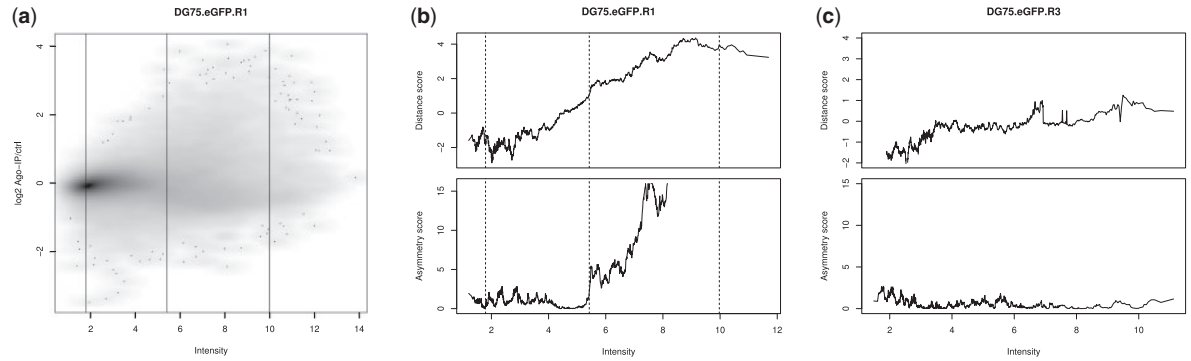

Fig. 2. Selecting expressed genes. **a** shows a density scatterplot of the $\log_2$ IP intensities against the $\log_2$ enrichment values of all genes for the first Jijoye RIP-chip replicate. The running window mixture models for the three indicated vertical lines are shown in Supplementary Figure S2. In **b** and **c**, the distance and asymmetry scores are plotted for all windows for replicate one and three of the DG75-eGFP experiment. Starting from intensity values of ≈5, the distribution seems to be a mixture of two normal distributions. In contrast to the first replicate, the third does not show the expected behaviour of the mixture of a background and target distribution, which was a consequence of poor RNA quality in this experiment

from a dataset, and we excluded the two additional DG75-eGFP replicates from further analyses accordingly.

We also noticed that the background distribution is not the same over the whole spectrum of intensity values. Therefore, we computed *z*-scores for each gene using mean and standard deviations obtained from the running window approach. This is very similar to well-known non-linear normalization techniques (Yang *et al.*, 2002), with the difference that the model for normalization is not fitted to all data but only to the background.

3.2 Determining microRNA targets

Computing *z*-scores from the raw enrichment values based on the fitted background distribution can be interpreted as a subtraction of this background. The background here does not consist of the unexpressed genes, but of the expressed but not targeted genes. This background subtraction step can have a great effect: For the four Jijoye RIP-chip replicates, we observe that without subtraction, it seems that the IPs of replicates three and four were more efficient than of the other two replicates, as there are more genes enriched >2-fold, for instance (see Fig. 3a and b). However, after background subtraction, replicates one and two show a larger fraction of enriched genes.

Obviously, if an IP was more efficient than another, its induced ranking of genes will better predict a gold standard of microRNA targets. HITS-CLIP is an experimental technique that is able to identify target sites of microRNAs with high confidence (Chi *et al.*, 2009). Thus, using the publicly available HITS-CLIP data for Jijoye (Riley *et al.*, 2012), we can construct a gold standard by taking all genes as true targets that have at least *n* HITS-CLIP target sites. Independent on the choice of *n*, replicates one and two induce rankings that are in better agreement with HITS-CLIP data (see Fig. 3c for $n=1$), and thus, background subtraction is a necessary step. We also repeated this analysis using in-house, unpublished PAR-CLIP data for Jijoye (∼14.000 sites on ∼5500 genes, will be published elsewhere) leading to the same conclusions (data not shown).

Furthermore, we propose that the fitted background distribution allows to compute valid FDRs for microRNA targets. For a cut-off *c*, the FDR is defined as the expected fraction of non-target genes. Obviously, a non-target gene should contain less HITS-CLIP target sites than target genes on average. If we compute the average number of HITS-CLIP target sites per gene for the set of targets defined by cut-off *c* on the RIP-chip data, the dependence on the corresponding FDR should thus be linear with a negative slope: For instance, if the FDR is twice as high, we expect twice as many non-target genes. Therefore, the average number of HITS-CLIP target sites per gene should decrease by a factor that is dependent on the true average number of HITS-CLIP targets sites per target gene and non-target gene. For all four replicates, the plot of the FDR against the fraction of HITS-CLIP target sites per gene is roughly a straight line (see Supplementary Fig. S3), and even if the enrichment/*z*-score distributions are different, the slopes and intercepts of linear fits to all four plots are very similar to each other (see Supplementary Fig. S4).

This also allows us to estimate the average number of HITS-CLIP target sites per target gene and non-target gene by taking the value of the linear fit at $\mathrm{FDR}=0\%$ and $\mathrm{FDR}=100\%$, respectively. Based on the RIP-chip data as a reference, we can estimate that HITS-CLIP produces ≈0.8 target sites per expressed target gene and ≈0.2 target sites per non-target gene (see Supplementary Fig. S4).

3.3 Taking replicates into account

As indicated above, IP efficiencies between replicate experiments may be very different from each other. These differences introduce bias into such a dataset, and no RIP-chip study known to us has properly accounted for that. Our mixture model approach also cannot remove this bias from RIP-chip data. The problem becomes obvious when we visually inspect scatterplots across replicate enrichment values/*z*-scores.

For replicates one and two of our Jijoye RIP-chip data, the main cloud of target genes roughly scatters around the main diagonal in Figure 4a, whereas for the comparison of replicates one and three, the diagonal is far away from the main cloud (Fig. 4b). The canonical way for summarizing replicates is to take the unweighted mean of the enrichment values/*z*-scores. This can geometrically be interpreted as an orthogonal

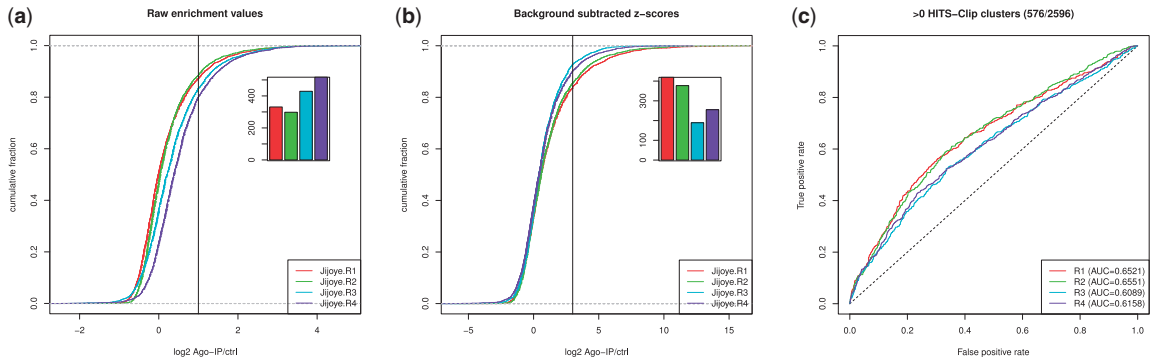

Fig. 3. Background subtraction is necessary. **a** and **b** show the enrichment distributions of expressed genes in the four Jijoye RIP-chip replicates. Raw enrichment values indicate that the IPs of replicates three and four were more efficient based on the number of genes enriched >2-fold (see corresponding inset). After background subtraction, however, replicates one and two show a larger fraction of enriched genes. Replicates one and two have significantly better correspondence to the HITS-CLIP experiment performed in Jijoye, which shows the need for background subtraction

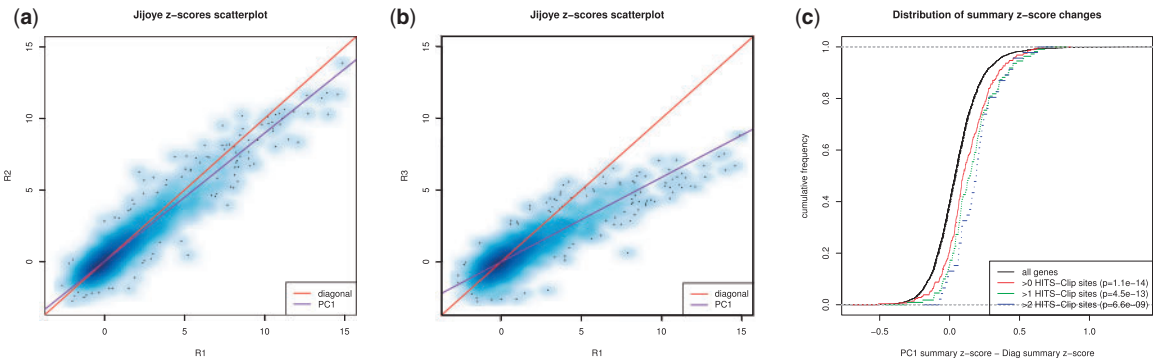

Fig. 4. Differing IP efficiencies require normalization before computing summary values. In contrast to replicates one and two of the Jijoye RIP-chip experiments, where IP efficiencies are very similar (**a**), replicate three is different (**b**). Normalizing replicates using the first principal component (PC1) significantly improves the summary z-score with respect to HITS-CLIP data as reference: The difference of the normalized score to the un-normalized score is significantly greater for genes with HITS-CLIP target sites (coloured distributions) in comparison with all differences (black distribution). The improvement is even more pronounced for genes with multiple HITS-CLIP target sites (**c**)

projection onto the diagonal vector $d = (0.25, 0.25, 0.25, 0.25)$ and measuring the distance of the projected point to the origin. Thus, all four-dimensional points lying on any hyperplane orthogonal to d would get the same summary value. Such a hyperplane would not cut the main cloud of target genes in the scatter plot of replicates one and three orthogonally, which is only a consequence of different IP efficiencies.

However, the first principal component of this point cloud defines such orthogonal hyperplanes, and we use the components of the corresponding rotation vector to compute a weighted mean accounting for all linear effects of differing IP efficiencies.

We can evaluate this additional step again by using the HITS-CLIP data as reference. We consider the differences between each normalized summary value and the corresponding un-normalized value. The difference for HITS-CLIP sites containing genes is statistically significantly greater than for other genes ($P < 10^{-14}$, Kolmogorov–Smirnov test, see Fig. 4c), and the more HITS-CLIP targets sites are found for a gene, the more pronounced is its positive change.

3.4 Determining differential microRNA targets

To find differential microRNA targets, experiments of different conditions must be compared, i.e. genes must be identified, which are more enriched in one condition in comparison with the other. Obviously, a similar problem as in the summarization of replicates plays a role: How can we account for differing IP efficiencies if we compare four replicates of the EBV-infected cell line Jijoye to the two replicates of the control cell line DG75-eGFP?

We can extend our method for summarizing replicates to the differential problem: The first principal component corresponds to the direction of greatest variance, which is the direction of common targets under the assumption that there are enough common targets (both are B cell lines). Differential microRNA targets exclusive to Jijoye should have positive enrichment values in the Jijoye RIP-chip replicates and smaller values in DG75-eGFP. These targets induce variance into the corresponding direction of the six-dimensional space such that the second principal component corresponds to the IP efficiency normalized direction of differential targets (see Fig. 5a).

To compare the PC2 normalized differential enrichment values to the un-normalized differential enrichment (i.e. subtract the enrichment mean of DG75-eGFP from the mean of Jijoye), we exploit the fact that microRNAs are able to downregulate expression of target mRNAs (Bartel, 2009; Guo *et al.*, 2010), and that mRNA levels were measured as well in the RIP-chip experiment: x fold downregulated genes get consistently and

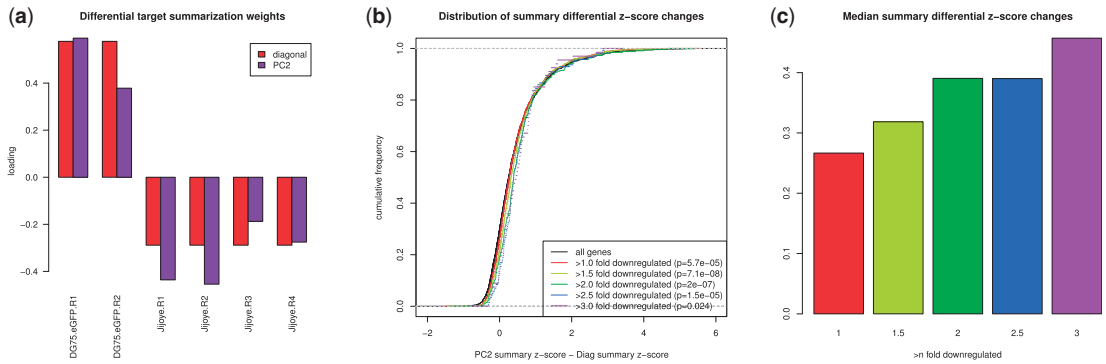

Fig. 5. Differing IP efficiencies require normalization before computing differential targets. The second principal component in **a** is able to discover the experimental structure. Its loadings can be used as weights to compute a differential enrichment value that is normalized for different IP efficiencies, in contrast to the standard way of subtracting the mean log enrichment in DG75-eGFP from the mean log enrichment in Jijoye (corresponding to weights indicated in red). The difference distribution of the normalized differential enrichment values and the un-normalized ones is shown in **b**. Differential targets are expected to be downregulated, and indeed, the difference is significantly greater than background for downregulated genes. As illustrated in **c**, this effect is more pronounced the higher the downregulation is

significantly higher scores after normalization as compared with all other genes, independent on the choice of x (see Fig. 5). Thus, after normalization, significantly more RIP-chip targets are downregulated than without normalization (independent on the particular threshold used to define RIP-chip targets and downregulated genes).

4 DISCUSSION

A similar approach to our Gaussian mixture modelling (GMM) has already been used in (Mukherjee *et al.*, 2009); however, GMM was applied to summarized enrichment values, and log odds ratios (LOD scores) were computed as the ratio of the two scaled mixture components. LOD scores were then used in two different ways: First, they were directly subjected to gene set enrichment analysis (Subramanian *et al.*, 2005), where they have no advantage over directly using enrichment values (as the weighted Kolmogorov–Smirnov statistic used for gene set enrichment analysis is non-parametric and only sensitive to the ranking of the genes). Second, the authors used a cut-off of LOD >0 to define a set of targets. However, choosing a cut-off based on the LOD is still arbitrary and not statistically meaningful in contrast to our FDRs. If we used the LOD to define a cut-off, we would get FDRs ranging from 5 to 15% in our experimental dataset.

We could show that our refined mixture modelling approach has several advantages: First, it allows us to filter unexpressed genes. When comparing two conditions (e.g. virus infected cells expressing viral microRNAs versus non-infected cells), expression of a gene targeted by cellular microRNAs below the detection limit of the microarray in one, but not the other cell line, would result in the misinterpretation of this to be a target of the viral microRNAs. Second, for experiments with poor IP efficiency, we observed extremely poor distance and asymmetry scores over the whole intensity range and could remove these bad replicates from further analyses. Third, it helps to compare experiments with each other (see Fig. 3c) and finally, we can compute valid FDRs.

Furthermore, the comparison of the RIP-chip FDRs to HITS-CLIP data revealed important properties of both the RIP-chip and HITS-CLIP techniques: Both are designed to identify microRNA targets, and, naturally, they agree significantly ($P < 2.2 \times 10^{-16}$, Kolmogorov–Smirnov test), a fact that is also reflected in the negative slope of the linear fit to the FDR against sites per gene plot. However, the agreement is not perfect, and we estimate an average number of ≈0.8 HITS-CLIP target sites per RIP-chip target gene and of ≈0.2 per RIP-chip non-target gene. Even if HITS-CLIP target sites may be erroneous, and the CLIP techniques may implicate additional bias (König *et al.*, 2010; Kishore *et al.*, 2011), we expect that not all of the ≈0.2 sites per background gene are true errors: Such inconsistencies may be due to differing experimental steps (e.g. different antibodies used for IP) or due to differences the Jijoye cell cultures have accumulated in the two laboratories since the cell line has been established. Also, as HITS-CLIP does not control for target mRNA abundance, it may find several weak sites on highly expressed genes that are biologically irrelevant (i.e. not contributing significantly to regulation of its expression). Such a gene should not be enriched in a RIP-chip experiment and could explain many cases of HITS-CLIP sites on background genes. Thus, even if CLIP techniques have several advantages (e.g. they are able to identify target sites instead of target genes), RIP-chip is still a useful complementary method.

The second, novel method introduced in this article is to use principal components to normalize for different IP efficiencies. Evaluation using HITS-CLIP data or the differential expression of target genes shows that the normalization improves results significantly. The normalization proposed can only account for linear bias between experiments. This very lenient normalization appears to be sufficient, as affine offsets are already removed by the mixture model approach, and non-linear effects are not recognizable in a visual inspection.

Our proposed methods do not include a way to compute statistical significance, e.g. like a t-test for standard DE experiment. However, this can be accomplished in a straight-forward way, as all available tests could directly be used after our linear normalization has been applied to a dataset.

5 CONCLUSION

In this article, we presented methods we developed to analyse RIP-chip data. In comparison with standard DE experiments, the additional IP step introduces special requirements for the data analyis. First, we use GMM to determine biologically significant target genes; second, we use a linear normalization technique based on PCA to remove bias introduced by the IP. The evaluation of both methods using independent data showed a significant improvement in comparison with standard approaches: The background of not enriched genes can be removed, valid FDRs can be calculated and the comparability of both replicates and differential experiments is improved.

Funding: This work was supported by the German Bundesministerium für Bildung und Forschung (NGFN-Plus #01GS0801)

Conflict of Interest: none declared.

REFERENCES

Bartel,D.P. (2009) MicroRNAs: target recognition and regulatory functions. *Cell*, **136**, 215–233.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.*, **57**, 289–300.

Birney,E. *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.

Chi,S.W. *et al.* (2009) Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature*, **460**, 479–486.

Dölken,L. *et al.* (2010) Systematic analysis of viral and cellular microRNA targets in cells latently infected with human gamma-herpesviruses by RISC immunoprecipitation assay. *Cell Host Microbe*, **7**, 324–334.

Fraley,C. and Raftery,A.E. (2002) Model-based clustering, discriminant analysis and density estimation. *J. Am. Stat. Assoc.*, **97**, 611–631.

Fundel,K. *et al.* (2008) Normalization strategies for mRNA expression data in cartilage research. *Osteoarthritis Cartilage*, **16**, 947–955.

Gentleman,R. (2005) *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer, New York.

Guo,H. *et al.* (2010) Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature*, **466**, 835–840.

Hafner,M. *et al.* (2010) Transcriptome-wide identification of RNA-Binding protein and MicroRNA target sites by PAR-CLIP. *Cell*, **141**, 129–141.

Hendrickson,D.G. *et al.* (2008) Systematic identification of mRNAs recruited to argonaute 2 by specific microRNAs and corresponding changes in transcript abundance. *PloS One*, **3**, e2126.

Ho,J. *et al.* (2011) ChIP-chip versus ChIP-seq: lessons for experimental design and data analysis. *BMC Genomics*, **12**, 134.

Irizarry,R.A. *et al.* (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics (Oxford, England)*, **4**, 249–264.

Irizarry,R.A. *et al.* (2006) Comparison of affymetrix GeneChip expression measures. *Bioinformatics (Oxford, England)*, **22**, 789–794.

Johnson,D.S. *et al.* (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science (New York, NY)*, **316**, 1497–1502.

Karginov,F.V. *et al.* (2007) A biochemical approach to identifying microRNA targets. *Proc. Natl. Acad. Sci. USA*, **104**, 19291–19296.

Kishore,S. *et al.* (2011) A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nat. Methods*, **8**, 559–564.

König,J. *et al.* (2010) iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat. Struct. Mol. Biol.*, **17**, 909–915.

Landthaler,M. *et al.* (2008) Molecular characterization of human argonaute-containing ribonucleoprotein complexes and their bound target mRNAs. *RNA (New York, NY)*, **14**, 2580–2596.

Malone,J.H. and Oliver,B. (2011) Microarrays, deep sequencing and the true measure of the transcriptome. *BMC Biology*, **9**, 34.

Marioni,J.C. *et al.* (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, **18**, 1509–1517.

Mukherjee,N. *et al.* (2009) Coordinated posttranscriptional mRNA population dynamics during t-cell activation. *Mol. Syst. Biol.*, **5**, 288.

Nishikura,K. (2006) Editor meets silencer: crosstalk between RNA editing and RNA interference. *Nat. Rev. Mol. Cell Biol.*, **7**, 919–931.

Park,P.J. (2009) ChIP–seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, **10**, 669–680.

Park,T. *et al.* (2003) Evaluation of normalization methods for microarray data. *BMC Bioinformatics*, **4**, 33.

Ren,B. *et al.* (2000) Genome-wide location and function of DNA binding proteins. *Science (New York, NY)*, **290**, 2306–2309.

Riley,K.J. *et al.* (2012) EBV and human microRNAs co-target oncogenic and apoptotic viral and human genes during latency. *EMBO J.*, **31**, 2207–2221.

Stoecklin,G. *et al.* (2008) Genome-wide analysis identifies interleukin-10 mRNA as target of tristetraprolin. *J. Biol. Chem.*, **283**, 11689–11699.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA*, **102**, 15545–15550.

Wang,Z. *et al.* (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10**, 57–63.

Witten,J.T. and Ule,J. (2011) Understanding splicing regulation through RNA splicing maps. *Trends Genet.*, **27**, 89–97.

Yang,Y.H. *et al.* (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.*, **30**, e15.

Zhu,L.J. *et al.* (2010) ChIPpeakAnno: a bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinformatics*, **11**, 237.