UNIVERSITA' DEGLI STUDI DI VERONA

DIPARTIMENTO DI BIOTECNOLOGIE

SCUOLA DI DOTTORATO DI SCIENZE INGEGNERIA MEDICINA
DOTTORATO IN BIOTECHNOLOGIE APPLICATE

XXIV° Ciclo

PERFORMANCE ASSESSMENT OF DIFFERENT MICROARRAY DESIGNS USING RNA-Seq AS REFERENCE

S.S.D. BIO/18

Coordinatore: Prof. MASSIMO DELLEDONNE
Firma

Tutor: Prof. MASSIMO DELLEDONNE
Firma

Co-Tutor: Dr. GIOVANNI MALERBA

Dottorando: Dott. NOEL DOUGBA DAGO
Firma

CONTENTS

Acknowledgments

First I would like to thank Prof. Massimo Delledonne for suggesting the subject and for supervising this thesis with such great interest. I appreciate his support and excellent advice and I always felt welcome to discuss my work with him.

I would also like to thank Dr Giovanni Malerba for his motivating support in statistic analysis as well as accepting to act as the co-supervisor.

Further thanks go to my colleagues Dr Alberto Ferrarini and Dr Paola Tononi for their practical help as well as for the pleasant working environment and the many enjoyable events and inspiring discussions we shared.

Finally, I would like to thank all my family and in particular my wife (Teresa Liberatore) for her constant and important support during the last three years.

ABSTRACT

In the past decade, the completion of sequencing of higher organisms has led to the development of whole transcriptome analysis techniques. Among the most important innovations in this field is the microarray technology. It allows to quantify the expression for thousand of genes simultaneously by measuring the hybridization from a tissue or cell of interest to probes immobilized on a solid surface. This powerful technology has applications in addressing many biological questions at genomic scale that were not approachable previously; however, the enormous size of microarray data sets leads to issues of experimental design and statistical analysis that are unfamiliar to many molecular biologists. The type of array used, the design of the biological experiment, the number of experimental replicates, and the statistical method for data analysis should all be chosen based on the scientific goals of the investigator. Here we compare two different strategies of array design (single replicate probe per transcript and multiple probes per transcript) in two highly customizable microarray platforms (CombiMatrix and NimbleGen). In this work we implemented a statistical methodology based on comparison of microarrays with RNA-Seq data to highlight the differences among different platforms and array designs and the causes of such differences. Our work showed that, the four analyzed microarray designs exhibited different advantages depending on the considered parameter (sensibility, specificity, accuracy and predictive positive values).Thus, our results provide insights and guidance that can be used by researchers for properly selecting the approach more suitable to their scientific goal.

1-INTRODUCTION

1.1- Gene Expression Microarrays

1.1.1. DEFINITION AND WORK FLOW OVERVIEW

Definition and application of Gene Expression Microarray

Microarray technology has revolutionized molecular biology. It uses hundreds to millions of highly organized probes on a limited solid surface to simultaneously interrogate the multiple RNA or DNA molecules, defined as target within individual sample (*Lockhart et al*., 1996; *Schena et al*., 1995). The microarray technologies are universal tools that can be applied throughout the life sciences (*Brown & Botstein*, 1999; *Lockhart & Winzeler*, 2000;*Young*, 2000). mRNA-expression profiling is the most frequent application. Parallel quantification of large numbers of messenger RNA transcripts using microarray technology promises to provide detailed insight into cellular processes involved in the regulation of gene expression. This should allow new understanding of signaling networks that operate in the cell and of the molecular basis beyond cellular process. The principle of a microarray experiment, is that mRNAs from a given cell line or tissue are labelled sample, and hybridized in parallel to a large number of DNA sequences, immobilized on a solid surface in an ordered array (see Figure 1 below). Number of probes immobilized can vary from hundreds designed to monitor the expression of few specific genes to hundreds of thousands representing complete transcriptomes. Presently, there are three major applications for DNA microarray-based mRNA expression profiling. One application is identification of genes differentially expressed, compared to a standard condition, in specific biological conditions and stimulus, e.g., in response to particular treatments, in particular cell types, or in particular mutants. Such genes are often considered candidates as players in the biological process of interest. For example, genes that are highly expressed in drug-resistant tumour tissues but not in drug-susceptible tumour tissues might be involved in the drug resistance mechanism. Even if these genes are not casually involved, their mRNA levels can be used as molecular markers for drug resistance. A second application of microarray-based gene

expression profiling is an extension of the first one and involves the comparison of expression levels of genes under many different biological conditions (*Stuart et al*., 2003). mRNA profiles of different genes across many biological conditions (samples) are compared to identify genes that are transcriptionally regulated in a similar manner. Genes co-expressed across many different conditions are likely to be involved in similar biological processes, such as the same biosynthetic pathway. This way of defining genes putative functions is often called a guilt-by-association approach. For example, eukaryotes generally have many cytochrome P450 genes. Which P450 genes are involved in a particular metabolic pathway? The ones whose mRNA profiles are similar to those of genes that are known to be involved in the pathway are good candidates. A third application of gene expression microarrays involves treating expression profiles as descriptions of collective behaviours. The state of the cell from which the sample was prepared is collectively characterized by determining the expression levels of tens of thousands of genes. Thus, in this application, mRNA profiles consisting of data for many genes in different samples are compared. For example, an expression profile of a drug-treated cell describes the effect of the drug, and drugs with similar modes of action can be identified by finding drug-treated cells with similar profiles. Due to the size and complexity of mRNA profile data, computational tools are required for analysis. These tools must be tailored according to the type of analysis being carried out. If the goal is to identify genes that show differential expression between different samples, statistical tools for significance tests and multiple tests correction are needed to sort genes based on the degree of likelihood that they are actually differentially expressed.

Microarray Expression Work-flow Overview

The work-flow of a microarray experiment encompasses the production of the arrays, the extraction of mRNA from tissue or cells samples of interest, labelling of the mRNAs and hybridization to the array, scanning, data processing, statistical analysis and validation (Figure 1).

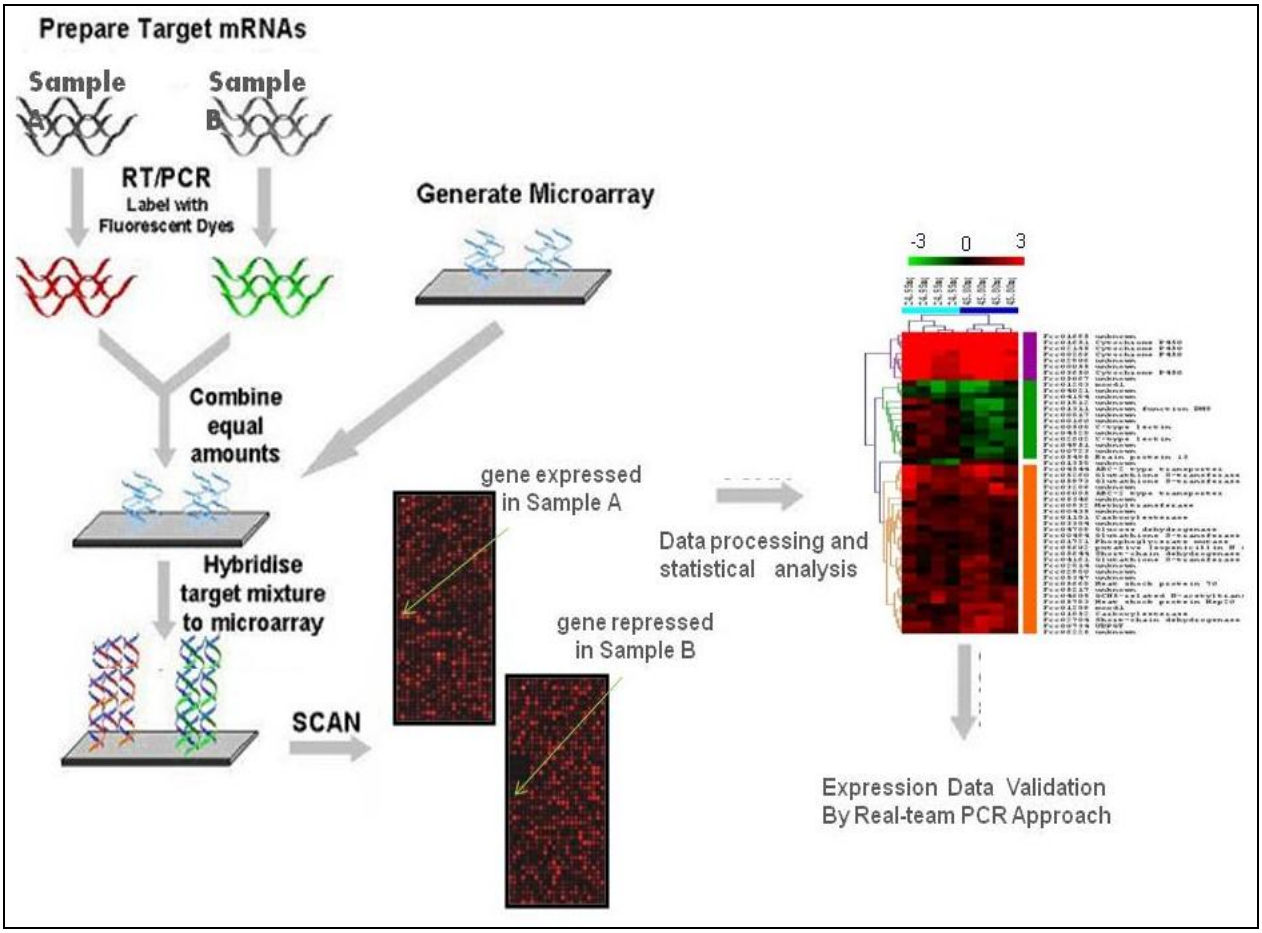

Figure 1: work-flow of microarray expression experiment (from array preparation to expression data analysis and validation)

1.1.2. MICROARRAY MANUFACTURING PROTOCOL

Oligonucleotide microarray are produced by *in situ* synthesis of oligonucleotide or by deposition of prefabricated oligo probes. The probes are designed to be complementary to sequences of known or predicted open reading frames and to represent a single gene or family of gene splice-variants. In spotted microarrays, the probes are oligonucleotides, cDNA or small fragments of PCR products that correspond to mRNAs. The probes are synthesized prior to deposition on the array surface and are then spotted onto a glass slide. The required array printers are widely available, so this technology is suitable for production of custom arrays in academic laboratories. However, the probe density is fairly limited compared with commercial *in situ* synthesized arrays. The advantages of this technology are its flexibility and relatively low cost. However, the cost advantage is diminishing these days as commercial arrays are becoming less expensive. In addition, there are associated

technical challenges as spotted arrays are less uniform than *in situ* synthesized arrays, making it difficult to compare data obtained from different arrays. Consequently, two-channel methods are often used (see below). Microarrays produced by *in situ* synthesis of oligos can be fabricated using a variety of technologies such as photolithography using pre-made masks, photolithography using dynamic micro-mirror devices (DMD), electrochemistry or ink-jet technology. Sequences may be shorter (25-mer probes produced by Affymetrix) or longer (60-mer probes produced by such the Agilent and NimbleGen) depending on the platform and purpose. They are all generally characterized by a high reproducibility of spot shape and concentration, small features sizes and high density. Affymetrix oligonucleotide probes are synthesized using semiconductor based photochemical synthesis. To target specific nucleotides to exact probe sites, photolithographic masks are used. Each photolithographic mask has a defined pattern of windows, which acts as a filter to either transmit or block light from specific features on the chemically protected microarray surface. Areas of the microarray surface in which light has been blocked will remain protected from the addition of nucleotides, whereas areas exposed to light will be deprotected, and specific nucleotides can be added. The pattern of windows in each mask directs the order of nucleotide addition. *In situ* probe synthesis is therefore accomplished through the cycling of masking, light exposure, and the addition of either A,C,T, or G bases to the growing oligonucleotide (see Figure 2). Because Affimetrix in situ-synthesized probes are short (20 to 25bp) , multiple probes per target are included to improve sensitivity, specificity, and statistical accuracy. Moreover Affimetrix technology requires the construction of a very expensive new set of photolithographic masks for each design and as such it's not flexible at all. Other high-density oligonucleotide microarrays include those manufactured by Roche NimbleGen (*Madison,* WI) and Agilent Technologies (*Palo Alto,* CA). Both platforms use longer oligonucleotide probes (60 to 100 bp), but NimbleGen uses mask-less photo-mediated synthesis, and Agilent employs inkjet technology for the *in situ* manufacturing of the probes. The Roche NimbleGen approach to in situ synthesis

is similar to that of the Gene-Chips described above, but photolithographic masks are replaced by virtual or digital masks in Roche NimbleGen's mask-less array synthesizer technology. Mask-less array synthesizer technology uses an array of programmable micro-mirrors to create digital masks that reflect the desired pattern of light to deprotect the features where the next nucleotide will be coupled (see Figure 2). Such method does not require the construction of photolithographic masks and as such has very little overhead costs for the production new designs. Electrochemistry methods based on microelectrode arrays have been also used for the *in situ* manufacturing of the probes by CombiMatrix. The probe synthesis is accomplished in a reaction chamber in which an array of electrodes directs acid deblocking to confined regions on a solid support. The method uses conventional DNA synthesis chemistry with an electrochemical deblocking step. Specific regions of a silicon slide are locally acidified by application of an electric field to microelectrodes, thus allowing nucleotide addition only at chosen sites. In practice, the oxidation of an electrolyte solution after the application of current to microelectrodes liberates acid at the anodes; concomitant reduction at the cathodes consumes acid. The ions and radicals generated by these redox reactions at the electrode surfaces move away from the electrodes and to the substrate through a combination of diffusion, migration and convection effects. During this transit time the electrode products may further react. Once the primary or secondary products reach the substrate, they may react to remove protecting groups (deblock), facilitating coupling of the next nucleotide (see Figure 2). CombiMatrix technology is the only *in situ* synthesis based technology for which a commercial synthesizer for in house microarray manufacturing is available. Thus, such technology has all the advantages associated with *in situ* oligo synthesis (reproducibility, high density) but at the same time it's extremely flexible and allows the production of arrays by academic laboratories.

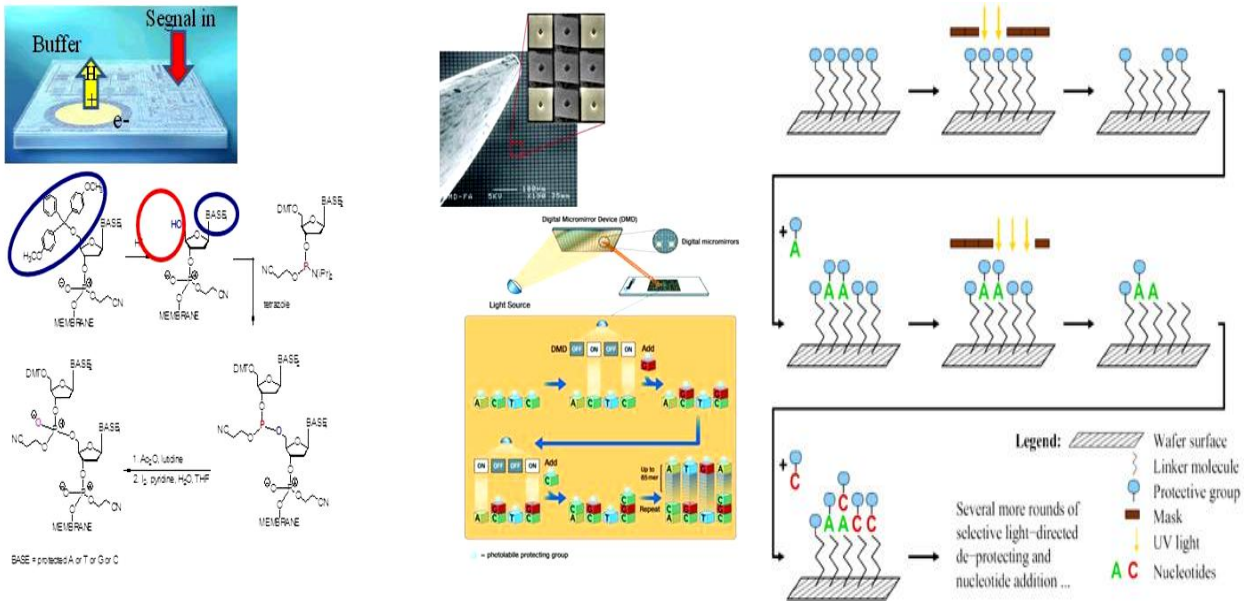

Figure2: from left to right: electrochemistry, photolithographic digital micro-mirror devices (DMD) and mask photolithographic technologies have been represented.

1.1.3. MICROARRAY DESIGN STRATEGIES

Microarray expression technology is strongly influenced by the number, identity and quality of the probes on the array. To trace and maintain the identity of the thousands of probes requires an informatic system throughout the fabrication process (*Donson et al*., 2002). Gene-specific oligonucleotide probes are currently used in microarrays to avoid cross-hybridization of highly similar sequences. If the probes are not optimized for sequence specificity, microarrays can generate false-positive data due to non-specific cross-hybridization to highly similar sequences, gene families (*Xu,W. et al*., 2001), or alternatively spliced variants (*Modrek,B. and Lee,C*., 2002). In addition, high local sequence similarity between different sequences also causes significant cross-hybridization. Long oligonucleotide probes are prone to cross-hybridization and thus often exhibit poor discrimination and hybridize to similar sequences. Studies have suggested that the percentage sequence homology is a reasonable predictor of cross-hybridization (*Evertsz,E.M.*, 2001) and to overcome this cross-hybridization problem, a general practice adopted by several laboratories is to design oligonucleotide probes targeting regions of low sequence similarity (*Kane,M.D et al.*, 2000). Literature data (*Hughes,T.R. et al, 2001*) indicate that longer

oligonucleotides provide significantly better detection sensitivity than shorter probes. Single or multiple probes per genes can be designed. Cheng-Chung Chou et al. (*Cheng-Chung et al.*,2004) demonstrated that a single 70mer or longer oligonucleotide probe for a gene could be sufficient for accurate expression measurement if the probe is validated experimentally. However, it has been reported that oligonucleotide probes binding to different regions of a gene yield different signal intensities (*Lockhart,D.J. et al*., 1996), and it is difficult to predict whether an oligonucleotide probe will bind efficiently to its target sequence and yield a good hybridization signal on the basis of sequence information alone (*Li, F.* and *Stormo,G.D.* ,2001). Because of this, multiple probes per gene have been used in oligonucleotide array designs to obtain reliable quantitative information of gene expression . Furthermore it has been shown that the measurement bias decreased with an increase in the number of probes per gene. Fewer probes per gene were required for the longer probes to achieve the same bias reduction as shorter probes. Several authors results (*Cheng-Chung Chou et al.*, 2004) show that a single long probe per gene selected using a computational approach without experimental validation may not accurately measure the true gene expression signal intensity and that multiple oligonucleotide probes per gene are required to obtain statistically reliable gene expression data. Another important parameter that can influence the hybridization intensity using multiple probe per transcript design in expression microarray study is the spacer between probes. The spacer effect is negligible for long probes while for short and medium probes much longer spacers were required to show a significant effect (*Cheng Chung et al.*,2004). However, previous reports have indicated that the addition of a spacer has a large effect on the hybridization signal intensity for 15–30mer oligonucleotides, but that the signal decreases with spacer length after an optimal length is reached (*Guo,Z. et al.*, 1994).

1.1.4. NUMBER OF CHANNELS

The standard microarray expression experimental paradigm compares mRNA abundance in two different biological samples, on the same or replicate microarrays.

Each microarray platforms has been optimized to work with either a single or dual colour detection system (single and two channels). In both cases, mRNA from cells or tissue is extracted, converted to DNA and labelled, hybridized to the DNA elements on the array surface of the array, and detected by fluorescence scanning. The high reproducibility of *in situ* synthesis of oligonucleotide chips allows accurate comparison of signals generated by samples hybridized to separate arrays in single colour experiments. In the case of spotted arrays, the process of gridding and the uniformity of spots is not accurate enough to allow comparison between different arrays. Thus spotted arrays generally are analyzed in two colour experiments. Two colour experiments essentially compare transcript abundance between two different biological samples on the same microarray. One florescent target is prepared from a reference mRNA and the second from RNA isolated from the treated cells or a disease tissue under investigation. The reference is either generated from normal cells or tissues, or a standard mRNA mix. After hybridization and washing, the slide is scanned using two different wavelengths, corresponding to the dyes used, and the intensity of the same spot in both channels is compared. This results in a measurement of the ratio of transcript levels for each gene is represented on the array. To be able to compare a large number of samples, the same reference RNA sometimes a mixture of all the samples of one experiment or a commercially available standard can be used. In single channel microarrays or one colour microarrays, the arrays provide intensity data for each probe or probe set indicating a relative level of hybridization with the labelled target. However, they do not truly indicate abundance levels of a gene but rather relative abundance when compared to other samples or conditions when processed in the same experiment. Each RNA molecule encounters protocol and batch-specific bias during amplification, labelling, and hybridization phases of the experiment making comparisons between genes for the same microarray uninformative. The comparison of two conditions for the same gene requires two separate single-dye hybridizations. One strength of the single-dye system lies in the fact that an aberrant sample cannot affect the raw data derived from

other samples, because each array chip is exposed to only one sample as opposed to a two-colour system in which a single low-quality sample may drastically impinge on overall data precision even if the other sample was of high quality. Another benefit is that data are more easily compared to arrays from different experiments so long as batch effects have been accounted for. A drawback to the one-colour system is that, when compared to the two-colour system, twice as many microarrays are needed to compare samples within an experiment.

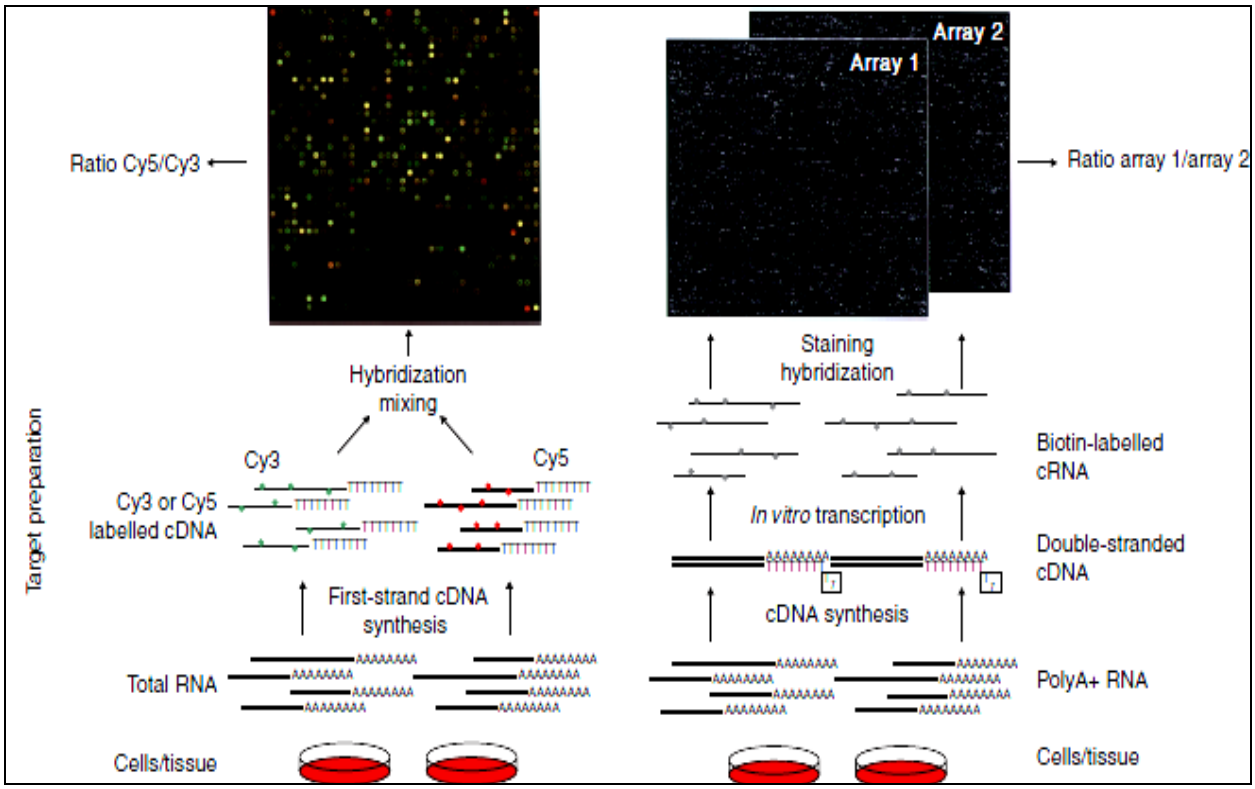

Figure3: image from *Nature Cell Biology* Vol3 (2001) show the schematic overview of two channels (left) and one channel (right) system in microarray target preparation.

1.1.5. TARGET LABELLING STRATEGIES

Several different cDNA labelling methods have been developed for microarray based gene expression analysis. Three popular methods or variations are employed in most laboratories to label target for use with microarrays: direct incorporation (direct labelling), indirect or post labelling using an amino-allyl method and chemical labelling. Direct labelling is the most straightforward approach to generate Cy3 or

Cy5 labelled cDNA. In the direct labelling scheme a Cy3 or Cy5 dye is incorporated during the first cDNA strand synthesis from total or poly A RNA samples. The post-labelling methodology is a two step process where first-strand cDNA is initially labelled with amino-allyl deoxyuridine triphosphate (AA-dUTP) followed by the chemical coupling of cyanine dyes to the nucleic acid. This technology is usually used for bacterial nucleic acid labelling. The chemical labelling based on the universal labelling system (ULS). The proprietary universal labelling system technology is based on the stable binding properties of platinum to bio-molecules. The ULS molecule consists of a platinum complex, a detectable molecule and a leaving group which is displaced upon reaction with the target. This ULS molecule forms a coordinative bond, firmly coupling the ULS to the target. A universal labelling system thereby enables one-step non-enzymatic labelling of nucleic acids to be achieved within 15-30 minutes. Universal labelling system labelling of RNA and DNA is compatible with all DNA microarray platforms and represents a fast system for nucleic acid labelling. An alternative target strategy labelling is the fluorescent linear amplification (or low RNA-input linear amplification) that generates a Cy3 or Cy5 labeled cDNA cRNA target from few total RNA amount exist. The mRNA is primed with an oligo dT T7 primer and then amplified linearly with a T7 RNA polymerase. Cy3 and Cy5 labelled cRNA are employed as the hybridization targets. This alternative target strategy labelling is used by Agilent microarray platform. Furthermore, the signal generated from each messenger (target) is largely independent of base composition or length of the transcript.

1.1.6. MICROARRAY DATA ANALYSIS

Gene expression microarrays are widely used as measurement tools in biological research. A wide range of methods for microarray data analysis have evolved, ranging from simple fold change (FC) approaches to testing for differential expression, to many complex and computationally demanding techniques (*Kerr M. k.*, 2003). Recognizing this allows an investigator to choose procedure more judiciously and methodically to direct their efforts more efficiently. In microarrays the

hybridization intensity is represented by the amount of fluorescence emission, which give an estimate of the relative amount of the different transcript that is represented. Several factors should be considered when setting up a microarray experiment:

(i) Design: the development of an experimental plan (experimental design) to maximize the quality and quantity of information. Experimental design affects the efficiency and internal validation of experiments (*Yang Y. H et al.*, 2002). In microarray analysis, two types of replication can be carried out: technical replication, when mRNA from a single biological case is used on multiple microarrays, and biological replication, when measurements are taken from multiple cases. Technical replicates allow only to estimate the effect of measurement variability whereas biological replicates allow to estimate the combination of measurement variability and biological difference between cases. Consequently, although almost all experiments that use statistical inference require biological replication, technical replicates are almost never required when the aim is to make inference about population that are based on sample data, as is the case in most microarray studies. However, there are some situations where technical replication is needed, such as quality-control studies. Additionally, if the number of cases available is finite or small, or if the cost of obtaining another case exceeds the cost of an array, then technical replicate replication might be useful in addition to biological replicates. How many biological or technical replicates are needed for an expression microarray experiment? Traditional approaches to analyzing statistical power are ill suited to microarray studies which test many hypotheses, use false discovery rate (FDR) estimates for inference, and often use classification technique that have thousands of transcript. Statisticians have therefore begun to provide tailor made solutions to calculate the probability of rejecting a null hypothesis that is false and to calculate the sample size requirements for microarrays. Although there is no consensus about which sample size determination procedures are best, there is consensus that power analysis should be done, that newer methods especially for microarray research should be used, and that more replicates generally provide greater power.

Recommending a specific number of replicates or the statistical thresholds necessary to conduct and analyze any particular expression profile experiment would be inappropriate. However, ongoing research into false discovery rate (FDR; *Benjamini and Hochberg*, 1995, 2002; *Tusher et al*., 2001) methodology has identified some critical considerations that can be used to help determine the appropriate sample size and statistical thresholds needed to fulfill experimental objectives. Factors such as the expected percentage of non-differentially expressed genes on the array (usually 90–95% on a non-specialty array), the level of differential gene expression and the significance threshold anticipated by the investigator to be used during the analysis phase to establish a candidate gene list (e.g. a 2-fold difference at $P < 0.001$), and the percentage of false positives the user is willing to accept (a 5–20% FDR) can be used to approximate the number of replicates required to reach a desired experimental power (*Pawitan et al*., 2005). However pooling biological samples can be useful. Variability among arrays can be reduced by pooling mRNA from biological replicates. Many investigators favour this strategy because sample size can be increased without purchasing more microarray. However there are caveat that apply to mRNA pooling. First pooling is not always beneficial for example in the contest of classification, pooling interferes with the ability to accurately assess inter-individual variation and co-variation. Second, one cannot simply analyze one pool per group as the analysis of multiple pools is required to estimate variance for inference testing. Third: the potential problem of the poisoned pool that is one outlier can yield misleading results. Finally, measurement from pools do not necessary correspond to mathematical average of measurement from individuals comprising the pool. Nevertheless, pooling can be beneficial when identifying differential expression is the sole goal, when biological variability is high relative to measurement error, and when biological samples are inexpensive relative to array expenditure. It is also important to know that microarray measurements can be greatly influenced by extraneous factors. If such factors co-vary with the independent variable for example, with different treatments that are applied to two sets of samples this might confound

the study and yield erroneous conclusions. Therefore it is crucial that such factors are minimized or, ideally, eliminated. For example, arrays should be used from a single batch and processed by one technician on the same day. However, this is difficult with large experiments and it is therefore important to orthogonalize extraneous factors (for example, by analyzing equal numbers of samples from two groups under assessment on each day of analysis), or to randomize cases to levels of these factors (*Kerr M. K.* 2003).

(ii) Preprocessing: Processing of the microarray image and normalization of the data to remove systematic variation. Several image-processing methods have been developed and are now available for expression microarray. These methods estimate the amount of RNA from fluorescent array images, while trying to minimize the extraneous variation that occurs owing to technical artefacts (*Nielsen, H. B. et al*, 2005, *Irizarry, R. A. et al,* 2003). For example robust multi-array average (RMA), corrects arrays for background using a transformation, normalizes them using a formula that is based on a normal distribution, and uses a linear model to estimate expression values on a log scale. However, for accurate comparisons both within and among experimental sets it is critical to consider issues such as data quality and processing method prior to data analysis. Contributing factors to low data quality include the length of the cDNA synthesized, the linearity of the target amplification, the efficiency of the fluorescent dye incorporation, and contaminations in the hybridization. Localized quality problems, such as uneven hybridization and contamination, can be identified early by examining the image files of the scanned microarray data, and either removed from the analysis or corrected by spatial correction (*Borevitz et al*., 2003). After condensing the data, a box plot can be used to visualize the detection range of each array, and to compare it with the known dynamic range of the array type. Data outliers with high background, low intensity, or narrow detection range can be identified (Figure 5). Hierarchical clustering of experiments can also be used to assess data quality by determining if replicate or biologically related samples cluster together. Determination of correlation

coefficients is another way to assess similarity or divergence between samples. These and related techniques can be used to assess biological and technical reproducibility between arrays and experiments as well as to track reproducible errors in the experimental process prior to actual data analysis. Another important preprocessing step is normalization, a process by which non biological variation is minimized and standardized and which allows comparisons between microarray experiments. It also generally makes data more consistent with the assumptions that underlie many inferential procedures. Normalization can be applied multiple times at different levels of analysis for different purposes. There are many different methods for normalizing microarray data. Users should be aware that certain condensing algorithms, such as MAS 5.0 or RMA, normalize the data during the condensing process. Depending on the hypothesis, experimental objectives, and experimental design, additional normalization may or may not be required. One way to account for experimental differences between arrays during normalization is to divide every value on the array by the arithmetic or logarithmic median of the entire array. This effectively establishes a common reference for array-to-array comparisons. This calculation is a linear transformation, specific to each array, so the relative expression level differences between genes on the same array do not change. Further self-normalization can be applied according to gene median. This self normalization is useful for pattern finding and for comparing arrays from different experiments. After normalizing each array to its own median, the value for each individual gene is divided by the median value of that gene across all arrays within an experiment. Consequently, the values of all genes across all arrays, regardless of the level at which the gene is expressed, are standardized to a common baseline where the gene median is equal to 1. This technique highlights the pattern of gene expression rather than the absolute magnitude and has been shown to increase cluster stability and accuracy (*Yeung et al*., 2004). A consequence of gene median normalization is that the absolute expression value of each gene and the relative expression difference between genes is lost. Array median followed by gene median normalization of each

experimental set is highly recommended before analyzing data from different microarray experiments. Subtracting the average value of a gene across all chips in all experiments and then dividing by the standard deviation is another method of standardizing data from multiple experiments (*Yeung et al*., 2004). The global normalization method is based on the assumption that the total amounts of labelled mRNA in all samples are similar. While this assumption is legitimate in most cases, it is not always valid. In some transcriptionally inactive samples, such as pollen and fruit, the total amount of RNA transcripts is significantly lower (*Becker et al*., 2003). When comparing these samples, an alternative method using internal controls is recommended (*Becker et al*., 2003).The internal controls can be genes with housekeeping functions that are constitutively expressed, or spiked cRNA controls (*Hill et al*., 2001). When using internal controls, it is important to validate the assumption that the control genes have a constant transcription level across samples.

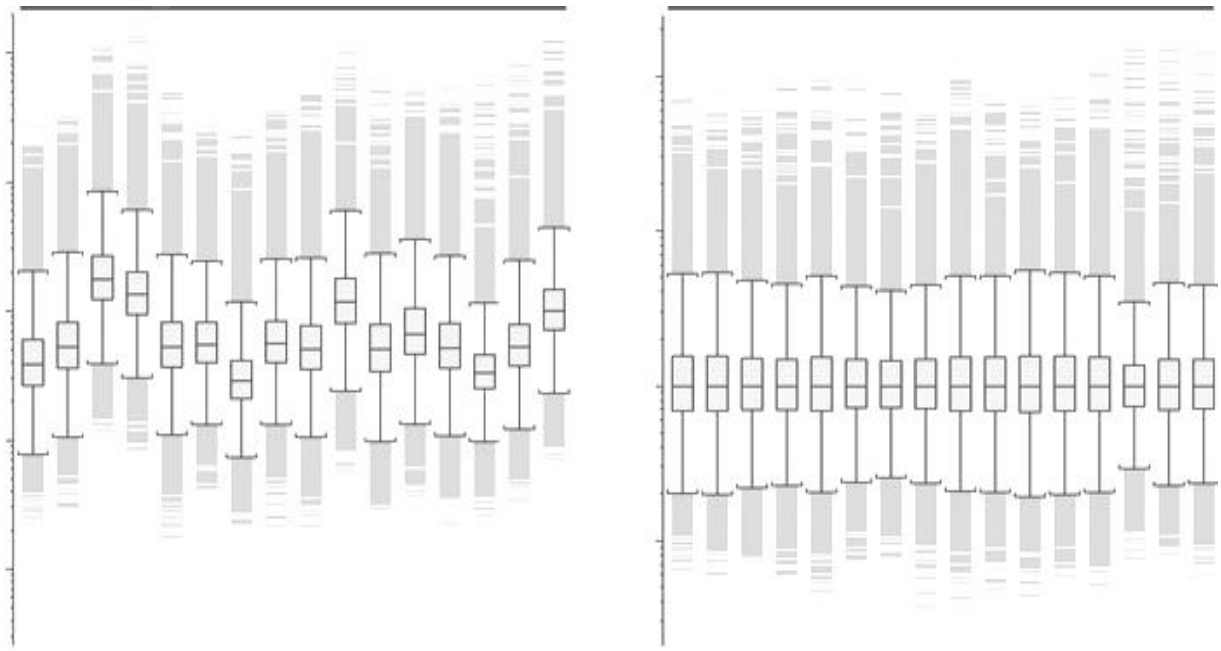

Figure 5: example of raw (left) and normalized (right) data before microarray statistical analysis. The figure also shows data outliers with high background, low intensity, or narrow detection.

(iii) Inference: Inference entails testing statistical hypotheses (these are usually about which genes are differentially expressed). Inference involves making conclusions about the truth of hypotheses that involve unobserved parameters about whole

populations, which are based on statistics obtained from samples. An example hypothesis is: there is a difference in gene expression between mice that are exposed to conditions A and B in the theoretical population of all mice that could have been exposed to conditions A and B. Importantly, there is a clear distinction between inference and the simple ranking of findings for follow-up. Methods are needed to minimize inferential errors that is, type 1 error (false-positive error) and type 2 error (false-negative error) and that estimate the long range error rate. Using fold change (FC), alone as a differential expression test is not valid. Fold change was the first method used to evaluate whether genes are differentially expressed, and is a reasonable measure of effect size. The popularity of fold change stems primarily from its simplicity. Moreover, many researchers use it because it often seems to work reasonably well for ranking results but not for true inference. This is presumably because all transcripts go through the same processing together, and therefore have similar variances. Thinking about fold change is conceptually useful because FC essentially assumes a constant variance across transcripts and, therefore, occupies one pole of a continuum of variance shrinkage. However, fold change is widely considered to be an inadequate statistical test (*Miller et al.*, 2001) because it does not incorporate variance and offers no associated level of confidence (*Hsiao, A et al.*, 2004). Basic statistical analysis uses parametric and non-parametric methods to determine whether differences in transcript value between samples are real or a result of random variation. Most parametric methods assume that the data being interrogated are normally distributed. While it has been reported that, after log transformation, this assumption is legitimate for data for most genes, those genes with low transcript levels may not follow a normal distribution (*Giles and Kipling*, 2003). Commonly used parametric methods are Student's and Welch's t-tests when comparing two groups, which assume equal or disequal variances, respectively, and the analysis of variance (*ANOVA*), which compares more than two groups. Common non-parametric methods include the Wilcoxon rank sum (Mann–Whitney) test and the Kruskal Wallis test when comparing two and more than two groups, respectively.

However, non-parametric methods suffer from a lack of statistical power when using the small sample sizes common to most microarray experiments. It is not practical to explore all statistical options, as more complex methods tend to address specific questions. Moreover, there is no standard statistical methodology that works for every microarray experimental design. Considering the numbers of probes per probe set, probe sets per array, arrays per sample, and samples per experiment, the number of individual data points being evaluated in a typical microarray study is in the millions if not billions. Consequently, the chances are very good that whatever is being searched for will be found at some level of significance whether it is real or not. Caution should be exercised before performing statistics without determining the proper tests or test parameters for a particular hypothesis and experimental design. Furthermore, interpretation of statistical information should include the understanding that: significance and confidence scores strictly apply to expression value differences of probe sets between arrays, and a statistical assessment of differential gene expression can not be used to prove or validate biological function.

(iv) Classification and clustering

The process of classification entails either placing genes into pre-existing categories (supervised classification), or developing a set of categories into which objects can subsequently be placed (unsupervised classification). Many classification algorithms are extensively used in microarray research. Supervised classification (often called prediction or discrimination) entails developing algorithms to assign objects to priority defined categories. Algorithms are typically developed and evaluated on a training data set and an independent test data set, respectively, in which the categories to which objects belong are known before they are used in practical applications. Many supervised classification algorithms are available, but all are susceptible to over fitting to some degree. The phenomenon of over fitting is shown in the Figure 6 which shows the effect of the complexity of the model used on its predictive accuracy. The smaller the sample and the larger the number of transcripts modelled, the more algorithms will capitalize on chance sample patterns and obtain predictive

functions that perform well with training data but poorly with new data. The great challenge is to determine the optimal degree of model complexity that a given data set can support. A common misconception is that the set of the most differentially expressed genes will necessarily give the best predictive accuracy. The gene list that is obtained from hypothesis testing does not necessarily give the best prediction. No one method for constructing prediction algorithms is widely accepted as superior or optimal. However, experience and intuition suggest that with the sample sizes that are typically available in microarray studies, simpler methods might out-perform more complex approaches. Cluster-analysis algorithms group objects on the basis of some sort of similarity metric that is computed for one or more features or variables. For example, genes (biological objects) can be grouped into classes on the basis of the similarity in their expression profiles across tissues, cases or conditions. Hierarchical cluster analysis graphically presents results in a tree diagram, and is probably the most common unsupervised classification algorithm in microarray analysis. Non-hierarchical clustering methods divide the cases (samples or genes) into a predetermined number of groups in a manner that maximizes a specific function (for example, the ratio of variability between and within clusters). Cluster-analysis approaches entail making several choices, such as which metric to use to quantify the distance or similarity among pairs of objects, what criteria to optimize in determining the cluster solution, and how many clusters to include in the solution. No consensus or clear guidelines exist to guide these decisions. Cluster analysis always produces clustering, but whether a pattern observed in the sample data characterizes a pattern present in the population remains an open question. Resampling based methods can address this last point, but results indicate that most clustering in microarray data sets are unlikely to reflect reproducible patterns or patterns in the overall population (*Garge N. et al.*, 2005).

(v) Validation and finding: the process of confirming the veracity of the inferences and conclusion drawn in the study. For genes that are not declared differentially expressed, it is possible that random measurement error has reduced the ability to

detect true differences and has produced an erroneous inference. So, one could argue that the genes that should be validated are those for which the test statistic was almost significant. This could be done by using a more precise gene-expression measure and/or a larger sample size. It makes sense to do this because random measurement error does not bias results away from, but only towards, the null hypothesis.

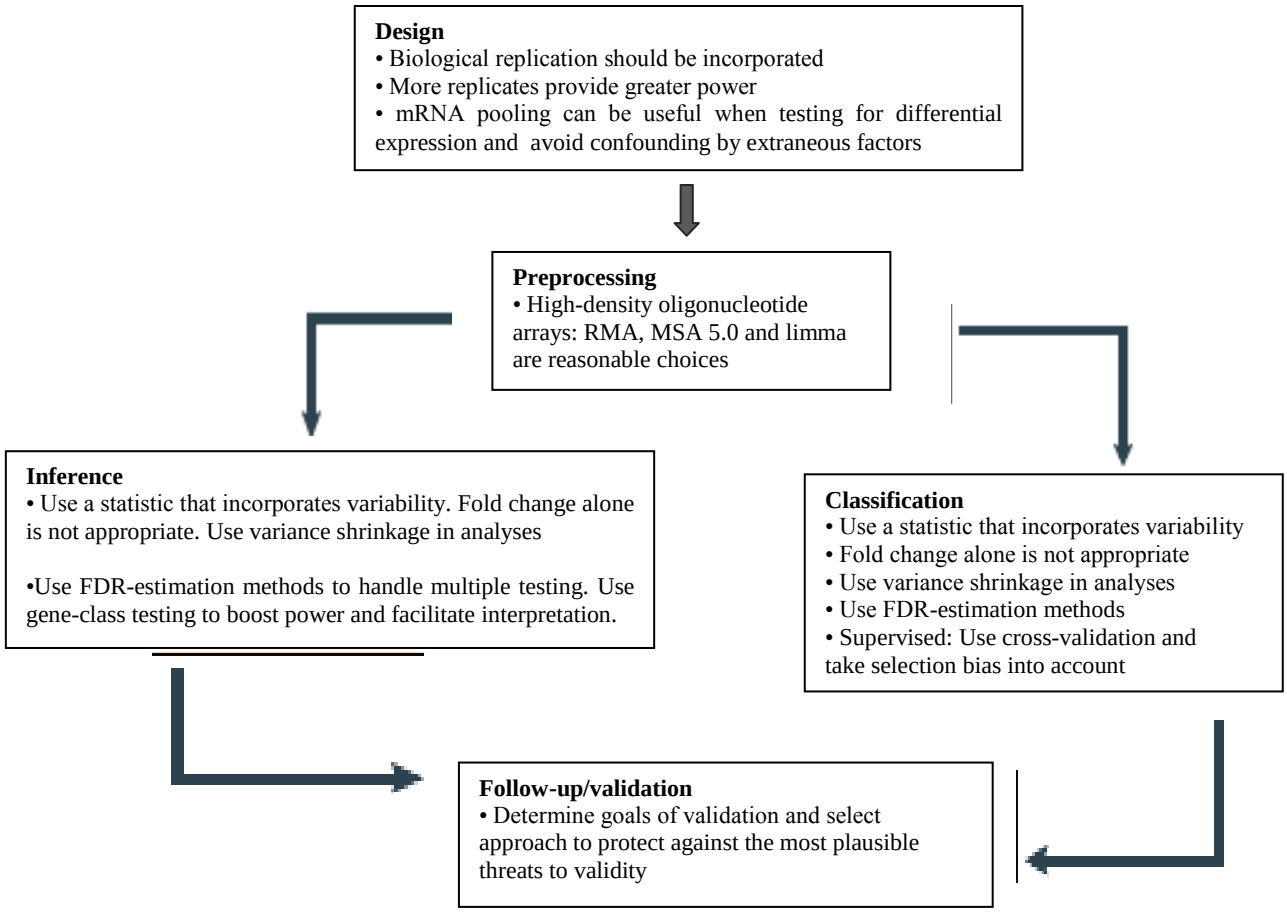

Figure 6: image from *Nature Review|Genetic* (2006) show the guidelines for the statistical analysis of microarray experiments. The flow chart indicates the guidelines for each relevant stage of a microarray study. FDR, false-discovery rate; RMA, robust multi-array average.

1.2- Assessment of Microarray Performance

1.2.1. COMPARISON OF MICROARRAY PLATFORMS (MAQC consortium)

While the multiplicity of microarray platforms offers an opportunity to expand the use of the methodology and make it more easily available to different laboratories, the comparison and integration of data sets obtained with different microarray

platforms is still challenging (*Park PJ. Et al.* , 2004). Sources of diversity arise from the technology features intrinsic to chip manufacturing, from the protocols used for sample processing and hybridization, from detection systems, as well as from approaches applied to data analysis. On one hand, the combined use of multiple platforms can overcome the inherent biases of each approach, and may represent an alternative that is complementary to RT-PCR for identification of the more robust changes in the gene expression profiles on the other hand, the comparison of data generated using different platforms may represent a significant challenge, particularly when considering very different systems. The publication of studies with dissimilar or altogether contradictory results, obtained using different microarray platforms to analyze identical RNA sample, has raised concerns about of the reliability of this technology. The Microarray Quality Control (MAQC) project was initiated to address these concerns, as well as other performance and data analysis issues. The MAQC project has generated a rich data set that, when appropriately analyzed, reveals promising results regarding the consistency of microarray data between laboratories and cross platforms. What is expected from microarrays? Searching for determinants of a phenotype using gene expression levels requires suitable exposure of the genome coupled with reasonable reproducibility, accuracy and sensitivity in the technology employed. These limitations matter less if microarrays are used for screening because changes in gene expression can be verified independently. However, the stakes were raised when microarrays were suggested as a diagnostic tool in molecular disease classification (*Wang, Y. et al*. , 2005) because regulatory agencies, such as the Food and Drug Administration (FDA), require solid, empirically supported data about the accuracy, sensitivity, specificity, reproducibility and reliability of diagnostic techniques. The first decade of microarray technology produced rather limited data pertinent to these issues.

(*i*) *Accuracy*: can be defined as the degree of conformity of the measured quantity to its actual value. Usually, measurements are affected by a bias, which makes the mean depart from the actual value. Given a set of measurements, the accuracy of the

instrument or technique is usually measured by comparing some measure of central tendency of the measurements (e.g. mean and median) to the actual value. An ideally accurate technique would have the mean exactly equal to the actual value.

(ii) Precision: also called reproducibility or repeatability is the degree to which repeated measurements of the same quantity will show the same or similar results. Usually, measurements are affected by an error that makes repeated measurements differ from each other. Given a set of measurements, the precision is usually measured by comparing some measure of dispersion (e.g. variance or standard deviation) with zero. An ideally precise technique would have all measurements exactly equal (zero variance). Accuracy and precision are completely independent. A technique can be accurate but not precise (the mean of several measurements is close to the actual value but the individual measurements vary considerably), precise but not accurate (the individual measurements are close to each other but their mean is far from the actual value) neither or both. If a result is both accurate and precise, it is valid.

(iii) Specificity: in the context of DNA microarrays, refers to the ability of a probe to bind to a unique target sequence. A specific probe will provide a signal that is proportional to the amount of the target sequence only. A non-specific probe will provide a signal that is influenced by the presence of other molecules. The specificity of a probe can be diminished by cross-hybridization, a phenomenon in which sequences that are not strictly complementary according to the Watson–Crick rules bind to each other. Cross-hybridization is also called non-specific hybridization. Microarray experiments generally depend on the hybridization intensity measurement for an individual probe to infer a transcript abundance level for a specific gene. This relationship raises several difficult issues, including: which gene corresponds to which probe, and how sensitive and specific the probe is.

Receiver Operating Characteristic (ROC) Curve Analysis

Receiver Operating Characteristics (ROC) graphs are a useful technique for organizing classifiers and visualizing their performance. ROC graphs are commonly

used in medical decision making, and in recent years have been increasingly adopted in the machine learning and data mining research communities. An ROC graph is a technique for visualizing, organizing and selecting classifiers based on their performance. ROC graphs have long been used in signal detection theory to depict the trade off between hit rates and false alarm rates of classifiers (*Egan*, 1975; *Swets et al.*, 2000). ROC analysis has been extended for use in visualizing and analyzing the behaviour of diagnostic systems (*Swets*, 1988). The medical decision making community has an extensive literature on the use of ROC graphs for diagnostic testing (*Zou*, 2002). Swets, Dawes and Monahan (2000) brought ROC curves to the attention of the wider public with their Scientific American article. An ROC curve is a two-dimensional depiction of classifier performance. To compare classifiers we may want to reduce ROC performance to a single scalar value representing expected performance. A common method is to calculate the area under the ROC curve, abbreviated AUC (*Bradley*, 1997; *Hanley* and *McNeil*, 1982). Since the AUC is a portion of the area of the unit square, its value will always be between 0 and 1.0. However, because random guessing produces the diagonal line between (0; 0) and (1; 1), which has an area of 0.5, no realistic classifier should have an AUC less than 0.5. The AUC has an important statistical property: the AUC of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance. This is equivalent to the Wilcoxon test of ranks (*Hanley* and *McNeil*, 1982).

Metrics for Microarray Quality Control

Reproducibility is the most readily assessable characteristic of any microarray platform. Microarray data consistency within each platform is evaluated by the intra-platform data repeatability and reproducibility (quantitative and qualitative signal measurement) . The precision of microarray detection is estimated by technical reproducibility. This is achieved through technical replication using the same target sample under the same labelling and hybridization conditions. The correlation coefficient and coefficient of variation (CV) are commonly used to estimate the

impact of technical variation on precision. A good microarray experiment should have a correlation coefficient greater than 0.99 and a CV less than 0.15 among technical replicates. In general, a poor correlation between technical replicates is indicative of substandard materials and operation error. Unfortunately, a platform can have an excellent reproducibility without necessarily producing measurements that are accurate or consistent with other platforms. This is because reproducibility only requires that a given probe binds to the same number of labelled transcripts in repeated measurements of the same sample. Badly designed probes that cross hybridize with several other transcripts, can easily provide highly reproducible and yet useless data. Therefore, reproducibility is a necessary but not sufficient requirement. In their appropriate sensitivity range, most microarray platforms produce highly reproducible measurements.

Cross-platform consistency

If microarray data were highly reproducible across various platforms and if they provided information about the absolute transcript levels, one could use appropriately normalized gene expression data without regard to the platform on which the data was obtained. This in turn would reduce the need to replicate experiments and would enable researchers to build universal gene-expression databases that would compile many different data sets from a variety of experimental conditions. This consideration is particularly relevant for clinical samples with limited amounts of mRNA. Owing to the relative scarcity of comprehensive, large scale, spike-in or independently measured gene expression data sets, cross-platform consistency has been used as a surrogate measure of microarray reliability. In this approach, aliquots from the same RNA sample, or RNA isolated from the same biological source are profiled on different microarray platforms. The consistency of these results is considered an indication of the reliability of all platforms compared. Lack of consistency can be caused by the inferior performance of at least one of the platforms, without a clear indication of the relative merit of each platform. Interpreting the cross-platform consistency as proof of accuracy and reliability is not necessarily warranted because

highly similar results across platforms could be simply caused by consistent cross - hybridization patterns without either platform measuring the true level of expression. Nevertheless, a high level of cross-platform consistency is desirable, because, if both platforms performed accurate measurements, cross-platform consistency would automatically follow. Cross-platform consistency is a necessary but insufficient requirement to validate the technology. Despite this limitation, cross-platform consistency studies produced several useful lessons for microarray users. Cross-platform comparison of various microarray platforms depends on the availability of data sets based on same RNA aliquots profiled on different microarray platforms. However one of the difficulties in the cross-platform comparison of microarray data is to ascertain that probes on the various platforms aimed at the same gene do in fact quantify the same mRNA transcript. The various strategies to match probes between different platforms can be constrained by the amount of information provided by the manufacturers of the given microarray. Furthermore, expression values generated on different platforms cannot be directly compared because unique labelling method and probe sequences will result in variable signals for probes that hybridize the same target. Alternatively the relative expression between a pair of sample types should be maintained across platform. For this raison in the MAQC project microarray data for comparability between platform have been examined by reviewing sample type B relative to sample type A (Case and Control) expression values with three different metrics: differential gene overlap, log ratio compression and log ratio rank correlation.

Sources of inaccuracy and inconsistencies in microarray measurements

As a reasonable approximation, signals produced by any given microarray probe can be considered as the composite of three signals: (i) specific signal produced by the originally targeted labelled transcript; (ii) cross-hybridization signal produced by transcripts that have a non-perfect but still significant sequence similarity with the probe; (iii) a nonspecific, background signal that is present in the absence of any significant sequence similarity. Ideally, high specificity microarray platform the

second and third components would be negligible relative to the targeted specific signal. However, even under such ideal conditions, microarray technology in its current state would face significant limitations for several reasons:

First, the relationship between probe sequences, target concentration and probe intensity is rather poorly understood. A given microarray probe is designed as a perfect complementary strand to a given region of the transcript. Based on the Watson–Crick pairing, the probe will capture a certain number of the transcripts. This number is proportional to the concentration of the transcript, but the actual relationship between transcript concentration and the number of molecules bound to the probe, and thus the signal produced, also depends on the affinity of the probe, or free energy change values, under the given hybridization conditions. This affinity is determined to a large extent by the actual nucleotide sequence stretch participating in the binding. This sequence-affinity relationship is not sufficiently understood. Although the sequence dependence of DNA–DNA hybridization in solutions has been studied in detail (*SantaLucia, J., Jr. et al*., 1996), DNA–RNA hybridization has received significantly less attention. Remarkably, the results of Sugimoto et al. (*Sugimoto, N. et al*., 2000) suggest that the sequence dependence of DNA–RNA hybridization can still hold surprises. For example, for certain sequences the binding energy of a DNA–RNA duplex can be stronger for a single mismatch than for the corresponding perfectly complementary strands (*Naef, F. and Magnasco, M.O*. 2003). The kinetics of hybridization are further complicated by the incorporation of modified nucleotides into the target transcripts during the most widely used labelling protocols. Furthermore, the results obtained in solutions cannot be directly applied to the hybridization of microarray probes attached to surfaces (*Peterson, A.W. et al*., 2002). Various researchers have tried to investigate the dependence of affinities on the microarray probe sequence (*Zhang, L. et al.*,2003) but no convincing model has emerged.

Second, splice variants constitute another dimension that can introduce difficulties in the microarray analysis. It is estimated that at least half of the human genes are

alternatively spliced, and might have many potential splice variants (*Modrek, B. et al.*, 2001). A given short oligonucleotide probe is targeted at either a constitutive exon (present in all splice variants) or at an exon specific for certain splice variants. In the former case, the probe intensity will reflect the concentration of all splice variants that are present in the sample, therefore obscuring expression changes occurring in certain splice variants. In the latter case, the specific splice variant will be measured, but other splice variants of the same gene will be ignored. Covering the various types of exons on short oligonucleotide-based arrays is necessary to dissect the splice variant associated composite signals. cDNA microarrays usually have a unique long probe with which the abundance of several splice variants can be measured. This might explain some of the discrepancies often observed between cDNA and short oligonucleotide microarrays.

Third, folding of the target transcripts (*Mir, K.U. and Southern, E.M*. 1999) and cross-hybridization (*Zhang, J. et al*., 2005) can also contribute to the variation between different probes targeting the same region of a given transcript. It has been shown previously that a large proportion of the microarray probes produce significant cross-hybridization signals (*Wu, C. et al*., 2005) for both oligonucleotide and cDNA microarrays. Even a limited stretch of sequence complementarities might be sufficient to enable binding between two unrelated sequences. However, evaluating the overall impact of cross-hybridization on the accuracy of microarray measurements is not easy. Furthermore, the impact of cross-hybridization strongly depends on the relative concentration and the relative affinities of the correct target and the cross-hybridizing

target(s). The latter must be present in sufficient quantities to interfere with specific signals. Cross-hybridization, in conjunction with splice variants, is probably a prime candidate to explain the discrepancies in differential gene expression calls between various microarray platforms, although no systematic study has yet been undertaken. Removing and/or redesigning the microarray probes prone to cross-hybridization is a reasonable strategy to increase the hybridization specificity and hence, the accuracy

of the microarray measurements. However, this requires a good understanding of cross-hybridization, towards which only limited progress has been made owing to the lack of appropriate experimental data. In light of the complexity of microarray signals described, issues such as the compression of expression ratios can be reasonably explained. The presence of cross-hybridization signals on a given probe, for example, might prevent the detection of large changes in gene expression levels because a probe will always produce a certain level of false signal, even if the true signal is much lower or perhaps undetectable.

Alternative mRNA Profiling Technologies

Profiling mRNAs of a few hundred genes by quantitative reverse transcription-PCR (qRT-PCR; reverse transcription followed by real-time PCR) has become a viable option. In principle, qRT-PCR has higher sensitivity and accuracy than microarrays. In practice, the accuracy of qRT-PCR in comparisons between different samples could be compromised since it is common to use only one or a very few genes for the purpose of between sample normalization. Typical microarray data sets have numerous relatively invariant genes that span a wide range of expression levels,
which is important for good between sample normalization. More recently, next generation massively parallel sequencing technologies, have made it feasible to quantitate mRNA by direct sequencing of cDNAs and counting each mRNA species (*Torres et al.,* 2008). Currently, next generation sequencing technology based mRNA profiling (RNA-Seq) is expensive and is focused on extremely high detection sensitivity. However, multiplexing many different samples could make it economically competitive with ordinary microarray based mRNA profiling. It is conceivable that the next generation sequencing technology will make many applications of DNA microarray technology obsolete in a relatively short time. As mentioned, microarrays have grown in popularity ever since but have been heavily criticized and challenged right from the start by other high-throughput methods such as the serial analysis of gene expression (SAGE; *Velculescu et al*. 1995) and the massive parallel signature sequencing (MPSS; *Brenner et al.* ,2000). Both of these

methods are based on the presence of a poly(A) stretch on mRNAs. The SAGE method was designed to generate a single short tag (14–27 bp; *Velculescu et al*. 1995, *Gowda et al*. ,2004) from every poly(A) bearing RNA molecule, concatemerize them to produce a library of tags and systematically tally each transcript in a sample by sequencing the library. MPSS is a different approach where cDNA molecules are immobilized individually on a bead and duplicated to populate its surface. The entire batch of beads is then subjected to several rounds of an ingenious DNA sequencing strategy involving type IIS restriction enzymes, which cut the DNA at a defined distance from their binding site, and a mixture of adaptors bearing different fluorescent dyes targeting the diverse potential combination of resulting cohesive ends. Both of these methods are often referred to as true transcriptomic methods since they do not require knowledge of the transcriptome to be performed. This aspect has become the main criticism toward microarray technologies where the method confines the user to investigate only known candidates whose probes are featured on the array and thus limits the opportunity for generating novel unexpected elements. However, although these methods are extremely powerful, the requirement for large amounts of starting material and the challenging identification of the short tags (for SAGE) or short reads (for MPSS), especially in the absence of a reference genome, prevented the wide use of these methods. These limitations are now being overcome with the development and availability of highly effective sequencing procedures, in addition to the fact that genomes of model species have been sequenced completely and several livestock as well as wild species have been sequenced with at least sufficient coverage to allow chromosome positioning and identification of genes from small DNA sequences. So far, the challenges arising from deep sequencing or RNA-Seq in the case of a transcriptome (*Wang et al*. ,2009) are linked to the efficient handling of the data flow, in addition for requiring the same considerations for the amount of starting material and the targeted RNA population being studied as microarrays. Apart from the potential bias introduced during sample processing, post-sequencing normalization can be challenging if the goal were to generate an overall

perspective of relative abundance values across transcripts as well as across developmental stages. Sample fragmentation introduces a bias related to the size of the transcripts as longer ones will generate more fragments and will thus be counted more often. Strategies have been proposed to account for this size dependent count, such as considering a defined interval from the poly(A) tail (*Robinson and Oshlack* 2010). So far, the issue of the fluctuating RNA composition and poly(A) length occurring during early development has not been addressed for RNA Seq. Although these highly effective sequencing methods represent huge amounts of sequences, on the order of giga-bases, most of them still generate relatively short reads that need to be sorted back onto a frame provided by a reference sequence such as Refseq or a complete genome. Deep sequencing undoubtedly provides a high-performance tool to study RNA abundance and its discovery potential is undeniable. However, the study of RNA abundance and of the impact of splice variants during development may not be undertaken with the ease that is currently expected. In the case of splice variants, the extent of fragmentation required to reduce sequence length and ensure complete coverage somehow limits the power of deep sequencing to discover novel transcript variants, as these fragments must lie directly at the junction of a variant for it to be identified. Third-generation deep-sequencing procedures will soon be available, and some are predicting the demise of microarrays. Until then, the cost of deep-sequencing runs is a limiting factor when considering numerous sample types in addition to technical replication. So far, the main bottleneck of highly effective sequencing remains data processing, normalization, and analysis. The development of benchmarked data processing tools to properly handle the tremendous datasets is underway. As both technologies are complementary, microarrays are still evolving and as such, whole genome tiling arrays may provide a valuable alternative to study transcript variants. They are not available yet for livestock species, are also expensive to run, and have been criticized for their lack of sensitivity due to inherent high background signals (*Wang et al*. ,2009). The development of the exon arrays or restricted tiling arrays solely containing probes targeting all transcripts known to be

expressed in the tissue of interest determined by deep sequencing may provide an acceptable alternative.

1.2.2. MICROARRAY VERSUS RNA-Seq

1.2.2.1 RNA-Seq Definition and Workflow Overview

RNA-Seq is a novel method for gene expression profiling by next-generation sequencing of transcripts. The technology has been applied to gain global views of the complex transcriptomes of mammalian samples, including human embryonic kidney and B-cells (*Sultan et al*., 2008), mouse embryonic stem cells (*Cloonan et al*., 2008), blastomeres (*Tang et al*., 2009), and different mouse tissues (*Mortazavi et al*., 2008). An advantage of RNA-Seq over other profiling technologies is that it allows a comprehensive assay of gene expression that is not reliant on probes for targets that must be specified in advance. It is particularly well suited for the *de novo* detection of splice junctions and allows genome-wide qualitative expression profiling of organisms with unknown genome sequence. Transcript detection obviously benefits from the digital nature of counting sequence reads. The observed identification rate increases with additional sequencing but is partly determined by the non-random nature of biological sequences and the highly skewed distribution of transcript abundances. We can extrapolate an expected achievable identification rate from the observed dependency on experimental parameters like read depth and read length. In addition, we examine the effects of random read sampling on the identification of low-copy number transcripts, as resulting from the distribution of reads mapped to different splice-forms. With many transcription factors being biologically active in low-copy numbers, this is particularly topical for studies of gene regulation. Increasingly, there has been an interest in applying RNA-Seq not only for qualitative transcriptome profiling but also for the quantification of gene expression (*Blow*, 2009; *Jiang and Wong*, 2009; *Shendure*, 2008; *Trapnell et al*., 2010; *Wilhelm et al*., 2008). Using raw read counts mapped to individual targets, however, can result in length-dependent bias (*Oshlack and Wakefield*, 2009). A common approach for the

quantification of gene expression in an RNA-Seq experiment thus computes the number of reads per kilo-base of exonic sequence per million mapped reads (RPKM) to produce a gene expression measure which overall correlates well with measurements from microarrays (*Mortazavi et al*., 2008). Such normalization is also necessary to allow the combination or the comparison of RNA-Seq runs. Unlike microarrays, which measure continuous probe intensities, RNA-Seq quantifies discreet, digital sequencing reads count aligning to a particular sequence. The digital nature of this process support an unlimited dynamic range, which enable researcher to quantify RNA activity at much high resolution, important for capturing subtle gene expression change associated with biological process. Methods for the estimation of the transcript's abundance using RNA-Seq data have been intensively studied, many of which are based on the assumption that the short-reads of RNA-Seq are uniformly distributed along the transcripts. However, the short-reads are found to be non-uniformly distributed along the transcripts, which can greatly reduce the accuracies of these methods based on the uniform assumption. Several methods are developed to adjust the biases induced by this non-uniformity, utilizing the short-read's empirical distribution in transcript. While standard microarray probes only cover around 20% of a gene on average, capturing only a portion of the biologically relevant data, RNA-Seq can profile the entire transcript. The sequencing data can also be re-analyzed as novel exon are discovered whereas the sample would have to be rerun on a microarray with updated probes. Figure 7 represents a global overview of RNA-Seq technology work-flow (see below).

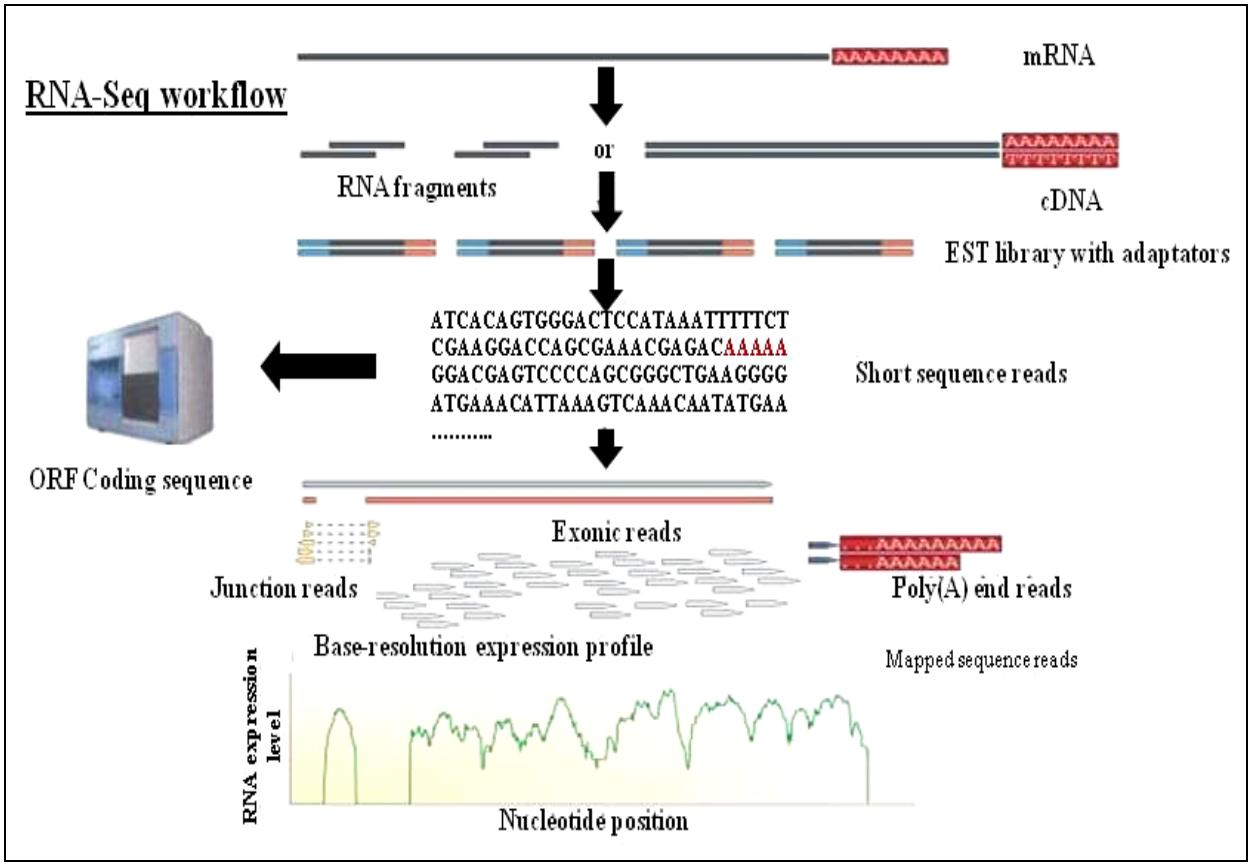

Figure 7: image from *Nature Reviews| genetics* (2010) shows the general overview of messenger RNA (mRNA) expression profiling using RNA-Seq technology

While earlier work has focused on reads that unambiguously identify a transcript, current developments extend data analysis to complex gene models of alternative splicing, also taking into account the many reads that may come from different splice-forms (*Griffith et al*., 2010; *Jiang and Wong*, 2009; Lee *et al*., 2011; *Mortazavi et al*., 2008). A popular recently emerging approach is to align reads to the genome, and then use this information to assemble transcripts *de novo* and calculate their abundances, as implemented by the Top-Hat/Cufflinks tools (*Trapnell et al*., 2009). Despite, or perhaps even because of, the fast pace of development of both the measurement technology and the associated novel analysis tools (*Datta et al*., 2010), the central question of measurement reliability or, of how precisely we can actually quantify transcript expression, has received relatively little attention beyond initial observations of overall good correlation (*Marioni et al*., 2008; *Wilhelm et al*., 2008). Simple correlation coefficients, however, can be misleading as they are dominated by

a small number of very highly expressed genes. Despite the excellent overall correlation, reproducibility seems to be lower for gene classes that are less strongly expressed (*Mortazavi et al*., 2008). It is therefore interesting to consider measurement precision for all targets individually. Similar to the early microarray data, however, there has been a lack of large RNA-Seq datasets with the necessary technical replicates. Now a comprehensive analysis of the reproducibility of gene expression level measurements by RNA-Seq has become possible and constitutes a necessary complement to characterizations of systemic measurement bias in next-generation sequencing (*Bullard et al*., 2010). Measurement precision in particular determines the power of any analysis to reliably identify relevant signals or changes, such as in screens for differential expression, independent of whether replicates are employed or not (*Anders and Huber*, 2010).

1.2.2.2 Statistical Analysis of RNA-Seq Expression Data

Reads count data is discrete and skewed and is hence not well approximated by a normal distribution (see material and method chapter). Thus, a test based on the negative binomial distribution, which can reflect these properties, has much higher power to detect differential expression. In fact, tests for differential expression between two experimental conditions should take into account both technical and biological variability. Recently, several authors have claimed that the Poisson distribution can be used for this purpose. However, tests based on the Poisson assumption (this includes the binomial test and the chi-squared test) ignore the biological sampling variance, leading to incorrectly optimistic p-values. In fact test for differential expression between two experimental conditions should take into account both technical and biological variant. Thus a test based on negative binomial distribution, which can reflect these properties , has much higher power to detect differential expression. Different software package implementing model based on the negative binomial distribution such as *edgeR* (*Mark Robinson, Davis Mc Carthy*, 2012), *bay-Seq* (*Thomas J. Hardcastle*, 2012) and *DESeq* (*Simon Andres*, 2010) have been released during the last couple of years. On of the most used is *DESeq* which is

an open source R package to analyze count data from high-throughput sequencing assay such as RNA-Seq abd test for differential expression. One of the new features of *DESeq* is the ability to estimate the variance in a local fashion, using different coefficients of variation for different expression strengths. This removes potential selection biases in the hit list of differentially expressed genes, and gives a more balanced and accurate result. *DESeq*'s applicability is not limited to RNA-Seq. Rather, it may be used for many kinds of count data derived from highly effective experiments. The *DESeq* package expects count data, as obtained from an RNA-Seq or other high-throughput sequencing experiments. The package *DESeq* provides a powerful tool to estimate the variance in such data and test for differential expression. The *DESeq* package expects count data, as obtained from the RNA-Seq experiment, in form of a matrix of integer values. Each column corresponds to a sample, typically one run on the sequencer. The core assumption of this method it that the mean is a good predictor of the variance. In fact, genes with a similar expression level also have similar variance across replicate. Hence, we need to estimate for each repining and veraison condition a function that allows to predict the variance from the mean. This estimation is done by the calculating for each gene, the sample mean and variance within replicates and then fitting a curve to this data. It is instructive to observe at which count level the biological noise starts to dominate the shot noise. To do that, one should check whether the base variance functions seem to follow the empirical variance well. To this end diagnostic functions were provided by *DESeq* package. Diagnostic function returns, for a specific condition, a data frame with four columns: the mean base level for each gene, the base variance as estimated from the count values of this gene only, and the fitted base variance (the predicted value from the local fit through the base variance estimates from all genes). A more convenient way to view the diagnostic data is given by a plot which show empirical cumulative density functions (ECDF) stratified by base level. Later, RNA-Seq data have been processed for differential analysis using *DESeq* package based on binomial negatives distribution. The output file of this analysis returns a data frame with p values, p

adjusted values, fold change values and logarithm transformation fold change values, general base mean, base mean and the variance of each investigated conditions (samples).

1.2.2.3 Interpretation of Differential RNA Abundance

The greatest current advantage of microarray is knowledge of biases in array data and mature analysis strategies and experimental designs for dealing with them. By comparison, sources of bias in sequence data are still being actively researched, and optimum analytical strategies still to be developed (*Auer PL.et al.*, 2010). Meanwhile RNA-Seq continues to evolve, so it will take some time to develop appropriate standards for this tool. One of the most important concerns about sequencing RNA is the depth of sequencing required to effectively sample the transcriptome. This equates to how many times to sequence a sample. For highly expressed genes, small amounts of sequencing are sufficient, but for the middle and low end of expression levels, it is clear that many reads are needed. A final consideration about arrays and sequencing is the quantity and size of the data. In expression microarrays the raw data are composed of image files, typically TIFF files that may be around 30 MB per array. These TIFF files are transformed into text files that contain fluorescence intensities for each gene. The Illumina instrument generates upwards of 3.5 Tb of data files but the sequence files (around 60 GB) are typically used as a starting point for analysis. These sequence files are orders of magnitude larger than those from arrays and because of these large file sizes, Python, Perl, Unix command line, and other scripting are necessary to sort and experiment with these files. Using spreadsheet software will not be an option and therefore bioinformatics support is necessary. For biologists unfamiliar with computer languages, there are growing alternatives for working with sequencing data. For example, many of the tools for sequencing data analysis are now available in Galaxy software, a web interface that provides a user friendly graphical interface (*Goecks J. et al.*, 2010). A key first question is whether, when used to ask exactly the same question, both techniques give the same answer. The typical RNA management needs to be accounted for in the

interpretation of RNA abundance fluctuations. All high-throughput methods, including microarrays and RNA-Seq, require pre-amplification of the sample to generate sufficient starting material. A source of technical bias is due to the lack of a consistent and reliable approach for data standardization. Data normalization is required to provide a comparable basis between samples and treatments but can also be a source of variation in the absence of an appropriate standard. As a consequence, over normalization can distort RNA abundance values. The impact of the presence of transcript variants can be an additional source of false abundance values. Surprisingly, it was recently estimated that 61% of differential expression detected by microarrays may in fact be associated with the probe detecting a shift in transcript processing, e.g. the presence of variants (*Kwan et al.*, 2008). The unique characteristics of some RNA samples, combined with the numerous options for sample and data processing, profoundly impact the resulting differential expressed gene list. However, in order to minimize the methodological differences and provide a strong basis for comparison, complete evaluation of the impacts of each source of methodological variations needs to be performed. Until then, compatibility between platforms is questionable. On a more positive note, the true gain of knowledge from RNA abundance studies resides in downstream interpretation. This objective may be somewhat buffered from the discrepancies that arise between platforms as the main consideration lies within relative abundance values or more precisely the extent and the nature of the change, rather than absolute transcript copy numbers.

1.2.2.4 RNA-Seq and Microarray Data Validation

Microarray and RNA-Seq expression analysis have revolutionized many facets of biology and will continue to be applied widely. However, significant questions remain with regard to the generation, analysis, and in particular, interpretation of expression data. Although the validation of microarray and RNA-Seq expression results obtained for specific genes using independent techniques is still considered a desirable component of any expression experiments and, the genes selected for validation, are usually identified from the microarray or RNA-Seq data. The

selection is based on the implicit assumption that there is a good correlation between the microarray or RNA-Seq data and actual mRNA levels in the cells or tissues under investigation. One fundamental issue that has not been adequately addressed is how well expression measurement tools score reflect actual mRNA levels in the sample being examined. To facilitate data comparison between expression platforms it is important that scientific community adopt consistent validation methodologies. This is especially important if expression technology play a role in the clinical setting (*Petricoin EF et al.*, 2002). Laboratory based validation of data provides independent, experimental verification of gene-expression levels, and typically begins with the same samples that were studied in the initial gene expression experiment(s). The methodology used varies depending on the scientific question, but commonly used techniques include real-time PCR, northern blot, ribonuclease protection assay, and *in situ* hybridization or immunehistochemistry using tissue microarrays (*Luo, J. et al.*, 2001). The objective is to confirm that the differential expression detected by array or RNA-Seq approaches (expression measurement tool) can be replicated by other means. In general, relative expression level detected from an RNA blot-analysis are similar to those measured by the microarray, although RNA blot analysis can be more sensible as a result of the radioactive labelling (*Taniguchi et al*., 2001). In contrast quantitative RT-PCR is usually more sensitive than microarray detection (*Czechowski et al*., 2004; *Dallas et al*., 2005; *Yuen et al*., 2002). However, while the change detected by the two methods were in the same direction, the dynamic range spanned five orders of the magnitude for the quantitative RT-PCR data but only three orders of magnitude for the Gene-Chip data (*Guimil et al.*, 2005). The correlation between the results form the two methods which are affected by the sequence selected for the probes and primers. Microarray probes and quantitative RT-PCR primers for the same region generally are in better agreement when testing genes with moderate transcript levels (*Etienne et al.*, 2004). Genes that showed poor detection correlation can be explained by different levels of detection, different subset of alternative transcripts being recognized, and probe sequence annotation errors (*Dallas et al.*,

2005). However, the choice of validation methodology remain a contentious issue (*Rockett JC et al.*, 2004). To date, quantitative RT-PCR is the method of validation that has been used in the majority of published microarray and RNA-Seq studies, presumably because it is a rapid, sensitive, high-throughput procedure that requires minimal amount of test material compared to techniques such as Northen blotting or ribonuclease protection assays. As in the case for many studies, quantitative RT-PCR is often the only feasible approach when rare or unique tissues are investigated. For this reason, it would appear likely that quantitative RT-PCR will continue to be used extensively for the validation of microarray and RNA-Seq expression data. Moreover, several studies have demonstrated a strong correlation between microarray and quantitative RT-PCR expression data (*Peter B. Dallas et al.* , 2005). It should be noted, however, that quantitative RT–PCR requires a significant up-front effort to optimize amplification conditions, and the method has potential pitfalls that must be carefully monitored. However, a major problem with validating RNA-Seq expression estimates is that there is no clear gold standard for expression estimation. Comparison of RNA-Seq to microarrays has suggested that the former technology is more accurate than the latter (*Bradford J. et al.*, 2010). Quantitative reverse transcription PCR (qRT-PCR) has served as a benchmark in numerous RNA-Seq studies but it is not a perfect expression measurement assay (*Fleige S, Pfaffl M*, 2006), and it is therefore unclear which technology currently produces the most accurate expression estimates. Nevertheless, at present authors believe it to be the best measure of expression aside from, perhaps, RNA-Seq itself (*Adam Roberts et al.*, 2011). Due to the previously demonstrated superiority of RNA-Seq over microarrays, and the problems with other expression measurement tools, several authors performed their benchmarking with respect to qRT-PCR (*Adam Roberts et al.*, 2011). Quantitative real-time PCR monitors the amount of amplicon generated as the reaction occurs. Usually, the amount of product is directly related to the fluorescence of a reporter dye. Because it detects the amount of product as the reaction progresses, real-time PCR provides a wide linear dynamic range, demonstrates high sensitivity,

and is very quantitative. The initial amount of template DNA is inversely proportional to a parameter measured for each reaction, the threshold cycle (Ct).

1.3 -Expression Microarray Design Based on CombiMatrix Platform

CombiMatrix is a custom high density oligonucleotide microarray based on a maximum of 90 thousand probes (90K) per slide. This platform is available in the Functional Genomics Centre of the University of Verona since 2007. CombiMatrix gene expression oligonucleotide microarray platforms are produced by *in situ* synthesis of oligonucleotide and it is based on ceramic slide (solid support). The advantages of this technology (CombiMatrix microarray platform) are its flexibility and relatively low cost. Design of CombiMatrix microarray oligos was performed by using OligoArray (*Rouillard et al.*, 2003), a programme that designs specific oligonucleotides at the genomic scale. OligoArray software for CombiMatrix oligos design is based on the selection of a single specific probe (35-40mer) per gene. This strategy had some advantages as the data produced are easy to analyze (no complex summarization algorithms involved) and could be validated easily by real time as primers could be designed in correspondence of the specific microarray probe. CombiMatrix platform is an open platform, is customizable and flexible and allows the in house production of custom microarrays with a dedicated synthesizer. Moreover, CombiMatrix microarray platform used an electrochemical technology based on microelectrode arrays for *in situ* oligo fabrication. The high reproducibility of CombiMatrix microarray platforms based on *in situ* synthesis of oligonucleotides allows accurate comparison of signals generated by samples hybridized to separate arrays in single color experiments. In the Functional Genomics Centre of the University of Verona, several bacteria, plants and animal microarray designs have been developed and improved using CombiMatrix platform due to it flexibility. In fact this Centre was also involved in the Grape Genome Sequencing initiative for the development of a transcriptomic platform based on CombiMatrix platform analysis as represented by Figure 8. Table 1 displays the overall overview of CombiMatrix microarray platform for expression analysis (see below).

Table 1: overview of CombiMatrix microarray design based on single specific replicate probe per transcript

	CombiMatrix microarray design
Platform Technology	RNA microarray hybridization
Probe length	35-40 mer
Substrate	Ceramic slide
Deposition	*In situ* synthesis
Detection	One color Cy5 Fluorescence
Software for background correction data normalization and data analysis	*Limma* package
Number of probe per transcript	1
Replicated probe per transcript	3

Despite its great flexibility CombiMatrix microarray platform does allow the hybridization only of a single sample per time limiting the number of samples that can be processed per day. NimbleGen microarray platform overcomes this limitation because it allows up to 12 samples per slide to be processed and as such guarantees a very highly effective hybridization procedure.

1.4- Expression Microarray Design Based on NimbleGen Platform

NimbleGen (12x135K) gene expression microarray platform is a customizable and high density oligonucleotide microarray that allows the processing of 12 different samples with a single slide for gene expression experiment. Probe selection strategies used for NimbleGen for microarray transcript-base design is based on a scoring algorithm developed by NimbleGen. NimbleGen microarray platform approach for *in situ* synthesis is similar to that of the Gene-Chips (see above). However, in this case, photolithographic masks are replaced by virtual or digital masks in NimbleGen's mask-less array synthesizer technology. Mask-less array synthesizer technology uses an array of programmable micro-mirrors to create digital masks that reflect the desired pattern of light to de-protect the features where the next nucleotide will be coupled. NimbleGen microarray expression platform includes 135 thousand probes for each independent hybridization room and it is based on a glass support.

Moreover, this microarray expression platform based on the 60 mer probe formats provides enhancement in sensitivity over the short probe formats partly due to the larger area available for hybridization (*Gary Hardiman*, 2004). The longer 60 mer probe formats are also more tolerant of sequence mismatches and are thus more suitable for the analysis of highly polymorphic regions (*Gary Hardiman*, 2004). The design method used by NimbleGen has several differences and peculiarities compared to the method developed for the CombiMatrix platform. First to avoid non-specific binding, highly repetitive element of the transcript are excluded from microarray designs. NimbleGen has developed a new method that utilizes a strategy similar to the Window Masker program, developed by Margolis (*Bioinformatics.*,2006), to identify these regions and exclude them from probe selection. NimbleGen's method is based on the frequency of 15 mers within the targets genome (smaller, less complex genomes may use shorter *n-mers*). The process compares the set of probes against a pre-computed frequency histogram of all possible 15 mer probes in the target genome. For each probe, the frequencies of the 15 mer comprising the probe are then used to calculate the average 15 mer frequency of the probe. The higher the average 15mer frequency, the more likely the probe is to lie within a repetitive region of the genome. For large eukaryotic genomes, only probes with an average 15 mer frequency of less than 100 are used. While the OligoArray-based protocol developed from the Centre of Functional genomics uses a BLAST comparison with all the possible targets to determine the specificity of the probes, Nimblegen uses a totally different approach. In particular two terms have been defined by NimbleGen to discriminate the cross-hybridization potential of any given probes: frequency and careful uniqueness. There is a subtle but significant difference between frequency and careful uniqueness. The frequency, also referred to as count or perfect matches, is defined as the number of times an oligo of a given size appears exactly within a set of comparison sequences. In fact probe generated for transcript-based designs are generally compared to the transcriptome. Like frequency, careful uniqueness is based on a comparison of an oligo against the source

transcriptome. Careful uniqueness, however, is evaluated based on the existence of close matches rather than exact matches. Uniqueness for a transcript-based design is a Boolean measure that classifies an oligo as either similar or dissimilar to other oligos of the same size in the transcriptome. Each oligo of a given size is compared to all other oligos of that size in the transcriptome, and a weighted score is calculated based on the number and position of mismatches between the two oligos. The resulting score is compared to a set threshold, and if the score exceeds the threshold, the oligo is classified as unique. The default weighting scheme is trapezoidal in shape, giving greater weight to mismatches in the center of the oligo than mismatches at either end. However, NimbleGen microarray platform photolithographic masks *in situ* oligonucleotide synthesis uses the digital micro-mirror device (DMD or digital light processor). NimbleGen microarray platform have described a technology for a centralized production facility that can hold synthesized microarray containing up to195,000 features using a DMD that creates digital masks to synthesize specific polymers (*Nuwaysir EF et al.*, 2002). The NimbleGen technology is mask-less, in that unlike Affimetrix Gene-Chips, no physical masks are involved in the synthesis. The digital micro-mirror device (DMD) pattern light by flipping mirrors on and off according to the instruction provided in a digital mask file. The advantage of this system is that custom high-density arrays can be created in a rapid and cost-effective manner. NimbleGen microarray platforms have been optimized to work with a single colour detection. Table 2 shows the overall overview of NimbleGen microarray platforms for expression analysis (see below).

Table 2: overall overview of NimbleGen microarray design based on 4 four different probe per transcript

	NimbleGen microarray design (*nmgd*)
Platform Technology	cDNA microarray hybridization
Probe length	60 mer
Substrate	Glass slide
Deposition	*In situ* synthesis
Detection	One color cy3 Fluorescence
Background correction, data Normalization and analysis	RMA module from Nimblescan software v2.5 (background correction and data normalization), *limma* package for differential analysis
Number of probe per transcript	4
Replicated probe per transcript	1

The NimbleGen microarray design is based on the 4 best probes on 3' end of transcripts discriminate by NimbleGen probe design algorithm. Ologonucleotide probes discriminated by NimbleGen algorithm are not necessarily specific. The experience on Combimatrix design showed that a single probe per transcript is reliable and validated well by Real-Time analysis. Therefore, NimbleGen microarray design with a single probe per transcript (the best probe per each transcript) (*nmgd1*) have been requested.

1.5 -Comparison Between NimbleGen and CombiMatrix Microarray Platforms

Microarray designs

The two developed NimbleGen microarray designs (*nmgd1* and *nmgd2*) were compared with the CombiMatrix microarray design based on a single specific replicate probe per transcript (*cmbd1*) to evaluate if it was possible to integrate the expression data from the two platforms. Thus, as described above the following microarray designs were available:

(i) one CombiMatrix microarray design with a single specific replicated probe (35-40 mer) per transcript (see Figure 8);

(ii) Two NimbleGen microarray designs based on single and multiple probes (60 mer) per transcript (see Figure 8).

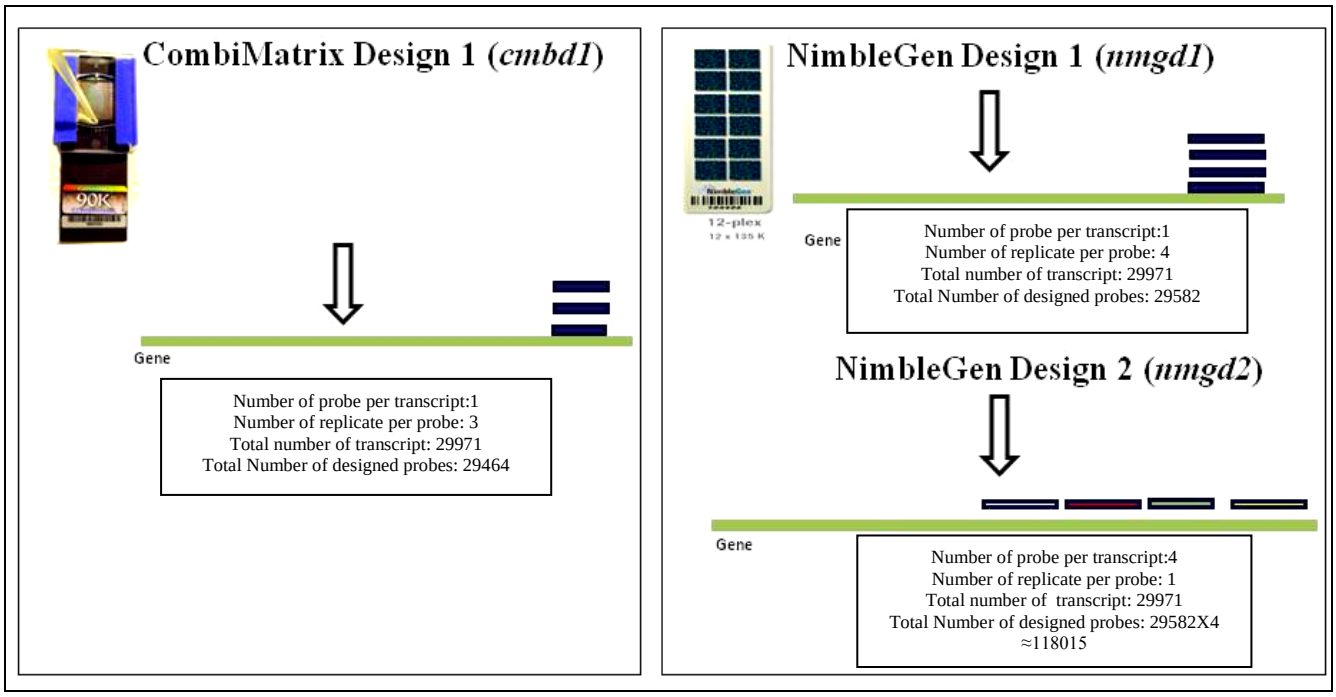

Figure 8: CombiMatrix microarray design with a single specific probe per transcript(left) and both NimbleGen microarray designs based on a single replicate and multiple probes per transcript (right).

Experimental Design

The three developed *cmbd1*(single specific replicate probe per transcript), *nmgd1*(best single replicate probe per transcript) and *nmgd2* (multiple probes per transcript) microarray designs have been compared in differential analysis by monitoring gene expression during veraison and ripening (two *Vitis vinifera* berry development stages) by analyzing samples collected in 2008. Each sample was processed in 3 technical replicates for both CombiMatrix and NimbleGen microarray hybridization (*cmbd1*, *nmgd1* and *nmgd2* microarray hybridization). Figure 9 summarizes the experimental design for this analysis.

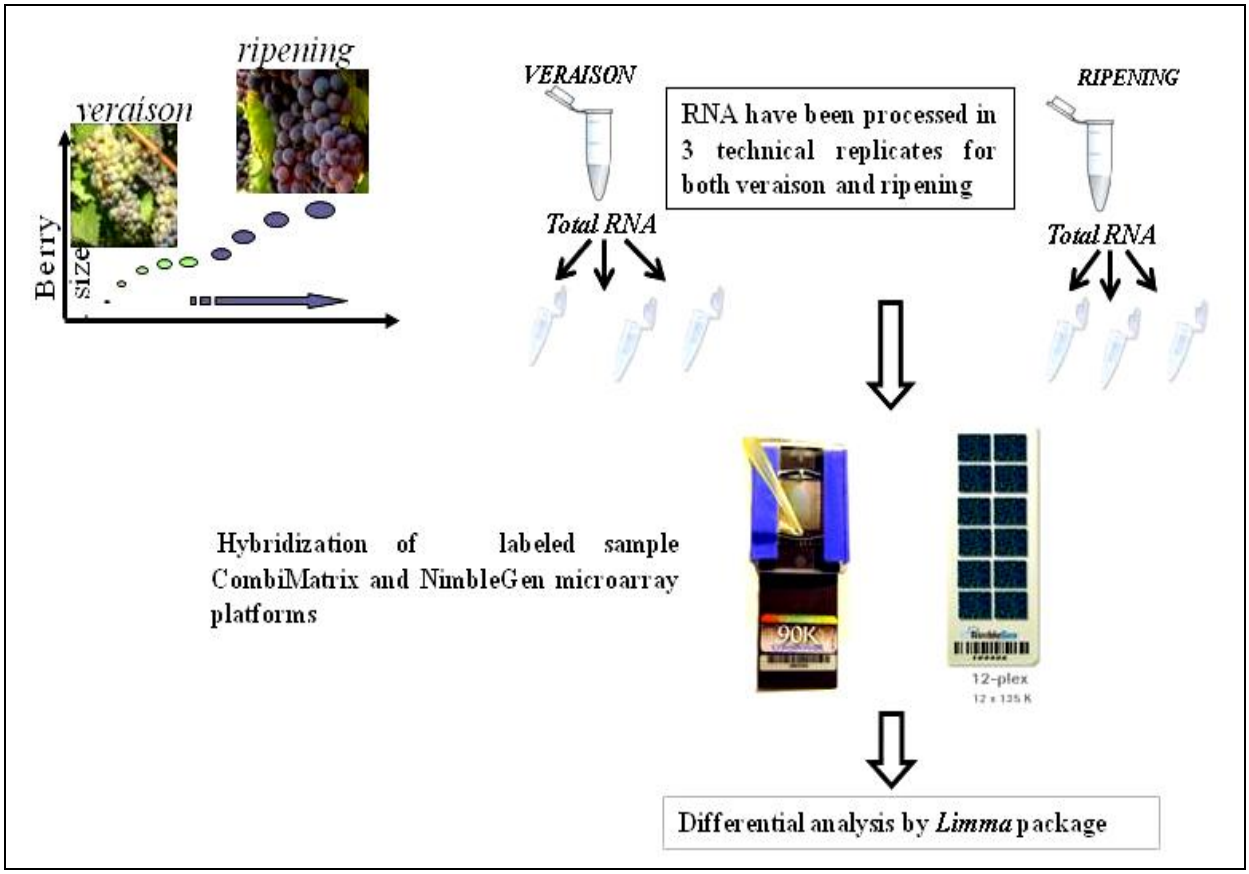

Figure 9: experimental design to assess the performance of NimbleGen microarray design based on a single replicate probe per transcript (*nmgd1*), NimbleGen microarray design based on multiple probes per transcript (*nmgd2*) and CombiMatrix microarray design with a single specific replicate probe per transcript (*cmbd1*).

Microarray Data pre-processing and normalization

Data pre-processing and intensity data normalization have been done as following:

(i) NimbleGen microarray design based on a single replicated probe per gene (*nmgd1*): Linear model microarray data analysis (limma) for background subtraction and quantile normalization. Probe replicates were averaged by median parameter.

(ii) NimbleGen microarray design based on different probes per transcript (*nmgd2*): Robust Multichip Average (RMA) module from NimbleScan v.2.5 for background subtraction and quantile normalization. Robust Multichip Average (RMA) module from NimbleScan software v.2.5 as described by Irizarry et al. (*Biostatistics* 2003; 4:249) also calculates a single gene call for each probe-set.

(iii) CombiMatrix microarray design based on a single specific replicated probe per transcript (*cmbd1*): Linear model microarray data analysis (limma) for background

subtraction and quantile normalization. Probe replicates were averaged by median parameter.

1.5.1. INTRA-PLATFORM DATA REPRODUCIBILITY

Intra-platform data repeatability was evaluated by calculating Pearson correlations among technical replicates. The following tables show the Pearson correlations values between the performed technical replicates. The three microarray designs exhibited a high intra-platform data reproducibility. Pearson correlation among technical replicate as indicated by Tables 3, 4 and 5 ranged from 0.95 to 0.99.

Table 3: Pearson correlation among technical replicate: NimbleGen microarray design 1 (*nmgd1*)

	ripening1	ripening2	ripening3	veraison1	veraison2	veraison3
ripening1	1					
ripening2	0.990588	1				
ripening3	0.992658	0.988461	1			
veraison1	0.862351	0.865495	0.869951	1		
veraison2	0.830671	0.832714	0.84424	0.957101	1	
veraison3	0.880394	0.884941	0.886276	0.979299	0.948402	1

Table 4: Pearson correlation among technical replicate: NimbleGen microarray design 2 (*nmgd2*)

	ripening1	ripening2	ripening3	veraison1	veraison2	veraison3
ripening1	1					
ripening2	0.996083	1				
ripening3	0.996175	0.997528	1			
veraison1	0.864171	0.866549	0.867993	1		
veraison2	0.836214	0.839868	0.843954	0.970139	1	
veraison3	0.879004	0.880356	0.880812	0.98517	0.958374	1

Table 5: Pearson correlation among technical replicate: CombiMatrix microarray design 1 (*cmbd1*)

	ripening1	ripening2	ripening3	veraison1	veraison2	veraison3
ripening1	1					
ripening2	0.986192	1				
ripening3	0.958337	0.958018	1			
veraison1	0.822092	0.818085	0.775005	1		
veraison2	0.821652	0.820419	0.771841	0.990649	1	
veraison3	0.821375	0.816802	0.78545	0.989307	0.986256	1

1.5.2. INTER- PLATFORM DATA REPRODUCIBILITY

In order to assess difference or similarities between the analyzed microarray designs based on CombiMatrix and NimbleGen microarray platforms, the only gene common to the three platforms were used in the comparison. NimbleGen and CombiMatrix microarray data were processed independently for the statistical analysis using the limma package (*Gordon K. Smyth et al.*, 2010) but the same procedure was applied. Differential expression ratios log2-fold change were compared between platforms to define cross-platform correlation (see table 6 below). Both NimbleGen microarray designs based on single replicate and multiple probes per transcript exhibited low correlation between them in differential analysis. It has been also observed that CombiMatrix microarray design displayed a very low correlation with both NimbleGen microarray design suggesting a low agreement between the two platforms.

Table 6: Person correlation between the three microarray designs in fold change profile

	Nmgd1 microarray design	*Nmgd2* microarray design	*Cmbd1* microarray design
NimbleGen microarray design based on the best single replicate probe per transcript (*nmgd1*)	--		
NimbleGen microarray design 2 based on different probes per transcript (*nmgd2*).	0.677	--	
CombiMatrix microarray design based on a single specific replicate probe per transcript (*cmbd1*).	0.292	0.323	--

1.5.3. DIFFERENTIAL ANALYSIS

Data coming from all the microarray datasets were analyzed with *limma* package. A false discovery rate (FDR) threshold of 5% and a fold change ratio (FC) >2 were chosen to discriminate significantly differentially expressed genes (DEG). The three

microarray designs (*cmbd1*, *nmgd1* and *nmgd2* microarray designs) called for a different number of significantly differentially expressed genes (DEG) as following:

(i) NimbleGen microarray design based on a single replicate probe per transcript (*nmgd1*) : 3781 genes.

(ii) NimbleGen microarray design with multiple probes per transcript (*nmgd2*): 7171 genes.

(iii) CombiMatrix microarray design based on a single specific replicate probe per gene (*cmbd1*): 5923 genes.

Moreover, NimbleGen microarray design based on different probes per transcript (*nmgd2*) called a higher number of significantly differentially expressed genes than NimbleGen microarray design based on a single replicate probe per transcript (*nmgd1*). Table 7 summarizes the overlap between genes called as differentially expressed form the three analyzed microarray designs. A limited overlap of significantly differentially expressed genes (DEG) was observed among the three microarray designs.

Table 7: overlap between gene called as significantly differentially expressed

	Nmgd1 microarray design	*Nmgd2* microarray design	*Cmbd1* microarray design
Nmgd1 microarray design	--	2371	1299
Nmgd2 microarray design	2371	--	1870
Cmbd1 microarray design	1299	1870	--

1.5.4. SINGLE PROBE DATA IN *nmgd2* MICROARRAY DESIGN

Analysis of *nmgd2* microarray design based on 4 different probes per transcript showed that for 6538 out of 29582 targets (or genes), the 4 probes show a different

behaviour. Moreover, 4146 of these 6538 targets (or genes) were detected as differentially expressed. 4909 genes out of 7171 significantly differentially expressed genes called by NimbleGen microarray design 2 (*nmgd2*) show a modulation not in agreement with NimbleGen microarray design 1 with a single replicate probe per transcript (*nmgd1*). In addition, for 3002 out this 4909 gene the 4 probes behave differently among them. 1807 out of this 3002 genes share a probe with NimblgeGen microarray design 1 based on the best single replicate probe per gene (*nmgd1*). It has been observed that, 1736 out of the 1807 probes shared with *nmgd1* behave in agreement with *nmgd1*.This analysis suggested that for a portion of significantly differentially expressed genes called by *nmgd2* microarray design probes on probe set behave differently among them.

Results summary

CombiMatrix and NimbleGen microarray platform/designs comparison analysis showed that:

(i) A direct integration of data between the two developed NimbleGen and CombiMatrix microarray designs is apparently not feasible.

(ii) Little overlap is observed between the significantly differentially expressed genes (DEG) generated by the developed CombiMatrix and NimbleGen microarray designs.

(iii) Moreover even the same platform (both *nmgd1* and *nmgd2* microarray designs based on the same platform) showed different results by changing the number of probes per target.

(iv) NimbleGene microarray design with different probes per transcript (*nmgd2* microarray design) discriminated a higher number of significantly differentially expressed genes (DEG).

2- AIM OF THE WORK

Transcriptional profiling using microarray technology is a powerful genomic tool that is widely used to characterize biological systems. Despite the increasing reliance on this technology by the scientific community, the issue concerning the reproducibility of microarray data between laboratories and across platforms has yet to be fully resolved. The issues of data reproducibility and reliability is crucial to the generation of, and ultimately to the utility of, large database of microarray results. Several consortiums such as Microarray Quality Control (MAQC) and Gene Expression Data (MGED) society have coordinated an impressive effort to develop guideline to assess the performance of different microarray technologies. The power of microarrays depends on the number, identity and quality of the probes on the array (*Joseph D. et al.*, 2006). Previous unpublished data produced by the Functional Genomics Centre of the University of Verona showed that CombiMatrix and NimbleGen microarrays designed on the same dataset of targets exhibited a small overlap of differentially expressed genes and a low correlation when the same samples were analyzed. Moreover, even the same platform (NimbleGen) showed considerable differences in the number of genes differentially expressed and in the correlation values when different numbers of probes were designed on each target gene.

In this thesis, we developed a CombiMatrix microarray design based on different probes per transcript in order to compare the two design strategies (single and multiple probes) based on the two NimbleGen and CombiMatrix microarray platforms (four chips).

Data obtained from the four arrays have been integrated to highlight the differences/similarities between the two design strategies and the two array platforms (NimbleGen and CombiMatrix) and to understand how the two design strategies vary depending on the array platforms.

Moreover, in this thesis, we set up and implemented a statistical methodology based on comparison of microarrays with a high-throughput sequencing-based transcriptomic platform to evaluate the performances and explain differences or similarities between the different design/platform combinations when used for differential expression analysis.

3- MATERIALS AND METHODS

3.1- CombiMatrix Microarray Oligonucleotide Design

The OligoArray 2.0 software (*Rouillard et al.*, 2003) was used to design 29464 and 83110 oligonucleotides for CombiMatrix design 1 and 2 respectively. Prior to running OligoArray 2.0, transcribed sequences of *Vitis vinifera* have been saved in a file in FASTA format. The designs were based on the 29971 annotations of the *Vitis vinifera* grape 12x assembly (*Jaillon et al*., 2007) using the following OligoArray 2.0 parameter settings:

- Oligonucleotide length range was set to 35 - 40 nt;
- Melting temperature (Tm) range was set to 80 - 86°C;
- GC content range was set to 40 - 60% according to the low GC content of the *Vitis vinifera* genome,
- Thresholds to reject oligonucleotides that can fold to form stable secondary structures and to start to consider putative cross-hybridizations were both set to 65°C;
- Oligonucleotides containing either AAAAA, TTTTT, GGGGG, CCCCC or longer homo-polymers were rejected;
- Maximum distance accepted between the 5’ end of an oligonucleotide and the 3’end of the input sequence: 1500bp;
- Maximum number of oligonucleotide per input sequence for *cmbd1* microarray design: 1;
- Maximum number of oligonucleotide per input sequence for *cmbd2* microarray design: 3;
- Number of replicated probe for *cmbd1*microarray design:3;
- Number of replicated probe for *cmbd2* microarray design:1;
- Minimum distance accepted between the 5’ end of two contiguous oligos for *cmbd2*:100.

3.2- NimbleGen Microarray Oligonucleotide Design

Probe selection strategies used for NimbleGen microarray transcript-base design is based on a scoring algorithm developed by Roche NimbleGen. These algorithm is based on different parameters such as: Melting temperature, probes uniqueness and frequency (cross-hybridization potential of the oligos). Under our request 29582 oligonucleotides for *nmgd1* and 118015oligonucleotides for *nmgd2* microarray designs have been designed by Roche NimbleGen. The designs were based on the 29971 annotations on *Vitis vinifera* grape 12x assembly (*Jaillon et al*., 2007) using the following design parameters:

- Oligo length: 60 nt;
- Maximum distance from 3'end: 1500 bp
- Melting temperature (Tm): 76 - 79 °C;
- GC content: 40 - 47 %;
- Thresholds to reject oligonucleotides that can fold to form stable secondary structures and to start to consider putative cross-hybridizations were both set to 65°C;
- Probes uniqueness is checked against grape transcripts database;
- On our request probe uniqueness was checked also against the grape genome assembly 12x sequences;
- Maximum number of oligonucleotide expected per input sequence for *nmgd1* microarray design: 1;
- Maximum number of oligonucleotide expected per input sequence for *nmgd2* microarray design: 4;
- Number of replicates per probe for *nmgd1* microarray design: 4;
- Number of replicates per probe for *nmgd2* microarray design: 1.

3.3- RNA Samples

RNA samples used for microarray hybridizations correspond to two stage of *Vitis vinifera* berry development, ripening and veraison, that have been analyzed by RNA-Seq in a previous work (*Zenoni et al.*,2010). In particular, *Vitis vinifera* berry clusters were collected at 10 and 15 week after flowering, corresponding at the veraison and

ripening stage respectively during the 2008 growing season, from a vineyard in the Verona province. The cluster were taken from three different plants at the vineyard's north site and three different plants on the south site. Ten berry were randomly selected from each cluster and pooled with berries from the other plant on the same vineyard site, resulting in two independent pools of 30 berries for each development stage. Total RNA was extracted form north and south samples at the two developmental stages using the method described by Zamboni et al. (*Zamboni et al*. 2008). North and south RNA samples were then pooled. The resulting two samples corresponding to ripening and veraison developmental stages were hybridized in triplicate.

3.4- Quality Control of RNA Samples

3.4.1. SPECTROPHOTOMETER RNA SAMPLES QUANTIFICATION

- The quantity of RNA was evaluated by a NanoDrop ND-1000 instrument (Thermo Scientific).
- The sample intensities along with the blank intensities are used to calculate the sample (nucleic acid) absorbance according to the following equation: **Absorbance = -log (Intensity$_{sample}$/Intensity$_{blank}$)**.
- **260/280** ratio of sample absorbance at 260 and 280 nm is used to assess the purity RNA or DNA (nucleic acid). A ratio of around 1.8 was accepted as pure for RNA. If the ratio is appreciably lower in either case, it may indicate the presence of protein, phenol or other contaminants that absorb strongly at or near 280 nm.
- **260/230** ratio of sample absorbance at 260 and 230 nm is a secondary measure of nucleic acid purity. A ratio of 1.8-2.2. was accepted. If the ratio is appreciably lower, this may indicate the presence of co-purified contaminants.

3.4.2. AGILENT 2100 BIO-ANALYZER ASSAY

The integrity of the RNA has been verified using the Agilent 2100 Bio-analyzer with the Eukaryote Total RNA assay, following manufacturer's protocol:

Pipette 550µl of RNA 6000 Nano gel matrix into a spin filter and centrifuge at 1500±20% rcf for 10 minutes at room temperature;

1. Vortex the RNA 6000 Nano dye concentrate for 10 seconds, spin down and add 1µl of dye into a 65µl aliquot gel of filtered gel;

2. Vortex the solution well and spin the tube containing the gel - dye mix at 13000 rcf for 10 minutes at room temperature;

3. Put a new RNA 6000 Nano chip on the chip priming station and pipette 9µl of gel-dye mix in the proper well;

4. Position the plunger of the syringe at 1 ml and then close the chip priming station;

5. Press the plunger until it is held by the clip. After 30 seconds exactly release clip and wait for more or less 5 seconds and slowly pull back plunger to 1ml position;

6. Open the chip priming station and pipette 9µl of gel-dye mix in the proper wells;

7. Pipette 5µl of RNA 6000 Nano marker in the 12 sample wells and in the ladder well ;

8. Heat the RNA samples and ladder for 2 minutes at 70°C then pipette 1µl of sample in each of the 12 sample wells;

9. Put the chip horizontally in the adapter of the IKA vortex and vortex for 1 minute at 2400 rpm followed by the running of the chip in the Agilent 2100 bio-analyzer within 5 minutes.

We considered RNA sample as intact for the microarray experiment when the integrity number (RIN) was major o equal 7 (for example sample A in figure 1).

Sample A: RIN =8 **Sample B: RIN =2.7**

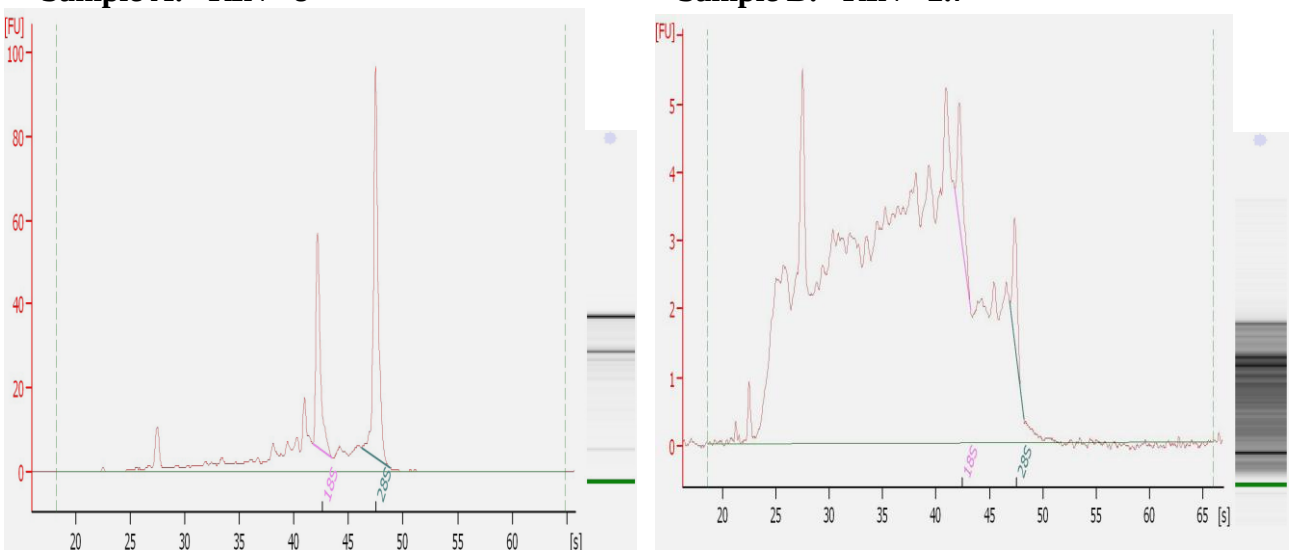

Figure 1: profile of an RNA *Vitis vinifera* berry sample run in Agilent 2100 bio-analyzer for an integrity control.

3.5- CombiMatrix Microarray Hybridizations

3.5.1. FIRST STRAND cDNA SYNTHESIS

1. Place a maximum volume of 10 µL of 2µg of total RNA into a nonstick, sterile, RNase-free, 0.2 mL tube.

2. Add 1 µL of T7 Oligo (dT) Primer;

3. Add Nuclease-free Water to a final volume of 12 µL, vortex briefly to mix, then centrifuge to collect the mixture at the bottom of the tube;

4. Incubate for 10 min at 70°C in a thermal cycler;

5. Centrifuge samples briefly to collect them at the bottom of the tube and place the mixture on ice;

6.At room temperature, prepare Reverse Transcription Master Mix in a nuclease-free tube (see table below);

Reverse Transcription Master Mix (for a single 20 µL reaction)

Amount	Component
2 µL	10X First Strand Buffer
4 µL	dNTP Mix
1 µL	RNase Inhibitor
1µL	ArrayScript

7. Transfer 8 µL of Reverse Transcription Master Mix to each RNA sample and mix

8. Place the samples in a 42°C incubator for 2 hours . After the incubation, centrifuge briefly to collect the reaction at the bottom of the tube;

9. Place the tubes on ice and immediately proceed to the second strand cDNA synthesis

3.5.2. SECOND STRAND cDNA SYNTHESIS

1. On ice, prepare a Second Strand Master Mix in a nuclease free tube in the order listed on the table below.

Second Strand Master Mix (for a single 100 µL reaction)

Amount	Component
63 µl	Nuclease Free Water
10 µl	10xSecond Strand Buffer
4 µl	dNTP Mix
2µl	DNase Polymerase
1 µl	RNase H

2. Mix well by gently vortexing. Centrifuge briefly to collect the Second Strand Master Mix at the bottom of the tube and place on ice;

3. Transfer 80µl of Second Strand Master Mix to each sample and mix**4**. Place the tubes in a 16°C thermal cycler. It is important to cool the thermal cycler block to 16°C before adding the reaction tubes because subjecting the reactions to temperatures >16°C will compromise aRNA yield;

4. Incubate for 2 hours in a 16°C thermal cycler. After the 2 hours incubation at 16°C, place the reactions on ice and proceed to cDNA Purification

3.5.3. cDNA PURIFICATION

1. Add 250µl of cDNA Binding Buffer to each sample and mix. Follow up with a quick spin to collect the reaction in the bottom of the tube. Proceed quickly to the next step;

2. Pipette the cDNA sample\cDNA Binding Buffer onto the center of the cDNA Filter Cartridge;

3. Centrifuge for 1 minute at 10,000 x g, and discard the flow-through and replace the cDNA Filter Cartridge in the wash tube;

4. Apply 500 µl Wash Buffer to each cDNA Filter Cartridge, centrifuge for 1 minute at 10,000 x g;

5. Discard the flow-through and spin the cDNA Filter Cartridge for an additional minute to remove trace amounts of Wash Buffer;

6. Transfer cDNA Filter Cartridge to a cDNA Elution Tube and apply 10µl of Nuclease-free Water (preheated to 50-55°C) to the center of the filter in the cDNA Filter Cartridge;

7. Leave at room temperature for 2 minutes and then centrifuge for 2 minutes at 10,000 x g;

8. Elute with a second 10µl of preheated Nuclease-free Water. Proceed directly to In Vitro Transcription to synthesize aRNA.

3.5.4. aRNA SYNTHESIS

1. At room temperature, prepare an IVT Master Mix by adding the following reagents to a nuclease-free microfuge tube in the order listed on the table below.

IVT Master Mix for a single reaction (40µL)

40µl rxn	Component
16 µl	Double-stranded cDNA in Nuclease-free Water
16 µl	T7 rNTP mix (75mM)
4 µl	T7 10X Reaction Buffer
4µl	T7 Enzyme Mix

2. Mix well by gently vortexing. Centrifuge briefly to collect the IVT Master Mix at the bottom of the tube and place on ice;

3. Transfer IVT Master Mix to each sample and mix thoroughly

4. Once assembled, place the tubes at 37°C and incubate for 14 hours.

5. Stop the reaction by adding Nuclease-free Water to each aRNA sample to bring the final volume to 100 µl. Mix thoroughly by gentle vortexing and proceed to the aRNA purification step.

3.5.5. aRNA PURIFICATION

This purification removes enzymes, salts and unincorporated nucleotides from the aRNA.

1. Check to make sure that each IVT reaction was brought to 100 µL with Nuclease-free Water. Add 350 µl of aRNA Binding Buffer to each aRNA sample. Proceed to the next step immediately;

2. Add 250 µl of ACS grade 100% ethanol to each aRNA sample and mix;

3. Pipette each sample mixture onto the center of the filter in the aRNA Filter Cartridge and centrifuge for 1 minute at 10,000 x g;

4. Discard the flow-through and replace the aRNA Filter Cartridge back into the aRNA Collection Tube;

5. Apply 650µl Wash Buffer to each aRNA Filter Cartridge and centrifuge for 1 minute at 10,000 x g;

6. Discard the flow-through and replace the aRNA Filter Cartridge back into the aRNA Collection Tube;

7. Apply 650µl 80% ethanol to each aRNA Filter Cartridge and centrifuge for 1 minute at 10,000 x g;

8. Discard the flow-through and spin the aRNA Filter Cartridge for an additional 1 minute to remove trace amounts of ethanol and Wash Buffer;

9. Transfer Filter Cartridge(s) to a fresh aRNA Collection Tube and to the center of the filter, add 100 µl Nuclease-free Water (preheated to 50-60°C)

10. Leave at room temperature for 2 minutes and then centrifuge for 2 minutes at 10,000 x g;

11. The aRNA will now be in the aRNA Collection Tube in ~100 µl of Nuclease-free Water.

3.5.6. ASSESSING aRNA YIELD AND QUALITY

The concentration of aRNA was determined by measuring its absorbance at 260 nm and multiplying the A260 by the dilution factor and the extinction coefficient. (1 A260 = 40 µg RNA/mL) as displayed by the following equation:

$$\textbf{Conc. of nucleic acid (ng/µl)} = \frac{\textbf{(OD260 - corr. factor) * dilution factor * 40}}{\textbf{cuvette length (cm)}}$$

For all RNAs OD260/280 should be >1.9 and OD260/230 should be >2.1.

3.5.7. ULS LABELING PROCEDURE

ULS labeling process was performed using 6µg of aRNA

1. Take 6 µg of purified aRNA;
2. Add 6 µl of Cy5 ;
3. Add 2 µl volume of 10x Labeling solution and adjust with RNase-free water to final volume (20µl) and mix by pipetting;
4. Label sample by incubation for 45 minutes at 85ºC and then place samples on ice, spin down to collect contents of tube before and proceed with purification using the KREApure column.

3.5.7.1. Dye removal using KREApure columns

Removal of free ULS label using KREApure columns:

1. Resuspend column material by vortexing;
2. Loosen cap ¼ turn and snap off the bottom closure and place the column in a 2 ml collection tube;
3. Pre-spin the column for 1 minute at 20800 xg and discard flow through and re-use collection tube;
4. Wash the column with 300 µl RNase free water and spin column for 1 minute at 20800 xg;

5. Discard collection tube and flow-through and put column in a new 1.5 ml micro-centrifuge tube;

6. Add ULS-labeled aRNA on to column bed, careful not to pipette on the sides of the column but directly on the column material;

7. Spin column for 2 minutes at 20800 xg. The flow through is purified labeled aRNA. At this point the degree of labeling (DOL) can be measured (see below)

3.5.7.2. Determination of the Degree of Labeling (DOL)

The degree of labeling (DOL) of Cy5-ULS labeled aRNA was determined according to the following equations:

- **Conc. of nucleic acid (ng/µl) = $\frac{(OD260 - (ODdye * corr.\ factor)) * dilution\ factor * 40}{cuvette\ length\ (cm)}$**

- **Degree of Labeling %(DoL) = $\frac{340 * Conc.\ of\ dye\ (pmol/\mu l) * 100\%}{Conc.\ of\ nucleic\ acid\ (ng/\mu l) * 1000}$**

A %(DoL) higher than 2 was accepted.

3.5.8. aRNA FRAGMENTATION

1. Prepare 5x RNA Fragmentation Solution (see table below);

Reagent	Volume for 10 ml	Final concentration
1M Tris Acetate pH 8.1	2ml	200mM
KOAc	0.49g	500mM
MgOAc	0.32g	150mM
Water	To 10 ml	
Total volume	10 ml	

2. Set up the RNA fragmentation reaction (see table below) using 4µg of the labeled aRNA for each hybridization CustomArray 90k microarrays

Reagent	Volume for 10 ml
Nuclease-free water + 4µg of Labeled aRNA	20.8µl
5x RNA Fragmentation Solution	5.2µl
Total volume	26 µl

3. Incubate at 95°C for 20 minutes. Place on ice.

3.5.9. COMBIMATRIX CUSTOMARRAY 90K ASSEMBLY

1.Align the Hybridization Cap over the slide so that the top edge of the slide is flush against the stop on the Hybridization Cap, and the Cap is centered over the semiconductor area;

2. Secure the Hybridization Cap in place with the Clips provided.

3.5.10. PRE-HYBRIDIZATION

1. Prepare fresh Pre-hybridization solution (see table below);

Reagent	For 120µl volume Cap	Final concentration
2x Hybridization Solution Stock	60 µl	6x SSPET, 0.05 Tween-20, 20mM EDTA
Nuclease-free water	41 µl	
50x Denhardt's Solution	12µl	5x
Salmon sperm DNA (10mg/ml)*	1 µl	100ng/µl
1%SDD	5 µl	0.05%
Total volume	120 µl	

*Heat Salmon sperm DNAsolution to 95°C for at least 5 minutes and then place in ice at least 1 minute before use.

2. Fill the hybridization chamber with nuclease-free water. Incubate at 65°C for 10 minutes. and then aspirate the water out of the hybridization chamber;

3. Fill the hybridization chamber with the Pre-hybridization Solution. Mix gently by pipetting. A small air bubble can be introduced to improve the mixing process;

4.Load the microarray onto the rotisserie in the hybridization oven and incubat at 45°C;

3.5.11. HYBRIDIZATION

1. Prepare hybridization solution (see table below)

Reagent	For 120µl volume Cap	Final concentration
2x Hybridization Solution Stock	60 µl	6x SSPET, 0.05 Tween-20, 20mM EDTA
DI Foramide (for RNA target only)	30 µl	25%
Labeled targets 2 to 8µg per sample	Varies (up to 24 µl)	20-80 ng/µl recommended
Salmon sperm DNA (10mg/ml)*	1 µl	100ng/µl
1%SDD	5 µl	0.04%
Nuclease-free water	To 120 µl	
Total volume	120 µl	

*Salmon sperm DNA solution should be heat-denatured at 95°C for 5minutes (as for preparation of the pre-hybridization solution).

2. Add the entire above prepared Fragmentation Reaction volume to 104µl of the prepared hybridization solution;

3. Denature the hybridization solution at 95°C for 3 minutes and then cool for 1 minute on ice;

4. Spin down the solution in a micro-centrifuge for 5 seconds at maximum speed to collect condensate;

5. Pipet the Pre-hybridization Solution out of the hybridization chamber;

6. Fill the hybridization chamber with the Hybridization Solution and mix gently with repeated pipetting. A small air bubble can be introduced to improve the mixing process if the microarray is rotated during hybridization;

7. load the microarray onto the rotisserie in the hybridization oven and incubate at 45°C the for 16 hours with gentle rotation;

3.5.12. HYBRIDIZATION WASHING

1.Preparation of Wash solutions:

	Solutions	For 10ml
6xSSPET Wash	6x SSPE, 0.05% Tween-20	3 ml 20X SSPE, 50 µl 10% Tween-20, 6.95 ml Nuclease-free water
3xSSPET Wash	3x SSPE, 0.05% Tween-20	1.5 ml 20X SSPE, 50 µl 10% Tween-20, 8.45 ml Nuclease-free water
0.5xSSPET Wash	0.5x SSPE, 0.05% Tween-20	250 µl 20X SSPE ,50 µl 10% Tween-20 ,9.7 ml Nuclease-free water
PBST Wash	2x PBS, 0.1% Tween-20	2 ml 10X PBS, 100 µl 10% Tween-20, 7.9 ml Nuclease-free water
PBS Wash	2xPBS	2 ml 10X PBS, 8 ml Nuclease-free water

1. Prior to starting the wash procedure, preheat the 6xSSPET wash Solution at 45° C;

2. Remove the microarray from the hybridization oven. Pipet the Hybridization Solution out of the chamber;

3. Using the pre-heated 6xSSPET Wash Solution, rinse the hybridization chamber, fill the chamber, and return the microarray to the hybridization oven for 5 minutes (with gentle rotation). Remove the 6xSSPET Wash Solution from the hybridization chamber;

4. Using the 3xSSPET Wash Solution, rinse the hybridization chamber, fill the chamber, and incubate the microarray at room temperature for 1 minute . Remove the 3xSSPET Wash Solution from the hybridization chamber;

5. Using the 0.5xSSPET Wash Solution, rinse the hybridization chamber, fill the chamber, and incubate the microarray at room temperature for 1 minute. Remove the 0.5xSSPET Wash Solution from the hybridization chamber.

6. Using the PBST Wash Solution, rinse the hybridization chamber, fill the chamber, and incubate the microarray at room temperature for 1 minute. Remove the PBST Wash Solution from the hybridization chamber;

7. Remove the PBST Wash Solution from the hybridization chamber;

8. Using the PBS Wash Solution, rinse the hybridization chamber, fill the chamber, and incubate the microarray at room temperature for 1 minute. Remove the PBS Wash Solution from the hybridization chamber. Repeat a second time;

9. Retain the PBS Wash Solution in the hybridization chamber until you are ready to proceed to the Imaging step.

3.5.13. IMAGING OF COMBIMATRIX CUSTOMARRAY 90K

1. Remove the PBS Wash Solution from the hybridization chamber;

2. Carefully remove the hybridization chamber from the microarray by removing the clips and lifting the Hybridization Cap off the slide surface;

3. Immediately cover the semiconductor microarray surface with around 90µl of Imaging Solution.

4. Using thin-tipped forceps, pick up a fresh Lifter-Slip and hold it so that the raised edges face the microarray. The raised edges can be detected by gently rubbing an edge with the tip of the forceps the raised edge will feel rougher than the glass surface;

5. Lay the LifterSlip at an angle onto the microarray so that it is centered over the semiconductor area (see Figure 2). First touch the Imaging Solution with one side of

the LifterSlip, then slowly lower the slip down, taking care not to introduce air bubbles;

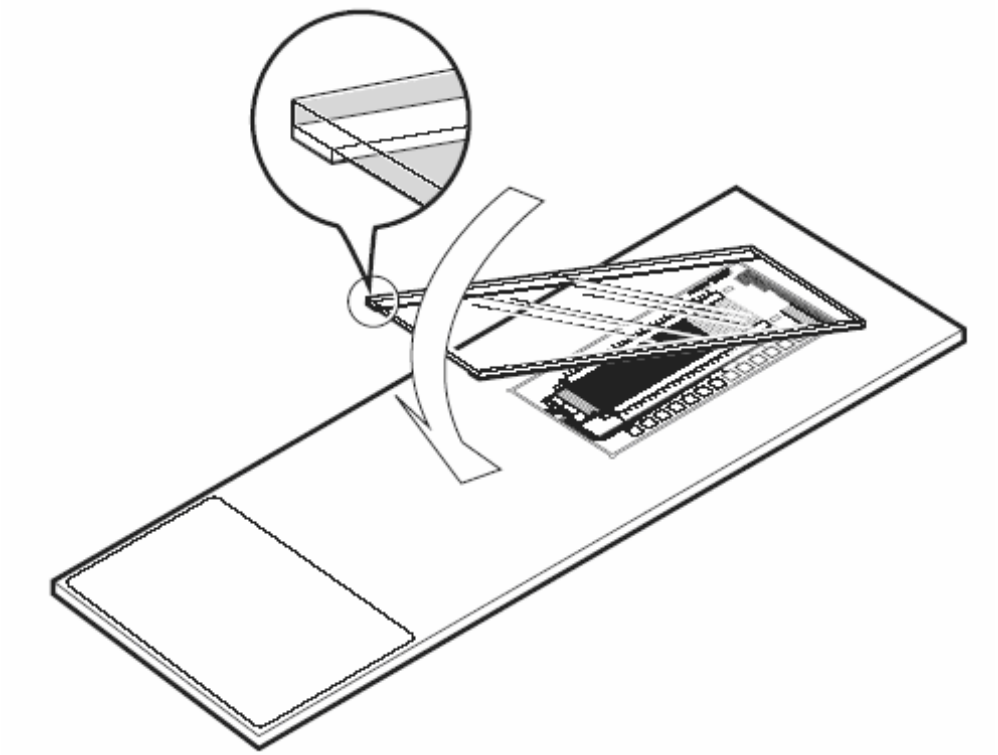

Figure 2: CombiMatrix 90K with Lifter-Slip cover-slip

6. Carefully remove any excess Imaging Solution from the edge of the LifterSlip using a lint-free tissue, until it is resting evenly over the microarray;

7. Load the CustomArray microarray into the scanner, taking care not to disturb the LifterSlip coverslip.

3.5.14. IMAGE SCANNING PARAMETERS AND DATA EXTRACTION

1. We performed CombiMatrix image scanning using an Axon Instruments GenePix 4200A scanner , by setting the following parameters:

- Wavelength= 635 (Cy5);
- PMT Gain=450 (could be increased if the normalized count at the 65.000 intensity is less than 1e-6) ;
- Power(%)=33,
- Pixel size=5;
- Lines to average=1

- Focus position(μm)= 100.

2. Data has been extracted using the software Gene Pix Pro 7.

3.6 -NimbleGen Microarray Hybridization Experiment

3.6.1. SAMPLE PREPARATION

3.6.1.1. First Strand cDNA Synthesis

We used the Invitrogen SuperScript Double-Stranded cDNA Synthesis Kit to synthesize double-stranded cDNA.

1. Combine components in a 0.2ml tube on ice according to the following table:

	Total RNA Amout
RNA	10μg
Oligo dT Primer	2μl
DEPC Water	To volume
Total	11μl

2. Heat each sample to 70°C for 10 minutes in a thermo-cycler. Briefly spin tubes in a micro-centrifuge and place them in an ice-water slurry for 5 minutes;

3. Add the following to each sample tube;

	Volume
The above reaction volume	11μl
5x First Strand Buffer	4μl
0.1M DTT	2μl
dNTP mix	1μl
Total volume	18μl

4. Mix gently and briefly spin the tubes in a micro-centrifuge and place samples in a thermo-cycler set at 42°C for 2 minutes.

5. Add 2 µl of SuperScript II RT and mix gently and, then incubate the samples at 42°C for 60 minutes;

8. Briefly spin the tubes in a micro-centrifuge. Place the samples on ice until the second strand synthesis.

3.6.1.2. Second Strand cDNA Synthesis

1. Add the following components to the above first strand reaction(s) in the indicated order;

Components	Volume
First strand cDNA reaction	20µl
DEPC Water	90.8µl
5x Second Strand Buffer	30µl
10mM dNTP Mix	3µl
10U/µl DNA Polymerase I	4µl
2U/µl RNAse H	0.2µl
Total	150µl

2. Mix gently and briefly spin the tubes in a micro-centrifuge and then incubate at 16°C for 2 hours;

3. Add 2 µl of 5 U/µl T4 DNA polymerase to each reaction. Incubate at 16°C for an additional 5 minutes;

4. Place the samples on ice and add 10µl of 0.5M EDTA. Proceed with the RNase A Cleanup step.

3.6.1.3. RNase A Cleanup

1. Add 1 µl of 4 mg/ml RNase A solution to the tubes from above Step 4;

2. Mix gently and briefly spin the tubes in a micro-centrifuge and then incubate samples at 37°C for 10 minutes;

3. Add 163 µl of phenol:chloroform:isoamyl alcohol to one set of 1.5 ml centrifuge tubes;

4. Transfer the samples to the tubes containing phenol:chloroform:isoamyl alcohol. Vortex well;

5. Centrifuge at 12,000 x g for 5 minutes and transfer the upper, aqueous layer to a clean, labeled 1.5 ml tube.

3.6.1.4. cDNA Precipitation

1. Add 16 µl (0.1 volume of the above step 5) of 7.5 M ammonium acetate to the samples. Mix by repeated inversion. Briefly spin the tubes in a micro-centrifuge;

2. Add 7 µl of 5 mg/ml glycogen to the samples. Mix by repeated inversion. Briefly spin the tubes in a micro-centrifuge;

3. Add 326 µl (2 volumes of above Step 3) of ice-cold absolute ethanol to the samples. Mix and centrifuge at 12,000 x g for 20 minutes;

4. Removed supernatant taking care not to disturb the pellet and add 500 µl of ice-cold 80% ethanol (v/v) and mix;

5. Centrifuge tubes at 12,000 x g for 5 minutes and removed supernatant.;

6. repeat steps 4 – 5 and dry the pellet in a DNA vacuum concentrator;

7. Rehydrate samples with 20 µl of VWR water and quantify the cDNA amount.

3.6.1.5. Spectrophotometric QC of cDNA

- Quantitate each cDNA sample according to the following formula:

cDNA Concentration (µg/ml) = A260 x 50 x Dilution Factor and verify that all samples meet the following requirements: sample concentration ≥ 100ng/µl, 260/A280 ≥ 1.8 and 260/A230 ≥ 1.8.

- Bioanalyzer-QC of cDNA

1.Transfer 250 ng cDNA to a sterile micro-centrifuge tube.

2. Analyze the samples using the Agilent Bioanalyzer and RNA 6000 Nano Kit. verify that all samples meet the following requirement for acceptance:

- Median size ≥ 400 bp when compared to a DNA ladder.

3.6.2. SAMPLE LABELING

We label the cDNA sample using the NimbleGen One-Color DNA Labeling Kit.

1. prepare the following solution;

Random Primer Solution	Amount
Random primer Buffer	998.25µl
β-Mercaptoethanol	1.75µl
Total	1ml

2. Briefly centrifuge Cy3-Random Nonamer and dilute the primer in 924 µl of Random Primer Buffer with β-Mercaptoethanol. Aliquot 40 µl individual reaction volumes in 0.2 ml thin-walled PCR tubes and store at -20°C, protected from light;

3. Assemble the following components in separate 0.2 ml thin-walled PCR tubes:

cDNA	1µg
Diluted Cy3-Random Nonamers from step 2	40 µl
Nuclease-free water	To volume
Total	80µl

4. Heat-denature samples in a thermo-cycler at 98°C for 10 minutes. Quick-chill in an ice-water bath for 2 minutes;

5. Prepare the following dNTP/Klenow master mix for each above prepared samples;

10mM dNTP Mix	10µl
Nuclease-free water	8µl
Klenow Fragment (3'->5' exo-) 50U/µl	2µl
Total	20µl

6. Add 20 µl of the dNTP/Klenow master mix prepared in step 5 to each of the denatured samples prepared in step 4. Keep on ice;

Reaction volume from step 4	80µl
dNTP/Klenow Master Mix	20µl
Total	100µl

7. Mix well and quick-spin to collect contents in bottom of the tube;

8. Incubate at 37°C for 2 hours in thermo-cycler with heated lid, protected from light;

9. Stop the reaction by addition of 10µl of Stop Solution (0.5M EDTA);

10. Add 11.5µl 5M NaCl to each tube. Vortex briefly, spin, and transfer the entire to a 1.5ml tube containing 110µl isopropanol;

11. Vortex well. Incubate for 10 minute at room temperature, protected from light;

12. Centrifuge at 12,000xg for 10 minutes. Remove supernatant with a pipette. Pellet should be pink;

13. Rinse pellet with 500 µl 80% ice-cold ethanol. Dislodge pellet from tube wall by pipetting a few times;

14. Centrifuge at 12.000xg for 10 minutes. Remove supernatant with a pipette and dry content in a SpeedVac on low heat until dry, protested from light;

15. Spin tubes briefly prior to open. Rehydrate pellets in 25µl Nuclease-free water per reaction;

16. Let at room temperature protected from light, for approximately 5 minute, then vortex and quick-spin;

17. Quantitate each sample using the NanoDrop. Based on the concentration, calculate the volume of Cy3-labeled cDNA sample required for each hybridization (4µg);

18.Dry in a SpeedVac on low heat, protected from light.

3.6.3 -HYBRIDIZATION AND WASHING

3.6.3.1. Prepare Sample

1. Set the Hybridization System to 42°C. With the cover closed, allow at least 3 hours for the temperature to stabilize;

2. Each sample to be hybridized to a12x135K array should be resuspended in a unique STC (3.3µl). Record which STC is used for each sample;

3. Vortex well and spin to collect content in bottom of the tube;

4.Using components from the NimbleGen Hybridization Kit, prepare the hybridization solution master mix according to the following table;

Hybridization Solution Master Mix to Hybridize a Single Slide	Amount 12x135K Array
2x hybridization Buffer	88.5µl
Hybridization Component A	35.4µl
Alignment Oligo	3.6µl
Total	127.5µl

5. Add 8.7µl of hybridization solution to each sample, vortex well and spin to collect contents in bottom of the tube and incubate at 95°C for 5 minutes, protected from light;

6. Place tubes at 42°C until ready for sample loading. Vortex prior to loading.

3.6.3.2. Prepare Mixers

1. Position the Precision Mixer Alignment Tool (PMAT) with its hinge on the left. Open the PMAT

2. Snap the mixer onto the two alignment pins on the lid of the PMAT with the tab end of the mixer toward the inside hinge and the mixer's adhesive gasket facing outward (Figure 3).

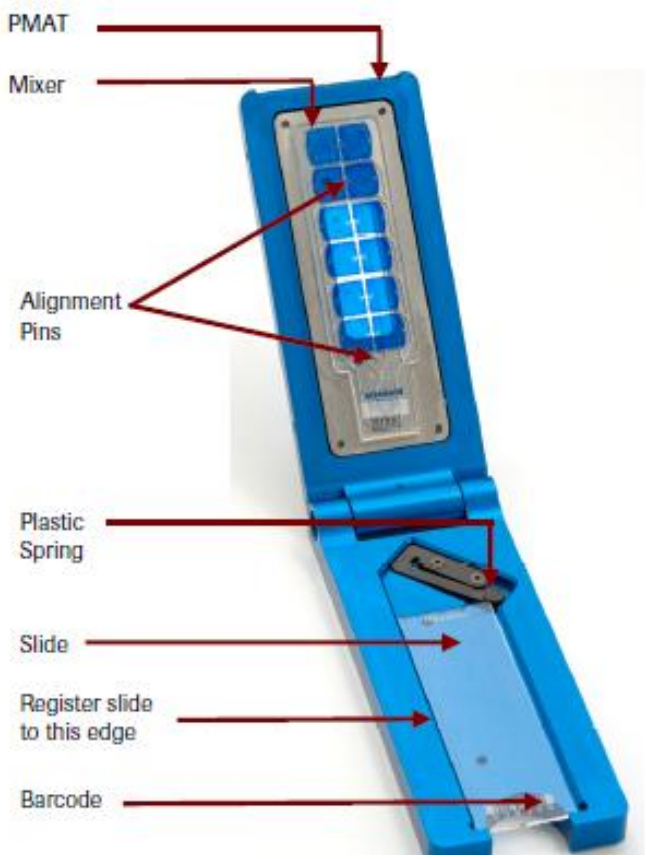

Figure 3: PMAT with HX12 Mixer and Slide. For photographic purposes only, blue coloring was used to show the location of the mixer's hybridization chambers. The hybridization chambers of the mixer you receive will not be blue.

3. While pushing back the plastic spring with a thumb, place the slide in the base of the PMAT so that the barcode is on the right and the corner of the slide sits against the plastic spring. The NimbleGen logo and barcode number should be readable. Remove your thumb and make sure the spring is engaging the corner of the slide and the entire slide is registered to the edge of the PMAT to the rightmost and closest to you. In addition, be sure that the slide is lying flat against the PMAT. Gently blow compressed nitrogen or argon gas across the mixer and slide to remove dust;

4. Using forceps, remove the backing from the adhesive gasket of the mixer and close the lid of the PMAT so that the gasket makes contact with the slide;

5. Lift the lid by grasping the long edges of the PMAT while simultaneously applying pressure with a finger through the window in the lid of the PMAT to free the mixer-slide assembly from the alignment pins;

6. Remove the mixer-slide assembly from the PMAT and place the mixer-slide assembly on the back of a 42°C heating block for 5 minutes to facilitate adhesion of the mixer to the slide.

7. Rub the Mixer Brayer over the mixer with moderate pressure to adhere the adhesive gasket and remove any bubbles. For HX12 mixers, first use a corner of the Mixer Brayer to rub the borders between the arrays and then rub around the outside of the arrays. The adhesive gasket will become clear when fully adhered to both surfaces;

8. Place the mixer-slide assembly in the slide bay of the Hybridization System.

3.6.3.3. Hybridization Step

1Slowly dispense the appropriate sample volume into the fill port. Load sample and seal mixer ports;

2. Close the bay clamp;

3. Turn on the Mixing Panel on the Hybridization System, set the mix mode to B, and press the mix button to start mixing. Confirm that the Hybridization System recognizes the slide in each occupied bay (its indicator light becomes green);

4. Hybridize samples at 42°C to the array(s) for 18 hours.

3.6.3.4. Wash Hybridized Array

1. Before removing the mixer-slide assemblies from the Hybridization System, prepare Washes I, II, and III according to the following tables. Note that you prepare two containers of Wash I.

Washing Multiple Slides	Wash I	Washes I,II, III
Water	135ml	247.5ml
10x Wash Buffer I ,II ,III	15ml	27.5ml
1M DTT	15ml	27.5ml
Total	150ml	275ml

2. To facilitate the removal of the mixer, heat the shallow dish containing Wash I to 42°C. Keep the remaining three wash solutions at room temperature;

3. Insert the Mixer Disassembly Tool into the shallow dish containing warm Wash I. we also insert a slide rack into the wash tank containing Wash I at room temperature;

4. Remove a mixer-slide assembly from the Hybridization System and load it into the Mixer Disassembly Tool immersed in the shallow dish containing warm Wash I;

5. With the mixer-slide assembly submerged, carefully peel the mixer off the slide. It is important to hold the Mixer Disassembly Tool flat while removing the mixer and to avoid any horizontal movement or scraping with the mixer across the slide;

6. Working quickly, discard the mixer and remove the slide from the Mixer Disassembly Tool;

7. Gently agitate the slide for 10 -15 seconds in the shallow dish containing warm Wash I to quickly remove the hybridization buffer;

8. Transfer the slide with the barcode at the top into a slide rack (Figure 10) in the wash tank that contains Wash I and Agitate vigorously for 10 - 15 seconds;

9. Wash for an additional 2 minutes in Wash I with vigorous (1 minute bar code top and 1 minute bar code down), constant agitation. If washing multiple slides, move the rack up and down with enough agitation to make foam appear;

10. Quickly blot the rack, several times using paper towels to minimize buffer carryover. Transfer the slide to Wash II and wash for 1 minute with vigorous, constant agitation clean the tops of the slide;

11. Transfer the slide to Wash III and wash for 15 seconds with vigorous, constant agitation;

12. Remove the slide from Wash III. Spin dry in a NimbleGen Microarray Dryer. For a NimbleGen Microarray Dryer, the recommended drying time is 2 minutes;

13. Remove the slide from the NimbleGen Microarray Dryer or other microarray dryer. Blot dry the edges to remove any residual moisture. Proceed immediately to the step for scanning the array.

3.6.4. IMAGE SCANNING PARAMETERS AND DATA EXTRACTION

1. Scanning was performed using an Axon GenePix 4400A scanner, by setting the following parameters:

- Wavelength= 532 (Cy3);
- PMT Gain=550 (could be increased if the normalized count at the 65.000 intensity are less than 1e-5) ;
- Power(%)=100;
- Pixel size=2.5;
- Lines to average=1;
- Focus position(μm) = 0.

3.Data extraction have been performed by *NimbleScan v2.5* software developed by Roche NimbleGen.

3.7- Microarray Data Pre-processing/Normalization

We obtained background-corrected of both CombiMatrix and NimbleGen microarray intensity data using Normexp_saddle (*Ritchie et al.*,2007) method provide by limma package from R bio-conductor . RMA (robust multi-chip average) provide the background correction algorithm for NimbleGen microarray data as implemented in the NimbleScan software (see Table below).

For CombiMatrix microarray design 1, CombiMatrix microarray design 2, NimbleGen microarray design 1 and NimbleGen microarray design 2 two different summarization methods of the normalized intensities of replicate probes have been tested: mean and median. For NimbleGen microarray design 2, the summarization of the normalized intensities of the different probes per transcript has been also performed by the Robust Multichip Average (RMA) algorithm from NimbleScan software (*Irizarry et al.*, 2003) (see table below). Probes from CombiMatrix and NimbleGen microarray designs based on different probes per transcript have been also analyzed separately (see table below).

Summary of all analyzed microarray data pre-processing:

Microarray design	Background correction	Probe average
NimbleGen design1 (*nmgd1*): single replicate probe per gene	Normexp-Saddle (*limma*)	Mean/median (*limma*)
NimbleGen design 2 (*nmgd2*): four different probes per gene	NimbleScan v.2.5 standard background correction (RMA)	RMA
NimbleGen design2 (*nmgd2*): four different probes per gene	Normexp-Saddle , (*limma*)	Mean/median (*limma*)
NimbleGen design2 (*nmgd2*): four different probes per gene	Normexp-Saddle (*limma*)	Probe have been analyzed separately
CombiMatrix design1 (*cmbd1*): single specific probe per gene	Normexp-Saddle (*limma*)	Mean/median (*limma*)
CombiMatrix design2 (*cmbd2*): three different probes per gene	Normexp-Saddle (*limma*)	Mean/median (*limma*)
CombiMatrix design2 (*cmbd2*): three different probes per gene	Normexp-Saddle (*limma*)	Probe have been analyzed separately

Both CombiMatrix microarray designs (*cmbd1* and *cmbd2*) intensity data have normalized using quantile method (*Bolstad et al*.,2003) provide by limma package from R bio-conductor.

NimbleGen microarray designs (*nmgd1* and *nmgd2*) intensity data have been normalized by quantile method (*Bolstad, et al*., 2003) provide by NimbleScan software.

3.8- Microarray and RNA-Seq Statistical Analysis

3.8.1. MICROARRAY STATISTICAL ANALYSIS

Analysis of differentially expressed genes for all analyzed microarray designs was performed using linear modeling and moderated statistical t-test based on the empirical Bayes methods, as implemented in the limma R package (*G.K. Smyth*, 2005). P-values were adjusted for multiple testing with the Benjamini and Hochberg method (1995).

Fisher's method which combines probability values from each test, have been used to calculate p-adjusted values of *cmbd2* and *nmgd2* microarray expression data.

3.8.2. RNA-Seq STATISTICAL ANALYSIS

The package *DESeq* (*Simon A. et al*, 2010) from R software was used to estimate the variance in RNA-Seq data and to test for differentially expression. The output file of this analysis returns a data frame with p values, p adjusted values, fold change values, logarithm transformation fold change values and the mean base values of each considered conditions.

3.9- Real Time RT-PCR

We design RT-PCR primer on gene region within 1 kb upstream of the 3'-end of each genes (see table below). As template for primer design we use the 12x grape genome assembly (*Jaillon et al.* ,2007).

Primer sequences

Forward primer sequences	Reverse primer sequences	Gene Names
TATCATGACATGCCAAATCG	ACTCGTCCACCGGTATTTG	JGVv301.10.t01
TTGAACCCAATGCCTTCA	CTCGCGTAAACATCCAACA	JGVv129.66.t01
CCCATTCTCGTTTTCTCAG	TATGGGGTTCTGGGAGACT	JGVv151.6.t01
CAGTAATGGCTCAGCAGGA	TGTGGAGGAACTTGGAGAG	JGVv4.362.t01
GCGATGAGAGTGTTCCAGT	CTCCCCTCAAACAGAGATG	JGVV61.51.t01
CTTCTCTCCTCATGGGCTCT	ACCCAGGGGAAGGTATATGA	JGVV44.63.t01
ACCACTGGCTCAAGGATTC	ACGAAAGAATAGCGGAGTTG	JGVV0.133.t01
CAAGGAGGATCACCTGATG	CTCCTGTAGATCCTGAACCA	JGVv0.270.t01
CACAGACGGAGGTGATCTT	GGAGTCGAGAGCCATCAG	JGVV20.224.t01
ACAAACTCGTACTCCCGAAT	TGTTTTAACAAGGCGGTATTC	JGVV1.1082.t01

3.9.1. DNAse REACTION

RNA samples have been treated with DNase using the Turbo DNA-free kit (Applied Biosystem) as following:

RNA Sample	5 µg
DNAse	1µl
Buffer 10x	2.5µl
Nuclease free Water to volume	25µl

1. Incubate at 37°C for 30 minutes in 500µl eppendorf;
2. Once incubation complete add 2.5µl DNase Inactivation Reagent and we incubate for 5 minutes at room temperature and then we precipitated the RNA treated by DNase by ethanol in order to remove any phenol contamination;
3. we gently mix and spin for 1.5 minute and transferred the supernatant in a new tube and control on agarose gel 0.8%

3.9.2. cDNA SYNTHESIS

Superscript II reverse Transcriptase Invitrogen kit for cDNA synthesis has been used for cDNA synthesis (3 different reactions have been performed for each considered berry development stage) as following: development stage) as following:

Mix For one reaction

Oligo dT 0.5mg/ml	1µl
RNA treated with DNAse	5µg
dNTPs Mix (10mM)	1µl
Nuclease free Water to volume	12µl

1. Incubate at 65°C for 5 minutes and then add 4µl of 5x First Strand Buffer + 2µl of 0.1M DTT + 2µl of water;
2. Incubate the above 20µl reaction at 42°C for 2 minutes and then add 1µl of SuperScript II RT at each reaction;

3. Incubation at 42°C for 50 minutes;

4. inactivation by incubation at 70°C for 15 minutes.

3.9.3. REAL TIME PCR REACTION

1.We performed RT-PCR reaction preparing the following reaction:

SYBR Green	12.5µl
Primer FOR. (10µM)	1µl
Primer REV. (10µM)	1µl
cDNA (1/10)	1/10 of total volume
Nuclease-free water	To 25µl final volume

2. perform 40 cycles of PCR amplification as follows:

Step	Temperature	Duration
Pre-heat	95°C	2 minutes
Denature	95°C	15 seconds
Anneal	55°C	30 seconds
extend	72°C	30 seconds

3.9.4. DATA ANALYSIS

Amplification efficiency was calculated from raw data using LingRegPCR software (*Ramakers et al*., 2003). The relative expression ratio value between ripening and veraison samples was calculated according to the Pfaffl equation (*Pfaffl MW*., 2001). The significance of difference in transcript levels between

ripening and veraison *Vitis vinifera* berry development was assessed by a classical t-test.

3.10- Accuracy, Sensibility, Specificity, and Positive Predictive Values (PPV)

		Number of differentially expressed genes by RNA-Seq	
		True	**False**
Number of differentially expressed genes by Microarray designs (test outcome)	**True**	**True positive**	**False positive**
	False	**False negative**	**True negative**

- Sensitivity : measures the proportion of actual positives which are correctly identified as such.

Sensitivity=Number of true positives÷ (Number of true positives +Number of false negatives)

- Specificity: measures the proportion of negatives which are correctly identified

Specificity=Number of true negatives÷(Number of true negatives + Number of false positives)

- In the fields of statistics, the accuracy of a measurement is the degree of closeness of measurements of a quantity to that quantity's actual (true) value.

Accuracy= (Number of true positives + Number of true negatives)÷ (Number of true positives + Number of false positives+ Number of false negatives + Number of true negatives)

- In statistics and diagnostic testing, the positive predictive value, or precision rate is the proportion of subjects with positive test results who are correctly diagnosed.

PPV= Number of true positives÷ Number of positives calls

3.11- ROC Curve Analysis

ROC (Receiver Operating Characteristic) curve analysis have been used to measure the ability of microarray platform in detecting differentially expressed genes assuming RNA-Seq as reference. ROC analysis have been performed using the R package ROC.

4- RESULTS

4.1- CombiMatrix Microarray Design Based on Different Probes per Transcript

(i) Probe design of CombiMatrix microarray design based on different probes per gene (*cmbd2*) was performed by OligoArray 2.0 software (*Rouillard et al.*, 2003) a programme that uses a thermodynamic approach to predict secondary structures and to calculate the specificity of probes. *Cmbd2* microarray design, was based on 29971 annotations of grape 12x genome assembly (*Jaillon et al.*, 2007) and included 3 different probes per transcript. A total number of 83110 probes for 29465 target gene models have been selected. A total, 25724 gene model transcripts were recognized by a probe set built by 3 unique probes, while 2197 gene model transcripts are recognized by a probe set built by 2 unique probes. 1544 gene model transcripts were recognized by a probe set built by a single replicate probe.

(ii) In order to determine the specificity of the designed probes a blast analysis (with no mismatches by setting Tm $\geq 85 \pm 15°$ C threshold) has been performed against all represented gene model transcripts and whole *Vitis vinifera* genome and the results have been reported in the following table:

Table 1: Number of gene models per probe with no mismatches for CombiMatrix design 2

Probe number	Number of gene model transcript matches
68936	1
6534	2
2630	3
5010	>3

For *cmbd2* microarray design based on multiple probes per transcript, 83.73% of selected probes were specific for the target gene models when compared with the

whole transcripts dataset, while 79.33% of selected probes were specific when compared to the whole genome.

Summary of the developed CombiMatrix and NimbleGen microarray designs

Figure 1 below shows the two microarray designs based on a single replicate probe per transcript and on multiple probes per transcript in CombiMatrix and NimbleGen microarray platforms (*cmbd1*, *cmbd2*, *nmgd1* and *nmgd2* microarray designs).

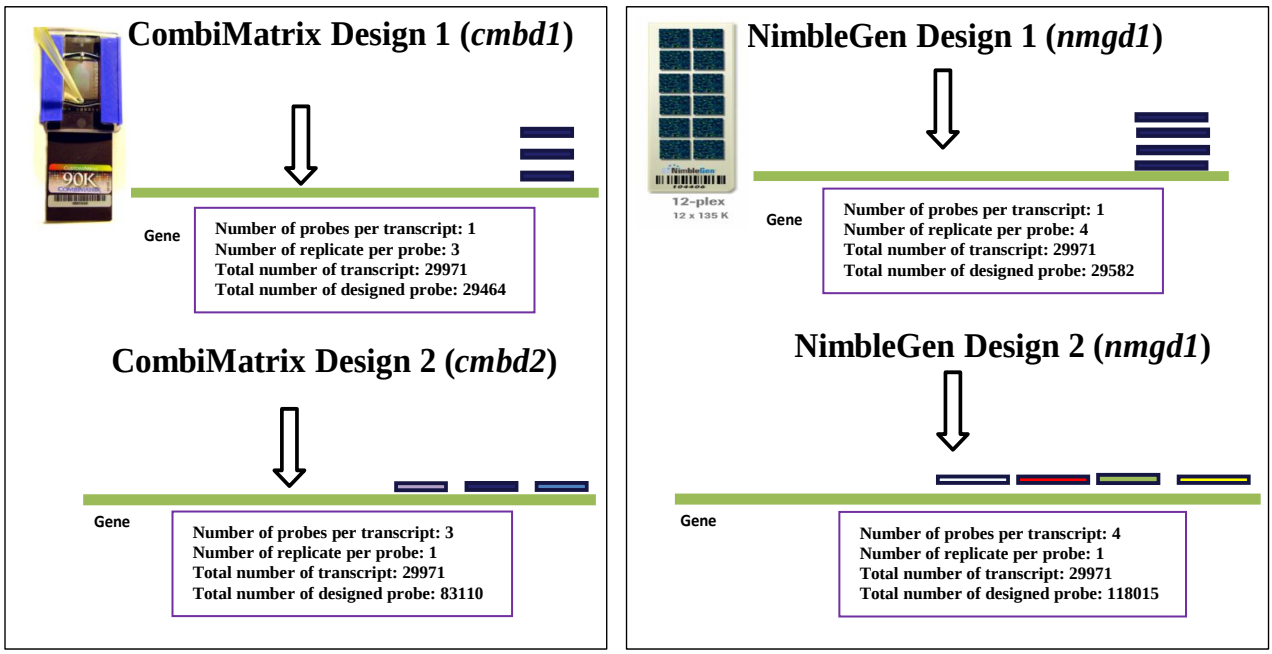

Figure 1: (left) CombiMatrix (*cmbd1* and *cmbd2*) and (right) NimbleGen (*nmgd1* and *nmgd2*) microarray designs based on a single replicate probe per transcript and on multiple probes per transcript.

Experimental Design for cmbd2 microarray design Hybridizations

The same ripening and veraison RNA samples used for previous microarray hybridization on *cmbd1*, *nmgd1* and *nmgd2* microarrays have been processed in 3 technical replicates on CombiMatrix microarray design based on multiple probes per transcript (*cmbd2*) (see Figure 2).

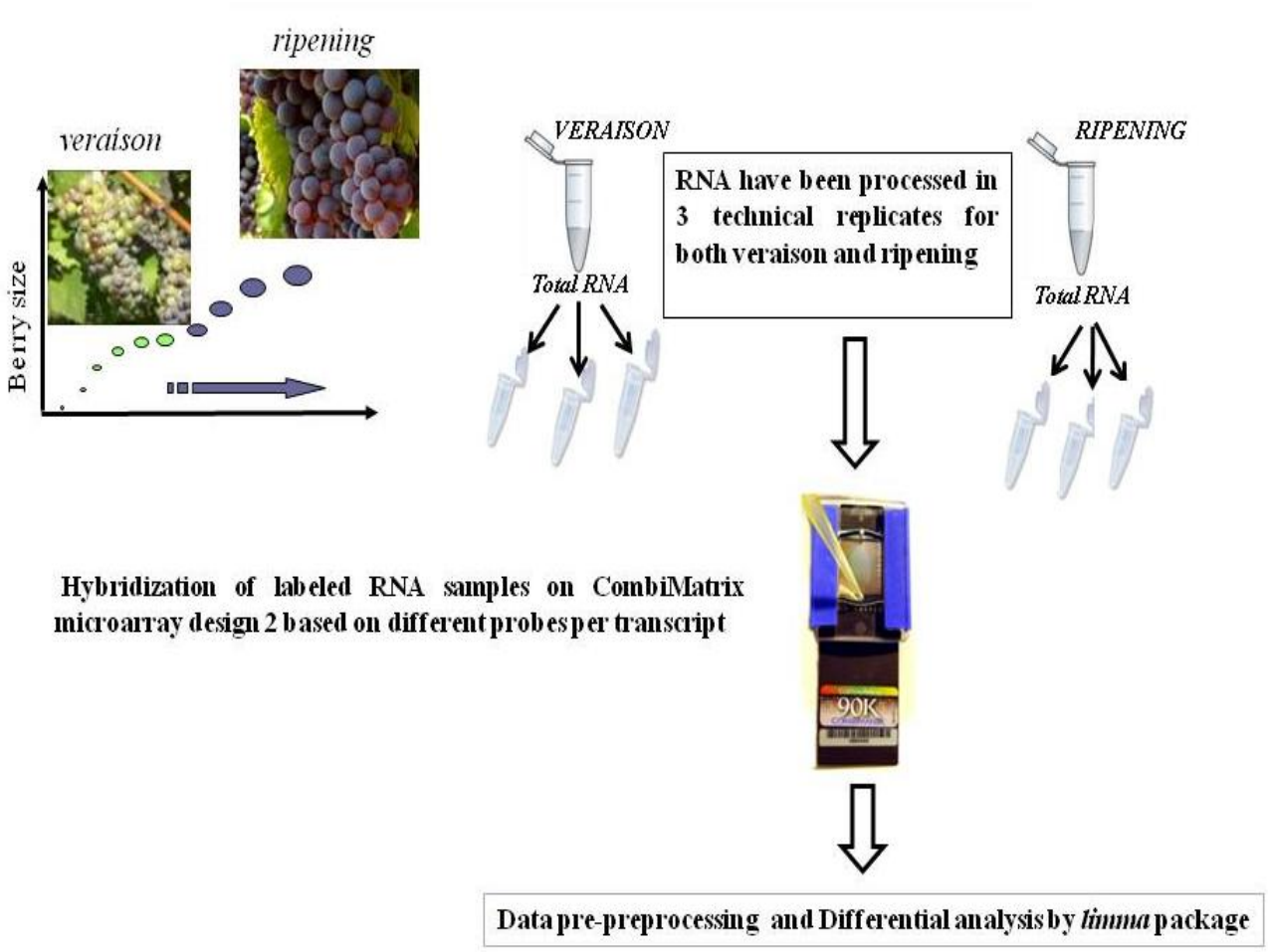

Figure 2: representation of the experimental design for the developed CombiMatrix microarray design 2 based on different probes per transcript (*cmbd2*) hybridization.

4.2- Quality Control of CombiMatrix Microarray Design 2 Hybridization

The precision of microarray detection is estimated by technical reproducibility. For *cmbd2* microarray design, parallel gene expression data were collected on each of the two considered total RNA samples (veraison and ripening *Vitis vinifera* berry developmental stages) in triplicate. Table 2 reports the Pearson correlation across technical replicates performed on *cmbd2* microarray platform. The results showed a concordance rate of 0.98-0.99 between sample replicates. This data indicates a good intra-platform data reproducibility in *cmbd2* microarray design based on different probes per transcript.

Table2: Intra-platform correlation between technical replicate: *cmbd2* gene expression experiment

	Ripening1	Ripening2	Ripening3	Veraison1	Veraison2	Veraison3
Ripening1	1					
Ripening2	0.9958461	1				
Ripening3	0.9893871	0.988672	1			
Veraison1	0.8647365	0.867462	0.871663	1		
Veraison2	0.8671879	0.870173	0.870293	0.988291	1	
Veraison3	0.8521861	0.857714	0.862808	0.984708	0.98335863	1

New microarray hybridizations have been performed with *cmbd1*, *nmgd1* and *nmgd2* microarray designs (see Figure 1of this chapter for *cmbd1*, *cmbd2*, *nmgd1* and *nmgd2* developed microarray designs and Figure 9 of the introduction section for the experimental design of the new performed microarray hybridizations), considering the same ripening and veraison RNA samples used for previous microarray hybridization as described in the introduction section. However, the intra-platform data reproducibility yielded by this new hybridization were comparable with the intra-platform data reproducibility obtained from the previous analyzed microarray expression data showed in the introduction chapter (see tables 3, 4 and 5 on pages 49-50).

4.3- Microarray Inter-platform Data Comparability

A new microarray expression data analysis have been performed, applying several combinations of background correction, and probes signal summarization (see Table 3 below). Expression values generated on different platforms cannot be directly compared because unique labelling methods and probe sequences will result in

variable signal for probes that hybridize to the same target. Alternatively, the relative expression variation between a pair of sample types should be maintained across platform. For this reason, I compared fold change values between different design/platforms. Fold change was calculated as the ratio of the ripening stage expression value to the veraison stage expression value. Furthermore, I used the logarithm (log2) transformation of the fold change values in order to avoid biases due to a compression of ratios between 0 and 1 and to represent data in a more intuitive symmetrical way centered around the 0 value.

Table 3: microarray data pre-processing and signal intensity normalization methods

Microarray designs	Background correction/data normalization	Probe average	Package and statistical method analysis
nmgd1 microarray design	Normexp_Saddle (*limma*)/ quantile	Mean/median (*limma*)	*Limma* (empirical Bayes methods)
nmgd2 microarray design	NimbleScan v. 2.5 RMA module/quantile	RMA module	*Limma* (empirical Bayes methods)
nmgd2 microarray design	Normexp_Saddle (*limma*)/quantile	Mean/median (*limma*)	*Limma* (empirical Bayes methods)
nmgd2 microarray design	Normexp_Saddle (*limma*)/quantile	Probe have been analyzed separately	*Limma* (empirical Bayes methods and P-adj. values have been averaged by Fisher combined probability test)
cmbd1 microarray desgn	Normexp_Saddle (*limma*)/quantile	Mean/median (*limma*)	*Limma* (empirical Bayes methods)
cmbd2 microarray design	Normexp_Saddle (*limma*)/quantile	Mean/median (*limma*)	*Limma* (empirical Bayes methods)
cmbd2 microarray design	Normexp_Saddle (*limma*)/quantile	Probe have been analyzed separately	*Limma* (empirical Bayes methods and P-adj. values have been averaged by Fisher combined probability test)

4.3.1. CORRELATION BETWEEN MICROARRAY PLATFORMS IN FOLD CHANGE PROFILE

I evaluated the correlation of fold changes among the four analyzed microarray designs and platforms between ripening and veraison stages (see Table 4). Each microarray intensity data (expression data) had been pre-processed and normalized with the parameters as showed in Table 3. The mean expression level of the three technical replicates measured for both *Vitis vinifera* developmental stages (ripening and veraison) was used to calculate the fold change between the two conditions. Only genes common among platforms (29432 genes) were used in the comparison. The results showed that:

(i) The inter-platform/design correlation ranged from 0.23 to 0.68 suggesting a low reproducibility of fold changes between the four analyzed microarray platforms/designs (see Table 4 below).

(ii) *Cmbd1* microarray design based on a single replicate probe per transcript exhibited the lower correlation with the other three analyzed microarrays. This result suggests that more genes discriminated as significantly differentially expressed by *cmbd1* microarray design based on a single specific replicated probe per transcript, and were not recognized as such by the other three analyzed microarray designs (*cmbd2*, *nmgd1* and *nmgd2* microarray designs).

(iii) Table 4 shows that, the correlation between *nmgd1* microarray design based on a single replicate probe per gene and *nmgd2* microarray design with multiple probes per gene is higher than the correlation between *cmbd1* (single replicate probe per gene) and *cmbd2* (multiple probes per gene) microarray designs.

Table 4: correlation between the four analyzed microarray platforms in fold change profile

	Cmbd1 mean	*Cmbd1* median	*Cmbd2* mean	*Cmbd2* median	*Nmgd1* mean	*Nmgd1* median	*Nmgd2* mean	*Nmgd2* median	*Nmgd2* RMA	*Cmbd2* fisher	*Nmgd2* fisher
Cmbd1 mean	1										
Cmbd1 median	0.980	1									
Cmbd2 mean	0.300	0.304	1								
Cmbd2 median	0.284	0.313	0.883	1							
Nmgd1 mean	0.278	0.279	0.534	0.509	1						
Nmgd1 median	0.288	0.285	0.535	0.512	0.981	1					
Nmgd2 mean	0.230	0.232	0.586	0.550	0.70	0.70	1				
Nmgd2 median	0.220	0.225	0.579	0.543	0.685	0.685	0.894	1			
Nmgd2 RMA	0.280	0.290	0.555	0.528	0.674	0.680	0.724	0.874	1		
Cmbd2 fisher	0.302	0.316	0.984	0.919	0.538	0.540	0.586	0.579	0.561	1	
Nmgd2 fisher	0.232	0.231	0.597	0.561	0.705	0.705	0.972	0.964	0.860	0.597	1

4.3.2. MICROARRAY DIFFERENTIAL EXPRESSION ANALYSIS

The four examined microarray designs have been processed for statistical analysis with *limma* package (*G.K. Smyth*, 2010), and we discriminated the statistically significant differentially expressed genes by setting a False Discovery Rate (FDR) ≤ 0.05. Results of this analysis (see Table 5) showed that:

(i) *cmbd1* microarray design based on a single specific replicate probe per gene called a higher number of significantly differentially expressed genes (DEG) than *cmbd2* microarray design with multiple probes per transcript when expression data of both

cmbd1 and *cmbd2* microarray designs were analyzed by the moderated statistical t-test (empirical Bayes methods).

(ii) Analyzing *cmbd2* and *nmgd2* microarray designs expression data with the Fisher's combined probability test, the number of significantly differentially expressed genes called by *cmbd2* microarray design increased, while those discriminated by *nmgd2* microarray design decreased (*nmgd2*.RMA and *nmgd2*.mean) or increased (*nmgd2*.median) as showed by table 5 (see below). This observation indicates that the statistical method used in microarray differential analysis influences the final expression result.

(iii) *Nmgd2* microarray design based on multiple probes per transcript called the highest number of significantly differentially expressed genes, while *cmbd2* microarray design with multiple probes per transcript exhibited the smallest number of significantly differentially expressed genes. Moreover, NimbleGen microarray designs based on long probes (60 mer) called for a higher number of significantly differentially expressed genes than CombiMatrix microarray designs based on medium probes (35-40mer). These results were in agreement with literature data which indicates that longer oligonucleotides provide better detection sensitivity than shorter probes (*Hughes,T.R. et al.*, 2001).

Table 5: significantly differentially expressed genes at a FDR ≤ 0.05 called by microarray design

Platforms	Total Number of Analyzed Genes	Differentially expressed genes at an FDR ≤0.05
CombiMatrix1.mean	29465	5248
CombiMatrix1.median	29465	5107
CombiMatrix2.mean	29432	3386
CombiMatrix2.median	29432	2464
NimbleGen1.mean	29582	11213
NimbleGen1.median	29582	11740
NimbleGen2.RMA	29549	15800
NimbleGen2.mean	29549	16664
NimbleGen2.median	29549	14723
CombiMatrix2.fisher	29432	4581
NimbleGen2.fisher	29549	15681

4.3.3. CORRELATION BETWEEN MICROARRAY DESIGNS FOR SIGNIFICANTLY DIFFERENTIALLY EXPRESSED GENES

Log2 fold change ratios of significantly differentially expressed genes (DEG) for each microarray were compared with the other microarrays.

Significantly Differentially Expressed Genes Called by each cmbd1 and cmbd2 Microarray Designs:

(i) Correlation data suggest a good agreement between *cmbd2* microarray design based on multiple probes per transcript and both NimbleGen microarray designs for differentially expressed genes discriminated by limma (moderated statistical t-test). In other words, more genes called as significantly differentially expressed by *cmbd2* microarray design based on multiple probes per transcript when expression

analysis was performed by the moderated statistical test and were recognized as such by *nmgd1* and *nmgd2* microarray designs. For significantly differentially expressed genes discriminated by limma (moderated statistical t-test), *cmbd1* microarray design based on a single replicate probe per transcript exhibited a low correlation with both *nmgd1* and *nmgd2* microarray designs. This result indicates that more genes discriminated as significantly differentially expressed by *cmbd1* microarray design were not confirmed by *nmgd1* and *nmgd2* microarray designs. In other words, correlation data suggest a low agreement between *cmbd1* microarray design based on a single specific replicate probe per transcript and both NimbleGen microarray designs (see Table 6 below).

(ii) For significantly differentially expressed genes (DEG) detected by *cmbd1* and *cmbd2* microarray design when expression data were analyzed by limma (moderated statistical t-test), Table 6 shows that, the correlation between *nmgd1* and *cmbd1* microarray designs was comparable with the correlation value observed between *nmgd2* and *cmbd1* microarray designs and that, the correlation value between *nmgd1* and *cmbd2* microarray designs was also comparable with those observed between *nmgd2* and *cmbd2* microarray designs. These results suggest that *nmgd1* and *nmgd2* microarray designs show a comparable performance in differential analysis.

(iii) For significantly differentially expressed genes discriminated by the Fisher's combined probability test, the correlation between *cmbd2* microarray design based on multiple probes per transcript and both NimbleGen microarray designs decreases (see Table 6). This result indicates that, the statistic test used in microarray expression data analysis could drastically influence the differential expression result. Moreover, this data suggests that more genes indentified as significantly differentially expressed by *cmbd2* microarray design when expression data were analyzed with the Fisher combined probability test and were not recognized as such by *cmbd1*, *nmgd1* and *nmgd2* microarray designs.

Table 6: correlation between all analyzed microarray designs in fold change profile for differentially expressed genes at a FDR ≤0.05 called by CombiMatrix microarray designs

	Cmbd1 mean (limma) (gene at an FDR ≤0.05)	*Cmbd1*median (limma) (gene at an FDR ≤0.05)	*Cmbd2* mean (limma) (gene at an FDR ≤ 0.05)	*Cmbd2* median (limma) (gene at an FDR ≤0.05)	*Cmbd2* Fisher (gene at an FDR ≤0.05)
Cmbd1 mean (limma)	--	0.98	0.70	0.73	0.56
Cmbd1 median (limma)	0.98	--	0.70	0.73	0.56
Cmbd2 mean (limma)	0.72	0.56	--	0.98	0.73
Cmbd2 median (limma)	0.72	0.56	0.96	--	0.72
Cmbd2 Fisher	0.54	0.65	0.73	0.72	--
Nmgd1 mean (limma)	0.47	0.47	0.81	0.82	0.57
Nmgd1 median (limma)	0.46	0.46	0.81	0.82	0.57
Nmgd mean (limma)	0.38	0.40	0.82	0.84	0.57
Nmgd2 median (limma)	0.41	0.40	0.82	0.84	0.57
Nmgd2 RMA (limma)	0.36	0.36	0.79	0.81	0.56
Nmgd2 Fisher	0.32	0.32	0.65	0.67	0.62

Significantly Differentially Expressed Genes Called by each nmgd1 and nmgd2 Microarray Designs:

(i) The good correlation between the expression values of genes detected as significantly differentially expressed by analysis with limma (moderated statistical t-test) in *nmgd1* or *nmgd2* microarrays suggests a good agreement between the two developed NimbleGen microarray designs.

(ii) Analyzing microarray expression data with limma (moderated statistical t- test) and considering significantly differentially expressed genes called by each *nmgd1* and *nmgd2* microarray designs, the correlation between NimbleGen and CombiMatrix microarray designs is low (*R* ranged from 0.38 to 0.62 as showed by the Table 7), indicating that more genes discriminated as significantly differentially expressed (DEG) by each *nmgd1* and *nmgd2* microarray designs were not recognized as such by *cmbd1* and *cmbd2* microarray designs. In other words CombiMatrix and NimbleGen microarray designs detect different change in gene abundance.

(iii) Considering significantly differentially expressed discriminated by each *nmgd1* and *nmgd2* microarray designs we showed that, the correlation between *cmbd1* and *nmgd1* microarray designs is different than those exhibits between *cmbd2* and *nmgd1*microarray designs and that the correlation between *cmbd1* microarray design and *nmgd2* microarray design is different than the correlation displayed between *cmbd2* and *nmgd2* microarray designs (see Table 7 below). These results suggest that both *cmbd1* and *cmbd2* microarray designs exhibit different performance in differential analysis.

Table 7: correlation between all analyzed microarray designs in fold change profile for differentially expressed genes at a FDR ≤0.05 called by NimbleGen microarray designs

	Nmgd1 mean (limma) (gene at an FDR ≤0.05)	*Nmgd1* median (limma) (gene at an FDR ≤0.05)	*Nmgd2* man (limma) (gene at an FDR ≤0.05)	*Nmgd2* median (limma) (gene at an FDR ≤0.05)	*Nmgd2* RMA (limma) (gene at an FDR ≤0.05)	*Nmgd2* Fisher gene at an FDR ≤0.05)
Cmbd1 mean (limma)	0.38	0.38	0.30	0.32	0.24	0.31
Cmbd1 median (limma)	0.38	0.38	0.30	0.33	0.25	0.31
Cmbd2 mean (limma)	0.62	0.60	0.60	0.62	0.50	0.56
Cmbd2 median (limma)	0.60	0.59	0.59	0.60	0.49	0.54
Cmbd2 Fisher	0.44	0.43	0.46	0.47	0.39	0.53
Nmgd1 mean (limma)	--	0.98	0.73	0.75	0.70	0.70
Nmgd1 median (limma)	0.98	--	0.73	0.75	0.71	0.69
Nmgd2 mean (limma)	0.81	0.79	--	0.96	0.90	0.82
Nmgd2 median (limma)	0.79	0.79	0.94	--	0.90	0.81
Nmgd2 RMA (limma)	0.75	0.74	0.90	0.92	--	0.79
Nmgd2 Fisher	0.70	0.68	0.82	0.82	0.89	--

4.4- Assessment of Microarray Design by Comparison With RNA-Seq Approach

Ripening and veraison grape berry RNA samples used for above described microarray experiment had been analyzed in two technical replicates in a previous RNA-Seq work. RNA-Seq sequencing was conducted using an Illumina Genome analyzer II machine yielding more than 59 million sequences reads, 36-44 bp in length as previously described (*Zenoni et al.* ,2010). Figure 3 shows the overview of RNA-Seq experimental design.

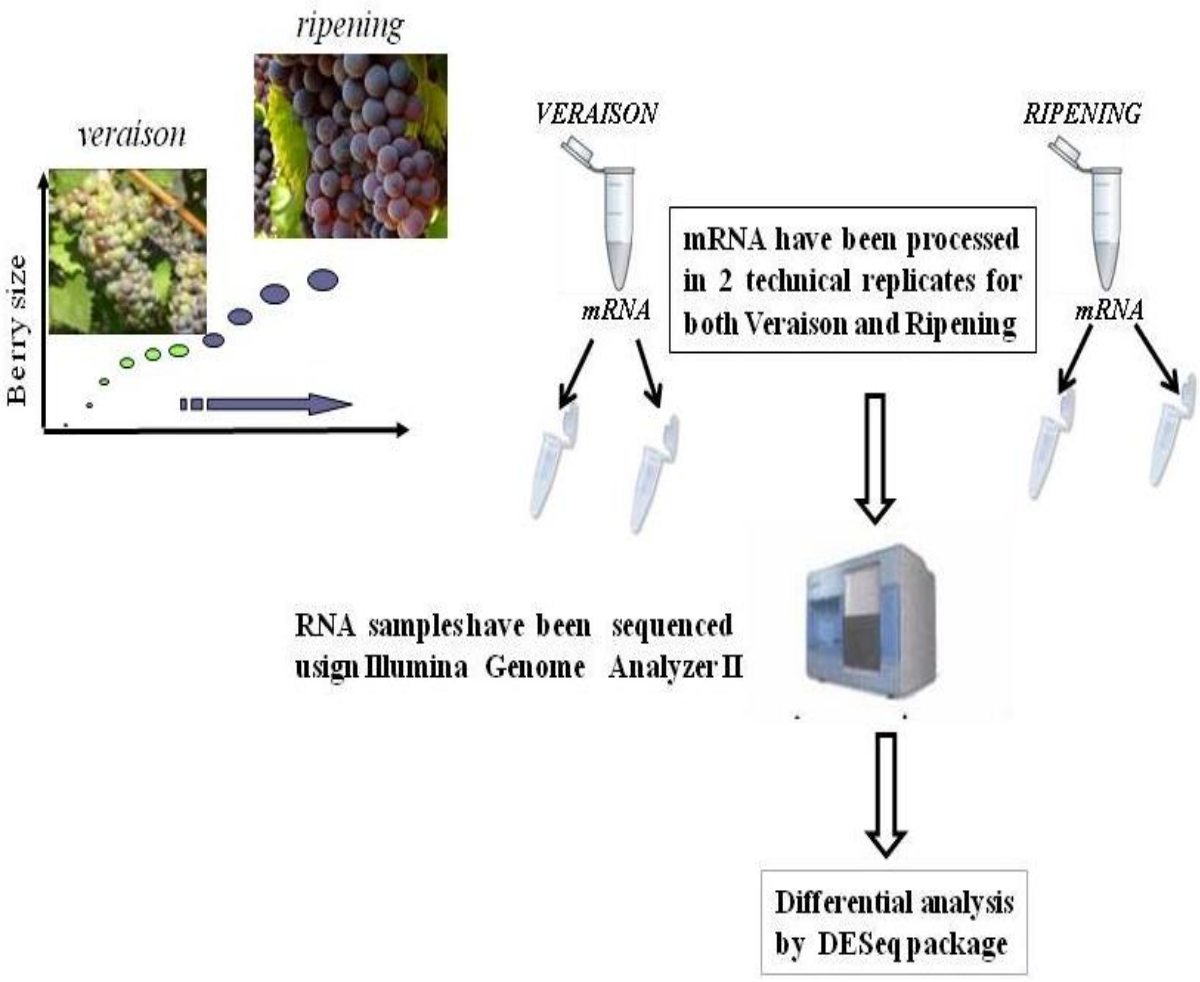

Figure 3: overview of the RNA-Seq experimental design. Ripening and veraison grape berry RNA samples had been analyzed in two technical replicates

4.4.1. RNA-Seq: EXPESSION DATA QUALITY CONTROL

In this work I re-analyzed RNA-Seq expression data by using *DESeq* R package. The package *DESeq* provides a powerful tool to estimate the variance in such data and test for differential expression. The core assumption of this method it that the mean is a good predictor of the variance. In fact gene with a similar expression level also have similar variance across replicates. Hence, we need to estimate for both ripening and veraison conditions a function that allows to predict the variance from the mean. This estimation is done by calculating for each gene the sample mean and variance within replicates and then fitting a curve to this data. It is instructive to observe at which count level the biological noise starts to dominate the shot noise. To do that, one should check whether the base variance functions seem to follow the empirical variance well. Figure 4 (see below) displayed the residual ECDF (Empirical cumulative density function) plot for veraison ("V") and ripening ("R") conditions. Green line on the residual ECDF graphic represents the expected data. For low reads counts (below 20 reads), the expression deviation become stronger in both ripening and veraison conditions. That could be due at the shot noise. Upper 20 reads count in both veraison and ripening condition, the ECDF curve follow the diagonal green line (expected data) well, indicating that variance is overestimated, which lead to too high p values and hence, an overestimation of the False Discovery Rate (FDR). For lower expressed genes, ECDF curves were below the green line, indicating the underestimation of variance, and testing for differential expression you will get too low p values and hence, too many false positives. This case may indicate a serious problem that might compromise our result. Therefore differential analysis of RNA-Seq data for gene expression major to 20 reads will allow an overestimation of the False Discovery Rate, while for gene expression under 20 reads count will include many false positives. However, genes with a read number under 20 have been included in the differential analysis. Having estimated the variance-mean dependence, it is now straightforward to look for differentially expressed genes.

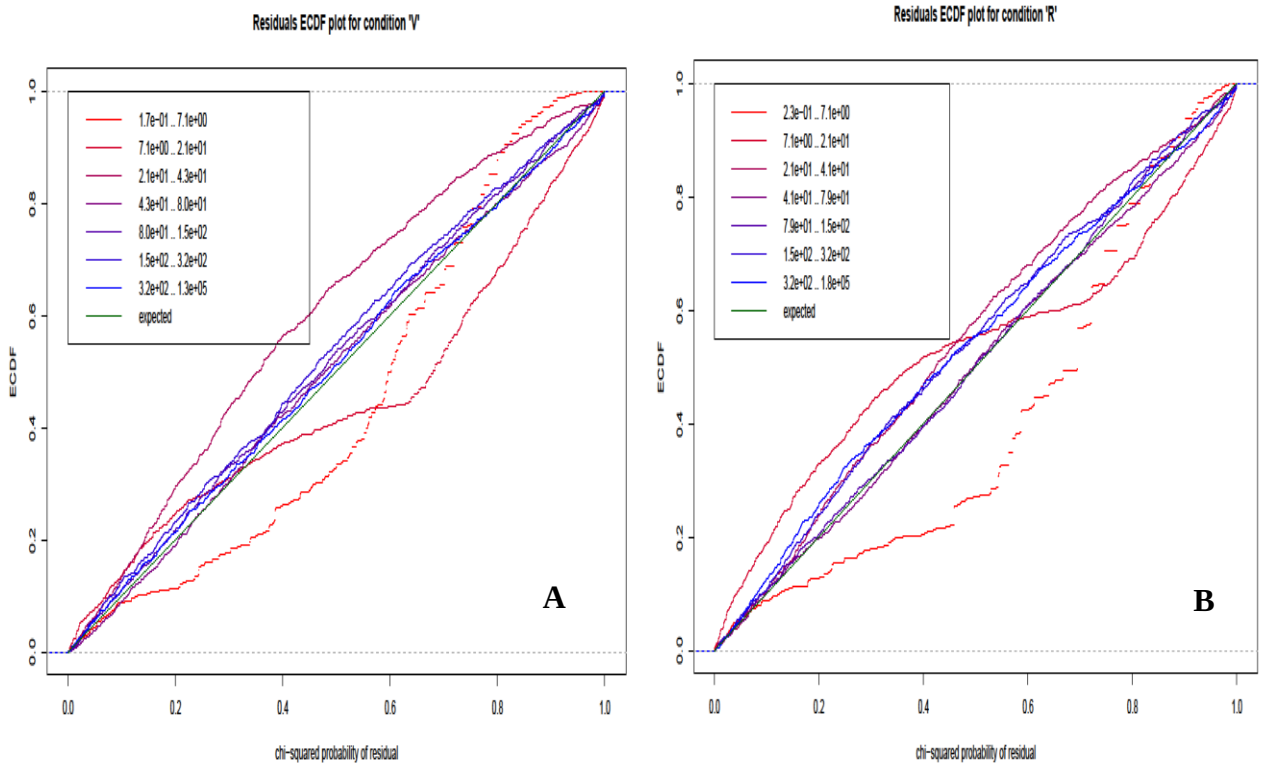

Figure 4: image of diagnostic plot to check the fit of the variance functions for the two performed technical replicates from veraison and ripening *Vitis vinifera* berry development stage.

4.4.2. RNA-Seq: DIFFERENTIAL EXPRESSION ANALYSIS

RNA-Seq data have been processed for differential analysis using *DESeq* package based on binomial negatives distribution. The analysis has been performed comparing ripening (R) and veraison (V) *Vitis vinifera* berry development stages. Figure 5 shows the plot of calculated log2-fold changes of each analyzed gene against their mean expression level. In this representation, the 9595 significantly differentially expressed genes (DEG) detected at a FDR ≤ 0.05 have been plotted in red (see below).

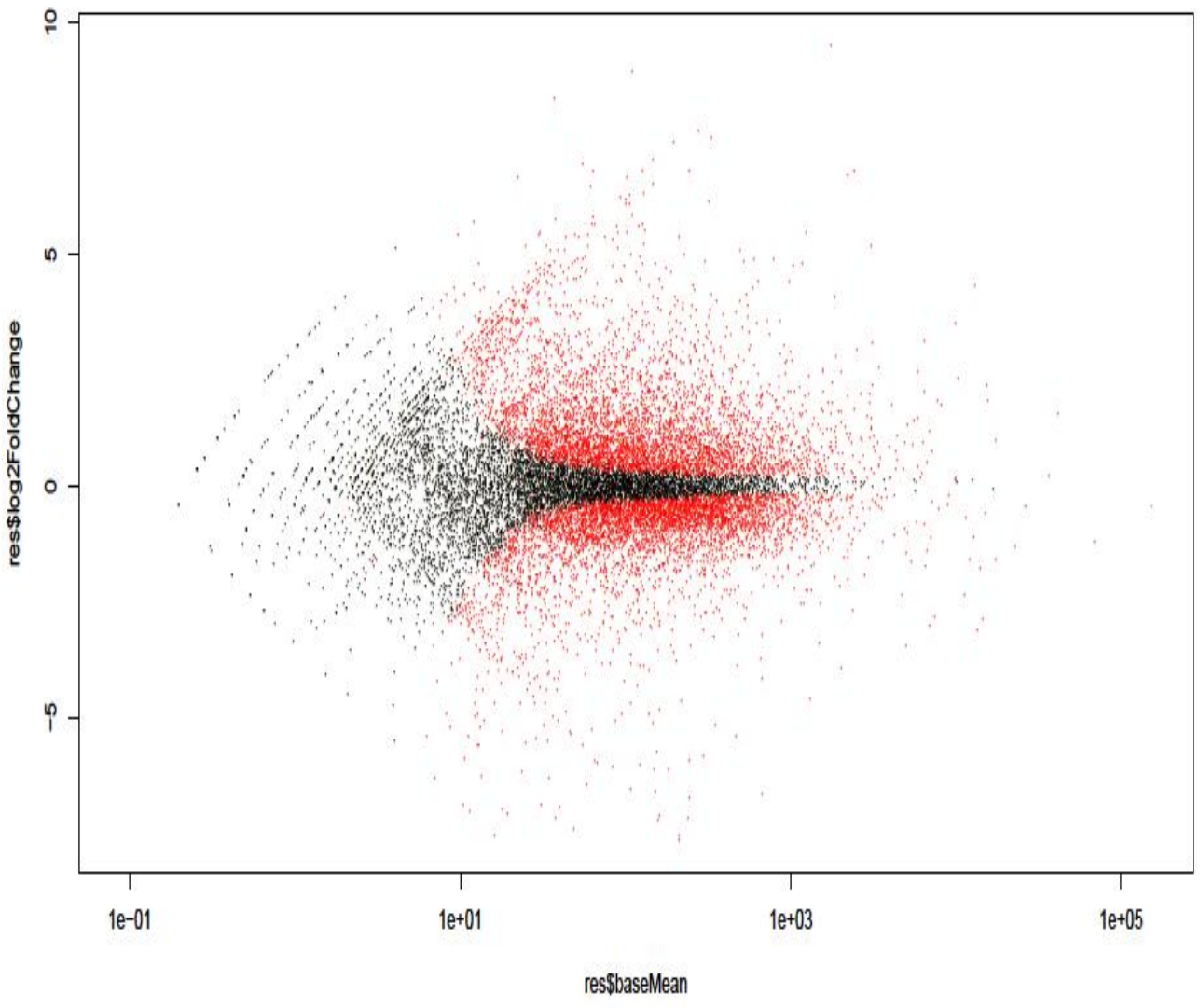

Figure 5: plot of log2-fold change values of all expressed genes across ripening and veraison conditions against their mean expression level. Red spot represent the 9595 genes discriminated as significantly differentially expressed at a FDR ≤ 0.05 by RNA-Seq expression analysis.

4.4.3. COMPARISON BETWEEN MICROARRAY AND RNA-Seq

Because different gene expression platforms utilize different technologies, quantitation and normalization method, the absolute signal values for each platform tend to be somewhat arbitrary and not suitable for Pearson correlation analysis across different platforms (*Yulei W. et al*, 2005). I therefore compared the log2-fold changes calculated between ripening and veraison *Vitis vinifera* berry development stages in microarray and RNA-Seq platforms by correlation analysis. A scatter plot was

generated between log2-fold change values determined by the two approaches (see Figures 6,7 and 8 below). Only 17850 genes expressed in all the 4 microarrays and RNA-Seq have been considered for the comparison analysis.

Correlation Between RNA-Seq and Microarray Designs Fold Change Values

The correlation values of fold change between each analyzed microarray designs and RNA-Seq are reported in Table 8 (see below) and show that:

(i) the correlation between *cmbd2* microarray design with multiple probes per transcript and RNA-Seq (*R*=0.56-0.59) is higher than those observed between *cmbd1* microarray design based on a single replicate probe per transcript and RNA-Seq (*R*=0.30). This result suggests that microarray design based on multiple medium probes (35-40 mer) per transcript and microarray design with a single replicate medium probe per transcript exhibit difference performance in differential analysis when RNA-Seq was assumed as reference.

(ii) Correlation between RNA-Seq and both NimbleGen microarray designs are comparable (*R*=0.72-0.73 and *R*=0.68-0.69), suggesting that multiple long oligos (60 mer) per transcript gave similar performance in differential analysis as a single replicated long oligos per gene (*Cheng-Chung Chou et al*., 2004).

(iii) NimbleGen microarray designs displayed a higher correlation with RNA-Seq than CombiMatrix microarray designs in fold change profile. This data indicates that long probes (NimbleGen microarray platform based on 60 mer oligo probes) performed better sensitivity than medium probes (CombiMatrix microarray platform based on 35-40mer oligo probes) in microarray expression analysis (*Hughes,T.R et al.*, 2001).

Table 8: correlation between microarray designs and RNA-Seq fold change values

	RNA-Seq
Cmbd1.mean microarray design	0.30
Cmbd1.median microarray design	0.30
Cmbd2.mean microarray design	0.58
Cmbd2.median microarray design	0.56
Nmgd1.mean microarray design	0.68
Nmgd1.median microarray design	0.69
Nmgd2.mean microarray design	0.73
Nmgd2.median microarray design	0.72
Nmgd2.RMA microarray design	0.72
Cmbd2.Fisher microarray design	0.59
Nmgd2.Fisher microarray design	0.73

Comparison Between Significantly Differentially Expressed Genes

- CombiMatrix Microarray Designs Analyzed by limma (moderated statistical t-test) and RNA-Seq:

I examined the ability of both RNA-Seq and Microarray technologies to indentify significantly differentially expressed genes at a False Discovery Rate ≤ 0.05 between repining and veraison *Vitis vinifera* berry development stages.

(i) The comparison between RNA-Seq and CombiMatrix microarray designs considering significantly differentially expressed genes called by both approaches showed that regardless of the microarray designs used, both *cmbd1* and *cmbd2* microarray designs displayed a high correlation with RNA-Seq (see Figure 5 and Table 9 below; *R*=0.82-0.89). These results indicate a good agreement between both

developed CombiMatrix microarray designs (*cmbd1* and *cmbd2* microarray designs) and RNA-Seq when a set of significantly differentially expressed gene (DEG) called by the two approaches overlapped.

(ii) *Cmbd1* microarray design based on a single specific replicate probe per transcript exhibits an higher number of significantly differentially expressed genes at a FDR≤0.05 confirmed by RNA-Seq than *cmbd2* microarray design with multiple probes per transcript. The result shows that *cmbd1* microarray design with a single specific replicate medium probe (35-40 mer) per transcript is more sensible than *cmbd2* microarray design based on different medium probes (35-40 mer) per gene in detecting significantly differentially expressed gene at a FDR≤0.05. Analysis of Figure 5 shows that,1021 and 64 significantly differentially expressed genes have been recognized only by *cmbd1* and *cmbd2* microarray designs respectively (green spot in figure 5), supporting that *cmbd2* microarray design based on different probes per transcript called less false positives as significantly differentially expressed genes in differential analysis when RNA-Seq was assumed as gold standard. This result suggests that *cmbd2* microarray design based on multiple probes per gene is more specific than *cmbd1* microarray design with a single specific replicate probe per transcript in differential analysis.

(iii) Figure 5 also shows that around 7000 genes (7109 and 7648 for *cmbd1* and *cmbd2* microarray design respectively) have been discriminated as significantly differentially expressed only by RNA-Seq (blue spot in Figure 5). This observation suggests that both CombiMatrix microarray designs are less sensitive than RNA-Seq in differential analysis (*John C. Marioni et al.*, 2008).

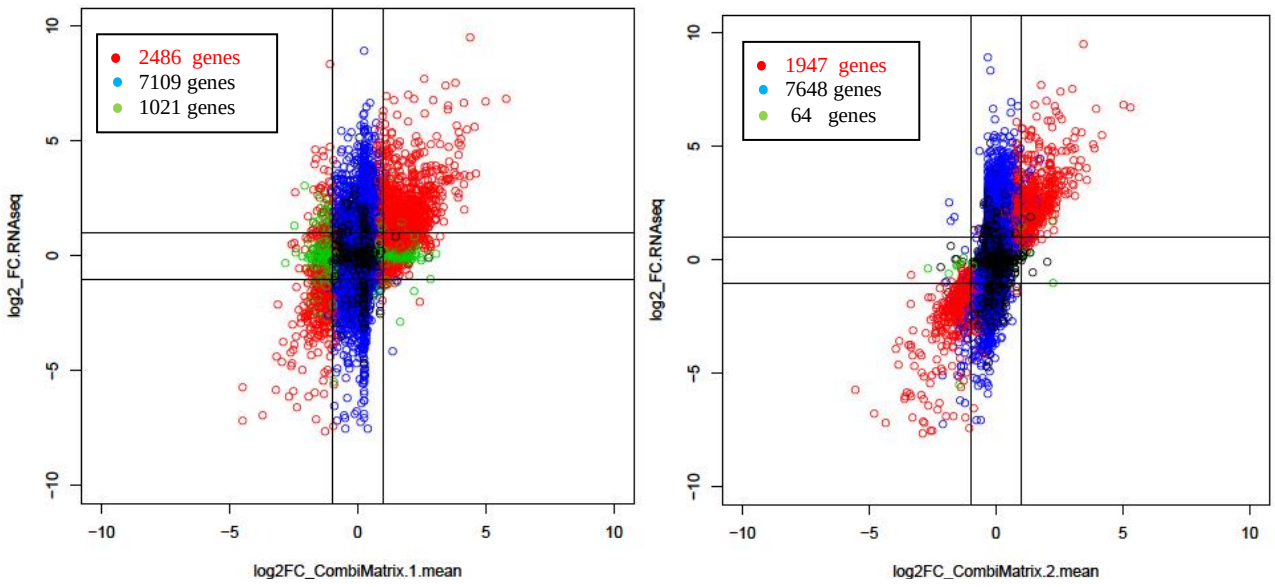

Figure 5: CombiMatrix1-mean and CombiMatrix2-mean microarray designs scatter plots compares the log2-fold change ratio of significantly differentially expressed (using ripening versus veraison). Genes identified as differentially expressed by both methods are plotted in red; genes identified as differentially expressed by either microarray or RNA-Seq are plotted in green and blue respectively; genes not identified as differentially expressed by either method are plotted in black.

Red spot: Significantly differentially expressed genes at an FDR≤0.05 by both RNA-Seq and Microarray

Blue spot: Significantly differentially expressed genes at an FDR≤0.05 discriminated only by RNA-Seq

Green spot: Significantly differentially expressed genes at an FDR≤0.05 discriminated only by Microarray

Black spot: Gene not significantly differentially expressed

- CombiMatrix Microarray Design Analyzed by Fischer's Combined Probability Test and RNA-Seq:

I also performed a comparison between *cmbd2* microarray design based on multiple probes per transcript and RNA-Seq when microarray expression data have been processed by the Fisher combined probability test.

(i) It is note worthy that the number of significantly differentially expressed genes at a false discovery rate (FDR) ≤0.05 called by *cmbd2* microarray design confirmed by RNA-Seq increases (Figure 6). This data indicates that the sensitivity of *cmbd2* microarray design based on different probes per transcript in detecting significantly differently expressed genes increases when microarray expression data were analyzed by the Fisher combined probability test.

(ii) Figure 6 shows that the number of the false positive significantly differentially expressed genes called by *cmbd2* microarray design when the statistical analysis was performed by the Fisher combined probability test assuming RNA-Seq tool as reference increases. This result indicates that *cmbd2* microarray design based on multiple medium probes (35-40 mer) per transcript analyzed by the Fisher combined t-test is less specific in differential analysis.

(iii) However, *cmbd2* microarray design based on different probes per transcript analyzed by Fischer combined probability test, is more sensitive and more specific than *cmbd1* microarray design with a single replicate probe per transcript in discriminating significantly differentially expressed genes at a false discovery rate (FDR) ≤ 0.05, when RNA-Seq was assumed as gold standard (see Figures 5 and 6). The data suggests that Fischer combined probability test could be an alternative statistical test for *cmbd2* microarray design expression data analysis. This data also suggests that *cmbd2* microarray design based on multiple medium probes per transcript performed better in differential analysis with respect to *cmbd1* microarray design based on a single replicate medium probe per transcript.

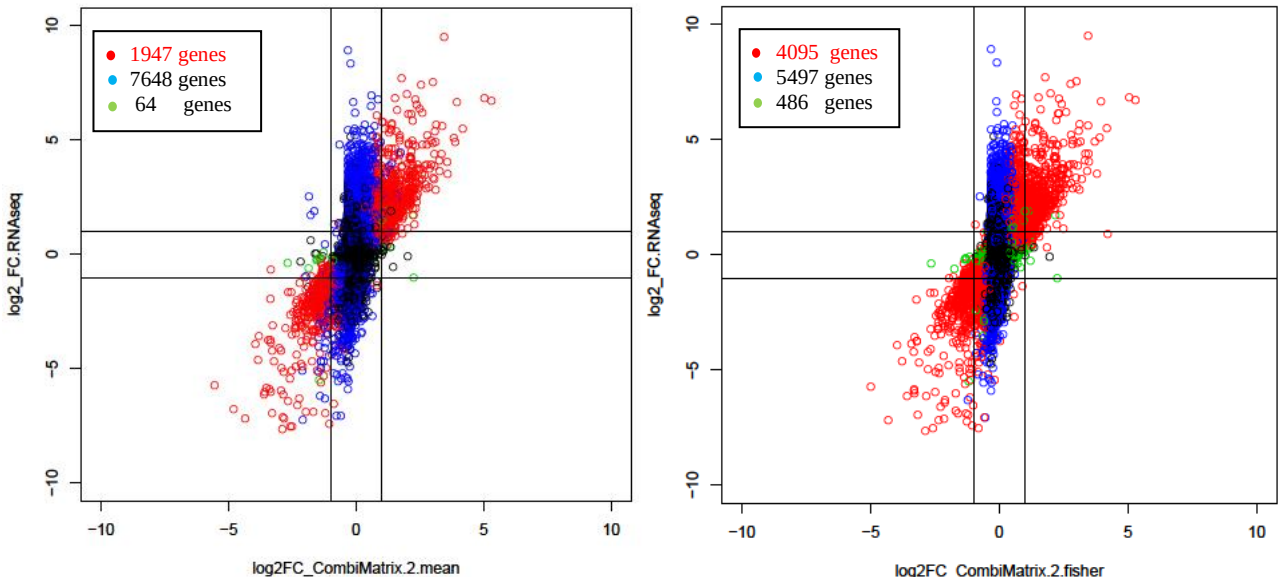

Figure 6: CombiMatrix2-mean and CombiMatrix2-Fisher microarray designs scatter plots compares the log2-fold change ratio of significantly differentially expressed (using ripening versus veraison). Genes identified as differentially expressed by both methods are plotted in red; genes identified as differentially expressed by either microarray or RNA-Seq are plotted in green and blue respectively; genes not identified as differentially expressed by either method are plotted in black.

Red spot: Significantly differentially expressed genes at an FDR≤0.05 by both RNA-Seq and Microarray

Blue spot: Significantly differentially expressed genes at an FDR≤0.05 discriminated only by RNA-Seq

Green spot: Significantly differentially expressed genes at an FDR≤0.05 discriminated only by Microarray

Black spot: Gene not significantly differentially expressed

NimbleGen Microarray designs and RNA-Seq

-NimbleGen Microarray Designs Analyzed by limma (moderated statistical t-test) and RNA-Seq

(i) Figure 7 showed that *nmgd2* microarray design based on different probes per transcript discriminated a higher number of significantly differentially expressed genes at a false discovery rate (FDR) ≤0.05 confirmed by RNA-Seq with respect to *nmgd1* microarray design with a single replicate probe per transcript, indicating that *nmgd2* microarray design based on multiple long probes (60 mer) per transcript is more sensitive than *nmgd1* microarray design with a single replicate long probe (60 mer) per transcript in discriminating significantly differentially expressed genes (DEG) at a false discovery rate (FDR) ≤0.05 when RNA-Seq was assumed as gold standard.

(ii) *Nmgd2* microarray design called a higher number of significantly differentially expressed genes that were not recognized by RNA-Seq. This result indicates that *nmgd2* microarray design based on different probes per transcript exhibits a less specificity in differential analysis with respect to *nmgd1* microarray design based on a single replicate probe per transcript.

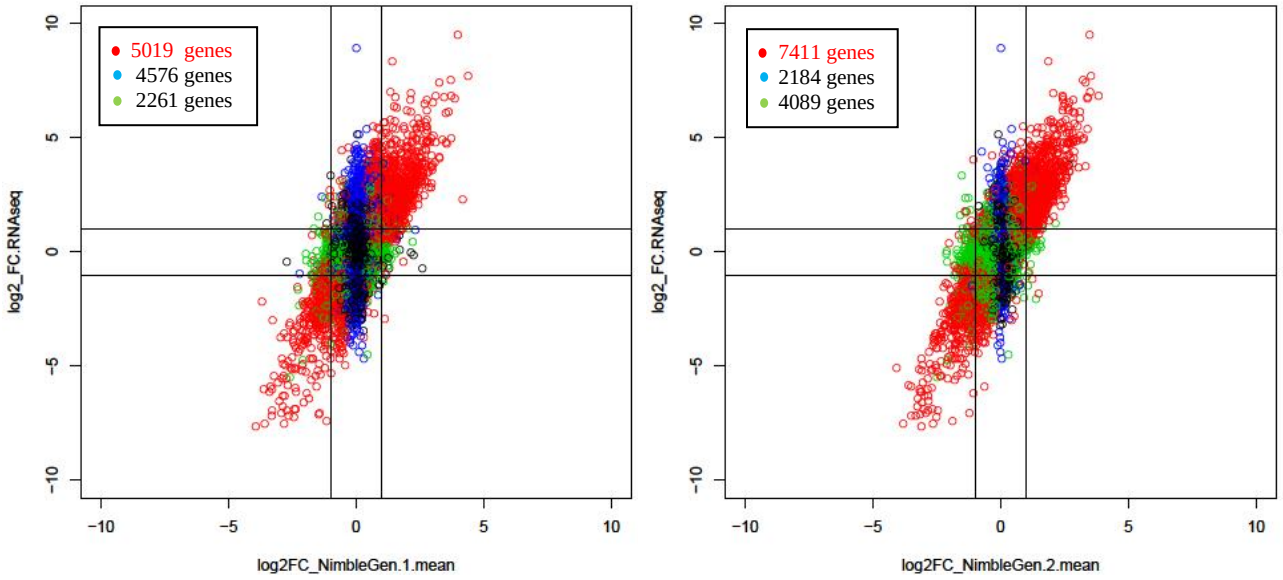

Figure 7: NimbleGen1-mean and NimbleGen2-mean microarray designs scatter plots compares the log2-fold change ratio of significantly differentially expressed (using ripening versus veraison) from NimbleGen1-mean and NimbleGen2-mean. Genes identified as differentially expressed by both methods are plotted in red; genes identified as differentially expressed by either microarray or RNA-Seq are plotted in green and blue respectively; genes not identified as differentially expressed by either method are plotted in black.

Red spot: Significantly differentially expressed genes at an FDR≤0.05 by both RNA-Seq and Microarray

Blue spot: Significantly differentially expressed genes at an FDR≤0.05 discriminated only by RNA-Seq

Green spot: Significantly differentially expressed genes at an FDR≤0.05 discriminated only by Microarray

Black spot: Gene not significantly differentially expressed

I next compared NimbleGen microarray design based on different probes per transcript with RNA-Seq considering microarray data for which internal replicate probes have been summarized by RMA algorithm as described by Irizarry et. al (*Irizarry et al.* ,2003) from NimbleScan software (NimbleGen data-pre-processing software; see materials and methods). It emerged from this analysis that processed *nmgd2* microarray design expression data with RMA module improves it specificity and decreases it sensitivity. The result indicates that different methods of data pre-

processing in microarray data analysis could influence the performance of microarray platforms in differential analysis. However, both described data analysis systems in Figure 8 (intensity data pre-processed by *limma* and by RMA module) displayed a similar performance (correlation: *R*=0.70-71) and a good concordance with RNA-Seq in differential analysis (see Figure 8 and Table 9 below).

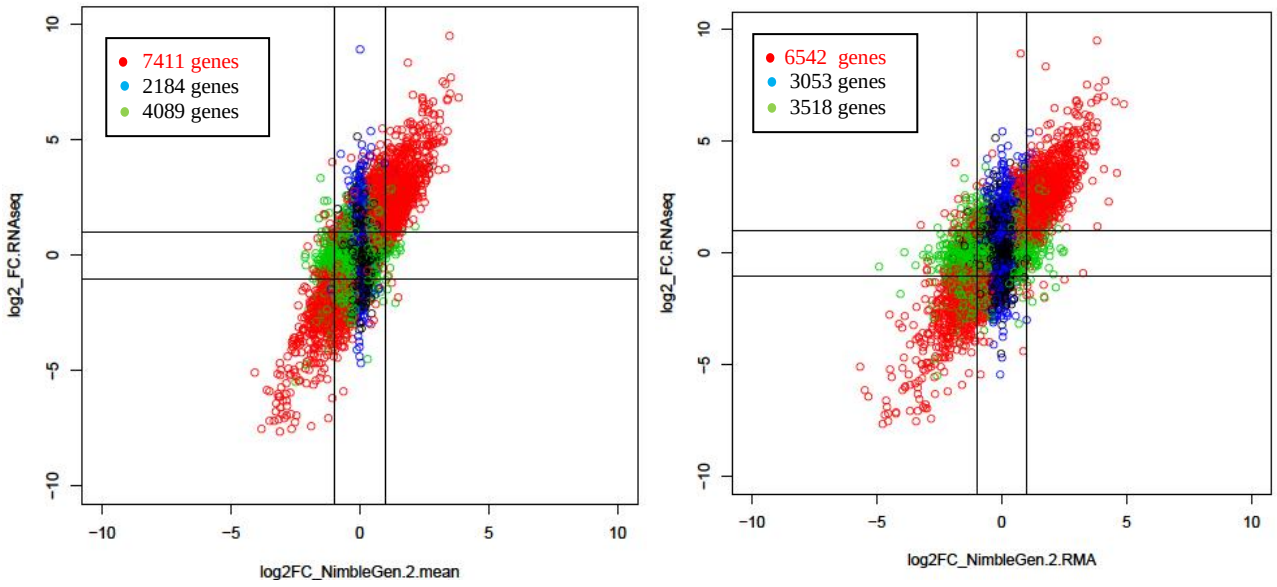

Figure 8: NimbleGen2-mean and NimbleGen2-RMA microarray designs scatter plots compares the log2-fold change ratio of significantly differentially expressed (using ripening versus veraison). Genes identified as differentially expressed by both methods are plotted in red; genes identified as differentially expressed by either microarray or RNA-Seq are plotted in green and blue respectively; genes not identified as differentially expressed by either method are plotted in black.

Red spot: Significantly differentially expressed genes at an FDR≤0.05 by both RNA-Seq and Microarray

Blue spot: Significantly differentially expressed genes at an FDR≤0.05 discriminated only by RNA-Seq

Green spot: Significantly differentially expressed genes at an FDR≤0.05 discriminated only by Microarray

Black spot: Gene not significantly differentially expressed

- NimbleGen Microarray Design 2 Pre-processed by RMA Algorithm and Analyzed by limma (moderated statistical t-test) and Fisher Combined Probability test and RNA-Seq.

Finally I performed a comparison between NimbleGen microarray design based on different probes per transcript (*nmgd2*) for expression data pre-processed by RMA module (*Irizarry et al.* ,2003) and analyzed by limma moderated statistical t-test and Fisher combined probability test. The sensibility of *nmgd2* microarray design based

on different probes per transcript increases drastically when the statistical analysis was performed by the Fisher combined t-test (see Figure 9). By contrast, *nmgd2* microarray design based on different probes per transcript exhibited a low specificity in differential analysis when statistical analysis was performed by Fisher combined probability test. This result indicates that *nmgd2* microarray design performance in differential analysis is more influenced by the Fisher combined probability test. Regardless, the statistical test applied for expression data analysis, *nmgd2* microarray design based on different probes per transcript exhibited a good concordance with RNA-Seq approach in differential analysis (see table 9: R=0.70-0.71) and discriminated the highest number of significantly differentially expressed genes confirmed by RNA-Seq.

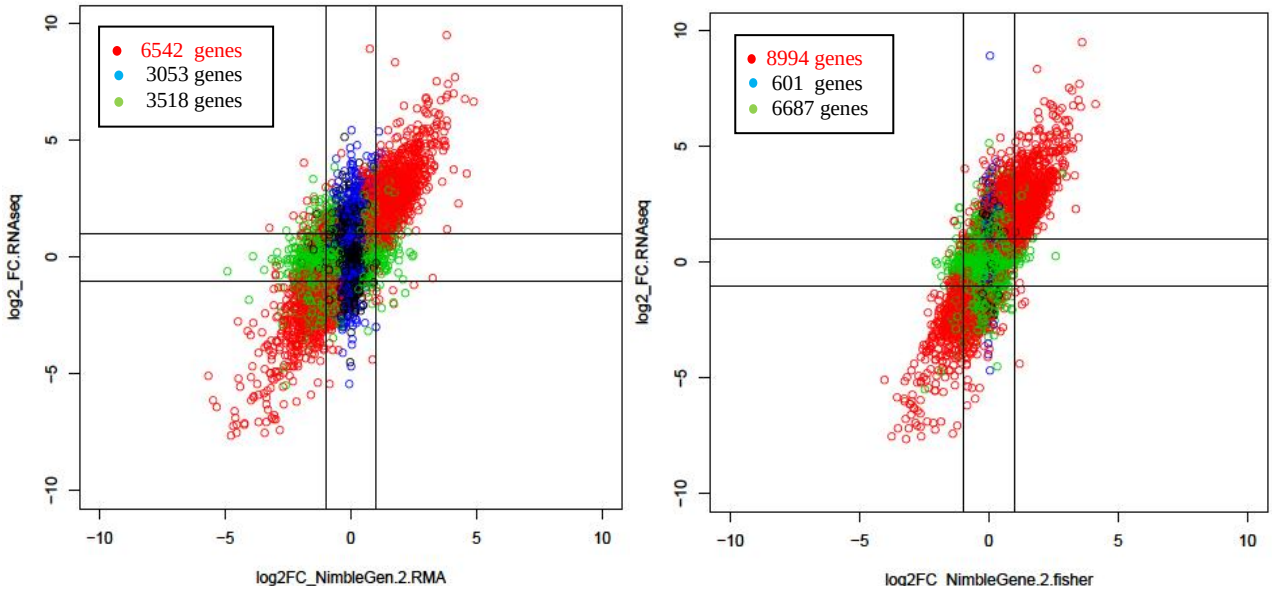

Figure 9: NimbleGen2-RMA and NimbleGen2-Fisher microarray designs scatter plots compares the log2-fold change ratio of significantly differentially expressed (using ripening versus veraison). Genes identified as differentially expressed by both methods are plotted in red; genes identified as differentially expressed by either microarray or RNA-Seq are plotted in green and blue respectively; genes not identified as differentially expressed by either method are plotted in black.

Red spot: Significantly differentially expressed genes at an FDR≤0.05 by both RNA-Seq and Microarray

Blue spot: Significantly differentially expressed genes at an FDR≤0.05 discriminated only by RNA-Seq

Green spot: Significantly differentially expressed genes at an FDR≤0.05 discriminated only by Microarray

Black spot: Gene not significantly differentially expressed

Table 9 (below) shows the general overview of the comparison (plot data) between the four analyzed microarray design/platforms and RNA-Seq approach in differential analysis.

Table9: summary of comparison between microarray designs and RNA-Seq in differential analysis

Microarray Platforms	Not DEG	DEG called only by Microarray	DEG called only by RNA-Seq	DEG called by both Microarray and RNA-Seq	Fold change correlation between array and RNA-Seq for DEG called by both approaches
cmbd1 mean	7234	1021	7109	2486	0.82
cmbd1 median	7276	985	7166	2429	0.82
cmbd2 mean	8191	64	7648	1947	0.88
cmbd2 median	8172	83	7748	1847	0.87
nimgd1 mean	5994	2261	4576	5019	0.70
nimgd1 median	5460	2795	3890	5705	0.69
nimgd2 mean	4166	4089	2184	7411	0.71
nimgd2 median	4494	3761	2559	7036	0.71
nimgd2 RMA	4737	3518	3053	6542	0.70
cmbd2 Fisher	7769	486	5497	4095	0.88
nimgd2 Fisher	1568	6687	601	8994	0.71

These results indicate that:

(i) Regardless of the microarray design used, NimbleGen microarray platform exhibited a lower specificity in detecting genes defined as differentially expressed by RNA-Seq analysis with respect to CombiMatrix microarray design, while CombiMatrix microarray design showed a lower sensitivity.

(ii) The sensitivity and the specificity of analyzed microarray designs based on different probes per transcript are influenced by the statistical test and the system or module used for expression data pre-processing.

(iii) The Fisher combined probability test improves the sensitivity of both *cmbd2* and *nmgd2* microarray designs based on multiple probes per transcript. The Fisher combined probability test could be an alternative statistical test for both *cmbd2* and *nmgd2* microarray data analysis. However, the Fisher combined probability test reduces the specificity of both *cmbd2* and *nmgd2* microarray designs considering RNA-Seq as gold standard.

(iv) A good correlation was observed between each analyzed microarray design and RNA-Seq when the set of significantly differentially expressed genes called by the two approaches overlapped.

4.5- Performance Assessment of microarray designs by Sensitivity, Specificity, Accuracy and Positive Predictive Values Parameters

I examined the True Positives Rate, the Accuracy, the Sensitivity and the Specificity (see materials and methods) of each analyzed microarray designs in discriminating significantly differentially expressed genes (DEG) assuming RNA-Seq tool as reference (see Table 10). Gene at a false discovery rate (FDR) ≤0.05 and with a fold change ratio (FC) ≥ 2 was considered as Significantly Differentially Expressed (DEG). 5020 genes were discriminated as significantly differentially expressed by RNA-Seq. Table 10 (see below) summarizes the sensitivity, the specificity, the accuracy and the positive predictive values parameters of each analyzed microarray designs.

(i) For statistical analysis performed by limma (moderated statistical t-test),both c*mbd1* and *cmbd2* microarray designs called a similar number of significantly differentially expressed genes with a fold change ratio ≥2 confirmed by RNA-Seq. Table 10 showed that, *cmbd2* microarray design based on multiple probes per gene exhibited a higher accuracy (76%) in discriminating significantly differentially expressed genes in differential analysis with respect to the other three microarray designs. Moreover, *cmbd2* microarray design based on multiple medium probes per transcript exhibited a higher positives predictive values ratio (PPV = 75.95-76.08) than the other three analyzed microarray designs. These results support the good agreement between CombiMatrix microarray design based on different probes per transcript (*cmbd2* microarray design) and RNA-Seq in differential analysis .

(ii)Table 10 shows that, the positive predictive values of both *nmgd1* and *nmgd2* microarray design are comparable (44%-49%) when statistical analysis was performed by limma moderated statistical t-test. The accuracy value of both *nmgd1* and *nmgd2* microarray designs in differential analysis are also comparable (64%-69%). However, *nmgd2* microarray design (multiple long probes per transcript) displayed the highest sensitivity (85%-88%) calling the highest number of significantly differentially expressed genes confirmed by RNA-Seq in comparison with the other three analyzed microarray designs. *Nmgd2* microarray design based on different long probes per transcript exhibited less specificity than *nmgd1* microarray design with a single long replicate probe per transcript.

(iii) The sensitivity of *cmbd2* and *nmgd2* microarray designs based on multiple probes per transcript increase when statistical analysis of their expression data was performed by the Fisher combined probability test, while their capacity to predict true positive significantly differentially expressed genes decrease (see Table 10). *Cmbd2* microarray data analyzed by the Fisher combined probability test exhibited a high specificity (specificity= 83.30%) in discerning significantly differentially expressed genes, while *nmgd2* microarray displayed a low specificity (specificity=15.86%). This result indicates that the number of false positive

significantly differentially expressed genes called by *cmbd2* microarray design is lower than those detected by *nmgd2* microarray design, indicating that the Fisher combined probability test is a good alternative statistical test for *cmbd2* microarray data analysis.

Table 10: positive predictive values (PPV), Sensitivity, Specificity and Accuracy of microarray designs in detecting significantly differentially expressed genes (FDR≤0.05 and |log2-fold change|≥1)

Microarray designs	Positive	Sensitivity	Specificity	Accuracy	True Positives	Positive Predictive Values
cmbd1.mean	3507	27.95	83.77	67.50	1454	41.46
cmbd1.median	3414	27.29	84.23	67.64	1420	41.59
cmbd2.mean	2011	29.31	96.16	76.68	1525	75.83
cmbd2.median	1930	28.31	96.39	76.55	1473	76.32
nimgd1.mean	7280	67.67	70.28	69.51	3520	48.35
nimgd1.median	8500	73.78	63.14	66.24	3838	45.15
nimgd2.mean	11500	88.95	45.66	58.27	4627	40.23
nimgd2.median	10797	85.91	49.97	60.44	4469	41.39
nimgd2.RMA	10060	85.10	55.46	64.10	4427	44.00
cmbd2.Fisher	4581	46.82	83.03	72.48	2435	53.15
nimgd2.Fisher	15681	96.89	15.87	39.48	5040	32.14

4.6- Assessment of Microarray Designs by ROC Curve Analysis

Previous results showed that the set of genes differentially expressed detected by the developed microarray designs and RNA-Seq technology did not overlap for a consistent portion of genes (green and blue spots in Figures 5, 6, 7, 8 and 9). To further explore these discrepancies, a Receiver Operating Characteristics (ROC) curve was constructed for each microarray design and platform assuming RNA-Seq expression data set as the gold standard. Each point on the ROC curve of a given microarray platform represents the sensitivity on *Y-axis* (True Positive Rate: TPR) and the specificity on *X-axis* (False Positive Rate: FPR).

(i) ROC curve analysis performed on the dataset of differentially expressed genes identified by the two developed CombiMatrix microarray designs showed that the area under curve (AUC) described by the tested log2-fold change values from *cmbd2* microarray design based on different probes per transcript (AUC=0.70) is higher than those discriminated by *cmbd1* microarray design based on a single specific replicate probe per transcript (AUC=0.57) (see Figure10 and Table 11 below). This data suggests that *cmbd2* microarray design with multiple medium probes per transcript performed better in discriminating significantly differentially expressed genes in differential analysis with respect to *cmbd1* microarray design based on a single specific replicate medium probe per transcript when RNA-Seq was assumed as gold standard.

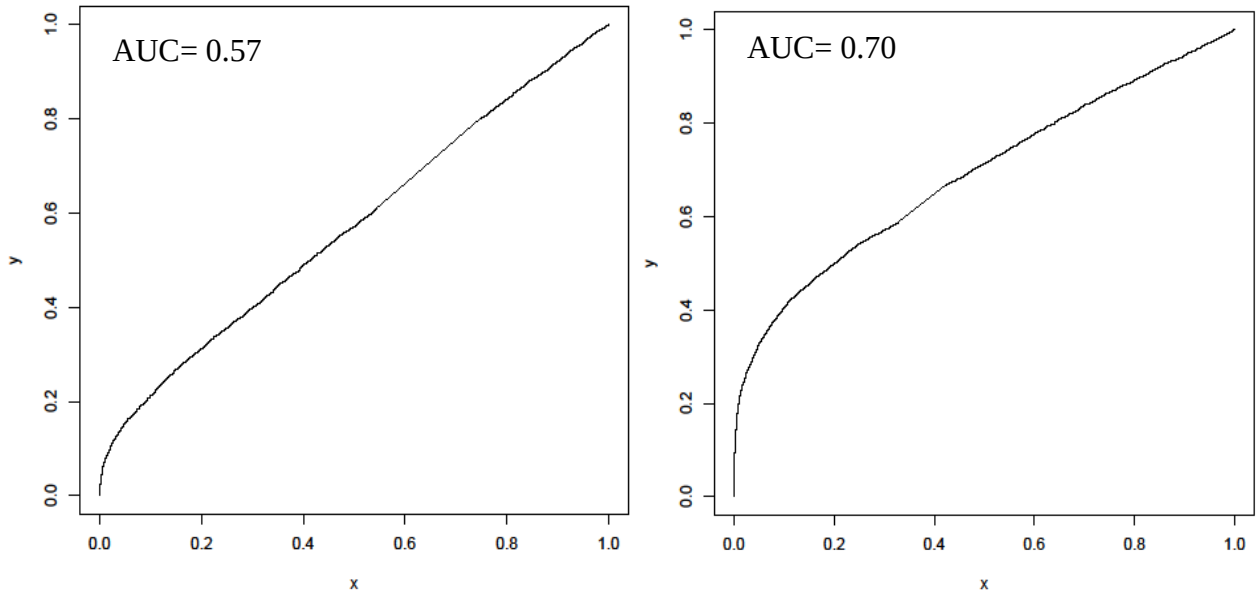

Figure 10: ROC curve to evaluate signal detection of both *cmbd1* and *cmbd2* microarray designs in differential analysis. On *X* and *Y-axis*, false positive rate (FPR) and true positive rate (TPR) of discriminated differentially expressed genes have been respectively represented.

(ii) ROC curve analysis performed on datasets of differentially expressed genes detected by NimbleGen microarray designs showed that the area under curve (AUC) described by the tested log2-fold change values relative to *nmgd2* microarray design based on different probes per transcript (AUC=0.81) is higher than the area under curve discriminated by the tested log2-fold change values from *nmgd1* microarray design based on a single replicate probe per transcript (AUC=0.75) (see Figure10 and Table 11 below). This data suggests that NimbleGen microarray design with multiple long probes per transcript performed better than NimbleGen microarray design based on a single long replicate probe per gene in detecting differentially expressed genes identified by RNA-Seq. However, regardless of the microarray design used, *nmgd2* microarray design with multiple long probes per transcript and *nmgd1* microarray design based on a single replicate long probe per transcript displayed a moderate accurate performance (*Ezio Bottarelli* and *Stefano Parodi*, 2003) in detecting significantly differentially expressed genes in differential analysis when RNA-Seq was considered as reference.

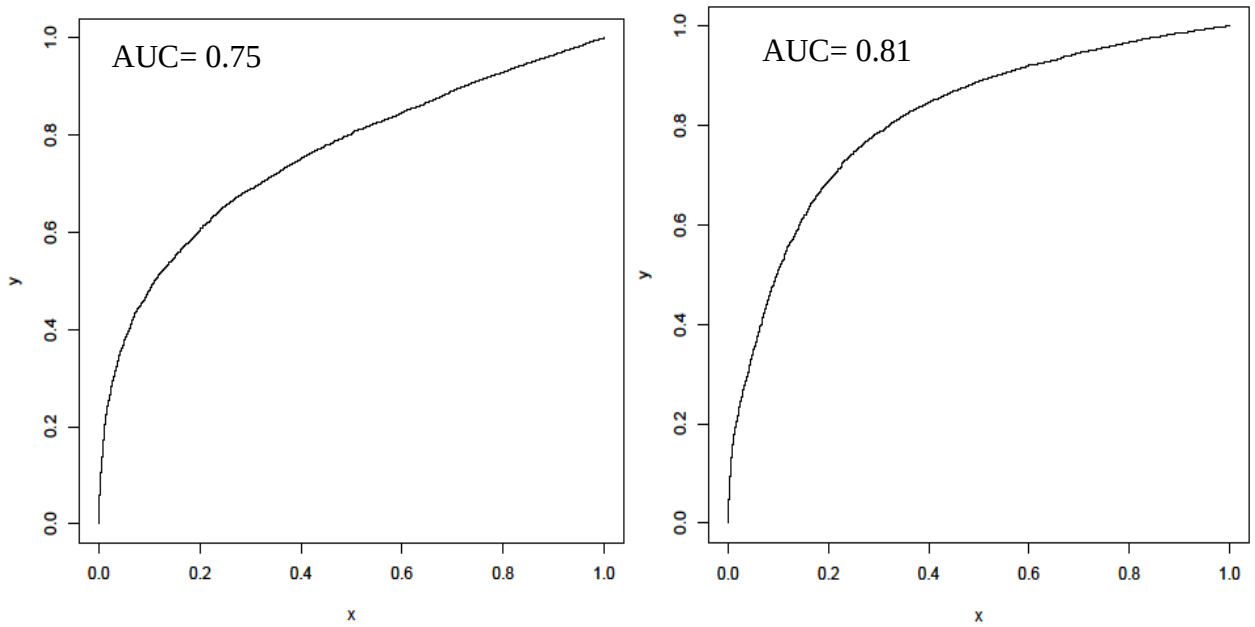

Figure 11: ROC curve analysis to evaluate signal detection of both *nmgd1*and *nmgd2* microarray designs in differential analysis. On *X* and *Y-axis*, false positive rate (FPR) and true positive rate (TPR) of discriminated differentially expressed genes have been respectively represented.

In summary ROC curve analysis shows that:

(i) NimbleGen microarray designs with long probes per transcript (60 mer) performed better in detecting significantly differentially expressed genes (DEG) identified by RNA-Seq analysis in comparison to CombiMatrix microarray designs based on medium probes (35-40 mer) .

(ii) Both, CombiMatrix and NimbleGen microarray designs with multiple probes per gene exhibited a higher Area Under Curve than their respective designs based on a single replicated probe per transcript (see Table 11). Therefore, regardless of the microarray platform used, microarray designs with multiple probes per transcript performed better in discriminating genes differentially expressed previously identified by RNA-Seq analysis.

(iii) However, regardless of the microarray design used, both NimbleGen microarray designs based on long probes (60 mer) exhibited a moderated accurate performance in detecting significantly differentially expressed genes when RNA-Seq was assumed as reference. By contrast the performance of microarray platform based on medium probe per transcript (CombiMatrix: 35-40 mer) is more influenced by the sort of the microarray design in differential analysis (see Table 11).

Table 11: Area Under Curve (AUC) of each analyzed microarray design

Microarray Designs	Area Under Curve (AUC)
CombiMatrix design1 mean (*cmbd1*.mean)	0.5706
CombiMatrix design1 median (*cmbd1*.median)	0.5712
CombiMatrix design2 mean (*cmbd2*.mean)	0.7022
CombiMatrix design2 median (*cmbd2*.median)	0.6887
NimbleGen design1 mean (*nmgd1*.mean)	0.7587
NimbleGen design1 median (*nmgd1*.median)	0.7601
NimbleGen design2 mean (*nmgd2*.mean)	0.8319
NimbleGen design2 median (*nmgd1*.median)	0.8131
NimbleGen design2 RMA (*nmgd2*.RMA)	0.8155

4.7- Validation of Differential Expression Data

In order to validate the expression data of microarray and RNA-Seq expression experiments, a set of 10 genes whose expression was in agreement and in disagreement among the two approaches, have been tested by Real Time PCR. 8 genes out the 10 selected for Real Time PCR analysis were in disagreement between all analyzed microarray design and RNA-seq. This analysis shows a good agreement between Real Time PCR and RNA-Seq in fold change profile (see Figure 13 and Table 12). For the 10 analyzed genes, Real Time PCR and RNA-Seq expression data agree for 8 genes and contrast for two genes, while Real Time-PCR and microarray

expression data (considering all the four analyzed microarray designs) agree for the 2 selected genes (JGVV1.1082 and JGVV301.10) which were in agreement between microarray and RNA-Seq (see Table 12). This result suggests that microarray technology failed in differentially analysis for a portion of significantly differentially expressed genes probably because of the low sensitivity and the limit of microarray technology to detect accurately small variation in differential analysis.

Table 12: log2-fold change values of the four analyzed microarray designs, RNA-Seq and RT-PCR

genes	RNA-Seq Gene expression level (RPKM)	log2-FC RNA-Seq	log2-FC *nmgd2*	log2-FC *nmgd1*	log2-FC *cmbd2*	log2-FC *cmbd1*	log2-FC.qRTPCR
JGVV1.1082	29.30	-6.17	-3.38	-1.94	-2.60	-2.65	-1.21
JGVV301.10	38.64	7.13	4.43	1.85	3.03	1.59	5.35
JGVV44.63	7.61	-0.30	-1.33	0.22	-1.33	-0.31	0.72
JGVV1.1126	2.46	-2.90	-0.24	0.19	-0.16	-0.22	-2.57
JGVV0.270	2.95	1.35	-0.07	0.00	0.08	0.18	4.96
JGVV0.133	2.77	1.34	-0.05	-0.11	0.05	0.67	4.99
JGVV151.6	2.20	-2.99	1.13	0.19	0.09	-0.02	0.35
JGVV129.66	4.54	1.48	-1.01	-0.46	0.45	-0.22	8.04
JGVV4.362	3.45	1.87	-1.22	-0.48	0.14	-0.16	2.11
JGVV61.51	6.45	-1.23	3.37	-1.12	0.44	-0.16	-1.06

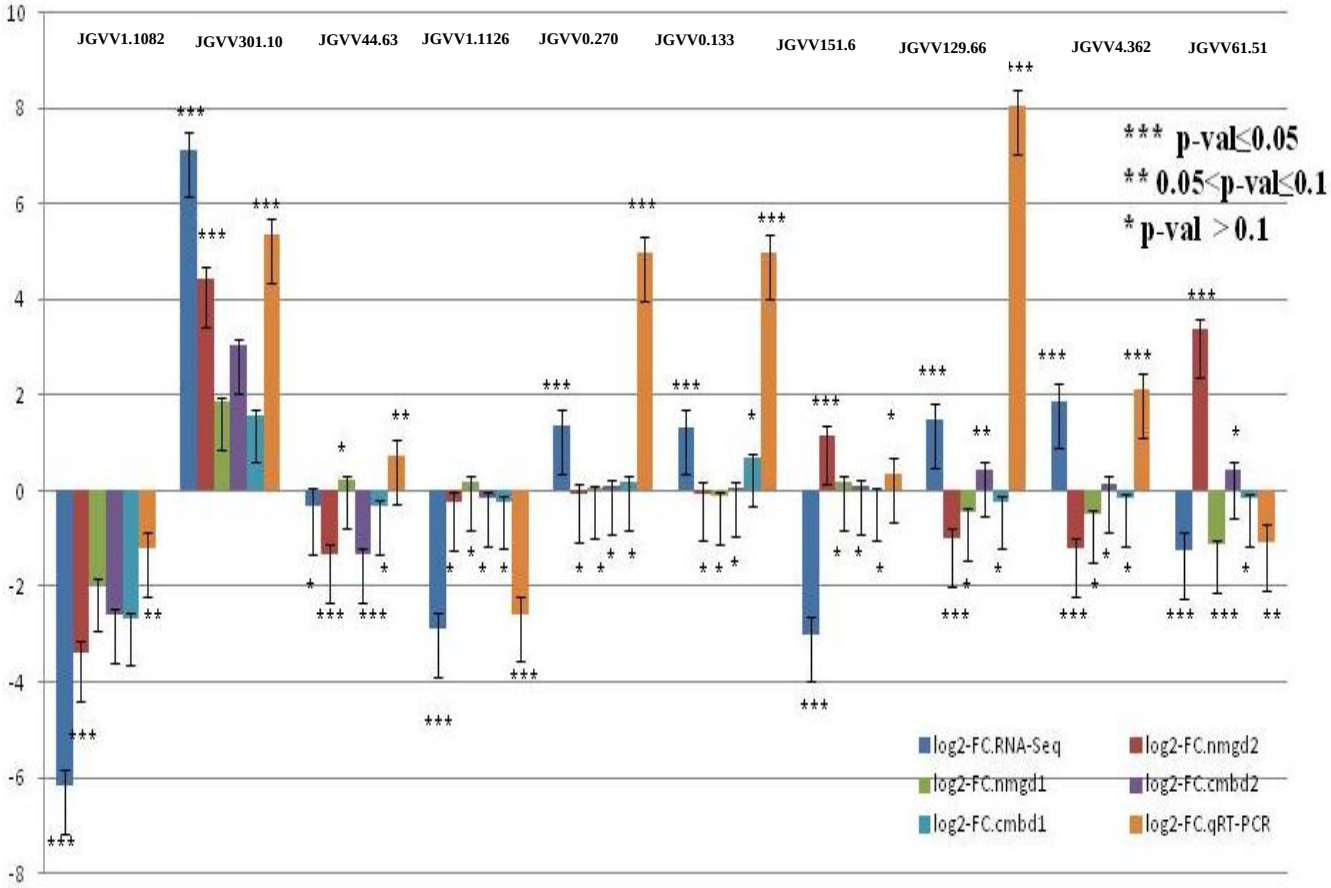

Figure12: RT-PCR, RNA-Seq and microarray log2-fold change values have been reported on *Y-axis* . (***)gene with p. adjusted values ≤ 0.05, (**) genes with 0.05<p. adjusted≤0.1, (*) genes with p. adjusted values >0.1.

5-Discussion and Conclusion

Although microarrays have been extensively used as discovery tools for biological and biomedical studies, the challenge remains whether this technology can be reliably applied in clinical practice and regulatory decision making, where high precision and accuracy in performance are required. A series of studies have been reported on evaluating performance across various commercial and homebrewed microarray platforms, However, most of these studies focused on evaluating the level of agreement across different microarray platforms. While these analyses emphasized critical issues such as the compatibility across different microarray platforms, they tended to result in conflicting conclusions because of the relative to relative nature of such approaches. What is lacking in these studies is a whole transcriptome gold standard data set that allows an evaluation of different microarray platforms based on a common ground truth. In this work we evaluated the performances of two strategies of array design (single replicated probe per transcript and different or multiple probes per transcript) based on CombiMatrix and NimbleGen microarray platforms assuming RNA-Seq tool as gold standard. The results from RNA-Seq were obtained on the same RNA samples used for microarray analysis as previously reported (*Zenoni et al.*, 2010). RNA-Seq appears to be an extremely promising technology measuring mRNA expression and identifying differentially expressed genes, comparable, and in some way superior, to existing array based approaches (*Marioni et al.*, 2008). This work allowed me to develop a statistical methodology based on sensitivity, specificity, accuracy and positive predictive value (PPV) parameters to evaluate the performances of different microarray design/platforms in differential analysis.

I showed that microarray designs based on multiple medium probes per transcript displayed a higher accuracy than microarray designs with a single specific replicate medium probe per transcript in differential analysis and that microarray design based on medium multiple probes per transcript calls less false positive significantly differentially expressed genes in comparison to microarray designs with a single

replicate medium probe per transcript. This result is in agreement with previous data (*Zhou and Abagyan*, 2002) showing that multiple short probes per gene were necessary to accurately measure the transcript abundance.

I also found that, microarray designs based on a single long probe per transcript are more specific than microarray designs based on multiple long probes per transcript, while microarray designs based on multiple long probes per transcript are more sensitive than microarray designs with a single long probe per transcript. Hence, microarray designs with a single long replicate probe per transcript discriminated less false positive significantly differentially expressed genes in differential analysis, while microarray designs based on multiple long probes per transcript called a higher number of significantly differentially expressed genes in differential analysis. This observation supported the high sensitivity of microarray design based on long probes as well as their low specificity as reported in literature (*Shingo Suzuki et al*., 2007) .

The work performed in the scope of this thesis also showed that regardless of the microarray design strategies used, microarray designs with long probes displayed a higher sensitivity but included more false positives significantly differentially expressed genes (low specificity) when RNA-Seq was considered as reference. By contrast microarray design with medium probes per transcript exhibited a low sensitivity and a high specificity in differential analysis when RNA-Seq was assumed as gold standard. Therefore I showed that long probes yield better signal intensity (high sensitivity) than medium probes in differential analysis (*Cheng Chung Chou et al*., 2004).

Our developed statistical methodology approach to assess the performance of microarray designs showed that, the positive predictive values (the capacity of each microarray design/platform to predict true positive significantly differentially expressed genes) is higher for *cmbd2* microarray designs based on multiple medium probes per transcript than those of *cmbd1* microarray designs based on a single replicate medium probe per transcript, while both NimbleGen microarray designs based on long probes per transcript exhibited a comparable positive predictive value

(PPV). Moreover, I also showed that microarray design based on multiple long oligos per transcript gave similar accuracy as microarray design with a single replicate long oligos per gene in differential analysis. These results are in agreement with Cheng Chung Chou et al., 2004, work that supported that accurate gene expression measurement can be achieved with multiple probes per gene and fewer probes are needed if longer probes rather than shorter probes are used.

Analyzing microarray expression data with Fisher's combined p-value test, the number of significantly differentially expressed genes detected by *cmbd2* and *nmgd2* microarray designs based on different probes per transcript confirmed by RNA-Seq increase. A good statistical method will have high power to detect differentially expressed genes but low false discovery rate (FDR) (*Ann Hess and Hari* Iyer, 2007). I therefore showed that regardless of the microarray platforms used, the Fisher combined probability method is a promising alternative to existing methods of testing for differential gene expression as supported by Ann Hess and Hari Iyer 2007 work.

Performing ROC curve analysis that combined microarray designs sensitivity and specificity in calling significantly differentially expressed genes assuming RNA-Seq as gold standard, I showed that the ability to discriminate significantly differentially expressed genes (DEG) confirmed by RNA-Seq is strongly evident between both analyzed microarray designs based on medium probes (35-40 mer) per transcript in comparison between both microarray designs based on long probes (60 mer) per transcript. However, the analysis showed that regardless of the microarray platforms and probes size used, *cmbd2* and *nmgd2* microarray designs with multiple probes per transcript exhibited a high performance in detecting significantly differentially expressed genes in comparison with their respective microarray designs based on a single replicate probe per transcript when RNA-Seq was assumed as gold standard. I therefore showed that the use of different oligo nucleotides per transcript provided an accurate measure of transcript abundance.

I evidenced some discrepancies between microarray and RNA-Seq approach in differential analysis. In fact I found that for genes with a high fold change ratio across

ripening and veraison development stages (FDR≤ 0.05 and fold change > 2) all analyzed microarray designs exhibited a good agreement with RNA-Seq, while for genes that displayed a low variation between ripening and veraison *Vitis Vinefera* development stages (low fold change ratio), the two approaches contrast. The disagreement observed between microarray and RNA-Seq could be due to some limitation of microarrays such as their low sensitivity evidencing small variation in differential analysis (*John C. Marioni et al.*, 2008).

The present work allows an assessment of the performance of two microarray design strategies based on two different microarray platforms on the largest reference data set (whole *Vitis vinifera* transcript model) of gene expression measurement. I showed that, the four analyzed microarray design/platform exhibit different performances (sensitivity, specificity, accuracy and predictive positive values) in differential analysis. However, this work proposes a statistical methodology based on comparison of microarrays with RNA-Seq data, which will help the investigator to choose the microarray platform/design that fits better the scientific goal.

Bibliography

A.A.V. Hill, J. Lu, M.A. Masino, O.H. Olsen and R.L. Calabrese. A Model of a Segmental Oscillator in the Leech Heartbeat Neuronal Network. Journal of Computational Neuroscience Volume 10, Number 3, 281-302, DOI: 10.1023/A:1011216131638; (2001).

Adam Roberts, Cole Trapnell, Julie Donaghey, John L Rinn and Lior Pachter :Improving RNA-Seq expression estimates by correcting for fragment bias. Genome Biology 2011, **12**:R22

Agarwal A, Koppstein D, Rozowsky J, Sboner A, Habegger L, Hillier LW, Sasidharan R, Reinke V, Waterston RH, Gerstein M: Comparison and calibration of transcriptome data from RNA-Seq and tiling arrays. *BMC Genomics* 2010, 11:383.
AlmutSchulze and Julian Downward Navigating gene expression using microarray a technology review. *Nature Cell Biology* Vol 3.; (2001 Aug).

Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer and Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* - **5**, 621 - 628 (2008); doi:10.1038/nmeth.1226.

Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, *et al.*: Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 2000, 403:503-511.

Anita Zamboni, Mariasole Di Carli, Flavia Guzzo, Matteo Stocchero, Sara Zenoni, Alberto Ferrarini, Paola Tononi, Ketti Tofalli, Angiola Desiderio, Kathryn S. Lilley, M. Enrico Pè, Eugenio Benvenuto, Massimo Delledonne and Mario Pezzotti. Identification of putative stage-specific grapevine berry biomarkers and omics data integration into networks. Plant Physiology November 2010 vol. 154 N°3 1439-1459.

Ann Hess and Hari Iyer. Fisher's combined p-value for detecting differentially expressed genes using Affymetrix expression arrays. *BMC Genomics* 2007, **8**:96 doi:10.1186/1471-2164-8-96 (2007).

Arbeitman MN, Furlong EE, Imam F, Johnson E, Null BH, Baker BS, Krasnow MA, Scott MP, Davis RW, White KP: Gene expression during the life cycle of Drosophila melanogaster. *Science* 2002, 297:2270-2275.

Blostad, B. M., Irizarry, R. A. Astrand, M. and Speed , T.P. A comparison of normalization method s for high density oligonucleotide array data based on variance and bias. *Bioinformatic* **19**, 185-193 (2003).

Brem RB, Yvert G, Clinton R, Kruglyak L: Genetic dissection of transcriptional regulation in budding yeast. *Science* 2002, 296:752-755.

Bradford J, Hey Y, Yates T, Li Y, Pepper S, Miller C: A comparison of massively parallel nucleotide sequencing with oligonucleotide microarrays for global transcription profiling. *BMC Genomics* 2010, 11:282.

Brian T. Wilhelm, Samuel Marguerat, Stephen Watt, Falk Schubert, Valerie Wood, Ian Goodhead, Christopher J. Penkett, Jane Rogers and Jürg Bähler. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* **453**, 1239-1243 (26 June 2008) | doi:10.1038/nature07002; (2008).

Brown PO, Botstein D. Exploring the new world of the genome with DNA microarrays. *Nat Genet*. (1 Suppl):33-7; (1999 Jan 21).

Chan ET, Quon GT, Chua G, Babak T, Trochesset M, Zirngibl RA, Aubin J, Ratcliffe MJ, Wilde A, Brudno M, Morris QD, Hughes TR: Conservation of core gene expression in vertebrate tissues. *J Biol* 2009, 8:33.

Cheng-Chung Chou, Chun-Houh Chen1, Te-Tsui Lee and Konan Peck. Optimization of probe length and the number of probes per gene for optimal microarray analysis of gene expression. Nucleic Acids Research, 2004, Vol. 32, No. 12 e99 doi:10.1093/nar/gnh099.

Chen, J.-N., van Eeden, F.J.M., Warren, K.S., Chin, A., Nüsslein-Volhard, C., Haffter, P., and Fishman, M.C. Development 124(21):4373-4382 (Journal); (1997).

Cloonan N, Forrest AR, Kolle G, Gardiner BB, Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G, Robertson AJ, Perkins AC, Bruce SJ, Lee CC, Ranade SS, Peckham HE, Manning JM, McKernan KJ, Grimmond SM. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat Methods*. 2008 Jul;5(7):613-9. Epub 2008 May 30.

Cole Trapnell, Lior Pachter and Steven L. Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25** (9): 1105-1111.; (2009).

David B. Allison, Xiangqin Cui, Grier P. Page and Mahyar Sabripour. *Nature Reviews Genetics* Vol 7 (2006).

De Lichtenberg U, Wernersson R, Jensen TS, Nielsen HB, Fausbøll A et al. (2005) New weakly expressed cell cycle-regulated genes in yeast *Yeast* **22**(15): 1191-1201.

Dobson CM, Wai T, Leclerc D, Wilson A, Wu X, Doré C, Hudson T, Rosenblatt DS, Gravel RA. Identification of the gene responsible for the cblA complementation group of vitamin B12-responsive methylmalonic acidemia based on analysis of prokaryotic gene arrangements. *Proc Natl Acad Sci* U S A.;99(24):15554-9.; (2002 Nov 26).

Dutta, A. and Bhattacharya, M. and Barat, P. and Mukherjee, P. and Gayathri, N. and Das, G. C, PhysRevLett.105.099602, doi:10.1103/*PhysRevLett*.105.099602, (2010 Aug).

Etienne Hollande, Sylvie Cantet, Ginette Ratovo, Ghislaine Daste, FranÃ§ois BrÃmont and Marjorie Fanjul. Growth of putative progenitors of type II pneumocytes in culture of human cystic fibrosis alveoli. *Biology of the cell* (2004) 96, (429-441).

Evertsz EM, Au-Young J, Ruvolo MV, Lim AC, Reynolds MA. Hybridization cross-reactivity within homologous gene families on glass cDNA microarrays. *Biotechniques*.**31**(5):1182, 1184, 1186 passim.; (2001 Nov).

Ezio Bottarelli and Stefano Parodi, 2003, Ann. Fac. Medic. Vet. di Parma (Vol. XXIII, 2003) - pag. 49 - pag. 68.

Fleige S, Pfaffl M: RNA integrity and the effect on the real time qRT-PCR performance. *Molecular Aspects of Medicine* 2006, 27:126-139.

Garge, N., Page, G.P., Spague, A. P., Gorman, B. S. and Allison, D. B. Reproducible cluster from microarray research : whither ? *BMC Bioinformatics* **6**, (Suppl. 2), S10 (2005).

Gary Hardiman Microarray platforms - comparison and contrast. *Pharmacogenomics* **5**, 487-502.; (2004).

Giles PJ, Kipling D. Normality of oligonucleotide microarray data and implications for parametric statistical analyses. *Bioinformatics*. **22**;19(17):2254-62; (2003 Nov).

Guo Z, Guilfoyle R A, Thiel A J, Wang R, and Smith L M. Direct fluorescence analysis of genetic polymorphisms by hybridization with oligonucleotide arrays on glass supports. Nucleic Acids Res.; 22(24): 5456–5465; (1994 December 11).

Goecks J, Nekrutenko A, Taylor J: Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* 2010, 11:R86.

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, *et al.*: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 1999, 286:531-537.

Griffith et al., 2010; Jiang and Wong, 2009; Lee et al., 2011; Mortazavi et al., 2008
Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D, Schmidt D, O'Keeffe S, Haas S, Vingron M, Lehrach H, Yaspo ML. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*. 2008 Aug 15;321(5891):956-60. (2008 Jul 3).

Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982; 143:29-36. (PMID 7063747).

Hardcastle TJ, KellyKA , baySeq: Empirical Bayesian Methods For Identifying Differential Expression in Sequence Count Data. *BMC Bioinformatics*; (2010).

Hatim T. Allawi and John SantaLucia, Jr. Department of Chemistry, Wayne State University, Detroit, Michigan 48202 *Biochemistry*, 1997, **36** (34), pp 10581–10594
DOI: 10.1021/bi962590c.

Hsiao, A. Worrall, D. S., Olefsky, J. M. and Subramaniam, S. Variance model posterior inference of microarray data: detecting gene expression change in 3T3-L1 adipocytes. *Bioinformatics* **20**, 3108-3127 (2004).

Hsiao HH, Yang MY, Chang JG, Liu YC, Liu TC, Chang CS, et al. Dihydropyrimidine dehydrogenase pharmacogenetics in the Taiwanese population. *Cancer Chemother Pharmacol*; **53**:445-51; (2004).

Hughes TR, Mao M, Jones AR, Burchard J, Marton MJ, Shannon KW, Lefkowitz SM, Ziman M, Schelter JM, Meyer MR, Kobayashi S, Davis C, Dai H, He YD, Stephaniants SB, Cavet G, Walker WL, West A, Coffey E, Shoemaker DD, Stoughton R, Blanchard AP, Friend SH, Linsley PS. Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. Nat **19**(4):342-7.; (2001 Apr).

Hui Jiang and Wing Hung Wong. Statistical inferences for isoform expression in RNA-Seq. Bioinformatics 25 (8): 1026-1032. doi: 10.1093/bioinformatics/btp113.; (2009).

Irizarry et al. *Nucleic Acids Res*. 2003; 31:e15 and *Biostatistics* 2003; 4:249

Hughes,T.R., Mao,M., Jones,A.R., Burchard,J., Marton,M.J., Shannon,K.W., Lefkowitz,S.M., Ziman,M., Schelter,J.M., Meyer,M.R. et al. (2001) Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat. Biotechnol*., **19**, 342–347.

Irizarry, R. A. Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.* **31**, e15 (2003).

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics.***4**(2):249-64.; (2003 Apr).

Irizarry, R. A. Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.* **31**, e15 (2003).

JA Swets. Measuring the accuracy of diagnostic systems.Science 3 June 1988: Vol. 240 no. 4857 pp. 1285-1293 DOI: 10.1126/science.3287615.

Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, et al. (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449: 463–467.

James H Bullard, Elizabeth Purdom, Kasper D Hansen, Sandrine Dudoit. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* 2010, 11:94.

Jay Shendure and Hanlee Ji**.** Next-generation DNA sequencing Nature *Biotechnology* **26**, 1135 - 1145 (2008) | doi:10.1038/nbt1486.

Jeremy Goecks, Anton Nekrutenko, James Taylor and the Galaxy. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life science *Genome Biology*, **11**:R86 doi:10.1186/gb-2010-11-8-r68; (2010).

John B. Welsh, Patrick P. Zarrinkar, Lisa M. Sapinoso, Suzanne G. Kern, Cynthia A. Behling, Bradley J. Monk, David J. Lockhart et al. 1996, Robert A. Burger§, and Garret M. Hampton. Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer; doi: 10.1073/pnas.98.3.1176 *PNAS*,vol. 98 no. 3 1176-1181 (2001 January 30).

John C. Marioni, Christopher E. Mason, Shrikant M. Mane, et al. RNA-Seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* 2008 18: 1509-1517 doi:10.1101/gr.079558108 (Jun 11, 2008).

Joseph D. Clarke and Tong Zhu. Microarray analysis of the transcriptomes as a stepping stone towards understanding biology system: practical consideration and perspectives. *The plant biology Journal* (2006) **45**, 630-650.

Justin O. Borevitz, David Liang, David Plouffe, et al. Large-Scale Identification of Single-Feature Polymorphisms in Complex Genomes.; *Genome Res.*13: 513-523; (2003).

Kai T, Williams D, Spradling AC: The expression profile of purified Drosophila germline stem cells. *Dev Biol* 2005, 283:486-502.

Kane MD, Jatkoe TA, Stumpf CR, Lu J, Thomas JD, Madore SJ. Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Res*. **15**;28(22):4552-7.; (2000 Nov).

Kerr MK. Design considerations for efficient and effective microarray studies. *Biometrics*.**59**(4):822-8; (2003 Dec).

Kwon, S.Y., Xiao, H., Glover, B.P., Tjian, R., Wu, C., Badenhorst, P. The nucleosome remodeling factor (NURF) regulates genes involved in Drosophila innate immunity. *Dev. Biol.* 316(2): 538-547.; (2008).

Leung YF, Cavalieri D. Fundamentals of cDNA microarray data analysis. *Trends Genet.***19** (11):649-59 (2003 Nov).

Li,F. and Stormo,G.D. Selection of optimal DNA oligos for gene expression arrays. *Bioinformatics*, **17**, 1067–1076.; (2001).

Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, Brown EL. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol*.**14**(13):1675-80.; (1996 Dec).

Lockhart DJ, Winzeler EA Genomics, gene expression and DNA arrays. *Nature*.405(6788):827-36.; (2000 Jun 15).

Luo, J. Dunggan DJ, Chen Y et al. Human prostate cancer and benign prostatic hyperplasia: molecular dissection by gene expression profiling. *Cancer Res.* **61**, 4683–4688 (2001).

Luo J, Nikolaev AY, Imai S, Chen D, Su F, Shiloh A, Guarente L, Gu W. Negative control of p53 by Sir2alpha promotes cell survival under stress. *Cell*.**107**(2):137-48. (2001 Oct 19).

Mangalathu S. Rajeevan,1 Daya G. Ranamukhaarachchi, Suzanne D. Vernon, and Elizabeth R. Unger. Use of Real-Time Quantitative PCR to Validate the Results of cDNA Array and Differential Display PCR Technologies. *Methods* **25**, 443–451 doi:10.1006/meth.2001.1266; (2001).

Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* 2008 Sep;**18** (9):1509-17. Epub 2008 Jun 11.

Mark D. ,Robinson Davis J. ,McCarthy, Gordon K. Smyth: edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**(1): 139-140; (2010).

Mei R., Earl Hubbell, Stefan Bekiranov, Mike Mittmann, Fred C. Christians, Mei-Mei Shen, Gang Lu, Joy Fang, Wei-Min Liu, Tom Ryder, Paul Kaplan, David Kulp, and Teresa A. Webster. Probe selection for high-density oligonucleotide arrays. Proc. *Natl. Acad. Sci.* U. S. A. 100, 11237–11242; (2003).

Miller MA, Nguyen VQ, Lee MH, Kosinski M, Schedl T, Caprioli RM, Greenstein D. A sperm cytoskeletal protein that signals oocyte meiotic maturation and ovulation. *Science*. **16**;291(5511):2144-7; (2001 Mar).

Mir, K.U. and Southern, E.M. (1999) Determining the influence of structure on hybridization using oligonucleotide arrays. *Nat. Biotechnol*. **17**, 788–792.

Mark D Robinson and Alicia Oshlack. A scaling normalization method for differential exression analysis of RNA-seq data. *Genome Biology*, **11**:R25 doi:10.1186/gb-2010-11-3-r25 (2010).

Modrek, B. et al. (2001) Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res*. **29**, 2850–2859.

Muller, R. A., Galecki, A. and Shmookler Reis, R.J. interpretation, design and analysis of gene array expression experiment. *J. Gerontol*. **A 56**, B52-B57 (2001).

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

Nielsen, H. B., Gautier, L. & Knudsen, S. Implementation of a gene expression index calculation method based on the PDNN model. *Bioinformatics* **21**, 687–688 (2005).

Nuwaysir EF, Huang W, Albert TJ et al.: Gene expression analysis oligonucleotide arrays produced by maskless photolithography. *Genome Res.* **12**, 1749-1755 (2002).

Park PJ, Cao YA, Lee SY, Kim JW, Chang MS, Hart R, Choi S: Current issues for DNA microarrays: platform comparison, double linear amplification, and universal RNA reference. *J Biotechnol* 2004, 112:225-245.

Paul L. Auer and Doerge R. W. Statistical Design and Analysis of RNA Sequencing Data; DOI: 10.1534/*genetics*.110.114983; (2010).

Pawitan Y, Bjöhle J, Amler L, Borg AL, Egyhazi S, Hall P, Han X, Holmberg L, Huang F, Klaar S, Liu ET, Miller L, Nordgren H, Ploner A, Sandelin K, Shaw PM, Smeds J, Skoog L, Wedrén S, Bergh J. Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast Cancer Res*. 2005;7(6):R953-64.; (2005 Oct 3).

Peter B Dallas, Nicholas G Gottardo, Martin J Firth, Alex H Beesley, Katrin Hoffmann, Philippa A Terry, Joseph R Freitas, Joanne M Boag, Aaron J Cumming and Ursula R Kees: Gene expression levels assessed by oligonucleotide microarray analysis and quantitative real-time RT-PCR- how well do they correlated? *BMC Genomics*, **6**:59 doi:10.1186/1471-2164-6-59.; (2005).

Peterson, A.W., Wolf, L.K., Georgiadis, R.M. 2002 Hybridization of mismatched or partially matched DNA at surfaces J. Am. *Chem. Soc*.12414601–14607.

Petricoin EF, Hackett JL, Lesko LJ, Puri RK, Gutman SI, Chumakov K, Woodcock J, Feigal DW Jr, Zoon KC, Sistare FD: Medical applications of microarray technologies: a regulatory science perspective. *Nat Genet* 2002, **32**: 474-479.

Pfaffl MW. (2001) A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res* **29**: e45.

Qin, L. X. and Kerr, K. F. Empirical evaluation of data transformation and raking statistic for microarray analysis. *Nucleic Acid Res*. **32**, 5471-5479 (2004).

Quackenbush J. Microarray data normalization and transformation. *Nat Genet*.**32** Suppl:496-501.; (2002 Dec).

Rajeevan, M.S., Vernon, S.D., Taysavang, N. & Unger, E.R. Validation of array based gene expression profiles by real-time (kinetic) RT-PCR. *J. Mol. Diagnosis* **3**, 26–31 (2001).

Ramakers C, Ruijter JM, Deprez RH, Moorman AF. (2003) Assumption-free analysis of quantitative real-time polymerase chain reaction (PCR) data. *Neurosci Lett* 339: 62–66.

Reimers M: Making informed choices about microarray data analysis. *PLoS Computat Biol* 2010, 6:e1000786.

Rockett JC, Hellmann GM: Confirming microarray data- it is really necessary? *Genomics*, **83**:541-549, (2004).

SantaLucia, J., Jr. et al. (1996) Improved nearest-neighbor parameters for predicting DNA duplex stability. *Biochemistry* **35**, 3555–3562.

Sara, Z., Alberto, F. Enrico, G., Luciano, X., Marianna Fasoli, Giovanni Malerba, Diana Bellin, Mario Pezzotti, Massimo Delledonne. Characterization of tranxscriptional complexity during berry development in Vitis viifera usign RNA-Seq. Plant Physiology doi:10.1104|pp. 109.149716; (2010).

Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*.;270(5235):467-70.; (1995 Oct 20).

Shingo Suzuki, Naoaki Ono, Chikara Furusawa, Akiko Kashiwagi1 andTetsuya Yomo. Experimental optimization of probe length to increase the sequence specificity of high-density oligonucleotide microarrays. *BMC Genomics* 2007, **8**:373 doi:10.1186/1471-2164-8-373.

Simon Anders and Wolfgang Huber: Differential expression analysis for sequence count data. *Genome Biology*, **11**:R106, (2010).

Smyth, G. K. (2005). Limma: linear models for microarray data. In: Bioinformatics and Computational Biology Solutions using R and Bioconductor, R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, W. Huber (eds.), Springer, New York, pages 397–420.

Smyth, G. K. (2005). Limma: linear models for microarray data. In: *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, R.

Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B: Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell* 1998, 9:3273-3297.

Stafford P. Liu P. Microarray technology comparison, statistical analysis and experimental design. *Microarray Methods and Application* –Nut and Bolt. DNA Press, Eagleville, PA USA 273-324 (2003).

Stewart J. P., Liu A. H. and Choi Y., Amplification factors for Spectral Acceleration in Tectonically active Regions, Bull. Seism. Soc. Am., Vol. 93, No. 1, 332-352. (2003).

Sydney Brenner, Maria Johnson, John Bridgham, George Golda, David H. Lloyd, Davida Johnson, Shujun Luo, Sarah McCurdy, Michael Foy, Mark Ewan, Rithy Roth, Dave George, Sam Eletr, Glenn Albrecht, Eric Vermaas, Steven R. Williams, Keith Moon, Timothy Burcham, Michael Pallas, Robert B. DuBridge, James Kirchner, Karen Fearon, Jen-i Mao, and Kevin Corcoran. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays *nature biotechnology* vol 18; (2000 Jun).

Sugimoto, N., Naoki Sugimoto, Mariko Nakano, and Shu-ichi Nakano, Thermodynamics-structure relationship of single mismatches in RNA–DNA duplexes. *Biochemistry* **39**, 11270–11281; (2000).

Sugimoto, N. ,Ozutsumi, K. and Matsuda, M. Purification by high performance liquid chromatography of *Clostridium perfringens* type a enterotoxin prepared from high toxin producers selected by a toxin-antitoxin halo. *European Journal of Epidemiology* Volume 1, Number 2, 131-138, DOI: 10.1007/BF00141806; (2000).

Taniguchi T, Ogasawara K, Takaoka A, Tanaka N. IRF family of transcription factors as regulators of host defense. *Annu Rev Immunol.* 2001;19:623-55.

Telonis-Scott M, Kopp A, Wayne ML, Nuzhdin SV, McIntyre LM: Sex-specific splicing in Drosophila: widespread occurrence, tissue specificity and evolutionary conservation. *Genetics* 2009, 181:421-434.

Thomas J Hardcastle and Krystyan A Kelly. baySeq: Empirical Bayesian methods for identifiying differential expression in sequance count data *BMC Bioinformatic*s 2010, 11:422 doi:10.1186/1471-2105-11-422.

Torres J.J., Marro J., Cortes J.M., Wemmenhove B. Instabilities in attractor networks with fast synaptic fluctuations and partial updating of the neurons activity Neural Networks 21 1272_1277; (2008).

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol.* 2010 May;28(5):511-5.; (2010 May 2).

Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci* U S A. 2001 Apr 24;98(9):5116-21. Epub 2001 Apr 17.

Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. *Science*. 1995 Oct. 20;270(5235):484-7.

Wang, Y., Klijn JG, Zhang Y, Sieuwerts AM, Look MP, Yang F, Talantov D, Timmermans M, Meijer-van Gelder ME, Yu J, Jatkoe T, Berns EM, Atkins D, Foekens JA. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* 365, 671–679; (2005).

Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*. 2009 Jan;10(1):57-63.

Wolfgang Becker, Axel Bergmann, Christoph Biskup, Laimonas Kelbauskas, Thomas Zimmer, Nikolaj Klöcker, Klaus Benndorf. High resolution TCSPC lifetime imaging; *Proc. SPIE* 4963 (2003).

Wu, C. et al. (2005) Sequence dependence of cross-hybridization on short oligo microarrays. *Nucleic Acids Res*. **33**, e84.

Xiaodi Su,Ying-Ju Wu,Wolfgang Knoll (Peterson, A.W. et al., 2002), Comparison of surface plasmon resonance spectroscopy and quartz crystal microbalance techniques for studying DNA assembly and hybridization, *Elsevier Biosensors and Bioelectronics*; (2005 November 15).

Xu W, Chen H, Du K, Asahara H, Tini M, Emerson BM, Montminy M, Evans RM. A transcriptional switch mediated by cofactor methylation. *Science*. 2001 Dec 21;294(5551):2507-11. Epub 2001 Nov.

Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res*.**30**(4):e15.; (2002 Feb 15).

Yoav Benjamini; Yosef Hochberg. Controlling the False Discovery Rate: A Powerful Approach to Mutiple Testing. *Journal of the royale Statistic Society*. Series B (Methodological), Vol 57. 57 N°1 (1995), 289-300.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A particular and powerful Approach to multiple. Journal of the Royal Statistic Society Series B (Methodological), Vol. 57, N°1, 289-300 ;(1995).

Young T, Peppard P Sleep-disordered breathing and cardiovascular disease: epidemiologic evidence for a relationship. Sleep.;23 Suppl 4:S122-6.; (2000 Jun 15).

Zhang J. et al. :Detecting false expression signals in highdensity oligonucleotide arrays by an in silico approach. *Genomics* **85**, 297–308, (2005).

Zhang, L., Miles MF, Aldape KD A model of molecular interactions on short oligonucleotide microarrays. *Nat. Biotechnol*. 21, 818–821; (2003 Jul).

Zhongxue Chen, Monnie McGee, Qingzhong Liu and Richard H. Scheuermann. A distribution free summarization method for Affymetrix GeneChip arrays *Bioinformatics* (2007) 23 (3): 321-327. doi: 10.1093/bioinformatics/btl609.

Zou L, Harkey MR, Henderson GL. Effects of herbal components on cDNA-expressed cytochrome P450 enzyme catalytic activity. *Life Sci.***16**;71(13):1579-89.; (2002 Aug).