Molecular characterization of disseminated cancer cells isolated from patients with luminal B type breast cancer

DISSERTATION ZUR ERLANGUNG DES DOKTORGRADES DER NATURWISSENSCHAFTEN (DR. RER. NAT.) DER FAKULTÄT FÜR BIOLOGIE UND VORKLINISCHE MEDIZIN DER UNIVERSITÄT REGENSBURG

vorgelegt von

Christoph Irlbeck

aus

Schwandorf

im Jahr

2019

Das Promotionsgesuch wurde eingereicht am:

31.10.2019

Die Arbeit wurde angeleitet von:

Prof. Dr. Christoph Klein

Unterschrift:

Table of contents

Table of figures

Table of tables

Table of abbreviations

Abbreviation	Meaning
APC	Allophycocyanin
APS	Ammonium persulfate
ASncmtRNA	Antisense non-coding mitochondrial RNA
ATP	Adenosine triphosphate
BC	Breast cancer
BM	Bone marrow
bp	Base pair
BR	Broad range
BSA	Bovine serum albumin
CCC	Circulating cancer cell
cDNA	Copy/complementary deoxyribonucleic acid
CDK	Cyclin-dependent kinase
CGH	Comparative genomic hybridization
CNA	Copy number alteration
CR	Complete response
CSC	Cancer stem cell
CT	Chemotherapy
DCC	Disseminated cancer cell
DCIS	Ductal carcinoma in situ
DEPC	Diethyl pyrocarbonate
dGTP	Deoxy guanosine triphosphate
DNA	Deoxyribonucleic acid
dNTP	Deoxynucleotide triphosphate
DTT	Dithiothreitol
eBC	Early breast cancer
EBCTCG	Early Breast Cancer Trialists' Collaborative Group
ECM	Extracellular matrix
eDCC	Early DCC
EDTA	Ethylenediaminetetraacetic acid
EMT	Epithelial to mesenchymal transition
EpCAM	Epithelial cell adhesion molecule
ER	Estrogen receptor
ET	Endocrine therapy
eWTA	Extended WTA (for miRNA isolation)
FBS	Fetal bovine serum
FCS	Fetal calf serum
gDNA	Genomic deoxyribonucleic acid
GO	Gene ontology
GTC	Guanidinium thiocyanate
H_2O	Water
HR	Hormone receptor
HBSS	Hank’s balanced salt solution
hcSeq	Half-cell sequencing
HD	Healthy donor
HS	High sensitivity
IHC	Immune histochemistry
IRS	Immune reactive score
ISCN	International System for Human Cytogenetic Nomenclature
ISH	In situ hybridization
LF	Long fragment (artificial RNA)
LP-Seq	LowPass-Sequencing
LumA	Luminal A breast cancer
LumB	Luminal B breast cancer
M0	Non-metastatic
M1	Metastatic
MACS	Magnetic activated cell sorting
mBC	Metastatic breast cancer

Abbreviation	Meaning
mCGH	Metaphase comparative genomic hybridization
ME	Microenvironment
MET	Mesenchymal to epithelial transition
MIB	MiRNA isolation buffer
miRNA	Micro RNA
MNC	Mononuclear cell
MRD	Minimal residual disease
mRNA	Messenger RNA
NA	Not applicable or not available (missing value)
ncRNA	Non-codingRNA
NCC	Non-cancer ($EpCAM^+$) cell
NEB	New England Biolabs
NET	Neoadjuvant endocrine therapy
NGS	Next generation sequencing
NLS	M-lauroylsarcosine
NPV	Negative predictive value
NST	No special type (of breast cancer), formerly called invasive ductal carcinoma
OPA	One Phor All buffer
OS	Overall survival
PAA	Polyacrylamide
PAP	Poly(A) polymerase
PBS	Phosphate buffered saline
PCA	Principal component analysis
PCR	Polymerase chain reaction
PFS	Progression-free survival
piRNA	PIWI-interacting RNA
PPV	Positive predictive value
PR	Progesterone receptor
PT	Primary tumor
QC	Quality control
qPCR	Quantitative PCR
qRT-PCR	Quantitative reverse transcription PCR
q-value	Adjusted p-value
RNA	Ribonucleic acid
rRNA	Ribosomal RNA
RT	Room temperature
SC	Single cell
SCE	Single cell equivalent
SD	Standard deviation
SDC	Sodium deoxycholate
SDS	Sodium dodecyl sulfate
siRNA	Small interfering RNA
Small-Seq	Small RNA Sequencing
SOP	Standard operating procedure
STA	Single tube amplification
sWTA	Standard WTA
TBE	Tris/Borate/EDTA buffer
TdT	Terminal Deoxynucleotidyl Transferase
TEMED	Tetramethylethylenediamine
TFS	ThermoFisher Scientific
TNBC	Triple negative breast cancer
TNM	Tumor-Node-Metastasis staging system
tRNA	Transfer RNA
WGA	Whole genome amplification
WTA	Whole transcriptome amplification

1. Introduction

1.1 Breast cancer

1.1.1 Epidemiology

Breast cancer (BC) is the most frequent type of cancer in females all across the world accounting for 24.2 % of all reported cancer incidences in 2018 (Bray et al., 2018). With 15 % of cancer-related deaths, it was also the leading cause of cancer-related mortality in women. Across both sexes, it ranked second (close behind lung cancer) regarding the number of incidences with a total of 2.1 million cases (11.6 % of all cancer cases worldwide) in 2018 according to the same study. With close to 627,000 (6.6 %) deaths, it occupied the fifth place for global cancer-related mortality across all cancer types in 2018. For comparison, lung cancer claimed the most victims with a number of 1.76 million (18.4 % of global cancer related deaths) in the same year.

BC is also the most common form of cancer in women in Germany, with a 5-year prevalence rate of 559 per 100,000 inhabitants (Bertz et al., 2010). According to the *Report on the Situation of Cancer in Germany,* 71,640 women were newly diagnosed with BC in 2013 at a mean age of 64.3, while 17,853 (roughly 25 % of new cases) died of the disease in the same year at a mean age of 72.6 years (Barnes et al., 2016). Although males can also suffer from breast cancer, this occurs only very rarely (682 cases in 2013). Table 1-1 provides an overview of the main characteristics of the epidemiology of BC in females. Overall, the 5- and 10-year survival rates for BC are high compared to other cancer types, but due to the large number of cases, this still translates into many deaths. Barnes and colleagues also reported that the frequency of BC doubled since 1970, while the number of deaths only increased by 40 %, as the numbers tumors diagnosed early (carcinoma in situ and stage I, see chapter 1.1.2) in women at the age of 50-69 increased in recent years, while the rate of late stage tumors (stage II and above) decreased since 2011. The relative reduction of mortality is due to the introduction of a country-wide mammography screening program for early detection of BC in this age group, which accounts for 45 % of newly diagnosed BC cases (followed by women aged >69 [37 % of cases] and <50 [18 % of cases]). Overall, mortality has decreased by roughly a quarter in the 50-69 age group and by about a third in the group <50 years of age between 1999 and 2013. The prevalent cause of cancer-related deaths is metastasis, which will be discussed in detail in chapter 1.2.

Table 1-1 Overview of breast cancer incidence, mortality, survival, and prevalence in Germany in 2013. Adapted from (Barnes et al., 2016). Numbers represent female cases only.

Parameter	Incidences	Deaths	Survival	Prevalence
Absolute number	71,640	17,853	-	-
Mean age	64.3	72.6	-	-
Rate per 100,000	174	43.4	-	-
5-year	-	-	88 %	315,740
10-year	-	-	82 %	551,960

1.1.2 Classification

BC is a very heterogenous disease, both clinically and genetically (Stingl and Caldas, 2007; Malhotra et al., 2010), and is usually classified using the following separate systems: histopathology, grade, receptor status, and stage.

From a histopathological standpoint, BC was historically classified into roughly 18 subtypes depending on histological features of the tumor (Tavassoli and Devilee, 2003; Malhotra et al., 2010). Broadly, BC is classified into two main classes: carcinoma in situ (mostly ductal carcinoma in situ, DCIS) and invasive (infiltrating) carcinoma (see Figure 1-1). Both of these are further subdivided into multiple subclasses based on their growth patterns and cytological features (Malhotra et al., 2010). The most frequent of these subclasses is the invasive ductal class now known as "no special type" (NST), which represents about 80 % of BC cases, while the remaining 20 % of cases comprise carcinoma in situ and several forms of so-called invasive special types: (Stingl and Caldas, 2007; Lakhani et al., 2012; Sinn and Kreipe, 2013).

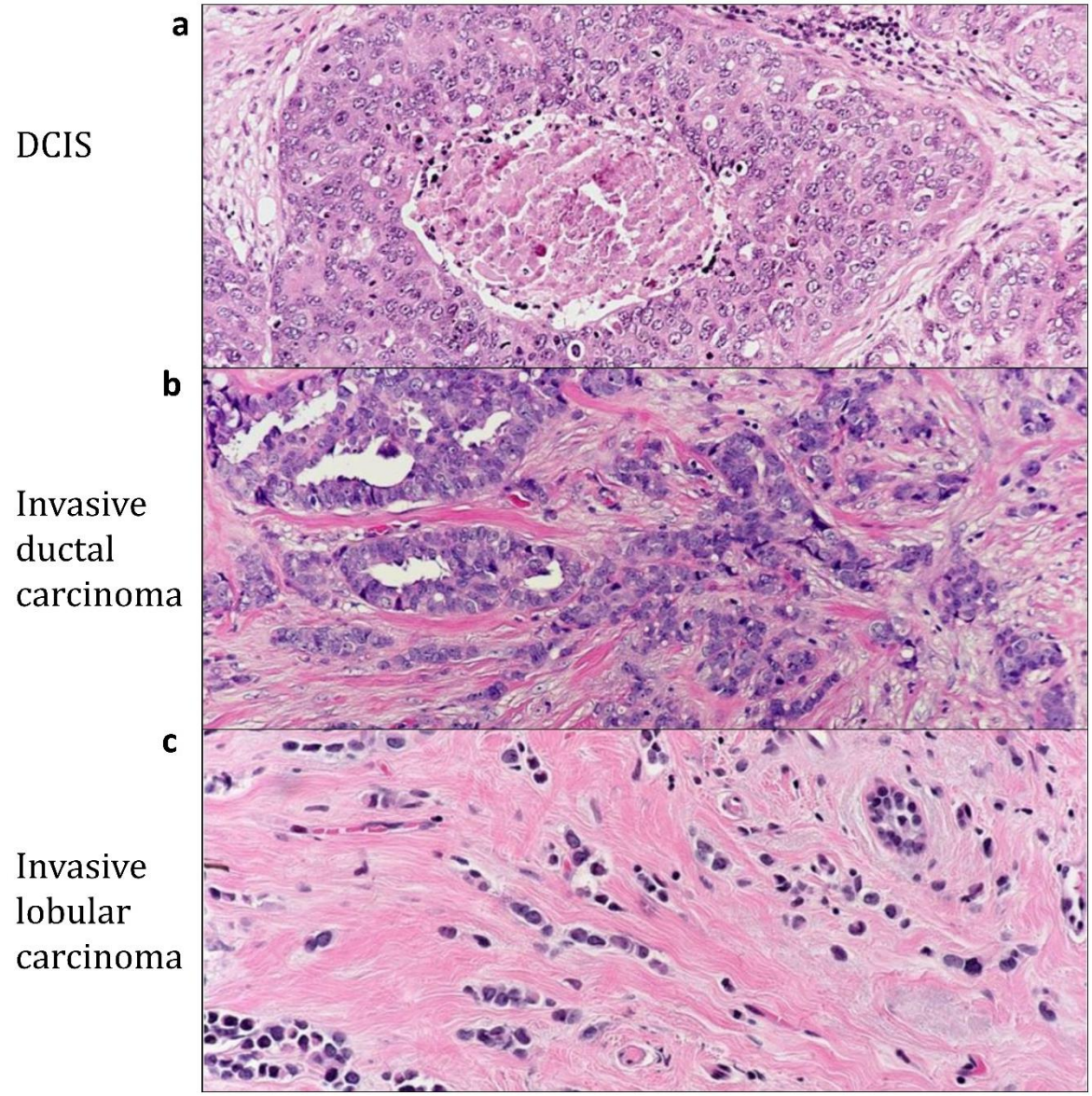


Figure 1-1 Hematoxylin and eosin staining of the three most common histopathological breast cancer classes. The images show exemplary images of hematoxylin and eosin stained breast cancer biopsies. (a) Ductal carcinoma in situ, (b) invasive ductal carcinoma also known as NST, and (c) invasive lobular carcinoma. Images adapted from: *https://pathology.jhu.edu/breast/types-of-breast-cancer/* (date of image retrieval: 07.08.2019).

The grading of breast tumors established by Elston and Ellis is based on the degree of differentiation observed in the cancer cells compared to healthy breast tissue (Elston and Ellis, 1991). The differentiation is assessed using three separate criteria: tubule formation (proportion of primary tumor [PT] forming normal ducts), nuclear pleomorphism (uniform structure of cell nuclei), and mitotic count (number of mitotic cells). Each of these criteria is first rated individually before the final grade is calculated from the three separate grades. Cancer cells with a high differentiation, meaning a close resemblance of the healthy cells, are assigned a low grade, while poor differentiation results in a high grade. Grades range from one (high differentiation) up to three (low differentiation). High grade is strongly correlated with poor patient outcome providing important prognostic information (Elston and Ellis, 1991).

The third method for classification is determination of the receptor status, which refers to the presence or absence of three important cell surface receptors: estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2 or ERBB2). ER and PR status are often summarized under the term hormone receptor (HR) status. The HR status is determined utilizing immunohistochemistry (IHC) and the so-called immune reactive score (IRS) established by Remelle and Stegner, which ranges from zero to twelve and is determined independently for ER and PR (Remmele and Stegner, 1987). A tumor is considered HR-positive, if it has a score of at least two for either of the receptors. HR-positive tumors depend on the respective hormones for their growth, which represents a good target for drug treatment (e.g. with tamoxifen, see chapter 1.1.4). Similar to the HR status, the HER2 status is also determined by IHC and results can be positive, equivocal, or negative (Wolff et al., 2013). In case of an equivocal result, an alternative IHC or in situ hybridization (ISH) assay is used to clarify the result. $HER2^+$ tumors respond well to targeted treatment with the monoclonal antibody trastuzumab which leads to good overall patient survival (see Table 1-3 in chapter 1.1.5).

Lastly, staging of BC is based on the TNM classification system, which currently exists in its eighth version since its establishment that came into effect in 2018 (Amin et al., 2017; Brierley et al., 2017). In the TNM system, three different variables are measured: the size of the primary tumor (“T”), presence of tumor cells in regional lymph nodes (“N”), and presence of metastases at distant sites (“M”). In brief, the system determines the tumor load of a patient and the cancer’s spread throughout the body. The three variables are first evaluated independently and indexed, e.g. M_0 (referred to as M0 throughout this thesis) indicates absence of distant metastasis, while M_1 (referred to as M1) indicates its presence. The combination of indexed T, N, and M variables is then used to assign a cancer stage. In comparison to older versions of the staging system, the current one considers not only the TNM characteristics, but also the tumor grade and the receptor status (see above), in order to reflect more closely what clinicians have already been doing when planning the treatment of a patient. The stages range from zero (carcinoma in situ), over one through three (growth within the breast and regional lymph nodes in different severities), to four (metastatic disease). Depending on the combination of variables, stages I-III are further subdivided into two or three substages depending on the combination of T and N status. The stages 0-III are often referred to as early BC (eBC), which is defined as every BC that has not spread to distant organs (Harbeck and Gnant, 2017). Metastatic breast cancer (mBC) is always classified as stage IV regardless of any of the other variables. Higher stages, regarding both the four main stages and respective substages, are correlated with a worse prognosis (Polyak and Metzger Filho, 2012).

1.1.3 Molecular subtypes

Several global gene expression studies using complementary deoxynucleic acid (cDNA) microarrays have revealed that BC can be divided into five to six molecular subtypes, namely Luminal A (LumA), Luminal B (LumB), HER2-enriched, basal-like, claudin-low, and a normal breast-like group based on their gene expression (Perou et al., 2000; Sørlie et al., 2001; Herschkowitz et al., 2007; Perou, 2010; Prat et al., 2010; Lehmann et al., 2011). There is evidence that the normal breast-like subtype is actually a technical artifact caused by contamination of samples with normal breast tissue, but so far there is no consensus on this matter (Weigelt et al., 2010). The basal-like and claudin-low subtypes can be summarized as triple negative breast cancer (TNBC), because they are both negative for ER, PR and HER2. Note that TNBC is a clinicopathological definition that overlaps to about 80 % with the basal-like molecular subtype defined by its gene expression. The remaining 20 % contain claudin-low and some special histological types. A recent study discovered that TNBC actually seems to consist of six transcriptomically distinct subtypes (Lehmann et al., 2011). Even more recently, Curtis and colleagues have shown, through copy number alteration (CNA) and expression analysis, that there are at least ten distinct molecular subtypes (Curtis et al., 2012). However, for simplicity I will refer to the basal-like and claudin-low subtypes as TNBC in the course of the thesis. Taken together, these studies underline the high complexity of BC and stress that BC is not a single disease, but many diseases occurring in the same tissue. Subtyping is important, because each subtype has a different prognosis and responds differently to treatment strategies (Sørlie et al., 2001; Kennecke et al., 2010; Polyak and Metzger Filho, 2012; Fallahpour et al., 2017; Howlader et al., 2018).

Following the initial gene expression studies, the established IHC-based markers ER, PR, and HER2 as well as marker of proliferation Ki-67 (KI67), the basal cytokeratins 5 and 6 (KRT5/6), and epidermal growth factor receptor (EGFR, also known as HER1 or ERBB1) were successfully linked to the molecular subtypes by several studies, enabling faster assignment of samples to the molecular subtypes without the need for gene expression microarrays (Kennecke et al., 2010; Lehmann et al., 2011; Eroles et al., 2012). However, the matching of the IHC markers with the molecular profiles is merely an approximation due to the low number of markers used. Therefore, the IHC marker patterns (see Table 1-2) apply to the majority of tumors of each molecular subtype, however it may happen that the subtype determined by gene expression microarray may be different from the one assigned by IHC in some rare cases. Regarding the frequency with which the subtypes occur, researchers have observed that the LumA subtype is by far the most frequent one accounting for 50-60 % of breast cancer cases (Eroles et al., 2012; Polyak and Metzger Filho, 2012). In contrast, LumB and TNBC make up around 10-20 %, while the Her2-enriched type represents 10-15 % (see Table 1-2).

Table 1-2 Immunohistochemical marker patterns observed in BC subtypes. Adapted from (Eroles et al., 2012). +/- means that the respective marker can be either positive or negative. ER and PR are summarized in the HR column. Expression of either of these two markers is sufficient for HR positivity.

Subtype	HR	HER2	CK5/6	EGFR	KI67	Frequency
LumA	+	-	-	-	≤14 %	50-60 %
LumB HER2$^-$	+	-	-	-	>14 %	10-20 %
LumB HER2$^+$	+	+	-	-	any	
HER2-enriched	-	+	+	-	>14 %	10-15 %
TNBC (basal-like)	-	-	+	+	>14 %	10-20 %
TNBC (claudin-low)	+/-	-	+	-	≤14 %	12-14 %
Normal breast-like	-	-	-	-	≤14 %	5-10 %

ER, PR, HER2, and KI67 status are determined in clinical routine diagnostics, which enables classification of patients into four subtypes according to the surrogate definitions of intrinsic subtypes from the 12th St. Gallen consensus conference (Goldhirsch et al., 2011): LumA, LumB, HER2-enriched, and TNBC. The St. Gallen criteria are based on work by Nielsen and colleagues as well as Cheang and colleagues (Nielsen et al., 2004; Cheang et al., 2009). There is also evidence to suggest that these four types are in fact the main ones, while all others might merely be heterogeneities occurring within these four main types (Cancer Genome Atlas Network, 2012).

There is evidence that the molecular subtypes represent abnormal cells stuck at distinct stages of mammary gland differentiation (Lim et al., 2009; Prat and Perou, 2009). According to this concept, the more aggressive TNBC and HER2-enriched subtypes, which display more mesenchymal phenotypes, are derived from stem cells or progenitor cells, while the less aggressive luminal types are derived from differentiated luminal or myoepithelial cells (see Figure 1-2). The gradient from luminal to mesenchymal phenotype is correlated with prognosis, with the basal-like type.

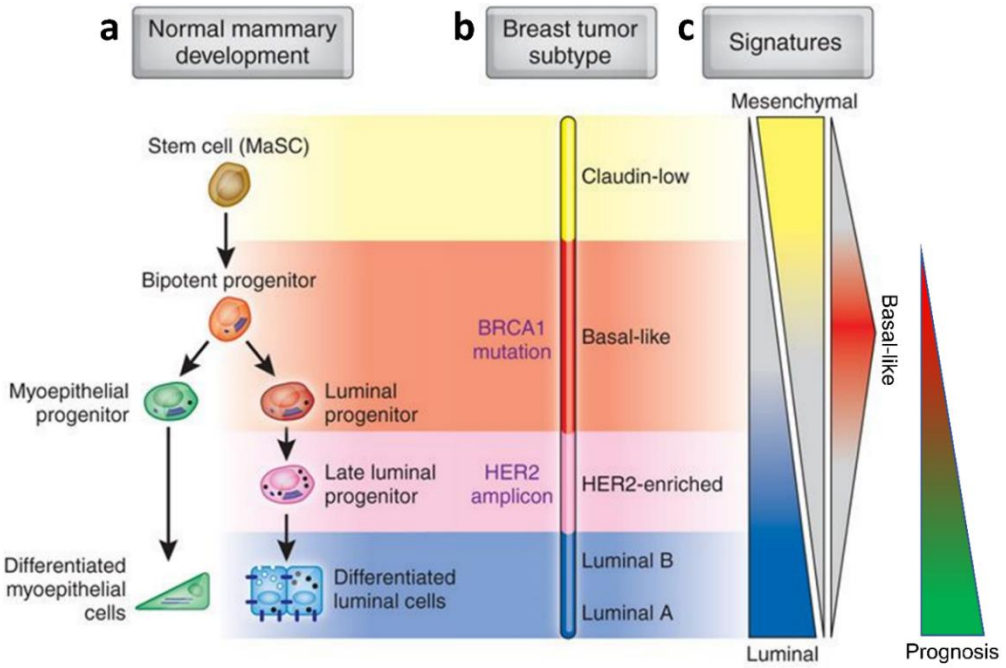


Figure 1-2 Relation of normal mammary development with BC subtype. The scheme illustrates how cells involved in normal mammary gland development may be related to the molecular intrinsic subtypes of BC. (a) Subpopulations of normal breast tissue and potential cells of origin for the intrinsic subtypes of breast cancer; these cells may represent a stage of developmental arrest for a tumor with an origin earlier in the differentiation hierarchy or, alternatively, transformation of a cell type at one specific stage of development. (b) The various breast tumor subtypes molecularly compared to subpopulations from normal breast tissue. (c) The defining expression patterns of luminal, mesenchymal or claudin-low, and basal-like cells. These molecular patterns may be best represented as gradients of expression, as opposed to a discrete 'on' or 'off' state of expression. The approximate locations of the differentiation blocks imposed by *BRCA1* loss and *HER2* amplification are suggested by their locations in the differentiation hierarchy. A higher degree of differentiation is linked to a better prognosis (basal-like and claudin-low combined as TNBC). Green = good prognosis, red = bad prognosis. Image and caption adapted from (Prat and Perou, 2009), which is based on (Lim et al., 2009).

1.1.4 Treatment

The choice of treatment for BC depends mainly on the stage of a patient's PT. EBC requires a different treatment approach than mBC, because eBC is considered curable, in contrast to mBC, which still cannot be cured with current methods (Harbeck and Gnant, 2017). Apart from the stage, the subtype and grading of the patient's PT also influence which treatment will result in the

best possible outcome. Due to the high heterogeneity of BC, treatment must be individually fine-tuned for each patient.

In eBC the treatment plan is based on the molecular subtype and locoregional tumor load (i.e. in the breast and adjacent lymph nodes). Here, I will focus on the luminal A and B subtypes only. Local treatment by breast conserving surgery and/or radiotherapy is usually preferred as a primary treatment. Alternatively, neoadjuvant systemic treatment is now considered standard of care for situations in which breast conservation surgery is not possible due to tumor size (Schmidt, 2014). In case of the LumA subtype, the primary surgery is followed up with endocrine therapy (ET), while LumB ($HER2^-$) patients usually first receive chemotherapy (CT) after primary surgery prior to ET (Harbeck and Gnant, 2017). $HER2^+$ LumB cancers are also first treated with CT followed by a combination of targeted anti-HER2 therapy and ET.

In ET, the selective ER modulator Tamoxifen decreases growth of BC cells by inhibiting binding of estrogen to its receptor (Harper and Walpole, 1967; Cole et al., 1971; Wang et al., 2004),over a course of five years (Early Breast Cancer Trialists' Collaborative Group, 2011). Ovarian suppression drugs together with an aromatase inhibitor may be used together with Tamoxifen to enhance efficacy (Pagani et al., 2014; Francis et al., 2015). As ET often causes side effects in the bones, bisphosphonates like zoledronic acid are administered in addition to prevent bone loss (Gnant et al., 2008; Coleman et al., 2013). Newer data also suggest bisphosphonates might even prevent bone and other metastases (Gnant and Clézardin, 2012). For targeted anti-HER2 therapy Trastuzumab is administered over the course of one year (Goldhirsch et al., 2013; Pivot et al., 2013). Trastuzumab may be combined with Pertuzumab for a dual HER2 blockade (Gianni et al., 2012) in the neoadjuvant setting. Both Trastuzumab and Pertuzumab are monoclonal antibodies that bind to the HER2 receptor on the cell membrane leading to growth inhibition of tumor cells (Carter et al., 1992; Franklin et al., 2004; Hudis, 2007; Lamond and Younis, 2014). Standard agents for CT in eBC are anthracyclines (e.g. Epirubicin, Doxorubicin), taxanes (e.g. Paclitaxel, Docetaxel), cyclophosphamide, and 5-fluorouracil. Adjuvant or neoadjuvant CT should only be considered if the estimated relapse risk of a patient is >10 % over the course of ten years, since low risk also means low benefit from CT (Harbeck and Gnant, 2017).

Despite being considered incurable with current treatment strategies (Harbeck and Gnant, 2017), a small number of long-term survivors of mBC have been observed in the past. Greenberg and colleagues observed that 16.6 % of mBC patients treated with adjuvant CT achieved complete responses and 3.1 % remained in complete response for more than 5 years (Greenberg et al., 1996). Therefore, it seems possible that metastatic disease may one day become a chronic disease controlled by sequential therapies. However, nowadays the main goals of therapy of mBC are still prolongation of survival, maintenance of quality of life, and palliation of symptoms (Harbeck and Gnant, 2017). Since each patient has an individual history of previous treatments and preferences, treatment of mBC is usually more individualized than in eBC. In general, systemic therapy is the primary choice for mBC and locoregional therapy (surgery and/or radiation) may be added in specific situations, e.g. on symptomatic metastases. In case of bone metastasis, bone-modifying drugs like bisphosphonates, which reduce the half-life of osteoclasts, or Denosumab, a monoclonal antibody that prevents osteoclast development by RANKL inhibition (Pageau, 2009), are considered as a standard maintenance therapy.

The main criterion for treatment decisions in mBC is how severe a patient's symptoms are. If a rapid response is needed, CT is the best choice in any BC subtype, as it kills all proliferating cells (Harbeck and Gnant, 2017). In case of $HER2^+$ cancer (ER^+ or ER^-), a targeted anti-HER2 therapy is performed in combination with CT using docetaxel and trastuzumab plus pertuzumab (Swain et al., 2015). In contrast, if the disease is progressing slowly, treatment is performed according to the molecular subtype as explained above. Once HR^+ cancers stop responding to ET, CT represents

the next line of treatment. ET resistance is more frequently observed in LumB BC, resulting in a worse prognosis compared to LumA (Szostakowska et al., 2019). However, so far it is not clear why LumB BC is more prone to becoming treatment resistant than the histopathologically similar LumA type.

1.1.5 Prognosis and survival

The most relevant prognostic factors of patient outcome are personal history of BC, higher grade, and higher TNM-stage (Paredes-Aracil et al., 2017). Additionally, the PT's molecular subtype determined by the histological receptor status can also provide valuable insight into how a patient will fare (Sørlie et al., 2001; Fallahpour et al., 2017).

Looking at different TNM stages in luminal ($HER2^-$), $HER2^+$ (ER^+ and ER^-), and TNBC tumors, Polyak and Metzger Filho discovered that both the stage and the subtype influence the survival of patients (Polyak and Metzger Filho, 2012). Regarding the BC subtype, the luminal ($HER2^-$) and $HER2^+$ types displayed good 5-year and 10-year survival results in stages I and II (see Table 1-3) thanks to ET and anti-HER2 therapy. The overall survival (OS) rates only started to decrease a lot at stage III and dropped drastically at stage IV (metastatic disease). Therefore, the metastatic state is one of the main prognostic factors of bad outcome. Compared to the other two subtypes, TNBC patients displayed a worse outcome across all four stages, which is most likely caused by the lack of viable treatment options. Additionally, the presence of disseminated cancer cells (DCC) in the BM is also robustly correlated with a worse prognosis in M0 patients (Early Breast Cancer Trialists' Collaborative Group, 2005; Banys et al., 2014).

Table 1-3 Overall survival of breast cancer subtypes depending on stage. Table adapted from (Polyak and Metzger Filho, 2012). * Preinvasive stage, ** estimated overall survival (OS) using HER2-targeting therapies

Subtype	Frequency (%)	Stage	5-year OS (%)	10-year OS (%)
DCIS*	NA	0	99	98
Luminal ($HER2^-$)	70	I	98	95
		II	91	81
		III	72	54
		IV	33	17
$HER2^+$ **	20	I	98	95
		II	92	86
		III	85	75
		IV	40	15
TNBC	10	I	93	90
		II	76	70
		III	45	37
		IV	15	11

Fallahpour and colleagues stratified their data differently. They separated luminal tumors into LumA and LumB, but did not take stages into account across all subtypes. They found robust discrepancies in OS between the four routinely used molecular subtypes (Figure 1-3; Fallahpour et al., 2017). Interestingly, despite their similar histopathological phenotype, LumA and LumB diverge strongly regarding survival (see Figure 1-3). On closer examination, LumB usually displays lower expression of ER, harbors *TP53* mutations more often, is associated with a higher metastatic relapse rate, and a higher propensity for development of anti-ER treatment resistance compared to LumA (Kittaneh et al., 2013; Szostakowska et al., 2019). However, the exact mechanisms underlying the higher rate of metastasis and treatment resistance remain poorly understood.

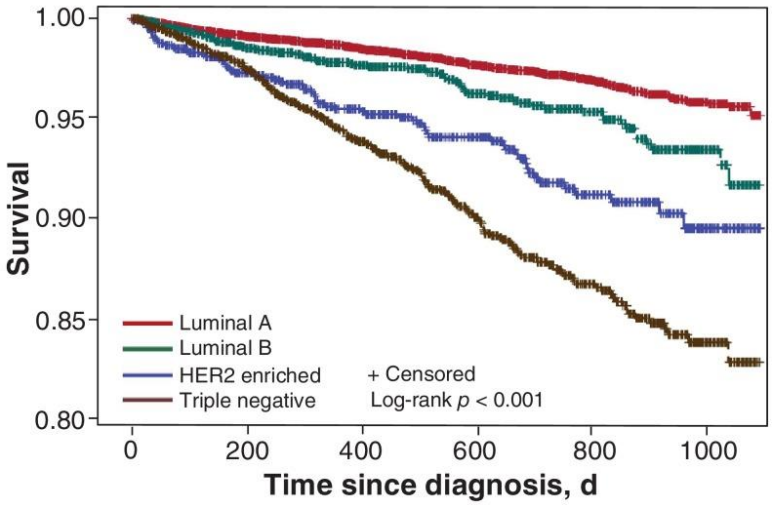


Figure 1-3 Overall survival of the four routinely used molecular subtypes. Kaplan-Meier plot of overall breast cancer survival over 1000 days stratified by molecular subtype. Treatment of patients was not included in the analysis. Figure from (Fallahpour et al., 2017).

1.2 Metastasis

The term "metastasis", which was coined almost 200 years ago (Recamier, 1829; Talmadge and Fidler, 2010), describes the multistep process of cells of a PT spreading to distant organs, surviving there, and finally forming a secondary tumor (Fidler, 1970; Chambers et al., 2002; Valastyan and Weinberg, 2011). Metastasis is considered highly ineffective, because at each step (outlined below) the probability of success decreases progressively (Luzzi et al., 1998). Luzzi and colleagues observed that less than three percent of cancer cells survived well enough to form micrometastases (4-16 cells), although 87 % of cells survived circulation and extravasation. In the end, only 0.02 % of injected cancer cells managed to form a macrometastasis (Luzzi et al., 1998). Interestingly, experiments in mice have shown that even normal mammary epithelial cells injected into a mouse's bloodstream could survive in the lungs in surprisingly large numbers (about 1.2 in 10,000 injected cells) for prolonged periods of time and still formed mammary glands when re-implanted into mammary fat tissue (Podsypanina et al., 2008), while only very few metastases arose from millions of injected cultured cancer cells, because the later steps become progressively harder for cancer cells to cope with (Fidler, 1970; Luzzi et al., 1998; Chambers et al., 2002; Klein, 2008). This indicates that survival in foreign tissues is not the only problem that cancer cells must face. Valastyan and Weinberg formulated a total of seven steps required for successful metastasis, namely the following (Figure 1-4; Valastyan and Weinberg, 2011): (1) local invasion into neighboring tissues by breaching of the basement membrane, (2) intravasation into the lumina of blood or lymphatic vessels, (3) survival in the circulation despite a variety of different stresses (e.g. shear stress, lack of anchorage to extracellular matrix [ECM]), (4) and arrest at a distant organ site, whereby each type of carcinoma seems to have a preference for a limited set of target organs (Fidler, 2003). During their trip through the circulation the cells are called circulating cancer cells (CCC). After arrest at a distant site, (5) cancer cells must undergo extravasation by crossing the layers of endothelial cells and pericytes lining the vessel lumen to enter the distant tissue and (6) survive in the foreign microenvironment (ME) where these CCCs are now referred to as DCCs. Finally, the last step of the metastatic cascade is (7) to carry out metastatic colonization (Valastyan and Weinberg, 2011). In order to perform intra- and extravasation, a cancer cell must likely first undergo partial epithelial to mesenchymal transition (EMT) at the PT site (Nieto et al., 2016; Cho et al., 2019), which renders them more invasive

(Thiery et al., 2009), followed by mesenchymal to epithelial transition (MET) at the target site to facilitate proliferation and thereby survival (Weinberg, 2007). Regarding systemic spread, the hematogenous pathway appears to be the predominant one (Gupta and Massagué, 2006). Once extravasation has been successful and the DCCs have found a way to survive in a foreign ME, it seems that the vast majority of them do not manage to grow in the target organ, but instead persist as single cells (SC) or microcolonies in a state of long-term dormancy while retaining viability without gain or loss of cells numbers (Fisher and Gebhardt, 1978; Chambers et al., 2002; Aguirre-Ghiso, 2007). Studies suggest that this may be caused by impairment of proliferation due to incompatibilities with the foreign ME through lack of certain signaling molecules (Barkan et al., 2008; Shibue and Weinberg, 2009; Barkan et al., 2010a).

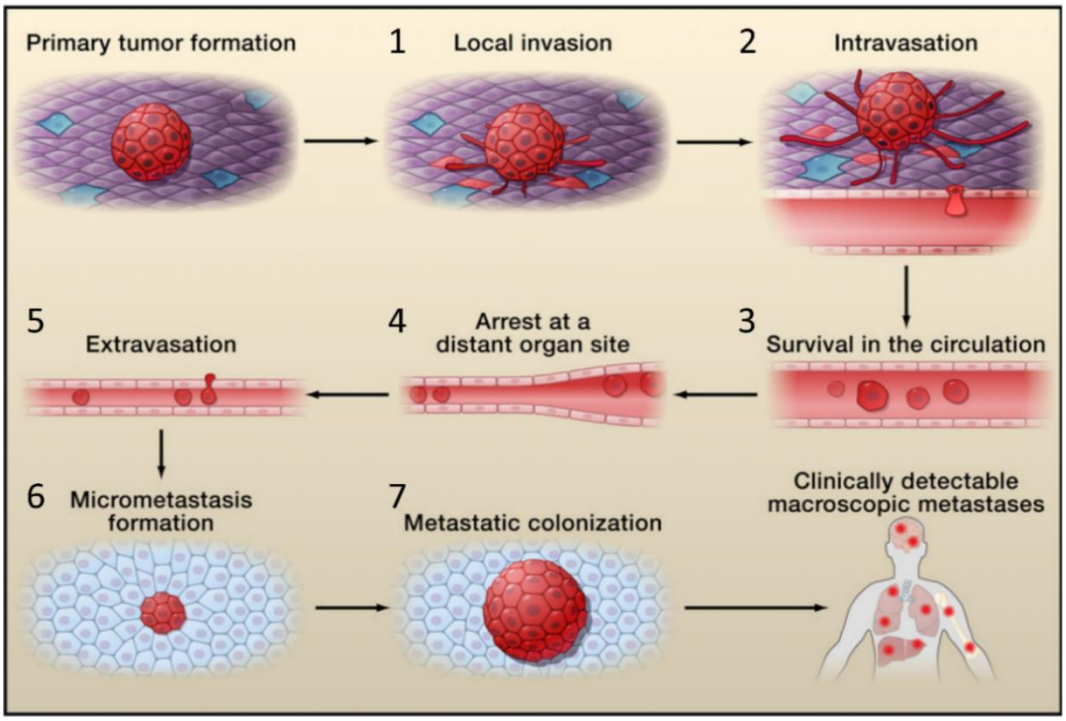


Figure 1-4 The metastasis cascade. Schematic illustration of the necessary steps from PT formation to clinically detectable metastasis. Figure adapted from (Valastyan and Weinberg, 2011).

As mentioned above, many types of cancer preferentially metastasize to specific organs, despite the fact that they could spread almost anywhere via the hematogenous or lymphatic vessels (Fidler, 2003). For example, BC metastases preferentially manifest in bone marrow (BM; 41 %), lungs (22 %), liver (7 %), and brain (7 %) according to Berman and colleagues (Berman et al., 2013). However, the target site preference varies depending on the BC subtype (Gong et al., 2017; Xiao et al., 2018). This phenomenon was first described in 1889 by Stephen Paget for different cancer entities and later became what is known today as the "seed and soil" hypothesis (Paget, 1889). This hypothesis states that metastatic cells (the seed) can only survive and grow within a certain ME (the soil; Chambers et al., 2002). It has been proposed by Psaila and Lyden that PTs can release systemic signals that lead to establishment of a pre-metastatic niche at metastatic sites by recruitment of hematopoietic cells, which migrate to the target organ from the BM to prepare the ME for the arriving DCCs (Psaila and Lyden, 2009). James Ewing challenged the seed and soil hypothesis in 1928. He suggested that the patterns of preferential metastasis could be sufficiently explained by circulatory patterns between PT and the target organs (Ewing, 1928). However, the two theories are not mutually exclusive (Chambers et al., 2002) and newer studies by Weiss and colleagues suggest that both factors play a role in determining where a cancer metastasizes, because only up to 66 % of metastases could be explained by circulatory patterns (Weiss et al., 1986; Weiss and Harlos, 1986; Weiss et al., 1988; Weiss, 1992; Chambers et al., 2002).

1.3 Early dissemination

Metastasis is observed in about 6 % of BC patients at diagnosis and later develops in 20-50 % of patients, who were initially diagnosed with eBC (Chambers et al., 2002; O'Shaughnessy, 2005; Cardoso and Castiglione, 2009; Lu et al., 2009). Despite the fact that about 90 % of cancer-related deaths are caused by metastases and not the PT (Bendre et al., 2003; Fidler, 2003; Weigelt et al., 2005; Loberg et al., 2007; Redig and McAllister, 2013), no significant progress has been made in the search for a cure for metastatic disease in the last decades and knowledge is fragmented (Kozłowski et al., 2015).

The main reason for this problem is that the classical linear or late dissemination model (Foulds, 1954; reviewed by Klein, 1998; Weinberg, 2008; Valastyan and Weinberg, 2011) is still stuck in the heads of many researchers. This model states that metastasis is a late event occurring only once the PT has grown large enough to acquire mutations permitting cells to leave the tumor and found secondary tumors at distant sites in the body. However, evidence has been accumulating that this is not the whole story, which is why a slow paradigm shift towards the early progression model is occurring (Klein, 2008). The early progression model is founded on the observation that many human cancers start metastasizing years before the PT has been detected (Friberg and Mattson, 1997). This implies that some not yet fully malignant tumor cells must have acquired the ability to leave the PT and survive elsewhere in early stages of tumor development. The early dissemination model states that cancer cells may leave the PT already before the PT can be clinically detected, settle down in a distant organ, and evolve independently of the PT (Klein, 2008). In fact, PTs were shown to differ genetically from metastases in up to 85% of patients (Stoecklein and Klein, 2010), suggesting that early dissemination is a widespread phenomenon. This has dramatic implications for patient treatment, because metastases may not respond to a treatment that was selected according to characteristics of the PT, ultimately leading to the high mortality rate of mBC patients still observed today (Bendre et al., 2003; Fidler, 2003; Weigelt et al., 2005; Loberg et al., 2007; Redig and McAllister, 2013).

Since early and late dissemination mechanisms are not mutually exclusive, both were combined into the parallel progression model (Figure 1-5; Klein, 2009), the idea of which already emerged in the 1950s (Collins et al., 1956). According to parallel progression, early DCCs (eDCC) disseminate when the PT has a size of 1-4 mm, settle down in a distant organ, and lie mostly dormant while slowly accumulating further mutations until they are finally able to grow into a macroscopic metastasis. Hanahan and Weinberg's seminal publication "Hallmarks of Cancer" defines six distinct traits each cancer must develop, in order to survive and grow: evasion of apoptosis, self-sufficiency in growth signals, insensitivity to anti-growth signals, sustained angiogenesis, limitless replicative potential, and tissue invasion (Hanahan and Weinberg, 2000, 2011). It is plausible that these hallmarks also apply to metastases. As mentioned earlier, survival of cells in foreign MEs is not a rare event (Podsypanina et al., 2008), therefore it is likely that many eDCCs survive at distant locations as micrometastases. There, it probably takes the eDCCs many years to develop the six hallmarks of cancer postulated by Hanahan and Weinberg, which would explain why many cancers relapse so late. In BC, for example, metastatic relapse may still occur at least 15 years after initial diagnosis (Early Breast Cancer Trialists' Collaborative Group, 2005).

The dormant state of DCCs in a foreign ME, in which they are undetectable to routine diagnostics, is often referred to as minimal residual disease (MRD; Klein, 2003). During MRD, the PT releases factors that can prepare pre-metastatic niches for the DCCs (Psaila and Lyden, 2009) and may also stimulate the DCCs themselves. Since DCCs most likely lie dormant in many organs in pre-metastatic niches prepared by the PT (Psaila and Lyden, 2009), several metachronous metastases may occur in each patient (see Figure 1-5). In rare cases, highly aggressive eDCCs may result in detectable macrometastases at diagnosis despite the PT still being in an early stage (Klein, 2009).

This phenomenon may explain why mBC is observed in only 6 % of patients at the time of PT diagnosis. It is assumed that BC-derived DCCs are mostly dormant in the BM (and other distant organs), meaning that they are in a kind of non-proliferative and non-productive hibernation-like state until they are re-activated by some external stimulus and begin to grow into macrometastases (Aguirre-Ghiso, 2007; Hüsemann et al., 2008; Klein, 2011; Hosseini et al., 2016; Yadav et al., 2018). Judging from the long relapse periods observed in BC (Early Breast Cancer Trialists' Collaborative Group, 2005), DCCs can survive in this dormant state for many years after primary surgery before waking up (Kang and Pantel, 2013). Several studies concluded that integrin signaling by the ECM is involved in the entry into and escape from dormancy (Barkan et al., 2008; Barkan et al., 2010b; Barkan et al., 2010a). Unfortunately, the exact mechanisms that cause DCCs to grow into macrometastases are still poorly understood (Kang and Pantel, 2013). Connected to MRD and dormancy is the theory of cancer stem cells (CSC), which are thought to be vital for initiating and sustaining a cancer (Dick, 2003; Morrison et al., 2008), similar to normal stem cells which initially give rise to an organ and constantly renew its tissue (Yoo and Hatfield, 2008). However, the concept of CSCs is still highly controversial and there is evidence to suggest that all cancer cells are able to sustain a tumor, which is why the CSC model may need to be revised (Yoo and Hatfield, 2008).

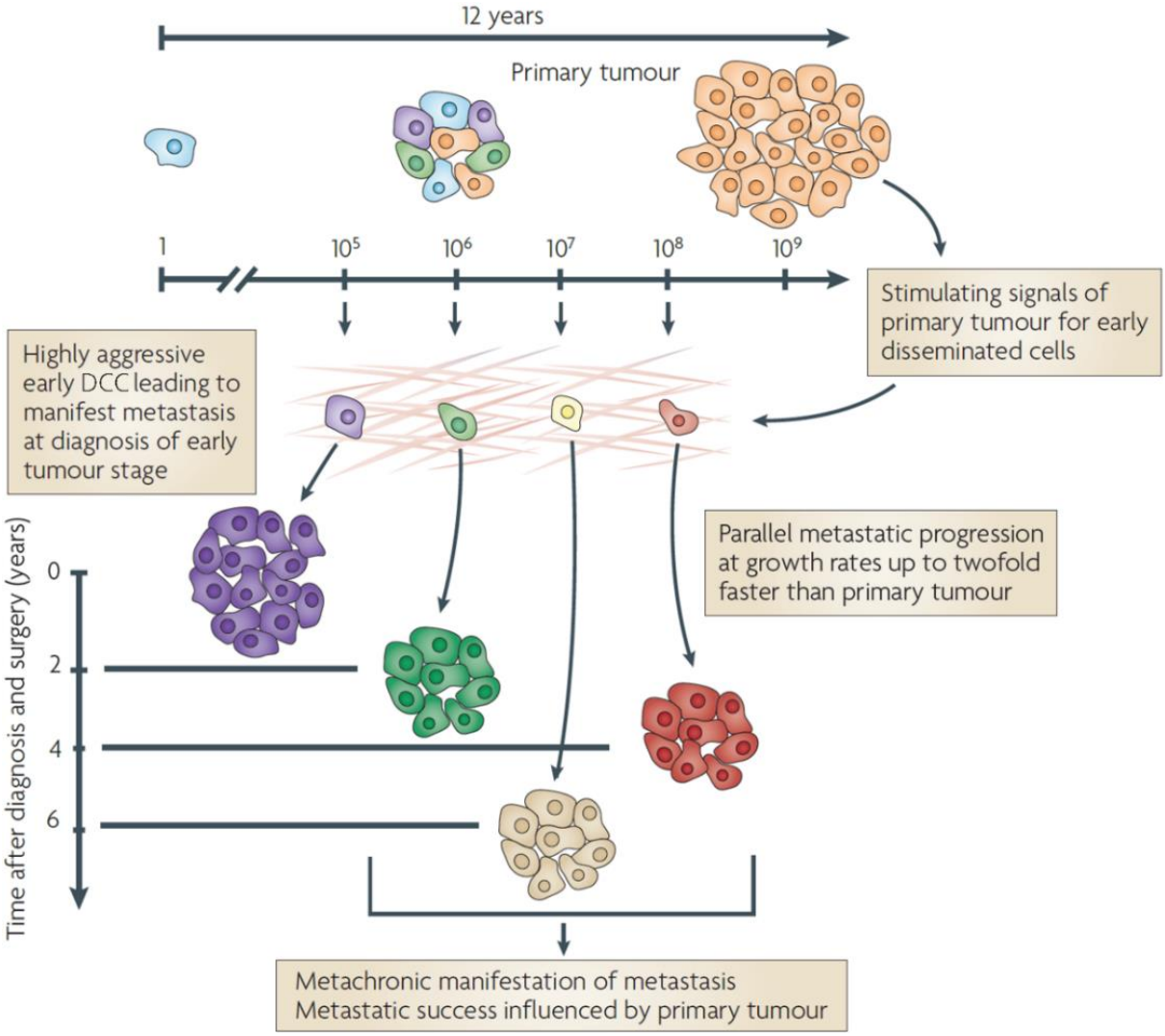


Figure 1-5 Parallel progression model. In the parallel progression model, several waves of disseminated tumor cells (DCCs) may disseminate before diagnosis and may progress in parallel at different rates in different organs. Factors secreted by the PT may stimulate colonization and account for the relationship of tumor size and probability of metastatic outgrowth. Figure and caption adapted from (Klein, 2009).

1.4 Detection and analysis of DCCs from patients

DCCs represent promising targets for investigation of the mechanisms of early dissemination, because their presence in the BM has been shown to have prognostic value for patients by the majority of available studies, underlining their important role in metastatic relapse (Cote et al., 1991; Harbeck et al., 1994; Schindlbeck et al., 2011; Domschke et al., 2013; Banys et al., 2014; Hartkopf et al., 2019). Since the BM represents the main target for BC dissemination, because it contains no epithelial cells, and because it is more easily accessible than lungs, liver, or brain, it represents the best organ to detect and isolate DCCs. Studies have found DCCs in the BM of about 20-30 % of eBC patients (Schlimok et al., 1987; Braun et al., 2005), where they likely undergo gradual somatic progression according to the parallel progression model (Klein, 2009).

For these reasons, BM aspirates are utilized for detection and isolation of DCCs. The applied method for isolation and subsequent whole genome amplification (WGA) and whole transcriptome amplification (WTA) of single DCCs was developed by our research group (Klein et al., 1999; Klein et al., 2002; Stoecklein et al., 2002; Klein et al., 2003; Hartmann and Klein, 2006). The WTA is a multistep process designed as a dual-omics approach to separate the messenger RNA (mRNA) from the genomic DNA (gDNA) of a single cell prior to global transcriptome and genome amplification, respectively (Klein et al., 2002; Hartmann and Klein, 2006). Building on the WTA, the WGA of the previously isolated supernatant is another multistep process that enables deterministic, i.e. reproducible, amplification of a cell's genome which is beneficial for comparative genomic hybridization (CGH) methods (Klein et al., 1999; Stoecklein et al., 2002). The whole workflow from the BM aspirate to downstream analysis is summarized in Figure 1-6. The details of the protocols are described in chapters 3.1 and 3.2. The WTA and WGA have been commercialized under the names *Ampli1* WTA WGA, but they will be referred to as WTA and WGA.

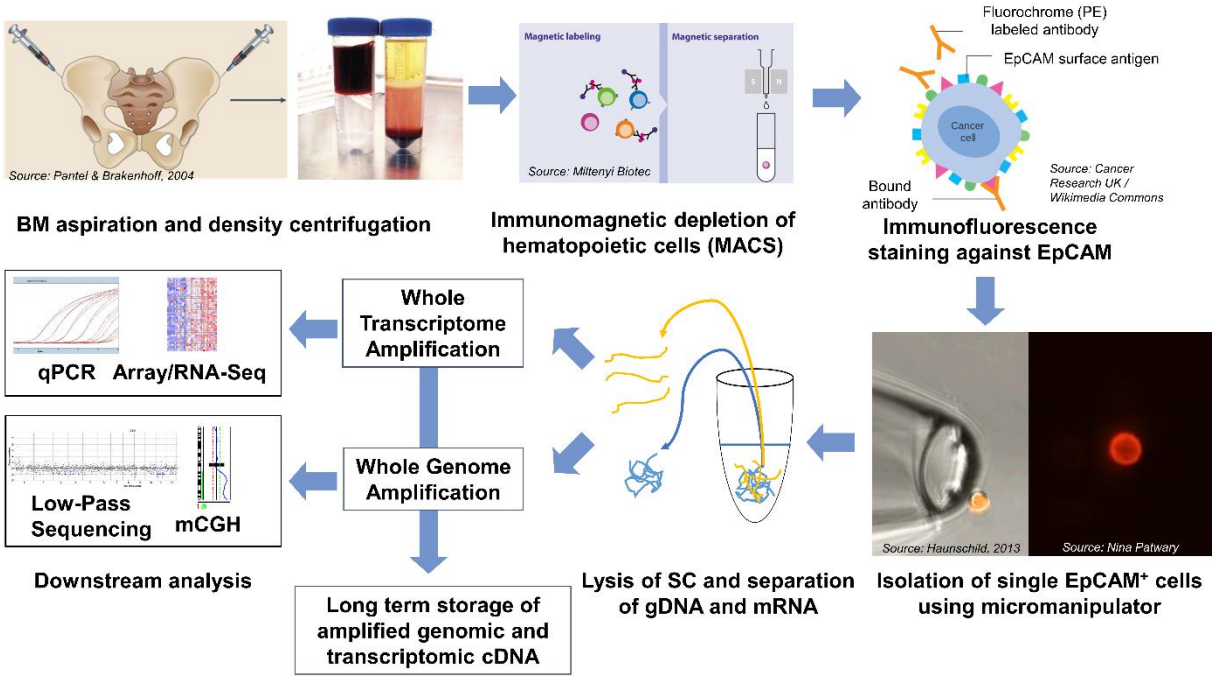


Figure 1-6 Workflow for isolation of single DCCs from bone marrow aspirates. The schematic illustrates how BM samples are processed, resulting in simultaneous isolation of genomic DNA and mRNA from the same DCC, and how the obtained material can further be used. Indicated components of the schematic were adapted from publications (Pantel and Brakenhoff, 2004; Haunschild, 2013), websites (Miltenyi Biotec [*https://www.miltenyibiotec.com/DE-en/products/macs-cell-separation/macs-cell-separation-strategies.html*; date of image retrieval: 06.08.2019] and Cancer Research UK [*https://www.cancerresearchuk.org/about-cancer/cancer-in-general/treatment/targeted-cancer-drugs/types/monoclonal-antibodies;* date of image retrieval: 06.08.2019]) or colleagues (Nina Patwary). MACS = magnetic activated cell sorting

BC – like all carcinomas – is derived from epithelial cells (Kirkham and Lemoine, 2001). Therefore, BC DCCs can be identified using specific epithelial markers like different cytokeratins (CK) or the epithelial cell adhesion molecule (EpCAM; Schlimok et al., 1987; Malzahn et al., 1998; Woelfle et al., 2005; Keller et al., 2019). We use EpCAM instead of CK as a marker, because – unlike CK - EpCAM is a surface marker, which allows us to detect living DCCs. This way we are able to isolate high quality mRNA, which would not be possible using CK antibodies. Unfortunately, EpCAM is not completely specific for DCCS, because evidence from healthy BM donors (HD) suggests that there is a small population of erythroid progenitor cells in the BM, which also expresses EpCAM (Bühring et al., 1996; Lammers et al., 2002; Gužvić et al., 2014). Therefore, there is a need to distinguish true DCCs from these HD-derived confounding $EpCAM^+$ non-cancer cells (NCC).

1.5 Micro RNAs and their role in breast cancer

Micro RNAs (miRNA) are a class of small, single stranded RNA molecules with a length of 19-24 bp that were first described in 1993 in *C. elegans,* but were also soon discovered in humans (Lee et al., 1993; Pasquinelli et al., 2000; Bhaskaran and Mohan, 2014). They are evolutionarily conserved and function as post-transcriptional regulators of translation by mRNA cleavage or repression (Felekkis et al., 2010). Moreover, they are important for animal development and also associated with a wide variety of diseases including cancers like gastric, liver, hepatocellular or prostate cancer (Calin et al., 2004; Wienholds and Plasterk, 2005; Lu et al., 2008; Lee and Dutta, 2009; Peng and Croce, 2016; Tan et al., 2018).

MiRNAs are transcribed by RNA polymerase 2 or 3 (Pol II, Pol III) from specific miRNA genes, each of which may contain several individual miRNA sequences (Figure 1-7; Peng and Croce, 2016; Mandujano-Tinoco et al., 2018). The resulting transcript is called pri-miRNA and is marked by methyltransferase-like 3 (METTL3) for processing by the Drosha/DGCR8 complex that turns the pri-miRNA into a so-called pre-miRNA. In this process, the individual miRNAs contained in the single pri-miRNA are separated into individual pre-miRNAs (not depicted in Figure 1-7). The pre-miRNA molecule is then exported from the nucleus via Exportin 5/RANGTP into the cytoplasm where the Dicer/TRBP complex removes the loop of the pre-miRNA generating a miRNA:miRNA duplex molecule. This duplex then serves as substrate for the assembly of the RNA-induced silencing complex (RISC) consisting – among others – of argonaute (AGO) proteins. One of these proteins' functions is to select one of the strands of the duplex for integration into the RISC and degrade the other. In the RISC, the integrated miRNA serves as a guide that identifies the intended target mRNA. Depending on whether the miRNA matches the target perfectly or displays mismatches, the mRNA is either cleaved and degraded or its translation is repressed, respectively. Additionally, recent studies have discovered that mature miRNAs may also serve as ligands that directly bind to toll-like receptors (TLR) to activate downstream signaling (Fabbri et al., 2012; He et al., 2013).

MiRNA dysregulation was first described in BC in 2005 by Iorio and colleagues, who reported mir-125b, mir-145, mir-21, and mir-155 as the most deregulated miRNA species (Iorio et al., 2005). Since then the number of publications covering miRNAs in BC has skyrocketed to about 600 in 2016 (Mandujano-Tinoco et al., 2018). The main reason for miRNA dysregulation is up- or down-regulation of proteins involved in miRNA biogenesis. Depending on the BC subtype, different steps of the miRNA biogenesis pathway may be dysregulated. For example, down-regulation of Drosha is associated with TNBC and generally with higher grade, tumor size and metastasis (Poursadegh Zonouzi et al., 2017), DGCR8 up-regulation is connected to high KI67 expression, ER-positivity, high grade, and metastasis of invasive ductal BC (Fardmanesh et al., 2016), while its down-regulation is linked to ER-negative BC without affecting patient outcome (Dedes et al., 2011), and

overexpression of nuclear export components like Exportin 5 is correlated with invasiveness and poor prognosis (Vaidyanathan et al., 2016), just to name a few. In their review, Mandujano-Tinoco and colleagues also argue that miRNA expression patterns are linked to the six hallmarks of cancer proposed by Hanahan and Weinberg in the year 2000 (Hanahan and Weinberg, 2000; Mandujano-Tinoco et al., 2018). Last but not least, some miRNAs were reported to predict resistance of BC to the most common systemic treatments (Campos-Parra et al., 2017), which makes them even more interesting in the context of LumB BC. However, the role of miRNAs in therapy resistance is still incompletely understood and requires further research.

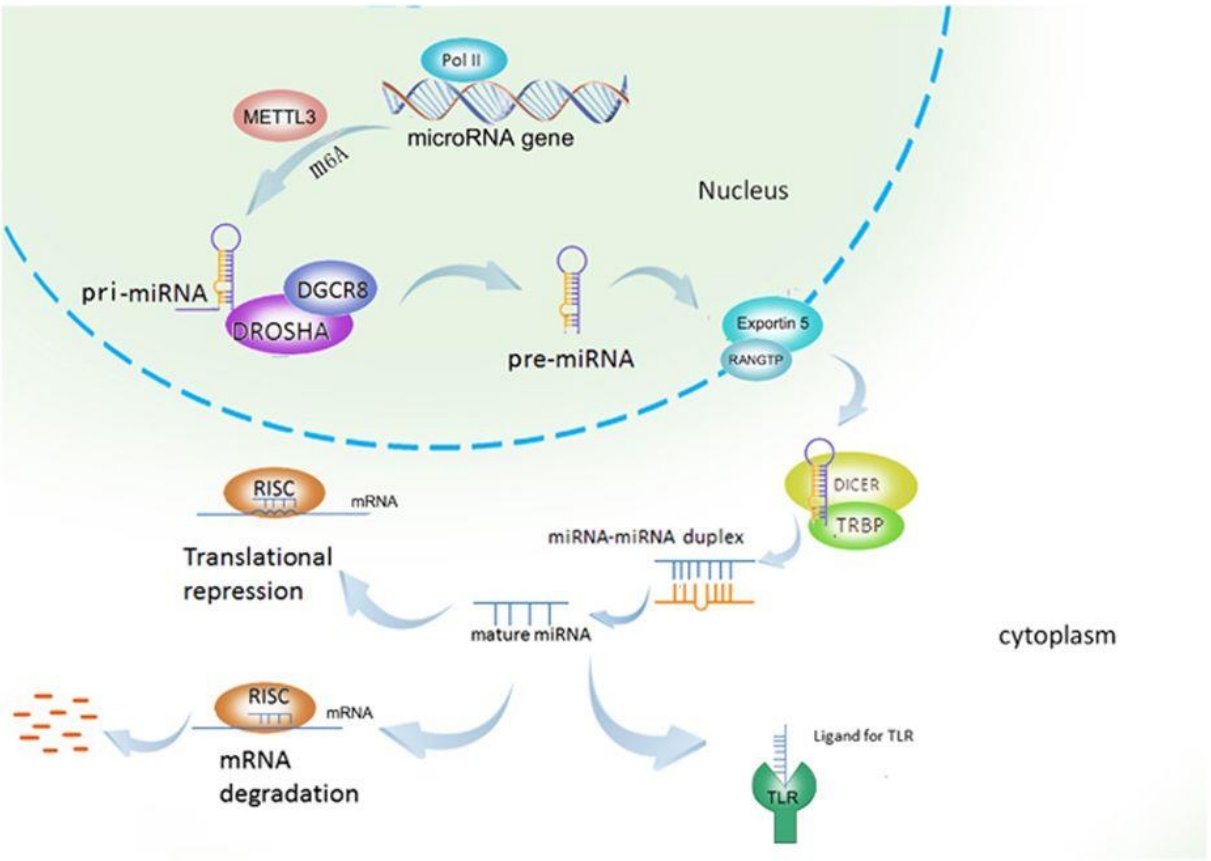


Figure 1-7 Overview of canonical biogenesis and functional mechanisms of miRNA. MiRNAs are transcribed from miRNA genes in the form of relatively long pri-miRNA molecules, which are then processed into pre-miRNA by the Drosha/DGCR8 complex inside the nucleus. Subsequently, the pre-miRNA is shuttled from the nucleus into the cytoplasm via Exportin 5/RANGTP where it is further processed (loop removal) into a miRNA:miRNA duplex by the Dicer/TRBP complex. Next, the RNA-induced silencing complex (RISC), partially consisting of argonaute (AGO) proteins, is assembled. In this process the miRNA:miRNA duplex is unwound, one of the miRNA strands (miRNA*) is degraded and the other single-stranded miRNA is loaded into the RISC. Strand selection is done by AGO and depends on 5' stability of the strand. Finally, RISC either cleaves (perfect match) or represses (imperfect match) the target mRNA. Apart from the mRNA regulating function, the mature miRNA can also serve as a ligand for Toll-like receptors (TLR) to trigger downstream signaling pathways. Figure from (Peng and Croce, 2016).

1.6 Current methods to quantify miRNAs from single cells

Since miRNAs were first described, many different approaches have been developed to detect them. However, most of them were not intended to profile miRNAs from SC amounts of RNA. Appropriate methods that allow quantification of miRNAs in SCs have only recently been established. They are all based on one of two underlying technologies: quantitative reverse transcription PCR (qRT-PCR) or RNA-Seq. In the following, I want to provide an overview of each of those technologies, discuss respective advantages and disadvantages, and give a few examples for each (Table 1-4).

Table 1-4 Comparison of qRT-PCR and sRNA-Seq technologies for measurement of miRNA in single cells.

Technology	Advantages	Disadvantages	Example assays
qRT-PCR	• Established method • Higher sensitivity and accuracy • Easier to use and customizable • Cheaper than RNA-Seq	• No discovery of novel miRNAs • No global miRNAome profiling / low throughput	• miRCURY LNA™ microRNA qPCR system (Qiagen, Lunn et al., 2008) • miScript Single Cell qPCR Kit (Qiagen) • miScript miRNA PCR arrays (Qiagen) • Two-tailed RT-qPCR (Androvic et al., 2017)
sRNA-Seq	• Can detect novel miRNAs • Global miRNAome profiling / high throughput	• Still under development • Significant computational support needed for analysis • More expensive	• Small-Seq (Faridani et al., 2016) • STA (Lee et al., 2017) • Half-cell sequencing (Wang et al., 2019)

QRT-PCR is a well-established method, which has been used for several decades and has also been adapted to miRNA detection more than ten years ago (Tang et al., 2006; Lunn et al., 2008). Depending on primer design it has a very high sensitivity and accuracy and is fairly easy to use. Additionally, it can be customized and – to date – is still cheaper than RNA-Seq. However, since PCR is a targeted technology, it only allows detection of known miRNAs. Despite the existence of several commercially available qRT-PCR solutions for single cell miRNA quantification (see Table 1-4), which include qPCR arrays for many different biological processes and diseases, also for BC (84 miRNAs covering many different pathways), researchers are still developing alternative approaches to decrease costs and improve performance. For example, Androvich and colleagues introduced a novel qRT-PCR method utilizing a two-tailed hemiprobe primer for reverse transcription followed by regular qPCR (Androvic et al., 2017). The two-tailed primer contains a stem-loop that functions as an elongation to make the resulting cDNA long enough for regular qPCR using two standard linear primers. This approach was shown to be sensitive enough to detect single cell amounts of miRNA, across an astonishing number of seven to eight orders of magnitude down to only ten copies depending on the target miRNA. Additionally, the reverse transcription step can be multiplexed, significantly increasing the throughput and lowering reagent costs. Overall, the method is very cheap compared to commercially available qRT-PCR technologies. The downsides of the two-tailed qPCR method are on the one hand that the design of the two-tailed primers is much more complicated and time-consuming than that of linear

primers and on the other hand that it has not been extensively tested by other research groups yet.

In contrast, SC RNA-Seq is a relatively new technology that has only become widely accessible at the beginning of this decade. In 2013, SC sequencing (both DNA and RNA) was crowned method of the year by *Nature methods* (Nature methods, 2013), however, method development is still ongoing (Chi, 2013). This applies even more so to small RNA-Seq (sRNA-Seq), which is required for miRNA profiling and differs significantly from the more established long RNA-Seq (for mRNA and long non-coding RNA). Several individual approaches, which specifically focus on uniform and robust cDNA generation from small RNAs (sRNA), have been published this decade (Jayaprakash et al., 2011; Viollet et al., 2011; Sorefan et al., 2012; Zhang et al., 2013; Song et al., 2014; Baran-Gale et al., 2015). As library generation from sRNAs is more challenging than from long RNAs - due to the smaller size resulting in higher amplification bias - the results generated by these protocols can vary a lot, requiring a systematic comparison to enable appropriate data interpretation (Dard-Dascot et al., 2018; Giraldez et al., 2018). Giraldez and colleagues have found that protocols using adapters containing degenerate bases (4N protocols) created less bias, but were still differing a lot from each other. However, none of these protocols have yet been tested in the SC setting. To my knowledge, there are currently only three published protocols that are able to sequence small RNAs (sRNA) derived from a single cell (Faridani et al., 2016; Lee et al., 2017; Wang et al., 2019). While the method of Faridani and colleagues (Small-Seq) focuses specifically on sRNA, the single tube amplification (STA) approach by Lee and the half-cell sequencing strategy by Wang go one step further and allow sequencing of both mRNA and sRNA from a single cell. In the latter method, the lysed cell is split in half, which supposedly avoids material loss and technical variation (Roden et al., 2015). They then perform long RNA-Seq on one and sRNA-Seq on the other half. While there have been other studies providing evidence that parallel sequencing of genome and transcriptome from single cells (Macaulay et al., 2015; Han et al., 2018), of epigenome and transcriptome (Angermueller et al., 2016; Clark et al., 2018), and even of genome, epigenome, and transcriptome (Hou et al., 2016) is possible, Lee and colleagues were the first to carry out a dual-omics sequencing approach on mRNA and miRNA (Lee et al., 2017).

All of the methods above have one common characteristic: they do not allow concomitant isolation of gDNA, mRNA, and miRNA, even though several of them enable some form of multi-omic analysis of SCs. Furthermore, many of these approaches are streamlined, single-purpose applications that do not generate pre-amplified material that can be archived for later analysis. This is appropriate when working with cell culture or any other sample that is available in high amounts, however, it is insufficient for our rare patient-derived DCC samples, because we require a triple-omics approach that functions more like a platform that allows long-term storage of amplified gDNA, mRNA, and miRNA for future research to build on like our established WTA does (see chapter 3.2.1). Therefore, there is still an unmet need for a technology able to provide gDNA, mRNA and miRNA profiling of a single cell.

1.7 Aims of the study

Although the LumA and LumB molecular subtypes are closely related due to their luminal phenotype, LumB BC displays a worse outcome than the LumA type, because of a higher propensity to metastasize (Buonomo et al., 2017) and to develop therapy resistance (Ades et al., 2014). However, there is still a lack of data on what causes this discrepancy. Therefore, there is an unmet clinical need for an in-depth comparison of these two subtypes, in order to improve both diagnostics and therapy. We hypothesize that the difference between the two subtypes must be caused by variations in gene expression or mutational landscape. Identification of these differences should help to explain the increased malignancy of the LumB subtype compared to LumA and to facilitate development of more effective treatment for LumB patients.

Since DCCs represent the basis for metastatic spread of BC, we decided to focus on the characterization of LumB DCCs to gain insight into the exact mechanisms that allow LumB DCCs to disseminate and form metastases more frequently than LumA DCCs. To tackle this problem as well as to identify new diagnostic and therapeutic targets, it was necessary to detect, isolate, and molecularly characterize DCCs in detail. Due to the presence of NCCs among $EpCAM^+$ cells isolated from patient BM, it was necessary to find a way to distinguish true DCCs from the NCCs to prevent contamination of downstream analyses with non-cancerous cells. This was a prerequisite before genomic and transcriptomic analyses of LumA and LumB DCCs were carried out.

Numerous studies have shown that miRNAs can play important roles in BC. Therefore, another aim of this thesis was to find a way to include these molecules in our established WTA protocol to facilitate simultaneous study of miRNAome, genome, and transcriptome of each single DCC.

With all the previous points in mind, the central questions of this thesis were the following:

- How can true DCCs be distinguished from $EpCAM^+$ NCCs?
- By which somatic mutations or differentially expressed genes do LumA and LumB subtype DCCs differ?
- How can isolation of miRNAs from single cells along with mRNA and gDNA be realized on the basis of our WTA?

2. Materials

2.1 Patient bone marrow samples

2.1.1 Cooperation partners

Human bone marrow (BM) aspirates of breast cancer patients were provided by the clinical cooperation partners listed in Table 2-1:

Table 2-1 List of clinical cooperation partners.

Name	Institution at time of sample submission
Dr. Brigitte Rack	Department of Gynecology and Obstetrics, University Medical Center Munich
Dr. Claus Lattrich	Department of Gynecology and Obstetrics, University Medical Center Regensburg
Dr. Daniel Oruzio	Department of Hematology and Oncology, Medical Center Augsburg
Dr. Matthias Maak	Department of Surgery, Medical Center rechts der Isar, Munich
Dr. Sebastian Winkler	Orthopedic Clinic of the University of Regensburg, Asklepios Clinic Bad Abbach
Dr. Stefan Buchholz	Department of Gynecology and Obstetrics, University Medical Center Regensburg
Dr. Thomas Blankenstein	Department of Gynecology and Obstetrics, University Medical Center Munich
Prof. Dr. Günter Schlimok	Department of Hematology and Oncology, Medical Center Augsburg
Prof. Dr. Helga Bernhard	Medical Department V – Oncology and Hematology, Medical Center Darmstadt
Prof. Dr. Karl Sotlar	Department of Gynecology and Obstetrics, University Medical Center Munich
Prof. Dr. Michael Nerlich	Department of Trauma Surgery, University Medical Center Regensburg

2.1.2 Sample acquisition

Bone marrow samples were received between August 2008 and December 2015. Bone marrow aspirates from M0- and M1-stage breast cancer patients were collected directly before primary tumor resection and screened for $EpCAM^+$ cells. In addition, we obtained bone marrow samples of cancer-free female healthy donors (HD) undergoing trauma or orthopedic surgery as controls. Detailed information on the sample numbers is provided in the results in chapter 4.1.

2.1.3 Subtype criteria for primary tumors

Using the receptor status of each patient's PT provided by the respective pathologist, each patient and all their isolated DCCs were assigned to appropriate molecular intrinsic subtypes according to the criteria listed in Table 2-2.

Table 2-2 Criteria for BC subtype determination. Criteria according to (Cheang et al., 2009; Goldhirsch et al., 2011). ER = estrogen receptor, PR = progesterone receptor, HER2 = human epidermal growth factor receptor 2

Subtype	HR status	KI67 status	HER2 amplification
Luminal A	+	<14 %	-
Luminal B	+	≥14 % Any if HER2+	-/+ If + then any KI67
Luminal undefined	+	Unknown	Negative or unknown
Triple Negative / Basal-like	-	any	-
HER2 enriched	-	any	+

2.1.4 Ethics

All aspects of the study were approved by the local ethics committee of the University of Regensburg (Regensburg, Germany; ethics vote number 07-079).

2.2 Reagents

2.2.1 Chemicals and commercial solutions

Table 2-3 List of used chemicals.

Name	Manufacturer	Catalog nr.
1 kb Plus DNA Ladder + Dye	New England Biolabs	N3200L
AB serum, human	Bio Rad	805135
Acetonitrile ≥99.9 %, LiChrosolv® gradient grade for liquid chromatography	VWR	1.00030.2500
Adenosine triphosphate (ATP) 100 mM	Roche Diagnostics	11140965001
Agarose LE	Anprotec	AC-GN-00009
Ammonium persulfate (APS)	Sigma Aldrich	A3678-100g
AMPure XP purification beads	Beckman Coulter	A63882
Bacto™ Peptone	Becton Dickinson	211677
Boric acid (H_3BO_3)	Sigma Aldrich	31146-500G
Bovine serum albumin (BSA) (20 mg/ml) (for PCR)	Roche Diagnostics	10711454001
BSA (for picking and cell culture)	Sigma Aldrich	B8667-5ml
BSA fraction V (for MACS buffer)	VWR	441555J
Chloroform	Roth	3314
Deoxy guanosine triphosphate 100 mM	GE Healthcare	28406521
Dithiothreitol (DTT, comes with SuperScript)	Thermo Fisher Scientific	11553117
dNTP Set ;100 mM each A, C, G, T; 4x 24 µM	GE Healthcare	28-4065-51
EcoRI buffer 10x	New England Biolabs	B7006S
Ehtanol absolute Mol. Bio.Grade 250 ml	VWR Chemicals	437443T
Elution buffer (Buffer EB)	Qiagen	19086
Ethanol absolut ≥99.8 %, AnalaR NORMAPUR®	VWR	20821.330

Name	Manufacturer	Catalog nr.
Ethidium Bromide Solution (10 mg/ml)	Sigma-Aldrich	E1510-10ML
Ethylenediaminetetraacetic acid (EDTA)	J.T. Baker	B-1073.1000
Expand Long Template Buffer 1	Roche Diagnostics	11759060001
FastStart dNTP mix	Roche Diagnostics	4738420001
FastStart PCR buffer with $MgCl_2$	Roche Diagnostics	4738420001
Fetal bovine serum (FBS) sera Plus	PAN Biotech	P30-3702
Formamid BioUltra, for molecular biology, ≥99,5 %	Sigma-Aldrich	47671-250ML-F
Gel loading dye 6x, purple, no SDS	New England Biolabs	B7025S
Gene expression buffer 1	Agilent Technologies	5288-5325
Gene expression buffer 2	Agilent Technologies	5288-5326
Hank's balanced salt solution 10x	Biochrom	L2045
Igepal CA-630 viscous liquid	Sigma-Aldrich	I3021-50ml
iQ™ SYBR® Green Supermix	Bio Rad	1708885
L-Glutamin	PAN Biotech	P04-80100
Low molecular weight ladder	New England Biolabs	N3233L
Magnesium chloride ($MgCl_2$) solution 1 M	Sigma-Aldrich	M1028-100ml
Mineral oil, for molecular biology, light oil	Sigma-Aldrich	M5904-500ML
Monopotassium phosphate (KH_2PO_4) for analysis	VWR International	1048731000
mTRAP™ Lysis Buffer	Active Motif	29011
PB buffer	Qiagen	28106
PE buffer	Qiagen	28106
Penicillin (10.000U/ml) / Streptomycin (10mg/ml)	PAN Biotech	P06-07100
Percoll™	GE Healthcare	17089101
Phenol	Roth	A980
Poly T gripNA™ Probe (=PNA)	Active Motif	29008
Potassium Chloride Solution 1M in H_2O for Molecular Biology	Sigma-Aldrich	60142-100ML-F
RPMI 1640	PAN Biotech	P04-17500
RT buffer 5x	Thermo Fisher Scientific	11553117
S-adenosylmethionine (SAM)	New England Biolabs	B9003S
Sodium chloride solution 5 M	Sigma Aldrich	71386-1L
SPRI Select beads	Beckman Coulter	B23317
Streptavidin Beads (from mTRAP™ Midi)	Active Motif	29010
SUPERase In™ RNase Inhibitor (20 U/µL)	Thermo Fisher Scientific	AM2696
Tetramethylethylenediamine (TEMED)	Roth	2367.3
Tris buffer pH 7.0 (1 M) for molecular biology	AppliChem	A5247,0500
Tris buffer pH 8.0 (1 M) for molecular biology	AppliChem	A4577,0500
Tris buffer pH 8.8 (1 M) for molecular biology	AppliChem	A4265,0500
Tris EDTA 1x pH 8.0 low EDTA for mol. Biology	AppliChem	A8569,0500
Tris ultrapure for biochemistry	AppliChem	A1086,1000
tRNA from *E. coli* MRE 600	Roche Diagnostics	10109541001
Trypan blue	Sigma Aldrich	T8154-20ml
Trypsin/ EDTA (10x) (0,05 % Trypsin/ 0,02 % EDTA)	PAN Biotech	P10-024100
TWEEN® 20, for molecular biology, viscous liquid	Sigma-Aldrich	P9416-50ml

Name	Manufacturer	Catalog nr.
Urea	Sigma Aldrich	51456-500G
Water for chromatography, LiChrosolv®, LC-MS grade (PCR-water)	Merck	1.15333.1000
Water UltraPure, DEPC-treated (DEPC-water)	Invitrogen	750023
Water, aqua ad iniectabilia (NGS-water)	Braun	2351744
Water, demineralized	Taken from tap	Not available

2.2.2 Custom buffers and solutions

Table 2-4 List of used custom buffers and solutions with composition.

Name	Components	Application
AB serum 10 %/ peptone 2 % solution	5 ml AB serum, human 5 ml 20% peptone solution 40 ml 1x phosphate buffered saline (PBS) pH 7.4	EpCAM staining
cDNA-Igepal wash buffer	500 µl Tris-HCL 1 M, pH 8,0 750 µl KCl 1 M 1000 µl DTT 0.1 M 25 µl 100 % Igepal CA-630 7725 µl DEPC-water	WTA
cDNA-Tween wash buffer	500 µl Tris-HCL 1 M, pH 8,0 750 µl KCl 1 M 1000 µl DTT 0.1 M 50 µl 100% Tween 20 7700 µl DEPC-water	WTA
dNTPs 10 mM	10 µl dATP 100 mM 10 µl dCTP 100 mM 10 µl dGTP 100 mM 10 µl dTTP 100 mM 60 µl DEPC-water	WTA
DTT 1 mM	10 µl DTT 0.1 M 990 µl DEPC-water	WTA
Formamide 20 %	10 ml Formamide 40 ml DEPC-water	WTA
Igepal 10 %	2 ml 100 % Igepal CA-630 18 ml DEPC-Water	WTA
KH_2PO_4 200 mM	200 µl KH_2PO_4 800 µl DEPC-water	WTA
MACS buffer	500 ml 1x PBS 2.5 g BSA Fraktion V 2 ml 0.5 mM EDTA	Processing of BM samples
$MgCl_2$ 40 mM	40 µl $MgCl_2$ 100 µl DEPC-water	WTA
One Phor All (OPA) buffer	5 ml 1 M Tris acetate 5 ml 1 M Magnesium acetate 1 ml 5 M Potassium acetate PCR-Water ad 1 L Sterile filtrate	WGA

Name	Components	Application
Peptone solution 20 %	100 g Bacto™ Peptone 500 ml 1x PBS pH 7.4 Sterile filtration with 0.45 µM filter	EpCAM staining
Percoll 100 %	100 ml Percoll stock 9 ml Hanks balanced salt solution Sterile filtrate	Processing of BM
PBS pH 7.4 10x	450 g Sodium chloride 71.65 g Disodium phosphate (Na_2HPO_4) 13.35 g Monopotassium phosphate (KH_2PO_4) Distilled water ad 5 L	Cell culture, processing of BM samples
Polyacrylamide (PAA) carrier	1250 µl Acrylamide 400 µl Tris-HCl 67 µl Sodium acetate 20 µl EDTA 8153 µl PCR-H_2O Mix, then add: 100 µl APS 10 µl TEMED Polymerize for 2 h, then add: 25 ml Ethanol 100% Mix and centrifuge, discard supernatant, then solve pellet in 45ml PCR-H_2O	WTA/WGA from supernatant
Tailing wash buffer	500 µl KH_2PO_4 100 µl DTT 25 µl Igepal 100 % 9365 µl DEPC-water	WTA
Tris/Borate/EDTA (TBE) buffer 10x	539 g Tris 275 g Boric acid 37 g EDTA 5 l Demineralized water	Agarose gel electrophoresis
Tween 10 %	2 ml 100 % TWEEN® 20 18 ml DEPC-Water	WTA

2.2.3 Enzymes

Table 2-5 List of used enzymes.

Name	Manufacturer	Catalog nr.	Application
AcuI	New England Biolabs	R0641S	Primer establishment
AluI	New England Biolabs	R0137S	Primer establishment
ApaLI	New England Biolabs	R0507S	Primer establishment
BanI	New England Biolabs	R0118S	Primer establishment
BbvI	New England Biolabs	R0173S	Primer establishment
BglI	New England Biolabs	R1043L	RNA-Seq
BpuEI	New England Biolabs	R0633S	RNA-Seq
BsmAI	New England Biolabs	R0529S	Primer establishment
DdeI	New England Biolabs	R0175S	Primer establishment
DNase, RNase-free	Qiagen	79254	Bulk miRNA isolation
DpnI	New England Biolabs	R0176S	Primer establishment

Name	Manufacturer	Catalog nr.	Application
FastStart Taq Polymerase	Roche Diagnostics	4738420001	WTA/WGA-QC, gradient PCR
FatI	New England Biolabs	R0650S	Primer establishment
HaeIII	New England Biolabs	R0108S	Primer establishment
HhaI	New England Biolabs	R0139S	Primer establishment
HinP1I	New England Biolabs	R0124S	Primer establishment
Hpy188I	New England Biolabs	R0617S	Primer establishment
Hpy188III	New England Biolabs	R0622S	Primer establishment
HpyCH4III	New England Biolabs	R0618S	Primer establishment
HpyCH4V	New England Biolabs	R0620S	Primer establishment
Invitrogen™ SuperScript® II Reverse Transcriptase	Thermo Fisher Scientific	11553117	WTA
MnlI	New England Biolabs	R0163S	Primer establishment
Mse I, recombinant, conc., 2500U 50000U/ml	New England Biolabs	R0525M	WGA, primer establishment
MslI	New England Biolabs	R0571S	Primer establishment
MwoI	New England Biolabs	R0573S	Primer establishment
NdeI	New England Biolabs	R0111S	Primer establishment
NlaIII	New England Biolabs	R0125S	Primer establishment
NlaIV	New England Biolabs	R0126S	Primer establishment
Pol Mix 5 U/µl (Expand Long Template enzyme mix)	Roche Diagnostics	11759060001	WTA, WGA, re-amplification, RNA-Seq
Poly(A) Polymerase (*E. coli*)	New England Biolabs	M0276L	miRNA experiments
Protease	Active Motif	29012	WTA
Proteinase K, recombinant	Roche	3115828001	WGA
RNase H	Thermo Fisher Scientific	EN0201	miRNA experiments (rRNA depletion)
Sau96I	New England Biolabs	R0165S	Primer establishment
T4 DNA Ligase 500U 5U/µl;	Roche	10799009001	WGA
Terminal Deoxynucleotidyl Transferase, Recombinant	Affymetrix	72033	WTA
Tsp45I	New England Biolabs	R0583S	Primer establishment
Tsp509I	New England Biolabs	Discontinued	Primer establishment
Tth111I	New England Biolabs	R0185S	Primer establishment
XmnI	New England Biolabs	R0185S	Primer establishment

2.2.4 Antibodies and microbeads

Table 2-6 List of used antibodies and microbeads.

Name	Manufacturer	Catalog nr.
CD11b-APC, human (Clone: M1/70.15.11.5)	Miltenyi Biotec	130-091-241
CD235a (Glycophorin A) MicroBeads, human	Miltenyi Biotec	130-050-501
CD326 (EpCAM)-PE, human (Clone: HEA-125)	Miltenyi Biotec	130-098-118
CD33-APC, human (Clone: AC104.3E3)	Miltenyi Biotec	130-091-731
CD45-APC, human (Clone: 5B1)	Miltenyi Biotec	130-110-633
Anti-APC-MicroBeads	Miltenyi Biotec	130-090-855

2.2.5 Oligonucleotides and primers

All oligonucleotides and primers were obtained from Eurofins Germany in HPSF grade except for hemiprobe and qPCR primers for miRNA detection, which were purchased in HPLC grade. The primers CP2-BglI-13C and CP2-BpuEI were obtained from Metabion.

Table 2-7 List of used oligonucleotides and primers. * These primers amplify a polymorphic DNA section on human chromosome 5. Precisely, this is a length polymorphism, i.e. the length may vary for each individual and also between the two alleles of one individual. # This primer was used starting from March 2012, introducing an additional restriction enzyme target site for BpuEI. WTAs performed with this primer are labeled "New SOP", while older WTAs performed with the CFL15CT24 primer are considered "Old SOP". SOP: standard operating procedure, for: forward primer, rev: reverse primer

Name	Base sequence (5'->3') (N=A/T/C/G ; V=A/C/G;)	T_A (°C)	Amplicon size (bp)	Application
18S block oligonucleotide	TAA TGA TCC TTC CGC AGG TT -ZNA-5	NA	-	Blocking of rRNA
18S rRNA rev	AAA CGG CTA CCA CAT CCA AG	58	112	qPCR
18S rRNA for	CAA TTA CAG GGC CTC GAA AG	58	112	qPCR
28S block oligonucleotide	GAC AAA CCC TTG TGT CGA GG-ZNA-5	NA	-	Blocking of rRNA
28S rRNA rev	GTG GAA TGC GAG TGC CTA GT	58	114	qPCR
28S rRNA for	CCT TTT CTG GGG TCT GAT GA	58	114	qPCR
5.8S block oligonucleotide	AAG CGA CGC TCA GAC AGG CG -ZNA-5	NA	-	Blocking of rRNA
5.8S rRNA rev	GAC TCT TAG CGG TGG ATC ACT C	58	109	qPCR
5.8S rRNA for	AGT GCG TTC GAA GTG TCG AT	58	109	qPCR
5S block oligonucleotide	AAA GCC TAC AGC ACC CGG TA -ZNA-5	NA	-	Blocking of rRNA
5S rRNA rev	TAC GGC CAT ACC ACC CTG A	58	102	qPCR
5S rRNA for	GGT ATT CCC AGG CGG TCT	58	102	qPCR
hACTB for	GCG TGA CAT TAA GGA GAA GCT G	58	378	WTA-QC
hACTB rev	CGC TCA GGA GGA GCA ATG AT	58	378	WTA-QC
CFL15CT24	$(CCC)_5$ GTC TAG ATT $(TTT)_7$ TVN	RT	-	WTA
CFL15CT24BpuEI#	CCC CCC CCC CCC CCC GTC TAG ACT TGA GTT $(TTT)_7$ TVN	RT	-	WTA
CFL5CN8	CCC CCC CCC CCC CCC GTC TAG ANN NNN NNN	RT	-	WTA
hKRT19 rev	TTC ATG CTC AGC TGT GAC TG	58	621bp	WGA-QC
hKRT19 for	GAA GAT CCG CGA CTG GTA C	58	621bp	WGA-QC
CP2	TCA GAA TTC ATG CCC CCC CCC CCC CCC	65	-	WTA
CP2_9C	TCA GAA TTC ATG CCC CCC CCC	55	-	WTA reamp
CP2-BglI-13C	TCA GAA TTC ATG $(CCC)_2$ CGG $(CCC)_2$	55	-	RNA-Seq
CP2-BpuEI	TCA GAA TTC ATG $(CCC)_5$ GTC TTG AGT TTT TT	55	-	RNA-Seq
hD5S2117 rev	ACT GTG TCC TCC AAC CAT GG	58	140bp*	WGA-QC
hD5S2117 for	CCA GGT GAG AAC CTA GTC AG	58	140bp*	WGA-QC
ddMSE11	TAA CTG ACAG ddC	65	-	WGA
hEF1alpha rev	TGC CCC AGG ACA CAG AGA CT	58	290	WTA-QC
hEF1alpha for	CTG TGT CGG GGT TGT AGC CA	58	290	WTA-QC

Name	Base sequence (5'->3') (N=A/T/C/G ; V=A/C/G;)	T_A (°C)	Amplicon size (bp)	Application
hEHMT2 rev	CGC CAT AGT CAA ACC CTA GC	58	153	qPCR
hEHMT2 for	GCA ACA TCA GCC GCT TCA	58	153	qPCR
hGAPDH for	CCA TCT TCC AGG AGC GAG AT	58	489	WTA-QC
hGAPDH rev	CAG TGG GGA CAC GGA AGG	58	489	WTA-QC
hAHNAK rev	CCC CAT TTC TCT GCC AAC CA	58	117	qPCR
hAHNAK for	AGT GTG TCT GGG CCT CAA G	58	117	qPCR
hAHSP rev	AAG TCC CTG TAC TTG GCC AG	56	118	qPCR
hAHSP for	TCA ACT ATT ACA GGC AGC AGG	56	118	qPCR
hCA1 rev	AGG GCT GTG TTC TTG AGG AA	58	123	qPCR
hCA1 for	GTG ATA ACG CTG TCC CCA TG	58	123	qPCR
hCD44 for	CGG ACA CCA TGG ACA AGT TT	58	150	qPCR
hCD44 rev	CCG TCC GAG AGA TGC TGT AG	58	150	qPCR
hKRT8 for	GCG TAC AGA GAT GGA GAA CGA	58	150	qPCR
hKRT8 rev	AGC TCC CGG ATC TCC TCT T	58	150	qPCR
hEpCAM rev	AGC CAC ATC AGC TAT GTC CA	55	161	qPCR
hEpCAM for	AAA GTT TGC GGA CTG CAC TT	55	161	qPCR
hJUN rev	CCC CGA CGG TCT CTC TTC AAA	58	101	qPCR
hJUN for	GGT GGC ACA GCT TAA ACA GAA AG	58	101	qPCR
hMKI67_rev	CAG ACC CAG CAA ATC CAA AGT	58	247	qPCR
hMKI67 for	GCG GAG TGT CAA GAG GTG T	58	247	qPCR
hMCM2 rev	ATG CAG AGA GGT TGT GGA TGT T	58	82	qPCR
hMCM2 for	GCC CAG CAG GAC ACT ATT GAG	58	82	qPCR
hPCNA for	CAC TCC ACT CTC TTC AAC GGT	58	118	qPCR
hPCNA rev	ATC CTC GAT CTT GGG AGC CA	58	118	qPCR
hKRAS rev	CTG AAT TAG CTG TAT CGT CAA GG	58	91bp	WGA-QC
hKRAS for	ATA AGG CCT GCT GAA AAT GAC	58	91bp	WGA-QC
Lib1	AGT GGG ATT CCT GCT GTC AGT	65	-	WGA
Long fragment rev	AGA CGT CAG GTG GCA CTT TT	58	206	qPCR
Long fragment for	AGC AAA AAC AGG AAG GCA AA	58	206	qPCR
hRAB7A rev	TTT TCA GGA TCT CGG GGA CT	58	190	qPCR
hRAB7A for	ACA GGC TAG TCA CAA TGC AG	58	190	qPCR
hREEP5 rev	CAG AGA GGG CAG GAA GTT TC	58	209	qPCR
hREEP5 for	GGG CTG AAC TGC TCT ACA AG	58	209	qPCR
hTP53 Exon2/3 rev	CAG CCC AAC CCT TGT CCT TA	58	301bp	WGA-QC
hTP53 Exon2/3 for	GAA GCG TCT CAT GCT GGA TC	58	301bp	WGA-QC
hVCP rev	CGA AGG TTG CTC TCA GAC TC	55	163	qPCR
hVCP for	TTG GTG TGA AGC CTC CTA GA	55	163	qPCR
ZNA1 block oligonucleotide	ACTCAACCAAGTCATTCTGA-ZNA-5	NA	-	Blocking of *LF* transcript

2.2.6 Commercial kits

Table 2-8 List of used commercial kits.

Name	Manufacturer	Catalog nr.	Application
Ampli1™ LowPass Kit (SET A+Set B) 2 x 48 reactions	Menarini Silicon Biosystems	WGLPAB	LowPass-Seq
Bioanalyzer DNA 1000 Kit	Agilent Technologies	5067-1504	RNA-Seq
Bioanalyzer DNA High Sensitivity Kit	Agilent Technologies	5067-4626	LowPass-Seq, RNA-Seq
Expand Long Template PCR System	Roche Diagnostics	11759060001 (Sigma Aldrich)	WTA, WTA reamp, WGA
FastStart™ Taq DNA Polymerase, dNTPack	Roche Diagnostics	4738420001 (Sigma Aldrich)	WTA QC PCR, gradient PCR
Invitrogen™ SuperScript® II Reverse Transcriptase	Thermo Fisher Scientific	11553117	WTA
KAPA Library Quantification Kit	Roche Diagnostics	07960298001	RNA-Seq
MiSeq® Reagent Kit v2 (50 cycles)	Illumina	MS-102-2001	RNA-Seq
MiSeq® Reagent Kit v3 (150 cycles)	Illumina	MS-102-3001	LowPass-Seq
QiaQuick PCR purification kit	Qiagen	28106	WTA reamp purification
Qubit® dsDNA BR Assay Kit	Thermo Fisher Scientific	Q32853	LowPass-Seq, RNA-Seq
Qubit® dsDNA HS Assay Kit	Thermo Fisher Scientific	Q32854	LowPass-Seq, RNA-Seq
RNeasy mini kit	Qiagen	74104	Isolation of bulk RNA
TruSeq DNA CD Indexes (96 Indexes, 96 Samples)	Illumina	20015949	RNA-Seq library preparation
TruSeq DNA PCR-Free High Throughput Library Prep Kit (96 samples)	Illumina	20015963	RNA-Seq library preparation

2.3 Cell culture

2.3.1 Cell lines

Table 2-9 List of used cell lines.

Name	Description	Application
DU145	Human prostate cancer cell line	miRNA experiments

2.3.2 Culturing media

Table 2-10 List of used cell culture media.

Name	Composition	Application
DU145 medium	500 ml RPMI 1640 50 ml FBS 5 ml Penicillin/Streptomcin 5 ml L-Glutamin	Propagation of DU145 cells

2.4 Consumables

Table 2-11 List of used consumables.

Name	Manufacturer	Catalog nr.
Rotilabo® syringe filter, PVDF, sterile, 22 µm	ROTH	P666.1
Rotilabo® syringe filter, PVDF, sterile, 45 µm	ROTH	P667.1
Injekt® 10 ml	B Braun	4606108V
96-well cell culture test plate	TPP	92097
Adhesive clear PCR seal	Biozym	600208
Adhesive sealing sheets	Thermo Fisher Scientific	AB0558
Cell culture flask T75	Greiner Bio-One	658175
Cell Strainer 40µm	Becton Dickinson	352340
Cellstar® serological pipette 10 ml	Greiner Bio-One	607180
Cellstar® serological pipette 2 ml	Greiner Bio-One	710180
Cellstar® serological pipette 25 ml	Greiner Bio-One	760180
Cellstar® serological pipette 5 ml	Greiner Bio-One	606180
Centrifuge tube 15 ml	Greiner Bio-One	188271
Centrifuge tube 50 ml	Greiner Bio-One	227261
Combitips advanced 0.1 ml	Eppendorf	0030089618
Combitips advanced 0.5 ml	Eppendorf	0030089634
Combitips advanced 1 ml	Eppendorf	0030089642
Erlenmeyer flask 250 ml DURAN	Schott	21 216 36
Erlenmeyer flask 500 ml DURAN	Schott	21 216 44
MACS Separation Columns LS	Miltenyi Biotec	130-042-401
MAXYMum Recovery™ PCR Tubes 0.2 ml	Axygen Scientific	11370145
Micro-hematocrit capillary, non-heparinized length 75 mm x 1.1/1.2 mm	Brand	749321
microTUBE-50 AFA Fiber Screw-Cap	Covaris	PN520166
Nitril BestGen® Powederfree gloves	Meditrade	1286
Nunc™ Lab-Tek™ Chamber Slides; 8 fields	Thermo Fisher Scientific	11367764
PCR SingleCap 8er-Soft Strips 0.2 ml, clear	Biozym	710970
PCR tube 0.2 ml, single tube	4titude Deutschland	4ti-0795
PCR tube 1.5 ml, graduated, non-sterile	Greiner Bio-One	616201
PCR tube 2 ml, graduated, non-sterile	Greiner Bio-One	623201
Protein LoBind Tube 0.5 ml	Eppendorf	022431064
Protein LoBind Tube 1.5 ml	Eppendorf	022431081
Protein LoBind Tube 2 ml	Eppendorf	022431102
Protein LoBind Tube 5 ml	Eppendorf	0030108302
Reagent reservoirs 10 ml	Integra	4331
SafeSeal Surphob 1250 µl (filter)	Biozym	VT0270

Name	Manufacturer	Catalog nr.
SafeSeal Surphob 20 µl (filter)	Biozym	VT0220
SafeSeal Surphob 200 µl (filter)	Biozym	VT0240
SafeSeal-Tips Professional 10 µl (filter)	Biozym	770010
Transparent 96-well PCR plate	Biozym	710884
White, skirted 96-well qPCR plate	Biozym	712282

2.5 Devices

Table 2-12 List of used devices.

Name	Manufacturer	Application
Autoclave 3150 EL	Systec	Cell culture
Bioanalyzer 2100	Agilent Technologies	LowPass-Seq, RNA-Seq
CellTram Pump	Eppendorf	DCC isolation
Centrifuge 5424	Eppendorf	WTA reamp purification
Centrifuge 5424R	Eppendorf	BM processing
Centrifuge 5810R	Eppendorf	Cell culture
Centrifuge Plate Fuge	Benchmark Scientific	PCR plate centrifugation
Centrifuge Rotina 380R	Hettich	BM processing
DMZ Universal Puller	Zeitz	DCC isolation
DNA Engine Peltier Thermal Cycler	Bio Rad	WTA
DNA Engine Tetrad2 Peltier Thermal Cycler	Bio Rad	PCRs
Electrophoresis chamber 40-1214	Peqlab	Agarose gel electrophoresis
Genetouch thermal cycler	Bioer	PCRs
HiSeq4000	Illumina	RNA-Seq
Hybridization oven PerfectBlot	Peqlab	WTA
Incubator Heraeus BB15	Thermo Fisher Scientific	Cell culture
Laminar flow bench Her Safe KS18	Thermo Fisher Scientific	BM processing, WTA, WGA, cell culture
LightCyler 480	Roche Diagnostics	qPCR
M220X	Covaris	RNA-Seq
Manual pipettes (2µl, 10µl, 20µl, 200µl, 1000µl)	Gilson	Molecular biology
Microscope Axiovert 200M	Zeiss	DCC isolation
Microscope CX23	Olympus	BM processing
Microscope IB inverted	Optech	Cell culture
Microscope IX81, inverted	Olympus	DCC isolation
Microwave	Micromaxx	Agarose gel electrophoresis
MiSeq	Illumina	LowPass sequencing
Multipette Stream	Eppendorf	PCRs
Nanodrop 2000c	Thermo Fisher Scientific	WTA reamp purification
Neubauer hemocytometer	Schubert & Weiss OMNILAB	BM processing
Patchman NP2 micromanipulator	Eppendorf	DCC isolation
PCR bench UVT-S-AR	Thermo Fisher Scientific	PCR/NGS
pH-meter PB-11	Sartorius	pH adjustment of buffers
Power Supply MP-250N	Kisker Biotech	Agarose gel electrophoresis
Qubit3 fluorometer	Thermo Fisher Scientific	NGS

Name	Manufacturer	Application
Research Pro 8-channel pipette	Eppendorf	PCRs and NGS
Roller Mixer SRT1	Stuart Scientific	WTA, BM staining
Scale AVW220D	Shimadzu	Generation of buffers
Scale PLS 510-3	Kern	Agarose gel electrophoresis
Thermomixer C	Eppendorf	NGS, supernatant WGA
UV Illuminator	Intas	Agarose gel imaging
Vortex/centrifuge PCV-2400	Grant-bio	WTA, WGA, PCR
Vortexer	VELP Scientifica	PCRs

2.6 Software and databases

Table 2-13 List of used databases and software.

Name	Provider/URL/Citation	Application
2100 Expert Software Version B.02.09SIO725 (SR1)	Agilent Technologies	Bioanalyzer
7zip	https://www.7-zip.de/	File compression/ decompression
Bbduk (part of bbmap) version from 08.03.2019	Bushnell, 2014	RNA-Seq analysis
biomaRt v2.40.3	Durinck et al., 2009	RNA-Seq analysis
cellSens Dimension 1.9	Olympus	Microscopy
Cytoscape 3.7.1	https://cytoscape.org/	Generation of GO term networks
Database for Annotation, Visualization, and Integrated Discovery (DAVID) v6.8	https://david.ncifcrf.gov/	RNA-Seq analysis
FastQC v0.11.5	Andrews, 2010	RNA-Seq QC
Featurecounts v1.6.4	Liao et al., 2014	RNA-Seq analysis
GIMP 2.10.12	https://www.gimp.org/	CNA counting
GraphPad Prism v6.07	https://www.graphpad.com/scientific-software/prism/	Plotting of data
Genecards	https://www.genecards.org/	Information on gene names and function
Human Protein Atlas	www.proteinatlas.org	Evaluation of LumB-associated candidate genes
IDT OligoAnalyzer Tool	https://eu.idtdna.com/pages/tools/oligoanalyzer	Primer analysis
InferCNV v1.0.3	Tickle et al., 2019	RNA-Seq analysis
ISCN 2009 Atlas of Genetics and Cytogenetics in Oncology and Haematology	http://atlasgeneticsoncology.org/ISCN09/ISCN09.html	Annotation of LowPass CNA data for *Progenetix*
LightCycler 480 Software 1.5	Roche	qPCR
Limma package for R v3.40.4	Ritchie et al., 2015	RNA-Seq analysis
Microsoft Office 2013/2016	Microsoft	Data management, analysis, and visualization

Name	Provider/URL/Citation	Application
MultiQC v1.7	Ewels et al., 2016	RNA-Seq QC
NCBI nucleotide database	https://www.ncbi.nlm.nih.gov/nucleotide/	Primer design
NCBI Primer-BLAST	https://www.ncbi.nlm.nih.gov/tools/primer-blast/	Primer design and analysis
NEBcutter v2.0	http://nc2.neb.com/NEBcutter2/	Primer establishment
pheatmaps v1.0.12	https://cran.r-project.org/web/packages/pheatmap/index.html	Clustering and generation of heat maps
Progenetix user data analysis tool	https://progenetix.org/cgi-bin/pgx_userfile.cgi?project=progenetix&genome=GRCh38	Generation of cumulative copy number profiles
PubMed	https://www.ncbi.nlm.nih.gov/pubmed	Literature search
Qualimap v2.2.1	García-Alcalde et al., 2012	RNA-Seq QC
Restriction Mapper v3	http://www.restrictionmapper.org/	Primer establishment
R base v3.6.0	R Core Team, 2014	RNA-Seq analysis
R-Studio v3.5.1	https://www.rstudio.com/	Annotation of LowPass RefSeq files
Scan v1.12.1	Wilbert and Lueke, 2019	RNA-Seq analysis
Scater v1.12.2	McCarthy et al., 2017	RNA-Seq analysis
scDD v1.8.0	Korthauer et al., 2016	RNA-Seq analysis
STAR v 2.6.1c	Dobin et al., 2013	RNA-Seq analysis
UCSC Goldenpath database	http://hgdownload.cse.ucsc.edu/goldenpath/hg38/database/	Annotation of LowPass RefSeq files
VassarStats	http://vassarstats.net/	Statistics
Webcutter 2.0	http://heimanlab.com/cut2.html	Primer establishment

3. Methods

3.1 Processing of primary human bone marrow samples and cell isolation

3.1.1 Bone marrow aspiration and shipment

Aspirates were taken from the left and right iliac crests of patients directly before primary tumor surgery at cooperating institutions and shipped to the Chair of Experimental Medicine and Therapy Research at the University Medical Center Regensburg in sterile containers, which were sometimes cooled during summer months, but usually at room temperature (RT), for further processing. In order to prevent clotting during shipping, heparin was added to the bone marrow samples. Clotted samples were discarded.

3.1.2 Processing of primary human bone marrow samples

Upon arrival, the BM sample was washed with Hank's balanced salt solution (HBSS), in order to remove fat and thrombocytes. For this purpose, the sample was filled up to 50 ml with HBSS and centrifuged at 170 x g (acceleration 9, brake 3; applies also to all following centrifugations) for 10 min at 4 °C and the supernatant was discarded. The washing was repeated a second time. Next, the cell pellet was re-suspended in 9 ml of HBSS and overlaid on 6 ml of 65 % Percoll solution (ρ = 1.083 g/cm^3). Erythrocytes and granulocytes were removed by centrifuging the sample at 1000 x g, for 20 min at 4 °C. After centrifugation, the interphase containing mononuclear cells (MNCs) was carefully collected and washed with phosphate buffered saline (PBS). The cell suspension was centrifuged at 500 x g, for 10 min at 4 °C. The number of MNCs and erythrocytes was determined by staining of 10 µl cell suspension with 10 µl Trypan blue and a hemocytometer. Dead MNCs were not included in the count. The sample was depleted of the majority of hematopoietic cells using negative immunomagnetic selection, in order to enrich the DCC-containing fraction, if the cell number was at least 10^7. To achieve this, the cell pellet was re-suspended in MACS buffer (90 µl per 10^7 MNC). Then, the cell suspension was incubated with allophycocyanin (APC)-conjugated antibodies against CD11b, which is expressed by monocytes, granulocytes, macrophages, and natural killer cells (10 µl per 10^7 MNC), CD33, which is found on myeloid cells (5 µl per 10^7 MNC), and CD45, the common antigen of leukocytes (5 µl per 10^7 MNC). After 15 min of incubation at 4 °C, the suspension was washed with ten times the volume of MACS buffer compared to the sample volume. The cell pellet was re-suspended in MACS buffer (60 µl per 10^7 erythrocytes) once more before anti-APC beads (20 µl per 10^7 MNC) and anti-CD235a/glycophorin beads (20 µl per 10^7 erythrocytes), which is expressed by mature erythroid cells, were added. Following a 15 min incubation at 4 °C, the suspension was washed with ten times the volume of MACS buffer compared to the sample volume. The cell pellet was re-suspended in 1 ml of MACS buffer, run through a 40 µm cell sieve, and then put on a MACS LS column, which was previously equilibrated with MACS buffer. Three times 3 ml of MACS buffer were used to wash the column and the eluate containing the marker-negative cell fraction was collected on ice. The eluate was then centrifuged at 500 x g and 4 °C for 5 min. Lastly, after removal of the supernatant and resuspension in 1-5 ml of 1x PBS depending on pellet size, the cells were counted again using a hemocytometer.

3.1.3 EpCAM staining

Using the previously determined cell number, the volume of cell suspension containing two million MNCs was calculated. Next, this volume was taken from the suspension and centrifuged at 500 x g at RT for 5 min. Then, the cells were re-suspended in 98 µl of blocking solution (10 % AB-serum/ 2 % peptone solution) followed by addition of 2 µl of anti-EpCAM-PE antibody (concentration proprietary and therefore unknown), gentle mixing, and incubation in the dark for 15 min on a roller at 4 °C. Following incubation, the solution was centrifuged at 500 x g at RT for 5 min and the supernatant was removed before addition of 1 ml 1x PBS. The solution was again centrifuged at 500 x g at RT for 5 min and the supernatant was removed. Lastly, the cells were resuspended in 100 ml 1x PBS per 1 x 10^6 cells.

3.1.4 Isolation of single EpCAM$^+$ cells

Each BM sample was manually screened for the presence of EpCAM$^+$ cells on an inverted fluorescence microscope (Olympus or Zeiss) equipped with a micromanipulator and a pump. The cell suspension containing 2 x 10^6 cells was split into equal portions of 30 µl containing ~ 3 x 10^5 of stained BM cells and distributed onto the fields of an 8-chamber microscope slide containing 170 µl of PBS. Single cells positively stained for EpCAM and with intact morphology were isolated using a glass capillary attached to the Patchman NP2 micromanipulator and coated in FBS. After capture, the single cell was transferred into a separate BSA coated chamber containing 200 µl of PBS, to ensure that only one cell was isolated. Then, selected single cells were isolated manually using a micropipette, by aspirating each single cell in 1 µl of PBS and transferring it to a 0.2 ml MAXYMum Recovery PCR tube containing 4 µl of mTRAP lysis buffer with 0.4 µl *E. coli* tRNA. These tubes possess a special coating that prevents binding of nucleic acids to the plastic surface, in order to minimize losses. The tubes were immediately stored at -80 °C. The tRNA was added to avoid loss of nucleic acids through unspecific binding to the tube wall. Additionally, a pool of cells was isolated by taking 1 µl of cell suspension and transferring it to the microtubes with tRNA and lysis buffer after screening and isolation of single cells. At last, 1 µl of cell-free PBS, from which individual cells were isolated, was also isolated as a negative control for subsequent WTA.

3.2 Amplification of genome and transcriptome of single cells

3.2.1 Whole transcriptome amplification (WTA)

3.2.1.1 General description

The WTA is a multistep process designed as a dual-omics approach to separate the mRNA from the genomic DNA (gDNA) of a single cell prior to global transcriptome and genome amplification, respectively (Klein et al., 2002; Klein, 2003; Hartmann and Klein, 2006). To this end, a protease digestion of the cell lysate, which degrades RNases and nucleic acid-binding proteins, is performed to free up the gDNA and mRNA. Second, the mRNA is captured using poly(T) peptide nucleic acids (PNA) conjugated to a biotin molecule and streptavidin-coated magnetic beads called mTRAP beads. Throughout several washing steps, the supernatant that contains the genomic DNA (gDNA) is separated and stored in pure ethanol at -20 °C for gDNA precipitation until the WGA is carried out. Third, the mRNA bound to the beads is reverse transcribed (solid phase) using two types of primers: on the one hand poly(T) primers with two random anchoring nucleotides at the 3' end and on the other hand random octamer primers. The two different kinds of primers are meant to

ensure complete reverse transcription of each mRNA molecule regardless of its size, but also introduces a non-deterministic fragmentation of the transcriptome by the random octamer primers. Both sets of primers possess a small multiple cloning site in their center containing restriction enzyme target sites and a stretch of 15 cytosines at the 5' end, which will be important for the final amplification step. After reverse transcription, poly(G) tails are added to the 3' ends of the all cDNA molecules. Lastly, the whole transcriptome is amplified using a single poly(C) primer containing an EcoRI target site to facilitate adapter removal for downstream applications, if necessary. The amplification with a single primer is possible thanks to the introduction of the poly(C) stretch contained in the reverse transcription primers and the poly(G) tailing. The amplification is performed with a single primer, as this provides a more homogenous amplification than the use of two different primers. After amplification, the final WTA product is stored at -20 °C and can be utilized for different analyses like quantitative polymerase chain reaction (qPCR) or RNA-sequencing (RNA-Seq), a form of next generation sequencing (NGS), at any time.

The described method has been made commercially available as a kit with the name *Ampli1* WTA kit.

3.2.1.2 Detailed protocol

Compositions of wash buffers are listed in Table 2-4, the components of the different reaction master mixes can be found in Table 3-1. After thawing of cell lysates, 1 µl of 1 µg/µl protease solution and 1 µl of 37.5 µM solution of Poly T gripNA™ Probe (peptide nucleic acids = PNA) was added to the samples. Proteolytic digestion was performed by incubating the samples for 10 min at 45 °C, followed by inactivation of the protease at 75 °C for 1 min, and annealing of PNAs to poly(A) tails of mRNAs, at 22 °C for 10 min. For capturing mRNAs, 4 µl of streptavidin-conjugated mTRAP beads were added and samples incubated at room temperature (RT) on a roller for 45 min. Next, PNA-mRNA complexes were precipitated on a magnet rack and the bead pellets were washed with 10 µl of cDNA-Igepal wash buffer, 20 µl of cDNA-Tween wash buffer, and again with 20 µl of cDNA-Igepal wash buffer. At each washing step, the DNA-containing supernatants were transferred to a separate MAXYMum Recovery PCR tube, containing 0.8 µl of polymerized 0.25 % polyacrylamide (PAA) as a carrier, for subsequent precipitation and whole genome amplification (WGA; chapter 3.2.2). Following the final washing, the beads were re-suspended in cDNA synthesis mix I and annealing of the primers was done at room temperature for 10 min. After addition of cDNA synthesis mix II, reverse transcription was carried out in a hybridization oven at 44 °C for 45 min while tubes were rotating. Following reverse transcription, beads were precipitated in magnetic racks, washed with 20 µl of tailing wash buffer, and re-suspended in 10 µl of tailing mix. The reaction mixture was covered with 40 µl of mineral oil, and the cDNA single strands released from beads by incubating the mixture at 94 °C for 4 min. The sample was immediately transferred onto ice and 0.8 µl Terminal Deoxynucleotidyl Transferase (TdT) were added. Addition of dGTPs to single stranded cDNA was performed by incubating the mixture for 60 min at 37 °C. After inactivation of TdT at 70 °C for 5 min, 35 µl of primary WTA mix I were added. Hotstart PCR was performed by heating the sample to 78 °C and adding 5.5 µl of primary WTA mix II. The primary WTA was carried out using the cycler program detailed in Table 3-2.

Table 3-1 Master mix compositions for WTA.

Name	Components for one reaction
cDNA mix I	2 µl 5x RT buffer 1 µl DTT 0.1 M 0.5 µl Igepal 10 % 0.5 µl DEPC-water 3 µl CFL15CT24BpuEI 100 µM 3 µl CFL15CN8 200 µM
cDNA mix II	2 µl 5x RT buffer 1 µl DTT 0.1 M 0.5 µl dNTP 10mM 0.5 µl DEPC-water 3 µl Super Script II
Tailing mix	1 µl $MgCl_2$ 40mM 1 µl DTT 1mM 1 µl dGTP 2mM 0.5 µl KH_2PO_4 200mM 6.5 µl DEPC-water
Primary WTA mix I	4 µl Expand Long Template Buffer 1 7.5 µl Formamide 20 % 24 µl DEPC-water
Primary WTA mix II	2.5 µl CP2 24 µM 1.75 µl dNTP 10 mM 1.5 µl DNA Pol mix

Table 3-2 Cycler program for primary WTA.

Step	Temperature (°C)	Duration (h:min:sec)	Cycles
1	78	00:00:30	1
2	94	00:00:15	
3	65	00:00:30	20
4	68	00:02:00	
5	94	00:00:15	
6	65	00:00:30	20
7	68	00:02:30 + 10 s per cycle	
8	68	00:07:00	1
9	4	Forever	1

3.2.2 Whole genome amplification (WGA) from supernatants

3.2.2.1 General description

The WGA of the previously isolated supernatant is another multistep process that enables deterministic, i.e. reproducible, amplification of a cell's genome which is beneficial for comparative genomic hybridization (CGH) methods (Klein et al., 1999; Stoecklein et al., 2002). First, the precipitated gDNA pellet derived from the WTA is washed several times to remove any residual RNA. Second, the gDNA solution is subjected to a protease digestion to destroy residual proteins or DNases that might have been accidentally introduced. Third, the gDNA is digested using the MseI restriction enzyme, which fragments the human genome into pieces that are 150-1500 bp in length. Fourth, a double-stranded adapter oligonucleotide with one of the strands lacking a phosphate (to prevent its ligation) is ligated to the gDNA. Following the ligation, the non-

ligated adapter strand is removed by heat denaturation creating an overhang of the so-called Lib1 oligonucleotide. Lastly, the final PCR amplification is performed using the excess Lib1 molecules as a primer. The polymerase first fills up the previously generated overhangs resulting in a Lib1 complementary sequence on the reverse strand. This enables exponential amplification of the gDNA fragments. Analogous to the WTA, the final WGA product is then stored at -20 °C and the amplified gDNA can be exploited for various downstream applications like CGH or sequencing.

The WGA has been made commercially available as a kit with the name *Ampli1* WGA kit.

3.2.2.2 Detailed protocol

The components of the different reaction master mixes can be found in Table 3-3. The 60 µl of DNA- and PAA carrier-containing supernatant from the WTA procedure were mixed with 120 µl of ice-cold absolute ethanol and left at -20 °C to precipitate overnight. The next day, the tubes were centrifuged at 4 °C for 45 min at 20,800 x g, the supernatant was removed, and the pellet was washed with 180 µl of 70 % ice-cold ethanol and incubated in a thermomixer (18 °C, 350 rpm) for 10 min. Next, the tubes were centrifuged at room temperature for 10 min at 20,800 x g. This washing step, starting from ethanol addition, was repeated two more times. After the final centrifugation, the pellet was air-dried and re-suspended in 3.5 µl of PCR-water followed by incubation in the thermomixer (18 °C, 350 rpm) for 18 h. On the subsequent day, the proteinase K digestion mix was added and the sample incubated in a PCR cycler at 42 °C for 15 h followed by enzyme inactivation at 80 °C for 10 min and subsequent cooling to 4 °C. After digestion of proteins and release of gDNA, the sample was further digested by addition of MseI digestion mix followed by incubation at 37 °C for 3 h. Meanwhile, the pre-annealing mix was placed in a thermal cycler and subjected to an annealing program starting at 65 °C and decreasing by 1 °C per minute down to 15 °C, in order to form a double-stranded adapter complex. After MseI digestion, the enzyme in the sample was inactivated at 65 °C for 5 min. Next, 1 µl ATP 10 mM and 1 µl T4 ligase 5U/µl were added to the pre-annealing mix and the resulting 5 µl of pre-annealing/ligation mix were added to the MseI-digested sample, followed by incubation of the sample at 15 °C overnight. After ligation of the adapter, 40 µl of primary WGA mix were added and the sample was subjected to the amplification program (Table 3-4).

Table 3-3 Master mix compositions for WGA.

Name	Components for one reaction
Proteinase K digestion mix	0.5 µl OPA 10x 0.13 µl Tween 10 % 0.13 µl Igepal 10 % 0.26 µl Proteinase K (10mg/ml) 1.28 µl PCR-water
MseI digestion mix	0.2 µl OPA 10x 0.2 µl MseI 50,000 U/µl 1.6 µl PCR-water
Pre-annealing mix	0.5 µl OPA 10x 0.5 µl Lib1 100 µM 0.5 µl ddMse11 100 µM 1.5 µl PCR-water
Primary WGA mix	3 µl Buffer 1 2 µl dNTPs 10 mM 1 µl DNA Pol Mix 35 µl PCR-water

Table 3-4 Cycler program for primary WGA.

Step	Temperature (°C)	Duration (h:min:sec)	Cycles
1	68	00:03:00	1
2	94	00:00:40	15
3	57	00:00:30	
4	68	00:01:30 + 1 s per cycle	
5	94	00:00:40	9
6	57 + 1 °C per cycle	00:00:30	
7	68	00:01:45 + 1 s per cycle	
8	94	00:00:40	23
9	65	00:00:30	
10	68	00:01:53 + 1 s per cycle	
11	68	00:03:40	1
12	4	Forever	1

3.3 Quality control (QC) of WTA and WGA products

3.3.1 WTA

Successful cell isolation and transcriptome amplification was confirmed by multiplex endpoint PCR for the three transcripts *ACTB, EEF1A1* (EF1α primers), and *GAPDH*. All primers are listed in Table 2-7. A master mix was prepared according to Table 3-5 and 9.5 µl of the master mix were deposited into a reaction tube, followed by addition of 0.5 µl of primary WTA product. To control for the purity of reagents and for functionality of the reaction, a negative and a positive control were included. The PCR reaction was performed according to the following protocol: cDNA was denatured at 95 °C for 4 min, followed by 32 cycles of 95 °C for 30 sec, 58 °C for 30 sec, and 72 °C for 90 sec. Lastly, a final elongation was carried out at 72 °C for 7 min followed by cooling to 4 °C. The amplified cDNA was then analyzed by gel electrophoresis (see chapter 3.4). Samples with two or three bands were considered to have high quality.

Table 3-5 Master mix for WTA quality control.

Reagent	Amount per reaction [µl]
10x FastStart PCR Buffer (with 20mM $MgCl_2$)	1
Primer mix (8 µM per primer)	1
dNTPs (from FastStart kit)	0.2
BSA (20 mg/ml)	0.2
FastStart Taq Polymerase (5 U/µl)	0.1
PCR-water	7

3.3.2 WGA

Analogous to the WTA, successful amplification of the genome from the supernatant was confirmed by endpoint PCR. For this purpose, the four genes *KRAS, KRT19,* and *TP53* Exon2/3, as well as a polymorphic DNA area on chromosome 5 using the D5S2117 primers were analyzed. All primers are listed in Table 2-7. For the PCR, a reaction master mix was prepared according to Table 3-6. Next, 9 µl of the master mix was deposited into a reaction tube and 1 µl of the primary WGA was added. To control for the purity of reagents and for functionality of the reaction, a negative and a positive control were included. The PCR reaction was performed as described in chapter 3.3.1. Lastly, the amplified DNA was loaded on an agarose gel for analysis (see chapter 3.4). Samples with three or four bands were considered to be of high quality.

Table 3-6 Master mix for WGA quality control.

Reagent	Amount per reaction [µl]
10x FastStart PCR Buffer (with 20mM $MgCl_2$)	1
Primer mix (8 µM per primer)	1
dNTPs (from FastStart kit)	0.2
BSA (20 mg/ml)	0.2
FastStart Taq Polymerase (5 U/µl)	0.1
PCR-water	6.5

3.4 Agarose gel electrophoresis

Gels were cast by dissolving different amounts of agarose in 100 ml 1x TBE buffer by heating in a microwave, resulting in agarose gels of different concentration. For the WTA-QC and gradient PCRs, 2 g of agarose were used for a 2% gel, while 1.5 g were used for WGA-QC for a 1.5% gel. Additionally, for restriction digestions 3 g of agarose were added for a 3% gel with better resolution. Next, 4 µl of 10 mg/ml ethidium bromide solution were added and the liquid gel mixed by shaking to distribute the ethidium bromide evenly. The liquid was then transferred from the Erlenmeyer flask to a gel tray equipped with two combs for 20 pockets each and left at RT for at least 20 min for polymerization. During polymerization, 3 µl of gel loading dye were added to each sample and the samples were mixed by pipetting. After polymerization, about 11 µl of each sample were loaded onto the gel to avoid foaming. Together with the samples, 8 µl of 1 kb DNA ladder were also loaded for comparison. Lastly, the DNA was separated at 160 V and 400 mA for 45 min and imaged using UV light.

3.5 Re-amplification of WTA products and quality control

3.5.1 Re-amplification

In order to re-amplify the primary WTA product, a master mix was prepared according to Table 3-7. For each sample, 49 µl were deposited into a 0.2 ml reaction tube and 1 µl of primary WTA was added. The amplification was performed in a PCR cycler according to the program described in Table 3-8. Following the re-amplification, another QC was performed according to chapter 3.3.1.

Table 3-7 Master mix for WTA re-amplification.

Reagent	Amount per reaction [µl]
Expand Long Template Buffer 1	5.0
CP2_9C primer (24 µM)	6.0
dNTPs (10 mM)	1.75
Formamide (20 %)	7.5
Pol Mix (5 U/µl)	1.5
PCR-water	27.25

Table 3-8 Cycler program for WTA re-amplification.

Step	Temperature (°C)	Duration (h:min:sec)	Cycles
1	95	00:01:00	1
2	94	00:00:15	5
3	55	00:01:00	
4	65	00:03:30	
5	94	00:00:15	3
6	55	00:01:00	
7	65	00:03:30 + 10 s per cycle	
8	65	00:07:00	1
9	4	Forever	1

3.5.2 Quality control of WTA re-amplification products

The quality control of the re-amplification was performed as described in chapter 3.3.1 with the following changes. The master mix contained only 6.5 µl of PCR-water per sample and 9 µl of the mix was deposited into a reaction tube. The re-amplified WTA was diluted 1:5 with PCR-water before adding 1 µl of the dilution to the master mix. After the PCR program described in chapter 3.3.1, the amplified cDNA was then analyzed by gel electrophoresis as described in chapter 3.4. Samples were considered high quality, if they displayed two or three bands on the agarose gel.

3.6 Purification of WTA products

Following re-amplification and QC, 25 µl of re-amplified WTA product were purified using the QIAquick PCR Purification Kit according to the manufacturer's instruction with several changes. No pH-indicator was added to the PB buffer. Purified cDNA was eluted from the purification column using PCR-water instead of the manufacturer's EB buffer. For elution, 20 µl PCR- water were pipetted on the silica membrane of the column, followed by 5 min incubation at room temperature prior to the final centrifugation (elution) step. To facilitate a more optimal distribution of water on the silica membrane of the purification column, samples were centrifuged at 500 rpm for 30 s before the final centrifugation at 17,900 x g for 60 s. The concentration of each purified sample was measured using NanoDrop 2000c (Thermo Fisher Scientific) utilizing 1 µl of the purified cDNA.

3.7 Quantitative real-time polymerase chain reaction (qPCR)

Quantitative PCR (qPCR) for selected genes was performed using the LightCycler 480 instrument with normalized template input amounts. Initially, a master mix was prepared according to Table 3-9 with the lights of the PCR bench turned off to avoid bleaching of the SyBr Green dye. The master mix volume was calculated so that each sample could be run in technical triplicates. Positive (one per master mix and plate) and negative controls (one per master mix) were also included in triplicates for each target transcript. A single qPCR reaction comprised 2.5 µl of template cDNA previously diluted to 1 ng/µl (total of 2.5ng cDNA per reaction) and 7.5 µl of the master mix. The approach using a normalized amount of template for absolute quantification was favored over relative quantification due to a lack of reliable housekeeping genes in the single cell setting. The qPCR was performed using the program shown in Table 3-10. Melting curves were examined to validate the specificity of PCR amplification. Samples whose melting curves did not

match that of the positive control were considered negative for the target transcript and were assigned a crossing point (Cp) value of 33 or 35 (depending on experiment, if compatibility with older measurements was required) to enable inclusion in downstream analyses. Cp values were determined with the LightCycler 480 Software using the second derivative maximum method applying the high sensitivity algorithm. All single cell WTA products and controls were measured and analyzed in technical triplicates. Cp values were transferred to *Microsoft Excel* for further analysis. First, Cp values were averaged across the technical replicates and then normalized between plates using the positive controls before further data processing. To enable normalization, the positive controls included on all plates were identical aliquots of a single stock solution of a sample previously known to be positive for the tested transcript. All analyses were carried out with normalized Cp values.

Table 3-9 Master mix for qPCR.

Reagent	Amount per reaction [µl]
iQ SYBR Green Supermix	5.0
Forward primer (8 µM)	0.5
Reverse primer (8 µM)	0.5
PCR-water	1.5

Table 3-10 Cycler program for qPCR. * Annealing temperature depends on the used primer pair. Variables available on the cycler but not listed in the table were all set to zero on the LightCycler 480.

Step	Temp. (°C)	Duration (h:min:sec)	Ramp rate (°C/s)	Acquisition mode	Acquisitions (per °C)	Cycles
Pre-incubation	95	00:05:00	4.4	None	0	1
Amplifi-cation	95	00:00:20	4.4	None	0	40
	Variable*	00:00:15	2.2	None	0	
	72	00:00:15	4.4	Single	0	
Melting curve	95	00:00:05	4.4	None	0	1
	55	00:01:00	2.2	None	0	
	95	00:00:00	0.11	Continuous	5	
Cooling	40	00:00:30	2.2	None	0	1

3.8 Design and establishment of primers

3.8.1 Primer design

Primer sequences were selected using the *NCBI PrimerBLAST online tool* (Ye et al., 2012), which also checks for specificity of each primer. In order to enable specificity checking, the NCBI Reference Sequence ID of the transcript of interest was obtained from the NCBI nucleotide database and used as input for *PrimerBLAST*. For primer sequence acquisition, most of the default *PrimerBLAST* settings were used with a few exceptions. The PCR product size was chosen to be between 80 bp and 200 bp and should span an exon/exon junction, whenever possible. Additionally, primer pairs were selected in such a way that they could simultaneously detect all isoforms of the respective target transcript. Whenever this was not possible, the primer pair detecting most of the isoforms was selected. The most promising primer sequences suggested by PrimerBLAST were also checked for dimer and hairpin formation with the *IDT Oligonucleotide Analyzer*. The final primers were ordered from Eurofins Germany.

3.8.2 Gradient PCR

To determine the optimal annealing temperature of a new primer pair, a mixture of WTA products of different cell lines of different cancer entities - called reference cDNA - was used as a template. All annealing temperatures were selected based on results of the gradient PCR performed on the reference cDNA. The gradient PCR was performed in a total volume of 10 µl comprising 0.5 µl of reference cDNA diluted 1:100 in water and 9.5 µl of master mix (Table 3-11). Thermal cycling was performed according to Table 3-12. Following the PCR, the amplified cDNA was analyzed by electrophoresis (chapter 3.4). The best annealing temperature for the tested primer pair was selected according to highest band intensity on the gel. If several bands displayed similar intensities, the temperature closest or equal to 58 °C was chosen. Temperatures which resulted in undesired bands were excluded. Primer pairs which did not produce a single band of expected size at any temperature were excluded as well.

Table 3-11 Master mix for gradient PCR.

Reagent	Amount per sample [µl]
10x FastStart PCR Buffer (with 20mM $MgCl_2$)	1
Forward primer (8 µM)	0.5
Reverse primer (8 µM)	0.5
dNTPs (from FastStart kit)	0.2
BSA (20 mg/ml)	0.2
FastStart Taq Polymerase (5 U/µl)	0.1
PCR-water	7

Table 3-12 Cycler program for gradient PCR.

Step	Temperature (°C)	Duration (h:min:sec)	Cycles
1	94	00:02:00	
2	55-66	00:00:30	1
3	72	00:02:00	
4	94	00:00:40	
5	55-66	00:00:30	11
6	72	00:00:20	
7	94	00:00:40	
8	55-66	00:00:30	28
9	72	00:00:30	
10	94	00:00:40	
11	55-66	00:00:30	1
12	72	00:07:00	
13	4	Forever	1

3.8.3 Standard curve qPCR

After determination of the best annealing temperature for the new primer pair, primer efficiency and sensitivity were assessed by a standard curve experiment on the LightCycler 480 using the optimal annealing temperature determined by the previous gradient PCR. The aforementioned reference cDNA was diluted with PCR-water to create a titration series ranging from 50 ng to 0.00005 ng in $\log_{10}$ increments. A negative control to assess primer dimer formation and purity of reagents was also included. The qPCR reaction was carried out according to chapter 3.7. Primer efficiency was calculated in *Microsoft Excel* using the Slope function on the $\log_{10}$ dilution factor

and the corresponding Cp value resulting from the respective dilution and then inserting the resulting slope into the following formula:

$$Efficiency\ (\%) = \left(10^{\frac{-1}{Slope}} - 1\right) x\ 100$$

Primers with efficiencies between 90-110 % that were able to detect the target sequence across at least four orders of magnitude were considered suitable for further testing. In cases, in which all possible primer pairs for the same transcript failed, also efficiencies between 80%-90% were accepted to enable qPCR analysis of the respective transcripts.

3.8.4 Restriction digestion of amplicons

Lastly, in order to confirm the identity of the amplified sequence fragments, the PCR product (either from gradient PCR or qPCR) was digested with adequate restriction endonucleases with each amplicon being separately digested using two different enzymes. The enzymes were selected by checking the known amplicon sequence with *NEBcutter2*, *Webcutter 2.0*, and *Restriction Mapper version 3*. Three separate reactions were prepared for each primer pair, one without any enzyme as a control and one for each selected restriction enzyme, according to Table 3-13. After distribution of the master mix to the reaction tubes, 2 µl of PCR-water or the respective enzyme were added to the master mix containing the PCR amplicon. The digestion was carried out by incubation of the reaction at the respective optimal temperature of each enzyme (37 °C in most cases) for 3 h followed by heat inactivation at 65 °C or 80 °C for 20 min depending on the enzyme. Afterwards, the digested DNA was analyzed by gel electrophoresis (chapter 3.4), with the difference that the low molecular weight (LMW) ladder was loaded instead of the 1 kb DNA ladder to achieve a better resolution. If the digested fragments matched the expected sizes for both enzymes, the primers were considered valid and used for further experiments.

Table 3-13 Master mix for restriction digestion.

Reagent	Amount per reaction [µl]
PCR/qPCR amplified cDNA	10
Enzyme-specific buffer (usually Cutsmart buffer)	3
BSA	0.3
PCR-water	15

3.9 Metaphase comparative genomic hybridization (mCGH)

The copy number alteration (CNA) profiles of EpCAM⁺ cells analyzed in this dissertation were all generated by Gundula Haunschild. A detailed description of the methodological approach and data analysis can be found in her dissertation (Haunschild, 2013).

3.10 Gene expression microarray

The gene expression microarray laboratory work of M0 versus HD cells was performed by Nina Patwary following the procedure of (Hartmann and Klein, 2006) and is described in Mrs. Patwary's dissertation (Patwary, in preparation).

3.11 LowPass-Sequencing for copy number alteration profiling

The LowPass-Sequencing (LP-Seq) method is based on work by (Buson et al., 2016; Ferrarini et al., 2017; Ferrarini et al., 2018). WGA products with at least one amplified marker (except *KRAS*) in the WGA-QC were used to prepare sequencing libraries. First, 5 µl of primary *Ampli1*™ WGA product were transferred into a new tube and cleaned up with 1.8x SPRIselect Beads according to manufacturer's instructions and eluted in 22 µl of nuclease free water, of which only 20 µl were transferred to a new tube to avoid aspiration of beads. Then, the libraries for LowPass-Seq were prepared with *Ampli1*™ LowPass kit for Illumina® platforms. Briefly, starting from 3 µl of purified primary *Ampli1*™ WGA product, we generated barcoded libraries compatible with Illumina® systems. The libraries were quantified using the Qubit dsDNA HS reagent kit and the Qubit 3.0 Fluorometer (see chapter 3.13.1). Additionally, the average fragment sizes of the libraries were assessed using the Agilent High Sensitivity DNA Kit on the Agilent 2100 Bioanalyzer (see chapter 3.13.2). The libraries were pooled in equimolar concentrations to obtain a 4 nM pool. Prior to sequencing, 5 µl of the pool were denatured by adding 5 µl of 0.2 M sodium hydroxide and incubation at RT for 5 min. Afterwards, the 10 µl of denatured library were added to 990 µl of ice-cold HT1 buffer. The MiSeq device was prepared according to the manufacturer's instructions. The *Ampli1*™ LowPass libraries were sequenced in single-end mode on a MiSeq device using the MiSeq Reagent Kit v3 (150 cycles) and the custom sequencing primer provided with the *Ampli1*™ LowPass kit for Illumina® platforms, which was diluted in ice-cold HT1 buffer (5 µl primer and 595 µl buffer). Lastly, 600 µl of sample pool and 600 µl of diluted primer were loaded onto the MiSeq reagent cartridge into well 17 (sample) and well 18 (primer), respectively.

The open source software Control-FREEC (Control-Free Copy number caller) was used to obtain copy-number calls, using the mode without reference sample and without contamination parameters. For the evaluation of quality metrics, only samples with more than 200,000 reads and a derivative log ratio spread (DLRS) < 0.50 were evaluated. CNA profiles obtained in the analysis were visualized as scatter plots in which red represents genomic gains and blue genomic losses.

Data analysis is described in chapters 3.16.2 and 4.3.1.

3.12 RNA-Sequencing

Due to a change in the sequence of the CFL15CT24 primer, which introduces a BpuEI target site into the WTA product, the library preparation of all WTAs generated from March 2012 onward needed to be performed slightly differently. WTAs generated before March 2012 were processed according to the Old SOP (see 3.12.1), while all others were treated according to the New SOP (chapter 3.12.2).

3.12.1 Old SOP library preparation

Two separate 1:5 dilutions were generated by mixing 1.5 µl primary WTA and 6 µl NGS-water. Next, two times five (five per WTA dilution, total of ten) separate re-amplification reactions (20 µl each) were prepared (see Table 3-14) and the re-amplification was performed with the previously described re-amplification program displayed in Table 3-8. Afterwards, all ten reactions were pooled in a new 1.5 ml reaction tube and the cDNA products were purified with 1.8x volume of Ampure XP beads according to the manufacturer's instructions. Briefly, the pellet was washed twice with 70 % ethanol and beads were dried by incubation at 37 °C on a thermomixer with open lid until beads became cracked. Afterwards, the cDNA was eluted with 41 µl NGS-water, of which

only 40 µl were transferred to a new tube to avoid aspiration of beads. Following re-amplification, WTA-QC was performed (see chapter 3.3.1) and only samples with at least one band on the agarose gel were processed further. Next, the WTA adapters were removed from the cDNA libraries by restriction digestion. For this purpose, digestion mix I (see Table 3-14) was added and the samples incubated at 37 °C for 1 h followed by heat inactivation of the enzyme at 65 °C for 20 min. Subsequently, the digestion mix II described in Table 3-14 was added and the samples incubated at 37 °C for 3 h followed by heat inactivation of the enzyme at 65 °C 20 min. The digested libraries were then purified with 1.8x volume of Ampure XP beads as described above and eluted in 16 µl Qiagen elution buffer. Next, the library concentration was determined with the Qubit HS reagent kit according (chapter 3.13.1) and the length distribution of purified cDNA populations was analyzed on the Bioanalyzer 2100 (chapter 3.13.2). Optimal Covaris settings for fragmentation of each purified cDNA sample to 350 bp insert size were determined on the basis of the average length distribution. Fragmentation was carried out according to the Covaris M220X manual using 55 µl of cDNA library diluted to 20 ng/µl. Subsequently, correct length distribution of fragmented cDNA was controlled on the Bioanalyzer 2100 using the DNA1000 chip. Sequencing libraries were subsequently prepared using the TruSeq DNA PCR-Free High Throughput Library Prep Kit for directional libraries by following the manufacturer's instructions. Lastly, the resulting libraries were quality-checked again on the Bioanalyzer 2100 using the HS DNA kit (chapter 3.13.2) and subsequently quantified with KAPA Library Quantification Kit for Illumina Platforms according to the kit manual.

Table 3-14 Master mix compositions for Old SOP RNA-Seq preparation.

Reagent	Components for one reaction
RNA-Seq re-amplification mix	2.0 µl Expand Long Template Buffer 1 1.2 µl CP2_BglI-13C primer (24 µM) 1.2 µl CP2_BpuEI-13C primer (24 µM) 0.7 µl dNTPs (10 mM) 3.0 µl Formamide (20 %) 0.6 µl Pol Mix (5 U/µl) 10.3 µl NGS-water 1 µl 1:5 dilution of WTA
RNA-Seq library digestion mix I	40 µl Amplified cDNA 5 µl 10x EcoRI buffer supplemented with 80 µM SAM 2.5 µl BpuEI (5U/µl) 2.5 µl NGS-water
RNA-Seq library digestion mix II	50 µl BpuEI digested cDNA library 1 µl 10x EcoRI buffer 2.5 µl BglI (10U/µl) 6.5 µl NGS-water

3.12.2 New SOP library preparation

Since WTA products generated starting from March 2012 already carried the BpuEI target site, the CP2-BpuEI-13C primer was skipped in the amplification and only five separate reactions were prepared from one 1:5 dilution of the WTA. The composition of the reaction mix is shown in Table 3-15, the PCR was carried out using the program displayed in Table 3-16. Afterwards, all five reactions were pooled in a new 1.5 ml reaction tube for a total volume of 100 µl. Except for differing bead (180 µl beads per sample) and ethanol volumes (300 µl for each washing step) due to the difference in re-amplified cDNA, the remaining library preparation procedure was identical to the Old SOP (chapter 3.12.1).

Table 3-15 Master mix for New SOP RNA-Seq re-amplification.

Reagent	Amount per reaction [µl]
Expand Long Template Buffer 1	2.0
CP2_BglI-13C primer (24 µM)	2.0
dNTPs (10 mM)	0.7
Formamide (20 %)	3.0
Pol Mix (5 U/µl)	0.3
NGS-water	11
1:5 dilution of WTA	1

Table 3-16 Cycler program for New SOP RNA-Seq re-amplification.

Step	Temperature (°C)	Duration (h:min:sec)	Cycles
1	94	00:02:00	1
2	94	00:00:15	8
3	68	00:04:00	
4	68	00:07:00	1
5	4	Forever	1

3.12.3 Sequencing run

For sequencing, the individual sequencing libraries were pooled in groups of twelve in equimolar ratios (10 nM per library) while avoiding overlapping of barcodes. The pools were then sequenced on the Illumina NovaSeq platform with 150bp paired-end reads by Genewiz Germany GmbH, generating 29 million read pairs per sample on average.

3.13 Sample concentration measurement and QC

3.13.1 Qubit concentration measurement

Sample concentrations were determined with the Qubit 3.0 fluorometer according to the manufacturer's instructions using either the broad range (BR) or high sensitivity (HS) reagent kit depending on the sample. Briefly, for each sample 199 µl of BR or HS buffer was mixed with 1 µl of BR or HS dye, respectively, to generate a working solution. Then, 198 µl of the working solution were mixed with 2 µl of the sample. Additionally, two standards were also generated by mixing 190 µl of the working solution with 10 µl of standard 1 or standard 2. Subsequently, the sample and the standards were incubated for 2min at RT before measurement on the Qubit device. Pipetting steps were performed with the lights of the PCR bench switched off to avoid bleaching of the dye. The Qubit fluorometer was first calibrated using the two included standards, before measuring the sample concentrations.

3.13.2 Bioanalyzer analysis of samples

The length distribution of fragments in DNA samples was assessed using the Bioanalyzer DNA High Sensitivity (HS) Reagent kit or DNA1000 kit, respectively, according to the manufacturer's instructions and corresponding program (HS or DNA1000). For HS analysis, samples were diluted either to 1.8 ng/µl (RNA Seq before Covaris) or 1:5 (RNA Seq to check final libraries) before being loaded onto the chip. For DNA1000 analysis samples were loaded onto the chip without dilution,

however the marker was diluted by mixing 30 µl of the marker with 51 µl NGS-water. In order to obtain the average fragment length of the sample, a smear analysis was performed using the 2100 Expert Software after the run. The upper and lower borders for this analysis were either 200-2000bp for RNA-Seq libraries or 300-3000 for LowPass-Seq libraries.

3.14 Cell culture

Adherent DU145 prostate cancer cells were cultured in T75 culturing flasks using 13 ml of the medium (see Table 2-10) with the incubator set to 37 °C and 5 % CO_2. The cells were passaged twice per week at around 90% confluency. For passaging, the medium was removed, cells were washed with 1x PBS and then treated with 2 ml 0.05 % trypsin/ETDA solution at 37 °C for 3 min, followed by addition of 10 ml fresh medium for trypsin inactivation. Afterwards, the cells were transferred to a 50 ml centrifuge tube and centrifuged at 500 x g for 5 min (acceleration 9, brake 3). Next, the supernatant was discarded and the pellet re-suspended in 10 ml of fresh medium. Lastly, the cells were re-seeded into a new T75 flask in 1:30-1:40 dilution depending on the concentration of cells. As cells were only used as source material for experimental WTAs, they were not counted but the confluency was merely estimated by eye.

3.15 Experimental protocols for miRNA isolation

3.15.1 Preliminary experiments

3.15.1.1 Experiments with an *in vitro*-transcribed RNA template

All blocking oligonucleotides were tested using the *Long fragment (LF)* RNA as a template, which were annealed to the template using the programs listed below (Table 3-17, Table 3-18, Table 3-19). The polyadenylation was performed by incubation of the template with poly(A) polymerase (PAP) at 37 °C for 30 min followed by inactivation at 65 °C for 20 min. Then, PNAs were added and annealed to the poly(A) tail by heating to 75 °C for 1 min and incubation at 22 °C for 10 min. Next, 4 µl of mTRAP beads were added and the standard WTA protocol (chapter 3.2.1) performed to obtain amplified cDNA.

The annealing programs for the blocking oligonucleotides are listed below (Table 3-17, Table 3-18, and Table 3-19).

Table 3-17 Annealing95 program for annealing of blocking oligonucleotides.

Step	Temperature (°C)	Duration (h:min:sec)	Cycles
1	95	00:00:45	1
2	74 - 1 per cycle	00:00:20	19
3	22	forever	1

Table 3-18 Annealing82 program for annealing of blocking oligonucleotides.

Step	Temperature (°C)	Duration (h:min:sec)	Cycles
1	82	00:00:45	1
2	74 - 1 per cycle	00:00:20	19
3	22	forever	1

Table 3-19 Annealing75 program for annealing of blocking oligonucleotides.

Step	Temperature (°C)	Duration (h:min:sec)	Cycles
1	75	00:00:45	1
2	74 - 1 per cycle	00:00:20	19
3	22	forever	1

Detailed experimental conditions for all conducted experiments are listed in Table 3-20. For the corresponding results see Table 5-1.

Table 3-20 Experimental details of preliminary experiments with *in-vitro* transcript. The main variable under investigation in each experiment is highlighted in bold. All experiments included appropriate negative controls (without template or without blocking oligonucleotides). After the listed steps, the standard WTA was performed starting from the reverse transcription. All experiments were conducted by Dr. Verena Lieb. The sequence of the *long fragment* (*LF*) transcript was is provided in chapter 12.3. oligo = oligonucleotide, RT buffer = reverse transcription buffer

Experiment	Experimental conditions
WTA 1	• Template: *LF* transcript RNA (8 pg, 42 pg, 100 pg) in water • Lysis buffer: RT buffer supplemented with ATP • RNase inhibitor: none • Cell lysis program: protease only, no PNAs, 22 °C step skipped • Blocking: **40 bp DNA oligo, 20,000 x excess**, Annealing75 (see Table 3-19) • Polyadenylation: 37 °C/65 °C/22 °C, 0.5 µl PAP in 5 µl total vol. • mRNA capture: 1 µl PNAs without GTC, PNA binding program, 4 µl mTRAP beads, incubation at RT for 45 min
WTA 2	• Template: *LF* transcript RNA (30 pg) diluted in water • Lysis buffer: RT buffer supplemented with ATP • RNase inhibitor: none • Cell lysis program: protease only, no PNAs, 22 °C step skipped • Blocking: 40 bp DNA oligo, **20/100/500/1000 x excess**, Annealing75 (see Table 3-19) • Polyadenylation: 37 °C/65 °C/22 °C, 0.5 µl PAP in 9 µl total vol. • mRNA capture: see WTA 1
WTA 3	• Template: *LF* transcript RNA (30 pg) diluted in water • Lysis buffer: RT buffer supplemented with ATP • RNase inhibitor: none • Cell lysis program: protease only, no PNAs, 22 °C step skipped • Blocking: **40 bp and 37 bp DNA oligos**, $10^2/10^3/10^4/5x10^4/10^5$ x excess, Annealing82 (see Table 3-18) • Polyadenylation: 37 °C/65 °C/22 °C, 0.5 µl PAP in 9 µl total vol. • mRNA capture: see WTA 1
WTA 4	• Template: *LF* transcript RNA (30 pg) diluted in water • Lysis buffer: RT buffer supplemented with ATP • RNase inhibitor: none • Cell lysis program: protease only, no PNAs, 22 °C step skipped • Blocking: **25 bp DNA oligo**, $10^3/10^4/2.5x10^4/5x10^4/7.5x10^4/10^5/5x10^5$ x excess, Annealing82 (see Table 3-18) • Polyadenylation: 37 °C/65 °C/22 °C, 0.5 µl PAP in 9 µl total vol. • mRNA capture: see WTA 1
WTA 5	• Template: *LF* transcript RNA (30 pg) diluted in water • Lysis buffer: RT buffer supplemented with ATP • RNase inhibitor: none • Cell lysis program: protease only, no PNAs, 22 °C step skipped • Blocking: **20 bp ZNA oligos**, $10^1/10^2/10^3/10^4/10^5/10^6$ x excess, Annealing82 (see Table 3-18) • Polyadenylation: 37 °C/65 °C/22 °C, 0.5 µl PAP in 9 µl total vol. • mRNA capture: see WTA 1
WTA 6	• Template: *LF* transcript RNA (30 pg) diluted in water • Lysis buffer: RT buffer supplemented with ATP • RNase inhibitor: none

Experiment	Experimental conditions
	• Cell lysis program: protease only, no PNAs, 22 °C step skipped • Blocking: **25 bp DNA or 20 bp ZNA oligos**, $10^3/10^4/10^5/10^6$ x excess, Annealing75 (see Table 3-19) • Polyadenylation: 37 °C/65 °C/22 °C, 0.5 µl PAP in 10 µl total vol. • mRNA capture: see WTA 1
WTA 7	• Template: *LF* transcript RNA (30 pg) diluted in water • Lysis buffer: RT buffer supplemented with ATP • RNase inhibitor: none • Cell lysis program: protease only, no PNAs, 22 °C step skipped • Blocking: **20 bp ZNA oligos**, **$10^6/2x10^6/10^7/1.15x10^7$ x** excess, Annealing75 (see Table 3-19) • Polyadenylation: 37 °C/65 °C/22 °C, 0.5 µl PAP in 10 µl total vol. • mRNA capture: see WTA 1
WTA 8	• Template: *LF* transcript RNA (30 pg) diluted in water • Lysis buffer: RT buffer supplemented with ATP • RNase inhibitor: none • Cell lysis program: protease only, no PNAs, 22 °C step skipped • Blocking: 20 bp ZNA oligos, $10^4/10^5/10^6$ x excess combined either with **Annealing95 or 82** (see Table 3-17 and Table 3-18) • Polyadenylation: 37 °C/65 °C/22 °C, 0.5 µl PAP in 10 µl total vol. • mRNA capture: see WTA 1

After the WTA, a 30 cycle PCR was carried out on the resulting cDNA using the CFL15CT24 primer and a reverse primer specific for the *LF* and the differences in yield between oligonucleotides were analyzed on an agarose gel (see chapter 3.4).

3.15.1.2 Preliminary experiments on single cells

Detailed experimental conditions for all conducted experiments are listed in Table 3-21. For corresponding results see Table 5-2.

Table 3-21 Experimental details of preliminary experiments with single cells and total RNA. The main variable under investigation in each experiment is highlighted in bold. All experiments included appropriate negative controls (without template or without blocking oligonucleotides). After the listed steps, the standard WTA was performed starting from the reverse transcription. All experiments were conducted by Dr. Verena Lieb. The sequence of the *LF* transcript was is provided in chapter 12.3. oligo = oligonucleotide, RT buffer = reverse transcription buffer

Experiment	Experimental conditions
WTA 9	• Template: **single DU145 cells** • Lysis buffer: RT buffer supplemented with ATP and 10 % Igepal • RNase inhibitor: none • Cell lysis program: protease diluted in water, no PNAs • Blocking: **four 20 bp ZNA oligos**, **$2.5x10^5/10^6$ x excess**, Annealing95 or 82 (see Table 3-17 and Table 3-18) • Polyadenylation: 37 °C/65 °C/22 °C, 0.5 µl PAP in 10 µl total vol. • mRNA capture: 1 µl PNAs without GTC, PNA binding program, 4 µl mTRAP beads, incubation at RT for 45 min
WTA 10	• Template: single DU145 cells • Lysis buffer: **RT buffer or PAP buffer** supplemented with ATP, 10 % Igepal and RNase inhibitor • RNase inhibitor: **combinations of none, tRNA, and SUPERase** • Cell lysis program: protease diluted in water, no PNAs • Blocking: four 20 bp ZNA oligos, **10^6 x excess**, Annealing95 or Annealing 82 (see Table 3-17 and Table 3-18) • Polyadenylation: 37 °C/65 °C/22 °C, 0.5 µl PAP in 10 µl total vol. • mRNA capture: **2 µl PNAs with 1M GTC**, PNA binding program, **8 µl mTRAP beads**, rolling at RT for 45 min
WTA 11	• Template: single DU145 cells

Experiment	Experimental conditions
	• Lysis buffer: **RT buffer or PAP buffer** supplemented with ATP, 10 % Igepal and RNase inhibitor • RNase inhibitor: **combinations of none, tRNA, and SUPERase** • Cell lysis program: **protease and 1.5 µl PNAs (like standard WTA)** • Blocking and polyadenylation skipped • mRNA capture: **6 µl mTRAP beads**, rolling at RT for 45 min
WTA 12	• Template: single DU145 cells • Lysis buffer: RT buffer or PAP buffer supplemented with ATP, 10 % Igepal, RNase inhibitor, and **40 mM $MgCl_2$; control with mTRAP** • RNase inhibitor: SUPERase or tRNA • Cell lysis program: **protease diluted in mTRAP**, no PNAs • Blocking: four 20 bp ZNA oligos, 10^6 x excess, Annealing95 (see Table 3-17) • Polyadenylation: 37 °C/65 °C/22 °C, 0.5 µl PAP in 7.4 µl total vol. • mRNA capture: **8 µl mTRAP beads**, rolling at RT for 45 min
WTA 13	• Template: 30 pg DU145 total RNA (=single cell equivalent) • Lysis buffer: **PAP buffer** supplemented with ATP, 10 % Igepal, and RNase inhibitor or **mTRAP with tRNA** • RNase inhibitor: **SUPERase** • Cell lysis program: protease diluted in water, no PNAs • Blocking: four 20 bp ZNA oligos, 10^6 x excess, Annealing95 or no annealing (see Table 3-17) • Polyadenylation: 37 °C/65 °C/22 °C, 0.5 µl PAP in 10 µl total vol. • mRNA capture: **2 µl PNAs with 1M GTC**, PNA binding program, **8 µl mTRAP beads**, rolling at RT for 45 min
WTA 14	• Template: **single DU145 cells or Pools of ten cells** • Lysis buffer: PAP buffer supplemented with ATP, 10 % Igepal, and RNase inhibitor • RNase inhibitor: SUPERase • Cell lysis program: protease diluted in water, no PNAs, • Blocking: four 20 bp ZNA oligos, 10^6 x excess, **Annealing75** (see Table 3-19) • Polyadenylation: 37 °C/**with or without 65 °C**/22 °C, 0.5 µl PAP in 10 µl total vol. • mRNA capture: 2 µl PNAs with 1M GTC, PNA binding program, **6 µl mTRAP beads**, rolling at RT for 45 min
WTA 15	• Template: single DU145 cells or Pools of ten cells • Lysis buffer: PAP buffer supplemented with ATP,10 % Igepal, and RNase inhibitor • RNase inhibitor: SUPERase • Cell lysis program: protease diluted in water, no PNAs, • Blocking: **no blocking oligos**, **Annealing75** (see Table 3-19) • Polyadenylation: 37 °C/with or without 65 °C/22 °C, 0.5 µl PAP in 10 µl total vol. • mRNA capture: 2 µl PNAs with 1M GTC, PNA binding program, 6 µl mTRAP beads, rolling at RT for 45 min
WTA 16	• Template: single DU145 cells or Pools of ten cells • Lysis buffer: PAP buffer supplemented with ATP, 10 % Igepal, and RNase inhibitor • RNase inhibitor: SUPERase • Cell lysis program: protease diluted in water, no PNAs • Blocking: **with and without 20 bp ZNA oligos at 10^6 x excess**, Annealing75 (see Table 3-19) • Polyadenylation: 37 °C/with or without 65 °C/22 °C, 0.5 µl PAP in 10 µl total vol. • mRNA capture: 2 µl PNAs with 1M GTC, PNA binding program, 6 µl mTRAP beads, rolling at RT for 45 min

3.15.1.3 Generation of single cell equivalents

To generate single cell equivalents (SCE), individually picked cells were thawed, pooled in a single tube, mixed, and then dispensed to the same number of new tubes as there were initial samples. This procedure ensured that all samples were homogenous and that the observed differences were due to the different treatments. This approach was employed numerous times throughout the eWTA experiments. To save time in later experiments, 15 or 21 SCs were immediately picked into one tube containing the amount of lysis buffer multiplied by the number of cells to be picked

and frozen at -80 °C. The cell pool was then thawed, mixed, and dispensed to up to 15 or 21 separate reaction tubes, depending on how many cells were contained in the pool.

3.15.2 Modifications of the lysis buffer and cell lysis procedure

3.15.2.1 Custom miRNA isolation buffer (MIB)

3.15.2.1.1 Buffer preparation

The buffers were prepared as follows: Tris, NaCl, and urea powders were dissolved in DEPC-water at RT. The concentrations of the chemicals were chosen in such a way that they would be diluted to 200 mM Tris, 200 mM NaCl, and various concentrations of urea at the following poly(A) tailing step. Subsequently, the pH of the solutions was adjusted using hydrogen chloride (HCl) and sodium hydroxide (NaOH) and a calibrated pH meter. The pH meter was re-calibrated every time a new batch of buffers was prepared. Following pH adjustment, sterile filtration of the buffers was carried out using a 22 µm filter and a 10 ml syringe to push the liquid through the filter. Several aliquots of the buffers were stored at -20 °C. Detergents were added to the buffer directly before each experiment.

3.15.2.1.2 Lysis experiments

In order to test the ability of the buffers to lyse cells, DU145 cells were cultured on an 8-field chamber slide for at least 2 h to let them attach to the surface. Then, the slide was placed under the Olympus IX81 microscope, the medium was removed from one field, movie acquisition was started, and an experimental buffer cooled to 4 °C was added to the same field. After up to five minutes of observation, the next buffer variant was tested on another field. This way, up to seven different experimental buffers were tested in one experiment. The eighth field was used for mTRAP buffer as a reference.

3.15.2.1.3 Poly(A) polymerase activity experiments

Detailed experimental conditions for all conducted experiments are listed in Table 3-22. For corresponding results see Table 5-6.

Table 3-22 WTA experiments testing the functionality of the custom buffer. Important parameters under investigation are highlighted in bold. All experiments included appropriate positive and negative controls. After the listed steps, the standard WTA was performed starting from the reverse transcription. The sequence of the *LF* transcript was is provided in chapter 12.3. MIB = miRNA isolation buffer, NLS = N-lauroylsarcosine

Experiment	Experimental conditions

PAP activity 1	• Template: 10 pg *LF* transcript
	• Started at the polyadenylation step followed by addition of beads and regular WTA procedure
	• Buffer: **PAP buffer only supplemented with ATP, SUPERase and Igepal (0.1 %)**
	• Treatments:
	  ○ 37 °C 30 min (pos. control)
	  ○ **30 °C 30/40/50 min**
	  ○ 30 °C 50 min without PAP (neg. control)

PAP activity 2	• Template: 10 pg *LF* transcript • Started at the polyadenylation step followed by addition of beads and regular WTA procedure • Buffer: PAP buffer as control, **MIB with various urea concentrations,** both supplemented with ATP, SUPERase and Igepal (0.1 %); MIB additionally supplemented with $MgCl_2$

Experiment	Experimental conditions
	• Treatments: o PAP buffer (pos. control/reference) o **MIB with 1/2/4/8 M urea** • Polyadenylation: 30 °C for 50 min, 0.5 µl PAP in 10 µl total vol. • mRNA capture: 2 µl PNAs with GTC, PNA binding program, 6 µl mTRAP beads, incubation at RT for 45 min
PAP activity 3	• Template: 10 pg *LF* transcript • Started at the polyadenylation step followed by addition of beads and regular WTA procedure • Buffer: PAP buffer as control, **MIB with various urea concentrations,** both supplemented with ATP, SUPERase and Igepal (0.1 %); MIB additionally supplemented with $MgCl_2$ • Treatments: o PAP buffer (pos. control/reference) o **MIB with 0/0.5/1/2/3/4/6/8 M urea** o PAP buffer without PAP (neg. control) • Polyadenylation: 30 °C for 50 min, 0.5 µl PAP in 10 µl total vol. • mRNA capture: 2 µl PNAs with GTC, PNA binding program, 6 µl mTRAP beads, incubation at RT for 45 min
PAP activity 4	• Template: 10 pg *LF* transcript • Started at the polyadenylation step followed by addition of beads and regular WTA procedure • Buffer: PAP buffer as control, **MIB with 3 M urea,** both supplemented with ATP and SUPERase; PAP buffer also received Igepal (0.1 %); MIB additionally supplemented with $MgCl_2$ • Treatments: o PAP buffer (pos. control/reference) o **MIB with 3 M urea and 0.1/0.5/0.05 % of NLS** o PAP buffer without PAP (neg. control) • Polyadenylation: 30 °C for 50 min, 0.5 µl PAP in 10 µl total vol. • mRNA capture: 2 µl PNAs with 1 M GTC, PNA binding program, 6 µl mTRAP beads, incubation at RT for 45 min
PAP activity 5	• Template: 10 pg *LF* transcript • Started at the polyadenylation step followed by addition of beads and regular WTA procedure • Buffer: PAP buffer as control, **MIB with 2/3 M urea,** both supplemented with ATP and SUPERase; PAP buffer also received Igepal (0.1 %); MIB additionally supplemented with $MgCl_2$ • Treatments: o PAP buffer (pos. control/reference) o **MIB with 2/3 M urea and 0.1 % of NLS** o PAP buffer without PAP (neg. control) • Polyadenylation: 30 °C for 50 min, 0.5 µl PAP in 10 µl total vol. • mRNA capture: 2 µl PNAs with 1 M GTC, PNA binding program, 6 µl mTRAP beads, incubation at RT for 45 min
WTA 19	• Template: DU145 single cells and cell pools • Buffer: **mTRAP with tRNA as control, MIB with 3 M urea and 0.01 % NLS,** supplemented with SUPERase • Treatment: standard WTA

3.15.2.2 eWTA with diluted mTRAP buffer

Detailed experimental conditions for all conducted experiments are listed in Table 3-23. For corresponding results see Table 5-7. Titration of mTRAP buffer was performed with DEPC-water.

Table 3-23 Experimental details for experiments on activity of PAP in diluted mTRAP buffer. Important parameters under investigation are highlighted in bold. All experiments included appropriate positive and negative controls. The sequence of the *LF* transcript was is provided in chapter 12.3.

Experiment	Experimental conditions
PAP activity 6	• Template: 10 pg *LF* transcript • Started at the polyadenylation step followed by addition of beads and regular WTA procedure • Buffer: **PAP buffer supplemented with ATP, SUPERase and NLS (0.01 %); mTRAP supplemented with ATP, SUPERase, NaCl, and $MgCl_2$**

Experiment	Experimental conditions

	• Treatments:
	  ○ PAP buffer (pos. control/reference)
	  ○ **mTRAP undiluted**
	  ○ **mTRAP 1:5/1:50/1:500/1:5000**
	  ○ PAP buffer without PAP (neg. control)
	• Polyadenylation: 30 °C for **30 min**, 0.5 µl PAP in 10 µl total vol.
	• mRNA capture: 2 µl PNAs with 1 M GTC, PNA binding program, 6 µl mTRAP beads, incubation at RT for 45 min

PAP activity 7	• Template: 10 pg *LF* transcript
	• Started at the polyadenylation step followed by addition of beads and regular WTA procedure
	• Buffer: see "PAP activity 6" above
	• Treatments:
	  ○ PAP buffer (pos. control/reference)
	  ○ **mTRAP undiluted**
	  ○ **mTRAP 1:2.5/1:5/1:10/1:20**
	  ○ PAP buffer without PAP (neg. control)
	• Polyadenylation: 30 °C for **20 min**, 0.5 µl PAP in 10 µl total vol.
	• mRNA capture: 2 µl PNAs with 1 M GTC, PNA binding program, 6 µl mTRAP beads, incubation at RT for 45 min

PAP activity 8	• Template: 10 pg *LF* transcript
	• Started at the polyadenylation step followed by addition of beads and regular sWTA procedure
	• Buffer: see "PAP activity 6" above
	• Treatments:
	  ○ PAP buffer (pos. control/reference)
	  ○ **mTRAP 1:5/1:6/1:7/1:8/1:9/1:10**
	  ○ PAP buffer without PAP (neg. control)
	• Polyadenylation: 30 °C for 20 min, 0.5 µl PAP in 10 µl total vol.
	• mRNA capture: 2 µl PNAs with 1 M GTC, PNA binding program, 6 µl mTRAP beads, incubation at RT for 45 min

PAP activity 9	• Template: 10 pg *LF* transcript
	• Started at the polyadenylation step followed by addition of beads and regular sWTA procedure
	• Buffer: see "PAP activity 6" above
	• Treatments:
	  ○ PAP buffer (pos. control/reference)
	  ○ **mTRAP undiluted**
	  ○ **mTRAP 1:2/1:3/1:4/1:5**
	  ○ PAP buffer without PAP (neg. control)
	• Polyadenylation: 30 °C for 20 min, 0.5 µl PAP in 10 µl total vol.
	• mRNA capture: 2 µl PNAs with 1 M GTC, PNA binding program, 6 µl mTRAP beads, incubation at RT for 45 min

3.15.2.3 Standard WTA with reduced mTRAP volume for cell isolation

15 single DU145 cells were isolated together (in 1 µl PBS using a micropipette) and transferred to one tube containing 45 µl mTRAP (3 µl per cell). Furthermore, 14 µl PBS were added to simulate isolation of cells one by one, because usually each SC is isolated in 1 µl of PBS. Then the solution was gently mixed and the tube frozen at -80 °C. A negative control of the lysis buffer and PBS was also included and processed identically to the samples. Next, 30 µl of protease/PNA/lysis buffer mix were prepared according to the standard WTA protocol (see chapter 3.2.1) and 2 µl of the mix were distributed to each of ten empty reaction tubes. Afterwards, 1 µl of mTRAP and 0.4 µl of tRNA were deposited into five of the same tubes that contained the lysis mix to generate standard WTA conditions with a total of 4 µl mTRAP and tRNA upon addition of cell lysate later on, while only 0.38 µl SUPERase were added to the other five tubes (3 µl mTRAP with SUPERase). Subsequently,

the cell pool was thawed, mixing, centrifuged, and 4 µl of the cell pool were distributed to each of the ten tubes. This resulted in five biological replicates of standard WTA samples containing 5.4 µl (4 µl mTRAP, 1 µl PBS, 0.4 µl tRNA) of cell lysate and five biological replicates of experimental samples containing only 4.375 µl (3 µl mTRAP and 0.375 µl SUPERase) of lysate. These ten samples were then subjected to the standard lysis program and the remaining standard WTA (see chapter 3.2.1).

3.15.3 Proof of principle of polyadenylation

15 single DU145 cells were isolated together (in 1 µl PBS using a micropipette) and transferred to a single tube containing a mixture of 45 µl mTRAP and 5.7 µl SUPERase (3 µl and 0.38 µl per cell, respectively). Furthermore, 14 µl PBS were added to simulate isolation of cells one by one. Then, the solution was gently mixed and the tube frozen at -80 °C. A negative control of the lysis buffer and PBS was also included and processed alongside the samples. Next, 15 µl of protease/lysis buffer mix were prepared from 14.25 µl of mTRAP buffer and 0.75 µl 1 µg/µl protease for later use. After preparation of the lysis mix, the 15-cell pool was thawed and 1 µl of *Long Fragment* RNA diluted to 761,428 copies per microliter (using DEPC-water) was added for a final amount of 50,761 copies per SCE. The pool was vortexed and then 4.4 µl were distributed to each of ten empty reaction tubes, before addition of 3 µl DEPC-water to each tube to account for the volume of blocking oligonucleotides in future experiments, thereby generating the SCEs for the extended WTA (eWTA). These ten samples were then subjected to lysis by addition of 1 µl of the prepared protease lysis mix to each SCE followed by incubation at 45 °C for 10 min and heating to 75 °C for 1 min. Subsequently, 5.88 µl of poly(A) tailing mix, consisting of 1.5 µl 10 mM ATP, 0.64 µl 5 M NaCl, 0.15 µl $MgCl_2$, 0.75 µl SUPERase, and 2.84 µl DEPC-water, were added to all tubes followed by 0.75 µl of PAP to five tubes (polyadenylation group) or 0.75 µl of DEPC-water to the control group. The tubes were then placed in a PCR cycler and incubated at 37 °C for 30 min and cooled to 4 °C. After the tailing reaction, 4 µl of GTC/PNA mix, consisting of 1 µl 5 M GTC and 3 µl 37.5 µM PNAs, were added and the PNAs annealed in a cycler by heating to 75 °C for 1 min to relax secondary structures followed by 22 °C for 10 min to anneal the PNAs. Lastly, 6 µl of streptavidin-conjugated beads were added and from that point the remaining protocol was performed according to the standard procedure (see chapter 3.2.1) with one modification: the first washing step after the reverse transcription was done with 20 µl of cDNA Igepal wash buffer instead of 10 µl.

3.15.4 Targeted rRNA depletion

15 single DU145 cells were isolated together (in 1 µl PBS using a micropipette) and transferred to a single tube containing a mixture of 45 µl mTRAP and 5.7 µl SUPERase (3 µl and 0.38 µl per cell, respectively). Furthermore, 14 µl PBS were added to simulate isolation of cells one by one, because usually each SC is isolated individually in 1 µl of PBS. Then, the solution was gently mixed and the tube frozen at -80 °C. A negative control of the lysis buffer and PBS was also included and processed alongside the samples. Before beginning the WTA procedure, the block oligonucleotide mix consisting of 500 mM of each of the 113 different DNA oligonucleotides (sequences in appendix chapter 12.5) provided by Dr. Balagopal Pai was prepared. The WTA experiment comprised three different treatments with five replicate samples each: a control group without addition of oligonucleotides and RNase H (group I), a group that received only the oligonucleotides (group II), and a group that received both the oligonucleotides and the enzyme (group III). Additionally, one negative control for each treatment group was also included. First,

the mTRAP/protease lysis mix was prepared from 26.5 µl mTRAP buffer and 1 µl of 1µg/µl protease solution. The dilution of the protease was increased compared to the standard WTA, because of the lower reaction volume at the lysis step due to the reduced mTRAP amount and lack of PNAs. Second, 15 µl of the lysis mix were added to the thawed 15-cell pool and lysis was carried out at 45 °C for 10 min followed by enzyme activation at 75 °C for 1 min. Subsequently, the lysate was vortexed, centrifuged, and SCEs were generated by distribution of the lysate to empty tubes in portions of 5.4 µl per tube. Afterwards, 2.63 µl of block oligonucleotide reaction mix, comprising 1.28 µl of the 500 mM block oligonucleotide mix, 0.8 µl SUPERase, and 0.55 µl DEPC-water, or a mix of 0.8 µl SUPERase and 1.83 µl DEPC-water were added to groups II and III or group I, respectively. Then, groups II and III were subjected to the Annealing75 program (Table 3-19), while group I was stored at 4 °C. After annealing, 5.5 µl of RNase H reaction mix, consisting of 0.47 µl 100 mM $MgCl_2$, 1.5 µl 100 mM DTT, 0.7 µl SUPERase, and 2.93 µl DEPC-water, were added to each sample prior to addition of 1.4 µl of 5U/µl RNase H (ThermoFisher Scientific) or 1.4 µl of DEPC-water to group III and groups I and II, respectively. Subsequently, group III was incubated at 37 °C (lid heated to 40 °C) for 30 min. Meanwhile, groups I and II were stored at 4 °C. After the rRNA depletion, 4 µl of GTC/PNA mix, comprising 1 µl 5 M GTC and 3 µl 37.5 µM PNA, were added to all groups and subjected to PNA annealing by heating to 75 °C for 1 min before incubation at 22 °C for 10 min. Lastly, 6 µl of streptavidin-conjugated beads were added and from that point the remaining protocol was carried out according to the standard procedure (compare chapter 3.2.1) with one modification: the first washing step after the reverse transcription was done with 20 µl of cDNA Igepal wash buffer instead of 10 µl.

3.15.5 Effect of rRNA blocking oligonucleotides

3.15.5.1 Blocking of *Long fragment* transcript spike-in in SCEs

The following treatment groups were included:

- Group I: polyadenylation and rRNA blocking
- Group II: polyadenylation without rRNA blocking

15 single DU145 cells were isolated together (in 1 µl PBS using a micropipette) and transferred to a single tube containing a mixture of 45 µl mTRAP and 5.7 µl SUPERase (3 µl and 0.38 µl per cell, respectively). Furthermore, 14 µl PBS were added to simulate isolation of cells one by one. Then, the solution was gently mixed and the tube frozen at -80 °C. A negative control of the lysis buffer and PBS was also included and processed according to the WTA protocol. Before beginning the WTA procedure, several preparations were made. First, 15 µl of protease/lysis buffer mix were prepared from 14.25 µl of mTRAP buffer and 0.75 µl of 1 µg/µl protease for later use. Second, the ZNA1 block oligonucleotide was diluted to 16.67 x 10^9 copies/µl (~1,000,000x excess compared to spike-in) by mixing 1 µl of the 100 µM stock with 3611.5 µl DEPC-water. Third, the *LF* RNA transcript was diluted to 761,428 copies/µl by blending 1 µl of a 10 ng/µl aliquot of the transcript with 15.6 µl DEPC-water. To start the eWTA, the 15-cell pool was thawed and 1 µl of the diluted *LF* RNA was added to it before mixing. Next, 4.38 µl of the pool were distributed to each of ten empty reaction tubes resulting in ten SCEs, each containing 50,761 copies of the *LF* RNA. Subsequently, 3 µl of the prepared ZNA1 dilution or 3 µl of DEPC-water were added to group I or group II, respectively. Furthermore, 1 µl of the protease/lysis buffer mix was added and the samples were lysed by incubation at 45 °C for 10 min and followed by 75 °C for 1 min. After lysis, 0.42 µl SUPERase were added and the blocking oligonucleotides were annealed using the program shown in Table 3-19. Once the annealing was done, 5.46 µl of poly(A) tailing mix, consisting of 1.5 µl 10 mM ATP, 0.64 µl 5 M NaCl, 0.15 µl $MgCl_2$, 0.75 µl SUPERase, 2.42 µl DEPC-water, were

added to all samples followed by 0.75 µl of PAP, which was also added to both groups. Polyadenylation was carried out by incubation at 37 °C for 30 min followed by cooling to 4 °C. Afterwards, 4 µl of GTC/PNA mix, comprising 1 µl 5 M GTC and 3 µl 37.5 µM PNAs, were added and the PNAs annealed in a cycler. The annealing program consisted of 75 °C for 1 min to relax secondary structures and 22 °C for 10 min to anneal the PNAs. Lastly, 6 µl of streptavidin-conjugated beads were added and from that point the remaining protocol was carried out according to the standard procedure (see chapter 3.2.1) with one modification: the first washing step after the reverse transcription was done with 20 µl of cDNA Igepal wash buffer instead of 10 µl.

3.15.5.2 Blocking of endogenous rRNAs with ZNA oligonucleotides

The following treatment groups were included:

- Group I: polyadenylation and rRNA blocking
- Group II: polyadenylation without rRNA blocking
- Group III: no polyadenylation and no rRNA blocking

After isolation of cells as described above (see chapter 3.15.5.1), 15 µl of protease/lysis buffer mix were prepared from 14.25 µl of mTRAP buffer and 0.75 µl 1 µg/µl protease for later use. Second, each of the four blocking oligonucleotides was diluted 1:10 using DEPC-water for a total volume of 7.5 µl per oligonucleotide and then the dilutions were mixed producing a total of 30 µl of block oligonucleotide mix. To start the eWTA, the 15-cell pool was thawed, 15 µl of the protease/lysis buffer mix were added and the pool was lysed by incubation at 45 °C for 10 min followed by 75 °C for 1 min. Next, 5.38 µl of the pool were distributed to each of 15 empty reaction tubes to generate SCEs. Subsequently, 3 µl of the prepared block oligonucleotide mix or 3 µl of DEPC-water were added to group I or groups II and III, respectively, and the blocking oligonucleotides were annealed using the program shown in Table 3-19, while groups II and III were stored at 4 °C. Once the annealing was done, 5.88 µl of poly(A) tailing mix, consisting of 1.5 µl 10 mM ATP, 0.64 µl 5 M NaCl, 0.15 µl $MgCl_2$, 0.75 µl SUPERase, and 2.84 µl DEPC-water, were added to all samples. Immediately afterwards, 0.75 µl PAP or 0.75 µl DEPC-water were added to groups I and II or group III, respectively. Polyadenylation was carried out by incubation at 37 °C for 30 min followed by cooling to 4 °C. From this point the experiment was conducted as described above (see chapter 3.15.5.1).

3.15.5.3 Comparison of two different sets of blocking oligonucleotides

The following treatment groups were included:

- Group I: no polyadenylation and no rRNA blocking (control)
- Group II: polyadenylation without rRNA blocking
- Group III: polyadenylation and blocking with Dr. Verena Lieb's four ZNA oligonucleotides
- Group IV: polyadenylation and blocking with Dr. Balagopal Pai's 113 DNA oligonucleotides

A total of 21 single DU145 cells were isolated together (in 1 µl PBS using a micropipette) and transferred to a single tube containing a mixture of 63 µl mTRAP and 7.98 µl SUPERase (3 µl and 0.38 µl per cell, respectively). Furthermore, 20 µl PBS were added to simulate isolation of cells one by one. Then, the solution was gently mixed and the tube frozen at -80 °C. A negative control of the lysis buffer and PBS was also included and processed alongside the samples. At the start of the eWTA, the protease/lysis buffer mix was prepared from 26.5 µl of mTRAP buffer and 1 µl 1 µg/µl

protease for later use. Second, Dr. Lieb's four blocking oligonucleotides were diluted 1:10 by adding 1 µl of each 100 µM stock solution to 36 µl of DEPC-water for a total 40 µl of block oligonucleotide mix. The mix of block oligonucleotides provided by Dr. Pai consisting of 500 mM of each of the 113 different oligonucleotides was re-used from the previous rRNA depletion experiment (see chapter 3.15.4). Following the preparations, the 21-cell pool was thawed, 21 µl of the protease/lysis buffer mix were added and the pool was lysed by incubation at 45 °C for 10 min and 75 °C for 1 min. Next, 5.38 µl of the vortexed pool were distributed to each of 20 empty reaction tubes to generate SCEs. Subsequently, 3 µl of the prepared block oligonucleotide mixes or 3 µl of DEPC-water were added to groups III and IV or groups I and II, respectively, and the blocking oligonucleotides were annealed using the program listed in Table 3-19. Once the annealing was done, 5.88 µl of poly(A) tailing mix, consisting of 1.5 µl 10 mM ATP, 0.64 µl 5 M NaCl, 0.15 µl $MgCl_2$, 0.75 µl SUPERase, and 2.84 µl DEPC-water, were added to all samples. Immediately afterwards, 0.75 µl PAP or 0.75 µl DEPC-water were added to groups II-IV or group I, respectively. All samples were subjected to the polyadenylation program by incubation at 37 °C for 30 min followed by cooling to 4 °C. From this point the experiment was conducted as described above (see chapter 3.15.5.1).

3.16 Statistics and bioinformatics

3.16.1.1 Descriptive statistics

The performance of the qPCR-based DCC signature was assessed by calculating several metrics from 2x2 confusion matrices as follows (letters in formulae are relate to Table 3-24):

$$Prevalence = \frac{Actual\ Positives}{Total\ Observations} = \frac{(A+C)}{(A+B+C+D)}$$

$$Accuracy = \frac{True\ Predictions}{Total\ Observations} = \frac{(A+D)}{(A+B+C+D)}$$

$$Misclassification\ rate = \frac{False\ Predictions}{Total\ Observations} = \frac{B+C}{(A+B+C+D)}$$

$$Sensitivity = \frac{True\ Positives}{True\ Positives + False\ Negatives} = \frac{A}{(A+C)}$$

$$Specificity = \frac{True\ Negatives}{True\ Negatives + False\ Positives} = \frac{D}{(B+D)}$$

$$Positive\ Predictive\ Value\ (PPV) = \frac{True\ Positive}{Predicted\ Positive} = \frac{A}{(A+B)}$$

$$Negative\ Predictive\ Value\ (NPV) = \frac{True\ Negatives}{Predicted\ Negatives} = \frac{D}{(C+D)}$$

Table 3-24 Exemplary confusion matrix for calculation of qPCR signature performance metrics.

		CNA profiling	
		DCC	**NCC**
qPCR signature	**DCC**	A	B
	NCC	C	D

3.16.1.2 Statistical tests

Statistical testing of frequency data (counts) of the patient or single cell cohorts was carried out on the VassarStats website for statistical computation (Lowry, 2004), where data were entered into confusion matrices to perform Chi-Square or Fisher's Exact Test (including Freeman-Halton extension for 2x3 tables). Correction for multiple testing of CNA data (chapters 4.3.3 and 4.3.4) was performed using the Benjamini-Hochberg method (Benjamini and Hochberg, 1995). Briefly, the p-values were first sorted and ranked. The smallest p-value got rank 1, the second one rank 2, and the largest got rank N. Then, each p-value was multiplied by N and divided by its assigned rank to give the adjusted p-values. Adjusted p-values < 0.1 were accepted as significant. All other statistical tests, like Student's T-Test, multiple T-Tests, ANOVA with multiple comparisons, and correlation analysis (Spearman correlation), were performed using *GraphPad Prism* version 6.07.

- Student's T-test was performed without pairing of samples while assuming Gaussian distribution (parametric test) and the same standard deviation. The p-value was calculated in a two-tailed manner with a confidence level of 95 %.
- Multiple T-tests were performed without assumption of consistent SD (fewer assumptions option) using false discovery rate (FDR) correction with FDR <5 %, unless stated otherwise.
- Two-way ANOVA analysis was done without sample pairing. Multiple comparisons were performed between means of every column or with a control column by comparing column means within each row (simple effects within rows) using one family per row (recommended option). Significances between groups were calculated with Tukey's post-hoc test.
- One-way ANOVA was calculated without pairing and assuming Gaussian distribution. Multiple comparisons were performed between means of every column or with a control column followed by Tukey's post-hoc test (comparison of means between all groups) or the Dunnett method (comparison with mean of a control column).

3.16.2 Automatic annotation of cytobands in RefSeq files and aberration filtering

For increased comprehensibility, the code is shown in several chunks, but in practice the whole script was run at once, as soon as all files were named and placed correctly. The code segments presented in the following were exported to Word from *R-Studio* (referred to as *R* throughout this chapter) using the "Knit to Word" function of the *Rmarkdown* package for *R*, in order to preserve the syntax highlighting for easier reading.

3.16.2.1 Preparation of files, project directory and importing of data into R

First, the required reference file called "cytoband.txt.gz", which links the genomic coordinates in the RefSeq file to the corresponding cytoband, was downloaded from the UCSC Goldenpath database (link to the website in Table 2-13). The file was decompressed using the free software *7zip*, the resulting text (.txt) file imported into *Excel* and the file renamed to "UCSC_goldenpath_hg38_cytoBand.xlsx". Then, headers matching those in the RefSeq files were manually added to the reference file (Figure 3-1), before placing the finished reference into the *R* project directory.

chrom	start	end	cytoband	stain
chr1	0	2300000	p36.33	gneg
chr1	2300000	5300000	p36.32	gpos25
chr1	5300000	7100000	p36.31	gneg
chr1	7100000	9100000	p36.23	gpos25
chr1	9100000	12500000	p36.22	gneg

Figure 3-1 Example of the UCSC Goldenpath cytoband reference file structure. The table is an excerpt of the first few rows of the UCSC Goldenpath reference table. The "chrom" column provides information on the chromosome, while the "start" and "end" columns represent the genomic coordinates (bases) of the corresponding cytoband listed in the "cytoband" column. The "stain" column indicates how the respective cytoband is affected by Giemsa staining. The "stain" column was not required for the annotation procedure.

Second, all the RefSeq.xlsx files to be analyzed were renamed so that they had unique names traceable to the samples they originated from, before being copied to the R project directory.

Third, the required packages were installed in *R* and a subfolder created in the project directory for the output data using the following commands (Figure 3-2).

```
install.packages("readxl")
install.packages("writexl")
install.packages("tibble")
dir.create("output")
```

Figure 3-2 Code for preparation of R-Studio and the output folder. The „install.packages" function installs the packages named in the brackets. This only has to be done once as long as neither R-Studio nor the packages are deleted. The "dir.create" command creates a folder in the working directory with the name provided in the brackets.

After preparation of the files and the project directory, the necessary *R*-packages (*readxl, writexl, tibble*) were activated and the reference table as well as the sample file names were imported into *R* (Figure 3-3).

```
library(readxl)
library (writexl)
library(tibble)
ref = read_excel("UCSC_goldenpath_hg38_cytoBand.xlsx")
dataFiles = list.files(pattern="*.xlsx")
dataFiles = dataFiles[!grepl("UCSC_goldenpath_hg38_cytoBand.xlsx", dataFiles)]
```

Figure 3-3 Code for loading of R-packages and importing of reference and sample file names. The "library" command loads the previously installed *R* packages named in brackets, thereby activating them for the current session. The bottom three lines create objects in *R*, which contain the data that is specified by the function and the given files. The "ref" object is filled with the content of the UCSC Goldenpath reference. The object "dataFiles" is a list containing the names of all .xlsx files present in the project directory. The final line removes the entry with the reference's name from the list to avoid analysis of the reference.

Subsequently, the annotation process was carried out (chapter 3.16.2.2).

3.16.2.2 Running the annotation and extracting aberrant entries

Following the preparations above, the RefSeq files were annotated as follows (Figure 3-4):

```
for (i in 1:length(dataFiles))
{
  sample = read_excel(dataFiles[i])
  sample$cytoband = NA
  for (k in 1:length(ref$start))
  {
    ref_chrom = ref$chrom[k]
    ref_range = c(ref$start[k], ref$end[k])
    ref_band = ref$cytoband[k]
    target_rows = which(sample$chrom == ref_chrom & sample$start >
    ref_range[1]-1 & sample$end < ref_range[2]+1)
    sample[target_rows, 11] = ref_band
  }
  miss = which(is.na(sample$cytoband))
  miss_previous = miss -1
  sample[miss, 11] = sample[miss_previous,11]
  sample_aberrant = subset(sample, sample$status == "gain" | sample$status == "loss")
  write_xlsx(sample_aberrant, paste0("output/",  "aberrant_annotated_",dataFiles[i]),
  col_names=TRUE, format_headers=FALSE)
}
```

Figure 3-4 Code for cytoband annotation of the RefSeq files. The script uses two nested "for" loops. First, the outer loop handles the selection of the input file one by one, according to the "dataFiles" list that was created previously. After selection of one file, the inner loop screens each row of the UCSC reference file containing one individual cytoband, checks the start/end (=range) genomic coordinates of that cytoband, and adds the corresponding cytoband into a newly generated "cytoband" column for all rows in the sample RefSeq file that fit the correct range. Then, it proceeds with the next cytoband from the reference. Once the inner loop finishes annotation of all cytobands, the outer loop takes over again and looks for potential missing values (NA), which may have occurred due to a slight difference in the genomes used for the reference and the LP-Seq analysis, which likely altered the genomic coordinate ranges of some cytobands by a small margin. The missing values, which occurred in roughly 1.9 % of rows, are filled by the outer loop using the cytoband from the last previous row that still contains cytoband information. Next, the script extracts all entries classified as gain or loss and writes them into a new .xlsx file in the output folder. This way, the amount of data for subsequent manual screening (chapter 4.3.1) is significantly reduced by removing all genomic areas without aberrations. The output file is named "aberrant_annotated_" followed by the name of the input file. Once the writing of the file is finished, the outer loop selects the next input file from the "dataFiles" list and the process repeats until all files in the project directory have been annotated.

The script took about one minute to annotate all 91 RefSeq files corresponding to the 91 CNA profiles, which were found to be good enough for analysis (see Figure 4-4). Figure 3-5 depicts an excerpt of an annotated RefSeq file.

chrom	start	end	gene	copy_number	status	alteration_start	alteration_end	cytoband
chr1	1853390	1935276	CFAP74	22.0	gain	1419414	2365690	p36.33
chr1	1950768	1962192	GABRD	22.0	gain	1419414	2365690	p36.33
chr1	1980640	1981509	LOC105378591	22.0	gain	1419414	2365690	p36.33
chr1	1981909	2144159	FAAP20,PRKCZ	22.0	gain	1419414	2365690	p36.33

Figure 3-5 Example of an annotated RefSeq file. The "chrom", "start", and "end" columns are identical to the ones in the reference (see Figure 3-1). The "gene" column lists all known genes located in the given genomic range, while the "copy_number" and "cytoband" columns contain what their names suggest. The "status" column indicates whether the respective locus has been amplified or deleted. Finally, "alteration_start" and "alteration_end" represent the start and end coordinates of a continuous aberration, which can span over numerous loci, i.e. rows in the table.

After the annotation procedure, the annotated RefSeq files and corresponding CNA profiles were manually evaluated to create the ISCN annotation for *Progenetix* (chapter 4.3.1).

3.16.3 RNA-Seq data analysis

First, the quality of the raw data was assessed with *FastQC* version 0.11.5 (Andrews, 2010) and the individual quality reports were summarized using *MultiQC* version 1.7 (Ewels et al., 2016). Afterwards, according to the quality, up to 35 of the initial bases at the 5' end of reads and the complete Illumina adaptors at the 3' end were trimmed using *bbduk,* which represents a part of *bbmap* (Bushnell, 2014), in the version from 08.03.2019 with the following settings: "ftm=5 ktrim=r k=23 mink=11 hdist=1 ftl=35 qtrim=rl trimq=10 minlen=50 tbo tpe". Bases with an average quality < 10 at both ends were also removed. After trimming, the reads were mapped to the GRCh38 reference genome (gene annotation version 96) using *STAR* version 2.6.1c (Dobin et al., 2013) with default settings except for the use of "twopassMode" mapping. Subsequently, the mapping quality was investigated with *Qualimap* version 2.2.1 (García-Alcalde et al., 2012).

Second, *Featurecounts (Liao et al., 2014)* was applied with paired-end settings to convert sequencing reads to gene counts. The counts from all cells were retrieved with the command "*awk '{ a[FNR] = (a[FNR] ? a[FNR] FS : "") $5 } END { for(i=1;i<=FNR;i++) print a[i] }' $(ls -1v *)*" to generate a count table. With this count table, the cell quality was assessed with *scater* version 1.12.2 (McCarthy et al., 2017) and one cell with a high number of expressed genes, which may have been a doublet, was removed from the dataset. The "runPCA" function of *scater* was applied to cluster the samples according to their gene expression. One cell which was annotated as a healthy donor-derived NCC clustered with DCCs in this principal component analysis (PCA) and was removed from the dataset. *InferCNV* version 1.0.3 (Tickle et al., 2019) confirmed that this cell's CNV status differed from other healthy donor cells.

Third, differentially expressed genes between proliferating and non-proliferating DCCs were identified with *scDD* version 1.8.0 (Korthauer et al., 2016). All differentially expressed genes with the GO annotation "GO: 0007049 Cell Cycle" were extracted using the function getBM from *biomaRt* version 2.40.3 (Durinck et al., 2009). The pheatmap function of *R* version 3.6.0 (R Core Team, 2014) from the package *pheatmap* version 1.0.12 (Kolde, 2019) was applied to cluster the resulting genes and create a heatmap using the package's default settings.

Fourth, the "trendVar" and "decomposeVar" commands from *Scan* version 1.12.1 (Wilbert and Lueke, 2019) were employed to separate biologically relevant data from the technical noise. Afterwards, "denoisePCA" from *Scan* was applied to the previously generated PCA. Next, the correlation of the principal component 1 (PC1) variable with the expression level of marker of proliferation Ki-67 *(MKI67)* of each cell and the KI67 status from the corresponding primary tumor was investigated using the Spearman correlation function of *R.* Next, a multiple linear model was fitted with PC1 using *lmFit* and differentially expressed genes under the linear model with only PC1 were retrieved with the *eBayes* function. Both *lmFit* and *eBayes* are tools from the *Limma* package version 3.40.4 (Ritchie et al., 2015) for *R*. Genes that were upregulated in PC1 positive and negative directions, respectively, were extracted. For the positive direction, the significance threshold was set to $q < 0.01$; for the negative direction, the threshold was set to $q < 0.05$ and $logFC < -0.01$. The cutoff for the negative direction was modified, because with $q < 0.01$ only four genes were obtained and a sharp decrease of gene numbers at $logFC < -0.01$ was observed for the negative direction.

Lastly, the two previously identified gene sets were analyzed with *DAVID* version 6.8 (Huang et al., 2009b) with the Biological Process GO annotation to obtain the top enriched functions.

3.16.4 Generation of gene ontology (GO) term networks

For creation of GO term networks, the *Cytoscape* open source software (see Table 2-13) was utilized (Shannon et al., 2003). First, the *BiNGO* (v3.0.3) and *yFiles Layout Algorithms* (v1.0.2) applications (apps) were installed utilizing Cytoscape's built-in app manager function. Subsequently, *BiNGO* was started via the Apps tab of *Cytoscape* and a list of differentially expressed genes was copied into the newly opened window using the "Paste Genes from Text" function. The exact settings used for the tool are shown in Figure 3-6. Following the calculation of the network, the *yFiles* radial layout was applied to the network via the layout tab of *Cytoscape*. Next, the tool panel (found under the "View" tab) was used to reduce the scale of the network (increasing size of text relative to the nodes) and to rotate it to prevent overlapping of text. Lastly, some of the nodes were manually arranged to allow the network to fit on a single page while remaining legible.

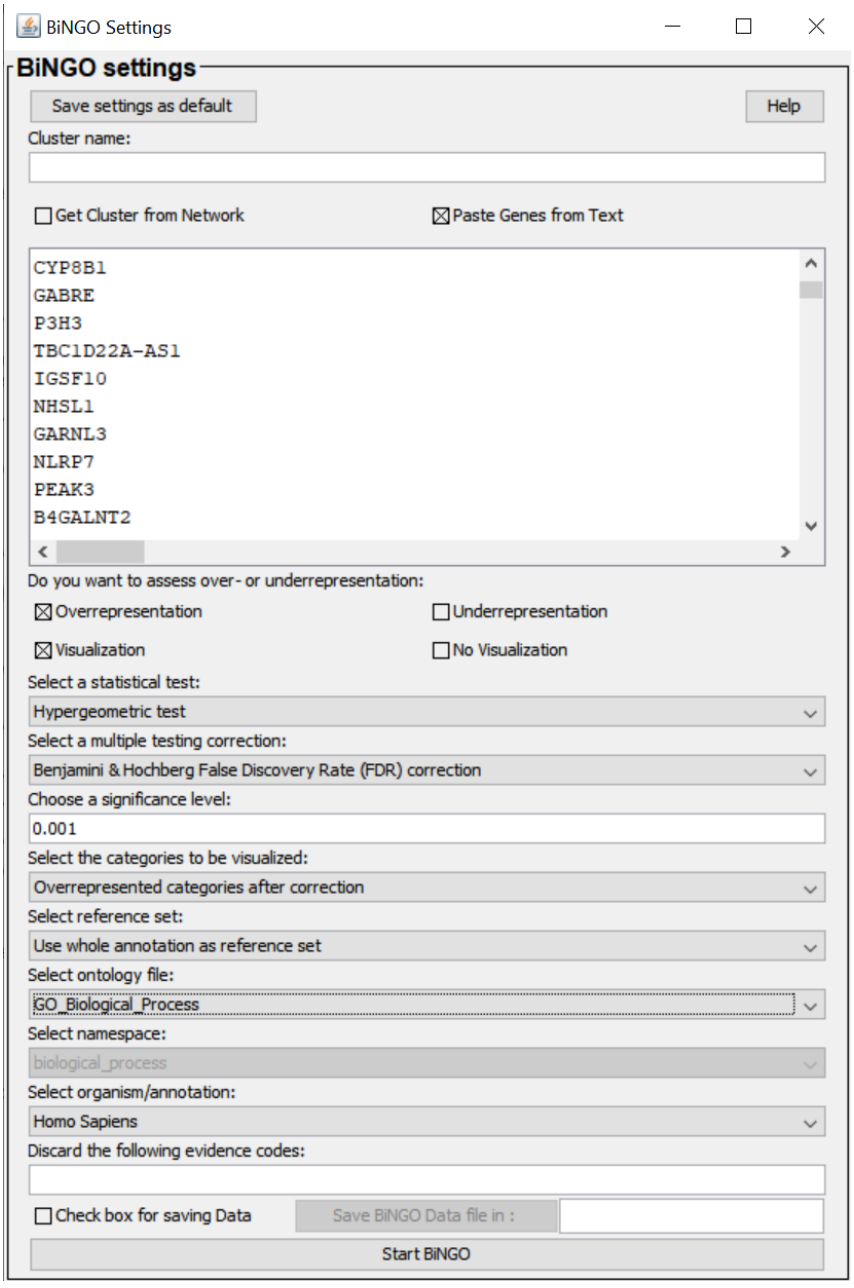


Figure 3-6 Settings used in the *BiNGO* tool to generate the GO term networks.

4. Results of transcriptomic and genomic characterization of DCCs

To identify factors contributing to the increased malignancy of LumB type breast cancer, both transcriptomic and genomic analyses of DCCs isolated from the BM of patients were performed. Chapter 4.1 provides an overview of the patient cohort utilized for these studies, while chapter 4.2 covers the identification of true DCCs among the EpCAM+ cells. After identification of the true DCCs, I began characterizing LumA and LumB DCCs. For this purpose, a detailed analysis of the observed copy number alterations (CNA, chapter 4.3), the proliferation state (chapter 4.4), and the overall gene expression profiles (chapter 4.5) of the EpCAM+ DCC collective was performed.

4.1 Overview of patient and single cell cohort

In the following sections, I will provide an overview of the total patient collective (chapter 4.1.1), clinical characteristics of patients included in the study (chapter 4.1.2), and subtype stratified numbers of isolated single EpCAM+ cells (chapter 4.1.3).

4.1.1 Overview of complete EpCAM+ BC patient cohort

Numbers of patient-derived BM samples received and screened, samples with detectable EpCAM+ cells in the BM (EpCAM+ patients), and EpCAM+ patients included in the final study are provided (Table 4-1). While M0 and HD groups did not differ in the rate of EpCAM-positivity (Fisher's exact test, p=0.3), M1 samples were EpCAM+ significantly more often than M0 samples (p=0.01).

Table 4-1 Overview of all processed BM samples. Percentages calculated relative to received BM samples. Statistics: Fisher's exact test on screened samples versus EpCAM+ samples; * M0 versus HD p=0.30; # M0 versus M1 p=0.01

Characteristic	M0	M1	HD
Received patient BM samples	313/313 (100 %)	20/20 (100 %)	52/52 (100 %)
Screened patient BM samples	247/313 (78.9 %) *#	18/20 (90 %) #	40/52 (76.9 %) *
EpCAM+ patient BM samples	100/313 (31.9 %) *#	13/20 (65 %) #	20/52 (38.5 %) *
EpCAM+ BM samples in study	88/313 (28.1 %)	13/20 (65 %)	16/52 (30.8 %)

Out of the 247 screened M0 BM samples (Table 4-1), 67 were classified as LumA and 113 as LumB (Table 4-2) according to the criteria listed in Table 2-2. Out of these BM samples, we detected EpCAM+ cells in 31 and 46, respectively. This did not represent a significant difference in the EpCAM-positivity rate (31/67=46.3 % versus 46/113=40.7 %, Fisher's exact test, p=0.53).

Table 4-2 Subtype stratification of screened and EpCAM+ M0 patients. Percentages were calculated relative to the total number of patients within either the screened or the EpCAM+ group, thereby indicating the frequency of each subtype in the respective patient collective. * Fisher's exact test on EpCAM+ and EpCAM- patients p=0.53.

Subtype	Screened (n=247)	EpCAM+ (n=100)
Luminal A (LumA)	67/247 (27.1 %)	31/100 (31 %) *
Luminal B (LumB)	113/247 (45.7 %)	46/100 (46 %) *
Luminal undefined	18/247 (7.3 %)	7/100 (7 %)
Her2 enriched	11/247 (4.5 %)	2/100 (2 %)
Triple negative / basal-like	26/247 (10.5 %)	12/100 (12 %)
No data	12/247 (4.9 %)	2/100 (2 %)

4.1.2 Characteristics of patients included in the study

Out of the 100 M0, 13 M1, and 16 EpCAM+ HDs, 88, 13, and 16, respectively, were included in the final study (Table 4-1). Twelve, seven, and four patients, respectively, were excluded, because the cells isolated from their BM either had insufficient quality or were lost during isolation. The characteristics and numbers of patients included in the downstream analyses of this study is provided below (Table 4-3). Interestingly, there was one M1 patient, whose PT was classified as non-malignant on initial diagnosis when the patient was still in the M0 stage. However, when we received the BM sample after the first distant metastasis was detected and the patient had reached M1 stage (now classified as LumA), a total of 17 EpCAM+ cells could be isolated from that patient's BM, of which eight had a good WTA quality. One of them was identified as a true DCC by CNA analysis later on (chapter 4.2.2), which indicates that even seemingly non-malignant tumors are able to disseminate to other organs. Overall, one of 13 M1 patients (8 %) was already in the M1 stage at initial diagnosis, while the remaining twelve (92 %) progressed to the M1 stage later on. Moreover, it is worth noting that only eight out of 88 M0 (9.9 %) and two out of twelve M1 patients (16.7 %) received neoadjuvant therapy. All other patients were untreated when the BM samples were taken. The KI67 status (Table 4-3) was stratified into three levels (low ≤10 %, medium 11-19 %, high ≥20 %) instead of two (low <13 %, high ≥14 %), because all LumA or LumB patients with a medium KI67 level were excluded for selection of cells for RNA-Seq later on (chapter 4.5.1).

Table 4-3 Clinical characteristics of EpCAM+ patients included in the study. Percentages were calculated relative to the total number of patients in each group (M0/M1/HD) given in the top row. Please note that percentages are rounded, meaning that in some cases they do not add up to exactly 100 %. In the cases of "Luminal undefined" and "No data" categories, the tumor slices could not be found in the sample archive anymore or the patients were unknown, probably due to typing errors during transfer of samples to our laboratory. NA = not applicable

Characteristic		M0 (n=88)	M1 (n=13)	HD (n=16)
Age at surgery (years)				
	Mean	54.6	62.3	66.2
	Standard deviation	13.5	13.0	12.9
	No data	0	1	0
Estrogen receptor status				
	Positive	75 (85 %)	11 (85 %)	NA
	Negative	11 (13 %)	2 (15 %)	NA
	No data	2 (2 %)	0 (0 %)	NA
HER2 amplification				
	Positive	10 (11 %)	1 (8 %)	NA
	Negative	71 (81 %)	12 (92 %)	NA
	No data	7 (8 %)	0 (0 %)	NA
Molecular intrinsic subtypes				
	LumA	29 (33 %)	1 (8 %)	NA
	LumB	39 (44 %)	6 (46 %)	NA
	Luminal undefined	7 (8 %)	4 (31 %)	NA
	HER2 enriched	1 (1 %)	1 (8 %)	NA
	Triple negative (TNBC)	10 (11 %)	1 (8 %)	NA
	No data	2 (2 %)	0 (0 %)	NA
Nodal status				
	0	59 (67 %)	4 (31 %)	NA
	1	20 (23 %)	4 (31 %)	NA
	2	5 (6 %)	1 (8 %)	NA
	3	2 (2 %)	2 (15 %)	NA
	No data	2 (2 %)	2 (15 %)	NA

Characteristic		M0 (n=88)	M1 (n=13)	HD (n=16)
Grading (Elston & Ellis)				
	1	6 (7 %)	1 (8 %)	NA
	2	57 (65 %)	2 (15 %)	NA
	3	22 (25 %)	7 (54 %)	NA
	No data	3 (3 %)	3 (23 %)	NA
Histology				
	Invasive ductal / NST	60 (68 %)	9 (69 %)	NA
	Invasive lobular	15 (17 %)	1 (8 %)	NA
	Invasive ductal+lobular	4 (5 %)	0 (0 %)	NA
	Invasive special types	2 (2 %)	2 (15 %)	NA
	Carcinoma *in situ*	4 (5 %)	0 (0 %)	NA
	No data	3 (3 %)	1 (8 %)	NA
KI67 expression in PT (%)				
	≤ 10 %	31 (35 %)	1 (8 %)	NA
	11-19 %	8 (9 %)	0 (0 %)	NA
	≥ 20 %	39 (44 %)	7 (54 %)	NA
	No data	10 (11 %)	5 (38 %)	NA

4.1.3 Single EpCAM$^+$ cell collective

Analogous to the patients, the isolated EpCAM$^+$ single cells were classified according to the subtype of the respective patient they originated from (Table 4-4). The table contains only cells with sufficient genome and transcriptome quality. The quality was assessed by multiplex endpoint PCR on three transcripts (chapter 3.3). Cells with at least one out of three bands on the agarose gel (chapter 3.4) were considered suitable for downstream analyses. Regardless of the number of bands, there were cells, of which there was only either a WTA or a WGA product available, since these procedures sometimes failed independently of each other. Due to this fact and stricter quality criteria for some applications, not all cells could be included in all downstream analyses.

Table 4-4 Numbers of single EpCAM$^+$ cells included in the study stratified by subtype. Percentages were calculated relative to the total number of patients in each group (M0/M1/HD) given in the top row. NA = not applicable

Subtype	M0 cells (n=304)	M1 cells (n=74)	HD cells (n=47)
Luminal A	106/304 (34.9 %)	15/74 (20.3 %)	NA
Luminal B	127/304 (41.8 %)	24/74 (32.4 %)	NA
Luminal undefined	30/304 (9.9 %)	27/74 (36.5 %)	NA
HER2 enriched	3/304 (1 %)	2/74 (2.7 %)	NA
Triple negative / basal-like	33/304 (10.9 %)	6/74 (8.1 %)	NA
No data	5/304 (1.6 %)	0/74 (0 %)	NA

As no difference between M0 LumA and LumB subtypes had previously been observed regarding the EpCAM-positivity rate (chapter 4.1.1), a two-sided Student's T-test was performed on the number of cells isolated per M0 LumA or LumB patient (Figure 4-1). Similar to the EpCAM-positivity rate, this did not result in a significant difference (p=0.37).

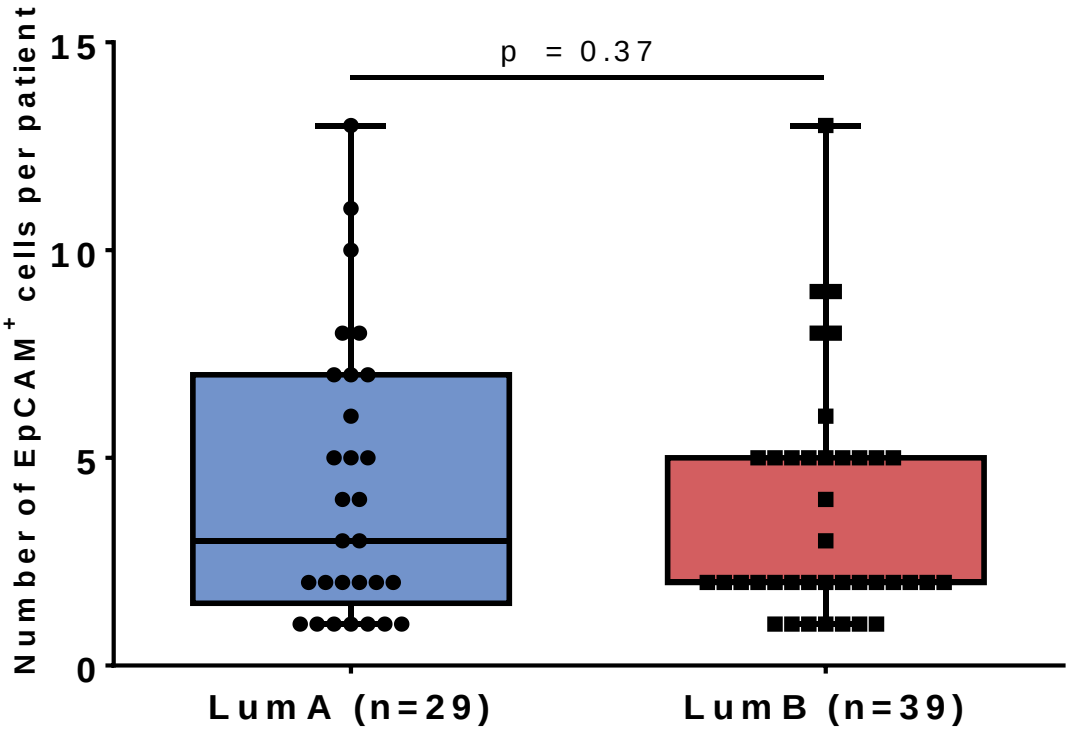


Figure 4-1 Numbers of EpCAM⁺ cells isolated from each M0 LumA and LumB patient included in the study. The box plot illustrates the numbers of cells isolated from in individual M0 LumA or M0 LumB patient included in the study. The whiskers represent the minimal and maximal values. Statistics: Student's T-test (chapter 3.16.1.2).

Since there was also a considerable number of EpCAM⁺ cells isolated from HDs (Table 4-4), a method to distinguish true cancer cells from the confounding non-cancerous EpCAM⁺ population identified in HDs had to be developed (chapter 4.2).

4.2 Identification of true DCCs

As shown above, 20 out of 40 (50 %) of the screened HD BM samples contained EpCAM⁺ cells (Table 4-1), which likely belong to the erythroid progenitor cell lineage according to the literature (Bühring et al., 1996; Lammers et al., 2002; Gužvić et al., 2014). Consequently, I needed to find a way to distinguish those EpCAM⁺ cells with a cancerous nature (i.e. DCCs) from the EpCAM⁺ non-cancer cells (NCCs). Initial attempts by the former PhD student Dr. Gundula Haunschild using endpoint PCR for targeted analysis of several epithelial markers – mostly cytokeratins (various *KRT* genes, referred to as CK)- , *ERBB2* wild type (WT) and the *ERBB2Δ16* mutant, as well as pyruvate kinase M 1/2 (*PKM1/2*) were inconclusive (Haunschild, 2013). Therefore, a gene expression signature for qPCR, which would allow us to distinguish DCCs from NCCs, was established in cooperation with two colleagues (chapter 4.2.1). For this purpose, we took a global approach by profiling the whole transcriptome of selected cells by microarray. In the end, this signature did worked quite well, but needed further conformation, so CNA profiling was also performed to distinguish cells according to the presence of genomic aberrations similar to what Gundula Haunschild did with metaphase comparative genomic hybridization (mCGH, chapter 4.2.2).

4.2.1 DCC identification by qPCR

Chapters 4.2.1.1 and 4.2.1.2 are based on Nina Patwary's thesis, which has not been submitted yet (Patwary, in preparation). I want to provide her data here to make the following experiments more comprehensible, since I was building on her work.

4.2.1.1 Identification of signature genes by microarray

A total of 25 DCCs from M0 patients, which were previously shown to have genomic aberrations by means of mCGH (Haunschild, 2013), and 22 HD-derived NCCs with unknown CNA status were chosen for gene expression profiling and microarrays of the single cells were prepared (Patwary, in preparation). The unsupervised hierarchical clustering analysis yielded a large number of differentially expressed genes (Figure 4-2). In total, there were 1060 differentially expressed genes, 570 of which were more highly expressed in M0 DCCs, while 490 were more highly expressed in the NCCs from the HDs. Interestingly, four of the NCCs clustered with the M0 DCCs, but apart from these cells, we observed a robust separation of M0 DCCs and NCCs (Figure 4-2).

After clustering, we selected six genes from the list of the differentially expressed genes according to their fold change and p-value. These were *alpha-hemoglobin stabilizing protein* (*AHSP*), *peptidylprolyl isomerase A* (*PPIA*), and *carboanhydrase 1* (*CA1*), which were more highly expressed in M0 DCCs (DCC-genes), as well as *AHNAK nucleoprotein* (*AHNAK*), *proto-oncogene C-Jun* (*JUN*), and *krueppel-like factor 6* (*KLF6*), which were more highly expressed in the NCCs (NCC-genes). Afterwards, the novel candidate genes were validated (chapter 4.2.1.2).

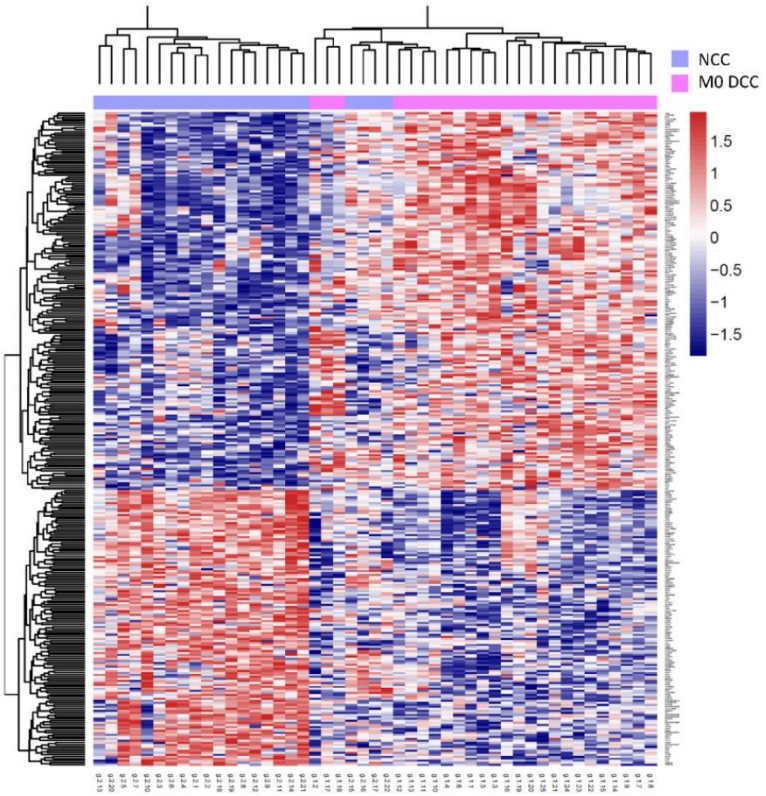


Figure 4-2 Heat map of the differentially expressed genes of DCCs versus HD cells. The heat map depicts the relative expression of differentially expressed genes of 25 aberrant M0 patient-derived DCCs and 22 HD-derived EpCAM$^+$ NCCs. The columns represent the individual cells, while the rows represent the differentially expressed genes. There are two separate groups of cells in the data as indicated by the different branches at the top of the figure. Both groups displayed up-regulated (red tiles) and down-regulated (blue tiles) genes. The two separate clusters comprise almost exclusively M0 DCCs (pink) or NCCs (purple). Adapted from (Patwary, in preparation).

4.2.1.2 Validation of signature genes

First, endpoint PCR as well as qPCR assays were established for all six previously identified genes (see chapter 4.2.1.1) using all cells analyzed by microarray (= training cohort). Taking together the data from both PCRs, we selected *AHSP, CA1, AHNAK*, and *JUN* for the final signature (Patwary, in preparation). *PPIA* and *KLF6* were eliminated from the candidate list, because either their differential expression between M0 DCCs and NCCs could not be confirmed by PCR (*PPIA*) or they were too frequently expressed in M0 DCCs (*KLF6*).

The qPCR revealed that the two DCC-genes *AHSP* and *CA1* were strongly expressed in all DCCs, but also in a few of the NCCs (left part of Figure 4-3). In those NCCs, in which the DCC-genes were expressed, the transcripts were present in amounts similar to those in DCCs. Unsurprisingly, two and three, respectively, of the *AHSP*- and *CA1*-expressing NCCs were among the NCCs clustering with M0 DCCs (Figure 4-2). As expected, the two NCC-genes *AHNAK* and *JUN* were not expressed in the majority of DCCs (right section of Figure 4-3). In a few DCCs, the NCC-genes were present at relatively low levels compared to NCC-genes. At the same time, several NCCs were negative for the NCC-genes. Nevertheless, there was a highly significant difference in the overall expression of each of the four genes between DCCs and NCCs (multiple T-tests, all p-values < 0.0001).

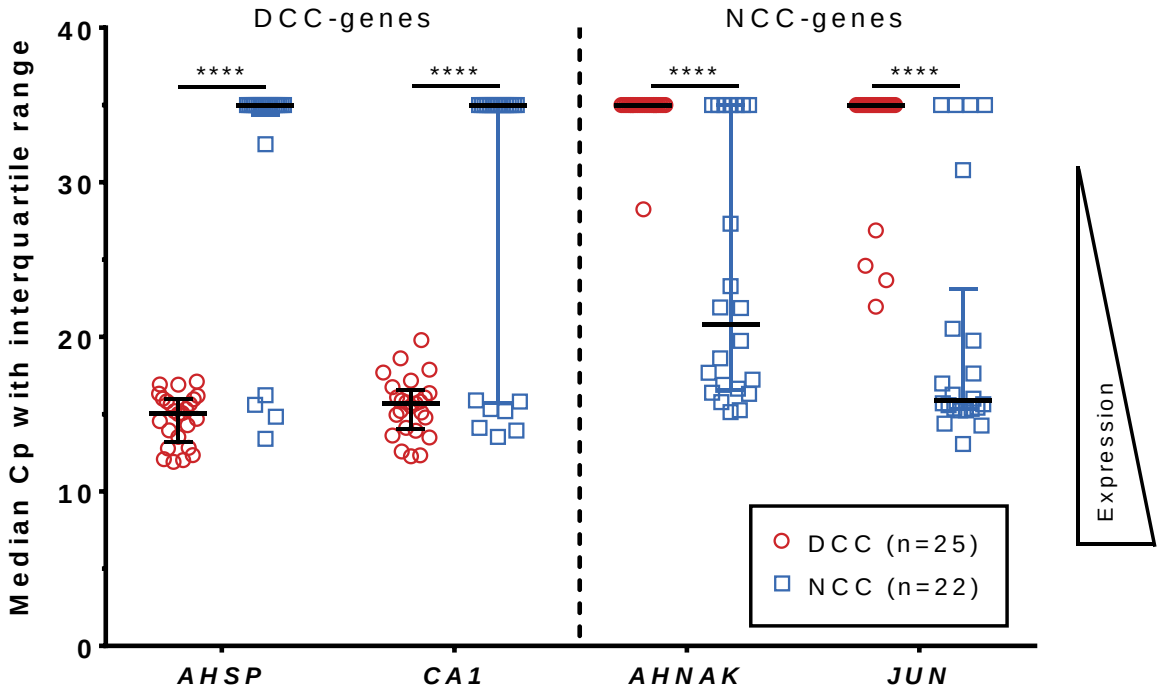


Figure 4-3 Absolute expression of the four DCC signature genes in the training cohort. The plot shows expression of the four DCC signature genes in the training cohort as median crossing point-(Cp)values with interquartile range represented by the whiskers. Expression in DCCs is represented by red circles, while expression in the NCCs is represented by blue circles. Expression is inversely correlated with Cp as indicated by the wedge on the right of the graph. Statistics: multiple T-tests (chapter 3.16.1.2); **** $p < 0.0001$. Adapted from (Patwary, in preparation)

Using the raw Cp values, the cells were classified into five different groups according to the criteria detailed in Table 4-5 below. The five classes were chosen to best reflect the continuous nature of the data, since many cells did not display a black-and-white expression profile, thereby preventing a binary separation. Out of the 25 M0 DCCs, 20 (80 %) were classified as DCCs, while the other five cells (20 %) had a DCC-like expression profile (Table 4-6). In contrast, 13 out of 22 (59.1 %) of NCCs expressed the NCC profile. Interestingly, four NCCs were classified as DCC, which matches the previous observation that four NCCs were clustering with the M0 DCCs (Figure 4-2).

Table 4-5 Cutoff Cp values for classification of EpCAM⁺ cells according to DCC signature. The table lists the different criteria for the five cell classes resulting from the DCC signature. The DCC and NCC classes represent the left and right extremes with expression of only DCC- or NCC-genes, while the other three classes represent different combinations of intermediate expression levels. The three intermediate classes possess two or three alternative conditions, indicated by the separate sub columns. The words "One" and "Other" mean in this case that either one of the two genes from a set (*AHSP/CA1* or *AHNAK/JUN*) needed to have a Cp value ≤ 30 while the other was > 30 for example. A Cp value below 25 was considered high expression, while Cp values above 30 were considered negative. Values between 25 and 30 were regarded as weakly positive. The blue and red wedges illustrate the rationale behind the classification that the five classes should represent the relative ratios of DCC genes to NCC genes instead of absolute differences to reproduce the continuous nature of the observed gene expression.

	DCC	**DCC-like**		**Undefined**			**NCC-like**		**NCC**
DCC genes (*AHSP/ CA1*)	Both Cp ≤ 25	Both Cp < 30	One Cp < 30 Other Cp ≥ 30	Both Cp ≤ 30	Both Cp ≥ 30	One Cp < 30 Other Cp ≥ 30	One Cp < 30 Other Cp ≥ 30	Both Cp ≥ 30	Both Cp ≥ 30
NCC genes (*AHNAK /JUN*)	Both Cp ≥ 30	One Cp < 30 Other Cp ≥ 30	Both Cp ≥ 30	Both Cp ≤ 30	Both Cp ≥ 30	One Cp < 30 Other Cp ≥ 30	Both Cp < 30	One Cp < 30 Other Cp ≥ 30	Both Cp ≤ 25

Table 4-6 Classification of DCCs and NCCs of the trainings set according to DCC signature gene expression. The percentages are relative to the total number of cells for each row given in the "Group" column.

Group	**DCC**	**DCC-like**	**Undefined**	**NCC-like**	**NCC**
M0 (n=25)	20/25 (80 %)	5/25 (20 %)	0/25 (0 %)	0/25 (0 %)	0/25 (0 %)
NCC (n=22)	4/22 (18.2 %)	1/22 (4.5 %)	1/22 (4.5 %)	3/22 (13.6 %)	13/22 (59.1 %)
Total (n=47)	24/47 (51.1 %)	6/47 (12.8 %)	1/47 (2.1 %)	3/47 (6.4 %)	13/47 (27.7 %)

To assess the performance of the qPCR-based DCC signature, several statistical metrics were calculated (formulas and definitions in chapter 3.16.1.1). In order to obtain a 2x2 confusion matrix to enable the calculations, the DCC-like, undefined, and NCC-like classes were excluded and the calculations were carried out only on the DCC and NCC cases from the training set (Table 4-7 left table), which represented a stringent signature. Overall, the prevalence of the mCGH-aberrant DCCs was 54.1 %, the accuracy 89.2 % and the misclassification rate 10.8 %. Moreover, the sensitivity of the DCC signature was 100 %, while the specificity was 76.5 %. Lastly, the positive predictive value (PPV) was 83.3 % and the negative predictive value (NPV) was 100%.

However, the calculation above ignored the existence of ten intermediate cases (see Table 4-6) and may therefore have overestimated the signature's performance. Consequently, a dataset with relaxed selection criteria was evaluated, which combined DCC and DCC-like classes as positives and NCC-like and NCC classes as negatives, in order to incorporate more samples and improve the performance estimation (Table 4-7 right table). By this approach, only one undefined cell was excluded from analysis. In this dataset, the prevalence of mCGH-aberrant DCCs was 54.4 %, the accuracy 89.1 % and the misclassification rate 10.9 %. Furthermore, the sensitivity of the signature was 100 %, while the specificity was 76.2 %. Additionally, the positive predictive value (PPV) and the negative predictive value (NPV) were 83.3 % and 100%, respectively. Interestingly, the relaxed signature brought about only marginal differences in the performance estimates compared to the conservative variant calculated above.

Table 4-7 Confusion matrices for assessment of qPCR signature performance on training set. The columns represent the origin of the analyzed cells: either mCGH-aberrant DCCs or HD-derived EpCAM+ NCCs. The latter were not tested by mCGH. In contrast, the rows represent the respective classifications resulting from the qPCR signature. The left table represents the dataset obtained with stringent selection criteria, while the right table displays the dataset of the relaxed criteria that included DCC-like and NCC-like cells.

Stringent		Source	
		DCC	NCC
qPCR signature	DCC	20	4
	NCC	0	13

Relaxed		Source	
		DCC	NCC
qPCR signature	DCC/DCC-like	25	5
	NCC/NCC-like	0	16

Judging from these data, the signature was considered sufficient for profiling of the whole single cell collective (chapter 4.2.1.3).

4.2.1.3 Application of the qPCR signature on the EpCAM+ cell collective

The expression in all available patient BM-derived EpCAM+ single cells was measured by qPCR in an attempt to identify additional DCCs for downstream analysis. The classification of the cells was performed using the previous criteria (Table 4-5).

The data revealed that 48 % of M0 EpCAM+ cells expressed the DCC pattern while 15 % displayed to the NCC pattern (Table 4-8). Interestingly, 17.5 % of NCCs showed the DCC pattern, while only 37.5 % of NCCs expressed the NCC expression pattern. Of note, only 19.6 % of the M1-derived EpCAM+ cells ended up in the DCC class compared to 28.6 % of M1 cells matching the NCC pattern. However, it was expected that the signature would not work well on M1 cells, since no M1 DCCs were included in the microarray to generate the signature. Together with the 32.7 % of M1 cells assigned to the NCC-like group, 61.8 % of the M1 cells expressed either the NCC or the NCC-like patterns, while only 22.5 % of M0 cells fell into the NCC or NCC-like categories. An analysis of the performance of the signature on this larger dataset will follow later (end of chapter 4.2.2.2).

Table 4-8 Classification of EpCAM+ cells according to DCC signature gene expression. The percentages are relative to the total number of cells for each group given in the "Group" column. Percentages are rounded. Therefore, they may not add up to exactly 100 %. The numbers include the 47 cells of the microarray training set (see Table 4-6).

Group	DCC	DCC-like	Undefined	NCC-like	NCC
M0 (n=272)	131/272 (48 %)	50/272 (18 %)	30/272 (11 %)	20/272 (7 %)	41/272 (15 %)
M1 (n=56)	11/56 (20 %)	5/56 (9 %)	6/56 (11 %)	18/56 (32 %)	16/56 (29 %)
NCC (n=40)	7/40 (18 %)	7/40 (18 %)	2/40 (5 %)	9/40 (23 %)	15/40 (38 %)
Total (n=368)	149/368 (41 %)	62/368 (17 %)	38/368 (10 %)	47/368 (13 %)	72/368 (20 %)

Taken together the qPCR signature looked promising, but the data on the whole EpCAM+ cell collective were not conclusive enough on their own. Additional evidence was required to confirm the accuracy of the qPCR signature. For that reason, the genomic aberrations of the EpCAM+ cells were investigated (chapter 4.2.2).

4.2.2 DCC identification by detection of genomic aberrations

4.2.2.1 LowPass-Sequencing results

As Dr. Haunschild had already performed mCGH on several of the M0 and M1 patient-derived EpCAM$^+$ cells to confirm their cancer origin by detection of CNAs (Haunschild, 2013), I also turned to CNA profiling to identify the true DCCs in the EpCAM$^+$ cell collective. For this purpose, the novel *Ampli1* LowPass-Sequencing (LP-Seq) technology was utilized, as it allowed profiling of all available WGA products due to its relatively low price compared to mCGH. In total, 282 WGA products were processed. The schematic below (Figure 4-4) outlines the selection process and numbers of profiles passing each step of the workflow. From 262 of these, CNA profiles were obtained for data analysis. Next, those profiles with sufficient quality were selected to make a definitive statement whether a cell had an aberrant or balanced genome. A few example profiles are shown below (Figure 4-5). Figure 4-5a depicts the high-quality CNA profile of an NCC with a balanced genome, while panels b-c display high quality profiles of an M0- and M1-stage DCC, respectively. Unfortunately, the majority of profiles were of too low quality for analysis as displayed in Figure 4-5d+e. Known artifacts like a minor gain in the beginning of the short (p) arm of chromosome 1 and small gains in telomeric or centromeric regions of all chromosomes were not counted as aberrations. Additionally, all profiles were discussed with several experienced postdocs, in order to draw from their knowledge of CNA profiles. In the end, 91 profiles were selected for further analysis, while 171 (73 low quality and 98 insufficient quality) profiles were excluded, because they either could not be interpreted clearly due to a high spread of the reads (low quality, Figure 4-5d) or were just noise (insufficient quality). The 91 high quality profiles are provided in the appendix (chapter 12.1.1).

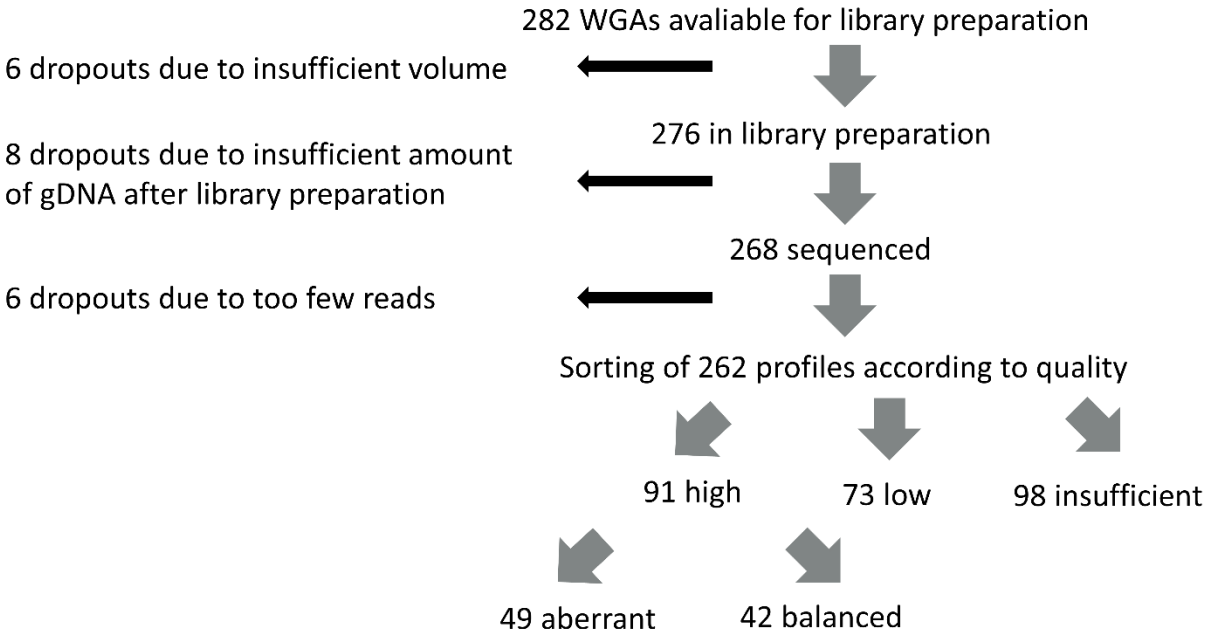


Figure 4-4 Schematic of the LP-Seq profile generation process and resulting cell numbers per step. The schematic illustrates the numbers of WGA samples initially processed and how many samples or profiles were lost or excluded at each step of the workflow.

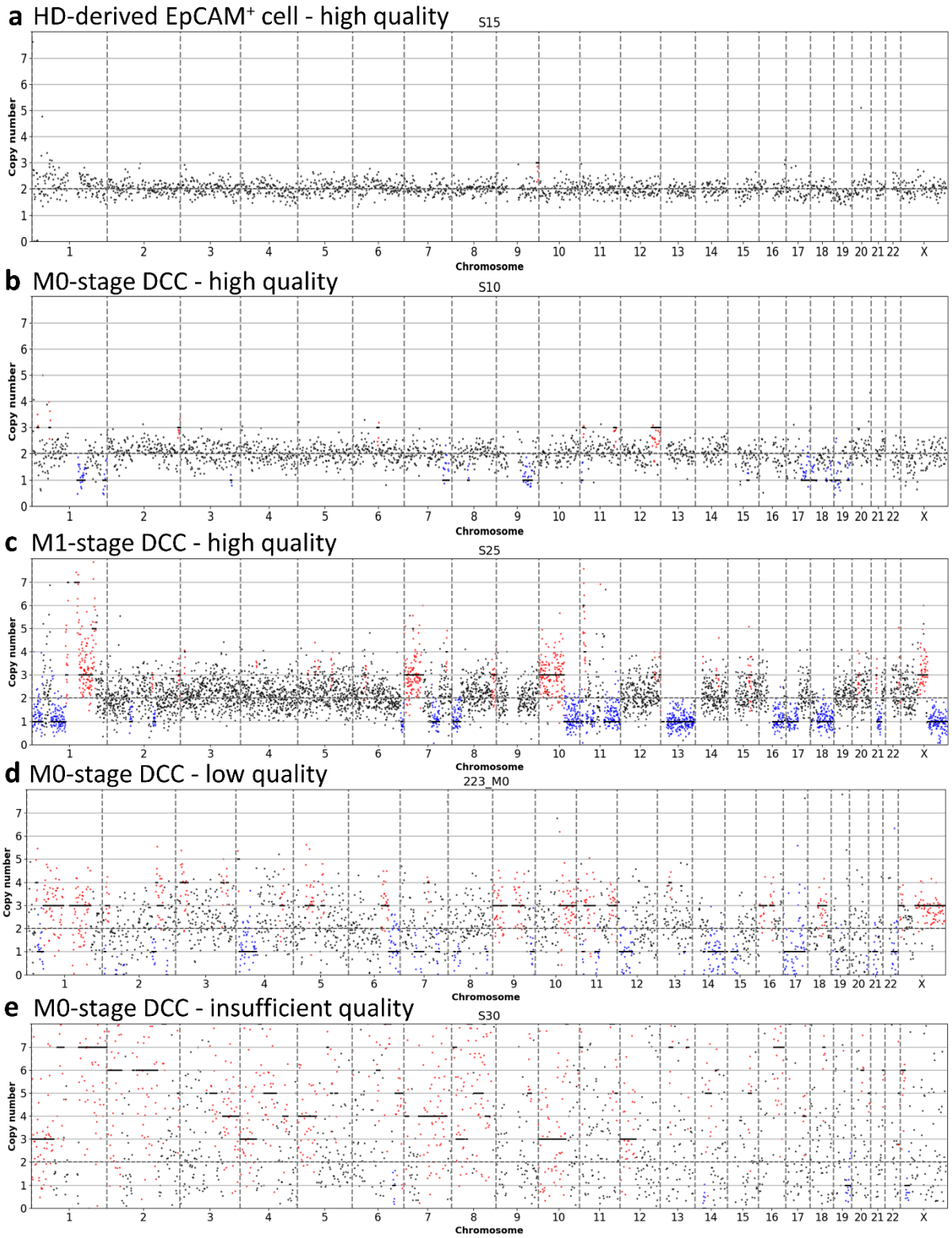


Figure 4-5 Exemplary LP-Seq profiles. The profiles (a-c) represent high quality profiles, while profiles (d-e) illustrate what low or insufficient quality profiles looked like. Each profile shows the distribution of the sequencing reads (dots) across the 22 autosomes and the X chromosome (X-axis) plotted against the calculated copy number (y-axis). (a) Control cell, i.e. HD-derived NCC with a balanced genome. This is illustrated by the reads clustering around the line indicating a normal copy number of two. (b) Profile from an M0 patient, which displays several smaller aberrations identifying it as a true DCC. Blue dots indicate genomic losses, while red dots represent genomic gains. (c) Profile from an M1 patient showing several large-scale gains and losses. (d) Low quality profile of an M0-derived cell. (e) Insufficient quality profile of an M0-derived cell. Profiles as shown in (d) and (e) were excluded from further analysis.

Next, I wanted to assess, whether the quality of the initial gDNA was connected to the quality of the CNA profiles. For that purpose, the profile quality was compared to the initial quality of the genome represented by the genomic integrity index (GII; ranging from one to three depending on the number of visible PCR bands from the WGA-QC [see chapter 3.3.2]). A robust link between the LP-Seq profile quality and the WGA quality was observed (Figure 4-6). A Chi^2 test on the numbers of profiles stratified by the GII and profile quality revealed a very strong association of the two variables (p=0, see Table 4-9).

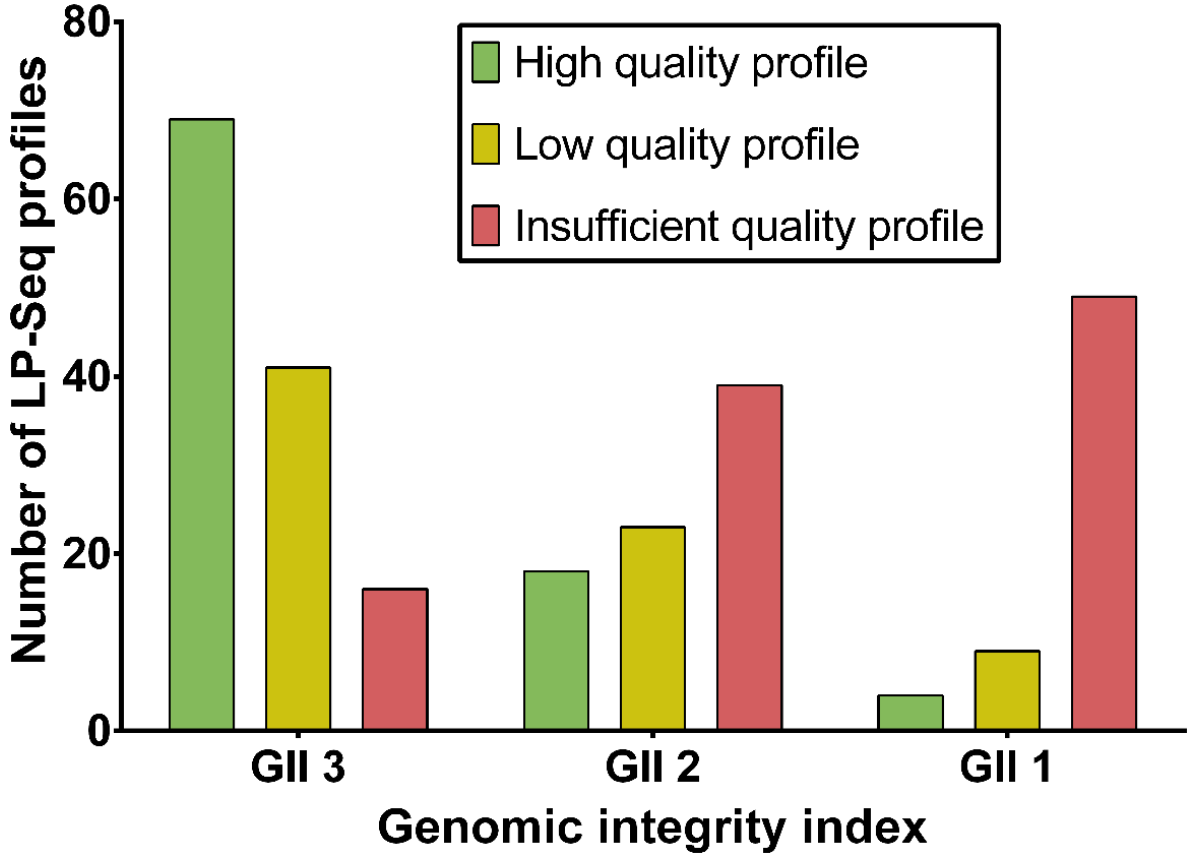


Figure 4-6 Link of LP-Seq profile quality with genomic integrity index. The bar plot illustrates the numbers of high, low, and insufficient quality LP-Seq-derived CNA profiles stratified by the GII of the original WGA products.

Table 4-9 Contingency table of GII and corresponding LowPass-Seq-derived CNA profile quality. Insufficient quality profile numbers in the table consist of 98 actual profiles with insufficient quality and six dropouts (see Figure 4-4), which had too few sequencing reads, resulting in a total of 104 insufficient profiles.

		Profile quality		
		High	**Low**	**Insufficient**
GII	**3**	69	41	16
	2	18	23	39
	1	4	9	49
Total		**91**	**73**	**104**

The final numbers of aberrant and balanced cells, both in total and stratified according to the BC subtype within M0, M1, and HD groups are listed below (Table 4-10). Fisher's exact test revealed that the ratio of aberrant to balanced cells in the M0 group was significantly different from that in the M1 group (p=0.001), while the ratio was similar between LumA and LumB cells within the M0 group (p=1).

Table 4-10 Overview of aberrant and balanced cells identified by LP-Seq. The table shows how many of the individual profiles out of the 91 analyzed ones were classified aberrant or balanced. The underlined data are the total cell counts for the M0, M1, and NCC groups, while the values below (italics) are counts stratified according to the BC subtype. Percentages were calculated relative to the total cell number of each line. Statistics: Fisher's exact test performed on cell counts (chapter 3.16.1.2) * LumA vs LumB aberrant and balanced cells (p=1) # M0 vs. M1 aberrant and balanced cells (p=0.001)

Group	Total cells	Aberrant	Balanced
M0 (total)	59	27/59 (45.8 %) #	32/59 (54.2 %) #
LumA	*17*	*8/17 (47.1 %) **	*9/17 (52.9 %) **
LumB	*27*	*14/27 (51.9 %) **	*13/27 (48.1 %) **
Lum undefined	*7*	*1/7 (14.3 %)*	*6/7 (85.7 %)*
TNBC	*7*	*3/7 (42.9 %)*	*4/7 (57.1 %)*
Missing data	*1*	*1/1 (100 %)*	*0/1 (0 %)*
M1 (total)	22	19/22 (86.4 %) #	3/22 (13.6 %) #
LumA	*1*	*1/1 (100 %)*	*0/1 (0 %)*
LumB	*5*	*5/5 (100 %)*	*0/5 (0 %)*
Lum undefined	*14*	*11/14 (78.6 %)*	*3/14 (21.4 %)*
TNBC	*2*	*2/2 (100 %)*	*0/2 (0 %)*
NCC (total)	10	3/10 (30 %)	7/10 (70 %)
Total	**91**	**49/91 (53.8 %)**	**42/91 (46.2 %)**

Following evaluation of the LP-Seq data, I examined how comparable the results were with the previous mCGH data of Dr. Haunschild, before combining the CNA data from both methods for further analysis (chapter 4.2.2.2).

4.2.2.2 Combination of mCGH and LowPass-Seq data

In order to select the maximum number of DCCs for further analyses, I aimed at combining the genomic aberration data from the LP-Seq with Gundula Haunschild's mCGH data (Haunschild, 2013). First, the classification results of 24 cells, for which both LP-Seq and mCGH data were available (for example profiles see Figure 4-7 and Figure 4-8 on the following pages) were compared. The two methods agreed in 87.5 % of cases (aberrant and balanced matches combined, Table 4-11) and the association of the results was highly significant (Fisher's exact test: p=0.0008). Among the 14 cells, which were aberrant in both mCGH and LP-Seq, was one sample that was tested twice using mCGH. It was aberrant in the first experiment, while it was balanced in the second. Since this particular sample was also aberrant in the LP-Seq analysis, the cell was finally classified as aberrant.

Table 4-11 Association of LP-Seq and mCGH results. The table represents a contingency table of the 24 single cells, which were tested both by mCGH and by LP-Seq and the resulting classification of the assess cells. The cohort comprised both M0 and M1 patient-derived EpCAM$^+$ cells. The data represent cell counts, the percentages are relative to the total number of analyzed cells. Matching results (positive and negative) are underlined. Statistics: Fisher's exact test was performed on the cell counts (chapter 3.16.1.2); p=0.0008

	LowPass aberrant	LowPass balanced
mCGH aberrant	14/24 (58.3 %)	1/24 (4.2 %)
mCGH balanced	2/24 (8.3 %)	7/24 (29.2 %)

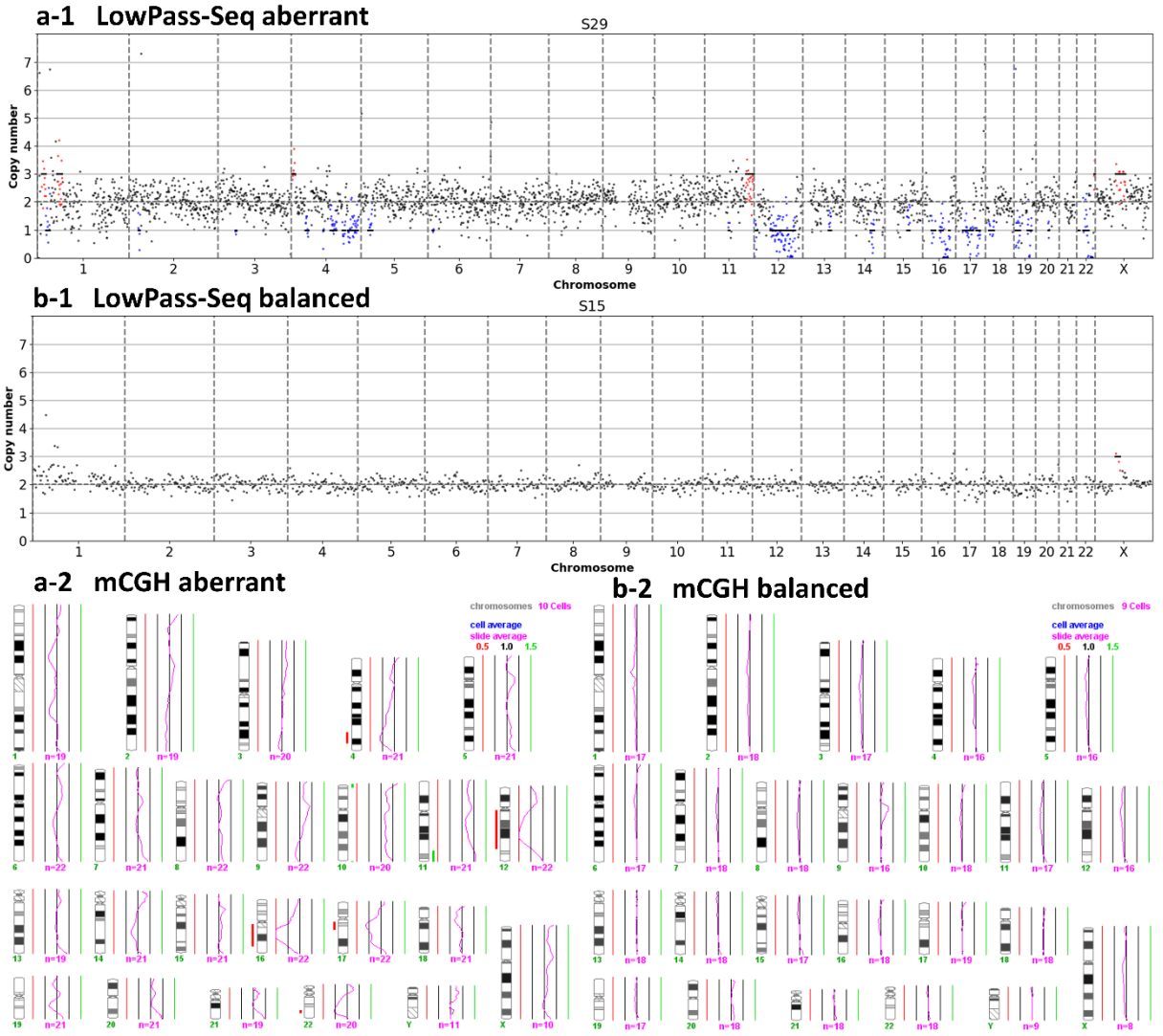


Figure 4-7 Examples of matching LP-Seq and mCGH profiles from the same single cells. The figure displays two matched pairs of LP-Seq and mCGH profiles that were in accordance. (a-1) Aberrant M0 LP-Seq profile. (a-2) Aberrant mCGH profile generated from the same single cell genome as the one shown in panel a-1. (b-1) Balanced M0 LP-Seq profile. (b-2) Balanced mCGH profile generated from the same single cell genome as the one shown in panel b-1.

Judging from the observation that the two methods were in accordance in the majority of cases (Table 4-11), I was confident that the mCGH and LP-Seq data were comparable. Therefore, the EpCAM⁺ cell collective was re-evaluated and all cells with aberrations confirmed by either mCGH or LP-Seq - and also those aberrant according to both methods – were accepted as true DCCs. The resulting numbers of cells are summarized in Table 4-12. Similar to the result of the LP-Seq data alone (compare Table 4-10), Fisher's exact test confirmed that M1 DCCs carried CNAs significantly more often than M0 DCCs (p=0.002), while LumA and LumB cells from M0 patients were similar (p=1). The cells identified as genomically aberrant were considered true DCCs, and only those were used for further detailed analyses (see chapters 4.3 and 4.4).

Regarding the three aberrant HD-derived EpCAM⁺ cells identified by CNA analysis (Table 4-12), it is noteworthy that all of them were among the four cells that previously clustered with the M0 DCCs in the microarray analysis (compare Figure 4-2). These three cells originated from two HDs of 60 and 82 years of age, respectively. In contrast, the fourth cell that clustered with the M0 DCCs had a balanced genome and originated from a HD of 71 years.

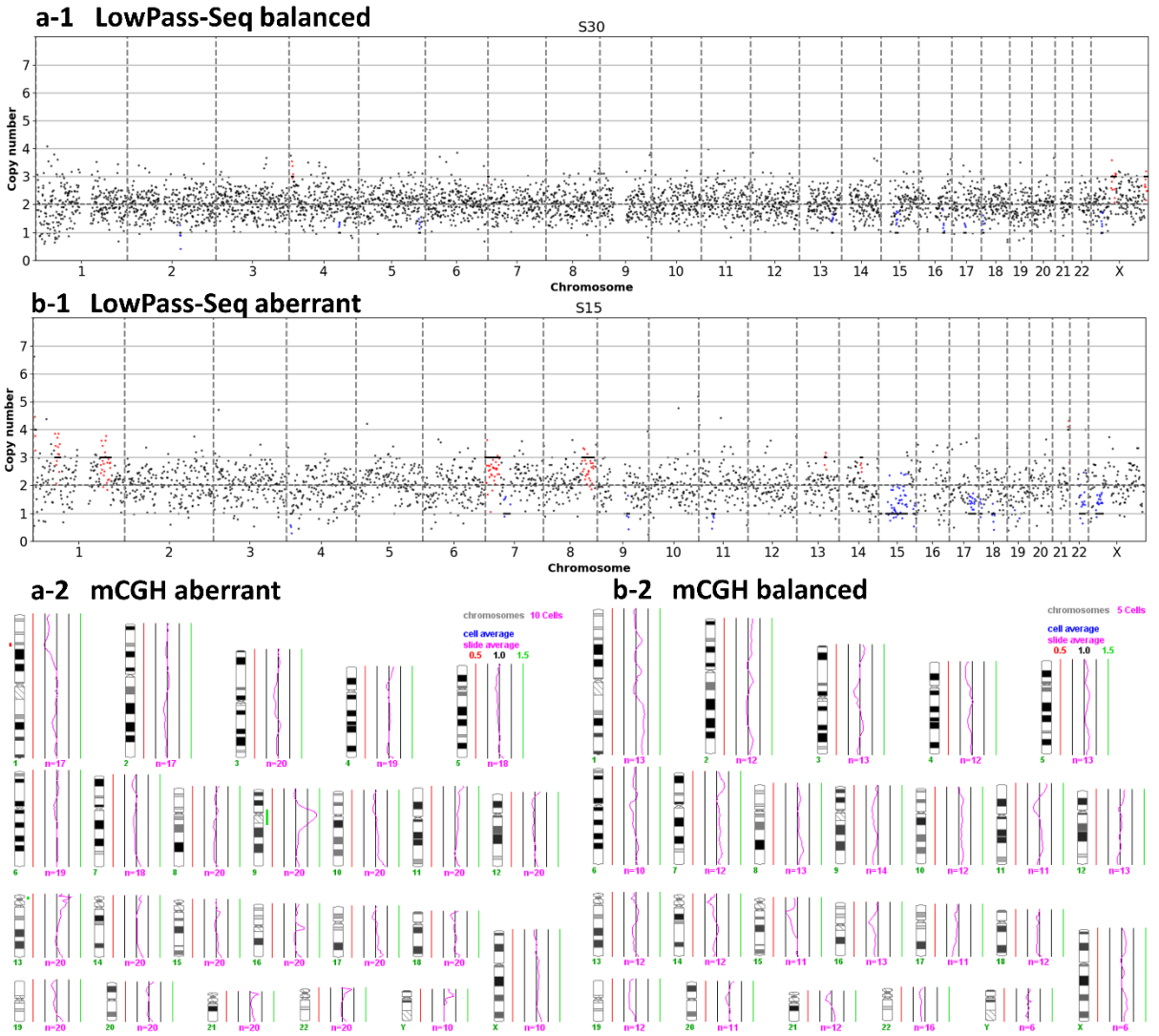


Figure 4-8 Examples of mismatching LP-Seq and mCGH profiles from the same single cells. The figure shows two matched pairs of LP-Seq and mCGH profiles that were in disagreement. (a-1) Balanced M0 LP-Seq profile. (a-2) Aberrant mCGH profile generated from the same single cell genome as the one shown in panel a-1. (b-1) Aberrant M0 LP-Seq profile. (b-2) Balanced mCGH profile generated from the same single cell genome as the one shown in panel b-1.

Table 4-12 Overview of aberrant and balanced cells identified by combination of mCGH and LP-Seq. The table displays the numbers of aberrant and balanced cells when c mCGH and LowPass CNA data were combined. Percentages are relative to the total number of cells in each row. Statistics: Fisher's exact test on cell counts (chapter 3.16.1.2); * LumA vs LumB (p=1) # M0 vs. M1, p=0.002

Group	Total cells	Aberrant	Balanced
M0 (total)	81	41/81 (50.6 %) #	40/81 (49.4 %) #
LumA	*26*	*14/26 (53.8 %)* *	*12/26 (46.2 %)* *
LumB	*32*	*18/32 (56.3 %)* *	*14/32 (43.8 %)* *
Lum undefined	*11*	*4/11 (36.4 %)*	*7/11 (63.6 %)*
TNBC	*11*	*4/11 (36.4 %)*	*7/11 (63.6 %)*
Missing data	*1*	*1/1 (100 %)*	*0/1 (0 %)*
M1 (total)	24	21/24 (87.5 %) #	3/24 (12.5 %) #
LumA	*1*	*1/1 (100 %)*	*0/1 (0 %)*
LumB	*14*	*11/14 (78.6 %)*	*3/14 (21.4 %)*
Lum undefined	*6*	*6/6 (100 %)*	*0/6 (0 %)*
TNBC	*3*	*3/3 (100 %)*	*0/3 (0 %)*
NCC (total)	10	3/10 (30 %)	7/10 (70 %)
Total	**115**	**65/115 (56.5 %)**	**50/115 (43.5 %)**

Lastly, the numbers of patients that the analyzed cells originated from were summarized (Table 4-13). Fisher's exact test did not reveal a significant difference in the frequency of patients with aberrant cells between the LumA and LumB subtypes (p=0.69). Interestingly, unlike the strong discrepancy previously observed in the single cells (Table 4-12), there was no difference between M0 and M1 patients (p=1). However, this may be due to the low number of M1 cases.

Table 4-13 Overview of patients with confirmed DCCs. The table shows the numbers of patients with aberrant or balanced DCCs corresponding to the cells from Table 4-12. Percentages are relative to the total number of patients in each row. Statistics: Fisher's exact test on cell counts (chapter 3.16.1.2); * LumA vs LumB (p=0.69); # M0 vs. M1 (p=1)

Group	Total Patients	Aberrant	Balanced
M0 (total)	44	30/44 (68.2 %) #	14/44 (31.8 %) #
LumA	*14*	*9/14 (64.3 %)* *	*5/14 (35.7 %)* *
LumB	*16*	*12/16 (75 %)* *	*4/16 (25 %)* *
Lum undefined	*6*	*4/6 (66.7 %)*	*2/6 (33.3 %)*
TNBC	*7*	*4/7 (57.1 %)*	*3/7 (42.9 %)*
Missing data	*1*	*1/1 (100 %)*	*0/1 (0 %)*
M1 (total)	9	6/9 (66.7 %) #	3/9 (33.3 %) #
LumA	*1*	*1/1 (100 %)*	*0/1 (0 %)*
LumB	*3*	*3/3 (100 %)*	*0/3 (0 %)*
Lum undefined	*4*	*1/4 (25 %)*	*3/4 (75 %)*
TNBC	*1*	*1/1 (100 %)*	*0/1 (0 %)*
NCC (total)	8	2/8 (25 %)	6/8 (75 %)
Total	**61**	**38/61 (62.3 %)**	**23/61 (37.7 %)**

Following the identification of true DCCs by CNA analysis, the results of the mCGH and LP-Seq were compared to the qPCR signature developed in chapter 4.2.1.1. In agreement with Table 4-11, mCGH and LP-Seq agree in most of the analyzed cases (Figure 4-9). In contrast, the qPCR showed several mismatches. Similar to previous results (Table 4-8), the qPCR signature seemed to perform particularly bad with M1 DCCs as expected, because no M1 patient-derived cells were included in the microarray for identification of the signature genes (see Figure 4-2).

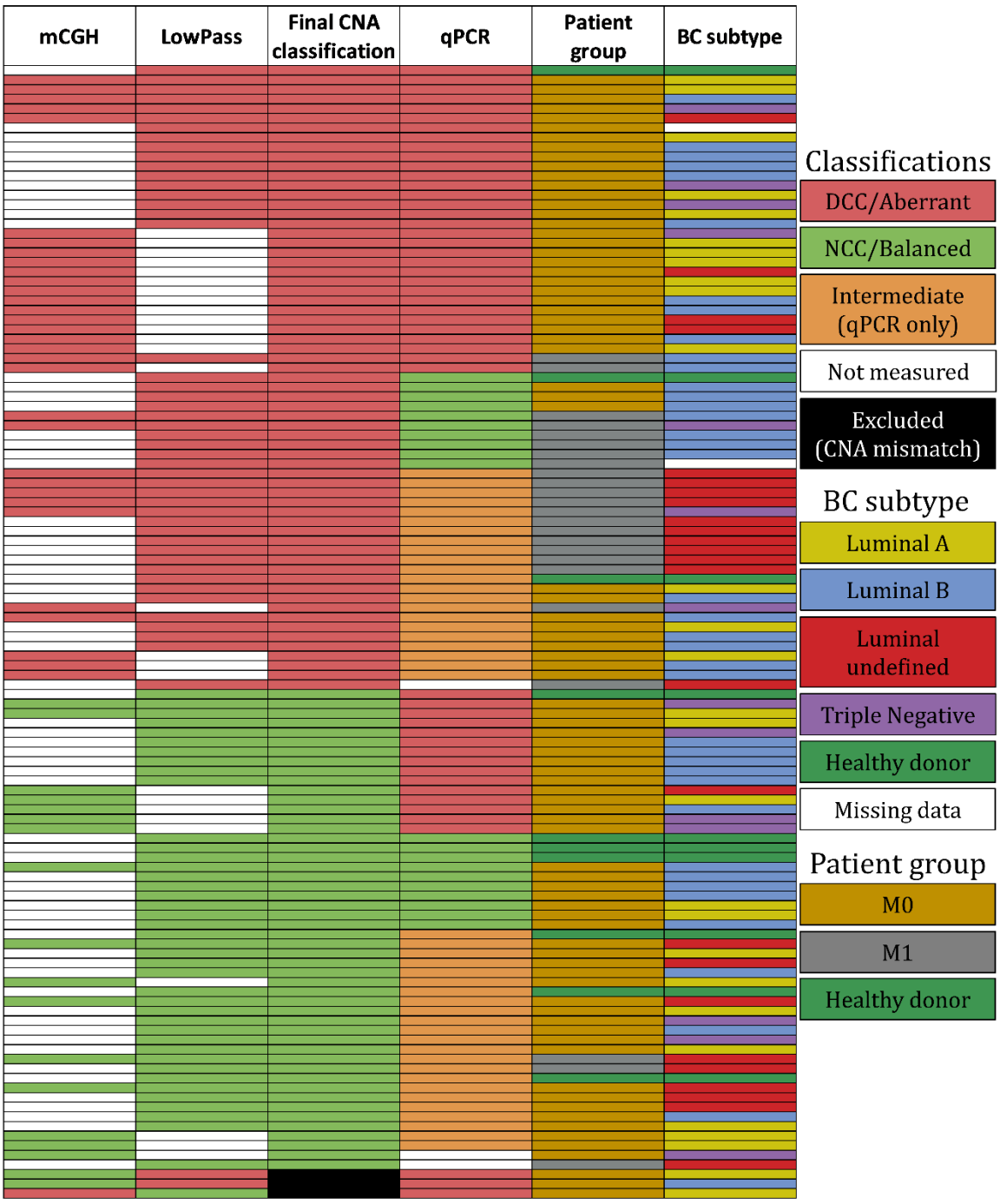


Figure 4-9 Comparison of LP-Seq, mCGH, and qPCR classifications of DCCs. The heat map illustrates the classification results of the three different methods for each cell. The rows represent individual cells, while the first four columns indicate the classification of each cell according to mCGH, LP-Seq, the overall CNA-based classification (taking together mCGH and LP-Seq), and qPCR according to the colors depicted in the legend. The DCC-like, undefined, and NCC-like classes of the qPCR (see Table 4-8) were summarized in the intermediate class (orange) to simplify the heat map. If LP-Seq and mCGH contradicted each other, the respective cell was excluded from further analysis. In addition to the classifications, the two columns on the right display the patient group each cell was derived from as well as the BC subtype of each cell.

Next, I calculated the same performance metrics as before (see training set data in chapter 4.2.1.2, formulas and definitions in chapter 3.16.1.1) on the set of CNA profiled cells (see Table 4-12). M1-derived DCCs were excluded and M0- and HD-derived EpCAM$^+$ classified as aberrant or balanced were combined into one dataset to evaluate the signature's capability to identify true M0 DCCs (Table 4-14, "Stringent" table). Again, the signature was assessed with stringent and relaxed selection criteria (see chapter 4.2.1.2). In the stringent variant, the metrics were the following: prevalence 57.6 %, accuracy 67.8 %, misclassification rate 32.2 %, sensitivity 88.2 %, specificity 40 %, PPV 66.7 %, and NPV 71.4 %. Interestingly, there were far more false positives of the qPCR signature (n=15) than false negatives (n=4, Table 4-14, "Stringent" table), which may indicate a higher sensitivity of the qPCR signature than the CNA profiling.

For the stringent signature (above), 34 intermediate cases were excluded from the calculation. By inclusion of 24 DCC-like and NCC-like classified cells into the relaxed criteria (Table 4-14, "Relaxed" table), this number was reduced to ten undefined cells. The following performance metrics were calculated: prevalence 53 %, accuracy 67.5 %, misclassification rate 32.5 %, sensitivity 90.9 %, specificity 41 %, PPV 63.5 %, and NPV 80 %. In this dataset the discrepancy between false positives (n=23) and false negatives (n=4) of the qPCR was even more pronounced than before.

Table 4-14 Confusion matrices for assessment of qPCR signature performance on test set. The columns of each table represent the result from the previous CNA analysis. In contrast, the rows represent the respective classifications resulting from the qPCR signature. M0- and HD-derived EpCAM$^+$ cells with robust CNA profiles were included. The left table represents the stringent signature, while the right table displays the data of the relaxed signature, which included DCC-like and NCC-like cells. Abr = aberrant, Bal = balanced

Stringent		**CNA profiling**	
		Abr	**Bal**
qPCR signature	**DCC**	30	15
	NCC	4	10

Relaxed		**CNA profiling**	
		Abr	**Bal**
qPCR signature	**DCC/DCC-like**	40	23
	NCC/NCC-like	4	16

Once the final DCCs had been selected according to their genomic aberration status, I continued with a detailed analysis of the CNA profiles of the DCCs (chapter 4.3).

4.3 CNA analysis of DCCs

Following the hypothesis that LumA and LumB subtypes differ in their mutational landscapes, the available CNA data were analyzed in more detail. In order to be able to plot the CNA data in a summarized form, I first created appropriate annotations of the profiles for the *Progenetix* online tool (chapter 4.3.1). Next, the profiles of cells, for which both mCGH and LP-Seq data were available, were compared in detail (chapter 4.3.2), since the CNA profiles had previously only been examined qualitatively (whether the genome was overall aberrant or balanced). Afterwards, individual CNAs in M0 vs. M1 DCCs and $EpCAM^+$ vs. CK^+ cells (chapter 4.3.3) as well as CNAs identified in LumA and LumB DCCs (chapter 4.3.4) were compared.

4.3.1 Annotation of profiles for *Progenetix*

Progenetix is an online repository for published chromosomal aberration data (Baudis and Cleary, 2001), which also includes a tool to upload user data for generation frequency plots that was supposed to be utilized to summarize the newly generated CNA data for further analysis. *Progenetix* accepts data both in genomic coordinates generated by modern next-generation-sequencing (NGS) platforms or Affymetrix arrays and in the International System for Human Cytogenetic Nomenclature (ISCN) style (ISCN rules for listing chromosomal rearrangements, 2001) used in older cytogenetics applications like mCGH. In order to be able to analyze the CNA data from both mCGH and LP-Seq together using *Progenetix*, it was necessary to convert one of the data types into the other. Since the LP-Seq analysis pipeline also provided so-called RefSeq files containing the genomic coordinates of all aberrations, I decided to take these files and add the cytoband information from a reference file, in order to convert the LP-Seq data to the ISCN format. For automation of this annotation process, a script was written in R-Studio (described in chapter 3.16.2).

Due to the LP-Seq analysis pipeline supplied by Menarini Silicon Biosystems sometimes calling false positive genomic aberrations, the *Progenetix*-compatible ISCN annotation was performed manually for all samples. In order to do this, I went through all 49 high quality aberrant LP-Seq profiles (see Figure 4-10 below) while also checking the corresponding RefSeq files, decided which gains or losses to accept as true aberrations, and then translated the start and end cytoband of each aberration to the ISCN format. To more easily connect the aberrations from the CNA profile and the genomic coordinates in the RefSeq file, the cytoband information from the latter were compared with chromosome ideograms found in the ISCN 2009 Atlas of Genetics and Cytogenetics in Oncology and Haematology (link in Table 2-13). Similar to chapter 4.2.2.1, small aberrations in centromeric and telomeric areas, as well as a small recurrent gain in the 1p arm of chromosome 1, were excluded. Additionally, the genomic coordinates from the RefSeq file were also taken into account and all aberrations with a length of less than 1 megabase were excluded. In the course of the annotation process, five of the aberrant profiles were excluded, because the data - despite being of good quality - were too noisy or contained too many potential artifacts to clearly decide on the validity of many of the genomic alterations. The schematic below illustrates the final workflow of LP-Seq profile selection that led to the 44 samples (Figure 4-10), which were finally annotated according to the ISCN guidelines and analyzed with *Progenetix*.

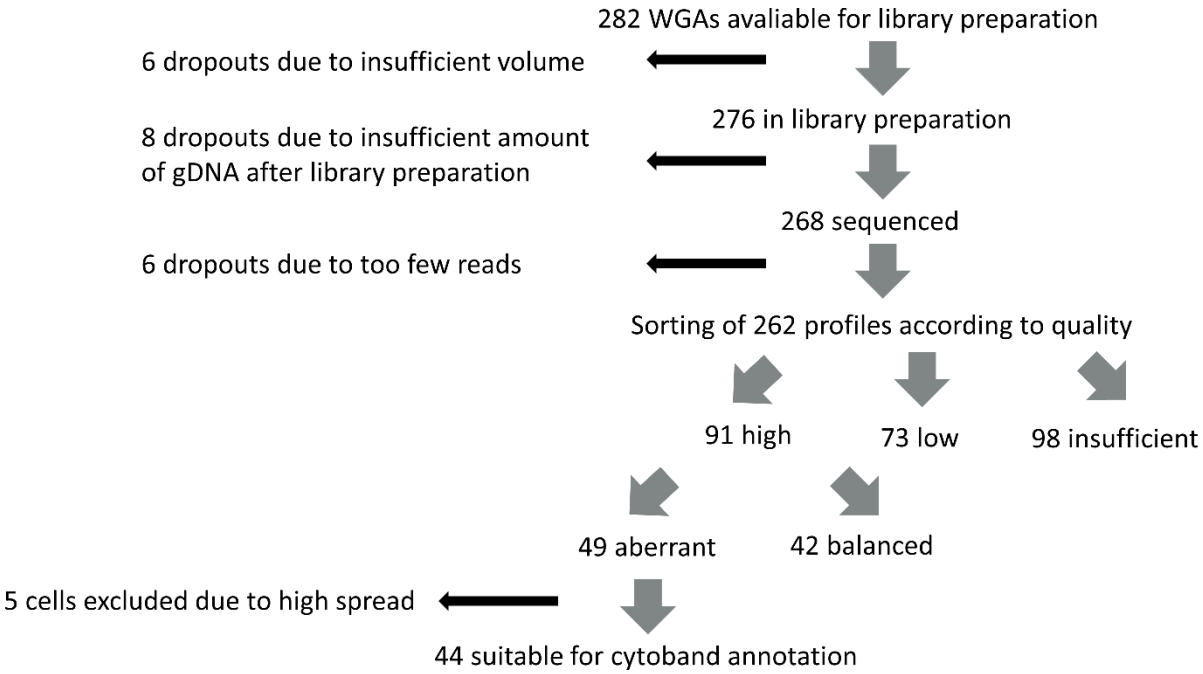


Figure 4-10 Schematic of LP-Seq profile generation process up to cytoband annotation. The schematic is a continuation of Figure 4-4 and also shows the selection of profiles for cytoband annotation and subsequent generation of frequency plots using *Progenetix*.

The final ISCN-style annotations – either the LP-Seq data alone or together with Gundula Haunschild's mCGH annotations (all annotations are provided in appendix chapter 12.1.2) - were then uploaded to the *Progenetix* user data tool (ISCN format option) to generate the desired frequency plots which will be discussed in chapters 4.3.3 and 4.3.4. First, however, those cases, in which LP-Seq and mCGH data were available of the same cells, were investigated to compare the two technologies in more detail (chapter 4.3.2).

4.3.2 Comparison of overlapping LowPass-Seq and mCGH profiles

Using the 14 profiles, which were aberrant according to both mCGH and LP-Seq (see Table 4-11), I wanted to assess how comparable the two technologies were. Unfortunately, three of the samples dropped out, because two of them had been excluded during the manual screening of the aberrations (see chapter 4.3.1) and for the third one the mCGH data were not annotated, as this sample had previously been tested both aberrant and balanced by Gundula Haunschild's mCGH and was therefore excluded by her (see chapter 4.2.2.2). Because of the ambiguous mCGH data, I decided to exclude this sample from this analysis as well. Subsequently, both the mCGH- and LP-Seq-derived ISCN annotations of the remaining eleven samples were uploaded to *Progenetix*.

The resulting frequency plots (Figure 4-11) revealed three things. First, both methods detected mostly the same CNAs, e.g. losses in 1p and gains in 1q, 7p gain, gain of the whole chromosome 12 or loss of the entire chromosome 13, to name only a few. There were a few CNAs, however, like the 1p gain or Xp loss, which were not detected by mCGH. Second, the LP-Seq data often displayed higher frequencies of some aberrations compared to mCGH, e.g. the gains of chromosomes 4 or 12. The same applied to the losses found in 4q or 16q. Third, the LP-Seq data looked more precise than the mCGH. For example, the gains of the 1q and 8q arms illustrated this, there were more nuances in specific areas of the chromosome than in the mCGH data.

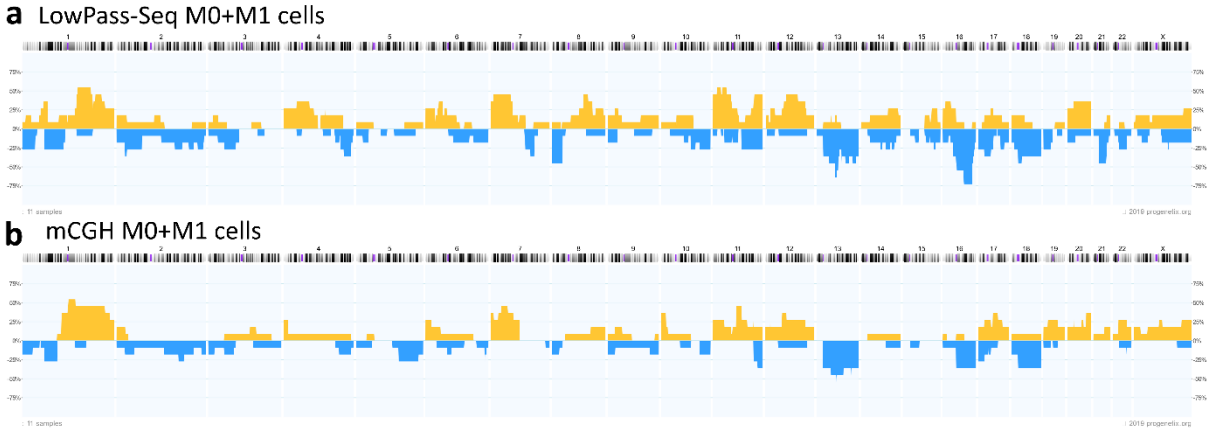


Figure 4-11 Frequency plots comparing LP-Seq and mCGH CNA profiles of the same EpCAM⁺ DCCs. The plots illustrate the frequency (in %) of a given CNA in the tested samples for each chromosome. Genomic gains are depicted in yellow, losses in blue. Each plot consists of n=11 separate samples, four of which originated from M0 patients and seven from M1 patients. (a) CNAs detected by LP-Seq. (b) CNAs detected by Dr. Haunschild's mCGH (Haunschild, 2013).

Next, I compared the overall similarity of the mCGH and LP-Seq-derived data by separate assessment of the whole available LowPass and mCGH data divided into M0 and M1-derived samples resulting in a total of four groups. Note that the previous samples with both LP-Seq and mCGH data available (see Figure 4-11) were included as well. The M0 and M1 datasets obtained either by LP-Seq or mCGH looked similar (Figure 4-12). Apart from the observations described in the previous paragraph, the most striking differences were the absence of the 8q gain in the M0 mCGH dataset (Figure 4-12b) and the lack of 3q aberrations in the M1 LP-Seq data (Figure 4-12c) compared to the M1 mCGH data (Figure 4-12d), which was missing the losses on chromosomes 21 and X.

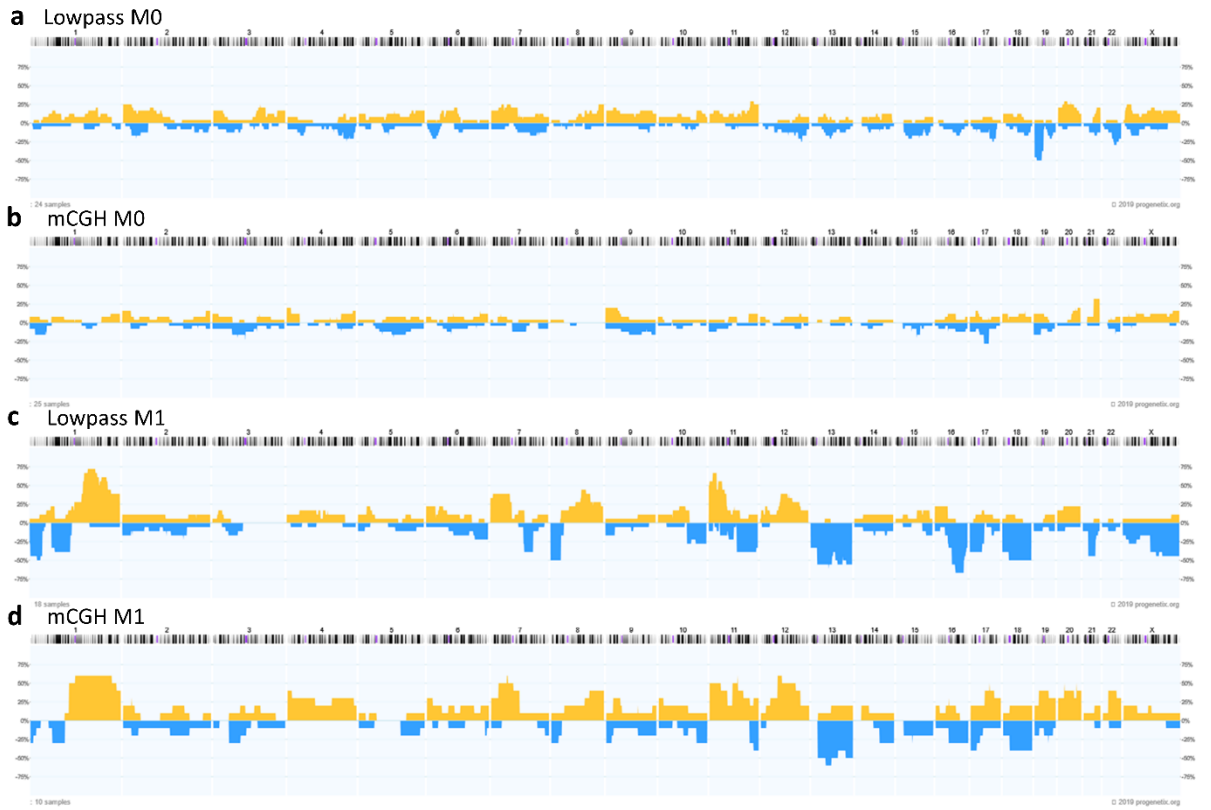


Figure 4-12 Frequency plots of all LP-Seq and mCGH CNA data of EpCAM⁺ DCCs – M0 versus M1. The plots display the frequency (in %) of a given CNA in the tested samples for each chromosome. Genomic gains are depicted in yellow, losses in blue. (a) LP-Seq M0 n=24. (b) mCGH M0 n=25. (c) LP-Seq M1 n=18. (d) mCGH M1 n=10.

Judging from these observations, I decided to combine the mCGH and LP-Seq datasets of the EpCAM+ DCCs for further analysis of the CNA data to increase sample numbers, since the methods were agreeing in the majority of cases.

4.3.3 CNAs in M0 versus M1 and EpCAM+ versus CK+ DCCs

In the course of her experiments, Dr. Haunschild discovered that M0 DCCs carried fewer aberrations than M1 DCCs both within the EpCAM+ DCC collective and within the CK+ DCC collective, while EpCAM+ and CK+ DCCs differed only slightly in three CNAs when comparing M0 or M1 DCCs between the two collectives (Haunschild, 2013). Therefore, I aimed to investigate whether this was still true with addition of the LP-Seq data to the EpCAM+ DCC collective.

Before comparing the CNA profiles of the EpCAM+ and CK+ collectives, mean and median numbers of cells per patient for each sample group were calculated to check whether there were any patients overrepresented by higher numbers of cells (Table 4-15). The CK+ collective was adapted from a previous publication of our group (Hosseini et al., 2016) and contained a few additional cells compared to the cells used by Gundula Haunschild (Haunschild, 2013).

Table 4-15 Patient and cell numbers of EpCAM+/CK+ collectives for CNA profiles stratified by metastatic status.

Group	Patients	Cells	Cells per patient (mean)	Cells per patient (median)
M0 EpCAM+	34	45	1.32	1
M1 EpCAM+	6	23	3.83	2.5
M0 CK+	27	45	1.67	1
M1 CK+	24	77	3.2	2.5

To check whether cell numbers were similar, pairwise comparisons using Student's t-test were performed on the number of cells isolated from each patient. The following pairs were analyzed:

- Pair I: EpCAM+ M0 versus CK+ M0
- Pair II: EpCAM+ M0 versus EpCAM+ M1
- Pair III: EpCAM+ M1 versus CK+ M1
- Pair IV: CK+ M0 versus CK+ M1

The T-tests revealed that the cell numbers of the M0 or M1 subgroups between the collectives (pairs I and III) were comparable, with p=0.15 and p=0.64 for pair I and pair III, respectively. However, M1 groups within each collective consisted of more cells per patient than the respective M0 groups (pairs II and IV). Specifically, in the EpCAM+ M1 group there was one patient with eleven cells, which made up almost half of the cells resulting in a p-value of 0.0006 for the comparison of pair II. In the CK+ M1 group there were two patients with ten and eleven cells, respectively. Despite the higher number of patients, there was still a highly significant difference for pair IV with p=0.008. The overrepresentation of one and two patients, respectively, in the EpCAM+ M1 and CK+ M1 groups needs to be kept in mind when interpreting the comparison of pairs II and IV.

For visualization of the CNAs, the LP-Seq data of the EpCAM+ M0 and M1 DCCs were combined with Gundula Haunschild's mCGH data of the EpCAM+ M0 and M1 DCCs and the frequency plots of the combined EpCAM+ collective were compared with those of the CK+ collective. In case of those

samples, of which both LP-Seq and mCGH data were available (compare Figure 4-11), I decided to use the more precise LP-Seq data for plotting and to exclude the mCGH to avoid including the same sample twice, thereby distorting the analysis. This resulted in a lack of CNAs in the 3q arm, although there were aberrations in this arm in the mCGH data. Moreover, the data of the X chromosomal CNAs of the CK$^+$ DCC cohort could not be used due to some of the samples having been hybridized to male references (sex mismatch).

For the statistical analysis of the CNA frequencies, Fisher's exact test to was performed for pairwise comparisons of the counts of the four sample groups and the p-values were corrected for multiple comparisons with the Benjamini and Hochberg method (chapter 3.16.1.2) to obtain adjusted p-values (q-values). For each pair, I screened the respective frequency plots by eye and selected those aberrations which appeared most different, counted the number of cells carrying the respective CNA in each group using the cluster plots generated by *Progenetix* with an overlaid grid for easier counting (example in Figure 4-13), and calculated the q-values. The grid was added using the software *GIMP*. All the chromosomal regions that were considered for each pair are listed in Table 4-16 below. In several cases, CNAs which looked like they covered a whole chromosomal arm in in the frequency plot (Figure 4-14), in fact comprised adjacent CNAs across the individual cells which were merged into a single long CNA for the frequency plots. In these cases, the partial CNAs on the indicated arms were also counted if they made up more than 50 % of the chromosomal arm, in order to simplify the analysis. When investigating losses of complete chromosomes, aberrations covering at least 50 % of the chromosome were counted. In cases, in which I focused on a specific locus, I labeled the respective CNA as centromeric (cen), terminal (ter), or internal (int) depending on the location of the CNA on the chromosomal arm, in order to enable identification on the frequency plot. The two separate losses in the 1p arm, which frequently occurred in EpCAM$^+$ M1 DCCs (compare Figure 4-14c), mostly occurred together so they were treated like a single aberration and all samples with an aberration in either of those loci were counted for the statistics. Due to the discrepancy between LowPass and mCGH data in the 3q chromosomal arm, this arm was excluded from the analysis. CNAs with $q<0.1$ were considered weakly significant, due to the relatively high noise inherent to single cell genomics.

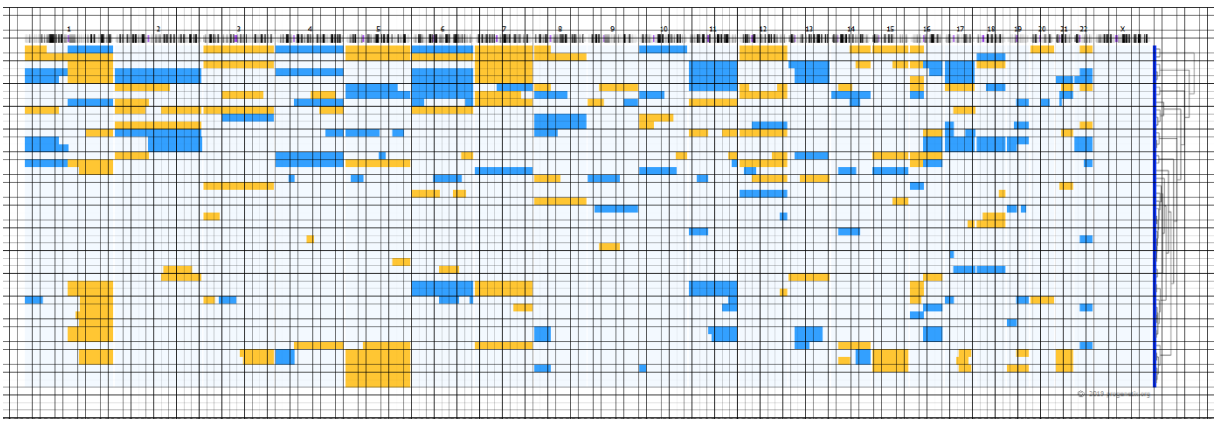


Figure 4-13 Example cluster plot of the CK$^+$ M0 DCC group. The plot shows all annotated aberrations across all chromosomes for each CK$^+$ M0 DCC. Each row represents one individual DCC, while the columns represent different chromosomal areas. Yellow bars indicate genomic gains, while blue bars represent genomic losses.

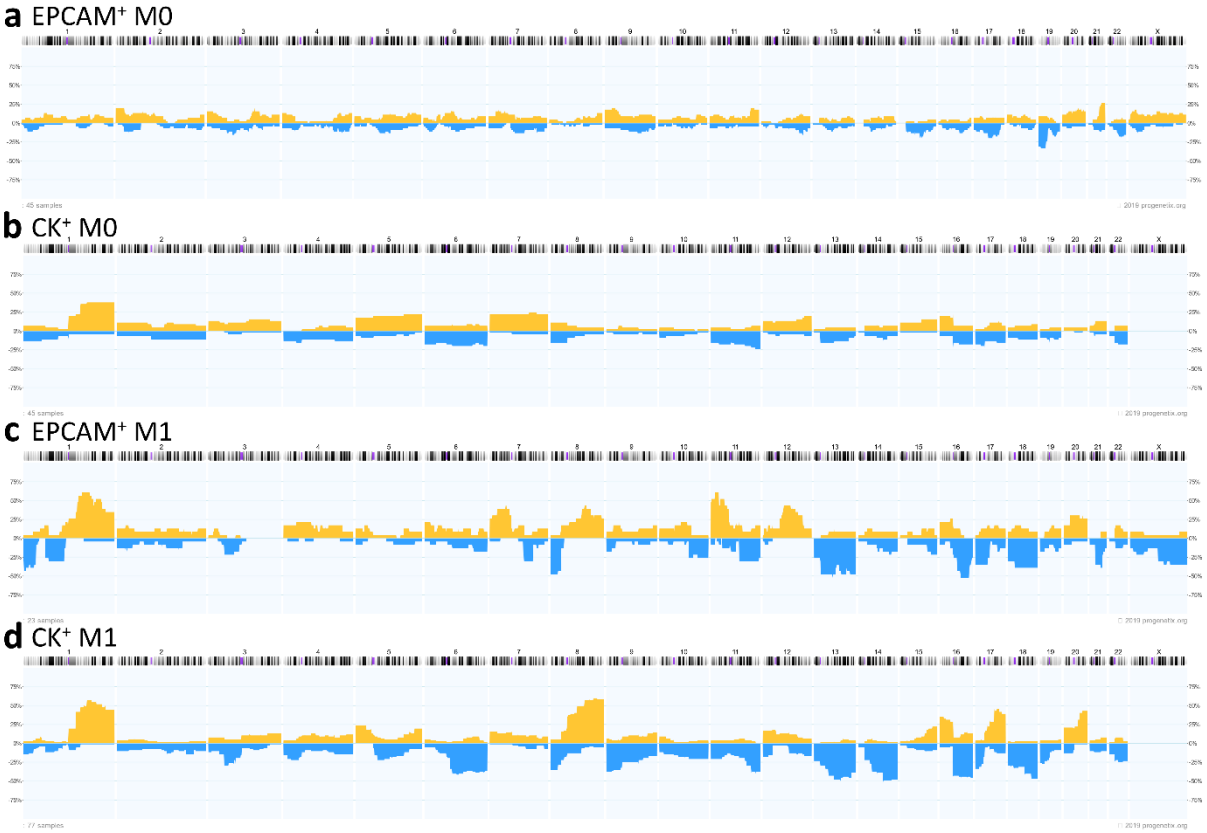


Figure 4-14 Frequency plots of EpCAM⁺ and CK⁺ M0 and M1 DCCs. The plots show the frequency (in %) of a given CNA in the tested samples for each chromosome. Genomic gains are depicted in yellow, losses in blue. In case of the EpCAM⁺ DCCs, the plots represent the combined data of the LP-Seq and the mCGH. The data of the X chromosome of the CK⁺ DCC cohort could not be used due to some of the samples having been hybridized to male references (sex mismatch), therefore the X chromosome is free of aberrations in the corresponding plots. (a) EpCAM⁺ M0 DCCs n=45. (b) CK⁺ M0 DCCs n=45. (c) EpCAM⁺ M1 DCCs n=23. (d) CK⁺ M1 DCCs n=77. ISCN annotations for CK⁺ cells were taken and adapted from the supplementary data of (Hosseini et al., 2016).

Table 4-16 CNAs selected for statistical analysis for each pairwise comparison. If no arm is given with the chromosome number, this means the whole chromosome was considered aberrant. The "ter" (=terminal) suffix indicates a CNA located close to the telomeres of a chromosomal arm, while "cen" represents CNAs located next to the centromere. CNAs with the "int" suffix are internal CNAs located in the middle of a chromosomal arm without reaching to the centromere or telomer.

Comparison	Gains	Losses
EpCAM⁺ M0 vs. CK⁺ M0 (Pair I)	1q, 5, 7, 8p, 9p, 12, 15, 16p, 20, 21qter	1p, 6, 8p, 11, 13, 15, 16q, 19p
EpCAM⁺ M0 vs. EpCAM⁺ M1 (Pair II)	1q, 7p, 8q, 10p, 11p, 12q, 16p, 20, 21qter	1p, 7q, 8p, 10q, 11q, 13, 14q, 16q, 17p, 18, 19p, 21qcen, Xq
EpCAM⁺ M1 vs. CK⁺ M1 (Pair III)	2, 4, 7p, 8q, 10p, 11p, 12q int, 15q, 17q, 18, 19q	1p, 4, 5q, 6q, 8p, 9, 10p, 14, 21qcen, 22
CK⁺ M0 vs. CK⁺ M1 (Pair IV)	1q, 5qcen, 7, 8q, 12qter, 16qcen, 16p, 17q, 20	3pcen, 5q int, 6q, 8p, 9, 10, 13, 14q, 16q, 17p, 18, 19p

The frequencies of CNAs in the four different pairs (Figure 4-14) and the significantly different CNAs for each pairwise comparison including the corresponding significance levels (Table 4-17) are provided below. The exact q-values for all assessed CNAs are listed in the appendix (chapter 12.1.3). The data revealed that there were fewer differences between EpCAM$^+$ DCCs and CK$^+$ DCCs (i.e. pairs I and III) than between M0 and M1 DCCs within each DCC collective (i.e. pairs II and IV).

First, pair I was compared, in which case only three significant differences were observed in the CNAs, only one of which (1q gain) had a highly significant q-value <0.01 (Table 4-17). The comparison of pair III provided similar results. Apart from five rather weakly significant differences (q<0.1), there was only one difference, a gain in 11p, which was highly significant with q<0.001. In a nutshell, the EpCAM$^+$ DCCs and CK$^+$ DCCs differed in three and six CNAs, respectively, while only one CNA per collective was highly significant.

In contrast, when M0 DCCs were compared to M1 DCCs, 14 (CK$^+$ collective, pair IV) and 16 (EpCAM$^+$ collective, pair II) significant differences, respectively, were found, many of which were highly significant with q<0.01 or even q<0.001. Regarding the CK$^+$ collective, the CK$^+$ M1 DCCs carried more aberrations in several chromosomes than the M0 cells (pair IV). Surprisingly, the 5q int gain was more frequent in M0 than M1 DCCs. Lastly, pair II displayed the highest number of significant differences in CNAs. For this pair, all CNAs were more frequent in M1 DCCs.

Table 4-17 Significantly different CNAs from pairwise comparisons of EpCAM$^+$ and CK$^+$ DCC collectives. The table shows all CNAs from Table 4-16 which were significantly different in the indicated pairwise comparisons. Non-significant results are not shown. Frequencies of selected aberrations were counted and compared using Fisher's exact test with correction for multiple comparisons employing the Benjamini and Hochberg method (Benjamini and Hochberg, 1995).

q-value	EpCAM$^+$ M0 vs. CK$^+$ M0	EpCAM$^+$ M0 vs. EpCAM$^+$ M1	EpCAM$^+$ M1 vs. CK$^+$ M1	CK$^+$ M0 vs. CK$^+$ M1
< 0.001	-	1q gain, 8p loss, 13 loss	11p gain	8q gain, 14q loss
< 0.01	1q gain	11p gain, 12q gain, 16q loss, 18 loss, Xq loss	-	3pcen loss, 13 loss, 17q gain
< 0.05	-	1p loss, 7p gain, 8q gain, 10q loss, 11q loss, 16p gain, 17p loss, 21qcen loss	-	5q int loss, 5qcen gain, 6q loss, 8p loss, 9 loss, 12qter gain, 16q loss, 19p loss, 20 gain
< 0.1	5 gain, 19p loss	-	6q loss, 7q int loss, 14 loss, 21qcen loss	-

After the comparison of the M0 and M1 DCCs, I proceeded to tackle the initial aim of the dissertation: to find out whether there were differences between the LumA and LumB subtypes (chapter 4.3.4).

4.3.4 CNAs in LumA versus LumB DCCs

To find potential differences in CNAs between LumA and LumB DCCs from M0 patients and to uncover hints as to why the LumB subtype is more malignant than the LumA subtype, M0 DCCs from both subtypes were visualized and statistically analyzed in the same way described before (compare 4.3.3). M1 patients were not included, because there was only one LumA subtype patient in the M1 group (compare Table 4-3). For completeness, CNAs of TNBC-derived DCCs and NCCs were also visualized (Figure 4-15c+d), but due to very low sample numbers, they were not analyzed any further.

As depicted in Figure 4-15a+b, both LumA and LumB DCCs carried very homogenously distributed aberrations in their genomes, but at first glance there were several loci which looked different between the two, e.g. the chromosomal arm 2p or the whole X chromosome. However, the statistical analysis revealed that none of the putative discrepancies was significant. All loci included in the analysis are marked "ns" (not significant) in the figure. The corresponding p- and q-values are provided below (Table 4-18).

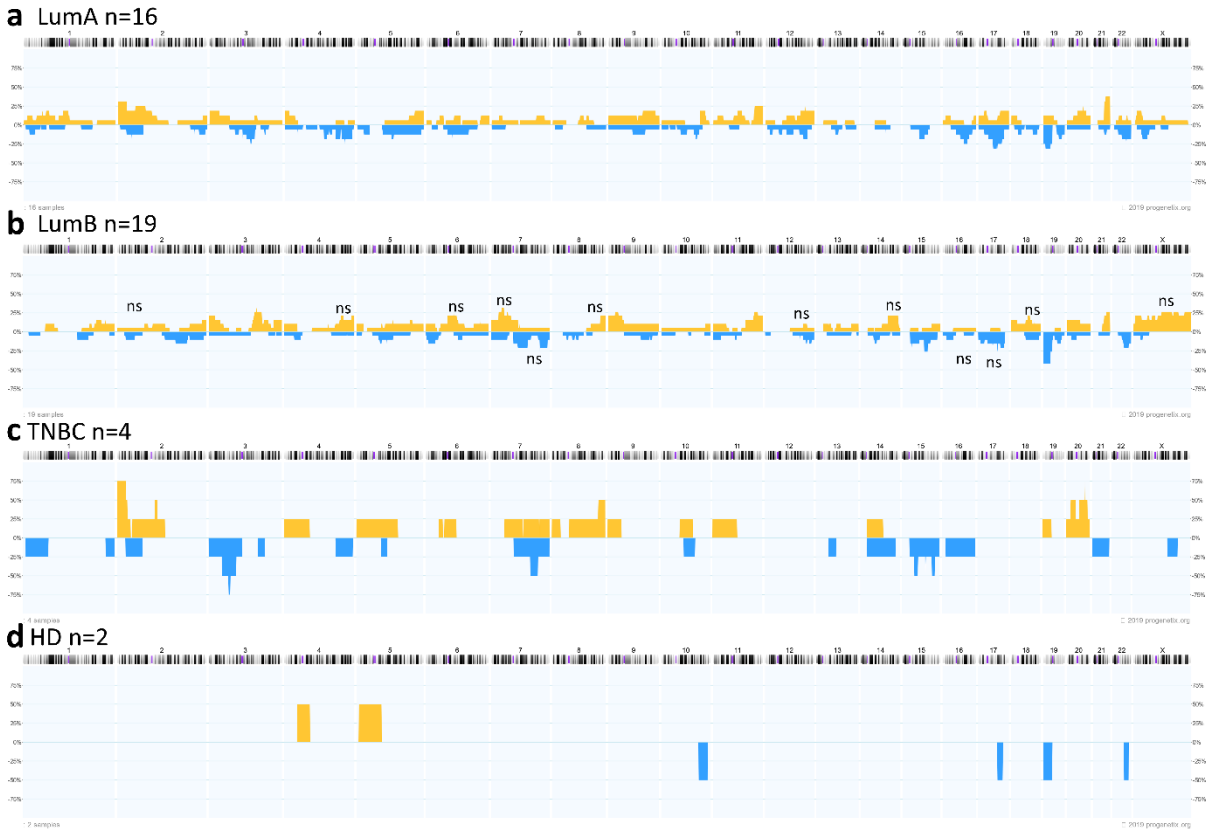


Figure 4-15 CNA profiles of BC subtype-stratified M0 DCCs and NCCs. The plots depict the frequency (in %) of a given CNA in the tested samples for each chromosome. The profiles were generated from a combination of mCGH and LP-Seq data. Genomic gains are depicted in yellow, losses in blue. (a) M0 LumA n=16, (b) M0 LumB n=19, (c) TNBC n=4, (d) HD-derived NCCs n=2. The "ns" (not significant) symbols indicate loci, which were compared between LumA and LumB (panels a and b) using Fisher's exact test, but had a q-value >0.1.

Table 4-18 P- and q-values of Fisher's exact test comparing LumA and LumB DCCs.

Aberration	p-value	Rank	q-value
2p gain	0.155795	5	0.373908
4qter gain	0.58504	10	0.702048
6qcen gain	1	12	1
7p gain	0.34683	8	0.520245
7q losses	0.009194	1	0.110328
8qter gain	0.34683	8	0.520245
12qter gain	0.311994	7	0.53484686
14qter gain	0.10879	4	0.32637
16q internal loss	0.155795	5	0.373908
17q loss	0.7003	11	0.76396364
18q internal gain	0.108785	2	0.65271
X gain	0.108785	2	0.65271

Overall, the CNA profiling did not reveal any differences between the LumA and LumB subtypes. Therefore, I went on with transcriptomic analysis by looking at proliferation marker expression in the EpCAM$^+$ DCCs to examine the proliferation status of the cancer cells in the BM (chapter 4.4), which may provide clues as to why the LumB subtype is more aggressive than the LumA type.

4.4 Proliferation status of DCCs

The current opinion in the scientific community is still that most DCCs located in the BM enter cellular dormancy (Chambers et al., 2002; Aguirre-Ghiso, 2007; Yadav et al., 2018). It was also shown that ER$^+$ BC displays higher dormancy scores than ER$^-$ BC and that a high dormancy score in ER$^+$ BC was correlated with a better recurrence-free survival compared to ER$^+$ BC with low dormancy scores (Kim et al., 2012). Moreover, a study found that metastases of luminal BC frequently contain non-proliferative HR$^-$ cells resistant to endocrine and chemotherapy (Ogba et al., 2014). Therefore, the goal was to investigate whether DCCs from the EpCAM$^+$ DCC collective were also dormant and specifically whether there were differences between LumA and LumB. To examine this, the expression of two proliferation markers was quantified: *marker of proliferation Ki-67* (*MKI67*) and *minichromosome maintenance complex component 2* (*MCM2*). To this end, cutoff Cp values were determined for both genes (chapter 4.4.1) before they were measured in the EpCAM$^+$ cells (chapter 4.4.2). Specifically, I looked at the expression levels in PTs and matched DCCs of M0 LumA and LumB patients, the frequency of proliferating M0 DCCs, and the correlation of the KI67 status in the M0-derived PTs compared with the *MKI67* expression in matched DCCs (chapter 4.4.2.1). The proliferation state of aberrant NCCs was also briefly checked (Figure 4-19). Furthermore, the frequency of proliferating M1 DCCs was examined (chapter 4.4.2.2).

4.4.1 Determination of a cutoff Cp value for proliferation markers

During their PhD projects, two other PhD students of our group isolated naïve as well as stimulated human CD8$^+$ T-cells during different cell cycle stages (Patwary, in preparation; Grujovic, 2019). I used these naïve cells as ideal representations of non-proliferating cells to determine cutoff Cp values (baseline expression in non-proliferating cells) for *MKI67* and *MCM2*, before quantifying their expression in the EpCAM$^+$ DCCs to enable a robust classification into proliferating and non-proliferating cells. The primers for both genes were designed and validated by Nina Patwary (Patwary, in preparation).

Two independent replicate measurements per gene were carried out by qPCR (measurements performed as described in chapter 3.7) on the same 16 naïve CD8$^+$ T-cell samples using technical triplicates for each sample and experiment (total of six technical replicates per sample across two qPCR experiments). The two replicate experiments for *MKI67* were normalized using the included calibrator. Then, the trimmed mean across all technical replicates was calculated for each of the two experiments separately. The second *MCM2* experiment could not be normalized due to a technical problem with the calibrator in the first experiment, but the data of both experiments were still used in the calculation of the cutoff (resulting in slightly lower Cp values for this experiment, see Figure 4-16 Run 2). Negative samples were assigned a Cp value of 33 to enable their inclusion in the analysis. This value was chosen during the establishment of the primers by Nina Patwary and was adopted for the proliferation marker measurements in this study, in order to keep the proliferation analyses consistent across different studies of our research group. The qPCR revealed that both proliferation markers were expressed weakly (Cp < 30) in a few naïve G0 CD8$^+$ T-cells, while the majority of cells displayed Cp values > 30 (Figure 4-16). Overall, both genes

displayed very high Cp values, indicating that the cells were not proliferating. In contrast, the CD8⁺ T-cells isolated during G1-, G2-, or S-phase were all expressing *MCM2* and most were also expressing *MKI67* (Figure 4-17). This result provided the necessary confidence that the G0 T-cells were suitable for establishment of a cutoff value. Therefore, the average of both trimmed means from the separate experiments for each gene was taken as the cutoff value. This way cutoff values of 31.8 and 31.6 were obtained for *MKI67* and *MCM2*, respectively. We decided not to perform additional measurements of *MCM2*, as in retrospect it turned out that the choice of the cutoff value was of minor importance, since almost all DCCs expressed *MKI67* and *MCM2* either at Cp values well below 30 or were completely negative (see Figure 4-19). Additionally, the measurement of *MKI67* and *MCM2* in the naïve CD8⁺ T-cells confirmed the high specificity of the qPCR primers for their respective target.

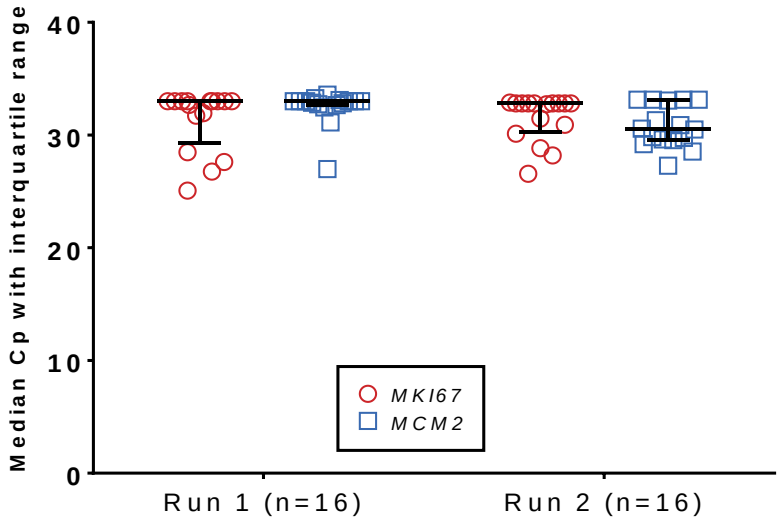


Figure 4-16 Expression of *MKI67* and *MCM2* in naïve CD8⁺ T-cells – cutoff determination. The plot depicts the expression of *MKI67* and *MCM2* in both qPCR runs as median Cp values with interquartile range (whiskers) in n=16 naïve (G0 stage) CD8⁺ T-cells. The data of the second *MCM2* experiment could not be normalized to the first run (= replicate experiment), consequently the resulting Cp values in the second run are slightly lower than in in the first one.

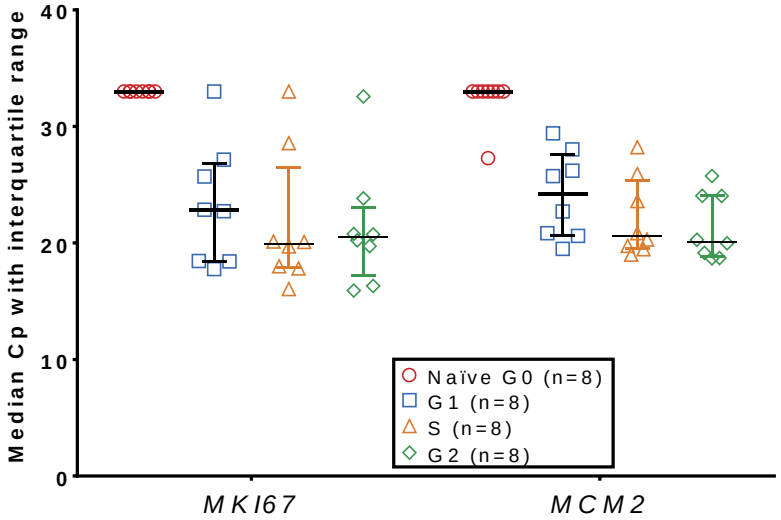


Figure 4-17 Expression of *MKI67* and *MCM2* in T-cells from different cell cycle stages. The plot illustrates the expression of *MKI67* and *MCM2* in naïve (G0 stage), G1-stage, S-stage, and G2-stage CD8⁺ T-cells (n=8 per group) as median Cp values with interquartile range (represented by the whiskers). The displayed data represent a separate experiment from the one shown in Figure 4-16, in which only eight biological replicates were measured per group.

After identification of the cutoff Cp values for *MKI67* and *MCM2*, the proliferation state of EpCAM⁺ DCCs was studied in detail (chapter 4.4.2).

4.4.2 Proliferation in EpCAM+ DCCs and NCCs

4.4.2.1 Proliferation in LumA versus LumB DCCs

Following the determination of the cutoff Cp values for the *MKI67* and *MCM2* proliferation markers, the expression of both genes was quantified in the aberrant EpCAM+ DCCs by qPCR and the data compared with the immunohistochemical KI67 status in the matched PTs. For this purpose, I initially looked at the PT and DCC data separately and then calculated the correlation between them to investigate whether there was a difference between the M0 LumA and M0 LumB subtypes as well as between M0 and M1 DCCs regardless of subtype.

By definition, LumA and LumB differ by their KI67 status in the PT with the exception of ER+/HER2+ BC, which is classified as LumB, even if the KI67 level is <14 % (see Table 2-2, Cheang et al., 2009; Goldhirsch et al., 2011). The percentages of KI67-expressing cells in the PTs of LumA, LumB, and TNBC patients are illustrated below (Figure 4-18). Available M0 TNBC patients were included as a reference, however, due to their low number they were not analyzed further. For comparison of the three subtypes, a one-way ANOVA was applied (chapter 3.16.1.2) which revealed significant differences between the means ($p=0.003$). Tukey's post-hoc test revealed that M0 LumA and M0 LumB PTs differed significantly in their KI67 status ($p<0.05$) as expected. Additionally, the data showed that, while LumA and TNBC diverged significantly ($p<0.01$), LumB was not different from TNBC ($p>0.05$). The similarity of LumB with the highly aggressive TNBC subtype (see Figure 1-3) DCCs hints at the more aggressive nature of the LumB subtype compared to LumA.

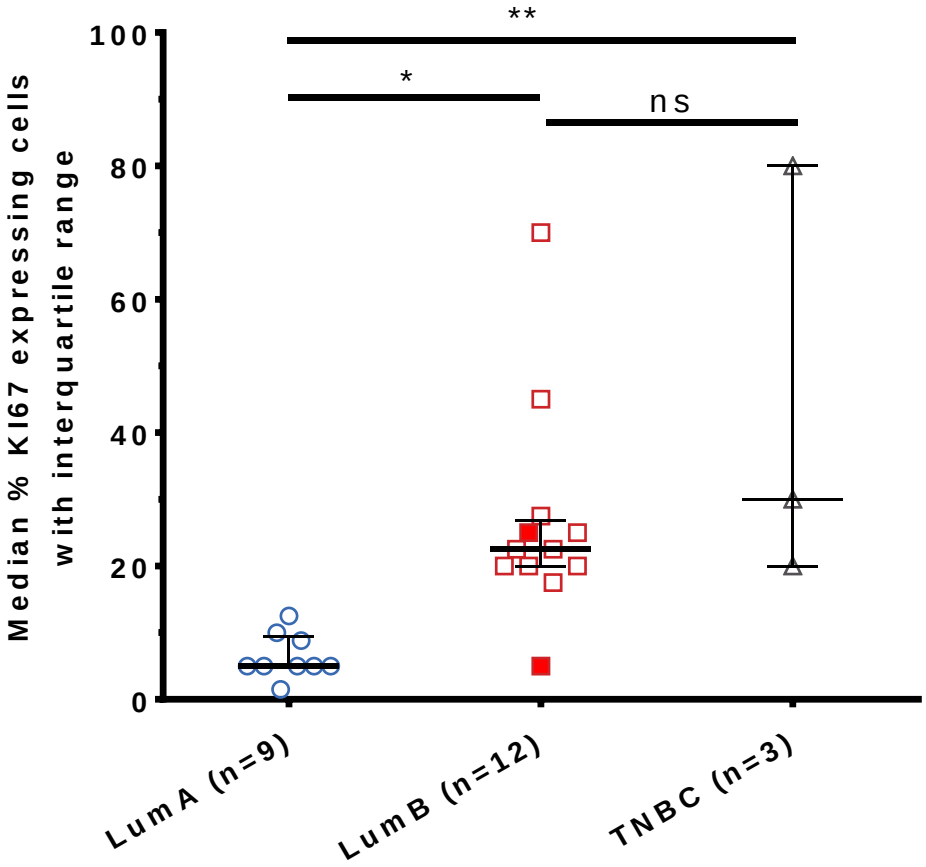


Figure 4-18 KI67 status in the PT of M0 EpCAM+ patients stratified by subtype. The graph depicts the median percentage of KI67-expressing cells found in the PTs derived from LumA, LumB, and TNBC patients, which were positive for true DCCs. The whiskers represent the interquartile ranges. The red filling of some data points of the LumB dataset indicates patients with HER2 amplification (n=2). Statistics: one-way ANOVA (chapter 3.16.1.2); ** $p<0.01$, * $p<0.05$, ns = not significant

In contrast to the PTs, the matched DCCs of neither subtype differed significantly from any of the others (Figure 4-19, two-way ANOVA with multiple comparisons, subtype effect p=0.72). Interestingly, there was not even a difference between DCCs and NCCs. Note that all NCCs included in the statistical analysis were balanced according to the CNA analysis (n=7). The three aberrant NCCs that were previously identified (see Table 4-12) were excluded, because the control cells should not have genomic aberrations. However, for visualization, the three genomically aberrant NCCs were included in Figure 4-19. For completeness, a repeated analysis including the aberrant NCCs was performed and provided similar results (two-way ANOVA with multiple comparisons, subtype effect p=0.39), but is not included in the figure. Of note, all three aberrant NCCs were non-proliferating (Figure 4-19).

Intriguingly, it looked like the cells from each subtype were diverging into two separate subgroups, those cells which were proliferating (Cp below cutoff value) and those which were not (Cp=33).

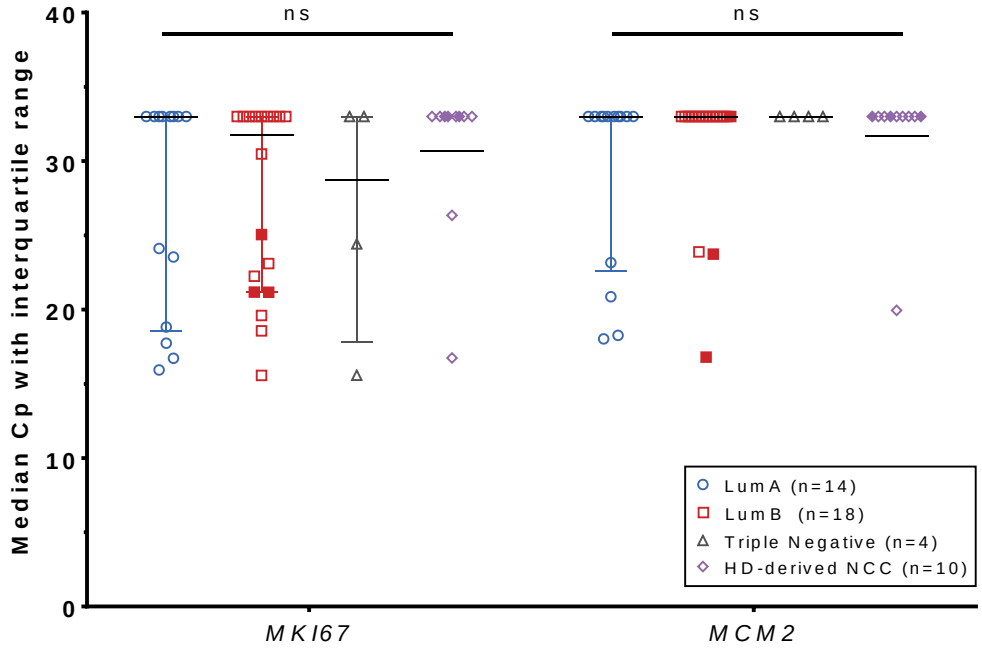


Figure 4-19 Expression of *MKI67* and *MCM2* in M0 EpCAM⁺ DCCs stratified by subtype and NCCs. The graph depicts the expression of *MKI67* and *MCM2* in DCCs across different BC subtypes and in NCCs derived from HDs (including aberrant NCCs) as median Cp values with interquartile range (represented by the whiskers). The red filling of some data points of the LumB dataset indicates cells with HER2 amplification (n=3). NCCs with non-filled symbols were balanced (n=7) according to CNA analysis, whereas the ones with purple filling were aberrant (n=3). Statistics: two-way ANOVA (chapter 3.16.1.2); ns = not significant

Following the individual examination of PT and DCC data, I wanted to investigate the correlation of proliferation in matched PTs and DCCs. Therefore, the percentage of KI67-expressing cells in the PT of each patient was plotted against the *MKI67* expression in the DCCs derived from the matched patients and the Spearman correlation was calculated.

The data revealed that there was no correlation between the two datasets, neither in any of the three subtypes (Figure 4-20b-d) nor overall (Figure 4-20a). The slightly reduced number of cells contained in Figure 4-20 compared to Figure 4-19 was caused by a lack of KI67 data of the PT of some patients. Furthermore, Figure 4-20a contains one DCC of the Luminal undefined type, which is why there is a discrepancy between the sum of data pairs in panels b-d and the number of pairs in panel a.

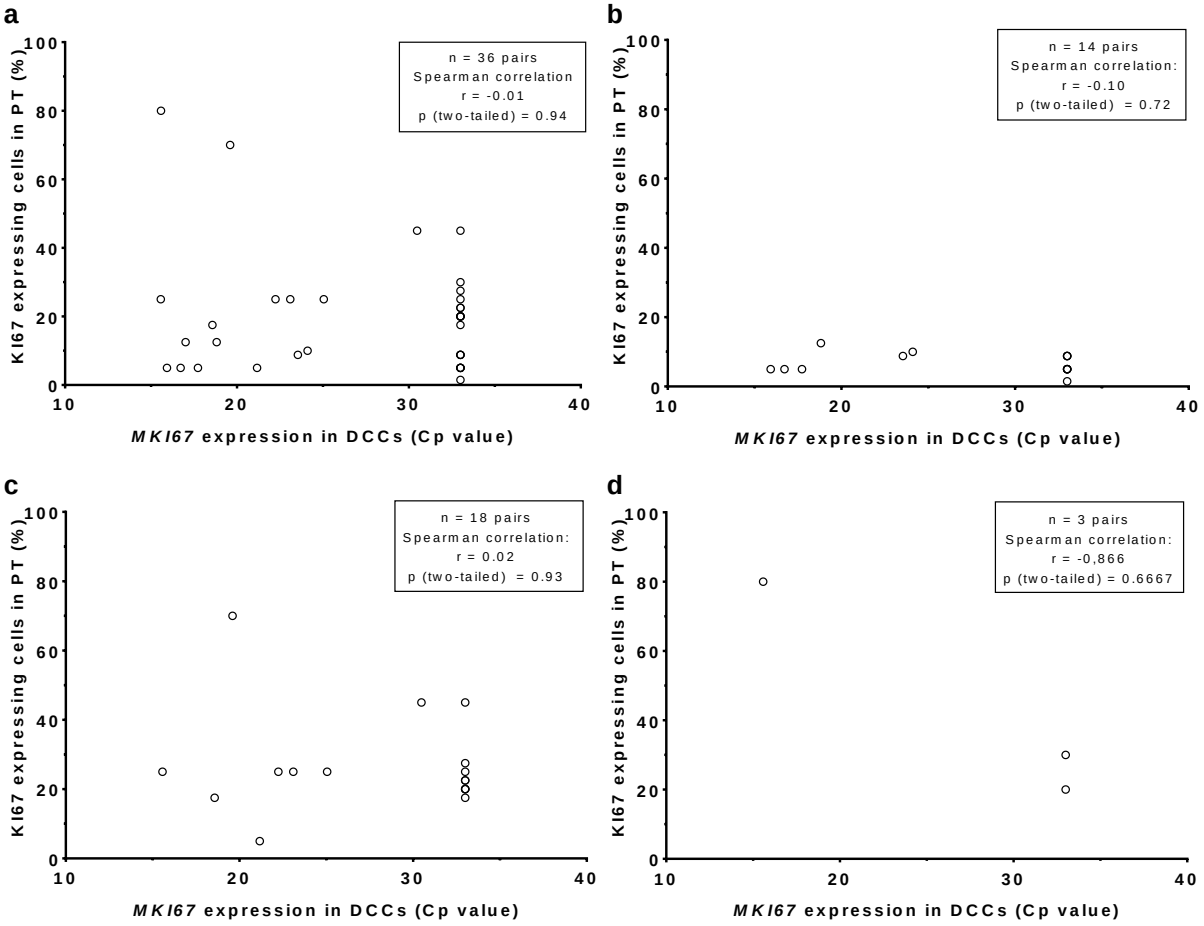


Figure 4-20 Correlation of KI67 level in M0 PT with *MKI67* expression in matched M0 DCCs stratified by subtype. The scatter plots show the percentages of KI67-expressing cells in PTs plotted against the *MKI67* expression (qPCR data). Additionally, each panel also contains the number of pairs used for the analysis, the calculated Spearman correlation coefficient r, and the corresponding two-tailed p-value of the correlation. (a) DCCs of all subtypes together, (b) LumA DCCs, (c) LumB DCCs, (d) TNBC DCCs. Statistics: Spearman correlation (chapter 3.16.1.2)

Next, I used the previously defined cutoff Cp values for *MKI67* and *MCM2* (see chapter 4.4.1) to classify the DCCs either as proliferating or non-proliferating. A cell was considered proliferating if at least one of the two proliferation markers displayed a Cp value below the respective cutoff value. Vice versa, cells with both markers above their corresponding thresholds were considered non-proliferating. Then, the LumA, LumB, and TNBC subtypes were compared according to their frequency of proliferating DCCs. Interestingly, there were only very few cells that had Cp values close to the cutoff value. Most of them were either completely negative with a value of 33 or clearly positive with a Cp <30 (see Figure 4-19).

The results revealed that LumA, LumB, and TNBC proliferated at identical rates of 50 % (Figure 4-21; Fisher's exact test, p=1). For the statistics, the raw numbers of cells were used instead of the percentages shown in Figure 4-21. The whole M0 DCC collective taken together displayed a proliferation rate of 51% ("All DCCs" group, Figure 4-21). Note that this group also contained four cells of the Luminal undefined subtype and one cell without subtype information leading to a higher number of cells than the sum of the three displayed subtypes.

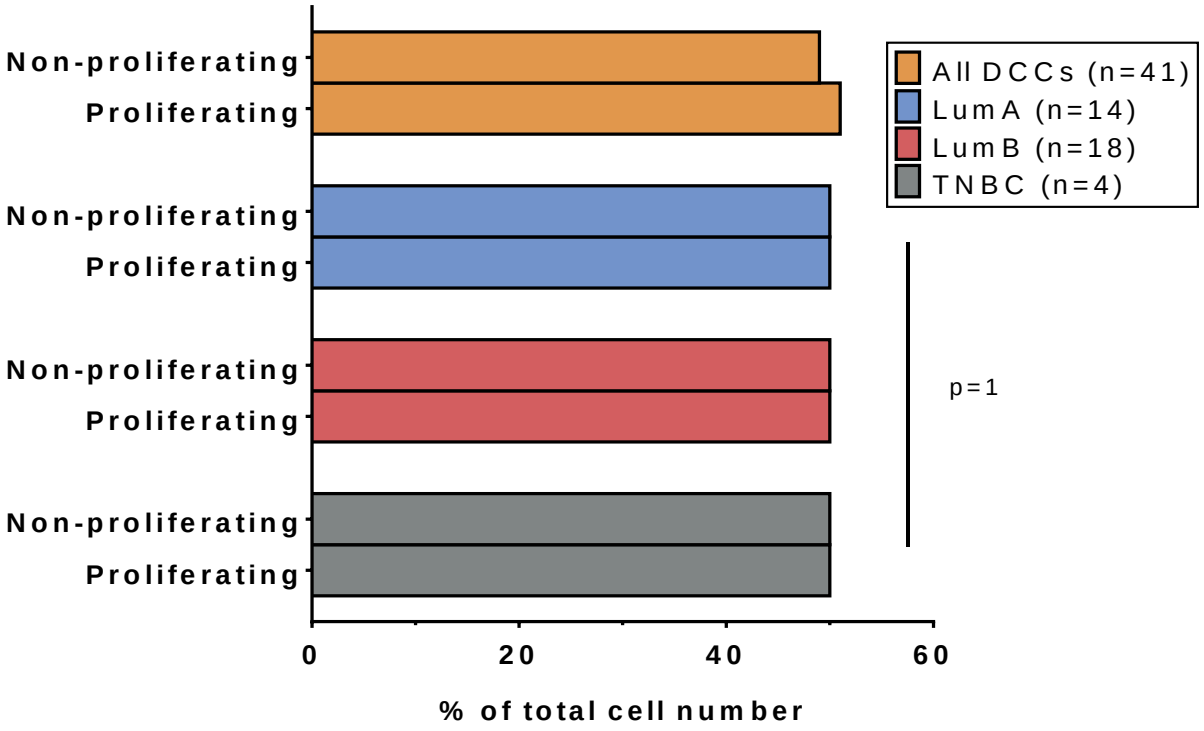


Figure 4-21 Frequency of proliferating and non-proliferating cells among M0 DCCs. The bars represent the percentages of proliferating or non-proliferating DCCs in each subtype or across DCCs of all subtypes. The raw counts instead of the displayed percentages were used for the statistics. Statistics: Fisher's exact test with Freeman-Halton extension for a 2x3 contingency table (chapter 3.16.1.2)

Intrigued by the previous result that DCCs were seemingly diverging into proliferating and non-proliferating subgroups, I decided to repeat the previous analyses with only the proliferating DCCs to see whether there would be a difference between the subtypes.

The expression of both *MKI67* and *MCM2* was the same across all BC subtypes, even when only the proliferating DCCs were considered (Figure 4-22). The two-way ANOVA analysis revealed no significant differences, neither an overall effect of the subtype (p=0.73) nor in pairwise comparisons of the subtypes. Similarly, the results of the correlation analysis of the PT proliferation status with the *MKI67* expression in the DCCs remained insignificant, although the p-value of the LumA subtype decreased by a large margin from p=0.72 to p=0.11 with the Spearman correlation coefficient increasing from r=-0.1 to r=0.67 (compare Figure 4-20b and Figure 4-23b).

Due to the observation that so many M0 DCCs were proliferating according to the qPCR data, I was curious whether same phenomenon would also apply to M1 DCCs. Therefore, the M0 DCCs were compared to the M1 DCC collective to see whether the metastatic state of the patient had any impact on the DCC proliferation rate (chapter 4.4.2.2).

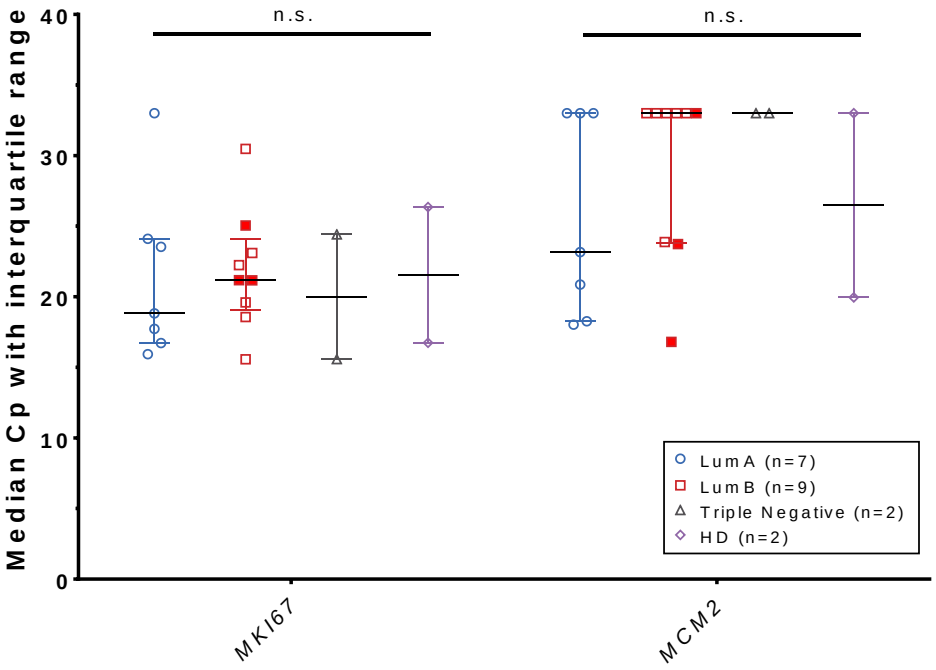


Figure 4-22 Expression of *MKI67* and *MCM2* in proliferating EpCAM⁺ DCCs stratified by subtype. The graph displays *MKI67* and *MCM2* expression in proliferating DCCs stratified by BC subtype and in NCCs derived from HDs. The data are shown as median Cp values, the whiskers represent the interquartile ranges. Shown NCCs were all balanced according to CNA analysis (aberrant NCCs were all non-proliferating). The red filling of some data points of the LumB dataset indicates patients with HER2 amplification. Statistics: two-way ANOVA (chapter 3.16.1.2); ns = not significant

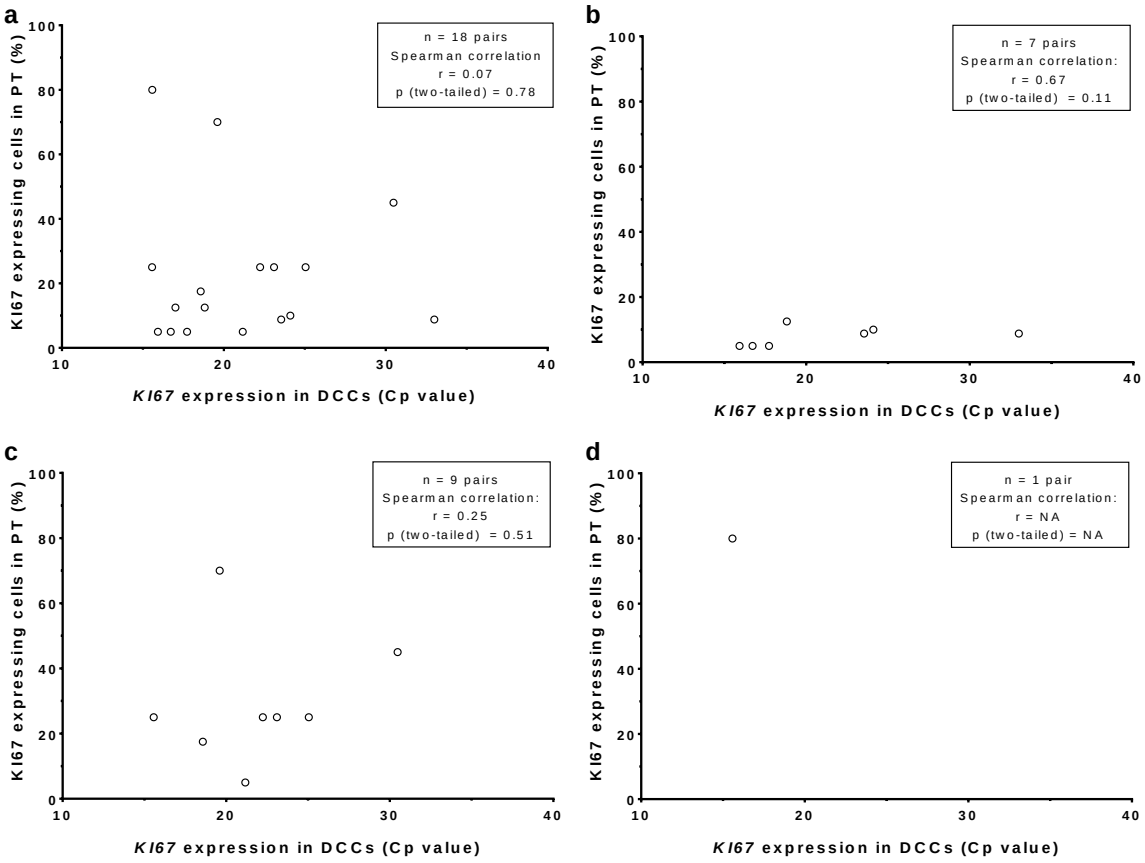


Figure 4-23 Correlation of KI67 level in PT with *MKI67* expression in proliferating DCCs stratified by subtype. The scatter plots show the percentage of KI67-expressing cells in the PT plotted against the *MKI67* expression and the calculated Spearman correlation coefficient r as well as the corresponding p-value of the correlation for different BC subtypes. (a) DCCs of all subtypes together, (b) LumA DCCs, (c) LumB DCCs, (d) TNBC DCCs. Statistics: Spearman correlation (chapter 3.16.1.2). NA = not applicable

4.4.2.2 Proliferation in M0 versus M1 DCCs

To investigate, whether the metastatic state of the patient had any impact on the DCC proliferation rate, I also quantified the expression of *MKI67* and *MCM2* in M1 DCCs, classified the cells into proliferating and non-proliferating ones, and looked at the frequency of proliferating cells compared to M0 DCCs.

Interestingly, the M1 collective contained significantly fewer proliferating cells than the M0 collective (Fisher's exact test, cell count used as input, p=0.011), with a frequency of 15 % compared to 51 %, respectively (Figure 4-24).

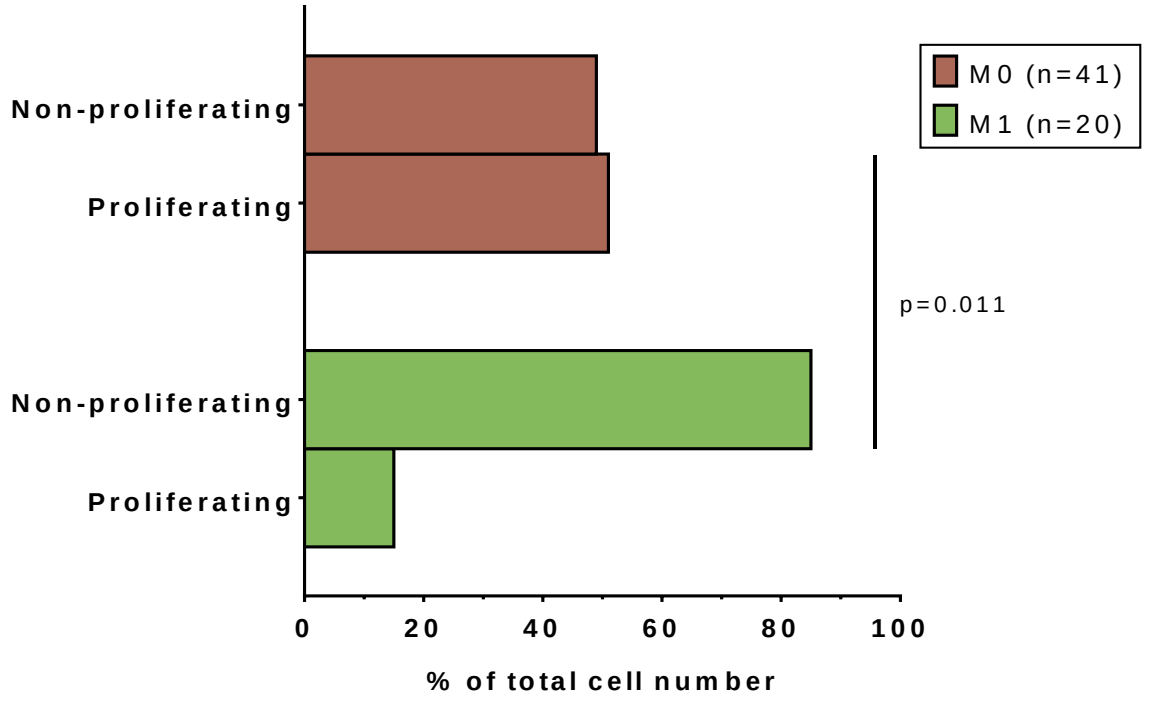


Figure 4-24 Frequency of proliferating and non-proliferating cells among EpCAM⁺ M0 and M1 DCCs. The bars represent the percentages of proliferating and non-proliferating DCCs from M0 or M1 patients. The raw counts instead of the displayed percentages were used for the statistics. Statistics: Fisher's exact test (chapter 3.16.1.2)

Since KI67 data of the PTs of most M1 patients were missing and those DCCs, for which I had the matched KI67 status of the PT data, were all non-proliferating, the correlation between the PT and the matched DCCs was not calculated and the proliferation marker expression of the M1 DCCs was not plotted as was done previously for the M0 DCC collective (see Figure 4-19).

As neither CNA profiles nor proliferation marker expression analyses yielded any differences between LumA and LumB, I proceeded with a global gene expression analysis (chapter 4.5), in order to identify discrepancies between the two subtypes.

4.5 Global gene expression of LumA and LumB subtype DCCs

As I was unable to identify differences between the LumA and LumB subtypes both in the rates of EpCAM⁺ cells in the BM (chapter 4.1.1) and the rate of true DCCs (chapter 4.2.2.2), as well as CNAs (chapter 4.3.4) and proliferation marker expression (chapter 4.4.2.1), I aimed to perform in-depth gene expression profiling of DCCs by RNA-Seq. To produce robust data for bioinformatic analysis (chapters 4.5.2-4.5.5), I first selected the best cells available at that time. The selection criteria are discussed in section 4.5.1. Lastly, I proposed ten candidate genes for future studies (chapter 4.5.6).

4.5.1 Sample selection

The aim for the sample selection was to have at least ten LumA and ten LumB M0 DCCs, five to ten M1 cells, which had previously been tested by mCGH, and ten HD-derived NCCs as controls. The samples were already selected early on in the project after histological data became available to enable the subtyping of patients (see Table 4-2 and Table 4-4) and the qPCR DCC signature genes had been measured in the EpCAM$^+$ cell collective (see chapter 4.2.1.3). For this reason, the LP-Seq data were not yet available when the final selection was made. Therefore, the majority of DCCs were primarily chosen according to their CNA status determined by mCGH experiments (Haunschild, 2013). Unfortunately, the desired numbers of LumA and LumB DCCs could not be reached using only the mCGH data, so the qPCR DCC signature (stringent variant, meaning only the cells expressing the DCC signature, no DCC-like cells) had to be included to narrow down the most promising cells to fill the gaps. For the selection of the remaining cells, the quality according to the QC-PCR and availability of both WTA and WGA products for the same cell were considered in addition to the M0 DCC signature result. With this approach, I ended up with ten LumA and twelve LumB DCCs. Note that only cells of LumA patients with a KI67 status in the PT ≤10 % and cells of LumB patients with a KI67 in the PT ≥20 % were included (see Table 4-3), in order to have a safety margin between the subtypes instead of a narrow cutoff at 14 %. Furthermore, three mCGH-confirmed M0 DCCs of other subtypes, two of the TNBC type and one of the Luminal undefined type were included due to mCGH-confirmed aberrations. In addition to the M0 DCCs, nine M1 cells that were aberrant according to mCGH and eleven NCCs (controls) were also included. For the NCCs, I had to rely on the M0 DCC signature result alone, since there were no mCGH data on any of these cells. I selected HD-derived EpCAM$^+$ cells that expressed the NCC pattern according to the qPCR, unfortunately, some of the cells dropped out during library preparation and due to the limited number of available NCCs, they had to be replaced with two NCC-like cells and one with undefined expression pattern (see Table 4-5). An overview of the final cells for RNA-Seq, which were retrospectively classified as aberrant or balanced using the LP-Seq data acquired later during the study, is provided below (Table 4-19).

Table 4-19 Overview of cells selected for deep RNA-Seq. The underlined data are the total cell numbers of the M0, M1, and NCC groups, while the values in italics represent the numbers stratified according to the BC subtype. Retrospective CNA classification as a combination of mCGH and LP-Seq data (only few of the selected cells had data from both methods) is provided in the three columns headed by "CNA". The original inclusion criteria (mCGH or qPCR signature) are given in parenthesis in the first column.

Cell type	Number of cells	CNA aberrant	CNA balanced	CNA unclear
M0 (mCGH aberrant)	15	14	0	1
LumA	*8*	*7*	*0*	*1*
LumB	*4*	*4*	*0*	*0*
Luminal undefined	*1*	*1*	*0*	*0*
TNBC	*2*	*2*	*0*	*0*
M0 (M0 DCC qPCR signature)	10	3	2	5
LumA	*2*	*1*	*0*	*1*
LumB	*8*	*2*	*2*	*4*
M1 (mCGH aberrant)	9	9	0	0
LumB	*3*	*3*	*0*	*0*
Luminal undefined	*3*	*3*	*0*	*0*
TNBC	*3*	*3*	*0*	*0*
NCC (M0 DCC qPCR signature)	11	1	1	9
Total	**45**	**27**	**3**	**15**

Following the sample selection, the preparation of the sequencing libraries was carried out according to the protocol described in section 3.12. After sequencing was done, the data were analyzed. First, the quality control (QC) of the raw data was performed (chapter 4.5.2).

4.5.2 Bioinformatic quality control of RNA-Seq data

4.5.2.1 Raw data QC

RNA-Seq data analysis was performed by Dr. Huiqin Koerkel-Qu, who first checked the overall quality of the data using the *FastQC* software (Andrews, 2010) to generate QC reports for each sample, followed by compilation of a summary report using *MultiQC* (Ewels et al., 2016). A more detailed description of the procedure, including all the following steps, is provided in chapter 3.16.3.

The numbers of sequencing reads obtained for each sample are illustrated below (Figure 4-25). Due to sample pooling, the expected maximum number of reads per sample was 29 million, but this value was only reached by a few of the samples. There were also three outliers with surprisingly high numbers of reads, namely >40 million and even >80 million in one case. Likewise, there were also several samples with less than 20 million reads and even one below ten million. The average number of reads across all samples was 24.5±13.1 million (median 21.5 million). The QC also showed a high amount of duplicated reads (Figure 4-25), however this is expected for deep RNA-Seq, because the duplicates represent highly expressed transcripts (Choy et al., 2015). Due to the paired-end sequencing, there were two data sets for each of the cells. The read counts of forward (Read 1) and reverse (Read 2) datasets of each sample were similar (Figure 4-25).

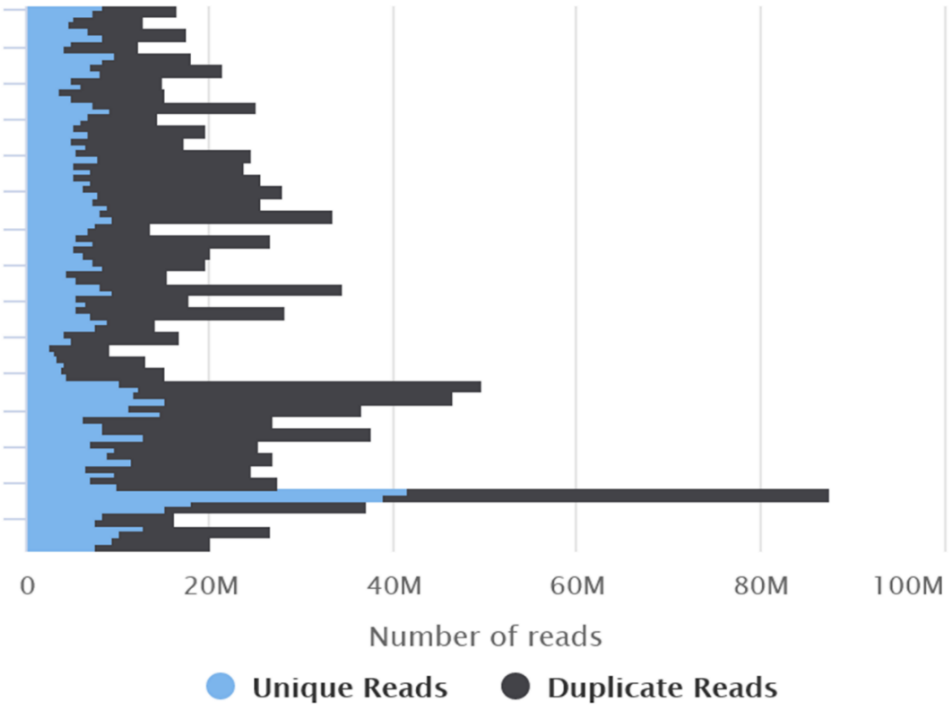


Figure 4-25 Raw sequence counts. The plot shows the total amount of sequencing reads obtained per sample (n=45) represented by the total length of each bar. Each bar corresponds to one sample and is vertically subdivided into number of unique reads (blue) and duplicated reads (grey) out of the total number of reads. Duplicated reads indicate that some transcripts were highly expressed, but in general they are not considered problematic in RNA-Seq (Choy et al., 2015). Additionally, as we performed paired-end sequencing, we have two datasets for each sample, called Read One and Read Two. Therefore, each sample's bar is horizontally subdivided to represent Read One and Read Two.

Furthermore, the analysis of Phred scores revealed that the sequencing reads of all samples had a high quality (Figure 4-26). The Phred score is calculated by the sequencing machine for every read and provides an estimate of the probability of calling a wrong nucleobase at any base position in a read with a score of 30 corresponding to a probability of 0.1 % to call a wrong base (Ewing et al., 1998; Ewing and Green, 1998). The average Phred scores were close to or above 30 for all samples and at all positions of the reads, indicating robust base calling (Figure 4-26).

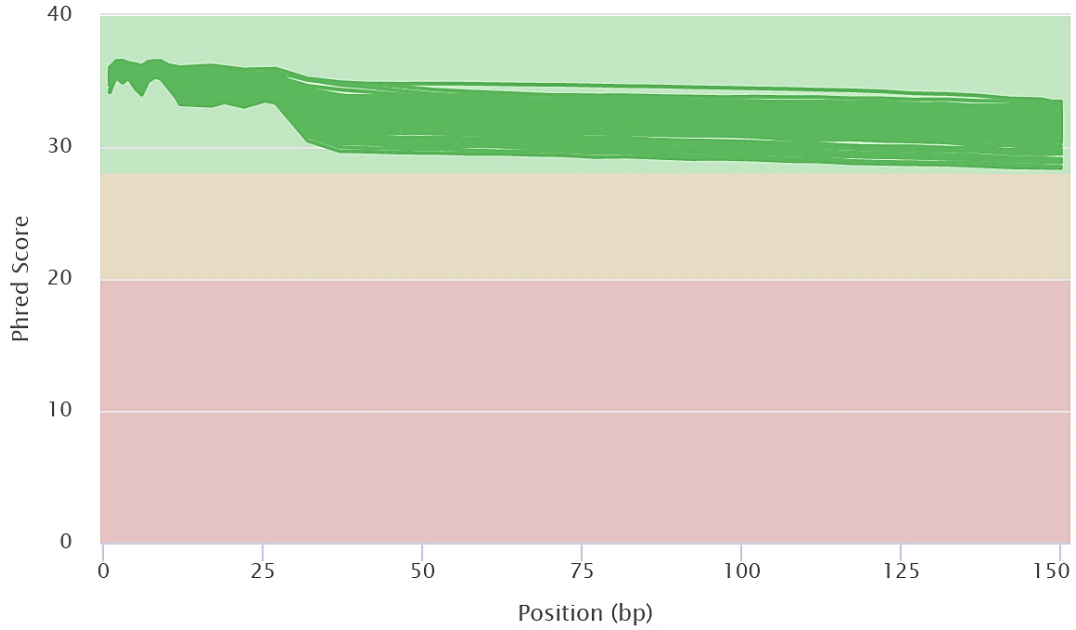


Figure 4-26 Average Phred scores. The plot illustrates the average distribution of Phred scores (y-axis) across all 150 base positions (x-axis) of the sequencing reads for all samples. Read One and Read Two of each sample are plotted individually, therefore there are 90 overlapping lines.

Another important quality metric is the GC content of the sequencing reads, (Figure 4-27). Depending on the organism under investigation, each of the four nucleobases should make up around 25 % of the total bases, meaning that the GC content should be around 50 %. When plotted, the GC content optimally follows a normal distribution around the 50 % mark. A shift of the normal distribution indicates a systematic bias in the sequencing library or source material, which is often caused by PCR amplification of the material. In contrast, an abnormal looking distribution may indicate a contamination or another kind of bias. In our data, we observed that most samples displayed a normal distribution around 50-55% (green [n=28] and yellow [n=37] lines in Figure 4-27) suggesting good to acceptable quality. Again, each sample's Read One and Read Two were treated separately leading to a total of 90 datasets. Apart from these datasets, there were also 25 with a skewed distribution peaking between 55-60 % of GC content (red lines). As GC bias removal is not widely accepted in the bioinformatics community, no correction was performed. However, the likely reason for this observation is that the QC was done on the raw reads before adapter trimming. Due to our WTA adapters (CP2 primer) containing long GC-stretches, the problematic results are probably caused by those adapters.

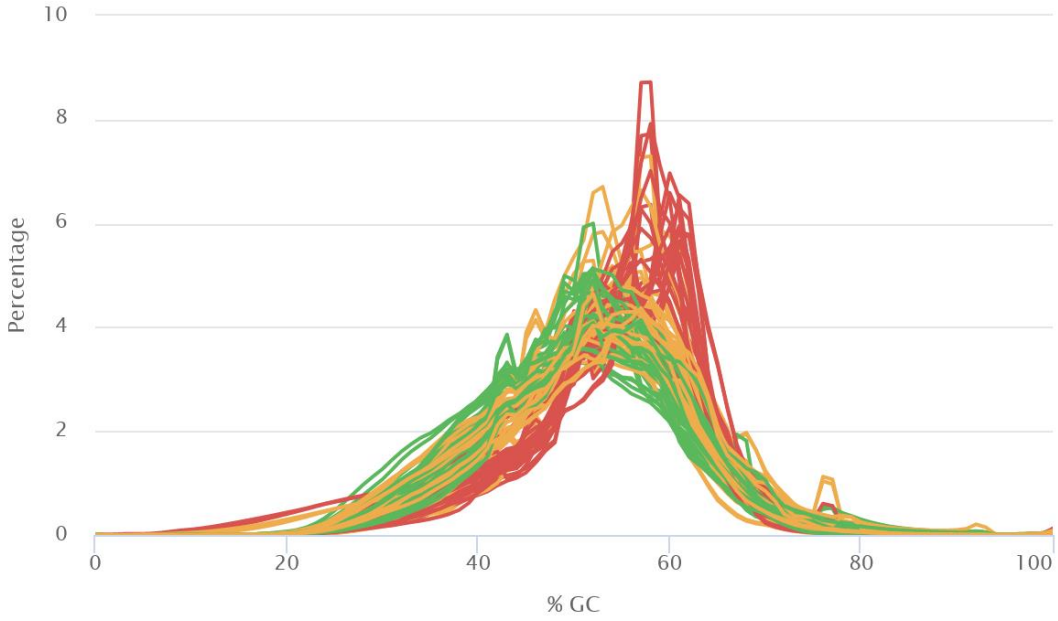


Figure 4-27 GC content per sequence. The graph depicts the percentage of GC content (x-axis) plotted against the percentage of reads with the indicated GC content (y-axis). The datasets shown in green matched the model distribution of FastQC, while the yellow ones still fit, but were slightly shifted to the right. In contrast, the datasets marked red displayed an abnormal shape.

Following evaluation of the raw read quality, trimming and mapping of the reads to the reference genome were performed (see chapter 3.16.3 for the protocol), followed by QC on the mapping (chapter 4.5.2.2).

4.5.2.2 Mapping QC

On average, only 40.2±15.3% of reads could be mapped to the human reference genome using the default settings of the *STAR* software. However, as we still had high numbers of mapped reads for most samples (average 5.61±3.08 million), this relatively low mapping rate was not problematic. The alignment was repeated with different settings and the mapping rate improved, but this did not significantly impact the downstream analyses, therefore, we decided to work with the datasets aligned using the standard settings to make the results more comparable to other studies. Note that upon mapping the Read One and Read Two datasets of each sequenced sample were combined, therefore, the number of datasets was down to the 45 individual cells from this point.

Additionally, we also looked at the gene coverage profiles of all samples. The sample with >80 million reads was excluded (see Figure 4-25), because it had such a high coverage that it masked all other samples and also expressed an abnormal number of 40,000 genes. The remaining 44 datasets revealed an overall bias towards the 5' region of transcripts, which was more pronounced in some samples than in others (Figure 4-28). The sample with the high coverage showed this bias to a far lesser extent (Figure 4-29).

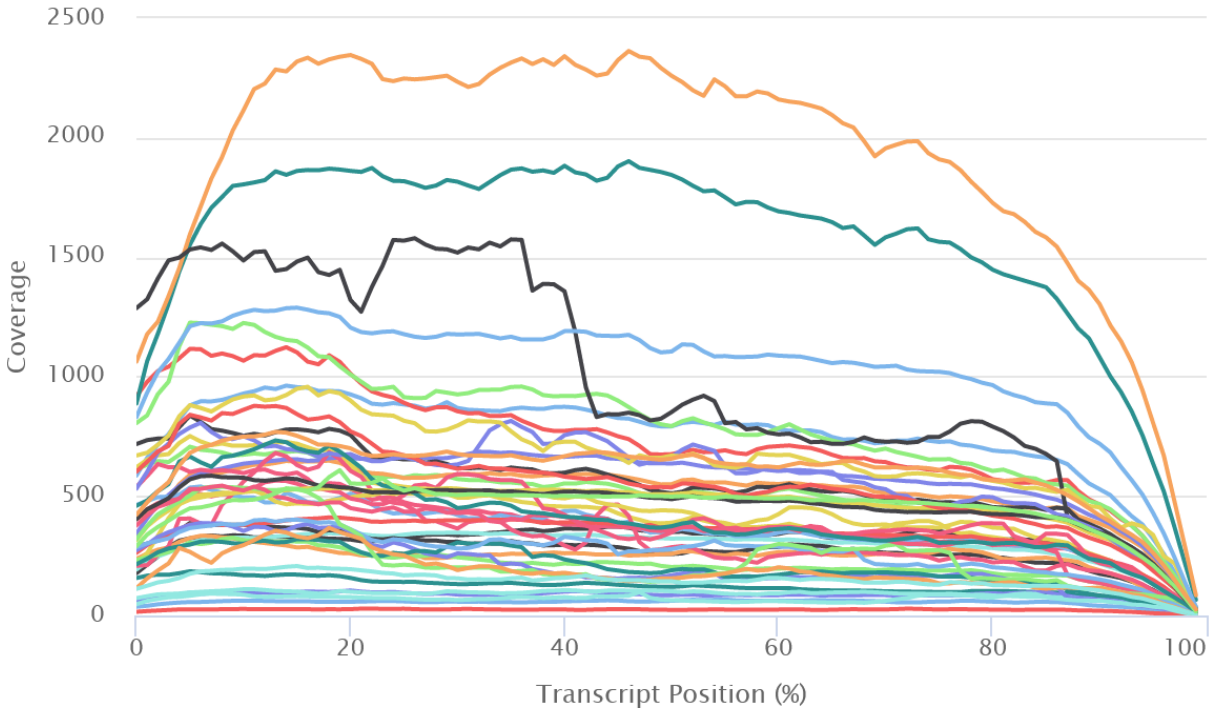


Figure 4-28 Gene coverage profiles of transcripts without outlier sample. The graph shows the coverage of genes across all mapped transcripts for each dataset. The sample with >80 million reads was excluded (see Figure 4-25), because it had such a high coverage that it masked all other samples.

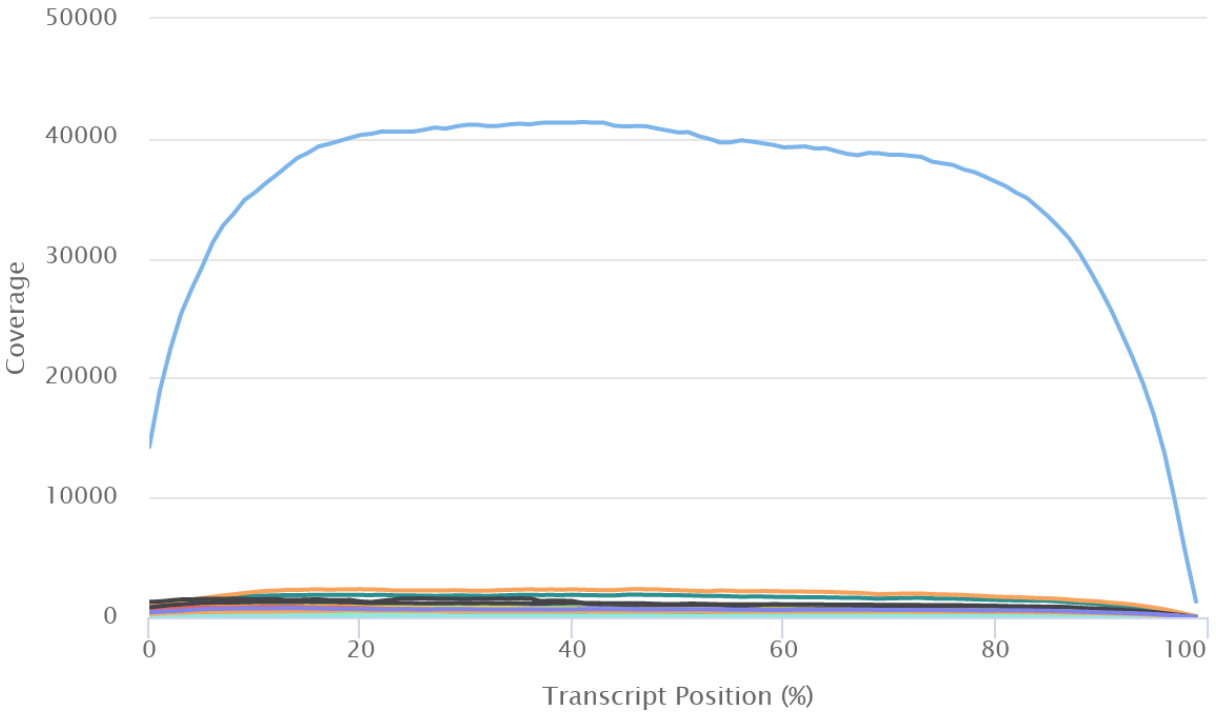


Figure 4-29 Gene coverage profiles of transcripts with all samples. The graph depicts the coverage of genes across all mapped transcripts for each dataset. The sample with >80 million reads (see Figure 4-25) masked all other samples, because of its extremely high coverage.

Taken together the QC indicated that the sequencing data were of sufficient quality for further analysis. Therefore, we first looked at the expression of proliferation-associated genes in the sequenced DCCs (chapter 4.5.3) to confirm the previous proliferation data obtained by qPCR (see chapter 4.4.2.2).

4.5.3 Expression of cell cycle-associated genes

In order to corroborate the previous finding that roughly half of M0 DCCs were proliferating (see Figure 4-21), we looked at the expression of numerous cell cycle-associated genes in the RNA-Seq data of all DCCs (M0 and M1). For this purpose, all genes belonging to the gene ontology (GO) term "cell cycle" (GO: 0007049, biological process, 671 genes) were analyzed and differentially expressed genes between proliferating and non-proliferating DCCs (proliferation status defined by qPCR; M0 and M1-derived DCCs combined) from these 671 genes were identified. Subsequently, unsupervised clustering was performed to generate a heat map using only the significantly different genes (details of the selection process are provided in chapter 3.16.3).

We observed three large clusters distributed across the two main branches of the clustering analysis (Figure 4-30). The first cluster (DCC-NP) consisted only of non-proliferating cells and made up the whole left branch of the clustering. In contrast, the second (DCC-Pa) and third (DCC-Pb) clusters consisted almost exclusively of proliferating DCCs. Each of these two clusters made up about half of the right branch of the clustering tree. The proliferating DCCs could be further subdivided into two populations of roughly equal size. LumA and LumB subtype DCCs were similarly distributed across the two main branches (left branch with DCC-NP and right branch with DCC-Pa and DCC-Pb) of the clustering (Fisher's exact test, two-tailed p=0.38). However, comparison of the two proliferating clusters DCC-Pa and DCC-Pb revealed a significantly different distribution of LumA and LumB (Fisher's exact test, two-tailed p=0.013). While DCC-Pa contained DCCs of all four different subtypes, DCC-Pb was almost exclusively made up of LumB-derived DCCs with the exception of one LumA-derived cell. Furthermore, the data also showed that six out of nine M1-stage DCCs ended up in the DCC-NP cluster containing only non-proliferating cells, which was in accordance with previous results that M1 DCCs were less proliferative than M0 DCCs (see Figure 4-24). Lastly, the two proliferating clusters consisted mostly of cells that were classified as DCCs according to the M0 DCC qPCR signature as compared to the non-proliferating cluster, which supports the validity of the signature.

Following the examination of the expression of cell cycle-associated genes, we examined whether the KI67 status of the PT was correlated to the overall gene expression in DCCs (chapter 4.5.4.).

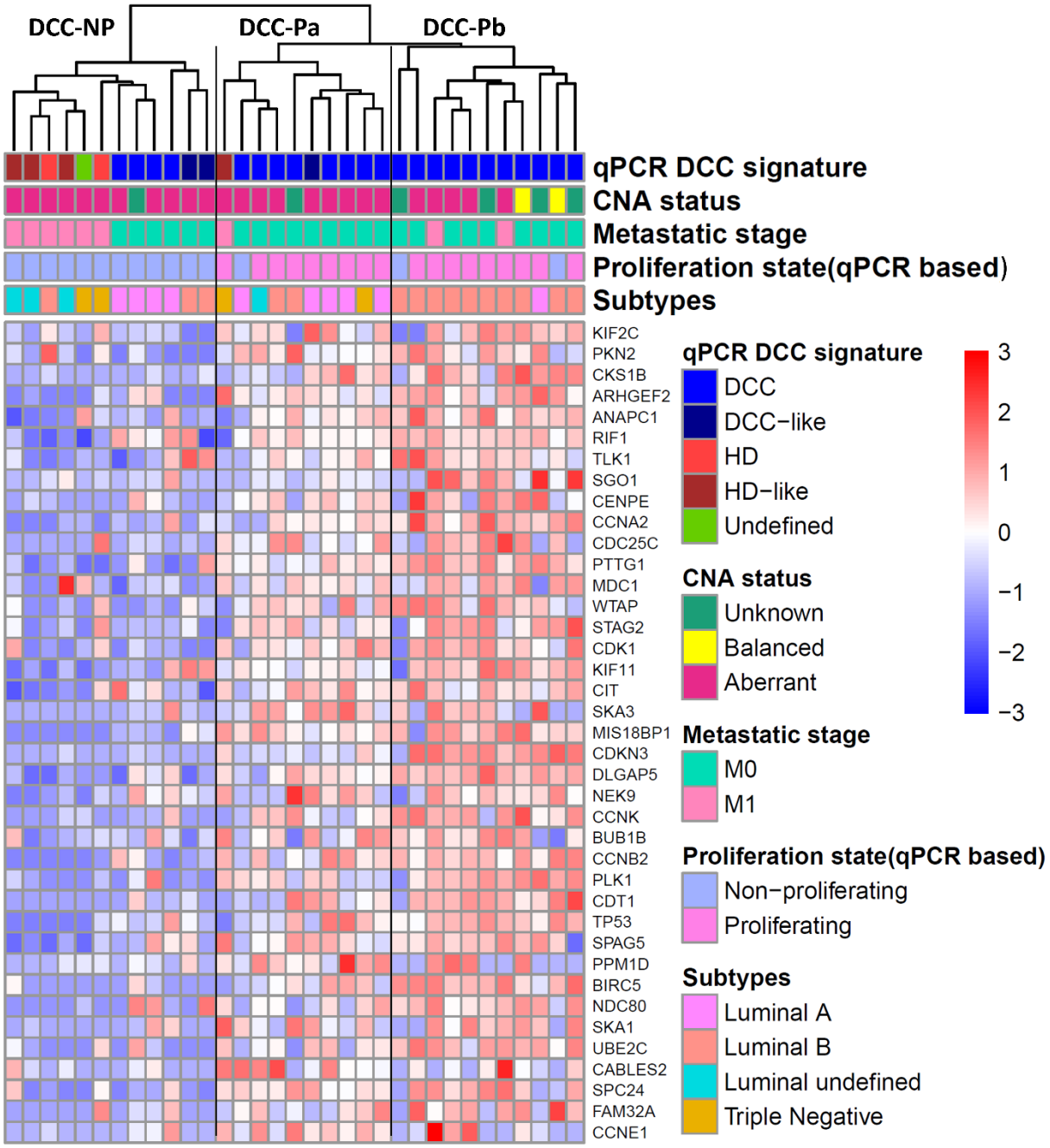


Figure 4-30 Expression of cell cycle-associated genes in proliferating and non-proliferating DCCs. The heat map displays the normalized log count of expression of the indicated genes (represented by the rows) in both proliferating and non-proliferating M0 and M1 DCCs (each cell is represented by one column). The proliferation state was determined according to the qPCR results (chapter 4.4). The colored bars on the top indicate the result of the M0 DCC qPCR signature, the CNA status, the metastatic state, the proliferation status according to qPCR, and the BC subtype. The branches on the top depict how similar the individual DCCs are to the other cells. Two vertical black lines were added to highlight the separation into three distinct clusters of cells. The names given to each cluster are displayed on top.

4.5.4 Correlation of the KI67 status of the PT with overall gene expression

So far, only the correlation of the PT KI67 status with *MKI67* expression was examined by qPCR (see Figure 4-20 and Figure 4-23). Therefore, we investigated whether the KI67 status of the PT was associated with the overall gene expression patterns of matched DCCs. Indirectly, this would also provide evidence for differences in the gene expression between LumA and LumB, as the subtypes are defined by different KI67 levels in the PT (see Table 2-2, Cheang et al., 2009;

Goldhirsch et al., 2011). For this reason, we first looked at the correlation of the KI67 level of the PT with the *MKI67* expression in ten sequenced LumA and 14 LumB DCCs (12 M0 and two M1 DCCs that clustered closely with M0 LumB DCCs) as a reference and then correlated the KI67 level of the PT with the principal component 1 (PC1), one of the variables derived from a dimension reduction analysis of the overall gene expression pattern (see chapter 3.16.3 for workflow), which represented most of the variance (14 %) in the gene expression.

The analysis uncovered that the KI67 status of the PT showed a tendency, but was not significantly correlated with the *MKI67* expression in the DCCs (r=0.38, p=0.06, Figure 4-31a). However, there was a robust correlation of the KI67 status with the overall gene expression pattern represented by PC1 (r=0.54, p=0.006, Figure 4-31b) leading to a separation of LumA and LumB M0 DCCs. Furthermore, there was also a strong correlation of the *MKI67* expression in DCCs with the PC1 in the DCCs (r=0.82, p=2.3 x 10^{-6}, Figure 4-31c), but this did not separate the LumA and LumB DCCs.

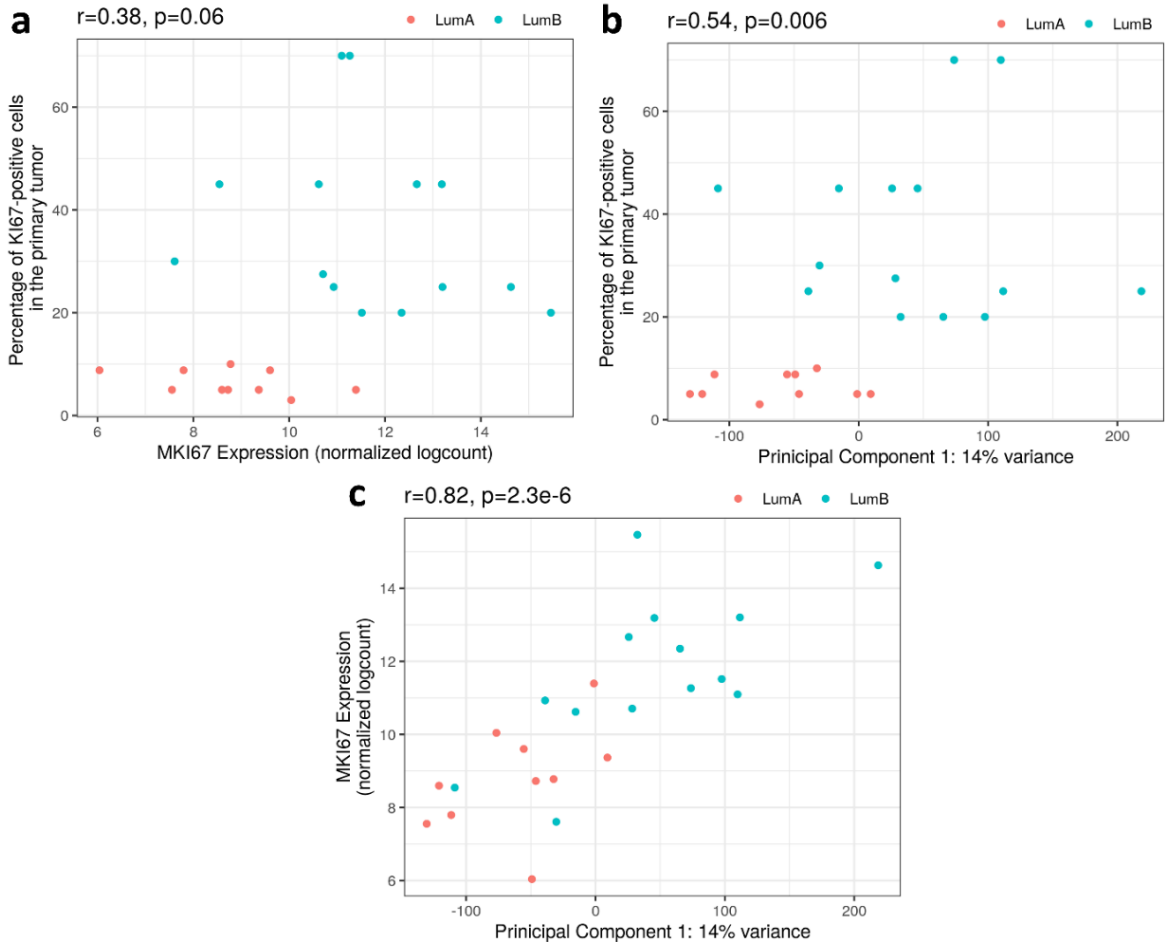


Figure 4-31 Correlation of KI67 status in PT with *MKI67* level in DCCs and overall gene expression in DCCs. The scatter plots illustrate the correlation of the percentage of KI67-expressing cells in the PT with (a) the *MKI67* expression in DCCs and (b) PC1 representing 14 % of the overall variation in the gene expression of the DCCs. Additionally, (c) depicts the correlation of *MKI67* expression in DCCs with PC1. The colors of the data points indicate the BC subtype of the DCCs according to the legend on the top right of each panel. The calculated correlation coefficient r and corresponding p-values are provided at the top of each panel (for details on analysis see chapter 3.16.3).

Since these data looked promising, we then proceeded with identification of genes associated with PC1, in order to carry out a gene ontology (GO) analysis (chapter 4.5.5).

4.5.5 Gene ontology analysis

4.5.5.1 Most relevant GO terms

As the LumA and LumB subtypes - represented by the KI67 status of the PT - were robustly correlated with PC1 (see Figure 4-31b), it was decided to use PC1-associated genes for further analyses. The procedure for the identification of the genes is described in chapter 3.16.3. Gene lists were created for both positive and negative directions of PC1, meaning one list for genes down-regulated in LumB (negative direction, 815 genes) and one with genes up-regulated in LumB (positive direction, 470 genes). The full gene lists are provided in appendix chapter 12.2.

Using the full gene lists, numerous biological pathways, which were overrepresented among the two sets of genes, were identified using the DAVID database (Huang et al., 2009b, 2009a). From the two resulting GO lists for LumB down- and LumB up-regulated genes, the top 25 and 50 GO terms, respectively, from a combination of all three GO categories ("biological process", "cellular component", and "molecular function") were screened and 20 terms that were considered most relevant for BC were selected from each list. The respective 20 GO terms per list were then plotted.

Among the GO terms overrepresented in LumB down-regulated genes were several ones related to both the plasma membrane and membranes in general, cell adhesion and extracellular matrix, and transmembrane transport including calcium ion channels (Figure 4-32). In contrast, many GO terms overrepresented in LumB up-regulated genes were related to splicing, ribosomes, and translation, indicating a higher proliferation (Figure 4-33). However, there were also GO terms similar to the down-regulated genes which were associated with cell-cell adhesion and extracellular matrix. By comparing the two figures, we observed that the GO terms related to LumB up-regulated genes were more focused on only a few terms with very high $-\log_2$ False discovery rate (FDR; Figure 4-33) as compared to the LumB down-regulated genes, among which the $-\log_2$ FDR values were less pronounced (Figure 4-32). The main reason for this was likely that we applied less strict criteria for selection of the LumB down-regulated genes, as the same criteria that we had used for the LumB up-regulated genes (FDR < 0.01) yielded only four candidates. Therefore, the cutoff was changed to an FDR < 0.05 and a log Fold change < -0.01.

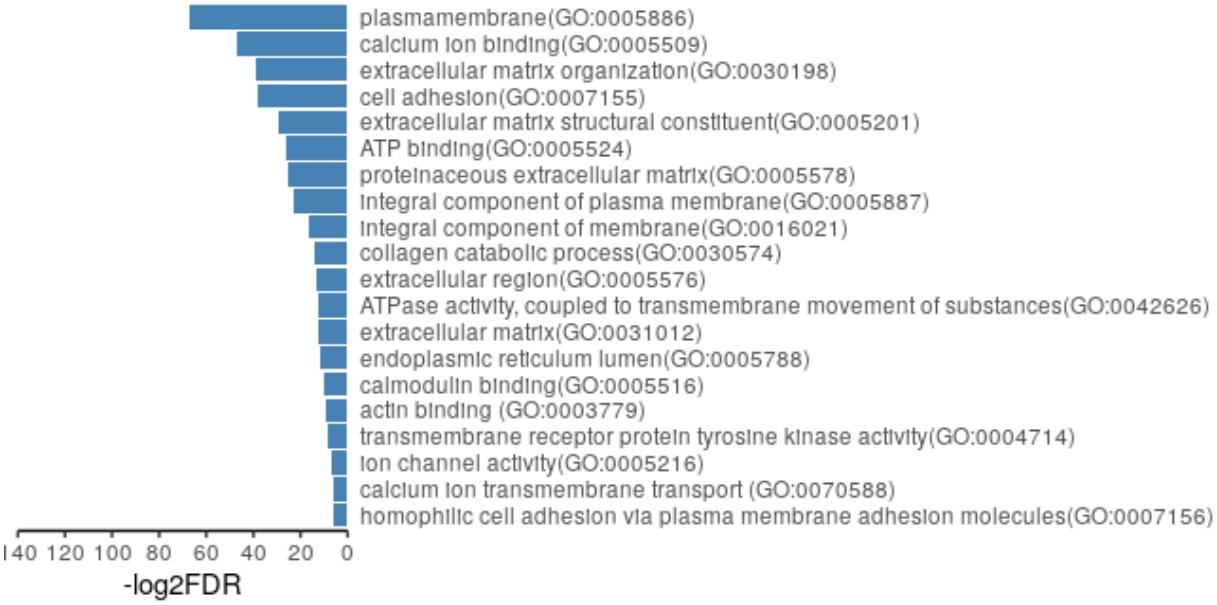


Figure 4-32 GO terms overrepresented in LumB down-regulated genes. The bar plot illustrates the negative $\log_2$ FDR (q-value) of selected GO terms provided by the DAVID database for the genes down-regulated in LumB DCCs.

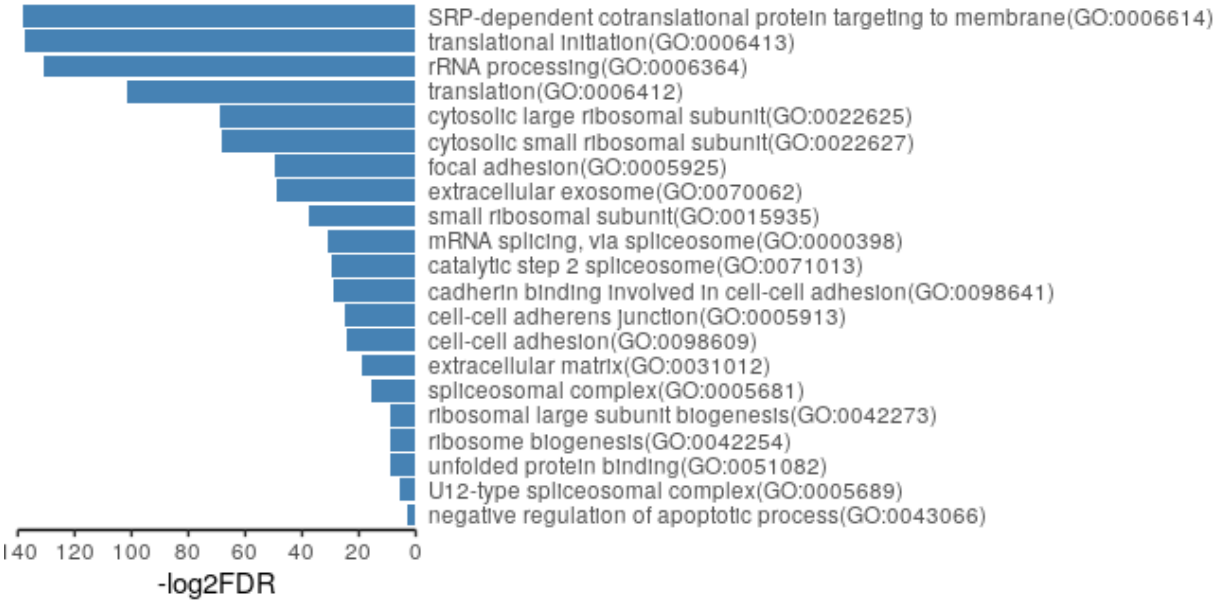


Figure 4-33 GO terms overrepresented in LumB up-regulated genes. The bar plot illustrates the negative $\log_2$ FDR (q-values) of selected GO terms provided by the DAVID database for the genes up-regulated in LumB DCCs.

4.5.5.2 GO term network

To visualize the GO terms characteristic for the LumB subtype in more detail, the two previously compiled gene lists (chapter 4.5.5.1) were used to generate two separate GO networks. To this end, a software called *Cytoscape* was used. The exact procedure and software settings for the creation of the networks is described in the methods (chapter 3.16.4). I used an FDR < 0.001 (as opposed to FDR < 0.01 for selection of LumB up-regulated genes above in chapter 4.5.5.1) as a threshold to reduce the number of resulting GO terms enough to be able to fit the whole network on one page while remaining legible. Note that this step represents a separate GO term analysis with another software, therefore the results may differ from the previous analysis (see chapter 4.5.5.1). Additionally, the networks contain all resulting GO terms without any pre-selection.

Wherever possible, thematically distinct clusters were spatially separated in order to highlight them more clearly.

Analysis of the 815 LumB down-regulated genes, uncovered enrichment of 109 GO terms (Figure 4-34). Due to this high number, I will only mention some of the main categories (first- or second-degree neighbors of the central node) of the network. These main GO categories were again *cell adhesion* and *extracellular structure organization,* but also *cell development, cell differentiation, regulation of biological process, signaling pathway, system process, transport, localization,* and *metabolic process.* These categories mostly matched those that were previously selected (see Figure 4-32), but provided many additional ones.

Figure 4-34 Network of GO terms overrepresented in LumB down-regulated genes. See next page for figure. The network, consisting of 109 nodes, illustrates the biological processes overrepresented in the 815 LumB down-regulated genes. The color intensity of the nodes reflects the q-value of the statistical test. Bright (yellow) nodes indicate low significance, while dark (orange) nodes represent highly significant GO terms. White nodes represent GO terms with q-values >0.001 and are mostly there to connect the colored nodes to the root of the network and to clarify relations between GO terms. The size of the nodes reflects the relative number of genes belonging to the respective GO term. The network was organized with the yFiles Radial Layout. Therefore, the "biological_process" root of the network is located in the center, while the more general main categories are close to the center and the more specific ones radiate outward.

For the 470 LumB up-regulated genes, an overrepresentation of 78 GO terms was observed (Figure 4-35). As before, only some of the main categories of the network will be mentioned. These main GO categories were *organelle organization, cellular component assembly, ribonucleoprotein complex biogenesis, cellular macromolecular complex subunit organization, regulation of biological process, biosynthetic process,* and several variations of *metabolic process.* These categories mostly matched those selected from the previous analysis (see Figure 4-33), but expanded the picture by a large margin.

Figure 4-35 Network of GO terms overrepresented in LumB up-regulated genes. See second next page for the figure. The network, consisting of 78 nodes, illustrates the biological processes overrepresented in the 470 LumB up-regulated genes. The color intensity of the nodes reflects the q-value of the statistical test. Bright (yellow) nodes indicate low significance, while dark (orange) nodes represent highly significant GO terms. White nodes represent GO terms with q-values >0.001 and are mostly there to connect the colored nodes to the root of the network and to clarify relations between GO terms. The size of the nodes reflects the relative number of genes belonging to the respective GO term. The network was organized with the yFiles Radial Layout. Therefore, the "biological_process" root of the network is located in the center, while the more general main categories are close to the center and the more specific ones radiate outward.

Following the GO analysis, ten promising target genes for future studies were selected from the two available gene lists (chapter 4.5.6).

Figure 4-34

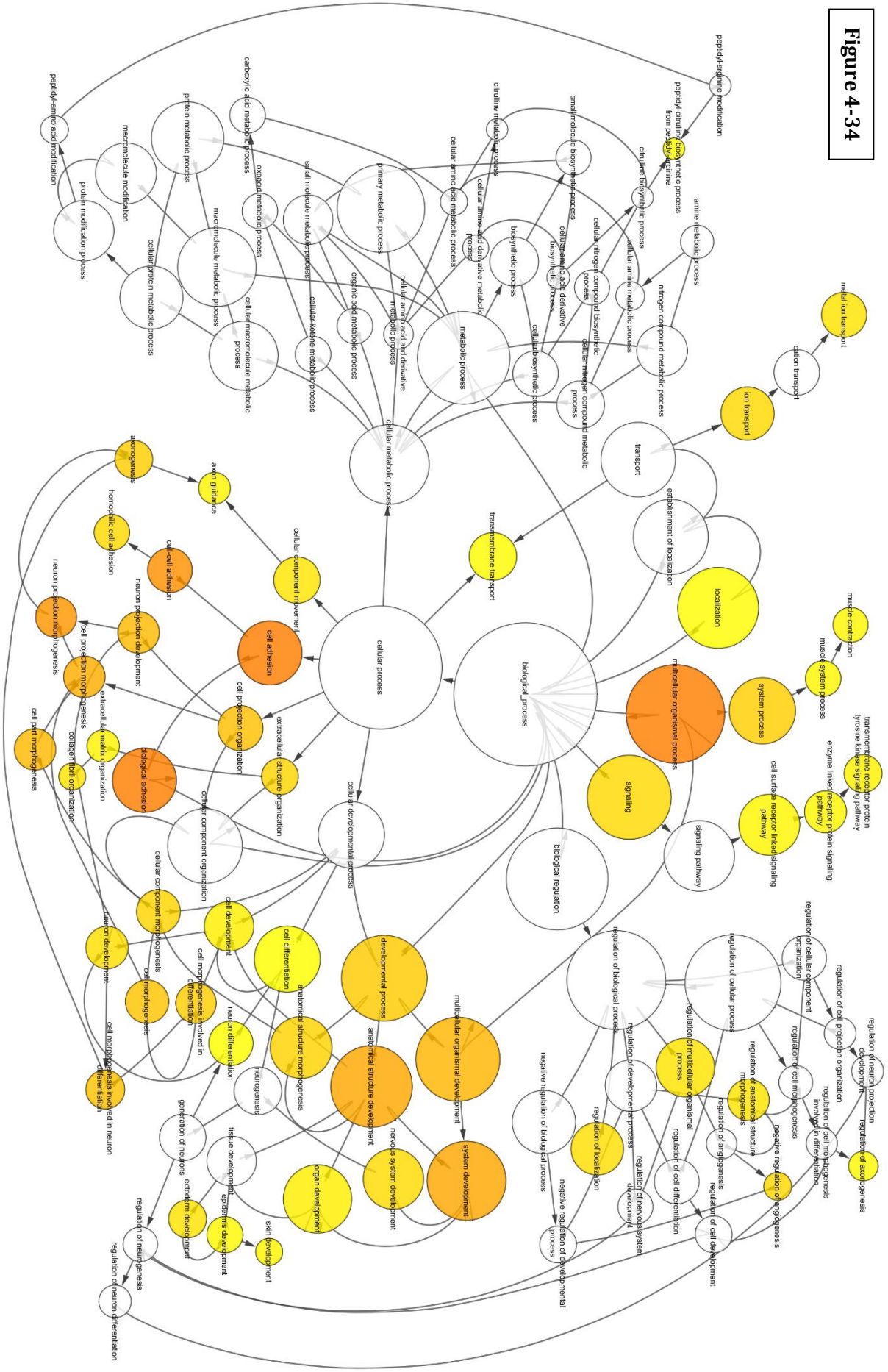
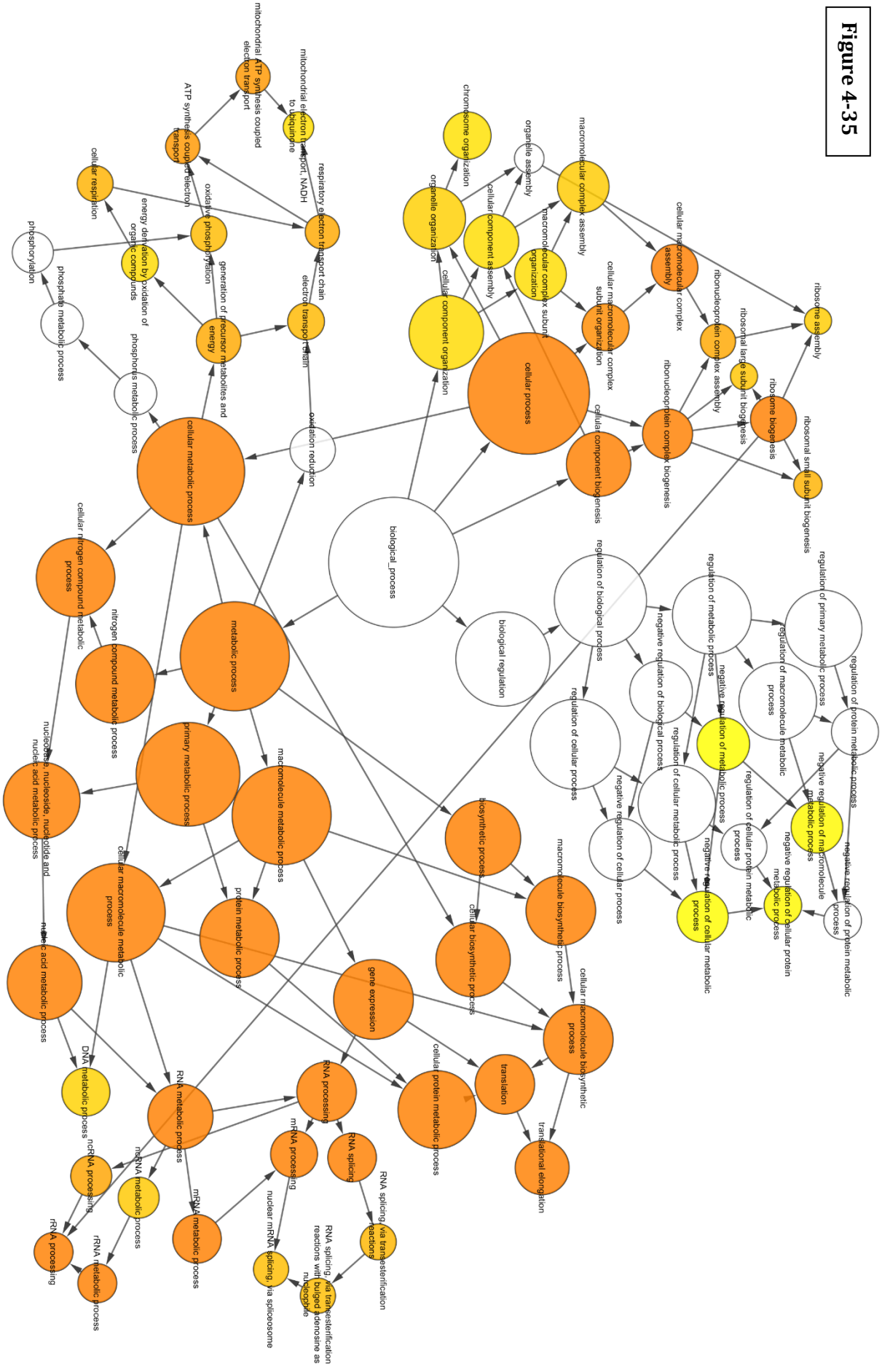

peptidyl-arginine modification
peptidyl-citrulline biosynthetic process from peptidyl-arginine
peptidyl-amino acid modification
metal ion transport
cation transport
ion transport
transport
establishment of localization
localization
transmembrane transport
amine metabolic process
citrulline biosynthetic process
nitrogen compound metabolic process
cellular amine metabolic process
cellular nitrogen compound biosynthetic process
cellular nitrogen compound metabolic process
small molecule biosynthetic process
cellular amino acid biosynthetic process
cellular amino acid metabolic process
citrulline metabolic process
cellular amino acid derivative metabolic process
small molecule metabolic process
primary metabolic process
metabolic process
cellular biosynthetic process
biosynthetic process
cellular metabolic process
carboxylic acid metabolic process
oxoacid metabolic process
organic acid metabolic process
cellular amino acid derivative metabolic process
protein metabolic process
macromolecule modification
protein modification process
macromolecule metabolic process
cellular protein metabolic process
cellular macromolecule metabolic process
cellular ketone metabolic process
cellular process
biological_process
multicellular organismal process
system process
muscle system process
muscle contraction
signaling
signaling pathway
transmembrane receptor protein tyrosine kinase signaling pathway
enzyme linked receptor protein signaling pathway
cell surface receptor linked signaling pathway
biological regulation
regulation of biological process
regulation of cellular process
regulation of multicellular organismal process
regulation of biological quality
regulation of developmental process
regulation of localization
regulation of cellular component organization
regulation of cell projection organization
regulation of anatomical structure morphogenesis
regulation of multicellular organismal process
regulation of cell morphogenesis
regulation of cell differentiation
regulation of neuron projection development
regulation of cell morphogenesis involved in differentiation
regulation of angiogenesis
regulation of axonogenesis
regulation of nervous system development
regulation of cell development
negative regulation of biological process
negative regulation of developmental process
developmental process
cellular developmental process
anatomical structure development
multicellular organismal development
system development
nervous system development
organ development
tissue development
epithelium development
skin development
ectoderm development
generation of neurons
neurogenesis
cell differentiation
anatomical structure morphogenesis
neuron differentiation
cell development
cellular component movement
cellular component organization
cellular component morphogenesis
cell morphogenesis
cell morphogenesis involved in differentiation
neuron development
cell morphogenesis involved in neuron differentiation
regulation of neurogenesis
regulation of neuron differentiation
cell adhesion
cell-cell adhesion
homophilic cell adhesion
biological adhesion
axonogenesis
axon guidance
cell projection organization
cell projection morphogenesis
neuron projection development
neuron projection morphogenesis
cell part morphogenesis
extracellular structure organization
extracellular matrix organization
collagen fibril organization

Figure 4-35


biological_process
cellular process
metabolic process
cellular metabolic process
primary metabolic process
biological regulation
regulation of biological process
regulation of cellular process
regulation of metabolic process
regulation of primary metabolic process
regulation of macromolecule metabolic process
regulation of protein metabolic process
negative regulation of biological process
negative regulation of cellular process
negative regulation of metabolic process
negative regulation of macromolecule metabolic process
negative regulation of protein metabolic process
negative regulation of cellular protein metabolic process
negative regulation of cellular metabolic process
regulation of cellular metabolic process
regulation of cellular protein metabolic process
cellular component organization
cellular component assembly
cellular component biogenesis
organelle organization
organelle assembly
chromosome organization
macromolecular complex assembly
macromolecular complex subunit organization
cellular macromolecular complex assembly
cellular macromolecular complex subunit organization
ribonucleoprotein complex biogenesis
ribonucleoprotein complex assembly
ribosome biogenesis
ribosomal large subunit biogenesis
ribosomal small subunit biogenesis
ribosome assembly
macromolecule metabolic process
cellular macromolecule metabolic process
nitrogen compound metabolic process
cellular nitrogen compound metabolic process
nucleobase, nucleoside, nucleotide and nucleic acid metabolic process
protein metabolic process
cellular protein metabolic process
biosynthetic process
cellular biosynthetic process
macromolecule biosynthetic process
cellular macromolecule biosynthetic process
gene expression
translation
translational elongation
RNA processing
mRNA processing
RNA splicing
RNA splicing, via transesterification reactions
RNA splicing, via transesterification reactions with bulged adenosine as nucleophile
nuclear mRNA splicing, via spliceosome
RNA metabolic process
mRNA metabolic process
ncRNA processing
ncRNA metabolic process
rRNA processing
rRNA metabolic process
DNA metabolic process
nucleic acid metabolic process
oxidation reduction
respiratory electron transport chain
mitochondrial electron transport, NADH to ubiquinone
mitochondrial ATP synthesis coupled electron transport
ATP synthesis coupled electron transport
oxidative phosphorylation
cellular respiration
energy derivation by oxidation of organic compounds
generation of precursor metabolites and energy
phosphorylation
phosphate metabolic process
phosphorus metabolic process

4.5.6 Proposed candidate genes for further investigation

With the previously identified overrepresented GO terms in mind, the five most promising genes (Table 4-20) from each list (up- or down-regulated genes), which could be further investigated in future studies, were selected. The main criteria for the genes were low q-values, high fold change, and their function, but it was also considered whether they were already known as prognostic markers in BC or other types of cancer according to the Human Protein Atlas' Pathology Atlas (Uhlen et al., 2005; Uhlen et al., 2017). Regarding gene function, I tried to focus on those related to previously identified GO terms (chapter 4.5.5). Therefore, the desired categories were mainly cell adhesion, extracellular matrix, mRNA splicing, metabolic processes, and ribosomes/translation. Additionally, genes related to the hallmarks of cancer like apoptosis resistance or invasion were also considered.

Table 4-20 List of suggested LumB DCC associated genes for further study. All gene symbols, full names, and functions were retrieved from the Genecards database (Stelzer et al., 2016). Five up-regulated and five down-regulated genes were selected.

Up/Down regulated	Gene symbol	Full name	Function or related pathways
↓	*ANKRD13B*	Ankyrin Repeat Domain 13B	Positive regulation of internalization of ligand-activated EGFR (HER1)
↓	*CD6*	T-Cell Differentiation Antigen CD6	Cell-cell adhesion, promotion of T-cell activation
↓	*FAT4*	FAT Atypical Cadherin 4	Calcium ion binding, inhibition of YAP1-mediated proliferation and differentiation, maintenance of planar cell polarity
↓	*IFNL2*	Interferon Lambda 2	activation of the JAK/STAT signaling, antitumor activity
↓	*TIRAP*	Toll/Interleukin-1 Receptor Domain-Containing Adapter Protein	Activation of NF-κB, MAPK1, MAPK3 and JNK, cytokine secretion and inflammatory response, positive regulation of production of TNF-alpha and interleukin-6
↑	*CTTN*	Cortactin	Regulation of interactions between components of adherens junctions, organization of cytoskeleton and cell adhesion structures of epithelial and carcinoma cells, modulates levels of potassium channels in the membrane
↑	*DKC1*	Dyskerin Pseudouridine Synthase 1	Ribosome biogenesis and telomere maintenance (isoform 1); cell-cell and cell-substrate adhesion, increased proliferation, cytokeratin hyper-expression (isoform 3)
↑	*FEM1B*	Fem-1 Homolog B	Regulation of apoptosis, replication stress-induced checkpoint signaling (activation of CHEK1)
↑	*MORF4L2*	Mortality Factor 4 Like 2	Chromatin organization, DNA repair, oncogene and proto-oncogene mediated growth induction
↑	*SMAD1*	SMAD Family Member 1	Mediation of bone morphogenetic protein (BMP) signaling, regulation of growth, apoptosis, morphogenesis

Three of these ten genes should be highlighted. The first is *FEM1B*, as this gene was previously identified as a differentially expressed gene between LumA and LumB in an older RNA-Seq dataset of the same cells, which was discarded due to serious quality issues. Nevertheless, out of seven identified genes from this old dataset, *FEM1B* was the only one whose differential expression between LumA and LumB could later be validated by qPCR (T-test, p=0.011; Figure 4-36). Note that this is the only one of the suggested genes (see Table 4-20) that has so far been validated by qPCR. The second gene I would like to highlight is *CD6*, because - according to the Human Protein Atlas - it is a well-known favorable prognostic marker for BC which was down-regulated in the LumB subtype DCCs (see Table 4-21). In contrast to that, the third gene is *MORF4L2*, which is known as an unfavorable marker in BC according to the Human Protein Atlas and was up-regulated in the LumB DCCs (Table 4-21).

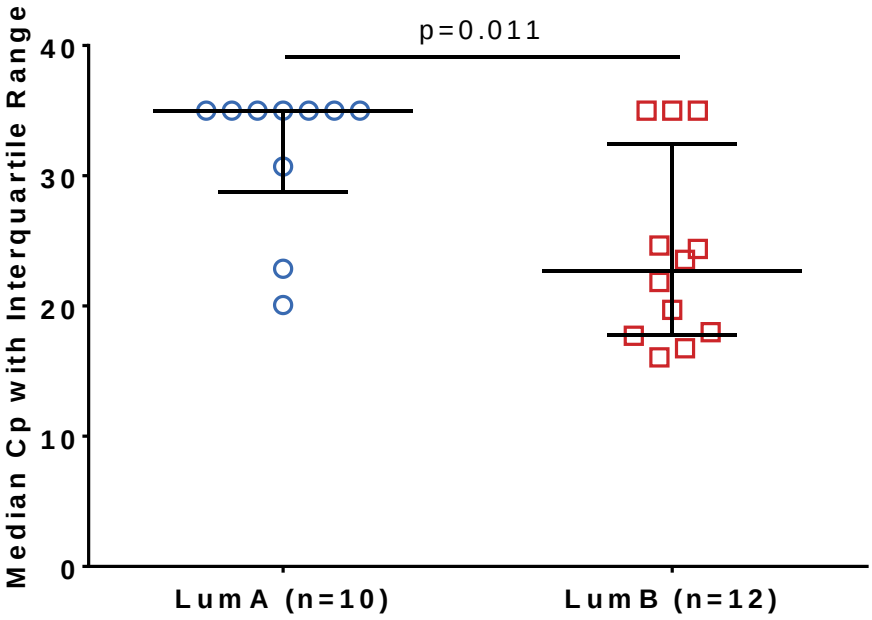


Figure 4-36 qPCR validation of expression of *FEM1B* in M0 DCCs. The graph shows the expression of *FEM1B* in M0 LumA and M0 LumB DCCs as median Cp values with interquartile range (represented by the whiskers). Statistics: T-test (chapter 3.16.1.2); p=0.011.

Table 4-21 Human Protein Atlas data on selected candidate genes. RNA expression given in the database was estimated relative to the other cancer entities tested for the same gene. The database also provided information on which cancer entities the gene of interest was found to be prognostic for. All ten genes had been assessed for their prognostic capacity across several cancer entities. fav = favorable marker, unfav = unfavorable marker

Up/Down regulated	Gene symbol	RNA expression in BC	Prognostic for
↓	*ANKRD13B*	Moderate	Renal (fav) and pancreatic (unfav) cancer
↓	*CD6*	Moderate	Breast cancer (fav)
↓	*IFNL2*	Very low	Not prognostic
↓	*FAT4*	Low	Renal cancer (fav)
↓	*TIRAP*	Moderate	Endometrial cancer (fav)
↑	*FEM1B*	High	Not prognostic
↑	*SMAD1*	High	Not prognostic
↑	*CTTN*	Moderate	Head and neck cancer/Liver cancer (unfav)
↑	*DKC1*	Moderate	Renal/liver/endometrial/head and neck cancer, melanoma (all unfav)
↑	*MORF4L2*	High	Breast cancer (unfav)

5. Results of method development for isolation of the miRNAome from single cells

In parallel to the work on the patient-derived DCCs (chapter 4), I conducted experiments to extend the established standard WTA protocol (sWTA, see chapter 3.2.1), in order to enable isolation of miRNAs and other non-coding RNAs (ncRNA) from single cells. This new extended WTA (eWTA) is aimed at isolation of miRNAs along with mRNA and gDNA from the same cell to obtain a more comprehensive image of the intrinsic state of each single cell at the time point of isolation. To this end, two additional steps are to be introduced into the protocol (see Figure 5-1): (1) an addition of blocking oligonucleotides complementary to the 3' end of the most abundant ribosomal RNAs (rRNA) to prevent their polyadenylation and (2) a poly(A) tailing (polyadenylation) step that will facilitate the capture of miRNAs together with mRNAs. The blocking of rRNAs is necessary, as rRNAs are highly abundant, making up as much as 80-90 % of a cell's RNA content (O'Neil et al., 2013) while not providing any valuable information on gene expression. The blocking oligonucleotides are added and annealed right after the protease digestion of the lysate, followed by the polyadenylation procedure. After the polyadenylation, the enzyme is deactivated and PNAs are added to enable the capture of miRNA together with the mRNA. Since the reverse transcription primers contain two random anchoring nucleotides causing reverse transcription to start directly at the beginning of the poly(A) tail, the additional polyadenylation of the mRNAs was not considered a problem for the downstream steps of the WTA.

The project is based on preliminary experiments done by Dr. Verena Lieb in close cooperation with Dr. Miodrag Guzvic, with whom I have also been working on this project. Dr. Lieb's work and some of the preliminary experiments on her most recent samples are briefly described in chapter 5.1. The preliminary data indicated that the implementation of the polyadenylation step required establishment of a novel buffer that enabled both proper cell lysis and functionality of the poly(A) polymerase (PAP), which is covered in chapter 5.2. After settling these issues, experiments to prove that the polyadenylation procedure was indeed working were performed (chapter 5.3). Following the proof of principle of the polyadenylation procedure, we investigated whether there was an increase in rRNA contamination due to the polyadenylation procedure (chapter 5.4). In the following, the effect of two types of rRNA blocking oligonucleotides on rRNA levels also examined (chapter 5.5). Lastly, the preliminary extended WTA (eWTA) protocol will be presented (chapter 5.6).

Due to the nature of method development, there were many failed attempts at all steps of the development process. Therefore, not every single experiment will be presented in detail. The main text will focus on the key experiments leading to the proposed protocol presented at the end. However, in order to prevent potential future researchers from repeating the failed experiments, all relevant experiments and the main results will be mentioned in the form of tables in each chapter. The individual experimental conditions are listed in separate tables in the corresponding methods chapters (see subchapters of chapter 3.15).

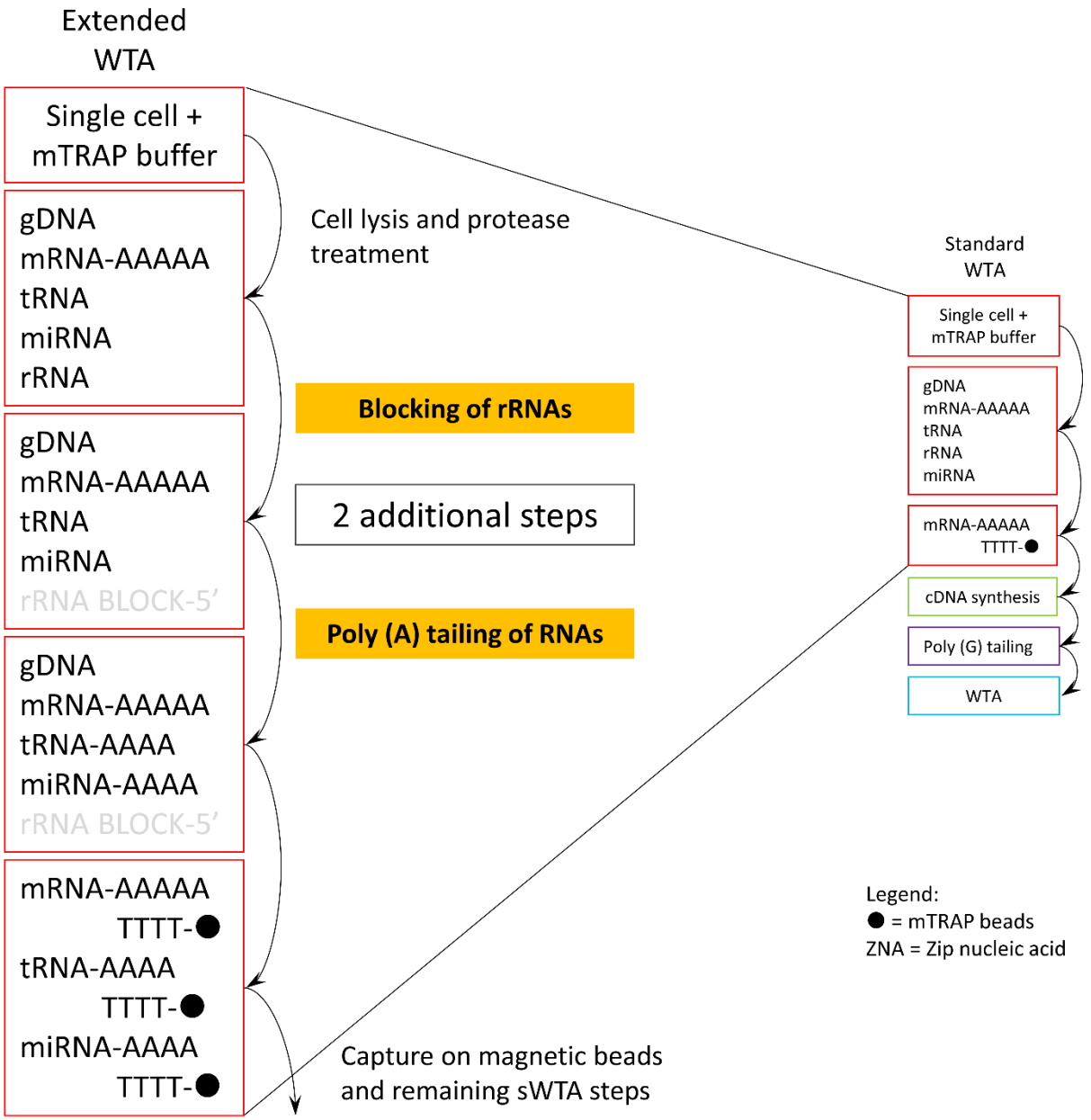


Figure 5-1 Schematic of the extended WTA protocol. The right column shows the main steps of the standard WTA protocol. The center of the illustration describes the newly introduced steps, while the left column highlights the changes in the early WTA steps between cell lysis and RNA capture.

5.1 Preliminary experiments

Dr. Verena Lieb was mainly working on the polyadenylation and blocking procedures. For this purpose, she performed numerous experimental WTAs using an *in vitro*-transcribed artificial RNA as her template (chapter 5.1.1). The sequence of the *in vitro*-transcribed RNA, called *Long Fragment* (*LF*), is provided in the appendix (chapter 12.3). After the best blocking oligonucleotides and annealing program for these oligonucleotides had been identified (chapter 5.1.1), she moved forward to experiments with single cells (chapter 5.1.2.). Since Dr. Lieb's work has not been published so far, I am unable to cite it in the following chapters. Afterwards, qPCR measurements targeting rRNA were conducted by me on some of Dr. Lieb's and also new samples (chapter 5.1.3).

5.1.1 Experiments with an *in vitro*-transcribed RNA template

Verena Lieb designed several variants of rRNA blocking oligonucleotides. First, she experimented with 40bp, 37bp, and 25bp DNA oligonucleotides, before also testing Zip nucleic acid (ZNA) oligonucleotides, which are DNA oligonucleotides conjugated to cationic spermine units. These units increase target affinity of the oligonucleotide by reducing electrostatic repulsion between the anionic DNA oligonucleotide and the *LF* RNA. All WTA experiments were conducted as described in chapter 3.15.1.1 to obtain amplified cDNA.

The conducted experiments are listed in Table 5-1. In the course of these experiments it was concluded that a 20 bp ZNA oligonucleotide with 5 spermine units provided the best blocking effect compared to DNA oligonucleotides. Next, the experiment was repeated with different concentrations of the ZNA oligonucleotide and it was observed that a blocking oligonucleotide concentration of roughly 1 million times the concentration of the target RNA combined with annealing of the oligonucleotide from 95 °C to 55 °C (Annealing95, detailed program in Table 3-17) provided the best results (Figure 5-2). Lower or higher concentrations from 10 times up to 15 million times the concentration of the template as well as annealing from 82 °C to 55 °C (Annealing82, detailed program in Table 3-18) resulted in worse performance (Table 5-1).

Table 5-1 Results of preliminary experiments with *in-vitro* transcript. Important variables and results are highlighted in bold. All experiments included appropriate negative controls (without template/without blocking oligonucleotides). After this step the standard WTA was performed starting from the reverse transcription. All experiments were conducted by Dr. Verena Lieb. For experimental details see Table 3-20. oligo = oligonucleotide, RT buffer = reverse transcription buffer

Experiment	Main variable under investigation	Result
WTA 1	Blocking: **40 bp DNA oligo, 20,000 x excess**, Annealing75 (see Table 3-19)	Endpoint PCR resulted in **weaker bands on the agarose gel when blocking oligos were added**, but no difference between various template amounts (PCR saturation); Sanger sequencing confirmed correct amplicon.
WTA 2	Blocking: 40 bp DNA oligo, **20/100/500/1000 x excess**, Annealing75 (see Table 3-19)	Endpoint PCR resulted in **no visible differences between blocking oligo concentrations** on gel.
WTA 3	Blocking: **40 bp and 37 bp DNA oligos**, $10^2/10^3/10^4/5x10^4/10^5$ x excess, Annealing82 (see Table 3-18)	Endpoint PCR resulted in **no visible differences between blocking oligos**, but showed concentration dependence for each oligo.
WTA 4	Blocking: **25 bp DNA oligo**, $10^3/10^4/2.5x10^4/5x10^4/7.5x10^4/10^5/5x10^5$ x excess, Annealing82 (see Table 3-18)	**Visible blocking effect of oligo at higher concentrations** according to endpoint PCR and agarose gel.
WTA 5	Blocking: **20 bp ZNA oligos**, $10^1/10^2/10^3/10^4/10^5/10^6$ x excess, Annealing82 (see Table 3-18)	Robust **reduction of band intensity starting from 10,000 x excess of the ZNA oligo.** Effect was strongest at the highest concentration.
WTA 6	Blocking: **25 bp DNA or 20 bp ZNA oligos**, $10^3/10^4/10^5/10^6$ x excess, Annealing75 (see Table 3-19)	Similar effect at 1,000x excess, but **ZNA oligos visibly more effective** from 10,000 x excess and above.
WTA 7	Blocking: **20 bp ZNA oligos**, **$10^6/2x10^6/10^7/1.15x10^7$ x** excess, Annealing75 (see Table 3-19)	**$10^6/2x10^6$ display equal effects**. Above $2x10^6$ x excess the band intensity on the agarose gel increased due to the high amount of oligo.
WTA 8	Blocking: 20 bp ZNA oligos, $10^4/10^5/10^6$ x excess combined either with **Annealing95 or 82** (see Table 3-17 and Table 3-18)	See Figure 5-2

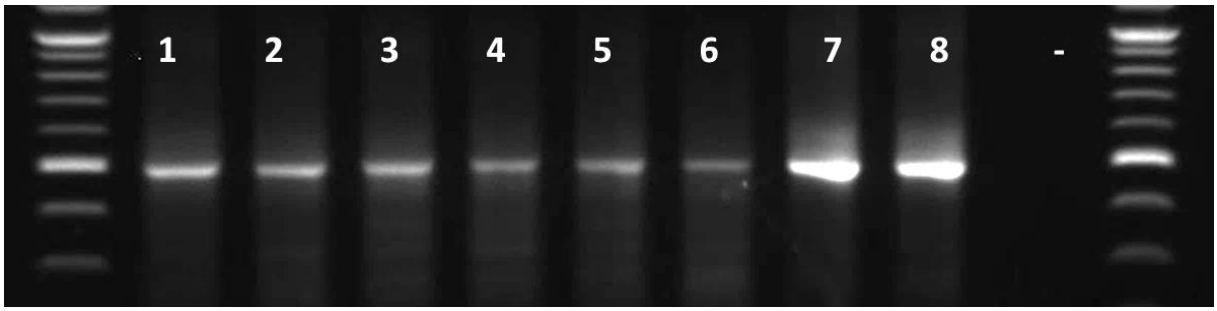


Figure 5-2 Agarose gel of experiment WTA 8 – ZNA oligonucleotide concentrations and different annealing. The image shows the *LF* amplicon bands resulting from the WTA 8 experiment (see Table 5-1). (1) 10^4 x, Annealing82; (2) 10^4 x, Annealing95, (3) 10^5 x, Annealing82; (4)10^5 x, Annealing95; (5)10^6 x, Annealing82; (6) 10^6 x, Annealing95; (7+8) controls without blocking oligonucleotide; (-) Negative control. Image by Verena Lieb

After these initial experiments with the *LF* RNA template, Verena Lieb moved on to SC material (chapter 5.1.2).

5.1.2 Single cell and total RNA experiments

For the experiments on SC material, Verena Lieb designed and established ZNA oligonucleotides for the *RNA28S (28S), RNA18S (18S), RNA5-8S (5.8S),* and *RNA5S (5S)* transcripts as well as PCR primers for the same rRNA transcripts (sequences are listed in Table 2-7). The main aims of these experiments were to test, (1) whether the novel blocking and polyadenylation steps were interfering with the WTA overall and (2) if rRNAs could still be detected after WTA. As a readout the regular WTA-QC PCR and in some cases endpoint PCR for the rRNAs was performed. Cells of the DU145 prostate cancer cell line were utilized as samples for the experiments. To test the impact of the additional steps, numerous experiments combining two different buffers (reverse transcription and PAP buffers) with three different PCR programs for the annealing of the block oligonucleotides and with two PCR programs for the polyadenylation were performed. Addition of several reagents like the RNase inhibitor SUPERase was also assessed. All relevant experiments are listed below (Table 5-2).

In the course of these experiments, it was discovered that the Annealing95 program (see Table 3-17), which had been the best program in the experiments using the *LF* RNA as a template (see Figure 5-2), was causing degradation of the RNA isolated from the cells, resulting in a lack of bands in the subsequent WTA-QC (see "WTA 13" in Table 5-2). In the end, it was determined – among several other variables of the protocol - that

- the Annealing75 program (see Table 3-19) was best for blocking in cells ("WTA 14")
- even low amounts of guanidinium thiocyanate (GTC) destroyed PAP ("WTA 12")
- PAP buffer supplemented with Igepal was working better than the reverse transcription buffer in the standard WTA ("WTA 11", sWTA with 1.5x upscaled volume)

Table 5-2 Preliminary experiments with single cells and total RNA. Important variables and results are highlighted in bold. All experiments included appropriate negative controls (without template/without blocking oligonucleotides). All experiments were conducted by Dr. Verena Lieb. For experimental details see Table 3-21. oligo = oligonucleotide, RT buffer = reverse transcription buffer

Experiment	Main variable under investigation	Result
WTA 9	Blocking: four 20 bp ZNA oligos, **$2.5x10^5/10^6$ x excess**	**Bad WTA quality** in most cases meaning only very few samples displayed more than one band in the WTA-QC, **rRNAs detectable in most samples** by endpoint PCR.
WTA 10	• Lysis buffer: RT buffer or PAP buffer	No bands visible on agarose gel at all.

Experiment	Main variable under investigation	Result
	• RNase inhibitor: **combinations of none, tRNA, and SUPERase** • Blocking: four 20 bp ZNA oligos, 10^6 x excess	
WTA 11	• Lysis buffer: RT buffer or PAP buffer • RNase inhibitor: combinations of none, **tRNA, and SUPERase** • **Blocking and polyadenylation skipped (sWTA procedure)**	Samples with **PAP buffer all displayed 3/3 bands in WTA-QC, therefore RNA is not degraded**, RT buffer samples only displayed 1/3 bands.
WTA 12	• Lysis buffer: control with mTRAP (established buffer of sWTA) included • RNase inhibitor: SUPERase or tRNA • Cell lysis program: **protease diluted in mTRAP for all samples**	RT buffer samples: 1/3 QC bands, PAP buffer samples: 0/3 bands. Addition of tRNA or SUPERase made no difference. Controls with mTRAP displayed 3/3 bands. **No rRNAs were detected** in any sample. **Even low amount of GTC in the mTRAP seems to destroy PAP.**
WTA 13	• Lysis buffer: PAP buffer or mTRAP with tRNA • Blocking: four 20 bp ZNA oligos, 10^6 x excess, **Annealing95 or no annealing** (see Table 3-17)	The **Annealing95 program is causing RNA degradation**, as samples which did not undergo the program displayed 3/3 bands on the gel, while those that underwent it showed no bands at all; no visible difference between mTRAP and PAP buffer.
WTA 14	• Blocking: four 20 bp ZNA oligos, 10^6 x excess, **Annealing75** (see Table 3-19) • Polyadenylation: 37 °C/**with or without 65 °C**/22 °C, 0.5 µl PAP in 10 µl total vol.	**Annealing 75 does not cause RNA degradation**, no difference between polyadenylation program with or without inactivation at 65 °C. All samples displayed three bands in the WTA-QC and rRNAs were also detected in all cases.
WTA 15	• Blocking: **no blocking oligos**, **Annealing75** (see Table 3-19)	All samples displayed **three bands in the WTA-QC and rRNAs were also detected in all cases**. Bioanalyzer profiles of samples prepared with PAP buffer were comparable to controls prepared with mTRAP.
WTA 16	• Blocking: **with and without four 20 bp ZNA oligos at 10^6 x excess**, Annealing75 (see Table 3-19)	**All samples displayed three** bands in the WTA-QC, but rRNAs were not quantified.

Ultimately, a protocol was assembled that was able to consistently produce samples with three bands in the WTA QC, but it remained unclear whether the rRNA blocking worked in this setting. Therefore, following Dr. Lieb's experiments, we decided that the next step was to quantify rRNAs by qPCR in samples produced by her latest protocols (experiments "WTA 14-16" see Table 5-2), in order to examine whether the blocking oligonucleotides were working in single cells.

5.1.3 qPCR analysis of preliminary eWTA experiments

First, I conducted a standard curve experiment as described in chapter 3.8.3 to assess the specificity and efficiency of the *28S, 18S, 5.8S,* and *5S* rRNA primers in the qPCR. The results are shown in the appendix (chapter 12.4.1).

Subsequently, the primers were used to measure the expression of the four rRNAs in the combined samples from experiment "WTA 14-16" (see Table 5-2 "qPCR 1"). The results suggested that the blocking was working (two-way ANOVA, $p=0.0012$ for the treatment variable, $p<0.0001$ for transcript variable), however the best combination of the blocking and polyadenylation programs was not distinguishable (Figure 5-3). This may have been caused by the low sample number (n=2 per combination). Interestingly, the eWTA protocol seemed to cause a tendency towards an increase in rRNA levels compared to the sWTA.

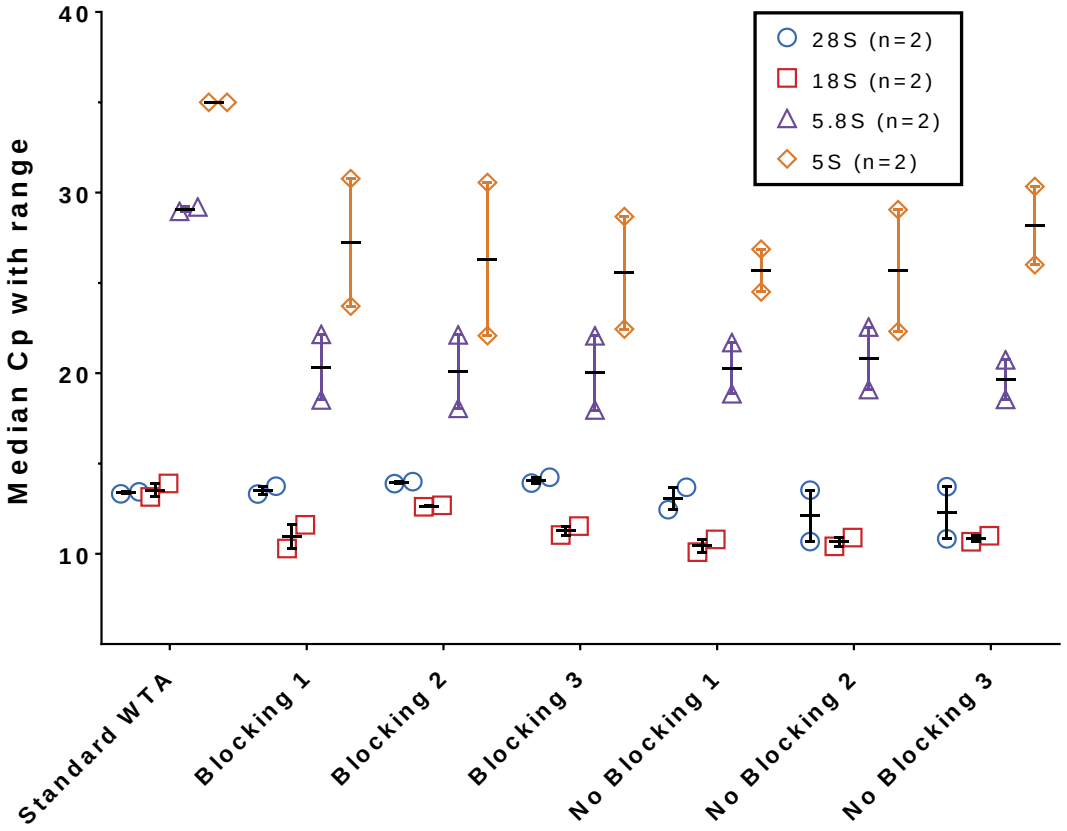


Figure 5-3 Quantification of rRNA levels in preliminary WTAs. The plot illustrates the expression levels of the *28S, 18S, 5.8S,* and *5S* rRNA transcripts across the combined samples of Verena Lieb's "WTA 14-16" experiments. Treatment groups: three main categories (sWTA versus eWTA with or without blocking of rRNAs). The eWTA categories are divided into three subgroups: (1) no annealing program/polyadenylation without PAP inactivation at 65 °C, (2) no annealing program and polyadenylation with PAP inactivation at 65 °C, (3) Annealing75 program and polyadenylation without PAP inactivation at 65 °C. Statistics: two-way ANOVA (chapter 3.16.1.2). All "blocking" and "no blocking" treatments were significantly different from the standard WTA for the *5.8S* and *5S* rRNAs at $p<0.05$ or $p<0.01$, however, the significances were not included in the graph to avoid cluttering. There were no significant differences between any of the blocking or no blocking treatments.

Due to the low sample numbers, Dr. Lieb's eWTA experiment "WTA 16" was replicated using single cell equivalents (SCE, see chapter 3.15.1.3 for details). SCEs were used to avoid the inherent transcriptional noise of SCs. However, the qPCR results of the new samples were very similar to the previous ones (see Figure 5-3 and Table 5-3 "WTA 17"). Furthermore, the PAP buffer treated groups again contained more rRNAs than the sWTA performed with the established mTRAP buffer. Therefore, our conclusion was that the PAP buffer with 0.1 % Igepal, which had been used up to that point, was not working as intended. Most likely, the cell lysis did not work properly.

To settle this issue, we decided to focus on the comparison of the established mTRAP buffer and the PAP buffer. A standard WTA was performed (as described in chapter 3.2.1) either with mTRAP or with PAP buffer with 0.1% Igepal, using ten SCEs per group. The *28S* and *18S* rRNAs were increased in the WTAs prepared with the PAP buffer compared to mTRAP buffer, however these differences were not significant due to several dropouts across the transcripts indicated by a Cp value of 35 (multiple T-tests, Figure 5-4). However, no sample dropped out completely (all four rRNAs lost), but each sample lost only one, two, or three (in one sample) of the examined transcripts. Nevertheless, the remaining transcripts were present at high amounts. This may indicate incomplete lysis and random loss of transcripts before amplification.

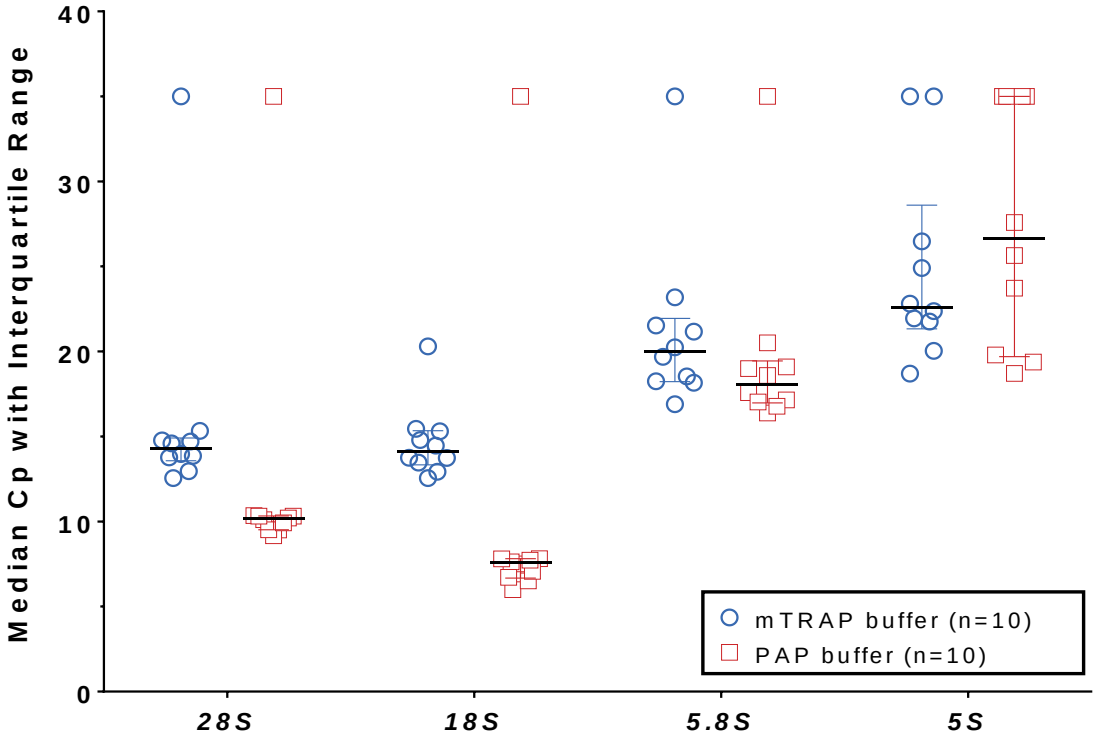


Figure 5-4 Levels of rRNAs in sWTAs prepared with mTRAP or PAP buffer. The graph illustrates the expression of four rRNAs in WTAs prepared either with the established mTRAP buffer (blue circles) or with PAP buffer (red squares). For each buffer, ten single cells were used. The qPCR data are shown as median Cp values with interquartile range (represented by whiskers). Statistics: multiple T-tests (chapter 3.16.1.2). No significant differences due to dropouts.

Based on the data presented so far, we concluded that the PAP buffer was likely unsuitable for the eWTA and needed to be replaced with another buffer that enabled a better cell lysis and did not lead to an increase of rRNA in absence of polyadenylation. The search for a better buffer and the necessary changes to the WTA protocol coming with it are covered in the next chapter (5.2).

Table 5-3 eWTAs and qPCR measurements on samples from Dr. Lieb's preliminary experiments. All experiments included appropriate positive and negative controls.

Experiment	Reaction conditions	Result
qPCR 1	• Samples: WTA 14-16 (see Table 5-2) • Targets: 28S, 18S, 5.8S, and 5S rRNAs • By combining the samples of the three experiments, there were two samples each for six combinations of blocking and polyadenylation	See Figure 5-3
WTA 17	• Repetition of "WTA 16" (see Table 5-2) • The only difference was the use of SCEs instead of SCs to reduce the biological variation; generation of SCEs is explained in chapter 3.15.1.2 • qPCR Targets: *28S, 18S, 5.8S*, and *5S* rRNAs	Similar result to qPCR 1, the group undergoing sWTA treatment with PAP buffer contained more rRNA than the samples undergoing sWTA with mTRAP (reference)
WTA 18	• Template: DU145 SCEs • Standard WTA with mTRAP and tRNA compared to standard WTA with PAP buffer supplemented with ATP, 10 % Igepal, and SUPERase	See Figure 5-4

5.2 Modifications of the lysis buffer and cell lysis procedure

As the preliminary data suggested that the cell lysis was not working well with the PAP buffer despite the addition of Igepal (see Figure 5-4), it was necessary to establish a better buffer for the eWTA, because the established mTRAP lysis buffer was too harsh for the PAP to function (see "WTA 12" in Table 5-2). Once the best buffer conditions had been identified (chapter 5.2.1), it was also necessary to adapt the lysis procedure, in order to improve the performance of the protocol (chapter 5.2.2).

5.2.1 Identification of optimal buffer conditions

The mTRAP buffer contains the potent protein denaturant guanidinium thiocyanate (GTC) along with proprietary ionic detergents, which further enhance protein denaturation. These properties are necessary to properly lyse cells, to free RNAs from protein complexes, and to inactivate RNases. In contrast, the PAP buffer itself contained no detergents or chaotropic agents like GTC. Although the detergent Igepal was added to the PAP buffer, the preliminary results suggested that this was still insufficient to make the protocol work consistently (see Figure 5-4). Since Dr. Lieb's data also indicated that even low amounts of GTC in the polyadenylation reaction destroyed the PAP (see "WTA 12" in Table 5-2), a custom buffer called miRNA isolation buffer (MIB) was developed (chapter 5.2.1.1). In the course of this development process, we realized that PAP was surprisingly resilient to denaturing conditions. Therefore, we decided to test the polyadenylation procedure in diluted mTRAP buffer as well, which worked exceptionally well (5.2.1.2). The implementation of a dilution required an adaptation of the lysis procedure (chapter 5.2.2).

5.2.1.1 Custom miRNA isolation buffer

The new custom buffer had to be able to facilitate both cell lysis and PAP functionality, therefore it was necessary to find a compromise between the mTRAP buffer, which is optimized for cell lysis and RNase inhibition, and the PAP buffer, which is optimized for PAP activity. The compositions of both commercially available buffers are shown in Table 5-4.

Table 5-4 Compositions of commercial mTRAP and Poly(A) Polymerase buffer.

mTRAP (stock)	10x PAP buffer (stock)
200 mM Tris (pH 7.5)	500mM Tris (pH 7.9)
200 mM NaCl	2500mM NaCl
25mM $MgCl_2$	100mM $MgCl_2$
500 mM GTC	
+ proprietary detergents (ionic and non-ionic)	

In order to avoid PAP denaturation by GTC, our idea was to replace the GTC found in the mTRAP buffer with a milder chaotropic reagent in combination with a milder detergent. Urea was chosen as a replacement for GTC, because urea is widely used as a protein denaturant for gel electrophoresis (Floyd et al., 1974) while at the same time the ammonium ion (NH_4^+) in the urea molecule is listed as a weaker chaotropic agent than the guanidinium ion in the Hofmeister series (Hofmeister, 1888). We decided to test whether a buffer containing urea would work, despite some concerns about additional problems caused by urea's temperature-dependent degradation (will be discussed in more detail at the end of this chapter). Regarding the detergent, several widely used non-ionic as well ionic ones were tested. Additionally, tRNA in the reaction was meant

to be avoided, because tRNA can be polyadenylated and would consequently be captured along with the miRNA. Therefore, the tRNA, which is normally used to coat the inside of the reaction tube and protect mRNA from RNases, was replaced with the SUPERase RNase inhibitor.

First, a wide variety of buffers were tested regarding their ability to properly lyse cells (buffer preparation is described in chapter 3.15.2.1.1). SUPERase was not included in these lysis experiments, because it is expensive and would not have contributed to the cell lysis. I then evaluated how fast each buffer lysed the cells and what the process looked like compared to the mTRAP buffer, because the aim was to find a buffer that was as close to mTRAP as possible (see chapter 3.15.2.1.2 for description of the procedure).

In mTRAP buffer, the cells were initially contracting before bursting (Figure 5-5, compare panels a and b) and the first cells started bursting after about 10 sec, while the majority of cells began after around 60 sec (Figure 5-5c). After 90 sec most of the cells had burst and the visibly perforated cell membranes remained on the slide until the end of the observation (Figure 5-5d). Unlike the mTRAP buffer, most of the MIB variants that contained urea did not cause the cells to burst, but they homogenously dissolved the cells at different speeds depending on urea and detergent concentration. Using this assay, each of the detergents was screened at various concentrations and in combination with different urea concentrations (see Table 5-5). Finally, urea with either sodium dodecyl sulfate (SDS), sodium deoxycholate (SDC), or N-lauroylsarcosine (NLS) – which are all anionic detergents – were identified as the best combinations, because the lysis process looked similar to the bursting that was observed in mTRAP buffer (see Figure 5-6), only a lot faster. For example, the MIB with 8 M urea and 0.03 % NLS lysed all cells within 48 sec.

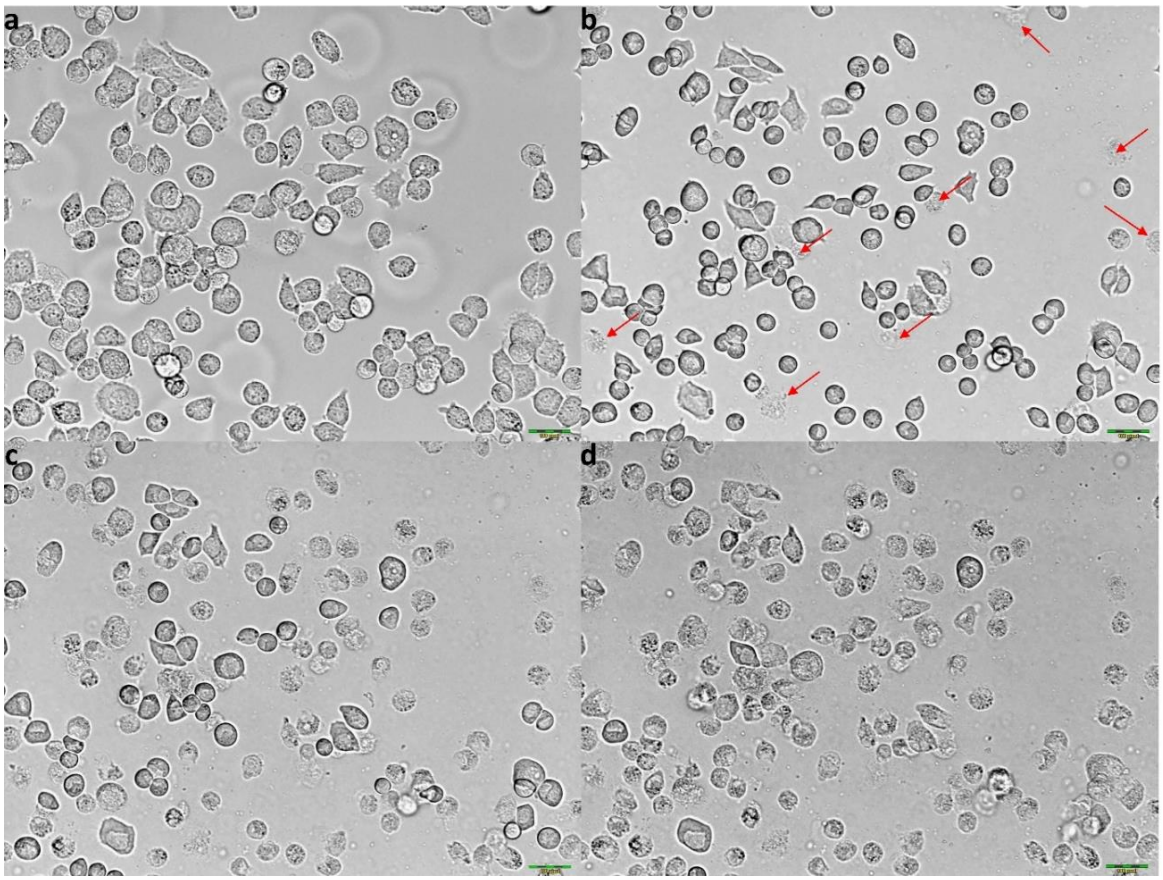


Figure 5-5 Overview of cell lysis by mTRAP buffer over a 90 sec time course. The images show the same field of view over a time course of 90 sec. The displayed DU145 cells were incubated in 4 °C cold mTRAP lysis buffer with the slide being at RT. (a) 0 sec timepoint, (b) 30 sec timepoint, (c) 60 sec timepoint, (d) 90 sec timepoint. Red arrows in panel (b) indicate bursting cells. Bursting cells in panels (c-d) are not indicated by arrows due to their high abundance. Panel (b) illustrates how the cells contracted upon addition of the lysis buffer

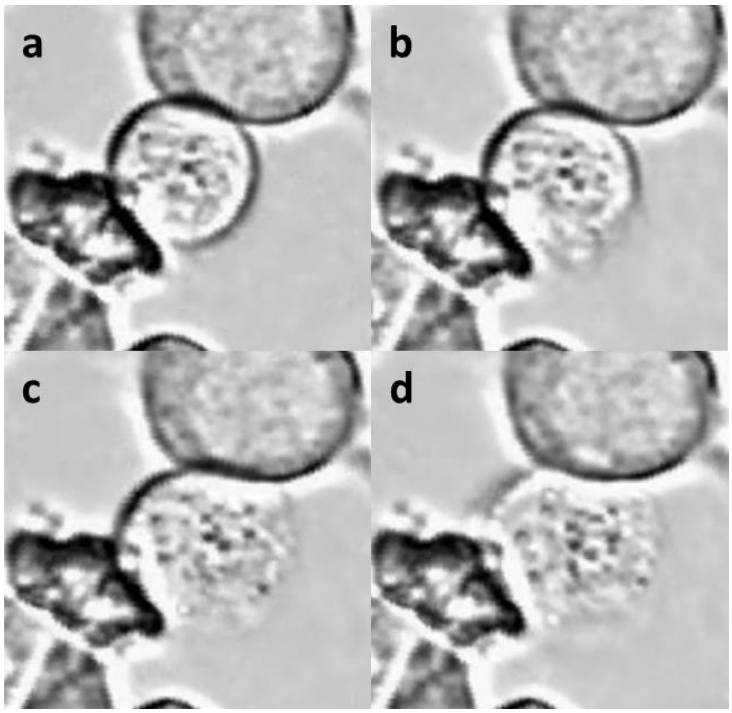


Figure 5-6 Single DU145 cell bursting in mTRAP lysis buffer. The image sequence illustrates the bursting of a single cell (central cell). The whole process took about 2sec. (a) Cell before bursting, (b) cell membrane breaks at the bottom side of the cell, (c) the cytosol is partly released, (d) the dead shell of the cell remains on the slide.

Next to the previously described lysis experiments, I also compared some of the most promising MIB variants with the mTRAP buffer in a standard WTA using SCEs. QPCR was performed on several mRNA transcripts to investigate whether their expression was comparable when different buffers were used. Overall, a combination of urea and either of the ionic detergents SDS, SDC, and NLS worked best. Since SDS performed worst of the three ionic detergents, it was excluded first (Table 5-5). Although SDC performed slightly better than NLS in the WTA, it had to be excluded as well, because the literature indicated that it possesses DNA-damaging properties (Merritt and Donaldson, 2009), which we wanted to avoid since one of the main points of the WTA protocol is that it also enables isolation of gDNA from each cell through the supernatant WGA protocol.

Table 5-5 Lysis experiment results and corresponding WTA data. The table shows relative lysis speed and expression levels of tested mRNAs in the WTAs which were evaluated in comparison to the mTRAP buffer. Each given combination of reagents was also tested with different concentrations of each component. Each of the detergents was tested at concentrations of 0.01 %, 0.05 %, 0.1 %, 0.5 %, and 1 %, while urea was tested at 0.5 M, 1 M, 2 M, 3 M, 4 M, 6 M, and 8 M. To keep the table concise, combinations of different detergent and urea concentrations are not included. Here, "MIB" represents the constant part of the buffer consisting of 200 mM Tris and 200 mM NaCl. The pH was adjusted to 7.5 unless indicated otherwise. SDS: sodium dodecyl sulfate, SDC: sodium deoxycholate, NLS: N-lauroylsarcosine

Buffer	Lysis speed	WTA mRNA level
mTRAP	Reference (~90s)	Reference
MIB + Igepal pH 7.5 or 7.9	Slightly faster	Decrease
MIB + GTC	No effect on cells	Not tested
MIB + GTC + Igepal	Fast	Equal
FCP buffer (commercial, Qiagen)	Not tested	Complete loss
MIB + urea (various amounts) + Igepal	Very fast	Slight decrease
MIB + urea + Tween 20	Slow	Decrease
MIB + urea + Triton X-100	Fast	Slight decrease
MIB + urea+SDS with/without Igepal	Very Fast (3rd fastest)	Equal (3rd best)
MIB + urea+SDC with/without Igepal	Very Fast (2nd fastest)	Equal (best)
MIB + urea+NLS with/without Igepal	Very Fast (fastest)	Equal (2nd best)

Due to concerns that urea from the WTA might interfere with the WGA from supernatants, a WGA was conducted as described in chapter 3.2.2 on some supernatants from WTAs prepared with MIB containing 1M/4M/8M urea and 0.5% Igepal each. The WGA QC revealed that the WGA had worked as usual and that the samples prepared from urea-containing MIB even tended to have more bands on the agarose gel than the mTRAP reference (Figure 5-7).

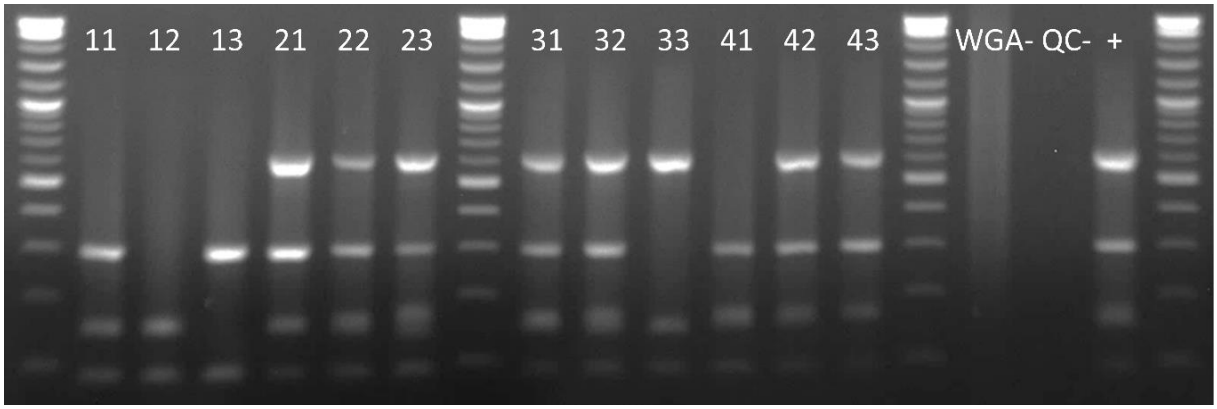


Figure 5-7 WGA-QC gel of supernatant WGA products made from eWTA samples. The image shows an agarose gel resulting from a WGA-QC PCR of eWTA supernatant WGA products. The displayed samples are biological replicates of (11-13) control group treated with mTRAP buffer, (21-23) group treated with MIB containing 1M urea, (31-33) group treated with MIB containing 4M urea, (41-43) group treated with MIB containing 8M urea. (WGA-) negative control of the WTA/WGA procedure, (QC-) negative control of the WGA-QC, and (+) positive control of the WGA-QC.

Please note that from this point, whenever concentrations of chemicals in the MIB are discussed, the concentrations relate to the level at the polyadenylation step. This means that the concentrations are initially higher at lysis, because the cell lysate will be diluted throughout the eWTA procedure due to addition of reagents. For example, MIB with 8 M urea and 0.03 % NLS during lysis roughly translates to 3 M urea and 0.01 % NLS at polyadenylation.

So far, the MIB was only tested regarding the lysis of cells. Next, a series of WTA experiments was conducted with the *LF* transcript as template to narrow down the best buffer composition for PAP functionality (list of experiments in Table 5-6). First, a reduction of the polyadenylation temperature was assessed, because of urea's tendency to degrade above 30 °C (see final paragraph of this chapter for details). The experiment showed that an increase of the incubation time from 30 min to 50 min together with a decrease in temperature from 37 °C to 30 °C resulted in the same yield of the *LF* transcript (see "PAP activity 1" in Table 5-6).

Subsequently, a similar experiment was conducted, in which PAP activity was assessed in MIB with different urea concentrations (see "PAP activity 2" in Table 5-6). The result was surprising: the lower concentrations of 1 M and 2 M resulted in a strong reduction of enzyme activity, while the high concentrations of 4M and 8 M permitted enzyme activity comparable to the control reaction with PAP buffer (see Experiment 1 in Figure 5-8). Due to this finding, the experiment was replicated two more times with more conditions to increase resolution (see "PAP activity 3" in Table 5-6). Experiment 2 ("PAP activity 3") was in agreement with Experiment 1 ("PAP activity 2") up to 4M of urea, which provided the same expression level as the PAP buffer in both cases, despite the absolute Cp values being different (Figure 5-8). At 6 M and 8 M urea the results worsened in Experiment 2. In contrast, Experiment 3 ("PAP activity 3" replicate) suggested that all urea concentrations were equal (Figure 5-8), however, this experiment was problematic, as half of the samples had to be discarded due to contamination. Therefore, the whole experiment may not be reliable due to the low number of biological replicates per treatment group.

A two-way ANOVA analysis on Experiments 2 and 3 (Experiment 1 excluded due to missing of some datapoints) revealed a strong effect of treatment ($p=0.0004$) and of experiment ($p<0.0001$).

However, the post-hoc test only showed significances between the PAP buffer (control) and the "MIB 0.5M urea" as well as "No PAP" control group of Experiment 2. There were no significant differences in Experiment 3. Furthermore, the control group that did not receive PAP also displayed surprisingly low Cp values in Experiment 2 and 3 indicating an unwanted passive carry-over of the target transcript to the final amplification step of the WTA.

After careful evaluation of the data – mainly considering Experiment 1 and 2 - we decided to conduct further experiments with the urea concentration of 3 M. An additional reason for this choice was that 8 M of urea during lysis, which is required to end up with 3 M urea during polyadenylation, is still feasible. However, more than 8 M is difficult to achieve, because the urea will not dissolve properly anymore.

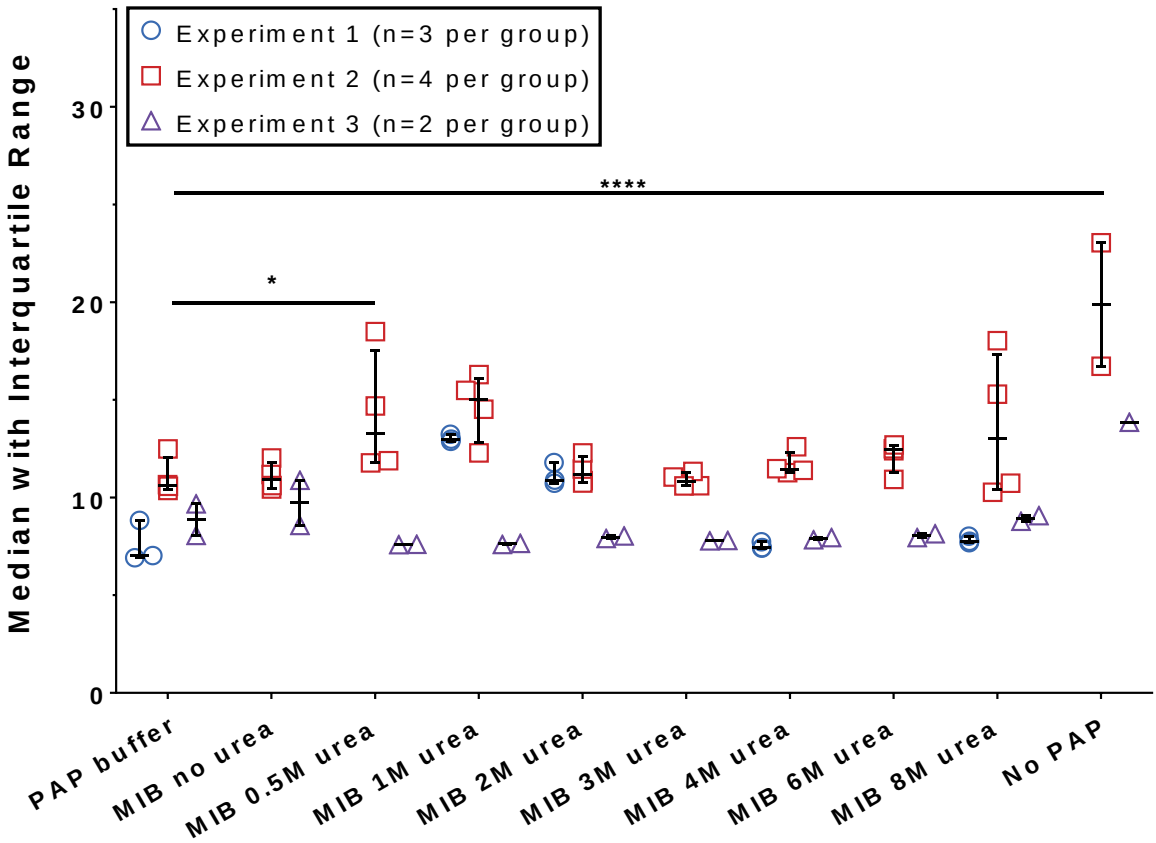


Figure 5-8 Activity of PAP in MIB with various concentrations of urea. The plot depicts the combined data of the experiments "PAP activity 2 "(Experiment 1) and "PAP activity 3" (Experiments 2+3, see Table 5-6) as median Cp values of the *LF* transcript expression level. Experiment 1 did not include the "MIB no urea", "MIB 0.5M urea", "MIB 3M urea", "MIB 4M urea", and "No PAP" groups, therefore no datapoints are displayed. In Experiment 3 the samples from one day (Experiments 2 and 3 were conducted over two days due to high sample numbers) had to be discarded due to contamination. Therefore, only two biological replicates are shown. Statistics: two-way ANOVA (chapter 3.16.1.2); **** $p<0.0001$, * $p<0.05$

Based on the previous experiment, we decided to move on and include NLS in the MIB as well to identify a final buffer composition. For this purpose, a urea concentration of 3 M was combined with different concentrations of NLS followed by assessment of PAP activity in each case (see "PAP activity 4" in Table 5-6). The data showed that an NLS concentration of 0.05 % resulted in an activity of the enzyme that was almost equal to the PAP buffer.

Afterwards, 2 M and 3 M urea were tested with an NLS concentration of 0.01 % in comparison to the PAP buffer (see "PAP activity 5" in Table 5-6). The data revealed that the two MIB variants were significantly different from each other (one-way ANOVA with multiple comparisons $p=0.03$), but neither of them was different from the PAP buffer (Figure 5-9). The "No PAP" control had to

be excluded from statistical analysis, because it only comprised a single sample. Despite the fact that he MIB with 2 M urea and NLS displayed a slightly lower Cp and may likely provide a better performance of the enzyme, we decided to continue testing the MIB with 3 M urea and 0.01 % NLS, because the higher urea concentration will provide a better cell lysis and RNase inhibition.

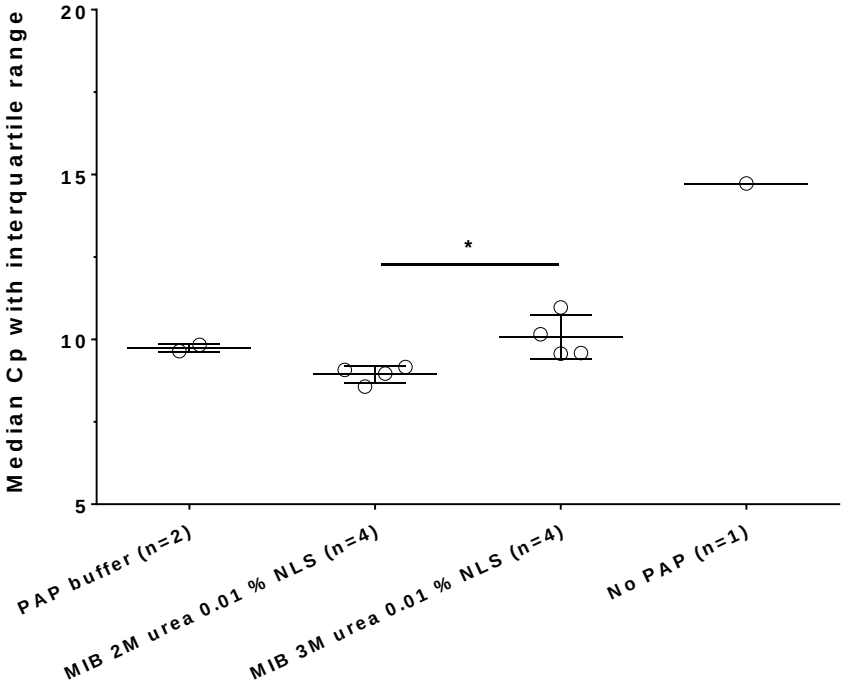


Figure 5-9 Activity of PAP in MIB with 0.01 % NLS and 2 M or 3 M urea. The plot illustrates the expression level of the *LF* transcript after reactions in PAP buffer or MIB with urea and NLS at indicated concentrations. The data are displayed as median Cp values with interquartile range (whiskers). Statistics: one-way ANOVA (chapter 3.16.1.2); "No PAP" control excluded from analysis, because there was only one sample. * $p<0.05$

Lastly, an sWTA experiment was conducted to compare the preferred MIB variant (3 M urea and 0.01 % NLS) with the established mTRAP buffer (see "WTA 19" in Table 5-6). Five SCEs and one cell pool were used per buffer condition, in order to examine whether this variation of the MIB was comparable to the established mTRAP buffer in the sWTA. The data revealed no significant differences between the groups, neither in the SCEs (Figure 5-10a; multiple T-tests) nor in the cell pools (Figure 5-10b; no test possible, only one datapoint per group). In the SCEs, the relative expressions varied rather strongly in cases of *EHMT2* and *REEP5*, with the expression level of the former appearing increased in MIB, while the level of the latter was decreased. However, as mentioned before, none of the changes were significant and therefore likely caused by noise inherent to the SCs used in the experiment.

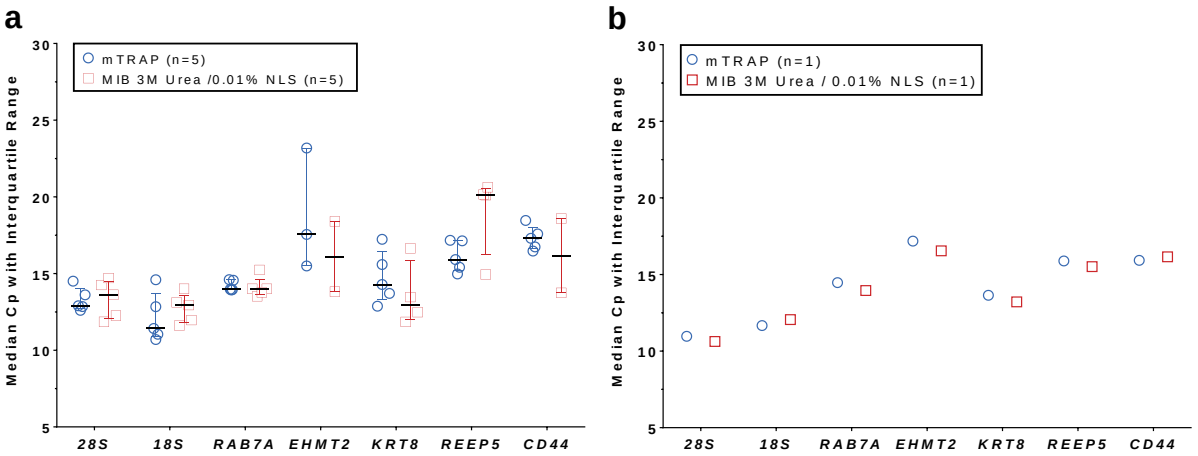


Figure 5-10 Comparison of final MIB and mTRAP buffer in SCs and pools. The graph shows the expression of several transcripts in sWTAs prepared either with the established mTRAP buffer or with the final MIB containing 3M urea and 0.01% NLS from experiment "WTA 19" (see Table 5-6). For each buffer, five single cells and one cell pool were prepared. The qPCR data are shown as median Cp values with interquartile range (whiskers). (a) Single cell Cp values, (b) Cell pool Cp value. Statistics: multiple T-tests (chapter 3.16.1.2); no significant results.

Taking together the presented experiments, we concluded that the MIB with 3M urea and 0.01% NLS (concentration at polyadenylation step) was fulfilling its task sufficiently well when compared to the mTRAP buffer. However, we knew from the beginning of the experiments that the presence of urea in the buffer posed additional problems that would have required more optimizations, but at the time when the experiments were planned, urea was the most promising alternative to the mTRAP buffer. The issue with urea is that it exists in equilibrium with ammonium cyanate ions and the rate at which these ions form increases with temperature (Hagel et al., 1971). These ions can modify amino acids of proteins through a carbamylation reaction (Stark et al., 1960), thereby potentially impairing or changing their function, which implies that high temperatures should be avoided in a protocol using urea when PAP is supposed to function in the solution. For that reason, the polyadenylation temperature was reduced to 30 °C (see experiment "PAP activity 1" in Table 5-6). Since one of the previous experiments had revealed that PAP can tolerate urea unexpectedly well (see Figure 5-8), we decided to also try other options for buffers. Upon re-examination of the preliminary data we noticed that the only experiment that indicated denaturation of PAP by GTC was "WTA 12" (see Table 5-2). However, in this experiment all samples had been subjected to the Annealing95 program, which later turned out to be the reason for RNA degradation (see "WTA 13" in Table 5-2). Unfortunately, there were no further investigations into the effect of GTC on PAP afterwards. Consequently, we decided to investigate PAP activity in different dilutions of the GTC-containing mTRAP buffer in hope of being able to avoid further protocol optimizations to accommodate the urea-containing MIB (chapter 5.2.1.2).

Table 5-6 WTA experiments testing the functionality of the custom buffer. Important parameters under investigation are highlighted in bold. All experiments included appropriate positive and negative controls. For experimental details see Table 3-22.

Experiment	Main variables under investigation	Result
PAP activity 1	Polyadenylation programs: • 37 °C 30 min (pos. control) • **30 °C 30/40/50 min** • 30 °C 50 min without PAP (neg. control)	**Tailing for 50 min at 30 °C resulted in equal amount of transcript being detected**, shorter incubation time provided a lower yield. Transcript was also detectable in the neg. control group, albeit at a far lower level.
PAP activity 2	• PAP buffer (pos. control/reference) • **MIB with 1/2/4/8 M urea**	See Figure 5-8
PAP activity 3	• PAP buffer (pos. control/reference) • **MIB with 0/0.5/1/2/3/4/6/8 M urea** • PAP buffer without PAP (neg. control)	See Figure 5-8 This experiment was conducted twice and represents an extension of "PAP activity 2"

Experiment	Main variables under investigation	Result
PAP activity 4	• PAP buffer (pos. control/reference) • **MIB with 3 M urea and 0.1/0.5/0.05 % of NLS** • PAP buffer without PAP (neg. control)	**Activity of PAP** was reduced when treated with 0.1 % or 0.5 % NLS, but **similar to control level with 0.05 % NLS**. The control without PAP also displayed robustly detectable levels of the transcript, albeit at a higher Cp (lower concentration). This indicated an undesired **passive carry-over of material.**
PAP activity 5	• PAP buffer (pos. control/reference) • **MIB with 2/3 M urea and 0.1 % of NLS** • PAP buffer without PAP (neg. control)	Both **MIB variants perform similarly to the PAP buffer**, but were significantly different from each other. See Figure 5-9
WTA 19	• Buffer: **mTRAP with tRNA as control, MIB with 3 M urea and 0.01 % NLS** • Treatment: standard WTA	**MIB with 3 M urea and 0.01 % NLS is equal to mTRAP** in the sWTA See Figure 5-10

5.2.1.2 Diluted mTRAP buffer

In parallel to the last few MIB experiments, I investigated the activity of PAP in diluted mTRAP buffer. The rationale behind these experiments was to dilute the mTRAP-containing cell lysate enough for PAP to function, which would enable polyadenylation without having to completely change the buffer for the eWTA, thereby also saving time and resources on the optimization of the protocol.

The activity of PAP in mTRAP was examined through several titration series (detailed protocol for this type of experiments in chapter 3.15.2.2, see Table 5-7 for individual experimental conditions and results). The first three experiments revealed that a dilution of mTRAP by 1:5 resulted in PAP performance comparable to the PAP buffer and that dilutions beyond this did not improve the activity (see “PAP activity 6”, “PAP activity 7”, and “PAP activity 8” in Table 5-7). This dilution corresponded to a GTC concentration of 44 mM in the polyadenylation reaction in this experimental setting.

Lastly, a fourth titration series was tested (see “PAP activity 9” in Table 5-7). The qPCR data revealed that, while the undiluted mTRAP displayed Cp values of the *LF* similar to those in the control without PAP, a 1:3 dilution was already similar (one-way ANOVA, overall $p<0.0001$ for *LF* expression) to the PAP buffer and that the 1:5 dilution of mTRAP led to Cp values even slightly lower than in the reference group (Figure 5-11a). Looking at the relative expression (Figure 5-11b), only the 1:5 dilution resulted in a capture rate of the *LF* higher than the reference. The whole experiment was replicated independently a second time and the results were similar.

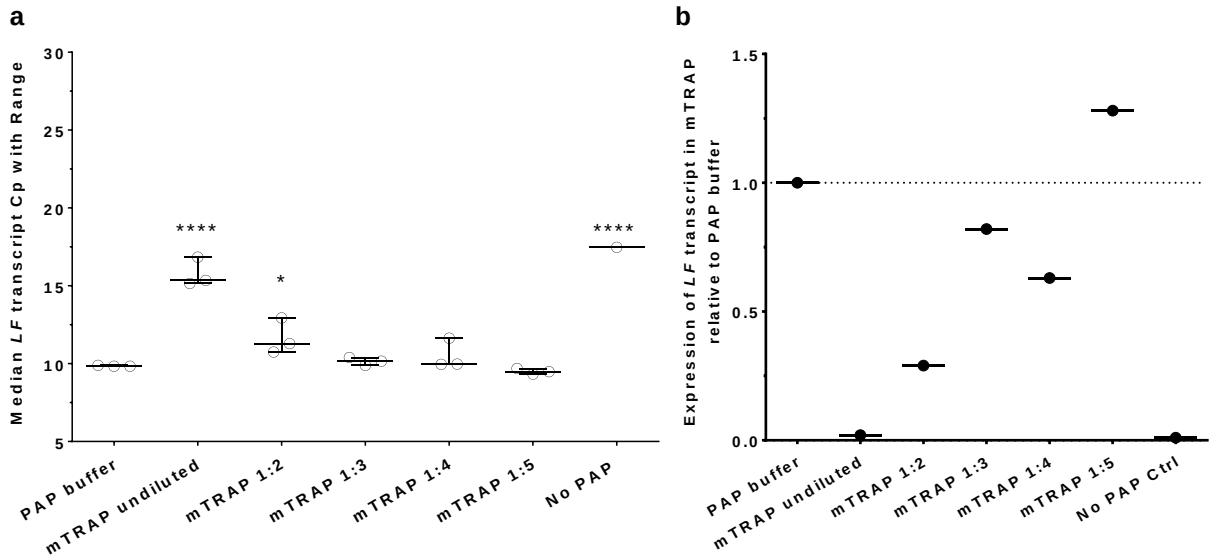


Figure 5-11 Activity of poly(A) polymerase in different dilutions of mTRAP buffer. The graph shows the levels of the *LF* transcript in PAP buffer (reference), undiluted mTRAP, different dilutions of mTRAP, and a control of mTRAP without PAP. Expression levels were measured by qPCR. Each treatment was performed in triplicates except for the control without PAP ("No PAP"). The qPCR data are shown as (a) median Cp values with range (whiskers) or (b) expression relative to the PAP buffer control. Statistics: one-way ANOVA (chapter 3.16.1.2); **** $p<0.0001$ * $p<0.05$

Table 5-7 Activity of PAP in diluted mTRAP buffer. Important parameters under investigation are highlighted in bold. All experiments included appropriate positive and negative controls. All experiments used the artificial *LF* RNA as a template. For experimental details see Table 3-23.

Experiment	Reaction conditions	Result
PAP activity 6	• PAP buffer (pos. control/reference) • **mTRAP undiluted** • **mTRAP 1:5/1:50/1:500/1:5000** • PAP buffer without PAP (neg. control) • Polyadenylation: 30 °C for **30 min**, 0.5 µl PAP in 10 µl total vol.	All **mTRAP dilutions (starting at 1:5) were equal to or slightly better than the PAP** buffer. The **mTRAP stock prevented PAP function completely** and was equal to the control without PAP. Polyadenylation was shortened to 30 min to prevent potential saturation of the reaction.
PAP activity 7	• PAP buffer (pos. control/reference) • **mTRAP undiluted** • **mTRAP 1:2.5/1:5/1:10/1:20** • PAP buffer without PAP (neg. control) • Polyadenylation: 30 °C for **20** min, 0.5 µl PAP in 10 µl total vol.	mTRAP 1:2.5 worked, but activity was reduced. **PAP activity was again equal to control levels starting from 1:5 dilution.** Polyadenylation was shortened even more to 20 min to prevent saturation of the reaction. Experiment replicated once with identical results.
PAP activity 8	• PAP buffer (pos. control/reference) • **mTRAP 1:5/1:6/1:7/1:8/1:9/1:10** • PAP buffer without PAP (neg. control)	Results identical to experiment "PAP activity 6" above. All dilutions were equally good. Experiment replicated once with identical results.
PAP activity 9	• PAP buffer (pos. control/reference) • **mTRAP undiluted** • **mTRAP 1:2/1:3/1:4/1:5** • PAP buffer without PAP (neg. control)	See Figure 5-11 Experiment replicated once with identical results.

Based on the results of the presented experiments, we decided to incorporate the 1:5 dilution of the mTRAP-containing cell lysate into the eWTA protocol to perform the polyadenylation and to discard the MIB, thereby saving time and resources on buffer optimization. In order to integrate the dilution step in an economical way, a small change to the original WTA protocol was necessary (chapter 5.2.2).

5.2.2 Modification of the lysis procedure

In order to save reagents, we tried reducing the lysis buffer used for picking from four to three microliters, because the required volume for the 1:5 dilution of the mTRAP-containing cell lysate depended mainly on the initial amount of mTRAP. The experiment was conducted as described in chapter 3.15.2.3. The experiment was conducted a second time, gene expression was measured by qPCR, and the Cp values of both experiments were normalized using the included calibrator sample. Subsequently, the data of both experiments were pooled for a combined analysis.

The data revealed that the expression levels of the *28S, 18S,* and *5.8S* rRNAs, as well as the mRNAs *RAB7A, EHMT2, KRT8, REEP5,* and *CD44,* all of which were measured by qPCR, were very similar between the two treatment groups (Figure 5-12, multiple T-tests, no significant results). The results of the two independent replicate experiments were the same.

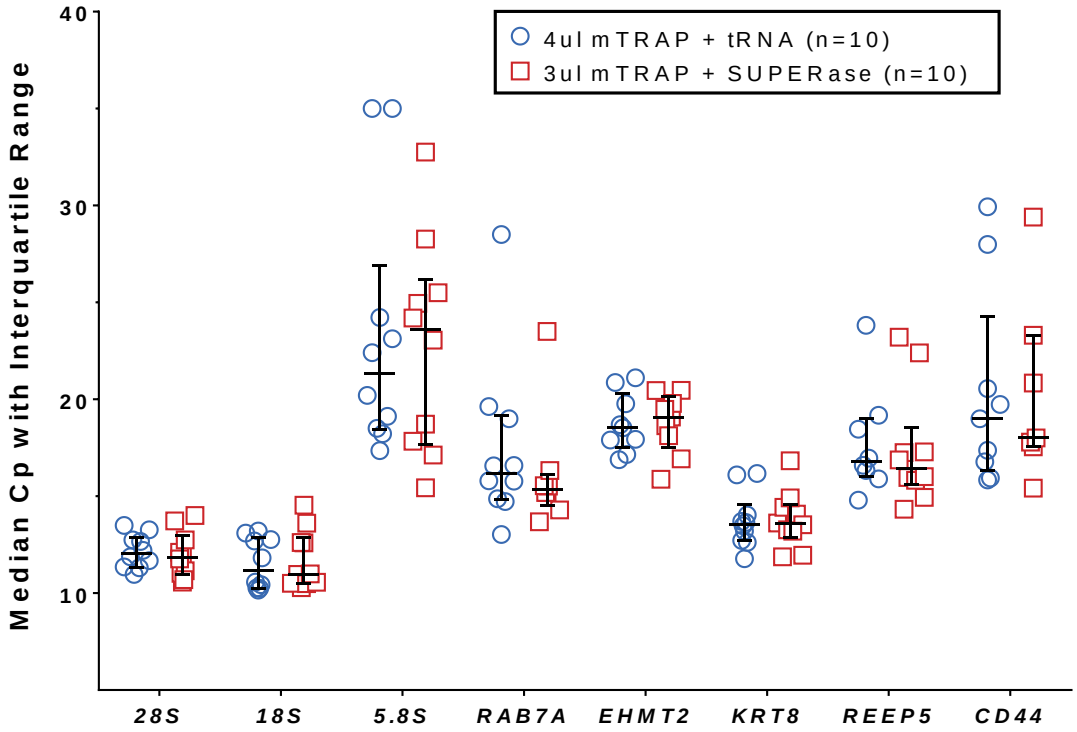


Figure 5-12 Expression of rRNAs and mRNAs in reduced mTRAP volume with SUPERase. The graph illustrates the expression of three rRNAs and five mRNAs in WTAs prepared with cells lysed in 4 µl mTRAP with tRNA or 3 µl mTRAP with SUPERase. The qPCR data are shown as median Cp values with interquartile range (whiskers). Statistics: multiple T-tests (chapter 3.16.1.2); no significant differences.

Based on the results of this experiment, all following eWTAs were performed with the reduced mTRAP volume. With a suitable buffer established, we aimed to prove that the poly(A) tailing procedure was working in principle (chapter 3.15.3).

5.3 Proof-of-principle of polyadenylation

Having identified the necessary buffer conditions for both cell lysis and polyadenylation (see chapter 5.2), the stage was set for a proof-of-principle experiment to investigate whether the polyadenylation was working as intended. For this purpose, an eWTA was carried out according to chapter 3.15.3. This experiment was conducted independently a total of three times, in order to replicate the results. The individual replicate results are displayed in addition to the combination of the data, in order to highlight the high variability between the replicates.

The data revealed a significant effect of the polyadenylation in two out of three individual replicates (Figure 5-13a+b; multiple T-tests). However, the *LF* RNA spike-in was only significantly different in the first replicate experiment (Figure 5-13a), although there was a strong tendency towards an increased level of the spike-in transcript in the second experiment as well (Figure 5-13b), but it was not significant. Here, only the *28S* and *18S* rRNAs were significantly affected by the procedure, but not the *LF* (Figure 5-13b). In the third replicate experiment (Figure 5-13c) there was a tendency towards lower Cp values for the *28S* and *18S* rRNAs, but they were not significant. After assessment of the individual replicates, the Cp values of the three individual experiments were normalized using the included calibrator sample and the data were pooled for a combined analysis. The joint analysis indicated a robust effect of the polyadenylation on the *28S* and *18S* rRNAs as well as on the artificial spike-in RNA (Figure 5-13d; multiple T-tests, FDR<0.001 and p=0.0005 for the *LF* transcript). The mRNAs displayed no significant differences in any of the datasets. Unsurprisingly, the *LF* spike-in was detectable in a few samples even without polyadenylation (Figure 5-13b+c), a phenomenon that had been observed before in controls without PAP (see Figure 5-9 and Figure 5-11). Surprisingly, there were also cases in which the *LF* spike-in was undetectable in the polyadenylation group (Figure 5-13b+c). This high variability of the results between the replicates and the surprisingly high levels of rRNAs in the non-tailed control group indicated that the eWTA procedure was not functioning in an optimal fashion yet.

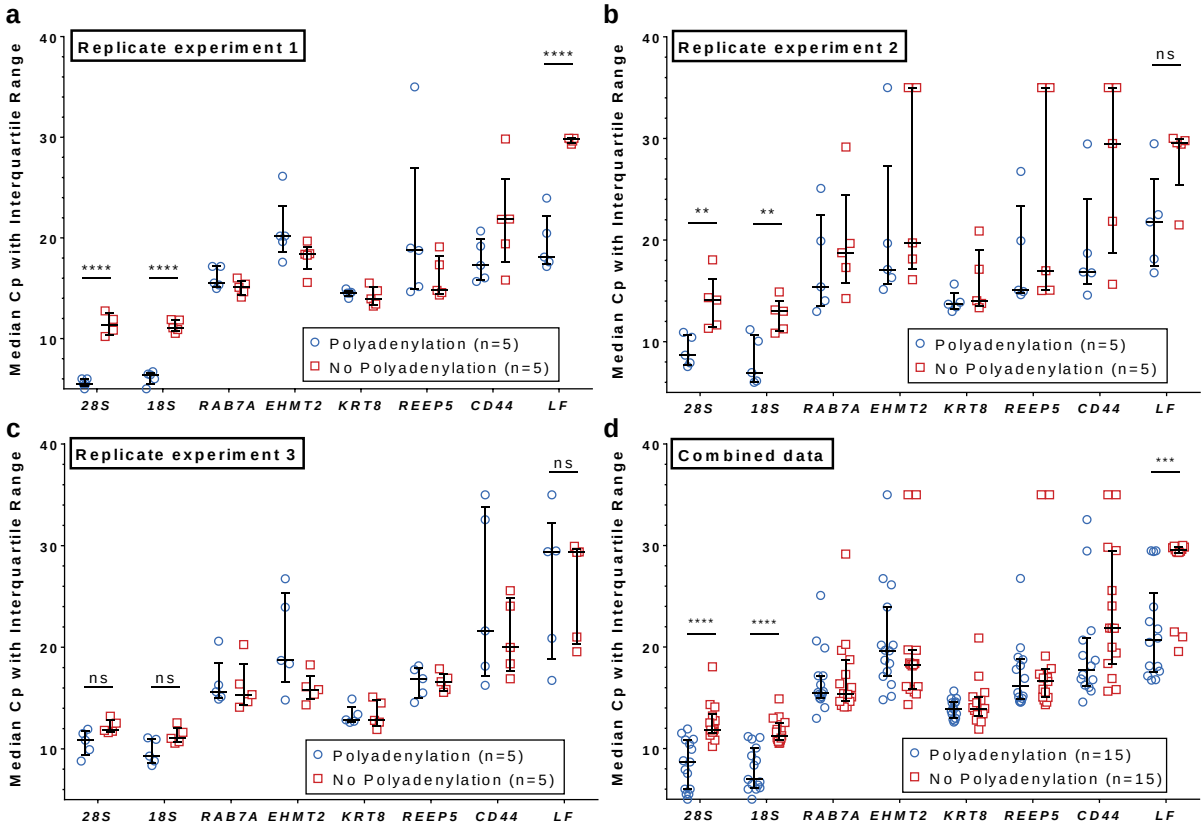


Figure 5-13 Effect of polyadenylation on *in vitro* transcript levels - SCEs dispensed before protease lysis. The graphs illustrate the expression of different transcripts and the levels of the *LF* transcript after the eWTA procedure. The qPCR data are shown as median Cp values with interquartile range (whiskers). (a-c) Individual replicate experiments 1-3, (d) combination of the three replicates from a-c. Statistics: multiple T-tests (chapter 3.16.1.2); **** FDR<0.0001, *** FDR<0.001, ** FDR<0.01, ns = not significant; FDR = false discovery rate

Since we observed a high variation of the Cp values between the biological replicates, we decided to repeat the experiment described above with a modification to the SCE generation procedure. So far, the picked cell pools were mixed and dispensed to individual tubes before the protease lysis step, meaning each sample was undergoing its own, separate lysis (experiment displayed in

Figure 5-13). After the change, the whole cell pool was lysed together in a single tube with the reaction volume upscaled to the number of contained cells. The SCEs were then dispensed to individual tubes afterwards (experiment displayed in Figure 5-14). This change was introduced to make the starting material for the eWTA more homogenous and thereby reduce the variation inherent to single cell amounts of RNA. This experiment was also conducted a total of three times.

Interestingly, we still observed quite a large variability in the data, despite the cells being separated into SCEs after the lysis (Figure 5-14). Moreover, there was only a significant difference in the *28S* rRNA in one of the three replicates (multiple T-tests, Figure 5-14c), while there were no differences in any other transcript in any of the individual replicate experiments (Figure 5-14a and b). Like before, the data of the replicate experiments were then normalized and analyzed together. As a result, the data revealed significant differences in the *28S* and *18S* rRNA levels between the polyadenylation and control groups, but not in the level of the *LF* spike-in transcript (p=0.047, insignificant due to FDR correction). Again, rRNAs were highly abundant in the control group without polyadenylation. Similarly, the spike-in was also detected in the control group, while it was undetectable in several samples of the polyadenylation group indicating problems with the eWTA procedure.

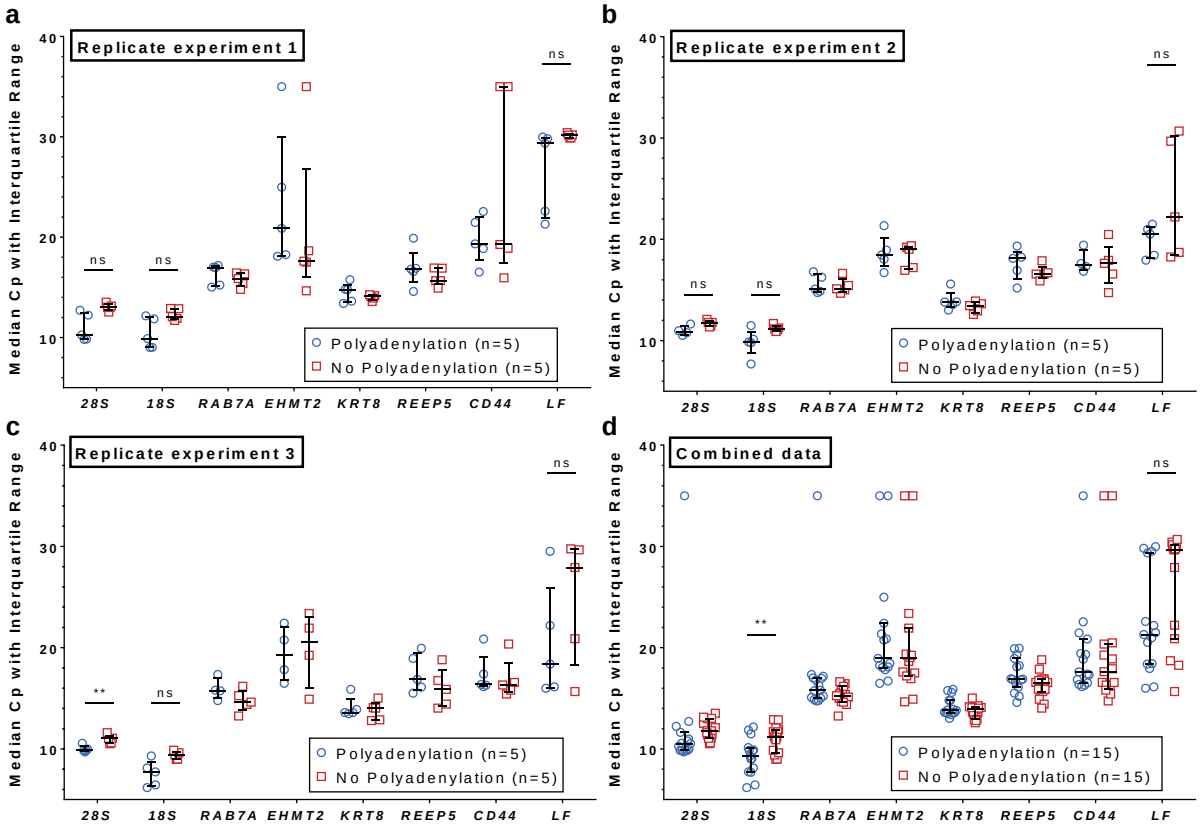


Figure 5-14 Effect of Poly(A) tailing on *in vitro* transcript levels - SCEs dispensed after protease lysis. The graphs illustrate the expression of different genes and the levels of the *LF* transcript after the eWTA procedure. Data are shown as median Cp values with interquartile range (whiskers). (a-c) Individual replicate experiments 1-3, (d) combination of the three replicates from a-c. Statistics: multiple T-tests (chapter 3.16.1.2); ** FDR<0.001, ns = not significant

The previous data have shown that the polyadenylation works in principle, albeit not in every single sample or experiment. They have also consistently shown that rRNAs were abundant in WTA products, both in the standard and eWTA, because they displayed far lower Cp values than any of the mRNAs (e.g. Figure 5-13, Figure 5-14). Furthermore, we learned that the experimental results are better when SCEs are dispensed before the protease lysis and undergo the lysis separately in small volumes (compared Figure 5-13 and Figure 5-14).

Therefore, we decided to look more closely at the observed rRNA contamination and potential ways to reduce it (chapter 5.4).

5.4 Investigation of rRNA contamination

All previous experiments performed with cell-derived RNA have consistently shown that rRNAs, especially *28S* and *18S* rRNAs, were indeed depleted but still present in WTAs without polyadenylation (sWTA) at high amounts, despite the included poly(A)-based selection step (for example comparing both groups in the sWTAs shown in Figure 5-12 or "No Polyadenylation" group in Figure 5-13). Therefore, we wanted to investigate why this happened and – more importantly – tested several different approaches to reduce the levels of rRNAs in the eWTA. To do this, I first compared rRNA levels of WTAs from DU145 cells with WTAs of patient-derived cells (chapter 5.4.1). Second, numerous different changes of the existing WTA protocol were explored to reduce the rRNA contamination (chapter 5.4.2). Lastly, targeted depletion of rRNAs was also investigated (chapter 5.4.3).

5.4.1 rRNA levels in patient RNA-Seq data

Since rRNA concentrations had only been quantified in sWTAs (see Figure 5-12) performed with cultured DU145 cells until this point, we were interested in the rRNA levels in patient cells processed with the sWTA to examine how the different sources of cells compared to each other. For this purpose, three $EpCAM^+$ and three $EpCAM^-$ cells with high-quality cDNA were randomly selected from the BC collective. Additionally, six $EpCAM^+$ cells were chosen from our chair's prostate cancer (PC) collective. The PC cells were isolated and processed in the same way as the BC cells. Furthermore, 13 control samples (sWTA treatment) of previous eWTA experiments done on DU145 cells, were added to the analysis for comparison. In these samples, the rRNAs had already been quantified. Unfortunately, the qPCR negative controls for the *5.8S* and *5S* rRNAs were contaminated in two and one out of three experiments, respectively. Therefore, the *5.8S* and *5S* rRNA data were only available from five and nine samples, respectively.

First, $EpCAM^+$ and $EpCAM^-$ control cells were compared to see whether these two BM-derived populations were distinct. However, there were no significant differences in the expression levels of rRNAs between those two types of cells (multiple T-tests, Figure 5-15).

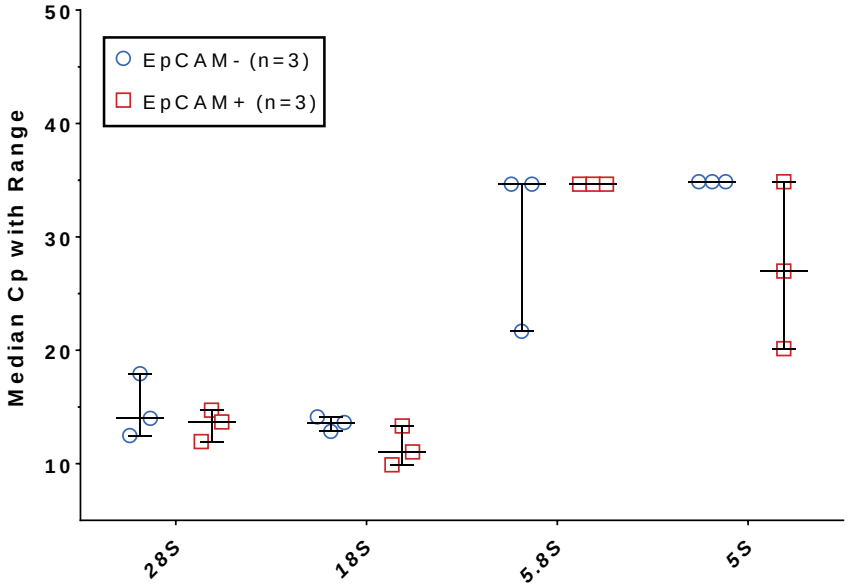


Figure 5-15 rRNA levels in EpCAM⁺ and EpCAM⁻ cells from the BM of BC patients. The plot illustrates the expression levels of four rRNA species in three randomly selected high quality EpCAM⁺ and three EpCAM⁻ cells. Data are displayed as median Cp values with range (whiskers). Statistics: multiple T-tests (chapter 3.16.1.2). There were no significant differences.

Next, I compared the PC, BC, and cultured DU145 cells. For this purpose, the EpCAM⁺ and EpCAM⁻ BC cells were pooled, as no difference had been observed between the two populations (see Figure 5-15). The two-way ANOVA uncovered a highly significant overall difference between the source materials ($p<0.0001$). The post-hoc test revealed significant differences in the generally less abundant *5.8S* and *5S* rRNAs (Figure 5-16). The highly abundant *28S* and *18S* rRNAs were not significantly different, however they displayed a tendency towards the cultured cells containing more of the rRNAs than the patient-derived cells.

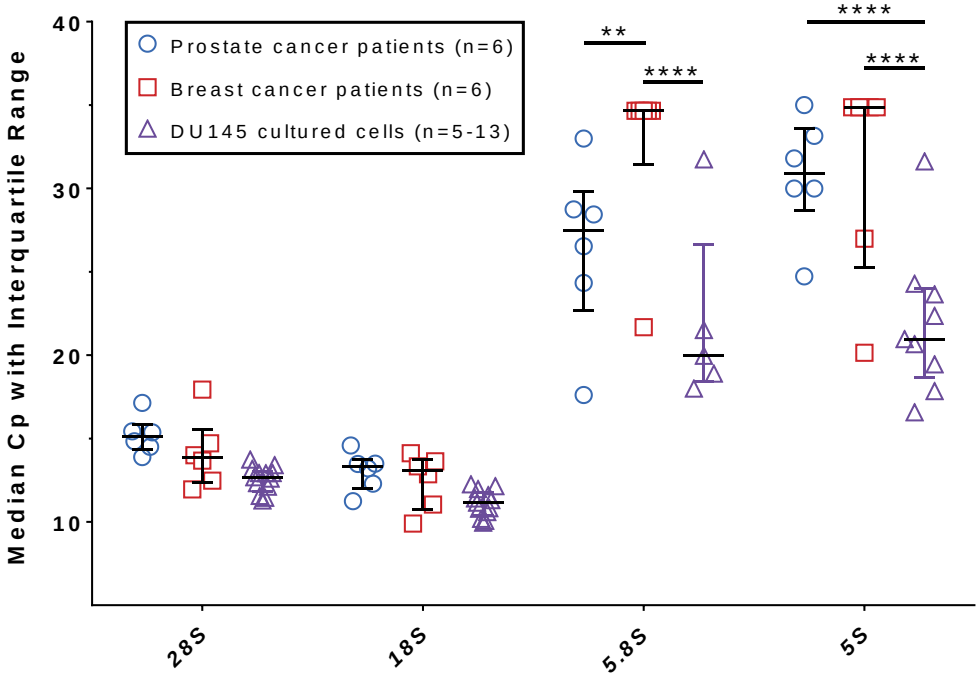


Figure 5-16 Comparison of rRNA levels in breast and prostate cancer patients as well as DU145 cells. The figure depicts the levels of four rRNA species in cells derived from PC patients, BC patients (EpCAM⁺ and EpCAM⁻ cells together), and DU145 cultured cells. Data are shown as median Cp values with interquartile range (whiskers). Statistics: two-way ANOVA (chapter 3.16.1.2); **** $p<0.0001$, ** $p<0.01$

In addition, the three groups were also examined in a pairwise manner by performing multiple T-test analysis on the same data. In this case, no significant differences were found between the PC and BC cells, but there were highly significant differences between PC and DU145 cells and BC and DU145 cells, respectively (Table 5-8). These data supported the idea that there was a tendency towards higher rRNA levels in the cultured cells compared to patient-derived cells.

Table 5-8 P-values of multiple T-test analysis on PC- and BC-derived cells versus DU145 cultured cells. Additional analysis of the data shown in Figure 5-16. Significant discoveries are marked with an asterisk (*).

rRNA species	BC versus PC	PC versus DU145	BC versus DU145
28S	0.3	1.18e-005*	0.026*
18S	0.5	0.0004*	0.021*
5.8S	0.07	0.21	0.011*
5S	0.91	0.0013*	0.005*

In order to assess whether the high rRNA levels detected by qPCR also appeared in other downstream applications, I examined the *28S* and *18S* rRNA levels in the sequencing data of the 45 cells that underwent RNA-Seq (see chapter 4.5.1) and correlated the qPCR results of the same cells with the percentage of reads mapping to rRNA genes. Analysis revealed a moderate correlation of qPCR and RNA-Seq for both *28S* and *18S* rRNA species (Spearman correlation, p=0.0012 and p=0.0004, respectively; Figure 5-17). The mean with standard deviation and median percentages of rRNA reads were 4.06±0.05 % and 1.48 %, respectively, suggesting a rather low rRNA contamination in the sequenced WTA samples.

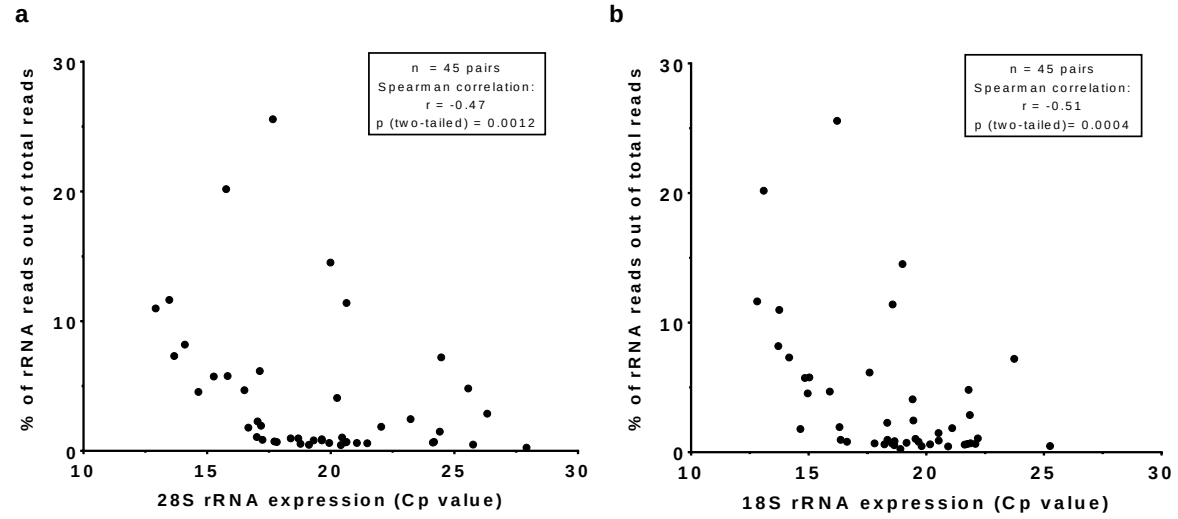


Figure 5-17 Correlation of *28S* and *18S* rRNA levels from qPCR and RNA-Seq experiments. The plots show the correlation of the expression of (a) *28S* rRNA and (b) *18S* rRNA levels measured by qPCR (x-axis) and by deep RNA-Seq (y-axis). Statistics: Spearman correlation (chapter 3.16.1.2); results of analyses are provided in the boxes in each panel.

Together, these data suggested that usage of DU145 cells led to an overestimation of rRNA contamination, because the cultured cells were not fully representative of the patient-derived cells. However, the result that the cell lines were not exactly the same as patient-derived material was expected. Additionally, the data showed that – despite the robust correlation of rRNA levels measured by qPCR and RNA-Seq - the seemingly high rRNA levels detected by qPCR were not problematic with regard to sequencing, as only a minority of cells contained >10% of rRNA-mapped reads, which is often regarded as a quality threshold by bioinformaticians. Encouragingly, the mean and median percentages of rRNA-mapping reads were rather low. Nevertheless, a variety of approaches to reduce the rRNA contamination was tested, in order to further improve the eWTA (chapter 5.4.2).

5.4.2 Modification of existing protocol to reduce rRNA contamination

In order to reduce the passive carry-over of rRNAs that seemed to occur despite the poly(A)-based selection and washing procedure (see Figure 5-12 or Figure 5-13), numerous different parameters with potential influence on the suspected sticking of rRNAs to either the beads or the mRNAs were changed. As none of the approaches showed any positive effect on mRNA and rRNA levels measured by qPCR, the experiments will not be described in detail.

The following changes were individually introduced into the standard WTA (see chapter 3.2.1) and the levels of rRNAs and several mRNAs were assessed to evaluate the effect compared to the standard WTA (Table 5-9):

Table 5-9 Modified parameters to reduce rRNA contamination. Each parameter was the only modified parameter per sample. There were no combination treatments of parameters from different rows of the table. Changes were assessed by qPCR targeting both rRNAs and mRNAs. All treatments were tested in DU145 single cell equivalents.

Parameter	Description of changes	Result
Amount of mTRAP beads	Tested volumes: 2 µl or 8 µl	No difference
Blocking of beads	With tRNA, different polymers (Ficoll, Polyvinylpyrrolidon, Heparin, Dextran sulfate), or DNA (human Cot-1 DNA, salmon sperm DNA, herring sperm DNA or a combination of all three)	No significant difference, but DNA seemed to be contaminated with rRNA
Cell lysate volume	Dilution of the cell lysate with mTRAP buffer (volume increased 3x) before addition of 8 µl beads	No difference
Bead capture of mRNA	Incubation time of sample with beads was shortened to 25 min	No difference
Wash buffer volume	cDNA-Tween and cDNA-Igepal volume doubled	No difference
Bead pre-washing	Pre-washing of beads and resuspension in cDNA-Igepal wash buffer before addition to samples	No difference
Re-isolation of mRNAs	Samples already bound to beads were heated to 94 °C for 4 min to release mRNAs. Then the mRNA-containing supernatant was diluted with 60 µl cDNA-Igepal or mTRAP and transferred to a fresh tube. The remaining beads were washed twice with cDNA-Igepal, the supernatant was removed and combined with the previously removed supernatant in the fresh tube. Finally, the complete supernatant was added to fresh beads and incubated again at RT to bind them. Afterwards, the WTA was continued normally.	No significant difference, but tendency to increasd rRNA amount and loss of mRNA
Bead capture temperature and novel pre-heating	Incubation of sample with beads performed at 42 °C or 50 °C (instead of RT) for 45 min. These treatments were both done with or without pre-heating of samples (after addition of beads but before 45 min incubation period) to 75 °C for 1 min	Samples with 50 °C incubation and pre-heating showed tendency towards overall loss of transcripts, no effect on the remaining samples
Wash buffer temperature	cDNA-Igepal wash buffer warmed to 44 °C for the first washing step after reverse transcription	No difference
PNAs	WTA performed without addition of PNAs	MRNA was completely lost without PNAs. There is no specificity problem with the PNAs. See Figure 5-18
Lysis and wash buffer composition	Addition of isostabilizing agents (betaine, dimethyl sulfoxide, tetramethylammonium chloride, tetraemethylammonium chloride) to mTRAP and wash buffers in different combinations and with or without Triton X-100	No difference

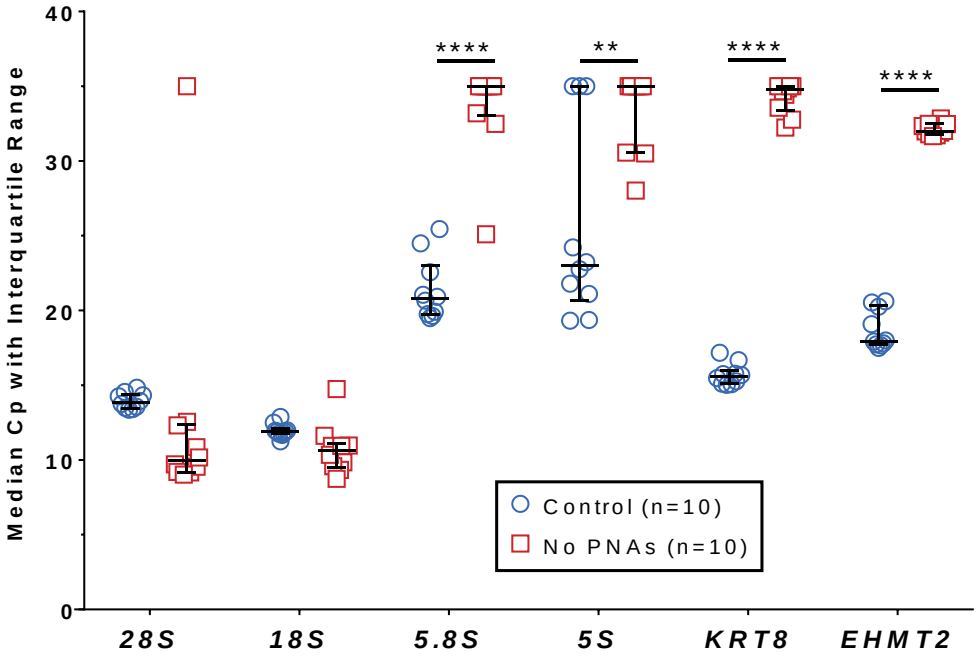


Figure 5-18 Expression of rRNAs and mRNAs in WTA without PNAs. The plot displays the expression levels of the four rRNAs (*28S, 18S, 5.8S,* and *5S*) as well as the two mRNAs *KRT8* and *EHMT2*. Data are displayed as median Cp values with range (whiskers). Statistics: multiple T-tests (chapter 3.16.1.2). **** $p < 0.0001$, ** $p < 0.01$

Since none of the listed approaches led a significant improvement in the rRNA or mRNA levels, I proceeded with an attempt to actively deplete rRNAs (chapter 5.4.3).

5.4.3 Targeted rRNA depletion

As changes to the WTA protocol did not result in a reduction of rRNA contamination, we attempted to deplete the rRNAs actively by using RNase H and a set of 113 DNA oligonucleotides (sequences provided in appendix chapter 12.5), both of which were kindly provided by Dr. Balagopal Pai. RNase H specifically degrades the RNA strand in a DNA/RNA hybrid, thereby enabling targeted depletion of RNA using complementary DNA oligonucleotides. The oligonucleotides were designed to target four different rRNAs, the cytoplasmic *28S* and *18S* species, as well as the mitochondrial *16S* and *12S* rRNAs. The principle of this approach as well as the oligonucleotide sequences were derived from the commercial NEBNext® rRNA Depletion Kit, which is based on a publication by Morlan and colleagues (Morlan et al., 2012) and was successfully tested by another independent research group (Adiconis et al., 2013). This type of rRNA depletion has also been routinely applied by Balagopal Pai, therefore it was considered a promising way to solve the rRNA contamination issue.

To test the functionality of the rRNA depletion, an eWTA with an additional rRNA depletion step was performed (protocol in chapter 3.15.4). The qPCR analysis of the WTAs revealed a significant effect of the treatment (two-way ANOVA, p=0.0013) and that *28S* and *18S* rRNA levels were significantly reduced by RNase H (Figure 5-19), however the effect was only weakly significant ($p<0.05$) in the case of the *18S* rRNA and only compared with the group that received only the oligonucleotides but not compared to the control group. In contrast, the mitochondrial rRNAs *16S* and *12S* as well as the tested mRNAs *KRT8* and *EHMT2* were not significantly affected (Figure 5-19).

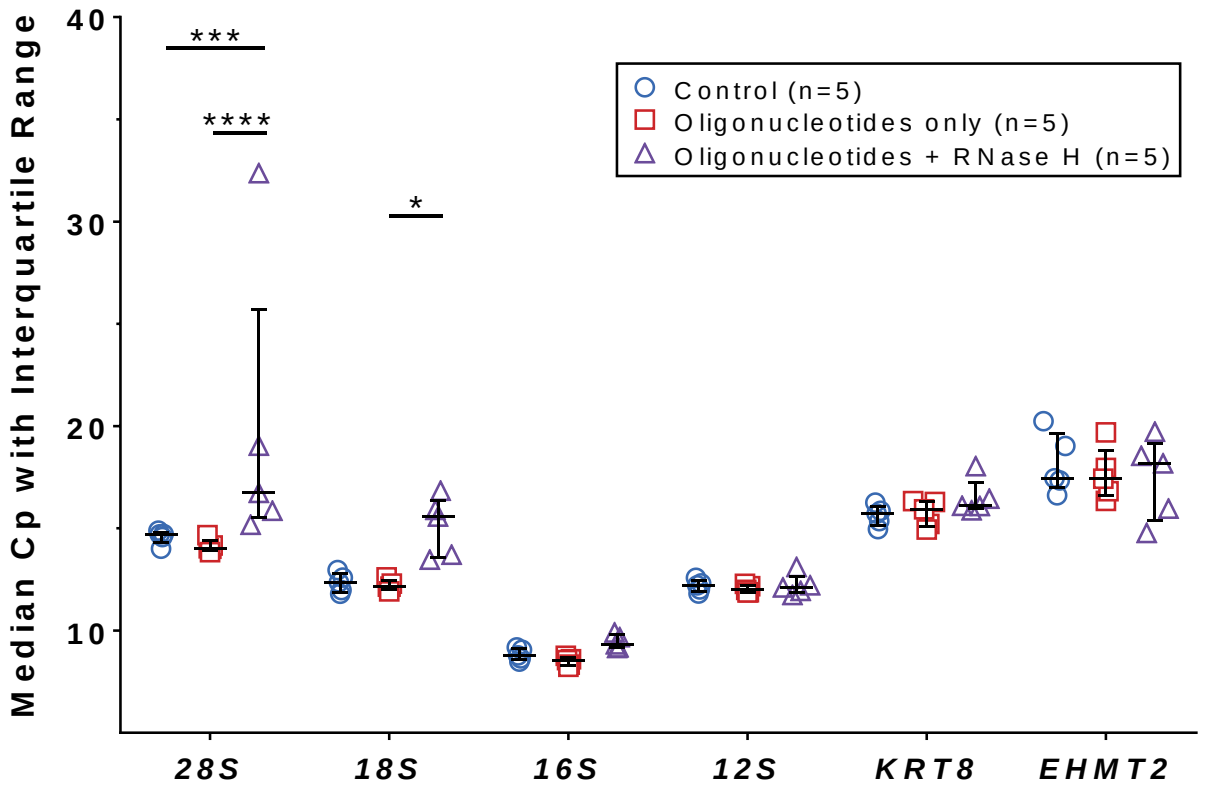


Figure 5-19 rRNA and mRNA levels in eWTAs after RNA depletion by RNase H – first attempt. The figure shows the levels of rRNAs and mRNAs detected in WTAs that were conducted with the rRNA depletion step. The qPCR data are depicted as median Cp values with interquartile range (whiskers). Statistics: two-way ANOVA (chapter 3.16.1.2); **** p<0.0001, *** p<0.001, * p<0.05

As the first attempt with the RNase H aliquot from Dr. Pai looked promising, we wanted to replicate these findings. Unfortunately, we had only received enough RNase H (ThermoFisher Scientific, TFS) for one experiment. Therefore, the experiment was repeated with a new aliquot of the RNase H enzyme with the exact same catalog number from TFS, as well as a cheaper version from New England Biolabs (NEB) for comparison. The experiment described above was replicated with two modifications: first, group II (only block oligonucleotides) was replaced by treatment with the NEB RNase H; and second, 1.5 µl of RNase H were added instead of 1.4 µl.

Despite having bought the exact same enzyme from TFS that was used in the first experiment, we were unable to replicate the previous result. Figure 5-20 shows how the previously significant effect on the *28S* and *18S* rRNAs (see Figure 5-19) had disappeared (two-way ANOVA) while both enzymes almost completely degraded the *KRT8* and *EHMT2* mRNAs, which we had measured as controls. Only one sample still contained the expected amount of *KRT8*. In contrast, both mRNAs were unaffected in the control group without RNase H.

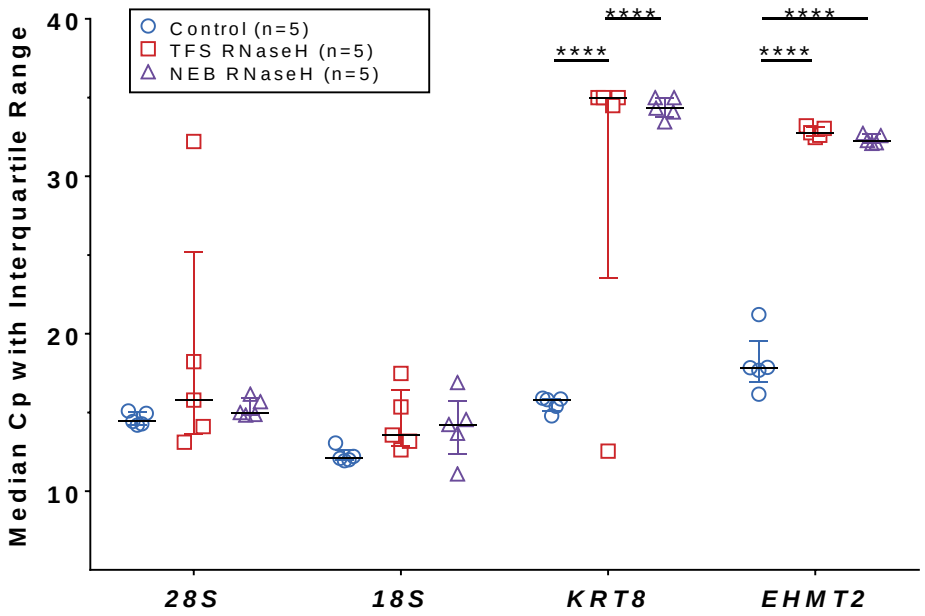


Figure 5-20 rRNA and mRNA levels in eWTAs after RNA depletion by RNase H – second attempt. The figure illustrates the levels of rRNAs and mRNAs detected in WTAs, in which rRNA was depleted with two different RNase H enzymes. The qPCR data are depicted as median Cp values with interquartile range (whiskers). Statistics: two-way ANOVA (chapter 3.16.1.2); **** p<0.0001; TFS = ThermoFisher Scientific, NEB = New England Biolabs

The rRNA depletion was attempted three more times with a slightly changed protocol every time. First, a heat inactivation step of the RNase H was included. Second, we tried to replicate the previous experiment with an extra control undergoing the heat inactivation together with the treatment group to control for heat degradation of mRNAs. In this experiment we also used a new aliquot of the blocking oligonucleotides from Dr. Balagopal Pai. Third, we performed the experiment on 30 pg of bulk RNA isolated from DU145 cells, which was treated with DNase, to check whether DNA contamination might cause unintentional priming of RNase H. However, none of these experiments changed the outcome: the mRNAs were still being completely degraded (see Figure 5-20).

Consequently, we decided to abandon the targeted rRNA depletion and to move on with testing of the rRNA blocking step in the eWTA protocol instead (chapter 5.5).

5.5 Effect of rRNA blocking oligonucleotides

5.5.1.1 Blocking of *Long fragment in vitro* transcript spike-in

In order to test the functionality of the blocking oligonucleotide approach, we utilized ZNA1, a ZNA blocking oligonucleotide designed by Verena Lieb targeting the 3' end of the *in vitro*-transcribed *LF* RNA. Using ZNA1 and the *LF* RNA as a spike-in, an eWTA experiment on SCEs was conducted with two treatment groups, one undergoing blocking and polyadenylation (group I) and one only undergoing polyadenylation as a control (group II), as described in chapter 3.15.5.1.

Analysis of the expression levels of *28S* and *18S* rRNAs as well as several mRNAs and the *LF* transcript revealed that the blocking of the *LF* RNA did not have any significant effect on any of the measured transcripts (Figure 5-21, multiple T-tests). However, we did observe that addition of the blocking oligonucleotide caused random dropouts of some SCEs (biological replicates) in some of the different transcripts. In case of dropouts (no output from the LightCycler instrument or wrong amplicon according to melting curve analysis), we assigned a Cp value of 35 to enable

inclusion of the samples into the analysis. These dropouts were mostly concentrated on cells of the second replicate experiment (Figure 5-21b) indicating a problem with that experiment. We repeated the joint analysis without the data from this run, but it did not change the result so this analysis was discarded.

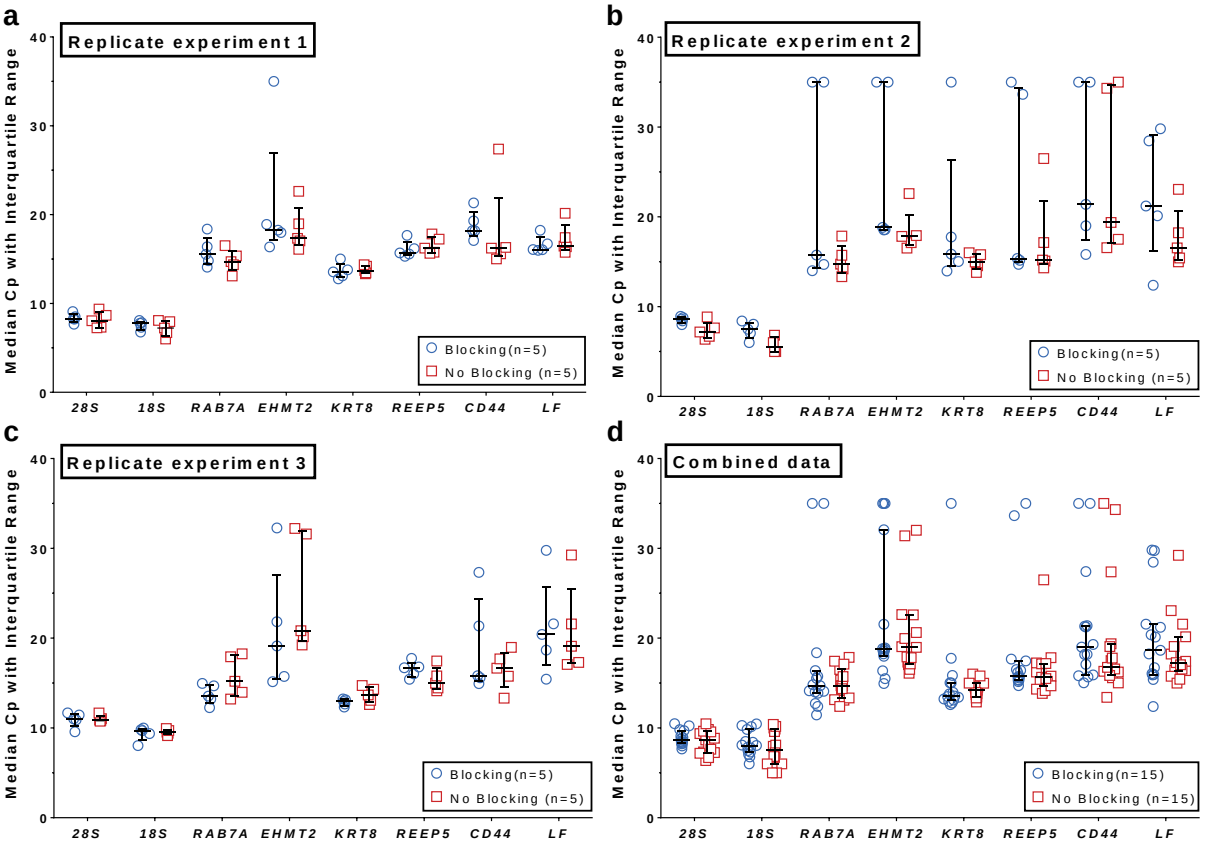


Figure 5-21 Effects of blocking on *LF* RNA spike-in and other transcripts in SCEs. The graphs illustrate the levels of different rRNAs and mRNAs and the *LF in vitro* transcript after the eWTA procedure with and without blocking of the *LF* RNA. The qPCR data are shown as median Cp values with interquartile range (whiskers). (a-c) Individual replicate experiments (d) Combination of the three replicates. Statistics: multiple T-tests (chapter 3.16.1.2)

Overall, this initial experiment of the *LF* transcript suggested that the principle of preventing a transcript from being polyadenylated via an oligonucleotide complementary to the 3' end of the target transcript was not working in our setting.

5.5.1.2 Blocking of endogenous rRNAs with ZNA oligonucleotides

Despite the failed blocking attempts on the artificial spike-in, we decided to investigate Verena Lieb's ZNA blocking oligonucleotides targeting the *28S, 18S, 5.8S,* and *5S* rRNAs. For this experiment, an additional control group without polyadenylation was included, resulting in three different treatment groups:

- Group I: polyadenylation and rRNA blocking
- Group II: polyadenylation without rRNA blocking
- Group III: no polyadenylation and no rRNA blocking

The experiment was conducted independently three times (for experimental procedure see chapter 3.15.5.2). Group III was done with four replicate samples instead of five in two of the replicate experiments conducted by a technician due to problems with the handling of the high

number of samples. This resulted in a maximal number of 13 (five from the first replicate and four from each of the other two experiments) samples for this treatment group. The levels of gene expression were measured by qPCR and the Cp values of each experiment were normalized to facilitate a pooled analysis of all samples. Unfortunately, there were contaminations in the *5.8S* and *5S* rRNA groups in two and one of the three replicate experiments, respectively, which is why these measurements were excluded from analysis resulting in only five and nine data points, respectively, for these transcripts.

There was no significant effect of the blocking procedure on any of the tested transcripts (Figure 5-22, two-way ANOVA). Fortunately, the mRNAs were not affected either. However, we observed a robust effect of the polyadenylation in groups I and II (blocking and no blocking with polyadenylation) across all four rRNAs compared to group III (no tailing and no blocking). Surprisingly, the direction of the effect was reversed for the *5S* rRNA.

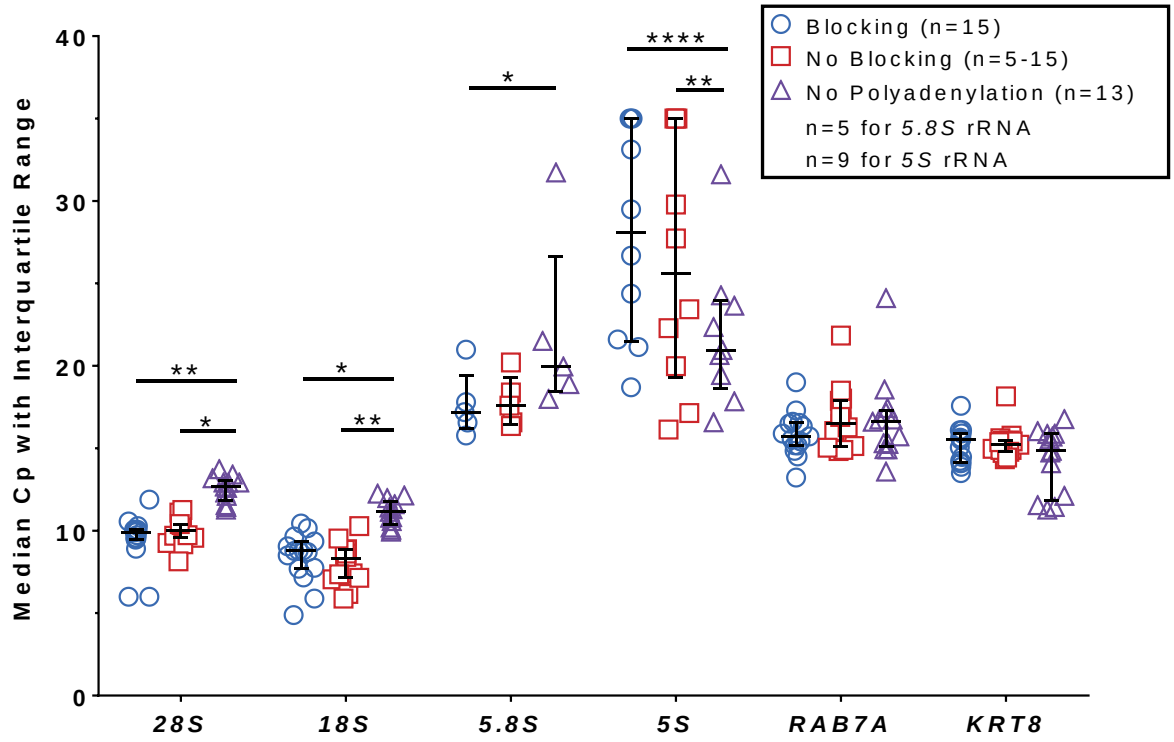


Figure 5-22 Effect of ZNA blocking oligonucleotides targeting *28S, 18S, 5.8S*, and *5S* rRNAs. The plot depicts the expression of four different rRNAs and two mRNAs in eWTAs performed with the ZNA blocking oligonucleotides. The qPCR data are displayed as median Cp values with interquartile range (whiskers). Due to contamination of the *5.8S* and *5S* rRNAs in some replicate experiments, nine and four data points, respectively, were excluded from analysis. Statistics: two-way ANOVA (chapter 3.16.1.2); **** p<0.0001; ** p<0.01, * p<0.05

In a nutshell, the presented experiment showed us that the blocking had no effect on transcript levels in its current form, however, it confirmed that the polyadenylation works, since the amounts of detected rRNAs were increased with the exception of the *5S* rRNA.

5.5.1.3 Comparison of two different sets of blocking oligonucleotides

Lastly, we wanted to investigate whether the set of 113 blocking DNA oligonucleotides of Dr. Balagopal Pai, which has previously been used in the rRNA depletion experiments (see 5.4.3), would display an effect, unlike Dr. Lieb's ZNA oligonucleotides. Therefore, a single eWTA experiment was performed that compared a control without blocking and without polyadenylation (group I) with polyadenylation alone (group II) and polyadenylation with either

Dr. Lieb's oligonucleotide's (group III) or Dr. Pai's oligonucleotides (group IV). The experiment was conducted as described in chapter 3.15.5.3. After the eWTA, the levels of gene expression of five rRNAs and two mRNAs were measured by qPCR. Since Dr. Pai's blocking oligonucleotides were targeting the mitochondrial *16S* and *12S* rRNAs instead of the cytoplasmic *5.8S* and *5S* species, I established qPCR primers for these two transcripts (see appendix chapter 12.4.2) and quantified these transcripts in the eWTA sample, too. The *5S* rRNA was skipped in this experiment, because it had previously provided odd results (see Figure 5-22).

The analysis revealed a highly significant effect of the treatment variable (two-way ANOVA, $p<0.0001$), but according to the post-hoc test neither the four ZNAs nor the 113 DNA blocking oligonucleotides exerted a significant effect on any of the examined transcripts (Figure 5-23). However, similar to the previous experiment (see Figure 5-22), there was an effect of the tailing procedure (groups II-IV compared to group I) in the *28S* and *5.8S* rRNAs. Furthermore, there was a weak tendency towards a higher Cp in the *28S*, *18S,* and *12S* rRNAs in samples treated with the DNA oligonucleotides.

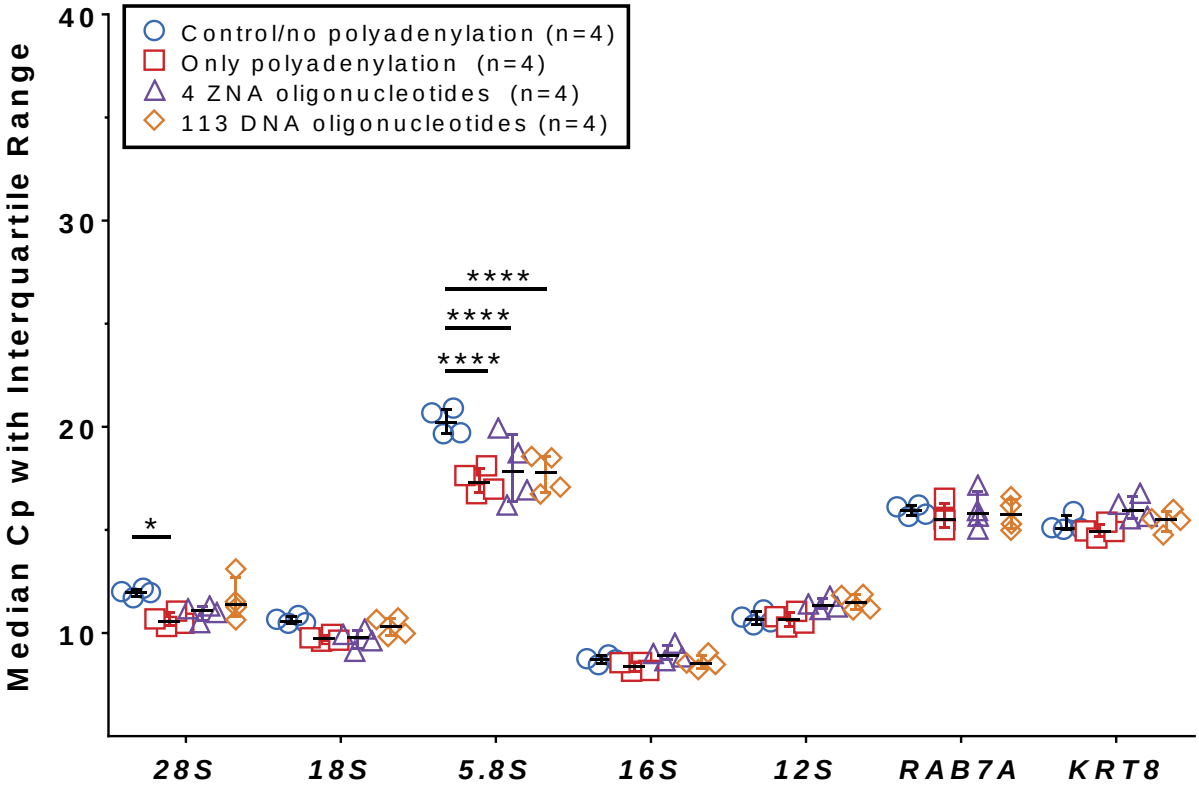


Figure 5-23 Comparison of two different blocking oligonucleotide sets in SCEs. The plot depicts the expression of five rRNA species and two mRNAs in eWTAs prepared with two different sets of blocking oligonucleotides. The qPCR data are displayed as median Cp values with interquartile range (whiskers). Statistics: two-way ANOVA (chapter 3.16.1.2); **** $p<0.0001$, * $p<0.05$

Taken together, the data presented in this chapter suggested that the blocking procedure in its current form was either not effective or masked by the high abundance of rRNAs in the cultured cells used for the experiments and the high inherent variation of the WTA itself (despite the usage of SCEs). However, since the passive rRNA contamination seemed to be a negligible problem for RNA-Seq (see Figure 5-17) and the polyadenylation had been shown to work on an artificial spike-in transcript as well as cell-intrinsic rRNAs (see Figure 5-13 and Figure 5-23), I want to present a preliminary eWTA protocol (chapter 5.6).

5.6 Proposed preliminary eWTA protocol

Taking together the data presented above, I propose a preliminary protocol for the eWTA. The rationale behind the changes to existing steps of the sWTA will be provided in the discussion (chapter 7.1). The reasons for exclusion of the blocking step from the protocol will be explained in the discussion as well (chapter 7.1.2). Please note that the described protocol still requires extensive optimization as well as final proof that it is able to detect endogenous miRNAs. However, the latter would have required development of another custom protocol as a readout due to incompatibility of our WTA adaptors with commercial solutions, which would have extended beyond the scope of this thesis. The preliminary eWTA protocol is performed as follows:

SCs are isolated as described in chapter 3.1.4 with two modifications. First, the lysis buffer, in which the cells are deposited, is prepared according to Table 5-10 (eWTA picking mix). Second, only 3.4 µl of the mix are deposited per tube to reduce the reaction volume later on. Following cell isolation and storage at -80 °C, the actual eWTA is performed. Before the thawing of cells, the eWTA lysis mix is prepared (Table 5-10). Subsequently, the SCs are thawed, 1 µl of the eWTA lysis mix is added to each SC and the cells are incubated in a PCR cycler for 10 min at 45 °C, followed by inactivation of the protease at 75 °C for 1 min. Note that the incubation at 22 °C for 10 min is skipped, because there are no PNAs to anneal yet. Following the protease lysis, 9.6 µl of Poly(A) mix are added (Table 5-10), the sample is incubated at 37 °C for 30 min and then cooled to 4 °C in a PCR cycler. By adding the indicated volume of Poly(A) mix, the mTRAP lysis buffer is finally diluted 1:5 (in the tailing reaction relative to the eWTA picking mix), which enables full functionality of the PAP. Lastly, 4 µl of PNA/GTC mix (Table 5-10) are added and the solution is placed in a PCR cycler for annealing of the PNAs. The PNAs are annealed by incubation at 75 °C for 1 min, followed by 22 °C for at least 10 min. The GTC is added to inactivate the PAP without a time-consuming and potentially RNA-damaging heat denaturation. Now that the sample has undergone polyadenylation and annealing of PNAs, it is ready for the capture of all polyadenylated molecules by addition of 4 µl of the streptavidin-conjugated mTRAP beads. From this point, the remaining WTA procedure is performed according to the sWTA protocol outlined in chapter 3.2.1.

Table 5-10 Master mix compositions for preliminary eWTA.

Name	Components for one reaction
eWTA picking mix	3 µl mTRAP lysis buffer 0.4 µl SUPERase
eWTA lysis mix	26.5 µl mTRAP lysis buffer 1 µl Protease solution (1 µg/µl)
Poly(A) mix	1.5 µl 10mM ATP 0.64 µl NaCl 5 M 0.15 µl MgCl2 0.5 M 0.75 µl SUPERase (20 U/µl) 0.75 µl *E. coli* Poly (A) Polymerase (5U/µl) 5.81 µl DEPC-water
PNA/GTC mix	1 µl GTC 5 M 3 µl Oligo(dT) PNA 37.5 µM

6. Discussion of transcriptomic and genomic characterization of DCCs

Although only 6 % of breast cancer (BC) patients display metastatic disease at the point of diagnosis, up to 50 % of early BC (eBC) patients progress to the metastatic stage in the course of the disease (Chambers et al., 2002; O'Shaughnessy, 2005; Cardoso and Castiglione, 2009; Lu et al., 2009). Despite several decades of research and the urgent need for better systemic therapies, metastatic disease still accounts for roughly 90 % of cancer-related deaths, as current treatment strategies fail to successfully eradicate disseminated cancer cells (DCCs) and circulating cancer cells (CCC) hiding in patients' bodies (Bendre et al., 2003; Fidler, 2003; Weigelt et al., 2005; Loberg et al., 2007; Hartkopf et al., 2011; Redig and McAllister, 2013), which is mainly due to a lack of understanding of the underlying molecular mechanisms. This problem also applies to the LumB subtype, which metastasizes much more frequently than the closely related LumA subtype (Buonomo et al., 2017) resulting in a strong reduction of overall survival (OS) compared to LumA patients (Cheang et al., 2009; Fallahpour et al., 2017). However, it is still unknown why LumB has a higher propensity to form macrometastases. Since DCCs represent the seed from which metastases develop and because their presence is correlated with patient survival (Cote et al., 1991; Harbeck et al., 1994; Schindlbeck et al., 2011; Domschke et al., 2013; Banys et al., 2014; Hartkopf et al., 2019), we decided to investigate DCCs from LumA and LumB in detail, in order to promote the development of better diagnostics and treatment strategies for LumB BC. Unfortunately, DCCs cannot easily be identified, due to the presence of $EpCAM^+$ non-cancer cells (NCC) in the bone marrow (BM) that probably belong to the erythroid progenitor cell lineage (Bühring et al., 1996; Lammers et al., 2002; Gužvić et al., 2014). Therefore, the aims of this thesis were (1) to investigate how true DCCs can be distinguished from $EpCAM^+$ non-cancer cells (NCC) and (2) how LumB DCCs differ from LumA DCCs on the genomic and transcriptomic levels by utilizing our chair's extensive collection of $EpCAM^+$ cells isolated from the BM of M0 and M1 patients.

6.1 Identification of true DCCs

A total of 100 out of 247 (40.5 %) screened M0 patient BM samples that we received contained $EpCAM^+$ cells, which is slightly above the maximum rate of tumor cells observed by others (20.5 %, 32 %, and 38 %; Schlimok et al., 1987; Harbeck et al., 1994; Schindlbeck et al., 2011). However, all mentioned studies were using cytokeratin (CK)-targeting antibodies to detect DCCs, in contrast to our EpCAM-based approach. It is known that EpCAM is also expressed on an erythroid progenitor cell population located in the BM (Bühring et al., 1996; Lammers et al., 2002; Gužvić et al., 2014), which likely increases the number of cells we detect (compared to CK-based detection methods) and may explain our higher EpCAM-positivity rate. Consequently, $EpCAM^+$ cells were also detected in 20 out of 40 (50 %) non-cancer patients (=healthy donor, HD) which was not significantly different from the frequency observed in M0 BC patients (see Table 4-1). This is in accordance with the 56 % EpCAM-positivity rate detected in male HDs (Gužvić et al., 2014). Due to the presence of these cells in the BM samples it was necessary to find a method to reliably distinguish $EpCAM^+$ NCCs from true DCCs to prevent them from distorting the results. Two different approaches were applied to achieve this distinction: qPCR (chapter 6.1.1) and copy number alteration (CNA) profiling (chapter 6.1.2).

6.1.1 A qPCR signature can identify true DCCs

The novel qPCR signature consisting of the four genes *AHSP, CA1, AHNAK,* and *JUN* was successfully established on the training set of cells (see Figure 4-3) with an accuracy of 89 % and a misclassification rate of 11 % (see Table 4-7), using true DCCs whose malignant origin was proven by CNA profiling and HD-derived EpCAM$^+$ cells (Patwary, in preparation; Haunschild, 2013). Among the 47 tested single cells, there were no false negatives (wrong classification as NCC although a cell was truly aberrant), but four presumable false positives (wrong classification as DCC despite the cell presumably having a balanced genome). Here, it is important to note that the EpCAM$^+$ HD-derived NCCs used as the control group, were not tested by mCGH, it was merely assumed these cells had balanced genomes. Intriguingly, three of these four false positive cells were later shown to have aberrant genomes despite originating from HDs (see chapter 4.2.2.2). However, when the M0 DCC signature was applied to a larger set of single cells, the accuracy and misclassification rate worsened. Interestingly, it was observed that the vast majority of the misclassifications were again "false" positives (see Table 4-14). If the signature were not working properly, one would expect that the misclassifications were similarly distributed to false positives and false negatives. However, the distribution did not seem random. It is important to note that QPCR is a very sensitive technology that can accurately quantify transcripts down to only ten copies (Androvic et al., 2017), so it is possible that the high number of "false" positive classifications by the qPCR may indicate that the gene expression signature can detect subtle characteristics of early DCCs (eDCC) which CNA profiling of the gDNA cannot, since metaphase comparative genomic hybridization (mCGH) and LowPass-Sequencing (LP-Seq) have rather high detection limits of 100 kb (LP-Seq; Ferrarini et al., 2018) and 10-20 Mb (mCGH; Bentz et al., 1998; Jeuken et al., 2002), respectively. As a consequence, aberrations below this limit will be missed and cells will be wrongly classified as NCCs with a balanced genome despite being tumor cells. Schardt and colleagues have previously demonstrated that tumor cells can indeed appear to be karyotypically normal (Schardt et al., 2005), which supports this hypothesis. Additionally, the evaluation of the CNA profiles was complicated and partly subjective, which makes it prone to errors (see chapter 6.1.2). Therefore, it would be interesting for future studies to examine these alleged "false" positive samples more closely, for example by whole genome sequencing with a better coverage than is offered by LP-Seq. If the cells were indeed found to carry small aberrations, it might prove that the M0 DCC signature is in fact better than LP-Seq and mCGH and could replace these technologies as the gold standard for identification of early BC DCCs.

The M0 DCC signature's performance was tested in two variations. First, a stringent signature, which only included cells displaying the DCC or NCC expression patterns, but not the DCC-like and NCC-like ones (see Table 4-8 and Table 4-14, "Stringent" table). Second, there was also a relaxed variant, which combined the DCC and DCC-like as well as the NCC and NCC-like expression patterns, in order to include more of the tested cells and to obtain a more precise performance estimation. The latter seemed to be a valid approach as well, since the performance metrics remained at a similar level compared to the stringent variant (Table 4-14, compare "Stringent" table and "Relaxed" table). Therefore, the relaxed M0 DCC signature is likely the better option for future studies, because it allows inclusion of more EpCAM$^+$ cells without losing accuracy.

The finding that the M0 DCC signature classified four HD-derived NCCs (17.5 %) as DCCs and that three of these cells had aberrant genomes not only supported the validity of the signature, but was also interesting in light of the early dissemination model (Klein, 2008). Since the mean age of the HDs at surgery (mostly hip replacement surgeries) was 66.2 ± 12.9 years (see Table 4-3) and the two HDs, from which the three NCCs with confirmed aberrations originated, were 60 and 82 years old, respectively, it is possible that these cells represented eDCCs of primary tumors (PT), which were still undetected at the point of BM aspiration. Unfortunately, our current ethics vote does

not permit inquiries about follow-up data of the HDs, so it is not possible to test this hypothesis for the existing samples, but a new ethics vote will enable collection of follow-up data for future HD samples. Furthermore, it is possible that the fourth cell that clustered with the DCCs and expressed the DCC signature, was falsely classified as balanced due to its aberrations being below the detection limit of the CNA profiling.

The comparison of mCGH, LP-Seq, and qPCR revealed how poorly the qPCR signature worked in M1 EpCAM$^+$ cells (see Table 4-8 and Figure 4-9). However, this was expected, because no M1 DCCs were included in the initial training set for the signature (see Table 4-6). The fact that the signature worked so poorly on M1-derived EpCAM$^+$ cells, which more frequently expressed the NCC and NCC-like patterns than DCC and DCC-like patterns (see Table 4-8), indicates that M1 DCCs are in fact very different from M0 DCCs, which is consistent with the observation that M0-stage DCCs and M1-stage DCCs differ significantly in their CNA profiles (Schmidt-Kittler et al., 2003).

6.1.2 CNA profiling unambiguously identifies true DCCs

For validation of the qPCR signature, CNA profiling by LP-Seq was applied to identify the true DCCs among the isolated EpCAM$^+$ cells resulting in 91 out of 262 CNA profiles being used for further analysis (see Figure 4-4). Subsequently, the comparison of overlapping LP-Seq and mCGH results available for a subset of cells showed that the two CNA profiling methods agreed in the majority of cases (see Figure 4-7 and Table 4-11). Therefore, the two datasets were combined, which increased the numbers of cells that could be clearly classified as DCC or NCC to a total of 115, which were then analyzed in more detail. However, the analysis of the CNA profiles had some limitations (see chapter 6.4).

To my knowledge, there is no published workflow on how to combine mCGH and LP-Seq data for CNA analysis. To facilitate joint analysis of all available aberrant (mCGH and LP-Seq) CNA profiles using *Progenetix*, an R script was developed to annotate the LP-Seq-derived RefSeq files containing the genomic coordinates of the aberrations with information of the corresponding cytobands (see chapter 4.3.1). It was done this way, because down-scaling of the more detailed Low-Pass-Seq data (max. resolution 100 kb; Ferrarini et al., 2018) was considered the best way to combine them with the available lower resolution mCGH data (max. resolution 10-20 Mb; Bentz et al., 1998; Jeuken et al., 2002) of Gundula Haunschild (Haunschild, 2013). The general functionality of the script was later confirmed by one of our bioinformaticians, who also pointed out a small error in the script (see chapter 6.4).

During the manual annotation process following the cytoband annotation (see chapter 4.3.1), five of the 49 aberrant profiles were excluded (see Figure 4-10), because - despite the clear classification as aberrant - they contained dozens of tiny aberrations which would have made it extremely difficult to decide which ones were real. Therefore, these profiles were not considered robust enough to include them in the detailed CNA analysis. Future studies might overcome this problem by adjusting the minimum size of aberrations in the bioinformatic pipeline to filter out these small aberrations.

The comparison of cumulative aberration profiles generated from the same cells either by LP-Seq or mCGH revealed that both technologies agreed on the major aberrations (see Figure 4-11), but that there were also several smaller alterations that were not overlapping. These non-overlapping CNAs between the two methods might be technical or software artifacts, which – due to the small sample size – appear relatively large in the frequency plot (see Figure 4-11) compared to the real aberrations, due to the plotting of the aberrations in per cent of cells, which carried a specific aberration. With higher sample numbers, it would probably be easier to discern true aberrations

from artifacts that occur at random locations, because the artifacts would make up smaller percentages of the total number of aberrations. However, artifacts that are fixed to specific loci would not be discernible this way, because they would likely accumulate at a rate comparable to real aberrations. This problem represents a limitation of the method and might be solved by a systematic comparison of CNA profiles derived from DCCs of different cancer entities (see chapter 6.4). It was also observed that some CNAs, which appeared in both mCGH and LP-Seq profiles, were more frequent in the LP-Seq-derived profiles. This may indicate a higher sensitivity of the NGS-based method compared to mCGH or that detection of small aberrations is easier with the RefSeq files provided by LP-Seq. Another explanation for the discrepancies – both in presence of aberrations and frequency - is that the aberrations were annotated by two different persons. Since the annotation process is partly subjective, it is possible that I may have been less restrictive than Dr. Haunschild, who annotated the mCGH profiles. Lastly, the LP-Seq profiles were more nuanced and precise compared to the mCGH-derived profiles. This is most likely due to the translation of the precise genomic coordinates from the RefSeq files into corresponding cytobands, which facilitates sharper CNA boundaries than the mCGH annotation, which involves comparison of the CNA profiles to printouts of chromosomal ideograms. Despite the existence of some non-overlapping aberrations, the overall accordance of both methods was considered robust, because the largest aberrations (spanning at least a chromosome arm) were in agreement. In addition, a recent publication by the developers of the LP-Seq method, in which they compared the LP-Seq protocol (in this case for the Ion Torrent sequencing platform) with array CGH (aCGH), showed a high concordance of the methods (Ferrarini et al., 2018). Taken together, the available data suggested that LP-Seq and mCGH were comparable and that the combination of the LP-Seq data of the current study and Dr. Haunschild's mCGH data was a valid approach to increase the number of samples for analysis.

6.2 Characterization of M0 and M1 DCCs

6.2.1 M0 DCCs carry fewer aberrations than M1 DCCs

Overall, CNA profiling revealed that 51 % of M0 EpCAM$^+$ cells were aberrant (LP-Seq and mCGH combined; see Table 4-12), while more than 88 % of M1-derived EpCAM$^+$ cells carried genomic aberrations (LP-Seq and mCGH combined; see Table 4-12), which represented a significant difference in the aberration rate between the M0 and the M1 collectives. This result is in accordance with two of our group's previous studies (Schmidt-Kittler, 2003; Haunschild, 2013).

Utilizing the combined data, M0 and M1 DCCs of the EpCAM$^+$ collective were compared to a set of cytokeratin-positive (CK$^+$) DCCs collected by a previous PhD student (Schmidt-Kittler, 2003). It was discovered that, while M0 and M1 DCCs within each collective differed strongly, there was no remarkable difference between the EpCAM$^+$ and CK$^+$ cells (see chapter 4.3.3), which was in accordance with both Dr. Schmidt-Kittler's (M0 versus M1 in CK$^+$ DCCs; Schmidt-Kittler, 2003) and Dr. Haunschild's results (M0 versus M1 in both EpCAM$^+$ and CK$^+$ cells; Haunschild, 2013). The pronounced difference between M0 and M1 DCCs also agrees with a recent publication of another group working on EpCAM$^+$ DCC pools (Magbanua et al., 2018a). Moreover, the fact that M0 DCCs carried fewer aberrations than the M1 DCCs supports the early dissemination and parallel progression models advocated by our group (Klein, 2008, 2009). Regarding the differences between EpCAM$^+$ and CK$^+$ collectives, the CNAs of the EpCAM$^+$ M0 DCCs appeared rather randomly distributed across the entire genome, while the CK$^+$-derived CNAs displayed an incipient enrichment of aberrations in a few loci, of which the most prominent was the gain in the 1q chromosome, which was also the only highly significant difference (see Figure 4-14 and Table

4-17). In contrast, the M1 DCCs of both collectives displayed many characteristic chromosomal aberrations (see Figure 4-14; amplifications: 1q, 8q, 16p, 17q, 20; deletions: 8p, 13q, 16q, 17p) that have been described in the literature (Kallioniemi et al., 1994; Nishizaki et al., 1997; Hermsen et al., 1998; Tirkkonen et al., 1998; Buerger et al., 1999; Roylance et al., 1999; Buerger et al., 2001; Haunschild, 2013). Furthermore, the loss of chromosome 11q, which was observed frequently in M1 DCCs of both collectives and also in CK^+ M0 DCCs (see Figure 4-14), is associated with a worse prognosis and early onset of BC (Gentile et al., 1999; Gentile et al., 2001). The observed incipient enrichment of known aberrations in CK^+ DCCs suggests that these cells are slightly more progressed than $EpCAM^+$ DCCs, however the differences were marginal and a double staining against CK and EpCAM found that the majority of DCCs were either $EpCAM^+/CK^+$ or $EpCAM^-/CK^-$ suggesting that $EpCAM^+$ and CK^+ cells represent two facets of the same tumor cell population (Haunschild, 2013).

6.2.2 M0 DCCs proliferate more frequently than M1 DCCs

Since proliferation in foreign microenvironments (ME) is crucial for accumulation of somatic mutations and, ultimately, growth of macrometastases (Klein, 2009), a targeted survey of proliferation marker expression in the DCC collective was carried out. Surprisingly, the data suggested that half of the M0 DCCs were proliferating (see Figure 4-21), since they were expressing the proliferation markers *MKI67* or *MCM2* or both (see Figure 4-21 and Figure 4-24). In contrast, the same applied only to ~15 % of M1-derived DCCs (see Figure 4-24). This discrepancy was surprising, because until now it was assumed that DCCs located in the BM should either form a growing micrometastasis, enter dormancy, or die when arriving in distant organs (Chambers et al., 2002; Aguirre-Ghiso, 2007). Additionally, the prevailing opinion is also that DCCs already possess all hallmarks of cancer (Hanahan and Weinberg, 2000, 2011). A proliferation rate of 50 % in M0 DCCs in parallel to absence of a detectable metastasis (hence the M0 status) did not fit this concept. Therefore, we initially wondered whether there may have been a problem with the qPCR, but the RNA-Seq data corroborated the qPCR results, because the expression of numerous cell cycle-associated genes matched the qPCR-defined proliferation status in the majority of cells (see Figure 4-30). Interestingly, the high rate of proliferation may be consistent with the concept of a steady-state proliferation that has been suggested by others (Holmgren et al., 1995; Uhr and Pantel, 2011; Yadav et al., 2018). This steady-state proliferation - or tumor mass dormancy as Yadav and colleagues called it - involves a constant turnover of M0 DCCs due to a balance of proliferation and cell death in proliferation-permissive distant organs, which prevents metastatic outgrowth. This steady-state proliferation may – at first glance - give the impression that the DCCs are quiescent, because they do not grow into a macrometastasis, however, the presented data demonstrate that half of eDCCs are in fact not dormant in the classical sense, meaning in a quiescent state, but are actively cycling (see Figure 4-21). However, functional proof that DCCs are indeed proliferating is still missing, which is a limiting factor of this study (see chapter 6.4). Overall, this phenomenon might suggest that eDCCs have not yet developed all hallmarks of cancer, since they do not form metastases. However, the high proliferation may actually drive acquisition of the necessary mutations for development of macrometastases, in the course of which only the fittest cells survive the foreign ME, while the less fit cells are constantly culled. This might also explain why BC recurrence can take many years (Early Breast Cancer Trialists' Collaborative Group, 2005) and supports the notion that cancer should be viewed as a pathogenic evolutionary process (Klein, 2013). Maybe because individual mutations associated with the hallmarks of cancer (Hanahan and Weinberg, 2000, 2011) may not immediately provide survival benefits by themselves, DCCs need to remain in this steady-state proliferation until all hallmarks of cancer have been acquired or at least until angiogenesis has been induced and the

DCCs are able to thrive (Holmgren et al., 1995). The result that M1 DCCs displayed a proliferation rate of only 15 % (see Figure 4-24) suggests that the M1 DCCs are able to reduce their proliferation rate after acquisition of all hallmarks. The reason why this happens will have to be investigated in more detail.

Although it is possible that DCCs are lodged in all organs of a patient (Klein, 2011), it seems that somatic progression by proliferation is only possible in certain organs with the right metastatic niche that allows DCC proliferation. Interestingly, the high proliferation rate observed in the DCCs (see Figure 4-21) coincides with BM being the most frequent metastatic site of HR^+ BC. Therefore, the rate of proliferation permitted by a target organ may ultimately be reflected by the preferred metastatic sites of BC that have already been observed by Paget in the 19th century, which were then transformed into his seed and soil hypothesis (Paget, 1889). To test this hypothesis, future studies would need to isolate DCCs from the BM of patients, who suffer from other cancers, which form less bone metastases, e.g. colon cancer (Patanaphan and Salazar, 1993; Assi et al., 2016), and compare the proliferation rates of the colon cancer DCCs with BC DCCs.

Expression analysis of cell cycle-associated genes revealed that 39 out of 671 of these genes were differentially expressed among the BM-derived DCCs and that the cells separated into proliferating and non-proliferating cells based on expression of these genes (see Figure 4-30). This separation coincided very well with the proliferation status determined by qPCR, confirming the validity of the qPCR-based classification into proliferating and non-proliferating cells, both for M0 and M1 patient-derived DCCs. In a recent publication, Magbanua and colleagues looked at pools of 20 $EpCAM^+$ DCCs and discovered two DCC clusters. One with high expression of *MKI67* and *CCNB1* as well as other proliferation and cancer stem cell-associated genes, which may correspond to the proliferating cells from this study, and one with higher levels of epithelial cytokeratins, vimentin (*VIM*) and estrogen receptor 1 (*ESR1*), which could correspond to the non-proliferating cells due to the more epithelial expression profile (Magbanua et al., 2018a). Additionally, they also observed a tendency towards a higher recurrence score and poorer survival of the patients, from whom the proliferating cell pools were isolated, compared to the non-proliferating pools, according to the Oncotype Dx signature (Paik et al., 2004). This result may support the hypothesis that proliferation drives somatic progression and – ultimately – metastasis, which translates into more frequent relapse in patients with more proliferating DCCs. Furthermore, Magbanua and colleagues think that their proliferating cluster, which overexpressed the putative stem cell marker *ALDH1A1* and the stemness and proliferation-promoting *TACC3* gene, may represent cancer stem cells (Magbanua et al., 2018a), which have been hypothesized by others (Visvader and Lindeman, 2008; Pantel and Alix-Panabières, 2014). Building on this, future studies might look for signs of stemness in the proliferating DCCs identified in the current study.

Interestingly, the proliferating cells additionally branched into two subclusters (see Figure 4-30), which differed in their expression strength of the displayed genes. This was not observed by Magbanua and colleagues, which may be due to the methodological differences (Magbanua et al., 2018a). First, they were analyzing DCC pools from M0 patients instead of SCs. Furthermore, these pools were isolated using FACS with EPCAM as a positive marker and CD45 as a negative selection criterion. As the current work has revealed, EPCAM is not a perfect marker for DCC isolation (see Table 4-1), as there are also $EpCAM^+$ NCCs in the BM. Additionally, previous studies of our group have shown that only ~20-28 % of HD-derived NCCs expressed CD45, meaning that CD45 is not a sufficient negative selection criterion to separate true DCCs from NCCs (Haunschild, 2013; Gužvić et al., 2014). Therefore, it is likely that Magbanua and colleagues analyzed a mixture of DCCs and NCCs, which will distort the observed gene expression patterns. Second, Magbanua and colleagues included BM leukocytes in their clustering analysis. The larger relative difference between DCCs

and the leukocytes will likely lead to a closer clustering of DCCs, whereby the two subclusters within proliferating DCCs, which were observed in the current study (see Figure 4-30), may have been missed. Third, they used a panel qPCR assay to measure a selected set of genes, which may have led to them missing important genes for further separation of the proliferating cell cluster. Nevertheless, the mere fact that others also discovered a separation of M0-derived DCCs into proliferating and non-proliferating populations corroborates the results of this work.

6.3 Differences between LumA and LumB subtypes

6.3.1 Representation of subtypes in study cohort

Looking at the subtype-stratified numbers of screened M0 patient BM samples (see Table 4-2, middle column), the incidence of luminal BC overall (LumA and LumB together = 72.8 % of cases) was in accordance with the literature (Kennecke et al., 2010; Eroles et al., 2012). Interestingly, the relative frequencies of LumA and LumB are reversed in our collective compared to the studies of Kennecke and Eroles, according to whom LumA should be roughly two times more abundant than LumB. In contrast, we observe that the LumB type appears more frequently (45.7 % of cases) than the LumA type (27.1 % of cases, see Table 4-2). In concordance with our data, there are at least three other recent studies with similar findings (Inic et al., 2014; Dai et al., 2015; Hashmi et al., 2018). Therefore, it is possible that the incidence rates of LumA and LumB subtypes have changed over time for unknown reasons, as Kennecke and colleagues used data originating from 1986-1992, while our patients were recruited between 2008 and 2015. It could be that nutrition and changes in lifestyle, which are known to have an influence on cancer incidence (Barnes et al., 2016; Cicco et al., 2019), have contributed to such a shift in subtype incidences, however, this remains to be investigated. Alternatively, it could be that there is some kind of bias in the selection of patients in the clinics, which provide our patient samples, resulting in a higher number of more aggressive LumB patient samples being delivered to our laboratory, despite the true incidence rates still being the same as described by Kennecke and colleagues. Yet another possibility could be that neoadjuvant endocrine therapy (NET) was increasingly used on LumA patients in recent years potentially leading to BC remission, which diminished the need for surgery and therefore BM samples were not acquired. This hypothesis might be supported by a study from Alba and colleagues, who discovered that NET in ER^+ $KI67^{low}$ (=LumA) resulted in a response comparable to chemotherapy, but with less toxic side-effects (Alba et al., 2012). Furthermore, patients receiving NET might become less fit for surgery or simply do not have their tumors excised for other reasons (Selli and Sims, 2019), which could also account for the lower frequency of LumA BM samples. This would also explain why only 9.9 % of M0 patients included in this study received NET (see chapter 4.1.2), simply because LumA patients with successful NET usually do not need to have surgery and consequently their BM is not aspirated. However, the underlying reason for the observed underrepresentation of the LumA subtype among the BM samples needs to be investigated in future studies, because the data of the current study do not allow to draw any conclusions about this matter.

Due to the higher number of LumB patients, the total number of isolated LumB subtype $EpCAM^+$ cells was also higher compared to LumA (see Table 4-4). Nevertheless, the comparison of LumA and LumB subtypes regarding the rate of EpCAM-positivity (presence of $EpCAM^+$ cells in the BM, see Table 4-2) or number of $EpCAM^+$ cells isolated per patient (see Figure 4-1) did not reveal any significant differences, suggesting that LumA and LumB cancers disseminate at equal rates. Furthermore, the CNA profiling revealed that the rate of true DCCs in the BM of LumA and LumB patients was also similar (see Table 4-12 and Table 4-13).

6.3.2 LumA and LumB DCCs display similar CNA profiles

No significant differences between LumA and LumB DCCs were detected in the frequency of any of the analyzed CNAs (see chapter 4.3.4), despite the fact that others have shown that LumB PTs displayed CNAs significantly more often than LumA PTs, at levels comparable to basal-like and HER2-enriched cancers (Creighton, 2012). However, it is important to keep in mind that the analyzed samples were eDCCs and not PTs. Judging from the strong discrepancy observed between M0 and M1 DCCs (see chapter 4.3.3), it may well be that LumA and LumB DCCs from M1 patients would display significant differences, and could provide hints as to which aberrations eDCCs need to obtain to be able to successfully colonize the BM. Unfortunately, the current study is limited in this regard, because there was only one LumA patient in the M1 group and from this patient only one true DCC was isolated (see Table 4-12 and Table 4-13). Consequently, more samples will have to be obtained, in order to tackle this question. Overall, the strong divergence between Creighton and colleagues' results derived from PTs and the results derived from the eDCCs is more consistent with the parallel progression model (Klein, 2009) than the linear or late dissemination model (Foulds, 1954; reviewed by Klein, 1998; Weinberg, 2008; Valastyan and Weinberg, 2011).

6.3.3 LumA and LumB DCCs express cell cycle genes differently

The qPCR analysis of *MKI67* and *MCM2* expression did not reveal any differences between LumA and LumB DCCs, neither in the positivity rates nor in the expression levels even when only proliferating DCCs were assessed (see Figure 4-21 and Figure 4-22). Interestingly, the qPCR data showed a separation of the cells into proliferating and non-proliferating populations in both subtypes to a similar extent (see Figure 4-19), which was consistent with the RNA-Seq results (see Figure 4-30). The qPCR also uncovered that the proliferation marker expression of DCCs was not correlated with the proliferation state of the matched PT, regardless of the BC subtype (see Figure 4-20), which suggests that eDCCs strongly diverge from their corresponding PTs, a phenomenon that was also recently described in melanoma-associated DCCs (Werner-Klein et al., 2018). In comparison, the *MKI67* expression of DCCs determined by RNA-Seq compared to the KI67 status of the PT revealed a tendency towards a weak correlation, while there were moderate and strong significant correlations between PT proliferation status and the overall expression pattern as well as between *MKI67* expression and overall gene expression, respectively (see Figure 4-31). These results suggest that, while a higher KI67 status of the PT only marginally translated to a higher expression of *MKI67* in matched DCCs, it was clearly correlated with changes in global gene expression. As the KI67 status of the PT determines whether a luminal BC patient belongs to the LumA or LumB subtype, this also translated into a robust separation of LumA and LumB subtypes according to the principal component 1 (PC1) variable, which represented the overall gene expression pattern (Figure 4-31b).

Regarding the identities of the two proliferating subclusters mentioned earlier (see 6.2.2), the data also unveiled that the proliferating DCC cluster A (DCC-Pa) consisted of DCCs of four different BC subtypes, while the proliferating DCC cluster B (DCC-Pb) comprised only LumB-derived DCCs with the exception of one LumA DCC (see Figure 4-30). This suggests a distinct proliferative behavior of the LumB DCCs compared to LumA DCCs. Unfortunately, the behavior compared to the TNBC subtype cannot be judged due to the low sample number. Based on the relative higher expression of cell cycle-associated genes in the DCC-Pb cluster than the DCC-Pa cluster, it is possible that LumB-derived DCCs proliferate more or faster than LumA DCCs, which would mean they undergo somatic progression more quickly. However, functional evidence for this is still missing to this point, therefore, the results need to be interpreted carefully (see chapter 6.4).

Consistent with this hypothesis is the finding that *cyclin- dependent kinase 1 (CDK1)* is among the differentially expressed genes, because deregulation of *CDK1* frequently leads to genomic and chromosomal instability (reviewed by Malumbres and Barbacid, 2009). Moreover, it has been reported that *CDK1* is the only essential cyclin-dependent kinase (CDK) in mammalian cells that is able to drive embryonic development in the absence of all interphase *CDK*s (Santamaría et al., 2007). Based on these findings, it may be that LumB BC DCCs can exploit this ability of *CDK1*, which is normally meant for embryonic development, to their advantage. Moreover, interphase CDKs may be dispensable for normal cells, but can be essential for tumor cells, which may have specific requirements for individual CDKs (Malumbres and Barbacid, 2009). For example, *CDK4* is not essential for mammary gland development, but for development of mammary gland tumors, which can be arrested in G0/G1 phase by *CDK4/6* inhibition (Finn et al., 2009; Malumbres and Barbacid, 2009). Future studies might take a closer look at the available data of this work and investigate whether any CDKs or associated cyclins are specific to either of the luminal subtypes, as it is already known that different BC subtypes display specific dependencies on the cell cycle and associated checkpoints (Cancer Genome Atlas Network, 2012; Thu et al., 2018). Additionally, a recent study described that antisense non-coding mitochondrial RNAs (ASncmtRNA) are specifically down-regulated in proliferating tumor cells, but not normal cells (Fitzpatrick et al., 2019). It is possible that LumA and LumB DCCs display differences in their expression of these RNAs. The authors also reported that knockdown of ASncmtRNA induces massive apoptosis in BC cell lines but not normal mammary cells. This property makes this type of RNA an especially interesting molecule to study in DCCs and would be an additional application of the novel extended WTA presented in this study (chapter 5.6) once the protocol has been finalized. In a nutshell, extensive research will be required to reveal how exactly LumA and LumB DCCs differ in their cell cycle behavior.

6.3.4 LumA and LumB DCCs display different overall expression profiles

Building on the observed correlation of the KI67 level of the PT with the PC1 variable (see chapter 4.5.4), we identified genes, which were associated with either the positive (LumB up-regulated) or negative (LumB down-regulated) direction of PC1. With the resulting lists, we performed a GO term analysis.

Many terms resulting from the LumB down-regulated genes were linked to membranes, the extracellular matrix (ECM), cell adhesion, and ion channels (see Figure 4-32 and Figure 4-34). This suggests that the communication of LumB DCCs with the ME may be different than that of LumA DCCs, which may permit stronger proliferation or reduce the selective pressure on the cells. A further possibility is that the changes to the ECM-related genes are associated with the alleged higher proliferation of LumB DCCs, as it is known that the ECM undergoes architectural changes during tumor cell proliferation, e.g. increased secretion of fibronectin and several collagens, through the interactions of the ME and resident cells (Malik et al., 2015; reviewed by Walker et al., 2018). Alternatively, it may also mean that LumB DCCs use a different way to prime a pre-metastatic niche (Chin and Wang, 2016). However, more research is needed to clarify, whether the observed changes in LumB DCCs' communication with their surroundings are inherent traits that were present before the dissemination or whether the changes in the DCCs may be induced by the pre-metastatic niche and represent an adaptation to the foreign ME. This might be done by studying circulating cancer cells (CCC), which have not been exposed to the BM niche yet. Magbanua and colleagues did this by comparing DCC pools to CCC pools isolated with the same method (Magbanua et al., 2018b; Magbanua et al., 2018a). They found that DCCs and CCCs represented two distinct populations of cancer cells with different gene expression patterns

(Magbanua et al., 2018a), which indicates that CCCs that settle down in a target organ (whereupon they are referred to as DCCs), undergo a change in their gene expression profile to adapt to the new ME. However, their study was limited to the genes included in their panel qPCR and more importantly by the fact that the CCCs were not collected from the same patients as the DCCs, which may account for much of the difference in gene expression. Others have studied the clinical and prognostic relevance of CCCs compared to DCCs (Schindlbeck et al., 2013; Rack et al., 2014; reviewed by Banys-Paluchowski et al., 2015) or the divergence of molecular features of CCCs from those of the corresponding PT (Boral et al., 2017), but to my knowledge there is currently no study that performed a comprehensive comparison of molecular features of matched DCCs and CCCs.

The GO terms resulting from genes up-regulated in LumB were mostly linked to a few common themes, namely ribosomes and translation, cell adhesion, ECM and exosomes, as well as splicing (see Figure 4-33 and Figure 4-35). The increase in genes linked to ribosome biogenesis and translation complements the previous finding that LumB DCCs displayed a higher expression of cell cycle-associated genes (see Figure 4-30), since the presumably higher proliferative activity requires an increase in protein production. The presence of cell adhesion-related genes among both the down-regulated and up-regulated genes in LumB DCCs suggests a complex interaction of the genes that are responsible for the same process, ultimately leading to a change of cell adhesion in LumB DCCs which likely also involves crosstalk with the ECM or neighboring cells, which is known to be crucial for survival of DCCs in the BM (Kaplan et al., 2006). However, it is unclear whether these changes led to an overall increase or decrease in cell adhesion compared to LumA DCCs. Similarly, ECM-related genes also contained some that were up- and some that were down-regulated. Therefore, the difference between the two luminal subtypes regarding the interaction of DCCs with the ECM is rather complex. Additionally, LumB DCCs also displayed an enrichment of splicing-related genes. Consequently, it seems that LumB DCC-specific characteristics involve an interaction with the ME, which is somehow associated with changes in splicing and proliferation. This is not surprising as Weaver and colleagues have demonstrated that the ECM is able to revert malignant BC cells back to a normal phenotype, indicating that ECM effects are dominant over a tumor cell's genotype (Weaver et al., 1997). This also means that tiny changes in the ME may exert significant effects on the proliferation of cancer cells (Walker et al., 2018). Lastly, one GO term called "negative regulation of apoptotic process" also appeared among the genes up-regulated in LumB (Figure 4-33), which suggests that apoptosis resistance may already have developed to some extent in LumB DCCs. Surprisingly, this was the only GO term directly linked to apoptosis, which may suggest that this hallmark of cancer was still under development, when the DCCs were isolated.

Overall, both the up- and down-regulated genes were associated with very heterogenous GO terms and, so far, we do not know what exactly this means. Therefore, more research will be necessary to clarify the complex interactions between DCCs and the ME which were indicated by the data. A first step for follow-up studies could be to examine whether the genes belonging to the splicing GO category are known for interactions with the ECM. If so, the respective genes may represent good targets to investigate the interaction of LumB DCCs with the ECM.

Finally, ten promising candidate genes (see Table 4-20) were selected for future investigations based on their low q-values and high fold changes as well as their functions and – in some cases – prognostic relevance in cancer. Hopefully, they may provide novel insights into the mechanisms governing the higher malignancy of LumB BC. Among these genes are the LumB down-regulated genes *ANKRD13B, CD6, FAT4, IFNL2*, and *TIRAP,* as well as the up-regulated genes *CTTN, DKC1, FEM1B, MORF4L2,* and *SMAD1.* Out of these ten, three are of special interest. The first two, *CD6* and *MORF4L2*, are known prognostic markers in BC (see Table 4-21), whose direction of regulation also fits its prognostic direction (*CD6* is a favorable marker and down-regulated, while

MORF4L2 is unfavorable and up-regulated) and which have been tested by immune histochemistry, which makes them promising candidates for playing a role in the higher metastatic propensity of LumB BC. Furthermore, *FEM1B* should also be highlighted, because – although it is not known to be prognostic in any cancer – it was already successfully validated as differentially expressed by qPCR (see Figure 4-36). Moreover, *FEM1B* is of interest, because it is a member of the death receptor protein family that associates with the other family members *Fas* and *tumor necrosis factor 1* (*TNF1*). Although it is known as a pro-apoptotic protein in colon cancer (Chan et al., 1999; Subauste et al., 2009; Lei et al., 2016), proteins often fulfill different – even opposite - functions depending on the cell type that expresses them, as was just recently shown for the tumor suppressor *p53*, which - contrary to its usual function - seems to promote tumor growth in liver cancer (Kim et al., 2019). Therefore, it is possible that *FEM1B* is somehow involved in negative regulation of apoptosis in BC or exerts a yet undiscovered function. Unfortunately, there are no studies investigating *FEM1B*'s role in BC patients so far. However, one study found it up-regulated in a *BRCA1* mutated cell line (Privat et al., 2018). In conclusion, further studies will be necessary to fully elucidate its role in LumB BC DCCs.

6.4 Limitations of the study

The study of differences between LumA and LumB DCCs was mainly limited by the sample numbers, as there was only a single LumA patient in the M1 group (see chapter 6.3.2). Therefore, LumA and LumB could not be compared in metastatic patients.

Another major limitation of the study is the partial subjectivity of the LP-Seq analysis. While the LP-Seq itself and the generation of CNA profiles was straight forward and easily done, the interpretation of the M0 patient-derived profiles was not. Similar to mCGH, the algorithms responsible for calling gains and losses are prone to errors. Therefore, each aberration needed to be carefully examined to decide whether it was real or just an artifact. By looking at all available profiles, I tried to gain an impression which small and frequently occurring aberrations were likely to be artifacts, in order to exclude those later in the analysis. Unfortunately, unlike in M1 patient-derived EpCAM$^+$ cells, there were only few recurring aberrations in the M0 patient-derived EpCAM$^+$ cells, which complicated the classification of the M0 cells, as individual small gains or losses appeared seemingly randomly across all chromosomes for each individual cell. For this reason, it was not clear whether these were of biological origin or merely technical artifacts. Future studies might improve this situation by modifying the selection criteria for true aberrations. For example, in a recent study Zhou and colleagues have performed whole genome sequencing with an average coverage of 1.5x (compared to ~0.5x in the current study) and used a cutoff size of 10 Mb to filter artifacts (Zhou et al., 2019) in contrast to the one megabase used in the current study. Furthermore, the decision-making process would benefit massively from a comparison of LP-Seq-derived CNA profiles from different cancer entities, because it would provide information on which of the few recurring aberrations are real and specific to BC and which are just technical noise. Without such knowledge, several small, but real CNAs may have been wrongly excluded as artifacts. Unfortunately, no such comparison had been done at the time the data were analyzed, so a certain degree of uncertainty remained in the classification. In an attempt to overcome this uncertainty and subjective impressions of the profiles, all profiles were discussed with three experienced postdocs. Afterwards, the consensus was used as the final classification for each cell's profile. Nevertheless, the classification of each profile as aberrant or balanced remained subjective to some extent, which is why all cells, which could not be classified into the aberrant or balanced groups without any doubts, were excluded.

Moreover, the bioinformatic pipeline provided by Menarini Silicon Biosystems for LP-Seq analysis used the hg37 version of the human genome for generation of the CNA profiles, while the cytoband annotation script used the hg38 version for the cytoband annotation. This resulted in a few hundred missing cytoband values (1.9 % of all genomic loci) for genomic loci which did not exist in the older reference used for generation of RefSeq files and profiles. In these cases, the cytoband information of the closest previous annotated locus was utilized to fill the gap. At the time the script was written, it was not known was causing the missing values and the mismatch in the genome versions was only noticed after the whole CNA analysis was already finished. However, as the deviation introduced by the different versions was considered marginal and would not have changed the results in a noticeable way, we decided to refrain from repeating the tedious manual CNA annotation procedure with RefSeq files re-annotated using the hg37 version of the human genome. Nevertheless, this issue represents a small limitation of the CNA analysis and needs to be kept in mind for future studies to avoid repeating this mistake.

Furthermore, the comparison of LP-Seq and mCGH needs to be assessed carefully, because the small sample numbers give more weight to small artifacts or less relevant real aberrations relative to the major aberrations, which are described in the literature (Kallioniemi et al., 1994; Nishizaki et al., 1997; Hermsen et al., 1998; Tirkkonen et al., 1998; Buerger et al., 1999; Roylance et al., 1999; Buerger et al., 2001; Haunschild, 2013). As discussed above, this problem could be reduced in future studies by a comparative analysis of CNAs observed across different cancer entities.

Another big limitation of this study is the lack of functional proof, since all results are based on genome and transcriptome analyses. For example, the result that the DCCs were proliferating requires proof that the gene expression really translates to a corresponding phenotype, since the measured expression levels of proliferation markers (see chapter 4.4.2) and cell cycle-associated genes (see Figure 4-30) must not necessarily translate to the same level of protein expression and a proliferating phenotype. This lack of functional proof also includes the non-proliferating and proliferating clusters identified by RNA-Seq (Figure 4-30), therefore the proliferation-related results need to be interpreted carefully, until functional proof has been provided. This gap in our knowledge will be closed by a new study investigating label retaining and non-label retaining cells as well as quiescent stem cells derived from cultured healthy mammary tissue isolated by another PhD student of our chair (Grujovic, 2019). These cells were isolated according to their phenotypes, which were either non-proliferating (label retaining) or proliferating (non-label retaining) as parts of mammospheres or single label retaining cells (quiescent stem cells). Sequencing of these three cell types and comparison with the expression profiles of the DCCS may provide a possibility to link the DCCs to one of these phenotypes. Nevertheless, the high concordance of qPCR and RNA-Seq (Figure 4-30) suggests that the proliferation data are valid, despite the lack of functional proof.

A further drawback of the current study is that there were not enough M0 EpCAM$^+$ cells that had been tested for genomic aberrations by mCGH for the RNA-Seq experiment and that the LP-Seq data were not yet available at that point. Therefore, the selection of additional M0 DCCs required usage of the M0 DCC qPCR signature (see chapter 4.2.1), which still requires proof that it is equal to or better than CNA profiling, to identify enough DCCs for the RNA-Seq. As there were only four mCGH-aberrant LumB DCCs compared to eight LumA ones (see Table 4-19), the majority of cells, which had to be selected according to the M0 DCC signature, belonged to the LumB subtype. Due to this higher uncertainty regarding the identity of the majority of LumB cells, twelve instead of the intended ten LumB cells were included in the RNA-Seq library preparation process, to increase the chances of having ten true DCCs in the end. Retrospective classification of all included cells using the LP-Seq profiles (after data analysis on all samples was already finished) revealed that only two out of eight LumB cells selected based on the M0 DCC signature had an aberrant genome, while two were balanced and the remaining four profiles were of too low quality for analysis.

Therefore, there were six LumB and eight LumA DCCs with confirmed aberrations (mCGH and LP-Seq taken together). However, the unclear status of four LumB cells meant there was still a chance that these cells were actually aberrant, because the qPCR signature had shown they were expressing the DCC pattern. Nevertheless, the uncertain CNA status and the inclusion of two balanced LumB cells needs to be considered during data interpretation.

Furthermore, there were some caveats regarding the RNA-Seq data. The bioinformatic QC discovered that the average yield of reads was lower than calculated, but there were still plenty of reads for all samples to enable proper analysis (see Figure 4-25). Additionally, there were also a few outliers regarding the number of reads. However, the high amount of duplicate reads relative to unique reads (see Figure 4-25) is considered normal for RNA-Seq experiments, because algorithms are unable to distinguish natural duplications (high gene expression) from PCR-induced artifacts. Therefore, duplications did not need to be removed, as this may only have worsened data quality (Parekh et al., 2016). The observed Phred scores were excellent across all samples indicating a low error rate in the base calling process (Figure 4-26), while the GC content per sequence was only good or sufficient for two thirds of samples (Figure 4-27). The remaining third displayed uneven distribution. The most likely reason for this is that the QC was performed prior to the trimming of reads, which means that our WTA adaptors were still included and may therefore have caused the observed bias in the GC content. Lastly, the observed shape of the read coverage profiles (Figure 4-28 and Figure 4-29) may have been caused by long mRNAs, because the longer the transcript the higher the chance of the mRNA strand being broken. Thereby the 5' ends of some transcripts may not have been captured. Regarding the small 3' bias we do not have a robust explanation as to why this occurs, because the WTA utilizes oligo(dT) primers, which should be able to fully capture the 3' ends of transcripts. The observed bias might indicate truncated transcripts, which are captured by the random octamer primers, or such transcripts that naturally occur without a poly(A) tail. The eWTA experiments have shown that such transcripts can be passively carried over to the final amplification despite the absence of a poly(A) tail (see Figure 5-13).

6.5 Conclusion

The M0 DCC qPCR signature was shown to be able to distinguish M0 DCCs from NCCs with sufficient accuracy. So far, CNA profiling is still the gold standard for reliable identification of NCCs among the isolated M0 EpCAM$^+$ cells. However, since it is much more expensive than qPCR, future research should investigate the false positive results of the M0 DCC signature more closely, in order to answer the question whether the signature is more precise at identification of true DCCs than CNA profiling. This would represent a major improvement in the detection of eDCCs, because it would reduce both cost and workload for this task. Furthermore, CNA profiling revealed that M1 DCCs contained significantly more genomic aberrations than M0 DCCs. This difference between M0 and M1 groups applied to DCCs from both the EpCAM$^+$ and CK$^+$ collectives, but EpCAM$^+$ and CK$^+$ from the same metastatic group differed only marginally, suggesting that EpCAM$^+$ and CK$^+$ cells represent similar populations of cells. Moreover, LumA and LumB DCCs could only be differentiated by global transcriptome profiling, as the two subtypes were similar in their rates of EpCAM-positivity, number of EpCAM$^+$ cells isolated per patient, expression of proliferation markers measured by qPCR, and genomic aberrations. The proliferation-related RNA-Seq results provide useful hints regarding the higher propensity of LumB BC to metastasize, however they do not permit any conclusions about the higher frequency of therapy resistance that has been reported for this subtype (Szostakowska et al., 2019). In order to obtain data on this characteristic of the LumB subtype, future studies should take a closer look at the non-proliferating DCCs, because these may be the ones surviving treatment, as chemotherapy targets only proliferating

cells. Moreover, RNA-Seq analysis identified more than 1000 differentially expressed genes between the LumA and LumB subtypes and a specific analysis of cell cycle-associated genes indicated a slightly increased expression of proliferation-related genes in LumB DCCs.

Overall, the presented data provide evidence for the early dissemination and parallel progression models advocated by our group (Klein, 2008, 2009). The following points speak in favor

- The M0 DCC signature worked very poorly on M1 DCCs, indicating a strong difference in gene expression between these two populations of cancer cells. This difference may be caused by somatic progression.
- M0 DCCs carried fewer aberrations than the M1 DCCs, which also implies occurrence of somatic progression.
- M0 DCCs displayed a higher proliferation rate than M1 DCCs, which suggest that the DCCs isolated from patient BM have disseminated early and require the high proliferation rates to acquire more mutations to facilitate survival.
- LumA and LumB DCCs did not differ in their CNA profiles, despite reported differences in PTs. Therefore, DCCs likely disseminated early.
- Three aberrant EpCAM$^+$ cells, which were expressing the M0 DCC signature, were detected in elderly healthy donors without overt disease.

Taken together, the results of this work are more consistent with the parallel progression model than the late dissemination model. This suggests that the former represents a more accurate description of biological reality than the latter. However, more evidence is necessary in order to support this claim.

There are at least two possible scenarios (see Figure 6-1) that may be consistent with the observed GO categories resulting from LumB up- and down-regulated genes. These scenarios could be investigated in future studies. One would be that LumB DCCs receive signals from the ME that induce changes in splicing, which in turn lead to more proliferation accompanied by increased ribosome biogenesis and translation. An alternative possibility is that differences in splicing variants in LumB DCCs cause a different response to the ME, which finally increases proliferation compared to LumA DCCs. Most likely the biological reality is a complex mixture of several processes. More research will be required to shed light onto the crosstalk between LumB DCCs and the ME of the BM, in order to facilitate development of better treatment of this BC subtype.

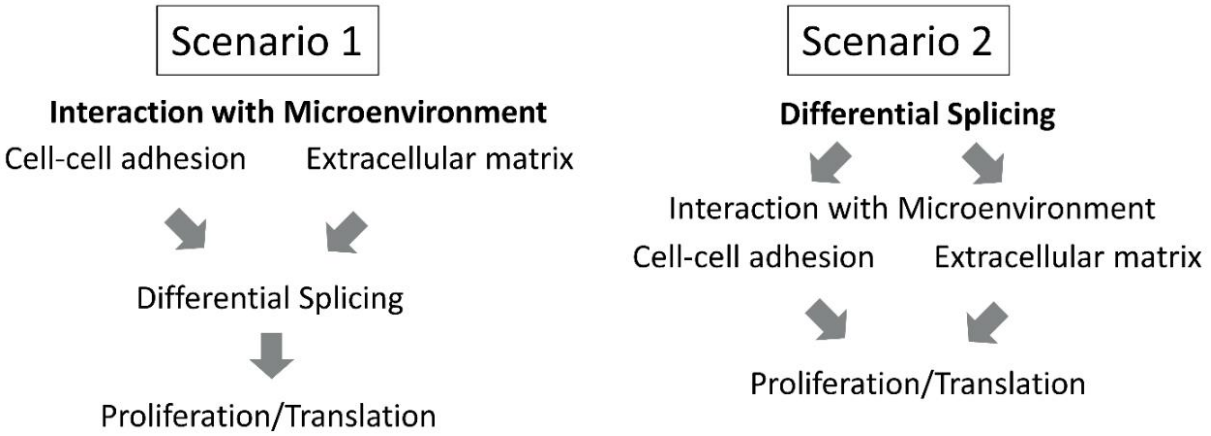


Figure 6-1 Proposed differences in biological processes in LumB BC DCCs compared to LumA DCCs. Both hypothetical scenarios lead to concomitant increases in proliferation and translation, but these increases are initiated in different ways. In scenario one, LumB DCCs carry inherent differences in their composition of cell-cell or cell-ECM interacting proteins, which cause a different reaction to the ME and lead to altered splicing of mRNAs, which in turn increases proliferation and translation. In scenario 2, inherent differential splicing in LumB DCCs leads to a different interaction with the BM ME and finally to increased proliferation and translation.

7. Discussion of method development for isolation of the miRNAome from single cells

Micro RNAs (miRNA) play important roles in BC (Iorio et al., 2005; Dedes et al., 2011; Fardmanesh et al., 2016; Vaidyanathan et al., 2016; Poursadegh Zonouzi et al., 2017; Mandujano-Tinoco et al., 2018). Specifically, they are associated with development of treatment resistance, the different molecular subtypes (Lv et al., 2014; Kurozumi et al., 2017), and can function as prognostic markers (Yerukala Sathipati and Ho, 2018). However, there are currently only three published methods for sequencing of miRNAs from single cells (SCs; Faridani et al., 2016; Lee et al., 2017; Wang et al., 2019) and none of them offers the possibility to isolate both the genome and transcriptome of the same cell along with the miRNA, which could provide important insights into the biology of the extremely rare disseminated cancer cells (DCC) studied in our lab. Therefore, the aim of this part of the current study was to find a way to adapt the existing WTA (see chapter 3.2.1) to facilitate isolation of the miRNAome along with transcriptome and genome.

To achieve this goal, a polyadenylation step in combination with rRNA blocking was added to our established WTA protocol (see Figure 5-1). So far, a preliminary version of this extended WTA (eWTA) protocol has been developed. Using artificial spike-in RNA, it was demonstrated that the polyadenylation step works in principle, however the rRNA blocking did not have any effect. To our surprise, the data also indicated that this step might not even be necessary.

7.1 Stepwise discussion of eWTA development

The following subchapters discuss the rationale for each new step and each introduced change compared to the sWTA in the order they appear in the eWTA protocol (see chapter 5.6). The whole eWTA procedure is summarized in Figure 7-1 at the end of this chapter.

7.1.1 Changes to existing steps - lysis buffer, picking, protease treatment

The qPCR measurements of rRNA levels suggested that the Igepal-containing Poly(A) polymerase (PAP) buffer, which was supposed to be used for cell lysis and polyadenylation, was likely too mild for cell lysis with only 0.1 % Igepal (see chapter 5.1.3). Others have reported good results with 0.25-0.3 % Igepal in lysis protocols for downstream RT-qPCR (Shatzkes et al., 2014; Le et al., 2015). RT-qPCR is of course not directly comparable to our WTA, yet, it supported the conclusion that the cell lysis may not have worked as intended. Additionally, it lacked a denaturing component like the guanidinium thiocyanate (GTC), which is contained in the mTRAP lysis buffer used in the regular WTA. However, GTC was initially believed to be too harsh for PAP to function (see “WTA 12” experiment in Table 5-2). Therefore, a custom buffer containing urea and different detergents was tested.

Although the final custom buffer with urea worked (see Figure 5-10), it was not used for the proposed eWTA protocol, because it caused new problems through its temperature sensitivity (see last paragraph of chapter 5.2.1.1). Nevertheless, the buffer experiments provided valuable insights into the lysis process, which may prove useful in future method development. More importantly, the experiments showed that PAP was surprisingly resilient to denaturing conditions, which prompted testing whether the enzyme would also work in diluted GTC-containing mTRAP buffer. A series of experiments revealed that a 1:5 dilution of the cell lysate was sufficient to facilitate optimal function of PAP (see chapter 5.2.1.2). In order to save reagents

for this dilution, the volume of lysis buffer, in which a cell is deposited after picking, was decreased from 4 µl to 3 µl (see Figure 5-12). Additionally, the *E.coli* tRNA, which was previously added to the lysis buffer to coat the tube and protect the RNA from RNases, was replaced by SUPERase. This was done, because the reaction tubes used nowadays are already pre-coated to prevent sticking of nucleic acids and because the bacterial tRNAs would have been polyadenylated and unintentionally captured. The tRNA was replaced by SUPERase to rule out RNA degradation as a potential reason for bad experimental results so that evaluation of the results would be less complicated. However, it was never tested in the final eWTA protocol, whether SUPERase is necessary to maintain RNA integrity. Consequently, optimization experiments should investigate whether this expensive reagent is really necessary.

Next, the PNAs were removed from the subsequent protease lysis reaction, because they were meant to be annealed after polyadenylation of the miRNAs. As a result, the PNA annealing step of the lysis cycler program was moved to a later point in the protocol. Additionally, the lack of PNAs and the reduced initial lysis buffer volume would have increased the protease concentration in the protease digestion. As a consequence, the amount of mTRAP buffer in the lysis mix was increased to maintain the protease concentration used in the sWTA (see Table 5-10).

7.1.2 Discarded new step - blocking

As rRNAs represent 80-90 % of cellular RNA (O'Neil et al., 2013; Wu et al., 2014), two sets of blocking oligonucleotides (Dr. Lieb's four ZNA and Dr. Pai's 113 DNA oligonucleotides) were tested to prevent rRNA polyadenylation. Regarding the annealing of the blocking oligonucleotides, Dr. Lieb's preliminary experiments revealed that annealing worked best in SCs by heating to 75 °C and gradually cooling down to 55 °C (see chapter 5.1.2). Although this was only shown with her own oligonucleotides, the same cycler program was also applied to Dr. Pai's DNA oligonucleotides later on, no major differences were expected due to the similar length of the oligonucleotides. However, no extensive testing was performed on this topic.

Unfortunately, no significant effect of either oligonucleotide set was observed (see Figure 5-23). Luckily, the passive rRNA contamination, which frequently occurred despite selection for poly(A) tailed transcripts (for example see Figure 5-15), did not turn out to be a big problem. Curiously, the polyadenylation increased the amount of captured rRNA by only one to two qPCR cycles for the *28S* and *18S* rRNAs and by roughly three qPCR cycles for the less abundant *5.8S* rRNA, while the mitochondrial rRNAs 16S and 12S as well as mRNAs were completely unaffected by the polyadenylation (see Figure 5-23). After the polyadenylation, there was only a difference of about five qPCR cycles between the most abundant *18S* rRNA and the assessed mRNAs *RAB7A* and *KRT8*, which corresponds to a 32-fold difference in expression. This difference was surprisingly low considering that a few different rRNAs account for up to 90 % (O'Neil et al., 2013) of a cell's total RNA, while thousands of different mRNAs represent only ~4 % of total RNA (Wu et al., 2014). This result suggested that the rRNA contamination was not as bad as was initially assumed. In addition, the RNA-Seq data of patient-derived DCCs revealed that the majority of cells only contained a little over 4 % of rRNA reads on average (see chapter 5.4.1). Since these samples were prepared using the standard WTA, the percentage of rRNA reads is expected to increase between two- and four-fold in case of the eWTA. However, this increased amount of rRNA reads can be overcome by increasing the sequencing depth to still be able to detect less abundant mRNA transcripts. Due to the aforementioned results and the lack of effect of the blocking (see Figure 5-21, Figure 5-22, and Figure 5-23), the blocking step was dropped from the proposed eWTA, in order to save valuable time and resources.

7.1.3 New step - polyadenylation and PNA annealing

Without the blocking step, the poly(A) tailing is performed directly after the protease lysis. For the polyadenylation it was important to increase the total volume of the reaction mix to 15 µl, in order to achieve the necessary 1:5 dilution of the mTRAP buffer contained in the cell lysate. The incubation temperature of 37 °C for the polyadenylation was adopted from the manufacturer's recommendations for the PAP, while the incubation time of 30 min was taken from Verena Lieb's experiments.

In order to deactivate the PAP after the polyadenylation without heating the solution, a high amount of GTC is added together with the PNAs. The concentration of the added GTC was chosen so that the final concentration would match the GTC level of the cell lysate in the sWTA going into the mRNA capture step, in order to keep the conditions as similar to the sWTA as possible. The amount of PNAs added to the solution was increased from 2 µl to 3 µl to account for the polyadenylated miRNAs, however the necessary amount of PNAs has not been tested systematically and should be assessed in future studies. Finally, the PNAs are annealed using the part of the lysis cycler program that was previously skipped (see last paragraph of chapter 7.1.1).

The presented experiments consistently showed that the poly(A) tailing procedure worked in principle. First and foremost, the spike-in experiments with the artificial *Long fragment (LF)* RNA confirmed that the protocol was able to detect exogenous sequences (see Figure 5-14). Furthermore, the spike-in experiments also confirmed that the undesired passive carry-over was not restricted to rRNAs only, but also occurred with the *LF* RNA in a few samples, therefore this may be a general problem of the bead-based capture strategy (see Figure 5-13 and Figure 5-14). Apart from the spike-in experiments, the blocking experiments consistently demonstrated that the polyadenylation increased the amount of detected rRNAs by one to three cycles depending on the type of rRNA and the experiment (see Figure 5-22 and Figure 5-23). This means that the amount of rRNA was increased two- to eight-fold in the eWTA compared to the sWTA, in which the rRNAs are usually depleted by around 96 % due to the poly(A)-specific capture strategy. This increase in rRNA was surprisingly low and suggests that the polyadenylation is still generally inefficient and that polyadenylation of rRNA seems to be specifically inefficient. However, this will be assessed by a future study using a customized RNA-Seq approach.

Unfortunately, I could not provide evidence that the eWTA can detect real miRNAs from SCs, as this would have required development of a separate custom protocol that could have been used as a readout, exceeding the scope of this dissertation. The reason for this is that our WTA adapters are not compatible with commercial miRNA detection kits or sequencing protocols. Therefore, this challenge remains to be overcome by a future study.

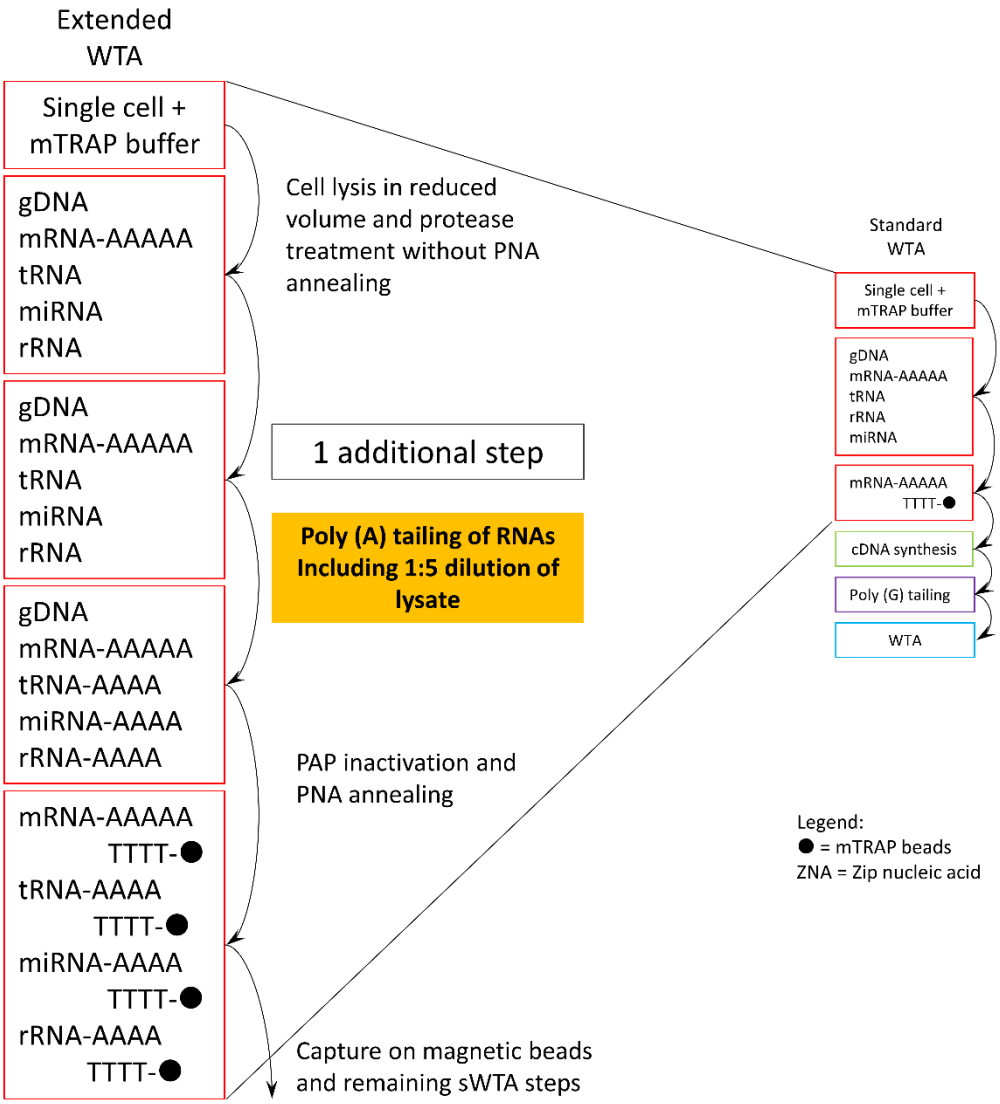


Figure 7-1 Schematic overview of proposed eWTA protocol. The scheme illustrates the preliminary eWTA protocol described in chapter 5.6 with special emphasis on the newly introduced step (left) compared the sWTA (on the right).

7.2 Advantages and disadvantages of the eWTA

The first and biggest advantage of the eWTA is that it – once established – would be one of very few triple-omic methods in combination with our WGA and the only one with the option of isolating gDNA, mRNA, and miRNA from a single cell. As mentioned previously (see chapter 1.6), the existing methods for global profiling of miRNAs from SCs are either single- or dual-omics approaches, with Small-Seq focusing on small RNAs (sRNA) only (Faridani et al., 2016; Hagemann-Jensen et al., 2018), while the other two – single tube amplification (STA) and half-cell sequencing (hcSeq) - allow parallel sequencing of miRNA and mRNA (Lee et al., 2017; Wang et al., 2019). In contrast, our current WTA/WGA combination allows profiling of gDNA and mRNA of the same cell, but misses miRNAs. While there is already at least one triple-omics technology for SCs called scTrio-seq that measures genome, transcriptome, and DNA methylome (Hou et al., 2016), however, scTrio-Seq is not relevant for us, since it does not allow isolation of miRNAs, therefore, it is not discussed further. In contrast, the novel eWTA protocol presented here would be the first triple-omic technology to be able to profile genome, transcriptome, and miRNAome. A drawback of the eWTA approach is that it has a low throughput and is hard to automate. An overview of the mentioned ncRNA-targeting methods is provided below (Table 7-1).

Table 7-1 Comparison of single cell-based global ncRNA profiling technologies. Based on (Faridani et al., 2016; Lee et al., 2017; Hagemann-Jensen et al., 2018; Wang et al., 2019) for small-Seq, STA, and hcSeq.

Method	Advantages	Disadvantages
Protocol as described here (eWTA/WGA)	• Triple-omic (gDNA, mRNA, ncRNA) • Universal capture of ncRNAs by poly(A) enrichment • Platform approach for multiple downstream applications • Fast and easy reamplification	• Depends on sequencing as a readout • Low throughput, because difficult to automate • 3' bias due to poly(A) enrichment strategy, mostly compensated by random octamer primers
STA	• Universal capture of ncRNAs by poly(A) enrichment • Pre-amplification makes it a potential platform approach • Simple workflow	• Dual-omic only (mRNA and ncRNA) • 3' bias due to poly(A) enrichment strategy • No random primers to compensate 3' bias
Small-Seq	• Suitable for automation due to skipping of size selection • Unique molecular identifier enables counting of sRNA transcripts per cell to remove PCR-induced bias • Uses standard reagents and devices	• Single-omic only (sRNA only) • Ligation-based bias • Streamlined, single-purpose process • Low throughput as of yet
hcSeq	• Throughput can be increased by microfluidics • Half-cell approach avoids material loss and technical variation	• Dual-omic only (mRNA and sRNA) • Ligation-based bias • Streamlined, single-purpose process • Low throughput as of yet

The second advantage of the eWTA is the poly(A) enrichment strategy of the WTA. Throughout previous chapters the eWTA was only referred to as a means for isolation of miRNAs, however, I want to stress that the polyadenylation and poly(A) enrichment approach enables inclusion of several kinds of non-coding RNAs (ncRNA), which is not possible with adapter ligation-based approaches that might miss some species of small RNAs. These consist of – to name the most prominent ones – miRNAs, PIWI-interacting RNAs (piRNA), and small interfering RNA (siRNA; Farazi et al., 2008). In addition, we may also be able to detect other ncRNAs in the form of transfer RNAs (tRNA) or fragments thereof, which have also been shown to play a role in disease, including BC (Goodarzi et al., 2015; Kirchner and Ignatova, 2015), and large intergenic non-coding RNAs (lincRNA), which exist in both poly(A) tailed and un-tailed form (Esteller, 2011; Zhang et al., 2014). The poly(A) enrichment approach of the eWTA is able to capture more of the ncRNAs than the adapter ligation-based small-Seq and hcSeq, because some ncRNAs can contain 2'-O-methyl modifications at their 3' end, which prevent adapter ligation (Dard-Dascot et al., 2018). Moreover, ligation-based methods also suffer from biases towards certain sequence compositions located at the ends of sRNAs (Hafner et al., 2011; Zhuang et al., 2012). Therefore, the more general poly(A) enrichment strategy seems to be better suited to capture as many ncRNAs as possible together with the mRNA. The 3' bias that is introduced by this strategy is compensated to a large extent by the use of random octamer primers, however this does not completely prevent it as was previously demonstrated (see Figure 4-28). The only other method that can match this versatility is Lee and colleagues' STA, as it also employs poly(A) tailing to capture ncRNAs (Lee et al., 2017).

Apart from the two points discussed above, the eWTA offers another advantage. Small-Seq and hcSeq were designed as streamlined processes to directly prepare SCs for sequencing without focusing on generating a useful intermediate product that could be used for other downstream applications. To do this the methods would have to be adapted. In contrast, our eWTA/WGA technology intentionally serves as a platform that turns each cell into a versatile intermediate product that allows a multitude of downstream analysis methods and can be stored over long periods of time. Additionally, our reamplification protocol (see chapter 3.5) enables a fast and easy amplification of the primary material, which exponentially increases the number of analyses that can be done with a single cell. This is especially useful for precious patient samples and rare cells like DCCs.

However, the multitude of possible downstream analyses is accompanied by some problems. As the eWTA/WGA approach generates no actual data by itself, it depends strongly on other methods, which often need to be customized, because the WTA adapters are not always compatible with downstream assays. This applies especially to the quantification of miRNAs, as commercially available miRNA quantification methods usually rely on the presence of universal adapters introduced during reverse transcription (Lunn et al., 2008). As our WTA products are already reverse transcribed, our samples would require whole miRNAome sequencing or customized PCR applications.

7.3 Conclusion

This study demonstrates that the newly implemented polyadenylation strategy of the eWTA and the overall protocol works in principle and that artificial transcripts can be detected after polyadenylation. However, it remains to be shown that the eWTA can actually detect endogenous miRNA from real SCs. Once the final proof has been provided, the protocol requires a comprehensive optimization before being applied to patient samples.

The final, optimized eWTA/WGA strategy will be suited for research on precious, rare samples like DCCs, because it is designed to maximize the amount of data that can be obtained from each individual cell by isolating the genome, transcriptome, and miRNAome. Furthermore, our approach is designed to preserve the primary material for as long as possible by means of reamplification. Additionally, the universal polyadenylation approach may even facilitate study of yet unknown types of ncRNAs, because such RNAs may already have been captured by the eWTA as long as the respective RNA can be polyadenylated.

Taken together, the eWTA method will enable the study of DCCs at an unprecedented level, as it opens up new possibilities to study the intricate crosstalk between the genome, it's messengers and regulators thereof.

8. Summary

Breast cancer (BC) accounts for almost a quarter of reported cancer incidences in women worldwide. It comprises several molecular subtypes, out of which the luminal A (LumA) and luminal B (LumB) types are the most common. Despite being quite similar from a histopathological point of view, the LumB type has a far worse prognosis due to a higher propensity to metastasize and develop therapy resistance. Despite decades of effort, metastasis is still responsible for ~90 % of cancer-related deaths. Metastatic relapse is caused by disseminated cancer cells (DCC) that can lie dormant in distant organs for several years before growing into macrometastases. The epithelial cell adhesion molecule (EpCAM) can be used as a marker to detect DCCs in patient bone marrow (BM). However, there are non-cancer cells (NCC) belonging to the erythroid progenitor lineage in the BM, which also express this marker, thereby representing a confounding factor in our $EpCAM^+$ single cell collective. Regarding treatment resistance, recent studies have implicated several miRNAs in resistance of BC to the most common systemic treatments.

The aim of this dissertation was (1) to identify a way to distinguish true DCCs from NCCs, (2) to identify genomic or transcriptomic differences between LumA and LumB DCCs accounting for LumB's increased malignancy, and (3) to develop a novel extended whole transcriptome amplification (eWTA) to isolate miRNAs along with mRNA and gDNA from single cells to further elucidate the underlying causes for LumB BC's higher aggressiveness.

The data revealed that separation of NCCs from true DCCs was possible using an M0 DCC qPCR signature consisting of four genes, but the distinction was most reliably done using copy number alteration (CNA) profiling. However, the data suggested that the qPCR signature might actually be more precise that the CNA profiling. A detailed analysis of the CNA profiles of true DCCs revealed no differences between LumA and LumB DCCs. Additionally, targeted proliferation marker analyses by qPCR did not reveal differences between LumA and LumB DCCs, neither regarding expression levels nor regarding the percentage of proliferating cells. In contrast to the comparison of LumA and LumB DCCs, there was a pronounced divergence of CNAs in DCCs derived from non-metastatic (M0) and metastatic (M1) BC patients with the latter carrying more genomic aberrations compared to the former. In line with the CNA profiles, targeted proliferation marker analyses showed a difference between M0 and M1 DCCs with the M1 DCCs displaying a lower percentage of proliferating cells. Interestingly, qPCR revealed that half of M0 cells were proliferating. However, there was no correlation between the KI67 status of the primary tumor (PT) and the expression of *MKI67* in matched DCCs.

Subsequent global transcriptomic profiling by RNA-Sequencing confirmed that all of the DCCs, which were classified as proliferating by qPCR, were expressing cell cycle-associated genes significantly more than the non-proliferating DCCs. Unsupervised hierarchical clustering identified two subgroups among the proliferating DCCs with differing expression of analyzed genes. These groups comprised either a mix of DCCs from all subtypes, or almost exclusively LumB DCCs, suggesting a higher proliferation of LumB DCCs. The RNA-Seq analysis also uncovered a correlation of the overall expression signature of DCCs with the KI67 status of the PT, which indirectly translated to differences between LumA and LumB subtypes in their overall gene expression. Gene ontology (GO) analysis identified several biological processes that were enriched among the up- and down-regulated genes in LumB compared to LumA. Genes related to membranes and transmembrane transport were associated with down-regulated genes, while splicing, ribosomes, and translation were overrepresented among the up-regulated genes. Cell adhesion and extracellular matrix (ECM) pathways were present in both gene lists, indicating a complex differential regulation of these processes in LumB compared to LumA.

In parallel to the previous experiments, a preliminary protocol for the novel eWTA was established. This protocol included a polyadenylation step to enable capture of single stranded RNAs. The polyadenylation required introduction of an additional dilution of the cell lysate, in order to prevent denaturation of the Poly(A) polymerase by the lysis buffer. Experiments on artificial long RNA as well as expression changes of ribosomal RNAs demonstrated that the employed polyadenylation strategy worked in principle. However, a detection of short RNAs in the range of miRNA could not be done, as this would have required development of a separate method compatible with our WTA adapters. Nevertheless, the first step towards an eWTA for isolation of the miRNAome alongside the genome and transcriptome of a single cell has been taken.

In conclusion, the data suggest that the main factors driving the increased malignancy of LumB cancer is likely a higher proliferation, because it enables faster accumulation of somatic mutations. Two hypothetical scenarios explaining the underlying mechanisms are the following: (1) LumB DCCs may interact with the microenvironment (ME) at metastatic target sites in a different manner than LumA DCCs, which leads to changes in mRNA splicing, which in turn increases proliferation. Or, (2) LumB DCCs may initially display differential mRNA splicing before arriving at the target site, causing altered interaction with the ME at the target site and in response proliferation is up-regulated in the DCCs. More research will be required to provide functional proof of the higher proliferation of LumB DCCs and to determine the exact mechanisms underlying this alleged higher proliferation profile of LumB DCCs. It is also possible that miRNAs are somehow involved in this. Therefore, it will be important to further develop the eWTA protocol, in order to aid in advancing our knowledge of the intricate crosstalk of miRNAs with their target mRNAs and concomitant genomic changes.

9. Zusammenfassung

Brustkrebs (BC) ist weltweit für fast ein Viertel aller Krebserkrankungen in Frauen verantwortlich. BC setzt sich aus mehreren molekularen Subtypen zusammen, von denen luminal A (LumA) und luminal B (LumB) die häufigsten sind. Trotz der starken histopathologischen Ähnlichkeit dieser Subtypen ist die Prognose bei LumB Patientinnen weitaus schlechter, was auf eine höhere Tendenz zur Metastasierung und Entwicklung von Therapieresistenzen zurückzuführen ist. Trotz jahrzehntelanger Anstrengungen verursachen Metastasen immer noch ~90 % aller krebsbedingten Todesfälle. Metastatische Rezidive werden von disseminierten Krebszellen (DCC), die zunächst in entfernten Organen schlummern bevor sie zu Makrometastasen anwachsen, verursacht. Zur Detektion von DCC im Knochenmark (BM) von Patienten kann das epitheliale Zelladhäsionsmolekül (EpCAM) als Erkennungsmerkmal verwendet werden. Jedoch existiert im BM auch eine EpCAM-exprimierende Population von nicht-Krebszellen (NCC), die zur Abstammungslinie der erythroiden Vorläuferzellen gehört und somit unser Zellkollektiv verunreinigt. Bezüglich der Therapieresistent konnten Studien kürzlich nachweisen, dass miRNAs an der Entwicklung von Resistenzen gegen viele systemische Standardbehandlungsmethoden beteiligt sind.

Das Ziel dieser Dissertation war es (1) echte DCCs von NCCs zu unterscheiden, (2) genomische oder transkriptomische Unterschiede zwischen LumA und LumB DCCs, die für die höhere Malignität von LumB verantwortlich sind, zu identifizieren und (3) eine erweiterte globale Transkriptomamplifikation (eWTA) zu entwickeln, die die gemeinsame Isolation von miRNA, mRNA und gDNA aus einzelnen Zellen ermöglicht. Diese neue Methode sollte einen detaillierteren Einblick in die zugrundeliegenden Ursachen der höheren Aggressivität von LumB BC gewähren.

Die Daten zeigten, dass eine Unterscheidung von NCCs und echten DCCs durch eine aus vier Genen bestehende M0 DCC Signatur mittels qPCR möglich war, eindeutige Ergebnisse jedoch nur durch die Analyse von Kopienzahlveränderungen (CNA) im Genom erzielt werden konnten. Allerdings wiesen die Daten darauf hin, dass die qPCR Signatur eventuell sogar genauer sein könnte als die CNA-Analyse. Eine tiefergehende CNA-Analyse der echten DCCs förderte keine signifikanten Unterschiede zwischen LumA und LumB zutage. Außerdem ergab eine gezielte Analyse von Proliferationsmarkern mittels qPCR ebenfalls keine Unterschiede zwischen den beiden Subtypen, weder im Expressionsniveau noch in der Expressionsfrequenz. Im Gegensatz dazu fanden sich in DCCs von nicht-metastatischen (M0) Patientinnen insgesamt weit weniger CNAs als in DCCs von metastatischen (M1) Patientinnen. Analog dazu ergab die gezielte Analyse von Proliferationsmarkern eine weitaus niedrigere Frequenz proliferierender Zellen in der Gruppe der M1 DCCs im Vergleich zu den M0 DCCs. Interessanterweise zeigte die qPCR, dass die Hälfte der M0 Zellen proliferierte und es keine Korrelation zwischen dem KI67 Status des Primärtumors (PT) und der Expression von *MKI67* in den gepaarten DCCs gab.

Eine nachfolgende globale Transkriptomanalyse mittels RNA-Sequenzierung bestätigte, dass alle DCCs, welche zuvor mittels qPCR als proliferierend klassifiziert wurden, eine Vielzahl von Zellzyklus-assoziierten Genen signifikant stärker exprimierten als nicht-proliferierende DCCs. Eine unbeaufsichtigte Clustering-Analyse identifizierte innerhalb der proliferierenden DCCs zwei Untergruppen mit unterschiedlicher Expressionsstärke der betrachteten Gene. Diese Gruppen bestanden entweder aus einer Mischung von DCCs aller BC Subtypen oder beinahe ausschließlich aus LumB DCCs, was auf eine höhere Proliferationsaktivität der LumB DCCs hindeutete. Die Sequenzierungsdaten enthüllten außerdem eine Korrelation der Gesamtexpressionsprofile der DCCs mit dem KI67 Status der gepaarten PTs, was indirekt Unterschiede zwischen LumA und LumB Subtypen in ihren Gesamtexpressionsprofilen bedeutete. Eine Genontologie-Analyse enthüllte mehrere biologische Prozesse, die unter den hoch- und herunter-regulierten Genen in

LumB im Vergleich zu LumA DCCs angereichert waren. Gene, die allgemein mit Membranen und Transmembrantransport assoziiert sind, fanden sich unter den herunterregulierten Genen, während Splicing, Ribosomen und Translation unter den hochregulierten Genen überrepräsentiert waren. Gene, die in Verbindung mit Zelladhäsion und der extrazellulären Matrix (ECM) stehen, kamen sowohl in den herunter- als auch den hochregulierten Genen vor, was auf eine komplexe differenzielle Regulation dieser Prozesse in LumB im Vergleich zu LumA DCCs hindeutet.

Parallel zu den beschriebenen Experimenten wurde ein vorläufiges Protokoll für die neue eWTA etabliert. Hierzu wurde eine Polyadenylierung in das Protokoll eingefügt, welche das Einfangen einzelsträngiger RNAs ermöglichen sollte. Die Polyadenylierung wurde durch die zusätzliche Einführung einer Verdünnung des Zelllysats ermöglicht, um eine Inaktivierung der Poly(A) Polymerase durch den Lysepuffer zu verhindern. Experimente mit einer künstlichen, langen RNA sowie Expressionsveränderungen endogener ribosomaler RNAs zeigten, dass die angewandte Polyadenylierungsstrategie prinzipiell funktionierte. Jedoch steht ein Funktionsnachweis an kurzen RNAs, ähnlich der Länge der miRNAs, noch aus, da solch ein Nachweis die Entwicklung einer separaten Methode zur Messung dieser RNAs erfordert hätte, da existierende Methoden nicht mit unseren WTA Adaptoren kompatibel sind. Nichtsdestotrotz stellt die vorläufige eWTA einen ersten Schritt in Richtung der Isolation des miRNAoms zusammen mit Genom und Transkriptom von Einzelzellen dar.

Zusammengenommen legen die Daten dieser Studie nahe, dass der Hauptfaktor für die höhere Malignität von LumB BC vermutlich eine höhere Proliferationsrate ist, da diese eine zügigere Anhäufung von somatischen Mutationen ermöglicht. Anhand der Daten sind zwei hypothetische Szenarien, die zu dieser höheren Proliferationsrate führen, denkbar: (1) LumB DCCs könnten anders auf die Mikroumgebung in metastatischen Zielorganen reagieren als LumA DCCs, wodurch zunächst Veränderungen im mRNA Splicing induziert werden, was dann schließlich zu einer erhöhten Proliferation führt. (2) Alternativ könnten LumB DCCS sich bereits vor ihrer Ankunft im Zielorgan in einem unterschiedlichen Splicing-Zustand befinden, wodurch sie anders mit der Mikroumgebung interagieren und als Antwort darauf stärker proliferieren als LumA DCCs. Es bedarf jedoch weiterer Forschungsarbeiten, um die exakten zugrundeliegenden Mechanismen der mutmaßlich stärkeren Proliferation von LumB DCCs zu bestimmen. Es ist durchaus denkbar, dass miRNAs zumindest teilweise daran beteiligt sind, weshalb es wichtig sein wird das vorläufige eWTA Protokoll weiter zu entwickeln, um ein tiefgreifendes Verständnis der feinen Interaktionen von miRNAs mit ihren Zieltranskripten und einhergehenden genomischen Veränderungen zu ermöglichen.

10. References

Ades, F., Zardavas, D., Bozovic-Spasojevic, I., Pugliano, L., Fumagalli, D., Azambuja, E. de, Viale, G., Sotiriou, C., and Piccart, M. (2014). Luminal B breast cancer: molecular characterization, clinical management, and future perspectives. Journal of clinical oncology: official journal of the American Society of Clinical Oncology; ISSN: 0732-183X *32*, 2794-2803.

Adiconis, X., Borges-Rivera, D., Satija, R., DeLuca, D.S., Busby, M.A., Berlin, A.M., Sivachenko, A., Thompson, D.A., Wysoker, A., and Fennell, T., et al. (2013). Comparative analysis of RNA sequencing methods for degraded or low-input samples. Nature methods *10*, 623-629.

Aguirre-Ghiso, J.A. (2007). Models, mechanisms and clinical evidence for cancer dormancy. Nat Rev Cancer *7*, 834-846.

Alba, E., Calvo, L., Albanell, J., La Haba, J.R. de, Arcusa Lanza, A., Chacon, J.I., Sanchez-Rovira, P., Plazaola, A., Lopez Garcia-Asenjo, J.A., and Bermejo, B., et al. (2012). Chemotherapy (CT) and hormonotherapy (HT) as neoadjuvant treatment in luminal breast cancer patients: results from the GEICAM/2006-03, a multicenter, randomized, phase-II study. Annals of oncology: official journal of the European Society for Medical Oncology *23*, 3069-3074.

Amin, M.B., Greene, F.L. and Edge, S.B. (2017). AJCC cancer staging manual (Schweiz, Chicago, IL: Springer; AJCC American Joint Committee on Cancer).

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Androvic, P., Valihrach, L., Elling, J., Sjoback, R., and Kubista, M. (2017). Two-tailed RT-qPCR: a novel method for highly accurate miRNA quantification. Nucleic acids research *45*, e144.

Angermueller, C., Clark, S.J., Lee, H.J., Macaulay, I.C., Teng, M.J., Hu, T.X., Krueger, F., Smallwood, S., Ponting, C.P., and Voet, T., et al. (2016). Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. Nature methods *13*, 229-232.

Assi, R., Mukherji, D., Haydar, A., Saroufim, M., Temraz, S., and Shamseddine, A. (2016). Metastatic colorectal cancer presenting with bone marrow metastasis: a case series and review of literature. Journal of gastrointestinal oncology *7*, 284-297.

Banys, M., Krawczyk, N., and Fehm, T. (2014). The role and clinical relevance of disseminated tumor cells in breast cancer. Cancers *6*, 143-152.

Banys-Paluchowski, M., Schneck, H., Blassl, C., Schultz, S., Meier-Stiegen, F., Niederacher, D., Krawczyk, N., Ruckhaeberle, E., Fehm, T., and Neubauer, H. (2015). Prognostic Relevance of Circulating Tumor Cells in Molecular Subtypes of Breast Cancer. Geburtshilfe und Frauenheilkunde *75*, 232-237.

Baran-Gale, J., Kurtz, C.L., Erdos, M.R., Sison, C., Young, A., Fannin, E.E., Chines, P.S., and Sethupathy, P. (2015). Addressing Bias in Small RNA Library Preparation for Sequencing: A New Protocol Recovers MicroRNAs that Evade Capture by Current Methods. Frontiers in genetics *6*, 352.

Barkan, D., El Touny, L.H., Michalowski, A.M., Smith, J.A., Chu, I., Davis, A.S., Webster, J.D., Hoover, S., Simpson, R.M., and Gauldie, J., et al. (2010a). Metastatic growth from dormant cells induced by a col-I-enriched fibrotic environment. Cancer Res *70*, 5706-5716.

Barkan, D., Green, J.E., and Chambers, A.F. (2010b). Extracellular matrix: a gatekeeper in the transition from dormancy to metastatic growth. European journal of cancer (Oxford, England : 1990) *46*, 1181-1188.

Barkan, D., Kleinman, H., Simmons, J.L., Asmussen, H., Kamaraju, A.K., Hoenorhoff, M.J., Liu, Z.-y., Costes, S.V., Cho, E.H., and Lockett, S., et al. (2008). Inhibition of metastatic outgrowth from single dormant tumor cells by targeting the cytoskeleton. Cancer Res *68*, 6241-6250.

Barnes, B., Kraywinkel, K., Nowossadeck, E., Schönfeld, I., Starker, A., Wienecke, A. and Wolf, U. (2016). Bericht zum Krebsgeschehen in Deutschland 2016 (Robert Koch-Institut).

Baudis, M., and Cleary, M.L. (2001). Progenetix.net: an online repository for molecular cytogenetic aberration data. Bioinformatics (Oxford, England) *17*, 1228-1229.

Bendre, M., Gaddy, D., Nicholas, R.W., and Suva, L.J. (2003). Breast cancer metastasis to bone: it is not all about PTHrP. Clinical orthopaedics and related research, S39-45.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological) *57*, 289-300.

Bentz, M., Plesch, A., Stilgenbauer, S., Döhner, H., and Lichter, P. (1998). Minimal sizes of deletions detected by comparative genomic hybridization. Genes, chromosomes & cancer *21*, 172-175.

Berman, A.T., Thukral, A.D., Hwang, W.-T., Solin, L.J., and Vapiwala, N. (2013). Incidence and patterns of distant metastases for patients with early-stage breast cancer after breast conservation treatment. Clin Breast Cancer *13*, 88-94.

Bertz, J., Dahm, S., Haberland, J., Kraywinkel, K., Kurth, B., and Wolf, U. (2010). Verbreitung von Krebserkrankungen in Deutschland. Entwicklung der Prävalenzen zwischen 1990 und 2010. Robert Koch Institut.

Bhaskaran, M., and Mohan, M. (2014). MicroRNAs: history, biogenesis, and their evolving role in animal development and disease. Veterinary pathology *51*, 759-774.

Boral, D., Vishnoi, M., Liu, H.N., Yin, W., Sprouse, M.L., Scamardo, A., Hong, D.S., Tan, T.Z., Thiery, J.P., and Chang, J.C., et al. (2017). Molecular characterization of breast cancer CTCs associated with brain metastasis. Nature Communications *8*, 196.

Braun, S., Vogl, F.D., Naume, B., Janni, W., Osborne, M.P., Coombes, R.C., Schlimok, G., Diel, I.J., Gerber, B., and Gebauer, G., et al. (2005). A pooled analysis of bone marrow micrometastasis in breast cancer. The New England journal of medicine *353*, 793-802.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians *68*, 394-424.

Brierley, J., Gospodarowicz, M.K. and Wittekind, C. (2017). TNM classification of malignant tumours (Chichester, West Sussex, UK, Hoboken, NJ: John Wiley & Sons Inc).

Buerger, H., Mommers, E.C., Littmann, R., Simon, R., Diallo, R., Poremba, C., Dockhorn-Dworniczak, B., van Diest, P.J., and Boecker, W. (2001). Ductal invasive G2 and G3 carcinomas of the breast are the end stages of at least two different lines of genetic evolution. The Journal of pathology *194*, 165-170.

Buerger, H., Otterbach, F., Simon, R., Schfer, K.-L., Poremba, C., Diallo, R., Brinkschmidt, C., Dockhorn-Dworniczak, B., and Boecker, W. (1999). Different genetic pathways in the evolution of invasive breast cancer are associated with distinct morphological subtypes. The Journal of pathology *189*, 521-526.

Bühring, H.J., Müller, T., Herbst, R., Cole, S., Rappold, I., Schuller, W., Zhu, X., Fritzsch, U., Faul, C., and Armeanu, S., et al. (1996). The adhesion molecule E-cadherin and a surface antigen recognized by the antibody 9C4 are selectively expressed on erythroid cells of defined maturational stages. Leukemia *10*, 106-116.

Buonomo, O.C., Caredda, E., Portarena, I., Vanni, G., Orlandi, A., Bagni, C., Petrella, G., Palombi, L., and Orsaria, P. (2017). New insights into the metastatic behavior after breast cancer surgery, according to well-established clinicopathological variables and molecular subtypes. PLoS One *12*, e0184680.

Bushnell, B. (2014). BBMap: a fast, accurate, splice-aware aligner (Lawrence Berkeley National Lab, Berkeley, CA (United States), available online at: https://sourceforge.net/projects/bbmap/).

Buson, G., Tononi, P., Forcato, C., Fontana, F., Medoro, G., Neves, R., Möhlendick, B., Stoecklein, N.H., and Manaresi, N. (2016). Abstract 2394: Scalable, rapid and affordable low-pass whole genome sequencing method for single-cell copy-number profiling on LM-PCR based WGA products. Cancer Res *76*, 2394.

Calin, G.A., Sevignani, C., Dumitru, C.D., Hyslop, T., Noch, E., Yendamuri, S., Shimizu, M., Rattan, S., Bullrich, F., and Negrini, M., et al. (2004). Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers. Proc Natl Acad Sci U S A *101*, 2999-3004.

Campos-Parra, A.D., Mitznahuatl, G.C., Pedroza-Torres, A., Romo, R.V., Reyes, F.I.P., López-Urrutia, E., and Pérez-Plasencia, C. (2017). Micro-RNAs as Potential Predictors of Response to Breast Cancer Systemic Therapy: Future Clinical Implications. International journal of molecular sciences *18*.

Cancer Genome Atlas Network (2012). Comprehensive molecular portraits of human breast tumours. Nature *490*, 61-70.

Cardoso, F., and Castiglione, M. (2009). Locally recurrent or metastatic breast cancer: ESMO clinical recommendations for diagnosis, treatment and follow-up. Annals of oncology : official journal of the European Society for Medical Oncology *20 Suppl 4*, 15-18.

Carter, P., Presta, L., Gorman, C.M., Ridgway, J.B., Henner, D., Wong, W.L., Rowland, A.M., Kotts, C., Carver, M.E., and Shepard, H.M. (1992). Humanization of an anti-p185HER2 antibody for human cancer therapy. Proc Natl Acad Sci U S A *89*, 4285-4289.

Chambers, A.F., Groom, A.C., and MacDonald, I.C. (2002). Dissemination and growth of cancer cells in metastatic sites. Nature Reviews Cancer *2*, 563-572.

Chan, S.L., Tan, K.O., Zhang, L., Yee, K.S., Ronca, F., Chan, M.Y., and Yu, V.C. (1999). F1Aalpha, a death receptor-binding protein homologous to the Caenorhabditis elegans sex-determining protein, FEM-1, is a caspase substrate that mediates apoptosis. J Biol Chem *274*, 32461-32468.

Cheang, M.C.U., Chia, S.K., Voduc, D., Gao, D., Leung, S., Snider, J., Watson, M., Davies, S., Bernard, P.S., and Parker, J.S., et al. (2009). Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer. Journal of the National Cancer Institute *101*, 736-750.

Chi, K.R. (2013). Singled out for sequencing. Nature methods *11*, 13 EP -.

Chin, A.R., and Wang, S.E. (2016). Cancer Tills the Premetastatic Field: Mechanistic Basis and Clinical Implications. Clin Cancer Res *22*, 3725-3733.

Cho, E.S., Kang, H.E., Kim, N.H., and Yook, J.I. (2019). Therapeutic implications of cancer epithelial-mesenchymal transition (EMT). Archives of Pharmacal Research *42*, 14-24.

Choy, J.Y.H., Boon, P.L.S., Bertin, N., and Fullwood, M.J. (2015). A resource of ribosomal RNA-depleted RNA-Seq data from different normal adult and fetal human tissues. Scientific data *2*, 150063.

Cicco, P. de, Catani, M.V., Gasperi, V., Sibilano, M., Quaglietta, M., and Savini, I. (2019). Nutrition and Breast Cancer: A Literature Review on Prevention, Treatment and Recurrence. Nutrients *11*.

Clark, S.J., Argelaguet, R., Kapourani, C.-A., Stubbs, T.M., Lee, H.J., Alda-Catalinas, C., Krueger, F., Sanguinetti, G., Kelsey, G., and Marioni, J.C., et al. (2018). scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. Nature Communications *9*, 781.

Cole, M.P., Jones, C.T.A., and Todd, I.D.H. (1971). A New Anti-oestrogenic Agent in Late Breast Cancer: An Early Clinical Appraisal of ICI46474. British journal of cancer *25*, 270-275.

Coleman, R., Boer, R. de, Eidtmann, H., Llombart, A., Davidson, N., Neven, P., Minckwitz, G. von, Sleeboom, H.P., Forbes, J., and Barrios, C., et al. (2013). Zoledronic acid (zoledronate) for postmenopausal women with early breast cancer receiving adjuvant letrozole (ZO-FAST study): final 60-month results. Annals of oncology : official journal of the European Society for Medical Oncology *24*, 398-405.

Collins, V.P., Loeffler, R.K., and Tivey, H. (1956). Observations on growth rates of human tumors. The American journal of roentgenology, radium therapy, and nuclear medicine *76*, 988-1000.

Cote, R.J., Rosen, P.P., Lesser, M.L., Old, L.J., and Osborne, M.P. (1991). Prediction of early relapse in patients with operable breast cancer by detection of occult bone marrow micrometastases. Journal of clinical oncology: official journal of the American Society of Clinical Oncology; ISSN: 0732-183X *9*, 1749-1756.

Creighton, C.J. (2012). The molecular profile of luminal B breast cancer. Biologics : targets & therapy *6*, 289-297.

Curtis, C., Shah, S.P., Chin, S.-F., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D., Lynch, A.G., Samarajiwa, S., and Yuan, Y., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature *486*, 346 EP -.

Dai, X., Li, T., Bai, Z., Yang, Y., Liu, X., Zhan, J., and Shi, B. (2015). Breast cancer intrinsic subtype classification, clinical use and future trends. American Journal of Cancer Research *5*, 2929-2943.

Dard-Dascot, C., Naquin, D., d'Aubenton-Carafa, Y., Alix, K., Thermes, C., and van Dijk, E. (2018). Systematic comparison of small RNA library preparation protocols for next-generation sequencing. BMC genomics *19*, 118.

Dedes, K.J., Natrajan, R., Lambros, M.B., Geyer, F.C., Lopez-Garcia, M.A., Savage, K., Jones, R.L., and Reis-Filho, J.S. (2011). Down-regulation of the miRNA master regulators Drosha and Dicer is associated with specific subgroups of breast cancer. European journal of cancer (Oxford, England: 1990) *47*, 138-150.

Dick, J.E. (2003). Breast cancer stem cells revealed. Proc Natl Acad Sci U S A *100*, 3547-3549.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics (Oxford, England) *29*, 15-21.

Domschke, C., Diel, I.J., Englert, S., Kalteisen, S., Mayer, L., Rom, J., Heil, J., Sohn, C., and Schuetz, F. (2013). Prognostic value of disseminated tumor cells in the bone marrow of patients with operable primary breast cancer: a long-term follow-up study. Annals of surgical oncology *20*, 1865-1871.

Durinck, S., Spellman, P.T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nature protocols *4*, 1184-1191.

Early Breast Cancer Trialists' Collaborative Group (2005). Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials. The Lancet *365*, 1687-1717.

Early Breast Cancer Trialists' Collaborative Group (2011). Relevance of breast cancer hormone receptors and other factors to the efficacy of adjuvant tamoxifen: patient-level meta-analysis of randomised trials. The Lancet *378*, 771-784.

Elston, C.W., and Ellis, I.O. (1991). Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. Histopathology *19*, 403-410.

Eroles, P., Bosch, A., Alejandro Pérez-Fidalgo, J., and Lluch, A. (2012). Molecular biology in breast cancer: Intrinsic subtypes and signaling pathways. Cancer Treatment Reviews *38*, 698-707.

Esteller, M. (2011). Non-coding RNAs in human disease. Nature reviews. Genetics *12*, 861-874.

Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics (Oxford, England) *32*, 3047-3048.

Ewing, B., and Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome research *8*, 186-194.

Ewing, B., Hillier, L., Wendl, M.C., and Green, P. (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome research *8*, 175-185.

Ewing, J. (1928). Neoplastic Diseases: A Treatise on Tumours. Third edition. p77–89.

Fabbri, M., Paone, A., Calore, F., Galli, R., Gaudio, E., Santhanam, R., Lovat, F., Fadda, P., Mao, C., and Nuovo, G.J., et al. (2012). MicroRNAs bind to Toll-like receptors to induce prometastatic inflammatory response. Proc Natl Acad Sci U S A *109*, E2110-6.

Fallahpour, S., Navaneelan, T., De, P., and Borgo, A. (2017). Breast cancer survival by molecular subtype: a population-based analysis of cancer registry data. CMAJ open *5*, E734-E739.

Farazi, T.A., Juranek, S.A., and Tuschl, T. (2008). The growing catalog of small RNAs and their association with distinct Argonaute/Piwi family members. Development (Cambridge, England) *135*, 1201-1214.

Fardmanesh, H., Shekari, M., Movafagh, A., Alizadeh Shargh, S., Poursadegh Zonouzi, A.A., Shakerizadeh, S., Poursadegh Zonouzi, A., and Hosseinzadeh, A. (2016). Upregulation of the double-stranded RNA binding protein DGCR8 in invasive ductal breast carcinoma. Gene *581*, 146-151.

Faridani, O.R., Abdullayev, I., Hagemann-Jensen, M., Schell, J.P., Lanner, F., and Sandberg, R. (2016). Single-cell sequencing of the small-RNA transcriptome. Nature biotechnology.

Felekkis, K., Touvana, E., Stefanou, C., and Deltas, C. (2010). microRNAs: a newly described class of encoded molecules that play a role in health and disease. Hippokratia *14*, 236-240.

Ferrarini, A., Buson, G., Bolognesi, C., Forcato, C., Tononi, P., Monaco, V.d., Mangano, C., Fontana, F., Medoro, G., and Manaresi, N. (2017). Abstract 3945: Precise copy-number profiling of single cells isolated from FFPE tissues by low-pass whole-genome sequencing. Cancer Res *77*, 3945.

Ferrarini, A., Forcato, C., Buson, G., Tononi, P., Del Monaco, V., Terracciano, M., Bolognesi, C., Fontana, F., Medoro, G., and Neves, R., et al. (2018). A streamlined workflow for single-cells genome-wide copy-number profiling by low-pass sequencing of LM-PCR whole-genome amplification products. PLoS One *13*, e0193689.

Fidler, I.J. (1970). Metastasis: quantitative analysis of distribution and fate of tumor emboli labeled with 125 I-5-iodo-2'-deoxyuridine. Journal of the National Cancer Institute *45*, 773-782.

Fidler, I.J. (2003). The pathogenesis of cancer metastasis: the 'seed and soil' hypothesis revisited. Nat Rev Cancer *3*, 453-458.

Finn, R.S., Dering, J., Conklin, D., Kalous, O., Cohen, D.J., Desai, A.J., Ginther, C., Atefi, M., Chen, I., and Fowst, C., et al. (2009). PD 0332991, a selective cyclin D kinase 4/6 inhibitor, preferentially inhibits proliferation of luminal estrogen receptor-positive human breast cancer cell lines in vitro. Breast cancer research : BCR *11*, R77.

Fisher, B., and Gebhardt, M.C. (1978). The evolution of breast cancer surgery: past, present, and future. Seminars in oncology *5*, 385-394.

Fitzpatrick, C., Bendek, M.F., Briones, M., Farfán, N., Silva, V.A., Nardocci, G., Montecino, M., Boland, A., Deleuze, J.-F., and Villegas, J., et al. (2019). Mitochondrial ncRNA targeting induces cell cycle arrest and tumor growth inhibition of MDA-MB-231 breast cancer cells through reduction of key cell cycle progression factors. Cell Death & Disease *10*, 423.

Floyd, R.W., Stone, M.P., and Joklik, W.K. (1974). Separation of single-stranded ribonucleic acids by acrylamide-agarose-urea gel electrophoresis. Analytical Biochemistry *59*, 599-609.

Foulds, L. (1954). The experimental study of tumor progression: a review. Cancer Res *14*, 327-339.

Francis, P.A., Regan, M.M., Fleming, G.F., Láng, I., Ciruelos, E., Bellet, M., Bonnefoi, H.R., Climent, M.A., Da Prada, G.A., and Burstein, H.J., et al. (2015). Adjuvant ovarian suppression in premenopausal breast cancer. The New England journal of medicine *372*, 436-446.

Franklin, M.C., Carey, K.D., Vajdos, F.F., Leahy, D.J., Vos, A.M. de, and Sliwkowski, M.X. (2004). Insights into ErbB signaling from the structure of the ErbB2-pertuzumab complex. Cancer cell *5*, 317-328.

Friberg, S., and Mattson, S. (1997). On the growth rates of human malignant tumors: implications for medical decision making. J Surg Oncol *65*, 284-297.

García-Alcalde, F., Okonechnikov, K., Carbonell, J., Cruz, L.M., Götz, S., Tarazona, S., Dopazo, J., Meyer, T.F., and Conesa, A. (2012). Qualimap: evaluating next-generation sequencing alignment data. Bioinformatics (Oxford, England) *28*, 2678-2679.

Gentile, M., Olsen, K., Dufmats, M., and Wingren, S. (1999). Frequent allelic losses at 11q24.1–q25 in young women with breast cancer: association with poor survival. British journal of cancer *80*, 843-849.

Gentile, M., Wiman, A., Thorstenson, S., Loman, N., Borg, A., and Wingren, S. (2001). Deletion mapping of chromosome segment 11q24-q25, exhibiting extensive allelic loss in early onset breast cancer. Int J Cancer *92*, 208-213.

Gianni, L., Pienkowski, T., Im, Y.-H., Roman, L., Tseng, L.-M., Liu, M.-C., Lluch, A., Staroslawska, E., La Haba-Rodriguez, J. de, and Im, S.-A., et al. (2012). Efficacy and safety of neoadjuvant pertuzumab and trastuzumab in women with locally advanced, inflammatory, or early HER2-positive breast cancer (NeoSphere): a randomised multicentre, open-label, phase 2 trial. The Lancet Oncology *13*, 25-32.

Giraldez, M.D., Spengler, R.M., Etheridge, A., Godoy, P.M., Barczak, A.J., Srinivasan, S., Hoff, P.L. de, Tanriverdi, K., Courtright, A., and Lu, S., et al. (2018). Comprehensive multi-center assessment of small RNA-seq methods for quantitative miRNA profiling. Nat Biotechnol *36*, 746 EP -.

Gnant, M., and Clézardin, P. (2012). Direct and indirect anticancer activity of bisphosphonates: a brief review of published literature. Cancer Treatment Reviews *38*, 407-415.

Gnant, M., Mlineritsch, B., Luschin-Ebengreuth, G., Kainberger, F., Kässmann, H., Piswanger-Sölkner, J.C., Seifert, M., Ploner, F., Menzel, C., and Dubsky, P., et al. (2008). Adjuvant endocrine therapy plus zoledronic acid in premenopausal women with early-stage breast cancer: 5-year follow-up of the ABCSG-12 bone-mineral density substudy. The Lancet Oncology *9*, 840-849.

Goldhirsch, A., Gelber, R.D., Piccart-Gebhart, M.J., Azambuja, E. de, Procter, M., Suter, T.M., Jackisch, C., Cameron, D., Weber, H.A., and Heinzmann, D., et al. (2013). 2 years versus 1 year of adjuvant trastuzumab for HER2-positive breast cancer (HERA): an open-label, randomised controlled trial. The Lancet *382*, 1021-1028.

Goldhirsch, A., Wood, W.C., Coates, A.S., Gelber, R.D., Thürlimann, B., and Senn, H.-J. (2011). Strategies for subtypes--dealing with the diversity of breast cancer: highlights of the St. Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2011. Annals of oncology : official journal of the European Society for Medical Oncology *22*, 1736-1747.

Gong, Y., Liu, Y.-R., Ji, P., Hu, X., and Shao, Z.-M. (2017). Impact of molecular subtypes on metastatic breast cancer patients: a SEER population-based study. Scientific reports *7*, 45411 EP -.

Goodarzi, H., Liu, X., Nguyen, H.C.B., Zhang, S., Fish, L., and Tavazoie, S.F. (2015). Endogenous tRNA-Derived Fragments Suppress Breast Cancer Progression via YBX1 Displacement. Cell *161*, 790-802.

Greenberg, P.A., Hortobagyi, G.N., Smith, T.L., Ziegler, L.D., Frye, D.K., and Buzdar, A.U. (1996). Long-term follow-up of patients with complete remission following combination chemotherapy for metastatic breast cancer. Journal of clinical oncology: official journal of the American Society of Clinical Oncology; ISSN: 0732-183X *14*, 2197-2205.

Grujovic, A. (2019). Regulation of cellular dormancy in disseminated breast cancer cells. Monography (Regensburg).

Gupta, G.P., and Massagué, J. (2006). Cancer metastasis: building a framework. Cell *127*, 679-695.

Gužvić, M., Braun, B., Ganzer, R., Burger, M., Nerlich, M., Winkler, S., Werner-Klein, M., Czyż, Z.T., Polzer, B., and Klein, C.A. (2014). Combined genome and transcriptome analysis of single disseminated cancer cells from bone marrow of prostate cancer patients reveals unexpected transcriptomes. Cancer Res *74*, 7383-7394.

Hafner, M., Renwick, N., Brown, M., Mihailović, A., Holoch, D., Lin, C., Pena, J.T.G., Nusbaum, J.D., Morozov, P., and Ludwig, J., et al. (2011). RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries. RNA (New York, N.Y.) *17*, 1697-1712.

Hagel, P., Gerding, J.J.T., Fieggen, W., and Bloemendal, H. (1971). Cyanate formation in solutions of urea: I. Calculation of cyanate concentrations at different temperature and pH. Biochimica et Biophysica Acta (BBA) - Protein Structure *243*, 366-373.

Hagemann-Jensen, M., Abdullayev, I., Sandberg, R., and Faridani, O.R. (2018). Small-seq for single-cell small-RNA sequencing. Nature protocols *13*, 2407-2424.

Han, K.Y., Kim, K.-T., Joung, J.-G., Son, D.-S., Kim, Y.J., Jo, A., Jeon, H.-J., Moon, H.-S., Yoo, C.E., and Chung, W., et al. (2018). SIDR: simultaneous isolation and parallel sequencing of genomic DNA and total RNA from single cells. Genome research *28*, 75-87.

Hanahan, D., and Weinberg, R.A. (2000). The Hallmarks of Cancer. Cell *100*, 57-70.

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

Harbeck, N., and Gnant, M. (2017). Breast cancer. The Lancet *389*, 1134-1150.

Harbeck, N., Untch, M., Pache, L., and Eiermann, W. (1994). Tumour cell detection in the bone marrow of breast cancer patients at primary therapy: results of a 3-year median follow-up. British journal of cancer *69*, 566-571.

Harper, M.J., and Walpole, A.L. (1967). A new derivative of triphenylethylene: effect on implantation and mode of action in rats. Journal of reproduction and fertility *13*, 101-119.

Hartkopf, A.D., Banys, M., Krawczyk, N., Wallwiener, M., Schneck, H., Neubauer, H., and Fehm, T. (2011). Circulating Tumor Cells in Early-Stage Breast Cancer. Geburtshilfe und Frauenheilkunde *71*, 1067-1072.

Hartkopf, A.D., Brucker, S.Y., Taran, F.-A., Harbeck, N., Au, A. von, Naume, B., Pierga, J.-Y., Hoffmann, O., Beckmann, M.W., and Rydén, L., et al. (2019). Abstract GS5-07: International pooled analysis of the prognostic impact of disseminated tumor cells from the bone marrow in early breast cancer: Results from the PADDY study. In General Session Abstracts (American Association for Cancer Research).

Hartmann, C.H., and Klein, C.A. (2006). Gene expression profiling of single cells on large-scale oligonucleotide arrays. Nucleic acids research *34*, e143.

Hashmi, A.A., Aijaz, S., Khan, S.M., Mahboob, R., Irfan, M., Zafar, N.I., Nisar, M., Siddiqui, M., Edhi, M.M., and Faridi, N., et al. (2018). Prognostic parameters of luminal A and luminal B intrinsic breast cancer subtypes of Pakistani patients. World Journal of Surgical Oncology *16*.

Haunschild, G. (2013). Nachweis und molekulare Charakterisierung EpCAM-positiver Zellen im Knochenmark von Mammakarzinom-Patientinnen. Monography (Regensburg).

He, S., Chu, J., Wu, L.-C., Mao, H., Peng, Y., Alvarez-Breckenridge, C.A., Hughes, T., Wei, M., Zhang, J., and Yuan, S., et al. (2013). MicroRNAs activate natural killer cells through Toll-like receptor signaling. Blood *121*, 4663-4671.

Hermsen, M.A.J.A., Baak, J.P.A., Meijer, G.A., Weiss, J.M., Walboomers, J.W.W., Snijders, P.J.F., and van Diest, P.J. (1998). Genetic analysis of 53 lymph node-negative breast carcinomas by CGH and relation to clinical, pathological, morphometric, and DNA cytometric prognostic factors. The Journal of pathology *186*, 356-362.

Herschkowitz, J.I., Simin, K., Weigman, V.J., Mikaelian, I., Usary, J., Hu, Z., Rasmussen, K.E., Jones, L.P., Assefnia, S., and Chandrasekharan, S., et al. (2007). Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. Genome biology *8*, R76.

Hofmeister, F. (1888). Zur Lehre von der Wirkung der Salze. Archiv f. experiment. Pathol. u. Pharmakol *24*, 247-260.

Holmgren, L., O'Reilly, M.S., and Folkman, J. (1995). Dormancy of micrometastases: balanced proliferation and apoptosis in the presence of angiogenesis suppression. Nat Med *1*, 149-153.

Hosseini, H., Obradovic, M.M.S., Hoffmann, M., Harper, K.L., Sosa, M.S., Werner-Klein, M., Nanduri, L.K., Werno, C., Ehrl, C., and Maneck, M., et al. (2016). Early dissemination seeds metastasis in breast cancer. Nature *540*, 552-558.

Hou, Y., Guo, H., Cao, C., Li, X., Hu, B., Zhu, P., Wu, X., Wen, L., Tang, F., and Huang, Y., et al. (2016). Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. Cell Res *26*, 304-319.

Howlader, N., Cronin, K.A., Kurian, A.W., and Andridge, R. (2018). Differences in Breast Cancer Survival by Molecular Subtypes in the United States. Cancer epidemiology, biomarkers & prevention: a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology *27*, 619-626.

Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009a). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic acids research *37*, 1-13.

Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature protocols *4*, 44-57.

Hudis, C.A. (2007). Trastuzumab--mechanism of action and use in clinical practice. The New England journal of medicine *357*, 39-51.

Hüsemann, Y., Geigl, J.B., Schubert, F., Musiani, P., Meyer, M., Burghart, E., Forni, G., Eils, R., Fehm, T., and Riethmüller, G., et al. (2008). Systemic spread is an early step in breast cancer. Cancer cell *13*, 58-68.

Inic, Z., Zegarac, M., Inic, M., Markovic, I., Kozomara, Z., Djurisic, I., Inic, I., Pupic, G., and Jancic, S. (2014). Difference between Luminal A and Luminal B Subtypes According to Ki-67, Tumor Size, and Progesterone Receptor Negativity Providing Prognostic Information. Clinical Medicine Insights. Oncology *8*, 107-111.

Iorio, M.V., Ferracin, M., Liu, C.-G., Veronese, A., Spizzo, R., Sabbioni, S., Magri, E., Pedriali, M., Fabbri, M., and Campiglio, M., et al. (2005). MicroRNA gene expression deregulation in human breast cancer. Cancer Res *65*, 7065-7070.

ISCN rules for listing chromosomal rearrangements (2001). Current protocols in human genetics *Appendix 4*, Appendix 4C.

Jayaprakash, A.D., Jabado, O., Brown, B.D., and Sachidanandam, R. (2011). Identification and remediation of biases in the activity of RNA ligases in small-RNA deep sequencing. Nucleic acids research *39*, e141.

Jeuken, J.W.M., Sprenger, S.H.E., and Wesseling, P. (2002). Comparative genomic hybridization: practical guidelines. Diagnostic molecular pathology : the American journal of surgical pathology, part B *11*, 193-203.

Kallioniemi, A., Kallioniemi, O.P., Piper, J., Tanner, M., Stokke, T., Chen, L., Smith, H.S., Pinkel, D., Gray, J.W., and Waldman, F.M. (1994). Detection and mapping of amplified DNA sequences in breast cancer by comparative genomic hybridization. Proc Natl Acad Sci U S A *91*, 2156-2160.

Kang, Y., and Pantel, K. (2013). Tumor cell dissemination: emerging biological insights from animal models and cancer patients. Cancer cell *23*, 573-581.

Kaplan, R.N., Psaila, B., and Lyden, D. (2006). Bone marrow cells in the 'pre-metastatic niche': within bone and beyond. Cancer metastasis reviews *25*, 521-529.

Keller, L., Werner, S., and Pantel, K. (2019). Biology and clinical relevance of EpCAM. Cell stress *3*, 165-180.

Kennecke, H., Yerushalmi, R., Woods, R., Cheang, M.C.U., Voduc, D., Speers, C.H., Nielsen, T.O., and Gelmon, K. (2010). Metastatic behavior of breast cancer subtypes. Journal of clinical oncology: official journal of the American Society of Clinical Oncology; ISSN: 0732-183X *28*, 3271-3277.

Kim, J., Yu, L., Chen, W., Xu, Y., Wu, M., Todorova, D., Tang, Q., Feng, B., Jiang, L., and He, J., et al. (2019). Wild-Type p53 Promotes Cancer Metabolic Switch by Inducing PUMA-Dependent Suppression of Oxidative Phosphorylation. Cancer cell *35*, 191-203.e8.

Kim, R.S., Avivar-Valderas, A., Estrada, Y., Bragado, P., Sosa, M.S., Aguirre-Ghiso, J.A., and Segall, J.E. (2012). Dormancy signatures and metastasis in estrogen receptor positive and negative breast cancer. PLoS One *7*, e35569.

Kirchner, S., and Ignatova, Z. (2015). Emerging roles of tRNA in adaptive translation, signalling dynamics and disease. Nature reviews. Genetics *16*, 98-112.

Kirkham, N., and Lemoine, N.R. (2001). Progress in pathology. p. 52 (London: Greenwich Medical Media).

Kittaneh, M., Montero, A.J., and Glück, S. (2013). Molecular profiling for breast cancer: a comprehensive review. Biomarkers in cancer *5*, 61-70.

Klein, C.A. (2003). The systemic progression of human cancer: a focus on the individual disseminated cancer cell--the unit of selection. Advances in cancer research *89*, 35-67.

Klein, C.A. (2008). Cancer. The metastasis cascade. Science (New York, N.Y.) *321*, 1785-1787.

Klein, C.A. (2009). Parallel progression of primary tumours and metastases. Nat Rev Cancer *9*, 302-312.

Klein, C.A. (2011). Framework models of tumor dormancy from patient-derived observations. Curr Opin Genet Dev *21*, 42-49.

Klein, C.A. (2013). Selection and adaptation during metastatic cancer progression. Nature *501*, 365-372.

Klein, C.A., Schmidt-Kittler, O., Schardt, J.A., Pantel, K., Speicher, M.R., and Riethmüller, G. (1999). Comparative genomic hybridization, loss of heterozygosity, and DNA sequence analysis of single cells. Proc Natl Acad Sci U S A *96*, 4494-4499.

Klein, C.A., Seidl, S., Petat-Dutter, K., Offner, S., Geigl, J.B., Schmidt-Kittler, O., Wendler, N., Passlick, B., Huber, R.M., and Schlimok, G., et al. (2002). Combined transcriptome and genome analysis of single micrometastatic cells. Nat Biotechnol *20*, 387-392.

Klein, C.A., Zohlnhöfer, D., Petat-Dutter, K., and Wendler, N. (2003). Gene expression analysis of a single or few cells. Current protocols in molecular biology *Chapter 25*, Unit 25B.8.

Klein, G. (1998). Foulds' dangerous idea revisited: the multistep development of tumors 40 years later. Advances in cancer research *72*, 1-23.

Kolde, R. (2019). pheatmap. Avalailable only at https://cran.r-project.org/web/packages/pheatmap/index.html.

Korthauer, K.D., Chu, L.-F., Newton, M.A., Li, Y., Thomson, J., Stewart, R., and Kendziorski, C. (2016). A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. Genome biology *17*, 222.

Kozłowski, J., Kozłowska, A., and Kocki, J. (2015). Breast cancer metastasis - insight into selected molecular mechanisms of the phenomenon. Postepy higieny i medycyny doswiadczalnej (Online) *69*, 447-451.

Kurozumi, S., Yamaguchi, Y., Kurosumi, M., Ohira, M., Matsumoto, H., and Horiguchi, J. (2017). Recent trends in microRNA research into breast cancer with particular focus on the associations between microRNAs and intrinsic subtypes. Journal of human genetics *62*, 15-24.

Lakhani, S.R., Ellis, I.O., Schnitt, S.J., Tan, P.H., and van Vijver, M.J. de (2012). WHO classification of the breast (Lyon: IARC Press).

Lammers, R., Giesert, C., Grünebach, F., Marxer, A., Vogel, W., and Bühring, H.-J. (2002). Monoclonal antibody 9C4 recognizes epithelial cellular adhesion molecule, a cell surface antigen expressed in early steps of erythropoiesis. Experimental Hematology *30*, 537-545.

Lamond, N.W., and Younis, T. (2014). Pertuzumab in human epidermal growth-factor receptor 2-positive breast cancer: clinical and economic considerations. International journal of women's health *6*, 509-521.

Le, A.V.-P., Huang, D., Blick, T., Thompson, E.W., and Dobrovic, A. (2015). An optimised direct lysis method for gene expression studies on low cell numbers. Scientific reports *5*, 12859.

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

Lee, Y.-H., Hsueh, Y.-W., Peng, Y.-H., Chang, K.-C., Tsai, K.-J., Sun, H.S., Su, I.-J., and Chiang, P.-M. (2017). Low-cell-number, single-tube amplification (STA) of total RNA revealed transcriptome changes from pluripotency to endothelium. BMC biology *15*, 22.

Lee, Y.S., and Dutta, A. (2009). MicroRNAs in cancer. Annual review of pathology *4*, 199-227.

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

Lei, J., Li, Q., Gao, Y., Zhao, L., and Liu, Y. (2016). Increased PKCα activity by Rack1 overexpression is responsible for chemotherapy resistance in T-cell acute lymphoblastic leukemia-derived cell line. Scientific reports *6*, 33717.

Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics (Oxford, England) *30*, 923-930.

Lim, E., Vaillant, F., Di Wu, Forrest, N.C., Pal, B., Hart, A.H., Asselin-Labat, M.-L., Gyorki, D.E., Ward, T., and Partanen, A., et al. (2009). Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. Nat Med *15*, 907 EP -.

Loberg, R.D., Bradley, D.A., Tomlins, S.A., Chinnaiyan, A.M., and Pienta, K.J. (2007). The lethal phenotype of cancer: the molecular basis of death due to malignancy. CA: a cancer journal for clinicians *57*, 225-241.

Lowry, R. (2004). VassarStats : website for statistical computation.

Lu, J., Steeg, P.S., Price, J.E., Krishnamurthy, S., Mani, S.A., Reuben, J., Cristofanilli, M., Dontu, G., Bidaut, L., and Valero, V., et al. (2009). Breast cancer metastasis: challenges and opportunities. Cancer Res *69*, 4951-4953.

Lu, M., Zhang, Q., Deng, M., Miao, J., Guo, Y., Gao, W., and Cui, Q. (2008). An analysis of human microRNA and disease associations. PLoS One *3*, e3420.

Lunn, M.-L., Mouritzen, P., Faber, K., and Jacobsen, N. (2008). MicroRNA quantitation from a single cell by PCR using SYBRÂ® Green detection and LNA-based primers. Nature methods *5*, A3 EP -.

Luzzi, K.J., MacDonald, I.C., Schmidt, E.E., Kerkvliet, N., Morris, V.L., Chambers, A.F., and Groom, A.C. (1998). Multistep Nature of Metastatic Inefficiency. The American Journal of Pathology *153*, 865-873.

Lv, J., Xia, K., Xu, P., Sun, E., Ma, J., Gao, S., Zhou, Q., Zhang, M., Wang, F., and Chen, F., et al. (2014). miRNA expression patterns in chemoresistant breast cancer tissues. Biomed Pharmacother *68*, 935-942.

Macaulay, I.C., Haerty, W., Kumar, P., Li, Y.I., Hu, T.X., Teng, M.J., Goolam, M., Saurat, N., Coupland, P., and Shirley, L.M., et al. (2015). G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. Nature methods *12*, 519 EP -.

Magbanua, M.J.M., Rugo, H.S., Hauranieh, L., Roy, R., Scott, J.H., Lee, J.C., Hsiao, F., Sosa, E.V., Van't Veer, L., and Esserman, L.J., et al. (2018a). Genomic and expression profiling reveal molecular heterogeneity of disseminated tumor cells in bone marrow of early breast cancer. NPJ breast cancer *4*, 31.

Magbanua, M.J.M., Rugo, H.S., Wolf, D.M., Hauranieh, L., Roy, R., Pendyala, P., Sosa, E.V., Scott, J.H., Lee, J.S., and Pitcher, B., et al. (2018b). Expanded Genomic Profiling of Circulating Tumor Cells in Metastatic Breast Cancer Patients to Assess Biomarker Status and Biology Over Time (CALGB 40502 and CALGB 40503, Alliance). Clin Cancer Res *24*, 1486-1499.

Malhotra, G.K., Zhao, X., Band, H., and Band, V. (2010). Histological, molecular and functional subtypes of breast cancers. Cancer biology & therapy *10*, 955-960.

Malik, R., Lelkes, P.I., and Cukierman, E. (2015). Biomechanical and biochemical remodeling of stromal extracellular matrix in cancer. Trends in biotechnology *33*, 230-236.

Malumbres, M., and Barbacid, M. (2009). Cell cycle, CDKs and cancer: a changing paradigm. Nature Reviews Cancer *9*, 153-166.

Malzahn, K., Mitze, M., Thoenes, M., and Moll, R. (1998). Biological and prognostic significance of stratified epithelial cytokeratins in infiltrating ductal breast carcinomas. Virchows Archiv : an international journal of pathology *433*, 119-129.

Mandujano-Tinoco, E.A., García-Venzor, A., Melendez-Zajgla, J., and Maldonado, V. (2018). New emerging roles of microRNAs in breast cancer. Breast Cancer Res Treat *171*, 247-259.

McCarthy, D.J., Campbell, K.R., Lun, A.T.L., and Wills, Q.F. (2017). Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. Bioinformatics (Oxford, England) *33*, 1179-1186.

Merritt, M.E., and Donaldson, J.R. (2009). Effect of bile salts on the DNA and membrane integrity of enteric bacteria. Journal of medical microbiology *58*, 1533-1541.

Morlan, J.D., Qu, K., and Sinicropi, D.V. (2012). Selective depletion of rRNA enables whole transcriptome profiling of archival fixed tissue. PLoS One *7*, e42882.

Morrison, B.J., Schmidt, C.W., Lakhani, S.R., Reynolds, B.A., and Lopez, J.A. (2008). Breast cancer stem cells: implications for therapy of breast cancer. Breast Cancer Research *10*, 210.

Nature methods (2013). Method of the Year 2013. Nature methods *11*, 1 EP -.

Nielsen, T.O., Hsu, F.D., Jensen, K., Cheang, M., Karaca, G., Hu, Z., Hernandez-Boussard, T., Livasy, C., Cowan, D., and Dressler, L., et al. (2004). Immunohistochemical and clinical characterization of the basal-like subtype of invasive breast carcinoma. Clin Cancer Res *10*, 5367-5374.

Nieto, M.A., Huang, R.Y.-J., Jackson, R.A., and Thiery, J.P. (2016). EMT: 2016. Cell *166*, 21-45.

Nishizaki, T., Chew, K., Chu, L., Isola, J., Kallioniemi, A., Weidner, N., and Waldman, F.M. (1997). Genetic alterations in lobular breast cancer by comparative genomic hybridization. Int. J. Cancer *74*, 513-517.

Ogba, N., Manning, N.G., Bliesner, B.S., Ambler, S.K., Haughian, J.M., Pinto, M.P., Jedlicka, P., Joensuu, K., Heikkilä, P., and Horwitz, K.B. (2014). Luminal breast cancer metastases and tumor arousal from dormancy are promoted by direct actions of estradiol and progesterone on the malignant cells. Breast cancer research : BCR *16*, 489.

O'Neil, D., Glowatz, H., and Schlumpberger, M. (2013). Ribosomal RNA depletion for efficient use of RNA-seq capacity. Current protocols in molecular biology *Chapter 4*, Unit 4.19.

O'Shaughnessy, J. (2005). Extending survival with chemotherapy in metastatic breast cancer. The oncologist *10 Suppl 3*, 20-29.

Pagani, O., Regan, M.M., Walley, B.A., Fleming, G.F., Colleoni, M., Láng, I., Gomez, H.L., Tondini, C., Burstein, H.J., and Perez, E.A., et al. (2014). Adjuvant exemestane with ovarian suppression in premenopausal breast cancer. The New England journal of medicine *371*, 107-118.

Pageau, S.C. (2009). Denosumab. mAbs *1*, 210-215.

Paget, S. (1889). The distribution of secondary growths in cancer of the breast. 1889. Cancer metastasis reviews *8*, 98-101.

Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M., Baehner, F.L., Walker, M.G., Watson, D., and Park, T., et al. (2004). A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. The New England journal of medicine *351*, 2817-2826.

Pantel, K., and Alix-Panabières, C. (2014). Bone marrow as a reservoir for disseminated tumor cells: a special source for liquid biopsy in cancer patients. BoneKEy reports *3*, 584.

Pantel, K., and Brakenhoff, R.H. (2004). Dissecting the metastatic cascade. Nature Reviews Cancer *4*, 448-456.

Paredes-Aracil, E., Palazón-Bru, A., La Folgado-de Rosa, D.M., Ots-Gutiérrez, J.R., Compañ-Rosique, A.F., and Gil-Guillén, V.F. (2017). A scoring system to predict breast cancer mortality at 5 and 10 years. Scientific reports *7*.

Parekh, S., Ziegenhain, C., Vieth, B., Enard, W., and Hellmann, I. (2016). The impact of amplification on differential expression analyses by RNA-seq. Scientific reports *6*, 25533 EP -.

Pasquinelli, A.E., Reinhart, B.J., Slack, F., Martindale, M.Q., Kuroda, M.I., Maller, B., Hayward, D.C., Ball, E.E., Degnan, B., and Müller, P., et al. (2000). Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. Nature *408*, 86-89.

Patanaphan, V., and Salazar, O.M. (1993). Colorectal cancer: metastatic patterns and prognosis. Southern medical journal *86*, 38-41.

Patwary, N. (in preparation). Identifizierung und molekulare Charakterisierung disseminierter Tumorzellen im Knochenmark von Mammakarzinom-Patientinnen. Monography (Regensburg).

Peng, Y., and Croce, C.M. (2016). The role of MicroRNAs in human cancer. Signal Transduction And Targeted Therapy *1*, 15004 EP -.

Perou, C.M. (2010). Molecular stratification of triple-negative breast cancers. The oncologist *15 Suppl 5*, 39-48.

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

Pivot, X., Romieu, G., Debled, M., Pierga, J.-Y., Kerbrat, P., Bachelot, T., Lortholary, A., Espié, M., Fumoleau, P., and Serin, D., et al. (2013). 6 months versus 12 months of adjuvant trastuzumab for patients with HER2-positive early breast cancer (PHARE): a randomised phase 3 trial. The Lancet Oncology *14*, 741-748.

Podsypanina, K., Du, Y.-C.N., Jechlinger, M., Beverly, L.J., Hambardzumyan, D., and Varmus, H. (2008). Seeding and propagation of untransformed mouse mammary cells in the lung. Science (New York, N.Y.) *321*, 1841-1844.

Polyak, K., and Metzger Filho, O. (2012). SnapShot: breast cancer. Cancer cell *22*, 562-562.e1.

Poursadegh Zonouzi, A.A., Shekari, M., Nejatizadeh, A., Shakerizadeh, S., Fardmanesh, H., Poursadegh Zonouzi, A., Rahmati-Yamchi, M., and Tozihi, M. (2017). Impaired expression of Drosha in breast cancer. Breast disease *37*, 55-62.

Prat, A., Parker, J.S., Karginova, O., Fan, C., Livasy, C., Herschkowitz, J.I., He, X., and Perou, C.M. (2010). Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. Breast cancer research : BCR *12*, R68.

Prat, A., and Perou, C.M. (2009). Mammary development meets cancer genomics. Nat Med *15*, 842 EP -.

Privat, M., Rudewicz, J., Sonnier, N., Tamisier, C., Ponelle-Chachuat, F., and Bignon, Y.-J. (2018). Antioxydation And Cell Migration Genes Are Identified as Potential Therapeutic Targets in Basal-Like and BRCA1 Mutated Breast Cancer Cell Lines. International journal of medical sciences *15*, 46-58.

Psaila, B., and Lyden, D. (2009). The metastatic niche: adapting the foreign soil. Nat Rev Cancer *9*, 285-293.

R Core Team (2014). R: A Language and Environment for Statistical Computing; available online at: http://www.R-project.org/ (Vienna, Austria: R Foundation for Statistical Computing).

Rack, B., Schindlbeck, C., Jückstock, J., Andergassen, U., Hepp, P., Zwingers, T., Friedl, T.W.P., Lorenz, R., Tesch, H., and Fasching, P.A., et al. (2014). Circulating tumor cells predict survival in early average-to-high risk breast cancer patients. Journal of the National Cancer Institute *106*.

Recamier, J.C.A. (1829). Recherches sur le traitement du cancer sur la compression methodique simple ou combinee et sur l'histoire generale de la meme maladie. 2nd edition.

Redig, A.J., and McAllister, S.S. (2013). Breast cancer as a systemic disease: a view of metastasis. Journal of internal medicine *274*, 113-126.

Remmele, W., and Stegner, H.E. (1987). Vorschlag zur einheitlichen Definition eines Immunreaktiven Score (IRS) für den immunhistochemischen Ostrogenrezeptor-Nachweis (ER-ICA) im Mammakarzinomgewebe. Der Pathologe *8*, 138-140.

Ritchie, M.E., Phipson, B., Di Wu, Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic acids research *43*, e47.

Roden, C., Mastriano, S., Wang, N., and Lu, J. (2015). microRNA Expression Profiling: Technologies, Insights, and Prospects. Advances in experimental medicine and biology *888*, 409-421.

Roylance, R., Gorman, P., Harris, W., Liebmann, R., Barnes, D., Hanby, A., and Sheer, D. (1999). Comparative genomic hybridization of breast tumors stratified by histological grade reveals new insights into the biological progression of breast cancer. Cancer Res *59*, 1433-1436.

Santamaría, D., Barrière, C., Cerqueira, A., Hunt, S., Tardy, C., Newton, K., Cáceres, J.F., Dubus, P., Malumbres, M., and Barbacid, M. (2007). Cdk1 is sufficient to drive the mammalian cell cycle. Nature *448*, 811-815.

Schardt, J.A., Meyer, M., Hartmann, C.H., Schubert, F., Schmidt-Kittler, O., Fuhrmann, C., Polzer, B., Petronio, M., Eils, R., and Klein, C.A. (2005). Genomic analysis of single cytokeratin-positive cells from bone marrow reveals early mutational events in breast cancer. Cancer cell *8*, 227-239.

Schindlbeck, C., Andergassen, U., Hofmann, S., Jückstock, J., Jeschke, U., Sommer, H., Friese, K., Janni, W., and Rack, B. (2013). Comparison of circulating tumor cells (CTC) in peripheral blood and disseminated tumor cells in the bone marrow (DTC-BM) of breast cancer patients. Journal of cancer research and clinical oncology *139*, 1055-1062.

Schindlbeck, C., Pfab, G., Jueckstock, J., Andergassen, U., Sommer, H., Janni, W., Friese, K., and Rack, B. (2011). Prognostic relevance of disseminated tumor cells in the bone marrow of patients with primary breast cancer--results of a standardized follow-up. Anticancer research *31*, 2749-2755.

Schlimok, G., Funke, I., Holzmann, B., Göttlinger, G., Schmidt, G., Häuser, H., Swierkot, S., Warnecke, H.H., Schneider, B., and Koprowski, H., et al. (1987). Micrometastatic cancer cells in bone marrow: in vitro detection with anti-cytokeratin and in vivo labeling with anti-17-1A monoclonal antibodies. Proc Natl Acad Sci U S A *84*, 8672-8676.

Schmidt, M. (2014). Chemotherapy in early breast cancer: when, how and which one? Breast care (Basel, Switzerland) *9*, 154-160.

Schmidt-Kittler, O. (2003). Von einzelnen disseminierten Tumorzellen zur Metastase: Genomische Analyse der minimalen Resterkrankung des Mammakarzinoms. Monography (München).

Schmidt-Kittler, O., Ragg, T., Daskalakis, A., Granzow, M., Ahr, A., Blankenstein, T.J.F., Kaufmann, M., Diebold, J., Arnholdt, H., and Muller, P., et al. (2003). From latent disseminated cells to overt metastasis: genetic analysis of systemic breast cancer progression. Proc Natl Acad Sci U S A *100*, 7737-7742.

Selli, C., and Sims, A.H. (2019). Neoadjuvant Therapy for Breast Cancer as a Model for Translational Research. Breast cancer : basic and clinical research *13*, 1178223419829072.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research *13*, 2498-2504.

Shatzkes, K., Teferedegne, B., and Murata, H. (2014). A simple, inexpensive method for preparing cell lysates suitable for downstream reverse transcription quantitative PCR. Scientific reports *4*, 4659 EP -.

Shibue, T., and Weinberg, R.A. (2009). Integrin beta1-focal adhesion kinase signaling directs the proliferation of metastatic cancer cells disseminated in the lungs. Proc Natl Acad Sci U S A *106*, 10290-10295.

Sinn, H.-P., and Kreipe, H. (2013). A Brief Overview of the WHO Classification of Breast Tumors, 4th Edition, Focusing on Issues and Updates from the 3rd Edition. Breast care (Basel, Switzerland) *8*, 149-154.

Song, Y., Liu, K.J., and Wang, T.-H. (2014). Elimination of ligation dependent artifacts in T4 RNA ligase to achieve high efficiency and low bias microRNA capture. PLoS One *9*, e94619.

Sorefan, K., Pais, H., Hall, A.E., Kozomara, A., Griffiths-Jones, S., Moulton, V., and Dalmay, T. (2012). Reducing ligation bias of small RNAs in libraries for next generation sequencing. Silence *3*, 4.

Sørlie, T., Perou, C.M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M.B., van de Rijn, M., and Jeffrey, S.S., et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A *98*, 10869-10874.

Stark, G.R., Stein, W.H., and Moore, S. (1960). Reactions of the Cyanate Present in Aqueous Urea with Amino Acids and Proteins. Journal of Biological Chemistry *235*, 3177-3181.

Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., Stein, T.I., Nudel, R., Lieder, I., and Mazor, Y., et al. (2016). The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. Current protocols in bioinformatics *54*, 1.30.1-1.30.33.

Stingl, J., and Caldas, C. (2007). Molecular heterogeneity of breast carcinomas and the cancer stem cell hypothesis. Nat Rev Cancer *7*, 791-799.

Stoecklein, N.H., Erbersdobler, A., Schmidt-Kittler, O., Diebold, J., Schardt, J.A., Izbicki, J.R., and Klein, C.A. (2002). SCOMP Is Superior to Degenerated Oligonucleotide Primed-Polymerase Chain Reaction for Global Amplification of Minute Amounts of DNA from Microdissected Archival Tissue Samples. The American Journal of Pathology *161*, 43-51.

Stoecklein, N.H., and Klein, C.A. (2010). Genetic disparity between primary tumours, disseminated tumour cells, and manifest metastasis. Int J Cancer *126*, 589-598.

Subauste, M.C., Ventura-Holman, T., Du, L., Subauste, J.S., Chan, S.-L., Yu, V.C., and Maher, J.F. (2009). RACK1 downregulates levels of the pro-apoptotic protein Fem1b in apoptosis-resistant colon cancer cells. Cancer biology & therapy *8*, 2297-2305.

Swain, S.M., Baselga, J., Kim, S.-B., Ro, J., Semiglazov, V., Campone, M., Ciruelos, E., Ferrero, J.-M., Schneeweiss, A., and Heeson, S., et al. (2015). Pertuzumab, trastuzumab, and docetaxel in HER2-positive metastatic breast cancer. The New England journal of medicine *372*, 724-734.

Szostakowska, M., Trębińska-Stryjewska, A., Grzybowska, E.A., and Fabisiewicz, A. (2019). Resistance to endocrine therapy in breast cancer: molecular mechanisms and future goals. Breast Cancer Res Treat *173*, 489-497.

Talmadge, J.E., and Fidler, I.J. (2010). AACR centennial series: the biology of cancer metastasis: historical perspective. Cancer Res *70*, 5649-5669.

Tan, W., Liu, B., Qu, S., Liang, G., Luo, W., and Gong, C. (2018). MicroRNAs and cancer: Key paradigms in molecular therapy. Oncology letters *15*, 2735-2742.

Tang, F., Hajkova, P., Barton, S.C., O'Carroll, D., Lee, C., Lao, K., and Surani, M.A. (2006). 220-plex microRNA expression profile of a single cell. Nature protocols *1*, 1154-1159.

Tavassoli, F.A. and Devilee, P. (2003). Pathology and genetics of tumours of the breast and female genital organs (Lyon: IARC Press).

Thiery, J.P., Acloque, H., Huang, R.Y.J., and Nieto, M.A. (2009). Epithelial-mesenchymal transitions in development and disease. Cell *139*, 871-890.

Thu, K.L., Soria-Bretones, I., Mak, T.W., and Cescon, D.W. (2018). Targeting the cell cycle in breast cancer: towards the next phase. Cell cycle (Georgetown, Tex.) *17*, 1871-1885.

Tickle, T., Tirosh, I., Georgescu, C., Brown, M., and Haas, B. (2019). infercnv of the Trinity CTAT Project (Klarman Cell Observatory, Broad Institute of MIT and Harvard, Cambridge, MA, USA, available online at: https://github.com/broadinstitute/inferCNV: Bioconductor).

Tirkkonen, M., Tanner, M., Karhu, R., Kallioniemi, A., Isola, J., and Kallioniemi, O.P. (1998). Molecular cytogenetics of primary breast cancer by CGH. Genes, chromosomes & cancer *21*, 177-184.

Uhlen, M., Björling, E., Agaton, C., Szigyarto, C.A.-K., Amini, B., Andersen, E., Andersson, A.-C., Angelidou, P., Asplund, A., and Asplund, C., et al. (2005). A human protein atlas for normal and cancer tissues based on antibody proteomics. Molecular & cellular proteomics : MCP *4*, 1920-1932.

Uhlen, M., Zhang, C., Lee, S., Sjöstedt, E., Fagerberg, L., Bidkhori, G., Benfeitas, R., Arif, M., Liu, Z., and Edfors, F., et al. (2017). A pathology atlas of the human cancer transcriptome. Science (New York, N.Y.) *357*.

Uhr, J.W., and Pantel, K. (2011). Controversies in clinical cancer dormancy. Proc Natl Acad Sci U S A *108*, 12396-12400.

Vaidyanathan, S., Thangavelu, P.U., and Duijf, P.H.G. (2016). Overexpression of Ran GTPase Components Regulating Nuclear Export, but not Mitotic Spindle Assembly, Marks Chromosome Instability and Poor Prognosis in Breast Cancer. Targeted oncology *11*, 677-686.

Valastyan, S., and Weinberg, R.A. (2011). Tumor metastasis: molecular insights and evolving paradigms. Cell *147*, 275-292.

Viollet, S., Fuchs, R.T., Munafo, D.B., Zhuang, F., and Robb, G.B. (2011). T4 RNA ligase 2 truncated active site mutants: improved tools for RNA analysis. BMC biotechnology *11*, 72.

Visvader, J.E., and Lindeman, G.J. (2008). Cancer stem cells in solid tumours: accumulating evidence and unresolved questions. Nat Rev Cancer *8*, 755-768.

Walker, C., Mojares, E., and Del Río Hernández, A. (2018). Role of Extracellular Matrix in Development and Cancer Progression. International journal of molecular sciences *19*.

Wang, D.-Y., Fulthorpe, R., Liss, S.N., and Edwards, E.A. (2004). Identification of estrogen-responsive genes by complementary deoxyribonucleic acid microarray and characterization of a novel early estrogen-induced gene: EEIG1. Molecular endocrinology (Baltimore, Md.) *18*, 402-411.

Wang, N., Zheng, J., Chen, Z., Liu, Y., Dura, B., Kwak, M., Xavier-Ferrucio, J., Lu, Y.-C., Zhang, M., and Roden, C., et al. (2019). Single-cell microRNA-mRNA co-sequencing reveals non-genetic heterogeneity and mechanisms of microRNA regulation. Nature Communications *10*, 95.

Weaver, V.M., Petersen, O.W., Wang, F., Larabell, C.A., Briand, P., Damsky, C., and Bissell, M.J. (1997). Reversion of the malignant phenotype of human breast cells in three-dimensional culture and in vivo by integrin blocking antibodies. The Journal of cell biology *137*, 231-245.

Weigelt, B., Mackay, A., A'hern, R., Natrajan, R., Tan, D.S.P., Dowsett, M., Ashworth, A., and Reis-Filho, J.S. (2010). Breast cancer molecular profiling with single sample predictors: a retrospective analysis. The Lancet Oncology *11*, 339-349.

Weigelt, B., Peterse, J.L., and van 't Veer, L.J. (2005). Breast cancer metastasis: markers and models. Nat Rev Cancer *5*, 591-602.

Weinberg, R.A. (2007). The biology of cancer (New York, NY: GS Garland Science).

Weinberg, R.A. (2008). The many faces of tumor dormancy. APMIS : acta pathologica, microbiologica, et immunologica Scandinavica *116*, 548-551.

Weiss, L. (1992). Comments on hematogenous metastatic patterns in humans as revealed by autopsy. Clinical & experimental metastasis *10*, 191-199.

Weiss, L., Grundmann, E., Torhorst, J., Hartveit, F., Moberg, I., Eder, M., Fenoglio-Preiser, C.M., Napier, J., Horne, C.H., and Lopez, M.J. (1986). Haematogenous metastatic patterns in colonic carcinoma: an analysis of 1541 necropsies. The Journal of pathology *150*, 195-203.

Weiss, L., and Harlos, J.P. (1986). The validity of negative necropsy reports for metastases in solid organs. The Journal of pathology *148*, 203-206.

Weiss, L., Harlos, J.P., Torhorst, J., Gunthard, B., Hartveit, F., Svendsen, E., Huang, W.L., Grundmann, E., Eder, M., and Zwicknagl, M. (1988). Metastatic patterns of renal carcinoma: an analysis of 687 necropsies. Journal of cancer research and clinical oncology *114*, 605-612.

Werner-Klein, M., Scheitler, S., Hoffmann, M., Hodak, I., Dietz, K., Lehnert, P., Naimer, V., Polzer, B., Treitschke, S., and Werno, C., et al. (2018). Genetic alterations driving metastatic colony formation are acquired outside of the primary tumour in melanoma. Nature Communications *9*, 595.

Wienholds, E., and Plasterk, R.H.A. (2005). MicroRNA function in animal development. FEBS letters *579*, 5911-5922.

Wilbert, J., and Lueke, T. (2019). Scan: Single-case data analyses for single and multiple baseline designs. Retrieved from https://CRAN.R-project.org/package=scan.

Woelfle, U., Breit, E., Zafrakas, K., Otte, M., Schubert, F., Müller, V., Izbicki, J.R., Löning, T., and Pantel, K. (2005). Bi-specific immunomagnetic enrichment of micrometastatic tumour cell clusters from bone marrow of cancer patients. Journal of immunological methods *300*, 136-145.

Wolff, A.C., Hammond, M.E.H., Hicks, D.G., Dowsett, M., McShane, L.M., Allison, K.H., Allred, D.C., Bartlett, J.M.S., Bilous, M., and Fitzgibbons, P., et al. (2013). Recommendations for human epidermal growth factor receptor 2 testing in breast cancer: American Society of Clinical Oncology/College of American Pathologists clinical practice guideline update. Journal of clinical oncology: official journal of the American Society of Clinical Oncology; ISSN: 0732-183X *31*, 3997-4013.

Wu, J., Xiao, J., Zhang, Z., Wang, X., Hu, S., and Yu, J. (2014). Ribogenomics: the science and knowledge of RNA. Genomics, proteomics & bioinformatics *12*, 57-63.

Xiao, W., Zheng, S., Yang, A., Zhang, X., Zou, Y., Tang, H., and Xie, X. (2018). Breast cancer subtypes and the risk of distant metastasis at initial diagnosis: a population-based study. Cancer management and research *10*, 5329-5338.

Yadav, A.S., Pandey, P.R., Butti, R., Radharani, N.N.V., Roy, S., Bhalara, S.R., Gorain, M., Kundu, G.C., and Kumar, D. (2018). The Biology and Therapeutic Implications of Tumor Dormancy and Reactivation. Frontiers in oncology *8*, 72.

Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S., and Madden, T.L. (2012). Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. BMC bioinformatics *13*, 134.

Yerukala Sathipati, S., and Ho, S.-Y. (2018). Identifying a miRNA signature for predicting the stage of breast cancer. Scientific reports *8*, 16138.

Yoo, M.-H., and Hatfield, D.L. (2008). The cancer stem cell theory: Is it correct? Molecules and cells *26*, 514-516.

Zhang, Y., Yang, L., and Chen, L.-L. (2014). Life without A tail: new formats of long noncoding RNAs. The international journal of biochemistry & cell biology *54*, 338-349.

Zhang, Z., Lee, J.E., Riemondy, K., Anderson, E.M., and Yi, R. (2013). High-efficiency RNA cloning enables accurate quantification of miRNA expression by deep sequencing. Genome biology *14*, R109.

Zhou, H., Wang, X.-J., Jiang, X., Qian, Z., Chen, T., Hu, Y., Chen, Z.-H., Gao, Y., Wang, R., and Ye, W.-W., et al. (2019). Plasma cell-free DNA chromosomal instability analysis by low-pass whole-genome sequencing to monitor breast cancer relapse. Breast Cancer Res Treat *178*, 63-73.

Zhuang, F., Fuchs, R.T., Sun, Z., Zheng, Y., and Robb, G.B. (2012). Structural bias in T4 RNA ligase-mediated 3'-adapter ligation. Nucleic acids research *40*, e54.

11. Acknowledgement

Now that I am at the end of my thesis, I want to thank all the people who helped me get to this point by aiding me in the lab, by giving me advice or simply by encouraging me to keep going.

First, I want to thank my supervisor Prof. Dr. Christoph Klein for entrusting this fascinating research project to me and for his constant feedback and advice. I am also grateful for the large degree of freedom and independence I was allowed to enjoy throughout the whole project especially regarding the method development part, which allowed me to be creative. I also want to thank him for giving me the opportunity to present my research on international conferences.

Second, I want to thank my mentors Prof. Dr. Gunter Meister and Prof. Dr. Wilko Weichert for the interesting discussions and their valuable feedback regarding my work. Additionally, I want to thank the members of the TransLUMINAL-B consortium, especially Prof. Dr. Carsten Denkert and Prof. Dr. Gero Brockhoff, who also provided useful advice during our progress meetings.

Third, I want to thank all the PostDocs that helped me with their advice in the course of my project. Namely, these are Dr. Miodrag Gužvić, Dr. Courtney König, Dr. Bernhard Polzer, Dr. Elisabeth Schneider, Dr. Giancarlo Feliciello, Dr. Huiqin Koerkel-Qu, Dr. Zbigniew Czyż, Dr. Hedayatollah Hosseini, and Dr. Stefan Kirsch. Particularly, I want to highlight Miodrag and Courtney for all the time they sacrificed for me by always being open to my questions and for proof-reading this thesis. At the same time, I also want to apologize to them, because they usually did not have a choice, since they were sitting next to me in the office. Furthermore, I want to specially thank Giancarlo for his invaluable help with the LowPass-Sequencing and for letting me use his computer for the analysis so many times. Finally, I want to express my gratitude towards Dr. Verena Lieb for laying the groundwork for the extended WTA and for staying available in case I had any questions.

Fourth, I also need to thank all the technicians of the AG Klein that helped me in the lab, both with their advice and with hands-on assistance in my experiments. Regarding the latter, I want to especially thank Isabell Blochberger and Sandra Grunewald, without whom I this dissertation would not have been possible.

Moreover, I want to thank Dr. Gundula Haunschild and Nina Patwary for collecting most of the patient samples that I used for this thesis as well as for providing some fundamental data that I could build on. Special thanks also go to Dr. Huiqin Koerkel-Qu for her extremely committed and quick analysis of my RNA-Seq data.

Furthermore, I want to express my gratitude to (in no particular order) Ana, Nina, Rezan, Mio, Elisabeth, Courtney, Heda, Tom, Theresa, the three Sandras, Julia, Christian, Isa, and Mani as well as all the other colleagues of the AG Klein for creating such a wonderful atmosphere to work in. It was a pleasure to be a part of this group and to gain many new friends, who will stay with me even after my PhD. Thanks especially to Tom, Christian, and Theresa for introducing me to the wonderful world of climbing and thank you Tom for organizing those unforgettable hiking trips.

I also want to thank all the patients and healthy donors, without whom none of my DCC-related work would have been possible. I also want to express my gratitude towards our clinical partners, who provided the samples. Particularly, I want to thank Prof. Dr. Karl Sotlar and Prof. Dr. Helga Bernhard for the re-staining of many primary tumors to provide me vital information on the patients' KI67 status.

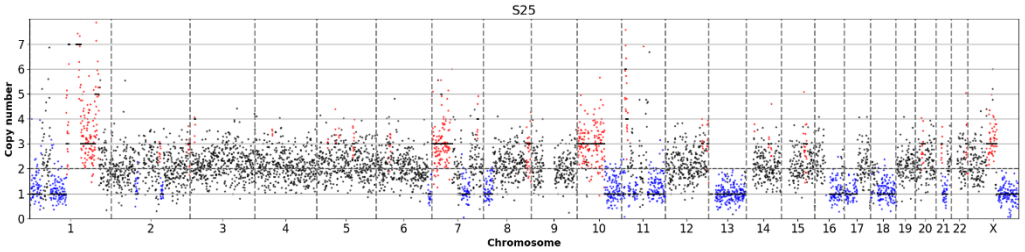
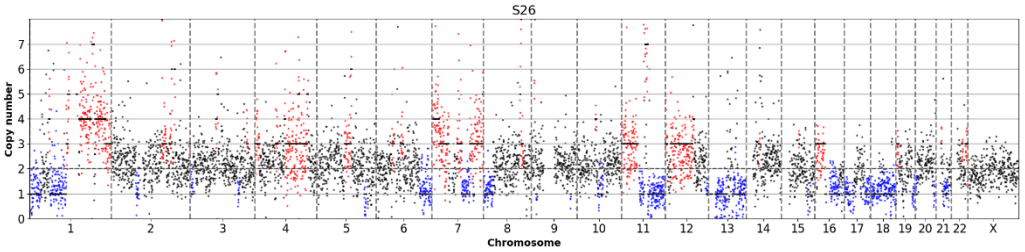
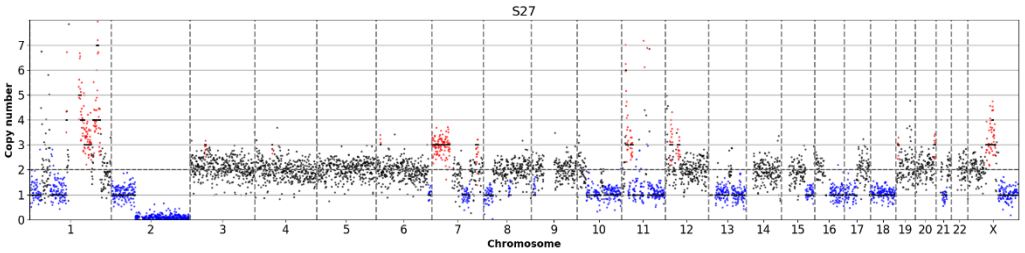
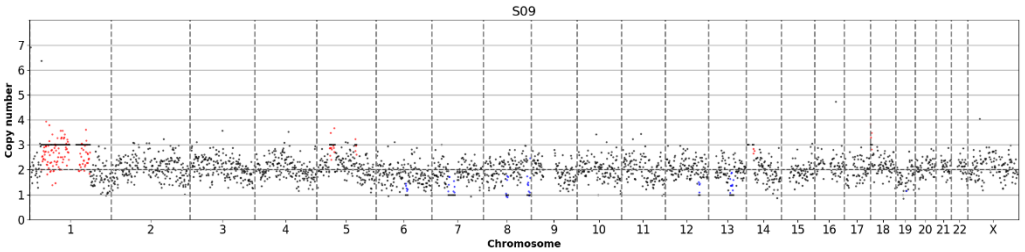
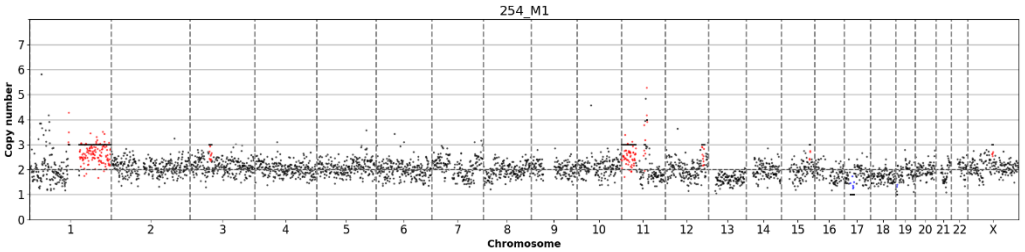
Last but not least, I want to thank my wife Julia as well as my parents Karin and Werner for always supporting me throughout my studies and for believing in me, even when times were difficult.

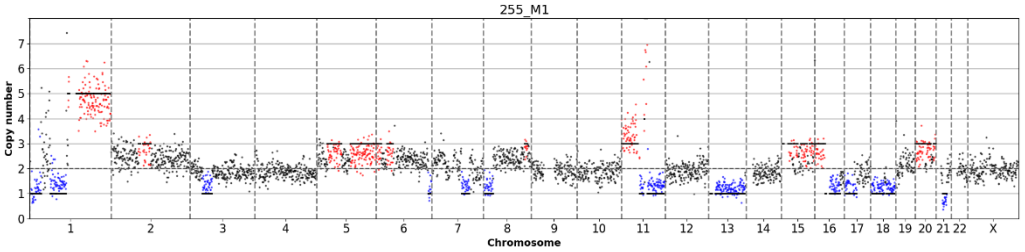
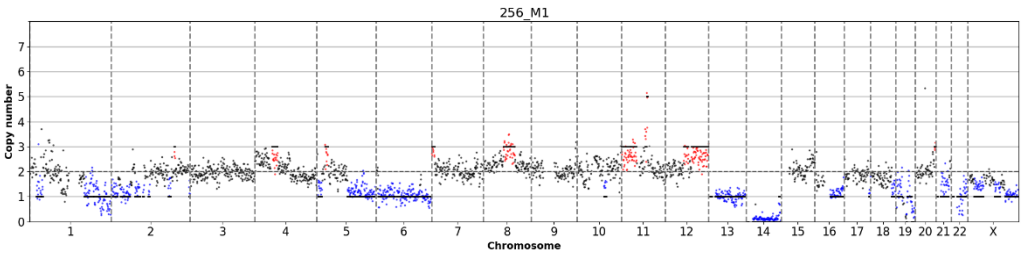
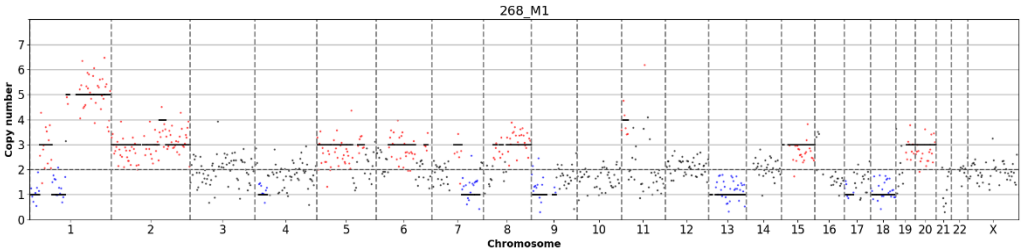
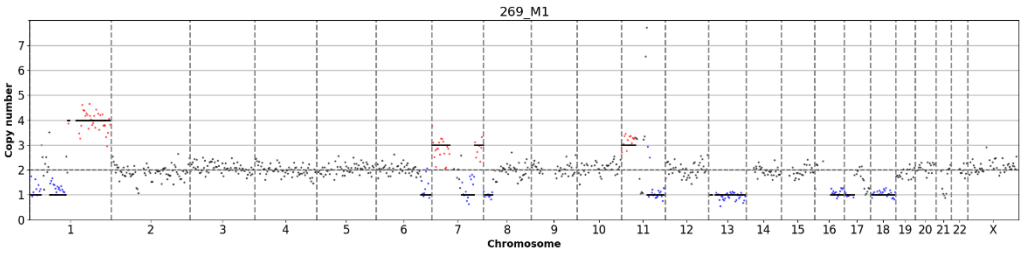
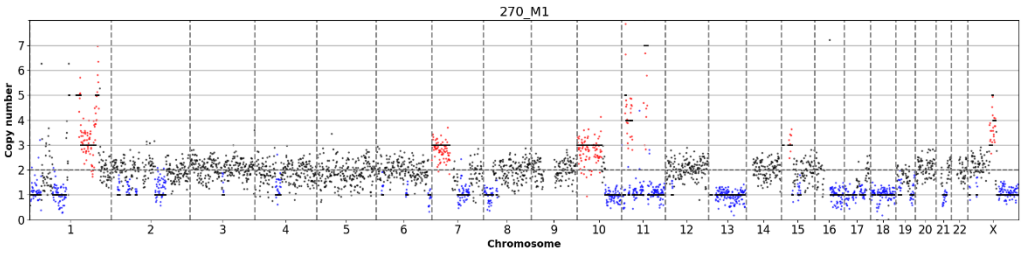
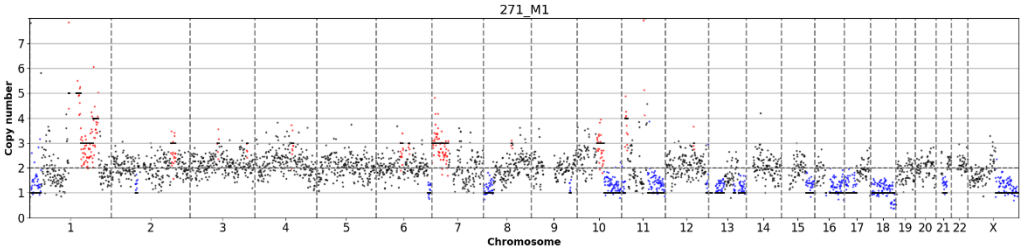
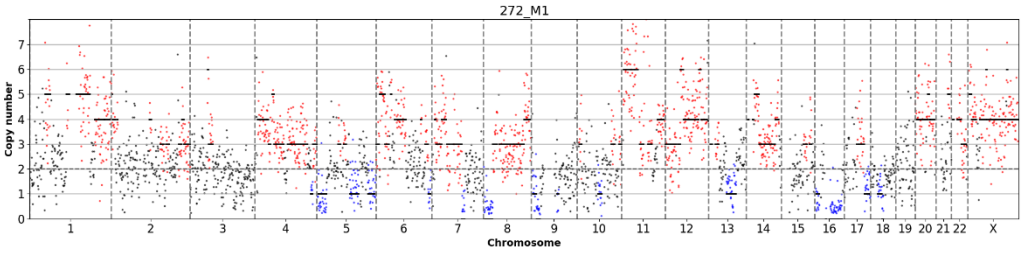
12. Appendix

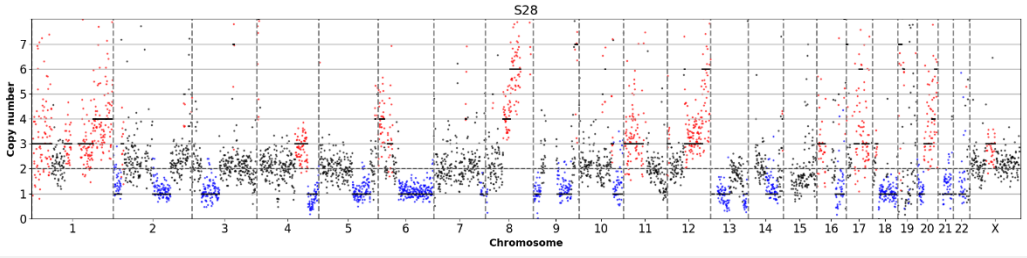
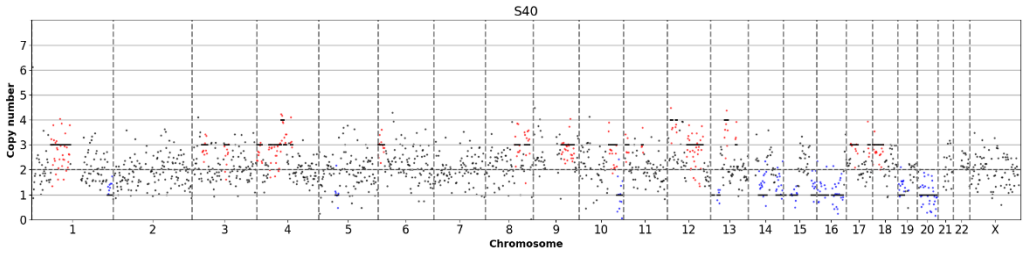
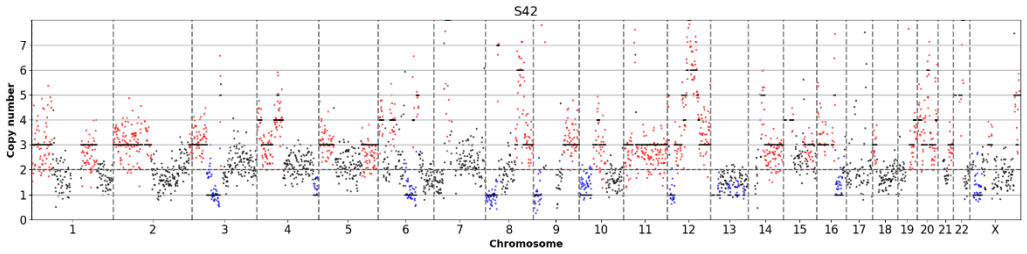
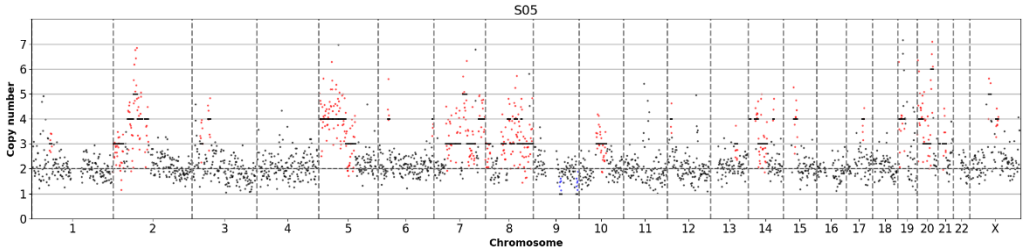
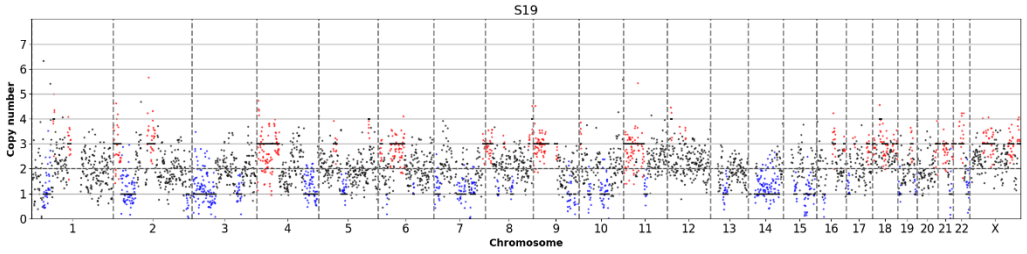
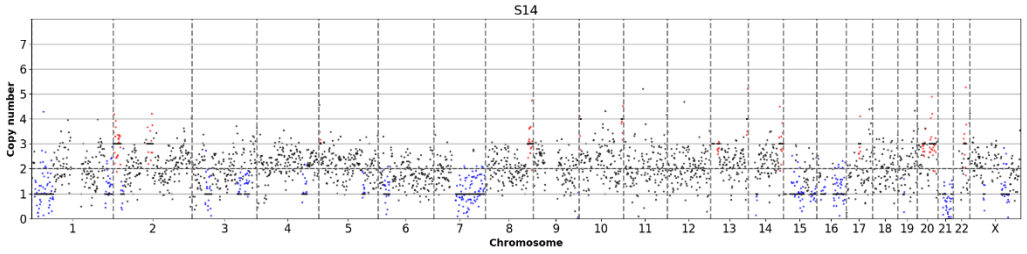
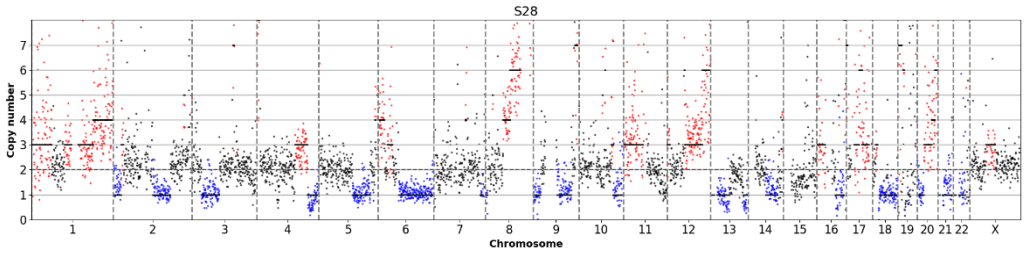
12.1 Copy number alteration analysis

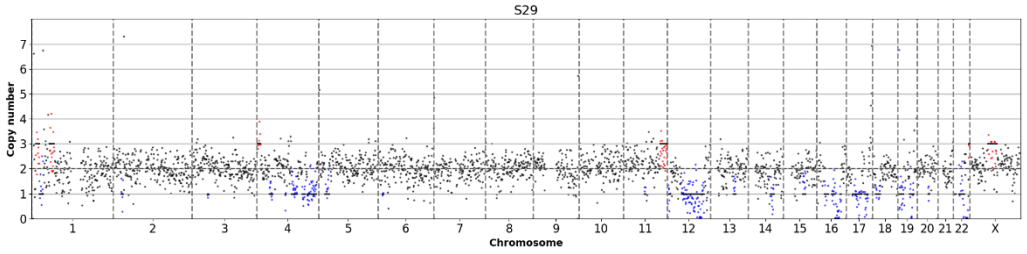
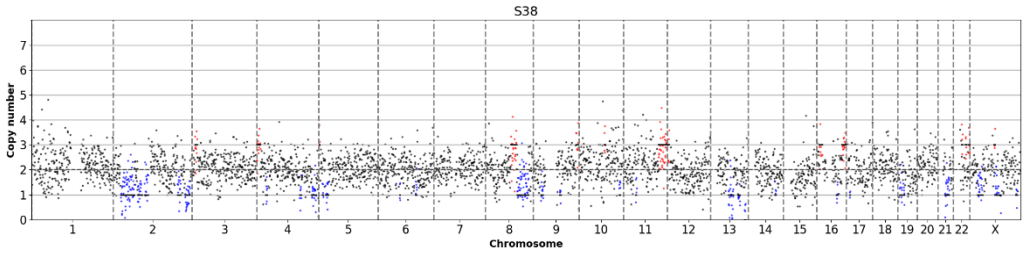
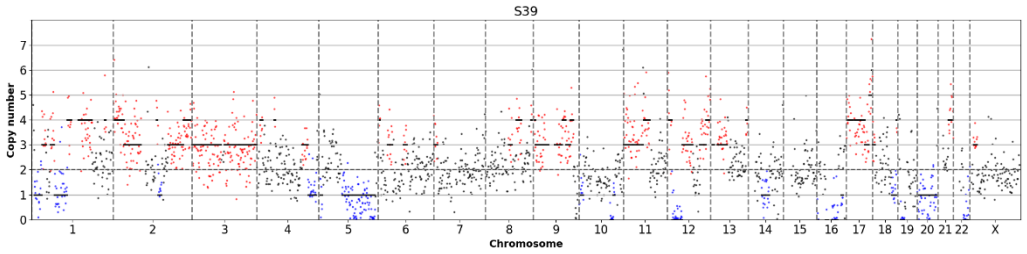
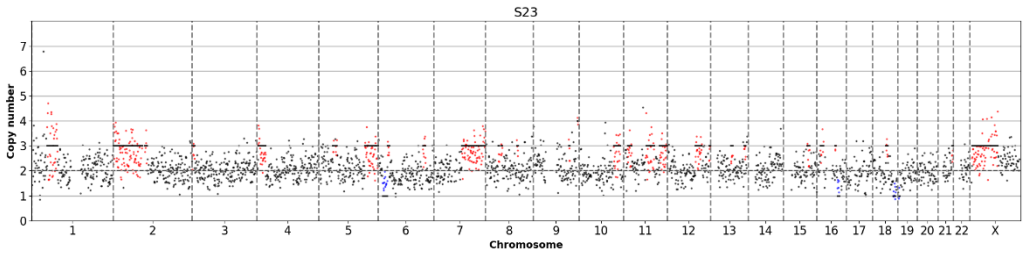
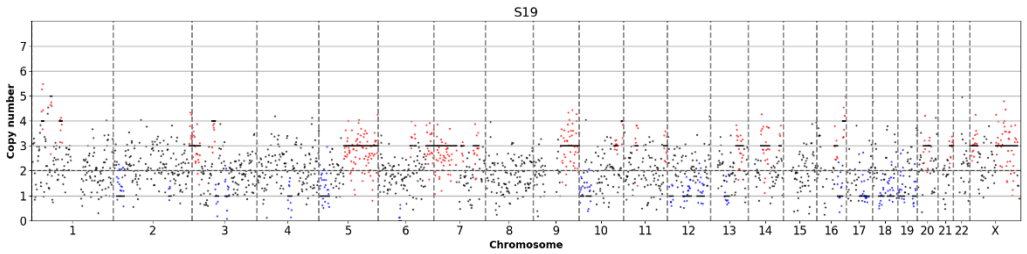
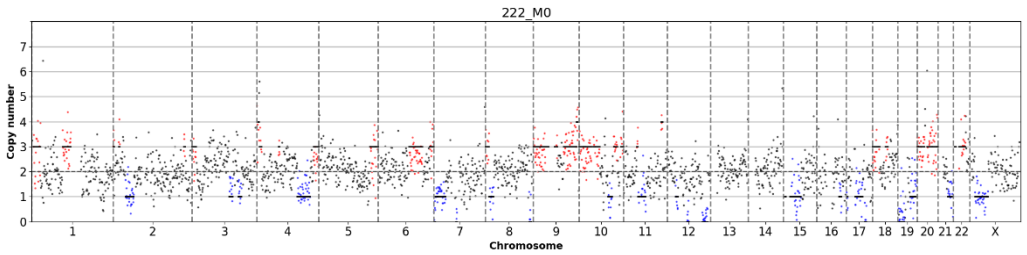
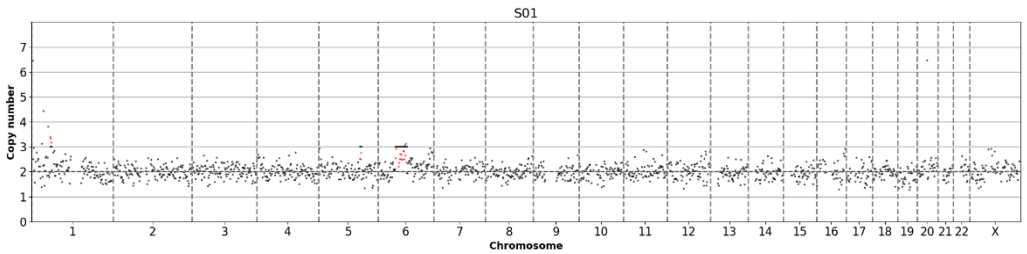
12.1.1 CNA profiles with sufficient quality for analysis

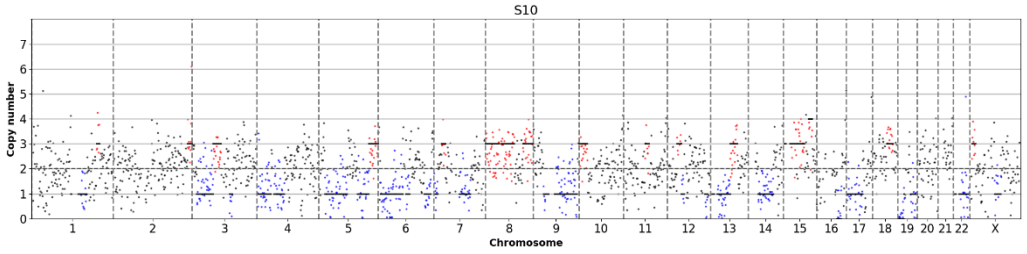
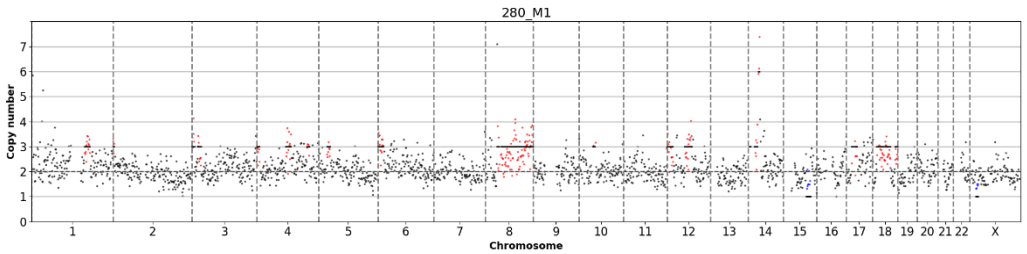
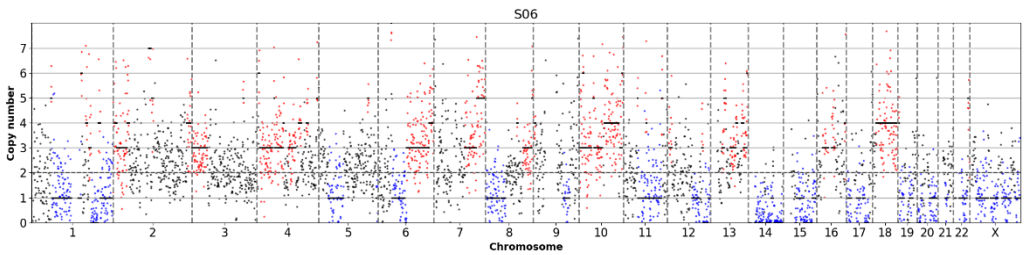
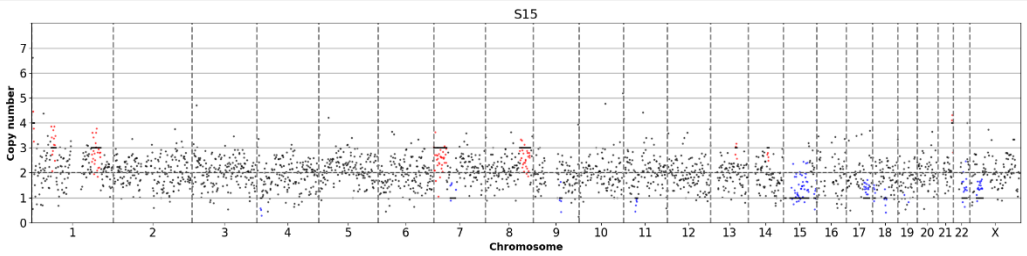
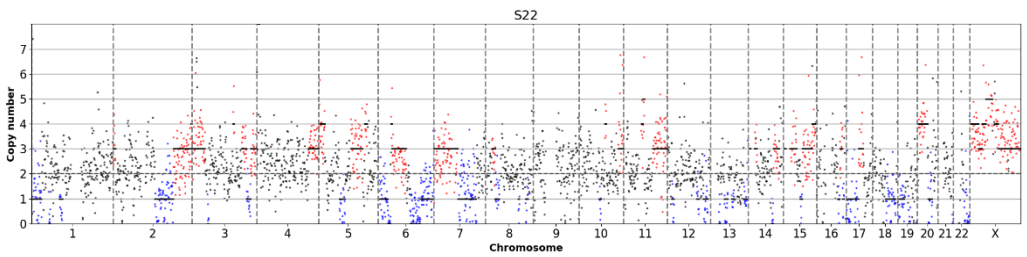
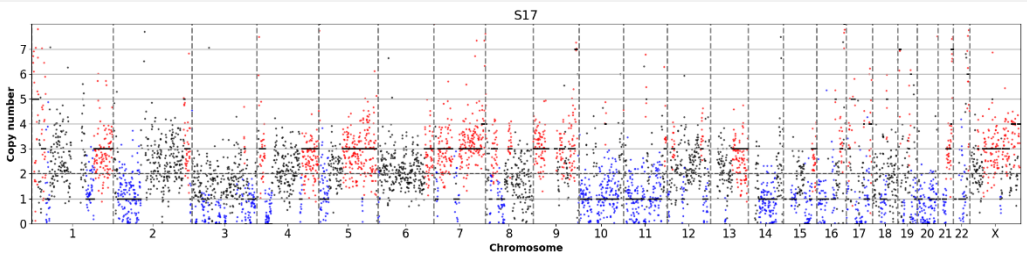
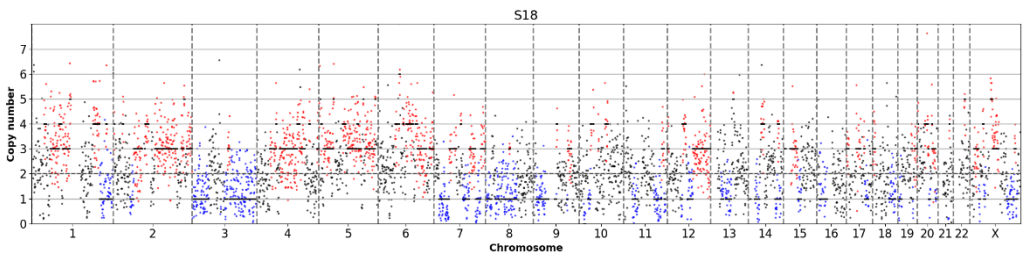
Table 12-1 Collection of aberrant LowPass-Seq profiles with sufficient quality for further analyses. The table provides all profiles, which could clearly be classified as aberrant, together with relevant sample information in the following order: sample ID, sample source (M0/M1/HD), BC subtype, read count, and derivate log ratio spread (DLRS). The profiles were sorted by BC subtype.

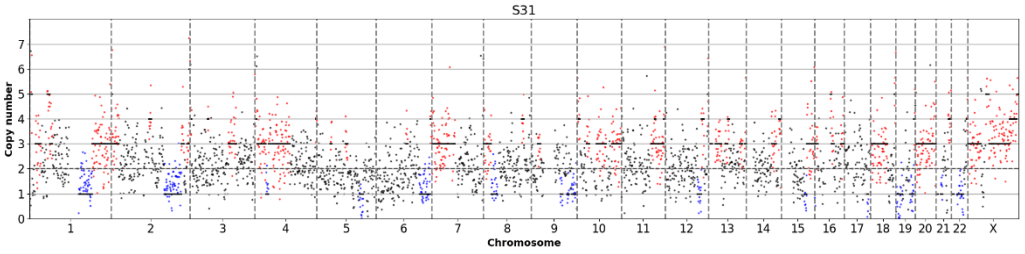
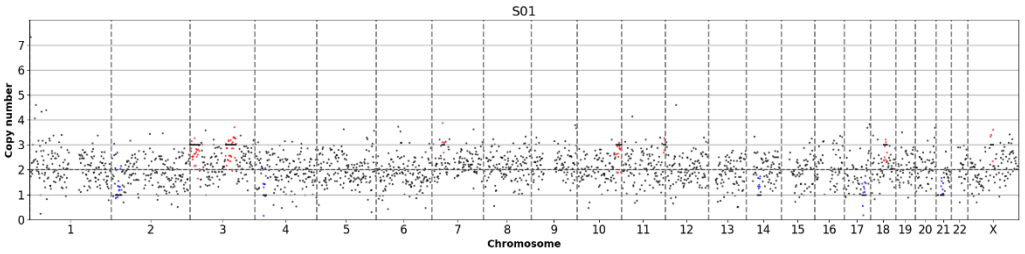
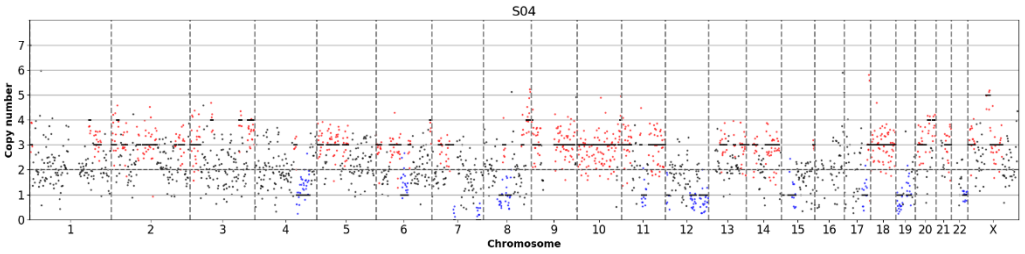
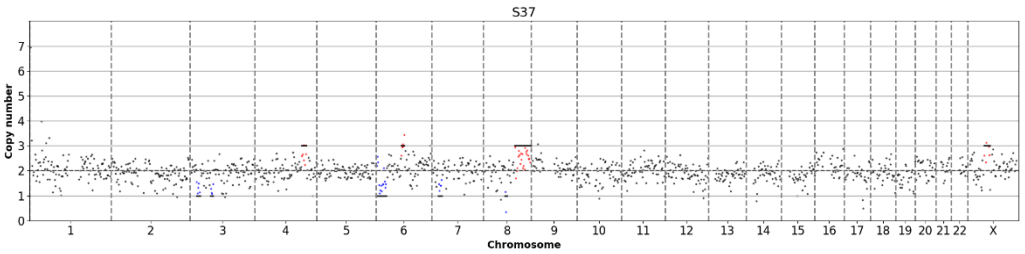
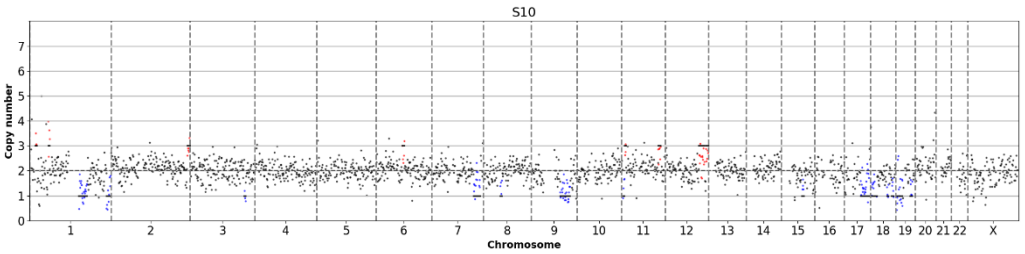
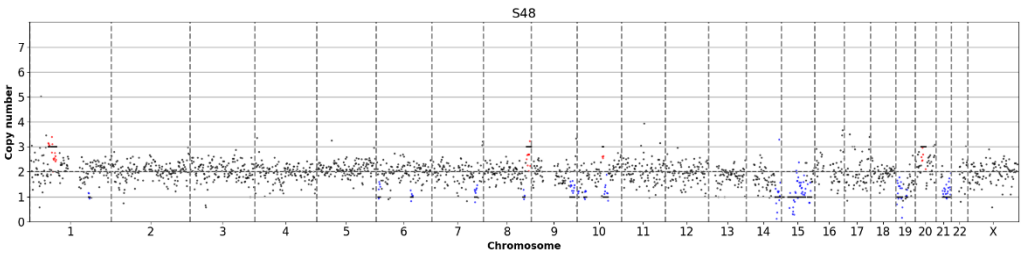
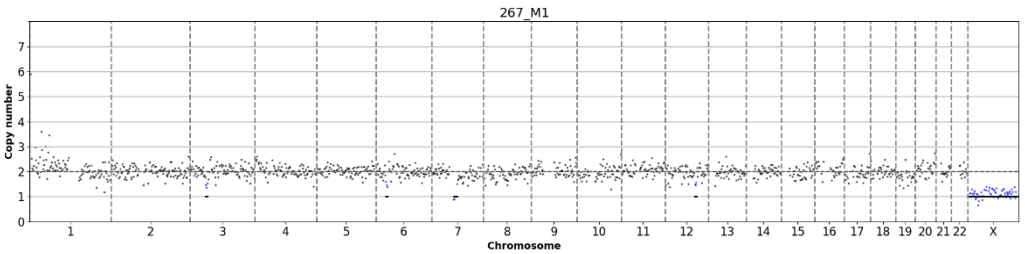
Sample	Profile
M1-01-09-689-3 M1 Lum undefined 1817461 0.388	
M1-01-09-689-4 M1 Lum undefined 2126722 0.455	
M1-01-09-689-5 M1 Lum undefined 1629083 0.408	
MC05-09-802-2 M0 Lum undefined 820858 0.321	
M1 01/09-689-11 M1 Lum undefined 1105037 0.250	

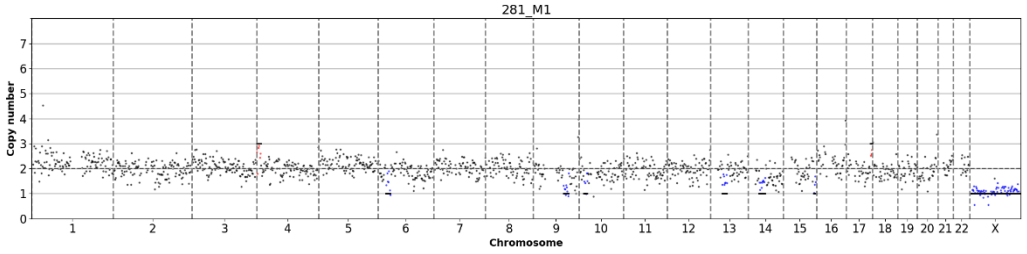
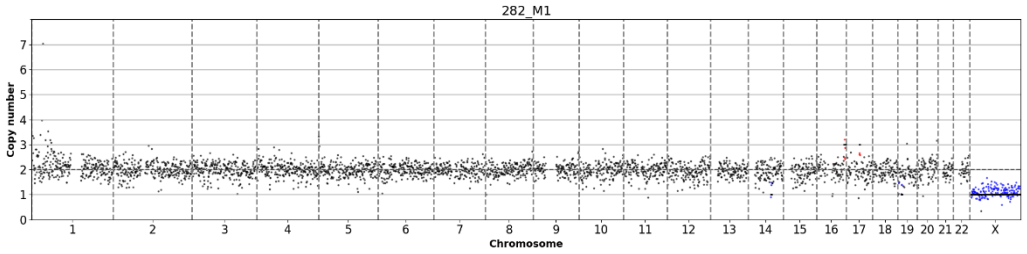
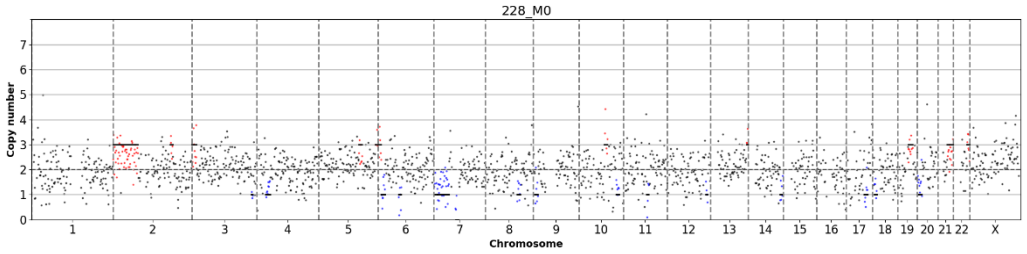
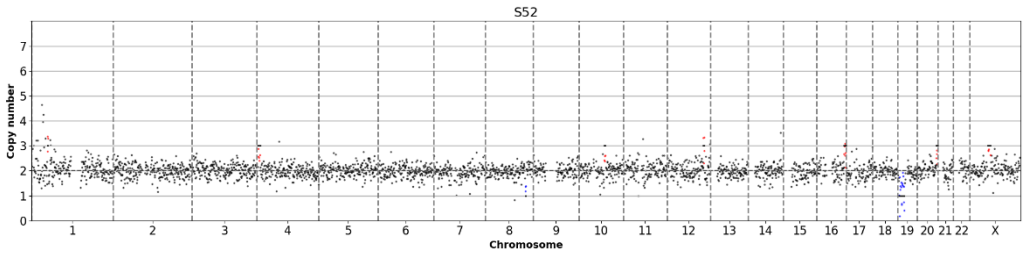
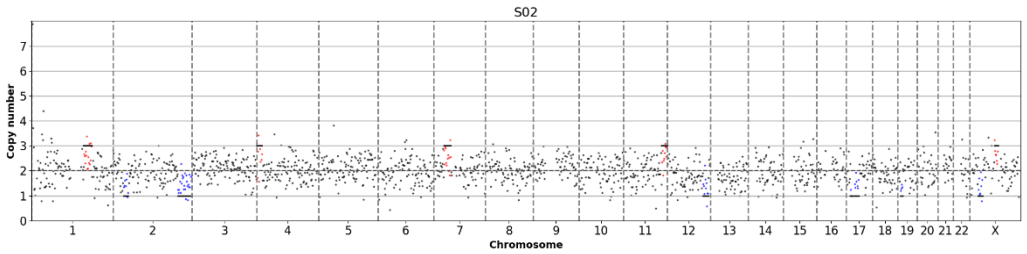
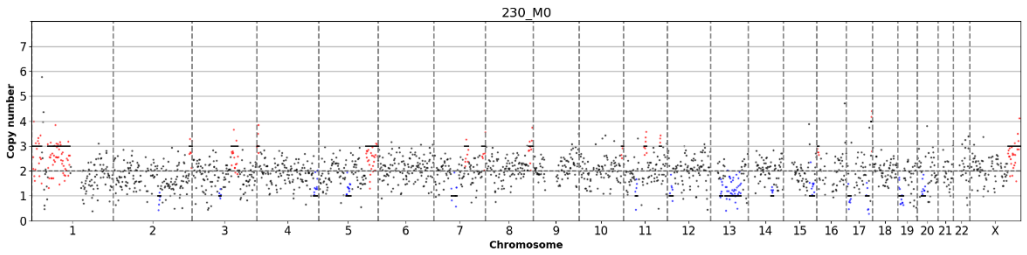
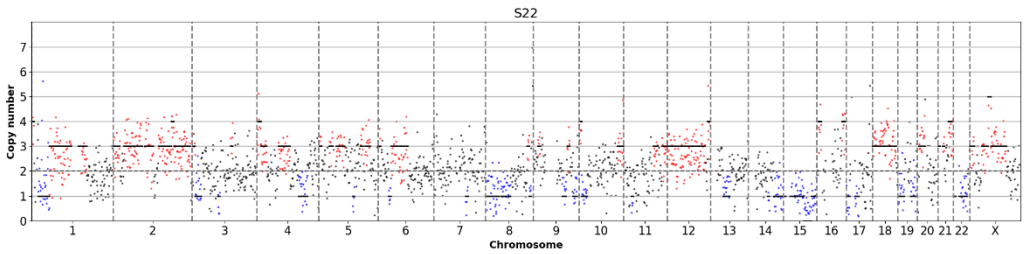
Sample	Profile
M1 01/09-689-7 M1 Lum undefined 1088362 0.284	
M1 01/09-689-9 M1 Lum undefined 838896 0.313	
M1 01/09-689-1 M1 Lum undefined 290906 0.454	
M1 01/09-689-10 M1 Lum undefined 280749 0.308	
M1 01/09-689-2 M1 Lum undefined 1093732 0.382	
M1 01/09-689-6 M1 Lum undefined 1050925 0.303	
M1 01/09-689-8 M1 Lum undefined 953616 0.509	

Sample	Profile
M1-03-09-773-2 M1 TNBC 1478932 0.392	
Met02-10-947-1 M1 TNBC 518147 0.537	
Met02-10-947-4 M1 TNBC 967750 0.418	
MC12-2094KM-12 M0 TNBC 742248 0.365	
MC-04-11-1125-5 M0 TNBC 1211662 0.454	
MC13-2308KM-1 M0 TNBC 835015 0.526	
M1-03-09-773-2 M1 TNBC 1478932 0.392	

Sample	Profile
MC03-09-741-3 M0 LumA 900066 0.493	
MC-11-10-1058-2 M0 LumA 1190225 0.406	
MC-11-10-1058-17 M0 LumA 680096 0.612	
MC-10-10-1030-6 M0 LumA 866199 0.326	
MC13-2336KM-1 M0 LumA 676121 0.538	
MC 11-2009 KM-2 M0 LumA 750579 0.502	
MC03-09-741-1 M0 LumA 493714 0.286	

Sample	Profile
MC12-2203KM-3 M0 LumA 628848 0.649	
M1-12-2232KM-27 M1 Luminal A 658699 0.285	
M1-03-11-1122-2 M1 LumB 1306475 0.799	
MC-02-11-1102-4 M0 LumB 741177 0.492	
MC-03-11-1119-3 M0 LumB 1018690 0.770	
MC-03-11-1119-4 M0 LumB 1638092 0.787	
MC-03-11-1119-5 M0 LumB 1414400 0.568	

Sample	Profile
MC-01-11-1090-4 M0 LumB 998382 0.471	
MC11-2012-KM-1 M0 LumB 715429 0.484	
MC11-2051KM-11 M0 LumB 655233 0.524	
MC13-2361KM-2 M0 LumB 399627 0.351	
MC13-2384KM-2 M0 LumB 646919 0.323	
MC13-2427KM-11 M0 LumB 612670 0.391	
MC14-2675KM-2 M1 LumB 400313 0.257	

Sample	Profile
MC14-2675KM-3 M1 LumB 460039 0.267	
MC14-2675KM-4 M1 LumB 1053849 0.243	
MC12-2081KM-4 M0 LumB 581248 0.513	
MC13-2477KM-13 M0 LumB 935147 0.268	
MC12-2081KM-2 M0 LumB 517406 0.443	
MC12-2096KM-1 M0 LumB 638545 0.442	
N13-2411KM-3 HD NCC 656905 0.517	

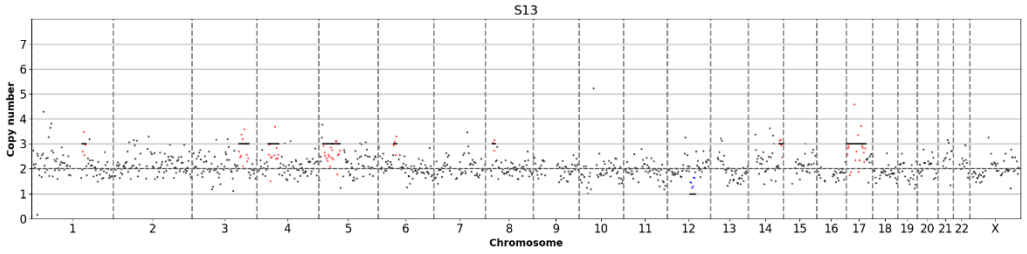
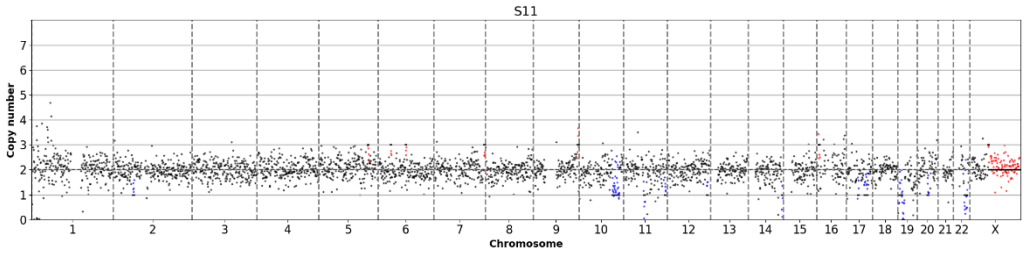
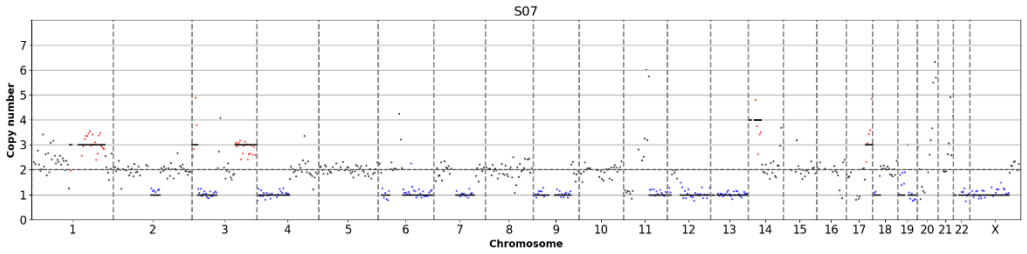
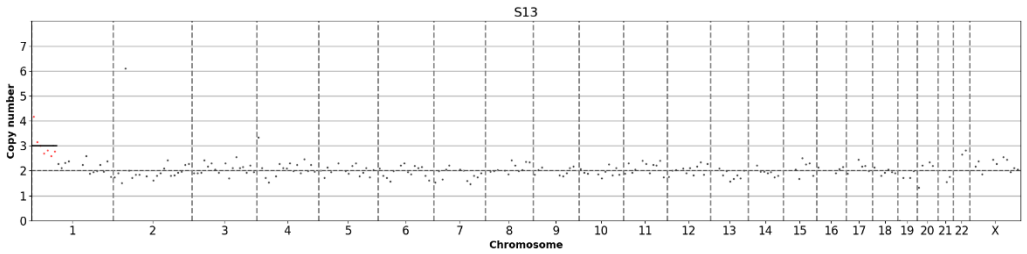
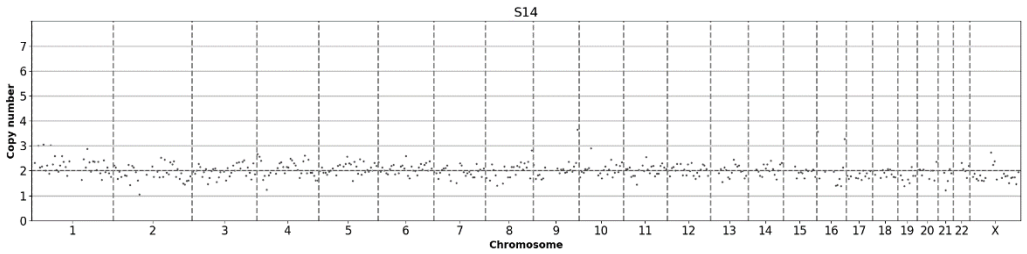
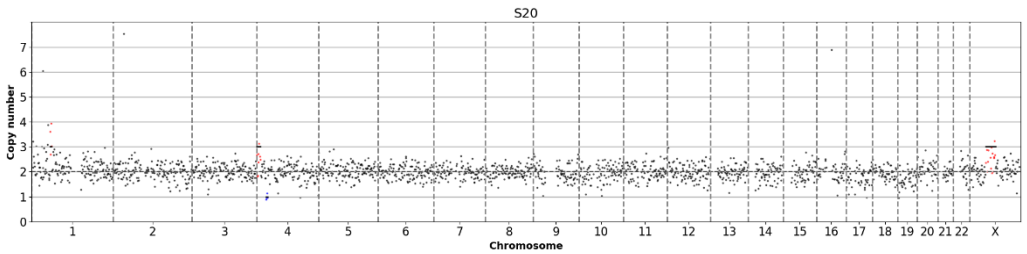
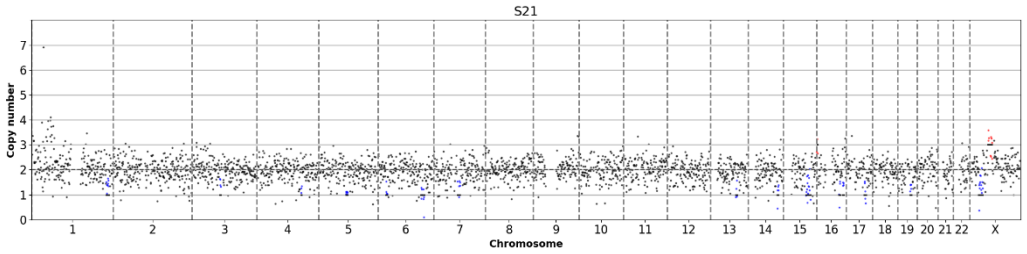
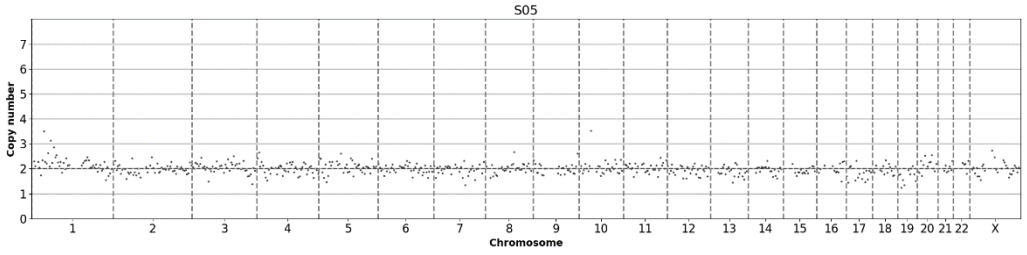
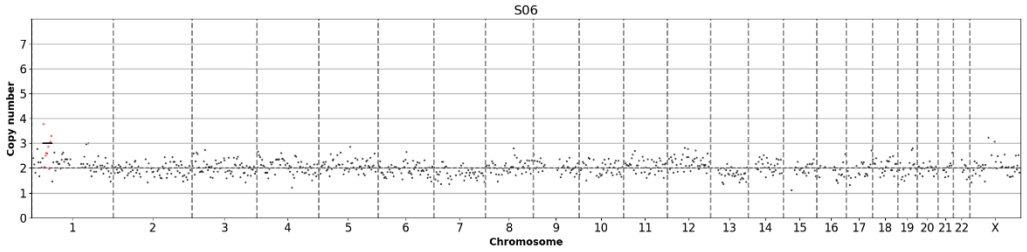
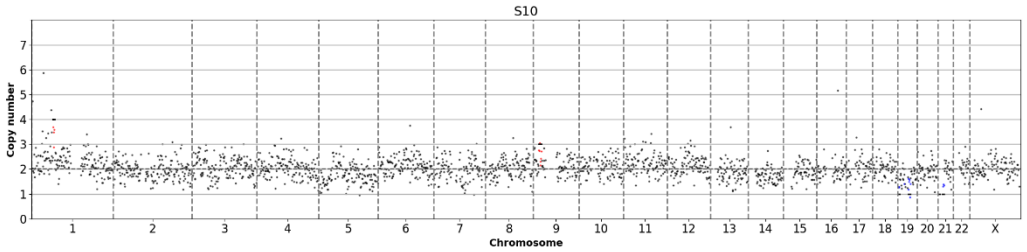
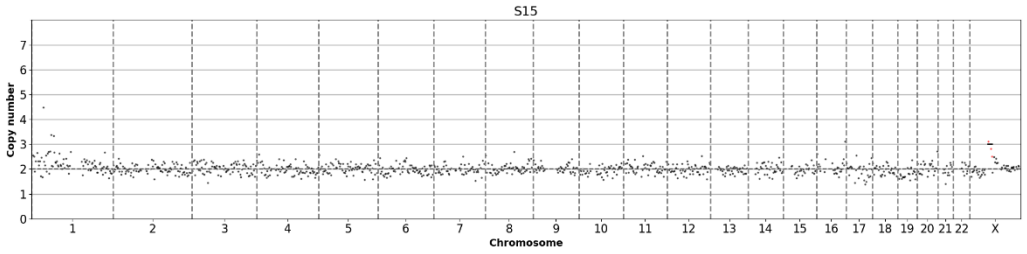
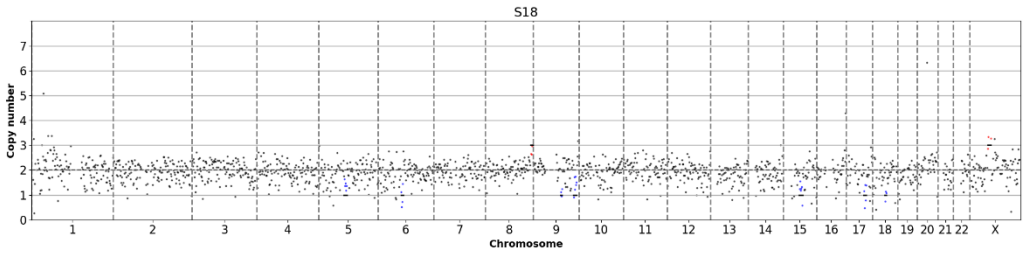
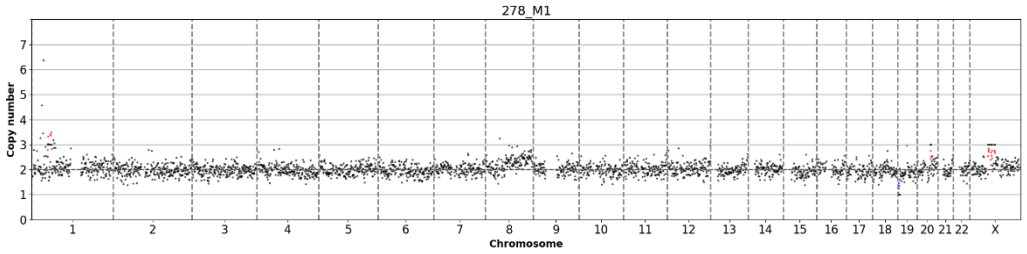
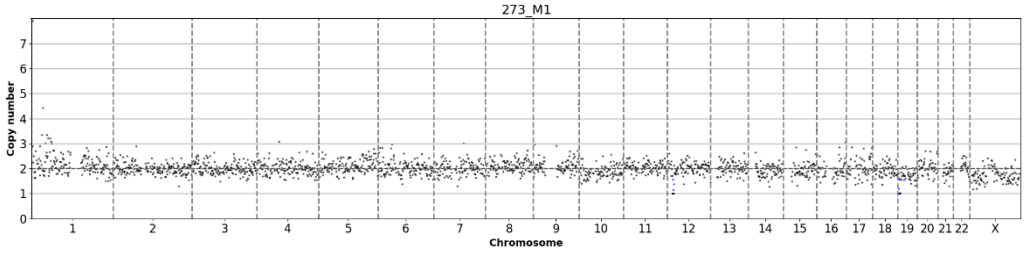
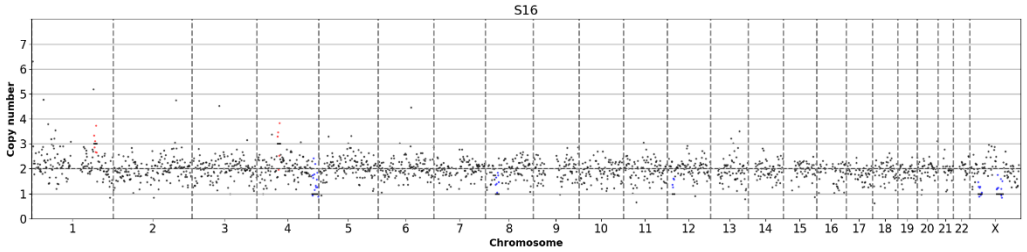
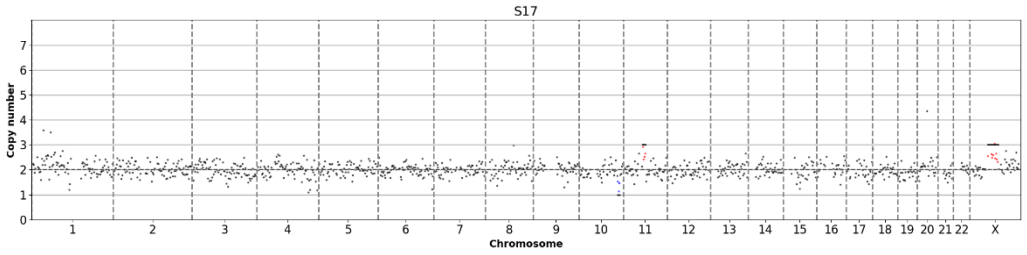
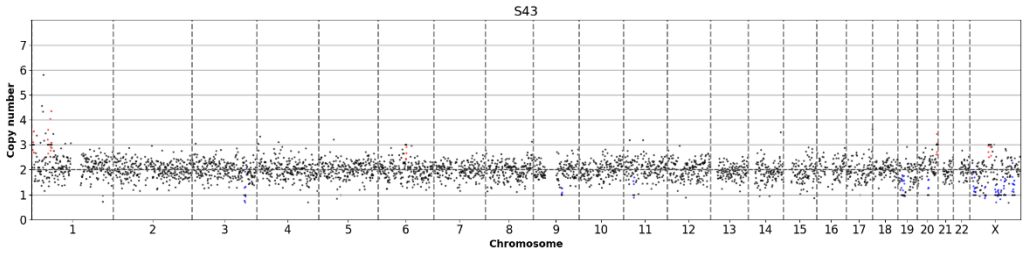
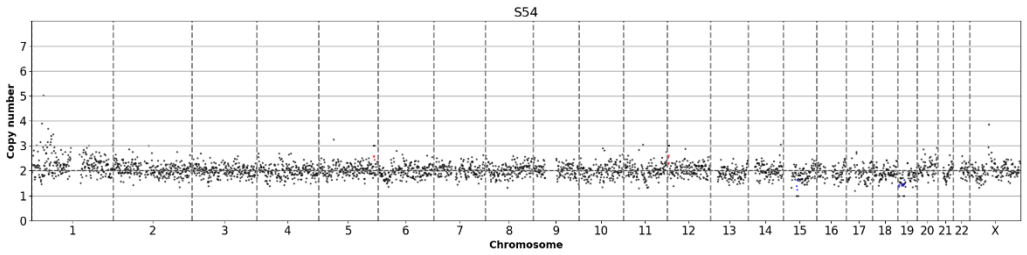
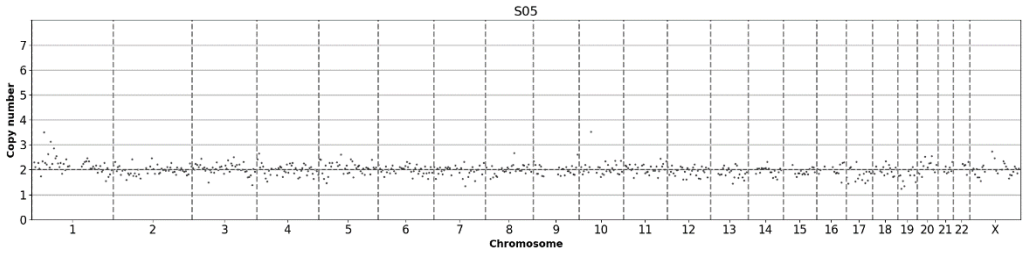
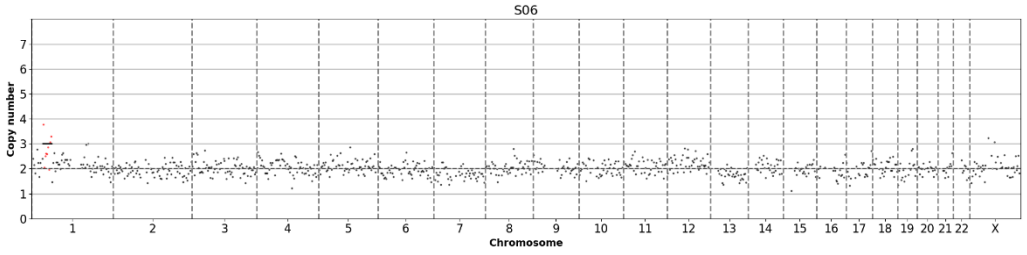
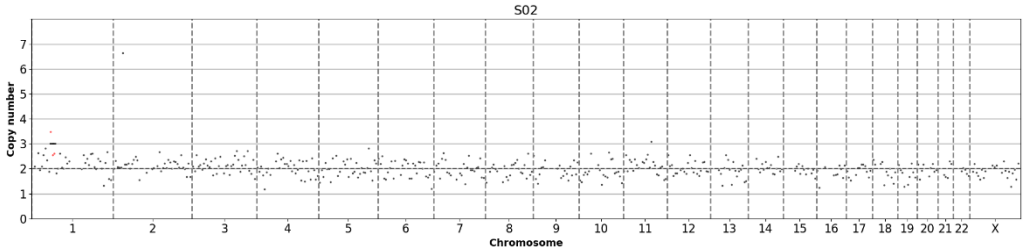
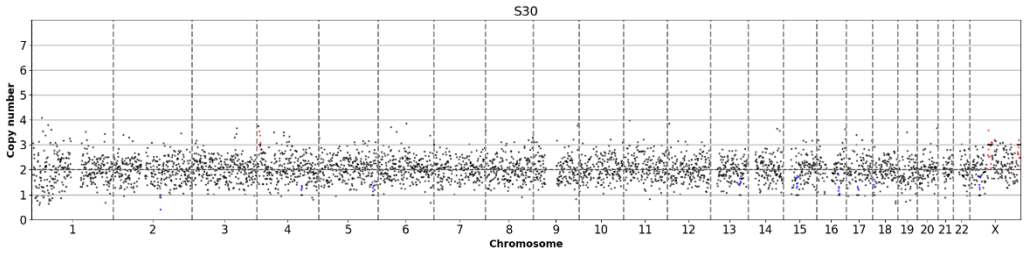
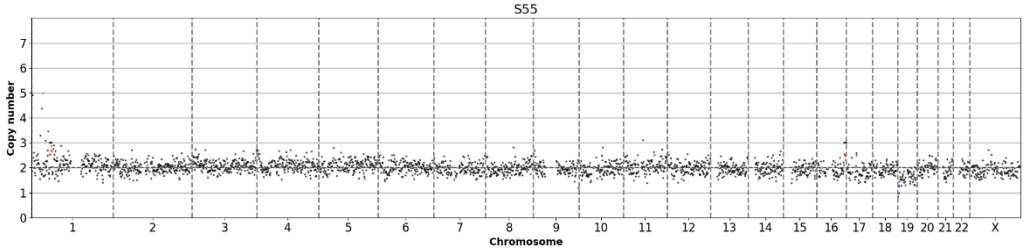
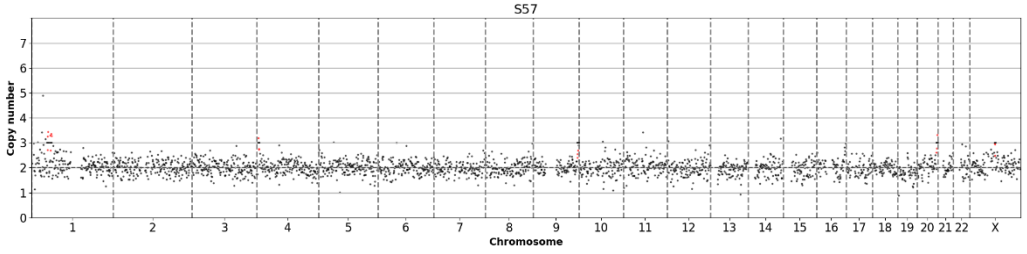
Sample	Profile
N13-2411KM-2 HD NCC 343723 0.328	
N12-2213KM-3 HD NCC 1028822 0.352	
MC04-08-356-2 M0 No data 215422 0.376	

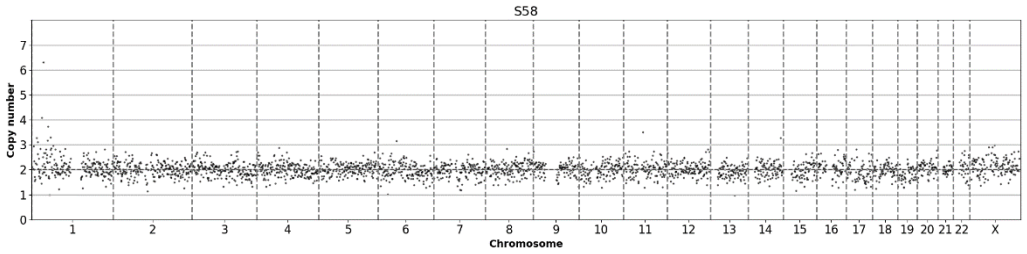
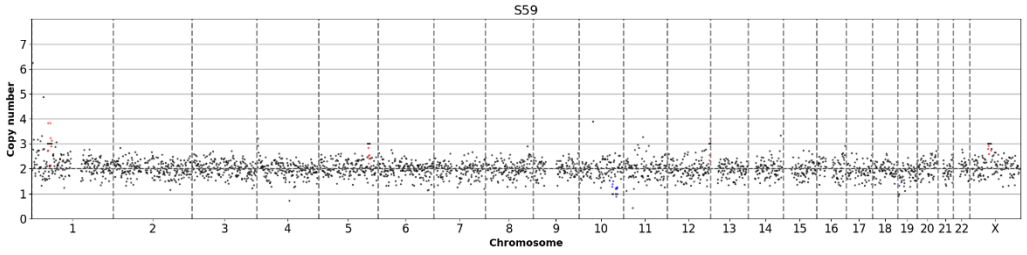
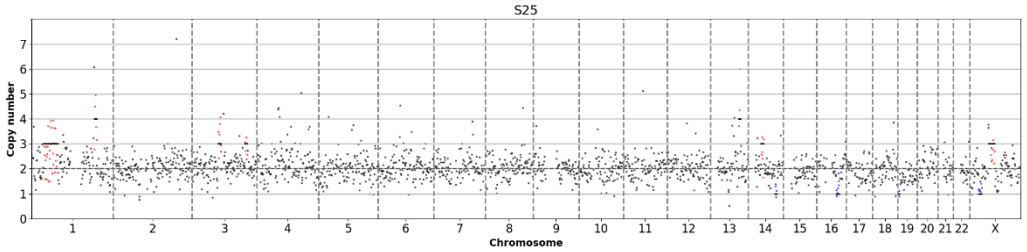
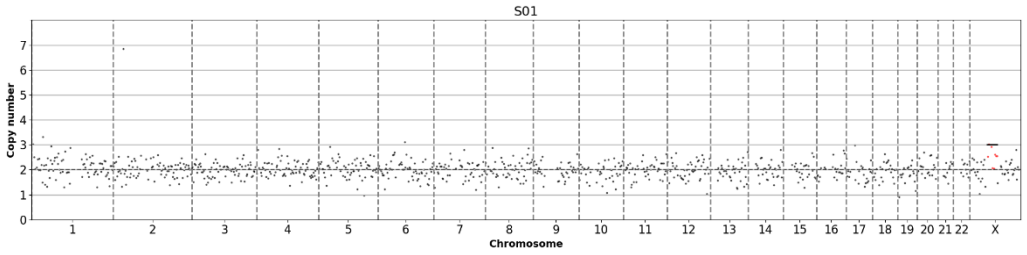
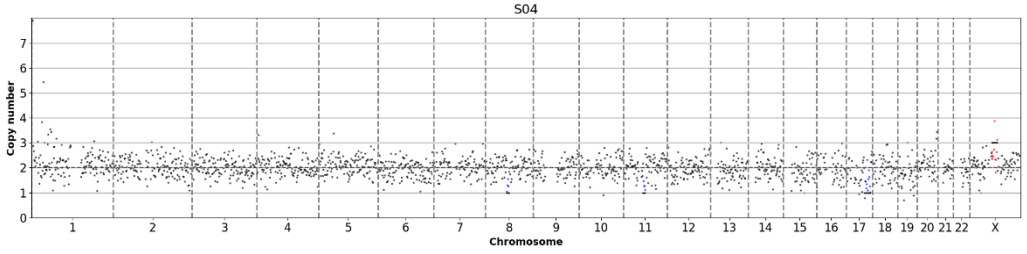
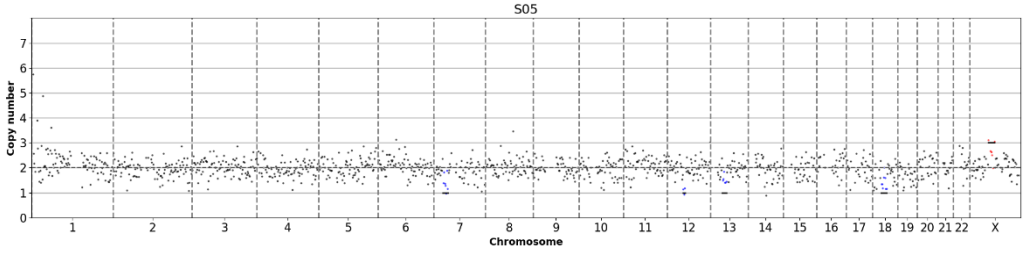
Table 12-2 Collection of balanced LowPass-Seq profiles with sufficient quality for further analyses. The table provides all profiles, which could clearly be classified as balanced, together with relevant sample information in the following order: sample ID, metastatic state, BC subtype, read count, derivate log ratio spread (DLRS). The profiles were sorted by BC subtype.

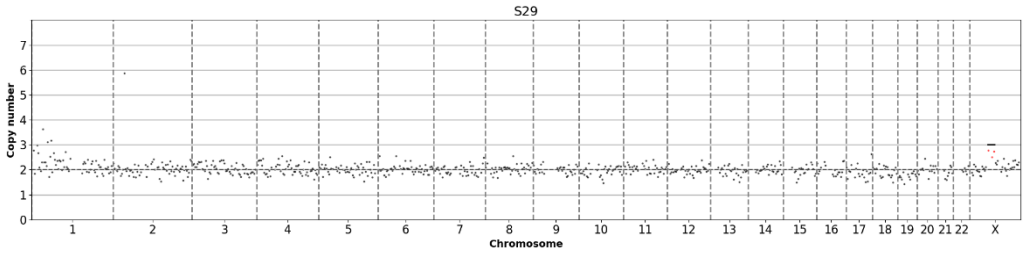
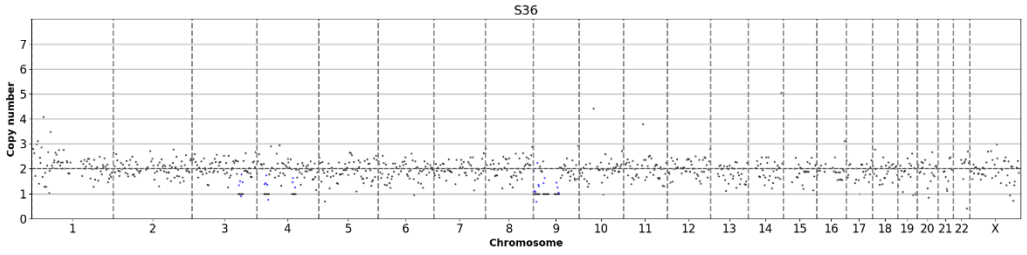
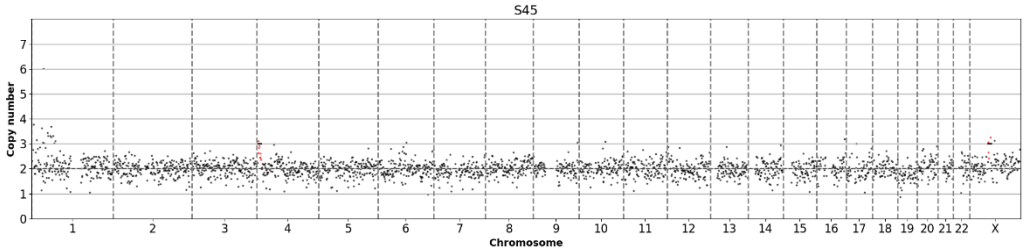
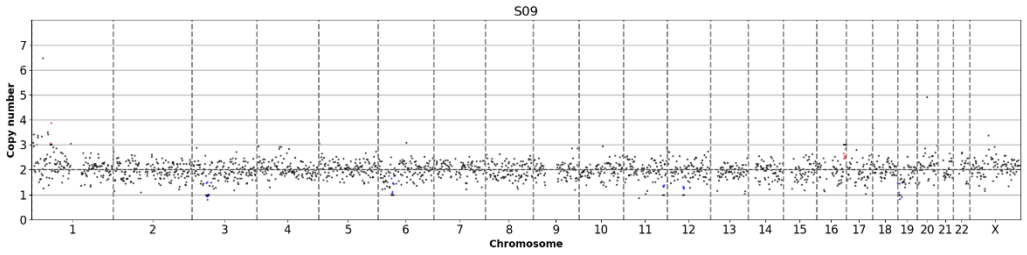
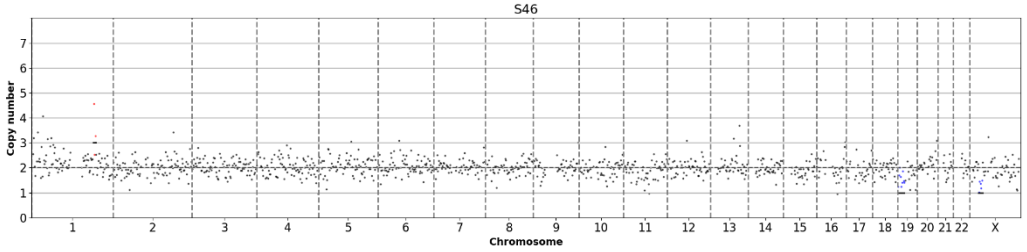
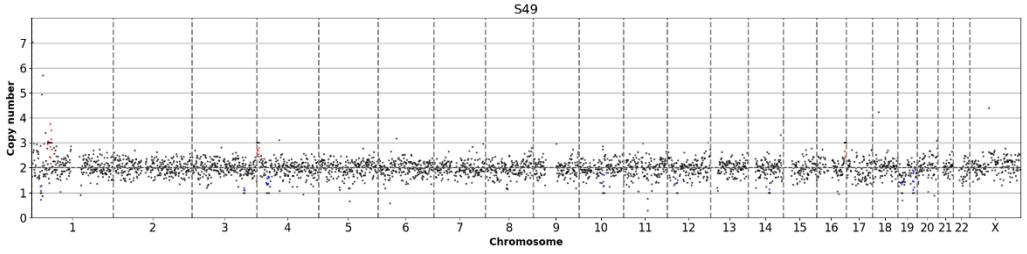
Sample info	Profile
MC08-08-524-1 M0 Lum undefined NA NA	
MC08-08-524-1 M0 Lum undefined NA NA	
MC-03-11-1117-3 M0 Lum undefined 630190 0.292	

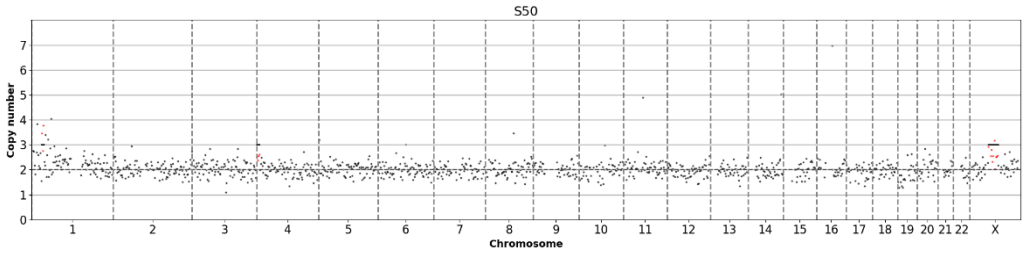
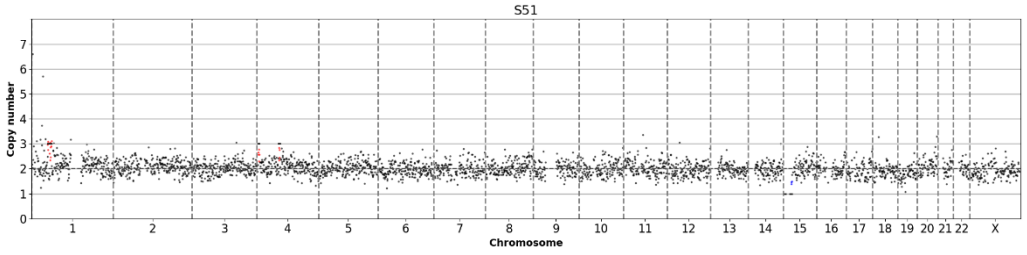
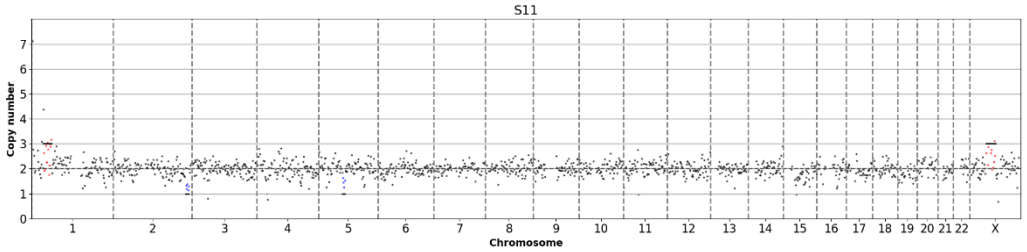
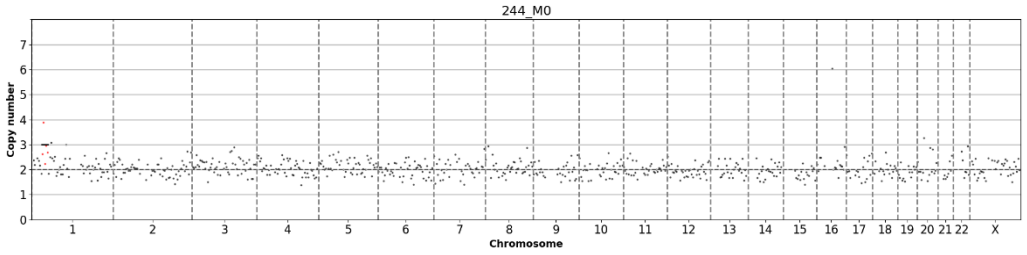
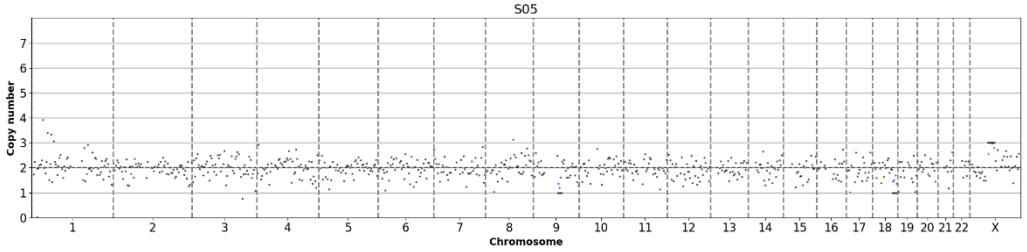
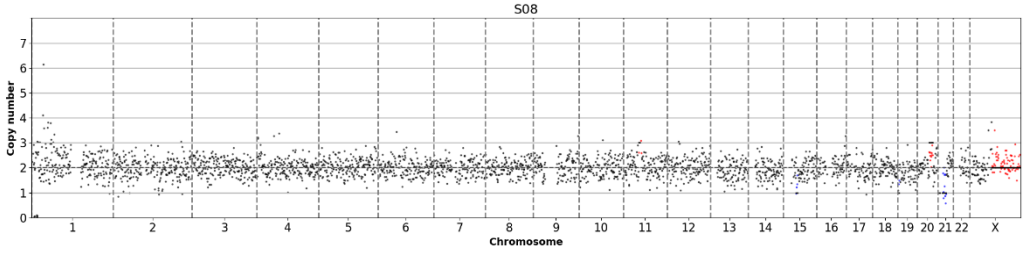
Sample info	Profile
MC-03-11-1117-5 M0 Lum undefined 935220 0.384	
MC03-09-746-6 M0 Lum undefined NA NA	
MC03-09-746-8 M0 Lum undefined 280367 0.289	
MC05-09-802-3 M0 Lum undefined 687988 0.323	
MC12-08-644-1 M1 Lum undefined 328945 0.257	
MC-03-11-1112-1 M0 Lum undefined 520178 0.449	

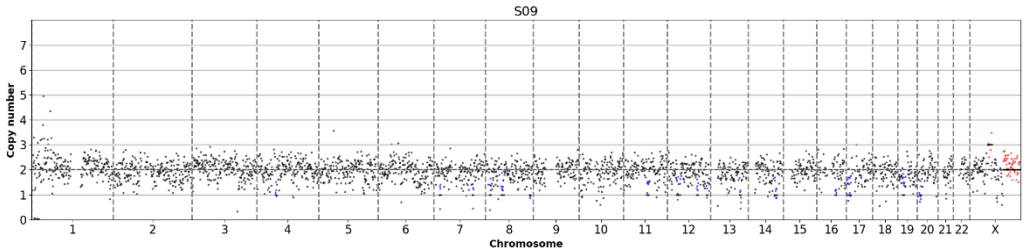
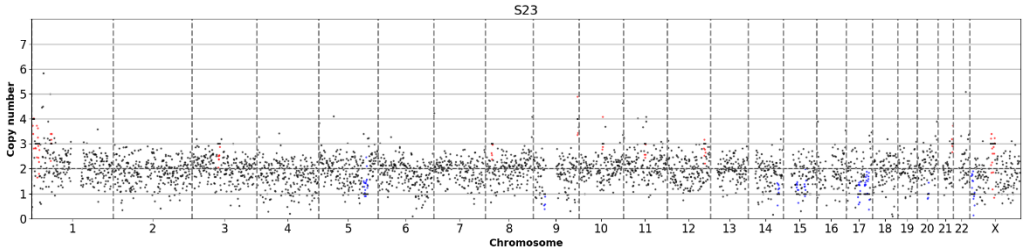
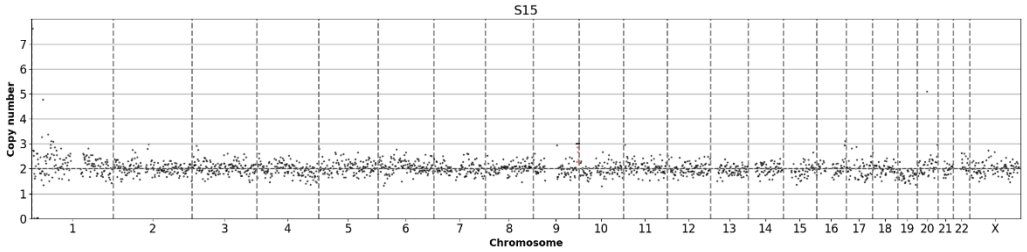
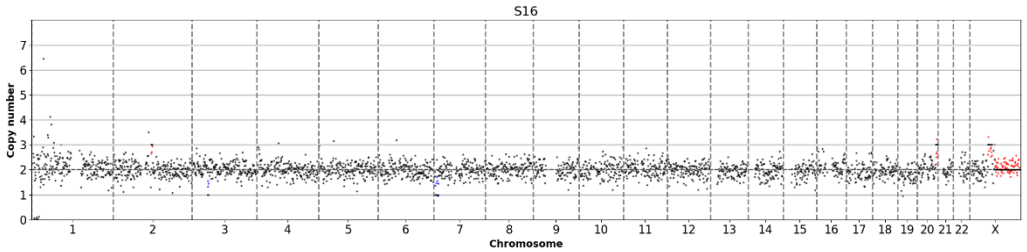
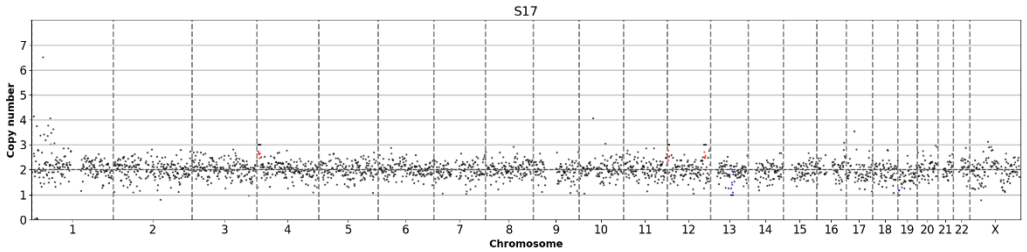
Sample info	Profile
M1-12-2175KM-2 M1 Lum undefined 960853 0.226	
M1 04/10-976-4 M1 Lum undefined 661883 0.248	
MC-05-10-983-1 M0 TNBC 591474 0.342	
MC-05-10-983-4 M0 TNBC 349536 0.306	
MC13-2365KM-1 M0 TNBC 1082757 0.298	
MC13-2512KM-1 M0 TNBC 968366 0.249	

Sample info	Profile
MC03-09-746-6 M0 LumA NA NA	
MC03-09-746-8 M0 LumA 280367 0.289	
MC03-09-741-8 M0 LumA NA NA	
MC07-09-843-1 M0 LumA 1239114 0.338	
MC13-2546KM-1 M0 LumA 853498 0.240	
MC15-0164KM-5 M0 LumA 760913 0.306	

Sample info	Profile
MC15-0164KM-6 M0 LumA 848091 0.277	
MC15-0164KM-9 M0 LumA 845726 0.321	
MC-11-1161-KM-Depl-3 M0 LumA 626985 0.402	
MC12-2081KM-1 M0 LumB 307637 0.411	
MC12-2081KM-12 M0 LumB 573913 0.377	
MC12-2081KM-13 M0 LumB 386418 0.375	

Sample info	Profile
MC-12-10-1067-3 M0 LumB 237467 0.253	
MC13-2361KM-1 M0 LumB 302982 0.438	
MC13-2367KM-1 M0 LumB 794655 0.299	
MC13-2384KM-1 M0 LumB 607474 0.311	
MC13-2427KM-1 M0 LumB 358536 0.366	
MC13-2477KM-1 M0 LumB 1026664 0.288	

Sample info	Profile
MC13-2477KM-2 M0 LumB 371398 0.347	
MC13-2477KM-11 M0 LumB 976599 0.241	
MC13-2477KM-12 M0 LumB 449446 0.265	
MC14-2698KM-1 M0 LumB 224097 0.422	
N12-2147KM-1 HD NCC 237841 0.478	
N12-2147KM-5 HD NCC 862959 0.317	

Sample info	Profile
N12-2165KM-2 HD NCC 889091 0.323	
N13-2462KM-2 HD NCC 1036318 0.434	
N13-2470KM-1 HD NCC 609048 0.248	
N13-2494KM-1 HD NCC 892496 0.266	
N13-2533KM-1 HD NCC 784449 0.298	

12.1.2 ISCN annotations of aberrant DCCs for *Progenetix*

Table 12-3 ISCN annotations of samples generated by LowPass-Seq.

Sample ID	Patient group	Cancer subtype	ISCN annotation for *Progenetix*
N12-2213KM-3	HD /NCC	-	rev ish dim (10q24.32q26.13, 17q22q24.1, 19p13.3p12, 22q12.3q13.2)
N13-2411KM-2	HD /NCC	-	rev ish enh (4p14q13.2, 5p15.31q12.3)
MC 01/11-1090-4	M0	LumB	rev ish enh (3q13.32q23, 7p22.3q11.22, 9p21.3p21.2, 11q14.3q25, 13q12.11q13.3, 13q14.3q21.31, 14q24.3q32.2, 15q25.1q26.3, 18p11.32q21.2, 20, 21q22.13q22.3, Xp22.32p11.4, Xp11.22q28) dim (1q21.2q31.1, 2q24.2q37.1)

Sample ID	Patient group	Cancer subtype	ISCN annotation for *Progenetix*
MC 02/11-1102-4	M0	LumB	rev ish enh (1q25.3q32.2, 7p22.3p14.3, 8q23.1q24.3) dim (15q12q24.3, 17q22q24.3, 18q12.2q23.3, 22q12.1q13.2, Xp22.12p11.4)
MC 03/11-1119-3	M0	LumB	rev ish enh (3p26.3p22.2, 3q21.2q21.3, 3q25.1q26.1, 3q26.32q29, 4q32.1q35.2, 5p15.33p15.1, 5q21.1q32, 6p21.2q14.3, 7p22.3q11.22, 11q11q13.1, 11q14.3q25, 14q24.3q32.2, 15q11.2q15.1, 15q21.3q26.3, 17q12q22, 20p13q11.22, X) dim (2q14.3q32.1, 3q26.1q26.32, 5q12.3q14.1, 6p25.2p21.2, 6q15q27, 7q11.23q32.3, 8p12p11.21, 12q24.12q24.31, 17p13.3q12, 18q12.3q22.3, 19p, 22q12.2q13.33)
MC 04/11-1125-5	M0	Triple Negative	rev ish enh (2p25.3p23.3, 2q12.1q14.3, 4p16.3q13.2, 6p21.2p21.1, 6p12.2q14.1, 8p23.3p21.3, 9p24.3p13.2, 11p15.5q13.1) dim (2p23.3p14, 3p26.3p13, 3q22.2q25.1, 4q31.21q35.1, 5q13.1q14.2, 7q31.1q31.33, 10q21.3q23.32, 13q13.3q14.3, 14q11.2q32.13, 15q14q15.1, 15q22.31q25.3)
MC 10/10-1030-6	M0	LumA	rev ish enh (1p34.3p31.1, 2p25.3q12.1, 4p16.3p15.2, 5q31.3q35.3, 7q21.11q36.3, 10q25.1q26.2, 11p15.3p14.3, 11q13.2q21, 11q23.1q25, 12q21.32q23.3, Xp22.31q21.1) dim (6p22.3p22.1, 16q21q22.1)
MC 11/10-1058-17	M0	LumA	rev ish enh (1p13.3q25.3, 2p25.3p11.2, 2q24.3q37.3, 3, 4q31.1q32.1, 6p22.1p21.1, 6q14.1q15, 8q13.3q21.13, 8q22.1q23.2, 9p24.3q33.3, 11p15.5q14.1, 12q13.11q21.2, 12q23.1q24.33, 13q12.11q14.3, 17, 21q22.11q22.3, Xp22.2p22.12) dim (1p36.31p35.3, 1p31.1p13.3, 4q32.1q35.1, 5q13.2q35.1, 10q23.33q25.1, 12p13.1q12, 14q21.1q23.3, 16q12.1q21, 18q21.32q22.3, 19p13.2p12, 20, 22q12.1q13.33)
MC 11/10-1058-2	M0	LumA	rev ish enh (8q21.13q22.1, 11q22.3q25, 16p13.2p12.3, 16q23.2q24.3) dim (2p23.3p14, 22p12q12.3, 2q33.1q33.3, 2q36.1q37.2, 4q32.3q34.3, 8q22.1q24.3, 13q14.3q21.32, 13q31.1q33.2, 19p13.3p12, 21q21.1q22.12, Xp22.11p11.4, Xq13.3q21.32
MC 11-2009 KM-2	M0	LumA	rev ish enh (1p21.3p12, 4q34.1q35.2, 5q33.3q35.3, 6q16.2q23.3, 9p24.3q21.13, 9q22.32q34.3, 10p15.3q21.2, 10q25.1q25.3, 18p11.32q11.2, 20, 22p11.22q13.1) dim (2p22.2p14, 3q13.2q21.2, 3q23q25.1, 4q28.1q32.2, 7p22.3p14.2, 8p23.1p21.2, 10q23.31q24.31, 11p11.2q13.3, 12q23.3q24.31, 15q11.2q21.3, 17q11.2q23.2, 19p13.3p12, 19q13.12q13.43, 21q22.11q22.3, Xp22.2p11.21)
MC03/09-741-1	M0	LumA	rev ish enh (6p12.1q15)
MC03/09-741-3	M0	LumA	rev ish enh (1p36.22p36.12, 1p32.3p31.3, 11q23.2q25) dim (4p14p12, 4q25q28.1, 4q31.1q31.23, 4q32.1q34.3, 5p14.3p13.3, 12q12q24.13, 14q24.1q24.3, 15q22.2q23, 16p12.2q23.1, 17p11.2q21.32, 17q22q24.2, 18p11.22q11.2, 19p13.3p13.11, 19q13.13q13.33, 22q)
MC04/08-356-2	M0	missing data	rev ish enh (1q21.3q41, 3q22.1q29) dim (2q14.1q22.3, 3p25.1p12.3, 4p16.3q23, 6p24.1p21.31, 6q14.1q27, 7q11.22q31.32, 9p24.3q33.1, 11q14.1q25, 12q12q24.33, 13, 19, 22, Xp22.32q24)
MC05/09-802-2	M0	Lum undefined	rev ish enh (1p35.2q25.3)
MC11-2012 KM-1	M0	LumB	rev ish enh (3p26.3p24.1, 3q13.13q23, 10q26.12q26.3, 18q12.3q21.1) dim (4p, 14q13.2q21.1, 17q22q24.2)
MC11-2051KM-11	M0	LumB	rev ish enh (1q31.3q41, 2p12q22.1, 3q26.32q29, 5p15.33p15.1, 5p13.3q13.1, 5q14.1q15, 6p25.3p22.3,

Sample ID	Patient group	Cancer subtype	ISCN annotation for *Progenetix*
			6p21.1q14.1, 7p15.3p14.3, 7p13q11.2, 8q24.22q24.3, 9, 10, 11p15.5p13, 11q14.1q25, 13q13.2q14.11, 13q32.1q34, 14q11.2q13.1, 14q23.1q32.31, 18, 20, 21q21.3q22.3, Xp22.32p22.12, Xq12q22.2) dim (4q28.2q32.3, 6q14.1q16.1, 7q33q36.1, 8p11.21q12.1, 8q13.1q21.2, 11q12.2q13.5, 12q21.2q22, 12q23.3q24.33, 19p13.3p12, 19q13.11q13.32, 22q12.3q13.33)
MC12-2081KM-2	M0	LumB	rev ish enh (1q23.2q25.3, 7p14.3p12.1) dim (2p24.1p21, 2q33.1q37.3)
MC12-2081KM-4	M0	LumB	rev ish enh (2p25.3p12, 19q13.11q13.32, 21q21.3q22.3) dim (4p15.1p13, 6p24.3p22.3, 7p22.3p12.3, 8q22.1q23.1)
MC12-2094KM-12	M0	Triple Negative	rev ish enh (2p25.3p22.3, 2p21q12.3, 5p15.33q22.1, 7p14.1q21.11, 7q21.13q32.2, 7q33q36.3, 8q11.21q24.3, 10q21.1q23.1, 14q11.2q23.1, 19p13.3p12, 20p13q13.13)
MC12-2096KM-1	M0	LumB	rev ish enh (3q13.33q23, 5q32q35.3, Xq23q28) dim (4q34.1q35.2, 5q14.2q15, 13q12.3q13.3, 13q14.11q31.3, 15q25.1q26.1, 17p13.1p12, 17q23.2q24.3, 19p13.3p13.11, 20p12.1p11.21)
MC13-2308KM-1	M0	Triple Negative	rev ish enh (2p25.3p24.1, 8q24.21q24.3, 20p12.1p11.21, 20q11.23q13.31) dim (1p36.23p31.3, 1q42.12q44, 3p22.2p14.3, 7q11.21q36.3, 15q11.2q22.2, 15q25.2q26.3, 16p13.13q24.3, 21, Xq21.33q24)
MC13-2336KM-1	M0	LumA	rev ish enh (3p26.3p24.2, 5q13.3q35.3, 6q24.3q27, 7p22.3q11.22, 7q31.32q34, 9q21.32q34.3, 10q25.1q26.11, 13q22.3q32.3, 14q21.1q23.3, 20p11.23q12, 21q22.12q22.3, Xp22.32p22.11, Xq21.1q27.3) dim (2p25.1p22.3, 3p13p12.3, 3q12.2q13.2, 4q22.3q25, 5p15.33p13.3, 10p15.3p13, 10p12.2p11.21, 12p13.31p12.3, 12q14.1q21.1, 12q22q24.13, 13q14.11q21.1, 17q21.2q24.2, 18q11.1q12.1, 18q12.3q22.1, 19p13.3p12, 19q13.31q13.43
MC13-2361KM-2	M0	LumB	rev ish enh (4q31.21q32.1, 8q22.1q24.3) dim (6p25.3p21.32)
MC13-2384KM-2	M0	LumB	rev ish enh (11q23.2q23.3, 12q23.3q24.33) dim (1q21.2q25.1, 1q42.3q44, 7q32.1q36.1, 9q22.2q33.1, 19p13.3p12)
MC13-2427KM-11	M0	LumB	rev ish enh (1p31.3p31.1, 8q24.22q24.3) dim (9q33.1q34.2, 10q22.2q22.3, 14q31.3q32.33, 15q11.2q24.1, 15q25.1q26.1, 19p13.3p13.11, 21q21.1q22.3)
MC13-2477KM-13	M0	LumB	rev ish dim (19p13.3p12)
M1 01/09-689-1	M1	Lum undefined	rev ish enh (1p35.3p31.3, 1p13.3q44, 2, 5p15.33q21.3, 5q23.2q31.3, 6p21.2q22.31, 6q24.3q25.3, 7q11.22q21.2, 8p12q12.1, 8q13.3q24.3, 11p15.5p15.1, 15, 19q, 20) dim (1p36.32p35.3, 1p31.3p13.3, 4p15.33p14, 7q21.2q35, 9p24.3q21.13, 13, 18)
M1 01/09-689-10	M1	Lum undefined	rev ish enh (1p13.1q44, 7p22.3p11.2, 7q32.3q36.3, 11p15.5p12) dim (1p36.32p35.1, 1p31.3p13.3, 6q23.3q27, 7q21.2q32.3, 8p23.3p21.1, 11q14.1q25, 13, 16q12.1q24.3, 17p13.3q11.2, 18)
M1 01/09-689-11	M1	Lum undefined	rev ish enh (1q21.3q44, 11p15.5p11.2)
M1 01/09-689-2	M1	Lum undefined	rev ish enh (1q21.2q32.3, 7p22.3p11.2, 10p15.3q22.2, 11p15.3p15.1, 11p14.2p13, 11q13.2q14.1, 15q11.2q13.3) dim (1p36.33p35.1, 1p31.1p13.2, 2q21.3q24.3, 4q13.1q21.21, 7q21.11q31.2,

Sample ID	Patient group	Cancer subtype	ISCN annotation for *Progenetix*
			8p23.3p21.1, 10q23.1q26.3, 11p15.5p15.4, 11p15.1p14.2, 11p13q13.2, 11q14.1q25, 13, 16q12.1q24.3, 17p13.3q12, 17q21.32q24.1, 18, 19q13.31q13.43, 21q21.1q22.11, Xq21.1q28)
M1 01/09-689-3	M1	Lum undefined	rev ish enh (1q21.1q32.3, 7p22.3p11.2, 10p, 10q11.21q23.1) dim (1p36.33p34.3, 1p31.3p13.2, 7q21.3q31.2, 8p23.3p21.1, 10q23.1q26.3, 11p15.5p15.4, 11p14.3p14.2, 11p13p11.12, 11q14.1q25, 13, 16q11.2q23.2, 17p13.3q21.1, 18q, 21q21.1q22.12, Xq21.1q28)
M1 01/09-689-4	M1	Lum undefined	rev ish enh (1q, 7p22.3p12.1, 7q21.11q21.2, 7q31.2q36.3, 11p15.5p11.12, 12p12.3q21.32, 16p13.3p11.2) dim (1p36.33p35.1, 1p31.3p13.2, 6q23.2q27, 7q21.2q31.2, 8p23.3p21.1, 11q11q13.2, 11q14.1q25, 13q12.11q21.1, 13q21.33q34, 16p11.2q23.1, 17p13.3q11.2, 17q23.2q25.3, 18, 21q21.1q22.2)
M1 01/09-689-5	M1	Lum undefined	rev ish enh (1q21.2q31.1, 1q31.3q41, 7p22.3p11.2, 11p15.3p15.1, 11p14.2p13, 12p13.32p13.31, 12p12.3p12.1) dim (1p36.33p35.1, 1p31.3p13.3, 2, 7q21.3q31.2, 8p23.3p21.1, 10p12.1q22.1, 10q23.1q26.3, 11p13q13.2, 11q14.1q25, 13q12.11q21.2, 13q21.33q34, 15q24.2q26.3, 16q, 17p13.3q12, 18, 21q21.1q22.11, Xq21.32q28)
M1 01/09-689-6	M1	Lum undefined	rev ish enh (1q21.1q32.2, 2q32.1q33.1, 7p22.3p12.1, 8q21.2q23.3, 10q21.2q23.1, 11p15.4p15.1) dim (1p36.33p35.1, 8p23.3p12, 10q23.1q26.3, 11q14.1q24.33, 13q12.11q14.2, 13q22.2q31.1, 13q31.3q34, 15q24.2q26.3, 16q, 17p13.3q12, 18, 21q21.1q22.11, Xq21.1q28)
M1 01/09-689-7	M1	Lum undefined	rev ish enh (1q, 11p15.5q12.1, 16p13.3p11.2) dim (1p36.32p35.1, 1p31.3p13.3, 3p22.2p14.1, 7q21.2q31.2, 8p23.3p12, 11q12.1q13.4, 11q14.1q25, 13, 16q12.1q23.1, 17p13.3q12, 18, 21q21.1q22.11)
M1 01/09-689-8	M1	Lum undefined	rev ish enh (1p33p31.3, 1q21.2q25.3, 1q32.1q44, 2q33.3q37.3, 4p16.3q22.1, 4q24q32.1, 6p25.3q15, 7p21.3q21.2, 8p21.1q24.3, 11p15.5p11.2, 11q23.1q25, 12q13.11q24.33, 14, 15q22.2q26.1, 17q12q23.2, 20, 22q11.21q12.2, X) dim (5p15.33p13.3, 8p23.3p21.1, 13q14.3q31.1, 16p13.3p13.12, 16p11.2q23.3, 17q23.2q25.3)
M1 01/09-689-9	M1	Lum undefined	rev ish enh (5p14.1p13.3, 8q12.3q22.1, 11p15.5p11.2, 11q13.4q14.1, 12q13.3q21.1, 12q21.31q24.33) dim (1p36.12p34.2, 1q24.1q44, 2p25.3p24.1, 2p23.2p16.1, 2p13.1q11.2, 2q14.1q14.2, 2q31.1q31.1, 5p15.33p15.1, 5q14.3q35.3, 6, 10q23.1q23.31, 13, 14, 16q12.1q24.3, 19, 21, 22, Xp22.13p11.3, Xq21.1q21.33, Xq23q28)
M1 03/09-773-2	M1	Triple Negativ	rev ish enh (1p21.2q44, 4q26q32.1, 6p25.3p22.3, 6p22.1p21.1, 8q11.23q23.1, 11p15.5q13.3, 12q13.13q24.33, 16p13.3p11.2, 17p11.2q25.3, 20p11.23q13.33) dim (2p25.3p23.3, 2q14.3q31.1, 3p24.1p12.2, 4q32.1q35.2, 5q21.2q33.3, 6q12q27, 9p24.3p21.3, 9q21.13q33.1, 10q25.1q26.3, 13q12.11q21.1, 13q32.3q34, 14q22.2q31.3, 16q21q24.1, 18q, 20p13p11.23, 21, 22q11.21q12.3)
M1-12-2232KM-27	M1	LumA	rev ish enh (1q23.3q25.2, 3p24.3p24.1, 4q22.1q24, 5p14.1p13.2, 6p25.3p22.3, 8p12q24.3, 12p13.33p12.3, 12q13.2q21.2, 18p11.21q21.2)
MC14-2675KM-2	M1	LumB	rev ish dim (X)

Sample ID	Patient group	Cancer subtype	ISCN annotation for *Progenetix*
MC14-2675KM-3	M1	LumB	rev ish dim (X)
MC14-2675KM-4	M1	LumB	rev ish dim (X)
Met02/10-947-1	M1	Triple Negativ	rev ish enh (1p32.1p12, 3p25.1p24.2, 4p14q22.1, 4q24q25, 8q21.3q23.1, 8q24.12q24.23, 9, 12q14.1q24.11, 13q14.11q14.3, 17p12q12, 18p11.32q12.1) dim (14q12q22.3, 14q31.3q32.33, 15q11.2q15.2, 15q25.3q26.3, 16, 19p13.3p12, 20)
Met02/10-947-4	M1	Triple Negativ	rev ish enh (1p36.33p31.3, 1q21.3q31.1, 2p25.3q14.1, 3p26.3p21.31, 4p16.3q21.21, 5p15.33p12, 5q31.1q35.3, 6p25.3p22.3, 6p21.31p12.1, 6q21q22.31, 8q21.3q24.3, 9q22.2q34.3, 10q11.21q23.1, 11, 12p13.32q23.33, 14q12q32.33, 15q22.2q23, 15q24.3q26.1, 16p13.3q13, 19q13.11q13.43, 20, 21q22.11q22.3, Xq26.3q28) dim (3p21.31p12.2, 6q14.1q22.2, 8p23.3p12, 9p24.3p21.3, 12p13.31p12.1, 16q13q23.1, Xp22.2p21.1)

Table 12-4 ISCN annotations of samples generated by mCGH. Annotations by (Haunschild, 2013).

Sample ID	Patient group	Cancer subtype	ISCN annotation for *Progenetix*
MC 11/10- 1049-2	M0	Lum undefined	rev ish dim (3q21qter, 5qcenq23, 9, 11q14qter, 15q24qter, 17pterp12, 21q, Xq26q27) enh(1pterp36.1, 10q, 19pterq13.2, 20q12qter, 22q12qter)
MC 01/11-1073-3	M0	Lum undefined	rev ish dim (3p13pcen, 17q21) enh(20q13.3)
MC 03/11-1112-5	M0	Lum undefined	rev ish dim (1pterp34.1, 9q34, 19) enh(4q31.3qter, 5pterp12, 5q14q23, 5q32q34, 9pterp21, 13q31q33, 17pterq21, 18qcenq21, 21q, Xp22.1p21)
MC 03/11-1117-4	M0	Lum undefined	rev ish dim (11p)
MC 10/10-1029-1	M0	LumA	rev ish dim (5qterq15.1) enh(1pterp33, 6pterp24)
MC 10/10-1030-11	M0	LumA	rev ish dim (1q22q25, 3qcenq13.3, 4q13q21, 4q31.1q31.2, 15q21) enh(6p21.2p21.1, 16q24, 17p11.2, 17q23qter, 19pterq13.1, 20qcenq13.1, 21q22, 22q11.2, 22q13)
MC 10/10-1030-3	M0	LumA	rev ish dim (6p12q21) enh(21q22)
MC 11/10-1058-2	M0	LumA	rev ish dim (2p23p21, 2q37, 4q31.3qter, 13q21q31) enh (1q12, 11qcenq13, 17p11.2q11.2)
MC 11/10-1058-3	M0	LumA	rev ish dim (3p14pcen, 3q25, 4pterp15.1, 4q25q26, 6q13q23, 9q21q22, 10q11.2q21, 15q21q22) enh (1q32qter, 2pterq21, 2q37, 11q12q13, 16pterq13, 17p, 17q22qter, 19pterq11, 21q22)
MC 12/10-1069-2	M0	LumA	rev ish dim (10q21q23, 16q) enh(2pterp24, 4p16, 6q25qter, 12pterp12, 12q24.1qter, 18pterp11.2, 20q13.1qter)
MC 01/11-1074-2	M0	LumA	rev ish dim (1p36.1p34.2, 2q34qter, 17q21)
MC 11-1194-4	M0	LumA	rev ish dim (5q13q15, 9, 17, 22q13, Xp21) enh (1p31pcen, 2pterp24, 2p16p13, 3pterp21, 3q21, 4pterp15.1, 7p15, 8pterp12, 11q23qter, 12q14qter, 21q)
MC03/09-741-3	M0	LumA	rev ish dim (4q31.2q34, 12qcenq24.1, 16q, 17q11.2q21, 22q13) enh(10pterp14, 11q23qter)
MC03/09-741-4	M0	LumA	rev ish dim (3qcenq13.3, 6p11.2q12) enh()
MC07/09-843-1	M0	LumA	rev ish dim (1p33p32)
MC 12/10-1059-2	M0	LumB	rev ish dim (5q12q31) enh(16p11.2)
MC 01/11-1088-10	M0	LumB	rev ish dim (5p12q12, 7qcenq22, 15q11.2, 16p13.1pcen, 16q21q23) enh (2q11.2q14.3, 2q33qter, 3q21q25, 6pterp22, 13q12)
MC 01/11-1088-2	M0	LumB	rev ish dim (2q21q32.2, 5pterp15.2, 7pterp14, 7qcenq21, 9qcenq31, 10q26, 12q24.3, 13q34,

Sample ID	Patient group	Cancer subtype	ISCN annotation for *Progenetix*
			14q11.2q21, 15q11.2q15, 17q21q24) enh (1q42qter, 3pterp21, 3q22qter, 6q14q16, 6q23q24, 11pterp12, 11q23qter, 16q, 18q21, 21q22, Xp11.4p11.2, Xq21q24, Xq27qter)
MC 01/11-1090-4	M0	LumB	rev ish dim (1q21q24, 2q31q33, 5q21qter, 6p21.3q12, 6q24q25, 9q33qter, 19pterp13.1, 22q11.2q12) enh(2pterp23, 4p16, 7p22, 7p14pcen, 10pterp13, 16qcenq13, 20q12qter, 21q22, Xpterp22.2, Xp11.4qter)
MC 03/11-1111-1	M0	LumB	rev ish dim (1p36.1p34.1, 7qcenq21, 7q33qter, 12p13, 15q22, 19) enh(4pterp15.1, 4q27q28, 4q35, 6q, 9pterp13, 14qterq21)
MC 03/11-1111-5	M0	LumB	rev ish dim (7q33qter, 12q23q24.1, 17pterq21, 19pterp13.1, 22q, Xpterp22.1) enh(3pterp25, 3p12, 4pterp15.1, 4q21q26, 4q31.3qter, 6qcenq13, 9p, 18q22qter)
MC 03/11-1119-1	M0	LumB	rev ish dim (8q12q13, 11p15, 15q22q24) enh (2q37, 18pterp11.2)
MC 03/11-1119-4	M0	LumB	rev ish dim (2p23p14, 3pterq13.2, 3q27qter, 10pterp11.2, 11pterq12, 14q21q24, 15q22, 16pterp12, 17qcenq21, 19q13.2q13.3, 20) enh (1p31p22, 1q32qter, 2q31qter, 4q31.1qter, 5q13qter, 6q25qter, 7p15p13, 7q21qter,9, 12q15q24.1, 13q21qter, 21q22, X)
MC 04/11-1125-5	M0	TNBC	rev ish dim (1pterp34.1, 3pterp14) enh (9pterq13, 12p13, 18, Xq26qter)
MC 05/10-983-5	M0	TNBC	rev ish dim (3p14pcen) enh (20q13.2qter)
M1 01/09-689-3	M1	Lum undefined	rev ish dim (1p32p22, 6q26, 10q25qter, 11q23qter, 13q, 16q, 17pterq11.2, 18, Xq24qter) enh (1p13q41, 3q13.3q21, 7pterq11.2, 10pterq22, 11q13, 20pterp12)
M1 01/09-689-4	M1	Lum undefined	rev ish dim (1p31p22, 11q23qter, 13q, 17pterp12) enh (1p13qter, 11q13q14)
M1 01/09-689-5	M1	Lum undefined	rev ish dim (1pterp36.1, 1p31p22, 2, 11q23qter, 13q, 16q, 17p, 18) enh (1p13q42, 3p21q26.2, 5p13pcen, 7pterq11.2, 10pterp14, 11pterp14, 11q13q22, 12pterq13, 17q21q23, 19, Xq13)
M1 01/09-689-8	M1	Lum undefined	rev ish dim (3pterp25, 3q21qter, 5pterp13, 5q23qter, 6q26qter, 7q36, 8pterp21, 9, 10q21q22, 13q14q21, 15q11.2q15, 16, 18q) enh (1p21qter, 2pterp24, 4pterq34, 6pterq22, 7pterq11.2, 8q23qter, 11, 12p12qter, 14q, 17q21, 19, 20, 21, 22, X)
M1 03/11-1122-2	M1	LumB	rev ish dim (1p36.3, 14q22qter, 15q21qter, 17, 19pterp13.2, 19q13.2q13.3) enh (2p23p21, 2q12q24, 2q36qter, 3p12, 4, 6q16qter, 7p13qter, 9pterq21, 10, 13q12qter, 16p11.2q21, 18)
M1 03/11-1122-4	M1	LumB	rev ish dim (1pterp36.2, 2p23p22, 3p21pcen, 6p22, 7pterp15, 7q32qter, 9q22qter, 10q22qter, 12p13, 14q24qter, 15q11.2qter, 19, 21q22, 22q11.2qter) enh (1p13qter, 2q22q31, 4pterq13, 4q28qter, 5pterp15.1, 5q31qter, 6qcenq13, 6q25qter, 8, 11pterp11.2, 11q14qter, 12p11.2q23, 13q21qter, 16qcenq12.2, 18q, 20p11.2q13.1, Xpterp21)
Met02/10-947-1	M1	TNBC	rev ish enh (6pterp23, 9q34, 17, 22q)
Met02/10-947-4	M1	TNBC	rev ish dim (8pterp22) enh (7p15p13, 9p21p13, 12q13q23, 19p13.1q13.1, 20p11.2q12, 20q13.2qter, 22qcenq12)
Met02/10-947-8	M1	TNBC	rev ish dim (3p21q12, 8pterp21, 9p24p23, 10pterp11.2, 13q) enh (2pterp23, 7p15p11.1, 8q22qter, 9p21pcen, 9q22qter, 11pterq13, 11q24qter, 12q13q22, 14qcenq13, 16p, 17q22qter, 19p13.1qter, 20, 21q22, 22q)

Sample ID	Patient group	Cancer subtype	ISCN annotation for *Progenetix*
M1 03/09-773-2	M1	TNBC	rev ish dim (2q21q31, 3p21p13, 5q23q34, 6qcenq16, 9pterp22, 10q25qter, 11q24qter, 13qcenq14, 13q32qter, 14q24qter, 18q, 19qcenq13.1) enh (1q, 4p16, 6pterp21.2, 8p11.2qter, 11pterp11.2, 12q13qter, 16pterp13.2, 17, 20q13.2qter)

12.1.3 CNA statistical results

The following tables list the raw data of all CNAs tested in the pairwise comparisons. The p-value columns represent the initial values calculated by Fisher's exact test. The ranks are based on these values and were used to perform the multiple comparisons correction. which resulted in the adjusted p-values listed in the last column.

Table 12-5 Adjusted p-values for pair I: M0 EpCAM$^+$ vs. M0 CK$^+$.

Aberration	p-value	Rank	Adjusted p-value
1q gain	0.00016	1	0.00288
1p loss	0.5	16	0.5625
5 gain	0.015	3	0.09
6 loss	0.1572	7	0.40422857
7 gain	0.1441	6	0.4323
8p gain	0.43	15	0.516
8p loss	0.05825	5	0.2097
9p gain	0.2998	13	0.41510769
11 loss	0.2998	13	0.41510769
12 gain	0.266	10	0.4788
13 loss	0.266	10	0.4788
15 gain	0.203	9	0.406
15 loss	0.62	18	0.62
16p gain	0.0496	4	0.2232
16q loss	0.55	17	0.58235294
19p loss	0.0086	2	0.0774
20 gain	0.266	10	0.4788
21qter gain	0.187	8	0.42075

Table 12-6 Adjusted p-values for pair II: M0 EpCAM$^+$ vs. M1 EpCAM$^+$.

Aberration	p-value	Rank	Adjusted p-value
1p loss	0.011006	12	0.02017767
1q gain	0.00001129	2	0.00012419
7p gain	0.01574	16	0.0216425
7q internal loss	0.3555	22	0.3555
8p loss	0.000008541	1	0.0001879
8q gain	0.0053893	10	0.01185646
10p gain	0.3269	21	0.34246667
10q loss	0.0148	14	0.02325714
11p gain	0.00155	7	0.00487143
11q loss	0.0053893	10	0.01185646
12q gain	0.00179991	8	0.00494975

Aberration	p-value	Rank	Adjusted p-value
13 loss	0.00004465	3	0.00032743
14q loss	0.1087	18	0.13285556
16p gain	0.0146	13	0.02470769
16q loss	0.00049997	5	0.00219987
17p loss	0.015738	15	0.0230824
18 loss	0.0005622	6	0.0020614
19p loss	0.0881	17	0.11401176
20 gain	0.11061	19	0.12807474
21qcen loss	0.004444	9	0.01086311
21qter gain	0.1161	20	0.12771
Xq loss	0.00046431	4	0.00255371

Table 12-7 Adjusted p-values for pair III: M1 EpCAM$^+$ vs. M1 CK$^+$.

Aberration	p-value	Rank	Adjusted p-value
1p loss	0.092	9	0.2248889
2 gain	1	22	1
4 gain	0.68	19	0.7873684
4 loss	0.1973	14	0.3100429
5 q loss	0.179	13	0.3029231
6q loss	0.0095	3	0.0696667
7p gain	0.037	7	0.1162857
7q int loss	0.0097	4	0.05335
8p loss	0.33	17	0.4270588
8q gain	0.098	10	0.2156
9 loss	0.0649	8	0.178475
10p gain	0.692	20	0.7612
10p loss	0.2864	16	0.3938
11p gain	0.00000233	1	5.126E-05
12q int gain	0.035	6	0.1283333
14 loss	0.013	5	0.0572
15q gain	0.678	18	0.8286667
17q gain	0.264	15	0.3872
18 gain	0.7009	21	0.7342762
19q gain	0.13267	12	0.2432283
21qcen loss	0.0052	2	0.0572
22 loss	0.11	11	0.22

Table 12-8 Adjusted p-values for pair IV: M0 CK$^+$ vs. M1 CK$^+$.

Aberration	p-value	Rank	Adjusted p-value
1q gain	0.45178	21	0.45178
3pcen loss	0.001655	5	0.006951
5q internal loss	0.024027	11	0.045869727
5qcen gain	0.0057755	7	0.0173265
6q loss	0.027729	12	0.04852575
7 gain	0.1108551	15	0.15519714
8p loss	0.022398	10	0.0470358

Aberration	p-value	Rank	Adjusted p-value
8q gain	4.02544E-09	1	8.45343E-08
9 loss	0.009431	8	0.024756375
10 loss	0.163794	19	0.181035474
12qter gain	0.03715	14	0.055725
13 loss	0.0011053	3	0.0077371
14q loss	0.000059545	2	0.000625223
15qcen gain	0.14327	17	0.176980588
16p gain	0.28849	20	0.3029145
16q loss	0.0053707	6	0.01879745
17p loss	0.1255	16	0.16471875
17q gain	0.0011053	3	0.0077371
18 loss	0.14834	18	0.173063333
19p loss	0.032146	13	0.051928154
20 gain	0.00969	9	0.02261

12.2 Differentially expressed genes between LumA and LumB

Table 12-9 List of 815 genes down-regulated in LumB. Genes are sorted by adjusted p-value.

Gene symbol	logFC	AveExpr	P.Value	adj.P.Val
SLC22A20P	-0.01703	1.46579	9.21406E-05	6.52088E-03
PLA2G4C	-0.01070	1.05730	1.91407E-04	9.94338E-03
LCT	-0.02213	2.19555	1.93829E-04	9.99053E-03
TGM4	-0.01117	1.20909	2.49240E-04	1.18654E-02
PLEK2	-0.01525	1.43099	2.58499E-04	1.21142E-02
SEMA6D	-0.01134	1.16230	2.74563E-04	1.27447E-02
CCDC102B	-0.01068	1.08988	2.92623E-04	1.29199E-02
CCDC36	-0.01414	1.43494	3.26097E-04	1.37513E-02
C2orf16	-0.01635	1.81489	3.57649E-04	1.43861E-02
C3orf67	-0.01658	1.48416	4.79394E-04	1.68305E-02
TLR6	-0.01644	1.45215	5.11597E-04	1.75438E-02
ANKRD13B	-0.01202	1.26185	5.26379E-04	1.78254E-02
OR10A2	-0.01672	1.60686	5.32022E-04	1.79898E-02
ANGPTL1	-0.01089	1.12884	5.54436E-04	1.83746E-02
IFNL2	-0.01215	1.22861	5.77802E-04	1.86880E-02
ZNF831	-0.01010	1.12509	6.28163E-04	1.96217E-02
CTNS	-0.01435	1.30325	6.31285E-04	1.96350E-02
ZNF729	-0.01035	1.06093	6.48373E-04	1.99228E-02
SERPINF1	-0.02008	1.63358	6.50803E-04	1.99228E-02
PKDREJ	-0.01285	1.42686	6.51650E-04	1.99228E-02
ST18	-0.01174	1.31772	6.55227E-04	1.99228E-02
CD6	-0.01002	1.02373	6.77436E-04	2.03448E-02
GRHL1	-0.01264	1.16722	7.00516E-04	2.06319E-02
LCTL	-0.01086	1.07129	7.42979E-04	2.12956E-02
AGBL3	-0.01124	1.16976	8.06366E-04	2.21381E-02
LRP4	-0.01971	2.12976	8.16543E-04	2.22489E-02
FYB1	-0.01056	1.19410	8.31312E-04	2.24280E-02
CHIT1	-0.01434	1.63488	9.15382E-04	2.37540E-02
KIF17	-0.01151	1.20970	9.17675E-04	2.37540E-02
CHRFAM7A	-0.01107	1.08770	9.23424E-04	2.37580E-02
CYP8B1	-0.01045	1.14287	9.27184E-04	2.37795E-02
GABRE	-0.01129	1.26947	9.36744E-04	2.39431E-02
P3H3	-0.01300	1.42865	9.49621E-04	2.40020E-02
TBC1D22A-AS1	-0.02232	2.13610	9.73876E-04	2.42784E-02

Gene symbol	logFC	AveExpr	P.Value	adj.P.Val
IGSF10	-0.01559	1.68407	9.75445E-04	2.42784E-02
NHSL1	-0.01455	1.62622	9.81924E-04	2.43725E-02
GARNL3	-0.01238	1.51127	1.00103E-03	2.46178E-02
NLRP7	-0.01096	1.22424	1.01647E-03	2.48252E-02
PEAK3	-0.02862	3.60802	1.02907E-03	2.50007E-02
B4GALNT2	-0.01012	1.13225	1.05414E-03	2.53325E-02
HYDIN	-0.01898	2.56836	1.06938E-03	2.55478E-02
SMO	-0.01133	1.24356	1.10641E-03	2.60468E-02
TRIO	-0.02042	2.68833	1.11258E-03	2.61199E-02
PADI6	-0.01349	1.34602	1.11613E-03	2.61282E-02
HIST1H1C	-0.01209	0.96432	1.11614E-03	2.61282E-02
TSBP1	-0.01004	1.07185	1.12766E-03	2.62718E-02
CARMIL3	-0.01291	1.50162	1.15371E-03	2.66475E-02
SH3RF2	-0.01013	1.09304	1.15617E-03	2.66475E-02
PLEKHG3	-0.01274	1.45214	1.18467E-03	2.69160E-02
FAM171A1	-0.01499	1.79456	1.20470E-03	2.72269E-02
TIRAP	-0.01028	0.93303	1.21740E-03	2.72697E-02
LRRC37A6P	-0.01108	1.19774	1.23396E-03	2.73552E-02
PROM2	-0.01199	1.39217	1.23466E-03	2.73552E-02
MROH9	-0.01167	1.31388	1.24010E-03	2.74308E-02
TSKS	-0.01360	1.44262	1.24032E-03	2.74308E-02
PRDM7	-0.01253	1.34151	1.26526E-03	2.76946E-02
SBF2	-0.01145	1.40531	1.29217E-03	2.80280E-02
TMPRSS2	-0.01011	1.18723	1.30645E-03	2.82285E-02
CYP4F11	-0.01041	1.12830	1.34490E-03	2.86286E-02
LRP1B	-0.01723	1.90499	1.35945E-03	2.87352E-02
PPFIA3	-0.01205	1.33901	1.36638E-03	2.87352E-02
BPIFB3	-0.01030	1.10702	1.41628E-03	2.92432E-02
SLC6A12	-0.01162	1.30194	1.43755E-03	2.93718E-02
NRCAM	-0.01066	1.29227	1.44606E-03	2.94431E-02
PPARGC1A	-0.01155	1.35051	1.44707E-03	2.94431E-02
PLD2	-0.01133	1.27750	1.45640E-03	2.95589E-02
LONRF3	-0.01207	1.32126	1.47394E-03	2.96928E-02
BTN1A1	-0.01557	1.23701	1.48510E-03	2.97596E-02
AFF2	-0.01399	1.58562	1.49188E-03	2.97596E-02
LAMC2	-0.01338	1.58556	1.50500E-03	2.98775E-02
RAPGEF4	-0.01010	1.14481	1.51219E-03	2.99203E-02
SHROOM2	-0.01398	1.56738	1.51874E-03	2.99651E-02
CCNB3	-0.01122	1.21488	1.52101E-03	2.99651E-02
ACTL8	-0.01094	1.19300	1.56103E-03	3.03698E-02
GUCY2F	-0.01055	1.20435	1.58264E-03	3.06681E-02
ARHGAP6	-0.01191	1.31254	1.58758E-03	3.07393E-02
NRG1	-0.01218	1.52659	1.59398E-03	3.07658E-02
HSP90AB4P	-0.01257	1.28443	1.59861E-03	3.08306E-02
ATP2C2	-0.01298	1.48854	1.60033E-03	3.08394E-02
RABEP2	-0.01215	1.28113	1.60745E-03	3.09005E-02
ERN2	-0.01290	1.43980	1.61276E-03	3.09325E-02
PCDHGB3	-0.01096	1.20675	1.62955E-03	3.10351E-02
WDR72	-0.01157	1.38949	1.63287E-03	3.10499E-02
OR2J3	-0.01085	1.18424	1.64852E-03	3.12002E-02
FAT4	-0.01731	2.05858	1.65141E-03	3.12002E-02
FCGR3B	-0.01076	1.17570	1.65713E-03	3.12002E-02
GLT8D2	-0.01395	1.08873	1.66810E-03	3.13195E-02
CCDC190	-0.01095	1.24540	1.66928E-03	3.13195E-02
PLA2G4E	-0.01043	1.14273	1.67143E-03	3.13197E-02
CILP	-0.01299	1.54162	1.67352E-03	3.13336E-02
CXCR1	-0.01124	1.24252	1.69163E-03	3.15288E-02
PDZD2	-0.01823	2.58562	1.69980E-03	3.15606E-02

Gene symbol	logFC	AveExpr	P.Value	adj.P.Val
CD209	-0.01090	1.19636	1.71281E-03	3.15860E-02
SHC4	-0.01082	1.23688	1.71859E-03	3.16334E-02
SPTBN4	-0.01436	1.86595	1.72198E-03	3.16334E-02
GABBR1	-0.01834	1.97038	1.73386E-03	3.17199E-02
TRPV1	-0.01286	1.23993	1.73602E-03	3.17199E-02
ADGRG5	-0.01051	1.18638	1.74505E-03	3.17964E-02
AC024940.2	-0.01474	1.93410	1.75437E-03	3.18474E-02
TSPAN4	-0.01706	1.48610	1.75773E-03	3.18846E-02
PDE3A	-0.01025	1.13940	1.76246E-03	3.19207E-02
MAPKBP1	-0.01351	1.64639	1.76364E-03	3.19207E-02
PSG3	-0.01124	1.28509	1.77560E-03	3.19364E-02
PCDH11X	-0.01303	1.44406	1.77583E-03	3.19364E-02
ACOT11	-0.01020	1.24296	1.78143E-03	3.19364E-02
MROH5	-0.01020	1.11686	1.78764E-03	3.19364E-02
UGT3A1	-0.01034	1.20518	1.79139E-03	3.19364E-02
CSMD2	-0.01938	2.36972	1.80233E-03	3.20363E-02
TECPR2	-0.01703	2.04616	1.80328E-03	3.20363E-02
LAMA4	-0.01363	1.60045	1.81824E-03	3.21009E-02
KIR3DL1	-0.01156	1.36415	1.82277E-03	3.21009E-02
MYH8	-0.01399	1.62508	1.83231E-03	3.21009E-02
UNC79	-0.01512	1.81679	1.85574E-03	3.22490E-02
PIK3C2B	-0.01553	1.93514	1.86353E-03	3.23153E-02
EPHA3	-0.01053	1.19698	1.87011E-03	3.23834E-02
DALRD3	-0.01559	1.60369	1.87783E-03	3.24481E-02
ADAMTS6	-0.01032	1.18770	1.89713E-03	3.26505E-02
LOXHD1	-0.01647	2.04025	1.89913E-03	3.26505E-02
CAPN8	-0.01315	1.47261	1.89964E-03	3.26505E-02
MPP3	-0.01079	1.27301	1.90900E-03	3.26638E-02
TGFBI	-0.01127	1.36590	1.91789E-03	3.27462E-02
CHRNA2	-0.01185	1.49896	1.92439E-03	3.27884E-02
AP000346.2	-0.01292	0.97234	1.92928E-03	3.28440E-02
STRCP1	-0.01116	1.23984	1.93073E-03	3.28440E-02
MET	-0.01195	1.32736	1.94915E-03	3.30031E-02
SLC5A9	-0.01150	1.29229	1.95822E-03	3.30649E-02
CARD11	-0.01252	1.45548	1.97721E-03	3.31510E-02
PTCHD4	-0.01094	1.22972	1.98022E-03	3.31510E-02
ITGAL	-0.01194	1.41984	2.00283E-03	3.32905E-02
SHANK2	-0.01181	1.61159	2.01961E-03	3.32905E-02
HEG1	-0.01264	1.40386	2.03818E-03	3.32905E-02
PIWIL2	-0.01141	1.30796	2.03915E-03	3.32905E-02
MUC4	-0.01106	1.36897	2.03981E-03	3.32905E-02
SEZ6	-0.01230	1.36975	2.04687E-03	3.32905E-02
DSCAML1	-0.01253	1.50411	2.07916E-03	3.32905E-02
F2	-0.01577	1.65351	2.08189E-03	3.32905E-02
PPP1R1B	-0.01036	1.26045	2.10091E-03	3.32905E-02
ANKRD30A	-0.01245	1.38648	2.10346E-03	3.32905E-02
VWA3B	-0.01284	1.41714	2.10672E-03	3.32905E-02
GALNT16	-0.01073	1.24383	2.12385E-03	3.32905E-02
CLCN1	-0.01167	1.39041	2.13031E-03	3.32905E-02
THBS2	-0.01029	1.18927	2.15114E-03	3.32905E-02
LRRK2	-0.01243	1.49708	2.15303E-03	3.32905E-02
COL11A2	-0.01357	1.50954	2.15428E-03	3.32905E-02
C11orf80	-0.01516	1.44804	2.15959E-03	3.32905E-02
SAGE4P	-0.01080	1.24419	2.16245E-03	3.32905E-02
AP3B2	-0.01078	1.25524	2.17496E-03	3.32905E-02
SORCS3	-0.01127	1.36749	2.17741E-03	3.32905E-02
ERICH3	-0.01194	1.34635	2.18545E-03	3.32905E-02
DNAH9	-0.01864	2.21593	2.18582E-03	3.32905E-02

Gene symbol	logFC	AveExpr	P.Value	adj.P.Val
GRIN2B	-0.01368	1.85895	2.20309E-03	3.32905E-02
CD163L1	-0.01368	1.52680	2.20381E-03	3.32905E-02
LAMP3	-0.01023	1.09221	2.20742E-03	3.32905E-02
NAV2	-0.01667	1.91411	2.21425E-03	3.32905E-02
GABBR2	-0.01078	1.20947	2.22789E-03	3.32905E-02
PIK3R6	-0.01188	1.37300	2.23313E-03	3.32905E-02
NDST1	-0.01307	1.54989	2.25509E-03	3.32905E-02
CRYBG2	-0.01138	1.27663	2.25791E-03	3.32905E-02
CASZ1	-0.01030	1.17667	2.27501E-03	3.32905E-02
AFAP1L2	-0.01007	1.13500	2.29074E-03	3.32905E-02
ANO3	-0.01006	1.14703	2.29138E-03	3.32905E-02
PELO	-0.01212	1.45157	2.29931E-03	3.32905E-02
FKBP9	-0.01706	1.74389	2.32850E-03	3.32905E-02
ANKRD35	-0.01210	1.40208	2.33125E-03	3.32905E-02
OR2L3	-0.01096	1.23406	2.33180E-03	3.32905E-02
XPNPEP2	-0.01078	1.21604	2.33619E-03	3.32905E-02
KCNB1	-0.01062	1.37034	2.33677E-03	3.32905E-02
IGSF1	-0.01326	1.48781	2.35284E-03	3.32905E-02
PCSK2	-0.01003	1.12868	2.35647E-03	3.32905E-02
PSD3	-0.01262	1.52485	2.36916E-03	3.32905E-02
NRXN3	-0.01417	1.66616	2.38231E-03	3.32905E-02
ITIH1	-0.01295	1.51214	2.39611E-03	3.32905E-02
ATP2A3	-0.01221	1.43596	2.40520E-03	3.32905E-02
B3GALT2	-0.01129	1.00717	2.40690E-03	3.32905E-02
OR10K1	-0.01030	1.12373	2.45587E-03	3.32905E-02
ASB16	-0.01356	1.29248	2.47656E-03	3.32905E-02
MYO3A	-0.01178	1.37543	2.49479E-03	3.32905E-02
SIM1	-0.01046	1.26042	2.49706E-03	3.32905E-02
MST1L	-0.01220	1.38197	2.50848E-03	3.32905E-02
MAG	-0.01040	1.18549	2.52012E-03	3.32905E-02
KRTAP10-11	-0.01097	1.23880	2.53020E-03	3.32905E-02
ALPK2	-0.01243	1.40615	2.55125E-03	3.32905E-02
CEP126	-0.01204	1.39659	2.55329E-03	3.32905E-02
SPHKAP	-0.01386	1.65473	2.57981E-03	3.32905E-02
CHD5	-0.01152	1.43304	2.58011E-03	3.32905E-02
SIGLEC11	-0.01060	1.13868	2.59807E-03	3.32905E-02
CREB5	-0.01076	1.32675	2.61067E-03	3.32905E-02
PCK1	-0.01068	1.16916	2.62779E-03	3.32905E-02
KIRREL1	-0.01332	1.55368	2.64392E-03	3.32905E-02
ABCC11	-0.01345	1.62551	2.65363E-03	3.32905E-02
PTPRN	-0.01214	1.42329	2.65441E-03	3.32905E-02
CGN	-0.01284	1.45355	2.65917E-03	3.32905E-02
ZNF624	-0.01237	1.37345	2.68358E-03	3.32905E-02
DOCK3	-0.01519	1.78471	2.70342E-03	3.32905E-02
OR1F1	-0.01028	1.20180	2.70874E-03	3.32905E-02
CFAP52	-0.01002	1.23081	2.71183E-03	3.32905E-02
COL14A1	-0.01284	1.57431	2.71548E-03	3.32905E-02
TRIM50	-0.01125	1.27412	2.73210E-03	3.32905E-02
EPHA8	-0.01135	1.24714	2.74602E-03	3.32905E-02
OR2L5	-0.01090	1.20369	2.75091E-03	3.32905E-02
PDZRN3	-0.01004	1.14500	2.75988E-03	3.32905E-02
KSR1	-0.01399	1.68076	2.76326E-03	3.32905E-02
ZC3H12B	-0.01033	1.13366	2.77678E-03	3.32905E-02
OR2A7	-0.01144	1.26396	2.79745E-03	3.32905E-02
PRSS16	-0.01122	1.24331	2.79962E-03	3.32905E-02
AXDND1	-0.01047	1.18753	2.80229E-03	3.32905E-02
SCN1A	-0.01238	1.46288	2.80970E-03	3.32905E-02
MUC17	-0.02203	2.97216	2.82976E-03	3.32905E-02

Gene symbol	logFC	AveExpr	P.Value	adj.P.Val
DENND2A	-0.01171	1.29776	2.83290E-03	3.32905E-02
ITGB5	-0.01126	1.36984	2.83589E-03	3.32905E-02
PCDHGA2	-0.01118	1.23084	2.83640E-03	3.32905E-02
ADCYAP1R1	-0.01008	1.23671	2.84208E-03	3.32905E-02
AL353807.3	-0.01280	1.38053	2.85212E-03	3.32905E-02
NSUN5P1	-0.02247	3.06522	2.86598E-03	3.32905E-02
ARHGAP40	-0.01128	1.25222	2.87176E-03	3.32905E-02
COL13A1	-0.01045	1.20212	2.87259E-03	3.32905E-02
KIF5C	-0.01127	1.43677	2.87616E-03	3.32905E-02
NFATC4	-0.01267	1.48806	2.87799E-03	3.32905E-02
DNAH6	-0.01503	1.87118	2.88262E-03	3.32905E-02
CYP4B1	-0.01080	1.20118	2.88841E-03	3.32905E-02
SRL	-0.01245	1.46918	2.90641E-03	3.32905E-02
CD163	-0.01019	1.13578	2.92415E-03	3.32905E-02
ADAMTS2	-0.01034	1.25623	2.92856E-03	3.32905E-02
TNNC2	-0.01231	1.00425	2.95129E-03	3.32905E-02
TP63	-0.01112	1.23963	2.94665E-03	3.32905E-02
CTNND2	-0.01108	1.37990	2.95039E-03	3.32905E-02
GRIP2	-0.01203	1.38952	2.95997E-03	3.32905E-02
ZNF222	-0.01155	1.20670	2.97105E-03	3.32905E-02
FBXO40	-0.01010	1.20578	2.97408E-03	3.32905E-02
SPOCD1	-0.01004	1.15789	2.97652E-03	3.32905E-02
ADAM22	-0.01136	1.26253	2.98396E-03	3.32905E-02
MUC12	-0.01120	1.28566	2.98504E-03	3.32905E-02
CHST4	-0.01126	1.25096	2.98752E-03	3.32905E-02
CORIN	-0.01157	1.38694	3.01546E-03	3.32905E-02
EGF	-0.01147	1.27553	3.01985E-03	3.32905E-02
PXDNL	-0.01032	1.20060	3.02116E-03	3.32905E-02
HHIPL2	-0.01055	1.16776	3.02207E-03	3.32905E-02
PCDHAC2	-0.01113	1.24232	3.02467E-03	3.32905E-02
ELAVL4	-0.01021	1.12841	3.03609E-03	3.32905E-02
NRK	-0.01234	1.44567	3.08173E-03	3.32905E-02
CHI3L1	-0.01060	1.18832	3.08183E-03	3.32905E-02
FANK1	-0.01256	1.14799	3.09405E-03	3.32905E-02
SCN3A	-0.01144	1.36150	3.11107E-03	3.32905E-02
KRT6A	-0.01110	1.24137	3.12590E-03	3.32905E-02
DUOX2	-0.01089	1.35932	3.13080E-03	3.32905E-02
LHX9	-0.01157	1.38632	3.13379E-03	3.32905E-02
MYH1	-0.01388	1.64898	3.13854E-03	3.32905E-02
CPNE2	-0.01306	1.55995	3.15032E-03	3.32905E-02
ACOXL	-0.01108	1.29341	3.15200E-03	3.32905E-02
PLCH1	-0.01261	1.62255	3.15723E-03	3.32905E-02
EPHB2	-0.01161	1.35510	3.17185E-03	3.32905E-02
ADGRF5P1	-0.01050	1.16703	3.18447E-03	3.32905E-02
MTMR11	-0.01004	1.20519	3.19358E-03	3.32905E-02
PZP	-0.01253	1.40484	3.20212E-03	3.32905E-02
ZNF34	-0.01059	1.18449	3.21152E-03	3.32905E-02
ARMH4	-0.01013	1.13891	3.22714E-03	3.32905E-02
ITGAD	-0.01016	1.20948	3.23535E-03	3.32905E-02
UNC80	-0.01558	1.88898	3.23674E-03	3.32905E-02
GAREM1	-0.01120	1.51016	3.24853E-03	3.32905E-02
KCNB2	-0.01034	1.16130	3.24887E-03	3.32905E-02
COL22A1	-0.01455	1.69108	3.25163E-03	3.32905E-02
SLIT2	-0.01189	1.46655	3.25612E-03	3.32905E-02
ALOX15	-0.01178	1.35342	3.26644E-03	3.32905E-02
DNM1	-0.01003	1.26431	3.28374E-03	3.32905E-02
KRT6B	-0.01067	1.19539	3.28505E-03	3.32905E-02
PCSK5	-0.01273	1.59336	3.29181E-03	3.32905E-02

Gene symbol	logFC	AveExpr	P.Value	adj.P.Val
AOAH	-0.01091	1.30864	3.29413E-03	3.32905E-02
SV2C	-0.01003	1.27673	3.29657E-03	3.32905E-02
ADAMTS9	-0.01412	1.64279	3.30452E-03	3.32905E-02
ABCC12	-0.01228	1.54981	3.31211E-03	3.32905E-02
STXBP5L	-0.01172	1.46583	3.31558E-03	3.32905E-02
FRAS1	-0.01708	2.15780	3.33183E-03	3.32905E-02
KIR3DL3	-0.01113	1.25081	3.34425E-03	3.32905E-02
ADAMTS18	-0.01186	1.41386	3.34818E-03	3.32905E-02
ADGRL3	-0.01230	1.38712	3.34980E-03	3.32905E-02
KRT76	-0.01051	1.17888	3.37204E-03	3.32905E-02
FLG2	-0.01932	2.30655	3.37350E-03	3.32905E-02
WFS1	-0.01174	1.37813	3.38001E-03	3.32905E-02
HUNK	-0.01087	1.22735	3.39424E-03	3.32905E-02
ARHGEF40	-0.01193	1.56531	3.39436E-03	3.32905E-02
PADI3	-0.01153	1.29774	3.39490E-03	3.32905E-02
SEZ6L	-0.01128	1.36718	3.39653E-03	3.32905E-02
FAM135B	-0.01224	1.54439	3.39901E-03	3.32905E-02
SCN4A	-0.01045	1.25237	3.40225E-03	3.32905E-02
COL24A1	-0.01245	1.44101	3.41530E-03	3.32905E-02
COL15A1	-0.01322	1.49385	3.42198E-03	3.32905E-02
ABCB5	-0.01060	1.26453	3.42339E-03	3.32905E-02
OR2T4	-0.01317	1.48175	3.42397E-03	3.32905E-02
DCST2	-0.01136	1.28503	3.43170E-03	3.32905E-02
NLRP12	-0.01121	1.36180	3.44333E-03	3.32905E-02
PXDN	-0.01183	1.33780	3.44486E-03	3.32905E-02
SEMA5B	-0.01029	1.19781	3.44948E-03	3.32905E-02
KDR	-0.01124	1.35134	3.45372E-03	3.32905E-02
MYH7	-0.01296	1.46525	3.46331E-03	3.32905E-02
EPB41L5	-0.01003	1.13770	3.46730E-03	3.32905E-02
PSMB9	-0.01323	1.09555	3.48664E-03	3.32905E-02
MUC16	-0.02194	2.89090	3.47187E-03	3.32905E-02
RPTN	-0.01342	1.51420	3.48501E-03	3.32905E-02
MYO1A	-0.01188	1.38800	3.50464E-03	3.32905E-02
CACNA2D3	-0.01259	1.42451	3.51111E-03	3.32905E-02
TGM7	-0.01159	1.35658	3.51450E-03	3.32905E-02
SYTL2	-0.01202	1.48324	3.51883E-03	3.32905E-02
TFCP2L1	-0.01034	1.16950	3.51974E-03	3.32905E-02
PLEKHA7	-0.01317	1.71040	3.52121E-03	3.32905E-02
ATP4A	-0.01050	1.20491	3.52649E-03	3.32905E-02
PIEZO2	-0.01431	1.86204	3.53753E-03	3.32905E-02
UNC13A	-0.01358	1.64138	3.54711E-03	3.32905E-02
ENAH	-0.01082	1.42210	3.55196E-03	3.32905E-02
SUPT20HL1	-0.01075	1.22209	3.56069E-03	3.32905E-02
NLRP9	-0.01095	1.28475	3.56253E-03	3.32905E-02
CYP4A22	-0.01163	1.39718	3.56442E-03	3.32905E-02
ZNF878	-0.01209	1.24122	3.57079E-03	3.32905E-02
FAM205A	-0.01095	1.24145	3.57125E-03	3.32905E-02
PPFIA4	-0.01503	1.83698	3.58275E-03	3.32905E-02
CCDC60	-0.01021	1.20117	3.58665E-03	3.32905E-02
F8	-0.01343	1.61534	3.59294E-03	3.32905E-02
MGAM2	-0.01526	1.90478	3.60050E-03	3.32905E-02
ROR1	-0.01030	1.21095	3.60511E-03	3.32905E-02
TNS4	-0.01032	1.19862	3.60940E-03	3.32905E-02
RAG1	-0.01078	1.22700	3.63551E-03	3.32905E-02
CDH16	-0.01140	1.29650	3.64027E-03	3.32905E-02
BRINP2	-0.01272	1.48132	3.65106E-03	3.32905E-02
CNGA3	-0.01057	1.19774	3.65522E-03	3.32905E-02
RFPL2	-0.01063	1.20273	3.66036E-03	3.32905E-02

Gene symbol	logFC	AveExpr	P.Value	adj.P.Val
LRRC15	-0.01003	1.13750	3.70429E-03	3.32905E-02
MYBPC2	-0.01243	1.41653	3.70677E-03	3.32905E-02
WDR64	-0.01207	1.50156	3.71057E-03	3.32905E-02
HSD17B7	-0.01356	1.32833	3.71253E-03	3.32905E-02
DUSP27	-0.01110	1.26525	3.72109E-03	3.32905E-02
HEPHL1	-0.01054	1.19263	3.72654E-03	3.32905E-02
KRT37	-0.01010	1.15169	3.72704E-03	3.32905E-02
RYR1	-0.01720	2.20834	3.72719E-03	3.32905E-02
SLC9A5	-0.01133	1.28371	3.72933E-03	3.32905E-02
MYO5B	-0.01447	1.64960	3.73006E-03	3.32905E-02
LRIG3	-0.01004	1.14340	3.74312E-03	3.32905E-02
FCRL4	-0.01125	1.31935	3.74475E-03	3.32905E-02
TGM1	-0.01060	1.24936	3.74856E-03	3.32905E-02
KCP	-0.01064	1.36339	3.77858E-03	3.32905E-02
SPATA31D1	-0.01241	1.41606	3.78980E-03	3.32905E-02
SLC26A9	-0.01260	1.48285	3.79044E-03	3.32905E-02
PLPPR2	-0.01076	1.30138	3.79769E-03	3.32905E-02
KIAA1549L	-0.01433	1.67746	3.81301E-03	3.32905E-02
GPLD1	-0.01688	2.30358	3.82337E-03	3.32905E-02
MYO1H	-0.01028	1.16488	3.84265E-03	3.32905E-02
ABCC9	-0.01167	1.33740	3.84585E-03	3.32905E-02
RGL1	-0.01041	1.26559	3.84838E-03	3.32905E-02
HEATR4	-0.01122	1.28206	3.86823E-03	3.32905E-02
IPPK	-0.01416	1.42100	3.87792E-03	3.32905E-02
FMNL2	-0.01026	1.31970	3.88108E-03	3.32905E-02
MAP3K15	-0.01021	1.25601	3.88683E-03	3.32905E-02
HIP1	-0.01256	1.61627	3.88880E-03	3.32905E-02
GPRASP1	-0.01089	1.29221	3.91172E-03	3.32905E-02
PRDM9	-0.01155	1.32101	3.92115E-03	3.32905E-02
VIT	-0.01000	1.18846	3.92294E-03	3.32905E-02
FCRL3	-0.01273	1.45503	3.93255E-03	3.32905E-02
TAP1	-0.01328	1.39213	3.93642E-03	3.32905E-02
ZNF559	-0.01165	1.16188	3.94701E-03	3.32905E-02
ACSM5	-0.01147	1.35058	3.94761E-03	3.32905E-02
RTL9	-0.01394	1.63760	3.95776E-03	3.32905E-02
AC139795.1	-0.01118	1.29938	3.97371E-03	3.32905E-02
TG	-0.01595	1.92990	3.99898E-03	3.32905E-02
MST1R	-0.01025	1.17295	4.00198E-03	3.32905E-02
GLI2	-0.01152	1.42925	4.00808E-03	3.32905E-02
TTLL6	-0.01076	1.23364	4.01106E-03	3.32905E-02
ABCA8	-0.01162	1.33697	4.01131E-03	3.32905E-02
EGFR	-0.01275	1.54334	4.01324E-03	3.32905E-02
AC126603.1	-0.01459	1.91652	4.04016E-03	3.32905E-02
ANO2	-0.01222	1.50902	4.04431E-03	3.32905E-02
PARD3B	-0.01211	1.57508	4.04775E-03	3.32905E-02
ZNF423	-0.01207	1.42477	4.04966E-03	3.32905E-02
PNPLA1	-0.01001	1.14905	4.05246E-03	3.32905E-02
COL27A1	-0.01189	1.68717	4.05435E-03	3.32905E-02
PCDHA1	-0.01104	1.30101	4.05550E-03	3.32905E-02
SCN8A	-0.01459	1.77008	4.05843E-03	3.32905E-02
TRPM2	-0.01083	1.28939	4.05847E-03	3.32905E-02
OR4M1	-0.01052	1.20759	4.08546E-03	3.32905E-02
GRM2	-0.01030	1.21458	4.08645E-03	3.32905E-02
ABLIM2	-0.01143	1.39625	4.08872E-03	3.32905E-02
TIAM1	-0.01425	1.67275	4.08963E-03	3.32905E-02
NINL	-0.01019	1.21521	4.09032E-03	3.32905E-02
SIRPB1	-0.01163	1.37216	4.09766E-03	3.32905E-02
PRKAG2	-0.01549	2.14555	4.10176E-03	3.32905E-02

Gene symbol	logFC	AveExpr	P.Value	adj.P.Val
KRT5	-0.01035	1.16163	4.11549E-03	3.32905E-02
SYT17	-0.01048	1.27954	4.11700E-03	3.32905E-02
USHBP1	-0.01003	1.13843	4.11719E-03	3.32905E-02
OR2L6P	-0.01096	1.30629	4.11876E-03	3.32905E-02
TTYH2	-0.01004	1.15409	4.12337E-03	3.32905E-02
CMYA5	-0.01419	1.63314	4.13409E-03	3.32905E-02
COL4A1	-0.01448	1.75117	4.13987E-03	3.32905E-02
ACOX2	-0.01047	1.28491	4.14591E-03	3.32905E-02
ROS1	-0.01220	1.49281	4.14970E-03	3.32905E-02
NBPF1	-0.01364	1.91867	4.15212E-03	3.32905E-02
REEP6	-0.01433	1.34074	4.16054E-03	3.32905E-02
GAS2L2	-0.01059	1.22373	4.16307E-03	3.32905E-02
MXRA5	-0.01482	1.71171	4.16595E-03	3.32905E-02
EFR3B	-0.01124	1.34247	4.16705E-03	3.32905E-02
NLRP4	-0.01087	1.25266	4.17083E-03	3.32905E-02
TRIM42	-0.01037	1.18502	4.17407E-03	3.32905E-02
PIPOX	-0.01039	1.19429	4.18630E-03	3.32905E-02
ADCY6	-0.01291	1.82043	4.18724E-03	3.32905E-02
KCNAB1	-0.01680	2.05304	4.19590E-03	3.32905E-02
KRT85	-0.01002	1.18756	4.19752E-03	3.32905E-02
SORCS2	-0.01106	1.35131	4.20441E-03	3.32905E-02
THBS1	-0.01328	1.66311	4.20810E-03	3.32905E-02
CCT8L1P	-0.01179	1.35853	4.21046E-03	3.32905E-02
EPHA1	-0.01264	1.45948	4.21295E-03	3.32905E-02
CADPS	-0.01193	1.55035	4.21455E-03	3.32905E-02
GPR179	-0.01240	1.46638	4.21960E-03	3.32905E-02
ADGRD1	-0.01157	1.37697	4.22796E-03	3.32905E-02
TECTA	-0.01333	1.76398	4.22847E-03	3.32905E-02
CAMK1G	-0.01011	1.19189	4.24716E-03	3.32905E-02
CCDC33	-0.01137	1.31678	4.24970E-03	3.32905E-02
TOGARAM2	-0.01128	1.40118	4.25533E-03	3.32905E-02
ADGB	-0.01024	1.25633	4.27807E-03	3.32905E-02
ANK2	-0.01552	1.88986	4.28760E-03	3.32905E-02
KRT34	-0.01198	1.41860	4.29015E-03	3.32905E-02
AHNAK2	-0.01910	2.47231	4.29463E-03	3.32905E-02
FCAMR	-0.01054	1.25363	4.29646E-03	3.32905E-02
COL9A1	-0.01040	1.28548	4.29692E-03	3.32905E-02
ZDBF2	-0.01159	1.37418	4.30322E-03	3.32905E-02
PCDH10	-0.01308	1.51384	4.32246E-03	3.32905E-02
NTRK3	-0.01108	1.49278	4.32563E-03	3.32905E-02
LPA	-0.01236	1.47425	4.34322E-03	3.32905E-02
CFH	-0.01162	1.42987	4.34673E-03	3.32905E-02
AMPH	-0.01114	1.28955	4.34800E-03	3.32905E-02
SLC26A8	-0.01145	1.36111	4.35034E-03	3.32905E-02
PCDHGA7	-0.01029	1.19130	4.36030E-03	3.32905E-02
ABCC3	-0.01329	1.75724	4.36457E-03	3.32905E-02
SVIL2P	-0.01172	1.42935	4.36492E-03	3.32905E-02
GRM7	-0.01166	1.50357	4.36641E-03	3.32905E-02
NOTCH4	-0.01437	1.77529	4.37437E-03	3.32905E-02
AC139495.3	-0.01328	1.42123	4.37498E-03	3.32905E-02
COL28A1	-0.01040	1.38739	4.37512E-03	3.32905E-02
NOS2	-0.01337	1.59097	4.38328E-03	3.32905E-02
OR2T2	-0.01144	1.36385	4.38407E-03	3.32905E-02
SLC12A5	-0.01229	1.47369	4.38784E-03	3.32905E-02
LRRC52	-0.01033	1.22794	4.39058E-03	3.32905E-02
ASAP2	-0.01079	1.25184	4.39633E-03	3.32905E-02
MYH2	-0.01422	1.64877	4.41610E-03	3.32905E-02
ITIH3	-0.01024	1.22610	4.43428E-03	3.32905E-02

Gene symbol	logFC	AveExpr	P.Value	adj.P.Val
SLC7A14	-0.01014	1.21171	4.44815E-03	3.32905E-02
PSG8	-0.01035	1.27736	4.45338E-03	3.32905E-02
GCNT3	-0.01019	0.93740	4.45460E-03	3.32905E-02
PTPN14	-0.01545	1.88314	4.45959E-03	3.32905E-02
PCDHGA12	-0.01033	1.19318	4.46239E-03	3.32905E-02
PEG3	-0.01245	1.44550	4.46437E-03	3.32905E-02
CD276	-0.01014	1.26206	4.47471E-03	3.32905E-02
FBXW10	-0.01294	1.54681	4.48677E-03	3.32905E-02
XKR6	-0.01228	1.38375	4.50460E-03	3.32905E-02
RORC	-0.01252	1.49160	4.50963E-03	3.32905E-02
HEPH	-0.01107	1.30678	4.51120E-03	3.32905E-02
ASXL3	-0.01168	1.45927	4.52857E-03	3.32905E-02
FAM205BP	-0.01255	1.49727	4.55032E-03	3.32905E-02
GRM5	-0.01098	1.30332	4.56471E-03	3.32905E-02
ADGRG4	-0.01403	1.82802	4.57642E-03	3.32905E-02
SYNJ2	-0.01178	1.51986	4.62085E-03	3.32905E-02
MUC20P1	-0.01040	1.20440	4.62533E-03	3.32905E-02
DQX1	-0.01040	1.24033	4.63011E-03	3.32905E-02
MEGF11	-0.01192	1.46513	4.63598E-03	3.32905E-02
PAMR1	-0.01008	1.20836	4.64077E-03	3.32905E-02
NELL1	-0.01042	1.32382	4.64520E-03	3.32905E-02
MYO5C	-0.01145	1.45651	4.65168E-03	3.32905E-02
CPNE9	-0.01002	1.25091	4.65581E-03	3.32905E-02
ZNF836	-0.01098	1.40539	4.66264E-03	3.32905E-02
ASTN1	-0.01459	1.79012	4.66955E-03	3.32905E-02
BCAN	-0.01033	1.19815	4.67776E-03	3.32905E-02
RYR2	-0.01778	2.37231	4.69032E-03	3.32905E-02
ZNF334	-0.01052	1.40991	4.69150E-03	3.32905E-02
MGAM	-0.01642	2.19421	4.69426E-03	3.32905E-02
SLC6A20	-0.01078	1.33646	4.70374E-03	3.32905E-02
GRM1	-0.01056	1.31463	4.71322E-03	3.32905E-02
ATP1A2	-0.01451	1.77641	4.72803E-03	3.32905E-02
GUCY2C	-0.01049	1.22310	4.73441E-03	3.32905E-02
NLRP8	-0.01103	1.33083	4.76769E-03	3.32905E-02
RASAL1	-0.01127	1.35154	4.77429E-03	3.32905E-02
NES	-0.01131	1.32375	4.78078E-03	3.32937E-02
ADCY8	-0.01056	1.27158	4.78767E-03	3.33226E-02
CYP4F3	-0.01036	1.19065	4.78989E-03	3.33247E-02
LIMCH1	-0.01162	1.67783	4.79485E-03	3.33274E-02
ENPP2	-0.01068	1.28084	4.80947E-03	3.33303E-02
TCHHL1	-0.01148	1.33957	4.81917E-03	3.33403E-02
REN	-0.01183	1.41543	4.82026E-03	3.33403E-02
BMPR2	-0.01663	2.38702	4.82915E-03	3.33403E-02
MYO16	-0.01058	1.27545	4.84145E-03	3.33403E-02
ABCC8	-0.01427	1.80597	4.85737E-03	3.33403E-02
NLRP5	-0.01133	1.32239	4.86795E-03	3.33403E-02
CEACAM5	-0.01039	1.25522	4.87128E-03	3.33403E-02
SLC28A1	-0.01049	1.22919	4.87623E-03	3.33403E-02
SUPT20HL2	-0.01063	1.24069	4.90127E-03	3.33403E-02
COL7A1	-0.01583	2.14087	4.90743E-03	3.33403E-02
LYZL2	-0.01157	1.34764	4.90764E-03	3.33403E-02
PPFIA2	-0.01056	1.30709	4.91418E-03	3.33403E-02
TGM6	-0.01063	1.28383	4.95232E-03	3.33403E-02
LDLRAD4-AS1	-0.01674	1.96733	4.97373E-03	3.33403E-02
WNK2	-0.01081	1.30497	4.98300E-03	3.33403E-02
KCNH4	-0.01077	1.26227	5.00423E-03	3.33403E-02
THSD1	-0.01036	1.21608	5.00851E-03	3.33403E-02
SH3PXD2B	-0.01054	1.31679	5.01261E-03	3.33403E-02

Gene symbol	logFC	AveExpr	P.Value	adj.P.Val
PKHD1	-0.01581	2.06480	5.03252E-03	3.33403E-02
WNK3	-0.01060	1.29198	5.03319E-03	3.33403E-02
ABCA1	-0.01283	1.73642	5.04291E-03	3.33403E-02
KCNU1	-0.01106	1.33515	5.04668E-03	3.33403E-02
UMODL1	-0.01228	1.44071	5.04721E-03	3.33403E-02
MERTK	-0.01153	1.40555	5.05641E-03	3.33403E-02
MMP2	-0.01080	1.30408	5.09458E-03	3.33403E-02
KIR2DL4	-0.01022	1.19963	5.10772E-03	3.33403E-02
SCNN1D	-0.02131	2.36968	5.10932E-03	3.33403E-02
CACHD1	-0.01060	1.31348	5.11780E-03	3.33403E-02
SALL1	-0.01196	1.40132	5.12031E-03	3.33403E-02
NOD2	-0.01130	1.40396	5.13155E-03	3.33403E-02
CES5A	-0.01010	1.18447	5.14731E-03	3.33403E-02
GALNT14	-0.01011	1.26576	5.15981E-03	3.33403E-02
UNC13C	-0.01212	1.61396	5.16164E-03	3.33403E-02
SCNN1G	-0.01017	1.31509	5.17264E-03	3.33403E-02
NYNRIN	-0.01235	1.53001	5.19329E-03	3.33403E-02
ENDOV	-0.01081	1.42947	5.19500E-03	3.33403E-02
COL3A1	-0.01092	1.39023	5.20695E-03	3.33403E-02
CEMIP	-0.01326	1.60089	5.21206E-03	3.33403E-02
GLI3	-0.01324	1.76929	5.22009E-03	3.33403E-02
CNGB1	-0.01020	1.27544	5.22973E-03	3.33403E-02
THSD7A	-0.01130	1.36790	5.23039E-03	3.33403E-02
ABCA13	-0.01551	1.94823	5.23753E-03	3.33403E-02
LPO	-0.01116	1.45468	5.23885E-03	3.33403E-02
PADI2	-0.01090	1.31974	5.24144E-03	3.33403E-02
CEACAM3	-0.01063	1.24205	5.24437E-03	3.33403E-02
NRAP	-0.01143	1.45863	5.24569E-03	3.33403E-02
SIGLEC1	-0.01175	1.48119	5.27895E-03	3.33403E-02
DSCAM	-0.01400	1.86667	5.29120E-03	3.33403E-02
PLA2G4D	-0.01005	1.27459	5.29657E-03	3.33403E-02
ACTN2	-0.01228	1.48329	5.29776E-03	3.33403E-02
MAGI3	-0.01235	1.56234	5.30316E-03	3.33403E-02
CLSTN2	-0.01101	1.33640	5.30960E-03	3.33403E-02
ZNF132	-0.01486	1.76146	5.35376E-03	3.33423E-02
DUOX1	-0.01463	1.92598	5.35393E-03	3.33423E-02
CAPN11	-0.01052	1.34297	5.38459E-03	3.33877E-02
NMNAT2	-0.01067	1.34803	5.40042E-03	3.33877E-02
CC2D2A	-0.01325	1.90474	5.40143E-03	3.33877E-02
ACCSL	-0.01006	1.25197	5.41929E-03	3.33910E-02
TRPM1	-0.01205	1.50096	5.43236E-03	3.34268E-02
EFCAB8	-0.01208	1.53305	5.45980E-03	3.34713E-02
TNFRSF8	-0.01032	1.26073	5.46064E-03	3.34713E-02
AC126323.1	-0.01020	1.30195	5.47790E-03	3.34713E-02
THBS4	-0.01101	1.34107	5.50019E-03	3.34713E-02
DYNC2H1	-0.01185	1.56796	5.50165E-03	3.34713E-02
PTPRU	-0.01416	1.82489	5.56816E-03	3.35565E-02
CECR2	-0.01227	1.59859	5.61622E-03	3.35804E-02
KIF26B	-0.01116	1.32197	5.62054E-03	3.35835E-02
ADGRB2	-0.01016	1.28064	5.62480E-03	3.35835E-02
ANPEP	-0.01095	1.42741	5.62609E-03	3.35835E-02
CTTNBP2	-0.01089	1.42890	5.65206E-03	3.36837E-02
ZFHX4	-0.01459	1.79554	5.67332E-03	3.37359E-02
FAT2	-0.01627	2.29289	5.69171E-03	3.37532E-02
WDR87	-0.01253	1.57228	5.70095E-03	3.37532E-02
JRKL	-0.01503	1.73531	5.71401E-03	3.37532E-02
NOS1	-0.01301	1.72881	5.73189E-03	3.37532E-02
B4GALNT3	-0.01031	1.30177	5.73681E-03	3.37532E-02

Gene symbol	logFC	AveExpr	P.Value	adj.P.Val
GRIA1	-0.01055	1.30472	5.73954E-03	3.37532E-02
PCNX2	-0.01370	1.73127	5.77656E-03	3.38047E-02
CACNA1G	-0.01410	1.87749	5.78010E-03	3.38047E-02
ARAP3	-0.01260	1.59992	5.78042E-03	3.38047E-02
SEMA4F	-0.01062	1.42294	5.81391E-03	3.38399E-02
GPER1	-0.01029	1.17981	5.82673E-03	3.38399E-02
CRYBG3	-0.01176	1.61368	5.83005E-03	3.38399E-02
GTF2IRD1	-0.01197	1.61434	5.85381E-03	3.38399E-02
SYT6	-0.01069	1.35163	5.85389E-03	3.38399E-02
USH2A	-0.01732	2.50242	5.85435E-03	3.38399E-02
BNIPL	-0.01078	1.30946	5.85882E-03	3.38498E-02
STAB2	-0.01410	1.87617	5.86466E-03	3.38501E-02
NRP2	-0.01159	1.41971	5.87736E-03	3.38521E-02
ZNF366	-0.01129	1.39169	5.90099E-03	3.38678E-02
CLCNKB	-0.01147	1.36113	5.91110E-03	3.38678E-02
FAT3	-0.01558	2.09792	5.94273E-03	3.38678E-02
TENM1	-0.01397	1.91638	5.94283E-03	3.38678E-02
MROH2B	-0.01026	1.26835	5.96366E-03	3.38683E-02
PLXNA2	-0.01535	2.12573	5.99852E-03	3.38771E-02
TMEM236	-0.01000	1.21899	6.00793E-03	3.38797E-02
COL5A2	-0.01149	1.50071	6.02517E-03	3.39043E-02
PKD1L2	-0.01462	1.87818	6.05308E-03	3.39523E-02
EYA2	-0.01007	1.29026	6.05674E-03	3.39523E-02
KIF21B	-0.01476	2.07219	6.05873E-03	3.39523E-02
COL16A1	-0.01385	1.95964	6.06892E-03	3.39622E-02
THSD4	-0.01124	1.68174	6.07293E-03	3.39622E-02
COL12A1	-0.01313	1.70066	6.08249E-03	3.39622E-02
CCT8L2	-0.01081	1.29390	6.11383E-03	3.39901E-02
SLC13A2	-0.01129	1.40443	6.15920E-03	3.40212E-02
SIPA1L2	-0.01470	1.93996	6.17303E-03	3.40212E-02
ADORA1	-0.01026	1.26764	6.18684E-03	3.40212E-02
F5	-0.01588	2.07269	6.28661E-03	3.41303E-02
C3	-0.01091	1.52315	6.29386E-03	3.41303E-02
ITGAM	-0.01294	1.59624	6.32151E-03	3.41303E-02
TLL1	-0.01025	1.35424	6.32955E-03	3.41530E-02
KRT33B	-0.01027	1.23296	6.37415E-03	3.42498E-02
FBN3	-0.01354	1.73409	6.38995E-03	3.42895E-02
GREB1	-0.01330	1.80552	6.39225E-03	3.42942E-02
CRIM1	-0.01577	2.05418	6.40986E-03	3.42942E-02
CACNA1S	-0.01481	1.96124	6.43712E-03	3.42942E-02
DMD	-0.01454	2.05542	6.45847E-03	3.42942E-02
UNC5D	-0.01052	1.34099	6.46374E-03	3.42942E-02
SYNPO2	-0.01319	1.88494	6.46387E-03	3.42942E-02
CFAP45	-0.01060	1.31610	6.47355E-03	3.43126E-02
ITPRID1	-0.01208	1.53156	6.49308E-03	3.43430E-02
MYBPC3	-0.01179	1.54684	6.49907E-03	3.43430E-02
KIFC3	-0.01162	1.38999	6.50213E-03	3.43454E-02
TKTL1	-0.01160	1.38428	6.50675E-03	3.43454E-02
TCHH	-0.01011	1.31146	6.54106E-03	3.43454E-02
PAPPA2	-0.01490	1.94893	6.55741E-03	3.43501E-02
PLEKHH1	-0.01381	1.82218	6.57859E-03	3.43912E-02
TENM3	-0.01113	1.42097	6.58562E-03	3.43959E-02
DYSF	-0.01456	1.99740	6.63910E-03	3.44585E-02
MAP1B	-0.01276	1.88738	6.67327E-03	3.45702E-02
XDH	-0.01277	1.73051	6.68015E-03	3.45702E-02
EFCAB6	-0.01078	1.39479	6.71537E-03	3.46445E-02
L1CAM	-0.01128	1.46547	6.72666E-03	3.46469E-02
RANBP17	-0.01255	1.80786	6.73043E-03	3.46469E-02

Gene symbol	logFC	AveExpr	P.Value	adj.P.Val
ABCC6	-0.01371	1.83892	6.74857E-03	3.46469E-02
FRMD4B	-0.01086	1.34421	6.77888E-03	3.46941E-02
C1orf56	-0.01205	1.44654	6.78075E-03	3.46941E-02
BTBD18	-0.01380	1.25568	6.82978E-03	3.47843E-02
KIR2DL1	-0.01053	1.31865	6.88606E-03	3.48962E-02
CRB1	-0.01008	1.37514	6.90499E-03	3.49337E-02
CPAMD8	-0.01118	1.49985	6.90945E-03	3.49337E-02
CUX2	-0.01092	1.59272	6.91833E-03	3.49337E-02
MECOM	-0.01089	1.42487	6.97806E-03	3.49838E-02
CDC42BPB	-0.01283	1.89001	6.98861E-03	3.49946E-02
AC004687.1	-0.02387	3.86095	7.00449E-03	3.50382E-02
CAPN9	-0.01106	1.41319	7.02863E-03	3.50988E-02
TDRD6	-0.01126	1.52746	7.09398E-03	3.52379E-02
IGFN1	-0.01537	2.14124	7.17032E-03	3.54465E-02
CRISPLD2	-0.01003	1.39382	7.17464E-03	3.54465E-02
FAM186A	-0.01046	1.35123	7.18542E-03	3.54465E-02
VCAN	-0.01266	1.83997	7.22528E-03	3.55241E-02
CSMD3	-0.01335	1.76981	7.27513E-03	3.56185E-02
ADGRL2	-0.01060	1.52820	7.29465E-03	3.56712E-02
FAAP100	-0.01454	1.88250	7.32093E-03	3.57210E-02
HDAC9	-0.01192	1.73799	7.35889E-03	3.57927E-02
NLRP3	-0.01205	1.50954	7.39951E-03	3.58607E-02
HDAC5	-0.01149	1.58383	7.40186E-03	3.58650E-02
CYP2A13	-0.01077	1.35776	7.41546E-03	3.58971E-02
MYO3B	-0.01192	1.64823	7.41720E-03	3.58971E-02
BOC	-0.01168	1.52609	7.42551E-03	3.59027E-02
CFAP61	-0.01087	1.60811	7.55106E-03	3.61149E-02
ITGA11	-0.01257	1.78431	7.58111E-03	3.61410E-02
SAGE1	-0.01106	1.39324	7.59615E-03	3.61655E-02
DSP	-0.01240	1.81033	7.62996E-03	3.62484E-02
OTOG	-0.01339	1.76951	7.67693E-03	3.63397E-02
LMOD2	-0.01288	1.24062	7.69319E-03	3.63646E-02
COL5A3	-0.01211	1.56347	7.72826E-03	3.64483E-02
SLC22A14	-0.01000	1.26664	7.77062E-03	3.65926E-02
PDGFA	-0.01059	1.63192	7.79544E-03	3.66378E-02
LTF	-0.01090	1.42652	7.80043E-03	3.66466E-02
CPNE6	-0.01098	1.44509	7.89945E-03	3.68974E-02
LBP	-0.01050	1.36516	7.98656E-03	3.70718E-02
COL11A1	-0.01125	1.45937	8.01066E-03	3.70855E-02
SCN5A	-0.01276	1.69810	8.01806E-03	3.71098E-02
IKBKE	-0.01245	1.48696	8.03439E-03	3.71547E-02
SSPO	-0.01296	1.82782	8.05521E-03	3.72275E-02
FAM160A1	-0.01043	1.50720	8.07621E-03	3.72406E-02
THSD7B	-0.01090	1.61099	8.12490E-03	3.73484E-02
SLC4A9	-0.01055	1.36016	8.15229E-03	3.73669E-02
CCDC136	-0.01115	1.45764	8.20112E-03	3.74673E-02
MMP14	-0.01064	1.30757	8.21332E-03	3.74975E-02
MOV10L1	-0.01239	1.69889	8.22975E-03	3.75093E-02
SCN2A	-0.01172	1.66536	8.23332E-03	3.75186E-02
SLIT3	-0.01322	1.96757	8.27314E-03	3.76368E-02
KIAA1210	-0.01122	1.42461	8.32083E-03	3.77329E-02
CFAP74	-0.01154	1.54696	8.32370E-03	3.77333E-02
ITGA10	-0.01304	1.76582	8.33416E-03	3.77385E-02
ADAMTS17	-0.01170	1.55051	8.34760E-03	3.77385E-02
MRC1	-0.01196	1.67598	8.36579E-03	3.77698E-02
ATP2A1	-0.01208	1.44275	8.43164E-03	3.78693E-02
INSRR	-0.01109	1.56093	8.43620E-03	3.78693E-02
MORN1	-0.01648	2.43354	8.48710E-03	3.79904E-02

Gene symbol	logFC	AveExpr	P.Value	adj.P.Val
SPTBN5	-0.01255	1.69531	8.50925E-03	3.80320E-02
STARD13	-0.01095	1.43265	8.51577E-03	3.80320E-02
PGBD4	-0.01774	2.11701	8.52161E-03	3.80512E-02
LRRC37A4P	-0.01231	1.78189	8.52660E-03	3.80595E-02
FRY	-0.01447	1.96729	8.56415E-03	3.81199E-02
PTPRF	-0.01089	1.72305	8.56920E-03	3.81242E-02
MYO10	-0.01293	1.70612	8.57434E-03	3.81355E-02
NPHS1	-0.01017	1.42210	8.58055E-03	3.81413E-02
SPINK5	-0.01139	1.66315	8.68186E-03	3.82905E-02
ATP8A2	-0.01187	1.63743	8.72213E-03	3.83796E-02
PTCH2	-0.01040	1.31330	8.75195E-03	3.84426E-02
DPP6	-0.01059	1.39159	8.77265E-03	3.84620E-02
ADAMTS16	-0.01181	1.59106	8.77952E-03	3.84620E-02
SVEP1	-0.01321	1.89273	8.79176E-03	3.84620E-02
COBL	-0.01228	1.71392	8.79789E-03	3.84620E-02
ZNF208	-0.01146	1.55391	8.82551E-03	3.84792E-02
KIAA1755	-0.01108	1.48940	8.88257E-03	3.86219E-02
C1orf167	-0.01008	1.33129	8.98928E-03	3.88533E-02
MYOM3	-0.01159	1.59561	9.00374E-03	3.88769E-02
SETBP1	-0.01104	1.57603	9.00429E-03	3.88769E-02
ARHGEF10L	-0.01282	1.80843	9.06446E-03	3.90199E-02
CNTNAP2	-0.01166	1.68804	9.07256E-03	3.90333E-02
MRVI1	-0.01005	1.36520	9.08090E-03	3.90333E-02
ITIH6	-0.01034	1.34962	9.10783E-03	3.90889E-02
PDE1C	-0.01018	1.49145	9.11739E-03	3.91123E-02
SCN10A	-0.01360	1.92255	9.12270E-03	3.91123E-02
JCAD	-0.01229	1.71015	9.15215E-03	3.91640E-02
KALRN	-0.01524	2.33216	9.22298E-03	3.92342E-02
PTPRT	-0.01210	1.79815	9.22841E-03	3.92342E-02
EPHA2	-0.01043	1.37822	9.22881E-03	3.92342E-02
MUC19	-0.01704	2.76919	9.23553E-03	3.92342E-02
SCAPER	-0.01208	1.68195	9.25321E-03	3.92982E-02
IRS4	-0.01117	1.50242	9.29348E-03	3.93871E-02
FER1L6	-0.01214	1.67999	9.30232E-03	3.93905E-02
LTBP2	-0.01108	1.56870	9.33722E-03	3.94698E-02
SLC6A1	-0.01073	1.61970	9.46384E-03	3.96277E-02
TRPM4	-0.01101	1.48806	9.48046E-03	3.96565E-02
CCDC168	-0.01477	2.37850	9.50748E-03	3.97015E-02
BNC2	-0.01038	1.46538	9.51933E-03	3.97244E-02
FSTL4	-0.01036	1.40534	9.52864E-03	3.97349E-02
PRKCB	-0.01249	1.70536	9.57402E-03	3.98132E-02
LAMB3	-0.01352	1.86332	9.64303E-03	3.99600E-02
ADAMTS12	-0.01137	1.64630	9.70010E-03	4.00138E-02
TIE1	-0.01331	1.59737	9.73428E-03	4.00864E-02
TNN	-0.01355	1.78340	9.76555E-03	4.01760E-02
CACNB1	-0.01041	1.40031	9.84522E-03	4.03344E-02
NWD1	-0.01032	1.42780	9.99081E-03	4.06594E-02
CHRND	-0.01099	1.24761	1.00417E-02	4.07679E-02
SULF2	-0.01316	1.81698	1.00818E-02	4.09035E-02
DNAH1	-0.01467	2.01976	1.01182E-02	4.09588E-02
ATP8B4	-0.01042	1.42767	1.01305E-02	4.09853E-02
RHPN2	-0.01054	1.39579	1.01759E-02	4.10932E-02
TDRD9	-0.01306	1.67178	1.01822E-02	4.11007E-02
PPP2R3A	-0.01077	1.43461	1.03152E-02	4.13217E-02
ALK	-0.01288	1.87913	1.03979E-02	4.14489E-02
GSN	-0.01129	1.54302	1.04083E-02	4.14756E-02
PRKCE	-0.01044	1.62132	1.04419E-02	4.15429E-02
RASIP1	-0.01013	1.34803	1.04587E-02	4.15736E-02

Gene symbol	logFC	AveExpr	P.Value	adj.P.Val
RNF215	-0.01126	1.31872	1.04615E-02	4.15736E-02
SERINC2	-0.01948	2.78246	1.04868E-02	4.16401E-02
AGBL2	-0.01023	1.26852	1.05095E-02	4.16635E-02
ADAMTS10	-0.01119	1.34316	1.05462E-02	4.17265E-02
OBSL1	-0.01221	1.73762	1.06077E-02	4.18450E-02
ADGRV1	-0.01434	2.41981	1.06509E-02	4.18965E-02
ABCB11	-0.01108	1.51551	1.06955E-02	4.19772E-02
COL5A1	-0.01218	1.76664	1.06990E-02	4.19842E-02
PTPRH	-0.01190	1.68953	1.07458E-02	4.20657E-02
PIGR	-0.01195	1.62475	1.08240E-02	4.21983E-02
RIMS2	-0.01030	1.52381	1.08339E-02	4.21983E-02
LY9	-0.01119	1.53821	1.08643E-02	4.22663E-02
FDXR	-0.01115	1.39127	1.10048E-02	4.26469E-02
CACNA1F	-0.01096	1.52426	1.10093E-02	4.26469E-02
GBP1	-0.01029	1.43557	1.10230E-02	4.26645E-02
MAGI1	-0.01181	1.74786	1.10599E-02	4.27005E-02
UNC45B	-0.01038	1.45135	1.10714E-02	4.27218E-02
C1orf116	-0.01035	1.39990	1.11400E-02	4.28092E-02
DCC	-0.01088	1.48474	1.11992E-02	4.29464E-02
GPR132	-0.01113	1.52958	1.12081E-02	4.29505E-02
APOB	-0.01396	2.06744	1.13099E-02	4.31609E-02
AMOT	-0.01074	1.44135	1.13511E-02	4.32369E-02
PLCB4	-0.01141	1.45504	1.13923E-02	4.33209E-02
PLXDC1	-0.01109	1.39037	1.14380E-02	4.34222E-02
PCDHGA5	-0.01008	1.36131	1.14406E-02	4.34222E-02
PCDH19	-0.01183	1.62352	1.14647E-02	4.34528E-02
HGSNAT	-0.01942	4.43840	1.15465E-02	4.36271E-02
HSPG2	-0.01321	2.07039	1.15751E-02	4.36881E-02
COL4A2	-0.01272	2.03767	1.15803E-02	4.37007E-02
INO80-AS1	-0.01140	0.94961	1.16997E-02	4.39821E-02
PARD3	-0.01066	1.55851	1.17042E-02	4.39821E-02
PPP1R13B	-0.01214	1.56616	1.18471E-02	4.42296E-02
PREX2	-0.01109	1.52713	1.20244E-02	4.45426E-02
TJP1	-0.01086	1.59088	1.20399E-02	4.45604E-02
MYH13	-0.01404	1.94351	1.20911E-02	4.46622E-02
COL4A6	-0.01319	1.85388	1.21476E-02	4.47893E-02
TNC	-0.01153	1.59849	1.21680E-02	4.48188E-02
AK5	-0.01045	1.08347	1.21918E-02	4.48375E-02
SCUBE3	-0.01170	1.82155	1.22068E-02	4.48502E-02
DNAH5	-0.01378	2.03663	1.23016E-02	4.49717E-02
USP6	-0.01117	1.83525	1.23179E-02	4.49960E-02
ULK2	-0.01330	2.13052	1.24003E-02	4.51821E-02
MYOM1	-0.01139	1.60669	1.25210E-02	4.54040E-02
MROH2A	-0.01169	1.76469	1.25251E-02	4.54040E-02
ATP10A	-0.01213	1.75489	1.25270E-02	4.54040E-02
PLEKHG5	-0.01049	1.30190	1.25589E-02	4.54819E-02
FER1L5	-0.01242	1.89039	1.25850E-02	4.55306E-02
TNR	-0.01269	1.84489	1.26015E-02	4.55594E-02
CACNA1A	-0.01212	2.19500	1.26829E-02	4.57058E-02
PTCH1	-0.01068	1.45867	1.27170E-02	4.57828E-02
MYLK	-0.01312	2.19484	1.29113E-02	4.61875E-02
LAMA2	-0.01293	1.85859	1.29825E-02	4.62886E-02
RGSL1	-0.01100	1.59529	1.30044E-02	4.63360E-02
STX1B	-0.01346	1.15863	1.30813E-02	4.64977E-02
MDGA1	-0.01012	1.52928	1.31172E-02	4.65510E-02
ROBO2	-0.01015	1.49040	1.32088E-02	4.67738E-02
NPR1	-0.01092	1.58780	1.32439E-02	4.68529E-02
EPHA4	-0.01155	1.56308	1.33252E-02	4.70360E-02

Gene symbol	logFC	AveExpr	P.Value	adj.P.Val
ARHGAP31	-0.01164	1.70121	1.33484E-02	4.70722E-02
ADGRE2	-0.01686	3.10336	1.33690E-02	4.70884E-02
SGSM1	-0.01111	1.65642	1.34329E-02	4.72115E-02
XIRP2	-0.01148	1.65573	1.35722E-02	4.74823E-02
PKD1L3	-0.01054	1.51640	1.35906E-02	4.75131E-02
FGD5	-0.01140	1.72821	1.36723E-02	4.76891E-02
MICOS13	-0.01412	1.61908	1.38103E-02	4.80001E-02
ABCA3	-0.01169	1.81910	1.38338E-02	4.80534E-02
MUC22	-0.01128	1.60600	1.38570E-02	4.81202E-02
KDM7A	-0.02095	3.19473	1.38838E-02	4.81672E-02
TACC2	-0.01181	1.76401	1.39666E-02	4.83274E-02
PTPRB	-0.01210	1.85172	1.40050E-02	4.84174E-02
DISP1	-0.01078	1.51139	1.40287E-02	4.84614E-02
OR2M3	-0.01201	1.76933	1.40507E-02	4.84960E-02
ITGAX	-0.01042	1.60251	1.40534E-02	4.84987E-02
NLGN4X	-0.01055	1.51101	1.40866E-02	4.85583E-02
NCKAP5	-0.01194	1.81195	1.41362E-02	4.86813E-02
CYP2A6	-0.01145	1.57870	1.42520E-02	4.89144E-02
FILIP1	-0.01008	1.41882	1.43137E-02	4.90565E-02
PADI1	-0.01099	1.54542	1.43195E-02	4.90565E-02
CLTCL1	-0.01094	1.58006	1.44399E-02	4.93474E-02
C4orf54	-0.01095	1.61488	1.44534E-02	4.93760E-02
IGSF22	-0.01122	1.65482	1.45197E-02	4.94993E-02
LRP2	-0.01396	2.42142	1.45646E-02	4.95555E-02
SPEG	-0.01101	1.72127	1.45955E-02	4.96190E-02
PTPRM	-0.01094	1.70670	1.46005E-02	4.96255E-02
FBN2	-0.01262	1.88171	1.47097E-02	4.98892E-02

Table 12-10 List of 470 genes up-regulated in LumB. Genes are sorted by adjusted p-value.

Gene symbol	logFC	AveExpr	P.Value	adj.P.Val
PCBP1	0.04281	8.90786	1.75738E-08	2.77584E-04
GANAB	0.03607	8.68838	3.57195E-08	2.77584E-04
ATP5PF	0.04590	8.19950	3.75893E-08	2.77584E-04
SNRPE	0.04005	8.68016	4.54757E-08	2.77584E-04
RPL36AL	0.03922	10.25176	5.93238E-08	2.89690E-04
MT-CO2	0.02366	18.27575	1.01699E-07	3.72853E-04
RPL10A	0.03508	11.45550	1.06896E-07	3.72853E-04
DCAF7	0.03713	7.71259	1.62280E-07	4.95278E-04
MT-ND1	0.02145	16.69997	2.25553E-07	5.31419E-04
ST13	0.03786	8.83696	2.36232E-07	5.31419E-04
DYNC1I2	0.03867	6.84911	2.39417E-07	5.31419E-04
CCNA2	0.03648	6.61307	2.80743E-07	5.34646E-04
NOLC1	0.03707	8.70361	2.95104E-07	5.34646E-04
HMGA1	0.03807	9.49501	3.20949E-07	5.34646E-04
SRP14	0.03666	10.20448	3.28461E-07	5.34646E-04
MYL12B	0.03997	8.89776	3.68653E-07	5.43173E-04
MT-ND4L	0.02791	12.67085	4.19619E-07	5.43173E-04
ATP5MG	0.03857	8.39254	4.53608E-07	5.43173E-04
GAPDH	0.02698	13.37899	4.65549E-07	5.43173E-04
RPL38	0.03257	8.18758	4.72929E-07	5.43173E-04
MORF4L2	0.03818	6.34129	4.95931E-07	5.43173E-04
RPL34	0.02718	12.00365	4.99694E-07	5.43173E-04
MT-CO1	0.01912	17.41200	5.13166E-07	5.43173E-04
COX17	0.03793	4.06104	5.33918E-07	5.43173E-04
DKC1	0.03246	9.99851	6.09552E-07	5.88124E-04
MRPL42	0.03919	7.78536	6.38625E-07	5.88124E-04

Gene symbol	logFC	AveExpr	P.Value	adj.P.Val
PSMB1	0.03671	10.21944	6.50367E-07	5.88124E-04
HNRNPA2B1	0.02114	14.56030	7.08725E-07	6.09978E-04
RPL6	0.02317	13.36514	7.24498E-07	6.09978E-04
SF3B2	0.02979	10.06147	7.82661E-07	6.35110E-04
MT-CO3	0.01828	17.77522	8.15705E-07	6.35110E-04
MRPL51	0.03683	9.57248	8.32385E-07	6.35110E-04
HMGB2	0.02306	13.71668	8.60993E-07	6.37031E-04
MINPP1	0.04095	6.25782	9.30243E-07	6.42082E-04
CDKN3	0.04319	4.87634	1.06360E-06	6.42082E-04
ERH	0.03305	10.83023	1.06524E-06	6.42082E-04
EBNA1BP2	0.03889	7.11128	1.08721E-06	6.42082E-04
DDX3X	0.03101	10.20837	1.09922E-06	6.42082E-04
NSA2	0.03344	7.43341	1.19798E-06	6.42082E-04
ZNF830	0.03677	3.33062	1.19994E-06	6.42082E-04
CALR	0.03372	10.41239	1.21410E-06	6.42082E-04
NDUFS5	0.03730	8.41927	1.22709E-06	6.42082E-04
MRPL32	0.04323	5.82968	1.25052E-06	6.42082E-04
LARP1	0.03510	9.56164	1.25966E-06	6.42082E-04
SMAD1	0.03536	3.97188	1.26397E-06	6.42082E-04
FAU	0.03735	8.16511	1.28892E-06	6.42082E-04
CYCS	0.03385	7.45220	1.29635E-06	6.42082E-04
CPSF1	0.04018	7.40683	1.30401E-06	6.42082E-04
HSP90AB1	0.02479	13.13367	1.30438E-06	6.42082E-04
RPS18	0.03107	10.22156	1.33543E-06	6.42082E-04
DAD1	0.04077	6.78940	1.35781E-06	6.42082E-04
PROSER1	0.03117	7.72172	1.36747E-06	6.42082E-04
VBP1	0.03727	5.28566	1.44342E-06	6.64952E-04
SLIRP	0.03527	8.75842	1.48266E-06	6.70381E-04
ZC3H13	0.03222	7.27199	1.60996E-06	7.14707E-04
VDAC3	0.03929	7.89563	1.67971E-06	7.32353E-04
RPL37A	0.02863	11.23615	1.71076E-06	7.32807E-04
YBX1	0.02221	11.91558	1.95781E-06	8.24170E-04
FEM1B	0.03630	6.17088	2.05194E-06	8.49157E-04
MT-CYB	0.02186	15.48561	2.09365E-06	8.51976E-04
MT-ND5	0.02243	14.92463	2.21062E-06	8.84830E-04
SSBP1	0.02875	10.18375	2.26601E-06	8.92370E-04
SNAP29	0.04166	5.25728	2.35426E-06	9.12408E-04
MT-ATP6	0.02054	15.83851	2.40251E-06	9.16558E-04
ANAPC1	0.02993	7.13664	2.53352E-06	9.24440E-04
C11orf58	0.03791	7.25175	2.54733E-06	9.24440E-04
PMP22	0.04015	5.87391	2.56115E-06	9.24440E-04
GUCD1	0.03077	10.25727	2.60177E-06	9.24440E-04
SAP18	0.03902	8.01232	2.61684E-06	9.24440E-04
PDAP1	0.03568	8.32208	2.65035E-06	9.24440E-04
RPL28	0.02037	12.26965	2.78560E-06	9.51970E-04
RFC4	0.03314	7.35368	2.80725E-06	9.51970E-04
PIN4	0.03663	5.41131	2.89856E-06	9.69470E-04
TLN1	0.02984	10.46650	3.02640E-06	9.98550E-04
RPL12	0.02180	13.53187	3.09657E-06	1.00808E-03
MT-ND4	0.02061	15.82931	3.24639E-06	1.04295E-03
ERP29	0.03208	8.75326	3.31662E-06	1.05167E-03
CANX	0.02707	9.54985	3.40452E-06	1.06570E-03
ANP32B	0.02968	10.93912	3.65331E-06	1.12910E-03
PFDN5	0.03633	8.23780	3.75282E-06	1.14536E-03
YIPF5	0.03781	3.58945	4.00210E-06	1.20636E-03
SF1	0.02897	9.42727	4.08653E-06	1.21679E-03
RPL18AP3	0.03153	6.50914	4.29323E-06	1.21817E-03
MORF4L1	0.03368	8.94812	4.35557E-06	1.21817E-03

Gene symbol	logFC	AveExpr	P.Value	adj.P.Val
TET3	0.03350	7.49610	4.36318E-06	1.21817E-03
YBX3	0.03564	8.58906	4.37975E-06	1.21817E-03
PPP2R5C	0.02957	7.61909	4.38662E-06	1.21817E-03
HSPD1	0.02617	10.74289	4.39053E-06	1.21817E-03
DYNLL1	0.03713	9.39150	4.78393E-06	1.31241E-03
AC116533.1	0.03022	7.64445	4.90131E-06	1.31969E-03
RPL7A	0.02087	12.00393	4.94494E-06	1.31969E-03
RAD21	0.03152	7.98624	4.97263E-06	1.31969E-03
TRIM33	0.02942	7.49798	5.22665E-06	1.36106E-03
GLOD4	0.03752	6.63561	5.24000E-06	1.36106E-03
MTDH	0.03095	8.07591	5.38327E-06	1.38356E-03
CTNNA1	0.03446	8.08819	5.48396E-06	1.39475E-03
TSPAN17	0.02944	9.40596	5.75074E-06	1.44753E-03
EIF3M	0.03011	8.79022	5.92226E-06	1.47549E-03
NDUFB4	0.03600	5.71309	6.22708E-06	1.52052E-03
MORF4L1P1	0.03230	5.79955	6.22754E-06	1.52052E-03
ODC1	0.03365	8.15786	6.35008E-06	1.53509E-03
ALAD	0.03262	9.63637	6.66090E-06	1.59444E-03
SUMO2	0.03450	9.01548	6.84963E-06	1.62369E-03
NCL	0.01798	14.33461	6.99202E-06	1.64151E-03
IQGAP2	0.02966	7.31564	7.12828E-06	1.65756E-03
UQCRB	0.03011	8.81885	7.26835E-06	1.66594E-03
SF3B3	0.02977	9.60175	7.35018E-06	1.66594E-03
MT-ND3	0.02720	9.46182	7.36899E-06	1.66594E-03
NCAPD2	0.03110	9.39712	7.63679E-06	1.69517E-03
MTCO1P12	0.02417	8.66767	7.63923E-06	1.69517E-03
RPL14	0.02037	13.67380	7.70660E-06	1.69517E-03
GTF3C5	0.03424	6.25786	7.82263E-06	1.70533E-03
FBH1	0.03194	7.34152	8.41541E-06	1.81456E-03
SEPT7	0.03196	7.30869	8.47233E-06	1.81456E-03
TMX4	0.03529	5.23818	8.73077E-06	1.83909E-03
PNRC2	0.03692	6.85914	8.79262E-06	1.83909E-03
EI24	0.03548	7.11091	8.81281E-06	1.83909E-03
ZNF787	0.03128	2.18408	9.03987E-06	1.87049E-03
YPEL5	0.03422	5.06688	9.30029E-06	1.90091E-03
C1QBP	0.03832	8.33944	9.41772E-06	1.90091E-03
SRSF3	0.03137	9.86354	9.43012E-06	1.90091E-03
SET	0.02812	9.45076	9.54046E-06	1.90091E-03
PPP1R9B	0.03500	6.29869	9.63233E-06	1.90091E-03
RPS15A	0.01971	14.07998	9.65404E-06	1.90091E-03
DANCR	0.04406	5.15088	9.78900E-06	1.90112E-03
RPS12	0.02769	12.05757	9.93203E-06	1.90112E-03
TPM3	0.03160	8.92172	1.00315E-05	1.90112E-03
RPL29	0.03396	9.68386	1.00629E-05	1.90112E-03
EEF1A1	0.01822	14.86227	1.00944E-05	1.90112E-03
RPS27L	0.03717	5.60492	1.01833E-05	1.90112E-03
TUBA1C	0.03419	9.97205	1.03261E-05	1.90112E-03
MTPN	0.03269	6.54792	1.03437E-05	1.90112E-03
NOC2L	0.03454	6.52394	1.04258E-05	1.90112E-03
ANP32E	0.02885	8.33420	1.04337E-05	1.90112E-03
TSR1	0.02995	7.65400	1.07844E-05	1.95045E-03
KTN1	0.03021	8.38239	1.13928E-05	2.04534E-03
MT-ATP8	0.02105	12.63082	1.26071E-05	2.22543E-03
H2AFZ	0.03297	9.08556	1.26686E-05	2.22543E-03
MYCBP2	0.03046	7.49917	1.26693E-05	2.22543E-03
HDAC6	0.03609	7.44906	1.28989E-05	2.24957E-03
STX17	0.03812	4.87456	1.30840E-05	2.26566E-03
SYF2	0.03019	6.93115	1.41075E-05	2.42569E-03

Gene symbol	logFC	AveExpr	P.Value	adj.P.Val
RPS24	0.01913	14.28955	1.43806E-05	2.44595E-03
KDM1A	0.03015	8.15954	1.44256E-05	2.44595E-03
RAB7A	0.03350	7.34396	1.49437E-05	2.51478E-03
DNAJC8	0.03520	7.25336	1.50376E-05	2.51478E-03
ARHGAP11A	0.03344	6.50675	1.53251E-05	2.54542E-03
XRCC5	0.02343	10.14593	1.54878E-05	2.55507E-03
PDCD10	0.03703	5.36662	1.57946E-05	2.58820E-03
YWHAQ	0.03160	9.03870	1.62902E-05	2.63352E-03
MED4	0.03567	5.70137	1.64382E-05	2.63352E-03
PSMA4	0.02894	9.16064	1.64769E-05	2.63352E-03
UBE2J1	0.03268	5.21894	1.65026E-05	2.63352E-03
MTHFD2	0.03413	5.67393	1.66129E-05	2.63390E-03
RPS14	0.03608	7.03028	1.70182E-05	2.66432E-03
MPP1	0.03335	8.64953	1.70230E-05	2.66432E-03
PARP1	0.02705	9.43940	1.71703E-05	2.67026E-03
RPL41P2	0.01341	1.26432	1.75320E-05	2.70924E-03
PBX2	0.03710	8.63425	1.80531E-05	2.77222E-03
RPL7	0.01944	13.88225	1.82528E-05	2.78538E-03
UBC	0.02781	9.01480	1.84013E-05	2.79060E-03
C8orf59	0.03491	3.41434	1.87327E-05	2.82331E-03
RPS4X	0.01829	14.96670	1.94064E-05	2.90692E-03
PPRC1	0.03676	7.81491	1.96197E-05	2.90758E-03
SNRPD1	0.03176	11.04809	1.96490E-05	2.90758E-03
ATP5MC1	0.03382	9.55567	1.97702E-05	2.90789E-03
RPS13	0.02054	12.93051	1.98985E-05	2.90923E-03
KCNH2	0.02479	11.57192	2.01702E-05	2.92369E-03
RPL5	0.02477	11.98014	2.02701E-05	2.92369E-03
MTATP6P1	0.02000	11.64615	2.03566E-05	2.92369E-03
PSMC3	0.03155	8.56566	2.05190E-05	2.92977E-03
SPN	0.02254	11.97616	2.18748E-05	3.10520E-03
MT-RNR2	0.01948	17.32662	2.22772E-05	3.14404E-03
RPS11	0.02962	9.16394	2.29097E-05	3.21473E-03
TPI1	0.03559	5.77928	2.32019E-05	3.22322E-03
PTTG1	0.03457	8.11279	2.32342E-05	3.22322E-03
SERBP1	0.02580	10.46712	2.37062E-05	3.25459E-03
MARCH8	0.03397	6.35193	2.37269E-05	3.25459E-03
GNB2	0.04151	6.03792	2.39598E-05	3.25918E-03
MBNL1	0.02776	9.54295	2.40274E-05	3.25918E-03
RPS27	0.02295	11.44530	2.43639E-05	3.28657E-03
SIGMAR1	0.03445	7.57996	2.45818E-05	3.29774E-03
RBMX	0.02335	9.90730	2.47479E-05	3.30188E-03
TOP2B	0.02773	7.92960	2.49336E-05	3.30858E-03
ATP5MC2	0.02979	8.53002	2.50851E-05	3.31070E-03
MFF	0.02972	4.49131	2.53746E-05	3.33089E-03
AC018475.1	0.01877	1.67982	2.59159E-05	3.38376E-03
FDPS	0.03204	7.19460	2.60787E-05	3.38690E-03
HNRNPA1	0.01984	14.07765	2.64484E-05	3.41100E-03
RPS20	0.02266	12.30227	2.65437E-05	3.41100E-03
CNBP	0.03265	8.20026	2.68533E-05	3.41999E-03
HLTF	0.03104	6.53055	2.68937E-05	3.41999E-03
RSL24D1	0.03186	6.46627	2.75304E-05	3.48156E-03
RIPOR1	0.03137	7.67357	2.76631E-05	3.48156E-03
BSG	0.02869	7.87697	2.79713E-05	3.50229E-03
S100A4	0.03118	7.29787	2.82817E-05	3.52309E-03
ANAPC16	0.03657	5.77776	2.86658E-05	3.55281E-03
PES1	0.02981	7.36258	2.90794E-05	3.57455E-03
RPS28	0.02706	9.80902	2.91340E-05	3.57455E-03
SLC31A1	0.03769	6.17426	2.93935E-05	3.57843E-03

Gene symbol	logFC	AveExpr	P.Value	adj.P.Val
TSN	0.03110	7.33172	2.94588E-05	3.57843E-03
MFHAS1	0.03063	6.59904	2.96566E-05	3.58463E-03
RBM8A	0.03142	7.28884	3.05242E-05	3.66589E-03
TGFBRAP1	0.03370	7.61289	3.06292E-05	3.66589E-03
ARPC3	0.02974	7.41564	3.08603E-05	3.67554E-03
RPL41	0.02849	8.65586	3.11662E-05	3.67921E-03
COX6C	0.03457	6.64424	3.13489E-05	3.67921E-03
PLRG1	0.03444	4.57747	3.14662E-05	3.67921E-03
G3BP1	0.02894	9.40086	3.14939E-05	3.67921E-03
PRRC2C	0.02862	9.59467	3.19745E-05	3.71757E-03
CBX5	0.02543	9.42042	3.24292E-05	3.75256E-03
RPS3	0.02491	11.26198	3.27000E-05	3.76605E-03
RPL22L1	0.03326	9.54441	3.30268E-05	3.78583E-03
ADSS	0.03667	5.56275	3.39276E-05	3.85925E-03
COX7C	0.02822	9.81005	3.39834E-05	3.85925E-03
UBXN4	0.02960	7.36738	3.42533E-05	3.85953E-03
POLR1D	0.03396	4.90513	3.43020E-05	3.85953E-03
MT-ND2	0.01930	15.29669	3.47095E-05	3.85965E-03
MT-TY	0.03442	4.62750	3.49766E-05	3.85965E-03
AC090498.1	0.01475	1.30365	3.53655E-05	3.85965E-03
CBX3	0.02792	7.35501	3.55920E-05	3.85965E-03
C6orf89	0.03552	6.73474	3.56545E-05	3.85965E-03
TUFM	0.03057	8.29428	3.58533E-05	3.85965E-03
RPL17	0.02991	6.72547	3.58781E-05	3.85965E-03
RPS9	0.02955	9.51475	3.59426E-05	3.85965E-03
SOD1	0.03433	7.19986	3.59754E-05	3.85965E-03
RPL14P1	0.02315	7.43762	3.60360E-05	3.85965E-03
PICALM	0.02774	7.96936	3.60768E-05	3.85965E-03
RPL27	0.02747	10.17980	3.62712E-05	3.85965E-03
DNAJC7	0.02981	7.51257	3.66635E-05	3.85965E-03
PDCD11	0.03365	6.36606	3.67836E-05	3.85965E-03
FTL	0.02820	10.39472	3.68958E-05	3.85965E-03
PBDC1	0.03547	4.77461	3.69327E-05	3.85965E-03
FKBP4	0.03090	8.80044	3.70420E-05	3.85965E-03
TFAM	0.03555	5.52423	3.71485E-05	3.85965E-03
COMMD10	0.02914	2.42337	3.73712E-05	3.86633E-03
FOXRED2	0.03330	5.76255	3.78261E-05	3.89688E-03
PYURF	0.03482	4.15257	3.81860E-05	3.90513E-03
CDC42P6	0.02486	2.94653	3.82260E-05	3.90513E-03
EPCAM	0.02912	9.02970	3.83987E-05	3.90643E-03
RPLP0	0.01447	14.61951	3.87092E-05	3.92167E-03
BBIP1	0.03415	3.65180	3.89615E-05	3.93092E-03
RPS23	0.02124	9.65429	3.98931E-05	4.00584E-03
CNOT2	0.03174	7.72257	4.00583E-05	4.00584E-03
SEC13	0.03254	7.07012	4.01962E-05	4.00584E-03
HNRNPR	0.02316	10.67902	4.05414E-05	4.01009E-03
HMGN1	0.02558	9.57403	4.05674E-05	4.01009E-03
ATP11C	0.03028	5.01266	4.22430E-05	4.15889E-03
PRR11	0.03235	7.93306	4.34023E-05	4.25586E-03
STAT5A	0.02899	8.64493	4.37770E-05	4.27543E-03
HNRNPA0	0.02617	8.52009	4.48574E-05	4.36350E-03
INPPL1	0.03241	6.73745	4.55882E-05	4.41699E-03
GOLPH3	0.03167	4.27520	4.74207E-05	4.56061E-03
RPL39	0.02776	4.71752	4.74756E-05	4.56061E-03
HNRNPH1	0.02076	12.54005	4.76309E-05	4.56061E-03
HNRNPF	0.02498	9.92041	4.81566E-05	4.57567E-03
APEX1	0.02779	8.39733	4.82337E-05	4.57567E-03
PSMG1	0.03468	5.82250	4.83504E-05	4.57567E-03

Gene symbol	logFC	AveExpr	P.Value	adj.P.Val
LASP1	0.03042	8.49971	4.89720E-05	4.61660E-03
PCNA	0.03178	9.20114	4.97499E-05	4.66532E-03
PRPF19	0.03113	9.11154	4.98709E-05	4.66532E-03
SKP1	0.02800	8.84366	5.03703E-05	4.69405E-03
STK38	0.03509	4.36811	5.06126E-05	4.69870E-03
DNAJC15	0.03250	9.06345	5.26155E-05	4.86614E-03
SELENOF	0.03371	7.25494	5.36728E-05	4.92825E-03
EMD	0.03336	6.04432	5.36908E-05	4.92825E-03
RPL23	0.02229	10.07192	5.45842E-05	4.96985E-03
PPIA	0.02025	12.64038	5.47822E-05	4.96985E-03
TUBA1B	0.02593	12.01979	5.50910E-05	4.96985E-03
COPS3	0.02788	8.11903	5.50977E-05	4.96985E-03
RPL6P27	0.02212	8.01752	5.53446E-05	4.96985E-03
RPS6	0.01620	15.12859	5.53653E-05	4.96985E-03
PRPF6	0.03042	7.72455	5.59360E-05	4.99734E-03
CCNI	0.02997	7.75175	5.61509E-05	4.99734E-03
ZCCHC10	0.03134	3.08425	5.62855E-05	4.99734E-03
SNRPF	0.02617	8.84552	5.68086E-05	5.02551E-03
LYAR	0.02785	7.57733	5.77516E-05	5.05025E-03
JARID2	0.02521	9.40818	5.78796E-05	5.05025E-03
HJURP	0.03192	5.93072	5.81798E-05	5.05025E-03
NUP50-DT	0.00847	0.38527	5.86652E-05	5.05025E-03
RPL11	0.01928	13.28094	5.83837E-05	5.05025E-03
RPS7	0.02510	11.41953	5.86044E-05	5.05025E-03
COX17P1	0.01084	0.67067	5.87234E-05	5.05025E-03
RPS15	0.02163	13.42543	5.87431E-05	5.05025E-03
RPS3A	0.02092	11.80680	5.91557E-05	5.06788E-03
H2AFY	0.02762	9.11123	5.98792E-05	5.11192E-03
PTGES3	0.02805	8.83884	6.12826E-05	5.21351E-03
RPP14	0.02919	2.34608	6.15280E-05	5.21621E-03
NPM1	0.01948	13.21526	6.20099E-05	5.23887E-03
MSN	0.03441	7.38704	6.26621E-05	5.27572E-03
DDX21	0.02577	9.84716	6.37651E-05	5.33207E-03
ILF2	0.03301	8.62761	6.37682E-05	5.33207E-03
TUBB	0.01904	13.75690	6.46942E-05	5.39104E-03
LBR	0.02477	8.17891	6.53589E-05	5.42790E-03
YWHAB	0.02876	8.21381	6.59357E-05	5.43744E-03
DERL1	0.03613	5.68544	6.61132E-05	5.43744E-03
LDHA	0.03069	8.88605	6.61419E-05	5.43744E-03
CCNK	0.03279	4.84934	6.69912E-05	5.48878E-03
RPL4	0.01746	13.44454	6.77513E-05	5.51772E-03
PEBP1	0.02832	9.78611	6.81716E-05	5.51772E-03
CCDC47	0.03055	6.76507	6.83150E-05	5.51772E-03
SAR1B	0.03293	4.96043	6.86155E-05	5.51772E-03
AC098583.1	0.01349	1.36340	6.86499E-05	5.51772E-03
FUBP3	0.03000	6.66739	6.87004E-05	5.51772E-03
HSPA8	0.02643	12.56591	6.89589E-05	5.52033E-03
NSFL1C	0.03496	5.49343	7.01038E-05	5.59008E-03
HNRNPDL	0.02585	10.99909	7.02881E-05	5.59008E-03
MTND2P28	0.01974	10.46770	7.10084E-05	5.62903E-03
DGKZP1	0.01821	1.33195	7.16411E-05	5.64758E-03
AC073869.1	0.03193	6.96466	7.18480E-05	5.64758E-03
CPSF2	0.03066	6.13837	7.19364E-05	5.64758E-03
HMGB1	0.01963	13.33747	7.22655E-05	5.65524E-03
COX7A2	0.02905	8.96373	7.26203E-05	5.65538E-03
RPL13AP5	0.02318	9.17941	7.27306E-05	5.65538E-03
DHFR	0.03159	7.48006	7.29895E-05	5.65750E-03
HNRNPA1P48	0.02199	7.51513	7.43120E-05	5.74178E-03

Gene symbol	logFC	AveExpr	P.Value	adj.P.Val
HDAC2	0.02693	8.34607	7.47910E-05	5.76056E-03
USP14	0.02765	7.45566	7.56045E-05	5.80490E-03
BTF3L4	0.03165	4.79455	7.59673E-05	5.81447E-03
H3F3AP4	0.02680	6.07328	7.65086E-05	5.82648E-03
CAPNS1	0.02877	8.27524	7.66014E-05	5.82648E-03
HNRNPH3	0.02343	11.94965	7.76618E-05	5.85708E-03
MAGOHB	0.03609	6.39774	7.77211E-05	5.85708E-03
MICU2	0.02753	4.53318	7.77728E-05	5.85708E-03
IKZF5	0.03428	4.98614	7.80846E-05	5.85708E-03
AHCY	0.03078	7.92643	7.82032E-05	5.85708E-03
DOCK8	0.02865	5.72362	7.85313E-05	5.86367E-03
SMAP2	0.03182	5.56451	7.88352E-05	5.86842E-03
ZFAND1	0.03279	3.96394	7.93230E-05	5.87015E-03
TMCO1	0.03024	6.99687	7.93394E-05	5.87015E-03
PFN1	0.03066	8.70756	8.02142E-05	5.90428E-03
RBM17	0.02613	8.94252	8.04597E-05	5.90428E-03
DYNC1I2P1	0.02438	2.24523	8.05261E-05	5.90428E-03
EMP3	0.03352	4.81797	8.23923E-05	6.02302E-03
DDX18	0.03019	7.77290	8.39274E-05	6.11693E-03
POLR1A	0.02859	8.43886	8.48823E-05	6.16811E-03
GLRX5	0.03508	6.66010	8.54533E-05	6.19118E-03
RPS27A	0.02820	9.72831	8.64497E-05	6.24484E-03
ECH1	0.03337	7.23173	8.77584E-05	6.29877E-03
TAPBP	0.02731	6.80585	8.79460E-05	6.29877E-03
RPL12P4	0.02360	7.71033	8.79702E-05	6.29877E-03
EIF3A	0.02162	11.94668	8.94617E-05	6.38683E-03
AL353593.3	0.02823	1.30144	9.02763E-05	6.42620E-03
METTL8	0.03138	6.49033	9.06333E-05	6.43286E-03
CAT	0.02280	9.68890	9.39814E-05	6.63193E-03
HES6	0.03664	4.11083	9.43481E-05	6.63768E-03
MED1	0.02368	8.72205	9.47447E-05	6.63768E-03
ANKLE2	0.02781	6.96438	9.51546E-05	6.63768E-03
NUCKS1	0.02782	10.43936	9.52196E-05	6.63768E-03
CD44	0.03052	7.22015	9.54221E-05	6.63768E-03
MKI67	0.02077	10.48646	9.70698E-05	6.72187E-03
RRS1	0.03500	4.46988	9.71830E-05	6.72187E-03
HNRNPM	0.02361	10.81693	9.79949E-05	6.74672E-03
PLAA	0.03256	5.48164	9.80949E-05	6.74672E-03
MYC	0.02881	7.96878	9.92930E-05	6.79173E-03
SLC25A5	0.03225	8.31557	9.93057E-05	6.79173E-03
BMS1	0.02818	7.55946	1.00116E-04	6.82800E-03
RPL30	0.02514	9.34956	1.00469E-04	6.83061E-03
MSH6	0.02981	7.29803	1.00713E-04	6.83061E-03
TCEAL9	0.03273	3.11218	1.01613E-04	6.87254E-03
EZR	0.02756	7.44537	1.04006E-04	7.01496E-03
PI4KA	0.03035	6.71708	1.04657E-04	7.03938E-03
TRIB2	0.02748	7.90026	1.05548E-04	7.07982E-03
TNFAIP8L1	0.03689	3.98868	1.07327E-04	7.17310E-03
TRIM44	0.03025	5.96750	1.07643E-04	7.17310E-03
YBX1P1	0.02353	6.22079	1.07820E-04	7.17310E-03
RPL41P1	0.02794	4.74044	1.08723E-04	7.21352E-03
HECTD3	0.03224	6.84608	1.09094E-04	7.21380E-03
SUB1	0.02444	10.20977	1.09634E-04	7.21380E-03
ADAR	0.03075	7.59345	1.09759E-04	7.21380E-03
TERF2	0.03470	4.57662	1.09909E-04	7.21380E-03
TYK2	0.03242	5.43865	1.12049E-04	7.30133E-03
TERF2IP	0.02810	5.72565	1.13033E-04	7.30133E-03
FADS2	0.02910	9.48785	1.13222E-04	7.30133E-03

Gene symbol	logFC	AveExpr	P.Value	adj.P.Val
HEBP1	0.02992	7.20178	1.13379E-04	7.30133E-03
AL136126.1	0.00898	0.63587	1.13410E-04	7.30133E-03
UBR4	0.02278	10.28525	1.13553E-04	7.30133E-03
TOMM20	0.02910	9.42765	1.13699E-04	7.30133E-03
AP2M1	0.02989	7.74360	1.13726E-04	7.30133E-03
EIF1B	0.02911	6.77678	1.13934E-04	7.30133E-03
ENY2	0.03318	6.53419	1.15081E-04	7.32959E-03
RPS21	0.02588	8.76306	1.15105E-04	7.32959E-03
FAM210B	0.03078	5.66230	1.15275E-04	7.32959E-03
CRTC3	0.03312	5.63164	1.18161E-04	7.49359E-03
EIF3I	0.02833	9.01605	1.18564E-04	7.49963E-03
SLC30A5	0.02974	4.99686	1.20812E-04	7.62211E-03
CTTN	0.03082	7.02562	1.21377E-04	7.63796E-03
NADK2	0.02637	6.12976	1.23044E-04	7.71369E-03
TGS1	0.03120	4.93767	1.23228E-04	7.71369E-03
EIF4HP1	0.02208	2.51305	1.23528E-04	7.71369E-03
FAM53B	0.03013	3.48666	1.24485E-04	7.75365E-03
NREP	0.03114	4.61499	1.25355E-04	7.78794E-03
IMP3	0.02829	2.26295	1.27131E-04	7.86896E-03
NPM1P35	0.01116	1.14246	1.27344E-04	7.86896E-03
CAST	0.02609	8.33984	1.27716E-04	7.86896E-03
PHB	0.03089	7.12327	1.28088E-04	7.86896E-03
EIF4H	0.02347	9.98236	1.28270E-04	7.86896E-03
DCAF11	0.03408	5.75148	1.28835E-04	7.87127E-03
MRFAP1	0.03126	6.38801	1.28953E-04	7.87127E-03
AKR1C3	0.02882	8.24730	1.29774E-04	7.90163E-03
E2F4	0.02857	8.79363	1.30227E-04	7.90951E-03
MTCO1P2	0.01016	0.90971	1.32901E-04	8.04890E-03
BABAM1	0.02955	3.77048	1.33181E-04	8.04890E-03
WNK1	0.02348	9.26028	1.34096E-04	8.08417E-03
FER	0.02807	3.15100	1.34920E-04	8.11381E-03
TP53BP2	0.03129	6.56388	1.37027E-04	8.22024E-03
CCND2	0.02516	9.33708	1.38149E-04	8.26527E-03
EEF1B2	0.02515	10.18378	1.38592E-04	8.26527E-03
AL450405.1	0.02544	4.38648	1.39235E-04	8.26527E-03
SEC16A	0.03030	6.80797	1.39262E-04	8.26527E-03
KLF13	0.03060	7.48395	1.39470E-04	8.26527E-03
BCLAF1	0.02304	9.87042	1.39878E-04	8.26941E-03
AC139256.2	0.02245	1.50119	1.42809E-04	8.41992E-03
GNAQ	0.02162	9.37204	1.43114E-04	8.41992E-03
CDV3	0.02993	7.04157	1.43702E-04	8.43423E-03
GCLC	0.02657	6.05715	1.45315E-04	8.48805E-03
YBX1P10	0.02342	5.40078	1.46441E-04	8.53339E-03
MT-ND6	0.02295	11.58770	1.46869E-04	8.53797E-03
NAA50	0.02819	6.98990	1.47907E-04	8.57792E-03
MTND4P12	0.02374	6.12661	1.49188E-04	8.61828E-03
SARS	0.03124	6.43209	1.49309E-04	8.61828E-03
GNL3L	0.02860	8.12712	1.50019E-04	8.63885E-03
NUSAP1	0.03234	6.27373	1.51077E-04	8.64170E-03
CAPZB	0.02903	6.32534	1.51188E-04	8.64170E-03
ANP32A	0.02312	11.13462	1.51277E-04	8.64170E-03
CRYZL1	0.02891	2.56336	1.51485E-04	8.64170E-03
ATG3	0.02876	8.52003	1.52672E-04	8.68912E-03
BUB3	0.03426	5.64565	1.53356E-04	8.70778E-03
HMGB3	0.02811	8.81301	1.55310E-04	8.77228E-03
NUP188	0.02585	7.44049	1.55460E-04	8.77228E-03
TMED2	0.03025	5.84805	1.55570E-04	8.77228E-03
PPP2CA	0.02512	7.35059	1.56618E-04	8.81101E-03

Gene symbol	logFC	AveExpr	P.Value	adj.P.Val
ARL6IP1	0.02638	9.07568	1.57044E-04	8.81468E-03
EIF4G1	0.02480	10.32098	1.58547E-04	8.85165E-03
CARM1	0.02668	7.53457	1.58619E-04	8.85165E-03
ARID1B	0.02467	8.47124	1.58941E-04	8.85165E-03
ZNF618	0.03046	6.21277	1.59153E-04	8.85165E-03
MTHFR	0.02863	7.11569	1.59712E-04	8.86257E-03
TUBB4B	0.02268	11.03697	1.61325E-04	8.91266E-03
RPL21	0.02028	12.18334	1.61345E-04	8.91266E-03
RAB1B	0.03381	4.46633	1.63218E-04	8.96617E-03
ACADSB	0.02998	4.41765	1.64347E-04	8.96617E-03
EXOSC2	0.02679	7.69122	1.64370E-04	8.96617E-03
SERPINB1	0.02959	2.87046	1.64959E-04	8.96617E-03
COA7	0.03017	4.40719	1.65012E-04	8.96617E-03
C11orf74	0.03304	5.34893	1.65140E-04	8.96617E-03
MT-TC	0.01723	1.56207	1.65163E-04	8.96617E-03
SNRPB2	0.03057	7.50750	1.65251E-04	8.96617E-03
SF3B5	0.03239	5.89326	1.66877E-04	9.02837E-03
PHB2	0.02886	7.24571	1.67501E-04	9.02837E-03
SNRPEP4	0.02048	2.13612	1.67507E-04	9.02837E-03
SUMO2P1	0.02225	3.08659	1.67998E-04	9.03488E-03
SLC41A1	0.02699	7.96877	1.69701E-04	9.10643E-03
YTHDF2	0.02604	6.43362	1.76534E-04	9.45231E-03
ATP5F1C	0.02810	7.92205	1.78436E-04	9.53323E-03
FAM32A	0.03236	4.41681	1.79530E-04	9.57020E-03
HNRNPA1P35	0.02159	3.94713	1.80234E-04	9.57020E-03
SHPRH	0.02929	5.07023	1.80304E-04	9.57020E-03
TTC1	0.03271	6.88590	1.83566E-04	9.72221E-03
ZFP91	0.03024	5.63912	1.84047E-04	9.72662E-03
HK1	0.02237	10.77822	1.84552E-04	9.73223E-03
CCDC59	0.03441	7.93776	1.86570E-04	9.81742E-03
AC005912.1	0.02283	6.69279	1.87075E-04	9.82286E-03
MRTO4	0.02943	6.17168	1.89472E-04	9.90788E-03
RPL12P8	0.01649	1.84581	1.89595E-04	9.90788E-03
RPL12P6	0.01603	1.95832	1.90196E-04	9.90788E-03
UBE2Z	0.03196	4.82176	1.90318E-04	9.90788E-03
DDX24	0.02583	8.77302	1.92316E-04	9.95217E-03
FAM199X	0.02629	2.92315	1.92391E-04	9.95217E-03
SNCA	0.03297	5.28955	1.93951E-04	9.99053E-03

12.3 Sequence of *Long fragment in vitro* transcript

5‘-GGGAAGGCCAAGUCGGCCGAGCUCGAAUUCGUCGACCUCGAGGGAUCCGGGCCCUCUAGAUGCGGC
CGCAUGCAUAAGCUUGAGUAUUCUAUAGUGUCACCUAAAUCCCAGCUUGAUCCGGCUGCUAACAAAG
CCCGAAAGGAAGCUGAGUUGGCUGCUGCCACCGCUGAGCAAUAACUAGCAUAACCCCUUGGGGCCUCU
AAACGGGUCUUGAGGGGUUUUUUGCUGAAAGGAGGAACUAUAUCCGGAUAACCUGGCGUAAUAGCG
AAGAGGCCCGCACCGAUCGCCCUUCCCAACAGUUGCGCAGCCUGAAUGGCGAAUGGAAAUUGUAAGCG
UUAAUAUUUUGUUAAAAUUCGCGUUAAAUUUUUGUUAAAUCAGCUCAUUUUUUAACCAAUAGGCCG
AAAUCGGCAAAAUCCCUUAUAAAUCAAAAGAAUAGACCGAGAUAGGGUUGAGUGUUGUUCCAGUUU
GGAACAAGAGUCCACUAUUAAAGAACGUGGACUCCAACGUCAAAGGGCGAAAAACCGUCUAUCAGGG
CGAUGGCCCACUACGUGAACCAUCACCCUAAUCAAGUUUUUUGGGGUCGAGGUGCCGUAAAGCACUA
AAUCGGAACCCUAAAGGGAGCCCCCGAUUUAGAGCUUGACGGGGAAAGCCGGCGAACGUGGCGAGAA
AGGAAGGGAAGAAAGCGAAAGGAGCGGGCGCUAGGGCGCUGGCAAGUGUAGCGGUCACGCUGCGCGU
AACCACCACACCCGCCGCGCUUAAUGCGCCGCUACAGGGCGCGUCCUGAUGCGGUAUUUUCUCCUUAC
GCAUCUGUGCGGUAUUUCACACCGCAUAUGGUGCACUCUCAGUACAAUCUGCUCUGAUGCCGCAUAG
UUAAGCCAGCCCCGACACCCGCCAACACCCGCUGACGCGCCCUGACGGGCUUGUCUGCUCCCGGCAUCC
GCUUACAGACAAGCUGUGACCGUCUCCGGGAGCUGCAUGUGUCAGAGGUUUUCACCGUCAUCACCGA
AACGCGCGAGACGAAAGGGCCUCGUGAUACGCCUAUUUUUAUAGGUUAAUGUCAUGAUAAUAAUGG
UUUCUUAGACGUCAGGUGGCACUUUUCGGGGAAAUGUGCGCGGAACCCCUAUUUGUUUAUUUUUCU
AAAUACAUUCAAAUAUGUAUCCGCUCAUGAGACAAUAACCCUGAUAAAUGCUUCAAUAAUAUUGAA
AAAGGAAGAGUAUGAGUAUUCAACAUUUCCGUGUCGCCCUUAUUCCCUUUUUUGCGGCAUUUUGCC
UUCCUGUUUUUGCUCACCCAGAAACGCUGGUGAAAGUAAAAGAUGCUGAAGAUCAGUUGGGUGCACG
AGUGGGUUACAUCGAACUGGAUCUCAACAGCGGUAAGAUCCUUGAGAGUUUUCGCCCCGAAGAACGU
UUUCCAAUGAUGAGCACUUUUAAAGUUCUGCUAUGUGGCGCGGUAUUAUCCCGUAUUGACGCCGGGC
AAGAGCAACUCGGUCGCCGCAUACACUAUUCUCAGAAUGACUUGGUUGAGU-3‘

12.4 Melting curves and efficiency of rRNA primers

12.4.1 Adopted from Verena Lieb

The primers for the 28S, 18S, 5.8S, and 5S rRNAs were adopted from Dr. Verena Lieb, who had already performed the gradient PCR and restriction digestion, therefore, I only had to conduct standard curve qPCR experiments, which are shown below.

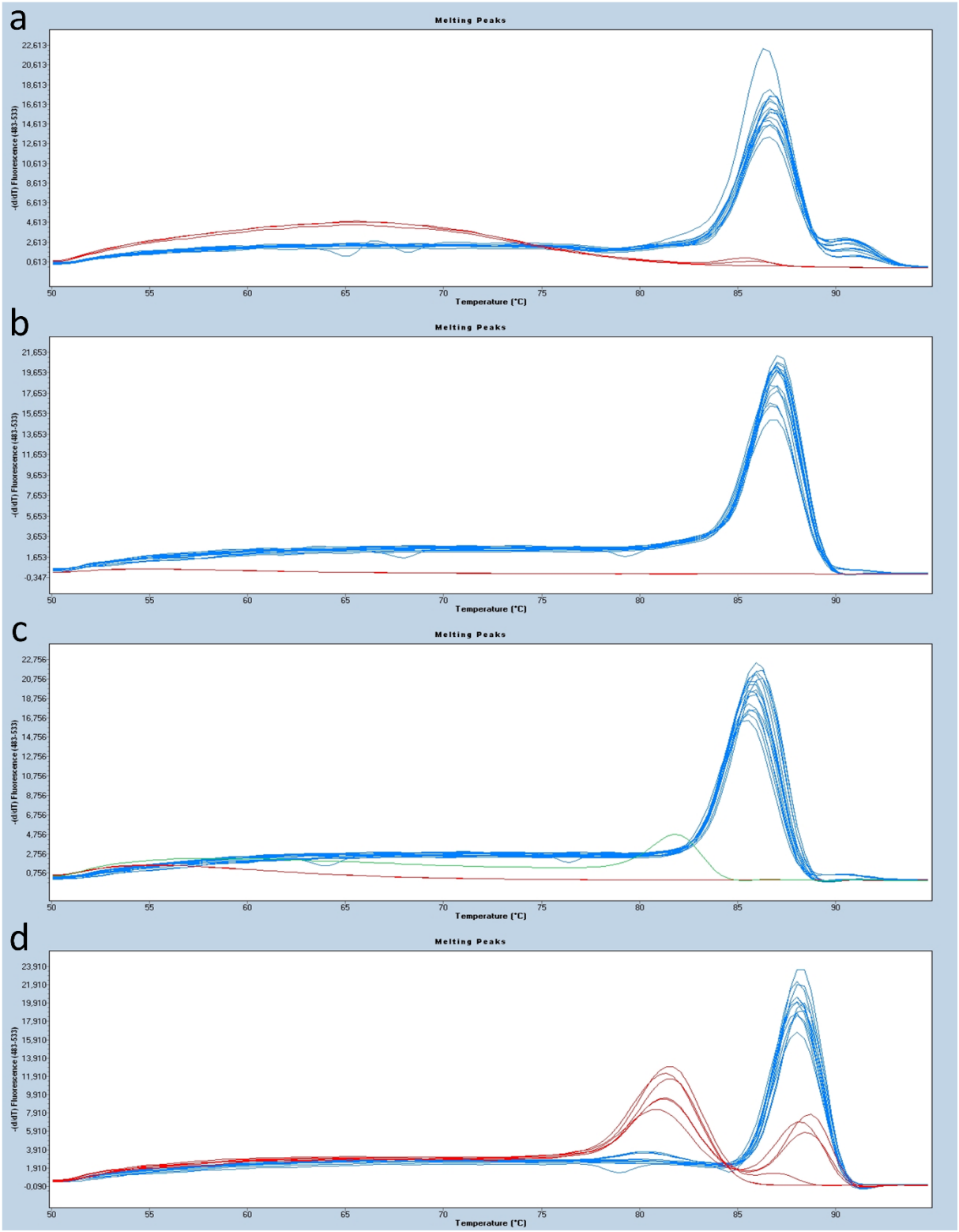


Figure 12-1 Melting curves of 28S, 18S, 5.8S, and 5S rRNA primer amplicons. (a) *28S* rRNA, (b) *18S* rRNA, (c) *5.8S* rRNA, (d) *5S* rRNA. Red lines indicate negative controls or samples with high dilution resulting in primer dimer formation.

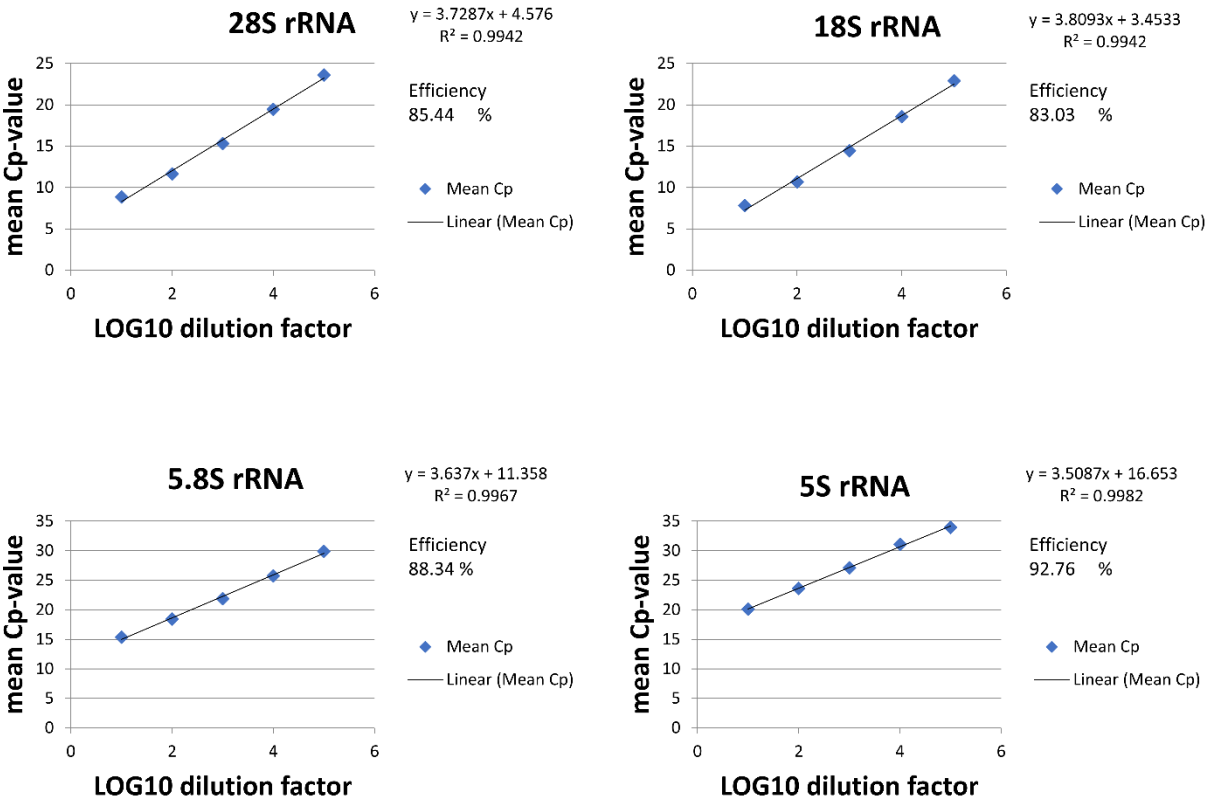


Figure 12-2 Efficiency and specificity of 28S, 18S, 5.8S, and 5S rRNA primers.

12.4.2 New primers

The primers for *16S* and *12S* rRNA were established in the course of this study. Figure 12-3 depicts the gel resulting from the gradient PCR, while Figure 12-4 illustrates the results of the restriction digestion. The melt curve diagrams and efficiency calculations obtained from the standard curve experiment are depicted in Figure 12-5 and Figure 12-6, respectively.

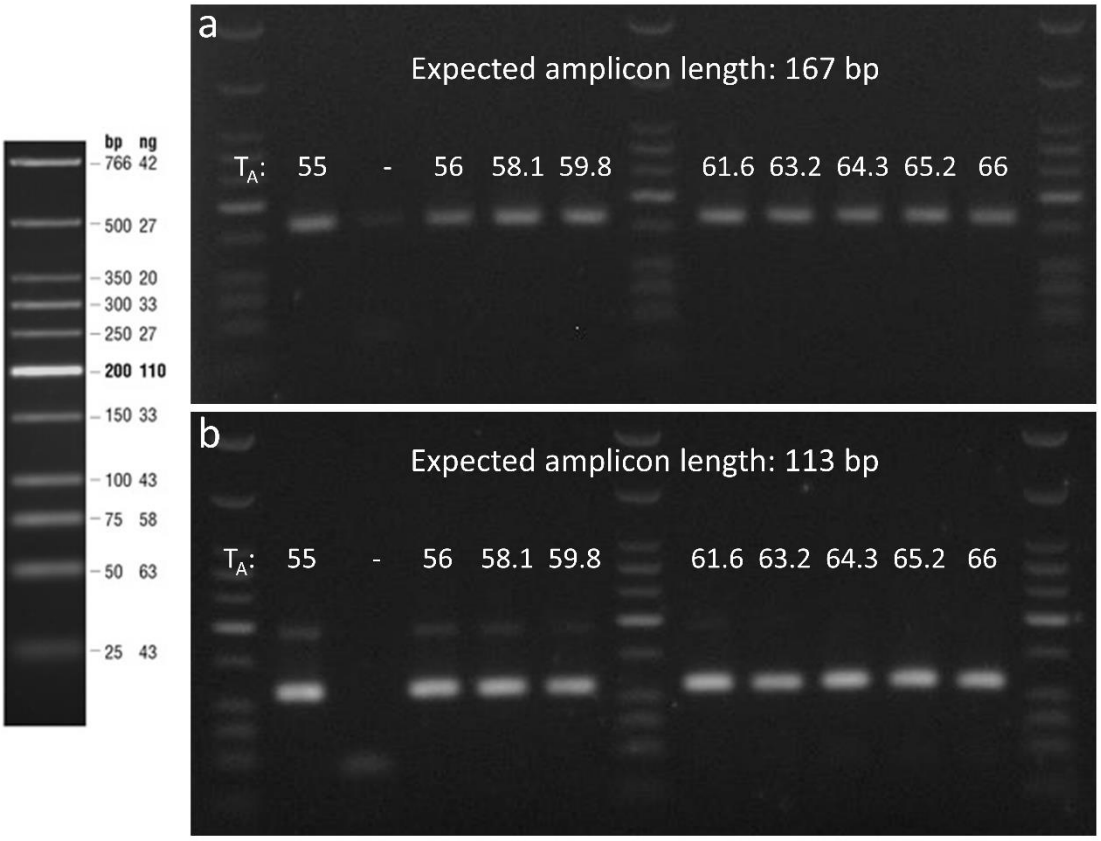


Figure 12-3 Gradient PCR agarose gels of *16S* and *12S* rRNA. The gel shows the intensity of bands depending on annealing temperature used in the PCR. (a) *16S* rRNA, (b) *12S* rRNA.

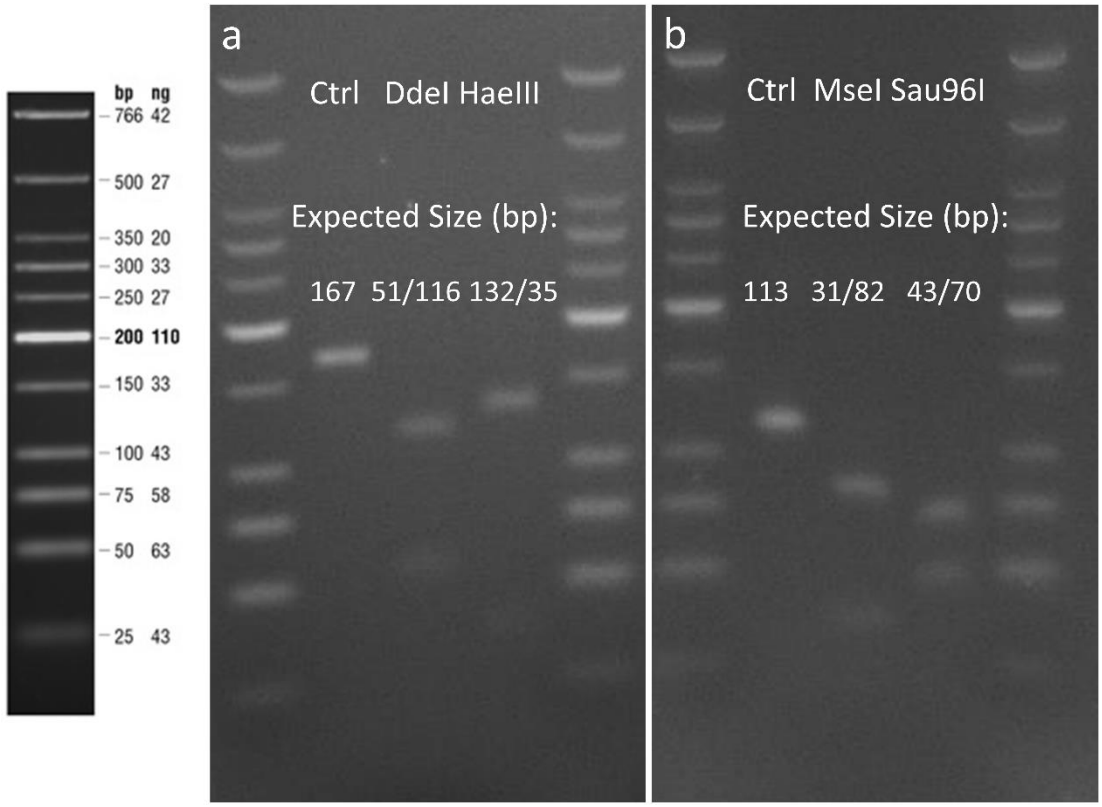


Figure 12-4 Restriction digestion agarose gels of *16S* and *12S* rRNA. The gel displays the bands resulting from restriction digestion of the 16S and 12S rRNA PCR amplicons.

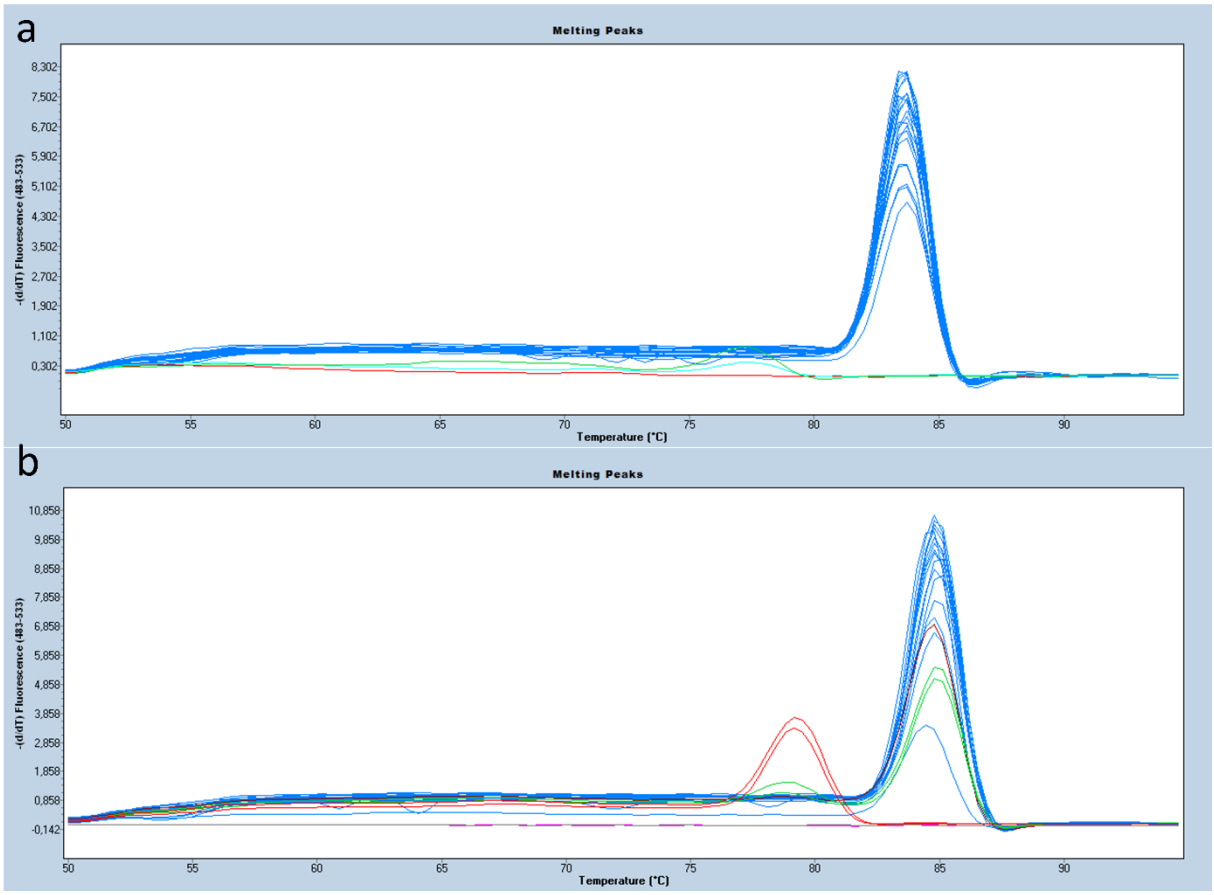


Figure 12-5 Melting curves of 28S, 18S, 5.8S, and 5S rRNA primer amplicons. (a) *16S* rRNA, (b) *12S* rRNA. Red lines indicate negative controls or samples with high dilution resulting in primer dimer formation.

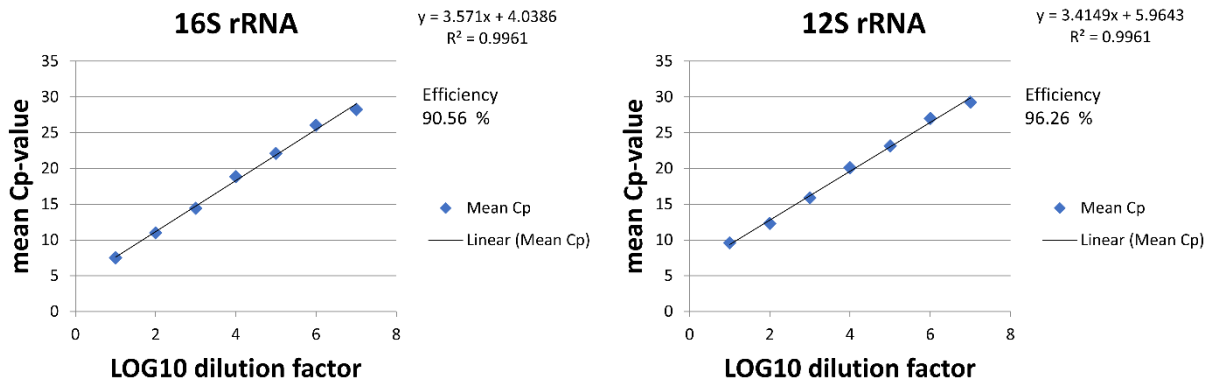


Figure 12-6 Efficiency and specificity of 16S and 12S rRNA primers.

12.5 Sequences of Dr. Pai's 113 blocking oligonucleotides

Table 12-11 List of Dr. Balagopal Pai's blocking oligonucleotides with sequences.

Oligonucleotide name	Oligonucleotide sequence 5'->3'
12S_1	GGT TTG GTC CTA GCC TTT CT
12S_10	AGC GCA AGT ACC CAC GTA AA
12S_11	TGG CAA GAA ATG GGC TAC AT
12S_12	GGT GGA TTT AGC AGT AAA CTA AGA
12S_13	CGT CAC CCT CCT CAA GTA T
12S_2	CAC CAC GAT CAA AAG GAA CA
12S_3	CAC GGG AAA CAG CAG TGA TT
12S_4	CCC AGG GTT GGT CAA TTT C
12S_5	AGC CGG CGT AAA GAG TGT T
12S_6	CCA GTT GAC ACA AAA TAG ACT ACG A
12S_7	TGC TTA GCC CTA AAC CTC AA
12S_8	GAA CAC TAC GAG CCA CAG CTT
12S_9	CGA TCA ACC TCA CCA CCT CT
16S_1	GCT AAA CCT AGC CCC AAA CC
16S_10	GAA AAC ATT CTC CTC CGC ATA
16S_11	GCC CAA TAT CTA CAA TCA ACC A
16S_12	AAA AGT AAA AGG AAC TCG GCA AA
16S_13	CAC CGC CTG CCC AGT
16S_14	CCG TGC AAA GGT AGC ATA ATC
16S_15	AAA TTG ACC TGC CCG TGA
16S_16	GAA GAC CCT ATG GAG CTT TAA TTT
16S_17	TGC ATT AAA AAT TTC GGT TGG
16S_18	TTC ACC AGT CAA AGC GAA CT
16S_19	GGG ATA ACA GCG CAA TCC TA
16S_2	ATT GAA ACC TGG CGC AAT AG
16S_20	ATG GTG CAG CCG CTA TTA AA
16S_21	GGA GTA ATC CAG GTC GGT TTC
16S_22	TAC TTC ACA AAG CGC CTT CC
16S_3	ACC AAG CAT AAT ATA GCA AGG A
16S_4	AAA GCT AAG ACC CCC GAA AC
16S_5	ATA GGT AGA GGC GAC AAA CCT
16S_6	AAC TTT AAA TTT GCC CAC AGA A
16S_7	TGT TAG TCC AAA GAG GAA CAG C
16S_8	GCC TAA AAG CAG CCA CCA AT
16S_9	TCC CAA ACA TAT AAC TGA ACT CCT C
18S_1	CTG GTT GAT CCT GCC AGT AG
18S_10	CCC GAA GCG TTT ACT TTG AA
18S_11	CCG CAG CTA GGA ATA ATG GA
18S_12	CGG GGG CAT TCG TAT TG
18S_13	GCA TTT GCC AAG AAT GTT TTC
18S_14	GAC GAT CAG ATA CCG TCG TAG TT
18S_15	CTT CCG GGA AAC CAA AGT CT
18S_16	AAG GGC ACC ACC AGG AGT
18S_17	CGG ACA GGA TTG ACA GAT TG
18S_18	GTG GAG CGA TTT GTC TGG TT
18S_19	GCG TCC CCC AAC TTC TTA G
18S_2	AAC TGC GAA TGG CTC ATT AAA
18S_20	CCG AGA TTG AGC AAT AAC AGG T
18S_21	TCA GCG TGT GCC TAC CCT AC
18S_22	GCA ATT ATT CCC CAT GAA CG
18S_23	CGC TAC TAC CGA TTG GAT GG
18S_24	CTG GCG GAG CGC TGA
18S_25	GTC GTA ACA AGG TTT CCG TAG G
18S_3	TTG GAT AAC TGT GGT AAT TCT AGA GC
18S_4	TGC ATT TAT CAG ATC AAA ACC AAC
18S_5	GTC TGC CCT ATC AAC TTT CG

Oligonucleotide name	Oligonucleotide sequence 5'->3'
18S_6	GGA GAG GGA GCC TGA GAA AC
18S_7	GGG GAG GTA GTG ACG AAA AA
18S_8	AAA TCC TTT AAC GAG GAT CCA TT
18S_9	TGC TGC AGT TAA AAA GCT CGT
28S_1	GAC CCG CTG AAT TTA AGC AT
28S_10	GCA GCA CTC GCC GAA TC
28S_11	ACC CCC GCG GGA ATC
28S_12	CCG GGG GAG GTT CTC TC
28S_13	CCG ACC CGT CTT GAA ACA
28S_14	GGG GCT CGC ACG AAA
28S_15	CGA GGC CTC TCC AGT CC
28S_16	ACG TGT TAG GAC CCG AAA GA
28S_17	AAC TCT GGT GGA GGT CCG TA
28S_18	TAG CTG GTT CCC TCC GAA GT
28S_19	GTT TTA TCC GGT AAA GCG AAT G
28S_2	CCA GGA TTC CCT CAG TAA CG
28S_20	CCT ATT CTC AAA CTT TAA ATG GGT A
28S_21	GGC CAC TTT TGG TAA GCA GA
28S_22	CCC AGA AAA GGT GTT GGT TG
28S_23	GGA ATC CGC TAA GGA GTG TG
28S_24	AGC GTC GGG CCC ATA
28S_25	TAC GCC GCG ACG AGT AG
28S_26	CAG GTG CAG ATC TTG GTG GT
28S_27	GTG AAC AGC AGT TGA ACA TGG
28S_28	CGA TGG CCT CCG TTG C
28S_29	AGG CGT CCA GTG CGG TA
28S_3	ACA TGT GGC GTA CGG AAG AC
28S_30	CCG GGG AGA GTT CTC TTT TC
28S_31	AAA GCG TCG CGG TTC C
28S_32	GTA CCC ATA TCC GCA GCA G
28S_33	AGG TAA GGG AAG TCG GCA AG
28S_34	CTG GGG CGC GAA GC
28S_35	ACC CCG CGC CCT CTC TCT
28S_36	CAG GGG AAT CCG ACT GTT TA
28S_37	CCC AGT GCT CTG AAT GTC AA
28S_38	AAC GGC GGG AGT AAC TAT GA
28S_39	CGA GAT TCC CAC TGT CCC TA
28S_4	GGT GTG AGG CCG GTA GC
28S_40	GGG GAA AGA AGA CCC TGT TG
28S_41	CGC CGG TGA AAT ACC ACT AC
28S_42	GCG GTA CAC CTG TCA AAC G
28S_43	TCA GGG AGG ACA GAA ACC TC
28S_44	CCT CAC GAT CCT TCT GAC CT
28S_45	CAG GGA TAA CTG GCT TGT GG
28S_46	GTG AAG CAG AAT TCG CCA AG
28S_47	AGA CCG TCG TGA GAC AGG TT
28S_48	GAA CCG CAG GTT CAG ACA TT
28S_49	TCT GTG GGA TTA TGA CTG AAC G
28S_5	AAG CGG GTG GTA AAC TCC AT
28S_50	GCG GAG CCT CGG TTG
28S_51	CCC TTC GTC CTG GGA AAC
28S_52	ACC TGG CGC TAA ACC ATT C
28S_53	CTC CCT CGC TGC GAT CTA T
28S_6	ACC GTA AGG GAA AGT TGA AA
28S_7	GGG GTC CGC GCA GT
28S_8	GCC CGG CGG ATC TTT
28S_9	GGG ACG GCT GGG AAG