Analysis of Affymetrix Exon Arrays

Karin Zimmermann Ulf Leser

Humboldt Universität zu Berlin
Institut für Informatik
Wissensmanagement in der Bioinformatik

29th June 2010

Exon arrays enable the monitoring of expression on a more fine-grained level than conventional 3' arrays. By targeting single exons alternative splicing events can be detected. However, the increased amount of data resulting from the denser coverage of the transcribed regions gives rise to new challenges in data analysis compared to 3' arrays. One must carefully decide which probes are considered for the final analysis to avoid measurements that are not reflecting biological reality. The most outstanding difference between gene level and exon level analysis emerges in the detection of differential expression. To decide whether an exon is differentially expressed between two conditions it must be set in relation to its corresponding gene. Therefore, completely new algorithms need to be applied. This work gives an overview on the analysis of Affymetrix exon arrays. Technical Design, Preprocessing and the detection of alternative splicing are dicussed and finally, a complete workflow is proposed.

Contents

1 Introduction

High-throughput analysis methods are nowaday established tools for biological research . Especially gene expression microarrays have now been used for over a decade to measure the gene-wise amount of RNA transcribed in a cell [34]. The standard type of chip in such analysis are the so called 3' microarrays which try to measure the expression of genes. In contrast, the recently introduced Exon Arrays are measuring the expression of single exons. This higher resolution is only possible due to advances in spotting technology allowing for an increased number of probes per array.

Exon arrays offer a double advantage to researchers. First, when used to measure gene expression, they are more accurate than 3' arrays due to a higher coverage and more evenly distributed probes [4, 23]. Many changes in gene expression reported by 3' arrays are actually caused by the presence of different isoforms of the gene [10]. Another imprecision in the gene expression measurement with 3' arrays is rooted in the fact that some genes lack poly-A tails. For instance, some histones have never been detected in 3' based gene expression due to the fact that primers were isolating the mRNA target by their poly-A tails which their primary transcripts do not have [7].

The second advantage of exon arrays is that exon expression can be measured for the first time on only one chip. As over 90% of the human genes are alternatively spliced [42], the importance of isoform detection is obvious. Exon arrays allow for the discovery of alternative promoter usage, alternative splicing and alternative transcript termination [10]. The relevance to health related studies is given by the fact that about 30% of all alternatively spliced transcripts are disease related [43]. It is already known that isoforms of the same gene may be specific to certain diseases or even disease stages. Even isoforms with antagonistic function, such as the pro- and contraapoptotic isoform of BCL-X, have been described [40].

The advantages of exon arrays come at the cost of more complex analysis. Due to the higher number of probes, the different amount of background probes and different subsets of probes reflecting different reliability. The most significant difference between the analysis of gene expression and exon expression is the fact that the later always has to be normalized to the expression of the corresponding gene. Otherwise differences between gene expression instead of differential exon expression might be measured.

In this report we give an overview on the current state-of-the-art in preprocessing and analysis of exon array. Chapter two compares the design of the Affymetrix Exon Array to the Affymetrix HGU 133 Plus 2 chip, the probably most widely used 3' arrays. Chapter three is concerned with preprocessing methods, especially for quality control, normalization, and filtering. Section four summarizes common methods used to detect differential gene expression as they can also be applied to exon arrays. Methods for detecting alternative splicing are discussed in Chapter five. We discuss the issue of multiple testing correction in Chapter six. In the last chapter, we summarize this report by proposing a concrete analysis workflow derived from published experiences.

2 GeneChip Human Exon 1.0 ST Array

This chapter characterises the Affymetrix Exon Array technology. As many aspects are similar to 3' microarrays, we concentrate on differences in comparison to the widely used

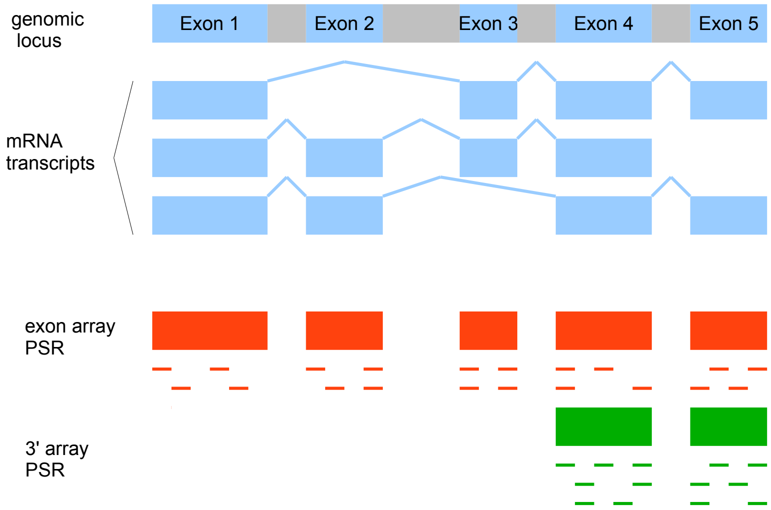

Figure 1: Probe coverage and distribution in Exon vs. 3' Arrays

HG U133 plus 2 array from Affymetrix. Major dissimilarities in the target extraction protocols, the array design and the probes on the chip are reviewed. Furthermore, the design of exon arrays raised new questions about the usage of background probes and control of the false positive rate, which also will be discussed.

2.1 Design Exon vs. 3' Arrays

The GeneChip Human Exon 1.0 ST Array from Affymetrix differs in several aspects from conventional arrays, like the Affymetrix HG U133 plus 2. One outstanding difference is the method used to generate the targets for amplification. In contrast to most 3' arrays, which use primers targeting the poly-A tail of mRNAs, the GeneChip Whole Transcript (WT) Sense Target Labelling Assay [6] applied for the exon array uses randomly attaching primers. Thus bias towards the 3' end of transcripts is avoided.

Another consequence of the different target generation protocols is the usage of mRNA as a target on 3' arrays, while DNA is hybridized to the exon arrays. Hence, the probes on the exon chip are selected to hybridize to the sense-strand, which is different from the 3' array where probes hybridize to the anti-sense strand.

The random primer technique leads to evenly distributed probes (see Figure 1). Each exon is covered by on average four probes (see Figure 2) which are later combined to one signal per exon. The higher coverage of the exon arrays requires a much higher number of probes. While, for example, the HGU 133 Plus 2 chip contains approximately 1,3 million probes, the exon array contains over 5,5 million probes (see Table 1) [4].

	3' arrays GeneChip HG U133 Plus 2.0	Exon arrays GeneChip Human Exon 1.0 ST Array
Probes per gene	$\sim 11-20$	~ 40
Probes per array	$\sim 1,300,000$	$\sim 5,500,000$
Probe sets per array	$54,000$	$1,400,000$
Background probes per array	$650,000$	$40,000$

Table 1: Comparison of probe number of HGU133Plus2 and exon arrays.

2.2 Probe types

Apart from confirmed exons, the exon array explicitely also contains uncertain and predicted exons to cover as many exons as possible. All probes on the array are divided into three categories according to their reliability [4]:

1. The first category are the *core* probes. These probes are derived from RefSeq [31] transcripts or full-length mRNAs.
2. The *extended* loci contain all core probes as well as all cDNA based loci. Among these are ESTs, mRNAs from Genbank which are not annotated as being full-length, as well as microRNA annotations [2].
3. The category *full* loci encompasses all probes from the extended loci plus loci derived from ab-initio gene predictions.

To avoid cross hybridization, sequences of all probes have been compared to each other. Sequence similarities to untranslated regions are ignored, i.e. not excluded from the set of probes, to avoid unnecessary rejection of thermodynamically favorable probes. Affymetrix classifies all probe sets into three categories according to their cross-hybridization potential:

- The about 1,25 million *unique* probe sets contain only probes that have no known potential for crosshybridization with probes of other probe sets.
- Approximately 70,000 probe sets are categorized as *similar*. These sets contain probes that are candidates for cross-hybridization, but all probes of the respective probe set interrogate the same genomic region, i.e. the same gene.
- Approximately 200,000 *mixed* probe sets exhibit inconsistent hybridizations, i.e., they might hybridize to different locations in the genome.

2.3 Background Probes

To estimate a reliable background signal, Affymetrix 3' arrays contain as many perfect match probes as mismatch probes. The mismatch probes differ in one base from the corresponding perfect match probe and are used to calculate the unspecific binding signal strength. As the exon arrays need a much larger number of probes to cover all exons, those arrays do not contain probe-specific background probes, but only a set of about 40.000 probes in total.

To account for the effects of GC-richness on hybridization strengths, background probes are binned according to their GC-content. 26 bins of different GC contents are defined, each containing approximately 1000 background probes. For each bin, a separate null distribution for probes with this GC content is calculated and used to estimate robust confidence values later. Furthermore, these 40k background probes are divided into genomic and antigenomic background probes, and either can be used to estimate a background signal. Binning differs according to wether a probe is of 'genomic' or 'antigenomic' origin:

- *Genomic background probes* match to regions that are not likely to be transcribed. To produce reasonable background probes mismatches have been introduced. Each bin of GC content from 0 to 0.25 is covered by about 1000 mismatch probes [4].

- *Antigenomic background probes* originate from sequences that are not found in the human, mouse or rat genome [4]. They are therefore not expected to cross hybridize with transcribed human DNA. The 26 bins range from a total absence of G and C nucleotides to a GC content of 100%.

2.4 Problems and Challenges

The much higher and much more fine-grained coverage of the genome achived with exon arrays does not only offer advantages, but also leads to a number of issues that need to be clarified during experiment design and data anaysis.

First, it must be decided which probes should be used for analysis: Only the more reliable (*core*) probes or also the more experimental *full* set of probes. This decision must be taken depending on the aim of the study. If the researcher is interested in reliable and well annotated data only, only core probes should be used for analysis. It has been shown that the signal produced by the extended and full probes does not correlate well with the core probe signals [44]. If, on the other hand, the study has more of an exploratory or experimental character, it might be valuable to also use the extended and full probes. In this scenario analysis results have to be interpreted with greater care.

Second, it must be decided how to compute background signals. As exon arrays do not have a mismatch probe for every perfect match an alternative background calculation model must be used (see above).

Third, as exon arrays contain a much higher number of features than conventional microarrays, data analysis implies a much higher number of statistical tests that are evaluated in parallel. This in turn may lead to a higher number of false positives. To avoid this phenomenon different strategies like filtering or intersecting results of different methods are used (see Chapter 4).

3 Preprocessing of Exon Arrays

As for all high-throughput techniques, preprocessing of the raw measurements of exon arrays is essential to reduce the effects of non-biological variations across experiments

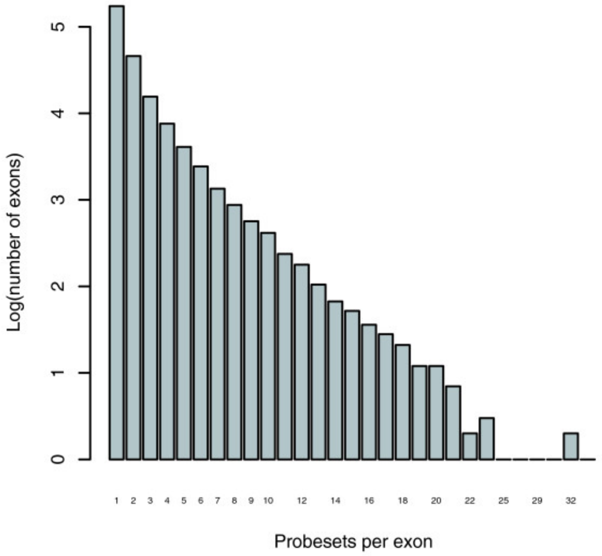

Figure 2: Histogram of the number of probe sets per exon [28].

[21]. Such preprocessing usually consists of a number of steps. First, the quality of the data is determined and potential outlier samples are removed. Second, values are normalized to increase comparability of values across samples. Third, all probe values within a probe set are summarized into a single value representing the expression of an exon. Finally, probes sets that are not likely to be differentially expressed are removed in a filtering step. Overall, precprocessing comes a long with a significant redution in data dimensionality which also reduces the probability of generating false positives during successive analysis.

3.1 Quality Control

Evaluation of the quality of the data being analyzed is crucial to the reliability of the findings concluded from the analysis [13]. Samples of low quality should be excluded from further analysis. As different methods for quality control focus on different aspects of the data, the application of several of them is recommended.

- **Hierarchical Clustering** groups similar elements together. It can also be used for quality control. Samples that are far away from the majority of the other samples show a bias potentially introduced for technical rather than biological reasons. If such bias is not removed by normalization these samples should be excluded from further analysis.

- **Array-Array Intensity Correlation** displays the correlation of all pairwise samples encoded as colors, scatterplots or numerical values. This method immediately reveals outliers, i.e. samples whose probe signals do not correlate at a certain threshold.

- **Boxplots** help to compare the intensity value distributions of the arrays by plotting various measures from descriptive statistics, like the median or different quartiles. Boxplots of samples that differ significantly in the median value or the hight of the box indicate non-comparable intensity value distributions.

- **Intensity distributions** of different arrays can depict major variations in the expression values. If the density curves differ much after normalization, the values of probes from different samples are not comparable.

- **Principal Component Analysis (PCA)** transforms the data from a high-dimensional space into a two or three dimensional one preserving as much of the variance as possible. This dimensionality reduction allows the visual inspection of high-dimensional data without losing too much of the variation in the original values. As in hierarchical clustering, samples that are far away from most of the other sample are considered outliers. In comparison to hierarchical clustering PCA preserves more information, but is also computationally more demanding.

- **Relative Log Expression (RLE)** is a model-based technique to identify samples with an unusual high amount of differential expression. As differential expression is thought to be a rare event, such samples usually are excluded from further analysis. RLE values are derived from a probe-level model. This is achieved by calculating the log ratio of every probe set on every array and the median expression of the same probe set from all other arrays. A boxplot visualizes the computed values for each array, which should be centered around zero. As for other boxplots deviating box height indicates potential outliers.

- The **Normalized Unscaled Standard Error (NUSE)** plot is calculated by using standard error estimates derived from a PLM fit. Normalization of theses values efectuates that the probe set medians of the error estimates equal one on every array. If these values are visualized in a boxplot per array, deviations in centering or unequalities in box height hint to outlier samples.

- **RNA degradation plot:** Natural RNA degradation starts from the 5' end of the RNA. By ordering the probes of each array according to their distance to the 5' end and plotting their average intensity values for every array, the degree of RNA degradation is visualized. Probes should display a lower average intensity the closer they are located near the 5' end of the RNA. The slope of the plot is proportional to the degradation effect and should be roughly the same for all samples. Outlier arrays can be detected by having a different slope than other arrays in the set.

3.2 Summarization and Normalization

In order to produce comparable values as a basis for further analysis normalization and summarization steps are performed. Normalization aims at removing technical bias, while summarization stabilizes the values through averageing.

The two most common methods for Affymetrix 3' GeneChips are RMA and PLIER. As they do not (necessarily) make use of mismatch probes, they also can be applied directly to exon arrays. Furthermore, we describe other methods that have been developed specifically for exon arrays.

3.2.1 RMA

Robust Multiarray Average (RMA) [21] is one of the widest used normalization methods when working with Affymetrix GeneChip data. RMA is popular with gene chips as well as with exon arrays. As opposed to other methods like MAS5.0 [20, 8] it only uses the Perfect Match (PM) probes on the chip. Basically, RMA consists of four steps: background correction, taking the logarithm, quantile normalization and summarization.

First, the values measured for every PM probe on every chip are background corrected in a way that no value is negative. The background corrected values are log2 transformed and normalized by quantile normalization. Quantile normalization ranks the values on every chip by their intensity. For every rank the average over all chips is computed. The actual values of this rank on every chip are then replaced by this average. To summarize the values of one probe set a linear model accounting for the probe affinity effect, the log scale expression level for the array and an error term are fitted. Median polish ensures robust parameter estimation.

3.2.2 PLIER

Probe Logarithmic Intensity Error (PLIER) [5] is another model-based signal estimator. Two main points differentiate PLIER from other methods: the accession of probe quality by measuring the feature response, and an error model accounting for different major components of intensity in the low and high intensity range.

The so-called *feature response* is a scaling factor accounting for systematic differences between features of a probe set. Once such a feature response factor is determined, non systematic differences in feature response can be identified. According to the quality of their response, features are up- or downweighted in their contribution to the final probe set signal. If, for example, a feature shows always twice the response intensity as the common features in the probe set while a second feature shows twice the intensity in the high but the same intensity in the low intensity range, after scaling the former should contribute more to the probe set signal than the latter. Scaling factors are computed by using all arrays in the set.

The second improvement in signal summarization is a different error model. In general, it is assumed that the error is proportional to the biggest component contributing to the observed signal. However, this component can be different in the low intensity range compared to the high intensity range, which should be accounted for in the error model. In high intensity ranges, the error is proportional to the background corrected intensities as the main signal is assumed to represent the actual target response. But in the low intensity ranges, background is the main source of signal, and accordingly the error should not be estimated on background corrected signals. The PLIER error model accounts for this different error dependencies by transitioning between the two [5, 39].

PLIER is only a summarization method. Usually quantile normalization is applied before summarization. Furthermore PLIER does not include variance stabilization and produces values close to zero. For all values the addition of 16 is recommended before

[5] taking the logarithm of the summarized values.

3.2.3 iterPlier

This variant of PLIER differs from the original method by iteratively discarding probes considered less representative for the main signal [3]. To achieve this, first PLIER is applied to all probes of a probe set. Only the 22 probes best correlating with the calculated PLIER signal are used to rerun PLIER. Next, the 11 probes correlating best with the newly computed PLIER signal are used to compute the final signal by applying the procedure a last time. If the probe set has only 11 or less probes, PLIER is run only once. This corresponds to the original PLIER signal.

3.2.4 GC Dependent Based Bias

A simple method used to remove the bias induced by the GC content of the probes is described by Numata et al. [24]. The basic idea is to categorize probes based on their GC content and remove the bias of each probe according to the category assigned (recall that this observation is also the basis for the selection of background probes in the exon arrays, see Section 2.3). First, every probe is assigned to a GC category. For every GC category the median value is computed. This median value is now subtracted from every probe that contributed to its computation. A variant of this approach is to compute the median for each GC category based on the antigenomic background probes also divided in GC categories instead of using the actual probes (again, see Section 2.3).

3.2.5 Model based Analysis of Tiling arrays (MAT)

Recently, a number of probe selection algorithms have been developed which try to identify and remove poorly performing probes [44]. This is especially important and promising for exon arrays, given the extremely high number of probes on these arrays. In MAT, probes are used as an estimate for gene expression which should highly correlate between the samples [22]. To remove 'absent' probes, the intensity of each probe is compared to the one predicted by a MAT background model. Only probes significantly different from background are kept.

3.2.6 Probe Selection by Correlation

Another method, called probe selection by correlation, for selecting representative and trustworthy probes is described in [44]. This algorithm selects a subset of highly correlated probes for each gene by applying hierarchical clustering to the probes in the probe set over all samples. Only those probes are used for the signal calculation that are highly similar in their expression across samples.

3.3 Other Probe Filtering methods

The previous two approaches try to identify probes that behave differently from the other probes in the same probe set. However, probe selection can also be performed based on different criteria.

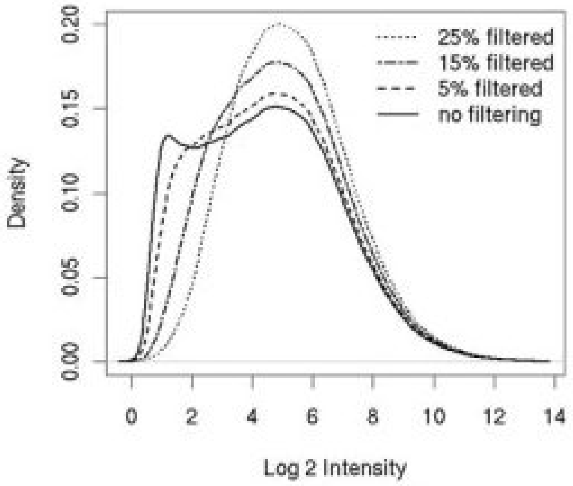

Figure 3: Intensity distribution after multiple filtering steps [37].

One technique is directly based on the sequence of probes. It consists of filtering all probes that can potentially cross hybridize, i.e. those that have multiple matches in the genome or contain one or more mismatching base [37], [17]. Recall that probes on the exon array are preclassified according to this criterium (see Section 3.2).

Another selection strategy is to find probes whose intensity is not considered above background. [37] suggestes to remove all probes ranging in the lowest intensity quartile of all samples, as they can be considered as not distinguishable from background. Following a similar idea, but using a more elaborate algorithm, the Detection Above Background (DABG) method assigns a p-value to every probe indicating if the probe is likely to have a expression level above background. The null distribution is calculated from a randomly chosen set of probes with the same GC content. [17] suggests to keep only probes with a p-value of less than 0.05.

Probes that display a great sample-to-sample variation may also be excluded from the analysis as they will potentially not produce a coherent signal among the classes [37]. The hope is that the observed signal distribution of an exon array converges more and more to the expected (real) signal distribution with every filtering step (see Figure 3).

3.4 Comparison of Different Preprocessing Procedures

The evaluation and comparison of normalization and summarization methods on real data is non-trivial, as there nothing is known about the "true" values that should be reproduced by those methods. Accordingly, only few works exist that compare different methods on the same data set. Johnson et al. compared MAT to the GC-binning method for background correction proposed by Affymetrix and report superior performance for MAT [22]. [17] and [29] applied RMA and PLIER, but report only marginal differences.

A way to get around the missing gold standard is to compare normalization and summarization by their effect on the detection of differential expression. [25] conducted such a comparison, eventhough with a different purpose. The authors applied different methods for detection of differential expression to datasets that were normalized with either RMA or PLIER. By using a gold standard set of differentially expressed genes, ROC curves were constructed for combinations of normalization and differential detection methods. The study showed that combinations using RMA in most of the cases performed better than the ones with PLIER.

A recent study [33] comparing the performance of RMA and PLIER on exon array data lead to a favoritism of RMA. PLIER seems to misestimate the gene-level signal and therefor in- and excludes exons not in consistency with the TISA (Tissue specific alternative splicing) database [27].

4 Differential Expression

After preprocessing and quality filtering, the most important question studied with exon arrays is which genes or exons are expressed differentially between two conditions. Given the high resolution of exon arrays, it is especially interesting to find isoforms that are expressed differently in the different conditions. It is important to precisely differentiate between these tasks:

- At the **gene level**, any method for detecing differential expression developed for 3' gene arrays can be applied. We will shortly review such techniques in Chapter 4.1. The only additional step necessary is the aggregation of exon intensities to gene intensities (see Chapter 4.1).

- On the **exon level**, the situation is more intricate and poses a challenge that does not exist for 3' gene arrays. Here, the aim is to detect exons that are included in a gene, i.e. a transcript of the gene in one condition whereas the exon is not included in the transcript present in the other condition. Imagine, differential exon expression would be computed analogously to differential gene expression, i.e. simply applying a t-test to the expression values of an exon in different conditions. If the gene containing the exon would be differentially expressed between the two conditions all exons of this gene would be deemed differentially expressed which is not what we are looking for. Consider the example given in Figure 4. Here, obviously exon2 is not contained in the transkript of tissue A and therefor differentially expressed between tissue A and tissue B. Nevertheless, this exon would be the only one *not* determined differentially expressed by the naive approach just sketched.

  The essential step in exon level analysis is therefore to normalize the exon level signal to the corresponding gene level signal to determine true differential exon expression.

Identifying different isoforms of a gene in different conditions requires that exon specific expression values are analyzed relative to the expression of the according gene (which must be determined first). This question usually is studied under the term *differential exon expression*; this means, that differential exon expression is defined as a different amount of expression of an exon in one or more conditions where the expression value has been set in relation to the expression value of the respective gene.

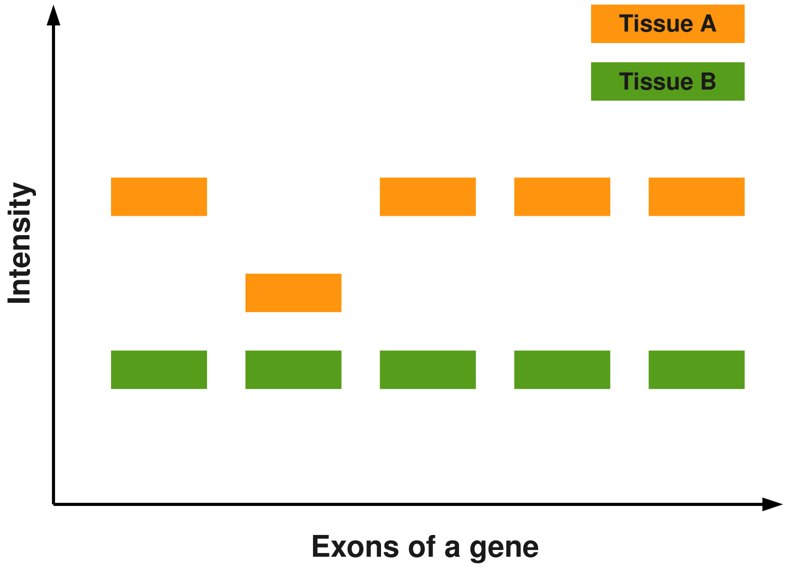

Figure 4: Different isoforms in two conditions. The expression intensity of a gene can be different between the conditions of interest. To this end, the expression intensities of each exon must be seen in relation to the expression levels of its corresponding gene. If exons are only analzyed in isolation, all but the second exon would be deemed differentially expressed. However, the more interesting fact is that only the second exon in tissue A is spliced out in comparison to tissue B.

A very simple approach to detect differential exon expression chooses only genes that do not show a significant change on the gene expression level while indicating significant changes on the exon expression level [10]. A drawback of this method is that it misses cases as depicted in Figure 4 where genes are expressed at different levels *and* display alternative splicing.

4.1 Gene level Analysis

To perform gene level analysis on exon array data a signal intensity for every gene/transcript has to be computed. A common approach is to aggregate the gene signal over all exon signals associated to it. Different aggregation functions, like mean or median, can be applied.

Once the gene level expression signal is determined, the analysis for differential gene expression on exon arrays is reduced to the analysis of differential gene expression of 3' arrays measuring the amount of mRNA per gene. The *t-test* is probably one of the most frequently used statistical tests for the determination of differences between two conditions. For more than two classes, *ANOVA* can be applied [36]. Further, non-parametric tests like *Significance analysis of microarrays (SAM)* [41] and *Rank Produkt* [14] are also used frequently because they do not make certain assumptions suchas a normal distribution or equality of variance. For more details, see [38, 36].

4.2 Exon level Analysis

Many different approaches have been developed for the analysis of differential exon expression. We present a selection of them in the following. Some methods like the Splicing Index, FIRMA, PAC and the correlation based approach compute scores indicating the probability of alternative splicing while MIDAS, ANOSVA and MADS use statistical tests. ARH implements an information theoretical approach. All these methods are described in the following subsections. As the number of false positives is a common problem in this field and the methods in use have different strengths and weaknesses, many papers suggested intersecting or combining different methods. Della Beffa et al. [17] suggest intersecting the results from MIDAS and Rank Product, while Shah et al. [37] propose to use a statistical test and a correlation measure.

4.2.1 Splicing Index (SI)

The splicing index [1] can be regarded as the exon level analog of the fold change. It is a measure for the difference of exon specific expression between samples. Let e be an exon part of a gene g in an experiment with m samples.

As a first step, the expression level of each exon e in sample k is normalized to its corresponding gene expression g.

$$\hat{e_k} = \frac{expr(e_k)}{expr(g_k)} \tag{1}$$

Based on these normalized measures, the Splicing Index $SI(e)$ measures the relative change between two conditions:

$$SI(e) = \frac{\hat{e}_{k_2}}{\hat{e}_{k_1}} \tag{2}$$

In the case of sets of samples multiple procedures are applicable. On the one hand the ratio of the medians of each set can be calculated. One can also compute the SI for all paired samples from two groups and take the median as an indicator of alternative splicing. The SI can only be applied in a two-condition scenario.

4.2.2 Pattern Based Correlation (PAC)

The basic idea of PAC [1] is that the expression level of an exon is proportional to the expression level of the gene it belongs to over all samples as long as no alternative splicing event is taking place. PAC therefore first computes an expected expression level per exon. Again, let exon e be part of gene g. Let β be the average of all exon expression / gene expression ratios over all samples and all exons of g. Then, the expected expression level e' of e in sample k is computed as:

$$e'_k = \beta \cdot expr(g_k) \tag{3}$$

Next, the correlation between the expression estimates and the actually measured expressions of all exons of a gene are computed. A low correlation indicates alternative splicing.

A drawback of this method is that it is not applicable to the comparison of two groups of samples, as the exon expression levels of the two classes may either correlate positively in direction (+1) or negatively (-1).

4.2.3 Microarray Detection of Alternative Splicing (MIDAS)

The idea of MIDAS [1] is similar to the idea of SI, but follows a more sophisticated approach. Like in other approaches, it assumes that if the ratio of an the expression level of an exon to the expression level of the corresponding gene is constant over all samples, then no differential exon expression has taken place. This assumption is formulated as null hypothesis. Next, for each exon the exon / gene expression ratios for all samples are computed and their variance is tested for statistical significance using, for instance, ANOVA.

4.2.4 MIDAS and Rank Product intersect

One of the main problems with statistical tests in application to exon array data (as used by MIDAS) is the high number of false positives. To reduce the false positive rate usually multiple testing correction procedures are applied (see also Section 5. Della Beffa et al. [17] argue that Bonferroni correction for multiple testing is too strict while FDR controlling methods like Benjamini-Hochberg can not be applied because of the non-uniformity of the raw p-value distributing produced by MIDAS. Also the independence of the tests, a common requirement for multiple testing correction methods, is not given when testing multiple exons of one gene.

Instead of using multiple testing correction, Della Beffa et al. propose to decrease the number of false positives by applying two methods, Rank Product [14] and MIDAS [1], and intersect their results. Rank Product is a simple, biologically motivated non-parametrical test. In this approach the genes in a sample are sorted decreasingly by expression values. The 'score for a gene is calculated by determining the geometric mean of its ranks, while significance is assessed by a permutation of the gene names.

4.2.5 Microarray analysis of differential splicing (MADS)

The MADS method [45] consists of three steps: background correction, summarization and detection of differential splicing events. For background correction, a sequence-specific linear model with 80 parameters is fit to predict the background intensity for each probe. The predicted background intensity is then subtracted from the observed signal. Genomic as well as antigenomic background probes are used to train the model. The specialty of this background model is a nucleotide and position specific model for the 25mer probes.

In the second step the probes with the highest correlation over all samples are selected for each gene by application of hierarchical clustering. The Li-Wong [26] model is fitted to these probes to compute an estimate of gene expression. Only probes are kept that show high correlation between the background corrected values and the corresponding gene signal estimates over all samples. Similar to iterPLIER, this procedure is repeated until the number of probes stabilizes.

To determine differential splicing, first the Splicing Index is calculated for every probe. A t-test is applied to determine the significance of the calculated Splicing Indices. The probe level p-values are transformed by $x = -2log(p)$ and are added up to a probe set p-value. The probe set p-values are then used to rank the probe sets, i.e. exons. The final results are filtered for potentially cross hybridizing probes.

4.2.6 Correlation and SI/LIMMA

In this approach the authors combine two methods for optimal detection of differentially expressed exons. On the one hand they compute the SI followed by an application of the moderated t-test implemented in the R/Bioconductor [18] package LIMMA (reffered to as LIMMA by the authors). On the other hand, they simply compute the correlation of the expression values of for exons of a gene/transcript in different conditions.

As shown by Shah and Pallas [37] the application of $SI/LIMMA$ yields good results if only one or few exons were alternatively spliced in a gene, but it produced a very low rank in alternative splicing to genes where many exons were alternatively spliced. Besides, this approach produces a high number of false positives. Hence extensive pre-filtering of the data (up to 25 %) is required.

The two following observations lead to using correlations between expression patterns of the exons of one gene between different samples or groups of samples [37] for the detection of alternative splicing. The first one is that the expression levels of exons in one gene vary but are similar in all samples as long as no alternative splicing took place (see Figure 5). The second observation is that approaches based on gene expression level estimates, like $SI/LIMMA$, tend to lead to false negatives if many exons are differentially spliced in one group but not in the other. The reason for this misclassification becomes clear from an example. Suppose experiments with two tissue types were performed. If many exons of one gene are differentially expressed, i.e. their expression level is higher in the one tissue than in the other, the expression estimate for the corresponding gene will be misestimated. Therefore the normalized exon expression signal calculated will be misleading or even false.

On the other hand, the correlation approach favors genes containing many alternative splicing events. Therefor the two methods are complementary and may be used both for optimal results.

To determine correlation all genes containing only one or two core probes are removed as their is no way to compute correlation for one probe and the correlation of two probes always results in 1 or -1. Genes are deemed likely to undergo differential splicing if they showed a correlation less than 0.8 and a multiple testing corrected p-value less than 0.05. The cutoff of 0.8 for correlation was chosen based on the observation, that 94% of the genes determined not tu untergo differential splicing by $SI/LIMMA$ had a correlation coefficient greater than 0.8.

4.2.7 Finding Isoforms using Robust Multichip Analysis (FIRMA)

FIRMA [32], another method for detecting differential splicing, has the major advantage that it also can be used if there are no predefined groups or if alternative splicing events are not consistent in the given conditions. The basic idea is to use a fitted linear model for expression estimation and deduce a score for alternative splicing for each exon from the model parameters. As the name suggests, Robust Multichip Analysis (RMA) is used to determine these parameters. The fitting of the linear model for expression estimates leads to the possibility to compute the difference between estimated and measured expression. This difference is taken as basis for the computation of a score for differential exon expression. By doing so the problem of alternative splicing detection essentially is converted to one of outlier detection. FIRMA is implemented in the aroma.affymetrix package of Bioconductor.

The model fitted by RMA for every gene g contains a chip effect term c_k for the kth chip, a probe effect term p_i for the ith probe and an error term ϵ_{ki}. An estimate for the background corrected and normalized expression level of a gene is computed as follows:

$$log_2(expr(g_{k,i})) = c_k + p_i + \epsilon_{ki} \tag{4}$$

For exon arrays the model can be adjusted by introducing $expr(e_j)$, the exon expression for exon j, the interaction d_{kj} between chip and exon as well as a new error term $\epsilon_{kji(j)}$.

$$log_2((expr(e_{k,j,i})) = c_k + expr(e_j) + d_{kj} + p_i + \epsilon_{kji} \tag{5}$$

As the parameter d_{kj} represents the difference between an exon in sample k and the expected expression for this exon, the parameter can be seen as a measure of differential splicing. Instead of fitting the exon level model, the gene level model is fitted with the exon array data to improve robustness. d_{kj} is then estimated by using the residuals r_{kji} of the fitted gene level model.

$$r_{kji} = y_{kji} - c_k - p_i \tag{6}$$

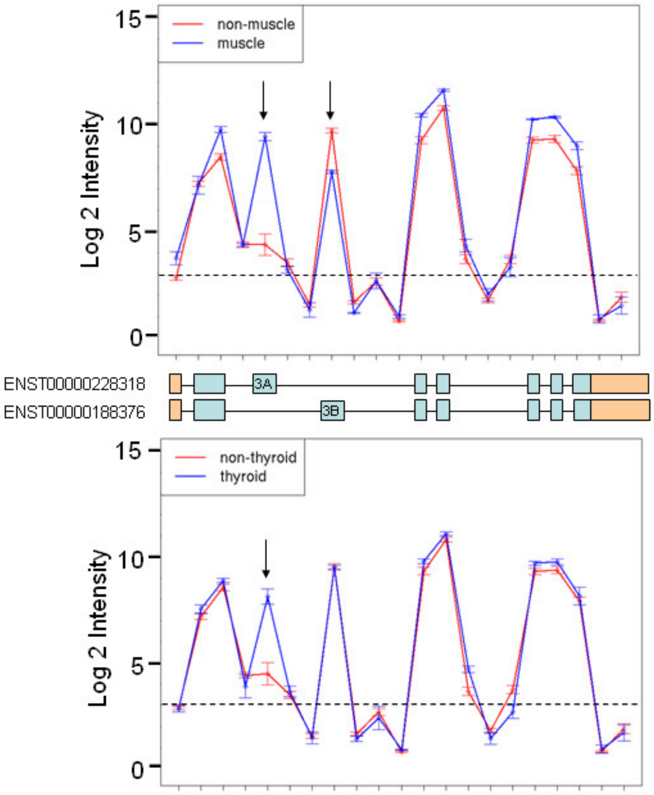

Figure 5: Correlation of exon expression patterns of one gene between two samples [37]. The expression profiles shown for the same gene in 'muscle, non-muscle' (upper Subfigure) and 'thyrois, non-thyroid' (lower Subfigure) show alternative splicing (AS). The AS events are marked by arrows. Both muscle and thyroid include the second exon oposed to their non-muscle and non-thyroid counterparts. The second arrow in the upper Subfigure marks the skipping of this exon in the muscle related transkript.

The actual value is computed by averaging over all four residuals in an exon.
The final score for alternative splicing is now calculated as F_{kj},

$$F_{kj} = median_{k \in exonj} \frac{r_{kji}}{s} \tag{7}$$

where s, an estimate of the standard error, is derived from the median absolute deviation (MAD) of the residuals. By introducing this parameter the score is made more comparable between different genes.

4.2.8 Alternative Splicing robust prediction method based on entropy (ARH)

ARH [35] is one of the most recent methods in this field. Unlike other approaches it is not based on correlation or statistical tests, but applies an information theoretic approach based on Shannon's entropy. An advantage of this method is that it overcomes problems like the dependency of a score on the number of exons or the inherent variability in exon expression intensities. A drawback of the method is that it can be applied only in the case of two different conditions.

ARH proceeds in two steps. In the first step, spliced genes are identified, while in the second step the exons of a gene are ranked by their likelihood of having undergone differential expression. As usual, let gene g have m exons, let k_1 and k_2 denote condition one and condition two, respectively, and let $expr(e_k 1)$ and $expr(e_k 2)$ denote the expression signal of exon e in the two conditions.

First, the exon splicing deviation e' between the two conditions is computed by subtracting the median log_2 ratio of the exon expressions from all median log_2 ratio exon expressions of the associated gene.

$$e' = log_2(\frac{expr(e_{k_1})}{expr(e_{k_2})}) - median_{i=1,\ldots,m} log_2(\frac{expr(e_{k_1})}{expr(e_{k_2})}) \tag{8}$$

The absolute value of the splicing deviation is turned into the probability $p(e)$ of exon e being differentially spliced.

$$p(e) = \frac{2^{|e'|}}{\sum_{i=1,\ldots,m} 2^{|e'||}} \tag{9}$$

Next, an entropy is computed for each gene, indicating whether the splicing probabilities are equally distributed.

$$H(p(e_1),\ldots,p(e_m)) = -\sum_{i=1}^{m} p(e_i) \cdot log_2(p(e_i)) \tag{10}$$

The theoretical maximum $max(H)$ of the entropy is $log_2(m)$. To make the entropy independent of the number of exons, it is subtracted from the theoretical maximum.

$$max(H) - H = log_2(m) - H(p(e_1),\ldots,p(e_m)) \tag{11}$$

Now the final score ARH for gene g can be computed to indicate whether g is alternatively spliced.

$$ARH = \frac{Q_{75}}{Q_{25}} \cdot (max(H) - H) \tag{12}$$

The weighting factor $\frac{Q_{0.75}}{Q_{0.25}}$ accounts for the strength of deviation within g. Here, Q_{xx} denotes the interquartile range of expression range of the xxth quartile.

By the computation of a background distribution for ARH values from various datasets, values larger than 0.03 are considered to be an indication for alternative splicing by the authors.

4.2.9 Two way ANOVA (ANOSVA)

The two way analysis of variance (ANOSVA) [15] can be used also for the detection of alternative splicing [24]. However, according to Affymetrix the method does not perform well on exon array data [1]. Let $\hat{e}$ denote the background corrected probe intensities for exon e of gene g in probe k. $\hat{e}$ is modelled as follows:

$$e\hat{=}\mu + \alpha_e + \beta_j + \gamma_{ej} + \epsilon_{ejk} \tag{13}$$

The index j denotes the different sample conditions, i.e. $j = 0$ denotes normal while $j = 1$ denotes diseased tissue. The parameter α_e accounts for the basic expression intensity of exon e. This is necessary as the same splice variation can occur in both conditions. Bacause the two conditions can have different average expression values per gene, the parameter β_j is included in the model to account for this. The term capturing differential splicing is γ_{ej}. It denotes interaction effects of the different exons (e) with the different conditions (j) . The basic signal is modelled by μ, the overall mean of the probe intensities. Finally, ϵ_{ejk} denotes a general error term with mean zero and standard deviation σ.

By applying statistical tests to different parameters of the model, significant differences in gene expression (β_j) and exon expression (γ_{ej}) can be detected. By testing α_e, general splicing variants that are present in all conditions can be found.

4.2.10 Probe Level Expression Change Averaging-Splicing Index (PECA-SI)

The PECA-SI [25] does not summarize the values of a probe set before calculating the splicing index. This approach determines the SI for every probe and computes a final exon-level SI by averaging over the probe-level SI's.

As shown by the authors this method performs well on synthetic data, but no results on real data have been published yet. When comparing the method to other common methods for the detection of differential splicing in combinations with summarization methods like RMA and PLIER, two interesting findings can be observed. First, PECA-SI performs best in most of the tested scenarios. Second, the combination of RMA with a method for detection of differential splicing outperforms in most scenarios combinations of PLIER with the respective method. The different scenarios were created by simulating data with different noise levels and different numbers of alternative spliced exons. Verification is performed by measuring the area under the ROC curve.

4.3 Isoforms

We presented various methods that detect differnetially expressed isoforms between different conditions. However, determining the actual isoforms present in the conditions is non-trivial. The problem can be devided into three subtasks. (1) To predict the set of different transcripts (known and unknown) present in the samples. (2) To predict the structure of these transcripts. (3) To quantify the relative concentrations of the predicted isoforms in the different samples.

SPACE is an algorithm solving these three problems [9]. Essentially, it performs non-negative matrix factorization and thereby produces two matrices. One of them can be interpreted as representing the relative concentration of each transcript, while the other one contains the transcript structure. The number of the isoforms of a gene is estimated by the internal factorization dimension.

SPACE has some characteristics that should be considered when applying the algorithm. First, SPACE assumes the probe signals to be proportional to the number of transcripts. If this is not the case, incorrect predictions of structure and quantity of a transcript can result. Second, although the algorithm was developed for exon junction array and exon array data, it seems to work better for exon junction array data. A third limitation is that SPACE is not able to predict different transcripts if only one experimental condition is used.

5 Correction for Multiple Testing

The more statistical tests are applied, the higher is the total expected number of false positives [16]. For instance, measuring differential expression in expon arrays on a probe level implies the execution of 1,4 million tests in parallel. With the (typical) significance level of 0.05, the expected number of false positives accordingly is 70 000, as 5% of the features are expected to be determined differentially expressed by pure chance. Clearly, this number is unacceptable, calling for a adaptation or correction of the significance thresholds of individual tests depending on the number of statistical tests performed in parallel. Such a correction should reduce the expected number of false positives without increasing the expected number of false negatives too much.

Methods for performing multiple testing cirrectoin can be devided into two classes: (1) Methods controlling the false discovery rate (FDR), and (2) methods controlling the family wise error rate (FWER). In general, methods controlling the FDR tend to lead to more false positives while methods controlling the FWER lead to more false negatives as they are much stricter in correction. All approached first perform individual tests, but then post-process the resulting p-values.

Benjamini-Hochberg [11] is a method for controlling the FDR. The FDR is the expected proportion of true null hypotheses rejected in the total number of rejections. Thus, FDR measures the expected proportion of incorrectly rejected null hypotheses, i.e. type I errors. Let V denote the number of null hypotheses that are rejected in multiple testing procedure and W the number of true null hypotheses rejected. The FDR is defined as:

$$FDR = E(T) \tag{14}$$

with

$$T = \begin{cases} W/V & \text{für} \quad V > 0 \\ 0 & \text{für} \quad V = 0 \end{cases} \tag{15}$$

To achieve a predefined false discovery rate $FDR = \alpha$, first the original p-values are sorted in ascending order, giving $p_1...p_n$. Next, the first position $i \in 1...n$ ist determined which fullfilles the equation:

$$p_i \leq \frac{i}{n}\alpha \tag{16}$$

Only p-values $p_1...p_i$ are considered as significant, i.e. the nullhypothesis for the tests 1 to i is rejected.

Bonferroni [12] is amethod for controlling the FWER. The FWER is the probability of having at least one false positive in the set of results considered as significant. With Bonferroni, every original p-value is multiplied by the number of statistical tests performed in parallel, i.e. the number of genes tested for differential expression. With N the number of genes tested and p the p-value of a given probe, one computes an adjusted p-value using:

$$p_{adjusted} = pN \tag{17}$$

Only if the adjusted p-value is smaller than the pre-chosen significance value, the probe is considered as differentially expressed.

6 Proposed Final Workflow

This work is part of the DFG-funded Transregio TRR54[1]. The general aim of the TRR54 is the exploration of various forms of lymphoma. Within the project, a large number of exon arrays will be measured for a series of different lymphoma subtypes. In the following, we suggest a workflow for the analysis of these arrays (see Fig.6). The purpose of the analysis workflow is twofold: First, we aim at detecting differentially expressed genes with subsequent functional analysis. Second, we will search for subtype-specific isoforms by studying differentially spliced exons.

Our analysis will concentrate on the core probes, as full and extended probes tend to reduce the quality of the data as the amount of random signal is potentially higher. Furthermore, it has been shown that the signal produced by the extended and full probes does not correlate well with the core probe signals [44].

6.1 Gene level analysis

To assess the perfomance of the data different quality control plots, in particular array-array correlation plot, boxplots of the raw intensities, and NUSE and RLE boxplots, are produced [19]. Note that quality control will be performed using only the core probes. Now, the data is normalized and summarized using RMA. Again quality control plots are generated, samples behaving peculiar after normalization are excluded from the data set.

To detect differentially expressed genes the moderated t-test is applied to the gene summary data. The correction for multiple testing is performed with Benjamini-Hochberg. Clustering of the resulting genes is applied to detect patterns inherent to tha data. On the gene level analysis genes should be excluded that are also present in the candidate set for differential splicing of the exon level analysis. These changes in gene expression could be due to different isoforms [10]. Finally, functional analysis is applied, i.e. the differentially expressed genes are tested for enriched pathways and GO terms.

6.2 Exon level analysis

The data is normalized and summarized using RMA, producing a signal for every exon. Again the quality control plots are applied to detect outlier samples. As the gene signal

[1] See http://www.trr54.de/

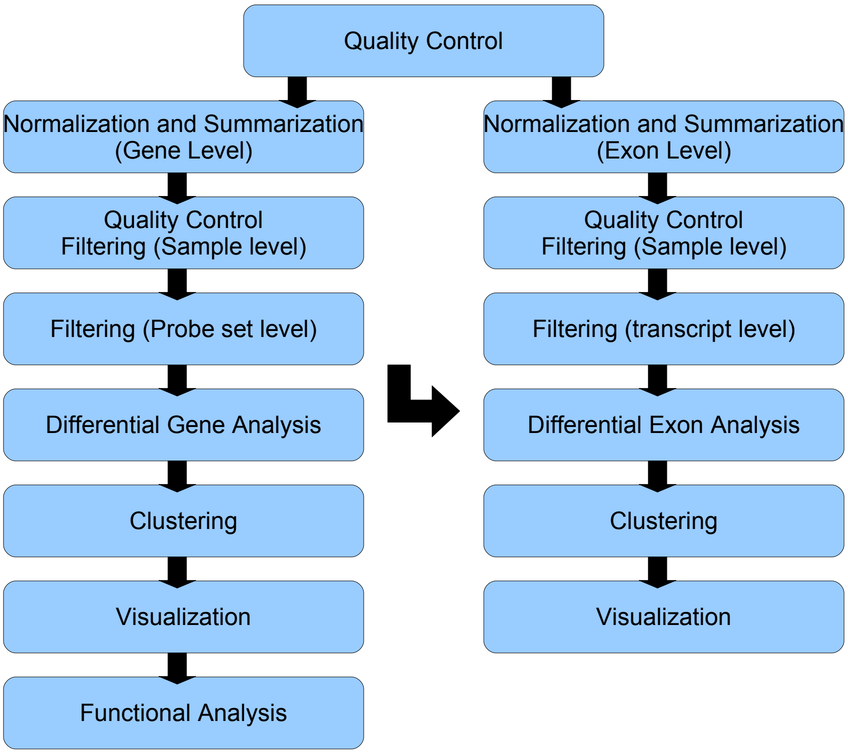
Quality Control
Normalization and Summarization (Gene Level)
Quality Control Filtering (Sample level)
Filtering (Probe set level)
Differential Gene Analysis
Clustering
Visualization
Functional Analysis
Normalization and Summarization (Exon Level)
Quality Control Filtering (Sample level)
Filtering (transcript level)
Differential Exon Analysis
Clustering
Visualization

Figure 6: Workflow. To determine the quality of the data and to detect outlier samples quality control plots are applied. Before normalization NUSE and PLIER while after normalization Array-Array correlation plots, boxplots and PCA are performed. Samples showing poor quality are filtered out in both data sets (gene and exon level). Another filtering step is done on the exon (probeset) and gene (transcript) level to avoid non-expressed as well as cross-hybridizing probe sets. Differential expression is determined with SI/MIDAS on the exon level and a moderated t-test on the gene level. Benjamini-Hochberg multiple testing correction is applied in both cases. Clustering and visualization of the results is subsequently performed. On the gene level funktional analysis detects GO term and pathway enrichment.

estimate is needed for the calculation of differential exon expression the same sample should be included in the gene level as in the exon level data set. Note that the quality control plots before normalization are the same as in the gene level data set.

On the probe set level filtering is applied as well. Following suggestions in [30, 40] a DAGB (see 3.3) p-value less than 0.05 in at least half of the samples is required to keep the probe sets for further analysis. The exon level data set should contain only exons of genes that are expressed above a certain threshold. Otherwise differential splicing detection methods tend to lead to more false positives. Differential splicing can only take place if the gene is expressed in both conditions. As alternative splicing can only occur in genes containing at least two exons all genes containing only one exon are filtered out. This additionaly reduces the number of statistical tests as about 7% of the genes are removed [46]. Probe sets with the potential of cross-hybridization are removed as well.

Subsequently, the detection of differentially expressed exons is performed. Therefore a SI for every exon is computed. Also MIDAS is performed to obtain a p-value for every exon. In a next step multiple testing correction is applied using the Benjamini Hochberg method. We consider exons to be differentially spliced if they have a p-value less than 0.05 and a gene-normalized fold change (log-ratio) between the different conditions of at least 0.5 [40]. The reason for this relatively small fold change is that alternative splicing events often occur in only a part of the samples of one condition. Hence, fold changes are expected to be less high than in differential gene expression.

To detect potential subclasses clustering of the differentially spliced exons is applied.

The isoforms significantly associated with the different conditions are then determined by applying the SPACE algorithm (see Section 4.3) to all genes/transcripts showing differential exon expression. An important step to reduce false positives is the vizualization of the candidate set for alternative splicing [37].

7 Acknowledgement

We would like to thank Dr. Michael Hummel and Dr. Dido Lenze for support and usefull discussions. We are also very greatfull for the generous funding the DFG provided.

References

[1] Affymetrix: Alternative Transcript Analysis Methods for Exon Arrays, 2005. http://www.affymetrix.com/support/technical/whitepapers/exon_alt_transcript_analysis_whitepaper.pdf.

[2] Affymetrix: Exon Probeset Annotations and Transcript Cluster Groupings, 2005. http://www.affymetrix.com/support/technical/technotes/exon_array_design_technote.pdf.

[3] Affymetrix: Gene Signal Estimates from Exon Arrays, 2005. http://www.affymetrix.com/support/technical/whitepapers/exon_gene_signal_estimate_whitepaper.pdf.

[4] Affymetrix: GeneChip Exon Array Design, 2005. http://www.affymetrix.com/support/technical/technotes/exon_array_design_technote.pdf.

[5] Affymetrix: Guide to Probe Logarithmic Intensity Error (PLIER) Estimation, 2005. http://www.affymetrix.com/support/technical/technotes/plier_technote.pdf.

[6] Affymetrix: Human Exon 1.0 ST Array and WT Sense Target Labeling Assay for Genome-Wide, Exon-Level Expression Analysis, 2006. http://www.affymetrix.com/support/technical/technotes/human_exon_wt_target_technote.pdf.

[7] M. Adesnik, M. Salditt, W. Thomas, and JE Darnell. Evidence that all messenger RNA molecules (except histone messenger RNA) contain Poly (A) sequences and that the Poly (A) has a nuclear function. *Journal of molecular biology*, 71(1):21, 1972.

[8] I. Affymetrix. Statistical algorithms description document. *Technical paper*, 2002.

[9] M. Anton, D. Gorostiaga, E. Guruceaga, V. Segura, P. Carmona-Saez, A. Pascual-Montano, R. Pio, L. Montuenga, and A. Rubio. SPACE: an algorithm to predict and quantify alternatively spliced isoforms using microarrays. *Genome Biology*, 9(2):R46, 2008.

[10] Amandine Bemmo, David Benovoy, Tony Kwan, Daniel Gaffney, Roderick Jensen, and Jacek Majewski. Gene expression and isoform variation analysis using affymetrix exon arrays. *BMC Genomics*, 9(1):529, 2008.

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

[12] J.M. Bland and D.G. Altman. Statistics notes: Multiple significance tests: the Bonferroni method. *British Medical Journal*, 310(6973):170, 1995.

[13] B. Bolstad, F. Collin, J. Brettschneider, K. Simpson, L. Cope, R. Irizarry, and TP Speed. Quality assessment of Affymetrix GeneChip data. *Bioinformatics and computational biology solutions using R and bioconductor*, pages 33–47, 2005.

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

[15] M.S. Cline, J. Blume, S. Cawley, T.A. Clark, J.S. Hu, G. Lu, N. Salomonis, H. Wang, and A. Williams. ANOSVA: a statistical method for detecting splice variation from expression data. *Bioinformatics*, 21(90001), 2005.

[16] X. Cui and G.A. Churchill. Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol*, 4(4):210, 2003.

[17] C. Della Beffa, F. Cordero, and R.A. Calogero. Dissecting an alternative splicing analysis workflow for GeneChip® Exon 1. 0 ST Affymetrix arrays. *BMC Genomics*, 9(1):571, 2008.

[18] R. Gentleman, V. Carey, D. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10):R80, 2004.

[19] B.E. Howard, B. Sick, and S. Heber. Unsupervised assessment of microarray data quality using a Gaussian mixture model. *BMC bioinformatics*, 10(1):191, 2009.

[20] E. Hubbell, W.M. Liu, and R. Mei. Robust estimators for expression analysis. *Bioinformatics*, 18(12):1585, 2002.

[21] R.A. Irizarry, B. Hobbs, F. Collin, Y.D. Beazer-Barclay, K.J. Antonellis, U. Scherf, and T.P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, 2003.

[22] W.E. Johnson, W. Li, C.A. Meyer, R. Gottardo, J.S. Carroll, M. Brown, and X.S. Liu. Model-based analysis of tiling-arrays for ChIP-chip. *Proceedings of the National Academy of Sciences*, 103(33):12457, 2006.

[23] K. Kapur, Y. Xing, Z. Ouyang, and W.H. Wong. Exon arrays provide accurate assessments of gene expression. *Genome Biology*, 8(5):R82, 2007.

[24] N. Kazuyuki, Y. Ryo, N. Masao, S. Ayumu, I. Seiya, and M. Satoru. ExonMiner: Web service for analysis of GeneChip Exon array data. *BMC Bioinformatics*, 9, 2008.

[25] E. Laajala, T. Aittokallio, R. Lahesmaa, and L. Elo. Probe-level estimation improves the detection of differential splicing in Affymetrix exon array studies. *Genome Biology*, 10(7):R77, 2009.

[26] C. Li and W.H. Wong. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proceedings of the National Academy of Sciences of the United States of America*, 98(1):31, 2001.

[27] S.J. Noh, K. Lee, H. Paik, and C.G. Hur. TISA: tissue-specific alternative splicing in human and mouse genes. *DNA research*, 13(5):229, 2006.

[28] M. Okoniewski, T. Yates, S. Dibben, and C. Miller. An annotation infrastructure for the analysis and interpretation of Affymetrix exon array data. *Genome biology*, 8(5):R79, 2007.

[29] M.J. Okoniewski and C.J. Miller. Comprehensive analysis of affymetrix exon arrays using bioconductor. *PLoS Comput Biol*, 4(2):e6, 02 2008.

[30] G. Paul, C. Tyson, S. Brian, S. Michelle, Y. Qing, V. James, S. Anthony, A. Tarif, S. Charles, D. Suzanne, et al. Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array. *BMC Genomics*, 7, 2006.

[31] K.D. Pruitt and D.R. Maglott. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Research*, 29(1):137, 2001.

[32] E. Purdom, K.M. Simpson, M.D. Robinson, J.G. Conboy, A.V. Lapuk, and T.P Speed. FIRMA: a method for detection of alternative splicing from exon array data. *Bioinformatics*, page btn284, 2008.

[33] Y. Qu, F. He, and Y. Chen. Different effects of the probe summarization algorithms PLIER and RMA on high-level analysis of Affymetrix exon arrays. *BMC bioinformatics*, 11(1):211, 2010.

[34] G. Ramsay. DNA chips: state-of-the art. *Nature biotechnology*, 16(1):40–44, 1998.

[35] A. Rasche and R. Herwig. ARH: predicting splice variants from genome-wide data with modified entropy. *Bioinformatics*, 26(1):84, 2010.

[36] L. Sachs and J. Hedderich. *Angewandte Statistik.* Springer Verlag, Berlin, Heidelberg, New York, 13 edition, 2009.

[37] S. Shah and J. Pallas. Identifying differential exon splicing using linear models and correlation coefficients. *BMC Bioinformatics*, 10, 2009.

[38] S. Siegel. Nonparametric statistics. *American Statistician*, 11(3):13–19, 1957.

[39] T.M. Therneau and K.V. Ballman. What does PLIER really do? *Cancer Informatics*, 6:423, 2008.

[40] K. Thorsen, K.D. Sorensen, A.S. Brems-Eskildsen, C. Modin, M. Gaustadnes, A.M.K. Hein, M. Kruhoffer, S. Laurberg, M. Borre, K. Wang, et al. Alternative splicing in colon, bladder, and prostate cancer identified by exon array analysis. *Molecular & Cellular Proteomics*, 7(7):1214, 2008.

[41] V.G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays, April 22 2008. US Patent 7,363,165.

[42] E.T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S.F. Kingsmore, G.P. Schroth, and C.B. Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, 2008.

[43] L. Xi, A. Feber, V. Gupta, M. Wu, A.D. Bergemann, R.J. Landreneau, V.R. Litle, A. Pennathur, J.D. Luketich, and T.E. Godfrey. Whole genome exon arrays identify differential expression of alternatively spliced, cancer-related genes in lung cancer. *Nucleic Acids Research*, 2008.

[44] Y. Xing, K. Kapur, and W.H. Wong. Probe selection and expression index computation of affymetrix exon arrays. *Plos One*, 1(1), 2006.

[45] Y. Xing, P. Stoilov, K. Kapur, A. Han, H. Jiang, S. Shen, D.L. Black, and W.H. Wong. MADS: a new and improved method for analysis of differential alternative splicing by exon-tiling microarrays. *Rna*, 14(8):1470, 2008.

[46] J. Zhu, F. He, S. Song, J. Wang, and J. Yu. How many human genes can be defined as housekeeping with current expression data? *BMC genomics*, 9(1):172, 2008.