UNIVERSITY OF TARTU
Institute of Computer Science
Computer Science Curriculum

Sander Tars

Description and application of gene expression data analysis method Barcode

Bachelor's Thesis (9 ECTS)

Supervisor: Anna Ufliand, MSc
Supervisor: Priit Adler, PhD

Tartu 2016

Description and application of gene expression data analysis method Barcode

Abstract:
The main goals of this thesis is to assert whether gene expression data analysis method Barcode offers improvement over the method fRMA and to visualise the difference clearly.

First, descriptive part of this thesis focuses on the gene expression data analysis method Barcode. Barcode is explained by presenting an overview of different Barcode versions. For each version a description of functionalities and possible uses are given with emphasis on new functionalities, compared to the older versions.

Second, practical part of this thesis compares Barcode and fRMA method(fRMA method output is the starting point for Barcode analysis). To compare these two methods human gene expression dataset of DNA microarray experiment results is used. The dataset E-TAB-145 contains expression data from 158 human tissue samples. Tissue samples are first manually clustered to use as reference in comparison of these two methods. Data is then analysed with both Barcode and fRMA. To visualise and compare the result two statistical methods are separately used: Principal component analysis and Hierarchical clustering. For the results of both statisical analysis methods a detailed analysis is given. In the analysis it is concluded that Barcode really does offer an improvement over fRMA. Barcode allows samples to be classified better into clusters - samples of the same tissue type are separated better from other samples compared to fRMA.

Keywords: Principal component analysis (PCA), Hierarchical clustering, Barcode, gene expression, microarray experiment data, frozen RMA (fRMA)

CERCS:B110 Bioinformatics, medical informatics, biomathematics, biometrics

Geenieksпressiooni andmete analüüsi meetodi Barcode kirjeldus ja rakendamine

Lühikokkuvõte: Käesoleva bakalaureuse töö peamised eesmärgid on üle kontrollida, kas geenieksпressiooni andmete analüüsi meetod Barcode täiustab meetodit fRMA ja tuua erinevused visuaalselt välja.

Esimene, kirjeldav osa keskendub geenieksпressiooni andmete analüüsi meetodil Barcode. Barcode'i kirjelduse käigus antakse ülevaade erinevatest Barcode'i versioonidest. Iga versiooni juures on kirjeldatud funktsionaalsused ja nende kasutamine. Põhirõhk on seejuures pandud uutele funktsionaalsustele võrreldes varasemate versioonidega.

Teises, praktilises osas võrreldakse meetodeid Barcode ja fRMA (fRMA meetodi väljund on Barcode analüüsi alguspunkt). Nende kahe meetodi võrdlemiseks kasutatakse inimese geenieksпressiooni andmehulka DNA kiibi eksperimentidest. Andmehulk tähisega E-TABM-145 sisaldab 158 inimese koenäidise ekspressiooniandmeid. Kõigepealt jaotatakse need koenäidised manuaalselt gruppidesse. Need manuaalselt loodud grupid on aluseks mõlema meetodi töö hindamisele. Seejärel töödeldakse algseid andmeid nii meetodiga Barcode kui ka meetodiga fRMA. Mõlema meetodi tulemuste visualiseerimiseks ja võrdlemiseks kasutatakse eraldi kahte statistilist meetodit: peakomponentanalüüs (principal component analysis) ja hierarhiline klasterdamine. Mõlema statistilise meetodi väljunditele on tehtud analüüs ja võrdlus Barcode'i ja fRMA vahel. Vastavate statistiliste meetodite väljundite võrdlusest saab järeldada, et Barcode on tõepoolest täiendab fRMA-d. Barcode võimaldab koenäidiseid apremini õigetesse klastritesse klassifitseerida - näidised, mis tulevad samast koest on kasutades Barcode'i paremini ülejnäänud näidistest eraldatud kui fRMA puhul.

Võtmesõnad: Peakomponentanalüüs (PCA), hierarhiline klasterdamine, Barcode, geenieksпressioon, DNA kiibi andmed, fRMA.

CERCS:B110 Bioinformaatika, meditsiiniinformaatika, biomatemaatika, biomeetrika

Contents

1 Introduction

Over time, analysis of human genome has produced large amount of data, including gene locations and sequences. In addition to this, it is important to know which genes are expressed in which tissue types, to understand better the processes in each individual tissue type. Understanding processes that take place in each tissue type, can be used as a foundation to create more specific and more effective drugs to cure and prevent diseases both on individual and collective level. Detecting gene expression differences by tissue types requires standardized method to be successfully used on large data sets. One tool that helps give answer to that is gene expression Barcode for microarray data [1–3].

Barcode is a method that allows to determine whether a gene is expressed in a given cell type or not. Also, it allows comparison across multiple cell types since it takes different expression modes across cel types into account. Barcode method can also be used to identify new genes in a particular tissue that has not been well studied. Since Barcode is more immune to lab/batch effect than other methods, it can be applied to a single DNA chip to identify cell type. This property can be used to more efficiently classify unknown cells as cancerous because reliable comparison of data across different studies is possible.

Because it is a reliable approximation of the transcriptome, the Barcode data has been used in epigenetic studies, to improve ChIP-seq [4] and ChIP-chip [5] data analysis and to investigate increased heterogeneity in cancer. The barcode data is an important part of the EpiViz [6] webtool, which links transcriptomic and epigenomic data. The main bottlenecks for Barcode tool to be more efficient are limited amount of public data and inconsistent user-supplied annotations and vocabulary used to describe samples, making computational curation of data annotation difficult.

In this thesis, to understand Barcode better, it is explained by presenting an overview of different Barcode versions. For each version a description of functionalities and possible uses are given with emphasis on new functionalities compared to the older versions.

Second, practical part of this thesis compares Barcode and fRMA [7] method(fRMA method output is the starting point for Barcode analysis). To compare these two methods human gene expression dataset of DNA microarray experiment results is used. The dataset contains expression data from 158 human tissue samples. Tissue samples are first manually clustered to use as a reference clustering in comparison of these two methods. Data is then analysed with both Barcode and fRMA. To visualise and compare the results, two statistical methods are separately used: Principal component analysis and Hierarchical clustering. For the results of both statisical analysis methods a detailed analysis is given. The analysis and visualisation is conducted using R and the relevant R packages.

There are three files added to the thesis which are mentioned in the appendices. Firstly, the R code (BcodeThesis.R) used in practical part for drawing plots and applying Barcode and fRMA methods. Secondly, the text file (celandtissue.txt) containing relevant information for R code to work, including manual clustering info. Finally, the text file (data.txt) containing Barcode processed data of sample tissues.

2 Relevant genetic background

2.1 Genomic structures

2.1.1 Nucleic acids

There are two types of nucleic acids: DNA and RNA. Deoxyribonucleic acid (DNA) is a biomolecule that carries genetic information. DNA consists of four different nucleotides: adenine - A, guanine - G, cytosine - C and thymine - T. DNA can be found in organisms mostly double-stranded, consisting of two anti-parallel DNA strands. The binding (annealing) of these two strands is complementary, meaning that the nucleotides that are connected are A-T and G-C, other pairings can only appear as a mistake. A-T connection has two hydrogen bonds whereas G-C bond has three, which makes G-C bonds stronger and more stable.

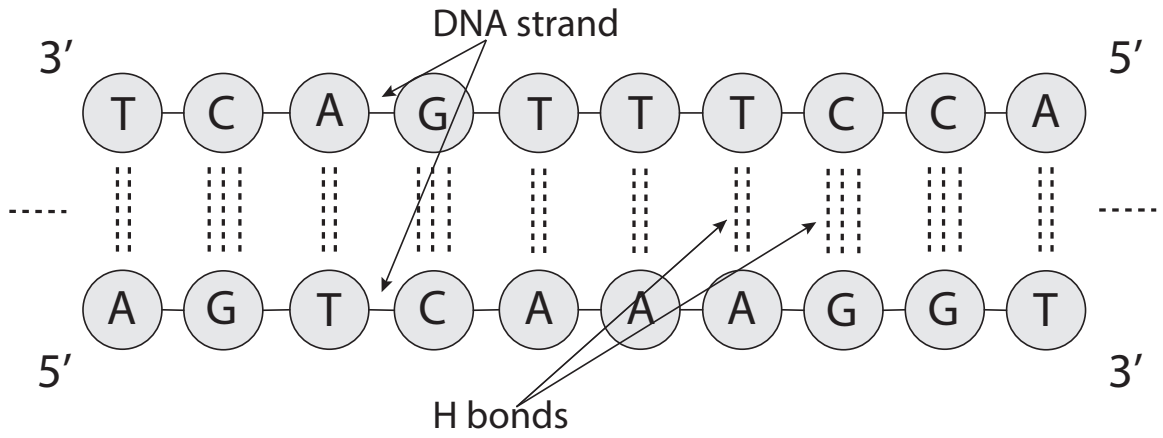


Figure 1: Part of double stranded DNA. It can be seen that G-C pairs have 3 hydrogen bonds whereas A-T pairs have 2 hydrogen bonds, which makes G-C pairs more stable. Also the two strands of DNA are anti-parallel - notice the 3' and 5' ends are reversed in the strands.

Ribonucleic acid (RNA) is a biomolecule that can act both as carrier of genetic information (only in viruses) and effector molecule for example, by catalysing biological reactions, controlling gene expression or mediating cellular signals. Unlike DNA, RNA is mostly found in single strands, which fold upon themselves to form complementary structure with itself. Similar to DNA, RNA consists of four different nucleotides, but instead of the T nucleotide, RNA has U (uracil) nucleotide, which means that binding pairs are A-U and G-C.

Both DNA and RNA are directional molecules. There are 5' ends and 3' ends in DNA and RNA. In double-stranded RNA (dsRNA) and dsDNA the two nucleic acid molecules are in opposite directions. In vivo DNA and RNA are synthesized from 5' end to 3' end. The relative positions of structures on the DNA strand, including genes, are referred to as upstream (towards 5' end) and downstream (towards 3' end).

2.1.2 Genome

Genetic material of organism is referred to as genome. It consists of DNA with the exception of some viruses, which have RNA genome. In eukaryotic (with nucleus, including

humans) organisms, genome is usually divided into one or more linear double-strands of DNA which is then further packed into higher structures. Genomes of all eukaryotic organisms contain two types of regions: gene regions, that consist of exons, introns and expression regulating sequences, and intergenic regions. The part of genome that is mostly under observation in genetic studies is exons. Exons are the parts of a genome that make up the mRNA.

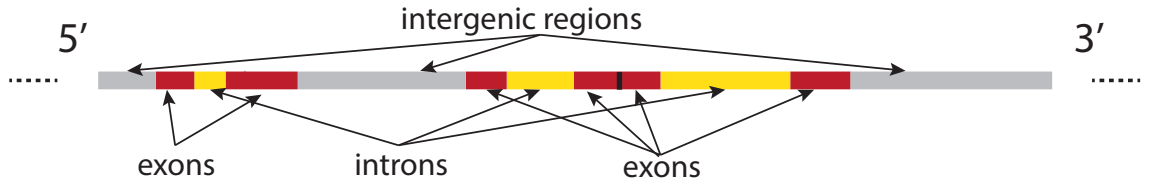


Figure 2: Genomic regions. Genome can be largely split into two: gene sequences and inter-genic sequences. Gene sequences consist of coding parts - exons and non-coding parts - introns and regulatory sequences.

2.1.3 Genome properties

The properties of primary and secondary structures of DNA/RNA are largely set by the nucleotide sequence. One such property, very important in genetic studies, is GC%. GC% means guanine-cytosine content in nucleic acid strand. DNA/RNA sequences with higher GC% provide more stable double-stranded molecule, for example, able to withstand higher temperatures. GC% may be measured in both shorter DNA sequences and the whole genome. GC% is calculated as follows: (G+C)/(A+T(U)+G+C) where A and T (U) are other two nucleotides present in DNA (RNA). In microarray studies, GC% affects sensitivity and binding specificity by increasing sensitivity through increase in signal intensity and decreasing binding specificity by being more prone to mismatch binding [8, 9].

2.1.4 Genes

Genes consist of sequence of nucleotides. The length and combination of nucleotides in a given gene defines the protein or some other gene product a gene codes. In addition to the sequence of a gene, it is important to know the location of the gene sequence in genome. In case of humans it is determined by giving number of chromosome and nucleotide range in this chromosome which belong to that gene. It is also added whether the gene is located on forward or reverse strand. For example, in Ensembl database for gene HLA-B the gene location is: Chromosome 6: 31,353,872-31,357,188 reverse strand [10].

2.2 Gene expression

What defines how organism looks like and functions, known as phenotype, is how its genes are expressed. One gene can have different variants, known as alleles, in different individuals. These variants code different gene products and when expressed, result in

different phenotypes. Gene is called expressed when it is used in the synthesis of a functional gene products. These gene products may be either proteins or functional RNA (for example tRNA, rRNA).

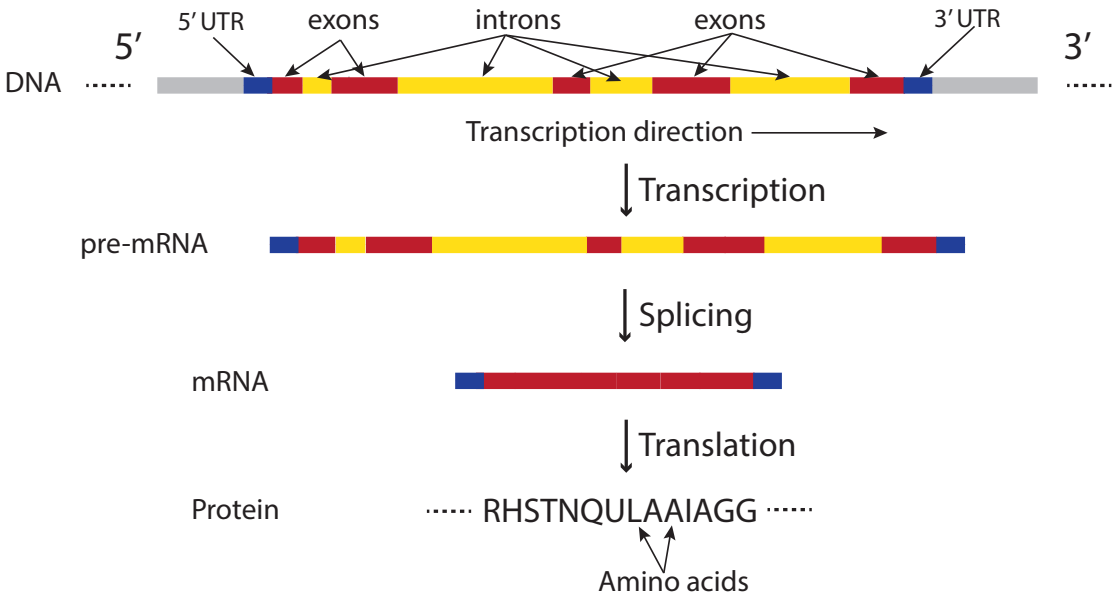


Figure 3: Gene expression. Genes are expressed mostly in direction DNA -> RNA -> protein. Firstly, based on DNA a pre-mRNA strand is transcribed. The pre-mRNA strand is basically the whole gene sequence. Then in the process of splicing, introns are cut away from pre-mRNA to produce mRNA. Finally mRNA in translated in ribosomes into chains of aminoacids that form proteins .

2.2.1 Expression mediators

Genes are expressed through RNA, more precisely, through messenger RNA (mRNA). mRNA is transcribed from the gene coded in DNA during a process called transcription. During transcription pre-mRNA, consisting of introns and exons is firstly transcribed in 5' -> 3' direction based on DNA template. The beginning and the end of the region to be transcribed are marked by untranslated regions (UTR-s) on the respective ends of the region. Then the introns are spliced away from the pre-mRNA and remaining exons are joined to form mRNA. Then the mRNA is translated into protein in cell components called ribosomes. In translation two other RNA types also play a key role: ribosomal RNA (rRNA) and transfer RNA (tRNA). Ribosomal RNA is one of the structural components of ribosomes. Transfer RNA mediates recognition of three subsequent nucleotides of mRNA, called codon, and provides the corresponding amino acid.

2.2.2 Splicing

Splicing is a process in which introns are cut away from the pre-mRNA and the remaining exon parts are joined to make up mRNA. Some regions of genome can code several genes because they have several different splicing sites. Which gene is expressed depends on which sites are used in splicing. When different sites are used, mRNA ends up with different sequence, meaning it will be used to translate different gene product, therefore a

different gene is expressed. The process of using different splicing sites to express different genes from the same region is known as alternative splicing.

2.2.3 Expression variability

Most genes are expressed differently across species, individuals and tissue types [11]. Also, in different populations and individuals, expression of the same gene in the same tissue can be different. This means that for the same gene in one tissue there may be several healthy expression modes. Therefore this within tissue variability has to be taken into account when defining whether the gene is expressed or silenced in the corresponding tissue [11].

There are some genes that are expressed very similarly across tissues. These are called high entropy genes.

2.2.4 Transcriptome

Transcriptome is a collection of gene transcripts (mRNA) of one cell, a cell population, organisms or even species. Transcriptome includes all genes that are being actively transcribed at any given time. Transcriptomes are profiled using DNA microarrays or RNA-seq. RNA-seq uses next generation sequencing methods to present RNA in a cell at a given moment.

Similar "-omes" are epigenome - collection of chemical compounds that interact with DNA [12], and proteome - collection of proteins expressed by a biological unit [13].

2.3 Unified representation of genetic data

Gene ontology (GO) [14] is an initiative to achieve unified representation of both genes and gene products across all species. This means that controlled vocabulary is developed and used to describe genes and gene products. Also, unified annotation is added to genes and gene products. Providing unified structure for the gene data makes it machine readable, allowing for much faster data analysis. GO helps represent gene and gene product properties by providing an ontology of terms for three domains: cellular component, molecular function, biological process. Each GO term has a term name, unique alphanumeric identifier, a definition with cited sources, and a namespace indicating what domain the term represents. In GO terms may also have synonyms. GO forms a directed acyclic graph where each term is in a defined relationship with one or many other terms. The relationships can be intra- or interdomain relationships and are species-neutral [15].

Annotating gene data using GO means assigning GO terms to the data. GO annotations form an annotation database, where there is also reference what was used to make the annotation (for example an article) and by whom the annotation was made. Both GO terms and annotations are dynamic databases, subject to changes and additions made by its collaborators.

Table 1: Example Gene Ontology term

id: GO:0016049
name: cell growth
namespace: biological_process
def: "The process in which a cell irreversibly increases in ..." [GOC:ai]
subset: goslim_generic
subset: goslim_plant
subset: gosubset_prok
synonym: "cell expansion" RELATED []
synonym: "cellular growth" EXACT []
synonym: "growth of cell" EXACT []
is_a: GO:0009987 ! cellular process
is_a: GO:0040007 ! growth
relationship: part_of GO:0008361 ! regulation of cell size

2.4 Usage of microarrays in genetic studies

2.4.1 DNA microarrays

DNA microarray, also referred to as Gene chip or DNA chip, is a series of microscopic DNA spots attached to a solid surface. In AffyMetrix Genechip arrays Each spot has $1.4 * 10^6$ short single-stranded DNA or RNA fragments [16], known as probes (also as oligos), attached to it. In gene experiments the probes can be designed differently, for example to identify unique transcripts or common transcript sequence segments.

Probes are used to hybridise a cDNA sample, called target, under specific conditions. cDNA samples are generated from mRNA by reverse transcription [16]. The phenomenon, called hybridisation, that microarray studies utilise is the ability of single-stranded DNA and RNA sequences to bind (anneal) to complementary strands. In case of microarrays, the hybridised single strands are probe and target.

Targets have usually some fluorophores attached to them, for example biotin [16] or molecules that are later used to bind fluorophores. The fluorophores or fluorophore binders used are usually small non-protein organic molecules so that they do not perturb the hybridisation.

The targets are then injected to microarray to hybridise, after which the weakly- and non-hybridised targets are washed away. When the fluorophore has not been previously attached to the targets then it is done in the process of washing. Probe-target hybridisation is detected and quantified by measuring fluorescence intensity caused by targets present on the spot after washing. More intense fluorescence means more probes have hybridised, meaning stronger binding [16]. Stronger binding generally means more precisely matched probe-target and therefore more probable expression of the corresponding gene.

Then the used fluorophore, fluorescence intensity, probe and target of each spot and wavelength used to get fluorescent reaction are all recorded, making up the raw microarray data open for further analysis [16].

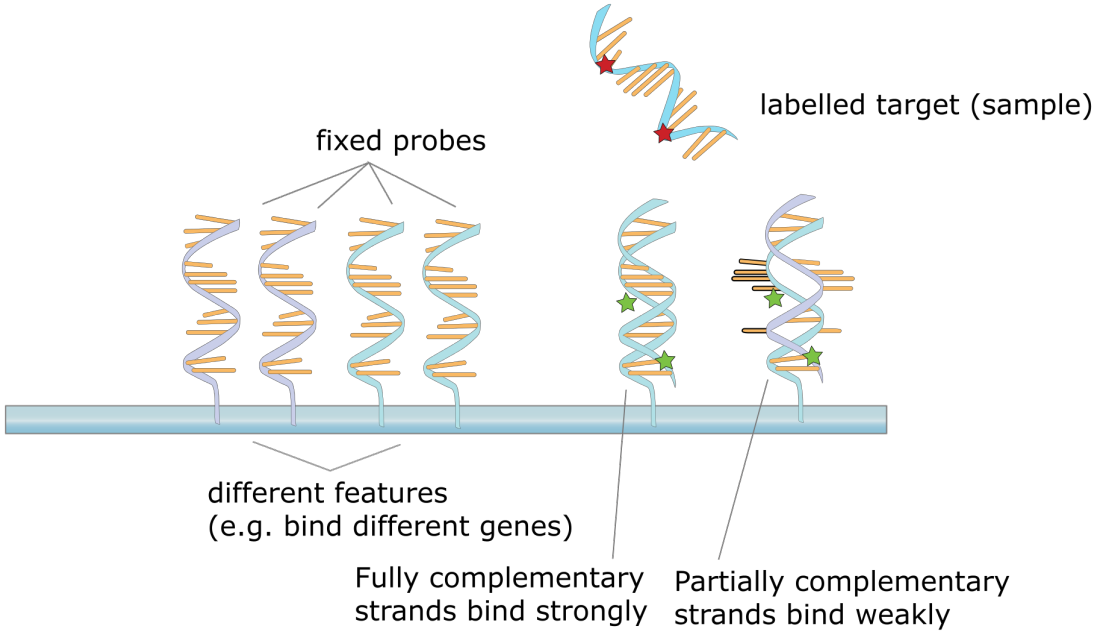


Figure 4: Microarray probe-target hybridisation [17]. Targets have fluorophores attached to them. The fluorophores are usually small non-protein organic molecules so that they do not perturb the hybridisation. The targets are added to microarray to hybridise, after which the weakly- and non-hybridised targets are washed away. Probe-target hybridisation is detected and quantified by measuring fluorescence intensity caused by targets present on the spot after washing. More intense fluorescence means more probes have hybridised, meaning stronger binding. Stronger binding generally means more precisely matched probe-target and therefore more probable expression of the corresponding gene.

2.5 Processing microarray data

For raw microarray data to present any real information, it has to be thoroughly processed first. The data is subject to many biases, like lab/batch effect, cross-hybridisations and poorly performing probe sets that have to be removed first. To remove these biases, negative control experiments are usually done. Negative control experiments ensure there is no effect where there should not be and also define the value of the experiment which refers to no effect. Even after trying to remove these biases, processing methods make mistakes, which can be used to compare their effectiveness. For example, false positive rate (false expressed gene calls made out of all expressed calls) can be used for that purpose.

2.5.1 Lab-batch effect

Lab effect or batch effect means that measured results from the samples depend on the lab/batch the samples came from [18]. Lab/batch effect is most commonly caused by small differentiations in how the batch of samples was made. The lab/batch effect can be found by running a control experiment on the sample from the same batch. If the effect is known then it can be computationally removed from the measurements which makes it possible to produce meaningful data.

"Lower levels" for lab/batch effect are array effect and probe effect. Array effect is caused by differences single arrays have, compared to each other. Probe effect is due to

differences in probe properties, for example different probe lengths or GC%.

2.5.2 Cross-hybridising

Cross-hybridising is a probe-target hybridisation phenomenon in which the target binds to the probe to which it was not intended to bind [19]. The wrong probe has rather similar nucleic acid sequence to the intended probe, resulting in a weaker hybridisation, but strong enough not to be washed away. This false binding causes background noise, interfering with measurement interpretation [19]. To define the background noise values, a negative-control experiment can be conducted. This means that all targets would provide only probe-target hybridisations that are by nature cross-hybridisations, giving the default background noise values to be used in calculations.

2.5.3 Poorly performing probe sets

Probe set is called poorly performing when it provides too many false-positive or false-negative results. The causes for a probe set to perform poorly are most likely incorrectly chosen probe and/or target sequences leading to cross-hybridisation or no hybridisation at all. The probe set can also appear to perform poorly when background noise is falsely read or expression-threshold values are set incorrectly.

2.5.4 Purified cell types

One way to reduce potential biases in microarray experiments is to use probe/target material obtained from purified cell types [20]. Purifying a cell type means isolating a cell population of same phenotype. By adding identifying markers to the cells of interest, they can be extracted from cell suspension to form a purified cell population with the phenotypic trait of interest [20], for example cancerous cells. Cell types can also be purified based on physical traits, for example such as weight - purified with centrifugation, or magnetic bead separation which uses antibody-antigen binding [20].

2.6 Other relevant methods of genetic studies

2.6.1 Next generation sequencing

Second generation sequencing, or as it is more frequently used, next generation sequencing (NGS) techniques offer an alternative to microarrays. NGS allows direct RNA, DNA analysis. NGS includes ultra-high throughput sequencing technologies, which allow for much faster sequence analysis than microarrays. [21] The information gathered from a sample with these new techniques is comparable to that in a single array enabling identification of differentially expressed genes. In addition to that, these new methods allow better discovery of low-expressed genes, alternative splice variants and new transcripts [22]. For RNA-s the technique that uses NGS is called RNA-seq.

2.6.2 ChIP technologies

ChIP-seq, short for chromatin immunoprecipitation sequencing is a method that analyses protein-DNA interactions [4]. Mostly it is used to analyse how chromatin-associated proteins like transcription factors and other regulate gene expression. The DNA sites that interact with these proteins can be isolated by chromatin immunoprecipitation.

Immunoprecipitation uses an antibody that specifically binds to a protein to precipitate it out of a solution. These sites are then used to combine a library of DNA sites bound to the protein of interest in vivo. Sequencing of the DNA fragments is done simultaneously using a genome sequencer. ChIP-chip [5], also known as ChIP-on-chip is technology similar to ChIP-seq, but in this case the DNA fragments are analysed using microarray technology, hence the "chip" part in the name.

2.7 Microarray analysis methods

2.7.1 RMA

RMA, short for robust multiarray average is a preprocessing algorithm for Affymetrix gene expression microarrays [23]. RMA allows background correction, normalisation and summarisation in a modular way. Normalisation and summarisation require several arrays to be simultaneously analysed. The ability to use information across samples allows RMA to produce a fitted parametric model for probe effects and improve outlier detection. Using this fitted parametric model allows the quality metrics to be set [7]. RMA cannot be used to process arrays individually or in small batches, which hinders its use in clinical studies. Also, data sets that are preprocessed separately cannot be compared [7]. RMA does not use information gathered from use of mismatch probes. Mismatch probes are different from perfect match probes by one base at central position to serve as control probe for cross-hybridisation.

2.7.2 Frozen RMA

Frozen RMA (fRMA), similarly to RMA, is a microarray data analysis tool. Unlike RMA, fRMA allows analysing individual or small microarray batches [7]. This is achieved by using information from large publicly available microarray databases. Based on the publicly available data, probe-specific effect estimates and variances are computed and frozen. When analysing new data sets, these frozen values are used to normalise and summarise the data from this new set. fRMA single array results are comparable with RMA batch results and when analysing multiple batches, fRMA outperforms RMA by removing batch effect.

2.7.3 PAM

Predictive Analysis of Microarrays (PAM) [24] is a statistical class prediction tool for gene expression data which uses nearest shrunken centroids method [25]. Nearest shrunken centroids method defines gene subsets that are best used to describe the corresponding gene classes. The method finds a standardised centroid for each gene class. The standardised centroid is the average gene expression value for each different gene in each class and it is divided by within-class standard deviation for the specific gene. Each new sample is classified by taking its gene expression profile and comparing it to class centroids. The closest class centroid, measured in squared distance, is the predicted class of that new sample. Shrunken centroid means that each class centroid is moved ("shrunk") towards the overall centroid for all classes by a user-defined threshold amount, always moving it towards zero, but never past it. For example, centroids with values of 2.7 and -3.8 with threshold value 2.0 would be moved to 0.7 and -1.8 respectively. Centroid with value 0.8

would be moved to zero. This kind of shrinking reduces effect of noisy genes and allows for some automatic gene selection [25].

3 Gene expression Barcode

Over time, analysis of human genome has produced large amount of data, including gene locations and sequences. It is important to know which genes are expressed in which tissue types, to understand better the processes in each individual tissue type. Understanding processes that take place in each tissue type, can be used as a foundation to create more specific and more effective drugs to cure and prevent diseases both on individual and collective level. Detecting gene expression differences by tissue types requires standardized method to be successfully used on large data sets. Without standardisation and large sets of reference data, studies are unable to reliably define expressed genes, because they are prone to produce false positives, especially medical studies, where data sets tend to be small. One tool that helps solve that is Gene Expression Barcode for microarray data [3]. Barcode eliminates data shortage by providing standardised data from public data sets. This means that by using Barcode, experiments can be conducted on only a few samples at once.

3.1 Barcode 1.0

3.1.1 Defining expression thresholds

In Barcode 1.0 [1] tool microarray experiment value thresholds that indicate gene expression are defined by gathering public datasets of microarray gene expression experiments in which genes are already stated as expressed or silenced. Raw data is obtained for more than hundred experiments and pre-processed with RMA 2.7.1.

Then for each gene the median log_2 expression estimate is computed and an empirical density smoother is used to estimate the expression distribution of that gene across tissues. Cross-validation [26] is used to choose the smoothness parameter. The mode with the lowest intensity is considered the expected intensity of an unexpressed gene. Estimates with even lower intensity are used to define the standard deviation of unexpressed genes. Estimates that are six or more standard deviations (defined by cross-validation assessment) larger than unexpressed mean correspond to expression. Only genes with two or more distribution modes are included to exclude repetitive information. Genes with only one mode are considered either expressed or unexpressed in all tissues.

Figure 5 shows the estimations for two human genes. The vertical line shows the cut-off between expressed and unexpressed intensity. Box plots on the right show clearly which genes are above the cut-off, meaning expressed.

Additionally, these genes have manufacturer provided expression calls: A - absent, M - marginal, P - present. The calls are marked with lines above and below the density plots. The orange lines represent absent calls and blue and green represent marginal and present calls respectively. Figure 5 shows that expression calls based on single-array data are inconsistent with Barcode and not useful.

3.1.2 Original prediction algorithm

In Barcode 1.0 [1] the expression barcodes are produced by first assigning each sample a specific barcode - a vector of zeros and ones, each number representing one gene. Based on whether the gene is considered expressed or not the number is set 1 or 0. For defining expression states of samples, previously defined expression thresholds are used. Then the barcode for each tissue is set by averaging the zeros and ones of the samples. This

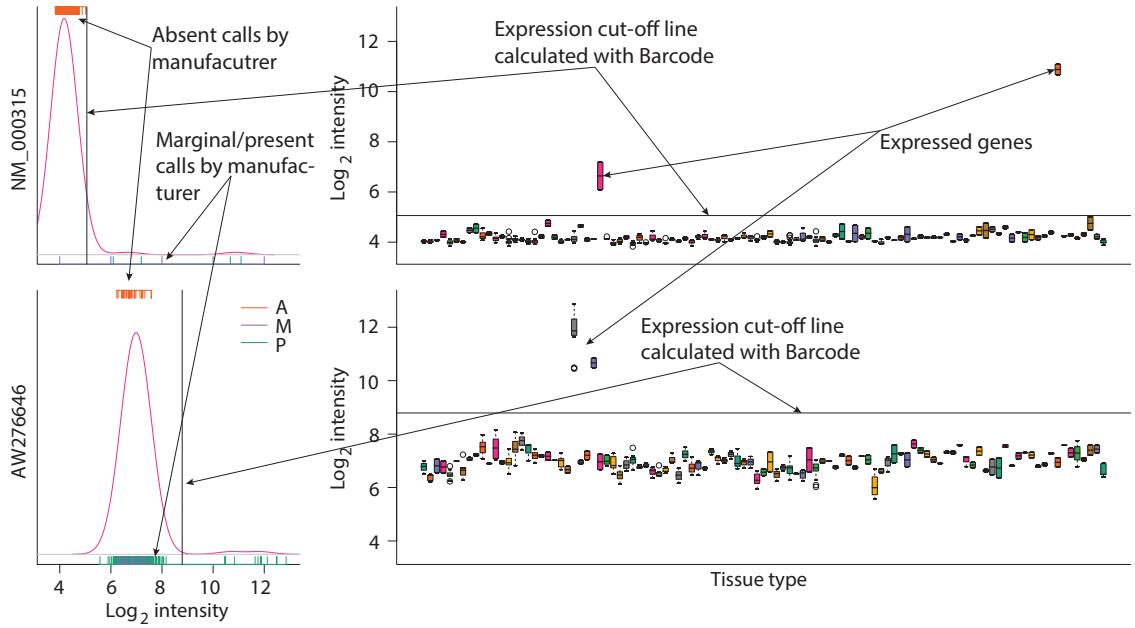


Figure 5: Defining of expression threshold, Barcode 1.0. The vertical line shows the cut-off between expressed and unexpressed intensity. Box plots on the right show clearly which genes are above the cut-off, meaning expressed. Additionally, these genes have manufacturer provided expression calls: A - absent, M - marginal, P - present. The calls are marked with lines above and below the density plots. The orange lines represent absent calls and blue and green represent marginal and present calls respectively. The manufacturer provided expression calls based on single-array data appear to be inconsistent with Barcode and not useful [1].

means that tissue barcode can contain any value between 0 and 1. For classifying a new sample, its barcode is computed and then the barcode is compared with existing tissue barcodes by calculating the Euclidean distance from each tissue barcode. The tissue type, that minimises this distance is the predicted tissue type. Euclidean distance between two barcodes, and therefore samples, can be interpreted as the number of genes that are expressed in one sample and not in the other.

3.1.3 Testing

Barcode 1.0 tool is tested on six datasets including clinical data, by comparing the barcodes to predefined ones, where state of expression is already known. Using odd-one-out cross validation, Barcode 1.0 compared to tool PAM [25], outperforms PAM in all cases but two where it performs just as well. Performance in this case is measured in precision percentage on distinguishing normal tissue types from the disease types. Barcode 1.0 is tested vs PAM also on independent data sets not included in the original database. Using prediction algorithm Barcode 1.0 shows far better results than PAM and what is more important, Barcode 1.0 shows consistent results with the testing done on the six datasets whereas PAM failed to show that consistency. The reason for PAM failing on independent datasets but Barcode 1.0 not, is that Barcode 1.0 greatly alleviates lab/batch effect by being more immune to the changes in the intensity values and therefore chance of false-positives or false-negatives is reduced.

3.2 Barcode 2.0

3.2.1 Probability of Expression

Probability of Expression (POE) [27], is a statistical algorithm that is used to model logarithmically transformed (log_2) intensities in Barcode 2.0 and 3.0 [2, 3] tool. In Barcode tool a specific version of POE is used to obtain an estimate of silenced and expressed intensity values for each gene. [2] The version is as follows:

$$\begin{aligned}(y_{ijg}|\theta_{jg}) &\sim N(\theta_{jg}, \sigma_g^2)\\ \theta_{jg}|\mu_g &\sim (1-p_g) * N(\mu_g, \tau_g^2) + p_g * U(\mu_g, S_g)\\ \mu_g &\sim N(\xi, \lambda^2)\\ \tau_g^2 &\sim IG(\alpha, \beta)\end{aligned} \tag{1}$$

In the model, y_{ijg} is the observed fRMA log_2 intensity for gene g in sample i and tissue type j. For each gene/tissue type there is an average observed intensity found, which is marked as θ_{jg}. It is assumed that these average observed intensities follow a combination of normal distribution for silenced values and an uniform distribution for expressed values. These are marked as $N(\mu_g, \tau_g^2)$ and $U(\mu_g, S_g)$ where S_g is the interval from the silenced mean to saturation. It is also assumed that silenced means (μ_g) are from a normal distribution and silenced variances (τ_g^2) are from an inverse gamma distribution.

3.2.2 Defining expression thresholds

In Barcode 2.0 tool defining expression thresholds is done by gathering public datasets of microarray gene expression experiments from which it is possible to create large database of (binding) intensities. The aim is to create silenced distribution for each gene, which means that there has to be some experiment in which all genes are silenced. For that negative control experiments with yeast are conducted. The yeast RNA hybridizes to the probes on the human microarrays similarly to cross-hybridizing human RNA, meaning any observed signal is the result of background noise and thus provides the experimental data in which all genes are silenced. Unlike barcode 1.0, the background estimation also takes into account the across-tissue distribution data from POE. By including these samples' data and using their average observed intensities it is possible to estimate the silenced distribution for each gene. Using Probability of Expression algorithm (POE) silenced and effectively expressed thresholds for each gene are set, allowing expressed genes to be marked as 1 when calculated value is above threshold and silenced as 0 when calculated value is below threshold.

POE takes into account variability within tissue, variability within expression state and uses information across genes to allow comparison across all genes and tissues. This addition, compared to Barcode 1.0, allows to have more tissue specific data which allows to more specifically identify tissue type of the target DNA, observe different expression modes of one gene across tissue types. POE also uses relevant expression info from other genes to call expression states.

3.2.3 Testing

To test Barcode 2.0, public data from GEO [28, 29] and Array Express [30] is used: HGU133a (13824 samples), HGU133plus2 (18656), Mouse4302 (9652). After manually

curated annotation 78, 131, 89 (respectively) different normal and cancer cell types (at least 5 for each tissue type) is left. In each cell type, for each gene the proportion of samples for which that gene is called expressed is computed. These values define estimated transcriptome for each cell type. For validation of these transcriptomes, genes expressed in CD4+ T cells, cerebellum, liver and skeletal muscle were grouped by gene ontology. Functional annotation clustering using DAVID [31, 32] shows that the most enriched biological groups are really those expected for a given tissue, for example muscle contraction related genes in skeletal muscle (Table 2).

Table 2: Functional annotation clustering using DAVID. Functional annotation clustering using DAVID shows that the most enriched biological groups are really those expected for a given tissue, for example muscle contraction related genes in skeletal muscle.

Liver		Skeletal Muscle	
GO Term	ES	GO Term	ES
Cellular ketone metabolism	26.2	Muscle Contraction	15.8
Monocarboxylic acid metabolism	16	Muscle Organ Development	9.1
Organic acid catabolism	15.7	Striated muscle tissue development	7.1
Steroid metabolism	11.4	Energy derivation by oxidation of organic compounds	5.9
Wound healing	10.7	Anatomical structure development	4.5

3.2.4 Performance

Barcode 2.0 shows to be very consistent in expression calls, where majority of genes are marked as 1 or 0 (expressed or silenced) and most of the non-calls are caused by minority of the genes allowing to assume that these are either due to poorly performing probe sets or these are high entropy genes.

Table 3: Comparison to other methods. Barcode 2.0 outperforms competing methods EBI , Bodymap, TiGER by finding more expressed genes (Expressed) and achieving lower false positive rate (FP%) [2].

Method	Tissue	Expressed	FP, %
Barcode	Kidney	761	13
TiGER	Kidney	320	13
EBI	Kidney	245	14
Barcode	Liver	695	21
TiGER	Liver	295	41
Bodymap	Liver	36	25

As seen in Table 3, Barcode 2.0 outperforms competing methods EBI [33] (determines whether gene is up- or down-regulated in cell type), Bodymap [34] (assesses expression

strength of genes in tissue type), TiGER [35](determines tissue type based on expression sequence tags which are short cDNA sequences that represent parts of expressed genes) by finding more expressed genes (Expressed) and achieving lower false positive rate (FP%). For these methods to be comparable, gene lists from each of the resources are obtained and with DAVID [31] gene identifiers are converted to Ensembl ID-s [36] for use with the second.gen RNA sequencing data which is then to be compared to the threshold rates and used to determine false positive rate.

3.3 Barcode 3.0

3.3.1 New data in Barcode 3.0

Barcode 3.0 triples the amount of data for platforms existent in Barcode 2.0 and thus improves the barcodes for these platforms. Barcode 3.0 also adds three new platforms. The change in the data used can be seen in Table 4.

Table 4: Changes in data for Barcode 3.0 [3]

Affymetric GeneChip	Barcode 2.0 sample number	Barcode 3.0 sample number
U133A	13824	23936
U133 plus 2.0	18656	63331
U133A 2.0	0	8528
Human Gene 1.0 ST	0	10309
MOE430 2.0	9652	32241
Mouse Gene 1.0 ST	0	10505

Addition of the newer, ST gene platforms makes it possible to extend barcode technology from 3' in vitro transcription to whole gene arrays. ST platforms represent microarrays where probes that are hybridised with targets from not only the 3' end but the entire gene sequence. These arrays have to be preprocessed a bit differently to distinguish between batch-effect susceptible probes and probes that target the exons involved in alternative splicing. For this, fRMA [7] implementation that includes both probe-effect and exon-effect parameters needs to be used.

3.3.2 Barcode 3.0 data quality

The data in public databases is submitted open vocabulary and open structure, therefore most of the public data used is not computationally curatable and has to be manually curated for the Barcode 3.0 [3]. Currently annotation data is collected and most useful text fields are identified. Then normal and tumor tissue type samples are manually identified for parameter estimation and also to be classified as tissue or purified cell type.

After the publication of Barcode 2.0, a single-array measure of quality was developed and used to show that 10% of publicly available HGU133a and HGU133plus2 microarray data is of poor quality. In Barcode 3.0 this quality measurement is used to filter poor quality arrays to improve estimates of the null mean and variance. User of Barcode 3.0

is allowed to set the quality threshold. Improved quality control and increase in input data provides improved estimates of barcode parameters and therefore better estimation of absolute gene expression.

Barcode 2.0 and 3.0 parameter estimates are similar. Estimates of the null means were highly correlated between platform existent in both Barcode versions and only 1% of the null means differed by >1. There were a few genes whose null mean changed by >2 between versions which could be due to some poor quality arrays in Barcode 2.0 training data or additional training data in Barcode 3.0 providing more accurate estimate of the null mean. Either way, it shows that considerable improvements can be made by improving quality of arrays and incorporating additional data.

3.3.3 Barcode 3.0 Bottom-Up research

Barcode data needs to be easily usable in studies that include only some genes at once. For this, Barcode 3.0 has new suite of data mining and analysis tools to allow researchers query the database for changes at individual gene level without being obscured by great amount of extraneous results.

Since each probe set works differently on an array, reliability and efficacy of each probe set must be taken into consideration. For this purpose, probe reliability evaluation is provided in graphical form and for efficacy evaluation user is provided with across tissue distribution of the corresponding gene, average entropy of probe set (reliability measurement) and a probe page to enable sharing among researchers. For example, there are nine probe sets for ESR1 gene on the u133 plus 2.0 microarray platform. When examining across-tissue distribution of these probe sets, only one of these probe sets, 205225_at achieves a z-score>5 (considered to be evidence of expression) in a variety of tissues. This is strong evidence that 205225_at is only one of these probe sets that can measure ESR1 expression [3].

By examining the distribution of average z-scores across tissues and cell types, abilities of different probe sets to detect gene expression can be compared and thus their suitability for the experiment can be evaluated by looking at expression distributions.

Barcode has two different search methods added. First, a researcher can identify the genes and experiments of interest and directly download the preprocessed data for analysis. Secondly, consensus data for tissues and purified cell types can be downloaded and compared, for example normal breast and breast tumors. To check for potential confounding effects from false positives, one could graph Affymetrix [37] control probe sets along with the gene of interest.

3.3.4 Barcode 3.0 Single-array results

As opposed to regular approach of checking expression of specific parameter in patients by pooling against each other patients who have the parameter expressed and those who do not, receiving patient specific (single-array) result in Barcode 3.0 tool can be done by looking at each sample independently and determining parameter status and then looking at other genes of interest for that sample. This removes the potential bias from pre-categorising patients allowing further subdivision to be easily and more correctly done and thus more differences to be determined.

As the genome, epigenome and proteome all interact with the transcriptome, the barcode estimations will be of interest to a broad community of researchers. The frma

R/BioC package [38] with the frmavecs data packages [39] for each supported platform, allows one to easily incorporate barcode data into one's own analyses.

3.4 Conclusion

Barcode tool allows to determine whether a gene is expressed or not in a given cell type. Also, it allows comparison across multiple cell types since it takes into account different expression modes. Barcode tool can also be used to identify new genes in a particular tissue that has not been well studied. Since Barcode tool is more immune to lab/batch effect than other tools, it can be applied to a single chip to identify cell type and it can be used more efficiently to classify cancer because it allows to perform reliable data-comparison across different studies. Additionally, it was shown that the genes are better clustered by tissue type rather than by species allowing to remove species-specific biases.

Because it is a reliable approximation of the transcriptome, the Barcode data has been used in epigenetic studies, to improve ChIP-seq and ChIP-chip data analysis and to investigate increased heterogeneity in cancer. The barcode data is an important part of the EpiViz [6] webtool, which links transcriptomic and epigenomic data. The main bottlenecks for Barcode tool to be more efficient are still somewhat limited public data and inconsistent user-supplied annotations and vocabulary used to describe samples, making computational curation of data annotation difficult. Barcode data can be incorporated in researchers analysis using the frma R/BioC package with frmavecs data packages for each supported platform.

4 Application of the Barcode tool

This section describes the application of Barcode tool - a practical part of this thesis. Firstly, the description of data used in application will be given and preprocessing of it is explained. Secondly, the applying of Barcode tool to data and obtaining comparison data is described. Thirdly, a detailed explanation is given on how differently processed data is compared in the application process. Lastly, interpretation and conclusions are presented.

4.1 Data description

The data used for application is obtained from a database of functional genomics experiments ArrayExpress [30]. The chosen data set of gene expression microarray experiment data, marked as E-TABM-145 in ArrayExpress, contains 158 different probe-target hybridisation intensity value vectors from 79 different human cell lines and tissues. Each vector has intensities for hybridisations with 22283 different probes. The samples are gathered with experiments done on Affymetrix GeneChip microarrays with probe set based on annotation package human genome hgu133a. Hence the type name Affymetrix GeneChip Human Genome HG-U133A (A-AFFY-33).

Each sample in this data set is represented with a CEL [40] file containing gene espression experiment data from the microarray of the respective sample. Each CEL file has a header in which the parameters of the file are described and an intensity section, which contains the calculated intensity of each pixel on the microarray. The inensity part of CEL files is fairly large with each containing more than 500 000 rows each of which contains five data points. The CEL files are accompanied by a sdrf file which has descriptive parameters for each of the CEL file eg. organism part, organism, clinical history.

4.2 Preprocessing data

Preprocessing of the CEL files is done in R [41]. CEL files are read into R using the ReadAffy function from the affy package [42]. Using ReadAffy, in this particular case without any arguments, produces an AffyBatch object, which is a class representation for the intensities from multiple arrays of the same CEL type.

From there fRMA analysis tool is applied to normalise and summarise the data based on frozen values which are already created in fRMA based on publicly available data sets. Finally, to obtain a matrix of gene-level exprerssion values, fRMA object is converted to ExprrssionSet object with the R command exprs(frmaobject).

4.3 Creating expression barcode

The barcode algorithm estimates which genes are expressed and which are unexpressed in given microarray data. Barcode uses fRMA-based expressionset objects as a starting point. Barcode object is created by determining whether the fRMA intensities from the new array are within the estimated distributions. Obtaining these fRMA intensities is described above.

By default the output of the barcode function is a vector of ones and zeros denoting which genes are estimated to be expressed (ones) and unexpressed (zeros). There are also

other options - LOD score [43] vector, z-score vector or p-value vector. In this thesis, the default option of zeros and ones is used.

4.4 Clustering processed samples and visualising the differences between fRMA and Barcode

In order to compare Barcode and fRMA, two visualisation tools were applied: principal component analysis (PCA) and hierarchical clustering.

4.4.1 Principal component analysis

PCA [44] is a tool for finding and visualising patterns in high-dimensional data. PCA finds the best possible characteristics, the ones that summarize the dataset as well as possible. The characteristics, called principal components, do not have to mean anything by themselves, these are just to bring out the patterns in data better. The first principal component bisects a data cloud with a straight line in a way that explains the most variance of the data. The second principal component cuts through the data orthogonal to the first, again in a way that covers most of the variance not explained by the first component. The third component would be orthogonal to the preceding components and fit the residuals from those, and so forth. The way principal components are designed, allows them to be compared and prioritised. Prioritising the components allows some more low-variance dimensions that do not carry any useful information to be easily dropped from further analysis.

In Figure 6 an example of dataset with two characteristics is shown. This dataset can be plotted as points in a plane. To bring out variation, PCA finds a new coordinate system in which every point has a new (x,y) value. The new axes, first and second principal components, do not actually mean anything by themselves, these are linear combinations of old characteristics that are chosen to explain as much variance as possible.

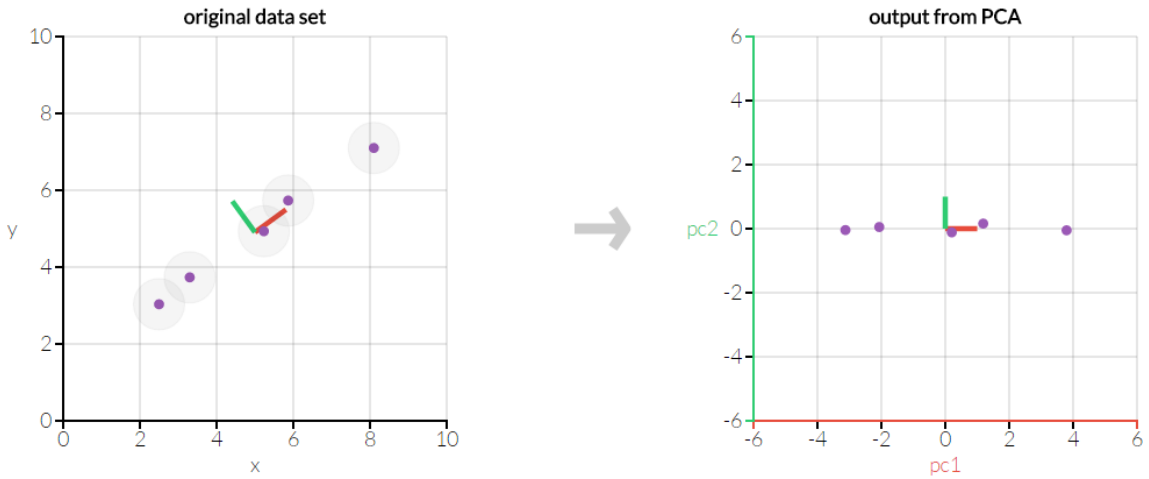


Figure 6: Principal components of 2D data. The dataset on the left has two characteristics and therefore can be plotted as points in a plane. To bring out variation, PCA finds a new coordinate system in which every point has a new (x,y) value. The new axes, first and second principal components, do not actually mean anything by themselves, these are combinations of old characteristics that are chosen to explain as much variance as possible [45].

With three dimensions, PCA is more useful, because it's hard to see through a cloud of data. In the Figure 7, it is shown how first two principal components project through 3D data. Basically what PCA has done is choosing the best angle to look at the data cloud and then project it to 2D using the first two principal components. The PCA transformation ensures that the horizontal axis PC1 has the most variation, the vertical axis PC2 the second-most, and a third axis PC3 the least.

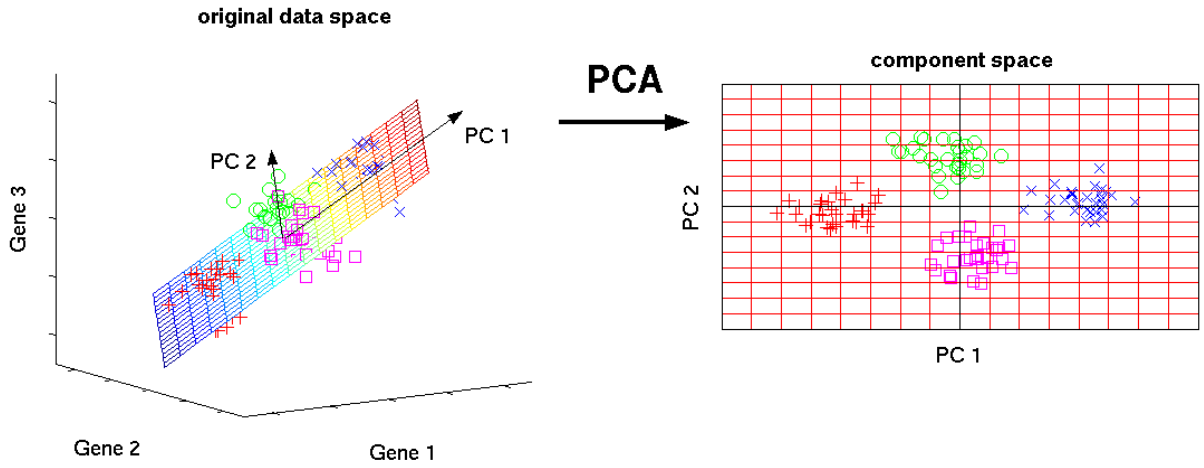


Figure 7: Finding first two principal components of 3D data. With three dimensions, PCA is more useful, because it's hard to see through a cloud of data. It is shown how first two principal components project through 3D data. Basically PCA has chosen the best angle to look at the data cloud and then projected it to 2D using the first two principal components. The PCA transformation ensures that the horizontal axis PC1 has the most variation, the vertical axis PC2 the second-most, and a third axis PC3 the least [46].

When performing PCA, it is useful to normalize the data first. Because PCA seeks to identify the principal components with the highest variance, if the data are not properly normalized, attributes with large values and large variances (in absolute terms) will end up dominating the first principal component when they should not. Normalizing the data gets each attribute onto more or less the same scale, so that each attribute has an opportunity to contribute to the principal component analysis. PCA in this thesis is done using R function "prcomp".

4.4.2 Plotting PCA

To compare and interpret the PCA plots better, the samples were manually grouped and labelled. Both morphological and functional properties of the tissues were taken into account when manual grouping was conducted. Functional properties were given higher priority in defining the group in which tissue belongs because these are expected to affect gene expression more than morphologic properties.

Manual grouping is provided only for samples that make up a group of at least 4 (2 unique tissue types). The groups are: Brain - 38 (samples), Blood - 24, Autonomous nervous system - 10, Covering epithel - 4, Blood processing - 6, Testis - 10, Lung - 4, Glandular cells - 16, Lymph system - 6, Adrenal system - 4, Tumor - 12. Each group is colour-coded. The samples that could not be grouped with the others are coloured as grey on the plots. Some of the samples are not grouped because of the lack of information

about the origin of the tissue, these are marked as "unclassified" and are not used in analysis of PCA plots.

The PCA plots themselves are designed so that the axis title is the number of the principal component it represents. For example, when axes titles are "1" and "2" then it means that the axes represent first and second principal components respectively.

4.4.3 fRMA PCA

Only first nine principal components were closely examined and plotted, because further components each explained 1% or less of total variance and did not provide any interesting information on the initial plots for all components. Of the first nine components only first, second, third and seventh component appear to separate some tissues. Rest of the nine components which do not show any separation are not displayed unless they appear on the same plot with the four mentioned components.

The only tissue that first component appears to separate is Blood (pink), visible on 1-2 plot top right corner. However, as 2-3 plot shows, second component does better job at separating Blood (middle right side). 2-3 plot also seems to separate Tumor and lymph system samples from other samples and also Autonomous nervous system and Brain samples, but these separations are much less clearer. Other tissue types do not appear to be separated clearly enough to be analysed.

As seen on 2-3 middle "row" and 3-4 plot middle "column", third component also does quite well in separating Blood, although not as well as second component. Third component provides the best separation of Brain (right side of the 3-4 plot and top left of the 2-3 plot), although the separation between Brain and Blood is obstructed because Brain cloud (blue) intersects with the Blood (pink) as seen on 3-4 plot.

Seventh component separates Testis as seen on 6-7 plot bottom left corner.

All in all, fRMA PCA plots do offer clear picture on the separation of Brain, Blood and Testis, but for other tissue types no clear separation appears. This means that fRMA PCA plots indicate that clustering of new (unknown) samples can be confidently done only for the three mentioned tissue groups but not for the others.

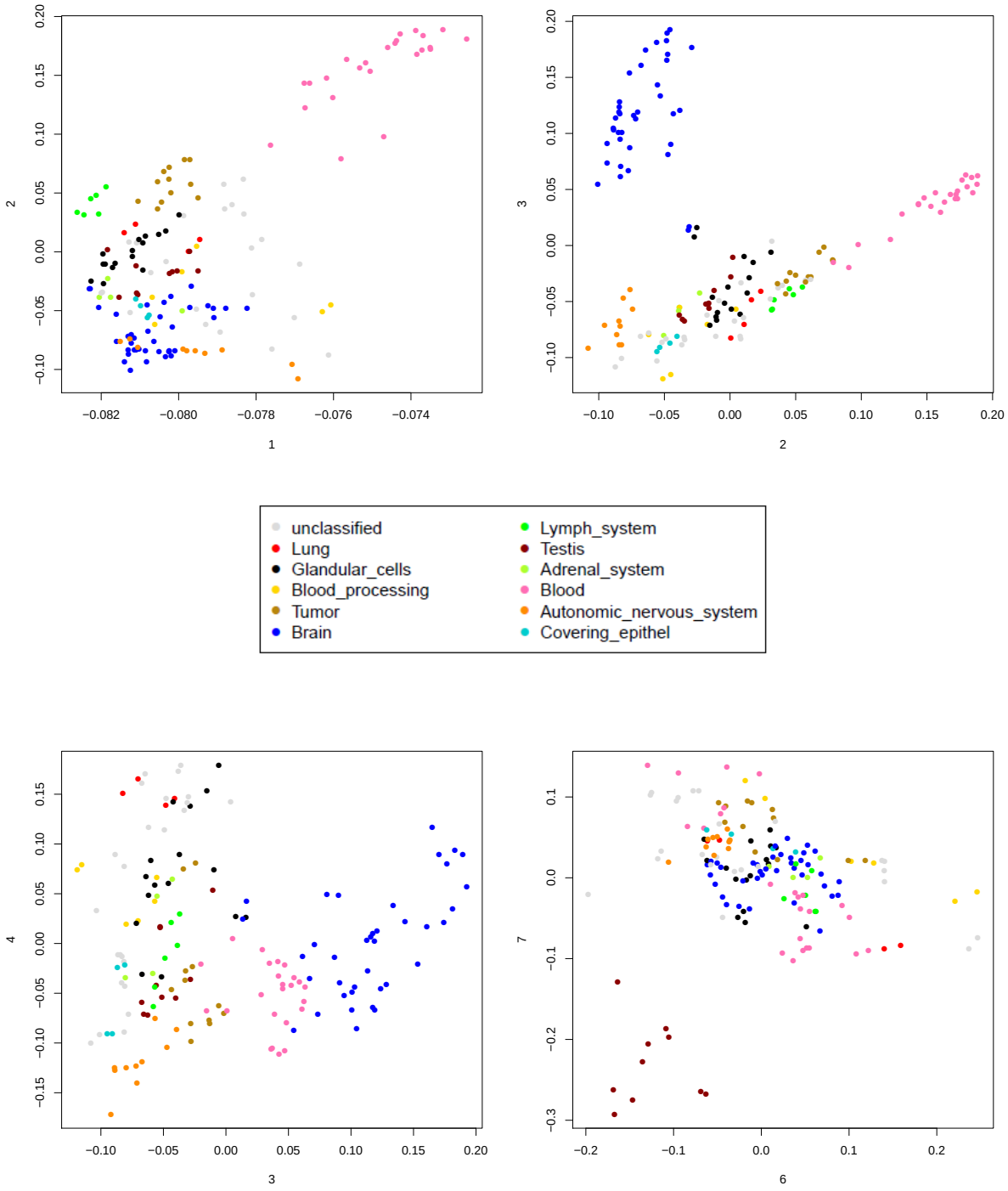


Figure 8: fRMA PCA, plots 1-2,2-3,3-4,6-7. First component separates Blood (pink), visible on 1-2 plot top right corner. However, as 2-3 plot shows, second component does better job at separating Blood (middle right side). 2-3 plot also seems to somewhat separate Tumor (brown) and lymph system (green) from other samples and Autonomous nervous system (orange) and Brain samples (blue). Third component also does well in separating Blood (plots 2-3, 3-4), although not as well as second component. Third component seems to do the best job at separating Brain (top left corner of 2-3, right side of 3-4). Seventh component separates Testis as seen on 6-7 plot bottom left corner. All in all, the separations between tissues are non-existent for most of the tissues and clearly visible only for tissues like Brain, Blood and Testis.

4.4.4 Barcode PCA

Only first nine principal components were closely examined and plotted for the same reasons as in fRMA PCA. On Barcode PCA plots it is clearer how tissue types are separated by components than it is on fRMA PCA plots. Of the first nine components only second, fifth and sixth components separate some tissues which is one less component than in fRMA case. Rest of the components are not plotted unless they appear on plots with the three mentioned components.

First component does not separate any tissues and does not appear to separate even groups of tissues. 1-2 and 2-3 plot show that second component separates three different tissues from each other and also from other tissues: Brain (blue) - bottom part of 1-2 plot and bottom left on 2-3, Autonomous nervous system (orange) - clear line in the middle of 1-2 plot, cluster in the top of 2-3, and Blood (pink) - top part of 1-2 and bottom right of 2-3. Also, second component separates Tumor and lymph system from other tissues. The second component of Barcode PCA corresponds to this of fRMA where second component separated the same tissues. In Barcode case however, the separation in Brain, Blood and Autonomous nervous system is visibly clearer.

On plots 5-6 and 6-7 it is visible that the fifth and sixth component both separate Testis from other tissues - bottom left of 5-6 plot and middle left of 6-7, and do the same for Tumor - top of the 5-6 plot and right side of 6-7.

All in all, as with fRMA, Barcode PCA plots separate clearly Brain, Blood and Testis. However, unlike fRMA, Barcode separates two more tissues clearly: Autonomous nervous system and Tumor. For other tissue types, the separation remains unclear.

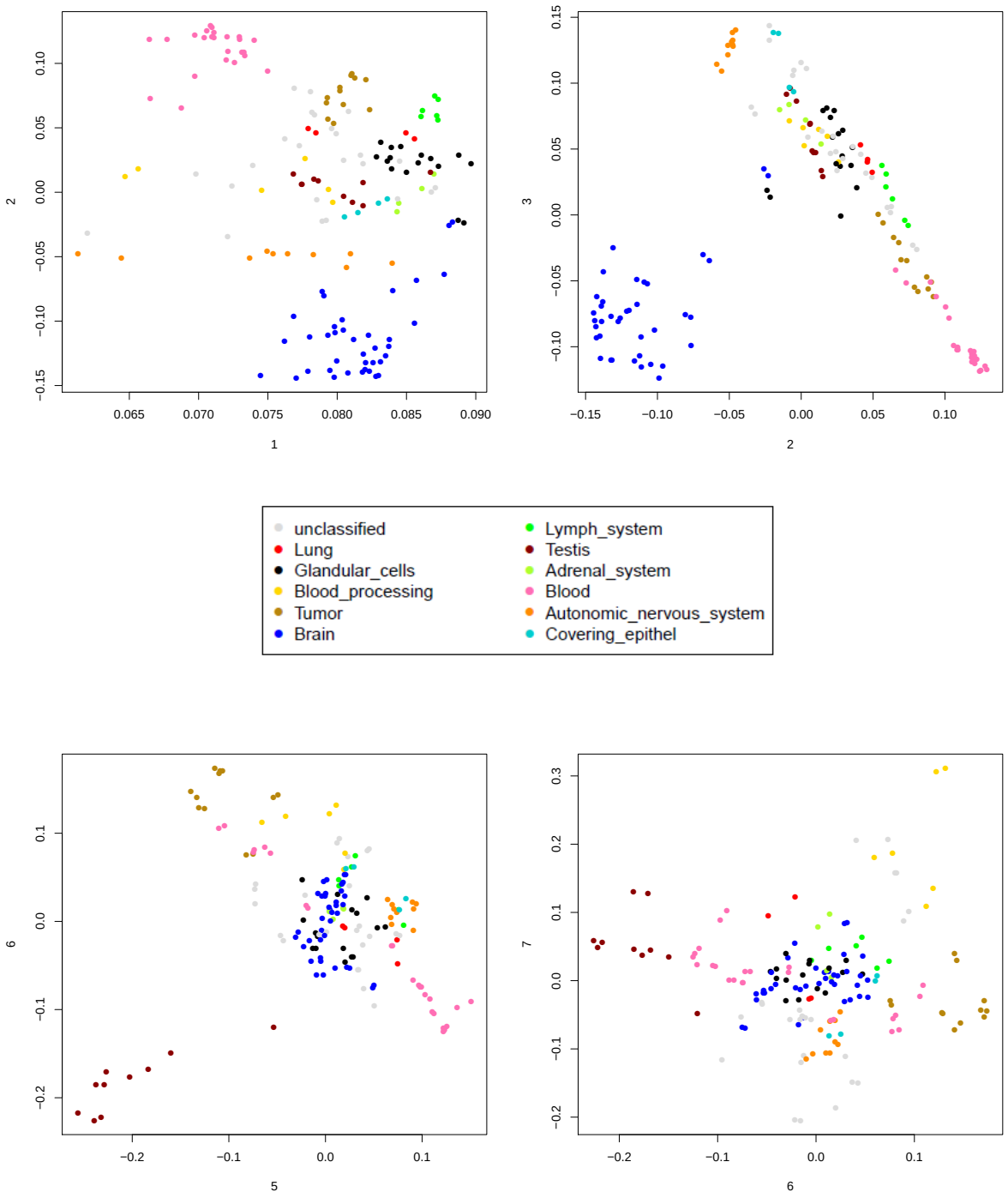
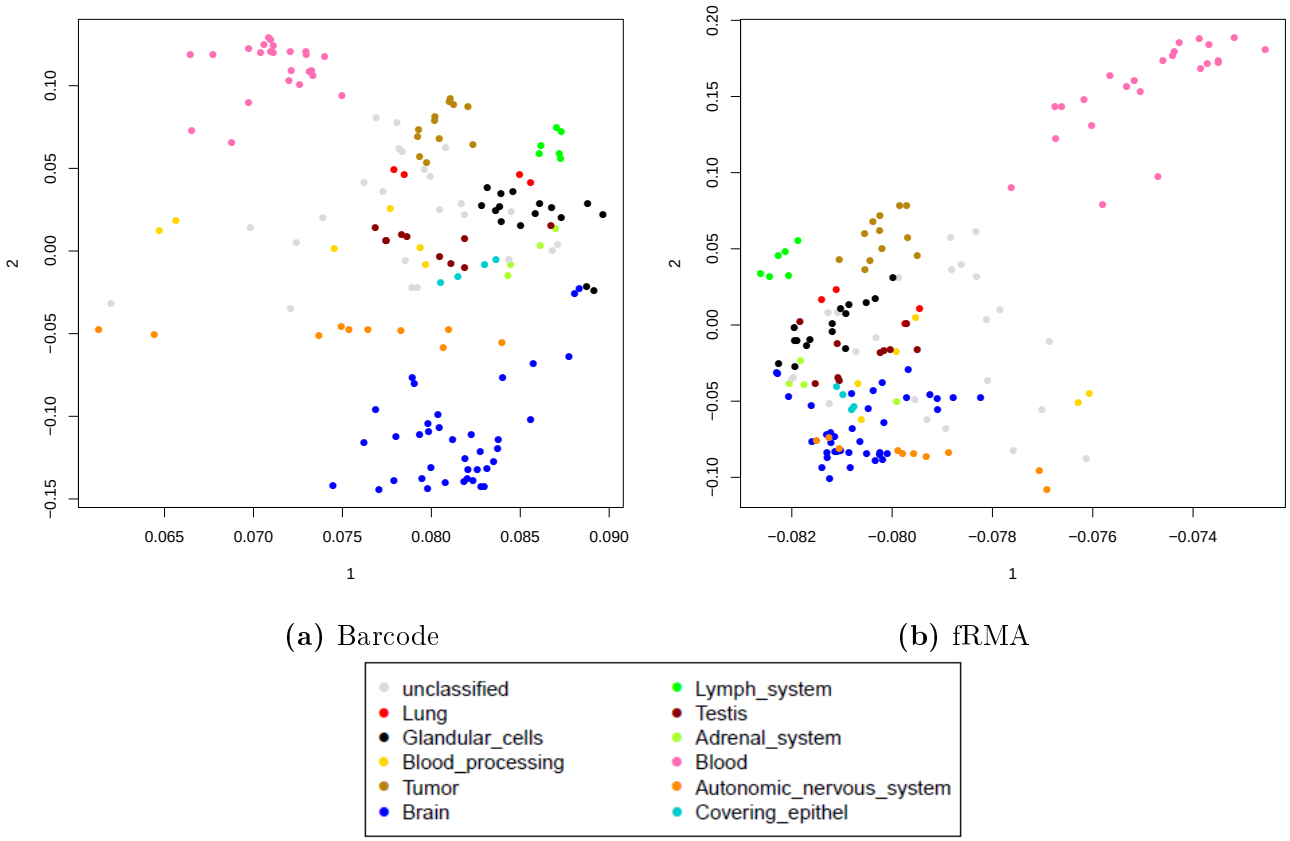


Figure 9: Barcode PCA, plots 1-2, 2-3, 5-6, 6-7. Plots 1-2 and 2-3 show that second component separates three different tissues: Brain (blue) - bottom part of 1-2 plot and bottom left on 2-3, Autonomous nervous system (orange) - clear line in the middle of 1-2 plot, cluster in the top of 2-3, and Blood (pink) - top part of 1-2 and bottom right of 2-3. Also, second component separates Tumor and lymph system from other tissues - towards right side of 2-3 plot. On plots 5-6 and 6-7 it is visible that the fifth and sixth component both separate Testis from other tissues - bottom left of 5-6 plot and middle left of 6-7, and do the same for Tumor - top of the 5-6 plot and right side of 6-7.

4.4.5 PCA comparison

Firstly, neither fRMA nor Barcode PCA plots offer satisfying results to whether the Barcode methodology is a clear improvement to fRMA or not. The PCA plots present themselves in too similar manner to draw any certain conclusions. There are however some differences.

As seen on Figure 10 the second component of Barcode PCA separates the same components as second fRMA component. On fRMA plot, bottom left corner, Brain and Autonomous nervous system are separated from other tissues but not from each other. In Barcode case however, the separation between Brain (blue), and Autonomous nervous system (orange) is visibly clearer - both in the bottom part of the plot. Separation for Blood (pink) is equal to fRMA.



(a) Barcode **(b)** fRMA

Figure 10: The second component of Barcode PCA separates the same components as second fRMA component. In Barcode case however, the separation between Brain (blue), and Autonomous nervous system (orange) is visibly clearer. Separation for Blood (pink) is equal.

None of the fRMA plots separates Tumor (brown), sixth component of Barcode PCA however separates Tumor clearly, as seen on Figure 11. This indicates that Barcode can be used to define unknown samples as Tumor or non-Tumor, which is important for medical use.

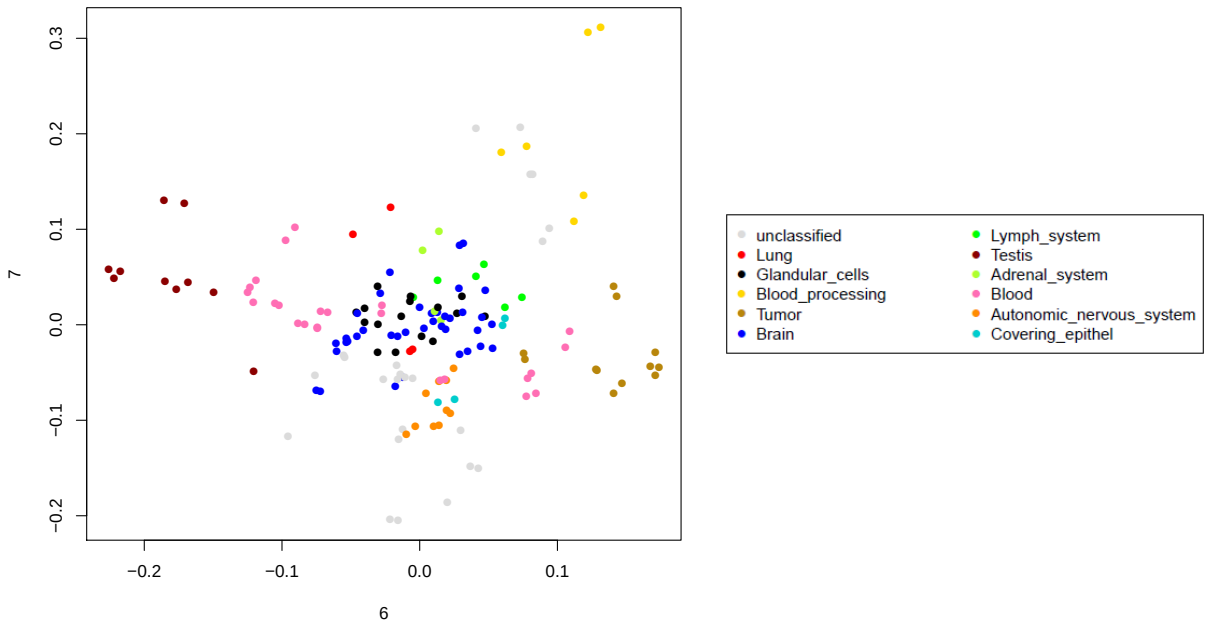


Figure 11: Sixth component of Barcode PCA separates clearly Tumor samples (brown) on the right side of the plot. This indicates that Barcode can be used to define unknown samples as Tumor or non-Tumor.

All in all, the PCA comparison of fRMA and Barcode methods remains unconclusive, because on one hand, comparison of Barcode and fRMA PCA plots shows Barcode to have better results in separating Brain, Blood, Testis, Tumor and Autonomous nervous system, last two being the difference clearly standing out between fRMA and Barcode. However, on the other hand it is not clear whether Barcode method improves, impairs or leaves unchanged the analysis of the tissues which are not separated by PCA.

Therefore, although Barcode has indications to better fRMA in Tumor sample detection and also for Brain, Blood, Testis and Autonomous nervous system, for conclusive results some other analysis method needs to be used too.

4.4.6 Hierarchical clustering analysis

Hierarchical clustering analysis (HCA) [47] is a clustering method that builds a hierarchy of clusters. The hierarchy can be built two ways - "bottom up" and "top down".

"Bottom up" means that each observation starts in its own cluster, and pairs of most similar clusters are merged as one moves up the hierarchy. "Top down" means that all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

In order to decide which clusters should be combined/where split, a measure of dissimilarity between sets of observations is required. Usually it is done by use of appropriate metric [47] (for example euclidean distance between pairs of observations) and linkage criterion (for example complete-linkage) setting the dissimilarity of sets as a function of the pairwise distances of observations in the sets. In general, the merges and splits are determined in a greedy manner. The results of hierarchical clustering are usually presented in a dendrogram.

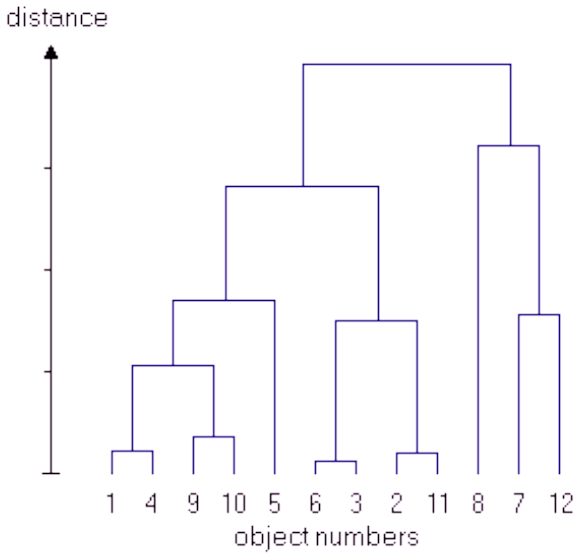


Figure 12: An example hierarchical clustering dendrogram [48].

Hierarchical clustering in this thesis is done with hclust function in R with correlation by Pearson correlation coefficient as distance measure. Hclust performs clustering by using a set of dissimilarities for the objects being clustered. Hclust implements iterative "bottom-up" clustering. At each stage when two objects i and j are clustered $(i \cup j)$, then the distances have to be recomputed between the new cluster and all other clusters. In hclust this is done by Lance-Williams dissimilarity update formula:

$$d(i \cup j, k) = \alpha_i d(i,k) + \alpha_j d(j,k) + \beta d(i,j) + \gamma |d(i,k) - d(j,k)| \tag{2}$$

$\alpha_i, \alpha_j, \beta$ and γ define the agglomerative criterion. Values of α, β and γ depend on the used clustering method. In case of complete-linkage method $\alpha = 0.5, \beta = 0$ and $\gamma = 0.5$. This results in

$$d(i \cup j, k) = \frac{1}{2} d(i,k) + \frac{1}{2} d(j,k) + \frac{1}{2} |d(i,k) - d(j,k)|$$

which can be rewritten as

$$d(i \cup j, k) = max\{d(i,k), d(j,k)\}$$

There are several different clustering methods provided in hclust. In this thesis complete-linkage method is used, because it avoids cluster chaining and separates clusters the most. Complete-linkage means that the distance between two clusters is defined by the greatest distance between any two elements from these clusters.

Hclust orders each subtree so that the tighter cluster is on the left (i.e visually, the cluster with the lower connection bar is tighter and therefore on the left). Single observations are the tightest clusters possible.

4.5 Plotting hierarchical clustering

To validate and compare the clusters of Barcode and fRMA, the same manual clustering was used as in PCA. The plots are based on hierarchical clustering and are designed to

check whether there are differences between fRMA and Barcode processing. The plots should be interpreted as follows. Firstly, labels of different tissue and cell types are coloured based on the manual labelling. The less disperse the samples of the same colour are, the better is the clustering. When looking at colours one has to also look whether the colours belong to the same larger cluster or are they in different bigger clusters side-by-side.

Secondly, the higher the bar connecting two samples (sample groups), the more different the samples are, meaning that new samples can be classified more confidently. The plots are drawn with ape package [49] from R.

4.5.1 fRMA clustering

Hierarchical clustering performed on fRMA data is shown on Figure 13. First thing to be noticed is that bars that connect samples are relatively low. This means the difference between the samples is not so big, which could lead to false-clustering and false classification of unknown samples. fRMA clustering provides two clearly separated clusters: Blood and Brain.

Brain samples appear very similar to each other and are clustered far from other tissues with the exception of olfactory bulb (responsible for the sense of smell) samples that are clustered very far from other brain tissues. This could be because olfactory bulbs function separately from other parts of brain and can be viewed as a relay station from cranial nerve I. However, because olfactory bulbs are located inside the skull and is often classified as part of brain, in this thesis olfactory bulb is manually grouped as a brain tissue. Because brain samples are so different from other samples, clustering samples as a brain can be done with high confidence. On the other hand, the similarity between brain tissues means when more exact tissue type is required, the clustering is not so confident at all.

Blood samples appear to be clustered better than brain samples. The cluster is rather far from other clusters, but the difference between different blood tissues is also greater and difference between same tissues is small relative to the difference of different blood tissues. This means that even the exact tissue type can be clustered confidently.

Other clusters that appear not so clearly, but are worth mentioning are Testis, Tumor and Lymph system. Testis samples are grouped together and relatively far from other tissues with the exception of one stranded sample, that cannot be explained. Tumor samples are also well clustered, however, there is a bronchial epithelial sample in this cluster, which should not be there and again, cannot be explained. Lymph system tissues are not so far apart from other tissues as Tumor and Testis, but still grouped together. There are two marrows samples also next to the lymph system cluster, which can be correct because some marrow cells have very similar functionality, but these marrow samples could not be grouped more specifically due to lack of information on them.

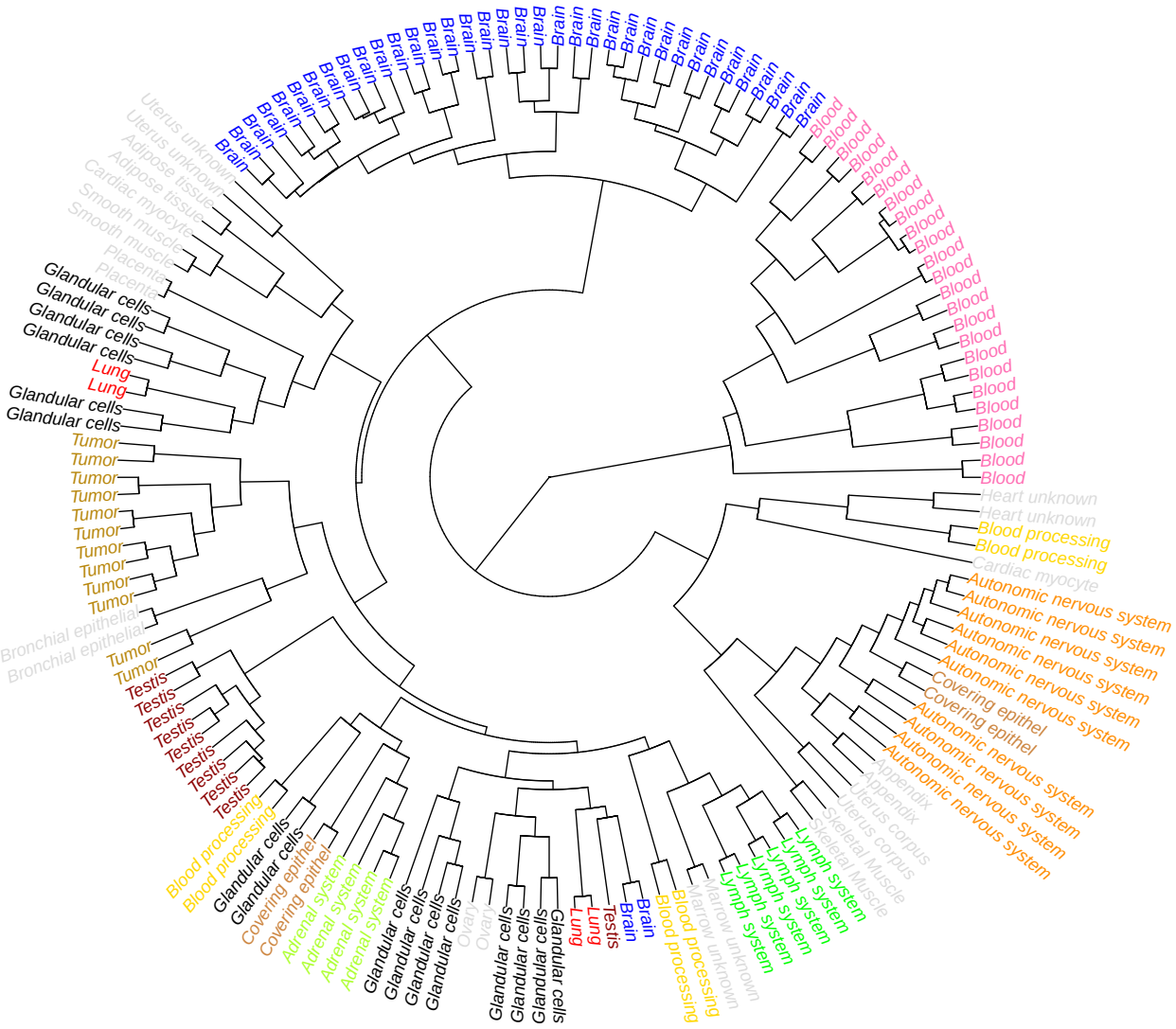


Figure 13: Hierarchical clustering of fRMA data. First thing to be noticed is that bars that connect samples are relatively low. This means the difference between the samples is not so big, which could lead to false-clustering and false classification of unknown samples. It is visible from the mixed order of colours that more similar tissues are not clustered well. The 150 to 315 degrees part of the clustering shows that tissues and tissue groups have very little difference. This has resulted for example in throwing glandular cell samples apart from each other. Also there are some tissues which appear to be displaced due to just having values that do not correspond to morpho-functional clustering, for example covering epithelial, blood processing and lung.

In Figure 13 it is visible from the mixed order of colours that more similar tissues are not clustered well. The 150 to 315 degrees part of the clustering shows that tissues and tissue groups have very little difference, because height of the bars connecting those clusters is small. This however means, that small differences in data can cause mixup of the clustering tree branches resulting in uninformative clustering. This has resulted for

example in throwing glandular cell samples apart from each other. Also there are some tissues which appear to be displaced due to just having values that do not correspond to morpho-functional clustering, for example covering epithelial, blood processing and lung.

4.5.2 Barcode clustering

Hierarchical clustering performed on Barcode data is visualised on Figure 14. First thing to be noticed is that bars that connect samples are high and even more so for different tissues. The difference between the samples is rather big, therefore reducing the probability of falsely clustering and classifying unknown samples. This allows to negate some of the problems that are present in fRMA clustering, therefore more meaningful clusters appear.

In case of Barcode clustering Brain samples again form their own cluster with the exception of olfactory bulb, which seems to support the hypothesis proposed for this anomaly in fRMA clustering section. Unlike fRMA clustering, Barcode separates Brain samples and different brain tissue types from each other almost as much as other tissue clusters. This still allows to confidently cluster samples as Brain but in addition allows to define the exact type of brain tissue confidently.

Blood samples again appear to be clustered better than brain samples, but the difference is much smaller. The cluster is still rather far from other clusters, with the difference between different blood tissues being enough for the exact tissue type to be clustered confidently.

Unlike fRMA, there are many more clusters that are grouped by their expected functional and morphologic traits. Most importantly Tumor samples are clearly apart from other tissues, which is useful for medical studies seeking to classify unknown samples as tumor on not tumor. Also Autonomous nervous system, Blood processing, Covering epithelial and Lung samples have achieved secluded clusters.

Glandular cell cluster appears deceptively to be close together when looking at labels. However, when looking at cluster tree, the glandular cell samples are divided into two clusters, where one group is more closer to adipose tissue, unknown uterus, testis and olfactory bulb samples. The clustering is much better than in fRMA where Glandular cells are scattered all over, but the division probably has some disruptive effect. Some part in division can be attributed to the fact that Glandular cell group is the most weakly bound of the predefined groups, but the effect should not be that big.

Based on the secluded colour groups on Figure 14, it is visible that samples are quite well clustered. Mostly the clusters make sense and are well secluded. However, there are still some anomalies such as cardiac myocyte samples being far apart, probably due to quality of data, or previously mentioned glandular cell division into two subclusters.

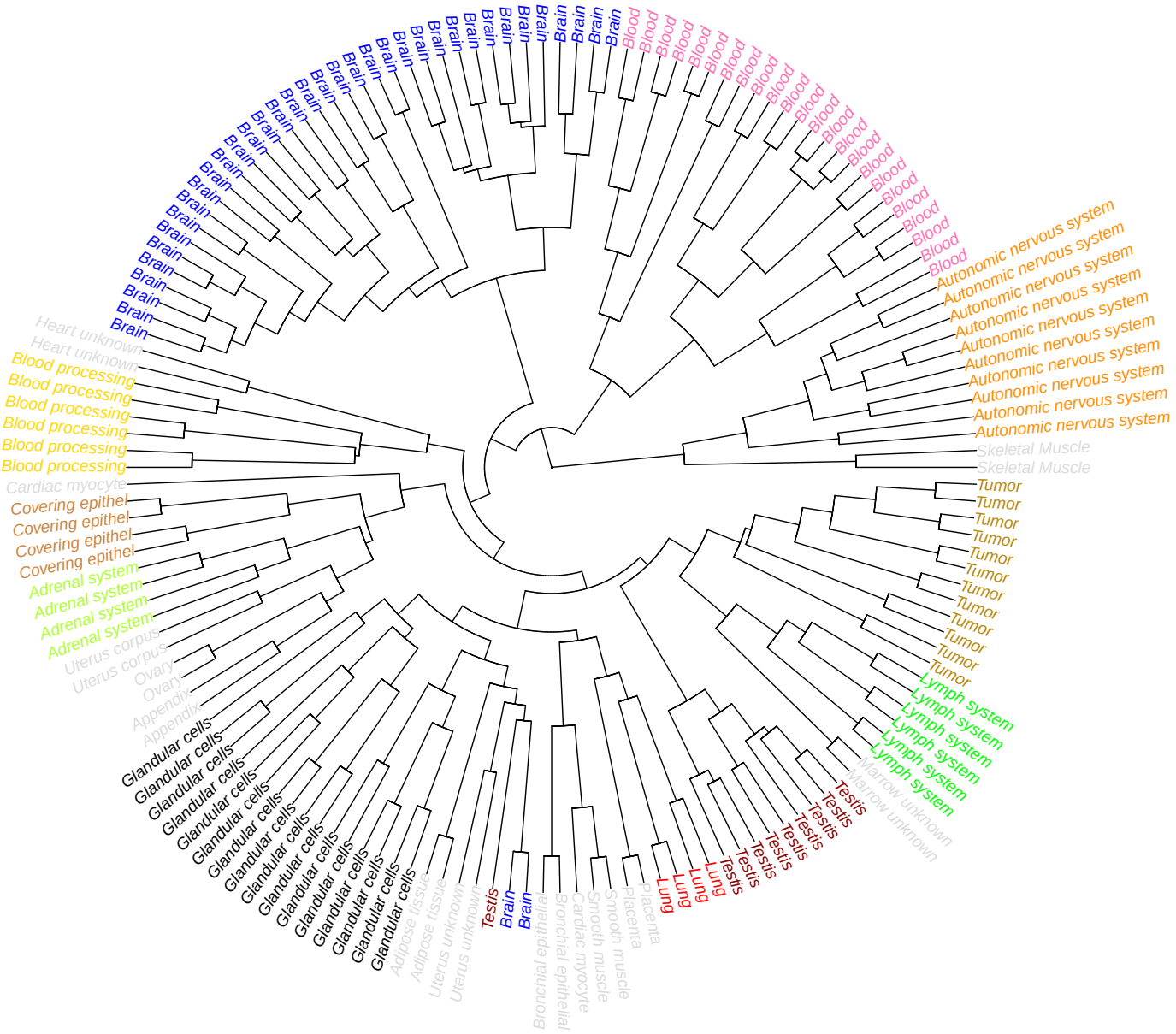


Figure 14: Hierarchical clustering of Barcode data. The bars that connect samples and show how much difference there is are high and even more so for different tissues. This allows to negate some of the problems that are present in fRMA clustering, therefore more meaningful clusters appear. Unlike fRMA, there are many more clusters that are grouped by their expected functional and morphologic traits. Most importantly Tumor samples are clearly apart from other tissues, which is useful for medical studies. Based on the secluded colour groups, it is visible that samples are quite well clustered. Mostly the clusters make sense and are well secluded.

4.5.3 Clustering comparison

Both fRMA and Barcode clusterings appear to be quite well grouped for functionally distinct tissue groups like brain and blood. However in case of smaller and more similar groups like glandular cells, blood processing tissues and lung, Barcode clustering clearly outperforms fRMA. The difference can be attributed to both Barcode separating samples more by converting expression values to binary and the improved expression detection algorithm in Barcode method.

Although visible in overall comparison of the plots, the ability of Barcode to separate samples better compared to fRMA is most drastically visible in Brain cluster. When in fRMA cluster the bars of all Brain samples appear to be more or less the same height, then Barcode cluster presents more jagged cluster, meaning that the difference of tissues is presented more clearly. Other cluster where it can be more clearly seen is Testis and Autonomous nervous system.

In terms of clustering tissues by their trait similarity Barcode outperforms fRMA. The difference is more clear in similar tissue groups. For example, in fRMA clustering the Glandular cells, Lung and Blood processing are so far apart that the improvement visible in Barcode cannot be attributed only to binarisation of the data. This presents also clear evidence that Barcode algorithm for expression detection is better than it is in fRMA.

All in all, based on the two clusterings it can be said that Barcode offers a considerable improvement to fRMA without any visible downsides.

4.6 Results

Both PCA and hierarchical clustering indicate that Barcode offers improvement in defining expression states of genes. However, the results of PCA visualisations alone are inconclusive. There are some small differences in how much Barcode and fRMA separate samples from different tissues and in that Barcode is visibly better in big sample groups. The revealed differences between Barcode and fRMA are not big enough to draw certain conclusions. More importantly, PCA plots cannot be used to determine whether the Barcode method compared to fRMA is better, worse or same on small sample groups. Therefore the differences that PCA reveals can be used as supporting results for more conclusive analysis results, but not sole evidence of Barcode superiority.

Hierarchical clustering offers far better overview of Barcode and fRMA differences than PCA. Firstly, because Barcode turns expression calls into binary, hierarchical clustering shows that each sample type is separated from other types clearly more than in fRMA case. Secondly, Barcode clusters sample types better based on the manual morpho-functional clustering of sample tissues created as a reference for this thesis. Both results are also supported by the finding of the PCA conducted in this thesis.

Separating each sample type more clearly both within and between tissue clusters allows new unknown samples to be clustered more precisely with more confidence and thus define the type of this unknown sample. This means that Barcode offers better results than fRMA in one of its main applications which is to define the tissue type of unknown samples for small sample batches. Also it enables to detect larger shifts in gene expression within one tissue for small sample batches which is very important in medical studies.

There is also one side-result found which was not the intended aim of this thesis. This results is also not mentioned in the original Barcode development articles.

5 Discussion

The thesis clearly shows that Barcode method is an improvement to fRMA method. This is supported by both PCA and hierarchical clustering results which show Barcode to be more efficient in separating samples within one tissue and also separating different tissue types from one another. This allows clearer lines to be drawn between tissue types and therefore observe anomalies in gene expression for a specific tissue better.

The manual grouping of sample tissues for reference in comparing Barcode and fRMA is simple on purpose. To avoid mistakes in grouping tissues, the tissue types which could not be confidently grouped were left out of the analysis, unless these were clearly in the wrong place. This makes it possible to claim fairly certainly that the tissues that are grouped are grouped correctly and thus the difference between Barcode and fRMA for these tissues can be assessed correctly. Any possible misgroupings are addressed in the respective part of the thesis.

During the thesis one side-result appeared: reduction in data size after Barcode processing. Most of the numbers used in gene expression microarray experiment data analysis are large floating point numbers. This kind of numbers take up quite a lot of data space. Barcode method allows this data to be binarised, meaning the space taken up would be much less. For data used in this thesis, Barcode output data takes up roughly ten times less space than both fRMA output data and non-standardised expression data. For large scale studies that base only on the knowledge of whether gene is expressed this reduction of data size would save up a lot of valuable space. Also the reduced data size would allow computational methods to retrieve the data with less effort and/or in less time.

The fact which slowed down the work was insufficient sample description. The reason that samples had to be grouped for reference manually and why some of the information on samples had to be manually added for clustering and PCA, was that samples did not have sufficient information of origin tissue type. Missing was information for linking the samples into larger groups and sometimes the sample description was too broad. For example, samples "marrow" did not have added specifications of which kind of bone marrow was the sample from. There were also plenty of brain samples, but none of these had information that the unifying group would be brain. The fact that uneven/insufficient information on samples hampers computational analysis and requires manual work was also mentioned in the original Barcode articles. The solution that was proposed there - introducing GO terms in more experiment data - would have also saved a lot time and effort in this thesis, therefore illustrating the need for unified representation of sample information for experimental data.

Barcode method has potential to be more widely used in not only bioinformatics studies but also in purely clinical and genetic studies. This would require the method to be implemented in software application other than as R package or a web application form. The software application would ease the use of Barcode method and would therefore widen the range of specialists that could be able to use the method.

References

[1] M. J. Zilliox and R. A. Irizarry. A gene expression bar code for microarray data. *Nat. Methods*, 4(11):911–913, Nov 2007.

[2] M. N. McCall, K. Uppal, H. A. Jaffee, M. J. Zilliox, and R. A. Irizarry. The Gene Expression Barcode: leveraging public data repositories to begin cataloging the human and murine transcriptomes. *Nucleic Acids Res.*, 39(Database issue):D1011–1015, Jan 2011.

[3] M. N. McCall, H. A. Jaffee, S. J. Zelisko, N. Sinha, G. Hooiveld, R. A. Irizarry, and M. J. Zilliox. The Gene Expression Barcode 3.0: improved data processing and mining tools. *Nucleic Acids Res.*, 42(Database issue):D938–943, Jan 2014.

[4] D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold. Genome-wide mapping of in vivo protein-DNA interactions. *Science*, 316(5830):1497–1502, Jun 2007.

[5] M. Zheng, L. O. Barrera, B. Ren, and Y. N. Wu. ChIP-chip: data, model, and analysis. *Biometrics*, 63(3):787–796, Sep 2007.

[6] F. Chelaru, L. Smith, N. Goldstein, and H. C. Bravo. Epiviz: interactive visual analytics for functional genomics data. *Nat. Methods*, 11(9):938–940, Sep 2014.

[7] M. N. McCall, B. M. Bolstad, and R. A. Irizarry. Frozen robust multiarray analysis (fRMA). *Biostatistics*, 11(2):242–253, Apr 2010.

[8] H. C. Yang, Y. J. Liang, M. C. Huang, L. H. Li, C. H. Lin, J. Y. Wu, Y. T. Chen, and C. S. Fann. A genome-wide study of preferential amplification/hybridization in microarray-based pooled DNA experiments. *Nucleic Acids Res.*, 34(15):e106, 2006.

[9] K. Kucho, H. Yoneda, M. Harada, and M. Ishiura. Determinants of sensitivity and specificity in spotted DNA microarrays with unmodified oligonucleotides. *Genes Genet. Syst.*, 79(4):189–197, Aug 2004.

[10] http://www.ensembl.org/homo_sapiens/gene/summary?db=core;g=ensg00000234745;r=6:31353872-31357188, Last visited 12.05.2016.

[11] E. Y. Alemu, J. W. Carl, H. Corrada Bravo, and S. Hannenhalli. Determinants of expression variability. *Nucleic Acids Res.*, 42(6):3503–3514, Apr 2014.

[12] B. E. Bernstein, A. Meissner, and E. S. Lander. The mammalian epigenome. *Cell*, 128(4):669–681, Feb 2007.

[13] M. S. Kim, S. M. Pinto, D. Getnet, R. S. Nirujogi, S. S. Manda, R. Chaerkady, A. K. Madugundu, D. S. Kelkar, R. Isserlin, S. Jain, J. K. Thomas, B. Muthusamy, P. Leal-Rojas, P. Kumar, N. A. Sahasrabuddhe, L. Balakrishnan, J. Advani, B. George, S. Renuse, L. D. Selvan, A. H. Patil, V. Nanjappa, A. Radhakrishnan, S. Prasad, T. Subbannayya, R. Raju, M. Kumar, S. K. Sreenivasamurthy, A. Marimuthu, G. J. Sathe, S. Chavan, K. K. Datta, Y. Subbannayya, A. Sahu, S. D. Yelamanchi, S. Jayaram, P. Rajagopalan, J. Sharma, K. R. Murthy, N. Syed, R. Goel, A. A. Khan, S. Ahmad, G. Dey, K. Mudgal, A. Chatterjee, T. C. Huang, J. Zhong, X. Wu, P. G. Shaw, D. Freed, M. S. Zahari, K. K. Mukherjee, S. Shankar, A. Mahadevan, H. Lam,

C. J. Mitchell, S. K. Shankar, P. Satishchandra, J. T. Schroeder, R. Sirdeshmukh, A. Maitra, S. D. Leach, C. G. Drake, M. K. Halushka, T. S. Prasad, R. H. Hruban, C. L. Kerr, G. D. Bader, C. A. Iacobuzio-Donahue, H. Gowda, and A. Pandey. A draft map of the human proteome. *Nature*, 509(7502):575–581, May 2014.

[14] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, 25(1):25–29, May 2000.

[15] C. R. Primmer, S. Papakostas, E. H. Leder, M. J. Davis, and M. A. Ragan. Annotated genes and nonannotated genomes: cross-species use of Gene Ontology in ecology and evolution research. *Mol. Ecol.*, 22(12):3216–3241, Jun 2013.

[16] D. D. Dalma-Weiszhausz, J. Warrington, E. Y. Tanimoto, and C. G. Miyada. The affymetrix GeneChip platform: an overview. *Meth. Enzymol.*, 410:3–28, 2006.

[17] https://commons.wikimedia.org/wiki/file:NA_hybrid.svg, Last visited 12.05.2016.

[18] C. Chen, K. Grennan, J. Badner, D. Zhang, E. Gershon, L. Jin, and C. Liu. Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PLoS ONE*, 6(2):e17238, 2011.

[19] C. Reilly, A. Raghavan, and P. Bohjanen. Global assessment of cross-hybridization for oligonucleotide arrays. *J Biomol Tech*, 17(2):163–172, Apr 2006.

[20] P. J. Amos, E. Cagavi Bozkulak, and Y. Qyang. Methods of cell purification: a critical juncture for laboratory research and translational science. *Cells Tissues Organs (Print)*, 195(1-2):26–40, 2012.

[21] M. Kircher and J. Kelso. High-throughput DNA sequencing–concepts and limitations. *Bioessays*, 32(6):524–536, Jun 2010.

[22] S. Zhao, W. P. Fung-Leung, A. Bittner, K. Ngo, and X. Liu. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PLoS ONE*, 9(1): e78644, 2014.

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

[24] http://statweb.stanford.edu/ tibs/pam/, Last visited 12.05.2016.

[25] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci. U.S.A.*, 99 (10):6567–6572, May 2002.

[26] Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, pages 1137–1143, 1995.

[27] Garrett E.S. Anbazhaghan R. Parmigiani, G. and E. Gabrielson. A statistical framework for expression-based molecular classification in cancer. *J. R. Statist. Soc. B*, 64(Part 4):717–736, 2002.

[28] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C. L. Robertson, N. Serova, S. Davis, and A. Soboleva. NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.*, 41(Database issue):D991–995, Jan 2013.

[29] R. Edgar, M. Domrachev, and A. E. Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, 30(1):207–210, Jan 2002.

[30] N. Kolesnikov, E. Hastings, M. Keays, O. Melnichuk, Y. A. Tang, E. Williams, M. Dylag, N. Kurbatova, M. Brandizi, T. Burdett, K. Megy, E. Pilicheva, G. Rustici, A. Tikhonov, H. Parkinson, R. Petryszak, U. Sarkans, and A. Brazma. ArrayExpress update–simplifying data submissions. *Nucleic Acids Res.*, 43(Database issue): D1113–1116, Jan 2015.

[31] d. a. W. Huang, B. T. Sherman, and R. A. Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*, 4(1): 44–57, 2009.

[32] d. a. W. Huang, B. T. Sherman, and R. A. Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, 37(1):1–13, Jan 2009.

[33] M. Lukk, M. Kapushesky, J. Nikkila, H. Parkinson, A. Goncalves, W. Huber, E. Ukkonen, and A. Brazma. A global map of human gene expression. *Nat. Biotechnol.*, 28(4):322–324, Apr 2010.

[34] O. Ogasawara, M. Otsuji, K. Watanabe, T. Iizuka, T. Tamura, T. Hishiki, S. Kawamoto, and K. Okubo. BodyMap-Xs: anatomical breakdown of 17 million animal ESTs for cross-species comparison of gene expression. *Nucleic Acids Res.*, 34 (Database issue):D628–631, Jan 2006.

[35] X. Liu, X. Yu, D. J. Zack, H. Zhu, and J. Qian. TiGER: a database for tissue-specific gene expression and regulation. *BMC Bioinformatics*, 9:271, 2008.

[36] F. Cunningham, M. R. Amode, D. Barrell, K. Beal, K. Billis, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fitzgerald, L. Gil, C. G. Giron, L. Gordon, T. Hourlier, S. E. Hunt, S. H. Janacek, N. Johnson, T. Juettemann, A. K. Kahari, S. Keenan, F. J. Martin, T. Maurel, W. McLaren, D. N. Murphy, R. Nag, B. Overduin, A. Parker, M. Patricio, E. Perry, M. Pignatelli, H. S. Riat, D. Sheppard, K. Taylor, A. Thormann, A. Vullo, S. P. Wilder, A. Zadissa, B. L. Aken, E. Birney, J. Harrow, R. Kinsella, M. Muffato, M. Ruffier, S. M. Searle, G. Spudich, S. J. Trevanion, A. Yates, D. R. Zerbino, and P. Flicek. Ensembl 2015. *Nucleic Acids Res.*, 43(Database issue):D662–669, Jan 2015.

[37] G. Liu, A. E. Loraine, R. Shigeta, M. Cline, J. Cheng, V. Valmeekam, S. Sun, D. Kulp, and M. A. Siani-Rose. NetAffx: Affymetrix probesets and annotations. *Nucleic Acids Res.*, 31(1):82–86, Jan 2003.

[38] http://bioconductor.org/packages/release/bioc/html/frma.html, Last visited 12.05.2016.

[39] Matthew N. McCall and Rafael A. Irizarry. *hgu133afrmavecs: Vectors used by frma for microarrays of type hgu133a.* R package version 1.5.0.

[40] http://media.affymetrix.com/support/developer/powertools/changelog/gcos-agcc/cel.html, Last visited 12.05.2016.

[41] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2015. URL `https://www.R-project.org/`.

[42] Laurent Gautier, Leslie Cope, Benjamin M. Bolstad, and Rafael A. Irizarry. affy—analysis of affymetrix genechip data at the probe level. *Bioinformatics*, 20(3):307–315, 2004. ISSN 1367-4803. doi: 10.1093/bioinformatics/btg405.

[43] N. E. MORTON. Sequential tests for the detection of linkage. *Am. J. Hum. Genet.*, 7(3):277–318, Sep 1955.

[44] Maimon Oded and Lior Rokach. Principal component analysis (PCA). In *Data Mining And Knowledge Discovery Handbook. 2nd ed.*, pages 57–61, 2010.

[45] http://setosa.io/ev/principal-component-analysis/, Last visited 12.05.2016.

[46] http://phdthesis-bioinformatics-maxplanckinstitute-molecularplantphys.matthias-scholz.de/, Last visited 12.05.2016.

[47] Maimon Oded and Lior Rokach. Hierarchical methods. In *Data Mining And Knowledge Discovery Handbook. 2nd ed.*, pages 278–279, 2010.

[48] http://www.statistics4u.com/fundstat_eng/cc_dendrograms.html, Last visited 12.05.2016.

[49] E. Paradis, J. Claude, and K. Strimmer. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20:289–290, 2004.

Appendices

A. R Code

The R code used in practical part for drawing plots and applying Barcode and fRMA methods is accompanied separately.

B. Tissue and sample information

The text file (celandtissue.txt) containing relevant information for R code to work, including manual clustering info is accompanied separately.

C. Barcode data

The text file (data.txt) containing Barcode processed data of sample tissues is accompanied separately

Non-exclusive licence to reproduce thesis and make thesis public

I, Sander Tars (date of birth: 12th of March 1994),

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to:

1.1 reproduce, for the purpose of preservation and making available to the public, including for addition to the DSpace digital archives until expiry of the term of validity of the copyright, and

1.2 make available to the public via the web environment of the University of Tartu, including via the DSpace digital archives until expiry of the term of validity of the copyright,

Description and application of gene expression data analysis method Barcode

supervised by **Anna Ufliand** and **Priit Adler**

2. I am aware of the fact that the author retains these rights.

3. I certify that granting the non-exclusive licence does not infringe the intellectual property rights or rights arising from the Personal Data Protection Act.

Tartu, 12.May 2016