UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

STATISTICAL METHODOLOGIES FOR THE ANALYSIS AND NORMALIZATION OF RIP-CHIP DATA

Emiliano Barreto Hernández

DOUTORAMENTO EM ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

(Especialidade de Bioestatística e Bioinformática)

2011

UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

STATISTICAL METHODOLOGIES FOR THE ANALYSIS AND NORMALIZATION OF RIP-CHIP DATA

Emiliano Barreto Hernández

Tese orientada pela Professora Doutora Lisete Maria Ribeiro de Sousa

DOUTORAMENTO EM ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

(Especialidade de Bioestatística e Bioinformática)

2011

To my loved girls: *Natalía, Ana María and Liliana*

"Only the knowledge that makes us better is useful"

Sócrates

Acknowledgements

I am grateful to dedicate this page to those who gave me all their support, confidence and believe in me. In particular, I would like to thank to:

- My supervisor, Professora Doutora Lisete Maria Ribeiro de Sousa, for her valuable support and knowledge and the opportunity to work with her. I am deeply grateful for her patience, dedication and her friendship.
- Professora Doutora Margarida Gama Carvalho from the Faculty of Sciences of the University of Lisbon, for being always available for answering my questions about the biological concepts involved with her RIP-Chip experiment.
- Department of Statistics and Operational Research and the Center of Statistics and Applications of the Faculty of Sciences, for providing the conditions that enabled the accomplishment of this thesis.
- Biotechnology Institute of the National University of Colombia, for the support and the resources that allowed me the achievement of this thesis.
- Bioinformatician Pedro Fernandes from the Gulbenkian Institute of Science, for his support and friendship.
- Portuguese Foundation for Science and Technology (FCT) that partially supported this work through national funds under the projects PEst-OE/MAT/UI0006/2011 and PTDC/MAT/64353/2006.
- EU ALBAN program for supporting me with a Fellowship (ALBAN fellowship No. E06D101266CO).
- My friends for giving me the strength and impulse needed to accomplish this goal.
- My family, without their love and sacrifice it would have been impossible to make this work.

Resumo

Nos últimos anos foram desenvolvidas técnicas de alto rendimento na investigação em biologia. Essas técnicas evoluíram fornecendo à comunidade científica instrumentos como: sequenciadores de alta capacidade, que permitem obter milhões de fragmentos de DNA ao mesmo tempo; espectómetros de massa em tandem que permitem a identificação de proteínas ou proteomas completos; ou hibridação de *microarrays*, usados para determinar a expressão dos genes através da identificação de mRNAs presentes na célula num momento específico.

Os *microarrays* constituem uma técnica usada para quantificar a expressão de genes e analisar fragmentos de genes, proteínas ou metabolitos. Também têm sido utilizados para clarificar elementos específicos do Dogma Central da Biologia Molecular, envolvidos no controle da transcrição; na busca de dados que expliquem como a expressão do gene começa a partir do DNA; ou como o mRNA em associação com os ribossomas é traduzido em proteínas no citoplasma da célula.

Dado o enquadramento biológico descrito acima, o Capítulo 1 introduz os aspectos da biologia relacionados com os dados RIP-Chip utilizados nesta tese, dados esses obtidos por Gama-Carvalho et al. [2006], em que se pretende identificar os mRNAs associados a PTB e U2AF^{65} em condições nativas. Estas duas proteínas de ligação a RNA fazem parte do controle pós-transcripcional da expressção genética em células eucariótas.

Este capítulo começa por introduzir conceitos de biologia molecular da célula tal como o dogma central da biologia molecular, onde os processos de transcrição e tradução são essenciais para manter a vida da célula e onde o controle de expressão genética é um aparelho fundamental na regulação da célula. Como parte do controle da expressão dos genes, o Capítulo 1 apresenta uma visão geral do controle pré- e pós-transcripcional da expressão dos genes. O *splicing* de pré-mRNAs é um passo essencial no controle da expressão pós-transcripcional dos genes e envolve factores de *splicing* tais como as proteínas PTB e U2AF^{65}, sendo U2AF^{65} exportada para o citoplasma e envolvida em outras funções celulares.

O Capítulo 1 mostra como foram obtidos os dados RIP-Chip das proteínas PTB e U2AF^{65} e apresenta uma breve descrição da metodologia utilizada por Gama-Carvalho et al. [2006] na sua experiência RIP-Chip. Mostra como a identificação de mRNAs associados a PTB e U2AF^{65}, em condições nativas, foi realizada por imunoprecipitação (IP) após a adição de um anticorpo monoclonal específico (Bb7 anti-PTB mAb ou anti-U2AF^{65} MC3), seguido de extração de RNA, poliadenilação, transcrição reversa, etiquetagem final e amplificação por PCR. Os cDNAs gerados foram hibridados com o GeneChip Affymetrix Human Genome U133 Plus 2.0 [Gama-Carvalho et al., 2006].

Este capítulo apresenta uma descrição da tecnologia de *microarrays*, em particular as características dos *microarrays* da Affymetrix utilizados na experiência RIP-Chip executada por Gama-Carvalho et al. [2006].

De seguida, o Capítulo 2 apresenta alguns dos métodos mais comuns de análise de dados de *microarrays* e os resultados de seu desempenho nos dados de Gama-Carvalho et al. [2006].

Para a correção de *background* foi utilizado o modelo linear robusto (RMA) de Irizarry et al. [2003a] e uma modificação do mesmo (GCRMA) proposta por Wu et al. [2004], apenas sobre PM (*Perfect Match*). A normalização foi realizada através da normalização quartílica e a sumariação das sondas foi feita usando a mediana *polish* [Irizarry et al., 2003a]. Alternativamente, os dados foram pré-processados usando o programa dChip: apenas para PM; usando o método de normalização *invariant set* [Li and Wong, 2001]; e o método baseado em modelos de Li and Wong [2001] para calcular os níveis de expressão.

Para efeitos de comparação foram utilizados os dados obtidos após a correção de *background* com RMA, a normalização quartílica e sumariação com a mediana *polish*. Com base nestes dados, foi feita a seleção de genes enriquecidos usando as seguintes bibliotecas do BioConductor: `limma` (ajusta um modelo linear para cada gene); `eBayes` (calcula a estatísticas T moderada, F e B - logaritmo das chances *a posteriori*); `decideTests` com um valor-p < 0.05 (baseia-se em testes múltiplos para determinar se cada estatística numa matriz de estatísticas T deve ser considerada significativamente diferente de zero [Smyth, 2004]); `RankProd` com FDR < 0.05 (teste não-paramétrico que deteta itens que são consistentemente classificados como estando no topo da lista [Breitling et al., 2004]). Estes resultados foram comparados com os resultados obtidos com o programa dChip considerando uma taxa de falsas

descobertas (FDR) < 0.05 e um valor-p < 0.05 [Li and Wong, 2003].

Os resultados apresentados no Capítulo 2 mostram como diferentes metodologias aplicadas aos dados de Gama-Carvalho et al. [2006] produziram resultados diferentes. Parte das diferenças devem-se sobretudo ao facto de mais de 20% dos mRNAs serem enriquecidos e os métodos de normalização comuns terem por base pequenas diferenças entre eles.

Como esta tese teve como principal objetivo o desenvolvimento de metodologias estatísticas para análise de baixo nível e seleção de genes enriquecidos em experiências RIP-Chip, o Capítulo 3 é dedicado a apresentar a implementação de um novo método de correção de *background* inspirado num método de hibridação não específica utilizado para pré-processamento de dados ChIp-Chip [Johnson et al., 2006]. Modelos de regressão linear foram usados para modelar em cada *microarray* a hibridação não específica, representando interações entre cada três nucleótidos consecutivos na sequência da sonda. As intensidades das sondas foram padronizadas usando sua intensidade prevista e a variância das sondas de intensidades previstas semelhantes. A nova abordagem aqui proposta utiliza a informação de cada *microarray* de forma independente, e os valores de intensidade padronizados não revelaram necessidade de normalização adicional. Assim, os *microarrays* podem ser directamente comparados [Barreto-Hernandez et al., 2011].

O Capítulo 3 apresenta também um *score* para a sonda; a definição de um valor de enriquecimento da sonda (ENRval) e respectivos valores-p para a seleção de genes enriquecidos [Barreto-Hernandez et al., 2011]. Os genes enriquecidos obtidos usando esta metodologia, tanto para os dados RIP-Chip de PTB como de $U2AF^{65}$, estão de acordo com os genes identificados experimentalmente por Gama-Carvalho et al. [2006].

Finalmente, o Capítulo 3 apresenta ao desenvolvimento da uma nova metodologia não-paramétrica baseada em postos (*ranks*), implementada para seleção de genes enriquecidos e aplicada aos *scores* propostos en este Capítulo. Esta metodologia tem em conta a variabilidade da intensidade padronizada em cada sonda, em vez de usar o valor de sumariação de cada sonda (ENRval).

Ainda neste capítulo, as metodologias desenvolvidas nesta tese para a seleção de genes enriquecidos são aplicadas aos dados da experiência Spike-In. Esta base de dados foi construída há alguns anos e é usada no desenvolvimento e comparação de métodos de análise de expressão diferencial de genes [Irizarry et al., 2003b]. A experiência Spike-In U133 engloba 42 transcritos adiciona-

dos a um complexo transcriptoma humano em concentrações que variam de 0.125pM a 512pM, correspondendo a 14 hibridações separadas com três repetições técnicas. Os transcritos foram incluídos na experiência sob a forma de um quadrado latino clássico [Irizarry et al., 2003b]. Para a análise comparativa, três diferentes hibridações Spike-In foram selecionadas (hibridações 1, 8 e 14) e usadas para simular diferenças de enriquecimento em experiências RIP-Chip através do seguinte procedimento: 1 como Controle e 8 como IP; 1 como IP e 14 como Controle.

As duas metodologias desenvolvidas nesta tese para seleção de genes enriquecidos, apresentam elevada exatidão quando aplicadas aos dados Spike-In U133.

Palavras-chave: *Microarrays*, RIP-Chip, Proteínas de associação a RNA, Bioinformática, pré-processamento.

Abstract

Pre-mRNA splicing is an essential step in the post-transcriptional gene expression control involving protein-splicing factors like PTB and U2AF^{65}; the last one is exported to the cytoplasm and involved in some other cellular functions. The identification of PTB- and U2AF^{65}-associated mRNAs under native conditions was performed by immunoprecipitation and hybridization on Chip (RIP-Chip) technology using the Affymetrix GeneChip® Human Genome U133 Plus 2.0.

The aim of this thesis is to develop statistical methodologies for low level analysis and enriched gene selection in RIP-Chip experiments. When the most common methodologies for quality assessment, low level analysis (background adjustment, normalization and summarization) and detection of differentially expressed genes (DEG), are applied to RIP-Chip data the obtained results differ. This probably happens because usually more than 20% of the mRNAs are enriched, while methods for normalization and identification of DEG are developed supposing that only a small proportion of genes (1% or 5%, say) express differently. Also, methods for detecting differentially expressed genes may not be the most adequate for gene enrichment selection.

In this thesis is implemented a background correction method inspired in a non-specific hybridization method used for pre-processing ChIP-Chip data. Linear regression models are used in each array to model the non-specific hybridization. Probe intensities on the array are standardized using their predicted intensity and the variance of similar predicted intensities. The standardized probe intensities showed no need for further normalization, so the scores could be directly compared. It is proposed a probe set score, a probe set enrichment value and its p-value for enriched gene selection. The genes selected using this new method are practically the same as the ones found experimentally. Additionally, a new methodology based on ranks is presented for enriched gene selection, being applied to the probe set scores proposed.

Both methodologies had high accuracy when applied to Spike-In U133 dataset, which is used to benchmark methodologies for analysing Affymetrix microarrays.

Keywords: Pre-processing RIP-Chip, RNA binding proteins, linear models,

bioinformatics.

Contents

List of Figures

List of Tables

Glossary

A	log intensities average of a certain probe between two arrays.
avgDiff	Average of the differences between PM and MM intensities
ChIP-Chip	Chromatin immunoprecipitation on chip
DEG	Differentially expressed genes
ENRval	Enriched value
FC	Fold change
FDR	False discovery rate
IM	Ideal Mismatch
IP	Immunoprecipitated sample
M	Difference of the log intensities of a certain probe between two arrays.
MAT	Model-based Analysis of Tiling-arrays
MM	Mismatch probe
mRNA	Messenger RNA
ncRNAs	non-coding RNAs
NSB	non-specific binding
NUSE	Normalized Unscaled Standard Error
PCR	Polymerase chain reaction
PM	Perfect mach probe
PSsco	Probe set score
PTB	Polypyrimidine-tract-binding protein
QA	Quality assessment
RBP	RNA binding proteins
RIP-Chip	Immunoprecipitation and hybridization on Chip
RLE	Relative Log Expression
RP	Rank products
rRNA	Ribosomal RNA
RT	reverse transcriptase
RT-PCR	Real time PCR
TCA	Trichloroacetic acid
TP	Target protein
tRNA	Transfer RNA
U2AF$^{6}5$	Heterodimeric splicing factor U2 snRNP auxiliary factor protein
VSN	Variance stabilization and normalization

Preface

In the past few years, the development of the high throughput techniques in biology research have given to the scientific community several instruments like: the high capacity sequencers which can get hundreds-thousands DNA fragments; tandem mass spectrometers that allow identification of cell proteins or proteome; microarrays hybridization and reading, used to determine gene expression through the identification of mRNA present in the cell at a moment.

Microarrays constitute a technique used to quantify the difference between gene expressions and to analyze their fragments as well as proteins or metabolites. They have also been used to clarify specific elements of Central Dogma of Molecular Biology, involved in transcription control, looking for massive data on how gene expression begins from DNA or how the kind and amount of mRNAs is modulated before it reaches the ribosome to start protein synthesis.

Given the biological framework described above, Chapter 1 introduces the biological aspects related to RIP-Chip data used in this thesis, obtained by Gama-Carvalho et al. [2006], which aims at looking for the Identification of PTB and U2AF^{65}-associated mRNAs under native conditions. These two RNA binding proteins are part of post-transcriptional gene expression control in eukaryotic cells. This chapter starts with cell molecular biology concepts like the Central dogma of the molecular biology, where transcription and translation process are essential for maintaining the cell's life and where the gene expression control is a fundamental cell regulatory apparatus. As part of the gene expression control, Chapter 1 presents an overview of the pre- and post-transcriptional gene expression control. The pre-mRNA splicing is presented as an essential step in the post-transcriptional gene expression control that involved protein-splicing factors like PTB and U2AF^{65}, which are the last to be exported to the cytoplasm and also implicated in additional cellular functions.

Chapter 1 presents how PTB and U2AF^{65} RIP-Chip data were obtained, giving a short description of the methodology used by Gama-Carvalho et al. [2006] in their RIP-Chip experiment. In this chapter is also shown how the identification of PTB and U2AF^{65}-associated mRNAs under native conditions was performed by immunoprecipitation (IP) after the addition of a specific monoclonal antibody (Bb7 anti-PTB mAb or anti-U2AF^{65} MC3), followed by an RNA extraction, polyadenilation, reverse transcription, end tagging and PCR amplification. The resulting cDNAs were hybridized on the Affymetrix GeneChip Human Genome U133 Plus 2.0. [Gama-Carvalho et al., 2006]

Besides, this chapter presents the microarray technology description, in particular the characteristics of the Affymetrix microarrays used in the RIP-Chip experiment run by Gama-Carvalho et al. [2006].

Chapter 2 presents the common microarray data analysis methods and the results of their performance on Gama-Carvalho et al. [2006] data.

Background correction was performed only on Perfect Match (PM), on raw intensity scale, using a robust linear model (RMA) [Irizarry et al., 2003a] and its modification (GCRMA) [Wu et al., 2004]. Normalization was performed considering quartile normalization and the probe set summarization using median polish [Irizarry et al., 2003a]. Alternatively, the data were pre-processed using dChip: only for PM; using invariant set normalization method [Li and Wong, 2001]; and the model-based method on Li and Wong [2001] for computing expression values.

For comparison reasons, pre-processed data using RMA is used to select enriched genes through the following Bioconductor libraries: `limma` (fits a linear model for every gene); `eBayes` (computes moderated t-statistics, F-statistic, and posterior log-odds B); `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]); `RankProd` with FDR < 0.05 (non-parametric test that detects items that are consistently highly ranked in a number of lists [Breitling et al., 2004]). The results were compared to the results obtained using the dChip program considering a false the discovery rate (FDR) < 0.05 and a p-value < 0.05) [Li and Wong, 2003].

The results presented in this chapter showed that the application of different

methodologies on the Gama-Carvalho et al. [2006] data, produced different results. Basically, part of the differences should be mainly because more than 20% of the mRNAs detected are differently enriched while the common normalization methods are based on small differences between them.

Knowing that the purpose of this thesis was to develop statistical methodologies for low-level analysis and enriched gene selection in RIP-Chip experiments, Chapter 3 is dedicated to present the implementation of a new background correction method, inspired in a non-specific hybridization method used for pre-processing ChIP-Chip data [Johnson et al., 2006]. Linear regression models were used to model, in each array, the non-specific hybridization, accounting for interactions between each three consecutive nucleotides into the probe sequence. Probe intensities on the array were standardized using their predicted intensity and the probes' variance for similar predicted intensities. The new approach proposed here uses the information for each array in the experiment independently, and the standardized probe intensity values showed no need for further normalization. Thus, the arrays could be directly compared between them [Barreto-Hernandez et al., 2011].

Chapter 3 also presents a probe set score; a probe set enrichment value (ENRval) and its respective p-value for enriched gene selection. The enriched genes obtained using the methodology implemented in this work for both PTB and U2AF^{65} matched with the experimental information [Barreto-Hernandez et al., 2011].

Finally, Chapter 3 is dedicated to the development of a new non-parametric methodology based on ranks, implemented for enriched gene selection on the probe set scores proposed on Chapter 3. It takes into account the standardized probe intensity variability in each probe set, instead of using a summarization value of each probe set (ENRval).

This chapter also shows how the methodologies developed for enriched gene selection in this thesis were benchmarked using the data of U133 Spike-In experiment. This database was created a few years ago and is used in the development and comparison of differential gene expression analysis methods [Irizarry et al., 2003b]. U133 Spike-In experiment comprises 42 spike transcripts in a complex human background at concentrations ranging from 0.125pM to 512pM, evolving fourteen separate hybridizations with three technical replicates. The spike transcripts were put in a classical latin square experiment [Irizarry et al., 2003b]. For benchmarking, three different Spike-In experiment hybridizations were selected (hybridizations 1, 8 and 14) and

used to simulate RIP-Chip enrichment differences, through the following procedure: 1 as Control and 8 as IP; 1 as IP and 14 as Control.

Both methodologies for enriched gene selection show high accuracy when they were applied to Spike-In U133 data set [Irizarry et al., 2003b].

Chapter 1

Biological concepts related to RIP-Chip data analysis

1.1 Introduction

In recent years, knowledge regarding molecular function of cells has raised exponentially, thanks to the development of, so-called, high throughput techniques. These techniques evolved from the needs that Human Genome Project created; providing to the scientific community with the new high capacity sequencers which can get hundred-thousands DNA fragments; tandem mass spectrometers that allow identification of cell proteins or proteome; or microarrays hybridization, used to determine gene expression through the identification of all mRNAs present in the cell at a specific moment.

This has allowed to go from just identifying some DNA fragments, RNAs or proteins, to the evaluation of all the molecules present at a specific cellular time under a particular conditions such as a normal or a pathologic state.

Especially, microarray is a technique used to quantify the difference between gene expressions and to analyze their fragments as well as proteins or metabolites. This technic has also been used to clarify specific elements of Central Dogma of Molecular Biology involved in the transcription control. Microarray techniques like ChIP-Chip or RIP-Chip have been applied, generating a massive amount of data providing insight on how gene expression is regulated from DNA or how the kind and amount of mRNAs are modulated before they reach the ribosome to start protein synthesis.

Given the biological framework described above, this chapter will introduce

the Biology concepts related to RIP-Chip data analysis.

1.2 DNA structure

The DNA molecule, as the primary repository of genetic information in living systems, is constrained to be stable and predictably structured. Since Watson and Crick [1953] proposed that the human Desoxyribonucleic Acid (DNA) is a double-stranded helix, where each strand is formed in a combination of 4 different units call nucleotides (a unit of phosphate, desoxyribose and purine or pirimidine, Figure 1.1): A (adenine), C (cytosine), T (thymine) and G (guanine); the knowledge and the analytic techniques related to the cell life cycle have advanced enormously specially with the recently introduction of the high throughput techniques for genomic, proteomic, etc.

The double helix of DNA has two polynucleotide strands wound around each other, assembled in the 5' to 3' direction by convention. The phosphate group bonded to the 5' carbon atom of one deoxyribose is covalently bonded to the 3' carbon of the next. The purine or pyrimidine attached to each deoxyribose is projected toward the axis of the helix and each base forms hydrogen bonds with the one directly opposite it, forming base pairs.

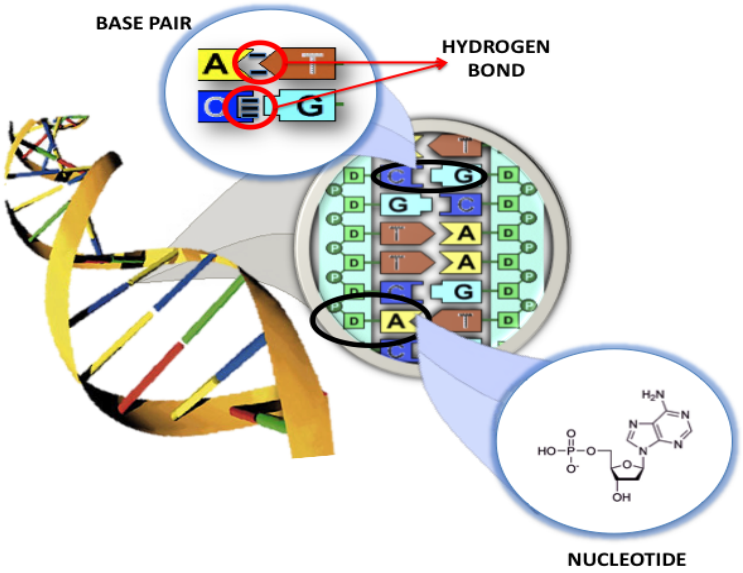


Figure 1.1: The double helix structure of DNA.

1.3 Central dogma of the molecular biology

Few years later Crick [1970] proposed the protein synthesis and the model known as the central dogma of the molecular biology (see Figure 1.2), and from that moment on many researchers around the world were able to study the mechanisms used by the cell for controlling the proteins production - molecules responsible for most of cell functions.

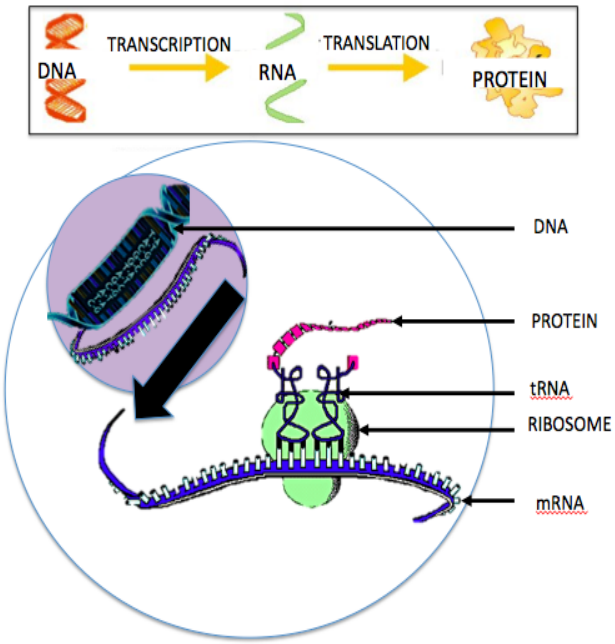


Figure 1.2: Central dogma of the molecular biology.

The Central Dogma as originally formulated by Crick, is a negative hypothesis, which states that information cannot flow downwards from the Protein to the DNA. Its complement, the Sequence Hypothesis is often conflated with the Central Dogma. Under it, DNA is transcribed to Ribonucleic Acid (RNA) (Transcription), and RNA is translated into protein (Translation). More abstractly, information flows upward from DNA, to RNA, to proteins, and, by extension, to the cell, and, finally, to multicellular systems. [Werner, 2005].

Basically, the central dogma of the molecular biology says that a gene does not participate directly in the protein synthesis; in eukaryotes, the DNA is enclosed inside of nuclear membrane while the protein-synthesizing machin-

ery is outside the structure, in the cytoplasm, and the two never meet. DNA sends copies of its encoded information to the cytoplasm, by the mRNA.

The protein synthesis is a two-step process: (1) Transcription, (2) Translation; where the proteins are synthesized as a polypeptide chain of amino acids that carry out the genetic instructions encoded in an organism's DNA.

1.3.1 Transcription

Transcription is the process, of synthesized RNA using DNA as a template, takes place in the nucleus. The synthesized RNA is a single-stranded polynucleotide where T is replaced by U (uracile), there are the same $3'$ to $5'$ phosphodiester bonds between successive nucleotides, and the sugar is a ribose (-OH on the $2'$ position, makes the molecule unstable relative to DNA).

Five general steps may describe this process (Figure 1.3):

1. RNA polymerase binds to a DNA region called the promoter (in eukaryotes, three of such polymerases exist) and then the two DNA strands are transitorily separated and placed at the site of RNA synthesis;

2. RNA polymerase adds bases (ATP, CTP, GTP and UTP) that pair with the DNA. The synthesis requires that the duplex DNA strands are separated providing a single-stranded template for directing the sequence of nucleotides to be assembled into RNA, and then come together again after the polymerase has passed;

3. RNA polymerase reaches a termination signal and the transcription stops;

4. The new RNA strand is released from the DNA template and

5. the mRNA leaves the nucleus and enters to the cytosol for completing the protein synthesis [Fuda et al., 2009].

Transcription produces different types of RNAs that in general may be divided in coding a non-coding RNAs, the last one is involved in the gene expression's regulation.

- **Messenger RNA (mRNA)** carries the genetic information coded in the DNA from the nucleus to the cytosol. This information corresponds to the primary amino acid sequence of the protein to be synthesized,

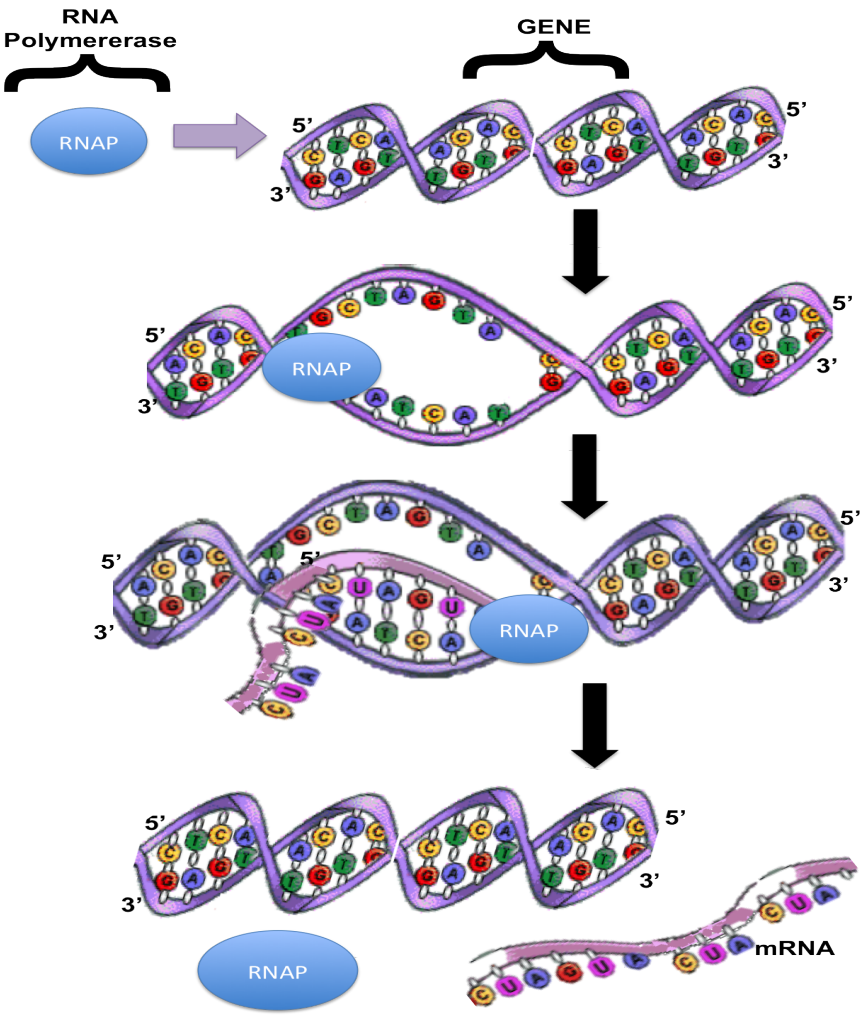


Figure 1.3: Transcription of DNA

transcribed from a specific genome region called gene.

Many eukaryotes genes have a large number of nucleotides, much more than the corresponding complementary mRNA. After a careful experimental analysis, it was possible to conclude that two types of regions constitute most of the genes in eukaryotes cells. The first type of region is the one that codify the protein and are and the second type is a DNA region that do not codify proteins, which are denominate 'introns'. The mature mRNA, is constituted only by exons, and for that reason is always smaller than the gene, concerning the chromosomal DNA, as it also takes the introns.

- **Transfer RNA (tRNA)** is a hairpin shape RNA that reads the code in the mRNA and carries the amino acid that will be used in the protein synthesis in the translation process.

- **Ribosomal RNA (rRNA)** is the catalytic component of a nucleoprotein called ribosome. The ribosome serves as the location where the protein synthesis takes place and it also carries the necessary enzymes for completing the process.

- **Regulatory RNAs** are part of gene expression down-regulation by being complementary to a gene's DNA or a part of a mRNA. For example the RNA interference (RNAi) is a system in the cell that gets involved in the control or regulation of the gene. Telling it in which moment it needs to be activated and how strong the activation should be. As part of this system the MicroRNAs (miRNA; 21-22 nt; found in eukaryotes) block the mRNA from being translated, or accelerate its degradation through an effector enzymatic complex responsible of breaking down mRNA which the miRNA is complementary to [Matzke and Matzke, 2004, Wu and Belasco, 2008]. Small interfering RNAs (siRNA; 20-25 nt) are considered a RNAi that works in a similar way to miRNAs, often produced by breakdown of viral RNA, and also by endogenous sources [Vazquez et al., 2004, Watanabe et al., 2008]. Some researchers have found that miRNAs and siRNAs can cause methylation of the genes they target, decreasing or increasing their transcription [Sontheimer and Carthew, 2005, Pushparaj et al., 2008, Doran, 2007].

Some other types of regulatory RNAs have been reported, like CRISPR RNAs, a regulatory system similar to RNA interference, found in prokaryotes [Horvath and Barrangou, 2010], or the Piwi-interacting RNAs (piRNA; 29-30 nt, found in animals) active in germline cells which possible function is thought to be a defense against transposons and play a role in gametogenesis [Horwich et al., 2007, Girard et al., 2006]. Also antisense RNAs are reported to be widespread, few of them as a transcription activators [Wagner et al., 2002] and the others down-regulate a gene, in some cases by binding to a mRNA and forming double-stranded RNA that is enzymatically degraded [Gilbert, 2010]. Long non-coding RNAs have been reported too, which principal function is gene regulation in eukaryotes. [Amaral and Mattick, 2008, Heard et al., 1999].

1.3.2 Translation

Translation is the process of converting mRNA into proteins. It takes place in the cytosol on top of a ribosome. Cytosol contains amino acids, tRNA, and ribosomes, which are needed for protein synthesis.

tRNA is essential to the translation process because each tRNA molecule carries an amino acid to the ribosome. tRNA recognizes the code and binds itself to the mRNA molecule by an anti-codon, a tRNA region that has three complementary bases which function is recognize and bound to a specific mRNA codon. In the step 1 of the translation, the starting codon (AUG) is recognized by the tRNA and the tRNA binds to the mRNA and brings an amino acid (methionine) with it.

The small ribosomal subunit in eukaryotes (40 S) is responsible of controlling base-pairing between the tRNA anticodon and each mRNA codon during protein synthesis. In order to be able to scan successfully from the $5'$ cap to the start codon, the small ribosomal subunit needs to be able to overcome intramolecular secondary structures that would, otherwise, block progress along the RNA. Detailed studies in mammalian and fungal systems have demonstrated that the inhibitory influence of stem-loop structures in the $5'$ untranslated regions ($5'$ UTRs) of mRNAs is related to the stability of these structures. RNA helicases are proteins that can disrupt stemloop structures and hence, there is the expectation that proteins of this type should be capable of facilitating 40 S binding and scanning. Most RNA helicases contain variations of a DEAD box motif along with six other highly conserved motif.

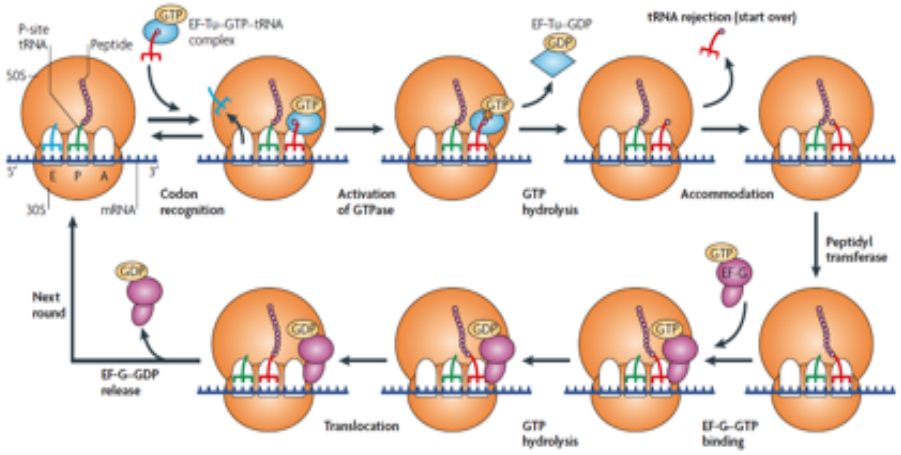


Figure 1.4: An overview of ribosomal structure and mRNA translation. [Source Steitz, 2008]

mRNA translation (Figure 1.4) is initiated with the binding of tRNAfmet to the P site. An incoming tRNA is delivered to the A site in complex with elongation factor $(EF) - Tu–GTP$. Correct codon-anticodon pairing activates the GTPase centre of the ribosome, which causes hydrolysis of GTP and release of the aminoacyl end of the tRNA from $EF - Tu$. Binding of tRNA also induces conformational changes in ribosomal (r)RNA that optimally orientates the peptidyl-tRNA and aminoacyl-tRNA for the peptidyl-transferase reaction to occur, which involves the transfer of the peptide chain onto the A site tRNA. The ribosome must then shift in the $3'$ mRNA direction so that it can decode the next mRNA codon. Translocation of the tRNAs and mRNA is facilitated by binding of the GTPase-EF-G, which causes the deacylated tRNA at the P site to move to the E site, and the peptidyl-tRNA at the A site to move to the P site upon GTP hydrolysis. The ribosome is then ready for the next round of elongation. The deacylated tRNA in the E site is released so that it can be bound the nex aminoacyl-tRNA to the A site. Elongation ends when a stop codon is reached, which initiates the termination reaction that releases the polypeptide [Steitz, 2008].

1.4 Central dogma of molecular biology Exceptions.

Although RNAs have been considered for a long time as informative macromolecules that carry messages from the genome to the proteome, they are becoming increasingly appreciated as regulatory macromolecules. [Lioliou et al., 2010].

In 1970, Temin's stunning finding of reverse transcriptase (RT) [Temin and Mizutani, 1970] revolutionized our understanding of genome function, because it showed for first time the existence of a second flow of information, from RNA to DNA, in sharp contrast with the central dogma that admitted DNA to RNA as the only possible informational direction.

Exceptions to the central dogma include reverse transcriptases, enzymes that synthesize DNA from RNA [Sciamanna et al., 2009] . Moreover, DNA is not static and does more than serves as an inert source of information. Rather, DNA is dynamic and modifiable. DNA modifications contribute to gene transcription regulation, such as chromatin organization and remodeling, genomic imprinting, DNA methylation, and other mechanisms. [Robinson, 2009].

Recent transcriptome analysis and different experimental approaches have identified a surprisingly large number of non-coding RNAs (ncRNAs) in eukaryotic cells. ncRNAs comprise microRNAs, anti-sense transcripts and other Transcriptional Units containing a high density of stop codons and lacking any extensive *Open Reading Frame*. As Central Dogma exception they have been shown to regulate gene expression by novel mechanisms such as RNA interference, gene co-suppression, gene silencing, imprinting and DNA demethylation. [Costa, 2005].

It is widely known today that the DNA-RNA connection is bi-directional, reverse transcription does exist. The old and discarded idea that even proteins can carry genetic information came back, at least temporarily, with the discovery of prions. Undoubtedly, there are biological functions which would be difficult to explain without the existence of reverse information flow from proteins back to nucleic acids, for example the immune response and the formation of long term memory, both requiring de novo protein synthesis. [Biro, 2004].

In conclusion relying on several sets of experimental results which indicated

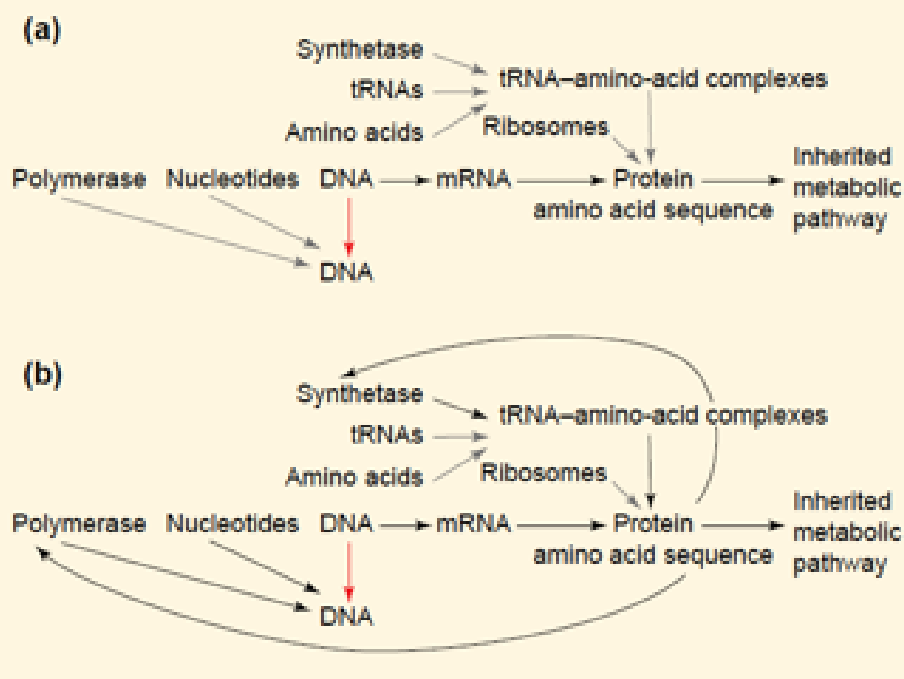


Figure 1.5: Molecular biology central dogma representation under their exceptions. [Source Thieffry and Sarkar, 1998]

that compounds other than DNA (e.g. DNA polymerase and aminoacyl-tRNA synthetase) affect the result of transcription, Commoner replaced the unidirectional flow of information prescribed by the Central Dogma with a more complex scheme that explicitly includes feedback from proteins to DNA and RNA (Figure 1.5). [Thieffry and Sarkar, 1998].

1.5 Gene expression control

The characteristics of organisms result largely from the dynamic interplay between DNA or RNA and the regulatory apparatus. The gene expression control is a fundamental process to bring the genome to life, and it pervades most of biology, from cell proliferation and differentiation to development.

Recent technological advances are transforming our understanding of how the DNA sequence of the genome is transcribed into its functional output of RNA and protein. Researchers are uncovering new layers of complexity on many levels, ranging from the mechanism by which genes are transcribed into RNA to how genetic changes can give rise to disease. [Eccleston and Skipper, 2009].

There are many forms to control the gene expression when a gene has to be expressed into a functional protein. Those forms of control have been associated to: Pre-transcriptional steps (DNA unpacking and transcription) and post-transcriptional steps (alternative splicing, mRNA transport, mRNA degradation, translation and post-translational protein modification)

1.5.1 Pre-transcriptional control

The eukaryotic DNA is packaged into chromatin, which is organized as DNA segments wound around a protein core of histones call nucleosomes. The chromatin structure is directly related to the control of gene expression because it may block RNA polymerase access to the promoter regions. RNA Polymerases can be three different types: I, II and III in eukaryotes, where RNA polymerase II (RNA Pol II) is part of the transcription mechanism of all protein genes, and the other two polymerases I and III are part of the transcription of the three rRNA genes (28S, 18S, and 5.8S), clustered as a pre-rRNA gene in mammalian genomes, and the tRNA, 5S rRNA and the U6 snRNA genes, respectively.

The combination of nucleosome positioning and dynamic modification of

DNA and histones has a key role in gene regulation and guides development and differentiation [Wang et al., 2008]. Chromatin states can influence transcription process directly by altering the packaging of DNA to allow or prevent access to DNA-binding proteins, or they can modify the nucleosome surface to enhance or impede recruitment of effector protein complexes. Recent advances suggest that this interplay between chromatin and transcription is dynamic and more complex than previously appreciated, and there is a growing recognition that systematic profiling of the epigenomes in multiple cell types and stages may be needed for understanding developmental processes and disease states. [He et al., 2008].

Regulatory proteins bind to DNA either to block or to stimulate transcription, modifying how the DNA interacts with RNA polymerases. Controlling the expression of eukaryotic genes requires general and specific transcription factors. The general transcription factors: TFIIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH (TF stands for transcription factor and II for the RNA Pol II), are required for transcription initiation facilitating the proper binding of RNA Pol II to the DNA [Lee and Young, 2000] (Figure 1.6); and the specific transcription factors (e.g sigma factors in prokaryotes) increase transcription in certain cells or in response to signals.

Methylation (the addition of -CH3) of DNA or histone proteins is associated with the control of gene expression too. The addition of acetyl groups to histone tails remodel the nucleosomes doing that DNA available for transcription [Berg et al., 2006]. Clusters of methylated cytosine nucleotides bind to a protein that prevents activators from binding to DNA. Methylated histone proteins are associated with inactive regions of chromatin.

1.5.2 Post-transcriptional control

The post-transcriptional mechanisms are an important part of the regulation of gene expression, making possible that the gene expression can be controlled precisely, mostly by modulation of mRNA stability [Pautz et al., 2006]. Post-transcriptional mechanisms of gene expression are intricate, and their elucidation is essential for obtain full understanding of how gene expression is regulated, which is the interplay between these mechanisms, and how their dysfunction is involving in numerous genetic disorders and cancer. [Chen and Manley, 2009].

Post-transcriptional mechanisms usually involves: RNA interference, alternative splicing, RNA editing and mRNA degradation. RNA interference

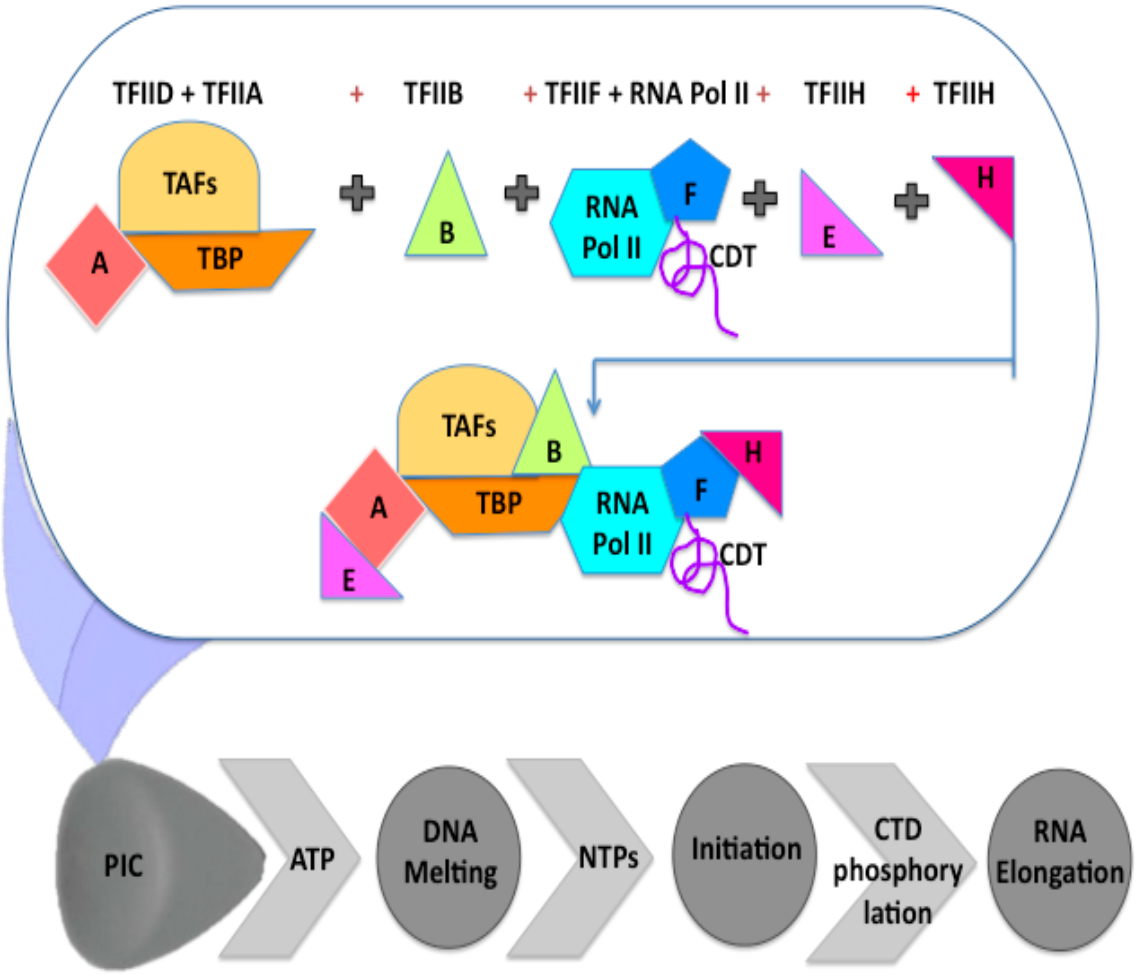


Figure 1.6: Assembly of the pre-initiation complex (PIC). TBP first binds to the promoter and then recruits TFIIB to join TFIID (and TFIIA if present). Before entering PIC, RNA Pol II and TFIIF are bound together, and recruited by TFIIB. Finally, RNA Pol II recruits TFIIE, which further recruits TFIIH to complete the PIC assembly.

involves the use of small RNA molecules. The enzyme Dicer chops double stranded RNA into small pieces of RNA by means micro-RNAs bind to complementary RNA to prevent translation and small interfering RNAs degrade particular mRNAs before translation. [Lodish et al., 2007].

In a recent publication, Brookes and Pombo [2009] found that the regulation of gene expression is essential for the generation of diverse cell types during the development and the adaptation process to the environmental signals. RNA Pol II transcribes genetic information and coordinates the recruitment of accessory proteins that are responsible for the establishment of active chromatin states and transcript maturation. RNA Pol II is post-translationally modified at active genes during transcription initiation, elongation and termination, and thereby recruits specific histone and RNA modifiers. RNA Pol II complexes are also located at silent genes in promoter-proximal paused configurations that provide dynamic transcriptional regulation downstream from initiation. [Brookes and Pombo, 2009].

Genes that encode various RNAs are transcribed in the nucleus (although some are transcribed in mitochondria) by RNA Pol I, II or III, been the primary transcripts of theses genes in eukaryotes cells, virtually never the mature active species. These primary RNAs usually undergo post-transcriptional processing: $5'$ capping, $3'$ polyadenylation, pre-mRNA splicing and in some cases RNA editing (alternative splicing). The rate of translation of mature mRNA depends of its concentration in the cytoplasm, which is determined by the RNA quality control, mRNA transport, storage and degradation mechanisms.

The splicing

Typically a pre-mRNA is a single strand composed by nucleotide's sequences called introns that will not be translated nor alternated with exons but will be translated into a protein. Introns are spliced out of pre-mRNAs to produce the mature mRNA that will be translated. Mature mRNA molecules have various half-lives depending on the gene and the tissue where they are been expressed, and the amount of polypeptides produced from a particular gene is influenced by the half-life of the mRNA molecules.

Maturation of mRNA transcripts requires sophisticated machineries to remove the introns and join the exons. This process known as 'splicing' is regulated by dedicated splicing factors and is a very significant source of proteome complexity because alternative exon-assembly (alternative splic-

ing) allows for a majority of genes to encode more than one protein.

Alternative splicing is a regulated mechanism and the product of a gene will frequently vary from one tissue to the other, dictated by the balance of splicing enhancer and repressor activities that are presented [Shi and Manley, 2007, Matter et al., 2002]. Recent data have shown that these factors contact the RNA Pol II as well as transcription factors and chromatin remodeling enzymes present inside the coding region of the gene [Allemand et al., 2008]. Splicing also responds to various signal transduction pathways activated by external stimuli. Such stimuli are able to modify the activity of the splicing factors or their expression level. Interestingly, the transcription's activity factor are also regulated by the same signal transduction pathways and even more, there has been some speculation on the fact that transcriptional machinery could also be involved in splicing. This would only be truth if splicing is initiated while the pre-mRNA is still under synthesis. Some authors also suggest that chromatin could function as an RNA-binding matrix displaying the immature transcripts to the spliceosomes. [Allemand et al., 2008].

Alternative splicing markedly affects human development, and the lack of regulation and control over the splicing has improve many human diseases. Although the mechanisms of alternative splicing have been studied extensively, in the last few years has been discovered the diversity and complexity of alternative splicing regulation by an intricate protein-RNA network. The progress in such area still increases, and the research and study of individual transcripts and through genome-wide approaches have provide a better picture of the mechanistic regulation of alternative pre-mRNA splicing. [Chen and Manley, 2009].

In addition, a vast number of alternative splicing events have been identified, adding to the complexity and ubiquitous expression of the human transcriptome [Tress et al., 2007, Wang et al., 2008]. This plasticity at the RNA level is even further accentuated by the presence of an immense number of inhibitory RNAs [Yelin et al., 2003, Katayama et al., 2005] and the recent discovery that tens of thousands of binding sites are present across the genome, as shown by genome-wide profiles of the DNA binding of mammalian transcription factors. [Robertson et al., 2007, Ponten et al., 2009].

Spliceosome

The spliceosome is a ribonucleoprotein machine which main function is to remove introns from pre-mRNA in a two-step reaction. Warkocki et al. [2009]

investigated the catalytic steps of splicing and they established an in vitro splicing complementation system. This system allows future mechanistic analyses of spliceosome activation and catalysis.

The spliceosome assembly Is carried out by a sequential formation of the A, B and C complexes, from the ensemble of two previous complexes known as E and H (Figure 1.7). Different substrate-specific proteins are part of H (colored in orange - 1.7) and they are not directly involved in the spliceosome assembly, but their bound proteins might influence the ability to progress to the E complex, the earliest complex committed to the splicing pathway. Although they have not traditionally been thought to contain splicing factors, the H complex formed on the c-src N1 exon contains U1 small nuclear ribonucleoprotein (snRNP). Within the E complex, all the consensus splice-site elements are recognized: the $5'$ splice site (indicated as GU) is recognized by U1 snRNP, the branch point (BP; pink circle) by SF1 and the polypyrimidine tract (PPT; pink rectangle) and $3'$ splice site (AG) by $U2AF^{65}$ and $U2AF^{3}5$, respectively (blue ellipses 65 and 35). E complexes have mainly been characterized as cross-intron complexes (left), but in multi-exon pre-mRNAs they form initially across exons (i.e. exon definition; right). Stabilizing interactions between factors bound to the $5'$ and $3'$ splice sites are indicated schematically by the double-headed red arrows. Such interactions might be mediated by SR proteins, which can bind to exon splicing enhancers (ESE), or formin-binding protein (FBP). Cross-exon interactions must be replaced by cross-intron interactions in the later splicing complexes (B and C) for splicing to occur. [Spellman and Smith, 2006].

RNA binding proteins (RBP)

The processes of RNA splicing, transport, capping, editing, and polyadenylation depend mainly on protein factors that recognize the pre-mRNA and assemble the appropriate pre-mRNA processing complexes. Many different protein factors that guide pre-mRNA modification pathways are composed of a limited number of conserved, modular RNA-binding domains. Of these, the RNA recognition motif (RRM) domain is by far the most abundant type of eukaryotic RNA-binding motif. The most conserved RRM signature sequence is an eight-residue motif called ribonucleoprotein 1 (RNP1). Besides the protein association with RNA, the protein-protein interactions are fundamental for extracting and obtaining the catalytic components to sites of RNA modification and to coordinate pre-mRNA processing with other cellular pathways. Further more, common protein interaction domains, such as SH2, SH3, and WW motifs, are rarely observed in pre-mRNA processing

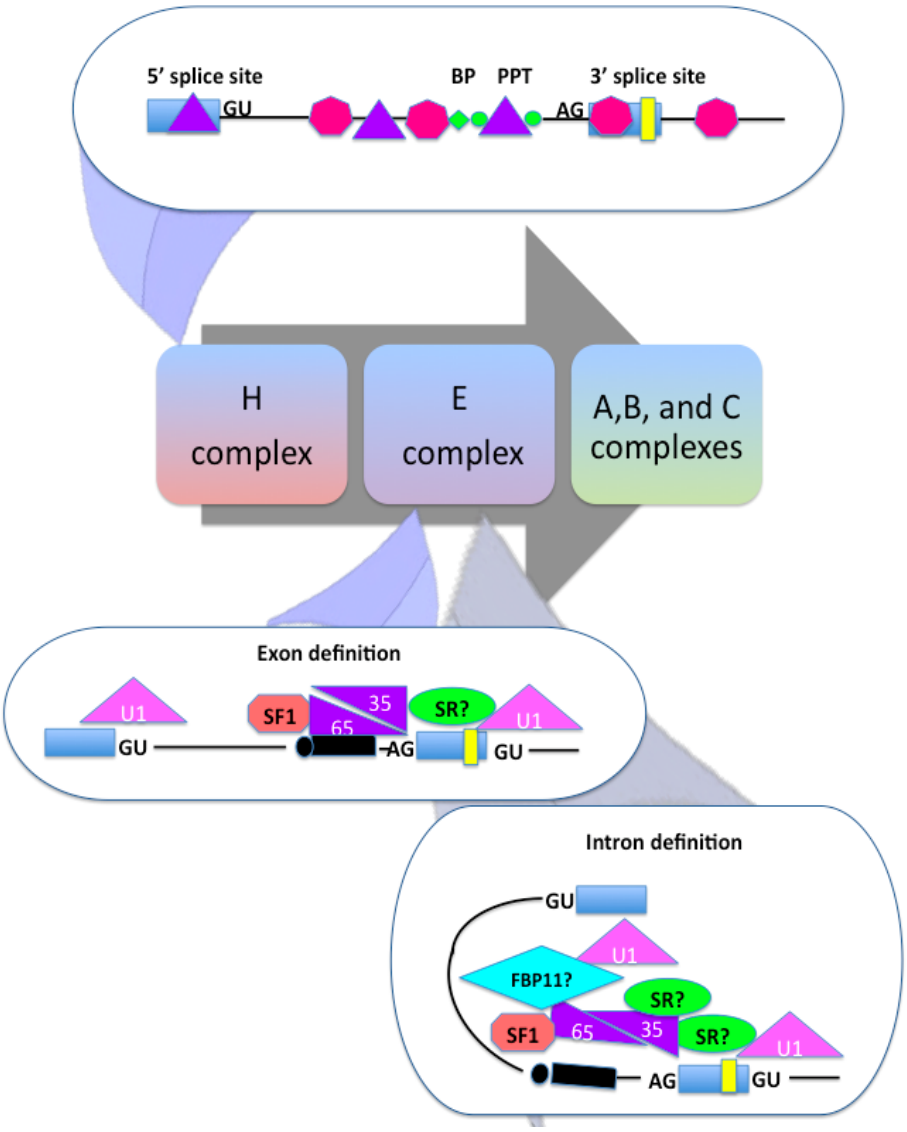


Figure 1.7: Spliceosome assembly: H and E complexes

factors [Shatkin and Manley, 2000, Zhou et al., 2002], where the ability to interact between each other probably relies the sequences previously thought to be involved in RNA binding. [Kielkopf et al., 2004].

Several RNA-binding proteins are characterized by having a modular structure and being composed of multiple repeats of few basic domains, which are arranged in different ways in order to satisfy their many and various functional requirements. Lately, researches have been investigating the cooperation between different modules that allows regulating the RNA-binding specificity and the biological activity of these proteins. They have also investigated how multiple modules cooperate with enzymatic domains to regulate the catalytic activity of enzymes that act on RNA. And so the results show how multiple modules define the fundamental structural unit that is responsible for biological function [Lunde et al., 2007], at least for several RNA-binding proteins.

RNA-binding proteins (RBPs) perform many essential functions in the post-transcriptional control of gene expression. Play essential functions in all post-transcriptional regulatory processes, including RNA processing, cellular localization, translation and mRNA decay. These proteins recognize RNA by using relatively few RNA-binding modules that combine to create versatile macromolecular binding surfaces to define their specificity [Auweter et al., 2006, Lunde et al., 2007, Chen and Varani, 2005].

RNA binding proteins (RBP) classes

At least nine families of RNA binding proteins have been identified using sequence-based analyses of RNA binding proteins, together with functional characterization of mutations that affect the specificity or affinity of RNA binding. But the number of experimentally determined structures for RNA-protein complexes is still relatively small and heavily biased (ribosomal proteins represent 50% of all RNA binding proteins in the Protein Data Bank [PDB]). However, some computational analyses of RNA-protein complexes have generated databases of RNA-protein contacts and provided valuable insights into the biophysical basis of interaction patterns between ribonucleotides and amino acids. [Terribilini et al., 2006].

Polypyrimidine-tract-binding protein

Polypyrimidine-tract-binding protein (PTB) is a repressive regulator of alternative splicing. Splicing decisions are usually the result of multiple an-

tagonistic and synergistic influences of regulatory proteins. Activators of the serine/arginine-rich (SR) protein family bind to splicing enhancers and consist of RNA-binding domains and 'RS' domains (enriched in Arg-Ser dipeptides) that contact other splicing factors and RNA. Repressor proteins such as PTB typically bind to splicing silencers and have RNAbinding domains, but lack RS domains. Most regulators seem to act at an early stage in the pathway of spliceosome assembly (Figure 1.7).

Although spliceosomes form across introns, in multi-exon pre-mRNAs, the early pre-spliceosomal 'E' and 'A' complexes (Figure 1.7) initially form cooperatively across exons in a process termed 'exon definition'; thus, regulators could influence the formation of cross-exon or cross-intron complexes. SR proteins are prime candidates to mediate cross-talk between 30 and 50 splice sites, but the interactions network in exon definition is not well defined. The recent work from the Valcarcel (www.crg.es) and Black (www.hhmi.ucla.edu) laboratories show that the splicing repressor PTB targets the interactions involved in exon and intron definition. [Spellman and Smith, 2006].

U2AF

During pre-mRNA splicing, $U2AF^{65}$ (an heterodimeric splicing factor U2 snRNP auxiliary factor -$U2AF^{65}$-) and some other important factors are responsible for making easy the sequential association between small nuclear RNP particles (snRNPs), including U1, U2, U4, U5, and U6 snRNPs, with the borders of intervening pre-mRNA sequences [Brow, 2002]. Following assembly of the functional spliceosome, the intron is excised as a branched lariat by two catalytic steps, and adjacent exons are joined together to form the spliced mRNA.

$U2AF^{65}$ was identified as a factor that binds to pre-mRNA consensus sequences at the $3'$ splice site (3' SS), and is required for stable association of the U2 snRNP core spliceosome particle with the pre-mRNA branch point sequence (BPS) during the first ATP-dependent step of the splicing process (Complex A) [Ruskin et al., 1988, Zamore and Green, 1989]. Soon, finding that both subunits are heavily needed in Drosophila melanogaster, Caenorhabditis elegans and Schizosaccharomyces pombe, the $U2AF^{65}$ in vitro proved to be very important.

Because $U2AF^{65}$ commits the pre-mRNA to the first critical ATP-dependent step of splicing, its binding is often regulated during alternative splicing. In humans, the products of five $U2AF^{3}$5-like open reading frames and the

single U2AF^{65} subunit may form distinct heterodimers with different functional activities. In addition to U2AF^{65}, other non-snRNP protein factors are required for building up Complex A, including Splicing Factor 1 (SF1) and Splicing Factor 3b (SF3b), a multisubunit component of the U2 snRNP. [Kielkopf et al., 2004].

1.6 RBPs immunoprecipitation: PTB and U2AF^{65}

The RBPs U2AF^{65} and PTB as was said above are part of the spliceosome and they have been studied by Gama-Carvalho et al. [2006], using RIP-Chip technology. This experiment will be described in the next paragraphs, since we applied the methodologies presented in this document to those data obtained by Gama-Carvalho et al. [2006] using RIP-Chip technology.

HeLa cells were used by Gama-Carvalho et al. [2006] as a eukaryotic type cell for the detection of the mRNA pools associated with U2AF^{65} and PTB RBPs during the post-transcriptional gene expression control. In this experiment (see Figure 1.8) an adherent HeLa cells were grown in Dulbecco's modified Eagle's medium (DMEM), 10% fetal calf serum (FCS) and penicillin/streptomycin.

For the microarray experiments two different samples were prepared, an Input sample with the total mRNAs and immunoprecipitated sample (IP) for a mRNA pool associated to a specific RBP (PTB or U2AF^{65}).

The Input sample corresponds to total RNA isolated from the HeLa cell lysate that was used as the starting point for the immunoprecipitation assay. [Gama-Carvalho et al., 2006]

The IP sample was obtained by the immunoprecipitation of PTB or U2AF^{65}-associated mRNAs performed over pre-cleared HeLa cell lysate. First, immune complexes were obtained by addition of a specific monoclonal antibody (Bb7 anti-PTB mAb or anti-U2AF^{65} MC3) over the pre-cleared HeLa cell lysate. Second, the immune complexes precipitation was made with a slurry with 50% of protein A/protein G agarose beads blocked with tRNA and RNAse free bovine serum albumin, followed by washes performed with lysis buffer. [Gama-Carvalho et al., 2006]

Complexes bounded to the beads were eluted with TES buffer (10 mmol/l Tris, 0.5 mol/l EDTA,0.5% SDS [pH 8.0]) after heating at 65°C for 10 min-

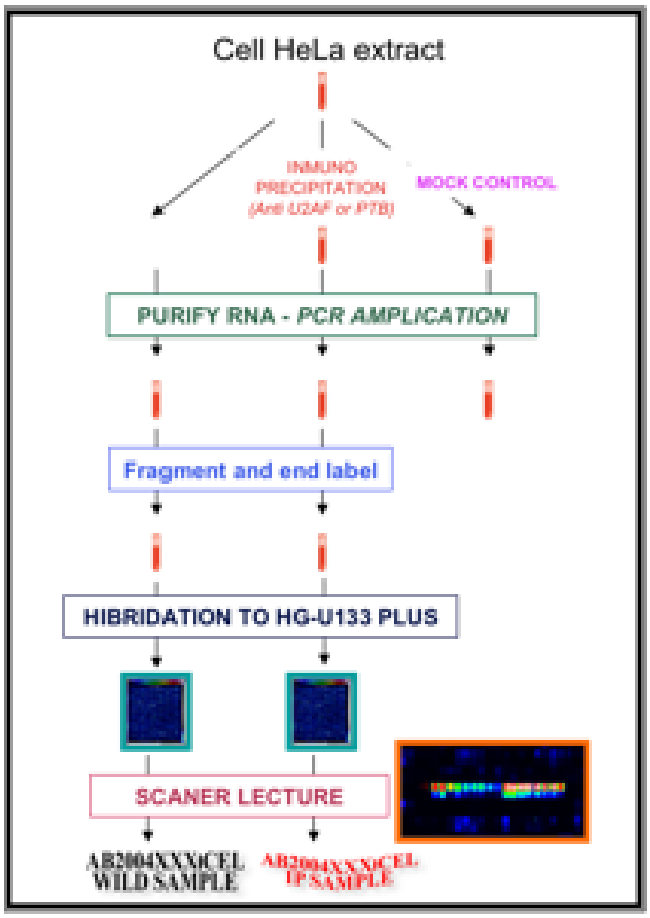


Figure 1.8: Immunoprecipitation of mRNAs bound to RBPs PTB and U2AF^{65}. Hibridization with GeneChip® Human Genome U133 Plus 2.0 microarrays.

utes and were Trizol extracted for posterior mRNA analysis. RNAs from the Input and IP samples were polyadenilated, reverse transcribed, end tagged and PCR amplified. Resulting cDNAs from Input and immunoprecipitated (IP) samples for microarray hybridization were labeled and hybridized to Affymetrix GeneChip® Human Genome U133 Plus 2.0 Arrays (Affymetric, Inc., Santa Clara, CA, USA), as was described by Gama-Carvalho et al. [2006].

A total of three independent immunoprecipitation experiments for U2AF^{65} and two for PTB were performed (Input and IP sample pairs, for each RBP), corresponding to 10 microarray hybridization datasets overall.

1.7 Microarray technology

Given that microarray hybridization is an important part of the RIP-Chip methology applied by Gama-Carvalho et al. [2006], which results correspond to the data sets used in this thesis, this section presents a short overview of the microarray technology.

Microarray technology is based on the parallel hybridization of complementary nucleotide strands (DNA or RNA) with a genomic or transcriptomic samples, that allows to detect the expression of hundred of thousands of genes at the same time [Schena, 1996]. Microarrays consist of thousands of DNA molecules named probes (representing different genes) that are gridded and immobilized onto a support such as nylon membrane, glass or silicon. Each position of the microarray grid is known as spot or cell depending on the technology used and corresponds to a specific gene or transcript.

The detection of the expression levels of all genes in cells taken from a cell culture (tissue sample, cell line culture, microorgaminsm culture, etc.) using microarrays is made by isolating the total mRNA (cell trascriptome at given condition). cDNA fluorescently or radioactively labelled is then prepared from the isolated mRNA pool (target molecules) and hybridized to the microarray. Afterwards, the unhybridized cDNA targets are washed away and the signals of the hybridized cDNAs are quantified by measuring the fluorescence or radioactive signal of each spot or cell.

The measure of the signal of the hybridized cDNA targets is proportional to the amount of isolated mRNA. The relative abundance of hybridized targets on a defined array spot can be determined, allowing to compare in parallel the transcription level of several thousands of different genes from one sample to other samples in different experimental conditions.

Different array types have been designed, which differ by the density (spots per square centimeter) of the array and more important, by the type of probes that are immobilized (synthetic oligonucleotides or cDNA). There are two basic sources for DNA probes on a microarray that define the microarray type. First, microarrays where each probe (oligonucleotides between 20 to 80 mer long) is individually synthesized in situ on a rigid surface (in general glass) using the photolithographic methods, one of them developed by Fodor et al. [1993] and initially commercialized by Affymetrix (described next section).

Second, microarrays where each probe is pre-synthesized or PCR amplified (usually 25-5000 mer long) and immobilized to a solid surface such as glass (or nylon) using a robot spotting, methodology proposed more than 30 years ago by Ed Southern and popularized by the Patrick Brown laboratory at Stanford University. The Stanford methodology has been extensively used by the academic research laboratories because it is more affordable in the sense that it allows flexible microarrays design and the equipment is cheaper in relationship with the in situ synthesis microarrays. The last ones require more sophisticated and expensive robotic equipment, but allows high density microarrays (hundred of thousand of spots per array) in contrast to 200 to 10,000 spots allowed by the Stanford arrays.

1.8 Affymetrix microarrays

Affymetrix microarrays are produced using a combination of photolithography and combinatorial chemistry. In each array are synthesized up to 1.3 million different probes. Each probe (25 mer oligonucleotide) is located in a specific area within the array, called probe cell and each cell may contain hundreds of thousands or even millions of copies of given oligonucleotide. [Affymetrix, 2009]

The probes correspond to 11 up to 20 unique regions that allow the specific hybridization of each one of the complementary sequences (genes, RNAs, expressed sequence tags [ESTs]) aimed to be detected in a sample. These probes are named perfect mach (PM). For each PM in the array an identical probe is included which differs only in the 13th position, where the original nucleotide at the PM is substituted by its complementary nucleotide (A by T or C by G, or vice versa) (see Figure 1.9). These probes are called mismatch (MM) and are used as a reference for later background correction. [Affymetrix, 2009].
Figure 1.10 shows how the Affymetrix arrays are made through a series of cycles where all the probes are synthesized simultaneously on the array surface. Initially, the glass substrate is coated with linkers containing photo labile protecting groups. Then a photolithographic mask is applied exposing selected glass areas (probes cells) to the ultraviolet light. This process removes the protective groups and allows the selective addition of a specific phosphoramidite nucleoside on the exposed sites. Four cycles using 4 different photolithographic masks, one for each nucleotide (ATGC) are made for synthesizing the first probe position of every probe included in the array.

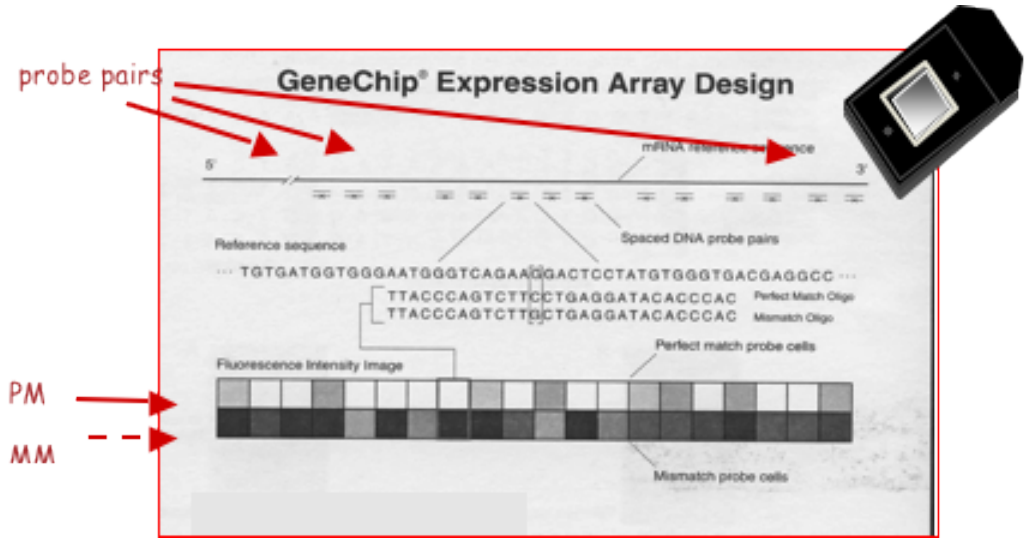


Figure 1.9: Affymetrix probe selection. 11 up to 20 unique region in the genomic target (gene, RNA, EST) are selected and used as a template of the perfect mach probes and the mismatch probe during the array design.

Thus cycles are repeated until the 25 nucleotides of each probe are synthesized [Affymetrix, 2009].

The Affymetrix array used by Gama-Carvalho et al. [2006] was the GeneChip® Human Genome U133 Plus 2.0. This array provides complete coverage of the Human Genome, allowing the analysis of more than 38,500 well-characterized human genes and 47,000 transcripts on a single array, using more than 54,000 probe sets and more than 1.2 million probes. [Affymetrix, 2007]

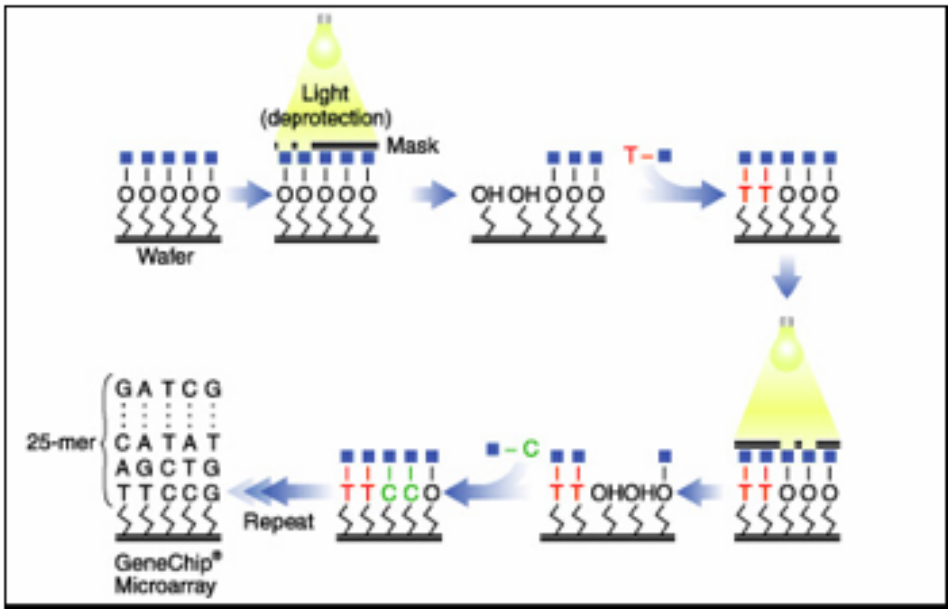


Figure 1.10: Synthesis of oligonucleotides on GeneChip® microarrays, based on the concept of photolithography. (Source: http://www.dkfz.de/gpcf/24.html)

Chapter 2

RIP-Chip experiment data analysis

2.1 Introduction

RNA Binding Proteins (RBP) are an important part of the gene expression regulation. They are involved in the intron elimination and in the synthesis of different proteins from a single gene through alternative splicing. It has also been established that one of the gene control expression level mechanisms of the cells after transcription is binding some RBPs to specific sites of the transcript sequences. This mechanism can alter the speed of transcript concentration decay, depending of the RBPs affinity and the cells environment, controlling the protein production and giving the cell the ability to change transcript levels in a fast way. [Mata et al., 2005].

In this context, it has been proved that some proteins bind differentially to mature mRNAs and travel with them to the cytoplasm, contributing to the translation level control of some genes [Gama-Carvalho et al., 2001]. That is the case of the U2 small ribonucleoprotein ($U2AF^{65}$) which is an auxiliary factor in the spliceosome assembly occurred inside the nucleus [Zamore and Green, 1989, Zamore et al., 1992], in charge of recognizing the AG dinucleotide at the $3'$ splice site [Merendino et al., 1999, Wu et al., 1999, Zorio and Blumenthal, 1999]; and the case of the Polypyrimidine Tract Binding Protein (PTB), component of the spliceosome [García-Blanco et al., 1989], involved in the alternative splicing regulation, suppressing the inclusion of alternative exons [Black, 2003]. The $U2AF^{65}$ and the PTB have been reported to bind differentially to mature mRNAs and travel with them to the cytoplasm, helping in the translation control level of certain groups of genes

and other cellular functions . [Gama-Carvalho et al., 2001]

The high throughput molecular technique known as RNA immunoprecipitation and hybridization on Chip (RIP-Chip) [Baroni et al., 2008, Keene et al., 2006], has been used for the simultaneous determination of the mRNAs that are binding to a specific RBP. Basically this technique consists of the immunoprecipitation of the a complex mRNA-RBP using RBP antibody, followed by the mRNA purification and later microarray hybridization. The analysis of the data obtained in this way is the aim of the present work.

2.2 The Gama-Carvalho et al. RIP-Chip experiment data analysis

The hybridization of the labeled-cDNA (obtained from the immunoprecipitated mRNAs or the Input samples) to an Affymetrix array is followed by the acquisition of fluorescence signals emitted by the labeled-cDNA hybridized, using laser scanning confocal microscopy. Typically the Affymetrix scanner produces an image that is stored as a file with the extension .dat.

The Affymetrix GCOS (GeneChip® Operating Software) suite computes cell intensity data from the image file, and a .cel file is obtained. It contains a single intensity value for each probe cell delineated by the grid (calculated by the Cell Analysis algorithm). [Affymetrix, 2004]

As it was mentioned above Gama-Carvalho et al. [2006] ran a total of three independent immunoprecipitation experiments for $U2AF^{65}$ and two for PTB (Input and IP sample pairs for each RBP), corresponding to a total of 10 microarray hybridization datasets that were laser scanned and processed (10 different .cel files), as is shown in Table 2.1.
The Affymetrix GCOS suite could be used to analyze the cell intensity data [Affymetrix, 2004], as well as many other tools like dChip [Li and Wong, 2001] or Bioconductor – R library specialized in microarray data analysis. [Gentleman et al., 2004]

In general these tools cover two main steps in the Affymetrix data analysis, low level analysis and gene selection. The low level analysis itself has different steps too: microarray quality assessment, background correction, normalization and summarization. Gene selection, may include different statistical and non-statistical methodologies. In the case of the RIP-Chip experiments,

Table 2.1: Gama-Carvalho et al. [2006] experiment .cel files. Input samples correspond to the wild sample and IP correspond to the immunoprecipitated sample

Sample	Experiment type	$U2AF^{65}$	PTB
1	Input	AB2004021301.cel	AB2004031115.cel
	IP	AB2004021302.cel	AB2004031116.cel
2	Input	AB2004021303.cel	AB2004031117.cel
	IP	AB2004021304.cel	AB2004031118.cel
3	Input	AB2004021305.cel	
	IP	AB2004021306.cel	

given the lab methodology applied, this step is named enriched gene selection. The next sections will give a short explanation of these general steps and some them will by applied to the Gama-Carvalho et al. [2006] data.

2.2.1 Low level analysis

Quality assessment

Exploratory data analysis is an important starting point during the microarray data analysis, being the quality assessment (QA) critical for obtaining highly reproducible results.

First step in general is to determine if any anomalies exist by taking a look at the image plots of the probe-level data (PM and MM). The idea is to look for spatial artifacts or other nonhomogeneous patterns in the image plots. The image plots across an experiment array set may help to see whether one or more arrays might appear abnormal or potentially defective, when they display spatial artifacts (for example scratches) or appear lighter or darker than the others.

$\log_2$ image plots

The Bioconductor libraries of R [Gentleman et al., 2004] were used for generating the $\log_2$ image plots for the 4 PTB and 6 $U2AF^{65}$ arrays of the Gama-Carvalho et al. [2006] data. The `ReadAffy()` command of the Bioconductor `Affy` library allows to read the .cel files that contain the intensity

data for each probe along with other important values, and an `affy` object is generated for further analysis. The R code is shown in appendix A.1.1. [Gentleman et al., 2004]

The $\log_2$ image plots in Figure 2.1 (obtained using the function `image()` – R code appendix A.1.1), appear similar to each other displaying no obvious anomalies, neither for the PTB arrays nor for U2AF^{65} arrays of the RIP-Chip Gama-Carvalho et al. [2006] data.

Boxplots and density plots

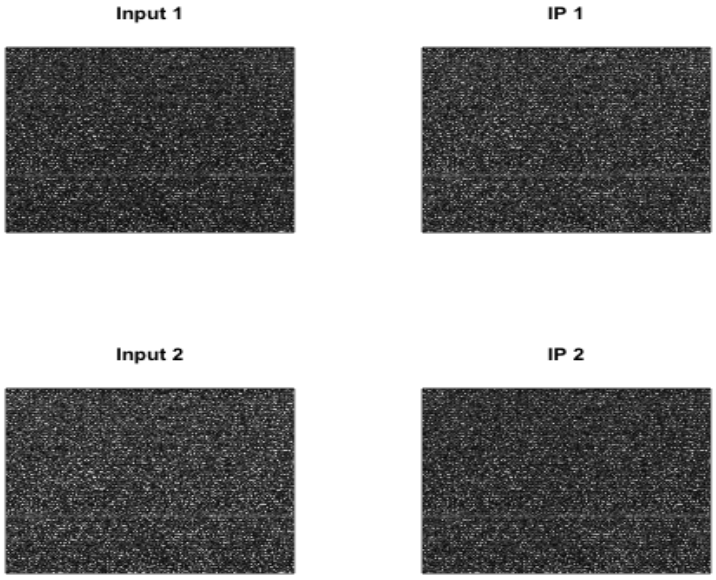
The boxplots and density plots of the probe-level data allow to determine the existence of potentially defective arrays too. For example, boxplots that stand out from the others, that show displaced boxes (interquartile ranges, IQR) or different ranges, or in the case of the density plots, densities that are removed from the others, that display bimodalities, show uniquely different shapes or other abnormalities. [Bolstad et al., 2005]

In the case of PTB and U2AF^{65} RIP-Chip experiments both Figures 2.2 and 2.3 (R code is shown in appendix A.1.1) show that there are no potential defective arrays in Gama-Carvalho et al. [2006] data. Density plots showed a similar and typical curve shape for Affymetrix arrays for both experiments (PTB and U2AF^{65}), not displaying any bimodalities and existing a significant overlap among the individual density plots (Figure 2.3). The probe-level data boxplots (Figure 2.2) do not show any array significantly standing out from the others. These results suggest good quality on the arrays.

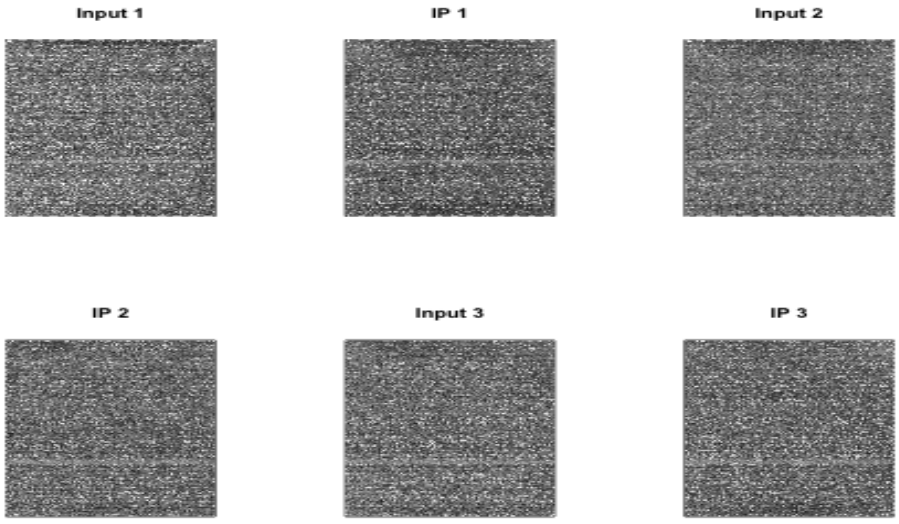
MA plots

MA plot is another exploratory plot for Affymetrix arrays quality assessment. M values represent, for each gene, the difference of the log intensities of a certain probe between two arrays. The M values are plotted against the average (denoted by A) of the same log intensities. When more than two arrays are compared, a reference array is created by taking the probe wise medians [Bolstad et al., 2005] across all arrays, and then each array in the experiment may be plotted versus the reference array.

Quality problems are evident if there is a high variability of the M values in one or more arrays relative to the others [Bolstad et al., 2005]. Another evidence of quality problem is the case where the MA-plot shows that the loess smoother oscillates wildly. MA plots of PTB and U2AF^{65} RIP-Chip

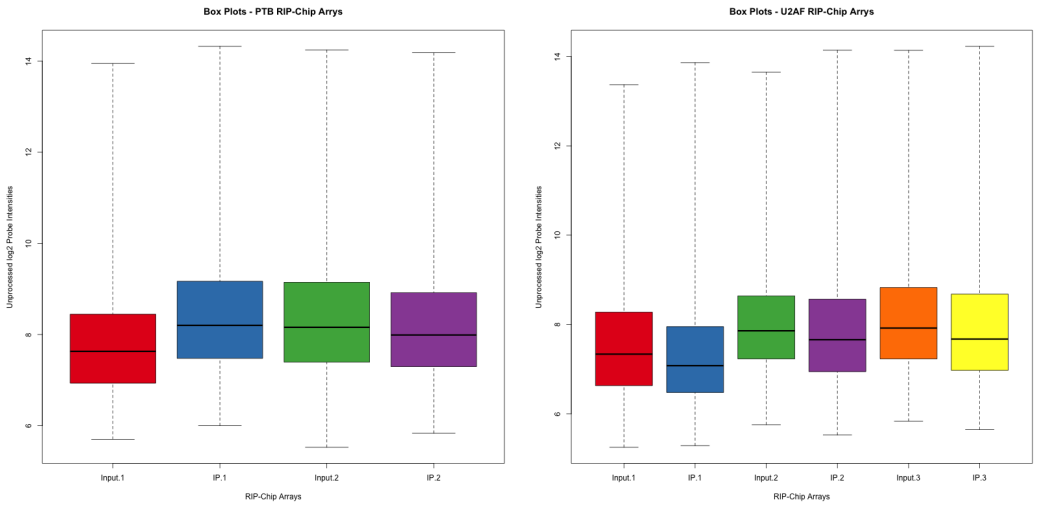


(a) PTB arrays



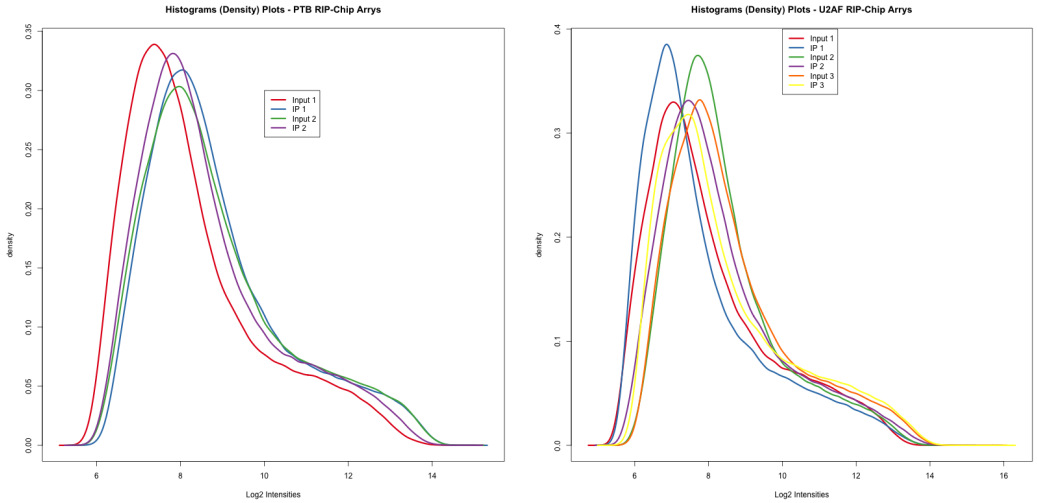
(b) U2AF^{65} arrays

Figure 2.1: Image plots of perfect match (PM) and mismatch (MM) probe intensities ($\log_2$) for PTB and U2AF^{65} RIP-Chips arrays of the Gama-Carvalho et al. [2006] data. Input correspond to the wild samples and IP to the immunoprecipitated samples.



(a) PTB arrays **(b)** U2AF^{65} arrays

Figure 2.2: Boxplots of PTB and U2AF^{65} RIP-Chips arrays of the Gama-Carvalho et al. [2006] data. Input correspond to the wild samples and IP to the immunoprecipitated samples.



(a) PTB arrays **(b)** U2AF^{65} arrays

Figure 2.3: Density plots of PTB and U2AF^{65} RIP-Chips arrays of the Gama-Carvalho et al. [2006] data. Input correspond to the wild samples and IP to the immunoprecipitated samples.

arrays (Figures 2.4 (a) and (b), respectively) where obtained using the R code shown in the appendix A.1.1 and they did not show evidence for the anomalies mentioned before.

RNA degradation plots

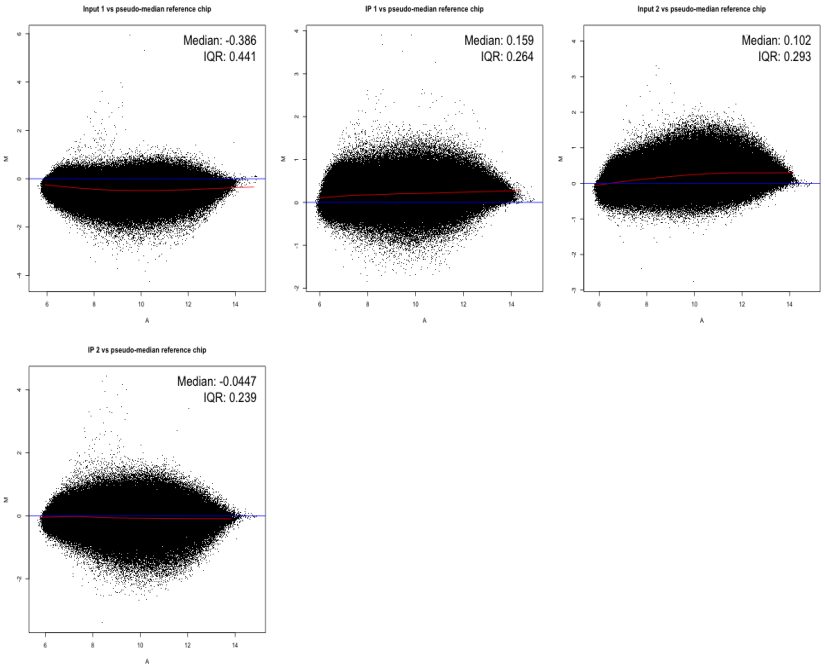
Another Affymetrix Chips quality test is the RNA degradation evaluation. RNA degradation is normally present during the hybridization experiments and is usually major in the 5' end of the RNA. Therefore, if the specific transcript probes are numbered sequentially from the 5' to 3' end of the RNA targeted, PM probes intensities at 5' end may be systematically minor than the 3' end probes intensities when the RNA degradation is high. In general, following 5' $\rightarrow$ 3' orientation, for the middle probe positions the degradation rate is constant and higher than zero, whereas for the ends the degradation rate may not constant assuming small values. [Alvord et al., 2007]

Figure 2.5 produced by the R code shown in appendix A.1.1 shows the RNA degradation for PTB and U2AF^{65} RIP-Chip arrays and their slopes are summarized in the Table 2.2. The curves have the expected behavior and are similar for each RIP-Chip experiment, PTB and U2AF^{65} (Figure 2.5). Slopes are close to 1, pointing out the RNA degradation trend to be moderate and the curves look reasonably parallels. Is important to say that each array type has its own characteristic degradation slope and there is no threshold for determining a bad array, reason because some authors consider that RNA degradation plots have narrowed utility. [Bolstad et al., 2005].

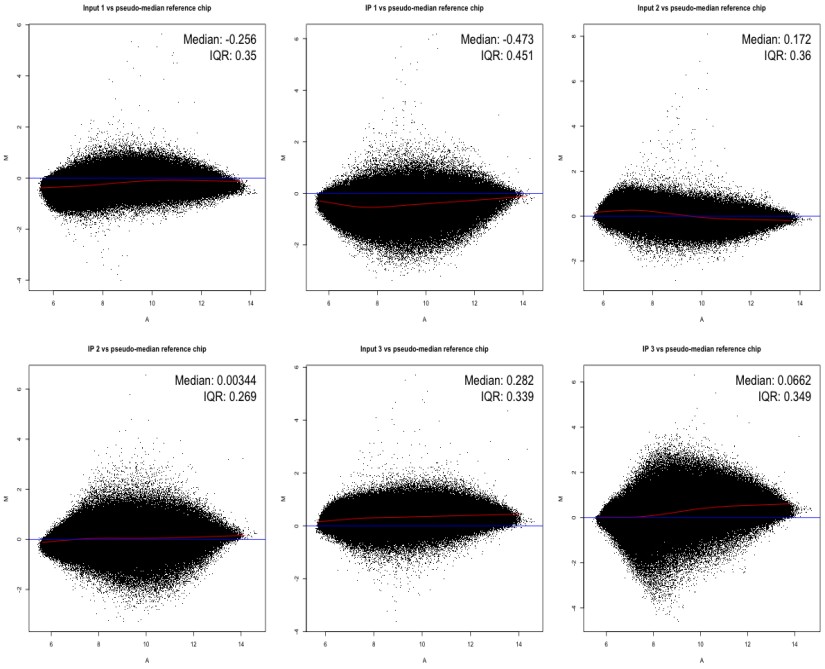
Probe-level models

Probe-level models (PLM) fitted to probe-intensity data [Bolstad, 2011], have been used in determining the quality of Affymetrix chips, providing parameter estimates for probe sets (genes) and chips. In general, the PLM procedures are powerful tools that allow to make visible the effects not appearing in other images or to detect artifacts that might otherwise be missed completely. [Bolstad, 2011].

`affyPLM` Bioconductor library provides different alternatives for fitting PLM through the function `fitPLM` [Bolstad, 2011]. For checking the Gama-Carvalho et al. [2006] RIP-Chip arrays quality, default PLM options were chosen that use the RMA and Quantile Normalization for background correction and normalization, respectively, frequently applied to this kind of analysis. [Bolstad et al., 2005, Bolstad, 2011]

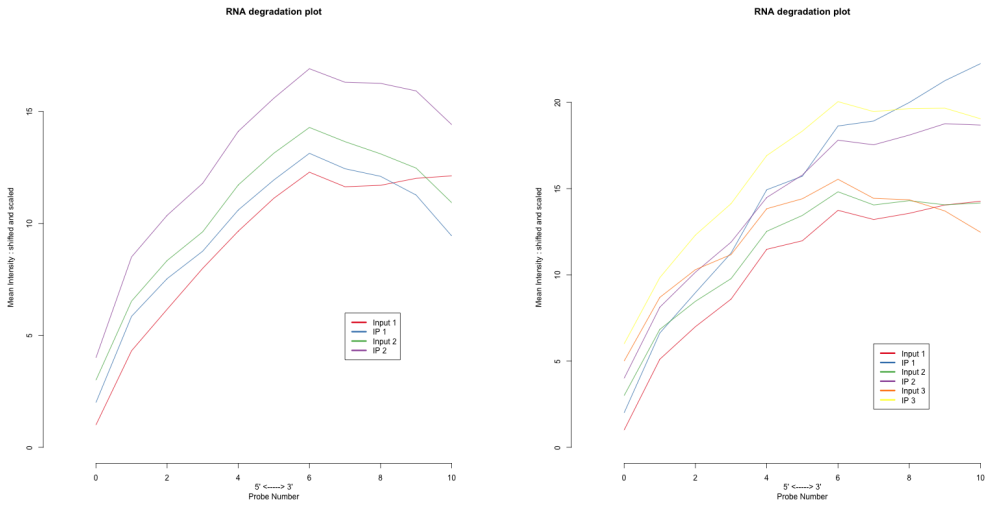


(a) PTB arrays



(b) U2AF^{65} arrays

Figure 2.4: MA plots of each PTB and U2AF^{65} RIP-Chips arrays of the Gama-Carvalho et al. [2006] data versus the synthetic (median) array, centered at zero. Input correspond to the wild samples and IP to the immunoprecipitated samples.



(a) PTB arrays **(b)** U2AF^{65} arrays

Figure 2.5: RNA degredation plos of each PTB and U2AF^{65} RIP-Chips arrays of the Gama-Carvalho et al. [2006] data versus the synthetic (median) array, centered at zero. Input correspond to the wild samples and IP to the immunoprecipitated samples.

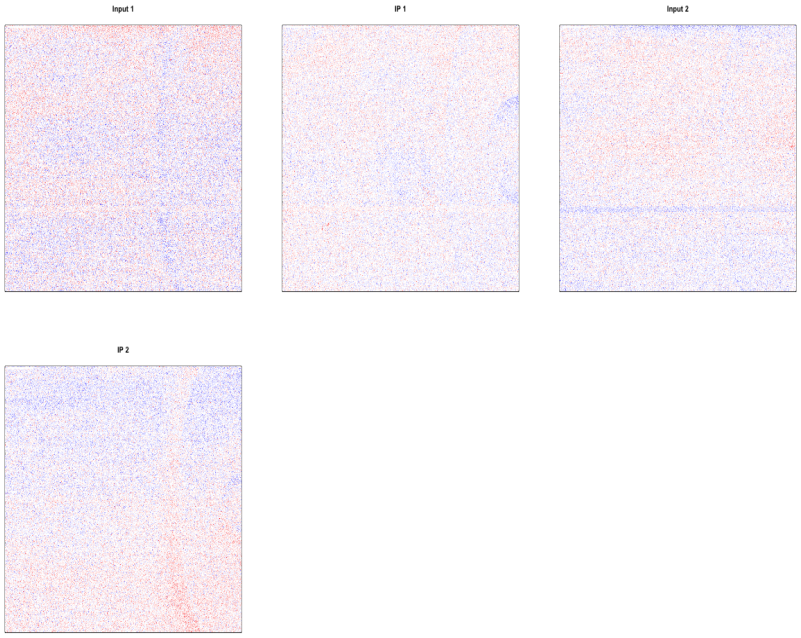
The `fitPLM` function of the `affyPLM` Bioconductor library, fitting a specified robust linear model to the probe level data which fits a specified robust linear model to the probe level data [Bolstad, 2011], returns the following linear PLM model (equation 2.1) for the background correction and probe-level data normalization [Irizarry et al., 2003b]:

$$\log_2 S_{gij} = \theta_{gi} + \varphi_{gj} + \varepsilon_{gij} \tag{2.1}$$

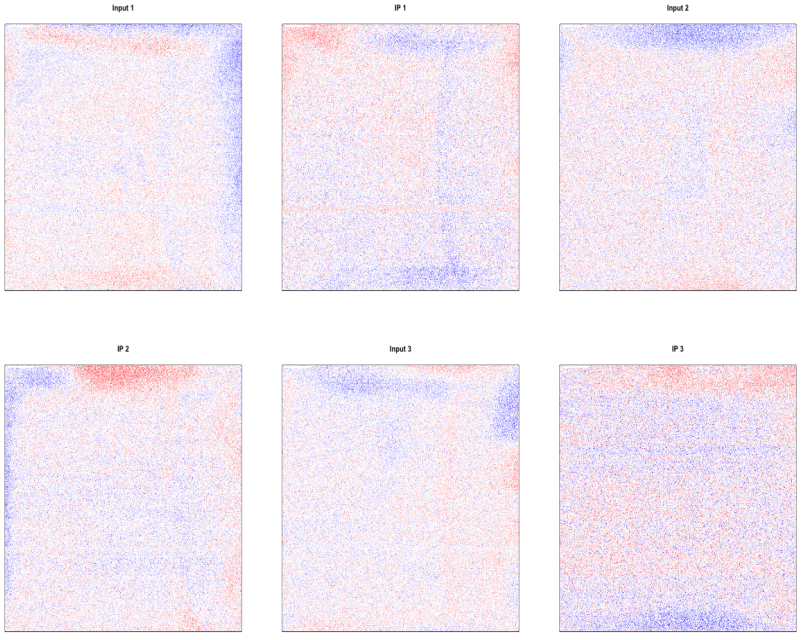
where, S_{gij} is the PM (perfect match) signal value for the j-th probe on the g-th gene on the i-th array; θ_{gi} is the expression level for the g-th gene on the i-th array; φ_{gj} is the j-th probe on the g-th gene; and ε_{gij} is the measurement error. [Irizarry et al., 2003b].

The `fitPLM` function produces an object containing information regarding the parameter estimates, standard errors, weights, residuals and signed residuals [Bolstad et al., 2005, Bolstad, 2011].

Figures 2.6, 2.7 and 2.8 (generated by the R code shown in the appendix A.1.1), show the array pseudo-images of the residuals, weights, and signed



(a) PTB arrays



(b) U2AF^{65} arrays

Figure 2.6: Chip pseudo-images base on residuals of the PLM fit arrays PTB and U2AF^{65} of the Gama-Carvalho et al. [2006] data. Input correspond to the wild samples and IP immunoprecipitated samples.

Table 2.2: Summary of the slopes for the RNA degradation plots for PTB and U2AF^{65} RIP-Chip experiments

PTB RIP-Chip arrays

	Input 1	**IP 1**	**Input 2**	**IP 2**
Slope	1.030000	0.7500	0.80300	1.01000
***p*-value**	0.000173	0.0092	0.00598	0.00172

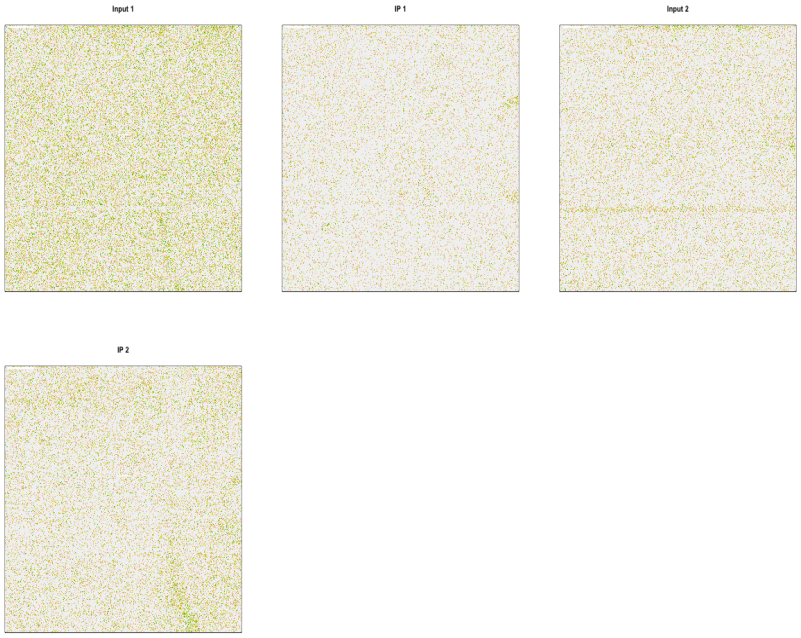
U2AF^{65} RIP-Chip arrays

	Input 1	**IP 1**	**Input 2**	**IP 2**	**Input 3**	**IP 3**
Slope	1.2100000	1.93000000	1.030000	1.400000	0.70700	1.280000
***p*-value**	0.0000715	0.00000075	0.000346	0.0000174	0.00875	0.000235

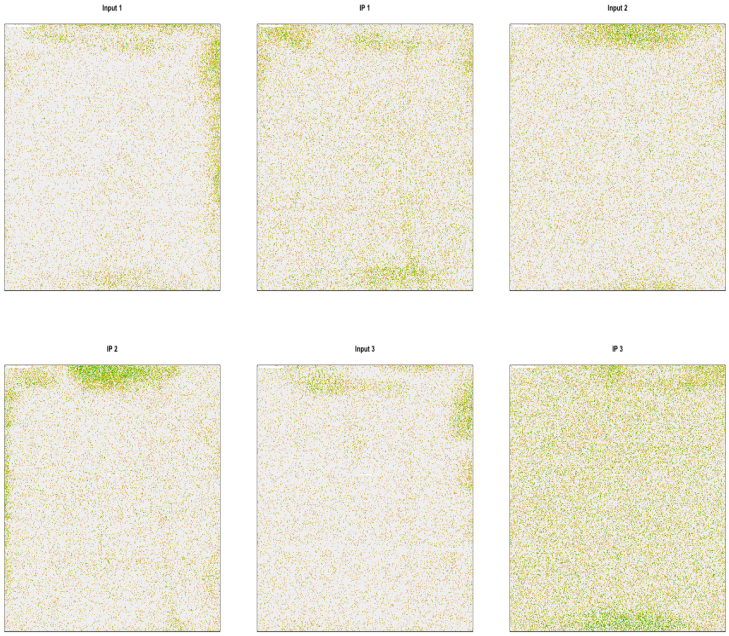
residuals based on PLM fit on PTB and U2AF^{65} RIP-Chip data. Figure 2.6 corresponds to the PLM Residuals plots, where the negative residuals are darker (blue in color plots) and positive residuals lighter (red in color plots) [Bolstad et al., 2005]. Both positive and negative residuals must be homogeneously spread out across the pseudo-images, as it is in general for the PTB microarrays. This behavior is not so evident in the U2AF^{65} pseudo images, where one can find some non-homogeneous areas, however it is not enough to consider these microarrays problematic.

The PLM Weights plot (Figure 2.7) obtained from the robust regression procedure use topographical coloring in the pseudo-images. Dark areas (green in color plots) represent low weights, corresponding to mis-performing probes associated with outliers; and light areas indicate high weights. No apparent spatial artifacts are present in the PTB and U2AF^{65} RIP-Chip plots, however the U2AF^{65} plot shows a few non-significant homogeneous areas but less than the residuals plots (Figure 2.6, b)

The PLM Signed Residuals pseudo-images (Figure 2.8), show the signs of the residuals, either -1 or +1 depending on whether the residual is positive or negative [Bolstad et al., 2005]. No apparent spatial artifacts are present

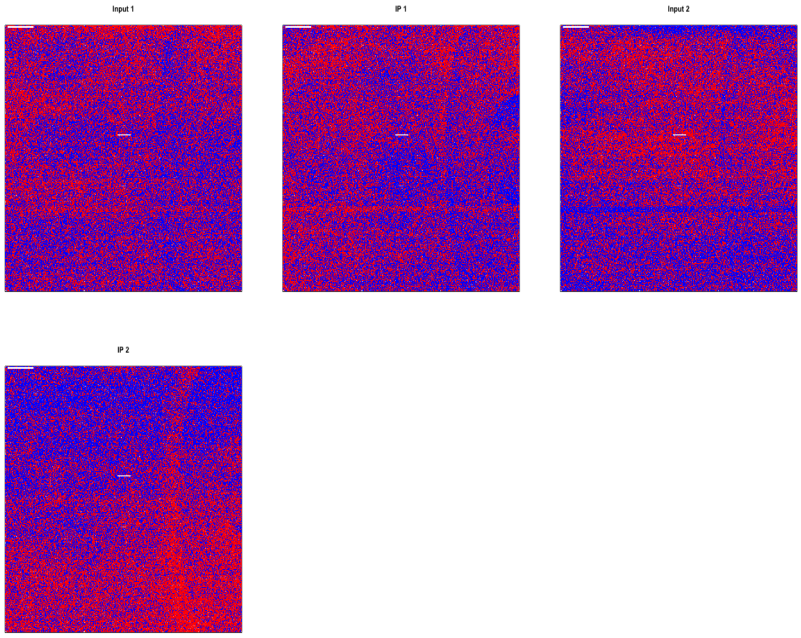


(a) PTB arrays

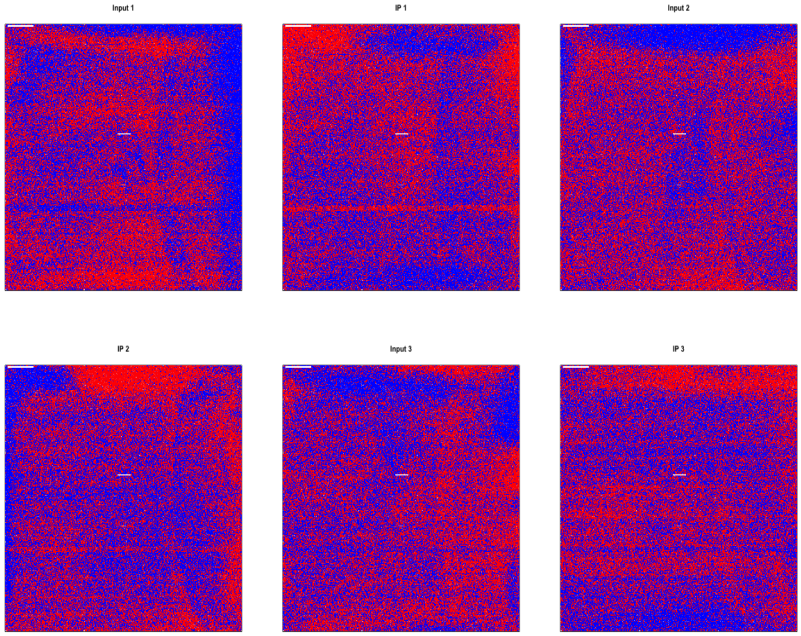


(b) U2AF^{65} arrays

Figure 2.7: Chip pseudo-images base on weights of the PLM fit arrays PTB and U2AF^{65} of the Gama-Carvalho et al. [2006] data. Input correspond to the wild samples and IP immunoprecipitated samples.



(a) PTB arrays



(b) U2AF^{65} arrays

Figure 2.8: Chip pseudo-images base on signed residuals of the PLM fit arrays PTB and U2AF^{65} of the Gama-Carvalho et al. [2006] data. Input correspond to the wild samples and IP immunoprecipitated samples.

in PTB and U2AF^{65} RIP-Chip plots, although U2AF^{65} plots show similar behavior than in the other PLM plots, but not significant enough to consider to remove some of the microarray datasets for further analysis.

Following the fitting of a PLM, it is possible derivate other useful array quality statistics like the Relative Log Expression (RLE) plots (Figure 2.9) or Normalized Unscaled Standard Error Plot (NUSE).

Relative Log Expression (RLE) plots

RLE first estimates, from the PLM fit, the log scale of expression $\hat{\theta}_{gi}$ for each gene g on each array i, and then the median value across arrays for each gene g, $m(g)$, is computed. The RLE is defined as follow:

$$RLE_{gi} = \hat{\theta}_{gi} - m(g) \tag{2.2}$$

These values are then displayed with a boxplot for each array. In a common gene expression experiment it is assumed that the majority of genes are not changing expression between the samples (for example, wild sample vs. treated sample comparison). The RLE boxplot should be centered at 0 and have small spread. The presence of a box that has relatively high spread and is not centered around 0, may indicate an array with quality problems [Bolstad et al., 2005].

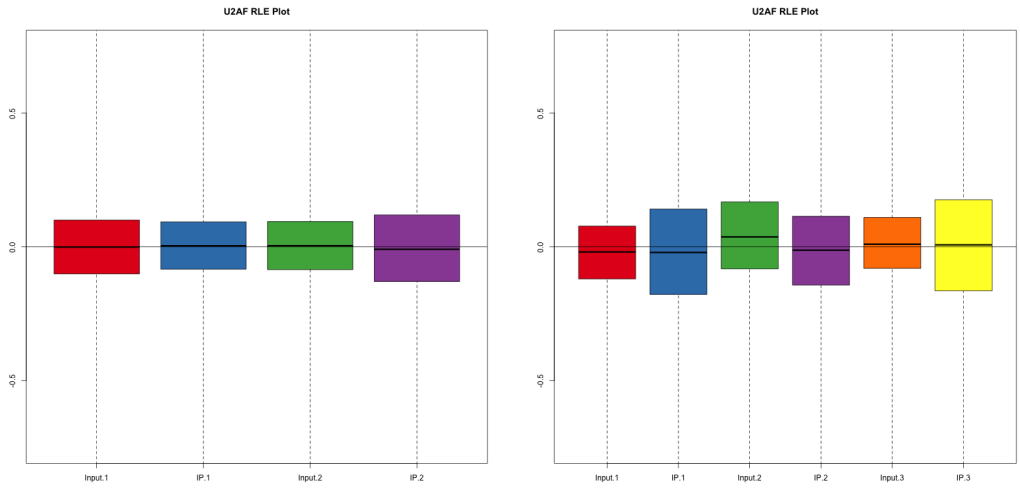
Figure 2.9 (generated by R code presented in appendix A.1.1), shows that the RLE plots for PTB and U2AF^{65} RIP-Chip arrays are centered around 0, with similar box sizes (i.e. IQRs), indicating no quality control problems.

Normalized Unscaled Standard Error (NUSE) plots

Normalized unscaled standard error (NUSE) plot is another graphical tool where the standard error of the estimates SE ($\hat{\theta}_{gi}$) are obtained for each gene g on each array i from the PLM fit. Since the variability may differ considerably among genes, the standard errors of the estimates are standardized as they are divided by the median of the standard errors across arrays, turning the NUSE median equal to 1 for each gene (equation 2.3).

$$NUSE_{gi} = \frac{SE(\hat{\theta}_{gi})}{med_{i'=1,\ldots,I}\{SE(\hat{\theta}_{gi'})\}} \tag{2.3}$$

If some NUSE medians are high or if there are boxes with higher IQRs relative to the others, it may indicate problems with some arrays. NUSE values



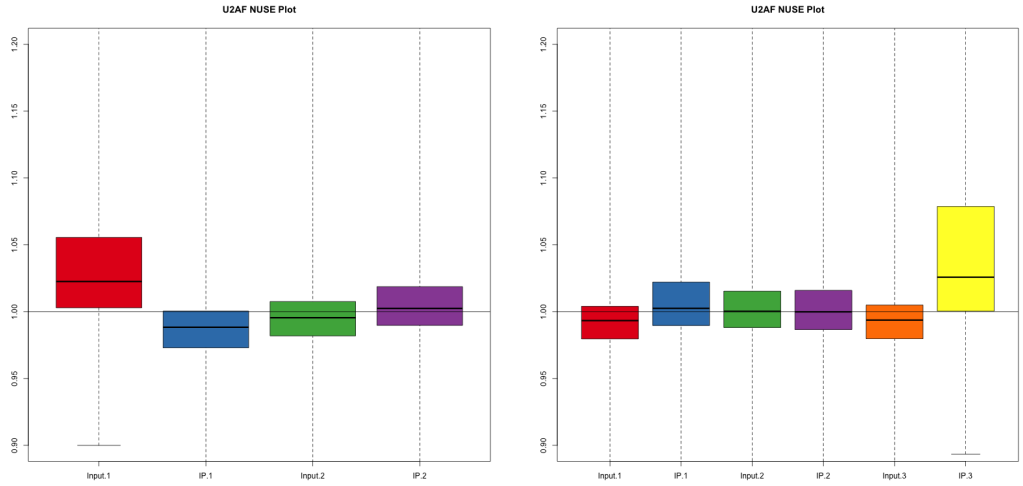
(a) PTB arrays **(b)** U2AF^{65} arrays

Figure 2.9: Relative Log Expression (RLE) Plot of the PLM fit arrays PTB and U2AF^{65} of the Gama-Carvalho et al. [2006] data. Input correspond to the wild sample and IP immunoprecipitated samples

are calculated only within a data set; therefore these values are not comparable across data sets [Bolstad et al., 2005].

It appears that the NUSE boxplots in Figure 2.10 are reasonably centered around the median 1, with relatively equal box sizes. Most of them do not seem to present any quality -control problems, but is evident (Figure 2.10) that the Input 1 from PTB microarray and the IP 3 from U2AF^{65} microarray are slightly different from the others. Both show high values of NUSE medians and boxes that have slightly more spread relative to the others, which could be a sign of bad quality.

Other quality controls may be applied to Affymetrix microarrays, but in general all control procedures serve as an indication wether the microarrays quality is good, or if one or more microarray data sets should be excluded in order to have more confident final results. There is no clear definition on how many of these procedures must indicate a good quality microarray data set, but some authors consider a poor quality microarray if it fails in two or more of theses procedures [Alvord et al., 2007]. For the PTB and U2AF^{65} RIP-Chip microarrays the last criteria was taken and all the chips were considered to have good quality and, therefore, all microarrays were



(a) PTB arrays **(b)** U2AF^{65} arrays

Figure 2.10: Normalized unscaled standard error (NUSE) plot for the PLM fit arrays PTB and U2AF^{65} of the Gama-Carvalho et al. [2006] data. Input correspond to the wild samples and IP to the immunoprecipitated samples.

used in further analysis.

After quality assessment it is necessary to obtain final expression values for each gene in each condition, as the observed intensities need to be adjusted to give accurate measurements of specific hybridization. This process usually involves three steps: background correction, normalization and summarization. [Alvord et al., 2007].

Background correction

Background correction is essential because the microarray raw intensities must be adjusted to remove the influence of the background signal, which may be the result of nonspecific hybridization, auto-fluorescence from the slides, noise in the optical detection system and other reasons.

Different methodologies have been developed to perform this step. For example, the MAS 5.0 algorithm [Affymetrix, 2009] calculates the signal as the anti-log of a robust average (Tukey biweight) of the values:

$$\log(PM_{ij} - IM_{ij}),\ j = 1, ..., J \tag{2.4}$$

where j is the probe number in the microarray i.

Although the mismatch (MM) probes are designed by Affymetrix to have into account the non-specific hybridization and are allowed to correct the perfect match (PM) signal, there is a large proportion of MM probes with higher signal than the corresponding PM probe. Ideal Mismatch (IM) (equation 2.4) is used for dealing with possible negative numbers and is defined as a quantity equal to MM when MM<PM, but adjusted to be less than PM when MM>PM, which occurs in near of 30% of all MM. [Irizarry et al., 2003b].

Li and Wong [2001] proposed a different approach which is used in their dChip software package [Li and Wong, 2003]. They proposed that the different probes in a probe set might have different affinities for the same gene, behavior that should remain constant across arrays (after normalization) and therefore the affinities can be modeled for a set of arrays. They proposed the following model:

$$PM_{ij} - MM_{ij} = \theta_i.\phi_j + \varepsilon_{ij} \tag{2.5}$$

In equation 2.5 [Li and Wong, 2001] the authors assume that the probe affinities (ϕ_j) influence the final signal in a multiplicative manner, and are constant across the arrays in the experiment. Therefore, fitting the model using multiple arrays allows to obtain θ_i for each array i, giving a summary statistic for the probe set and detecting probes that do not have a good fit to the model and may be defective [Li and Wong, 2001]. Some authors have reported that the Li and Wong [2001] model has strong mean variance dependence. [Irizarry et al., 2003b].

Another probe affinity modeling approach is the robust multichip average (RMA) [Irizarry et al., 2003b]. This method has been implemented in the library `affy` of Bioconductor [Gentleman et al., 2004] and is one of the most used for background correction and normalization of Affymetrix microarrays data sets. RMA is different from other methods since it uses only the PM probe intensities on each array, because of the high proportion of MM probes that have higher intensities than the corresponding PM probe. RMA is an additive model for the log transformation, background correction and normalization of the PM intensities, following the equation [Irizarry et al., 2003b]:

$$T(PM_{ij}) = e_i + \alpha_j + \varepsilon_{ij},\ i = 1, \ldots, I,\ j = 1, \ldots, J \tag{2.6}$$

where T represents the transformation (for background correction, normalization, and $\log_2$) of the PM intensities, e_i is the $\log_2$ scale expression value found on array $i = 1, \ldots, I$, α_j represents the log scale affinity effects for probes $j = 1, \ldots, J$, and ε_{ij} represents the random error. The robust linear fitting procedure using polish median [Tukey, 1977], is considered to estimate the log scale expression values e_i.

RMA has been modified to include different propensities of probes in background correction in order to undergo non-specific binding (NSB), type of background often underestimated. This modification of RMA, called GCRMA, uses probe sequence information to estimate probe affinity to the non-specific binding (NSB). Each probe affinity is obtained summarizing the sequence information for the base types (A,T,G or C) at each position (1-25) along the each probe. The parameters of the position-specific base contributions to the probe affinity are estimated in a NSB experiment in which only NSB but no gene-specific binding is expected. The probe affinities allow to estimate the relationship between the probe sequences and the quantity of NSB, through the estimation of the function:

$$NSB = h(\text{affinity}) \tag{2.7}$$

which estimated by fitting a loess curve to $\text{MM}_{\text{intensities}} \approx \text{MM}_{\text{affinities}}$ or using any list of negative controls (NC) instead of MM. The background adjusted intensity is computed as the posterior mean of specific binding given the observed intensities and the probe sequences. [Wu et al., 2004].

Normalization

Normalization is the adjustment for the differences of the overall probe intensities among arrays. Those differences are due to a variety of technical reasons like: different efficiencies of reverse transcription, labeling, hybridization reactions, physical problems with the arrays, reagent batch effects and laboratory conditions. Without normalization it is almost impossible to directly compare measurements from different microarrays. The simplest form of normalization involves multiplicatively transforming all the intensities on each array by a factor so that all arrays have the same median probe intensity. There are different linear and non-linear methods hat have been applied for normalizing Affymetrix microarrays.

Scale Normalization is a linear method that picks a baseline microarray, and all the other arrays in the experiment are scaled according to the mean intensity of the chosen microarray. It is like to use a selected baseline microarray

for fitting a linear regression with each of the other microarrays in the experiment, without an intercept term. Then, uses the fitted regression line as the normalizing relationship. Affymetrix uses a trimmed mean calculated after removing the lowest and highest 2% of the data.

Some non-linear normalization methods have been implemented and they tend to have better performance that the scaling ones. Some examples are cross-validated splines [Schadt et al., 2001], median lines, loess smoothers [Bolstad et al., 2003], quantile normalization [Bolstad et al., 2003] and variance stabilization and normalization (VSN) .[Huber et al., 2002, 2003].

Quantile normalization is one of the most reported non-parametric methods used for normalizing Affymetrix microarrays, which imposes the same empirical distribution of intensities to each array [Bolstad et al., 2003].

This method produces, for different data sets, the same distribution by transforming the quantiles of each set to have the same value. Bolstad et al. [2003] proposed the next algorithm for the implementation of quantile normalization:

1. Given I datasets of length N, create X of dimension $N \times I$, where each dataset is a column,
2. Set $d = (1/\sqrt{I}, \ldots, 1/\sqrt{I})$, a vector of dimension I,
3. Sort each column of X to give X_{sort},
4. Project each row of X_{sort} onto d to get X'_{sort},
5. Get X_{norm} by rearranging each column of X'_{sort} to have the same ordering as original X.

The projection is equivalent to taking the average of the quantiles in a particular row of X_{sort} and substituting each individual element in that row by this value, for example, if $q_j = (q_{j1}, \ldots, q_{jI})$ is a row in X_{sort}, then the corresponding row in X'_{sort} is given by:

$$q'_j = proj_d q_j = \frac{q_j.d}{d.d} d = \frac{1}{\sqrt{I}} \sum_{i=1}^{I} q_{ji} d = (\frac{1}{I} \sum_{i=1}^{I} q_{ji}, \ldots, \frac{1}{I} \sum_{i=1}^{I} q_{ji}) \qquad (2.8)$$

Another well known normalization method applied to Affymetrix microarrays is Cyclic Loess [Dudoit et al., 2002]. This approach is based upon the

MA plot, being M the difference between the log expression values of a specific probe under two different conditions and A the average of the same log expression values.

For any two arrays i, i', with probe intensities x_{ji} and $x_{ji'}$ where $j = 1, \ldots, J$ represents the probe, $M_j = \log_2(x_{ji}/x_{ji'})$ and $A_j = \frac{1}{2}\log_2(x_{ji} \times x_{ji'})$ are calculated and the normalization curve is fitted to the MA plot using loess method (locally weighted polynomial regression) [Bolstad et al., 2003]. Based on the normalization curve fitted, new values for M_j are calculated – $\hat{M}_j$ – and therefore the normalization adjustment is $M'_j = M_j - \hat{M}_j$. Adjusted probe intensities are given by $x'_{ji} = 2^{A_j + \frac{M'_j}{2}}$ and $x'_{ji'} = 2^{A_j - \frac{M'_j}{2}}$.

Cyclic Loess is carried out in a pairwise manner and when there are more than two arrays the adjustment for each two arrays is recorded. After looking at all pairs of arrays, a set of adjustments which may be applied to the set of arrays is: selected, applied and the process is repeated. Usually, one or two complete iterations through all pairwise combinations are needed to achieve the normalization. It is a time consuming method because it works in a pairwise manner. [Bolstad et al., 2003].

Summarization

Summarization is the final step in pre-processing Affymetrix microarray data. As it was mentioned before, in the Affymetrix microarray design each transcript is represented in the microarray by a set of 11 to 20 different probes (probe set) that hybridized specifically with that transcript product. Summarization is the process of combining the multiple probe intensities for each probe set to produce an expression value that will serve as an indicator of the level of expression of a specific transcript.

For probe set summarization in a single microarray it has been used the average of the differences – **avgDiff** – between PM and MM intensities of each probe corresponding to a specific probe set, method no longer recommended for use due to many flaws motivated by the fact of some MM probe intensities being higher than the corresponding PM probe intensities. The MAS 5.0 algorithm [Affymetrix, 2009] for summarization, in a single microarray, uses robust average in log scale (One-Step Tukey biweight) to combine the probe intensities of a specific probe set k, as follow:

$$\frac{1}{n_k}\sum_{j=1}^{n_k}\log_2(PM_j - IM_j) \tag{2.9}$$

where n_k (usually between 11 and 20) is the number of probes in probe set k.

In the case of probe set summarization using multiple microarrays, good examples are the mutiplicative model implemented in dChip software package [Li and Wong, 2003] and the robust multichip linear model fit on the log scale, implemented in RMA [Irizarry et al., 2003b].

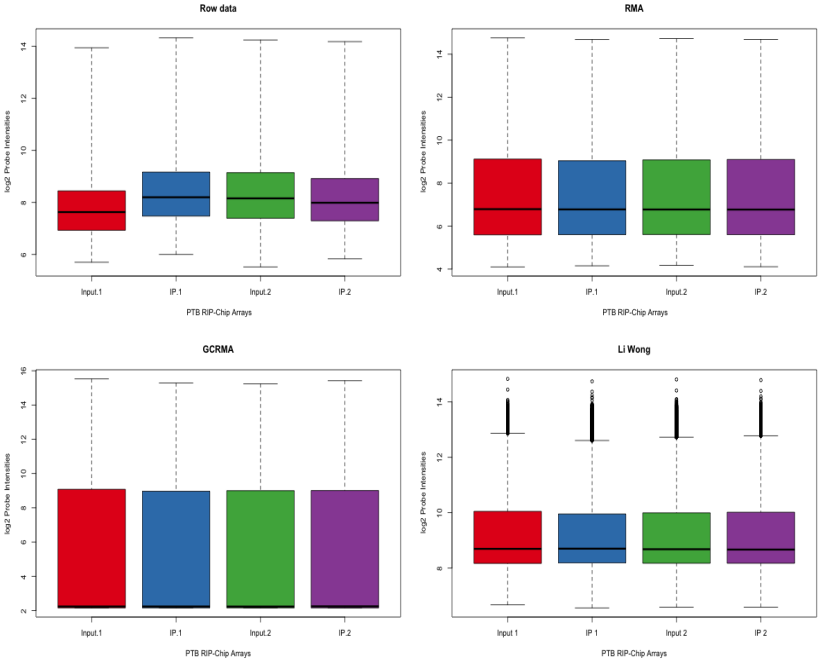
PTB and U2AF^{65} background correction, normalization and summarization

As was described previously on this chapter, there are many different ways for pre-processing Affymetrix microarrays data. Three different approaches were chosen for pre-processing PTB and U2AF^{65} RIP-Chip data [Gama-Carvalho et al., 2006], basically because they have been used in the analysis of RIP-Chip microarray data or because they were the most reported for the analysis of GeneChip® U133 2.0 plus Affymetrix microarrays.

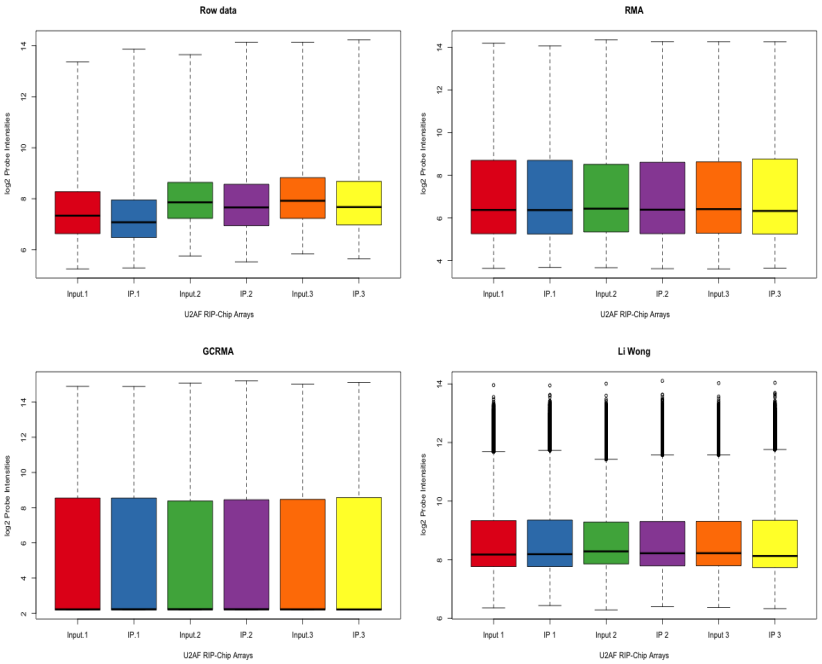
Two probe affinity modeling approaches were used for data pre-processing: RMA [Irizarry et al., 2003b] and its modification GCRMA [Wu et al., 2004]. These two methods consider only PM on raw intensity scale, quartile normalization, and probe set summarization done by the fitted model obtained via median polish algorithm. Also, the data were pre-processed using only PM , invariant set normalization method [Li and Wong, 2001] and the model-based method of Li and Wong [2001] for computing expression values.

The boxplots of the three different approaches: Li-Wongo's Model-Based, RMA and GCRMA applied to PTB and U2AF^{65} RIP-Chip data, are showed in the Figure 2.11. The library affy of Bioconductor allows to make each pre-processing approach by the R code shown in the appendix A.1.2.

Figure 2.11 shows that each of the three pre-processing methods produced different data distributions, where Li-Wong's Model-Based and RMA performed similarly while the GCRMA method was significantly different. These differences may lead to different results in the posterior enriched gene selection affecting the quality of the biological conclusions from the RIP-Chip experiment.



(a) PTB arrays



(b) U2AF^{65} arrays

Figure 2.11: Probe values distribution of PTB and U2AF^{65} RIP-Chip data before (raw data) and after Li-Wongo's Model-Based, RMA and GCRMA pre-processing.

2.2.2 High level analysis (*Enriched Gene Selection*)

The aim in most part of the microarray experiments is finding the differences between gene expression levels under two different cell conditions (*treatment* versus *control* or *disease* versus *healthy*), for example, genes responsible for diseases like cancer are expressed in different levels when the diseased is present and when it is not. In general, the conditions or states that are being compared have few differences between them and the lab experiment is running under the same conditions.

On another hand, the RIP-Chip allows to identify RNAs associated with specific RBPs, by comparing two different samples: the IP sample, (obtained by immunoprecipitation of RNAs bounded to a specific RBP, purification of this subset of RNAs, amplification by PCR, reverse transcription, labeling and hybridization), with the Input sample (where all mRNAs expressed by the HeLa cells are amplified by PCR, reverse transcribed, labeled and hybridized). It is evident that these two samples are processed in different lab conditions and that the proportions of mRNAs that are immunoprecipitated (IP sample) are higher than the Input sample. Which is why the genes that express the immunoprecipitated mRNAs are named enriched genes, and thanks to this can be detected by comparing the Input samples with the IP samples.

Many methods have been proposed for the determination of differentially expressed genes and some of then have been applied to RIP-Chip data. The goal of these methods is to identify genes that are differently expressed in the two samples or more, generally using a very low number of microarrays. An important feature of the methods is the capability to access the number of false positives [Irizarry et al., 2003b], because it allows to know whether the results obtained using those methods are reliable.

First studies considered as a criterion to detect genes of interest the fold-change (FC) [DeRisi et al., 1997], in which the gene expression levels under two different conditions are compared by the difference of the means of the transformed intensities (in general $\log_2$). The FC has the disadvantage of the cutoff values being arbitrary and consequently not providing a significant estimate for the observed changes in the presence of biological and experimental variation. Theses variations usually differ from gene to gene and are evident in some studies that have shown a false positive rate of 60-70%, which is the main reason for using statistical tests to access the differential expression of the genes.

The statistical tests look at various properties of the distributions of a gene's expression levels under different conditions, but the mean and the median are the most often considered through parametric tests, such as the t-test and moderated t-statistic [Smyth, 2004] and non-parametric tests, such as the Rank Products test [Breitling et al., 2004].

For assessing the differential expression of a gene under two experimental conditions, the t-test has been one of the most used methods, as well as ANOVA for more than two groups. For more general trend tests are frequently applied linear models, parametric method with convenient interpretability of the model parameter. The linear models are usually computed for each gene and allow to identify genes of interest, but due the lack of information of their biological relationship, their joint biological behavior is studied through more sophisticated modeling approaches like gene network models. in the next sections some of the most used will be briefly described.

t-test

The t-test is based on the standardized expression level, using the mean and the variance of the mean expression levels for each gene (of the treatment and control samples). When the expression level change is large in comparison with the variance of the mean difference, it is possible to assume that there is a real difference in gene levels and therefore they are differentially expressed. On the other hand, even if the difference is large, but the gene values have high variance, they will not be treated as differentially expressed.

The t-score is computed using the following statistic:

$$T = \frac{M_c - M_t}{\sqrt{\frac{S_c^2}{n_c} + \frac{S_t^2}{n_t}}} \tag{2.10}$$

where S_c^2 , S_t^2 are the variances in control and treatment samples respectively, M_c, M_t are the mean levels in control and treatment samples respectively and n_c, n_t are the dimension of control and treatment samples respectively.

It is possible to calculate a p-value for each t-score in order to access the chance for a false positive (the chance is the p-value itself). We should dismiss genes with p-values higher than some cutoff bound, being 0.01 and 0.05 the most commonly used. The calculation of the p-value depends as usually of the kind of samples. Considering that the variances are unknown and that

the hypothesis of normally distributed data holds, it is used the t-Student distribution. If the samples are very small or not normally distributed it is advisable to obtain the p-value by bootstrap or permutation/randomization methods. [Wit and McClure, 2004].

Lineal models

The central idea of this method, included in the library `limma` of Bioconductor, is to fit a linear model to the expression data for each gene in order to estimate the variability in the data. For statistical analysis and assessing differential expression, `limma` uses a moderate t-statistic and an empirical Bayesian method. This results in more stable inference and improved power, especially for experiments with small numbers of arrays. With Affymetrix microarrays, linear modeling is an ordinary ANOVA or multiple regression, where the is fitted for every gene. [Smyth, 2004].

This method assumes a linear model $E[y_g] = \alpha_g X$ where y_g contains the expression data for the gene g , X is the design matrix and α_g is a vector of coefficients. Here, y_g^T is the g-th row of the expression matrix and contains the log-intensities. The contrasts of interest are given by $\beta_g = C^T \alpha_g$ where C is the contrasts matrix. The coefficients component of the fitted model contains estimated values for the α_g. The contrast step, allows the fitted coefficients α_g to be compared in many ways depending of the questions to be answered, and estimate the β_g. [Smyth, 2004].

The library `limma` of Bioconductor [Smyth, 2004] produces a moderated t-statistic that uses a weighted average of s_0^2 (global variance estimate s_0^2 using all genes' variances) and s_g^2, instead of the single gene estimated variance s_g^2. This statistic can be shown to follow a t-distribution under the null hypothesis (also, under normality distribution) with the degrees of freedom depending on the data.

`limma` also implements an empirical Bayesian approach to select the differential expressed genes. The posterior odds of any particular gene g with respect to contrast β_g can be computed, given the joint distribution of $\tilde{T}_g$ and S_g^2, representing the mean and the standard deviation of the expression values for gene g, respectively. It is a target measure for each gene to be differentially expressed:

$$O_g = \frac{p(\beta_g \neq 0 | \tilde{T}_g, S_g^2)}{p(\beta_g = 0 | \tilde{T}_g, S_g^2)} \tag{2.11}$$

where $\tilde{T}_g$ and S_g^2 are independent and the distribution of S_g^2 does not depend on β_g . [Smyth, 2004].

Following Lonnstedt and Speed [2002], the logarithm of posterior odds:

$$B_g = \log O_g \tag{2.12}$$

can be calculated and it is useful for ranking genes with respect to their differential expression evidence. [Smyth, 2004].

Rank products

Rank products (RP) is a non-parametric method based on the ranks of the expression values changes. The method is based on the calculation of the rank products (RP) from replicate experiments, which is statistically rigorous and can be used to provide reliable significance thresholds to distinguish significantly regulated genes. At the same time, it provides a statistical way to determine the significance level for each gene and allows for the flexible control of the false discovery rate (FDR). [Breitling et al., 2004]

The assumptions made for the RP method are relatively weak: (1) relevant expression changes affect only a minority of genes, (2) measurements are independent between replicate arrays, (3) most changes are independent of each other, and (4) measurement variance is about equal for all genes. [Breitling et al., 2004]

In general, for each gene g in each replicate i with N_i genes, when $N_i = N$ for all replicates it is possible to calculate the combined probability as a rank product using the geometric mean rank [Breitling et al., 2004]:

$$\bar{r}_g^{up} = (\prod_{i=1}^{I} r_{gi}^{up})^{\frac{1}{I}}, \tag{2.13}$$

where r_{gi}^{up} is the position of the gene g in the list of genes in the i-th replicate sorted by decreasing FC, being $r_{gi}^{up} = 1$ when the gene is the most up-regulated. In the same way it is possible to calculate $\bar{r}_g^{down}$ but from the gene list sorted by increasing intensities. These $\bar{r}$ values (RP values) can be used to sort the genes according to the likelihood of observing them so high on the lists of differentially expressed genes just by chance.

The significance level of the RP values is calculated using a permutation-based estimation procedure that allows to determinate how likely a given

RP value, or better, is observed in a random experiment. For that, p permutations are generated from I rank lists of length N and the rank products for all the N genes for each permutation are calculated. After that, c_g is set to count how many times the rank products of the genes in the permutations are smaller or equal to the observed $\bar{r}_g^{up.or.down} = min(\bar{r}_g^{up}, \bar{r}_g^{down})$ and it is used for obtaining the expected value for the rank product by: $E(RP_g) = c_g/p$. Finally, FDR is calculated using the equation: $FDR_g = E(RP_g)/\bar{r}_g^{up..or..dwon}$. [Breitling et al., 2004]

The multiple testing problem

The Multiple Testing Problem arises when researchers try to identify which genes are differentially expressed between two conditions or between classes, using a few replicates, testing simultaneously a large number of hypothesis, making more likely to find extreme differential expression values, even if all null hypothesis are true [Wit and McClure, 2004]. For example, using t-tests for independent samples when 20,000 genes are examined, with the aim of identifying each differential expressed gene (under the null hypothesis of genes not being differentially expressed), may lead to 5% of the genes to have p-values <0.05. This would imply that 1000 genes would be identified as being significant at that p-value level, resulting in an expected number of 1000 false positives.

Various strategies have been proposed to deal with this problem: the significance analysis of microarrays (SAM) [Tusher et al., 2001]; the false discovery rate (FDR) [Benjamini and Hochberg, 1995]; the positive false discovery rate (pFDR) [Storey, 2002]; and the Q-value, a Bayesian posterior p-value or pFDR analogue of the p-value [Storey, 2003].

Bonferroni correction

The Bonferroni correction is a method for controlling the multiple testing problem associated with the familywise error rate (FWER) in microarray experiments. This conservative method classifies each gene with a p-value less than q/N as differential expressed, being q the desired FWER while N is the number of genes are being tested. For example, a cutoff of 0.00001 for the p-value should be chosen in order to have a FWER of 0.01 level over 1000 comparisons.

Bonferroni correction indeed reduces chance for false positives but may cause a large number of false negatives – genes that may be differential expressed

but were not select due to the small p-value threshold. [Wit and McClure, 2004].

False discovery rate

The false discovery rate (FDR) [Benjamini and Hochberg, 1995] is the expected proportion, $E(Q)$, of false positives among the genes declared as differential expressed:

$$Q = \frac{V}{V+S} \tag{2.14}$$

where V is the number of false positives, S the number of true positives and $V + S = R$ the number of genes declared to be significant (known). Q expected value is:

$$E(Q) = E\left(\frac{V}{V+S}\right) = E\left(\frac{V}{R}\right) \tag{2.15}$$

To control FDR, ensuring that it is less than a given threshold q, let $H_1, \cdots, H_N$ be the null hypotheses in increasing order of their p-values $p_1, \ldots, p_N$. For a given q, find the largest i, say j, such that

$$p_i \leq i * \frac{q}{N} \tag{2.16}$$

Then, reject (declare differentially expressed genes) all $H_{(i)}$ for $i = 1, \ldots, j$. It will guarantee that the false positives amount q is not exceeded. The FDR assumes that the gene expression of genes on the microarray is independent but in many cases their expressions are correlated.

Enriched gene selection for PTB and U2AF^{65} RIP-Chip pre-processed data

The enriched gene selection was performed using the expression data obtained by the RMA method applied to PTB and U2AF^{65} RIP-Chip data. As it was mentioned before, RMA performed similar to Li-Wong's model-based and better than the GCRMA pre-processing methods. RMA is one of the most cited pre-processing methods given its accuracy and performance on Affymetrix microarray data when it is associated with the gene selection methods used in this section. [Kadota et al., 2009].

Three different enriched gene selection methods with different approaches were applied to the PTB and U2AF^{65} pre-processed data (R code showed in the appendix A.1.2):

1. Fold change (FC $>$ 1.5), t-test and FDR ($<$ 0.05) by using dChip software [Li and Wong, 2003].

2. Fit linear models and the statistics B calculated by the empirical Bayesian approach. Moreover, multiple testing procedures (p-value $<$ 0.05) by using the Bioconductor library `limma` [Smyth, 2004].

3. The non-parametric test Rank Products with FDR $<$ 0.05 by using the R library `RankProd` [Hong et al., 2006].

These three differentially expressed gene selection methods are commonly used and were applied with FDR $<$ 0.05 as threshold for selecting significant enriched genes. Table 2.3 shows the differences in the number of enriched genes selected by each method, being Rank Products (RP) the most conservative, selecting only 134 and 449 genes as enriched genes that bind their expressed mRNAs to the RBPs PTB and U2AF^{65}, respectively, when more than 3000 were expected [Gama-Carvalho et al., 2006].

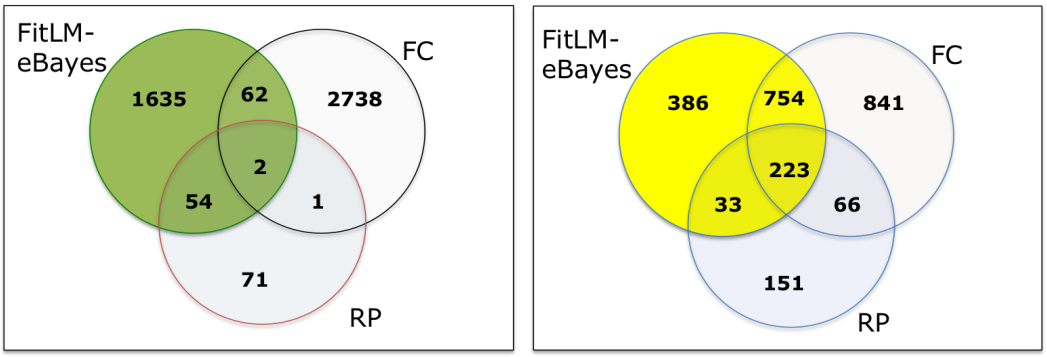
This means that each applied method has different false positive and false negative associated levels. If they are too conservative like RP, the false negative error is likely higher that in the others and vice versa. In general, the applied methods may have a high false negative percentage.

Table 2.3: Enriched genes select that bind their expressed mRNAs to PTB and U2AF^{65}.

Method	**PTB Genes**	**U2AF^{65} Genes**
Linear model - eBayes	1753	1396
Fold Change t-statistic	2803	1884
Rank Products	134	449

Figure 2.12 shows differences not only in the number of genes selected but also in the identification of those genes.

Figure 2.12 shows that only 2 and 223 enriched genes were selected simultaneously by the three methods applied to PTB and U2AF^{65} RIP-Chip pre-processed data respectively and most of the enriched genes were selected from one specific method. 93.3% and 27,65% of the genes were selected by fitting a linear model – eBayes, applied to PTB and U2AF^{65} RIP-Chip pre-processed data, respectively.



(a) PTB **(b) U2AF^{65}**

Figure 2.12: Venn diagram showing enrichmed genes selected by fitting a linear model – eByaes, Fold change (FC) – *t*-test, and Rank Products (RP).

The differences in the selected genes in others microarray experiments by applying different gene selection methods, have been already reported [Breitling et al., 2004] and could involve problems in the confidence of the results. This is very important when those results must support biological conclusions and further experiments.

Is important to point out that in PTB and U2AF^{65} RIP-Chip experiments more that 20% of the genes should be enriched and therefore, the assumption of a few expression differences between experimental conditions is not fulfilled, likely affecting the performance of this methods when applied for enriched gene selection.

Chapter 3

New methods for RIP-Chip data analysis

3.1 Introduction

The systematic identification of RNA targets has provided clues to unsuspected functions of well-known RBPs and 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 immunoprecipitation 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 Chromatin immunoprecipitation on chip (ChIP-Chip) [Hanlon and Lieb, 2004] and is called RNA immunoprecipitation on chip (RIP-Chip).

Chromatin immunoprecipitation on chip (ChIP-Chip) is a popular technique to study the in vivo targets of DNA-binding proteins at the genome level. In the first step of this technique, the target protein (TP) is cross-linked by a formaldehyde fixation with the DNA site where it is binding to in vivo environment. It is followed by cell lysis and the DNA fragmentation using sonication. Then, double-stranded DNA fragments (1 kb or less in length) cross-linked to the TP are immunoprecipitated out of the set of DNA fragments by using a specific antibody to the TP and after purification, the DNA complex are split and the DNA strands are purified. The next step are amplification, denaturation, labeling of the single-stranded DNA fragments with a fluorescent tag and finally, hybridization to a DNA microarray. [Hanlon and Lieb, 2004].

It is evident how similar the ChIP-Chip and the RIP-Chip techniques are, both performed in different ways to the DNA expression experiments. In particular, the RIP-Chip technique the immunoprecipitated (IP) 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. As it was presented in Chapter 2, the main problem is related to the pre-processing methods, like RMA [Irizarry et al., 2003b], dChip [Li and Wong, 2003], etc., implemented with the assumption that there are few differences in the expression levels between conditions, which is not true in the RIP-Chip experiments, and there are not specific methods developed for this kind of experiments.

Due the lack of methods implemented to pre-processed RIP-Chip data and its similarities with the ChIP-Chip technique, in particular when in the hybridization step are used Affymetrix tiling microarrays (closely related to the Affymetrix GeneChip microarrays used in RIP-Chip), a pre-processing method developed in this thesis, applied to U2AF^{65} RIP-Chip microarray data, is presented in this chapter, following the spirit of the algorithm proposed by Johnson et al. [2006] for analyzing ChIP-Chip data obtained by sample hybridization with Affymetrix tiling microarrays.

This chapter presents a simple linear model that estimates the baseline probe behavior using the probe sequence for background correction of Affymetrix microarray data. This model is based on the previous research made by Johnson et al. [2006], Naef and Magnasco [2003] and Wu et al. [2004], that proposed different unspecific hybridization models, and additionally this new model as in the Johnson et al. [2006] algorithm uses the information associated with each microarray for its own pre-processing. Basically, a linear model is fitted (using an iterative procedure) in order to predict each probe intensity and standardize it afterwards. In this way it takes the advantage of high density of the Affymetrix microarrays, where each U133 2.0 plus microarray contains more than 1.2 million probes, and the overfit of the models is not a problem. Thus, the assumption of a small number of differences usually applied for the pre-processing methods is not necessary.

After pre-processing, the enriched genes selection is a crucial step of the of RIP-Chip data analysis and any other kind of microarray data analysis, therefore two new methodologies are presented in this chapter. The first one is inspired in the Johnson et al. [2006] methodology designed for the ChIP-Chip data analysis. The idea is to summarize the probe set intensity on the basis of the standardized probe intensities, resulting in a statistic which we

named *ENRval* – enriched value.

The motivation for developing the second method, was due to the fact of the methods for Affymetrix data analysis using a unique value to represent the behavior of the levels of the probes within a probe set. However, it is generally known that the variability within the probe set is high, even if the probes intensities have been corrected. An alternative methodology is presented for enriched gene selection, which takes into account the probe intensity variability in each probe set. This is a non-parametric test based on ranks, and a permutation test is performed for obtaining a FDR for each standardized probe intensity.

Finally, the evaluation of the methods presented here is made using the data of U133 Spike-In experiment [Cope et al., 2004].

3.2 Sequence-specific affinity models estimation

The probe behavior estimation model proposed in this thesis was motivated by the idea of taking advantage of the more than 1.2 million 25-mer oligonucleotide probes present on each Affymetrix U133 Plus 2.0 array, in order to obtain an accurate and robust prediction of probe sequence effects on non-specific hybridization background. Finally, for their own pre-processing not requiring the assumption of small differences between microarray conditions.

The Naef and Magnasco [2003] research on the labeling and effective binding in oligonucleotide microarrays serves as motivation for the implementation of a different approach to estimate sequence-specific affinity models for pre-processing Affymetrix microarrays . Naef and Magnasco [2003] proposed a model where they found that the variation within a PM probe set is sequence dependent, by fitting brightness B of PM probes ([RNA] concentration: the median of the PM brightnesses) against their own sequence composition:

$$\ln\left(\frac{B}{[RNA]}\right) = \sum_{lk} S_{lk} A_{lk} = \sum_{l\alpha} S_{l\alpha} \sum_{k} A_{lk} P_{k\alpha} \tag{3.1}$$

where $l = A, C, G, T$ is the letter index and $k = 1, \ldots, 25$ the position along the 25-mer probe; S_{lk} is a Boolean variable equal to 1 if the probe sequence has letter l at site k and 0 otherwise; and thus A_{lk}'s are per-site, per-letter

affinities. Note that $\sum_l S_{lk} = 1$ for all k. In the last equality Naef and Magnasco [2003] use an expansion of the spatial dependence in orthonormal polynomials $P_{k\alpha}$ on set $\{1, \ldots, 25\}$.

Naef and Magnasco [2003] did not include nearest-neighbor interactions along the transcript length in their model because they found it does not improve the fit enough to justify the increase in the number of parameters, but found that the insertion of position-dependent affinities produces a strong improvement in their model.

Wu et al. [2004] introduced the Non-Specific Background (NBS) adjustment in their GCRMA method for pre-processing Affymetrix microarrays. They fitted the model of the Equation 3.1 to log intensity data using a spline with 5 degrees of freedom, instead of the polynomial of degree 3 proposed by Naef and Magnasco [2003], and used these affinity estimates to describe NSB noise in GCRMA assuming that:

$$\begin{aligned} PM &= O_{PM} + N_{PM} + E \\ MM &= O_{MM} + N_{MM} + \Phi E \end{aligned} \tag{3.2}$$

where O represents the optical noise, N represents NSB noise, E is a quantity proportional to RNA expression (the quantity of interest) and the parameter $\Phi \in (0, 1)$ accounts for the fact that for some probe pairs the MM detects signal. The large amount of data allows a very precise estimate of model parameters and the GCRMA model, as a background adjustment procedure, is formalized as the statistical problem of predicting E given that the observed PM and MM and the affinity estimates to describe NSB noise [Wu et al., 2004]. It is important to annotate that the GCRMA assumes that there are a small number of differences between microarrays under different conditions, and uses all microarrays in the experiment for the pre-processing step.

Johnson et al. [2006] inspired in the previous work of Naef and Magnasco [2003] and Wu et al. [2004], propose an algorithm called Model-based Analysis of Tiling-arrays (MAT), where they use a probe behavior model fitting as part of the algorithm for pre-processing, independently for each Affymetrix tiling microarray. They propose the following tiling array probe affinity model [Johnson et al., 2006]:

$$\log(PM_i) = \alpha n_{iT} + \sum_{j=1}^{25} \sum_{k \in \{A,C,G\}} \beta_{jk}\ I_{ijk} + \sum_{k \in \{A,C,G,T\}} \gamma_k n_{ik}^2\ + \delta \log(c_i) + \varepsilon_i, \tag{3.3}$$

where:

- PM_i Perfect Match (PM) probe intensity value of the probe i;
- n_{ik} is the number of times nucleotide k occurring in probe i;
- α is the baseline value (intercept or constant) based on the number of T nucleotides on the probe, e.g., 25α is the baseline when the probe sequence is a run of 25 T nucleotides;
- 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 (except T, which is already modeled in α) at each position j;
- γ_k is the effect of nucleotide count squared;
- c_i is the number of times that the sequence of probe i appears in the genome. Affymetrix tiling array libraries provide the 25-mer sequence of every probe, which we mapped to the non-repeat-masked newest (May 2004) version of the human genome assembly;
- δ is the effect of the log of the probe copy number;
- ε_i is the probe-specific error term, assumed to follow a normal distribution.

The Model is fitted by ordinary least squares to each array separately using all of the probes on a tiling array. After parameter estimation, the model can predict the probe i baseline intensity, $\hat{m}_i$, given its probe sequence and copy number of its sequence in the genomic regions that is represented in the tiling microarray. After that it is possible to correct and standardize the probe values, eliminating the need of data normalization. [Johnson et al., 2006]

Attending the technical similarities between RIP-Chip and ChIP-Chip techniques and the pre-processing methods limitations, 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 the previous Sequence-Specific probe behavior models for gene expression microarrays proposed by Naef and Magnasco [2003] and Wu et al. [2004]), it is proposed the following linear Sequence-Specific affinity model for background correction of gene expression Affymetrix arrays [Barreto-Hernandez et al., 2011]:

$$\log(PM_i) = \sum_{j=1}^{25} \sum_{k\in\{A,C,G,T\}} \beta_{jk}\ I_{ijk} + \sum_{k\in\{A,C,G,T\}} \gamma_k n_{ik}^2\ + \varepsilon_i \tag{3.4}$$

where:

- PM_i Perfect Match probe intensity value;
- n_{jk} is the number of times nucleotide k occurring 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.

The linear sequence-specific affinity model for background correction (Equation 3.4) was applied to each microarray in the U2AF^{65} RIP-Chip experiment data (R code shown in the appendix A.2.1). This final model was obtained after testing different approaches with base in the Equation 3.3 which results are showed in the Table 3.1.

Given the differences between gene expression and tiling microarrays, the number of times that the sequence of probe i appears in the genome (c_i) and the effect of the probe copy number (δ) in the Equation 3.3 were not included in the model (Equation 3.4). Due to that, the gene expression Affymetrix microarrays are designed using probes that only recognize one specific transcript and therefore, their probe signals correspond to a one transcript for each probe and not multiple genome fragments that may increase the magnitude of the probe signal like could occur with tiling microarrays.

As it is shown in the Table 3.1, when the model setup is done using the effect of nucleotide count squared (γ_k) had better fit (R^2= 0.02681) than when the model was fitted without it (R^2= 0.01853), although none of them represents

Table 3.1: Results for different combinations of the presence or not of γ_k, $n_{jT}\alpha$ and thymine (T) affinity terms in the linear affinity model [Equation 3.4] fitted to the U2AF^{65} Input sample 1.

Model including:	**Multiple R^2**	**p-value**
γ_k, $n_{jT}\alpha$ and no T affinities	0.02681	$< 2.2 \times 10^{-16}$
no γ_k, $n_{jT}\alpha$ and no T affinities	0.01853	$< 2.2 \times 10^{-16}$
no γ_k, no $n_{jT}\alpha$ and no T affinities	0.01853	$< 2.2 \times 10^{-16}$
γ_k, no $n_{jT}\alpha$ and T affinities	0.02681	$< 2.2 \times 10^{-16}$

a good model. Additionally, the effect of the thymine count ($n_{jT}\alpha$) were removed from the model and the thymine affinities effect was included due the equal model fit results (Table 3.1).

The linear model (Equation 3.4), 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 adenine, thymine, cytosine, guanine (ATCG) nucleotides count. This model was fitted using a random sample of 100,000 sequences (the calculations using the R code placed in appendix A.2.1 are computing demanding in terms of memory and time) and accounted for 3 to 4% of the variation in the arrays (based on the multiple R^2 of the model).

The linear sequence-specific affinity model (Equation 3.4) was modified trying to improve its fit by taking into account the nucleotide position-specific interaction, replacing the expression:

$$\sum_{j=1}^{25} \sum_{k\in\{A,C,G,T\}} \beta_{jk}\ I_{ijk} \tag{3.5}$$

by

$$\sum_{j=1}^{26-it} \sum_{k_1,\ldots,k_{it}\in\{A,C,T,G\}} \beta_{jk_1\ldots k_{it}}\ I_{ijk_1}\ldots I_{i(j+it-1)k_{it}} \tag{3.6}$$

where it is the interaction nucleotide number with the nucleotide k in the probe i. This modification originates the model:

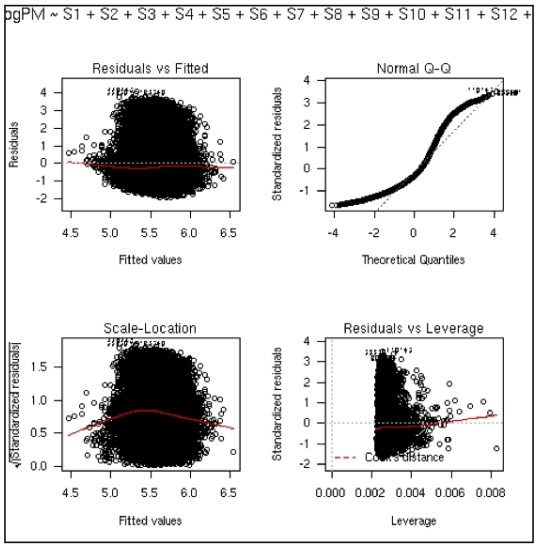
$$\log(PM_i) = \sum_{j=1}^{26-it} \sum_{k_1,\ldots,k_{it}\in\{A,C,T,G\}} \beta_{jk_1\ldots k_{it}}\ I_{ijk_1}\ldots I_{i(j+it-1)k_{it}} + \\ + \sum_{k\in\{A,C,G,T\}} \gamma_k n_{ik}^2 + \varepsilon_i \qquad (3.7)$$

where, in order to improve the model fit, the contribution of each nucleotide in each sequence position to the binding affinity (Equation 3.5) was replaced by contribution of each nucleotide in each sequence position to the binding affinity plus its interaction with the neighbor nucleotides (Equation 3.6).

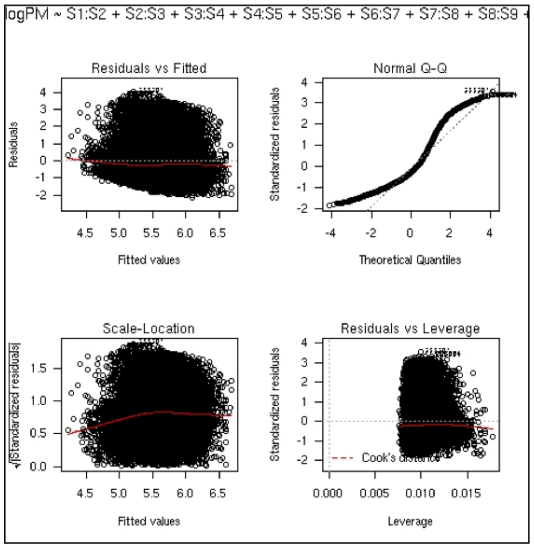
Table 3.2 shows the results after applying this model (R code shown in the appendix A.2.1) with different setups: 2, 3 and 4 interacting nucleotides in comparison with the original no-interaction model. We obtained (Figure 3.1) better fitting results when the number of interacting nucleotides is increased (more than 58% for 4 interacting nucleotides) which is evident in the Figure 3.1d, where the Q-Q plot (Quantile-Quantile plot) of the model where interact 4 nucleotides shows the better fit. 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 20,000 and 8000 probes, respectively, from 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 has less than 10,000 probes.

Table 3.2: Results of the no-interaction model (Equation 3.4), and the nucleotide interaction (2, 3 and 4 interacting nucleotides) model (Equation 3.7) applied to the U2AF^{65} Input sample 1.

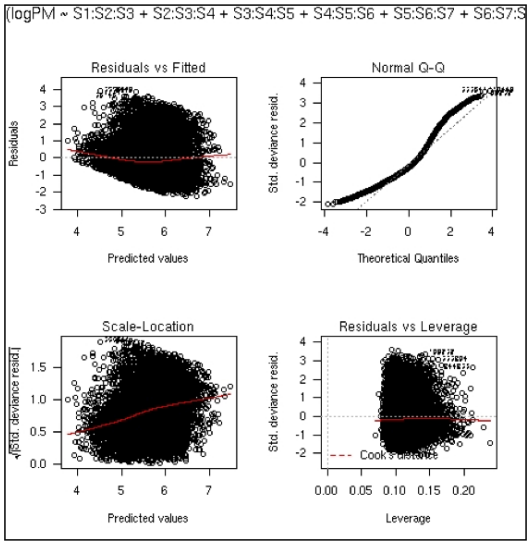
Model	**Multiple R^2**	**p-value**
No-interaction	0.02681	$< 2.2 \times 10^{-16}$
2 Nucleotides interaction	0.06918	$< 2.2 \times 10^{-16}$
3 Nucleotides interaction	0.124	$< 2.2 \times 10^{-16}$
4 Nucleotides interaction: 8K sample	0.5861	7.48×10^{-11}



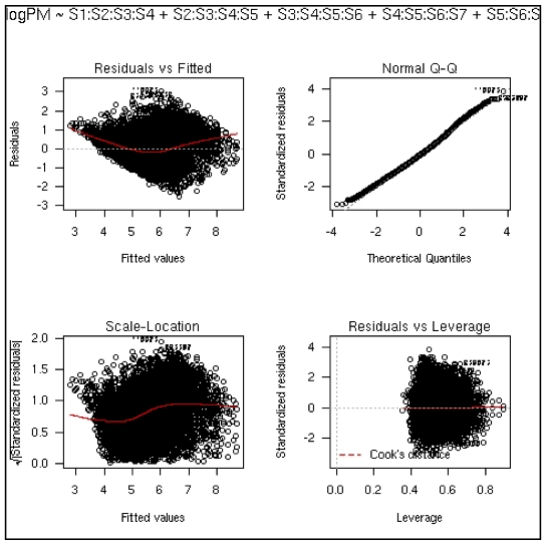
(a) No-interaction



(b) 2 Nucleotides interaction



(c) 3 Nucleotides interaction



(d) 4 Nucleotides interaction

Figure 3.1: Residuals plots of the fitted model, Equation 3.7, using different numbers of interacting nucleotides applied to the U2AF^{65} Input sample 1.

The 3 nucleotides interaction model was used for trying to improve the correlation between predicted probe intensities and the observed values, by the implementation of the following iterative process:

1. Linear model parameters estimation.
2. If the difference between multiple R squared (R^2) of the fitted model and the multiple R^2 of the fitted model in the previous iteration is $>$ 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.

With the idea of modeling the non-specific hybridization in terms of probe sequence, the iterative process was implemented for removing outliers after fitting the 3 nucleotides interaction model (Equation 3.7), using an initial random sample of 20,000 probes sequences. The process stops when there are no significant differences between the last fitted model and the previous one which includes the removed outliers ($R^2_v - R^2_{v-1} < 0.001$, where v is the iteration number).

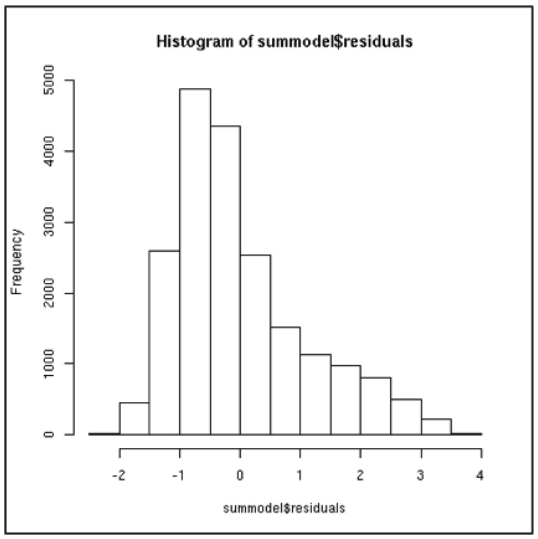
The linear model parameter estimation and the posterior probe baseline intensity estimation, $\hat{m}_i$, are calculated using the `lm` R library (code shown in the appendix A.2.1) and applied to each microarray data set independently.

After the iterative process the model accounted for 42 to 58% of the variation in the arrays explained (Table 3.3), 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.

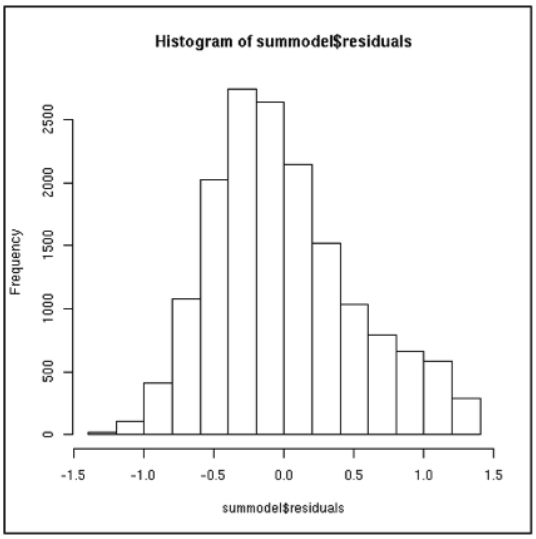
The residuals (Figures 3.2, 3.3 and 3.4) of the fitted model after the iterative process, on each U2AF^{65} RIP-Chip microarrays, were approximately normally distributed.

3.3 Probe standardization

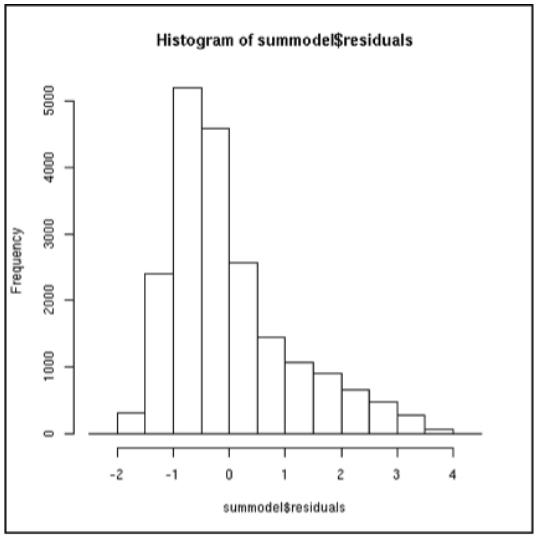
After predicting each probe intensity $\hat{m}_i$ using the iterative process, a standardization for each probe is made on each array independently, using the Equation [Johnson et al., 2006]:



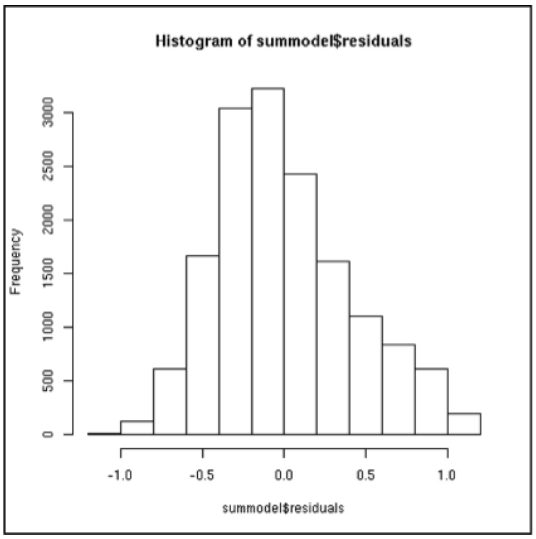
(a) Input 1 iteration 1



(b) Input 1 final iteration

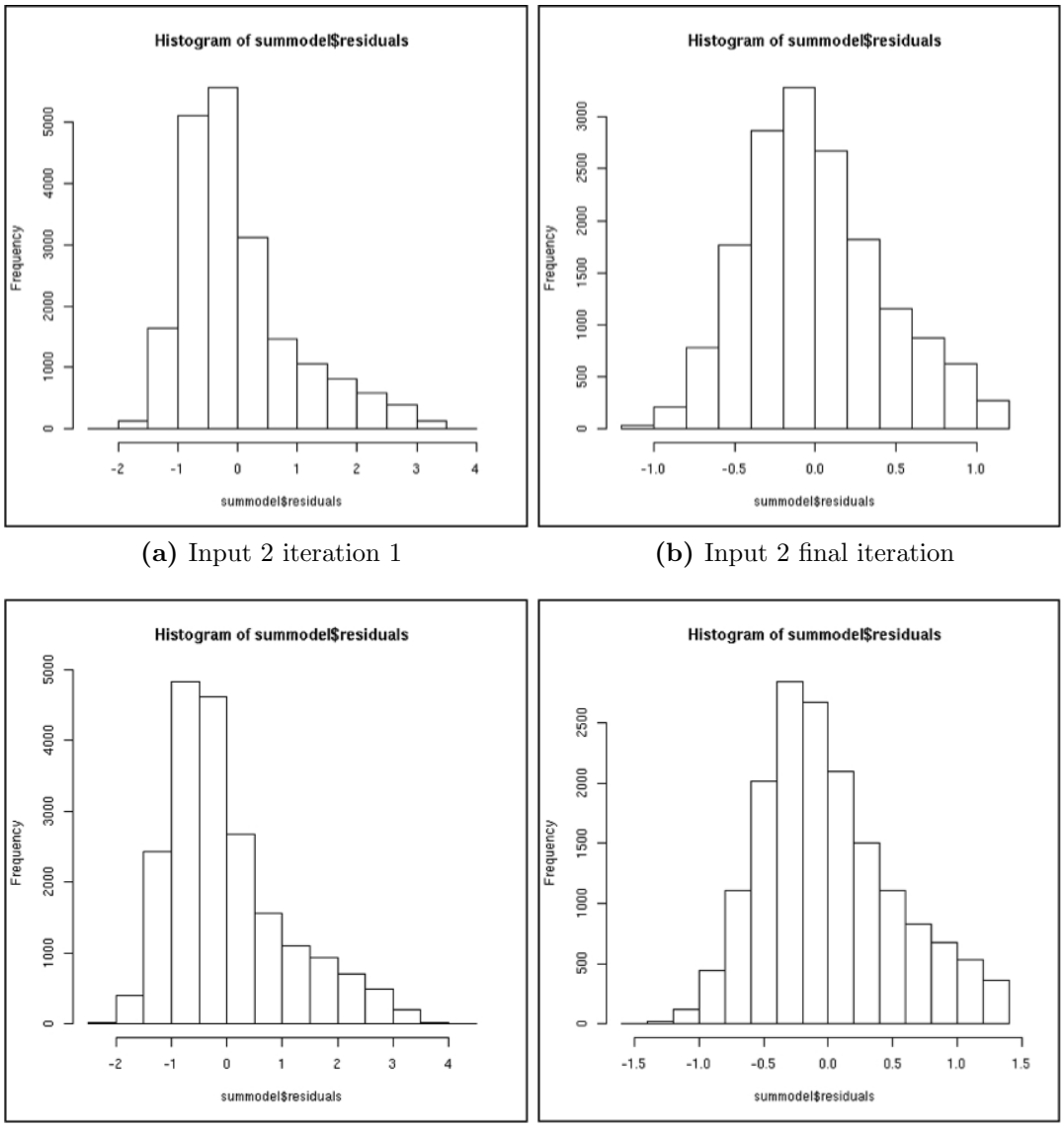


(c) IP 1 iteration 1



(d) IP 1 final iteration

Figure 3.2: Residuals histograms of the iterative process fitting the 3 nucleotides interaction linear model to the U2AF^{65} RIP-Chip samples. Part A



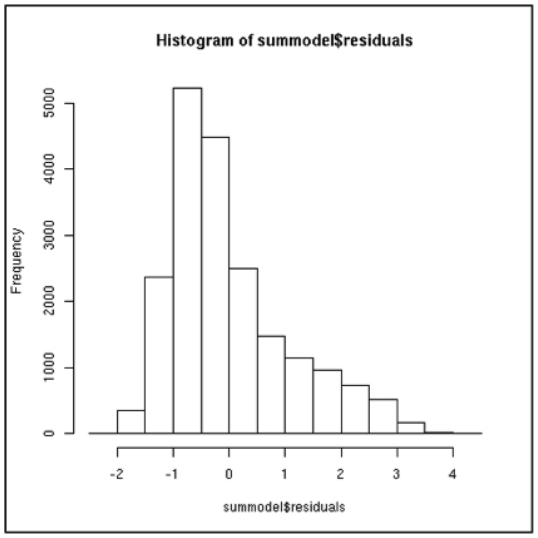
(a) Input 2 iteration 1

(b) Input 2 final iteration

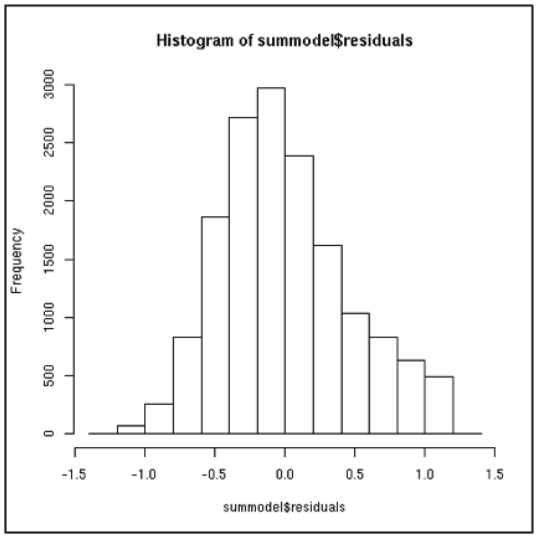
(c) IP 2 iteration 1

(d) IP 2 final iteration

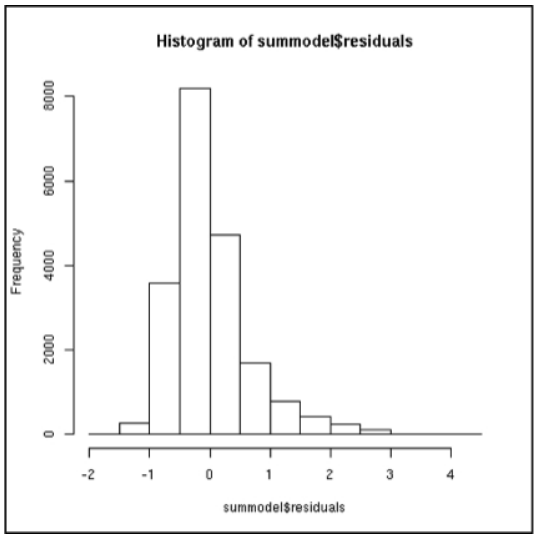
Figure 3.3: Residuals histograms of the iterative process fitting the 3 nucleotides interaction linear model to the U2AF^{65} RIP-Chip samples. Part B



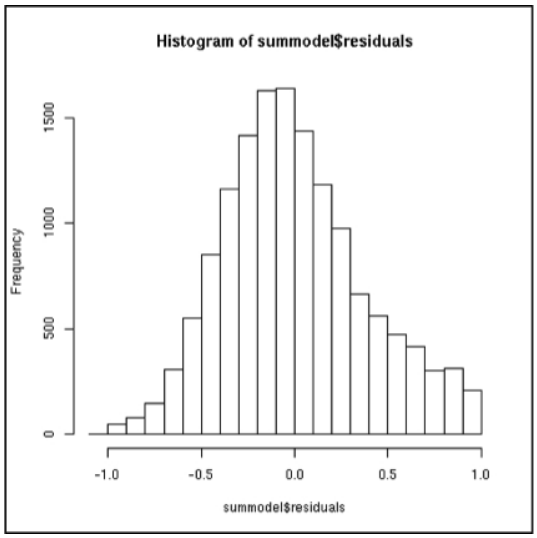
(a) Input 3 iteration 1



(b) Input 3 final iteration



(c) IP 3 iteration 1



(d) IP 3 final iteration

Figure 3.4: Residuals histograms of the iterative process fitting the 3 nucleotides interaction linear model to the U2AF^{65} RIP-Chip samples. Part C

Table 3.3: Results of the iterative process using 3 nucleotides interaction linear model, applied to Gama-Carvalho et al. (2006) U2AF RIP-Chip experiment.

Microarray	Iteration number	R^2	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

$$t_i = \frac{\log(PM_i) - \hat{m}_i}{SD_{i\_affinity\ bin}} \tag{3.8}$$

where $\hat{m}_i$ is the predicted probe i baseline intensity and $SD_{i\_affinity\ bin}$ is the observed sample variance estimated within each affinity bin.

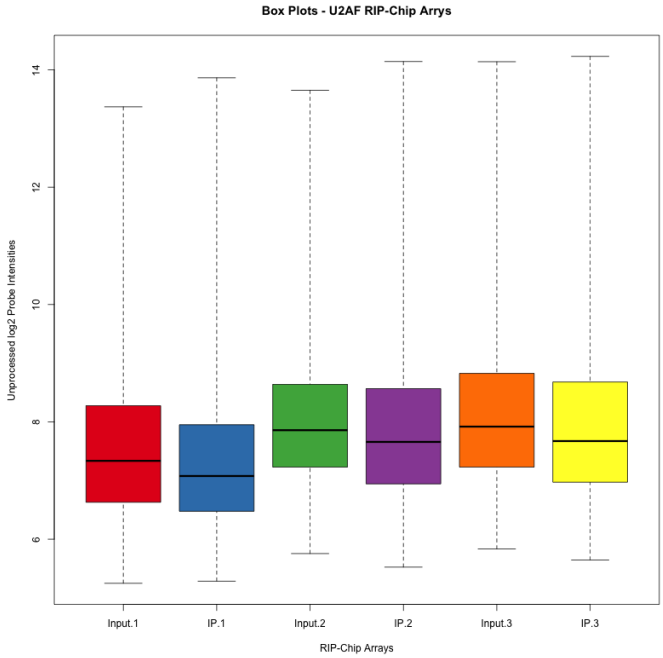
Following Johnson et al. [2006] suggestion, using the R code shown in the appendix A.2.2, all probes on an array were divided into bins containing 3000 probes predicted to have similar intensities $\hat{m}_i$. Each probe's variance was estimated from the sample variance of all probes in its specific bin. Figure 3.5 shows that normalization of U2AF[65] RIP-Chip microarray data is evidently required. However, after probes standardization (Figure 3.5), the t-values calculated using the Equation 3.8 do not require further normalization and could be compared directly.

3.4 Probe set summarization

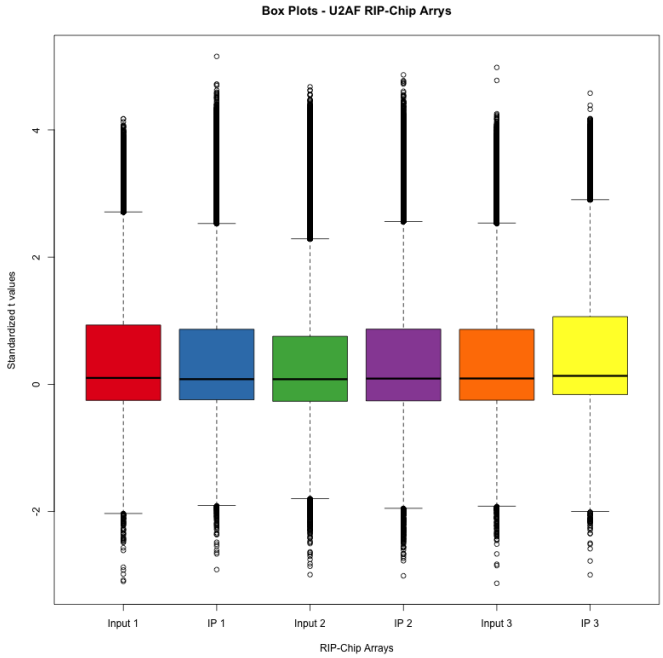
A probe set score ($PSsco$) is calculated as follows [Johnson et al., 2006]:

$$PSsco = \sqrt{n_p} \times TM_s \tag{3.9}$$

where TM_s is the trimmed-mean of the t-values of the probe set s, obtained by removing the top 10% and bottom 10% t-values, and $n_p \in \{11 \times$



(a) PM intensities U2AF^{65} RIP-Chip microarrays



(b) Standardized t-values U2AF^{65} RIP-Chip microarrays

Figure 3.5: Boxplots of PM intensities and standardized t-values U2AF^{65} RIP-Chip microarrays.

$0.8, \ldots, 20 \times 0.8\}$ is the number of probes used to calculate TM_s.

For multiple replicates, $PSsco$ are calculated pooling all t-values for a specific probe set across all replicates. The higher the number of replicates, the higher will be the prediction confidence.

3.5 Enriched value (*ENRval*)

We propose the probe set enriched value (*ENRval*), to be calculated as follows:

$$ENRval = PSsco_{IP} - PSsco_{Input} \tag{3.10}$$

When there are more than 2 experimental replicates, the *ENRvals* are divided by SD_{Input} (standard deviation of the 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 around the *ENRvals* median and the negative values correspond to the non enriched genes. This symmetry is imposed by forcing the second half distribution to be the "mirror" of the lower half distribution (*ENRvals* smaller than the median), under the assumption of these *ENRvals* corresponding to non enriched values.

The R code shown in the appendix A.2.3 is used for processing the standardized probe values t_i of U2AF^{65} and to make the calculation of probe set scores ($PSsco$), enriched values (*ENRvals*) and p-values. Table 3.4 shows the number of genes selected using the proposed enriched values (*ENRvals*) at different p-value levels applied to 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 the expected number to be associated with U2AF^{65} (3200), sharing a high percentage (66%) of the genes reported in Gama-Carvalho et al. [2006]. Selected genes also include the genes: ahsa2, gas2l1, cdkn1b, ZNF174 and AHSA2, verified by real time PCR (RT-PCR) as targets of the U2AF^{65} by the same authors.

It is possible that some of those differences were due to the application of normalization methods without having into account the differences in the

Table 3.4: Enriched genes selected using *ENRvals* at different p-values level. *ENRvals* calculated from the standardized t-values.

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

data distribution between samples.

ENRvals allows a consistent enriched gene selection, but it is important to benchmark the proposed method, which is presented in the Section 3.7 of this chapter, using the U133 Spike-In data set – standard microarray experiment results used in the development and comparison of differential expression analysis methods. [Cope et al., 2004]

It is evident that there is space for improving the unspecific hybridization prediction and in this way improving the background correction. Most important, is that there is the possibility of obtaining better approaches of using the data of the high density microarrays for its own pre-processing, facilitating the microarrays comparison when technics like RIP-Chip are applied.

3.6 New rank based approach for enriched gene selection

ChIP-Chip experiments have used Tiling Affymetrix microarrays, which are designed in a different way to the GeneChip® Affymetrix microarrays used in RIP-Chip technology. The Tiling microarrays probes are short fragments (25 pb in Affymetrix) designed to cover the entire genome or contiguous regions of the genome.

Johnson et al. [2006] methodology implements the *PSsco* value to represent the consecutive hybridized probes values in DNA window of ≈ 600 pb. These values are used to determine the sites where a specific protein binds to the DNA, main purpose of the ChIP-Chip technology.

In another hand, the GeneChip® Affymetrix microarrays, as was explained before, contain a probe set of 11 up to 20 different probes representing transcripts (gene, RNA, etc) and in the case of the GeneChip® U133 2.0 plus, there are more than 54,000 of these transcripts.

An alternative methodology is presented in this section for enriched gene selection, based in the standardized probe intensity (t-values) variability in each probe set, corresponding to a specific gene, in both immunoprecipitated (IP) and Input samples, instead of the summarization of each probe set and the posterior *ENRval* calculation.

The first step is to calculate the difference between the t-values of each probe pair, IP and Input samples, as follows:

$$dt_{ijk} = t_{ijk(IP)} - t_{ijk(Input)} \tag{3.11}$$

for:

$i = 1, \ldots, M$	M is the total array number -IP or Input-;
$j = 1, \ldots, N$	N is the total probe set number in the array;
$k = 1, \ldots, n$	n is the probe number in each probe set.

These dt_{ijk} values will be used for enriched gene selection using a rank based approach, where different ranks for each gene and their the total rank RT are calculated.

The following step is to calculate the $\overline{dt}_{ij}$ mean of each probe set (gene) in each experiment as follows:

$$\overline{dt}_{ij} = \frac{\sum_{k=1}^{n} dt_{ijk}}{n} \tag{3.12}$$

for:

$$i = 1, \ldots, M$$
$$j = 1, \ldots, N$$

and the $\overline{dt}_{j}$ mean of each probe set across all experiments as follows:

$$\overline{dt}_{j} = \frac{\sum_{i=1}^{M} \sum_{k=1}^{n} dt_{ijk}}{M * n} \tag{3.13}$$

for:

$$j = 1, \ldots, N$$

Then, the mean of the differences between dt_{ijk} and its probe set mean, $\overline{dt}_{ij}$, across all experiments pairs are calculated as follows:

$$Tpss_j = \frac{\sum_{i=1}^{M} \sum_{k=1}^{n} |dt_{ijk} - \overline{dt}_{ij}|}{M * n} \tag{3.14}$$

for:

$$j = 1, \ldots, N$$

and the mean of the differences between the means of a probe set across all experiments $\overline{dt}_{ij}$ and their mean $\overline{dt}_j$, for each gene as follows:

$$Tps_j = \frac{\sum_{i=1}^{M} |\overline{dt}_{ij} - \overline{dt}_j|}{M} \tag{3.15}$$

for:

$$j = 1, \ldots, N$$

Finally four different rankings are proposed corresponding to different variability measure approaches, having into account the variability among probes in different experiments and among probe set levels, that are calculate as follows:

$$\begin{array}{lcl}
RP_{1_j} & \rightarrow & \text{position of } \overline{dt}_j - Tpss_j \\
 & & \text{sorting } \{\overline{dt}_{j'} - Tpss_{j'}\}_{j'=1,\ldots,N} \text{ by increasing values,} \\
RP_{2_j} & \rightarrow & \text{position of } \overline{dt}_j + Tpss_j \\
 & & \text{sorting } \{\overline{dt}_{j'} + Tpss_{j'}\}_{j'=1,\ldots,N} \text{ by increasing values,} \\
RP_{3_j} & \rightarrow & \text{position of } \overline{dt}_j - Tps_j \\
 & & \text{sorting } \{\overline{dt}_{j'} - Tps_{j'}\}_{j'=1,\ldots,N} \text{ by increasing values,} \\
RP_{4_j} & \rightarrow & \text{position of } \overline{dt}_j + Tps_j \\
 & & \text{sorting } \{\overline{dt}_{j'} + Tps_{j'}\}_{j'=1,\ldots,N} \text{ by increasing values.}
\end{array} \tag{3.16}$$

for:

$$j = 1, \ldots, N$$

and the total rank is as follows:

$$RP_{T_j} = \prod_{l=1}^{4} RP_{l_j} \tag{3.17}$$

for:

$$j = 1, \ldots, N$$

RP_{T_j} values are put in increasing order, where the top values (smallest) correspond to the enriched genes.

The expected number of times each RP_{T_j} is greater than other $RP_{T.}$ corresponds to the E-value for gene j. If one considers m permutations – each permutation i' consisting of permutating separately each RP_{l_j} and recalculating RP_{T_j}, this time called $RP_{T.Permut\_i'_j}$ – the E-value can be approximated as follows:

$$E - value_j = \frac{z_j}{m} \tag{3.18}$$

for:

$$j = 1, \ldots, N$$

where:

$$\begin{aligned} &m \text{ is the total number of permutations and} \\ &z_j = \#\left\{j' = 1, \ldots, N : RP_{T.Permut\_i'_{j'}} \leq RP_{T_j} \text{ for } i' = 1, \ldots, m\right\} \end{aligned}$$

The false discovered rate (FDR) for each probe set is calculated as follows:

$$FDR_j = \frac{E - value_j}{rank_j} \tag{3.19}$$

for:

$$j = 1, \ldots, N$$

and

$rank_j$ – position of transcripts (genes) j in the list of all genes sorted by increasing RP_{T_j} value, corresponding to the number of genes accepted as significantly enriched.

The R code shown in the appendix A.2.4 is used for processing the standardized probe values t_i of U2AF^{65} and to make the calculation of this new rank based approach for enriched gene selection.

The different rank lists and the RP_{Tj} for each transcript j and its corresponding E-values and FDR were calculated. Table 3.5 shows the number of genes selected at different E-value levels and FDR < 0.05.

Table 3.5: Enriched genes selected using the new rank based approach at different E-values level. Rank indexes calculated from the standardized t-values.

E-value	Enriched Genes
< 0.001	483
< 0.01	1219
< 0.05	2530
< 0.1	3618

The number of enriched genes selected is 2530 for a E-value < 0.05, a very small number in comparison to the expected number of mRNAs to be binding with U2AF^{65} (≈20% of the transcripts). However, these results share a higher percentage (87%) of the genes reported in [Gama-Carvalho et al., 2006] than *ENRval* method (66%) with the same reported genes. Concerning this non-parametric enriched gene selection, *ENRval* method shares 78.3% of the genes selected.

The list of enriched genes include the genes: ahsa2, gas2l1, cdkn1b, ZNF174 and AHSA2, verified by real time PCR (RT-PCR) as targets of the U2AF^{65} by Gama-Carvalho et al. [2006].

It is possible that some of those differences are due to the application of normalization methods without having into account the differences in the data distribution between samples, but also the effect of the high variability of the probe intensity levels in each probe set.

This new non parametric method for enriched gene selection is consistent with previous reposts [Gama-Carvalho et al., 2006], but as it was mentioned for the *ENRval* method it is important to benchmark this method. That is what will be done in the next section.

3.7 Methods benchmark

The developed methods were benchmarked using the data of U133 Spike-In experiment, which was made some years ago and is used in the development

and comparison of differential expression analysis methods [Cope et al., 2004].

The U133A Spike-In dataset was produced by an experiment that has three technical replicates of 14 separate hybridizations of 42 spiked transcripts selected, based on their absent expression in background total RNA isolated from a HeLa cell line (ATCC CCL-13). Thirty of the spiked transcripts correspond to cDNA clones isolated from total RNAs of a lymphoblast cell line (and are not expressed in the HeLa cell line), eight of the spiked transcripts are made from artificial sequences, and the remaining four spiked transcripts are Affymetrix eukaryotic controls are that available as part of a poly A spike control kit. There are fourteen groups of three genes each. The concentration of each gene within a group is spiked at the same concentration. A cyclic latin square was used for the group concentrations (Tables 3.6 and 3.7).

Specifically, the dosing pattern (Tables 3.6 and 3.7) for Group 1 across 14 arrays is 0, 1/8, 1/4, 1/2, 1, 2, 4, . . . , 256, and 512 pM, for Group 2 is 1/8, 1/4, 1/2, 1, . . . , 512, and 0 pM, and so on until Group 14, which has a pattern of 512, 0, 1/8, 1/4, . . . , 128, and 256 pM. Each pattern appears on three replicate arrays, yielding a total of 42 arrays.

There are 11 probes per probe set for each of 38 genes and 20 probes per probe set for the four Affymetrix eukaryotic controls. Additional information about this experiment is available at the Affymetrix website, `www.affymetrix.com/support/technical/sample data/datasets.affx`.

For benchmarking the methods presented in this chapter, three different Spike-In experiment hybridizations are selected: hybridizations 1, 8 and 14 and used to simulate RIP-Chip enrichment differences (see the values in red, Tables 3.6 and 3.7). Both combinations (hybridization 1 as Input and 8 as IP; 1 as IP and 14 as Input) were processed independently using the R code shown in the Appendices A.2.3 and A.2.4 .

The results on Table 3.8 showed that the developed method has better accuracy when the gene expression differences between conditions (Input and IP) are high, as one can see in the combination: 1 (as Input) and 8 (as IP); where 20/21 and 19/21 Spike-in genes were classified correctly by *ENRval* and the new ranks approach methods respectively. When the concentration differences between conditions (IP and Input) are small, the accuracy decrease as one can see in the combination: 14 (as Input) and 1 (as IP), where only 21/39 and 19/39 Spike-in genes were classified correctly by *ENRval* and the new ranks approach methods respectively.

Table 3.6: Spike-In U133 experiments 1-7 description.

Group ID	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Gene ID	203508_at 204563_at 204513_s_at	204205_at 204959_at 207655_s_at	204836_at 205291_at 209795_at	207777_s_at 204912_at 205569_at	207160_at 205692_s_at 212827_at	209606_at 205267_at 204417_at	205398_s_at 209734_at 209354_at
EXP 1	**0**	**0.125**	**0.25**	**0.5**	**1**	**2**	**4**
EXP 2	0.125	0.25	0.5	1	2	4	8
EXP 3	0.25	0.5	1	2	4	8	16
EXP 4	0.5	1	2	4	8	16	32
EXP 5	1	2	4	8	16	32	64
EXP 6	2	4	8	16	32	64	128
EXP 7	4	8	16	32	64	128	256
EXP 8	**8**	**16**	**32**	**64**	**128**	**256**	**512**
EXP 9	16	32	64	128	256	512	0
EXP 10	32	64	128	256	512	0	0.125
EXP 11	64	128	256	512	0	0.125	0.25
EXP 12	128	256	512	0	0.125	0.25	0.5
EXP 13	256	512	0	0.125	0.25	0.5	1
EXP 14	**512**	**0**	**0.125**	**0.25**	**0.5**	**1**	**2**

Table 3.7: Spike-In U133 experiments 8 -14 description.

Group ID	**8**	**9**	**10**	**11**	**12**	**13**	**14**
	206060_s_at	207641_at	203471_s_at	AFFX-r2-TagA_at	AFFX-r2-TagD_at	AFFX-r2-TagG_at	AFFX-LysX-3_at
Gene ID	205790_at	207540_s_at	204951_at	AFFX-r2-TagB_at	AFFX-r2-TagE_at	AFFX-r2-TagH_at	AFFX-PheX-3_at
	200665_s_at	204430_s_at	207968_s_at	AFFX-r2-TagC_at	AFFX-r2-TagF_at	AFFX-DapX-3_at	AFFX-ThrX-3_at
EXP 1	**8**	**16**	**32**	**64**	**128**	**256**	**512**
EXP 2	16	32	64	128	256	512	0
EXP 3	32	64	128	256	512	0	0.125
EXP 4	64	128	256	512	0	0.125	0.25
EXP 5	128	256	512	0	0.125	0.25	0.5
EXP 6	256	512	0	0.125	0.25	0.5	1
EXP 7	512	0	0.125	0.25	0.5	1	2
EXP 8	**0**	**0.125**	**0.25**	**0.5**	**1**	**2**	**4**
EXP 9	0.125	0.25	0.5	1	2	4	8
EXP 10	0.25	0.5	1	2	4	8	16
EXP 11	0.5	1	2	4	8	16	32
EXP 12	1	2	4	8	16	32	64
EXP 13	2	4	8	16	32	64	128
EXP 14	**4**	**8**	**16**	**32**	**64**	**128**	**256**

Clearly, these two methods have good accuracy when the differences in concentrations of the transcripts is high, which is the common behavior of the enriched genes in the RIP-Chip experiments. However, they require adjustments to make them more sensitive and precise. Those adjustments could be made through: improving the background correction model; applying other types of trimming for the summarization and calculation of the *ENRvals*; or finding other ways to measure the variability of the probe set trying to improve the sensibility of the new ranks based approach for enriched gene selection.

Additionally, further work could consider the mixture of part of the methodology proposed in this thesis with other methods, which is common in microarray data analysis.

Table 3.8: Benchmark results of the *ENRval* and new ranks based approach for enriched gene selection.

Spike-In	**Exp1**	**Exp8**	**Exp14**	***ENRval***	RP_T	***ENRval***	RP_T
				Ex 8-1	**Ex 8-1**	**Ex 1-14**	**Ex 1-14**
203508_at	0	8	512	22	22	22298	22298
204563_at	0	8	512	21	23	22299	22299
204513_s_at	0	8	512	19	17	22300	22300
204205_at	0,125	16	0	18	21	6054	8643
204959_at	0,125	16	0	16	3	12943	15242
207655_s_at	0,125	16	0	13	7	18259	15410
204836_at	0,25	32	0,125	14	16	3399	5370
205291_at	0,25	32	0,125	20	20	9846	10247
209795_at	0,25	32	0,125	17	15	8728	4973
207777_s_at	0,5	64	0,25	10	8	2341	5809
204912_at	0,5	64	0,25	15	12	8264	6766
205569_at	0,5	64	0,25	9	6	4855	6113
207160_at	1	128	0,5	6	13	3511	4591
205692_s_at	1	128	0,5	7	11	37	49
212827_at	1	128	0,5	11	18	5001	3905
209606_at	2	256	1	8	14	62	133
205267_at	2	256	1	3	4	74	116
204417_at	2	256	1	4	2	33	22
205398_s_at	4	512	2	5	10	36	87
209734_at	4	512	2	1	1	506	408
209354_at	4	512	2	2	5	125	296
206060_s_at	8	0	4	22263	22263	30	37
205790_at	8	0	4	22260	22262	58	44
200665_s_at	8	0	4	22264	22264	27	35
207641_at	16	0,125	8	22261	22260	39	128
207540_s_at	16	0,125	8	22266	22268	44	31
204430_s_at	16	0,125	8	22265	22266	32	28
203471_s_at	32	0,25	16	22267	22265	10	24
204951_at	32	0,25	16	22268	22267	12	14
207968_s_at	32	0,25	16	22262	22261	42	184
AFFX-r2-TagA_at	64	0,5	32	22270	22272	20	32
AFFX-r2-TagB_at	64	0,5	32	22269	22271	19	30
AFFX-r2-TagC_at	64	0,5	32	22298	22300	9	8
AFFX-r2-TagD_at	128	1	64	22283	22291	1	2
AFFX-r2-TagE_at	128	1	64	22295	22285	2	10
AFFX-r2-TagF_at	128	1	64	22286	22284	6	5
AFFX-r2-TagG_at	256	2	128	22291	22278	7	11
AFFX-r2-TagH_at	256	2	128	22288	22296	14	6
AFFX-DapX-3_at	256	2	128	22289	22277	4	12
AFFX-LysX-3_at	512	4	256	22271	22274	35	50
AFFX-PheX-3_at	512	4	256	22280	22275	23	27
AFFX-ThrX-3_at	512	4	256	22278	22297	21	7

Chapter 4

Conclusions and remarks

4.1 Conclusions

4.1.1 RIP-Chip data analysis

- Microarray quality controls must be applied to Affymetrix microarrays and serve as an indication wether the microarrays quality is good, or if one or more microarray data sets should be excluded in order to have more confident final results. For the PTB and U2AF^{65} RIP-Chip microarrays all the chips were considered to have good quality.

- The three pre-processing methods applied to PTB and U2AF^{65} RIP-Chip microarray data produced different data distributions, that may lead to different results in further enriched gene selection affecting the quality of the biological conclusions from the RIP-Chip experiment.

- The three differentially expressed gene selection methods: fitting a linear model - eBayes, Fold Change t-statistic and Rank Products (RP), show differences in the number of enriched genes selected, being Rank Products the most conservative. Therefore, each applied method has different false positive and false negative associated levels. In general, the applied methods may have a high false negative percentage.

- These three differentially expressed gene selection methods selected simultaneously a few amount of common genes and most of the enriched genes were selected for one specific method, which involves problems in the accuracy of the results, important aspect when those results must support biological conclusions and further experiments. It is important to point out that in the RIP-Chip experiments more that 20% of the

genes should be enriched and therefore, the assumption of few expression differences between experimental conditions is not fulfilled, likely affecting the performance of this methods when applied for enriched gene selection.

4.1.2 New methods for RIP-Chip data analysis

- A linear Sequence-Specific affinity model for background correction of gene expression Affymetrix arrays [Barreto-Hernandez et al., 2011] was implemented. It 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 adenine, thymine, cytosine, guanine (ATCG) nucleotides count. This model is computing demanding in terms of memory and time and accounted for 3 to 4% of the variation in the arrays.

- The contribution of each nucleotide in each sequence position to the binding affinity plus its interaction with the neighbor nucleotides were included in the Sequence-Specific affinity model obtaining better fitting results when the number of interacting nucleotides is increased, but rises the computing demanding in terms of memory and time.

- The iterative process implemented using the Sequence-Specific affinity model with 3 nucleotides interaction accounted for 42 to 58% of the variation in the arrays explained, improved the model fit. Mainly, this process which residuals were approximately normally distributed removes high intensity values 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.

- The t-values obtained after standardization for each probe made on each array independently do not require further normalization and could be compared directly.

- A probe set score (*PSsco*) was proposed as a summarization value of the probe set in the Affymetrix microarray and calculated pooling all t-values from a specific probe set across all replicates. The higher the number of replicates, the higher will be the prediction confidence.

- The probe set enriched value (*ENRval*) and its p-value was proposed and allowed a consistent enriched gene selection using the U2AF^{65} RIP-Chip microarray data. Selected genes include some verified experimentally as targets of the U2AF^{65}.

- A new rank based approach was implemented for enriched gene selection based in taking into account the standardized probe intensity (t-values) variability in each probe set, corresponding to a specific gene, instead of the usual summarization of each probe set and the posterior *ENRval* calculation. The method selected less enriched genes that *ENRval* method, but shared a higher percentage of the genes reported in [Gama-Carvalho et al., 2006] than the genes share by the *ENRval*. The list of enriched genes includes the genes verified by real time PCR (RT-PCR) as targets of the U2AF^{65} by Gama-Carvalho et al. [2006].

- The method proposed uses the standardized probe intensity levels from the individual information of each array, making possible to compare arrays from different conditions in a more adequate way for RIP-Chip experiments.

- The developed methods were benchmarked using the data of U133 Spike-In experiment [Cope et al., 2004], showing that they have better accuracy when the gene expression differences are high for both of them: *ENRval* and the new ranks approach methods. When the concentration differences are small, the accuracy decreases.

4.2 Some remarks

- 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 probe t-values.

- The enriched gene selection method requires adjustments to make it more sensitive and precise. Those adjustments could be made to improve the background correction model, to apply other types of trimming for the summarization and calculation of the *ENRvals* or try other ways to measure the variability of the probe set trying to improve the sensibility of the new ranks approach for enriched gene selection.

- The further work could consider the mixture of part of the methodology proposed in this thesis with other methods, which are commonly applied in microarray data analysis.

- It is evident that there is space for improving the unspecific hybridization prediction and in this way improving the background correction. Most important, is that there is the possibility of obtaining better approaches of using the data of the high density microarrays for its own pre-processing, facilitating the microarrays comparison.

Appendix A

R Scripts

A.1 RIP-Chip data analysis

A.1.1 Quality Assesment

Affymetrix .CEL files reading and image plots

```
#  Affymetrix libraries
#
library(affy)
library(affyPLM)
library('RColorBrewer')
bs.cols <- brewer.pal (6, 'Set1')

#  Affymetrix read data
#

da=ReadAffy()
colnames(exprs(da))<-  c("Input 1","IP 1","Input 2",
"IP 2","Input 3","IP 3")
sampleNames(da) <- c("Input 1","IP 1","Input 2",
"IP 2","Input 3","IP 3")
ls.names <- sampleNames(da)

#  Affymetrix quality control
#
#  Images
```

```
Sys.setenv("DISPLAY"=":0")
#jpeg(file="chipsU.jpeg",bg="white", quality=100,
 res=10000)
png(file="chipsU.png", width = 800, height = 800 )
par(mfrow=c(3,2))
image (da)
dev.off()
```

Boxplot and density plot

```
# Construct color boxplots
Sys.setenv("DISPLAY"=":0")
png(file="unp-boxplotU.png", width = 800,
height = 800 )
boxplot(da, col=bs.cols, ylab='Unprocessed log2
Probe Intensities',
xlab='RIP-Chip Arrays')
title(main="Box Plots - U2AF RIP-Chip Arrys")
dev.off()

# Construct density plots
Sys.setenv("DISPLAY"=":0")
png(file="unp-DensityU.png", width = 800,
height = 800 )
hist (da, col=bs.cols, lty=1, xlab="Log2 Intensities",
lwd=3)
legend (10,.4, legend=ls.names, lty=1,  col=bs.cols,
lwd=3)
title(main="Histograms (Density) Plots - U2AF
RIP-Chip Arrys")
dev.off()
```

MAplots

```
# MA plots
Sys.setenv("DISPLAY"=":0")
png(file="unMAplotU.png", width = 1200,
height = 800 )
par(mfrow=c(2, 3))
MAplot(da)
dev.off()
```

RNA degradations plots

```
#
# RNA degradation
#
rnadeg=AffyRNAdeg(da)
Sys.setenv("DISPLAY"=":0")
png(file="RNAdegraU.png", width = 800, height = 800 )
plotAffyRNAdeg(rnadeg,col=bs.cols)
legend (7, 6, legend=ls.names, lty=1,  col=bs.cols, lwd=3)
dev.off()
write.table(summaryAffyRNAdeg(rnadeg,
signif.digits = 3),file="RNA_Degradation_Sumary.txt")
```

PLM plots

```
#
#PROBE-LEVELMODELS
#
PLMda <- fitPLM(da)

# residuals PLM
#
Sys.setenv("DISPLAY"=":0")
png(file="WeightPLMU.png", width = 1200, height = 800 )
par(mfrow=c(2, 3))
image (PLMda, type="resids")
dev.off()

# weights PLM
#
Sys.setenv("DISPLAY"=":0")
png(file="ResiduesPLMU.png", width = 1200, height = 800 )
par(mfrow=c(2, 3))
image (PLMda, type="weights")
dev.off()

# Signed Residulas PLM
#
Sys.setenv("DISPLAY"=":0")
png(file="SignResiPLMU.png", width = 1200, height = 800 )
```

```
par(mfrow=c(2, 3))
image (PLMda, type="sign.resids")
dev.off()

# RLE boxplots PLM
#
Sys.setenv("DISPLAY"=":0")
png(file="RLEU.png", width = 800, height = 800 )
RLE(PLMda, col=bs.cols, main="U2AF RLE Plot")
dev.off()

# NUSE boxplots PLM
#
Sys.setenv("DISPLAY"=":0")
png(file="NUSEU.png", width = 800, height = 800 )
NUSE(PLMda, col=bs.cols, main="U2AF NUSE Plot")
dev.off()
```

A.1.2 Preprocessing

Backgroung correction, Normalization and summarization

```
# Preprocesing RMA
#
da2=rma(da)
e2=exprs(da2)

# Preprocesing GCRMA
#
da4 <- gcrma(da)

# Preprocesing Li Wong
#
da5 <- expresso(da,bg.correct=FALSE, normalize.method=
"invariantset", pmcorrect.method="pmonly", summary.method="liwong")
da6=exprs(da5)

#Boxplots
Sys.setenv("DISPLAY"=":0")
png(file="boxplotNU.png", width = 1200, height = 800 )
par(mfrow=c(2, 2))
```

```
boxplot(da, col=bs.cols, ylab=' log2 Probe Intensities', xlab=
' U2AF RIP-Chip Arrays')
title(main="Row data")
boxplot(da2, col=bs.cols, ylab=' log2 Probe Intensities', xlab=
' U2AF RIP-Chip Arrays')
title(main="RMA")
boxplot(da4, col=bs.cols, ylab='log2 Probe Intensities', xlab=
'U2AF RIP-Chip Arrays')
title(main="GCRMA")
boxplot(log2(da6), col=bs.cols, ylab=' log2 Probe Intensities',
xlab='U2AF RIP-Chip Arrays')
title(main="Li Wong")
dev.off()
```

Enriched gene selection

```
# Limma- eBayes
library(limma)

design <- model.matrix(~ -1+factor(c(1,2,1,2,1,2)))
colnames(design) <- c("C", "TU")
fit <- lmFit(e2, design)
contrast.matrix <- makeContrasts(TU-C, levels=design)
fit2 <- contrasts.fit(fit, contrast.matrix)
fit2 <- eBayes(fit2) results <- decideTests(fit2)

# Rank Products
library(RankProd)
e2.origen=c(1,1,1,1,1,1)
e2.cl=c(0,1,0,1,0,1)
RP.adv.out <- RPadvance(e2,e2.cl,e2.origen,num.perm=100,
gene.names=rownames(e2),logged=TRUE,rand=123)
topRP=topGene(RP.adv.out,cutoff=0.05,logged=TRUE,logbase=2,
gene.names=row)
```

A.2 New methods for RIP-Chip data analysis

A.2.1 Sequence-specific affinity model

Sequence-specific affinity model

```
# Model setup

resample <- function(x, size, ...)
       if(length(x) <= 1) { if(!missing(size) && size == 0) x[FALSE]
       else x        } else sample(x, size, ...)

#Read tables
logPM1 <- read.table("logPM.txt")

AN1 <- read.table("A_C_F_pm.txt")
TN1 <- read.table("T_C_F_pm.txt")

GN1 <- read.table("G_C_F_pm.txt")
CN1 <- read.table("C_C_F_pm.txt")
S251 <- read.table("SEQ_ARRAY_1_25_pm_C_F.txt")

logPM <- logPM1[,j]

AN <- AN1[,1]^2
TN <- TN1[,1]^2
GN <- GN1[,1]^2
CN <- CN1[,1]^2
S1 <- S251[,1]
S2 <- S251[,2]
S3 <- S251[,3]
S4 <- S251[,4]
S5 <- S251[,5]
S6 <- S251[,6]
S7 <- S251[,7]
S8 <- S251[,8]
S9 <- S251[,9]
S10 <- S251[,10]
S11 <- S251[,11]
S12 <- S251[,12]
S13 <- S251[,13]
S14 <- S251[,14]
S15 <- S251[,15]
```

```
S16 <- S251[,16]
S17 <- S251[,17]
S18 <- S251[,18]
S19 <- S251[,19]
S20 <- S251[,20]
S21 <- S251[,21]
S22 <- S251[,22]
S23 <- S251[,23]
S24 <- S251[,24]
S25 <- S251[,25]

n=100000
Set300 <- sample(ss<-1:length(logPM),n)

logPM <- logPM[Set300]
# para reducir memoria
AN <- AN[Set300]
TN <- TN[Set300]
GN <- GN[Set300]
CN <- CN[Set300]

S1 <- S1[Set300]
S2 <- S2[Set300]
S3 <- S3[Set300]
S4 <- S4[Set300]
S5 <- S5[Set300]
S6 <- S6[Set300]
S7 <- S7[Set300]
S8 <- S8[Set300]
S9 <- S9[Set300]
S10 <- S10[Set300]
S11 <- S11[Set300]
S12 <- S12[Set300]
S13 <- S13[Set300]
S14 <- S14[Set300]
S15 <- S15[Set300]
S16 <- S16[Set300]
S17 <- S17[Set300]
S18 <- S18[Set300]
S19 <- S19[Set300]
S20 <- S20[Set300]
```

```
S21 <- S21[Set300]
S22 <- S22[Set300]
S23 <- S23[Set300]
S24 <- S24[Set300]
S25 <- S25[Set300]

#final model setup
#?k, njT ? and no T affinities
model <- lm(logPM ~  TN + S1 + S2 + S3 + S4 + S5 + S6  + S7
+ S8 + S9 + S10 + S11+ S12  + S13 + S14  + S15  + S16  + S17
 + S18  + S19 + S20  + S21 + S22 + S23+S24+S25 + AN + TN
 + GN + CN)
#no ?k, njT ? and no T affinities
model <- lm(logPM ~  TN + S1 + S2 + S3 + S4 + S5 + S6  + S7
+ S8 + S9 + S10 + S11+ S12  + S13 + S14  + S15  + S16  + S17
+ S18  + S19 + S20  + S21 + S22 + S23 +S24+S25)
#no ?k, no njT ? and no T affinities
model <- lm(logPM ~   S1 + S2 + S3 + S4 + S5 + S6  + S7 + S8
 + S9 + S10 + S11+ S12  + S13 + S14  + S15  + S16  + S17 + S18
 + S19 + S20  + S21 + S22 + S23 +S24+S25)
#?k, no njT ? and T affinities
model <- lm(logPM ~  S1 + S2 + S3 + S4 + S5 + S6  + S7 + S8
+ S9 + S10 + S11+ S12  + S13 + S14  + S15  + S16  + S17 + S18
 + S19 + S20  + S21 + S22 + S23 +S24+S25 + AN + TN + GN + CN)

summary(model)
Sys.setenv("DISPLAY"=":0")
jpeg(file="ALL_MM_seq_noTN_3NUcleo.jpeg",bg="white",
 quality=100, res=10000)
opar <- par(mfrow = c(2,2), oma = c(0, 0, 1.1, 0))
plot(model, las = 1)      # Residuals, Fitted, ...
par(opar)
dev.off()
```

Sequence interaction model setup

```
# different probe interactions. 1 to 4 probes interactions
model <- lm(logPM ~ S1 +S2 + S3 + S4 + S5 + S6 + S7 + S8
+ S9 + S10 + S11 + S12 + S13 +S14 + S15 + S16 + S17 + S18
 + S19 + S20 + S21 + S22 + S23 + S24 + S25 + AN2 + TN2
 + GN2 + CN2,subset=Set300)
```

```
summary(model)
Sys.setenv("DISPLAY"=":0")
jpeg(file="ALL_PM_seq_noTN.jpeg",bg="white", quality=100,
 res=10000)
opar <- par(mfrow = c(2,2), oma = c(0, 0, 1.1, 0))
plot(model, las = 1)      # Residuals, Fitted, ...
par(opar)
dev.off()

model <- lm(logPM ~ S1:S2 + S2:S3 + S3:S4 + S4:S5 + S5 : S6
+ S6 : S7 + S7 : S8 + S8  : S9 + S9 : S10 + S10 : S11 + S11 : S12
 + S12 : S13 + S13 :S14 + S14 : S15 + S15 : S16 + S16 : S17
 + S17 : S18 + S18 : S19 + S19 : S20 +  S20 : S21 + S21 : S22
 + S22 : S23 + S23 : S24 + S24 : S25 +  AN2 + TN2 +
 GN2 + CN2,subset=Set300)

summary(model)
Sys.setenv("DISPLAY"=":0")
jpeg(file="ALL_PM_seq_noTN_2NUcleo.jpeg",bg="white",
quality=100, res=10000)
opar <- par(mfrow = c(2,2), oma = c(0, 0, 1.1, 0))
plot(model, las = 1)      # Residuals, Fitted, ...
par(opar)
dev.off()

Set300 <- sample(ss<-1:length(logPM),20000)

model <- lm(logPM ~ S1:S2:S3 + S2:S3:S4 + S3:S4:S5 +
S4:S5:S6 + S5 : S6:S7 + S6 : S7:S8 +  S7 : S8:S9 + S8  : S9:S10
+ S9 : S10:S11 + S10 : S11:S12 +  S11 : S12:S13 +
S12 : S13:S14 + S13 :S14:S15 + S14 : S15:S16 + S15 : S16:S17
  + S16 : S17:S18 + S17 : S18:S19 + S18 : S19:S20 + S19 : S20:S21 +
  S20 : S21:S22 + S21 : S22:S23 + S22 : S23:S24 + S23 : S24:S25
   + S24 : S25 + AN2 + TN2 + GN2 + CN2,subset=Set300)

summary(model)
Sys.setenv("DISPLAY"=":0")
jpeg(file="ALL_PM_seq_noTN_3NUcleo.jpeg",bg="white",
quality=100, res=10000)
opar <- par(mfrow = c(2,2), oma = c(0, 0, 1.1, 0))
```

```
plot(model, las = 1)      # Residuals, Fitted, ...
par(opar)
dev.off()

Set300 <- sample(ss<-1:length(logPM),10000)

model <- glm(logPM ~ S1:S2:S3 + S2:S3:S4 + S3:S4:S5 +
S4:S5:S6 + S5 : S6:S7 + S6 : S7:S8 + S7 : S8:S9 + S8  : S9:S10
+ S9 : S10:S11 + S10 : S11:S12 + S11 : S12:S13 + S12 : S13:S14
+ S13 :S14:S15 + S14 : S15:S16 + S15 : S16:S17 + S16 : S17:S18
+ S17 : S18:S19 + S18 : S19:S20 + S19 : S20:S21 + S20 : S21:S22
+ S21 : S22:S23 + S22 : S23:S24 + S23 : S24:S25 + S24 : S25 +
AN2 + TN2 + GN2 + CN2,subset=Set300)

summary(model)
Sys.setenv("DISPLAY"=":0")
jpeg(file="ALL_PM_glm_seq_noTN_3NUcleo.jpeg",bg="white",
quality=100, res=10000)
opar <- par(mfrow = c(2,2), oma = c(0, 0, 1.1, 0))
plot(model, las = 1)      # Residuals, Fitted, ...
par(opar)
dev.off()

Set300 <- sample(ss<-1:length(logPM),8000)
model <- lm(logPM ~ S1 :S2 : S3 : S4  + S2 : S3 : S4 : S5 +
S3 : S4 : S5 : S6  + S4 : S5 : S6 : S7  + S5 : S6 : S7 : S8   +
S6 : S7 : S8  : S9  + S7 : S8  : S9 : S10  + S8  : S9 : S10 : S11
 + S9 : S10 : S11 : S12  + S10 : S11 : S12 : S13  +
  S11 : S12 : S13 :S14  + S12 : S13 :S14 : S15  +
  S13 :S14 : S15 : S16  + S14 : S15 : S16 : S17  +
  S15 : S16 : S17 : S18  + S16 : S17 : S18 : S19
+ S17 : S18 : S19 : S20 + S18 : S19 : S20 : S21
+ S19 : S20 : S21 : S22 + S20 : S21 : S22 : S23 +
S21 : S22 :S23: S24 + S22 :S23: S24 : S25 + AN + TN + GN +
CN,subset=Set300)

summary(model)
Sys.setenv("DISPLAY"=":0")
jpeg(file="../PM/ALL_PM_glm_seq_noTN_4NUcleo.jpeg",
bg="white", quality=100, res=10000)
opar <- par(mfrow = c(2,2), oma = c(0, 0, 1.1, 0))
```

```
plot(model, las = 1)      # Residuals, Fitted, ...
par(opar)
dev.off()
```

3 nucleotides interaction model implementation

```
resample <- function(x, size, ...)
       if(length(x) <= 1) { if(!missing(size) && size == 0) x[FALSE]
       else x } else sample(x, size, ...)


logPM1 <- read.table("logU2AFpm.txt")

AN1 <- read.table("ANpm.txt")
TN1 <- read.table("TNpm.txt")

GN1 <- read.table("GNpm.txt")
CN1 <- read.table("CNpm.txt")
S251 <- read.table("SEQ_ARRAY_1_25_pm.txt")

for(j in 1: dim(logPM1)[2]){
logPM <- logPM1[,j]
print(paste ("slide_",as.character(j)))
# para reducir memoria
AN <- AN1[,26]^2
TN <- TN1[,26]^2
GN <- GN1[,26]^2
CN <- CN1[,26]^2


S1 <- S251[,1]
S2 <- S251[,2]
S3 <- S251[,3]
S4 <- S251[,4]
S5 <- S251[,5]
S6 <- S251[,6]
S7 <- S251[,7]
S8 <- S251[,8]
S9 <- S251[,9]
S10 <- S251[,10]
S11 <- S251[,11]
```

```
S12 <- S251[,12]
S13 <- S251[,13]
S14 <- S251[,14]
S15 <- S251[,15]
S16 <- S251[,16]
S17 <- S251[,17]
S18 <- S251[,18]
S19 <- S251[,19]
S20 <- S251[,20]
S21 <- S251[,21]
S22 <- S251[,22]
S23 <- S251[,23]
S24 <- S251[,24]
S25 <- S251[,25]

n=20000
Set300 <- sample(ss<-1:length(logPM),n)
x=2
xx=0
y <-0
resumen <- matrix(0,100,8)
print ("Itera_  R.squared     Adj R.squared     Eliminados")

while (x>0.001){


logPM <- logPM[Set300]
# para reducir memoria
AN <- AN[Set300]
TN <- TN[Set300]
GN <- GN[Set300]
CN <- CN[Set300]


S1 <- S1[Set300]
S2 <- S2[Set300]
S3 <- S3[Set300]
S4 <- S4[Set300]
S5 <- S5[Set300]
S6 <- S6[Set300]
S7 <- S7[Set300]
```

```
S8 <- S8[Set300]
S9 <- S9[Set300]
S10 <- S10[Set300]
S11 <- S11[Set300]
S12 <- S12[Set300]
S13 <- S13[Set300]
S14 <- S14[Set300]
S15 <- S15[Set300]
S16 <- S16[Set300]
S17 <- S17[Set300]
S18 <- S18[Set300]
S19 <- S19[Set300]
S20 <- S20[Set300]
S21 <- S21[Set300]
S22 <- S22[Set300]
S23 <- S23[Set300]
S24 <- S24[Set300]
S25 <- S25[Set300]


model <- lm(logPM ~ S1:S2:S3 + S2:S3:S4 +
S3:S4:S5 + S4:S5:S6 + S5 : S6:S7 + S6 : S7:S8
+ S7 : S8:S9 + S8  : S9:S10 + S9 : S10:S11 +
S10 : S11:S12 + S11 : S12:S13 + S12 : S13:S14
+ S13 :S14:S15 + S14 : S15:S16 + S15 : S16:S17 +
S16 : S17:S18 + S17 : S18:S19 + S18 : S19:S20
+ S19 : S20:S21 + S20 : S21:S22 + S21 : S22:S23
+ S22 : S23:S24 + S23 : S24:S25 + AN + TN + GN
 + CN)
summodel <- summary(model)
quantilResiduals <- quantile (summodel$residuals)
dquantiles <- quantilResiduals[4]-quantilResiduals[2]

# outliers limits
MenorLimite <- quantilResiduals[2]-(1.5* dquantiles)
MayorLimite <- quantilResiduals[4]+(1.5* dquantiles)


Set300 <-as.numeric(names(resample(residuals(model)
[residuals(model)< MayorLimite & residuals(model)>
 MenorLimite])))
```

```
x1 <- n - length(Set300)
n <- length(Set300)
y <- y+1
resumen[y,1:5] <- quantile (summodel$residuals)
#print ("Residual Quantiles:")
#print (quantile (summodel$residuals))

print (paste(as.character(y),"     ",as.character
(summodel$r.squared),"     ",as.character(summodel
$adj.r.squared),"     ",as.character(x1)))
resumen[y,6] <- summodel$r.squared
if (y > 1){
x <-resumen[y,6]-resumen[y-1,6]
}
if (resumen[y,6]>=0.75 & xx==0) {
xx=1
}
resumen[y,7] <- summodel$adj.r.squared
resumen[y,8] <- x1

if (y==1 | x<=0.001 | xx==1){
if (y==1) {
res <- paste ("PM_LIM_15_Histo_Array_",
as.character(j),"_Itera1.jpeg")
}
if (x<=0.001) {
res <- paste ("PM_LIM_15_Histo_Array_",
as.character(j),"_Itera_Final.jpeg")
}
if (xx==1){
res <- paste ("PM_LIM_15_Histo_Array_",
as.character(j),"_Itera_R0.75.jpeg")
save (model,file =paste ("PM_LIM_15_
model_Array_",as.character(j),
"_Itera_R0.75.RData") )
xx <-2
}
Sys.setenv("DISPLAY"=":0")
jpeg(file=res,bg="white", quality=100, res=10000)
hist (summodel$residuals)
dev.off()
```

```
}
}
save (model,file =paste ("PM_LIM_15_model_Array_
",as.character(j),"_Itera_Final.RData") )
write.table (resumen[1:y,], paste("PM_LIM_15_resumen_Array_
",as.character(j),".txt"))
}
```

A.2.2 Probe standardization: j values

```
logPM1 <- read.table("logU2AFpm.txt")
logPM1p <- read.table("PM_LIM_15_model_Array_ALLarrays
_Pred_PM.txt")
n1 <- round(nrow/200,0)

nrow <- dim(logPM1)[1]  #probes number
ncol <- dim(logPM1)[2]  # arrays number
J <- matrix(0,nrow,ncol)  #j values
des <- matrix(0,round(nrow/n,0),7*ncol) # bins number

for(j in 1: ncol){
print(paste ("slide_",as.character(j)))

#calculo de los sd de los bin de 3000 cada uno

binLogPm1 <- matrix(0,nrow,4)
binLogPm1[,1] <- 1:nrow
binLogPm1 [,2]<- logPM1[,j]
binLogPm1 [,3]<- logPM1p[,j]
binLogPm1<- binLogPm1[order(binLogPm1[,3]),]

for(z in 0 : ((nrow/n1)-1)) {
p <- z*n1+1
p1 <- ifelse (z == round(((nrow/n1)-1),0), nrow,
(z+1)*n1)
binLogPm1 [p:p1,4]<- sd (binLogPm1 [p:p1,2])
des[(z+1),(7*(j-1)+1):(j*7-2)] <- quantile(binLogPm1
 [p:p1,2])
des[(z+1),(j*7-1)] <- sd (binLogPm1 [p:p1,2])
des[(z+1),(j*7)] <- mean (binLogPm1 [p:p1,2])
```

```
}

#j values calculation
binLogPm1<- binLogPm1[order(binLogPm1[,1]),]

J[,j] <- (logPM1[,j]-logPM1p [,j] )/binLogPm1[,4]

}

write.table(des,"PM_LIM_15_model_Array_ALLarrays_
IntervalsCorrectLogPM.txt")
write.table(J,"PM_LIM_15_model_Array_ALLarrays_Jmat_
Corect.txt")
```

A.2.3 *ENRval* values calculation

```
resample <- function(x, size, ...)
       if(length(x) <= 1) { if(!missing(size) && size == 0)
       x[FALSE] else x
       } else sample(x, size, ...)

PS <- read.table (paste(PathTablas,"HG-U133A_tag_
ProbeSequenceOrder.txt",sep=""), skip = 1)
PS<- PS[,2]
names (PS)<- 1:length(PS)

#Read $t$ values
J <- read.table(paste(PathResults,"ALLarrays_Jmat_Corect1.txt",
sep=""))

J <- J[,J11]

IP <-levels(PS)
nProbeSet <- length(IP)

Exp <- c("T","I","T","I","T","I")
nInput <- length(resample(Exp[Exp == "T"]))
nIp <- length(resample(Exp[Exp == "I"]))
muestra <- c(1,1,2,2,3,3)
nArray <-dim(J)[2]
```

```
J_Ind_trim <- matrix(NA, nProbeSet,(nIp*3)) #individual PSsco
J_trim <- matrix(NA, nProbeSet,11)  #ENRval replicates

for (i in 1:nProbeSet){
        print (i)
Set300 <-sort(as.numeric(names(resample(PS[PS ==
IP[i]]))))
nProbes <- length(Set300)
nProbes10 <- round ((nProbes*0.1),0)

#trimmed PSsco

J1 <- rep (0, (nProbes*nInput)) # arrays Input
J2 <- rep (0, (nProbes*nIp)) # arrays Ip
c <-0 # array possition J1
c1 <- 0 # array possitio J2
c3 <- 0 # array possitio J_Ind_trim
for (j in 1: nArray){
#PSsco Ip individuals
Jaux <- J[Set300,j]
Jaux <- Jaux[order(Jaux)]

if(Exp[j]=="T"){
J1[(c+1):(c+ nProbes)] <- Jaux
c <- c + nProbes
}
else{
c3 <- c3 +1
J_Ind_trim[i,c3]<-sqrt(nProbes-(nProbes10*2))
*mean(Jaux[(1+ nProbes10): (
nProbes-nProbes10)])
J2[(c1+1):(c1+ nProbes)] <- Jaux #
c1 <- c1 + nProbes
}

}
nProbes <- length(J1)
nProbes10 <- round ((length(J1)*0.1),0)
Jaux <- J1
Jaux <- Jaux[order(Jaux)]
```

```
#PSsco  J_trim  col 1 = Input
J_trim[i,1]<-sqrt(length(J1)-(nProbes10*2))*mean(Jaux[(1
+ nProbes10):
(nProbes-nProbes10)])

#PSsco J_trim  col2 = sd trimmed Input
J_trim[i,2]<- sd(Jaux[(1+ nProbes10): (nProbes-nProbes10)])

nProbes <- length(J2)
nProbes10 <- round ((length(J2)*0.1),0)
Jaux <- J2
Jaux <- Jaux[order(Jaux)]

#PSsco J_trim col3 = Ip
J_trim[i,3]<-sqrt(length(J2)-(nProbes10*2))*mean(Jaux[(1
+ nProbes10):
(nProbes-nProbes10)])

#PSsco J_trim col6 = IP - Input
J_trim[i,6]<- J_trim[i,3] - J_trim[i,1]

#PSsco J_trim col9 = (IP - Input)/ sd trimmed Input
J_trim[i,9]<- J_trim[i,6] / J_trim[i,2]
}

# Pvalue and FDR individual array Ip
for (i in 1:nIp){
MedIp <- median(J_Ind_trim[,i])
SDIp <- sd (resample (J_Ind_trim[J_Ind_trim[,i] <=  MedIp,i]))
J_Ind_trim[,(nIp+(2*i-1))]<-1-pnorm(J_Ind_trim[,i], mean=
MedIp, sd=SDIp, log =  FALSE) #Pvalue
for (j in 1:nProbeSet){
if (J_Ind_trim[j,i]> MedIp){ #FDR >s 0
J_Ind_trim[j,(nIp+(2*i))] <- length (resample
(J_Ind_trim[J_Ind_trim[,i] < -
(J_Ind_trim[j,i]-MedIp),i])) / length(resample
 (J_Ind_trim[J_Ind_trim[,i] >= J_Ind_trim[j,i],i]))
}
}
}
#Pvalie  J_trim  col 4 = Ip
```

```
MedIp2 <- median(J_trim[,3])

#SDIp <- sd (resample (J_trim[J_trim[,3] <= MedIp2,3]))
SDIp <- sd (c(resample (J_trim[J_trim[,3] <= MedIp2,3]),(
-resample (J_trim[J_trim[,3] <=
MedIp2,3])+(2* MedIp2))))
J_trim[,4]<-1-pnorm(J_trim[,3], mean=MedIp2, sd=SDIp,
log = FALSE)
#Pvalie  J_trim  col 8 = Ip - Input
MedIp3 <- median(J_trim[,6])

#SDIp <- sd (resample (J_trim[J_trim[,6] <= MedIp3,6]))
SDIp <- sd (c(resample (J_trim[J_trim[,6] <= MedIp3,6]),(
-resample (J_trim[J_trim[,6] <= MedIp3,6])+(2* MedIp3))))
J_trim[,7]<-1-pnorm(J_trim[,6], mean=MedIp3, sd=SDIp,
log = FALSE)

#Pvalie  J_trim  col 10 = Ip - Input/ sd Input
MedIp5 <- median(J_trim[,9])

#SDIp <- sd (resample (J_trim[J_trim[,9] <= MedIp5,9]))
SDIp <- sd (c(resample (J_trim[J_trim[,9] <= MedIp5,9]),
(-resample (J_trim[J_trim[,9] <= MedIp5,9])+(2* MedIp5))))
J_trim[,10]<-1-pnorm(J_trim[,9], mean=MedIp5, sd=SDIp, log =
 FALSE)

#calculo FDR
for (i in 1:nProbeSet){
if (J_trim[i,3]> MedIp2){ #FDR > 0
J_trim[i,5]<-length(resample (J_trim[J_trim[,3]
 < -(J_trim[i,3]-MedIp2),3]))/length(resample
 (J_trim[J_trim[,3] >= J_trim[i,3],3]))
}
if (J_trim[i,6]> MedIp3){ #FDR > 0
J_trim[i,8]<-length(resample (J_trim[J_trim[,6]
 < -(J_trim[i,6]-MedIp3),6]))/length(resample
 (J_trim[J_trim[,6] >= J_trim[i,6],6]))
}
if (J_trim[i,9]> MedIp5){ #FDR > 0
J_trim[i,11]<-length(resample (J_trim[J_trim[,9]
< -(J_trim[i,9]-MedIp5),9]))/length(resample
```

```
 (J_trim[J_trim[,9] >= J_trim[i,9],9]))
}
}
# tables tx
write.table(J_Ind_trim,paste(PathResults,"PSsco_ind.txt",sep=""))
#write.table(J_trim,"./AnovaENR_trimmed/PSsco_c1.txt")
write.table(J_trim,paste(PathResults,"PSsco_c1_SD.txt",sep=""))

# summary table  of  probe set
#write.table(probeset,"PM_Probeset_J_NoIP.txt")
#plot (J[Set300,4],type = "l",col = "red")
#points(J[Set300,4], cex = .5, col = "dark red")
# Sys.setenv("DISPLAY"=":0")
# jpeg(file=res,bg="white", quality=100, res=
10000)
# hist (summodel$residuals)
# dev.off()
```

A.2.4 New rank based approach for enriched gene selection.

```
resample <- function(x, size, ...)
       if(length(x) <= 1) { if(!missing(size) && size == 0) x[FALSE] else x
       } else sample(x, size, ...)


PS <- read.table ("/Users/ebarretoh/Documents/POSGRADO
/estatistica/Datos_Margarida
/LocalMAT/spikeIN/Tablas/HG-U133A_tag_
ProbeSequenceOrder.txt", skip = 1)
PS<- PS[,2]
names (PS)<- 1:length(PS)
J <- read.table("/Users/ebarretoh/Documents/POSGRADO/
estatistica/Datos_Margarida
/LocalMAT/spikeIN/results/PM/PM_LIM_15_model_Array_
ALLarrays_Jmat_Corect1.txt")

J11 <-c(14,1,28,15,42,29)
J <- J[,J11]
IP <- levels(PS)
```

```
nProbeSet <- length(IP)

Exp <- c("T","I","T","I","T","I")
nInput <- length(resample(Exp[Exp == "T"]))
nIp <- length(resample(Exp[Exp == "I"]))
muestra <- c(1,1,2,2,3,3)
nArray <-dim(J)[2]

J_trim <- matrix(0, nProbeSet,9)  # array ranks results
names (J_trim) <- nProbeSet

for (i in 1:nProbeSet){
print(i)
Set300 <-sort(as.numeric(names(resample(PS[PS ==
IP[i]]))))
nProbes <- length(Set300)
nProbes10 <- round ((nProbes*0.1),0)

#trimmed PSsco

J1 <- rep (0, (nProbes*nInput)) # arrays Input
J2 <- rep (0, (nProbes*nIp)) # arrays Ip
c <-0 # posicion arreglo J1
c1 <- 0 # posicion arreglo J2

for (j in 1: nArray){
Jaux <- J[Set300,j]
if(Exp[j]=="T"){
J1[(c+1):(c+ nProbes)] <- Jaux
c <- c + nProbes
}
else{
J2[(c1+1):(c1+ nProbes)] <- Jaux #
c1 <- c1 + nProbes
}
}
J1 <- J2 -J
nProbes <- length(J1)
nProbes10 <- round ((length(J1)*0.1),0)
Jaux <- J1
Jaux <- Jaux[order(Jaux)]
```

```
          J1 <- Jaux[(1+ nProbes10): (nProbes-nProbes10)]

#  col1 TPj mean
J_trim[i,1]<-mean(J1)

c <-0 # option array J1
J_trim[i,2] <-0
J_trim[i,3] <-0
for (j in 1: (nArray/2)){
Jaux1<-J1[(c+1):(c+ nProbes)]
Jaux2 <- mean(Jaux1)
#col2 TPsjnAr
J_trim[i,2]<-J_trim[i,2] + abs(Jaux2 - J_trim[i,1])
#col3 TPssj
J_trim[i,3]<-J_trim[i,3] + (sum(abs(Jaux1-Jaux2))
/nProbes)
c <- c + nProbes
}
#col2 TPsj
J_trim[i,2]<-J_trim[i,1] - (J_trim[i,2]/(nArray/2))
#col3 TPssj
J_trim[i,3]<-J_trim[i,1]- (J_trim[i,3]/(nArray/2))
#col4 TPsj
J_trim[i,4]<-(2*J_trim[i,1]) - J_trim[i,2]
#col5 TPssj
J_trim[i,5]<-(2*J_trim[i,1]) - J_trim[i,3]
}
# Total Rank  RT
J_trim[,6] <- (nProbeSet+1- rank (J_trim[,2]))*(nProbeSet+1-rank
(J_trim[,3]))*(nProbeSet+
1- rank (J_trim[,4]))*(nProbeSet+1- rank (J_trim[,5]))/nProbeSet^4
# Estipular M
M=100

#----- permutations-------------------------
rp<-matrix(0, nProbeSet,M)
for(t in 1:M)
{
print (t)
rp[,t] <- (nProbeSet+1-rank (resample(J_trim[,2]))) *
(nProbeSet+1-rank (resample
```

```
(J_trim[,3]))) * (nProbeSet+1-rank (resample(J_trim[,4])))
*(nProbeSet+1-rank (resample(J_trim[,5])))/nProbeSet^4
}


# average expected value" E(rp)~x(rp)/M
#
J_trim[,7]<-0
for(i in 1: nProbeSet)
{
print(i)
# col5 E value
count<-0
for(j in 1: M)
count <- count + length(resample (rp[rp[,j] <= J_trim[i,6],j]))
J_trim[i,7] <- count/M
}


#  "false discovery rate" q=E(rp)/rank

#col6 FDR ---  q
J_trim[,8] <- J_trim[,7]/ rank (J_trim[,6])

# p-value
J_trim[,9] <- J_trim[,7]/nProbeSet
# tablas txt

write.table(J_trim,"/Users/ebarretoh/Documents/POSGRADO/
estatistica/Datos_Margarida
/LocalMAT/spikeIN/results/PM/Ranks_Subset_Suma_resta
_trim.txt")
```

References

Affymetrix I. 2004. Genechip® expression analysis data analysis fundamentals. Technical report, Affymetrix Inc, Santa Clara, CA.

Affymetrix I. 2007. Genechip human genome u133 arrays data sheet. Technical report, Affymetrix Inc, Santa Clara, CA.

Affymetrix I. 2009. Genechip expression analysis techcical manual. Technical report, Affymetrix Inc, Santa Clara, CA.

Allemand E., Batsche E. and Muchardt C. 2008. Splicing, transcription, and chromatin: a menage a trois. *Current opinion in genetics & development*, 18(2):145–51.

Alvord G. W., Roayaei J. A., Quiñones O. A. and Schneider K. T. 2007. A microarray analysis for differential gene expression in the soybean genome using bioconductor and r. *Brief Bioinform*, 8(6):415–31.

Amaral P. P. and Mattick J. S. 2008. Noncoding rna in development. *Mammalian genome : official journal of the International Mammalian Genome Society*, 19(7-8):454–92.

Auweter S. D., Oberstrass F. C. and Allain F. H. 2006. Sequence-specific binding of single-stranded rna: is there a code for recognition? *Nucleic acids research*, 34(17):4943–59.

Baroni T. E., Chittur S. V., George A. D. and Tenenbaum S. A. 2008. Advances in rip-chip analysis : Rna-binding protein immunoprecipitation-microarray profiling. *Methods in molecular biology*, 419:93–108.

Barreto-Hernandez E., Gama-Carvalho M. and Sousa L. 2011. Preprocessing optimization of rna immunoprecipitation microarray data. *Journal of Computational Biology*, DOI: 10.1089/cmb.2010.0020.

Benjamini Y. and Hochberg Y. 1995. Controlling the false discovery rate - a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B-Methodological*, 57(1):289–300.

Berg J., Tymoczko J. L. and Stryer L. 2006. *Biochemistry*. W. H. Freeman, San Francisco, 6th edition.

Biro J. C. 2004. Seven fundamental, unsolved questions in molecular biology. cooperative storage and bi-directional transfer of biological information by nucleic acids and proteins: an alternative to "central dogma". *Medical hypotheses*, 63(6):951–62.

Black D. L. 2003. Mechanisms of alternative pre-messenger rna splicing. *Annu Rev Biochem*, 72:291–336.

Bolstad B. M., Irizarry R. A., Astrand M. and Speed T. P. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–93.

Bolstad B. M., Collin F., Brettschneider J., Simpson K., Cope L., Irizarry R. A. and Speed T. P. 2005. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, chapter Quality assessment of Affymetrix GeneChip Data. Springer.

Bolstad B. M. B. *affyPLM: methods for fitting probe-level models*. Bioconductor, r package version 1.28.5 edition, 2011.

Breitling R., Armengaud P., Amtmann A. and Herzyk P. 2004. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett*, 573 (1-3):83–92.

Brookes E. and Pombo A. 2009. Modifications of rna polymerase ii are pivotal in regulating gene expression states. *EMBO reports*, 10(11): 1213–9.

Brow D. A. 2002. Allosteric cascade of spliceosome activation. *Annual review of genetics*, 36:333–60.

Chen M. and Manley J. L. 2009. Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nature reviews. Molecular cell biology*, 10(11):741–54.

Chen Y. and Varani G. 2005. Protein families and rna recognition. *The FEBS journal*, 272(9):2088–97.

Cope L. M., Irizarry R. A., Jaffee H. A., Wu Z. and Speed T. P. 2004. A benchmark for affymetrix genechip expression measures. *Bioinformatics*, 20(3):323–31.

Costa F. F. 2005. Non-coding rnas: new players in eukaryotic biology. *Gene*, 357(2):83–94.

Crick F. 1970. Central dogma of molecular biology. *Nature*, 227(5258): 561–3.

DeRisi J. L., Iyer V. R. and Brown P. O. 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278 (5338):680–6.

Doran G. 2007. Rnai - is one suffix sufficient? *J RNAi Gene Silencing*, 3(1):217–9.

Dudoit S., Yang Y. H., Callow M. J. and Speed T. P. 2002. Statistical methods for identifying genes with differential expression in replicated cdna microarray experiments. *Stat. Sin.*, 12(1):111–139.

Eccleston A. and Skipper M. 2009. Transcribing the genome. *Nature*, 461(7261):185.

Fodor S. P., Rava R. P., Huang X. C., Pease A. C., Holmes C. P. and Adams C. L. 1993. Multiplexed biochemical assays with biological chips. *Nature*, 364(6437):555–6.

Fuda N. J., Ardehali M. B. and Lis J. T. 2009. Defining mechanisms that regulate rna polymerase ii transcription in vivo. *Nature*, 461(7261): 186–92.

Gama-Carvalho M., Carvalho M. P., Kehlenbach A., Valcarcel J. and Carmo-Fonseca M. 2001. Nucleocytoplasmic shuttling of heterodimeric splicing factor u2af. *J Biol Chem*, 276(16):13104–12.

Gama-Carvalho M., Barbosa-Morais N. L., Brodsky A. S., Silver P. A. and Carmo-Fonseca M. 2006. Genome-wide identification of functionally distinct subsets of cellular mrnas associated with two nucleocytoplasmic-shuttling mammalian splicing factors. *Genome Biol*, 7(11):R113.

García-Blanco M. A., Jamison S. F. and Sharp P. A. 1989. Identification and purification of a 62,000-dalton protein that binds specifically to the polypyrimidine tract of introns. *Genes Dev*, 3(12A):1874–86.

Gentleman R. C., Carey V. J., Bates D. M. and others . 2004. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5:R80.

Gerber A. P., Herschlag D. and Brown P. O. 2004. Extensive association of functionally and cytotopically related mrnas with puf family rna-binding proteins in yeast. *PLoS Biol*, 2(3):E79.

Gilbert S. F. 2010. *Developmental Biology*. Sinauer Associates, Inc., Sunderland, MA, 9th edition.

Girard A., Sachidanandam R., Hannon G. J. and Carmell M. A. 2006. A germline-specific class of small rnas binds mammalian piwi proteins. *Nature*, 442(7099):199–202.

Hanlon S. E. and Lieb J. D. 2004. Progress and challenges in profiling the dynamics of chromatin and transcription factor binding with dna microarrays. *Curr Opin Genet Dev*, 14(6):697–705.

He Y., Vogelstein B., Velculescu V. E., Papadopoulos N. and Kinzler K. W. 2008. The antisense transcriptomes of human cells. *Science*, 322(5909):1855–7.

Heard E., Mongelard F., Arnaud D., Chureau C., Vourc'h C. and Avner P. 1999. Human xist yeast artificial chromosome transgenes show partial x inactivation center function in mouse embryonic stem cells. *Proceedings of the National Academy of Sciences of the United States of America*, 96(12):6841–6.

Hong F., Breitling R., McEntee C. W., Wittner B. S., Nemhauser J. L. and Chory J. 2006. Rankprod: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics*, 22(22): 2825–7.

Horvath P. and Barrangou R. 2010. Crispr/cas, the immune system of bacteria and archaea. *Science*, 327(5962):167–70.

Horwich M. D., Li C., Matranga C., Vagin V., Farley G., Wang P. and Zamore P. D. 2007. The drosophila rna methyltransferase, dmhen1, modifies germline pirnas and single-stranded sirnas in risc. *Curr Biol*, 17(14):1265–72.

Huber W., Heydebreck A., Sültmann H., Poustka A. and Vingron M. 2002. Variance stabilization applied to microarray data calibration

and to the quantification of differential expression. *Bioinformatics*, 18 Suppl 1:S96–104.

Huber W., Heydebreck A., Sueltmann H., Poustka A. and Vingron M. 2003. Parameter estimation for the calibration and variance stabilization of microarray data. *Stat Appl Genet Mol Biol*, 2:Article3.

Irizarry R. A., Hobbs B., Beazer-Barclay Y., Antonellis K. J., Scherf U. and Speed T. 2003a. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2): 249–64.

Irizarry R. A., Bolstad B. M., Collin F., Cope L. M., Hobbs B. and Speed T. P. 2003b. Summaries of affymetrix genechip probe level data. *Nucleic Acids Res*, 31(4):e15.

Johnson , W.E.L. W. M. C. A., Gottardo R., Carroll J. S., Brown M. and Liu X. S. 2006. Model-based analysis of tiling-arrays for chip-chip. *Proc Natl Acad Sci U S A*, 103(33):12457–62.

Kadota K., Nakai Y. and Shimizu K. 2009. Ranking differentially expressed genes from affymetrix gene expression data: methods with reproducibility, sensitivity, and specificity. *Algorithms Mol Biol*, 4:7.

Katayama S., Tomaru Y., Kasukawa T., Waki K., Nakanishi M., Nakamura M., Nishida H., Yap C. C., Suzuki M., Kawai J., Suzuki H., Carninci P., Hayashizaki Y., Wells C., Frith M., Ravasi T., Pang K. C., Hallinan J., Mattick J., Hume D. A., Lipovich L., Batalov S., Engstrom P. G., Mizuno Y., Faghihi M. A., Sandelin A., Chalk A. M., Mottagui-Tabar S., Liang Z., Lenhard B. and Wahlestedt C. 2005. Antisense transcription in the mammalian transcriptome. *Science*, 309(5740): 1564–1566.

Keene J. D., Komisarow J. M. and Friedersdorf M. B. 2006. Ripchip: the isolation and identification of mrnas, micrornas and protein components of ribonucleoprotein complexes from cell extracts. *Nature protocols*, 1(1):302–7.

Kielkopf C. L., Lucke S. and Green M. R. 2004. U2af homology motifs: protein recognition in the rrm world. *Genes & development*, 18(13): 1513–26.

Lee T. I. and Young R. A. 2000. Transcription of eukaryotic protein-coding genes. *Annual review of genetics*, 34:77–137.

Li C. and Wong W. H. 2001. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc Natl Acad Sci U S A*, 98(1):31–6.

Li C. and Wong W. H. 2003. *The analysis of gene expression data: methods and software*, chapter DNA-Chip Analyzer (dChip), pages 120–141. Springer, New York.

Lioliou E., Romilly C., Romby P. and Fechter P. 2010. Rna-mediated regulation in bacteria: from natural to artificial systems. *New biotechnology*, 27(3):222–35.

Lodish , HF.and BerkA., Kaiser C., Krieger M., Scott M., Bretscher A., Ploegh H. and Matsudaira P. 2007. *Molecular Cell Biology*, chapter Chapter 8: Post-transcriptional Gene Control, Molecular Cell Biology. WH Freeman, San Francisco.

Lonnstedt I. and Speed T. 2002. Replicated microarray data. *Statistica Sinica*, 12(1):31–46.

Lunde B. M., Moore C. and Varani G. 2007. Rna-binding proteins: modular design for efficient function. *Nature reviews. Molecular cell biology*, 8(6):479–90.

Mata J., Marguerat S. and Bähler J. 2005. Post-transcriptional control of gene expression: a genome-wide perspective. *Trends Biochem Sci*, 30(9):506–14.

Matter N., Herrlich P. and Konig H. 2002. Signal-dependent regulation of splicing via phosphorylation of sam68. *Nature*, 420(6916):691–5.

Matzke M. A. and Matzke A. J. 2004. Planting the seeds of a new paradigm. *PLoS biology*, 2(5):E133.

Merendino L., Guth S., Bilbao D., Martínez C. and Valcárcel J. 1999. Inhibition of msl-2 splicing by sex-lethal reveals interaction between u2af35 and the 3' splice site ag. *Nature*, 402(6763):838–41.

Naef F. and Magnasco M. O. 2003. Solving the riddle of the bright mismatches: labeling and effective binding in oligonucleotide arrays. *Phys Rev E Stat Nonlin Soft Matter Phys*, 68(1 Pt 1):011906.

Pautz A., Linker K., Hubrich T., Korhonen R., Altenhofer S. and Kleinert H. 2006. The polypyrimidine tract-binding protein (ptb)

is involved in the post-transcriptional regulation of human inducible nitric oxide synthase expression. *The Journal of biological chemistry*, 281(43):32294–302.

Ponten F., Gry M., Fagerberg L., Lundberg E., Asplund A., Berglund L., Oksvold P., Bjorling E., Hober S., Kampf C., Navani S., Nilsson P., Ottosson J., Persson A., Wernerus H., Wester K. and Uhlen M. 2009. A global view of protein expression in human cells, tissues, and organs. *Molecular systems biology*, 5:337.

Pushparaj P. N., Aarthi J. J., Kumar S. D. and Manikandan J. 2008. Rnai and rnaa–the yin and yang of rnaome. *Bioinformation*, 2(6): 235–7.

Robertson G., Hirst M., Bainbridge M., Bilenky M., Zhao Y., Zeng T., Euskirchen G., Bernier B., Varhol R., Delaney A., Thiessen N., Griffith O. L., He A., Marra M., Snyder M. and Jones S. 2007. Genome-wide profiles of stat1 dna association using chromatin immunoprecipitation and massively parallel sequencing. *Nature methods*, 4(8):651–7.

Robinson V. L. 2009. Rethinking the central dogma: noncoding rnas are biologically relevant. *Urologic oncology*, 27(3):304–6.

Ruskin B., Zamore P. D. and Green M. R. 1988. A factor, u2af, is required for u2 snrnp binding and splicing complex assembly. *Cell*, 52 (2):207–19.

Schadt E. E., Li C., Ellis B. and Wong W. H. 2001. Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. *J Cell Biochem Suppl*, Suppl 37:120–5.

Schena M. 1996. Genome analysis with gene expression microarrays. *BioEssays : news and reviews in molecular, cellular and developmental biology*, 18(5):427–31.

Sciamanna I., Vitullo P., Curatolo A. and Spadafora C. 2009. Retrotransposons, reverse transcriptase and the genesis of new genetic information. *Gene*, 448(2):180–6.

Shatkin A. J. and Manley J. L. 2000. The ends of the affair: capping and polyadenylation. *Nature structural biology*, 7(10):838–42.

Shi Y. and Manley J. L. 2007. A complex signaling pathway regulates srp38 phosphorylation and pre-mrna splicing in response to heat shock. *Molecular cell*, 28(1):79–90.

Smyth G. K. 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, 3:Article3.

Sontheimer E. J. and Carthew R. W. 2005. Silence from within: endogenous sirnas and mirnas. *Cell*, 122(1):9–12.

Spellman R. and Smith C. W. 2006. Novel modes of splicing repression by ptb. *Trends in biochemical sciences*, 31(2):73–6.

Steitz T. A. 2008. A structural understanding of the dynamic ribosome machine. *Nature reviews. Molecular cell biology*, 9(3):242–53.

Storey J. 2002. A direct approach to false discovery rates. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 64: 479–498.

Storey J. 2003. The positive false discovery rate: A bayesian interpretation and the q-value. *Annals of Statistics*, 31(6):2013–2035.

Temin H. M. and Mizutani S. 1970. Rna-dependent dna polymerase in virions of rous sarcoma virus. *Nature*, 226(5252):1211–3.

Terribilini M., Lee J. H., Yan C., Jernigan R. L., Honavar V. and Dobbs D. 2006. Prediction of rna binding sites in proteins from amino acid sequence. *RNA*, 12(8):1450–62.

Thieffry D. and Sarkar S. 1998. Forty years under the central dogma. *Trends in biochemical sciences*, 23(8):312–6.

Tress M. L., Martelli P. L., Frankish A., Reeves G. A., Wesselink J. J., Yeats C., Olason P. I., Albrecht M., Hegyi H., Giorgetti A., Raimondo D., Lagarde J., Laskowski R. A., Lopez G., Sadowski M. I., Watson J. D., Fariselli P., Rossi I., Nagy A., Kai W., Storling Z., Orsini M., Assenov Y., Blankenburg H., Huthmacher C., Ramirez F., Schlicker A., Denoeud F., Jones P., Kerrien S., Orchard S., Antonarakis S. E., Reymond A., Birney E., Brunak S., Casadio R., Guigo R., Harrow J., Hermjakob H., Jones D. T., Lengauer T., Orengo C. A., Patthy L., Thornton J. M., Tramontano A. and Valencia A. 2007. The implications of alternative splicing in the encode protein complement. *Proceedings of the National Academy of Sciences of the United States of America*, 104(13):5495–500.

Tukey J. W. 1977. *Exploratory data analysis*. Addison-Wesley series in behavioral science. Addison-Wesley Pub. Co., Reading, Mass. ISBN 0201076160.

Tusher V. G., Tibshirani R. and Chu G. 2001. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A*, 98(9):5116–21.

Vazquez F., Vaucheret H., Rajagopalan R., Lepers C., Gasciolli V., Mallory A. C., Hilbert J. L., Bartel D. P. and Crete P. 2004. Endogenous trans-acting sirnas regulate the accumulation of arabidopsis mrnas. *Molecular cell*, 16(1):69–79.

Wagner E. G., Altuvia S. and Romby P. 2002. Antisense rnas in bacteria and their genetic elements. *Advances in genetics*, 46:361–98.

Wang E. T., Sandberg R., Luo S., Khrebtukova I., Zhang L., Mayr C., Kingsmore S. F., Schroth G. P. and Burge C. B. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221): 470–6.

Warkocki Z., Odenwalder P., Schmitzova J., Platzmann F., Stark H., Urlaub H., Ficner R., Fabrizio P. and Luhrmann R. 2009. Reconstitution of both steps of saccharomyces cerevisiae splicing with purified spliceosomal components. *Nature structural & molecular biology*, 16 (12):1237–43.

Watanabe T., Totoki Y., Toyoda A., Kaneda M., Kuramochi-Miyagawa S., Obata Y., Chiba H., Kohara Y., Kono T., Nakano T., Surani M. A., Sakaki Y. and Sasaki H. 2008. Endogenous sirnas from naturally formed dsrnas regulate transcripts in mouse oocytes. *Nature*, 453(7194):539–43.

Watson J. D. and Crick F. H. 1953. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–8.

Werner E. 2005. Genome semantics, in silico multicellular systems and the central dogma. *FEBS letters*, 579(8):1779–82.

Wit E. and McClure J. D. 2004. *Statistics for microarrays: design, analysis, and inference*. John Wiley & Sons, Chichester, England. ISBN 0470849932 (hbk. : alk. paper). URL `http://www.loc.gov/catdir/description/wiley041/2004045909.html`.

Wu L. and Belasco J. G. 2008. Let me count the ways: mechanisms of gene regulation by mirnas and sirnas. *Molecular cell*, 29(1):1–7.

Wu S., Romfo C. M., Nilsen T. W. and Green M. R. 1999. Functional recognition of the 3' splice site ag by the splicing factor u2af35. *Nature*, 402(6763):832–5.

Wu Z., Irizarry R., Gentleman R., Martinez-Murillo F. and Spencer F. 2004. A model-based background adjustment for oligonucleotide expression arrays. *Journal of the American Statistical Association*, 99 (468):909–917.

Yelin R., Dahary D., Sorek R., Levanon E. Y., Goldstein O., Shoshan A., Diber A., Biton S., Tamir Y., Khosravi R., Nemzer S., Pinner E., Walach S., Bernstein J., Savitsky K. and Rotman G. 2003. Widespread occurrence of antisense transcription in the human genome. *Nature biotechnology*, 21(4):379–86.

Zamore P. D. and Green M. R. 1989. Identification, purification, and biochemical characterization of u2 small nuclear ribonucleoprotein auxiliary factor. *Proc Natl Acad Sci U S A*, 86(23):9243–7.

Zamore P. D., Patton J. G. and Green M. R. 1992. Cloning and domain structure of the mammalian splicing factor u2af. *Nature*, 355 (6361):609–14.

Zhou Z., Licklider L. J., Gygi S. P. and Reed R. 2002. Comprehensive proteomic analysis of the human spliceosome. *Nature*, 419(6903):182–5.

Zorio D. A. and Blumenthal T. 1999. Both subunits of u2af recognize the 3' splice site in caenorhabditis elegans. *Nature*, 402(6763):835–8.