The Effect of miRNAs to the Regulation of Triple Negative Breast Cancer

Master's Thesis in Technology

University of Turku
Department of Biochemistry
Biotechnology

May 2014

Pauli Tikka

UNIVERSITY OF TURKU
Department of Biochemistry / Faculty of Mathematics and Natural Science

PAULI TIKKA: The effect of miRNAs to the regulation of triple negative breast cancer

Master's Thesis in Technology, 90 p.
Biotechnology
May 2014

Triple negative breast cancer (TNBC) forms a specific subgroup of breast cancer. The roles of microRNAs (miRNAs) on the carcinogenesis of TNBC cells were studied with statistical and mathematical methods using the expression signals of messenger RNAs (mRNAs) and miRNAs. The target genes of miRNAs were found by analysing the most relevant target gene methods and internet databases.

An enrichment analysis, with Fisher exact tests, was conducted to miRNAs that were significantly dysregulating their target genes in TNBC. These tests revealed 21 enriched miRNAs. A hierarchical clustering analysis was then performed to specific set of target gene pairs by using their enriched miRNA related Hamming distances with Ward's method. The hypergeometric enrichment tests for these clusters yielded many biological processes, molecular functions, and pathways indicating that miRNAs had multitude of regulative tasks in TNBC.

Clearly observable regulation between the target genes and miRNAs in TNBC cells was searched by mixed integer programming (MIP) modelling, which is an extended version of linear programming. MIP employs both real and integer variables when it minimizes the error arising from the comparison of the real expression signals of mRNAs to their miRNA related estimates. The gene of Ataxin 1 protein (*ATXN1*) resulted in the most reliable correlation (82%) in MIP using TNBC samples. Ataxin 1 is a substantial component of the notch signalling pathway regulating differentiation. MIP model also showed that the down-regulation of *ATXN1* in TNBC is mostly due to hsa-miR-96-5p, and down-regulation of a gene of Leucine Zipper Protein 1 (*LUZP1*) by an enriched miRNA hsa-miR-29b-3p. MIP models and enrichment tests yielded compatible results compared to literature. In ensuing studies histone modifications, transcription factors, and the distance dependency of the target gene sites of miRNAs should be employed in the model. This could give a more proper description of the potency of some of the well correlated enriched miRNAs to work as biomarkers.

Keywords: miRNA; target genes of miRNAs; triple negative breast cancer; differential expression; enrichment analysis; mixed integer programming

TURUN YLIOPISTO
Biokemian laitos / Matemaattisluonnontieteellinen tiedekunta

PAULI TIKKA: mikroRNA:iden vaikutus kolmoisnegatiivisen rintasyövän säätelyssä

Diplomityö, 90 s.
Biotekniikka
Toukokuu 2014

Kolmoisnegatiivisten rintasyöpien (*engl.* triple negative breast cancer, TNBC) solut muodostavat erityisen rintasyöpäsolujen alaryhmän. mikroRNA:iden (miRNA) rooleja TNBC solujen karsinogeneesissä tutkittiin tilastotieteellisin ja matemaattisin menetelmin käyttäen miRNA- ja lähetti-RNA (*engl.* messenger RNA, mRNA) ekspressiosignaaleja. miRNA:iden kohdegeenit selvitettiin analysoimalla oleellisia kohdegeenianalyysin menetelmiä ja internetin datapankkeja.

Rikastustesti Fisherin tarkalla testillä tehtiin niille TNBC:ssä esiintyville miRNA:ille, jotka merkittävästi säätelevät kohdegeenejään normaalista poikkeavasti. Rikastustestit paljastivat 21 rikastunutta miRNA:ta. Hierarkiallinen ryhmittyminen selvitettiin sitten spesifiselle joukolle kohdegeenipareja käyttäen niiden miRNA:iden välisiä Hammingin etäisyyksiä Wardin menetelmän avulla. Hypergeometriset rikastustestit näille ryhmille tuottivat useita biologisia prosesseja, molekulaarisia tehtäviä ja metaboliateitä, jotka kertoivat, että miRNA:illa on suuri joukko säätelyllisiä tehtäviä TNBC:ssä.

Selkeästi todennettavaa säätelyä TNBC:n miRNA:iden ja geenien välillä etsittiin sekoitetun kokonaislukuoptimoinnin (*engl.* mixed integer programming, MIP) mallinnuksella, joka on kehittynyt versio lineaarisesta mallinnuksesta. MIP pystyy käyttämään sekä reaali-, että kokonaislukuja minimoidessaan virhettä, mikä muodostuu verrattaessa oikeita mRNA-ekspressiosignaaleja miRNA:iden avulla saatuihin arvioihin. Kaikkein merkityksellisin korrelaatio (82 %) MIP:ssa saatiin Ataksiini 1 proteiinin geenille (*ATXN1*) käyttäen TNBC näytteitä. Tämä proteiini on merkittävässä roolissa notch-signalointireittillä, joka säätelee solun erikoistumista. Hsa-miR-96-5p toimi mallin mukaan *ATXN1*:n alassäätelyn tärkeänä osana, kuten myös hsa-miR-29b-3p LUZP1:n geenin (*engl.* gene of Leucine Zipper Protein 1, *LUZP1*) alassäätelyssä. MIP-mallit ja rikastustestit tuottivat yhteneviä tuloksia verrattuna kirjallisuuteen. Jatkotutkimuksissa MIP-mallinnusta voitaisiin kehittää siten, että se ottaisi huomioon myös histonimodifikaatiot, transkriptiotekijät ja miRNA:iden kohdegeenipaikkojen etäisyysriippuvuuden. Tällöin selviäisi varmemmin joidenkin hyvin MIP:ssä korreloivien rikastuneiden miRNA:iden potentiaalinen toimivuus biomarkkereina.

Avainsanat: mikroRNA; miRNA:iden kohdegeenit; kolmoisnegatiivinen rintasyöpä; differentiaalinen ekspressio; rikastusanalyysi; sekoitettu kokonaislukuoptimointi

Making this Master's Thesis has been an exceptional project for me. Not only was it conducted in Germany, Jena, but also within the field of bioinformatics, which has become more and more intriguing for me during this project. The University of Turku provided a good starting point and assistance for initiating my work in Germany. Especially I would like to thank ERASMUS for providing funding, M.Sc. in Tech. Juho Terrijärvi for assisting and supervising my thesis work, and Prof. Dr. Tero Soukka for helping me in various matters during my studies. In addition, the whole department of Biotechnology has been very supportive all these years.

Furthermore, I want to thank Prof. Dr. Rainer König for accepting me to do the practical part of my thesis in Jena, and also Dr. Vijaykumar Muley in Hans Knöll Institute for his friendship and help in coding. I have also received invaluable comments to my English language from my father Kari Tikka and brother Petri Tikka. Finally, I would like to express my gratitude to all of my family and friends that have supported me far beyond measure.

Turku 2.5.2014

Pauli Tikka

TABLE OF CONTENTS

ABBREVIATIONS

α	Limit set to the miRNA amount
β_i	Constants that compose the estimation model backbone
β_o	A primary constant, which balances the estimation arising from miRNA expression signal values
g_s	Gene expression signals in a patient sample number s
k	Maximum number of miRNAs that can be present at any given time according to the target gene list
m_{si}	miRNA expression signals in a patient sample number s that has amount of i many miRNA expression signals
x_j	Binary variable for each miRNA *j*
2PHACTR2	Phosphatase and Actin Regulator
AIB1	Gene of Amplified in Breast 1 protein
AMPK	5' Adenosine Monophosphate-Activated Protein Kinase
NCOA3	Gene of Nuclear Receptor Co-Activator 3 protein
AGO2	Gene for Argonaute 2 protein
AGO2	Argonaute 2 protein
Akt	Serine-Threonine Protein Kinase
APC	Adenomatous Polyposis Coli
ATXN1	Gene of Ataxin 1 protein
ATXN1	Ataxin 1 protein
AU	Approximately Unbiased
BACH2	Gene of BTB and CNC Homology 1, Basic Leucine Zipper Transcription Factor 2
BACH2	BTB and CNC (Carney complex) Homology 1, Basic Leucine Zipper Transcription Factor 2
BCL-X_L	B-Cell Lymphoma Extra-Large
BP	Bootstrap Probability
B-RAF	Protein of V-RAF Murine Sarcoma Viral Oncogene Homolog B
BRCA1	Gene of Breast Cancer 1 protein
BRCA2	Gene of Breast Cancer 2 protein
BTB	A similar protein motif found from Broad-Complex (BR-C), Tramtrack (ttk), and Bric à Brac (bab) proteins

BTG1	B-Cell Translocation Gene 1, Anti-Proliferative
C8orf33	Chromosome 8 Open Reading Frame 33
CASP9	Caspase 9, Apoptosis-Related Cysteine Peptidase
CBX6	Gene of Chromobox Homolog 8 protein
CBX6	Chromobox Homolog 8 protein
ccjp	Cell-Cell Junction Proteins
CELF2	Gene of CUGBP (CUG binding protein), Elav-like Family Member 2 protein
CELF2	CUGBP (CUG binding protein), Elav-like Family Member 2 protein
CPS1	Gene Carbamoyl Phosphate Synthetase 1
CtBP	C-terminal-Binding Protein 1
cDNA	Complementary DNA
C-KIT	Mast/Stem Cell Growth Factor Receptor
ds	Double-stranded
DEG	Differentially Expressed Gene
DEmR	Differentially Expressed miRNA
DGCR8	Digeorge Syndrome Critical Region 8 nuclear protein
Dicer	Gene for Endoribonuclease in the RNAse III Family
Dicer	Endoribonuclease in the RNAse III Family
Drosha	Ribonuclease III dsRNA-Specific Endoribonuclease
E2F	Transcription Factor Activating Adenovirus E2
EGF	Epidermal Growth Factor
EGFR	Epidermal Growth Factor Receptor
EmR	Enriched miRNA
EMT	Epithelial-Mesenchymal Transition
ERK	Extracellular Signal-Regulated Kinase
ER	Estrogen Receptor
Err	Error term that is minimized in linear programming
EVH1	Enabled / Vasodilator-stimulated Phosphoprotein (VASP) Homology 1 protein
Exp5	Exportin-5 karyopherin protein
FASN	Gene of Fatty Acid Synthase
FDR	False Discovery Rate
FOXO1	Gene of Forkhead Box O1 protein

FOXO1	Forkhead box O1 protein
FOXO3a	Gene of Forkhead Box O3A protein
GATA3	Gene of Trans-Acting T-Cell-Specific Transcription Factor 3
GATA3	Trans-Acting T-Cell-Specific Transcription Factor 3
GO	Gene Ontology
HBEGF	Gene of Heparin-Binding EGF-Like Growth Factor
HER2/neu	Gene of Tyrosine-Protein Kinase Erbb-2 Receptor
HER2/neu	Tyrosine-Protein Kinase Erbb-2 Receptor
HOXB	Gene of Homeobox protein
hsa	*Homo sapiens*
HTS	High-Throughput Screening
IGF1R	Oncogene of Insulin-Like Growth Factor 1 Receptor
IGFBP3	Gene of Insulin-Like Growth Factor-Binding Protein 3
Jab1	Gene of C-Jun Activation Domain-Binding Protein 1
Jab1	C-Jun Activation Domain-Binding Protein 1
KEGG	Kyoto Encyclopaedia of Genes and Genomes
LCA	Lysophosphatidic acid
LCOR	Gene of Ligand-Dependent Corepressor protein
LCOR	Ligand-Dependent Corepressor protein
LOH	Loss of heterozygosity
LPA	Lysophosphatidic acid
LUZP1	Leucine Zipper Protein 1
MAPK	Mitogen-Activated Protein Kinase
MDM2	Gene of Mouse Double Minute 2 Homolog
MDM2	Mouse Double Minute 2 Homolog protein
MEK	Mitogen-Activated Protein Kinase
MIP Gap	Gap between the optimum and current solver value in MIP
MIP	Mixed integer programming
miRNA	microRNA
MMP	Matrix MetalloProtease
mRNA	messenger RNA
MTAP	Methylthioadenosine Phosphorylase
NCBI	National Center for Biotechnology Information
NF-κB	Nuclear Factor- κB

NPC-A-5	Gene of Niemann-Pick Disease Type C Line A-5 protein
p16	Cyclin-Dependent Kinase Inhibitor 2A
p70S6K	Ribosomal Protein S6 Kinase, 70kDa, Polypeptide 1
PACT	Protein Kinase, Interferon-Inducible dsRNA Dependent Activator
PARP1	Poly (ADP-Ribose) Polymerase 1
Pasha	DGCR8 equivalent in vertebrates
PCC	Pearson Correlation Coefficient
PDGFC	Gene of Platelet Derived Growth Factor C
PDK1	Pyruvate Dehydrogenase Kinase, Isozyme 1
PI3K	Gene of Phosphatidylinositol-4,5-Bisphosphate 3-Kinase
PI3K	Phosphatidylinositol-4,5-Bisphosphate 3-Kinase
PIP3	Phosphatidylinositol (3,4,5)-Triphosphate
PR	Progesterone Receptor
pre-miRNA	Precursor miRNA
pri-miRNA	Primary miRNA
PP1	Phosphoprotein Phosphatase,
PTCH1	Gene of Protein Patched Homolog 1
PTCH1	Protein Patched Homolog 1
PTEN	Phosphatase and Tensin Homolog
P value	Probability that two distributions come from the same source
Q value	FDR corrected P value
QKI	Gene of Quaking Homolog, Kh Domain RNA Binding protein
RAF	Rapidly Accelerated Fibrosarcoma
Ral-GEF	Rat Sarcoma protein (RAS) Like Guanine Exchange Factor
RAS	Rat Sarcoma virus
RAS	Rat Sarcoma protein
RB	Retinoblastoma protein
RIN	RNA integrity number
RISC	RNA-Induced Silencing Complex
RPKM	Reads per Kilobase per Million Mapped
rRNA	Ribosomal RNA
RSV	Rous Sarcoma Virus
siRNAs	Small interfering RNA
SMAD4	Gene of Mothers Against Decapentaplegic Homolog 4 protein

SNP	Single-Nucleotide Polymorphism
SNP309	Single-Nucleotide Polymorphism at Gene Location 309
snRNAs	Small nuclear RNA
SOCS	Gene of Suppressor of Cytokine Signalling protein
SOCS	Suppressor of Cytokine Signalling protein
SPRED1	Gene of Prouty-Related, EVH1 Domain-Containing Protein 1
SRC	V-Src Avian Sarcoma (Schmidt-Ruppin A-2) Viral Oncogene Homolog
TCGA	The Cancer Genome Atlas
TGF-β	Transforming Growth Factor B
TNBC	Triple Negative Breast Cancer
TP53	Gene of Tumour Protein p53
TP53	Tumour Protein p53
TRBP	Transactivation Region (TAR) RNA Binding Protein
tRNA	Transfer RNA
VEGF	Vascular Endothelial Growth Factor
WHO	World Health Organization
Wnt	Wingless-Type Mouse Mammary Tumour Virus (MMTV) Integration Site Family protein
ZEB1	Gene of Zinc Finger E-Box Binding Homeobox 1
ZEB2	Gene of Zinc Finger E-Box Binding Homeobox 2
ZNF703	Zinc Finger Protein 703
Z-score	Standard score obtained from raw scores by standardizing

1 Introduction

1.1 Breast cancers – diverse cancer landscape

Human cancers caused 8.2 million deaths in the year 2012. The most common cancer type in women is breast cancer. (World Health Organization 2012.) Most types of cancer cells develop during a long period of time and are caused by genetic mutations in their genome. Cancer generally does not produce symptoms at the early stages of its development. The alterations in the genome of cancer cells are usually caused by environmental factors, such as tobacco, radiation, environmental pollutants, stress, imbalanced diet, insufficient physical activity and infections (Anad et al., 2008). Cancer is therefore mainly a non-hereditary disease. A cancerous cell has lost some of its normal capabilities, such as staying in its normal position in the organism, ceasing its dividing phase, attaching to other cells, and most of all, dying when required. This indicates that cancer cells have learned how to bypass or modify important cellular defensive mechanisms, such as apoptosis and DNA repair. Consequently, there should be multiple stages before normal cells form cancerous ones. Eventually these stages will lead to abnormal growth of cells, which is called neoplasia, and finally to a subset of neoplasm, which is a lumped tumour. This subset of neoplasm is typically a malignant one, in other words cancer.

Cancer is not a locally expressed disease; subsequently it can spread, i.e. metastasise, to other parts of the body. The behaviour of cancer is characterized by its tendency for proliferation and malignancy. This tendency is enabled by resisting cell death that lead to replicative immortality, evading growth suppressors, sustaining proliferative signalling, activating invasion and the metastasis, inducing angiogenesis (i.e formation of blood vessels), reprogramming of energy metabolism, and evasion of immune destruction (Hanahan and Weinberg 2000; Warburg 1956; Hanahan and Weinberg 2011). The classification of cancers is derived from how similar the cancer cells appear compared to the normal cells, for example sarcoma resembling connective tissue and carcinoma looking like epithelial cells, e.g. colon or breast.

Breast cancer occurs in certain breast tissues. Usually they are located at the lobules that provide milk for the milk ducts, which is then lobular carcinoma,

or just in the inner lining of the milk ducts, i.e. ductal carcinoma. Most breast cancers are detected by mammography (X-ray) or by a physical examination as part of a screening programme. (Hunt et al*.,* 2008.) Late first pregnancy, obesity, family history and genetics, as well as age are well-known risk factors for breast cancer (Salehi et al*.,* 2008). The incidence of breast cancer (relative to age) is almost three-fold higher in developed than in developing parts of the world. This disease is treated mainly with surgery, radiation or medication, e.g. immune-therapy, chemotherapy, and hormonal therapy (Florescu 2011, Pegram et al., 2004, Locker 1998). Breast cancer can be classified into six subgroups according to the molecular subtype compared to a normal breast cell (Perou et al., 2000; Herschkowitz et al., 2007; Hu et al*.,* 2011). These subgroups can be identified by tumour grade and the receptor status. Tumour grade is a tumour differentiation status; a poorly-differentiated cell receives a high tumour grade. The receptor status denotes if the receptor is present (+) or not (-) in the cancer cell. The receptors in this case are Estrogen (ER), Progesterone (PR) and Tyrosine-Protein Kinase Erbb-2 Receptor (HER2/neu; Table 1). Evidently, there are other classification systems of breast cancer, such as using histopathology, TNM classification, and DNA classification.

In this diverse landscape of breast cancers, the triple negative breast cancer (TNBC) is a special type of breast cancer that does not express ER, PR, and HER2/neu, and does not require these receptors for its growth. It is a very heterogeneous group of breast cancers (Perou 2011; Lehmann et al., 2011). A recent study by Lehmann et al., (2011) depicted six different subtypes of TNBC: two basal-like, an immunomodulatory, a mesenchymal, a mesenchymal stem–like, and a luminal androgen receptor subtype. TNBCs cover 15% of all types of breast cancer (Cleator et al., 2007). Medication of TNBC remains a difficult task, while it lacks above mentioned proteins, such as ER, which would be expressed in ER+ breast cancer cells, and could be targeted by a hormone therapy drug, such as tamoxifen (Jordan 2006). Yet TNBC can be treated with chemotherapy, possibly with Poly (ADP-Ribose) Polymerase 1 (PARP1) inhibitors (Virág and Szabó 2002) or by targeting the highly expressed Epidermal Growth Factor Receptor (EGFR; Cleator et al., 2007).

Table 1. The classification of breast cancers according to the receptor status and tumour grade.

Subgroups of breast cancer	Grade	ER[1]	PR	HER2/neu	Other
Luminal A	low	+			
Luminal B	high	+			
Luminal ER-/AR+		-			Androgen+
ERBB2/HER2+				++	
Normal breast-like					
Basal-like, i.e. TNBC		-	-	-	
Claudin-low		(-)	(-)	(-)	low ccjp[2]

1) The receptor statuses are indicated by +/-. 2) ccjp = Cell-Cell Junction Proteins.

1.2 Cancer research

The public cancer research tries to improve the methods for diagnostics, treatments, and prevention of cancer. But most of all, it tries to recognize the causes of cancer. These causes may be mutations or viruses, which lead to genetic changes in the cells. If these genetic changes create cancer, then it is imperative to find out the consequences of those genetic changes the biology of the cell. The properties of cancer cells are marked by these changes. The cancer cells can even employ novel genetic events, which lead to further progression of the cancer. (Weinberg 2013.) There are several ways to conduct research for recognizing the causes of cancer. Two prominent ways are laboratory and clinical experiments. In addition, systems biology and especially biomedical research often focus to cancer research. For example, genetics, environmental factors and also matters of nutrition play a part in the cancer research (Willett 2002). Typical cancer research protocol for finding potential drug targets begins from preliminary laboratory experiments for the samples of clinical observations. An essential part of this research is the searching of how the mechanisms of carcinogenesis operate. Genetic and also epigenetic changes are then readily explored. For the analysis of tumour initiation, cultures of mammalian and bacterial cells are incorporated. These cells do not reveal the complexity of tumour formation in human, so an animal model such as mouse, may be utilized. The mice are mostly tested in the context of altering the tasks of genes. If the preliminary research is successful to explain some of the effectors of cancer, the clinical trials will duly follow. The idea is to assess the safety and efficiency of the new therapeutic drug in real case scenarios, in order to validate the findings for a large scale production (Gibbs 2000).

The other methods of treatments for cancers in general, similarly as in the case of breast cancer, could range from radiation therapy, surgery, immunotherapy, hormone therapy to a combination of the previously mentioned treatments. Fairly recently, more and more therapies have emerged from biotechnology research, such as immunotherapy with antibodies, e.g. Trastuzumab, and gene therapy (Dranoff 2011; Slamon et al., 2001; Nichols 1988). Cancer can be screened with various methods, such as biomarkers from blood or urine samples or by medical imaging or with more invasive types of methods, e.g. tissue specimen.

1.3 The effect of pathways and genes in cancers

There are several cellular mechanisms and pathways in breast cancer that have gone astray in a way that produces the typical behaviours of cancer, such as excessive proliferation and apoptosis avoiding tactics. Oncogenes facilitate the cell reproduction and growth, whereas tumour suppressor genes inhibit cell division and survival. New oncogenes or normal oncogenes are excessively expressed, i.e. up-regulated, and tumour suppressor genes are under-expressed, or in other words down-regulated, or mutated in cancer. These mutations in the chromosome usually appear to be extensive. Nonetheless, they may also be small, such as point or somatic mutations. If these mutations appear in the gene's promoter region they can affect its expression. The product integrity and operation characteristics may change if the mutations are in the coding sequence of the gene. Thus, cancer pathways usually either activate or inactivate genes. For example, the activated genes can be Rat Sarcoma virus (*RAS*), a gene of Phosphatidylinositol-4,5-Bisphosphate 3-Kinase (*PI3K*), and a gene of Cyclin-Dependent Kinase Inhibitor 2A (*p16*), which are all oncogenes. The ones that are inhibited could be a gene of Adenomatous Polyposis Coli (*APC*) and a gene of Retinoblastoma Protein (*RB*), which are tumour suppressors, and also Transcription Factor Activating Adenovirus E2 (E2F).

Several key pathways are fundamentally abbreviated in cancer, such as cellular apoptosis (including caspase cascade), Serine-Threonine Protein Kinase (Akt) signalling, Epidermal Growth Factor (EGF) pathway, Tumour Protein P53 (TP53)

pathway, a pathway involved in the inhibition of Matrix MetalloProteases (MMPs), and a pathway involved in the interactions of Vascular Endothelial Growth Factor (VEGF) family of ligands and receptors (Weinberg 2013). Recently, especially the increased activities into and out of the mitochondrial glycine biosynthetic pathways were identified as indicators for increasing proliferation rates in cancer cells (Jain et al., 2012). In additionIn, the higher expression of this particular pathway was linked with higher mortality in breast cancer patients.

The protein of V-Src Avian Sarcoma (Schmidt-Ruppin A-2) Viral Oncogene Homolog (SRC) is a non-receptor tyrosine kinase that phosphorylates specific tyrosine residues in other proteins. SRC is associated with cancer proliferation by promoting other signals, such as *RAS* genes. There are three major downstream signalling cascades that arise from activated *RAS*: Rapidly Accelerated Fibrosarcoma (RAF) kinase, Phosphatidylinositol-4,5-Bisphosphate 3-Kinase (PI3K) and Rat Sarcoma protein (RAS) Like Guanine Exchange Factor (Ral-GEF). RAF signalling cascade is usually named Mitogen-Activated Protein Kinase (MAPK) pathway that can be found in both normal and neoplastic mammalian cells. PI3K pathway depends on kinases that phosphorylate phosphatidylinositol to Phosphatidylinositol (3,4,5)-Triphosphate (PIP3). PIP3 levels are normally kept low by mostly Phosphatase and Tensin Homolog (PTEN). PI3K/Akt and RAS/Mitogen-Activated Protein Kinase/Extracellular Signal-Regulated Kinase (RAS/MEK/ERK) pathways protect the cells from premature apoptosis. From time to time, the genes along these pathways are mutated in a way that turns them permanently on. As a result, the cell cannot destroy itself in an appropriate time. Likewise, RAS can also bind to Ral-GEF that will guide Ral for transforming GDP to GTP so that some downstream targets are activated. This causes a wide impact on the cell, changing the cytoskeleton and cellular motility. (Weinberg 2013.)

Wingless-Type Mouse Mammary Tumour Virus (MMTV) Integration Site Family protein factors (Wnt factors, notably the ligand binding protein family factors) control a pathway, which provides the cells the means to stay in a comparatively undifferentiated state, which is essential for some of the cancer cells.

Nuclear Factor-κB (NF-κB) signalling system relies on these dual-address proteins to migrate to nucleus and work as transcription factors, which activate many genes, i.e. at least 150 genes, including the ones in cell proliferation and the anti-apoptotic system.

The Transforming Growth Factor β (TGF-β) signalling pathway starts from TGF-β binding to its receptor, which will then attach to another receptor. This causes phosphorylation of cytoplasmic SMAD proteins. These proteins can now make complexes that can activate many genes. The role of this pathway in cancer is prominent. For example in carcinomas, such as breast cancers, it has a major effect on the pathogenesis, in other words the disease creation mechanism, of the cancer cell. To illustrate this in the early stages of cancer, TGF-β arrest cell proliferation, and in the later stages it helps the tumour cell to increase its invasiveness. (Weinberg 2013.)

In the same way, some of the genes are so called driver genes for a specific type of cancer, such as a gene of Zinc Finger Protein 703 (*ZNF703*), which has been observed in luminal A type of breast cancer (Curtis et al., 2012). Naturally also some viruses, such as DNA virus, which is independent of host and retrovirus, ordinarily a Rous Sarcoma Virus (*RSV*), also play a part in disrupting the normal function of the genomes. So to express this matter in rudimentary terms, many types of genes, DNA/RNA sequences and functions, and naturally also proteins, can hence regulate the proliferation of cancerous cells (Figure 1).

Besides the above mentioned pathways and mechanisms, cancer is also driven by epigenetic alterations, which are modifications that do not involve changes in the nucleotide sequence. These epigenetic changes can be DNA methylations, changes in some specific enzymes (e.g. DNase), histone modifications, such as acetylation and phosphorylation of histone residues (Baylin and Ohm 2006), and anti-inflammatory circuits mediated by microRNAs (miRNAs; Iliopoulos 2014). These epigenetic changes usually affect in such a moderate state so that the most relevant biological explainer of the gene expression is still a transcription factor (Cheng et al., 2012). Nevertheless, transcription factors are one part of a bigger control mechanism of the production of proteins, where the epi-

genetic changes play a part. This mechanism may also affect the genes of miRNAs (Chen et al., 2012).

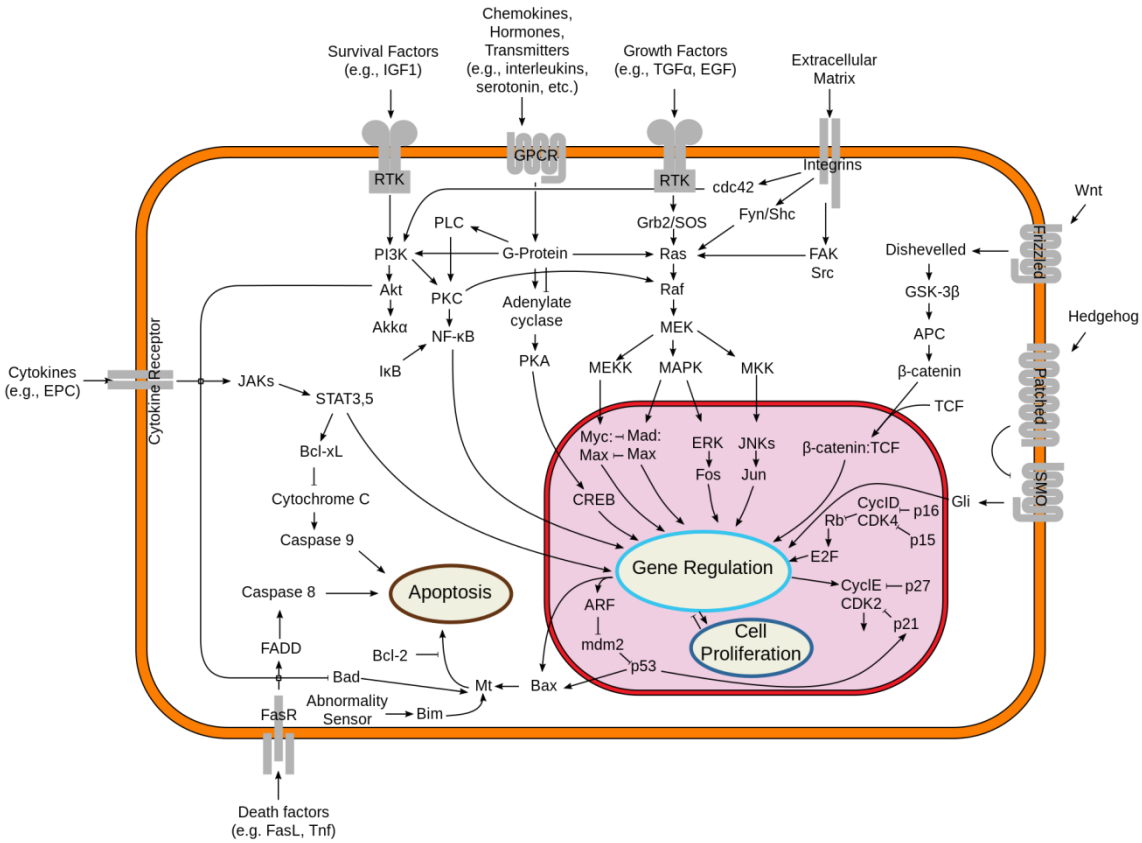

Figure 1. Signalling pathways and their proteins that are prominent for most of the cancer cells including breast cancer. (Picture taken from Douglas and Weinberg 2000).

Albeit the regulation of cancer is partially well known, some matters still need elucidation. These matters could be the operations of pathways as a whole due to differences in superficially similar proteins. An example of this is RAF and its close cousin, a serine/theorine-protein kinase, which is a protein of V-RAF Murine Sarcoma Viral Oncogene Homolog B (B-RAF). It is equally relevant to understand the roles of phosphotyrosine phosphatases in the regulation and the signal processing operation by individual signal-transducing proteins. Moreover, the signal transducing ability of proteins is affected by post-translational modifications, such as phosphorylation. However, there are many other more complex modifications as well as intracellular concentrations and localizations. A proper elucidation of the connections between these matters is a complex task. Not to mention that none of the above mentioned intracellular pathways work in isolation. Signalling cascades are likely to function in a closely

tuned, dynamic equilibrium, where negative and positive regulators work simultaneously and all the time counterbalancing one another. (Weinberg 2013.)

1.4 Molecular mechanisms and genes in breast cancer

The role of some of the important genes and molecular mechanisms in the regulation of breast cancer cells is quite well depicted in literature. For example, the oncogene of Insulin-Like Growth Factor 1 Receptor (*IGF1R*) that is up-regulated (Curtis et al., 2012), while tumour suppressor genes *BRCA1* and *BRCA2* (Breast Cancer Genes 1 and 2) are down-regulated by inherited mutations, could potentially lead to breast cancer. There are also other genes down-regulated by mutations, such as a gene of Methylthioadenosine Phosphorylase (*MTAP*) or a gene of Mothers Against Decapentaplegic Homolog 4 (*SMAD4*; Curtis et al., 2012). Obviously, some of the mutations can be point or somatic mutations, such as for the gene of Trans-Acting T-Cell-Specific Transcription Factor 3 (*GATA3*), so that its product regulates the luminal epithelial cell differentiation in the mammary glands (Hosein et al., 2006), and for the gene of Tumour Protein P53 (*TP53*) in breast cancer (The Cancer Genome Atlas 2012). When analysing the breast cancer cell sample, Trans-Acting T-Cell-Specific Transcription Factor 3 (GATA3) has emerged as a robust and unbiased evaluator of ER status, tumour differentiation, and clinical outcome. GATA3 spontaneously steers the expression of the gene of ER and other genes associated with epithelial differentiation. The loss of GATA3 leads to loss of differentiation and poor prognosis due to cancer cell metastasis and invasion. (Hosein et al., 2008.) Besides this gene, the high activity of V-Src Avian Sarcoma (Schmidt-Ruppin A-2) Viral Oncogene Homolog (*SRC*) is also prominent in breast cancer (Summy and Gallick 2006), leading to above mentioned signalling cascades.

The expression of the anti-apoptotic B-Cell Lymphoma Extra-Large (BCL-X_L) in human mammary epithelial cells results in suppression of apoptosis that is normally done during anoikis. This anoikis is an excavation of lumina, or channels, in globular aggregates of epithelial cells, acini, when some cells are not properly attached to the ducts. Equally important – Single-Nucleotide Polymorphisms (SNPs) can occur at crucial places in the gene as the promoter.

If SNPs happen to the promoter of a gene of Mouse Double Minute 2 Homolog protein (*MDM2*), the proliferation of cancer could be elevated, since Mouse Double Minute 2 Homolog protein (MDM2) is a component of the TP53 pathway. This polymorphism is called SNP309, which is common in homozygous configurations. It increases the risk of developing breast cancer. (Weinberg 2013; Economopoulos et al., 2010.)

The typical regulative features of TNBC consist of elevated proliferative rate and abundant signalling through MAPK and Akt pathways, and overexpression of EGFR and Mast/Stem Cell Growth Factor Receptor (C-KIT). By the same token, these features could be high DNA damage rate, dense mutations of *TP53*, phenotypical similarity to *BRCA1*-associated cancers, and with some likelihood also some defective DNA repair pathways (Cleator et al., 2007). It has been shown by Ossovskaya et al., (2011) that TNBC has considerable differences (Figure 2) in the regulation of its genes compared to other breast cancer types. For example, inferring from the Figures 1 and 2; Caspase 9, Apoptosis-Related Cysteine Peptidase (CASP9) does not lead directly to apoptosis as it did with in the general case (Figure 1).

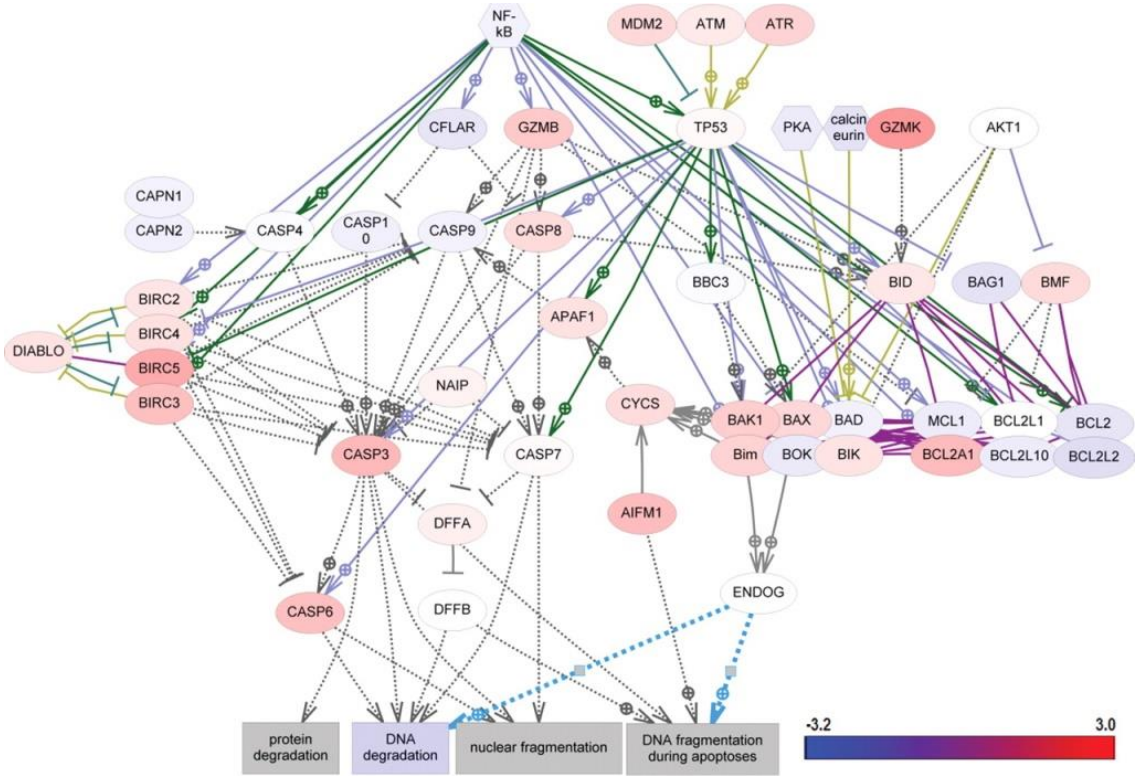

Figure 2. Gene expression changes in the apoptotic pathways of TNBC. The colours red and blue in the scale denote up-regulation and down-regulation, respectively. (Picture taken from Ossovskaya et al., 2011.)

1.5 Functions of miRNAs

1.5.1 Basic tasks of miRNAs

miRNAs are around 22 nucleotides long non-coding RNAs that have fairly been recently discovered (Lee et al., 1993). Their importance to the gene regulation is extensive. The human genome has been estimated to encode over one thousand miRNAs. This information can be inferred from miRBase, which is an internet database for experimentally validated and predicted miRNAs and their sequences (miRBase 2014). In April 2014, it had 2,578 unique miRNAs, although all of them have not been experimentally validated. miRNAs may regulate even 30% of human genes (Lewis et al., 2005). These molecules can assist in the transcriptional and post-transcriptional regulation of gene expression (Chen and Rajewsky 2007). miRNAs target protein-coding genes and reduce their expression in a specific way. They affect translation efficiency of messenger RNAs (mRNA) or adjust the levels of mRNA in the cytoplasm, or both. miRNAs attach to the complementary sequences of mRNAs so that they are silenced. The mammalian miRNAs can recognize as little as seven nucleotides of the sequence of their target mRNAs, or to represent this in more

common terms, the target genes of miRNAs. This thesis considers mRNAs when describing target genes of miRNAs, on account of this convention has also been loosely used in the literature by John et al., (2004), unless otherwise stated. So, particularly, these miRNA-target gene consortiums cannot be translated into proteins by ribosomes. Usually these blocks of target genes and miRNAs are degraded by cellular mechanisms (Bartel 2009). To illuminate further miRNAs' complex role, for example in breast cancer, miRNAs rather act as modulators of miRNA-to-target gene interactions than on-off molecular switches (Dvinge et al., 2013). According to several studies, miRNAs have various tasks in negative regulation and in positive regulation of genes (i.e. transcript degradation, translational suppression, and transcriptional and translational activation; Poy et al., 2004; Lim et al., 2005; Place et al., 2007, respectively). Subsequently, this involvement in the regulation of genes implycates that miRNAs are part of many biological processes (Sun and Lai 2013; Carthew 2006). This combinatorial regulation is therefore a typical phenomenon of miRNA regulation (Krek et al., 2005).

1.5.2 miRNA-related cellular mechanisms

Finding the genes that produce miRNAs is not a straightforward matter, while converting cellular RNA to complementary DNA (cDNA) can also detect other noncoding RNAs than miRNAs. These other noncoding RNAs include ribosomal RNAs (rRNAs), transfer RNAs (tRNAs), small nuclear RNAs (snRNAs), mRNAs and most of all small interfering RNAs (siRNAs; Ambros et al., 2003). SiRNAs are double-stranded (ds) exogeneous RNAs, whereas miRNAs are endogeneous single-stranded RNAs. The majority of these non-miRNA sequences can fortunately be checked out by matching all sequences to the information in the annotated databases.

For proper regulation and modelling analyses, it is maybe more valuable to understand how miRNAs are formed rather than elucidating how the right miRNAs and their genes are searched *per se*. The cellular process of producing miRNAs is highly conserved (Heimberg et al., 2008). The RNase III enzyme complex, Ribonuclease III dsRNA-specific Endoribonuclease-Digeorge Syndrome Critical Region 8 nuclear protein (Drosha-DGCR8 or "Pasha" in inverte-

brates), processes the transcribed long miRNA precursors, i.e. primary miRNAs (pri-miRNAs), to form around 70-base short precursor miRNAs (pre-miRNAs; Denli et al., 2004). Karyopherin protein, Exportin-5 (Exp5), exports the pre-miRNAs from the nucleus to an Endoribonuclease in the RNAse III Family (Dicer) that has two partner proteins: Transactivation Region (TAR) RNA Binding Protein (TRBP) and Protein Kinase, Interferon-Inducible dsRNA Dependent Activator (PACT). The pre-miRNA is processed in this complex (Dicer-TRBP-PACT) to mature miRNA, which can be incorporated into an argonaute-containing RNA-Induced Silencing Complex (RISC; Du and Zamore 2005). Mature miRNA can also go to nucleus. Typically, miRNA-RISC complex attaches to mRNA. Translational repression and mRNA cleavage or degradation is followed by the binding of the silencing complex (Petersen et al., 2006; Bartel 2004). Translational repression and mRNA degradation will occur if miRNAs pair imperfectly to mRNAs of protein-coding genes. The mRNA cleavage requires almost perfect pairing of miRNAs to mRNAs (Figure 3).

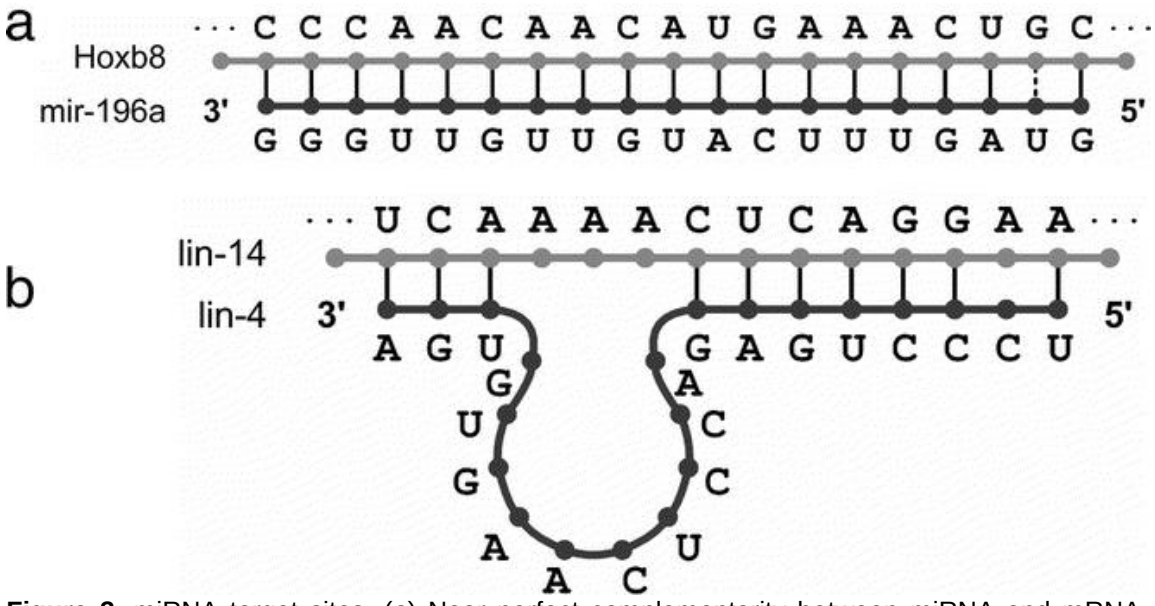

Figure 3. miRNA target sites. (*a*) Near perfect complementarity between miRNA and mRNA gives mRNA cleavage, whereas (*b*) less complementarity gives translational suppression and mRNA degradation. (Picture taken from Saetrom et al., 2007).

There are many mechanism that explain the mRNA cleavage and degradation (Jing et al., 2005), but usually miRNAs direct endonuclease cleavage, which is also called as slicing, to mRNA. There should be perfect base-pairing at least in the seed region (6-8 nucleotides), and a specific Argonaute 2 protein (AGO2 for

mammalians) present for inducing the slicing activity in degradation (Tan et al., 2009). The prompt cleavage starts with removal of poly (A) tail of mRNA by deadenylation so that mRNA can be degraded by exosome or other specific enzymes (Dcp1 and Dcp2) that facilitate 5'-to-3' degradation by exoribo-nuclease (XRN1; Orban and Izaurralde 2005; Valencia-Sanchez et al., 2006). Processing bodies, i.e. dynamic entities of agglomerate of enzymes, are also required for miRNA-mediated gene silencing by providing a functional site for mRNA turnover (Sheth and Parker 2003; MacFarlane and Murphy 2010; Figure 4). The mRNA degradations and cleavages may also lead to a cascade of more cleavages, while some of the cleaved mRNAs might be miRNAs of their own (Tan et al., 2009).

1.5.3 Classification of miRNAs

The names of miRNAs have their peculiarities, which depict their functions. Pre-miRNAs that lead to similar mature miRNAs are denoted with an additional dash-number suffix. For example, the pre-miRNAs hsa-mir-129-1 and hsa-mir-129-2 form an identical mature miRNA, hsa-miR-129. The distinction is that they are located at different places in the genome. Whereas, miRNAs with almost similar sequences are annotated with an additional lower case letter, such as miR-365a that would be closely related to miR-365b. The origin of the species may in similar fashion be indicated with a three-letter prefix. For instance, hsa-miR-3609 is a human (*Homo sapiens*) miRNA. Likewise, two mature miRNAs can be formed from opposite ends of the same pre-miRNA. If they did form so, this would give the names of mature miRNAs an additional suffix (-3p or -5p; Griffiths-Jones et al., 2006.)

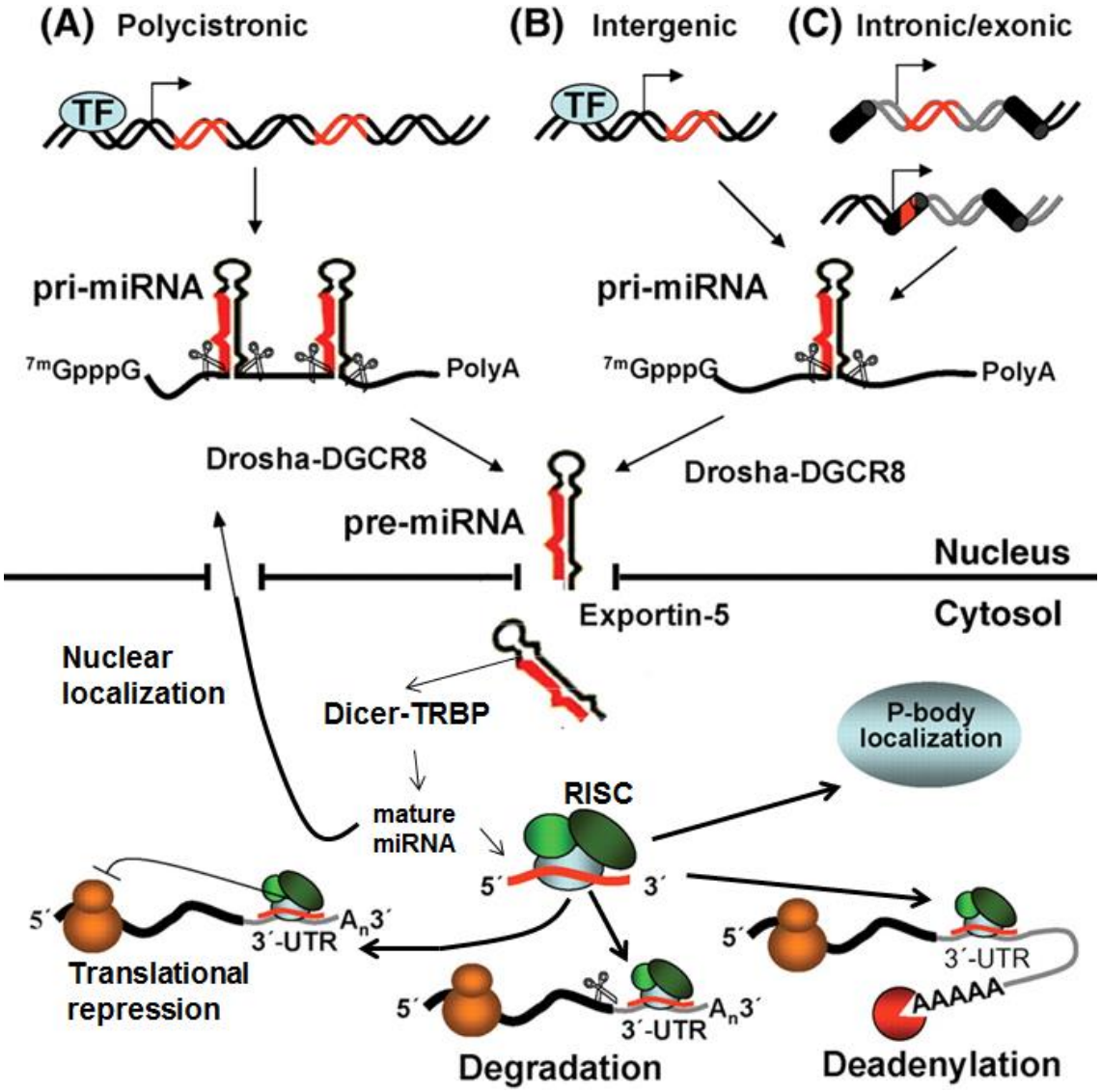

Figure 4. Genomic organization, biogenesis and functions of miRNAs (red). (*A*) polycistronic primary transcripts cleaved into multiple miRNAs; (*B*) intergenic regions transcribed as independent transcriptional units; (*C*) intronic sequences (in grey) of protein-coding or -non-coding transcription units or exonic sequences (black cylinders) of non-coding genes. mRNA targets are black curves. Scissors denote the cleavage on pri-miRNA or mRNA. (Picture modified from Fazi and Nevi 2008.)

1.6 Target gene analysis

For every target gene, there can be many different miRNAs, and similarly a particular miRNA may have multiple different mRNA targets (Rajewsky 2006). Thus, miRNA expression is commonly analysed with mRNA expression data to check how miRNAs can be related to mRNA levels (Gupta et al., 2011). This challenges the quantification of miRNA regulation using High-Throughput Screening (HTS), which is also usually not an easy task on account of the errors due to a larger variance than from mRNA quantification, and these issues arise from specialities of HTS (Zeng and Mortazavi 2012; Maier et al.,

2007; Macarron 2006). Proper databases, which predict target genes based on their base sequence, should be used to pair miRNA and mRNA data (Griffiths-Jones et al., 2008). This is usually done after quantification of interesting miRNAs, i.e. the highly expressed ones. The estimates of the amount of different target genes can differ quite considerably. The method of predicting these targets affects also the estimates (Thomson et al., 2011). Regardless that the task is not straightforward, few additional notions have been suggested for the construction of analysis tools that would elegantly find target genes from multiple sources (Artmann et al., 2012; Muniategui et al., 2012; Dong et al., 2013; Tuschen 2013).

1.7 Tumour suppressor genes, oncogenes and miRNAs

The loss of genetic information is essential for the development of many cancers and requires the inactivation of tumour suppressor genes and activation of oncogenes. Usually both copies of a tumour suppressor gene must be lost in order to affect the phenotype of the cell. The inactivation of the function of tumour suppressor gene can happen through epigenetic silencing of genes via promoter methylation, miRNAs (Mavrakis et al., 2011), or genetic mutation. Besides this, there are other inactivation mechanisms, such as the ones related to the loss of heterozygosity (LOH) at the locus of tumour suppressor gene. These mechanisms comprise of inappropriate chromosomal segregation or gene conversion coming from a switch in template strand during DNA replication. Equally important similar mechanisms are the loss of a chromosomal region that harbours the gene and the mitotic recombination. LOH occurs more often than the epigenetic silencing procedures or genetic mutations. The loss of tumour suppressor genes happens more often during the development of a tumour than the activation of proto-oncogenes into oncogens. (Weinberg 2013.) In this context miRNAs have been noticed to deactivate oncogene expression, such as the *RAS* gene. Not to mention the genes that are located in chromosomal regions and encode miRNAs, which occur in cancer. These genes could also be interpreted as tumour suppressors or oncogens. Oncogenic miRNAs may down-regulate tumour suppressors or other genes, such as the ones involved in cell differentiation. Likewise, tumour suppressor miRNAs may down-regulate different oncogenes. (Shenouda and

Alahari 2009). Furthermore, these oncogenic miRNAs tend to cleave target genes more frequently than the tumour suppressor miRNAs (Wang et al., 2010). The loss of more than a dozen of these genes of tumour suppressor miRNAs has been associated with the formation of many cancers (Esquela-Kerscher and Slack 2006; Weinberg 2013). This means that miRNAs could be used as potential therapeutic drugs (Chan et al., 2011, Martin et al., 2014) by adding these miRNAs or even synthetic ones of them, similarly as with siRNAs (Elbashir et al., 2001). Nonetheless, it is not clear whether the changes in miRNA expression are a cause or effect of the cancer for many miRNA species, which could, as gene regulators, be employed as potential diagnostic and prognostic candidates, and new therapeutic targets (Fu et al., 2011).

1.8 miRNAs in breast cancer

According to study by Volinia et al., (2006), a number of miRNAs are deregulated in human breast cancer. A further research discovered a number of miRNAs that were differentially expressed in breast tumour biopsies and that miRNA expression correlated with HER2 and ER status (Mattie et al., 2006). This differential expression and previous studies by Esquela-Kerscher and Slack (2006) demonstrated that tumour suppressor and oncogenic miRNAs are directly involved in oncogenesis of breast cancer (Table 2). It has also been proven that there are 133 miRNAs associated with normal breast cells and breast cancer cells (Blenkiron et al., 2007). Some of these 133 miRNAs, e.g. miR-150, miR-30a-3p, and miR-199a, are especially associated with clinicopathological factors, such as grade, stage, and Nottingham Prognostic Index. This study also showed that miRNAs are expressed in a coordinated fashion. An example of this is the miRNAs located on a specific chromosome location (C7q22.1; hsa-miR-25, hsa-miR-93 and hsa-miR-106b) that are all highly expressed in high grade tumours.

Table 2. The role of miRNAs in breast cancer. (Table modified from Fu et al., 2011).

Tumour suppressor miRNAs	Targets	Functional pathways
miR-206	ESR1	ER signalling
miR-17-5p	AlB1, CCND1, E2F1	Proliferation
miR-125a,b	HER2, HER3	Anchorage-dependent growth
miR-200c	BMl1, ZEB1, ZEB2	TGF-β signalling
let-7	H-RAS, HMGA2, LIN28, PEBP1	Proliferation, differentiation
miR-34a	CCND1, CDK6, E2F3, MYC	DNA damage, proliferation
miR-31	FZD3, ITGA5, M-RIP, MMP16, RDX, RHOA	Metastasis
miR-335	SOX4, PTPRN2, MERTK, TNC	Metastasis
miR-27b	CYP1B1	Modulation of the response of tumour to anti-cancer drugs
miR-126	IRS-1	Cell cycle progression from G1/G0 to S
miR-101	EZH2	Oncogenic and metastatic activity
miR-145	miR-145 in TP53-mediated repression of c-Myc	Suppresses cell invasion and metastasis
miR-146a/b	NF-κB	Negatively regulate nuclear factor-κB, and impaired invasion and migration capacity
miR-205	ErbB3 and VEGF-A expression	Inhibits tumour cell growth and cell invasion

Oncogenic miRNAs	Targets	Functional pathways
miR-21	BCL-2, TPM1, PDCD4, PTEN, MASPIN	Apoptosis
miR-155	RHOA	TGF-β signalling
miR-10b	HOXD10	Metastasis
miR-373 / 520c	CD44	Metastasis
miR-27a	Zinc finger ZBTB10, Myt-1	Cell cycle progression G2-M checkpoint regulation
miR221 / 222	p27Kip1	Tamoxifen resistance

Nevertheless, Blenkiron et al., (2007) also suggested that there are not that many cases, where miRNA expressions are generally linked to the host gene expression, whilst there were high correlated exceptions, such as miR-205/ a gene of Niemann-Pick Disease Type C Line A-5 protein (*NPC-A-5)* and miR-10a/genes of certain Homeobox proteins (*HOXB2-HOXB6*). The reason for this could be that the deregulation of genes required for miRNA biogenesis in different tumour subtypes may lead to global changes in miRNA expression. This comprises the down-regulation of the gene of the Endoribonuclease in the RNAse III Family (*Dicer*) and the gene of Argonaute 2 protein (*AGO2*). In

addition, pre-miRNA processing by Dicer might be deviated due to a specific inhibitor activity (Obernosterer et al., 2006).

miRNAs are not just a research tool to find out the regulation of gene expression; they can likewise be used as biomarkers for cancer classification, response to therapy, and prognosis (Fu et al., 2011; Martin et al., 2014; Lowery et al., 2009). These miRNA biomarker identifiers are analysed from samples to find out different breast cancer statuses. For example, ER status is analysed with miR-26a/b, miR-30 family, miR-29b, miR-155, miR-342, miR-206, and miR-191 (Blenkiron et al., 2007; Sempere et al., 2007; Mattie et al., 2006; Iorio et al., 2005). PR status can in like manner be analysed with let-7c, miR-29b, miR-26a, miR-30 family, and miR-520g (Iorio et al., 2005; Lowery et al., 2009). Finally, HER2/neu status is investigated with miR-520d, miR-181c, miR-302c, miR-376b, and miR-30e (Blenkiron et al., 2007; Lowery et al., 2009). miRNAs may thus elucidate changes in important regulatory networks that could drive oncogenecis (Chan et al., 2011). Some of the above mentioned miRNAs are in Figure 5.

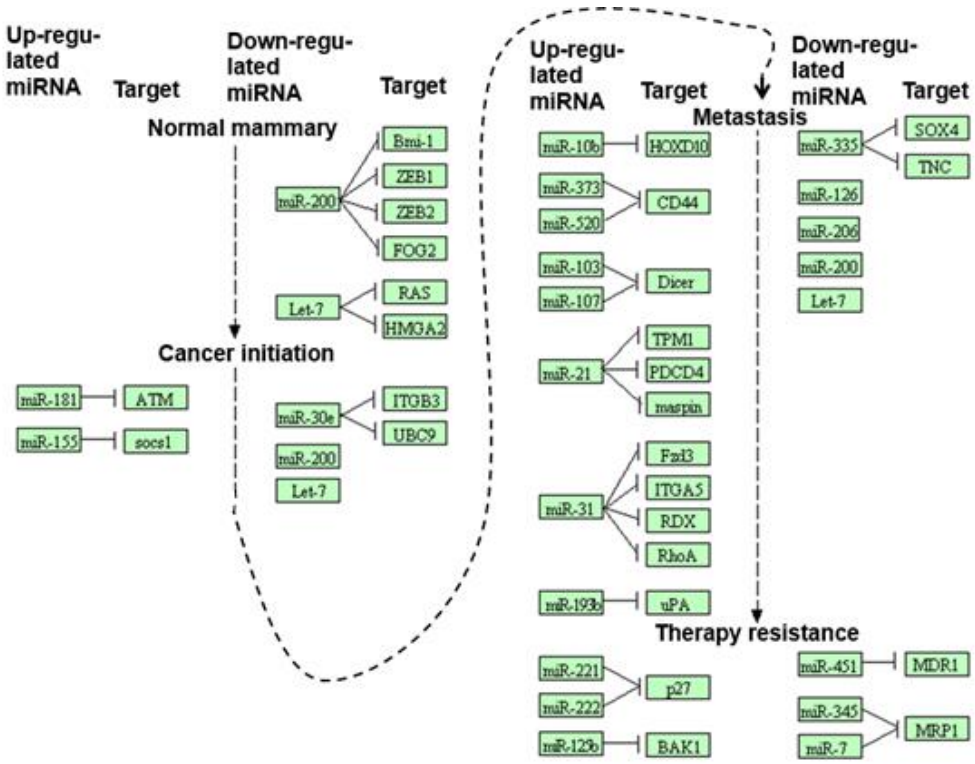

Figure 5. Breast cancer associated miRNAs. (Figure modified from Kyoto Enclycopedia of Genes and Genomes 2014).

1.9 Mixed integer programming as a tool for analysing gene regulation

There are numerous ways to analyse gene expression signals. Many of these methods are somewhat novel, for example statistical models using transcription factors and histone modifications, a biophysical interaction model applied to miRNAs, a hypergraph-based learning method using miRNAs, dynamic models together with both miRNAs and transcription factors, and linear programming evaluated together with transcription factors and miRNAs (Cheng and Gerstein 2012; Cheng et al., 2012; Khorshid et al., 2013; Kim et al., 2013; Lai et al., 2012; Nazarov et al., 2013; Setty et al., 2012, respectively). A linear programming method has been established to be an efficient tool for inferring miRNA-mRNA regulative relationships (Tuschen 2013). Linear programming is an adequately robust method for relating the mRNA expression signals to miRNA expression signals with a relatively reasonable correlation. Notwithstanding, dynamic models are most likely the most biologically realistic models (Lai et al., 2012), and in following studies they could be an interesting comparative method for checking the linear programming results. This would mean that there should be the same time series data available for both dynamic models and linear programming models. In subsequent studies additional factors, such as transcription factors, histone modifications, and distance relationship of miRNA target sites to the efficiency and coordination of degradation should be added in the model. In that case the nature and the amount of the governing equations in the linear programming would change.

Linear programming is generally used to optimize an objective function with constraints and variables (Williams 1999). In economics it is implemented to maximize the profit coming from products including the costs of making and delivering them. In gene expression analysis it is applied for minimizing the error between the real signals of mRNAs and the estimated ones by relating transcription factors to miRNAs or to other explainers of the signals. Most of the constraining equations are typical linear equations. These equations confine a space or plane that has optimal points, from which the best is chosen, that is to say it is optimized, according to a proper mathematical algorithm, such as simplex algorithm. The linear programming variables may be preceded or multiplied by some constants, i.e. coefficients, that are usually continuous real

variables. In some cases, they may also be just binary variables (0/1), or in more general terms, integer variables. The linear programming model is converted to a mixed integer programming (MIP) model, if the model has both real and integer variables, which is the reason for the use of word ‘mixed’. MIP models can be optimized with branch-and-cut algorithm.

Linear programming is usually performed with a computer program. R programming language can be utilized for this, since it has an additional package called *Gurobi* (Gurobi 2014), which is also an efficient standalone program. The application of the package is not straightforward, since it requires a specific form of input, which is more complex than in the standalone version of this program namely the linear programming files. The efficiency of this package in R comes from its incorporation of the powerful commands in R language. It is possible to construct loops that handle the optimization between many miRNAs and mRNAs with just one page of coding. It gives the modelling results in orderly fashion with all the relevant output that the modeller designs. One can also insert additional parameters, such as the applied optimization method and MIP Gap. This MIP Gap is a portion value between optimal and current solver result. Setting this value high can speed the obtaining of the results. On the other hand, it may lower the correlations between real and estimated values of gene expression.

1.10 Motivation

This work was a part of a research project of the German Cancer Research Center in Heidelberg, Germany. The project studied the role of TNBC genes in the urea cycle by experimental and analytical methods. The urea cycle happens in hepatocytes and renal cells and it produces urea from ammonia. Dysregulation of some of the urea cycle genes has been linked to assist the development of cancer. For example, the down-regulation of a urea cycle gene Carbamoyl Phosphate Synthetase 1 (*CPS1*) by DNA methylation has been shown to occur in human hepatocellular carcinoma (Liu et al., 2011). Apparently, some enzymes in the urea cycle, such as argininosuccinate lyase, can increase the activity of other substrates, such as arcinosuccinate, when the gene of the enzyme is down-regulated, e.g. also by short hairpin RNAs (Zheng et al.,

2013). This was one of the reasons why other (epigenetic) regulation mechanisms besides the DNA methylation were considered in this work; mainly the role of miRNAs in the regulation of TNBC, for elucidating the relevance of miRNAs in oncogenesis. There was collaboration between the University of Jena, the University of Turku, the Heidelberg University, and the Hans Knöll Insitute at Jena, where this research was carried out. For the system biological part of this project, miRNAs and mRNAs were investigated by exploring the relationships between expression signals from The Cancer Genome Atlas (TCGA).

The effect of miRNAs to the regulation of TNBC was studied with statistical and mathematical methods. The main method was based on linear programming models explaining transcript levels by the regulation of miRNAs. The method was extended to MIP, which employs a certain amount of miRNAs and their target genes. This approach could give the most correlated genes to the expression signals of miRNAs. Enrichment analyses were likewise performed for the purpose of finding the relevant miRNAs in TNBC.

2 Materials and methods

2.1 Evaluation of TCGA's expression signals

TCGA has expression data for mRNAs, and miRNAs in many cancer types with multitude of different samples. This sample data has come from different institutions that have measured the patient samples with a next-generation RNA sequencing device, HiSeq 2000 RNA Sequencing (Illumina, USA). The expression data is freely available and it is length normalized with Reads per Kilobase (mRNA/miRNA) per Million mapped reads (RPKM) method on account of the RNA sequencing protocol (Mortazavi et. al., 2008). This protocol tries to gain sequence coverage for all of the transcripts, where the longer transcripts will have more reads than shorter ones (Oshlack and Wake 2009; Figure 6).

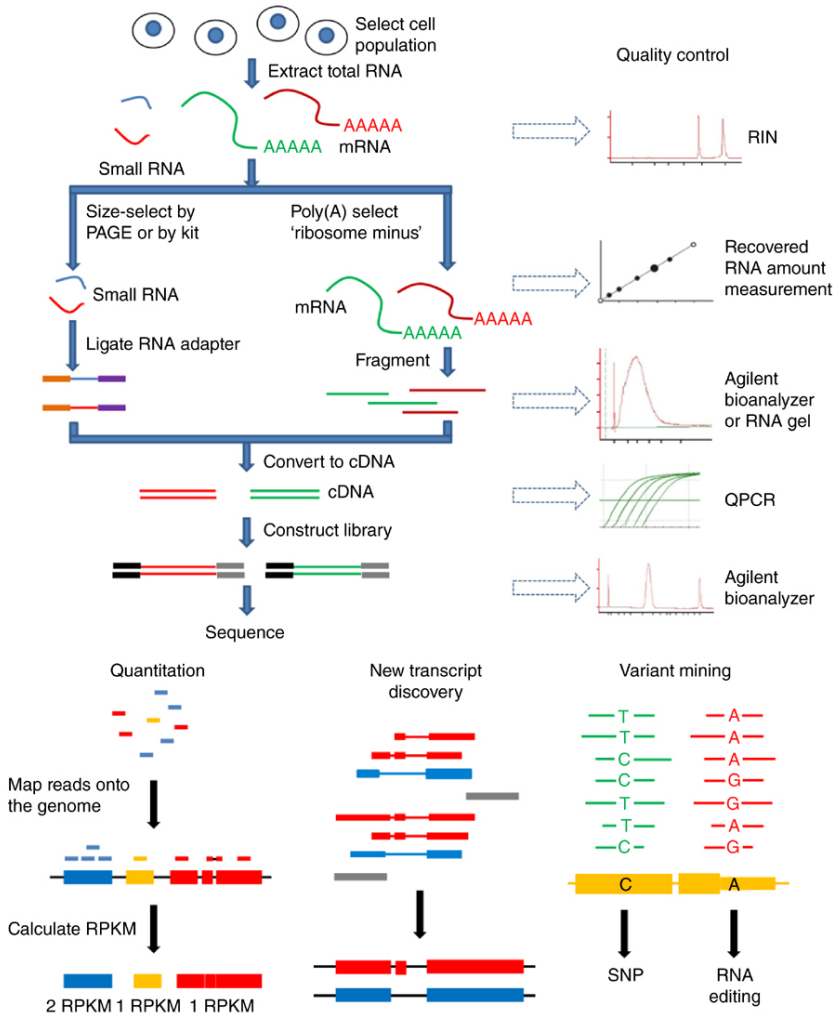

Figure 6. The principle of RNA sequencing protocol. In this picture Agilent bioanalyzer was used, where RIN is RNA integrity number for assessing the extent of RNA degradation. (Picture taken from Zeng and Mortazavi 2012.)

The data from TCGA is organized in files for each patient and further annotated in a separate file. This so called annotation file can be utilized to infer specified knowledge regarding the samples, such as which samples were from tumour patients and which from normal patients. TCGA has information for 20,531 breast cancer associated mRNAs in each of its 1,100 samples. Similarly, it has 1,046 breast cancer related miRNAs in each of its 770 samples that were not all the same as with mRNA case. Throughout the materials and methods, all the data preparation and procedures can be executed with R programming language (R, 2013), unless otherwise stated. These coding procedures are widely used in the literature (Matloff 2011; Zuur et al., 2009; Hahne et al., 2008). First, TCGA's mRNA Entrez IDs, which are generally used by the National Center for Biotechnology Information (NCBI) database, were changed to the real gene symbol names. According to annotation file, the expression information was correctly available only for some of these samples. One could then divide these correctly annotated samples to normal and tumour samples. The accession number names of miRNAs (starting with MIMAT) were converted to standard miRNA names with a list provided by the miRBase (miRBase 2014). Each of these miRNAs were listed for each sample, and each miRNA had several reads per million values, whereas mRNAs had just one. These read entries were summed for each miRNA as introduced by the Broad Institute (Broad Institute 2014).

The samples between miRNAs and mRNAs were matched. The same annotation regarding to tumour types and other information in the annotation file could also be then related to miRNAs. The total amount of TNBC and normal samples was analysed with information from the annotation file for mRNA data and then matched to miRNA data. The rest of the matching samples could be then deduced to be other breast cancer types. The workflow of this thesis is depicted at Figure 7.

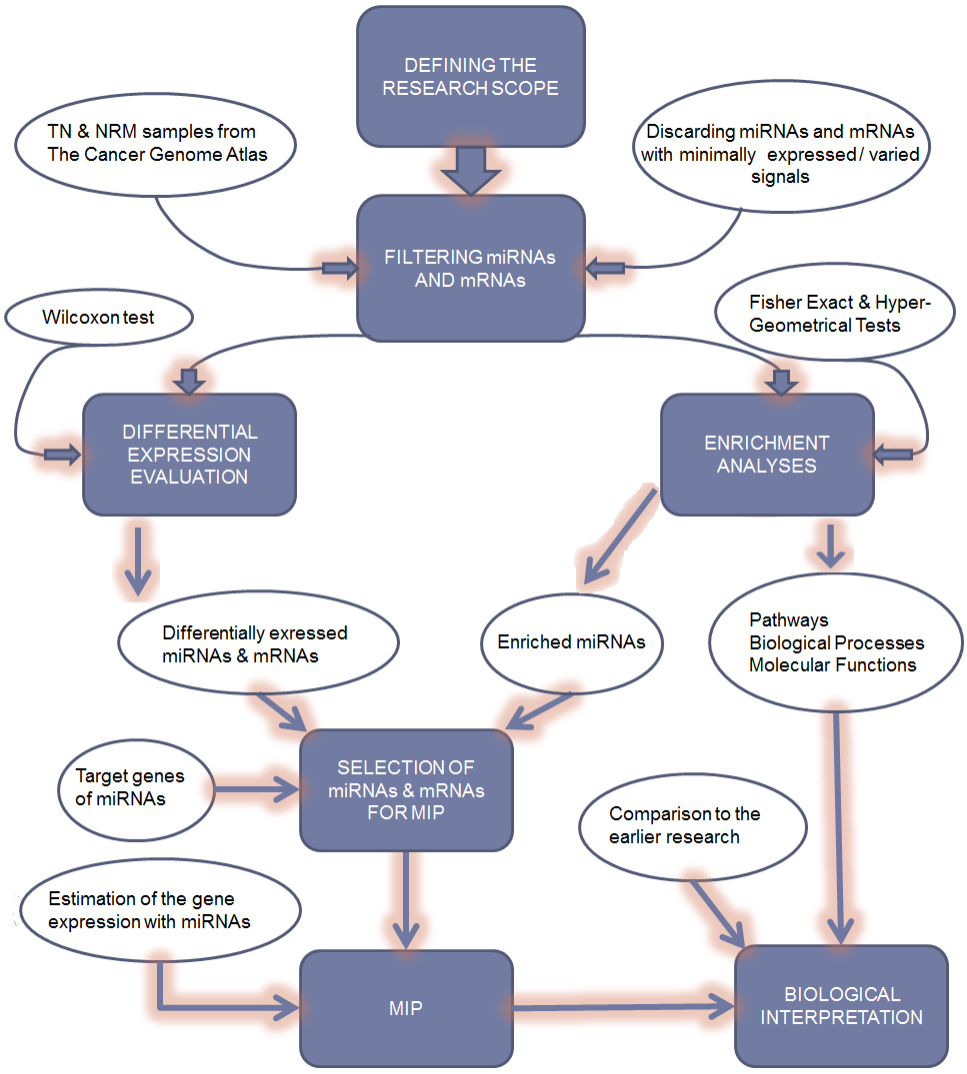

Figure 7. Workflow of the experimental part of thesis. First the research was defined to consist of the regulative role of miRNAs in TNBC. Secondly, appropriate data was downloaded from TCGA and filtered. This data contained the expression signals for mRNAs and miRNAs. The aberrantly expressed miRNAs and mRNAs were evaluated with Wilcoxon tests. Enriched miRNAs were obtained by evaluating differentially expressed target genes by Fisher exact tests. The genes for MIP were selected from differentially expressed genes, which were regulated by at least two enriched miRNAs according to the target gene list. MIP model correlated the target gene expression signals with miRNA expression signals by minimizing the error between real and miRNA estimated values of gene expression. Finally, the biological relevance of the results of MIP model could be analysed.

Most of the minimally expressed miRNAs and mRNAs were discarded. It was assumed that these expression values could arise from background noise or other unintended sources. More precisely, it was required that 10% of the initial expression values of miRNAs and mRNAs must be above five and ten, respectively. These filtered mRNAs were now regarded as the reference gene list. There are three levels of data with different types of processing status available in TCGA. It was suitable to use the most processed data that TCGA

had. This turned out to be the level number three mRNA expression data, which has been normalized according to the information provided by TCGA (2014). This mRNA expression data was duly Z-transformed. The normalisation for the processed miRNA expression data was performed with *justvsn* package in R, which is similar to *vsn* package (Huber et al., 2002) and then Z-transformed.

2.2 Cluster analysis

Cluster analysis was carried out for three different types of cases: 1) TNBC samples with miRNAs or mRNAs; 2) specific genes explained below; 3) miRNAs with TNBC samples or normal samples. The clustering analysis for the first case with TNBC samples assisted the categorizing between different tumour groups and a normal group using either miRNA or mRNA expression data. The previously filtered data of mRNAs and miRNAs with matching samples was limited by selecting 75% quantiles of both expression data. This limitation was performed to get the most informative group of mRNAs and miRNAs (Tuschen 2013). mRNA expression data was even further filtered by reducing the low varied expression data. A histogram of mRNA variances was plotted and mRNAs with near zero (0.8) variance were discarded.

TNBC samples were clustered using *hclust* package in R according to Ward's method (Hartigan 1975) with the Euclidean distances from the expression data. The stable clusters of the Ward's clustering results were searched with *pvclust* package in R that employs a multiscale bootstrapping method (Suzuki and Simodaira 2006). The second clustering case requires some definitions. For example, the enriched miRNAs (EmRs) are those miRNAs that can be specifically observed at certain conditions of TNBC's gene expression. Equally important is to note that EmR is a differentially expressed miRNA (DEmR), which differentially expressed target genes are more significantly involved in TNBC compared to the genes that are not targets. In addition, a differentially expressed gene (DEG), or DEmR is a gene (or miRNA) that is specifically found to be aberrantly expressed in certain conditions, such as in TNBC compared to normal case. The methods how DEGs, differentially expressed target genes, DEmRs, and EmRs have been searched are depicted in the next chapter (2.3). The second case consisted of differentially expressed target genes with at least

two affecting EmRs in TNBC. The genes in the case number two were collected to a list that is hence forward called the cooperation list. The designated name to this list comes from a finding that some miRNAs could have a synergistic cooperative role in regulation of their target genes (Chen et al., 2012).

Hamming distances were calculated according to a certain comparison for the genes in the cooperation list. In this comparison, the incidences of same EmRs and the absence of same EmRs, i.e. patterns, were evaluated for each gene pair in the list. If these patterns were nearly the same, then the distance between evaluated genes would be small. In practice, this was accomplished by constructing a matrix of zeros and ones, regarding the patterns, and comparing two rows EmRs for two different genes. Finally, this distance information was utilized in a similar clustering approach as with the case of TNBC samples (using Ward's method). The difference was that instead of getting clusters of TNBC, normal and other breast cancer samples, one got clusters of genes. These gene clusters could be analysed visually according to the major branches of the cluster dendrogram. The distances between leaves of the cluster dendrogram image can be calculated by using the heights mentioned in the y-axis. Alternatively, the closeness of the leaves is evaluated with two different kinds of P values from *pvclust*, Approximately Unbiased (AU) and Bootstrap Probability (BP), obtained with bootstrapping (Nboot=500, α=0.99).

2.3 Differential expression analysis

In order to find out the aberrantly expressed miRNAs and mRNAs in TNBC, differential expression analyses were performed for both of these cases. These analyses were statistical tests, where the miRNA (or mRNA) data from normal samples was compared to TNBC samples. The tests were carried out for the filtered expression data. In this case, Wilcoxon test was employed, on account of unequal tail sizes witnessed in the data of histograms for both miRNAs and mRNAs (Wilcoxon 1945; Bauer 1972). Wilcoxon test is a statistical hypothesis test for comparing two related samples. The test calculates the positive and negative differences between the expression values of TNBC samples and the median of normal samples, and adds these distances together

yielding a sum of signed ranks. If the resulting sum is significant, then the P value will be low. In that case, the expression signals from TNBC samples behave differently from normal samples. These signals, either arising from miRNA or mRNA data, are differentially expressed. The P values of these tests were False Discovery Rate (FDR) corrected with Bonferroni method of multiple testing correction, and utilized for inferring the differential expression. These corrected P values are better known as the Q values. If the normal and TNBC cases were different, then the Q value was low. The Q values' lower cut-off limit was set to 0.05, or in mathematical terms to $Q \leq 0.05$, which is commonly used in statistics.

2.4 Target gene analysis

The target genes for the unique mature miRNAs from TCGA's breast cancer samples were obtained by the usage of a predefined prediction tool (Tuschen 2013). Notably, it selected the most overlapping target genes of the most experimentally validated targets from the predicted ones, which were obtained by different methods and internet databases. According to Tuschen (2013) only TargetScan (2012), PicTar (2014) and the probability of interaction by target accessibility (PITA 2014) databases and their methods were sufficient for obtaining these target genes. All of the targets were then collected into a target gene list. Only those miRNAs were selected to this target gene list that had at least one of these target genes.

2.5 Enrichment analyses

2.5.1 Analysis methods and data

The enrichment analyses were executed with Fisher exact tests (Agresti 2002) and the hypergeometric tests (Ahlblom 1993) using different data and tools for both of these methods. The first test was employed to find out Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways for differentially expressed genes in TNBC, and EmRs in TNBC. The second test was conducted in order to find out Gene Ontological (GO) terms of molecular functions and biological processes for specific sets of genes (chapter 2.5.3). This test was also applied for assessment of KEGG pathways associated with these gene sets. The tests were performed case-by-case either with R, or GeneCodis internet tool

(Carmona-Saez et al., 2007; Tabas-Madrid et al., 2012). The methods of these tests were different not just for the mathematical point of view, but also for the way how data was employed and what it consisted of.

2.5.2 Fisher exact tests for differentially expressed genes

The enrichment analysis with Fisher exact tests included finding EmRs and KEGG pathways in TNBC. The reason for searching EmRs, and not solely employing DEmRs, was that these miRNAs would regulate the differentially expressed target genes more likely than other TNBC related DEGs due to the nature of this test. These miRNAs were selected among the DEmRs using the test in a specific way. The regulation direction of DEmRs had a special role in this test, owing to the fact that up-regulated miRNAs in TNBC were redeemed to be down-regulating their target genes, and vice versa. As an illustration – the less miRNAs degrading targets the more there would be these targets. The up-regulation of miRNA in TNBC was defined as a positive value arising from extracting the median expression levels of this miRNA in TNBC from normal samples. Similarly, the down-regulation was defined as a negative value arising from this extraction. This definition applied in the same fashion to the target genes. Needless to mention, the two-tailed Fisher exact test employed here did not use miRNA nor mRNA expression values, but the amount of genes in four different cases. Each of the DEmRs was evaluated one-by-one with this test so as to search if it was EmR. The contingency table of this test (Table 3) received these amounts of genes from: a) the oppositely regulated differentially expressed target genes compared to the evaluated DEmR, b) the remaining differentially expressed target genes with similar regulation direction as with the evaluated DEmR, c) DEGs that were oppositely regulated compared to the evaluated DEmR, d) all the other genes in the reference gene list that were not the target genes or differentially expressed, and their regulation direction was not considered.

Table 3. The principle of the contingency tables of Fisher exact tests used in the enrichment analysis for finding EmRs in TNBC from the differentially expressed ones.

	Down-regulated / Up-regulated genes	Remaining genes	Sum
Differentially expressed target genes	a	b	(a+b)
DEGs[1]	c	d	(c+d)
Tot.	(a+c)	(b+d)	(a+b+c+d)

1) DEGs are the differentially expressed genes.

The up-regulation and down-regulation of the target genes or miRNAs in TNBC was calculated by extracting median values of TNBC samples from median values of normal samples. There was one particular reason why this information of up-regulation or down-regulation was not redeemed to be sufficient one *per se* to deduce the involvement of the target gene or miRNA in TNBC. When these extractions of medians would be close to zero, some of the target genes or miRNAs, which were deduced to be down-regulated could actually be up-regulated, and vice versa. In addition, the variances of these two sample types could also be considerably different (Auer and Doerge 2010; Wolfgang et al., 2002). So, with the intention of receiving conservative estimates, DEGs or DEmRs were utilized to make sure that these target genes or miRNAs were really dysregulated in TNBC. Nevertheless, the negative values and the positive values, resulting from these median extractions, were implemented as such, without any minimum cut-off limits for the median differences, to deduce the down-regulation and up-regulation of these differentially expressed target genes and DEmRs. Accordingly, the genes that were differentially expressed were incorporated in a, b, and c (Table 3) rather than genes that were not differentially expressed, because then it was certain that these genes were involved in TNBC, and thus also the miRNAs that potentially regulated them. The P value of this test was FDR corrected with Benjamini Hochberg (BH) method of multiple testing correction. The evaluated miRNA was considered to be EmR in TNBC, if the calculated Q value, that is to say the corrected P value, for this miRNA from the test was $Q \leq 0.05$.

The aberrantly regulated KEGG pathways in TNBC were obtained with Fisher exact tests using R with DEGs. In similar fashion, as with the enrichment test for

DEmRs, the amount of selected genes in TNBC, as defined below, was considered. This time the regulation was not in the main role, but the participation of DEGs in an evaluated pathway. The contingency table received the following amounts of genes: a) DEGs in the considered pathway, b) DEGs that were not in this pathway, c) genes in the pathway (but not differentially expressed), d) all the other genes in the reference gene list that were not inthe pathway and not differentially expressed. The P values of these tests were also FDR corrected with BH method of multiple testing correction. Significantly dysregulated KEGG pathways in TNBC were considered to be the ones with $Q \leq 0.05$.

2.5.3 The hypergeometric tests for specific genes

The enrichment tests for specific sets of genes (A-M; Table 4) were accomplished by the hypergeometric tests. This type of test was employed not only for the reason that it could give comparative results against the results from Fisher exact tests, but also because of the efficiency of tool utilized with this method.

The tool requested the genes of interests, i.e. A-M, and the reference gene list for the calculation of the P values, which were FDR corrected with BH method of multiple testing correction. The resulting Q values for GO molecular functions, GO biological processes, and KEGG pathways were requested to be below 0.05 so as to recognize their involvement in TNBC or other biological scenarios associated to these gene sets. The actual regulative role of miRNAs can be fully related to genes only later in MIP analysis. This is why these tests were redeemed as complementary to MIP analysis.

Table 4. The gene sets for the hypergeometric tests.

Gene set[1]	Description of the gene set
A	Down-regulated genes that are targets of up-regulated EmRs
B	Up-regulated genes that are targets of down-regulated EmRs
C	Down-regulated DEGs that are targets of up-regulated miRNAs
D	Up-regulated DEGs that are targets of down-regulated miRNAs
E	Down-regulated genes in the cooperation list[2]
F	Up-regulated genes in the cooperation list
G	Genes in the cooperation list
H	Down-regulated genes of group[3] one that are targets of up-regulated EmRs
I	Up-regulated genes of group one that are targets of down-regulated EmRs
J	Down-regulated genes of group two that are targets of up-regulated EmRs
K	Up-regulated genes of group two that are targets of down-regulated EmRs
L	Down-regulated genes of group three that are targets of up-regulated EmRs
M	Up-regulated genes of group three that are targets of down-regulated EmRs

1) The gene set symbols (A-M) are used to refer to the genes in this set. 2) The cooperation list is a list of 247 genes selected from DEGs, which are regulated by at least two different EmRs. 3) Groups one to three are the clustering results for the cooperation list using Ward's method with Hamming distances calculated for gene pairs with similar EmR patterns.

In order to see the significance of certain genes in these enrichment tests compared to the unenriched ones, also odds ratios were checked. If the odds ratio were high, then the genes were more prominent in explaining the test result. The odds ratios were calculated for DEGs at the enrichment analysis for KEGG pathways. They were also calculated for the specific sets of genes (A-M) at the enrichment analyses for GO molecular functions, GO biological processes and KEGG pathways. The odds ratios can be calculated as in the equation 1:

$$odds\ ratio = \frac{a/b}{c/d} \quad (1)$$

, where a, b, c, and d are the values in the corresponding contingency table (Table 3).

The DEGs and EmRs can be visualized in a cellular pathway image using R's *pathview* package (Luo 2013). The DEGs that are regulated by EmRs can be also indicated in this picture.

2.6 Linear and mixed integer programming

A linear programming model solves the most favourable outcome of the linearly dependent parameters of its function. An appropriate algorithm, such as branch-and-cut, is utilized for the optimization, which in this case is minimization of the objective function of this model. This function in equation 2 is a modified version of an equation introduced by Tuschen (2013). The linear or MIP problems are subjected to certain constraints, such as defined in equations 3-6. Albeit not all of these constraints are applicable for both of the problem cases, as will be discussed later. The error term to be minimized is defined below:

$$err = \left| g_s - (\beta_o + \sum_{i=1}^{k} m_{si}\beta_i) \right| \quad (2)$$

, where the error (err) is between real and estimated signal values, g_s = gene expression signals in a patient sample number s, β_o = a constant, which balances the estimation arising from miRNA expression signal values, m_{si} = miRNA expression signals in a patient sample number s that has amount of i many miRNA expression signals (where k is the maximum number of miRNAs that can be present at any given time according to the target gene list), and β_i = i many constants for miRNA expression signals that compose the backbone of the estimation model.

Only the filtered mRNA and miRNA values were employed in this modelling. The constraints for linear or MIP models can be noticed from the equations 3-6 below.

$$\beta_o + \sum_{i=1}^{k} m_{si}\beta_i - err \leq g_s \quad (3)$$

$$-\beta_o - \sum_{i=1}^{k} m_{si}\beta_i - err \leq -g_s \quad (4)$$

$$\beta_i \leq 0 \quad (5)$$

$$\sum_{j=0}^{k} x_j \leq \alpha \quad (6)$$

, where x_j = a binary variable for each miRNA. According to this formulation miRNAs reach their maximum frequency only if $\alpha \cong k$, where α = a limit set to the miRNA amount. This value is an integer, transforming the linear programming model to a MIP model. These coefficients are fitted so that the objective function (err) will reach its minimum value. The other terms in equations 3-5 are depicted after equation 2.

This kind of modelling results in β coefficients, which may be employed with real miRNA expression signals for the estimation of real expression signals of genes with certain accuracy, i.e. correlation. Different sample sizes s can also be used to depict the correlations between miRNAs and genes in different scenarios. These different samples are: TNBC samples, normal samples, luminal A (breast cancer type) samples, random samples or various sample subsets, including all samples.

One method to check the consistency of the model, is to do a cross validation, such as leave-one-out, where the model is trained with some subsets of samples and validated with the remaining subset so that all subsets are validated in the end. In this case, only five different sample subsets of TCGA expression data were applied, while there could have been more subsets, since this method was merely redeemed as an additional way to check the data. The mean of the correlations from these validated subsets was calculated to estimate the overall performance of the model.

The amount of miRNAs employed by the model was limited so as to check that there would not be any overfitting of miRNAs. This overfitting phenomenon arises from the model trying to fit too many miRNAs explaining mRNA expression signals that are actually not necessarily present in the real biological cases of gene regulation. MIP modelling was suitable for limiting the amount of miRNAs used by the model. The limitation comes from those constraints that are set to integers, and here more specifically binary variables that can have values of either zero or one. A miRNA is therefore either affecting or not affecting to the regulation of a gene, and not something between. This does not remove the fact though, that some constraints such as equation 5, are still real

variables. This is why MIP model had the word ‘mixed’ designated for describing it. The relevant results of this work were obtained with MIP modelling.

Linear programming was initially chosen as a basis for analysing the effect of miRNAs to the regulation of TNBC genes. This primary modelling approach provided a good training platform in the first stages of developing the models. The modelling was later expanded in three-folded fashion: 1) limiting the amount of miRNAs applied by the model for the purpose of avoiding overfitting of too many miRNAs explaining the expression signals of mRNAs, 2) forcing the coefficients of linear equations to be below zero so that the effect of degradation by miRNAs to genes could be analysed, and 3) using different sample subsets and cross validation. The first expansion of this model changed it from linear programming to MIP.

MIP model for the genes and miRNAs was constructed with R’s *Gurobi* package (Gurobi 2014). The Gurobi program can also be executed as a stand-alone academic version, where a certain type of model file can be optimized using standard linear programming algorithms. This program is a very efficient linear programming solver that has interfaces for many programming languages such as also Perl. R’s *Gurobi* interface requires the objective function as a vector with the size of the total amount of variables and the error term ‘variables’ in the model. This vector has as many error term ‘variables’, as there are patient samples. In order to give the right form of objective function to R’s *Gurobi* interface, then only the error term ‘variables’ are considered by giving them value of one, and setting the other (real) variables to zero. The reason for this is that the interface calculates the β coefficients as ‘variables’ and considers the expression values as constants. The amount of equations, incorporated in this model constructing, is the sample size times two so that the absolute value in equation 2 is taken into consideration as in constraints 3-4. The maximum or minimum values for each variable can be given separately of the constrain equations involved as vectors. All of the vectors and matrices are then given to the *Gurobi* function of R with proper parameters, such as MIP Gap value (0-1; a gap between the optimum and current solver value). An example code of MIP modelling is given in appendix A, which consists only of this example code.

3 Results and discussion

3.1 Clustering analyses

3.1.1 Preparations for clustering analysis

According to TCGA's annotation file (TCGA 2012), the expression information was correctly available for 921 samples, which was further divided into 104 normal and 817 tumour samples. Some samples were discarded due outliers. There were 567 matching samples between miRNA and mRNA data, from which 70 were normal samples and 95 TNBC samples according to the same annotation file. During the preparation of mRNA data, it was noticed that the expression values were relatively high with mean value of 1,158 across all the samples and mRNAs. It seemed that the normalisation from TCGA was conducted with some other way than expected. Usually high values of expression data would not work properly *per se*, e.g. in differential expression analysis, in linear programming, in distance dependency of miRNA target gene site analysis with Michaelis-Mentes related equations or in heat map analysis using R programming language, whether the values would be normalized or not. Subsequently, a log2 transformation was also applied to mRNA data that TCGA had seemingly normalized before Z-transformation. Furthermore, the most of the minimally expressed miRNAs and mRNAs were discarded resulting to 16,458 mRNAs and 332 miRNAs. This filtered data was limited by selecting 75% quantiles of both the expression data, and it yielded 4,115 mRNAs and 83 miRNAs. A histogram variances of mRNAs was plotted and mRNAs with near zero (0.8) variance were discarded yielding 710 mRNAs, before the clustering analysis to mRNAs was performed.

3.1.2 Clustering of samples

The breast cancer associated samples from TCGA could be organized into several groups according to the cluster analysis. The expression values of selected miRNAs with different sample types were adopted to perform this analysis. The sample types were TNBC, normal, and other breast cancer related samples. The medians for miRNAs' expression signals were usually higher in TNBC, that is to say 173 miRNAs out of 322 miRNAs, than in normal breast cancer samples, which foretold a plain clustering result. Accordingly, the selected sample types showed relatively clear clustering to three distin-

guished groups as seen in Figure 8. miRNA expression values in these three sample types clustered without noticeable separate islands of subgroups or in other words outliers. Whilst, there were not that many stable groups, there still were some, demonstrating that miRNA expressions could be applied for obtaining some significant TNBC groups. There were slightly more stable groups for the similar cluster analysis done with mRNA expressions suggesting that mRNAs, which quite often form to transcription factors, could have a better regulative role in some of the sample types of these groups, such as in TNBC (Figure 9). Nevertheless, the clustering analysis with miRNAs is in general more difficult than with mRNAs, because there are less miRNAs than mRNAs. This causes different variances, which have an effect on the implementation of Ward's method (Hartigan 1975). Obviously, miRNA expression signal values are often lower than mRNA expression signal values. The Z-scored medians for all miRNAs and mRNAs expression signals, in this case, were 5.31 and 8.50, respectively.

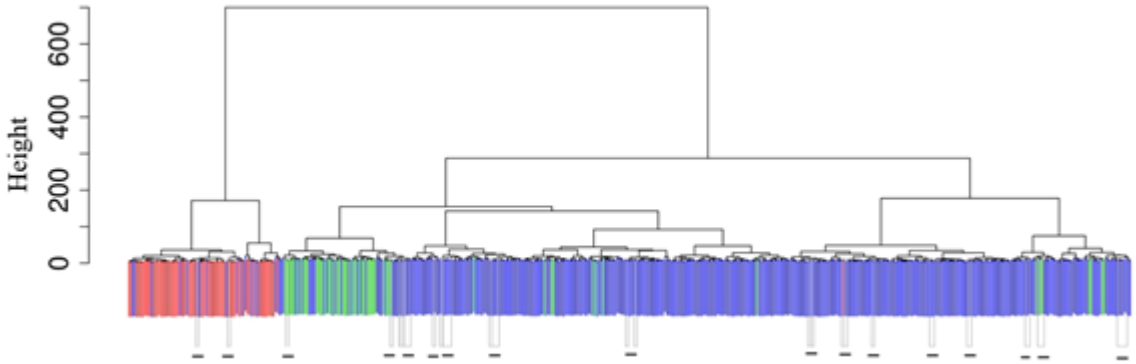

Figure 8. Clustering dendrogram based on filtered miRNA expression values (83 miRNAs) in terms of total expression level from 567 samples. Ward's method was used with Euclidian distances between these expression values. TNBC samples (54) are in red, normal samples (70) in green and other breast cancer cell samples in blue (443). The grey boxes (highlighted with black lines) below dendrograms leaves indicate highly significant clusters, which are strongly supported by the data having AU ≥ 0.99.

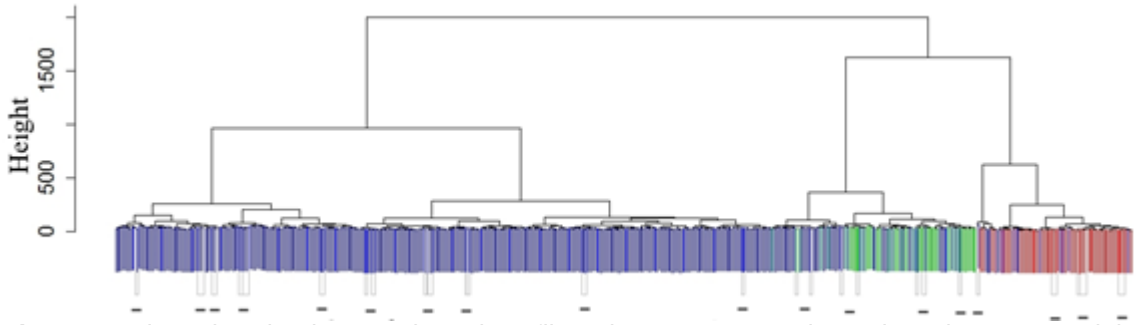

Figure 9. Clustering dendrogram based on filtered mRNA expression values (710 mRNAs) in terms of total expression level and variance from 567 samples. Ward's method was used with Euclidian distances between these expression values. TNBC samples (54) are in red, normal samples (70) in green, and other breast cancer cell samples in blue (443) colour. The grey boxes (highlighted with black lines) below dendrogram leaves indicate highly significant clusters, which are strongly supported by the data having AU ≥ 0.99.

3.1.3 Clustering of the genes in the cooperation list

The Hamming distances were calculated comparing the similarities of EmR patterns for each target gene pair in the list. If EmR was present in both of the genes or if the same EmR were not present, then the genes were close to each other. That is to say, the genes were at a close distance to each other, and subsequently receiving value of zero for the distance calculation instead of one. Finally, all the zeros and ones were summed for every gene pair. For example, the summed distance between a gene of Ataxin 1 protein (*ATXN1*) and a gene of Quaking Homolog, Kh Domain RNA Binding protein (*QKI*) was two, and the summed distance between *ATXN1* and a gene of GATA (sequence) Zinc Finger Domain Containing 2B (*GATAD2B*) was zero. In this case *ATXN1* was closer to *GATAD2B* than *QKI*. The results were compiled in a distance matrix that was adopted in a hierarchical clustering according to Ward's method (Figure 10). Three distinct groups of genes emerged. These genes could be further studied with the hypergeometric enrichment analyses. As a preliminary notion, these target genes for each EmR in TNBC, should occur more likely than the ones that are not the target genes. Thus, EmRs would be expected to regulate these target genes, in contrast to some other TNBC related genes. The first group was the largest, comprising of 147 genes, including a gene of Ligand-Dependent Corepressor protein (*LCOR*). This Ligand-Dependent Corepressor protein (LCOR) recruits C-terminal-Binding Protein 1 (CtBP) corepressor, and function as an attenuator of progesterone-regulated transcription in breast cancer (Palijan et al., 2009). The second group

was the smallest with ten genes, including B-Cell Translocation Gene 1, Anti-Proli-ferative (*BTG1*), which product inhibits breast cancer cell growth (Zhu et al., 2013). Finally, the third group, which was the second largest with 90 genes, including *QKI*, which represses a gene of Forkhead Box O1 (*FOXO1*; Yu et al., 2014). Forkhead box O1 protein (FOXO1) is an essential tumour suppressor for controlling cell proliferation. Differentially expressed target genes of these groups are evidently at least partly regulated by similar types of EmRs because of their closeness in the clustering.

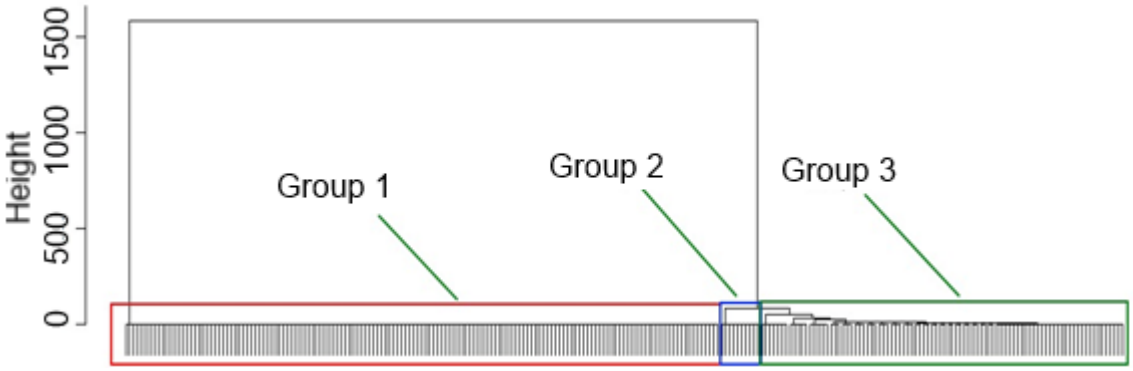

Figure 10. The clustering dendrogram of the genes in the cooperation list resulting three groups. It has been calculated by Wards' method. The method applied Hamming distances arising from miRNA related similarities of the genes.

3.1.4 Clustering of EmRs

The clustering patterns of EmRs were evaluated between TNBC and normal samples. These patterns could elucidate the expression similarities of miRNAs within the same sample types and between the different sample types. This could tell about an identical regulative role of the closely clustered miRNAs. As expected, there were differences between the clusters of EmR expression signals coming from either TNBC samples (Figure 11) or the normal ones (Figure 12). The miRNA expression signals in TNBC samples were observed to cluster into three different groups, whereas in normal samples they clustered into five different groups, if compared to TNBC case. It is noticeable in TNBC case (Figure 11) that the most up-regulated miRNA (hsa-miR-301b) grouped together with another also up-regulated miRNA (hsa-miR-877-5p) in the last group number three. The two other groups in TNBC case did not have such a clear distinction between up-regulated and down-regulated ones even if the most down-regulated miRNA (hsa-miR-186-5p) was in the first group of this

clustering. The clustering of miRNA expression signals from the normal samples behaved quite differently compared to TNBC samples. To illustrate this matter further (Figure 12); the first group was completely up-regulated, the second group comprised from up-regulated miRNAs except hsa-miR-200c-3p. In spite of that the last three groups were equally down-regulated and up-regulated including the most down-regulated miRNA hsa-miR-186-5p in the last group number five.

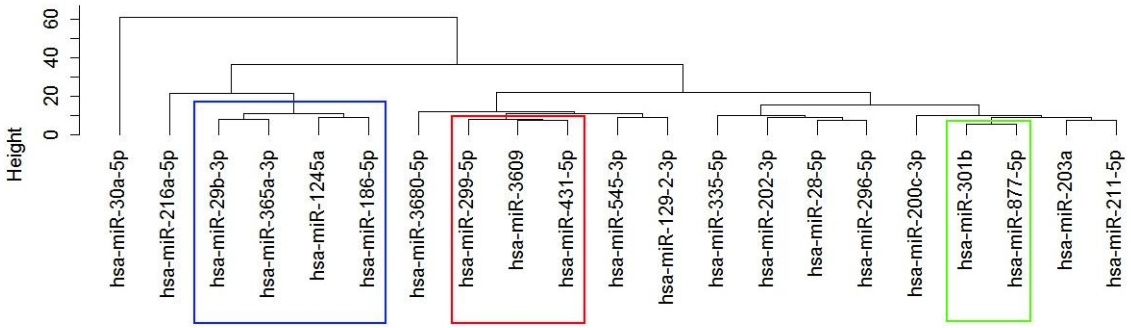

Figure 11. The clustering dendrogram of EmRs in TNBC samples. The expression signals of these samples were analysed with the average clustering method calculating the Euclidian signal distances. The three stable groups highlighted in blue, red, and green resulted from significant P values for robustness (P value ≤ 0.05) from a multiscale bootstrapping approach (Suzuki and Shimodaira 2006) performed in an R package *pvclust*.

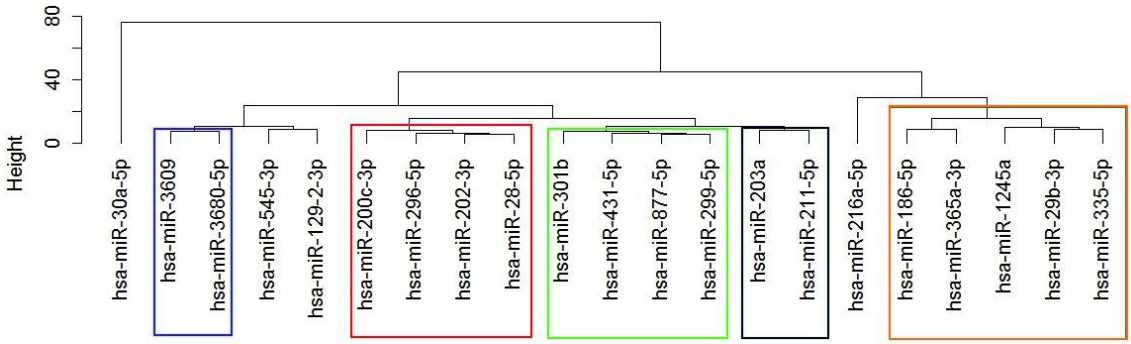

Figure 12. The clustering dendrogram of EmRs in normal samples. The expression signals of these samples were analysed with the average clustering method calculating the Euclidian signal distances. The five stable groups highlighted in blue, red, green, black, and orange resulted from significant P values for robustness (P value ≤ 0.05) from a multiscale bootstrapping approach (Suzuki and Shimodaira 2006) performed in an R package *pvclust*.

3.2 Differential expression analysis

A comparative analysis was performed using the data of the filtered expression signals from TNBC samples against the normal ones. This differential expression analysis was performed with Wilcoxon tests due to the skewness of the data noticed in histogram images. Both miRNA and mRNA signals were studied, and the selected results can be seen in Tables B1 and B2 of appendix B, respectively. The tables in appendix B depict the results of this differential expression analysis. There were 37 DEmRs and 1,963 DEGs. DEmR with the lowest Q value was hsa-miR-425-3p (Q value 0.0002) and DEG with the lowest Q value was gene in Chromosome 8 Open Reading Frame 33 (*C8orf33*; Q value 0.0001).

Hsa-miR-425 has been identified by Romero-Cordoba et al., (2012) with a previous differential expression analysis to be significantly involved (Q value 0.0068) in breast cancer. This miRNA was described to be up-regulated and to have three experimentally validated target genes: *Dicer*, *SMAD3*, and the gene of Platelet Derived Growth Factor C (*PDGFC*). The up-regulation of specific genes in TNBC, such as *C8orf33*, has been found out by Sasamoto et al., (2012) to be a prediction of the up-regulation of a gene of C-Jun Activation Domain-Binding Protein 1 (*Jab1*). C-Jun Activation Domain-Binding Protein 1 (Jab1) has a strategic role in breast cancer progression, for the reason that it is a target of HER2 (Hsu et al., 2008). Sasamoto et al., (2012) also reported that *C8orf33* plays a major role in mRNA processing. The protein of the second most differentially expressed gene, Chromobox Homolog 6 (*CBX6*), is a part of a Polycomb protein complex, which induces differentiation processes. Consequently, the down-regulation of *CBX6* in cancer, with for example miRNAs, is essential. Chromobox Homolog 6 protein (CBX6) has also been linked to hematologic malignancies due to the protein complex, implicating its greater role in the epigenetic regulation (Martin-Perez et al., 2010).

3.3 Enrichment analyses

3.3.1 Fisher exact tests with R

The enrichment analysis for finding EmRs out of DEmRs was carried out so as to evaluate the significance of the differentially expressed target genes of these miRNAs, and accordingly also the miRNAs themselves, in TNBC. The enrichment test calculated the Q values, with different gene frequencies, for each DEmRs. The lowest Q values were used to infer EmRs. This yielded 21 EmRs in TNBC (Table 5). The regulation direction adopted in this test was redeemed to be important, and accordingly they were assessed by extracting the medians of TNBC and normal samples. This yielded a maximum median value for up-regulation (8.70) and a minimum median value for down-regulation (-9.83). In most of the cases these medians were close to each other (0.55 and -0.72; 50% quantiles for medians in up-regulation and down-regulation, respectively). This lack of variation (Auer and Doerge 2010; Wolfgang 2002) was one of the reasons why differentially expressed target genes were adopted, instead of ordinary target genes, in this Fisher exact test for finding EmRs.

The most down-regulated miRNA was hsa-miR-186-5p and the most up-regulated one was hsa-miR-301b. Interestingly, hsa-miR-200c-3p was observed as the third most down-regulated miRNA, owing to the fact that miR-200 family has been shown to inhibit genes of Zinc Finger E-Box Binding Homeobox 1 and 2 proteins (*ZEB1* and *ZEB2*, respectively). The proteins of these genes are transcriptional repressors of E cadherin, and the down-regulation of these directs the Epithelial-Mesenchymal Transition (EMT). (Park et al., 2008.) The distribution of the target genes and differentially expressed target genes was also assessed (Table 5: last three columns). The average predicted amount of the target genes for EmRs was observed to be 422, whereas with the differentially expressed ones it was 48.

Table 5. EmRs in TNBC.

miRNA name	Q value[1]	Regulation of miRNA[2]	Target genes	Differentially expressed target genes	Ratio of differentially expressed target genes / Target genes
hsa-miR-186-5p	$4.215 \cdot 10^{-22}$	↓	108	16	0.15
hsa-miR-30a-5p	$5.255 \cdot 10^{-21}$	↓	1134	125	0.11
hsa-miR-200c-3p	$1.277 \cdot 10^{-14}$	↓	896	112	0.12
hsa-miR-211-5p	$2.280 \cdot 10^{-09}$	↓	602	68	0.11
hsa-miR-216a-5p	$1.792 \cdot 10^{-07}$	↓	229	29	0.13
hsa-miR-335-5p	$1.866 \cdot 10^{-05}$	↓	994	124	0.12
hsa-miR-299-5p	$2.716 \cdot 10^{-05}$	↓	873	97	0.11
hsa-miR-365a-3p	$1.152 \cdot 10^{-04}$	↓	105	14	0.13
hsa-miR-431-5p	$3.921 \cdot 10^{-03}$	↓	175	22	0.13
hsa-miR-301b	$6.936 \cdot 10^{-55}$	↑	179	23	0.13
hsa-miR-29b-3p	$2.815 \cdot 10^{-43}$	↑	105	20	0.19
hsa-miR-203a	$3.443 \cdot 10^{-43}$	↑	196	18	0.092
hsa-miR-545-3p	$1.238 \cdot 10^{-28}$	↑	426	48	0.11
hsa-miR-202-3p	$1.216 \cdot 10^{-25}$	↑	322	41	0.13
hsa-miR-3609	$6.701 \cdot 10^{-22}$	↑	220	21	0.095
hsa-miR-129-2-3p	$8.590 \cdot 10^{-22}$	↑	53	8	0.15
hsa-miR-877-5p	$1.719 \cdot 10^{-10}$	↑	452	44	0.097
hsa-miR-28-5p	$9.230 \cdot 10^{-10}$	↑	435	48	0.11
hsa-miR-296-5p	$8.160 \cdot 10^{-08}$	↑	572	45	0.079
hsa-miR-1245a	$4.024 \cdot 10^{-06}$	↑	740	84	0.11
hsa-miR-3680-5p	$1.285 \cdot 10^{-05}$	↑	52	6	0.12

1) The Q value is from Fisher exact test with two-sided tail. 2) Regulation: ↓ = down, ↑ = up (TNBC vs. Normal).

Interestingly enough, the ratio of differentially expressed target genes to the target genes had an average of 0.12, with quite small variance (0.000571; Table 5). This could denote that DEGs, and thereupon the miRNAs that are regulating them, reach an energetically favourable balance point, from which after the production of more mRNAs and miRNAs would not necessarily lead to a typical behaviour of TNBC, such as cancer cell proliferation.

Another enrichment analysis, using Fisher exact test, was performed for finding KEGG pathways in TNBC with DEGs. This analysis showed two significant pathways (Table 6). It was not quite unexpected to notice the proximal tubule bicarbonate reclamation as the most dysregulated pathway, since it is

connected to the urea cycle in human kidney, and evidently, according to the initial estimations, also involved in TNBC. Moreover, this pathway has been modelled with a porcine cell to be essential for a transepithelial drug transport in human kidney (Schlatter et al., 2006). These proximal tubular epithelial cells that were modelled play a major role in the synthesis of ammonia, implying yet again the involvement of the urea cycle. The other enriched pathway in TNBC, calcium signalling pathway, has been discovered to be pivotal in breast cancer (Davis et al., 2013). This pathway controls the starting of EMT that leads to an invasive type of breast cancer and therefore to the metastasis. The differences between the pathways examined here and in literature, e.g. in purine metabolism (Ossovskaya et al., 2011) are due to different samples (Cureline, Inc., San Francisco) and enrichment method, i.e. gene set enrichment analysis algorithm (Sivachenko et al., 2007).

Table 6. Significant KEGG pathways (Q ≤ 0.05) with DEGs in TNBC.

Pathway name	Q value[1]	Odds ratio	A[2]	B	C	D
Proximal tubule bicarbonate reclamation	$3.133 \cdot 10^{-14}$	0.07504	4	1959	384	14111
Calcium signalling pathway	0.007507	0.3664	13	1950	259	14236

1) The Q value is from Fisher exact test with two-sided tail. 2) A-D are the specific gene frequencies for the contingency table used in the test: A: DEGs in pathway, B: DEGs not in pathway, C: non-DEGs in pathway, D: non-DEGs not in pathway.

3.3.2 The hypergeometric tests with GeneGodis

GO molecular functions, GO biological processes, and KEGG pathways were searched for the selected sets of genes (A-M; Table 4) with the hypergeometric tests. These gene sets in general had various amounts of different genes: A: 148; B: 51; C: 556; D: 404; E: 177; F: 70; G: 247; H: 95; I: 26; J: 8; K: 1; L: 45; M: 24. The gene sets from H to M consisted of dysregulated genes of the three groups observed in the clustering analysis.

The hypergeometric enrichment tests yielded some interesting insights, with Q ≤ 0.05, for some of the evaluated sets. The regulative role of the miRNAs selected to govern these sets is in contrast disputable and cannot be established by these enrichment tests. For this reason, the results of these tests were considered to be complementary to the linear programming results, and can be found from tables C1-C15 in appendix C, which consist only of the re-

sults of these tests. Here merely some general aspects and main results of these tests are discussed.

The three hypergeometric enrichment analyses for genes in the gene sets A and B, that were regulated with opposite direction compared to their EmRs, yielded only three tables collected in appendix C's Tables C1-C3. In the optimal case, there would have been six tables arising from the three evaluations for each of these two sets. For instance, there was not a significant enrichment for the gene set A's GO molecular functions and KEGG pathways nor for the gene set B's GO biological processes. When the miRNAs were up-regulated in TNBC, as in set A, solely GO biological processes were significantly dysregulated. The neural crest cell development had both the lowest Q value (0.036) and also odds ratio (6.1), while the regulation of glucose import had the highest odds ratio (70), although somewhat higher Q value (0.049). The differences between these two Q values were not that high, suggesting that the regulation of glucose import could be more significant for down-regulated genes regulated by up-regulated EmRs. These EmRs might then have an inhibiting role in the regulation of glucose import in TNBC. Many types of cancer turn up their rate of glucose import (Warburg 1956; Dakubo 2010), so inferring from previous result, TNBC cells should have an alternative energy producing or importing mechanism instead. Continuing to the gene set B, where the up-regulated target genes regulated by down-regulated EmRs caused GO molecular functions and KEGG pathways listed in Tables C2 and C3 in appendix C. These functions were mostly enzyme activities, such as for myosin phosphatase and FAT10 activating enzyme, suggesting that these EmRs could enhance some reactions involved in breast cancer migration and invasion (Schwappacher et al., 2013), if their amount would be sufficiently low in TNBC. Apparently there was only one significant KEGG pathway, ubiquitin mediated proteolysis, observed for this set. This could demonstrate that these miRNAs might also play a part in migration of TNBC cells (Rossi and Loda 2003).

The tests for dysregulated DEGs in the gene sets C and D associated with the reversely regulated miRNAs, which could also be found in normal breast cells, yielded only GO molecular function for the gene set D. The result for

these down-regulated miRNAs is depicted in Table C4 in appendix C. This could demonstrate that normal miRNAs do not degrade DEGs, but have a different kind of function. Merely genes having a protein binding function, such as gene of Insulin-Like Growth Factor-Binding Protein 3 (*IGFBP3*; Martin et al., 2014; Lin et al., 2014) were now more prominently up-regulated. This was assumedly so on account of the lower activities of normal miRNAs in TNBC according to the tests. Subsequently, adding specific normal miRNAs to TNBC cell could diminish the functions of it.

The genes in the cooperation list were evaluated in three different setups in the gene sets E-G: down-regulated genes, up-regulated genes, and all the genes in the list regardless of their regulation direction. GO biological processes were noticed only, when the genes were down-regulated, which is seen in Table C5 of appendix C, or when all of the genes were adopted, as can be observed in appendix C's Table C6. Apparently, the genes in insulin receptor signalling pathway, and histone-serine phosphorylation were down-regulated in TNBC. This insulin receptor signalling pathway resulted when all the genes in this list were tested; demonstrating the importance of insulin, similarly as in the gene set D. Disrupting insulin receptor has also been identified earlier as a potential way to disrupt the cancer progression (Sachdev and Yee 2007). Nevertheless, this pathway was not on the top of the list according to the Q value rather than according to odds ratio. An example, how miRNAs affect certain genes in a significant biological process discovered during these enrichments, the insulin signalling pathway, is shown in Figure 13.

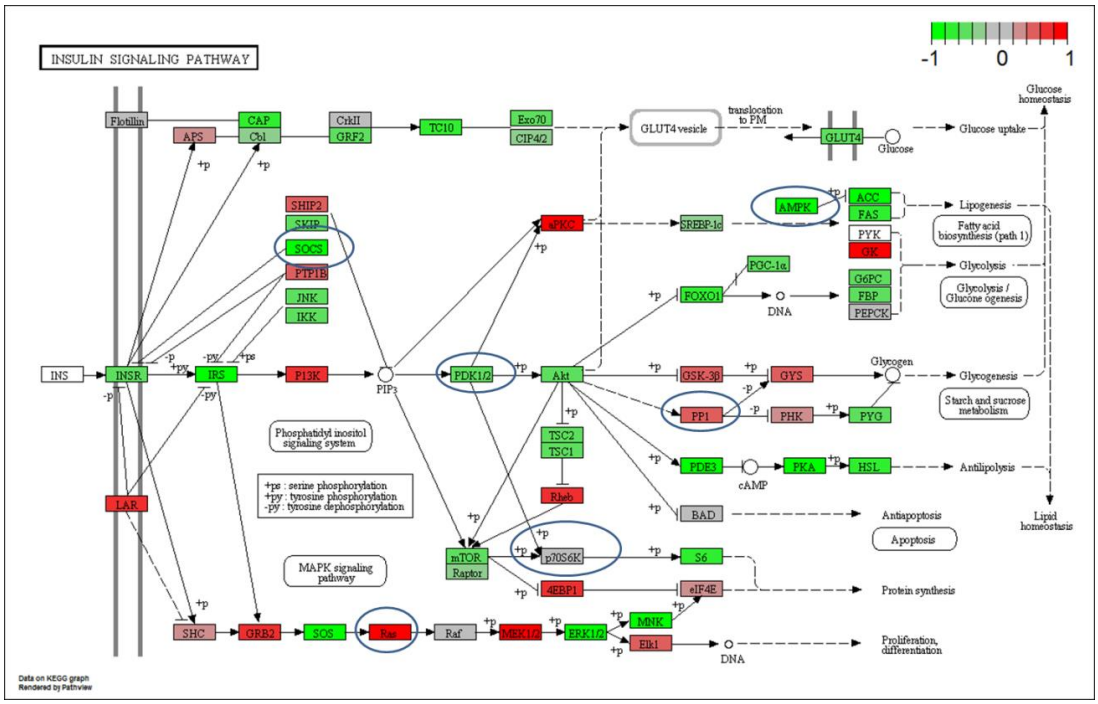

Figure 13. Aberrant regulation of genes in TNBC evaluated with TCGA's expression data for the insulin signalling pathway. This pathway was clearly significant in enrichment tests for specific sets of genes (E and G). Red is up-regulation and green is down-regulation, scaled from -1 to 1. The circled genes of these proteins are affected by EmRs. These proteins are Phosphoprotein Phosphatase (PP1), 5' Adenosine Monophosphate-Activated Protein Kinase (AMPK), Ribosomal Protein S6 Kinase, 70kDa, Polypeptide 1 (p70S6K), Pyruvate Dehydrogenase Kinase, Isozyme 1 (PDK1), Suppressor of Cytokine Signalling protein (SOCS), and RAS. (Figure modified from KEGG 2014).

The gene set H and I yielded significant enrichment only in some GO molecular functions, which are listed in Tables C7 and C8 in appendix C. Myosin phosphatase activity and receptor tyrosine kinase binding functions were the most common ones somewhat similarly as with the gene sets B and D as can be noticed from appendix C's Tables C2 and C4. In addition to previous findings by Schwappacher et al., (2013), it has been shown (Kim and Adelstein 2011) that lysophosphatidic acid (LPA)-induced migration requires also myosin phosphatase activity in breast cancer cells. This could mean that there might be a special synergistic regulation pattern behind especially up-regulated genes and miRNAs, regardless of miRNA's regulation status or type.

The enrichment tests for the gene set J, that is the group two's down-regulated genes that are targets of up-regulated EmRs, substantiated all of the enrichment facilities as can be noticed from Tables C9-C11 in appendix C. Even though the gene amount in this gene set J was small, namely eight genes, many GO molecular functions, such as different enzyme activities and

molecular bindings emerged. Not to mention GO biological processes, for example nuclear export and positive regulation of endothelial cell differentiation and negative regulation of insulin-like growth factor receptor signalling pathway, which also appeared alongside with KEGG pathways (e.g. RNA degradation). It is interesting to see such phenomena as negative regulation of cell proliferation and regulation of apoptotic process for these genes; notwithstanding, due to the small gene set size it is not indispensably feasible to distinguish a real incorporation of EmRs in this case to the regulation of TNBC. If there should be such a connection, then the insulin-like growth factor receptor signalling pathway should be considered, since it is also found in TNBC (Davison et al., 2011). The enrichment tests revealed that it is down-regulated by some of the genes in the group two. Consequently, if the previously mentioned pathway needs to be up-regulated, then these genes in the group two should as well be down-regulated. Actually, this turns out to be the case for most of the genes in the group two as seen in Table D1 of appendix D, which otherwise contains supplementary MIP results. The corresponding miRNAs should be therefore up-regulated. This could be the case, while some of the miRNAs associated to the regulation of these genes, such as *ATXN1* in this group two, were also later observed to have a relatively good correlation (0.55) in MIP to explain the down-regulation of *ATXN1*, as can be noted from Table D4 in appendix D. *ATXN1* gene is also the most correlated gene in MIP for TNBC samples as seen in Table 7.

The hierarchical cluster group three, comprising the gene sets L and M, had more genes, tot. 69, than in the previous case, tot. 9. The gene set L, with down-regulated genes that were targets of up-regulated EmRs, indicated many GO biological processes by the tests as listed at Table C12 in appendix C. If these EmRs only degraded these genes then they would be participating in processes, such as neural crest cell development, maternal process involved in parturition, and positive regulation of odontogenesis in TNBC. The activities in neural crest cell development could mean that these EmRs are a part of EMT, and subsequently also in the initiation of the metastasis of cancer progression (Klein 2008). Finally, the tests for the gene set M, consisting of up-regulated genes targeted by down-regulated EmRs, yielded quite clear results, due to

sample size, as seen in Tables C13-C15 of appendix C. There were precisely one GO molecular function and one KEGG pathway: histone-lysine N-methyltransferase activity, and lysine degradation, respectively. This clearly expresses that up-regulated EmRs do not have such a high role in the regulation of the genes in group three. There were also many GO biological processes described for these genes, e.g. histone lysine methylation and negative regulation of tyrosine phosphorylation of Stat3 protein. However, the Q values, besides the most relevant process, were close to 0.05. In contrast, the odds ratios were high, and there was just one different significant gene per every GO biological process that GeneCodis internet tool had taken into account, as can be found from Table C14 in appendix C. This demonstrated that the remaining values in the table for these GO biological processes could be biased.

3.4 Results of mixed integer programming models

3.4.1 Primary analysed groups of genes

The relevant linear programming modelling results for different sets of genes, mainly in the cooperation list, were obtained by restricting the amount of miRNAs as input for the model, and also by different sample subsets and cross validation. In real biological processes, for example due to steric hindrances and requirements for performances near the optimal energy balances, the amount of miRNAs affecting one gene cannot be as vast as the predicted amount of miRNAs affecting this gene. Therefore, linear programming as such was not as relevant as MIP when modelling the regulation of TNCB cells with miRNAs. MIP provided a way for limiting the amount of miRNAs applied in the linear programming model. Accordingly, the amount of miRNAs was limited to twenty two and their expression signals extracted from TNBC samples for MIP evaluation (Table 7). The amount of twenty two miRNAs was derived from 25% quantile of miRNA amount in the cooperation list. This enabled the examination of the role of miRNAs to the regulation of the target genes, or that is to say mRNAs, in TNBC. MIP evaluation produced sensible correlations of the real gene signal values to their miRNA estimates. *ATXN1* was the most well correlated gene to its estimate with a Pearson correlation coefficient (PCC) of 0.817. Ataxin 1 protein (ATXN1) is a component of notch signalling pathway

(Tong et al., 2011), which is a very major pathway in cell-to-cell communication and controlling cell differentiation processes (Dontu et al., 2004).

Table 7. Results of MIP, restricting the amount of miRNAs to 22, with the cooperation list of genes using TNBC samples.

No[1]	Gene symbol	PCC	miRNA freq. input to model	EmR freq. input to model	miRNA freq. used by model
1	*ATXN1*	0.817	63	7	22
2	*QKI*	0.777	100	7	22
3	*TP53INP1*	0.712	42	4	17
4	*SPRED1*	0.656	45	5	20
5	*ZFAND5*	0.611	33	3	16
6	*HBEGF*	0.606	16	2	12
7	*LUZP1*	0.587	39	4	18
8	*SALL1*	0.549	38	3	17
9	*SOCS5*	0.535	20	3	10
10	*ZC3H12C*	0.533	28	2	12

1) The top 10 of all 247 genes are shown according to the PCCs in this list.

Subsequently, it could be important for the regulation of TNBC to affect this gene with miRNAs. MIP was initially also tested for investigating the regulative relationships behind all EmRs and all possible target genes in TNBC as seen in appendix D's Table D2. The target genes were later specified to differentially expressed ones, to comprise the cooperation list, and analysed with all samples, as seen in Table 8. Nonetheless, this former mentioned initial result showed the gene of CUGBP (CUG binding protein), Elav-like Family Member 2 protein (*CELF2*) as a special gene in miRNA-mediated regulation in TNBC with 0.79 PCC to its estimate. CUGBP (CUG binding protein), Elav-like Family Member 2 protein (CELF2) takes part in many posttranscriptional events. It inhibits the apoptosis of breast cancer cells (Mukhopadhyay et al., 2003). Another important gene resulted from this initial test (Table D2 in Appendix D), was a gene for BTB and CNC (Carney complex) Homology 1, Basic Leucine Zipper Transcription Factor 2 (*BACH2*), where BTB is a similar protein motif found from Broad-Complex (BR-C), Tramtrack (ttk), and Bric à Brac (bab) proteins. This gene has been identified as an important regulator in similar cancer as breast cancer due to mutations in *BRCA*s, namely the ovarian cancer, as experimented by Motamed-Khorasani et al., (2007).

Table 8. Results of MIP (restricting the miRNAs to 22) with the cooperation list of genes using all samples.

No[1]	Gene symbol	PCC	miRNA freq. input to model	EmR freq. input to model	miRNA freq. used by model
1	*QKI*	0.750	100	7	20
2	*LUZP1*	0.667	39	4	20
3	*PHACTR2*	0.593	39	6	17
4	*SOCS5*	0.593	20	3	11
5	*FBN1*	0.587	18	2	12
6	*ZC3H12C*	0.576	28	2	18
7	*SALL1*	0.573	38	3	20
8	*ZFAND5*	0.566	33	3	20
9	*KCNJ2*	0.553	28	2	13
10	*ADAMTS9*	0.548	6	3	5

1) The top 10 of all 247 genes are shown according to the PCCs in this list.

Noticeable in MIP results shown in Table 8 is that the *ATXN1* was not the top correlated gene, to its estimate by miRNAs, rather than *QKI*, and the order of the genes was different compared to Table 7. As an example, the negative β miRNA modelling coefficients obtained from MIP for *QKI* are given in appendix D's Table D3. The differences between the correlation coefficients of common genes and miRNA amounts applied by the model at Table 8 and appendix D's Table D2, are due to different MIP Gap parameters, 0.05 and 0.01 respectively. One could note from these results that there were not that many genes in this cooperation list that yielded correlations above 0.5; precisely 25 genes. This could mean that miRNAs act specifically in some cases, such as with *QKI* and a gene of Leucine Zipper Protein 1 (*LUZP1*), to their gene targets, but not in the cellular level synergistically as a bulk system for many targets, as would have been expected when using all the samples. In earlier experiments by Motamed-Khorasani et al., (2007) it has been observed that the BTB and CNC Homology 1, Basic Leucine Zipper Transcription Factor 2 (BACH2) was up-regulated in ovarian cancer, which also has mutations in BCRA, similarly as in breast cancers. None of these most correlated genes to their miRNA estimates were common tumour suppressor genes, such as *BRCA1*, but may act as enzyme inhibitors, such as Phosphatase and Actin Regulator (PHACTR2). These results were then checked by employing different sample subsets; all samples, (MIP model trained with TNBC and normal samples and validated with)

luminal A samples, and random samples, causing different correlations distributions in MIP (Figure 14).

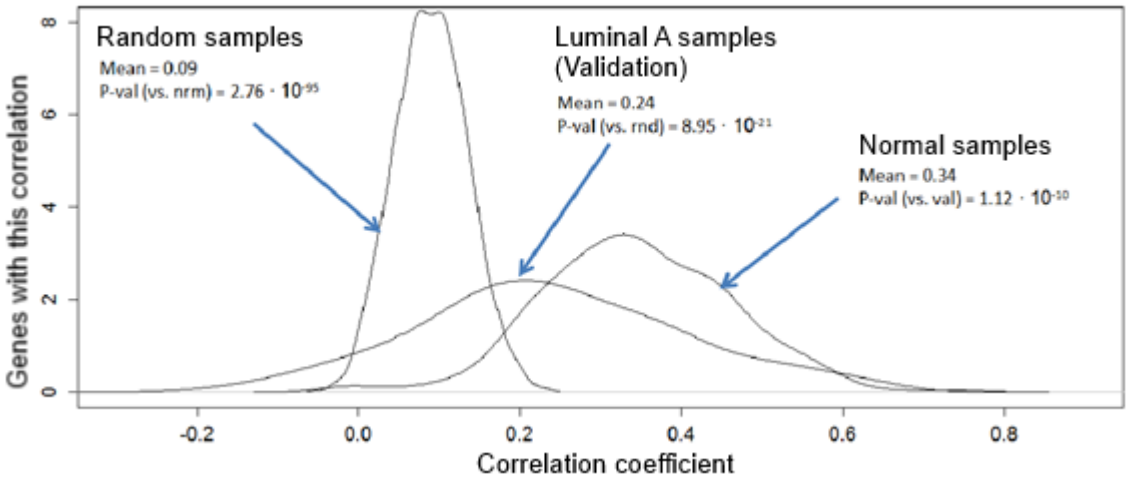

Figure 14. The correlation coefficient distributions of the three different sample types using the genes in the cooperation list with limiting the miRNA input to 22 for the MIP model. The random sample correlations were compared to normal ones (P value = 2.76 · 10-95), luminal A correlations were compared to random ones (P value = 8.95 · 10-21), and normal sample correlations were also compared to luminal A equivalents (P value = 1.12 · 10-10).

The effect of restricting the amount of miRNAs and cross validation to MIP was investigated by inputting different amounts of miRNAs and using five-fold cross validation. This type of validation was executed in order to reduce the overfitting and consequently to model the gene expression properly. Overfitting originates from employing all, or almost all of, the miRNA expression values. Supposedly, this tested the model, so as to see if it works coherently for explaining the results of a smaller subset or a completely different sample source. The result of this was that the correlations with certain miRNA amounts reached an optimum point in MIP with cross validation, after which they started to decrease.

According to earlier *in silico* experiments in the Heidelberg University (Schacht 2013), these correlations would be worse for most of the cases, compared to those, where the cross validation had not been performed. This phenomenon was studied with the cooperation list and some of the observations are collected at Table 9. These results showed the predicted outcome as described above, that there was an optimal point or a plateau, which was reached after a certain amount of miRNAs (Table 9).

Table 9. The cross validation results for the top five genes according to MIP model using all samples and the cooperation list.

Gene symbol / PCCs[1]	3 mRs[2]	5 mRs	7 mRs	10 mRs	12 mRs
QKI	0.54	0.53	0.68	0.62	0.60
ADAMTS9	0.52	0.52	0.54	0.56	0.57
SOCS5	0.46	0.50	0.53	0.56	0.55
LUZP1	0.44	0.49	0.53	0.54	0.54
HBEGF	0.43	0.49	0.52	0.53	0.53

1) The PCC values are mean of all sample groups. 2) The results show five different miRNA restriction amounts (3, 5, 7, 10, and 12).

MIP should not use more miRNAs as this optimum point for producing the most accurate predictions. In most of the cases, it was possible to notice an optimum correlation value before the last miRNA amount studied, such as for *QKI* (0.68 in 7 mRs) and a gene of Suppressor of Cytokine Signalling protein 5 (*SOCS5*; 0.56 in 10 mRs). One can also notice some of the correlations reaching the plateau before the last miRNA amount, such as with *LUZP1* (0.54 in 10 mRs) and a gene of Heparin-Binding EGF-Like Growth Factor (*HBEGF*; 0.53 in 10 mRs). This result tells that it is not required to use all miRNAs in the cross validation models, as noticed in the earlier research (Schacht 2013). Otherwise the result, of the predicted miRNAs that regulate the genes according to model, would be the same as the induced miRNAs from the target gene list *per se*.

The correlations started to drop a bit after an appropriate increase of the frequency of miRNAs inserted to the model, such as with the gene of Sprouty-Related, EVH1 Domain-Containing Protein 1 (*SPRED1*), where EVH1 is an Enabled / Vasodilator-stimulated Phosphoprotein (VASP) Homology 1, in the cooperation list, as seen in Figure 15. This gene was selected, because it showed a clear peak in the PCC versus miRNA frequency image. This meant that there could be an optimal amount of miRNAs for every given gene in MIP. The amount is not constant for all of the genes, and should be evaluated case-by-case. In the case of *SPRED1*, this limit was around 20 miRNAs. It would be illuminating to find the exact role of this gene in breast cancer, since it is important in leukemia (Olsson et al., 2014).

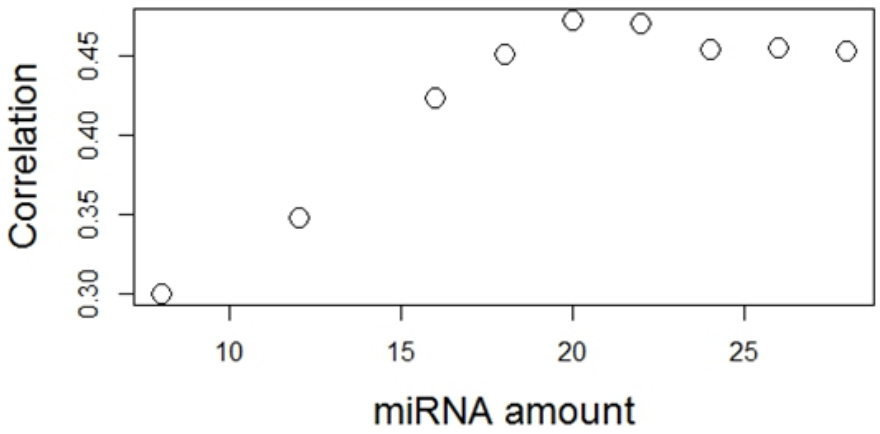

Figure 15. The cross validation MIP results for *SPRED1*. This gene had high correlation after the first modelling without cross validation, and using just a constant miRNA amount. The model reached its optimum plateau at around 20 miRNAs.

3.4.2 Genes in the urea cycle

A list, where the relevant urea cycle genes (according to the preliminary predictions), was collected and analysed as a comparison for the cooperation list. MIP results, denoting the correlations of real values to their estimates, for these genes showed that miRNAs did not explain the regulation of the urea cycle genes as can be noticed from Tables D5 and D6 and Figure D1 in appendix D. The only exception to this could come from the gene of a Fatty Acid Synthase (*FASN*). The model correlated its signals with 0.551 PCC to its estimates derived from signals of miRNAs (Table D6 in Appendix D), such as the most down-regulated EmR hsa-miR-186-5p. This is unusual, although not uncommon (Place et al., 2007), since miRNAs should be regularly up-regulated. The reason for this is that they could then efficiently down-regulate their target genes, unless the mechanism of making this miRNA would be hindered in some other TNBC related fashion. It has been shown that silencing of *FASN* leads to down-regulation of a gene of Tyrosine-Protein Kinase Erbb-2 Receptor (*HER2/neu*) and then to apoptosis in breast cancers (Menendez et al., 2004).

3.4.3 Inferring modelling results with real expression patterns

The real expression patterns of *QKI*, *LUZP1*, and also *ATXN1* correlated very well to their estimated patterns constituted by MIP that was modelled with real miRNA patterns (Table 7, Table 8). These real expression patterns were investigated further (Figure 16-Figure 18) with some of the miRNAs that MIP predicted as good explainers of the real signals of these genes. In some cases,

there were many miRNAs employed by MIP, e.g. twenty two for *ATXN1*. Therefore, only selected miRNAs, which possibly regulate these genes, are discussed. This selection of miRNAs can be utilized for inferring biological relevance and for evaluating the modelling approach. Especially the highly dysregulated miRNAs regulating these genes were supported by MIP, resulting in non-zero β coefficients, while the others were not necessary so. The discourse of this investigation is started from the miRNAs that had the best evidence to be the regulators of specific genes according to MIP, such as seen in Figure 16.

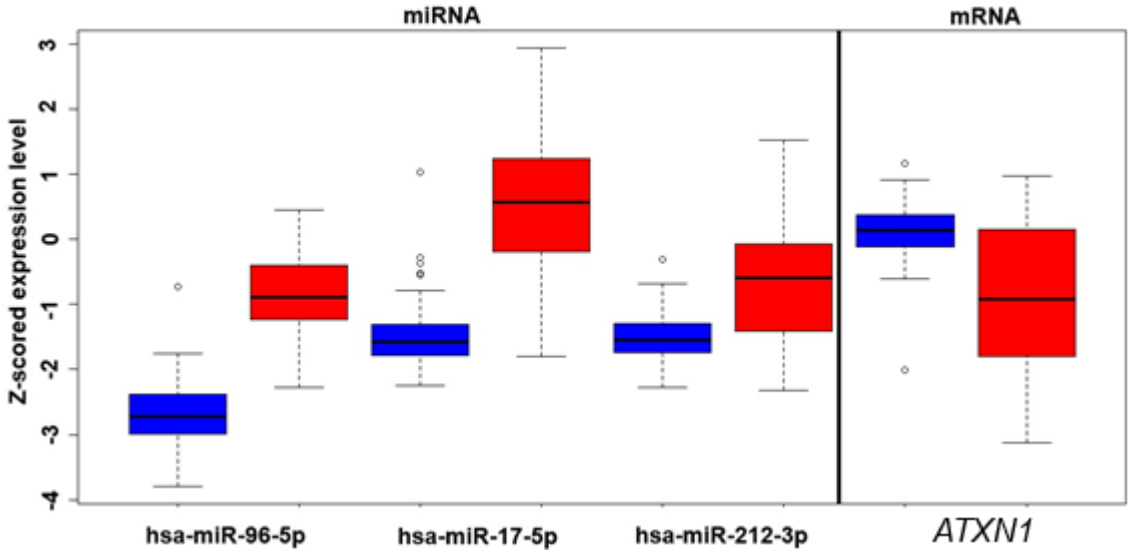

Figure 16. Boxplots for the expression signal patterns of three most up-regulated non-EmRs (hsa-miR-96-5p, hsa-miR-17-5p, hsa-miR-212-3p) in TNBC regulating *ATXN1* according to MIP. This gene had the best correlation in MIP using TNBC samples (Table 7). *ATXN1* is up-regulated in normal samples (blue), where miRNAs are down-regulated. In tumour samples *ATXN1* is down-regulated (red), where miRNAs are up-regulated.

In this case the best miRNA was the most up-regulated miRNA, hsa-miR-96-5p. This miRNA could have a synergistic role in the regulation of many genes. According to Lin et al., (2010), this miRNA induces cell proliferation by down-regulating a gene of transcriptional factor called Forkhead Box O3A *FOXO3a*, in breast cancer. It was also regulating the most down-regulated gene *ATXN1* (Figure 16), which has a high correlation to its miRNA estimate according to MIP with TNBC samples (Table 7).

Moreover, hsa-miR-96-5p was also observed as a part of the regulation of *LUZP1* (Figure 17), and has been experimentally validated to be a part of miRNA coordinated regulation of FOXO1 in breast cancers (Guttilla and White

2009). The second evaluated miRNA to regulate *ATXN1* (Figure 16) was hsa-miR-17-5p. This miRNA has also been discovered in earlier experiments to be associated to breast cancer progression, as regulating its cell proliferation by inhibiting the translation of a gene of Nuclear Receptor Co-Activator 3 (NCOA3), that is to say a gene of Amplified in Breast 1 protein (AIB1; Hossain et al., 2006).

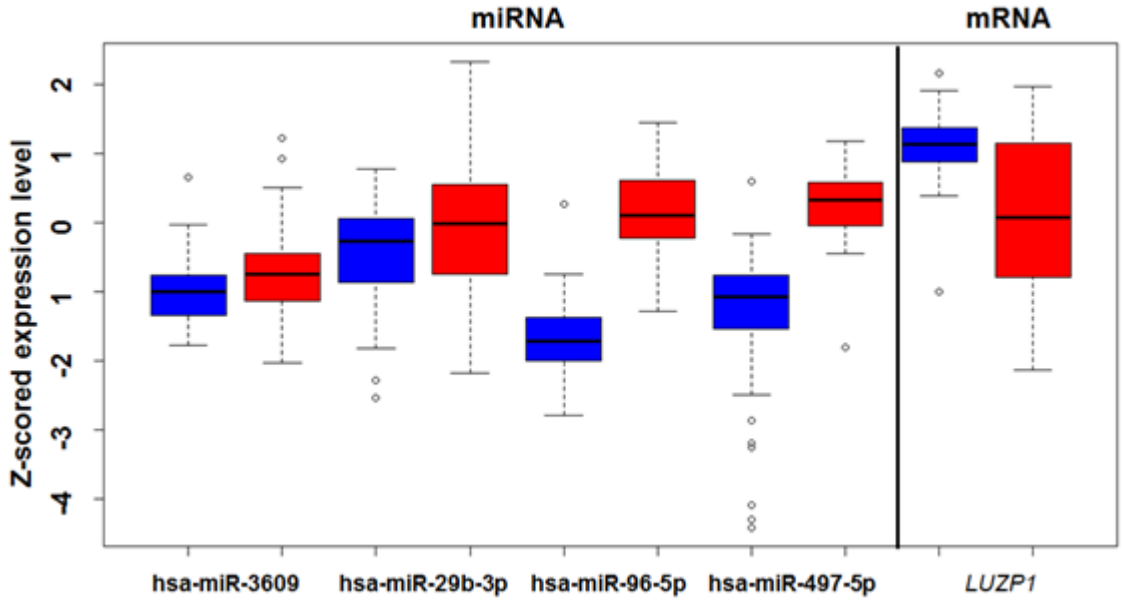

Figure 17. Boxplots for the expression signal patterns of four dysregulated miRNAs (hsa-miR-3609 (EmR), hsa-miR-29b-3p (EmR), hsa-miR-96-5p and hsa-miR-497-5p) in TNBC possibly regulating *LUZP1* according to MIP. The gene had the second best correlation in MIP with all samples (Table 8). *LUZP1* is up-regulated in normal samples (blue), where miRNAs are down-regulated. In tumour samples (red) *LUZP1* is down-regulated, and miRNAs are up-regulated.

Furthermore, hsa-miR-17-5p was confirmed to be tumour suppressor miRNA in earlier research (Fu et al., 2011; Table 2). The last evaluated miRNA (Figure 16) for this most correlated gene (*ATXN1*) according to MIP was hsa-miR-212-3p. Moreover probably regulating *ATXN1*, miR-212 has been observed to target the gene of Protein Patched Homolog 1 (*PTCH1*). Protein Patched Homolog 1 (PTCH1) is a receptor in the hedgehog pathway, especially found in non-small cell lung cancer. (Li et al., 2012a.)

There was also an example case (Figure 17), where the two most up-regulated EmRs (hsa-miR-3609, hsa-miR-29b-3p) and the two most up-regulated other miRNAs (hsa-miR-96-5p and hsa-miR-497-5p) were estimated to down-regulate a gene (*LUZP1*) according to MIP. Hsa-miR-3609 has been noticed to associate with Luminal B, subtype HER2+, type of breast cancer (Li et al., 2012b), but

hsa-miR-29b-3p could also operate in other types of tissues (Tuschen 2013), and in breast cancer subgroup Luminal B, subtype Ki67+ that requires Antigen KI-67 for its growth (Li et al., 2012b). Hsa-miR-29b-3p has also been assumed for regulating *QKI* gene (Figure 18).

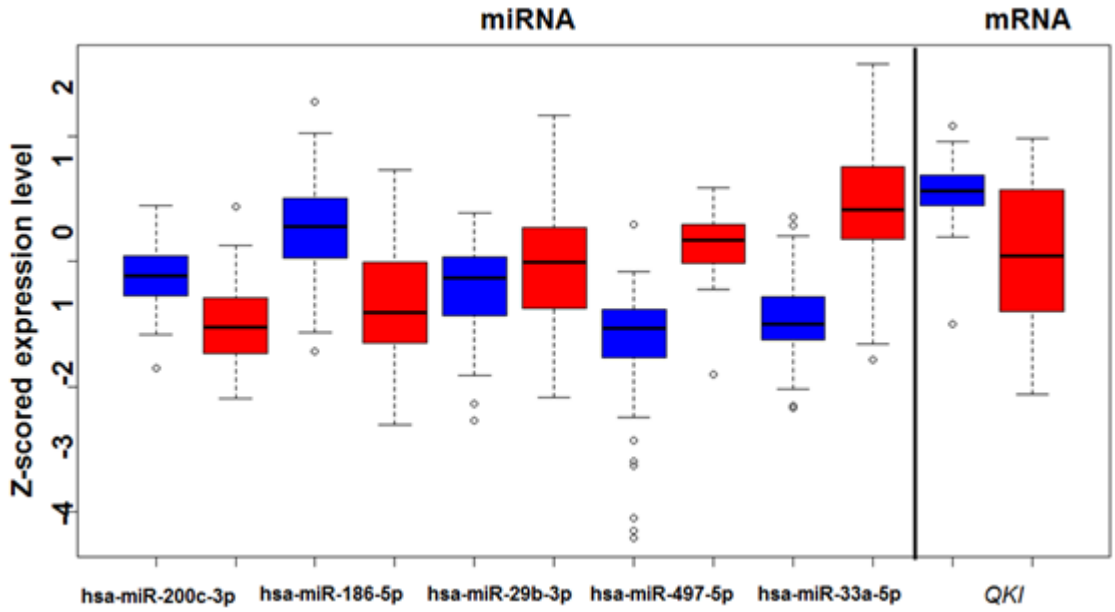

Figure 18. Boxplots for the expression signal patterns of five dysregulated miRNAs (hsa-miR-200c-3p (EmR), hsa-miR-186-5p (EmR), hsa-miR-29b-3p (EmR), hsa-miR-497-5p, hsa-miR-33a-5p) in TNBC and their possibly regulated, and well correlated gene (*QKI*) according to MIP using all samples (Table 8). *QKI* is up-regulated in normal samples (blue), where some of miRNAs are up-regulated, and vice versa for the other cases in tumour samples (red).

Evidently, the two most up-regulated non-EmRs (hsa-miR-96-5p and hsa-miR-497-5p) had a higher difference in expression signals in TNBC than EmRs (Figure 17). This could be the result of cooperativeness TNBC related miRNAs. It was again interesting to notice hsa-miR-497-5p as one of the pre-dicted regulators of *LUZP1* for the reason that this miRNA was also estimated to regulate *QKI* (Figure 18) albeit in a different fashion. Unsuitably though, hsa-miR-497-5p did not yield non-zero β coefficient, yet the others did; such as hsa-miR-96-5p with -0.15, which is slightly lower than the typical β coefficient from 19 β coefficients with median value of -0.20, and maximum of -0.66. This demonstrated that the restriction of 22 miRNAs was not necessarily applicable *LUZP1*, which already had a small amount of predicted affecting miRNAs, i.e. 33 miRNAs.

Even though literature (Zearo et al., 2014; Park et al., 2008; Lee et al., 2013; Li et al., 2012b), as will be soon explained in more details, gives indications that five miRNAs (hsa-miR-200c-3p, hsa-miR-186-5p, hsa-miR-29b-3p, hsa-miR-497-5p, hsa-miR-33a-5p) might be part of regulation of breast cancer, nevertheless they are not regulating a particular DEG (*QKI*) in TNBC according to the results of MIP (Figure 18). For example, the above mentioned basic degrading behaviour of miRNAs was not visible in boxplots for hsa-miR-200c-3p and hsa-miR-186-5p, since they were down-regulated in TNBC, and therefore these miRNAs did not degrade *QKI* (Figure 18). As an expected consequence, EmRs hsa-miR-200c-3p and hsa-miR-186-5p did not yield non-zero β coefficients in MIP. This could point out that these miRNAs have other roles than down-regulation of particular genes (Kolacinska et al., 2014). For example, hsa-miR-186-5p has been discovered to be one of the most DEmR in breast cancer serum as such (Zearo et al., 2014). The regulative role of hsa-miR-200c-3p has been linked to EMT as mentioned in the enrichment analysis (Park et al., 2008). Furthermore, if these down-regulated EmRs in TNBC would be in higher level, then some normal cellular functions would work properly, and TNBC would not survive. On the contrary, it is more likely that the model predicted correctly that these miRNAs did not associate with this gene, due to their non-participation in the modelling result. Consequently, omitting these nonregulating miRNAs from the model would render the results of the modelling correspond more to real biological scenarios. However, the down-regulation of these EmRs, and some others miRNAs, could implicate an increased activity of some genes needed for tumour proliferation.

Initially, it also appeared that there was one up-regulated EmR, hsa-miR-29b-3p in Figure 18, down-regulating *QKI*. Yet this was not the case, since it yielded zero valued β coefficients and the Z-score difference of TNBC compared to normal was 0.22. This was a low value seeing that the most highly up-regulated miRNA had the Z-score difference of 4.3, and the median of all 55 up-regulated miRNAs was 0.9.

Although it seemed counterintuitive, the last pair of highly up-regulated miRNAs in TNBC (hsa-miR-497-5p and hsa-miR-33a-5p) was again according to MIP

probably not associated in down-regulation of *QKI* (Figure 18). Both of these miRNAs had β coefficients with value of zero. Nonetheless, the expression patterns of these miRNAs were clearly elevated in TNBC compared to *QKI*'s opposite patterns. This could denote that the modelling mentioned in literature (Lee et al., 2013) for hsa-miR-497-5p as a part of a breast cancer associated network would not necessary be applicable to TNBC case. On the other hand, hsa-miR-33a-5p has been experimentally validated to be associated to TNBC (Li et al., 2012b), implicating that the limit of 22 miRNAs in MIP modelling is not indispensably sufficient enough. A regulative liaison between *QKI* and hsa-miR-33a-5p, regardless of the size of the miRNA limit, could not be definitely established by these results. This miRNA has not also been mentioned in TNBC experiments for this gene (Li et al., 2012b). The most relevant results in MIP yielded when TNBC samples were applied, for example in the case of *ATXN1* (Figure 16), although implementing the method to all samples gave some valuable insights for the mechanisms of MIP and breast cancer, as can be noticed especially when evaluating *LUZP1* and *QKI* (Figure 17 and Figure 18).

4 Conclusions

During the analysis of this work the miRNAs were confirmed to be a part of the regulation of the genes of TNBC, e.g. hsa-miR-29b-3p for *LUZP1*, and hsa-miR-212-3p for *ATXN1*. The hsa-miR-29 miRNA family has earlier been observed to be cardinal in the regulation of cancer pathways (Jacobsen et al., 2013). Consequently, this miRNA could be used as a potential preliminary biomarker indicating cancer. Notwithstanding, other miRNA biomarkers should also be considered if the connection to TNBC is to be established. For instance, the 21 EmRs discovered in this analysis did not work as a coherent group that would have induced regulation in all of the pathways. They rather affected as certain combinations (enriched, and/or normal miRNAs) to limited frequency of genes of specific pathways, such as to a gene called *ATXN1* (Table 7, Figure 16), which is a component of an important notch signalling pathway in cancer. The enrichment analyses for special sets of genes revealed that miRNAs have putatively other roles, such as regulators of insulin growth factor signalling pathway and in the initiation of the metastasis of cancer. These genes were sometimes clustered according to their EmRs in the clustering analysis.

The most illustrating example findings (*ATXN1*, *QKI*, and *LUZP1*) were shown with MIP modelling. It turned out to be an efficient tool to analyse the regulative incorporations behind miRNAs and genes. In spite of this fact, the amount of regulating miRNAs employed in the model for predicating the expression of particular genes must be assessed case-by-case. The reason for this is that some of the target genes might have a small amount of predicted affecting miRNAs, as observed during analyses for LUZP1 gene (Figure 17). This demonstrates that the limits should be sometimes lowered or discarded. The modelling approach also predicted quite efficiently, which miRNAs were not relevant for the regulation of genes (Figure 18). In those cases, the β coef-ficients were mostly zero. It is clear that miRNAs affect many genes in TNBC, for example the genes in the cooperation list and its subgroups, nevertheless the synergistic role of these miRNAs to the overall regulation, for example in the oncogenesis, proliferation, and the metastasis of TNBC, is elusive.

Indeed, the genes of the urea cycle were implied to be associated with TNBC according to the enrichment analysis. On the contrary to this analysis, MIP showed that the miRNAs were not greatly involved to the regulation of this pathway, except perhaps *FASN* with 0.551 PCC to its regulating miRNAs, such as the most down-regulated EmR hsa-miR-186-5p. In later studies, it would be interesting to check MIP results with dynamic models for the reason that they have been described models with close equivalency to biological scenarios (Lai et al., 2012). Similarly, it could be illuminating to add more variables to MIP, besides miRNAs, such as transcription factors, histone modifications, and also the distance dependency of miRNAs to the degradation of targets. Although the model is still incomplete, it could be said with a reasonable caution, that some of the miRNAs mentioned in this analysis could not be omitted without hindering the key tasks of TNBC. This work demonstrated that biologically relevant results, depicting the regulative role of miRNAs in TNBC, can be obtained by using enrichment tests and MIP models.

References

Agresti, A. (2002) Categorical data analysis. 2nd edition, p. 91–101. Wiley, New York.

Ahlblom, A. (2003) Biostatistics for epidemiologists, p. 80–83. CRC Press, Boca Raton.

Ambros, V., Bartel, B., Bartel, D.P., Burge, C.B., Carrington, J.C., Chen, X., Dreyfuss, G., Eddy, S.R., Griffiths-Jones, S., Marshall, M., Matzke, M., Ruvkun, G. & Tuschl, T. (2003) A uniform system for microRNA annotation. *RNA*. **9**:277–279.

Anad, P., Kunnumakkara., A.B., Sundaram, C., Harikumar, K.B., Tharakan, S.T., Lai, O.S., Sung, B. & Aggarwal, B.B. (2008) Cancer is a preventable disease that requires major lifestyle changes. *Pharm. Res.* **25**:2097–2116.

Artmann, S., Jung, K., Bleckmann, A. & Beissbarth, T. (2012) Detection of simultaneous group effects in microRNA expression and related target gene sets. *PLoS One.* **7**:38365-38375.

Auer, P.L. & Doerge, R.W. (2010) Statistical design and analysis of RNA sequencing data. *Genetics*. **185**:405–416.

Bartel, D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell.* **116**:281–297.

Bartel, D.P. (2009) MicroRNAs: target recognition and regulatory functions. *Cell.* **136**:215–233.

Baylin, S.B. & Ohm, J.E. (2006) Epigenetic gene silencing in cancer - a mechanism for early oncogenic pathway addiction? *Nat. Rev. Cancer.* **6**:107–216.

Blenkiron, C., Goldstein, L.D., Thorne, N.P., Spiteri, I., Chin, S.F., Dunning, M.J., Barbosa-Morais, N.L., Teschendorff, A.E., Green, A.R., Ellis, I.O., Tavaré, S., Caldas, C. & Miska, E.A. (2007) MicroRNA expression profiling of human breast cancer identifies new markers of tumour subtype. *Genome Biol.* **8**:214-229.

Broad institute's genome data analysis center (GDAC). 2014. [WWW-document]. https://confluence.broadinstitute.org/display/GDAC/Documentation (Read 15.1.2014).

Carmona-Saez, P., Chagoyen, M., Tirado, F., Carazo, J.M. & Pascual-Montano, A. (2007) GENECODIS: A web-based tool for finding significant concurrent annotations in gene lists. *Genome Biol.* **8**:3–13.

Carthew, R.W. (2006) Gene regulation by microRNAs. *Curr. Opin. Genet. Dev.* **16**:203–208.

Chan, E., Prado, D.E. & Weidhaas, J.B. (2011) Cancer microRNAs: from subtype profiling to predictors of response to therapy. *Trends Mol. Med*. **17**:235–243.

Chen, J., Wang, G., Lu, C., Guo, X., Hong, W., Kang, J. & Wang, J. (2012) Synergetic cooperation of microRNAs with transcription factors in iPS cell generation. *PLoS One.* **7**:40849–40862.

Chen, K. & Rajewsky, N. (2007) The evolution of gene regulation by transcription factors and microRNAs. *Nat. Rev. Genet.* **8**:93–103.

Cheng, C. & Gerstein, M. (2012) Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells. *Nucleic Acids Res.* **40**:553–568.

Cheng, C., Alexander, R., Min, R., Leng, J.K., Yip, Y., Rozowsky, J., Yan, K.K. Dong, X. Djebali, S. Ruan, Y. Davis, C.A. Carninci, P.Lassman, T., GingeRas, T.R.,Guigó, R., Birney, E., Weng, Z., Snyder, M. & Gerstein, M. (2012) Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome Res.* **22**:1658–1667.

Cleator, S., Heller, W. & Coombes, R.C. (2007) Triple-negative breast cancer: therapeutic options. *Lancet Oncol.* **8**:235–244.

Dakubo, G.D. (2010) Mitochondrial genetics and cancer, p. 42. Springer-Verlag, Berlin.

Davis, F.M., Azimi, I., Faville, R.A., Peters, A.A., Jalink, K., Putney, J.W., Goodhill, G.J., Thompson, E.W., Roberts-Thomson, S.J. & Monteith, G.R. (2013) Induction of epithelial-mesenchymal transition (EMT) in breast cancer cells is calcium signal dependent. *Oncogene.* doi: 10.1038/onc.2013.187. [Epub ahead of print]

Davison, Z, de Blacquière, G.E., Westley, B.R. & May, F.E. (2011) Insulin-like growth factor-dependent proliferation and survival of triple-negative breast cancer cells: implications for therapy. *Neoplasia.* **13**:504–515.

Denli, A.M., Tops, B.B., Plasterk, R.H., Ketting, R.F. & Hannon, G.J. (2004) Processing of primary microRNAs by the microprocessor complex. *Nature.* **432**:231–235.

Dong, F., Ji, Z.B., Chen, C.X., Wang, G.Z. & Wang, J.M. (2013) Target gene and function prediction of differentially expressed microRNAs in lactating mammary glands of dairy goats. *Int. J. Genomics*. **2013**:917342–917354.

Dontu, G., Jackson, K.W., McNicholas, E., Kawamura, M.J., Abdallah, W.M. & Wicha, M.S. (2004) Role of notch signalling in cell–fate determination of human mammary stem/progenitor cells. *Breast Cancer Res.* **6**:605–615.

Dranoff, G. (2011) Cancer immunology and immunotherapy, p. 2–4, Springer-Verlag, Berlin.

Du, T. & Zamore, P.D. (2005) microPrimer: the biogenesis and function of microRNA. *Development*. **132**:4645–4652.

Dvinge, H., Git, A., Gräf, S., Salmon-Divon, M., Curtis, C., Sottoriva, A., Zhao, Y., Hirst, M., Armisen, J., Miska, E.A., Chin, S.F., Provenzano, E., TuRashvili, G., Green, A., Ellis, I., Aparicio, S. & Caldas, C. (2013) The shaping and functional consequences of the microRNA landscape in breast cancer. *Nature.* **497**:378–382.

Elbashir, S., Harborth, J., Lendeckel, W., Yalcin, A., Weber, K & Tuschl T (2001) Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature.* **411**:494–498.

Esquela-Kerscher, A. & Slack, F.J. (2006) Oncomirs — microRNAs with a role in cancer. *Nat. Rev. Cancer.* **6:**259–269.

Fazi, F. & Nervi, C. (2008) MicroRNA: basic mechanisms and transcriptional regulatory networks for cell fate determination. *Cardiovasc. Res.* **79**:553–561.

Florescu, A., Amir, E., Bouganim, N., Clemons, M. (2011) Immune therapy for breast cancer in 2010 — hype or hope? *Curr. Oncol.* **18**:9–18.

Fu, S.W., Chen, L. & Man, Y.G. (2011) miRNA biomarkers in breast cancer detection and management. *J. Cancer.* **2**:116–122.

Gibbs, J.B. (2000) Mechanism-based target identification and drug discovery in cancer research. *Science.* **287**:1969–1972.

Griffiths-Jones, S., Grocock, R.J., van Dongen, S., Bateman, A. & Enrigh,t A.J. (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.* **34**:140–144.

Griffiths-Jones, S., Saini, H.K., van Dongen, S. & Enright, A.J. (2008) miRBase: tools for microRNA genomics. *Nucleic Acids Res.* **36**:154–158.

Gupta, A., Nagilla, P., Le H.S., Bunney, C., Zych, C., Thalamuthu, A., Bar-Joseph, Z., Mathavan, S. & Ayyavoo, V. (2011) Comparative expression profile of miRNA and mRNA in primary peripheral blood mononuclear cells infected with human immunodeficiency virus (HIV-1). *PLoS One.* **6**:22730-22741.

Gurobi optimization, Inc. 2014. Gurobi optimizer reference manual. [WWW-document]. < http://www.gurobi.com >. (Read 15.12.2013).

Guttilla, I.K. & White, B.A (2009) Coordinate regulation of FOXO1 by miR-27a, miR-96, and miR-182 in breast cancer cells. *J. Biol. Chem.* **284**:23204–23216.

Hahne, F., Huber, W., Gentleman, R. & Falcon, S. (2008) Bioconductor case studies, Springer Science plus Business Media, LLC, New York.

Hanahan, D. & Weinberg, R.A. (2000) The hallmarks of cancer. *Cell.* **100**:57–70.

Hanahan, D. & Weinberg, R.A. (2011) Hallmarks of cancer: the next generation. *Cell.* **144**:646–674.

Hartigan, J. A. (1975) Clustering algorithms. Wiley, New York.

Heimberg, A.M., Sempere, L.F., Moy, V.N., Donoghue, P.C. & Peterson, K.J. (2008) MicroRNAs and the advent of vertebrate morphological complexity. *Proc. Natl. Acad. Sci. USA*. **105**:2946–2950.

Herschkowitz, J.I., Simin, K., Weigman, V.J., Mikaelian, I., Usary, J., Hu, Z., Rasmussen, K.E., Jones, L.P., Assefnia, S., Chandrasekharan, S., Backlund, M.G., Yin, Y., Khramtsov, A.I., Bastein, R., Quackenbush, J., Glazer, R.I., Brown, P.H., Green, J.E., Kopelovich, L., Furth, P.A., Palazzo, J.P., Olopade, O.I., Bernard, P.S., Churchill, G.A., Van Dyke, T. & Perou, C.M. (2007) Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumours. *Genome Biol.* **8**:76–92.

Hosein, K.M., Kim, J.W., Bechis, S.K. & Werb, Z. (2008) GATA-3 and the regulation of the mammary luminal cell fate. *Curr. Opin. Cell Biol.* **20**:164–170.

Hosein, K.M., Slorach, E.M., Sternlicht, M.D. & Werb. Z. (2006) GATA-3 maintains the differentiation of the luminal cell fate in the mammary gland. *Cell.* **127**:1041–1055.

Hossain, A., Kuo, M.T. & Saunders, G.F. (2006) Mir-17-5p Regulates Breast Cancer Cell Proliferation by Inhibiting Translation of AIB1 mRNA. *Mol. Cell Biol.* **26**:8191–8201.

Hsu, M.C., Chai, C.Y., Hou, M.F., Chang, H.C., Chen, W.T. & Hung, W.C. (2008). Jab1 is overexpressed in human breast cancer and is a downstream target for HER-2/neu. *Mod. Pathol*. **21**:609–616.

Hu, R., Dawood, S., Holmes, M.D., Collins, L.C., Schnitt, S.J., Cole, K., Marotti, J.D., Hankinson, S.E., Colditz, G.A. & Tamimi, R.M. (2011) Androgen receptor expression and breast cancer survival in postmenopausal women. *Clin. Cancer Res.* **17**:1867–1874.

Hunt, K.K., Robb, G.L., Strom, E.A. & Ueno, N.T. (2008) Breast Cancer, 2nd edition, p. 5–11, 122. Springer, New York.

Iliopoulos, D. (2014) MicroRNA circuits regulate the cancer-inflammation link. *Sci. Signal.* **7**:8-9.

Iorio, M.V., Ferracin, M., Liu, C.G., Veronese, A., Spizzo, R., Sabbioni, S., Magri, E., Pedriali, M., Fabbri, M., Campiglio, M., Menard, S., Palazzo, J.P., Rosenberg, A., Musiani, P., Volinia, S., Nenci, I., Calin, G.A., Querzoli, P., Negrini, M. & Croce, C. M. (2005) MicroRNA gene expression deregulation in human breast cancer. *Cancer Res.* **65**:7065–7070.

Jacobsen, A., Silber, J., Harinath, G., Huse, J.T., Schultz, N. & Sander, C. (2013) Analysis of microRNA-target interactions across diverse cancer types. *Nat. Struct. Mol. Biol.* **20**:1325–1332.

Jain, M., Nilsson, R., Sharma, S., Madhusudhan, N., Kitami, T., Souza, A.L., Kafri, R., Kirschner, M.W. Clish, C.B. & Mootha, V.M. (2012) Metabolite profiling identifies a key role for glycine in rapid cancer cell proliferation. *Science.* **336**:1040–1044.

Jing, Q., Huang, S., Guth, S., Zarubin, T., Motoyama, A., Chen, J., Di Padova, F., Lin S.C., Gram, H. & Han, J. (2005) Involvement of microRNA in AU-rich element-mediated mRNA instability. *Cell.* **120**:623–634.

John, B., Enright, A.J., Aravin, A., Tuschl, T., Sander, C. & Marks, D.S. (2004) Human microRNA targets. *PLoS Biol.* **2**:363-380.

Jordan, V.C. (2006) Tamoxifen (ICI46,474) as a targeted therapy to treat and prevent breast cancer. *Br. J. Pharmacol.* **147**:269–276.

Khorshid, M., Hausser, J., Zavolan, M. & van Nimwegen, E. (2013) A biophysical miRNA-mRNA interaction model infers canonical and noncanonical targets. *Nat. Methods.* **10**:253–260.

Kim, J.H. & Adelstein, R.S. (2011) LPA_1-induced migration requires nonmuscle myosin II light chain phosphorylation in breast cancer cells. *J. Cell Physiol*. **226**:2881–2893.

Kim, S.J., Ha, J.W. & Zhang, B.T. (2013) Constructing higher-order miRNA-mRNA interaction networks in prostate cancer via hypergraph-based learning. *BMC Syst. Biol.* **7**:47–62.

Klein, C.A. (2008) Cancer: the metastasis cascade. *Science.* **321**:1785–1787.

Kolacinska, A., Morawiec, J., Fendler, W., Malachowska, B., Morawiec, Z., Szemraj, J., Pawlowska, Z., Chowdhury, D., Choi, Y.E., Kubiak, R., Pakula, L. & Zawlik, I. (2014) Association of microRNAs and pathologic response to preoperative chemotherapy in triple negative breast cancer: preliminary report. *Mol. Biol. Rep*. doi: 10.1007/s11033-014-3140-7. [Epub ahead of print]

Krek, A, Grün, D., Poy, M.N., Wolf, R., Rosenberg, L., Epstein, E.J., MacMenamin, P., da Piedade, I., Gunsalus, K.C., Stoffel, M. & Rajewsky, N. (2005) Combinatorial microRNA target predictions. *Nat. Genet.* **37**:495–500.

Kyoto enclycopedia of genes and genomes (KEGG). 2014. [WWW-document]. <http://www.genome.jp/kegg/>. (Read 10.3.2014).

Lai, X., Schmitz, U., Gupta, S.K., Bhattacharya, A., Vera, J., Wolkenhauer, O. & Kunz, M. (2012) Computational analysis of target hub gene repression regulated by multiple and cooperative miRNAs. *Nucleic Acids Res.* **40**:8818–8834.

Lee, C.H., Kuo, W.H., Lin, C.C., Oyang, Y.J., Huang, H.C. & Juan, H.F. (2013) MicroRNA-regulated protein-protein interaction networks and their functions in breast cancer. *Int. J. Mol. Sci.* **14**:11560–11606.

Lee, R.C., Feinbaum, R.L. & Ambros, V. (1993) The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell.* **75**:843–854.

Lehmann, B.D., Bauer, A.J., Chen, X., Sanders, M.E., Chakravarthy, A.B., Shyr, Y. & Pietenpol, J.A. (2011) Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J. Clin. Invest.* **121**:2750–2767.

Lewis, B.P., Burge,C.B. & Bartel, D.P. (2005) Conserved seed pairing, often flanked byadenosines, indicates that thousands of human genes are microRNA targets. *Cell*. **120**:15−20.

Li, J.Y., Jia, S., Zhang, W.H., Zhang, Y., Kang, Y. & Li, P.S. (2012b) Differential distribution of microRNAs in breast cancer grouped by clinicopathological subtypes. *Asian Pacific J. Cancer Prev*. **14**:3197–3203.

Li, Y., Zhang, D., Chen, C., Ruan, Z., Li, Y. & Huang, Y. (2012a) miR-212 displays tumour promoting properties in NSCLC cells and targets the hedgehog pathway receptor PTCH1. *Mol. Biol Cell.* **23**:1423–1434.

Lim, L.P., Lau, N.C., Garrett-Engele, P., Grimson, A., Schelter, J.M., Castle, J., Bartel, D.P., Linsley, P.S. & Johnson, J.M. (2005) Microarray analysis shows that some microRNAs down-regulate large numbers of target mRNAs. *Nature.* **433**: 769–773.

Lin, H., Dai, T., Xiong, H., Zhao, X., Chen, X., Yu, C., Li, J., Wang, X. & Song L. (2010) Unregulated miR-96 induces cell proliferation in human breast cancer by downregulating transcriptional factor FOXO3a. *PLoS One*. **5**:15797–15806.

Lin, M.Z., Marzec, K.A., Martin, J.L. & Baxter, R.C. (2014) The role of insulin-like growth factor binding protein-3 in the breast cancer cell response to DNA-damaging agents. *Oncogene.* **33**:85–96.

Liu, H., Dong, H., Robertson, K. & Chen Liu (2011) DNA methylation suppresses expression of the urea cycle enzyme carbamoyl phosphate synthetase 1 (CPS1) in human hepatocellular carcinoma. *Am. J. Pathol.* **178**:652–661.

Locker, G.Y. (1998) Hormonal therapy of breast cancer. *Cancer Treat. Rev.* **24**:221–240.

Lowery, A.J., Miller, N., Devaney, A., McNeill, R.E, Davoren, P.A., Lemetre, C., Benes, V., Schmidt, S., Blake, J., Ball, G. & Kerin, M.J. (2009) MicroRNA signatures predict estrogen receptor, progesterone receptor and HER2/neu receptor status in breast cancer. *Breast Cancer Res.* **11**:27–44.

Luo, W. & Brouwer, C. (2013) Pathview: an R/Bioconductor package for pathway based data integration and visualization. *Bioinformatics.* **29**:1830–1831.

Macarron, R. (2006) Critical review of the role of HTS in drug discovery. *Drug Discov. Today*. **11**:277–279.

MacFarlane, L.A. & Murphy, P.R. (2010) MicroRNA: biogenesis, function and role in cancer. *Curr. Genomics.* **11**:537–561.

Maier, W.F., Stöwe, K. & Sieg, S. (2007) Combinatorial and high-throughput materials science. *Angew. Chem. Int. Ed*. **46**:6016–6067.

Martin, J.L., de Silva, H.C., Lin, M.Z., Scott, C.D. & Baxter, R.C. (2014) Inhibition of insulin-like growth factor-binding protein-3 signalling through sphingosine kinase-1 sensitizes triple-negative breast cancer cells to EGF receptor blockade. *Mol. Cancer Ther.* **13**:316–228.

Martin-Perez, D., Piris, M.A. & Sanchez-Beato, M. (2010) Polycomb proteins in hematologic malignancies. *Blood*. **116**:5465-5475.

Matloff, N. (2011) The art of R programming: a tour of statistical software design, No Starch Press, San Fransisco.

Mattie, M.D., Benz, C.C., Bowers, J., Sensinger, K., Wong, L., Scott, G.K., Fedele, V., Ginzinger, D., Getts, R., & Haqq, C. (2006) Optimized high-throughput microRNA expression profiling provides novel biomarker assessment of clinical prostate and breast cancer biopsies. *Mol. Cancer.* **5**:24-37.

Mavrakis, K.J., Leslie C.S. & Wendel, H.G. (2011) Cooperative control of tumour suppressor genes by a network of oncogenic microRNAs. *Cell Cycle*. **10**:2845–2489.

Menendez, J.A., Vellon, L., Mehmi, I., Oza, B.P., Ropero, S., Colomer, R. & Lupu, R. (2004) Inhibition of fatty acid synthase (FAS) suppresses HER2/neu (erbB-2) oncogene overexpression in cancer cells. *Proc. Natl. Acad. Sci. USA*. **101**:10715–10720.

miRBase. The microRNA database. 2014. [WWW-document]. <http://www.mirbase.org/>. (Read 1.1.2014).

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat. Methods.* **5**:621–628.

Motamed-Khorasani, A., Jurisica, I., Letarte, M., Shaw, P.A., Parkes, R.K., Zhang, X., Evangelou, A., Rosen, B., Murphy, K.J. & Brown, T.J. (2007) Differentially androgen-modulated genes in ovarian epithelial cells from BRCA mutation carriers and control patients predict ovarian cancer survival and disease progression. *Oncogene*. **26**:198–214.

Mukhopadhyay, D., Jung, J., Murmu, N., Houchen, C.W., Dieckgraefe, B.K, & Anant, S. (2003) CUGBP2 plays a critical role in apoptosis of breast cancer cells in response to genotoxic injury. *Ann. NY Acad. Sci.* **1010**:504–509.

Muniategui, A., Pey, J., Planes, F.J., & Rubio, A. (2012) Joint analysis of miRNA and mRNA expression data. *Brief. Bioinform.* **14**:263–278.

Nazarov, P.V., Reinsbach, S.E., Muller, A., Nicot, N., Philippidou, D., Vallar., L., & Kreis, S. (2013) Interplay of microRNAs, transcription factors and target genes: linking dynamic expression changes to function. *Nucleic Acids Res.* **41**: 2817–2831.

Nichols, E.K. (1988) Human gene therapy, p. 9–15, Institute of Medicine, National Academy of Science, Harvard University Press, Massachusetts.

Obernosterer, G., Leuschner, P.J., Alenius, M., & Martinez, J. (2006) Post-transcriptional regulation of microRNA expression. *RNA.* **12**:1161–1167.

Olsson, L., Castor, A., Behrendtz, M., Biloglav, A., Forestier, E., Paulsson, K. & Johansson, B. (2014) Deletions of IKZF1 and SPRED1 are associated with poor prognosis in a population-based series of pediatric B-cell precursor acute lymphoblastic leukemia diagnosed between 1992 and 2011. *Leukemia*. **28**:302–310.

Orban, T.I. & Izaurralde, E. (2005) Decay of mRNAs targeted by RISC requires XRN1, the Ski complex, and the exosome. *RNA* **11**:459–469.

Oshlack, A. & Wakefield, M.J. (2009) Transcript length bias in RNA-seq data confounds systems biology. *Biol. Direct.* **4**:14-23.

Palijan, A., Fernandes, I., Verway, M., Kourelis, M., Bastien, Y., Tavera-Mendoza, L.E., Sacheli, A., Bourdeau, V., Mader, S. & White, J.H. (2009) Ligand-dependent corepressor LCoR is an attenuator of progesterone-regulated gene expression. *J. Biol. Chem.* **284**:30275–30287.

Park, S.M., Gaur, A.B., Lengyel, E. & Peter, M.E. (2008) The miR-200 family determines the epithelial phenotype of cancer cells by targeting the E-cadherin repressors ZEB1 and ZEB2. *Genes Dev.* **22**:894–907.

Pegram, M.D., Gottfried E. Konecny, G.E., O'Callaghan, C., Beryt, M., Pietras, R. & Slamon, D.J. (2004) Rational combinations of Trastuzumab with chemotherapeutic drugs used in the treatment of breast cancer. *J. Natl. Cancer. Inst.* **96:**739–749.

Perou, C.M. (2011) Molecular stratification of triple-negative breast cancers. *Oncologist.* **16**:61–70.

Perou, C.M. , Sorlie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Rees, C.A., Pollack, J.R. Ross, D.T., Johnsen, H. & Akslen, L.A. (2000) Molecular portraits of human breast tumours. *Nature.* **406**:747–752.

Petersen, C.P., Bordeleau, M.E., Pelletier, J., & Sharp, P.A. (2006) Short RNAs repress translation after initiation in mammalian cells. *Mol. Cell.* **21**:533–542.

PicTar. An algorithm for the identification of microRNA targets. 2014. [WWW-document]. <http://pictar.mdc-berlin.de/>. (Read 15.1.2014).

PITA. Probability of interaction by target accessibility. 2014. [WWW-document]. <http://genie.weizmann.ac.il/pubs/mir07/mir07_prediction.html>. (Read 15.1.2014).

Place, R.F., Li, L.C., Pookot, D., Noonan, E.J. & Dahiya, R. (2008) MicroRNA-373 induces expression of genes with complementary promoter sequences. *Proc. Natl. Acad. Sci. USA*. **105**:1608-1613.

Poy, M.N., Eliasson, L., Krutzfeldt, J., Kuwajima, S., Ma, X., Macdonald, P.E., Pfeffer, S., Tuschl, T., Rajewsky, N., Rorsman, P., & Stoffel, M. (2004) A pancreatic islet-specific microRNA regulates insulin secretion. *Nature.* **432**:226–230.

R. The R Project for Statistical Computing. 2013. [WWW-document]. <http://www.r-project.org/> (Read 15.9.2013).

Rajewsky, N. (2006) MicroRNA target predictions in animals. *Nat. Genet.* **38**:8–13.

Romero-Cordoba, S., Rodriguez-Cuevas, S., Rebollar-Vega, R., Quintanar-Jurado, V., Maffuz-Aziz, A., Jimenez-Sanchez, G., Bautista-Piña, V., Arellano-Llamas, R. & Hidalgo-Miranda, A. (2012) Identification and pathway analysis of microRNAs with no previous involvement in breast cancer. *PLoS One.* **7**:31904-31916.

Rossi S. & Loda M. (2003) The role of the ubiquitination–proteasome pathway in breast cancer. Use of mouse models for analyzing ubiquitination processes. *Breast Cancer Res.* **5**:16–22.

Sachdev D. & Yee D. (2007) Disrupting insulin-like growth factor signalling as a potential cancer therapy. *Mol. Cancer Ther.* **6**:1–12.

Salehi, F., Turner, M.C., Phillips, K.P., Wigle, D.T., Krewski, D. & Aronson, K.J., (2008) Review of the etiology of breast cancer with special attention to organochlorines as potential endocrine disruptors. *J. Toxicol. Environ. Health* **11**:276–300.

Sasamoto, M.M., Vu, T.T., Claret, F.X. & Edgerton, M.E. (2012) Functional correlates of Jab1 networks in triple negative breast cancer. *Mod. Pathol.* **25**:450-458.

Schacht, V.T. (2013) Analyzing the regulation of metabolic pathways and MITF regarding melanoma cell lines. Bachelor Thesis, Ruprecht-Karls-Universität Heidelberg, Germany.

Schlatter, P., Gutmann, H. & Drewe, J. (2006) Primary porcine proximal tubular cells as a model for transepithelial drug transport in human kidney. *Eur. J. Pharm. Sci.* **28**:141–154.

Schwappacher, R., Rangaswami, H., Su-Yuo, J., Aaron Hassad, A., Spitler R. & Casteel, D.E. (2013) cGMP-dependent protein kinase Ib regulates breast cancer cell migration and invasion via interaction with the actin/myosin-associated protein caldesmon. *J. Cell Sci.* **126**:1626–1636.

Sempere, L.F., Christensen, M., Silahtaroglu, A., Bak, M., Heath, C.V., Schwartz, G., Wells, W., Kauppinen, S. & Cole, C.N. (2007) Altered microRNA expression confined to specific epithelial cell subpopulations in breast cancer. *Cancer Res.* **67**:11612–11620.

Setty, M., Helmy, K., Khan, A.A., Silber, J., Arvey, A., Neezen, F., Agius, P., Huse, J.T., Holland, E.C. & Leslie, C.S. (2012) Inferring transcriptional and microRNA-mediated regulatory programs in glioblastoma. *Mol. Syst. Biol.* **8**:605–620.

Shenouda, S.K. & Alahari, S.K. (2009) MicroRNA function in cancer: oncogene or a tumor suppressor? *Cancer Metastasis Rev*. **28**:369-378.

Sheth, U. & Parker, R. (2003) Decapping and decay of messenger RNA occur in cytoplasmic processing bodies. *Science*. **300**:805-808.

Sivachenko, A.Y., Yuryev, A., Daraselia, N. & Mazo, I. (2007) Molecular networks in microarray analysis. *J. Bioinform Comput. Biol.* **5**:429–456.

Slamon, D.J., Leyland-Jones, B., Shak, S., Fuchs, H., Paton, V., Bajamonde, A., Fleming, T., Eiermann, W., Wolter, J., Pegram, M., Baselga, J. & Norton, L. (2001) Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. *N. Engl. J. Med.* **344**:783–792.

Summy, J.M. & Gallick, G.E. (2006) Treatment for advanced tumours: SRC reclaims center stage. *Clin. Cancer Res.* **12**:1398–1401.

Sun, K. & Lai, E.C. (2013) Adult-specific functions of animal microRNAs. *Nat. Rev. Genet.* **14**:535–548.

Suzuki, R. & Shimodaira, H (2006) Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics.* **22**:1540–1542.

Tabas-Madrid, D., Nogales-Cadenas, R. & Pascual-Montano, A. (2012) GeneCodis3: a non-redundant and modular enrichment analysis tool for functional genomics. *Nucleic Acids Res*. **40**:478–483.

Tan, G.S., Garchow, B.G., Liu, X., Yeung, J., Morris, J.P., Cuellar, T.L., McManus, M.T. & Kiriakidou, M. (2009) Expanded RNA-binding activities of mammalian Argonaute 2. *Nucleic Acids Res*. **37**:7533-7545.

TargetScan. Prediction of microRNA targets. Release 6.2: June 2012. [WWW-document]. < http://www.targetscan.org/>. (Read 15.1.2014).

The cancer genome atlas (TCGA) network (2012) Comprehensive molecular portraits of human breast tumours. *Nature.* **490**:61–70.

Thomson, D.W., Bracken, C.P. & Goodall, G.J. (2011) Experimental strategies for microRNA target identification. *Nucleic Acids Res.* **39**:6845–6853.

Tong, T., Gui, H., Jin, F., Heck, B.W., Lin, P., Ma, J., Fondell, J.D. & Tsai, C.C. (2011) Ataxin-1 and brother of ataxin-1 are components of the notch signalling pathway. *Embo Rep.* **5**:428–435.

Tuschen, M. (2013) Regulatory role of microRNAs in Glioblastoma multiforme. Bachelor Thesis, Ruprecht-Karls-Universität Heidelberg, Germany.

Valencia-Sanchez, M.A., Liu, J., Hannon, G.J. & Parker, R. (2006) Control of translation and mRNA degradation by miRNAs and siRNAs. *Genes Dev.* **20:**515–524.

Virág, L. & Szabó, C. (2002) The therapeutic potential of poly(ADP-ribose) polymerase inhibitors. *Pharmacol. Rev*. **54**:375-429.

Volinia, S., Calin, G.A., Liu, C.G., Ambs, S., Cimmino, A., Petrocca, F., Visone, R., Iorio, M., Roldo, C., Ferracin, M., Prueitt, R.L., Yanaihara, N., Lanza, G., Scarpa, A., Vecchione, A., Negrini, M., Harris, C.C. & Croce, C.M. (2006) A microRNA expression signature of human solid tumours defines cancer gene targets. *Proc. Natl. Acad. Sci. USA.* **103**:2257–2261.

Wang, D., Qiu, C., Zhang, H., Wang, J., Cui, Q., & Yin, Y. (2010) Human microRNA oncogenes and tumour suppressors show significantly different biological patterns: from functions to targets. *PLoS One.* **5**:13067–13073.

Warburg, O. (1956) On the origin of cancer cells. *Science.* **123**:309–314.

Weinberg, R.A. (2013) The biology of cancer, 2nd edition, p. 25–26, 180–202, 206, 223–224, 227–228, 361. Garland Science, Taylor & Francis Group, LLC, New York.

Wilcoxon, F. (1945) Individual comparisons by ranking methods. *Biometrics Bull.* **1**:80–83.

Willett, W.C. (2002) Balancing life-style and genomics research for disease prevention. *Science.* **296**:695–698.

Williams, H.P. (1999) Model building in mathematical programming, 4th edition, p. 7., John Wiley & Sons, Ltd, United Kingdom.

Wolfgang, H., Heydebreck, A., Sueltmann, H., Poustka, A. & Vingron, M. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics.* **18**:96–104.

World health organisation (WHO). 2012. [WWW-document]. Clobocan 2012: estimated cancer incidence, mortality and prevalence worldwide 2012. <http://globocan.iarc.fr/Pages/fact_sheets_cancer.aspx>. (Read 2.1.2014).

Yu, F., Jin, L., Yang, G., Ji, L., Wang, F., Lu, Z. (2014) Post-transcriptional repression of FOXO1 by QKI results in low levels of FOXO1 expression in breast cancer cells. *Oncol. Rep.* **31**:1459–1465.

Zearo, S., Kim, E., Zhu, Y., Zhao, J.T., Sidhu, S.B. & Robinson, B.G. (2014) MicroRNA-484 is more highly expressed in serum of early breast cancer patients compared to healthy volunteers. *BMC Cancer.* **14**:200–206.

Zeng, W & Ali Mortazavi, A. (2012) Technical considerations for functional sequencing assays. *Nat. Immunol.* **13**:802–807.

Zhu, R., Zou, S.T., Wan, J.M., Li, W., Li, X.L. & Zhu, W. (2013) BTG1 inhibits breast cancer cell growth through induction of cell cycle arrest and apoptosis. *Oncol. Rep.* **30**:2137–2144.

Zuur, A.F., Ieno, E.N. & Meesters E.H.W.G. (2009) A beginner's guide to R, Springer, New York.

Appendix A

An example function in R programming language for performing MIP.

The input of the function in this code consist of mRNA data (e11_zs_t and e_zs_t2), miRNA data (data_zs_t, data_zs_t2), the amount of miRNAs used (mir_amount) and the list of genes that are evaluated (list).

```
fi_loop2 =
function (e11_zs_t,  e11_zs_t2,  data_zs_t, data_zs_t2, mir_amount,  list) {
  library("gurobi")
  library("Matrix")

  # Making the initial vector for MIP results, gene names, and miRNAs:
  mg_list=vector(mode="list", length=length(list))
  names(mg_list) <- f_ent_gsymb(list)
  gene_nam=vector(mode="list", length=length(list))
  mirs=vector(mode="list", length=length(list))
  mi=mir_amount

# Evaluating all the genes in the list with a specific MIP function (f_limo_restr2),
and obtaining the miRNAs that regulate them (f_mir_limoo):
  for (i in 1:length(list)) {
    gene_nam[[i]] = as.vector((list)[i])
    mirs[[i]] = f_mir_limoo(tush_nl, as.vector(unlist(gene_nam[[i]])))
    mg_list[[i]] = f_limo_restr2(e11_zs_t, e11_zs_t2,data_zs_t,
data_zs_t2,as.vector(unlist(gene_nam[[i]])), as.vector(unlist(mirs[[i]])),mi)  }

  #Here the function returns the calculated values as the specified list:
  return(mg_list)
}
```

Appendix B

The selected results from differential expression analysis.

Table B1. Differentially expressed miRNAs in TNBC according to Wilcoxon test.

miRNA	Q value[1]
hsa-miR-425-3p	0.0002
hsa-miR-3609	0.0002
hsa-miR-190a-5p	0.0002
hsa-miR-1245a	0.0003
hsa-miR-30a-3p	0.0005
hsa-miR-431-5p	0.0005
hsa-miR-129-2-3p	0.0005
hsa-miR-572	0.0007
hsa-miR-365a-3p	0.0008
hsa-miR-335-5p	0.0009
hsa-miR-299-5p	0.0009
hsa-miR-202-3p	0.0014
hsa-miR-150-3p	0.0016
hsa-miR-493-5p	0.0019
hsa-miR-3667-5p	0.0021
hsa-miR-186-5p	0.0040
hsa-miR-28-5p	0.0041
hsa-miR-30a-5p	0.0049
hsa-miR-200c-3p	0.0049
hsa-miR-296-5p	0.0050
hsa-miR-29b-3p	0.0051
hsa-miR-602	0.0054
hsa-miR-203a	0.0063
hsa-miR-489-3p	0.0066
hsa-miR-216a-5p	0.0092
hsa-miR-1909-5p	0.0105
hsa-miR-191-3p	0.0109
hsa-miR-3124-5p	0.0112
hsa-miR-216b-5p	0.0154
hsa-miR-211-5p	0.0172
hsa-miR-654-3p	0.0187
hsa-miR-3158-3p	0.0208
hsa-miR-301b	0.0213
hsa-miR-548e-3p	0.0213
hsa-miR-511-5p	0.0257
hsa-miR-877-5p	0.0388
hsa-miR-3680-5p	0.0430
hsa-miR-545-3p	0.0485

1) The Q values are Bonferroni corrected results from that test. miRNAs are sorted according to ascending The Q values.

Table B2. Top 25 genes of 1963 differentially expressed mRNAs in TNBC according to Wilcoxon test.

Gene	Q value[1]
C8orf33	$1.01 \cdot 10^{-4}$
CBX6	$1.01 \cdot 10^{-4}$
GKAP1	$1.01 \cdot 10^{-4}$
NAGA	$1.01 \cdot 10^{-4}$
PITRM1	$1.01 \cdot 10^{-4}$
ZNF577	$1.01 \cdot 10^{-4}$
SP5	$1.03 \cdot 10^{-4}$
KCNK15	$1.04 \cdot 10^{-4}$
MLLT3	$1.04 \cdot 10^{-4}$
PEMT	$1.04 \cdot 10^{-4}$
TTC19	$1.04 \cdot 10^{-4}$
ZNF596	$1.04 \cdot 10^{-4}$
MS4A15	$1.05 \cdot 10^{-4}$
C10orf108	$1.07 \cdot 10^{-4}$
ASPSCR1	$1.07 \cdot 10^{-4}$
C16orf53	$1.07 \cdot 10^{-4}$
C1orf175	$1.07 \cdot 10^{-4}$
DAPK1	$1.07 \cdot 10^{-4}$
MMP17	$1.07 \cdot 10^{-4}$
MRPL45	$1.07 \cdot 10^{-4}$
NCAPH2	$1.07 \cdot 10^{-4}$
PECR	$1.07 \cdot 10^{-4}$
SEC24B	$1.07 \cdot 10^{-4}$
SYDE2	$1.07 \cdot 10^{-4}$
C19orf2	$1.10 \cdot 10^{-4}$

1) The Q values are Bonferroni corrected results from that test. Genes are sorted according to ascending the Q values.

Appendix C

The selected results of enrichment analyses with hypergeometric tests for specific gene sets (A-M), in order to evaluate GO molecular functions, GO biological processes and KEGG pathways.

Table C1. GO biological processes of down-regulated genes that are targets of up-regulated EmRs in TNBC (gene set: A).

GO ID	GO term	Q value	Sig.genes	Odds ratio
GO:0014032	neural crest cell development	0.036	3	6.1
GO:2000505	regulation of energy homeostasis	0.041	2	42
GO:0060137	maternal process involved in parturition	0.041	2	9.1
GO:0042482	positive regulation of odontogenesis	0.041	2	26
GO:0001821	histamine secretion	0.041	2	52
GO:0048511	rhythmic process	0.043	3	42
GO:0008286	insulin receptor signalling pathway	0.046	7	21
GO:0016049	cell growth	0.048	4	52
GO:0045055	regulated secretory pathway	0.049	2	42
GO:0046324	regulation of glucose import	0.049	2	70
GO:0014824	artery smooth muscle contraction	0.049	2	52

Table C2. GO molecular functions of up-regulated genes that are targets of down-regulated EmRs in TNBC (gene set: B)

GO ID	GO term	Q value	Sig.genes	Odds ratio
GO:0017018	myosin phosphatase activity	0.044	1	310
GO:0050115	myosin-light-chain-phosphatase activity	0.044	1	310
GO:0019780	FAT10 activating enzyme activity	0.044	1	310
GO:0008917	lipopolysaccharide N-acetylglucosaminyltransfeRase activity	0.044	1	310
GO:0008457	β-galactosyl-N-acetylglucosaminyl-galactosylglucosyl-ceramide β-1,3-acetylglucosaminyltransfeRase activity	0.044	1	310

Table C3. KEGG pathway for up-regulated genes that are targets of down-regulated EmRs in TNBC (gene set: B).

GO ID	GO term	Q value	Sig.genes	Odds ratio
Kegg:04120	Ubiquitin mediated proteolysis	0.027	4	9.3

Table C4. GO molecular functions of up-regulated DEGs that are targets of down-regulated miRNAs in TNBC (gene set: D).

GO ID	GO term	Q value	Sig.gens	Odds ratio
GO:0005515	protein binding	0.012	144	1.3

Table C5. GO biological processes of down-regulated genes in cooperation list (gene set: E).

GO ID	GO term	Q value	Sig.genes	Odds ratio
GO:0008286	insulin receptor signalling pathway	0.007	9	6.6
GO:0035404	histone-serine phosphorylation	0.047	2	87

Table C6. GO biological processes of genes in cooperation list (gene set: G).

GO ID	GO term	Q value	Sig.genes	Odds ratio
GO:0016568	chromatin modification	0.0091	13	3.9
GO:0014032	neural crest cell development	0.012	4	5.3
GO:0008286	insulin receptor signalling pathway	0.018	10	21

Table C7. GO molecular functions of group one's down-regulated genes that are targets of up-regulated EmRs (gene set: H).

GO ID	GO term	Q value	Sig.genes	Odds ratio
GO:0030971	receptor tyrosine kinase binding	0.037	3	24

Table C8. GO molecular functions of group one's up-regulated genes that are targets of down-regulated EmRs (gene set: I).

GO ID	GO term	Q value	Sig.genes	Odds ratio
GO:0017018	myosin phosphatase activity	0.014	1	8.4
GO:0050115	myosin-light-chain-phosphatase activity	0.014	1	16
GO:0019780	FAT10 activating enzyme activity	0.014	1	100
GO:0008917	lipopolysaccharide N-acetylglucosaminyltransfeRase activity	0.014	1	69
GO:0008457	β-galactosyl-N-acetylglucosaminylgalactosylglucosyl-ceramide β-1,3-acetylglucosaminyltransfeRase activity	0.014	1	150
GO:0019901	protein kinase binding	0.034	3	88
GO:0052650	NADP-retinol dehydrogenase activity	0.035	1	620
GO:0070215	MDM2 binding	0.037	1	620
GO:0061133	endopeptidase activator activity	0.037	1	88
GO:0032452	histone demethylase activity	0.037	1	210
GO:0017124	SH3 domain binding	0.041	3	69
GO:0016538	cyclin-dependent protein kinase regulator activity	0.041	1	120
GO:0016812	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds, in cyclic amides	0.041	1	69
GO:0005212	structural constituent of eye lens	0.041	1	120
GO:0005522	profilin binding	0.043	1	620
GO:0004745	retinol dehydrogenase activity	0.043	1	620
GO:0016810	hydrolase activity, acting on carbon nitrogen (but not peptide) bonds	0.043	1	620

Table C9. GO molecular functions of down-regulated genes of group two that are targets of up-regulated EmRs (gene set: J).

GO ID	GO term	Q value	Sig.genes	Odds ratio
GO:0047325	inositol tetrakisphosphate 1-kinase activity	0.0043	1	74
GO:0052725	inositol-1,3,4-trisphosphate 6-kinase activity	0.0043	1	50
GO:0052726	inositol-1,3,4-trisphosphate 5-kinase activity	0.0043	1	110
GO:0034046	poly(G) RNA binding	0.0065	1	320
GO:0043422	protein kinase B binding	0.013	1	970
GO:0008266	poly(U) RNA binding	0.013	1	320
GO:0042809	vitamin D receptor binding	0.031	1	1900
GO:0043621	protein self-association	0.042	1	1900
GO:0003712	transcription cofactor activity	0.05	1	1900

Table C10. GO biological processes of down-regulated genes of group two that are targets of up-regulated EmRs (gene set: J).

GO ID	GO term	Q value	Sig.genes	Odds ratio
GO:0051168	nuclear export	0.017	1	21
GO:0045603	positive regulation of endothelial cell differentiation	0.017	1	13
GO:0043569	negative regulation of insulin-like growth factor receptor signalling pathway	0.017	1	100
GO:0045663	positive regulation of myoblast differentiation	0.017	1	64
GO:0045778	positive regulation of ossification	0.018	1	190
GO:0007292	female gamete generation	0.018	1	51
GO:0042981	regulation of apoptotic process	0.019	2	88
GO:0032957	inositol trisphosphate metabolic process	0.02	1	84
GO:0090003	regulation of establishment of protein localization in plasma membrane	0.02	1	240
GO:0060385	axonogenesis involved in innervation	0.02	1	390
GO:0045671	negative regulation of osteoclast differentiation	0.024	1	130
GO:0046324	regulation of glucose import	0.025	1	320
GO:0008285	negative regulation of cell proliferation	0.027	2	190
GO:0008542	visual learning	0.028	1	970
GO:0060079	regulation of excitatory postsynaptic membrane potential	0.029	1	970
GO:0042326	negative regulation of phosphorylation	0.029	1	320
GO:0008344	adult locomotory behavior	0.034	1	320
GO:0010468	regulation of gene expression	0.041	1	640

Table C11. KEGG pathways for down-regulated genes of group two that are targets of up-regulated EmRs (gene set: J).

GO ID	GO term	Q value	Sig.genes	Odds ratio
Kegg:03018	RNA degradation	0.0023	2	27
Kegg:04070	Phosphatidylinositol signalling system	0.045	1	60

Table C12. GO biological processes of down-regulated genes of group three that are targets of up-regulated EmRs (gene set: L).

GO ID	GO term	Q value	Sig.genes	Odds ratio
GO:0014032	neural crest cell development	0.0016	3	4.6
GO:0060137	maternal process involved in parturition	0.0032	2	6.5
GO:0042482	positive regulation of odontogenesis	0.0032	2	43
GO:0001821	histamine secretion	0.004	2	11
GO:0014824	artery smooth muscle contraction	0.0045	2	19
GO:0007585	respiratory gaseous exchange	0.0048	3	30
GO:0030818	negative regulation of cAMP biosynthetic process	0.008	2	8.4
GO:0042310	vasoconstriction	0.0093	2	23
GO:0048016	inositol phosphate-mediated signalling	0.011	2	86
GO:0051482	elevation of cytosolic calcium ion concentration involved in G-protein signalling coupled to IP3 second messenger	0.012	2	14
GO:0090023	positive regulation of neutrophil chemotaxis	0.025	2	46
GO:0051545	negative regulation of elastin biosynthetic process	0.029	1	18
GO:0042313	protein kinase C deactivation	0.029	1	86
GO:0031583	activation of phospholipase D activity by G-protein coupled receptor protein signalling pathway	0.029	1	30
GO:0043179	rhythmic excitation	0.029	1	22
GO:0042045	epithelial fluid transport	0.029	1	40
GO:0060005	vestibular reflex	0.029	1	170
GO:2000108	positive regulation of leukocyte apoptosis	0.029	1	14
GO:0043112	receptor metabolic process	0.029	1	20
GO:0071417	cellular response to organic nitrogen	0.029	1	86
GO:2000647	negative regulation of stem cell proliferation	0.029	1	76
GO:0032642	regulation of chemokine production	0.029	1	98
GO:0070563	negative regulation of vitamin D receptor signalling pathway	0.029	1	86
GO:0035921	desmosome disassembly	0.029	1	86
GO:0001544	initiation of primordial ovarian follicle growth	0.029	1	86
GO:0051321	meiotic cell cycle	0.029	1	170
GO:0001958	endochondral ossification	0.03	2	86
GO:0001701	in utero embryonic development	0.03	4	110
GO:0016568	chromatin modification	0.031	4	86
GO:0007517	muscle organ development	0.031	3	86
GO:0007205	activation of protein kinase C activity by G-protein coupled receptor protein signalling pathway	0.032	2	86
GO:0050885	neuromuscular process controlling balance	0.033	2	340
GO:0000122	negative regulation of transcription from RNA polymeRase II promoter	0.037	5	110
GO:0048661	positive regulation of smooth muscle cell proliferation	0.038	2	170
GO:0070101	positive regulation of chemokine-mediated signalling pathway	0.039	1	170

Table C12. Continued. GO biological processes of Down-regulated genes of group three that are targets of up-regulated EmRs (gene set: L).

GO ID	GO term	Q value	Sig.genes	Odds ratio
GO:0048520	positive regulation of behavior	0.039	1	86
GO:0051771	negative regulation of nitric-oxide synthase biosynthetic process	0.039	1	110
GO:0018023	peptidyl-lysine trimethylation	0.039	1	140
GO:0006933	negative regulation of cell adhesion involved in substrate-bound cell migration	0.039	1	86
GO:0035404	histone-serine phosphorylation	0.039	1	340
GO:0033578	protein glycosylation in Golgi	0.039	1	230
GO:0003094	glomerular filtration	0.039	1	340
GO:0048144	fibroblast proliferation	0.039	1	86
GO:0001649	osteoblast differentiation	0.041	2	110
GO:0001569	patterning of blood vessels	0.041	2	86
GO:0002062	chondrocyte differentiation	0.041	2	69
GO:0007266	Rho protein signal transduction	0.047	2	170
GO:0042759	long-chain fatty acid biosynthetic process	0.047	1	110
GO:0032269	negative regulation of cellular protein metabolic process	0.047	1	340
GO:0030185	nitric oxide transport	0.047	1	340
GO:0034392	negative regulation of smooth muscle cell apoptosis	0.047	1	340
GO:0060585	positive regulation of prostaglandin-endoperoxide synthase activity	0.047	1	340
GO:0001667	ameboidal cell migration	0.047	1	86
GO:0060429	epithelium development	0.047	1	110
GO:0048659	smooth muscle cell proliferation	0.047	1	86
GO:0001547	antral ovarian follicle growth	0.047	1	170
GO:0042593	glucose homeostasis	0.048	2	86
GO:0046415	urate metabolic process	0.048	1	86
GO:2000505	regulation of energy homeostasis	0.048	1	170
GO:0010957	negative regulation of vitamin D biosynthetic process	0.048	1	170
GO:0060298	positive regulation of sarcomere organization	0.048	1	170
GO:0060627	regulation of vesicle-mediated transport	0.048	1	340
GO:0010452	histone H3-K36 methylation	0.048	1	110
GO:0040019	positive regulation of embryonic development	0.048	1	86
GO:0007176	regulation of epidermal growth factor-activated receptor activity	0.048	1	340
GO:0043084	penile erection	0.048	1	170
GO:0014826	vein smooth muscle contraction	0.048	1	110
GO:0045321	leukocyte activation	0.048	1	170
GO:0030072	peptide hormone secretion	0.048	1	340
GO:0010839	negative regulation of keratinocyte proliferation	0.048	1	340
GO:0048752	semicircular canal morphogenesis	0.048	1	340

Table C12. Continued. GO biological processes of down-regulated genes of group three that are targets of up-regulated EmRs (gene set: L).

GO ID	GO term	Q value	Sig.genes	Odds ratio
GO:0010839	negative regulation of keratinocyte proliferation	0.048	1	340
GO:0018027	peptidyl-lysine dimethylation	0.048	1	340
GO:0033629	negative regulation of cell adhesion mediated by integrin	0.048	1	110
GO:0042305	specification of segmental identity, mandibular segment	0.048	1	340
GO:0030335	positive regulation of cell migration	0.049	3	340

Table C13. GO molecular functions of up-regulated genes of group three that are targets of down-regulated EmRs (gene set: M).

GO ID	GO term	Q value	Sig.genes	Odds ratio
GO:0018024	histone-lysine N-methyltransfeRase activity	0.022	2	61

Table C14. GO biological processes of up-regulated genes of group three that are targets of down-regulated EmRs (gene set: M).

GO ID	GO term	Q value	Sig.genes	Odds ratio
GO:0034968	histone lysine methylation	0.012	2	14
GO:0042518	negative regulation of tyrosine phosphorylation of Stat3 protein	0.043	1	130
GO:0060235	lens induction in camera-type eye	0.043	1	160
GO:0060996	dendritic spine development	0.043	1	130
GO:0046425	regulation of JAK-STAT cascade	0.043	1	320
GO:0071679	commissural neuron axon guidance	0.043	1	320
GO:0045716	positive regulation of low-density lipoprotein particle receptor biosynthetic process	0.043	1	110
GO:0072369	regulation of lipid transport by positive regulation of transcription from RNA polymeRase II promoter	0.043	1	640
GO:0048679	regulation of axon regeneration	0.043	1	210
GO:0043516	regulation of DNA damage response, signal transduction by TP53 class mediator	0.043	1	160
GO:0021536	diencephalon development	0.043	1	130
GO:2000677	regulation of transcription regulatory region DNA binding	0.043	1	210
GO:0010988	regulation of low-density lipoprotein particle clearance	0.043	1	110
GO:0050878	regulation of body fluid levels	0.046	1	130
GO:0046426	negative regulation of JAK-STAT cascade	0.046	1	110
GO:0034067	protein localization in Golgi apparatus	0.046	1	210
GO:0006979	response to oxidative stress	0.046	2	160
GO:0018026	peptidyl-lysine monomethylation	0.047	1	160
GO:0045197	establishment or maintenance of epithelial cell apical/basal polarity	0.047	1	640
GO:0022038	corpus callosum development	0.047	1	640
GO:0021797	forebrain anterior/posterior pattern specification	0.047	1	320
GO:0072385	minus-end-directed organelle transport along microtubule	0.047	1	160
GO:0018125	peptidyl-cysteine methylation	0.047	1	320

Table C15. KEGG pathways of up-regulated genes of group three that are targets of down-regulated EmRs (gene set: M).

GO ID	GO term	Q value	Sig.genes	Odds ratio
Kegg:00310	Lysine degradation	0.031	2	26

Appendix D

The supplementary MIP results and regulation pattern analysis.

Table D1. Regulation pattern in group two out of the three groups according to the Hamming distance analysis to the cooperation list genes at TNBC.

Gene symbol	Regulation magnitude[1]
APPL1	-0.385
ATXN1	-0.725
BTG1	-0.639
EPC2	-0.502
FNDC4	-0.820
GATAD2B	0.367
IMPDH1	0.882
ITPK1	-0.671
NPTX1	-2.019
TOB2	-0.332

1) Regulation magnitude is the subtraction between median values of gene expression signals at TNBC and normal samples.

Table D2. MIP results with restricting the amount of miRNAs to 22 for an extended cooperation list, i.e. differential expression of genes was not considered, and using all of samples.

No[1]	Gene symbol	PCC	miRNA freq. input to model[2]	EmR freq. input to model	miRNA freq. used by model
1	*CELF2*	0.790	70	8	19
2	*QKI*	0.775	100	7	22
3	*TGFBR2*	0.747	26	5	12
4	*RUNX1T1*	0.739	56	5	19
5	*NFIB*	0.730	82	9	20
6	*BACH2*	0.729	50	5	19
7	*TCF4*	0.724	59	8	20
8	*CREB5*	0.716	51	7	19
9	*FZD4*	0.708	36	5	23
10	*CCND2*	0.706	49	6	19

1) The results are sorted according to descending PCCs. 2) Some of the input miRNAs were EmRs, and the model used some of the 22 miRNAs that it was restricted to out of its total input of miRNAs.

Table D3. The linear modelling β coefficients' values for *QKI* gene.

No	miRNAs	Values of β coefficients
1	hsa-miR-19a-3p	-0.007
2	hsa-miR-3200-3p	-0.017
3	hsa-miR-4295	-0.033
4	hsa-miR-221-3p	-0.063
5	hsa-miR-186-5p	-0.065
6	hsa-miR-130b-3p	-0.070
7	hsa-miR-200b-3p	-0.073
8	hsa-miR-143-3p	-0.076
9	hsa-miR-23c	-0.097
10	hsa-miR-16-5p	-0.097
11	hsa-miR-33a-5p	-0.110
12	hsa-miR-548a-3p	-0.111
13	hsa-miR-3619-5p	-0.117
14	hsa-miR-181c-5p	-0.127
15	hsa-miR-101-3p	-0.128
16	hsa-miR-497-5p	-0.134
17	hsa-miR-3202	-0.153
18	hsa-miR-200c-3p	-0.183
19	hsa-miR-452-5p	-0.217
20	hsa-miR-361-5p	-0.246

Table D4. MIP results, using 22 miRNAs as a restricting case, for the genes in the group two of the Cooperation showing the connection of miRNAs in the negative regulation of insulin-like growth factor receptor signalling pathway.

No[1]	Gene symbol	PCC	miRNA freq. input to model[2]	EmR freq. input to model	miRNA freq. used by model
1	*ATXN1*	0.55	63	7	17
2	*EPC2*	0.45	30	6	10
3	*GATAD2B*	0.43	52	5	15
4	*BTG1*	0.38	18	3	10
5	*NPTX1*	0.35	26	3	12
6	*TOB2*	0.35	22	2	6
7	*IMPDH1*	0.29	15	3	8
8	*APPL1*	0.22	25	4	6
9	*FNDC4*	0.18	11	2	2
10	*ITPK1*	0.13	6	2	2

1) The results are sorted according to descending PCCs. 2) Some of the input miRNAs were EmRs, and the model used some of the 22 miRNAs that it was restricted to out of its total input of miRNAs.

Table D5. Results of MIP (restricting the miRNAs to 10) with Heidelberg list of genes using all samples.

No[1]	Gene symbol	PCC	miRNA freq. input to model	EmR freq. input to model	miRNA freq. used by model
1	*CAD*	0.52	5	1	3
2	*FASN*	0.42	20	1	10
3	*GLS*	0.26	10	1	4
4	*LDHA*	0.25	7	0	6
5	*GLUD1*	0.25	11	0	5

1) The top 5 of all 22 genes are shown according to the PCCs in this list.

Table D6. Results of MIP (restricting the miRNAs to 10) with Heidelberg list of genes using TNBC samples.

No[1]	Gene symbol	PCC	miRNA freq. input to model	EmR freq. input to model	miRNA freq. used by model
1	*FASN*	0.551	20	1	10
2	*SMS*	0.357	5	0	2
3	*ACADL*	0.276	4	0	2
4	*HK1*	0.268	8	0	5
5	*PKM2*	0.251	11	0	8

1) The top 5 of all 22 genes are shown according to the PCCs in this list.

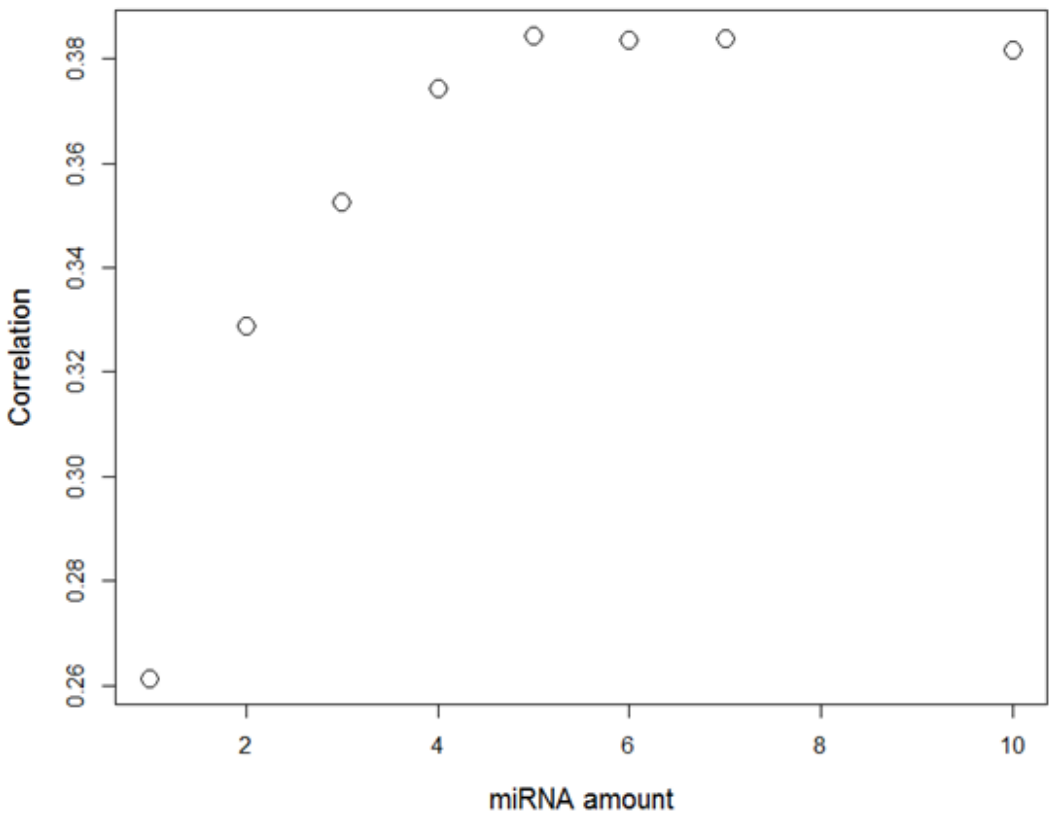

Figure D1. The effect of the amount of miRNAs in MIP predicting *FASN* of Heidelberg list.