Analysis of microarray gene expression data sets

© Lars M.T. Eijssen, Schimmert, 2006

ISBN-10: 90-9021327-9

ISBN-13: 978-90-9021327-9

Cover design

Lars Eijssen

Illustrations Chapter 1

Mike Gerards

Printed by

Drukkerij Econoom BV Beek-L

Analysis of microarray gene expression data sets

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit Maastricht,
op gezag van de Rector Magnificus,
Prof.mr. G.P.M.F. Mols,
volgens het besluit van het College van Decanen,
In het openbaar te verdedigen
op dinsdag 19 december 2006 om 16.00 uur

door

Lars Maria Theo Eijssen

geboren te Schimmert op 30 mei 1976

Promotor

Prof.dr. J.P.M. Geraedts

Copromotores

dr. P.J. Lindsey

dr. H.J.M. Smeets

Beoordelingscommissie / Assessment committee

Prof.dr. J.C.S. Kleinjans (voorzitter)

Prof.dr. F.C.P. Holstege (UMC Utrecht)

Prof.dr. C.A.J. Klaassen (Universiteit van Amsterdam)

Prof.dr. Y.M. Pinto

Prof.dr. E.O. Postma

The studies presented in this thesis were performed at the Department of Genetics and Cell Biology, Cardiovascular Research Institute Maastricht (CARIM), Maastricht University, Maastricht, the Netherlands

‘μαιευτική τέχνη’
-Socrates-

Table of contents

Chapter 1	
General introduction	9
Chapter 2	
A novel stepwise analysis procedure of genome-wide expression profiles identifies transcript signatures of thiamine genes as classifiers of mitochondrial mutants in yeast	59
Chapter 3	
Myocardial gene expression reveals maladaptive processes in cardiac myosin binding protein C knock-out mice	79
Chapter 4	
Affymetrix expression chip data analysis: the gain of modeling	125
Chapter 5	
Multivariate normal probe modeling for Affymetrix expression chip data	147
Chapter 6	
The use of spikes in Affymetrix chip expression data analysis	177
Chapter 7	
General discussion	205
Summary	217
Samenvatting	223
Dankwoord	229
Curriculum Vitae	233

Chapter 1

General introduction

From genetics to genomics

Since its start but especially during the last decades the field of genetics has gone through major changes. Up till the nineties of the last century focus within genetics was on chromosomal abnormalities and monogenic diseases, disorders characterized by one of the classical Mendelian patterns of inheritance. The research in the area of monogenic disease was devoted to find the causing locus (gene) and to unravel the function of the protein encoded by the gene and the mechanism of the protein pathways it functioned in. Sometimes the situation was characterized by genetic heterogeneity, where defects in more than one gene can cause the same clinical phenotype, or pleiotropy (phenotypic heterogeneity), where defects in one gene can cause one of several clinical phenotypes.

Since about 15 years, a development has taken place from genetics to so-called genomics, which is characterized by the simultaneous study of many genes and/or gene products at the same time. The human genome project, by which almost all human genes have become known [1, 2], has been one of the biggest stimuli to the genomics approach. In parallel with the sequencing of the human genome, many other eukaryotic genomes have been sequenced, e.g. those of the mouse [3] and yeast [4]. Furthermore, for many genes it has become known in which molecular pathways their gene products are involved.

The transition to genomics has been fuelled by an enormous increase in the technological possibilities available for genetic and molecular studies. Especially the field of gene expression analysis has profited from the development of several technical platforms, amongst which the microarray or chip technology has been one of the major achievements [5-7]. In general, microarrays are glass slides that contain particular molecules (probes) attached to their surface, each of which can specifically bind a particular target molecule. Microarrays are used for functional genomics research (retrieving differentially expressed genes), target discovery, biomarker determination, pharmacology, toxicology (to find effects of respectively drugs and toxic compounds), predicting disease prognosis, and subclassifying disease [8].

A corollary of the availability of high-throughput technology for research is an increased focus on far more frequent population diseases such as cardiovascular disease, diabetes, obesity, and others. These are caused by the interaction of several genes and the environment and require genome-wide approaches to detect the genes involved. These diseases are especially complex to study, since all of the contributing genes and effects interact with each other and only partially explain the disease [9, 10].

A direct result from this increase in scale is that experimental outcome can no longer be judged by eye and computational systems have become needed to interpret the results produced by these novel techniques. Per study typically tens of thousands or even hundreds of thousands of values are produced. Fortunately, at the same time also the computational power of the available hardware platforms has strongly increased, as dictated by Moore's law [11]. However, the increase in number of

features of the different microarray platforms is consistently a step ahead of the increase in computer power, keeping the analysis of the biggest chips a great challenge. For many chips, data analysis is possible on desktop or laptop machines, but for some a dedicated (server) system is needed. Besides computational power, of course also analytical methods are needed to process results. Before elaborating on those, a more detailed description of the microarray platform and its background is given. In Figure 1, an overview of the total experimental procedure for a gene expression microarray study is presented.

Figure 1 Overview of the general complete procedure followed when performing a gene expression microarray experiment. Throughout this text all steps will be discussed.

Gene expression

In order for the cell to produce proteins, DNA (deoxyribonucleic acid) is first transcribed into mRNAs (messenger ribonucleic acids) that are transported outside the nucleus to be translated into proteins. If this happens the gene is said to be expressed. At the DNA and mRNA level each triplet of bases codes for a certain amino acid, the building block of proteins. Besides the code to be translated, the mRNA molecule also contains control sequences in the untranslated regions (UTRs) at both ends of the molecule. As long as the mRNA is not degraded by the cellular machinery, it can be used to build more copies of the protein it encodes.

Since the completion of the human genome project it is clear that the great complexity of our genome does not lie in the number of genes (around 30000), but in the much larger number of proteins produced from those (more than 300000). The clue to this order of magnitude difference is alternative splicing, which beholds that from a single gene several protein products can be made. On the DNA level the sequences within a gene encoding parts of the mRNA (so called exons) are interrupted by sequences that are not transcribed (introns) and have to be split out. Alternative splicing is then performed by also leaving out one or more of the exons, to produce several different types of mRNA from the same genetic locus (gene).

Based on the processes of transcription and translation, a diversity of methods can be used to monitor the molecular functioning of the cell or parts of it. In accordance with the word *genome*, used for the entire DNA - nuclear, mitochondrial, chloroplastic - in the cell, the collection of all RNAs present in the cell is called the *transcriptome* and the collection of all proteins the *proteome*. Several systems have been developed to massively monitor either of those collections in parallel, where microarrays abound in each group. DNA microarrays have been developed to either measure common variances (single nucleotide polymorphisms, SNPs) on a genome-wide basis, or to sequence long stretches of DNA to find mutations in specific regions or genes. To detect expression levels of tens of thousands of RNA molecules or even the whole known transcriptome of the organism at hand, gene expression arrays have been developed. Finally, also for the detection of the proteome, microarrays are available. Because the protein is the eventual functional product within the cell, a proteomic array most directly measures the number of effective molecules present. However, because of the far more demanding complexity of proteome measurements, the mRNA expression array has been the first type of array to be used on a wide scale. The next sections discuss the gene expression platform in more detail, where the word '(micro)array' is meant to refer to gene expression microarray.

Microarrays

In very basic terms, a microarray is a slide of glass that contains many probes, in this case sequences of nucleotides, attached to its surface. Each of these probes specifically recognizes a certain mRNA molecule by hybridization. The copies of the probe recognizing a certain transcript are spotted together at the same position on the slide. When labeled sample material is brought onto the slide, transcripts bind to their respective probes. Because it is known which probe is at which location, scanning label intensities gives a parallel measurement of the abundances of all transcripts recognized. Figure 2 shows examples of scans of slides of two microarray platforms.

Figure 2 Zoomed parts of scans of two different array types: a) a custom home-made slide; b) a commercial Affymetrix GeneChip.

At the start of the era of microarray technology, when arrays were still limited to detecting a few thousand transcripts, the probes represented a sample of all genes in a certain genome or a thematic collection of genes (e.g. those expressed in a certain tissue, related to a certain biological process, or expected to be related to a type of disease). With the rapid increase in the number of probes per slide, at this moment most arrays are general, covering most or all of a certain genome. These arrays enable, for example, to detect differences between samples from patients suffering from a certain disease and samples from healthy controls on a genome-wide scale. Several types of microarray systems have been developed, considering both the way the probes to detect the different transcripts are constructed and the number of samples that can be screened on the same slide. These are discussed next, after an overview of the general procedure used to perform a microarray experiment.

Laboratory protocol

The general laboratory protocol (standard operating procedure, SOP) used for doing microarray studies contains several steps [8, 12], where each platform and system has its own specificities. However, a general outline would consist of the following steps (cf. Figure 1). First RNA is extracted from the sample (e.g. biopsy, cell culture) at hand. This can be total RNA (also including other types of RNA than mRNA) or total mRNA. Next, this material is amplified and labeled with a fluorescent dye. Labeling can be done by incorporating labeled nucleotides, or by attaching a tag (to be) stained with fluorescence. These steps generate enough copies of labeled material to perform the next step, hybridization of the labeled and amplified sample to the slide containing the probes to detect many transcripts. Finally, after washing non-specifically bound material from the slide, the slides are scanned with an array laser scanner. This scanner radiates the label, making it to fluoresce and detects this fluorescence. Scanning results in an image of spots, from which an intensity can be computed for each spot. This is done by an image quantitation algorithm that composes the scanned pixel intensities that make up each specific spot into a single spot intensity. For this quantitation many high-quality software packages are available and usually supplied with the scanners, that only minimally differ in the results they produce [13]. They vary in the way the spots are found and how they are defined or in other words how the (rim of) background is separated from the foreground [14]. The calculated spot intensities are the output of the wet bench (laboratory) and the input to the dry bench (bioinformatics analysis), where further processing by one of the several microarray analysis methods should lead to a clearer view on the ongoing processes. Each of the technical steps in the wet bench procedures can introduce variation that needs to be taken into account properly by later analysis.

cDNA versus oligonucleotide arrays

One type of array is the cDNA (complementary DNA) slide, for which the probes are complete cDNAs complementary to the mRNAs they are intended to detect. Thus the entire sequence of the mRNA at hand is recognized. This is a very straightforward approach but it suffers from cross-hybridization due to sequence similarity between the different mRNAs present in the hybridized sample. First of all, many mRNAs will contain domains or stretches of sequence that are very aspecific or generally present over many RNAs, since they may correspond to very general protein features such as transmembrane or other domains. Furthermore, several mRNAs are transcribed from families of genes that share most of their sequence. Finally, the proteins produced by any of the many alternatively spliced genes generally share most of their sequences. The other type of array is the oligonucleotide slide. An oligonucleotide (or oligo) is a short sequence of nucleotides. On this type of slide the probe sequences used are in general between 20 and 80 bases long. Systems using both short and long oligos exist, but on a single system all oligos are generally of the same length. Each oligo is designed carefully to be specific for the mRNA to be detected. This means that in

principle cross-hybridization should occur less often for oligo arrays than for cDNA arrays. However, in some cases it is still difficult to design an oligo fulfilling all requirements, especially in case of gene families or alternative splicing where the (few) unique parts of the sequences may not contain a suitable target location for oligo design. Also an oligo that was thought to be unique at the time of design, might be shown not to be at a later point when more transcripts have been discovered [8]. An advantage of oligo-based arrays is that it is also possible to design several different oligos per transcript, which, if used together, can lead to more reliable measurements.

Recent types of arrays specifically designed to measure the many alternatively spliced gene products, are exon-specific and exon-junction arrays, for example available for the Affymetrix platform. In the first type, specific probes are used for each exon to detect splicing variants, whereas in the second type the creation and loss of specific exon junctions by alternative splicing is exploited by targeting those junctions.

For both the cDNA and oligo types of array, all probes with the same sequence, are attached to the slide in a certain small predefined area. This ensures that it is known exactly which transcript is recognized at which position on the slide. For in-house spotting of slides by an arrayer (or spotter), several cDNA and oligo collections are available. These either contain all transcripts or general subsets of a certain genome, or specific subsets of transcripts for developmental stage (embryonic, fetal, adult), tissue or pathology. A problem with these sets is that they often come without warranty, so it is not guaranteed that probes for the genes indicated are indeed contained and that their quality is good. When tested, over a third of the probes in a cDNA clone set did not correspond to the gene intended, or was contaminated [15]. For oligos, also the specificity may be an issue. This means that either the clone collection is best screened in-house before being used, or an expensive sequence verified collection must be purchased.

Single color versus dual color arrays

Another important distinction between array platforms considers the number of samples measured on the same slide. The two most frequently used techniques have one and two samples hybridized to a single slide respectively (Figure 3).

In case of one sample, the expression matrix of spot intensities resulting from one slide contains an absolute expression value for each transcript. These values have to be combined with those from other slides to get to group or sample differences in expression. In case of two samples, two distinguishable labels which fluoresce at a different wavelength (dyes or 'colors') are used. By scanning, an image is obtained for each of the two labels separately and as such an absolute expression value is obtained for each of the two samples. These can be directly used or combined into paired values, where for each transcript a relative change in expression, or a fold change, can be computed. The advantage of co-hybridizing samples is that the experimental variability introduced between two samples is less than when the results

of two slides have to be compared. As such an important design question for dual color arrays is which samples have to be paired to be on a single slide (cf. section on experimental design for microarray studies) [16]. On the other hand, a disadvantage of dual color slides is that two different dyes have to be used, that will differ in their properties. Indeed the two most commonly used dyes, Cy3 and Cy5, vary considerably in their properties [16, 17]. One specific dye-effect to be taken into account is the instability of Cy5-dye that has been reported to be affected by atmospheric ozone [18].

Figure 3 Dual and single colors array experiments: a) in the dual color array case, two samples are amplified and labeled with different dyes (colors) and are co-hybridized to the same slide; b) in the single color array case only one sample is amplified, labeled, and hybridized per slide. For dual color arrays, after combining scans of the intensities of the two dyes, both absolute and relative expression is retrieved: absolute by the intensity of the spot and relative by its color, where red indicates upregulation, yellow indicates no change and green indicates downregulation. For single color arrays only absolute intensities can be determined and slides have to be combined to compute relative expression.

Also arrays with three or more colors are possible, though at the moment the technical performance of those platforms is still somewhat less [19] and they have not yet been widely used. The several types of arrays generally used are depicted in Figure 4.

Figure 4 Types of microarray platforms can be divided in those that measure variations in the DNA (SNP or resequencing arrays), mRNA expression, or protein abundance. Here focus is on gene expression arrays, which can use cDNA or oligonucleotide probes to recognize specific transcripts. Oligo-based arrays exist in two types: short and long oligo, depending on the length in number of bases of the oligos used. Furthermore either one or two samples are hybridized to the same slide, resulting in single or dual color arrays. For the several types or arrays, both custom home-made and commercial platforms exist.

Experimental design for microarray studies

Experimental design [20] is a very important first step when performing array studies and consists of several aspects. One of the aspects is the number of samples to be used and whether or not to pool some of them. Also, it is needed to take disturbing effects into account, to be able to control for them upon analysis. When using dual color slides, it is important to determine which samples are paired for hybridization on the same slide. Furthermore, when custom-made chips are used, the layout of the spots on the slides is a relevant design issue.

Observations

Collecting multiple observations is important, both between slides and within slides since estimates of variation are needed for any statistical technique to be applied [16]. Observational units between slides are now discussed (repeats within slides are covered in the section on slide design for custom-made arrays), where two types must be distinguished. A replication refers to redoing the complete procedure for another sample to obtain an independent extra measurement (unit). On the other hand, a repeat refers to redoing the procedure for the same sample (or another part

of the same sample) to obtain a second measurement on this sample. Hence, the need to have the observational unit clearly defined in order to be able to distinguish between replicates and repeats. Because for microarray studies the unit of investigation will generally be an organism, say a patient, a replication will involve a sample from another patient. On the other hand, a repeat may start with another sample from the same patient (a biological repeat) or may repeat the procedure for the same sample (a technical repeat).

Variation between replicates is generally much higher than between repeats. In an experiment both types of observational units can be collected alongside. The question of how many observations are needed is difficult to answer, because classical power or sample size calculations are difficult to adapt to array data [14]. For studies searching for genes with a different expression between groups, several techniques have been developed to estimate sample size and/or power [21-25]. Overall the need for replicates is generally established [26-28], whereas different numbers needed have been given [29-32]. It has been suggested that if more groups are present less replicates can be used per group [31]. The number needed may also depend on the biological material at hand: the more variable the material the more replicates needed [33]. Also the number required depends on the goal of the study, and is different for studies classifying samples as compared to studies searching for genes that differ in expression [34-36]. Finally, using Power Atlas, the number of replicates needed for a study can be predicted on the basis of available data sets [37].

In contrast to replicates, the need for repeats is not well-established [14, 16]. However one study suggests that twofold changes in expression could be reliably detected with three repeats available [38]. Correlation observed from repeated spots on the same array is generally over 95%, but between arrays (in case of duplos) it often drops to 60-80%. Between samples, correlation may even be half of this and between labs even lower [27]. A form of repeat specific to dual color slides is the dye-swap experiment, i.e. redoing measurements with reversed dyes [16, 17, 27, 39, 40]. As mentioned above, differences between the dyes can be considerable, thus dye swaps are important. Hence, if a two color experiment is carried out with repeats, these should be done as dye swaps. A variant of a dye swap would be to replicate with two other samples (rather than repeat the same samples) after switching dyes between groups. This variant uses two times less slides, but yields more variable results. In addition, it should be ensured that an equal number of samples from each experimental group is assigned to both dyes.

Pooling

Whether or not pooling is a good idea has been investigated thoroughly [14, 26, 27, 41, 42]. However, the authors of these studies have come to different conclusions. One of the reasons to pool is to reduce cost (by reducing the number of slides needed by grouping samples), especially if array cost is much higher than sample cost. Sometimes pooling may be necessary due to limited amounts of RNA.

However, with equal amounts of slides, using n individual cases is better than n pools of all n cases. For clustering in order to find new subtypes of disease, pooling cannot be performed, because it may combine individuals from several subclasses. For classification studies pooling is not advisable, because no estimate of group variability can be given any more, unless still more pools per group are used. In contrast, others indicate that pooling makes classification and prediction more reliable [40], improves efficiency and reduces cost [43]. The opposite is indicated as well [44] and some give formulas to calculate when pooling is a better option and when not [45]. In general, pooling always will lead to loss of information and is as such not advisable, but may have practical and technical advantages in some cases. One frequently used application of pooling, using a common reference pool to compare each sample to, is discussed in the section on study design for dual color arrays.

Disturbing effects

Two types of disturbing effects can be present: those introduced by drawing samples from the population under study, and those introduced by the application of the study protocol.

The first include differences in age, gender, BMI (body mass index), or any property of the population leading to differences in expression of genes of interest between the groups investigated. To control for these, each sample may be paired to a random sample from the other group, but this may not be the best option to take due to small sample sizes. The pairing may also be done to control-samples matched for age, gender, *et cetera*. This may not always be possible, due to the availability of too few samples, where especially controls are often difficult to obtain. Also, all important disturbing effects need to be known and it may be difficult to match for each of those at the same time. Another option is to restrict inclusion criteria to, for example, only males or only adults. In cases where assignment to groups can be determined by the researcher, randomized assignment can be used. However, when dealing with a small number of samples it may not be easy to randomize all disturbing effects over the groups. Still, to prevent adverse effects, careful determination of design is important.

The second type of disturbing effect includes differences in amount, composition or quality of starting material, differences in amplification, batches, labeling or hybridization efficiency, and overall scanning differences. All these effects can occur between slides, or in case of dual color slides between channels and slides. Also within slides several of these differences (e.g. hybridization or scanning effects) may be present, between regions of the slide. Further within-slide differences may result from scratches or dirt on the slides, which may even lead to the exclusion of certain regions of the slide from analysis. Furthermore, if custom-made slides are used, differences between spotting runs - i.e. between metagrids - or pin-effects - i.e. differences between subgrids - may be present (cf. section on slide design for

custom-made arrays). These effects need to be identifiable to be taken into account upon analysis.

It is very important to prevent the occurrence of confounders within the data set [26]. Confounding means that the effect of a certain disturbance cannot be distinguished from the effect of interest, due to perfect co-occurrence between the two. This happens for example if all patients are children and all controls adults, if all patient samples are labeled with the Cy3 dye and all controls with the Cy5 dye, if all patient samples are run on one day and all controls on another and so on.

In short, correction for disturbing effects can (partially) be done by choice of a proper design. Also in the analysis one can control for several disturbing effects as discussed in the section considering analysis methods.

Study design for dual color arrays

Although experimental study design [20, 26, 46, 47] options for dual color arrays are discussed here, some of the design aspects also apply to single color arrays. In a dual color array (a coupled design), some variations are inherently minimized for by the design by considering relative differences between pairs of samples hybridized to the same slide. As such, proper choice of which samples to co-hybridize is of great importance and leads to several experimental design options [16], several of which are shown in Figure 5.

A first possibility is pairing each sample to another sample and doing pairwise comparisons. The issues mentioned in the previous section implicate though, that a straightforward pairwise comparison is not a good option, and more complicated comparisons have to be carried out. A reference design in which each sample is paired to the same reference material is an alternative option. The reference material may be a basic or null sample (which may be totally unrelated to the samples studied), a pool of all samples, or a pool of all controls. The advantage of such a design is clearly that each sample is compared to the same reference, which makes it an intuitive design, and optimal when the goal of the study is clustering [48]. Another reason to use this design is that it leads to an easy and straightforward analysis [16]. However, a reference design has several drawbacks as well. First of all, variance in the reference is not taken into account as only measurements for one sample or pool are available. Also, such a design can bias results and affect power to detect genes that already were differentially expressed in the references sample or reference pool by chance. In case of a pool this can be caused by even one outlier sample within the pool. Furthermore, such a design does not generally directly make the comparisons one is interested in, leading to a greater variance for the comparisons of interest. A reference design also poses some difficulties on comparing data between researchers, because different reference materials are generally used [17, 47]. On the other hand, within one institute it eases both increasing sample size by adding new samples to the ones already examined and testing new samples, because the same reference material can be re-used to compare to [27].

Figure 5 Example study designs: a) a pairwise comparison; b) a dye flip experiment, where each experiment is repeated with the dyes reversed; c) a reference design, where both patients and controls are compared to a common reference pool; d) a loop design, where each sample is compared to one sample of each of two other groups. Each sample is labeled once in both dyes. Gray indicates control samples, black indicates samples of the disease under investigation, dark gray indicates reference material in c) and samples of a second disease group in d). The sample at the end of an arrow is labeled with Cy5, and the one at the head of an arrow is labeled with Cy3.

A loop design is an efficient alternative. Such a design links all samples to each other indirectly. A sample from group (or level) 1 is paired with one from group 2, one from 2 with one from 3, one from n with one from $n + 1$ and one from the last sample group, N , is paired with one from 1 again. This is especially intuitive for temporal or dilution experiments, where each level is compared to the next one, and the last to the first to close the circle. However, also for group comparison studies a loop design may be one of the most efficient ones to use. It allows indirectly to compare all samples to each other and provides measurements of each sample with both dyes (each sample is measured twice, if properly done once with each dye), but it also requires a specific and more involved analysis to be done [27]. A disadvantage of a loop design is that many comparisons are made indirectly, i.e. between slides [16]. Because there is more variation between slides than within, this leads to more variability in estimated differences than when direct comparisons are used.

Furthermore the path between two samples may become quite long in loop designs, thus they are not to be used (without modifications) with more than about ten sample groups. This in contrast to reference designs, for which the path between all samples is equally long [27]. Also a loop design best contains replicates, because loss of one array in the chain would greatly reduce quality. Efficiency of loop designs can be increased even more when three (or more) color arrays are used, with the same number of samples (and fewer arrays). Thus, if technical performance of those platforms is improved, they will be better than dual color arrays [19]. Overall, the same laboratory and analysis procedure as used for two color arrays will extend to three-color arrays [49], but design choices and analyses get more complex.

Another design with direct comparisons and equal path lengths is the complete (block) design in which each sample is paired with a sample from each other group. This design leads to measurements with the lowest variance, but will cost many slides and material (if more than three sample groups are to be compared). To decrease the number of experiments needed, balanced incomplete block designs can be used, which means that some combinations are not directly made, but the remaining ones are balanced over all groups and over the dyes.

Especially in the case of factorial experiments [16] many designs are possible, and the one chosen needs to be tuned to the research question at hand. It has even been suggested to analyze both channels completely separately in this case, with the use of replication, and no common reference design [17]. In general it can be stated that the comparisons of most interest should be made on the same slide, and for the others indirect balancing designs should be used as much as possible [27]. It is clear that the choice of experimental design strongly influences the way the data can be analyzed and what information can be retrieved from the experiment.

Slide design for custom-made arrays

As discussed above, each probe is spotted at a certain position on a microarray slide. Important aspects of experimental design for custom-made arrays, whether spotted in-house or ordered from an outside company, are the layout of the spots on the slide and the incorporation of controls. It is best to spread out all controls over the slide as much as possible [16, 27]. Additionally, if some probes are spotted multiple times, or in case of oligo arrays several probes are spotted to detect the same transcript, then these must also be spread out as much as possible. It still is advised to estimate correlations of repeated spots [50] because these must always be considered to be related [27].

Furthermore it is of importance to spread repeats over the printing needles of the array spotter. Because slides are spotted with a print head consisting of a certain number (e.g. 16) of needles working in parallel, collections of spots that have been printed with the same needle are present (Figure 6). Because these needles may vary in their physical printing properties, one should ensure that a possibly confounding print-needle effect can be captured by the design.

Figure 6 Layout of a custom home-spotted slide. The exact layout depends on the number of printing needles of the spotter used. Here a spotter with sixteen needles is considered. It may happen that the total number of probes cannot be spotted in one run, for instance in this example case three runs were needed. Within each run of the spotter, the sixteen needles complete one metagrid. The first time the head of the spotter is placed on the slide for run 1, the spots indicated in black are printed. The next time, the gray spots are printed, and so on. The number of spots in a subgrid is determined by the spot size, by the distance between the spots, and by the distance between the needles of the head. It is important to spread out control or repeat spots evenly throughout the slide, metagrids and subgrids.

Commercial arrays

Besides custom-made arrays, commercial predefined arrays can be used. These arrays are pre-designed and contain certain sets of probes, from specific subsets to genome-wide collections. They generally come with a full system of reagents, equipment and basic analysis software. Commercial arrays can be of any type, e.g. single or dual color, cDNA or short or long oligo. Commercial producers of arrays include companies like Affymetrix, Agilent, Amersham, Illumina, and many more. The specificities of the many commercial platforms are out of scope for this text. In Highlight 1 focus will be on the Affymetrix GeneChip expression Array platform as an illustrative example. The advantage of a commercial platform as compared to a custom-made system is that it is quicker in setup, does not need as much testing and tuning and is more standardized. It also facilitates comparison of data between labs. A disadvantage is that it is generally less flexible, although some companies now provide the option of providing the design for arrays they will then custom-build.

Highlight 1: The Affymetrix GeneChip expression Array platform

Affymetrix gene expression chips [51] were one of the first types of commercial platforms to become available in the beginning of the nineties of the previous century and are the overall most frequently used array platform nowadays. The Affymetrix protocol and system is well-standardized and leads to reproducible and reliable results [52, 53]. The complete laboratory and scanning procedure takes on average a few days to run. Many chips are available for several organisms. The chips are oligo-based and contain eleven to twenty oligos (or probes) to cover each transcript and the collection of all oligos detecting the same transcript is called a probeset. In some cases one transcript may also have several probesets, e.g. in the case of control or 'spike' probesets. In general, a spiked-in RNA (commonly called a spike) is a control that is added to the sample at any point during the SOP to check the performance of one or more steps performed during the experiments. On Affymetrix chips, both 'poly-A' spikes (controlling for amplification, labeling and hybridization) and hybridization spikes (only controlling for this) are used as controls. Depending on the type of chip, these spikes have several probesets each, to cover the five prime, the mid and the three prime areas of their sequences. The total number of oligos on one chip depends on the chip type and can range from a few thousand to several tens of thousands. Also, to be able to analyze even larger numbers of transcripts, several chips sets are available, where the probes are divided over more than one chip.

Quality control

Quality control can be divided into several aspects. Here, within and cross-platform consistency, between laboratory consistency and technical and biological validation are discussed.

Comparisons between laboratories and platforms

Variability in microarray results between laboratories is larger than within one and the same laboratory, even if the same platform is used. However, the amount of variability differs between the platforms. Besides between-laboratory comparisons, comparisons between the different platforms and between several specific array-types within one platform also give valuable insight. This showed first of all that implementation of standardized protocols could greatly increase reproducibility between laboratories [54].

Results between published lists of differential genes are often very discordant [55, 56]. Some authors indicate that this can be mainly explained by the fact that differences in analysis methods are not taken into account and that if this is done, concordance gets much better [57]. Between three commercial platforms, results showed considerable differences for the same hybridized RNA. However for each platform an alternative technique could verify results [56]. This may be explained by

the fact that each platform misses other classes of genes, but the ones retrieved are correct [58]. Another study showed similar results between three (different) platforms if only direction of change was considered and magnitude ignored [59]. In another study comparing two commercial platforms, results were considered to be very consistent, especially for genes having higher expression values [60]. Affymetrix chips have also been compared to another gene expression screening technique, Serial Analysis of Gene Expression (SAGE) [61], where the two correlated well both in absolute expression values as in calculated differential expression values [62]. Correlation was found to be best for genes with higher expression levels or greater changes in expression. Large differences between Affymetrix and dual color cDNA arrays could be explained by the difference in the way gene expression is measured [8]. Also differences were found between cDNA arrays and custom-made oligo arrays or Affymetrix GeneChips [63]. However, in other studies good consistence between both platforms was found [52, 64]. In this respect, it is important to compare relative expression differences and not absolute values, to ensure similar pre-processing, and to include several laboratories in the comparison [52].

The most important explanations though, found to be the major cause of discordance by many comparative studies, have been differences, inconsistencies and errors in probe annotation. The problem with annotation for cDNA and oligo clone collections has already been mentioned, but also for commercial platforms the same issues arise. A report showing larger consistency within platforms than between them could track this down to probe annotation quality because on two types of arrays sequences were not always correct, differential expression was variable, probes were not very specific for the intended target, fold-change calculations discrepant, and low specificity of probes for several isoforms was detected [55]. This may also be indicated by the finding that consistency was fairly good between several platforms where probe sequences were similar [65]. Also predictor (classifier) stability has been shown to be mainly below optimality due to probe sequence differences between platforms [66]. A further comparison of cDNA microarrays versus Affymetrix chips [67] showed probe-specific effects to play a role in causing low correlation. Also between generations of the same platform (Affymetrix) reproducibility appeared mainly dependent on probe similarity (and expression level of the transcript) [53, 68, 69].

Validation

Originally, arrays were mainly intended to be used for screening purposes. After drawing general conclusions (e.g. on a pathway level), results for specific genes need to be verified with other techniques. Two types of verification must be distinguished: verification of the measured mRNA expression levels themselves, and verification of the biological significance of changes in expression. The second type of verification is generally done at the protein level, where detection of protein abundance and functional tests or measurements of protein activity, can provide much information. However, these tests are not always possible due to non-availability of specific tests

for the protein at hand. Also the cost of doing such experiments can be prohibitively high, especially when several proteins need to be tested. Furthermore, especially with biopsies, material will frequently only be available to do a very limited number of tests.

The other kind of validation, verifying the change in gene expression itself with a different method, is frequently performed but some debate is going on on the use of this common practice. Mostly, quantitative real-time reverse transcriptase PCR (qPCR, RT-PCR) is used [70]. This technique allows to measure differences in expression between several genes by amplifying products and measuring quantities over time. As a standard to compare to, a reference gene or ‘housekeeping gene’ must be included in the experiment. The expression level of a gene under study is then computed relative to the expression level of the housekeeping gene. This directly indicates a problem: how to find a suitable housekeeping gene? Such a gene must be expressed constantly over all the samples, which means that its expression may not differ in the tissue(s) considered (even if tissue types differ between say biopsies), or between the several groups of samples (e.g. diseased versus controls). It must also be equally expressed for each level of relevant covariates for the experiment, e.g. sex or age. Some authors stress the importance of validation of differential genes with qPCR or other techniques as northern or western blots, or *in situ* hybridization [16, 71], while others think this is of less value because random noise would lead to false negatives instead of false positives [14]. Overall strong correlation was found between microarray (Affymetrix chip) and qRT-PCR results, but over ten percent of genes showed bad correlations [72]. This would demand cautious interpretation of array data and use of complementary techniques. However, for qPCR there is much doubt whether results are better than those from array measurements [14]. It has been suggested that with the use of controls, such as spiked in RNAs, to check if the experiment went well, or correction for those in the data analysis, no such validation is needed [16, 73].

Analysis methods for gene expression array data: general aspects

Analysis of micro array data [8] is not trivial. The data have a typical structure, with thousands to tens of thousands of variables and only tens to hundreds of samples. This poses a real challenge on analysis. Because of the development of arrays, more and more data-driven discovery is initiated instead of hypothesis-driven research [58]. Furthermore, the longer the system is in use, the more challenging the samples selected for study get. In the early phase, focus was mainly on distinguishing samples with clear differences in expression and characterizing the genes or processes involved. In the early days examples included cancer studies, developmental studies, tissue comparisons and studies related to inflammation. Also cell cycle studies were performed, because it was relatively easy to obtain synchronized samples, making expression more uniform and differences easier to detect. However, nowadays arrays are more and more used to detect more subtle differences, for example for comparing samples of patients with neurological,

muscular or many other diseases or syndromes to controls, or even for distinguishing between subtypes of a certain disease or between phenotypically similar patients. Of course, differences in gene expression are expected to be much smaller then, and as such harder to detect. Also the sample material under investigation can be challenging (cf. Highlight 2).

Highlight 2: Challenges posed by the material to be analyzed

The study material may be very challenging. Especially when small biopsies have to be used, very limited amounts of RNA can be extracted. Furthermore biopsies contain different cell types, and this cellular composition may differ between several biopsies in the same study. If this effect is not taken into account, results may be misinterpreted [8]. Microdissection [74] can help, but this is not always feasible or possible [8]. Also sample material is sometimes obtained from a different tissue as where the disease manifests itself most prominently, e.g. a blood or skin sample for a neuromuscular or central nervous system (CNS) disorder. Also the number of samples is often problematic, and the numbers required according to a standard sample size calculation may be impossible to reach. To obtain more samples or more (uniform) material than available from humans, frequently model systems or organisms are used, mostly cultured cell lines, yeast, or closer relatives such as the mouse. This approach is based on the strong homology between genomes and genes. Yeast is of special interest because of its unicellularity (ease of experimentation) and eukaryotic nature at the same time. As such it has been used extensively for genome-wide functional genomics research [75-80] and comparative genomics research [81, 82]. Maps of homology to the human genome have been created. Of course, when differences get more subtle and material less uniform, differences in expression get harder to detect and analysis methods need to be more specifically tuned or developed for these applications.

The most common goals of microarray experiments are: firstly the determination of differentially expressed genes between subsets of samples; secondly finding out which subgroups of genes or samples are present in the data set; and finally construction of gene signatures to classify samples. For all three goals several methods have been developed. Generally, before these are applied, pre-processing of array data is done to ensure proper control for disturbing effects and statistical validity of the further procedures. As such, analysis is generally done in a stepwise way, although combining pre-processing and the rest of the analysis in one step is possible. The purpose of correction for disturbing effects is to distinguish the biological variability between the samples, which is the outcome of interest, from the technical variability introduced by the procedure. Common pre-processing steps include flagging of low quality and/or low intensity spots, background correction, and normalization and filtering. Flagging spots can be done automatically by finding spots

that are too small, too variable in intensity within the spot, contain clear overlay of an aspecific signal, have too low a signal, or have the same signal as the background or even lower. For the last two cases, spots are often labeled 'absent', for the others 'poor'.

Signal intensity correction

After quantitation and flagging, the next step is signal intensity correction, where background signal has to be removed from the specific probe signal. Background signal consists of three contributing effects: overall slide background, local background per slide region or per probe, and specific probe background caused by aspecific binding (cross-hybridization). Correction can be done for each of those. Correction for overall background is done by correcting for the mean background, the mean intensity of empty control spots or the mean intensity of a certain percentage of the lowest intensity spots (assuming these correspond to non-present transcripts). Regional correction can be done by the same methods, but then based on a certain region of the slide. If done for each spot separately, a more specific option is correcting for the (distance-weighted) background intensity of (all spots in) a certain surrounding area. Correction for aspecific binding can be done by incorporating reference spots for each spot, such as the mismatch probes in Affymetrix chips (these are oligos that are the same as the oligo designed to bind the intended transcript, apart from a change in the middle base). This last type of correction is generally done after normalization. Ruling out aspecific binding as much as possible is very important, because it is hard to separate aspecific signal from real signal. An interesting type of array in this respect is the so-called zip-code array [83]. Here the oligos are designed not to bind any mRNA in the sample material. The target molecules in the sample are created by linking the mRNA to a complement sequence for the zip-code. Besides minimizing the risk of aspecific binding, this allows all kinds of assays to be done on the same slide. An extreme development of this technique is the use of the isoform of DNA not present in nature, L-DNA, to create the zips. This rules out any possible aspecific binding and also makes targets extremely resistant to nucleases, which only work on natural R-DNA

Normalization

Normalization is done to smooth out effects of disturbing factors of several kinds (cf. section on disturbing effects). It must be applied with great care because it has profound influence on results [57] and many procedures are available [13]. Total intensity normalization scales slides in such a way that the average (or median) intensity of each of them is equal. This is a simple and effective procedure. However it is not suitable for all studies. The underlying assumption is that the average expression of all transcripts on each slide should be the same. Thus the method cannot be used in cases where this assumption does not hold, such as slides with specific collections of genes spotted on them that are all expected to change, or studies in which vast differences in overall expression can be anticipated

(developmental stages, tissues, some types of cancer cells). Other normalization methods are based on the use of subsets of transcripts that are assumed to be equal in expression in all samples (housekeeping genes [84]), or based on a certain number of spikes/controls (RNA from another organism, like a bacterium, that is spiked into the sample at known concentration at one or more moments during the protocol). Such internal controls are often used, but occasionally not advised to standardize data because they are also subject to noise and introduce correlation [27].

The types of normalization mentioned above are all total slide or global normalizations, i.e. they apply to all measurements at equal effect. More refined methods are available as well. One of these is normalization based on intensity, where the value of each transcript is normalized based on a correction for all other samples, or certain reference controls, in the same range of expression. This is done because the correction needed may not be the same for lowly and highly expressed genes. Logged ratios (cf. section on determination of differentially expressed genes) may deviate from zero for low-intensity spots [40, 85], which can be visualized by plotting the logged ratio versus the mean intensity (Figure 7c). For this reason more specific ways of normalizing data are available, amongst which are LOESS normalization and its weighted variant LOWESS [86]. These methods are of special use and perform a local regression based on the intensity. They estimate for each data point the best-fit average to be subtracted, with in the weighted case more weight on its near points (having similar intensities). Another option is to try to make the distribution of probe expression levels the same for each array in the data set, as is done in the quantile method [87]. Other methods have also been developed to tackle situations in which many genes are differential, but without the use of spikes [88].

One can also perform local normalizations for parts of the slide [13], for each of the pins separately (where LOESS performs well [89]), or for normalizing dye effects (which also has been done using LOESS [40]). For dual color arrays, normalization may best be done per channel. This approach is required if no dye swap or randomization is done [16]. For local normalizations, the assumptions for the technique chosen should still hold, for example that most genes in the local group must be non-differentially expressed. Normalization can also be done in steps, first within the slide, then between slides [58]. Besides the mean, also the variance can be controlled for between slides, subgrids, or any parts of the array [40].

Filtering

Filtering of bad data can have strong influence on the results obtained [57]. Some filtering is already done at the level of flagging spots in the beginning of the procedure. However, certain types of filtering can only be done after all data have been collected. Because of the higher variability of lower expressed transcripts [90], these are sometimes discarded from analysis by imposing a required minimal expression, and only those measurements are considered that differ enough from

background [13]. This can be done by setting a minimum absolute expression level, or by using more advanced techniques, such as discarding a pre-selected percentage of all spots or removing all measurements within two standard deviations from background level. On the other side of the intensity range, saturation may be a problem, and spots saturated in all groups must be discarded, because it cannot be determined anymore if they were different between groups or not. Finally, apart from expression level, genes with minimal variance between all samples are also sometimes discarded [57].

Figure 7 Visualizations of microarray data: a) a heat map, showing gene and sample profiles, after reshuffling both based on the results of a hierarchical clustering, dark (red) indicates low expression, bright (white) indicates high expression; b) a volcano plot, showing differentially expressed genes by plotting the $\log_2(\text{fold change})$ against the $-\log_{10}$ of the p-value of the statistical test used; c) an intensity versus $\log_2(\text{fold change})$ plot showing that fold changes are more variable for lower intensities, the circled dots indicate differentially expressed genes; d) an intensity versus intensity plot for two slides or groups of slides showing which genes have different expression (circled dots) between the two.

Detection and removal of outliers is an analysis step that is also frequently incorporated to get to more reliable results. This can be done by removing all values further than a certain number of standard deviations (mostly 2) away from the mean, or by plotting diagnostic statistics on outlier proneness of each value. Furthermore, slides that are too dissimilar from the other slides in the experiment can be identified and discarded from further analysis, for example if the overall intensity is too different from the other slides. Also (Pearson) correlation has been suggested to select outlying slides, where those with correlations below 0.8 should be discarded [58].

Determination of differentially expressed genes

A common goal of microarray studies is finding which genes are differentially expressed between several groups of samples. The most basic approach is computing direct (mean) fold changes (FC), but this has been considered not to be adequate [14, 91, 92] because also variability must be estimated. For this reason, also the old practice of putting a cut-off on the ratios is not valid, because small changes can in fact be detected [17] and large ones need not be significant. Generally, ratios are log-transformed to normalize data by creating a continuous spectrum of values where up- and downregulated genes are similarly treated [13]. In addition, logging values allows to model multiplicative effects on the original scale by using additive models, which are easier to specify [31]. For slides that contain multiple oligos for the same transcript, mostly a method is first used to compute a common intensity, or transcript expression. Simply taking the mean can do this, but generally more robust methods are used that are less sensitive to outliers (removing the highest and lowest values first, removing all 'outlying' values first, taking the median or mode, or using other robust statistics).

Frequently, a list of (most) differentially expressed genes is created by carrying out a Student t-test for each gene. A problem that arises however is multiple testing, and the question remains how to correct for this. If no correction is used, many false positives occur [71]. A basic solution may be to do family-wise error-rate (FWER) control [31]. A very simple example of this type of correction is the Bonferroni correction, which divides the selected cut-off on the p-values by the number of tests done to ensure that the chosen significance level is reached over all tests together. This correction is conservative because it is based on the assumption of independence between genes, which does not really hold [57, 93]. As a consequence often no genes are retrieved any more. Also more elaborate FWER control methods are available [94]. These include the Westfall and Young step-down adjustment methods which are less conservative and more powerful than the Bonferroni correction [95, 96]. Thus without correction, often a long list containing many false positive genes is retrieved, where it depends on the experience of the biologist which ones are picked out. On the other hand, with a conservative correction often no genes are retrieved [97].

A good alternative to multiple testing with or without FWER-control methods, may be using the false discovery rate (FDR). The FDR controls the number of false positives [98, 99], but allows more false positives than FWER control methods, and has as such more power [31]. For this reason, it has frequently been applied to microarray data [100]. However, using this technique, dependencies between transcripts are a problem as well [14]. The FDR was used originally by controlling it and keep it below a certain level. Later, FDR modeling was done as well, by mixture models that consider genes to belong to one of two groups (differentially or non-differentially expressed) and control the FDR in a powerful way by calculating this statistic only for the latter group. This FDR-estimation was found to be better than either regular FDR-controlling or Bonferroni or other FWER control methods [14]. Another method based on the FDR, called significance analysis of microarrays (SAM) [101], has been developed to overcome difficulties in reaching significant p-values for non-parametric methods (that have less power than parametric tests because they make no assumptions on the underlying distribution). SAM combines results of all genes to reach higher significance. For genes with a test statistic greater than a certain threshold, SAM permutes measurements to estimate the FDR [58]. A disadvantage is that confounding factors such as sex and age cannot be incorporated. SAM is also a good gene-reduction technique for clustering and classification (cf. section on clustering and classification).

Receiver operating characteristics (ROC) curves [102], originally from signal analysis, can be used to estimate the (relative) quality of several expression measures because they can estimate precision and consistency of fold changes (over a range), and sensitivity and specificity of the differentiability calls. These curves visualize both chances on false positives and on false negatives by plotting sensitivity against power ($1 - \text{specificity}$). If the plot contains data points for several measures, the one giving the best cut-off between sensitivity and specificity can be selected.

Another issue is the difficulty estimating the variance for each transcript, because this may be small by chance, yielding highly significant p-values [31]. A common variance for all genes can be computed and added as variability information to the FCs. However, this is of no use because transcript ranking will remain the same as the ordinary FC ranking (because all get weighted by the same variance estimate). Because generally lower expressed values are (relatively) more variable [90], variation can be made intensity-dependant. This reduces the risk of overestimating the number of differentially expressed genes for low intensity. An alternative when working with FCs is using an intensity dependent threshold on the FCs [85], but still no information on the variability is used then. This information is included in a Z score for differential expression [85]. The standard t-test assumes equal variance in all subgroups, which is relaxed in the Welch-modification of the t-test. Methods that estimate the variance based on a combination of transcript-specific and overall variance may improve the quality of analysis, such as the regularized t-test [103] and the B-test that indicates the relative chances of being differentially expressed or not [104]. Variance shrinkage, where variance information from the transcript considered

is combined with (weighted) information from other genes, is especially useful with small sample sizes [101, 105]. Microarray data can be visualized in an intuitive way by a volcano plot (cf. Figure 7b), which combines FC and significance information by plotting logged ratios versus $-\log_{10}$ p-values, to get a quick overview of interesting genes [31]. Another basic visualization tool is a plot of the (mean) intensity of each transcript for one slide (or group of slides) compared to another (cf. Figure 7d), but this does not show information on variability.

Besides Student t-tests, several non-parametric tests, such as the Wilcoxon rank-sum test, have been frequently used to select differentially expressed genes. In contrast to the t-test, these tests do not assume normality of the data, are more flexible, work with any new test statistic, and are more robust to outliers, but are more computationally intensive [106]. This category of tests includes permutation-based testing, and bootstrapping-based (jack-knife, cross-validation) testing. The first type estimates significance by rerunning the same test on random permutations of the data set at hand, the second by rerunning after leaving out one or more samples. A comparison of three non-parametric tests (non parametric t-test, Wilcoxon rank-sum test, heuristic correlation based test) showed that all had low false positive rates and retrieved most differentially expressed genes, where the Wilcoxon test was most conservative [107].

For studies consisting of two or ordered groups, studies that contain a time or concentration component, or studies that have any expected expression profile between the samples, the correlation coefficient of a representative pattern can be used to determine which transcripts are of interest.

Modeling gene expression is a good alternative to doing tests, because it is more informative. When considering more than two groups, analysis of variance (ANOVA) is a good option. Mixed models with both fixed and random effects are considered to be general and powerful for microarray experiments [31]. Fixed effects are those assumed to be measured without error. They can obtain the same values if a further experiment is performed (e.g. sex, age, patient group) and primary interest is in their mean. Random effects are those assumed to represent a random sample of all possible values for that effect in the population. They will have different values each time, and primary interest is in their variability. Sources of variation can be taken into account using these methods, provided enough observations are available. It is important to characterize each source of variation, because statistical significance will otherwise be overestimated. A two stage ANOVA approach can consist of first building a normalization model that accounts for experiment-wide systematic effects that could bias gene expression inferences, and then building gene-specific models [92].

Building models for gene expression, such as linear models [108], has advantages as compared to testing that covariates can be incorporated into the models and that models can be rated with a criterion as the Akaike Information Criterion (AIC) [109] or the Bayesian Information Criterion (BIC) [110] to find out which genes are differentially expressed, avoiding sequences of hypothesis testing [58] and as such

issues with multiple testing. These measures compute a complexity-penalized fit of the model at hand. The AIC sums the minus log-likelihood (LL) with a complexity penalty of the number of parameters in the model. The BIC weights the penalty on the number of parameters with half the natural logarithm of the number of samples. The standard AIC has been criticized to suffer from multiple collinearity and to be biased for complicated models because the log-likelihood increases faster than the complexity [58], but its complexity penalty can be weighted to be stricter [111]. If the AIC or BIC gets better (lower) when an expression difference between groups for the transcript at hand is included in the model, then this transcript is concluded to be differentially expressed.

Also Bayesian approaches may be used, that compute a posterior probability distribution [58]. These approaches are not covered here, an overview can be found in [112].

However, any *per gene* analysis technique may miss important effects on pathways (small effects on all genes, instead of a big effect on one or a few). Gene Set Enrichment Analysis (GSEA) co-evaluates groups of genes according to prior knowledge regarding common function, location, or regulation. Consistence of experimental results between different studies indeed increases as compared to when single gene analysis approaches are used [97].

A final interesting fact to note is that also lack of change can be very important but this has been ignored so far [83]. For example, less than expected numbers or even complete lack of differentially expressed genes within certain pathways could also provide useful information.

Highlight 3: analysis of Affymetrix data sets

Several methods have been developed for Affymetrix GeneChip expression Array analysis. These deal with normalization, background correction, and summarizing probe intensities into probeset (gene) intensities. In the Affymetrix Microarray Suite (MAS) several algorithms have been incorporated. The MAS4 [113] algorithm AvgDiff computes probeset intensity by subtracting the mismatch probe intensity (MM) from the perfect match probe intensity (PM) and then averaging these differences. Because this could result in negative values for non-expressed probes (MM > PM for a third of genes [114]) and a linear algorithm did not perform well enough, this was changed in the MAS5 version [115-117]. In MAS5, an algorithm based on logged values that computes the intensity differently for lowly expressed transcripts is used. The logged difference between the PM and a value based on the MM but never bigger than the PM is used to compute a robust statistic, the median-related Tukey biweight that reciprocally weights data points on the basis of their distance to the median. Also outliers are detected and removed first. After computing their intensity, transcripts get an absent or present call.

An alternative procedure is the analysis method of the DNA-Chip Analyzer (dChip) software [118], which consists of an adaptive pixel selection algorithm (APS), leading to less variable spot intensities, and a normalization procedure based on a set of invariant genes, i.e. genes with the same intensity (PM - MM) ranking as the overall chips.

The robust multi-array average (RMA) method [114], in contrast, does not use the MM values. This method computes a robust multi-array average of background-corrected, intensity-based normalized, and log-transformed PM values. RMA outperforms dChip, which in turn outperforms MAS5 in a study performed by the RMA authors [102], where accuracy of RMA was similar but it obtained better precision than the other methods. Also FCs were more consistently estimated over the entire concentration range and the method was more sensitive and specific to determine differentially expressed genes. In contrast, another comparison between MAS5, RMA and dChip showed as different results between those as between Affymetrix and other platforms [56].

Another approach is the Model-Based Expression Indexes (MBEI, also known as Li & Wong) method [119], based on the fact that since variation between several probes of a probeset is higher than the variation in one probe across arrays, it is important to properly handle this. MBEI works with probe-based analysis models for PM - MM differences that have much smaller residuals. They correct for artifacts and do probe selection, leaving out outlier pairs. The models also give estimates of standard error, reduce variability for lowly expressed transcripts and have been extended to work for PM-only chips with equal performance [120]. A comparative study showed that MBEI was best for finding co-expressed genes, and MAS for finding differentially expressed genes [121]. Another study comparing all methods for Affymetrix Chip data analysis showed a high false negative rate for all, but that MAS performed best for the pre-processing steps (correction for background *et cetera*) [122].

The GLog Average (GLA) method, based on the mean generalized logarithm of the PM probes, is a simple alternative for the more commonly used methods, reducing variability and increasing detection efficiency [123].

Another method that has been recently introduced, is Factor Analysis for Robust Microarray Summarization (FARMS) [124]. The authors have shown that it is more sensitive, specific and faster than MAS5, RMA and MBEI and several other methods on test data. But it is very new, and more experience is needed.

As discussed before, Affymetrix chips contain controls for amplification, labeling and hybridization. In principle, these can be used to correct for between-slide (between-experiment) differences in these technical steps.

With many methods available, the best way to analyze Affymetrix chips is not yet clear. The AffyComp framework has been developed to test methods on standard data sets [125], as has been done to evaluate new methods [123], although still a rather limited number of data sets is available [14].

Clustering and classification

Besides finding which genes are differentially expressed, clustering genes and/or samples, or finding gene profiles to classify samples are two common other goals of microarray studies. Clustering aims at automatic determination of the groups or subtypes present within the samples, or of groups of genes that react similarly and may as such have similar function or be related to a disease studied in a specific way. Classification aims at finding a set of genes, of which the expression patterns can be used to classify a set of samples correctly, in order to be also able to classify future samples. A goal of microarray studies not covered in this text is reconstruction of regulatory genetic networks and proteomic or metabolic pathways. This field requires very specific methods and the types of (sub)networks that can be recovered are generally highly restricted to prevent the procedures of becoming computationally highly intensive.

In order to successfully accomplish clustering or classification (or network reconstruction), generally a method to reduce the list of genes to consider needs to be applied first. Mostly, a set of differentially expressed genes based on the results of the analysis methods discussed above is used. Without pre-selection of genes, the methods will not be able to find interesting clusters or classifiers due to the adverse signal-to-noise ratio. This is caused by the fact that often only a small percentage of all genes is affected by the process or disease under investigation, and it is very hard to detect those directly with clustering or classification techniques between an overload of unrelated genes. On the other hand, due to the vast number of measurements, it is always possible to cluster all samples or classify all samples correctly, because the number of samples is smaller than the number of independent gene expression profiles in the data set, where even unaffected genes vary in expression due to random noise. This leads to the fact that a number of genes equal to the number of samples can then be chosen to classify. Thus unless applied with great care, seemingly correct but overly specific and meaningless results can be obtained. In other words, any set of not confounded covariates of size equal to the number of degrees of freedom will yield a saturated model and hence a perfect fit that however cannot be generalized.

A distinction is made between unsupervised and supervised analysis methods [8, 57]. For clustering generally unsupervised methods are used. This means that the method is not provided with any information on the labeling or grouping of the samples. Results are obtained, only on the basis of the expression profiles of the (pre-selected) set of genes offered. On the other hand, classification is generally performed with supervised methods, where first a training set is offered including labeling information. The method then searches for a classifier that is able to separate the samples in each group from the samples in the other groups. One should always validate these results to prevent overfitting (overspecify to the data set at hand) using a different data set. For pre-selection either supervised methods, such as any of the techniques described in the previous paragraph on retrieving differential genes,

or unsupervised methods, such as those that find the genes to have most variation in expression (differences in intensity) within the data set at hand, can be used. In the optimal case, the (sub)list of genes to cluster or classify with is selected on the basis of another data set than the one the clustering or classification profile is constructed with. However, in practice this is often not feasible. To optimize, starting from a pre-selected list of genes a clustering or profile can be built by using the entire list as input and then removing the ones that are not needed, or by ordering the list first (e.g. from most significantly or strongly changed or most variable to least) and then adding genes one by one, at the same time removing the ones that are not needed anymore. It is not necessarily the case that all most differentially expressed or variable genes are the best to classify with, because expression profiles are not independent and as such several related genes may not bring extra information if one of them is already in the classifier.

Quality of results of analysis can be evaluated in several ways. Of course performance on the data set at hand gives a good first indication, but estimating performance on independent data sets is essential to prevent overfitting and overspecificity. If this is not possible, performance of unsupervised techniques must always be estimated with resampling [35, 126, 127] and supervised techniques must be validated by cross-validation (cf. section on classification techniques) [14]. Also, ROC curves can be used to visualize and select the result - e.g. clustering, classifier, model - that leads to the best cut-off between sensitivity and specificity.

A general exploratory technique, often used in sample clustering or classification [128-135], is principal component analysis (PCA). It is generally, but not exclusively, used in the gene pre-selection step. PCA and its variants, singular value decomposition (SVD) and multidimensional scaling (MS), are non-supervised multivariate data visualization and reduction techniques [136] based on an attempt to capture as much as possible of the variability within the data set, with as few variables as possible (using linear combinations of the original variables) [93]. PCA reduces the dimensionality of the data by transforming to new orthogonal variables, or principal components, ranked by the amount of variation they explain within the data [58]. As such, it is important to first center the data, because otherwise the first component (the one that explains most of the variance in the data set) may represent the vector pointing towards the mean expression [8]. The linear nature of the procedure has been considered its major disadvantage [58]. For clustering the use of only the first components of PCA is not advisable because these do not capture most of the cluster structure [137]. PCA is also applied as part of network reconstruction methods [138, 139] and for finding transcription-factor binding motifs [140].

Clustering techniques

Clustering was one of the first methods applied to gene expression data sets. Several techniques are available and have been used to cluster sets of genes or samples [57] based on similarities in expression profiles. Gene clustering may provide insight in which genes perform similar functions or belong to the same or related biological pathways, and the clustering of samples may give insight into subtypes of disease or other properties of the samples at hand [93]. The first issue to address when clustering is how to measure the (dis)similarity of several samples. Two measures that both perform well, are the Euclidean distance and the Pearson correlation coefficient, of which the first is scale-dependent and the second is not [57, 141]. Euclidean distance may be better for logged ratios, Pearson correlation for single color systems [142]. Though in tumor classification Pearson correlation may be best anyway [57]. When using Euclidian distance, normalization of the data may be performed first, but a disadvantage remains that negatively correlated genes are missed [8]. Correlation assumes normally distributed data and is sensitive to outliers. Moreover it is difficult to detect a regulating effect over a small part of the expression range [8]. For this purpose the Spearman correlation has been suggested because it is more robust to outliers than the Pearson correlation [58].

Two basic subtypes of clustering can be distinguished: hierarchical clustering and partitioning [141]. Both have been used for clustering microarray data. Hierarchical clustering techniques (Figure 8d) include single (nearest neighbor), average and complete (furthest neighbor) linkage, and the centroid, median and Ward methods, that mainly differ in the way they calculate the distance between two clusters. All of these produce trees (dendrograms). They were the first techniques to be used [143] and have been applied frequently thereafter [144-146]. Dendrograms (Figure 8e) are model free and make results easier to interpret (because the clustering process is visualized) [58]. Furthermore, by considering different levels of the tree, one can find the clustering retrieved for several numbers of clusters. In fact a clustering is provided for each number of clusters between one and the number of samples. The length of the connection between two branches in the tree gives information on their similarity [8]. Hierarchical clustering can be done in two ways: by joining the elements one by one (agglomerative) or by splitting them up one by one (divisive). Divisive algorithms have been used less frequently, partly because optimal division requires more computational time [126]. A disadvantage of hierarchical clustering is that local optima are used, and not a global optimal clustering is produced [8]. Also some subtypes, like single linkage and the median method, may suffer from chaining, i.e. linking together clusters via chains of intermediates.

Figure 8 Several clustering techniques: a) k -means clustering, where each point is assigned to the closest cluster center and centers are moved until they are stable; b) Kohonen Self-Organizing Map clustering, which resembles a) but where the cluster centers are linked and move together; c) clique partitioning, where the points to be clustered and their relationships are represented by a graph, upon which fully connected subgraphs or cliques are determined as clusters; d) hierarchical clustering where points are sequentially added to a cluster (or removed from one big cluster) leading to a set of nested clusters; e) dendrogram corresponding to hierarchical clustering; the dotted line indicates how to obtain a clustering in four clusters.

The group of partitioning techniques contains k -means clustering [141] (Figure 8a), and its fuzzy counterpart, fuzzy- k -means clustering. In these algorithms each gene or sample is assigned to the nearest of k randomly initiated cluster centers (centroids). Then the centers are recomputed on the basis of the average of the objects assigned to the cluster they represent. As soon as the cluster centers are stable, the clustering is completed. The fuzzy version is a generalization of the 'normal' (or crisp) algorithm. In fuzzy logic each object may have partial membership in more than one set or not belong to any set at all [147], which holds as well for the cluster memberships in

fuzzy-*k*-means algorithm. Fuzzy logic fits well with human judgment, is general and computationally efficient and accounts for noise in the data set and has also been used to interpret expression patterns [148]. Fuzzy-*k*-means has been used frequently [149-152], sometimes with adaptations [153]. A related technique is the Kohonen self organizing map (SOM) [8, 141, 154-156], a neural network technique that starts with a grid of cluster centers (Figure 8b). At every iteration a cluster center is moved towards a close gene, at decreasing rate with every iteration, but also all neighboring cluster centers are moved somewhat (due to the linked structure) and as such a 'map' of the data space is created. The SOM method carries a risk of overfitting. Also ANOVA [17] has been used. Finally several graph approaches have been applied to microarray data. The clique partitioning algorithm clusters by finding completely connected components in a graph representation of the data at hand (Figure 8c). An advantage of this technique is that it automatically detects the number of clusters present. Relevance networks work by relating objects based on features and only the ones fulfilling a certain criterion are kept. Disadvantage of these techniques is that it is computationally intensive to compute cliques within a graph [8]. However, some efficient methods are available to perform clique partitioning [157].

For most other clustering techniques, a number of clusters expected or desired needs to be computed or selected. If this cannot be done, clustering must be repeated for several numbers of clusters and then the best result selected based on cluster performance [93, 158] according to a certain specified criterion. In general, quality of several clusterings can be evaluated and compared by many criteria, such as overall fit, penalized by number of clusters. Another option is evaluating which clustering created the most internally compact and externally well-separated clusters [141]. If knowledge is present on subtypes of samples, or genes with similar functions, this can of course also be used to validate whether the clustering was successful [141]. Finally, quality of clustering can be estimated by adding noise, either estimated or sampled, to the data and evaluate the stability of the result [127, 159]. Also PCA can be used to compare the quality of clusterings in a visual way by plotting the data according to the first two components and see which clustering offers the best separation [126].

For visualization, the gene expression matrix is often represented in color, where each position in this matrix gets an appropriate background color (Figure 9). For dual color arrays, generally green is used for down- and red for upregulated genes, and color intensity corresponds to magnitude of difference in expression level. Some other color schemes are also in use for people with red-green color blindness [57]. For one color arrays often a color scheme from red, via orange and yellow, to white is used for low to high expression. Such a plot is called a heat map (cf. Figure 7a). These plots are not only a nice visualization of the data, but they can also help in interpreting results. After clustering with respect to genes, samples, or both, these plots show clear patterns after reordering the genes and/or samples based on the clusters they are in.

Unsupervised clustering has the advantage to be easy and hypothesis free, making few assumptions and always producing results. It takes an unbiased approach [57]. On the other hand, no estimate of significance or validity can be given, and such a general approach mostly does not directly answer the questions that the biologist has [14]. Also less is known about the advantages and disadvantages of several methods and their validity [126, 160]. Results are often not reproducible with the sample sizes at hand [35]. Some comparative studies have been done [126, 142, 161, 162]. These indicate that single linkage is not a good technique for microarray data sets, that complete linkage is better than average linkage, that partitioning techniques are better than hierarchical ones, where SOM may be somewhat better than (non-fuzzy) k means. Rerunning non-deterministic algorithms (e.g. those with random initial centroid positions), testing on randomized data and running several methods is highly advisable [141]. Overall there is more need for good evaluation of already existing methods than for the development of new methods [141].

Figure 9 Several samples (of two different patient groups) are each co-hybridized with control material to a dual color array. Scanned relative intensities are collected and log-transformed for each of them and combined in a matrix. This matrix can be visualized by a color scheme, where red indicates up- and green downregulation. The intensity indicates the magnitude of difference in expression. Note that this is just an example created to better illustrate the idea in black and white and does not match the real scheme.

Classification techniques

Besides clustering, a further frequent goal of gene expression studies is to find molecular diagnostic classifiers, sets of genes that can reliably (or more reliably than the current gold-standard) distinguish between diseases or between subtypes of disease. Although cluster analysis is appropriate for class discovery, it is not best suited for class prediction [163]. In contrast to clustering, classification makes use of supervised techniques, as indicated above. Data sets with known sample groupings are needed as training data sets. Classification has been used for all kinds of samples, but especially to separate tumor samples into subtypes. One of the first studies investigating this used a weighted voting and signal to noise ratio technique [164].

Classification techniques that were frequently used include discriminant analysis (DA), support vector machines (SVMs) and hidden Markov models (HMM). Discriminant analysis [36, 165, 166] searches for a linear weighted combination of genes that most effectively separates the sample groups from each other. It has also been used after PCA [167]. A disadvantage is the inherent linear nature of DA [58], just as for PCA. SVMs [168-170] do the opposite as PCA, they try to expand the number of features by adding functions of one or more of the genes as extra expression profiles (variables), for example multiplications between expression patterns of two genes. It may be that these extra variables can classify the data, whereas any linear combination of the contributing genes could not [8]. The goal is to find a separating plane in the data space of increased dimension that maximally separates the data between the classes of samples, attempting to avoid overfitting [8]. HMM are based on the assumption that gene expressions have been generated by underlying Markov chains [171]. In a hidden Markov chain, each state has a certain probability of being transformed into each of a set of next states, given by the transition matrix. HMM-based methods are most suitable for time series. Besides these techniques, also genetic algorithms, which mutate solutions in order to find the best one have been used [172, 173]. Occasionally rule-based expert systems have been used [173], but others question real availability within short time of this type of methods [83]. A simple but effective approach is k -nearest neighbor prediction, which bases classification of a sample on the classes of the k nearest samples [93].

As discussed above, each of the steps taken in a procedure to build a classifier - pre-selection of a list of genes, building and training the profile, and testing the profile - should be done on different data sets (or different subsets of samples of the same data set). Unfortunately, in most cases not many samples are available and this rule gets compromised. In this case, for validation, the following procedure is followed. One or more samples are left out of the data set. After this, a classifier is constructed and optimized based on the remaining samples. Then the classifier is tested on the sample(s) that has/have been left out. This technique is called bootstrapping, jack-knifing, or cross-validation. The procedure is generally repeated for each of the samples, where in turn each one is left out (leave-one-out) or, alternatively,

subgroups are left out (leave- k -out) [57, 163]. Of course, this is only a less good alternative if no independent test set is available, and it still risks overestimating extrapolation quality of the classifier. This risk is especially present, since for proper cross-validation it would be of major importance to repeat the entire procedure with the selected sample(s) left out, which is generally not done. In contrast, mostly the pre-selection of differential genes is done on all samples and then the classifier is tested with cross-validation on subsets of these samples. This is not entirely correct and influences results [163] because it is not independent [57] and selection bias will result [174].

Because of all issues mentioned, classification for microarray data sets should be done with care, to prevent overspecific and unstable classifiers. In a reanalysis of seven large cancer classification studies, the genes in the prognostic profile were highly unstable and dependent on the training sets used. Also, most studies did not perform better than expected by chance [175]. Furthermore, classical markers were found to be equally powerful for breast cancer prognosis as molecular portraits [176]. Most studies include few patients and it is unclear how results can be applied to more and less uniform patient groups in several settings and with analyses done at several institutes [57]. Furthermore, the biological interpretation of the classifying lists of genes is frequently difficult. However, this would be a problem for elucidating the mechanism of the disease, but not for its classification. If a profile reliably works, this is enough [57]. A more complete discussion of the use of molecular portraits of cancer, is given in [177]. Some of the first studies were on suppression of melanoma [178], leukemia [164], and breast cancer [179]. The most well known study is for breast cancer prognosis, where a seventy gene classifier was developed to predict which treatment can best be offered to patients suffering from this type of cancer [180]. By the same group, the classifier was further evaluated and found to be better than standard techniques [181]. However others criticize the methods used or the predictive value of the classifier or could not confirm it [163, 175, 176]. Optimization and standardization of the technique as well as clinical trials are required before microarrays can be used as diagnostic tools [182]. As such, the seventy gene profile will be evaluated now in a nation-wide clinical trial in the Netherlands [57]. This is very important in light of the contrast between the issues discussed with respect to generalization and the certainty required to use new technology for diagnostics.

Gene annotation: tools and databases

After a list of differential genes or a clustering of genes is obtained, the next step to be taken is the interpretation of the results collected. The procedure to follow is composed of selecting proper identifiers (IDs) for the genes, mapping the genes to biological pathways or otherwise related groups, determining which pathways or groups are important within the current study, further annotating genes in these pathways to learn as much as possible on their functions, and finally validating the results by further experimentation.

Highlight 4: Packages and tools for analysis

Several general statistical analysis packages are commercially available, including SAS, Stata, S-Plus and SPSS. Also many specific commercial systems and packages are available, e.g. Spotfire or the Biodiscovery package (including Imagine to locate and flag spots, and Genesight to analyze data). A package of special interest is the R package for statistical computing [183]. The R software is a freely available clone of the S-Plus package. As open-source initiative it is also a place to disseminate new approaches. Many libraries specific for bioinformatics, and as such for microarray analysis, have become available through the Bioconductor project [184, 185]. Also a specific package for Affymetrix chip analysis is available, called the Affy library [186]. Several statistical libraries for R are very useful to be used in microarray analysis as well, amongst which libraries for general and flexible model fitting (glm, growth, rmutl, repeated) [187, 188]

For a start, the IDs of the transcripts or probes on the array are not always correct and specific, as indicated in the section on quality control. Because this is also the case for Affymetrix GeneChips and moreover the annotation files get outdated over time, several studies have been conducted investigating probe specificity and annotation for these types of chips and ways to improve this. To assist in proper analysis, genes containing too low quality probes could be excluded from analysis [189], or incorrect or aspecific probes removed from the probesets they belong to [190]. A further step would be to totally rebuild probesets, because incorrect probes for one transcript may be correct for another, where probes may regroup into new probesets [190, 191]. In a large study in which all Affymetrix probes for many types of chips were verified against several databases, new annotation files or chip description files (CDFs) were created that are in accordance with the current builds of the public genome databases [191]. These updated CDF files are rebuild every few months and made publicly available. Depending on the database chosen, these CDFs are tuned to a specific research focus, being more or less restrictive in the gene considered and being more or less specific for transcript variants. Probes may not be included in any probeset anymore, and probesets do not necessarily contain an equal number of probes.

Even when it is correctly known which probe(set) corresponds to which gene, the identifier (ID) used may still cause problems for finding further information on the functions of the genes. Many IDs are used throughout public databases, and several array systems and clone sets use their own. Thus, it is generally required to map ID systems to each other, and several tools have become available to do so [192]. Several databases are available to find additional information on the genes retrieved. In public repositories at NCBI [193], UCSC [194] and Ensembl [195] all kind of public information is present on gene sequence, function, protein function, genomic conservation, relationships to diseases, literature, synopses and much more. Gene

Ontology (GO) has been another development of major importance with respect to gene annotation [196, 197]. It contains annotations on three 'axes' (annotations of three kinds) for each transcript: biological process, subcellular localization, and biochemical function. The ontology is organized as a tree-structure, enabling searches to be done at several levels within the tree, from very global to very specific. The tree is predefined, imposing certain annotations to be chosen from when entering new information. A gene can have multiple annotations on each axis, and the same annotation can occur at several branches in the tree, because it can be considered to be an instance of several more global items. Another very relevant database, containing pathway information, is KeGG (Kyoto Encyclopedia of Genes and Genomes) [198]. Furthermore, domain information can give insight into protein function, where many tools and databases are available, e.g. Pfam [199]. Comparison with other species can give information with respect to gene functioning, and conservation of a gene or one of its domains over many species increases confidence in the relevance of a change within it. Many species specific databases are available, for example for yeast in the *Saccharomyces* Genome Database [200] and the Yeast Proteome Database [201]. Of course many more databases and tools are available, of which overviews are present on the Internet (e.g. [202]) and a yearly update on all databases available is published by the journal Nucleic Acids Research [203].

For automatically annotating sets of genes, e.g. a list of differentially expressed genes between sample groups or a cluster of genes, several tools are available. These tools generally can do more than just annotation: they also contain algorithms to decide which specific annotations (biological functions and so on) are overrepresented in the list of genes given as input as compared to a reference set chosen. Generally they provide p-values indicating the significance of overrepresentation of each annotation. Both web tools such as DAVID / EASE [192, 204, 205] and standalone tools such as genMAPP / MAPPfinder [206, 207] and Gostat [208] are available. Most tools mainly work with GO annotations, but with DAVID also many others can be selected. A specific feature of genMAPP is that it also presents graphical schemes of each pathway, coloring the differential genes according to a user-definable coloring scheme based on, for example, FCs. To be able to do so, this tool - in contrast to the others - can be provided with more input than only the gene IDs. All these tools annotate genes after gene expression analysis has been done and it has been determined which genes are differentially expressed. In contrast, Gene Set Enrichment Analysis [97] directly considers sets of genes together upon the determination of differential genes, as discussed above. An issue with several algorithms that determine annotation overrepresentation is present in the difficulty how to determine which classes are significant. To determine this, genes (which are not independent) are often permuted instead of samples and interdependence between results is introduced (if one pathway gets more differential, another may get less without any changes in the last one) [14].

Finally, to assist in completing annotations even further, tools are available to automatically search literature. Amongst these are Pubmatrix [209] and TXTGate [210]. These tools can link genes based on the number of times they are referred to together, or can find all papers citing a certain gene or biological term. These tools return many non-relevant papers (in which it is mentioned that a gene is not involved in a certain process, or where a word that was coincidentally also a gene name was used, and so on), thus they require a lot of manual intervention, but they can certainly assist in finding relevant papers.

Towards a common practice? Guidelines

The MGED (Microarray Gene Expression Data) society has as goal the development of common protocols [211]. Criteria have been setup for microarray studies, data and their annotations to comply to the MIAME (Minimum Information About a Microarray Experiment) recommendations [212] and publications have appeared that give general guidelines for array studies [213, 214]. Public availability of microarray data will be very useful [215]. Within this perspective, ArrayExpress [216, 217], Gene Expression Omnibus (GEO) [218] and Centre for Information Biology Gene Expression Database (CIBEX) [219] have been suggested as repositories for obligatory submission of array data [220]. Even a language for microarray description has been developed, the MicroArray Gene Expression Markup Language (MAGE-ML) [221].

At the moment many methods are available for microarray data analysis, and the need for good quality assessments of those and comparison among them is higher than the need for even more new methods [14]. Moreover, meta-analyses may provide more confidence. Finally, re-analyzing existing data sets every few years is important because methods are expected to develop quickly [8] and results to improve. All of these objectives and the testing of new methods will be facilitated by making microarray data publicly available and using common standards for saving and describing them.

Aims and contents of this thesis

Since the development of the first microarray platform, many methods have been developed to analyze the data sets obtained and to interpret results. However, many issues and questions still remain to be answered. One of the important aspects that still needs improvement is the reproducibility of the results obtained within the area of gene profiling. Current gene profiles have appeared to be very unstable, especially due to overfitting and overspecificity. Also relatively small data sets, containing mainly alike samples, have been used for building these profiles. Gene profiling is not only a very important tool for research, but will most likely become a major technique in near future for diagnosis and prognosis as well, especially to determine the best treatment to offer to each patient. This stresses the importance of reliable and stable molecular classifiers that are treatment center independent and are applicable in different sample handling protocols. Therefore, the first aim of this thesis was to

develop a method that uses more concise gene profiles, corrects for overfitting and prevents the models from using too many degrees of freedom (Chapter 2).

Another very important aspect, both as a direct goal and as pre-selection step for clustering and classification, is reliable determination of differentially expressed genes between groups of samples. Several issues related to this topic were studied. One of the aims was to develop a modeling approach explicitly taking experimental and slide design into account by correcting for covariates directly in the models built to extract differentially expressed genes (Chapters 3 to 6). Another fact in gene expression analysis is that most algorithms are performed in a stepwise way, where several pre-processing steps precede the application of the technique to determine which genes are differentially expressed. The combination of these steps into one prevents possible loss of information between them and may improve results. This was another aim of this thesis (Chapters 3 and 5). There is mainly need for validation of and comparison between the many expression analysis methods available. One aspect of validation considers the different assumptions generally made. Thus, using the modeling approach introduced, another aim was to study the influence of the assumptions of normality of data and homoscedasticity (equal variance within all sample groups) (Chapter 4). A final aim was to further study the aspect of quality control for microarray analysis. Part of this is the question whether and how spike information should be incorporated in the analysis. This was investigated by computing statistics on the usage of spikes to correct for between-array variations in gene expression models (Chapter 6).

References

- 1 International Human Genome Sequencing Consortium: **Initial sequencing and analysis of the human genome**. *Nature* 2001, **409**:860-921.
- 2 Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, et al: **The sequence of the human genome**. *Science* 2001, **291**:1304-1351.
- 3 Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, et al: **Initial sequencing and comparative analysis of the mouse genome**. *Nature* 2002, **420**:520-562.
- 4 Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG: **Life with 6000 genes**. *Science* 1996, **274**:546, 563-547.
- 5 Schena M, Shalon D, Davis RW, Brown PO: **Quantitative monitoring of gene expression patterns with a complementary DNA microarray**. *Science* 1995, **270**:467-470.
- 6 Cheung VG, Morley M, Aguilar F, Massimi A, Kucherlapati R, Childs G: **Making and reading microarrays**. *Nat Genet* 1999, **21**:15-19.
- 7 Lipshutz RJ, Morris D, Chee M, Hubbell E, Kozal MJ, Shah N, Shen N, Yang R, Fodor SP: **Using oligonucleotide probe arrays to access genetic diversity**. *Biotechniques* 1995, **19**:442-447.
- 8 Butte A: **The use and analysis of microarray data**. *Nat Rev Drug Discov* 2002, **1**:951-960.
- 9 Gambaro G, Anglani F, D'Angelo A: **Association studies of genetic polymorphisms and complex disease**. *Lancet* 2000, **355**:308-311.

- 10 Glazier AM, Nadeau JH, Aitman TJ: **Finding genes that underlie complex traits.** *Science* 2002, **298**:2345-2349.
- 11 Moore GE: *Electronics Magazine* 1965, **April 19, 1965**.
- 12 Hegde P, Qi R, Abernathy K, Gay C, Dharap S, Gaspard R, Hughes JE, Snesrud E, Lee N, Quackenbush J: **A concise guide to cDNA microarray analysis.** *Biotechniques* 2000, **29**:548-550, 552-544, 556 passim.
- 13 Quackenbush J: **Microarray data normalization and transformation.** *Nat Genet* 2002, **32 Suppl**:496-501.
- 14 Allison DB, Cui X, Page GP, Sabripour M: **Microarray data analysis: from disarray to consolidation and consensus.** *Nat Rev Genet* 2006, **7**:55-65.
- 15 Halgren RG, Fielden MR, Fong CJ, Zacharewski TR: **Assessment of clone identity and sequence fidelity for 1189 IMAGE cDNA clones.** *Nucleic Acids Res* 2001, **29**:582-588.
- 16 Yang YH, Speed T: **Design issues for cDNA microarray experiments.** *Nat Rev Genet* 2002, **3**:579-588.
- 17 Jin W, Riley RM, Wolfinger RD, White KP, Passador-Gurgel G, Gibson G: **The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*.** *Nat Genet* 2001, **29**:389-395.
- 18 Fare TL, Coffey EM, Dai H, He YD, Kessler DA, Kilian KA, Koch JE, LeProust E, Marton MJ, Meyer MR, Stoughton RB, Tokiwa GY, Wang Y: **Effects of atmospheric ozone on microarray data quality.** *Anal Chem* 2003, **75**:4672-4675.
- 19 Woo Y, Krueger W, Kaur A, Churchill G: **Experimental design for three-color and four-color gene expression microarrays.** *Bioinformatics* 2005, **21 Suppl 1**:i459-467.
- 20 Kerr MK, Churchill GA: **Experimental design for gene expression microarrays.** *Biostatistics* 2001, **2**:183-201.
- 21 Pan W, Lin J, Le CT: **How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach.** *Genome Biol* 2002, **3**:research0022.
- 22 Zien A, Fluck J, Zimmer R, Lengauer T: **Microarrays: how many do you need?** *J Comput Biol* 2003, **10**:653-667.
- 23 Pawitan Y, Michiels S, Koscielny S, Gusnanto A, Ploner A: **False discovery rate, sensitivity and sample size for microarray studies.** *Bioinformatics* 2005, **21**:3017-3024.
- 24 Müller P, Parmigiani G, Robert C, Rousseau J: **Optimal sample size for multiple testing: The case of gene expression microarrays.** *J Amer Stat Assoc* 2004, **99**:990-1001.
- 25 Tsai CA, Wang SJ, Chen DT, Chen JJ: **Sample size for gene expression microarray experiments.** *Bioinformatics* 2005, **21**:1502-1508.
- 26 Kerr MK: **Design considerations for efficient and effective microarray studies.** *Biometrics* 2003, **59**:822-828.
- 27 Churchill GA: **Fundamentals of experimental design for cDNA microarrays.** *Nat Genet* 2002, **32 Suppl**:490-495.
- 28 Lee ML, Whitmore GA: **Power and sample size for DNA microarray studies.** *Stat Med* 2002, **21**:3543-3570.
- 29 Lee ML, Kuo FC, Whitmore GA, Sklar J: **Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations.** *Proc Natl Acad Sci U S A* 2000, **97**:9834-9839.
- 30 Pavlidis P, Li Q, Noble WS: **The effect of replication on gene expression microarray experiments.** *Bioinformatics* 2003, **19**:1620-1627.
- 31 Cui X, Churchill GA: **Statistical tests for differential expression in cDNA microarray experiments.** *Genome Biol* 2003, **4**:210.
- 32 Yeung KY, Medvedovic M, Bumgarner RE: **From co-expression to co-regulation: how many microarray experiments do we need?** *Genome Biol* 2004, **5**:R48.
- 33 Wei C, Li J, Bumgarner RE: **Sample size for detecting differentially expressed genes in microarray experiments.** *BMC Genomics* 2004, **5**:87.

- 34 Dobbin K, Simon R: **Sample size determination in microarray experiments for class comparison and prognostic classification.** *Biostatistics* 2005, **6**:27-38.
- 35 Garge NR, Page GP, Sprague AP, Gorman BS, Allison DB: **Reproducible clusters from microarray research: whither?** *BMC Bioinformatics* 2005, **6** Suppl 2:S10.
- 36 Hwang D, Schmitt WA, Stephanopoulos G, Stephanopoulos G: **Determination of minimum sample size and discriminatory expression patterns in microarray data.** *Bioinformatics* 2002, **18**:1184-1193.
- 37 Page GP, Edwards JW, Gadbury GL, Yeliseti P, Wang J, Trivedi P, Allison DB: **The PowerAtlas: a power and sample size atlas for microarray experimental design and research.** *BMC Bioinformatics* 2006, **7**:84.
- 38 Black MA, Doerge RW: **Calculation of the minimum number of replicate spots required for detection of significant gene expression fold change in microarray experiments.** *Bioinformatics* 2002, **18**:1609-1616.
- 39 Tseng GC, Oh MK, Rohlin L, Liao JC, Wong WH: **Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects.** *Nucleic Acids Res* 2001, **29**:2549-2557.
- 40 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* 2002, **30**:e15.
- 41 Kendzierski C, Irizarry RA, Chen KS, Haag JD, Gould MN: **On the utility of pooling biological samples in microarray experiments.** *Proc Natl Acad Sci U S A* 2005, **102**:4252-4257.
- 42 Kendzierski CM, Zhang Y, Lan H, Attie AD: **The efficiency of pooling mRNA in microarray experiments.** *Biostatistics* 2003, **4**:465-477.
- 43 Peng X, Wood CL, Blalock EM, Chen KC, Landfield PW, Stromberg AJ: **Statistical implications of pooling RNA samples for microarray experiments.** *BMC Bioinformatics* 2003, **4**:26.
- 44 Shih JH, Michalowska AM, Dobbin K, Ye Y, Qiu TH, Green JE: **Effects of pooling mRNA in microarray class comparisons.** *Bioinformatics* 2004, **20**:3318-3325.
- 45 Zhang SD, Gant TW: **Effect of pooling samples on the efficiency of comparative studies using microarrays.** *Bioinformatics* 2005, **21**:4378-4383.
- 46 Yang MC, Yang JJ, McIndoe RA, She JX: **Microarray experimental design: power and sample size considerations.** *Physiol Genomics* 2003, **16**:24-28.
- 47 Simon R, Radmacher MD, Dobbin K: **Design of studies using DNA microarrays.** *Genet Epidemiol* 2002, **23**:21-36.
- 48 Dobbin K, Simon R: **Comparison of microarray designs for class comparison and class discovery.** *Bioinformatics* 2002, **18**:1438-1445.
- 49 Forster T, Costa Y, Roy D, Cooke HJ, Maratou K: **Triple-target microarray experiments: a novel experimental strategy.** *BMC Genomics* 2004, **5**:13.
- 50 Smyth GK, Michaud J, Scott HS: **Use of within-array replicate spots for assessing differential expression in microarray experiments.** *Bioinformatics* 2005, **21**:2067-2075.
- 51 Lipshutz RJ, Fodor SP, Gingeras TR, Lockhart DJ: **High density synthetic oligonucleotide arrays.** *Nat Genet* 1999, **21**:20-24.
- 52 Irizarry RA, Warren D, Spencer F, Kim IF, Biswal S, Frank BC, Gabrielson E, Garcia JG, Geoghegan J, Germino G, et al: **Multiple-laboratory comparison of microarray platforms.** *Nat Methods* 2005, **2**:345-350.
- 53 Nimgaonkar A, Sanoudou D, Butte AJ, Haslett JN, Kunkel LM, Beggs AH, Kohane IS: **Reproducibility of gene expression across generations of Affymetrix microarrays.** *BMC Bioinformatics* 2003, **4**:27.
- 54 Bammler T, Beyer RP, Bhattacharya S, Boorman GA, Boyles A, Bradford BU, Bumgarner RE, Bushel PR, Chaturvedi K, Choi D, et al: **Standardizing global gene expression analysis between laboratories and across platforms.** *Nat Methods* 2005, **2**:351-356.
- 55 Kothapalli R, Yoder SJ, Mane S, Loughran TP, Jr.: **Microarray results: how accurate are they?** *BMC Bioinformatics* 2002, **3**:22.

- 56 Tan PK, Downey TJ, Spitznagel EL, Jr., Xu P, Fu D, Dimitrov DS, Lempicki RA, Raaka BM, Cam MC: **Evaluation of gene expression measurements from commercial microarray platforms.** *Nucleic Acids Res* 2003, **31**:5676-5684.
- 57 Quackenbush J: **Microarray analysis and tumor classification.** *N Engl J Med* 2006, **354**:2463-2472.
- 58 Liang Y, Kelemen A: **Associating phenotypes with molecular events: recent statistical advances and challenges underpinning microarray experiments.** *Funct Integr Genomics* 2006, **6**:1-13.
- 59 Petersen D, Chandramouli GV, Geoghegan J, Hilburn J, Paarlberg J, Kim CH, Munroe D, Gangi L, Han J, Puri R, Staudt L, Weinstein J, Barrett JC, Green J, Kawasaki ES: **Three microarray platforms: an analysis of their concordance in profiling gene expression.** *BMC Genomics* 2005, **6**:63.
- 60 Barnes M, Freudenberg J, Thompson S, Aronow B, Pavlidis P: **Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms.** *Nucleic Acids Res* 2005, **33**:5914-5923.
- 61 Velculescu VE, Zhang L, Vogelstein B, Kinzler KW: **Serial analysis of gene expression.** *Science* 1995, **270**:484-487.
- 62 Ishii M, Hashimoto S, Tsutsumi S, Wada Y, Matsushima K, Kodama T, Aburatani H: **Direct comparison of GeneChip and SAGE on the quantitative accuracy in transcript profiling analysis.** *Genomics* 2000, **68**:136-143.
- 63 Woo Y, Affourtit J, Daigle S, Viale A, Johnson K, Naggert J, Churchill G: **A comparison of cDNA, oligonucleotide, and Affymetrix GeneChip gene expression microarray platforms.** *J Biomol Tech* 2004, **15**:276-284.
- 64 Larkin JE, Frank BC, Gavras H, Sultana R, Quackenbush J: **Independence and reproducibility across microarray platforms.** *Nat Methods* 2005, **2**:337-344.
- 65 Barczak A, Rodriguez MW, Hanspers K, Koth LL, Tai YC, Bolstad BM, Speed TP, Erle DJ: **Spotted long oligonucleotide arrays for human gene expression analysis.** *Genome Res* 2003, **13**:1775-1785.
- 66 Stec J, Wang J, Coombes K, Ayers M, Hoersch S, Gold DL, Ross JS, Hess KR, Tirrell S, Linette G, Hortobagyi GN, Fraser Symmans W, Pusztai L: **Comparison of the predictive accuracy of DNA array-based multigene classifiers across cDNA arrays and Affymetrix GeneChips.** *J Mol Diagn* 2005, **7**:357-367.
- 67 Kuo WP, Jenssen TK, Butte AJ, Ohno-Machado L, Kohane IS: **Analysis of matched mRNA measurements from two different microarray technologies.** *Bioinformatics* 2002, **18**:405-412.
- 68 Hwang KB, Kong SW, Greenberg SA, Park PJ: **Combining gene expression data from different generations of oligonucleotide arrays.** *BMC Bioinformatics* 2004, **5**:159.
- 69 Elo LL, Lahti L, Skottman H, Kylaniemi M, Lahesmaa R, Aittokallio T: **Integrating probe-level expression changes across generations of Affymetrix arrays.** *Nucleic Acids Res* 2005, **33**:e193.
- 70 Rajeevan MS, Ranamukhaarachchi DG, Vernon SD, Unger ER: **Use of real-time quantitative PCR to validate the results of cDNA array and differential display PCR technologies.** *Methods* 2001, **25**:443-451.
- 71 Fathallah-Shaykh HM: **Microarrays: applications and pitfalls.** *Arch Neurol* 2005, **62**:1669-1672.
- 72 Dallas PB, Gottardo NG, Firth MJ, Beesley AH, Hoffmann K, Terry PA, Freitas JR, Boag JM, Cummings AJ, Kees UR: **Gene expression levels assessed by oligonucleotide microarray analysis and quantitative real-time RT-PCR -- how well do they correlate?** *BMC Genomics* 2005, **6**:59.
- 73 van Bakel H, Holstege FC: **In control: systematic assessment of microarray performance.** *EMBO Rep* 2004, **5**:964-969.
- 74 Emmert-Buck MR, Bonner RF, Smith PD, Chuaqui RF, Zhuang Z, Goldstein SR, Weiss RA, Liotta LA: **Laser capture microdissection.** *Science* 1996, **274**:998-1001.
- 75 Grunewald B, Winzler EA: **Treasures and traps in genome-wide data sets: case examples from yeast.** *Nat Rev Genet* 2002, **3**:653-661.

- 76 Wu LF, Hughes TR, Davierwala AP, Robinson MD, Stoughton R, Altschuler SJ: **Large-scale prediction of *Saccharomyces cerevisiae* gene function using overlapping transcriptional clusters.** *Nat Genet* 2002, **31**:255-265.
- 77 Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, Dow S, Lucau-Danila A, Anderson K, Andre B, et al: **Functional profiling of the *Saccharomyces cerevisiae* genome.** *Nature* 2002, **418**:387-391.
- 78 Jelinsky SA, Samson LD: **Global response of *Saccharomyces cerevisiae* to an alkylating agent.** *Proc Natl Acad Sci U S A* 1999, **96**:1486-1491.
- 79 Ge H, Liu Z, Church GM, Vidal M: **Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*.** *Nat Genet* 2001, **29**:482-486.
- 80 Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, et al: **Functional discovery via a compendium of expression profiles.** *Cell* 2000, **102**:109-126.
- 81 Andrade MA, Sander C, Valencia A: **Updated catalogue of homologues to human disease-related proteins in the yeast genome.** *FEBS Lett* 1998, **426**:7-16.
- 82 Steinmetz LM, Scharfe C, Deutschbauer AM, Mokranjac D, Herman ZS, Jones T, Chu AM, Giaever G, Prokisch H, Oefner PJ, Davis RW: **Systematic screen for human disease genes in yeast.** *Nat Genet* 2002, **31**:400-404.
- 83 Hoheisel JD: **Microarray technology: beyond transcript profiling and genotype analysis.** *Nat Rev Genet* 2006, **7**:200-210.
- 84 Eickhoff B, Korn B, Schick M, Poustka A, van der Bosch J: **Normalization of array hybridization experiments in differential gene expression analysis.** *Nucleic Acids Res* 1999, **27**:e33.
- 85 Yang IV, Chen E, Hasseman JP, Liang W, Frank BC, Wang S, Sharov V, Saeed AI, White J, Li J, Lee NH, Yeatman TJ, Quackenbush J: **Within the fold: assessing differential expression measures and reproducibility in microarray assays.** *Genome Biol* 2002, **3**:research0062.
- 86 Cleveland WS: **Robust locally weighted regression and smoothing scatterplots.** *J Amer Stat Assoc* 1979, **74**:829-836.
- 87 Bolstad BM, Irizarry RA, Astrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.** *Bioinformatics* 2003, **19**:185-193.
- 88 Chua SW, Vijayakumar P, Nissom PM, Yam CY, Wong VV, Yang H: **A novel normalization method for effective removal of systematic variation in microarray data.** *Nucleic Acids Res* 2006, **34**:e38.
- 89 Smyth GK, Speed T: **Normalization of cDNA microarray data.** *Methods* 2003, **31**:265-273.
- 90 Rocke DM, Durbin B: **A model for measurement error for gene expression arrays.** *J Comput Biol* 2001, **8**:557-569.
- 91 Budhraja V, Spitznagel E, Schaiff WT, Sadovsky Y: **Incorporation of gene-specific variability improves expression analysis using high-density DNA microarrays.** *BMC Biol* 2003, **1**:1.
- 92 Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, Afshari C, Paules RS: **Assessing gene significance from cDNA microarray expression data via mixed models.** *J Comput Biol* 2001, **8**:625-637.
- 93 Slonim DK: **From patterns to pathways: gene expression data analysis comes of age.** *Nat Genet* 2002, **32 Suppl**:502-508.
- 94 Dudoit S, Yang YH, Callow MJ, Speed TP: **Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments.** *Stat Sin* 2002, **12**:111-139.
- 95 Westfall PH, Young SS: *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment.* New York: Wiley; 1993.
- 96 Jung SH, Bang H, Young S: **Sample size calculation for multiple testing in microarray data analysis.** *Biostatistics* 2005, **6**:157-169.

- 97 Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proc Natl Acad Sci U S A* 2005, **102**:15545-15550.
- 98 Benjamini Y, Hochberg Y: **Controlling the false discovery rate - a practical and powerful approach to multiple testing.** *JR Stat Soc Ser B* 1995, **57**:289-300.
- 99 Li SS, Bigler J, Lampe JW, Potter JD, Feng Z: **FDR-controlling testing procedures and sample size determination for microarrays.** *Stat Med* 2005, **24**:2267-2280.
- 100 Tsai CA, Hsueh HM, Chen JJ: **Estimation of false discovery rates in multiple testing: application to gene microarray data.** *Biometrics* 2003, **59**:1071-1081.
- 101 Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response.** *Proc Natl Acad Sci U S A* 2001, **98**:5116-5121.
- 102 Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP: **Summaries of Affymetrix GeneChip probe level data.** *Nucleic Acids Res* 2003, **31**:e15.
- 103 Baldi P, Long AD: **A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes.** *Bioinformatics* 2001, **17**:509-519.
- 104 Lönnstedt I, Speed T: **Replicated microarray data.** *Stat Sin* 2002, **12**:31-46.
- 105 Cui X, Hwang JT, Qiu J, Blades NJ, Churchill GA: **Improved statistical tests for differential gene expression by shrinking variance components estimates.** *Biostatistics* 2005, **6**:59-75.
- 106 Xu R, Li X: **A comparison of parametric versus permutation methods with applications to general and temporal microarray gene expression data.** *Bioinformatics* 2003, **19**:1284-1289.
- 107 Troyanskaya OG, Garber ME, Brown PO, Botstein D, Altman RB: **Nonparametric methods for identifying differentially expressed genes in microarray data.** *Bioinformatics* 2002, **18**:1454-1461.
- 108 Lindsey JK: *Introductory Statistics: A modelling approach.* New York: Oxford University Press; 1995.
- 109 Akaike H: **Information theory and an extension of the maximum likelihood principle.** In *Second International Symposium on Inference Theory; Budapest: Akadémiai Kiadó.* Edited by Petrov BN, Csàki F. 1973: 267-281.
- 110 Schwarz G: **Estimating the Dimension of a Model.** *Annals of Statistics* 1978, **6**:461-464.
- 111 Atkinson AC: **A note on the generalized information criterion for choice of a model.** *Biometrika* 1980, **67**:413-418.
- 112 Yang D, Zakharkin SO, Page GP, Brand JP, Edwards JW, Bartolucci AA, Allison DB: **Applications of Bayesian statistical methods in microarray data analysis.** *Am J Pharmacogenomics* 2004, **4**:53-62.
- 113 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* 1996, **14**:1675-1680.
- 114 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* 2003, **4**:249-264.
- 115 Affymetrix: *Affymetrix Microarray Suite User Guide.* 5 edition. Santa Clara, CA: Affymetrix; 2002.
- 116 Hubbell E, Liu WM, Mei R: **Robust estimators for expression analysis.** *Bioinformatics* 2002, **18**:1585-1592.
- 117 Liu WM, Mei R, Di X, Ryder TB, Hubbell E, Dee S, Webster TA, Harrington CA, Ho MH, Baid J, Smeekens SP: **Analysis of high density expression microarrays with signed-rank call algorithms.** *Bioinformatics* 2002, **18**:1593-1599.
- 118 Schadt EE, Li C, Ellis B, Wong WH: **Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data.** *J Cell Biochem Suppl* 2001, **Suppl 37**:120-125.

- 119 Li C, Wong WH: **Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection.** *Proc Natl Acad Sci U S A* 2001, **98**:31-36.
- 120 Li C, Hung Wong W: **Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application.** *Genome Biol* 2001, **2**:RESEARCH0032.
- 121 Harr B, Schlotterer C: **Comparison of algorithms for the analysis of Affymetrix microarray data as evaluated by co-expression of genes in known operons.** *Nucleic Acids Res* 2006, **34**:e8.
- 122 Choe SE, Boutros M, Michelson AM, Church GM, Halfon MS: **Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset.** *Genome Biol* 2005, **6**:R16.
- 123 Zhou L, Rocke DM: **An expression index for Affymetrix GeneChips based on the generalized logarithm.** *Bioinformatics* 2005, **21**:3983-3989.
- 124 Hochreiter S, Clevert DA, Obermayer K: **A new summarization method for Affymetrix probe level data.** *Bioinformatics* 2006, **22**:943-949.
- 125 Cope LM, Irizarry RA, Jaffee HA, Wu Z, Speed TP: **A benchmark for Affymetrix GeneChip expression measures.** *Bioinformatics* 2004, **20**:323-331.
- 126 Datta S, Datta S: **Comparisons and validation of statistical clustering techniques for microarray gene expression data.** *Bioinformatics* 2003, **19**:459-466.
- 127 Kerr MK, Churchill GA: **Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments.** *Proc Natl Acad Sci U S A* 2001, **98**:8961-8965.
- 128 West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, Zuzan H, Olson JAJ, Marks JR, Nevins JR: **Predicting the clinical status of human breast cancer by using gene expression profiles.** *Proc Natl Acad Sci U S A* 2001, **98**:11462-11467.
- 129 Crescenzi M, Giuliani A: **The main biological determinants of tumor line taxonomy elucidated by a principal component analysis of microarray data.** *FEBS Lett* 2001, **507**:114-118.
- 130 Misra J, Schmitt W, Hwang D, Hsiao LL, Gullans S, Stephanopoulos G, Stephanopoulos G: **Interactive exploration of microarray gene expression patterns in a reduced dimensional space.** *Genome Res* 2002, **12**:1112-1120.
- 131 Alter O, Brown PO, Botstein D: **Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms.** *Proc Natl Acad Sci U S A* 2003, **100**:3351-3356.
- 132 Alter O, Brown PO, Botstein D: **Singular value decomposition for genome-wide expression data processing and modeling.** *Proc Natl Acad Sci U S A* 2000, **97**:10101-10106.
- 133 Raychaudhuri S, Stuart JM, Altman RB: **Principal components analysis to summarize microarray experiments: application to sporulation time series.** *Pac Symp Biocomput* 2000:455-466.
- 134 Wall ME, Dyck PA, Brettin TS: **SVDMAN--singular value decomposition analysis of microarray data.** *Bioinformatics* 2001, **17**:566-568.
- 135 Landgrebe J, Wurst W, Welzl G: **Permutation-validated principal components analysis of microarray data.** *Genome Biol* 2002, **3**:RESEARCH0019.
- 136 Krzanowski WJ: *Principles of Multivariate Analysis - A User's Perspective.* New York: Oxford University Press Inc.; 1996.
- 137 Yeung KY, Ruzzo WL: **Principal component analysis for clustering gene expression data.** *Bioinformatics* 2001, **17**:763-774.
- 138 Yeung MK, Tegner J, Collins JJ: **Reverse engineering gene networks using singular value decomposition and robust regression.** *Proc Natl Acad Sci U S A* 2002, **99**:6163-6168.
- 139 Price ND, Reed JL, Papin JA, Famili I, Palsson BO: **Analysis of metabolic capabilities using singular value decomposition of extreme pathway matrices.** *Biophys J* 2003, **84**:794-804.

- 140 Liu Y, Vincenti MP, Yokota H: **Principal component analysis for predicting transcription-factor binding motifs from array-derived data.** *BMC Bioinformatics* 2005, **6**:276.
- 141 D'Haeseleer P: **How does gene expression clustering work?** *Nat Biotechnol* 2005, **23**:1499-1501.
- 142 Gibbons FD, Roth FP: **Judging the quality of gene expression-based clustering methods using gene annotation.** *Genome Res* 2002, **12**:1574-1581.
- 143 Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci U S A* 1998, **95**:14863-14868.
- 144 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.
- 145 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.
- 146 Wen X, Fuhrman S, Michaels GS, Carr DB, Smith S, Barker JL, Somogyi R: **Large-scale temporal gene expression mapping of central nervous system development.** *Proc Natl Acad Sci U S A* 1998, **95**:334-339.
- 147 Zadeh LA: **Fuzzy Sets.** *J of Information and Control* 1965, **8**:338-353.
- 148 Woolf PJ, Wang Y: **A fuzzy logic approach to analyzing gene expression data.** *Physiol Genomics* 2000, **3**:9-15.
- 149 Asyali MH, Alci M: **Reliability analysis of microarray data using fuzzy c-means and normal mixture modeling based classification methods.** *Bioinformatics* 2005, **21**:644-649.
- 150 Dembele D, Kastner P: **Fuzzy C-means method for clustering microarray data.** *Bioinformatics* 2003, **19**:973-980.
- 151 Gasch AP, Eisen MB: **Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering.** *Genome Biol* 2002, **3**:RESEARCH0059.
- 152 Wang J, Bo TH, Jonassen I, Myklebost O, Hovig E: **Tumor classification and marker gene prediction by feature selection and fuzzy c-means clustering using microarray data.** *BMC Bioinformatics* 2003, **4**:60.
- 153 Belacel N, Cuperlovic-Culf M, Laflamme M, Ouellette R: **Fuzzy J-Means and VNS methods for clustering genes from microarray data.** *Bioinformatics* 2004, **20**:1690-1701.
- 154 Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR: **Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation.** *Proc Natl Acad Sci U S A* 1999, **96**:2907-2912.
- 155 Toronen P, Kolehmainen M, Wong G, Castren E: **Analysis of gene expression data using self-organizing maps.** *FEBS Lett* 1999, **451**:142-146.
- 156 Nikkila J, Toronen P, Kaski S, Venna J, Castren E, Wong G: **Analysis and visualization of gene expression data using self-organizing maps.** *Neural Netw* 2002, **15**:953-966.
- 157 Rutten JHGC: **Polyhedral clustering.** Ph.D. thesis, Universiteit Maastricht, 1998.
- 158 Tibshirani R, Walter G, Hastie T: **Estimating the number of clusters in a dataset via the gap statistic.** *J Roy Statist Soc B* 2001, **63**:411-423.
- 159 McShane LM, Radmacher MD, Freidlin B, Yu R, Li MC, Simon R: **Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data.** *Bioinformatics* 2002, **18**:1462-1469.
- 160 Yeung KY, Haynor DR, Ruzzo WL: **Validating clustering for gene expression data.** *Bioinformatics* 2001, **17**:309-318.
- 161 Handl J, Knowles J, Kell DB: **Computational cluster validation in post-genomic data analysis.** *Bioinformatics* 2005, **21**:3201-3212.
- 162 Gat-Viks I, Sharan R, Shamir R: **Scoring clustering solutions by their biological relevance.** *Bioinformatics* 2003, **19**:2381-2389.

- 163 Simon R, Radmacher MD, Dobbin K, McShane LM: **Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification.** *J Natl Cancer Inst* 2003, **95**:14-18.
- 164 Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.** *Science* 1999, **286**:531-537.
- 165 Dudoit S, Fridlyand J, Speed T: **Comparison of discrimination methods for the classification of tumors using gene expression data.** *J Amer Stat Assoc* 2002, **97**:77-87.
- 166 Cheok MH, Yang W, Pui CH, Downing JR, Cheng C, Naeve CW, Relling MV, Evans WE: **Treatment-specific changes in gene expression discriminate in vivo drug response in human leukemia cells.** *Nat Genet* 2003, **34**:85-90.
- 167 Mendez MA, Hodar C, Vulpe C, Gonzalez M, Cambiazo V: **Discriminant analysis to evaluate clustering of gene expression data.** *FEBS Lett* 2002, **522**:24-28.
- 168 Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M, Jr., Haussler D: **Knowledge-based analysis of microarray gene expression data by using support vector machines.** *Proc Natl Acad Sci U S A* 2000, **97**:262-267.
- 169 Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D: **Support vector machine classification and validation of cancer tissue samples using microarray expression data.** *Bioinformatics* 2000, **16**:906-914.
- 170 Chow ML, Moler EJ, Mian IS: **Identifying marker genes in transcription profiling data using a mixture of feature relevance experts.** *Physiol Genomics* 2001, **5**:99-111.
- 171 Ji X, Li-Ling J, Sun Z: **Mining gene expression data using a novel approach based on hidden Markov models.** *FEBS Lett* 2003, **542**:125-131.
- 172 Ooi CH, Tan P: **Genetic algorithms applied to multi-class prediction for the analysis of gene expression data.** *Bioinformatics* 2003, **19**:37-44.
- 173 Pan KH, Lih CJ, Cohen SN: **Analysis of DNA microarrays using algorithms that employ rule-based expert knowledge.** *Proc Natl Acad Sci U S A* 2002, **99**:2118-2123.
- 174 Ambroise C, McLachlan GJ: **Selection bias in gene extraction on the basis of microarray gene-expression data.** *Proc Natl Acad Sci U S A* 2002, **99**:6562-6566.
- 175 Michiels S, Koscielny S, Hill C: **Prediction of cancer outcome with microarrays: a multiple random validation strategy.** *Lancet* 2005, **365**:488-492.
- 176 Eden P, Ritz C, Rose C, Ferno M, Peterson C: **"Good Old" clinical markers have similar power in breast cancer prognosis as microarray gene expression profilers.** *Eur J Cancer* 2004, **40**:1837-1841.
- 177 Chung CH, Bernard PS, Perou CM: **Molecular portraits and the family tree of cancer.** *Nat Genet* 2002, **32 Suppl**:533-540.
- 178 DeRisi J, Penland L, Brown PO, Bittner ML, Meltzer PS, Ray M, Chen Y, Su YA, Trent JM: **Use of a cDNA microarray to analyse gene expression patterns in human cancer.** *Nat Genet* 1996, **14**:457-460.
- 179 Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lonning PE, Borresen-Dale AL, Brown PO, Botstein D: **Molecular portraits of human breast tumours.** *Nature* 2000, **406**:747-752.
- 180 van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH: **Gene expression profiling predicts clinical outcome of breast cancer.** *Nature* 2002, **415**:530-536.
- 181 van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, et al: **A gene-expression signature as a predictor of survival in breast cancer.** *N Engl J Med* 2002, **347**:1999-2009.
- 182 Reis-Filho JS, Westbury C, Pierga JY: **The impact of expression profiling on prognostic and predictive testing in breast cancer.** *J Clin Pathol* 2006, **59**:225-231.
- 183 R Development Core Team: *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing; 2004.

- 184 Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, et al: **Bioconductor: open software development for computational biology and bioinformatics.** *Genome Biol* 2004, **5**:R80.
- 185 Bioconductor [<http://www.bioconductor.org>]
- 186 Gautier L, Cope L, Bolstad BM, Irizarry RA: **affy--analysis of Affymetrix GeneChip data at the probe level.** *Bioinformatics* 2004, **20**:307-315.
- 187 Lindsey JK: *Models for Repeated Measurements.* Oxford: Oxford University Press; 1999.
- 188 Lindsey JK: R modeling libraries [<http://popgen.unimaas.nl/~jlinsey/rcode.html>]
- 189 Harbig J, Sprinkle R, Enkemann SA: **A sequence-based identification of the genes detected by probesets on the Affymetrix U133 plus 2.0 array.** *Nucleic Acids Res* 2005, **33**:e31.
- 190 Carter SL, Eklund AC, Mecham BH, Kohane IS, Szallasi Z: **Redefinition of Affymetrix probe sets by sequence overlap with cDNA microarray probes reduces cross-platform inconsistencies in cancer-associated gene expression measurements.** *BMC Bioinformatics* 2005, **6**:107.
- 191 Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, Bunney WE, Myers RM, Speed TP, Akil H, Watson SJ, Meng F: **Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data.** *Nucleic Acids Res* 2005, **33**:e175.
- 192 DAVID / EASE [<http://david.niaid.nih.gov/david/ease.htm>]
- 193 NCBI website [<http://www.ncbi.nlm.nih.gov/>]
- 194 UCSC [<http://genome.ucsc.edu/>]
- 195 Ensembl [<http://www.ensembl.org/index.html>]
- 196 Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25**:25-29.
- 197 Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, et al: **The Gene Ontology (GO) database and informatics resource.** *Nucleic Acids Res* 2004, **32**:D258-261.
- 198 **KEGG, Kyoto Encyclopedia of Genes and Genomes.**
- 199 Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer EL, Bateman A: **Pfam: clans, web tools and services.** *Nucleic Acids Res* 2006, **34**:D247-251.
- 200 Balakrishnan R, Christie KR, Costanzo MC, Dolinski K, Dwight SS, Engel SR, Fisk DG, Hirschman JE, Hong EL, Nash R, et al: **Saccharomyces Genome Database (SGD)** [<http://www.yeastgenome.org/>]
- 201 Yeast Proteome Database (YPD) [<http://proteome.incyte.com/>]
- 202 ABIM Online analysis tools [<http://www.up.univ-mrs.fr/~wabim/english/logligne.html>]
- 203 Galperin MY: **The Molecular Biology Database Collection: 2006 update.** *Nucleic Acids Res* 2006, **34**:D3-5.
- 204 Hosack DA, Dennis G, Jr., Sherman BT, Lane HC, Lempicki RA: **Identifying biological themes within lists of genes with EASE.** *Genome Biol* 2003, **4**:R70.
- 205 Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA: **DAVID: Database for Annotation, Visualization, and Integrated Discovery.** *Genome Biol* 2003, **4**:R60.
- 206 Dahlquist KD, Salomonis N, Vranizan K, Lawlor SC, Conklin BR: **GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways.** *Nat Genet* 2002, **31**:19-20.
- 207 Doniger SW, Salomonis N, Dahlquist KD, Vranizan K, Lawlor SC, Conklin BR: **MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data.** *Genome Biol* 2003, **4**:R7.
- 208 Beissbarth T, Speed TP: **GOstat: find statistically overrepresented Gene Ontologies within a group of genes.** *Bioinformatics* 2004, **20**:1464-1465.
- 209 Becker KG, Hosack DA, Dennis G, Jr., Lempicki RA, Bright TJ, Cheadle C, Engel J: **PubMatrix: a tool for multiplex literature mining.** *BMC Bioinformatics* 2003, **4**:61.

- 210 Glenisson P, Coessens B, Van Vooren S, Mathys J, Moreau Y, De Moor B: **TXTGate: profiling gene groups with text-based information**. *Genome Biol* 2004, **5**:R43.
- 211 Microarray Gene Expression Data (MGED) organization [<http://www.mged.org>]
- 212 Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, et al: **Minimum information about a microarray experiment (MIAME)-toward standards for microarray data**. *Nat Genet* 2001, **29**:365-371.
- 213 Ball CA, Sherlock G, Parkinson H, Rocca-Sera P, Brooksbank C, Causton HC, Cavalieri D, Gaasterland T, Hingamp P, Holstege F, Ringwald M, Spellman P, Stoeckert CJ, Jr., Stewart JE, Taylor R, Brazma A, Quackenbush J: **The underlying principles of scientific publication**. *Bioinformatics* 2002, **18**:1409.
- 214 Ball CA, Sherlock G, Parkinson H, Rocca-Sera P, Brooksbank C, Causton HC, Cavalieri D, Gaasterland T, Hingamp P, Holstege F, Ringwald M, Spellman P, Stoeckert CJ, Jr., Stewart JE, Taylor R, Brazma A, Quackenbush J: **Standards for microarray data**. *Science* 2002, **298**:539.
- 215 Perou CM: **Show me the data!** *Nat Genet* 2001, **29**:373.
- 216 Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, Holloway E, Kapushesky M, Kemmeren P, Lara GG, Oezcimen A, Rocca-Serra P, Sansone SA: **ArrayExpress--a public repository for microarray gene expression data at the EBI**. *Nucleic Acids Res* 2003, **31**:68-71.
- 217 Parkinson H, Sarkans U, Shojatalab M, Abeygunawardena N, Contrino S, Coulson R, Farne A, Lara GG, Holloway E, Kapushesky M, Lilja P, Mukherjee G, Oezcimen A, Rayner T, Rocca-Serra P, Sharma A, Sansone S, Brazma A: **ArrayExpress--a public repository for microarray gene expression data at the EBI**. *Nucleic Acids Res* 2005, **33**:D553-555.
- 218 Edgar R, Domrachev M, Lash AE: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository**. *Nucleic Acids Res* 2002, **30**:207-210.
- 219 Ikeo K, Ishii J, Tamura T, Gojobori T, Tateno Y: **CIBEX: center for information biology gene expression database**. *C R Biol* 2003, **326**:1079-1082.
- 220 Ball CA, Brazma A, Causton H, Chervitz S, Edgar R, Hingamp P, Matese JC, Parkinson H, Quackenbush J, Ringwald M, Sansone SA, Sherlock G, Spellman P, Stoeckert C, Tateno Y, Taylor R, White J, Winegarden N: **Submission of microarray data to public repositories**. *PLoS Biol* 2004, **2**:E317.
- 221 Spellman PT, Miller M, Stewart J, Troup C, Sarkans U, Chervitz S, Bernhart D, Sherlock G, Ball C, Lepage M, et al: **Design and implementation of microarray gene expression markup language (MAGE-ML)**. *Genome Biol* 2002, **3**:RESEARCH0046.

Chapter 2

A novel stepwise analysis procedure of genome-wide expression profiles identifies transcript signatures of thiamine genes as classifiers of mitochondrial mutants in yeast

L.M.T. Eijssen^{1}, P.J. Lindsey¹, R. Peeters⁴, R.L. Westra⁴, R.G.E. van Eijsder², M. Bolotin-Fukuhara³, H.J.M. Smeets^{1,2}, R.F.M. Vlietinck^{1,5}*

¹Department of Population Genetics, Genomics & Bioinformatics, Maastricht University, P.O. Box 616, 6200 MD, Maastricht, the Netherlands

²Department of Clinical Genetics, Maastricht University, P.O. Box 616, 6200 MD, Maastricht, the Netherlands

³Institut de Génétique et Microbiologie (CNRS UMR 8621), Batiment 400, Université Paris-Sud, 91405 Orsay Cedex, France.

⁴Department of Mathematics, Maastricht University, P.O. Box 616, 6200 MD, Maastricht, the Netherlands

⁵Department of Human Genetics, Leuven University, Herestraat 49, Box 602, B-3000, Leuven, Belgium

** Corresponding Author:*

L.M.T. Eijssen, Department of Genetics and Cell Biology, Maastricht University, P.O. Box 616, 6200 MD Maastricht, the Netherlands. Tel.: +31-43-3881995. Fax: +31-43-3884573. E-mail: Lars.Eijssen@gen.unimaas.nl

submitted

Abstract

To extract functional information on genes and processes from large expression data sets, analysis methods are required that can both computationally deal with these amounts of data and are tunable to specific research questions. To satisfy both requirements, a stepwise procedure that combines elements from both principal component analysis and discriminant analysis, was developed to specifically retrieve genes involved in processes of interest and classify samples based upon those genes.

In a global expression data set of 300 gene knock-outs in *Saccharomyces cerevisiae*, the procedure successfully classified knock-outs with equal protein localization by expression signatures of limited numbers of genes. The genes discriminating "mitochondrion" from the other subgroups were evaluated in more detail.

The thiamine pathway turned out to be one of the processes involved and was successfully evaluated in a logistic model to predict whether yeast knock-outs were mitochondrial or not. Further, this pathway is biologically related to the mitochondrial system. Hence, this strongly indicates that our approach is effective and efficient in extracting meaningful information from large microarray experiments and assigning functions to yet uncharacterized genes.

Keywords

principal component analysis, discriminant analysis, logistic regression, *Saccharomyces cerevisiae*, gene expression microarray analysis

Introduction

The completion of the sequence of several genomes, from yeast [1] to man [2], and the development of high-throughput genetic technology have enabled creation of huge data sets with genomic information. These extensive functional genomics studies enable large-scale data mining and put a challenge on bioinformatics. For finding biological functions of new genes, analysis methods need to be able to both handle these data sets and to tune to specific research questions. Many analysis techniques can extract information from microarray data [3].

The novel approach described in this paper combines a non-supervised preprocessing step (principal component analysis, PCA) with a supervised step that builds classification functions for groups within the data at hand (discriminant analysis, DA). PCA is often used in bioinformatics as a first data reduction step in sample classification or clustering [4, 5]. It is also applied as part of network reconstruction methods [6] and metabolic network research [7]. Some studies apply this technique (or adaptations of it) to gene clustering or classification [8, 9]. Discriminant analysis is used to find statistically significant group differences in expression studies [10], sometimes on its own [11] or following PCA [12]. In the current study, PCA results were used in a different way as done before, as the eigenvectors were not used directly, but genes were ranked by their weighed participation in all eigenvectors. Highly ranked genes were considered to be influential for the systems under investigation and used to build discriminant functions.

To evaluate our method in finding new information on genes or pathways from genome wide expression data, a *Saccharomyces cerevisiae* deletion mutant data set [13] was used. Yeast data sets are attractive as they are extensive, the annotation of the yeast genome is good, and the biological material under investigation is relatively well characterized and uniform. The yeast genome has been sequenced [1] and has been used extensively [14] in functional genomics [15-23], proteomics [24, 25], and comparative genomics research [26, 27]. The data set contains whole genome gene expression profiles of 300 experiments, mainly involving gene knock-out strains, but also some chemically treated strains. It was shown before [13] that many classes of mutants displayed recognizable and distinct expression profiles.

We conclude that the method presented in the current research is effective and efficient in extracting meaningful information from large microarray experiments and assigning functions to yet uncharacterized genes. It was able to distinguish mitochondria-related gene knock-outs from other knock-outs and identified thiamin related genes as classifiers to make this distinction.

Materials and methods

Stepwise data reduction and gene classification method

The first step in the method was performing a PCA, a non-supervised multivariate data reduction method [28]. Standard implementations of PCA first compute the covariance matrix from which the principal components are then calculated. As this matrix has a number of rows and columns equal to the number of genes in the data set, this can create computational problems in high throughput studies. Therefore, an implementation using matrix decompositions [29] was used. The code was written for and was run in the Matlab package, version 6.1. Within the procedure, complete PCA results were used to reduce the size of the data set by keeping the most informative genes as variables, instead of transforming them to components composed of several genes as is generally the purpose of PCA and also done in other studies that combine PCA and DA [12]. For each gene and each component the product from the contribution of that gene to that component and the eigenvalue of the component was computed (Table 1). All genes were ranked according to their maximum values and the best scoring ones from the resulting list were chosen. With this step, the number of genes was reduced to a computationally feasible amount of 200 genes.

Table 1 Computation of scores for genes: k indicates the required number of genes; contr. = contribution.

Gene name	Eigenvalue comp 1 X_1	Contr. in comp 1 Y_1	$X_1 * Y_1$	Eigenvalue comp 2 X_2	Contr. in comp 2 Y_2	$X_2 * Y_2$	...	...	Max ($X_i * Y_i$)
Gene 1	X_{11}	Y_{11}	$X_{11} * Y_{11}$	X_{21}	Y_{21}	$X_{21} * Y_{21}$	...	...	$\text{Max}(X_{11} * Y_{11}, X_{21} * Y_{21}, \dots)$
Gene 2	X_{11}	Y_{11}	$X_{12} * Y_{12}$	X_{22}	Y_{22}	$X_{22} * Y_{22}$	...	...	$\text{Max}(X_{12} * Y_{12}, X_{22} * Y_{22}, \dots)$
...									
Final Selection				The k genes with the maximum values of the last column					

The next step was DA, a multivariate supervised technique that selects the optimal set of contributing genes to distinguish between two or more subgroups of samples in the data set [28]. First discriminating functions were built in a (forward) stepwise way with the *proc stepdisc* method in SAS version 8, using the default SAS P-value criterion of 0.15. Classification performance of the functions was evaluated with the *proc discrim* method that enables resubstitution and cross validation testing. Resubstitution tests the fit of the function to the training set by evaluating how well it classifies the samples in that set. If no independent data set is available, cross validation is used to test the extrapolation capacity of the function built.

Functions were built to distinguish between several subgroups of the samples where one of those was of special interest. Each group was compared pairwise to each of the other ones. Experiments belonging to both groups (due to multiple annotations) of a particular comparison were left out for that one. The comparisons not involving the group of special interest were used to filter out genes that are non-specific for the distinctions of interest. The rank for a gene was calculated by counting the number of times it was present in a function that discriminates the group of special interest from another group and then subtracting the weighted number of times it was present in discriminant functions between other groups. This ranking measure is given by the following formula for comparing one specific group, assigned index 1, to n others (a total of $n + 1$ groups):

$$\sum_{j=2}^{n+1} a_{1j} - \left(\sum_{i=2}^{n+1} \sum_{j=2}^{n+1} a_{ij} \times \frac{2}{n-1} \right), \quad (1)$$

where group index $i \neq j$, $n \geq 2$, and $a_{ij} = 1$ if the gene is present in the function comparing group i to group j and 0 otherwise. This formula can take values between minus and plus the number of groups (n) to be compared to.

The high-ranked genes were characterized by looking at their description and Gene Ontology (GO) annotations [30] using the *Saccharomyces* Genome Database [31]. Genes that were related to the research question at hand or pathways that were prominent in the results by the presence of multiple members were selected for further evaluation. To evaluate the method and the results, the classification performance of logistic models [32] that only used the selected genes or pathways was checked. These models to predict class membership for each of the samples were built in a stepwise way, choosing each time the model with the lowest Akaike Information Criterion (AIC) [33]. The AIC scores a model by balancing fit and complexity, penalizing the minus log-likelihood with the number of parameters the model contains.

In case of two outcomes, coded zero and one, the general logistic formula to compute the probability for the outcome to be one is:

$$P(\text{outcome}) \sim \frac{e^X}{1 + e^X},$$

where X indicates the linear combination of the intercept and the genes in the model, where each term is multiplied by the coefficient of that gene. So for each of the samples, the predicted classification can be computed by filling in the expression values for that experiment of the genes in the model. How well the model performed, could be rated by evaluating the predictions it made. Furthermore, important genes were identified within each model. The models were computed with the statistical package R version 2.1.0 [34].

Data

The publicly available data set used contains expression profiles of 6,325 genes from 300 gene knock-outs and chemical treatment experiments in yeast as described in [13] and available from Rosetta Inpharmatics [35]. The yeast strains are compared to a reference strain in a two-color cDNA microarray experiment. Strains are cultured on Synthetic Complete medium containing 2% glucose and grown overnight to mid-log phase. After dilution, strains are grown for several hours until equal concentrations are reached for experimental and control cultures.

The logged (base 10) ratio expression values are in the -2 to 2 range. 331 genes lacking seven values or more were removed for the data set. The remaining missing values were filled by an imputing approach that replaced them by the mean value of that gene in the other experiments (0.05% of all values in the data set). This resulted in a data set containing 5,994 genes. For each knock-out experiment, ontology information was added by looking up the cellular component entry of the specific gene in the Yeast Proteome Database [36]. Five subgroups of experiments were used: mitochondrial, lysosome / vacuole / peroxisome (lvp), cell membrane, cytoplasm, and cytoskeleton. The group of mitochondrial knock-out experiments was selected as group of special interest and the other four as controls. This allowed filtering out genes that were non-specific for the mitochondrial distinctions. Also, the four other groups were chosen to be of differing size, in order to better estimate the general performance of the procedure. As indicated, groups were chosen to be distinct with respect to cellular component, but this may not be the case for cellular function. For example, some cytoplasmic proteins have a mitochondrial function. An overview of the design is given in Figure 1.

For respectively estimating the effects of a somewhat variable but quite well classified subgroup, a highly variable subgroup with possible misclassifications, and equal size subgroups, additional modeling was done. The mitochondrial group was respectively compared with the group of all experiments in one of the four other classes ('other4'), with the group of all other experiments combined ('rest', this indicates 'other4' plus all experiments in the original data set that were not used in any subgroups), and with an equally sized group consisting of the cell membrane and cytoskeleton groups.

Figure 1 Study design: comparing each group versus all the others. The numbers indicate the number of experiments in each group.

Results

To get a first idea about the information present, PCA was used on the data set and the genes that contributed most to the three components with the highest eigenvalues were evaluated. The analysis showed that component one was quite variable, a mean effect possibly, but a quite large number of transcription factors appeared to be included. The second component involved many genes related to amino acid metabolism. The third component contained many glucose metabolism and cell stress genes.

Then a top set of 200 genes was extracted from the PCA analysis by applying the gene selection method, using the eigenvalues and gene coefficients. These genes were used as input for the following DA step. Pairwise discriminating functions were built to classify between each tuple of the five subgroups: mitochondrial, lvp, cell membrane, cytoplasm, and cytoskeleton. The number of genes present in each of the discriminant functions tended to increase with larger and more variable groups and ranged between 7 and 51 (data not shown). Resubstitution and cross validation were performed, to evaluate the internal and external classification quality of the functions. Some classification errors were made in the mitochondrion versus rest experiment only, with five mitochondrial knock-outs and three of the others misclassified on resubstitution, six and eight on cross validation. All the other classifiers made no errors on either resubstitution or cross validation.

Computation of the ranking measurement described in Equation (1) yielded a list of genes with potential mitochondrial function. The complete analysis table contained all 200 genes that were originally selected from the PCA analysis (not shown). As the number of groups (n) that the mitochondrial one was compared to, equaled 4 in this case, this value was the maximum possible score. Table 2 presents all the genes with a score higher than 1 (maximum value reached was 3) together with their description. These are the genes that are most overrepresented in functions that discriminate between mitochondrial samples and others compared to those between two classes of non-mitochondrial samples.

Table 2 Top ranked genes (rank > 1): the Score column indicates the results of the ranking measurement: the number of occurrences of a gene in a discriminant function that compares the mitochondrial group to one of the other four groups, minus the scaled number of occurrences between the other four groups.

Gene Code	Gene Name	Description	Score
YNL274C		Putative hydroxyisocaproate dehydrogenase	3.0
YOR348C	PUT4	Proline-specific permease (also capable of transporting alanine and glycine)	3.0
YBL044W		Hypothetical ORF	2.0
YJL116C	NCA3	Regulates with NCA2p mitochondrial expression of subunits 6 (Atp6p) and 8 (Atp8p) of the Fo-F1 ATP synthase	2.0
YJR159W	SOR1	Sorbitol-induced sorbitol dehydrogenase	2.0
YLR031W		Hypothetical ORF	2.0
YOL055C	THI20	Hydroxymethylpyrimidine phosphate kinase, involved in the last steps in thiamine biosynthesis; member of a gene family with THI21 and THI22; functionally redundant with Thi21p	2.0
YAL063C	FLO9	Lectin-like protein with similarity to Flo1p, thought to be expressed and involved in flocculation	1.3
YBR240C	THI2	Zinc finger protein of the Zn(II)2Cys6 type, probable transcriptional activator of thiamine biosynthetic genes	1.3
YDR533C	HSP31	Possible chaperone and cysteine protease; member of the DJ-1/ThiJ/Pfpl superfamily	1.3
YER096W	SHC1	Sporulation-specific activator of Chs3p (chitin synthase III), required for the synthesis of the chitosan layer of ascospores; has similarity to Skt5p; transcriptionally induced at alkaline pH	1.3
YGL096W	TOS8	Target of SBF	1.3
YJL221C	FSP2	Unknown function, expression is induced during nitrogen limitation; similar to maltase (alpha-D-glucosidase)	1.3
YLR327C	RBF9	Hypothetical ORF	1.3
YMR107W	SPG4	Protein required for survival at high temperature during stationary phase	1.3

NCA3 is clearly a gene with a mitochondrial function. The GO biological process annotation of this gene is mitochondrion organization and biogenesis. It regulates the expression of ATP6p and ATP8p, which are both mitochondrially encoded subunits of complex V [37]. Two further genes in the list, THI2 and THI20, are involved in the thiamine pathway (Figure 2). The thiamine system is functionally related to mitochondria since thiaminepyrophosphate acts as cofactor of some key enzymes involved in the metabolic distribution between fermentation and respiration [38].

Figure 2 Overview of the thiamin pathway: HMP = hydromethylpyrimidine, HET = hydroxyethylthiazole, P = phosphate, PP = pyrophosphate.

To test the relevance of the thiamin pathway further, a thiamine gene list was created and used for the logistic regression modeling (Table 3). Using the stepwise logistic modeling procedure described above, separate models were built to distinguish the mitochondrial samples from the other4 group and from each of its four classes separately.

Table 3 Thiamine genes used as input for the logistic modeling validation approach.

Gene code	Gene name	Gene code	Gene name
YMR108W	ILV2 / THI1	YLR044C	PDC1
YBR240C	THI2	YDR081C	PDC2
YDL080C	THI3	YLR134W	PDC5
YGR144W	THI4	YGR087C	PDC6
YFL058W	THI5	YCR020C	PET18 / HIT2
YPL214C	THI6	YBR092C	PHO3
YLR237W	THI10 / THI7	YIL119C	RPI1
YJR156C	THI11	YMR095C	SNO1
YNL332W	THI12	YNL334C	SNO2
YDL244W	THI13	YFL060C	SNO3
YOL055C	THI20	YMR096W	SNZ1
YPL258C	THI21	YNL333W	SNZ2
YPR121W	THI22	YFL059W	SNZ3
YOR143C	THI80	YPR074C	TKL1
YIL125W	KGD1	YBR117C	TKL2
YER178W	PDA1	YGR096W	TPC1
YBR221C	PDB1		

Table 4 presents all models and the coefficients of the genes they include. The outcome of a logistic model indicates the probability for a sample to be mitochondrial. Hence, ideally the prediction should be close to 1 for a mitochondrial knock-out experiment and close to 0 otherwise. To ensure proper modeling with the numbers of genes and samples under consideration, only main effect models were built.

For the mitochondrion versus cell membrane comparison, a model without the intercept performed better. The mitochondrion versus other4 model showed that higher expression of SNO1, SNO3 and SNZ3 increased the probability of classifying the sample as mitochondrial, in order of their effects. The genes ILV2/THI1, KGD1, THI5, and SNZ2 decreased this probability with similar effect sizes. Several genes participated in more than one of the specific group models. Higher expression of KGD1 decreased the probability of mitochondrial classification in the comparisons with cytoplasm and with skeleton. THI5 decreased the probability in the comparisons against cytoplasm and lvp. The SNO2 gene showed opposite effects in the comparison against lvp and plasmamembrane. Maybe this gene was selective for the other groups and not the mitochondrial one, further indicated by its absence in the mitochondrion versus other4 model. Besides this gene, between different models, the coefficients and specifically the direction of the effect of all genes were consistent. In Figure 3, the prediction results are presented for each of the models, showing their good performance.

Table 4 Coefficients of the thiamine genes significantly present in the logistic regression models.

Gene	Model	Mitochondrion versus other4	Mitochondrion versus cytoplasm	Mitochondrion versus lyp	Mitochondrion versus skeleton	Mitochondrion versus plasma- membrane
Intercept		-1.59	-0.77	3.05	3.35	-
ILV2/THI1		-19.50	-13.90	-	-	-
THI4		-	13.75	-	-	-
THI5		-20.81	-13.93	-22.60	-	-
THI6		-	-	-	-	43.93
THI11		-	-	-	-	-38.03
THI13		-	-	-	-	35.23
THI22		-	-	-	-	-36.78
THI80		-	-	29.43	-	-
KGD1		-18.77	-26.50		-27.28	-
PDC2		-	-	-	-	33.41
PDC5		-	-	-	-43.91	-
PDC6		-	-	-	27.98	-
SNO1		4.21	4.43	-	-	-
SNO2		-	-	37.12	-	-19.94
SNO3		19.59	13.01	-	-	-
SNZ2		-17.35	-	-	-	-
SNZ3		34.80	-	21.26	-	-

The other (unknown) genes listed in Table 2 should be functionally evaluated for a mitochondrial function as well. Indicatively, several of these genes encode proteins that have scores indicating mitochondrial functions for the MitoP2 predictor [39], a tool computing overall probability for mitochondrial function. These include YNL274C, that was also predicted to be mitochondrial by localization [40], HSP31, that was also predicted by the domain-based predictor MITOPRED [41], and SHC1, the product of which also interacts with those of the mitochondrial gene ILV3 and the thiamin gene THI6 according to the MitoP2 database [39]. In addition MITOPRED indicated that RBF9 might encode a protein with a mitochondrial function. Finally PSORTII [42], based on targeting sequence information, indicated evidence for SOR1 to be mitochondrial. Also, this gene shows strong homology to XYL2, a gene encoding a xylitol-dehydrogenase and its expression responds in a xylitol-specific way.

Figure 3 Predictions of sample classes made by several logistic regression models: mito = mitochondrion; pluses indicate mitochondrial samples, triangles the respective other group; mitochondrial samples should have predictions near a probability of 1 and the respective other class near 0.

Discussion

Using PCA, biologically meaningful components were generated, indicating that at least some of the information present in the data set could be retrieved by unsupervised methods. The relevant information present in the data set apparently outweighed the noise level (signal-to-noise ratio). Within the proposed approach, PCA results were used indirectly, because only including genes present in certain major components may overstate certain processes and ignore others. When focusing on more specific classification questions with the DA approach, correctly classifying functions could be built. If DA is applied directly to the entire data set instead of after the suggested PCA step, results are rather poor, probably because the number of non-related genes is much higher than the number of related ones. In the current analyses, only some classification errors were made by the function discriminating the mitochondrion from the 'rest' group. This was not surprising as the latter group had a large internal variation as compared to the between group distance. For the mitochondrial misclassifications in this comparison, these experiments may well be closer in expression to the widely varying and possibly biased 'rest' group, as to their own group. Furthermore, there may be genes in the latter group which are (also) mitochondrial, but of which the function is still unknown. The effect of group uniformity was also clearly demonstrated by the mitochondrial versus the 'other4' group comparison, where all the items were correctly classified. This indicated that with properly constructed groups, the system performed well.

For the DA step, relatively loose constraints (SAS defaults) were used for genes to enter the functions. Consequently the number of genes in each function was still quite large, but being too strict from start on may result in directly losing all genes. With no independent data set available, validity was checked by cross validation, an alternative method that may underestimate the probability of errors. Thus, the initial discriminant functions might still be somewhat overspecific. However, comparing all mitochondrial classification functions to all others would filter out possible false positive genes. This approach and the logistic models that were constructed next prevented eventual overspecificity to the data set at hand. The main purpose of building the discriminant functions was to find a list of known or unknown discriminating genes, important in mitochondrial genetic pathways. When evaluating biologically relevant genes in the ranked list of results, it should be realized that in *Saccharomyces cerevisiae*, glucose is exclusively fermented and the genes necessary for mitochondrial function and biogenesis are glucose repressed. In this data set, the experiments were done with yeast strains cultured on glucose, so it was hard to identify specific genes of the mitochondrial energy system. This explained that not many genes of this system were in the top list (Table 2). However, the DA results contained three genes that control the synthesis of the two moieties necessary to make thiamine, plus their regulator. These genes are another way to indicate a discrimination between respiration (which requires mitochondrial functions)

and fermentation. Two of them (THI20 and THI2) were specifically in the top list of genes. Thiamine is a cofactor in glycolysis and the pentose phosphate pathway associated to enzymes that are key steps in determining the fate of pyruvate and the fueling of the Krebs cycle. It is synthesized from two different moieties, hydroxyethylthiazole and hydromethylpyrimidine (Figure 2). In this synthesis several genes play a role: THI4, THI20, THI21 and THI22. Thiamine can also be imported via THI10 (THI7) if in dephosphorylated state (via the phosphatase PHO3). This import route takes a phosphorylation step (by THI6) into thiamine monophosphate and a transformation (by THI80) into thiamine pyrophosphate. Two transcriptional activators are involved, THI2 and THI3 [38]. In addition, THI4 is also involved in mitochondrial genome stability [43].

The results from the logistic regression modeling based on the thiamin pathway (Figure 3) confirmed that the genes of this pathway were quite well able to separate mitochondrial knock-outs from the others, not fully but certainly for a large part. Each of the models was different with regard to the specific thiamin genes it contained (Table 4). The thiamin genes present in the top of the DA list were not in the logistic models themselves, which indicates that several components of the thiamine system could independently be useful in modeling the characteristics of mitochondrial samples. Especially, as both DA and logistic regression models performed well, this was further evidence of the central role of the pathway and the consistency of our procedure. The evidence suggests that the thiamine system as a whole was biologically influenced by mitochondrial dysfunction.

The data used have been generated in yeast, but for human mitochondrial disease the conservation in pathways between yeast and man has often been used to identify common pathogenic pathways [26, 27]. This may also be the case for the thiamine pathway, as thiamine has been used as a therapy in mitochondrial disease with varying success [44-47]. Thiamine supplementation can improve lactate and pyruvate plasma levels. Yeast cells have a phenotype in a thiamine related mitochondrial deficiency when grown on fermentative carbon sources like glucose [48]. As is summarized in OMIM [1], it has been shown that defects in thiamine metabolism may underlie a well-known mitochondrial disorder, Leigh syndrome. This may be related to inhibition of the biochemical formation of thiamine triphosphate. Thiamine derivatives have also been used in treatment of this disease.

Alternative pathways may be identified by adjusting our approach. So far, logistic regression was only used to confirm that the results were meaningful and that the thiamine genes contributed to classifying the samples. To build models that perform as good as possible some other options could be taken into account for future improvements. For instance, other significant pathways could be added to make fully classifying models. This can be done by evaluating other genes from the top list or by repeating the entire procedure, leaving out the thiamine system, to find other genes

or pathways. Gene interaction terms could also be considered to be included in the models. Another possible extension is to try to build common models to classify more groups at the same time. DA can be used for multiple group classification, but the results get more difficult to interpret. Multinomial modeling can be used to validate the results in this case.

The presented method can be applied quite generically and has the advantage of being able to classify and to find genes related to the process under consideration at the same time, reducing lists of genes. In contrast to methods like neural networks and some other clustering techniques, it is no black box algorithm. If applied to samples with defects in different biological processes, genes related to those can be unraveled. By adding several other subgroups, selected genes can be filtered to be specific for the process at hand. The method can also be used to find key genes that distinguish several clinical conditions from each other. Furthermore, unknown genes can be annotated with potential functions. For instance, in the example presented some relatively unknown genes were present in the top of the ranked gene list. Several of those were also predicted to be mitochondrial by other studies and existing tools based on different criteria, confirming our results. The less known genes would be good candidates for further evaluation and functional testing.

In conclusion, the proposed method can classify samples based on high-throughput microarray gene-expression data. Furthermore, it has proven effective as a tool for finding pathways involved in a certain condition under investigation or for finding biological functions of the genes present in the data set at hand. It is an easy and computationally fast and feasible way of extracting meaningful information from large microarray experiments.

Acknowledgements

The authors wish to thank Dr. Torik Ayoubi, Dr. Aimée Paulussen and Mike Gerards for their useful comments on this manuscript.

This research was supported by the Cardiovascular Research Institute Maastricht (CARIM), the Netherlands.

References

- 1 Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG: **Life with 6000 genes.** *Science* 1996, **274**:546, 563-547.
- 2 International Human Genome Sequencing Consortium: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409**:860-921.
- 3 Butte A: **The use and analysis of microarray data.** *Nat Rev Drug Discov* 2002, **1**:951-960.
- 4 West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, Zuzan H, Olson JAJ, Marks JR, Nevins JR: **Predicting the clinical status of human breast cancer by using gene expression profiles.** *Proc Natl Acad Sci U S A* 2001, **98**:11462-11467.
- 5 Crescenzi M, Giuliani A: **The main biological determinants of tumor line taxonomy elucidated by a principal component analysis of microarray data.** *FEBS Lett* 2001, **507**:114-118.
- 6 Yeung MK, Tegner J, Collins JJ: **Reverse engineering gene networks using singular value decomposition and robust regression.** *Proc Natl Acad Sci U S A* 2002, **99**:6163-6168.
- 7 Price ND, Reed JL, Papin JA, Famili I, Palsson BO: **Analysis of metabolic capabilities using singular value decomposition of extreme pathway matrices.** *Biophys J* 2003, **84**:794-804.
- 8 Misra J, Schmitt W, Hwang D, Hsiao LL, Gullans S, Stephanopoulos G, Stephanopoulos G: **Interactive exploration of microarray gene expression patterns in a reduced dimensional space.** *Genome Res* 2002, **12**:1112-1120.
- 9 Alter O, Brown PO, Botstein D: **Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms.** *Proc Natl Acad Sci U S A* 2003, **100**:3351-3356.
- 10 Hwang D, Schmitt WA, Stephanopoulos G, Stephanopoulos G: **Determination of minimum sample size and discriminatory expression patterns in microarray data.** *Bioinformatics* 2002, **18**:1184-1193.
- 11 Cheok MH, Yang W, Pui CH, Downing JR, Cheng C, Naeve CW, Relling MV, Evans WE: **Treatment-specific changes in gene expression discriminate in vivo drug response in human leukemia cells.** *Nat Genet* 2003, **34**:85-90.
- 12 Mendez MA, Hodar C, Vulpe C, Gonzalez M, Cambiazo V: **Discriminant analysis to evaluate clustering of gene expression data.** *FEBS Lett* 2002, **522**:24-28.
- 13 Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, et al: **Functional discovery via a compendium of expression profiles.** *Cell* 2000, **102**:109-126.
- 14 Grunewald B, Winzler EA: **Treasures and traps in genome-wide data sets: case examples from yeast.** *Nat Rev Genet* 2002, **3**:653-661.
- 15 Wu LF, Hughes TR, Davierwala AP, Robinson MD, Stoughton R, Altschuler SJ: **Large-scale prediction of *Saccharomyces cerevisiae* gene function using overlapping transcriptional clusters.** *Nat Genet* 2002, **31**:255-265.
- 16 Wang W, Cherry JM, Botstein D, Li H: **A systematic approach to reconstructing transcription networks in *Saccharomyces cerevisiae*.** *Proc Natl Acad Sci U S A* 2002, **99**:16893-16898.
- 17 Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N: **Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data.** *Nat Genet* 2003, **34**:166-176.
- 18 Jelinsky SA, Estep P, Church GM, Samson LD: **Regulatory networks revealed by transcriptional profiling of damaged *Saccharomyces cerevisiae* cells: Rpn4 links base excision repair with proteasomes.** *Mol Cell Biol* 2000, **20**:8157-8167.
- 19 Ge H, Liu Z, Church GM, Vidal M: **Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*.** *Nat Genet* 2001, **29**:482-486.

- 20 Giaever G, Flaherty P, Kumm J, Proctor M, Nislow C, Jaramillo DF, Chu AM, Jordan MI, Arkin AP, Davis RW: **Chemogenomic profiling: identifying the functional interactions of small molecules in yeast.** *Proc Natl Acad Sci U S A* 2004, **101**:793-798.
- 21 Jelinsky SA, Samson LD: **Global response of *Saccharomyces cerevisiae* to an alkylating agent.** *Proc Natl Acad Sci U S A* 1999, **96**:1486-1491.
- 22 Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, Dow S, Lucau-Danila A, Anderson K, Andre B, et al: **Functional profiling of the *Saccharomyces cerevisiae* genome.** *Nature* 2002, **418**:387-391.
- 23 Guelzim N, Bottani S, Bourguin P, Kepes F: **Topological and causal structure of the yeast transcriptional regulatory network.** *Nat Genet* 2002, **31**:60-63.
- 24 Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, et al: **A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*.** *Nature* 2000, **403**:623-627.
- 25 Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y: **A comprehensive two-hybrid analysis to explore the yeast protein interactome.** *Proc Natl Acad Sci U S A* 2001, **98**:4569-4574.
- 26 Andrade MA, Sander C, Valencia A: **Updated catalogue of homologues to human disease-related proteins in the yeast genome.** *FEBS Lett* 1998, **426**:7-16.
- 27 Steinmetz LM, Scharfe C, Deutschbauer AM, Mokranjac D, Herman ZS, Jones T, Chu AM, Giaever G, Prokisch H, Oefner PJ, Davis RW: **Systematic screen for human disease genes in yeast.** *Nat Genet* 2002, **31**:400-404.
- 28 Krzanowski WJ: *Principles of Multivariate Analysis - A User's Perspective.* New York: Oxford University Press Inc.; 1996.
- 29 Peeters R: **On the computation of principal components of tall data arrays.** Maastricht, the Netherlands: Department of mathematics, Maastricht University; 2004.
- 30 Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25**:25-29.
- 31 Balakrishnan R, Christie KR, Costanzo MC, Dolinski K, Dwight SS, Engel SR, Fisk DG, Hirschman JE, Hong EL, Nash R, et al: *Saccharomyces Genome Database (SGD)* [<http://www.yeastgenome.org/>]
- 32 Lindsey JK: *Introductory Statistics: A modelling approach.* New York: Oxford University Press; 1995.
- 33 Akaike H: **Information theory and an extension of the maximum likelihood principle.** In *Second International Symposium on Inference Theory; Budapest: Akadémiai Kiadó.* Edited by Petrov BN, Csàki F. 1973: 267-281.
- 34 R Development Core Team: *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing; 2004.
- 35 Rosetta Inpharmatics [<http://www.rii.com/>]
- 36 Yeast Proteome Database (YPD) [<http://proteome.incyte.com/>]
- 37 Pelissier P, Camougrand N, Velours G, Guerin M: **NCA3, a nuclear gene involved in the mitochondrial expression of subunits 6 and 8 of the Fo-F1 ATP synthase of *S. cerevisiae*.** *Curr Genet* 1995, **27**:409-416.
- 38 Hohmann S, Meacock PA: **Thiamin metabolism and thiamin diphosphate-dependent enzymes in the yeast *Saccharomyces cerevisiae*: genetic regulation.** *Biochim Biophys Acta* 1998, **1385**:201-219.
- 39 Andreoli C, Prokisch H, Hortnagel K, Mueller JC, Munsterkotter M, Scharfe C, Meitinger T: **MitoP2, an integrated database on mitochondrial proteins in yeast and man.** *Nucleic Acids Res* 2004, **32**:D459-462.
- 40 Sickmann A, Reinders J, Wagner Y, Joppich C, Zahedi R, Meyer HE, Schonfisch B, Perschil I, Chacinska A, Guiard B, Rehling P, Pfanner N, Meisinger C: **The proteome of *Saccharomyces cerevisiae* mitochondria.** *Proc Natl Acad Sci U S A* 2003, **100**:13207-13212.

- 41 Guda C, Fahy E, Subramaniam S: **MITOPRED: a genome-scale method for prediction of nucleus-encoded mitochondrial proteins.** *Bioinformatics* 2004, **20**:1785-1794.
- 42 Nakai K, Horton P: **PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization.** *Trends Biochem Sci* 1999, **24**:34-36.
- 43 Machado CR, Praekelt UM, de Oliveira RC, Barbosa AC, Byrne KL, Meacock PA, Menck CF: **Dual role for the yeast THI4 gene in thiamine biosynthesis and DNA damage tolerance.** *J Mol Biol* 1997, **273**:114-121.
- 44 Lou HC: **Correction of increased plasma pyruvate and plasma lactate levels using large doses of thiamine in patients with Kearns-Sayre syndrome.** *Arch Neurol* 1981, **38**:469.
- 45 Mastalgia F, Thompson PL, Papadimitrou JM: **Mitochondrial myopathy, cardiomyopathy, lactic acidosis and response to prednisone and thiamine.** *Aust NZ J Med* 1980, **10**:660-664.
- 46 Matthews PM, Ford B, Dandurand RJ, Eidelman DH, O'Connor D, Sherwin A, Karpati G, Andermann F, Arnold DL: **Coenzyme Q10 with multiple vitamins is generally ineffective in treatment of mitochondrial disease.** *Neurology* 1993, **43**:884-890.
- 47 Sato Y, Nakagawa M, Higuchi I, Osame M, Naito E, Oizumi K: **Mitochondrial myopathy and familial thiamine deficiency.** *Muscle Nerve* 2000, **23**:1069-1075.
- 48 Marobbio CM, Voza A, Harding M, Bisaccia F, Palmieri F, Walker JE: **Identification and reconstitution of the yeast mitochondrial transporter for thiamine pyrophosphate.** *Embo J* 2002, **21**:5653-5661.

Chapter 3

Myocardial gene expression reveals maladaptive processes in cardiac myosin-binding protein C knock-out mice

L.M.T. Eijssen^{1^}, B.J.C. van den Bosch^{1^}, N. Vignier^{2,3}, P.J. Lindsey¹, C.M.M. van den Burg¹, L. Carrier^{2,3,4}, P. A. Doevendans^{5,6}, G.J. van der Vusse⁷, H.J.M. Smeets^{1}.*

¹Dept. of Genetics and Cell Biology, Cardiovascular Research Institute Maastricht (CARIM), Maastricht University, P.O. Box 616, 6200 MD Maastricht, the Netherlands

²Institut National de la Santé et de la Recherche Médicale, U582, IRF14, Groupe Hospitalier Pitié-Salpêtrière, Paris, F-75013, France

³Université Pierre et Marie Curie, Paris, F-75013, France

⁴Institute of Experimental and Clinical Pharmacology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

⁵Dept. of Cardiology of the Heart Lung Centre Utrecht (UMCU), P.O. Box 85500, 3508 GA Utrecht, the Netherlands

⁶Interuniversity Cardiology Institute of the Netherlands (ICIN), P.O. Box 19258, 3501 DG Utrecht, the Netherlands

⁷Dept. of Physiology, Cardiovascular Research Institute Maastricht (CARIM), Maastricht University, P.O. Box 616, 6200 MD Maastricht, the Netherlands

[^]These authors contributed equally to this work

**Corresponding Author:*

Dr. H.J.M. Smeets, Department of Genetics and Cell Biology, P.O. Box 616, 6200 MD Maastricht, the Netherlands. Tel.: +31-43-3875843 / 3875803. Fax: +31-43-3877877. E-mail: Bert.Smeets@molcelb.unimaas.nl

submitted

Abstract

Familial hypertrophic cardiomyopathy (FHC) is an autosomal dominant disease, which is characterized by left ventricular hypertrophy (LVH) predominantly affecting the interventricular septum, and associated with myocardial disarray and interstitial fibrosis. Cardiac myosin binding protein C (cMyBP-C) mutations are common causes of FHC. A cMyBP-C knock-out mouse model has been developed, in which heterozygotes display asymmetric septal hypertrophy, as in FHC, and homozygotes develop eccentric LVH. Gene expression profiling was performed on left ventricular heart RNA of 9 week-old heterozygous and homozygous mice to identify the molecular processes in, respectively, the pre-hypertrophic and the early hypertrophic heart. The (partial) lack of the cMyBP-C protein lead mainly to a clearly increased energy demand, to activation of the JNK and p38 subsystems of the MAPK signaling pathway and deactivation of ERK 1/2 signaling and to induction of apoptotic processes. The current findings strongly suggest that pathological processes are already initiated at an early phase in the cMyBP-C KO mice, even before hypertrophy occurs, and persists after hypertrophy has been established. Changes in cardiac structure, contractile proteins, protein turnover and in Ca^{2+} handling genes were less prominent, which may be due to the time point of 9 weeks, as hypertrophy is either not yet present (heterozygotes) or completed (homozygotes) at that point. Many of the changes were more pronounced in the homozygous KO mice, which may also provide explanations as to why these mice develop hypertrophy more quickly and eventually demonstrate a more severe phenotype. It will be worthwhile to study these processes in patients to investigate if the same processes are involved in humans and may relate to the pathology observed and the severity of clinical manifestation.

Keywords

Cardiac MyBP-C knock-out mouse; microarray; hypertrophy; cardiomyopathy

Introduction

Familial hypertrophic cardiomyopathy (FHC) is an autosomal dominant disease, which is characterized by left ventricular hypertrophy (LVH) with myocardial disarray and interstitial fibrosis [1]. Mutations in cardiac myosin binding protein C (cMyBP-C) are one of the most common causes, accounting for at least 20% of all mutations leading to FHC [2-6]. Cardiac MyBP-C is a large and abundant myofibrillar protein, which interacts with myosin heads, actin and titin, indicating its important structural role in the cardiac contractile machinery [7, 8]. In contrast to other sarcomeric proteins involved in FHC, considerably less is known about the molecular mechanism by which cMyBP-C gene mutations lead to the disease [8]. Most of the cMyBP-C gene mutations result in a frameshift and are expected to produce a C-terminal truncated protein. In four cases analyzed, the truncated protein was not detected in myocardial tissue of the patients, suggesting that haploinsufficiency is involved [9-11]. It is not yet understood how the partial lack of functional cMyBP-C leads to impaired contractile behavior and cardiac hypertrophy and which adaptive and maladaptive processes are involved.

The aim of the present study was to investigate early alterations in gene expression in pre-hypertrophic and hypertrophic heart due to defects in the cMyBP-C protein and to identify the molecular pathological processes involved, which might help explain the clinical phenotype in FHC patients with cMyBP-C mutations. We used a recently developed cMyBP-C knock-out (KO) mouse model with inactivation of one or two alleles of the cMyBP-C gene [12]. The heterozygous KO mice developed asymmetric septal hypertrophy associated with fibrosis at 10-11 months of age, without impairment of LV function, similar to human FHC. The homozygous KO mice showed eccentric LVH with decreased fractional shortening at 3-4 months of age and developed impaired diastolic relaxation [12]. Elucidation of the molecular events occurring in the hearts of the heterozygous and homozygous cMyBP-C KO mice may provide insight into the development of LVH associated with FHC in patients with cMyBP-C mutations. For this purpose, we performed microarray analysis to screen thousands of different genes simultaneously in hearts of heterozygous and homozygous cMyBP-C KO mice at the age of nine weeks compared with age-matched wild-type (WT) mice. Left ventricular RNA was hybridized to microarrays containing a 15K cDNA collection of embryonic origin [13, 14]. For each gene, Gaussian models were fitted to determine the differential expression of the gene under investigation. The differentially expressed genes were further evaluated using GenMAPP [15] and MAPPfinder [16], online databases at the NCBI website [17] and the PubMatrix tool [18]. The processes, in which these genes are involved, were compared between the early hypertrophic homozygous KO hearts and the pre-hypertrophic heterozygous KO hearts to assess similarities and differences in cardiac remodeling due to defects in cMyBP-C.

Materials and methods

Cardiac MyBP-C null mice

Heterozygous and homozygous cMyBP-C KO mice were sacrificed at nine weeks of age and the hypertrophy status of five homozygous KO, seven heterozygous KO and six wild-type (WT) mice was determined by calculating left ventricular weight to body weight ratio (LVW/BW). LV samples were additionally prepared from: three WT (1 male, 2 female), three heterozygous KO (1 male, 2 female), and three homozygous KO (2 male, 1 female) mice. Experiments were performed in accordance with the recommendations of the Guide for the Care and Use of Laboratory Animals published by the US National Institutes of Health (NIH publication No. 85-23, revised 1996).

Microarray preparation and design

Microarray experiments were performed according to the MIAME (Minimum Information About a Microarray Experiment) recommendations [19] and according to the same procedure as in previous work [20]. Inserts of the NIA Mouse 15K cDNA Collection, containing 15000 different cDNA clones derived from embryonic and fetal cells and tissues [13, 14] were amplified by PCR, resuspended in 50% DMSO and spotted onto UltraGAPS (Gamma Amino Propyl Silane)[™] Coated Slides (Corning, NY) using the Microgrid II arrayer (Biorobotics, UK). Each slide consisted of three meta-grids, subdivided in 16 sub-grids, corresponding to the area printed by each pin of the arrayer. Out of the total of 19200 spots, 15120 spots corresponded to the cDNA clones. The remaining 4080 spots were allocated to spikes, positive and negative controls, DMSO, or remained empty. Each of these products was spotted in positions chosen uniformly across the sub-grid and therefore the slide. Because of the spikes, possible variations between slides can be taken into account during the data analysis, allowing precise estimation of the differences in gene expression between the groups of mice.

RNA extraction, labeling and hybridization

Total left ventricular RNA was extracted using TRIzol reagent (Invitrogen, Carlsbad, CA) according to the manufacturer's protocol, followed by purification with the RNeasy clean-up kit (Qiagen, Valencia, CA). RNA quality was determined using the Bioanalyser (Agilent, Palo Alto, CA). As the Cy5-dyes have been reported to be unstable due to the amount of atmospheric ozone [21], we decided to only use Cy3 labeling and compare the samples hybridized on different slides. In previous experiments using the same slides and cDNA collection, reproducibility was checked by performing replicate hybridizations for 5 samples [20]. Aminoallyl labeling and cDNA hybridization were performed according to the protocol described by Hegde et al. [22] and slightly adjusted for use in our own laboratory. Slides were scanned with a dual-laser scanner (Affymetrix / GMS 418, Santa Clara, CA).

Microarray data analysis

The scanned microarray images were processed and quantified using ImaGene software (BioDiscovery, Boston, MA). Spots flagged by this software as being empty, negative, or poor were excluded from further analysis. The remaining probes were analyzed using a Gaussian linear regression including labeling efficiency and, depending on the number of repetitions available, sex of the animals and one or more of eight spikes into the model. Fitness of the models was evaluated by the Akaike Information Criterion (AIC) [23], and the best one determined whether the gene was significant in any or both treatment groups. If for a certain gene sufficient data was available for only one of the KO groups, this gene was analyzed only for that particular group. Only genes that were differentially expressed in one or both of the KO groups were evaluated further. First GenMAPP was used to evaluate which Gene Ontology (GO) [24] categories were significantly overrepresented in the set of genes that could be analyzed for all groups. For genes with consistent effects (e.g. not some differential and some non-differential expression) for all of their respective probes, the murine gene name and description were retrieved using LocusLink (in the meantime replaced by Entrez Gene) at the NCBI website. Homologene was used to find human homologues for the genes. Then gene descriptions were extended by information from the nucleotide and, if necessary, MeSH and OMIM entries of both the mouse genes and their human homologues (NCBI). Annotations were further completed by performing automatic PubMed literature searches with PubMatrix for the group of genes analyzed for both KO groups and for the ones analyzable for one group only. Incorrect links were deleted from the result, e.g. abstracts containing the names of both gene and process but not in relation to each other.

Real-time PCR

Real-time quantitative PCR (TaqMan®, Applied Biosystems) was performed using the SYBR® Green PCR Core Reagents (Applied Biosystems). Primers were designed using Primer Express® software (Applied Biosystems) based on GenBank sequences. A dissociation curve was created to check for non-specific amplification. In addition, serial dilutions from the cDNA were analyzed for each target gene to assess PCR efficiency. Transcription values were related to levels of 18S rRNA. The results were analyzed using a Gaussian linear regression, similar to the microarray analysis, with the adjustment of adding 18S rRNA as a reference gene. The AIC values of several models were used to assess differences, if any, between the controls and one or both groups of KO mice.

A more detailed description of the methods is presented in Supplementary Material 1.

Results and Discussion

Induction of hypertrophy

It has been shown previously that the heterozygous KO mice partially lacking cMyBP-C do not show LV hypertrophy up to the age of eleven months [12]. We confirmed that the LVW/BW ratio did not change in the heterozygous KO vs. WT (4.32 ± 0.86 vs. 4.34 ± 0.43 mg/g, mean $\pm$ standard deviation). In contrast, the LVW/BW ratio increased by 48% in the homozygous KO vs. WT mice (6.44 ± 0.46 vs. 4.34 ± 0.43 mg/g, $p < 0.0001$, Student t-test). This indicates that the cMyBP-C homozygotes already developed a prominent degree of hypertrophy at nine weeks of age. In fact these mice have reached their maximal level by that age [25].

From probes to differential genes

Because of the flagging of some spots on our slides, the number of analyzable samples - and as such degrees of freedom - might differ for each gene. Therefore, the maximum number of covariates possible was used. A list of 8188 probes could be analyzed, of which 870 were differentially expressed in at least one of the two KO groups. LocusLink Identifiers (LL IDs) were retrieved and the probes with known LL IDs were retained, of which a few corresponded to the same gene. About 16% of the genes were represented by more than one probe. Of these, 87% showed similar expression values, i.e. the fold change direction remained the same within a small range. The remaining genes showed some of their probes differentially expressed and others not, or upregulation of some probes while others were downregulated. Therefore these were discarded from further analysis. Due to this conservative approach, some differentially expressed genes may have been discarded, but the ones retrieved were more probable to be true differentially expressed genes. Also, genes for which the intensities were near the saturation value for all groups were removed from further analysis, because possible changes in expression may be obscured. These pre-processing steps ensured good quality of the expression data for the genes selected for further analysis. In addition, the variation in the expression of eight spikes was low both within slides and between slides. In previous experiments using the same slides and cDNA collection [20], replicate hybridizations were performed for 5 samples and these results showed that between slide variability (when possible adjusting for labeling efficiency and spikes in the model) was close to zero. Figure 1 gives an overview of the number of genes after each step and in each group for the complete procedure. The fact that more down- than upregulated genes could only be analyzed for one of the two KO groups, can be explained by flagging of negative spots, as genes that were in fact downregulated in the other group as well, may have been below detection limit. Overall, the numbers of downregulated genes was higher than the number of upregulated genes in heterozygotes (145 and 90) and almost equal in homozygotes (122 and 129). The total number of differentially expressed genes was slightly higher in the homozygous group. The fold changes

were relatively moderate in the present data set. Only fourteen genes had a fold change below 0.33 and eight above 3. For all differential genes and their respective fold changes, see Supplementary Data Table 1.

Expression differences of 27 genes sampled over the most prominent differential processes, were verified by real-time PCR (Supplementary Data Table 2) with 18S rRNA as reference gene. The results were comparable with respect to the direction of the changes (with the exception of *Ikbke* and *F11r*), although the levels often differed and a differential gene in one type of analysis could not always be retrieved for the other. This was not unexpected as the number of samples and the expression differences were relatively small. Both methods are not the same with respect to the reference to which the gene expression levels are compared (all genes on the array vs the 18S rRNA gene in the RT-PCR). RT-PCR is highly dependent on the choice of the reference gene, which is less of an issue for array data. Obviously, this will be less of a problem if gene expression differences are bigger. However, overall results per biological process or pathway were the same and based on a substantial number of genes per process. Moreover, other investigators also confirmed array results for partly the same clone set and microarray platform with real-time PCR [26-28].

Overview of processes present

GenMAPP results provided a good first indication of the most significant processes (Local MAPPs and GO IDs) in the set of genes with reliable differential expression between WT and KO mice, containing processes or pathways representing translation, cell replication, Wnt- and G-protein signaling, cell structure and cell-environment interaction, RNA handling, protein ubiquitination and folding, energy conversion, oxygen transport, GTPases, and calcium handling (Supplementary Data Table 3). For both KO groups, the significant processes were similar. In some cases, the number of differentially expressed genes per process or pathway was relatively small and these were discarded if this number was too low. GO annotations were completed with information from Nucleotide, MeSH and OMIM entries of the differentially expressed genes. Also, literature links between these genes and the processes retrieved so far were obtained by using PubMatrix, with as keywords representing the processes: 'apoptosis', 'myosin' and 'actin', 'mitosis', 'mitochondrion', 'serine threonine kinase', 'ubiquitin', and 'zinc finger' and 'ring finger'. The ones manually checked to be correct were merged with the other annotations. Pubmatrix missed some genes of which the name was not recognized. Checking the NCBI website manually provided the most functional information, but this could only be achieved because of the moderate number of differential genes. Finally all genes that could only be analyzed for one of the two KO groups were also annotated using PubMatrix. Based on all annotations, the genes in the most prominent processes or pathways are presented in Tables 1 to 5, together with their descriptions. In Supplementary Data Table 4 an overview of genes and their relationships to these processes is provided.

Energy conversion pathways (Table 1)

The present findings indicated an upregulation of genes encoding proteins involved in mitochondrial fatty acid oxidation, including *Acads* and *Echs1*, as an early response in both homozygous and heterozygous hearts. This finding underscored the importance of fatty acid oxidation in cardiac energy conversion [29] and most likely reflected a functional adaptation to an increase in energy demand of the cMyBP-C KO hearts. This notion was supported by the enhanced expression of *Ndufc2* and *Suc1g2*, encoding enzymes involved in the mitochondrial respiratory chain and citric acid cycle, respectively. In the homozygous KO mice, increased expression of *Ckmt1* (involved in energy transfer from mitochondrial ATP to cytoplasmic phosphocreatine for fast generation of energy) and *Aldoc* (encoding the glycolytic enzyme aldolase) strongly suggested a further increase in energy demand. This increase is most likely caused by a decline in efficiency of the cardiac contractile machinery due to the (partial) lack of cMyBP-C. The initial upregulation of fatty acid handling genes in the present model of cardiomyopathy is at variance with observations in end-stage failing hypertrophied hearts mostly showing a downregulation of fatty acid-metabolizing enzymes [30-32].

Table 1 Differential genes in “mitochondrial / energy conversion”; HET = heterozygotes, HOM = homozygotes, nc = no change; a dash indicates that the fold change could not be reliably computed for that group.

MITOCHONDRIAL / ENERGY CONVERSION			
Mouse Gene Name	Description	HET	HOM
D10Etd214e	DNA segment, Chr 10, ERATO Doi 214, expressed	down	
Mrpl19	mitochondrial ribosomal protein L19	down	
Acads	acyl-Coenzyme A dehydrogenase, short chain	up	
Echs1	enoyl Coenzyme A hydratase, short chain, 1, mitochondrial	up	
Ndufc2	NADH dehydrogenase (ubiquinone) 1, subcomplex unknown, 2	up	
Mrpl15	mitochondrial ribosomal protein L15	up	
Suc1g2	succinate-Coenzyme A ligase, GDP-forming, beta subunit	up	
Timm22	translocase of inner mitochondrial membrane 22 homolog (yeast)	-	down
Txn2	thioredoxin 2	-	down
Afg3l2	AFG3(ATPase family gene 3)-like 2 (yeast), paraplegin-related	-	up
Alas2	aminolevulinic acid synthase 2, erythroid	down	-
Ndufs3	NADH dehydrogenase (ubiquinone) Fe-S protein 3	down	-
Ppargc1a	peroxisome prolifer. active. receptor, gamma, coactivator 1 alpha	down	-
Mrpl52	mitochondrial ribosomal protein L52	up	-
Pdhx	pyruvate dehydrogenase complex, component X	up	-
Aldoc	aldolase 3, C isoform	nc	up
Ckmt1	creatine kinase, mitochondrial 1, ubiquitous	nc	up
Perp	PERP, TP53 apoptosis effector	nc	up
Btf3	basic transcription factor 3	nc	up
Acad8	acyl-Coenzyme A dehydrogenase family, member 8	up	nc
1810063B05Rik	RIKEN cDNA 1810063B05 gene	down	nc
Uqcrrh	ubiquinol-cytochrome c reductase hinge protein	up	up

Contractile, structural and calcium handling proteins (Table 2)

Apart from changes in energy metabolism to compensate for the (partial) lack of cMyBP-C, also alterations in contractile, structural and calcium related proteins occur. Genes related to integrins and MLP (muscle LIM protein) are associated with the Z-disc and interact with the extracellular matrix, thereby acting as stress sensors in the hypertrophic response [33, 34]. The integrin-mediated LIM domain-containing gene Fhl1 that interacts with cMyBP-C [35] was found to be upregulated in the homozygous KO mice. Myh10, a sarcomere-related gene was also upregulated, showing a more than three-fold change. Due to cross hybridization that may occur using cDNA arrays, the apparent upregulation of this MHC gene could be caused by upregulation of other myosins, most likely beta-myosin. It is known from haemodynamic overload studies in rodents that alpha-MHC switches to beta-MHC [33]. The fact that the expression of only a few contractile and sarcomeric genes was altered and that the changes were relatively small, may be due to the fact that hypertrophy has already reached its maximum at 9 weeks of age in the homozygous KO mice [25] and that the heterozygous mice have not developed hypertrophy yet at this age. In line with this, Fhl1 and Myh10 showed no change in gene expression in heterozygous mice, but analysis of younger 6-weeks old homozygous cMyBP-C KO mice, which were in the phase of developing hypertrophy, showed a 40-fold increase in beta-MHC protein vs. WT (unpublished data). Several other cell structure related genes were clearly differentially expressed in both hetero- and homozygous KO mice, including genes encoding proteins related to microtubules or involved in cell-cell or cell-extracellular matrix interaction. Also, the expression of a number of Ca^{2+} -binding proteins was found to be affected, both in homozygous as in heterozygous hearts. Since Ca^{2+} is involved in both the contractile activity of the cardiac muscle cells and in the hypertrophic response, it is likely that adaptations in Ca^{2+} handling in hearts lacking cMyBP-C contribute to the altered cardiac phenotype [36].

Protein synthesis and degradation (Table 3)

A number of zinc finger proteins were found to be up- as well as downregulated. Since these proteins are involved in the regulation of gene expression and protein synthesis, the present findings indicate alterations in the rate of protein synthesis in cMyBP-C KO hearts, which has been reported before in cardiac hypertrophy [33, 37]. This notion is also supported by the fact that several RNA processing factors were found to be upregulated. Several genes encoding proteins involved in protein ubiquitination also showed differential expression, in most cases an upregulation, strongly suggesting an increase in protein turnover in the affected heart. These changes were present in both homo- and heterozygotes, albeit to a lesser extent in the latter, e.g. changes in expression of ubiquitination genes could be demonstrated in both directions. This may be explained by the fact that increased production of proteins may lead to more erroneous and misfolded ones, which have to be cleared by the proteasome system [38].

Table 2 Differential genes in “cell structure”; see Table 1 for abbreviations.

CELL STRUCTURE			
Mouse Gene Name	Description	HET	HOM
lfrd1	interferon-related developmental regulator 1	down	
Fndc5	fibronectin type III domain containing 5	down	
Mapt	microtubule-associated protein tau	down	
BC021523	cDNA sequence BC021523	down	
Actl6a	actin-like 6A	down	
Foxo3	forkhead box O3	down	
Abi2	abl-interactor 2	up	
P42pop	Myb protein P42POP	up	
Mastl	microtubule associated serine/threonine kinase-like	-	down
5730409F24Rik	RIKEN cDNA 5730409F24 gene	-	down
Bbs4	Bardet-Biedl syndrome 4 homolog (human)	-	down
Waspip	Wiskott-Aldrich syndrome protein interacting protein	-	down
Spg4	spastic paraplegia 4 homolog (human)	-	up
Spred1	sprouty protein with EVH-1 domain 1, related sequence	down	-
Smoc1	SPARC related modular calcium binding 1	up	-
Pdlim1	PDZ and LIM domain 1 (elfin)	up	-
Large	like-glycosyltransferase	up	-
Lgals4	lectin, galactose binding, soluble 4	up	down
Cugbp2	CUG triplet repeat,RNA binding protein 2	down	up
Abi1	abl-interactor 1	nc	up
Fhl1	four and a half LIM domains 1	nc	up
Smtn	Smoothelin	nc	up
D2Ert435e	DNA segment, Chr 2, ERATO Doi 435, expressed	nc	up
Actn4	actinin alpha 4	nc	up
1200009I06Rik	RIKEN cDNA 1200009I06 gene	nc	up
Myh10	myosin heavy chain 10, non-muscle	nc	up
Stk17b	apoptosis-inducing serine/threonine kinase 17b	nc	up
Bag2	Bcl2-associated athanogene 2	nc	up
D19Ert4144e	DNA segment, Chr 19, ERATO Doi 144, expressed	up	nc
Fmr1	fragile X mental retardation syndrome 1 homolog	down	nc
F11r	F11 receptor	down	nc
Smarce1	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, e, 1	down	nc
2310057H16Rik	RIKEN cDNA 2310057H16 gene	down	nc
Arhgap21	Rho GTPase activating protein 21	down	nc
Lad1	Ladinin	down	down

MAP kinase signaling (Table 4)

Several genes related to MAPK signaling, including the key gene MAPK9, were upregulated in the pre-hypertrophic (HET) and early hypertrophic (HOM) stages. Upregulation of MAPK signaling, that consists of the JNK, p38 and ERK subsystems [39], is known to be involved in cardiac hypertrophy. In both homozygotes and heterozygotes, the JNK and p38 pathways were upregulated whereas the ERK pathway was downregulated. These changes were less pronounced in heterozygotes than in homozygotes. Overexpression of the ERK1/2 subsystem is thought to be related to physiological hypertrophy, whereas overexpression of the JNKs and p38 subsystems is associated with pathological or maladaptive hypertrophy [36, 37]. Our findings strongly suggest that the overall genetic adaptation in hearts (partially) lacking cMyBP-C is maladaptive in nature and is established before hypertrophy occurs.

Table 3 Differential genes in “zinc finger protein” and “ubiquitin”; see Table 1 for abbreviations.

ZINC FINGER PROTEIN			
Mouse Gene Name	Description	HET	HOM
Zc3hdc5	zinc finger CCCH type domain containing 5	down	
Zfp281	zinc finger protein 281	down	
Zfp51	zinc finger protein 51	up	
Rad18	RAD18 homolog (S. cerevisiae)	up	
Rnf139	ring finger protein 139	up	
D5Ert689e	DNA segment, Chr 5, ERATO Doi 689, expressed	-	down
5430400N05Rik	RIKEN cDNA 5430400N05 gene	-	down
Rnf41	ring finger protein 41	-	down
Zfp334	zinc finger protein 334	-	down
Rnf35	ring finger protein 35	-	down
S100a6	S100 calcium binding protein A6 (calcyclin)	-	down
Akap8	A kinase (PRKA) anchor protein 8	-	down
Prdm5	PR domain containing 5	-	down
Zfp503	zinc finger protein 503	-	up
Zzz3	zinc finger, ZZ domain containing 3	down	-
Zfp367	zinc finger protein 367	down	-
Pclo	piccolo (presynaptic cytomatrix protein)	down	-
Zdhhc5	zinc finger, DHHC domain containing 5	up	-
Pdlim1	PDZ and LIM domain 1 (elfin)	up	-
Plagl1	pleiomorphic adenoma gene-like 1	down	up
Aebp2	AE binding protein 2	nc	up
Fhl1	four and a half LIM domains 1	nc	up
Rnf24	ring finger protein 24	nc	up
Trim25	tripartite motif protein 25	nc	up
Rpl37	ribosomal protein L37	up	nc
Zfp219	zinc finger protein 219	nc	down
BB114266	expressed sequence BB114266	down	nc
Phf13	PHD finger protein 13	up	up
UBIQUITIN			
Mouse Gene Name	Description	HET	HOM
Zc3hdc5	zinc finger CCCH type domain containing 5	down	
Ube2j2	ubiquitin-conjugating enzyme E2, J2 homolog (yeast)	up	
Ube2r2	ubiquitin-conjugating enzyme E2R 2	up	
Rnf139	ring finger protein 139	up	
Gga3	golgi assoc.,gamma adaptin ear containing,ARF binding protein 3	-	down
Btrc	beta-transducin repeat containing protein	-	down
S100a6	S100 calcium binding protein A6 (calcyclin)	-	down
Vps28	vacuolar protein sorting 28 (yeast)	-	up
Usp38	ubiquitin specific protease 38	-	up
Tceb1	transcription elongation factor B (SIII), polypeptide 1	-	up
Gga2	golgi assoc.,gamma adaptin ear containing,ARF binding protein 2	down	-
Sec63	SEC63-like (S. cerevisiae)	up	-
Birc6	baculoviral IAP repeat-containing 6	down	up
Rnf24	ring finger protein 24	nc	up
Trim25	tripartite motif protein 25	nc	up
Asb6	ankyrin repeat and SOCS box-containing protein 6	down	nc
Mapk9	JNK, mitogen activated protein kinase 9	up	up

Table 4 Differential genes in “mitosis / map kinases”; see Table 1 for abbreviations.

MITOSIS / MAP KINASES			
Mouse Gene Name	Description	HET	HOM
Igtp	interferon-gamma-inducible GTPase	down	
Mphosph9	M-phase phosphoprotein 9	down	
Pbk	PDZ binding kinase	up	
Rad18	RAD18 homolog (S. cerevisiae)	up	
Cdk2	cyclin-dependent kinase 2	up	
C430014M02Rik	RIKEN cDNA C430014M02 gene	up	
Btrc	beta-transducin repeat containing protein	-	down
Gap43	growth associated protein 43	-	down
Gnai3	guanine nucleotide binding protein, alpha inhibiting 3	-	down
Tacc3	transforming, acidic coiled-coil containing protein 3	-	down
Ikbke	inhibitor of kappaB kinase epsilon	-	down
Kit	c-Kit oncogene	-	up
Slk	STE20-like kinase	-	up
Mad2l1bp	MAD2L1 binding protein	-	up
Ets1	E26 avian leukemia oncogene 1, 5' domain	-	up
Ccnb1	cyclin B1	down	-
Sf3b1	splicing factor 3b, subunit 1	down	-
Runx2	runt related transcription factor 2	down	-
Fyn	Fyn proto-oncogene	down	-
Mapk8ip3	mitogen-activated protein kinase 8 interacting protein 3	down	-
Spred1	sprouty protein with EVH-1 domain 1, related sequence	down	-
Cdc37	cell division cycle 37 homolog (S. cerevisiae)	nc	up
Fhl1	four and a half LIM domains 1	nc	up
Tbrg4	transforming growth factor beta regulated gene	nc	up
Skb1	SKB1 homolog (S. pombe)	nc	up
Bag2	Bcl2-associated athanogene 2	nc	up
Trim25	tripartite motif protein 25	nc	up
Ets2	E26 avian leukemia oncogene 2, 3' domain	up	nc
D19Ert144e	DNA segment, Chr 19, ERATO Doi 144, expressed	up	nc
F11r	F11 receptor	down	nc
Mapk9	JNK, mitogen activated protein kinase 9	up	up
Xpo1	exportin 1, CRM1 homolog (yeast)	up	up
Cse1L	CAS, chromosome segregation 1-like	up	up
Trpc4ap	transient receptor potential cation channel - subfamily C - member 4 associated protein	up	up

Apoptosis (Table 5)

One of the processes associated with the JNK and p38 pathways is apoptosis. Both in homozygous and heterozygous hearts, the expression of genes encoding proteins involved in one of the major apoptotic pathways, i.e. the release of cytochrome C from mitochondria and the increase in caspase activity, was enhanced. In addition, a number of other pro-apoptotic genes, especially in the homozygous hearts, were upregulated and anti-apoptotic genes were found to be downregulated. However, a number of genes encoding proteins with anti-apoptotic properties were found to be upregulated, like Txndc5, known to be involved in the protection of endothelial cells against apoptosis under hypoxic conditions. Maybe spreading of endothelial cells, being part of the new formation of blood vessels known to occur in the hypertrophying myocardium [33], is facilitated due to the protective properties of Txndc5 proteins. The current results show a tendency towards apoptosis in hearts of cMyBP-C mice, which has been recently suggested to be a hallmark of maladaptive hypertrophy [40].

Table 5 Differential genes in “apoptosis”; see Table 1 for abbreviations.

APOPTOSIS			
Mouse Gene Name	Description	HET	HOM
Igtp	interferon-gamma-inducible GTPase	down	
Foxo3	forkhead box O3	down	
Txndc5	thioredoxin domain containing 5	up	
Wdr3	WD repeat	-	down
S100a6	S100 calcium binding protein A6 (calcyclin)	-	down
Tacc3	transforming, acidic coiled-coil containing protein 3	-	down
Txn2	thioredoxin 2	-	down
Prdm5	PR domain containing 5	-	down
Kit	c-Kit oncogene	-	up
Slk	STE20-like kinase	-	up
Ets1	E26 avian leukemia oncogene 1, 5' domain	-	up
Runx2	runt related transcription factor 2	down	-
Fyn	Fyn proto-oncogene	down	-
Rab6ip2	Rab6 interacting protein 2	down	-
Ctnnb1	catenin, beta like 1	up	-
Pacsin2	protein kinase C and casein kinase substrate in neurons 2	up	-
Birc6	baculoviral IAP repeat-containing 6	down	up
Cugbp2	ETR-3, CUG triplet repeat, RNA binding protein 2	down	up
Plagl1	pleiomorphic adenoma gene-like 1	down	up
Cdc37	cell division cycle 37 homolog (S. cerevisiae)	nc	up
Ppic	peptidylprolyl isomerase C	nc	up
Perp	PERP, TP53 apoptosis effector	nc	up
Stk17b	apoptosis-inducing serine/threonine kinase 17b	nc	up
Bag2	Bcl2-associated athanogene 2	nc	up
Btf3	basic transcription factor 3	nc	up
Ets2	E26 avian leukemia oncogene 2, 3' domain	up	nc
Anxa5	annexin A5	down	nc
Fmr1	fragile X mental retardation syndrome 1 homolog	down	nc
Mapk9	JNK, mitogen activated protein kinase 9	up	up
Uqcrrh	ubiquinol-cytochrome c reductase hinge protein	up	up
Cse1l	CAS, chromosome segregation 1-like	up	up

Other processes

Upregulation of mitosis-related genes such as cell cycle progression genes and several checkpoint and control genes was observed in both KO groups, especially in homozygotes, indicating enhanced cell division activity. This most likely indicates proliferation of cell types other than cardiomyocytes, such as endothelial cells or fibroblasts, probably responsible for the observed myocardial fibrosis, which could be considered a hallmark of maladaptive hypertrophy and cardiac failure [37].

Conclusion

In this study, both homozygous and heterozygous cMyBP-C hearts were investigated. Homozygous animals displayed maximal cardiac hypertrophy ($\approx 50\%$ increase in mass), whereas overall signs of hypertrophy were absent in the heterozygous animals at the age of nine weeks. Despite the striking difference in phenotype, the number of differentially expressed cardiac genes did not differ substantially between hetero- and homozygous animals, although the direction of the alterations did show some remarkable differences just as the specific genes changed (see Supplementary Data Table 1). The (partial) lack of the

cMyBP-C protein lead to an increased energy demand, to activation of the JNK and p38 subsystems of the MAPK signaling pathway and deactivation of ERK 1/2 signaling and to induction of apoptotic processes. This strongly suggests that a pathological pathway is already initiated at an early phase in the cMyBP-C KO mice, even before hypertrophy occurs (heterozygotes), and persists after hypertrophy has been completed (homozygotes). Changes in cardiac structure and contractile proteins and in Ca^{2+} handling genes were probably less prominent in our data set, as hypertrophy was either not yet present (heterozygotes) or already completed (homozygotes). As changes in cardiac structure require the formation and degradation of an adapted spectrum of proteins, altered protein turnover was also observed in the hearts. Many of the changes were more pronounced in the homozygous KO mice, which may explain why these mice develop hypertrophy more quickly and eventually demonstrate a more severe phenotype. Findings in cMyBP-C KO mice are also relevant for patients with FHC, as these show similar phenotypes as the heterozygous mice. Clinical studies have shown that many, but not all, patients with FHC due to cMyBP-C mutations have a relatively benign prognosis [41, 42]. Moreover, it has been shown that individuals with two mutations (homozygotes, compound heterozygotes or double-heterozygotes) develop a more severe form of LVH, often associated with arrhythmias and a bad prognosis [6]. It would be worthwhile to study the processes identified in both mice models in these patients to investigate if the same processes are involved in humans and if these relate to the pathology observed and the severity of clinical manifestation.

Acknowledgements

We would like to thank the Leids Universitair Medisch Centrum (LUMC) for the distribution of the NIA 15k set. Jeroen Aerssens, Rob Janssen en Erika Timmer are acknowledged for their assistance with the microarray experiments. This research was supported by the Netherlands Heart Foundation (99.122) and the Cardiovascular Research Institute Maastricht (CARIM), The Netherlands, and by the European Commission (MEXT FUGEN-FHC #014051).

References

- 1 Bashyam MD, Savithri GR, Kumar MS, Narasimhan C, Nallari P: **Molecular genetics of familial hypertrophic cardiomyopathy (FHC).** *J Hum Genet* 2003, **48**:55-64.
- 2 Carrier L, Bonne G, Barend E, Yu B, Richard P, Niel F, Hainque B, Cruaud C, Gary F, Labeit S, Bouhour JB, Dubourg O, Desnos M, Hagege AA, Trent RJ, Komajda M, Fiszman M, Schwartz K: **Organization and sequence of human cardiac myosin binding protein C gene (MYBPC3) and identification of mutations predicted to produce truncated proteins in familial hypertrophic cardiomyopathy.** *Circ Res* 1997, **80**:427-434.
- 3 Niimura H, Bachinski LL, Sangwatanaroj S, Watkins H, Chudley AE, McKenna W, Kristinsson A, Roberts R, Sole M, Maron BJ, Seidman JG, Seidman CE: **Mutations in the gene for cardiac myosin-binding protein C and late-onset familial hypertrophic cardiomyopathy.** *N Engl J Med* 1998, **338**:1248-1257.
- 4 Bonne G, Carrier L, Richard P, Hainque B, Schwartz K: **Familial hypertrophic cardiomyopathy: from mutations to functional defects.** *Circ Res* 1998, **83**:580-593.
- 5 Alders M, Jongbloed R, Deelen W, van den Wijngaard A, Doevendans P, Ten Cate F, Regitz-Zagrosek V, Vosberg HP, van Langen I, Wilde A, Dooijes D, Mannens M: **The 2373insG mutation in the MYBPC3 gene is a founder mutation, which accounts for nearly one-fourth of the HCM cases in the Netherlands.** *Eur Heart J* 2003, **24**:1848-1853.

- 6 Richard P, Charron P, Carrier L, Ledeuil C, Cheav T, Pichereau C, Benaiche A, Isnard R, Dubourg O, Burban M, Gueffet JP, Millaire A, Desnos M, Schwartz K, Hainque B, Komajda M: **Hypertrophic cardiomyopathy: distribution of disease genes, spectrum of mutations, and implications for a molecular diagnosis strategy.** *Circulation* 2003, **107**:2227-2232.
- 7 Winegrad S: **Myosin binding protein C, a potential regulator of cardiac contractility.** *Circ Res* 2000, **86**:6-7.
- 8 Flashman E, Redwood C, Moolman-Smook J, Watkins H: **Cardiac myosin binding protein C: its role in physiology and disease.** *Circ Res* 2004, **94**:1279-1289.
- 9 Rottbauer W, Gautel M, Zehelein J, Labeit S, Franz WM, Fischer C, Vollrath B, Mall G, Dietz R, Kubler W, Katus HA: **Novel splice donor site mutation in the cardiac myosin-binding protein-C gene in familial hypertrophic cardiomyopathy. Characterization Of cardiac transcript and protein.** *J Clin Invest* 1997, **100**:475-482.
- 10 Moolman JA, Reith S, Uhl K, Bailey S, Gautel M, Jeschke B, Fischer C, Ochs J, McKenna WJ, Klues H, Vosberg HP: **A newly created splice donor site in exon 25 of the MyBP-C gene is responsible for inherited hypertrophic cardiomyopathy with incomplete disease penetrance.** *Circulation* 2000, **101**:1396-1402.
- 11 Vignier N, Perrot A, Schulte HD, Richard P, Sebillon P, Schwartz K, Osterziel KJ, Carrier L: **Cardiac myosin-binding protein C and familial hypertrophic cardiomyopathy: from mutations identification to human endomyocardial proteins analysis (Abstract), 74th Scientific Sessions of the American Heart Association, 11-14 November 2001, Anaheim, CA, USA.** *Circulation* 2001, **104**, (suppl.):II-1.
- 12 Carrier L, Knoll R, Vignier N, Keller DI, Bausero P, Prudhon B, Isnard R, Ambroisine ML, Fiszman M, Ross J, Jr., Schwartz K, Chien KR: **Asymmetric septal hypertrophy in heterozygous cMyBP-C null mice.** *Cardiovasc Res* 2004, **63**:293-304.
- 13 Tanaka TS, Jaradat SA, Lim MK, Kargul GJ, Wang X, Grahovac MJ, Pantano S, Sano Y, Piao Y, Nagaraja R, Doi H, Wood WH, 3rd, Becker KG, Ko MS: **Genome-wide expression profiling of mid-gestation placenta and embryo using a 15,000 mouse developmental cDNA microarray.** *Proc Natl Acad Sci U S A* 2000, **97**:9127-9132.
- 14 Kargul GJ, Dudekula DB, Qian Y, Lim MK, Jaradat SA, Tanaka TS, Carter MG, Ko MS: **Verification and initial annotation of the NIA mouse 15K cDNA clone set.** *Nat Genet* 2001, **28**:17-18.
- 15 Dahlquist KD, Salomonis N, Vranizan K, Lawlor SC, Conklin BR: **GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways.** *Nat Genet* 2002, **31**:19-20.
- 16 Doniger SW, Salomonis N, Dahlquist KD, Vranizan K, Lawlor SC, Conklin BR: **MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data.** *Genome Biol* 2003, **4**:R7.
- 17 NCBI website [<http://www.ncbi.nlm.nih.gov/>]
- 18 Becker KG, Hosack DA, Dennis G, Jr., Lempicki RA, Bright TJ, Cheadle C, Engel J: **PubMatrix: a tool for multiplex literature mining.** *BMC Bioinformatics* 2003, **4**:61.
- 19 Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, et al: **Minimum information about a microarray experiment (MIAME)-toward standards for microarray data.** *Nat Genet* 2001, **29**:365-371.
- 20 van den Bosch BJC, Lindsey PJ, van den Burg CMM, van der Vlies SA, Lips DJ, van der Vusse GJ, Ayoubi TA, Doevendans PA, Smeets HJM: **Early and transient gene expression changes in pressure overload-induced cardiac hypertrophy in mice.** *Genomics* 2006:In press.
- 21 Fare TL, Coffey EM, Dai H, He YD, Kessler DA, Kilian KA, Koch JE, LeProust E, Marton MJ, Meyer MR, Stoughton RB, Tokiwa GY, Wang Y: **Effects of atmospheric ozone on microarray data quality.** *Anal Chem* 2003, **75**:4672-4675.
- 22 Hegde P, Qi R, Abernathy K, Gay C, Dharap S, Gaspard R, Hughes JE, Sniesrud E, Lee N, Quackenbush J: **A concise guide to cDNA microarray analysis.** *Biotechniques* 2000, **29**:548-550, 552-544, 556 passim.
- 23 Akaike H: **Information theory and an extension of the maximum likelihood principle.** In *Second International Symposium on Inference Theory; Budapest: Akadémiai Kiadó.* Edited by Petrov BN, Csàki F. 1973: 267-281.
- 24 Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25**:25-29.

- 25 Cazorla O, Szilagyi S, Vignier N, Salazar G, Kramer E, Vassort G, Carrier L, Lacampagne A: **Length and protein kinase A modulations of myocytes in cardiac myosin binding protein C-deficient mice.** *Cardiovasc Res* 2006, **69**:370-380.
- 26 Hashimoto M, Koda M, Ino H, Yoshinaga K, Murata A, Yamazaki M, Kojima K, Chiba K, Mori C, Moriya H: **Gene expression profiling of cathepsin D, metallothioneins-1 and -2, osteopontin, and tenascin-C in a mouse spinal cord injury model by cDNA microarray analysis.** *Acta Neuropathol (Berl)* 2004.
- 27 Duque G, Macoritto M, Kremer R: **Vitamin D treatment of senescence accelerated mice (SAM-P/6) induces several regulators of stromal cell plasticity.** *Biogerontology* 2004, **5**:421-429.
- 28 Mu X, Zhao S, Pershad R, Hsieh TF, Scarpa A, Wang SW, White RA, Beremand PD, Thomas TL, Gan L, Klein WH: **Gene expression in the developing mouse retina by EST sequencing and microarray analysis.** *Nucleic Acids Res* 2001, **29**:4983-4993.
- 29 van der Vusse GJ, Glatz JF, Stam HC, Reneman RS: **Fatty acid homeostasis in the normoxic and ischemic heart.** *Physiol Rev* 1992, **72**:881-940.
- 30 Frey N, Olson EN: **Cardiac hypertrophy: the good, the bad, and the ugly.** *Annu Rev Physiol* 2003, **65**:45-79.
- 31 Stanley WC, Chandler MP: **Energy metabolism in the normal and failing heart: potential for therapeutic interventions.** *Heart Fail Rev* 2002, **7**:115-130.
- 32 van Bilsen M, Smeets PJ, Gilde AJ, van der Vusse GJ: **Metabolic remodelling of the failing heart: the cardiac burn-out syndrome?** *Cardiovasc Res* 2004, **61**:218-226.
- 33 Ritter O, Neyses L: **The molecular basis of myocardial hypertrophy and heart failure.** *Trends Mol Med* 2003, **9**:313-321.
- 34 Tarone G, Lembo G: **Molecular interplay between mechanical and humoral signalling in cardiac hypertrophy.** *Trends Mol Med* 2003, **9**:376-382.
- 35 McGrath MJ, Cottle DL, Nguyen MA, Dyson JM, Coghill ID, Robinson PA, Holdsworth M, Cowling BS, Hardeman EC, Mitchell CA, Brown S: **Four and a half LIM protein 1 binds myosin-binding protein C and regulates myosin filament formation and sarcomere assembly.** *J Biol Chem* 2006, **281**:7666-7683.
- 36 Selvetella G, Hirsch E, Notte A, Tarone G, Lembo G: **Adaptive and maladaptive hypertrophic pathways: points of convergence and divergence.** *Cardiovasc Res* 2004, **63**:373-380.
- 37 Lips DJ, deWindt LJ, van Kraaij DJ, Doevendans PA: **Molecular determinants of myocardial hypertrophy and failure: alternative pathways for beneficial and maladaptive hypertrophy.** *Eur Heart J* 2003, **24**:883-896.
- 38 Hedhli N, Pelat M, Depre C: **Protein turnover in cardiac cell growth and survival.** *Cardiovasc Res* 2005, **68**:186-196.
- 39 Garrington TP, Johnson GL: **Organization and regulation of mitogen-activated protein kinase signaling pathways.** *Curr Opin Cell Biol* 1999, **11**:211-218.
- 40 van Empel VP, De Windt LJ: **Myocyte hypertrophy and apoptosis: a balancing act.** *Cardiovasc Res* 2004, **63**:487-499.
- 41 Charron P, Dubourg O, Desnos M, Isnard R, Hagege A, Bonne G, Carrier L, Tesson F, Bouhour JB, Buzzi JC, Feingold J, Schwartz K, Komajda M: **Genotype-phenotype correlations in familial hypertrophic cardiomyopathy. A comparison between mutations in the cardiac protein-C and the beta-myosin heavy chain genes.** *Eur Heart J* 1998, **19**:139-145.
- 42 Charron P, Dubourg O, Desnos M, Bennaceur M, Carrier L, Camproux AC, Isnard R, Hagege A, Langlard JM, Bonne G, Richard P, Hainque B, Bouhour JB, Schwartz K, Komajda M: **Clinical features and prognostic implications of familial hypertrophic cardiomyopathy related to the cardiac myosin-binding protein C gene.** *Circulation* 1998, **97**:2230-2236.

Supplementary Material 1

Microarray preparation and design

Inserts of the NIA Mouse 15K cDNA Collection were PCR amplified directly from the glycerol stocks. The insert size of eight clones per 96-well plates was verified as a quality check. After purification, PCR products were resuspended at a concentration of 200 ng/ μ l in 50% DMSO and spotted in three separate runs. As the Microgrid II arrayer (Biorobotics, UK) is mounted with a 16 pin head, each slide consisted of three meta-grids, subdivided in 16 sub-grids, corresponding to the area printed by each pin. Each sub-grid has 20 rows and columns, yielding 400 spots per sub-grid, 6400 per meta-grid, and a total of 19200 spots. Eight different spikes were printed four times by each of the 16 pins in all three meta-grids. Five positive controls were spotted six times by each pin in every meta-grid. Three negative controls and DMSO were spotted four times by every pin in each of the three meta-grids. Finally, seven spots were left empty in each of the 16 sub-grids of the three meta-grids.

RNA labeling and hybridization

Aminoallyl labeling and cDNA hybridization were performed according to the protocol by Hegde et al. [22] and slightly adjusted for use in our laboratory. In short, ten micrograms of total RNA from each sample was reverse transcribed into cDNA using random primers in the presence of aminoallyl dUTP. Following purification, products were coupled to Cy3 NHS-ester (Amersham, Piscataway, NJ). Labeled cDNAs were purified and labeling efficiency was determined. Slides were prehybridized in 1% BSA (bovine serum albumin), 50% formamide, 5X SSC and 0.1% SDS for 45 min at 42°C. Slides were then washed and dried. Labeled cDNA was resuspended in 60 μ l hybridization mix, containing 50% formamide, 5X SSC, 0.1% SDS and 8 μ g mouse CotI-DNA and hybridized to the arrays at 42°C for 16 hours. After hybridization, slides were washed for 4 min at 42°C in a 1X SSC and 0.2% SDS solution, followed by a 4 min wash at room temperature in a 0.1X SSC and 0.2% SDS solution and finally washed manually for 4 min at room temperature in a 0.1X SSC solution. Slides were then dried by centrifugation for 5 min at 50 g.

Microarray data analysis

To ensure reliable results, several quality controls were built into the analysis procedure. For all time points, any probe that was flagged by the ImaGene software (BioDiscovery) was excluded from further analysis as well as any spot with higher background than actual foreground intensity. In addition, probes that had less than two observations remaining per control, heterozygous knock-out or homozygous knock-out, were excluded from further analysis. This means that only reliable results were computed, and in some cases only one of the two knock-out groups could be analyzed.

Then the genes meeting the criteria were analyzed using a Gaussian linear regression $N(\mu, \sigma^2)$ that fits a Gaussian distribution with mean μ and variance σ^2 controlling for the labeling efficiency (“cDNA”), sex (“SEX”), housekeeping genes (“HKG”, when degrees of freedom permitted). When necessary a random effect to handle the repeats was included, $MVN(\mu, \Sigma)$ where Σ is the covariance matrix

$$\begin{pmatrix} \sigma^2 + \delta & \delta & \dots & \delta \\ \delta & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \delta \\ \delta & \dots & \delta & \sigma^2 + \delta \end{pmatrix}$$

with δ as both the extra component of variance across subjects and the common covariance among responses on the same subject. The inference criterion used for comparing the models is their ability to predict the observed data, i.e. models are compared directly through their minimized minus log-likelihood. To penalize for model complexity, the number of estimated parameters is added to this value, a form of the Akaike information criterion (AIC) [23]. For each gene, a model $E(y) = \text{cDNA} + \text{SEX} + \text{HKG}_1 + \dots + \text{HKG}_n$ containing the relevant covariates mentioned above was fitted in order to obtain a reference AIC. Then other models containing a common or group-specific KO-effect (KO) were fitted ($E(y) = \text{cDNA} + \text{SEX} + \text{HKG}_1 + \dots + \text{HKG}_n + \text{KO}$). The gene under consideration was found to be differentially expressed if the AIC of any model including a KO effect decreased compared to the model not containing this effect. This resulted in several possibilities, no effect (non-differential gene), a common effect in both KO groups versus the WT mice, an effect in heterozygotes only, an effect in homozygotes only, or an effect between all groups in the analysis. The last possibility means that both homozygotes and heterozygotes were differential with respect to the WT mice, but with significantly different effect sizes.

Real-time PCR protocol

Reverse transcription into cDNA was performed in a 40 μl volume with 1.25 μg total RNA as a template. After denaturing at 65°C for 10 min, the RNA, the first strand buffer, 1.25 mM dNTPs, 200 U Superscript II (Invitrogen), RNAsin and a mixture of 0.75 μg oligo(DT) primer and 0.75 μg random primers (Invitrogen) were incubated for 1 hr at 42°C followed by 5 min at 95°C. The cDNA was amplified in duplo in 20 μl SYBR Green PCR buffer, with 4 mM MgCl_2 , 1.25 mM dNTP mix, 0.25 U Amperase®UNG, 0.625 U AmpliTaq Gold® DNA Polymerase and 1.25 pmol of the forward and reverse primer. PCR conditions were first, 50°C for 2 min, 95°C for 10 min, followed by 40 cycles of 95°C for 15 sec/60°C for 1 min. Next, samples were heated to 95°C for 15 sec, 60°C for 1 min, and then heated to 95°C in 20 min, followed by cooling to 4°C to create a dissociation curve.

Real-time PCR data analysis

Similarly to the microarray analysis, the real-time PCR results were analyzed using a Gaussian linear regression $N(\mu, \sigma^2)$ with mean μ and variance σ^2 including the reference ribosomal RNA 18s and the repeat effect. Again, the AIC was used to assess whether there was a difference between the controls and one or both of the knock-out groups (KO effect). For each gene, a model $E(y) = \text{rRNA18s}$ containing the above mentioned covariate was fitted in order to obtain a reference AIC. Then further models $E(y) = \text{rRNA18s} + \text{KO}$, containing common or specific KO effects were fitted. Hence, the gene under consideration was found to be differentially expressed in any or both KO groups if the AIC of the corresponding KO model decreased compared to the reference model.

Supplementary Data Table 1 Complete list of differential genes; HET and HOM refer to heterozygotes and homozygotes respectively.

Sheet 1: common effect HET and HOM.

locuslink id*	gene name	HET and HOM	gene description	human homologue
217331	Zc3hdc5	0.2744	zinc finger CCH type domain containing 5	Zc3hdc5
68512	1110019J04Rik	0.3691	RIKEN cDNA 1110019J04 gene	?
52637	D10Etd214e	0.3862	DNA segment, Chr 10, ERATO Doi 214, expressed	C10orf70
16145	Igtp	0.4018	interferon gamma induced GTPase	?
15982	Ifrd1	0.4700	interferon-related developmental regulator 1	Ifrd1
384061	Fndc5	0.5728	fibronectin type III domain containing 5	Fndc5
269702	Mphosph9	0.5777	M-phase phosphoprotein 9	Mphosph9
17762	Mapt	0.5977	microtubule-associated protein tau	Mapt
52427	D4Etd111e	0.6242	DNA segment, Chr 4, ERATO Doi 111, expressed	?
223752	BC021523	0.6293	cDNA sequence BC021523	KIAA0767
56456	Actl6a	0.6368	actin-like 6A	Actl6a
76894	0610027B03Rik	0.6537	RIKEN cDNA 0610027B03 gene	FLJ33979
99889	Arfp1	0.6755	ADP-ribosylation factor interacting protein 1	Arfp1
226442	Zfp281	0.7050	zinc finger protein 281	ZNF281
52243	D3Etd229e	0.7872	DNA segment, Chr 3, ERATO Doi 229, expressed	?
56284	Mrp119	0.8301	mitochondrial ribosomal protein L19	Mrp119
68756	Foxo3	0.8663	forkhead box O3	FOXO3A
11409	Acads	1.1400	acyl-Coenzyme A dehydrogenase, short chain	Acads
101313	AV348454	1.1838	expressed sequence AV348454	?
93747	Echs1	1.2089	enoyl Coenzyme A hydratase, short chain, 1, mitochondrial	Echs1
140499	Ube2j2	1.2237	ubiquitin-conjugating enzyme E2, J2 homolog (yeast)	Ube2j2
11991	Hnrpd	1.2293	heterogeneous nuclear ribonucleoprotein D	Hnrpd
68902	1190009E20Rik	1.2655	RIKEN cDNA 1190009E20 gene	?
104930	AU015738	1.2744	expressed sequence AU015738	?
22709	Zfp51	1.2911	zinc finger protein 51	?
97131	C77096	1.2976	expressed sequence C77096	?
52033	Pbk	1.3288	PDZ binding kinase	TOPK
217737	Ahsa1	1.3288	AHA1, activator of heat shock 90kDa protein ATPase homolog 1 (yeast)	Ahsa1
14707	Gng5	1.3298	guanine nucleotide binding protein (G protein), gamma 5 subunit	Gng5
338355	C430014M02Rik	1.3327	RIKEN cDNA C430014M02 gene	KIAA0674
105245	Txndc5	1.3448	thioredoxin domain containing 5	Txndc5
68695	1110033O09Rik	1.3692	RIKEN cDNA 1110033O09 gene	MGC45386
80385	Tusc2	1.4225	tumor suppressor candidate 2	Tusc2
67615	Ube2f2	1.4247	ubiquitin-conjugating enzyme E2R 2	Ube2f2
215095	Astl	1.4264	astacin-like metalloendopeptidase (M12 family)	LOC440890
58186	Rad18	1.4350	RAD18 homolog (S. cerevisiae)	Rad18
217030	Ap1gpp1	1.4452	AP1 gamma subunit binding protein 1	?
28078	D13Wsu14e	1.4496	DNA segment, Chr 13, Wayne State University 14, expressed	?
50782	Rgs11	1.4962	regulator of G-protein signaling 11	Rgs11

locuslink id*	gene name	HET and HOM	gene description	human homologue
68197	Ndufc2	1.5012	NADH dehydrogenase (ubiquinone) 1, subcomplex unknown, 2	Ndufc2
66894	Wwp2	1.5109	WW domain containing E3 ubiquitin protein ligase 2	Wwp2
12566	Cdk2	1.5296	cyclin-dependent kinase 2	Cdk2
329165	Abi2	1.5337	abi-interactor 2	Abi2
320181	E230011A21Rik	1.5811	RIKEN cDNA E230011A21 gene	FLJ35838
52469	D11Etd99e	1.5953	DNA segment, Chr 11, ERATO Doi 99, expressed	HSPC009
52594	D11Etd437e	1.7070	DNA segment, Chr 11, ERATO Doi 437, expressed	?
66899	Flp111	1.7187	FLP1 like 1 (S. cerevisiae)	Flp111
53625	B3gnt1	1.7569	UDP-GlcNAc:betaGal beta-1,3-N-acetylglucosaminyltransferase 1	B3gnt1
75841	Rnf139	1.7579	ring finger protein 139	Rnf139
232934	P42pop	1.7846	Myb protein P42POP	LOC339344
11657	Alb1	1.8107	albumin 1	Alb
228061	Agps	2.0807	alkylglycerone phosphate synthase	Agps
27395	Mpr15	2.2189	mitochondrial ribosomal protein L15	Mpr15
230767	lqcc	2.5405	lq motif containing C	?
20917	Sucig2	2.6002	succinate-Coenzyme A ligase, GDP-forming, beta subunit	Sucig2

* II 100959 replaced with 269702

Sheet 2: only data for HET.

locuslink id*	gene name	HET	HOM	gene description
110532	Adarb1	0.1906	NA	adenosine deaminase, RNA-specific, B1
232664	4921511K06Rik	0.2622	NA	RIKEN cDNA 4921511K06 gene
97268	C230066G23Rik	0.2733	NA	RIKEN cDNA C230066G23 gene
52359	D7Erd495e	0.2817	NA	DNA segment, Chr 7, ERATO Doi 495, expressed
77856	D730048J04Rik	0.3167	NA	RIKEN cDNA D730048J04 gene
26895	Cops7b	0.3773	NA	COP9 (constitutive photomorphogenic) homolog, subunit 7b (Arabidopsis thaliana)
104464	AU015099	0.3824	NA	expressed sequence AU015099
101752	AU019521	0.3881	NA	expressed sequence AU019521
97784	C76472	0.3926	NA	expressed sequence C76472
109151	Chd9	0.3940	NA	chromodomain helicase DNA binding protein 9
24083	Gtlf3b	0.3948	NA	gene trap locus F3b
67525	6330414G21Rik	0.4079	NA	RIKEN cDNA 6330414G21 gene
98711	Rdh10	0.4103	NA	retinol dehydrogenase 10 (all-trans)
243369	Scospondin	0.4307	NA	subcommissural organ spondin
16391	Isgf3g	0.4311	NA	interferon dependent positive acting transcription factor 3 gamma
66902	Mtap	0.4405	NA	methylthioadenosine phosphorylase
20441	Siat6	0.4416	NA	sialyltransferase 6 (N-acetylglucosaminide alpha 2,3-sialyltransferase)
70544	5730437N04Rik	0.4597	NA	RIKEN cDNA 5730437N04 gene
13481	Dpm2	0.4605	NA	dolichol-phosphate (beta-D) mannosyltransferase 2
66980	Zdhnc6	0.4645	NA	zinc finger, DHHC domain containing 6
97078	C79860	0.5049	NA	expressed sequence C79860
52413	D4Erd767e	0.5054	NA	DNA segment, Chr 4, ERATO Doi 767, expressed
213773	Tb13	0.5070	NA	transducin (beta)-like 3
212123	BC057552	0.5087	NA	cDNA sequence BC057552
66131	Tipin	0.5150	NA	timeless interacting protein
19221	Ptgrn	0.5250	NA	prostaglandin F2 receptor negative regulator
14360	Fyn	0.5325	NA	Fyn proto-oncogene
74685	4930451C15Rik	0.5328	NA	RIKEN cDNA 4930451C15 gene
102143	D130072O21Rik	0.5328	NA	RIKEN cDNA D130072O21 gene
14815	Nr3c1	0.5335	NA	nuclear receptor subfamily 3, group C, member 1
77252	9430038I01Rik	0.5412	NA	RIKEN cDNA 9430038I01 gene
74158	1300006C06Rik	0.5427	NA	RIKEN cDNA 1300006C06 gene
58246	Slc35b4	0.5508	NA	solute carrier family 35, member B4
18032	Nfix	0.5546	NA	nuclear factor I/X
26875	Pcdo	0.5591	NA	piccolo (presynaptic cytomatrix protein)
20911	Sixbp2	0.5626	NA	syntaxin binding protein 2
238673	Zfp367	0.5637	NA	zinc finger protein 367
208228	5330417K06Rik	0.5801	NA	RIKEN cDNA 5330417K06 gene
94109	Csmd1	0.5833	NA	CUB and Sushi multiple domains 1
73327	1700040I03Rik	0.5934	NA	RIKEN cDNA 1700040I03 gene
21942	Tnfrsf9	0.5986	NA	tumor necrosis factor receptor superfamily, member 9
68712	1110036D12Rik	0.5987	NA	RIKEN cDNA 1110036D12 gene

locuslink id*	gene name	HET	HOW	gene description
66206	1110059E24Rik	0.6015	NA	RIKEN cDNA 1110059E24 gene
11881	Arsp	0.6046	NA	arylsulfatase B
111173	Rab6p2	0.6064	NA	Rab6 interacting protein 2
70695	3830408D24Rik	0.6095	NA	RIKEN cDNA 3830408D24 gene
20591	Jard1c	0.6139	NA	jumonji, AT rich interactive domain 1C (Rbp2 like)
12393	Runx2	0.6159	NA	runt related transcription factor 2
68349	Ndufs3	0.6232	NA	NADH dehydrogenase (ubiquinone) Fe-S protein 3
66628	5730409G07Rik	0.6246	NA	RIKEN cDNA 5730409G07 gene
52711	D11Etd729e	0.6283	NA	DNA segment, Chr 11, ERA10 Doi 729, expressed
19062	Pps	0.6327	NA	putative phosphatase
98166	C87583	0.6376	NA	expressed sequence C87583
17156	Man1b	0.6389	NA	mannosidase 1, beta
103554	Psme4	0.6468	NA	proteasome (prosome, macropain) activator subunit 4
20768	Sephs2	0.6485	NA	selenophosphate synthetase 2
225861	B930037P14Rik	0.6525	NA	RIKEN cDNA B930037P14 gene
319969	A830046J09Rik	0.6574	NA	RIKEN cDNA A830046J09 gene
66404	2410001C21Rik	0.6608	NA	RIKEN cDNA 2410001C21 gene
105670	Chc1l	0.6674	NA	chromosome condensation 1-like
59092	Pebp4	0.6681	NA	poly(C) binding protein 4
108857	Ankhd1	0.6750	NA	ankyrin repeat and KH domain containing 1
30957	Mapk8ip3	0.6805	NA	mitogen-activated protein kinase 8 interacting protein 3
67669	0610007P06Rik	0.6808	NA	RIKEN cDNA 0610007P06 gene
12526	Cd8b	0.6809	NA	CD8 antigen, beta chain
72709	C1qntf6	0.6839	NA	C1q and tumor necrosis factor related protein 6
67636	4930469P12Rik	0.6901	NA	RIKEN cDNA 4930469P12 gene
72928	2900042C004Rik	0.7037	NA	RIKEN cDNA 2900042C004 gene
114715	Spred1	0.7113	NA	sprouty protein with EVH-1 domain 1, related sequence
18157	Nppa	0.7133	NA	natriuretic peptide precursor type A
70382	Kctd2	0.7198	NA	potassium channel tetramerisation domain containing 2
14211	Smc2l1	0.7212	NA	SMC2 structural maintenance of chromosomes 2-like 1 (yeast)
58523	Statp1	0.7276	NA	signal transducer and activator of transcription interacting protein 1
217830	9030617O03Rik	0.7400	NA	RIKEN cDNA 9030617O03 gene
93697	Nang2	0.7423	NA	NMDA receptor-regulated gene 2
74105	Gga2	0.7431	NA	golgi associated, gamma adaptin ear containing, ARF binding protein 2
71169	4933425L03Rik	0.7484	NA	RIKEN cDNA 4933425L03 gene
69870	2010003119Rik	0.7533	NA	RIKEN cDNA 2010003119 gene
67387	Unc50	0.7582	NA	unc-50 homolog (C. elegans)
66642	Cnnb1l	0.7589	NA	catenin, beta like 1
81898	St3b1	0.7597	NA	splicing factor 3b, subunit 1
108946	Zz23	0.7609	NA	zinc finger, ZZ domain containing 3
105590	AU017455	0.7687	NA	expressed sequence AU017455
319166	Hist1h2ae	0.7718	NA	histone 1, H2ae
94186	Sirn3	0.7723	NA	striatin, calmodulin binding protein 3
66546	2010010M04Rik	0.7789	NA	RIKEN cDNA 2010010M04 gene

locuslink id*	gene name	HET	HOM	gene description
105425	B930053N05Rik	0.7822	NA	RIKEN cDNA B930053N05 gene
13709	Eif1	0.7849	NA	E74-like factor 1
218461	Pde8b	0.7865	NA	phosphodiesterase 8B
107272	Psat1	0.7904	NA	phosphoserine aminotransferase 1
317717	Sec22l2	0.7905	NA	SEC22 vesicle trafficking protein-like 2 (S. cerevisiae)
69724	Rnaseh2a	0.7973	NA	ribonuclease H2, large subunit
76707	Clasp1	0.8012	NA	CLIP associating protein 1
73002	2900076A13Rik	0.8022	NA	RIKEN cDNA 2900076A13 gene
11656	Alas2	0.8150	NA	aminolevulinic acid synthase 2, erythroid
16442	Itp1r5	0.8164	NA	inositol 1,4,5-triphosphate receptor 5
19017	Ppar1a	0.8200	NA	peroxisome proliferative activated receptor, gamma, coactivator 1 alpha
20853	Stau1	0.8209	NA	stau1 (RNA binding protein) homolog 1 (Drosophila)
69623	2310012P17Rik	0.8262	NA	RIKEN cDNA 2310012P17 gene
52249	D5Erd301e	0.8329	NA	DNA segment, Chr 5, ERATO Doi 301, expressed
57170	Dolpp1	0.8567	NA	dolichyl pyrophosphate phosphatase 1
288697	Ccnb1	0.8618	NA	cyclin B1
216618	4931428D14Rik	0.8857	NA	RIKEN cDNA 4931428D14 gene
78334	2700084L06Rik	0.8966	NA	RIKEN cDNA 2700084L06 gene
66442	2600017H08Rik	1.0663	NA	RIKEN cDNA 2600017H08 gene
23970	Pacsln2	1.0871	NA	protein kinase C and casein kinase substrate in neurons 2
52647	D15Erd509e	1.0889	NA	DNA segment, Chr 15, ERATO Doi 509, expressed
99633	Lphn2	1.1048	NA	latrophilin 2
54132	Pdlim1	1.1698	NA	PDZ and LIM domain 1 (elfin)
100194	Al666798	1.2700	NA	expressed sequence Al666798
140740	Sec63	1.3022	NA	SEC63-like (S. cerevisiae)
27402	Pdhx	1.3509	NA	pyruvate dehydrogenase complex, component X
16563	Klf2a	1.4813	NA	kinesin family member 2A
114570	Crip3	1.4861	NA	cysteine-rich protein 3
103537	Mbtd1	1.4931	NA	mbt domain containing 1
16795	Large	1.4958	NA	like-glycosyltransferase
213819	Cast1	1.5235	NA	O-acetyltransferase
64918	Bhmt2	1.5548	NA	betaine-homocysteine methyltransferase 2
228136	Zdhhc5	1.6059	NA	zinc finger, DHHC domain containing 5
64075	Smoc1	1.6411	NA	SPARC related modular calcium binding 1
78646	2210038L17Rik	1.6867	NA	RIKEN cDNA 2210038L17 gene
68836	Mtph52	1.9128	NA	mitochondrial ribosomal protein L52
101261	AU024342	1.9823	NA	expressed sequence AU024342
69150	Snx4	2.0537	NA	sorting nexin 4
17150	Mfap2	2.1999	NA	microfibrillar-associated protein 2
17064	C1qr1	2.2817	NA	complement component 1, q subcomponent, receptor 1
330281	9330117B14	2.3526	NA	hypothetical protein 9330117B14
97369	C78806	3.0224	NA	expressed sequence C78806
51786	Cpsf2	5.6766	NA	cleavage and polyadenylation specific factor 2

* || 99830 replaced with 19221

Sheet 3: only data for HOM.

locuslink id*	gene name	HET	HOM	gene description
56489	Ikbpke	NA	0.2094	inhibitor of kappaB kinase epsilon
19895	Rpia	NA	0.2377	ribose 5-phosphate isomerase A
15468	Hmtt111	NA	0.2573	heterogeneous nuclear ribonucleoprotein methyltransferase-like 1 (<i>S. cerevisiae</i>)
99814	AU022436	NA	0.2613	expressed sequence AU022436
70408	Poi3f	NA	0.2855	polymerase (RNA) III (DNA directed) polypeptide F
56399	Akap8	NA	0.2671	A kinase (PRKA) anchor protein 8
27060	Tcig1	NA	0.2933	T-cell, immune regulator 1
51940	D1Etd705e	NA	0.2987	DNA segment, Chr 1, ERATO Doi 705, expressed
21975	Top3a	NA	0.3486	topoisomerase (DNA) III alpha
71813	2610028J07RIK	NA	0.3571	RIKEN cDNA 2610028J07 gene
107305	5730409F24RIK	NA	0.3709	RIKEN cDNA 5730409F24 gene
17178	Fxyd3	NA	0.4255	FXVD domain-containing ion transport regulator 3
105833	4933417K04RIK	NA	0.4260	RIKEN cDNA 4933417K04 gene
73805	4930406D14RIK	NA	0.4352	RIKEN cDNA 4930406D14 gene
58901	A1449595	NA	0.4403	EST A1449595
68631	Cry11	NA	0.4416	crystallin, lambda 1
102774	Bbs4	NA	0.4638	Bardet-Biedl syndrome 4 homolog (human)
269252	Gt3c4	NA	0.4639	general transcription factor IIIC, polypeptide 4
212514	D16Etd480e	NA	0.4643	DNA segment, Chr 16, ERATO Doi 480, expressed
99229	AU041474	NA	0.4771	expressed sequence AU041474
68988	1500019O16RIK	NA	0.4893	RIKEN cDNA 1500019O16 gene
109785	Pgm3	NA	0.5116	phosphoglucuronase 3
12583	Cdo1	NA	0.5164	cysteine dioxygenase 1, cytosolic
66686	Dcbid1	NA	0.5186	discoilin, CUB and LCCL domain containing 1
11775	Ap3b2	NA	0.5237	adaptor-related protein complex 3, beta 2 subunit
104836	Cbl11	NA	0.5256	Castias B-lineage lymphoma-like 1
71860	1700019F09RIK	NA	0.5419	RIKEN cDNA 1700019F09 gene
66526	2210012G02RIK	NA	0.5436	RIKEN cDNA 2210012G02 gene
243676	E330021D16RIK	NA	0.5448	RIKEN cDNA E330021D16 gene
280635	Emilin3	NA	0.5469	elastin microfibril interfacar 3
21335	Tacc3	NA	0.5567	transforming, acidic coiled-coil containing protein 3
14673	Gna12	NA	0.5613	guanine nucleotide binding protein, alpha 12
109284	C030046I01RIK	NA	0.5619	RIKEN cDNA C030046I01 gene
100087	1110001A12RIK	NA	0.5649	RIKEN cDNA 1110001A12 gene
14679	Gna13	NA	0.5703	guanine nucleotide binding protein, alpha inhibiting 3
227394	Sico4c1	NA	0.5740	solute carrier organic anion transporter family, member 4C1
51902	Rnt24	NA	0.5748	ring finger protein 24
70075	2210401J11RIK	NA	0.5790	RIKEN cDNA 2210401J11 gene
99801	AU021001	NA	0.5817	expressed sequence AU021001
52587	D13Etd476e	NA	0.5861	DNA segment, Chr 13, ERATO Doi 476, expressed
75901	Mitc1	NA	0.5933	MAD homolog 4 interacting transcription coactivator 1

locuslink id*	gene name	HET	HOM	gene description
106970	AU015140	NA	0.5966	expressed sequence AU015140
52474	D14Erttd16e	NA	0.6172	DNA segment, Chr 14, ERATO Doi 16, expressed
56551	Txn2	NA	0.6212	thioredoxin 2
66540	3110001A13Rik	NA	0.6433	RIKEN cDNA 3110001A13 gene
74068	Asz1	NA	0.6466	ankyrin repeat, SAM and basic leucine zipper domain containing 1
75104	Mmd2	NA	0.6493	monocyte to macrophage differentiation-associated 2
101637	Al616248	NA	0.6541	expressed sequence Al616248
99169	AU015228	NA	0.6589	expressed sequence AU015228
14728	Lilrb4	NA	0.6736	leukocyte immunoglobulin-like receptor, subfamily B, member 4
260296	Rnf35	NA	0.6768	ring finger protein 35
67121	Mastl	NA	0.6796	microtubule associated serine/threonine kinase-like
56322	Timm22	NA	0.6838	translocase of inner mitochondrial membrane 22 homolog (yeast)
239985	Arid1b	NA	0.6868	AT rich interactive domain 1B (Swi1 like)
215280	Wasip	NA	0.6884	Wiskott-Aldrich syndrome protein interacting protein
330010	4833412E22Rik	NA	0.6924	RIKEN cDNA 4833412E22 gene
260302	Gga3	NA	0.6956	golgi associated, gamma adaptin ear containing, ARF binding protein 3
209225	5430400N05Rik	NA	0.6976	RIKEN cDNA 5430400N05 gene
252830	Obox6	NA	0.7056	oocyte specific homeobox 6
66622	5730410119Rik	NA	0.7060	RIKEN cDNA 5730410119 gene
70561	5730420B22Rik	NA	0.7294	RIKEN cDNA 5730420B22 gene
77291	9430079G20Rik	NA	0.7301	RIKEN cDNA 9430079G20 gene
105972	AU021889	NA	0.7413	expressed sequence AU021889
229841	Cenpe	NA	0.7446	centromere protein E
17984	Ndn	NA	0.7453	necdin
52397	D5Erttd689e	NA	0.7575	DNA segment, Chr 5, ERATO Doi 689, expressed
12234	Btrc	NA	0.7593	beta-transducin repeat containing protein
51812	Mcrrs1	NA	0.7684	microspherule protein 1
218518	Marveld2	NA	0.7804	MARVEL (membrane-associating) domain containing 2
18573	Pde1a	NA	0.7820	phosphodiesterase 1A, calmodulin-dependent
269399	6720461J16Rik	NA	0.7833	RIKEN cDNA 6720461J16 gene
12289	Cacna1d	NA	0.7838	calcium channel, voltage-dependent, L type, alpha 1D subunit
70779	Prdm5	NA	0.7904	PR domain containing 5
52614	Emr4	NA	0.7958	EGF-like module containing, mucin-like, hormone receptor-like sequence 4
69894	2010107G23Rik	NA	0.8015	RIKEN cDNA 2010107G23 gene
70423	2700063A19Rik	NA	0.8096	RIKEN cDNA 2700063A19 gene
67588	Rnf41	NA	0.8146	ring finger protein 41
73336	1700036D21Rik	NA	0.8151	RIKEN cDNA 1700036D21 gene
16468	Jarid2	NA	0.8162	jumonji, AT rich interactive domain 2
20200	S100a6	NA	0.8204	S100 calcium binding protein A6 (calcyclin)
97836	C87198	NA	0.8259	expressed sequence C87198
228876	Zfp334	NA	0.8335	zinc finger protein 334
75909	4930579A11Rik	NA	0.8339	RIKEN cDNA 4930579A11 gene
28085	D11Wsu48e	NA	0.8532	DNA segment, Chr 11, Wayne State University 48, expressed

locuslink id*	gene name	HET	HOM	gene description
106336	AU022167	NA	0.8603	expressed sequence AU022167
69684	2310044P18Rik	NA	0.8683	RIKEN cDNA 2310044P18 gene
99049	D030054H19Rik	NA	0.8704	RIKEN cDNA D030054H19 gene
269470	Wdr3	NA	0.8820	WD repeat domain 3
102654	AU015680	NA	0.8841	expressed sequence AU015680
99174	AU015850	NA	0.8981	expressed sequence AU015850
14432	Gap43	NA	0.9030	growth associated protein 43
17242	Mdk	NA	0.9070	midkine
75678	1810043M15Rik	NA	0.9627	RIKEN cDNA 1810043M15 gene
52680	D13Erd787e	NA	1.0259	DNA segment, Chr 13, ERATO Doi 787, expressed
67487	Dhx40	NA	1.1266	DEAH (Asp-Glu-Ala-His) box polypeptide 40
97519	C77901	NA	1.1325	expressed sequence C77901
225339	E230022H04Rik	NA	1.1809	RIKEN cDNA E230022H04 gene
67923	Tceb1	NA	1.1970	transcription elongation factor B (SII), polypeptide 1
66591	Mad21bp	NA	1.2171	MAD2L1 binding protein
66067	0610037H22Rik	NA	1.2650	RIKEN cDNA 0610037H22 gene
76281	Tax1bp3	NA	1.2650	Tax1 (human T-cell leukemia virus type I) binding protein 3
52098	D12Erd7e	NA	1.2770	DNA segment, Chr 12, ERATO Doi 7, expressed
218820	Zfp503	NA	1.2827	zinc finger protein 503
20874	Slk	NA	1.2876	STE20-like kinase (yeast)
66914	Vps28	NA	1.3029	vacuolar protein sorting 28 (yeast)
76719	1700081L11Rik	NA	1.3049	RIKEN cDNA 1700081L11 gene
14545	Gdapl	NA	1.3077	ganglioside-induced differentiation-associated-protein 1
50850	Spq4	NA	1.3245	spastic paraplegia 4 homolog (human)
67390	4833420N02Rik	NA	1.3276	RIKEN cDNA 4833420N02 gene
55978	0610009H04Rik	NA	1.3332	RIKEN cDNA 0610009H04 gene
229521	Syt11	NA	1.3465	synaptotagmin 11
69577	2310010B21Rik	NA	1.3813	RIKEN cDNA 2310010B21 gene
22417	Wnt4	NA	1.3919	wingless-related MMTV integration site 4
72017	Nqo3a2	NA	1.4237	NAD(P)H:quinone oxidoreductase type 3, polypeptide A2
70681	3830405G04Rik	NA	1.4419	RIKEN cDNA 3830405G04 gene
19340	Rab3d	NA	1.4604	RAB3D, member RAS oncogene family
12828	Col4a3	NA	1.4679	procollagen, type IV, alpha 3
52715	D11Erd707e	NA	1.4730	DNA segment, Chr 11, ERATO Doi 707, expressed
26561	Mmp23	NA	1.5858	matrix metalloproteinase 23
69597	Afg3l2	NA	1.6203	AFG3L(ATPase family gene 3)-like 2 (yeast)
23871	Ets1	NA	1.6526	E26 avian leukemia oncogene 1, 5' domain
98758	Hnrpf	NA	1.8032	heterogeneous nuclear ribonucleoprotein F
77766	Elp4	NA	1.8866	elongation protein 4 homolog (S. cerevisiae)
16590	Kit	NA	1.9229	kit oncogene
78473	1700091G21Rik	NA	2.3353	RIKEN cDNA 1700091G21 gene
74841	Usp38	NA	2.6263	ubiquitin specific protease 38

* locuslink 75376 replaced with 330010

Sheet 4: different effect HET and HOM.

locuslink id	gene name	HET	HOM	gene description	human homologue
18685	Phtf1	0.6385	0.3848	putative homeodomain transcription factor 1	Phtf1
68523	1110019N10Rik	0.6840	0.5832	1110019N10Rik	CGI-128
52132	D7Erd462e	0.8255	0.6214	DNA segment, Chr 7, ERATO Doi 462, expressed	MGC20255
16763	Lad1	0.5679	0.6302	ladinin	Lad1
14265	Fmr1	0.3514	1	fragile X mental retardation syndrome 1 homolog	Fmr1
72323	Asb6	0.3762	1	ankyrin repeat and SOCS box-containing protein 6	Asb6
67892	1810063B05Rik	0.4628	1	RIKEN cDNA 1810063B05 gene	C1orf31
11973	Atp6v1e1	0.4714	1	ATPase, H+ transporting, V1 subunit E isoform 1	Atp6v1e1
66932	Tceb3bp1	0.5347	1	transcription elongation factor B polypeptide 3 binding protein 1	Tceb3bp1
67951	2310057H16Rik	0.5390	1	RIKEN cDNA 2310057H16 gene	MGC4083
16456	F11r	0.5938	1	F11 receptor	F11r
57376	Smorce1	0.6331	1	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily e, member 1	Smorce1
68449	1110003P22Rik	0.6463	1	RIKEN cDNA 1110003P22 gene	DKFZP434P1750
74589	4933428M03Rik	0.6515	1	RIKEN cDNA 4933428M03 gene	?
66768	4933428G09Rik	0.6599	1	RIKEN cDNA 4933428G09 gene	MGC29898
235047	BB114266	0.6768	1	expressed sequence BB114266	ZNF91
71435	Afhgap21	0.7133	1	Rho GTPase activating protein 21	Afhgap21
11776	Ap3d1	0.7647	1	adaptor-related protein complex 3, delta 1 subunit	AP3D1
11747	Anxa5	0.8296	1	annexin A5	Anxa5
52410	D8Erd769e	0.7069	1.1763	DNA segment, Chr 8, ERATO Doi 769, expressed	?
22634	Plagl1	0.5177	1.3049	pleiomorphic adenoma gene-like 1	Plagl1
14007	Cugbp2	0.4995	1.3734	CUG triplet repeat, RNA binding protein 2	Cugbp2
12211	Birc6	0.7873	1.3895	baculoviral IAP repeat-containing 6	Birc6
94230	Cpsf1	0.4659	1.4820	cleavage and polyadenylation specific factor 1	Cpsf1
80744	BC003993	1	0.2576	cDNA sequence BC003993	KIAA1604
69890	Zfp219	1	0.3218	zinc finger protein 219	ZNF219
72157	Pgm2	1	0.6374	phosphoglucosmutase 2	PGM1
52176	D8Erd124e	1	0.7388	DNA segment, Chr 8, ERATO Doi 124, expressed	?
28023	D5Wsu150e	1	0.8226	DNA segment, Chr 5, Wayne State University 150, expressed	?
51885	D2Erd435e	1	1.2027	DNA segment, Chr 2, ERATO Doi 435, expressed	?
229709	Ahcy1	1	1.3051	S-adenosylhomocysteine hydrolase-like 1	Ahcy1
12716	Ckmt1	1	1.3199	creatine kinase, mitochondrial 1, ubiquitous	Ckmt1
77044	Arid2	1	1.3373	AT rich interactive domain 2 (Arid-rfx like)	Arid2
12539	Cdc37	1	1.3581	cell division cycle 37 homolog (S. cerevisiae)	Cdc37
26611	Rcn2	1	1.3898	reticulocalbin 2	Rcn2
21379	Tbrg4	1	1.3951	transforming growth factor beta regulated gene 4	Tbrg4
29856	Smtn	1	1.4072	smoothelin	Smtn
104718	Ttc7b	1	1.4257	tetratricopeptide repeat domain 7B	Ttc7b
100764	1110008J03Rik	1	1.4373	RIKEN cDNA 1110008J03 gene	FLJ14827
11676	Aldoc	1	1.5009	aldolase 3, C isoform	Aldoc
21399	Tcea1	1	1.5046	transcription elongation factor A (SII) 1	Tcea1

locuslink id	gene name	HEI	HOM	gene description	human homologue
14199 Fhl1		1	1.5064	four and a half LIM domains 1	Fhl1
76568 1500035H01Rik		1	1.5220	RIKEN cDNA 1500035H01 gene	FLJ21827
108116 Slco3a1		1	1.5297	solute carrier organic anion transporter family, member 3a1	Slco3a1
19038 Ppic		1	1.5307	peptidylprolyl isomerase C	Ppic
74190 1200009106Rik		1	1.5515	RIKEN cDNA 1200009106 gene	C14orf73
69082 2610312B22Rik		1	1.5540	RIKEN cDNA 2610312B22 gene	LEREPO4
98267 Skt17b		1	1.6149	serine/threonine kinase 17b (apoptosis-inducing)	Skt17b
50795 Sh3bgr		1	1.6334	SH3-binding domain glutamic acid-rich protein	Sh3bgr
217069 Trim25		1	1.6376	tripartite motif protein 25	?
18220 Nucb1		1	1.6526	nucleobindin 1	Nucb1
72043 Sulf2		1	1.6796	sulfatase 2	Sulf2
15374 Hn1		1	1.7013	hematological and neurological expressed sequence 1	Hn1
11308 Abi1		1	1.7079	abl-interactor 1	Abi1
66855 D8Erd325e		1	1.7395	DNA segment, Chr 8, ERATO Doi 325, expressed	KIAA1049
218490 Btf3		1	1.7586	basic transcription factor 3	Btf3
245867 5330414D10Rik		1	1.7682	RIKEN cDNA 5330414D10 gene	C20orf36
11569 Aebp2		1	1.7793	AE binding protein 2	Aebp2
27374 Skb1		1	1.7875	SKB1 homologue (S. pombe)	Skb1
60595 Actn4		1	1.8356	actinin alpha 4	ACTN1
227399 AW555814		1	2.0252	expressed sequence AW555814	KIAA0433
103674 A1316828		1	2.0665	expressed sequence A1316828	ZCWC1
213539 Bag2		1	2.0855	Bcl2-associated athanogene 2	Bag2
241520 D430039N05Rik		1	2.1272	RIKEN cDNA D430039N05 gene	KIAA1946
64058 Perp		1	2.1770	PERP, TP53 apoptosis effector	Perp
72839 Rnf24		1	2.4486	ring finger protein 24	Rnf24
107829 Fmip		1	2.5733	Fms interacting protein	C22orf19
77579 Myh10		1	3.0892	myosin heavy chain 10, non-muscle	Myh10
70380 Mospd1		1.3546	0.7789	motile sperm domain containing 1	Mospd1
16855 Lgals4		1.1045	0.8137	lectin, galactose binding, soluble 4	Lgals4
21871 Atp6v0a2		1.0957	0.9068	ATPase, H ⁺ transporting, lysosomal V0 subunit a isoform 2	Atp6v0a2
51991 D8Erd67e		1.2438	1	DNA segment, Chr 8, ERATO Doi 67, expressed	?
52004 D19Erd144e		1.4563	1	DNA segment, Chr 19, ERATO Doi 144, expressed	DOC-1R
52075 D5Erd66e		1.5416	1	DNA segment, Chr 5, ERATO Doi 66, expressed	D5Erd66e
52542 D15Erd154e		1.6108	1	DNA segment, Chr 15, ERATO Doi 154, expressed	D15Erd154e
67281 Rpl37		1.6463	1	ribosomal protein L37	Rpl37
23872 Ets2		1.7018	1	E26 avian leukemia oncogene 2, 3' domain	Ets2
66948 Acad8		1.7683	1	acyl-Coenzyme A dehydrogenase family, member 8	Acad8
320871 B230206H07Rik		1.7913	1	RIKEN cDNA B230206H07 gene	?
66378 2310028O11Rik		2.2877	1	RIKEN cDNA 2310028O11 gene	2310028O11Rik
66817 9030409E16Rik		2.9480	1	RIKEN cDNA 9030409E16 gene	LOC124491
15476 Hs3st1		1.2863	1.1088	heparan sulfate (glucosamine) 3-O-sulfotransferase 1	Hs3st1
103573 Xpo1		1.0841	1.1404	exportin 1, CRM1 homologue (yeast)	Xpo1
210925 D14Erd231e		1.6462	1.2381	DNA segment, Chr 14, ERATO Doi 231, expressed	FLJ10871

locuslink id	gene name	HET	HOM	gene description	human homologue
110750	Frp1	1.1277	1.2585	chromosome segregation 1-like (<i>S. cerevisiae</i>)	Cse11
106906	Al450241	2.1905	1.3462	expressed sequence A1450241	?
66576	Uqcrh	1.6565	1.3771	ubiquinol-cytochrome c reductase hinge protein	Uqcrh / LOC440567
108654	4933403F05Rik	1.0558	1.4094	RIKEN cDNA 4933403F05 gene	C18orf19
140546	Pmp1p1	1.2869	1.4689	prion protein interacting protein 1	PRNPIP
56407	Trpc4ap	1.1168	1.7927	transient receptor potential cation channel, subfamily C, member 4 associated protein	Trpc4ap
26420	Mapk9	1.4101	1.8970	mitogen activated protein kinase 9	Mapk9
50797	Copb2	1.2687	2.1745	coatamer protein complex, subunit beta 2 (beta prime)	Copb2
108735	2010005O13Rik	2.3114	4.0969	RIKEN cDNA 2010005O13 gene	LOC375035
230936	Phf13	7.8721	5.0380	PHD finger protein 13	Phf13
97923	C80171	11.4967	43.6920	expressed sequence C80171	?

Chapter 3

Supplementary Data Table 2 Results from real-time PCR validation runs; nc= no change.

gene ID	process		microarray*	RT-PCR r18s	correspondence**
Nrap / 18175 / H3020G05	cytoskeleton	heterozygotes	nc	nc	yes
		homozygotes	nc	up	?
Acs11 / 14081 / H3138H04	FA metabolism	heterozygotes	nc	up	?
		homozygotes	nc	nc	yes
Eno1 / 13806 / H3027E07/8/9	glycolysis	heterozygotes	nc	nc	yes
		homozygotes	nc	up	?
Fndc5 / 384061 / H3149B06	cell structure	heterozygotes	down	down	yes
		homozygotes			yes
Foxo3 / 68756 / H3035A11	sarcomeric	heterozygotes	down	down	yes
		homozygotes			yes
Fhl1 / 14199 / H3141F12	sarcomeric	heterozygotes	nc	nc	yes
		homozygotes	up	up	yes
Ckmt1 / 12716 / H3047D01	mito energy production	heterozygotes	nc	nc	yes
		homozygotes	up	up	yes
Ikbke / 56489 / H3155H09	NF-kappaB induction	heterozygotes	/	up	?
		homozygotes	down		no
Phf13 / 230936 / H3081H09	zinc-finger process	heterozygotes	up	nc	?
		homozygotes	up	up	yes
Pgc1 / 19017 / "PGC.1"	energy production	heterozygotes	down	nc	?
		homozygotes	/	down	?
Cpsf2 / 51786 / H3030A02	RNA processing	heterozygotes	up	nc	?
		homozygotes	/	up	?
Cugbp2 / 14007 / H3059H03	contractile	heterozygotes	down	nc	?
		homozygotes	up	up	yes
Nd2 / 17717 / "ND2"	oxphos	heterozygotes	saturated	nc	?
		homozygotes	/	nc	?
Cytb / 17711 / "CYTB"	oxphos	heterozygotes	saturated	nc	?
		homozygotes	saturated	nc	?
CoxII / 17709 / "COII"	oxphos	heterozygotes	/	nc	?
		homozygotes	saturated	up	?
Nrf1 / 18181 / H3056C03	mito regulation	heterozygotes	nc	nc	yes
		homozygotes	/	nc	?
Ant1 / 11739 /	mito import A-nucleotide	heterozygotes	/	nc	?
		homozygotes	/	nc	?
Dci / 13177 / H3075D06	mito FA metabolism	heterozygotes	nc	nc	yes
		homozygotes	nc	down	?
Tpi / 21991 / H3149C10	glycolysis FA biosynthesis	heterozygotes	nc	nc	yes
		homozygotes	nc	up	?
Echs1 / 93747 / H3047D05	mito FA metabolism	heterozygotes	up	nc	?
		homozygotes		nc	?
Uqcrh / 66576 / H3013F07	mito cytoC-reductase	heterozygotes	1.6565	nc	?
		homozygotes	1.3771	nc	?
Ets2 / 23872 / H3028G09	transcription / cell cycle	heterozygotes	up	up	yes
		homozygotes	nc		?

Chapter 3

gene ID	process		microarray*	RT-PCR r18s	correspondence**
F11r / 16456 / H3121F06, H3151D05	endothelium / ERK	heterozygotes	down	up	no
		homozygotes	nc	up	?
Mapk9 / 26420 / H3065B09	jnk / apoptosis	heterozygotes	up	up	yes
		homozygotes	up	up	yes
Casp3 / 12367 /	induction apoptosis	heterozygotes	/	down	?
		homozygotes	/	up	?
Igtp / 16145 / H3157D2	gtpase anti-apoptosis	heterozygotes	down	down	yes
		homozygotes		nc	?
Pdhb / 68263 / H3107H10	glycolyse, TCA	heterozygotes	nc	up	?
		homozygotes	nc	up	?

* / indicates that for the group under consideration not enough samples were analyzable to calculate a fold change

** Between microarray and RT-PCR; ? Indicates that either the gene was only measured on one of the platforms, or the gene did not reach significance on both because of small sample and effect sizes

Supplementary Data Table 3 Output file from genMAP/MAPInder.

all processes for which at least two genes were differential and the permute P was smaller than, respectively, 1 (for local MAPs) and 0.1 (for Gene Ontologies), are shown first annotations for all differential genes are given, then separate results are given for both knock-out groups and for up- and downregulation

OVERALL										differential									
MAP Name										Number Changed		Number Measured		Number On MAPP		Z Score		PermuterP	
Mm_Translation_Factors										5		28		46		2.896		0.017	
Mm_DNA_replication										3		14		40		2.611		0.034	
Mm_Ribosomal_Proteins										7		71		80		1.671		0.108	
Mm_Wnt_Signaling										2		13		62		1.555		0.16	
Mm_G_Protein_Signaling										2		17		74		1.125		0.274	
Mm_Proteasome_Degradation																			
Mm_TGF_Beta_Signaling_Pathway										4		42		64		1.153		0.278	
Mm_Sphingomycolipid_metabolism										2		18		51		1.036		0.589	
GOID										GO Name		GO Type	Number Changed	Number Measured	Number in GO	Z Score	PermuterP		
8143										poly(A) binding		F	2	2	3	5.436	0.004		
5519										cytoskeletal regulatory protein binding		F	2	2	9	5.436	0.005		
7026										microtubule stabilization		P	2	2	7	5.436	0.005		
8213										protein amino acid alkylation		P	2	3	10	4.289	0.005		
6376										mRNA splice site selection		P	2	7	13	3.97	0.006		
3727										single-stranded RNA binding		F	2	3	8	4.289	0.007		
245										spliceosome assembly		P	3	9	15	3.328	0.015		
3723										RNA binding		F	16	137	321	2.639	0.015		
15991										ATP hydrolysis coupled proton transport		P	2	3	7	4.289	0.017		
15988										energy coupled proton transport, against electrochemical gradient		P	2	3	7	4.289	0.017		
6457										protein folding		P	6	34	109	2.724	0.024		
16021										integral to membrane		P	20	484	3419	-2.229	0.025		
15669										gas transport		C	2	4	9	3.585	0.026		
5833										hemoglobin complex		C	2	4	6	3.585	0.026		
5344										oxygen transporter activity		F	2	4	10	3.585	0.026		
15671										oxygen transport		P	2	4	16	3.585	0.026		
6396										RNA processing		P	9	67	196	2.415	0.027		
15934										large ribosomal subunit		C	4	19	32	2.641	0.029		
5834										heterotrimeric G-protein complex		C	2	4	30	3.585	0.03		
16070										RNA metabolism		P	9	69	204	2.318	0.03		
16020										membrane		C	29	639	4110	-2.18	0.036		
6397										mRNA processing		P	6	41	101	2.197	0.038		
3729										mRNA binding		F	4	21	41	2.399	0.04		
377										RNA splicing, via transesterification reactions with bulged adenosine as nucleophile		P	4	22	63	2.289	0.045		
398										nuclear mRNA splicing, via spliceosome		P	4	22	64	2.289	0.045		
375										RNA splicing, via transesterification reactions		P	4	22	63	2.289	0.045		
16071										mRNA metabolism		P	6	43	103	2.067	0.047		
5623										cell		C	91	1614	7894	-2.012	0.047		

GOID	GO Name	GO Type	Number Changed	Number Measured	Number in GO	Z Score	PermuteP
5604	basement membrane	C	3	14	50	2.323	0.051
8380	RNA splicing	P	4	23	67	2.185	0.051
5829	cytosol	C	12	110	221	2.012	0.051
3927	heterotrimeric G-protein GTPase activity	F	2	6	53	2.716	0.054
5544	calcium-dependent phospholipid binding	F	2	6	25	2.716	0.055
5874	microtubule	C	2	6	39	2.716	0.056
4840	ubiquitin conjugating enzyme activity	F	3	14	34	2.323	0.056
8639	small protein conjugating enzyme activity	F	3	14	31	2.323	0.056
3995	acyl-CoA dehydrogenase activity	F	2	7	14	2.417	0.058
3924	GTPase activity	F	6	46	189	1.883	0.062
19538	protein metabolism	P	39	476	1758	1.851	0.064
8289	lipid binding	F	4	24	115	2.086	0.069
7275	development	P	6	195	1326	-1.95	0.071
4872	receptor activity	F	3	129	1374	-1.923	0.075
226	microtubule cytoskeleton organization and biogenesis	P	2	7	27	2.417	0.077
6725	aromatic compound metabolism	P	2	8	70	2.169	0.079
16637	oxidoreductase activity\, acting on the CH-CH group of donors\, other acceptors	F	2	8	16	2.169	0.08
6445	regulation of translation	P	3	15	34	2.177	0.081
6818	hydrogen transport	P	3	17	60	1.919	0.087
15992	proton transport	P	3	17	59	1.919	0.087
3676	nucleic acid binding	F	41	508	2068	1.801	0.092
5543	phospholipid binding	F	2	8	24	2.169	0.094
5509	calcium ion binding	F	9	82	594	1.752	0.094
6446	regulation of translational initiation	P	2	8	13	2.169	0.098
16830	carbon-carbon lyase activity	F	2	9	28	1.958	0.098
16829	lyase activity	F	4	28	119	1.735	0.098

HOMOZYGOTES										downregulated
MAP Name										Permutep
Mm_Ribosomal_Proteins										0.042
GOLD	GO Name	GO Type	Number Changed	Number Measured	Number On MAPP	Z Score	Permutep			
5833	hemoglobin complex	C	2	2	4	6	9.883	0.001		
15671	oxygen transport	P	2	2	4	16	9.883	0.001		
15669	gas transport	P	2	2	4	9	9.883	0.001		
5344	oxygen transporter activity	F	2	2	4	10	9.883	0.001		
5829	cytosol	C	5	5	110	221	3.841	0.006		
16568	chromatin modification	P	2	2	18	55	4.341	0.01		
6325	establishment and/or maintenance of chromatin architecture	P	2	2	35	107	2.834	0.03		
6323	DNA packaging	P	2	2	39	118	2.622	0.043		
7001	chromosome organization and biogenesis (sensu Eukarya)	P	2	2	41	171	2.527	0.047		
3677	DNA binding	F	6	6	291	1695	1.956	0.052		
5830	cytosolic ribosome (sensu Eukarya)	C	2	2	43	48	2.438	0.056		
6997	nuclear organization and biogenesis	P	2	2	43	178	2.438	0.056		
16043	cell organization and biogenesis	P	4	4	167	636	1.89	0.063		
6355	regulation of transcription\, DNA-dependent	P	5	5	231	1801	1.884	0.073		
45449	regulation of transcription	F	5	5	235	1295	1.842	0.081		
3924	GTPase activity	F	2	2	46	189	2.314	0.084		
3700	transcription factor activity	F	3	3	100	920	2.064	0.084		
6351	transcription\, DNA-dependent	P	5	5	237	1298	1.821	0.085		
6350	transcription	P	5	5	245	1359	1.741	0.091		
8151	cell growth and/or maintenance	P	10	10	626	2761	1.766	0.098		
HOMOZYGOTES										upregulated
MAP Name										Permutep
Mm_Translation_Factors										0.008
Mm_DNA_replication										0.021
Mm_TGF_Beta_Signaling_Pathway										0.189
Mm_Proteasome_Degradation										0.422
Mm_Ribosomal_Proteins										0.752
GOLD	GO Name	GO Type	Number Changed	Number Measured	Number in GO	Z Score	Permutep			
8143	poly(A) binding	F	2	2	2	3	6.665	0.002		
6457	protein folding	P	6	6	34	109	3.856	0.002		
3727	single-stranded RNA binding	F	2	2	3	8	5.32	0.003		
5834	heterotrimeric G-protein complex	C	2	2	4	30	4.502	0.009		
8639	small protein conjugating enzyme complex	F	3	3	14	31	3.163	0.012		
4840	ubiquitin conjugating enzyme activity	F	3	3	14	34	3.163	0.012		
3729	mRNA binding	F	4	4	21	41	3.339	0.013		
3723	RNA binding	F	11	11	137	321	2.207	0.017		
16021	integral to membrane	C	11	11	484	3419	-2.468	0.022		
3927	heterotrimeric G-protein GTPase activity	F	2	2	6	53	3.504	0.024		
6397	mRNA processing	P	5	5	41	101	2.507	0.028		

GOID	GO Name	GO Type	Number Changed	Number Measured	Number in GO	Z Score	PermuteP
6376	mRNA splice site selection	P	2	7	13	3.165	0.032
8565	protein transporter activity	F	8	84	227	2.395	0.034
16071	mRNA metabolism	P	5	43	103	2.384	0.037
5623	cell	C	60	1614	7894	-2.037	0.037
6512	ubiquitin cycle	P	4	30	68	2.448	0.04
3754	chaperone activity	F	5	47	168	2.157	0.043
245	spliceosome assembly	P	2	9	15	2.65	0.048
5794	Golgi apparatus	C	7	76	288	2.137	0.048
3924	GTPase activity	F	5	46	189	2.211	0.054
19897	extrinsic to plasma membrane	C	2	9	43	2.65	0.056
16020	membrane	C	19	639	4110	-1.94	0.061
6396	RNA processing	P	6	67	196	1.898	0.062
4842	ubiquitin-protein ligase activity	F	3	23	47	2.072	0.068
16070	RNA metabolism	P	6	69	204	1.82	0.069
3674	molecular_function	F	96	2076	10534	1.886	0.072
19898	extrinsic to membrane	C	2	11	42	2.27	0.073
377	RNA splicing\, via transesterification reactions with bulged adenosine as nucleophile	P	3	22	63	2.163	0.077
375	RNA splicing\, via transesterification reactions	P	3	22	63	2.163	0.077
398	nuclear mRNA splicing\, via spliceosome	P	3	22	64	2.163	0.077
16881	acid-D-amino acid ligase activity	F	3	25	56	1.903	0.083
8380	RNA splicing	P	3	23	67	2.072	0.085
9408	response to heat	P	2	11	29	2.27	0.088
19538	protein metabolism	P	28	476	1758	1.885	0.088
9266	response to temperature	P	2	12	33	2.113	0.098

HETEROZYGOTES										downregulated
MAP Name										
no significant MAPs										
GOID	GO Name	GO Type	Number Changed	Number Measured	Number On MAPP	Z Score	Permutep			
6396	RNA processing	P	5	67	196	4.005	0.001			
16070	RNA metabolism	P	5	69	204	3.918	0.002			
6376	mRNA splice site selection	P	2	7	13	5.829	0.003			
245	spliceosome assembly	P	2	9	15	5.06	0.006			
3723	RNA binding	F	6	137	321	2.791	0.011			
5604	basement membrane	C	2	14	50	3.894	0.02			
6397	mRNA processing	P	3	41	101	3.039	0.021			
16071	mRNA metabolism	P	3	43	103	2.931	0.023			
377	RNA splicing\, via transesterification reactions with bulged adenosine as nucleophile	P	2	22	63	2.898	0.038			
398	nuclear mRNA splicing\, via spliceosome	P	2	22	64	2.898	0.038			
375	RNA splicing\, via transesterification reactions	P	2	22	63	2.898	0.038			
8380	RNA splicing	P	2	23	67	2.808	0.04			
3676	nucleic acid binding	F	13	508	2068	2.117	0.04			
6139	nucleobase\, nucleoside\, nucleotide and nucleic acid metabolism	P	11	418	1926	2.009	0.057			
5856	cytoskeleton	C	4	107	522	1.898	0.072			
7017	microtubule-based process	P	2	28	118	2.429	0.076			
HETEROZYGOTES										upregulated
MAP Name										
Mm_DNA_replication										
	Mm_Ribosomal_Proteins		3	14	40	4.544	0.001			
	Mm_Proteasome_Degradation		4	71	80	1.767	0.094			
			2	42	64	0.95	0.601			
GOID	GO Name	GO Type	Number Changed	Number Measured	Number in GO	Z Score	Permutep			
15934	large ribosomal subunit	C	3	19	32	4.213	0.006			
30529	ribonucleoprotein complex	C	8	131	281	3.333	0.006			
3995	acyl-CoA dehydrogenase activity	F	2	7	14	4.928	0.014			
6725	aromatic compound metabolism	P	2	8	70	4.559	0.015			
16637	oxidoreductase activity\, acting on the CH-CH group of donors\, other acceptors	F	2	8	16	4.559	0.021			
8565	protein transporter activity	F	5	84	227	2.541	0.024			
16627	oxidoreductase activity\, acting on the CH-CH group of donors	F	2	10	28	3.987	0.029			
5840	ribosome	C	5	90	159	2.364	0.03			
19538	protein metabolism	P	16	476	1758	2.203	0.036			
3735	structural constituent of ribosome	F	5	98	266	2.149	0.044			
5829	cytosol	C	5	110	221	1.864	0.076			
6605	protein targeting	P	3	50	176	1.97	0.077			
5057	receptor signaling protein activity	F	2	21	84	2.407	0.078			
16337	cell-cell adhesion	P	2	24	182	2.164	0.079			
8289	lipid binding	F	2	24	115	2.164	0.091			
16829	lyase activity	F	2	28	119	1.894	0.094			
5737	cytoplasm	C	21	734	2532	1.798	0.098			

Supplementary Data Table 4 Relationship of differential genes to the processes mentioned in the text of the main manuscript; HET = heterozygotes, HOM = homozygotes, nc = no change; a dash indicates that the fold change could not be reliably computed for that group.

Mouse Gene Name	Relationship to processes in manuscript	HET	HOM
2310057H16Rik	plays a role in microtubule-based movement	down	nc
Abi1	related to Rac induced cytoskeleton reorganization in reaction to growth factors and to cell motility [1, 2]	nc	up
Abi2	related to Rac induced cytoskeleton reorganization in reaction to growth factors and to cell motility [1, 2]		up
Acad8	catalyzes the first dehydrogenation step in the beta-oxidation of fatty acyl-CoA derivatives	up	nc
Acads	catalyzes the first intramitochondrial step in the beta-oxidation of fatty acids		up
Actl6a	chromatin-related		down
Actn4	actin-binding protein involved in the regulation of cell motility [3, 4]	nc	up
Aldoc	encodes the glycolytic enzyme aldolase	nc	up
Arhgap21	potentially cytoskeletal Rho-GTPase activating protein involved in cell differentiation [5]	down	nc
Asb6	product contains a SOCS box, that inhibits cytokine signaling [6], targets proteins for ubiquitination by ubiquitin protein ligases [7] and may inhibit JAK/STAT signaling [8]	down	nc
Bag2	encodes SERCA isoform in spindle fibers [9, 10] and is controlled by the p38 system [11]; involved in the apoptotic process [12]	nc	up
Birc6	prevents apoptosis by ubiquitination of SMAC and caspase-9 [13]	down	up
Btrc	promotes degradation of beta-catenin and I kappa B [14]; activated by increased JNK at a non-transcriptional level [15]	-	down
Cdc37	encodes a protein partaking in complexes related to TNF-mediated induction of IKK and NF-kappaB [16], which is being involved in hypertrophy [17] and MAPK signaling [18]	nc	up
Cdk2	regulates levels of the cadherin-associated pro-hypertrophic component beta-Catenin over the course of the cell cycle and regulates the interaction of beta-Catenin with Axin [19-21]		up
Ckmt1	involved in energy transfer from mitochondrial ATP to cytoplasmic phosphocreatine for fast generation of energy	nc	up
Ctnnbl1	linked to induction of apoptosis [22]	up	-
Cugbp2	binds to a splicing motif in cardiac troponin T (cTNT) and in this way positively regulating the expression of the embryonic form of troponin T; also related to the disease related myotonin dystrophy protein kinase gene (DMPK) [23, 24]	down	up
D19Ert144e	homologous to human Doc1r, a MAP kinase substrate that controls microtubule organization [25]	up	nc
Echs1	catalyzes the second intramitochondrial step in the beta-oxidation of fatty acids		up
Ets1	product is activated by ERK1/2 [26]	-	up
Ets2	stimulates the CD13/aminopeptidase N complex via a Ras/MAPK mediated pathway; a potent regulator of angiogenesis [27]	up	nc
F11r	required for the integrin-mediated angiogenic activity of bFGF [28]; functioning through the ERK system [29]	down	nc
Fhl1	protein interacting with Erk2 and previously shown also to be upregulated in (integrin mediated) cardiomyopathy [30, 31]; family member Fhl2 has been found to inhibit Erk2 [32]; interacts with cMyBP-C [33]; genes related to integrins and MLP (muscle LIM protein) are found to be associated with the Z-disc and interact with the extracellular matrix, thereby acting as stress sensors in the hypertrophic response [34, 35]	nc	up
Fmr1	linked to the Rac1 GTPase pathway and cytoskeleton remodeling [36]	down	nc
Fndc5	strongly expressed in adult heart [37]		down
Foxo3	activates sarcomeric atrogen-1 [38], which prevents calcineurin-dependent hypertrophy [38, 39]		down
Fyn	related to ERK activation [40]	down	-
Gap43	involved in mitigation of MAPK activation [41]	-	down
Gnai3	involved in ERK activation [42]	-	down
Igtp	anti-apoptotic gene; counteracts [43, 44] initiation of the proteolytic activity of caspase-3 following the release of cytochrome c from the mitochondrion [43, 44]		down
Lad1	basement membrane related protein [45]	down	down
Large	myodystrophy gene	up	-
Lgals4	member of a cell-cell and cell-matrix interaction protein family	up	down
Mapk8ip3	related to JNK [46]	down	-

Mouse Gene Name	Relationship to processes in manuscript	HET	HOM
Mapk9	MAP kinase signaling, key gene; absence causes defect in release of cytochrome c from mitochondria [47]	up	up
Mapt	microtubule-related protein		down
Ndufc2	protein product involved in the mitochondrial respiratory chain		up
Nucb1	calcium-binding	nc	up
Pbk	related to the MAPKK family [48]		up
Perp	highly expressed in apoptotic cells as compared to cell-cycle arrested cells and a target of p53 [49]	nc	up
Plagl1	plays a stimulating role in apoptosis and cell cycle arrest [50, 51]	down	up
Ppic	related to DNA degradation in apoptosis [52]; encodes a protein that is associated with calcineurin	nc	up
Rab6ip2	anti-apoptotic I κ B kinase regulatory subunit [53]	down	-
Rcn2	calcium-binding	nc	up
Runx2	linked to the ERK system and angiogenesis (het) [54]; anti-apoptotic gene	down	-
Slk	product cleaved by caspase 3 [55]; initiating apoptosis through actin disassembly and cell adhesion modification [56]	-	up
Smoc1	calcium-binding basement membrane protein [57]	up	-
Smtn	actin binding smooth muscle-cell gene, present at the Z-disc level of cardiomyocytes [58-60]	nc	up
Spred1	inhibitor of cell migration and ERK [61]	down	-
Stk17b	able to phosphorylate myosin light chain and is involved in apoptotic signaling [62]	nc	up
Suc1g2	protein product involved in the citric acid cycle		up
Tacc3	centrosome component whose absence triggers p53-mediated apoptosis [63]	-	down
Trpc4ap	related to TNF, IKK and NF- κ B activation [64]		
Txn2	anti-apoptotic mitochondrial gene	-	down
Txndc5	protects endothelial cells against apoptosis [65]		up
Ube2j2	family member of Ube2r2		up
Ube2r2	member of family that regulates beta-catenin function		up
Uqcrh	accelerates apoptosis by enhancing the release of cytochrome c from mitochondria [66]	up	up
Xpo1	regulates export of Hog1 MAPK from the nucleus [67]	up	up
Zc3hdc5	part of the ubiquitin ligase complex		down

Supplementary references

- 1 Courtney KD, Grove M, Vandongen H, Vandongen A, LaMantia AS, Pendergast AM: **Localization and phosphorylation of Abl-interactor proteins, Abi-1 and Abi-2, in the developing nervous system.** *Mol Cell Neurosci* 2000, **16**:244-257.
- 2 Stradal T, Courtney KD, Rottner K, Hahne P, Small JV, Pendergast AM: **The Abl interactor proteins localize to sites of actin polymerization at the tips of lamellipodia and filopodia.** *Curr Biol* 2001, **11**:891-895.
- 3 Menez J, Le Maux Chansac B, Dorothee G, Vergnon I, Jalil A, Carlier MF, Chouaib S, Mami-Chouaib F: **Mutant alpha-actinin-4 promotes tumorigenicity and regulates cell motility of a human lung carcinoma.** *Oncogene* 2004, **23**:2630-2639.
- 4 Daniliuc S, Bitterman H, Rahat MA, Kinarty A, Rosenzweig D, Lahat N, Nitza L: **Hypoxia inactivates inducible nitric oxide synthase in mouse macrophages by disrupting its interaction with alpha-actinin 4.** *J Immunol* 2003, **171**:3225-3232.
- 5 Basseres DS, Tizzei EV, Duarte AA, Costa FF, Saad ST: **ARHGAP10, a novel human gene coding for a potentially cytoskeletal Rho-GTPase activating protein.** *Biochem Biophys Res Commun* 2002, **294**:579-585.
- 6 Kile BT, Metcalf D, Mifsud S, DiRago L, Nicola NA, Hilton DJ, Alexander WS: **Functional analysis of Asb-1 using genetic modification in mice.** *Mol Cell Biol* 2001, **21**:6189-6197.
- 7 Kile BT, Schulman BA, Alexander WS, Nicola NA, Martin HM, Hilton DJ: **The SOCS box: a tale of destruction and degradation.** *Trends Biochem Sci* 2002, **27**:235-241.

- 8 Yasukawa H, Hoshijima M, Gu Y, Nakamura T, Pradervand S, Hanada T, Hanakawa Y, Yoshimura A, Ross J, Jr., Chien KR: **Suppressor of cytokine signaling-3 is a biomechanical stress-inducible gene that suppresses gp130-mediated cardiac myocyte hypertrophy and survival pathways.** *J Clin Invest* 2001, **108**:1459-1467.
- 9 Liu JX, Thornell LE, Pedrosa-Domellof F: **Distribution of SERCA isoforms in human intrafusal fibers.** *Histochem Cell Biol* 2003, **120**:299-306.
- 10 Ogata T, Yamasaki Y: **Ultra-high-resolution scanning electron microscopic studies on the sarcoplasmic reticulum and mitochondria of the rat intrafusal muscle fibers. Part II: The extracapsular region.** *Arch Histol Cytol* 1992, **55**:117-124.
- 11 Ueda K, Kosako H, Fukui Y, Hattori S: **Proteomic identification of Bcl2-associated athanogene 2 as a novel MAPK-activated protein kinase 2 substrate.** *J Biol Chem* 2004, **279**:41815-41821.
- 12 Lips DJ, deWindt LJ, van Kraaij DJ, Doevendans PA: **Molecular determinants of myocardial hypertrophy and failure: alternative pathways for beneficial and maladaptive hypertrophy.** *Eur Heart J* 2003, **24**:883-896.
- 13 Hauser HP, Bardroff M, Pyrowolakakis G, Jentsch S: **A giant ubiquitin-conjugating enzyme related to IAP apoptosis inhibitors.** *J Cell Biol* 1998, **141**:1415-1422.
- 14 Nakayama K, Hatakeyama S, Maruyama S, Kikuchi A, Onoe K, Good RA, Nakayama KI: **Impaired degradation of inhibitory subunit of NF-kappa B (I kappa B) and beta-catenin as a result of targeted disruption of the beta-TrCP1 gene.** *Proc Natl Acad Sci U S A* 2003, **100**:8752-8757.
- 15 Spiegelman VS, Stavropoulos P, Latres E, Pagano M, Ronai Z, Slaga TJ, Fuchs SY: **Induction of beta-transducin repeat-containing protein by JNK signaling and its role in the activation of NF-kappaB.** *J Biol Chem* 2001, **276**:27152-27158.
- 16 Chen G, Cao P, Goeddel DV: **TNF-induced recruitment and activation of the IKK complex require Cdc37 and Hsp90.** *Mol Cell* 2002, **9**:401-410.
- 17 van Empel VP, De Windt LJ: **Myocyte hypertrophy and apoptosis: a balancing act.** *Cardiovasc Res* 2004, **63**:487-499.
- 18 Grammatikakis N, Lin JH, Grammatikakis A, Tschlis PN, Cochran BH: **p50(cdc37) acting in concert with Hsp90 is required for Raf-1 function.** *Mol Cell Biol* 1999, **19**:1661-1672.
- 19 Park CS, Kim SI, Lee MS, Youn CY, Kim DJ, Jho EH, Song WK: **Modulation of beta-catenin phosphorylation/degradation by cyclin-dependent kinase 2.** *J Biol Chem* 2004, **279**:19592-19599.
- 20 Kim SI, Park CS, Lee MS, Kwon MS, Jho EH, Song WK: **Cyclin-dependent kinase 2 regulates the interaction of Axin with beta-catenin.** *Biochem Biophys Res Commun* 2004, **317**:478-483.
- 21 Hardt SE, Sadoshima J: **Negative regulators of cardiac hypertrophy.** *Cardiovasc Res* 2004, **63**:500-509.
- 22 Jabbour L, Welter JF, Kollar J, Hering TM: **Sequence, gene structure, and expression pattern of CTNNBL1, a minor-class intron-containing gene--evidence for a role in apoptosis.** *Genomics* 2003, **81**:292-303.
- 23 Ladd AN, Charlet N, Cooper TA: **The CELF family of RNA binding proteins is implicated in cell-specific and developmentally regulated alternative splicing.** *Mol Cell Biol* 2001, **21**:1285-1296.
- 24 Lu X, Timchenko NA, Timchenko LT: **Cardiac elav-type RNA-binding protein (ETR-3) binds to RNA CUG repeats expanded in myotonic dystrophy.** *Hum Mol Genet* 1999, **8**:53-60.
- 25 Terret ME, Lefebvre C, Djiane A, Rassinier P, Moreau J, Maro B, Verlhac MH: **DOC1R: a MAP kinase substrate that control microtubule organization of metaphase II mouse oocytes.** *Development* 2003, **130**:5169-5177.
- 26 Dittmer J: **The biology of the Ets1 proto-oncogene.** *Mol Cancer* 2003, **2**:29.
- 27 Petrovic N, Bhagwat SV, Ratzan WJ, Ostrowski MC, Shapiro LH: **CD13/APN transcription is induced by RAS/MAPK-mediated phosphorylation of Ets-2 in activated endothelial cells.** *J Biol Chem* 2003, **278**:49358-49368.

- 28 Naik MU, Mousa SA, Parkos CA, Naik UP: **Signaling through JAM-1 and alphavbeta3 is required for the angiogenic action of bFGF: dissociation of the JAM-1 and alphavbeta3 complex.** *Blood* 2003, **102**:2108-2114.
- 29 Naik MU, Vuppalanchi D, Naik UP: **Essential role of junctional adhesion molecule-1 in basic fibroblast growth factor-induced endothelial cell migration.** *Arterioscler Thromb Vasc Biol* 2003, **23**:2165-2171.
- 30 Gaussin V, Tomlinson JE, Depre C, Engelhardt S, Antos CL, Takagi G, Hein L, Topper JN, Liggett SB, Olson EN, Lohse MJ, Vatner SF, Vatner DE: **Common genomic response in different mouse models of beta-adrenergic-induced cardiomyopathy.** *Circulation* 2003, **108**:2926-2933.
- 31 Chu PH, Ruiz-Lozano P, Zhou Q, Cai C, Chen J: **Expression patterns of FHL/Slim family members suggest important functional roles in skeletal muscle and cardiovascular system.** *Mech Dev* 2000, **95**:259-265.
- 32 Purcell NH, Darwis D, Bueno OF, Muller JM, Schule R, Molkentin JD: **Extracellular signal-regulated kinase 2 interacts with and is negatively regulated by the LIM-only protein FHL2 in cardiomyocytes.** *Mol Cell Biol* 2004, **24**:1081-1095.
- 33 McGrath MJ, Cottle DL, Nguyen MA, Dyson JM, Coghill ID, Robinson PA, Holdsworth M, Cowling BS, Hardeman EC, Mitchell CA, Brown S: **Four and a half LIM protein 1 binds myosin-binding protein C and regulates myosin filament formation and sarcomere assembly.** *J Biol Chem* 2006, **281**:7666-7683.
- 34 Ritter O, Neyses L: **The molecular basis of myocardial hypertrophy and heart failure.** *Trends Mol Med* 2003, **9**:313-321.
- 35 Tarone G, Lembo G: **Molecular interplay between mechanical and humoral signalling in cardiac hypertrophy.** *Trends Mol Med* 2003, **9**:376-382.
- 36 Schenck A, Bardoni B, Langmann C, Harden N, Mandel JL, Giangrande A: **CYFIP/Sra-1 controls neuronal connectivity in Drosophila and links the Rac1 GTPase pathway to the fragile X protein.** *Neuron* 2003, **38**:887-898.
- 37 Ferrer-Martinez A, Ruiz-Lozano P, Chien KR: **Mouse PeP: a novel peroxisomal protein linked to myoblast differentiation and development.** *Dev Dyn* 2002, **224**:154-167.
- 38 Sandri M, Sandri C, Gilbert A, Skurk C, Calabria E, Picard A, Walsh K, Schiaffino S, Lecker SH, Goldberg AL: **Foxo transcription factors induce the atrophy-related ubiquitin ligase atrogin-1 and cause skeletal muscle atrophy.** *Cell* 2004, **117**:399-412.
- 39 Li HH, Kedar V, Zhang C, McDonough H, Arya R, Wang DZ, Patterson C: **Atrogin-1/muscle atrophy F-box inhibits calcineurin-dependent cardiac hypertrophy by participating in an SCF ubiquitin ligase complex.** *J Clin Invest* 2004, **114**:1058-1071.
- 40 Kawabe J, Okumura S, Lee MC, Sadoshima J, Ishikawa Y: **Translocation of caveolin regulates stretch-induced ERK activity in vascular smooth muscle cells.** *Am J Physiol Heart Circ Physiol* 2004, **286**:H1845-1852.
- 41 Huang ZY, Wu Y, Burke SP, Gutmann DH: **The 43000 growth-associated protein functions as a negative growth regulator in glioma.** *Cancer Res* 2003, **63**:2933-2939.
- 42 Bao X, Lu C, Frangos JA: **Mechanism of temporal gradients in shear-induced ERK1/2 activation and proliferation in endothelial cells.** *Am J Physiol Heart Circ Physiol* 2001, **281**:H22-29.
- 43 Adams JW, Pagel AL, Means CK, Oksenberg D, Armstrong RC, Brown JH: **Cardiomyocyte apoptosis induced by Galphaq signaling is mediated by permeability transition pore formation and activation of the mitochondrial death pathway.** *Circ Res* 2000, **87**:1180-1187.
- 44 Zhang HM, Yuan J, Cheung P, Luo H, Yanagawa B, Chau D, Stephan-Tozy N, Wong BW, Zhang J, Wilson JE, McManus BM, Yang D: **Overexpression of interferon-gamma-inducible GTPase inhibits coxsackievirus B3-induced apoptosis through the activation of the phosphatidylinositol 3-kinase/Akt pathway and inhibition of viral replication.** *J Biol Chem* 2003, **278**:33011-33019.
- 45 Motoki K, Megahed M, LaForgia S, Uitto J: **Cloning and chromosomal mapping of mouse ladinin, a novel basement membrane zone component.** *Genomics* 1997, **39**:323-330.

- 46 Matsuura H, Nishitoh H, Takeda K, Matsuzawa A, Amagasa T, Ito M, Yoshioka K, Ichijo H: **Phosphorylation-dependent scaffolding role of JSAP1/JIP3 in the ASK1-JNK signaling pathway. A new mode of regulation of the MAP kinase cascade.** *J Biol Chem* 2002, **277**:40703-40709.
- 47 Tournier C, Hess P, Yang DD, Xu J, Turner TK, Nimnual A, Bar-Sagi D, Jones SN, Flavell RA, Davis RJ: **Requirement of JNK for stress-induced activation of the cytochrome c-mediated death pathway.** *Science* 2000, **288**:870-874.
- 48 Gaudet S, Branton D, Lue RA: **Characterization of PDZ-binding kinase, a mitotic kinase.** *Proc Natl Acad Sci U S A* 2000, **97**:5167-5172.
- 49 Attardi LD, Reczek EE, Cosmas C, Demicco EG, McCurrach ME, Lowe SW, Jacks T: **PERP, an apoptosis-associated target of p53, is a novel member of the PMP-22/gas3 family.** *Genes Dev* 2000, **14**:704-718.
- 50 Spengler D, Villalba M, Hoffmann A, Pantaloni C, Houssami S, Bockaert J, Journot L: **Regulation of apoptosis and cell cycle arrest by Zac1, a novel zinc finger protein expressed in the pituitary gland and the brain.** *Embo J* 1997, **16**:2814-2825.
- 51 Hoffmann A, Ciani E, Boeckardt J, Holsboer F, Journot L, Spengler D: **Transcriptional activities of the zinc finger protein Zac are differentially controlled by DNA binding.** *Mol Cell Biol* 2003, **23**:988-1003.
- 52 Montague JW, Hughes FM, Jr., Cidlowski JA: **Native recombinant cyclophilins A, B, and C degrade DNA independently of peptidylprolyl cis-trans-isomerase activity. Potential roles of cyclophilins in apoptosis.** *J Biol Chem* 1997, **272**:6677-6684.
- 53 Ducut Sigala JL, Bottero V, Young DB, Shevchenko A, Mercurio F, Verma IM: **Activation of transcription factor NF-kappaB requires ELKS, an IkappaB kinase regulatory subunit.** *Science* 2004, **304**:1963-1967.
- 54 Qiao M, Shapiro P, Kumar R, Passaniti A: **Insulin-like growth factor-1 regulates endogenous RUNX2 activity in endothelial cells through a phosphatidylinositol 3-kinase/ERK-dependent and Akt-independent signaling pathway.** *J Biol Chem* 2004, **279**:42709-42718.
- 55 Sabourin LA, Tamai K, Seale P, Wagner J, Rudnicki MA: **Caspase 3 cleavage of the Ste20-related kinase SLK releases and activates an apoptosis-inducing kinase domain and an actin-disassembling region.** *Mol Cell Biol* 2000, **20**:684-696.
- 56 Wagner S, Flood TA, O'Reilly P, Hume K, Sabourin LA: **Association of the Ste20-like kinase (SLK) with the microtubule. Role in Rac1-mediated regulation of actin dynamics during cell adhesion and spreading.** *J Biol Chem* 2002, **277**:37685-37692.
- 57 Vannahme C, Smyth N, Miosge N, Gosling S, Frie C, Paulsson M, Maurer P, Hartmann U: **Characterization of SMOC-1, a novel modular calcium-binding protein in basement membranes.** *J Biol Chem* 2002, **277**:37977-37986.
- 58 Kramer J, Aguirre-Arteta AM, Thiel C, Gross CM, Dietz R, Cardoso MC, Leonhardt H: **A novel isoform of the smooth muscle cell differentiation marker smoothelin.** *J Mol Med* 1999, **77**:294-298.
- 59 Johansson B, Eriksson A, Ramaekers F, Thornell L: **Smoothelin in adult and developing human arteries and myocardium.** *Histochem Cell Biol* 1999, **112**:291-299.
- 60 van der Loop FT, Schaart G, Timmer ED, Ramaekers FC, van Eys GJ: **Smoothelin, a novel cytoskeletal protein specific for smooth muscle cells.** *J Cell Biol* 1996, **134**:401-411.
- 61 Miyoshi K, Wakioka T, Nishinakamura H, Kamio M, Yang L, Inoue M, Hasegawa M, Yonemitsu Y, Komiya S, Yoshimura A: **The Sprouty-related protein, Sprad, inhibits cell motility, metastasis, and Rho-mediated actin reorganization.** *Oncogene* 2004, **23**:5567-5576.
- 62 Sanjo H, Kawai T, Akira S: **DRAKs, novel serine/threonine kinases related to death-associated protein kinase that trigger apoptosis.** *J Biol Chem* 1998, **273**:29066-29071.
- 63 Piekorz RP, Hoffmeyer A, Duntsch CD, McKay C, Nakajima H, Sexl V, Snyder L, Reh J, Ihle JN: **The centrosomal protein TACC3 is essential for hematopoietic stem cell function and genetically interfaces with p53-regulated apoptosis.** *Embo J* 2002, **21**:653-664.

- 64 Soond SM, Terry JL, Colbert JD, Riches DW: **TRUSS, a novel tumor necrosis factor receptor 1 scaffolding protein that mediates activation of the transcription factor NF-kappaB.** *Mol Cell Biol* 2003, **23**:8334-8344.
- 65 Sullivan DC, Huminiecki L, Moore JW, Boyle JJ, Poulsom R, Creamer D, Barker J, Bicknell R: **EndoPDI, a novel protein-disulfide isomerase-like protein that is preferentially expressed in endothelial cells acts as a stress survival factor.** *J Biol Chem* 2003, **278**:47079-47088.
- 66 Okazaki M, Ishibashi Y, Asoh S, Ohta S: **Overexpressed mitochondrial hinge protein, a cytochrome c-binding protein, accelerates apoptosis by enhancing the release of cytochrome c from mitochondria.** *Biochem Biophys Res Commun* 1998, **243**:131-136.
- 67 Ferrigno P, Posas F, Koepp D, Saito H, Silver PA: **Regulated nucleo/cytoplasmic exchange of HOG1 MAPK requires the importin beta homologs NMD5 and XPO1.** *Embo J* 1998, **17**:5606-5614.

Chapter 4

Affymetrix expression chip data analysis: the gain of modeling

L.M.T. Eijssen^{1}, P.J. Lindsey¹, R.G.E. van Eijsder², B. Schroen³, Y.M. Pinto³, H.J.M. Smeets^{1,2,4}, J.P.M. Geraedts²*

¹ Department of Population Genetics, Genomics, and Bioinformatics, Maastricht University, P.O. Box 616, 6200 MD, Maastricht, the Netherlands

² Department of Clinical Genetics, Maastricht University, P.O. Box 616, 6200 MD, Maastricht, the Netherlands

³ Department of Molecular Cardiology, Maastricht University, P.O. Box 616, 6200 MD, Maastricht, the Netherlands

⁴ Genome Center, Maastricht University, P.O. Box 616, 6200 MD, Maastricht, the Netherlands

**To whom correspondence should be addressed:*

L.M.T. Eijssen, Department of Genetics and Cell Biology, P.O. Box 616, 6200 MD Maastricht, The Netherlands. Tel.: +31-43-3881995. Fax: +31-43-3884573. E-mail: Lars.Eijssen@gen.unimaas.nl

submitted

Abstract

Background: Analysis of large datasets that result from microarray gene expression studies has become a major challenge in the post-genome world. Often, basic assumptions are made within standard statistical frameworks. Well-known hypotheses amongst these are Gaussian distribution and equality of dispersion within the experimental subgroups. The effect of these assumptions on the quality of gene expression modeling was investigated by building different statistical models and evaluating their likeliness through the Akaike Information Criterion (AIC). Besides evaluating the AIC, several other criteria were taken into consideration: the Bayesian Information Criterion (BIC), the total minus log-likelihood, and the total number of non-fitting or non-converging models for each of the settings and distributions.

Results: Two datasets performed on the Affymetrix Genechip Expression Array platform, one containing human cardiac samples and one containing human mitochondrial disease samples, were used to validate the different models. Four models were fitted for each of ten different continuous statistical distributions. Results clearly showed that relaxing the assumption of equal dispersion was a very effective approach, where the Gaussian model remained the best fitting one and the extra cost in degrees of freedom needed is very low. Besides improved model fit, results also improved from a biological viewpoint for test cases. Under the frequent assumption of equal dispersion, modeling could be improved by using other distributions than the Gaussian, but the effect was not as strong as relaxing the assumption.

Conclusions: We conclude that a general procedure using statistical models that allow for differences in dispersion among (treatment) groups improves model fit considerably and is very feasible in practice.

Keywords

distributions, location and dispersion, normality, group effect, microarray

Background

Microarray gene expression analysis has become a technique of major importance for retrieving global or specific genetic differences among several groups of biological samples, e.g. disease versus healthy samples. Several platforms are used, from custom-made cDNA and oligo-arrays to standardized commercial platforms [1]. The commercial Affymetrix Genechip Expression Array system is a reproducible and reliable platform [2] which has been increasingly employed. Many analysis methods have been developed to extract biologically relevant differential genes [3-6] or pathways [7-10] from the large datasets resulting from microarray studies. These methods generally consist of several steps, from background correction, summarizing probe expression levels into probeset intensities, and normalization [11], to evaluating differential probeset expression. Especially, the first three steps have been investigated thoroughly before, for instance the MAS5 (Affymetrix Microarray Suite) [3] or the robust multi-array average (RMA) method [5]. This paper focuses on the fourth step, the probeset (gene) expression modeling, where some elements of correction generally done by preprocessing are incorporated into this step.

The procedures used for this particular part of microarray data analysis are based on several assumptions [1, 12]. The major ones are Gaussian ('normal') distribution of the data and equal dispersion among subgroups. It is uncertain however, that normality assumptions will not be violated, even if logged ratios are used. Several distributions, other than the Gaussian, allow for heavier tails or skewness, possibly enabling better modeling of differential genes. Further, it is not unlikely that the assumption of equal dispersion may not be appropriate and allowing for differences in dispersion would then improve model fit considerably.

Analysis methods have already been developed in order to allow for unequal variances among groups such as Welch's t-test. However, Welch's t-test assumes that the data are normally distributed and can only draw a comparison between two groups. Further, performing tests is less informative than modeling approaches. Although, for instance a method such as Linear Models for Microarray Data (LIMMA) [6] is flexible enough to allow comparisons among several groups, it does not allow the choice of distribution.

In the current research, we propose a flexible approach allowing choosing among a large choice of distributions. This approach also allows modeling both the location and/or dispersion of the chosen distribution by specifying a linear (or non-linear) formula. If the design requires (for instance in the presence of technical repeats or time course experiments), several covariance structures can also be chosen. Hence, this approach permits us to investigate whether relaxation of the Gaussian and equal dispersion assumptions do in fact lead to better modeling by comparing several models for each of ten common statistical distributions. Although it is clear that allowing for a different distribution for every single probeset would necessarily lead to the best total model fit, in this paper, focus is on which overall assumptions are best used and whether improvement of fit is possible when still keeping a uniform and

ready applicable procedure for the entire dataset. This means that one set of settings should to be used for all probesets. Because features of the experimental design are available, these are taken into account in the models built. Thus, a minimal amount of preprocessing is used for the three foregoing steps.

To judge the fit of the different models considered, criteria such as the AIC [13] and BIC [14] are used rather than asymptotic inferences because they are suitable for small sample sizes. In addition, they are able to compare non-nested models and do not require any corrections such as the ones required when performing multiple tests [15-17].

To judge which statistical models performed best, two differing Affymetrix Genechip Expression Array datasets at hand were selected as test sets, enabling us to draw conclusions. Different types of chips were run for the two studies, each considering a different human disease and containing three subgroups of samples (two disease groups and a control group). We define what the best overall assumptions are to use in a ready-applicable gene expression analysis method and what the associated cost is, in degrees of freedom needed.

Methods

Distributions and models

First, two reference models were fitted: a model containing no covariates ('intercept', containing 2 parameters) and a model containing only hybridization and labeling spike intensities ('covariate', containing 11 parameters). Then several additional models were created: a model including a location effect for the groups ('location model', 13 parameters), a model allowing for differences in dispersion among groups ('dispersion model', 13 parameters), and a model including both ('location and dispersion model', 15 parameters). These three models also contain hybridization and labeling spike intensities and are intended to be compared to the reference models. Each of the models used the coefficients of the variables from a suitable reference model as initial estimates for the fitting procedure. Neither non-linear formulae nor covariance structures will be needed for any of the models mentioned.

All models were created for each probeset of both datasets and for ten distributions. Five two-parameter and five three-parameter distributions were considered: the Gaussian, logistic, gamma, Cauchy, and Laplace and the Student T, generalized logistic, generalized gamma, power exponential, and skewed Laplace respectively [18, 19].

Before building models for all probesets, a run on a subset of fifty probesets was done for both datasets in order to determine the best parameter settings for the computational digit precision and the maximum step to take between iterations. Both parameters needed to be well-chosen to enhance numerical convergence. To improve convergence further in the complete runs, two rounds of iterations were done while fitting the models, a first round with a maximum of 500 steps and a second with a maximum of 100 steps starting from the estimates obtained from the first round.

Evaluation of the models

Models were rated with the Akaike Information Criterion (AIC) [13] that is computed by summing twice the minus log-likelihood (-2LL) of the model and twice the number of parameters it contains, yielding lower values for better-fitting and less complex models, thus a goodness-of-fit measure penalized for model complexity. AIC values were only used as a measure to rank the different models whereas absolute values were not interpreted.

The Bayesian Information Criterion (BIC) [14] is computed by summing twice the minus log-likelihood (-2LL) of the model and the number of parameters it contains multiplied by the log of the number of observations. BIC and -2LL totals were used to confirm the conclusions drawn from the AIC results.

Model convergence was checked by comparing the -2LL values of the five models within one distribution with each other. The -2LL value of a model extended with more variables was required to always be equal to or lower than that of more basic ones to ensure proper convergence. If this was not the case within a small tolerance value (of 1 LL unit) to account for small random and rounding effects, the -2LL value and model coefficients were replaced by those of the more basic model and the AIC recomputed. As an additional built-in check, any model with an infinite parameter value or an infinite AIC, was discarded from further analysis. For the generalized logistic distribution a further built-in check was performed on its family parameter, because it was required to remain finite (cut-off value 50) to ensure a valid distribution.

Finally, the model fitting process was timed for each of the models and distributions.

Programming environment

All code was implemented in the R package [20], with use of the `gnlr` function part of the `gnlm` library [21, 22] and the `affy` library [23]. Probeset intensities were computed using the `expresso` method of the `affy` library, with no or “MAS” background correction, constant normalization, no use of mismatch intensities and the “MAS” summarization method.

The analyses were done on a 2.8 GHz dual processor server (5571 bogomips), having 4 GB of internal memory and operating on Redhat linux V9 (Shrike).

Data

Two human Affymetrix Genechip Expression Array datasets from the Genome Center lab at Maastricht University were used. For a cardiac study considering hypertrophy caused by valvular disease (aorta stenosis), twenty-six samples were run on hgu133a chips, with nine and ten samples in two disease groups and seven (coronary artery bypass grafting) controls. Each chip contained expression values of 22215 probe sets and several spikes.

For a mitochondrial study considering subtypes of MELAS disease and Leigh syndrome [24], twenty samples were run on hgu95av2 chips, with seven samples in

each of two disease groups and six controls. Each chip contained expression values of 12558 probe sets and several spikes.

Expression values were log scaled (not resulting in negative values) before fitting the models. Spike intensities were controlled for and used as covariates in the models. The datasets, with permuted probesets and masked labels because they represent ongoing research, are available upon request from the first author. The datasets are further referred to as respectively the ‘cardio-set’ and the ‘mito-set’.

Results

Parameter settings and missing values

First, a test run was performed on 50 probesets to determine settings for the computational precision and the maximum step size to be taken between modeling iterations. This resulted in the settings given in Supplementary Material (Table S1) being selected. These settings differed among different distributions and models, especially for the maximum step size parameter.

In order to obtain the total AIC score of all probeset models for each distribution and type of model, Not Available (NA) values in the AICs were removed. These represent models that could not be built (did not fit or converge) or did not meet the criteria set. The total number of NA values for each group is presented in Table 1 and Supplementary Material (Figure S1).

Table 1 Total number of non existing (NA) models over all probesets: per distribution and model type.

	Intercept Model	Covariate Model	Location Model	Dispersion Model	Location and dispersion Model
cardio-set					
Gaussian	0	0	0	0	0
Logistic	0	0	0	0	0
Gamma	0	0	0	6	23
Cauchy	0	0	0	0	0
Laplace	0	0	55	0	86
Student T	0	476	619	357	2771
Generalized Logistic	2734	3016	1427	651	601
Generalized Gamma	1	0	22	37	23
Power Exponential	10871	1135	8900	8730	8742
Skewed Laplace	0	0	3	2	6
mito-set					
Gaussian	0	0	0	0	0
Logistic	0	0	0	29	2
Gamma	0	0	0	4	50
Cauchy	0	0	0	1	0
Laplace	0	0	9	0	0
Student T	0	21	114	135	1402
Generalized Logistic	1893	1560	828	367	368
Generalized Gamma	2	0	23	33	26
Power Exponential	5896	581	4758	4809	4832
Skewed Laplace	0	0	10	6	7

Fit of the models

All probesets for which there was at least one model that did not exist for any of the ten distributions were removed from the dataset, leaving 3889 out of 22215 probe sets for analysis for the cardio-set and 2401 out of 12558 for the mito-set. Especially the power exponential, but also the Student T and the generalized logistic distributions contained many models that did not fit or converge. All distributions and model types were rated by their total AIC values over all probesets. The results contained few outliers, such that only three further AIC values had to be removed from the summary for the cardio-set (all generalized gamma, intercept model). For the mito-set nine values had to be removed (one generalized logistic intercept model, eight generalized gamma intercept model).

Because the number of analyzable probesets was quite restricted mainly because of difficulties building the three-parameter distribution models, and the fit of the two-parameter distributions was better for the analyzable probesets, results were first calculated for the latter only. This allowed analysis of 22047 probe sets for the cardio-set and 12464 for the mito-set, after removing the missings (less than 1% in both cases). The AIC results of these analyses are presented in Figure 1 and Supplementary Material (Table S2). The BIC and minus log-likelihood results for both analyses are presented in the Supplementary Material (respectively Tables S3 and S4).

Figure 1 Summed AIC value over all probeset models: for each of the model types and per distribution for the run on the two-parameter distributions; the cardio-set on the left and the mito-set on the right.

For the intercept model, the Gaussian and logistic models were best, with more or less equal overall AIC. For the covariate model and the location model, the Cauchy distribution was clearly the best. The other distributions followed with respectable distance and close to each other. Lower AIC values were obtained when allowing for a group difference in dispersion, where for the cardio-set the best results were

obtained for the gamma distribution, in the model without a location effect. For the mito-set, the Cauchy distribution was slightly better than the gamma for this model. In the model containing both location and dispersion effects, clearly the best results were obtained for the Gaussian distribution for both datasets, with the gamma and logistic following at respectable distance. From Supplementary Material (Table S3), it can be seen that using the BIC does not change the model ranking for either datasets. This is to be expected when dealing with small sample sizes (respectively 26 and 20 samples for the cardio and mito-sets). The model fit presented in Supplementary Material (Table S4) also yields the same model ranking for both datasets. Diagnostic plots of both datasets are presented in Figure 2. It can be seen from the scatterplot of the AIC values of the location and dispersion model (for the Gaussian distribution) versus the values of the location model that most models fit much better. Further, for a random probeset, it can be seen from the quantile-quantile plot for the residuals of the location and dispersion model and the residual versus fitted value plot (both for the Gaussian distribution) that the model fits well.

For the analyses on all distributions including the three-parameter ones, the resulting data are presented in Supplementary Material (Table S5 and Figure S2). These results show that the model ranking remains the same for both datasets. For the intercept model, the logistic and the Gaussian distribution were close and performed best again. For both models with equal dispersion, the basic covariate model and the location model, the Cauchy distribution was again best, leaving the logistic, Gamma and Gaussian, far behind. Also in this analysis, the AIC values were much lower for the models with a group difference in dispersion. The situation without a location effect resulted in the best fit for the Gamma distribution, in this case for both datasets. The model allowing for group differences in both dispersion and location, resulted in the lowest AIC values and the Gaussian was the best model again, with the Student T, the gamma and the logistic following.

Timing of the models

The total times to fit all the probeset models were computed for each category. When only considering the two-parameter distributions, the fitting times for the Gaussian location and dispersion model were at most twice the fitting times for the standard less-performing location effect model.

Discussion

In gene expression analysis, normality of the data under consideration is frequently assumed, as is equal dispersion for all experimental groups [1, 12]. However, these assumptions may influence the correctness of the results of the analyses. Comparing the fit of several statistical distributions, we showed that increasing modeling freedom

Figure 2 Diagnostic plots: the first row shows the AIC values for the location and dispersion model versus the location model, the second row shows the quantile-quantile plot for the residuals of a random probeset, and the third row shows the residuals versus the fitted values. The left and right columns respectively correspond to the cardio- and mito-set. All AIC values, quantiles, residuals, and fitted values correspond to models for the Gaussian distribution.

by considering models allowing for group differences in both dispersion and location, improved modeling considerably. In our case the Gaussian was the best fitting distribution, indicating that the widely used Gaussian framework worked well, if enough freedom was offered to the modeling process. In two tested datasets, the extra time cost for fitting a Gaussian model including a group difference in dispersion in comparison to the model without was under twofold.

Imposing the assumption of equal dispersion, the Gaussian was not the best distribution to use to model a group effect. The Cauchy distribution fitted best in this case, indicating that robustness to more measurements in the tails of the distribution was required. Hence, alternative distributions such as a heavy tailed or skewed one should also be routinely fitted. Indeed, other specific tissues or organisms may very well not require unequal variances or not be normally distributed.

Of course, allowing a different statistical distribution and a model with different characteristics for each probeset would improve overall model fit even more. However, this option was not incorporated, because the current goal was to design a method that improved modeling by selecting better common assumptions from a broad set of possibilities, but was still generally and easily applicable. Allowing for different distributions for each probeset would require trying all possibilities for each probeset in each analysis making it a lengthy and difficult procedure. Furthermore, the counts in Supplementary Material (Table S2) on how often each model type performed best show that for about 80% of all probesets a model including both location and dispersion fitted best. Of these more than 80% were Gaussian models, indicating the clear optimality of this model and the possibility using it for modeling all probesets in a general procedure. Hence, fitting a number of common continuous distributions to a random subset of probes should provide an indication of which distributions require thorough investigation.

Another issue for many microarray experiments is that covariates and their interactions cannot be taken into account due to small sample sizes. Whereas the models in this study took into account covariates, they might have performed even better if interactions among covariates, spikes, and group effects could have also been included. For the same reason, gene-gene interactions were not taken into account, because this would require even more samples or the input of *a priori* knowledge regarding which interactions are present or not, and this was beyond the scope of the current study.

Modeling results also depend on the foregoing steps in the analysis procedure, the background correction and normalization methods used, and the method used to get from per probe expression levels to probeset intensities. Because the models built took the experimental design into account by including appropriate covariates, basic settings without extensive preprocessing were used to compute probeset intensities.

Overall, several of the three-parameter distributions showed many non-converging models, mainly the power exponential and the generalized logistic. Furthermore, the -2LL, AIC, or BIC values of these two distributions and the skewed Laplace were not as good as for the other models, indicating their relatively small usability for modeling

this type of data. Also, the three-parameter distributions took more time, indicating that these models were more difficult to fit than the others. The generalized gamma distribution took especially long to run for the more complex models, but log-transformation of the data might have been disadvantageous to the fit of this distribution, just as for the basic gamma distribution.

Although the complete calculations for all distributions contained far less probesets than the calculations on the two-parameter distributions only, both the model ranking and timing results were relatively similar between the two analyses, with only close models sometimes changing places. This confirmed the stability of the results. Also between the two datasets, results with respect to the number of non-fitting models and model ranking were very similar, stressing this stability even more. This is strengthened by taking into account that both sets consider very different biological material and were created employing different types of chips.

Modeling was improved by performing a calibration run on a subset of probesets to optimize the settings of several parameters, because some sensitivity to these settings was observed. This was especially the case for the maximum step size to take between iterations. The best setting for the computational precision of the model fitting was quite stable, except for the Gaussian and Student T distributions. Because summing is sensitive to outliers, removal of outliers was very important and this was very effectively done by the built-in checks that prevented models from non-convergence and from going out of bounds during the fitting procedure. Hence, only very few outlying values had to be removed later on from the results. Also proper tuning of the settings of the procedures was quite important for fitting this many models in an automated way. For example, if not constrained, the family parameter of the generalized logistic distribution sometimes would get very large, resulting in an extremely low -2LL value, but a non-valid distribution. For the power exponential distribution, the -2LL values would sometimes go to positive infinite if not carefully guided.

Evaluating for each probeset which model has the lowest AIC or BIC value can also pinpoint differential genes, by selecting those probesets for which the location effect is included in this best model. To determine whether better fitting models also lead to better biological predictions, the specifically differential genes for MELAS disease (thus not for the other syndrome in the dataset, Leigh syndrome) in the mito-set were retrieved by selecting those probesets for which the best model contained a significant MELAS group variable and a non-significant surf1 variable. These genes were uploaded into David / Ease [9], both for a Gaussian analysis not allowing for a group difference in variance (the standard homoscedastic approach) and an analysis allowing for it (a location and dispersion or heteroscedastic model). This tool evaluates which Gene Ontology [8] annotations are overrepresented in the uploaded gene set, a useful approach to filter false positives genes from true differential pathways. The GO annotations on the 'biological process' and 'cellular component' axes at the most specific level were selected to be returned and evaluated for known relevance to the disease. Because the gene list for the heteroscedastic analysis was

longer because more modeling options were available with respect to the group effect, more processes were retrieved by this analysis. Rating these extra processes according to their (known) relationship to the disease, could give an indication of the biological relevance of the more extensive modeling. The list of processes with p-values below 0.05, presented in Supplementary Material (Table S6), showed indeed that the location and dispersion model analysis retrieved more relevant processes in addition to the ones already retrieved by the standard analysis. Both analyses showed transmission of nerve impulse / synaptic transmission, known to be of relevance. Less-related processes retrieved by the homoscedastic analysis disappeared from the results of the heteroscedastic analysis, which showed several processes related to muscle structure and functioning instead. These are very plausible and worth investigating in further biological study, indicating the usefulness of extending models to allow for a group difference in dispersion.

Conclusions

In conclusion, we advise to use a modeling approach allowing for differences in dispersion among groups, because this can greatly improve model fit to probeset expression profiles. In our case the Gaussian distribution is a very good choice that also enables the use of a standard analysis package. The extra cost of modeling dispersion differences comprises only one degree of freedom (for two treatment groups, one extra for each further group), which does not make requirements on sample size more demanding. Also the extra time cost (under twofold compared to the basic model) is very moderate. One should consider using other two-parameter distributions besides the Gaussian, like the Cauchy, logistic or Gamma, when modeling other tissue types, organisms, or under the assumption of equal dispersion in all groups. On the other hand, using three-parameter distributions is of limited use. In each case, doing a test run to determine parameter settings is very helpful in building well-fitting models.

Author's contributions

LE and PL implemented the statistical modeling and evaluated all models by applying them to the datasets and drafted the manuscript, RvE and HS provided the mitochondrial data and helped with analyzing and interpreting results on these data, BS and YP *idem* for the cardiac data, JG offered guiding assistance from a clinical genetics perspective.

Acknowledgements

The authors wish to thank Professor James Lindsey and Mike Gerards for their useful comments on the manuscript. The authors would also like to thank the referees for their valuable help in clarifying the different issues discussed.

This research was supported by the Cardiovascular Research Institute Maastricht (CARIM).

References

- 1 Butte A: **The use and analysis of microarray data.** *Nat Rev Drug Discov* 2002, **1**:951-960.
- 2 Irizarry RA, Warren D, Spencer F, Kim IF, Biswal S, Frank BC, Gabrielson E, Garcia JG, Geoghegan J, Germino G, et al: **Multiple-laboratory comparison of microarray platforms.** *Nat Methods* 2005, **2**:345-350.
- 3 Affymetrix: *Affymetrix Microarray Suite User Guide*. 5th edition. Santa Clara, CA: Affymetrix; 2002.
- 4 Huber W, von Heydebreck A, Vingron M: **Analysis of microarray gene expression data.** In *Handbook of Statistical Genetics*. 2nd edition: Wiley; 2003
- 5 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* 2003, **4**:249-264.
- 6 Smyth GK: **Linear models and empirical bayes methods for assessing differential expression in microarray experiments.** *Stat Appl Genet Mol Biol* 2004, **3**:Article3.
- 7 Becker KG, Hosack DA, Dennis G, Jr., Lempicki RA, Bright TJ, Cheadle C, Engel J: **PubMatrix: a tool for multiplex literature mining.** *BMC Bioinformatics* 2003, **4**:61.
- 8 Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, et al: **The Gene Ontology (GO) database and informatics resource.** *Nucleic Acids Res* 2004, **32**:D258-261.
- 9 Hosack DA, Dennis G, Jr., Sherman BT, Lane HC, Lempicki RA: **Identifying biological themes within lists of genes with EASE.** *Genome Biol* 2003, **4**:R70.
- 10 Tanabe L, Scherf U, Smith LH, Lee JK, Hunter L, Weinstein JN: **MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling.** *Biotechniques* 1999, **27**:1210-1214, 1216-1217.
- 11 Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, et al: **Bioconductor: open software development for computational biology and bioinformatics.** *Genome Biol* 2004, **5**:R80.
- 12 Quackenbush J: **Microarray analysis and tumor classification.** *N Engl J Med* 2006, **354**:2463-2472.
- 13 Akaike H: **Information theory and an extension of the maximum likelihood principle.** In *Second International Symposium on Inference Theory; Budapest: Akadémiai Kiadó*. Edited by Petrov BN, Csàki F. 1973: 267-281.
- 14 Schwarz G: **Estimating the Dimension of a Model.** *Annals of Statistics* 1978, **6**:461-464.
- 15 Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *Journal of the Royal Statistical Society B* 1995, **57**:289-300.
- 16 Dudoit S, Yang YH, Callow MJ, Speed TP: **Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments.** *Statistica Sinica* 2002, **12**:111-139.
- 17 Westfall PH, Young SS: *Resampling-based multiple testing: examples and methods for p-value adjustment*. Wiley; 1993.
- 18 Johnson NL, Kotz S, Balakrishnan N: *Continuous Univariate Distributions 1*. 2nd edition. New York: John Wiley & Sons; 1994.
- 19 Johnson NL, Kotz S, Balakrishnan N: *Continuous Univariate Distributions 2*. 2nd edition. New York: John Wiley & Sons; 1995.
- 20 R Development Core Team: *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing; 2004.
- 21 Lindsey JK: *Models for Repeated Measurements*. Oxford: Oxford University Press; 1999.
- 22 Lindsey JK: R modeling libraries [<http://popgen.unimaas.nl/~jlindsey/rcode.html>]
- 23 Gautier L, Cope L, Bolstad BM, Irizarry RA: **affy--analysis of Affymetrix GeneChip data at the probe level.** *Bioinformatics* 2004, **20**:307-315.
- 24 DiMauro S: **Mitochondrial diseases.** *Biochim Biophys Acta* 2004, **1658**:80-88.

Supplementary Material

Table S1 Settings determined to be best based on a test run on 50 probesets, for each distribution and model type for both datasets; ndigit = number of significant digits; maxstep = maximum stepsize to take between two iterations of fitting.

Cardio-set					
Distribution		Covariate Model	Location Model	Dispersion Model	Location and dispersion Model
Gaussian	ndigit	16	16	8	12
	<i>maxstep</i>	<i>100</i>	<i>1</i>	<i>1</i>	<i>10</i>
Logistic	ndigit	16	16	16	16
	<i>maxstep</i>	<i>20</i>	<i>50</i>	<i>1</i>	<i>0</i>
Gamma	ndigit	16	16	16	16
	<i>maxstep</i>	<i>1000</i>	<i>1</i>	<i>10</i>	<i>10</i>
Cauchy	ndigit	16	16	16	16
	<i>maxstep</i>	<i>1000</i>	<i>1</i>	<i>10</i>	<i>1</i>
Laplace	ndigit	16	16	16	16
	<i>maxstep</i>	<i>1000</i>	<i>50</i>	<i>20</i>	<i>50</i>
Student T	ndigit	8	8	12	16
	<i>maxstep</i>	<i>100</i>	<i>1</i>	<i>1</i>	<i>100</i>
Generalized Logistic	ndigit	8	8	8	8
	<i>maxstep</i>	<i>100</i>	<i>50</i>	<i>0</i>	<i>0</i>
Generalized Gamma	ndigit	16	16	16	16
	<i>maxstep</i>	<i>1</i>	<i>10</i>	<i>100</i>	<i>10</i>
Power Exponential	ndigit	16	16	16	16
	<i>maxstep</i>	<i>50</i>	<i>1000</i>	<i>50</i>	<i>50</i>
Skew Laplace	ndigit	16	16	16	16
	<i>maxstep</i>	<i>20</i>	<i>1000</i>	<i>20</i>	<i>1</i>
Mito-set					
Gaussian	ndigit	12	16	8	16
	<i>maxstep</i>	<i>0</i>	<i>10</i>	<i>10</i>	<i>10</i>
Logistic	ndigit	16	16	16	16
	<i>maxstep</i>	<i>10</i>	<i>1000</i>	<i>1</i>	<i>20</i>
Gamma	ndigit	16	16	16	16
	<i>maxstep</i>	<i>0</i>	<i>1</i>	<i>10</i>	<i>10</i>
Cauchy	ndigit	16	16	16	16
	<i>maxstep</i>	<i>1000</i>	<i>1</i>	<i>10</i>	<i>1</i>
Laplace	ndigit	16	16	16	16
	<i>maxstep</i>	<i>20</i>	<i>20</i>	<i>1</i>	<i>1</i>
Student T	ndigit	16	16	16	16
	<i>maxstep</i>	<i>50</i>	<i>50</i>	<i>1</i>	<i>1000</i>
Generalized Logistic	ndigit	8	8	8	8
	<i>maxstep</i>	<i>10</i>	<i>10</i>	<i>100</i>	<i>0</i>
Generalized Gamma	ndigit	16	16	16	16
	<i>maxstep</i>	<i>10</i>	<i>50</i>	<i>100</i>	<i>10</i>
Power Exponential	ndigit	16	16	16	16
	<i>maxstep</i>	<i>0</i>	<i>50</i>	<i>0</i>	<i>0</i>
Skew Laplace	ndigit	16	16	16	16
	<i>maxstep</i>	<i>100</i>	<i>100</i>	<i>50</i>	<i>0</i>

Figure S1 a) Total number of non existing (NA) models over all probesets, per distribution for each model type for the cardio-set; model order: 1 intercept model, 2 covariate model, 3 location model, 4 location and dispersion model, 5 dispersion model .

Figure S1 b) Total number of non existing (NA) models over all probesets, per distribution for each model type for the mito-set; for model order: see Figure s1 a).

Table S2 Summed AIC value over all probeset models: for each of the distributions and model types for the run on the two-parameter distributions; the best model row indicates for each dataset, the percentage and number of all probesets for which the best of all twenty-five models was a model of the specified type .

	Intercept Model	Covariate Model	Location Model	Dispersion Model	Location and dispersion Model
cardio-set					
Gaussian	-94125.12	-143768.18	-135971.66	-191755.95	-2957523.47
Logistic	-93582.55	-147182.96	-147544.39	-399086.28	-1244042.65
Gamma	-25058.23	-142993.66	-140494.15	-543757.04	-1549501.48
Cauchy	75346.46	-300166.17	-236083.38	-227593.52	-181790.70
Laplace	-46187.92	44254.44	27369.35	-31726.36	-130650.28
best model (%)	2.64 (581)	14.22 (3135)	1.42 (314)	5.17 (1140)	76.55 (16877)
mito-set					
Gaussian	-20901.68	-131187.83	-178223.05	-279765.35	-1855055.83
Logistic	-23195.13	-138292.26	-190482.95	-303207.87	-872364.92
Gamma	38331.61	-132900.34	-180543.25	-362684.96	-1051219.61
Cauchy	46213.71	-390535.82	-414698.65	-387040.37	-399861.16
Laplace	17022.39	-52989.99	-81033.95	-90417.38	-192526.25
best model (%)	0.03 (4)	2.64 (329)	0.78 (98)	8.46 (1054)	88.08 (10979)

Table S3 Summed BIC value over all probeset models: for each of the distributions and model types for the run on the two-parameter distributions.

	Intercept Model	Covariate Model	Location Model	Dispersion Model	Location and dispersion Model
cardio-set					
Gaussian	-38650.61	161341.62	224612.65	168828.36	-2541464.65
Logistic	-38108.04	157926.84	213039.92	-38501.98	-827983.84
Gamma	30416.28	162116.14	220090.16	-183172.74	-1133442.67
Cauchy	130820.97	4943.63	124500.93	132990.78	234268.12
Laplace	9286.59	349364.24	387953.65	328857.95	285408.54
mito-set					
Gaussian	3919.94	5331.05	-16882.55	-118424.85	-1668893.73
Logistic	1626.48	-1773.38	-29142.45	-141867.38	-686202.82
Gamma	63153.23	3618.53	-19202.76	-201344.47	-865057.50
Cauchy	71035.33	-254016.95	-253358.16	-225699.87	-213699.06
Laplace	41844.00	83528.89	80306.54	70923.11	-6364.14

Table S4 Twice the minus log-likelihood (-2LL), summed over all probeset models: for each of the distributions and model types for the run on the two-parameter distributions.

	Intercept Model	Covariate Model	Location Model	Dispersion Model	Location and dispersion Model
cardio-set					
Gaussian	-91156.56	-314401.09	-354596.83	-382488.97	-1809466.74
Logistic	-90885.27	-316108.48	-360383.19	-486154.14	-952726.33
Gamma	-56623.11	-314013.83	-356858.07	-558489.52	-1105455.74
Cauchy	-6420.77	-392600.08	-404652.69	-400407.76	-421600.35
Laplace	-67187.96	-220389.78	-272926.33	-302474.18	-396030.14
mito-set					
Gaussian	-35378.84	-202697.91	-251143.52	-301914.67	-1114487.92
Logistic	-36525.57	-206250.13	-257273.47	-313635.94	-623142.46
Gamma	-5762.19	-203554.17	-252303.63	-343374.48	-712569.80
Cauchy	-1821.14	-332371.91	-369381.33	-355552.18	-386890.58
Laplace	-16416.81	-163599.00	-202548.97	-207240.69	-283223.12

Table S5 Summed AIC value over all probeset models, for each of the distributions and model types for the run on all distributions.

	Intercept Model	Covariate Model	Location Model	Dispersion Model	Location and dispersion Model
cardio-set					
Gaussian	18382.24	7868.46	9640.05	-1048.50	-454595.31
Logistic	18139.44	7507.51	7923.40	-40057.14	-175284.86
Gamma	23045.92	7862.89	8495.60	-83724.68	-231148.04
Cauchy	46901.44	-19405.01	-6958.12	-5326.16	4115.54
Laplace	24246.39	36791.87	36437.63	24149.97	10616.09
Student T	32956.08	15005.67	16345.02	12318.41	-265663.48
Generalized Logistic	165008.43	141976.78	125162.79	125727.75	112280.18
Generalized Gamma	21838.90	39444.80	15594.83	-29406.53	-45146.45
Power Exponential	111975.45	94657.40	103716.86	107097.20	116843.11
Skew Laplace	57503.56	76490.99	89267.06	89216.90	101480.85
mito-set					
Gaussian	30658.76	15240.19	5237.17	-27597.94	-307404.27
Logistic	28649.61	13033.35	2225.94	-31128.41	-124491.17
Gamma	34117.87	14663.69	3961.87	-69672.27	-176867.88
Cauchy	37319.18	-37837.32	-37859.39	-34370.22	-33289.57
Laplace	30915.07	23449.21	20226.29	17309.75	-132.33
Student T	36105.65	4974.03	-10109.23	-2800.11	-181433.50
Generalized Logistic	84019.64	78326.10	71424.00	71179.51	66433.05
Generalized Gamma	31881.09	40738.63	19279.75	-17367.44	-39927.22
Power Exponential	62594.39	64585.44	71032.84	72868.13	79765.49
Skew Laplace	46471.58	55848.75	64101.82	63933.46	70199.95

Figure S2 Summed AIC value over all probeset models, for each of the model types and per distribution for the run on all distributions; the cardio-set on the left and the mito-set on the right.

Table S6 a) Significant processes according to David / Ease, with p-values < 0.05; based on an analysis allowing for homoscedastic (equal variance / dispersion in all groups) Gaussian models only.

TERM	COUNT	%	PVALUE
<i>TRANSMISSION OF NERVE IMPULSE</i>	9	5.1	0.006178
DNA REPAIR	7	4	0.016478
<i>SYNAPTIC TRANSMISSION</i>	8	4.6	0.017569
REGULATION OF EPIDERMAL GROWTH FACTOR RECEPTOR ACTIVITY	2	1.1	0.039933
INTEGRAL TO GOLGI MEMBRANE	3	1.7	0.04305
GOLGI MEMBRANE	4	2.3	0.046214

Table S6 b) Significant processes according to David / Ease, with p-values < 0.05; based on an analysis allowing for homoscedastic and heteroscedastic (differences in group variance / dispersion) Gaussian models.

TERM	COUNT	%	PVALUE
MYOFIBRIL	11	1.5	3.09E-06
STRIATED MUSCLE CONTRACTION	9	1.2	2.06E-05
ACTIN CYTOSKELETON	29	4	0.000133
FATTY ACID METABOLISM	16	2.2	0.000353
RESPONSE TO PEST/PATHOGEN/PARASITE	35	4.8	0.001457
REGULATION OF MUSCLE CONTRACTION	7	1	0.001694
ENZYME LINKED RECEPTOR PROTEIN SIGNALING PATHWAY	16	2.2	0.002127
COENZYME BIOSYNTHESIS	14	1.9	0.002594
NUCLEAR ORGANIZATION AND BIOGENESIS	20	2.8	0.002788
NUCLEOTIDE BIOSYNTHESIS	14	1.9	0.004037
<i>TRANSMISSION OF NERVE IMPULSE</i>	21	2.9	0.005617
PROTEIN MODIFICATION	77	10.6	0.005954
<i>SYNAPTIC TRANSMISSION</i>	20	2.8	0.008844
FATTY ACID OXIDATION	6	0.8	0.009178
PROTEIN KINASE CASCADE	19	2.6	0.012337
MUSCLE MYOSIN	5	0.7	0.019313
NUCLEOSIDE TRIPHOSPHATE BIOSYNTHESIS	8	1.1	0.019342
NEGATIVE REGULATION OF APOPTOSIS	10	1.4	0.022465
PHOSPHORYLATION	51	7	0.024794
REGULATION OF PROGRAMMED CELL DEATH	18	2.5	0.029534
ATP METABOLISM	7	1	0.031266
CARBOHYDRATE BIOSYNTHESIS	9	1.2	0.031418
NEGATIVE REGULATION OF NUCLEOBASE, NUCLEOSIDE, NUCLEOTIDE AND NUCLEIC ACID METABOLISM	11	1.5	0.034292
PROTEIN COMPLEX ASSEMBLY	10	1.4	0.035868
NUCLEOSIDE TRIPHOSPHATE METABOLISM	8	1.1	0.04176
POSITIVE REGULATION OF I-KAPPAB KINASE/NF-KAPPAB CASCADE	8	1.1	0.04176
INFLAMMATORY RESPONSE	14	1.9	0.049099
REGULATION OF I-KAPPAB KINASE/NF-KAPPAB CASCADE	8	1.1	0.049244

Chapter 5

Multivariate normal probe modeling for Affymetrix expression chip data

L.M.T. Eijssen^{1}, M. Gerards¹, R.G.E. van Eijsder², B. Schroen³, P.J. Lindsey¹*

¹ Department of Population Genetics, Genomics, and Bioinformatics, Maastricht University, P.O. Box 616, 6200 MD, Maastricht, the Netherlands

² Department of Clinical Genetics, Maastricht University, P.O. Box 616, 6200 MD, Maastricht, the Netherlands

³ Department of Molecular Cardiology, Maastricht University, P.O. Box 616, 6200 MD, Maastricht, the Netherlands

**To whom correspondence should be addressed:*

L.M.T. Eijssen, Department of Population Genetics, Genomics, and Bioinformatics, P.O. Box 616, 6200 MD Maastricht, The Netherlands. Tel.: +31-43-3881995. Fax: +31-43-3884573. E-mail: Lars.Eijssen@gen.unimaas.nl

in preparation

Abstract

Several analysis methods have been developed to extract biologically relevant differential genes or pathways from the large data sets resulting from Affymetrix studies. Frequently used methods include the MAS algorithms from Affymetrix, and the alternative RMA procedure. These methods include several steps to get from scanned probe intensities to gene expression levels, such as normalization. Before, we developed a modeling approach that allows incorporating these steps in a one-step approach. We showed how these models can incorporate study design directly to control for disturbing effects and what settings are best to use.

On Affymetrix chips each transcript is recognized by a specific probeset, which consists of a number of different probes. As such, another step taken by methods to analyze Affymetrix data is the summarization of individual probe intensities into common probeset (i.e. gene) intensities. The focus of the current research is to incorporate this step into modeling of probeset expression as well. To model expression of a probeset, the intensities of the separate probes it contains are included into multivariate normal (MVN) models with a random effect. This approach is compared to another approach that first computes a summarized intensity for each probeset and uses this in a subsequent univariate normal (UVN) modeling procedure. To find out what may be gained from the multivariate per-probe approach, two differing Affymetrix Genechip expression Array data sets were selected, considering different chips and diseases. Both model types were built for each of the probesets on the chips and overall results with respect to differentially expressed genes compared. This showed that the multivariate approach retrieved more concise lists of differential genes. This was anticipated as summarizing probe values in a pre-processing step leads to loss of information on between probe variability. This indicated that the genes retrieved are more reliable. However, when evaluating results at the level of biological processes, the multivariate approach still recovered the processes known to be important for the diseases studied and results between both approaches were concordant. The MVN modeling was computationally well-affordable and the number of probesets for which no model could be fitted, was very limited and smaller than for the UVN approach. We conclude that incorporating separate probe intensities into MVN models is a good alternative to summarizing probe intensities and leads to statistically more reliable results.

Keywords

Microarray, differential genes and processes, summarization, probeset

Introduction

Microarray gene expression analysis has become a technique of major importance for retrieving global or specific genetic differences between several groups of biological samples. Amongst these, studies comparing disease samples versus healthy controls form a major group. In this field, several platforms are used, from custom-made cDNA and oligo-arrays to standardized commercial platforms [1]. The commercial Affymetrix Genechip expression Array system is a reproducible and reliable platform [2] that is frequently used. Affymetrix Chips contain probesets measuring the expression levels of several gene transcripts. Each probeset consists of a number of 25mer probes, located at different positions along the sequence of the gene under consideration.

Many analysis methods have been developed to extract biologically relevant differential genes or pathways from the large data sets resulting from microarray studies. Such methods include several steps to get from scanned spots to gene expression levels: combining probe intensities into probeset (gene) intensities, background correction, and normalization. For the Affymetrix platform, analysis methods include the method implemented in the Affymetrix Microarray Suite (MAS5) [3], and a frequently used alternative, the robust multi-array average (RMA) method [4]. These and other techniques differ in the way they combine the scanned probe intensities for each sample into a probeset expression value for each probeset on the chip. For example, MAS5 uses mismatch intensities (probes equal to the intended probe, except for the middle base) to correct perfect match intensities, whilst RMA does not. After these steps, a further technique has to be applied on the resulting intensities to decide which genes or processes are differentially expressed between the groups within the data set. In the current research, an approach is used that models gene expression profiles to determine which genes are differential. The models used intend to combine as much as possible elements of all steps into one step, by adapting to the study design and using several spikes as covariates.

We demonstrated before that modeling the expression of Affymetrix probesets can best be done using the Gaussian distribution with inclusion of a variance effect besides a mean effect (heteroscedastic modeling) [5]. Furthermore it has been shown that the Affymetrix chip description file (CDF) annotations, containing probeset IDs (which gene corresponds to each of the probesets) are not always correct or in accordance with the current builds of the public genome databases [6]. In Dai *et al* updated versions of CDFs for several chip types are created, based on the information present in several public genome databases. These updated CDF files are rebuild every few months and made publicly available [7].

The aim of the current study is to investigate whether the expression of a certain probeset (i.e. gene) can be modeled better when the separate probes it contains are included into multivariate normal (MVN) models with a random effect, instead of first computing a summarized intensity per sample for each probeset and use this in a subsequent univariate normal modeling procedure (further referred to as 'UVN'

modeling). We model the expression profile of each probeset (gene) over all samples, based on the heteroscedastic approach and using the updated CDFs.

To evaluate the effectiveness of both approaches, two differing Affymetrix Genechip expression Array data sets were selected. Different types of chips were run for the two studies, each considering a different human disease and containing three subgroups of samples (two disease groups and a control group). We compare probeset based UVN modeling and probe based MVN modeling and find out what may be gained from a multivariate per-probe approach.

Materials and Methods

Data

Two available human Affymetrix Genechip expression Array data sets from the Genome Center lab at Maastricht University were used.

For a study considering cardiac hypertrophy caused by valvular disease (aorta stenosis), twenty-six samples were run on hgu133a chips, with nine and ten samples in two disease groups with different prognosis (further referred to as 'valv1' and 'valv2') and seven (coronary artery bypass grafting) controls. This data set is further referred to as the 'cardio-set'.

For a study considering two mitochondrial neuromuscular syndromes, MELAS disease and Leigh syndrome [8], twenty samples were run on hgu95av2 chips, with seven and six samples in the two disease groups (further referred to as 'MELAS' and 'Leigh') respectively and six controls. This data set is further referred to as the 'mito-set'.

The data sets, with permuted probesets and masked labels as they represent ongoing research, are available upon request from the first author.

Both studies used two-cycle amplification protocols. All chips contained several spikes, of which the intensities were controlled for by incorporating them as covariates in the models. The spike BioB was excluded from analysis because of its low expression. As they were not used in the experiments under consideration, the spike TrpnX and, when present, sense probes of the spikes were also excluded. Before being used as input into the modeling procedure, expression values were log transformed, to obtain additive models where the group difference represents the fold change between those groups on the original scale.

Chip Description Files

The UniGene based updates of the CDFs for the u133a and u95av2 chips from Dai *et al* [7], version 6, were used. Besides the spikes, the u133a chip contains 11732 updated probesets, the u95 contains 8113 updated probesets, with UniGene-based annotations.

Programming environment

All code was implemented in the R software version 2.0.1 [9], with use of the `glnm` and `growth` libraries [10, 11], and the `affy` library [12].

The analyses were done on 2.8 GHz dual processor servers, having 4 GB of internal memory and operating on linux, Redhat V9 (Shrike) or Fedora Core 3.

Univariate normal reference models

First, for each of the probesets in the updated CDFs, intensity was computed for each sample using the `expresso` function of the `affy` library, with “mas” background correction for the cardio-set and no background correction for the mito-set, constant normalization, no use of mismatch intensities and the “mas” summarization method. These settings were chosen to minimize pre-processing, as all models took the study design into account by incorporating spike covariates to control for slide differences and corrected for several variations directly. Thus the `expresso` function was only used for the purpose of calculating a summarized value for each probeset.

The resulting values were logged and used as response for the heteroscedastic UVN models to fit the expression profile of each probeset. Modeling was done stepwise: first building an intercept, covariate, group mean effect, and group variance effect model, and eventually a model including both mean and variance effects. Spikes were taken into account as covariates and were modeled by selecting the probeset that fitted best to the data at hand for each of them. More elaborate models used the coefficients of the more basic ones as initial parameter estimates. All models were created for each probeset in the data sets separately.

Before building models for all probesets, a run on a subset of 50 probesets was done for both data sets in order to determine the best parameter settings for the computational digit precision and the maximum step to take between each iteration. To improve convergence further in the complete runs, two rounds of iterations were done, a first round with a maximum of 500 steps and a second with a maximum of 100 steps starting from the initial estimates obtained from the first round. For the cardio set, different initial values were used to improve fitting.

The final model for a probeset was obtained by simplifying the model containing all spikes and both a mean and variance effect, removing all non-significant variables. To do so, first unneeded spike covariates were removed from the model, and then the mean and variance effects were simplified if possible.

Multivariate normal models

Modeling was performed according to a similar procedure as for the UVN approach, but without first performing the `expresso` step, using logged probe intensities directly as response instead. To model the correlations between the separate probes, a MVN model including a random effect, was built to fit the expression profile of each probeset. To ensure a consistent procedure for these models, for each of the spikes the probe (instead of probeset) fitting best to the data was selected and included into the model. The models were simplified as described above for the UVN models.

Thus, two complete models, a UVN and a MVN model, were created for each probeset of the two data sets under consideration. In order to be able to optimally compare modeling results, all other aspects of the modeling procedure were kept the same for both analysis types.

Evaluation of the models

Models were rated with the Akaike Information Criterion (AIC) that is computed by summing the minus loglikelihood (-LL) of the model and the number of parameters it contains, yielding lower values for better-fitting and less complex models [13], thus a goodness-of-fit measure penalizing complexity. The penalty of the AIC value was modified (weighed) to match a five percent significance level of a χ^2 test [14]. Model convergence was checked by comparing the -LL values of all models with the more basic ones for the same probeset, that were built during the procedure. The -LL value of a model extended with more variables was required to always be equal to or lower than that of more basic ones to ensure proper convergence. If this was not the case within a small tolerance value (of 1 unit), to account for small random and rounding effects, the -LL value and model coefficients were replaced by those of the more basic model and the AIC recomputed. As an additional check, any model with an infinite parameter value or an infinite AIC, was discarded from further analysis and the number of non-fitting or non-converging models was recorded.

Biological validation

To validate biological relevance of the results obtained for both data sets, the differential genes in each patients group were evaluated using David/Ease version 2.1 (2005 release) [15]. This tool determines whether specific biological pathways or Gene Ontologies (GO) [16] are overrepresented in a set of genes as compared to a proper reference set. Differential processes at level 4 for both the biological process (BP) and cellular component (CC) axes within GO were retrieved. For all differential genes, cut-off values on their ratios of 0.80 and 1.25 (4/5 and 5/4) were used in order to remove small yet statistically significant changes and focus on the most biologically promising ones. For each set of genes, the results obtained from the UVN analysis were compared to the results for the MVN analysis.

Results

On two Affymetrix data sets, a UVN model and a MVN model were fitted for each probeset and the results compared. Optimal settings of several modeling parameters, as determined from a test-run on a subset of genes, were different within each model type for intermediate models including different subsets of variables. The settings for both model types are shown in Supplementary Table S1.

Table 1 shows the numbers of differential genes for both model types and data sets, for each disease group. Standard errors for the group effects could not always be computed, but if the adapted AIC got lower for a model when including a group effect,

the effect could be called significant at that level. As is clear from the Table, the number of genes found to be differentially expressed by the UVN analysis, was much higher than for the MVN analysis. However, there was also a small group of genes for which the opposite was true.

Table 1 Number of significant probesets for both data sets and types of analysis; for the mito-set there were no probesets for which no model could be fitted e.g. because of non-convergence, for the cardio-set this was the case for 652 probesets for the UVN analysis and none for the MVN analysis; results based on AICs computed with penalty correction to match a 5% significance of a χ^2 test and analyses done with cut-off point on the ratios of 0.80 and 1.25, as described in the text; nc = no change. See Supplementary Material, Tables S2 and S3 for results for other settings.

Mito-set (8113 probesets)					
UVN analysis			MVN analysis		
MELAS	Leigh	Number of probesets	MELAS	Leigh	Number of probesets
up	up	286	up	up	35
up	nc	289	up	nc	31
up	down	36	up	down	0
nc	up	807	nc	up	203
nc	nc	5621	nc	nc	7730
nc	down	413	nc	down	102
down	up	69	down	up	0
down	nc	476	down	nc	7
down	down	116	down	down	5
total differential					
1272	1727	2492	78	345	383
Cardio-set (11080 and 11732 probesets)					
UVN analysis			MVN analysis		
Valv1	Valv2	Number of probesets	Valv1	Valv2	Number of probesets
up	up	539	up	up	101
up	nc	314	up	nc	141
up	down	0	up	down	0
nc	up	79	nc	up	15
nc	nc	9664	nc	nc	11470
nc	down	54	nc	down	1
down	up	0	down	up	0
down	nc	209	down	nc	3
down	down	221	down	down	1
total differential					
1283	893	1416	246	118	262

The results are summed in Table 2 to clarify the similarities and differences between the two analyses methods. The fraction of discordant genes between analysis methods (e.g. upregulated in one and downregulated in the other or *vice versa*) as compared to the total number of differential genes in both analysis types was very limited (less than 5%), with the notable exception of the MELAS group in the mito-data, where over a third of differential genes changes direction between both analyses types. When considering genes with similar effects in both analyses (e.g. both upregulated or both downregulated) the effect sizes were generally stronger for the UVN analysis than for the MVN analysis (data not shown) in these data sets for all disease types.

Table 2 Number of significant probesets for both data sets and types of analysis focused on the differences between the two types of analysis; results based on AICs computed with penalty correction to match a 5% significance of a χ^2 test and analyses done with cut-off point on the ratios of 0.80 and 1.25, as described in the text; nc = no change; NA = not available (model did not converge). See Supplementary Material, Tables S4 for results without cut-off points.

MELAS UVN	MELAS MVN	number	Leigh UVN	Leigh MVN	number
up	up	20	up	up	139
up	nc	590	up	nc	1019
up	down	1	up	down	4
nc	up	34	nc	up	96
nc	nc	6799	nc	nc	6236
nc	down	8	nc	down	54
down	up	12	down	up	3
down	nc	646	down	nc	513
down	down	3	down	down	49
NA	up	0	NA	up	0
NA	nc	0	NA	nc	0
NA	down	0	NA	down	0
total: 8113			total: 8113		
Valv1 UVN	Valv1 MVN	number	Valv2 UVN	Valv2 MVN	number
up	up	99	up	up	44
up	nc	754	up	nc	574
up	down	0	up	down	0
nc	up	90	nc	up	50
nc	nc	9704	nc	nc	10137
nc	down	3	nc	down	0
down	up	2	down	up	0
down	nc	427	down	nc	273
down	down	1	down	down	2
NA	up	51	NA	up	22
NA	nc	601	NA	nc	630
NA	down	0	NA	down	0
total: 11732			total: 11732		

The resulting David / Ease output of differential processes for each patient group versus the controls is presented in Tables 3 to 6. As the number of significant genes was much smaller for the MVN analysis, also the number of differential processes retrieved was smaller for this analysis and the associated P-values were lower. Overall, although the lists of specific processes differed for both data sets, processes belonging to the same general categories were present in these lists for the UVN and MVN analyses. For the cardiac data set, processes related to contraction, cytoskeleton, metabolism, cellular junctions and collagens were retrieved for both patient groups. For the mitochondrial data set, muscle-related genes (including development and contraction), complement system, metabolism, cell-cell junctions, apoptosis, and transport were retrieved. In addition, cytoskeleton, cell movement, and mitochondria were found for the Leigh group, and cell proliferation / organization for the MELAS group.

As it includes a random effect, the model fitting procedure took around two times longer for the MVN modeling as compared to the UVN modeling. For the mitochondrial data set, all models could be fitted with the procedure described. For the cardiac data set, this was also the case for the MVN analysis. On the other hand, for the UVN analysis, models for 652 probesets could not be fitted (5.56%) for this data set even after (automated) selection of several starting point.

Table 3a Most significant differential processes by DAVID, for BP / CC level 4, UVN modeling, MELAS group.

SYSTEM	TERM	COUNT	P-VALUE
GOTERM_BP_4	MUSCLE DEVELOPMENT	32	6.7E-08
GOTERM_BP_4	CELLULAR MACROMOLECULE METABOLISM	282	9.67E-07
GOTERM_BP_4	ORGANELLE ORGANIZATION AND BIOGENESIS	77	1.13E-05
GOTERM_BP_4	PROTEIN METABOLISM	259	2.33E-05
GOTERM_BP_4	INTRACELLULAR TRANSPORT	58	2.75E-05
GOTERM_BP_4	CELL CYCLE	84	5.73E-05
GOTERM_BP_4	RNA METABOLISM	54	8.11E-05
GOTERM_BP_4	LIPID METABOLISM	63	0.000154
GOTERM_BP_4	PROGRAMMED CELL DEATH	53	0.000271
GOTERM_BP_4	INTRACELLULAR SIGNALING CASCADE	98	0.000395
GOTERM_BP_4	CELL-CELL ADHESION	26	0.000959
GOTERM_BP_4	CELLULAR LIPID METABOLISM	49	0.001034
GOTERM_CC_4	INTEGRAL TO PLASMA MEMBRANE	124	0.001462
GOTERM_BP_4	DNA METABOLISM	64	0.001583
GOTERM_BP_4	ORGANIC ACID METABOLISM	46	0.001647
GOTERM_BP_4	NUCLEOBASE, NUCLEOSIDE, NUCLEOTIDE AND NUCLEIC ACID METABOLISM	262	0.001864
GOTERM_BP_4	REGULATION OF CELL PROLIFERATION	32	0.00194
GOTERM_CC_4	CONTRACTILE FIBER	10	0.002097
GOTERM_BP_4	AMINE METABOLISM	35	0.002584
GOTERM_BP_4	HEMOPOIESIS	13	0.002974
GOTERM_BP_4	CARBOHYDRATE METABOLISM	47	0.003025
GOTERM_BP_4	MUSCULOSKELETAL MOVEMENT	23	0.003621
GOTERM_BP_4	NEGATIVE REGULATION OF CELLULAR PHYSIOL. PROCESS	33	0.00385

Table 3a (continued)

SYSTEM	TERM	COUNT	P-VALUE
GOTERM_CC_4	ENDOPLASMIC RETICULUM	52	0.004009
GOTERM_BP_4	NEUROGENESIS	43	0.005893
GOTERM_BP_4	POSITIVE REGULATION OF CELLULAR PHYSIOLOGICAL PROCESS	30	0.007957
GOTERM_CC_4	ACTIN CYTOSKELETON	31	0.009616
GOTERM_BP_4	PROTEIN TRANSPORT	46	0.011272
GOTERM_BP_4	ESTABLISHMENT OF PROTEIN LOCALIZATION	46	0.012162
GOTERM_BP_4	AMINO ACID AND DERIVATIVE METABOLISM	29	0.012231
GOTERM_BP_4	BLOOD VESSEL MORPHOGENESIS	10	0.012331
GOTERM_BP_4	VESICLE-MEDIATED TRANSPORT	33	0.014716
GOTERM_BP_4	REGULATION OF MUSCLE CONTRACTION	7	0.014962
GOTERM_BP_4	STRIATED MUSCLE CONTRACTION	7	0.017407
GOTERM_BP_4	CELLULAR BIOSYNTHESIS	91	0.017666
GOTERM_BP_4	RESPONSE TO WOUNDING	37	0.019296
GOTERM_BP_4	RESPONSE TO PEST, PATHOGEN OR PARASITE	47	0.019422
GOTERM_BP_4	REGULATION OF PROGRAMMED CELL DEATH	28	0.024889
GOTERM_BP_4	ALCOHOL METABOLISM	26	0.027014
GOTERM_CC_4	CYTOSOL	35	0.029085
GOTERM_BP_4	RESPONSE TO DNA DAMAGE STIMULUS	23	0.03329
GOTERM_BP_4	RESPONSE TO EXTERNAL BIOTIC STIMULUS	48	0.033525
GOTERM_CC_4	INTEGRAL TO NUCLEAR INNER MEMBRANE	3	0.034573
GOTERM_CC_4	NUCLEAR MEMBRANE	13	0.034955
GOTERM_CC_4	VACUOLE	16	0.035842
GOTERM_BP_4	HEMOSTASIS	13	0.03802
GOTERM_BP_4	POLYSACCHARIDE METABOLISM	8	0.038166
GOTERM_CC_4	SPLICEOSOME COMPLEX	9	0.03953
GOTERM_BP_4	SECRETORY PATHWAY	16	0.048041
GOTERM_BP_4	REGULATION OF BIOSYNTHESIS	14	0.049812
GOTERM_CC_4	DYSTROPHIN-ASSOCIATED GLYCOPROTEIN COMPLEX	4	0.052148
GOTERM_BP_4	ACUTE-PHASE RESPONSE	7	0.052273
GOTERM_CC_4	ALPHA-KETOGLUTARATE DEHYDROGENASE COMPLEX (SENSU EUKARYOTA)	3	0.054591
GOTERM_BP_4	REGULATION OF PROTEIN METABOLISM	10	0.05612
GOTERM_BP_4	SULFUR METABOLISM	8	0.060246
GOTERM_CC_4	GOLGI APPARATUS	41	0.06386
GOTERM_BP_4	B-CELL DIFFERENTIATION	4	0.073283
GOTERM_CC_4	TRICARBOXYLIC ACID CYCLE ENZYME COMPLEX	3	0.077612
GOTERM_BP_4	IMMUNE CELL ACTIVATION	11	0.079808
GOTERM_CC_4	PROTEASOME COMPLEX (SENSU EUKARYOTA)	7	0.079846
GOTERM_BP_4	TRANSPORT	172	0.080125
GOTERM_BP_4	CYTOPLASM ORGANIZATION AND BIOGENESIS	9	0.092236
GOTERM_CC_4	SMALL NUCLEOLAR RIBONUCLEOPROTEIN COMPLEX	5	0.094561
GOTERM_BP_4	NEGATIVE REGULATION OF COAGULATION	4	0.099372

Table 3b Most significant differential processes by DAVID, for BP / CC level 4, MVN modeling, MELAS group.

SYSTEM	TERM	COUNT	P-VALUE
GOTERM_BP_4	HUMORAL IMMUNE RESPONSE	6	0.000544
GOTERM_BP_4	MUSCLE DEVELOPMENT	5	0.003793
GOTERM_BP_4	RESPONSE TO PEST, PATHOGEN OR PARASITE	7	0.014900
GOTERM_BP_4	RESPONSE TO EXTERNAL BIOTIC STIMULUS	7	0.019563
GOTERM_BP_4	REGULATION OF CELL PROLIFERATION	5	0.021914
GOTERM_BP_4	ORGANIC ACID METABOLISM	6	0.027572
GOTERM_BP_4	NEGATIVE REGULATION OF CELLULAR PHYSIOLOGICAL PROCESS	5	0.028830
GOTERM_BP_4	BIOPOLYMER CATABOLISM	7	0.053411
GOTERM_BP_4	ORGANELLE ORGANIZATION AND BIOGENESIS	7	0.056879
GOTERM_BP_4	CELLULAR CATABOLISM	8	0.068387
GOTERM_BP_4	SKELETAL DEVELOPMENT	3	0.089297

Table 4a Most significant differential processes by DAVID, for BP / CC level 4, UVN modeling, Leigh disease group.

SYSTEM	TERM	COUNT	P-VALUE
GOTERM_CC_4	INTEGRAL TO PLASMA MEMBRANE	198	8.81E-09
GOTERM_BP_4	NEGATIVE REGULATION OF CELLULAR PHYSIOLOGICAL PROCESS	58	2.33E-08
GOTERM_BP_4	NEUROGENESIS	74	5.18E-08
GOTERM_BP_4	CELL-CELL ADHESION	43	5.63E-08
GOTERM_BP_4	REGULATION OF CELL PROLIFERATION	51	7.14E-07
GOTERM_BP_4	MUSCLE DEVELOPMENT	34	6.52E-06
GOTERM_BP_4	CELLULAR MACROMOLECULE METABOLISM	361	2.06E-05
GOTERM_BP_4	RESPONSE TO PEST, PATHOGEN OR PARASITE	74	4.7E-05
GOTERM_BP_4	INFLAMMATORY RESPONSE	37	6.97E-05
GOTERM_BP_4	CELL CYCLE	107	8.18E-05
GOTERM_BP_4	INTRACELLULAR SIGNALING CASCADE	131	8.96E-05
GOTERM_BP_4	PROTEIN METABOLISM	337	8.99E-05
GOTERM_BP_4	INTRACELLULAR TRANSPORT	71	9.74E-05
GOTERM_BP_4	PROGRAMMED CELL DEATH	69	0.000108
GOTERM_BP_4	RESPONSE TO EXTERNAL BIOTIC STIMULUS	76	0.000115
GOTERM_BP_4	BLOOD VESSEL MORPHOGENESIS	16	0.000122
GOTERM_CC_4	ACTIN CYTOSKELETON	48	0.000131
GOTERM_CC_4	PROTEASOME COMPLEX (SENSU EUKARYOTA)	14	0.000183
GOTERM_BP_4	POSITIVE REGULATION OF CELLULAR PHYSIOLOGICAL PROCESS	44	0.000225
GOTERM_BP_4	RESPONSE TO WOUNDING	56	0.000321
GOTERM_BP_4	NEGATIVE REGULATION OF METABOLISM	31	0.000329
GOTERM_BP_4	ORGANELLE ORGANIZATION AND BIOGENESIS	91	0.000348
GOTERM_BP_4	REGULATION OF PROGRAMMED CELL DEATH	42	0.0009
GOTERM_BP_4	HEMOPOIESIS	16	0.002034
GOTERM_BP_4	REGULATION OF SIGNAL TRANSDUCTION	29	0.002189
GOTERM_BP_4	CELLULAR BIOSYNTHESIS	127	0.002476
GOTERM_BP_4	AMINE METABOLISM	44	0.003101
GOTERM_BP_4	ORGANIC ACID METABOLISM	57	0.003672
GOTERM_BP_4	VESICLE-MEDIATED TRANSPORT	45	0.003945
GOTERM_BP_4	PROTEIN TRANSPORT	61	0.006095

Table 4a (continued)

SYSTEM	TERM	COUNT	P-VALUE
GOTERM_BP_4	ESTABLISHMENT OF PROTEIN LOCALIZATION	61	0.00664
GOTERM_BP_4	AMINO ACID AND DERIVATIVE METABOLISM	38	0.006973
GOTERM_CC_4	ENDOPLASMIC RETICULUM	68	0.007028
GOTERM_BP_4	POLYSACCHARIDE METABOLISM	11	0.009818
GOTERM_BP_4	HEMOSTASIS	18	0.01006
GOTERM_BP_4	CELLULAR CATABOLISM	104	0.011583
GOTERM_BP_4	REGULATION OF PROTEIN METABOLISM	14	0.015006
GOTERM_CC_4	VACUOLE	22	0.016235
GOTERM_BP_4	NUCLEOBASE, NUCLEOSIDE, NUCLEOTIDE AND NUCLEIC ACID METABOLISM	337	0.017579
GOTERM_CC_4	CYTOSOL	48	0.018306
GOTERM_BP_4	HUMORAL IMMUNE RESPONSE	25	0.018358
GOTERM_BP_4	REGULATION OF I-KAPPAB KINASE/NF-KAPPAB CASCADE	15	0.018427
GOTERM_BP_4	CELL MIGRATION	13	0.020472
GOTERM_CC_4	CONTRACTILE FIBER	10	0.020766
GOTERM_BP_4	MACROMOLECULE CATABOLISM	91	0.024106
GOTERM_BP_4	TRANSPORT	238	0.025322
GOTERM_BP_4	REGULATION OF CELL ORGANIZATION AND BIOGENESIS	8	0.026166
GOTERM_BP_4	CELL MATURATION	5	0.032272
GOTERM_BP_4	NERVE MATURATION	5	0.032272
GOTERM_BP_4	POSITIVE REGULATION OF SIGNAL TRANSDUCTION	15	0.034473
GOTERM_BP_4	LIPID METABOLISM	67	0.035639
GOTERM_BP_4	SULFUR METABOLISM	10	0.047428
GOTERM_CC_4	PLASTID	5	0.052181
GOTERM_BP_4	DNA METABOLISM	72	0.057086
GOTERM_CC_4	CYTOSKELETON	107	0.059759
GOTERM_BP_4	COFACTOR METABOLISM	26	0.065576
GOTERM_BP_4	ALCOHOL METABOLISM	31	0.069308
GOTERM_CC_4	REPLICATION FORK	6	0.069661
GOTERM_CC_4	LAMELLIPODIUM	4	0.073835
GOTERM_BP_4	RESPONSE TO UNFOLDED PROTEIN	10	0.076418
GOTERM_BP_4	ORGANIC ACID TRANSPORT	12	0.080842
GOTERM_BP_4	BIOPOLYMER BIOSYNTHESIS	5	0.087233
GOTERM_BP_4	CELLULAR LIPID METABOLISM	51	0.090189
GOTERM_BP_4	PHOSPHORUS METABOLISM	114	0.098179
GOTERM_BP_4	NEUROTRANSMITTER TRANSPORT	8	0.099595

Table 4b Most significant differential processes by DAVID, for BP / CC level 4, MVN modeling, Leigh disease group.

SYSTEM	TERM	COUNT	P-VALUE
GOTERM_CC_4	ACTIN CYTOSKELETON	21	2.29E-07
GOTERM_BP_4	MUSCLE DEVELOPMENT	15	1.24E-06
GOTERM_CC_4	CYTOSKELETON	35	0.000109
GOTERM_BP_4	STRIATED MUSCLE CONTRACTION	6	0.000244
GOTERM_BP_4	HUMORAL IMMUNE RESPONSE	11	0.000963
GOTERM_CC_4	CONTRACTILE FIBER	6	0.001147
GOTERM_BP_4	CELL MIGRATION	7	0.001478
GOTERM_BP_4	CELLULAR MORPHOGENESIS	10	0.003771
GOTERM_BP_4	NEGATIVE REGULATION OF METABOLISM	10	0.003771
GOTERM_BP_4	CELLULAR CATABOLISM	29	0.003955
GOTERM_CC_4	INTRACELLULAR NON-MEMBRANE-BOUND ORGANELLE	44	0.003976
GOTERM_BP_4	NEGATIVE REGULATION OF CELLULAR PHYSIOLOGICAL PROCESS	13	0.006777
GOTERM_BP_4	RESPONSE TO PEST, PATHOGEN OR PARASITE	18	0.010271
GOTERM_BP_4	REGULATION OF PROTEIN METABOLISM	6	0.010612
GOTERM_BP_4	PROGRAMMED CELL DEATH	17	0.012391
GOTERM_CC_4	BASAL LAMINA	4	0.013056
GOTERM_BP_4	CELL ION HOMEOSTASIS	6	0.016026
GOTERM_BP_4	RESPONSE TO EXTERNAL BIOTIC STIMULUS	18	0.017823
GOTERM_BP_4	CELLULAR MACROMOLECULE METABOLISM	75	0.018136
GOTERM_BP_4	REGULATION OF PROGRAMMED CELL DEATH	11	0.026192
GOTERM_BP_4	REGULATION OF CELL SIZE	7	0.028988
GOTERM_BP_4	REGULATION OF CELL GROWTH	6	0.030562
GOTERM_BP_4	MACROMOLECULE CATABOLISM	23	0.03443
GOTERM_BP_4	REGULATION OF MYOBLAST DIFFERENTIATION	2	0.035619
GOTERM_BP_4	REGULATION OF CELL MOTILITY	3	0.036634
GOTERM_BP_4	ORGANELLE ORGANIZATION AND BIOGENESIS	20	0.046225
GOTERM_BP_4	REGULATION OF CELL PROLIFERATION	10	0.054196
GOTERM_BP_4	CELLULAR BIOSYNTHESIS	28	0.055456
GOTERM_BP_4	REGULATION OF HEART CONTRACTION RATE	3	0.058701
GOTERM_BP_4	PROTEIN METABOLISM	67	0.070235
GOTERM_CC_4	MITOCHONDRIAL MEMBRANE	11	0.086423
GOTERM_BP_4	REGULATION OF EMBRYONIC DEVELOPMENT	2	0.086691
GOTERM_BP_4	MACROMOLECULE BIOSYNTHESIS	18	0.091051
GOTERM_CC_4	MITOCHONDRIAL INNER MEMBRANE	10	0.095375

Table 5a Most significant differential processes by DAVID, for BP / CC level 4, UVN modeling, Valv1 group.

SYSTEM	TERM	COUNT	P-VALUE
GOTERM_BP_4	CELLULAR MACROMOLECULE METABOLISM	261	9.58E-07
GOTERM_BP_4	PROTEIN METABOLISM	236	7.42E-05
GOTERM_BP_4	PROGRAMMED CELL DEATH	51	0.000112
GOTERM_BP_4	INTRACELLULAR SIGNALING CASCADE	93	0.000163
GOTERM_BP_4	VESICLE-MEDIATED TRANSPORT	36	0.000617
GOTERM_BP_4	MUSCLE DEVELOPMENT	22	0.00077
GOTERM_BP_4	POSITIVE REGULATION OF CELLULAR PHYSIOLOGICAL PROCESS	31	0.00117
GOTERM_BP_4	REGULATION OF PROGRAMMED CELL DEATH	31	0.001244
GOTERM_CC_4	CONTRACTILE FIBER	9	0.003975
GOTERM_CC_4	CYTOSKELETON	77	0.006387
GOTERM_BP_4	PROTEIN TRANSPORT	44	0.00644
GOTERM_BP_4	ESTABLISHMENT OF PROTEIN LOCALIZATION	44	0.007129
GOTERM_BP_4	HEMOSTASIS	14	0.008851
GOTERM_BP_4	CARBOHYDRATE METABOLISM	41	0.012715
GOTERM_BP_4	RESPONSE TO WOUNDING	35	0.01516
GOTERM_BP_4	INTRACELLULAR TRANSPORT	43	0.016712
GOTERM_CC_4	EXTRINSIC TO PLASMA MEMBRANE	10	0.017054
GOTERM_BP_4	REGULATION OF SIGNAL TRANSDUCTION	19	0.020304
GOTERM_CC_4	ACTIN CYTOSKELETON	27	0.021865
GOTERM_BP_4	SECRETORY PATHWAY	16	0.024706
GOTERM_BP_4	REGULATION OF G-PROTEIN COUPLED RECEPTOR PROTEIN SIGNALING PATHWAY	6	0.025239
GOTERM_BP_4	RNA METABOLISM	40	0.026286
GOTERM_BP_4	CELLULAR LIPID METABOLISM	39	0.030438
GOTERM_BP_4	NEGATIVE REGULATION OF CELLULAR PHYSIOLOGICAL PROCESS	27	0.034558
GOTERM_BP_4	VITAMIN TRANSPORT	4	0.040646
GOTERM_CC_4	GOLGI APPARATUS	38	0.047953
GOTERM_BP_4	MACROMOLECULE CATABOLISM	62	0.048238
GOTERM_BP_4	ORGANIC ACID METABOLISM	36	0.048718
GOTERM_BP_4	NUCLEOBASE, NUCLEOSIDE, NUCLEOTIDE AND NUCLEIC ACID METABOLISM	225	0.058253
GOTERM_BP_4	POSITIVE REGULATION OF CELL DIFFERENTIATION	4	0.059106
GOTERM_BP_4	RESPONSE TO UNFOLDED PROTEIN	8	0.061205
GOTERM_BP_4	CELLULAR CATABOLISM	68	0.061347
GOTERM_BP_4	POSITIVE REGULATION OF CASPASE ACTIVITY	5	0.062236
GOTERM_BP_4	POLYSACCHARIDE METABOLISM	7	0.069676
GOTERM_BP_4	PHOSPHORUS METABOLISM	80	0.070959
GOTERM_BP_4	CELLULAR BIOSYNTHESIS	79	0.071808
GOTERM_BP_4	BIOPOLYMER CATABOLISM	52	0.078875
GOTERM_BP_4	SKELETAL DEVELOPMENT	13	0.082094
GOTERM_BP_4	LIPID METABOLISM	45	0.082805
GOTERM_BP_4	TRANSPORT	157	0.098642
GOTERM_CC_4	MEMBRANE COAT	7	0.099691

Table 5b Most significant differential processes by DAVID, for BP / CC level 4, MVN modeling, Valv1 group.

SYSTEM	TERM	COUNT	P-VALUE
GOTERM_CC_4	CYTOSKELETON	29	8.304E-06
GOTERM_CC_4	ACTIN CYTOSKELETON	15	8.907E-06
GOTERM_BP_4	MUSCLE DEVELOPMENT	11	2.062E-05
GOTERM_BP_4	CELLULAR MACROMOLECULE METABOLISM	63	3.519E-05
GOTERM_BP_4	PROTEIN METABOLISM	54	0.0019166
GOTERM_CC_4	CONTRACTILE FIBER	5	0.0019269
GOTERM_CC_4	INTRACELLULAR NON-MEMBRANE-BOUND ORGANELLE	33	0.0020156
GOTERM_BP_4	REGULATION OF HEART CONTRACTION RATE	4	0.0020931
GOTERM_BP_4	MACROMOLECULE BIOSYNTHESIS	18	0.0021517
GOTERM_BP_4	CARBOHYDRATE METABOLISM	14	0.0028989
GOTERM_BP_4	RESPONSE TO UNFOLDED PROTEIN	5	0.0047962
GOTERM_BP_4	REGULATION OF MUSCLE CONTRACTION	4	0.0056544
GOTERM_BP_4	CELLULAR BIOSYNTHESIS	23	0.0090086
GOTERM_BP_4	REGULATION OF BIOSYNTHESIS	6	0.0128287
GOTERM_BP_4	INTRACELLULAR TRANSPORT	13	0.013596
GOTERM_BP_4	CELLULAR MORPHOGENESIS	7	0.0157498
GOTERM_BP_4	BIOPOLYMER BIOSYNTHESIS	3	0.0206558
GOTERM_BP_4	PROTEIN TRANSPORT	12	0.025982
GOTERM_BP_4	ESTABLISHMENT OF PROTEIN LOCALIZATION	12	0.0266807
GOTERM_CC_4	MEMBRANE COAT	4	0.027213
GOTERM_CC_4	PYRUVATE DEHYDROGENASE COMPLEX	2	0.02883
GOTERM_BP_4	LIPID METABOLISM	13	0.0428714
GOTERM_BP_4	CYTOKINESIS	5	0.050693
GOTERM_BP_4	STRIATED MUSCLE CONTRACTION	3	0.054266
GOTERM_BP_4	MACROMOLECULE CATABOLISM	16	0.0544456
GOTERM_BP_4	REGULATION OF CELL ADHESION	3	0.0604565
GOTERM_BP_4	NEUROGENESIS	10	0.0627048
GOTERM_BP_4	REGULATION OF CELL SIZE	5	0.0660058
GOTERM_BP_4	AMINE METABOLISM	8	0.0730376
GOTERM_CC_4	COLLAGEN TYPE IV	2	0.0840403
GOTERM_BP_4	VESICLE-MEDIATED TRANSPORT	8	0.0864376
GOTERM_BP_4	CELLULAR LIPID METABOLISM	10	0.090119
GOTERM_BP_4	REGULATION OF JAK-STAT CASCADE	2	0.0904015
GOTERM_CC_4	SHEET-FORMING COLLAGEN	2	0.0973494
GOTERM_BP_4	AMINO ACID AND DERIVATIVE METABOLISM	7	0.0981783

Table 6a Most significant differential processes by DAVID, for BP / CC level 4, UVN modeling, Valv2 group.

SYSTEM	TERM	COUNT	P-VALUE
GOTERM_BP_4	CELLULAR MACROMOLECULE METABOLISM	183	2.84E-05
GOTERM_BP_4	PROTEIN METABOLISM	174	2.89E-05
GOTERM_BP_4	REGULATION OF PROGRAMMED CELL DEATH	27	9.94E-05
GOTERM_CC_4	CYTOSKELETON	66	0.000118
GOTERM_BP_4	PROGRAMMED CELL DEATH	37	0.000554
GOTERM_BP_4	POSITIVE REGULATION OF CELLULAR PHYSIOLOGICAL PROCESS	25	0.000563
GOTERM_BP_4	ORGANELLE ORGANIZATION AND BIOGENESIS	46	0.003094
GOTERM_BP_4	INTRACELLULAR SIGNALING CASCADE	63	0.003861
GOTERM_BP_4	HEMOSTASIS	12	0.003889
GOTERM_BP_4	PROTEIN TRANSPORT	34	0.004028
GOTERM_BP_4	PHOSPHORUS METABOLISM	66	0.004107
GOTERM_BP_4	ESTABLISHMENT OF PROTEIN LOCALIZATION	34	0.004305
GOTERM_BP_4	INTRACELLULAR TRANSPORT	34	0.006209
GOTERM_BP_4	MUSCLE DEVELOPMENT	15	0.008511
GOTERM_BP_4	VESICLE-MEDIATED TRANSPORT	24	0.010231
GOTERM_BP_4	REGULATION OF SIGNAL TRANSDUCTION	15	0.016814
GOTERM_BP_4	TRANSPORT	119	0.018913
GOTERM_CC_4	ACTIN CYTOSKELETON	21	0.021949
GOTERM_CC_4	MICROTUBULE CYTOSKELETON	19	0.0223
GOTERM_BP_4	REGULATION OF G-PROTEIN COUPLED RECEPTOR PROTEIN SIGNALING PATHWAY	5	0.029087
GOTERM_BP_4	NEGATIVE REGULATION OF METABOLISM	14	0.029955
GOTERM_CC_4	INTRACELLULAR NON-MEMBRANE-BOUND ORGANELLE	83	0.038532
GOTERM_BP_4	RNA METABOLISM	29	0.038926
GOTERM_BP_4	REGULATION OF MUSCLE CONTRACTION	5	0.040497
GOTERM_BP_4	REGULATION OF CELL ACTIVATION	5	0.049325
GOTERM_BP_4	CELLULAR LIPID METABOLISM	28	0.050012
GOTERM_BP_4	RESPONSE TO WOUNDING	24	0.053399
GOTERM_BP_4	CELLULAR MORPHOGENESIS	13	0.05973
GOTERM_CC_4	EXTRINSIC TO PLASMA MEMBRANE	7	0.065848
GOTERM_BP_4	NUCLEOBASE, NUCLEOSIDE, NUCLEOTIDE AND NUCLEIC ACID METABOLISM	159	0.068253
GOTERM_BP_4	CELLULAR BIOSYNTHESIS	57	0.073491
GOTERM_CC_4	MEMBRANE COAT	6	0.078963
GOTERM_BP_4	COFACTOR METABOLISM	14	0.085588
GOTERM_BP_4	POSITIVE REGULATION OF METABOLISM	8	0.086835
GOTERM_BP_4	POSITIVE REGULATION OF CASPASE ACTIVITY	4	0.087212
GOTERM_BP_4	REGULATION OF IMMUNE RESPONSE	6	0.09283

Table 6b Most significant differential processes by DAVID, for BP / CC level 4, MVN modeling, Valv2 group.

SYSTEM	TERM	COUNT	P-VALUE
GOTERM_CC_4	CYTOSKELETON	17	8.68E-05
GOTERM_BP_4	MUSCLE DEVELOPMENT	7	0.000335
GOTERM_BP_4	CARBOHYDRATE METABOLISM	10	0.001246
GOTERM_BP_4	CELLULAR MORPHOGENESIS	6	0.003106
GOTERM_BP_4	CELLULAR MACROMOLECULE METABOLISM	31	0.004409
GOTERM_CC_4	ACTIN CYTOSKELETON	7	0.00642
GOTERM_BP_4	MACROMOLECULE CATABOLISM	12	0.006702
GOTERM_CC_4	INTRACELLULAR NON-MEMBRANE-BOUND ORGANELLE	18	0.00797
GOTERM_BP_4	REGULATION OF PROTEIN METABOLISM	4	0.008993
GOTERM_BP_4	NEGATIVE REGULATION OF SIGNAL TRANSDUCTION	3	0.009808
GOTERM_BP_4	CELLULAR CATABOLISM	12	0.015702
GOTERM_BP_4	REGULATION OF CELL ADHESION	3	0.016433
GOTERM_BP_4	BIOPOLYMER CATABOLISM	10	0.017158
GOTERM_BP_4	REGULATION OF CELL GROWTH	4	0.018611
GOTERM_BP_4	LIPID METABOLISM	9	0.019257
GOTERM_CC_4	CONTRACTILE FIBER	3	0.027513
GOTERM_BP_4	REGULATION OF CELL SIZE	4	0.03872
GOTERM_CC_4	COLLAGEN TYPE IV	2	0.041363
GOTERM_BP_4	PROTEIN METABOLISM	26	0.043301
GOTERM_BP_4	RESPONSE TO UNFOLDED PROTEIN	3	0.044309
GOTERM_BP_4	REGULATION OF JAK-STAT CASCADE	2	0.045335
GOTERM_CC_4	SHEET-FORMING COLLAGEN	2	0.048092
GOTERM_BP_4	NEGATIVE REGULATION OF METABOLISM	4	0.076568
GOTERM_BP_4	REGULATION OF WNT RECEPTOR SIGNALING PATHWAY	2	0.077998
GOTERM_BP_4	REGULATION OF PHOSPHORUS METABOLISM	2	0.083333
GOTERM_BP_4	NEUROGENESIS	6	0.09804

Discussion

Comparing the results of both UVN and MVN modeling with respect to differential genes and differential processes gave several insights into the characteristics of both model types. First of all, as the total number of genes that was found to be differentially expressed was much higher in the UVN analysis than in the MVN analysis, this showed that the inclusion of between-probe variability into the model led to many probesets not being proven significantly differentially expressed anymore. As the UVN loses probe variability information by its stepwise approach of concatenating first and then only using a summarized value, this analysis would be expected to be overconfident. The MVN analysis on the other hand, uses a one-step approach and does not suffer from this drawback. Instead, it directly used the observed data and took the probe variance into account. Of course, the number of genes retrieved by the UVN analysis will also get smaller if a stricter significance level is used, but characterizing variables and variance better would be more correct than and preferred above summarizing and being more strict, assuming that no information had been lost doing so. Furthermore, the non self-evident decision of

what summarizing method to use to get to one value for each probeset does not have to be taken for the MVN analysis and will not influence results. Besides the effect of probe variability, there may also be an influence of the random effect variance that is only present in the multivariate model.

For both analysis types, occurrence of false positives was further reduced by considering results at a pathway level. False positives could be assumed to occur randomly over all biological processes and to be filtered out by this approach. For this reason a penalty score matching a non-stringent five percent significance level of a χ^2 test was used in this study. Furthermore, the AIC was chosen as a general criterion to rate the models for this explorative study, since statistical tests are not well-suited because they are done appropriate for what specific assumptions are made with respect to the data. A further advantage of the AIC is that it does not suffer from multiple testing, as techniques like t-tests and FDR do, and no correction needs to be applied.

In addition to the many genes that were not significant anymore in the MVN analysis, a much smaller group of non-differential genes in the UVN analysis got differential in the MVN analysis. This could be explained by the fact that a MVN analysis has a larger sample size available, as it uses more data points for each chip, where the UVN analysis only uses one measurement per chip. This implies that for those probesets where group differences in expression were in fact large enough in comparison to probe variability, smaller effects could be estimated.

For the genes that were differential in both analyses types, most showed concordant effects. Evaluation of several genes that were discordant between both analysis types showed that the effect estimated with the MVN modeling and not the UVN modeling was concordant to the effect computed from the uncorrected raw data both at probe and probeset levels. The fact that probe and probeset data showed the same effect indicated that the summarization method did not cause these differences, but the modeling itself. Maybe the fit of the UVN model was sometimes worse, and it tried to make up for this by adding many covariates. Indeed the number of spikes in the UVN models was much higher than in the MVN models (data not shown). Besides, also spike probeset expression estimates might have been less reliable for the UVN analysis as the variability of these probes was not taken into account as well. Finally, that the MVN analysis took into account the design of the study in a better way might add to the correctness of that analysis type and improve results.

Because the number of false positive differential genes decreased for a MVN analysis, also the number of processes retrieved by it decreased and got more reliable. However, at a process level the influence of false positives, and as such the difference between the data sets, would be mitigated by random occurrence of false positives over the processes in contrast to the real positives. These would be filtered out by the common pathway approach. Indeed, despite the decrease in number of processes for the MVN analysis as compared to the UVN analysis, the more general categories they belonged to remained more or less the same. It must be noted that for the heteroscedastic UVN analysis as it was done now, the retrieved differential

processes for all patient groups versus the controls, were already very biologically plausible. We showed before [5] that by allowing the modeling to be more flexible compared to a standard homoscedastic modeling approach, the list of biological processes retrieved got more extensive and that plausible biological processes apparently lost in a homoscedastic approach were retrieved again. Although, when going from a UVN analysis to a MVN analysis, the list of retrieved processes generally got shorter again, the processes appeared to remain very relevant for the diseases under study in this case. As expected, the MVN analysis did not introduce extra processes completely different from the ones in the UVN modeling. Thus the major effect of switching to a MVN approach is focusing on the most reliable subset of the processes retrieved by the UVN approach. However, as the general categories of processes retrieved by both analyses were the same, the results did not infer that a UVN analysis is not useful at all. As indicated, the influence of the less optimal determination of differential genes by the UVN analysis may be counterbalanced by focusing on common pathways. The UVN analysis runs faster, needs less processing power and does not require the use of a package allowing for multivariate analyses, and one might consider using a UVN analysis if any of those factors is critical against the price of a higher chance on false positive results. It must be stated though, that demands on time or computational power were not detected to be problematic for any of the two modeling approaches in any of the two test sets used, thus the advantage would be of minor importance. Furthermore, some more factors show the better suitability of a MVN analysis as compared to a UVN analysis. These include the easier fit of the models (less non-fitting or non-converging models and less tuning of initial estimates needed) and the lesser sensitivity within each data set to the initial modeling parameters chosen for the computational precision and the maximum stepsize to take between iterations. Moreover, the updated CDF files by Dai et al [7] were used, for which the probesets are already cleaned from non-specific or incorrect probes, making them more uniform. It will be likely that a MVN analysis improves results more when the original Affymetrix probesets are used. Though both analyses types will be affected by less consistent probe sets, the UVN will take again the higher risk of being overconfident.

To test the influence of overconfidence and the effect of switching from a UVN to a MVN analysis at the level of specific genes instead of processes, the differential genes with annotations containing 'muscle' were selected for the MELAS and Leigh disease samples as a representative example and shown in Table 7 and Supplementary Table S5. It could be stated that the extra genes found in the UVN analysis were plausible as well, supporting that a pathway approach may really separate the real differential genes from the false positives. The list of differential genes was also determined for a relaxed version of the MVN analysis, using a standard AIC and without the originally used cut-off points on the ratios of 0.80 and 1.25 (see same Tables). Even with these settings fewer genes were detected to be differential as compared to the UVN analysis with adapted AIC and with cut-offs. This

stressed once more the effect of taking into account the variability between the probes.

Of course, when these techniques are applied in a study with a more biological focus, it would be of special interest to focus on differences and similarities between the patient groups within each study and on the effect of specific genes or processes. However, as in this study results were only collected to estimate the quality of the two analysis types with respect to processes known or likely to be involved, and it is not yet known what the specific genetic effects or differences between patient groups should be, these factors were not further investigated here.

Table 7a Differential genes for MELAS disease that contain the word ‘muscle’ in their GO annotations at the biological process or the cellular component axis for the UVN analysis; results based on AICs computed with penalty correction to match a 5% significance of a χ^2 test and analyses done with cut-off point on the ratios of 0.80 and 1.25.

Hs.104672	downregulated in ovarian cancer 1
Hs.118845	troponin C, slow
Hs.134602	nebulin
Hs.134602	nebulin
Hs.140950	EF hand calcium binding protein 2
Hs.160562	insulin-like growth factor 1 (somatomedin C)
Hs.172684	vesicle-associated membrane protein 5 (myobrevin)
Hs.178023	myogenic factor 5
Hs.181768	myogenic factor 3
Hs.182421	troponin C2, fast
Hs.201379	phosphorylase kinase, alpha 1 (muscle)
Hs.20516	histone deacetylase 4
Hs.300772	tropomyosin 2 (beta)
Hs.336994	metastasis suppressor 1
Hs.410970	myosin, light polypeptide 5, regulatory
Hs.438782	histone deacetylase 5
Hs.438953	sarcoglycan, beta (43kDa dystrophin-associated glycoprotein)
Hs.439463	AE binding protein 1
Hs.476448	filamin B, beta (actin binding protein 278)
Hs.483454	calponin 3, acidic
Hs.490203	caldesmon 1
Hs.49582	protein phosphatase 1, regulatory (inhibitor) subunit 12A
Hs.495912	dystrophin (muscular dystrophy, Duchenne and Becker types)
Hs.500483	actin, alpha 2, smooth muscle, aorta
Hs.503998	transgelin
Hs.506502	myosin binding protein C, slow type
Hs.513490	aldolase A, fructose-bisphosphate
Hs.521653	muscle, skeletal, receptor tyrosine kinase
Hs.524488	diacylglycerol kinase, alpha 80kDa
Hs.534028	myosin, heavy polypeptide 8, skeletal muscle, perinatal
Hs.73454	troponin T3, skeletal, fast
Hs.75535	myosin, light polypeptide 2, regulatory, cardiac, slow
Hs.7879	interferon-related developmental regulator 1
Hs.927	myosin binding protein H
Hs.98303	caveolin 3

Table 7b Differential genes for MELAS disease that contain the word ‘muscle’ in their GO annotations at either the biological process or the cellular component axis for the MVN analysis; results based on AICs computed with penalty correction to match a 5% significance of a χ^2 test and analyses done with cut-off point on the ratios of 0.80 and 1.25.

Hs.118127	actin, alpha, cardiac muscle
Hs.200841	laminin, alpha 2 (merosin, congenital muscular dystrophy)
Hs.440895	myosin, heavy polypeptide 3, skeletal muscle, embryonic
Hs.534028	myosin, heavy polypeptide 8, skeletal muscle, perinatal
Hs.7879	interferon-related developmental regulator 1

Table 7c Differential genes for MELAS disease that contain the word ‘muscle’ in their GO annotations at either the biological process or the cellular component axis for the MVN analysis; results based on AIC values with standard penalty and analyses done without cut-off points on the ratios.

Hs.118127	actin, alpha, cardiac muscle
Hs.196054	histone deacetylase 9
Hs.200841	laminin, alpha 2 (merosin, congenital muscular dystrophy)
Hs.284244	fibroblast growth factor 2 (basic)
Hs.410970	myosin, light polypeptide 5, regulatory
Hs.440895	myosin, heavy polypeptide 3, skeletal muscle, embryonic
Hs.465929	calponin 1, basic, smooth muscle
Hs.483454	calponin 3, acidic
Hs.491359	lamin A/C
Hs.500483	actin, alpha 2, smooth muscle, aorta
Hs.503998	transgelin
Hs.505705	myosin, light polypeptide 6, alkali, smooth muscle and non-muscle
Hs.534028	myosin, heavy polypeptide 8, skeletal muscle, perinatal
Hs.7879	interferon-related developmental regulator 1

In conclusion we consider a MVN analysis to be a better option than a UVN analysis on theoretical and practical grounds, as it directly uses the collected data and abolishes the need for a pre-processing step summarizing probe values into probeset intensities and the choice under what assumptions this should be done. Also this type of analysis leads to easier fitting models, suggesting it is more appropriate for the data at hand. This does not take away that a UVN analysis using heteroscedastic models gives quite good results. Indeed the results when focusing at general biological processes retrieved were similar between the two types of analyses and as expected, the MVN analysis did not introduce entirely new processes. But, though a less computationally intensive alternative, a UVN analysis will more likely be overconfident by losing between probe variability information. As microarray analysis is frequently a first step in a research path, performing a MVN analysis may pay itself back by picking out a stricter list of more reliable changed genes, leading to a more specific set of candidates for further validation testing on the lab. By this kind of modeling in general, much information can be gained from the specific and less

known genes and biological processes that are retrieved between patient groups and in common between them. It is beyond the scope of this study though, to investigate in detail the biology of the diseases at hand. A future extension of this study could be to test a further improvement of the MVN analysis. As this analysis now assumes no order between the probes, it could be tested if taking order into account instead of including a random effect would improve these models still further. Also this would allow estimating the effect of RNA degradation and its influence on the results of the analysis.

Acknowledgements

This research was supported by the Cardiovascular Research Institute Maastricht (CARIM).

References

- 1 Butte A: **The use and analysis of microarray data.** *Nat Rev Drug Discov* 2002, **1**:951-960.
- 2 Irizarry RA, Warren D, Spencer F, Kim IF, Biswal S, Frank BC, Gabrielson E, Garcia JG, Geoghegan J, Germino G, et al: **Multiple-laboratory comparison of microarray platforms.** *Nat Methods* 2005, **2**:345-350.
- 3 Affymetrix: *Affymetrix Microarray Suite User Guide.* 5 edition. Santa Clara, CA: Affymetrix; 2002.
- 4 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* 2003, **4**:249-264.
- 5 Eijssen LMT, Lindsey PJ, van Eijnden RGE, Schroen B, Pinto YM, Smeets HJM, Geraedts JPM: **Affymetrix chip data analysis: the gain of better modeling.** (*submitted*).
- 6 Harbig J, Sprinkle R, Enkemann SA: **A sequence-based identification of the genes detected by probesets on the Affymetrix U133 plus 2.0 array.** *Nucleic Acids Res* 2005, **33**:e31.
- 7 Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, Bunney WE, Myers RM, Speed TP, Akil H, Watson SJ, Meng F: **Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data.** *Nucleic Acids Res* 2005, **33**:e175.
- 8 DiMauro S: **Mitochondrial diseases.** *Biochim Biophys Acta* 2004, **1658**:80-88.
- 9 R Development Core Team: *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing; 2004.
- 10 Lindsey JK: *Models for Repeated Measurements.* Oxford: Oxford University Press; 1999.
- 11 Lindsey JK: [<http://popgen.unimaas.nl/~jlindsey/rcode.html>]
- 12 Gautier L, Cope L, Bolstad BM, Irizarry RA: **affy--analysis of Affymetrix GeneChip data at the probe level.** *Bioinformatics* 2004, **20**:307-315.
- 13 Akaike H: **Information theory and an extension of the maximum likelihood principle.** In *Second International Symposium on Inference Theory; Budapest: Akadémiai Kiadó.* Edited by Petrov BN, Csàki F. 1973: 267-281.
- 14 Atkinson AC: **A note on the generalized information criterion for choice of a model.** *Biometrika* 1980, **67**:413-418.
- 15 Hosack DA, Dennis G, Jr., Sherman BT, Lane HC, Lempicki RA: **Identifying biological themes within lists of genes with EASE.** *Genome Biol* 2003, **4**:R70.
- 16 Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25**:25-29.

Supplementary Material

Table S1 Settings determined to be optimal based on a test run on 50 probesets, for both model types and both data sets; ndigit = number of significant digits; maxstep = maximum stepsize to take between two iterations of fitting

Model type		Covariate Model	Location Model	Dispersion Model	Location and dispersion Model
Cardio-set					
UVN	ndigit	8	16	12	8
	maxstep	1000	1	10	10
MVN	ndigit	16	16	16	16
	maxstep	50	50	50	50
Mito-set					
UVN	ndigit	8	8	12	8
	maxstep	100	1	10	1
MVN	ndigit	24	24	24	24
	maxstep	0	0	0	0

Table S2 Number of significant probesets for both data sets and types of analysis; for the mito-set there were no probesets for which no model could be fitted e.g. because of non-convergence, for the cardio-set this was the case for 652 probesets for the UVN analysis and none for the MVN analysis; results based on AICs computed with penalty correction to match a 5% significance of a χ^2 test and analyses done with no cut-off points on the ratios; nc = no change.

Mito-set (8113 probesets)					
UVN analysis			MVN analysis		
MELAS	Leigh	Number of probesets	MELAS	Leigh	Number of probesets
up	up	686	up	up	48
up	nc	820	up	nc	75
up	down	217	up	down	0
nc	up	1255	nc	up	300
nc	nc	2229	nc	nc	7402
nc	down	838	nc	down	252
down	up	362	down	up	0
down	nc	1228	down	nc	24
down	down	478	down	down	12
total differential					
3791	3836	5884	159	612	711

Cardio-set (11080 and 11732 probesets)					
UVN analysis			MVN analysis		
Valv1	Valv2	Number of probesets	Valv1	Valv2	Number of probesets
up	up	1032	up	up	363
up	nc	748	up	nc	295
up	down	20	up	down	0
nc	up	632	nc	up	156
nc	nc	6505	nc	nc	10897
nc	down	385	nc	down	3
down	up	28	down	up	0
down	nc	900	down	nc	16
down	down	830	down	down	2
total differential					
3558	2927	4575	676	524	835

Table S3 Number of significant probesets for both data sets and types of analysis; for the mito-set there were no probesets for which no model could be fitted e.g. because of non-convergence, for the cardio-set this was the case for 543 probesets for the UVN analysis and for 12 probesets for the MVN analysis; results based on AIC values with standard penalty and analyses done with no cut-off points on the ratios; nc = no change.

Mito-set					
UVN analysis			MVN analysis		
MELAS	Leigh	Number of probesets	MELAS	Leigh	Number of probesets
up	up	718	up	up	164
up	nc	971	up	nc	220
up	down	343	up	down	8
nc	up	901	nc	up	407
nc	nc	1707	nc	nc	6711
nc	down	765	nc	down	473
down	up	632	down	up	12
down	nc	1229	down	nc	74
down	down	847	down	down	44
total differential					
4740	4206	6406	522	1108	1402

Cardio-set					
UVN analysis			MVN analysis		
Valv1	Valv2	Number of probesets	Valv1	Valv2	Number of probesets
up	up	1970	up	up	721
up	nc	841	up	nc	309
up	down	38	up	down	0
nc	up	848	nc	up	613
nc	nc	4107	nc	nc	10010
nc	down	561	nc	down	8
down	up	54	down	up	1
down	nc	1036	down	nc	54
down	down	1734	down	down	4
total differential					
5673	5205	7082	1089	1347	1710

Table S4 Number of significant probesets for both data sets and types of analysis focused on the differences between the two types of analysis; results based on AICs computed with penalty correction to match a 5% significance of a χ^2 test and analyses done with no cut-off points on the ratios; nc = no change; NA = not available (model did not converge).

MELAS non-multi	MELAS multi	number	Leigh non-multi	Leigh multi	number
up	up	42	up	up	197
up	nc	1676	up	nc	2069
up	down	5	up	down	37
nc	up	49	nc	up	134
nc	nc	4253	nc	nc	4014
nc	down	20	nc	down	129
down	up	32	down	up	17
down	nc	2025	down	nc	1418
down	down	11	down	down	98
NA	up	0	NA	up	0
NA	nc	0	NA	nc	0
NA	down	0	NA	down	0
total: 8113			total: 8113		
Valv1 non-multi	Valv1 multi	number	Valv2 non-multi	Valv2 multi	number
up	up	298	up	up	182
up	nc	1502	up	nc	1510
up	down	0	up	down	0
nc	up	243	nc	up	268
nc	nc	7270	nc	nc	7884
nc	down	9	nc	down	1
down	up	7	down	up	8
down	nc	1747	down	nc	1225
down	down	4	down	down	2
NA	up	110	NA	up	61
NA	nc	537	NA	nc	589
NA	down	5	NA	down	2
total: 11732			total: 11732		

Table S5a Differential genes for Leigh disease that contain the word 'muscle' in their GO annotations at the biological process or the cellular component axis for the UVN analysis; results based on AICs computed with penalty correction to match a 5% significance of a χ^2 test and analyses done with cut-off points on the ratios of 0.8 and 1.25.

Hs.118845	troponin C, slow
Hs.133892	tropomyosin 1 (alpha)
Hs.134602	nebulin
Hs.134602	nebulin
Hs.140950	EF hand calcium binding protein 2
Hs.149387	myosin VI
Hs.182421	troponin C2, fast
Hs.183713	endothelin receptor type A
Hs.200841	laminin, alpha 2 (merosin, congenital muscular dystrophy)
Hs.201379	phosphorylase kinase, alpha 1 (muscle)
Hs.20516	histone deacetylase 4
Hs.2681	gastrin
Hs.269577	histone deacetylase 7A
Hs.282938	collagen, type IV, alpha 4
Hs.35937	myogenic factor 6 (herculin)
Hs.410970	myosin, light polypeptide 5, regulatory
Hs.413238	phosphoglycerate mutase 2 (muscle)
Hs.432642	mitogen-activated protein kinase 12
Hs.434285	myotubularin 1
Hs.438782	histone deacetylase 5
Hs.438953	sarcoglycan, beta (43kDa dystrophin-associated glycoprotein)
Hs.439463	AE binding protein 1
Hs.474667	like-glycosyltransferase
Hs.476448	filamin B, beta (actin binding protein 278)
Hs.483104	mutated in colorectal cancers
Hs.483454	calponin 3, acidic
Hs.500483	actin, alpha 2, smooth muscle, aorta
Hs.503998	transgelin
Hs.506502	myosin binding protein C, slow type
Hs.50889	myosin light chain 2
Hs.512709	troponin I, cardiac
Hs.513490	aldolase A, fructose-bisphosphate
Hs.514746	GATA binding protein 6
Hs.516966	BCL2-like 1
Hs.521653	muscle, skeletal, receptor tyrosine kinase
Hs.524484	integrin, alpha 7
Hs.534028	myosin, heavy polypeptide 8, skeletal muscle, perinatal
Hs.534313	early growth response 3
Hs.7879	interferon-related developmental regulator 1
Hs.94865	TEA domain family member 4
Hs.98303	caveolin 3

Table S5b Differential genes for Leigh disease that contain the word 'muscle' in their GO annotations at either the biological process or the cellular component axis for the MVN analysis; results based on AICs computed with penalty correction to match a 5% significance of a χ^2 test and analyses done with cut-off point on the ratios of 0.80 and 1.25.

Hs.104672	downregulated in ovarian cancer 1
Hs.133892	tropomyosin 1 (alpha)
Hs.149387	myosin VI
Hs.200841	laminin, alpha 2 (merosin, congenital muscular dystrophy)
Hs.336994	metastasis suppressor 1
Hs.371199	sarcoglycan, epsilon
Hs.432642	mitogen-activated protein kinase 12
Hs.440895	myosin, heavy polypeptide 3, skeletal muscle, embryonic
Hs.471419	desmin
Hs.476448	filamin B, beta (actin binding protein 278)
Hs.483454	calponin 3, acidic
Hs.490203	caldesmon 1
Hs.50889	myosin light chain 2
Hs.513490	aldolase A, fructose-bisphosphate
Hs.534028	myosin, heavy polypeptide 8, skeletal muscle, perinatal
Hs.73454	troponin T3, skeletal, fast
Hs.7879	interferon-related developmental regulator 1
Hs.85937	myosin binding protein C, fast type

Table S5c Differential genes for Leigh disease that contain the word ‘muscle’ in their GO annotations at either the biological process or the cellular component axis for the MVN analysis; results based on AIC values with standard penalty and analyses done without cut-off points on the ratios.

Hs.104672	downregulated in ovarian cancer 1
Hs.146070	TRK-fused gene
Hs.149387	myosin VI
Hs.160562	insulin-like growth factor 1 (somatomedin C)
Hs.172684	vesicle-associated membrane protein 5 (myobrevin)
Hs.183713	endothelin receptor type A
Hs.187338	myosin, light polypeptide 1, alkali; skeletal, fast
Hs.190086	myosin regulatory light chain MRCL3
Hs.196054	histone deacetylase 9
Hs.200841	laminin, alpha 2 (merosin, congenital muscular dystrophy)
Hs.201379	phosphorylase kinase, alpha 1 (muscle)
Hs.233240	collagen, type VI, alpha 3
Hs.284244	fibroblast growth factor 2 (basic)
Hs.432642	mitogen-activated protein kinase 12
Hs.440895	myosin, heavy polypeptide 3, skeletal muscle, embryonic
Hs.444403	protein phosphatase 1, regulatory (inhibitor) subunit 12B
Hs.463412	sarcoglycan, alpha (50kDa dystrophin-associated glycoprotein)
Hs.466088	tropomyosin 4
Hs.476448	filamin B, beta (actin binding protein 278)
Hs.483454	calponin 3, acidic
Hs.490203	caldesmon 1
Hs.491359	lamin A/C
Hs.49582	protein phosphatase 1, regulatory (inhibitor) subunit 12A
Hs.500483	actin, alpha 2, smooth muscle, aorta
Hs.503998	transgelin
Hs.505705	myosin, light polypeptide 6, alkali, smooth muscle and non-muscle
Hs.50889	myosin light chain 2
Hs.511504	transcription factor 12 (HTF4, helix-loop-helix transcription factors 4)
Hs.512709	troponin I, cardiac
Hs.513490	aldolase A, fructose-bisphosphate
Hs.513941	myosin, heavy polypeptide 2, skeletal muscle, adult
Hs.517586	myoglobin
Hs.523403	troponin I, skeletal, fast
Hs.524484	integrin, alpha 7
Hs.534028	myosin, heavy polypeptide 8, skeletal muscle, perinatal
Hs.534085	troponin T1, skeletal, slow
Hs.55777	Fukuyama type congenital muscular dystrophy (fukutin)
Hs.73454	troponin T3, skeletal, fast
Hs.7879	interferon-related developmental regulator 1

Chapter 6

The use of spikes in Affymetrix chip expression data analysis

L.M.T. Eijssen^{1}, P.J. Lindsey¹*

¹ Department of Population Genetics, Genomics, and Bioinformatics, Maastricht University,
P.O. Box 616, 6200 MD, Maastricht, the Netherlands

**To whom correspondence should be addressed:*

L.M.T. Eijssen, Department of Population Genetics, Genomics, and Bioinformatics,
P.O. Box 616, 6200 MD Maastricht, The Netherlands. Tel.: +31-43-3881995. Fax:
+31-43-3884573. E-mail: Lars.Eijssen@gen.unimaas.nl

in preparation

Abstract

A very important aspect of microarray data analysis is controlling for several disturbing effects, both of biological and technical origin. Various techniques are available to assist in correcting for these effects, amongst which are normalization and background correction. Another frequently used procedure is adding spiked in RNA to the sample at several stages of the protocol, to be used as control. Eventually, spike quantities are detected by specific spike probes on the array.

The Affymetrix platform detects transcript expression by a probeset containing several probes that are created to specifically recognize this transcript. In addition to the probesets to detect transcripts, probesets for two types of spikes are present on Affymetrix GeneChips. These spike types are hybridization spikes that are added to the sample before hybridization and 'poly-A' spikes that are added before amplification. The last group of spikes controls for amplification, labeling and hybridization.

Before, we developed an approach to model gene expression that directly incorporates study design to control for disturbing effects. We constructed both univariate models based on summarized probeset intensities, as well as multivariate models based on probe intensities. Here we investigate the use of Affymetrix spikes as covariates in both types of models.

Two differing Affymetrix Genechip expression Array data sets were selected, considering different chip types and diseases. For both model types and data sets overall statistics on spike, spike probeset and, in the multivariate case, spike probe use were computed. Strong differences in frequency of use of several spike probes and probesets as covariates in both types of models fitting expression profiles appeared to be present. However, these differences were not the same between the data set. Thus it is advisable to initially incorporate spike information for all spikes. During analysis, it can then be determined which ones are most needed for the data set at hand and which ones may be discarded. The frequency of use of the several spike types can provide insight into how well a certain step of the laboratory protocol was performed. Also, for the poly-A spikes, the five prime probesets were more frequently used. Overall, the univariate models contained more spikes per model than the multivariate models. However, almost all models built contained spike information. These results show that statistics on spike use can provide additional information on the experiment, and that the use of spikes as covariates improves gene expression modeling and as such detection of differential genes.

Keywords

Microarray, modeling, hybridization and labeling controls

Introduction

Microarray gene expression analysis has become a technique of major importance for retrieving global or specific genetic differences between several groups of biological samples. Amongst these, studies comparing disease samples versus healthy controls form a major group. One very important aspect of analysis methods to determine which genes are differentially expressed between the groups of samples is correcting for several disturbing effects. Amongst these are biological differences in the consistence of the sample, the quantity of material sampled, variances in the laboratory protocol followed or materials used or other external influences. These factors can result in adverse effects, including within and between slide differences. Various techniques are available to assist in correcting for these effects, amongst which are normalization and background correction. Another frequently used procedure is adding spiked in RNA to the sample at several stages of the protocol. Eventually, spike quantities are detected by specific spike probes on the array.

In the microarray field, several platforms are used, from custom-made cDNA and oligonucleotide-arrays to standardized commercial platforms [1]. The commercial Affymetrix Genechip expression Array system is a reproducible and reliable platform [2] that is widely used. Affymetrix Chips contain probesets measuring the expression levels of several gene transcripts. Each probeset consists of a number of 25mer probes, located at different positions along the sequence of the gene under consideration. Furthermore, the chips contain probes for two types of spikes: hybridization spikes that are added to the sample before hybridization (BioB, Bioc, BioDn, and CreX) and 'poly-A' spikes that are added before amplification (DapX, LysX, PheX, and ThrX). The last group of spikes controls for amplification, labeling and hybridization.

The aim of the current research was to investigate how these spikes perform when used as covariates in models for the expression profile of each probeset (gene) on the chip. Two types of models were used, a univariate normal (UVN) model including both a group mean and group variance effect [3], and a multivariate normal (MVN) model including both of these effects as well as a random effect [4]. To model the expression profile of a probeset, the UVN model uses a summarized value for each sample as response, whereas the MVN model uses all probe intensities for that probeset directly. Also, two differing Affymetrix Genechip expression Array data sets were selected. Different types of chips were run for the two studies, each considering a different human disease and containing three subgroups of samples (two disease groups and a control group). For both model types and data sets overall statistics on spike, spike probeset and, in the MVN case, spike probe use as covariates were computed. These calculations provided informative insights into characteristics of spike use and into the analysis of Affymetrix Genechip expression Arrays.

Materials and Methods

Data

Two available human Affymetrix Genechip expression Array data sets from the Genome Center lab at Maastricht University were used.

For a study considering cardiac hypertrophy caused by valvular disease (aorta stenosis), twenty-six samples were run on hgu133a chips, with nine and ten samples in two disease groups with different prognosis and seven (coronary artery bypass grafting) controls. This data set is further referred to as the 'cardio-set'.

For a study considering two mitochondrial neuromuscular syndromes, MELAS disease and Leigh syndrome [5], twenty samples were run on hgu95av2 chips, with seven and six samples in the two disease groups respectively and six controls. This data set is further referred to as the 'mito-set'.

Both studies used two-cycle amplification protocols. The data sets, with permuted probesets and masked labels as they represent ongoing research, are available upon request from the first author.

Chip Description Files

UniGene based updates of the CDFs for the hgu133a and hgu95av2 chips by Dai *et al* [6], version 6, were used. Besides the spikes, the hgu133a chip contains 11732 updated probesets, the hgu95 contains 8113 updated probesets, each with a UniGene-based annotation. For the spikes, the composition of the several probesets has not changed as compared to the original Affymetrix CDFs.

Spike characteristics

Both chip types considered contain probesets for several spikes that can be divided into two types: controls for amplification, labeling, and hybridization, called 'poly-A' spikes, and pure hybridization controls. Spikes of the first type are added before the labeling takes place and can be used as controls for the protocol from that point on. The second category is only added at the end, directly before the slides are hybridized with sample material. The spikes have different concentrations. Poly-A spikes include DapX, LysX, PheX, ThrX and TrpnX, which are detected by three probesets each (5', middle (M) and 3'). Hybridization spikes include BioB, BioC, BioDn and CreX, which are detected by two probesets each (5' and 3'). All spike probesets consist of twenty different anti-sense probes. The hgu95av2 Chip also contains sense probes for the hybridization spikes. These were excluded from analysis as they were not used in the experiments under consideration. The spike TrpnX was also excluded as it is no longer in the Affymetrix kit. Finally, the spike BioB was excluded from analysis because of its low expression.

Programming environment

All code was implemented in the R software version 2.0.1 [7], with use of the `glnm` and `growth` libraries [8, 9], and the `affy` library [10].

The analyses were done on 2.8 GHz dual processor servers, having 4 GB of internal memory and operating on Linux, Redhat V9 (Shrike) or Fedora Core 3.

Statistical models

For both data sets, expression values were log transformed, to obtain models where the group difference represents the fold change between those groups on the original scale. For each of the probesets in the updated CDFs, the logged intensity profile was modeled with two different approaches, a univariate normal (UVN) model taking a summarized probeset intensity for each sample as response and a multivariate normal (MVN) model taking the intensities of all probes directly as response for each probeset [3, 4]. In both cases, spike intensities were controlled for by using them as covariates in the models. This was done in a similar way for both model types, but adapted to the specificities of each, as for each spike, the best spike probeset (in case of the UVN model) or the best spike probe (in case of the MVN model), was included into the model. The best spike probe(set) for each spike was determined by sequentially including each of them into a model by itself. As such, after selection, each model contained seven spike covariates, one for each of the three hybridization spikes and one for each of the four poly-A spikes. Also, each model contained both a group effect for the mean and for the variance. Furthermore, the MVN model contained a random effect to model probe variability within the probeset at hand.

The summarized probeset intensity to be used as response for the UVN models was computed by using the `expresso` function of the `affy` library, with “mas” background correction for the cardio-set and no background correction for the mito-set, constant normalization, no use of mismatch intensities and the “mas” summarization method. These settings were chosen to minimize pre-processing, as the models took study design into account and entailed controls for differences between slides and variability in the protocol steps. Clearly, for the MVN models, no application of the `expresso` function was needed and probe intensities were used directly.

Modeling was done stepwise: first an intercept, covariate, mean group effect, and variance group effect model were built, and eventually a model including both group effects for the mean and variance. This allowed more elaborate models to use the coefficients of the more basic ones as initial parameter estimates. Before building models for all probesets in the two Affymetrix data sets, a run on a subset of 50 probesets was done for each data set in order to determine the best parameter settings for the computational digit precision and the maximum step to take between each iteration of the model fitting procedure. To improve convergence further in the complete runs, two rounds of iterations were done, a first round with a maximum of 500 steps and a second with a maximum of 100 steps starting from the initial estimates obtained from the first round. For the cardio set, different initial values had to be used to improve model convergence for the UVN model.

The final model for each probeset was obtained by simplifying the model containing all spikes and both a mean and variance effect, removing all non-significant variables. To do so, first unneeded spikes were removed, and then mean and variance group effects were removed if necessary. For the MVN models, the random effect was always kept as an essential parameter.

Rating of the models

Models were evaluated with the Akaike Information Criterion (AIC) that is computed by summing the minus loglikelihood (-LL) of the model and the number of parameters it contains, yielding lower values for better-fitting and less complex models [11], thus a goodness-of-fit measure penalizing complexity. The penalties of the AIC values were weighted to match a five percent significance level of a χ^2 test [12]. Model convergence was checked by comparing the -LL values of all models with the more basic ones for the same probeset, as built during the procedure. The -LL value of a model extended with more variables was required to always be equal to or lower than that of more basic ones to ensure proper convergence. If this was not the case within a small tolerance value (of 1 unit), to account for small random and rounding effects, the -LL value and model coefficients were replaced by those of the more basic model and the AIC recomputed. As an additional check, any model with an infinite parameter value or an infinite AIC, was discarded from further analysis and the number of non-fitting or non-converging models was recorded.

Estimating spike usability

The final simplified models built were used to calculate how often each of the spikes was used. It was also recorded how often each of the probesets of each spike was used. This permitted to estimate the usefulness of 3', M (middle) and 5' probesets. Furthermore, it was computed how many models required spikes. For those who did, it was further computed how many spikes were needed per model, separately for the two types of spikes. For the MVN analysis, the total use of each spike probe was computed as well.

To better estimate the usage frequency of each spike, the above calculations on spike usage were modified as follows. First, for both analyses all spike probe(set) counts were weighted by the total number of spikes in the model, e.g. a model having n spike probe(set)s resulted in increasing the count for each of its spikes by $1 / n$. Additionally, for the MVN analysis, all spike probes that were not used frequently (in less than one percent of models built for a data set) were not accounted for in the spike counts. Weighting was also applied to these filtered results. All computations were done for both the cardio-set and the mito-set and then results were compared between them and between the two analysis methods.

Results

For both the UVN and the MVN model, the optimal settings of several modeling parameters, as determined from a test-run on a subset of genes, were allowed to be different for intermediate models including different subsets of variables, in order to get the best fit (lowest AIC). Settings for both the UVN and MVN model types and data sets can be found in the supplement to [4]. After determination of these settings, models were built for all probesets of both the cardio- and the mito-set, where all models could be fitted except for 652 models of the cardiac data set with the UVN analysis (5.56%) even after (automated) selection of several starting values for all of the coefficients.

Both unweighted counts and the counts weighted by the reciprocal of the number of spike variables present in the model at hand are presented in Figure 1, for each of the spikes and for each of the spike probesets separately. From these results it is clear that spikes were not used with equal frequency as covariates in the models and there was a clear preference for some. The same held for the spike probesets. Furthermore, differences between the data sets were present and the order of the spikes with respect to overall use was not the same between them. In the mito-set the poly-A spikes were used more frequently than the hybridization spikes. This is especially apparent from the MVN analysis. On the other hand, in the cardio-set both types were used almost in equal amounts. In the MVN analysis of the cardio-set, the spikes CreX and PheX were the less frequently used, and ThrX the most frequently.

Figure 1 Total counts for each spike and each probe set, both unweighted and weighted by the reciprocal of the number of spike variables in the model. The top row figures are at the same scale to stress differences in total spike use, the others are at different scales to focus on the relative use of each spike probeset.

Differences were also present between the UVN and MVN models. Many more spikes were used in the UVN models than in the MVN models. In Figure 2, information on the spike usage per model is presented. In the UVN analysis, mostly five or six spikes - several of each type - were used per model, while in the MVN analysis mostly only one or two were used. Remarkably, the UVN models for the cardio-set contain any number of spikes from both types with roughly similar frequency, while the models for the mito-set mostly contain two hybridization controls and four poly-A controls. Furthermore, it was of note that in the mito-set, the MVN models generally included no hybridization spike and one spike of the poly-A type. In the cardio-set mostly one spike of any of both types was used. For both modeling approaches, only a small number of models did not need any spike information to be taken into account. In the cardio-set the variation in the number of spikes used per model was higher than in the mito-set. Considering totals over all spikes, in the UVN analysis the percentage of non-significant spike variables was double for the cardio-set as compared to the mito-set (45 versus 23%), a difference that disappeared when considering the MVN analysis (82 versus 84%). This also indicates that especially in the mito-set many more spikes were used in the models for the UVN analysis than for the MVN analysis, an over 3.5-fold difference as compared to a less than two-fold difference in the cardio-set.

Overall, for most spikes there was a strong difference in magnitude between the 3', M and 5' probeset counts in both data sets for the MVN models. For the poly-A spikes, there was a clear preference for the 5' probesets and for the cardio-set this was also the case for the hybridization spikes. In the cardio-set some of the 5' counts even increased when comparing the MVN to the UVN analysis, whereas the 3' probeset counts strongly decreased.

Weighting did not influence spike probeset counts relatively to each other beyond small differences. However, the relative loss in weighted count as compared to the plain count was somewhat larger for the less frequently used probesets, emphasizing differences between spikes, spike types and 3'/5' differences. This is also reflected in the fact that the decreases in probeset counts ranged between 10% and 45% for the MVN analysis, where probesets with high and low frequency of use are present. On the other hand, for the UVN analyses these decreases were in the same range for each probeset within each data set (approximately 75% for the cardio-set and 81% for the mito-set).

Figure 2 Histograms of number of spikes in each model, for all spikes and per spike class. Also the presence of spike classes in the models is shown.

For the MVN analysis counts per spike probe were collected as well, which are presented in Figure 3. For reference the probeset counts for this analysis are included again in this Figure. When considering the per probe results, it appeared that there was a clear preference for certain probes within the spike probesets but that results differed between both data sets considered. Besides the large differences between both data sets, some probes or regions of probes were apparently important for both data sets, the most clear example being the first half of the 5' probes for ThrX. Overall, differences in frequency of use between probes were large and it was clear that many of the spike probes were never or almost never used. Both at probeset and probe levels, the differences in frequencies of use were somewhat more extreme for the cardio-set than for the mito-set. To filter out probes that were used in a very low frequency, all probes were removed that were present in less than one percent of models. Probeset and probe counts after filtering are also presented in Figure 3, where the names of all probesets that completely lost their counts are cleared out. Filtering low count probes resulted in a stronger decrease in total use for less frequently used probesets and as such resulted in emphasizing differences, just as weighting did. Combination of weighting and filtering showed a cooperative emphasizing effect (data not shown), although, like weighting of the non-filtered data, weighting of the filtered data set did not have a large additional effect on relative probe use. The mito-set lost some more significant spike variables by filtering, indicating that the models for this data set contained more low frequency spike probes than those for the cardio-set. Though, when weighting the filtered set the effect was reversed, indicating that those low frequency spike probes were overrepresented in models containing more spike variables. In the cardio-set, only 5' probesets were left without zero counts after filtering, for both the poly-A probesets and BioC. For the other spikes the 5' probesets were highly overrepresented as well. For the mito-set, in the poly-A spike group only PheX was left with a non-zero 3' count, whereas in the hybridization spike group only BioC was left with a non-zero 5' count. Although several spike probesets got zero counts summed over their containing probes, no spike lost all of its count. When considered at a probe level, only six percent of probes for the cardio-set and nine percent of probes for the mito-set had a non-zero count after filtering. Furthermore, even after filtering, relative differences in probe use for the probes with non-zero counts were still large. A few probes contributed the majority of the total count for each spike and in fact in many cases the total count for a probeset was contributed by only one or two of its probes. Finally, in several of the cases where in fact several probes contributed to the total probeset count, we observed a tendency for regions of adjacent spike probes to be all frequently used, while the others were (almost) never used. This effect was more noticeable for probesets that had higher total counts than for those that were only used in a few models.

Figure 3 Total counts for each spike probeset and probe, both before and after filtering out the probes which are used in less than one percent of models for the respective data set. The top row figures are at the same scale to stress differences in total spike (probese) use, the others are at different scales to focus on the relative use of each spike probe. Probes are ordered along the sequence of the spike from 5' to 3'. As the order of the probesets is the same for all subplots, all labels for probeset with zero counts are indicatively left out.

A complete table of all results, basic, weighted, filtered and both, and summarized by probe, probeset, spike, spike type and spike position in the sequence, is presented in Supplementary Data Table 1.

Discussion

The current results show that the several spike probe(set)s are not used in a random way as covariates in models for the expression profiles of all genes on Affymetrix chips. At a spike level, clear differences are present between their usage frequencies. However, the effect is not consistent between data sets, as demonstrated by the example sets used. Together with the fact that none of the spikes had a zero count (never used in any model), this leads to the conclusion that it would not be wise to exclude certain spikes from the laboratory protocol or the data analysis. As demonstrated by the mito-set, it may occur however, that one of the spike types (hybridization or poly-A) is far more useful as a correction than the other type. Differences occurring between the data sets could relate to biological differences between the samples or to the consistency with which the laboratory protocols were followed. The fact that in the mito-set, hybridization spikes were used in a very low frequency probably indicates that hybridization was done very consistently in this study. This is extra relevant information that spike use frequencies could provide for data sets analyzed. Also, the MVN analysis of the cardio-set showed that, whereas both types of spike were needed, more often one or the other spike type was incorporated into a single model than both types. This may indicate that dependent on the transcript considered, the probes designed to detect that transcript had been well-optimized for amplification or hybridization.

Also for the specific spike probesets large differences in use were observed. For the poly-A spikes, the 5' probesets were far more frequently used as covariates in the models built, as compared to the 3' probesets. This could be explained by the fact that the 5' probesets may correlate better to the differences in amplification of transcripts within the protocol, whereas the 3' probesets get a signal underestimating sample transcript variability in amplification. This would happen because the 3' end of the spike transcript does not suffer from incomplete amplification during the protocol and always has high quality measurements. As such the 5' probesets would make better controls for the quantity of labeled transcript. This is also indicated for the hybridization spikes, by the fact that the cardio-set mainly uses 5' probes, whereas the mito-set, which has apparently a very good hybridization, mainly uses 3' probes.

When considering the frequency of use per spike probe, it is striking that many of those are (almost) never used. On the other hand, not rarely only one or very few probes per spike probeset were used extremely frequently and made up almost the entire count for that probeset. If a few probes were used for a spike probeset, these tended to be adjacent. This indicates that spikes targeting certain regions of the control sequence are more relevant than those targeting other parts of that sequence.

It may of course be that those probes are more specific than the others, but for spikes specificity should not be a major problem. Thus, certain regions of the control sequences probably match better to the sample transcripts with respect to their probe binding kinetics.

Another striking difference was in the number of spike variables needed per model, as compared between the UVN models and the MVN models. This may be because of the fact that the MVN models use the probe data directly for each gene on the chip while the UVN models use a summarized gene value for each of the samples, and may indicate that the MVN models and the assumptions these are based on, fit better to the data at hand. The UVN models will need several spikes of each type, to fit well to the expression profiles at hand. Differences in spike (probeset) use were also far clearer for the MVN models as these contain far less spike variables than the UVN models. However, it is also shown that even for the best suited analyses, the use of spike variables in the models still remains necessary, as only very few models contained no spike variables at all. Weighting spikes counts by the reciprocal of the number of spike variables in the model had a minor effect on spike counts. This indicates that likely no spike probe(set) were specifically used jointly in the models. An effect it shows though, is that spike probe(set)s that are used in lower frequency, lose more of their count than the ones used in higher frequency. This may indicate that the less used spikes are relatively more often used in models with more spike variables. These probably are again models that fit less easily to the expression profiles at hand and as such the spike variables may be used aspecifically in an attempt to still improve model fit.

In conclusion, strong differences in frequency of use of several spike probes and probesets as covariates in models fitting expression profiles are present. However, as these differences are not the same between several data sets, it is advisable to include all spikes in the experiments and not reject several beforehand. When analyzing it may appear that for the study at hand some spike probesets can be discarded, although several will always be needed for correction of design and protocol features. Data can then be re-analyzed after removing the spikes probe(set)s that were used in low frequency. This would also restrain the number of spikes variables to be incorporated in each model in a natural way, preventing overcomplicated and aspecific models. Future extensions of this work could test the effect of such procedure or of restraining the total number of spike variables per model. Furthermore, a non-used spike as TrpnX (discarded in this study) may also be included in the analysis, to evaluate its use as a background control.

Acknowledgements

This research was supported by the Cardiovascular Research Institute Maastricht (CARIM).

References

- 1 Butte A: **The use and analysis of microarray data.** *Nat Rev Drug Discov* 2002, 1:951-960.
- 2 Irizarry RA, Warren D, Spencer F, Kim IF, Biswal S, Frank BC, Gabrielson E, Garcia JG, Geoghegan J, Germino G, Griffin C, Hilmer SC, Hoffman E, Jedlicka AE, Kawasaki E, Martinez-Murillo F, Morsberger L, Lee H, Petersen D, Quackenbush J, Scott A, Wilson M, Yang Y, Ye SQ, Yu W: **Multiple-laboratory comparison of microarray platforms.** *Nat Methods* 2005, 2:345-350.
- 3 Eijssen LMT, Lindsey PJ, van Eijdsden RGE, Schroen B, Pinto YM, Smeets HJM, Geraedts JPM: **Affymetrix chip data analysis: the gain of better modeling.** (*submitted*) 2006.
- 4 Eijssen LMT, Gerards M, van Eijdsden RGE, Schroen B, Lindsey PJ: **Multivariate normal probe modeling for Affymetrix expression chip data.** (*in preparation*) 2006.
- 5 DiMauro S: **Mitochondrial diseases.** *Biochim Biophys Acta* 2004, 1658:80-88.
- 6 Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, Bunney WE, Myers RM, Speed TP, Akil H, Watson SJ, Meng F: **Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data.** *Nucleic Acids Res* 2005, 33:e175.
- 7 R Development Core Team: *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing; 2004.
- 8 Lindsey JK: *Models for Repeated Measurements.* Oxford: Oxford University Press; 1999.
- 9 Lindsey JK: [<http://popgen.unimaas.nl/~jlindsey/rcode.html>]
- 10 Gautier L, Cope L, Bolstad BM, Irizarry RA: **affy--analysis of Affymetrix GeneChip data at the probe level.** *Bioinformatics* 2004, 20:307-315.
- 11 Akaike H: **Information theory and an extension of the maximum likelihood principle.** In *Second International Symposium on Inference Theory; Budapest: Akadémiai Kiadó.* Edited by Petrov BN, Csàki F. 1973: 267-281.
- 12 Atkinson AC: **A note on the generalized information criterion for choice of a model.** *Biometrika* 1980, 67:413-418.

Supplementary Data Table 1 Complete results on spike counts.

cardio-set, UVN									
basic					weighted				
		M					M		
	5'	3'	total	5'	3'	total		5'	3'
BioC	2648.00	4033.00	6681.00	BioC	643.19	975.28	BioC	1618.47	
BioDn	3451.00	3302.00	6753.00	BioDn	888.72	848.60	BioDn	1737.32	
CreX	4475.00	3184.00	7659.00	CreX	1235.99	876.83	CreX	2112.82	
subtotal	10574.00	10519.00	21093.00	subtotal	2767.90	2700.71	subtotal	5468.61	
DapX	1902.00	1793.00	5012.00	DapX	394.14	387.23	DapX	300.29	1081.65
LysX	1824.00	985.00	5415.00	LysX	406.45	209.25	LysX	595.58	1211.29
PheX	1735.00	1527.00	5606.00	PheX	389.43	326.60	PheX	521.88	1237.92
ThrX	2885.00	1059.00	5574.00	ThrX	651.70	343.38	ThrX	225.46	1220.54
subtotal	8346.00	5935.00	21607.00	subtotal	1841.72	1266.47	subtotal	1643.20	4751.39
total	18920.00	5935.00	42700.00	total	4609.62	1266.47	total	4343.92	10220.00

cardio-set, MVN									
basic					weighted				
		M					M		
	5'	3'	total	5'	3'	total		5'	3'
BioC	2109.00	107.00	2216.00	BioC	1889.37	80.12	BioC	1969.48	
BioDn	1696.00	590.00	2286.00	BioDn	1282.70	497.70	BioDn	1780.40	
CreX	622.00	317.00	939.00	CreX	369.65	190.22	CreX	559.87	
subtotal	4427.00	1014.00	5441.00	subtotal	3541.72	768.03	subtotal	4309.75	
DapX	1868.00	60.00	1931.00	DapX	1242.07	20.50	DapX	1263.90	
LysX	2501.00	8.00	2559.00	LysX	1905.03	6.25	LysX	1933.70	
PheX	715.00	50.00	762.00	PheX	412.23	5.73	PheX	430.57	
ThrX	3916.00	15.00	3939.00	ThrX	2990.83	6.17	ThrX	3001.08	
subtotal	9000.00	36.00	9191.00	subtotal	6550.17	19.48	subtotal	6629.25	
total	13427.00	36.00	14632.00	total	10091.88	19.48	total	10939.00	

cardio-set, MVN									
filtered					filtered and weighted				
		M					M		
	5'	3'	total	5'	3'	total		5'	3'
BioC	1926.00	0.00	1926.00	BioC	1807.33	0.00	BioC	1807.33	
BioDn	1474.00	483.00	1957.00	BioDn	1164.95	446.00	BioDn	1610.95	
CreX	345.00	129.00	474.00	CreX	175.32	91.67	CreX	266.98	
subtotal	3745.00	612.00	4357.00	subtotal	3147.60	537.67	subtotal	3685.27	
DapX	1830.00	0.00	1830.00	DapX	1278.12	0.00	DapX	1278.12	
LysX	2334.00	0.00	2334.00	LysX	1851.23	0.00	LysX	1851.23	
PheX	499.00	0.00	499.00	PheX	309.98	0.00	PheX	309.98	
ThrX	3687.00	0.00	3687.00	ThrX	2906.40	0.00	ThrX	2906.40	
subtotal	8350.00	0.00	8350.00	subtotal	6345.73	0.00	subtotal	6345.73	
total	12095.00	0.00	12707.00	total	9493.33	0.00	total	10031.00	

mito-set, UVN							
basic				weighted			
5'	M	3'	total	5'	M	3'	total
BioC	1872.00	994.00	2866.00	BioC	354.93	185.25	540.18
BioDn	4151.00	1568.00	5719.00	BioDn	762.06	286.62	1048.68
CreX	3568.00	3244.00	6812.00	CreX	666.73	583.46	1250.19
subtotal	9591.00	5806.00	15397.00	subtotal	1783.72	1055.33	2839.05
DapX	2753.00	1469.00	2668.00	DapX	491.79	259.75	1235.68
LysX	4688.00	1334.00	1128.00	LysX	845.18	245.60	1297.49
PheX	2813.00	2939.00	1448.00	PheX	501.07	265.54	1298.38
ThrX	2838.00	1270.00	3189.00	ThrX	534.87	567.89	1335.41
subtotal	13092.00	7012.00	28537.00	subtotal	2372.90	1524.29	5166.95
total	22683.00	14239.00	43934.00	total	4156.62	2579.61	8006.00

mito-set, MVN							
basic				weighted			
5'	M	3'	total	5'	M	3'	total
BioC	276.00	126.00	402.00	BioC	209.30	74.75	284.05
BioDn	98.00	236.00	334.00	BioDn	57.23	174.58	231.82
CreX	49.00	228.00	277.00	CreX	31.85	119.35	151.20
subtotal	423.00	590.00	1013.00	subtotal	298.38	368.68	667.07
DapX	1203.00	5.00	1245.00	DapX	1007.88	2.20	1042.17
LysX	607.00	28.00	1663.00	LysX	504.53	904.88	1436.42
PheX	1186.00	258.00	2410.00	PheX	1059.20	826.62	2127.02
ThrX	2238.00	21.00	2613.00	ThrX	2053.73	18.67	2373.33
subtotal	5234.00	312.00	7931.00	subtotal	4625.35	289.07	6978.93
total	5657.00	902.00	8944.00	total	4923.73	657.75	7646.00

mito-set, MVN							
filtered				filtered and weighted			
5'	M	3'	total	5'	M	3'	total
BioC	110.00	85.00	195.00	BioC	90.50	53.17	143.67
BioDn	0.00	97.00	97.00	BioDn	0.00	84.75	84.75
CreX	0.00	116.00	116.00	CreX	0.00	77.17	77.17
subtotal	110.00	298.00	408.00	subtotal	90.50	215.08	305.58
DapX	1174.00	0.00	1174.00	DapX	1023.25	0.00	1023.25
LysX	453.00	870.00	1323.00	LysX	385.42	806.67	1192.08
PheX	996.00	830.00	1983.00	PheX	913.67	152.00	1804.33
ThrX	1893.00	84.00	1977.00	ThrX	1765.42	79.33	1844.75
subtotal	4516.00	1784.00	6457.00	subtotal	4087.75	152.00	5864.42
total	4626.00	1784.00	6865.00	total	4178.25	367.08	6170.00

probe name	cardio-set				mito-set			
	basic	weighed	filtered	filtered and weighed	basic	weighed	filtered	filtered and weighed
AFFX-Bioc-5_at (1)	1.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Bioc-5_at (2)	2.00	2.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Bioc-5_at (3)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Bioc-5_at (4)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Bioc-5_at (5)	761.00	734.08	761.00	739.83	110.00	82.00	110.00	90.50
AFFX-Bioc-5_at (6)	778.00	685.57	778.00	715.17	1.00	1.00	0.00	0.00
AFFX-Bioc-5_at (7)	0.00	0.00	0.00	0.00	5.00	3.83	0.00	0.00
AFFX-Bioc-5_at (8)	203.00	177.12	203.00	195.50	15.00	11.00	0.00	0.00
AFFX-Bioc-5_at (9)	32.00	25.58	0.00	0.00	8.00	4.53	0.00	0.00
AFFX-Bioc-5_at (10)	10.00	7.50	0.00	0.00	58.00	44.87	0.00	0.00
AFFX-Bioc-5_at (11)	2.00	0.75	0.00	0.00	37.00	33.53	0.00	0.00
AFFX-Bioc-5_at (12)	0.00	0.00	0.00	0.00	6.00	2.75	0.00	0.00
AFFX-Bioc-5_at (13)	1.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Bioc-5_at (14)	1.00	0.50	0.00	0.00	3.00	2.50	0.00	0.00
AFFX-Bioc-5_at (15)	17.00	10.42	0.00	0.00	1.00	1.00	0.00	0.00
AFFX-Bioc-5_at (16)	2.00	1.33	0.00	0.00	17.00	10.37	0.00	0.00
AFFX-Bioc-5_at (17)	5.00	4.50	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Bioc-5_at (18)	110.00	91.67	0.00	0.00	6.00	4.08	0.00	0.00
AFFX-Bioc-5_at (19)	184.00	147.35	184.00	156.83	6.00	5.33	0.00	0.00
AFFX-Bioc-5_at (20)	0.00	0.00	0.00	0.00	3.00	2.50	0.00	0.00
AFFX-Bioc-3_at (1)	29.00	14.58	0.00	0.00	3.00	0.83	0.00	0.00
AFFX-Bioc-3_at (2)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Bioc-3_at (3)	7.00	5.83	0.00	0.00	2.00	1.33	0.00	0.00
AFFX-Bioc-3_at (4)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Bioc-3_at (5)	23.00	19.83	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Bioc-3_at (6)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Bioc-3_at (7)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Bioc-3_at (8)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Bioc-3_at (9)	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Bioc-3_at (10)	5.00	5.00	0.00	0.00	7.00	6.25	0.00	0.00
AFFX-Bioc-3_at (11)	0.00	0.00	0.00	0.00	9.00	5.83	0.00	0.00
AFFX-Bioc-3_at (12)	0.00	0.00	0.00	0.00	18.00	13.08	0.00	0.00
AFFX-Bioc-3_at (13)	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Bioc-3_at (14)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Bioc-3_at (15)	2.00	2.00	0.00	0.00	1.00	1.00	0.00	0.00
AFFX-Bioc-3_at (16)	2.00	1.50	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Bioc-3_at (17)	6.00	4.50	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Bioc-3_at (18)	15.00	13.50	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Bioc-3_at (19)	5.00	4.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Bioc-3_at (20)	11.00	7.37	0.00	0.00	85.00	45.42	85.00	53.17

probe name	cardio-set				mito-set			
	basic	weighed	filtered	filtered and weighed	basic	weighed	filtered	filtered and weighed
AFFX-BioDn-5_at (1)	4.00	3.50	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-BioDn-5_at (2)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-BioDn-5_at (3)	0.00	0.00	0.00	0.00	16.00	11.03	0.00	0.00
AFFX-BioDn-5_at (4)	26.00	16.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-BioDn-5_at (5)	1350.00	1013.40	1350.00	1058.95	0.00	0.00	0.00	0.00
AFFX-BioDn-5_at (6)	10.00	8.50	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-BioDn-5_at (7)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-BioDn-5_at (8)	124.00	101.08	124.00	106.00	0.00	0.00	0.00	0.00
AFFX-BioDn-5_at (9)	12.00	5.07	0.00	0.00	8.00	4.00	0.00	0.00
AFFX-BioDn-5_at (10)	8.00	6.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-BioDn-5_at (11)	3.00	2.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-BioDn-5_at (12)	0.00	0.00	0.00	0.00	2.00	1.33	0.00	0.00
AFFX-BioDn-5_at (13)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-BioDn-5_at (14)	5.00	2.45	0.00	0.00	57.00	31.83	0.00	0.00
AFFX-BioDn-5_at (15)	72.00	62.17	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-BioDn-5_at (16)	1.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-BioDn-5_at (17)	17.00	15.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-BioDn-5_at (18)	21.00	19.00	0.00	0.00	3.00	1.53	0.00	0.00
AFFX-BioDn-5_at (19)	0.00	0.00	0.00	0.00	0.00	5.50	0.00	0.00
AFFX-BioDn-5_at (20)	43.00	28.03	0.00	0.00	3.00	2.00	0.00	0.00
AFFX-BioDn-3_at (1)	17.00	12.33	0.00	0.00	2.00	1.50	0.00	0.00
AFFX-BioDn-3_at (2)	0.00	0.00	0.00	0.00	2.00	1.17	0.00	0.00
AFFX-BioDn-3_at (3)	1.00	1.00	0.00	0.00	2.00	0.70	0.00	0.00
AFFX-BioDn-3_at (4)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-BioDn-3_at (5)	7.00	6.50	0.00	0.00	1.00	0.20	0.00	0.00
AFFX-BioDn-3_at (6)	139.00	120.53	139.00	136.00	0.00	0.00	0.00	0.00
AFFX-BioDn-3_at (7)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-BioDn-3_at (8)	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00
AFFX-BioDn-3_at (9)	0.00	0.00	0.00	0.00	36.00	21.78	0.00	0.00
AFFX-BioDn-3_at (10)	0.00	0.00	0.00	0.00	2.00	1.50	0.00	0.00
AFFX-BioDn-3_at (11)	1.00	0.50	0.00	0.00	21.00	16.67	0.00	0.00
AFFX-BioDn-3_at (12)	1.00	1.00	0.00	0.00	9.00	6.33	0.00	0.00
AFFX-BioDn-3_at (13)	36.00	21.67	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-BioDn-3_at (14)	3.00	3.00	0.00	0.00	15.00	7.87	0.00	0.00
AFFX-BioDn-3_at (15)	0.00	0.00	0.00	0.00	10.00	6.00	0.00	0.00
AFFX-BioDn-3_at (16)	38.00	34.08	0.00	0.00	97.00	83.12	97.00	84.75
AFFX-BioDn-3_at (17)	2.00	1.50	0.00	0.00	30.00	20.08	0.00	0.00
AFFX-BioDn-3_at (18)	145.00	117.75	145.00	129.00	2.00	2.00	0.00	0.00
AFFX-BioDn-3_at (19)	199.00	177.33	199.00	181.00	2.00	2.00	0.00	0.00
AFFX-BioDn-3_at (20)	1.00	0.50	0.00	0.00	4.00	2.67	0.00	0.00

probe name	cardio-set					mito-set				
	basic	weighed	filtered	filtered and weighed		basic	weighed	filtered	filtered and weighed	
AFFX-Crex-5 at (1)	0.00	0.00	0.00	0.00	0.00	3.00	1.67	0.00	0.00	0.00
AFFX-Crex-5 at (2)	0.00	0.00	0.00	0.00	0.00	3.00	1.45	0.00	0.00	0.00
AFFX-Crex-5 at (3)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Crex-5 at (4)	35.00	23.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Crex-5 at (5)	44.00	33.83	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Crex-5 at (6)	6.00	4.83	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Crex-5 at (7)	9.00	7.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Crex-5 at (8)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Crex-5 at (9)	0.00	0.00	0.00	0.00	0.00	2.00	2.00	0.00	0.00	0.00
AFFX-Crex-5 at (10)	0.00	0.00	0.00	0.00	0.00	2.00	0.67	0.00	0.00	0.00
AFFX-Crex-5 at (11)	59.00	38.67	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Crex-5 at (12)	0.00	0.00	0.00	0.00	0.00	28.00	17.40	0.00	0.00	0.00
AFFX-Crex-5 at (13)	345.00	171.98	345.00	0.00	175.32	3.00	1.83	0.00	0.00	0.00
AFFX-Crex-5 at (14)	1.00	0.50	0.00	0.00	0.00	5.00	4.33	0.00	0.00	0.00
AFFX-Crex-5 at (15)	8.00	5.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Crex-5 at (16)	2.00	1.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Crex-5 at (17)	55.00	42.50	0.00	0.00	0.00	3.00	2.50	0.00	0.00	0.00
AFFX-Crex-5 at (18)	7.00	5.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Crex-5 at (19)	6.00	3.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Crex-5 at (20)	45.00	30.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Crex-3 at (1)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Crex-3 at (2)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Crex-3 at (3)	7.00	2.48	0.00	0.00	0.00	36.00	20.98	0.00	0.00	0.00
AFFX-Crex-3 at (4)	1.00	0.50	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
AFFX-Crex-3 at (5)	3.00	2.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Crex-3 at (6)	0.00	0.00	0.00	0.00	0.00	5.00	2.33	0.00	0.00	0.00
AFFX-Crex-3 at (7)	1.00	0.50	0.00	0.00	0.00	3.00	1.83	0.00	0.00	0.00
AFFX-Crex-3 at (8)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Crex-3 at (9)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Crex-3 at (10)	4.00	2.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Crex-3 at (11)	19.00	12.17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Crex-3 at (12)	16.00	8.62	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
AFFX-Crex-3 at (13)	129.00	84.17	129.00	0.00	91.67	5.00	3.08	0.00	0.00	0.00
AFFX-Crex-3 at (14)	31.00	21.33	0.00	0.00	0.00	116.00	64.08	116.00	77.17	0.00
AFFX-Crex-3 at (15)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Crex-3 at (16)	1.00	0.25	0.00	0.00	0.00	1.00	0.25	0.00	0.00	0.00
AFFX-Crex-3 at (17)	0.00	0.00	0.00	0.00	0.00	4.00	2.50	0.00	0.00	0.00
AFFX-Crex-3 at (18)	103.00	55.37	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Crex-3 at (19)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Crex-3 at (20)	2.00	0.83	0.00	0.00	0.00	56.00	22.28	0.00	0.00	0.00

probe name	cardio-set				mito-set			
	basic	weighed	filtered	filtered and weighed	basic	weighed	filtered	filtered and weighed
AFFX-DapX-5_at (1)	22.00	15.83	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-DapX-5_at (2)	0.00	0.00	0.00	0.00	387.00	316.42	387.00	329.17
AFFX-DapX-5_at (3)	0.00	0.00	0.00	0.00	99.00	84.03	99.00	93.00
AFFX-DapX-5_at (4)	4.00	2.50	0.00	0.00	3.00	2.50	0.00	0.00
AFFX-DapX-5_at (5)	1.00	1.00	0.00	0.00	3.00	3.00	0.00	0.00
AFFX-DapX-5_at (6)	0.00	0.00	0.00	0.00	105.00	94.62	105.00	96.75
AFFX-DapX-5_at (7)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-DapX-5_at (8)	1.00	0.50	0.00	0.00	4.00	2.33	0.00	0.00
AFFX-DapX-5_at (9)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-DapX-5_at (10)	2.00	1.00	0.00	0.00	4.00	2.83	0.00	0.00
AFFX-DapX-5_at (11)	1599.00	1068.28	1599.00	1116.62	200.00	176.42	200.00	180.17
AFFX-DapX-5_at (12)	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00
AFFX-DapX-5_at (13)	0.00	0.00	0.00	0.00	1.00	0.50	0.00	0.00
AFFX-DapX-5_at (14)	3.00	2.20	0.00	0.00	1.00	1.00	0.00	0.00
AFFX-DapX-5_at (15)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-DapX-5_at (16)	0.00	0.00	0.00	0.00	160.00	140.87	160.00	143.33
AFFX-DapX-5_at (17)	0.00	0.00	0.00	0.00	2.00	1.33	0.00	0.00
AFFX-DapX-5_at (18)	5.00	3.33	0.00	0.00	6.00	5.00	0.00	0.00
AFFX-DapX-5_at (19)	231.00	147.42	231.00	161.50	223.00	173.53	223.00	180.83
AFFX-DapX-5_at (20)	0.00	0.00	0.00	0.00	4.00	2.50	0.00	0.00
AFFX-DapX-M_at (1)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-DapX-M_at (2)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-DapX-M_at (3)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-DapX-M_at (4)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-DapX-M_at (5)	1.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-DapX-M_at (6)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-DapX-M_at (7)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-DapX-M_at (8)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-DapX-M_at (9)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-DapX-M_at (10)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-DapX-M_at (11)	0.00	0.00	0.00	0.00	5.00	3.67	0.00	0.00
AFFX-DapX-M_at (12)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-DapX-M_at (13)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-DapX-M_at (14)	2.00	0.83	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-DapX-M_at (15)	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00
AFFX-DapX-M_at (16)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-DapX-M_at (17)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-DapX-M_at (18)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-DapX-M_at (19)	0.00	0.00	0.00	0.00	1.00	0.50	0.00	0.00
AFFX-DapX-M_at (20)	0.00	0.00	0.00	0.00	30.00	26.92	0.00	0.00

probe name	cardio-set					mito-set				
	basic	weighed	filtered	filtered and weighed		basic	weighed	filtered	filtered and weighed	
AFFX-DapX-3_at (1)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-DapX-3_at (2)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-DapX-3_at (3)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-DapX-3_at (4)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-DapX-3_at (5)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-DapX-3_at (6)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-DapX-3_at (7)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-DapX-3_at (8)	35.00	12.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-DapX-3_at (9)	2.00	0.75	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-DapX-3_at (10)	15.00	4.33	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
AFFX-DapX-3_at (11)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-DapX-3_at (12)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-DapX-3_at (13)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-DapX-3_at (14)	1.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-DapX-3_at (15)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-DapX-3_at (16)	1.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-DapX-3_at (17)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-DapX-3_at (18)	0.00	0.00	0.00	0.00	0.00	3.00	0.95	0.00	0.00	0.00
AFFX-DapX-3_at (19)	6.00	2.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-DapX-3_at (20)	0.00	0.00	0.00	0.00	0.00	1.00	0.25	0.00	0.00	0.00
AFFX-LysX-5_at (1)	145.00	93.43	145.00	101.45	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-LysX-5_at (2)	118.00	84.92	118.00	89.67	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-LysX-5_at (3)	19.00	12.83	0.00	0.00	1.00	1.00	0.00	0.00	0.00	0.00
AFFX-LysX-5_at (4)	90.00	60.23	0.00	0.00	4.00	4.00	0.00	0.00	0.00	0.00
AFFX-LysX-5_at (5)	0.00	0.00	0.00	0.00	100.00	97.53	100.00	98.25	0.00	0.00
AFFX-LysX-5_at (6)	1.00	0.50	0.00	0.00	39.00	35.58	0.00	0.00	0.00	0.00
AFFX-LysX-5_at (7)	4.00	2.20	0.00	0.00	353.00	276.33	353.00	287.17	0.00	0.00
AFFX-LysX-5_at (8)	4.00	3.00	0.00	0.00	3.00	2.33	0.00	0.00	0.00	0.00
AFFX-LysX-5_at (9)	47.00	34.67	0.00	0.00	16.00	14.08	0.00	0.00	0.00	0.00
AFFX-LysX-5_at (10)	2.00	1.50	0.00	0.00	51.00	40.33	0.00	0.00	0.00	0.00
AFFX-LysX-5_at (11)	2071.00	1611.75	2071.00	1660.12	2.00	1.25	0.00	0.00	0.00	0.00
AFFX-LysX-5_at (12)	0.00	0.00	0.00	0.00	10.00	8.00	0.00	0.00	0.00	0.00
AFFX-LysX-5_at (13)	0.00	0.00	0.00	0.00	10.00	9.33	0.00	0.00	0.00	0.00
AFFX-LysX-5_at (14)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-LysX-5_at (15)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-LysX-5_at (16)	0.00	0.00	0.00	0.00	2.00	1.33	0.00	0.00	0.00	0.00
AFFX-LysX-5_at (17)	0.00	0.00	0.00	0.00	6.00	5.33	0.00	0.00	0.00	0.00
AFFX-LysX-5_at (18)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-LysX-5_at (19)	0.00	0.00	0.00	0.00	5.00	5.00	0.00	0.00	0.00	0.00
AFFX-LysX-5_at (20)	0.00	0.00	0.00	0.00	5.00	3.08	0.00	0.00	0.00	0.00

probe name	cardio-set				mito-set			
	basic	weighed	filtered	filtered and weighed	basic	weighed	filtered	filtered and weighed
AFFX-LysX-M_at (1)	0.00	0.00	0.00	0.00	68.00	58.33	0.00	0.00
AFFX-LysX-M_at (2)	0.00	0.00	0.00	0.00	104.00	101.08	104.00	101.08
AFFX-LysX-M_at (3)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-LysX-M_at (4)	0.00	0.00	0.00	0.00	37.00	12.85	0.00	0.00
AFFX-LysX-M_at (5)	0.00	0.00	0.00	0.00	163.00	123.08	163.00	132.17
AFFX-LysX-M_at (6)	0.00	0.00	0.00	0.00	421.00	402.17	421.00	406.00
AFFX-LysX-M_at (7)	1.00	0.25	0.00	0.00	7.00	7.00	0.00	0.00
AFFX-LysX-M_at (8)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-LysX-M_at (9)	0.00	0.00	0.00	0.00	182.00	162.62	182.00	167.42
AFFX-LysX-M_at (10)	4.00	4.00	0.00	0.00	21.00	17.58	0.00	0.00
AFFX-LysX-M_at (11)	3.00	2.00	0.00	0.00	4.00	3.33	0.00	0.00
AFFX-LysX-M_at (12)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-LysX-M_at (13)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-LysX-M_at (14)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-LysX-M_at (15)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-LysX-M_at (16)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-LysX-M_at (17)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-LysX-M_at (18)	0.00	0.00	0.00	0.00	18.00	15.25	0.00	0.00
AFFX-LysX-M_at (19)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-LysX-M_at (20)	0.00	0.00	0.00	0.00	3.00	1.58	0.00	0.00
AFFX-LysX-3_at (1)	0.00	0.00	0.00	0.00	3.00	2.00	0.00	0.00
AFFX-LysX-3_at (2)	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-LysX-3_at (3)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-LysX-3_at (4)	1.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-LysX-3_at (5)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-LysX-3_at (6)	3.00	2.00	0.00	0.00	22.00	22.00	0.00	0.00
AFFX-LysX-3_at (7)	10.00	6.83	0.00	0.00	1.00	1.00	0.00	0.00
AFFX-LysX-3_at (8)	0.00	0.00	0.00	0.00	2.00	2.00	0.00	0.00
AFFX-LysX-3_at (9)	4.00	1.50	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-LysX-3_at (10)	1.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-LysX-3_at (11)	2.00	0.58	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-LysX-3_at (12)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-LysX-3_at (13)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-LysX-3_at (14)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-LysX-3_at (15)	2.00	0.83	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-LysX-3_at (16)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-LysX-3_at (17)	14.00	4.67	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-LysX-3_at (18)	9.00	2.92	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-LysX-3_at (19)	2.00	0.58	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-LysX-3_at (20)	1.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00

probe name	cardio-set					mito-set				
	basic	weighed	filtered	filtered and weighed		basic	weighed	filtered	filtered and weighed	
AFFX-PheX-5_at (1)	206.00	107.10	206.00	120.03		441.00	408.75	441.00	409.42	
AFFX-PheX-5_at (2)	11.00	7.50	0.00	0.00		43.00	37.08	0.00	0.00	
AFFX-PheX-5_at (3)	0.00	0.00	0.00	0.00		7.00	6.33	0.00	0.00	
AFFX-PheX-5_at (4)	7.00	4.00	0.00	0.00		122.00	107.00	122.00	107.50	
AFFX-PheX-5_at (5)	0.00	0.00	0.00	0.00		340.00	280.58	340.00	306.75	
AFFX-PheX-5_at (6)	1.00	0.50	0.00	0.00		29.00	26.42	0.00	0.00	
AFFX-PheX-5_at (7)	80.00	40.70	0.00	0.00		4.00	2.67	0.00	0.00	
AFFX-PheX-5_at (8)	6.00	3.83	0.00	0.00		0.00	0.00	0.00	0.00	
AFFX-PheX-5_at (9)	0.00	0.00	0.00	0.00		23.00	21.70	0.00	0.00	
AFFX-PheX-5_at (10)	0.00	0.00	0.00	0.00		1.00	1.00	0.00	0.00	
AFFX-PheX-5_at (11)	293.00	186.18	293.00	189.95		31.00	29.83	0.00	0.00	
AFFX-PheX-5_at (12)	110.00	61.92	0.00	0.00		5.00	3.83	0.00	0.00	
AFFX-PheX-5_at (13)	1.00	0.50	0.00	0.00		0.00	0.00	0.00	0.00	
AFFX-PheX-5_at (14)	0.00	0.00	0.00	0.00		17.00	17.00	0.00	0.00	
AFFX-PheX-5_at (15)	0.00	0.00	0.00	0.00		93.00	89.50	93.00	90.00	
AFFX-PheX-5_at (16)	0.00	0.00	0.00	0.00		22.00	21.25	0.00	0.00	
AFFX-PheX-5_at (17)	0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00	
AFFX-PheX-5_at (18)	0.00	0.00	0.00	0.00		5.00	4.00	0.00	0.00	
AFFX-PheX-5_at (19)	0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00	
AFFX-PheX-5_at (20)	0.00	0.00	0.00	0.00		3.00	2.25	0.00	0.00	
AFFX-PheX-M_at (1)	0.00	0.00	0.00	0.00		6.00	6.00	0.00	0.00	
AFFX-PheX-M_at (2)	0.00	0.00	0.00	0.00		413.00	388.03	413.00	394.25	
AFFX-PheX-M_at (3)	0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00	
AFFX-PheX-M_at (4)	0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00	
AFFX-PheX-M_at (5)	0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00	
AFFX-PheX-M_at (6)	0.00	0.00	0.00	0.00		4.00	3.33	0.00	0.00	
AFFX-PheX-M_at (7)	0.00	0.00	0.00	0.00		12.00	11.50	0.00	0.00	
AFFX-PheX-M_at (8)	0.00	0.00	0.00	0.00		2.00	2.00	0.00	0.00	
AFFX-PheX-M_at (9)	0.00	0.00	0.00	0.00		58.00	45.07	0.00	0.00	
AFFX-PheX-M_at (10)	0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00	
AFFX-PheX-M_at (11)	0.00	0.00	0.00	0.00		327.00	241.87	327.00	258.67	
AFFX-PheX-M_at (12)	0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00	
AFFX-PheX-M_at (13)	8.00	2.07	0.00	0.00		54.00	44.20	0.00	0.00	
AFFX-PheX-M_at (14)	0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00	
AFFX-PheX-M_at (15)	4.00	2.17	0.00	0.00		90.00	84.62	90.00	85.75	
AFFX-PheX-M_at (16)	0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00	
AFFX-PheX-M_at (17)	3.00	1.50	0.00	0.00		0.00	0.00	0.00	0.00	
AFFX-PheX-M_at (18)	0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00	
AFFX-PheX-M_at (19)	0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00	
AFFX-PheX-M_at (20)	0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00	

probe name	cardio-set				mito-set			
	basic	weighed	filtered	filtered and weighed	basic	weighed	filtered	filtered and weighed
AFFX-PheX-3_at (1)	1.00	0.20	0.00	0.00	9.00	7.33	0.00	0.00
AFFX-PheX-3_at (2)	0.00	0.00	0.00	0.00	2.00	2.00	0.00	0.00
AFFX-PheX-3_at (3)	0.00	0.00	0.00	0.00	38.00	34.33	0.00	0.00
AFFX-PheX-3_at (4)	0.00	0.00	0.00	0.00	7.00	7.00	0.00	0.00
AFFX-PheX-3_at (5)	3.00	1.33	0.00	0.00	7.00	7.00	0.00	0.00
AFFX-PheX-3_at (6)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-PheX-3_at (7)	2.00	0.58	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-PheX-3_at (8)	3.00	1.45	0.00	0.00	3.00	3.00	0.00	0.00
AFFX-PheX-3_at (9)	1.00	0.50	0.00	0.00	3.00	2.20	0.00	0.00
AFFX-PheX-3_at (10)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-PheX-3_at (11)	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00
AFFX-PheX-3_at (12)	0.00	0.00	0.00	0.00	157.00	149.67	157.00	152.00
AFFX-PheX-3_at (13)	0.00	0.00	0.00	0.00	2.00	2.00	0.00	0.00
AFFX-PheX-3_at (14)	2.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-PheX-3_at (15)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-PheX-3_at (16)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-PheX-3_at (17)	0.00	0.00	0.00	0.00	19.00	19.00	0.00	0.00
AFFX-PheX-3_at (18)	1.00	1.00	0.00	0.00	1.00	1.00	0.00	0.00
AFFX-PheX-3_at (19)	6.00	2.75	0.00	0.00	9.00	5.67	0.00	0.00
AFFX-PheX-3_at (20)	13.00	3.78	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-ThrX-5_at (1)	814.00	591.00	814.00	615.25	9.00	9.00	0.00	0.00
AFFX-ThrX-5_at (2)	638.00	478.42	638.00	504.83	49.00	45.50	0.00	0.00
AFFX-ThrX-5_at (3)	77.00	49.03	0.00	0.00	160.00	130.17	160.00	136.50
AFFX-ThrX-5_at (4)	1337.00	1014.50	1337.00	1022.33	175.00	160.58	175.00	163.17
AFFX-ThrX-5_at (5)	69.00	42.57	0.00	0.00	96.00	64.92	96.00	71.33
AFFX-ThrX-5_at (6)	16.00	11.50	0.00	0.00	3.00	1.70	0.00	0.00
AFFX-ThrX-5_at (7)	1.00	0.50	0.00	0.00	183.00	176.83	183.00	179.00
AFFX-ThrX-5_at (8)	898.00	751.57	898.00	763.98	87.00	83.83	87.00	86.33
AFFX-ThrX-5_at (9)	40.00	29.42	0.00	0.00	192.00	180.50	192.00	182.50
AFFX-ThrX-5_at (10)	3.00	1.33	0.00	0.00	38.00	36.50	0.00	0.00
AFFX-ThrX-5_at (11)	0.00	0.00	0.00	0.00	109.00	105.20	109.00	105.33
AFFX-ThrX-5_at (12)	0.00	0.00	0.00	0.00	12.00	12.00	0.00	0.00
AFFX-ThrX-5_at (13)	0.00	0.00	0.00	0.00	9.00	8.25	0.00	0.00
AFFX-ThrX-5_at (14)	0.00	0.00	0.00	0.00	261.00	256.33	261.00	256.83
AFFX-ThrX-5_at (15)	1.00	0.50	0.00	0.00	61.00	60.25	0.00	0.00
AFFX-ThrX-5_at (16)	21.00	19.50	0.00	0.00	630.00	582.25	630.00	584.42
AFFX-ThrX-5_at (17)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-ThrX-5_at (18)	0.00	0.00	0.00	0.00	55.00	50.17	0.00	0.00
AFFX-ThrX-5_at (19)	1.00	1.00	0.00	0.00	46.00	35.50	0.00	0.00
AFFX-ThrX-5_at (20)	0.00	0.00	0.00	0.00	63.00	54.25	0.00	0.00

probe name	cardio-set					mito-set				
	basic	weighed	filtered	filtered and weighed		basic	weighed	filtered	filtered and weighed	
AFFX-Thx-M_at (1)	0.00	0.00	0.00	0.00	0.00	1.00	0.50	0.00	0.00	0.00
AFFX-Thx-M_at (2)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Thx-M_at (3)	2.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Thx-M_at (4)	7.00	4.17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Thx-M_at (5)	0.00	0.00	0.00	0.00	0.00	76.00	55.37	0.00	0.00	0.00
AFFX-Thx-M_at (6)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Thx-M_at (7)	0.00	0.00	0.00	0.00	0.00	84.00	76.00	84.00	79.33	0.00
AFFX-Thx-M_at (8)	0.00	0.00	0.00	0.00	0.00	4.00	3.50	0.00	0.00	0.00
AFFX-Thx-M_at (9)	0.00	0.00	0.00	0.00	0.00	34.00	24.42	0.00	0.00	0.00
AFFX-Thx-M_at (10)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Thx-M_at (11)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Thx-M_at (12)	0.00	0.00	0.00	0.00	0.00	3.00	2.33	0.00	0.00	0.00
AFFX-Thx-M_at (13)	0.00	0.00	0.00	0.00	0.00	1.00	0.33	0.00	0.00	0.00
AFFX-Thx-M_at (14)	0.00	0.00	0.00	0.00	0.00	42.00	39.03	0.00	0.00	0.00
AFFX-Thx-M_at (15)	0.00	0.00	0.00	0.00	0.00	17.00	15.50	0.00	0.00	0.00
AFFX-Thx-M_at (16)	0.00	0.00	0.00	0.00	0.00	57.00	51.28	0.00	0.00	0.00
AFFX-Thx-M_at (17)	0.00	0.00	0.00	0.00	0.00	14.00	13.00	0.00	0.00	0.00
AFFX-Thx-M_at (18)	1.00	1.00	0.00	0.00	0.00	7.00	6.33	0.00	0.00	0.00
AFFX-Thx-M_at (19)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Thx-M_at (20)	0.00	0.00	0.00	0.00	0.00	14.00	13.33	0.00	0.00	0.00
AFFX-Thx-3_at (1)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Thx-3_at (2)	0.00	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
AFFX-Thx-3_at (3)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Thx-3_at (4)	0.00	0.00	0.00	0.00	0.00	4.00	3.33	0.00	0.00	0.00
AFFX-Thx-3_at (5)	2.00	0.50	0.00	0.00	0.00	10.00	9.50	0.00	0.00	0.00
AFFX-Thx-3_at (6)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Thx-3_at (7)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Thx-3_at (8)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Thx-3_at (9)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Thx-3_at (10)	1.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Thx-3_at (11)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Thx-3_at (12)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Thx-3_at (13)	1.00	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Thx-3_at (14)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Thx-3_at (15)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Thx-3_at (16)	9.00	2.83	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Thx-3_at (17)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Thx-3_at (18)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Thx-3_at (19)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AFFX-Thx-3_at (20)	0.00	0.00	0.00	0.00	0.00	6.00	4.83	0.00	0.00	0.00

Chapter 7

General discussion

Overview of this thesis

Analysis of gene expression microarray data is complex and many issues remain to be solved. In this thesis several of those are investigated.

A one-step modeling procedure is developed, in order to avoid pre-processing steps that may adversely influence results of later steps or lose part of the information present in the data set (e.g. by summarizing values). Furthermore, the approach developed can directly take experimental design into account in the analysis. Several assumptions and modeling options are evaluated, including normality of the data at hand, and equal dispersion within all groups in the data set. Also the use of RNA spikes as controls within this framework is investigated.

Furthermore a method was developed that allows construction of gene profiles for classification that are not overspecific. This is very important for the reliable use of microarrays for classification, especially with diagnostic applications emerging.

Details of the analysis issues may differ dependent on specific flavor of array and analysis methods used, and on the type of study done and its goals. Also, since the field of genetics is in quick development, new applications and systems rapidly emerge and even more different aspects occur when considering other types of arrays (exon, tiling, proteomic) and other fields of application. However, the flexibility and ease of extension of the approaches presented facilitates applying them to new areas.

A one-step modeling approach

Generally, microarray data are analyzed following a stepwise procedure [1]. Pre-processing steps that are generally taken include normalization, background correction, mismatch correction, and if applicable summarizing multiple probe values into a single 'gene' value. In this thesis, it is shown that it is a good alternative to perform the analysis in one go. Normalization [2] is done to smooth out disturbing effects. Using a modeling approach, a preferable alternative to control for these effects is incorporating them directly into the gene expression models, which can be done if the experiment has been properly designed [3-5]. Also, with respect to intensities of multiple probes, the summarization step should be avoided. A previously published approach used a per-probe testing procedure where differential expression of each probe was determined irrespective of probeset membership, after which probe calls are combined for each probeset and significance determined [6]. In contrast, here a modeling approach is used and furthermore the probes are not considered irrespective of the probeset they belong to. The probe values are directly used in multivariate models to fit the probeset expression profile. This prevents losing information on the probe variability and introducing bias due to the choice of summarization method.

Background and mismatch correction

In this thesis, both normalization and summarization were incorporated in the common modeling approach, which was not yet done for background correction. However the models could be extended to include variables controlling for this as well. Furthermore, within the current Affymetrix experiments, mismatch probe intensities were not taken into account. Mismatch probes have been developed for probe-specific signal correction (correction for aspecific binding). Affymetrix methods use their intensities to correct for aspecific binding, but there has been no agreement if this improves analysis results. For example, the robust multi-array average (RMA) method [7], which is a commonly used alternative to the Affymetrix Microarray Suite (MAS) algorithm [8], does not use intensities of these probes. One third of mismatch intensities is higher than the corresponding perfect match intensities [7], especially for non-expressed transcripts. Because positive intensities are generally required, the difference cannot be used and more involved computations need to be used. It is of interest to note that mismatch probes are not present any more for each of the transcripts on several of the newer chip types developed by Affymetrix. Instead, mismatch intensities are only computed for certain controls to be used as corrections. This of course allows doubling of the number of genes to be included on the chip. Perfect match only algorithms can still be used, whereas the others need to be adapted with respect to how background correction is performed.

Design considerations

As discussed, this thesis shows that it is important to choose a design [3-5] suited for the research question at hand and to be able to apply a one-step procedure. This must be done before starting any experiment. Several disturbing effects, either biological or technical in nature, can then be controlled for. In general, a good rule of thumb is to use approximately half the maximum degrees of freedom available to prevent overfitting the data at hand. Within the experiments presented, covariates such as labeling efficiency and spike intensities were taken into account. Many others can be considered, e.g. batch, hybridization day. The important question is which ones are the most relevant. This question should be answered because the number of degrees of freedom available upon analysis is determined by the number of observation that can be collected, which is often a strong limiting factor for design in the microarray field. A possible approach is to take into account the factors that show the most variation within the dataset.

The degrees of freedom offered by a study performed, should always be exploited. However, one should take care with power calculations because the degrees of freedom should be accounted for at their proper levels. It is clear that if - for instance - one decides to control for a biological property such as age of the patient, it is of no use to analyze many repeats of the same sample material. In other word, only the total number of arrays done cannot tell how many degrees of freedom are available at each level. Thus, the replicates and repeats done should correspond to the covariates controlled for.

Also it would be of great interest to use the current modeling approach for gene selection in a study with well-characterized material to specifically study which of the many possible disturbing effects are most important for different array types and analysis goals. Results could then be compared for several of those array types and the variability of several factors estimated for several data sets. However, taking into account all possible factors would require a vast number of arrays, and even for a moderate number of factors the number of arrays required would already be prohibitively large, especially for commercial platforms.

Extensions to multivariate modeling

The use of multivariate models allows for the direct inclusion of multiple measurements on one unit (i.e. gene), for instance by incorporating either more probesets or more probes within one set, into the model. In this thesis, the use of a normally distributed random effect model assuming no order between the probes [9] was investigated. Also non-normally distributed random effects could be evaluated. Furthermore, the procedure could be extended to include the order of the probes along the sequence, which can be incorporated into the models for instance by including a first order autoregressive (AR1) dependence. This could provide additional insights into sequence specific effects and possibly correct for five prime versus three prime differences. However, the same approach can be more generally applied. Apart from being used to model probe intensities, ordered effects can be used to model longitudinal data such as time series and concentration range experiments. Trends can be estimated or functions of several types can be fitted to the profiles at hand.

Evaluating assumptions

The modeling approach used also allowed investigating the effect of two basic assumptions on the results of the analysis when determining which genes are differentially expressed between sample groups. These are normality of the data, and equal dispersion between groups. It is shown that relaxation of the latter strongly improves results. Furthermore the Gaussian distribution performs well for the data sets considered in this thesis. Although the difference in performance between the normal heteroscedastic models and all others was large, it may still be that other statistical distributions perform better than the Gaussian for other data sets, for example considering different patient groups. For example for data sets with more extreme differences, distributions with heavier tails may fit better. Based on these results, it is advisable to perform analysis methods that check for a difference in dispersion between groups. Also checking different distributions, e.g. on a subset of genes, to determine which one fits best may be a sensible step to take.

The programs used to develop the models discussed, are very flexible. Indeed, similar to modeling the location, the dispersion can also be modeled to depend on, for instance, disturbing factors or time. Furthermore, non-linear regression or incorporating non-linear group effects on location and dispersion are also possible within this framework.

Creating generalizable classifiers

A procedure for finding relevant distinguishing gene profiles that are not too specific is proposed. The method makes use of a principal component analysis implementation developed, to select genes in an unsupervised way and prevent too specific genes to be selected, by penalizing genes when they are also differentially expressed between other subgroups of samples than the distinction of interest. Generally, principal component analysis (PCA, [10]) is used to transform variables into components composed of several genes, which is also done in a study that combines PCA and discriminant analysis (DA) [11]. In contrast, within the procedure presented here, complete PCA results were used to reduce the size of the data set by keeping the most informative genes as variables for input in the DA, in order not to overstate certain processes and ignore others. To further ensure generalization, in the logistic models built to classify samples, the gene lists retrieved as classifiers were restricted from becoming too long and consume too many degrees of freedom. As an illustration, one pathway of interest is selected. For further applications, several pathways can best be selected by applying the procedure repeatedly to create an optimal classifier.

Besides the availability of general modeling functions, building more complex types of models is also facilitated by the increase of computational power. Thus for analysis of differential genes, complexity of analysis has become much less of an issue. But in some instances computational complexity is still a major issue, such as for PCA techniques, where enormous matrices (covariance matrices with the number of genes as both number of rows and columns) need to be processed (e.g. inverted). Here an alternative method using matrix decompositions is proposed that enables this step. However, in this thesis this implementation is applied to PCA, requiring the DA to be run separately. Both of these could also be performed together in one efficient algorithm, where the system would in one go pre-select genes and create classification profiles. This would not influence results, but be a more elegant mathematical solution.

The classification procedure developed is intended for data sets with several subgroups, to be able to use those as controls to get to real specific genes. In other data sets, random groupings may be used as controls. However, because these do not represent real biological groupings, they may make a less optimal correction. However, the PCA pre-selection step could be replaced with an alternative one based on the modeling approach described as discussed in the next section.

Future extensions

Combining it all

In the future it would be interesting to apply all methods investigated as one complete procedure to a few data sets. A list of differentially expressed genes selected by the modeling approach proposed in this thesis can be used as input into the classification procedure. This would make the procedure applicable to data-sets where no control groups can be created, without getting overspecific. This would show how all methods developed perform when jointly used. Eventually even an interactive tool could be developed. Unfortunately time did not allow doing this within the current project. Also it may be investigated if results are the same for other species because here only human data sets were used. Furthermore, the investigations to determine the best modeling options (with respect to distribution, how variance information is used, whether probes need to be modeled separately) were done apart from each other and only on one type of platform (Affymetrix GeneChips, [12]). It would be of interest to extend this to investigate several issues for several types of arrays together. For instance, it would be interesting to check different multivariate distributions besides the Gaussian.

Analyzing groups of genes together

The same approach used for the multivariate modeling of probes and probesets could also be used for analyzing groups of transcripts / genes together. Such a group could be defined based on known annotations (e.g. Gene Ontology annotations [13]), protein interaction data, clustering, and so on. The commonly taken approach nowadays is to first select differentially expressed genes based on the assumption of independence between them and then search for overrepresented groups of in fact dependent genes. It has been indicated however, that co-evaluating predefined groups of genes based on prior knowledge (Gene Set Enrichment Analysis, GSEA) [14], or co-evaluating generated candidate groups of genes [15], is very important. Consistence of experimental results between different studies indeed got higher than when single gene analysis approaches are used. The artificial procedure mentioned above, which moreover requires an algorithm to detect differential groups properly, could also be avoided within the modeling framework, by using groups of genes directly in the suggested modeling procedure. However, even further extending this to a model with two levels including probes, transcripts, and pathways would be far more difficult because software to do so is hardly available, especially if ordered effects must be taken into account. If packages become available that allow multi-level modeling, further reduction of the complete analysis procedure to a one-step method will get possible.

Application to new developments:

Exon chip

The modeling techniques described in this thesis can also be extended to be applied for newly developed chip types. Because most genes - probably at least three quarters of multi-exon genes [16] - are alternatively spliced, analysis techniques to detect splice variants have been published for classical oligo-based chips with several oligos per transcript, such as Affymetrix GeneChips [17]. However, recently specific arrays to detect these variant have been developed, including exon and exon junction arrays. Exon arrays contain specific probes for each exon, where the total number of probes per exon is in general limited due to the shorter sequence available for designing them and the available space on the chip. Exon junction arrays [16] are similar to exon arrays, but the oligos are designed to target sequences at the boundaries of several exons to specifically detect presence of alternatively splices variants. For each junction, this array type contains (overlapping) probes including the exon boundaries. Also, both exon probes and junction probes can be present on the same array [18]. For analysis of data from these types of arrays, algorithms have already been published [19]. Exon-based arrays can be expected to gain popularity soon and become widely used to detect gene expression. They pose further challenges for modeling, because they have fewer probes per 'feature' and sometimes more probe sequence overlap, but most importantly, they add an extra exon-level. The multivariate models used for probes and probesets can similarly be used for analyzing probes and exons or exons and transcripts, but not for all of them together. However, the approach described can be extended to deal with this situation as well. A second level of nesting would permit to incorporate probes, exons and transcripts or, alternatively, exons, transcripts and pathways. In case of modeling probes, exons, transcripts, and pathways, a third level will be needed. However, as indicated above, tools for fitting these kinds of models need to be developed.

Tiling array

More challenges will be generated from the tiling array. This array type contains probes designed along the complete genome to detect whether or not any part of the genome is expressed. This offers estimation of expression in a way unbiased by current knowledge on the locations of genes. Probes may overlap, or be positioned directly adjacent, or be spaced at regular intervals [20]. Tiling arrays have been constructed using oligos or PCR products. A disadvantage of the PCR-based type is that strand specificity is not present [21]. Resequencing arrays, which are however applied to genomic DNA, can be considered to be an extreme type of tiling array, with probes built for every base. The expression tiling arrays will allow for new transcripts to be discovered and methods to do so already emerge [22, 23]. These arrays have already shown that many untranslated RNAs appear to be present [20]. However, partially other sets of disturbing effects will play a role and methods for classical expression arrays will need to be adapted or extended to analyze tiling arrays [21].

The characteristics of the data sets produced by this array system and their magnitude will require much from analysis methods. For real discovery of alternatively spliced variants, an algorithm is needed to compute which combinations of exons are expressed. Here some exons may be part of several alternatively spliced gene products. Their total expression has to be decomposed into its constituent parts and a technique to return the most likely combination must be applied, which is far from easy to do.

Proteomic array

Whereas this thesis focuses on mRNA expression arrays, some discussion has been going on about whether these arrays really measure the quantity of interest. In some cases, where RNA itself is the functional end molecule this is very likely the case, such as for ribosomal (rRNA), transfer (tRNA), small inhibitory (siRNA) RNA types and so on. For mRNA this will depend on the research question at hand. Interest may be in genetic regulation. The protein 'in between' step in cases where the protein product controls activity of other genes (e.g. transcription factors) can be implicitly incorporated in the inferred model. Thus no measurements are included, but effects are modeled through predicting protein concentration from mRNA concentration. In cases where interest is in the final protein products, the relevant question is whether mRNA levels correlate with protein levels or not. Different opinions have been given on this [24]. The correlation may be insufficient especially for low abundant proteins [25] but it may be good for all expression levels as well [26].

Proteomic studies have been done using two-dimensional gel electrophoresis or chromatography, in combination with mass spectrometry [27]. Also yeast two-hybrid screens (that measure which proteins can interact with one another [28]) have been used. Studies have been performed combining protein interaction data or measurements with gene expression data to explore relationships [29, 30]. Combination of several data types will also allow genetic and protein networks, where both of the components are measured, to be reconstructed. Also for other types of studies it will provide more data, but analysis methods to deal with the combined data sets still need to be developed [28]. However, combining data, performing computational simulations of biological systems may even become within reach in the future [28].

Proteomic arrays [27, 28, 31] have many parallels with gene expression arrays, but it has been difficult to translate the profiling from gene expression arrays to proteomic arrays [32]. It is far more complex and costly to produce this kind of arrays. Specificity is mostly obtained by using antibodies, but also other peptides are being used. Proteins are more three-dimensional in nature than mRNAs which will make detection harder because they must be kept in their native form. Also proteins cannot be amplified [28]. Analysis methods for this kind of chips partly face the same challenges as with gene expression arrays, but will undoubtedly also face new ones and other covariates may have to be considered. It will depend on the systems that finally make it as a standard in the proteomics field, what the exact implications for analysis are

going to be. An initial goal of proteomic arrays, the characterization of all proteins expressed by a certain cell or system, has not yet been technically reached. However other goals, such as defining protein structure, activities and interactions, evaluating protein-DNA interactions, and finding targets of small molecules or antibodies are investigated now with more success [28]. Here proteins from clone collections can be attached to the array [31]. Each of these goals will have its own specific modeling options, but generally the properties of the resulting data sets will be similar to those from expression arrays. Especially, as soon as the technical challenges with this type of array have been solved and the arrays get in use to detect quantities of proteins present in a biological sample, the approach developed in this thesis should be adapted for and tested on the resulting data sets.

Diagnostics

Gene expression microarrays will also become very important for diagnostic purposes, increasing scale for this application as well. Concentrations of several molecules in blood have since long been used as a diagnostic for diseases such as infections, organ dysfunction (liver, heart) and many more. Generally a few to a maximum of tens of concentrations, based on common experience, were determined and interpreted by the medical doctor. With the use of microarrays for diagnostics far more concentrations can be measured. However, a direct consequence of this technology is that these results cannot be judged by eye by the clinician any more. In contrast to genetic research, arrays are not yet a standard diagnostic tool, but classifiers developed and implemented by microarray technology are being tested on larger scale for this purpose [33]. One may be tempted to introduce the word 'diagnostomics' for this type of medical diagnostic practice. So both in research as in medical practice, advanced possibilities have become available through the development of high-throughput techniques amongst which the microarray is one of the most significant.

Even more than in research, development of standardized techniques for sample and laboratory handling, data formats, information to be saved and also analysis is of great importance. Reliability and consistency between hospitals is a major criterion for a diagnostic tool to be accepted. This indicates that classifying profiles characterizing certain diseases will more than ever have to be generalizable, also between institutions, and as such the methods developed here can help improving quality. If all quality criteria are met, similar pre-formatted classifier chips containing modest numbers of genes can really be launched for many diseases, where results can be obtained by completely automated analysis methods and systems. Also, when diagnostic chips get in more common use, this will lower the cost. All these aspects will allow several types of microarrays to become the gold-standard for certain diagnostic procedures.

References

- 1 Leung YF, Cavalieri D: **Fundamentals of cDNA microarray data analysis.** *Trends Genet* 2003, **19**:649-659.
- 2 Quackenbush J: **Microarray data normalization and transformation.** *Nat Genet* 2002, **32 Suppl**:496-501.
- 3 Churchill GA: **Fundamentals of experimental design for cDNA microarrays.** *Nat Genet* 2002, **32 Suppl**:490-495.
- 4 Yang YH, Speed T: **Design issues for cDNA microarray experiments.** *Nat Rev Genet* 2002, **3**:579-588.
- 5 Kerr MK: **Design considerations for efficient and effective microarray studies.** *Biometrics* 2003, **59**:822-828.
- 6 Aimone JB, Gage FH: **Unbiased characterization of high-density oligonucleotide microarrays using probe-level statistics.** *J Neurosci Methods* 2004, **135**:27-33.
- 7 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* 2003, **4**:249-264.
- 8 Hubbell E, Liu WM, Mei R: **Robust estimators for expression analysis.** *Bioinformatics* 2002, **18**:1585-1592.
- 9 Lindsey JK: *Models for Repeated Measurements.* Oxford: Oxford University Press; 1999.
- 10 Butte A: **The use and analysis of microarray data.** *Nat Rev Drug Discov* 2002, **1**:951-960.
- 11 Mendez MA, Hodar C, Vulpe C, Gonzalez M, Cambiazo V: **Discriminant analysis to evaluate clustering of gene expression data.** *FEBS Lett* 2002, **522**:24-28.
- 12 Lipshutz RJ, Fodor SP, Gingeras TR, Lockhart DJ: **High density synthetic oligonucleotide arrays.** *Nat Genet* 1999, **21**:20-24.
- 13 Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, et al: **The Gene Ontology (GO) database and informatics resource.** *Nucleic Acids Res* 2004, **32**:D258-261.
- 14 Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proc Natl Acad Sci U S A* 2005, **102**:15545-15550.
- 15 Xiao Y, Frisina R, Gordon A, Klebanov L, Yakovlev A: **Multivariate search for differentially expressed gene combinations.** *BMC Bioinformatics* 2004, **5**:164.
- 16 Johnson JM, Castle J, Garrett-Engle P, Kan Z, Loerch PM, Armour CD, Santos R, Schadt EE, Stoughton R, Shoemaker DD: **Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays.** *Science* 2003, **302**:2141-2144.
- 17 Fan W, Khalid N, Hallahan AR, Olson JM, Zhao LP: **A statistical method for predicting splice variants between two groups of samples using GeneChip expression array data.** *Theor Biol Med Model* 2006, **3**:19.
- 18 Fehlbaum P, Guihal C, Bracco L, Cochet O: **A microarray configuration to quantify expression levels and relative abundance of splice variants.** *Nucleic Acids Res* 2005, **33**:e47.
- 19 Shai O, Morris QD, Blencowe BJ, Frey BJ: **Inferring global levels of alternative splicing isoforms using a generative model of microarray data.** *Bioinformatics* 2006, **22**:606-613.
- 20 Mockler TC, Chan S, Sundaresan A, Chen H, Jacobsen SE, Ecker JR: **Applications of DNA tiling arrays for whole-genome analysis.** *Genomics* 2005, **85**:1-15.
- 21 Royce TE, Rozowsky JS, Bertone P, Samanta M, Stolc V, Weissman S, Snyder M, Gerstein M: **Issues in the analysis of oligonucleotide tiling microarrays for transcript mapping.** *Trends Genet* 2005, **21**:466-475.
- 22 Frey BJ, Morris QD, Hughes TR: **GenRate: a generative model that reveals novel transcripts in genome-tiling microarray data.** *J Comput Biol* 2006, **13**:200-214.

- 23 Huber W, Toedling J, Steinmetz LM: **Transcript mapping with high-density oligonucleotide tiling arrays.** *Bioinformatics* 2006, **22**:1963-1970.
- 24 Grunenfelter B, Winzeler EA: **Treasures and traps in genome-wide data sets: case examples from yeast.** *Nat Rev Genet* 2002, **3**:653-661.
- 25 Gygi SP, Rochon Y, Franza BR, Aebersold R: **Correlation between protein and mRNA abundance in yeast.** *Mol Cell Biol* 1999, **19**:1720-1730.
- 26 Futcher B, Latter GI, Monardo P, McLaughlin CS, Garrels JI: **A sampling of the yeast proteome.** *Mol Cell Biol* 1999, **19**:7357-7368.
- 27 Lueking A, Cahill DJ, Mullner S: **Protein biochips: A new and versatile platform technology for molecular medicine.** *Drug Discov Today* 2005, **10**:789-794.
- 28 Patterson SD, Aebersold RH: **Proteomics: the first decade and beyond.** *Nat Genet* 2003, **33 Suppl**:311-323.
- 29 Ge H, Liu Z, Church GM, Vidal M: **Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*.** *Nat Genet* 2001, **29**:482-486.
- 30 Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R, Hood L: **Integrated genomic and proteomic analyses of a systematically perturbed metabolic network.** *Science* 2001, **292**:929-934.
- 31 Bertone P, Snyder M: **Advances in functional protein microarray technology.** *Febs J* 2005, **272**:5400-5411.
- 32 Mitchell P: **A perspective on protein microarrays.** *Nat Biotechnol* 2002, **20**:225-229.
- 33 Quackenbush J: **Microarray analysis and tumor classification.** *N Engl J Med* 2006, **354**:2463-2472.

Summary

Within the last few decades, genetics has gone through major changes in scale, leading to the birth of genomics. A genomics technique of major importance is the gene expression microarray, a glass slide containing spotted probes to simultaneously measure abundance of many gene transcripts (mRNAs) within one or more samples. Computational tools and algorithms needed to be developed to analyze the data obtained, facilitated by a major increase in computational power. **Chapter 1** discusses the microarray technology and analysis of the resulting data sets with respect to determination of differentially expressed genes, clustering of samples and genes, and finding genetic profiles to classify the samples.

In **Chapter 2** of this thesis, focus is on classification. A novel method is described that can handle large data sets and can be tuned to specific research questions. It is applied to select gene-classifiers to discriminate mitochondrial defects (knock-outs) from other types of knock-outs in yeast. An important aspect of the method is that classifiers are prevented from being overspecific to the data set at hand. As a first step, principal component analysis (PCA) is applied, in order to select the genes that explain most of the variability within the data set. This is determined by ranking all genes based on their maximal eigenvalue-weighted contribution in the components. Use of an algorithm based on matrix decompositions prevents computation of a very large covariance matrix. Next, discriminant analysis (DA) is applied to build classifiers, in which only genes included in the limited set selected by the previous step may be incorporated. In order to allow enough genes to fulfill criteria, a non-strict cut-off is used for genes to enter the functions. However, further discriminant functions are built that distinguish between other subgroups of yeast knock-outs than the mitochondrial one. This allows selecting genes that are specifically valuable to discriminate the mitochondrial subgroup, removing highly variable genes or those specific to the particularities of the data set. This approach leads to discovery of genes related to the condition being studied. Then, classifiers are built based on those genes, using logistic modeling. Proper control of the degrees of freedom consumed prevents overspecificity of these models as well. The thiamine system appears to be a pathway of interest for the discrimination of mitochondrial yeast knock-outs. Also, models based on it perform well. The same procedure could be used to add genes from one or more other pathways, improving modeling even further.

Modeling is also the focus of the next chapters, but in this case linear modeling in order to determine which genes are differentially expressed. In **Chapter 3** a modeling approach is used that allows taking into account several factors, dependent on the experimental design chosen. By incorporating features of the design directly as covariates, this modeling can also be used instead of normalizing for several factors, reducing the number of steps to take within the analysis procedure. In this chapter, a cMyBP-C knock-out mouse model is studied, for which the heterozygous mouse

shows a resembling phenotype to patients with familial hypertrophic cardiomyopathy (FHC) and the homozygous mouse develops clear hypertrophy at early age. Gene expression models are built that take labeling efficiency and, depending on the number of degrees of freedom available, sex of the animals and one or more spike (control RNA) covariates into account. Expression is compared in nine week old wild-type, heterozygous knock-out and homozygous knock-out mice. The most relevant findings are differential expression of energy related pathways, showing an increased energy demand; differential expression of the MAPK pathways, with upregulation of the JNK and p38 subsystems and downregulation of the ERK1/2 subsystem; and differential expression of apoptotic-related genes with activation of the major apoptotic route leading to caspase activity and release of cytochrome C from mitochondria. These findings are strongly linked to a pathological form of hypertrophy and are in the heterozygous mice already eminent before hypertrophy develops. Changes in cell structure and contraction related processes are present, but less prominent, possibly explained by the pre-hypertrophic (heterozygotes) or full hypertrophic state (homozygotes, where changes in gene expression may well not be detectable any more). Overall findings are similar between the two knock-out groups, with more prominent changes in the homozygous mice, which may or may not be a time effect. The usefulness of the approach used for detecting differentially expressed genes (and pathways) is clearly shown by these findings.

In **Chapter 4**, the value of modeling gene expression is evaluated more formally, based on Affymetrix GeneChip data sets. Two of the several commonly made assumptions in microarray data analysis are evaluated: the normality (Gaussian distribution) of the gene expression profiles and the equal dispersion of expression values for the several sample subgroups under study. The consequences of these assumptions are investigated by building several models for the expression profile of each gene and comparing their performance by the Akaike Information Criterion (AIC). This criterion corrects the minus log-likelihood (-LL) for the number of model parameters. Models are also compared using the Bayesian Information Criterion (BIC) and the unpenalized -LL, confirming results for several criteria. Also the number of non-fitting, e.g. non-converging, models is recorded. Models with and without a difference in dispersion are built for each of ten different statistical distributions: the Gaussian, Cauchy, gamma, logistic, Laplace (two parameter distributions) and the Student t, power exponential, generalized gamma, generalized logistic, skewed Laplace (three parameter distributions). Two Affymetrix data sets, considering two different kinds of human disease (cardiovascular and mitochondrial disease respectively) and using different types of chips (the hgu133a and the hgu95av2 chip respectively) are used as test data sets. The heteroscedastic Gaussian is the best fitting model amongst the ones tested. For MELAS disease, results are further evaluated, and the number of knowingly related processes retrieved increases in a heteroscedastic Gaussian approach. Also, the extra cost in computational time and degrees of freedom needed for the heteroscedastic as compared to the homoscedastic Gaussian model is very moderate. Even though the differences are

large in the test sets used, performing a test run on a subset of genes to determine the best settings and distribution for the data set at hand remains of use. In contrast, the value of the more complicated three parameter distributions appears to be very limited. Allowing for different models and different distributions for each gene separately would improve overall model fit even more, but an easy applicable method with an overall set of best settings was aimed for.

Affymetrix chips contain probesets to detect each different transcript, which consist of several probes each. Affymetrix provides information on how to group the probes. This grouping has been improved by others, because poor annotation has been considered to be the major limiting factor in quality of data analysis results for many platforms. Based on this grouping information, existing analysis methods use several approaches to calculate summarized probeset intensities from the contributing probe intensities. How this step is performed influences further results and leads to loss of between-probe variability information. As such, modeling is extended further in **Chapter 5**, finding out whether a multivariate normal (MVN) model directly based on the probe intensities performs better than a univariate normal (UVN) model based on summarized probeset intensities. To evaluate this modeling extension, the same data sets as in the previous chapter are used. However, in this case, it is not possible to use evaluation measures such as the AIC, BIC or -LL to find out whether a MVN model performs better than a UVN model, because the data input to the models differs. For this reason, further biologic interpretation and validation of the results is needed. Clearly fewer genes are found to be differentially expressed by the MVN modeling approach, as expected because between-probe variance is taken into account. Determining which biological pathways are affected shows that, although fewer pathways are significant because of the smaller number of differentially expressed genes, relevant changes are kept when switching to a MVN analysis. Also, the types of pathways retrieved remain the same between both analyses, and no novel ones are introduced by the MVN analysis, supporting that the last filters out non-significant findings. Furthermore, a MVN analysis leads to less non-fitting models. Where a UVN analysis already produces good results, especially in combination with a pathway approach, a MVN modeling approach still helps in focusing and performing as much a one-step and statistically sound analysis as possible.

Another design feature of Affymetrix GeneChips is the availability of a number of probesets that detect certain types of spiked-in RNA. Hybridization spikes (BioB, BioC, BioDn, CreX) are added to the sample before hybridization, while poly-A spikes (DapX, LysX, PheX, ThrX) are already added earlier, to control for amplification, labeling and hybridization. In **Chapter 6** it is investigated how these spikes perform when used as covariates in both the UVN and MVN models described above. Again, the same data sets are used, where both chip types contain two probesets (five prime and three prime) for each of the hybridization spikes and three probesets (five prime, middle and three prime) for each of the poly-A spikes. For each spike, spike probeset, and - in the MVN case - probe, statistics are computed on how often they are used as

a model covariate, where the hybridization spike BioB is excluded because of its low signal intensity. It is clear that the UVN models need much more spike covariates than the MVN models. The frequency of use of each particular spike, spike probeset, and spike probe differs strongly, but results vary between the data sets used. Furthermore, the use of spikes can provide information on the experiment or the samples under consideration. It is thus advisable to incorporate all spikes, where some may later be discarded from analysis, dependent on the specific data set. The five prime probesets are more frequently used than the three prime ones, indicating a higher correspondence to sample quality and the implementation of the protocol. At the spike probe level, generally only a few (out of the twenty per probeset) make up almost the total count for their probeset. Moreover, these probes tend to be adjacent on the sequence. It may be good to limit the allowable amount of spike covariates per model, because more spike covariates appear to be more aspecifically used. However, the fact that spike covariates are used by almost all models confirms the usefulness of incorporating spikes into the experimental design.

This thesis focuses on analyzing gene expression data sets, mainly by modeling approaches. It shows that these are a good alternative to statistical testing. Modeling is more flexible and more powerful to understand the biological processes taking place. Furthermore, models do not give rise to the issues related to multiple testing, because they can be evaluated with a criterion instead of a test. Therefore, whereas tests are developed to evaluate specific hypotheses, models are especially suited for exploratory studies, such as most microarray studies. Modeling facilitates taking into account the experimental design, by incorporating several disturbing effects as covariates. Proper design is very important, including sample sizes, pairing of the samples for two-color arrays, planning of the laboratory experiments, and allowing for recording information on covariates. This thesis also shows how modeling enables a (as much as possible) one-step analysis procedure for microarray analysis, which is preferable above sequences of steps that all make assumptions and lose information. Finally, the modeling approach set-up and applied in this thesis is very general and flexible, just as the techniques used (the R package with several statistical modeling libraries), making it more adaptable to novel developments, such as proteomic arrays, exon-specific arrays, tiling arrays or even others yet to emerge. When genetic classifiers are built, it is very important to prevent overspecificity to the data set. Specific care should be taken not to exploit all degrees of freedom and ideally several different data sets must be used for building and testing the classifier. Building classifiers that reliably work for different data sets and even in several settings will become of even more importance in the application of the microarray technology for diagnosis or 'diagnostomics'. The final chapter, **Chapter 7**, discusses and summarizes the contents and findings of this thesis, as well as its application to other (new) fields and possible further developments.

Samenvatting

Gedurende de laatste decennia heeft de genetica grote veranderingen in schaal doorgemaakt, wat geresulteerd heeft in het ontstaan van genomics. Een genomics techniek van groot belang is de genexpressie microarray, een glaasje waarop probes geprint zijn om tegelijkertijd de mate van aanwezigheid van vele gentranscripten (mRNAs) in een of meer monsters te meten. Computerprogrammatuur en algoritmes moesten ontwikkeld worden om de verkregen gegevens te analyseren, vergemakkelijkt door een belangrijke toename in rekenkracht. **Hoofdstuk 1** bespreekt de microarray techniek en de diverse methodes om de resulterende datasets te analyseren met betrekking tot bepaling van differentieel tot expressie komende genen, clustering van monsters en genen, en het vinden van genetische profielen om monsters te classificeren.

In **Hoofdstuk 2** van dit proefschrift ligt de nadruk op classificatie. Een nieuwe methode wordt beschreven die grote datasets kan verwerken en die afgestemd kan worden op specifieke onderzoeksvraagstellingen. Ze wordt toegepast om genetische classificatoren te vinden om mitochondriale defecten (knock-outs) van andere typen gist knock-outs te onderscheiden. Een belangrijk aspect van de methode is dat ze voorkomt dat classificatoren overspecifiek zijn voor de dataset onder beschouwing. Als een eerste stap wordt principal component analysis (PCA) toegepast om de genen te selecteren die de meeste variabiliteit in de dataset verklaren. Dit wordt bepaald door alle genen in de dataset te ordenen op grond van hun maximale eigenwaarde-gewogen deelname in de componenten. Het gebruik van een algoritme dat is gebaseerd op matrix decomposities, maakt dat geen enorm grote covariantiematrix berekend hoeft te worden. Hierna wordt discriminant analyse (DA) toegepast om classificatoren te bouwen, waarin alleen genen opgenomen mogen worden die in de beperkte verzameling, geselecteerd door de vorige stap, aanwezig zijn. Om voldoende genen aan de criteria te laten voldoen, wordt een niet-strikt afkappunt gebruikt voor genen om in de functies terecht te komen. Daarnaast worden echter discriminant functies gebouwd die onderscheid maken tussen andere groepen gist knock-outs dan de mitochondriale. Dit maakt het mogelijk om alleen die genen te houden die specifiek waardevol zijn om mitochondriale knock-outs van de andere te onderscheiden, waarbij hoog variabele genen of die welke gerelateerd zijn aan de eigenaardigheden van de dataset, verwijderd worden. Deze aanpak leidt tot de opsporing van genen die gerelateerd zijn aan de bestudeerde conditie. Ook worden waardevolle classificatoren gebouwd gebaseerd op deze genen, middels logistische modellering. Correcte beperking van de gebruikte vrijheidsgraden voorkomt ook voor deze modellen dat ze overspecifiek raken. Het thiamine systeem blijkt een belangrijk proces te zijn voor de discriminatie van mitochondriale gist knock-outs. Erop gebaseerde modellen presteren ook naar behoren. Dezelfde methode zou gebruikt kunnen worden om meer genen van een of meer andere processen toe te voegen, waardoor de modellering nog beter wordt.

Modellering is ook het kernpunt van de volgende hoofdstukken, maar in dit geval lineaire modellering om te bepalen welke genen differentieel tot expressie komen. In **Hoofdstuk 3** wordt een modelleeraanpak gebruikt die toestaat om met verschillende factoren rekening te houden, afhankelijk van het experimenteel ontwerp dat gekozen is. Door ontwerpaspecten direct als covariaten mee te nemen, kan deze modellering ook gebruikt worden in plaats van normalisatie voor verscheidene factoren, waardoor het aantal stappen in de analyseprocedure verminderd wordt. In dit hoofdstuk wordt een cMyBP-C knock-out muizenmodel bestudeerd, waarvoor de heterozygote muis eenzelfde fenotype laat zien als patiënten met familiale hypertrofe cardiomyopathie (FHC) en de homozygote muis een duidelijke hypertrofie ontwikkelt op jonge leeftijd. Genexpressiemodellen worden gebouwd die labeling-efficiëntie en, afhankelijk van het aantal beschikbare vrijheidsgraden, geslacht van de dieren en een of meer spike (controle RNA) covariaten meenemen. Expressie wordt vergeleken in negen weken oude wildtype, heterozygote knock-out, en homozygote knock-out muizen. De meest relevante bevindingen zijn differentieële expressie van energie gerelateerde processen, die een verhoogde energiebehoefte laten zien; differentieële expressie van de MAPK processen, met upregulatie van de JNK en p38 subsystemen en downregulatie van het ERK1/2 subsysteem; en differentieële expressie van apoptose gerelateerde genen met activatie van de belangrijke apoptotische route die tot caspase activiteit en het vrijkomen van cytochroom C uit de mitochondria leidt. Deze bevindingen zijn sterk gekoppeld aan een pathologische vorm van hypertrofie en zijn in de heterozygote muizen al duidelijk aanwezig voordat hypertrofie optreedt. Veranderingen in celstructuur en contractie gerelateerde processen zijn aanwezig, maar minder uitgesproken, mogelijk verklaard door de prehypertrofe staat (heterozygoten) dan wel volledige hypertrofe staat (homozygoten, waar veranderingen in genexpressie niet meer detecteerbaar hoeven te zijn). Over het geheel genomen zijn de bevindingen dezelfde in de twee knock-out groepen, met meer prominente veranderingen in de hypertrofe homozygote muizen, wat al dan niet een tijdseffect kan zijn. Het nut van de gebruikte aanpak voor de opsporing van differentieel tot expressie komende genen (en processen) wordt duidelijk aangetoond door deze bevindingen.

In **Hoofdstuk 4** wordt de waarde van het modelleren van genexpressie meer formeel geëvalueerd, gebaseerd op Affymetrix GeneChip datasets. Twee van de verscheidene geregeld gedane aannames in microarray data analyse worden geëvalueerd: de normaliteit (Gaussiaanse verdeling) van de genexpressieprofielen en de gelijke dispersie van expressiewaardes voor de verschillende groepen monsters die bestudeerd worden. De gevolgen van deze aannames worden onderzocht door verschillende modellen te bouwen voor het expressieprofiel van elk gen en hun prestaties te vergelijken middels het Akaike Information Criterion (AIC). Dit criterium corrigeert de minus log-likelihood (-LL) voor het aantal parameters in het model. Modellen worden ook vergeleken met het Bayesian Information Criterion (BIC) en de standaard -LL, de resultaten bevestigend voor verschillende criteria. Ook het aantal niet-fittende, bijvoorbeeld niet convergerende, modellen wordt

bijgehouden. Modellen met en zonder een verschil in dispersie worden gebouwd voor elk van tien verschillende statistische verdelingen: de Gaussiaanse, Cauchy, gamma, logistische, Laplace (verdelingen met twee parameters) en de Student t, power exponential, gegeneraliseerde gamma, gegeneraliseerde logistische, skewed Laplace (verdelingen met drie parameters). Twee Affymetrix datasets, die twee verschillende types humane ziektebeelden betreffen (respectievelijk cardiovasculair en mitochondriaal) en die gebruik maken van verschillende types chips (respectievelijk de hgu133a en de hgu95av2 chip) worden gebruikt als test datasets. Het heteroscedastische Gaussiaanse model is het best passende van de geteste modellen. Voor het MELAS syndroom worden resultaten verder geëvalueerd en het aantal bekend verband houdende processen dat gevonden wordt neemt toe in een heteroscedastische Gaussiaanse aanpak. Ook zijn de extra kosten met betrekking tot rekentijd en vrijheidsgraden heel beperkt voor de heteroscedastische ten opzichte van de homoscedastische Gaussiaanse modellering. Ondanks dat de verschillen groot zijn voor de geteste datasets, blijft het van nut om voor andere datasets een testronde op een deel van de genen te doen om de beste modelleerinstellingen en de best passende verdeling te selecteren. De waarde van de meer gecompliceerde verdelingen met drie parameters blijkt daarentegen erg beperkt. Het mogelijke gebruik van een ander model en een andere verdeling voor elk gen op zich zou in totaal bekeken tot nog betere modellen leiden, maar het doel was een gemakkelijk toepasbare methode te vinden en de algehele beste instellingen.

Affymetrix chips bevatten probesets om elk verschillend transcript te detecteren, die elk uit verscheidene probes bestaan. Affymetrix voorziet in informatie hoe deze probes gegroepeerd moeten worden. Deze groepering is verbeterd door anderen, omdat slechte annotatie beschouwd is als de meest limiterende factor met betrekking tot de kwaliteit van de analyseresultaten voor veel platforms. Gebaseerd op deze groeperings-informatie, gebruiken bestaande analysemethoden verschillende technieken om samengevatte probesetintensiteiten te berekenen vanuit de bijdragende probe intensiteiten. Hoe deze stap wordt uitgevoerd, beïnvloedt verdere resultaten en leidt tot verlies van informatie over de variabiliteit tussen de probes. Daarom wordt in **Hoofdstuk 5** de modellering verder uitgebreid en wordt bekeken of een multivariaat normaal (MVN) model dat direct gebaseerd is op probe intensiteiten beter presteert dan een univariaat normaal model (UVN) gebaseerd op samengevatte probeset intensiteiten. Om deze modeluitbreiding te evalueren worden dezelfde datasets als in het vorige hoofdstuk gebruikt. Het is echter in dit geval niet mogelijk om evaluatiematen zoals de AIC, BIC of -LL te gebruiken om te beoordelen of een MVN model beter presteert dan een UVN model, omdat de gemodelleerde data verschillen. Om deze reden is verdere biologische interpretatie en validatie van de resultaten nodig. Duidelijk minder genen worden als differentieel tot expressie komend aangewezen door de MVN modellering, zoals ook verwacht omdat probe variantie meegenomen wordt. Bepaling van welke biologische processen aangedaan zijn, laat zien dat relevante veranderingen bewaard blijven bij het overgaan naar een MVN analyse, hoewel minder processen significant zijn vanwege het kleinere aantal

differentieel tot expressie komende genen. Ook de gevonden typen van processen blijven dezelfde tussen de beide analyses, en geen nieuwe worden geïntroduceerd door de MVN analyse, ondersteunend dat deze laatste niet-significante bevindingen uitfiltert. Daarnaast leidt een MVN analyse tot minder niet-passende modellen. Waar een UVN analyse al tot goede resultaten leidt, met name in combinatie met een procesaanpak, helpt een MVN aanpak nog steeds om verder te focussen en zoveel mogelijk een éénstaps en statistisch correcte analyse te verrichten.

Een ander ontwerpaspect van Affymetrix GeneChips ligt in de beschikbaarheid van een aantal probesets die bepaalde typen spiked-in RNA detecteren. Hybridisatie spikes (BioB, BioC, BioDn, CreX) worden aan het monster toegevoegd vóór hybridisatie, terwijl poly-A spikes (DapX, LysX, PheX, ThrX) al eerder toegevoegd worden om te controleren voor amplificatie, labeling en hybridisatie. In **Hoofdstuk 6** wordt uitgezocht hoe deze spikes presteren als ze gebruikt worden als covariaten in zowel de UVN als MVN modellen die hierboven beschreven zijn. Opnieuw worden dezelfde datasets gebruikt, waar beide chiptypes twee probesets bevatten (vijf accent en drie accent) voor elk van de hybridisatie spikes en drie probesets (vijf accent, midden en drie accent) voor elk van de poly-A spikes. Voor elke spike, spike probeset, en - in het geval van MVN modellering - probe, worden statistieken bijgehouden met betrekking tot hoe vaak ze gebruikt worden als modelcovariaat, waarbij de hybridisatie spike BioB wordt uitgesloten vanwege haar lage signaalsterkte. Het is duidelijk dat de UVN modellen veel meer spike covariaten nodig hebben dan de MVN modellen. De gebruiksfrequentie van elke specifieke spike, spike probeset en spike probe verschilt sterk, maar resultaten variëren tussen de twee gebruikte datasets. Verder kan het gebruik van spikes informatie verschaffen over het experiment of de monsters onder beschouwing. Het is dan ook aan te raden om alle spikes mee te nemen, waarna sommige dan bij de verdere analyse buiten beschouwing gelaten kunnen worden, afhankelijk van de specifieke dataset. De vijf accent probesets worden vaker gebruikt dan de drie accent probesets, indicierend dat ze meer beïnvloed worden door de monsterkwaliteit en de uitvoering van het protocol. Op het spike probe niveau, zorgen er in het algemeen slechts een paar (van de twintig per probeset) voor vrijwel het totale gebruik van de betreffende probeset. Bovendien neigen deze probes aangrenzend te zijn op de sequentie. Het is wellicht verstandig om het toegestane aantal spike covariaten per model te beperken, omdat meer spike covariaten meer specifiek gebruikt lijken te worden. Hoe dan ook, het feit dat spike covariaten door vrijwel al de modellen gebruikt worden, bevestigt het nut van het opnemen van spikes in het ontwerp van het experiment.

Dit proefschrift betreft analyse van genexpressie datasets, in het bijzonder door modellering. Het laat zien dat dit een goed alternatief vormt voor statistisch testen. Modellering is meer flexibel en krachtiger om de biologische processen die zich afspelen te leren begrijpen. Bovendien leidt modellering niet tot de problemen gerelateerd aan meervoudig testen, omdat modellen geëvalueerd kunnen worden met behulp van een criterium in plaats van een test. Daarom zijn modellen, terwijl testen ontwikkeld zijn om een specifieke hypothese te onderzoeken, bijzonder

geschikt voor verkennende studies, zoals de meeste microarray studies. Modelling maakt het gemakkelijker om het experimenteel ontwerp mee te nemen, door diverse versturende effecten als covariaten op te nemen. Goed ontwerp is heel belangrijk, inclusief de monsteraantallen, het koppelen van de monsters voor tweekleuren arrays, het plannen van de laboratoriumexperimenten, en het mogelijk maken informatie over covariaten te vergaren. Daarnaast, toont dit proefschrift ook aan hoe modelling een (zoveel mogelijk) éénstaps procedure voor microarray analyse mogelijk maakt, die te verkiezen is boven opeenvolgingen van stappen die elk aannames maken en informatie verliezen. Tenslotte is de modelleeraanpak die in dit proefschrift wordt opgezet en toegepast, zeer generiek en flexibel, net als de gebruikte technieken (het R-pakket met verscheidene statistische modelleerbibliotheken), wat haar beter aanpasbaar maakt aan nieuwe ontwikkelingen zoals proteoom (eiwit) arrays, exonspecifieke arrays, tiling arrays of zelfs andere typen die nog gaan verschijnen. Wanneer genetische classificatoren gebouwd worden, is het heel belangrijk overspecificiteit voor de gebruikte dataset te voorkomen. In het bijzonder moet opgelet worden niet alle vrijheidsgraden op te gebruiken en idealiter dienen verschillende datasets gebruikt te worden om de classifier te bouwen en te testen. Het bouwen van classificatoren die betrouwbaar werken voor verschillende datasets en zelfs in verschillende settings zal nog belangrijker worden bij de toepassing van de microarray technologie voor diagnostische doeleinden, 'diagnostomics' te noemen. Het laatste hoofdstuk, **Hoofdstuk 7**, beschrijft de inhoud en bevindingen van dit proefschrift in het kort en bediscussieert ze, net zoals de toepassing ervan op andere (nieuwe) velden en mogelijke verdere ontwikkelingen.

Dankwoord

Prof.dr. Geraedts, ik wil jou graag als eerste bedanken voor je inzet om samen met me tot een mooi proefschrift te komen. Beste Joep, ik ben heel blij dat jij bereid was in de laatste fase van mijn promotietraject, de rol van promotor op je te nemen en me de daarbij behorende ondersteuning en begeleiding te bieden. Ook bedank ik mijn beide copromotoren, Patrick Lindsey en Bert Smeets, voor jullie begeleiding en dat ik altijd wist wat ik aan jullie had. Patrick, thanks for getting (pulling?) me into statistics! Ik bedank ook jou, Bob Vlietinck, voor je rol als begeleider eerder tijdens mijn promotietraject. Naast deze mensen wil ik graag iedereen bedanken die aan het tot stand komen van dit proefschrift heeft bijgedragen of me op welke wijze dan ook gesteund heeft de afgelopen jaren.

Ik bedank mijn collega's voor de jaren plezierige samenwerking. Als eerste wil ik de 'AIO-kamer' bedanken. Bianca en Rudy, wij vormden een jaar of twee geleden het AIO groepje op 'popgen' en ik vind nog steeds dat we echt een topteam waren met ons drietjes. We hebben veel plezier gehad op de kamer waar we toen zaten, maar ik vond het ook heel fijn dat ik altijd wist bij jullie terecht te kunnen als dat even nodig was. Bianca ook dank voor je ervarings-informatie over 'promotiezaken'. Rudy dank voor je hulp met posters en onze leuke gezamenlijke congresbezoeken, veel succes met het afronden van jouw proefschrift! Lorraine, jij bent helaas al een hele tijd op ruime afstand aan het werk, maar gelukkig spreek ik je toch nog wel eens. Jij ook heel veel succes met de laatste loodjes van jouw proefschrift! Nicole, wij hebben ons prima vermaakt, bij het sporten (the place to talk), maar ook bij het organiseren van het vakgroepuitstapje waarbij we ons gedwongen zagen bier te gaan drinken, maar toch nog onder de modder kwamen te zitten... Ruben, Florence en An, jullie kwamen later ons AIO-groepje versterken en droegen elk bij aan de sfeer op de kamer en het behoud van een hechte AIO groep (waarbij An me net - niet - heeft weten te ontlopen). Het was fijn op een kamer te zitten met jullie allemaal! Ook nu een aantal van ons geen aio meer is, en ieder ook zijn eigen weg begint te gaan, hoop ik dat wij toch contact blijven houden met elkaar! Ook de andere collega's wil ik bedanken voor de leuke tijd op de afdeling: Aimée (Molo5?), Anja, Bieke (succes met het huis), Caroline (dank voor de RT-data hulp), Ellen, Erika (thx voor het uitgaan, we moeten naar Luik!), Fons (knepeke?), JosD (zwick), Liesbeth, Marij, Marion (heb je al weer een outfit klaar voor Carnaval?), Miroslav, Rob, Roselie (onze raseigen Patsy), Rosy (bedankt voor je gastvrijheid), Ton (dank voor 'instant' IT assistentie, zelfs van thuis), Torik (dank voor je adviezen bij een aantal van de papers). Ik bedank ook de voormalige collega's en stagiaires, meer of minder nadrukkelijk aanwezig. Een aantal stagiaires heb ik gelukkig na jullie stage nog wel vaker op de uni gezien en gesproken: Anne (hyperder dan ik, ongelofelijk!), Blanche (jij ook heel erg veel succes met je verdediging!!!), Debby (proficiat met je dr.-titel!), Jeske (nu snel bijpraten!), Joost, Joyce, Leentje, Marlon, Olaf, Sabine. Verder bedank ik de mensen

van klinische genetica (3X), moleculaire genetica, het RoMa overleg, en de AIO's van de andere afdelingen binnen en buiten onze vakgroep, zonder dat ik jullie allemaal bij naam kan noemen. Beste (ex)AIO's: de tijd ligt al weer ver achter ons dat we nog met zo weinig waren als promovendi van de secties van de vakgroep genetica en celbiologie en de moleculaire cardiologie, dat we zelfs allemaal samen bij iemand thuis konden eten. Deze bijeenkomsten vond ik altijd heel gezellig en ze versterkten de band tussen ons allemaal. Bij het 'voorbereiden' van onze inbreng in het capgroepoverleg deelden we altijd nuttige informatie, maar meestal nog véél meer ander interessant nieuws!

Ook wil ik de mensen van buiten onze eigen groep bedanken die met me hebben samengewerkt aan de diverse hoofdstukken: Ralf Peeters, dank voor je technische ondersteuning en uitleg (uit het hoofd!) over matrixbewerkingen; Ronald Westra, dank voor onze overleggen en zeker ook voor de gesprekken over allerlei andere interessante wetenschappelijke onderwerpen dan die waarover het eigenlijk moest gaan; Ger van der Vusse, dank voor je heldere inzichten over hoe de muizenpaper verbeterd kon worden zowel tekstueel als qua inhoud; Monique Bolotin-Fukuhara, *merci* for our discussions and your help related to the molecular biology of yeast; Lucie Carrier, thanks for always providing me with information on your mouse model and for discussing our findings; Jim Lindsey, thanks for revising several of my manuscripts; Koos Vrieze, dank voor de begincontacten met genetica voor een stage, die tot deze aanstelling geleid heeft.

Verder wil ik natuurlijk mijn vrienden bedanken voor de leuke dingen die we de afgelopen jaren gedaan hebben. Tim, ik ben blij dat jij al die jaren sinds onze studietijd zo een enorm hechte vriend gebleven bent, bij wie ik altijd terecht kon met alles en die altijd 's nachts beschikbaar was om de (vaak te) lange werkdag nog even ontspannend af te sluiten met een spelletje of het afwerken van een paar keys of sts. Lianne en Dennis, bedankt dat ik bij jullie zo vaak welkom was voor een avondje rustig hangen, en Lianne, ook bedankt voor de vele verhelderende discussies en gesprekken in de restaurants of kroegen van Maastricht. Ik wil (dr.) Mark graag bedanken voor het doorlezen van delen van mijn proefschrift. Verder bedank ik ook alle anderen die eraan meegeholpen hebben dat de afgelopen jaren ook hun ontspanning kenden nu en dan, Boukje en Stef, Cliff, Fiona (dank ook voor vet-support!), Joanna (Cockney-Upon-Tyne), Stan, Tessa, het donderdag-sportgroepje, en de BLAFfers Marcel en Susan, Mark en Katrien, Patrick, Kim, Marie-Pauline. Lars Eijssen, bedankt voor de diepe discussies over van alles en nog wat, we lijken duidelijk in meer op elkaar dan alleen onze naam!

Ik ben ook dankbaar voor de geweldige familie die ik om me heen heb, neven en nichten, ooms en tantes, bedankt voor de vele gezelligheid. Ik hoop dat we als familie zo met elkaar blijven omgaan zoals het nu is, echt top! Sander en Marcel, sorry dat ik zo weinig tijd had de laatste jaren voor 'Belsj' of Café Leen en Pieter, ook bezoeken voor nachtelijke koffie en/of shoarma bij de burens konden helaas niet meer de laatste tijd. Nicole en Robin, ik hoop dat ook wij elkaar nu weer makkelijker kunnen zien,

ondanks de afstand tussen Hilversum en zuid-Limburg. Chris (not accidentally categorized as a family member here), thanks for your visits and nice discussions, sharing your British points of view on many topics.

Uiteraard wil ik ook jullie, Mike en Jorg, mijn paranimfen, van harte bedanken. Jullie hebben er zeker mede voor gezorgd dat het resultaat er nu ligt! Zowel praktisch als mentaal zijn jullie heel belangrijk geweest. Mike bedankt voor het maken van de illustraties voor de introductie van dit proefschrift, Jorg voor het verzamelen en ordenen van welke informatie ik dan ook nodig had! Naast jullie rol als paranimf wil ik ook zeker niet nalaten jullie als persoon te bedanken. Mike, nadat jij stage gelopen had bij ons op de afdeling merkte ik snel dat, hoewel dit misschien niet meteen bleek, wij ontzettend goed met elkaar overweg konden. Het verbaast me dan ook niks dat we sindsdien ontzettend goede vrienden geworden en gebleven zijn. Bedankt PICie, voor alle onzin die we samen ondernomen hebben en dat ik altijd dingen met je kon bespreken waar ik mee zat (of volgens jou mee zat...). Ook jij, Gonda, wil ik bedanken, in eerste plaats natuurlijk voor je vriendschap, maar zeker ook dat ik nooit te veel was als een derde persoon die geregeld de hele dag in huis rondhing. Daarnaast ook bedankt voor het uitlenen van je vriend als ik hem oppiepte als wat moest gebeuren!

Jorg, ik bedank je voor de leuke dingen die we altijd samen gedaan hebben, waaronder zeker ook de vakanties die we vroeger ondernomen hebben. Jij bent altijd een geweldige broer geweest! Ook de laatste tijd, terwijl ik nooit vond dat mijn schema het toeliet, heb je me nu en dan even gedwongen te stoppen, even rust te nemen en een paar dagen met je ergens naar toe te gaan. Ook jou, Nicole, wil ik bedanken voor de dagjes uit en de etentjes met Jorg en mij. Fijn dat ook jij aandrong dat ik gewoon meemoest. Jullie samen hebben er toch voor gezorgd dat ik soms mijn gedachten even kon verzetten (ook al smokkelde ik al eens werkspullen mee voor onderweg, sorry daarvoor...). Tot slot bedank ik jullie, pa en ma, zonder jullie steun bij alles wat ik tot nu toe in mijn leven ondernomen heb, had het natuurlijk nooit gekund allemaal. De (over)ambitieuze keuzes die ik geregeld gemaakt heb, zullen ongetwijfeld - naast mezelf - ook jullie soms nogal belast hebben. Dank jullie wel voor altijd achter me te staan. Maar natuurlijk zeker ook bedankt voor jullie steun en advies bij mijn promotie en alles daar rond, echt in alle opzichten!

Er zijn ongetwijfeld nog vele anderen die ik dank verschuldigd ben en die ik hier nog niet bij name genoemd heb. Sorry dat je naam hier niet staat, maar ik denk ook aan jou, alleen net even niet op dit moment... Bij deze dus alsnog: bedankt!

Lars, drL, drLRC, LRC, drLars, dokter Lars

Curriculum Vitae

Lars Maria Theo Eijssen werd geboren op 30 mei 1976 in het zuid-Limburgse dorp Schimmert. Van 1988 tot 1994 doorliep hij het Voorbereidend Wetenschappelijk Onderwijs (VWO), afdeling Gymnasium, aan het Stella Maris College te Meerssen. Uit een brede interesse aan vervolgstudies viel zijn keuze vervolgens op de opleiding Geneeskunde. Hiertoe volgde hij eerst een verkorte cursus scheikunde aan het James Boswell Instituut van de Universiteit Utrecht. Vervolgens begon hij enthousiast aan de medische opleiding aan de Universiteit Maastricht, toen nog Rijksuniversiteit Limburg. Deze opleiding verschaftte hem theoretische, maar zeker ook praktische medische kennis. Omdat ook interesse voor meer technische vakken bleef, begon Lars in 1995 aan de opleiding Kennistechnologie aan de Universiteit Maastricht en aan de Universiteit Hasselt, toentertijd Limburgs Universitair Centrum geheten, in het Belgische Diepenbeek. In het kader van de laatste opleiding studeerde hij een maand informatica aan Baylor University te Waco, Texas, VS en nam hij deel aan een aantal projecten op het gebied van logistiek en automatisering voor verschillende bedrijven.

Na het doen van een wetenschapsstage bij de afdeling Epidemiologie van de Universiteit Maastricht, haalde Lars het Doctoraal geneeskunde *cum laude* in het jaar 1999. Nadat hij een stage bij de afdeling Genetica en Celbiologie van dezelfde universiteit had afgerond, mocht hij ook zijn doctoraal Kennistechnologie, specialisatie bedrijfs- en toegepaste wiskunde, in ontvangst nemen, tevens *cum laude*. Dit laatste diploma gaf ook aanleiding tot officiële toekenning van een gelijkwaardigheidverklaring met het Licentiaat Toegepaste Informatica door het Vlaamse Departement Onderwijs.

Omdat hij nog steeds grote interesse voor de wetenschap had, en zijn steentje bij wilde dragen aan de medische kennis en adequate patiëntenzorg, begon Lars in 2000 aan een promotietraject als Assistent in Opleiding bij de kersverse sectie Populatiegenetica, Genomics & Bioinformatica van de Universiteit Maastricht. Hier hielp hij naast zijn eigen onderzoek mee met het opzetten van een aantal andere activiteiten, waaronder veel ICT taken binnen de groep. Ook behaalde hij enkele aanvullende certificaten op biologisch en wiskundig gebied. Het doorlopen promotietraject leidde in 2006 tot de publicatie en verdediging van het proefschrift 'Analyse van microarray genexpressie datasets' met Prof.dr. J.P.M. Geraedts als promotor. Nu hij bij het afronden van zijn promotie weer nieuwe keuzes kan maken, is het voorsnog onduidelijk welke wegen hij de komende jaren zal gaan bewandelen.

Curriculum Vitae

Lars Maria Theo Eijssen was born on May 30th 1976 in the southern Dutch village Schimmert. From 1988 until 1994 he attended secondary school, Gymnasium, at the Stella Maris College in Meerssen. From a broad interest of scientific programs he chose to be a medical student. In order to do so, he first did an intensive course in chemistry at the James Boswell Institute of Utrecht University. After this, he started enthusiastically with medical training at Maastricht University. Here he acquired theoretical, but certainly also practical medical knowledge. Because he was also still interested in more technical fields, Lars started in 1995 with the program of Knowledge Engineering at both Maastricht University and the Belgian Hasselt University. As part of this program, he followed a course in computer science at Baylor University, Waco, Texas, USA for a month and participated in a number of projects related to logistics and information technology for several companies.

After completion of a scientific internship at the Department of Epidemiology of Maastricht University, Lars received the Masters of Medicine *cum laude* in the year 1999. After finishing an internship at the Department of Genetics and Cell Biology of the same university, he also obtained the Masters of Science in Knowledge Engineering, specialization business and applied mathematics, *cum laude*. Based on this, he was also granted an official certificate of equality to the Belgian Licentiate Applied Computer Science by the Flemish Department of Education.

Because he still possessed a major interest in science and wanted to contribute to medical knowledge and quality patient care, Lars started in 2000 as a PhD student at the brand new Department of Population Genetics, Genomics & Bioinformatics at Maastricht University. Apart from his own project, he also assisted in setting up several other activities, amongst which many ICT related tasks within the group. He also obtained some additional certificates in biological and mathematical fields. In 2006, the completion of the PhD project lead to publication and defense of the PhD thesis 'Analysis of microarray gene expression data sets' with Prof.dr. J.P.M. Geraedts as promotor. Being in the position to make new choices with the finishing of his PhD, it is not yet clear what his further trajectory for the coming years will be.