

**Visual and Statistical-based Cross-Platform Normalization on gene
expression data of Oral cancer.**

A Major Project report submitted

in partial fulfillment of the requirement for the degree of

Master of Technology

In

Bioinformatics

Submitted by

Anjali Chaudhary

(2K14/BIO/01)

Delhi Technological University, Delhi, India

Under the supervision of

Prof. B. D. Malhotra



Department of Biotechnology
Delhi Technological University
(Formerly Delhi College of Engineering)
Shahbad Daultapur, Main Bawana Road,
Delhi-110042, INDIA



CERTIFICATE

This is to certify that the M. Tech. dissertation entitled “**Visual and Statistical-based Cross-Platform Normalization on gene expression data of Oral cancer**”, submitted by **Anjali Chaudhary (2K14/BIO/01)** in partial fulfillment of the requirement for the award of the degree of Master of Engineering, Delhi Technological University (Formerly Delhi College of Engineering), is an authentic record of the candidate’s own work carried out by her under my guidance.

The information and data enclosed in this dissertation are original and has not been submitted elsewhere for honoring of any other degree.

Prof. B. D. Malhotra

(Project Mentor)

Department of Biotechnology

Delhi Technological University

Delhi- 110042

Dr. D. Kumar

Professor and Head

Department of Bio-Technology

Delhi Technological University

Delhi- 110042

DECLARATION

I certify that the project report entitled “**Visual and Statistical-based Cross-Platform Normalization on gene expression data of Oral cancer**” submitted by me is in partial fulfilment of the requirement for the award of the degree of Master of Technology in Bioinformatics, Department of Biotechnology, Delhi Technological University. It is a record of original research work carried out by me under the supervision of Prof. B.D. Malhotra Department of Biotechnology, Delhi Technological University, Delhi.

The matter embodied in this project report is original and has not been submitted for the award of any Degree/Diploma.

Date :

Anjali Chaudhary

2k14/BIO/01

Department of Biotechnology
Delhi Technological University
Delhi-110042

ACKNOWLEDGEMENT

I express my deepest sense of gratitude to Dr. B. D. Malhotra, Professor, Department of Biotechnology, Delhi Technological University, who has been a wonderful supervisor throughout the period of my project. I owe all the successful benchmarks of my entire work to his priceless counseling and thorough support.

I acknowledge my work to Dr. Yasha hasija for her continuous encouragement, support and critical suggestions throughout the work.

I would like to give my special thanks to Dr. S. Ramachandran; Senior Principal Scientist, Professor of the AcSIR in the Faculty of Biological Sciences, for providing me the system to work and for his valuable advises.

Lastly, I would like to thank my classmates, as well as my seniors, who have helped me, directly or indirectly, during the course of my project and have aided in making it a success.

Anjali Chaudhary

2K14/BIO/01

CONTENTS

TOPIC	PAGE NO
LIST OF FIGURES	1
LIST OF TABLES	2
LIST OF ABBREVIATIONS	2
1. ABSTRACT	3
2. INTRODUCTION	4-6
3. REVIEW OF LITERATURE	7-12
4. MATERIALS	19-20
5. METHODOLOGY	21-39
6. RESULTS	40-41
7. DISCUSSIONS AND CONCLUSION	42-44
8. FUTURE PERSPECTIVES	45
9. REFERENCES	46-57

LIST OF FIGURES

Figure No.	Figure Name	Page No.
1	The workflow process in a microarray experiment	11
2	Outline of two microarray integration methods	18
3	Window of R studio showing different tools	19
4	Homepage of bioconductor.	20
5	QC plot	24
6	RNA Degradation Plot for 8 arrays (4 arrays of each tumor and control).	25
7	Density plot: - (A) Plot of raw data i.e. before normalization and (B) plot of expression data set after normalization.	26-27
8	Boxplot - (A) Plot of raw data i.e. before normalization and (B) plot of expression data set after normalization.	28
9	Dendrogram showing hierarchical clustering of samples	29
10	MDS plot	32
11	RLE plot	33
12	GeneWise Box plot	34
13	Screenshot of list of differentially expressed probes with log-fold change and p-value	37
14	Heatmap of differentially expressed genes	38
15	List of differentially expressed genes with annotation	39

LIST OF TABLES

Table No.	Table Name	Page No.
1	Output of statistical testing of filtered data	40
2	The annotation of probe I.D. for differentially expressed genes.	41

LIST OF ABBREVIATIONS

S.NO.	ABBREVIATION	FULL FORM
1.	OSCC	Oral squamous cell carcinoma
2	HNSCC	Head and neck squamous cell carcinoma
3.	cDNA	Complementary DNA
4	HPV	Human Papilloma Virus
5	PM	Perfect Match
6	MM	Mis-Match
7	RMA	Robust Multi-Array Average
8	DEG	Differentially Expressed Genes
9	ECM	ExtraCellular Matrix

1. ABSTRACT

Oral squamous cell carcinoma is the sixth most common cancer worldwide. The increasing epidemiological relevance of this cancer emphasizes the need to identify predictive tumor markers. There are limited studies to associate the expression changes in OSCC using clinically relevant variables. Many studies showed inconsistent cancer biomarkers due to bioinformatics artifacts. In this work we use multiple data sets from microarrays in order to improve the reliability of cancer biomarkers. Combining a large number of gene expression datasets originating from different labs could be beneficial for the discovery of new biological insights and could increase the statistical power of gene expression analysis, but then this data should be combined in a consistent manner. We perform a Cross-Platform Normalization method which integrates and cross-annotates multiple data sets related to oral cancer. This Cross-Platform Normalization was done to determine differential gene expression in oral cancer using the open-source R programming environment in conjunction with the open-source Bioconductor software. Cross-Platform Normalization is a powerful tool for analyzing microarray experiments by combining data from multiple studies. Functionalities for combining outputs from different methods and for data transformation are also available in the package. Moderate t-statistics is used to find DEG using Limma package of Bioconductor. In this microarray analysis expression profile of samples were used to identify DEG using 352 samples of which 69 was normal while 283 was tumor. Total 16 genes are found to be differentially expressed, seven genes are found to be upregulated (MMP1, MMP12, CXCL8, SPP1, PTHLH, MMP3 and MMP10), while nine genes are found to be downregulated (ENDOU, MAL, CRNN, SCEL, TGM3, CLCA4, KRT4, CRISP3 and KRT13). All these genes are previously shown to be involved in OSCC and hence, can be used as the potential biomarker to detect oral cancer.

Keywords:

Oral squamous cell carcinoma, microarray, Cross-Platform Normalization, R-package, implementation, visualization

2. INTRODUCTION

Head and neck squamous cell carcinoma (HNSCC) is the fifth most common cancer worldwide [1]. Oral squamous cell carcinoma (OSCC), a common subtype of HNSCC, the sixth most frequent cancer in the world [2]. OSCC is a major cause of morbidity and mortality worldwide, accounting for more than 275,000 new cases and over 120,000 deaths every year [3]. Although there have been improvements in the therapeutic modalities, OSCC-associated morbidity and mortality remain high and have not changed in over three decades [4].

In developing countries including India, controlling the devastating, widespread consequences of oral cancer requires interventions in persons at-risk ideally before the disease becomes invasive but certainly before it becomes locally advanced or metastatic [5]. Once the neoplastic process sets in, it is rather difficult to control and endangers the life of the host. Therefore, detection of a malignancy before it arises would be the best possible mode of preventing the dreaded disease in its earliest form or by intervening before it reaches uncontrollable proportions. Advances in the analysis of molecular alterations in cells undergoing malignant transformation have increasingly revealed the mechanisms that lead to the occurrence and progression of malignancies [6].

Malignant cells have different histologic and biochemical behavior as compared to their normal counterparts. Earlier the most common determinant or marker of carcinomatous transformation in a tissue was the histopathologic presence or absence of epithelial dysplasia. However, the expanding field of oncology has revealed new and more specific markers that would help to determine the degree of cell alteration and enable a better understanding of the degree of malignant transformation of these cells. Data obtained from clinical examination and routine histopathologic studies are not always accurate about the potential or risk (to varying degrees) of the lesion in question becoming malignant [7].

In recent years, there have been a number of approaches to the problem of precancerous tissue with the aim to establish a more fundamental biochemical basis of understanding [8]. Several abnormal cellular products are synthesized by the neoplastic cells and also by the body in the presence of such an abnormal situation. Such cellular products can be detected in the various body fluids and on the surface of the cancer cells either by biochemical methods or by immunochemistry. These products that are detected and measured are known as 'tumor markers'.

Also, the gene which shows differential expression apart from their normal/baseline expression in particular tissue and leads to tumor are also referred as “tumor marker”. These tumor markers can be effectively made use of for early screening and detection of cancer. Diagnosis can be aided by the use of these and clinical staging can be better applied in the light of the revelations by these markers.

A number of published gene signatures validated using independent samples have been shown to serve as significant predictors of clinical outcome [9–13]. However, the development of prognostic signatures that are robust and stable (e.g., the same biomarkers are identified in both discovery and validation sets) [14] has proven challenging [15–17].

Published prognostic gene signatures derived from internal validation often show little overlap with genes identified by other study groups [12]. Potential causes of small reproducibility include differences in sample collection methods, processing protocols, and microarray platforms, patient heterogeneity, and small sample sizes [27]. Due to the difficulty of acquiring samples, particularly from human tissue and the associated costs, microarray experiments from single-institution patient cohorts are often composed of small sample sizes. Predictive models trained on the gene signatures identified from these smaller-sized individual studies are less robust [12, 17]. Michiels et al. [18] re-analyzed data from nine studies predicting cancer prognosis and found an unstable misclassification rate for the gene signature (defined as the 50 genes for which expression was most highly correlated with outcome) using training sets derived using a re-sampling approach, with performance increasing as the size of the training set increases.

Integration of multiple microarray data sets has been advocated to improve gene signature selection [19]. Increasing sample sizes increases the statistical power to obtain a more precise estimate of integration of (differential) gene expression and to assess the heterogeneity of the overall estimate, as well as to reduce the effects of individual study-specific biases [20–23]. Meta-analysis is most commonly applied for the purpose of detecting differentially expressed (DE) genes [24] which may serve as a candidate gene signature or be used as features in classification models or classifiers to further refine a clinically useful gene signature [25]. Supervised classification techniques (also known as prediction analysis or supervised machine learning) are the most commonly used methods in microarray analysis that lead to the

identification of clinically useful biomarkers (i.e., gene signatures providing improved discrimination between two or more patient groups) [24]. Classification methods for gene signature selection are beyond the scope of this work and have been reviewed elsewhere [26].

3. REVIEW OF LITERATURE

3.1. Oral squamous cell carcinoma (OSCC)

Oral squamous cell carcinomas (OSCC) are cancers originating from the squamous epithelium in the oral cavity. Locations include the lip, mobile tongue, and buccal mucosa, floor of the mouth, gingiva, hard palate and soft palate. OSCC belongs to a larger subgroup of tumors termed head and neck squamous cell carcinomas (HNSCC), comprising of carcinomas arising in the oral cavity, oropharynx, larynx, hypopharynx, nasal cavity, nasopharynx, salivary glands and the ear [92], where OSCCs are the most common oral malignancy with a poor 5-year survival rate [92,28-30].

3.1.1. Epidemiology and Etiological factors

In 2008, more than 260.000 new cases of oral cavity cancers were predicted worldwide and over 130.000 of these patients were estimated to die from the disease (approximately 50%). More than 60% of these cases occur in the developing countries, where the male population by far displays the highest prevalence [29]. Gender, race, and age have all been associated with differences in OSCC incidence, mortality, site, grade, histological type and tumor stage at diagnosis [31]. As with many other types of cancer, OSCC most commonly occurs in the middle-aged and elderly population [32, 33]. The male population has traditionally had a higher incidence in OSCC, typically 1:2 compared to women [33]. In 2001, the highest mortality rates for OSCC were reported to be in France, the Indian subcontinent, Brazil and central/eastern Europe [32]. The lowest survival rates have been ascribed patients of African-American origin living in the United States [31]. Also among South-African Indians, living in Natal, the mortality rates from OSCC were high [32-34]. Most often, such differences in mortality rates are explained by cultural traditions, ethnic differences and socioeconomic circumstances [32]. Certain risk factors such as tobacco use, alcohol consumption, and human papillomavirus (HPV) infections, increases the HNSCC incidence [29, 32]. Furthermore, heavy consumption of alcohol combined with smoking functions synergistically, multiplying the risk of developing OSCC [29, 32, and 35]. A high percentage of oropharyngeal cancers are HPV positive (90% in Sweden, 60% in the USA), and HPV is thought to be a major cause of cancers in the oropharynx [36], though far less important

for the development of cancers in the oral cavity. Other risk factors believed to have an impact on the development of OSCC are poor oral hygiene, gastro-esophageal reflux disease, dietary factors, use of marijuana and environmental contaminants such as paint fumes, plastic by-products, and gasoline fumes [37].

3.1.2. Stage, Histopathology and Grade of Primary Tumor

Important prognostic indicators that are known to affect regional metastasis and therefore outcome, include the size of the primary tumor, site, T stage, and grade, depth of invasion, biological tumor markers, perineural invasion and patient compliance [38-39].

The TNM classification of oral squamous cell carcinoma [40] provides a reliable basis for patient prognosis and therapeutic planning. There are a number of clinically detected or small undetectable primary tumors that display biological aggressiveness, with early regional metastasis and death. Typically, T1-T2 lesions are often associated with a risk of regional metastasis of 10% to 30% respectively, especially to lymph nodes, whereas several studies have shown a clear correlation between increasing tumor thickness and an increased risk of cervical metastasis [41-42]; T3-T4 lesions have a significantly higher risk of regional neck disease [37, 43].

3.1.3. Diagnosis

Efforts have been made to elucidate tumor-related factors that could influence the appearance of metastases in oral squamous cell carcinoma [44]. The samples are studied using hematoxylin and eosin staining and reviewed according to World Health Organization histological criteria [40]. Although a number of studies have investigated the potential of these biomarkers in oral squamous cell carcinoma, there is no agreement on a reliable predictor of prognosis. Various histopathological parameters, keratinization, mode of invasion, and lymphocyte infiltration have been described as being predictors of lymph node metastasis [45]. Tumor thickness is an important prognostic factor in carcinomas of the oral cavity. The treatment of tumors smaller than 3 mm might need to be less aggressive than if the tumor is larger than 5 mm [47].

The use of biomarkers could help to avoid the unnecessary surgical treatment of metastasis-free patients [46]. Although the TMN staging system is used routinely, the technique accurately

determines only the size and location of the tumor and does not predict their metastatic potential. The further clinical examination can only identify regional metastasis with an accuracy of 70%. Although the use of various forms of imaging can improve this percentage, the microscopic disease cannot be detected by these methods [46].

Several proteins and genes are candidates for use as predictors of metastasis due to the heterogeneity of the cells [46, 48]. Some studies have tried to relate the expression of proteins in primary tumors with the occurrence of metastasis, for this purpose, many key proteins have been searched with the intention of establishing more reliable prognostic factors of OSCC.

3.1.4. Early detection and its importance

Oral cancer patients commonly seek treatment at the advanced stage of the disease, thereby diminishing the chances of therapeutic success [49-50]. With advances in research, clinical outcomes have improved due to facilitation of early detection of lesions. Visual and cytological techniques that are routinely used to detect OSCC exhibit limited predictabilities. Recently, various staining methods have been used to evaluate oral dysplasias. Each of these methods, however, has its limitations. Nearly thirty percent of oral cancers do not arise from premalignant lesions [51], and histologically normal oral epithelium has been observed to develop into tumors in many instances [52]. Conversely, only a small proportion of premalignant tissues develop into cancer. The use of a variety of molecular and biochemical techniques provides abundant information regarding preneoplastic changes of the oral cavity in a laboratory setting. The practical applications of these strategies, however, remain to be evaluated at large scale in diagnosis at the clinic. Thus, the application of these biochemical and molecular methods for more accurate detection of potential oral lesions is of great importance.

3.2. Microarray

Microarray technology is a powerful tool for simultaneously evaluating the expression level of thousands of genes in a cell [53] and, hence, the information that is encoded in the DNA [54]. A microarray is a microscopic slide that contains an ordered series of DNA, RNA proteins or tissues. The DNA microarrays are the most common [55]. A DNA microarray is generally a glass slide or a silicon chip in which thousands of gene sequences are printed. One very spot many copies of a specified DNA sequence are chemically bonded to the surface of the slide [53].

The genes immobilized onto the slide are called the DNA probe. Over this DNA probe, the target DNA or the target RNA (depending on the microarray platform) obtained from the cell under study is hybridized (hydrogen bonded). The amount of hybridization is measured and related to the presence and expression of certain genes in the cell. [55-58].

Microarrays are microscope slides that contain an ordered series of samples (DNA, RNA, protein, tissue). The type of microarray depends on upon the material placed onto the slide: DNA, DNA microarray; RNA, RNA microarray; protein, protein microarray; tissue, tissue microarray. Since the samples are arranged in an ordered fashion, data obtained from the microarray can be traced back to any of the samples. This means that genes on the microarray are addressable. The number of ordered samples on a microarray can number into the hundreds of thousands [59]. The typical microarray contains several thousands of addressable genes. The most commonly used microarray is the DNA microarray. The DNA printed or spotted onto the slides can be chemically synthesized long oligonucleotides or enzymatically generated PCR products. The slides contain chemically reactive groups (typically aldehydes or primary amines) that help to stabilize the DNA onto the slide, either by covalent bonds or electrostatic interactions. An alternative technology allows the DNA to be synthesized directly onto the slide itself by a photolithographic process. This process has been commercialized and is widely available. By orderly arranging samples, the microarray provides a large-scale medium for matching known and unknown DNA segments based on base-pairing rules.

Figure: 1 shows the workflow process in a microarray experiment. The experimental process varies depending on the microarray platform that is used.

3.2.1. Types of microarrays:

Microarrays can be broadly classified according to at least three criteria:

- i. The length of the probes: arrays can be classified into “cDNA) arrays” which use long probes of hundreds of base pairs, and “oligonucleotide arrays,” which use short probes (50 bps or less).
- ii. Manufacturing methods include: “deposition” of previously synthesized sequences and “in-situ synthesis.” Usually, cDNA arrays are manufactured using deposition, while oligonucleotide arrays are manufactured using in-situ technologies. In-situ technologies

include: “photolithography” (eg, Affymetrix, Santa Clara, CA), “ink-jet printing” (eg, Agilent, Palo Alto, CA), and “electrochemical synthesis” (eg, Combimatrix, Mukilteo, WA).

- iii. The number of samples: “Single-channel arrays” (Affymetrix GeneChip) analyze a single sample at a time, whereas “multiple-channel arrays” can analyze two or more samples simultaneously.

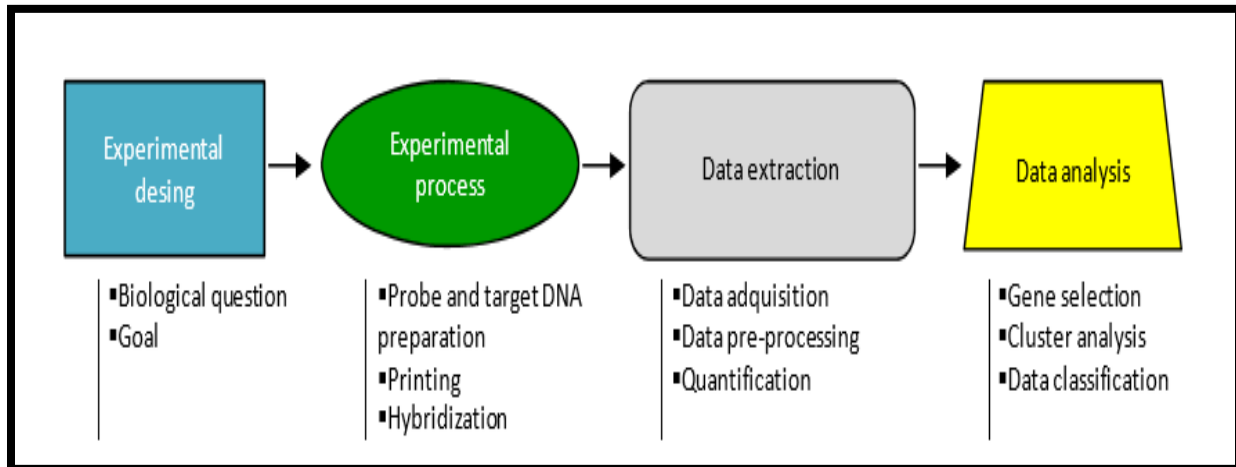


Figure 1: The workflow process in a microarray experiment [91].

The cDNA arrays apply to glass slides (or nylon membranes) spots of complimentary DNAs (cDNAs), which are generated in biological labs by reverse transcription (so that they only include the protein-coding part of the genome) [60].

The oligonucleotide arrays (often referred to as the Affymetrix arrays) place many thousands of gene-specific oligonucleotides (called probes) synthesized directly on a silicon chip. The probes are about 25 base pairs long, and 20 probe-pairs (one perfect fact and one mismatch) are often used to represent each gene (like a 20 digit barcode). In order to compare two types of cells (e.g., a cancer cell versus a normal cell), for example, the biologist first extracts the DNA materials from all the cells and labels those from one cell type (say, cancer cell) by fluorescence cy5 (red) and the other cell type by cy3 (green). The microarray is then exposed to the mixture of the two DNA samples for hybridization. When mRNA for a gene is more abundant in the cancer cell than in the normal cell, for example, the array spot corresponding to that gene will show a red color. Numerically, a vector of length G is reported, where G is the number of spots (genes) in

the array, and each entry of the vector records the ratios of the fluorescence intensities (cy5/cy3). Where each column corresponds to a cell type (e.g., lymphoma cell, leukemia cell, normal cell, etc.) or a treatment and each row corresponds to a gene [61]. Thus, through the use of DNA microarrays, one can monitor simultaneously the expression levels of thousands of genes in different types of cells.

DNA microarrays can be used to determine:

- The expression levels of genes in a sample, commonly termed expression profiling.
- The sequence of genes in a sample commonly termed mini-sequencing for short nucleotide reads, and mutation or SNP analysis for single nucleotide reads.

3.2.2. Cancer application

Cancer is a family of primarily genetic diseases in which altered gene expression is the main molecular characteristic. Therapeutic efficacy and clinical prognosis are highly variable even for cancers that are classified under the same category based on symptomatic and conventional diagnosis. Currently, in an effort to account for their variable clinical behavior, apparently similar cancer classes are being evaluated for subtle differences in gene expression. Microarray technology has facilitated unforeseen advances with this approach; in fact, it has been proposed that genomic profiles can serve as a diagnostic tool, and microarray technology has even led to some refinement in tumor classification. The types and numbers of applications for microarray experiments are quite variable and constantly increasing. Microarrays used to monitor the expression level of genes in the comparison between two conditions remains one of the most widespread uses of microarrays.

One of the most exciting areas of application is the diagnosis of clinically relevant diseases. The oncology field has been especially active and to an extent successful in using microarrays to differentiate between cancer cell types. The ability to identify cancer cells based on gene expression represents a novel methodology that has real benefits. In difficult cases where a morphological or an antigen marker is not available or reliable enough to distinguish cancer cell types, gene expression profiling using microarrays can be extremely valuable. Programs to predict clinical outcome and to design individual therapies based on expression profiling results are well underway. However, a major advantage of the microarray is the huge amount of

molecular information that can be extracted and integrated to find common patterns within a group of samples. As we will show here, microarrays could be used in combination with other diagnostic methods to add more information about the tumor specimen by looking at thousands of genes concurrently. This new method is revolutionizing cancer diagnostics because it not only classifies tumor samples into known and new taxonomic categories, and discovers new diagnostic and therapeutic markers, but it also identifies new subtypes that correlate with treatment outcome.

3.3. Bioconductor

Bioconductor is an open source and open development software project to provide tools for the analysis and comprehension of genomic data. Bioconductor is built on Open Source Platform, R programming language, but does contain contributions in other programming languages. Most Bioconductor components are distributed as R packages. Technically R is an expression language with a very simple syntax. It is case sensitive. R provides facilities like data manipulation, calculation, and graphical display. The Bioconductor project was started in the fall of 2001 and is overseen by the Bioconductor core team, based primarily at the Fred Hutchinson cancer Research Center with other members coming from the various US and international institutions [62]. The main goal of the Bioconductor project is the creation of a durable and flexible software development and deployment environment that meets these new conceptual, computational and inferential challenges [63] Other goals of Bioconductor is to provide widespread access to a broad range of powerful statistical and graphical methods for the analysis of genomic data. To provide a common software platform that enables the rapid development and deployment of extensible, scalable, and interoperable software. To further scientific understanding by producing high-quality documentation and reproducible research. To train researchers on computational and statistical methods for the analysis of genomic data [64].

3.3.1. Features of the Bioconductor

i) The R Project for Statistical Computing

Bioconductor is built on R language. So R provides a broad range of advantages to the Bioconductor. Some advantages of R are as below.

- It contains a high-level interpreted language in which one can easily and quickly prototype new computational methods.
- It includes a well-established system for packaging together software components and documentation.
- It can address the diversity and complexity of computational biology and bioinformatics problems in a common object-oriented framework.
- It provides on-line computational biology and bioinformatics data sources.
- It supports a rich set of statistical simulation and modeling activities. It contains cutting edge data and model visualization capabilities.
- It has been the basis for pathbreaking research in parallel statistical computing.
- It is under very active development by a dedicated team of researchers with a strong commitment to good documentation and software design [62].

ii) Documentation and reproducible research

Each Bioconductor package contains at least one vignette, which is a document that provides a textual, task-oriented description of the package's functionality. These vignettes come in several forms. Many are simple "How-to" that is designed to demonstrate how a particular task can be accomplished with that package's software. Others provide a more thorough overview of the package or might even discuss general issues related to the package [62].

iii) Statistical and graphical methods

The Bioconductor project aims to provide access to a wide range of powerful statistical and graphical methods for the analysis of genomic data. Analysis packages are available for preprocessing array data, identifying differentially expressed genes; graphical analyses, plotting genomic data. In addition, the R package system itself provides implementations for a broad range of state-of-the-art statistical and graphical techniques, including cluster analysis, resampling, etc [62].

iv) Open source

The Bioconductor project has a commitment to fully open source discipline, with distribution via a Source Forge-like platform. All contributions are expected to exist under an open source

license such as Artistic 2.0, GPL2, or BSD. There are many different reasons why open-source software is beneficial to the analysis of microarray data and to computational biology in general [62]. Reasons for deciding to release software under an open-source license are as follows

- To encourage reproducibility, extension and general adherence to the scientific method
- To ensure that the code is open to public scrutiny and comment
- To provide full access to algorithms and their implementation
- To provide to users the ability to fix bugs without waiting for the developer, and to extend and improve the supplied software
- To encourage good scientific computing and statistical practice by exhibiting fully appropriate tools and instruction
- To provide a workbench of tools that allow researchers to explore and expand the methods used to analyze biological data
- To ensure that the international scientific community is the owner of the software tools needed to carry out research
- To promote reproducible research by providing open and accessible tools with which to carry out that research [63]

Bioconductor is consisting of many packages, BioC 2.5, latest version consisting of 352 packages and designed to work with R 2.10.z, was released in 2009/10/26.

3.4. Cross-Platform Normalization

Cross-platform normalization (also termed “data merging”) [65] considers all data from experiments across different microarray platforms as a single data set from the same experiment. Direct integration of data sets performed on different microarray platforms may introduce undesirable batch effects due to systematic multiplicative biases [65, 67 and 71]. The level of difficulty present to combine multiple datasets has been termed “dataset complexity” [70]. For example, integrating different Affymetrix platforms is less complex to analyze by meta-analysis or cross-platform normalization than datasets performed across very different platforms. Studies using low complexity datasets, mainly from the Affymetrix platform, have directly merged the studies to construct a gene signature [68, 72-74].

Cross-platform transformation and normalization methods have been developed with an aim to remove the artifactual differences between data from different microarray platforms while preserving the underlying biological differences between conditions. This step is essential, as non-biological differences (“batch effects”) in the gene signature discovery data can obscure real biological differences found between clinical groups. Early attempts at cross-platform merging applied straightforward transformation methods of location and scale (mean and variance) to process the gene expression data from different studies. Batch mean centering [50] is a simple transformative method that standardizes the expression of each gene to have the same center (mean expression). Probe sets can be further transformed to have the same variance or distributions on different platforms [5, 76]. While these methods are relatively easy and intuitive, the batch mean centering method has been shown to have only marginal improvement compared to uncorrected data for cross-platform integration of Illumina and Affymetrix data [67]. The probability of expression (POE), a model-based transformation that is estimated based on a method that adopts an underlying mixture distribution that transforms each data value into the range $[-1,1]$ has been used for cross-platform merging based on a unified scale as an alternative to using gene-specific summaries [77-78]. While this transformation has been applied for identifying meta-signatures, it has been found to be difficult to compare to other normalization methods [66]. Over the past decade, a number of more complex cross-platform normalization methods have been published and their performance has been compared in several studies [64, 67]. Four cross-platform normalization methods found to be generally effective in a comparative review by Rudy and Palafer [79] are:-

3.4.1. Empirical Bayes (EB) method, known as Combat [80],

3.4.2. Cross-Platform Normalization (XPN) method [66],

3.4.3. Distance Weighted Discrimination (DWD) [81],

3.4.4. Gene Quantiles (GQ) method developed as part of the WebArrayDB service [82].

Combat have been found to perform well in previous analysis, the user must be cautious when applying this method to data sets that are unbalanced (e.g., different subtypes within each of the batches) as these methods will not be able to distinguish batch effects from biologically relevant signals [69].

3.5. Comparison of Meta-Analysis vs. Cross-Platform Normalization

Directly-merged microarray data (or applying cross-platform normalization) has been argued to have better performance than meta-analysis for the identification of robust biomarkers on the premise that “deriving separate statistics and then averaging is often less powerful than directly computing statistics from aggregated data” [87]. In a comparative study, Taminau et al. [85] found significantly more differentially-expressed genes using cross-platform normalization than meta-analysis. An additional advantage of cross-platform normalization is that it allows prediction models applied to a subset of studies to be applied across additional studies from other platforms [86]. While cross-platform normalization has been applied in multiple studies [88-90], it has less frequently been used in the literature compared to meta-analysis [79]. A recent comprehensive systematic literature review of studies applying microarray integration methods found that only 27% of the studies directly merged microarray data and this subset of studies were mostly performed on the same platform [86]. One major limitation of existing cross-platform normalization is that they require that every treatment group or sample type be represented on each platform to allow differentiation of treatment effects from platform effects. Furthermore, cross-platform normalization methods do not guarantee the elimination of laboratory or batch effects across experiments and Rung and Brazma [84] have argued that microarray meta-analysis provides better control of between-laboratory heterogeneity, which can be estimated using Cochran’s Q statistic and be correspondingly adjusted.

Gene signature discovery for prognostic and diagnostic purposes is improved with the knowledgeable selection and appropriate application of integration methods on microarray data performed on multiple platforms. While no consensus for the best implementation of cross-platform integration is currently available, previous benchmarking and comparative analyses have established the strengths and limitations of many of the existing methods. The recent evidence suggesting improved performance of cross-platform normalization methods over meta-analysis may lead to an increasing proportion of studies in the literature implementing the former method. Further refinement of existing methods and development of new methods for cross-platform normalization and classification to exploit the vast quantity of microarray data currently available are expected. As elimination of platform-specific bias becomes well-established with

these methods, future studies addressing the performance of prognostic signature discovery in light of the existing biological heterogeneity will become a central focus.

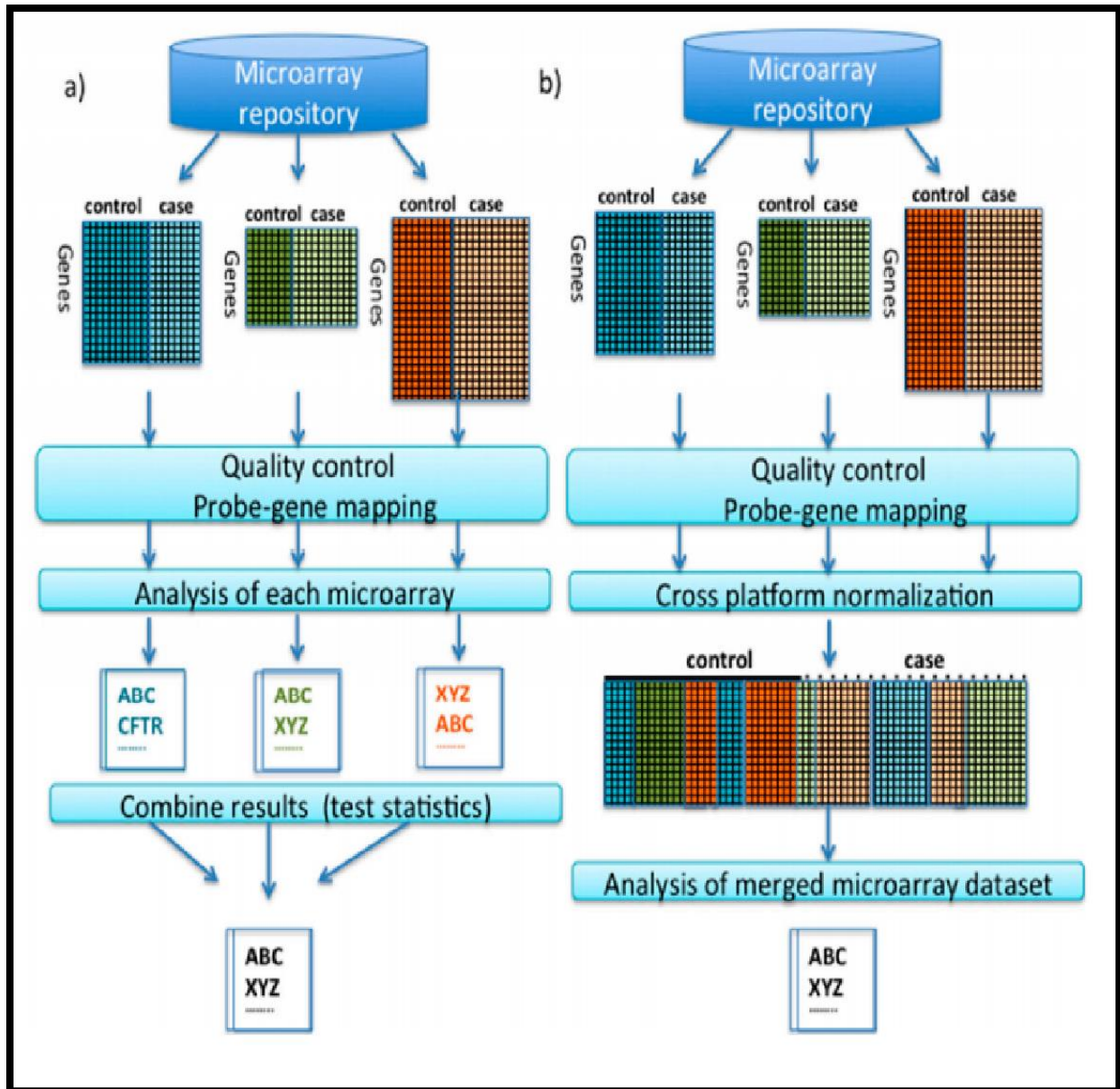


Figure 2: Outline of two microarray integration methods: (a) meta-analysis (“late integration”). Individual case-cohort microarray studies are pre-processed and each study is used to identify ranked gene lists which are then combined in the final step; (b) Cross-platform merging and normalization (“early integration”). After pre-processing of individual studies, a single unified case-cohort dataset is generated (“clustered” into cases and cohorts, indicating removal of batch to batch variation) [83].

4. MATERIALS

4.1. R 3.2.3

R is a free software environment (publically available) used for graphics and statistical computing. It runs and compiles on variety of OS (UNIX platforms, MacOS and Windows. Before downloading R, preferred CRAN mirror was chosen (i.e. INDIA). R software is available at <https://www.r-project.org/> .

4.2. RStudio

It is a set of integrated tools designed for the user which enable them to be more productive with R. It includes an editor for syntax highlighting (supports direct code execution), tools for plotting, history, a console, as well as provides debugging and workspace management.

Codes to run RStudio server on LINUX.

```
$ wget https://download2.rstudio.org/rstudio-server-rhel-0.99.902-x86_64.rpm
```

```
$ sudo yum install --nogpgcheck rstudio-server-rhel-0.99.902-x86_64.rpm
```

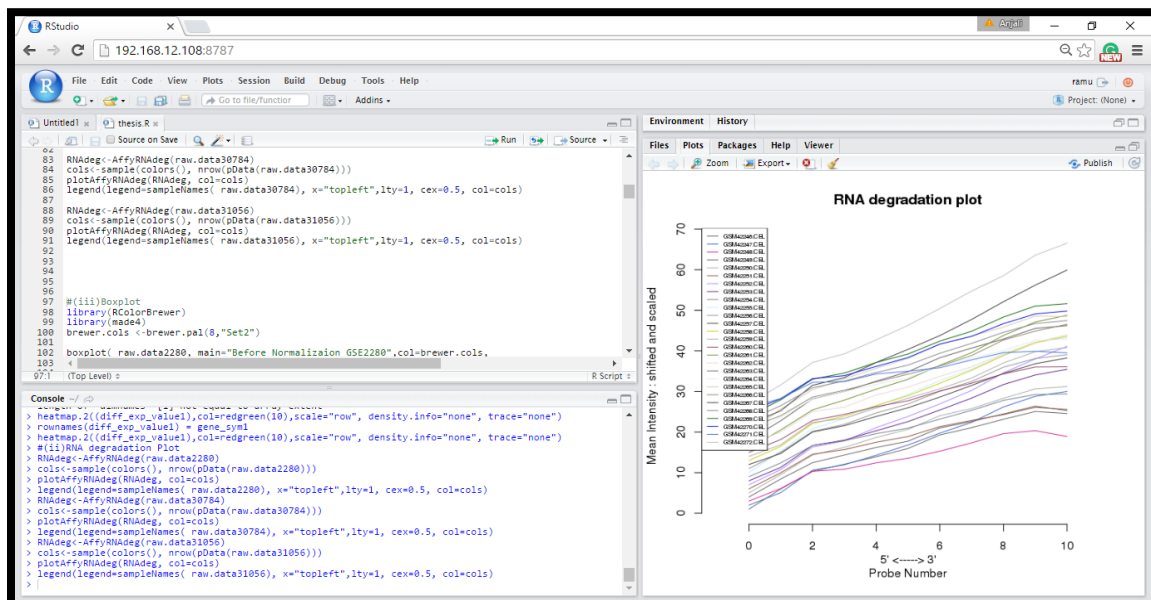


Figure 3: Window of R studio showing different tools. Upper left corner is provided with the tools for writing the source codes, and there is a console at the right bottom of the window, left upper corner is for history and the left end corner is for plots, packages, help etc.

4.3. Bioconductor

Open development and open source software; provides tools for the comprehension and analysis of high-throughput genomic data. This software uses the R language (statistical programming).

Working directory (the path where all data was stored and fetched) was set.

After installing the latest release of R Studio (version-3.2.2), the latest version of Bioconductor was also installed by using R Studio.

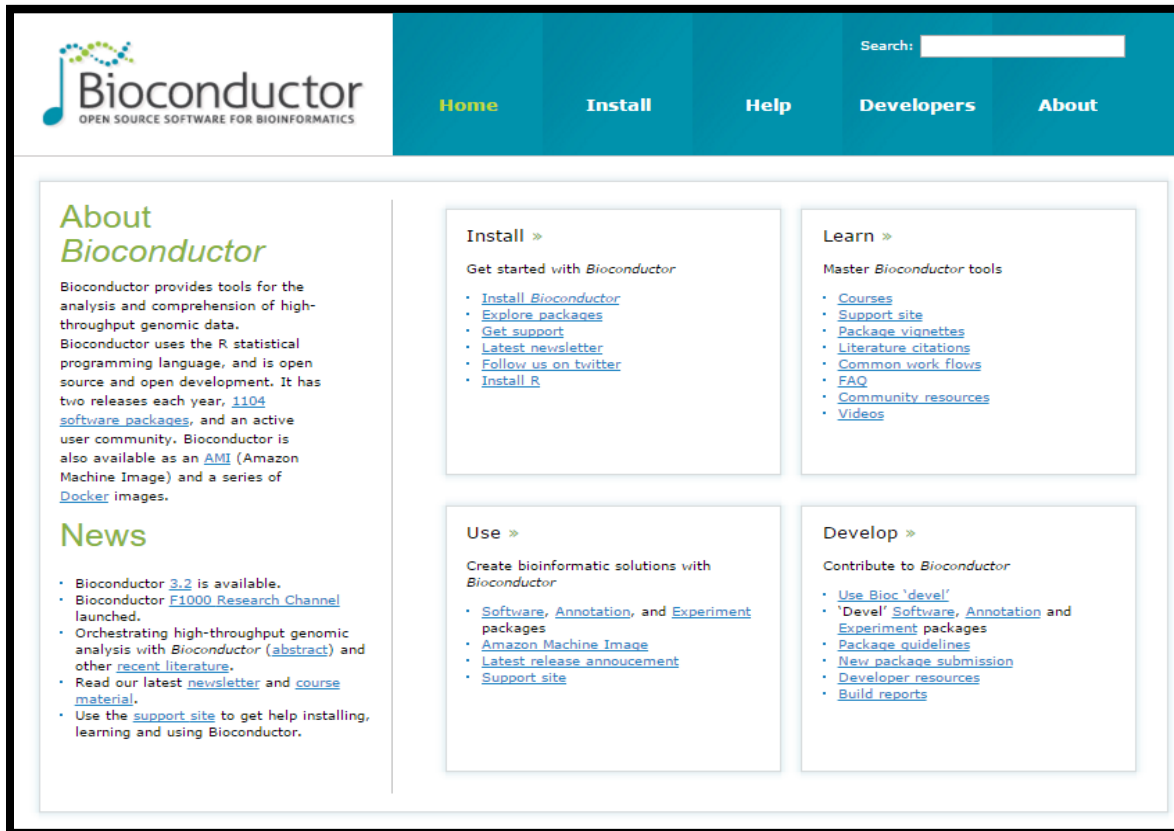
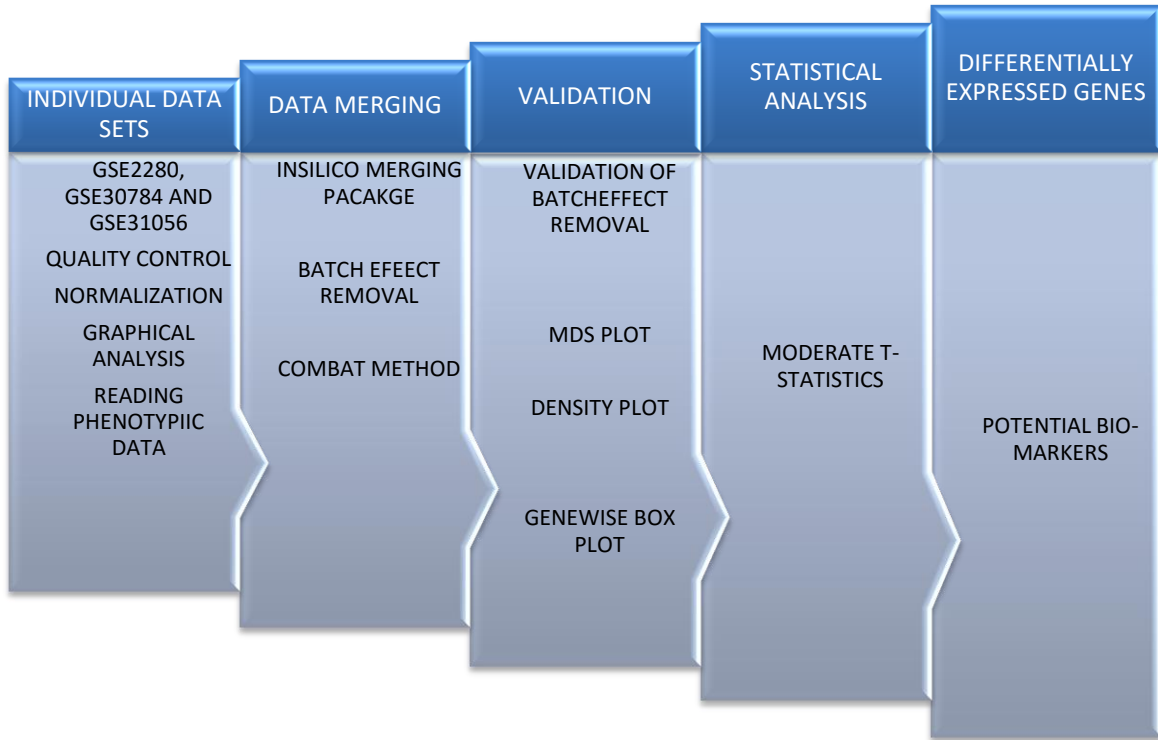


Figure 4: Homepage of biconductor. It consists of packages for various analyses with instructions of loading and using that package.

4.3. System Requirement: RStudio Server v0.99 requires RedHat or CentOS version 5.4 (or higher) as well as an installation of R. You can install R for RedHat and CentOS using the instructions on CRAN: <https://cran.rstudio.com/bin/linux/redhat/README>.

METHODOLOGY

WORKFLOW



5.1. Data collection

The microarray data (supplementary file i.e. raw data) was retrieved from GEO database (NCBI).

5.1.1. Accession number: GSE2280 (27 samples), GSE30784 (229 samples) and GSE31056 (96samples).

5.1.2. Platform: GPL96; Affymetrix Human Genome U133A Array [HG-U133A], GPL570; Affymetrix Human Genome U133 Plus 2.0 Array [HG-U133_Plus_2] and GPL10526; Affymetrix GeneChip Human Genome HG-U133 Plus 2 Array [HG-U133_Plus_2, Brainarray Version 12].

5.1.3. Experiment type: Expression profiling by array

5.1.4. Organism: *Homo sapiens*

5.1.5. Supplementary file: GSE_RAW.tar

5.1.6. Phenotypic data: Define sample or label sample with its characteristics.

5.2. Loading and Extracting Data

Package requires extracting the data from supplementary files (GSE2280_RAW.tar) information is “GEOquery”.

- A. GEOquery package was downloaded from Bioconductor- biocLite ("GEOquery").
- B. GEOquery library was loaded- library (GEOquery).
- C. The raw data (supplementary files) was also downloaded directly by R.
- D. Data was extracted from TAR zip files to a new folder
- E. .CEL files were listed
- F. Pasting .CEL files in new folder

OR

Download manually to working directory of R Studio from GEO database.

5.3. Reading Affymetrix data

#Reading data
> library(simpleaffy)
> setwd("/home/.....")
> raw.data2280 <- read.affy(covdesc="GSE2280.txt", path = "/home/ramu/project/GSE2280/data2280")
#info in raw.data2280: AffyBatch object size of arrays=712x712 features (19 kb) cdf=HG-U133A (22283 affyids) number of samples=27 number of genes=22283 annotation=hgu133a

5.4. Removal of extended, diffuse and compact blemishes

“Harshlight” package is used to detect extended, diffuse and compact blemishes on microarray chips. Harshlight automatically marks the areas in a collection of chips (affybatch objects) and a corrected AffyBatch object is returned, in which the defected areas are substituted with NAs or the median of the values of the same probe in the other chips in the collection. The new version handles the substitute value as the whole matrix to solve the memory problem.

```
> abatch.harshlight <- Harshlight(affy.object= raw.data2280, my.ErrorImage = NULL,
extended.radius = 10, compact.pval = 0.01, diffuse.bright = 40, diffuse.dark = 35, diffuse.pval =
0.001, diffuse.connect = 8,percent.contiguity = 50, report.name = 'R.report.ps', na.sub = FALSE)
```

5.5. Data Normalization

Normalization was done using robust multi-array average (RMA) method, which transformed the raw data (having probe intensity value) into expression value of each gene.

5.5.1. This involves three steps

5.5.1.1. Background adjustment: It reduces noise and observed intensities require adjustment for accurate measurements of specific hybridization.

5.5.1.2. Normalization: Without this, it is not possible to compare measurements of hybridizations from different array due to many obscuring sources of variation i.e. different efficiencies of transcription (reverse), labeling, reagent batch effects, physical problems of the arrays and laboratory conditions.

5.5.1.3. Summarization: It is needed because all transcripts are represented by multiple probes. For each gene, the normalized probe intensities and background adjusted, need to be summarized into expression set with one value.

5.5.2. Extracting expression values

For Affymetrix data, the expression values are already log₂-transformed which could be extracted as follow-

```
> eset2280<-rma(raw.data2280)
>Eset2280=exprs(eset2280, normalize=T)
> write.exprs(Eset,file="Eset2280.xls")
```

5.6. Quality control

Quality control of Affymetrix uses simple graphical exploration methods for quality assessment, before and after the normalization. This is performed for raw data (.CEL files).

5.6.1. Examining the expression

This is used to examine the expression of the control samples (control genes), see figure: 5.

```
# OC plot
> library(simpleaffy)
> aqc<-qc(raw.data2280)
> plot(aqc)
```

GSE2280

GSE30784

GSE31056

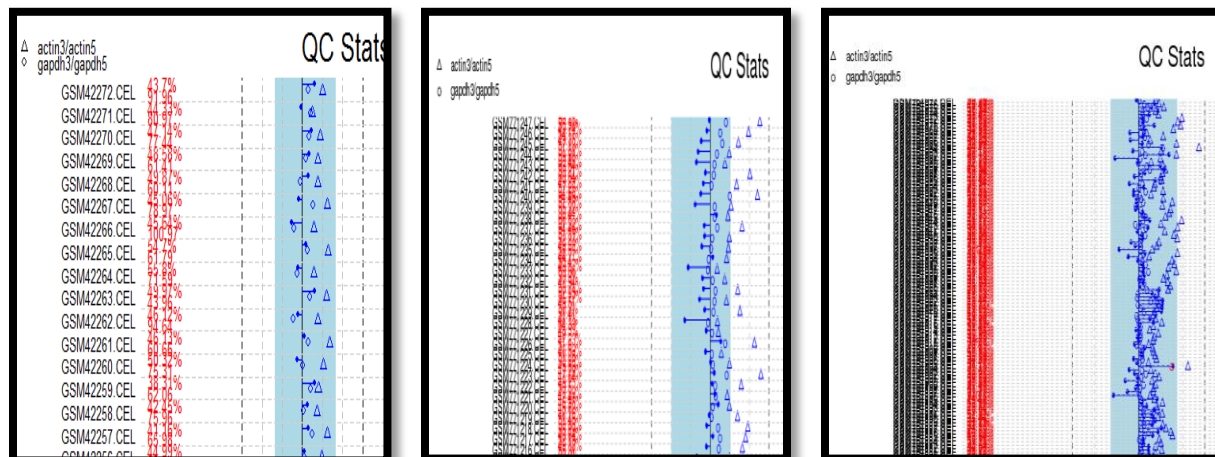


Figure 5: QC plot- Different chips are separated by vertical grey lines, the red numbers on the left report the number of probesets , and the average background on the chip. The blue region in the middle denotes the area where scaling factors are less than 3-fold of the mean scale factors of all chips. Bars that end with a point denote scaling factors for the chips.

Here, all the scaling factors lie within acceptable range. However, all of the chips which are showing very deviant control gene expression are samples of cancerous tissue. It represents the large gene expression changes displayed by the cancer tissue.

5.6.2. RNA degradation plot

This is to check wheatear there are big differences or no differences or slight difference in RNA degradation between different sample arrays. The slope of the lines which shows the amount of degradation is not that important, but if one or more lines showing very different slopes than the others, then there are expression manifestation.

```
# RNA degradation Plot
> RNAdeg<-AffyRNAdeg(raw.data2280)
> cols<-sample(colors(), nrow(pData(raw.data2280)))
> plotAffyRNAdeg(RNAdeg, col=cols)
> legend(legend=sampleNames(raw.data2280), x="topleft",lty=1, cex=0.5, col=cols)
```

There is no clear guideline of knowing how large a slope (degradation) must be to decide a bad chip. The slope of samples appears to be reasonably parallel (figure: 6).

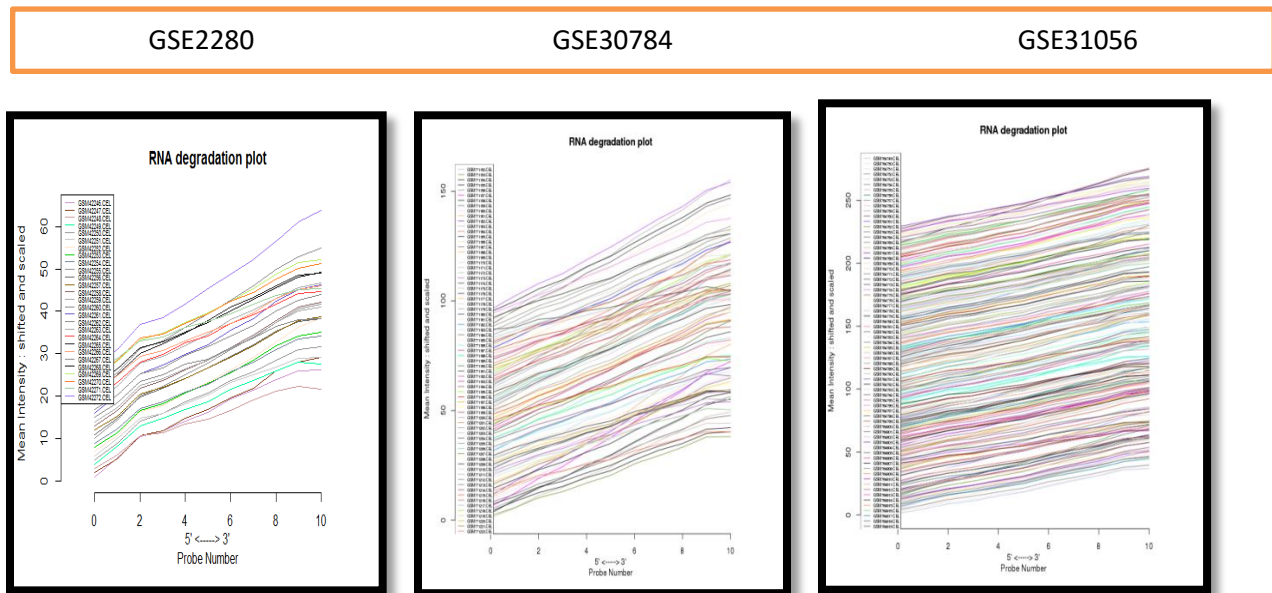


Figure 6: RNA Degradation Plots.

5.6.3. Density plot

For better comparison density plots of the distribution of all arrays are superposed in a single graph, which allows identification of arrays having weird distribution. The distributions of raw PM log-intensities must not be identical but still totally different. The distributions of normalized log-intensities are identical, normalization makes the distributions be even. This density plot allows the checking of normalization step.

```
# Density plot before normalization
```

```
> hist(raw.data2280,col=brewer.cols,lty=1,  
+ xlab='Log(base2)  
+ intensities',lwd=2, main="Before Normalization")
```

```
> legend(legend=sampleNames(raw.data2280), x="topright",  
+ lty=1, cex=0.5, col=brewer.cols, lwd=2)
```

```
# Density plot after normalization
```

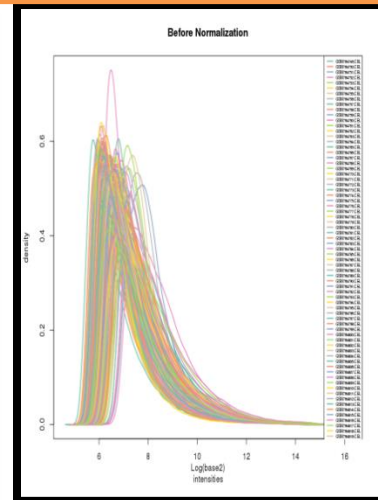
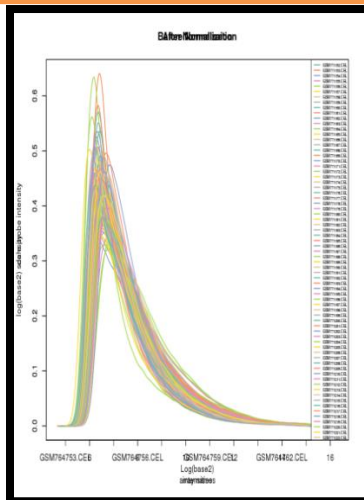
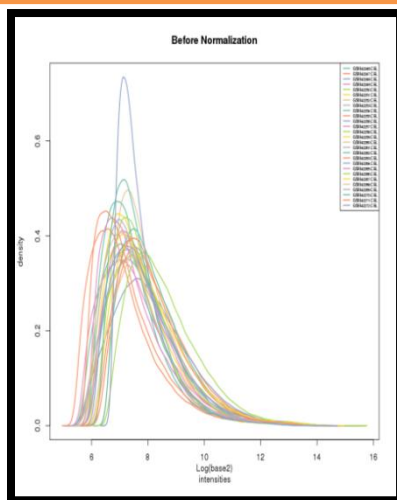
```
> plotDensity(Eset,  
+ col=brewer.cols,lty=1,  
+ xlab='Log(base2)  
+ intensities',lwd=2, main="After Normalization")
```

```
> legend(legend=sampleNames(raw.data2280), x="topright",  
+ lty=1, cex=0.5, col=brewer.cols, lwd=2)
```

[A] GSE2280

GSE30784

GSE31056



[B] GSE2280

GSE30784

GSE31056

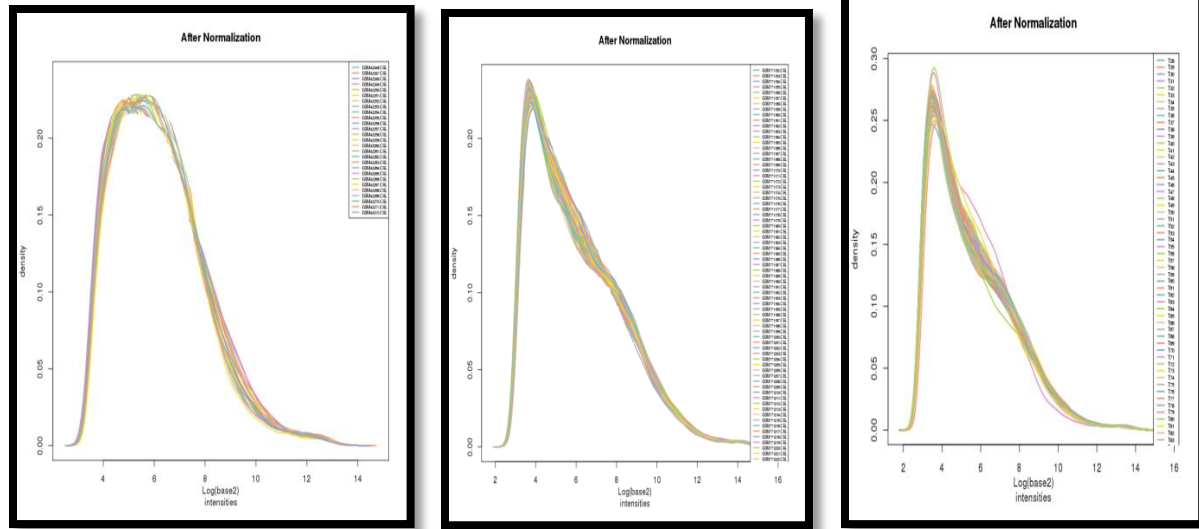


Figure 7: Density plot: - (A) Plot of raw data i.e. before normalization and (B) plot of expression data set after normalization.

5.6.4. Box plot

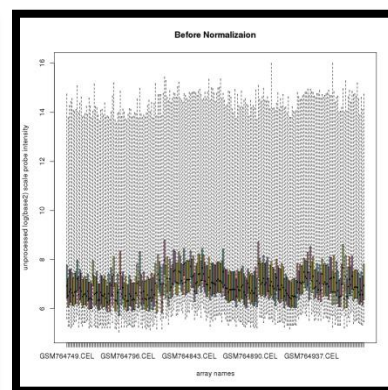
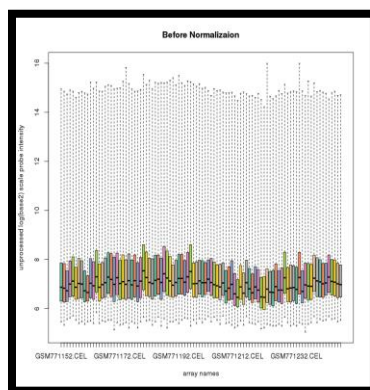
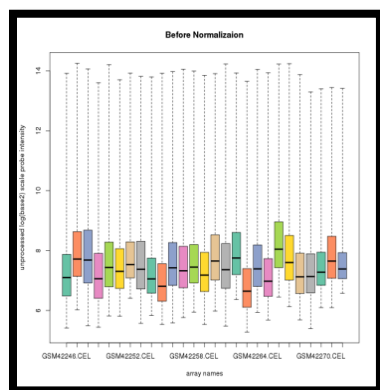
Boxplots of distribution (log-intensity) are plotted for comparison. The distributions of raw PM log-intensities must not be identical but also not totally different. The distributions of normalized probeset log-intensities must be comparable if they are not identical, normalization makes the distributions even. These boxplots allow the checking of the normalization step.

Box plots before normalization
> library(RColorBrewer)
> library(made4)
> brewer.cols <- brewer.pal(8, "Set2")
> boxplot(raw.data2280, main="Before Normalization", col=brewer.cols, + ylab="unprocessed log(base2) scale probe intensity", + xlab="array names")
Box plots after normalization
> brewer.cols <- brewer.pal(8, "Set2")
> boxplot(, main="Before Normalization", col=brewer.cols, + ylab="unprocessed log(base2) scale probe intensity", xlab="array names")

[A] GSE2280

GSE30784

GSE31056



[B] GSE2280

GSE30784

GSE31056

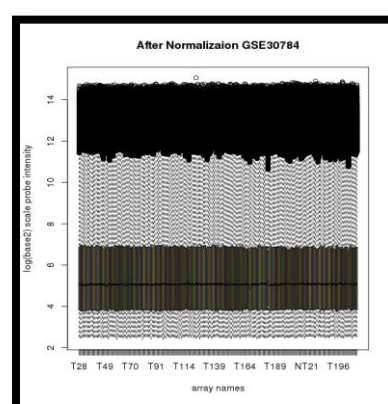
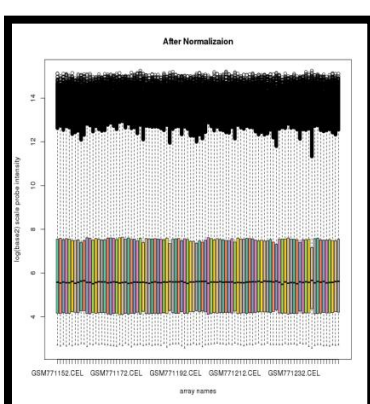
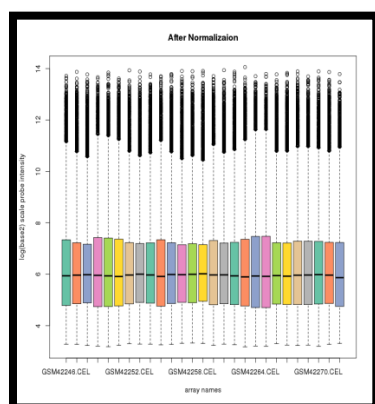


Figure 8: Boxplot - (A) Plot of raw data i.e. before normalization and (B) plot of expression data set after normalization.

5.6.5. Hierarchical clustering

Hierarchical clustering produces a dendrogram to see whether the samples of the same group are clustered together or not.

Dendrogram of expression set

```
> dat.dist<-dist(t(Eset))
```

```
> plot(hclust(dat.dist))
```

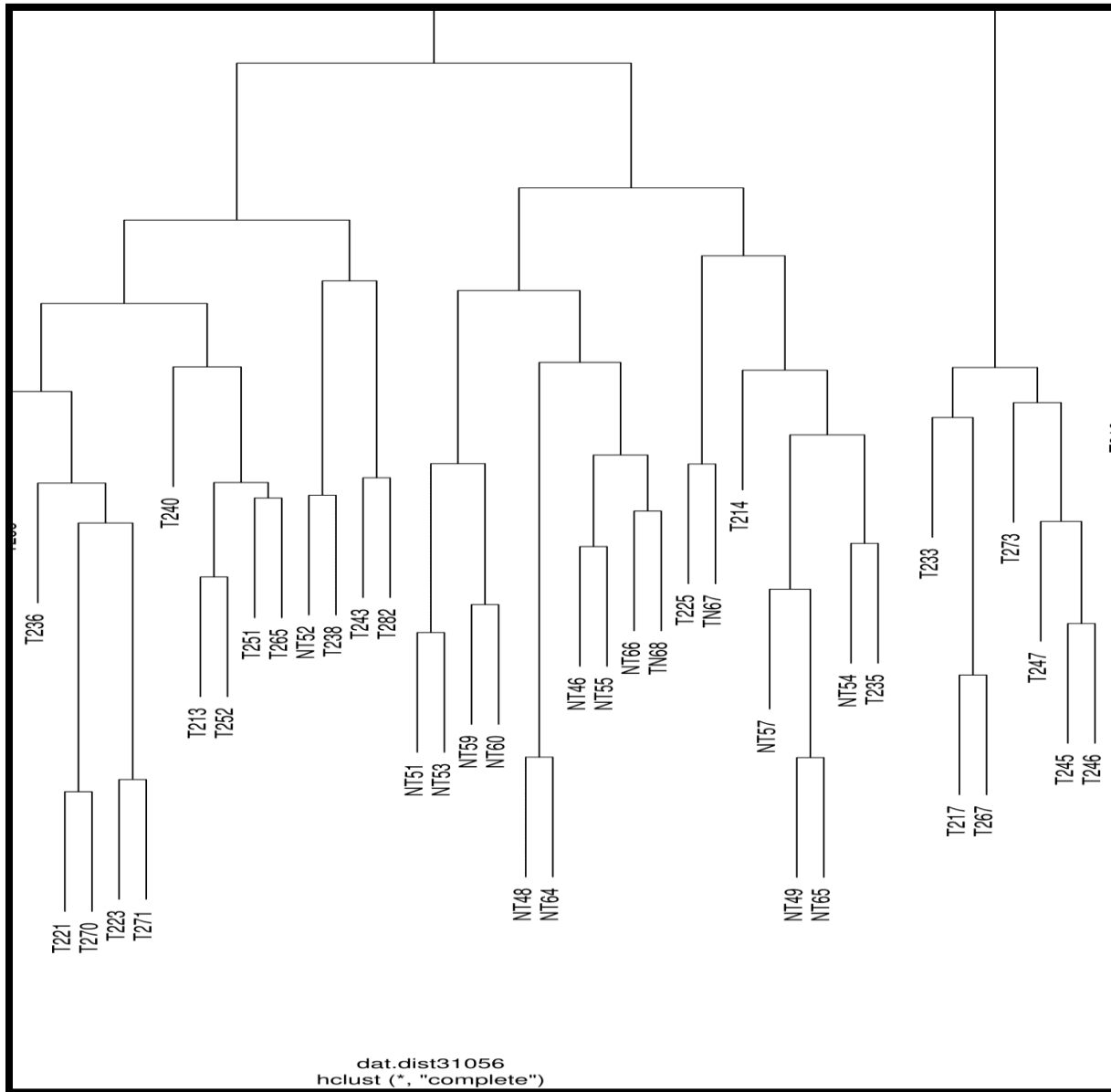


Figure 9: Dendrogram showing hierarchical clustering of samples (GSE31056)

Now, all the samples (normal and cancer) group nicely (biological grouping), indicating that the data is clean. Hence, ready for the further analyses. There are few exception in this clustering, non-tumor sample no 52 is clustered with tumors while non-tumor sample no 214, 225 and 235 are clustered with tumor samples, which should be removed from the analysis.

5.7. Cross-Platform Normalization

Combining a large number of gene expression datasets originating from different labs could be beneficial for the discovery of new biological insights and could increase the statistical power of gene expression analysis, but then this data should be combined in a consistent manner. “inSilicoMerging” package was used, this package provides different methods for data merging from which we have used Combat method (eBayes method) which removes Batch effect. Accessing uniformly represented data is only the first step when combining and integrating gene expression data sets since the use of different experimentation plans, platforms, and methodologies by different research groups introduce undesired batch effects in the gene expression measurements thus severely hindering downstream analysis.

5.7.1. Data preparation before merging: Before merging the datasets, there is preparation of datasets. In this the phenotypic information of samples are incorporated with their expression data.

```
## Data preparation (for individual datasets)
#Incorporating phenotypic data with expression data
> pData2280 <-read.AnnotatedDataFrame("GSE2280.txt",header=T)
> colnames(eset2280) <- row.names(pData2280)
> ALLSet2280 <- new("ExpressionSet", exprs = exprs(eset2280), phenoData = pData2280, annotation = "hgu133a")
> pData(ALLSet2280)
```

5.7.2. Dataset Merging: Different datasets each from different platforms are then merged. Due to differences in their platforms there are batcheffects which should be removed.

```
##Dataset merging
> library(inSilicoMerging)
> esets = list(ALLSet2280, ALLSet30784, ALLSet31056);
#MERGING WITHOUT BATCH CORRECTION
> eset_NONE = merge(esets, method="NONE");
```

```
#MERGING WITH BATCH CORRECTION
> eset_COMBAT = merge(esets, method="COMBAT")
#EXPRESSIONSET
> Esets<- exprs(eset_COMBAT)
```

5.7.3. Validation of batch correction: Validation should be done to check the removal of batcheffect. There are several validation tools enabling the inspection of the integration process, these packages enable researchers to fully explore the potential of combining gene expression data for downstream analysis.

For the visual inspection of merging results, five qualitative validation methods are provided. In additionl. These quantitative indices provide a more accurate evaluation of the batch effect removal and they are very effective tools for comparing the results of different methods.

plotMDS

creates a double-labeled Multidimensional Scaling (MDS) plot.

```
# MDS plot before batch correction
> plotMDS(eset_NONE,
  + colLabel = "characteristics",
  + symLabel = "Study",
  + main = "NONE (No Transformation)")
# MDS plot after batch correction
> plotMDS(eset_COMBAT,
  + colLabel = "characteristics",
  + symLabel = "Study",
  + main = "COMBAT")
```

It is intuitively clear from the MDS plots that samples cluster by study without any transformation and by disease after performing COMBAT.

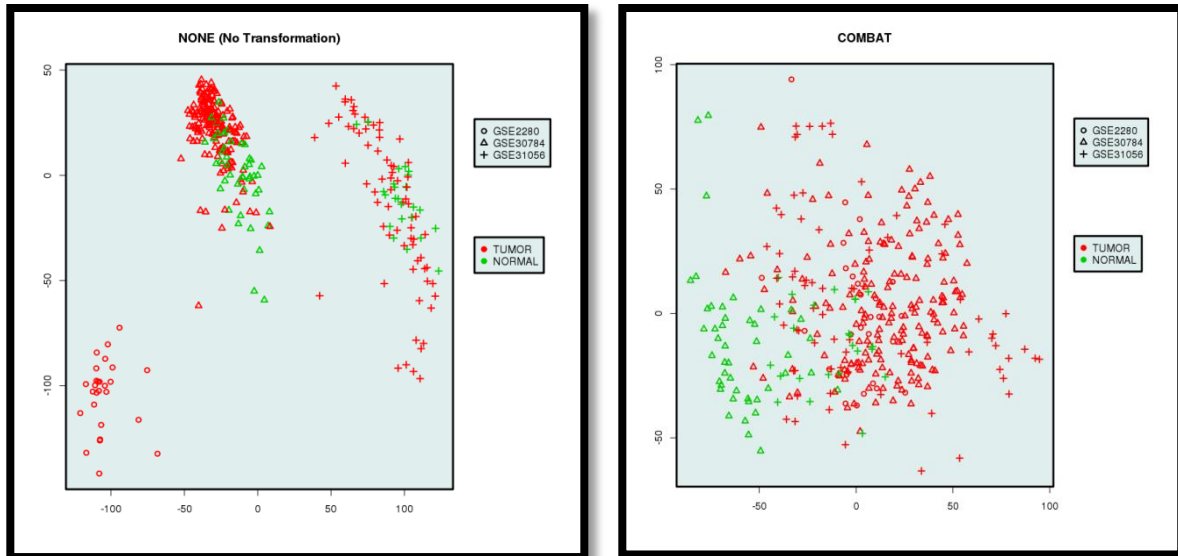


Figure 10: In these MDS plots samples are labeled by color based on the target biological variable of interest and are labeled by symbol based on the study they originate from. On the left the two data sets are merged without any transformation and on the right the two data sets are merged by using the **COMBAT** method.

plotRLE

creates a relative log expression (RLE) plot, initially proposed to measure the overall quality of a data set [24] but also useful in this context.

```
# RLE plot before batch correction
```

```
> plotRLE(eset_NONE,
  + colLabel = "characteristics",
  + symLabel = "Study",
  + main = "NONE (No Transformation)")
```

```
# RLE plot after batch correction
```

```
> plotRLE(eset_COMBAT,
  + colLabel = "characteristics",
  + symLabel = "Study",
  + main = "COMBAT")
```

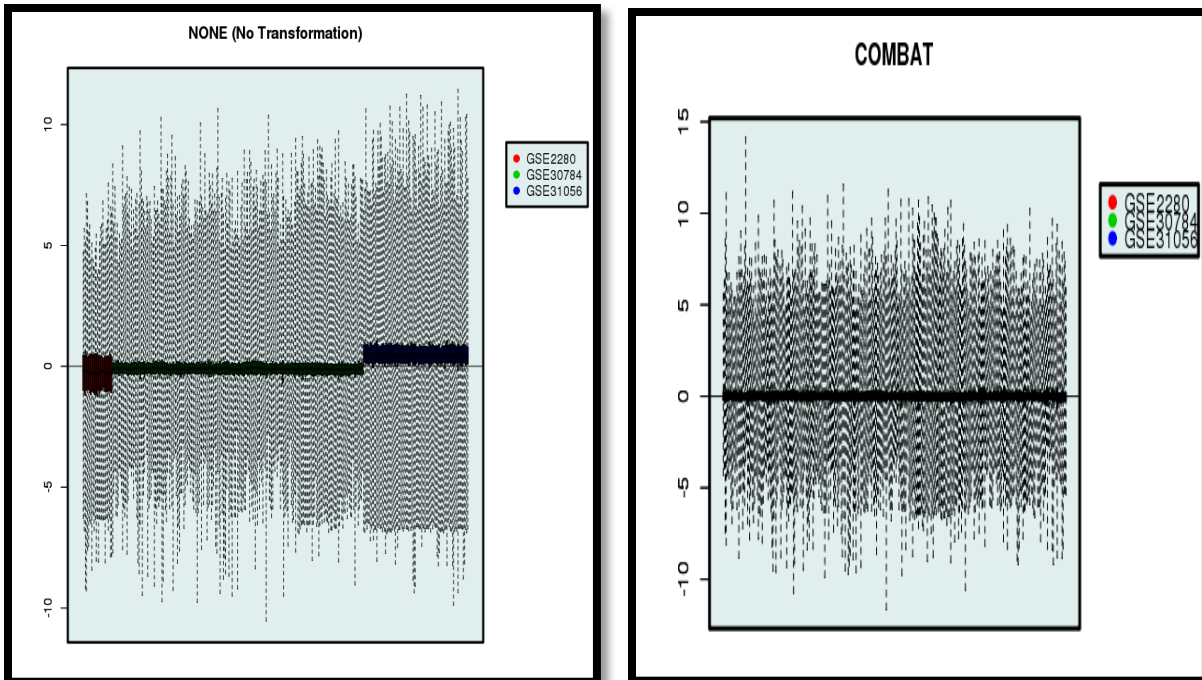


Figure 11: In these relative log expression plots samples are colored by study. On the left the two data sets are merged without any transformation and on the right the two data sets are merged by using the COMBAT method. After applying COMBAT the mean of the RLE is approximately 0 for all genes which indicates a good batch effect removal.

plotGeneWiseBoxPlots

provides a local visualization by looking at the box plots of a specific gene across all samples.

```
# GeneWiseBoxPlot before batch correction
> gene = sample(rownames(exprs(eset_NONE)), 100)
> plotGeneWiseBoxPlot(eset_NONE,
  + batchLabel = "Study",
  + colLabel = "characteristics",
  + gene = gene,
  + main = "NONE (No Transformation)");
# GeneWiseBoxPlot after batch correction
```

```

> plotGeneWiseBoxPlot(eset_COMBAT,
  + batchLabel = "Study",
  + colLabel = "characteristics",
  + gene = gene,
  + main = "COMBAT")

```

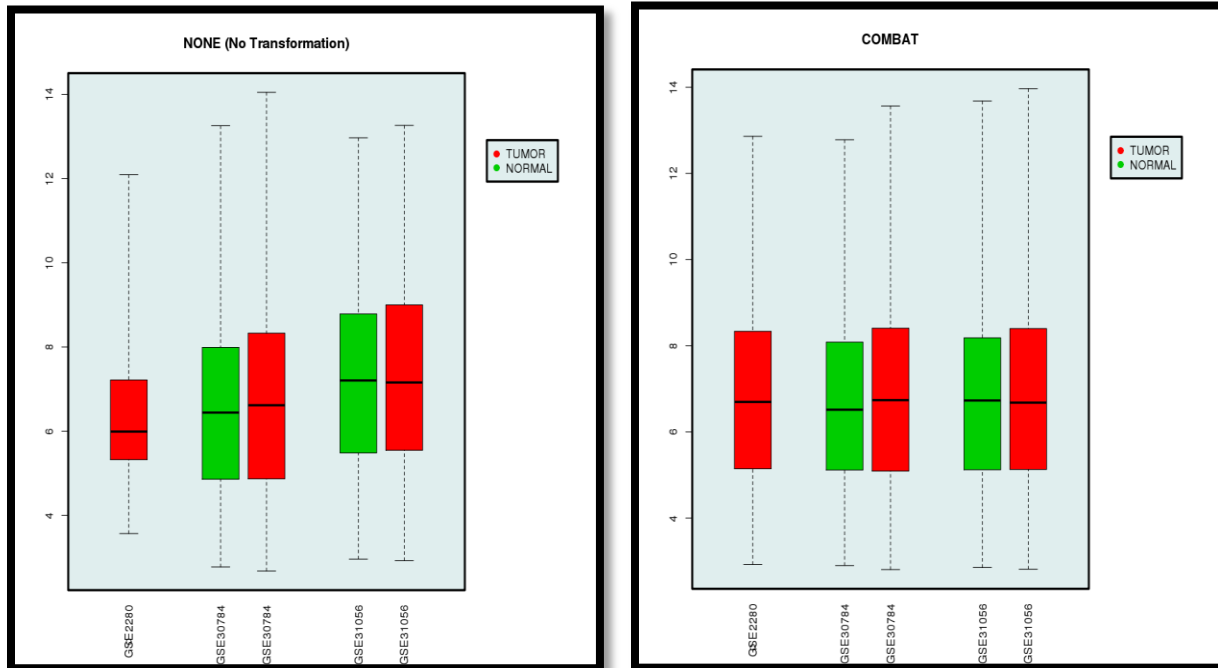


Figure 12: GeneWiseBoxplots of the randomly selected gene are grouped by study and colored by the target biological variable of interest. On the left the two data sets are merged without any transformation and on the right the two data sets are merged by using the COMBAT method. After batch effect removal the distribution of the gene is much more similar between studies than without.

5.8. Filtering Data

Uninformative data was removed such as control probesets and other internal controls as well as removing genes with low variance, which will be unlikely to pass statistical tests for differential expression, or are expressed uniformly close to background detection levels.

# Data filtering	
> filtereddata<- nsFilter(eset_COMBAT, require.entrez = FALSE, remove.dupEntrez = FALSE)	
# What got removed	
>filtereddata\$filter.log	
\$numLowVar [1] 11108	\$feature.exclude [1] 62
>filteredEset<- exprs(filtereddata\$eset)	

5.9. Statistical analysis for differential expression

Statistical analysis microarray data is still under development. There are no strict guidelines/rules of thumb when to apply or not to apply some tests and certain other tests. Limma is one of the widely used tools (package limma) for the statistical analysis, which implements linear models (analyzing very complicated datasets). The samples were coded with "T" and "N", the tumor i.e. cancer samples are coded as "T" and control samples as "NT".

extract information about the samples:
>samples <- c(eset2280\$characteristics, eset30784\$characteristics, eset31056\$characteristics)
convert into factors
> samples <- as.factor(samples)
set up the experimental design
> design <- model.matrix(~0 + samples)
> colnames(design) <- c("TUMOR", "NORMAL")
inspect the experiment design
> design
design TUMOR NORMAL 1 0 1 2 0 1 . .


```

.
344  1  0
.
.
350  0  1
351  1  0
352  0  1
attr("assign")
[1] 1 1
attr("contrasts")
attr("contrasts")$samples
[1] "contr.treatment"

```

5.9.1. Differential gene expression analysis:

The analysis was done by using the `lmFit()` command followed by `eBayes()`. The `lmFit()` get the design matrix, and a data matrix. The analysis was carried out using the filtered data.

```

# library(limma)
# fit the linear model to the filtered expression set
>fit <- lmFit(filteredEset, design)
>contrast.matrix <- makeContrasts("TUMOR-NORMAL", levels = colnames(design))
# Now the contrast matrix is combined with the per-probeset linear model fit
>fit_model <- contrasts.fit(fit, contrast.matrix)

```

Now the differential expression was calculated by empirical Bayes shrinkage of the standard errors towards a common value, by computing the moderated t-statistics, moderated F-statistic, and log-odds.

```

>ebayes_fit <- eBayes(fit_model)
# return the top results for given contrast
>probeset.list <- topTable(ebayes_fit, coef=1, p.value=0.05, lfc=1)
>probeset.list2 <- probeset.list1[(probeset.list1$adj.P.Val <= 0.05) & (abs(probeset.list1$logFC)
>= 3), ]
>dim(probeset.list2)
[1] 16 6

```

```

> probeset.list2 <- probeset.list1[(probeset.list1$adj.P.Val <= 0.05) & (abs(probeset.list1$logFC) >=
3), ]
> probeset.list2
      logFC  AveExpr      t    P.Value  adj.P.Val      B
206605_at  3.136642  7.007989 14.409184 2.306581e-37 2.134933e-34 74.13823
204777_s_at 4.090923  9.740338 12.323597 2.781421e-29 3.634500e-27 55.75682
204475_at  -4.834662  8.945692 -12.054122 2.855362e-28 2.952818e-26 53.45754
204580_at  -3.633232  8.828346 -11.738389 4.256423e-27 3.216060e-25 50.79052

```

Figure 13: Screenshot of list of differentially expressed probes with log-fold change and p-value.

The list of Differentially Expressed probes with $\text{adj.P.Val} \leq 0.05$ and fold change ≥ 3 was created, and then a heat map of the expression was made.

```

# Heat map of Differentially expressed genes
>final_probes = rownames(probeset.list2)
> exp_value = Esets[final_probes,]
> test123 = read.delim("annotations1.txt")
> gene_sym = test123[,2]
> diff_exp_value = exp_value
>rownames(diff_exp_value) = gene_sym
>heatmap.2((diff_exp_value),col=redgreen(10),scale="row", density.info="none", trace="none")

```

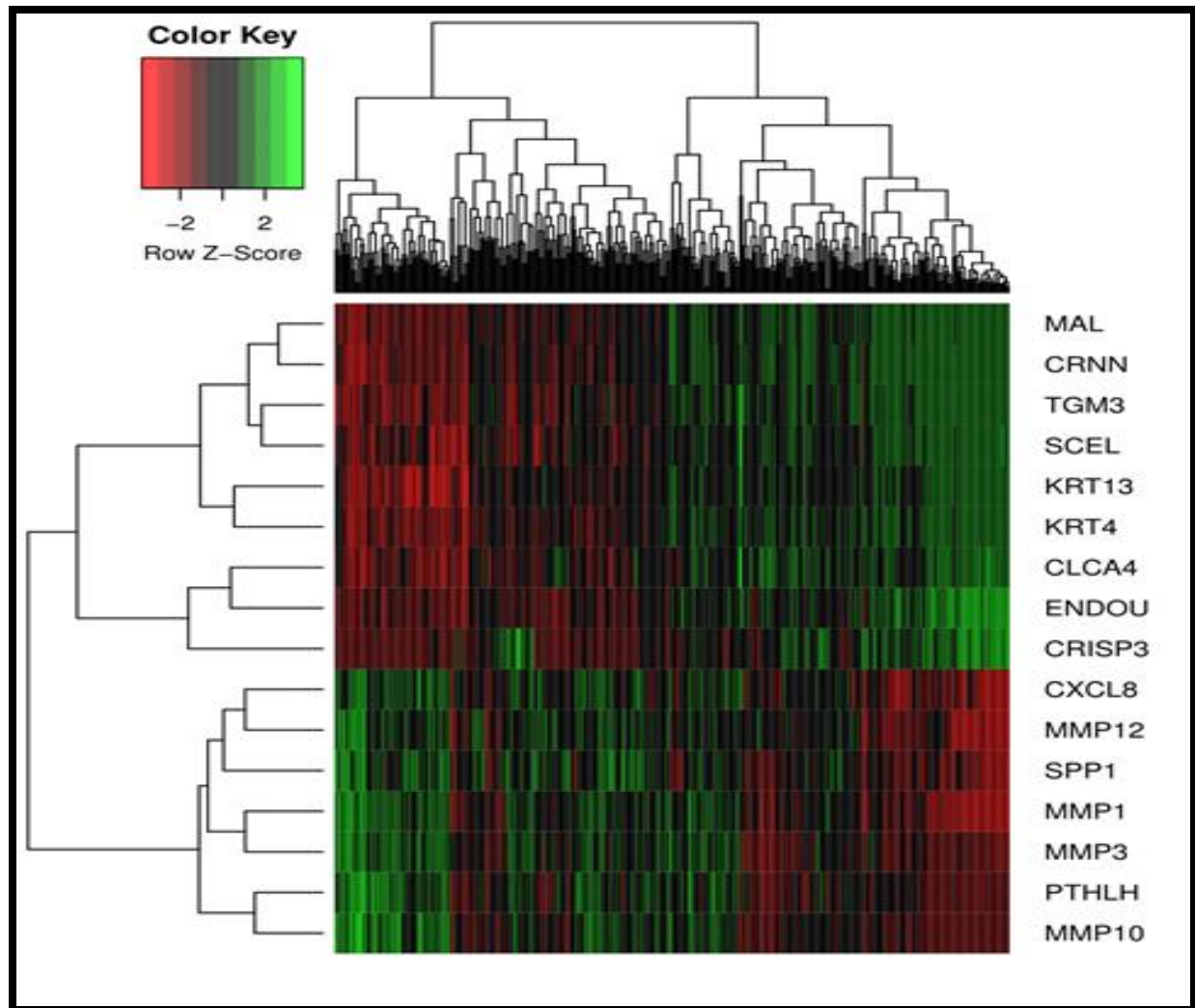


Figure 14: Heat map of DEG showing sample relationship by column dendrogram.

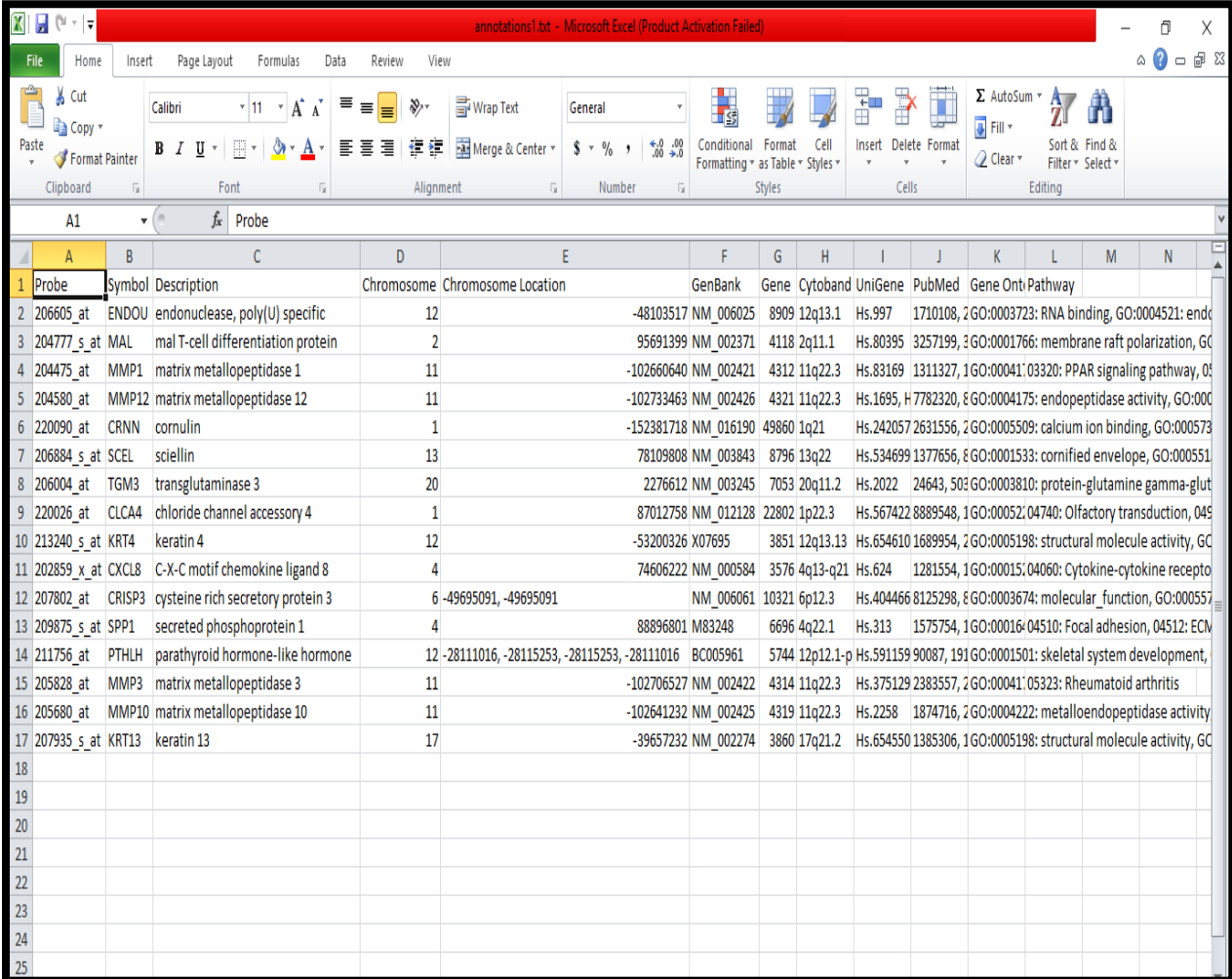
5.10. Annotating the results with associated gene symbols

Bioconductor have annotation packages for many types of chip, and can be used directly for annotation. As an input, it takes gene names as a vector, which could be extracted from matrix. Output is an HTML or a text file containing the annotations.

```
# library(hgu133a.db) #library(annaffy)
>genes<-rownames(probeset.list2)
> test123 = read.delim("annotations1.txt")
> annot.cols<-aaf.handler()
```

```
>annot.table<-aafTableAnn(genes, "hgu133a.db", annot.cols)
```

```
>saveHTML(annot.table1, "annotations1.html")
```



Probe	Symbol	Description	Chromosome	Chromosome Location	GenBank	Gene	Cytoband	UniGene	PubMed	Gene Ont Pathway
206605_at	ENDOU	endonuclease, poly(U) specific	12		-48103517 NM_006025	8909	12q13.1	Hs.997	1710108, 2	GO:0003723: RNA binding, GO:0004521: endo
204777_s_at	MAL	mal T-cell differentiation protein	2		95691399 NM_023371	4118	2q11.1	Hs.80395	3257199, 3	GO:0001766: membrane raft polarization, GO
204475_at	MMP1	matrix metalloproteinase 1	11		-102660640 NM_02421	4312	11q22.3	Hs.83169	1311327, 1	GO:00041: 03320: PPAR signaling pathway, 0
204580_at	MMP12	matrix metalloproteinase 12	11		-102733463 NM_02426	4321	11q22.3	Hs.1695, H	7782320, 8	GO:0004175: endopeptidase activity, GO:00
220090_at	CRNN	cornulin	1		-152381718 NM_016190	49860	1q21	Hs.242057	2631556, 2	GO:0005509: calcium ion binding, GO:000573
206884_s_at	SCEL	sciellin	13		78109808 NM_003843	8796	13q22	Hs.534699	1377656, 8	GO:0001533: cornified envelope, GO:000551
206004_at	TGM3	transglutaminase 3	20		2276612 NM_003245	7053	20q11.2	Hs.2022	24643, 503	GO:0003810: protein-glutamine gamma-glut
220026_at	CLCA4	chloride channel accessory 4	1		87012758 NM_012128	22802	1p22.3	Hs.567422	8889548, 1	GO:00052: 04740: Olfactory transduction, 04
213240_s_at	KRT4	keratin 4	12		-53200326 X07695	3851	12q13.13	Hs.654610	1689954, 2	GO:0005198: structural molecule activity, GC
202859_x_at	CXCL8	C-X-C motif chemokine ligand 8	4		74606222 NM_000584	3576	4q13-q21	Hs.624	1281554, 1	GO:00015: 04060: Cytokine-cytokine recepto
207802_at	CRISP3	cysteine rich secretory protein 3	6	-49695091, -49695091	NM_006061	10321	6p12.3	Hs.404466	8125298, 8	GO:0003674: molecular_function, GO:000557
209875_s_at	SPP1	secreted phosphoprotein 1	4		88896801 M83248	6696	4q22.1	Hs.313	1575754, 1	GO:00016: 04510: Focal adhesion, 04512: ECM
211756_at	PTH1H	parathyroid hormone-like hormone	12	-28111016, -28115253, -28115253, -28111016	BC005961	5744	12p12.1-p	Hs.591159	90087, 191	GO:0001501: skeletal system development,
205828_at	MMP3	matrix metalloproteinase 3	11		-102706527 NM_02422	4314	11q22.3	Hs.375129	2383557, 2	GO:00041: 05323: Rheumatoid arthritis
205680_at	MMP10	matrix metalloproteinase 10	11		-102641232 NM_02425	4319	11q22.3	Hs.2258	1874716, 2	GO:0004222: metalloendopeptidase activity
207935_s_at	KRT13	keratin 13	17		-39657232 NM_002274	3860	17q21.2	Hs.654550	1385306, 1	GO:0005198: structural molecule activity, GC
18										
19										
20										
21										
22										
23										
24										
25										

Figure 15: List of DEG with annotation.

. 6. RESULTS

Differentially expressed genes from linear model analysis

Top sixteen genes were identified by statistical testing of filtered expression set (Table: 1). The first column of the table is the probe set I.D., The next column (logFC) shows a fold (log₂-based) change between the groups. Down-regulated genes are shown with positive sign, and negative values are indicating up-regulation. The t is the moderated t-statistics and the B is the log-odds showing that the gene is differentially expressed. The p-value columns contain the p-value corrected for multiple comparisons (adj.P.Val) and raw p-value (P.Value) using false discovery rate.

	logFC	t	P.Value	adj.P.Val	B
206605_at	3.136642	7.007989	14.40918	2.31E-37	2.13E-34
204777_s_at	4.090923	9.740338	12.3236	2.78E-29	3.63E-27
204475_at	-4.83466	8.945692	-12.0541	2.86E-28	2.95E-26
204580_at	-3.63323	8.828346	-11.7384	4.26E-27	3.22E-25
220090_at	4.370781	9.522771	11.54485	2.20E-26	1.52E-24
206884_s_at	3.119697	9.366338	11.2825	1.99E-25	1.24E-23
206004_at	3.57463	9.720079	10.97813	2.50E-24	1.20E-22
220026_at	3.367255	8.011276	10.83141	8.36E-24	3.63E-22
213240_s_at	4.314265	10.41289	10.64071	3.97E-23	1.50E-21
202859_x_at	-3.08216	9.098558	-10.5723	6.92E-23	2.50E-21
207802_at	4.286134	6.63631	10.47631	1.50E-22	5.12E-21
209875_s_at	-3.13834	8.22097	-10.4012	2.75E-22	8.89E-21
211756_at	-3.13545	7.144302	-10.2204	1.17E-21	3.37E-20
205828_at	-3.35283	8.052793	-9.81981	2.75E-20	6.07E-19
205680_at	-3.27414	6.964798	-9.3747	8.45E-19	1.49E-17
207935_s_at	3.291211	11.53424	8.24633	3.24E-15	3.27E-14

Table 1: Output of statistical testing of filtered data based on slandered deviation.

Annotation table (Table: 2) have gene names, its description, chromosome number on which that gene is present, GenBank accession number and UniGene I.D.

Probe	Symbol	Description	Chromosome	GenBank	UniGene
206605_at	ENDOU	endonuclease, poly(U) specific	12	NM_006025	Hs.997
204777_s_at	MAL	mal T-cell differentiation protein	2	NM_002371	Hs.80395
204475_at	MMP1	matrix metalloproteinase 1	11	NM_002421	Hs.83169
204580_at	MMP12	matrix metalloproteinase 12	11	NM_002426	Hs.1695, Hs.709832
220090_at	CRNN	cornulin	1	NM_016190	Hs.242057
206884_s_at	SCEL	sciellin	13	NM_003843	Hs.534699
206004_at	TGM3	transglutaminase 3	20	NM_003245	Hs.2022
220026_at	CLCA4	chloride channel accessory 4	1	NM_012128	Hs.567422 Hs.654610,
213240_s_at	KRT4	keratin 4	12	X07695	Hs.731814
202859_x_at	CXCL8	C-X-C motif chemokine ligand 8	4	NM_000584	Hs.624
207802_at	CRISP3	cysteine rich secretory protein 3	6	NM_006061	Hs.404466
209875_s_at	SPP1	secreted phosphoprotein 1	4	M83248	Hs.313
211756_at	PTH1H	parathyroid hormone-like hormone	12	BC005961	Hs.591159
205828_at	MMP3	matrix metalloproteinase 3	11	NM_002422	Hs.375129
205680_at	MMP10	matrix metalloproteinase 10	11	NM_002425	Hs.2258
207935_s_at	KRT13	keratin 13	17	NM_002274	Hs.654550

Table 2: The annotation of probe I.D. for differentially expressed genes.

The values obtain from table 1; the negative sign shows the up-regulation of genes (MMP1, MMP12, CXCL8, SPP1, PTH1H, MMP3 and MMP10), while positive sign indicates the down-regulation of genes (ENDOU, MAL, CRNN, SCEL, TGM3, CLCA4, KRT4, CRISP3 and KRT13).

7. DISCUSSION AND CONCLUSION

OSCC is often associated with loss of eating and speech function, disfigurement and psychological distress. The development of OSCC is strongly associated with smoking and excessive alcohol consumption [93]. The prevention and management of this disease is likely to benefit from the identification of molecular markers and targets [94-95].

ENDOUB gene encodes a protein with protease activity and is expressed in the placenta. The protein may be useful as a tumor marker. *Karagoz K et. al.* report ENDOUB downregulation to be associated with esophageal cancer for the first time [96].

The protein encoded by MAL gene is a highly hydrophobic integral membrane protein belonging to the MAL family of proteolipids. The protein plays a role in the formation, stabilization and maintenance of glycosphingolipid-enriched membrane microdomains. Downregulation of MAL causes membrane to destabilized [97].

MMP1/3/10/12 genes encode a member of the peptidase M10 family of matrix metalloproteinases (MMPs). Matrix metalloproteinases (matrix metalloproteinase, MMPs), also called matrixins, are zinc-dependent endopeptidases and the major proteases in ECM degradation. MMPs are capable of degrading several extracellular molecules and a number of bioactive molecules. These genes found to be highly overexpressed in OSCC [98-110].

According to *yen et. al.* *MMP10* displayed the best sensitivity for oral cancer detection with any controls. MMP1 and MMP10 were suitable markers for cancer detection with gingiva and margin as controls. Using neck tissue as the control, only MMP10 was suitable for cancer detection. With margin and neck controls, there were no significant differences for MMP1, MMP10 and MMP12 in different stages, invasion and locations or different habits. Therefore, MMP1 and MMP10 but not MMP12 are potential oral cancer markers [111].

CRNN gene encodes a member of the "fused gene" family of proteins, which contain N-terminus EF-hand domains and multiple tandem peptide repeats. This gene, also known as squamous epithelial heat shock protein 53, may play a role in the mucosal/epithelial immune response and epidermal differentiation. Survival factor that participates in the clonogenicity of squamous esophageal epithelium cell lines attenuates deoxycholic acid (DCA)-induced apoptotic cell death

and release of calcium. When overexpressed in oral squamous carcinoma cell lines, regulates negatively cell proliferation by the induction of G1 arrest [112, 113].

TGM3; Transglutaminases are enzymes that catalyze the crosslinking of proteins by epsilon-gamma glutamyl lysine isopeptide bonds. This is a candidate tumor suppressor gene whose downregulation results in HNCC [115].

The protein encoded by CLCA4 gene belongs to the calcium sensitive chloride conductance protein family and was found to be downregulated in OSCC. [97,116]

KRT4, The protein encoded by this gene is a member of the keratin gene family. Downregulation of this gene is associated with morphological changes in the affected oral epithelium. [117]

The protein encoded by CXCL8 gene is a member of the CXC chemokine family. This chemokine is one of the major mediators of the inflammatory response. This gene is associated with GO term “negative regulation of keratinocyte proliferation” which is related to processes associated with multiplication or reproduction of keratinocytes; these processes ultimately increase the cell population. Malignant oral keratinocytes express 5–50 times more EGFR than do their healthy counterparts [118]; therefore, activation of EGFR enhances proliferation and the metastatic potential of keratinocytes [119].

The human cysteine-rich secretory protein (CRISP) family is a group of glycoproteins. *Wen-Chang Ko et. al.* suggest that the CRISP3 gene is a novel tumor suppressor gene particular to OSCC, and inactivation of the CRISP3 gene may play one or more roles in the carcinogenesis of OSCCs [120].

SPP1, which codes for osteopontin. This secreted glycoprotein has an important role in determining the oncogenic potential of many cancers and its increased expression is reported to correlate with tumor progression and metastasis [121, 122]

The protein encoded by PTHLH gene is a member of the parathyroid hormone family, PTHLH is up-regulated in OSCCs. Therefore, it could play a role in the pathogenesis of OSCC by affecting cell proliferation and cell cycle, and the protein levels of PTHLH might serve as a prognostic indicator for evaluating patients with HNSCCs. [123].

The protein encoded by KRT13 gene is a member of the keratin gene family. Although the loss of keratin 13 (KRT13) is reportedly linked to malignant transformation of oral epithelial cells, the molecular mechanisms through which KRT13 is repressed in oral squamous cell carcinoma (OSCC) remain unclear. [117,124].

All genes that are differentially expressed are previously shown to play role in oral cancer except ENDOU and SCEL, which are related to esophageal cancer.

8. FUTURE PERSPECTIVES

Due to small sample size the results cannot be proved significant so the analysis should be done using large sample size and using more raw data from different studies would provide significant information of DEG in OSCC. The results obtained till now using the large sample size from our current analysis combined with recent studies reveal that finding DEG using R and Bioconductor gives quite good results. The resulting genes can be experimentally verified using RT-PCR. These DEG which are potential biomarker of OSCC would allow to detect oral cancer early and can be used as drug target for the treatment of oral cancer.

REFERENCES

1. Sotiriou C, Lothaire P, Dequanter D, Cardoso F, Awada A. Molecular profiling of head and neck tumors. *Curr Opin Oncol* 2004; 16(3): 211-14.
2. Lippman SM, Hong WK. Molecular markers of the risk of oral cancer. *N Engl J Med* 2004; 344: 1323–1326.
3. Parkin DM, Bray F, Ferlay J, Pisani P. Global cancer statistics, 2002. *CA Cancer J Clin* 2005; 55: 74–108.
4. D.M. Parkin, F. Bray, J. Ferlay, P. Pisani, Global cancer statistics, *CA Cancer J Clin* 2002; 74–108.
5. Sudbo J. Novel management of oral cancer: A paradigm of predictive oncology. *Clin Med Res* 2004; 2(4):233-42.
6. Schliephake H. Prognostic relevance of molecular markers of oral cancer- a review. *Int J Oral Maxillofac Surg* 2003; 32:233-45.
7. Chimenos-Küstner E, Font-Costa I, López-López J. Oral cancer risk and molecular markers. *Med Oral Patol Oral Cir Buccal* 2004; 9: 377-84.
8. Scully C, Burkhardt A. Tissue markers of potentially malignant human oral epithelial lesions. *J Oral Pathol Med* 1993; 22:246-56.
9. Director's Challenge Consortium for the Molecular Classification of Lung A; Shedden, K.; Taylor, J.M.; Enkemann, S.A.; Tsao, M.S.; Yeatman, T.J.; Gerald, W.L.; Eschrich, S.; Jurisica, I.; Giordano, T.J.; et al. Gene expression-based survival prediction in lung adenocarcinoma: A multi-site, blinded validation study. *Nat. Med.* 2008, 14, 822–827.
10. Van Laar, R.; Flinchum, R.; Brown, N.; Ramsey, J., Riccitelli, S., Heuck, C.; Barlogie, B.; Shaughnessy, J.D., Jr. Translating a gene expression signature for multiple myeloma prognosis

into a robust high-throughput assay for clinical use. *BMC Med. Genom.* 2014, 7, oi: 10.1186/1755-8794-7-25.

11. Gesthalter, Y.B.; Vick, J.; Steiling, K.; Spira, A. Translating the transcriptome into tools for the early detection and prevention of lung cancer. *Thorax* 2015, 70, 476–481.

12. Shen, R.; Chinnaiyan, A.M.; Ghosh, D. Pathway analysis reveals functional convergence of gene expression profiles in breast cancer. *BMC Med. Genom.* 2008, 1, doi: 10.1186/1755-8794-1-28.

13. Shi, L.; Campbell, G.; Jones, W.D.; Campagne, F.; Wen, Z.; Walker, S.J.; Su, Z.; Chu, T.M.; Goodsaid, F.M.; Pusztai, L.; et al. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nat. Biotechnol.* 2010, 28, 827–838.

14. Simon, R. Genomic biomarkers in predictive medicine: An interim analysis. *EMBO Mol. Med.* 2011, 3, 429–435.

15. Diamandis, E.P. Cancer biomarkers: can we turn recent failures into success? *J. Natl. Cancer Inst.* 2010, 102, 1462–1467.

16. Baker, S.G. Improving the biomarker pipeline to develop and evaluate cancer screening tests. *J. Natl. Cancer Inst.* 2009, 101, 1116–1169.

17. Cruz, J.; Wishart, D. Applications of machine learning in cancer prediction and prognosis. *Cancer Inf.* 2006, 2, 59–77.

18. Michiels, S.; Koscielny, S.; Hill, C. Prediction of cancer outcome with microarrays: A multiple random validation strategy. *Lancet* 2005, 365, 488–492.

19. Hamid, J.S.; Hu, P.; Roslin, N.M.; Ling, V.; Greenwood, C.T.; Beyene, J. Data integration in genetics and genomics: Methods and challenges. *Hum. Genom. Proteom.* 2009, 2009, doi:10.4061/2009/869093.

20. Taminau, J.; Lazar, C.; Meganck, S.; Nowé, A. Comparison of merging and meta-analysis as alternative approaches for integrative gene expression analysis. *ISRN Bioinform.* 2014, 2014, doi:10.1155/2014/345106.
21. Ramasamy, A.; Mondry, A.; Holmes, C.C.; Altman, D.G. Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS Med.* 2008, 5, e184.
22. Hu, P.; Greenwood, C.M.; Beyene, J. Integrative analysis of multiple gene expression profiles with quality-adjusted effect size models. *BMC Bioinform.* 2005, 6, doi:10.1186/1471-2105-6-128.
23. Shabalin, A.A.; Tjelmeland, H.; Fan, C.; Perou, C.M.; Nobel, A.B. Merging two gene-expression studies via cross-platform normalization. *Bioinformatics* 2008, 24, 1154–1160.
24. Tseng, G.C.; Ghosh, D.; Feingold, E. Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Res.* 2012, 40, 3785–3799.
25. Hu, P.; Wang, X.; Haitzma, J.J.; Furmli, S.; Masoom, H.; Liu, M.; Imai, Y.; Slutsky, A.S.; Beyene, J.; Greenwood, C.M.; et al. Microarray meta-analysis identifies acute lung injury biomarkers in donor lungs that predict development of primary graft failure in recipients. *PLoS ONE* 2012, 7, e45506.
26. Perez-Diez, A.; Morgun, A.; Shulzhenko, N. Microarrays for cancer diagnosis and classification. *Adv. Exp. Med. Biol.* 2007, 593, 74–85.
27. Chang, C.; Wang, J.; Zhao, C.; Fostel, J.; Tong, W.; Bushel, P.R.; Deng, Y.; Pusztai, L.; Symmans, W.F.; Shi, T. Maximizing biomarker discovery by minimizing gene signatures. *BMC Genom.* 2011, 12, doi:10.1186/1471-2164-12-S5-S6.
28. Siegel R, Naishadham D, Jemal A. Cancer statistics, 2012. *CA Cancer J Clin* 2012; 62: 10-29.
29. Jemal A, Bray F, Center MM, Ferlay J, Ward E, et al. Global cancer statistics. *CA Cancer J Clin* 2011; 61: 69-90.

30. Barnes L, Evenson JW, Reichart P, Sidransky D. Pathology and Genetics of Head and Neck Tumours; Kleihues P, Sobin LH, editors. Lyon 2005; IARC Press. 430 p.
31. Funk GF, Karnell LH, Robinson RA, Zhen WK, Trask DK, et al. Presentation, treatment, and outcome of oral cavity cancer: a National Cancer Data Base report. *Head Neck* 2002; 24: 165-180.
32. Johnson N et. al. Tobacco use and oral cancer: a global perspective. *J Dent Educ* 2001; 65: 328-339.
33. Neville BW, Day TA. Oral cancer and precancerous lesions. *CA Cancer J Clin* 2002; 52: 195-215.
34. Van Wyk CW, Stander I, Padayachee A, Grobler-Rabie AF. The areca nut chewing habit and oral squamous cell carcinoma in South African Indians. A retrospective study. *S Afr Med J* 1993; 83:425-429.
35. Mao L, Hong WK, Papadimitrakopoulou VA. Focus on head and neck cancer. *Cancer Cell* 2004; 5.
36. Marur S, D'Souza G, Westra WH, Forastiere AA. HPV-associated head and neck cancer: a virus-related cancer epidemic. *Lancet Oncol* 2010; 11: 781-789.
37. Ayas B, Saleem K, Azim W, Shaikh A: A clinicopathological study of oral cancers. *Biomedica* 2011; 27: 29–32.
38. Jalisi S: Management of the clinically negative neck in early squamous cell carcinoma of the oral cavity. *Otolaryngol Clin North Am* 2005; 38: 37–46.
39. McGuirt W, Johnson J, Meyers E. The management of the clinically negative HPV sample from the neck. *Arch Otolaryngol Head Neck Surg* 1995; 121: 278–282.
40. Pindborg JJ, Reichart PA, Smith CJ, van der Waal I, Sobin LH: Histological Typing of Cancer and Precancer of the Oral Mucosa: In Collaboration with L.H.Sobin and Pathologists in 9 Countries (WHO of Tumours), 2nd Ed. Springer, 1997.

41. O'Brien CJ, Lauer CS, Fredricks S, Clifford AR, McNeil EB, Bagia JS, Koulmandas C: Tumor thickness influences prognosis of T1 and T2 oral cavity cancer – but what thickness? *Head Neck* 2003; 25: 937–945.
42. Veness M, Morgan G, Sathiyaseelan Y, Gebiski V: Anterior tongue cancer and the incidence of cervical lymph node metastases with increasing tumor thickness: Should elective treatment to the neck be standard practice in all patients? *ANZ J Surg* 2005; 75: 101–105.
43. Po Wing Yuen A, Lam KY, Lam LK, Ho CM, Wong A, Chow TL, Yuen WF, Wei WI: Prognostic factors of clinically stage I and II oral tongue carcinoma - a comparative study of stage, thickness, shape, growth pattern, invasive front malignancy grading, Martinez-Gimeno score, and pathologic features. *Head Neck* 2002; 24: 513–520.
44. Fan S, Tang QL, Lin YJ, Chen WL, Li JS, Huang ZQ, Yang ZH, Wang YY, Zhang DM, Wang HJ, Dias-Ribeiro E, Cai Q, Wang L: A review of clinical and histological parameters associated with contralateral neck metastases. *Int J Oral Sci* 2011; 3: 180–191.
45. Lopes M, Nikitakis N, Reynolds M, Ord R: Biomarkers predictive of lymph node metastasis in oral squamous cell carcinoma. *J Oral Maxillofac Surg* 2002; 60: 142–147.
46. Arellano-Garcia ME, Li R, Liu X, Xie Y, Yan X, Loo JA, Hu S: Identification of tetranectin as a potential biomarker for metastatic oral cancer. *Int J Mol Sci* 2010; 11: 3106–3121.
47. Goda H, Nakashiro KI, Oka R, Tanaka H, Wakisaka H, Hato N, Hyodo M, Hamakawa H: One-step nucleic acid amplification for detecting lymph node metastasis of head and neck squamous cell carcinoma. *Oral Oncol Epub ahead to print*, 2012.
48. Menezes M, Lehn C, Gonçalves A: Epidemiological and histopathological data and E-cadherin-like prognostic factors in early carcinomas of the tongue and floor of the mouth. *Oral Oncol* 2007; 43: 656–661.
49. Sciubba JJ. Oral cancer. The importance of early diagnosis and treatment. *Am J Clin Dermatol* 2001;2(4):239–51.
50. McGurk M, Chan C, Jones J, O'Regan E, Sherriff M. Delay in diagnosis and its effect on outcome in head and neck cancer. *Br J Oral Maxillofac Surg* 2005;43(4):281–4.

51. Tanaka T, Tanaka M. Oral carcinogenesis and oral cancer chemoprevention: a review. *Pathol Res Int* 2011;2011:431246
52. Hall GL, Shaw RJ, Field EA, Rogers SN, Sutton DN, Woolgar JA, et al. P16 promoter methylation is a potential predictor of malignant transformation in oral epithelial dysplasia. *Cancer Epidemiol Biomarkers Prev* 2008;17(8):2174–9
53. Higgs, P.G. and T.K. Attwood. *Bioinformatics and Molecular Evolution*, ed. B.S. Ltd. 2006; Blackwell Publishing.
54. Yang, Y.H., et al., Normalization for cDNA microarray data: a robust composite method addressing single and multiple slides systematic variation. *Nucleic Acids Research* 2002; 30: p. e15.
55. Paoli, S., et al., Integrating gene expression profiling and clinical data. *International Journal of Approximate Reasoning* 2008; 47: p. 58T69.
56. Moffitt, R., et al., Effect of Outlier Removal on Gene Marker Selection Using Support Vector Machines. *Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference* 2005; 1: p. 917T920.
57. Olsen, S.H., D.G. Thomas and D.R. Lucas, Cluster analysis of immunohistochemical profiles in synovial sarcoma, malignant peripheral nerve sheath tumor, and Ewing sarcoma. *Modern Pathology* 2006; 19: p. 659T668.
58. Mramor et al. Visualization based cancer microarray data classification analysis. *Bioinformatics* 2007; 23: p. 2147T2154.
59. Daly, M. J, Rioux,, Schaffner, S. F., Hudson, T. J., Lander, E. S . "High-resolution haplotype structure in the human genome." *Nature Genetics* 2001; 29:229-232.
60. Schadt EE, Li C, Su C, Wong WH. Analyzing high-density oligonucleotide gene expression array data. *J CELL BIOCHEM* 2000; 80: 192-202.

61. Tseng GC, Oh MK, Rohlin L, Liao JC, Wong WH. Issues in the cDNAMicroarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *NUCLEIC ACIDS RES* 2001; 29: 2549-2557.
62. <http://www.bioconductor.org/overview>.
63. Robert C. G.; Vincent J. C.; Douglas M B.; Ben B.; Marcel D.; Sandrine D.; Byron E.; Laurent G.; Yongchao G.; Jeff G.; Kurt H.; Torsten H.; Wolfgang H.; Stefano I.; Rafael I.; Friedrich L.; Cheng L.; Martin M.; Anthony J R.; Gunther S.; Colin S.; Gordon S.; Luke T.; Jean YH Y.; Jianhua Z. Bioconductor: open software development for computational biology and bioinformatics *Genome Biology*, 2004 5:R80.
64. Rudy, J.; Valafar, F. Empirical comparison of cross-platform normalization methods for gene expression data. *BMC Bioinform.* 2011, 12, doi:10.1186/1471-2105-12-467.
65. Taminau, J.; Lazar, C.; Meganck, S.; Nowé, A. Comparison of merging and meta-analysis as alternative approaches for integrative gene expression analysis. *ISRN Bioinform.* 2014, doi: 10.1155/2014/345106.
66. Shabalín, A.A.; Tjelmeland, H.; Fan, C.; Perou, C.M.; Nobel, A.B. Merging two gene-expression studies via cross-platform normalization. *Bioinformatics* 2008, 24, 1154–1160.
67. Turnbull, A.K.; Kitchen, R.R.; Larionov, A.A.; Renshaw, L.; Dixon, J.M.; Sims, A.H. Direct integration of intensity-level data from Affymetrix and Illumina microarrays improves statistical power for robust reanalysis. *BMC Med. Genom.* 2012, 5, doi:10.1186/1755-8794-5-35.
68. Konstantinopoulos, P.A.; Cannistra, S.A.; Fountzilias, H.; Culhane, A.; Pillay, K.; Rueda, B.; Cramer, D.; Seiden, M.; Birrer, M.; Coukos, G.; Zhang, L.; et al. Integrated analysis of multiple microarray datasets identifies a reproducible survival predictor in ovarian cancer. *PLoS ONE* 2011, 6, e18202, doi:10.1371/journal.pone.0018202.
69. Hughey, J.J.; Butte, A.J. Robust meta-analysis of gene expression using the elastic net. *Nucleic Acids Res.* 2015, doi:10.1093/nar/gkv229.
70. Campaign, A.; Yang, Y.H. Comparison study of microarray meta-analysis methods. *BMC Bioinform.* 2010, 11, doi:10.1186/1471-2105-11-408.

71. Sims, A.H.; Smethurst, G.J.; Hey, Y.; Okoniewski, M.J.; Pepper, S.D.; Howell, A.; Miller, C.J.; Clarke, R.B. The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets—Improving meta-analysis and prediction of prognosis. *BMC Med. Genom.* 2008, 1, doi:10.1186/1755-8794-1-42.
72. Xu, L.; Tan, A.C.; Winslow, R.L.; Geman, D. Merging microarray data from separate breast cancer studies provides a robust prognostic test. *BMC Bioinformatics* 2008, 9, doi:10.1186/1471-2105-9-125.
73. Liu, C.C.; Hu, J.; Kalakrishnan, M.; Huang, H.; Zhou, X.J. Integrative disease classification based on cross-platform microarray data. *BMC Bioinform.* 2009, 10, doi:10.1186/1471-2105-10-S1-S25.
74. Lee, Y.; Scheck, A.C.; Cloughesy, T.F.; Lai, A.; Dong, J.; Farooqi, H.K.; Liao, L.M.; Horvath, S.; Mischel, P.S.; Nelson, S.F. Gene expression analysis of glioblastomas identifies the major molecular basis for the prognostic benefit of younger age. *BMC Med. Genom.* 2008, 1, doi:10.1186/1755-8794-1-52.
75. Deshwar, A.G.; Morris, Q. PLIDA: Cross-platform gene expression normalization using perturbed topic models. *Bioinformatics* 2014, 30, 956–961.
76. Jiang, H.; Deng, Y.; Chen, H.S.; Tao, L.; Sha, Q.; Chen, J.; Tsai, C.J.; Zhang, S. Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes. *BMC Bioinform.* 2004, 5, doi:10.1186/1471-2105-5-81.
77. Shen, R.; Ghosh, D.; Chinnaiyan, A.M. Prognostic meta-signature of breast cancer developed by two-stage mixture modeling of microarray data. *BMC Genom.* 2004, 5, doi:10.1186/1471-2164-5-94.
78. Parmagiani, G.; Garret-Mayer, E.S.; Anbazhagan, R.; Gabrielson, E. A cross-study comparison of gene expression studies for the molecular classification of lung cancer. *Clin. Cancer Res.* 2004, 10, 2922–2927.
79. Rudy, J.; Valafar, F. Empirical comparison of cross-platform normalization methods for gene expression data. *BMC Bioinform.* 2011, 12, doi:10.1186/1471-2105-12-467.

80. Johnson, W.E.; Li, C.; Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 2007, 8, 118–127.
81. Huang, H.; Lu, X.; Liu, Y.; Haaland, P.; Marron, J.S. R/DWD: Distance-weighted discrimination for classification, visualization and batch adjustment. *Bioinformatics* 2012, 28, 1182–1183.
82. <http://www.webarraydb.org/webarray/index.html>. (Accessed on 12 May 2015).
83. Christopher J. Walsh, Pingzhao Hu, Jane Batt and Claudia C. Dos Santos. Microarray Meta-Analysis and Cross-Platform Normalization: Integrative Genomics for Robust Biomarker Discovery. *Microarrays* 2015, 4, 389-406; doi:10.3390/microarrays4030389.
84. Rung, J.; Brazma, A. Reuse of public genome-wide gene expression data. *Nat. Rev. Genet.* 2013, 14, 89–99.
85. Taminau, J.; Lazar, C.; Meganck, S.; Nowé, A. Comparison of merging and meta-analysis as alternative approaches for integrative gene expression analysis. *ISRN Bioinform.* 2014, 2014, doi:10.1155/2014/345106.
86. Tseng, G.C.; Ghosh, D.; Feingold, E. Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Res.* 2012, 40, 3785–3799.
87. Xu, L.; Tan, A.C.; Winslow, R.L.; Geman, D. Merging microarray data from separate breast cancer studies provides a robust prognostic test. *BMC Bioinformatics* 2008, 9, doi:10.1186/1471-2105-9-125.
88. Warnat, P.; Eils, R.; Brors, B. Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes. *BMC Bioinform.* 2005, 6, doi:10.1186/1471-2105-6-265.
89. Fielden, M.R.; Nie, A.; McMillian, M.; Yi, Y.; Morrison, C.; Yang, P.; Sun, Z.; Szoke, J.; Gerald, W.L.; Watson, M.; et al. Interlaboratory evaluation of genomic signatures for predicting carcinogenicity in the rat. *Toxicol. Sci.* 2008, 103, 28–34.

90. Lu, Y.; Lemon, W.; Liu, P.Y.; Yi, Y.; Morrison, C.; Yang, P.; Sun, Z.; Szoke, J.; Gerald, W.L.; Watson M; et al. A gene expression signature predicts survival of patients with stage I non-small cell lung cancer. *PLoS Med.* 2006, 3, e467.
91. Cristina Botella Perez. *Virgili multivariate classification of gene expression microarray data.* ISBN 2010; 978-84-693-5427-8/DL:T-1418-2010
92. Leemans CR, Braakhuis BJ, Brakenhoff RH. The molecular biology of head and neck cancer. *Nat Rev Cancer* 2011; 11: 9-22.
93. La Vecchia C: Epidemiology and prevention of oral cancer. *Oral Oncol.* 33:302–312. 1997.
94. Sabichi AL, Demierre MF, Hawk ET, Lerman CE and Lippman SM: Frontiers in cancer prevention research. *Cancer Res.* 63:5649–5655. 2003.
95. Spafford MF, Koch WM, Reed AL, et al: Detection of head and neck squamous cell carcinoma among exfoliated oral mucosal cells by microsatellite analysis. *Clin Cancer Res.* 7:607–612. 2001.
96. Karagoz K, Lehman HL, Stairs DB, Sinha R1, Arga KY. Proteomic and Metabolic Signatures of Esophageal Squamous Cell Carcinoma. *Curr Cancer Drug Targets.* 2016 Feb 2.
97. Bundela S, Sharma A, Bisen PS. Potential therapeutic targets for oral cancer: ADM, TP53, EGFR, LYN, CTLA4, SKIL, CTGF, CD70. *PLoS One.* 2014 Jul 16;9(7):e102610. doi: 10.1371/journal.pone.0102610.
98. Tanis T, Cincin ZB, Gokcen-Rohlig B, Bireller ES, Ulasan M, Tanyel CR, Cakmakoglu B. The role of components of the extracellular matrix and inflammation on oral squamous cell carcinoma metastasis. *Arch Oral Biol.* 2014 Nov;59(11):1155-63. doi: 10.1016/j.archoralbio.2014.07.005.
99. Kalfert D, Ludvikova M, Topolcan O, Windrichova J, Malirova E, Pesta M, Celakovsky P. Analysis of preoperative serum levels of MMP1, -2, and -9 in patients with site-specific head and neck squamous cell cancer. *Anticancer Res.* 2014 Dec;34(12):7431-41.

100. Impola U, Uitto VJ, Hietanen J, Hakkinen L, Zhang L, Larjava H, Isaka K, Saarialho-Kere U. Differential expression of matrilysin-1 (MMP-7), 92 kD gelatinase (MMP-9), and metalloelastase (MMP-12) in oral verrucous and squamous cell cancer. *J Pathol.* 2004 Jan;202(1):14-22.

110. Stott-Miller M, Houck JR, Lohavanichbutr P, Méndez E, Upton MP, Futran ND, Schwartz SM, Chen C. Tumor and salivary matrix metalloproteinase levels are strong diagnostic markers of oral squamous cell carcinoma. *Cancer Epidemiol Biomarkers Prev.* 2011 Dec;20(12):2628-36. doi: 10.1158/1055-9965.EPI-11-0503.

111. Yen, Ching-Yu et. al. Matrix metalloproteinases (MMP) 1 and MMP10 but not MMP12 are potential oral cancer markers. *Biomarkers.* 2009 Jun;14(4):244-9. doi: 10.1080/13547500902829375.

112. Salahshourifar et. el. Downregulation of CRNN gene and genomic instability at 1q21.3 in oral squamous cell carcinoma. *Clin Oral Investig.* 2015 Dec;19(9):2273-83. doi: 10.1007/s00784-015-1467-7.

113. Pawar H et. al. Downregulation of cornulin in esophageal squamous cell carcinoma. *Acta Histochem.* 2013 Mar;115(2):89-99. doi: 10.1016/j.acthis.2012.04.003.

SCEL: The protein encoded by this gene is a precursor to the cornified envelope of terminally differentiated keratinocytes. Downregulation of this gene leads to esophageal cancer [114].

114. Corona W et. al. Analysis of Sciellin (SCEL) as a candidate gene in esophageal squamous cell carcinoma. *Anticancer Res.* 2004 May-Jun;24(3a):1417-9.

115. Wu X1, Cao W, Wang X, Zhang J, Lv Z, Qin X, Wu Y, Chen W. TGM3, a candidate tumor suppressor gene, contributes to human head and neck cancer. *Mol Cancer.* 2013 Dec 1;12(1):151. doi: 10.1186/1476-4598-12-151.

116. Ye H1, Yu T et. al. Transcriptomic dissection of tongue squamous cell carcinoma. *BMC Genomics.* 2008 Feb 6;9:69. doi: 10.1186/1471-2164-9-69.

117. Sakamoto K et. al. Down-regulation of keratin 4 and keratin 13 expression in oral squamous cell carcinoma and epithelial dysplasia: a clue for histopathogenesis. *Histopathology*. 2011 Mar;58(4):531-42. doi: 10.1111/j.1365-2559.2011.03759.
118. Wong DTW, Todd R, Tsuji T, Donoff RB. Molecular biology of human oral cancer. *Crit Rev Oral Biol Med*. 1996;7:319–28. 94.
119. Meyer-Hoffert U, Wingertzahn J, Wiedow O. Human leukocyte elastase induces keratinocyte proliferation by epidermal growth factor receptor activation. *J Invest Dermatol*. 2004;123:338–45.
120. Wen-Chang Ko et.al. Copy number changes of CRISP3 in oral squamous cell carcinoma. *Oncology Letters*, 2011; DOI: 10.3892/ol.2011.418.
121. Senger DR, Asch BB, Smith BD, Perruzzi CA, Dvorak HF (1983) A secreted phosphoprotein marker for neoplastic transformation of both epithelial and fibroblastic cells. *Nature* 302(5910): 714–71.
122. Brown LF, Papadopoulos-Sergiou A, Berse B, Manseau EJ, Tognazzi K, Perruzzi CA, Dvorak HF, Senger DR (1994) Osteopontin expression and distribution in human carcinomas. *Am J Pathol* 145(3): 610–623.
123. Lv ZI, Wu X et. al.Parathyroid hormone-related protein serves as a prognostic indicator in oral squamous cell carcinoma. *J Exp Clin Cancer Res*. 2014 Dec 18;33:100. doi: 10.1186/s13046-014-0100-y.
124. Naganuma K, Hatta M1, Ikebe T, Yamazaki J. Epigenetic alterations of the keratin 13 gene in oral squamous cell carcinoma. *BMC Cancer*. 2014 Dec 20;14:988. doi: 10.1186/1471-2407-14-988.