A Two-stage Normalization Method For Robust Differential Expression Analysis in Microarray Experiments

By

Shirin Manafi B.Sc. in Medical Engineering, Azad University, Tehran, Iran, 2011

A Thesis Presented to the School of Graduate Studies at Ryerson University in partial fulfilment of the requirements for the degree of Master of Applied Science in the program of Mechanical and Industrial Engineering

Toronto, Ontario, Canada, August 2014 c

AUTHOR'S DECLARATION FOR ELECTRONIC SUBMISSION OF A THESIS

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I authorize Ryerson University to lend this thesis to other institutions or individuals for the purpose of scholarly research

I further authorize Ryerson University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my thesis may be made electronically available to the public.

Abstract

A Two-stage Normalization Method for Robust Differential Expression Analysis in Microarray Experiments

Master of Applied Science

Mechanical and Industrial Engineering

Ryerson University

Abstract - In this research, we introduce an approach to improve the reliability of genetic data analysis. Consistency of the results obtained from microarray data analysis strongly relies on elimination of non-biological variations during data normalization process. Instability in Housekeeping Gene (HKG) expression after performing common normalization methods might be an indication of inefficiency potentially resulting in sampling bias in differential expression analysis. This research aims to reduce the sampling bias in microarray experiments proposing a two-stage normalization algorithm. Proposed approach consists of non-linear Quantile normalization at the first stage and linear HKG based normalization at the second stage. We tested the efficiency of the two-stage normalization method using publicly available microarray datasets obtained from the experiments mainly in the field of reproductive biology. Results show that combined Robust Multiarray Average (RMA) and HKG normalization method reduces the sampling bias in experiments when variations in HKG expression is observed after RMA normalization.

Acknowledgement

I would like to express my sincere gratitude to my supervisor, Dr. Ayse Bener for her continuous encouragement and support throughout my research. She was always motivating and generous with her invaluable time. It was a great privilege to work with her.

I would also like to acknowledge the Industrial Engineering department and the School of Graduate Studies at Ryerson University for their support in terms of financial aid, and work experience as a teaching and graduate assistant.

Thanks are also to my colleagues in the Data Science Lab and medical data Research Group. I am honored to be a part of this group where a team spirit truly prevails. I would especially like to thank Dr. Asli Uyar for her great help.

My special thanks go to my parents for their great support and encouragement during my Master's studies, and to my brother, Shahriar, and his family who are always a great source of love and motivation.

Table of Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION & BACKGROUND

1.1 Introduction

Functions of genes and their products have always been considered the key to understand living organisms. Researchers have developed different methods to provide a complete picture of these genes and their variations in different tissues [1]. There are traditional methods such as Reverse Transcriptase Polymerase Chain Reaction (RT-PCR) that work on one gene at a time [2]. More recent microarray transcriptomics technology enabled scientists to inspect the gene expression of whole genome on a single chip so the modulation of thousands of genes could be pictures simultaneously [1]. Microarray technology enables genome-wide transcription profiling that provides a spectra of whole genome expression at once [3]. Microarrays are commonly used for identifying the genes of interest via differential expression analysis, which are over- or under-expressed in the investigated biological condition compared with another one, i.e. cancer vs. normal tissues [4].

In microarray technology, profiles of all genes' expression are examined together [5]. One of the main objectives of microarray experiments is to evaluate which genes are differentially expressed in different conditions [6]. Since microarray data analysis is a complex multi stage experimental procedure, different systematical and technical variations could be introduced into the data. To handle these non-biological variations, different normalization methods are proposed [7].

Progressive technology such as microarray experiments has enabled extensive mining of biological data in parallel. Microarray experimental design consists of specific chips that provide a medium for hybridization of gene sequences in order to be monitored, detect the mutations differentiate the genotypes. The whole genome of any organism such as human, mouse, bacteria, etc. could be analyzed at once with the aid of microarray technology. The microarray data analysis procedure could reveal the key features both in the process of any organism development and in terms of human health such as disease diagnosis, drug development, pathological sciences and terminal illnesses. In functional genomics, with all the above mentioned advantages, microarray technology has become a reliable and popular protocol [3] [8].

Sample size in microarray experiments is one of the critical factors affecting the results of differential expression analysis. Some of the available methods for sample size calculation are limited in a way that in order to be able to use them, the variation between samples should be known [9]; some other methods may be complicated, as they consider different errors in their

calculations [10]. Also, variations between different samples make the choice of samples for conducting a microarray experiment very important.

In a general sense one of the most important concepts in biology is reproduction. Simply, it means making a copy, a likeliness, and thereby facilitation of continuation of species. Reproduction has greater significance to living organisms than only in terms of the continuation of generation in animals and plants. The origin of life and the evolution of organisms is the main aspect of reproduction biology. At the beginning of time, there must have been some primitive ability of chemical systems to produce copies of themselves. Recently, transcriptomics technologies, have been used extensively in reproductive biology [11]. Transcriptomics is the study of the complete set of RNAs (transcriptome) encoded by the genome of a specific cell or organism at a specific time or under a specific set of conditions. Sampling of follicular cells is confronted by various challenges such as low amount of oocyte and cumulus/granulosa cell samples and potential contamination from adjacent cells. Collection of those samples is a time consuming process requiring specialized techniques. Therefore, choice of sample size in gene expression analysis of follicular cells is critical to provide a trade-off between the effort in sample collection and power of results [11].

The multi stage microarray experimental procedure is prone to undesired systematical and technical variations. This challenges the precision of differential expression analysis. Various normalization methods have been proposed to eliminate the non-biological variations in the microarray data [12]. Robust Multichip Average (RMA) normalization method is commonly utilized to eliminate the non-biological variations between samples. This method perform the preprocessing in three steps, background adjustment, quantile normalization, and summarization [4] [13]. The assumption behind quantile normalization is that the total expression should be the same for all samples in an experiment [14]. Microarray experiments in reproductive biology are not excepting from these challenges. Actually due to low quantity of follicular cells per condition in most of the microarray experiments in the field of reproductive biology, the risk of sampling bias is predicted to be higher.

In this research, we investigate the consistency of microarray differential expression analysis in association with RMA, sampling bias and stability of housekeeping genes. We propose a two-stage normalization method with an aim to reduce the sampling bias in microarray experiments.

1.2 Background

1.2.1 Basic Concepts of Reproductive Biology

Development of any living organism is due to interaction between plenty of molecules via indefinite chain of complex chemical reactions and constant exchange of matter and energy with surroundings. Proteins and nucleic acids are the main actors in biochemistry. Most endeavors in molecular biology studies are toward understanding the structure and functions of proteins and nucleic acids. Living organisms consist of structural proteins building the tissues and of enzymes to facilitate the chemical reactions. The unit of protein structure is nucleic acids. There are two kinds of nucleic acids in living organisms, ribonucleic acid (RNA) and deoxyribonucleic acid (DNA). The unit of nucleic acids are nucleotide molecules consisting of phosphate, sugar and base. RNA is a single stranded three dimensional structure formed from sequence of nucleotides. Different types of RNA have different functionalities in making the proteins. DNA is a double stranded helix consisting nucleotides as well. Chromosome is the long DNA double helix that have different coding parts. The coding parts consist of genes which represent as codes to form proteins. The whole set of chromosomes is called the genome of a living organism [4].

1.2.2 Microarray Data Analysis

Microarrays consist of transparent slides on which probes are positioned in specific order [15]. Each probe consists of ~25 base nucleotide corresponding to specific sequences on the genome. The arrangement of samples on microarray chip is pre-assigned, therefore the origin of the data obtained from any probe is known and can be further analyzed, accordingly. Therefore one can address the genes of interest precisely [16]. A typical microarray contains several thousands of addressable genes [17].

1.2.3 Experimental procedure

Microarray experiments start with sample collection, separation of RNA and reverse transcription. The next step is amplification of transcripts of interest which is done by Polymerase Chain Reaction (PCR) procedure [18]. In order to label both test and control samples, they are fluorescent dyed during the process of reverse transcription. The reverse transcription step is the process during which a double stranded DNA is made from single stranded RNA molecule by using enzymes. Based on the type of experiment one or two dyes are used. Then each sample is hybridized to a microarray chip. The hybridized samples are

illuminated by a laser light and as a result, a specific emission spectra from slide probes are pictured. To measure the emissions from probes, the chip is scanned by a laser microscope [19]. The measured emissions are transferred to an analytic software. A value is given to the amount of gene expression of each probe and therefore a large dataset is produced [20]. The dataset is in form a large matrix with each row consisting of each gene expression values and each column related to one of the examined samples; additionally the condition of each sample is known [1] [21].

Obtaining biological samples is a costly and time-consuming process so that using a large sample size may not be feasible. On the other hand, a very small sample size may result in weak inferences from the experiment. Therefore, there is a trade-off between precision and cost of experiment in order to find out the appropriate sample size [20].

1.2.4 Data Analysis

Once the appropriate number of samples is gathered, microarray data should be preprocessed using different methods. Exploratory data analysis needs to be conducted to check for data quality to identify missing values, redundancies, outliers, and noise as well as to understand the distribution, mean and spread of data points [19].

During the preprocessing of raw data, the intensity at the background must be corrected at each spot. The bias fluorescence from background could have plenty of sources, such as unwanted binding of samples to the chip surface, the deposits from previous washes or the scanner noise [22]. The background adjustment is performed by various algorithms in preprocessing stage [22], for example, in Robust Multi-array Analysis (RMA) algorithm, signal and noise distributions are convoluted [13]. After elimination of background noise, the data should be normalized. Normalization is performed to eliminate the technical variations among the samples in order to identify the desired biological variations [22]. Hybridization process does not happen equally for all spots. This process may differ in several aspects such as the initial quantity of hybridized RNA, the hybridization time or the sample volume [23]. These variations may lead to some scaling differences for fluorescent intensity levels of different spots. Also, the physical differences in arrays or between the scanners may affect the readings from probes [22]. Overall, normalization process ensures that at the end we will be analyzing the information gathered from equalized spots. According to reviewed studies, the utilized normalization method affect the differential expression analysis significantly, therefore choice of appropriate normalization method is critical [23] [24]. After the preprocessing, data shall be further analyzed

so that the genes of interest could be detected that is the genes that may have expressed differentially.

In microarray experiments, typically, expression levels are measured for plenty of genes at the same time [25]. Since the number of genes being investigated are far more than the number of samples, multiplicity complications could be generated [25] [26]. As for any multiple testing problem there are various solutions to forfeit this issue [22]. Benjamini and Hochberg [27] proposed False Discovery Rate (FDR) as a solution and the Bonferroni Method has been developed to control the family-wise error rate [22] [28].

1.2.5 Sampling Bias in Microarray Experiments

Before a microarray experiment is conducted scientists need to decide on the number of replicates that they will use in their experiment. It is necessary to gather the right number of replicates to ensure the power of experimental results in identifying differentially expressed genes [29]. The objectives of the study, available resources and the technology reliability as defined by chip accuracy and hybridization failure rate, are the factors that determine the sample size [30].

Determination of sample size, considering the intrinsic complication of microarray experiments, and involvement of large amounts of data, is a significant issue [10]. However, researchers mostly use rules of thumb for sample size instead of proved formulas on the basis of experiment's objective [10]. These rules of thumb are based on some common assumptions, as per the following [31] [32];

If the objective of the study is to locate huge differences (greater than 2-fold) between conditions, the assumption is that as for each condition, three samples is enough. As in other cases, to identify smaller differences, five samples per condition is needed to get consistent results about biological differences between samples. The integrity of results is sustained more if six control sample and six treated samples are gathered. This approach would result in increased accuracy in estimation of p-values and FDRs. In experiments that four or more conditions are being examined and the conditions are drastically different, it is assumed that with about four samples per condition, reasonable understanding of biological variations between samples can be obtained [30] [33].

As in any statistical analysis, microarray data analysis suffers from common challenges which includes choosing relevant test statistics, sample size determination, outlier identification and accordingly significance of results [22] [34]. In order to detect differentially expressed

genes, fold change criteria could be an appropriate approach. For different experiments, different cutoffs are selected, but mostly 2-fold change is thought to be a reliable threshold [22] [35]. On the other hand, the cutoff selection is related to the genes that show large fold changes and thus the genes with smaller variation between samples may be overlooked, despite the fact that they are as important. Considering these cautions about cutoff values, the significance of differential expression analysis cannot be assessed reliably [36] [37].

This approach for determining sample sizes in microarray experiments, could induce inconsistency in results threatening the validity of clinical inferences. Though there are plenty of theoretical methods to calculate the quantity of samples needed in an aforementioned experiment, these methods are unemployed in all the studies that are being performed in this area [10] [38].

1.2.6 Sampling Bias in Reproductive Biology Microarray Experiments

Microarray data analysis is one of the recent developed tools for expression profiling of thousands of genes. This tool has aided reproductive biologists to have more accurate spectra of gene expression during key developmental events. Microarray experiments provide biologists with lists of tens of thousands of genes which have been modulated, mutated or regulated during different developmental stages. The development of technological tools such as microarray data analysis, has made an evolution in reproductive biology studies from hypothetical state to more descriptive discoveries [39].

Success of any expression profiling experiment depend on different factors such as quality of samples, precision of experimental design, accurate choice of statistical strategies for data analysis and etc. [7] . In the field of reproductive biology, additional factors contributes to the risk of microarray results being inconsistent. Among these additional factors are the nature of the objective of study, the intrinsic complexity of developmental key events, the complexity of the tissue of interest and difficult interpretation of the biological patterns [39]. All these challenges are introduced to microarray experiments in the field of reproductive biology. Therefore sampling bias in reproductive biology experiments may lead to inaccurate clinical inferences in a more deep sense [39].

1.2.7 Two-Stage Normalization Method

The entire multi-stage microarray experiment is prone to errors and non-biological variations. During hybridization stages, variations from different amounts of RNA, different hybridization

times or the volume of samples, could be introduced into the data. Also physical differences between either arrays or scanners might be source of variation. Reducing these unwanted variation as much as possible would lead to more accurate and reliable expression analysis results [40]. One of the major steps in preprocessing of microarray data analysis and specifically in reduction of the variations, is normalization of raw data. The main goal of any normalization method is to eliminate the variations which have been introduced into data from different sources. Different methods and algorithms are being used for normalization of microarray data. We should choose the method that performs the best in evaluating different normalization method for any kind of microarray analysis [24].

Robust Multi-Array Analysis (RMA) algorithm, is one of the most popular preprocessing methods [13]. RMA algorithm relies on Quantile normalization in order to remove the technical variations. Quantile method makes the distribution of all probes intensities on a set of arrays, identical [41]. This method first performs the background adjustment for the arrays [22], then the perfect match probes are background corrected. Then the intensity values are transformed to log 2 values and at the end they are normalized by quantile algorithm and finally summarized [12].

It is a biological assumption that a treatment would result in up or down regulation of a number of genes and the other genes' expression would remain stable. On the other hand, there is supposed to be equal amounts of ribonucleic acid (RNA) on each array; therefore the sum of all expressions among the samples of one condition should be the same. Since RNA amount is not fully under control and these assumptions might not hold, thus to measure the actual expression values, use of a control would be helpful [42].

1.2.8 Housekeeping gene normalization

Another popular normalization method is normalization against internal controls or housekeeping genes. These genes are the ones with hypothetically stable and consistent expression values in all cells. Also in some experiments normalization is performed against external controls, which is based on using genes from other organisms [42]. Endogenous genes, also known as housekeeping or control genes are those that are responsible in basic cellular functions in all tissues and organisms [43]. It is assumed that in all cells, the housekeeping genes are expressed uniformly regardless offstage of differentiation, type of tissue or stage of development [42] [43] [44].Therefore these genes could be considered as perfect candidates to be used as experimental controls and to fulfill computational objectives [43] [45] [46] [47] [48]

[49]. For any organism, common housekeeping genes can be identified with the aid of advanced transcriptomic technologies [50]. Large-scale expression data profiling is used to examine the expression values of all genes and identify the housekeeping [50]. Different statistical algorithms such as geNorm and Normfinder are tools to identify housekeeping genes [51]. In geNorm algorithm a pair of housekeeping genes are chosen based on the M-value of multiple candidate genes. M-value is an estimation of pairwise variation for each gene [52]. Normfinder also ranks candidate housekeeping genes based on their expression stability [53].

Another issue in normalization against housekeeping genes is considering single gene as an internal reference versus using multiple housekeeping genes. Vandesompele et al. [52] proposed an allegedly more accurate normalization algorithm that is to select a set of genes with minimum variation between the samples and by calculation the geometric mean of them, perform the normalization against this mean. The overall algorithm of normalization against housekeeping genes is to first identify the most stable genes and determine a normalization factor per array by calculating the geometric mean of those stable genes' expressions and then divide all the expression values of the array by the normalization factor. The number of genes to use for calculation of geometric mean is dependent on the practical considerations [54].

1.2.9 Sampling bias

The expressions of housekeeping genes are expected to be stable after RMA normalization. However, the fundamental assumption of RMA approach may not hold in case of transcriptome-wide up-regulation or down-regulation of gene expression, or when there are differences in the concentration or purity of the samples. Instability in the HKG expression after RMA may indicate that kind of variation and may result in sampling bias in the differential expression analysis [55].

Chapter 2
LITERATURE REVIEW

2.1 Literature Review

2.1.1 Sample size calculation in microarray experiments

Microarray experiments have been performed for more than a decade now. There are plenty of studies regarding the design of the experiment. One of the major issues in the design of these experiments has been the number of replicates required to achieve power and consistency in the results. Different methods have been suggested to calculate the number of required samples. Some of these methods are based on traditional significance and power estimations as in any statistical analysis [56] [57] [58] [59] [60] [61].

Dobbin and Simon presented a straight formula for calculation of number of required replicates [10]. Based on their formula, in order to calculate the number of replicates, the intra-sample variation must be known [62]. They also have mentioned that the variation could be estimated from previously performed similar experiments. This approach raises some questions since the similarity of experiments is by itself an issue to be considered. On the other hand, as the authors mentioned in their article that formula is not suitable for small sample sizes the variation between samples could be estimated or the estimation would not be reliable due to small number of samples. Formula (1) below is the formula for sample size calculation derived by Dobbin and Simon [10]. As mentioned above, in this formula, the variation is needed to be known.

$$n/m = 4\left[\frac{z_{\alpha/2}+z_{\beta}}{\delta}\right]^2\left(\tau_g^2+\frac{\alpha_g^2}{m}\right) \quad (1)$$

In this formula, the required technical replicates of each sample is indicated by *m*, and the whole number of microarray by *n*. The 100α/2th and 100βth percentiles of the normal distribution are depicted by $z_{\alpha/2}$ and z_{β} and the distance of group means is shown as δ. The study's objective is to determine this distance. The variance between measurements for one gene among all samples of one group is annotated by$\left(\tau_g^2+\frac{\alpha_g^2}{m}\right)$. Since this variation must be known for each gene of interest, plenty of estimations should be made from previous experiments [10].

Another approach for sample size calculation has been suggested by Pawitan et al. It is only based on FDR control [56]. They have argued that the latter approach works better since FDR is a better scale than P-value [56]. Other calculations for sample size such as Lee and Whitmore proposed in 2002 is based on absolute false positives [31] [56]. This scale though better than

traditional scales, it cannot directly control the FDR [56]. In 2004, Müller et al. suggested a theoretical approach for the optimal sample size in a way to achieve a certain FDR by the maximum number of differentially expressed genes [56].

In other methods, Lin and Hsueh proposed a formula to calculate the sample size with regards to the number of arrays needed to reach a definite sensitivity with 95% significance level [63] [64]. This approach could be argued in the way 95% sensitivity is usually unachievable and the probability of detection is less than 50% [65]

In addition to theoretical methods, some software packages such as R (R package) and dchip are available for sample size calculation [34]. Determination of sample size in these software is also based on the known variation between samples [21].

Although there are plenty of methods for determination of sample size in microarray studies, practically only in few experiments these methods have been utilized. In field (clinical) experiments, there are not many researchers and biologists who calculate their sample size with the established calculation methods. This may be due to impracticality of small number of available biological replicates, unknown variations between samples, etc. In practice usually researchers use the available samples for their experiments. Normally the sample size calculated by the established formulas is more than 25 samples and this many might have not been gathered in the sample collection step [10]. On the other hand, for most sample size calculations, the variation should be known and it is usually estimated from previous similar experiments. This would not be a reliable approach since in most cases similarity between experiments is an issue of relativity. Considering this random use of already acquired samples, it may be a source of variation in the microarray experiment resulting in inconsistency in the experiment results.

2.2 Normalization of microarray data

This research main objective is to tackle the inconsistency in microarray data analysis results. This inconsistency is generally handled in preprocessing stage in microarray data analysis and specifically in normalization step. Different normalization methods have been introduced for microarray data.

Having seen the existing sampling bias in microarray data analysis, we targeted normalization, the main step in microarray experiments that is responsible for reducing the sampling bias. The main objective of normalization is to eliminate non-biological variation in microarray experiments that compromise the expression level measurements [24].

Normalization algorithms aim to adjust the samples for their non-biological differences. There are different sources for these differences such as amounts of sample and laser apparatus settings [8] [24]. Normalization is not the only step in microarray data analysis, responsible for removing systematic variations, however normalization is the most important step to eliminate non-biological variations in the earlier stage [66].

To choose the right normalization method for microarray experiment, certain decisions should be made [48].

These methods basically are categorized based on the level they are performed at:

I. Probe level

The microarray samples are hybridized on the probes of the microarray chip. If the normalization is performed on the initial values corresponding to the amount of emitted fluorescent dye from each of the probes, it is called normalization at probe level.

II. Expression level

After gathering the initial emission measures, they are summarized and the expression values for genes are calculated. If normalization happens at this level on the calculated expression values, it is called normalization at expression level.

Also it should be determined which genes to be utilized for normalization [24]:

I. All genes [24]
II. Housekeeping genes [24]
III. Different controls [24]
IV. Rank invariant genes [49]

It depends on the data category, the image analysis software, etc. to select the most appropriate method that provides the best results [67]. Below are these methods:

Scale normalization:

As the simplest approach, normalization of data is scaling, e.g. force the median of sample differences to be 0 [67].

Lowess:

This normalization method is mostly performed in two-color microarray experiments [67].

Quantile:

In this approach, not only the medians (50% quantile) is adjusted but also all the quantiles will be uniformed.

VSN:

This transformation is similar to using the natural log transformation

There are reviews on microarray normalization methods provided by Quackenbush [68] and Bilban et al. [69]. Some extensions are suggested in order to perform global and intensity-dependent normalizations [24]. As an instance, Kepler et al. [70] proposed to make the regression locally in order to estimate the normalized intensities. In another study, Wang et al. [71] proposed to perform normalizations with multiple iterations in order to both determine the normalized valued and internal controls [24].

Workman et al. [72] suggested a non-linear model to normalize microarray data based on distribution analysis [24]. Chen et al. [73] proposed to perform normalization both locally and globally [24]. In order to make corrections for spatial heterogeneity, Edwards [74] proposed non-linear LOWESS normalization.

Chapter 3

PROBLEM STATEMENT

3.1 Problem Statement and Proposed Solution

As in any statistical analysis, microarray data analysis is prone to different challenges such as sample size, outliers and statistical significance. Determination of sample size in microarray data analysis is very significant, since microarray experiments have intrinsic complexity and include large amount of data. On one hand, in microarray experiments, the process of sample gathering is expensive and involves high complications, on the other hand small sample sizes might lead to inaccurate results. Small sample sizes increase the risk of sampling bias, since it is more likely to have outliers and/or low-quality samples when there are few samples.

When it comes to reproductive biology, the sampling bias problem becomes even more critical. The main reproductive follicular cells that are examined in microarray experiments are oocytes, granulosa, cumulus and endometrium cells [75]. The mentioned cells tend to get contaminated by nearby cells and also collection of reproductive follicular cells is very timely and requires complicated techniques [76]. Thus, sample size determination is important in reproductive microarray experiments both from power of results point of view and complexity of sample gathering [77].

There are some assumptions for choosing sample size in microarray data analysis [21]. Based on the goals of the microarray experiment, such as to find large differences between two conditions, 3 samples per condition is assumed to be enough; or to find small differences, 5 samples per condition would be adequate; or if the conditions are not physiologically, drastically different, then 4 samples per condition is assumed to be adequate. If these ‘rules of thumb’ are used to determine sample size in microarray experiments, then possible inconsistency in the results would endanger the validity of corresponding clinical inferences.

3.2 Sampling bias in microarray data analysis

The initial goal of this research, has been to experimentally show that increasing number of replicates and using different combinations of available samples may influence microarray results. We provide different examples in which using both different number of replicates and different subsets of available samples, shows non-negligible oscillations in microarray results, which the number of differentially expressed genes. These examples are proofs of inconsistency

in microarray results and due to our approach, proofs of sampling bias in microarray data analysis.

3.3 Tackling the sampling bias

To tackle sampling bias in microarray data analysis, we targeted the preprocessing of raw data in microarray data analysis. Raw microarray data is challenged by undesired non-biological variations. Non-biological variations such as technical, systematical and human errors are the main sources of inconsistency in microarray results. In preprocessing step in microarray data analysis, the undesired variations are supposed to be removed at normalization stage. Different normalization methods is proposed for elimination of non-biological variations.

3.4 Common available normalization methods

One of the most popular and robust normalization methods in microarray data analysis is Robust Multichip Average (RMA) [9]. We evaluated the sampling bias in microarray differential expression analysis, in which we normalized the raw microarray data with RMA algorithm. Our results show that though the data were normalized with RMA, the differential expression analysis results (number of differentially expressed genes) under different conditions of sample choice, showed oscillation. It is expected that using different number of samples and different combinations of them (subsets), the number of differentially expressed genes remains the same or at least on the same range; though having seen the oscillation in differential expression analysis is a proof of inconsistency.

Also to investigate the efficiency of RMA algorithm in elimination of non-biological variations in microarray raw data, we examined the variation in housekeeping genes' expression. Housekeeping genes are the ones which are related to basic cell functions and are supposed to have stable expressions in all samples and conditions. The results of this investigation also confirmed that in plenty of microarray experiments sampling bias exists.

3.5 Proposing a two-stage normalization method to tackle sampling bias

We propose a two-stage normalization method to eliminate the non-biological variations and consequently reduce the sampling bias in microarray data analysis. The first stage normalization is using RMA algorithm at probe level and after investigation of variation of housekeeping genes' expression after RMA, the second stage is to normalize the microarray data using housekeeping genes normalization algorithm at expression level.

Chapter 4
METHODOLOGY

4.1 Methodology

In order to implement our proposed solution, a precise procedure should have been followed. The overall flowchart of our methodology is given in Figure 1. The procedure is given in Algorithm 1.

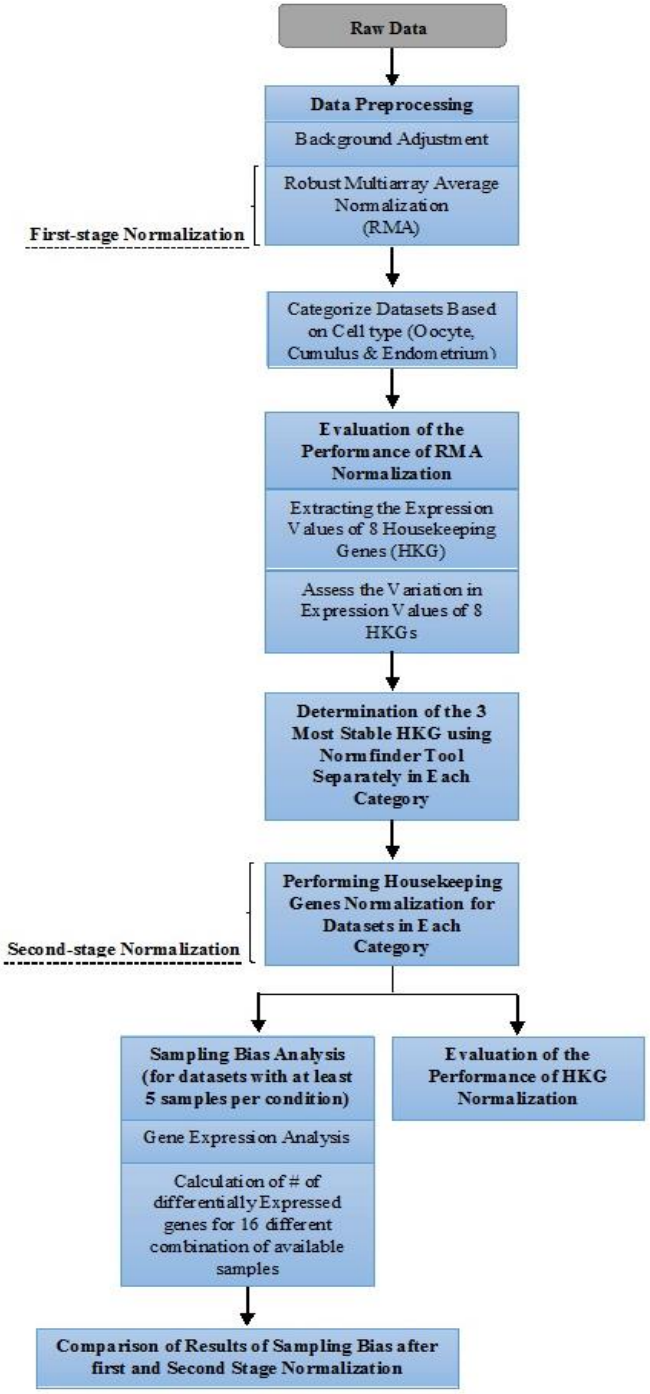

Figure 1 - Overall Flowchart of performed procedure

ALGORITHM 1

```
BEGIN
 INPUT raw data
  IF raw data == "Human Genome" THEN
       "Configure the corresponding platform to Human Genome"
  ELSE
   OUTPUT "Is it mouse genome?"
       "Configure the corresponding platform to Mouse Genome"
  ELSE IF
   OUTPUT "Is it cattle genome?"
       "Configure the corresponding platform to Cattle Genome"
  END IF
 END INPUT


 BEGIN PREPROCESSING
  FOR each sample
       "Perform RMA normalization"  %%First-Stage  Normalization%%
  END FOR
 END PREPROCESSING


 BEGIN SAMPLING BIAS ANALYSIS
  FOR each experiment
    IF sample == "Normal" THEN
      "Put sample in 'Normal' group"
    ELSE
      "Put sample in 'Treated' group"
    END IF
    IF count of samples per group => "5"
```

```
    FOR 16 different combinations of samples in 'Treated' group
     PERFORM Differential Expression Analysis
    END FOR
   END IF
END SAMPLING BIAS ANALYSIS


BEGIN ASSESSMENT OF VARIATION
 FOR each Housekeeping Gene
     "Extract the expression value in all samples per group"
 END FOR
 FOR all HKG in each genome
     "Use 'NormFinder' tool and Rank HKGs based on their variation"
 END FOR
END ASSESSMENT OF VARIATION


BEGIN SECOND-STAGE NORMALIZATION
 FOR each 'organism' AND 'cell type'
  "Choose the 3 most stable HKGs"
 FOR each sample
  "Calculate the 'geometric mean' of 3 most stable HKGs"
                                    %%Normalization Factor%%
  "Divide all the expression values by the 'Normalization Factor'"
                                   %%Second-stage Normalization%%
 END FOR
 END FOR
END SECOND-STAGE NORMALIZATION
```

```
 REPEAT SAMPLING BIAS ANALYSIS


 BEGIN EVALUATION OF THE EFFICIENCY OF PROPOSED SOLUTION
 FOR each experiment
   "Compare the results of sampling bias analysis"
 END FOR
 END EVALUATION
END
```

Each of these steps are explained thoroughly in the remainder of this chapter. At the end of chapter the performance measures of our research and the proposed solution is given.

4.2 Data gathering

4.2.1 Public Microarray Data Repositories

The first step in any data analysis project is to gather related data precisely. In the field of microarray studies, various public data repositories are available online. These repositories are the archive of data from microarray samples which have been used in different experiments.

In order to simplify, we imported our data from ArrayExpress database [78]. ArrayExpress is the database of European Bioinformatics Institute (EBI) which includes data from genomic experiments that can be searched and downloaded. In this database there are genome expression profiles from microarray sequencing studies. EBI is regarded as a section of European Molecular Biology Laboratory (EMBL) [78].

EMBL-EBI consists of publicly available data from scientific experiments. It also facilitate basic computational biology research and provides training programs, both for academic and industry researchers [78].

Since the datasets in ArrayExpress includes the ones both from microarray and sequencing experiments, we had to separate microarray datasets. ArrayExpress consists of the data from 41639 microarray experiments. This quantity is resulted when we sort ArrayExpress experiments based on the technology used, which was Array Essay. This technology is the underlying technology for microarray experiments. Each of these experiments had been performed in different organisms such as human, cattle, mouse, bacteria and etc. In each organism category there are experiments related to different tissue types such as reproductive tissue, muscular tissue, blood and etc.

4.2.2 Selection of experiments from different organisms and different tissue types

Experiments in ArrayExpress database are sorted based on technology of experiment, platform, organism, tissue types and the date of publishing. The overall procedure of our solution is the same for all microarray experiments, however since the second stage of our solution is based on housekeeping genes, we had to be very careful to implement our solution each time on a group of experiments that have been done on the same organism and tissue type. Considering

these cautions, our selection of data (experiments) in terms of organism and tissue type is as listed below:

- Organism: Human (Homo sapiens)
    - Tissue types: Oocyte
      Cumulus
      Endometrium
      Peripheral Blood
      Lymph
- Organism: Cattle (Bos Tauras)
    - Tissue type: Oocyte
- Organism: Mouse (Mus Musulus)
    - Tissue type: Oocyte

For each of these categories, we imported between 5 and 8 datasets. In our study we selected the experiments with only biological samples in different conditions. We didn't consider technical samples. Each biological samples is gathered from only one source (human, mouse, cattle); on the other hand, technical samples are several samples gathered from one source. In the process of downloading data, based on samples description, we made sure about the tissue type. Also we studied the description of each original experiment to find out how the samples were gathered and determine the control (normal) samples versus treated or mutated ones. A sample profile of ArrayExpress experiments is shown in Figure 2.

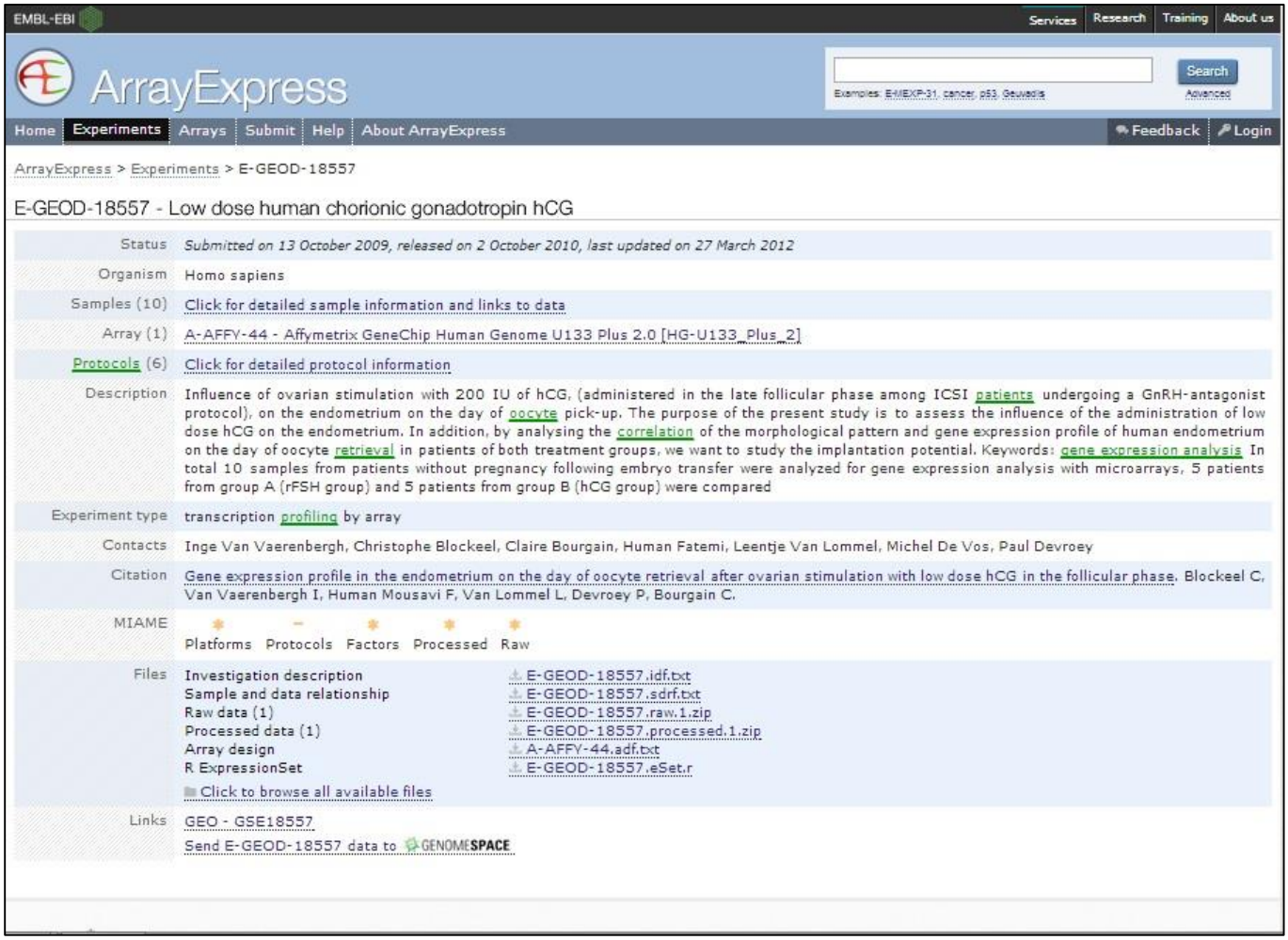

Figure 2 – Sample profile of ArrayExpress experiment

4.3 Platform Configuration

Any microarray experiment has been performed on a specific platform. The platform is the design of the microarray chip that the experimental samples were hybridized on. Anyone who wishes to perform a microarray analysis would buy these chip platforms to hybridize the samples that they have extracted from organisms on them. We need to figure out the platform in order to normalize the data and we need to have the platform description. We also need to know exactly which probe on the chip corresponds to any specific gene in order to determine expression value of each housekeeping gene in each sample. Any platform file is like a map that is compatible with downloaded files from samples' data. The detailed description of this map is provided in any platform's annotation file. For each platform that we used, we downloaded the

corresponding annotation file that was not easily accessible in all cases. Since we have implemented our solution on different organisms, the platforms were different for each organism. Though even for each organism, there are different platforms, to narrow down our research, for each organism we chose the experiments that their platform were the same. The platforms we used for each organism are as follows:

- A-AFFY-44 - Affymetrix GeneChip Human Genome U133 Plus 2.0 [HG-U133_Plus_2]
- A-AFFY-45 - Affymetrix GeneChip Mouse Genome 430 2.0 [Mouse430_2]
- A-AFFY-128 - Affymetrix GeneChip Bovine Genome Array Bovine

4.4 Performing the first stage normalization, Robust Multichip Average (RMA)

At first stage, the arrays of chosen datasets were normalized by Robust Multichip Average algorithm [9]. The intensity values are transformed to log 2 values and at the end they are normalized by quantile algorithm and finally summarized. Quantile algorithm which is based on the idea that plotting of the quantiles of probe intensities gives a straight line along a unit vector $(1/\sqrt{n}\ldots 1/\sqrt{n})$. As a result, if in a quantile plot we project the n-dimensional data points on a straight diagonal line, then the all the datasets will have identical distributions [12]. The quantile normalization for a set of data vectors can be performed as shown in algorithm 2.

ALGORITHM 2

```
BEGIN
 INPUT N datasets of length p
 FORM Matrix X
       Dim(X, 1) == p
       Dim(X, 2) == N
         "Put each dataset as a column"

 X_SORT == SORT Matrix X
 FOR each row of X_SORT
         "Calculate the average"
 END FOR
 FOR each element of the row
         "Substitute the element by the average"
 X'_SORT == Substituted Matrix from X_SORT
 END FOR

 FOR each column of X'_SORT
         "Rearrange the column to have the same order of original X"
 END FOR

 X_NORM == "Rearranged X'_SORT"
END
```

It is to be noted that in order to perform RMA normalization, each sample's data file should be uploaded with its corresponding platform file. Then the three steps of RMA is performed as mentioned above.

4.5 Statistical analysis with regards to sampling bias, after first-stage normalization

The final step in any microarray data analysis is differential expression analysis. After preprocessing, data should be further analyzed so that the genes of interest could be detected. This step is critical statistical analysis in any microarray data analysis.

After performing RMA normalization in previous step, the expression values resulted from RMA normalization were used to find out the number of differentially expressed genes in each experiment. The test used for evaluating the number of genes expressed was t-test with P-value of 0.05.

4.5.1 Evaluation of statistical significance of differentially expressed genes

The main purpose of microarray experiments is to detect the genes with different expression levels between two sample groups. It is demanded to understand which genes are up-regulated (increased in expression) or down-regulated (decreased in expression) between two sample groups [77]. There are multiple samples in each group. The average of expression levels between the samples in each group is calculated. Therefore, in order to assess the difference of expression levels, the target is to make comparison between the means of the sample groups. In the experiments that there are two sample groups to be examined, the case would represent some kind of a t-test [77]. As one of the popular hypothesis tests, the two-sample t-test is applied to investigate whether the distance between mean values of two groups is significant or random. In our analysis, Genes with corrected false discovery rate of p-value less than 0.05 and |Fold Change| > 2 were considered as differentially expressed [70]. The traditionally accepted P-value for something to be significant is $P < 0.05$. So if there is less than a 5% chance that the gene expression values from two samples came from the same group, then it is considered a significant difference between the two expression values.

In differential expression analysis, first we took the average of the log expression levels in treated group and subtracted the average of log expression levels in control group from it to reach an expression log ratio for each gene [71]. Then we needed to determine which of these expression ratios is different from 1, significantly. Our approach was to implement the two-sample t-test to evaluate the significance of differential expression analysis [72]. Two-sample t-test is suitable for comparing between two groups from different experimental treatments or populations. The necessary conditions for applying two-sample t-test are as follows [62]:

- Samples in each group are gathered from distinct populations
- The responses from samples in each group are independent from each other
- The distribution of samples in each group are normal

In our research the above mentioned conditions are fulfilled. In all our experiments two distinct sample groups are compared; also the comparison variables are independent from each other since we are comparing the gene expression values in two completely different cell conditions. As for the third condition, all our raw data from our samples are normalized with RMA normalization algorithm at the first stage; the last step of RMA normalization is to make the distribution of samples similar and normal [63]. Another issue in utilization of two-sample t-test, is the variation between different datasets. In order to forfeit this issue, in all of our experiments, we selected five samples per conditions to conduct our analysis. Even if there were more than five samples per conditions, we selected five of them.

Actually with this approach we are performing thousands of t-test together; some of them may affirm to be significant, even though they are actually not. We need to adjust for the simultaneously performed tests, so we used a correction, as Holm-Bonferroni proposed. As they suggested, we sorted the p-values from smallest to largest and multiplied the smallest p-value by K (total number of genes). Then we multiplied the next smallest p-value by (K-1), the next smallest by (K-2) and so forth [71].

4.5.2 Sampling bias in microarray data analysis

As mentioned in previous section, determination of sample size in presence of experimental variations may lead to inconsistent results. Though there are plenty of theoretical methods to calculate the number of required samples for an experiment, these methods are not used in all the studies that are being performed in this area. In reproductive biology microarray experiments, sample collection is more complicated and contamination of samples is much more likely; therefore in reproductive cells-related microarray experiments, low quantity of samples are witnessed more. We investigated this matter at the beginning of our analysis and the low quantity of samples per condition in reproductive tissue microarray experiments was witnessed (Chapter 5, section 5.1).

The initial goal of this research project is to provide examples of how increasing the number of replicates and also using different subsets of available samples can affect the microarray experiments. We provide examples of how using different combinations of replicates in a microarray analysis could result in different outcomes. This is exactly where the inconsistency originates and hence, we recommend that these existing variations should be considered before conducting the study.

In this research in order to consider sampling bias in differential expression analysis, the following approach was followed. For each experiment which originally had at least 5 samples per condition, analysis was performed with different number of replicates (3 to 5) and in each condition, all possible combinations of replicates that is sixteen, were considered and the analysis was repeated.

When there are differences in the concentration or purity of the samples, the expression of housekeeping genes (HKG) might show considerable variations, even after RMA normalization. Instability in the HKG expression after RMA may indicate that kind of variation and may result in sampling bias in the differential expression analysis [32].

4.6 Assessment of the variation of housekeeping genes expression values

The most popular internal controls in microarray data analysis are so-called housekeeping or endogenous genes. Housekeeping genes are the ones with hypothetically stable and consistent expression values in all cells. Also in some experiments normalization is performed against external controls, which is based on using genes from other organisms. Large-scale expression data profiling is used to examine the expression values of all genes and to identify the housekeeping genes [66].

We evaluated the possible variation – instability – of common housekeeping genes in each organism after RMA normalization. We chose these housekeeping genes based on available literature [79] [80]. The list of housekeeping genes used for each organism in order to evaluate the expression variation, is given in the tables below. We performed the evaluation of possible instability of HKG separately for the tissues of interest. We calculated the mean and standard deviation of each housekeeping gene in each dataset for each tissue type.

Table 1 - Most stable HKGs in Human Genome

HUMAN ORGANISM

#	Symbol	Description
1	ACTB	Actin, beta
2	GAPDH	Glyceraldehyde-3-phosphate dehydrogenase
3	HPRT1	Hypoxanthine guanine phosphoribosyl transferase 1
4	GUSB	Glucuronidase, beta, b
5	SDHA	Succinate dehydrogenase
6	TBP	TATA box binding protein
7	YWHAZ	Tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, zeta polypeptide
8	B2M	Beta-2-microglobulin

Table 2 - Most stable HKGs in Mouse Genome

MOUSE ORGANISM

#	Symbol	Description
1	ACTB	Actin, beta
2	GAPDH	Glyceraldehyde-3-phosphate dehydrogenase
3	GUSB	Glucuronidase, beta, b
4	SDHA	Succinate dehydrogenase
5	B2M	Beta-2-microglobulin
6	H2AFZ	H2A histone family, member Z
7	HPRT1	Hypoxanthine guanine phosphoribosyl transferase 1
8	EEF1E1	Eukaryotic translation elongation factor 1 epsilon 1
9	PPIA	Peptidylprolyl isomerase A

Table 3 - Most stable HKGs in Cattle Genome

CATTLE ORGANISM

#	Symbol	Description
1	ACTB	Actin, beta
2	GAPDH	Glyceraldehyde-3-phosphate dehydrogenase
3	HPRT1	Hypoxanthine guanine phosphoribosyl transferase 1
4	GUSB	Glucuronidase, beta, b
5	YWHAZ	Tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, zeta polypeptide
6	B2M	Beta-2-microglobulin
7	HMBS	Hydroxymethylbilane synthase
8	UBC	Ubiquitin C
9	PKG1	Phosphoglycerate kinase 1

4.7 Determination of most stable housekeeping genes in each tissue type

We determined the three most stable genes by using the Normfinder tool, developed by Andersen et al. [53]. It is an Excel add-on application to determine the most stable genes from a predefined list. The analysis of variance is performed with a model-oriented comparison of expression values. Then a stability measure is defined and the deviation of each HKG from this stability measure is evaluated. These deviations are then used to calculate the stability value for each gene. It means that the lowest this deviation, the most stable the gene is. In Normfinder, ranking of most stable genes is reported as the gene with the lowest deviation value, indicating the most stable [53]. Schematic results of Normfinder tool is shown in Figure 3.

We used Normfinder on each dataset to extract the stability values. We ranked eight HKGs with respect to the stability values of each of the eight HKG that were extracted from Normfinder results. For each dataset, the eight HKGs were ranked according to their Normfinder

stability values; then in each category the ranks for each HKG was averaged to find the ranks for each cell type.

Gene name	**Stability value**			**Best gene**	SDHA
ACTB	0.015			**Stability value**	0.006
GAPDH	0.016				
HPRT1	0.018			**Best combination of two genes**	SDHA and TBP
GUSB	0.038			**Stability value for best combination of two genes**	0.007
SDHA	0.006				
TBP	0.013				
YWHAZ	0.016				
B2M	0.081				
Intragroup variation					
Group identifier	1	2			
ACTB	0.001	0.002			
GAPDH	0.001	0.002			
HPRT1	0.001	0.004			
GUSB	0.010	0.005			
SDHA	0.000	0.000			
TBP	0.000	0.002			
YWHAZ	0.002	0.001			
B2M	0.028	0.038			
Intergroup variation					
Group identifier	1	2			
ACTB	-0.005	0.005			
GAPDH	-0.008	0.008			
HPRT1	0.017	-0.017			
GUSB	-0.007	0.007			
SDHA	-0.014	0.014			
TBP	-0.009	0.009			
YWHAZ	-0.012	0.012			
B2M	0.039	-0.039			

Figure 3 - Schematic results of Normfinder tool

4.8 Performing the second stage normalization using most stable housekeeping genes

The overall algorithm of normalization against housekeeping genes is to first identify the most stable genes and determine a normalization factor per array by calculating the geometric mean of those stable genes' expressions and then divide all the expression values of the array by the normalization factor. The number of genes to use for calculation of geometric mean is dependent on the practical considerations.

In second stage, second normalization method was applied at gene expression level. First a normalization factor was calculated for each sample by calculating the geometric mean of the expression of three most stable housekeeping genes – determined by Normfinder previously – and dividing all expression values of that sample by the normalization factor. In formula (2), NFj corresponds to the normalization factor for sample j and Eij is the expression of ith housekeeping gene in jth sample.

$$NF_j = \left(\prod_{i=1}^{n} E_{ij} \right)^{1/n} \qquad for\ i = 1, \dots, 3 \qquad (2)$$

4.9 Repeating Statistical analysis with regards to sampling bias, after second-stage normalization

In this research, the differential expression analysis was performed once after first stage (RMA), and once after second stage (RMA + HKG). Similar procedure as section 4.4 was performed after second-stage normalization done in previous step. The expression values resulted from RMA+HKG normalization were used to find out the number of differentially expressed genes in each experiment. Similarly, the test used for evaluating the number of genes expressed was t-test with P-value of 0.05.

4.10 Comparing the sampling bias results after one-stage and second-stage normalization

In an attempt to reduce the sampling bias in microarray experiments, we proposed a two-stage normalization of microarray expression data, taking into account the stability of housekeeping gene expression across the biological replicates. Proposed method includes non-linear RMA normalization at probe level expression values followed by linear HKG based normalization at gene expression level. At probe level in which the values are the reported amounts of emitted light, relative non-linear normalization should be performed. The systematic variations introduced into the microarray data are from different sources, therefore performing linear normalization at probe level is not capable of elimination of systematic variations. However after performing RMA, the expression values of all samples included in the microarray experiment, are normally distributed, therefore linear normalization is justified to be utilized at expression level.

In order to assess the efficiency of our proposed solution, we compared the results from sampling bias after first-stage normalization (section 4.4) and second-stage normalization (section 4.8).

4.11 Performance measures

Microarray data analysis is a multistage procedure and each stage of it has different performance measures. In the first place, data is gathered from biologic samples. This step was not under our control because it is performed in highly professional laboratories and by expert scientists. We used the raw data that they have gathered through years in different microarray experiments.

After importing the raw data, the essential preprocessing stage is applied. We proposed a two-stage normalization method in this stage. The suitable performance measure in normalization stage is the evaluation of variation after normalization [81]. To test the efficiency of normalization process, the evaluation of variation shall be performed on control genes. In our research, we used Normfinder tool to analyze the variation of housekeeping genes, as controls [53]. Normfinder is a reliable tool in analysis of variation in housekeeping genes' expression values. Normfinder both calculate the inter-sample and intra-sample variation of housekeeping genes and based on these measurements list the most stable genes [53].

After the preprocessing stage is completed, the differential expression analysis is performed. This is the statistical stage of microarray data analysis. As mentioned in chapter 4, methodology, in order to detect the differential expressed genes, we used two-sample t-test. Based on literature, t-test is one of the appropriate tools to detect the differences between two sample groups [82]. Additionally, as benchmarked in the protocol of microarray experiments through years, most of microarray data analysis are based on two-sample t-test in differential expression analysis stage [83]. Since in all of our experiments our differential expression analysis was between two distinct sample groups, t-test would have been the most efficient tool.

In our analysis, to declare that a gene is differentially expressed, we considered the absolute fold change of that gene between two sample groups to be bigger than twice. To assure the significance of this consideration, we applied a false discovery rate-corrected p value of less than 0.05. It is traditionally accepted that threshold of p-value to be significant is less than 5% [24].

Though the values used in differential expression analysis could be varied, we kept them constant in all our experiments in order to assess the efficiency and influence of normalization of raw data, solely.

Chapter 5

RESULTS & DISCUSSION

5.1 Results and Discussion

5.2 Sampling bias results

The first step of our methodology was to gather data from ArrayExpress database [78]. The chosen datasets for further analysis in this study are from experiments performed on human, cattle and mouse organism. Their raw data have been acquired, then preprocessed and analyzed.

In this research, the public data repository from European Molecular Biology Laboratory – European Bioinformatics Institute (EMBL-EBI) website was chosen [78]. At first, we wanted to figure out the distribution of number of replicates per condition in microarray experiments. The keyword "oocyte" was searched to narrow this analysis. The search query resulted in 220 experiments. Taking into account the experiments in which the conditions were well described, the distribution of number of replicates per condition for more than 180 experiments were determined. This figure represents a kind of exponential distribution and is highly skewed to the right. We can conclude that in real life microarray experiments, experts use few number of samples per conditions (3-5) to conduct their analysis.

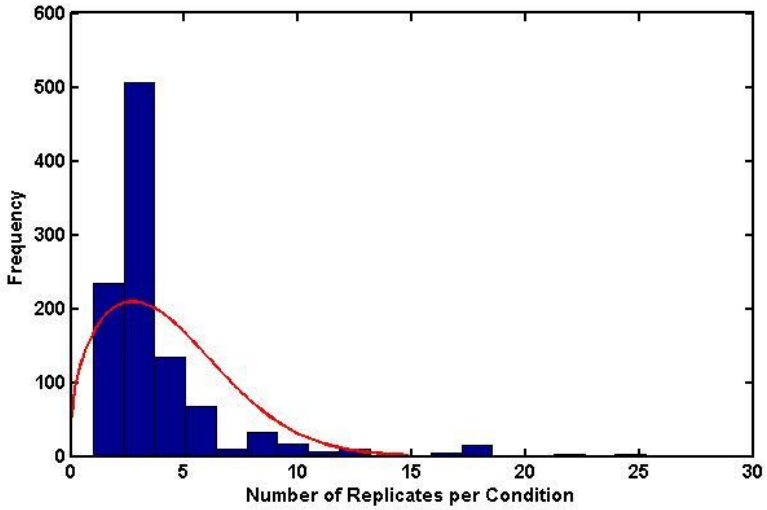

Figure 4 - The Distribution of number of replicates per condition in 180 experiments of EMBL-EBI database. Altogether, number of available conditions for approximately 1047 different conditions were counted and the figure shows the distribution fitted to histogram of distribution (with mean of 4.41 and standard deviation of 0.71) (Conditions with more than 30 replicates were excluded from the graph to provide a better illustration.)

Results from Figure 4 shoes that, the distribution of number of samples per condition is highly skewed to low number of replicates, and this shows that the experiments are being performed with low number of samples.

In this research in order to consider sampling bias in differential expression analysis, we performed the sampling bias analysis, mentioned in section 4.4.2, to provide some examples of how using different combinations of replicates in a microarray analysis could result in different outcomes.

In Figure 5, the results from expression analysis of five of the experiments are reported. The experiments have been repeated for different number of replicates and also for all possible different combinations of replicates. This consists of sixteen different conditions. In Figure 5, the number of differentially expressed genes in each of these sixteen combinations is illustrated. It is seen that by changing both the number of replicates and the combination of available replicates different outcomes are obtained. The number of differentially expressed genes was consistent with the original studies, the minor differences might be due to the choice of software, methods or parameters.

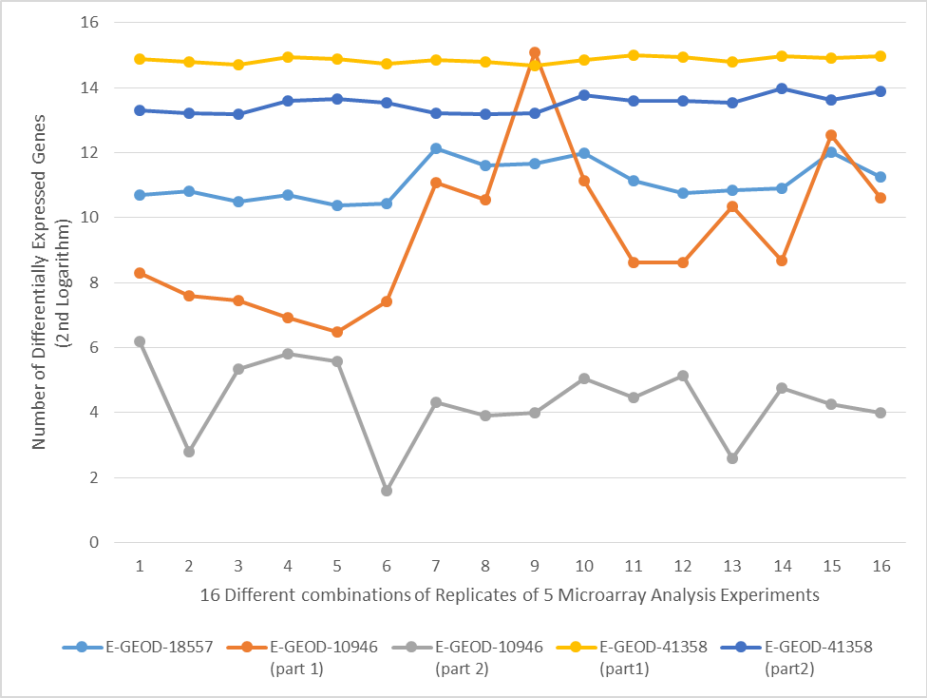

Figure 5 - The Number of Differentially Expressed Genes versus 16 Different Combinations of Replicates for 5 Different Microarray Experiments

The results showed that using different number of replicates, the number of expressed genes will either increase or decrease and no regular pattern is seen. Using different combinations also lead to different number of expressed genes. This finding is especially critical since biased results obtained from microarray experiments would lead to inaccurate inferences in subsequent pathway analysis.

5.3 Results from Assessment of the variation of housekeeping genes expression

To assess the possible variation of housekeeping genes after RMA normalization, we followed the procedure mentioned in section 4.5. It should be noted that we did this evaluation separately for each organism with regards to common housekeeping genes in that organism also separately for each tissue of interest. The result of this assessment for human organism in three tissue types is given below in Figure 6.

The points in Figure 6 show the mean expression value of each housekeeping gene after normalizing with RMA, for all datasets in each category (tissue type). The error bars demonstrate the standard deviation of the expression values.

In figure 6-a, the HKG expression values for the six datasets of "Oocyte" category are shown. It can be seen that YWHAZ expression shows significant variation, in most of the datasets, on the other hand, TBP and B2M show lower variation. Figure 6-b demonstrates the HKG expression values in category of five "cumulus" datasets. In this category, GAPDH expression shows lower variation in comparison to oocyte category, however the variation in B2M is much higher than oocyte datasets. In figure 6-c the expression values of eight HKG are shown in five datasets of "endometrium" category. In this category, overall the variation in HKG expression values is lower than the two other.

Similarly, for cattle and mouse organisms, the expression values of common housekeeping genes showed non-negligible variation after RMA normalization. These genes are supposed to have stable and uniform expression in all samples specifically that the RMA normalization step has already been performed.

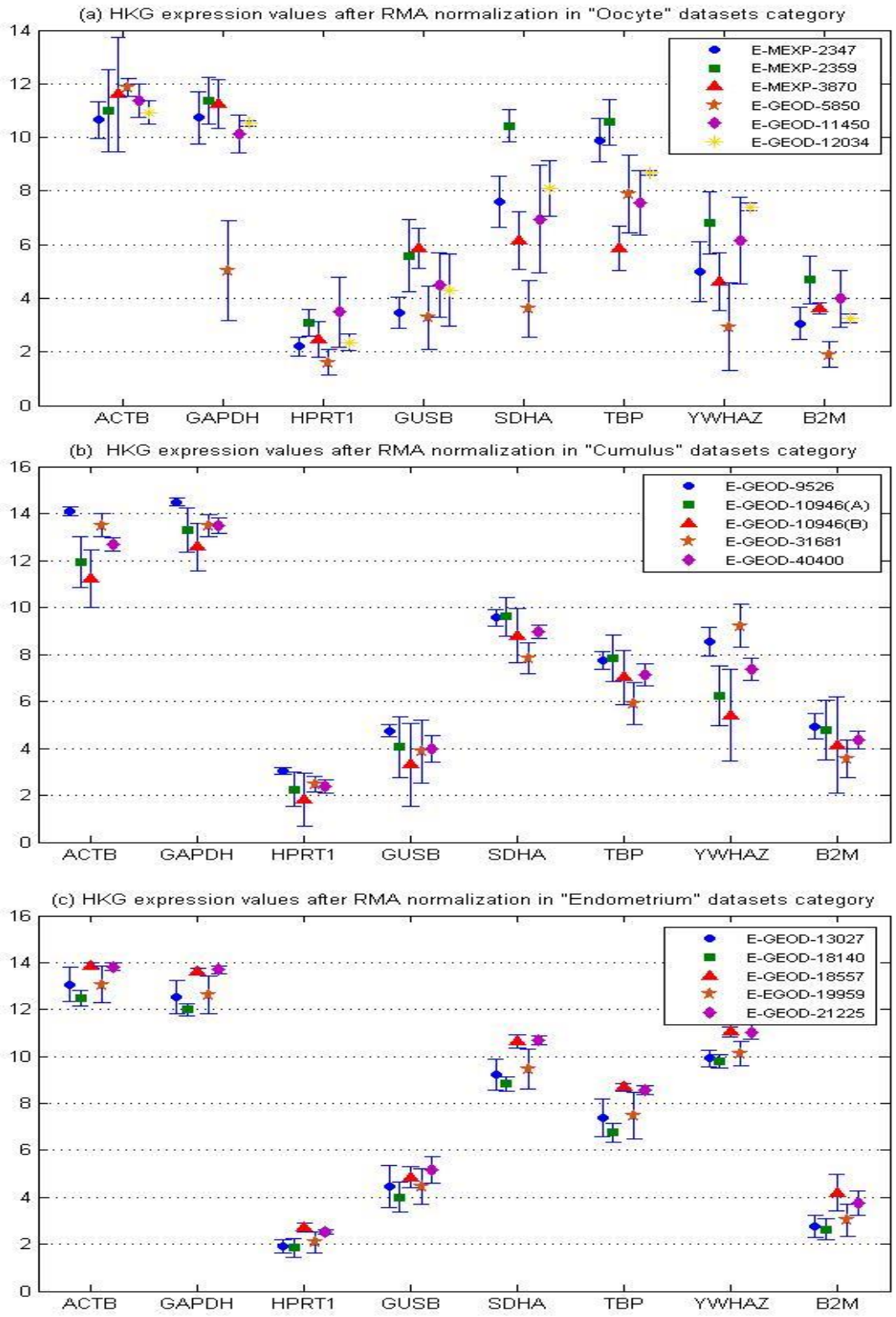

Figure 6 - House Keeping Genes expression after RMA normalization, 6-a, HKG Expression values of Oocyte datasets category. 6-b, HKG Expression values of Cumulus datasets category. 6-c, and HKG Expression values of Endometrium datasets category

5.4 Determination of most stable housekeeping genes in each tissue type

As mentioned in section 4.7, to determine the most stable housekeeping genes in any tissue type, we used Normfinder tool [53] for at least five experiments of that tissue type. Using the results of Normfinder, we ranked all common housekeeping genes in each organism, with regards to their stability, in the corresponding tissue type.

For human organism, the result of housekeeping genes' stability and their ranking is given below in Figure 7 and Table 1.

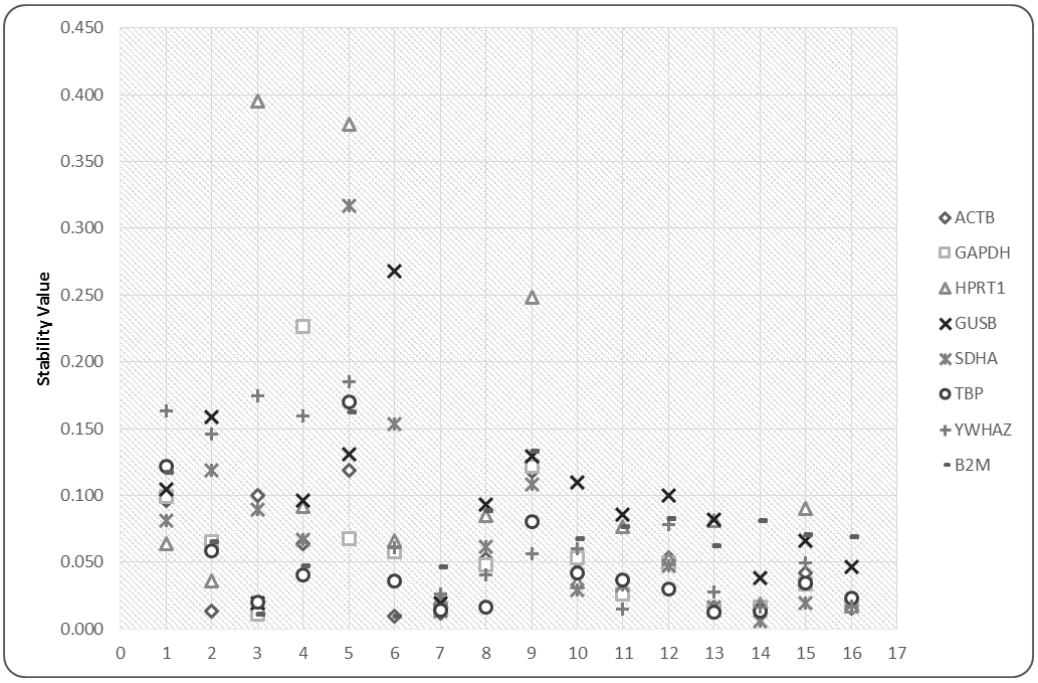

Figure 7 - Stability Values of Housekeeping Genes after RMA

Table 4 - HKG Ranking Based on Normfinder Results

Tissue Type	Oocyte							Cumulus						Endometrium						Overall Average Rank
	E-MEXP-2347	E-MEXP-2359	E-MEXP-3870	E-GEOD-5850	E-GEOD-11450	E-GEOD-12034	*Average Rank*	E-GEOD-9526	E-GEOD-10946 (A)	E-GEOD-10946 (B)	E-GEOD-31681	E-GEOD-40400	*Average Rank*	E-GEOD-13027	E-GEOD-18140	E-GEOD-18557	E-GEOD-19959	E-GEOD-21225	*Average Rank*	
ACTB	3	1	6	3	2	1	***2.67***	1	4	4	4	3	***3.20***	5	3	3	4	1	***3.20***	**3.000**
GAPDH	4	4	1	8	1	4	***3.67***	4	3	5	5	2	***3.80***	3	2	4	2	2	***2.60***	**3.375**
HPRT1	1	2	8	5	8	6	***5.00***	6	6	8	2	6	***5.60***	4	7	6	8	3	***5.60***	**5.375**
GUSB	5	8	3	6	3	8	***5.50***	5	8	6	8	8	***7.00***	8	8	7	6	7	***7.20***	**6.500**
SDHA	2	6	5	4	7	7	***5.17***	2	5	3	1	4	***3.00***	2	4	1	1	5	***2.60***	**3.688**
TBP	7	3	4	1	5	3	***3.83***	3	1	2	3	5	***2.80***	1	1	2	3	6	***2.60***	**3.125**
YWHAZ	8	7	7	7	6	5	***6.67***	7	2	1	6	1	***3.40***	6	5	5	5	4	***5.00***	**5.125**
B2M	6	5	2	2	4	2	***3.50***	8	7	7	7	7	***7.20***	7	6	8	7	8	***7.20***	**5.813**

In figure 7, the stability values of all eight HKGs - extracted from Normfinder - for all sixteen datasets from human organism, are plotted. On the horizontal axis, numbers 1-16 are annotating sixteen experiments. Numbers 1-6 correspond to the six "oocyte" datasets; numbers 7-11 are the five "cumulus" datasets and numbers 12-16 correspond to the five "endometrium" datasets. For each experiment, the most stable housekeeping gene is ranked as 1st and the least stable one is ranked as 8th. Then for all the experiments on one cell type these ranks are averaged for each HKG in order to find out the three most stable genes in the corresponding cell type. The determined three most stable housekeeping genes are to be used for next-stage normalization in the experiments of that cell type.

Table 1 shows the ranking results, using Normfinder stability values, with the average rank for each housekeeping gene in each category and overall. According to the averaged ranks, for oocyte category, the three most stable HKG in ranking were ACTB, GAPDH and B2M; for "cumulus" category, SDHA, TBP and ACTB and for "Endometrium" category, SDHA, TBP and GAPDH. Additionally, the overall rank for each HKG was calculated as the average of each HKG's ranks in all sixteen datasets. With regards to overall average rank, the three most stable genes across all sixteen datasets were ACTB, TBP and GAPDH.

5.5 Performing the second stage normalization using most stable housekeeping genes

5.5.1 Calculation of Normalization Factors

In order to perform the second-stage normalization based on housekeeping genes, for each organism, we found the three most stable housekeeping genes with regards to ranking of results from Normfinder stability values (section 5.4). To determine the normalization factors, for each sample, we calculated the geometric mean of 3 most stable genes' expression values.

5.5.2 Performance of second stage normalization

For implementation of second-stage normalization, we divided all expression values of any sample by the corresponding normalization factor calculated as mentioned in previous section. The results of this step are shown together with the differential expression analysis of selected datasets and also sampling bias results.

5.6 Repeating the Statistical analysis with regards to sampling bias, after second-stage normalization

Similar to section 5.2, we repeated the differential expression analysis of each experiment dataset with regards to sampling bias. The number of differentially expressed genes in each of sixteen combinations of available replicates was determined. In the next section, the results of this step are shown before and after second-stage normalization to compare the results and to determine the efficiency of second-stage normalization.

5.7 Comparing the sampling bias results after one-stage and second-stage normalization

To evaluate the performance of two-stage normalization method, the differential expression analysis was performed i) after RMA and ii) after RMA and HKG normalization.

In the following figures the results from expression analysis of selected datasets – from different organisms and tissue types– are shown. The number of differentially expressed genes

(scaled by Log 2) are reported before and after applying second stage normalization. The second stage normalization was performed by applying a normalization factor calculated from expressions of three most stable HKG expressions in each category to each sample. The number of housekeeping genes used for second-stage normalization depends on practical conditions. Two or three HKGs can be used for calculation of normalization factor. To perform the second-stage normalization more reliably, we used three most stable housekeeping genes to calculate normalization factor.

- ***Organism: human, Tissue type: oocyte***

E-GEOD-5850	HKG Expression values after RMA normalization									
	c1	c2	c3	c4	c5	m1	m2	m3	m4	m5
ACTB	11.85782	11.88655	11.9256	11.91462	12.15946	10.9799	12.03773	11.91351	12.24564	12.05573
GAPDH	3.732396	2.741243	1.750927	5.797959	6.563384	7.56764	3.41541	3.877315	6.801594	6.591599
HPRT1	1.635095	2.272416	1.433101	1.743296	1.483875	1.405752	1.910323	1.941364	1.571263	2.252108
GUSB	4.650752	2.929576	2.770791	2.629006	3.34069	2.021668	3.529802	4.839217	4.246946	4.925333
SDHA	3.534166	3.013954	2.868877	2.840847	3.07227	4.175638	3.298335	3.315453	5.236586	5.686747
TBP	8.240454	6.922446	8.133628	8.762368	8.485239	5.315039	8.150898	7.827516	9.619674	9.482792
YWHAZ	2.230368	2.611605	2.868877	1.851651	2.224857	1.468459	6.413231	4.494839	5.236586	3.227511
B2M	1.833157	2.402523	1.992767	1.743296	1.911827	1.436813	2.049621	2.630831	1.959226	2.473204

	three most stable HKG Expression values and normalization factors									
GAPDH	3.732396	2.741243	1.750927	5.797959	6.563384	7.56764	3.41541	3.877315	6.801594	6.591599
TBP	8.240454	6.922446	8.133628	8.762368	8.485239	5.315039	8.150898	7.827516	9.619674	9.482792
B2M	1.833157	2.402523	1.992767	1.743296	1.911827	1.436813	2.049621	2.630831	1.959226	2.473204
NF	3.834536	3.572385	3.050256	4.457479	4.739656	3.866242	3.849822	4.306085	5.042182	5.366968

E-GEOD-5850	HKG Expression values after RMA & HKG normalization									
	c1	c2	c3	c4	c5	m1	m2	m3	m4	m5
ACTB	3.092373	3.327342	3.909705	2.67295	2.565474	2.839941	3.126828	2.766669	2.42864	2.246283
GAPDH	0.973363	0.767343	0.574026	1.300726	1.384781	1.957363	0.887161	0.900427	1.348939	1.228179
HPRT1	1.915138	2.180579	2.408766	1.904401	1.804373	1.965337	2.063431	1.954042	1.755355	1.638917
GUSB	1.212859	0.820062	0.90838	0.589797	0.704838	0.522903	0.916874	1.123809	0.842283	0.917712
SDHA	0.921667	0.843681	0.940536	0.637322	0.648205	1.080025	0.85675	0.769946	1.038556	1.059583
TBP	2.149009	1.937766	2.66654	1.965768	1.790265	1.37473	2.117214	1.81778	1.90784	1.766881
YWHAZ	0.581653	0.731054	0.940536	0.415403	0.469413	0.379816	1.665852	1.043834	1.038556	0.601366
B2M	0.478065	0.672526	0.653311	0.391095	0.403368	0.37163	0.532394	0.610956	0.388567	0.46082

	One Stage	Two Stage
# of Differentially Expressed Genes (16 different combinations of samples)	252	213
	254	220
	246	227
	315	307
	367	340
	391	953
	8543	2018
	15640	706
	8723	654
	15394	1050
	328	254
	360	221
	372	271
	545	421
	15949	1055
	710	323
Mean	4274.313	577.0625
STD	6295.316	491.1013
# of Common Genes	69	26

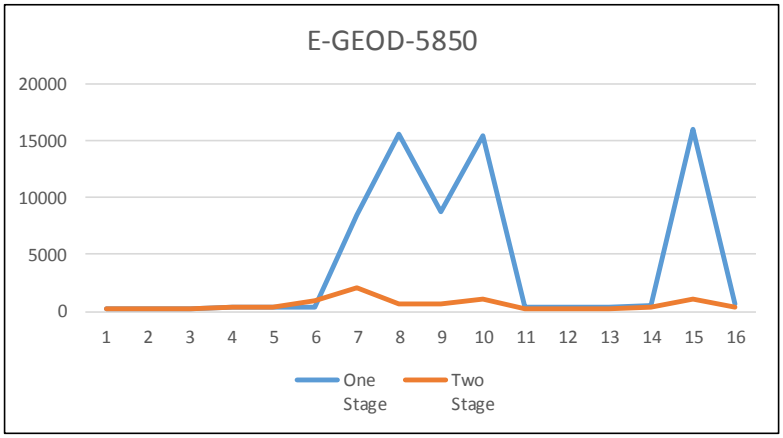

Figure 8 – Expression analysis results from human oocyte experiments

- ***Organism: human, Tissue type: cumulus***

E-GEOD-31681	HKG Expression values after RMA normalization									
	c1	c2	c3	c4	c5	m1	m2	m3	m4	m5
ACTB	13.8281	13.05693	13.1689	13.4153	12.83646	13.83447	13.32773	13.78564	13.79564	12.68108
GAPDH	14.15944	12.73716	13.17545	13.60819	13.15195	13.56081	13.29733	13.5058	13.99176	12.5382
HPRT1	2.14585	2.421824	1.963571	2.726763	2.527545	2.613074	2.59926	2.826749	2.607878	2.11009
GUSB	3.095278	5.312293	2.218856	6.108262	2.669767	3.720979	2.824313	5.043417	3.088045	6.076126
SDHA	8.261859	7.858795	7.777291	8.339737	6.962398	7.50594	7.714645	8.063764	8.593196	7.710584
TBP	5.467731	6.411614	4.16382	6.832568	4.536928	5.949186	5.587892	6.239966	6.67522	5.987951
YWHAZ	10.32525	8.318005	8.030007	9.276065	8.009587	9.712125	7.77517	9.947737	9.920648	7.724294
B2M	3.995263	4.007746	2.372962	4.258788	2.916284	3.13531	3.729776	4.944763	4.217208	3.229733

	three most stable HKG Expression values and normalization factors									
SDHA	8.261859	7.858795	7.777291	8.339737	6.962398	7.50594	7.714645	8.063764	8.593196	7.710584
TBP	5.467731	6.411614	4.16382	6.832568	4.536928	5.949186	5.587892	6.239966	6.67522	5.987951
YWHAZ	10.32525	8.318005	8.030007	9.276065	8.009587	9.712125	7.77517	9.947737	9.920648	7.724294
NF	7.755238	7.483662	6.382813	8.085373	6.324754	7.569356	6.946371	7.939895	8.286798	7.091554

E-GEOD-31681	HKG Expression values after RMA & HKG normalization									
	c1	c2	c3	c4	c5	m1	m2	m3	m4	m5
ACTB	1.783067	1.744725	2.063181	1.659206	2.029559	1.827695	1.918662	1.73625	1.664773	1.788195
GAPDH	1.825791	1.701995	2.064208	1.683063	2.079441	1.791541	1.914285	1.701005	1.68844	1.768047
HPRT1	1.122303	1.043366	1.27231	1.077768	1.344347	1.299617	1.307872	1.163413	1.167251	1.222158
GUSB	0.399121	0.709852	0.34763	0.755471	0.422114	0.491585	0.406588	0.635199	0.372646	0.856811
SDHA	1.065326	1.050127	1.218474	1.03146	1.100817	0.991622	1.110601	1.015601	1.036974	1.087291
TBP	0.705037	0.856748	0.652349	0.845053	0.717329	0.785957	0.804433	0.7859	0.805525	0.844378
YWHAZ	1.33139	1.111489	1.258067	1.147265	1.266388	1.283085	1.119314	1.25288	1.197163	1.089224
B2M	0.51517	0.535533	0.371774	0.526728	0.461091	0.414211	0.536939	0.622774	0.508907	0.455434

	One Stage	Two Stage
# of Differentially Expressed Genes (16 different combinations of samples)	1989	1983
	2031	1925
	828	1933
	5784	1773
	3525	1927
	2723	1546
	3008	2382
	898	1632
	903	1551
	3418	1392
	3441	2250
	1577	2098
	1584	1861
	4200	1807
	1759	1813
	2598	2194
Mean	2516.625	1879.188
STD	1351.542	270.0294
# of Common Genes	261	315

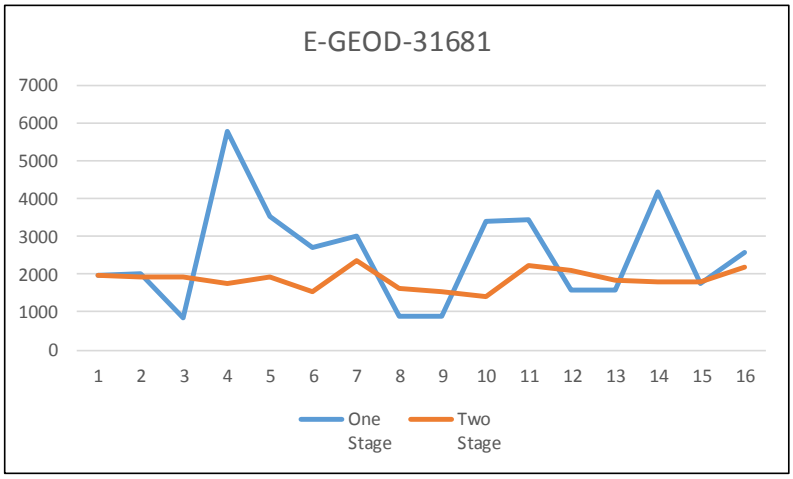

Figure 9 - Expression analysis results from human cumulus experiments

- ***Organism: human, Tissue type: Endometrium***

E-GEOD-19959	HKG Expression values after RMA normalization									
	c1	c2	c3	c4	c5	m1	m2	m3	m4	m5
ACTB	12.32358	12.32848	13.08287	12.22051	12.51565	13.94165	13.9027	14.00967	13.85187	13.75201
GAPDH	11.80284	11.97106	12.51188	11.77091	12.05109	13.48953	13.841	13.781	13.44233	13.14658
HPRT1	1.644336	1.609586	2.054068	1.538246	1.767719	2.384428	2.621083	2.99717	2.233195	2.113789
GUSB	4.259819	4.341887	4.004103	2.819841	5.036299	4.30564	5.409072	5.302037	5.593195	4.758799
SDHA	8.948237	8.680037	9.308511	8.38829	8.829136	10.38124	10.73157	11.05718	10.01161	10.0579
TBP	6.169033	6.908792	7.210191	6.37614	7.137658	8.887878	8.662814	9.010668	8.641234	7.538625
YWHAZ	9.999937	9.993162	9.881229	9.453401	9.894537	10.62545	11.12431	11.08534	10.23993	10.43664
B2M	2.011729	3.189112	2.525651	2.112913	2.769494	3.168217	3.690591	4.634283	3.240114	3.171882

	three most stable HKG Expression values and normalization factors									
GAPDH	11.80284	11.97106	12.51188	11.77091	12.05109	13.48953	13.841	13.781	13.44233	13.14658
SDHA	8.948237	8.680037	9.308511	8.38829	8.829136	10.38124	10.73157	11.05718	10.01161	10.0579
TBP	6.169033	6.908792	7.210191	6.37614	7.137658	8.887878	8.662814	9.010668	8.641234	7.538625
NF	8.669226	8.954034	9.434448	8.570649	9.123612	10.75675	10.87668	11.11461	10.51602	9.989353

E-GEOD-19959	HKG Expression values after RMA & HKG normalization									
	c1	c2	c3	c4	c5	m1	m2	m3	m4	m5
ACTB	1.421532	1.376863	1.386712	1.425856	1.371787	1.296083	1.278212	1.260474	1.317216	1.376667
GAPDH	1.361464	1.336946	1.326191	1.373397	1.320869	1.254051	1.272538	1.2399	1.278271	1.316059
HPRT1	1.17194	1.188839	1.207414	1.212931	1.220595	1.189821	1.086609	1.119625	1.212329	1.203168
GUSB	0.491372	0.484908	0.424413	0.329011	0.552007	0.400273	0.497309	0.477033	0.531874	0.476387
SDHA	1.032184	0.9694	0.986651	0.978723	0.967724	0.96509	0.986658	0.994833	0.952034	1.006862
TBP	0.711601	0.771584	0.764241	0.743951	0.782328	0.82626	0.796457	0.810705	0.821721	0.754666
YWHAZ	1.153498	1.116051	1.047357	1.102997	1.084498	0.987793	1.022767	0.997367	0.973745	1.044777
B2M	0.232054	0.356165	0.267705	0.246529	0.303552	0.294533	0.339312	0.416954	0.308112	0.317526

	One Stage	Two Stage
# of Differentially Expressed Genes (16 different combinations of samples)	36207	11763
	34275	10227
	26224	7294
	31631	11293
	22091	8010
	30714	10219
	41034	17021
	29651	12154
	28862	10268
	25316	11126
	43249	16435
	39072	13078
	37699	12457
	36582	13350
	41744	16682
	43826	17104
Mean	34261.06	12405.06
STD	6737.441	3078.368
# of Common genes	16083	3199

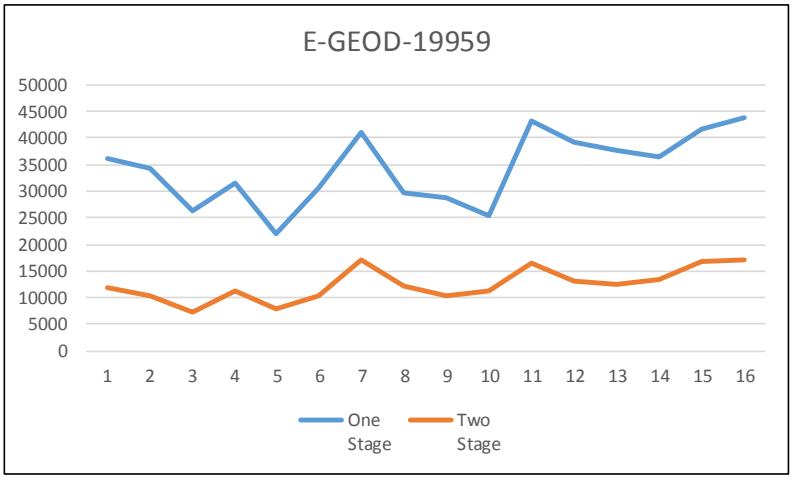

Figure 10 - Expression analysis results from human endometrium experiments

- ***Organism: human, Tissue type: lymphatic tissue***

E-MEXP-561	HKG Expression values after RMA normalization									
	c1	c2	c3	c4	c5	m1	m2	m3	m4	m5
ACTB	13.85505	13.74814	14.08605	14.08635	14.08511	13.39168	13.36608	13.79157	13.75598	13.82039
GAPDH	13.76705	13.67401	13.95432	13.94491	13.96152	13.1452	12.60064	13.26879	13.05511	13.28511
HPRT1	2.883177	3.213302	3.468444	3.019722	3.066779	2.97773	3.058862	3.48052	3.081499	3.094056
GUSB	5.244005	5.510057	5.727275	5.658711	6.180042	5.739623	5.328235	6.161439	5.649478	5.354249
SDHA	9.613894	9.795394	10.22832	10.0943	10.0573	7.632765	8.034452	8.635521	8.843195	8.421394
TBP	7.690392	7.472504	7.922342	7.359043	7.804066	4.373405	6.809763	7.969793	7.973088	6.904973
YWHAZ	11.72091	11.71193	12.01178	11.75162	11.73957	7.968484	9.426657	10.39628	8.978049	10.05235
B2M	4.756083	4.379642	5.502987	5.239885	4.996738	4.34594	5.57524	5.191091	6.991736	6.147718

	three most stable HKG Expression values and normalization factors									
ACTB	13.85505	13.74814	14.08605	14.08635	14.08511	13.39168	13.36608	13.79157	13.75598	13.82039
GAPDH	13.76705	13.67401	13.95432	13.94491	13.96152	13.1452	12.60064	13.26879	13.05511	13.28511
SDHA	9.613894	9.795394	10.22832	10.0943	10.0573	7.632765	8.034452	8.635521	8.843195	8.421394
NF	12.24004	12.25708	12.6212	12.5631	12.55235	11.03472	11.06074	11.64782	11.66703	11.56351

E-MEXP-561	HKG Expression values after RMA & HKG normalization									
	c1	c2	c3	c4	c5	m1	m2	m3	m4	m5
ACTB	1.131945	1.121648	1.116063	1.121248	1.12211	1.213594	1.208425	1.184048	1.179047	1.195172
GAPDH	1.124755	1.1156	1.105625	1.10999	1.112264	1.191258	1.139222	1.139165	1.118974	1.148882
HPRT1	0.235553	0.262159	0.274811	0.240364	0.244319	0.269851	0.276551	0.298813	0.26412	0.267571
GUSB	0.428431	0.449541	0.453782	0.450423	0.492342	0.520142	0.481725	0.528978	0.484226	0.46303
SDHA	0.785447	0.799162	0.810408	0.803488	0.801229	0.691704	0.726394	0.741385	0.757964	0.728273
TBP	0.628298	0.609648	0.627701	0.585766	0.621722	0.396331	0.61567	0.684231	0.683386	0.597134
YWHAZ	0.957588	0.955523	0.951715	0.935407	0.935249	0.722128	0.852263	0.892552	0.769523	0.869316
B2M	0.388568	0.357315	0.436011	0.417085	0.398072	0.393842	0.504057	0.445671	0.599273	0.531648

	One Stage	Two Stage
# of Differentially Expressed Genes (16 different combinations of samples)	8570	14462
	13389	11957
	14098	11296
	6999	10267
	7291	9906
	12353	10899
	14510	20375
	14573	19920
	18280	18107
	16052	17907
	13923	18036
	14265	17683
	18020	16440
	13160	15556
	18988	23237
	18205	21221
Mean	13917.25	16079.31
STD	3733.786	4213.844
# of Common genes	4655	5288

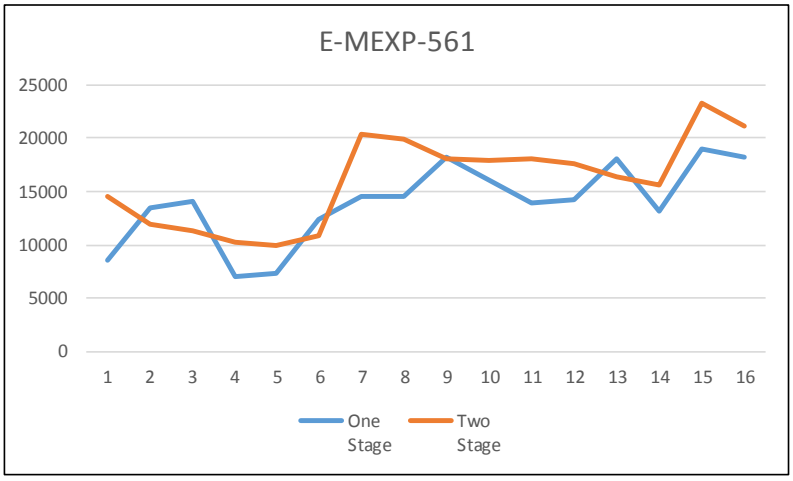

Figure 11 - Expression analysis results from human lymphatic tissue experiments

- ***Organism: human, Tissue type: Peripheral blood***

E-MEXP-3582	HKG Expression values after RMA normalization									
	c1	c2	c3	c4	c5	m1	m2	m3	m4	m5
ACTB	13.85763	13.79734	13.45677	13.18031	13.60317	13.44107	13.12547	13.88252	13.76569	13.40575
GAPDH	12.51021	12.01822	12.11781	11.74576	11.1199	9.599173	10.28684	12.56026	12.49757	11.20027
HPRT1	3.074319	2.930862	2.831211	2.686687	2.595961	2.183625	1.96858	2.534069	2.650169	2.250939
GUSB	7.757971	7.958579	7.710963	6.987252	7.06384	7.59034	7.331546	7.182556	6.538058	6.306407
SDHA	10.84036	10.68453	10.43726	10.19214	9.986967	9.900439	8.775019	10.9471	10.36734	10.085
TBP	9.997831	9.56749	9.295821	8.950917	8.643103	9.436122	8.082522	10.02827	9.577295	9.48813
YWHAZ	10.91473	10.97352	10.36908	10.00264	10.20554	10.47646	9.815496	10.78649	10.39439	10.05683
B2M	10.17301	10.92017	10.37452	9.997002	9.450781	9.270803	8.823591	9.351527	9.209281	9.939791

	three most stable HKG Expression values and normalization factors									
ACTB	13.85763	13.79734	13.45677	13.18031	13.60317	13.44107	13.12547	13.88252	13.76569	13.40575
HPRT1	3.074319	2.930862	2.831211	2.686687	2.595961	2.183625	1.96858	2.534069	2.650169	2.250939
SDHA	10.84036	10.68453	10.43726	10.19214	9.986967	9.900439	8.775019	10.9471	10.36734	10.085
NF	7.729661	7.559889	7.353596	7.119825	7.065192	6.623519	6.097783	7.275485	7.231797	6.726311

E-MEXP-3582	HKG Expression values after RMA & HKG normalization									
	c1	c2	c3	c4	c5	m1	m2	m3	m4	m5
ACTB	1.792786	1.825072	1.829957	1.851212	1.925378	2.029294	2.152498	1.908123	1.903494	1.993032
GAPDH	1.618468	1.589734	1.647875	1.649726	1.5739	1.449256	1.686981	1.726381	1.728141	1.665143
HPRT1	0.39773	0.387686	0.38501	0.377353	0.36743	0.329677	0.322835	0.348302	0.366461	0.334647
GUSB	1.003663	1.052737	1.048598	0.98138	0.999809	1.145968	1.20233	0.987227	0.904071	0.937573
SDHA	1.402436	1.413319	1.41934	1.431515	1.413545	1.49474	1.439051	1.504656	1.433577	1.499336
TBP	1.293437	1.265559	1.264119	1.257182	1.223336	1.424639	1.325485	1.378365	1.324331	1.410599
YWHAZ	1.412058	1.451545	1.410069	1.404899	1.444481	1.581706	1.609683	1.48258	1.437317	1.495148
B2M	1.3161	1.444488	1.410809	1.404108	1.337654	1.399679	1.447016	1.285348	1.273443	1.477748

	One Stage	Two Stage
# of Differentially Expressed Genes (16 different combinations of samples)	3024	4864
	5797	3809
	6664	4936
	3566	4828
	3501	5493
	6057	4824
	5080	5929
	4457	5449
	10006	5644
	6152	8387
	4811	5415
	4675	5754
	7973	5142
	5092	6371
	7151	7052
	6652	6590
Mean	1030.25	1323.188
STD	1322.733	816.7808
# of Common genes	1147	1331

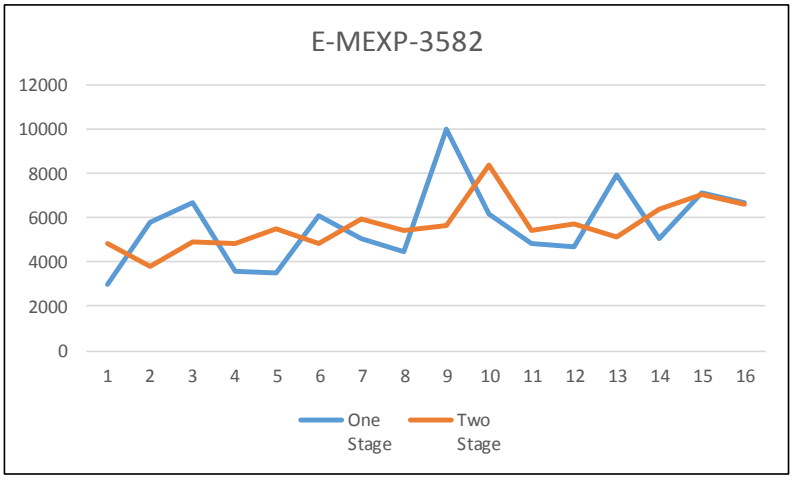

Figure 12 - Expression analysis results from human peripheral experiments

- ***Organism: Cattle, Tissue type: oocyte***

E-GEOD-57907	HKG Expression values after RMA normalization									
	c1	c2	c3	c4	c5	m1	m2	m3	m4	m5
ACTB	13.22861	13.13412	13.0998	13.14546	13.41376	12.98983	13.16213	13.05384	12.60921	12.68814
GAPDH	12.3547	11.95822	12.01707	12.49011	12.47447	11.67836	12.02491	12.29992	11.45745	12.01942
HPRT1	10.20535	10.15745	10.726	10.48587	10.99931	10.57597	10.62328	10.75612	9.677967	9.868044
GUSB	7.6556	7.644421	7.254935	7.814039	7.788011	7.038474	7.99628	7.610865	7.314156	7.252623
YWHAZ	6.85343	7.443593	8.06469	8.050067	7.894918	8.411339	8.136174	8.292222	6.377478	6.483743
B2M	12.76064	12.51671	12.21303	12.72984	12.82913	12.25251	13.14104	12.96461	12.0822	12.48646
HMBS	7.739002	7.088294	7.590513	7.276498	7.451348	7.361588	7.42912	6.747902	6.64588	7.038869
UBC	12.02536	11.95858	12.04976	11.71082	13.03738	11.86668	12.15113	12.20349	11.51133	12.12944
PKG1	10.45598	10.12293	10.54568	11.03551	10.87806	10.37445	10.76433	10.27668	9.933898	10.10917

	three most stable HKG Expression values and normalization factors									
ACTB	13.22861	13.13412	13.0998	13.14546	13.41376	12.98983	13.16213	13.05384	12.60921	12.68814
GAPDH	12.3547	11.95822	12.01707	12.49011	12.47447	11.67836	12.02491	12.29992	11.45745	12.01942
PKG1	10.45598	10.12293	10.54568	11.03551	10.87806	10.37445	10.76433	10.27668	9.933898	10.10917
NF	11.95557	11.67144	11.84075	12.19116	12.20979	11.63189	11.9435	11.81674	11.27973	11.55222

E-GEOD-57907	HKG Expression values after RMA & HKG normalization									
	c1	c2	c3	c4	c5	m1	m2	m3	m4	m5
ACTB	1.106481	1.125321	1.106332	1.078279	1.098607	1.116743	1.102033	1.10469	1.117865	1.098329
GAPDH	1.033384	1.024571	1.014891	1.024522	1.021678	1.003995	1.006816	1.040889	1.015756	1.040442
HPRT1	0.853606	0.870282	0.905855	0.860121	0.90086	0.909223	0.889461	0.910244	0.857997	0.854212
GUSB	0.640337	0.654968	0.612709	0.64096	0.63785	0.605102	0.669509	0.644075	0.648434	0.627812
YWHAZ	0.573241	0.637761	0.681096	0.66032	0.646606	0.723128	0.681222	0.701735	0.565393	0.561255
B2M	1.067339	1.072421	1.03144	1.044187	1.050725	1.053355	1.100267	1.097139	1.071143	1.08087
HMBS	0.647313	0.607319	0.64105	0.596867	0.610277	0.63288	0.622022	0.571046	0.589188	0.609309
UBC	1.005837	1.024602	1.017652	0.9606	1.067781	1.020186	1.017384	1.032729	1.020533	1.049966
PKG1	0.874569	0.867325	0.890626	0.905207	0.89093	0.891898	0.901271	0.869671	0.880686	0.875084

	One Stage	Two Stage
# of Differentially Expressed Genes (16 different combinations of samples)	693	987
	374	506
	435	600
	1659	893
	1745	1007
	2778	1108
	1016	938
	1328	1603
	1192	871
	4068	1943
	727	611
	796	838
	881	545
	2967	1134
	1843	1265
	1487	794
Mean	1030.25	1323.188
STD	1322.733	816.7808
# of Common genes	84	73

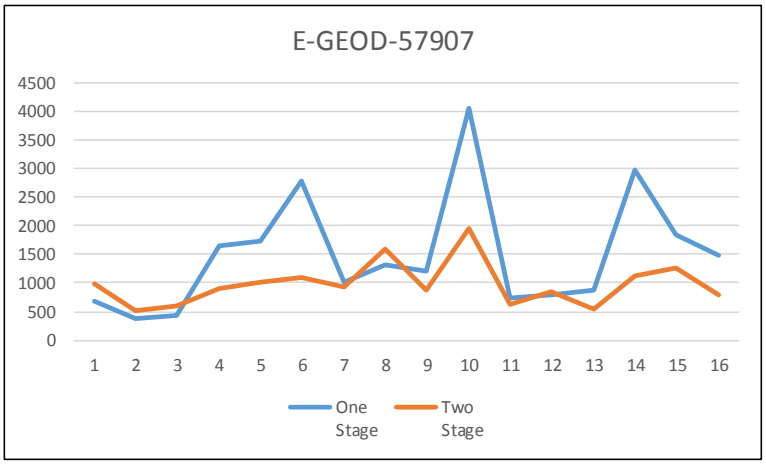

Figure 13 - Expression analysis results from cattle skin experiments

The results showed that after normalization with RMA method, there exists some non-negligible variation in the expression values of housekeeping genes. This shows that the data is not efficiently normalized. By implementing a second-stage normalization based on housekeeping genes' expression values, the non-biological variations will be decreased.

Chapter 6

THREATS TO VALIDITY

&

CONTRIBUTIONS

6.1 Threats to Validity

There are certain threats to the validity of this research. These threats are categorized as external, internal, construct and statistical threats. Following the threats in all the categories are explained.

In regard to external validity, this study has been testing random datasets from standard microarray databases available [78]. Therefore it can be argued that our sampling procedure in this research was completely randomized and from this point of view, there is no concern regarding the external validity. Also, as our unit of analysis in this research was three different genomes, human, cattle and mouse genomes, it can be argued that we covered different genomes; however mentioned genomes are all from mammalian species, therefore utilizing other genomes such as drosophila or yeast could have yielded other results. This could impose a threat to the external validity of this research. On the other hand, the datasets were taken from different tissues types and in this regard, the threat to external validity of this study is forfeited. However more experiments on more organisms and tissue types could result in different outcomes. The last threat to external validity of this study is regarding the type of microarray experiments which was RNA assay. Other microarray types might not hold the same conclusions. Regarding Internal validity, the main concern would be certain conditions in these kinds of experiments which are not under our control and may affect the cause and effect relationship. In this study, although the effect of number of replicates and different combinations of them were tested, other factors such as number of control samples that we did not take into account, could have also imposed some bias into our procedure, as well. Therefore, this could compromise our internal validity. Additionally, we are not sure the other extraneous variables in this research are under our control, since we gather the data from available archival datasets online.

There were no threats to construct validity identified in this research. We used two well-defined constructs in this study and there were no vague definition of them.

When it comes to statistical threats to validity of this study, there are some areas of concern that we should take into account. In preprocessing step, at first stage, Quantile normalization method and at second stage, housekeeping genes normalization method have been used in the experiments of this paper. Though the threat of only using RMA normalization in preprocessing

stage was omitted by adding additional normalization application, but again changing preprocessing methods and approaches could have produced different results and thus this can be considered as a threat to the statistical validity of this research. On the other hand, for gene expression analysis stage, one of the statistical test tools (two sample t test) were utilized and the samples used for the test were adequate. Therefore in this regard, no certain threat is identified. We kept statistical parameters constant in all experiments in order to solely determine the effect of normalization on microarray data analysis. As we considered certain P-Value threshold and fold change in this research, choosing other thresholds might alter the results. This could be considered as a statistical threat to our research. Another threat to statistical validity of this paper is related to choice of housekeeping genes, we addressed only eight or nine of the most popular housekeeping genes for each organism in its analysis; choice of other internal controls could affect our results.

6.2 Contributions

6.2.1 Theoretical contributions

Microarray data analysis is one of the main procedures in genetic studies in order to anticipate the functions of different genes. Also microarray data analysis, is a strong tool to investigate the responsible genes in different diseases. Large quantity of research and studies are found in literature related to microarray data analysis, is a solid proof of importance of this analysis. As in any statistical method, microarray procedure is challenged by different biases and noises. Genetic researchers appreciate any tool or algorithm that can benefit microarray data analysis in order to make the results more consistent and reliable.

Our research hugely contributes to consistency of microarray data analysis by making big improvements in one of the major stages of the analysis, normalization. Removing the non-biological variations between samples could hugely benefit the whole analysis and lead to more accurate results.

Our proposed two-stage normalization method improve the quality of microarray data considerably. With accurate evaluation of expression values of certain control genes, the efficiency of first stage normalization is tested and second stage normalization is applied consequently. This precise normalization has benefits in preprocessing of microarray data.

6.2.2 Practical contributions

In practice, implementing our proposed method in all compatible microarray experiments, would lead to a more reliable protocol for any microarray experiment and its analysis. With better and more consistent data preprocessing, scientists may rely on microarray results with more confidence and make more accurate clinical inferences. The clinical inferences from microarray data analysis could greatly influence the medical interpretations, finding the responsible genetic factors of fatal diseases and many more medical advances.

Chapter 7

CONCLUSIONS

&

FUTURE WORK

7.1 Conclusions

Analysis of distribution of number of replicates per condition in more than 200 microarray experiments showed that in practice not too many samples per condition are available for microarray experiments. Therefore, although there are methods to calculate the sample size for microarray studies, practically researches tend to use 3-5 samples for their research. This approach in sample size determination may lead to the conclusion that not many studies follow the sample size calculation methods.

Looking at the results obtained from further testing of five datasets that we chose, we can see that the number of differentially expressed genes oscillates as the number of replicates was changed. In some experiments the number of expressed genes decreased as the number of replicates were increased. In some others it goes up and then decreases. Results like this show that sampling bias may cause inconsistent results and inferences. We have also seen that using different combinations of replicates produce different outcomes. The results show that considerable variations exist in samples of microarray data. Either inter-sample variations, possible outliers, distance of clusters could affect the results. This sampling bias could yield unreliable results. Therefore the corresponding inference from these results might be compromised.

It is noteworthy that, microarray experiments usually do not provide absolute conclusions alone. Rather, microarray results are useful for reducing the search space for determination of significant up- or down-regulations of gene expression. Despite the demonstrated variations occurring as a result of sampling bias, it is likely that the genes representing highest differential expressions would be preserved among analysis of different subsets within the same dataset. On the other hand, subsequent functional analysis take into account the entire list of differentially expressed genes, and therefore meticulous care should be taken to provide reliable and consistent results. An illustration of samples needs to be done in order to detect possible outliers and low-quality samples. Moreover, the rule of thumb that says three replicates is sufficient, must be questioned as different number of replicates could provide inconsistent results. Additional sampling could assist the analysis in making more consistent and reliable conclusions. We suggest that assessment of inter-sample variance prior to differential expression analysis is a

crucial step in microarray experiments and proper handling of that variance may require alternative normalization and/or statistical test methods.

If sampling bias exists in the data gathering, as we showed, then the results and furthermore the clinical inferences may be compromised. Sampling bias is mainly due to non-biological variation between samples. The aim of normalization is to eliminate the technical variations among samples.

RMA normalization might be able to remove all the variations when the purity of samples is not determined. The expression values of popular and common housekeeping genes were analyzed. The results showed that after normalization with RMA method, there exists some non-negligible variation in the expression values of housekeeping genes. This shows that the data is not efficiently normalized.

By applying a second normalization after RMA, which included the use of housekeeping genes normalization, the mentioned variations were decreased and therefore the oscillations in expression analysis results were also less afterwards. This statement is true about the datasets which the variation of housekeeping genes' expressions were high after RMA. For the dataset with low variation (standard deviation < 0.01) in expression of those genes, application of second normalization method did not have much effect on expression analysis results.

Microarray studies are performed to evaluate other related experiments and narrow the gene search space in determination of significant up- or down-regulations of gene expression. Therefore to have consistent results, one should reduce the non-biological variations as much as possible. Normalization is the most significant tool to decrease the variations in data. By using alternate normalization methods or even additive methods consequently, the variation could be reduced and sampling bias could partially be handled.

7.2 Future Work

7.2.1 Data analysis

This research has been done to improve the quality of microarray data and therefore making the outcome of these experiments more consistent. Our approach to fulfill this goal was to implement our proposed method in plenty of randomly selected datasets from one of the main microarray public databases (ArrayExpress) and validate our proposed solution.

In the future, it would be great to implement our proposed solution for all datasets in ArrayExpress and also other microarray public repositories. To fulfill this goal, first of all an automatic complete model may be developed from our proposed method. Then to implement the model, large scale processing tools is needed to analyze all data-sets on clusters of commodity hardware. Since there are different platforms in different microarray public databases, the model shall be modified according to different platforms.

It is suggested to implement our method on other organisms as well, to further prove the efficiency of the solution. Also as we mentioned in our methodology, we considered only biological replicates. In future studies, use of technical replicates as well as biological replicates is suggested. Another aspect in the future work is regarding the quality of samples. We suggest to repeat the differential expression analysis after assessing the samples' quality via vector analysis, Principle Component Analysis (PCA), etc. and excluding the outliers. In our solution, we have suggested to perform non-linear normalization at probe level and linear normalization at expression level; it will be good to implement linear normalization at both stages and repeat the analysis.

Our approach in differential expression analysis was to compare the number of differentially expressed genes when different number of samples and different combinations of available samples are considered in the analysis. However in some examples, the number of differentially expressed genes has increased after performing second-stage normalization, which is not in favor of biologists. This can be considered as a limitation of our study. In the future, both increasing the consistency of microarray data analysis, and the quantity of differentially expressed genes

should be considered together. By finding the appropriate fold change in expression analysis, the results of microarray data analysis might be more favorable to biologists.

Another work in future is regarded to the use of other statistical methods rather than t-test. Yuan's method could result in more robust conclusions. The performance of different statistical methods should be evaluated and compared with each other.

Finally the performance of our solution (two-stage normalization) could be compared with other normalization methods and previous studies. Our solution is a novel proposition with regard to using additive normalizations; therefore it will be appropriate to evaluate the performance of our method further and make comparisons with previous works. Also, using the available consistency measures would be helpful to evaluate the performance of our method.

7.2.2 Microarray experimental procedure

In this study we provided plenty of examples showing the sampling bias in microarray data analysis. We proposed a two-stage normalization method to tackle this problem. Though our solution properly works, it has its own restrictions like any preprocessing method. As in future we look forward to provide better insights for actual microarray laboratories and with their cooperation, improve the data gathering process to reduce the sampling bias before reaching the data preprocessing stage. In the light of our analysis, the consequences of unwanted variation may be foreseen and therefore improve the whole microarray experimental procedure and avoid inconsistent clinical inferences.

Bibliography

[1] Brazma, A., Robinson, A., Cameron, G., & Ashburner, M. (2000). One-stop shop for microarray data. Nature, 403(6771), 699-700.

[2] Sessmentseries, M. (2008). Viruses In Food: Scientific Advice To Support Risk Management Activities

[3] http://popyomics.biol.soton.ac.uk/~nat/downloads/refs.bib

[4] Montag, M. (Ed.). (2014). A Practical Guide to Selecting Gametes and Embryos. CRC Press.

[5] Gadbury, G. L., Page, G. P., Edwards, J., Kayo, T., Prolla, T. A., Weindruch, R., ... & Allison, D. B. (2004). Power and sample size estimation in high dimensional biology. Statistical Methods in Medical Research, 13(4), 325-338.

[6] Peng, X., & Stromberg, A. J. (2003). Microarray Experiment Design and Statistical Analysis. In A Beginner's Guide to Microarrays (pp. 243-275). Springer US.

[7] Allison, D. B., Cui, X., Page, G. P., & Sabripour, M. (2006). Microarray data analysis: from disarray to consolidation and consensus. Nature Reviews Genetics, 7(1), 55-65.

[8] Do, J. H., & Choi, D. (2006). Normalization of microarray data: single-labeled and dual-labeled arrays. Molecules and cells, 22(3), 254.

[9] Nguyen, D. V., Bulak Arpat, A., Wang, N., & Carroll, R. J. (2002). DNA microarray experiments: biological and technological aspects. Biometrics, 58(4), 701-717.

[10] Dobbin, K., & Simon, R. (2005). Sample size determination in microarray experiments for class comparison and prognostic classification. Biostatistics, 6(1), 27-38.

[11] http://essaymania.com/53906/reproduction-process

[12] Bolstad, B. M., Irizarry, R. A., Åstrand, M., & Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics, 19(2), 185-193.

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

[14] Goldstein, D. R., & Delorenzi, M. (2004). Statistical design and data analysis for microarray experiments. Understanding Lipid Metabolism with Microarrays and Other Omic Approaches, 1.

[15] Nwana, N. Microarray Cancer Data Visualization: A Comparative Study.

[16] Smith, G. W., & Rosa, G. J. M. (2007). Interpretation of microarray data: trudging out of the abyss towards elucidation of biological significance. Journal of animal science, 85(13 suppl), E20-E23.

[17] Carter, N. P. (2007). Methods and strategies for analyzing copy number variation using DNA microarrays. Nature genetics, 39, S16-S21.

[18] Hofnagel, O., Luechtenborg, B., Eschert, H., Weissen-Plenz, G., Severs, N. J., & Robenek, H. (2006). Pravastatin Inhibits Expression of Lectin-Like Oxidized Low-Density Lipoprotein Receptor-1 (LOX-1) in Watanabe Heritable Hyperlipidemic Rabbits A New Pleiotropic Effect of Statins. Arteriosclerosis, thrombosis, and vascular biology, 26(3), 604-610.

[19] Jørstad, T. S., Langaas, M., & Bones, A. M. (2007). Understanding sample size: what determines the required number of microarrays for an experiment?. Trends in plant science, 12(2), 46-50.

[20] Brown, P. O., & Botstein, D. (1999). Exploring the new world of the genome with DNA microarrays. Nature genetics, 21, 33-37.

[21] Qiu, W., Lee, M. L. T., & Whitmore, G. A. (2008). Sample Size and Power Calculation in Microarray Studies Using the size power package. Technical report, Bioconductor.

[22] http://www3.it.nuigalway.ie/agolden/bioconductor/version1/MicroArrayAnalysis.pdf

[23] http://www.bioinformaticstutorials.com/

[24] Park, T., Yi, S. G., Kang, S. H., Lee, S., Lee, Y. S., & Simon, R. (2003). Evaluation of normalization methods for microarray data. BMC bioinformatics, 4(1), 33.

[25] Xu, W. W., & Carter, C. J. (2010). Parallel multiplicity and error discovery rate (EDR) in microarray experiments. BMC bioinformatics, 11(1), 465.

[26] Baldi, P., & Hatfield, G. W. (2002). DNA microarrays and gene expression: from experiments to data analysis and modeling. Cambridge University Press.

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

[28] Dudoit, S., Yang, Y. H., Callow, M. J., & Speed, T. P. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Statistica sinica, 12(1), 111-140.

[29] Wang, S. J., & Chen, J. J. (2004). Sample size for identifying differentially expressed genes in microarray experiments. Journal of Computational Biology, 11(4), 714-726.

[30] http://discover.nci.nih.gov/microarrayAnalysis/Experimental.Design.jsp

[31] Lee, M. L. T., & Whitmore, G. A. (2002). Power and sample size for DNA microarray studies. Statistics in medicine, 21(23), 3543-3570.

[32] Tibshirani, R. (2006). A simple method for assessing sample sizes in microarray experiments. Bmc Bioinformatics, 7(1), 106.

[33] Ray chaudhuri, S., Stuart, J. M., & Altman, R. B. (2000). Principal components analysis to summarize microarray experiments: application to sporulation time series. In Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing (p. 455). NIH Public Access.

[34] Li, C., & Wong, W. H. (2003). DNA-chip analyzer (dChip). In The Analysis of Gene Expression Data (pp. 120-141). Springer New York.

[35] DNA microarray data analysis. CSC-Scientific Computing, 2003.

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

[37] Morris, D., Golden, A., & Hinde, J. BioconductorBuntu Users Manual.

[38] Wolfinger, R. D., Gibson, G., Wolfinger, E. D., Bennett, L., Hamadeh, H., Bushel, P., ... & Paules, R. S. (2001). Assessing gene significance from cDNA microarray expression data via mixed models. Journal of Computational Biology, 8(6), 625-637.

[39] White, C. A., & Salamonsen, L. A. (2005). A guide to issues in microarray analysis: application to endometrial biology. Reproduction, 130(1), 1-13.

[40] Pieterse, B., Jellema, R. H., & van der Werf, M. J. (2006). Quenching of microbial samples for increased reliability of microarray data. Journal of microbiological methods, 64(2), 207-216.

[41] Paul, S., Kim, S. J., Park, H. W., Lee, S. Y., An, Y. R., Oh, M. J., ... & Hwang, S. Y. (2011). Impact of miRNA deregulation on mRNA expression profiles in response to environmental toxicant, nonylphenol. Molecular & Cellular Toxicology, 7(3), 259-269.

[42] Hannah, M. A., Redestig, H., Leisse, A., & Willmitzer, L. (2008). Global mRNA changes in microarray experiments. Nature biotechnology, 26(7), 741-742.

[43] Eisenberg, E., & Levanon, E. Y. (2013). Human housekeeping genes, revisited. Trends in Genetics, 29(10), 569-574.

[44] Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., ... & Venter, J. C. (1995). The minimal gene complement of Mycoplasma genitalium. Science, 270(5235), 397-404.

[45] Koonin, E. V. (2000). How Many Genes Can Make a Cell: The Minimal-Gene-Set Concept 1. Annual review of genomics and human genetics, 1(1), 99-116.

[46] Thellin, O., Zorzi, W., Lakaye, B., De Borman, B., Coumans, B., Hennen, G., ... & Heinen, E. (1999). Housekeeping genes as internal standards: use and limits. Journal of biotechnology, 75(2), 291-295.

[47] Robinson, M. D., & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol, 11(3), R25.

[48] Dheda, K., Huggett, J. F., Bustin, S. A., Johnson, M. A., Rook, G., & Zumla, A. (2004). Validation of housekeeping genes for normalizing RNA expression in real-time PCR. Biotechniques, 37, 112-119.

[49] Rubie, C., Kempf, K., Hans, J., Su, T., Tilton, B., Georg, T. & Schilling, M. (2005). Housekeeping gene variability in normal and cancerous colorectal, pancreatic, esophageal, gastric and hepatic tissues. Molecular and cellular probes, 19(2), 101-109.

[50] Ganapathi, M., Srivastava, P., Sutar, S. K., Kumar, K., Dasgupta, D., Singh, G. P. & Brahmachari, S. K. (2005). Comparative analysis of chromatin landscape in regulatory regions of human housekeeping and tissue specific genes. BMC bioinformatics, 6(1), 126.

[51] Szabo, A., Perou, C. M., Karaca, M., Perreard, L., Quackenbush, J. F., & Bernard, P. S. (2004). Statistical modeling for selecting housekeeper genes. Genome biology, 5(8), R59.

[52] Vandesompele, J., De Preter, K., Pattyn, F., Poppe, B., Van Roy, N., De Paepe, A., & Speleman, F. (2002). Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. Genome biology, 3(7), research0034.

[53] Andersen, C. L., Jensen, J. L., & Ørntoft, T. F. (2004). Normalization of real-time quantitative reverse transcription-PCR data: a model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. Cancer research, 64(15), 5245-5250.

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

[55] Manafi, S., Uyar, A., & Bener, A. (2013, September). Sampling bias in microarray data analysis: A demonstration in the field of reproductive biology. In Health Informatics and Bioinformatics (HIBIT), 2013 8th International Symposium on (pp. 1-7). IEEE.

[56] Pawitan, Y., Michiels, S., Koscielny, S., Gusnanto, A., & Ploner, A. (2005). False discovery rate, sensitivity and sample size for microarray studies. Bioinformatics, 21(13), 3017-3024.

[57] Pan, W., Lin, J., & Le, C. T. (2002). How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach. Genome Biol, 3(5), 1-0022.

[58] Yang, M. C., Yang, J. J., McIndoe, R. A., & She, J. X. (2003). Microarray experimental design: power and sample size considerations. Physiological Genomics, 16(1), 24-28.

[59] Zien, A., Fluck, J., Zimmer, R., & Lengauer, T. (2003). Microarrays: how many do you need?. Journal of Computational Biology, 10(3-4), 653-667.

[60] Gadbury, G. L., Page, G. P., Edwards, J., Kayo, T., Prolla, T. A., Weindruch, R., ... & Allison, D. B. (2004). Power and sample size estimation in high dimensional biology. Statistical Methods in Medical Research, 13(4), 325-338.

[61] Tsai, C. A., Wang, S. J., Chen, D. T., & Chen, J. J. (2005). Sample size for gene expression microarray experiments. Bioinformatics, 21(8), 1502-1508.

[62] Ferré, M. C. J. (2008). Missing data matrix factorization addressing the structure from motion problem (Doctoral dissertation, Universitat Autònoma de Barcelona).

[63] Müller, P., Parmigiani, G., Robert, C., & Rousseau, J. (2004). Optimal sample size for multiple testing: the case of gene expression microarrays. Journal of the American Statistical Association, 99(468), 990-1001.

[64] Lin, W. J., Hsueh, H. M., & Chen, J. J. (2010). Power and sample size estimation in microarray studies. BMC bioinformatics, 11(1), 48.

[65] Murakami, Y., & Mizuguchi, K. (2014). Homology-based prediction of interactions between proteins using Averaged One-Dependence Estimators. BMC bioinformatics, 15(1), 213.

[66] Benito, M., Parker, J., Du, Q., Wu, J., Xiang, D., Perou, C. M., & Marron, J. S. (2004). Adjustment of systematic microarray data biases. Bioinformatics, 20(1), 105-114.

[67] http://www.dkfzheidelberg.de/mga/home/hsueltma/Protocols/Normalization_of_microarray_data.pdf

[68] Quackenbush, J. (2002). Microarray data normalization and transformation. Nature genetics, 32, 496-501.

[69] Bilban, M., Buehler, L. K., Head, S., Desoye, G., & Quaranta, V. (2002). Normalizing DNA microarray data. Current Issues in Molecular Biology, 4, 57-64.

[70] Kepler, T. B., Crosby, L., & Morgan, K. T. (2002). Normalization and analysis of DNA microarray data by self-consistency and local regression. Genome biol, 3(7), 1-12.

[71] Wang, Y., Lu, J., Lee, R., Gu, Z., & Clarke, R. (2002). Iterative normalization of cDNA microarray data. Information Technology in Biomedicine, IEEE Transactions on, 6(1), 29-37.

[72] Workman, C., Jensen, L. J., Jarmer, H., Berka, R., Gautier, L., Nielser, H. B., ... & Knudsen, S. (2002). A new non-linear normalization method for reducing variability in DNA microarray experiments. Genome biol, 3(9), 1-16.

[73] Chen, Y. J., Kodell, R., Sistare, F., Thompson, K. L., Morris, S., & Chen, J. J. (2003). Normalization methods for analysis of microarray gene-expression data. Journal of biopharmaceutical statistics, 13(1), 57-74.

[74] Edwards, D. (2003). Non-linear normalization and background correction in one-channel cDNA microarray studies. Bioinformatics, 19(7), 825-833.

[75] Seli, E., Robert, C., & Sirard, M. A. (2010). OMICS in assisted reproduction: possibilities and pitfalls. Molecular human reproduction, 16(8), 513-530.

[76] Chen, H., & Tzeng, C. (2006). Applications of microarray in reproductive medicine. Chang Gung medical journal, 29(1), 15.

[77] Reinke, V., & White, K. P. (2002). Developmental genomic approaches in model organisms. Annual review of genomics and human genetics, 3(1), 153-178.

[78] European Bioinformatics Institute, www.ebi.ac.uk

[79] Lawson, M. J., & Zhang, L. (2008). Housekeeping and tissue-specific genes differ in simple sequence repeats in the 5′ -UTR region. Gene, 407(1), 54-62.

[80] de Kok, J. B., Roelofs, R. W., Giesendorf, B. A., Pennings, J. L., Waas, E. T., Feuth, T., ... & Span, P. N. (2004). Normalization of gene expression measurements in tumor tissues: comparison of 13 endogenous control genes. Laboratory investigation, 85(1), 154-159.

[81] Cressie, N. A. C., & Whitford, H. J. (1986). How to Use the Two Sample t‐Test. Biometrical Journal, 28(2), 131-148.

[82] Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Statistical applications in genetics and molecular biology, 3(1).

[83] http://www.mathworks.com/help/bioinfo/ref/mattest.html