universität
wien

MASTERARBEIT / MASTER’S THESIS

Titel der Masterarbeit / Title of the Master‘s Thesis

„Analysing FGFs and FGFRs in Human Hepatocellular Carcinoma“

verfasst von / submitted by

Andreas Unterberger

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2020 / Vienna 2020

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 066 875
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Bioinformatik
Betreut von / Supervisor:	Ao.Univ.-Prof.in Dr.in Bettina Grasl-Kraupp (MedUni Wien)

Erklärung zur Verfassung der Arbeit

Hiermit erkläre ich, die vorgelegte Arbeit selbständig verfasst und ausschließlich die angegebenen Quellen und Hilfsmittel benutzt zu haben. Alle wörtlich oder dem Sinn nach aus anderen Werken entnommenen Textpassagen und Gedankengänge sind durch genaue Angabe der Quelle in Form von Anmerkungen bzw. In-Text-Zitationen ausgewiesen. Mir ist bekannt, dass jeder Fall von Plagiat zur Nicht-Bewertung der gesamten Lehrveranstaltung führt und der Studienprogrammleitung gemeldet werden muss. Ferner versichere ich, diese Arbeit nicht bereits andernorts zur Beurteilung vorgelegt zu haben.

Wien, 17.08.2020

Andreas Unterberger, BSc

Acknowledgement

The completion of this study would not have been possible without the participation and assistance of other people. I would like to express my sincere gratitude to Ao.Univ.-Prof.in Dr.in Bettina Grasl-Kraupp and Univ.-Prof. Mag. Dr. Thomas Rattei for providing me the opportunity to do my thesis and project at the Institute of Cancer Research at the Medical University of Vienna as well as for their support and advice. I extend special thanks to DI Thomas Mohr for his excellent mentoring. I am grateful for his advice, his support and for giving me the freedom to make mistakes and learn from them. Furthermore, I want to thank Marie Theres-Bräuer for her help during the development of the pipeline and her support. Finally, special thanks to my friends and family, who motivated me and helped me keep up my spirits.

Abstract

Recent studies suggest members of the FGF family as possible targets for new pharmaceuticals in human hepatocellular carcinoma (HCC), a cancer type with low chances of survival due to high resistance against drugs. In this study we analysed the interaction of the FGF family in HCC with an analysis pipeline programmed mainly in R. Data was obtained from the TCGA/GDC database. Besides the pre-processing steps in this pipeline, we also implemented DeMixT, which is a deconvolution algorithm performing an in-silico microdissection on expression data. Microdissection in-vitro is a necessary step to get refined samples for sequencing in cancer, but it is also expensive and time-consuming. Computational approaches, on the other side, are relatively cheap and provide good results. Here we calculated gene expression profiles where no infiltration of stromal and immune cells occurred. We performed differential gene expression analysis, gene set enrichment analysis, gene regulatory network analysis and survival analysis on the given and on the deconvolved data. In both the convoluted and the deconvolved data, the expression of FGF12 and FGF13 is upregulated and the expression of FGF2 is downregulated. To get deeper insight into the resulting gene sets of the gene set enrichment analysis with focus on the members of the FGF family, we created mutual information networks of these gene sets using Aracne-AP. Comparing the deconvolved normal and the deconvolved tumour compartment showed changes of the interaction between FGFs and all other genes. Furthermore, results of the survival analysis of both the convoluted and the deconvolved data overlap on FGF18 and FGFR3. According to these, low chances of survival correlate with low expression of FGF18 and high expression of FGFR3.

Zusammenfassung

Jüngste Studien legen nahe, dass Mitglieder der FGF-Familie mögliche Ziele für neue Arzneimittel gegen humanen hepatozellulären Krebs (HCC) sein können, einem Krebstyp mit geringen Überlebenschancen aufgrund hoher Resistenz gegen Medikamente. In dieser Studie wurde die Interaktion der FGF-Familie in HCC mittels einer Analysepipeline, welche in der Programmiersprache R geschrieben wurde, untersucht. In dieser Studie wurden Daten aus der TCGA/GDC-Datenbank verwendet. Zusätzlich zur Standardisierungsprozedur wurde DeMixT, ein Programm zur in-silico Mikrodissektion, implementiert. Die Mikrodissektion in vitro ist eine Technik um Zielzellen von umliegenden Gewebe zu befreien und so Proben zu erhalten die bei der Sequenzierung eine höhere Genauigkeit besitzen. Diese Methode ist jedoch kostenintensiv und zeitaufwendig. Computergestützte Ansätze sind dagegen relativ billig und liefern gute Ergebnisse. Es wurden Genexpressionsprofile berechnet, bei denen keine Infiltration von Stroma- und Immunzellen auftrat. Sowohl an den gegebenen Daten wie auch an den Ergebnissen des DeMixT Algorithmus wurde eine differentielle Genexpressionsanalyse, eine Gen-Set-Anreicherungsanalyse (GSEA), eine Genregulations-Netzwerkanalyse und eine Überlebensanalyse durchgeführt. Das reguläre Genexpressionsprofil wie auch das mittels DeMixT berechnete zeigen eine Überexpression von FGF12 und FGF13 und eine Unterexpression von FGF2. Um einen tieferen Einblick in die resultierenden Gen-Sets der GSEA mit Schwerpunkt auf den Mitgliedern der FGF-Familie zu erhalten, wurden mithilfe von Aracne-AP MI-Netzwerke von diesen Gen-Sets erstellt. Der Vergleich der berechneten normalen und der berechneten Tumor-Komponente, des DeMixT Algorithmus, zeigte Veränderungen der Wechselwirkung zwischen FGFs und allen anderen Genen. Darüber hinaus einigen sich die Ergebnisse der Überlebensanalyse des regulären Genexpressionsprofils und des mit DeMixT berechneten Genexpressionsprofils auf FGF18 und FGFR3. Demnach korrelieren niedrige Überlebenschancen mit einer niedrigen Expression von FGF18 und einer hohen Expression von FGFR3.

Table of Contents

List of Figures

List of Tables

1 Introduction

1.1 Hepatocellular Carcinoma

According to statistics of the Global Cancer Observatory (GCO) from 2018, hepatocellular carcinoma (HCC) is the sixth most common cancer type. Among all other cancer types, it is the fourth most common which leads to death. Around 70% of new cases happen to be in Asia. In general, HCC occurs about twice as often in men than in women.[1]

Chronic liver disease and cirrhosis are the most relevant risk factors in HCC development. Due to high amounts of damaged tissue, repair mechanisms are activated and cells show higher proliferation rates. High activity of cell division cycles therefore lead to accumulation of mutations in the genome with some mutations being carcinogenic. The premalignant cells accumulate more and more mutations which provide a selective advantage. Known risk factors for HCC are viral hepatitis infections and excessive alcohol consumption. Other important risk factors can be aflatoxins, smoking and gender.[2]

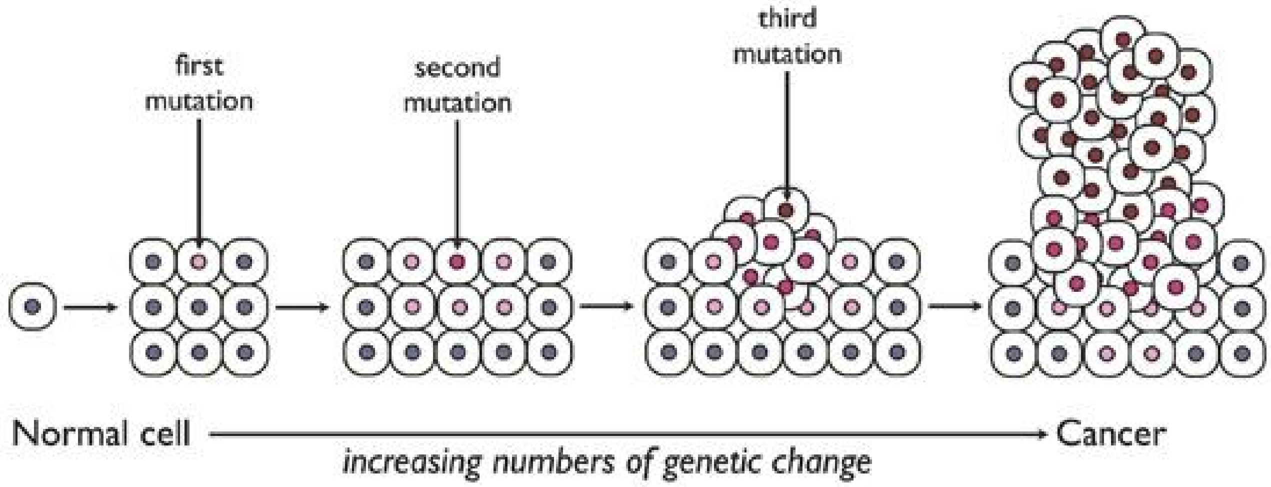


Figure 1.1.1: Schematic view of the development of cancer cells. Figure adapted from: https://www.carolina.com/teacher-resources/Interactive/the-effects-of-cell-cycle-deviation-on-cancer-development/tr38703.tr[56]

Early diagnosis can guarantee better chances for therapy and survival of patients. Radiodiagnostics and serological markers are used for detection. There are several therapy options available to deal with hepatocellular cancer. Among those options surgical resection and liver transplantation show the best chances for survival of patients. Other options are transarterial-chemo-embolization, transarterial radiation, percutaneous local ablation, microwave ablation and systemic therapy. According to GCO only 7% of people worldwide survive HCC.[1] This is due to pre-existing liver damage and its high resistance against pharmaceuticals.

1.2 Hallmarks of Cancer

Hanahan and Weinberg[3] suggested certain characteristics every tumour shows, namely the so-called hallmarks of cancer. The first important hallmark for cancer cells is sustaining proliferative signalling. While in normal tissue the cell cycle is under control, cancer cells manage to deregulate certain pathways to keep growing and multiplying. Cancer cells not only need to keep proliferation high, but also have to prevent control mechanisms which negatively affect proliferation. This leads to the next hallmark, which is called evading growth suppressors. It has been proven that cancer cells were able to inactivate different tumour suppressor genes like RB (retinoblastoma associated) protein, which is an important gatekeeper in the growth-and-division cycle. In this way, cell proliferation is not controlled any more.

The next crucial step is to avoid apoptosis. Tumour cells have lots of different strategies to avoid cell death. The most common way is knocking down the tumour suppressor TP53. This protein plays an important role in inducing apoptosis. It senses critical mutations during the cell cycle and can signal for cell death if necessary. Losing this function leads to multiplication of damaged cells. Another important characteristic of cancer is replicative immortality. Typically, cells lose part of their genetic information during the process of DNA replication. To avoid loss of important information, chromosomes possess multiple tandem hexanucleotide repeats at their ends, the so-called telomeres. These telomeres get shortened every cell cycle, meaning the cell has a given life span. Cancer cells can re-elongate these telomeres by activating the enzyme. In this way, the cancer cell becomes immortal.

Another important step for tumour tissue to keep growing is to ensure sufficient supply of nutrients and oxygen. Usually, the formation of new blood vessels from existing ones (=angiogenesis) is strictly regulated. Tumours, However, manage to increase the formation by upregulating certain genes like VEGF-A or members of the FGF family. Probably the most dangerous ability of cancer is to leave its primary site and spread to other parts of the body. Metastasis is a multistep process that includes the alteration of the cell-to-cell adhesion molecule E-cadherin. The mutation of this molecule allows cancer cells to leave the tissue and spread in the body.[3]

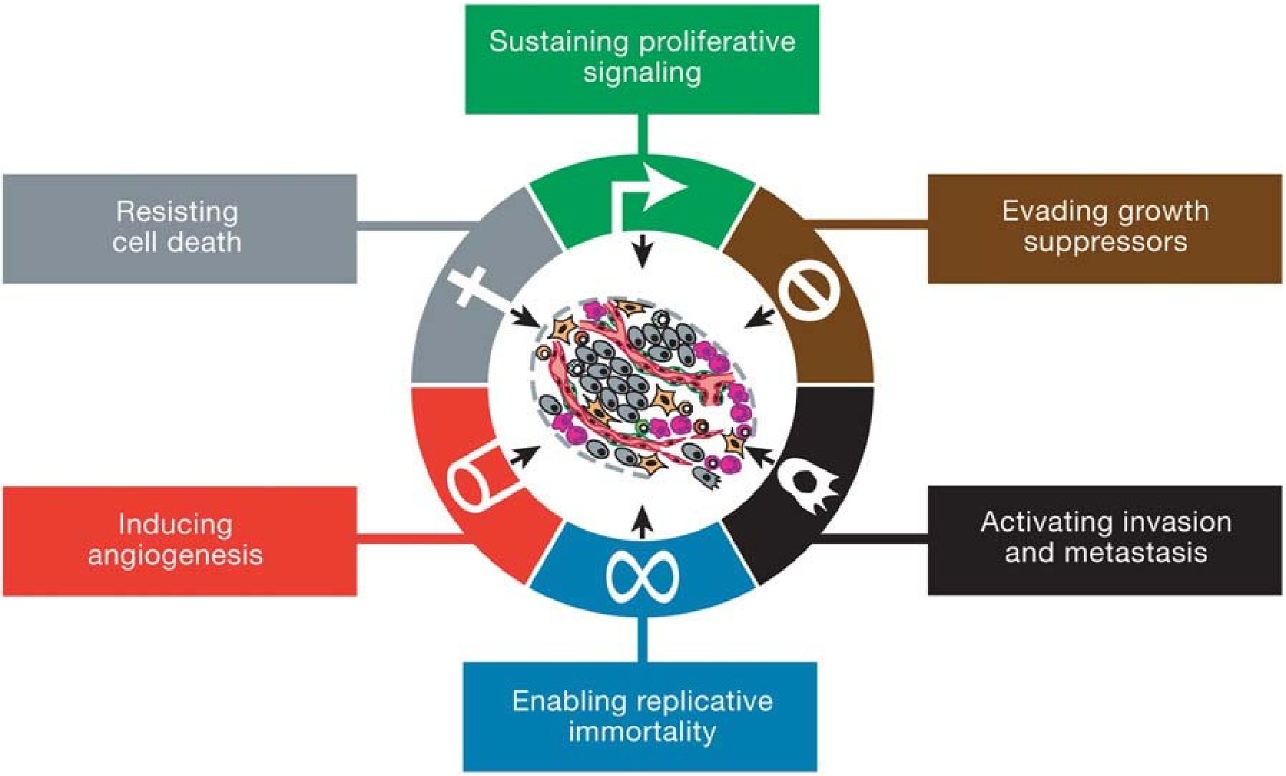

Sustaining proliferative signaling
Resisting cell death
Evading growth suppressors
Inducing angiogenesis
Activating invasion and metastasis
Enabling replicative immortality

Figure 1.2.1: Schematic view on the hallmarks of cancer. Figure adapted from: D. Hanahan and R. A. Weinberg, 'Hallmarks of cancer: the next generation', Cell, vol. 144, no. 5, pp. 646–674, Mar. 2011.[3]

1.3 Role of the FGF family in Tumour Progression

1.3.1 Basics

Fibroblast Growth Factors (FGFs) are a large family of secreted signalling proteins with 23 known members in humans. They play important roles in early stages of life during embryogenesis and organogenesis. Later on in adults they show homeostatic functions like wound repair, tissue maintenance, regeneration and metabolism. From these 22 proteins 18 are ligands and four are corresponding receptors with tyrosine kinase activity (FGFR1 – 4). The FGFR protein structure consists of the following domains: Two intracellular tyrosine-kinase domains (TK1 and TK2) for interaction with other mediates, one transmembrane domain (TM) and three immunoglobulin like domains on the outside of the cell (IgI, IgII, IgIII). Regulation of the ligand-receptor interaction is established through protein or proteoglycan cofactors and by extracellular binding proteins. Ligands bind on immunoglobulin-like domain III. FGFR1 – 3 can show two major splice variants in this domain, which are referred to as IIIb and IIIc. These splice variants have an impact on the ligand-binding specificity. Activation of FGFRs is coupled to intracellular signalling pathways like RAS-MAPK, PI3K-AKT and PLCɣ. For example, these pathways play important roles in regulating cell growth and proliferation.[4]

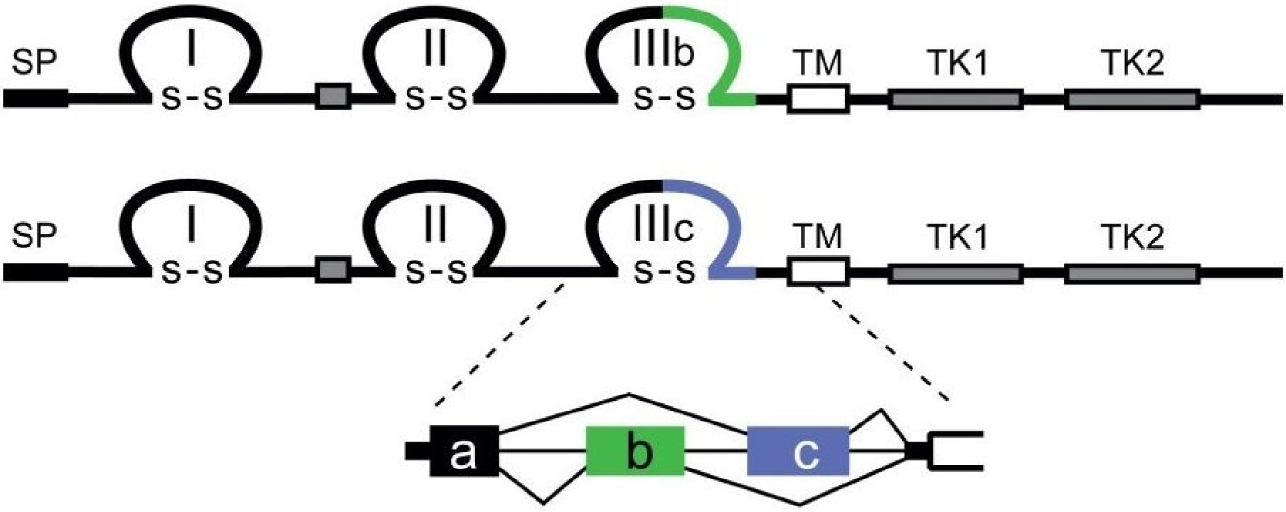


Figure 1.3.1: Schematic view on the structures of the two splice variants among the FGFRs. Original figure adapted from: D. M. Ornitz and N. Itoh, 'The Fibroblast Growth Factor signalling pathway', Wiley Interdiscip. Rev. Dev. Biol., vol. 4, no. 3, pp. 215–266, May 2015.[4]

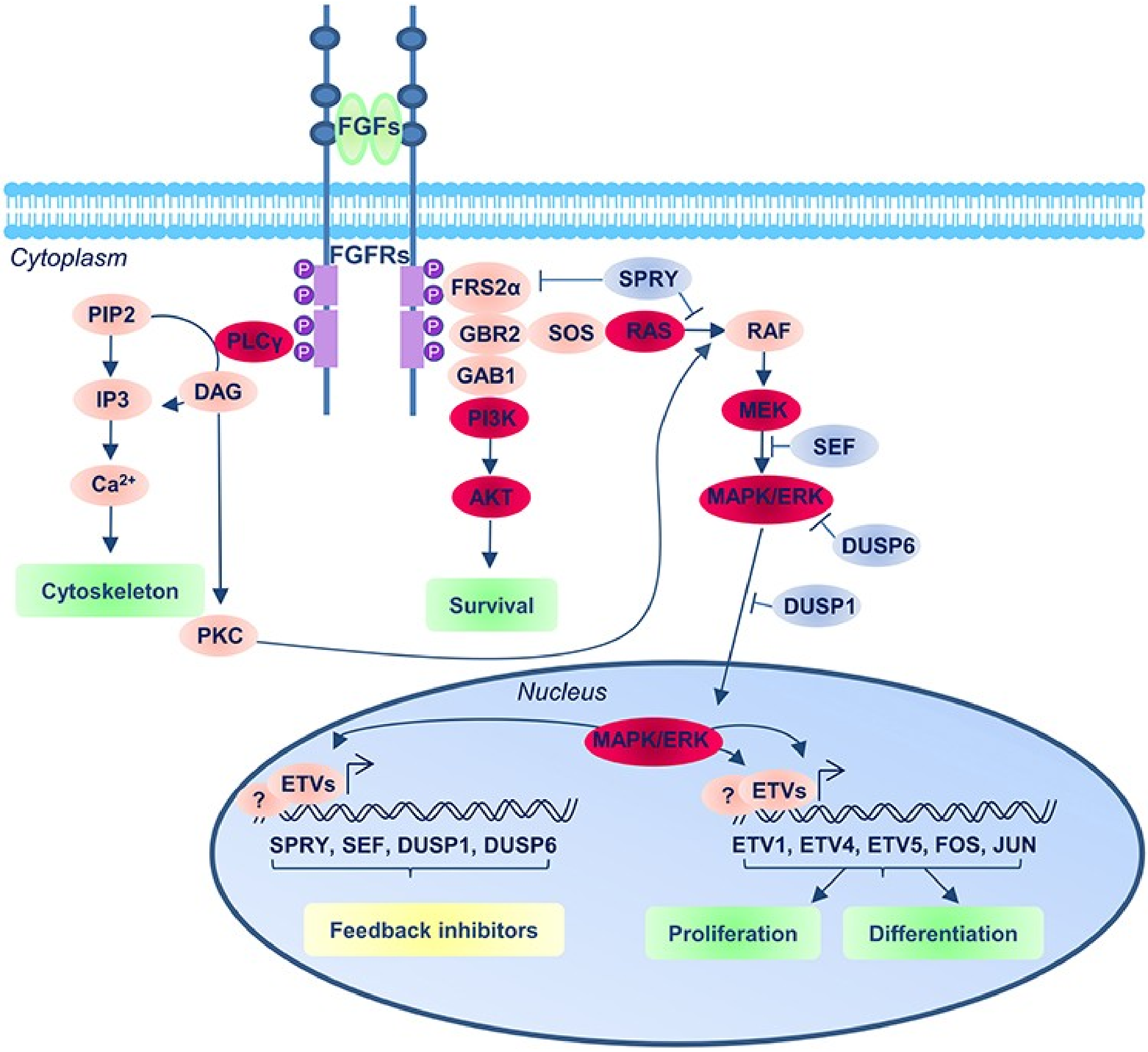


Figure 1.3.2: Simplified overview showing the different pathways FGFRs are involved in. Figure adapted from: R. Diez del Corral and A. V. Morales, 'The Multiple Roles of FGF Signalling in the Developing Spinal Cord', Front. Cell Dev. Biol., vol. 5, 2017.[57]

1.3.2 FGF Family in Hepatocellular Carcinoma

In the last years, several studies suggested new evidence for the importance of the FGF family in HCC. Tsunematsu et al.[5] showed that serum FGF2 levels were high in patients with chronic hepatitis C infection or liver cirrhosis and decreased during tumour progression. Stimulation with FGF2 led to higher expression of the membrane-bound major histocompatibility complex class I related chain A (MICA), which is a natural killer cell activating molecule. They concluded that FGF2 may play an important role in eliminating HCC cells by innate immunity.[5] According to another independent study, at least one of the members of the FGF8 subfamily (FGF8, FGF17 and FGF18) were upregulated in HCC patients. Due to the enhanced survival of HCC cells, the results of this study led to the conclusion that members of the FGF8 subfamily promote malignant behaviour and neoangiogenesis in hepatic tumours.[6] More studies also suggest FGFR3 and FGFR4 as potential therapeutic targets.[7], [8] Moreover, FGF19 is suggested to be an important factor for proliferation, cell survival and evasion of HCC.[9] In another recent study, the FGF9-FGFR3-IIIb/IIIc axis is considered to be a potential target for therapy.[10]

1.4 Aims of the Project

The aims of this project are to answer the following questions: Which differences exist in the gene expression of the FGF family between normal and cancer tissue? Is the expression of FGFs and FGFRs relevant for prognosis? How does high and low expression of FGFs and FGFRs impact biological processes and pathways?

To achieve this goal, high throughput sequencing data from the TCGA/GDC database will be analysed using bioinformatics tools. These tools will be implemented in an analysis pipeline using the R programming language and will be focused on a given set of genes, viz. the FGF family. Furthermore, this project shall give a small overview of the available software, how it works and what the possible benefits and limitations are.

2 Mathematical Background

2.1 Workflow

The workflow for the analysis pipeline was established step by step. Figure 2.1.1 shows the workflow steps with data pre-processing in light-blue, survival analysis in salmon-pink and differential gene-expression analysis, gene set enrichment analysis and gene regulatory network analysis in darker pink. The analyses in darker pink are grouped together because the results from every previous analysis step was used either as input or only as additional information for the next analysis step. (see Materials and Methods for more details)

Pre-procession of the data included downloading, filtering and standardization of the data as well as the deconvolution of it into three compartments. Deconvolution was integrated in this study with the idea of using the resulting data in an ensemble approach. In this way, it was possible to compare results from analyses on the convoluted and the deconvolved data and see in which ways they differ and which results show an overlap. The following chapters give insight into the mathematical principles of the software tools which were considered for this pipeline as well as the final choice for each.

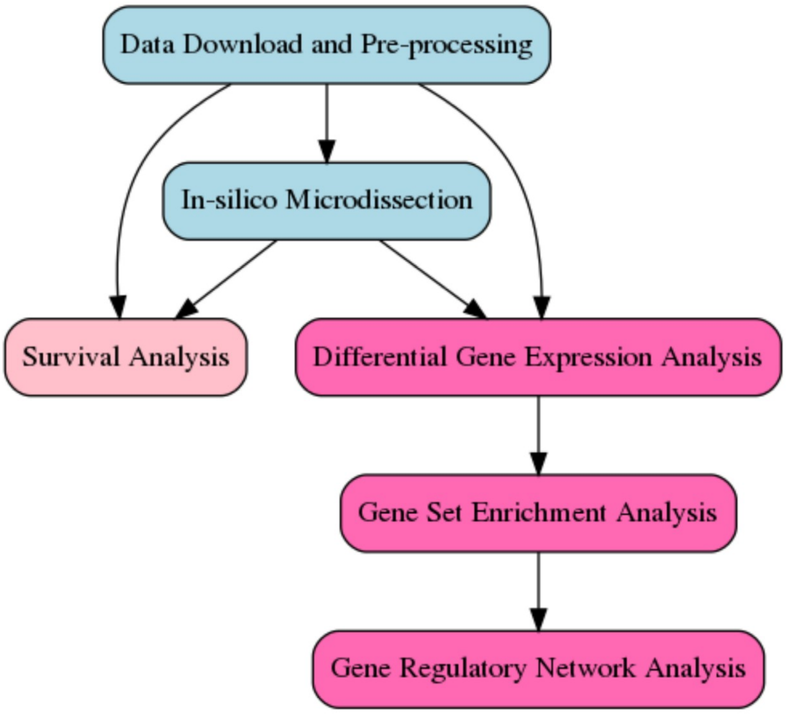

Data Download and Pre-processing
In-silico Microdissection
Survival Analysis
Differential Gene Expression Analysis
Gene Set Enrichment Analysis
Gene Regulatory Network Analysis

Figure 2.1.1: Schematic view of the workflow.

2.2 The Problem of Compartment Mixing in Tumour Samples

Every tumour is a heterogeneous tissue. It consists of tumour cells (also divided into stem cells and their offspring), stromal cells and immune cells (either attacking the tumour or supporting it). Therefore, analysing gene expression data can be difficult. Microdissection offers a solution to this issue.

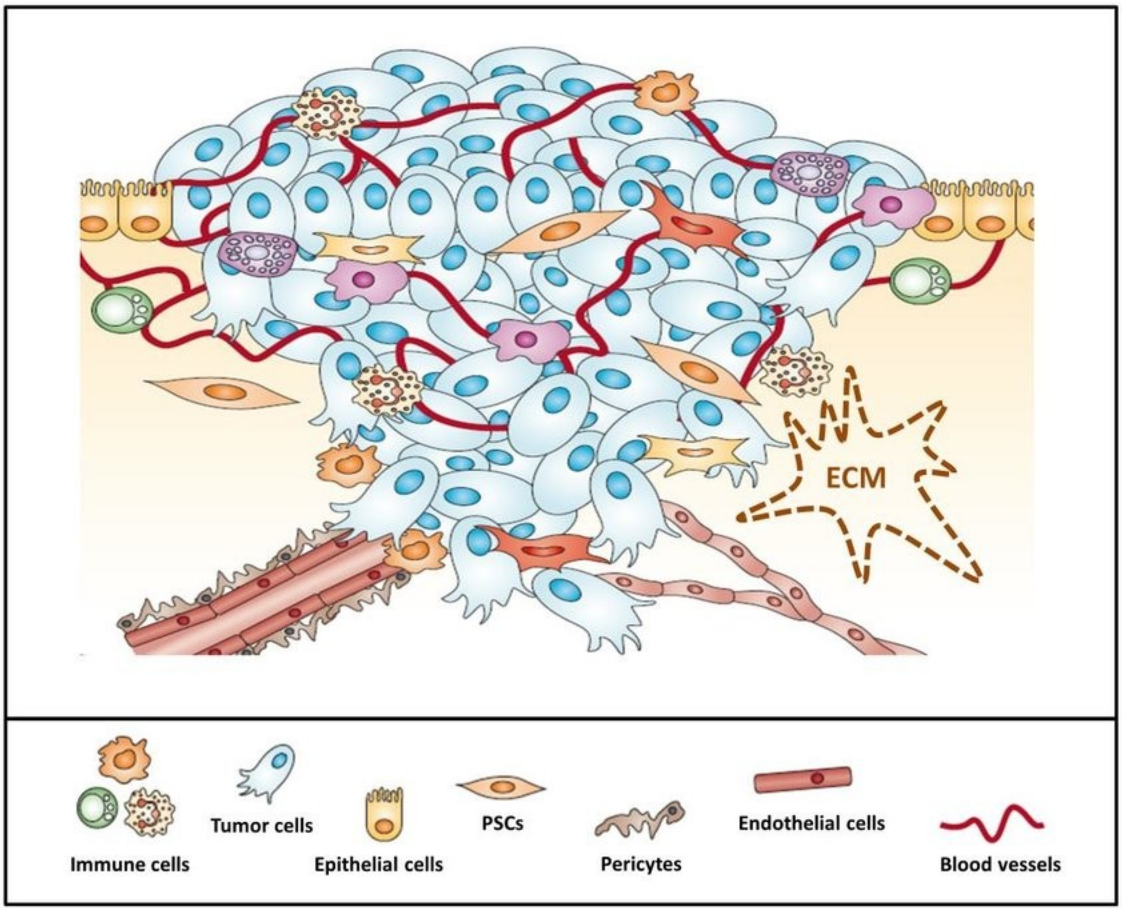


Figure 2.2.1: Schematic view on the tumour microenvironment. Figure adapted from: J. Kuen, 'Influence of 3D tumor cell/fibroblast co-culture on monocyte differentiation and tumor progression in pancreatic cancer', 2017[58]

2.2.1 Laser Microdissection

Laser Microdissection can be performed in the following ways. The first is Laser Capture Microdissection. A thermoplastic membrane attached to a plastic cap rests on the cells that need to be microdissected. The target cells are identified by using an inverted bright-field light microscope. Then the membrane on the identified cells is briefly melted by a low power, narrow beam infra-red laser. After some cooling, the cells are attached to the membrane and can easily be lifted off the tissue section.

The second approach is Laser Cutting Microdissection. Here the cells of interest are cut out of the surrounding tissue by using a narrow-beam ultraviolet laser. The laser "draws around" the cells of interest, and then they get collected in a tube. Depending on the used system, the cells may either be collected by "catapulting" them out of the tissue, because of gravity or by using a fine stainless-steel needle.[11]

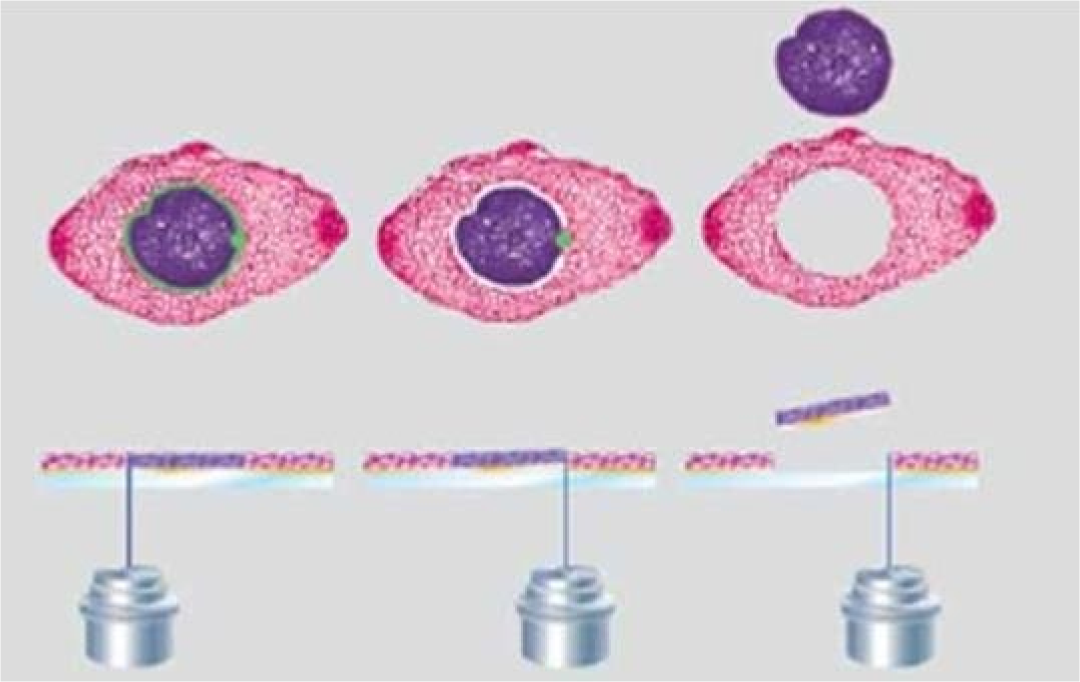


Figure 2.2.2: Simple illustration of a laser microdissection. Figure adapted from: M. Brazier, 'Microdissection of Alzheimer Brain Tissue for the Determination of Focal Manganese Accumulation', vol. 124, 2017, pp. 109–118.[59]

These approaches are needed to guarantee high level of purity and quality of cells for further analyses, but they happen to be expensive and time-consuming. The cells need to be prepared with expensive systems by specially trained personal. Therefore, algorithmic approaches are preferable.

2.2.2 In-silico Microdissection

Two general approaches for algorithms exist: Algorithms that estimate tumour purity and algorithms that calculate deconvolved gene expression profiles. Both algorithms need estimations as input. One algorithm which estimates purity is ABSOLUTE, for example. It uses somatic DNA copy number data to calculate the fraction of tumour cells in a tumour sample and based on that calculates the tumour purity[12]. Another approach is offered by the ESTIMATE package. The ESTIMATE algorithm performs a single-sample gene set enrichment analysis (ssGSEA), uses this data to calculate stromal- and immune-scores and combines those two scores to the so-called ESTIMATE-score, which then makes it possible to infer tumour purity for microarray data[13].

A common algorithm for deconvolving gene expression profiles is CIBERSORT, in which a machine learning approach is used. By providing an input matrix as reference, the CIBERSORT algorithm can deconvolve mixed datasets by using linear support vector regression[14]. A further approach is provided by the ISOPure algorithm. It is based on a statistical model in which the tumour profiles are multinomial distributed and the non-cancerous profiles are represented as convex combination. It has two major steps. In the first step, the “complete likelihood” function undergoes a maximum a posteriori (MAP) estimation by using numerical optimization. In step two the estimated values of their model get fixed and MAP estimation is used to optimize the cancer profiles[15]. The DeMixT algorithm was quite new at the time of this study and showed better performance than CIBERSORT according to Wang et al.[16] Therefore DeMixT has been chosen for this study.

2.2.3 Final Choice: DeMixT

In the DeMixT algorithm, the expression levels of the normalized measured data are modelled as linear combination of two or three components where one component is unknown:

$$Y_{ig}=\pi_{1,i}N_{1,ig}+\pi_{2,i}N_{2,ig}+\left(1-\pi_{1,i}-\pi_{2,i}\right)T_{1,ig}$$

The model is described as convolution of the log2-transformed normal-distributions of the components. Deconvolution of a dataset with three components is done in two steps, which both use Iterated Conditional Modes (ICM) for optimization.

Parameter Estimation using Iterated Conditional Modes (ICM)

The parameters of the unknown component can be divided into gene-wise and sample-wise sets and are conditionally independent. Therefore, the parameters of the unknown component can be derived iteratively (ICM) by using a golden section search in combination with successive parabolic interpolations.

Step 1

The two known components are merged and the third component's parameters are estimated relatively quickly in an artificial two-component setting using ICM. This method is called Gene Set-based Component Merging (GSCM).

Step 2

After fixing the parameters of the unknown component, the parameters of the two known components are estimated. Finally, three deconvolved expression profiles can be calculated by using the obtained parameters.[16]

2.3 Reverse Engineering: Gene Regulatory Networks

Determining relations between genes and thereby constructing a network from this information can be an intense computational task. Several approaches are available. In general, a network can be represented by an adjacency matrix with some sort of association measure between the nodes. The most recent approaches for reconstructing gene regulatory networks are correlation networks, polynomial and spline regression networks and mutual information (MI) networks.

2.3.1 Correlation Networks

One package, which constructs networks using correlation as association measure is the so-called WGCNA package (Weighted Gene Co-expression Network Analysis). Here the adjacency matrix is constructed by using either the Pearson-, Spearman- or biweight-midcorrelation.

The co-expression similarity s_{ij} is defined as: $s_{ij}=|cor(x_i,x_j)|$, where x is the expression profile. The according adjacency matrix A_{ij} can be defined either as unweighted network using a hard threshold τ with the expression $A_{ij}=\begin{cases} 1 & \text{if } s_{ij} \geq \tau \\ 0 & \text{otherwise} \end{cases}$ or as weighted network using the power of s_{ij} in the following expression $A_{ij}=s_{ij}^{\beta}$, $\beta \geq 1$. Later, the adjacency matrix can be transformed using the topological overlap measure (TOM). This transformation can help to create a more robust network by filtering out spurious or weak connections:

$$A_{TOM}(A^{original})_{ij}=\frac{\sum_{l\neq i,j} A_{il}^{original} * A_{l,j}^{original} + A_{ij}^{original}}{min(\sum_{l\neq i} A_{il}^{original}, \sum_{l\neq j} A_{jl}^{original}) - A_{ij}^{original} + 1}$$

The advantages of WGCNA are the relatively easy and fast calculations of the adjacency matrix and the straightforward implementation. A major disadvantage though is the fact that these correlation measures can only find linear relationships. Genes do not necessarily rely on linear relationships. A more robust approach was needed for this challenge[17], [18].

2.3.2 Polynomial and Spline Regression Networks

The WGCNA package also offers two other ways of calculating association measures. The first one uses polynomial regression:

$$E(y)=\beta_0 1+\beta_1 x+\beta_2 x^2 \ldots+\beta_d x^d = M\beta$$

This leads to:

$$R^2=cor(y,\hat{y})^2=cor(y,M\hat{\beta})^2$$

Here R² is the explained variance.

Spline regression can be seen as a local variant of the polynomial regression model. In this case, local means that the model tries to fit to a subinterval with range x. This is done by introducing so-called knots, which are transformed by a hockey stick function $()_+$, as additional parameters to the polynomial regression model.

E.g. some variable s → $(s)_+=\begin{cases} s & \text{if } s \geq 0 \\ 0 & \text{if } s < 0 \end{cases}$

The final model for a cubic spline with two knots, for example, would look like this:

$$E(y)=\beta_0 1+\beta_1 x+\beta_2 x^2+\beta_3 x^3+\beta_4 (x-knot_1)_+^3+\beta_5 (x-knot_2)_+^3$$

Because the networks constructed with a correlation coefficient showed no relationship between any of the genes and the workflow for polynomial and spline regression networks seemed to be still under development at the time of this study[18], the final choice was to use mutual information as association measurement for network construction.

2.3.3 Final Choice: Mutual Information Networks

Mutual Information (MI) is a statistical measure for the dependency of two variables. In contrast to correlation measures (e.g. Pearson correlation), it can determine non-linear relationships, which is of great value for analysing gene expression data. A major disadvantage though is the inability to tell if the found relationship is positively or negatively orientated. There are a few software packages which are used for estimating mutual information. For this project, the Aracne-AP package has been chosen[19]. The basic Aracne algorithm (Algorithm for the Reconstruction of Accurate Cellular Networks) is an extension of the RELNET algorithm[18].

Mutual Information

Mutual information for a pair of random variables x and y is defined as:

$$MI(x,y)=Entropy(x)+Entropy(y)-Entropy(x,y)$$

For a discrete variable, the entropy is defined as:

$$Entropy(t)=-\log(p(t_i))=-\sum_i p(t_i)\log(p(t_i))$$

The computation of MI is not a simple task. It requires estimation of the joint and marginal expression probability densities. To handle this problem, a Gaussian Kernel estimator was implemented. For a set of two-dimensional measurements $\vec{z}_i=\{x_i,y_i\},i=1....M$, the joint probability distribution (JPD) can be approximated with:

$$f(\vec{z})=\frac{1}{M}\sum_i h^{-2}*G(h^{-1}|\vec{z}-\vec{z}_i|)$$

Here G is the bivariate standard normal density.

Considering $f(x)$ and $f(y)$ as marginals of $f(\vec{z})$ the MI can be calculated in the following way:

$$MI(\{x_i\},\{y_i\})=\frac{1}{M}\sum_i \log\frac{f(x_i,y_i)}{f(x_i)*f(y_i)}$$ [20], [21]

Adaptive Partitioning

In the original Aracne implementation, the estimation was achieved by dividing the gene expression space into discrete bins of fixed size which thereby was called fixed bandwidth (FB) estimation. In this approach the number of bins had to be chosen in a preprocessing step. Aracne-AP, on the other side, uses an Adaptive Partitioning estimator, which chooses the number of bins automatically. The AP-algorithm divides the space recursively into quadrants at the means of the data. This recursion stops either if a uniform distribution between the new quadrants is met or if the split of a quadrant creates new splits with less than three data points[19], [21], [22].

Threshold for Mutual Information

Because MI is always non-negative even variables from random samples show positive results even though they are truly mutually independent. The algorithm handles this issue by shuffling the gene expression across all samples, calculating the MI for those genes and assigning a p-value to a threshold MI_0 by estimating the fraction of the estimates below MI_0 . This is done for different sample sizes and numerous gene pairs[20].

Data Processing Inequality

If genes g_1 and g_3 interact only through gene g_2 then the DPI states the following:

$$MI(g_{1,} g_3) \leq min[MI(g_{1,} g_2); MI(g_{2,} g_3)]$$

The DPI allows checking if the least of three MI values might be due to indirect interaction or not. In this scenario, Aracne starts with a network graph where each pair of genes with a MI value greater than MI_0 gets an edge. Then the algorithm examines all gene triplets and removes the edge with the smallest MI value[20].

2.4 Analysis Tools

2.4.1 Differential Gene Expression Analysis using Limma/Voom

Limma was chosen for this study because it performed well against other DGE analysis tools[23]. Furthermore, it is possible to correct for batch effects and comparing two (or even more) experimental conditions can be easily implemented. In the Limma approach, the gene expression is seen as a linear model:

$$E(y_{gi})=x_i^T*\beta_g$$

x_i^T is the vector of covariates and β_g represents $\log_2$ -fold-changes between experimental conditions. Using matrix terms, it can be written like this:

$$E(y_g)=X\beta_g$$

y_g is the vector of log-cpm values for gene-expression *g* and *X* is the design matrix[24].

Extremely variable data gets modelled by the Voom approach. In this case the fitted log-cpm values $\hat{\mu}_{gi}$ are converted to fitted counts:

$$\hat{\lambda}_{gi}=\hat{\mu}_{gi}+\log_2(R_i+1)-\log_2(10^6)$$

Here R_i is the geometric mean. Additionally, the LOWESS curve can be used to define a piecewise linear function $lo()$. The predicted square-root standard deviation of y_{gi} is the function value $lo(\hat{\lambda}_{gi})$. Precision weights are calculated by simply using the inverse of this function value[25].

The linear models for each gene are then put into an empirical Bayes framework, which allows borrowing information between genes and therefore moderate the residual variances. Final variance estimates for each gene are a compromise between the estimated variance for each gene and the global variability from the pooled ensemble of all genes[24], [26].

2.4.2 Gene Set Enrichment Analysis

The basic idea of GSEA is to check if certain genes are over- or under-represented in a gene set (e.g. pathway). This is done by a Kolmogorov-Smirnoff-like statistic. First, one has to supply a pre-ranked list of genes, for example according to their log-fold changes. Then the algorithm starts a running sum. Every gene of the pre-ranked gene list, which can be found in the gene set, increases the sum by a certain weighted value and every other gene decreases the sum in a similar way. This sum describes a curve where the optimum is the enrichment score (ES).

The next step is to check if the calculated ES is significant in comparison to random ESs. For this task, the gene names of the pre-ranked gene list are shuffled up to n times (e.g. n = 1000) and ESs are calculated. The p-value is estimated by comparing the original ES to the distribution of the ESs of the randomly permuted gene lists.

In the last step, every distribution of ESs, from the random gene lists, is normalized by its mean. This provides a null distribution of normalized enrichment scores (NESs). Finally, the FDR can be calculated[27].

2.4.3 Survival Analysis

Kaplan-Meier Curve

The Kaplan-Meier curve is a stepwise function which describes the descending probability of survival over a given time period. Estimating the intervals of this function is done with the following formula:

$$S(t_i)=S(t_{i-1})(1-\frac{d_i}{n_i})$$

$S(t_i)$ is the probability of survival at time t_i , d_i is the number of events (an event is categorical variable, e.g. “death”) at t_i and n_i is the number of patients alive before t_i .

Comparing two curves for significant difference can be achieved by performing the log-rank (LR) test:

$$LR=\frac{(O_1-E_1)^2}{E_1}+\frac{(O_2-E_2)^2}{E_2}$$

O_1 and O_2 are the total numbers of observed events and E_1 and E_2 are the calculated expected numbers of events for each group. LR then gets compared with the critical value of a chi-square distribution with one degree of freedom[28]–[30].

Stratification into high and low expressing Genes

Dividing a given continuous variable X (e.g. gene expression) into two categorical groups (e.g. low and high) can be achieved with a maximally selected rank statistic. $M = max_{\mu} |S_{\mu}|$, where M is the maximum of the standardized statistic and can be used as estimate for the unknown cut-point[31].

3 Material and Methods

3.1 Data

Samples of this dataset were obtained in the following way: Biopsy bio-specimens were gathered via surgical resection from HCC patients who had no treatment before the resection (e.g. ablation, chemotherapy, etc.). From the obtained tumour samples only some came with adjacent normal tissue samples as well. Every tumour and their adjacent normal tissue (if available) were controlled for quality. The controlled frozen section slides were either prepared by the Biospecimen Core Resource or by the Tissue Source Site. Sections of the samples were stained with haematoxylin and eosin and independently reviewed. Tumour specimens which showed histological characteristics of HCC and adjacent tissue specimens with no tumour cells were confirmed for further analysis. RNA and DNA samples were extracted from tissue specimens with a modified DNA/RNA AllPrep kit from QIAGEN[32].

Sequencing data was stored in BAM files and converted into FASTQ files with Biobambam[33] for further processing. Splice junction detection and alignment of the sequences was done using the STAR 2-pass alignment tool[34]. Quality assessment took place before alignment using FASTQC[35] and after the alignment using Picard Tools[36]. After these steps, HTSeq was used for gene expression quantification. HTSeq produces three outputs: first a raw read count expression matrix, second a FPKM normalized expression matrix and third a FPKM-UQ (upper quartile) normalized expression matrix. For this study the raw expression matrix as well as the clinical information about the patients (e.g. age, gender, etc.) and the annotation information about the sequenced genes (e.g. gene location on chromosome) were downloaded from the TCGA/GDC database[37].

3.2 Pre-processing

Filtering and Normalization

At the beginning gene IDs of the raw expression matrix needed to be mapped from Ensembl IDs to Entrez IDs, because most of the used software works with Entrez IDs. To achieve this goal, the AnnotationDbi package was used, which allows mapping from one ID to another[38]. Due to many Ensembl entries mapping to one Entrez entry, the raw expression matrix got filtered by choosing the maximum interquartile range (IQR) for every gene with the same Entrez entry. This procedure was implemented in the MapByIQR() function.

Further, filtering was done with respect to the counts per million (CPM). This filtering procedure had four steps: First, the values of the raw expression matrix were transformed to CPM using the cpm() function of the edgeR package[39], [40]. In the next step, a boolean matrix with the dimensions of the CPM-transformed expression matrix was created. Every entry of the boolean matrix was set to one (or TRUE) for each of their corresponding CPM-value being greater or equal than a given threshold. Next, a boolean vector was built with length equal to the number of rows of the CPM expression matrix. The row sums of the boolean matrix were divided by the total number of columns and every calculated value greater or equal than a second threshold (e.g. 10%) changed the according position in the boolean vector from zero to one. Finally, the raw expression matrix was filtered using the boolean vector. We implemented these steps in the FilterByCPM() function.

The filtered expression matrix was normalized with the calcNormFactors() function of the edgeR package. Here the TMM normalization (trimmed mean weight) is implemented[41]–[43]. This normalization method takes into account that most genes are not differentially expressed. The so-called TMM factor is computed, which can provide an estimate for correction of the library sizes. Normalization factors are re-scaled by the mean of the normalized library sizes. Finally, the raw read counts are divided by these factors[41]. Depending on the range of its library size, the normalized expression matrix either gets transformed to counts per million or by the Voom approach.

Building an Expression Set

An expression set is a data object with an expression matrix as core and clinical and annotation information as metadata[44]. The standardized gene expression matrix needed to be aligned sample-wise with the clinical information and gene-wise with the annotation information. First, the clinical data got filtered for all empty columns, columns with redundant information and columns with less valuable information content. Then the clinical information was filtered for patients with primary tumour and/or normal tissue. Finally, the expression matrix, the clinical information and the annotation information were aligned and stored in an expression set.

3.3 Deconvolution

DeMixT needs as input a normalized expression matrix and, depending on the setting, either one or two subsets of the expression matrix as starting guess. In this case a three-component setting was chosen, meaning two components needed to be provided as a starting guess. Creating the first subset was done by using only the tumour samples of the expression matrix and a table provided by TCGA, which contained information about the sample wise tumour purity[37]. Tumour purities were estimated from different algorithms and combined to a consensus. This consensus estimation was used to create the subset of the tumour expression matrix with high tumour purity samples. The second subset was created by selecting all normal tissue samples from the expression matrix with low immune score according to the ESTIMATE algorithm. Finally, DeMixT was started with the expression matrix and the two subsets.

3.4 Differential Gene Expression Analysis

In general, the Limma analysis pipeline consists of three main steps. First, a design matrix which divides the standardized expression matrix into groups has to be constructed. Then, a linear model to the expression matrix according to the design matrix has to be applied and, finally, moderated variances using the empirical Bayes framework have to be calculated[45].

For this study, two similar pipelines were created. One for the convoluted and the second one for the deconvolved data. The idea was to compare differentially expressed genes between tumour and normal tissue. In the convoluted data were 355 tumour samples, of which 49 samples also had adjacent normal tissue entries. These two groups were used to construct the design matrix in the Limma pipeline in two scenarios.

In the first scenario all 355 tumour samples were compared against the 49 normal tissue samples and in the second scenario a pairwise comparison was done. Put differently, 49 tumour samples were compared against their corresponding normal tissue samples. In the case of the deconvolved data, the computed tumour- and normal tissue expression matrices were used as groups. After this, the linear model was applied in both pipelines.

The Limma package provides two useful functions for comparing groups with each other. The first one is makeConstrasts(). This function constructs a contrast matrix using the coefficients of the linear model. The second one is the contrasts.fit() function. It re-calculates the coefficients of the linear model according to the contrast matrix.

Finally, empirical Bayes can be applied to the new linear model. The result is a table where every gene has the following values: the log-fold change, average expression, t-statistic, p-value, adjusted p-value and the B-statistic.

3.5 Gene Set Enrichment Analysis

The gene set enrichment analysis was performed with the three results from the differential gene expression analysis using the gseGO and the gseKEGG function of the clusterProfiler package[46]. These two functions perform the f-GSEA algorithm[47] on the gene sets from the gene ontology (GO) and the Kyoto encyclopedia of genes and genomes (KEGG) database respectively[48], [49]. Filtering these results for the genes of interest (FGFs and FGFRs) was done by simply selecting those significant gene sets which included at least one of the genes of interest.

To provide useful information for the gene regulatory network analysis, a consensus was made from all three GSEA results by intersecting the filtered significant gene sets with each other. This consensus was saved as an R-object.

3.6 Gene Regulatory Network Analysis

Constructing gene regulatory networks was done with the Aracne-AP software[19]. Aracne-AP is a collection of Java classes, which are gathered in a JAR file. It needs two inputs, a gene expression matrix and a list of regulators (transcription factors). Here, the list of regulators equals the genes of interest. The output is a table with four columns stored in a text-file: the regulator gene, the target gene, their calculated mutual information (MI) and the according p-value.

The Aracne-AP pipeline consists of three steps. First, a threshold for the MI, which depends mainly on the sample size is calculated, then bootstrap networks of the input matrix are constructed and finally all bootstrap networks are consolidated and combined into a final network.

This pipeline can be used to construct a network from a standardized expression matrix which is not log-transformed. However, after the deconvolution step in the analysis pipeline, there is not only one expression matrix but in total three valuable expression matrices: The convoluted data, the tumour compartment and the normal tissue compartment. Therefore, it was of great interest to get networks for each of these data sets. Furthermore, to get a better insight into the interaction of genes in each pathway, the consensus gene sets of the gene set enrichment analysis were used to build subsets for every data set. For this task, the Aracne-AP pipeline was parallelized in a bash script using GNU Parallel[50].

Some resulting networks were empty and removed afterwards. Adjacency matrices were constructed from these network tables with the mutual information value as edge weight. To compare networks against each other, row sums for each gene of the adjacency matrix were calculated and divided by the maximum of all row sums. Compensating for the missing information about the orientation of the relationship between two genes was achieved by adding the information of the DGE analysis. The information for the convoluted networks was adapted from the results of the DGE analysis with all samples. For the deconvolved compartments, the DGE results and their inverted form were used for the tumour compartment networks and the normal compartment networks respectively. Finally, graph objects were built with the adjacency matrix as core and row sums, symbol identifiers and information about regulation as vertex attributes. The final graph objects were saved as GML file, which made it possible to visualize them with Cytoscape[51].

3.7 Survival Analysis

Computing and visualizing survival curves for all genes of interest for the convoluted and deconvolved tumour samples was performed with the Survminer package[52]. The main process consists of five steps. First, the needed columns get extracted out of the data sets. Second, the cut-point of the gene expression dividing it into low and high expressing genes was evaluated with the surv_cut() function. Third, for each gene of interest all samples get divided into low and high expressing genes. Then, the survival curve was drawn and the p-value, indicating for a significant difference between the curves, was calculated. Finally, all p-values were gathered and were corrected using Benjamini-Hochberg[53] to address for multiple hypothesis testing.

4 Results

4.1 Differential Gene Expression Results

The Limma package offers a great function to explore the landscape of the expression data and to get an idea of how to design the analysis. This function is called plotMDS() and it shows the leading log fold changes of the provided expression data. The following MDS (Multidimensional scaling) plots show the leading log fold changes for the whole and the pairwise generated convoluted data set as well as for the deconvolved compartments.

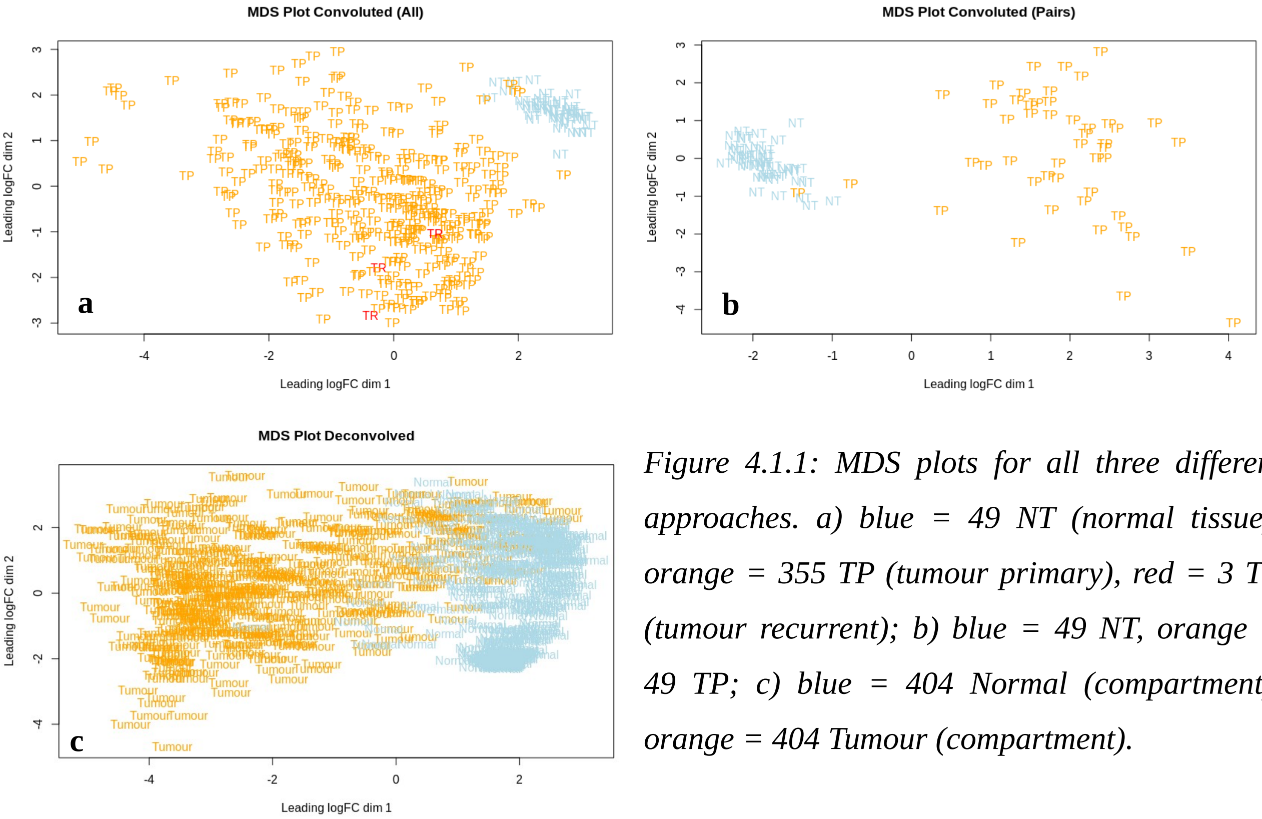


Figure 4.1.1: MDS plots for all three different approaches. a) blue = 49 NT (normal tissue), orange = 355 TP (tumour primary), red = 3 TR (tumour recurrent); b) blue = 49 NT, orange = 49 TP; c) blue = 404 Normal (compartment), orange = 404 Tumour (compartment).

The results of the DGE analysis are shown in the following heat maps. Here, upregulated genes are presented in orange and downregulated genes are presented in blue. Tables of the differential gene expression results for all three figures are provided in the supplementary material section.

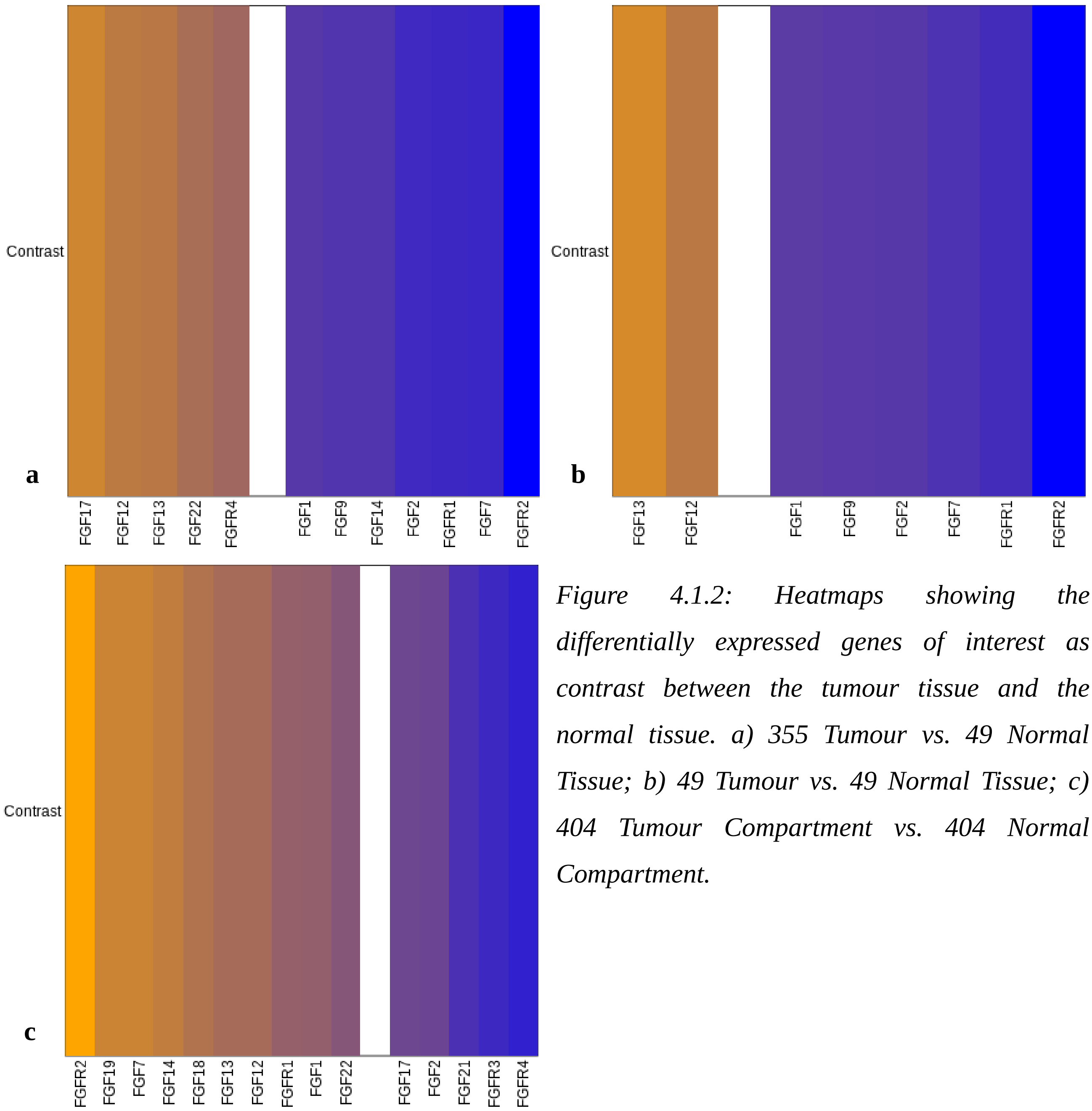


Figure 4.1.2: Heatmaps showing the differentially expressed genes of interest as contrast between the tumour tissue and the normal tissue. a) 355 Tumour vs. 49 Normal Tissue; b) 49 Tumour vs. 49 Normal Tissue; c) 404 Tumour Compartment vs. 404 Normal Compartment.

4.2 Gene Set Enrichment Results

The following histograms show the differences between the GSEA results of the three DGE analysis approaches and the normalized enrichment scores. Due to a low number of enriched KEGG gene sets, these histograms depict only the GO gene sets. While for the convoluted data the majority of the gene sets are significantly downregulated, they are equally distributed in case of the deconvolved compartments.

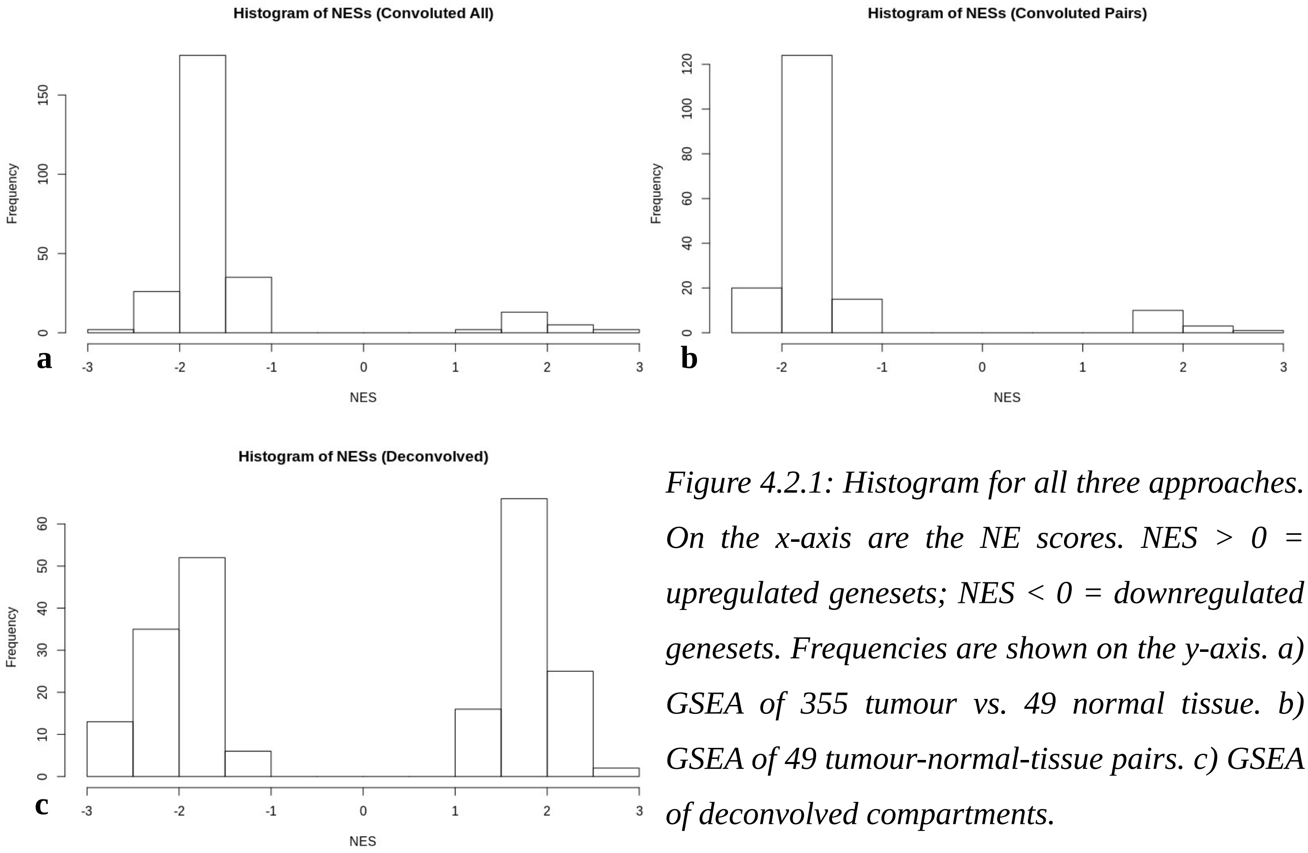


Figure 4.2.1: Histogram for all three approaches. On the x-axis are the NE scores. NES > 0 = upregulated genesets; NES < 0 = downregulated genesets. Frequencies are shown on the y-axis. a) GSEA of 355 tumour vs. 49 normal tissue. b) GSEA of 49 tumour-normal-tissue pairs. c) GSEA of deconvolved compartments.

Three gene sets have been chosen to show further the differences between the results of the convoluted and deconvolved data, namely “regulation of angiogenesis”, “regulation of vasculature development” and “PI3K-Akt signalling pathway”. Furthermore, the top 50 enriched gene sets for all GSEA results are provided in the supplementary material. While “regulation of angiogenesis” and “regulation of vasculature development” occur in all three different analysis approaches, “PI3K-Akt signalling pathway” only occurs in all convoluted samples and the deconvolved compartments. The following graphs show the pre-ranked list and the running score for the convoluted data set with all samples, the pairwise comparison and the comparison between the deconvolved tumour and normal compartment.

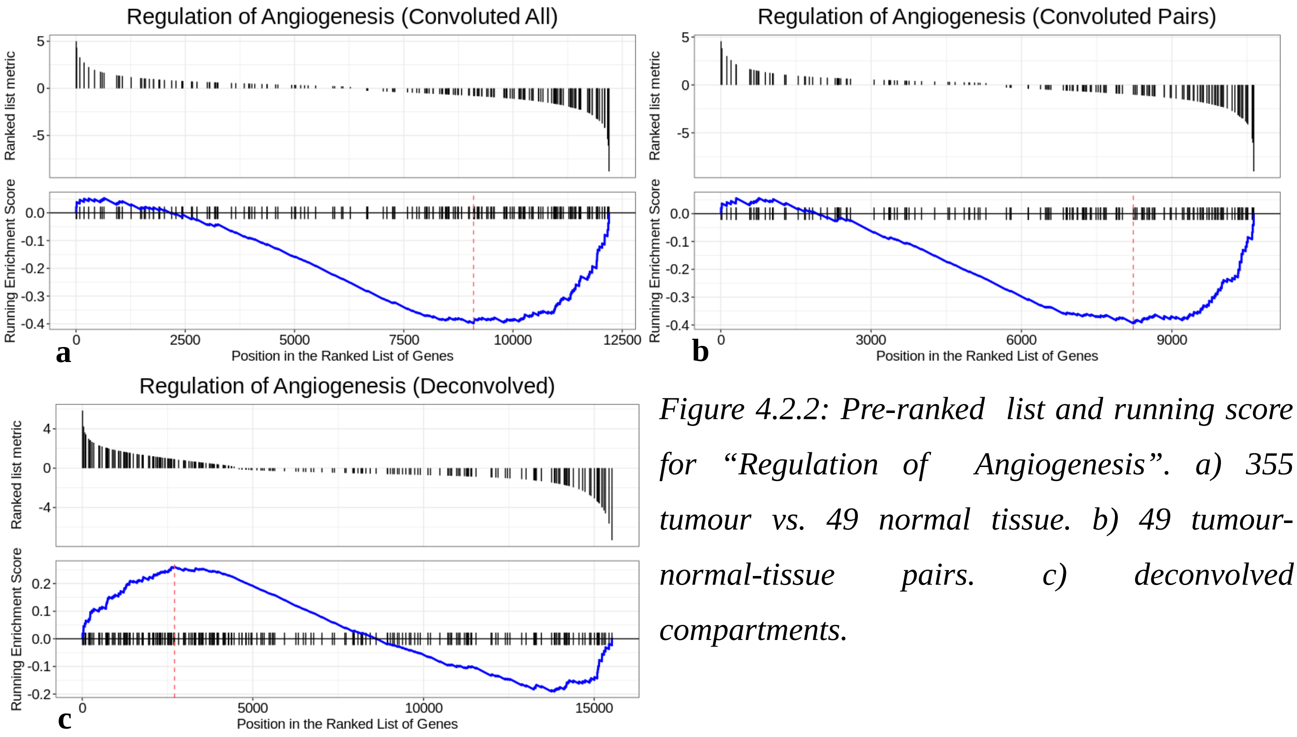


Figure 4.2.2: Pre-ranked list and running score for “Regulation of Angiogenesis”. a) 355 tumour vs. 49 normal tissue. b) 49 tumour-normal-tissue pairs. c) deconvolved compartments.

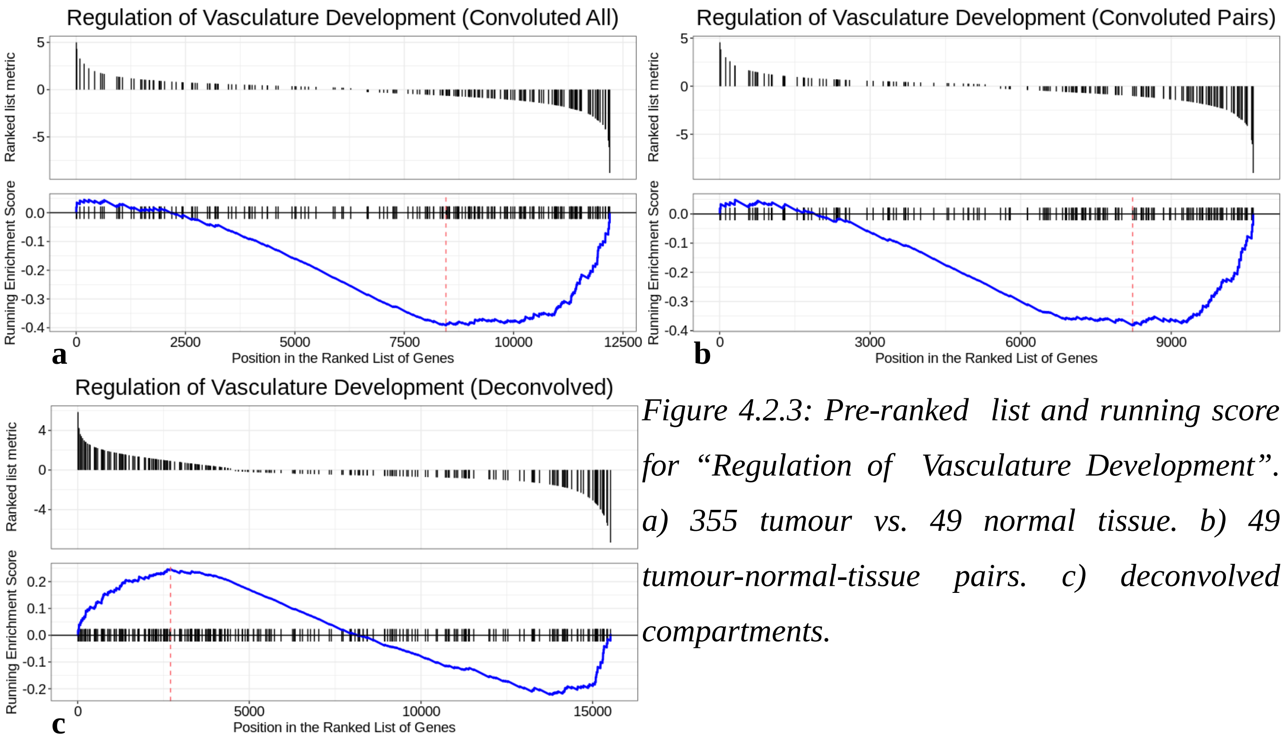


Figure 4.2.3: Pre-ranked list and running score for "Regulation of Vasculature Development". a) 355 tumour vs. 49 normal tissue. b) 49 tumour-normal-tissue pairs. c) deconvolved compartments.

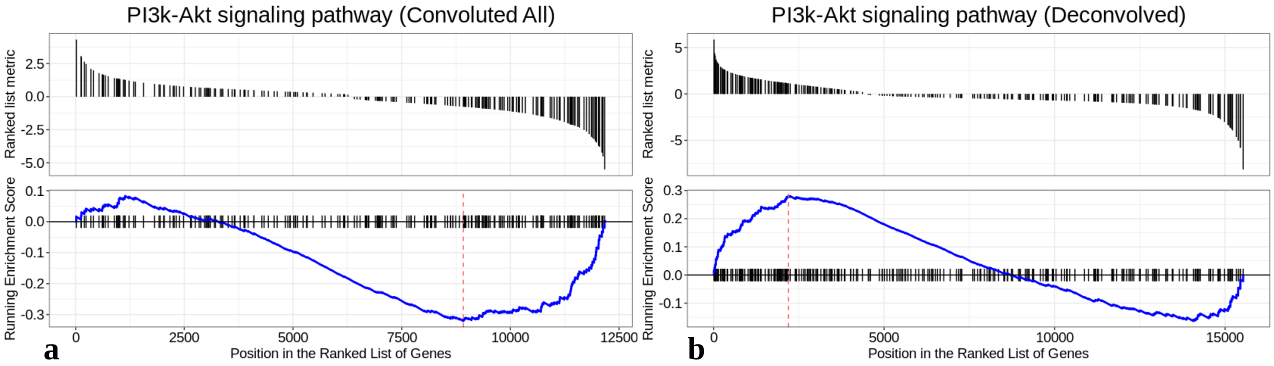


Figure 4.2.4: Pre-ranked list and running score for "PI3k-Akt signalling pathway". a) 355 tumour vs. 49 normal tissue. b) deconvolved compartments.

4.3 Gene Regulatory Network Results

To provide a better insight into the above shown gene sets (regulation of angiogenesis, regulation of vasculature development and PI3K-Akt signalling pathway), networks of the deconvolved normal and tumour compartment are presented here. Networks of the convoluted data are provided in the supplementary materials section. The size of the nodes (=genes) in the deconvolved normal and tumour compartment networks are weighted with the normalized row sums, meaning that the bigger the node, the greater the normalized row sum. Similarly to the plots of the DGE analysis, the colouring of the nodes encodes the gene regulation. Orange nodes represent upregulated genes, lightblue nodes show downregulated genes and white nodes are not significantly deregulated. The transparency and width of the edges is weighted by the mutual information value between two nodes. The darker and broader the edge is, the greater the mutual information value. These networks do not depict real pathways. They only show association between the genes of interest and other genes in the given genes sets.

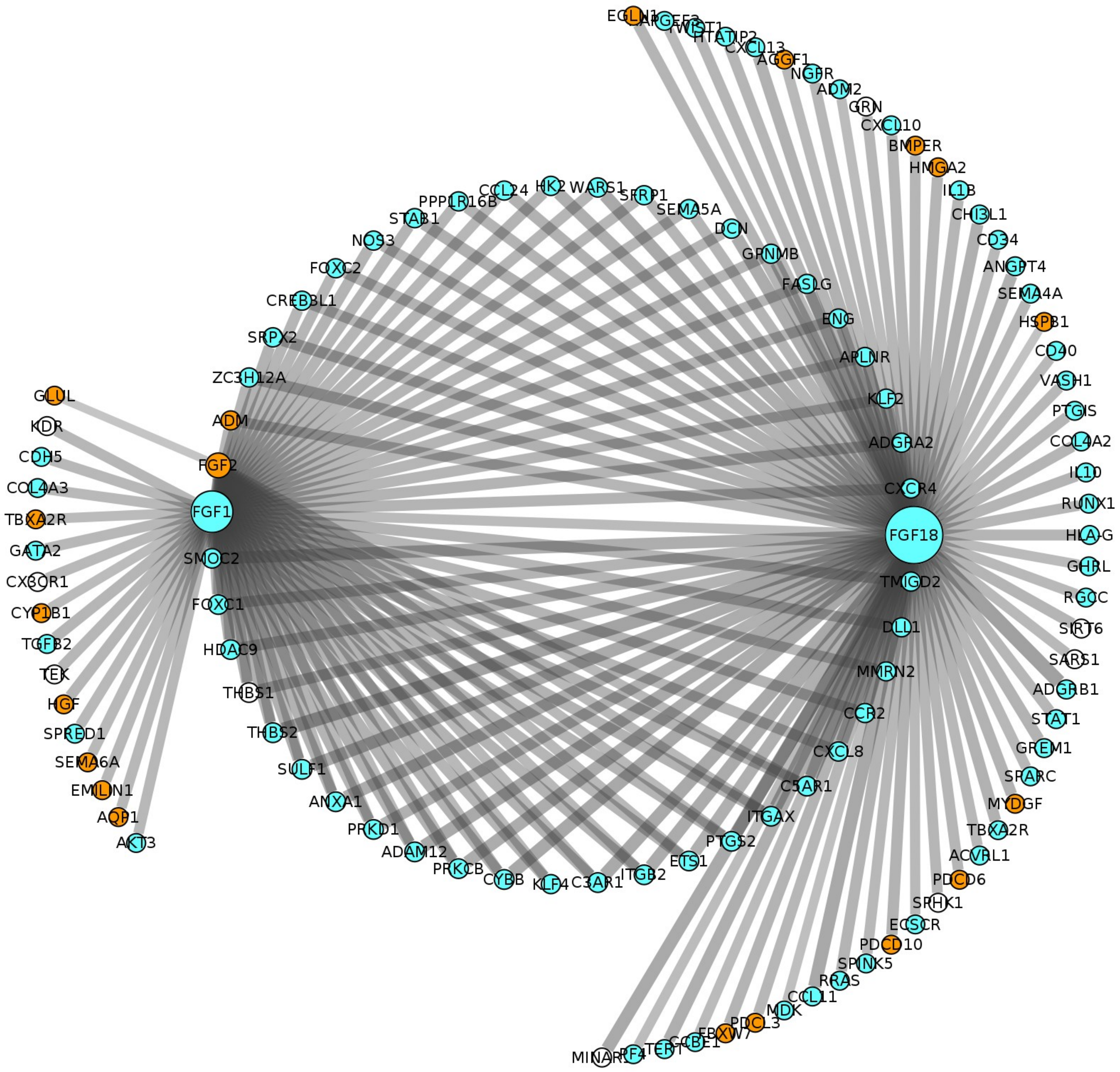


Figure 4.3.1: MI network of "regulation of angiogenesis" from the deconvolved normal compartment. The following members of the FGF family are included in this network: FGF18 (down, greatest row sum), FGF2 (up, smallest row sum), FGF1 (down, second greatest row sum).

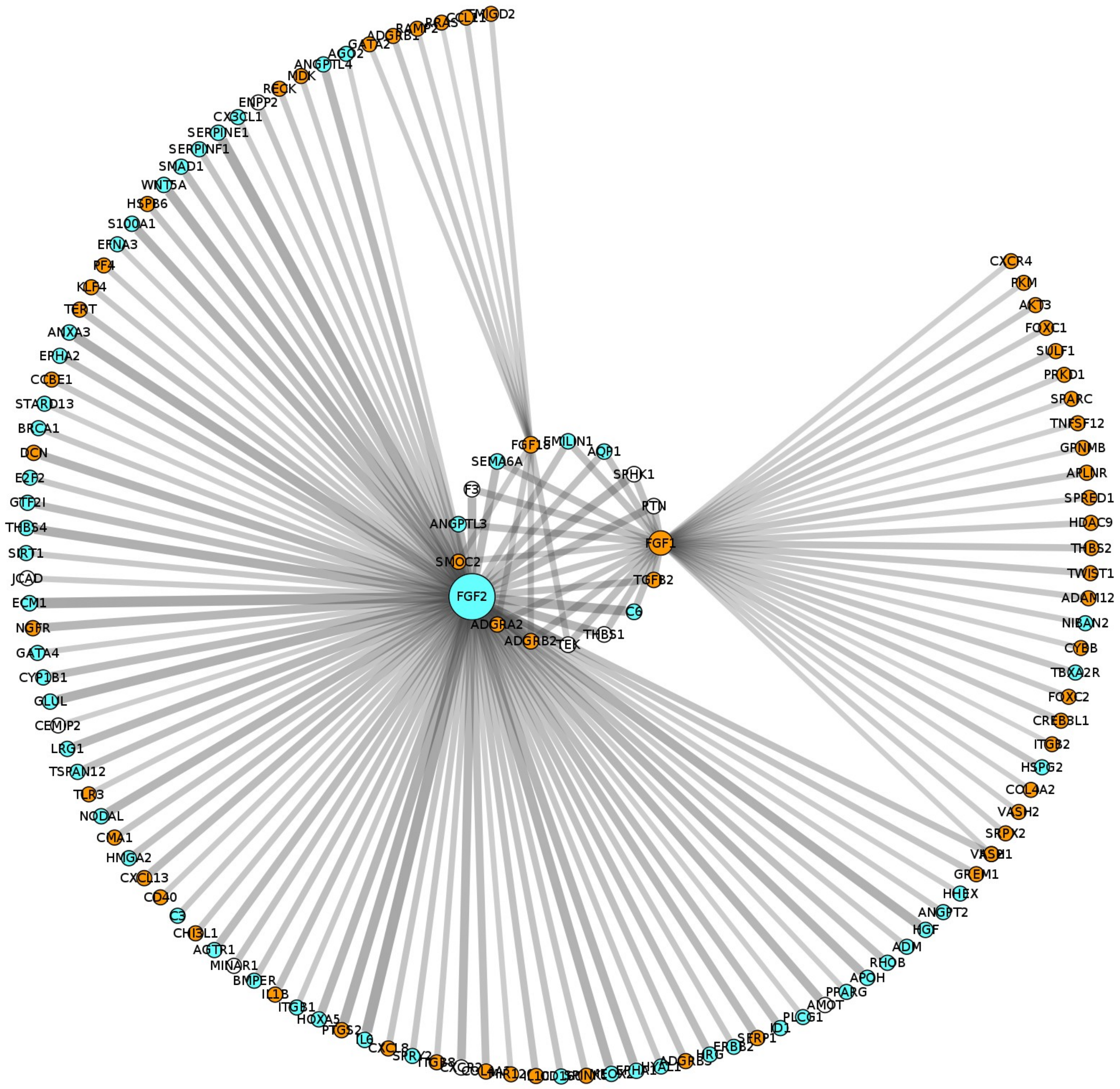


Figure 4.3.2: MI network of "regulation of angiogenesis" from the deconvolved tumour compartment. The following members of the FGF family are included in this network: FGF18 (up, smallest row sum), FGF2 (down, greatest row sum), FGF1 (up, the second greatest row sum).

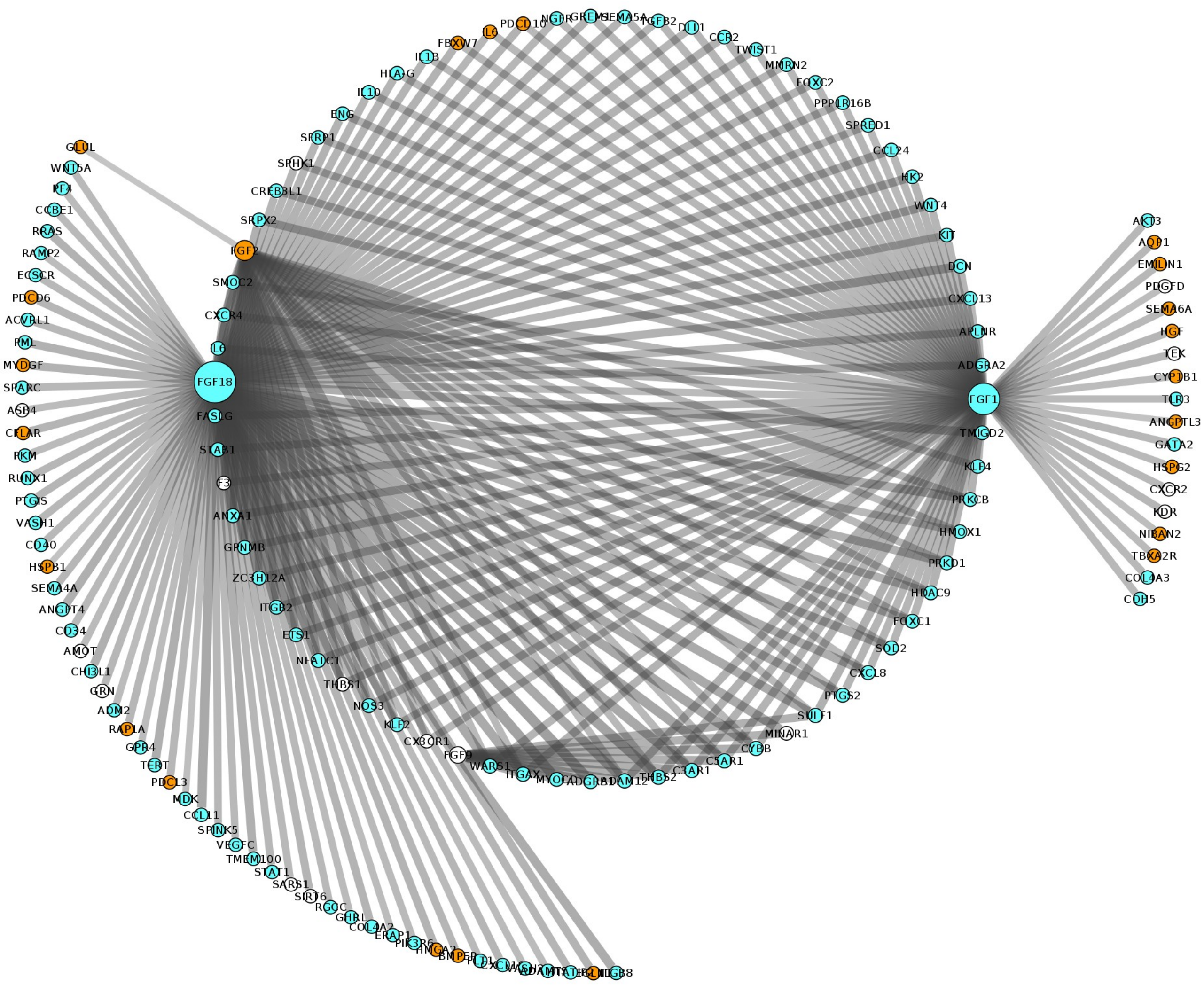


Figure 4.3.3: MI network of "regulation of vasculature development" from the deconvolved normal compartment. The following members of the FGF family are included in this network: FGF18 (down, greatest row sum), FGF2 (up, second smallest row sum), FGF1 (down, second greatest row sum), FGF9 (not deregulated, smallest row sum).

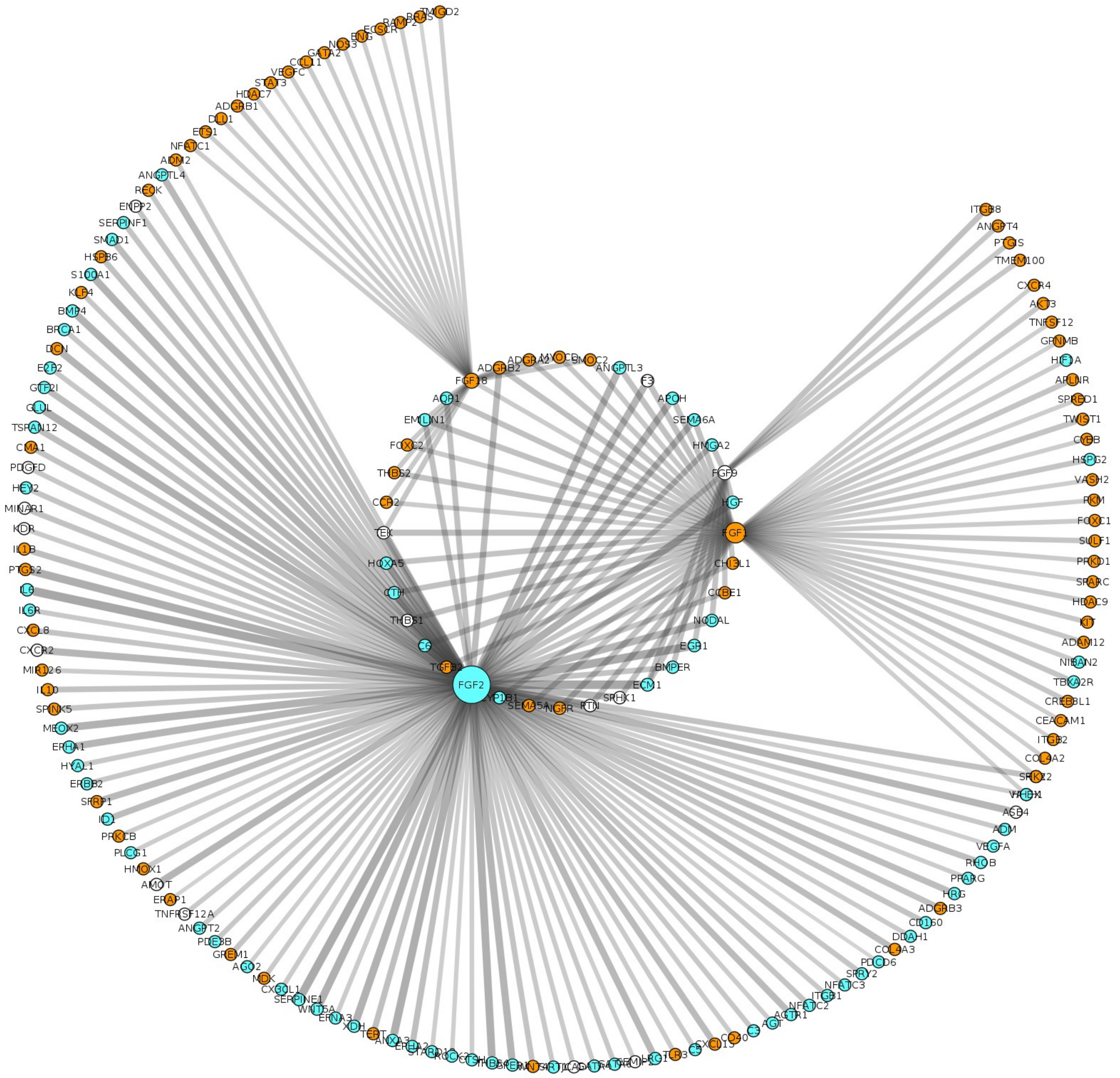


Figure 4.3.4: MI network of "regulation of vasculature development" from the deconvolved tumour compartment. The following members of the FGF family are included in this network: FGF18 (up, second smallest row sum), FGF2 (down, greatest row sum), FGF1 (up, second greatest row sum), FGF9 (not deregulated, smallest row sum).

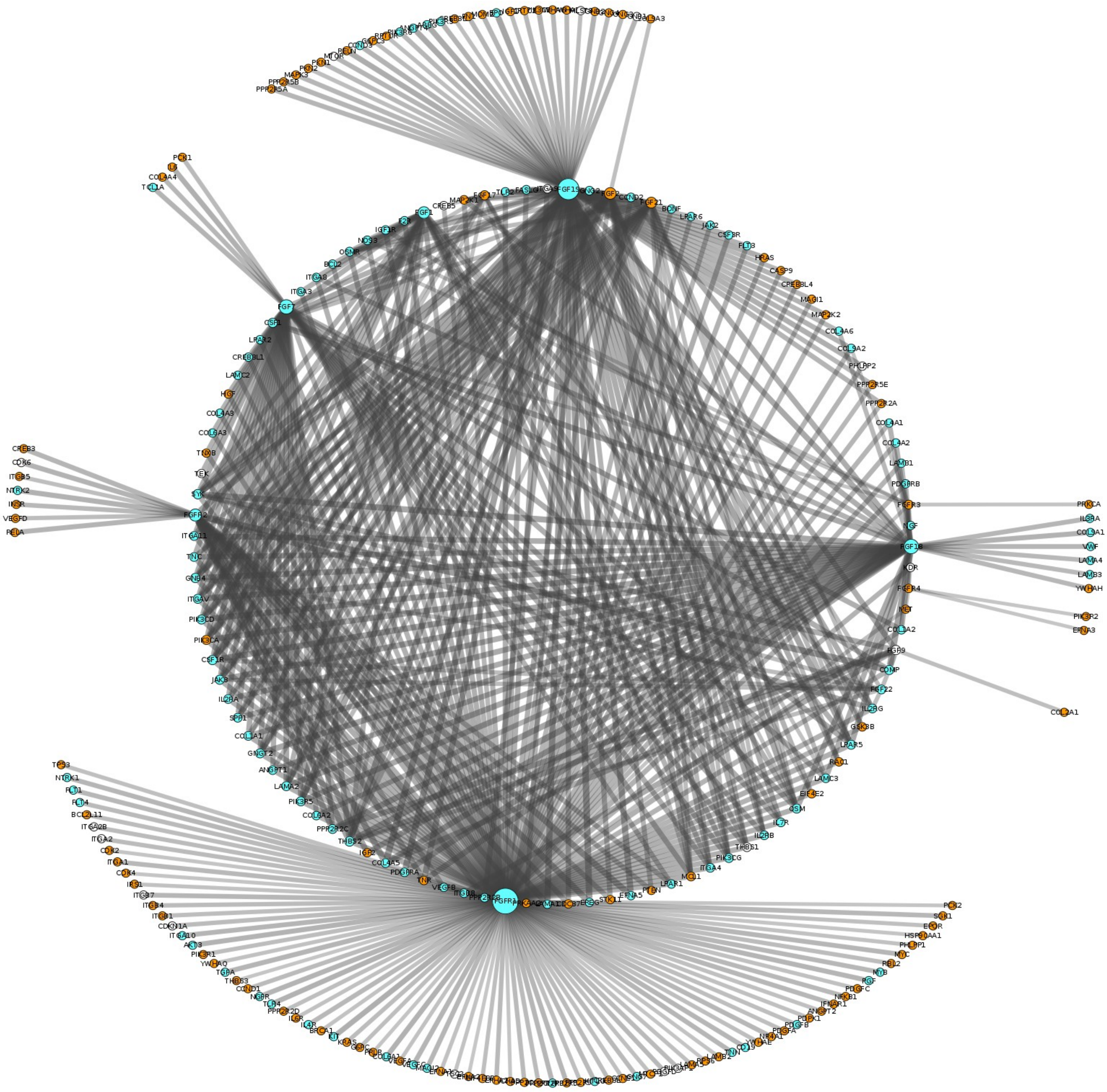


Figure 4.3.5: MI network of "PI3K-Akt signalling pathway" from the deconvolved normal compartment. The following members of the FGF family are included in this network: FGF1 (down), FGF2 (up), FGF7 (down), FGF9 (not deregulated), FGF17 (up), FGF18 (down), FGF19 (down, second greatest row sum), FGF21 (up), FGF22 (down), FGFR1 (down, greatest row sum), FGFR2 (down), FGFR3 (up), FGFR4 (up).

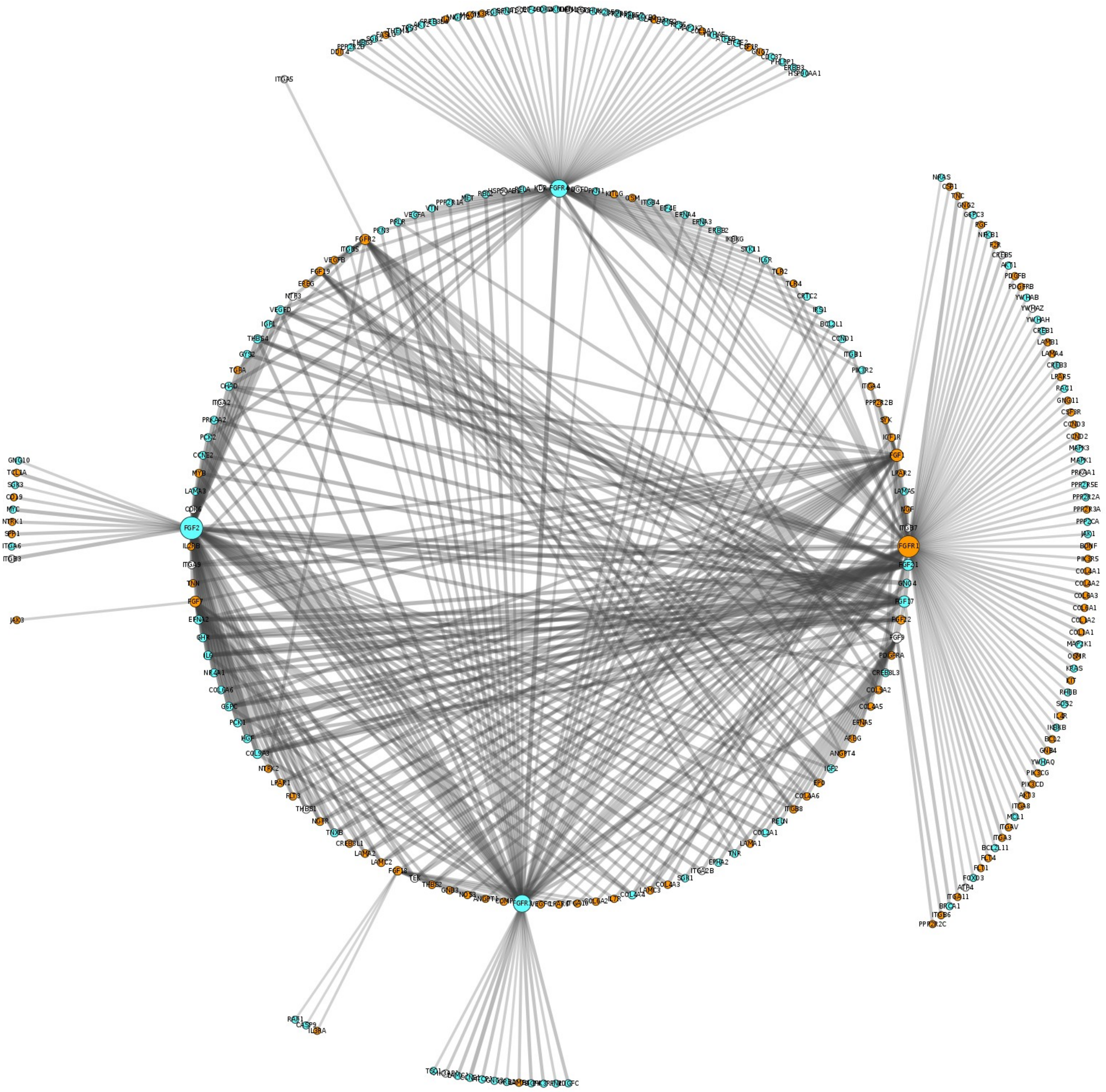


Figure 4.3.6: MI network of "PI3K-Akt signalling pathway" from the deconvolved tumour compartment. The following members of the FGF family are included in this network: FGF1 (up), FGF2 (down, greatest row sum), FGF7 (up), FGF9 (not deregulated), FGF17 (down), FGF18 (up), FGF19 (up), FGF21 (down), FGF22 (up), FGFR1 (up, second greatest row sum), FGFR2 (up), FGFR3 (down), FGFR4 (down, third greatest row sum).

4.4 Survival Analysis Results

In this section, only those results are shown which occur in both the convoluted and the deconvolved tumour data. All other results showing a significant difference between high and low expressing genes are provided in the supplementary material. The max-rank plots show the computed distributions after estimating the optimal cut-point. The survival plots show the comparison between the estimated low (= blue colour) and high expressing genes (= salmon-pink colour). Adjusted p-values are given for each plot individually.

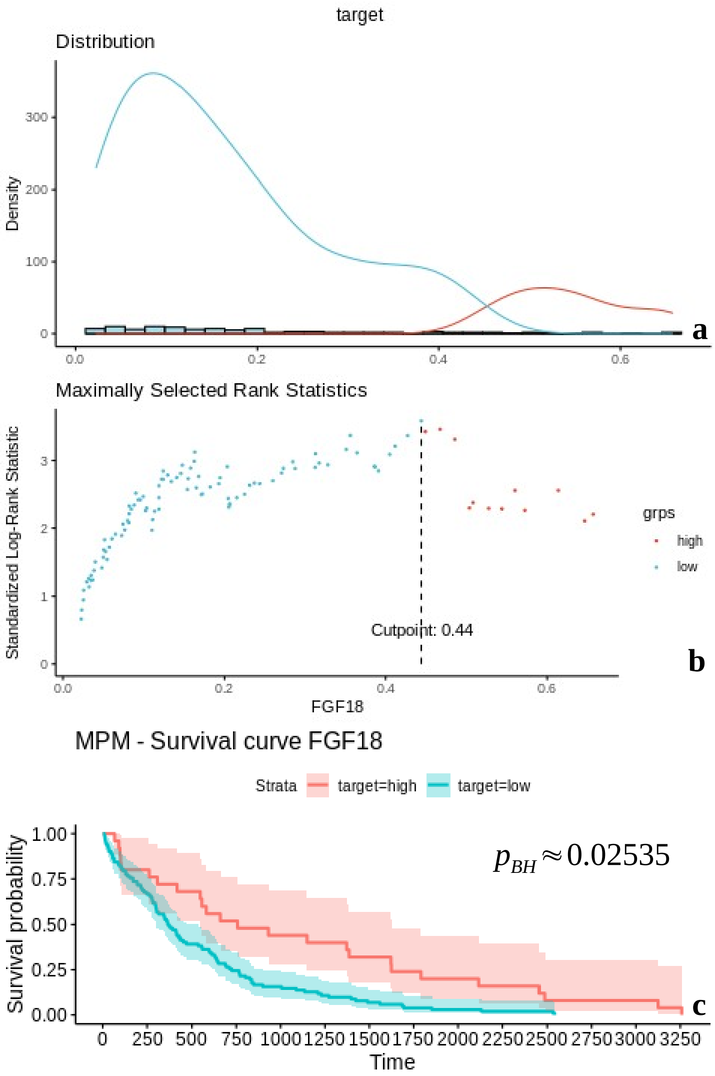


Figure 4.4.1: Max-rank plots and survival plot of FGF18 in the convoluted data set. a) Gene expression distribution divided by the cut-point. b) Maximal rank statistic at 0.44. c) Survival plot → x-axis = time (in days), y-axis = survival probability.

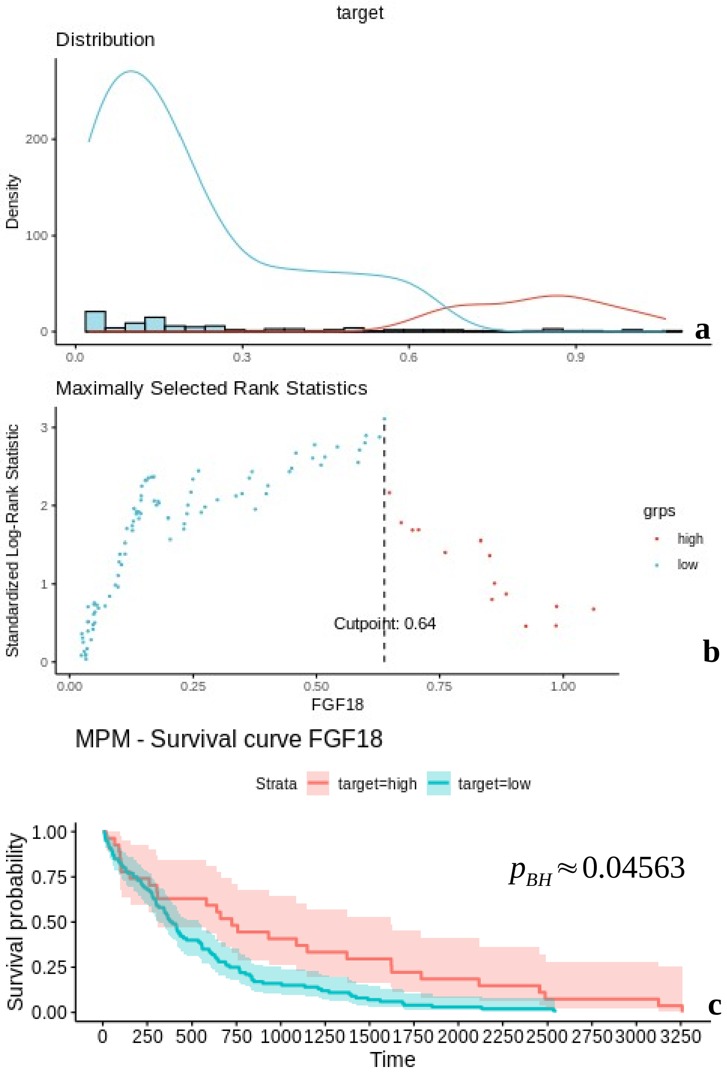


Figure 4.4.2: Max-rank plots and survival plot of FGF18 in the deconvolved data set. a) Gene expression distribution divided by the cut-point. b) Maximal rank statistic at 0.64. c) Survival plot → x-axis = time (in days), y-axis = survival probability.

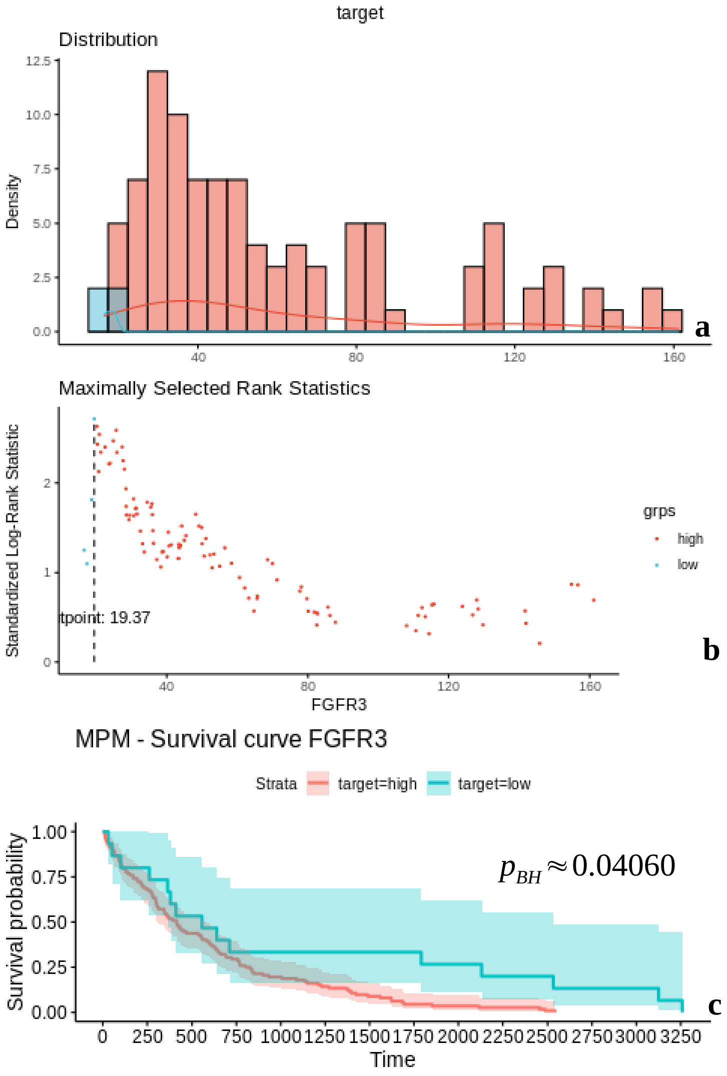


Figure 4.4.3: Max-rank plots and survival plot of FGFR3 in the convoluted data set. a) Gene expression distribution divided by the cut-point. b) Maximal rank statistic at 19.37. c) Survival plot → x-axis = time (in days), y-axis = survival probability.

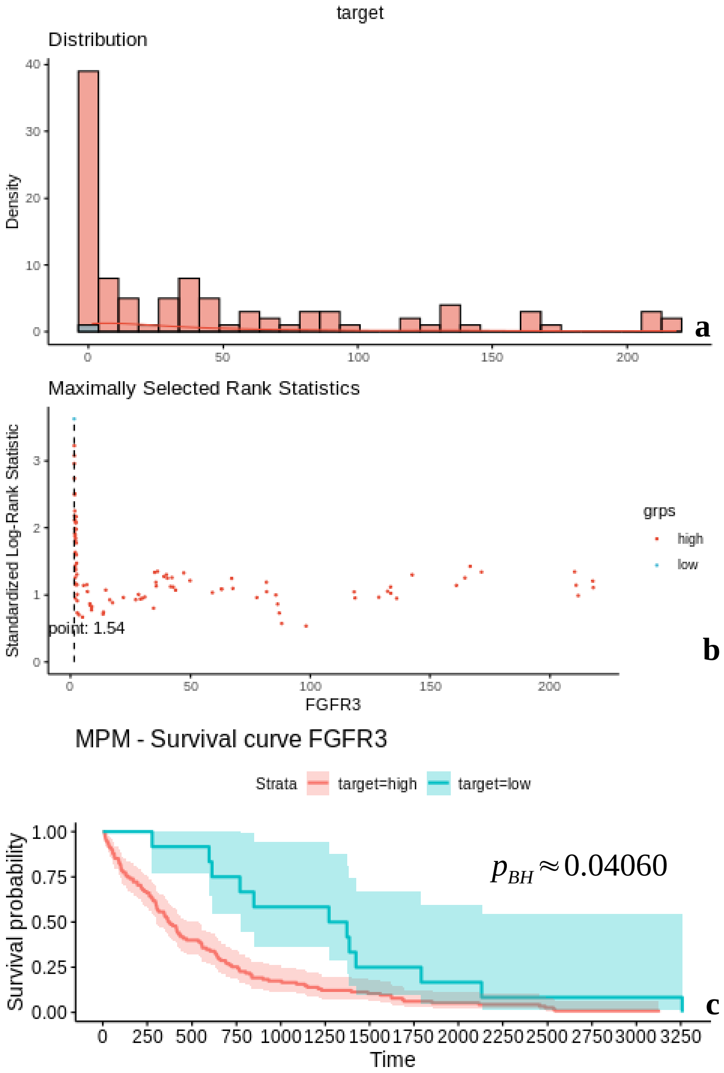


Figure 4.4.4: Max-rank plots and survival plot of FGFR3 in the deconvolved data set. a) Gene expression distribution divided by the cut-point. b) Maximal rank statistic at 1.54. c) Survival plot → x-axis = time (in days), y-axis = survival probability.

5 Discussion

5.1 Software

In this thesis, we have implemented an analysis pipeline in R and some parts as bash script to analyse the gene expression data of HCC with respect to the FGF family. First, the expression matrix was mapped from Ensembl to Entrez ID, then filtered for noisy genes considering the counts-per-million values, normalized by TMM and transformed to log-counts per million.

Because of multiple values for one gene or even genes without a complementary, mapping from one gene identifier to another can be challenging. The avereps() function of the Limma package does provide a solution for this problem. However, it can only be applied after the log-transformation step, meaning that all calculations of the normalization procedure are partly performed on genes with multiple identifiers or genes which get removed anyway. MapByIQR() on the other side deals with this issue by mapping from one gene identifier to another before the normalization. This method works for cases where only a few genes with multiple identifiers occur. However, it is inefficient for data with larger sets of multiple gene identifiers.

In general, the filterByExprs() function of the edgeR package should be good enough for filtering. Nevertheless, we tried a different approach, namely FilterByCPM(). In the case of the TCGA HCC data set, ~25 000 genes get filtered and ~17 000 genes remain with the FilterByCPM() approach. The filterByExprs() function removes around 5000 more genes by default, but by relaxing the parameters this function can be used as well.

Deconvolving data to get refined expression matrices for further analysis steps is a great accomplishment in the analysis of RNA-Seq data. The DeMixT algorithm does a great job, but has some disadvantages. Despite the documentation, it was not entirely clear how to implement the DeMixT software correctly. It took a lot of time and several trials with the help of different forums to get the deconvolution working properly. Especially getting the right information for the two compartments as starting guess in addition to the expression matrix as input was difficult. For this study, the additional information was provided using the Estimate package, which gives hints for purity in form of the stromal- and immune-score as well as downloaded tumour purity data from the TCGA database.

Another issue with the DeMixT algorithm is its gene filter function. Before the actual deconvolution process begins, DeMixT removes all genes with at least one value equal to zero. In our case, this step would have removed around 5000 genes. One way of avoiding this is to add pseudo counts, but genes with lots of small values lead to a program crash. Therefore, the GeneSaver() function was created. It removes all genes which have more than a given percentage of zero-values (default is 25%) and adds a pseudo count to the remaining expression matrix. Furthermore, it can keep genes of interest if a list, containing the gene identifiers, is provided. Due to lack of time, it was not possible to test different deconvolution methods and look for an overlap in results. However, it is clear that in-silico micro dissection is a promising approach to deal with data from heterogeneous tissues.

Limma offers an easy-to-use framework to perform differential gene expression analysis. The implementation of the analysis steps is easy and the software package allows for designing even more complex models, depending on the asked research question and available data. The results of the DGE analysis can be provided for gene set enrichment analysis. A good tool for this kind of analysis is the ClusterProfiler package. It provides lots of functions for enrichment analysis, post-processing of the results and visualization.

Calculating gene regulatory networks seemed challenging at first. The results of the WGCNA software package suggested that no linear relations could be found between almost all the genes. Therefore, Aracne-AP was used, a Java software which constructs gene regulatory networks with mutual information as association measurement. Using Java software within R is theoretically possible but was problematic for this study. That is why the construction of the MI networks was implemented in a bash script. The general idea of calculating gene regulatory networks with gene sets considering the genes of interest from the data sets was computationally expensive and needed improvements. Using only significantly enriched gene sets from the previous GSEA and parallelize computations with GNU Parallel made it possible to calculate gene regulatory networks in a reasonable amount of time. Nevertheless, for future work, I would consider the software package SJ-Aracne[54]. It is a python implementation of the Aracne-AP algorithm, which has a smaller memory footprint than the original Aracne-AP. Moreover, switching between R and Python should be less problematic. At the time of this study, it was, however, not possible to install it and make it work properly.

The Survminer package is a good choice for survival analyses. It is well documented and the implementation is straightforward. Survival analysis of gene expression data has its downside though, namely the categorization of continuous variables which leads to loss of information[55].

5.2 Biology

In this study, we analysed the role of the FGF family in HCC. The original convoluted data set with 355 tumour samples and 49 normal tissue samples makes differential expression analysis difficult due to unbalanced group sizes. Nevertheless, the model used by the Limma software is robust enough for this type of issue. Moreover, three different approaches for analysis were used: first, using the whole data set and compare 355 tumour samples against 49 normal tissue samples; second, a more conservative approach to compute differential gene-expression with pairs (meaning 49 normal tissue samples with their according tumour samples) and third, comparing the computationally generated tumour and normal tissue compartments from the DeMixT software. In this study, the DeMixT algorithm calculated gene expression profiles where no infiltration of stromal and immune cells occurs. The results of the different approaches show only overlap on certain genes. FGF12 and FGF13 are in all three approaches significantly overexpressed and FGF2 is significantly under expressed. Gene expression of those three genes is different between normal and cancer tissue in all three approaches suggesting that these genes might be potential candidates for further research.

The results of the gene set enrichment analysis show differences between the convoluted and the deconvolved data. More gene sets of the deconvolved data are upregulated than they are for the approaches with the convoluted data. Three of these differing gene sets are “regulation of angiogenesis”, “regulation of vasculature development” and “PI3K-Akt signalling pathway”. While the first two gene sets were significantly enriched for all three analysis approaches, “PI3K-Akt signalling pathway” was not enriched in the case of the pairwise compared data. Interestingly, compared to the results of the convoluted data, the results for the deconvolved data suggest up-regulation of these three gene sets.

Gene regulatory networks allow for a deeper insight into the interactions of the FGF family with all other genes in the above-mentioned gene sets. In both “regulation of angiogenesis” and “regulation of vasculature development” interaction with FGF18 decreases and interaction with FGF2 increases from normal tissue to tumour tissue. Here, interaction means the amount of mutual information. Moreover, FGF18 is upregulated and FGF2 is downregulated in the tumour compartment. “PI3K-Akt signalling pathway” shows the same changes for FGF2. Furthermore, FGF19, FGFR1 and FGFR4 are the most affected genes of the FGF family. FGF19 shows the second highest interaction after FGFR1 in the normal compartment, but in contrast to FGFR1 the interaction of FGF19 decreases in the tumour compartment. FGFR4 has the third highest interaction in the tumour compartment. Additionally, FGFR1 and FGF19 are upregulated and FGFR4 is downregulated in the tumour compartment. According to the GSEA results of the deconvolved data, all changes of the interaction between the FGF family members and the other genes contribute to an upregulation of the three before mentioned gene sets, namely “regulation of angiogenesis”, “regulation of vasculature development” and “PI3K-Akt signalling pathway”. Comparatively, the GSEA results of the convoluted data show that all changes in the networks (see Supplementary Material, Gene Regulatory Networks) contribute to a downregulation of the three before mentioned gene sets.

Results of the survival analysis overlap on two genes for both convoluted and deconvolved data. Lower chances for survival are associated with low expression of FGF18 and high expression of FGFR3. The results of the survival analysis show that those two members of the FGF family might be potential targets for therapy.

To get reliable results, an ensemble approach was used by simply creating different kinds of data sets. In this way it was possible to look for overlaps in results and also to compare the different approaches. Especially, comparing the results between the convoluted and the deconvolved data. Even though the MDS “landscapes” of the whole convoluted and the deconvolved data look similar, the results of all analysis steps show differences between those data sets. The best example here for are the results of the gene set enrichment analysis. Expression profiles generated by the DeMixT algorithm depend on the initially provided data. As a result, it is important to think about what data (or information) you have available and what the generated expression matrices shall represent. In the case of this study, the expression profiles shall represent gene expression of tumour and normal tissue without infiltration of immune cells or stromal cells respectively.

6 Conclusion

The Results of the developed analysis pipeline shall give researchers different insights into the mechanisms of the FGF family in cancer and help investigate possible new targets for therapy. We addressed the power as well as the limitations of the used computational methods. Hopefully, the results of this study provide new hints for further research and can be proved in future work.

7 Bibliografy

[1] 'Global Cancer Observatory'. http://gco.iarc.fr/ (accessed Mar. 16, 2020).

[2] J. Balogh *et al.*, 'Hepatocellular carcinoma: a review', *J. Hepatocell. Carcinoma*, vol. 3, pp. 41–53, Oct. 2016, doi: 10.2147/JHC.S61146.

[3] D. Hanahan and R. A. Weinberg, 'Hallmarks of cancer: the next generation', *Cell*, vol. 144, no. 5, Art. no. 5, Mar. 2011, doi: 10.1016/j.cell.2011.02.013.

[4] D. M. Ornitz and N. Itoh, 'The Fibroblast Growth Factor signaling pathway', *Wiley Interdiscip. Rev. Dev. Biol.*, vol. 4, no. 3, Art. no. 3, May 2015, doi: 10.1002/wdev.176.

[5] H. Tsunematsu *et al.*, 'Fibroblast growth factor-2 enhances NK sensitivity of hepatocellular carcinoma cells', *Int. J. Cancer*, vol. 130, no. 2, Art. no. 2, 2012, doi: 10.1002/ijc.26003.

[6] C. Gauglhofer *et al.*, 'Up-regulation of the fibroblast growth factor 8 subfamily in human hepatocellular carcinoma for cell survival and neoangiogenesis', *Hepatology*, vol. 53, no. 3, Art. no. 3, 2011, doi: 10.1002/hep.24099.

[7] C. Gauglhofer *et al.*, 'Fibroblast growth factor receptor 4: a putative key driver for the aggressive phenotype of hepatocellular carcinoma', *Carcinogenesis*, vol. 35, no. 10, Art. no. 10, Oct. 2014, doi: 10.1093/carcin/bgu151.

[8] J. Paur *et al.*, 'Fibroblast growth factor receptor 3 isoforms: Novel therapeutic targets for hepatocellular carcinoma?', *Hepatology*, vol. 62, no. 6, Art. no. 6, 2015, doi: 10.1002/hep.28023.

[9] N. Zheng, W. Wei, and Z. Wang, 'Emerging roles of FGF signaling in hepatocellular carcinoma', *Transl. Cancer Res.*, vol. 5, no. 1, Art. no. 1, Feb. 2016.

[10] J. Paur *et al.*, 'Interaction of FGF9 with FGFR3-IIIb/IIIc, a putative driver of growth and aggressive behavior of hepatocellular carcinoma', *Liver Int.*, vol. n/a, no. n/a, doi: 10.1111/liv.14505.

[11] G. I. Murray, 'An overview of laser microdissection technologies', *Acta Histochem.*, vol. 109, no. 3, pp. 171–176, Jun. 2007, doi: 10.1016/j.acthis.2007.02.001.

[12] S. L. Carter *et al.*, 'Absolute quantification of somatic DNA alterations in human cancer', *Nat. Biotechnol.*, vol. 30, no. 5, Art. no. 5, May 2012, doi: 10.1038/nbt.2203.

[13] K. Yoshihara *et al.*, 'Inferring tumour purity and stromal and immune cell admixture from expression data', *Nat. Commun.*, vol. 4, Oct. 2013, doi: 10.1038/ncomms3612.

[14] A. M. Newman *et al.*, 'Robust enumeration of cell subsets from tissue expression profiles', *Nat. Methods*, vol. 12, no. 5, Art. no. 5, May 2015, doi: 10.1038/nmeth.3337.

[15] G. Quon, S. Haider, A. G. Deshwar, A. Cui, P. C. Boutros, and Q. Morris, ‘Computational purification of individual tumor gene expression profiles leads to significant improvements in prognostic prediction’, *Genome Med.*, vol. 5, no. 3, Art. no. 3, Mar. 2013, doi: 10.1186/gm433.

[16] Z. Wang *et al.*, ‘Transcriptome Deconvolution of Heterogeneous Tumor Samples with Immune Infiltration’, *iScience*, vol. 9, pp. 451–460, Nov. 2018, doi: 10.1016/j.isci.2018.10.028.

[17] P. Langfelder and S. Horvath, ‘WGCNA: an R package for weighted correlation network analysis’, *BMC Bioinformatics*, vol. 9, p. 559, Dec. 2008, doi: 10.1186/1471-2105-9-559.

[18] L. Song, P. Langfelder, and S. Horvath, ‘Comparison of co-expression measures: mutual information, correlation, and model based indices’, *BMC Bioinformatics*, vol. 13, no. 1, Art. no. 1, Dec. 2012, doi: 10.1186/1471-2105-13-328.

[19] A. Lachmann, F. M. Giorgi, G. Lopez, and A. Califano, ‘ARACNe-AP: gene network reverse engineering through adaptive partitioning inference of mutual information’, *Bioinformatics*, vol. 32, no. 14, Art. no. 14, Jul. 2016, doi: 10.1093/bioinformatics/btw216.

[20] A. A. Margolin *et al.*, ‘ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context’, *BMC Bioinformatics*, vol. 7, no. Suppl 1, Art. no. Suppl 1, Mar. 2006, doi: 10.1186/1471-2105-7-S1-S7.

[21] R. Steuer, J. Kurths, C. O. Daub, J. Weise, and J. Selbig, ‘The mutual information: Detecting and evaluating dependencies between variables’, *Bioinformatics*, vol. 18, no. suppl_2, Art. no. suppl_2, Oct. 2002, doi: 10.1093/bioinformatics/18.suppl_2.S231.

[22] K.-C. Liang and X. Wang, ‘Gene Regulatory Network Reconstruction Using Conditional Mutual Information’, *EURASIP J. Bioinforma. Syst. Biol.*, vol. 2008, no. 1, Art. no. 1, Jun. 2008, doi: 10.1155/2008/253894.

[23] J. Costa-Silva, D. Domingues, and F. M. Lopes, ‘RNA-Seq differential expression analysis: An extended review and a software tool’, *PLOS ONE*, vol. 12, no. 12, p. e0190152, Dec. 2017, doi: 10.1371/journal.pone.0190152.

[24] M. E. Ritchie *et al.*, ‘limma powers differential expression analyses for RNA-sequencing and microarray studies’, *Nucleic Acids Res.*, vol. 43, no. 7, Art. no. 7, Apr. 2015, doi: 10.1093/nar/gkv007.

[25] C. W. Law, Y. Chen, W. Shi, and G. K. Smyth, ‘voom: Precision weights unlock linear model analysis tools for RNA-seq read counts’, *Genome Biol.*, vol. 15, no. 2, Art. no. 2, Feb. 2014, doi: 10.1186/gb-2014-15-2-r29.

[26] G. K. Smyth, ‘Linear models and empirical bayes methods for assessing differential expression in microarray experiments’, *Stat. Appl. Genet. Mol. Biol.*, vol. 3, p. Article3, 2004, doi: 10.2202/1544-6115.1027.

[27] A. Subramanian *et al.*, ‘Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles’, *Proc. Natl. Acad. Sci. U. S. A.*, vol. 102, no. 43, Art. no. 43, Oct. 2005, doi: 10.1073/pnas.0506580102.

[28] E. L. Kaplan and P. Meier, ‘Nonparametric Estimation from Incomplete Observations’, *J. Am. Stat. Assoc.*, vol. 53, no. 282, Art. no. 282, Jun. 1958, doi: 10.1080/01621459.1958.10501452.

[29] M. K. Goel, P. Khanna, and J. Kishore, ‘Understanding survival analysis: Kaplan-Meier estimate’, *Int. J. Ayurveda Res.*, vol. 1, no. 4, Art. no. 4, 2010, doi: 10.4103/0974-7788.76794.

[30] J. T. RICH, J. G. NEELY, R. C. PANIELLO, C. C. J. VOELKER, B. NUSSENBAUM, and E. W. WANG, ‘A PRACTICAL GUIDE TO UNDERSTANDING KAPLAN-MEIER CURVES’, *Otolaryngol.--Head Neck Surg. Off. J. Am. Acad. Otolaryngol.-Head Neck Surg.*, vol. 143, no. 3, Art. no. 3, Sep. 2010, doi: 10.1016/j.otohns.2010.05.007.

[31] T. Hothorn and B. Lausen, ‘Maximally Selected Rank Statistics in R’, p. 9.

[32] A. Ally *et al.*, ‘Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma’, *Cell*, vol. 169, no. 7, pp. 1327-1341.e23, Jun. 2017, doi: 10.1016/j.cell.2017.05.046.

[33] G. Tischler, *gt1/biobambam*. 2020.

[34] A. Dobin *et al.*, ‘STAR: ultrafast universal RNA-seq aligner’, *Bioinforma. Oxf. Engl.*, vol. 29, no. 1, pp. 15–21, Jan. 2013, doi: 10.1093/bioinformatics/bts635.

[35] S. W. Wingett and S. Andrews, ‘FastQ Screen: A tool for multi-genome mapping and quality control’, *F1000Research*, vol. 7, p. 1338, Sep. 2018, doi: 10.12688/f1000research.15931.2.

[36] *broadinstitute/picard*. Broad Institute, 2020.

[37] A. Colaprico *et al.*, ‘TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data’, *Nucleic Acids Res.*, vol. 44, no. 8, pp. e71–e71, May 2016, doi: 10.1093/nar/gkv1507.

[38] H. Pagès, M. Carlson, S. Falcon, and N. Li, *AnnotationDbi: Manipulation of SQLite-based annotations in Bioconductor*. Bioconductor version: Release (3.10), 2020.

[39] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, ‘edgeR: a Bioconductor package for differential expression analysis of digital gene expression data’, *Bioinformatics*, vol. 26, no. 1, pp. 139–140, Jan. 2010, doi: 10.1093/bioinformatics/btp616.

[40] D. J. McCarthy, Y. Chen, and G. K. Smyth, ‘Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation’, *Nucleic Acids Res.*, vol. 40, no. 10, pp. 4288–4297, May 2012, doi: 10.1093/nar/gks042.

[41] M. D. Robinson and A. Oshlack, ‘A scaling normalization method for differential expression analysis of RNA-seq data’, *Genome Biol.*, vol. 11, no. 3, Art. no. 3, Mar. 2010, doi: 10.1186/gb-2010-11-3-r25.

[42] M.-A. Dillies *et al.*, ‘A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis’, *Brief. Bioinform.*, vol. 14, no. 6, Art. no. 6, Nov. 2013, doi: 10.1093/bib/bbs046.

[43] P. Li, Y. Piao, H. S. Shon, and K. H. Ryu, ‘Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data’, *BMC Bioinformatics*, vol. 16, no. 1, Art. no. 1, Oct. 2015, doi: 10.1186/s12859-015-0778-7.

[44] W. Huber *et al.*, ‘Orchestrating high-throughput genomic analysis with Bioconductor’, *Nat. Methods*, vol. 12, no. 2, Art. no. 2, Feb. 2015, doi: 10.1038/nmeth.3252.

[45] C. W. Law *et al.*, ‘RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR’, *F1000Research*, vol. 5, Dec. 2018, doi: 10.12688/f1000research.9005.3.

[46] G. Yu, L.-G. Wang, Y. Han, and Q.-Y. He, ‘clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters’, *OMICS J. Integr. Biol.*, vol. 16, no. 5, Art. no. 5, Mar. 2012, doi: 10.1089/omi.2011.0118.

[47] A. A. Sergushichev, ‘An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation’, *bioRxiv*, p. 060012, Jun. 2016, doi: 10.1101/060012.

[48] M. Ashburner *et al.*, ‘Gene ontology: tool for the unification of biology. The Gene Ontology Consortium’, *Nat. Genet.*, vol. 25, no. 1, Art. no. 1, May 2000, doi: 10.1038/75556.

[49] M. Kanehisa, S. Goto, M. Furumichi, M. Tanabe, and M. Hirakawa, ‘KEGG for representation and analysis of molecular networks involving diseases and drugs’, *Nucleic Acids Res.*, vol. 38, no. Database issue, Art. no. Database issue, Jan. 2010, doi: 10.1093/nar/gkp896.

[50] O. Tange, *GNU Parallel 20150322 ('Hellwig’)*. Zenodo, 2015.

[51] P. Shannon *et al.*, ‘Cytoscape: a software environment for integrated models of biomolecular interaction networks’, *Genome Res.*, vol. 13, no. 11, pp. 2498–2504, Nov. 2003, doi: 10.1101/gr.1239303.

[52] A. KASSAMBARA, *kassambara/survminer*. 2020.

[53] Y. Benjamini and Y. Hochberg, ‘Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing’, *J. R. Stat. Soc. Ser. B Methodol.*, vol. 57, no. 1, pp. 289–300, 1995.

[54] A. Khatamian, E. O. Paull, A. Califano, and J. Yu, ‘SJARACNe: a scalable software tool for gene network reverse engineering from big data’, *Bioinformatics*, vol. 35, no. 12, Art. no. 12, Jun. 2019, doi: 10.1093/bioinformatics/bty907.

[55] C. Bennette and A. Vickers, ‘Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents’, *BMC Med. Res. Methodol.*, vol. 12, no. 1, Art. no. 1, Feb. 2012, doi: 10.1186/1471-2288-12-21.

[56] ‘The Effects of Cell Cycle Deviation on Cancer Development | Carolina.com’. https://www.carolina.com/teacher-resources/Interactive/the-effects-of-cell-cycle-deviation-on-cancer-development/tr38703.tr (accessed Mar. 16, 2020).

[57] R. Diez del Corral and A. V. Morales, ‘The Multiple Roles of FGF Signaling in the Developing Spinal Cord’, *Front. Cell Dev. Biol.*, vol. 5, 2017, doi: 10.3389/fcell.2017.00058.

[58] J. Kuen, ‘Influence of 3D tumor cell/fibroblast co-culture on monocyte differentiation and tumor progression in pancreatic cancer’, Julius-Maximilians University Würzburg, 2017.

[59] M. Brazier, ‘Microdissection of Alzheimer Brain Tissue for the Determination of Focal Manganese Accumulation’, vol. 124, 2017, pp. 109–118.

8 Supplementary Material

8.1 Differential Gene Expression Tables

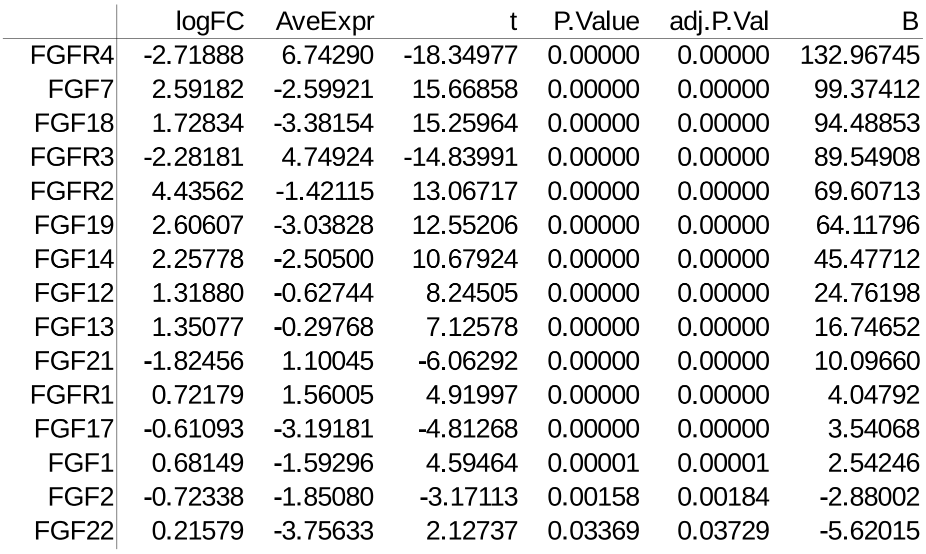
Table 8.1.1: DGE results of all convoluted samples. This table shows the results for the DGE analysis of all convoluted samples for all members of the FGF family. It displays the following information: log fold change, average expression of gene, t-statistic, p-value, adjusted p-value and B-statistic from the empirical Bayes.

	logFC	AveExpr	t	P.Value	adj.P.Val	B
FGFR1	-1.12283	3.07380	-5.59284	0.00000	0.00000	7.70047
FGF17	1.34932	-3.22026	4.46249	0.00001	0.00003	2.36421
FGFR2	-2.17551	4.15811	-4.44327	0.00001	0.00003	2.28277
FGF12	1.03079	0.57533	3.80138	0.00017	0.00037	-0.25183
FGF2	-1.09900	1.04851	-3.56303	0.00041	0.00085	-1.09956
FGF13	0.97119	1.42993	3.46735	0.00058	0.00118	-1.42539
FGF7	-1.17761	-1.35327	-3.29857	0.00106	0.00206	-1.97971
FGF1	-0.70016	-0.59007	-2.84944	0.00460	0.00801	-3.32641
FGF9	-0.79267	-4.83089	-2.82236	0.00500	0.00864	-3.40160
FGFR4	0.53394	7.26522	2.64447	0.00850	0.01407	-3.87820
FGF22	0.69619	-3.89885	2.63582	0.00871	0.01440	-3.90062
FGF14	-0.79677	0.06039	-2.28352	0.02292	0.03494	-4.75290

Table 8.1.2: DGE results of the pairwise comparison. This table shows the results for the DGE analysis of the paired convoluted samples for all members of the FGF family. It displays the following information: log fold change, average expression of gene, t-statistic, p-value, adjusted p-value and B-statistic from the empirical Bayes.

	logFC	AveExpr	t	P.Value	adj.P.Val	B
FGF13	1.67275	1.43291	5.64987	0.00000	0.00000	6.67505
FGFR1	-1.13909	3.49859	-5.57401	0.00000	0.00000	6.35160
FGFR2	-2.47691	4.85919	-4.73963	0.00001	0.00003	2.95696
FGF12	1.10739	0.22731	3.90059	0.00018	0.00051	-0.09727
FGF1	-0.70763	-0.31942	-2.74412	0.00721	0.01428	-3.55670
FGF9	-0.78688	-4.51761	-2.71164	0.00790	0.01548	-3.63941
FGF7	-0.96910	-0.78919	-2.57413	0.01154	0.02171	-3.98019
FGF2	-0.81616	1.57959	-2.55844	0.01204	0.02256	-4.01809

Table 8.1.3: DGE results of the deconvolved data. This table shows the results for the DGE analysis of the deconvolved compartments for all members of the FGF family. It displays the following information: log fold change, average expression of gene, t-statistic, p-value, adjusted p-value and B-statistic from the empirical Bayes.



	logFC	AveExpr	t	P.Value	adj.P.Val	B
FGFR4	-2.71888	6.74290	-18.34977	0.00000	0.00000	132.96745
FGF7	2.59182	-2.59921	15.66858	0.00000	0.00000	99.37412
FGF18	1.72834	-3.38154	15.25964	0.00000	0.00000	94.48853
FGFR3	-2.28181	4.74924	-14.83991	0.00000	0.00000	89.54908
FGFR2	4.43562	-1.42115	13.06717	0.00000	0.00000	69.60713
FGF19	2.60607	-3.03828	12.55206	0.00000	0.00000	64.11796
FGF14	2.25778	-2.50500	10.67924	0.00000	0.00000	45.47712
FGF12	1.31880	-0.62744	8.24505	0.00000	0.00000	24.76198
FGF13	1.35077	-0.29768	7.12578	0.00000	0.00000	16.74652
FGF21	-1.82456	1.10045	-6.06292	0.00000	0.00000	10.09660
FGFR1	0.72179	1.56005	4.91997	0.00000	0.00000	4.04792
FGF17	-0.61093	-3.19181	-4.81268	0.00000	0.00000	3.54068
FGF1	0.68149	-1.59296	4.59464	0.00001	0.00001	2.54246
FGF2	-0.72338	-1.85080	-3.17113	0.00158	0.00184	-2.88002
FGF22	0.21579	-3.75633	2.12737	0.03369	0.03729	-5.62015

8.2 GSEA Results

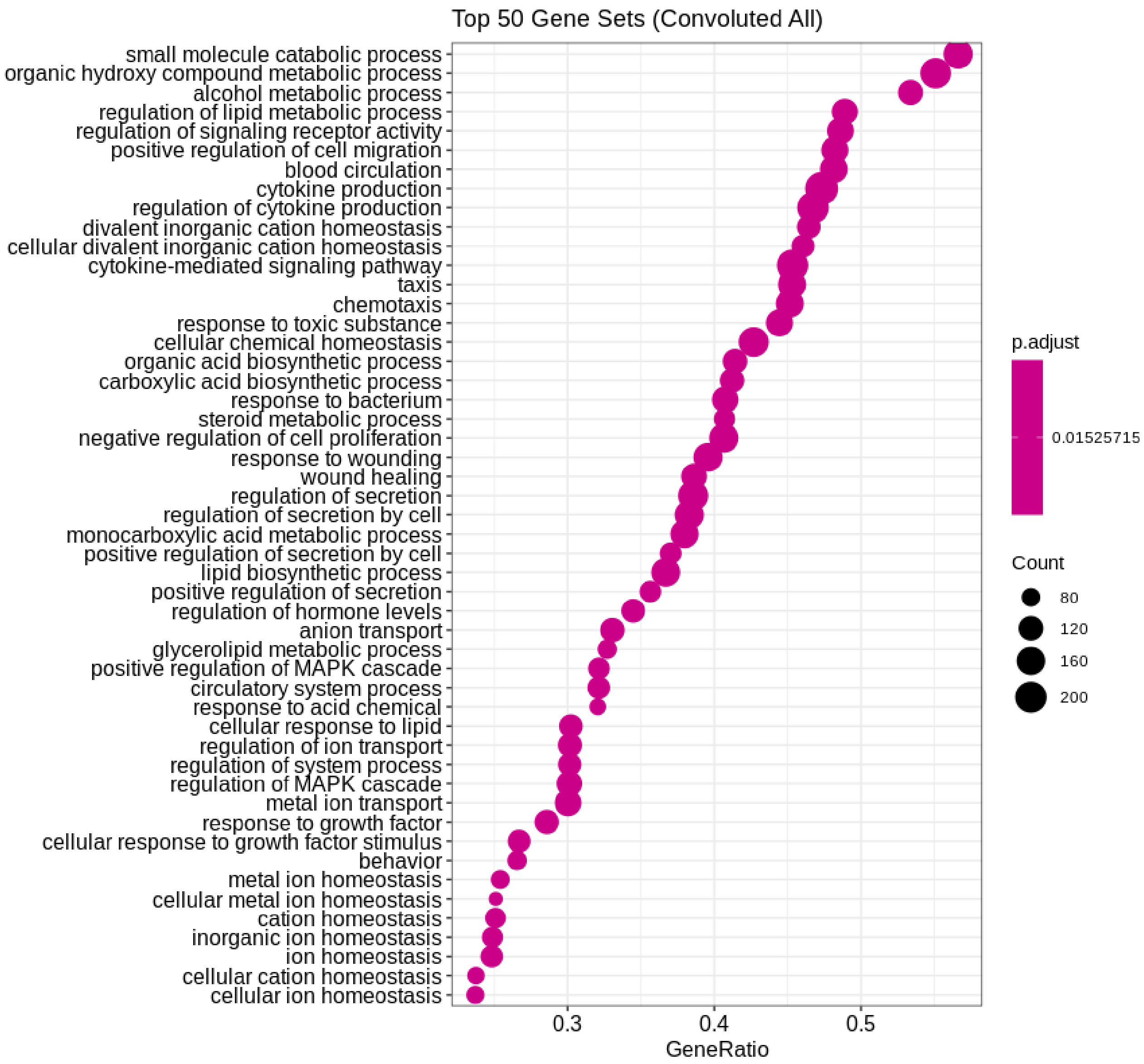


Figure 8.2.1: Dotplot of the top 50 gene sets with all convoluted samples. The x-axis shows the percentage of genes which contribute to the enrichment score. Descriptions for all gene sets are on the y-axis. Dot colour resembles the adjusted p-value and dot size is equal to the total number of included genes.

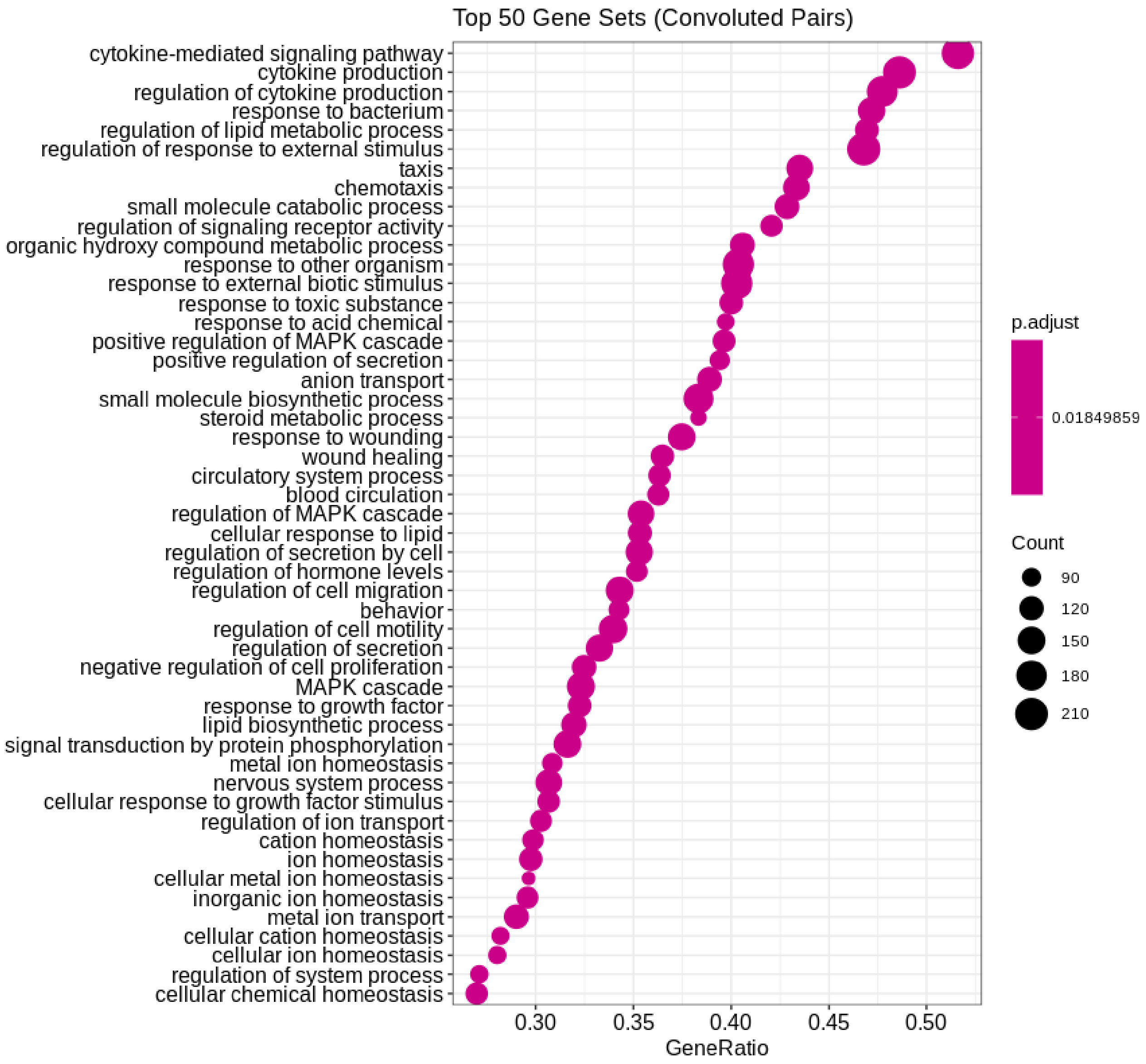


Figure 8.2.2: Dotplot of the top 50 gene sets with the paired convoluted samples. The x-axis shows the percentage of genes which contribute to the enrichment score. Descriptions for all gene sets are on the y-axis. Dot colour resembles the adjusted p-value and dot size is equal to the total number of included genes.

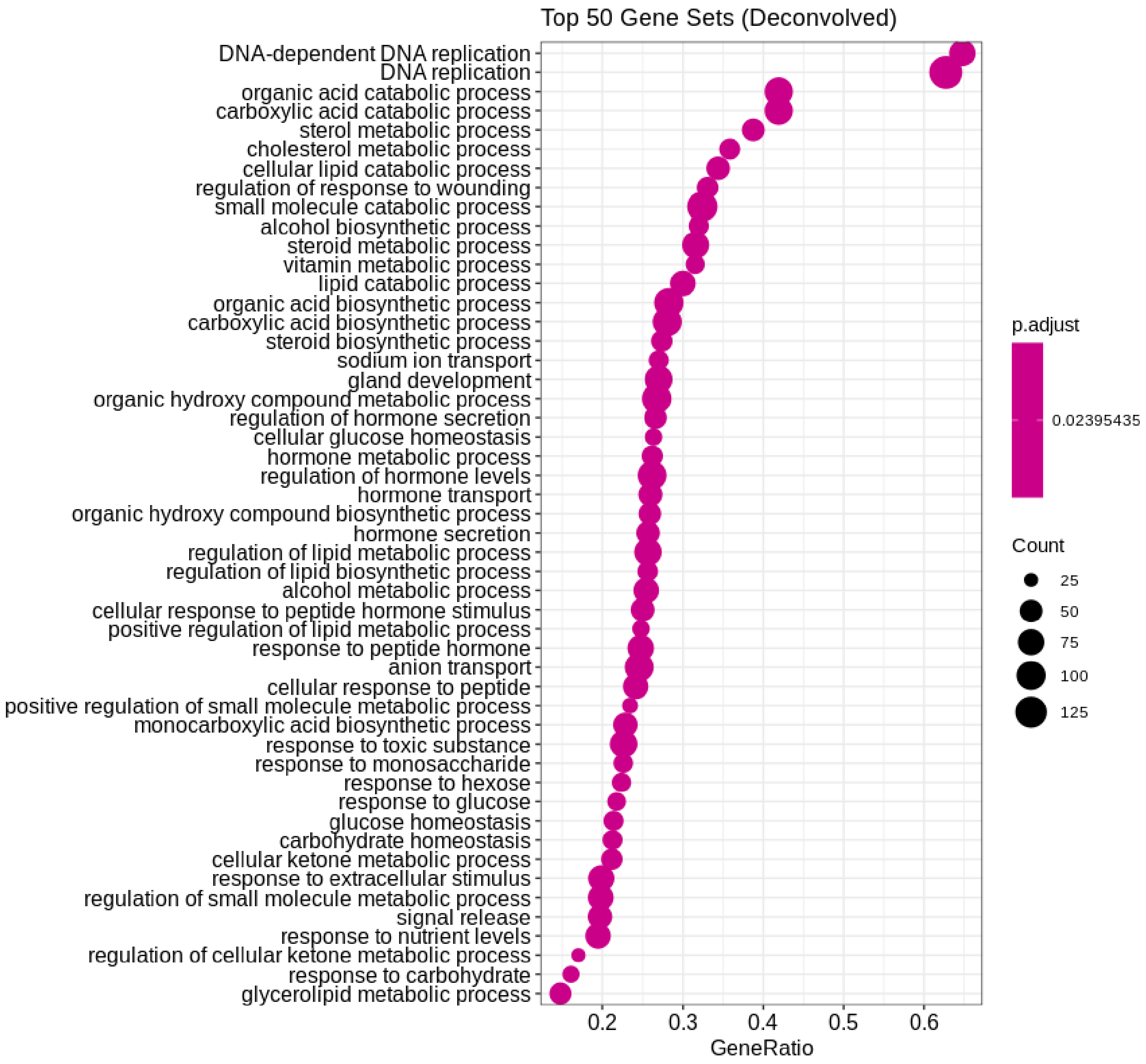


Figure 8.2.3: Dotplot of the top 50 gene sets with the deconvolved compartments. The x-axis shows the percentage of genes which contribute to the enrichment score. Descriptions for all gene sets are on the y-axis. Dot colour resembles the adjusted p-value and dot size is equal to the total number of included genes.

Table 1: All GSEA results of the analysis with the KEGG database. The following information is displayed: description of the gene set, enrichment score, normalised enrichment score, p-value, adjusted p-value, signal strength of the genes, the used analysis approach. Note that no results could be found for the paired convoluted samples.

Description	ES	NES	p-value	p.adjust	Signal (%)	Approach
Proteoglycans in cancer	-0.35143	-1.49885	0.00604	0.02957	30	convoluted all
PI3K-Akt signaling pathway	-0.32071	-1.45100	0.00772	0.03407	29	convoluted all
Rap1 signaling pathway	0.28966	1.53353	0.00476	0.01736	26	deconvolved
Regulation of actin cytoskeleton	0.27715	1.45280	0.00488	0.01736	21	deconvolved
PI3K-Akt signaling pathway	0.28002	1.58228	0.00730	0.02316	24	deconvolved
Ras signaling pathway	0.27093	1.45746	0.01026	0.02873	25	deconvolved
Pathways in cancer	0.23894	1.39923	0.01163	0.03110	25	deconvolved

8.3 Gene Regulatory Networks (Convoluted)

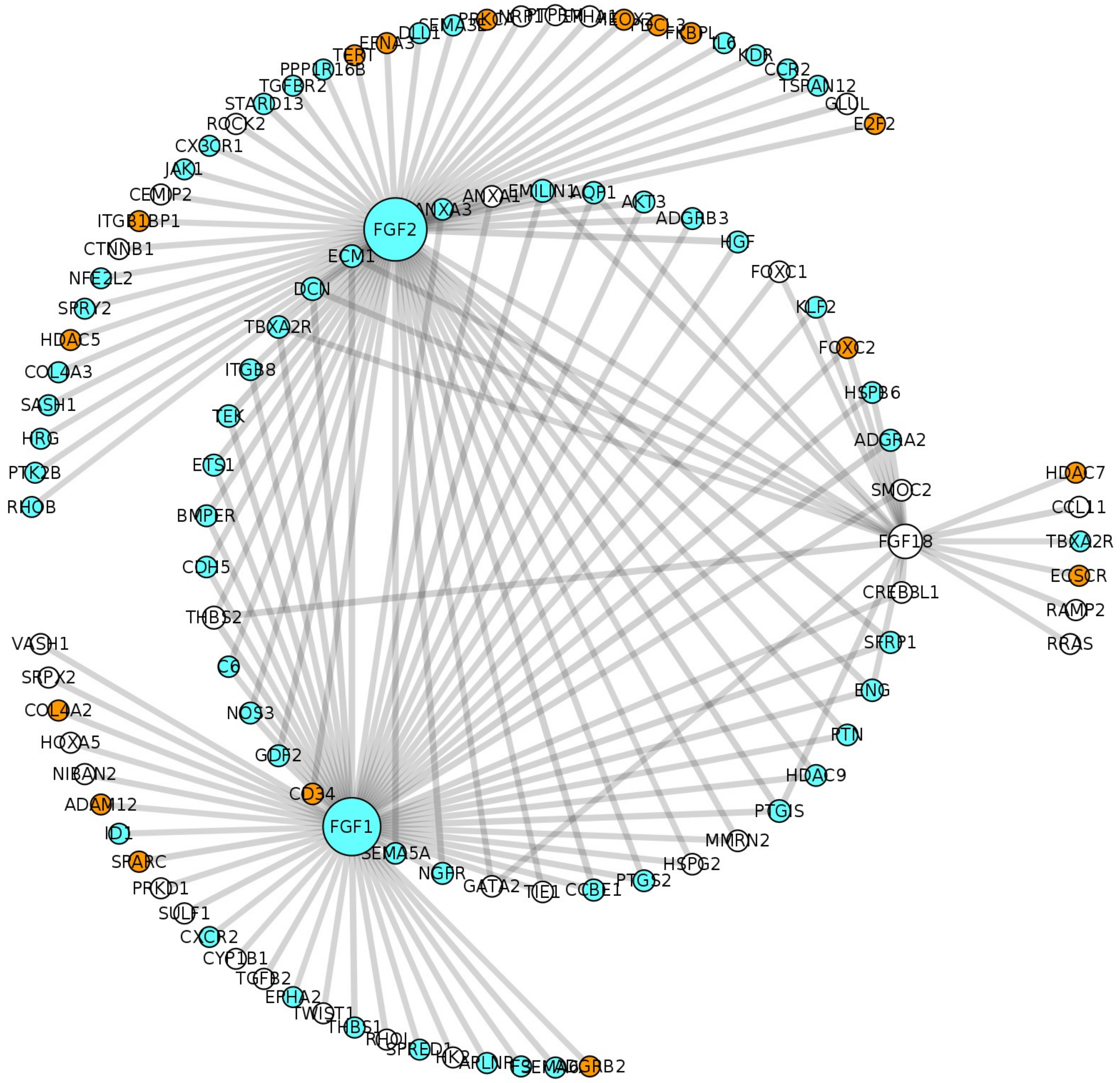


Figure 8.3.1: MI network of "regulation of angiogenesis" from the convoluted data set. The following members of the FGF family are included in this network: FGF18 (not deregulated, smallest row sum), FGF2 (down, greatest row sum), FGF1 (down, second greatest row sum).

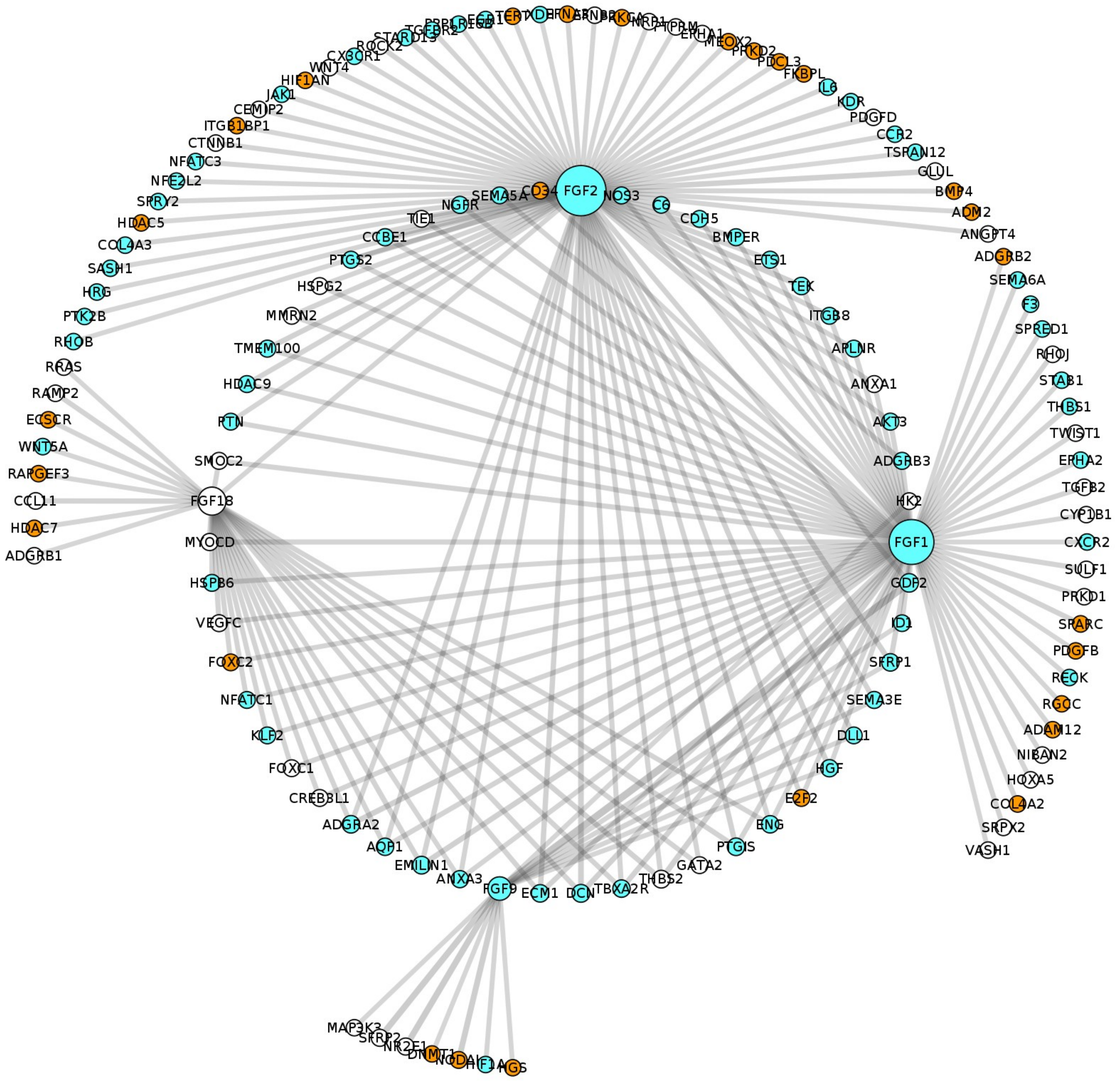


Figure 8.3.2: MI network of “regulation of vasculature development” from the convoluted data set. The following members of the FGF family are included in this network: FGF18 (not deregulated, second smallest row sum), FGF2 (down, greatest row sum), FGF1 (down, second greatest row sum), FGF9 (down, smallest row sum).

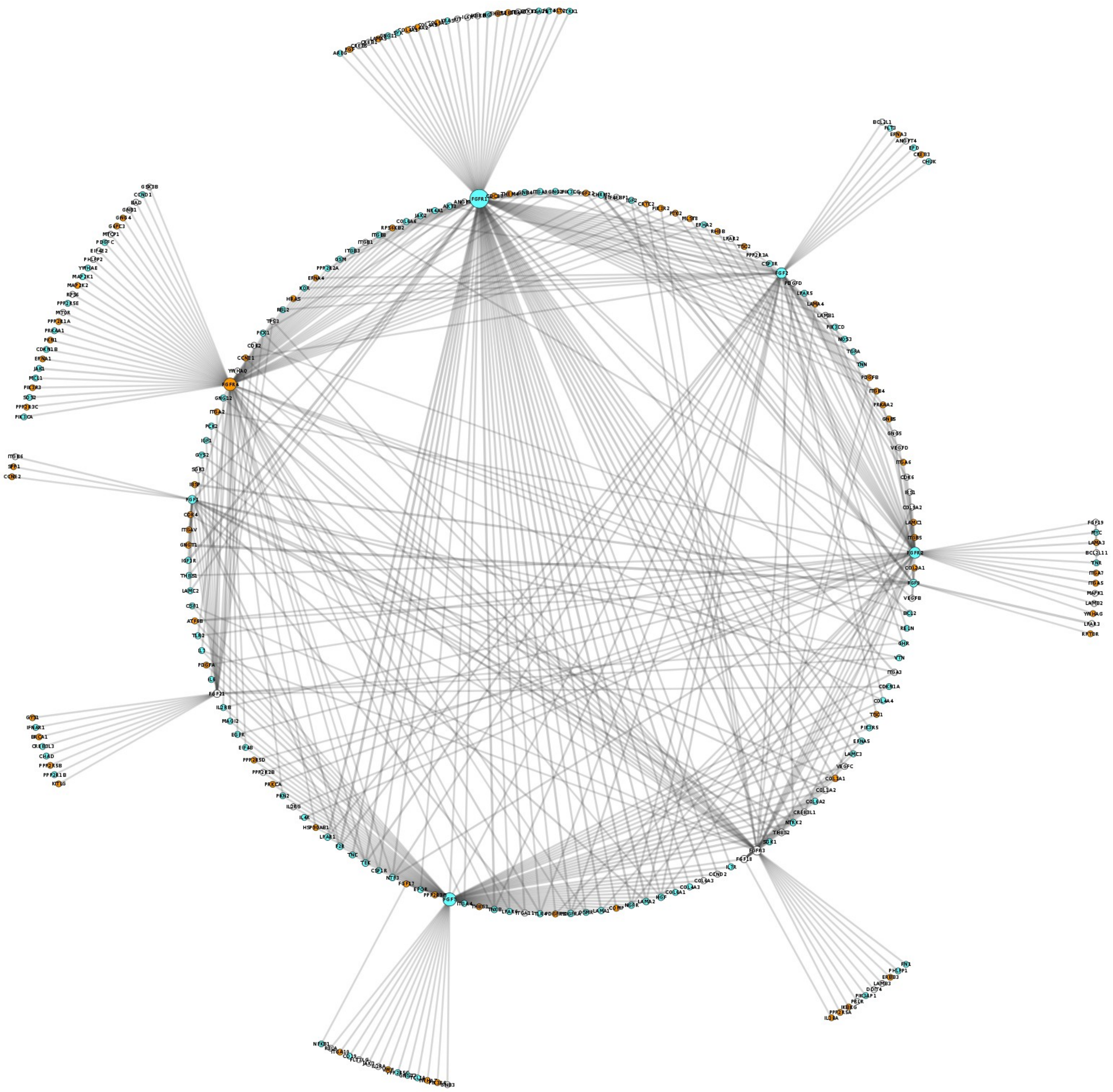


Figure 8.3.3: MI network of "PI3K-Akt signalling pathway" from the convoluted data set. The following members of the FGF family are included in this network: FGF1 (down), FGF2 (down), FGF7 (down, second greatest row sum), FGF9 (down), FGF17 (up), FGF18 (not deregulated), FGF19 (up), FGF21 (not deregulated), FGF22 (up), FGFR1 (down, greatest row sum), FGFR2 (down), FGFR3 (not deregulated), FGFR4 (up, third greatest row sum).

8.4 Survival Analysis Results

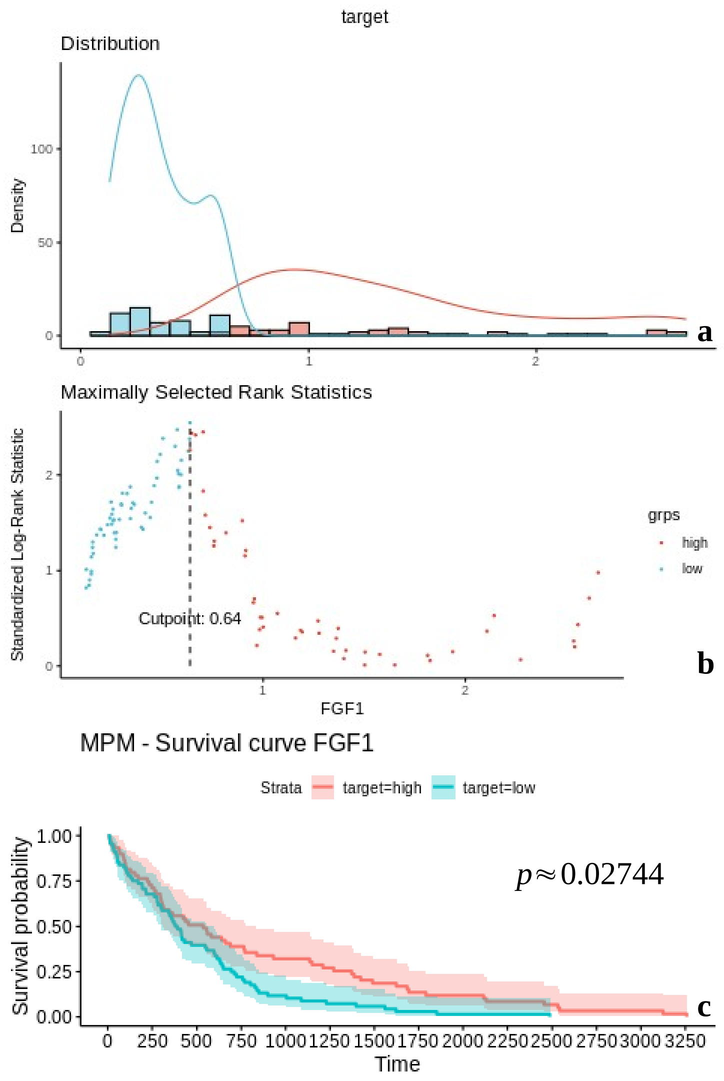


Figure 8.4.1: Max-rank plots and survival plot of FGF1 in the convoluted data set. a) Gene expression distribution divided by the cut-point. b) Maximal rank statistic at 0.64. c) Survival plot → x-axis = time (in days), y-axis = survival probability.

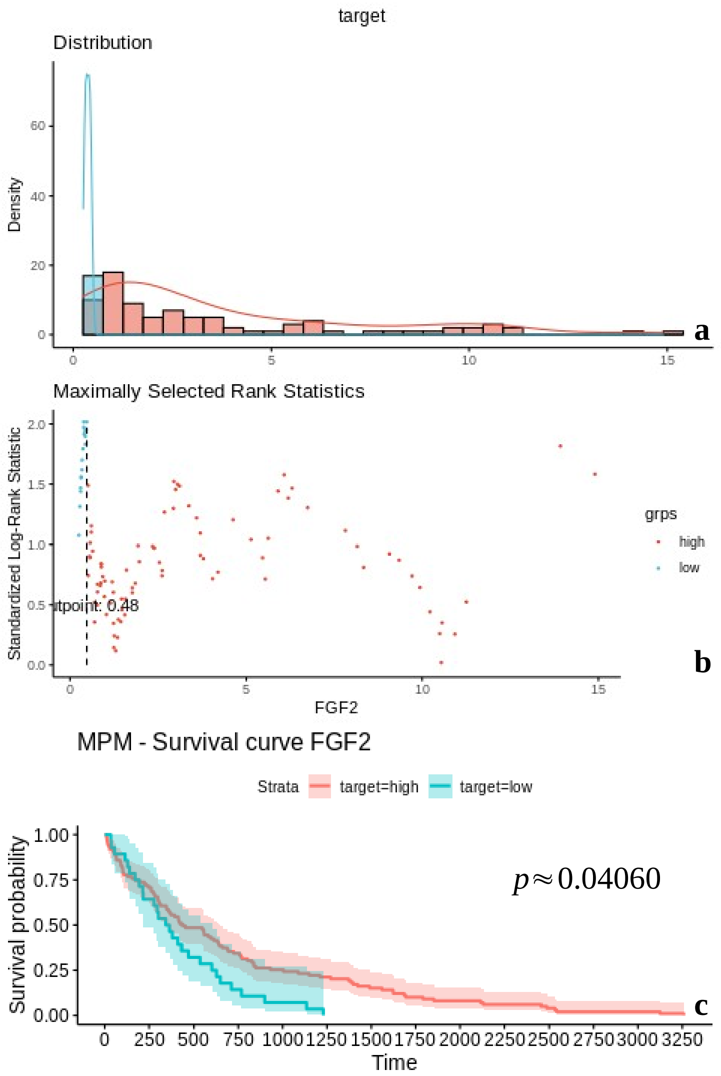


Figure 8.4.2: Max-rank plots and survival plot of FGF2 in the convoluted data set. a) Gene expression distribution divided by the cut-point. b) Maximal rank statistic at 0.48. c) Survival plot → x-axis = time (in days), y-axis = survival probability.

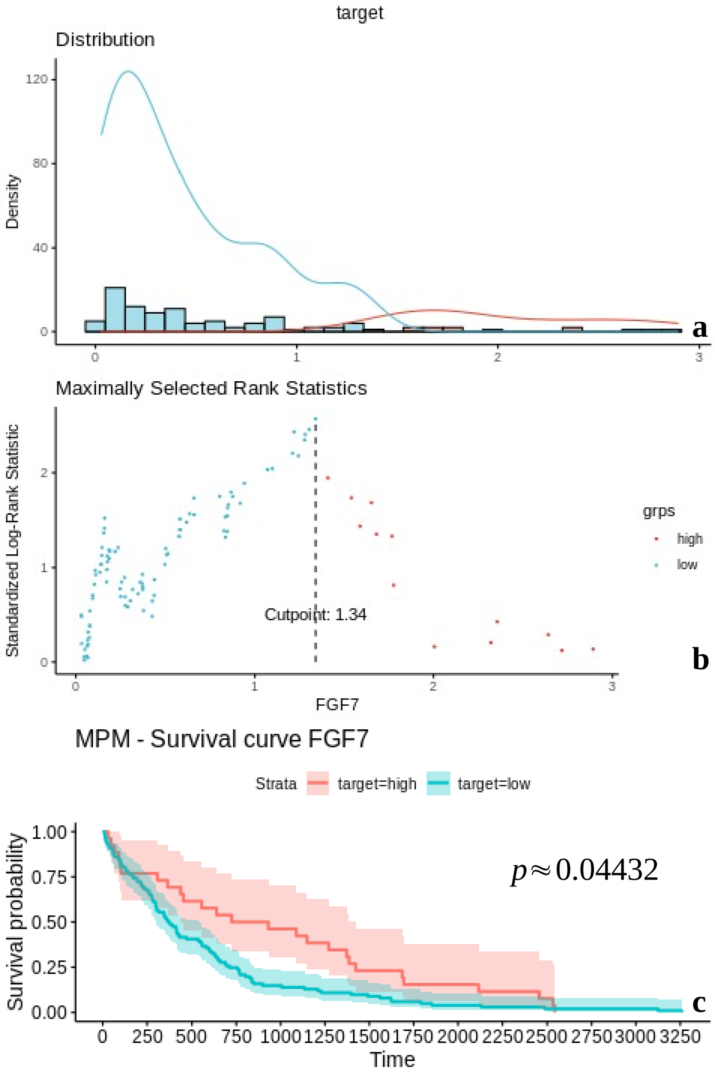


Figure 8.4.3: Max-rank plots and survival plot of FGF7 in the convoluted data set. a) Gene expression distribution divided by the cut-point. b) Maximal rank statistic at 1.34. c) Survival plot → x-axis = time (in days), y-axis = survival probability.

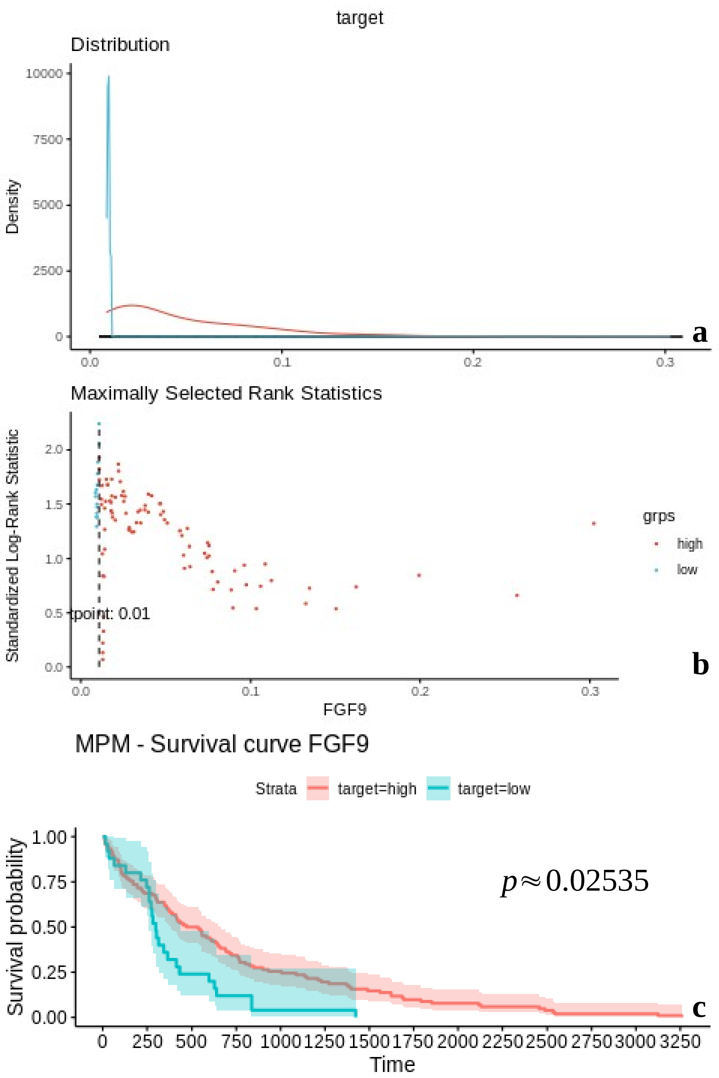


Figure 8.4.4: Max-rank plots and survival plot of FGF9 in the convoluted data set. a) Gene expression distribution divided by the cut-point. b) Maximal rank statistic at 0.01. c) Survival plot → x-axis = time (in days), y-axis = survival probability.

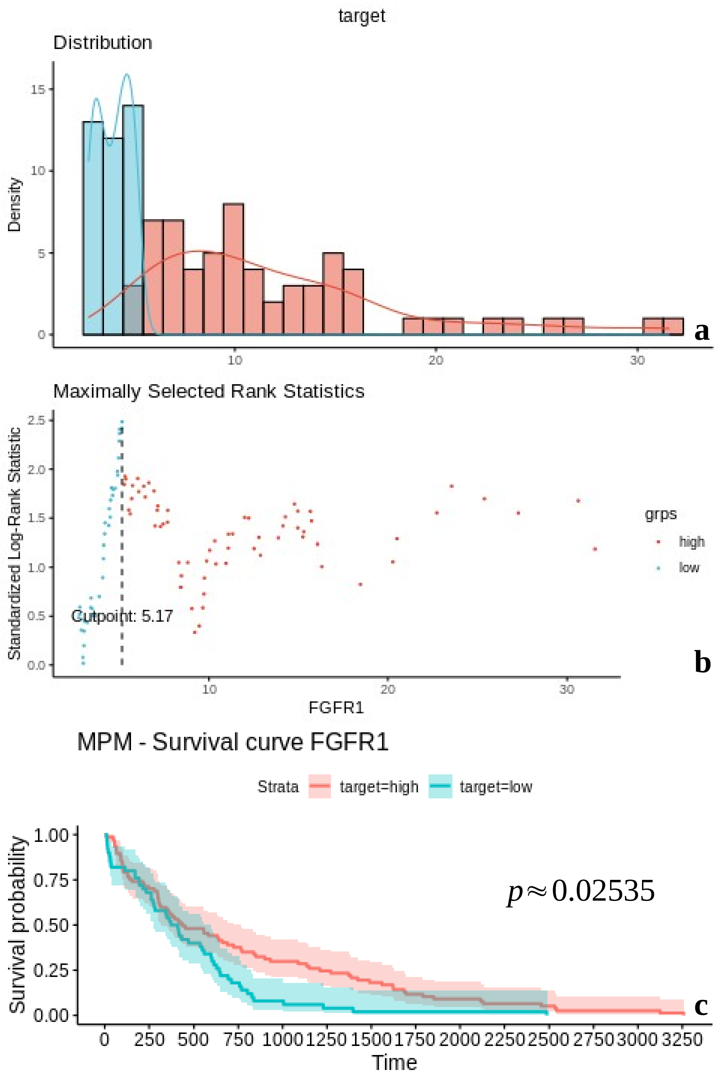


Figure 8.4.5: Max-rank plots and survival plot of FGFR1 in the convoluted data set. a) Gene expression distribution divided by the cut-point. b) Maximal rank statistic at 5.17. c) Survival plot → x-axis = time (in days), y-axis = survival probability.

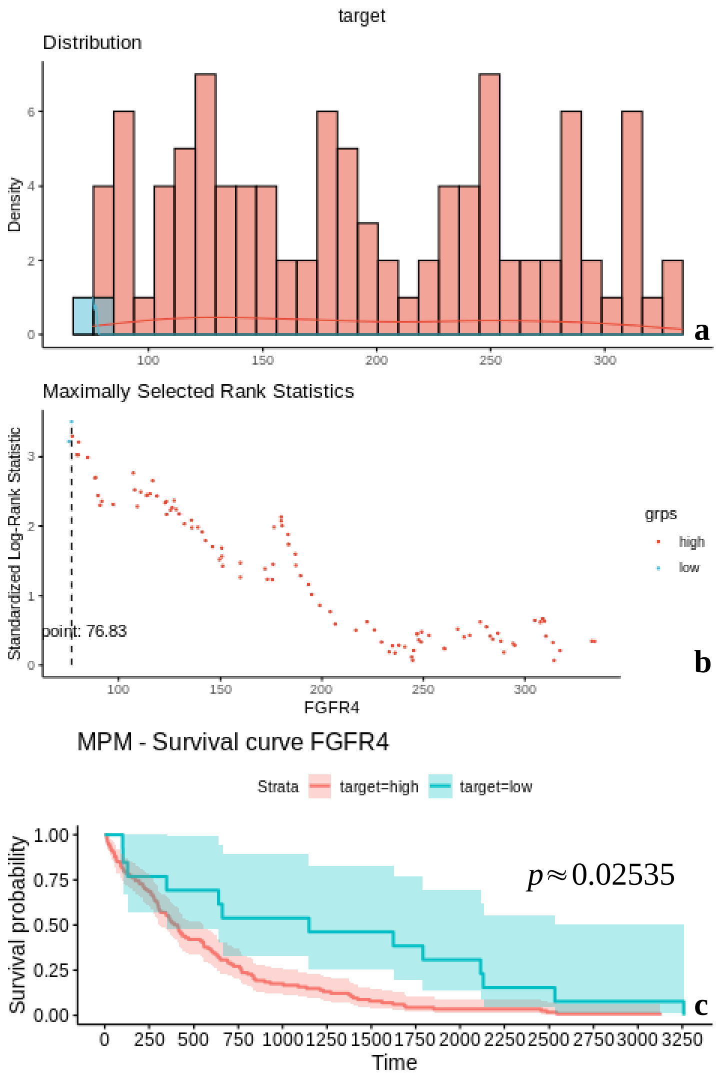


Figure 8.4.6: Max-rank plots and survival plot of FGFR4 in the convoluted data set. a) Gene expression distribution divided by the cut-point. b) Maximal rank statistic at 76.83. c) Survival plot → x-axis = time (in days), y-axis = survival probability.