Pre-processing Optimization of RNA Immunoprecipitation Microarray Data

Emiliano Barreto-Hernadez [1,3]. Corresponding author
[1] Instituto de Biotecnología
Universidad Nacional de Colombia
Bogotá, Colombia
Tel: +57 1 3165000 Ext 16956
Fax: +57 1 3165415
e-Mail: ebarretoh@unal.edu.co

Margarida Gama-Carvalho[2]
[2] Inst. de Medicina Molecular
Faculdade de Medicina, Universidade de Lisboa
Av. Professor Egas Moniz, 1649-028 Lisboa, Portugal
Tel: +35121794 0157
Fax: +351217951780
e-Mail: m.gamacarvalho@fm.ul.pt

Lisete Sousa[3]
[3] DEIO and CEAUL
Faculdade de Ciências da Universidade de Lisboa
Bloco C6 -Piso 4
Campo Grande, 1749-016 Lisboa, Portugal
Tel: +351217500235
Fax: +351217500081
e-Mail: lmsousa@fc.ul.pt

Financial support: Projects FCT/POCI2010 and PTDC/MAT/64353/2006, and ALBAN fellowship No. E06D101266CO.

Keywords: Affymetrix, binding proteins, linear models, bioinformatics.

Abbreviations:	RBP: RNA-binding proteins ChIP-chip: Chromatin immunoprecipitation on chip RIP-chip: RNA immunoprecipitation on chip RMA: Robust linear model PM: Perfect match

ABSTRACT

Pre-mRNA splicing is an essential step in the post-transcriptional gene expression control involving protein-splicing factors like U2AF, which is exported to the cytoplasm and implicated in additional cellular functions. Identification of U2AF-associated mRNAs under native conditions was performed by immunoprecipitation and hybridization to Affymetrix GeneChip. Normalization and gene selection methods were performed, but the results were not reliable as they were different for different procedures, mainly because more than 20% of the mRNAs detected are differently enriched and the common normalization methods are based on small differences between them. We implemented a background correction method inspired in a non-specific hybridization method used for pre-processing data from ChIP-Chip technology. In this work, linear regression models are used to model in each array the non-specific hybridization, accounting for interactions between each three consecutive nucleotides into the probe sequence. Every probe intensity on the array was standardized using its predicted intensity and the probes' variance for similar predicted intensities. The standardized probe intensity values showed no need for further normalization and could be directly compared. We propose a probe set score, and a probe set enrichment value (ENRval) and its respective p-value for gene enrichment selection.

INTRODUCTION

Gene expression control is one of the most important mechanisms for the cell function and its missregulation at any level can lead to disease. It integrates intrinsic and environmental information and takes place at two main levels: transcriptional and post-transcriptional.

Transcriptional regulation level has been highly studied, because of technical and historical reasons: it is one of the most important step of gene expression, and there are well-established methods for traditional studies (Kadonaga, 2004) and recent genome-wide approaches such as expression profiling (Lockhart and Winzeler, 2000) or location analyses of transcription factors and global chromatin remodeling (Hanlon and Lieb, 2004).

Although the post-transcriptional regulation level has been less studied, it is responsible for a substantial complexity of the control of gene expression, including the processing (through splicing), exportation, localization, turnover and translation of mRNAs (Mata et al. 2005).

Post-transcriptional control is mediated by small interfering RNAs and microRNAs, associated with protein effector complexes, that control the degradation and translation of target transcripts and it is also possible that these small RNAs regulate a large subset of mRNAs in a combinatorial manner. (Tomari and Zamore, 2005).

Several RNA-binding proteins (RBPs) are also part of the Post-transcriptional control level, determining the fate of the tagged transcripts and seem to coordinately regulate specific subsets of mRNAs (Mata et al. 2005). The small ribonucleoprotein U2AF, for example, was found to be an auxiliary factor in the spliceosome ensemble inside the nucleus cell (Zamore and Green, 1989; Zamore et al. 1992), which is responsible for the AG dinucleotide recognition at the 3' splicing site and has been reported as going to the cytoplasm together with mature mRNA sets, with evidence to participate at the translation control level of some mRNA sets (Gama-Carvalho et al. 2001).

The systematic identification of RNA targets has provided clues to unsuspected functions of well-known RBPs and, as shown in figure 1, they have been identified using microarrays. To achieve this, the RBP is purified together with its associated RNAs after immunoprecipitation using an epitope tag or antibodies raised against the RBP. The RNAs from the immunoprecipitate are then isolated, PCR purified, labeled and hybridized to DNA microarrays

(Gerber et al. 2004; Gama-Carvalho et al. 2006). This technology is analogous to ChIP-chip (chromatin immunoprecipitation on chip, Hanlon and Lieb (2004)) and is sometimes called RIP-chip (RNA immunoprecipitation on chip).

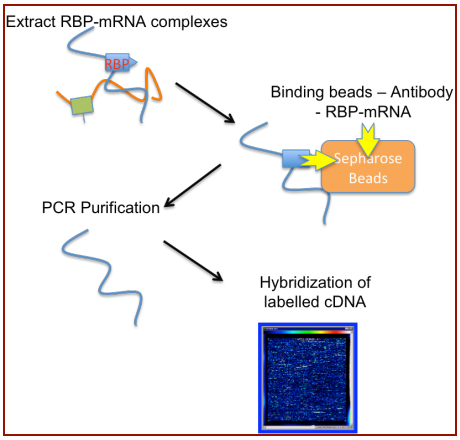

Figure 1. Systematic identification process of mRNA targets of a specific RBP using RIP-chip.

An important remark is that RIP-chip (and ChIP-chip) experiments are not performed in the same way as the DNA expression experiments. In the RIP-chip experiment the immunoprecipitated and input samples are processed in different ways (Gama-Carvalho et al. 2006) and are hybridized separately, obtaining two microarrays with strong differences in the gene expression levels. The main problem is that common pre-processing methods, like RMA (Irizarry et al. 2003), Dchip (Li and Wong, 2003), etc., were implemented with the assumption that there are only few differences in the expression levels between conditions and there are not specific methods developed for this kind of experiments.

In this framework, we explore how the data analysis of RIP-chip experiment is affected when some of the common pre-processing and gene selection methods are applied to it. We present an alternative method for pre-processing inspired in ChIP-chip experiment data analysis (Johnson et al., 2006; Gibbons, 2005) that takes into account the differences between conditions.

METHODS

We used a RIP-chip data from the microarrays obtained for identifying U2AF65 associated with postsplicing mRNP complexes, made by Gama-Carvalho et al. (2006). The data correspond to three independent immunoprecipitation experiments performed to identify U2AF65-associated mRNAs, where the polyadenylated RNA from the input and immunoprecipitated samples was reverse transcribed, end-tagged, and amplified by PCR within the linear range. Each of the resulting cDNAs samples were then labeled and hybridized to the Affymetrix 133U Plus 2.0 microarrays, which provided comprehensive coverage of the whole transcribed human genome (Affymetrix).

The six data set from the experiment were processed using methods form Bioconductor package (Gentleman et al., 2004) and scripts wrote in R program (R Development Core Team, 2008).

Pre-processing Analysis. For Background correction was used only PM on raw intensity scale using a Robust linear model (RMA) (Irizarry et al. 2003) and its variation (cg-RMA) (Wu et al. 2004), where is introduced the estimation of the non-specific hybridization effect. Normalization was performed using quartile normalization and the probe set summarization using median polish (Irizarry et al. 2003). Also, the data were pre-processing using the DChip program (Li and Wong, 2003) where the Invariant Set Normalization method (Li and Wong,

2001) was used to normalize arrays at probe cell level to make them comparable, and the model-based method (Li and Wong, 2001) was used for probe-selection and computing expression values. These expression levels were attached with standard errors as measurement accuracy, which were subsequently used to compute 90% confidence intervals of fold changes (Li and Wong, 2001).

Gene Enrichment Selection: For comparison reasons we used the data obtained after RMA background correction, quartile normalization and median polish summarization to perform gene selection (enriched genes) using the following Bioconductor libraries: Limma (fits a linear model for every gene); eBayes (computes moderated t-statistics, moderated F-statistic, and log-odds for differential expression by empirical Bayes shrinkage of the standard errors towards a common value); decideTests with p-value < 0.05 (computes multiple testing procedures for determining whether each statistic in a matrix of t-statistics should be considered significantly different from zero (Smyth, 2004)); Rank Products with FDR <0.05 (non-parametric test that detects items that are consistently highly ranked in a number of lists (Hong et al., 2006)). The results from theses two methods were compared to the results obtained using the DChip program considering PFR <0.05 and p-value <0.05) (package implementing a model-based expression analysis of oligonucleotide arrays that allows probe-level analysis on multiple arrays making possible to assess standard errors for the expression indexes and automatic probe selection (Li and Wong, 2003)).

Sequence-specific affinity Models Estimate. The probe behavior estimation models in the present work takes advantage of each Affymetrix U133 Plus 2.0 array containing more than 1.2 million 25-mer oligonucleotide probes, allowing for an accurate and robust prediction of probe sequence effects.

In analogy with the work of Johnson et al. (2006) on tiling arrays data analysis, where the authors proposed a probe affinity model for background correction and Sequence-Specific probe behavior models for gene expression microarrays (Naef and Magnasco, 2003; Wu et al. 2004), we propose the following linear Sequence-Specific affinity model for background correction if affymetrix arrays:

(1) $$\log(PM_i) = \sum_{j=1}^{25} \sum_{k \in (A,C,G,T)} \beta_{jk} I_{ijk} + \sum_{k \in (A,C,G,T)} \gamma_k n_{ik}^2 + \varepsilon_i$$

- PM_i Perfect Match probe intensity value;
- n_{ik} Number of nucleotide k in probe i;
- I_{ijk} Indicator function such that $I_{ijk} = 1$ if the nucleotide at position j in probe i is k, and $I_{ijk} = 0$ otherwise;
- β_{jk} Effect of each nucleotide k at each position j;
- γ_k is the effect of nucleotide count squared;
- ε_i is the probe-specific error term, assumed to follow a normal distribution.

We also modified the above model for taking into account the nucleotide position-specific interaction, replacing the expression:

(2) $$\sum_{j=1}^{25} \sum_{k \in (A,C,G,T)} \beta_{jk} I_{ijk}$$

by

(3) $$\sum_{j=1}^{26-it} \sum_{k_1,\ldots,k_{it} \in \{A,C,G,T\}} \beta_{jk_1\ldots k_{it}} I_{ijk_1} \ldots I_{i(j+it-1)k_{it}}$$

it Interaction nucleotide number with the nucleotide k in the probe i.

Finally we improved the correlation between predicted probe intensities and the observed values, using a 3 nucleotides interaction model, a random sample of 20000 PM sequences and the implementation of the following iterative process:

1. Linear model parameters estimation.
2. If the difference between multiple R square (R^2) of the fitted model and the multiple R^2 of the fitted model in the previous iteration is higher than 0.001, the outliers are removed from the sample and the process returns to step 1.
3. Probe baseline intensity is estimated and the process is stopped.

The linear model parameter estimation and the posterior probe baseline intensity estimation, m_i, are calculated using the *lm* library of the R program, and applied to each RIP-chip and input array sample independently.

Probe Standardization. Following Johnson et al. (2006) we made a probe standardization for each array. The probes in the array were divided into ''affinity bins'' each containing ~3000 probes with similar $\hat{m}_i$, and the observed sample variance was estimated within each affinity bin and used as the probe variance for each probe in the bin.

Individual probes in each array were then standardized using the following equation (Johnson et al., 2006):

(4) $$t_i = \frac{\log(PM_i) - \hat{m}_i}{SD_{i\_affinity\_bin}}$$

$\hat{m}_i$ Probe *i* baseline intensity predicted
$SD_{i\_affinity\_bin}$ Observed sample variance was estimated within each affinity bin.

Probe Set Summarization. Following Johnson et al. (2006), we proposed a probe set score ($PSsco$), calculated as follows:

(5) $$PSsco_s = \sqrt{n_p} * TM_s$$

TM_s Trimmed-mean of the t values into the probe set s, removing the top 10% and bottom 10% t values into the probe set s.
n_p Probe number used to calculate TM_s

For multiple replicates, $PSsco$ are calculated pooling all ***t*** values for a specific probe set across all replicates. Having many replicates, higher will be the prediction confidence.

Enriched Values ($ENRval$). We propose the probe set enriched value ($ENRval$), to be calculated as follow:

(6) $$ENRval = PSsco_{IP} - PSsco_{Input}$$

When there are more than 2 experimental replicates, the $ENRval$ values are divided by SD_{Input} (standard deviation of the $\boldsymbol{t}$ values used in $PSsco_{Input}$ calculation). This reduces the score in very noisy regions or where the Inputs samples give inconsistent results.
A p-value for each $ENRval$ is calculated assuming the null distribution to be symmetric about the $ENRval$**s** median, according to the $ENRval$**s** smaller than the median.

RESULTS

We applied some of the most used pre-processing and gene selection methods to the RIP-chip data to identify postsplicing mRNA – U2AF[65] complexes in Hella cells (Gama-Carvalho et al., 2006).

Pre-processing Analysis. Following the usual approach for Genechip Affymetrix data analysis, were applied the following methods for normalization: dChip, RMA and GC-RMA; whose boxplots are showed in the Figure 2.

Although pre–processing methods produced different data distributions, it is clear that dChip and RMA performed similarly while GC-RMA method performance was significantly different. These differences may lead to different results in the posterior enrichment gene selection.

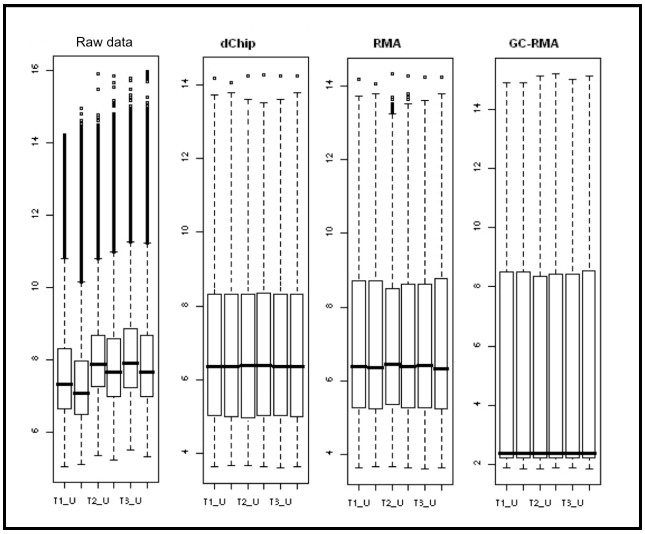

Figure 2. Probe value distribution in the six arrays samples before (Raw data) and after dChip, RMA and GC-RMA normalization.

Enriched Gene Selection: For comparison reasons we used the data obtained from RMA method, one of the most popular pre-processing methods due to its accuracy and performance in genechip Affymetrix microrarray data analysis. Table 1 shows the results of Limma: eBayes and Decide Tests (Smyth, 2004)), dChip (Li and Wong, 2003) and Rank Products (Hong et al. 2006). These three gene selection methods were applied using similar setups and represent some of the most used actually. As usual there were differences in the number of enrichment gene selected, being Rank products the most conservative, selecting only 449 genes as enriched genes when more than 3000 were expected (Gama-Carvalho et al., 2006).

There were some differences in the enriched genes number among different gene selection methods, mainly due to their conservative degree and accuracy. But, as Figure 3 shows, there were strong differences not only in the number of genes selected but also in the identification of those genes.

Methods	Genes
Limma :eBayes (eB)	1396
DChip - Fold Change (dC)	1884
Rank Porduct (RP)	449

Table 1. Enriched genes select by Limma: eBayes, Dchip an Rank Products methods implemented in Bioconductor in R.

Only 223 enriched genes were selected simultaneously by the three methods and within 27,65% and 44,64% of the genes were selected only by one specific method. It is important to keep in mind that the common pre-processing methods work correctly when there are few genes expressed differentially between different samples conditions and not, like in this case, differences at the level of 20% or more.

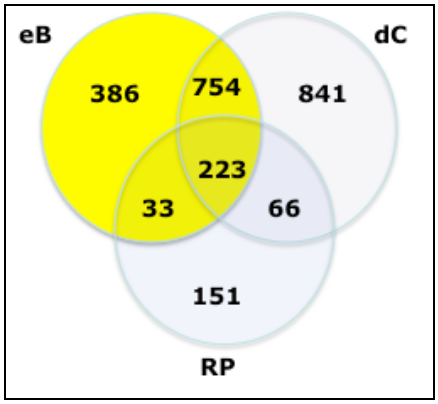

Figure 3. Benn diagram showing enriched genes selected by Limma: eBayes (eB), Dchip (dC) and Rank Products (RP) methods implemented in Bioconductor in R.

New alternative methodology. Attending the technical similarities between RIP-chip and ChIP-chip and the pre-processing methods limitations, we implemented a methodology in which the data of each array is used for its own pre–processing allowing its posterior comparison with the other arrays.

Background correction: The linear Sequence-Specific affinity model for background correction (1) was applied to each array in the U2AF experiment data from Gama-Carvalho et al. (2006). This model, inspired in Naef and Magnasco (2003) and Johnson et al. (2006), fits binding affinities to the sequence composition by examining the PM signal intensity, the contribution of each nucleotide in each sequence position and the effect of ATCG nucleotides count. This model accounted for 3 to 4% of the variation in the arrays (based on the multiple R^2 of the model).

In order to improve model fitting, a modification in equation (1) was introduced in which the contribution of each nucleotide in each sequence position to the binding affinity (2) was replaced by contribution of each nucleotide in each sequence position to the binding affinity plus its interaction with the neighbors nucleotides (3). Table 2 shows the results after applying this model with different setups: 2, 3 and 4 interacting nucleotides in comparison with the original no interaction model. We obtained better fitting results when the number of interacting nucleotides is increased (more than 58% for 4 interacting nucleotides), however increasing the number of interacting nucleotides turns out to be more computing demanding in terms of memory and time. This is the reason why fitting for 3 and 4 nucleotides was only possible using a random sample of 20000 and 8000 probes, respectively, from the each array data set. When the number of probes in the random sample is much smaller than the total number of probes in

the array, is introduced certain variability in the parameters estimation, especially when the sample is has less that 10000 probes.

Model	Multiple R-Squared	p-value
No interaction	0.02681	< 2.2e-16
2 Nucleotides interaction	0.06918	< 2.2e-16
3 Nucleotides interaction	0.124	< 2.2e-16
4 Nucelotides interaction : 8K sample	0.5861	7.48E-11

Table 2. Multiple R^2 obtained from the Input 1 sample fitted models: No interaction equation 1 and with nucleotide interaction equation 3 (2, 3 and 4 interacting nucleotides).

As the idea behind this process was to model the non-specific hybridization in terms of probe sequence, we implemented an iterative process for removing outliers after fitting the 3 nucleotides interaction model, using an initial sample of 20000 probes. The process stops when there are not significant differences between the last fitted model and the previous one which includes the removed outliers ($R_i^2 - R_{i-1}^2 < 0{,}001$, where i is the iteration number).
After the iterative process, as shows the table 3, the model accounted for 42 to 58% of the variation in the arrays, removing 18 to 28% probes in the probe sample. Mainly, this process removes high intensity values, which correspond to the specific hybridized probes, and increases the proportion of non-specific hybridized probes in the sample data, making the final fitted model more specific to the non-specific hybridization.

Array	Itera	R.squared	Outliers
Input1	18	0.421259224	3950
Input2	12	0.482873124	3634
Input3	16	0.48839486	4286
IP1	15	0.437264942	4500
IP2	18	0.41968973	3680
IP3	21	0.587910312	5631

Table 3. Iterative outlier remove process using 3 nucleotides interaction linear model fitting to the 6 arrays in Carvalho et al. (2006) U2AF RIP-Chip experiment.

The residuals were approximately normally distributed as shown by the residual Q-Q plot of the Input 1 sample (Figure 4).

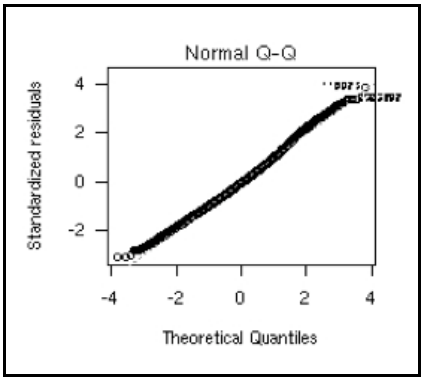

Figure 4. Q-Q plot of the residuals after iteration of 3 nucleotides interaction model fitted to Input sample 1.

Probe Standardization. After predicting each probe intensity using the specific fitted model for each array, the probe standardization was made in each array. Following Johnson et al. (2006) suggestion, all probes on an array were divided into bins containing ~3000 probes predicted to have similar intensities. Each probe's variance was estimated from the sample variance of all probes in its specific bin. Figure 2 shows that normalization is evidently required, however after probes standardization (Figure 5), the t values calculated using the equation (4) don't require further normalization and could be compared directly.

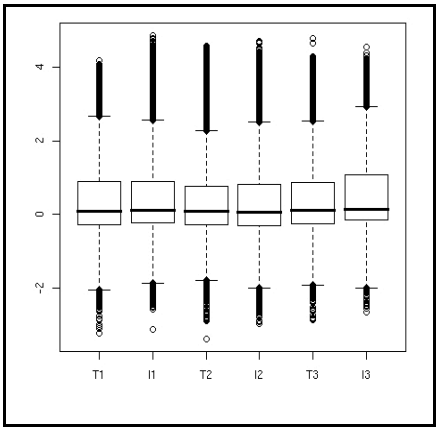

Figure 5. t values distribution of U2AF RIP-Chip experiment (Carvalho et al., 2006).

Enriched Genes. Finally, Table 4 shows the number of genes selected using our enrichment values (***ENRval)*** at different p-values levels applied to the standardized probes.

Enriched Genes	P value
762	< 0.001
1622	< 0.01
3200	< 0.05
4584	< 0.1

Table 4. Enriched genes selected at different p-values level. Enrichment values calculated from the standardized probes.

It is important to remark that the number of enriched genes selected with a p-value less than 0.05 is approximately close to the expected to be associated with *U2AF* (3200) sharing a high percentage (66%) of the genes reported in Gama-Carvalho et al. (2006). It is possible that some of those differences were due to the application of normalization methods without having into account the differences in the data distribution between samples. ***ENRval*** allows a consistent enriched gene selection.

Remarks

Common pre-processing and gene selection methods applied to RIP-Chip data may produce different results, maybe because of the high amount of differentially expressed genes between experimental conditions.
The method proposed standardizes the probe intensity level using the individual information of each array, making possible to compare arrays from different conditions in a more adequate way for RIP-Chip experiments.
A future direction will be to improve the Sequence-Specific probe affinity model, making it more accurate and less computing power demanding.

Develop more accurate non parametric gene selection methods using the ***t*** probes values.

REFERENCES

GAMA-CARVALHO, Margarida; CARVALHO, Marcos P.; KEHLENBACH, Angelika; VALCARCEL, Juan and CARMO-FONSECA, Maria. Nucleocytoplasmic shuttling of heterodimeric splicing factor U2AF. *Journal of Biological Chemistry,* April 2001, vol. 276, no. 16, p. 13104-13112.

GAMA-CARVALHO, Margarida; BARBOSA-MORAIS, Nuno L.; BRODSKY, Alexander S.; SILVER, Pamela A. and CARMO-FONSECA, Maria. Genome-wide identification of functionally distinct subsets of cellular mRNAs associated with two nucleocytoplasmic-shuttling mammalian splicing factors. *Genome Biology,* November 2006, vol. 7, no. 11, p. R113.

GENTLEMAN, Robert C.; CAREY, Vincent J.; BATES, Douglas M.; BOLSTAD, Ben: DETTLING, Marcel; DUDOIT, Sandrine; ELLIS, Byron; GAUTIER, Laurent; GE, Yongchao; GENTRY, Jeff; HORNIK, Kurt; HOTHORN, Torsten; HUBER, Wolfgang; IACUS, Stefano; IRIZARRY, Rafael; LEISCH, Friedrich; LI, Cheng; MAECHLER, Martin; ROSSINI, Anthony J.; SAWITZKI, Gunther; SMITH, Colin; SMYTH, Gordon; TIERNEY, Luke; YANG, Jean Y. H. and ZHANG, Jianhua. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, September 2004, vol. 5, no. 10, p. R80.

GERBER, André P.; HERSCHLAG, Daniel and BROWN, Patrick O. Extensive association of functionally and cytotopically related mRNAs with Puf family RNA-binding proteins in yeast. *PLoS Biology*, March 2004, vol. 2, no.3, p. e79.

GIBBONS, Francis D.; PROFT, Markus; STRUHL, Kevin and ROTH, Frederick P. Chipper: discovering transcription-factor targets from chromatin immunoprecipitation microarrays using variance stabilization. *Genome Biology*, November 2005, vol. 6, no. 11, p. R96.

HANLON, Sean E. and LIEB, Jason D. Progress and challenges in profiling the dynamics of chromatin and transcription factor binding with DNA microarrays. *Current Opinion in Genetics & Development*, December 2004, vol. 14, no.6, p. 697–705.

HONG, Fangxin; BREITLING, Rainer; MCENTEE, Connor W.; WITTNER, Ben S.; NEMHAUSER, Jennifer L. and CHORY, Joanne. RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics*, November 2006, vol. 22, no. 22, p. 2825-2827.

IRIZARRY, Rafael A.; HOBBS, Bridget C.; BEAZER-BARCLAY, Yasmin D.; ANTONELLIS, Kristen J.; SCHERF, Uwe and SPEED, Terence P. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, April 2003, vol. 4, no. 2, p. 249-264.

JOHNSON, Evan W.; LI, Wei; MEYER, Clifford A.; GOTTARDO, Raphael; CARROLL, Jason S.; BROWN, Myles and LIU, Shirley. Model-based analysis of tiling-arrays for ChIP-chip. *Proceedings of the National Academy of Science of the United States of America*, August 2006, vol. 103, no. 33, p. 12457-12462.

KADONAGA, James T. Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors. *Cell*, January 2004, vol. 116, no. 2, 247–257.

LI, Cheng and WONG, Wing H. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proceedings of the National Academy of Science of the United States of America*, January 2001, vol. 98, no. 1, p. 31-36.

LI, Cheng and WONG, Wing H. DNA-Chip Analyzer (dChip). In: Parmigiani, G., Garrett, E. S., Irizarry, R. and Zeger S. L. eds. *The analysis of gene expression data: methods and software*. New York, Springer, 2003, p. 120-141.

LOCKHART, David and WINZELER, Elizabeth A. Genomics, gene expression and DNA arrays. *Nature*, June 2000, vol. 405, p. 827–836.

MATA, Juan; MARGUERAT, Samuel and BAHLER, Jurg. Post-transcriptional control of gene expression: a genome-wide perspectiva. *Trend in Biochemical Sciences*, September 2005, vol. 30, no. 9, p. 506-514.

NAEF, Felix and MAGNASCO Marcelo O. Solving the riddle of the bright mismatches: Labeling and effective binding in oligonucleotide arrays. *Physical Review*, July 2003, vol. 68, p. 011906.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, R Foundation for Statistical Computing, 2008.

SMYTH, Gordon K. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, February 2004, vol. 3, no. 1, Article 3.

TOMARI, Yukihide and ZAMORE, Phillip D. Perspective: machines for RNAi. *Genes & Development*, March 2005, vol. 19, no. 5, p. 517–529.

WU, Zhijin; IRIZARRY, Rafael A.; GENTLEMAN, Robert C.; MARTINEZ-MURILLO, Francisco and SPENCER, Forrest. A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. *Journal of the American Statistical Association*, December 2004, vol. 99, no. 468, p. 909-917.

ZAMORE, Phillip D. and GREEN, Michael R. Identification, purification, and biochemical characterization of U2 small nuclear ribonucleoprotein auxiliary factor. *Proceedings of the National Academy of Science of the United States of America*, December 1989, vol. 86, no. 23, p. 9243-9247.

ZAMORE, Phillip D.; PATTON, James G. and GREEN, Michael R. Cloning and domain structure of the mammalian splicing factor U2AF. *Nature*, February 1992, vol. 355, p. 609-614.