UNIVERSITY OF TORINO

Department of Clinical and Biological Sciences

Doctoral School in Complex Systems in Medicine and Life Sciences

Ph.D. in **COMPLEX SYSTEMS IN POST-GENOMIC BIOLOGY**

XXIII cycle

Computational Methods in Transcriptomics

Tutors: **Prof. Raffaele Calogero**

Prof. Frank Klawonn

Candidate: **Cristina Della Beffa**

Coordinator: **Prof. Michele Caselle**

Academic years: 2008-2010 BIO/11

Contents

Introduction

I started Ph.D. in Complex Systems in Post-Genomic Biology in January 2008. The first two years I worked at the Department of Clinical and Biological Sciences of San Luigi Hospital at Orbassano (Torino), under the supervision of Prof. Raffaele Calogero. I worked at the definition of the optimization of the analysis workflow for the detection of alternative splicing events (ASEs) by mean of ***Affymetrix*** exon array data analysis. This work resulted in the following publication: C. Della Beffa, F. Cordero, R.A. Calogero, ***Dissecting an alternative splicing analysis workflow for GeneChip® Exon 1.0 ST Affymetrix arrays***. *BMC Genomics* 2008.

Since January 2008 until the end of March 2009 I continued the evaluation of several published statistics for alternative splicing events detection to further improve the pipe-line analysis. In the meanwhile, from October 2008 to January 2009 I collaborated at a project on amyotrophic lateral sclerosis, leading in March 2009 to the publication of an article at the International Conference on Complex, Intelligent and Software Intensive Systems: S.C. Lenzken, S. Vivarelli, F. Zolezzi, F. Cordero, C. Della Beffa, R.A. Calogero, Silvia Barabino, ***Genome-Wide Search for Splicing Defects Associated with Amyotrophic Lateral Sclerosis (ALS)***, CISIS 2009.

Since April 2009 until February 2010 I started a project on the development of a quantitative analysis workflow for non-coding RNA quantification by Next-Generation Sequencing (NGS). This resulted in an extension of ***oneChannelGUI*** library, that now allows the quantitative analysis of non-coding NGS data.

oneChannelGUI was originally a software package for single channel microarray data analysis developed in our lab and it is maintained by our group as our contribution to the Bioconductor project.

Since March 2010 I moved to Helmholtz Zentrum für Infektionsforschung (Helmholtz Centre for Infection Research) in Braunschweig, where I have worked under the supervision of Prof. Frank Klawonn. The project I am involved concerns the analysis of several experimental data sets of short hairpin RNAs (shRNAs) for the detection of liver regeneration biomarkers. My actual task is defining the optimal statistics to select significantly regulated shRNAs.

On the basis of the work I did during my Ph.D. training, the thesis is divided into three main topics:

1. Alternative splicing events detection methods.

2. Next-Generation Sequencing for non-coding RNAs analysis.

3. Regulation detection in short hairpin RNA sequencing reads.

Before explaining in brief the contents of this thesis, it is important to say that these past three years I have created algorithms and analyzed data mostly using R code and environment. R environment can be freely downloaded from the Bioconductor web site www.bioconductor.org, which contains many "libraries", sets of algorithms to be used for different kinds of analysis of genomic (and in part also proteomic) experimental data. I have also written routines in C code, interfaced with R environment to speed up some algorithms.

Each chapter is organized in this way: it begins with a general biological description of the phenomenon that deals with the developed computational tool; which instruments were involved in the experimental production of the analyzed data (microarrays, next-generation sequencing); which analytical methods were employed; the results obtained, explained in detail (publications, software, study of real data). In the following a brief summary of the thesis topics.

Chapter 1 begins with a biological introduction (**section 1.1**) to alternative splicing mechanisms and presenting GeneChip® Exon 1.0 ST platform because all the data analyzed to detect splicing came from experiments with this type of microarrays. Always in introductory part of the first chapter, there is a brief description of ***oneChannelGUI***, a graphical interface for pre-processing (quality control, filtering, study design, probe set summary and normalization) and analysis (statistical evaluation, ASEs detection, biological classification) of microarray and deep sequencing data. **Section 1.2** deals with techniques to detect ASEs. The final section presents in detail two articles I contributed to in 2008 and 2009. In the first article: C. Della Beffa, F. Cordero, R.A. Calogero, ***Dissecting an alternative splicing analysis workflow for GeneChip® Exon 1.0 ST Affymetrix arrays.*** *BMC Genomics* 2008, the performances of some statistical methods to detect ASEs at exon-level in microarray data are evaluated. These methods were subsequently implemented in the R library ***oneChannelGUI***. The second article: S.C. Lenzken, S. Vivarelli, F. Zolezzi, F. Cordero, C. Della Beffa, R.A. Calogero, S. Barabino, ***Genome-Wide Search for Splicing Defects Associated with Amyotrophic Lateral Sclerosis (ALS).***

International Conference on Complex, Intelligent and Software Intensive Systems; CISIS 2009, was published as part of the CISIS conference. In this study the pipe-line presented in the article of 2008 is applied to amyotrophic lateral sclerosis models to identify alternative splicing events.

Chapter 2 begins a new part of the thesis with a biological description (**section 2.1**) of non-coding RNAs (specifically focused on microRNAs) and of a recent high-throughput technology called Next-Generation Sequencing (NGS) (**section 2.2**) devoted to the generation of massive DNA/RNA sequences data. The sequences I analyzed were obtained with one of the most recent Next-Generation Sequencing technology, called SOLiD, developed by Applied Biosystems company. Following **section 2.3** deals with ***ncSOLID*** R library, through which next-generation sequences data for quantification of non-coding RNAs are analyzed. This library became part of the library ***oneChannelGUI***.

Chapter 3 deals with the project I am actually working at Helmholtz Centre for infective diseases research in Germany. Short hairpin RNAs (shRNAs) (**section 3.1**) are a type of silencing RNAs that, in this case, have been used as biomarkers to support liver regeneration. The purpose of this study was to detect regulated (up/down) shRNAs between a normal and a disease condition. ShRNAs were first sequenced with Illumina Genome Analyzer (**section 3.2**), another NGS technology, and then analyzed (**section 3.3**) with the most recent published methods, using different kinds of data normalization and filtering techniques to reduce the noise of the data set.

1. Alternative splicing detection methods

This first chapter is organized in seven sections and several subsections. The introduction explains the mechanism of alternative splicing, showing which are the main types of detected events; then array platform used for genomic experiments is presented. Subsequently ***oneChannelGUI*** software part of the Bioconductor open-source project is briefly described. In the "Methods" section, first the workflow for the analysis is dissected step by step, from pre-processing (summarization, normalization, filtering) to the proposed methods for the statistical analysis of the genomic data. In the section "Results and conclusions", the two articles I contributed to during the Ph.D. period are presented in detail, starting with the involved type of experiments, which resulting data were pre-processed before being subjected to a deep analysis from a statistical point of view to detect alternative splicing events (ASEs).

1.1 Introduction

Alternative splicing is a process by which the exons of the RNA produced by gene transcription (pre-mRNA) are joint in multiple combinations during RNA splicing (mRNA). The resulting different mRNAs may be translated into different protein isoforms.

Alternative splicing is a widespread phenomenon in eukaryotes, greatly increasing the diversity of proteins that can be encoded by the genome. Abnormal variations in splicing might contribute to the development of cancer or genetic diseases.

Several types of alternative ASEs are commonly known [1], between them “exon skipping”, where an exon may be spliced out of the primary transcript or retained, is the most common in mammalians.

A high-throughput approach to investigate splicing is DNA microarray-based analysis. The array platform produced by ***Affymetrix*** is a 1.28 cm^2 silicon chip divided into micro-cells (features) on which DNA fragments constituted of 25 base pairs (probes) are synthesized to be hybridized with cDNA or cRNA sequences (targets), previously labeled with fluorescent molecules to get a bright signal, proportional to expression level. In this way the global expression of the transcriptome is available and it is possible to grasp information on the expression of thousands of genes at the same time, with a unique microarray and this can reveal the presence of alternatively spliced mRNAs.

In the studies shown in the articles in **section 1.3**, data were produced with exon arrays. These arrays (*GeneChip*® Exon 1.0 ST) contain over 1.4 million probe sets (constituted by up to four probes each), spread across exons from all known genes, enabling two complementary levels of analysis: gene expression (gene-level) and alternative splicing (exon-level).

The pipe-line to detect alternative splicing events in microarrays experiments, (**section 1.3)**, was implemented as part of the software ***oneChannelGUI***, coded in R.

R is a programming language and environment, suitable for people working with statistics because it is rich of libraries, sets of algorithms, to statistically analyze data coming from biological experiments. Twice a year these libraries are updated and new libraries come out. ***oneChannelGUI*** was developed by Prof. Raffaele Calogero, Dr. Francesca Cordero and Dr. Remo Sanges and became part of the Bioconductor [3] libraries in October 2007 [4]. My contribution was related to the addition of the code for the analysis at exon-level presented in the paper described later on. ***oneChannelGUI*** is based on two previous software packages:

- ***limma*** (linear models for microarray data) [5]: is a generalisation of Lönnstedt and Speed model [6], a parametric empirical Bayesian approach using a mixture of normal distributions and a conjugate prior, deriving a simple expression for the posterior odds of differential expression for each gene. The posterior odds expression is a useful means of ranking genes with respect to their differential expression [5].
- ***affylmGUI*** : a graphical interface to analyze data from ***Affymetrix*** microarrays using ***limma***.

oneChannelGUI is a R library that extends the capabilities of ***affylmGUI*** graphical interface. This library was developed to simplify the use of Bioconductor tools for beginners having limited or no experience in writing R code [4]. This library allows a complete analysis of different type of data sets, from pre-processing (quality control, filtering) to differential expression detection, biological interpretation and classification. ***Affymetrix*** 3' IVT, Human Gene 1.0 ST and exon arrays were first implemented [4].

1.2 Methods

Generally splice detection methods are based on similar hypothesis [7]:

- the exons that constitute a gene are assumed to be proportional to each other across different samples;
- a model to predict exon response is fit;
- a statistic to measure how much biased is the data with respect to the model, is used: a p-value is computed to establish the significance of the obtained results.

The steps that precede splicing events detection are schematically shown below:

Data → Summarization → Normalization → Filters → Statistics → ASEs detection

Raw data coming from replicated biological experiments are the fluorescent signals of the probe sets describing exons. They must be summarized to get a mean value (expression) representing each exon/gene. Then exon intensities must be normalized with respect to their respective gene intensity to make exon expressions independent from the gene they constitute. These values are filtered to remove the lowest and noisiest values that are most likely to be not significant after further statistical analysis and that interfere with ASEs detection, constituting false values. The expression of a transcriptome obtained with microarrays experimental data is proportional to the fluorescence intensity obtained from hybridization of transcripts with DNA probes on microarrays. Once experimentally obtained the bright signal (due to a fluorescent dye on the probes) there are some techniques to obtain gene expression (called summarization methods, which compute the mean intensity of the probe sets) and methods to filter the signal, that can be contaminated by:

- background noise: noise due to experimental background, for example to unspecific hybridization of the probes with sequences different from those of genes complementary to them;
- bias: system errors that can be deleted normalizing the signals with respect to those of a reference array (for example the array with mean expression values);
- outliers: extreme values (very high/low with respect to the mean values of the signals) in some replicates of an experimental sample.

After having filtered out the signal with background adjustment techniques (to delete unspecific hybridization), normalization and research of extreme values (which are likely mistakes and hence to be deleted) it is possible to analyze alternative splicing events using different detection methods.

The following two subsections deals with the pre-processing phase and the statistical analysis of the data, respectively.

1.2.1 Summarization, normalization and filters

Summarizing, normalizing and filtering the data is important before performing deep statistical analysis of any genomic data set. In this section these three techniques of handling data, are presented in detail.

Summarization

A general summarization model, to get gene/exon expression from probe set intensities, formulated by Li and Wong [8], is based on the hypothesis that the intensity measured for arrays $j = 1\ldots J$ and probes $k = 1\ldots K$ is

$$y_{jk} = \left(\vartheta_j * \phi_k\right) + \varepsilon_{jk} \qquad (1)$$

- $y_{jk} = PM_{jk} - MM_{jk}$ difference of intensities;
- ϑ_j expression value for array *j*;
- ϕ_k PM probe affinity (cross-hybridization);
- ε_{jk} error.

This is the base on which are constructed the following two algorithms,

- RMA (Robust Multi-array Analysis, Irizarry 2003) [9], [10]

$$log_2\left(PM_{jk}\right) = \vartheta_j + \phi_k + \varepsilon_{jk} \qquad (2)$$

- PLIER (Probe Logarithmic Intensity Error, *Affymetrix* 2004)

Similar to RMA but keeps into account MM probes yet [11] [12],

$$log_2\left(PM_{jk} - MM_{jk}\right) = \vartheta_j * \phi_k * \varepsilon_{jk} \qquad (3)$$

Normalization

To get exon expressions independent from their gene expression, we compared the performance of the ASEs detection methods with the following ***Splice Index*** (*SI*),

$$SI = log_2\left(\frac{exon}{gene}\right) \qquad (4)$$

where *exon* means exon expression, *gene* means gene expression.

Filters

The following five filters were used to remove the noisy elements:

- Background correction removes intensity signals low with respect to a threshold (intensities lower than 1 in our case are transformed into 0).

- Cross-hybridization correction deletes the probe sets in which every probe perfectly/partially matches more than one sequence of the transcript.

- Delta Splicing Index considers only the intensities which difference between SI of treatment and control is higher than a fixed threshold.

$$SI_T = log_2\left(\frac{T}{\overline{T}}\right) \qquad SI_C = log_2\left(\frac{C}{\overline{C}}\right) \qquad \Delta SI = \left|SI_T - SI_C\right| \tag{5}$$

  Here the Splice Index is computed as the log_2 of the ratio between a value (T or C) with respect to its mean value. Delta Splice Index is the absolute difference between the SI of treatment and control cases.

- Multiple mRNAs retention is a filter to retain only genes associated to more than one transcript in the ENSEMBL database.
- Detection Above BackGround (DABG) compares each probe signal to a distribution of background probes with the same G/C content [13]. A DABG p-value representing the probability that the signal intensity is part of the null distribution is computed and only probes with a p-value lower than a p-value cutoff are retained. We also decided to consider only 90% of values filtered by DABG.

Hereafter are presented seven statistical methods, known from literature, which performances were evaluated at exon-level on a benchmark experiment.

1.2.2 Detection methods

In this section some published methods (MiDAS, Rank Product, OS, ORT, MADS, FIRMA, SPACE) for alternative splicing events detection are described, tested on a benchmark experiment on exon probe sets and on real data.

MiDAS

Proposed by ***Affymetrix***, MiDAS (Microarray Detection of Alternative Splicing) [7] is an ANOVA (analysis of variance) based method. This detection statistic is based on the logarithm of the Splicing Index (presented in the previous normalization description in **section 1.2.1**), a basic metric for the analysis of ASEs: it is a measure of how much exon specific expression differs between two samples [7]. The first step is to normalize the exon-level signals with respect to the gene-level signals and then take the logarithm of this ratio, mathematically transformed into the difference between logged signal of each exon and its gene).

$$log_2\left(\frac{e_{ijk}}{g_{jk}}\right) = log_2(e_{ijk}) - log_2(g_{jk}) \quad (6)$$

- i exon, j array, k gene;
- e_{ijk} exon-level expression;
- g_{jk} gene-level expression.

SI is used to remove the gene-level differential expression, in the estimation of ASEs. If the Splicing Index of an exon is constant in all the experiments, then we can say that this exon is not spliced.

A model for possible splicing is:

$$e_{ijk} = \alpha_{ik} * p_{ijk} * g_{jk} \tag{7}$$

- i exon, j array, k gene;
- e_{ijk} exon-level expression;
- α_{ik} ratio of exon signal to its gene signal in the sample where it is maximally expressed;
- $0 \leq p_{ijk} \leq 1$ proportionate expression of exon i of gene k in sample j;
- g_{jk} gene-level expression.

Dividing both sides of the model by g_{jk} we obtain the Splicing Index and taking the logarithm reduces this to an additive model:

$$log_2(e_{ijk}) - log_2(g_{jk}) = log_2(\alpha_{ik}) + log_2(p_{ijk}) \tag{8}$$

Gene-level analysis

MiDAS includes an error term ε_{ijk} and possible interactions γ_{ik} comparing:

$$log_2(e_{ijk}) - log_2(g_{jk}) = log_2(\alpha_{ik}) + log_2(p_{ijk}) + \gamma_{ik} + \varepsilon_{ijk} \tag{9}$$

wondering if $log_2(p_{ijk}) = \gamma_{ik} = 0$ across samples and exons.

Exon-level analysis

MIDAS considers the situation an exon at a time [7], so that $log_2(\alpha_{ik})$ constant and it is appropriate to consider the model excluding interactions:

$$log_2(e_{ijk}) - log_2(g_{jk}) = log_2(\alpha_{ik}) + log_2(p_{ijk}) + \varepsilon_{ijk} \tag{10}$$

to test the hypothesis of no alternative splicing by testing for the constant effects model $log_2\left(p_{ijk}\right)=0$ for all j samples.

Let us define Sensitivity and Specificity statistical measures:

- TP – *condition present* & *positive result*
- TN – *condition absent* & *negative result*
- FP – *condition absent* & *positive result* (type I error)
- FN – *condition present* & *negative result* (type II error)
- TRUEs = TP + FN
- FALSEs = TN + FP
- True Positive Rate $\quad \text{Sensitivity} = \dfrac{\text{TP}}{\text{TP}+\text{FN}}$
- True Negative Rate $\quad \text{Specificity} = \dfrac{\text{TN}}{\text{TN}+\text{FP}}$
- False Positive Rate $\quad 1\text{-}\text{Specificity} = \dfrac{\text{FP}}{\text{FP}+\text{TN}}$
- False Discovery Rate $\quad \text{FDR} = \dfrac{\text{FP}}{\text{FP}+\text{TP}}$

A Receiver Operating Characteristic (ROC) curve is a graphical plot of the Sensitivity versus (1 - Specificity): it measures how well a statistic differentiates true alternatives from false positives.

To do that we need a known set that does not exhibit alternative splicing (the null set) to be compared with a known set that does exhibit alternative splicing (the alternative set). MIDAS shows considerable improvement in the ROC curves when using exon-level detection over gene-level detection [7].

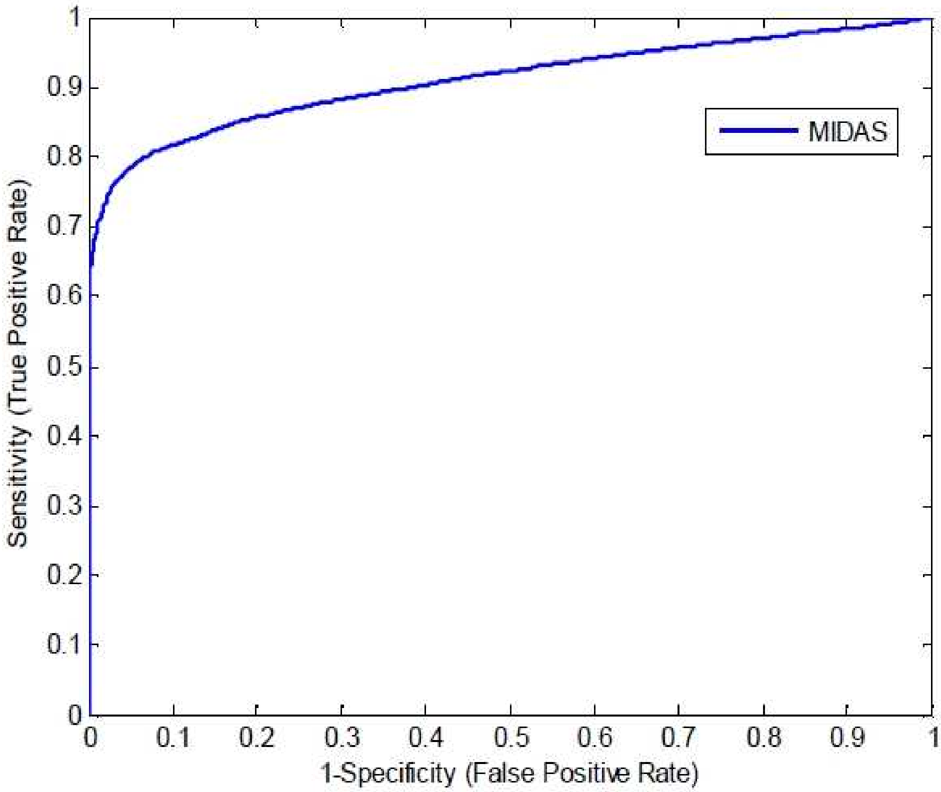


Fig. 1 Exon-level MiDAS on a colon cancer data set, from [7].

Rank Product

Rank Product [14] is a method usually used to detect gene differential expression in microarray data; it is a non-parametric statistic that detects items (genes/exons) that are consistently highly/lowly ranked (outliers) with respect to their differential expression in a number of lists, for example replicate experiments. Rank Product is based on the assumption that the probability of finding a specific gene among the top/down r of n items in a list is $p = r / n$. Computing this gene rank probability for every experiment and multiplying these results leads to the definition of Rank Product [15]

$$\mathrm{RP}_k = \prod_i \frac{r_k^i}{n_i} \tag{11}$$

- r_k^i is the rank of gene k in replicate i ;
- n_i total number of genes in replicate i.

For single-channel arrays, e. g. ***Affymetrix*** GeneChip arrays, the Rank Product values are calculated over all possible pair-wise comparisons between samples. Therefore the Rank Product value cannot be used directly to assess the significance of an observed expression change because we are interested in the combined probability that a gene shows a certain expression pattern, across all the arrays. A simple permutation-based estimation procedure provides a very useful way to determine how likely it is to observe a given Rank Product value or better in a random experiment.

The step by step procedure of the Rank Product algorithm is the following:

1. Compute the fold change (FC) between each pair of intensities belonging to samples in different conditions (e. g. treatment and control), for each gene. Associate to each FC a position (rank) in the list, according to the FC increasing value. Then compute the Rank Product of these ranks, as shown in (11).
2. Generate p permutations of the elements (genes) of the data set within each sample, respectively. Then repeat what already done for the original data set at step 1: compute for each gene the FC between the gene intensity of each couple of new samples. Subsequently, associate ranks to these FCs and compute Rank Products as in (11).
3. Compare the Rank Products computed in step 1 and 2: count how many times the Rank Products of the permuted gene intensities (computed at step 2) are smaller or equal to the Rank Product of the gene intensities in the original data set (step 1). Call this result c.
4. Calculate the mean value for the Rank Product of each gene as c / p.
5. Calculate the p-value as $(c / p) / (n * p)$, where n is the total number of items in each sample.

Only those genes which have p-value lower than a certain threshold (commonly 0.05) are retained because considered differentially expressed.

Advantages of Rank Product over previous statistical techniques:

- simple: a few weak assumptions on data (equal variance for all the genes);
- intuitive: the method is based on the idea that relevant changes should always be large, while small changes may have statistical but rarely biological significance [15].
- significant results with small data sets: a few replicates because Rank Product does not rely on estimating the measurement variance for each single gene.

In conclusion, Rank Product represents a powerful test statistics for defining differentially expressed genes in microarray experiments and its use could potentially be extended to proteomic data analysis and high-throughput sequencing techniques.

OS & ORT

Although OS (Outlier Sum) [16] and ORT (Outlier Robust T-test) [17] were developed with the aim of outlier identification in cancer samples, they were tested on their capability of ASEs detection, interpreting ASEs as exonic outliers. OS and ORT are two statistical methods based on scaling and centering of resulting intensities from experiments, i. e. on the data standardization.

The classical standardization is $t = \frac{x - \bar{x}}{\sigma}$

where $\bar{x}$ is the mean value of x and σ is the standard deviation of x.

Tibshirani and Hastie [16] define the t-statistic to be used in OS, with the median value instead of the mean and the median absolute deviation (mad) instead of the variance,

$$x_{ij}^{'} = \frac{x_{ij} - med_j}{mad_j} \qquad (12)$$

Instead Baolin Wu [17] defines the t-statistic to be used in ORT in the following way,

$$x_{ij}^{''} = \frac{x_{ij} - med_{1j}}{median\left\{ \left| x_{ij} - med_{1j} \right|_{i \le n_1}, \left| x_{ij} - med_{2j} \right|_{i \ge n_1} \right\}} \qquad (13)$$

with x_{ij} gene expression,

$med_j = median_i(x_{ij})$

$mad_j = median_i\{|x_{ij} - med_j|\}$

$med_{1j} = median_i(x_{1j})$ 1 – normal tissue sample

$med_{2j} = median_i(x_{2j})$ 2 – disease tissue sample

n_1 number of genes of the normal sample, n_2 number of genes in the disease sample, $n_1 + n_2 = n$ total number of genes.

Baolin Wu [17] re-defines OS, so the final statistics are respectively:

- <u>OS – Tibshirani & Hastie</u>:

$$W_i = \sum_{j \in C_2} x_{ij}^{'} \cdot I[x_{ij}^{'} > q_{75}(i) + IQR(i)] \qquad (14)$$

$$W_i^{'} = \sum_{j \in C_2} x_{ij}^{'} \cdot I[x_{ij}^{'} < q_{25}(i) - IQR(i)] \qquad (15)$$

where I is the indicator function and IQR the interquantile range as follows, so that values greater than the limit $IQR(x_{ij}^{'}) = q_{75}(i) - q_{25}(i)$ are defined to be outliers in the usual statistical sense [16] [17]. Then they set the outlier sum to the larger of W_i , $W_i^{'}$ in absolute value. This is called “two-sided outlier-sum statistic” [16] and explicitly looks for outliers in group 2, treating group 1 as reference.

- OS – Baolin Wu:

$$T_j = \frac{\sum_{i \in R} \left(x_{ij} - med_j \right)}{mad_j} \tag{16}$$

where R is the set of “outlier disease samples” defined by the following heuristic criterion [17]:

$$R = \left\{ \; i > n_1 : x_{ij} > q_{75}(x_{kj}) + IQR(x_{kj}) \right\} \tag{17}$$

This is equivalent to “OS – Tibshirani & Hastie”, since the subtraction and scaling would not change the order of the observed values [17].

- ORT – Baolin Wu:

$$T_j^{'} = \frac{\sum_{i \in U_j} \left(x_{ij} - med_{1j} \right)}{median\left\{ \; \left| x_{ij} - med_{1j} \right|_{i \le n_1}, \left| x_{ij} - med_{2j} \right|_{i \ge n_1} \right\}} \tag{18}$$

where U_j is the set of the disease sample in which there is an outlier,

$$U_j = \left\{ \; i > n_1 : x_{ij} > q_{75}(x_{kj}) + IQR(x_{kj}) \right\} \tag{19}$$

These techniques hypothesize that only some disease samples contain outliers. When the sum of all the intensities overcomes an ‘a priori’ limit, there happened alternative splicing. ORT is a method consequent to OS and better than this one: while OS gives good results with a few samples, ORT works better with many samples.

MADS

MADS (Microarray Analysis of Differential Splicing) [18] is a method to discover differential alternative splicing from exon microarray data, similar to MiDAS (previously described) because based on Splicing Index (**section 1.2.1**) as the ratio of its background-corrected probe intensity to the estimated gene expression index [18]. Then two separate one-sided t-tests are used to assess whether the Splicing Indices of a probe are significantly higher or lower in one sample group over another group [18] and these constitute the p-values for individual probes. Then p-values are transformed via the formula $x = -2\,log_{\,2}\left(p\right)$ Fisher's method).

Under the null hypothesis that the exon targets are not differentially spliced, the p-values follow a uniform [0,1a] distribution, and the transformed p-values follow a χ_2^2 distribution with 2 degrees of freedom. The sum of the transformed p-values follows a χ_{2k}^2 distribution, where k is the number of probes. This sum of the transformed p-values is used to calculate a probe-set-level p-value, which is used to rank all probe sets [18].

Therefore the main distinction between MADS and MiDAS is that MADS calculates splicing indices and p-values of individual probes separately, prior to the summarization of a probe-set-level p-value. By contrast, ***Affymetrix***'s approach first calculates an overall exon-level expression index (from four probes per probe set), prior to SI calculation and statistical testing. MADS software is available at http://biogibbs.stanford.edu/zyxing/MADS/

FIRMA

FIRMA (Finding Isoforms using Robust Multichip Analysis) [19] algorithm detects alternatively spliced exons in individual samples, without replication or pre-defined groups in the samples, from GeneChip Human Exon 1.0 ST data. It does not take into account the fact that in some probe sets, some/all probes overlap in sequence, introducing additional correlation which may bias alternative splicing detection [19]. This algorithm is sample-by-exon specific: each exon and sample pairing is given a score that is comparable across either samples, genes or exons [19]. This score derives from previous information from the estimation step, based on RMA summarization. For an exon array, a more general additive model can be considered, including the possibility of alternative splicing or different levels of expression per exon,

$$log_2\left(PM_{ijk(j)}\right) = c_i + e_j + d_{ij} + p_{k(j)} + \varepsilon_{ijk(j)} \tag{20}$$

- e_j is the relative change in exon expression for exon j;
- d_{ij} is the interaction between chip and exon giving the relative change for sample i in exon j;
- $p_{k(j)}$ is the nested relative probe effect for the k-th probe in exon j;
- $\varepsilon_{ijk(j)}$ error.

Large values of this parameter d_{ij} point out differential alternative splicing. Rather than estimate d_{ij} explicitly, it is proposed to fit the standard RMA model in

$$log_2\left(PM_{ik}\right) = c_i + p_k + \varepsilon_{ik} \tag{21}$$

for an exon array. In this way, the problem of detecting alternative splicing is considered as a problem of outlier detection.

Let define $r_{ijk} = y_{ijk} - \hat{c}_i - \hat{p}_k$ as the residuals from fitting the standard model in first equation. Then for each exon *j* and sample *i*, a summary score based on the four residuals (one for each probe) from exon *j* and sample *i* gives a measure of the discrepancy d_{ij} in the expression of the exon in that sample. Several scoring functions could be used (mean, median, lower quartile, minimum of the absolute residuals), the median of the residuals in an exon gave the best tradeoff between sensitivity to the size and sign of the residuals and robustness to the small number of probes [19].

This gives a final score statistic,

$$F_{ij} = \underset{k \in exon\ j}{median}\left(\frac{r_{ijk}}{s} \right) \tag{22}$$

The estimate of the standard error *s* is the median absolute deviation of the residuals and this helps in comparing the scores between different genes. The term e_j is not estimated separately, because it is comprised into the probe estimates.

Two main differences between MiDAS and FIRMA are the type of summarization used to get exon/gene signals and the fact that MiDAS requires samples with replicates while FIRMA does not. But these two techniques were both tested on a reference data set with replicated experiments in each condition.

FIRMA algorithm is implemented in the ***aroma.affymetrix*** R library.

SPACE

SPACE (Splicing Prediction And Concentration Estimation) is an algorithm to predict and quantify alternatively spliced isoforms using microarrays. It has been developed to [20]

1. Estimate the number of different transcripts expressed under several conditions.
2. Predict the precursor mRNA (pre-mRNA) splicing structure.
3. Quantify the transcript concentrations including unknown forms.

This algorithm applies 'non-negative matrix factorization' (NMF) to the matrix of data [20]. NMF is a factorization for non-negative multivariate data. Given a matrix of non-negative data V, NMF finds an approximate factorization $V \approx W \cdot H$ into matrices with non-negative elements W and H.

When applied to microarray data, NMF separates the data matrix for each gene into the product of two positive components corresponding to the structure of the gene transcripts and their individual concentrations, respectively.

SPACE includes also an algorithm to determine the internal dimension of the factorization that is an estimate of the number of transcripts of each gene. SPACE original algorithm is written with MatLab 7.1 and is freely available online as additional file of its reference paper [20]. SPACE was implemented in R code to be evaluated with the benchmark experiment, afterwards presented (**section 1.3.3**).

The following table summarizes the different shown methods with the relative summarization techniques (if defined within the method), filters and statistical tools.

	Summarization	Filters	Statistic
MiDAS	PLIER	-	ANOVA based on SI
Rank Product	-	-	signal ranking
OS	-	-	t-test
ORT	-	-	t-test
SPACE	-	-	expressions factorization
MADS	similar to PLIER	crosshyb, BG	based on SI after BG
FIRMA	RMA	BG	score ranking

Table 1. The seven above-mentioned methods for alternative splicing events detection are based on different statistics and suggested to be used in association to some particular summarization and filtering tools.

In the next section, each method performance will be evaluated on a reference benchmark experiment, specifying how the data are previously summarized, if possible and which different filters where used to avoid noisy and unclear results in the detection of alternative splicing events. At the beginning, analyses with MiDAS and Rank Product were performed in Windows XP operating system. MADS, FIRMA and SPACE were also run in Windows XP. But Rank Product resulted to be quite slow and with the purpose of increasing the number of permutations (from 100 to 1000 or 10000), it was better to run it on UNIX server. The entire project was developed in R environment, on Windows XP or UNIX environment.

1.3 Results and conclusions

Dissecting an alternative splicing analysis workflow for GeneChip® Exon 1.0 ST Affymetrix arrays

In the article here presented [13] an exon-level data analysis workflow is dissected to test the performance of each step and optimizing the detection of ASEs. Tissues comparison is characterized by big changes in isoforms expression, which might not be the case in other situations. In tissues comparison only part of the TPs is known on the basis of published data, while in a spike-in data set true positives are known.

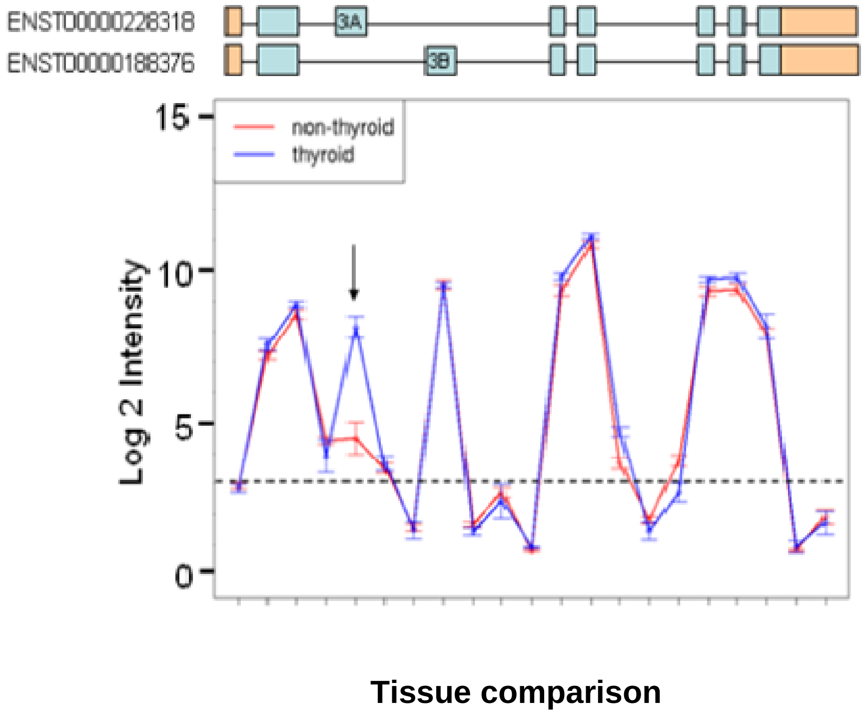


Fig. 2 Tissue-specific SLC25A3 transcripts from [21]. The expression plot shows the mean *log_2* intensity signals (with standard error bars) of core probe sets targeting SLC25A3 exons in the thyroid compared to non-thyroid tissue (bottom). The probe sets are plotted from left to right by genomic location (5' to 3'). The horizontal dashed line shows the mean *log_2* intensity of the negative control probe sets. Probe sets with intensities below this line are most likely unexpressed. In this case these probe sets are targeting either intronic regions or UTRs (in orange). Ensembl transcripts for SLC25A3 are shown below the plot. Probe sets with Benjamini-Hochberg corrected p-values less than 0.0001 are indicated by a black arrow.

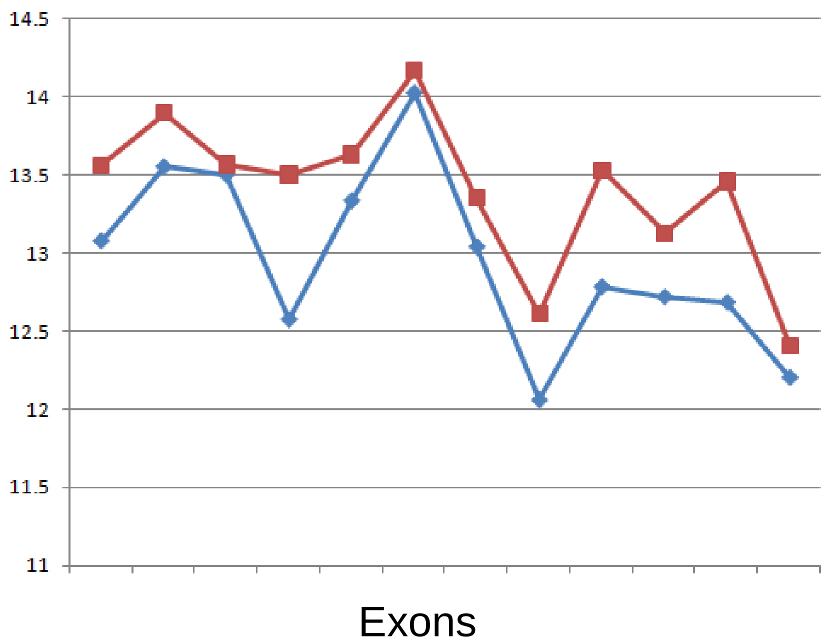


Fig. 3 Example of significant change in exonic *log_2* intensity between treatment (32.2, red) and control case (128, blue), in the benchmark experiment. It is evident that the exon in forth position is differentially expressed, with a *log_2* (FC) equal to 1.

Tissue splicing events are not the ideal instrument to test an analysis workflow. Hence, a semi-synthetic exon-skipping benchmark experiment from GeneChip® Exon 1.0 ST microarray data was built up for this evaluation. The final results point out that summarization methods (RMA, PLIER) do not affect the efficacy of statistical tools in detecting ASEs. However, data pre-filtering is mandatory if the detected number of false ASEs is meant to be reduced. MiDAS and Rank Product methods efficiently detect true ASEs but they suffer from the lack of multiple test error correction. The intersection of MiDAS and Rank Product results efficiently moderates the detection of false ASEs.

The last subsection concerns an attempt of improvement of the pipe-line with some other available statistical tools.

1.3.1 Introduction

GeneChip® Exon 1.0 ST is a new microarray platform developed and marketed by ***Affymetrix*** [22]. This microarray platform changes the conventional view of transcript analysis since it allows the evaluation of the expression level of a transcript by querying each exon component. This enables the study of specific alterations in splicing patterns such as those found in association with cancers [22].

The GeneChip® Exon 1.0 ST microarray platform is based on methods quite different from the 3' IVT arrays expression detection. Whilst the conventional ***Affymetrix*** GeneChips feature a probe set consisting of 11–20 probes selected from the 3' end of the mRNA sequence, the new all-exon arrays have 4 probes selected from each putative exonic region. To generate the target, Exon 1.0 ST arrays use T7 linked random hexamers for cDNA synthesis, instead of those of all previous ***Affymetrix*** expression arrays, which employed an oligo-dT linked T7 and thus required an intact poly-A tail. Importantly, this new WT Sense Target Labeling Assay generates DNA targets and therefore results in DNA/DNA duplex formation during hybridization, as opposed to DNA/RNA hetero-duplexes in conventional arrays. It has been shown that there is close agreement between the conventional ***Affymetrix*** 3' IVT arrays and the new Exon 1.0 ST arrays [23]. Furthermore, Exon 1.0 ST sensitivity of gene expression detection was shown to be in the same range of 3' IVT arrays [2]. Though at gene-level 3' IVT and Exon 1.0 ST show similar behavior, Exon 1.0 ST technology raises some issues about the computational instruments to be used for the analysis of exon-level data. ***Affymetrix*** proposed an analysis workflow based on pre-filtering of the expression data [7], transformation of exon-level intensity data in gene-level normalized values called Splice Index (**section 1.2.1**) and statistical validation based on an ANOVA based method based on measuring differences between an exon-level signal and aggregated gene-level signal called MiDAS (**section 1.2.2**).

There has however been no way to date of defining the efficacy of this workflow or of different statistical methods in the detection of alternative splicing events. The ideal instruments to evaluate the effect of data pre-processing and the efficacy of different statistical methods on differential expressions are benchmark spike-in experiments [24], where a limited number of transcripts are spiked-in at various concentrations in a common mRNA background.

In spike-in based experiments it is therefore possible to investigate differential expression sensitivity as a function of the false discovery rate (1-specificity). In this study a semi-synthetic exon-skipping experiment, encompassing 268 exon skipping events, was generated starting from the Latin-square spike-in experiment of Abdueva [2]. The semi-synthetic exon-skipping data set was used to evaluate the effects of data pre-processing as well as the performance of two statistical methods, MiDAS [7] and Rank Product [15], on ASEs detection.

1.3.2 Methods

Exon-skipping events were generated using experimental data, kindly provided by Abdueva [2]. MiDAS p-values were calculated using the software provided by ***Affymetrix*** in the APT tools (http://www.affymetrix.com). Rank Product (**section 1.2.2**) is a non-parametric statistics that detects items that are consistently highly ranked in a number of lists and the significance of the detection is assessed by a non-parametric permutation test [15]. RP was coded in R, modifying the available implementation (Bioconductor [3] ***RankProd*** package [14]), to be used for ASEs detection.

Specifically, in ASEs detection RP is run on the lists made by SIs (RP_{SI}) or intensities (RP_I) for all exon data set without considering their association to a specific gene and the significance of the detection is assessed using 500 permutations of those lists. Gene-level implementation of RP, i. e. running RP only on the subset of exons belonging to a specific gene, is computationally demanding and it is characterized by a very poor sensitivity. The modified RP method as well as all the filtering procedures are embedded in the Bioconductor ***oneChannelGUI*** [4] package.

1.3.3 Results

A benchmark experiment to validate ASEs detection methods

Exon skipping events were generated using the experimental data, kindly provided by Abdueva [2]. The Abdueva data set is a Latin-square experiment encompassing 25 genes, selected as ideal spike-in genes due to their expression absence in HeLa cells, which represents the mRNA background of the experiment.

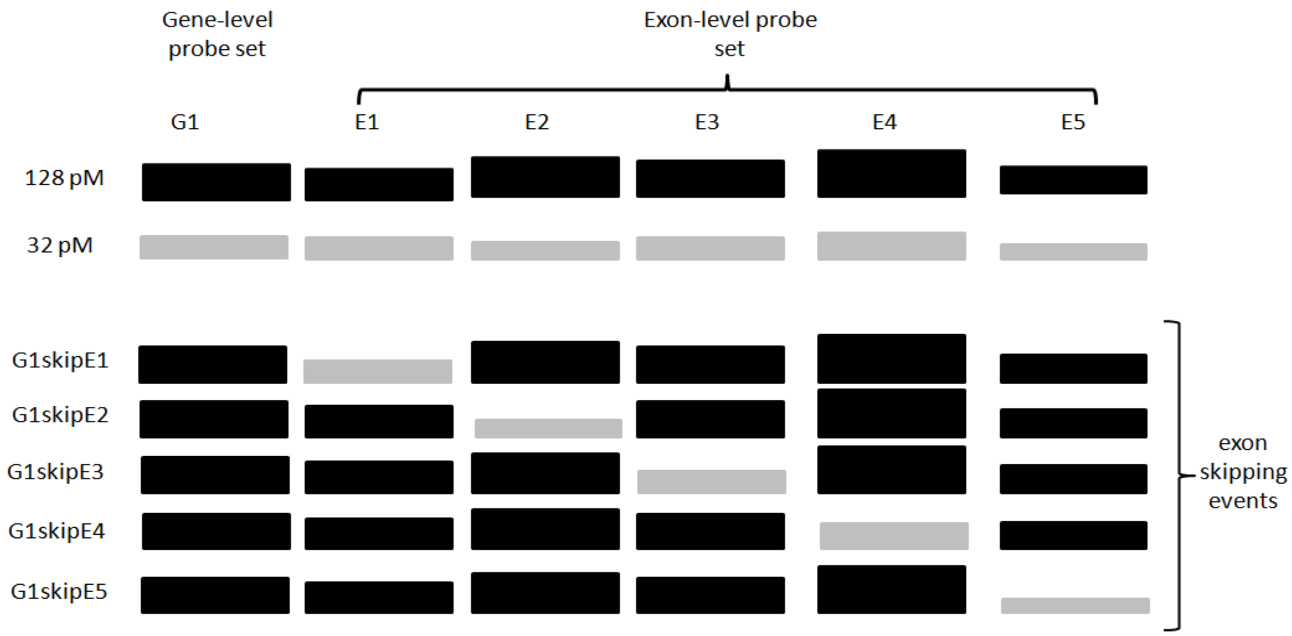


Fig. 4. Example of a set of exon skipping events, from [13]. The gene-level probe set (gene) G1 is made of 5 exon-level probe sets (exons) E1, E2, E3, E4, E5. Exon-level probe set signals associated with 128 pM spike-in are black whereas signals associated with 32 pM spike-in are gray. New genes are created combining exon-level expressions derived from different spike-in concentrations. In this specific case, the combination of 128 and 32 pM spike-in signals for gene G1 are used for the generation of 5 new genes (G1skipE1, G1skipE2, etc) each one characterized by a skipping event, given by the spike-in at 32 pM, in one of the 5 exons of gene G1. The unspliced exons are instead given by the 128 pM spike-in. For the sake of simplicity only one out of the three technical replicates is shown.

The spike-in concentrations were 0, 2, 32, 128 and 512 pM and the 25 genes were grouped in 5 subsets. Each experimental point was technically replicated three times for a total of 15 arrays. To build the exon skipping benchmark experiment 4 out of the 5 groups of spike-in genes (20 out of 25 genes) were used.

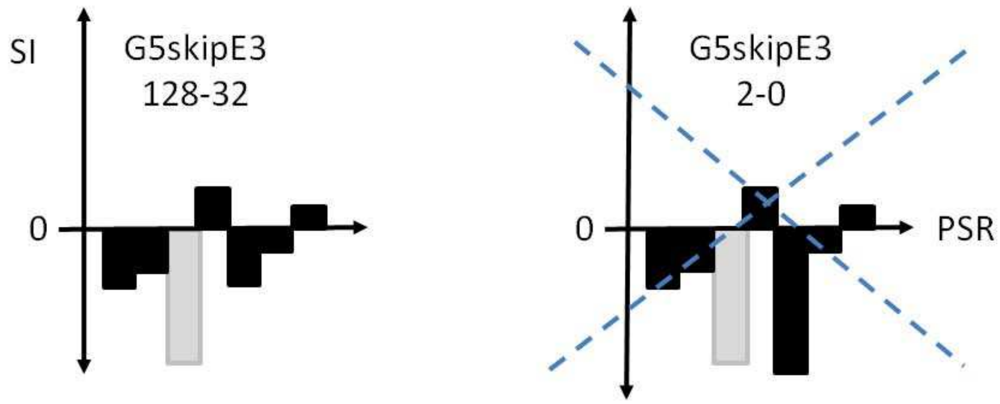


Fig. 5. Example of a set of exon-skipping cleaning procedure, from [13]. The cleaning procedure, applied to all new genes characterized by a skipping event, retains only those where the synthetic skipping event represents the smallest intensity or SI value within the exons belonging to the gene. Here, it is shown the example of gene G5, which is made of 7 exons and therefore produces 7 new genes, G5skipE1, G5skipE2, etc. In G5skipE3 gene, exon E3 should be the only exon characterized by the smallest SI. G5skipE3 gene is retained in the set 128-32, since E3 (gray) is characterized by the smallest SI within all 7 exons (black). The gene is instead removed in the set 2-0 since exon E5 has a SI smaller than the one of exon E3.

We focused on those because they were all part of the Exon 1.0 ST core annotation subset. The overall idea of the generation of synthetic exon skipping events is based on the availability of exon-level signals for spike-in genes. Therefore, it is possible to create new genes characterized by skipping events combining, for the same gene, exon-level expressions derived from different spike-in concentrations. An example is given in **Fig. 4**, where the combination of 128 and 32 pM spike-in signals for gene G1 are used for the generation of 5 new genes each one characterized by a skipping event in one of the 5 exons of gene G1.

In our semi-synthetic data set the new genes, characterized by skipping events, are generated using different associations of spike-in concentrations to evaluate the effect of signal intensity in the detection of alternative splicing. For each exon of the 20 genes we produced three sets of synthetic exon skipping events: 128-32, 32-2, 2-0. Specifically in the exon skipping set called 128-32 any of the new genes has all exons signals given by the $\log_2$ intensity (log2I) measured upon a spike-in of 128 pM unless the exon skipped, which has the log2I measured upon a spike-in of 32 pM (**Fig. 4**, G1skipE1, G1skipE2, etc.). The gene-level log2I is instead the one measured for the 128 pM spike-in (**Fig. 4**). Same design applies to the other two sets of exon skipping events, 32-2 and 2-0.

This semi-synthetic benchmark experiment embeds a total of 268 exon skipping events. Furthermore, the skipping events were manually inspected, in each of the three exon-skipping sets (128-32, 32-2, 2-0), in order to retain only those genes where the skipping event represents the smallest intensity signal or Splice Index (**section 1.2.1**) within each synthetic gene (**Fig. 5**).

This cleaning procedure yields:

- a total of 172 skipping events out of the original 268 for the 128-32 group, 195 for the 32-2 group and 179 for the group 2-0, if intensity data are used.
- a total of 174 skipping events out of the original 268 for the 128-32 group, 193 for the 32-2 group and 164 for the group 2-0, if SI data are used.

To identify exon-skipping events a comparison between two different conditions, i. e. unspliced versus spliced, is needed. Detection of exon-skipping events for the subset

128-32 was done comparing it to the unspliced set spiked in at 512 pM (called 512), for the subset 32-2 comparing it to the unspliced set spiked-in at 128 pM (called 128) and for the subset 2-0 comparing it to the unspliced set spiked-in at 32 pM (called 32). These comparisons embed a certain level of differential expression at gene-level. The expected gene-level differential expression is given by $log_2(128/512) = -2$ for the comparison of the 512 versus the 128-32 subset and by $log_2(32/128) = -2$ for the comparison 128 versus 32-2 subset. It is instead $log_2(2/32) = -4$ for the comparison 32 versus 2-0 subset.

RMA versus PLIER summarization

RMA and PLIER algorithms were used to combine the intensities belonging to the probes of each probe set to form one expression measure for each gene/exon-level probe set (summarization). The effect of these summarization methods on detection of alternative splicing events was investigated using MiDAS. A Receiver Operating Characteristic (ROC) curve was used to evaluate the effect of intensity summaries on alternative splicing detection (**Fig. 6**, continue lines). Our data suggest that the efficacy of detecting exon skipping events is not affected by summarization methods. On the other hand the reduction of the complexity of the data set, e. g. selecting only those ENSEMBL [25] genes associated with more than one transcript isoform (multiple mRNAs filter), strongly increases the sensitivity of the test (**Fig. 6,** dashed lines). Comparing the ROC curves of the 3 groups of data (**Fig. 6**, **Fig. 7**, **Fig. 8**) it is evident that multiple mRNAS filter throws out many more false values, after MIDAS analysis of the data set summarized with PLIER or RMA.

Filtering approaches to moderate multiple testing errors

A critical issue, highlighted in **Fig. 2**, is the important number of multiple testing errors that are accumulated if the full set of Exon 1.0 core data is used for the detection of ASEs. To moderate this critical issue, we decided to reduce the complexity of the data set filtering non-informative data (TN) before statistical analysis, using annotation and intensity based filters.

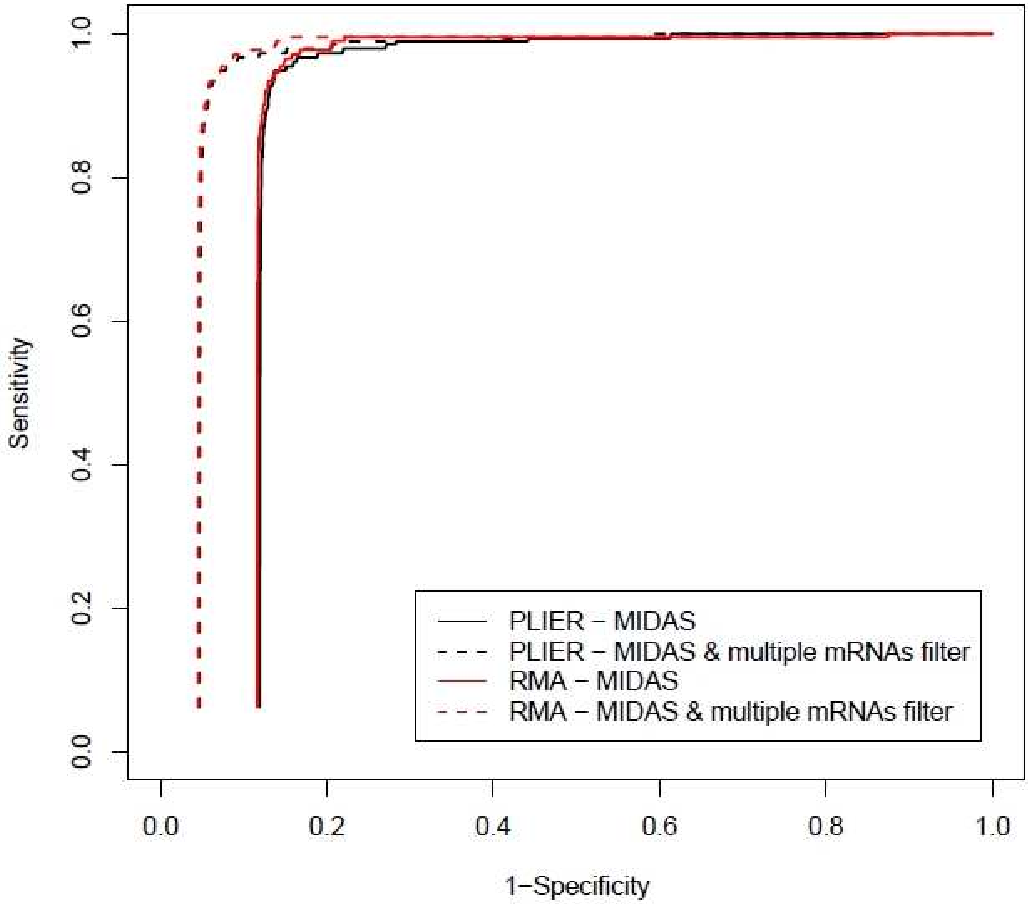


Fig. 6 [13] MiDAS exon skipping detection using RMA or PLIER summarization. ROC curves were used to identify the effect of data summarization on the detection of ASEs. ASEs were detected using MiDAS on the full core Exon 1.0 ST data set (continuous lines) using RMA (red line) or PLIER (black line). The same analysis was also applied to a subset of the core Exon 1.0 ST data set by encompassing only those gene/exon-level probe sets passing the multiple RNAs filter (dashed lines), i. e. those exons of genes associated to more than one mRNA isoform in ENSEMBL database.

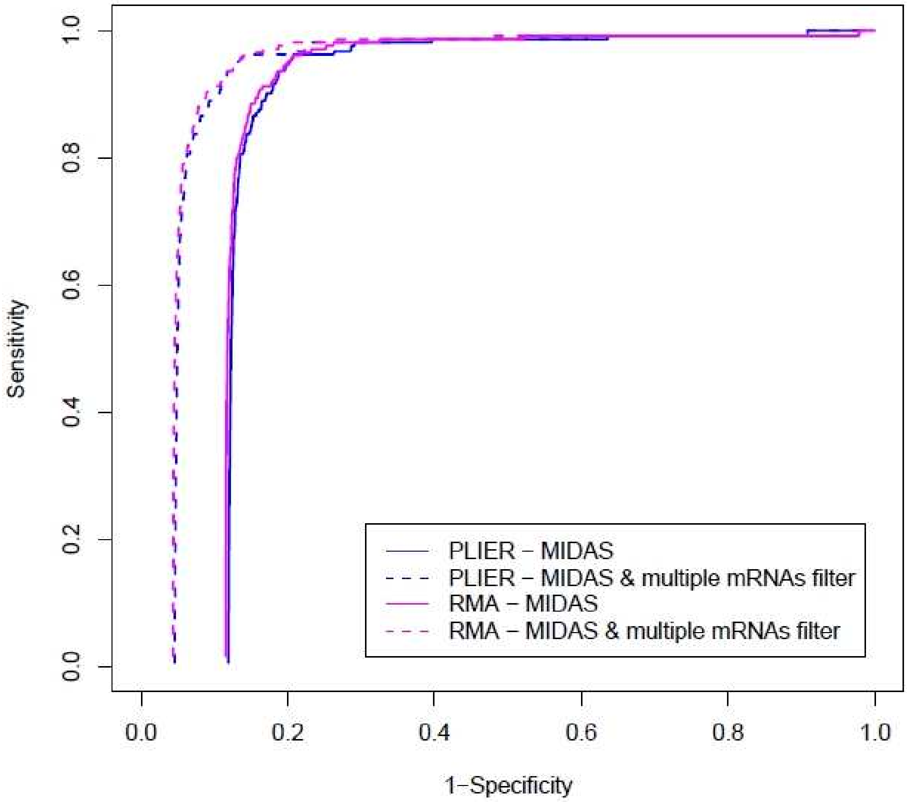


Fig. 7 Comparison of ROC curves of **128-32** group on the results obtained with MIDAS analysis, with data summarized in two different ways with PLIER or RMA.

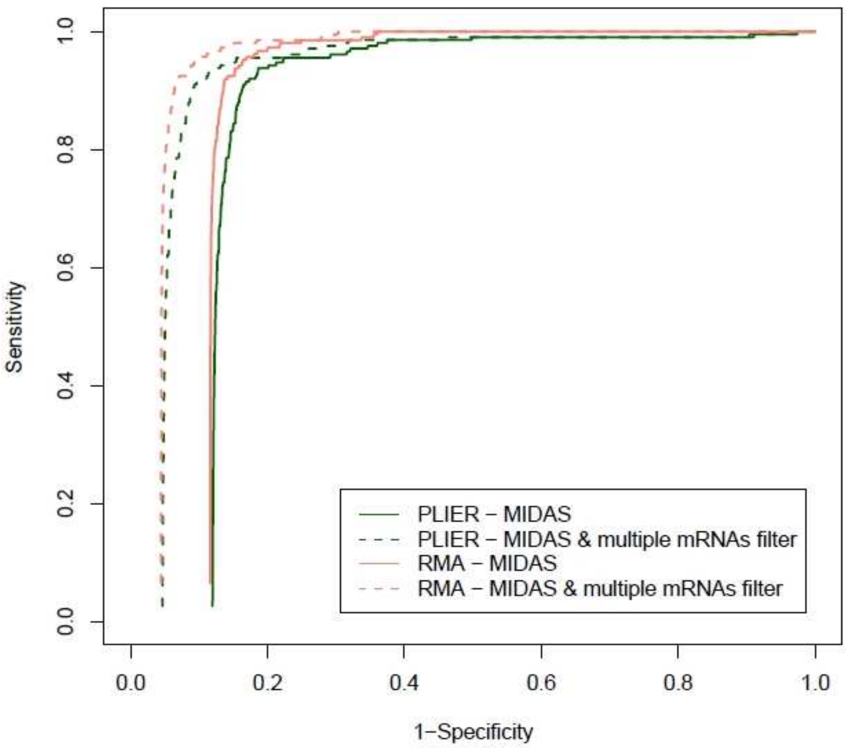


Fig. 8 Comparison of ROC curves of **2-0** group on the results obtained with MIDAS analysis, with data summarized in two different ways with PLIER or RMA.

Cross hybridization filter

We investigated the effect of removing those probe sets characterized by a certain level of probe promiscuity among transcribed sequences (cross hybridization filter). Specifically, using the exon-level probe set annotation information provided by ***Affymetrix***, we removed all probe sets where all the probes in the probe set perfectly match more than one sequence in the putatively transcribed array design content as well as those where the probes either perfectly match or partially match more than one sequence in the putatively transcribed array design content. This filter could have an important effect on the correct association of the gene expression signal. However, it affects a very limited number of exon-level probe sets and therefore it does not produce an important reduction of the size of non-informative data (**Table 2**). True Positives (TP), i. e. the semi-synthetic skipped genes previously described, are not affected by this filter since their exon-level probe sets are not annotated within the cross-hybridizing probe sets.

Multiple mRNAs filter

This filter uses the ***Affymetrix*** annotation that links each gene-level probe set to a specific GeneBank (GB) accession number (ACC), which represents the target sequence used to design the probes associated to a gene-level probe set. Then, Entrez Gene Ids (EGs) are retrieved querying with these ACCs a specific organism oriented Bioconductor annotation package (org.Hs.eg.db, org.Mm.eg.db or org.Rn.eg.db).

EGs are used to query ENSEMBL database and all ENSEMBL transcripts associated to any of them are retrieved. Subsequently, the filter procedure retains only those EGs associated to more than one ENSEMBL transcript. The EGs, retained by this filtering procedure, are mapped again to their gene-level probe sets. Multiple mRNAs filter strongly reduces the number of core exons because it retains only exons of genes which are linked to multiple transcripts in the ENSEMBL database and for this reason it results to be more effective than the other filters as shown both in **Table 2** and in **Fig. 2**. The new genes, with skipping events, generated in our data set are not affected by this filter since they do not exist in nature.

	128.32 vs 512		**32.2 vs 128**		**2.0 vs 32**	
	TP	**TN**	**TP**	**TN**	**TP**	**TN**
Multiple mRNAs	**172 (1.00)**	**71037 (0.31)**	**195 (1.00)**	**71307 (0.31)**	**179 (1.00)**	**71037 (0.31)**
Cross-hybridization	172 (1.00)	228264 (1.00)	195 (1.00)	228264 (1.00)	179 (1.00)	228264 (1.00)
DABG ≤ 0.05 (in 90% arrays)	172 (1.00)	197951 (0.86)	185 (0.95)	197951 (0.86)	170 (0.95)	197951 (0.86)

Table 2. Effect of annotation and intensity based filters on the selection of TP and reduction of unspliced exon set (TN). The effects of filtering by means of annotation (Cross Hybridization/Multiple mRNAs filters) or intensity signal (DABG filter) are evaluated using exon-skipping events at various concentrations.

DABG filter

In EXON 1.0 ST GeneChips, to determine if a given probe signal is detected above background (DABG), its intensity is compared to a distribution of background probes with the same G/C content.

		Fraction of TPs	**Enrichment of TPs**
Multiple mRNAs		0.727612	**268.8188**
DABG	***P-value:*** 0.01	0.3656716	56.2037
	P-value: 0.0001	0.3656716	66.25635
	P-value: 0.00001	0.3656716	71.07364
Cross-hybridization		0.727612	72.7612
Splicing Index	***Threshold:*** 0.001	0.727612	75.5265
	Threshold: 0.003	0.727612	81.04057
	Threshold: 0.005	0.5932836	70.70971

Table 3. Fraction of TPs is the number of detected TPs divided by the total number of known positive values. Enrichment of TPs is the number of TPs detected with respect to the total number of expected TPs. In red is shown the best enrichment in TPs and in blue the greatest fraction of TPs. It is evident from the values above reported that multiple mRNAs filter detects many more TP than any other filter, also more than Splicing Index filter with threshold 0.003. Because of this enrichment results it was decided to use as filter only multiple mRNAs as an important step of the pie-line to reduce the number of non informative probes.

A p-value is computed representing the probability that the signal intensity is part of the null distribution. Specifically the DABG p-value filter, used in this work, is designed to retain only probe sets characterized by a DABG p-value ≤ 0.05 in all the arrays. Although this filter reduces the data set under analysis (**Table 2**), it is much less effective than multiple mRNAs filter (**Table 2**). Increasing the stringency of this filter affects the total number of non-informative data (TN), which is reduced, but also part of the TPs are lost. DABG p-values could be useful in the detection and removal of low intensity signals which could produce misleading results when alternative splicing events are evaluated using the Splice Index, where signal intensity component is lost to remove the bias due to the presence of gene-level differential expression. However, in our data set a filter based on this approach is much less effective than that based on multiple mRNAs filter (**Table 2**).

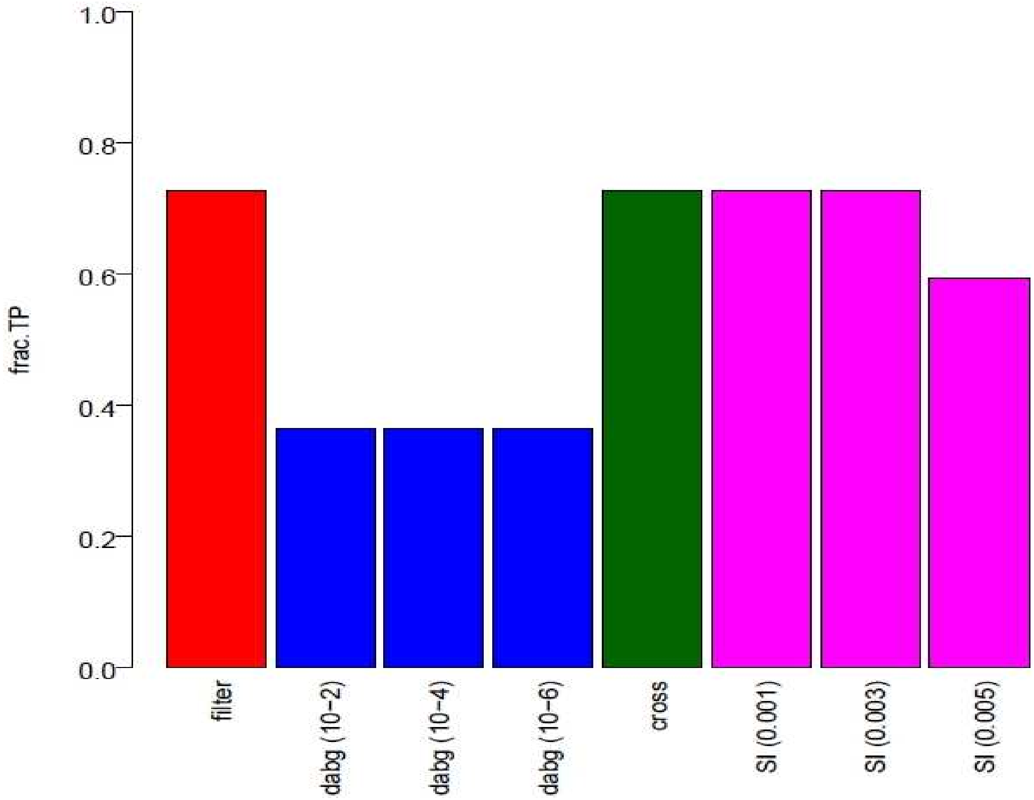


Fig. 9 Comparison of the **fraction of TPs** (frac.TP on y axis, Sensitivity) of the different filters that were used, trying to reduce the data set in analysis (**32-2** group). 'Filter' is the multiple mRNAs filter, 'cross' is the cross-hybridization filter. DABG and SI were tried with different thresholds.

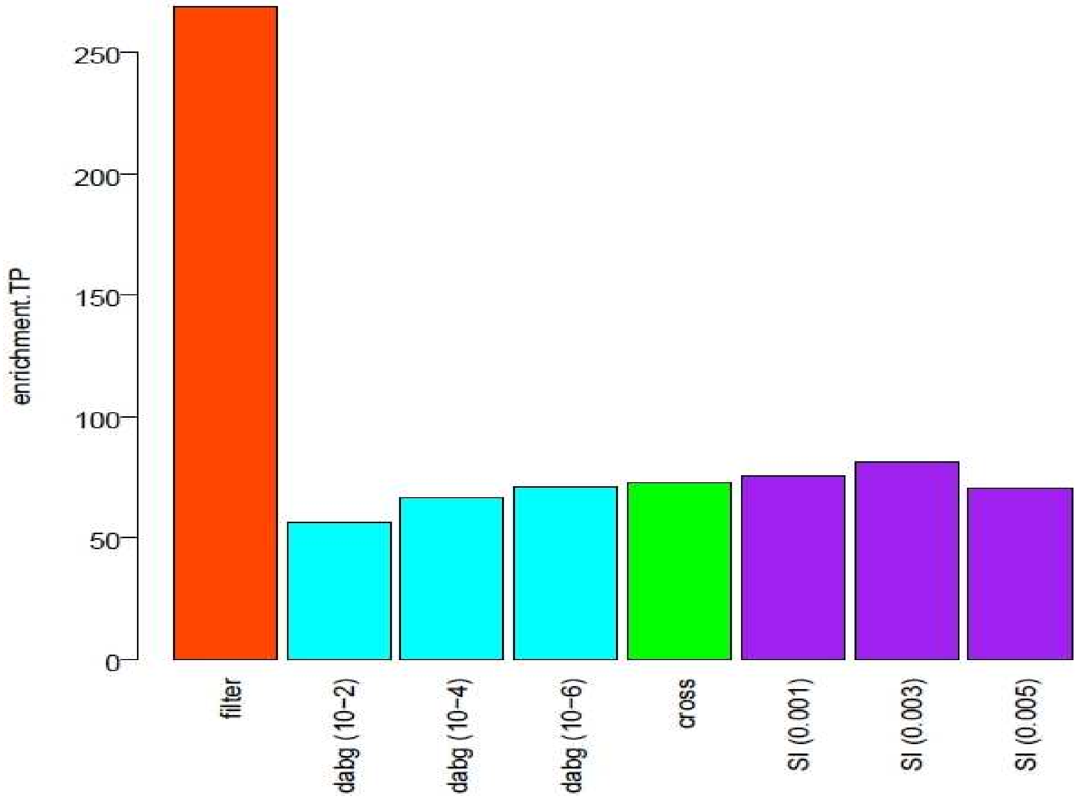


Fig. 10 Comparison of the **enrichment of TPs** (enrichment.TP on y axis) of the different filters that were used to reduce the data set in analysis (**32-2** group). 'Filter' is the multiple mRNAs filter, 'cross' is the cross-hybridization filter. DABG and SI were tried with different thresholds.

Efficacy of MiDAS and Rank Product in the detection of ASEs

We evaluated the efficacy of the detection of ASEs, using a linear model based algorithm (MiDAS) and a permutation based algorithm (Rank Product). MiDAS was applied on SI transformed data as RP was applied instead using SI, RP_{SI}, or directly to exon intensity signals, RPI. RP was implemented both at gene-level and exon-level. Gene-level implementation of RP indicates that the analysis is performed gene by gene and the permutations are generated within the list of exons of the same gene. Exon-level implementation of RP considers instead exons as items of a unique list, independent from their association with a gene. The exon-level implementation of RP has better sensitivity than that of the gene-level version and is faster since permutations are calculated only once and not gene by gene. Both MiDAS and RP seem to be effective in the detection of alternative splicing events independently from the presence of a certain level of gene differential expression and with limited dependency on gene-level intensity (**Fig. 11**). RP seems to perform slightly better than MiDAS. RPI (**Fig. 13**) gives the most homogeneous results independently of the intensity signals associated with ASEs (**Fig. 12**). Independently from the statistics in use, at p-value ≤ 0.05 (**Table 4**) the TPs detection is reasonably efficient for both methods, but is associated with a significant amount of False Positive values (FPs). We also evaluated, at the three intensity ranges under study, the number of TPs and Fps that can be detected intersecting all probe sets characterized by a p-value ≤ 0.05 both for MiDAS and RP (**Table 4**). The integration of the two statistical procedures improves the reduction of FPs without greatly affecting the sensitivity (**Table 4**).

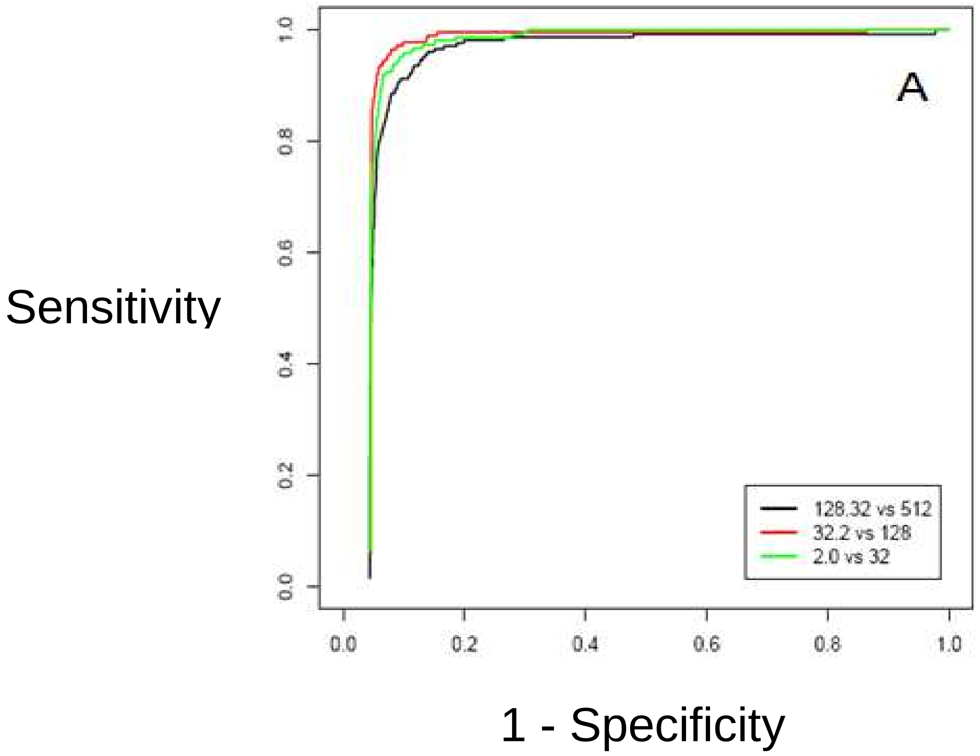


Fig. 11. ROC curves were used to detect the efficacy of MiDAS in the detection of ASEs.

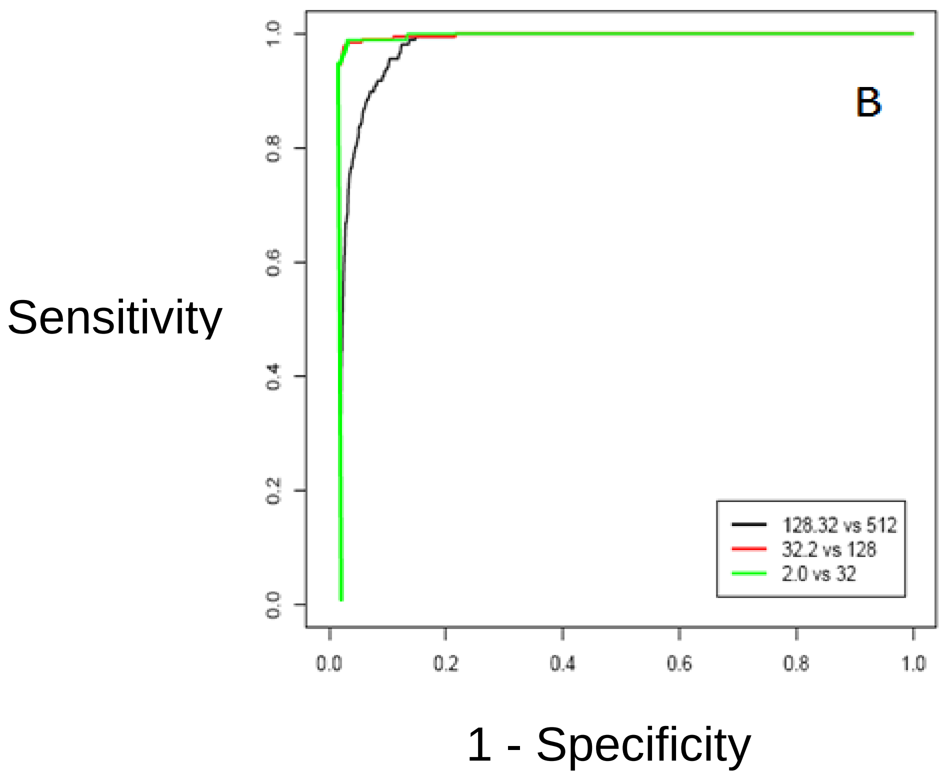


Fig. 12 ROC curves were used to detect the efficacy of RP_{SI} in the detection of ASEs. RP was calculated using exon signal normalized with respect to gene signal, i. e. SI.

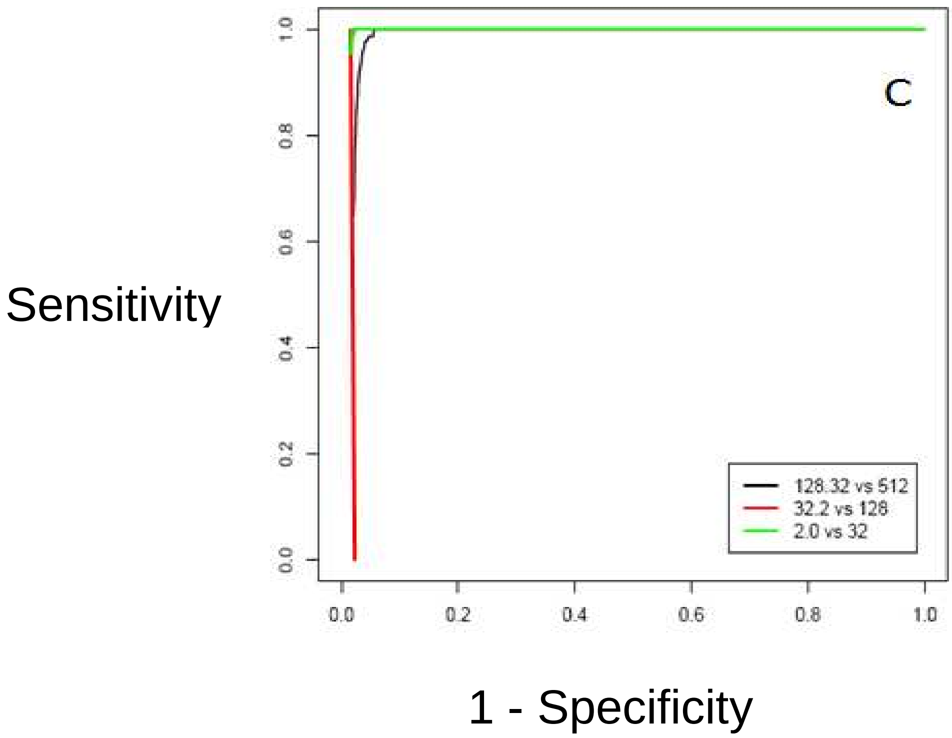


Fig. 13 ROC curves were used to detect the efficacy of RP_I in the detection of ASEs. RP_I was calculated using exon intensity signal without any further normalization.

The availability of a new instrument to study the behavior of transcription isoforms within a specific biological context, e. g. different cancer isolates, tissues, and differentiation/development stages, creates new opportunities for biologists. However, workflow for the detection of alternative splicing events using this new microarray technology has still to be investigated in order to define the importance of each analysis step and its strength and weakness. Our data point out that a major problem in ASEs detection is due to the multiple testing problem. In statistics, family-wise error rate (FWER) is the probability of making one or more false discoveries (FP) among all the hypotheses when performing multiple pairwise tests. Since FWER controlling procedures are often too conservative in high dimensional screening studies [26], they are rather weak if applied to exon-level analysis where the number of tests increases more than 10 times with respect to gene-level. For example, the human core data set is made of 22011 gene-level probe sets and 287329 exon-level probe sets. A better balance between the raw p-values and the stringent FWER-adjusted p-values may be provided by false discovery rate controlling and related procedures [26]. Benjamini & Hochberg [26] and Benjamini & Yekutieli [26] have developed efficient FDR controlling procedures currently called BH and BY. However, such approaches cannot be used to moderate multiple testing problems in exon-level analysis since, generally, the raw p-value distribution obtained with MiDAS is not uniform in the non significant range. Furthermore, in the case of BH, the assumption that the tests are independent is not fulfilled since exons belonging to the same gene are clearly correlated.

	128.32vs512		**32.2vs128**		**2.0vs32**	
	TP	**FP**	**TP**	**FP**	**TP**	**FP**
MIDAS + multiple mRNAs filter	119 (0.68)	2416 (0.03)	176 (0.91)	2319 (0.03)	138 (0.84)	2338 (0.03)
Rank Product	174 (1.00)	12941 (0.18)	193 (1.00)	11883 (0.17)	164 (1.00)	9989 (0.14)
Intersect MIDAS & Rank Product	**119 (0.68)**	**436 (0.006)**	**176 (0.91)**	**424 (0.006)**	**138 (0.84)**	**375 (0.005)**

Table 4: MiDAS and RP alternative splicing detection. RP_I is the Rank Product calculated using the intensity signals without SI calculation. Statistical analyses done using MiDAS or RP_I, calculated using intensity signals, at p-value ≤ 0.05 are contaminated by a significant number of Fps due to the multiple test problem. The intersection of the results using the two methods significantly reduces the number of Fps.

On the basis of the impracticality of applying conventional methods to moderate FWER, the reduction (filtering) of the data set size of previous statistical testing is, in our opinion, mandatory.

Our data point out that a significant reduction of the data set size can be realized by considering only probe sets associated with at least two alternative spliced isoforms in the ENSEMBL database (multiple RNAs filter). However, this approach limits the strength of the analysis since it cannot be applied in the case of the identification of non-annotated isoforms. If a study focuses on the discovery of non-annotated isoforms, an intensity filter, e.g. DABG p-values filter, can be used although its effect is not as strong as that based on annotation (**Table 2**). In this case, it would be necessary to clean the results of the large amounts of false positives, validating data by using alternative technologies such as the high-throughput re-sequencing techniques, e. g. Solexa (Illumina) or SOLiD (Applied Biosystems). These would however increase the complexity of the analysis due to the high computational demands of these techniques. We also investigated the performance of two statistical methods, one based on linear model analysis (MiDAS), developed by ***Affymetrix*** for the detection of ASEs, and another non-parametric (RP). Both methods, applied at exon-level and thus not taking into consideration the association of an exon to a specific gene, perform quite well in the detection of the true exon skipping events embedded in our data set (**Fig. 11**). The amount of FPs associated to an arbitrary p-value threshold of 0.05 is in both cases very high (**Table 2**) and the application of a more stringent p-value threshold reduces the number of FPs but also impacts negatively on TP rate. However, since the two statistics used for ASEs detection are based on completely different assumptions, it is feasible that random events (FPs) contaminating the TPs will not be the same. Therefore, the intersection of the results obtained by both statistics, given an arbitrary p-value threshold, effectively reduces FPs (**Table 4**). Since at the present time statistics specifically devoted to the detection of ASEs which also address the multiple test problem are not available, our approach represents an efficient temporary solution for moderating FWER.

1.3.4 Conclusions

The semi-synthetic data set presented here represents a suitable instrument for testing the efficacy of new statistics for exon-level analysis. Furthermore, it allowed us to test the efficacy of a basic workflow (**Fig. 14**) for ASEs using a GeneChip Exon 1.0 ST platform. However, our data highlights that more work is needed to design powerful instruments for ASE detection which must take into account the multiple testing problem.

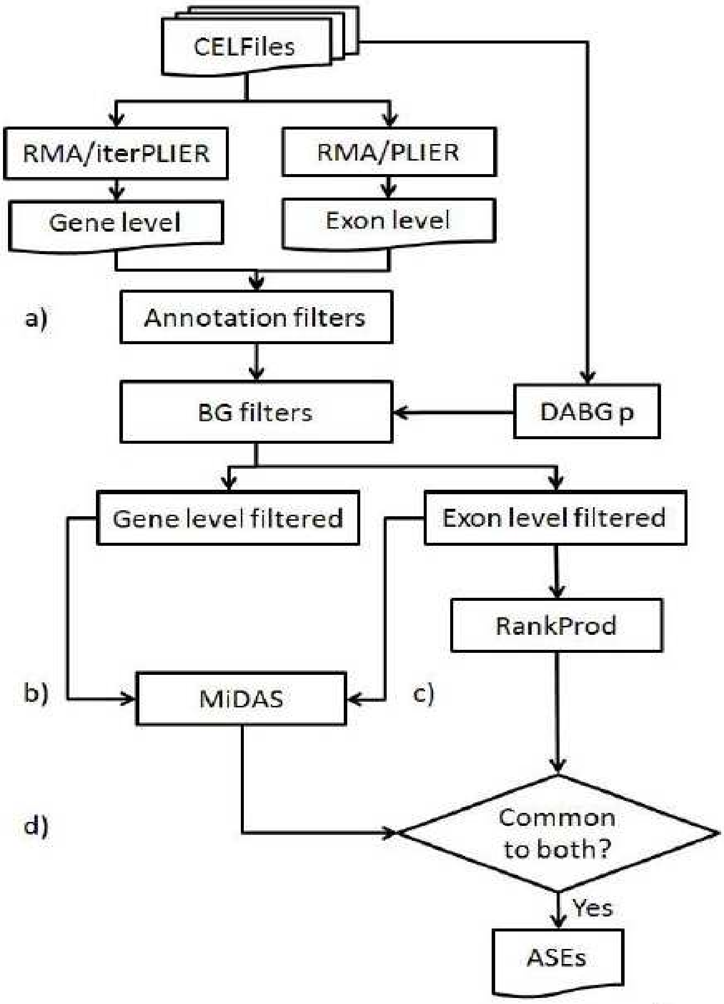


Fig. 14. Workflow for exon-level analysis. Workflow proposed for the detection of ASEs. **a)** The number of probe sets to be considered for the analysis is reduced on the basis of ENSEMBL isoform knowledge (multiple RNAs filter). Eventually, a filter based on the quality of the intensity signal (DABG filter) might be considered as an additional filter. **b-c)** Statistical analysis is done using a model based algorithm (MiDAS) and a non-parametric algorithm (RP). d) Intersection of data derived by the two statistical analyses, using a common arbitrary p-value threshold (e. g. 0.05), is used to reduce the number of FPs.

1.3.5 Pipe-line improvements

The performance of other methods was evaluated while trying testing MiDAS and Rank Product: OS and ORT. Then between the end of the year 2008 and March 2009 I worked, in chronological order, with MADS, FIRMA and SPACE. These methods did not give the expected good results. In the following table are shown the summarization techniques used to get mean intensities from the raw files and which filters were applied to reduce the number of false values. The following sections briefly explain why OS, ORT, MADS, FIRMA and SPACE methods were no more taken into account. Then follows a brief description of ***limma*** R package and ***FEVS,*** showing that intersecting MiDAS and RP results is the best choice because in this way a greater number of TPs is detected than with ***limma*** and ***FEVS***.

	Summarization	**Filters**	**Statistic**
MiDAS	PLIER, **RMA**	**SI, BG, DABG, crosshyb, multiple mRNAs**	ANOVA based on SI
Rank Product	**PLIER, RMA**	**SI, BG, DABG, crosshyb, multiple mRNAs**	signal ranking
OS	**RMA**	-	t-test
ORT	**RMA**	-	t-test
FIRMA	**RMA**	BG	score ranking
MADS	similar to PLIER	crosshyb, BG	based on SI after BG filter
SPACE	**RMA**	-	expressions factorization

Table 5. The seven methods for alternative splicing events detection are based on different statistics and suggested to be used in association to some particular summarization and filtering tools. defined within the method. Making a comparison with **Table 1**, in bold are highlighted the new summarization and filtering tools that were used, beyond the one already proposed in association with the method description.

OS – Outlier Sum

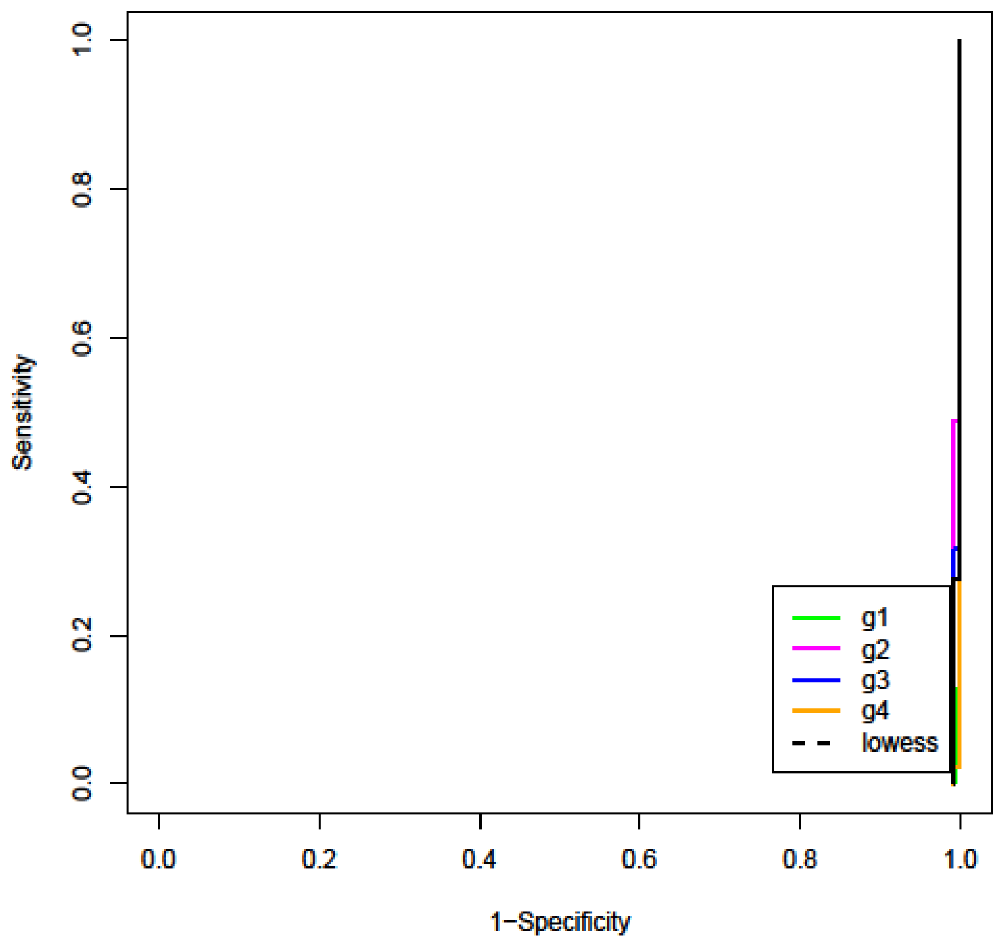


Fig.15 OS results (Baolin Wu version), 1000 points and 100 permutations. Too low Specificity.

As already mentioned in **section 1.2.2**, there are two versions of Outlier Sum statistic, one proposed by Tibshirani and Hastie [16] and an equivalent one proposed by Baolin Wu [17]. Tibshirani and Hastie statistic was first used (implemented in R code from the formula in [16]) but then it was worth wondering whether it was different or faster to use the version of Baolin Wu, which has a structure more similar to Outlier Robust T-test, directly including data standardization in the outlier sum detection. Both versions were applied to the semi-synthetic data set 100 and 1000 times, each time permuting the labels of the treated samples. The first OS version was slow and surprisingly found less TPs than the second one. Both versions were tried at gene-level and the results with both methods were quite poor: too few TPs were found with a very low Specificity.

ORT – Outlier Robust T-test

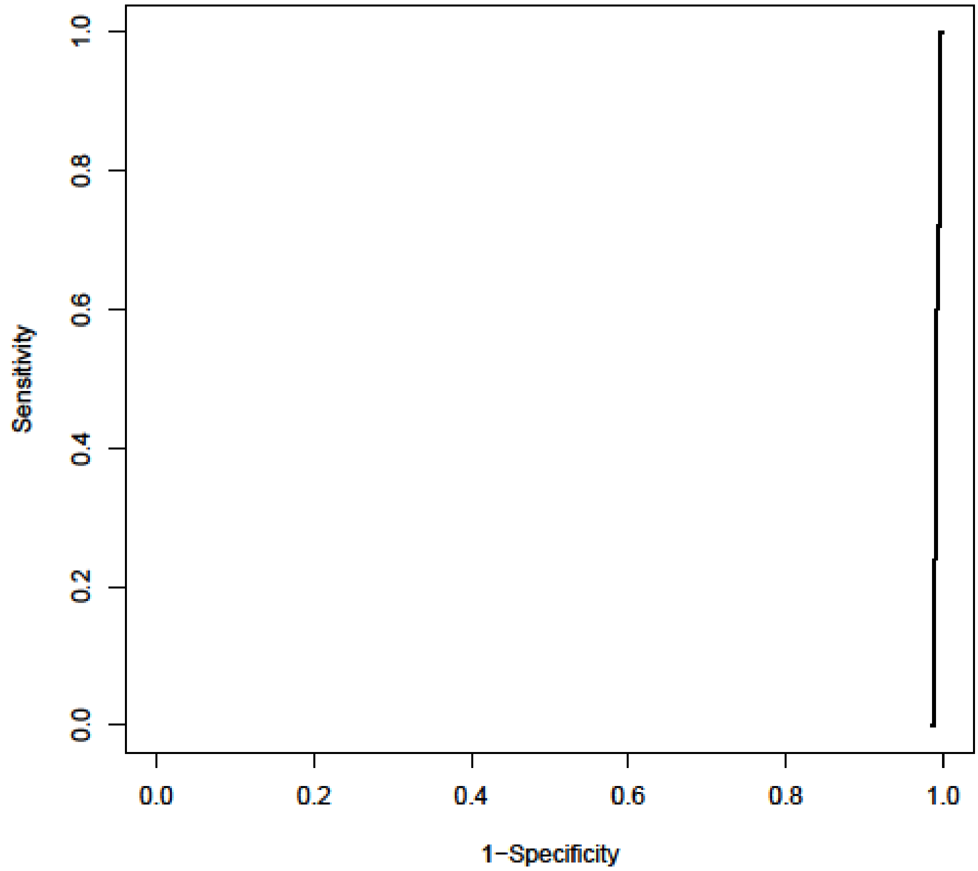


Fig.16 ORT with 1000 points and 100 permutations. Too low Specificity.

ORT was run with 1000 points and 100 permutations (OS and ORT were run also with 1000 permutations but the results were the same and the analysis was too slow). Both the results of OS methods and ORT are scores, not p-values and are not in the range [0,1a]. To get p-values instead of scores the genes were divided in classes with respect to the number of exons that constitute them (e. g. class 1 - genes with 12 exons), resulting in 103 gene classes. For each class of genes random genes (1000 exon permutations) were generated with the same number of exons, taken within the exons of that class of genes. Simulations were done using 1000 points (0.1-100 by 0.1) as thresholds. At each point the number of scores were overcoming this threshold and to them was associated 1, while 0 to the ones which did not. Then p-values were computed multiplying each 1/(total number of values).

As shown in **Fig.15** and **Fig.16** the results were unsatisfactory and after several months of trials, increasing the number of permutations and points, with and without SI normalization of the data, analyses with OS and ORT (one week long) were definitively abandoned.

MADS - Microarray Analysis of Differential Splicing

The data set provided by MADS article [18], constituted of three control replicates and three treatment ones, was analyzed with our algorithm, filtering out the genes with multiple mRNAs and applying then MiDAS and Rank Product and intersecting

their results to verify whether the number of true positive and negative values found was greater than the number found with MADS.

The semi-synthetic data set (**section 1.3.3**) could not be analyzed in MADS algorithm because it takes only CEL files (not modified raw data). The data set in article [18] came from a mouse experiment realized with Mouse Exon 1.0ST arrays and MADS analysis provided p-values related to each transcript; p-values under the threshold 0.05 were TPs, above were TNs. Therefore we used the MADS data set to evaluate the performance of our pipe-line. MADS data set contains: 40 TPs (i. e. validated splicing events detected with MADS), 23 TNs gold-standard (i. e. already known) values (exons). At exon-level, 24 TPs, 20 TNs were found with MADS. Filtering out these exons with the multiple mRNAs filter: 19 TPs and 14 TNs were found. Then running MiDAS: 4 TPs, no TNs with MiDAS ; while with Rank Product 7 TPs and 1 TN. In conclusion, MADS finds more TPs than the intersection of MiDAS and Rank Product, but MADS results are contaminated by a large number of TNs.

FIRMA – Finding Isoforms using Robust Multichip Analysis

The semi-synthetic data set and the one provided by MADS reference article [18] were analyzed with FIRMA algorithm, implemented in the Bioconductor [3] library ***aroma.affymetrix***. FIRMA is a method for detecting alternatively spliced exons in individual samples without replication and it was used with samples with replicates. FIRMA results were difficult to be interpreted because a threshold to compute TPs and TNs could not be defined.

SPACE - Splicing Prediction And Concentration Estimation

This algorithm deconvolutes the exon array data in the transcription isoforms associated to them. SPACE was written in MatLab and it was translated into R code. We tested it with the semi-synthetic data set but it produces a higher number of transcripts than the real ones, with the semi-synthetic data set and the data used to validate SPACE [20]. SPACE did not work well when increasing the number of its internal iterations (normally set to 1). Another problem was that a parameter representing the predicted maximum number of transcripts of a gene had to be equal to 10 (default value), if else different results were obtained as many times SPACE was run. Even the author of the program agreed with the above-mentioned problems and we decided to consider SPACE results not reliable.

Comparison of MIDAS/RP intersection with limma and FEVS

The benchmark experiment (**section 1.3.3**) was tested on ***limma*** R package [21] for differential expression analysis for microarray data and ***FEVS - Filtering Enhanced Variable Selection*** [27]. ***FEVS*** is a new multiple testing strategy for identifying differentially expressed variables, based on the combination of several filtering methods, instead of focusing only on a particular one. The authors of ***FEVS*** proved that it controls the FDR and that it gains sensitivity in the detection of truly differentially expressed elements. The following table, showing the results obtained running the analysis of the benchmark experiment on MiDAS, RP, ***limma*** and ***FEVS***, points out that ***FEVS*** has a very good control of FDR but fails to detect many TPs. A better performance is given by ***limma***, although in case of high expression levels the detection of TPs is very poor, which might be due to the increasing of variance as a consequence of high expression levels. The intersection between MiDAS and RP still detects the highest number of TPs but is less efficient in controlling the FDR with respect to the other two methods.

	Splicing set 128.32 vs 512		Splicing set 32.2 vs 128		Splicing set 2.0 vs 32	
	TP (Sensitivity)	FP (1-Specificity)	TP (Sensitivity)	FP (1-Specificity)	TP (Sensitivity)	FP (1-Specificity)
MiDAS	119 (0.68)	2416 (0.03)	176 (0.91)	2319 (0.03)	138 (0.84)	2338 (0.03)
RP_I	174 (1.00)	12941 (0.18)	193 (1.00)	11883 (0.17)	164 (1.00)	9989 (0.14)
MiDAS & RP_I intersection	119 (0.68)	436 (0.006)	176 (0.91)	424 (0.006)	138 (0.84)	375 (0.005)
FEVS $p \leq 0.05$ u=10	1 (0.006)	41 (0.0005)	26 (0.13)	9 (0.0001)	23 (0.13)	18 (0.0002)
Limma $p \leq 0.05$ BH	18 (0.10)	182 (0.002)	136 (0.70)	71 (0.0008)	90 (0.50)	153 (0.0018)

Table 6. The number of true (TPs) and false positive values (FPs) found with five methods of detection of alternative splicing events is here shown. It is evident that the intersection of the results obtained with MiDAS and RP_I gives a high number of TPs, decreasing much the number of FPs.

Genome-wide Search For Splicing Defects Associated with Amyotrophic Lateral Sclerosis

In this article we presented [28] a study in which we tried to detect gene and isoform specific events associated to the Amyotrophic Lateral Sclerosis (ALS).

SOD1 enzyme is a powerful antioxidant that protects the body from damage caused by superoxide, a toxic free radical. It has been proposed that defects in splicing of some mRNAs, induced by oxidative stress, can play a role in ALS pathogenesis. Alterations of splicing patterns have also been observed in ALS patients and in ALS murine models, suggesting that alterations in the splicing events can contribute to ALS progression.

Using Exon 1.0 ST GeneChips, the SH-SY5Y neuroblastoma cell line has been profiled after treatment with paraquat, which by inducing oxidative stress alters the patterns of alternative splicing. Furthermore, the same cell line stably transfected with wt and ALS mutant SOD has also been profiled. The integration of the two ALS models efficiently moderates ASE false discovery rate, one of the most critical issues in high-throughput ASEs detection.

1.3.6 Introduction

Amyotrophic Lateral Sclerosis (ALS) is a progressive, usually fatal, neurodegenerative disease caused by the degeneration of motor neurons, the nerve cells in the central nervous system that control voluntary muscle movement [29]. As a motor neuron disease, the disorder causes muscle weakness and atrophy throughout the body as both the upper and lower motor neurons degenerate, ceasing to send messages to muscles [29]. ALS is one of the most common neuromuscular diseases worldwide, and people of all races and ethnic backgrounds are affected. One to 2 people per 100,000 develop ALS each year. ALS most commonly strikes people between 40 and 60 years of age, but younger and older people can also develop the disease. Men are affected slightly more often than women. A definitive cause for ALS is not clear and the onset of the disease has been linked to several factors, including: exposure to viruses, neurotoxins, or heavy metals; genomic mutations; immune system and enzymatic abnormalities. "Familial ALS" accounts for approximately 5%–10% of all ALS cases and is caused by genetic factors. Of these, approximately 1 in 10 are linked to a mutation in copper/zinc superoxide dismutase (SOD1), an enzyme responsible for scavenging free radicals. This enzyme is a powerful antioxidant that protects the body from damage caused by superoxide, a toxic free radical. Free radicals are highly reactive molecules produced by cells during normal metabolism. Free radicals can accumulate and cause damage to DNA and proteins within cells. Although it is not yet clear how the mutant SOD1 gene mutation leads to motor neuron degeneration, selective vulnerability of motor neurons likely arises from a combination of several mechanisms, including protein misfolding, mitochondrial dysfunction, oxidative damage, defective axonal transport, excitotoxicity, insufficient growth factor signaling, and inflammation [29].

Furthermore, alterations of splicing patterns have also been reported in ALS patients and in ALS murine models, suggesting that alterations in pre-mRNA splicing events can contribute to ALS progression [30] [31]. Recently, it has been proposed that defects in splicing of some mRNAs, induced by oxidative stress, can play a role in ALS pathogenesis. The recent commercialization by ***Affymetrix*** of Exon 1.0 ST GeneChips (Exon GeneChips) allows the definition of both transcription patterns and of alternative pre-mRNA maturation events. Using this microarray platform we have profiled two ex vivo ALS models to identify the mRNA isoforms associated with ALS disease.

1.3.7 Methods

Benchmark experiment

A semi-synthetic exon-skipping events data set [13] was used to evaluate the limits of statistical methods used in the detection of alternative splicing events (ASEs). ASEs were detected using the model based method developed by ***Affymetrix***: MiDAS. MiDAS is an ANOVA based method measuring differences between an exon-level signal and aggregated gene-level signal. MiDAS p-values were calculated using the software provided by ***Affymetrix*** in the APT tools. Data were analyzed on R 2.7.0 and Bioconductor 2.2 [3]. Gene/exon-level expression summaries were generated using RMA algorithm [10] and quantile [32] sketch normalization by means of ***Affymetrix*** APT tools as suggested by Della Beffa et al. in [13].

ALS experiments

Alternative splicing events (ASEs) were detected using two experimental models: paraquat neurodegeneration model and ALS SOD model. Paraquat model: paraquat treatment on SH-SY5Y neuroblastoma cell line was carried out as described by Maracchioni and coworkers [33]. ALS SOD model: SODSH-SY5Y stably transfected with SOD1 wt and ALS mutant SOD1 G93A were grown as those for the paraquat experiment. Four prototypic situations were investigated: Paraquat experiment: SH-SY5Y neuroblastoma cell line with (para.t) and without (para.n) paraquat treatment. SOD experiment SHSY5Y stably transfected with wt (sod.n) and ALS mutant SOD1 G93A (sod.t). Each condition was replicated 5 times. After extraction and quality check 1.5 μgs total RNA was subjected to removal of ribosomal RNA following the procedure suggested by ***Affymetrix***. The resulting total RNA was then used to created the biotin-labeled library to be hybridized on GeneChip® Exon 1.0 ST human microarrays following the procedure described by the manufacturer. GeneChips hybridization, washing and scanning was done as suggested by the manufacturer. The resulting CEL files were analyzed using ***oneChannelGUI*** 1.6.5 [4]. ASEs events were detected by MiDAS comparing para.t versus para.n and comparing sod.t versus sod.n. Only ASEs characterized by showing a MiDAS p-value ≤ 0.05 in both experimental models were selected. Exon-level probe set mapping was performed by BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi), interrogating ASPIC database [34] and the genome browser at NCBI.

1.3.8 Results

ALS models microarray data analysis

The results obtained with our benchmark experiment clearly indicate a strong lack of specificity of MiDAS method, that cannot be moderated by conventional type I error correction methods. Della Beffa et al. [13] has shown that the integration of results generated by parametric and non-parametric ASEs detection methods strongly reduce type I errors. However, in our experimental framework we have to handle not only type I errors but also need to discriminate between the ALS-specific neurodegerative effect induced by paraquat and its intrinsic toxicity effect. To address both the moderation of multiple testing errors in ASEs detection and extrapolation of the ALS-specific events from the toxicity effect of paraquat we used a biological approach. Transcription profiling was done on two ex vivo models for ALS: paraquat treatment and SOD1 expression in SH-SY5Yneuroblastoma cell line. Four prototypic situations were investigated: Paraquat experiment: SH-SY5Y neuroblastoma cell line with (para.t) and without (para.n) paraquat treatment. SOD experiment SH-SY5Y stably transfected with wt (sod.n) and ALS mutant SOD1 G93A (sod.t). Gene/exon-level expression summaries were generated using RMA algorithm [10] and quantile [32] sketch normalization, analyzing Paraquat and SOD experiments independently. Sample group homogeneity was confirmed by Principal Component Analysis and hierarchical clustering (not shown). ASEs were detected in each of the two experiments using the model based algorithm MiDAS ($p \leq 0.05$). ASEs detected in paraquat and SOD experiments were respectively 2778 and 1974. 105 exon-level probe sets alternatively spliced and associated with 82 gene-level probe sets were found in common between the ASEs detected in the two model systems. Subsequently, the presence of a common trend between the paraquat and SOD experiments was detected at exon-level using the integrative correlation coefficient (IC) [36] applied on SI. 49 exon-level probe sets showed a common trend between paraquat and SOD experiments as instead 56 probe sets showed an opposite trend. Analyzing only the subset of probe sets characterized by a common trend within the two model systems, 35 out of 49 probe sets were associated with reference sequence transcripts and more than one exon probe set was mapping on the same transcript. Within those transcripts we have identified 7 ASEs associated with 5’ end exons and 7 ASEs associated with internal coding exons of known/predicted mRNA isoforms. We have also detected 1 ASE associated 5’ end exons and 3 ASEs associated with internal coding exons of genes where the splicing event is associated with an exon in common with all the known and predicted transcript isoforms. It is notable that the level of the ASEs is relatively low in intensity, in general it is represented by a variation of approximately 50% SI signal. We have also evaluated the SI mean centered distribution of the ALS associated ASEs in the tissue data set provided by ***Affymetrix*** and encompassing 10 adults tissues. The ALS associated exon-level probes show a complex pattern in the analyzed tissues.

1.3.9 Conclusions

Our results, generated using a semi-synthetic data set and real data, allows the generation of general and ALS-specific conclusions.

General conclusions

In a biologically defined framework, ASEs are represented more prevalently by changes in the ratios between the transcribed isoforms than the appearance of new isoforms. This results in relatively small exon-level expression changes, as observed by ourselves in ALS ASEs. Furthermore, the measurement of exon-level expression based only on 4 probes is less stable than that performed at gene-level (> 10 probes). This results in a relatively high fluctuation of the raw exon-level intensity signal measured in different arrays. The above mentioned criticalities combined with the lack of a specifically devoted statistical framework highlights the need for performing exon-level analysis using a high number of replicates, i.e. in our case we used five replications for a cell line based experiment. And the mandatory need for reducing false discovery rates in ASEs detection by taking advantage of biological instruments, i.e. intersection of data derived from different experimental models. ALS specific conclusions: Our analysis indicates that ASEs are part of the ALS phenotype in ex-vivo models of the pathology. However, the presence of common splicing events characterized by opposite trends in the two models might have two possible explanations: i) the paraquat model suffers from the lack of a non optimal setting of the paraquat treatment to simulate a chronic ALS effect as can be simulated by the stable transfection of mutant SOD gene. Ii) the mechanisms of action of paraquat and SOD in neurodegeneration, although both linked to oxidative stress, only partially overlap. This observation causes speculation as to whether the deregulation of the balanced expression of gene isoforms is involved in ALS and cannot be linked to the specific functionality of a few gene isoforms.

A literature search of the combination of gene name ALS associated ASEs and the "neurodegeneration" keyword highlighted the importance of NDRG2 and SOX9 genes in neurodegeneration. Specifically NDRG2 is particularly interesting since it is associated with Alzheimer's disease [37] and the first member of the NDRG family has been thoroughly studied as an intracellular protein associated with stress response, cell growth, and differentiation. A nonsense mutation in the NDRG1 gene causes hereditary motor and sensory neuropathy, Charcot-Marie-Tooth disease type 4D [38]. SOX9 is instead associated with demyelinating diseases [39]. Furthermore, a search of the OMIM database shows other links between ALS transcripts and specific brain functions. LMO3 and GPM6B are expressed in glial cells, and PREX1 in mouse cerebellum. Furthermore, PREX1 and PREX2 are regulators of Purkinje cell morphology and cerebellar function since Prex1/Prex2 double knockout mice are ataxic and have reduced basic motor activity, abnormal posture and gait, and impaired motor coordination at a young age [40]. It is notable that although the number of splicing events associated with ALS is limited, they are not equally distributed respectively in the 5' end, coding and 3' end of the gene, but 50% of the events are localized in the 5' UTR region, suggesting the presence of a deregulation not only at splicing but also at the transcriptional level. We are actually investigating the characteristics of the putative promoter regions of the 7 genes characterized by 5'UTR splicing events.

2. Next-Generation Sequencing of non-coding RNAs

In this section a work dealing with high-throughput sequencing is presented. First a biological description of RNAs which are not translated into proteins is given. Then Next-Generation sequencing methodology is described in general, then only the specific technology used for experiments under analysis is described in mode details. The last section concerns a software package that I developed for non-coding RNAs sequences analysis.

2.1 Introduction

Most types of RNA molecules do not codify for protein products and are called non-coding RNAs. They constitute a large family of abundant and functionally important RNAs such as transfer RNA and ribosomal RNA, as well as microRNA, small nucleolar RNA, short interfering RNA. Recent bioinformatic studies suggest the existence of thousands of non-coding RNA [41]. The software package presented afterwards is specifically focused on the analysis of microRNAs (miRNAs), a class of post-transcriptional regulators. They consist of 22 short nucleotide RNA sequences that bind to complementary sequences in the 3’ end of multiple target mRNAs, usually silencing them. MicroRNAs target 60% of all genes, are abundantly present in all human cells and are able to repress hundreds of targets each. More than 700 miRNAs have been identified in humans and over 800 more are predicted to exist. Due to their abundance and far-reaching potential, miRNAs can have very different functions in physiology, from cell differentiation to the regulation of fat metabolism. They display different expression profiles from tissue to tissue, reflecting the diversity in cellular phenotypes suggesting a role in tissue differentiation and maintenance.

2.2 Methods

Studying genome sequences has become fundamental for basic research in biology and medicine. In general, sequencing a molecule of DNA/RNA means splitting this molecule into segments to determine the order of its nucleotide bases (A - adenine, C - cytosine, G - guanine, T - thymine). DNA sequencing has accelerated biological research and discovery: the fast speed of sequencing has been fundamental in the sequencing of the human genome (Human Genome Project). In the 70's shotgun sequencing first appeared: it was then possible to split long DNA strands into short (100-1000 bp) partially overlapping segments. Then these segments were sequenced using the chain termination method [42] (Sanger method). Computer programs then used the overlapping ends of different reads to combine them into a continuous sequence. This sequencing method was successful but quite expensive and new low-cost technologies were needed. Although shotgun sequencing was the most advanced technique for sequencing genomes (1995–2005), other technologies started surfacing, called Next-Generation Sequencing.

The term Next-Generation Sequencing indicates high-throughput methods characterized by massive parallel production and subsequent analysis of millions of short-length (25-500 bp) sequences of genome (called “reads”), in a short time (day). Microarrays had the unchallenged primacy in Transcriptomics analysis in the last ten years and they were also widely used in other biological areas. But their main limitation is that they are available only for some organisms. Instead, using Next-Generation Sequencing the whole transcriptome of any organism could be potentially sequenced, allowing the identification of each transcript as well as their expression profile. For RNA and microRNA expression profiling, RNA sequencing (RNA-seq) has significant advantages compared with microarray methods: it detects more efficiently common and rare transcripts.

NGS differs from shotgun Sanger sequencing because does not need in vivo cloning. Other differences between the single techniques are reported in the following table [43] [44] [45] [46] [47].

	Read length	**Throughput**	**Timing**	**Accuracy**
Sanger	1000 bp	0.5 MB	1 day	0.99.999
454	200-500	400-600 MB/run	10 hours	0.99995
Solexa	35-150	10GB	4 days	0.99995
SOLiD	35-75	20 GB/run	5 days	0.99988

Three main Next-Generation Sequencing technologies raised in the last decade:

- **454** pyrosequencing (Roche, 2005) [48].
- **Solexa** reversible terminator sequencing (Illumina, 2006) [49].
- **SOLiD** sequencing by ligation (Applied Biosystems, 2007) [50].

These platforms enable:

- at genomic level: whole genome re-sequencing and de novo sequencing;
- at transcriptomic level: small RNA analysis, gene expression profiling and whole transcriptome analysis;
- at epigenomic level: chromatin immunoprecipitation sequencing (ChIP-Seq) and methylation analysis.

2.2.1 SOLiD by Applied Biosystems

SOLiD - Sequencing by Oligonucleotide Ligation and Detection - is one of the most recent Next-Generation Sequencing technologies, developed in 2007 by Applied Biosystems, that since November 2008 constitute, together with Invitrogen Corporation, the Life Technologies company.

The SOLiD™ sequencing system is based on sequential ligation of dye labeled oligonucleotide probes where each probe queries two base positions at a time. SOLiD™ System enables parallel sequencing of clonally amplified DNA fragments linked to beads. This system uses four fluorescent dyes to encode for the sixteen possible two-base combinations (**Fig.17**).

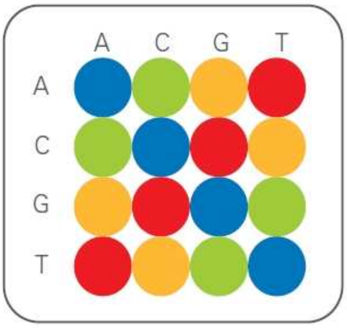


Fig.17 [51] Each reverse, complement and reversed complement couple of bases is represented by the same color (e.g. TA, AT, GC, CG).

It is possible to convert data from color space to the corresponding nucleotides or "base space" knowing the identity of the first base in the read.

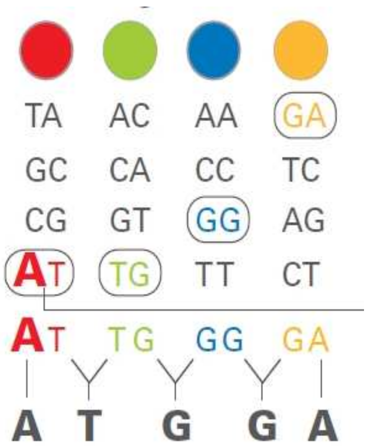
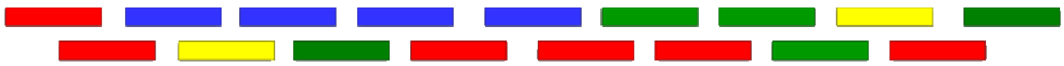


Fig.18 [51] The color space sequence is decoded starting from the first known base (A) to get a final sequence in the base space. Because each couple of adjacent colors must have a base in common, it is easy to translate the sequence of colors in consecutive di-bases, then into a unique sequence in the base space.

Each color can be also associated to a number (0 - blue, 1 - green, 2 - yellow, 3 - red) because it is computationally convenient to keep the original color space sequence and translate it into bases only at the end of the analysis of a data set, after pre-processing.

C 3 3 0 2 0 1 0 3 0 3 1 3 1 1 2 3 1

C G C C T T G G C C G T A C A G C A



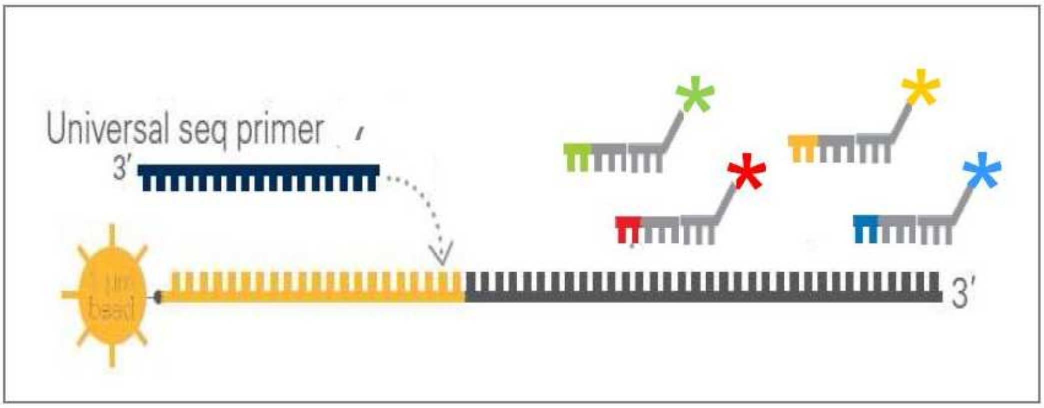
In the SOLiD System the conversion of a sequence of DNA from the color space into nucleotide base space is usually done after having aligned the sequence to a reference genome transcribed in color space (with tools like SHRIMP- SHort Read Mapping Package [53]).

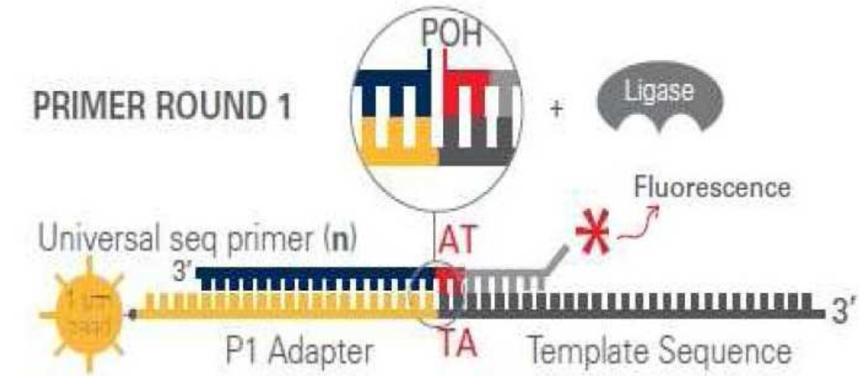
Advantages of this system:

- ability to detect complicated genomic variations.
- complete large-scale sequencing and tag experiments more cost effectively than previously possible.
- double check: since each base is interrogated twice in independent reactions, the information about each base is included in two adjacent pieces of color space data.
- higher accuracy for SNP detection and 99.94% base-calling accuracy.

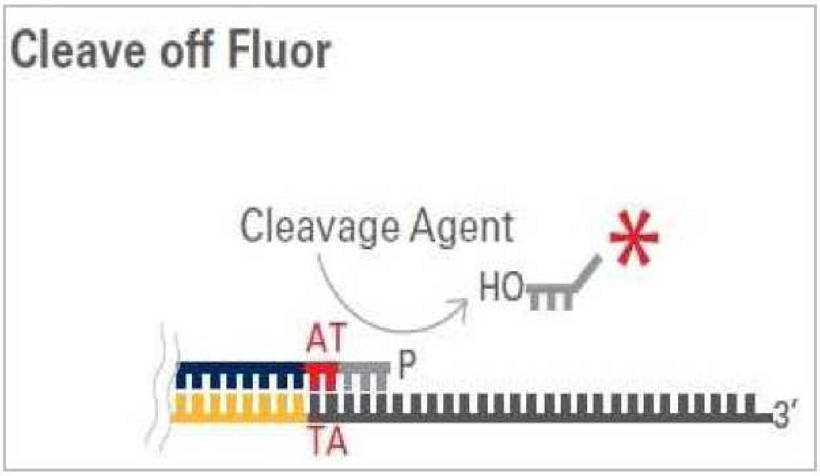
Sequencing by ligation

The DNA strand in analysis is fragmented into segments to which edges are artificially tied two universal primers (adapter P1, P2). Then these segments are put in a pool of beads in microreactors and it is likely that each bead will be ligated only to a unique sequence. Then the sequence on each bead is amplified (several copies of the same sequence are tied to the bead).

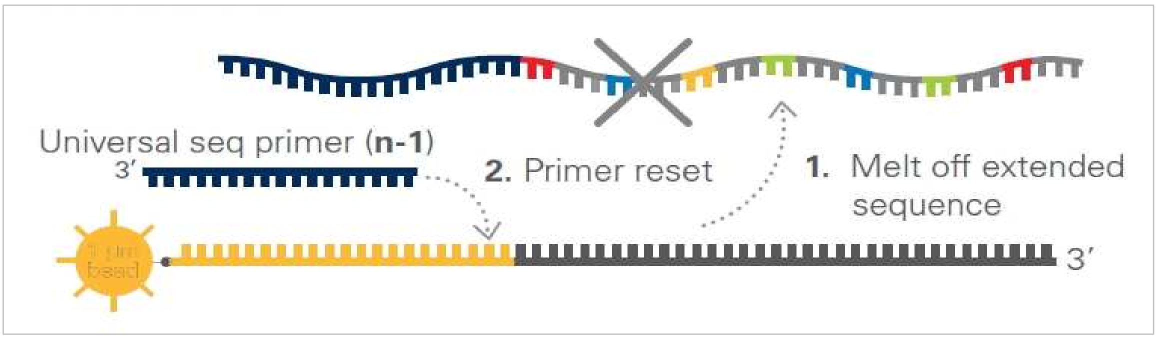




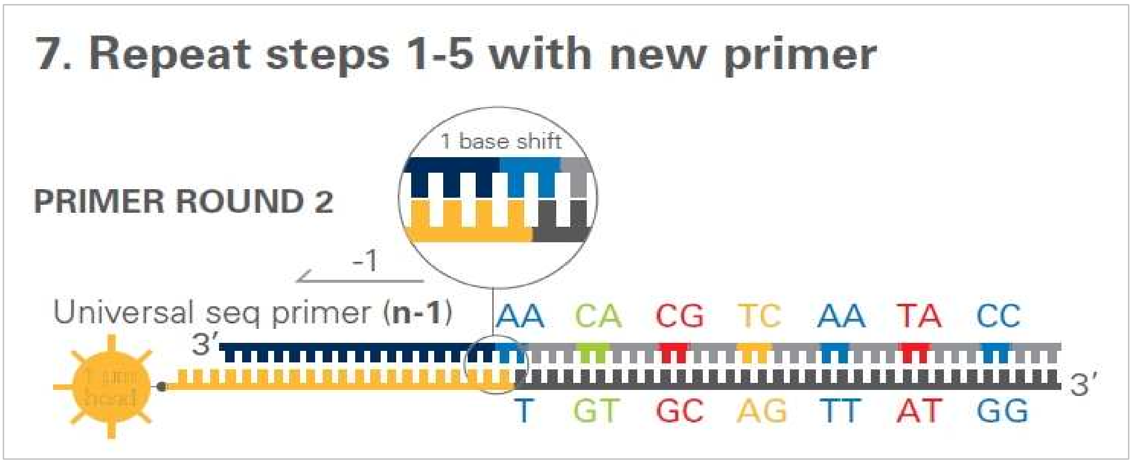
On each sequence of the bead a universal primer n nucleotides long hybridizes with the P1 adapter of the sequence and then one of the four possible fluorescent dye-labeled di-bases ligates to the sequence [51].



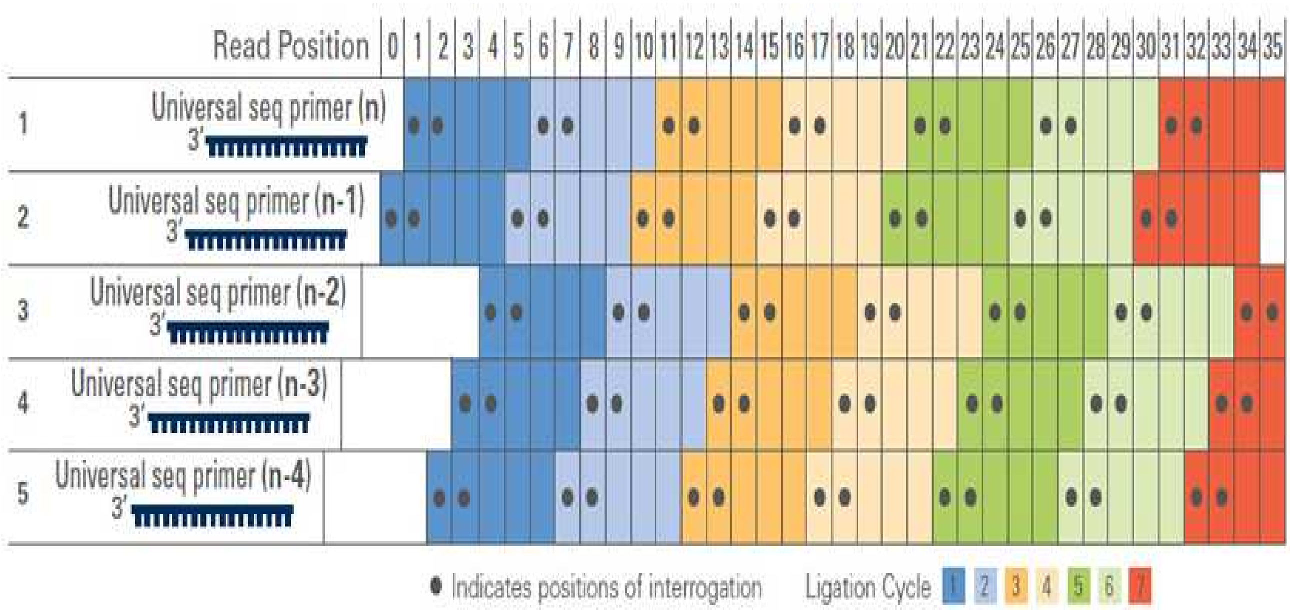
Several cycles of ligation, detection and cleavage are needed to get the complete sequence in the color space.



Then the universal primer on the original sequence is reset to be long (n-1) nucleotides to repeat the ligation step with the fluorescent di-bases and obtain a new sequence in color space [51].



The number of ligation cycles depends on the length of the sequence to be analyzed, e.g. if the sequence is 35 nucleotides long, there will be needed 7 ligation cycles repeated 5 times [51].



Once this table is ready, knowing the first base from the universal primer, the colors can be aligned to a reference genome translated in the color space.

2.3 Results

I worked on this project during the second year of Ph.D., in Italy. Almost all the work was done in R (some parts were written in C, then interfaced with R). The project concerns some algorithms, part of a package called ***ncSOLID,*** for the quantitative secondary analysis of non-coding transcriptome sequencing data generated with SOLiD System platform. The package was initially designed as stand-alone package but subsequently was integrated in ***oneChannelGUI***.

2.3.1 ncSOLID

ncSOLID was built with the aim of organizing RNA-seq data into a data structure that allows the statistical detection of differential expression for non-coding RNAs (ncRNAs), e.g. microRNAs. The library had a R interface to SOCS software [52] which was used to map and quantify sequence data.

SOCS (Short Oligonucleotides in Color Space) is a program to map color space sequence data to a reference genome. It allows mapping of color space data in a more flexible, mismatch-tolerant (0-5 possible mismatches) context to maximize the number of usable sequences within a given data set. The higher is the tolerance (high number of allowed mismatches), the lower the amount of rejected sequences, but the longer becomes the computational runtime.

SOCS maps at lower tolerances first, reducing the data to be mapped at higher tolerances.

During the mapping process, if a read maps to two or more [52]:

- non-identical genomic substrings within the maximum tolerance, quality scores and mismatch counts are used to get the optimal match (unambiguous matching).
- identical genomic substrings, all matching locations are considered ambiguous (ambiguous matching).

Once the number of optimal matches is determined, coverage maps of each reference chromosome are computed.

ncSOLID library allows to run multiple instances of SOCS. Coverage data produced by SOCS are segmented to select peaks of ncRNA expression, which are then organized in an ExpressionSet object [52], a data structure usually employed to collect data from microarrays. It binds together expression measurements with covariate and administrative data, convenient for results manipulation.

This structure of the expression level of transcripts is then quantified as coverage per million reads to the transcriptome. Those data are used to detect differentially expressed ncRNAs using several types of statistics (for example Rank product, presented in the first chapter, **section 1.2.2**). In the following, the workflow to analyze ncRNAs with ***ncSOLID*** is shown.

SOLiD output files are in .csfasta (color space fasta) for the sequences and .qual format for the scores associated to each sequence in the .fasta file.

The sequences of the .csfasta file (and respective scores file. qual) are first trimmed (trimSOLID.R), to remove the adapter P2 at the end of each sequence.

Trimming the sequences has three reasons:

- since non-coding RNAs are quite short (18-28 bp) and reads length on SOLiD system are longer (35-75 bp), it is possible that reads are at least partially contaminated by P2 adaptor.
- while the current matching tools provided on the SOLiD are designed for reads of equal length, ***ncSOLID*** library allows trimming: adaptor removal, to get the desired read length.
- trimming can substitute mismatching at the end of the sequence (3’).

This trimming algorithm was evaluated on a data set constituted by four .csfasta files (sequences long 35 bp) and four .qual files, two treatment and two control cases, that were mapped on the chromosome 22.

Given a sequence of given length (e. g. 35 bp), if one hypothesizes a minimum length (20 bp) not to be overcome and a step of trimming (5 bp), trimSOLID.R cut sequences from the end to get sequences of the desired length (35, 30, 25, 20 bp).

<u>Example:</u> Three sequences of one of the four files in input are trimmed of five elements, starting from the end of the sequence.

35 bp:

T33230310310120332321003330221330113
T33232130123212022023133333322313132
T03211231113210120233303333322313133

30 bp:

T332303103101203323210033302213
T332321301232120220231333333223
T032112311132101202333033333223

25 bp:

T3323031031012033232100333
T3323213012321202202313333
T0321123111321012023330333

20 bp:

T33230310310120332321
T33232130123212022023
T03211231113210120233

The reads in the .csfasta files always begin with **T** (that is not included in the length of the read, 35 in the example above) and are constituted of numbers from 0 to 3 that substitute the four colors that identify each couple of nucleotides.

They will be translated in the base space later on; during this pre-processing step it is better to keep this format to avoid errors of translation in the final sequence. The sequences in the .qual files contains sequences of quality scores, 35 scores for each read, if this is 35 bp long, for example. Here after parts of two of the files in input, in .csfasta and .qual format respectively, are shown: each sequence is preceded by the respective identification number.

<u>File.csfasta</u>

- ***920_6_719_F3***
  T33230310310120332321003330221330113

- ***920_7_366_F3***
  T33232130123212022023133333322313132

- ***920_8_370_F3***
  T03211231113210120233303333322313133

- ***920_8_721_F3***
  T01032201322220030211103310121333113

- ***920_11_374_F3***
  T30100313223202232201333333323313303

- ***920_11_1252_F3***
  T33211130130120211100330212330313113

- ***920_11_1559_F3***
  T13130113320230022223233333333213323

<u>File.qual</u>

- ***920_6_719_F3***
  8 12 18 6 16 19 4 4 18 23 5 21 7 26 24 8 26 5 16 25 13 24 7 22 18 24 21 27 24 6 6 13 4 26 24

- ***920_7_366_F3***
  14 23 20 27 7 9 27 17 26 25 4 14 23 4 22 6 6 7 6 18 4 14 7 13 21 4 4 5 6 4 4 7 4 4 10

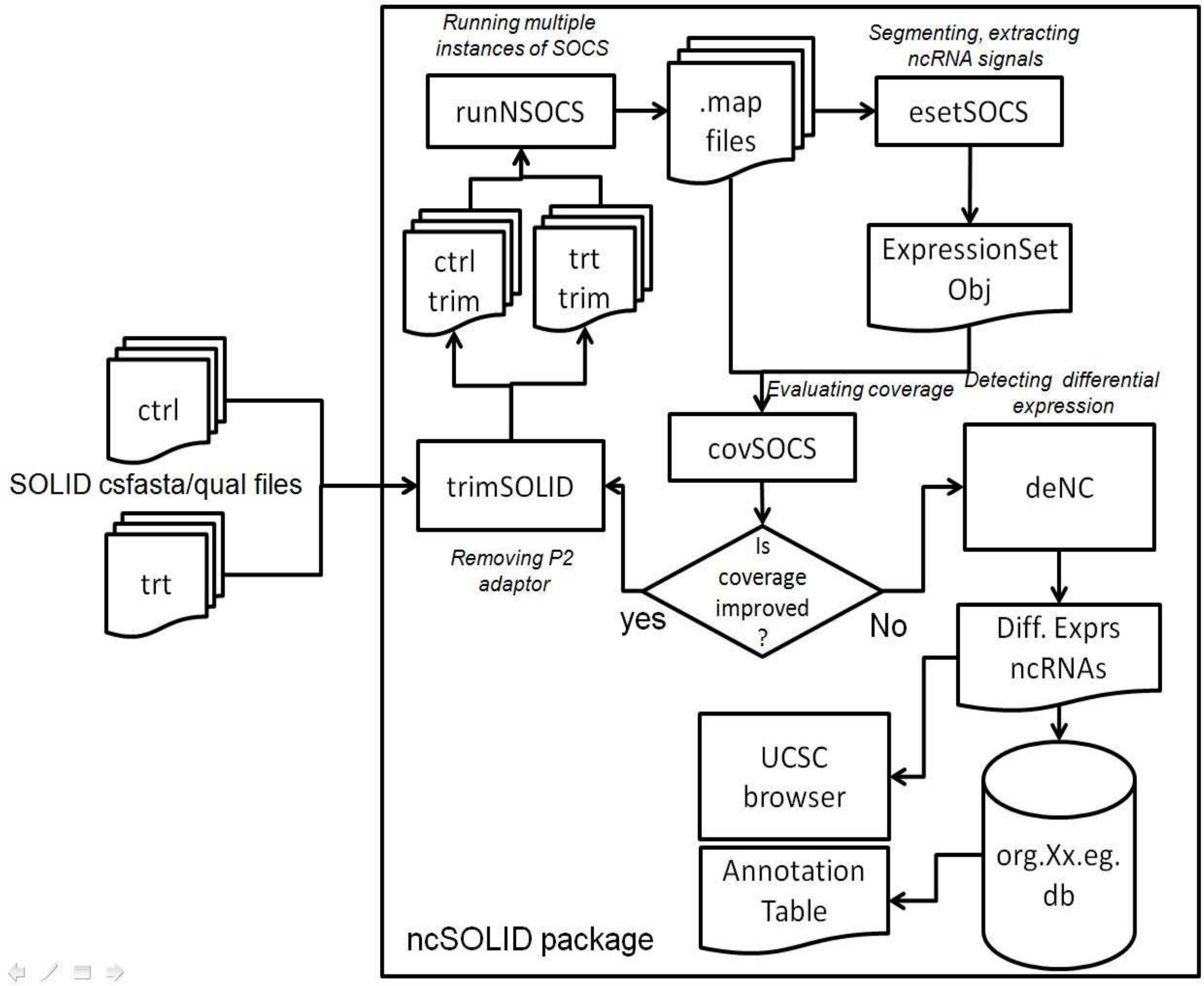
- ***920_8_370_F3***
  12 20 16 9 12 20 10 21 18 17 7 11 21 4 12 9 13 6 6 6 4 10 15 17 14 4 4 4 12 4 4 4 4 4 4

- ***920_8_721_F3***
  21 15 4 12 6 5 20 5 4 4 4 6 4 27 7 10 6 20 6 19 9 10 20 24 6 16 4 15 21 6 4 10 4 4 10

- ***920_11_374_F3***
  27 25 21 27 27 25 19 9 18 27 27 20 10 4 12 5 6 12 18 4 4 6 12 5 4 4 4 4 4 4 4 4 4 4 20

- ***920_11_1252_F3***
  27 27 22 25 26 27 27 6 26 21 26 26 7 12 12 25 15 11 15 20 14 21 6 6 4 5 4 7 22 10 10 6 10 6 13

- ***920_11_1559_F3***
  4 7 4 10 4 6 5 7 6 8 4 6 4 15 4 4 4 12 4 4 4 4 7 4 4 4 4 4 6 4 4 4 4 4 6



Then SOCS must be run N times (runNSOCS.R), because N are the samples in analysis, choosing a mismatch tolerance, to get for each sample a ".map" file, that contains the mapping of the reads in the sample to the reference chromosome of the genome. The sequences of the genome in the .map files are segmented, i.e. the regions where the reads are concentrated are divided from the regions of the genome not covered by reads, to get the expressions of ncRNAs (esetSOCS.R), organized in an ExpressionSet structure. Then the coverage of the chromosome in analysis with the resulting ExpressionSet must be evaluated (covSOCS.R) to verify whether it must be improved (and repeat the previous steps of the process increasing the trimming factor) or if it can be further analyzed to detect differential expression between the control and treatment case.

The library ***ncSOLID*** was created to become part of the libraries of Bioconductor but it was not submitted because in the meanwhile other tools for the secondary analysis of RNA-seq came out and performed better than SOCS. These tools are now part of ***oneChannelGUI***, as described below.

2.3.2 Extension of the R library oneChannelGUI

oneChannelGUI was first built up by Prof. Raffaele Calogero and co-workers, in 2007 [4] as graphical interface to analyze data sets from microarray experiments and recently (2010) it has been extended to analyze the results obtained with Next-Generation sequencing technologies, focusing in particular on microRNA analysis. The aim of this extension of the library is the secondary analysis of non-coding RNAs, short reads aligned against the relative reference genome.

The raw data resulting from a Next-Generation Sequencing experiment cannot be directly statistically analyzed. The data in input to ***oneChannelGUI*** can come from the following different mapping tools, freely available online: SHRIMP (SHort Read Mapping Package) [53], miRanalyser [54], MicroRazers [55], miRExpress[56], miRProf (web tool).

When the data in input are loaded in the program, if they do not come from one of the above-mentioned tools for the primary analysis, they are reorganized as ExpressionSet, as done in ***ncSOLID*** (**section 2.3.1**). This structural reorganization can be done with ***Genominator*** or a segmentation approach based on ***chipseq***, both R libraries. It is then possible to normalize the data and this is done by the program using a method that will be described hereafter in the third chapter [57] and that is part of a R library called ***edgeR*** [58], for the detection of differential expression of short reads. For a quality control, principal component analysis and hierarchical clustering are available, as already for the data from microarray experiments. And it is also available a multidimensional scaling plot, provided again by ***edgeR*** package. As statistics, Rank Product (**section 1.2.2**) and ***edgeR*** (**section 3.2.2**) are implemented. Rank Product was evaluated on a semi-synthetic data set in chapter 1 and it was demonstrated that it is a powerful technique to detect differential expression among data from microarray experiments, but it can be used also in short reads analysis, as well as ***edgeR***. In the next chapter there will be given an example of the use of these two methods, combined to find regulation in some shRNAs samples.

2.4 <u>Conclusions</u>

While defining the project, ***ncSOLID*** was meant to be a new R package in Bioconductor. After the development of the tool it seemed more convenient to include this software in the already existing ***oneChannelGUI*** package, extending its functions to the secondary analysis of non-coding RNAs.

3. Short hairpin RNAs modeling

This last chapter deals with the analysis of shRNAs experiments, which were made with the purpose of finding some regulated shRNAs that could be used as biomarkers during liver regeneration.

Short hairpin RNAs belong to the class of non-coding RNAs and are concerned in **section 3.1**. Before discussing the analysis of regulation of shRNAs data sets and the related results obtained with different statistical softwares, a technical description of these tools is given.

3.1 Introduction

Non-protein-coding RNA molecules with hairpin shape silencing gene expression are called short hairpin RNAs (shRNAs). shRNAs can be synthesized in vitro or transcribed in vivo to suppress the expression of target genes in cultured cells [59]. They are part of the RNA interference system (RNAi), an evolutionally conserved gene silencing mechanism present in a variety of eukaryotic species. It has been widely used as a novel effective tool for functional genomics studies, displaying a great potential in treating human diseases, e.g. cancer treatment, and there has been a recent development in the use of RNAi in the prevention and treatment of viral infections [60] [61]. A deep analysis of data sets of shRNAs with the aim to detect biomarkers for liver regeneration was made and it will be presented later on in this thesis. A previous study (2008) was made by Lars Zender et al. [62] to identify tumor suppressor genes relevant to human cancer (liver), establishing the feasibility of in vivo RNAi screens. A project focused on the deep analysis of data (counts) from shRNAs experiments recently began with a paper (2010) by Frank Klawonn, Torsten Wüstefeld and Lars Zender [63], where the future perspective is to determine the cause of variations between experiments in different conditions, which is the topic presented in the next sections: defining the optimal statistics to select significantly regulated shRNAs.

3.2 Methods

"Next-Generation" is the term with which high-throughput sequencing technologies raised in the last five years, are indicated. As reported at the beginning of **section 2.2**, these short sequences can be produced in different ways by the three different equipments: 454 by Roche, Solexa by Illumina, SOLiD by Applied Biosystems. In the next section the Solexa technology is described to understand how shRNAs were sequenced and subsequently analyzed from a statistical point of view in **section 3.3**.

3.2.1 Solexa technology by Illumina

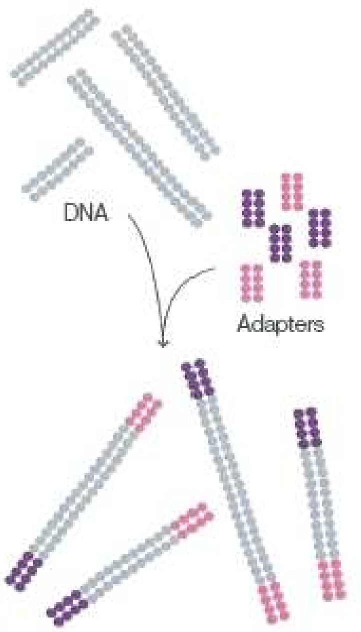
Solexa sequencing uses four fluorescently labeled nucleotides instead of associating a di-base to each color as done in SOLiD. The sample preparation methods used differ slightly from that used in the SOLiD system, but the basic goal is the same: generating large numbers of unique "polonies" (polymerase generated colonies) that can be simultaneously sequenced. These parallel reactions occur on the surface of a "flow cell" (a microscope slide) which provides a large surface area for many thousands of parallel chemical reactions.

As in the SOLiD Next-Generation technology (**section 2.2.1**), Solexa uses sequences that in average are 35 bp long, but instead of sequencing by ligation they are chemically synthesized.

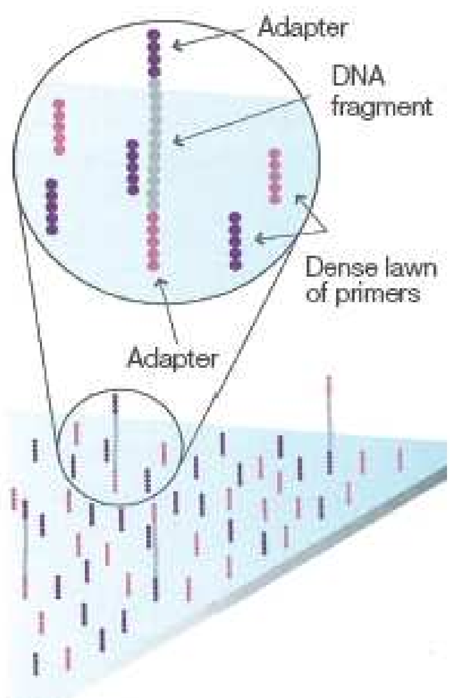
Sequencing by synthesis

Solexa's strategy is the amplification of DNA on an array followed by synthesis by incorporation of modified nucleotides linked to colored dyes [64]. The first step to prepare the DNA library to be sequenced is to randomly fragment DNA and ligate two adapters to the ends of the fragments.

1. Preparation of DNA sample

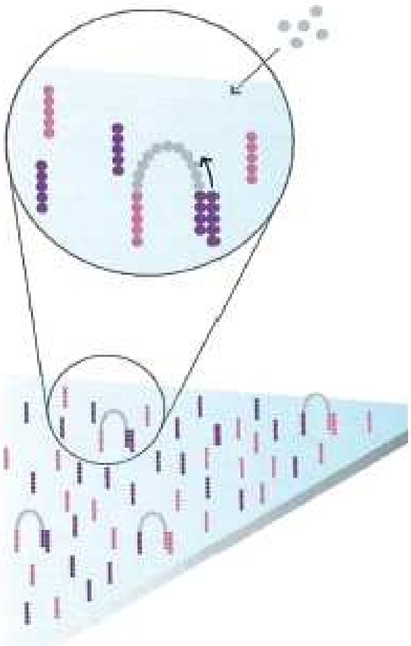


2. Attach DNA to solid surface

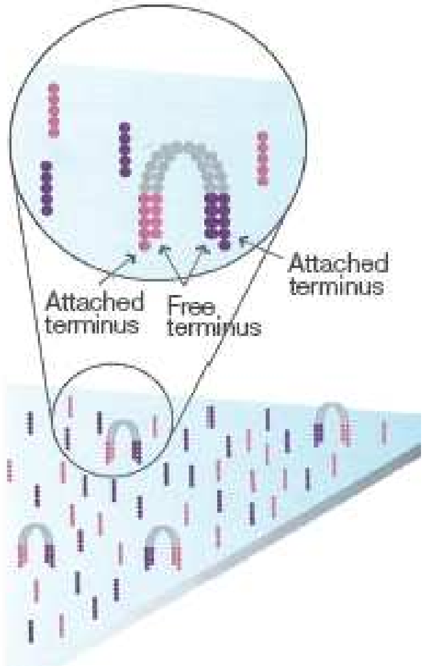


Single-stranded fragments must be then attached randomly to the solid surface of the flow cell channels, which are already partially covered by primers that will be used in the following phase of the process [65].

3. Bridge amplification

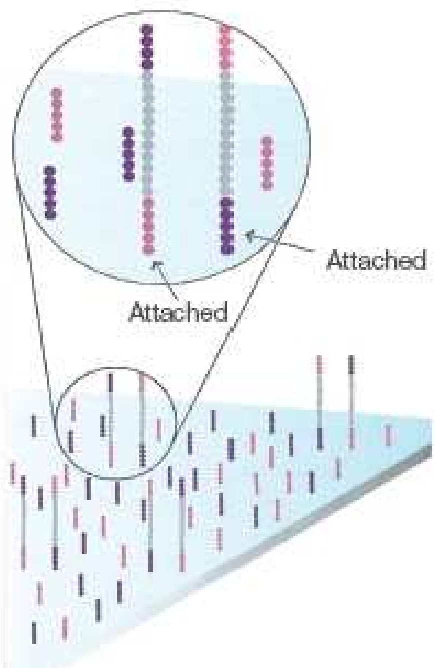


4. Fragments become double stranded

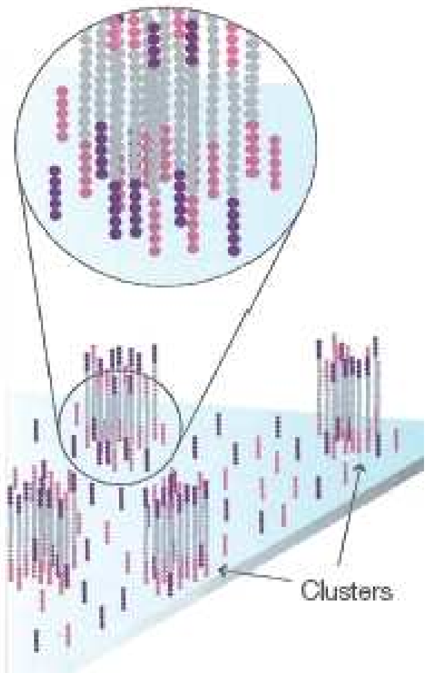


Add unlabeled nucleotides and enzyme to start solid-phase bridge amplification: the enzyme incorporates nucleotides to build double-stranded bridges on the substrate.

5. Denaturation of double-stranded molecules

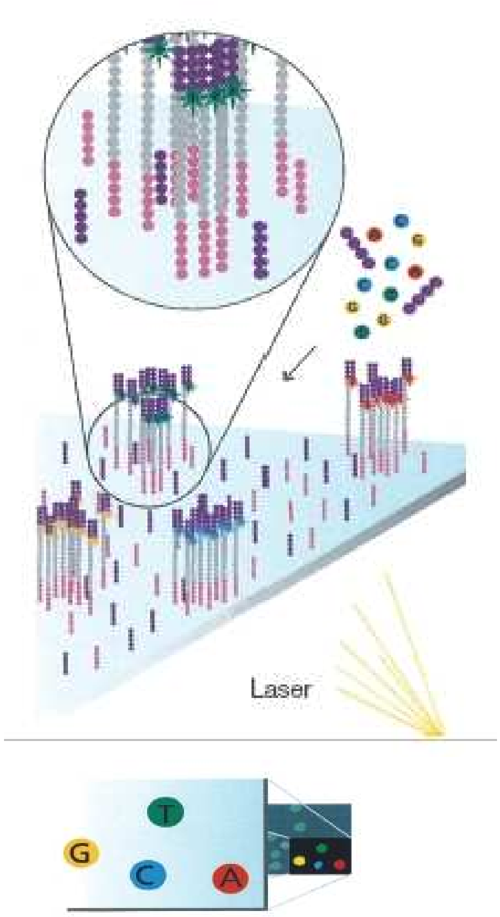


6. Amplification

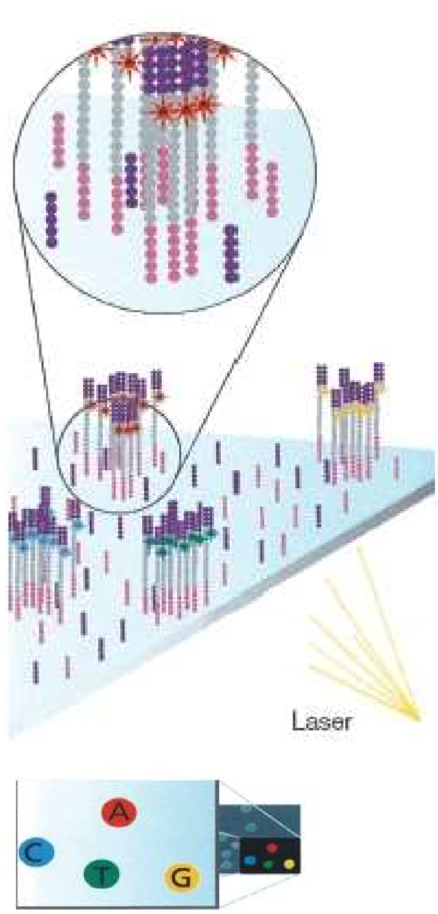


Denaturation implies that single-stranded fragments are now attached to the surface; several millions of dense clusters of double-stranded DNA are generated.

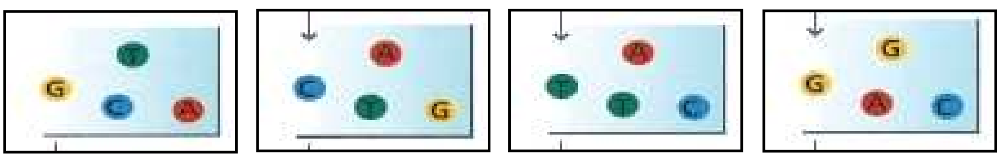
7. Determination of the first base



8. Determination of the second base



The first sequencing cycle begins by adding four labeled reversible terminators, primers and DNA polymerase. After laser excitation, the emitted fluorescence from each cluster is captured and the first base identified. The next cycle of sequencing repeats the incorporation of four labeled reversible terminators, primers and DNA polymerase. Then the image is captured with the laser as before and the second base is recorded [65].The sequencing cycles are repeated to determine the sequence of bases in a fragment, one base at a time. The results are aligned and compared to a reference and sequencing differences are found (TAAG is the upper sequence in the picture).



= TAAG

3.2.2 Detection methods

While the first project described in this thesis with alternative splicing events detection in microarray data and the second one with the secondary analysis of non-coding RNAs, this last project gives the impression of representing a combination of the previous two parts of this manuscript, dealing with the detection of differential expression in RNAs sequences. Hereafter are presented the methods used in the analysis of some shRNAs data. Both ***edgeR*** [57] and ***DESeq*** [66] are based on the hypothesis that the reads can be approximated as Negative Binomial distributions [67] [68] [69] [70] [71] [58] [66], while Rank Product is a non-parametric method.

Both ***edgeR*** and ***DESeq*** authors state that the read counts follow a Multinomial distribution [71] [72] which can be approximated by a Poisson model [73] [74] [75]. This is because the Multinomial is the multivariate case of a Binomial distribution which converges to a Poisson when the sample size is large and the probability of success/failure is close to zero.

In a Poisson distribution the unique parameter for the model is the mean μ that is equal to the variance *v*. But the authors of ***edgeR*** [58] and others [69] [70] [68] agree that the assumption of Poisson distribution for the read counts is too tight because the variance of the data is greater than the mean actually. This is called "overdispersion problem" [58] [66] and can be solved if the count data are modeled as Negative Binomial distributions with $v = \mu + \alpha\mu^2$, where the proportionality constant α is estimated from the data and $\mu << \sigma^2$

edgeR

This is the first and very recent (2010) library created in the R environment for the analysis of differential expression of count data produced by Next-Generation sequencing technologies; in particular ***edgeR*** [58] has the aim of deeply analyzing RNA-seq reads. subsequent other libraries are based on ***edgeR*** (empirical analysis of DGE in R): ***DESeq*** [66], described afterwards; ***baySeq*** [76] and ***DEGSeq*** [72], whose performance was first evaluated at the same time of ***edgeR*** and ***DESeq*** but subsequently abandoned because considered less reliable than the other two statistics to focus mostly the attention on the three statistics presented in this section. ***edgeR*** includes a new method of normalization of short reads so that they become comparable between different samples of the same or different conditions.

Some widespread previous scaling techniques are:

- adjusting counts to reads per kilobase per million mapped (RPKM technique), i.e. normalizing for RNA length and for the total read number in the measurement [77];
- data standardization by dividing each list of counts by the total number of reads in the list [78];
- performing an hyper-geometric test computing p-values to account for sample biases [71];
- quantile normalization [79] [80], as already used to normalize data resulting from microarray experiments [32].

The authors of ***edgeR*** in the related article on the description of this method find the second method, even if intuitive, too simple for many biological applications [57].

They assess that the number of reads that should map to a gene depends not only on the expression level and length of the gene but also the composition of the RNA population that is being sampled [57]. Hence if a large number of genes are highly expressed in one condition, fewer tags are available for the remaining genes in that sample. This artifact, if not adjusted, distorts the results of the differential expression analysis and results in higher false positive rates . This is what they tried to account for proposing a new normalization method, able to compare Next-Generation Sequencing data across samples, estimating a suitable scaling factor from the raw data. They started from two basic hypothesis:

1. A gene with the same expression level in two samples should not be detected as differentially expressed.

2. The amount of reads mapping a certain gene depends on the expression features of the whole sample rather than only on the gene expression level.

Let Y_{gk} be the observed count for gene g in the sample k, summarized from the raw reads ; let μ_{gk} be the true unknown expression level, L_g the length of gene g; N_k the total number of reads for library k. The expected values of Y_{gk} can be estimated as:

$$\hat{Y}_{gk} = \frac{\mu_{gk} L_g}{S_k} N_k \qquad \text{with} \qquad S_k = \sum_{g=1}^{G} \mu_{gk} L_g \tag{23}$$

S_k is the total RNA output of a sample k.

While N_k is known, S_k is unknown and can extremely vary from sample to sample [57] and cannot be directly estimated. The relative RNA production of two samples k,

k' is $f_k = \frac{S_k}{S_{k'}}$ and can be estimated using a weighted trimmed mean (average after removing the upper and lower x% of the data) of the logarithm of the expression ratios. The gene log fold changes are defined as follows [57]:

$$M_g = log_2\left(\frac{Y_{gk} / N_k}{Y_{gk'} / N_{k'}}\right) \tag{24}$$

and the absolute expression levels as:

$$A_g = \frac{1}{2} log_2\left(\left(Y_{gk} / N_k\right) \bullet \left(Y_{gk'} / N_{k'}\right)\right) \tag{25}$$

the proposed normalization is the trimmed mean of M_g and A_g values and is called by the authors of ***edgeR*** “TMM normalization” (Trimmed Mean of M values): by default the trim for M_g is 30% and for A_g is 5%.

After trimming, a weighted mean of M_g and A_g is taken, with weights computed as the inverse of the approximate asymptotic variances, computed using the delta method presented in [81]. The normalization factors for sample k with reference sample r, are

$$log_2(TMM_k^r) = \frac{\sum_g w_{gk}^r M_{gk}^r}{\sum_g w_{gk}^r} \tag{26}$$

with fold change $M_{gk}^r = log_2\left(\frac{Y_{gk} / N_k}{Y_{gr} / N_r}\right)$ and weights $w_{gk}^r = \frac{N_k - Y_{gk}}{N_k Y_{gk}} + \frac{N_r - Y_{gr}}{N_r Y_{gr}}$

with Y_{gr} and $Y_{gk} > 0$.

edgeR works only with data sets with replicated experiments.

If the replicates are only two the “effective” library size (the total sum of counts of a sample) is determined dividing the sample taken as reference by $\sqrt{f_k}$ and multiplying the non reference one by $\sqrt{f_k}$

Then the ***sage.test*** algorithm from the R library ***statmod*** can be used to compute the p-value for Fisher´s exact test for each gene.

Normalization factors across several samples can be computed selecting one sample as reference and computing TMM normalization factor for each non reference sample [57] to determine the “effective” library size (total number of counts in each sample) of the samples in each condition. Then the statistical analysis is performed following the method proposed in [71] using a likelihood ratio test to evaluate the differences in expression between libraries: they compute the maximum likelihood estimates under the under the null hypothesis that each gene must have the same mean expression in different samples. The standard likelihood ratio statistic,

$$D = -2\,ln\left(\frac{L_{H_0}}{L_{H_1}}\right) \tag{27}$$

where L_{H0} is the likelihood for the null model and L_{H1} the likelihood for the alternative model. *D* was computed and *p*-values for each gene were obtained exploiting the fact that, under the null hypothesis, this statistic has an approximated χ^2 distribution with one degree of freedom.

DESeq

This method is based on the previously described R package ***edgeR*** and similarly based on the hypothesis that each single read count can be described by a Negative Binomial distribution,

$$K_{ij} \approx NB(\mu_{ij}, \sigma_{ij}^2) \qquad \text{with} \qquad \mu << \sigma^2 \tag{28}$$

which unique and unknown parameters are the mean and the variance of the reads that must be estimated from the data. This distribution can be also parametrized, as suggested in [67] with respect to the probability (*p*) and the number of failures before a success (*r*) as

$$p = \frac{\mu}{\sigma^2} \qquad r = \frac{\mu^2}{\sigma^2 - \mu} \tag{29}$$

Normally the number of replicates is low and so there is an evident need for further modeling assumptions [66] and the authors of ***DESeq*** hypothesize that:

1. The mean μ_{ij} of gene *i* in sample *j*, is proportional to the library size s_j

$$\mu_{ij} = q_{i,\rho(j)} * s_j \tag{30}$$

with $q_{i,\rho(j)}$ gene abundance, $\rho(j)$ experimental condition of sample *j*

2. The variance σ^2 is constituted by two terms:

$$\sigma_{ij}^2 = \mu_{ij} + \left(s_j^2 * v_{i,\rho(j)}\right) \tag{31}$$

μ_{ij} is the mean and the other term is the raw variance.

3. The raw variance is proportional to the gene abundance

$$v_{i,\rho(j)} = v_\rho\left(q_{i,\rho(j)}\right) \tag{32}$$

Then the model must be fitted to the data, estimating: s_j $q_{i,\rho(j)}$ $v_{i,\rho(j)}$ follows:

$$\hat{s}_j = \underset{i}{median} \frac{k_{ij}}{\left(\prod_{t=1}^{m} k_{it}\right)^{1/m}} \tag{33}$$

$$\hat{q}_{i\rho} = \frac{1}{m_\rho} \sum_{j:\rho(j)=\rho} \frac{k_{ij}}{\hat{s}_j} \tag{34}$$

with m_ρ number of replicates of condition ρ

$$\hat{v}_\rho\left(\hat{q}_{i\rho}\right) = w_\rho\left(\hat{q}_{i\rho}\right) - z_{i\rho} \tag{35}$$

with $$w_{i\rho} = \frac{1}{m_\rho - 1} \sum_{j:\rho(j)=\rho} \left(\frac{k_{ij}}{\hat{s}_j} - \hat{q}_{i\rho}\right)^2$$

$$z_{i\rho} = \frac{\hat{q}_{i\rho}}{m_\rho} \sum_{j:\rho(j)=\rho} \frac{1}{\hat{s}_j}$$

where the denominator of $\hat{s}_j$ can be imagined as a “pseudo-reference” sample obtained computing the geometric mean across samples [66]. Once these parameters are estimated, under the null hypothesis that $q_{iA} = q_{iB}$ i.e. the gene abundance is equal in the two conditions A and B, it is possible to test the data for differential expression, defining as test statistic the total counts in the two conditions

$$K_{iA} = \sum_{j:\rho(j)=A} K_{ij} \qquad K_{iB} = \sum_{j:\rho(j)=B} K_{ij} \tag{36}$$

with $K_{iS} = K_{iA} + K_{iB}$.

The p-value of a couple of observed count sums is the sum of all the probabilities less or equal to p(k_{iA}, k_{iB}), given the total sum k_{iS} , with a, b with values [0, k_{iS}a]

$$p(k_{iA}, k_{iB}) = \frac{\sum_{\substack{a+b=k_{iS} \\ p(a,b) \le p(k_{iA},k_{iB})}} p(a,b)}{\sum_{a+b=k_{iS}} p(a,b)} \tag{37}$$

This approach is similar to that adopted in the article [69] that is of the same authors of ***edgeR***. $p(a,b)$, assuming that the counts are independent, is the product of the single marginal probabilities of a and b. To compute these two probabilities first is possible to compute the average of the counts rescaled with s_j

$$\hat{q}_{i0} = \sum_{j:\rho(j)=\{A,B\}} \left(\frac{k_{ij}}{s_j} \right) \tag{38}$$

and the mean and variance of the data in condition A are respectively (from equations(30), (31))

$$\hat{\mu}_{iA} = \sum_{j \in A} \left(\hat{q}_{i0} * \hat{s}_j \right) \tag{39}$$

Rank Product

This method [14] has been already mentioned in the previous two chapters (described in **section 1.2.2**) : it was used in the pipe-line to detect alternative splicing events in data sets produced with microarray platforms and it can complete the extension of the R library ***oneChannelGUI*** to analysis of non-coding RNA-seq. Rank Product can be used with a great variety of different types of data in input because it is a non-parametric statistic based on fold change and is being used now in RNA interference [82], as well as in metanalysis [83]. Rank Product was used with 100 permutations and 50 iterations were made to be sure not to loose any regulated shRNAs.

3.3 Results

Here is presented an analysis of seven shRNAs data sets, all produced using Illumina Genome Analyzer technology and kindly provided by Marina Pesic and Ramona Rudalska, Ph.D. students of the group of Prof. Lars Zender and Dr. Torsten Wüstefeld.

3.3.1 Short hairpin RNAs experiments

The data sets produced by R. Rudalska and M. Pesic come from in vitro experiments on a set of short hairpin RNAs that targets genes of chronic liver disease.

Hereafter the number of reads in each samples of each of the seven experiments and the number of treatment (T) and control (C) cases is shown:

- Data sets with replicated experiments

  R1. 1980 shRNAs: 4 C, 4 T (from ***mouse*** liver cell lines)

  R2. 1830 shRNAs: 3 C, 3 T (from ***mouse*** liver cell lines)

  R3. 1911 shRNAs: 3 C, 4 T (from ***mouse*** liver cell lines)

- Data sets without replicated experiments

  R4. 20440 shRNAs: 1 C, 1 T (from ***human*** liver cell lines)

  M1. 234 shRNAs: 1 C, 1 T (from ***mouse*** liver)

  M2. 230 shRNAs: 1 C, 1 T (from ***mouse*** liver)

  M3. 236 shRNAs: 1 C, 1 T (from ***mouse*** liver)

Data sets with no replicates were analyzed in a simpler way than the ones with replicates.

3.3.2 Filters and data normalization

The analysis is illustrated based on the R1 data set, constituted by four samples for each condition (treatment and control cases).

Filters

The first approach was to not consider the rows of the shRNAs data sets containing too many zeros, that are troublesome from a computational point of view. The following example shows which type of row was filtered out in first analysis, resulting in 1854 shRNAs after the filter, from a starting total number of 1980.

T1	T2	T3	T4	C1	C2	C3	C4	
8	0	0	0	0	0	0	0	deleted
0	0	0	0	8	0	0	0	deleted
8	8	0	0	0	0	0	0	taken
0	0	0	0	8	8	0	0	taken
8	0	0	0	8	0	0	0	taken

Only those rows presenting, in at least one of the two conditions, at least one element different from zero, were retained. An alternative more stringent filter was,

T1	T2	T3	T4	C1	C2	C3	C4	
8	0	0	0	0	0	0	0	deleted
0	0	0	0	18	0	0	0	deleted

8	9	0	0	0	0	0	0	deleted
0	0	0	0	18	19	0	0	deleted
8	0	0	0	18	0	0	0	deleted
8	9	0	0	18	19	0	0	taken
8	9	10	0	18	19	0	0	taken
8	9	0	0	18	19	20	0	taken
8	9	10	0	18	19	20	0	taken
8	9	10	11	18	19	20	0	taken
8	9	10	0	18	19	20	21	taken
8	9	10	11	18	19	20	21	taken

This filter considers only those rows in which, in both conditions, half or more elements were different from zero. This resulted in 1709 remaining shRNAs. But from a biological point of view there was a loss of too much information. And the last and most successful approach was to keep only those shRNAs that had half or more element different from zero only in the treatment, because if the control is initially 0 or very low and the treatment is very high this is interesting from a biological point of view and must be further analyzed.

T1	**T2**	**T3**	**T4**	**C1**	**C2**	**C3**	**C4**	
8	0	0	0	0	0	0	0	deleted
0	0	0	0	18	0	0	0	deleted
8	9	0	0	0	0	0	0	taken
0	0	0	0	18	19	0	0	deleted
8	0	0	0	18	0	0	0	deleted

8	9	0	0	18	19	0	0	taken
8	9	10	0	18	19	0	0	taken
8	9	0	0	18	19	20	0	taken
8	9	10	0	18	19	20	0	taken
8	9	10	11	18	19	20	0	taken
8	9	10	0	18	19	20	21	taken
8	9	10	11	18	19	20	21	taken

After this filter 1802 shRNAs remained. If a data set is without replicated experiments of the same condition zeros do not mean anything from a statistical point of view and all the rows with zeros can be ignored. Another filter was tried, on the p-values of the data set after the statistical analysis: deletion of all the regulated shRNAs in which range(T) intersected range(C), with range = (min, max), like in the following example:

	T1	**T2**	**T3**	**T4**	**C1**	**C2**	**C3**	**C4**	
shRNA.1	100	300	400	110	200	180	20	10	deleted
shRNA.2	100	15	50	8	5	10	20	2	deleted

In the first example 200 and 180 are in the range(T) = (100, 400), while in the second one 15 and 8 are in the range(C) = (2, 20) so they seem too noisy to be considered reliable. This filter subsequent to the statistical analyses was removed because biologically it was too stringent.

Dealing with zeros

Instead in the experiments with replicates, even after the (third) preliminary filter applied to reduce the size of the data set throwing out unuseful noisy data, there will be still some read counts equal to zero in some samples. Hence this raised the need to better deal with these zeros, which can cause computational problems.

The first approach was to create four different types of data sets to be subsequently subdued to differential expression analysis:

– Raw: the raw original data set

	T1	T2	T3	T4	C1	C2	C3	C4
shRNA.1	8	10	0	0	18	20	0	0
shRNA.2	8	10	0	0	0	0	0	0

– LC: filtered data set where to all the reads was added a pseudo-count of 1 (called sometimes Laplace Correction (LC) [84] [85] [86]), as also proposed in [63]. This method allows to account for the following problem: fold-change (FC) between low values is less statistically significant then the fold-change between high values, which are more reliable. For example: FC = 10/2 = 5; FC = 11/3 = 3.6, while FC = 100/20 = 5 and FC = 101/21 = 4.8.

	T1	T2	T3	T4	C1	C2	C3	C4
shRNA.1	9	11	1	1	19	21	1	1
shRNA.2	9	11	1	1	1	1	1	1

– NA: filtered data set in which 0 was changed into NA (not available elements) if the other elements of the shRNA were different from 0 (shRNA.1); otherwise zeros were unchanged (shRNA.2). Then the pseudo-count 1 was added everywhere. Changing 0 into NA means pointing out that 0 means that nothing was detected, because of experimental noise or other reasons.

	T1	T2	T3	T4	C1	C2	C3	C4
shRNA.1	8	10	NA	NA	18	20	NA	NA
shRNA.1	9	11	NA	NA	19	21	NA	NA
shRNA.2	8	10	NA	NA	0	0	0	0
shRNA.2	9	11	NA	NA	1	1	1	1

– Mean: the filtered data set in which each 0, was substituted with the mean of the other non-zero elements of the same condition , if the other elements of the shRNA were different from 0 (shRNA.1); otherwise zeros were unchanged (shRNA.2). Then the pseudo-count 1 was added.

	T1	T2	T3	T4	C1	C2	C3	C4
shRNA.1	8	10	9	9	18	20	19	19
shRNA.1	9	11	10	10	19	21	20	20
shRNA.2	8	10	9	9	0	0	0	0
shRNA.2	9	11	10	10	1	1	1	1

Then the results obtained analyzing these four data sets with the three statistics in **section 3.2.2**, were compared.

Initially there was the idea to not consider in the analysis the columns of the data set (with replicated experiments) that contained more than 50% of zeros, as suggested by Prof. Raffaele Calogero, because considered unreliable, too noisy. But in the case considered here, all the samples had more than 50% of non-zero elements. So not considering C2, C3 columns that contained 20% of zeros, was tried. But lately this approach was abandoned because it was leading to a great loss of information.

In the case of data sets without any replicated experiment dealing with zero was not a problem: seen that the elements equal to zero were a few (10-15) they were removed from the data set, instead of adding a pseudo-count of 1.

Normalization

edgeR and ***DESeq*** include a normalization method: ***edgeR*** uses TMM normalization, taking one sample of a certain condition as reference to compute TMM of the other samples; ***DESeq*** corrects the library size multiplying it by a scale factor estimated from the raw data. While Rank Product has not a normalization method.

A comparison between the raw data and the rescaled ones with the standard normalization (dividing each sample by the related library size, as mentioned in **section 3.2.2**), ***edgeR*** and ***DESeq*** proposed ones was made. The conclusion was that these three types of normalization are equivalent and so data to be submitted to Rank Product were previously rescaled with ***DESeq*** normalization, the most recent method. Actually the first analyses were run comparing RP results obtained on data sets with standard and DESeq normalization: the results were the same.

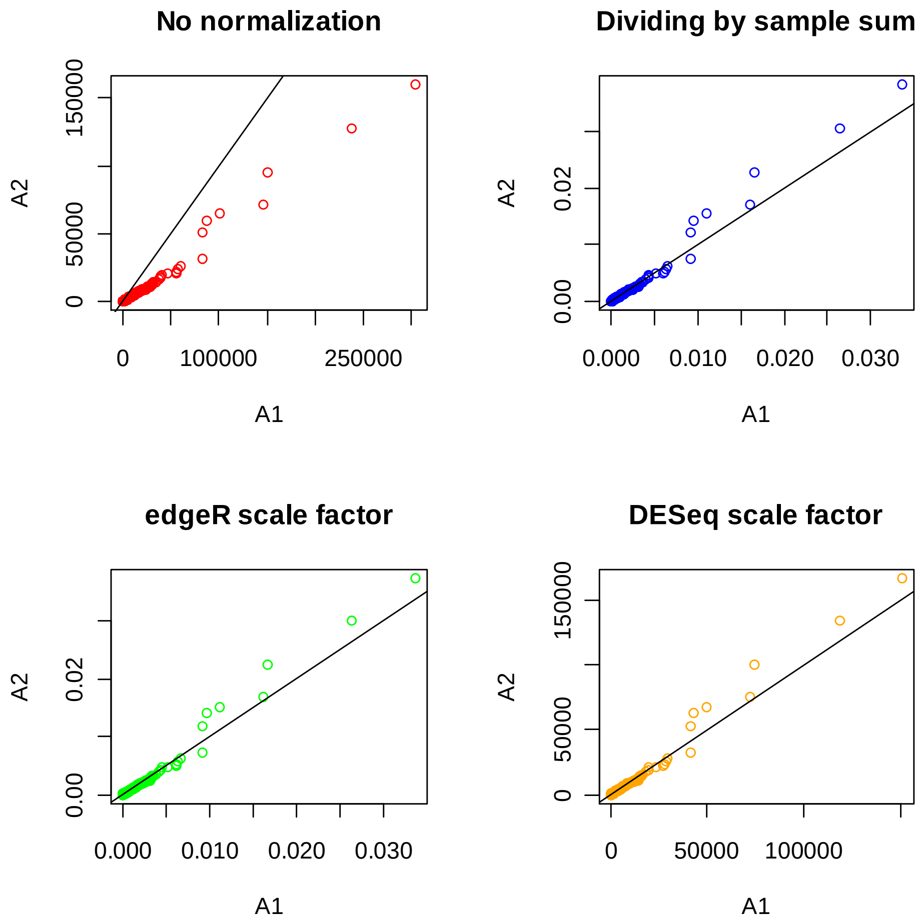


Fig.19 A is the control: sample A1 and A2 are compared with a qq-plot, before and after three types of normalization (dividing by sample sum, ***edgeR*** and ***DESeq*** factors).

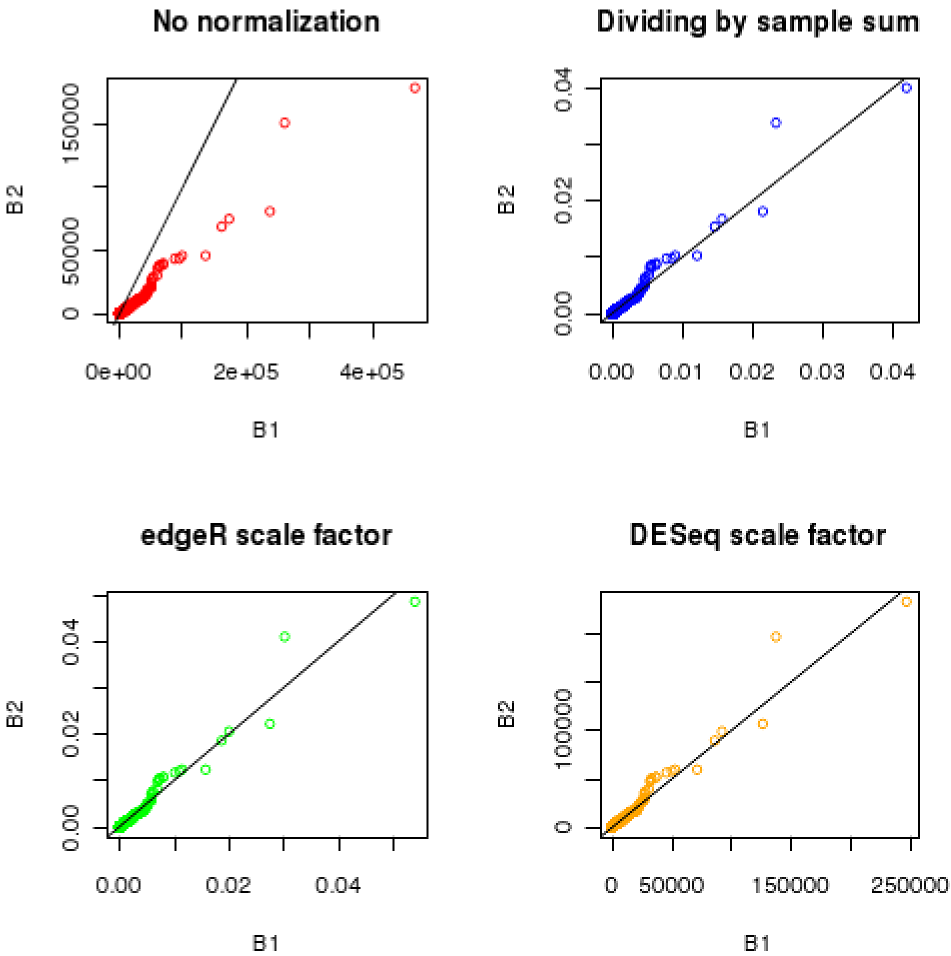


Fig.20 B is the treatment case: sample B1 and B2 are compared with a q-plot, before and after three types of normalization.

3.3.3 Regulation detection

Within all the results obtained with ***DESeq***, ***edgeR*** and ***RP*** only those with $log_2(FC) \geq 1$ and p-values ≤ 0.05 or p-adjusted ≤ 0.1 respectively, were retained. The p-values were adjusted with ***p.adjust*** algorithm of the R library ***stats***, with the method of Benjamini Hochberg (BH) [35], to account for multiple testing problem.

For each of the four possible data sets (Raw, LC, NA, Mean) log_2(FC) was computed (the same log_2(FC) of LC was assigned to NA). And only those regulated shRNAs which presented a consistent log_2(FC) were retained, i.e. only those for which (log2FC.LC/log2FC.Mean) > 0 so that the two values had the same sign to point out uniquely up (+) or down (-) regulation.

Data sets with replicates

To each of the three data sets were respectively applied one or more statistics:

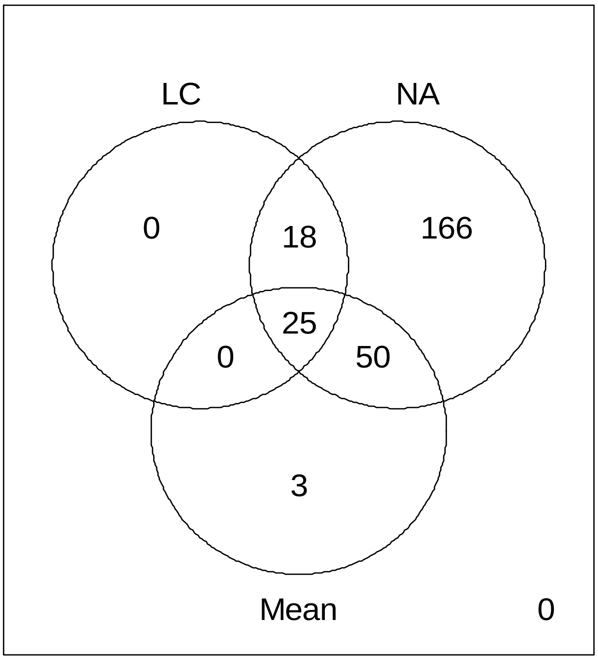
- Raw: ***DESeq, edgeR, RP.***
- LC: ***DESeq, edgeR, RP***.
- NA: ***RP*** was the only method applied because ***DESeq*** and ***edgeR*** can only analyze data without NA (for undefined elements), while ***RP*** can not consider NA in the analysis.
- Mean: ***DESeq, edgeR, RP***.

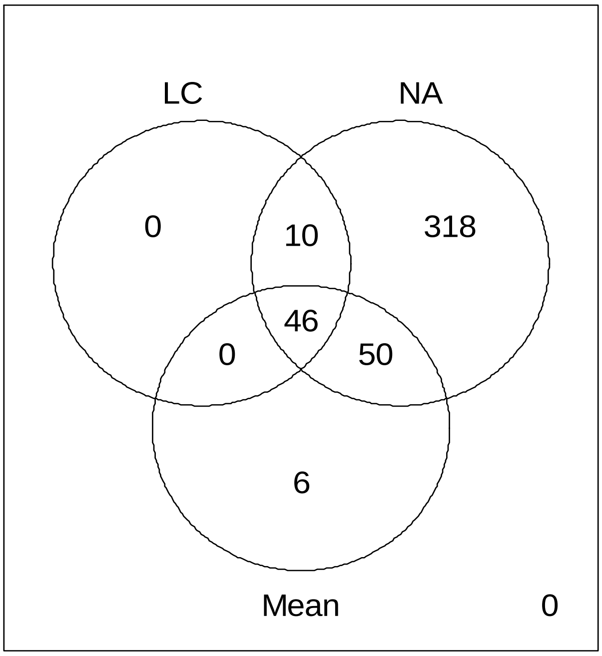
Then the intersection of the values (p-values, p-values adjusted) found with at least two methods (***edgeR*** and ***DESeq***; ***edgeR*** and ***RP***; ***DESeq*** and ***RP***) was taken.

Hereafter the Venn Diagrams of the results obtained is presented for the data sets R1, R2, R3. PVALUES label points out the number of shRNAs with p-value ≤ 0.05, while PADJUSTED label points out the number of shRNAs with p-value adjusted with BH method that is ≤ 0.1.

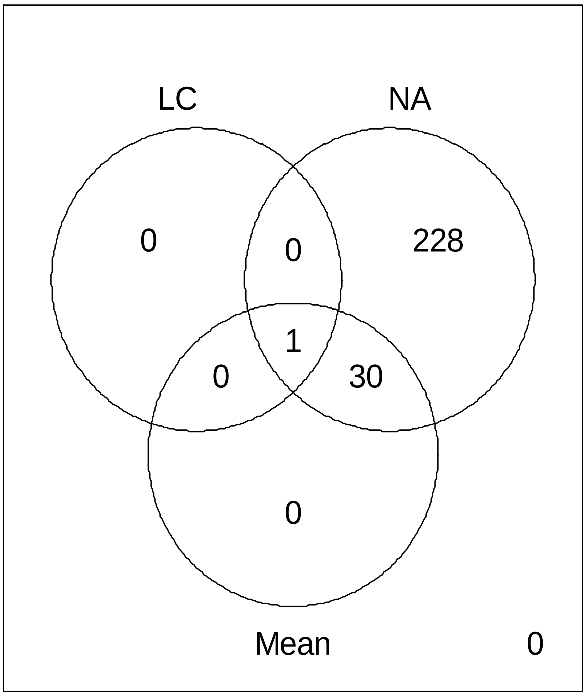
R1) **UP** **DOWN**

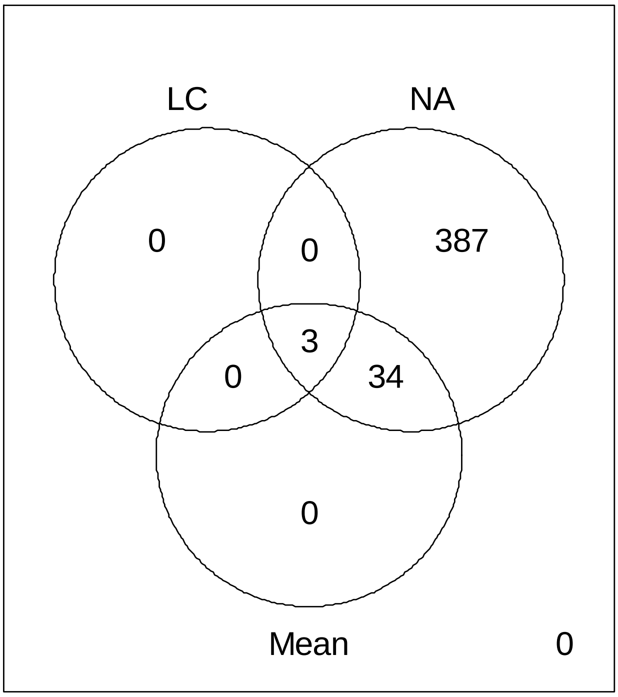
ShRNAs with p-value ≤ 0.05





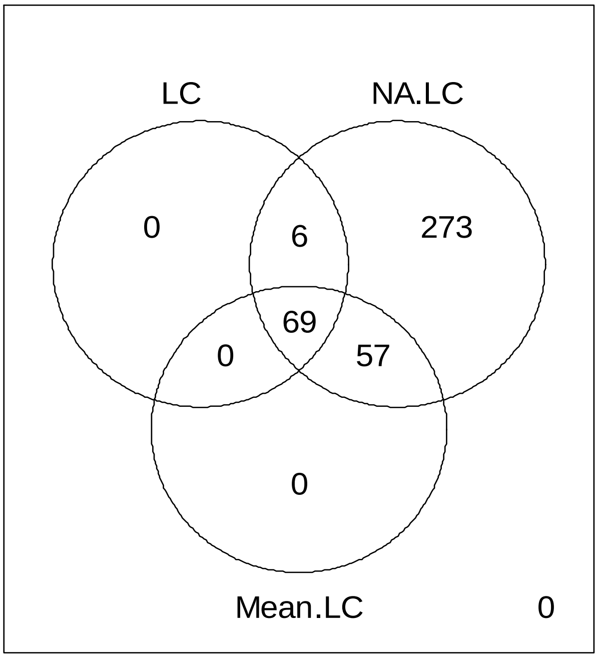
ShRNAs with p-value adjusted ≤ 0.1

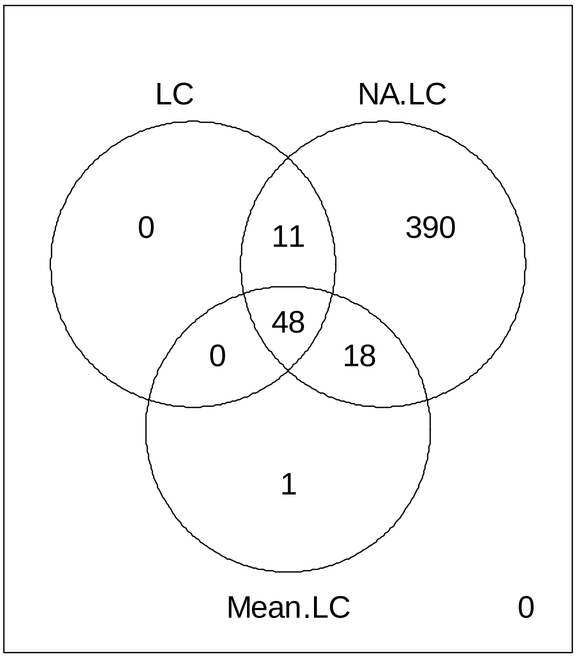




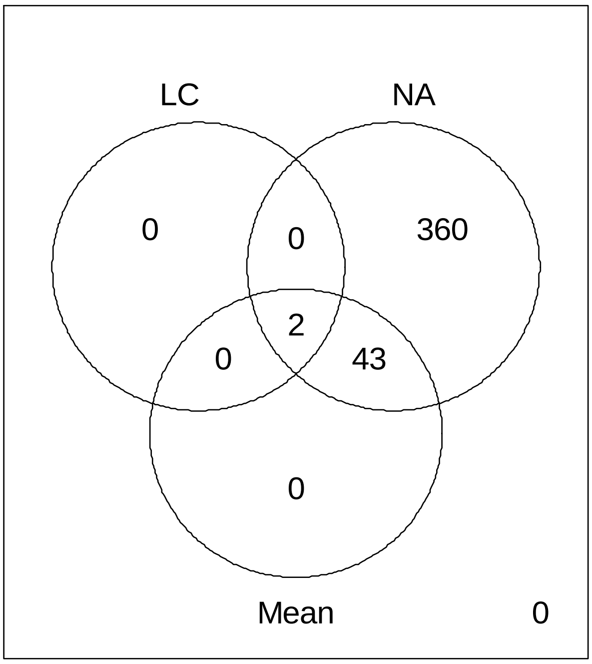
R2) **UP** **DOWN**

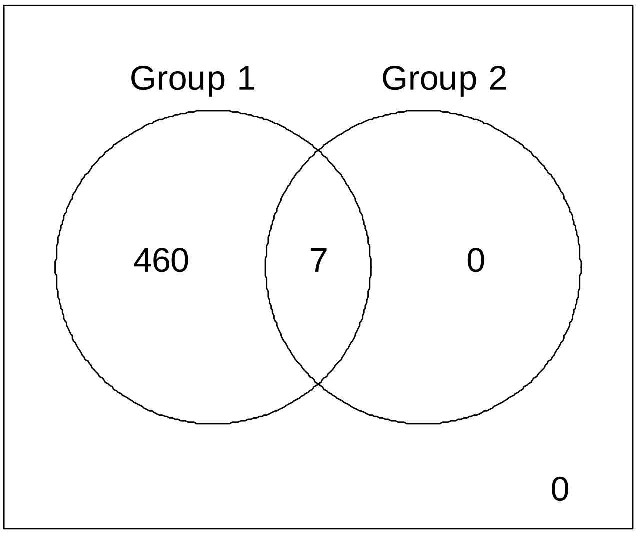
ShRNAs with p-value ≤ 0.05





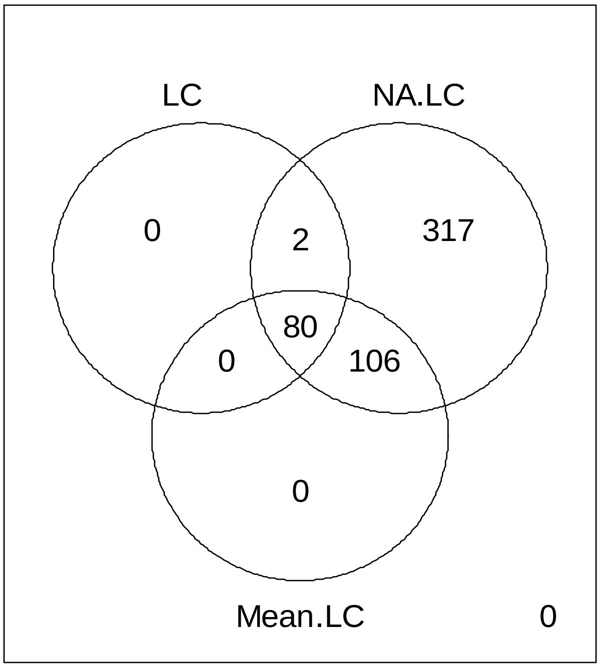
ShRNAs with p-value adjusted ≤ 0.1

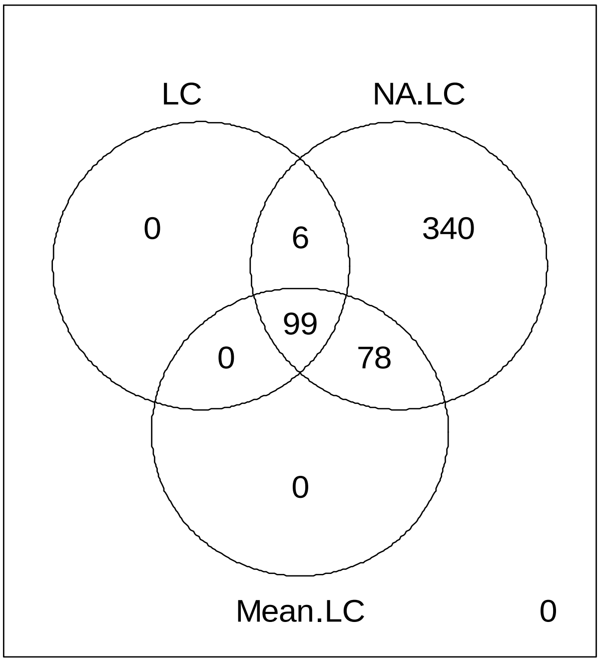




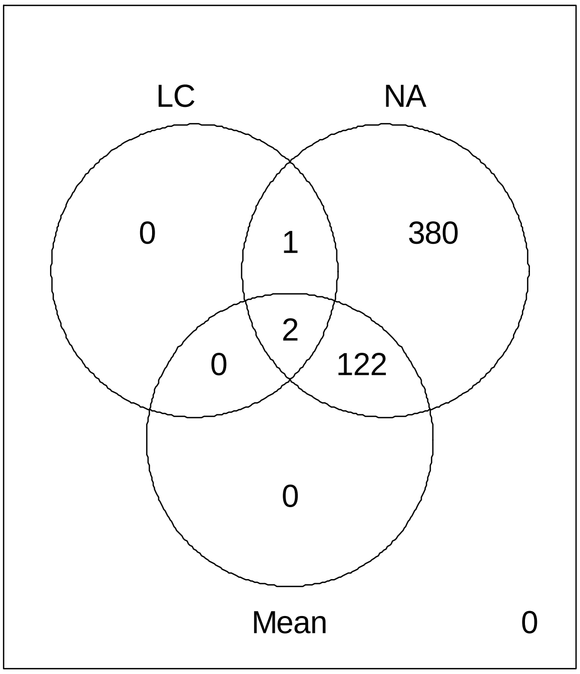
R3) UP DOWN

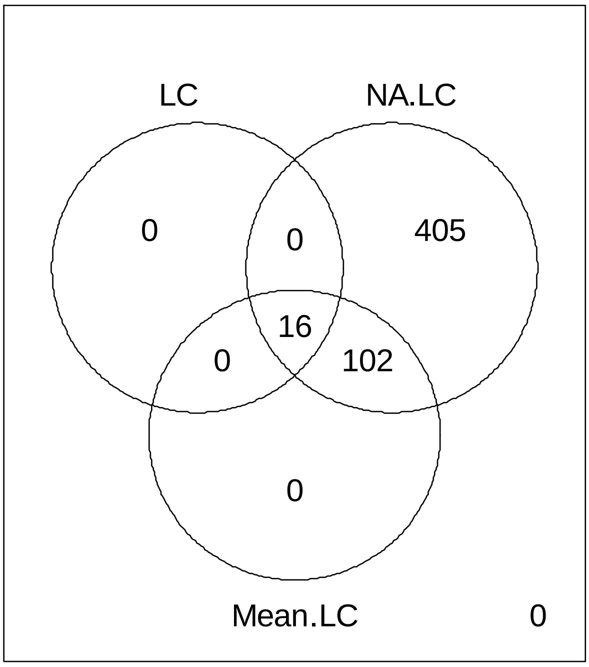
ShRNAs with p-value ≤ 0.05





ShRNAs with p-value adjusted ≤ 0.1





If the intersection of the results obtained with at least two methods between LC, NA and Mean is considered, the significant values are the following ones:

	P-values		P-values adjusted	
	UP	**DOWN**	**UP**	**DOWN**
R1	93	106	31	37
R2	132	77	45	7
R3	188	183	125	118

Table 7. P-values and p-values adjusted with BH method of the data sets with replicates (R1, R2, R3) are shown respectively for the up- and down-regulated shRNAs.

Data sets without replicates

To each of the three data sets were respectively applied: ***DESeq, edgeR, RP***, without any previous transformation of zeros because these were removed. ***edgeR*** does not work without replicated samples and it resulted unuseful. The union of the results found with ***DESeq*** and ***RP*** constituted the number of regulated shRNAs, as summarized in the table below:

	P-values		P-values adjusted	
	UP	**DOWN**	**UP**	**DOWN**
R4	1020	1945	0	79
M1	15	14	2	1
M2	12	19	0	1
M3	11	14	0	0

Table 8. P-values and p-values adjusted with BH method of the data sets with replicates (R4, M1, M2, M3, M4) are shown respectively for the up- and down-regulated shRNAs.

3.4 Conclusions

The results of this study are shown in the tables of the previous **section 3.3.3** and looking at them it is evident that Benjamini & Hochberg p-value adjustment cut off many data.

The pipe-line used to analyze the data without replicates is more simple than the one used for the other analysis and also less reliable because of this lack of available replication: hence no comparison through samples of the same condition can be made.

NA and Mean modified data sets give the best results and to reduce the number of false values the intersection of the common values is considered. NA seems to find much more regulated shRNAs than Mean but many of them are false values. Furthermore, the shRNAs whose regulation is detected by two or three different types of modified data sets instead of only one, are surely more reliable.

Afterwards part of the results, obtained for the three data sets with replicates (R1, R2, R3), is presented. All the presented shRNAs hereafter are up-regulated (if the control is lower than the treatment) or down-regulated (if the control is higher than the treatment) respectively and present a p-value lower or equal to 0.05, i.e. they are statistically significant. ShRNA IDs written in red (up-regulated) and green (down-regulated) point out that the results have the p-value adjusted with BH that is less or equal to the 0.1 threshold, beyond having p-value lower or equal to 0.05. ShRNA IDs highlighted in gray were detected simultaneously with LC, NA, Mean methods.

R1. Up-regulated shRNAs

	C1	C2	C3	C4	T1	T2	T3	T4	log2FC
u.R1.shRNA.1	0	0	0	0	1364,03	0	930,82	0	8,44
u.R1.shRNA.2	0	0	0	0	0	0	2015,43	111,80	8,33
u.R1.shRNA.3	0	3,06	0	0,42	126,17	233,79	2573,35	224,81	8,29
u.R1.shRNA.4	0	0	0	0	1930,22	0	0	115,41	8,27
u.R1.shRNA.5	0	0	0	0	973,85	7,34	368,16	633,56	8,23
u.R1.shRNA.6	0	0	7,97	4,63	475,53	1129,10	197,33	3756,87	8,18
u.R1.shRNA.7	0	0	3,98	0	2088,73	0	672,91	99,78	8,08
u.R1.shRNA.8	0	0	0	0,84	1396,37	351,21	185,97	54,10	8,06
u.R1.shRNA.9	0	0	0	0	0	737,01	561,71	0	7,62
u.R1.shRNA.10	0	0	0	0	331,33	165,64	0	502,52	7,24
u.R1.shRNA.11	2,75	0	0	6,31	1951,95	1,05	0	210,38	7,11
u.R1.shRNA.12	0	0	0	0	0	32,50	782,23	0	6,95
u.R1.shRNA.13	0	0	0	0	1,59	793,62	8,04	0	6,93
u.R1.shRNA.14	0	0	11,95	15,15	821,71	0	3174,34	1,20	6,89
u.R1.shRNA.15	0	3,06	0	0	1051,78	24,11	1,42	0	6,80
u.R1.shRNA.16	0	9,18	0	7,16	1256,42	17,82	0	1065,15	6,67
u.R1.shRNA.17	4,81	0	0	0,42	1139,26	0	0,47	0	6,59
u.R1.shRNA.18	0	0	0	0	0	1,05	548,93	2,40	6,39
u.R1.shRNA.19	0	0	0	11,79	1497,63	0	0	30,05	6,38
u.R1.shRNA.20	0	0	0	4,63	11,13	699,26	191,65	0	6,33
u.R1.shRNA.21	0	0	0	7,16	344,59	8,39	118,30	623,94	6,32
u.R1.shRNA.22	0	0	0	0	0	205,48	0,47	299,35	6,26
u.R1.shRNA.23	3,43	6,12	3,98	1,68	28,63	1370,22	194,97	30,05	6,22
u.R1.shRNA.24	0	0	0	2,95	680,16	0	4,26	0	6,17
u.R1.shRNA.25	0	0	0	0	0	453,95	0,95	16,83	6,16
u.R1.shRNA.26	0	0	0	29,04	109,74	12,58	0	2141,12	5,99
u.R1.shRNA.27	0	0	0	16,00	662,67	0	0	770,61	5,99
u.R1.shRNA.28	0	0	0	0	368,97	0	0,47	0	5,81
u.R1.shRNA.29	0	0	0	0	0	354,35	0,47	0	5,76
u.R1.shRNA.30	2,06	0	0	0	21,21	443,46	0	0	5,75
u.R1.shRNA.31	23,35	0	0	25,68	0	2878,83	0,95	0	5,70
u.R1.shRNA.32	0	16,84	0	12,63	0	91,21	1382,74	227,22	5,56
u.R1.shRNA.33	8,93	22,96	0	1,26	0	5,24	1737,65	0	5,46
u.R1.shRNA.34	0	0	0	0	2,65	0	273,52	0	5,40
u.R1.shRNA.35	0	0	0	0	0,53	242,17	0	0	5,21
u.R1.shRNA.36	0	12,24	0	7,58	19,61	31,45	614,24	300,55	5,20
u.R1.shRNA.37	0	22,96	0	0	0	3,15	1055,27	0	5,17
u.R1.shRNA.38	10,99	18,37	0	2,95	6,89	501,12	26,97	800,66	5,10
u.R1.shRNA.39	0	0	7,97	7,16	0,53	692,97	0	43,28	5,09
u.R1.shRNA.40	7,55	0	0	2,53	15,37	0	112,63	403,94	5,00
u.R1.shRNA.41	0	0	0	0	0	105,89	83,29	0	4,86
u.R1.shRNA.42	10,30	0	0	19,36	0	131,05	541,83	348,64	4,82
u.R1.shRNA.43	58,37	0	0	0	0,53	23,06	1166,48	583,07	4,77
u.R1.shRNA.44	0	0	0	115,34	0	0	1412,08	1897,07	4,76
u.R1.shRNA.45	0	19,90	0	0	427,29	30,40	0	253,66	4,75
u.R1.shRNA.46	26,10	42,86	39,85	14,73	20,68	6,29	2426,65	1014,66	4,74
u.R1.shRNA.47	4,12	0	0	0	2,12	0	0	278,91	4,73
u.R1.shRNA.48	0	12,24	0	0	0	460,24	1,89	0	4,62
u.R1.shRNA.49	0	1,53	0	21,05	220,01	2,10	170,36	292,13	4,56
u.R1.shRNA.50	0	0	0	0	5,83	138,39	0,95	0	4,49
u.R1.shRNA.51	0	0	0	0	0	0	0,47	143,06	4,47
u.R1.shRNA.52	0	0	27,89	18,10	21,74	199,19	408,86	522,96	4,46

R1. Down-regulated shRNAs

	C1	C2	C3	C4	T1	T2	T3	T4	log2FC
d.R1.shRNA.1	0	1951,51	0	149,43	0	2,10	24,61	0	-6,14
d.R1.shRNA.2	4166,44	0	286,89	0,84	0,53	0	58,68	1,20	-6,13
d.R1.shRNA.3	1,37	0	314,78	3,79	0	0	0,47	1,20	-6,05
d.R1.shRNA.4	0	39,80	769,03	88,40	1,06	9,44	0,47	0	-5,99
d.R1.shRNA.5	0	0	0	1133,57	0	0	0,47	14,43	-5,97
d.R1.shRNA.6	3112,98	0	442,29	39,57	0,53	69,19	0,47	0	-5,62
d.R1.shRNA.7	3874,58	2029,57	573,78	39,15	0	22,02	116,88	0	-5,52
d.R1.shRNA.8	1905,70	0	1494,22	515,64	5,83	4,19	1,42	94,97	-5,16
d.R1.shRNA.9	0	0	306,81	385,99	0	2,10	0	14,43	-5,14
d.R1.shRNA.10	0,69	0	800,90	0	1,59	0	0	20,44	-5,00
d.R1.shRNA.11	0	0	3363,00	0	0	18,87	0,95	82,95	-4,99
d.R1.shRNA.12	1499,15	134,69	2860,94	1514,09	0	187,66	17,04	0	-4,85
d.R1.shRNA.13	0	203,57	23,91	1120,10	0	38,79	0	4,81	-4,85
d.R1.shRNA.14	0	0	0	564,05	3,18	2,10	13,72	0	-4,68
d.R1.shRNA.15	0	47,45	0	737,47	0	26,21	3,31	0	-4,59
d.R1.shRNA.16	0	0	87,66	169,64	0,53	0	0,47	8,42	-4,38
d.R1.shRNA.17	41,89	0	0	2464,98	2,12	51,37	61,04	4,81	-4,36
d.R1.shRNA.18	439,51	4,59	0	4,21	11,13	3,15	5,21	0	-4,32
d.R1.shRNA.19	0	939,79	0	0	36,05	10,48	0	0	-4,25
d.R1.shRNA.20	26718,22	4,59	0	571,20	1077,76	0	193,55	207,98	-4,20
d.R1.shRNA.21	0	739,28	0	1,26	25,98	0	0	13,22	-4,14
d.R1.shRNA.22	0	0	0	844,81	0	6,29	0,47	39,67	-4,10
d.R1.shRNA.23	2133,01	0	1442,42	308,12	0,53	152,01	21,77	55,30	-4,06
d.R1.shRNA.24	1916,00	16851,85	167,35	244,98	3,71	679,35	0,95	462,85	-4,06
d.R1.shRNA.25	3558,68	130,10	0	460,50	164,87	0	66,72	21,64	-4,02
d.R1.shRNA.26	0	35,20	0	32,83	0,53	0	0	1,20	-3,90
d.R1.shRNA.27	5576,31	315,30	0	78,71	2,65	47,18	226,67	123,83	-3,89
d.R1.shRNA.28	0	691,83	3,98	160,80	27,57	30,40	0	0	-3,82
d.R1.shRNA.29	0	1270,39	31,88	150,69	0	1,05	0,95	100,98	-3,78
d.R1.shRNA.30	7830,19	7,65	5140,13	719,79	431,00	528,38	16,09	26,45	-3,77
d.R1.shRNA.31	3490,69	261,73	3,98	130,49	0	28,31	107,42	152,68	-3,74
d.R1.shRNA.32	257,53	7374,41	0	345,16	0	364,83	7,10	228,42	-3,73
d.R1.shRNA.33	111,94	0	7,97	444,50	0	23,06	4,26	18,03	-3,55
d.R1.shRNA.34	0	737,75	18799,31	100,60	56,19	678,30	2,37	949,74	-3,54
d.R1.shRNA.35	3,43	0	21664,24	657,07	531,72	6,29	0	1395,75	-3,53
d.R1.shRNA.36	0	5969,32	0	598,14	83,76	484,35	0	0	-3,52
d.R1.shRNA.37	34652,11	1,53	729,18	17,26	746,43	944,58	1479,75	36,07	-3,46
d.R1.shRNA.38	1,37	1790,80	147,43	0	1,06	174,03	4,73	1,20	-3,40
d.R1.shRNA.39	98,20	1763,25	135,48	0,42	0	189,76	1,42	0	-3,37
d.R1.shRNA.40	2703,00	0	7833,71	54,72	196,15	95,40	1,42	738,15	-3,36
d.R1.shRNA.41	0	45790,81	4654,01	560,26	203,04	2786,58	823,87	1226,24	-3,34
d.R1.shRNA.42	0	953,56	852,70	242,04	3,71	6,29	65,78	128,64	-3,31
d.R1.shRNA.43	0	55,10	2219,42	1198,39	162,75	0	88,49	98,58	-3,30
d.R1.shRNA.44	136,66	454,59	0	285,39	54,60	32,50	0	0	-3,29
d.R1.shRNA.45	0	1809,16	0	0,42	0	138,39	0	44,48	-3,29
d.R1.shRNA.46	22,66	1152,54	4952,85	3023,13	755,44	28,31	77,13	82,95	-3,27
d.R1.shRNA.47	1315,79	322,96	8662,51	436,09	954,77	27,26	0,47	143,06	-3,25
d.R1.shRNA.48	28,84	1317,84	1625,71	55,14	0,53	0	300,02	15,63	-3,25
d.R1.shRNA.49	19688,77	194,39	9168,55	2841,71	652,06	1611,35	481,73	663,61	-3,22
d.R1.shRNA.50	2225,03	1,53	1390,62	753,05	189,79	6,29	265,47	6,01	-3,22
d.R1.shRNA.51	23,35	4845,86	3,98	72,40	27,04	264,19	197,80	40,87	-3,22
d.R1.shRNA.52	9940,53	0	3426,75	252,56	36,58	351,21	644,99	453,23	-3,19

R2. Up-regulated shRNAs

	C1	C2	C3	T1	T2	T3	log2FC
u.R2.shRNA.1	0,89	1,12	1,05	17,45	3860,88	0	9,48
u.R2.shRNA.2	0	2,24	0	857,55	570,42	59,50	8,34
u.R2.shRNA.3	0	0	0	560,90	126,20	0	8,20
u.R2.shRNA.4	0	0	0,35	304,13	136,30	119,00	7,70
u.R2.shRNA.5	0	5,61	0	1353,64	257,45	7,68	7,67
u.R2.shRNA.6	3,56	0	0	4,99	1171,13	0	7,64
u.R2.shRNA.7	0	0	0	0	447,59	1,92	7,59
u.R2.shRNA.8	0	0	0	4,99	429,92	1,92	7,55
u.R2.shRNA.9	0	1,12	0	191,95	412,25	13,44	7,48
u.R2.shRNA.10	0	2,24	1,05	855,06	0,84	5,76	7,26
u.R2.shRNA.11	0	0	0	266,74	1,68	1,92	6,87
u.R2.shRNA.12	0,89	4,49	0,35	797,72	0	1,92	6,64
u.R2.shRNA.13	0	2,24	0	371,44	18,51	19,19	6,49
u.R2.shRNA.14	0	0	0,70	0	231,37	3,84	6,30
u.R2.shRNA.15	0,89	1,12	3,50	573,36	15,14	0	6,24
u.R2.shRNA.16	0	0	0,70	176,99	35,34	5,76	6,19
u.R2.shRNA.17	5,33	1,12	0,70	413,82	266,70	0	6,17
u.R2.shRNA.18	0	0	1,40	104,70	148,07	1,92	6,11
u.R2.shRNA.19	0	0	0	2,49	27,76	124,76	6,08
u.R2.shRNA.20	0	1,12	0	4,99	215,38	0	6,02
u.R2.shRNA.21	0	0	0,70	94,73	92,55	1,92	5,99
u.R2.shRNA.22	0	0	5,60	0	353,36	143,95	5,98
u.R2.shRNA.23	0	0	0	0	120,31	15,36	5,90
u.R2.shRNA.24	2,67	2,24	6,29	57,34	682,32	13,44	5,80
u.R2.shRNA.25	0	0	0	2,49	120,31	0	5,76
u.R2.shRNA.26	0	3,37	0	294,16	9,25	0	5,75
u.R2.shRNA.27	0	0	0,35	127,14	0	9,60	5,71
u.R2.shRNA.28	0	2,24	12,59	406,34	399,63	72,94	5,68
u.R2.shRNA.29	0	1,12	0	34,90	132,93	0	5,63
u.R2.shRNA.30	0	0	0	0	0,84	105,57	5,56
u.R2.shRNA.31	0	0	0	0	98,44	3,84	5,51
u.R2.shRNA.32	0	6,73	6,29	643,16	5,05	36,47	5,49
u.R2.shRNA.33	0	0	0	77,28	12,62	0	5,33
u.R2.shRNA.34	0	0	0,35	57,34	0,84	38,39	5,23
u.R2.shRNA.35	0	0	5,60	0	266,70	23,03	5,21
u.R2.shRNA.36	0	0	3,85	82,27	133,77	7,68	5,20
u.R2.shRNA.37	0	0	0	0	15,99	63,34	5,16
u.R2.shRNA.38	3,56	0	0	0	183,41	19,19	5,13
u.R2.shRNA.39	0	3,37	0	184,47	3,37	0	5,07
u.R2.shRNA.40	0	2,24	0	59,83	89,18	0	5,07
u.R2.shRNA.41	0	8,98	0	368,95	0,84	3,84	5,06
u.R2.shRNA.42	0	8,98	0	368,95	0,84	3,84	5,06
u.R2.shRNA.43	0	0	0	0	61,42	11,52	5,05
u.R2.shRNA.44	7,11	12,35	0,35	393,88	322,23	3,84	5,03
u.R2.shRNA.45	2,67	0	19,23	625,71	20,19	80,61	4,91
u.R2.shRNA.46	0,89	12,35	0,70	458,69	15,14	0	4,88
u.R2.shRNA.47	2,67	0	0	19,94	116,95	1,92	4,84
u.R2.shRNA.48	0	0	0	0	40,38	21,11	4,82
u.R2.shRNA.49	0	0	0	57,34	0	1,92	4,77
u.R2.shRNA.50	0	0	0	2,49	55,53	0	4,74
u.R2.shRNA.51	0	5,61	2,45	229,35	0,84	38,39	4,72
u.R2.shRNA.52	0	0	0	0	1,68	53,74	4,68

R2. Down-regulated shRNAs

	C1	C2	C3	T1	T2	T3	log2FC
d.R2.shRNA.1	61,35	1671,35	12,24	2,49	0	1,92	-7,50
d.R2.shRNA.2	1148,75	2246,05	123,10	7,48	3,37	9,60	-7,10
d.R2.shRNA.3	951,37	282,86	86,73	2,49	0	1,92	-7,10
d.R2.shRNA.4	563,71	974,30	129,04	0	5,05	1,92	-7,09
d.R2.shRNA.5	413,44	2840,95	0	4,99	10,94	3,84	-7,02
d.R2.shRNA.6	288,08	1094,40	0,35	0	0,84	5,76	-6,87
d.R2.shRNA.7	22,23	16699,99	18,53	52,35	34,49	67,18	-6,72
d.R2.shRNA.8	59,57	1359,30	0	7,48	0	1,92	-6,60
d.R2.shRNA.9	1417,27	920,42	396,22	2,49	21,87	3,84	-6,35
d.R2.shRNA.10	521,03	6120,79	170,66	0	79,93	3,84	-6,26
d.R2.shRNA.11	0	1515,32	0	12,46	2,52	0	-6,23
d.R2.shRNA.12	5092,03	9224,39	670,74	164,53	27,76	32,63	-6,03
d.R2.shRNA.13	106,70	411,94	0	0	0,84	1,92	-6,02
d.R2.shRNA.14	254,29	671,23	2,10	0	11,78	1,92	-5,62
d.R2.shRNA.15	157,38	1260,52	61,20	17,45	0	7,68	-5,61
d.R2.shRNA.16	1044,72	2537,89	236,40	24,93	2,52	47,98	-5,57
d.R2.shRNA.17	3054,15	9375,93	366,84	107,19	79,09	80,61	-5,56
d.R2.shRNA.18	223,17	6206,09	25,18	0	56,37	76,78	-5,54
d.R2.shRNA.19	3243,53	2040,64	221,36	114,67	3,37	0	-5,48
d.R2.shRNA.20	781,54	3347,18	46,16	0	27,76	61,42	-5,47
d.R2.shRNA.21	2160,58	4345,05	144,78	14,96	99,28	34,55	-5,43
d.R2.shRNA.22	500,58	3048,61	0	17,45	0	65,26	-5,34
d.R2.shRNA.23	305,86	3305,65	161,21	27,42	18,51	49,90	-5,22
d.R2.shRNA.24	146,71	4688,52	229,41	64,81	33,65	32,63	-5,22
d.R2.shRNA.25	3721,88	1091,03	209,82	32,41	74,88	30,71	-5,13
d.R2.shRNA.26	0	4464,03	0,35	99,72	0	24,95	-5,10
d.R2.shRNA.27	317,42	1269,50	46,16	4,99	4,21	34,55	-5,06
d.R2.shRNA.28	16,00	1419,91	57,00	34,90	5,89	1,92	-4,96
d.R2.shRNA.29	1138,08	1229,10	246,89	0	38,70	42,23	-4,92
d.R2.shRNA.30	445,45	1708,39	87,78	4,99	2,52	61,42	-4,92
d.R2.shRNA.31	110,25	106,63	14,34	0	0,84	1,92	-4,87
d.R2.shRNA.32	1664,44	569,09	197,23	54,84	0,84	23,03	-4,86
d.R2.shRNA.33	304,08	239,08	85,33	2,49	6,73	7,68	-4,83
d.R2.shRNA.34	4023,30	13193,42	371,74	9,97	440,86	174,66	-4,80
d.R2.shRNA.35	145,82	3113,71	92,67	77,28	4,21	36,47	-4,77
d.R2.shRNA.36	814,44	2796,05	152,12	122,15	0	13,44	-4,74
d.R2.shRNA.37	1143,42	1146,03	176,95	32,41	36,18	24,95	-4,64
d.R2.shRNA.38	1737,35	11975,55	438,88	314,10	49,64	214,97	-4,60
d.R2.shRNA.39	1186,98	2030,53	54,90	24,93	23,56	88,29	-4,53
d.R2.shRNA.40	2329,51	2614,21	108,76	74,79	37,02	103,65	-4,52
d.R2.shRNA.41	1579,09	4827,71	110,86	0	250,72	28,79	-4,52
d.R2.shRNA.42	220,50	197,55	16,79	0	6,73	7,68	-4,47
d.R2.shRNA.43	522,81	1256,04	54,20	57,34	18,51	3,84	-4,43
d.R2.shRNA.44	1244,78	1708,39	177,30	137,11	4,21	0	-4,42
d.R2.shRNA.45	546,81	1,12	62,25	7,48	15,99	0	-4,41
d.R2.shRNA.46	548,59	1302,06	39,52	0	62,26	23,03	-4,39
d.R2.shRNA.47	1312,35	3100,24	222,76	107,19	0,84	109,40	-4,38
d.R2.shRNA.48	3284,43	4653,73	102,81	181,98	173,31	30,71	-4,36
d.R2.shRNA.49	2255,71	1249,30	333,97	12,46	135,45	34,55	-4,35
d.R2.shRNA.50	3570,73	12633,31	1341,13	4,99	10,10	859,88	-4,32
d.R2.shRNA.51	885,57	3413,41	141,98	172,01	0	46,07	-4,31
d.R2.shRNA.52	4008,18	8674,39	137,09	27,42	519,10	119,00	-4,26

R3. Up-regulated shRNAs

	C1	C2	C3	T1	T2	T3	T4	log2FC
u.R3.shRNA.1	0,00	0,00	0,00	0,00	23467,85	25494,06	0,00	13,47
u.R3.shRNA.2	0,00	0,00	0,00	0,96	17170,01	8841,16	0,00	12,56
u.R3.shRNA.3	0,00	0,00	0,00	43,99	5613,10	4028,81	0,00	11,14
u.R3.shRNA.4	0,00	0,00	0,00	0,00	4995,95	3651,11	0,95	10,97
u.R3.shRNA.5	0,00	0,00	0,00	3,83	438,52	1056,08	5052,00	10,57
u.R3.shRNA.6	0,00	0,00	0,00	20,08	5915,64	4,44	544,70	10,56
u.R3.shRNA.7	0,00	0,00	0,00	2416,78	0,00	2904,59	12,32	10,28
u.R3.shRNA.8	0,00	0,00	0,00	297,44	1094,30	3288,22	23,68	10,10
u.R3.shRNA.9	3,92	0,00	0,00	321,35	1347,76	5,92	8576,94	10,07
u.R3.shRNA.10	0,00	0,00	0,00	350,99	0,00	0,00	4000,49	9,98
u.R3.shRNA.11	0,00	0,00	1,17	69,82	383,00	1528,58	3790,18	9,94
u.R3.shRNA.12	0,00	0,00	0,00	0,00	4127,75	0,00	72,94	9,93
u.R3.shRNA.13	0,00	0,00	0,00	0,00	580,14	700,60	2190,18	9,66
u.R3.shRNA.14	0,00	0,00	0,00	1917,55	1351,78	1,48	99,47	9,61
u.R3.shRNA.15	0,00	0,00	0,00	2005,54	1307,52	1,48	0,00	9,59
u.R3.shRNA.16	0,00	0,00	0,00	0,96	2,41	3246,75	5,68	9,57
u.R3.shRNA.17	0,00	0,00	4,68	7403,37	6,44	54,80	0,00	9,47
u.R3.shRNA.18	0,00	0,00	0,00	2411,05	377,37	39,99	0,95	9,36
u.R3.shRNA.19	0,00	0,00	3,51	4248,26	0,80	0,00	1431,38	9,31
u.R3.shRNA.20	0,00	0,00	0,00	187,45	0,00	5,92	2374,90	9,22
u.R3.shRNA.21	0,00	0,00	0,00	0,00	2120,20	0,00	406,40	9,20
u.R3.shRNA.22	0,00	16,51	2,34	127,20	6976,95	8140,56	1392,54	9,14
u.R3.shRNA.23	0,00	0,00	0,00	0,00	2410,67	4,44	0,00	9,13
u.R3.shRNA.24	0,00	2,25	3,51	2043,79	1918,24	1764,08	985,20	9,13
u.R3.shRNA.25	1,31	3,00	0,00	1589,51	3644,98	14,81	0,00	9,03
u.R3.shRNA.26	0,00	0,00	0,00	41,12	439,33	453,24	1221,08	8,97
u.R3.shRNA.27	0,00	0,00	0,00	167,37	1973,76	0,00	0,00	8,96
u.R3.shRNA.28	0,00	7,51	0,00	0,00	1809,61	2,96	5035,90	8,90
u.R3.shRNA.29	0,00	0,00	0,00	0,00	1206,95	622,10	178,09	8,87
u.R3.shRNA.30	0,00	0,00	0,00	658,95	1248,79	48,88	50,21	8,87
u.R3.shRNA.31	0,00	4,50	0,00	0,96	1915,02	1164,21	1592,43	8,83
u.R3.shRNA.32	1,31	0,00	0,00	0,00	613,13	1541,91	444,29	8,75
u.R3.shRNA.33	0,00	0,00	0,00	882,74	0,00	956,84	0,00	8,74
u.R3.shRNA.34	0,00	0,00	0,00	0,00	1810,42	2,96	0,00	8,72
u.R3.shRNA.35	0,00	0,00	2,34	2,87	2977,94	1,48	0,00	8,65
u.R3.shRNA.36	0,00	0,75	0,00	43,99	0,00	4,44	2068,92	8,64
u.R3.shRNA.37	10,46	0,00	2,34	241,01	5555,17	2283,98	397,87	8,63
u.R3.shRNA.38	1,31	0,00	0,00	0,00	0,00	2208,44	178,09	8,63
u.R3.shRNA.39	0,00	0,00	2,34	1134,27	1789,50	0,00	0,00	8,62
u.R3.shRNA.40	0,00	0,00	0,00	0,96	0,00	0,00	1688,10	8,62
u.R3.shRNA.41	0,00	0,00	0,00	1616,29	0,00	4,44	2,84	8,56
u.R3.shRNA.42	0,00	0,00	0,00	1129,49	31,38	305,12	154,41	8,56
u.R3.shRNA.43	2,61	0,00	0,00	1920,42	0,80	881,30	0,00	8,49
u.R3.shRNA.44	0,00	27,77	0,00	1256,69	929,35	17,77	11425,50	8,37
u.R3.shRNA.45	0,00	0,00	0,00	384,47	453,01	552,48	1,89	8,34
u.R3.shRNA.46	2,61	0,00	3,51	5,74	1380,75	1692,99	760,69	8,27
u.R3.shRNA.47	0,00	2,25	0,00	858,83	0,00	648,76	656,49	8,21
u.R3.shRNA.48	3,92	4,50	2,34	28,69	571,29	730,22	3919,97	8,14
u.R3.shRNA.49	3,92	4,50	2,34	28,69	571,29	730,22	3919,97	8,14
u.R3.shRNA.50	0,00	0,00	0,00	268,74	15,29	909,44	0,00	8,12
u.R3.shRNA.51	0,00	0,00	1,17	496,36	613,13	493,23	0,00	8,10
u.R3.shRNA.52	0,00	0,00	0,00	0,00	0,00	528,78	598,70	8,04

R3. Down-regulated shRNAs

	C1	C2	C3	T1	T2	T3	T4	log2FC
d.R3.shRNA.1	4896,19	1107,84	0	0	1,61	5,92	0	-9,42
d.R3.shRNA.2	2650,09	135,85	0	0,96	0	1,48	0	-9,13
d.R3.shRNA.3	1,31	3,75	2279,03	2,87	2,41	0	0	-8,33
d.R3.shRNA.4	10073,46	1727,80	3328,45	50,69	0	16,29	8,53	-7,98
d.R3.shRNA.5	1835,58	600,45	1,17	0	0	7,41	1,89	-7,91
d.R3.shRNA.6	7887,50	4263,22	380,23	0	0,80	0	67,26	-7,85
d.R3.shRNA.7	26,15	0	1395,73	1,91	0	2,96	0	-7,71
d.R3.shRNA.8	2426,52	0	0	3,83	0	0	9,47	-7,53
d.R3.shRNA.9	2677,54	2,25	0	14,35	0,80	0	0	-7,53
d.R3.shRNA.10	1337,46	3934,47	2459,19	3,83	9,66	39,99	0	-7,48
d.R3.shRNA.11	0	174,88	2511,84	7,65	0	8,89	0	-7,43
d.R3.shRNA.12	2672,31	0	51,48	0	0	10,37	6,63	-7,42
d.R3.shRNA.13	4004,54	2811,62	98,27	35,39	0	14,81	2,84	-7,33
d.R3.shRNA.14	0	0	760,46	0	2,41	0	0,95	-7,08
d.R3.shRNA.15	753,06	0	10,53	0,96	0	0	2,84	-7,00
d.R3.shRNA.16	0	44,28	1634,39	0	0	11,85	1,89	-6,97
d.R3.shRNA.17	504,65	0	0	0,96	0	0	0,95	-6,80
d.R3.shRNA.18	0	0	17526,73	0	8,85	0	202,72	-6,76
d.R3.shRNA.19	1,31	623,72	538,17	0	1,61	8,89	0	-6,73
d.R3.shRNA.20	0	0,75	1735,01	0,96	19,31	0	0	-6,57
d.R3.shRNA.21	1,31	0	549,87	1,91	1,61	1,48	0	-6,33
d.R3.shRNA.22	0	655,24	10138,62	0	0,80	189,59	0	-6,21
d.R3.shRNA.23	465,43	2371,79	1121,96	15,30	3,22	47,40	5,68	-6,12
d.R3.shRNA.24	3,92	3,75	25600,43	562,35	0	0	0,95	-5,91
d.R3.shRNA.25	1519,19	1504,89	1,17	14,35	0	48,88	0,95	-5,88
d.R3.shRNA.26	0	1020,77	792,04	4,78	0	32,59	0	-5,86
d.R3.shRNA.27	4303,94	109,58	797,89	97,55	22,53	0	0	-5,81
d.R3.shRNA.28	768,75	3,00	1614,50	53,56	0,80	0	0	-5,77
d.R3.shRNA.29	31,38	1123,60	60,84	2,87	0	23,70	0,95	-5,68
d.R3.shRNA.30	1115,21	3014,27	0	0	99,77	2,96	0,95	-5,67
d.R3.shRNA.31	2,61	1001,26	0	0,96	2,41	20,74	0	-5,57
d.R3.shRNA.32	1,31	0	6074,28	117,64	18,51	29,62	2,84	-5,55
d.R3.shRNA.33	3613,63	294,22	2,34	0	57,93	0	58,73	-5,43
d.R3.shRNA.34	5866,27	3761,84	18,72	91,81	39,43	179,22	0,95	-5,35
d.R3.shRNA.35	0	117,09	1443,69	45,91	0	5,92	0,95	-5,19
d.R3.shRNA.36	1261,63	0	0	13,39	0	28,14	0,95	-5,18
d.R3.shRNA.37	1842,12	86,32	861,07	73,64	0	0	28,42	-5,13
d.R3.shRNA.38	19851,45	14146,68	1132,49	849,27	175,41	299,20	13,26	-5,13
d.R3.shRNA.39	2580,79	0	0	92,77	2,41	1,48	0	-5,09
d.R3.shRNA.40	1460,36	0	0	46,86	6,44	0	0	-5,09
d.R3.shRNA.41	1893,10	412,06	2,34	35,39	24,14	0	28,42	-5,06
d.R3.shRNA.42	7680,93	6016,54	0	82,25	10,46	448,80	5,68	-5,05
d.R3.shRNA.43	4296,09	12,76	1977,18	61,21	42,65	148,12	0	-5,03
d.R3.shRNA.44	0	0	1016,67	28,69	10,46	0	0	-4,97
d.R3.shRNA.45	1,31	1236,93	10665,09	234,31	245,41	0	26,52	-4,96
d.R3.shRNA.46	8996,17	10881,71	1704,59	68,86	4,02	1014,61	1,89	-4,72
d.R3.shRNA.47	2182,04	68,30	527,64	2,87	20,12	114,05	0,95	-4,71
d.R3.shRNA.48	1936,25	382,79	0	6,69	4,02	102,20	2,84	-4,69
d.R3.shRNA.49	0	683,77	4886,80	0	3,22	281,42	0	-4,69
d.R3.shRNA.50	320,31	21,02	90,08	0	0,80	11,85	5,68	-4,69
d.R3.shRNA.51	0	48,04	5185,13	73,64	0	0	195,15	-4,68
d.R3.shRNA.52	1,31	114,84	6697,85	333,78	0,80	2,96	38,84	-4,58

Acknowledgments

I desire first of all to thank Prof. Raffaele Calogero with whom I began a collaboration at San Luigi Hospital of Orbassano (Torino) in the end of 2007 and then he became my supervisor when I started the Ph.D. in Complex Systems in Post-Genomic Biology in January 2008. I want to thank also Prof. Guido Forni, Prof. Federica Cavallo and Irene Merighi who welcame me and shared their offices with me at San Luigi Hospital and at the Molecular Biotechnology Centre of Torino, where I worked while doing the Ph.D. in Italy. I desire to thank Dr. Francesca Cordero, with whom I collaborated during my Ph.D.

I want to thank Prof. Frank Klawonn who hosted and supervised me the last year of Ph.D. since March 2010, when I moved at the Helmholtz Zentrum für Infektionsforschung - Helmholtz Centre for Infection Research in Braunschweig (Germany) for an internship of nine months.

Another acknowledgment goes to Dr. Lothar Jänsch and his Cellular Proteomics group that welcame me at the Helmholtz Centre as new member of the group; Thorsten Johl and Christoph Gernet who shared with me their office; Zofia Magnowska who shared with me these last months of Ph.D. during which both of us had to redact our thesis.

I thank Dr. Torsten Wüstefeld and Prof. Lars Zender, leader of the group of Chronic Infections and Cancer at Helmholtz Centre, and their collaborators Ramona Rudalska and Marina Pesic, who kindly provided me the data that I analyzed and presented in the last chapter of this work.

I thank my family and all the friends that have always supported me during all my education and in the three years of Ph.D. I desire to thank in particular my husband Riccardo, who have made possible to begin our new life as a married couple, living together in Germany since March 2010 and who supported me while writing this manuscript.

References

[1] M. Sammeth, S. Foissac, R. Guigó, **A general definition and nomenclature for alternative splicing events.** *PLoS Comput Biol.* 2008, 4 (8): e1000147.

[2] D. Abdueva, M.R. Wing, B. Schaub, T.J. Triche, **Experimental Comparison and Evaluation of the Affymetrix Exon and U133Plus2 GeneChip Arrays.** *PLoS ONE* 2007, 2 (9): e913.

[3] R.C. Gentleman, V.J. Carey, D.M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Yongchao Ge, J. Gentry, K. Hornik, et al. **Bioconductor: open software development for computational biology and bioinformatics.** *Genome Biology* 2004, 5: R80.

[4] R. Sanges, F. Cordero, R.A. Calogero, **oneChannelGUI: a graphical interface to Bioconductor tools, designed for life scientists who are not familiar with R language.** *Bioinformatics* 2007, 23: 3406-3408.

[5] G. K. Smyth, **Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments.** *Stat. Appl. Genet. Mol. Biol.* 2004, 3.

[6] I. Lönnstedt, T. P. Speed, **Replicated microarray data.** *Statistica Sinica* 2002, 12: 31-46.

[7] **Alternative Transcript Analysis Methods for Exon Arrays**. Affymetrix GeneChip® Exon Array Whitepaper Collection 2005.

[8] Cheng Li, Wing Hung Wong, **Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection.** *Proc Natl Acad Sci USA* 2001, 98 (1): 31-36.

[9] M.J. Okoniewski, J.C. Miller, **Comprehensive Analysis of Affymetrix Exon Arrays Using BioConductor.** *PLoS Comput Biol* 2008, 4 (2): e6.

[10] R.A. Irizarry, B.M. Bolstad, F. Collin, L.M. Cope, B. Hobbs, T.P. Speed, **Summaries of Affymetrix GeneChip probe level data.** *Nucleic Acids Res* 2003, 31: e15.

[11] **Guide to Probe Logarithmic Intensity Error (PLIER) Estimation.** *Affymetrix* Technical Note (2004).

[12] T.M. Therneau, K.V. Ballman, **What Does PLIER Really Do?** *Cancer Inform.* 2008, 6: 423-431.

[13] C. Della Beffa, F. Cordero, R.A. Calogero, **Dissecting an alternative splicing analysis workflow for GeneChip® Exon 1.0 ST Affymetrix arrays.** *BMC Genomics* 2008, 28 (9): 571.

[14] F. Hong, R. Breitling, C.W. McEntee, B.S. Wittner, J.L. Nemhauser, J. Chory, **RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis.** *Bioinformatics* 2006, 22 (22): 2825-2827.

[15] R. Breitling, P. Armengaud, A. Amtmann, P. Herzyk, **Rank Products: A simple, yet powerful, new method to detect differentially regulated genes**

in replicated microarray experiments. *FEBS Lett.* 2004, 573 (1-3): 83-92.

[16] R. Tibshirani, T. Hastie, **Outlier sums for differential gene expression analysis.** *Biostatistics* 2007, 8 **(**1): 2-8.

[17] **Baolin Wu, Cancer outlier differential gene expression detection.** *Biostatistics* 2007, 8 **(**3): 566-575.

[18] Xing Yi, P. Stoilov, K. Kapur, H. Areum, Jiang Hui, Shen Shihao,D.L. Black, **Hung Wong Wing, MADS: A new and improved method for analysis of differential alternative splicing by exon-tiling microarrays.** *RNA* 2008, 14: 1470-1479.

[19] E. Purdom, K.M. Simpson, M.D. Robinson, J.G. Conboy, A.V. Lapuk, T.P. Speed, **FIRMA: a method for detection of alternative splicing from exon array data.** *Bioinformatics* 2008, 24 (15): 1707-1714.

[20] M.A. Anton, D. Gorostiaga, E. Guruceaga, V. Segura, P. Carmona-Saez, A. Pascual-Montano, P. Pio, L.M. Montuenga, A. Rubio, **SPACE: an algorithm to predict and quantify alternatively spliced isoforms using microarrays.** *Genome Biology* 2008, 9: R46.

[21] S.H. Shah, J.A. Pallas, **Identifying differential exon splicing using linear models and correlation coefficients.** *BMC Bioinformatics* 2009, 10: 26.

[22] P.J. Gardina, T.A. Clark, B. Shimada, M.K. Staples, Q. Yang, J. Veitch, A. Schweitzer, T. Awad, C. Sugnet, S. Dee, et al., **Alternative splicing and**

differential gene expression in colon cancer detected by a whole genome exon array. *BMC Genomics* 2006, **7**: 325.

[23] M.J. Okoniewski, Y. Hey, S.D. Pepper, C.J. Miller, **High correspondence between Affymetrix exon and standard expression arrays.** *Biotechniques* 2007, 42 (2): 181-185.

[24] S.E. Choe, M. Boutros, A.M. Michelson, G.M. Church, M.S. Halfon, **Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset.** *Genome Biol* 2005, 6 (2): R16.

[25] P. Flicek, B.L. Aken, K. Beal, B. Ballester, M. Caccamo, Y. Chen, L. Clarke, G. Coates, F. Cunningham, T. Cutts, et al., **Ensembl 2008.** *Nucleic Acids Res* 2008, D707-714.

[26] F. Bretz, J. Landgrebe, E. Brunner, **Multiplicity issues in microarray experiments.** *Methods Inf Med* 2005, 44 (3): 431-437.

[27] L. Lusa, E.L. Korn, L.M. McShane, **A class comparison method with filtering-enhanced variable selection for high-dimensional data sets.** *Stat Med* 2008, 27 (28): 5834-5849.

[28] S.C. Lenzken, S. Vivarelli, F. Zolezzi, F. Cordero, C. Della Beffa, R.A. Calogero, S. Barabino, **Genome-Wide Search for Splicing Defects Associated with Amyotrophic Lateral Sclerosis (ALS).** International Conference on Complex, Intelligent and Software Intensive Systems; CISIS 2009, 795-799.

[29] S. Boillée, C. Vande Velde, D.W. Cleveland, **ALS: a disease of motor neurons**

and their nonneuronal neighbors. *Neuron* 2006, 52: 39-59.

[30] M. Pantelidou, S.E. Zographos, C.W. Lederer, T. Kyriakides, M.W. Pfaffl, N. Santama, **Differential expression of molecular motors in the motorcortex of sporadic ALS.** *Neurobiol Dis.* 2007, 26: 577-589.

[31] J. Robertson, M.M. Doroudchi, M.D. Nguyen, H.D. Durham, M.J. Strong, G. Shaw, J.P. Julien, W.E. Mushynski, **A neurotoxic peripherin splice variant in a mouse model of ALS.** *J Cell Biol.* 2003, 160: 939-949.

[32] B.M. Bolstad, R.A. Irizarry, M. Astrand, T.P. Speed, **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.** *Bioinformatics* 2003, 19: 185-193.

[33] A. Maracchioni, A. Totaro, D.F. Angelini, A. Di Penta, G. Bernardi, M.T. Carrì, T. Achsel, **Mitochondrial damage modulates alternative splicing in neuronal cells: implications for neurodegeneration.** *J Neurochem* 2007, 100: 142-153.

[34] P. Bonizzoni, R. Rizzi, G. Pesole, **ASPIC: a novel method to predict the exon intron structure of a gene that is optimally compatible to a set of transcript sequences.** *BMC Bioinformatics* 2005, 6: 244.

[35] Y. Benjamini, Y. Hochberg, **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *J. R. Statist. Soc. B* 1995, 57: 289-300.

[36] G. Parmigiani, E.S. Garrett-Mayer, R. Anbazhagan, E. Gabrielson, **A crossstudy comparison of gene expression studies for the molecular**

classification of lung cancer. *Clin Cancer Res.* 2004, 10: 2922-2927.

[37] C. Mitchelmore, S. Büchmann-Møller, L. Rask, M.J. West, J.C. Troncoso, N.A. Jensen, **NDRG2: a novel Alzheimer's disease associated protein.** *Neurobiol Dis.* 2004, 16: 48-58.

[38] T. Okuda, K. Kokame, T. Miyata, **Differential expression patterns of NDRG family proteins in the central nervous system.** *J Histochem Cytochem.* 2008, 56: 175-182.

[39] B. Nait-Oumesmar, N. Picard-Riéra, C. Kerninon, A. Baron-Van Evercooren, **The role of SVZ-derived neural precursors in demyelinating diseases: from animal models to multiple sclerosis.** *J Neurol Sci.* 2008, 265: 26-31.

[40] S. Donald, T. Humby, I. Fyfe, A. Segonds-Pichon, S.A. Walker, S.R. Andrews, W.J. Coadwell, P. Emson, L.S. Wilkinson, H.C. Welch, **P-Rex2 regulates Purkinje cell dendrite morphology and motor coordination.** *Proc Natl Acad Sci U S A.* 2008, 105: 4483-4488.

[41] H. van Bakel, C. Nislow, B.J. Blencowe, T.R. Hughes. **Most "dark" matter transcripts are associated with known genes.** *PLoS Biol* 2010, 8 (5): e1000371.

[42] F. Sanger, A.R. Coulson. **A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase.** *J. Mol. Biol.* 1975, 94 (3): 441-

[43] J. Shendure, H. Ji, **Next-generation DNA sequencing.** *Nature Biotechnology* 2008, 26 (10).

[44] A.E. Men, P. Wilson, K. Siemering, S. Forrest, **Sanger DNA Sequencing,** Next-Generation Genome Sequencing: Towards Personalized Medicine (2008).

[45] **Roche 454 Sequencing.** System Features for GS FLX Titanium Series. (2008) http://www.454.com/products-solutions/systemfeatures.asp

[46] Feature Report: **Next-generation sequencing: Synergy with Microarrays**, *University Health Network Microarray Centre* (2009).

[47] O. Harismendy, P.C. Ng, R.L. Strausberg, XiaoyunWang, T.B. Stockwell, K.Y. Beeson, N.J. Schork, S.S. Murray, E.J. Topol, S. Levy, K.A. Frazer, **Evaluation of next generation sequencing platforms for population targeted sequencing studies.** *Genome Biology* 2009, 10:R32.

[48] M. Margulies, M. Egholm, W.E. Altman, S. Attiya, J.S. Bader, L.A. Bemben, J. Berka, M.S. Braverman, Y.J. Chen, Z. Chen, S.B. Dewell, et al., **Genome sequencing in microfabricated high-density picolitre reactors.** *Nature* 2005, 437 (7057): 376-380.

[49] S. Bennett, **Solexa.** *Ltd. Pharmacogenomics* 2004, 5: 433-438.

[50] J. Shendure, G.J. Porreca, N.B. Reppas, Xiaoxia Lin, J.P. McCutcheon, A.M. Rosenbaum, M.D. Wang, Kun Zhang, R.D. Mitra, G.M. Church, **Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome.** *Science* 2005, 309 (5741): 1728-1732.

[51] Di-Base Sequencing and the Advantages of Color-Space Analysis in the SOLiD System. Application Note.

[52] B. Ondov, A. Varadarajan, K.D. Passalacqua, N.H. Bergman, **Efficient**

mapping of Applied Biosystems SOLiD sequence data to a reference genome for functional genomic applications. *Bioinformatics* 2008, 24 (23): 2776-2777.

[53] S.M. Rumble, P. Lacroute, A.V. Dalca, M. Fiume, A. Sidow, **SHRiMP: Accurate Mapping of Short Color-space Reads.** *PLoS Comput Biol* 2009, 5 (5): e1000386.

[54] M. Hackenberg, M.Sturm, D. Langenberger, J.M. Falcon-Perez, A.M. Aransay, **miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments.** *Nucl. Acids Res.* 2009, 37(2): W68-W76.

[55] A.K. Emde, M. Grunert, D. Weese, K. Reinert, S.R. Sperling, **MicroRazerS: Rapid alignment of small RNA reads.** *Bioinformatics* 2010, 26(1): 123-124.

[56] Wei-Chi Wang, Feng-Mao Lin, Wen-Chi Chang, Kuan-Yu Lin, Hsien-Da Huang, Na-Sheng Lin, **miRExpress: Analyzing high-throughput sequencing data for profiling microRNA expression.** *BMC Bioinformatics* 2009, 10: 328.

[57] M.D. Robinson, A. Oshlack, **A scaling normalization method for differential expression analysis of RNA-seq data.** *Genome Biology* 2010, 11:R25.

[58] M.D. Robinson, D.J. McCarthy, G.K. Smyth, **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.** *Bioinformatics* 2010, 26 (1): 139-140.

[59] P.J. Paddison, A.A. Caudy, E. Bernstein, G.J. Hannon, D.S. Conklin, **Short hairpin RNAs (shRNAs) induce sequence-specific silencing in mammalian cells.** *Genes Dev.* 2002, 16: 948-958.

[60] Yan Ma, Chu-Yan Chan, Ming-Liang He, **RNA interference and antiviral therapy.** *World J Gastroenterol* 2007, 13 (39): 5169-5179.

[61] S. Q. Harper, P.D. Staber, X. He, S.L. Eliason, I.H. Martins, Q. Mao, L. Yang, R.M. Kotin, H.L. Paulson, B.L. Davidson, **RNA interference improves motor and neuropathological abnormalities in a Huntington's disease mouse model.** *PNAS* 2005, 102 (16): 5820-5825.

[62] L. Zender, W. Xue, J. Zuber, C.P. Semighini, A. Krasnitz, Beicong Ma, P. Zender, et al., **An Oncogenomics-Based In Vivo RNAi Screen Identifies Tumor Suppressors in Liver Cancer.** *Cell* 2008, 135: 852-864.

[63] F. Klawonn, T. Wüstefeld, L. Zender, **Statistical Modelling for Data from Experiments with Short Hairpin RNAs.** *Advances in Intelligent Data Analysis IX- 9th International Symposium.* Lecture Notes in Computer Science 2010, 6065: 79-90.

[64] N. Hall, **Advanced sequencing technologies and their wider impact in microbiology.** *The Journal of Experimental Biology* 2007, 209: 1518-1525.

[65] Illumina Sequencing Technology report.

[66] S. Anders, W. Huber, **Differential expression analysis for sequence count** data. Available from *Nature Precedings* (2010).

[67] A.C. Cameron, P.K. Trivedi, **Book: Regression Analysis of Count Data.** *Regression Analysis of Count Data* (1998) Econometric Society Monograph No.30, Cambridge University Press.

[68] Jun Lu, J.K. Tomfohr, T.B. Kepler, **Identifying differential expression in multiple SAGE libraries: an overdispersed log-linear model approach.** *BMC Bioinformatics* 2005, 6: 165.

[69] M.D. Robinson, G.K. Smyth, **Small-sample estimation of negative binomial dispersion, with applications to SAGE data.** *Biostatistics* 2008, 9 (2): 321-332.

[70] M.D. Robinson, G.K. Smyth, **Moderated statistical tests for assessing differences in tag abundance.** *Bioinformatics* 2007, 23 (21): 2881-2887.

[71] J.C. Marioni, C.E. Mason, S.M. Mane, M. Stephens, Y. Gilad, **RNA-seq: An assessment of technical reproducibility and comparison with gene expres sion arrays.** *Genome Res* 2008, 18 (9): 1509–1517.

[72] Likun Wang, Zhixing Feng, Xi Wang, Xiaowo Wang, Xuegong Zhang, **DEGseq: an R package for identifying differentially expressed genes from RNA-seq data.** *Bioinformatics* 2010, 26 (1): 136-138.

[73] J. Panaretos, E. Xekalaki, **On Generalized Binomial and Multinomial D** 38: 223-231.

[74] B. Roos, **On the Rate of Multivariate Poisson Convergence.** *Journal of J*

[75] B. Roos, **Metric multivariate Poisson approximation of the generalized multinomial distribution.** *Teor. Veroyatnost. i Primenen.* 1998, 43: 404-413.

[76] T. J. Hardcastle, K.A. Kelly, **baySeq: Empirical Bayesian methods for i** *Bioinformatics* 2010, 11: 422.

[77] A. Mortazavi, B.A. Williams, K. McCue1, L. Schaeffer, B. Wold, **Mapping and quantifying mammalian transcriptomes by RNA-Seq.** *Nature Methods* 2008, 5: 621-628.

[78] M. Sultan, M.H. Schulz, H. Richard, A. Magen, A. Klingenhoff, M. Scherf, M. Seifert, T. Borondina, et al., **A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome.** *Science* 2008, 321 (5891): 956-960.

[79] J.H. Bullard, E.A. Purdom, K.D. Hansen, S. Dudoit, **Evaluation of Statistical Methods for Normalization and Differential Expression in mRNA-Seq Experiments.** *U.C. Berkeley Division of Biostatistics Working Paper Series*

2009, Working Paper 247.

[80] N. Cloonan, A.R. R. Forrest, G. Kolle, B.B.A. Gardiner, G.J. Faulkner, M.K. Brown, D.F. Taylor, et al., **Stem cell transcriptome profiling via massive-scale mRNA sequencing.** *Nature Methods* 2008, 5 (7): 613-619.

[81] G. Casella, R.L. Berger, **Statistical Inference.** Duxbury Press 2002.

[82] J.A. Koziol, **Comments on the rank product method for analyzing replicated experiments.** *FEBS Letters* 2010, 584: 941-944.

[83] F. Hong, R. Breitling, **A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments.** *Bioinformatics* 2008, 24 (3): 374-382.

[84] E. Hüllermeier, S. Vanderlooy **Why fuzzy decision trees are good rankers.** IEEE Transactions on Fuzzy Systems archive 2009, 17 (6): 1233-1244.

[85] A. Juan , H. Ney, **Reversing and Smoothing the Multinomial Naive Bayes Text Classifer.** In Proceedings of the 2nd Int. Workshop on Pattern Recognition in Information Systems, 2002.

[86] S.M. Kielbasa, D. Gonze, H. Herzel, **Measuring similarities between transcription factor binding sites.** *BMC Bioinformatics* 2005, 6: 237.