SelGenes: a tool for selecting marker genes in heterogeneous samples

Kristian Samdal

Department of Informatics
University of Bergen

Abstract

SelGenes is a tool for selecting marker genes for the dominating cell type in heterogeneous samples. Based on a framework from an existing algorithm, SelGenes selects cell-type specific marker genes for the dominating cell-type and uses these marker genes to estimate cell-type proportions in the sample and cell-type specific expression profiles. We test the performance of SelGenes on a benchmark set and validate the marker genes against an external database and further apply SelGenes to a real data set containing gene expression data from cancer samples. Compared to an existing method the results from the test were consistently better for SelGenes.

Acknowledgements

I would like express my gratitude to my thesis advisor Professor Inge Jonassen for his remarks, guidance and enthusiasm during the development of this thesis. Furthermore I would like to thank Konstantina Dimitrakopoulou and the others at Computational Biology Unit and Center for Cancer Biomarkers for their inspiration and helpful comments. And finally I would like to thank the authors of the UNDO algorithm for their work, without which this thesis would not be possible.

Contents

List of Figures

List of Tables

Chapter 1

Background

In cancer research sample heterogeneity is a major obstacle as most samples contain not only tumour cells, but a range of other cell-types from the microenvironment. And although the microenvironment has an important role in the tumour formation and progression it can also be similar to normal tissue. Separating the signals is therefore crucial in order to make any discoveries that are related to the tumour and to find important changes in the tumour's microenvironment.

Different approaches of separating these signals have been developed, with cell sorting [39], expression deconvolution [42] and single cell sequencing [7] being the current main categories. This thesis focuses on expression deconvolution.

Many previous studies into gene expression deconvolution have tried deconvolving using *a priori* knowledge about either the tissue/cell-type proportions or pure tissue/cell-type expression profiles [15, 19, 35, 38]. Other studies have tried deconvolving without *a priori* knowledge [14, 23, 24]. Common to some of these studies is the use of so called tissue/cell-type specific marker genes in order to estimate tissue/cell-type proportions and expression profiles.

We develop SelGenes, a tool for selecting marker genes based on estimated pure tissue/cell-type expression values. SelGenes then use the selected marker genes to again estimate tissue/cell-type specific expression profiles.

We test performance by applying SelGenes to a benchmark set comparing the results to the results from a similar approach and try to validate selected marker genes against an external database. We further apply SelGenes to a invasive breast carcinoma dataset from The Cancer Genome Atlas.

In this chapter I will describe some of the basics in molecular biology, statistics and other terms used later in this paper.

1.1 Molecular Biology

Deoxyribonucleic acid (DNA) is the hereditary material in almost all known living organisms. DNA is made up from four chemical bases: guanine (G), cytosine (C), adenine (A) and thymine (T). These bases are called nucleic acids. These bases pair with each other, A with T and C with G, to form base pairs. Together with a sugar and a phosphate these bases form nucleotides, which are arranged in two long strands forming a double helix. The order of these nucleotides determine the information available for building and maintaining an organism [12].

A change in the nucleotide sequence is called a **mutation**. Mutations can result from DNA copying mistakes made when the cell is dividing, exposure to radiation or chemicals called mutagens, infections by viruses and some other causes. Some mutations are inherited, these are called germ line mutations, while somatic mutations occur in the body and are not passed on to the next generation [11].

A segment of a DNA molecule containing the information used to synthesise a protein or another biological product is called a **gene**. DNA molecules tend to be large, as each cell contains thousands of genes, therefore DNA molecules are highly condensed in **chromosomes**. Humans have 46 chromosomes, 22 pairs and two sex chromosomes.

The only known functions for DNA is storage and transmission of biological information [5]. The process of translating from DNA to product involves ribonucleic acid (RNA). RNA, similarly to DNA, is made up from four bases: guanine (G), cytosine (C), adenine (A) and uracil (U). Unlike DNA, RNA is single stranded and is therefore less stable than DNA and more prone to degradation. The synthesis of RNA from DNA is known as **transcription** [27]. The RNA molecule can then be translated into a protein, through a process called **translation**, or it can play other important roles in the cell, *e.g.* turn genes on or off. If a gene is switched on in a cell, it means that the cell expresses that gene. The process by which the information in a gene is used to synthesize the gene product is called gene expression [10].

In eukaryotes the initial RNA sequence transcribed from the DNA template usually contains non-coding regions, called **introns**, and protein coding regions, called **exons**. The removal of the introns is done in a process called **splicing**. Through alternative splicing one gene can encode for several proteins [2] .

An RNA molecule directly transcribed from the DNA sequence in a gene is called a transcript and the collection of all the gene readouts is called the **transcriptome**.

There are many types of RNA, messenger RNA (mRNA), ribosomal RNA (rRNA) and transfer RNA (tRNA) are all involved in the translation process. In addition to these there are many non-protein coding types of RNA that play vital roles in the cell [17].

Nearly every cell in the body has the same set of genes, but each cell only use a fraction of the genes at any given time. The genes used, or expressed, are what makes cell-types different from each other, *i.e* blood cells from brain cells. This gene expression can be measured through RNA levels in the cell [25]. This definition of gene expression will be the one used for this thesis unless otherwise stated. Using all the gene expression data from a cell-type, it is possible to set up a **gene expression profile** for that cell-type. [26] These profiles can be used to recognize a cell-type from measured gene expression typically through mRNA abundance.

A biomarker is a biological molecule found in tissues, blood or other body fluids that signifies a normal or abnormal state, it can also be used as a sign of a condition or disease [13]. A **marker gene** could be considered an example of a biomarker. We define a marker gene as a gene that is highly expressed in one cell-type, or tissue, and in small amounts or not at all in others.

1.2 Cancer

Cancer is the generic term used for a large group of diseases that involves formation of one or more tumours. Tumours form from normal cells that has lost their ability to control and regulate cell division and growth [34]. If a normal cell is damaged or has been mutated, it will normally be repaired during the cell cycle before it can divide, if it cannot be repaired the cell is scheduled for programmed cell death (apoptosis). If the genes that regulate these processes are mutated in a way that their proteins no longer function in the way that they should, then that could lead to damaged cells continuing to divide, and ultimately formation of tumours. Since cancer cells have lost or reduced function of the processes that detect and repair mutations, they are unstable and acquire new mutations more easily. This is called a cascading effect.

Cancers are not just a mass of malignant cells, they are a complex mix of malignant and non-transformed cells. These cells interact to form the tumour microenvironment (TME) [8]. For tumours to become life threatening they must develop four characteristics: (1) the ability to move, metastasise, (2) degrade the extra cellular matrix (ECM) , (3) survive in blood and (4) establish itself in new tissues. The TME is of critical importance for gaining these characteristics [9].

The TME is comprised of tumour cells, tumour stroma, blood vessels, infiltrating inflammatory cells and a variety of associated tissue cells. The composition of the TME is unique to each tumour and changes in the course of tumour progression. The tumour dominates at all times the creation and shape. There are also immune cells present in the tumour. The inflammatory cells usually contribute to tumour progression, largely because of the tumour's ability to create mechanisms for avoiding immune intervention. The tumour not only evades the immune response, but also manage to benefit from infiltrating cells [36].

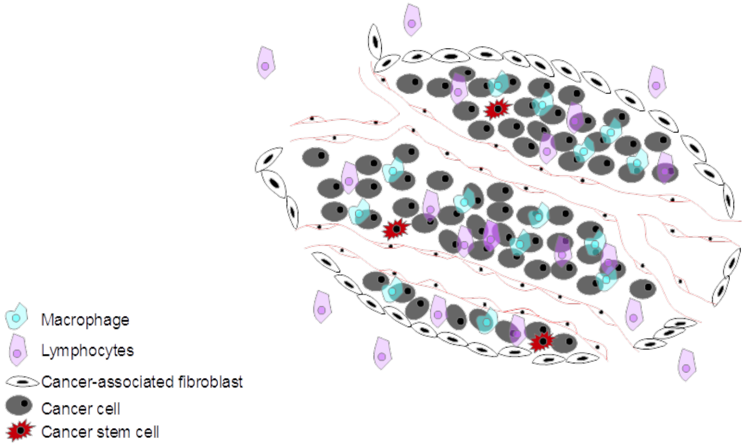


Figure 1.1: A simple figure of the TME. The cancer cells dominate, but other cell-types are present as well. This is a simplified figure with only a few of the cell-types present, there are many more cell-types in the TME. This figure was inspired by figures in [3, 8]

1.3 Expression Deconvolution

The fact that not all the genes are active at all times or expressed in different amounts explain the differences in behaviour of cell-typess [4]. Many biological samples contain more than one cell-type. For example, a cancer biopsy may not contain only cancer cells, but also a wide array of cell-types from the TME. In such a mixed sample the total amount of each mRNA measured depends on the composition of the sample [15]. This means that if we want to make any conclusions about a certain cell-type, *e.g* cancer cells, we first have to separate the different gene expression for each cell-type. One way of doing this is through a method termed **expression deconvolution**.

Expression deconvolution can work in three ways: (1) estimate cell-type proportions with the use of expression data from reference cell-types,(2) estimate cell-type expression profiles with the use of reference cell-type proportions and (3) estimate both cell-type proportions and expression profiles. The first two require some *a priori* knowledge about either the proportions or reference expression profiles, the third does not require any previously acquired information. Since getting both cell-type proportions and reference expression profiles is difficult, the third way is the most desirable, but also the most challenging.

The first way, estimating cell-type proportions, is by far the most common approach since expression data from reference cell-types is readily available from several databases. PERT [38] and ccSAM [35] are examples of this type of deconvolution.

The second way, estimating cell-type expression profiles from reference cell-type proportions, is not that common as many cell-types already have a gene expression profile built from pure samples of that particular cell-type. ESTIMATE [19] is an example of this type of deconvolution.

The third way, estimating both gene expression profiles and cell-type proportions, is not that common yet, but some approaches have been made. DeMix [14], UNDO [24] and CAM [23] are examples of this type of deconvolution.

Most approaches are based on the assumption of linearity, the assumption that the expression of each gene is a weighted average of expression values for pure population of those cell-types.

$$gene_i = \sum\nolimits_j w_j * p_j$$

where p_j is the expression value of pure population of cell-type j and w_j is the proportion of each pure population.

These approaches are usually also based on a framework proposed by Venet et al. [4, 42]. Venet et al. said that gene expression for a given cell-type had to be non-negative and their approach uses non-negative least squares (NNLS).

1.3.1 Unsupervised Deconvolution (UNDO)

Unsupervised Deconvolution (UNDO) is an algorithm that uses raw measured gene expression data to find cell-type specific marker genes, estimate the cell-type proportions in each sample and uncovers pure cell expression profiles [24]. The algorithm falls into the third category mentioned above as it does not require any knowledge about proportions or reference cell-type expression profiles.

UNDO takes in two or more samples in a matrix as input, where each row represents a probe/gene and each column is the expression value for that gene in that sample. If there are more than two samples, they apply eigenvalue decomposition (PCA) to reduce the dimensions to two, where the first two principal eigenvalues are used as a transformation matrix. UNDO then consists of three main steps: filtering, marker gene selection and deconvolution.

First a vector norm is calculated for each row in the matrix and sorted using this vector norm. The filtering step then filters away a percentage from the top and bottom, where the percentages is given as input parameters. Using this filtered set, the ratios between samples is found by dividing one column on the other. The marker genes are then selected as the genes having a ratio close to the maximum and minimum ratio. The rational behind this is that cell-type specific marker genes should be located around the radii of the scatter section corresponding to genes that are maximally expressed in one sample and minimally expressed in the other.

Using these marker genes, UNDO estimates the mixing proportions and finds the cell-type specific expression values for the entire set. This method got good results both when tested on several benchmark datasets and when compared to other deconvolution algorithms.

If the composition of the samples are 50% of each source, then UNDO will not perform as expected.

UNDO will be described in detail later in this thesis.

1.4 Statistics

Outliers

Almost all large datasets contain outliers. An outlier is a data point that lies at an abnormal distance from other data points in a random sample from a population.

Outliers can occur as results of equipment variance, background noise, amplification steps, when creating the array, equipment bias and a whole range of other factors. Outliers can have a large influence on any analysis performed on the sample and should be identified and removed before any analysis. What constitutes an outlier should be determined before the analysis is performed in order to avoid testing bias. One common way of defining and removing outliers is using the interquartile range (IQR) [30].

The IQR is the difference between the upper and lower quartile in a dataset, $IQR = Q_3 - Q_1$ where Q_3 and Q_1 is the upper and lower quartile respectively. The IQR can be used to define outliers as anything outside the following interval:

$$[Q_1 - k * IQR, Q_3 + k * IQR]$$

Where k is a constant, often 1.5.

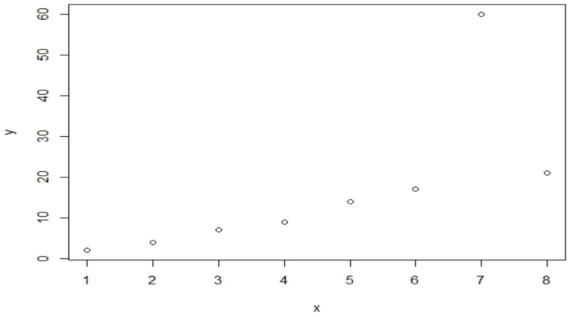


Figure 1.2: Example of a data set containing an outlier. An outlier can have major effects on any analysis of the dataset. In this example the mean of the y coordinates with the outlier is 16.75, while without it is 10.5. So by just removing one element we changed the result by a significant margin.

1.5 Technology

In this section I describe the technology used to collect gene expression data and some alternatives to expression deconvolution.

First I describe the two most prevalent methods for gathering gene expression data: microarrays and RNA-seq.

Microarray

Microarrays are used to measure gene expression values in a sample. Microarrays consists of many probes, usually cDNA molecules or oligonucleotide sequences, bound to a solid surface [21]. The target for each probe is a specific gene. The mRNA sample is then typically fluorescently labelled and hybridized to the probe microarray. A successful hybridization will increase the fluorescence intensity for the probe over the background level, and can be measured by a scanner. There are several different methods using this method, and they can be distinguished by the nature of the probes, the solid surface the probes are bound to and other characteristics.

RNA-seq

In contrast to microarray methods all sequencing approaches determine the cDNA sequence directly [41]. In general a population of RNA is converted to a library of cDNA fragments with adaptors attached to one or both ends. Each molecule is then sequenced in a high throughput manner to obtain short sequences from one or both ends. The reads are typically 30-400 bp, depending on the DNA sequencing technology used. After sequencing, the resulting reads are either aligned to a reference genome or transcripts, or assembled *de novo* without genomic sequence to produce a genome-scale transcription map that has both transcriptional structure and/or level of expression for each gene.

Microarray technology uses a probe of cDNA to detect presence of mRNA, and is therefore limited to detecting transcripts corresponding to existing sequences. In contrast, RNA-seq is not limited in such a way and can reveal the precise location of transcriptional boundaries, to a single base resolution.

RNA-Seq can also reveal sequence variations (for example, SNPs) in the transcribed regions and has been shown to be highly accurate for quantifying expression levels by established methods (quantitative PCR and spike-in RNA) [1, 37].

Because there are no cloning steps, and with some of the developed technology there is no amplification step either, RNA-seq requires less RNA sample.

Next I describe two alternatives to expression deconvolution: Single cell sequencing and cell sorting.

Single cell sequencing

Advances in DNA and RNA sequencing technology has scaled up in throughput and down in the amount of DNA or RNA is required for analysis [7]. These advances has now made it possible to analyse the DNA or RNA content of individual cells. First a cell must be isolated from the surrounding tissue. This can be done in several ways (Table 1 [7]), which can be classified in two ways, unbiased (randomized) or biased (targeted). An unbiased sample better represents the composition of the tissue, but a targeted sample is necessary for isolating rare cell-types.

Cell sorting

An alternative to gene expression deconvolution is **cell sorting**. Since cell-types have different physiological, immunological and functional properties, etc. , we can differentiate between them and therefore also sort them accordingly.

Fluorescence Activated Cell Sorting (FACS) [22] and other similar approaches can be used to physically separate defined cell-types from a heterogeneous sample before gene expression analysis is preformed [39]. One drawback with this kind of method is that the sorting process could introduce stress onto the cells and could therefore change the gene expression profiles.

Chapter 2

Aims of study

The aim of the work done in this study was to find improved methods for marker gene selection in heterogeneous samples, find and improve weaknesses in existing methods and evaluate the improvements made and compare them to the existing methods. We chose to focus specifically on UNDO.

The study objectives include:

i) To investigate UNDO closely in order to identify weaknesses in the algorithm;

ii) To develop and evaluate modifications made to the UNDO algorithm;

iii) To develop a new method for identifying and selecting marker genes based on the UNDO algorithm;

iv) To evaluate the marker genes found in this new method using an external database.

Chapter 3

Methods & Materials

3.1 Materials

3.1.1 Dataset

Benchmark data

To assess SelGenes and UNDO, a public gene expression dataset GSE19830 [35] was downloaded from the Gene Expression Omnibus (GEO) [31] website.[1] This dataset was generated from rat microarray experiments with Affymetrix Rat Genome 230 2.0 Arrays. The data we used were 12 mixed samples of brain and liver tissues in four proportions. The downloaded datasets are RMA normalized [33] meaning there is no way of getting back to the raw gene expression values, but using exponentiation we can still use the samples. RMA use log base 2 so to reverse that, exponentiation base 2 is used: $\mathbf{E}_i = 2^{sample_i}$, where $\mathbf{E}_i$ is the expression value for $gene_i$ used in our experiments and $sample_i$ is the expression value in the downloaded data.

The 12 mixed samples all contain liver and brain tissues, along with lung tissue in small proportions. The proportions of the samples are:

Sample	Brain	Liver	Lung
1-3	25%	70%	5%
4-6	34%	65%	1%
7-9	35%	60%	5%
10-12	70%	25%	5%

Table 3.1: Sample numbers and composition in the Affymetrix Rat Genome dataset

[1] http://www.ncbi.nlm.nih.gov/geo/

As UNDO is best suited for analysis on samples consisting of two cell-typess, the samples with the smallest amount of lung tissue were chosen. During testing the lung tissue components were simply ignored.

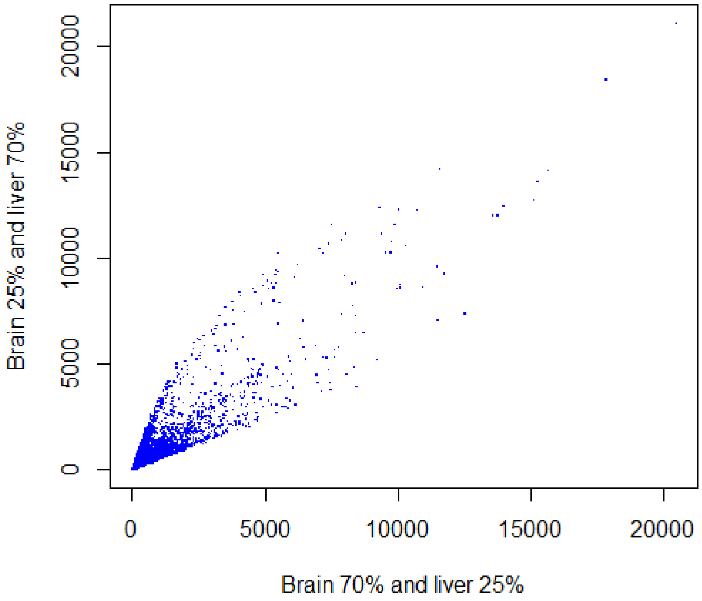


Figure 3.1: Scatter plot of sample 10 and sample 1 from the table above. In this example sample 10 is the x-coordinate and sample 1 is the y-coordinate.

As the microarray data uses probes and not genes, we need to translate from probe ids to gene names. This was done using the R packages *rat232.db* (version 3.2.3) and *annotate* (version 1.48.0). As the rat genome is updated with more and more recent information the matching from probe to gene may yield both 1 to 1 matchings and 1 to many matchings as well as 1 to none matching. When there is more than one possible match the first matching in the annotation file is used.

Real cancer patient data

To test SelGenes on real patient data we downloaded a dataset from The Cancer Genome Atlas (TCGA) [29][2]. This dataset was downloaded in July 2014 and consists of 1172 breast invasive carcinoma (BRCA) samples, of which 1052 were primary tumour samples, 113 were control samples and 7 were metastasis samples. The data is RSEM normalized RNA-seqV2 measured at level 3 (gene-level) and processed using TCGA assembler [40]. Breast cancer can be classified into 5 subtypes: Basal-like,luminal A (LumA), luminal B (LumB), HER2-enriched and normal-like [18]. Using the Supplementary table from this study [28], where they found the subtype of breast cancer for each sample, we chose a subset consisting of 30 basal-like samples and 30 LumA samples. These subtypes labels were found using microarray data.

3.1.2 Expression Atlas/Reference marker genes

Expression Atlas (EA) is a value-added database for querying differential gene expression across tissues, cell-types etc [32]. EA is an extension on the previous version Gene Expression Atlas launched by the European Bioinformatics Institute in 2008 [20]. EA introduces the concept of baseline expression, a concept that measures the abundance of each gene and splice variant in healthy or untreated tissues or cell-types.

As there does not exist a list of known marker genes for all tissues in rats, assessing the potential marker genes found with any method is challenging. So in order to assess SelGenes we downloaded lists from the EA website[3] of genes that were highly expressed in liver tissue and in brain tissues in adult rats. The list were found by entering search queries with *rattus norwegicus* and the tissue in question, and then choosing the baseline experiment that suited our parameters.

The downloaded lists were then filtered, keeping only genes that had an expression value at least c and those genes that had an expression value d times higher in liver than in brain and *vice versa*. We also removed all duplicate genes, keeping the first instance of that gene. This left us with one list of reference marker genes for liver tissue and one for brain tissue.

3.2 Methods

In this section I will first give a detailed description of how UNDO works. Then I will show some modifications on the UNDO algorithm itself, as well as showing a method of creating clearer radii for the UNDO algorithm by removing outliers, possibly improving the results. Then I will show a new method for selecting marker genes based on the estimated gene expression values

[2] https://tcga-data.nci.nih.gov/tcga/

[3] https://www.ebi.ac.uk/gxa/home

returned from UNDO. Finally in this chapter I will give a description of how we assessed this method on both the benchmark set and on the real cancer patient data.

3.2.1 UNDO

UNDO is a gene expression deconvolution algorithm that uses raw expression data from two or more samples to deconvolve them. UNDO does this by selecting tissue specific marker genes to estimate the mixing proportions. The idea is that marker genes should be located around the radii of the scatter section corresponding to the genes that are minimally expressed in one sample and maximally expressed in the other. This theory is based on two theorems given in the article [24].

UNDO uses a linear latent model of raw measured expression data:

$$\begin{bmatrix} x_{sample1}(i) \\ x_{sample2}(i) \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} s_{tumour}(i) \\ s_{stroma}(i) \end{bmatrix} \rightarrow \mathbf{x}(i) = \mathbf{a}_1 s_{tumour}(i) + \mathbf{a}_2 s_{stroma}(i) \tag{3.1}$$

for genes $i = 1, ..., n$. Where $x_{sample1}(i)$ and $x_{sample2}(i)$ are the gene expression values in heterogeneous samples and a_{jk} are the mixing proportions combined in the mixing matrix. To estimate the mixing matrix UNDO first selects cell-type specific marker genes from a filtered set.

Filtering in UNDO is done by using the Frobenius norm also known as the Gaussian norm:

$$||A_i||_F = \sqrt{\sum_j a_{i,j}^2}$$

for gene $i = 1, ..., n$ Where $a_{i,j}$ is the gene expression for gene i and sample j.

The genes are then sorted according to their norms and a percentage of the genes are filtered from the top and bottom of this sorted list. The proportion to be filtered is given as input parameter, the default being 40% from the bottom and 10% from the top. Genes removed in the filtering step are not used in the marker gene selection step, but are used in the consecutive steps.

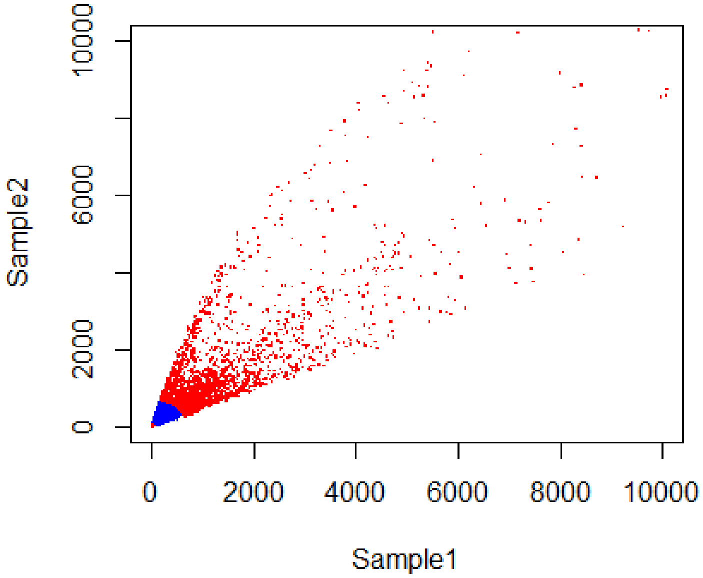


Figure 3.2: Example of how the data is filtered. The points marked in red are what UNDO filters away and the points marked in blue are the points that are kept. This plot was obtained using sample 10 and sample 1 in Table 3.1.

After the filtering is done, ratios between expression values are calculated for gene i in the filtered list with the following formula:

$$Ratio_i = Sample2_i / Sample1_i$$

The marker gene are then selected in the following way:

$$MG1 = \{gene_i | min(ratio) + eps1 * min(ratio) \geq ratio_i \geq min(ratio)\}$$
$$MG2 = \{gene_i | max(ratio) - eps2 * max(ratio) \leq ratio_i \leq max(ratio)\}$$

Where eps1 and eps2 are input parameters.

Then the sets MG1 and MG2 are the selected marker genes for the dominant cell-type or tissue in sample 1 and sample 2 respectively. We can see from this definition that UNDO always selects at least one marker gene along each of the radii.

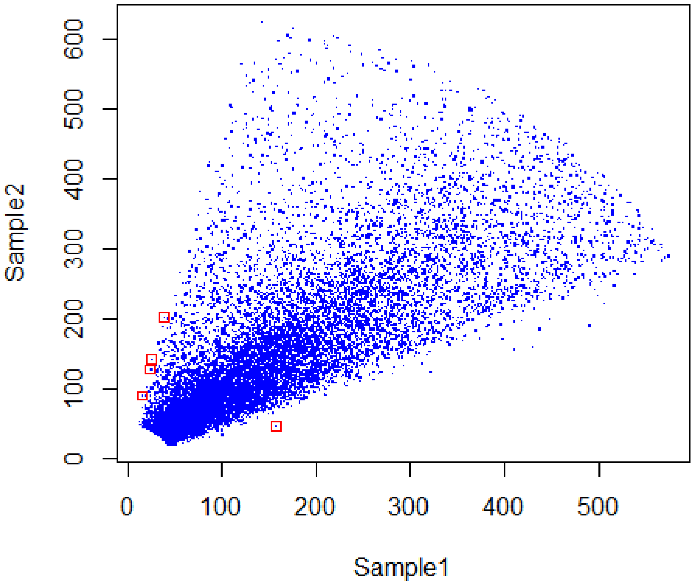


Figure 3.3: Example of how the data is filtered and marker genes selected in UNDO. The genes highlighted in red are the ones used as marker genes. This plot was obtained using sample 10 and sample 1 from Table 3.1.

The next step is to estimate the cell-type proportions in the samples. This is done using the marker genes in the following way:

$$\hat{\mathbf{a}}_1 = \begin{bmatrix} \hat{a}_{11} \\ \hat{a}_{21} \end{bmatrix} = \frac{1}{n_{MG-sample1}} \sum\nolimits_{i \in MG-sample1} \frac{x(i)}{||x(i)||}$$
$$\hat{\mathbf{a}}_2 = \begin{bmatrix} \hat{a}_{12} \\ \hat{a}_{22} \end{bmatrix} = \frac{1}{n_{MG-sample2}} \sum\nolimits_{i \in MG-sample2} \frac{x(i)}{||x(i)||}$$

Where $n_{MG-sample}$ is the number of genes chosen as marker genes for the respective sample and $||.||$ depicts the vector norm.

The two vectors are then combined to create the estimated $\hat{A} = \begin{bmatrix} \hat{a}_{11} & \hat{a}_{12} \\ \hat{a}_{21} & \hat{a}_{22} \end{bmatrix}$ from 3.1. As $\hat{a}_1$ is found using only genes that are highly expressed in sample 1 it will be inflated in one end and $\hat{a}_2$ for the same reason will be inflated in the other end. To avoid this the $\hat{A}$ need to be scaled. This scale is found by taking the inverse of $\hat{A}$ and multiplying it with a 2x1 identity matrix to get a 1x2 vector. The elements in this vector are then put on the diagonal of a 2x2 matrix and multiplied with $\hat{A}$ to get A.

$$A = \begin{bmatrix} scale & 0 \\ 0 & scale \end{bmatrix} \hat{A}$$

The cell specific expression values are found for the original dataset by using matrix inversion:

$$A^{-1} \begin{bmatrix} x_{sample1}(i) \\ x_{sample2}(i) \end{bmatrix} = \begin{bmatrix} s_{tumour}(i) \\ s_{stroma}(i) \end{bmatrix}$$

In practice they use non-negative least squares, in order uphold the assumption that all genes must have a non-negative expression value.

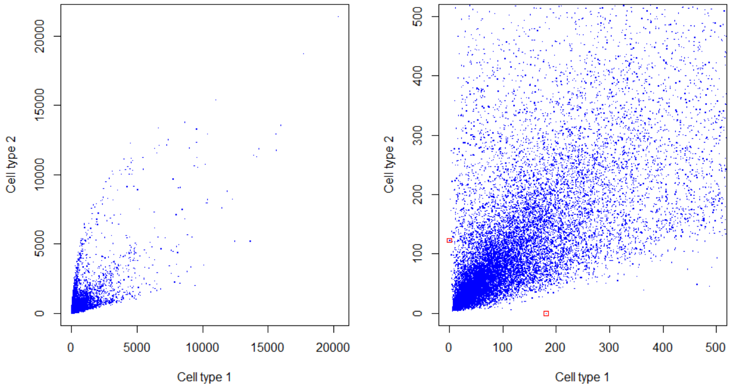


Figure 3.4: The estimated cell-spesific expression we get from an execution of the UNDO algorithm, *i.e.* $s_{sample1}(i)$ and $s_{sample2}(i)$ plotted against eachother. This is the result after running the dataset from 3.1 through UNDO with the default values. The plot to the right show the same plot zoomed in and with marker genes from UNDO highlighted in red. These plots were obtained by using sample 10 and sample 1 from Table 3.1.

3.2.2 UNDO symmetry

The marker gene selection in UNDO is done in the following way:

$$MG1 = \{gene_i | min(ratio) + eps1 * min(ratio_i)) \geq ratio_i \geq min(ratio)\}$$
$$MG2 = \{gene_i | max(ratio) - eps2 * max(ratio_i) \leq ratio_i \leq max(ratio)\}$$

We can see from this definition that the genes in MG1 will have a ratio in the interval [0,1) and the genes in MG2 will have a ratio in the interval [1,∞). This means that comparing sample X against sample Y, does not necessarily give the same number of genes as when comparing sample Y against sample X. This can have a significant impact on the calculation of the cell specific gene expression values and can be modified to select the same number and the same genes.

One way to modify this is to create two sets of ratios:

$$A_i = sample1_i/sample2_i$$
$$B_i = sample2_i/sample1_i$$

for $i = 1, ...n$. And then let marker genes be selected in the following way:

$$MG1 = \{gene_i|max(A_i) - eps1 * max(A_i)) \leq A_i \leq max(A_i)\}$$
$$MG2 = \{gene_i|max(B_i) - eps2 * max(B_i)) \leq B_i \leq max(B_i)\}$$

By doing this the interval in which the ratios fall has changed to $[1, \infty)$ for both sets. Using this procedure should get the same number of, and the same, genes executing sample X against sample Y and *vice versa*.

3.2.3 Removing outliers

Since outliers can have a large impact on any analysis performed on a data set, they should be removed from the set. We propose to use IQR to define outliers.

To find the upper and lower quartiles we first needed to be able to sort our data. We chose to sort on the ratio between samples. In order to guarantee that the order in which the samples are put together does not change what points are considered outliers, we make some changes to the IQR definition. This is because the IQR is based on the assumption that the points are normally distributed, with both sides of the mean being equal. In our set we have that the median should be around 1, as we expect that the samples will have a lot of genes in common. But below the median the interval is between 0 and 1 and above it between 1 and ∞. That is why we use two ratios A and B (as defined above) and define outliers as:

$$Outlier = \big\{gene_i|(A_i < ((Q_1(A)) - 1.5 * (Q_3(A) - Q_1(A)) \vee (B_i < ((Q_1(B)) - 1.5 * (Q_3(B) - Q_1(B))\big\}$$

where Q_1 and Q_3 are the lower and upper quartiles respectively. As we do not wish to remove too many of the points we chose the interval [0,1) instead of [1,∞).

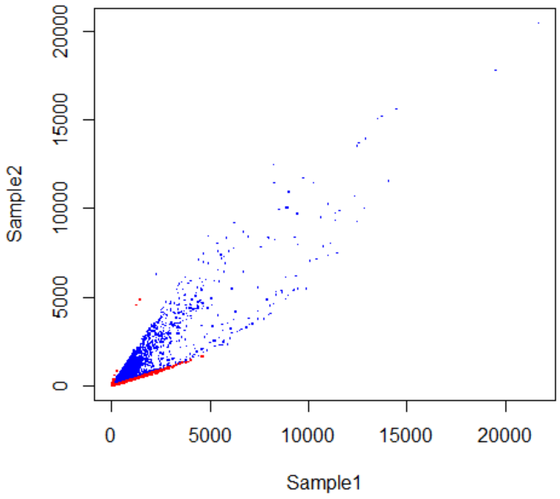


Figure 3.5: Scatter plot of sample 4 and sample 10 from Table 3.1 with the outliers highlighted in red.

To demonstrate the impact removing outliers can have on UNDO, we compared sample 4 and sample 10 from Table 3.1 with and without outlier detection and show the results in Figure 3.6.

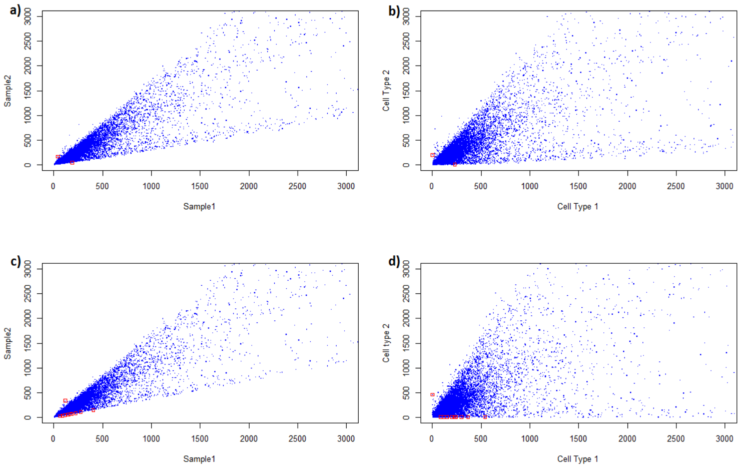


Figure 3.6: Scatter plots of expression values before and after deconvolution with UNDO both with and without outlier detection. The samples used were S4 and S10 from Table 3.1. The plots are have been zoomed in on in order to show the marker genes selected by UNDO. **a)** Scatter plot of the two samples without outlier detection. Marker genes selected by UNDO are highlighted in red. **b)** Scatter plot of the cell specific expression values returned from UNDO without outlier detection. Marker genes selected by UNDO are highlighted in red. **c)** Scatter plot of the two samples with outlier detection. Marker genes selected by UNDO are highlighted in red. **d)** Scatter plot of the cell specific expression values returned from UNDO with outlier detection. Marker genes selected by UNDO are highlighted in red.

3.2.4 Marker gene selection

There is no outlier detection in UNDO, although the filtering process will more than likely remove all points considered outliers. This is why marker gene selection in UNDO is done on the filtered set. It is done indiscriminatingly without setting any clear criteria other than that the ratio must be either the minimum/maximum or relatively close to it.

Since we do outlier detection as part of the preprocessing we propose new a selection method, SelGenes, by letting the original UNDO algorithm, with the symmetry modification, select candidate genes, calculate the slopes of the radii and estimate the proportions in the manner described previously. And then select marker genes based on the estimated cell specific expression data. This means we can now set clear criteria for what a marker gene is.

We now define a marker gene for the dominant cell-type in sample 1 as having an expression value at least x times higher in cell-type 1 than in cell-type 2, while never having an expression value above y in cell-type 2. And *vice versa* for a marker gene for the dominant cell-type in sample 2. Where both x and y are input parameters with default value: $x = 10$ and $y = 20$.

$$MG1 = \{gene_i | (exprval1 > x * exprval2) \wedge (exprval2 < y)\}$$
$$MG2 = \{gene_i | (exprval2 > x * exprval1) \wedge (exprval1 < y)\}$$

where *exprval* are the cell-type specific expression values for $gene_i$ in sample 1 and 2 respectively.

Using this method of selection we get the example scatter plot in Figure 3.7 when executing with sample 10 and sample 1 from Table 3.1.

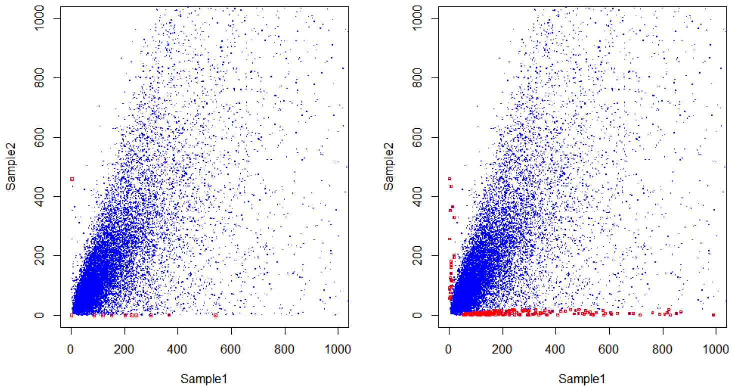


Figure 3.7: **a)** Scatter plot of the cell specific expression values returned from UNDO with the candidate genes used in the UNDO algorithm highlighted in red. **b)** Scatter plot of the cell specific expression values returned from UNDO with the marker genes selected with SelGenes highlighted in red.

3.2.5 Estimating tissue proportions

UNDO uses the marker genes it finds to estimate the mixing proportions of the samples. Having selected marker genes we can do the same estimation using the marker genes selected with SelGenes.

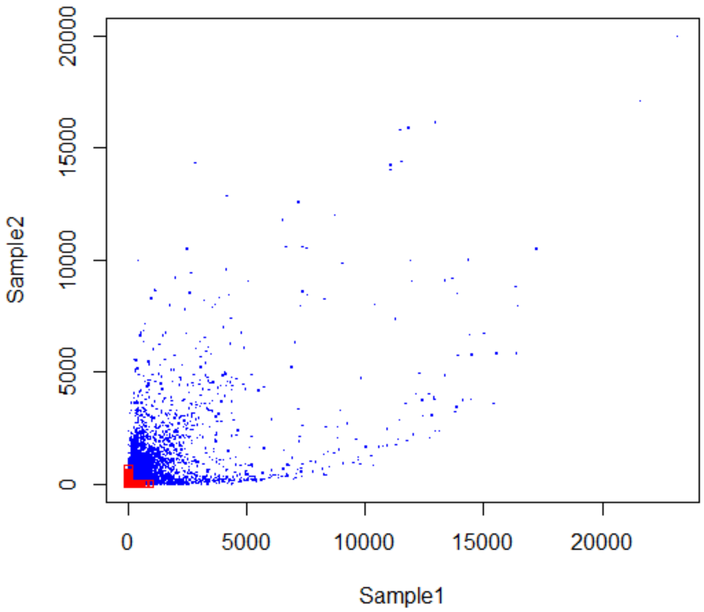


Figure 3.8: Scatter plot of estimated cell-type specific expression values after deconvolving with the marker genes found with SelGenes. This plot was obtained using sample 4 and sample 10 from Table 3.1.

In UNDO, in order to measure the estimated tissue proportions found they adopt a performance measure [6]:

$$ind_\alpha(P) = \frac{1}{2}\left[\sum_i\left(\sum_j \frac{|p_{ij}|^\alpha}{max_k|p_{ik}|^\alpha} - 1\right) + \sum_j\left(\sum_i \frac{|p_{ij}|^\alpha}{max_k|p_{kj}|^\alpha} - 1\right)\right] \tag{3.2}$$

Where p_{ij} is elements of $P = Aest^{-1} * A$. As can be seen from 3.2 that the first sum is small when there is one dominating element in each column of P. And the second sum is small when there is one dominating element in each row. Note that both criteria is 0 if and only if P is the product of a matrix and its true inverse, *i.e.* 2x2 identity matrix.

In UNDO they call their performance measure E1 and it is based on Equation 3.2, with some changes. They remove $\frac{1}{2}$ at the beginning, set $\alpha = 1$ and both i and j go from 1 to 2 as the mixing matrix is a 2x2 matrix.

Another measure of performance frequently used is Root Mean Square Error (RMSE). RMSE represents the standard deviation of the differences between the estimated values and the observed values. RMSE is defined as:

$$RMSE(X, \hat{X}) = \sqrt{\frac{\sum_{i=1}^{n}(\hat{x}_i - x_i)^2}{n}} \tag{3.3}$$

where X are the observed values, $\hat{X}$ are the estimated values and n is the total number of values.

3.2.6 Assessing the marker genes

In order to assess the genes found in both SelGenes and UNDO as marker genes we compared them to the reference list from EA. Then we formulate our hypothesis. We wish to determine if there is a larger than expected overlap between the list of genes chosen by SelGenes and EA. We therefore set up our hypothesis as following:

- H_0: The list of genes drawn are independent from the reference marker gene list.
- H_a: There is some correlation between the two lists.

To determine this we use a hypergeometric test.

In a hypergeometric distribution there is a population of N individuals to be sampled, where each individual can be classified as a success or a failure, and there are M successes in the population. Then a sample of n individuals is selected without replacement such that each subset of size n is equally likely to be chosen. [16] Then the random variable X is the number of successes in the sample. This distribution is given by:

$$P(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

Where $\binom{M}{x}$ is notation for M choose x and x is the number of successes we got from drawing n individuals.

Our method and UNDO each give a list of genes that we believe to be marker genes. In order to assess them as such we compare them to the list of reference marker genes from EA. This means we now can set up a hypergeometric distribution where:

- The population is the total number of genes.
- The samples are the number of genes found with SelGenes (or the UNDO method).
- The number of successes are the number of genes that give a match in the reference list.

The formula for finding $P(X \leq x)$ is the sum of all probabilities below x: $P(X \leq x) = \sum_0^x P(X = i)$. Meaning that since we wish to find the probability that we have an improbably large overlap we need to use:

$$P(X \geq x) = 1 - P(X \leq (x - 1))$$

The p-value obtained from these test are then used to determine whether or not we reject the null hypothesis at a certain confidence level.

3.2.7 Real cancer patient data

Having seen that SelGenes works well on a benchmark set, in the sense that it recovers known marker genes for each cell-type, we wanted to try SelGenes on real cancer patient data. We chose to analyse a set of breast cancer expression data coming from two different cancer subtypes and analyse whether the marker genes identified has any relation to the cancer subtypes and their marker genes. We also wanted to see if SelGenes manages to find expression profiles for the dominating cell-types and if it can find differences between subtype 1 compared to subtype 2.

Chapter 4

Results

4.1 Benchmark data

In order to test all of the methods outlined in Chapter 3, we need a dataset were the ground truth is known. For this purpose we used the public gene expression dataset GSE19830 outlined in Section 3.1.1. Using this dataset we tried to validate the modifications made to the UNDO algorithm, compare the results from SelGenes with the results from UNDO and evaluate genes picked by SelGenes as marker genes using the lists of genes downloaded from EA as outlined in Section 3.1.2.

4.1.1 Modifications of UNDO

To assess our modifications of UNDO outlined in Section 3.2.2 we looked at the number of genes found when comparing sample i against sample j and sample j against sample i. As ordering of samples should not change the results, we expect to find the same number of genes, and the same genes, in both executions. These results were obtained using all default values in the UNDO algorithm, that is $eps1 = eps2 = 0.01$ and filtering 40% from the bottom and 10% from the top. We found that the number of genes picked in UNDO in many cases change depending on the ordering of the samples as expected, but with our modification the identity of the genes and the number of genes always stay the same. Table 4.1 show the results from some of the comparisons.

	Number of genes found							
	UNDO				UNDO Symmetric			
	Forward		Reverse		Forward		Reverse	
Sample	MG1	MG2	MG1	MG2	MG1	MG2	MG1	MG2
S1-S4(25%-34%)	2	3	3	2	2	3	3	2
S1-S7(25%-35%)	1	2	2	1	1	2	2	1
S1-S10(25%-70%)	1	4	4	1	4	1	1	4
S4-S7(34%-35%)	1	4	4	1	1	4	4	1
S4-S10(34%-70%)	5	2	2	7	7	2	2	7
S7-S10(35%-70%)	14	1	1	19	19	1	1	19

Table 4.1: Table of some of the results from the experiments with changed UNDO selection criteria. Each line describes which samples were used (sample number is from Table 3.1 with brain tissue percentage in brackets) and the number of genes were found when these samples were compared in UNDO, with and without the symetry modification. The forward column shows how many genes were found when they were compared in the order written and reverese shows how many genes were found when they were compared in the opposite order.

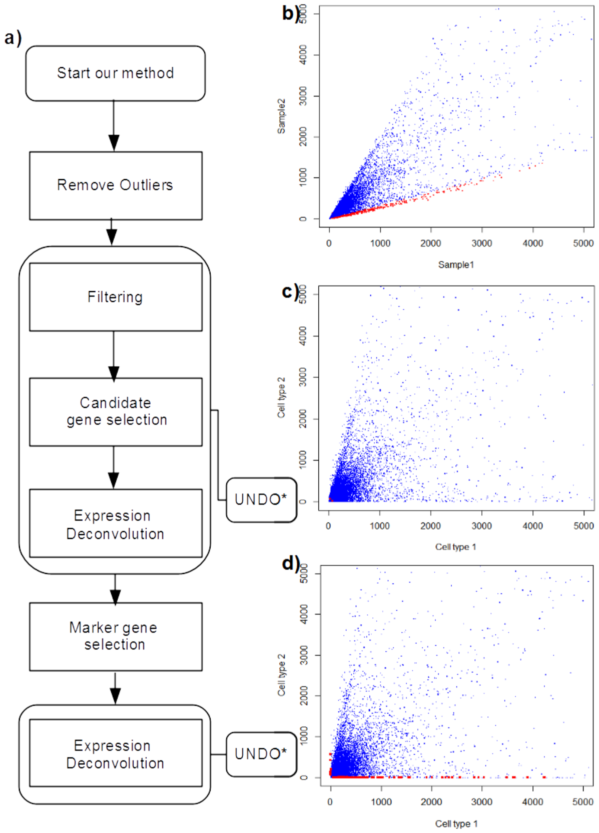
4.1.2 Marker genes

An overview of how SelGenes operates is seen in Figure 4.1. The major steps being: (1) removal of outliers, (2) executing UNDO with symmetry modification, (3) selecting marker genes based on the estimated expression values for the dominating cell-type in sample 1 and 2, and (4) using the selected marker genes to estimate expression values for the dominating cell-type in sample 1 and 2.

In order to test SelGenes on the benchmark set, we compared two and two samples against each other. Since comparing a sample to itself makes no sense and the order in which the samples are compared gives the same results, we were left with 66 separate comparisons, in the following called experiments. We used the following parameters in all of the experiments. For UNDO we used all default values, meaning filtering percentage was 40% from the bottom and 10% from the top and both epsilon values set to 0.01. For SelGenes, we used $x = 10$ and $y = 20$ leaving the set definitions:

$$MarkersLiver = \big\{gene_i | (exprvalLiver > 10 * exprvalBrain) \wedge (exprvalBrain < 20)\big\}$$
$$MarkersBrain = \big\{gene_i | (exprvalBrain > 10 * exprvalLiver) \wedge (exprvalLiver < 20)\big\}$$

where *exprval* is the estimated explained tissue expression values for $gene_i$. The UNDO algorithm should pick marker genes for the dominating cell-type in each sample. As the real mixing proportions are known we could determine which of the marker gene sets were for brain by



*The UNDO methods used have been altered using changes discussed in this paper

Figure 4.1: **a)**Flow chart of the algorithm. After outlier removal UNDO is executed, with all three steps, and marker gene selection is then done using the results. After marker gene selection just the expression deconvolution step of UNDO is executed. **b)** Scatter plot of the input data with outliers highlighted in red. **c)** Scatter plot of the estimated expression values returned from UNDO with candidate genes highlighted in red. **d)** Scatter plot of estimated expression values returned from SelGenes with marker genes highlighted in red.

checking which of the samples had the highest proportion of brain. If the two samples had the same amount of brain the first set of marker genes were chosen to be for liver and the other for brain.

All of the experiments were done both with and without outlier detection and removal. With outlier detection turned on between 0 and 820 data points were removed.

Table 4.2 shows the number of genes found by each method both with and without outlier detection. The rest of this table is included in Appendix Table A.1. As the table shows, the total number of genes found by both UNDO and by SelGenes are increased significantly by removing outliers. This is expected as removing outliers makes the radii in the scatter plot even more pronounced and more genes will have a ratio within the range of marker gene definitions.

	Number of genes found							
	SelGenes				UNDO			
	O.D off		O.D on		O.D off		O.D on	
Sample	Brain	Liver	Brain	Liver	Brain	Liver	Brain	Liver
S1-S4(25%-34%)	1	23	29	38	1	1	10	1
S1-S7(25%-35%)	1	0	0	30	0	0	5	0
S1-S10(25%-70%)	70	0	90	291	1	0	1	0
S4-S7(34%-35%)	13	0	0	73	0	1	23	0
S4-S10(34%-70%)	49	5	77	110	1	1	1	0
S7-S10(35%-70%)	54	0	82	106	2	0	2	1

Table 4.2: Total number of genes found using SelGenes and UNDO, both when outlier detection (O.D) was turned on and off. The sample numbers are the same as those from Table 3.1 with the brain tissue percentage in brackets.

4.1.3 Comparing to reference marker gene sets

As we assume the genes selected in SelGenes are marker genes we wanted to try to validate them as such using the lists of highly expressed genes in brain and liver tissue downloaded from EA. In order to filter the lists down to genes that could be considered marker genes we used $c = 5$ and $d = 10$ from Section 3.1.2. Meaning that we kept only genes that had an expression value at least 5 and an expression value 10 times higher in brain than in liver and *vice versa*. This resulted in 1465 reference marker genes for brain tissue and 548 reference marker genes for liver tissue.

Then using these lists we tried to validate the genes picked in both SelGenes and UNDO as marker genes. Trying to match each gene name in our lists to the reference list we recorded how many matches we got. Table 4.3 shows the number of hits on the same samples from Table

4.2. The rest of this table is shown in the Appendix Table A.2.

	Number of genes found							
	SelGenes				UNDO			
	O.D off		O.D on		O.D off		O.D on	
Sample	Brain	Liver	Brain	Liver	Brain	Liver	Brain	Liver
S1-S4(25%-34%)	0	0	0	0	0	1	4	0
S1-S7(25%-35%)	0	0	0	0	0	0	2	0
S1-S10(25%-70%)	29	0	36	180	0	0	0	0
S4-S7(34%-35%)	0	0	0	2	0	0	0	0
S4-S10(34%-70%)	26	0	34	32	0	0	0	0
S7-S10(35%-70%)	27	0	36	60	0	0	0	0

Table 4.3: The number of matches we got when comparing the genes from Table 4.2 to the reference lists from EA. The sample numbers are the same as those from Table3.1 with the brain tissue percentage in brackets.

In order to evaluate if the overlap with the set of reference marker genes were larger than expected we performed hypergeometric tests. With confidence levels 90% and 95% we created heatmaps (Fig.4.2) to show when H_0 were rejected, with red breakpoint for 95% and yellow breakpoint for 90%.

We found that without outlier detection UNDO (Fig.4.2**a**) has a larger than expected overlap with the reference marker gene list for the dominant cell-type in sample 1 if the sample compositions differ greatly, otherwise there is no observable pattern present. When SelGenes was used on the same set (Fig.4.2**b**) we observe the same pattern as with UNDO, but to a greater extent. We also see that in a number of the experiments that UNDO only had a large overlap with the reference marker genes for the dominating cell-type in sample 1, SelGenes had a large overlap for the dominating cell-type in sample 2 as well.

We also found that with outlier detection UNDO (Fig.4.2**c**) has a larger than expected overlap with the reference marker gene list for the dominant cell-type in sample 1 when the sample compositions differ greatly, same as when outlier detection was turned off, but to a somewhat greater extent. Otherwise there is no observable pattern as the hypergeometric test gave some low p-values here and there but it seems random.

When SelGenes was used on the same set (Fig.4.2**d**) we found that it gave a larger than expected overlap with the reference marker gene list for the dominant cell-type in both sample 1 and sample 2 for most of the samples. The exceptions being S4,S5 and S6 (Table 3.1), especially when compared against S7,S8 and S9 (Table 3.1). This can be explained with the similar composition of the samples and that therefore fewer genes will be chosen as marker genes as their ratios will be close to 1. If we accept this explanation, then we can see that for most of the samples SelGenes, with outlier detection turned on, has a larger than expected overlap with the reference marker gene lists for the dominating cell-type in both sample 1 and sample 2.

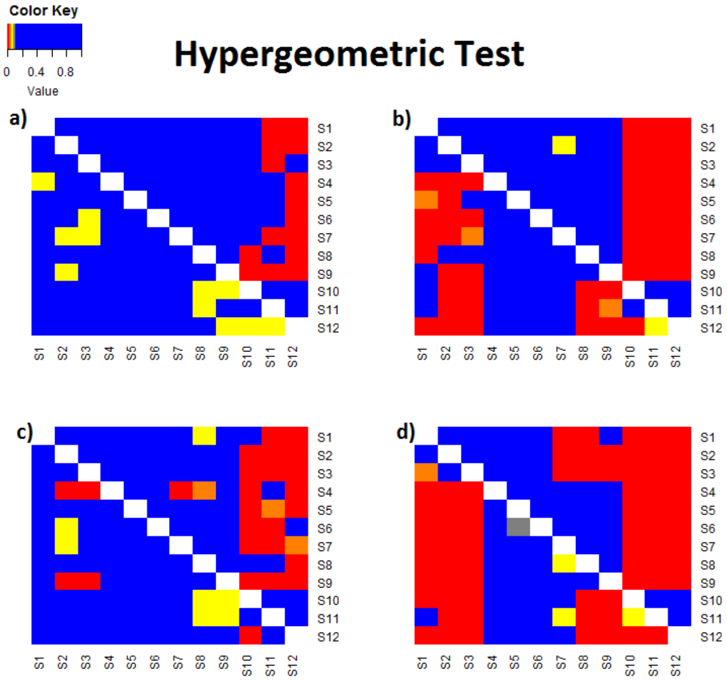


Figure 4.2: Heatmaps of the returned p-values from the hypergeometric tests. The upper triangle matrix is the p-value for marker genes from cell-type 1 and the lower triangle matrix is for marker genes from cell-type 2. **a)** P-values from the hypergeometric test when the marker genes came from executing UNDO without outlier detection. **b)** P-values from the hypergeometric test when the marker genes came from executing SelGenes without outlier detection. **c)** P-values from the hypergeometric test when the marker genes came from executing UNDO with outlier detection. **d)** P-values from the hypergeometric test when the marker genes came from executing SelGenes with outlier detection.

4.1.4 Estimating new mixing proportions

Using the marker genes found with SelGenes, expression deconvolution was performed again. In order to measure the performance of SelGenes we used the E1 criterion in the UNDO package and RMSE, both explained in Section 3.2.5.

The results from the error estimation when using the E1 criterion are shown in Figure 4.3. With the E1 criterion we found that the best estimations were those that involved S10, S11 and S12 (Table 3.1). This is expected as these are the samples that differ the most in composition from the others and therefore will give the clearest radii. We also expect to get a good estimated mixing matrix when using true marker genes in the estimation, and we can see from Figure 4.2 that in many cases these samples have a good overlap with the reference marker genes.

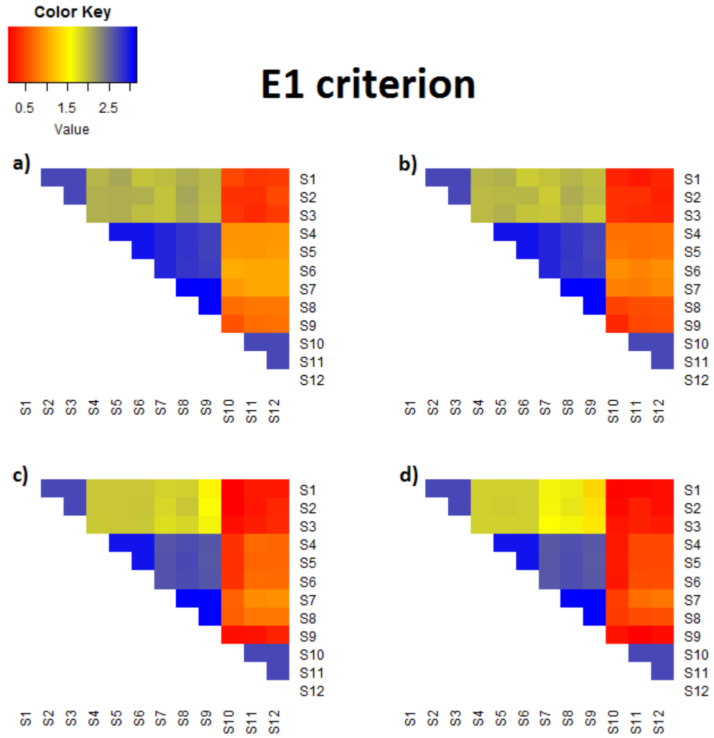


Figure 4.3: Measure of performance using the E1 criterion. **a)** The E1 criterion for the first estimation using candiate genes from UNDO without outlier detection. **b)** The E1 criterion for the second estimation using marker genes found with SelGenes without outlier detection. **c)** The E1 criterion for the first estimation using candiate genes from UNDO with outlier detection. **d)** The E1 criterion for the second estimation using marker genes found with SelGenes with outlier detection.

The results from the error estimation when using RMSE are shown in Figure 4.4. With RMSE we found that the best estimations are found for samples S10, S11 and S12 (Table 3.1), the same as with the E1 criterion. This is again expected as estimating tissue proportions with true marker genes should lead to a good estimation and with these samples we found a larger than expected overlap with the reference marker gene lists.

We should also note that the values from the two methods of error estimation have a different range, with RMSE returning values between 0.0347 and 0.4266 and the E1 criterion returning values between 0.0859 and 3.1667.

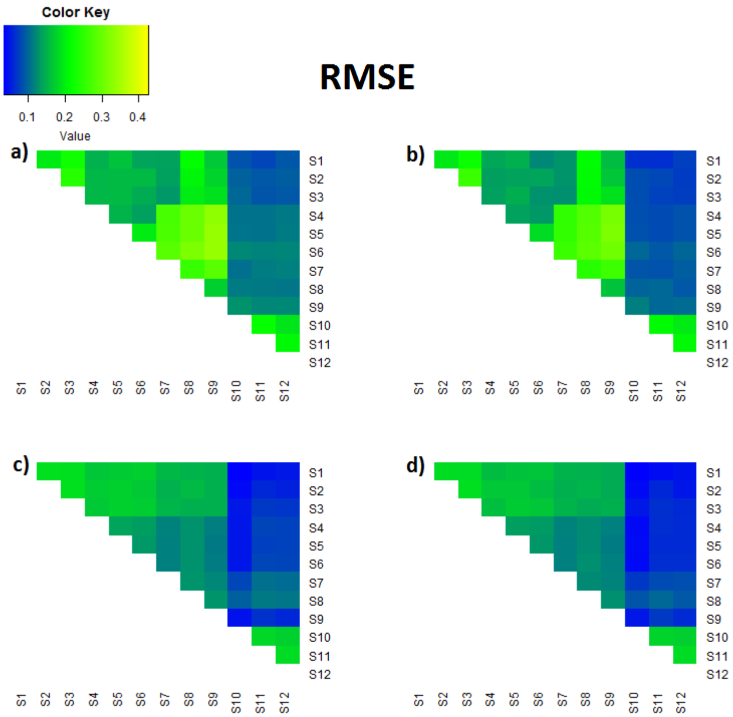


Figure 4.4: Measure of performance using RMSE. **a)** RMSE for the first estimation using candiate genes from UNDO without outlier detection. **b)** RMSE for the second estimation using marker genes found with SelGenes without outlier detection. **c)** RMSE for the first estimation using candiate genes from UNDO with outlier detection. **d)** RMSE for the second estimation using marker genes found with SelGenes with outlier detection.

4.2 Real cancer patient data

Having validated SelGenes potential of selecting marker genes on the benchmark set, we applied SelGenes on the TCGA dataset described in Section 3.1.1. In order to apply SelGenes to this dataset we had to add 1 to all expression values, as SelGenes uses the ratio between samples and dividing by 0 is not valid.

One of our hypotheses was that SelGenes selects cell-type specific marker genes, and if it does then we should expect to to find tumour-type specific marker genes when comparing one subtype of breast cancer to a different subtype of breast cancer. If this hypothesis is correct then we expect these tumour-type specific marker genes to occur more frequently when we compare type 1 samples to type 2 samples, than when we compare randomly selected samples. This means that if SelGenes finds genes that occur more frequently when comparing type 1 samples against type 2 samples than when the samples are compared randomly, then SelGenes is probably capable of selecting subtype specific marker genes.

In order to test this, we compared each basal-like sample with all LumA samples. For each comparison we collected all marker genes in two separate lists, one for the basal-like samples and one for the LumA samples. Then we counted how many times each gene occurred within each of the two lists. Then we found how many genes occurred a certain number of times and created a histogram. After that we randomized the samples and compared each of the 30 first samples against the all of the 30 last samples and collected all of the marker genes in two separate lists. Then we counted how many times each gene occurred within each list and how many genes occurred a certain number of times. We did this randomized comparison 100 times, and found the average number of genes that occurred a certain number of times and added this to our histogram. Figure 4.5 show the distribution of how many genes occurred x number of times in both marker gene sets.

As can be seen from Figure 4.5, the distributions are similar for both marker gene sets. There are however some genes that were selected around 300 or more times as marker genes when comparing basal-like samples against LumA samples and this did not occur often when randomized. This is easiest to spot when looking at marker genes for LumA (Fig.4.5**b)**, and this could mean that there are more marker genes for LumA tumours than basal-like tumours.

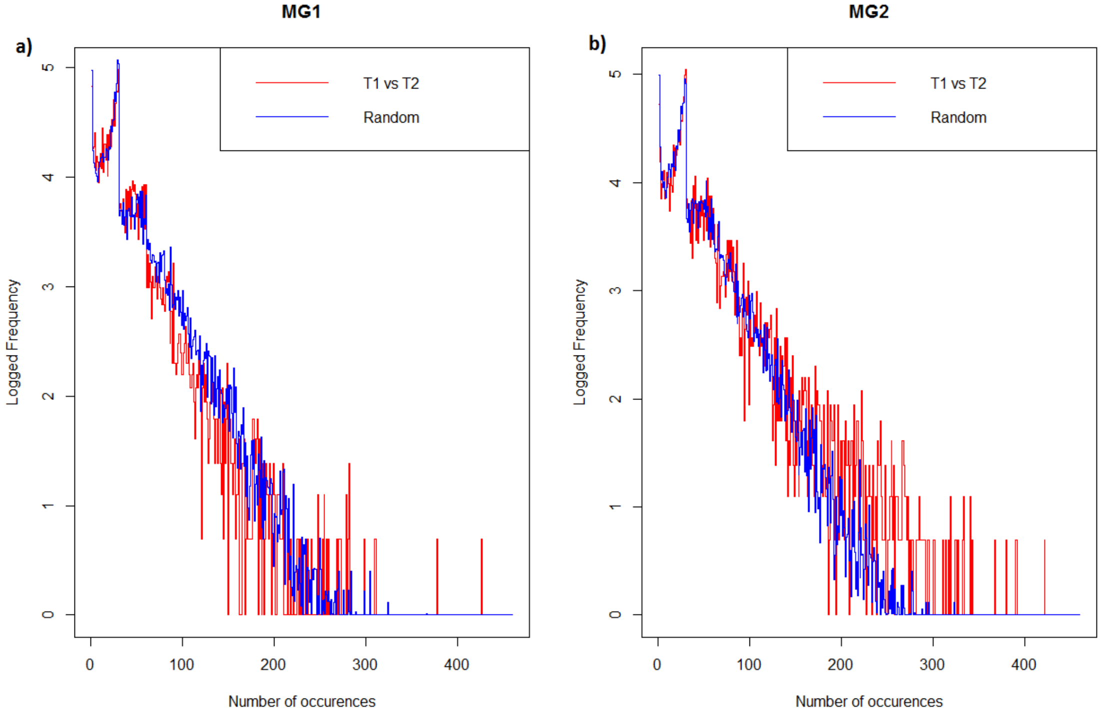


Figure 4.5: Diagram of how many genes occur x number of times as marker genes when using SelGenes. The red line shows the results when each basal-like sample (T1) were compared to all of the LumA samples (T2) and the blue line show the average results from 100 executions were the samples were randomized. To be able to take the logarithm base 2 of the frequencies we had to add 1. **a)** The logarithm base 2 of how many genes occur x number of times as marker genes for sample 1. **b)** The logarithm base 2 of how many genes occur x number of times as marker genes for sample 2

Another of our hypotheses was that SelGenes could find expression profiles for the dominating cell-type in the samples. And that SelGenes could find out if the expression profiles for subtype 1 and subtype 2 are noticeably different from each other.

In order to test this hypothesis we compared each of the basal-like samples to all of the LumA samples keeping the estimated expression values for the comparison that gives the highest estimated proportion for the basal-like samples as this should be purest expression profile. Then we did the same with the LumA samples, comparing each of the LumA samples to all of the basal-like samples and keeping the estimated expression values for the comparison that gives the highest estimated proportions for the LumA samples as this should be the purest expression profile. Then created heatmaps of the expression values for both the original samples and the estimated expression values returned from SelGenes. For both these datasets we logged the data, and for the estimated expression values we had to add 1 to all of the values as it is not possible to take the logarithm of 0 (already added 1 to the original data in order to apply SelGenes). The heatmaps are seen in Figure 4.6.

We found that SelGenes manages to find expression profiles for the dominating cell-type in the samples and that they are similar for all of the samples. However from Figure 4.6**b)** we cannot see any clear differences between the two subtypes, although some of the basal-like samples do have some differences from the others. As these differences are not found for all of the basal-like samples we cannot claim that there is a clear difference in the expression profiles between the two subtypes.

From Figure 4.6**b)** we can also note that most of the samples have a similar gene expression profile for the dominating cell-type with a lot of the genes having the same colour.

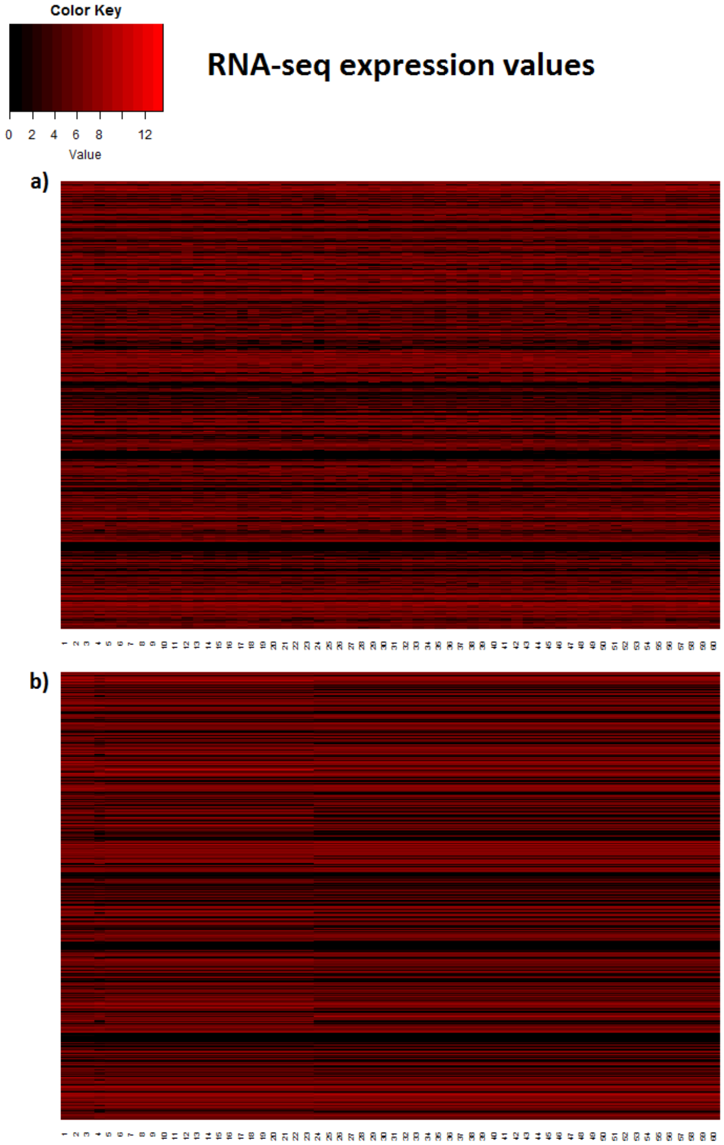


Figure 4.6: Logged expression values for the real patient cancer data. Each column represents one patient and each row represents one gene. Patient ID for each column is found in Appendix Table A.3. **a)** Heatmap of logged expression values from the original dataset, as downloaded, with 1 added to all the expression values. **b)** Heatmap of the logged expression values as returned from SelGenes. In order to log these values 1 was added to all of them as some of them were 0.

Chapter 5

Discussion

UNDO

As can be seen from Equation 3.1, UNDO divides the samples into two cell-types, tumour and stroma. This is an oversimplification, as there are multiple cell-types in the TME. This means that UNDO finds marker genes for the dominating cell-types in sample 1 and 2. UNDO also finds the expression profile for the dominating cell-types as the marker genes are used to calculate the expression profiles.

As SelGenes uses results from a modified version of UNDO to select marker genes, this oversimplification still persist in SelGenes. This can become an issue if the samples are very similar in composition as seen with some of the benchmark samples. But in real biological samples we do not expect this to be a major problem, as the probability for having two samples from two different subjects be similar in composition is small. On the other hand real samples are likely to contain more than two cell-types.

Outliers

Outliers can be caused by error as stated in Section 1.4, but not necessarily. It might be that some of the genes that are removed, in reality are good marker genes. This is one reason why we have to be careful in defining what is an outlier, in order to avoid removing exactly what we are looking for. And why we should always determine if outlier removal makes sense based on the biology and visual inspection of plots and not just the statistics.

Marker genes

We developed a method, SelGenes, for selecting marker genes for the dominating cell-type in a sample based on the estimated expression values after deconvolution. As marker genes are not generally known for tissues or cell-types, validating the genes selected as true marker genes is challenging. Using a benchmark set and an independent external database of genes highly expressed in different tissues we assessed SelGenes' ability to select genes that are highly expressed in one tissue and not in the other. As this is the definition often used for marker genes, we show SelGenes' ability to select marker genes.

On the TCGA dataset we found that when applying SelGenes on two different subtypes, some genes had a higher occurrence rate than when the subtypes were randomized. However as this was true only for a few genes and mostly only for one of the marker gene sets this could be considered a coincidence and more testing should be done on this before any conclusions are drawn.

Estimating cell-type proportions and expression profiles

SelGenes' ability to estimate cell-type proportions in samples was evaluated using both the E1 criterion and RMSE. Both methods gave the same picture, with estimation being best for the same samples, confirming that estimating cell-type proportions with marker genes works well. This also validates SelGenes' ability to select true marker genes as the estimation was clearly better when the overlap between genes selected by SelGenes and the reference marker gene lists from EA was larger than expected.

The lung tissue component in the benchmark set, that was ignored during testing, make it so that we expect both the E1 criterion and the RMSE to be different from 0. This is because the real mixing proportions did not add up to 100% and SelGenes estimates mixing proportions that do.

On the TCGA dataset SelGenes found expression profiles for the dominating cell-types, but could not find specific differences between the subtypes. We also saw from Figure 4.6 that many of the samples had a similar expression profile. This could stem from the fact that all of these samples are from breast cancer, and that they therefore have a similar expression profile. It would be interesting to compare the expression profiles found with these breast cancer samples to other expression profiles found with other cancer types.

Conclusions

SelGenes show good potential when selecting cell-type specific marker genes in heterogeneous samples after initial testing on both benchmark data and real patient cancer data. With marker gene validation through an external database SelGenes consistently found more marker genes compared to UNDO. And as finding accurate cell-type specific expression profiles are dependent on selecting good cell-type specific marker genes, SelGenes show good potential as an expression deconvolution algorithm.

For further validation of SelGenes' ability to select marker genes, we should apply SelGenes to other benchmark sets were the ground truth is known. And we should also compare the results from SelGenes to results from other expression deconvolution algorithms.

Although we have chosen to focus on cancer samples, expression deconvolution in general can be used to deconvolve other heterogeneous samples. As shown with our benchmark set, expression deconvolution can be used for tissues and cell-types, but with SelGenes the assumption is that there is a dominating tissue or cell-type.

Appendix A

Tables

Total number of marker genes found

Table A.1: The table shows the total number of marker genes we found using UNDO and SelGenes both with and without outlier detection (O.D). The sample column describe which two samples were used and how much brain percantage in each sample respectively.

Sample	Number of genes found							
	SelGenes				UNDO			
	O.D off		O.D on		O.D off		O.D on	
	Brain	Liver	Brain	Liver	Brain	Liver	Brain	Liver
S1-S2(25%-25%)	1	6	13	91	0	0	19	0
S1-S3(25%-25%)	6	2	9	93	1	0	1	0
S1-S4(25%-34%)	1	23	29	38	1	1	10	1
S1-S5(25%-34%)	2	22	31	28	1	1	5	1
S1-S6(25%-34%)	3	24	27	33	1	0	6	1
S1-S7(25%-35%)	1	0	0	30	0	0	5	0
S1-S8(25%-35%)	1	4	15	51	1	0	1	0
S1-S9(25%-35%)	2	5	14	79	1	0	1	0
S1-S10(25%-70%)	70	0	90	291	1	0	1	0
S1-S11(25%-70%)	104	2	119	281	1	1	1	0
S1-S12(25%-70%)	51	4	66	291	1	0	1	0
S2-S3(25%-25%)	6	1	12	92	1	1	15	1

Continued from previous page

Sample	Number of genes found							
	SelGenes				UNDO			
	O.D off		O.D on		O.D off		O.D on	
	Brain	Liver	Brain	Liver	Brain	Liver	Brain	Liver
S2-S4(25%-34%)	3	23	31	36	1	0	3	1
S2-S5(25%-34%)	2	26	31	36	1	0	4	1
S2-S6(25%-34%)	3	21	26	35	1	1	9	1
S2-S7(25%-35%)	3	3	4	54	0	1	0	1
S2-S8(25%-35%)	0	1	6	55	0	0	4	0
S2-S9(25%-35%)	6	4	10	84	2	1	11	2
S2-S10(25%-70%)	62	11	73	310	1	1	1	2
S2-S11(25%-70%)	83	20	96	312	1	2	1	2
S2-S12(25%-70%)	36	4	45	299	1	0	1	2
S3-S4(25%-34%)	1	23	31	47	1	0	10	1
S3-S5(25%-34%)	0	22	29	50	3	0	7	3
S3-S6(25%-34%)	8	22	25	33	1	1	8	1
S3-S7(25%-35%)	2	0	0	36	0	1	7	0
S3-S8(25%-35%)	0	4	15	64	1	0	9	1
S3-S9(25%-35%)	5	9	21	95	1	1	12	1
S3-S10(25%-70%)	47	19	58	316	1	2	1	1
S3-S11(25%-70%)	68	25	82	292	1	1	1	1
S3-S12(25%-70%)	39	10	49	315	1	1	1	1
S4-S5(34%-34%)	0	2	3	73	0	0	16	0
S4-S6(34%-34%)	14	13	25	79	2	1	14	2
S4-S7(34%-35%)	13	0	0	73	0	1	23	0
S4-S8(34%-35%)	17	3	108	13	1	1	16	1
S4-S9(34%-35%)	26	1	28	135	0	2	26	0
S4-S10(34%-70%)	49	5	77	110	1	1	1	0
S4-S11(34%-70%)	51	6	70	95	1	1	1	0
S4-S12(34%-70%)	25	6	39	121	2	1	2	0
S5-S6(34%-34%)	2	2	17	102	1	1	25	0
S5-S7(34%-35%)	17	0	0	74	0	1	24	0
S5-S8(34%-35%)	23	1	15	130	0	2	25	0
S5-S9(34%-35%)	27	3	31	155	2	1	32	2
S5-S10(34%-70%)	48	7	76	122	2	1	2	2

Continued from previous page

Sample	Number of genes found							
	SelGenes				UNDO			
	O.D off		O.D on		O.D off		O.D on	
	Brain	Liver	Brain	Liver	Brain	Liver	Brain	Liver
S5-S11(34%-70%)	48	10	74	104	1	2	1	2
S5-S12(34%-70%)	24	12	42	121	1	2	1	2
S6-S7(34%-35%)	6	0	0	97	0	1	23	0
S6-S8(34%-35%)	16	3	25	97	1	1	14	1
S6-S9(34%-35%)	18	1	24	141	1	1	25	1
S6-S10(34%-70%)	24	2	45	95	1	1	1	1
S6-S11(34%-70%)	32	3	50	110	1	1	1	1
S6-S12(34%-70%)	19	2	38	105	1	1	1	1
S7-S8(35%-35%)	0	0	5	71	0	0	12	0
S7-S9(35%-35%)	0	1	21	100	1	0	12	1
S7-S10(35%-70%)	54	0	82	106	2	0	2	1
S7-S11(35%-70%)	29	0	50	106	1	0	1	1
S7-S12(35%-70%)	16	0	31	140	1	0	1	1
S8-S9(35%-35%)	5	0	0	139	0	1	27	0
S8-S10(35%-70%)	55	14	64	137	3	1	3	0
S8-S11(35%-70%)	34	12	39	145	0	1	0	0
S8-S12(35%-70%)	32	11	39	152	1	0	1	0
S9-S10(35%-70%)	57	3	65	135	1	1	1	0
S9-S11(35%-70%)	34	2	50	141	1	1	1	0
S9-S12(35%-70%)	26	2	40	136	1	1	1	0
S10-S11(70%-70%)	14	12	25	99	1	1	15	1
S10-S12(70%-70%)	4	7	115	17	1	1	29	1
S11-S12(70%-70%)	3	5	106	10	1	2	22	2

Number of matches with Expression Atlas

Table A.2: The table shows the number of matches found when comparing marker genes from UNDO and SelGenes with those from Expression Atlas both with and without outlier detection (O.D). The sample column describe which two samples were used and how much brain percantage in each sample respectively.

Sample	Number of genes found							
	SelGenes				UNDO			
	O.D off		O.D on		O.D off		O.D on	
	Brain	Liver	Brain	Liver	Brain	Liver	Brain	Liver
S1-S2(25%-25%)	0	0	0	2	0	0	0	0
S1-S3(25%-25%)	1	0	1	10	0	0	0	0
S1-S4(25%-34%)	0	0	0	0	0	1	4	0
S1-S5(25%-34%)	0	0	0	0	0	0	0	0
S1-S6(25%-34%)	0	0	0	0	0	0	2	0
S1-S7(25%-35%)	0	0	0	0	0	0	2	0
S1-S8(25%-35%)	0	0	1	9	0	0	0	0
S1-S9(25%-35%)	0	0	1	32	0	0	0	0
S1-S10(25%-70%)	29	0	36	180	0	0	0	0
S1-S11(25%-70%)	48	0	52	167	1	0	1	0
S1-S12(25%-70%)	27	4	35	163	1	0	1	0
S2-S3(25%-25%)	1	0	1	0	0	0	4	0
S2-S4(25%-34%)	0	0	0	0	0	0	2	0
S2-S5(25%-34%)	0	0	0	0	0	0	1	0
S2-S6(25%-34%)	0	0	0	0	0	0	4	0
S2-S7(25%-35%)	1	2	1	13	0	1	0	0
S2-S8(25%-35%)	0	0	0	0	0	0	1	0
S2-S9(25%-35%)	0	0	0	0	0	1	6	0
S2-S10(25%-70%)	25	7	27	185	0	0	0	0
S2-S11(25%-70%)	38	10	41	195	1	1	1	0
S2-S12(25%-70%)	20	3	22	181	1	0	1	0
S3-S4(25%-34%)	0	0	0	0	0	0	6	0
S3-S5(25%-34%)	0	0	0	0	0	0	5	0
S3-S6(25%-34%)	0	0	0	0	0	1	5	0
S3-S7(25%-35%)	0	0	0	0	0	1	3	0

Continued from previous page

Sample	Number of genes found							
	SelGenes				UNDO			
	O.D off		O.D on		O.D off		O.D on	
	Brain	Liver	Brain	Liver	Brain	Liver	Brain	Liver
S3-S8(25%-35%)	0	0	0	0	0	0	5	0
S3-S9(25%-35%)	0	0	0	0	0	0	6	0
S3-S10(25%-70%)	23	7	24	189	0	0	0	0
S3-S11(25%-70%)	29	10	34	173	1	0	1	0
S3-S12(25%-70%)	23	3	25	184	0	0	0	0
S4-S5(34%-34%)	0	0	1	0	0	0	1	0
S4-S6(34%-34%)	0	1	3	0	0	0	1	0
S4-S7(34%-35%)	0	0	0	2	0	0	0	0
S4-S8(34%-35%)	0	1	1	6	0	1	0	1
S4-S9(34%-35%)	0	0	4	0	0	0	0	0
S4-S10(34%-70%)	26	0	34	32	0	0	0	0
S4-S11(34%-70%)	25	0	32	22	0	0	0	0
S4-S12(34%-70%)	17	0	25	45	2	0	2	0
S5-S6(34%-34%)	0	0	1	0	0	0	1	0
S5-S7(34%-35%)	0	0	0	0	0	0	1	0
S5-S8(34%-35%)	0	0	0	1	0	0	0	0
S5-S9(34%-35%)	0	2	5	1	0	0	0	0
S5-S10(34%-70%)	23	0	32	46	0	0	0	0
S5-S11(34%-70%)	18	0	33	37	0	0	0	0
S5-S12(34%-70%)	18	0	24	43	1	0	1	0
S6-S7(34%-35%)	0	0	0	1	0	0	1	0
S6-S8(34%-35%)	0	1	6	1	0	0	0	0
S6-S9(34%-35%)	0	0	5	0	0	0	1	0
S6-S10(34%-70%)	13	0	21	26	0	0	0	0
S6-S11(34%-70%)	18	0	21	19	0	0	0	0
S6-S12(34%-70%)	13	0	24	30	1	0	1	0
S7-S8(35%-35%)	0	0	0	0	0	0	1	0
S7-S9(35%-35%)	0	0	0	0	0	0	2	0
S7-S10(35%-70%)	27	0	36	60	0	0	0	0
S7-S11(35%-70%)	17	0	28	47	1	0	1	0
S7-S12(35%-70%)	10	0	19	64	1	0	1	0

Continued from previous page

Sample	Number of genes found							
	SelGenes				UNDO			
	O.D off		O.D on		O.D off		O.D on	
	Brain	Liver	Brain	Liver	Brain	Liver	Brain	Liver
S8-S9(35%-35%)	0	0	0	0	0	0	2	0
S8-S10(35%-70%)	23	10	26	60	2	1	2	0
S8-S11(35%-70%)	18	6	20	56	0	1	0	0
S8-S12(35%-70%)	18	5	20	60	1	0	1	0
S9-S10(35%-70%)	25	3	30	78	1	1	1	0
S9-S11(35%-70%)	16	1	22	65	1	0	1	0
S9-S12(35%-70%)	13	2	22	69	1	1	1	0
S10-S11(70%-70%)	0	2	5	1	0	0	1	0
S10-S12(70%-70%)	0	1	1	2	0	1	0	1
S11-S12(70%-70%)	0	1	1	1	0	1	0	1

Patient IDs used when testing SelGenes

Table A.3: The patient IDs used when testing SelGenes on the real patient cancer data. The number corresponds to the ones used for marking columns in Figure 4.6.

Number	TCGA patient ID
1	TCGA-AO-A0J5-01A-11R-A034-07
2	TCGA-A8-A0A7-01A-11R-A00Z-07
3	TCGA-AN-A0XL-01A-11R-A10J-07
4	TCGA-AR-A0TR-01A-11R-A084-07
5	TCGA-A8-A096-01A-11R-A00Z-07
6	TCGA-A7-A13D-01A-13R-A12P-07
7	TCGA-E9-A22E-01A-11R-A157-07
8	TCGA-B6-A0RH-01A-21R-A115-07
9	TCGA-EW-A1J3-01A-11R-A13Q-07
10	TCGA-E2-A1IH-01A-11R-A13Q-07
11	TCGA-C8-A1HM-01A-12R-A137-07
12	TCGA-AR-A5QQ-01A-11R-A28M-07

Continued from previous page

Number	TCGA patient ID
13	TCGA-AO-A03N-01B-11R-A10J-07
14	TCGA-AQ-A04H-01B-11R-A10J-07
15	TCGA-E2-A1IF-01A-11R-A144-07
16	TCGA-D8-A1Y0-01A-11R-A14M-07
17	TCGA-AR-A1AH-01A-11R-A12D-07
18	TCGA-BH-A0HW-01A-11R-A034-07
19	TCGA-E2-A1IG-01A-11R-A144-07
20	TCGA-BH-A1EN-11A-23R-A13Q-07
21	TCGA-E2-A1B4-01A-11R-A12P-07
22	TCGA-AN-A03X-01A-21R-A00Z-07
23	TCGA-BH-A0C0-01A-21R-A056-07
24	TCGA-E9-A5UO-01A-11R-A28M-07
25	TCGA-E2-A10B-01A-11R-A10J-07
26	TCGA-BH-A18K-11A-13R-A12D-07
27	TCGA-D8-A27K-01A-11R-A16F-07
28	TCGA-B6-A0X5-01A-21R-A109-07
29	TCGA-E2-A15J-01A-11R-A12P-07
30	TCGA-BH-A18I-01A-11R-A12D-07
31	TCGA-B6-A0RU-01A-11R-A084-07
32	TCGA-BH-A0BW-11A-12R-A115-07
33	TCGA-E2-A1IK-01A-11R-A144-07
34	TCGA-B6-A0RQ-01A-11R-A115-07
35	TCGA-E2-A1LS-11A-32R-A157-07
36	TCGA-B6-A40B-01A-11R-A239-07
37	TCGA-A8-A08Z-01A-21R-A00Z-07
38	TCGA-BH-A0B5-11A-23R-A12P-07
39	TCGA-E9-A1N5-11A-41R-A14D-07
40	TCGA-BH-A0C3-01A-21R-A12P-07
41	TCGA-B6-A0IK-01A-12R-A056-07
42	TCGA-BH-A0DG-11A-43R-A12P-07
43	TCGA-BH-A0E6-01A-11R-A034-07
44	TCGA-B6-A0IM-01A-11R-A034-07
45	TCGA-E2-A1LK-01A-21R-A14D-07
46	TCGA-BH-A18R-01A-11R-A12D-07

Continued from previous page

Number	TCGA patient ID
47	TCGA-E2-A1LH-11A-22R-A14D-07
48	TCGA-BH-A1EV-01A-11R-A137-07
49	TCGA-C8-A134-01A-11R-A115-07
50	TCGA-BH-A18J-11A-31R-A12D-07
51	TCGA-A7-A13E-01A-11R-A12P-07
52	TCGA-AC-A5XU-01A-11R-A28M-07
53	TCGA-BH-A1EV-11A-24R-A137-07
54	TCGA-BH-A1EU-11A-23R-A137-07
55	TCGA-BH-A0H5-11A-62R-A115-07
56	TCGA-C8-A12T-01A-11R-A115-07
57	TCGA-AO-A03M-01B-11R-A10J-07
58	TCGA-B6-A0IP-01A-11R-A034-07
59	TCGA-D8-A142-01A-11R-A115-07
60	TCGA-D8-A1XL-01A-11R-A14M-07

Bibliography

[1] Mortazavi A, Williams B.A, McCue K, Schaeffer L, and Wold B. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nature Methods*, 5:621–628, 2008.

[2] Alberts B, Johnson A, Lewis J, Raff M, Roberts K, and Walter P. *Molecular Biology of the Cell*. Garland Science, 2002.

[3] Hanahan D and Weinberg R.A. The hallmarks of cancer. *Cell*, 100(1):57–70, 2000.

[4] Venet D, Pecasse F, Maenhaut C, and Bersini H. Separation of samples into their constituents using gene expression data. *Bioinformatics*, 17:S279–S287, 2001.

[5] Nelson D.L, Cox M.M, and Lehninger A.L. *Lehninger: Principles of biochemistry*. W. H. Freeman, 2008.

[6] Moreau E. A generalization of joint-diagonalization criteria for source separation. *IEEE Transactions On Signal Processing*, 49(3):530–541, 2001.

[7] Shapiro E, Biezuner T, and Linnarsson S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews Genetics*, 14:618–630, 2013.

[8] Balkwill F, Capasso M, and Hagemann T. The tumor microenvironment at a glance. *J Cell Sci*, 125:5591–5596, 2012.

[9] Mbeunkui F and Johann Jr D.J. Cancer and the tumor microenvironment: a review of an essential relationship. *Cancer Chemotherapy and Pharmacology*, 63(4):571–582, 2008.

[10] GeneticsHomeReference. Gene expression. `http://ghr.nlm.nih.gov/glossary=geneexpression`. Accessed: 2015-08-24.

[11] GeneticsHomeReference. Mutation. `http://ghr.nlm.nih.gov/glossary=mutation`. Accessed: 2015-08-24.

[12] GeneticsHomeReference. What is DNA? `http://ghr.nlm.nih.gov/handbook/basics/dna`. Accessed: 2015-07-20.

[13] National Cancer Institute. NCI dictionary of cancer terms. `http://www.cancer.gov/publications/dictionaries/cancer-terms?cdrid=45618`. Accessed: 2016-05-26.

[14] Ahn J, Yuan Y, Parmigiani G, Suraokar M.B, Diao L, Wistuba I.I, and Wang W. Demix: deconvolution for mixed cancer transcriptomes using raw measured data. *Bioinformatics*, 29(15):1865–1871, 2013.

[15] Clarke J, Seo P, and Clarke B. Statistical expression deconvolution from mixed tissue samples. *Bioinformatics*, 26(8):1043–1049, 2010.

[16] Devore J.L and Berk K.N. *Modern Mathematical Statistics with Applications*. Springer, 2012.

[17] Mattick J.S and Makunin I.V. Non-coding RNA. *Hum. Mol. Genet.*, 14:R17–R29, 2006.

[18] Parker J.S, Mullins M, Cheang M.C.U, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He X, Hu Z, Quackenbush J.F, Stijlenman I.J, Palazzo J, Marron J.S, Nobel A.B, Mardis E, Nielsen T.O, Ellis M.J, Perou C.M, and Bernard P.S. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol*, 27(8):1160–1167, 2009.

[19] Yoshihara K, Shahmoradgoli M, Martinez E, Vegesna R, Kim H, Torres-Garcia W, Trevin V, Shen H, Laird P.W, Levine D.A, Carter S.L, Getz G, Stemke-Hale K, Mills G.B, and Verhaak R.G.W. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nature Communications*, 4(2612), 2013.

[20] Kapushesky M, Adamusiak T, Burdett T, Culhane A, Farne A, Filippov A, Holloway E, Klebanov A, Kryvych N, Kurbatova N, Kurnosov P, Malone J, Melnichuk O, Petryszak R, Pultsin N, Rustici G, Tikhonov A, Travillian R.S, Williams E, Zorin A, Parkinson H, and Brazma A. Gene expression atlas updatea value-added database of microarray and sequencing-based functional genomics experiments. *Nucl. Acids Res.*, 40(D1):D1077–D1081, 2011.

[21] Miller M and Tang Y. Basic concepts of microarrays and potential applications in clinical microbiology. *CMR*, 22(4):611–633, 2009.

[22] Julius M.H, Masuda T, and Herzenberg L.A. Demonstration that antigen-binding cells are precursors of antibody-producing cells after purification with a fluorescence-activated cell sorter. *Proc. Nat. Acad. Sci*, 69(7):1934–1938, 1972.

[23] Wang N, Hoffman E.P, Chen L, Chen L, Zhang Z, Liu C, Yu G, Herrington D.M Clark R, and Wang Y. Mathematical modelling of transcriptional heterogeneity identifies novel markers and subpopulations in complex tissues. *Sci Rep*, 6:18909, 2015.

[24] Wang N, Gong T, Clarke R, Chen L, Shih T.M, Zhang Z, Levine D.A, Xuan J, and Wang Y. UNDO: a bioconductor r package for unsupervised deconvolution of mixed gene expressions in tumor samples. *Bioinformatics*, 31(1):137–139, 2014.

[25] Nature. Gene expression is analyzed by tracking RNA. `http://www.nature.com/scitable/topicpage/gene-expression-is-analyzed-by-tracking-rna-6525038`. Accessed: 2015-08-24.

[26] Nature. Gene expression profiling. `http://www.nature.com/subjects/gene-expression-profiling`. Accessed: 2016-01-07.

[27] Nature. Ribonucleic acid / RNA. `http://www.nature.com/scitable/definition/ribonucleic-acid-rna-45`. Accessed: 2015-07-24.

[28] The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490:61–70, 2012.

[29] The Cancer Genome Atlas Research Network, Weinstein J.N, Collisson E.A, Mills G.B, Shaw K.R.M, Ozenberger B.A, Ellrott K, Shmulevich I, Sander C, and Stuart J.M. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45:1113–1120, 2013.

[30] Purplemath. Box-and-whisker plots: Interquartile ranges and outliers. `http://www.purplemath.com/modules/boxwhisk3.htm`. Accessed: 2016-05-20.

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

[32] Petryszak R, Burdett T, Fiorelli B, Fonseca N.A, Gonzalez-Porta M, Hastings E, Huber W, Jupp S, Keays M, Kryvych N, McMurry J, Marioni J.C, Malone J, Megy K, Rustici G, Tang A.Y, Taubert J, Williams E, Mannion O, Parkinson H.E, and Brazma A. Expression atlas updatea database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. *Nucl. Acids Res.*, 42(D1):D926–D932, 2013.

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

[34] Weinberg R.A. *The biology of cancer.* Garland Science, 2014.

[35] Shen-Orr S.S, Tibshirani R, Khatri P, Bodian D.L, Staedtler F, Perry N.M, Hastie T, Sarwal M.M, M Davis, and Butte A.J. Cell typespecific gene expression differences in complex tissues. *Nature Methods*, 7:287–289, 2010.

[36] Whiteside T.L. The tumor microenvironment and its role in promoting tumor growth. *Oncogene*, 27:5904–5912, 2008.

[37] Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, and Snyder M. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, 320(5881):1344–1349, 2008.

[38] Qiao W, Quon G, Csaszar E, Yu M, Morris Q, and Zandstra P.W. PERT: A method for expression deconvolution of human blood samples from varied microenvironmental and developmental conditions. *PloS Comput Biol*, 8(12):e1002838, 2012.

[39] Zhong Y, Wan Y.W, Pang K, Chow L.M.L, and Liu Z. Digital sorting of complex tissues for cell type-specific gene expression profiles. *BMC Bioinformatics*, 14:89, 2013.

[40] Zhu Y, Qiu P, and Ji Y. Tcga-assembler: An open-source pipeline for TCGA data downloading, assembling, and processing. *Nat. Methods*, 11(6):599–600, 2014.

[41] Wang Z, Gerstein M, and Snyder M. RNA-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10:57–63, 2009.

[42] Yingdong Z and Richard S. Gene expression deconvolution in clinical samples. *Genome Medicine*, 2:93, 2010.