University of Stavanger

Faculty of Science and Technology

MASTER'S THESIS

Study program/ Specialization: biological chemistry	Spring semester, 2016. **Restricted Access**
Writer: Tia Tidwell	…… (Writer's signature)
Faculty supervisor: Peter Ruoff External supervisor(s): Oddmund Nordgård & Kjersti Tjensvoll	
Thesis title: Detection and Characterization of Circulating Tumor Cells in Early Breast Cancer Patients	
Credits (ECTS): 60	
Key words: biologisk kjemi, molekylær og cellebiologi, breast cancer, circulating tumor cells, CTCs, biomarkers, qPCR, gene expression, NGS	Pages: 104 + enclosure: 43 Stavanger, 15.06.16 Date/year

UNIVERSITY OF STAVANGER

Detection and Characterization of Circulating Tumor Cells in Early Breast Cancer Patients

by

Tia Tidwell

A thesis submitted in partial fulfillment for the
degree of Master of Science in Biological Chemistry

in the
Department of Mathematics and Natural Sciences
Faculty of Science and Technology

Faculty Supervisor: Peter Ruoff
External Supervisor: Oddmund Nordgård
Co-supervisor: Kjersti Tjensvoll

June 2016

Declaration of Authorship

I, TIA TIDWELL, declare that this thesis titled, 'Detection and Characterization of Circulating Tumor Cells in Early Breast Cancer Patients' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

"Research is what I'm doing when I don't know what I'm doing."

Wernher van Braun

Abstract

BACKGROUND: Detection of circulating tumor cells (CTCs) has demonstrated prognostic significance in metastatic breast cancer. This is less studied in early breast cancer due to the rarity of such cells in early disease and challenges in CTC detection, but shows strong clinical value as well. The purpose of this study was to collect CTCs in early breast cancer patients by use of an enhanced immunomagnetic enrichment method (MINDEC), detect and characterize them by multi-marker quantitative PCR (qPCR), and compare the results with clinicopathological data.

PATIENTS AND METHODS: CTCs were analyzed in 170 peripheral blood samples from 133 early-stage breast cancer patients. Blood samples from 30 healthy female volunteers were analyzed by the same methods as the patient group. CTC detection and characterization was performed using the MINDEC negative enrichment method (multi-marker depletion of leukocytes) followed by multi-marker qPCR. The multi-marker panel was selected based on previous literature, differential expression by serial analysis of gene expression (SAGE) data, and analysis of cell lines, breast tumor samples, and healthy controls. CTC status and clinicopathological factors were analyzed for statistical associations. The markers selected were *EPCAM*, *ERBB2*, *KRT8*, *KRT19*, *SCGB2A2*, *SNAI1*, *SNAI2*, *TWIST1*, and two novel markers, *LUM* and *CCDC80*.

RESULTS: Circulating tumor cells were detected in at least one blood sample in 35 of 133 (26.3%) of the patients and in 37 of 170 (21.8%) total samples. Of the CTC-positive patients, 7 (20%) were positive for more than one marker, 9 (24.3%) expressed only epithelial markers, 22 (59.5%) expressed only EMT markers, and 6 (16.2%) expressed both. Of the 35 CTC-positive patients, *LUM* was detected in 12 (34.3%) and *CCDC80* detected in 10 (28.6%). CTC-status and individual markers were not significantly associated with any clinicopathological features.

CONCLUSIONS: Detection and characterization of CTCs by the presented approach was feasible and revealed heterogeneous gene expression in CTC fractions from early breast cancer patients, with over 60% expressing EMT markers alone or with epithelial markers. Two novel extracellular matrix (ECM) markers (*CCDC80* and *LUM*) were selected for the panel and had the highest detection rates of all markers. Our detection rate of CTCs was similar to that observed with other methods in early-stage breast cancer, while allowing for expanded analysis of CTC characteristics. The clinical significance of these findings remains to be seen and will await further data on the clinical outcome for these patients.

Acknowledgements

I would like to express my deepest appreciation to my supervisors, Oddmund Nordgård and Kjersti Tjensvoll, for their guidance and help over the last several months. They were both energetic and excited about the project and supportive when I needed assistance, especially when it came to areas in which I had less knowledge and experience. I am thankful for their patience with my lab work, but mostly my writing!

I would like to thank Satu Oltedal for taking the time to train me on techniques and always be there as a sample processing back up and lab companion. Many thanks to Siri Lunde for dependable delivery of samples, great lab company, and being so generous and patient with compiling and sharing data. Thanks to Morten Lapin for technical training and being a sounding board for many questions and comments. Finally, I would like to thank the rest of the lab group for support throughout the thesis and also to the molecular biology lab for creating such a welcoming work environment.

Contents

List of Figures

List of Tables

Abbreviations

DCIS	ductal carcinoma *in situ*
ERBB2	human epidermal growth factor receptor 2
ER	oestrogen receptor
PR	progesterone receptor
CTC	circulating tumor cell
EMT	epithelial-mesenchymal transition
EPCAM	epithelial cellular adhesion molecule
KRT	keratin
ECM	extracellular matrix
DTC	disseminated tumor cell
MET	mesenchymal-epithelial transition
ALDH1	aldehyde dehydrogenase 1
RBC	red blood cell
WBC	white blood cell
FACS	fluorescence-activated cell sorting
NGS	next-generation sequencing
qPCR	quantitative polymerase chain reaction
MINDEC	multimarker immunomagnetic negative depletion enrichment of CTCs
WHO	World Health Organization
RNA	ribonucleic acid
mRNA	messenger RNA
DNA	deoxyribonucleic acid
cDNA	complementary DNA
cfDNA	cell-free DNA
ctDNA	circulating tumor DNA

MIC	metastasis initiating cell
CSC	cancer stem cell
MBC	metastatic breast cancer
PBCB	Prospective Breast Cancer Biobank
ISET	isolation by size of epithelial tumor cells
PBMCs	peripheral blood mononuclear cells
ISPs	Ion Sphere Particles
BAM	binary sequence alignment
VCF	variant call format
IGV	Integrative Genomics Viewer
CTM	circulating tumor microemboli

Dedicated to those who know me best.

Chapter 1

Introduction

1.1 | Breast cancer

Cancer is a global health concern, with 8.2 million deaths attributed to the disease in 2012 [7]. Breast cancer in particular cancer in both incidence and mortality in women [7]. In the United States, it is the second most common cause of death after heart disease, with breast cancer being the second most fatal cancer for women [8]. There were an estimated 231,840 new cases and 40,370 deaths of breast cancer in the United States in 2015 [8]. In Norway, 3,090 women are diagnosed with breast cancer every year (average 2010-2014), with 255 of them coming from Rogaland [9]. While there are geographical differences in incidence, mortality does not differ as widely (15.4% in less-developed regions versus 14.3% in more-developed regions) [7]. The incident difference among regions could be due to environmental risk factors of breast cancer, differences in genetic mutation rates [10], or screening rates. On average, incidence rates have been increasing, but mortality has largely been dropping in most countries since the mid-1990s (with the exception of the Asian and South American regions who already have the lowest rates) [7]. Survival has been greatly increased because of the strong focus on breast cancer screening, treatments, and research. However, age is the number one factor in cancer risk, and as the population lives longer, the number of people diagnosed with breast cancer is guaranteed to climb. Therefore, the need for effective treatments and subsequent reduction of mortality is of grave concern.

1.1.1 | Risk factors

The risk factors for breast cancer are similar to cancer in general (age, environment, and lifestyle), with some specific risks due to being a predominantly female cancer (male breast cancer generates 0.5% to 1% of cases [11]). Geography and environment, age, family history, onset of menarche and menopause, pregnancy history, and previous benign breast disease are all well-established risk factors of breast cancer [12]. Other

risk factors include post-menopausal hormone therapy, obesity, alcohol consumption, cigarette smoking, and exposure to ionizing radiation [12]. Hereditary mutations in *BRCA1* and *BRCA2* genes are the most significant genetic risk factors, conferring lifetime risks of 65-80% and 45-85%, respectively, in those that carry them [13]. Some of these risks cannot be modified, but others can be reduced by a change in lifestyle. The most impactful changes to reduce risk are to increase physical activity, eat a healthy low-calorie diet, and to reduce or avoid alcohol consumption [14].

1.1.2 | Diagnosis and classification of breast cancer

Breast cancer is usually found by the presence of a lump either by the patient or their physician, and at this point of detection, 50% of cases have spread to local lymph nodes [12]. However, regular mammograms can detect the tumors before they are felt by the patient and are usually at the ductal carcinoma *in situ* (DCIS) stage (Figure 1.1), or only at about 1 cm for an invasive carcinoma [12].

The presence of a tumor is not enough to yield a diagnosis of breast cancer. In addition to a clinical examination, the normal course of action dictates diagnostic imaging and a tissue biopsy for analysis of prognostic variables such as tumor staging, histological classification, and molecular markers [15]. Additional investigation into presence of lymph node and distant metastases is performed as well. Other experimental tests like genotyping or expression profiling may be done. The clinical course is based on these variables and what they may say about treatment response or whether the disease is operable.

International standards have been created by the American Joint Committee on Cancer for the staging of breast cancer [3]. The cancer is staged based on the state of the tumor (T), lymph nodes (N) and metastases (M) for a combined TNM classification or staging. Tumors are classified depending on the extent of local invasion and size of the tumor (Table 1.1). Carcinoma in situ (usually ductal, DCIS) is a pre-cancerous lesion with some cellular abnormalities, but is localized and considered benign (Figure 1.1). Invasive or infiltrating carcinoma is a malignant tumor with edges of the tumor invading through the basement membrane and into the surrounding tissue. Further classification is made on the basis of lymph involvement and distant metastasis (Table 1.1). Overall staging of the breast cancer by combining these factors aids in categorizing the disease and making a judgment of risk and operability. Stage groups I-III are designated by many different combinations of T and N classifications, without metastatic disease, while stage IV cancer is characterized by the presence of any metastasis regardless of T or N classification [3].

Further analysis is performed histopathologically on tissue samples from the tumor. Histologic grade is designated by how differentiated or abnormal the cells are, ranging from

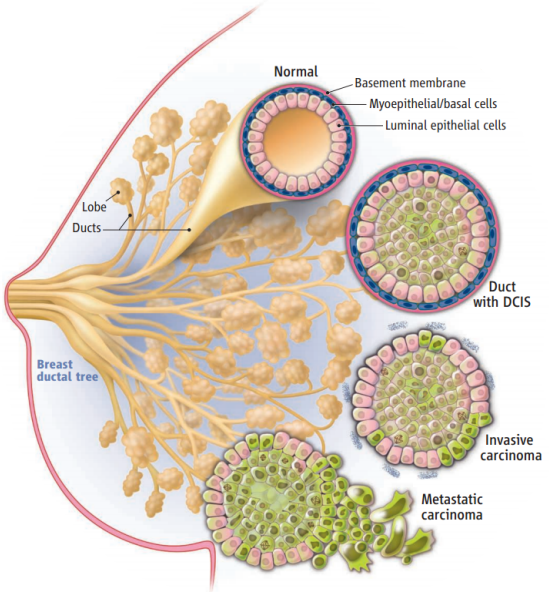

FIGURE 1.1: Anatomy of breast cancer progression.[16] Reprinted with permission from AAAS.

TABLE 1.1: Cancer TNM Staging. From AJCC Breast Cancer Staging 7th Edition [3]

Primary Tumor (T)		Lymph Nodes (N)	
Tx	Primary tumor cannot be assessed	Nx	Regional lymph nodes cannot be assessed
T0	No evidence of primary tumor	N0	$\leq$ 0.2 mm of cluster of less than 200 cells
Tis	Carcinoma in situ	N1	$>$ 0.2-2mm tumor deposit or more than 200 cells
T1	Tumor $\leq$ 20 mm in greatest dimension	N2	Metastases in 4-9 nodes with at least one tumor deposit $>$ 2.0 mm
T2	Tumor $>$ 20 mm but $\leq$ 50 mm in greatest dimension	N3	Metastases in $\geq$ 10 nodes with at least one tumor deposit $>$ 2.0 mm
T3	Tumor $>$ 50 mm in greatest dimension		
T4	Tumor of any size with direct extension to the chest wall and/or to the skin		
Metastases (M)			
M0	No clinical or radiographic evidence of distant metastases		
cM0-(i+)	No clinical or radiographic evidence of distant metastases, but deposits of molecularly or microscopically detected tumor cells in circulating blood, bone marrow, or other non-regional nodal tissue that are no larger than 0.2 mm in a patient without symptoms or signs of metastases		
M1	Distant detectable metastases as determined by classic clinical and radiographic means and/or histologically proven larger than 0.2 mm		

TABLE 1.2: Molecular classification of breast cancers [4, 5].

Subtype	**Molecular characteristics**	Prevalence
Claudin Low	ER^-, $Claudin^-$, $KRT3/4/7^{low}$, $vimentin^+$, $E\text{-}cad^{low}$, $Zeb1^+$	12-14%
Basal Like (Triple-negative)	ER^-, PR^-, $HER2^-$, $KRT5/14^+$, $EGFR^+$	15-20%
Her2 enriched	$HER2^+$, ER^-	10-15%
Luminal A	ER^{high}, $HER2^{low}$	40%
Luminal B	ER^{high}, $HER2^{low}$, $Proliferation^{high}$	20%

grades 1 to 4 with 4 being the highest and most undifferentiated grade [17]. The specialized type of the carcinoma such as tubular, medullary, mucinous is decided histologically, or designated as ductal if there is no special type [12]. Proliferation is documented by expression of the Ki-67 protein as it is present only during active phases of the cell cycle [18]. Based on the data from Sørlie *et al.* [4], the St Gallen expert panel of 2011 [19] also recommended inclusion of the molecular classification of breast cancer for prognostic and predictive assessment. This molecular classification further divides patients into four subtypes of breast cancer based on analyses of oestrogen (ER) and progesterone receptors (PR), and overexpression and/or amplification of the human epidermal growth factor receptor 2 (*ERBB2*/*HER2*) oncogene. The four subtypes are luminal A, luminal B, *ERBB2*-overexpression (*ERBB2*+) and basal-like breast cancer (Table 1.2). These subtypes are significantly correlated with overall survival; with basal-like and *ERBB2*+ subtypes predicting the shortest overall and relapse-free survival [4]. Furthermore, the classification between luminal subtypes reveals differential survival outcomes despite the similar hormonal receptor expression [4].

Genotyping is new, but is still not a well-established clinical practice due to its novelty and lack of validation. In a survey of physicians, most stated that the main hurdles to use were their lack of knowledge and also inaccessibility to the testing [20]. However, 10% of cancers are familial and caused by inherited mutations, with 30% of these being mutations in the well-known *BRCA1* and *BRCA2* genes.[12] The remaining 60% are due to novel and unique mutations with further research into these genetic factors showing great promise in the clinic. Easton *et al.* performed a large review of studies on gene panels and evaluated them for evidence of personal risk prediction. They found the highest risk prediction to be truncated or missense mutations in *BRCA1/2*, *PALB2*, *PTEN*, and *TP53*, with 2-4x increased risk in six genes (*CHEK2*, *ATM*, *NF1*, *STK11*, *CDH11*, and *NBN*), and 100 additional single nucleotide polymorphisms (SNPs) associated with low risk [21]. Kurian *et al.* found 42 gene mutations in *BRCA1/2*-mutation-negative patients that conferred significant additional risk, with 15 prompting treatment changes [22]. On analysis of 86 known risk variants, the top 25% of patients at risk comprised approximately 50% of future cancer cases, making a strong case for preventative genotyping to screen for increased risk; this could spur a reduction in non-genetic risk factors and select for those that would benefit from early mammography screening [23]. Furthermore, Lips *et al.* sequenced triple-negative breast cancer cases and found amplifications, mutations, and chromosomal copy number changes to be associated with clinical outcomes, such as relapse and poor chemotherapy response [24]. In whole genome sequencing of

560 breast cancers, Nik-Zainal *et al.* found 93 driver mutations in cancer genes [25]. An incomplete picture remains though, with recurring mutations also found in non-coding regions as well [25]. It's important to be mindful that however promising the practice of sequencing is, it must be well-validated before widespread and consistent clinical use can occur. More large-scale studies like these need to happen to optimize the predictive value and reduce any harmful clinical outcomes.

1.1.3 | Treatment

Current clinically accepted treatments for breast cancer include surgical removal of the tumor, adjuvant (post-surgery) therapies (cytotoxic chemotherapy, radiation, endocrine therapy), and neoadjuvant (pre-surgery) therapy for large non-metastatic tumors, with one or a mixture of these methods combined depending on the case [26]. Chemotherapy targets and destroys fast growing cells such as cancer cells. Endocrine or hormone-blocking therapies target the hormone-dependent (ER+/PR+) breast cancers and are grouped into two categories: selective estrogen receptor modulators (i.e. tamoxifen) and aromatase inhibitors (i.e. letrozole) [27]. Subtype specific treatments are also available, such as herceptin which targets *HER2*-expressing tumors [27].

The problem with some of these treatments are that they are very general, not targeted to the individual patient or tumor, and affect the entire system with unspecific consequences (from immune depletion to causing new cancers) [26]. Another challenge is that cancers can become resistant to certain therapies; the cells with which the treatment is effective will be destroyed leaving the resistant population behind to thrive and disseminate. This is why targeting treatments reflecting the heterogeneous nature of cancer is vital. As is monitoring of response to treatments to detect such resistance and treat accordingly.

Another challenge to the precise treatment of breast cancer is differentiating low risk patients, without infiltrating tumors, that may be able to avoid aggressive clinical solutions. DCIS is non-infiltrating, but has the potential to progress to infiltrating carcinoma [16]. The decision for clinical action in DCIS cases with no other residual diseases is difficult and can be decided upon through use of molecular markers in addition to the histological findings [5]. The concern of overtreatment in cases that will not progress further raises the need for better prognostic and predictive biomarkers. Over-treatment is a major concern in breast cancer, from unnecessary surgeries to toxic systemic therapies resulting in undue physical, financial, and emotional costs. Treatment decisions can be enhanced with biomarkers in addition to current staging alone. With new biomarkers in mind, more personalized treatments are being developed. Clinical trials are currently using or have used pathway and molecular inhibitors, histone deactylase (HDACs) inhibitors, and poly ADP ribose polymerase (PARP) inhibitors for *BRCA1/2* and *PALB2* deficient cancers [28].

1.1.4 | Disease progression and metastasis

In 90% of cases, fatality of cancer is caused by metastasis of the primary tumor to other organs in the body [29]. This is why it is imperative to catch breast cancer at an early stage before any tumor cells have colonized elsewhere. However, the mechanism of metastasis in cancers is not completely understood. At the basic level, tumor cells spread by detaching from the primary tumor and travel either through the lymphatic or circulatory system. This is why lymph nodes are removed and tested for presence of cancer cells in breast cancer.

In breast cancer, the most common metastatic sites are the lungs, bones, liver, and brain [26]. The preference of cancers for certain organ sites is still a topic of discussion and there are many hypotheses. The most prominent is the "seed and soil" hypothesis by Stephen Pagat in 1889, which states that metastasis formation in certain organs is due to the hospitability of that location to the specific cancer and not due to chance or circulatory patterns [26]. This has been largely proven over the past 100 years, with many studies showing the selective, and usually inefficient, metastasis formation by tumor cells [30]. To further understand the metastatic process, we must elucidate the properties of these metastasis-forming cells and how they interact with other cells in the body.

1.2 | Circulating tumor cells (CTCs)

Circulating tumor cells (CTCs) are cells that have detached from the primary tumor and are circulating in the bloodstream, comprising one of the first steps of metastasis formation. They have been described clinically as long ago as 1869 [31], with sporadic reports being published up until the 1950s [32], and more consistent attention up to present. With recent technological advancements, the isolation and characterization of CTCs have moved to the forefront of cancer biomarker research. In 2013, publications on CTCs broke 1000 articles with over 1100 every year since then and over 500 already in 2016 (PubMed search [33]: circulating tumor cells OR CTCs). The AJCC even include the presence of CTCs in their 7th edition staging standards, with cM0(i+) described as lack of "clinical or radiographic evidence of distant metastases, but deposits of molecularly or microscopically detected tumor cells in circulating blood, bone marrow, or other non-regional nodal tissue that are no larger than 0.2 mm in a patient without symptoms or signs of metastases [3]". (Table 1.1)

It is generally accepted that these cells are responsible for the formation of metastasis. However, the mechanism of extravasation and details of colonization remain unclear and unproven. It is known that they are consistently discovered in the blood of advanced and early breast cancer patients and their relative number does carry prognostic significance [34–38]. The specific characteristics of the CTCs are currently investigated by many

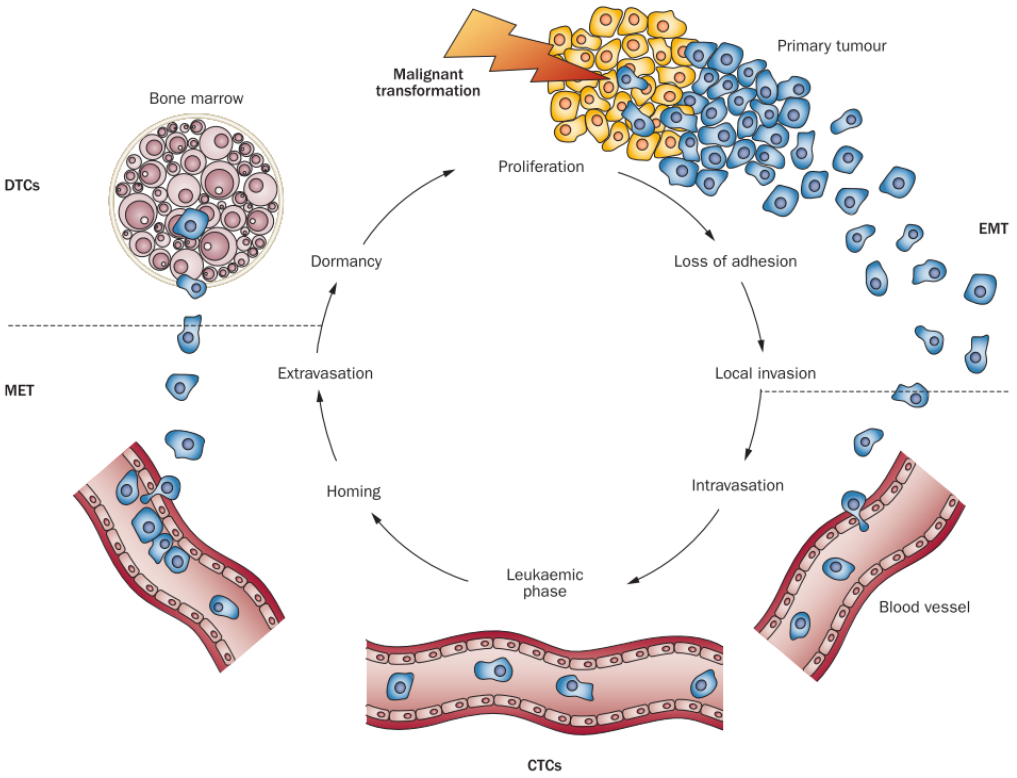

FIGURE 1.2: The metastatic cycle [41]. Reprinted by permission from Macmillan Publishers Ltd: Nature, copyright 2012

different methods with an effort to find qualities of CTCs that can yield even more information on their metastatic potential and mechanism.

1.2.1 | Biology of CTCs

Knowledge of the basic biology of circulating tumor cells is essential in order to successfully isolate and characterize the cells. From the start, a CTC is a primary tumor cell. It becomes a CTC once it has made the journey into the circulatory system. To make this journey, it may undergo numerous changes and can exhibit phenotypes ranging from those similar to the primary tumor to a cell with a divergent phenotype.

At the basic level, a CTC retains the identifiable phenotype of a cancer cell despite its potential for variable molecular profiles. Once in circulation, there are some physical qualities that can differentiate it from the surrounding blood cells. The majority of CTCs are larger than most blood cells [39]. An additional assumption is that it will express the same epithelial markers as the primary tumor. There are many CTC-enrichment methods that rely on this quality (Table 1.3), but it is proven that CTCs are heterogeneous and many exhibit divergent phenotypes from epithelial cells [40].

1.2.1.1 | Epithelial to mesenchymal transition

Most cancers are of epithelial tissue (carcinomas) [26], and thus the cells detached from the primary tumor are of epithelial origin. However, CTCs have been found to exhibit phenotypes divergent from the normal epithelial cells.

It is hypothesized, that the CTC life cycle begins when the tumor cells start to become increasingly invasive and motile through the expression of a more mesenchymal phenotype that allows for these qualities (Figure 1.2). This is referred to as the epithelial-mesenchymal transition (EMT) and presents similarly to the wound healing process with similar recruitment of stromal elements [26]. In EMT, the cytoskeleton of the cell is reorganized and many epithelial markers like E-cadherin, epithelial cell adhesion molecule (EpCAM), claudins, and keratins are expressed at much lower levels [42]. In place of these molecules, mesenchymal markers are expressed such as N-cadherin and vimentin that allow for the weak cell adhesion and loose attachment to the extracellular matrix (ECM) for greater motility [43].

However, CTCs cannot survive in circulation for long. The circulatory system is an inhospitable place for an visiting cell and thus causes a natural filtering of what populations make it through. As mentioned above, a CTC is larger than most blood cells and this could affect travel through small capillaries. Shear forces alone can destroy the cells as well, if they may not be flexible enough to survive them. In addition, the body has natural reaction to cells in the wrong location or expressing foreign/mutated (tumor-specific) markers, and many CTCs will be destroyed by innate immunity. The CTCS that survive these challenges are those that have favorable phenotypes [39]. This may be because they are more mesenchymal-like, stem-cell-like, or have recruited the environment to act in their favor. While most CTCs are destroyed (by internal or external actions), some make it to distant sites, extravasate, and for instance enter the bone marrow. In the bone marrow, they can exist in a dormant state for years. The presence of disseminated tumor cells (DTCs) are confirmed in multiple cancers [37], and in this case, EMT remains while the cell is dormant and before colony formation [44, 45]. The reverse process of EMT, mesenchymal-epithelial transition (MET), is thought to occur when the cell either leaves circulation or its dormant state, and adapts to a region of the body to form metastasis (Figure 1.2). This phenotypic change is very important for tumor cell survival in distant sites.

1.2.1.2 | Cancer stem cells

Cancer stem cells are tumor cells with greater tumorigenic potential than the majority of cells present in the tumor [26, 46]. These breast cancer cells express stem cell markers like CD44, CD47, CD133, ALDH1. CD44 is a marker that is specific to bone cell populations, CD47 is inhibitory to phagocytic cells [46], and CD133 is prominin protein with unknown

function [47]. Aldehyde dehydrogenase isoform 1 (ALDH1) is also targeted as a stem cell marker in breast cancer [48]. Stem cell characteristics are found in many of these studies to be concurrently expressed in both CTCs and DTCs. The proportion of stem cell-like cells expressing CD44 and ALDH1 within breast tumors has also been shown to be of prognostic significance [5].

1.2.1.3 | Tumor microenvironment

Cells exist within a complicated system and rely on intracellular and extracellular interactions for their function. They are inherently fairly elastic, having to exhibit many functions depending on the current needs. It is intuitive that cancer cells would behave in the same way and that some of these adaptive pathways are co-opted and used in a tumor supporting manner. They can also be used by the tumor cells in circulation to evade the immune system, maintain EMT, and to prepare metastatic sites [49]. Inhibition of immune cells in the tumor microenvironment. The down-regulation of cell death and MHC class I genes in CTCs or formation of circulating tumor microemboli (CTM) by recruitment of host cells can both aid in evading immune detection [50]. These host cells may include fibroblasts, leukocytes, endothelial cells, pericytes, and platelets [50]. Due to the already favorable environment local to the primary tumor, CTCs may return from distant sites to reintegrate, known as tumor "self-seeding" [51]. Over-expression of proteins and molecules in these recruitment and niche-forming pathways are potential targets when it comes to CTC isolation, characterization, and even therapeutic targeting.

1.2.2 | CTC enrichment/isolation

CTCs are very rare when compared to other cells present in blood. They are only a few among millions of red blood cells (RBCs), white blood cells (WBCs), platelets, and other molecules. CTCs have been detected in small numbers in 31-67% in metastatic breast cancer patients [52] and 20.2% in early breast cancer patients [34], but there are some cases of very high CTC capture up to 100,000 cells [53]. Because of this, the main focus in CTC research is on the development of specific and sensitive enrichment methods to capture the few cells present.

In the first recorded presence of CTCs, it was possible to visualize them directly in the blood of very advanced cancer patients by microscopy because of the extremely high tumor load present [31]. A later report isolated CTCs by hemolysis of the blood, centrifugation, and fixing of the pellet in paraffin for analysis of sections [32]. These morphological analyses were abandoned due to occasional confusion with normal cells in circulation, to be replaced by immunocytological tests instead [54]. Both morphology and immunocytology are still very commonly used, but in concert with more specific

TABLE 1.3: Selected methods for CTC enrichment that have been tested in breast cancer patients.

Method	**Principle**	**References**
Density Gradient Centrifugation	Isolation of PBMCs and CTCs based on density	Mikhitarian *et al.* 2008[58], Shen *et al.* 2009[59], Obermayr *et al.* 2010[60], Van der Auwera *et al.* 2010[61], Joosse *et al.* 2012[42]
Size-based isolation	Separate CTCs based on size by microfiltration	
	ISET	Farace *et al.* 2011[62]
	Parsortix	Hvichia *et al.* 2016 [63]
	ScreenCell	Desitter *et al.* 2011[64]
FACS	Separation cell sorting by immunofluorescent detection of surface proteins	Vishnoi *et al.* 2015[65]
Positive immunomagnetic enrichment	Isolation of CTCs by magnetic beads coated with CTC-specific antibodies	Markou *et al.* 2011[66], Molloy *et al.* 2011[67], Strati *et al.* 2011[68], Albuquerque *et al.* 2012[69], Nadal *et al.* 2012[70]
	Cell Search: EPCAM	Cristofanilli *et al.* 2004[71], Hayes *et al.* 2006[72], Van der Auwera *et al.* 2010[61], Franken *et al.* 2012[73], Lucci *et al.* 2012[74], Fisher *et al.* 2013[53], Baccelli *et al.* 2013[46], Shiomi-Mouri *et al.* 2014[75], Farace *et al.*2011[62]
	AdnaTest: EPCAM and MUC1	Aktas *et al.* 2009[48], Van der Auwera *et al.* 2010[61], Strati *et al.* 2013[76]
Negative Immunomagnetic enrichment	Depletion of PBMCs by magnetic beads coated with PBMC-specific antibodies	Liu *et al.* 2011[77], Giordano *et al.* 2012[78], Markiewicz *et al.* 2014[79]
CTC chips	Separation of magnetically labeled cells by microfluidics	
	LiquidBiopsy	Strauss *et al.* 2015[57]
	CTC iChip	Ozkumur *et al.* 2013[56], Yu *et al.* 2014[80], Aceto *et al* 2014[81]
None	Extraction of total RNA from blood and proceed to detection methods	Kuniyoshi *et al* 2015[82]

tests [55–57]. More recently, better methods have been developed that use our enhanced knowledge of the molecular qualities of CTCs and the primary tumor whence they came. A summary of methods used for CTC enrichment in breast cancer patients is shown in Table 1.3, with more complete descriptions in the following sections.

1.2.2.1 | Positive enrichment

Positive enrichment is a method that selects specifically for CTCs in a sample, by a number of different methods. The most popular is by immunomagnetic beads selecting

for epithelial markers, leaving behind all blood cells that should not be expressing epithelial markers. Current tests using this method include the AdnaTest, CellSearch, the ^{pos}CTC iChip (also ^{Hb}CTC-Chip and $^{\mu p}$CTC-Chip). All use selection by anti-EPCAM antibodies, but the AdnaTest also used anti-MUC1. CellSearch is an FDA-approved device [83] and is currently being used in interventional trials [84]. The CTC Chip uses a 3-step microfluidics separation process after the bead coating for more pure cell population, enabling whole blood samples to be purified to CTC-populations in one chip. Fluorescence-activated cell sorting (FACS) is also used for some positive selection and is dependent on fluorescent labeling of extracellular surface proteins. This can be used to sort CTCs from blood cells; most commonly CTCs are distinguished by high EpCAM labeling (or other epithelial marker like keratin) and low CD45 labeling, while blood cells are identified by the opposite (EpCAM low and CD45 high). The main advantage of these methods is the lack of contaminating blood cells after enrichment. The major disadvantage to these methods is the potential loss of CTCs that have undergone EMT and either express epithelial markers at low levels or not at all.

1.2.2.2 | Negative enrichment

Negative enrichment is based on the methodology of removing all cells that are not of a CTC phenotype in order to leave a more heterogeneous CTC population behind. There are a few different methods currently implemented. The ^{neg}CTC iChip uses magnetic beads targeting CD45 and CD15 (leukocyte common and granulocyte antigens, respectively) to deplete the sample of white blood cells (WBC) after hydrodynamic cell sorting to remove red blood cells, platelets, and other blood molecules [56]. Other negative enrichment methods rely on a similar immunomagnetic bead depletion, but vary in their targeting. The most basic example is of only targeting CD45-positive cells [77, 79].

The advantage of this method is that it allows the collection of all CTCs, regardless of phenotype. With the heterogeneity of CTCs and limits of EpCAM-dependent capture being considered, this is the best possible approach [85]. The disadvantages is that it can leave more non-CTCs cells behind, because blood cells can vary in their CD45 expression depending on their differentiation state [86]. Including more lineage-specific antigens can enhance the procedure and allow for better depletion. The MINDEC method used in this project is an example of this and uses five antibodies targeted to specific blood cells [87].

1.2.2.3 | Physical selection methods

Methods targeting the differential physical properties of CTCs from normal blood cells range from simple filtration to sorting by dielectrophoresis. Microfiltration based on size

is one of the oldest methods of enrichment. Modern methods have been developed that also target other physical properties such as deformability, density, and electrical properties. Separation using a density gradient, a common method for depleting erythrocytes from the whole blood sample, is used as a first step in some enrichment and as the only method in others [58–61]. Size-based isolation of CTCs by microfiltration is performed in the Isolation by Size of Epithelial Tumor cells (ISET) [62] and ScreenCell methods [64], while the Parsortix [63] system separates on both size and deformability (CTCs are less deformable). Dielectrophoresis is another avenue, with tumor cells being sorted and collected based on their attraction to an electric field [88]. Capturing a more diverse population of cells is the advantage to these non-molecular methods, however they are plagued by the same problem that all the enrichment methods face. Some CTCs may be lost due to size and phenotypic variability and some blood cells may be included for the same reasons.

1.2.3 | Detection and Characterization of CTCs

Shortcomings in enrichment methods can be overcome by sensitive and specific detection techniques. Once obtained, there are many ways to detect and characterize the cells. Many studies are using pure count of cells (such as with CellSearch/CellSpotter) without further characterization and this has to be associated with worse prognosis. The CellSearch system is also the only FDA-approved method. In order to achieve better prediction of prognosis and improved clinical guidance for treatment decision-making, more information needs to be obtained and validated. In most cases, detection and characterization methods rely on the known biology of CTCs. This can involve known signaling pathways[89], expression of transcription factors (*SNAIL, ZEB, TWIST*) [43], and stem cell markers [46–48]. Great potential also lies in the search for novel sources, targets, and mechanisms of action in the tumor cells.

1.2.3.1 | Immunocytology

In many methods, markers on the surface of cells are used for further detection and characterization of the population collected. Use of antibodies and immunofluorescence in flow cytometry or microscopic analysis is used for enumeration of collected CTCs in some. From CellSearch enrichment, the CellSpotter Analyzer is used to stain for nuclei, CD45, and keratins (KRT; 8/18/19) and then nucleated cells that are CD45-negative/keratin-positive are considered CTCs (by microscopic examination) [71]. The same idea is used with flow cytometry and cell sorting. The markers used in some of these studies also include EpCAM, other keratins (7/8), stem cell markers, and more [46, 77, 78, 81]. Additionally, in situ hybridization is used in a couple studies to analyze the cytogenetic profiles of CTCs and in these cases compare them to CTC-established cell cultures or xenografts [53, 70].

TABLE 1.4: Selected methods for CTC detection and characterization that have been tested in breast cancer patients.

Method	**Principle**	**References**
Immunocyto-chemical, microscopy	Detection of surface proteins specific to PBMC and CTCs to distinguish and identify populations – microscopic characterization	Cristofanilli *et al.* 2004, Hayes *et al.* 2006, Van der Auwera *et al.* 2010, Joosse *et al.* 2012, Franken *et al.* 2012, Lucci *et al.* 2011, Nadal *et al.* 2012, Strauss *et al.* 2015, Fisher *et al.*2013, Ozkumur *et al.* 2013, Markiewicz *et al.* 2014, Shiomi-Mouri *et al.* 2014
Immunocyto-chemical, FACS	Detection of surface proteins specific to PBMC and CTCs to distinguish and identify populations – sorting and counting of cell populations	Liu *et al.* 2011, Giordano *et al.* 2012, Aceto *et al* 2014
FISH	Analysis of cytogenic profile by fluorescent nucleic acid probes	Nadal *et al.* 2012
Comparitive Genomic Hybridization	Detection of chromosomal abnormalities through competitive FISH of target and reference samples	Fisher *et al.*2013
EPISPOT	Short-term cell culture in antibody-coated plates to detect tumor cell-specific surface proteins	Alix-Panabieres 2012[90]
RT-qPCR	Detection of CTCS by gene expression profiles	
	Array: high number of targets assayed in sample at once	Vishnoi *et al* 2015
	Multi-marker: sample analyzed with multiple targets, at same time (multi-plex) or not	Mikhitarian *et al.* 2008, Aktas *et al.* 2009, Shen *et al.* 2009, Obermayr *et al.* 2010, Van der Auwera *et al.* 2010, Markou *et al.* 2011, Molloy *et al.* 2011, Strati *et al.* 2011, Strati *et al.* 2013,Giordano *et al.* 2012 Albuquerque *et al.* 2012, Markiewicz *et al.* 2014, Kuniyoshi et al 2015
	Single-plex: analysis of sample with by one target only	Strati *et al.* 2013
NGS	Analysis of mutation (DNA) and/or expression (RNA) profiles of CTCs	Strauss *et al.* 2015, Yu *et al.* 2014, Aceto *et al* 2014
Cell Culture	Creation of CTC-cell lines for monitoring and testing of phenotype and genotype	Yu *et al.* 2014
Xenografts	Injection of subsets of CTCs to identify metastasis-inducing-CTCs	Baccelli *et al.* 2013, Yu *et al.* 2014

Advantages include being able to numerate the CTCs and confirm their presence by visualization. The disadvantages are that the cells themselves are only observed and no other information is obtained outside of surface protein presence and morphology. Some cytological methods allow for further characterization (cell sorting) but others do not (fixation of cell on slide). Also, some variation between studies may be observed when cells are counted by subjective manual methods such as microscopy, or due to differences in labels and probes used for visualization.

1.2.3.2 | Gene Expression

Gene expression or mRNA measurements can be useful to indirectly detect and subtype CTCs after enrichment or detection or after no enrichment at all [82]. In the case of negative depletion or no enrichment, there must be a way to demonstrate the presence CTCs in a pool of other cells. This can be done with varying gene assays for epithelial, EMT, and other markers. The AdnaTest relies on this method after enrichment and uses a multi-plex assay for HER2, MUC1, and EPCAM [91]. Multi-marker qPCR assays such as this are very popular due to the large amount of information obtained about the CTCs, with many studies using custom panels.

Since this is a relatively new method and is continuing to be studied, the methods vary considerably in both design and results [76]. Different genes are targeted; with the attempt to find the best mixture to capture all CTCs and yield the most relevant information. Different primer and probe kits are used (i.e. SYBR green or Taqman).

Advantages are the options available and the flexibility. qPCR analysis is relatively cheap, simple to carry out, and sensitive. As low as 3 copies can be detected with a well-developed assay [92]. Also, this can be a very powerful investigative method, allowing for new candidates to be found on large scales by arrays or sequencing. The disadvantage is that you cannot enumerate the CTCs or visualize morphology with this method. It is important to be aware that capture of CTC-fractions and subsequent analysis of gene expression yields information on potentially a pool of cells and not individual cells. Multiple genes can be expressed, but there is no way to know if they are concurrent in one cell or separately over multiple cells.

1.2.3.3 | Gene Sequencing

The clinical relevance of tumor sequencing in breast cancer is well established (see section 1.1.2). The same benefits can gained from analyzing the genome of CTCs. It has been shown that CTCs can exhibit similar mutations to primary tumors and metastases [93, 94], predicating its use as a liquid biopsy of disease stage, classification, and prediction of response. Some studies have revealed mutations in CTCs that are not identified in the primary but still are clinically actionable [95].

1.2.4 | Clinical Relevance and CTCs as biomarkers

A biomarker has been defined by the World Health Organization (WHO) to be "any substance, structure or process that can be measured in the body or its products and influence or predict the incidence of outcome or disease [96]." Cancer biomarkers are a popular area of research; with special interest in the potential of a "liquid biopsy" in order to have easy access, larger volumes, and almost unlimited time-points for cancer monitoring. This is less invasive compared to surgical procedures or biopsies which yield tumor samples, but are a very limited resource due to difference in tissue sizes and the standard pathology tests that need to be done. Also, if these a "liquid biopsy" can be done prior to surgical removal or biopsy of the tumor, it has even more power as a clinical biomarker. Urine and blood from cancer patients are the most heavily studied due to ease of sampling. Avenues of analysis and targets include: proteomics, transcriptomics (RNA, mRNA, miRNA, CTCs), genomics (CTCs, ctDNA, lncDNA), platelets, and exosomes. Blood is the fluid of choice due to being the circulatory highway of the body.

With the power to form fatal metastases, knowledge of CTCs can be useful clinical tool. They have been shown to have prognostic value for overall survival and some studies are focused on companion diagnostic use of CTCs to improve metastatic treatment outcomes [38, 84]. First-line screening and diagnostics are a more difficult level to reach, but if the methods are optimized, it could be possible in the future.

1.2.4.1 | Prognostic Value

It has already been shown that the CTC load in a patient is a significant prognostic factor in overall and relapse-free survival [72–75]. This has been demonstrated on a large scale using CellSearch. Pooled analyses have been performed on numerous studies on CTC enumeration [34, 35] and detection [52] in metastatic and non-metastatic breast cancer, finding significant associations with overall and relapse-free survival. On a smaller scale, preliminary research has been able to go into even more detail. Specific types of CTCs have been tied to more aggressive cancers and a possible poorer prognosis [58, 59, 67]. This is intuitive since not all cancer cells will survive for implantation. There must be characteristics that some cancer cells have that enable them to survive longer in the bloodstream. Qualities that may effect their survival include deformability, EMT, stem-cell characteristics, and the CTC microenvironment.

1.2.4.2 | Screening and Diagnostics

Early diagnostics and cancer detection from a blood test is one moonshot goal in cancer diagnostics. To achieve this, method sensitivity and specificity have to greatly improve.

This may not be realistic due to the low CTC burden in very early cases, however early detection of relapse is possible [36]. In the same vein, this would be useful for tracking the effectiveness of treatment regiments.

1.2.4.3 | Personalized Medicine

As described, the current practice of tumor characterization relies on small biopsy tissue samples for a clinical judgment on the status of the entire tumor. Tumor sizes vary widely and one small piece of the tumor does not give a whole picture. Tumors are heterogeneous in genotype and histopathology due to nature of clonal evolution and subclonal diversity [97–99]. Problems can arise if treatment is based on only one region of the tumor. A less aggressive cell type could be destroyed leaving the opportunity for the more aggressive cell type to thrive.

As with primary tumor characterization, CTCs could play a role in choosing a targeted treatment. Micrometastases and CTCs left in the body after primary tumor removal can be targeted by adjuvant treatment. However, CTCs can differ in many ways from the primary tumor. In this aspect, it would be useful to consider the characteristics of both. If not done, the primary could be eradicated leaving DTCs to grow and cause later relapse or metastasis [36]. The predictive value of CTCs lies in both information on resistant clones and treatment response in general. Clinical trials have concentrated on these features by measuring treatment response as a function of CTC count and also treatment based on characteristics of both the primary tumor and CTC (with respect to HER2 expression) [84]. Further possibility lies in targeted destruction of CTCs themselves to prevent metastasis [100]. The European CANCER-ID consortium is also focused on the validation of liquid biopsies in cancer [101]. If CTC and other biomarker analysis ultimately allows for less aggressive systemic treatments, it will enhance the quality of life for breast cancer patients.

1.2.5 | Challenges and limitations in CTC analysis

The reward of information gained from the isolation and analysis of circulating tumor cells is great, but challenges in the process are numerous. CTC characteristics currently being analyzed in cancer patients include phenotypic and genomic heterogeneity, EMT-like properties, resistance to anoikis in circulation(self-destruction upon loss of ECM-adhesion), metastatic potential, and single-cell or clustering properties (Figure 1.3) [93]. General hurdles to obtaining this information include the detection of such rare cells, overcoming bias in the methods, and translation into a clinical setting. The methods that struggle in one areas, such as with detection of rare cells (negative depletion of leukocytes), excel in other areas like selection bias, and vice versa with positive selection. However, the methods as a whole are limited by their lack of standardization. Further

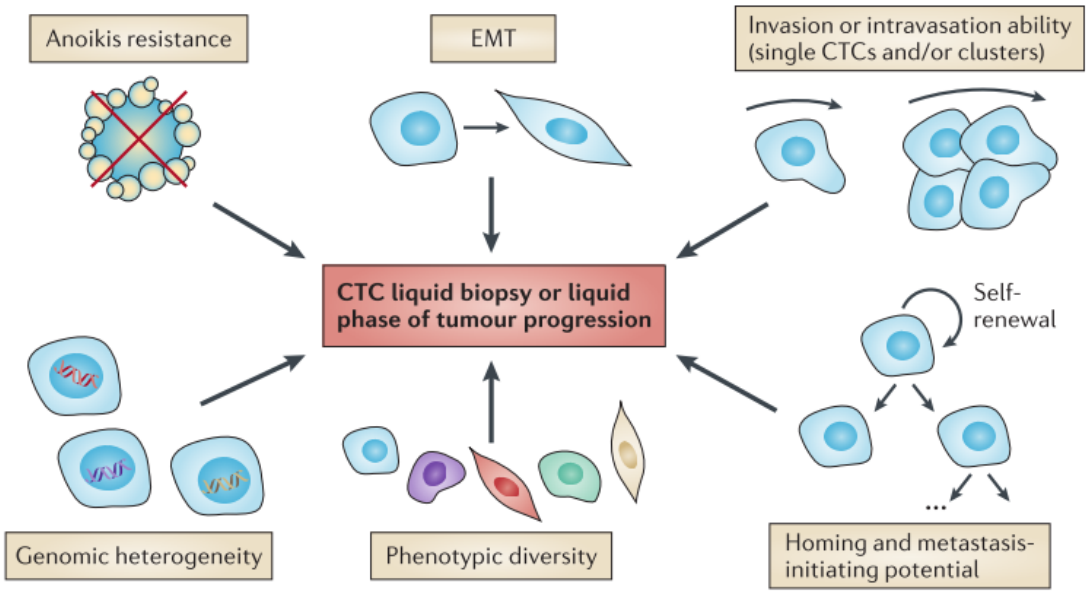

FIGURE 1.3: CTC characteristics as currently described. Reprinted with permission from Macmillan Publishers Ltd: Nature, copyright 2014 [93]

challenge lies in interpreting the meaning of CTCs once detected. Analysis of single cells to understand CTC populations and subsequent assays to ascertain function could be solutions to the problem.

1.2.5.1 | Rarity of cells

A small number of CTCs (commonly between 1-10) are found in the majority of cancer patients [93]. This could be due to the nature of the location of the sampling, the nature of the tumor, or the systemic environment. Portal veins have been considered as an option and found to contain much higher number of cells [102], however this is clearly more invasive than a typical venipuncture. The main reason liquid biopsies are sought after is the ease of sample retrieval. Another option that is less invasive than arterial sampling is leukaphoresis. Several liters of blood can be filtered and collected at one time. In a comparison with peripheral blood and the CellSearch workflow, this method was found to collect a much higher number of CTCs and while revealing significant associations with TNM stage and metastasis-free survival [53].

1.2.5.2 | Capture Bias

All methods are based on assumptions on the cell populations being collected or removed. There is no method that is 100% effective or precise due to basic biological variation. This is further complicated by the heterogeneity of individual tumors between and within patients which is further reflected in the CTC populations. It is difficult to define CTCs

by morphology or molecular and genetic properties. If we would point to a characteristic present in all tumor cells, we would be solving a much larger problem in the cancer treatment.

All of the methods beyond pure enumeration showcase this heterogeneity of CTCs. This is not surprising given the variability of cells in the primary tumor, and the innate ability for cancer cells to adapt to their environment. The best methods going forward will be the ones that allow for capture and detection across a wide variety of cellular characteristics. The more details known, perhaps the better we can understand the cancer and provide more personalized and effective treatments.

There are ways to control for this in both the enrichment and characterization steps to the best of our ability. This can be done by first not selecting CTCs based on EpCAM, as this is known to be a overly-selective property and excludes many cells that may be the most predictive [85, 103]. The selective nature of different methods is made clear in many comparison studies [56, 61, 77, 81]. In addition, the characterization methods should also be inclusive enough to analyze and gain information from as many cells as possible.

1.2.5.3 | Functional Characteristics

Despite the evidence demonstrating the clinical relevance of CTC numbers, the functional characteristics of CTCs are not as intensely investigated. Surface receptors present on the cells, along with gene expression, can give some idea of what is happening within the cell on a molecular level, but how that effects the function of the cell is unknown. The CTCs with the most clinical value are those that survive circulation, dissemination, and go on to form metastasis. Functional assays are needed to find the specific characteristics that support these actions. Some studies have been done that investigate these features, such as metastatic initiating cells (MICs) in xenografts [46], and growing and monitoring cell cultures from CTCs [80, 90].

1.2.5.4 | Lack of standardization and translational medicine

Medical decisions can hopefully be enhanced with the input from CTC science, but many challenges and limitations remain for their translation to the clinic [104]. The methods presented here present only a snapshot of the hundreds of publications every year in CTC analysis. With so many methods and techniques being used, it makes comparison and standardization in the field more difficult. Biologics is a complicated medical field, but to be used in the clinic a CTC method must be rigorously proven and validated and for this, a standard and routine set of methods must be developed. Unfortunately, we are still trying to arrive at what the best methods may be for the most clinical value. The

best method will ultimately be easy, effective, and minimize inaccuracies, with function being more important than novelty.

1.3 | Purpose

The purpose of this project was to:

- evaluate the performance and feasibility of a new negative enrichment method [87] for the collection of CTCs in early breast cancer patients,
- develop and validate a new multi-marker mRNA panel for detection of CTCs by qPCR,
- characterize the CTCs in terms of both epithelial (*EPCAM*, *ERBB2*, *KRT8*, *KRT19*, *SCGB2A2*) and mesenchymal-like (*CCDC80*, *LUM*, *SNAI1*, *SNAI2*, *TWIST1*) characteristics, and
- investigate potential associations between CTC findings and clinicopathological patient characteristics.

Breast cancer is a leading cause of death in women worldwide with incidence that is only going to grow. While overall mortality has decreased, this has not been the case in the most aggressive cancers. This shows that the clinical designation of high-risk cases is not working in addition to or combination with ineffective treatments. Being able to identify patients who are at higher risk of relapse or non-response to treatment is important for the reduction of mortality. This will also reduce the overtreatment of those in a lower risk group. With so many women being diagnosed in enhanced screenings, more of them are being subjected to intense treatment regimens that may not be helpful and even harmful. Outside of health effects, cancer patients may have significant negative physical, financial, and emotional outcomes after intervention. Any alleviation of this burden is warranted.

Current methods do not identify with great accuracy those who are not going to progress further (and should get less treatment) and those that are at true risk of worse outcomes (need better treatment) and both could benefit from more personalized treatments. Patient stratification based on biomarkers (specifically CTCs) hold promise for achieving this level of precision medicine. Presently, there is evidence that CTC count in metastatic breast cancer patients predicts treatment response, progression-free and overall survival. In early-stage breast cancer, CTC number has been associated with reduced survival as well. Some trials have even started to cater treatment based on *HER2* expression of CTCs. However, CTCs are rare cells and detection is difficult. Many current methods in CTC detection are biased to only epithelial CTC populations and investigation into the relevance of other CTC characteristics is limited.

In order to address these challenges, a multi-marker negative enrichment method (MINDEC) was used to collect heterogeneous CTCs in this project. Furthermore, a multi-marker mRNA panel was selected for the detection and characterization of CTCs with variable properties, from standard epithelial to EMT marker expression. Included in the marker panel were two novel markers that have not yet been investigated in breast cancer CTCs.

The patient samples analyzed in this project are the first included in the Prospective Breast Cancer Biobank (PBCB) study. The PBCB study consists of samples from 300 breast cancer patients every 6 months for 10 years following diagnosis. CTCs are to be analyzed alongside circulating cell-free DNA (cfDNA) for comparison with diagnostics, treatment, and outcome. This project, as a part of the larger PBCB study, will aid in the investigation of the predictive and prognostic power of both CTC presence in early-stage patients and their relevant CTC characteristics, as well as improve understanding the role of CTCs in metastasis formation.

Chapter 2

Materials and Methods

2.1 | Materials

2.1.1 | Patient and control blood samples

In total, 170 breast-cancer diagnosed patient samples were analyzed in this project. They came from 133 patients at three time points (Table 2.1). Control samples were obtained from 30 healthy female volunteers.

Clinicopathological characteristics of all patients were recorded and are summarized in Table 3.8. This is data from the baseline visit (Visit 1). The median age of the PBCB patients was 60 (range: 25-85). In contrast, the median age of the control group was 48.5 (range: 33-61). Of the 133 patients, 17 were diagnosed with DCIS (13.7%), and the other 116 with infiltrating breast carcinomas. Infiltrating ductal carcinomas (IDC) were diagnosed in 87 (74.4%) of the patients with the remainder diagnosed with invasive lobular (ILC), mucosal (IMC), papillary (IPC), tubular carcinomas (ITC), and other. 53.8% of patients had Stage 1 tumors, 30.8% had Stage 2, and 1.7% had Stage 3 tumors. Seven of the DCIS tumors (17 total) were designated accordingly as *is*, while 9 were undetermined. 13 of all the patients were histopathologically triple-negative. The adjuvant therapies prescribed included chemotherapy in 78 patients (59.5%), herceptin in 8 (12.9%), and endocrine therapy in 81 (61.8%).

TABLE 2.1: Number of patient samples at each timepoint

Visit Number	Timepoint	Samples
Visit 1	Baseline	117
Visit 1.5	6 months	41
Visit 2	1 year	12
Total		170

TABLE 2.2: Patient clinicopathological characteristics at baseline visit. Fisher's exact test for categorical variables. *Kruskal-Wallis rank sum test for continuous variables.

n	131
Age (median [IQR])	60.00 [53.00, 65.50]
Diagnosis (%)	
DCIS	17 (13.0)
IDC	96 (73.3)
ILC	8 (6.1)
IMC	3 (2.3)
IPC	1 (0.8)
ITC	2 (1.5)
other	4 (3.1)
T Stage (%)	
1	73 (55.7)
2	40 (30.5)
3	2 (1.5)
is	7 (5.3)
undetermined	9 (6.9)
Tumor 1 Size(median [IQR])	16.00 [12.00, 26.75]
Multifocal (%)	16 (12.2)
N Stage (%)	
N0	89 (67.9)
N1	23 (17.6)
N2	5 (3.8)
N3	1 (0.8)
undetermined	13 (9.9)
Metastasis (%)	19 (28.4)
Grade (%)	
1	20 (15.3)
2	47 (35.9)
3	48 (36.6)
DCIS	16 (12.2)
ER Status (%)	
neg	16 (12.2)
pos	99 (75.6)
undetermined	16 (12.2)
PR Status (%)	
neg	34 (26.0)
pos	79 (60.3)
undetermined	18 (13.7)
HER2 Status(%)	
neg	104 (79.4)
pos	11 (8.4)
undetermined	16 (12.2)
Ki67(median [IQR])	31.00 [19.00, 44.00]
Resection (%)	101 (77.1)
Mastectomy(%)	35 (26.7)

2.1.2 | Breast Tumor Samples

The following breast tumor tissue samples were obtained from Asterand Bioscience and used in the validation of candidate mRNA markers.

Sample	Biosample Confirmed Diagnosis	Tumor Grade
Br1	Invasive ductal & lobular carcinoma	2
Br2	Invasive ductal & lobular carcinoma	3
Br3	Invasive ductal & lobular carcinoma	1
Br4	Invasive ductal carcinoma	3
Br5	Invasive ductal carcinoma	1
Br6	Invasive ductal carcinoma	2
Br7	Lobular carcinoma	2
Br8	Lobular carcinoma	1
Br9	Lobular carcinoma	2/3
Br10	Medullary carcinoma	3

2.1.3 | Cell Culture

The following cell culture lines were used in this study. European Collection of Authenticated Cell Cultures (ECACC) catalog numbers are listed for each and the formulations of media for each follow.

Cell Line	ECACC Cat #
MDA-MB-231	92020424
MCF-7	86012803
SDM	11120712
ZR-75-1	87012601

Medium formulations for each cell line:

ZR-75-1:

- RPMI 1640
- 10 % FBS
- 2 mM Glutamine
- Antibiotics (penicillin/streptomycin)

MCF-7:

- EMEM
- 10 % FBS
- 2 mM Glutamine
- Antibiotics (penicillin/streptomycin)

MDA-MB-231:

- L-15
- 15 % FBS
- 2 mM Glutamine
- Antibiotics (penicillin/streptomycin)

SDM103T2:

- DMEM:HAMS F12 (1:1)
- 15 % FBS
- 2 mM Glutamine
- Antibiotics (penicillin/streptomycin)

2.1.4 | Prepared solutions

0.9% NaCl, 1000 ml

- 9 g NaCl
- 1000 ml MilliQ water

1X PBS, 1000 ml

- 2 PBS tablets
- 1000 ml MilliQ water

Staining Buffer, 50 ml 0.5% BSA, 2 mM EDTA

- 1X PBS to 50 ml

- 0.25 g BSA
- 1 ml 100 mM EDTA

Isolation Buffer 0.1% BSA, 2 mM EDTA

- 1X PBS to 50 ml
- 10 ml Staining Buffer
- 800 μl 100 mM EDTA

RLT + β-mercaptoethanol (1% v/v)

- 50 ml RLT BUffer
- 500 μl β-mercaptoethanol

2.1.5 | Kits

Kit	Manufacturer	Catalog Number	Use
QIAGEN AllPrep DNA/RNA/Protein Mini Kit	QIAGEN	80004	RNA/DNA purification
SSIV First Strand	Invitrogen	18091050	Reverse transcription
High Capacity	Applied Biosystems	4368814	Reverse transcription
Ampliseq Cancer HotSpot Panel v2	Life Technologies	4475346	Sequencing
Ion Xpress Barcode Adapters	Life Technologies	4471250	Sequencing
Qubit dsDNA HS Assay Kit	ThermoFisher	Q32851/54	Sequencing
Ion PI Template OT2 200 Kit v3	Life Technologies	4488318	Sequencing
Ion PI Sequencing Kit v3	Life Technologies	4488315	Sequencing

2.1.6 | Primers and probes for PCR

Taqman Gene Expression Assays were used for pre-amplification of targets and also for quantitative measurement of targets. The Taqman assays used are described in Table 2.3. KRT19L is a custom designed assay that spans exon boundaries that will not amplify genomic DNA (F-GATGAGCAGGTCCGAGGTTACT, R-TCTTCCAAGGCAGCTTTCATG, probe-TTCAGGGTCTTGAGATTG). All other assays are ready-to-order from ThermoFisher. All assay with _m1 in the assay ID span exon boundaries and will not amplify genomic DNA. Any assay IDs containing _g1 span exon boundaries but may amplify genomic DNA.

2.1.7 | Reagents

See Table 2.4.

TABLE 2.3: Taqman Gene Expression Assays. *denotes use of assay in gene expression measurement of PBCB samples.[6]

Gene Symbol	Gene Aliases	Gene Name	Taqman Assay ID	Amplicon Length
*BCR**	*ALL, BCR1, CML, D22S11, D22S662, PHL*	breakpoint cluster region	Hs01036532_m1	112 bp
*CCDC80**	*DRO1, HBE245, SSG1, URB*, okuribin	coiled-coil domain containing 80	Hs00277341_m1	69 bp
*EPCAM**	*DIAR5, EGP-2, EGP314, EGP40, ESA, HNPCC8, KS1/4, KSA, M4S1, MIC18, MK-1, TACSTD1, TROP1*	epithelial cell adhesion molecule	Hs00158980_m1	64 bp
*ERBB2**	*CD340, HER-2*, H*ER-2/neu, HER2, MLN 19, NEU, NGL, TKR1*	v-erb-b2 avian erythroblastic leukemia viral oncogene homolog 2	Hs01001580_m1	60 bp
KRT7	*CK7, K2C7, K7, SCL*	keratin 7	Hs00559840_m1	95 bp
*KRT8**	*CARD2, CK-8, CK8, CYK8, K2C8, K8, KO*	keratin 8	Hs01595539_g1	164 bp
KRT16	*CK16, FNEPPK, K16, K1CP, KRT16A, NEPPK*	keratin 16	Hs04194235_g1	94 bp
*KRT19L**	*CK19, K19, K1CS*	keratin 19	AI70M8O	96 bp
SERPINH1	*AsTP3, CBP1, CBP2, HSP47, OI10, PIG14, PPROM, RA-A47, SERPINH2, gp46*	serpin peptidase inhibitor, clade H (heat shock protein 47), member 1, (collagen binding protein 1)	Hs01060397_g1	114 bp
*SCGB2A2**	*MGB1, UGB2*	secretoglobin, family 2A, member 2 (mammoglobin)	Hs00935948_m1	107 bp
*SNAI1**	*SLUGH2, SNA, SNAH, SNAIL, SNAIL1, dJ710H13.1*	snail family zinc finger 1	Hs00195591_m1	66 bp
*SNAI2**	*SLUG, SLUGH1, SNAIL2, WS2D*	snail family zinc finger 2	Hs00950344_m1	86 bp
TFF3	*ITF, P1B, TFI*	trefoil factor 3	Hs00902278_m1	64 bp
*TWIST1**	*ACS3, BPES2, BPES3, CRS1, SCS, TWIST, bHLHa38*	twist basic helix-loop-helix transcription factor 1	Hs00361186_m1	115 bp
ZEB1	*AREB6, BZP, DELTAEF1, FECD6, NIL2A, PPCD3, RP11-472N13.4, TCF8, ZFHEP, ZFHX1A*	zinc finger E-box binding homeobox 1	Hs00232783_m1	63 bp

Table 2.4: Reagents used in experiments

Material	**Manufacturer**	**Catalog Number**	**Use**
DMEM / HAMS F12	Sigma	D6421	Cell culture
EMEM	Sigma	M2279	Cell culture
L-Glutamine 200 mM	Sigma	G7513	Cell culture
L15	Sigma	L5520-500ML	Cell culture
PBS	Sigma	D8537	Cell culture
RPMI-1640	Sigma	R0883-500ML	Cell culture
0.25% Trypsin-EDTA	Sigma	T4049	Cell culture
Fetal Bovine Serum	Sigma	F7524	Cell culture / CTC enrichment
Bovine Serum Albumin	Sigma	A7030-10G	CTC enrichment
CD235a (GYPA)	eBioscience	13-9987-80	CTC enrichment
CD16	eBioscience	13-0168-80	CTC enrichment
CD163	eBioscience	13-1639-82	CTC enrichment
CD19	eBioscience	13-0199-82	CTC enrichment
Dynabeads	Invitrogen	11308D	CTC enrichment
EDTA	Merck	1.08418.0250	CTC enrichment
Human CD45 – Biotin Conjugated	Life Technologies	MHCD4515	CTC enrichment
Lymphoprep	Axis Shield	1114545	CTC enrichment
Phosphate Buffered Saline Tablets	Sigma	P4417-100TAB	CTC enrichment
RLT Buffer	QIAGEN	79216	CTC enrichment
SepMate Tubes 50 mL	StemCell Technology	15450	CTC enrichment
Sodium Chloride	Sigma	S3014-1KG	CTC enrichment
Trypan Blue 0.4\%	Sigma	T8174	CTC enrichment
CD236 EpCAM FITC	Miltenyi Biotech	130-098-113	Flow cytometry
CD45 APC	Miltenyi Biotech	130-098-143	Flow cytometry
CytoFlex Daily QC	Beckman Coulter	B53230	Flow cytometry
FcR Blocking Reagent	Miltenyi Biotech	130-059-901	Flow cytometry
Taqman Pre-Amplification Master Mix	Applied Biosystems	4369016	Pre-amplication of templates
Taqman Gene Expression Master Mix	Applied Biosystems	4369016	qPCR
0.1 M DTT	Invitrogen	Y00147	Reverse transcription
5X FS Buffer	Invitrogen	Y0232T	Reverse transcription
dATP	GE Healthcare Life Sciences	28406501U	Reverse transcription
dCTP	GE Healthcare Life Sciences	28406511	Reverse transcription
dGTP	GE Healthcare Life Sciences	28406521	Reverse transcription
dTTP	GE Healthcare Life Sciences	28406531	Reverse transcription
M-MLV 200 U/μl	Invitrogen	28025-013	Reverse transcription
Random Primers	Invitrogen	58878	Reverse transcription
Rnase OUT	Invitrogen	10777-019	Reverse transcription
RQ1 Dnase	Promega	M610A	Reverse transcription
RQ1 Stop Solution	Promega	M199A	Reverse transcription
QIAshredder	QIAGEN	79656	Cell lysate homogenization

2.2 | Methods

2.2.1 | Cell Culture

Four cell lines were used in this study (see Table 2.1.3 in Materials). The expression of genes in the prospective mRNA panel was investigated to find the best calibrator cell candidate, but also to establish which panel markers would be well suited for further analysis. In addition, ZR-75-1 was used in spiking experiments to measure recovery rate of the enrichment method. These and SDM103T2 cells were also used for spiking experiments to measure qPCR sensitivity of various gene expression assays.

2.2.1.1 | Aseptic Technique

All the following techniques were performed according to aseptic technique. They were done exlusively in a dedicated cell culture room, negatively pressurized relative to the adjoined staging area, and requiring use of gowns and shoe covers for further protection. Hands were thoroughly washed prior to wearing gloves and gloves were also sterilized with an 70% ethanol solution. All work was performed in a laminar flow hood which was sterilized before and after use with the ethanol solution and UV decontamination. Reagents, media, bottles, and solutions to be used in the hood were sterilized prior to placing them inside. Items were handled carefully and mindfully to avoid contamination from any non-sterile surface or cross-contamination between any reagants. Any spills were immediately cleaned and sterilized. Cultures were also closely monitored for any macroscopic and microscopic signs of bacterial contamination.

2.2.1.2 | Resuscitation of frozen culture

The cryotube containing the cell stock was quickly thawed in a 37℃ water bath (about 2 minutes). The contents of the tube were then mixed with warm media and transferred to a T25 flask for a total volume of 12 ml. The flask was then incubated at 37℃ and at 5% CO_2 (no CO_2 for MDA-MB-231 cell line).

2.2.1.3 | Subculturing

Media was carefully removed from the flask to be subcultured, to ensure the adherent cells were undisturbed. Warm 1X PBS (1-3 mL) was added to rinse any serum from the cells (as serum inactivates trypsin) and removed. Warm trypsin was then added and the flask was incubated at 37℃ for 3-5 minutes. The flask was ready when the cells were loosened when rocked. For some highly adherent cultures, this took more than 5 minutes and was checked every minute past for detachment. The cells were then collected by

adding warm, fresh media in a volume at least equal to the amount of trypsin used and an easily divisible volume for splitting. It was thoroughly mixed by pipetting up and down several times to ensure a suspension of single-cells. The cell suspension was then transferred to a new flask containing fresh media. The volume transferred depended on the split size, and this size in turn depended on the rate of growth of the cell culture being split.

2.2.1.4 | Harvest and counting of cells

Cells were harvested when the confluence was at least 70%. The cells were rinsed with warm 1X PBS and then with 1-3 mL warm trypsin to detach the cells. The cells were incubated with the trypsin for 3-5 minutes. Fresh media was added to flask to collect and resuspend the cells and the cells were subsequently counted. Cells were mixed 1:1 with Trypan blue (50 μl each) and counted using a Bürker counting chamber (Figure 2.1). To count cells using the chamber, 20 μl of the cell suspension/Trypan blue mix was pipetted under the slide cover of the chamber. If possible, 200 cells were counted for each sample and then divided by the number of squares taken to reach 200 for the average cells per square. The following equation was used to calculate cells per ml:

$$\text{Cells per ml} = \text{Average cells per } 4x4 \text{ (1 mm) square} \times \text{dilution factor} \times 10{,}000$$

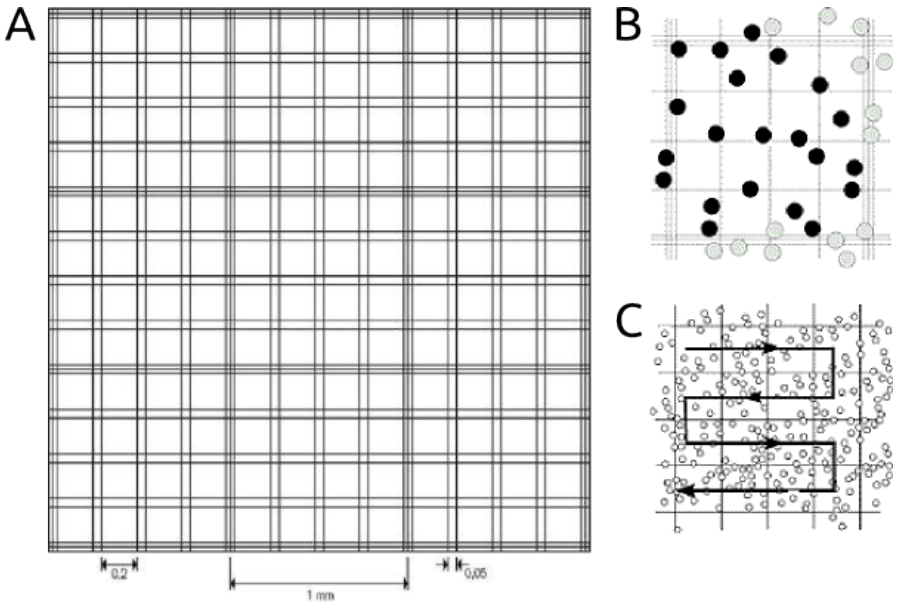

FIGURE 2.1: Counting cells with Bürker counting chamber. (A) Layout and dimensions of the Bürker counting chamber (Image: Sigma-Aldrich) (B) and (C) demonstrate counting methods (Images: Laboroptik) (B) To avoid recounting, cells touching the lines are counted for the current square if they touch the top and left boundaries. (C) General order of counting to prevent missing squares.

The cells were then used directly for spiking experiments or stored for RNA/DNA extraction. For later extraction of nucleic acids, the cell suspension was centrifuged (10

min, 200xg)and the pellet was resuspended in RLT+β-mercaptoethanol) at a volume of 600 μl per 1×10^6 cells. The lysate was then stored at -80°C until further processing.

2.2.2 | Flow Cytometry

For analysis of cell populations by FACS after enrichment, the final resuspension step in the enrichment procedure was replaced with resuspension in 100 μl staining buffer. To this resuspension, 25 μl FcR blocking reagent (Miltenyi Biotech) and 2.5 μl of each stain (EpCAM-FITC and CD45-APC, Miltenyi Biotech) was added. The samples were incubated in darkness at room temperature for 20 minutes and subsequently washed with 1 ml staining buffer and centrifuged. Finally, the pellet was resuspended in 500 μl staining buffer.

The prepared samples were then analyzed on the flow cytometer (CytoFLEX, Beckman Coulter Inc.). Daily QC was performed before analysis. All samples were recorded for 100 seconds at a flow rate of 30 μl per minute, resulting in a total analysis volume of 50 μl. Between test samples and the control, a flush of sheath fluid was run to remove any residual cells. Selection of populations was done by comparison to controls.

2.2.3 | Collection of Blood Samples

All blood samples and clinical information were obtained with informed consent from patients and healthy donors. Patient and volunteer samples were gathered as a part of the Prospective Breast Cancer Biobank (PBCB) project, with approval from the Regional Committee for Medical and Health Research Ethics (REK) (reference: 2015/2010/REK vest). A peripheral blood volume of 9 ml was collected from the antecubital vein in Vacuette EDTA tubes under sterile conditions. The blood was obtained in the middle of the venipuncture with the first few milliliters discarded to avoid epithelial contamination. Blood samples were enriched for CTCs on the same day as collection.

The PBCB samples used in this study were collected from the period of February 2015 to February 2016, starting from patient ID 154 Visit 1 (V1). Baseline samples (V1) were taken prior to surgery. Some samples were from the same patient over multiple visits (Table 2.1). However, due to the short timeline of this study this is only a small portion of the samples. There are no samples that cover three visits from one patient at this point. The Visit 2 (V2) samples consist of patients that were included in the biobank previous to the new enrichment method. Thus, the first timepoint for analysis of these patients is Visit 1.5 (V1.5).

The clinical data analyzed and presented here reflects clinicopathological status at the first visit (Table 2.1). Data analysis was performed blind to patient clinicopathological

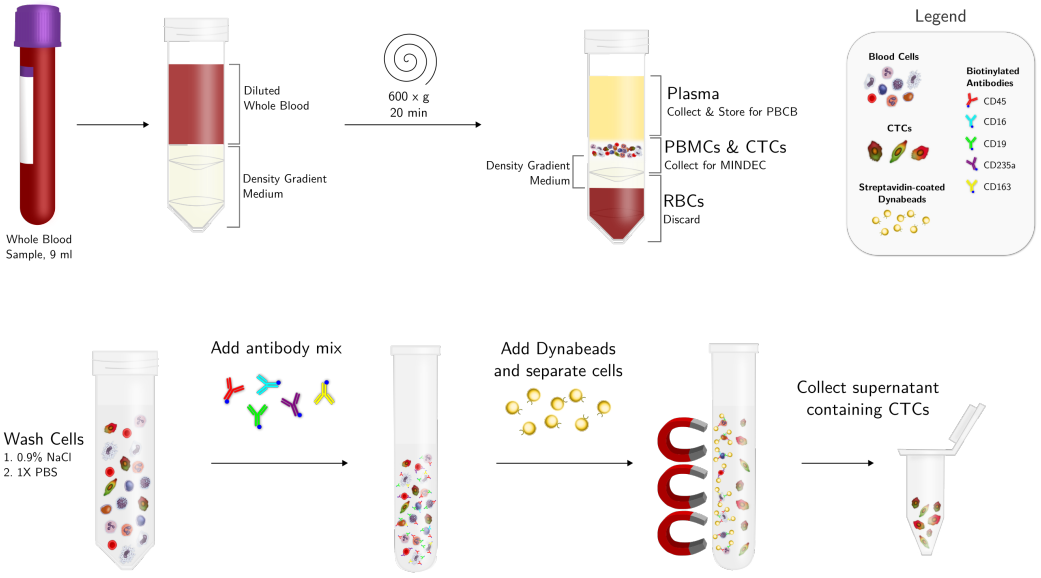

FIGURE 2.2: CTC enrichment workflow.

status. In addition to the PBCB samples, 30 control blood samples were collected from healthy donors.

2.2.4 | CTC Enrichment

CTCs were enriched from peripheral blood by a combination of density gradient centrifugation and multimarker immunomagnetic negative depletion enrichment of CTCs (MINDEC) (Figure 2.2). This specific method was developed by a current Ph.D. student in the lab, Morten Lapin [87]. In it, cells covered in biotin-conjugated antibodies are negatively selected by streptavidin-coated magnetic beads. The cells bound to the magnetic beads are immobilized to the walls of the tube by magnetic racks, leaving the supernatant with free cells (containing CTCs) available for collection. The detailed procedure follows.

2.2.4.1 | Removal of Erythrocytes by Density Gradient

Erythrocytes were removed from the whole blood sample by density gradient. The whole blood sample was mixed 1:1 with 0.9% sodium chloride and overlayed on 15 ml density gradient media (Lymphoprep) in a 50 ml SepMate tube. After centrifugation at $600xg$ (20℃, 20 minutes if within 2 hours of sampling and 30 minutes thereafter, brake off), the top plasma layer was collected and stored (-80℃). The remaining fluid above the SepMate filter consisted of residual plasma and the buffy coat, containing peripheral blood mononuclear cells (PBMCs). This was poured into a new 50 ml tube. This fraction was then washed with cold 0.9% sodium chloride (40 ml), centrifuged (10 min, $200xg$, 4℃), and the pellet was washed by resuspension in PBS. Finally, centrifugation

TABLE 2.5: Antibodies used for negative enrichment of leukocytes. Volume per $1x10^7$ cells.

Antibody	**Volume**	**Concentration**	**Target**
CD45	4 μl	unknown	leukocytes
CD16	4 μl	0.5 mg/ml	NK cells, monocytes, macrophages
CD19	2 μl	0.5 mg/ml	B-lymphocytes
CD163	1 μl	0.5 mg/ml	monocytes, macrophages
CD235a (GYPA)	4 μl	0.5 mg/ml	erythrocytes
Total	15 μl		

was repeated and the pellet was resuspended in 1 ml Isolation Buffer (section 2.1.4). An aliquot (5 μl) of the suspension was stained (1:20 dilution in Trypan Blue) and counted using a Bürker counting chamber (section 2.2.1.4).

2.2.4.2 | MINDEC: Immunomagnetic depletion of leukocytes

The resuspended cells were centrifuged (10 min, 200xg, 4°C) and resuspended in 100 μl isolation buffer (all volumes are adjusted to the concentration of PBMCs, and unless otherwise noted were per $1x10^7$ cells). A mix was prepared using biotin-conjugated antibodies and added to the cell suspension (Table 2.5). The antibody mix and cell suspension were mixed thoroughly by pipetting up and down with a P-100 pipet set to full volume.

The suspension and antibody mix were incubated at 4°C for 20 minutes and subsequently mixed with 2 ml isolation buffer and centrifuged (10 min, 200xg, 4°C). The pellet was resuspended in 900 μl isolation buffer and 100 μl Dynabeads (pre-washed and buffer exchanged to isolation buffer) were added. This mixture was incubated at 4°C in a tube inverting instrument (HulaMixer, Invitrogen) for 15 minutes. Isolation buffer was added and the solution was placed in a magnetic rack (Dynamag, Life Technologies) for 3 minutes to collect the bound cells and beads on the side of the tube and leave unbound cells in the supernatant. This was performed twice with the supernatant from each step being pooled and a third magnetic incubation performed on the pooled supernatant. Finally, this supernatant was collected and centrifuged in a 15 ml conical tube (10 min, 200xg rpm, 4°C) and the pellet was resuspended in 350 μl RLT+β-mercaptoethanol and stored at -80°C.

2.2.5 | RNA/DNA Extraction

Extraction of RNA and genomic DNA was performed by following the QIAGEN AllPrep DNA/RNA/Protein Mini Kit protocol [105] for purification from animal or human cells (protein was not saved from these samples). All samples going through extraction in

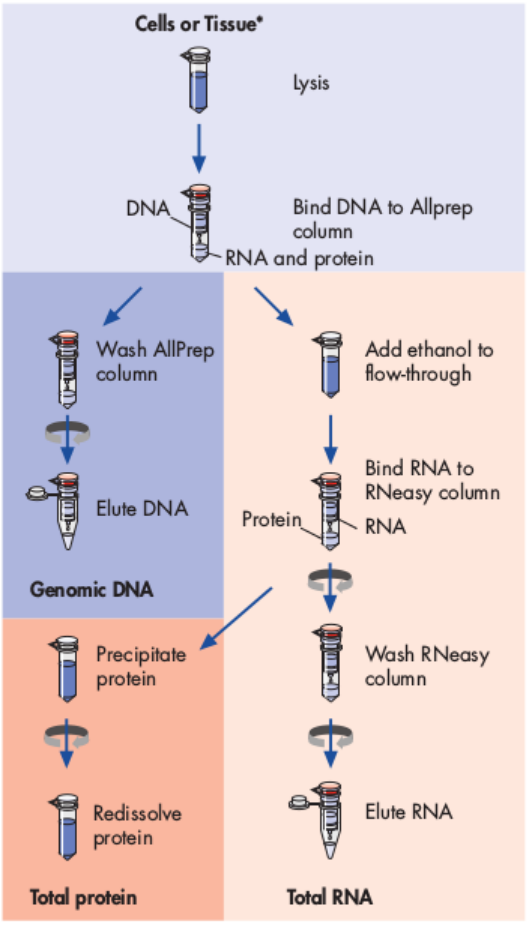

FIGURE 2.3: QIAGEN Allprep DNA/RNA/Protein Mini Kit workflow [105].

this project were cell samples and were lysed with RLT+β-mercaptoethanol and homogenized using the QIAshredder homogenizer columns. The RLT lysate was first thawed and transferred to the QIAshredder column, then centrifuged at maximum speed for 2 minutes. The flow-through was transferred to the AllPrep DNA spin column and nulceic acid purification was performed as described by the manufacturer and summarized in Figure 2.3.

2.2.5.1 | Purification of Genomic DNA

DNA was eluted from the Allprep column in 100 μl of EB buffer (two separate elutions in 50 μl EB buffer). The DNA was then frozen at -80℃ for later use (non-PBCB samples stored at -20℃).

2.2.5.2 | Purification of Total RNA

RNA was eluted from the RNeasy column with RNase-free water (MilliQ). For the cell line and breast tissue samples, this was completed in two elutions with 50 μl for 100

μl total. For MINDEC samples, only one elution with 30 μl RNAse-free water was performed to have as high concentration as possible. The RNA was then stored at -80℃ (non-PBCB samples stored at -20℃).

2.2.5.3 | Nucleic acid quantification

Before freezing, RNA and DNA purifications were measured for concentration and quality with the NanoDrop 2000c (ThermoScientific). MINDEC samples were consistently below the detectable limit (2 ng/μl [106]) due to low quantity of cells and therefore low amounts of nucleic acids. It was not standard practice in the study to analyse these samples by NanoDrop. All other samples were analyzed however and their concentrations were measured and given. Ratios of A_{260}/A_{280} and A_{260}/A_{230} were recorded as well. An A_{260}/A_{280} ratio of of 1.8 is considered pure for DNA while a ratio of 2.0 is considered pure for RNA. An A_{260}/A_{230} ratio of 1.8-2.2 reflects acceptable quality. A much lower number could mean that there are contaminants present in the sample [106].

2.2.6 | cDNA Synthesis

The purified RNA from the samples were reverse transcribed to produce cDNA for later pre-amplification and quantification. Three methods were performed with different samples. The M-MLV method was the established lab protocol and was performed for the initial testing of cell line expression, breast tissues, and the first normal controls. This method included a DNase treatment. It has been shown that DNase treatment before pre-amplification is unnecessary [107], so this was not included in the other methods. Additional methods included The High Capacity cDNA Synthesis Kit and SuperScript IV First-Strand Synthesis System (SSIV), which were tested for compatibility with the Taqman Gene Expression qPCR. The High Capacity cDNA Synthesis Kit was chosen from the two and thus used for the sensitivity samples, calibrator cell cDNA synthesis, and all PBCB patient/control samples.

2.2.6.1 | M-MLV Method

DNAse Treatment Total volume of the DNase treatment reaction was 10 μl. This consisted of 5X First Strand Synthesis Buffer (FSS, 2 μl), RQ1 DNase (1 μl), and RNase OUT RNAse inhibitor (0.25 μl). For cell line samples, 1 ug of RNA and MilliQ water was added to a total of 7 μl. For MINDEC samples (patient and controls, validation experiments) the maximum volume of the RNA sample (7 μl) was used. This mix was then incubated at 37℃ for 30 minutes. Immediately after, 1 μl of RQ1 stop solution was added and incubated at 65℃ for 10 minutes.

Reverse Transcription To each DNAse-treated reaction, 0.2 μl of Random Primers (Invitrogen), 0.4 μl of 25 mM dNTPs (GE Healthcare Life Sciences), and 0.4 μl milliQ water was added. This was then incubated at 65℃ for 5 minutes followed by at least 2 minutes on ice. Next added to the mix was 5X FSS (2 μl), 0.1 M DTT (2 μl), RNase OUT (1 μl), and MilliQ H2O (2 μl). After a 3 minute incubation at 37℃, 1 μl of M-MLV reverse trascriptase (Sigma) was added to each reaction. To one tube, water was added in place of M-MLV as a no-enzyme control (NEC). The samples were incubated in the flow hood at room temperature for 10 minutes, followed by 1 hour at 37℃, and a final incubation at 65℃ for 15 minutes to inactivate the enzyme. MINDEC samples were then stored at the final reaction volume of 20 μl, while the rest of the samples were diluted to 10 ng/μl by the addition of 80 μl MilliQ water. These samples were stored at -20℃.

2.2.6.2 | High Capacity cDNA Synthesis Kit

The protocol for the High Capacity cDNA Synthesis Kit was followed for this procedure. The final reverse transcribed samples were diluted to 10 ng/μl with the addition of 80 μl water and stored at -20℃. MINDEC samples were not diluted and they were stored at -80℃.

2.2.6.3 | SSIV Kit

The protocol for the SSIV system was followed for this procedure. Reverse transcribed samples were diluted to 10 ng/μl with 80 μl MilliQ water and stored at -20℃. This protocol was not used with any MINDEC samples.

2.2.7 | Gene expression analysis

With the small amount of cells collected in the MINDEC procedure and the resulting low concentration of total RNA, pre-amplification was necessary in order to quantify gene expression of multiple transcripts. Taqman Gene Expression Assays were used for pre-amplification of targets and also for quantitative measurement of targets. The Taqman assays used are described in Table 2.3 (see section 2.1.6). This is a list of all assays considered at the preliminary stage. The genes marked with an asterisk are assays that were used in the final gene expression measurements of PBCB samples.

2.2.7.1 | Pre-Amplification

The TaqMan PreAmp Master Mix was used for pre-amplification of the cDNA and the manual followed. All assays were pooled and diluted to a total 100X in TE buffer.

TABLE 2.6: Pre-Amplification Thermocycler Settings. cDNA volume varies: maximum volume used with MINDEC samples and volume to reach 1 μg for others.

Step		**Temperature**	**Time**
Taq Enzyme Activation		95℃	10 min
Pre-Amplification	Denature	95℃	15 s
14 cycles	Anneal	60℃	4 min

TABLE 2.7: PCR reaction mix reagents and volumes for 96- and 384-well plates.

Reagents	**96-well**	**384-well**
Taqman Gene Expression Master Mix (2X)	10.0 μl	5.0 μl
Taqman Gene Expression Assay (20X)	1.0 μl	0.5 μl
cDNA template	varies	varies
nuclease-free H2O	up to 9 μl	up to 4.5 μl
Total	20 μl	10 μl

To each pre-amplification reaction the following was added: 25 μl of pre-amplification Master Mix, 12.5 μl of the pooled Taqman assays, 2.5 μl MilliQ water, and 10 μl of the cDNA sample to be pre-amplified. The samples were then pre-amplified in the thermocycler with the program in Table 2.6. After pre-amplification, 950 μl of MilliQ water was added to each reaction for a 1:20 dilution. The pre-amplified samples were then stored at -20℃.

2.2.7.2 | Real-time quantitative PCR

TaqMan Gene Expression Assays include both the primers and hydrolysis probes in one. Quantification relies on Taq DNA polymerase cleaving a dual-labeled fluorescent probe for detection by the qPCR instrument. For quantification of the pre-amplified DNA, the TaqMan Gene Expression Assays Protocol was followed. The measurement of gene expression of the targets was done by using the same general protocol for each experiment. Changes between runs included the samples being analyzed, assays used, and plate size/layout.

Each assay target was amplified using the master mix as described in Table 2.7. On all plates, the calibrator cell and no-template control (NTC) were run. The no-enzyme control (NEC) for each batch of reverse transcription was also run when necessary to ensure no amplification of residual genomic DNA. After plates were loaded, they were sealed with foil and centrifuged for 1 minute to collect all liquid in the bottom of the wells. The sample and run information was entered into the LightCycler 480 software. The cycling conditions of the PCR program are shown in Table 2.8.

TABLE 2.8: Real-time PCR Program Settings

Step		**Temperature**	**Time**
UGD Activation		50℃	2 min
Taq Enzyme Activation		95℃	10 min
Amplification	Denature	95℃	15 s
40 cycles	Anneal	60℃	1 min

2.2.8 | Amplification Efficiency

To test amplification efficiency, cDNA was diluted in a series of 4-fold dilutions: undilute, 1/4, 1/16, 1/64, 1/256 (10 ng, 2.5 ng, 0.625 ng, 0.15625 ng, 0.0391 ng). The samples were all run in triplicates on 384-well plates. To evaluate assay efficiency, the cell line for each assay was chosen based on the highest expressing cell line for that particular assay. The cell line cDNA was then diluted as described for qPCR. Efficiency was measured for all assays (n=15). To measure pre-amplification efficiency, dilutions of the calibrator cell cDNA were made prior to pre-amplification. The samples were then pre-amplified and run on the LC480 in triplicates. Three assays were used as to analyze the pre-amplification efficiency: *BCR*, *KRT8*, and *KRT19*.

To calculate efficiency, a standard curve was produced from the resulting Cq values plotted against log concentration. The slope of this curve was used in the following equations:

$$E = 10^{\frac{-1}{slope}}$$

For the percent efficiency:

$$\%E = (10^{\frac{-1}{slope}} - 1)\times 100$$

2.2.9 | Next Generation Sequencing

An evaluation of a sequencing method was performed on samples using the Ion Proton and a targeted gene panel. The workflow of the entire process is summarized in Figure 2.5. The first step requires the creation of a library from the genomic DNA by amplification of targeted regions and subsequent purification using magnetic beads. Next, emulsion PCR is performed to obtain clonal amplification of specific templates on individual Ion Sphere Particles (ISPs). The template-positive ISPs are enriched with the Ion One Touch and magnetic beads. Finally, the templates are sequenced on the Ion Proton, which calls bases by detecting the pH change upon each base addition. Kits used for this procedure are listed in section 2.1.5.

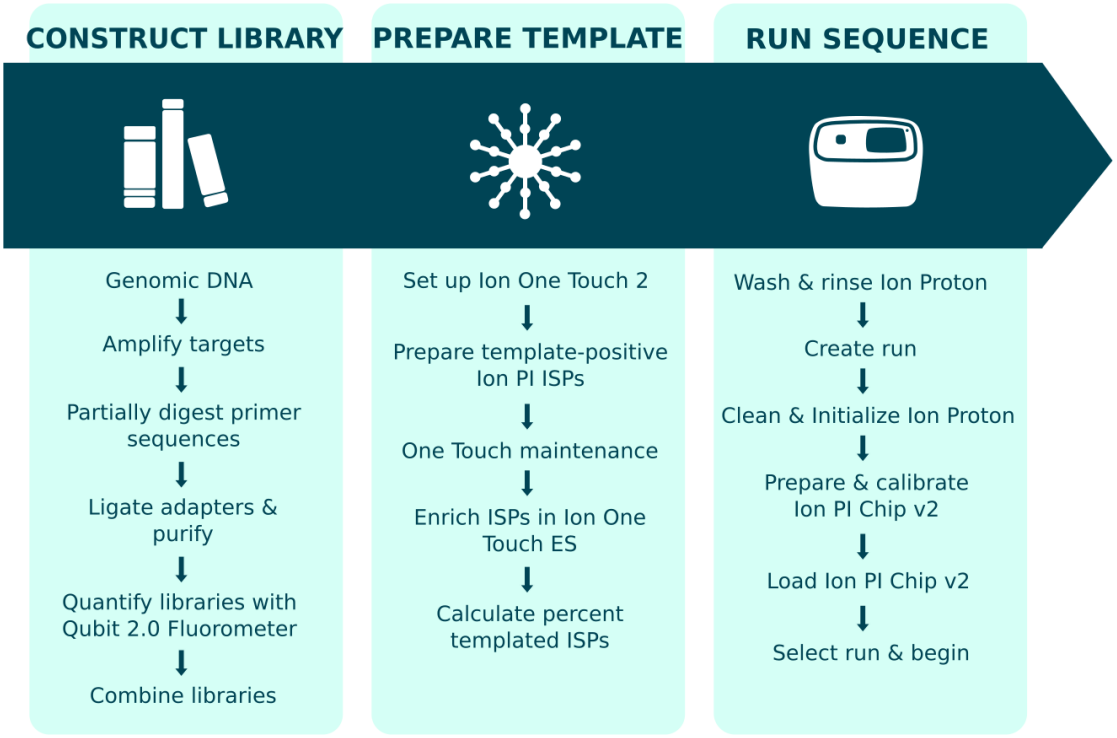

FIGURE 2.4: Next Generation Sequencing workflow [108].

TABLE 2.9: Cancer Hotspot Panel v2 Gene Coverage. 50 target genes with a total of 207 amplified regions [109]. Listed as: *GENE* (number of amplicons)

ABL1 (4)	*EZH2* (1)	*JAK2* (1)	*PTEN* (8)
AKT1 (2)	*FBXW7* (5)	*JAK3* (3)	*PTPN11* (2)
ALK (2)	*FGFR1* (2)	*KDR* (9)	*RB1* (10)
APC (7)	*FGFR2* (4)	*KIT* (9)	*RET* (5)
ATM (17)	*FGFR3* (5)	*KRAS* (3)	*SMAD4* (9)
BRAF (2)	*FLT3* (4)	*MET* (6)	*SMARCB1* (4)
CDH1 (3)	*GNA11* (1)	*MLH1* (1)	*SMO* (5)
CDKN2A (2)	*GNAQ* (1)	*MPL* (1)	*SRC* (1)
CSF1R (2)	*GNAS* (2)	*NOTCH1* (3)	*STK11* (5)
CTNNB1 (1)	*HNF1A* (2)	*NDM1* (1)	*TP53* (8)
EGFR (8)	*HRAS* (2)	*NRAS* (3)	*VHL* (3)
ERBB2 (3)	*IDH1* (1)	*PDGFRA* (4)	
ERBB4 (8)	*IDH2* (1)	*PIK3CA* (11)	

2.2.9.1 | Library Construction

The construction of the template library was performed following the Ampliseq Library Preparation User Guide. The Cancer Hotspot Panel was used to create the library, which amplifies gene regions commonly mutated in various cancers. The panel covers 50 genes with a total of 207 primer pairs (Table. 2.9)

Target Amplification A cell line sample and two spiked normal blood samples were evaluated in this experiment. A 1:10 dilution of cell line cDNA was made and 10 ng added to the target mix. 12 μl (maximum volume) of each the spiked samples was added. To prepare the libraries for amplification the following was added to each DNA template: 5X Ion AmpliSeq HiFi Mix (4 μl), 5X Ion Ampliseq Primer Pool (4 μl), and nuclease-free water to a total of 20 μl. The samples were mixed by vortexing and subsequently spun down and were amplified in the thermocycler.

Digest Primer Sequences The primer sequences were partially digested by adding 2 μl of FuPa Reagent to a total reaction colume of 22 μl. The samples were mixed, spun down, and again placed in the thermocycler.

Ligate Adapters & Amplify Since multiple libraries were prepared and would be run on a single chip, a unique barcoded adapter was used for each library. 2 μl of the 1:4 barcode adapter mix and 4 μl Switch Solution was added to each digested amplicon library. DNA Ligase (2 μl) was then added to each sample and the samples were placed in the thermocycler for ligation.

Purify Libraries For purification of the libraries, 45 μl of Agencourt AMPure XP Reagent was added to each, incubated for 5 minutes at room temperature, and subsequently placed in a magnetic rack. To wash the beads, 70% ethanol was added to each tube, removed and the tube incubated for 5 minutes to dry. Immediately following drying, the libraries were amplified for quantification.

Library Amplification, Purification and Quantification Libraries were amplified by first adding Platinum PCR Supermix High Fidelity (50 μl) and Library Amplification Primer Mix (2 μl) to each bead pellet and replaced in the magnet for 2 minutes. The sample libraries were placed in the thermocycler and following amplification, the libraries were purified by AmpPure beads with Agencourt AMPure XP Reagent. The first round was at a 0.5X bead-to-sample ratio for the removal of any residual high molecular-weight DNA. The second round was at a 1.2X bead-to-original-sample-volume ratio. Here the amplicons bind to the beads and primers are left in solution. The bead pellet was saved and the amplicons were eluted from the beads.

Library concentrations were measured on the Qubit 2.0 Fluorometer (Life Technologies). A fresh 1:200 working dilution of the Qubit dsDNA HS reagent was prepared. Each amplified library aliquot was combined with 190 μl of dye reagent and incubated for 2 minutes.

Combine Libraries After quantification, the libraries were diluted accordingly (to 15 ng/ml) for an even mix of 3.5 μl each for template preparation. Nuclease-free water was added to a final volume of 100 μl for template-positive ISP preparation.

2.2.9.2 | Template Preparation

Template-positive ISPs were prepared by emulsion PCR for clonally amplified DNA, and subsequent enrichment of the template-positive particles using the Ion One Touch system. The Ion PI Template OT2 200 guide was followed for this technique. After emulsion PCR and before enrichment of template-positive ISPs, the percent of templated ISPs were measured in the Qubit fluorometer. The ISP sample was measured by inserting the sample into the Qubit and under the Ion option, AF 488 was selected. The value was recorded and then the AF 647 fluorescence was measured and recorded. These values were entered into a spreadsheet containing the factor calculator to determine the percent of templated ISPs in the unenriched sample. The sample was then enriched for template-positive ISPs and stored at 4°C for the sequencing run.

2.2.9.3 | Run Sequence

For the sequencing chip preparation and run, the Ion PI Sequencing 200 guide was followed. First, the instrument and was initialized to obtain the proper pH in the sequencing. The chip was also prepared with multiple washes and calibrated by the instrument to ensure the correct pH was present in the chip. The chip was then loaded; 55 μl of the ISP solution was foamed and injected into the chip. Further dispersion of the ISP foam into the chip wells was performed by centrifugation. The chip was placed in the instrument and the sequencing run was started.

2.2.10 | Data Analysis

2.2.10.1 | Multimarker mRNA Panel

Candidate markers for the multimarker mRNA panel were chosen from previously used markers in literature (see Appendix A) and by searching for new markers in the Cancer Genome Project's (CGAP) serial analysis of gene expression (SAGE) database (http://cgap.nci.nih.gov/SAGE) [110]. Individual queries were made for each relevant marker from literature with the purpose of finding genes highly expressed in normal breast/breast neoplasms as well as low/no expression in white blood cells (WBCs). To search for novel and differentially expressed mRNAs, a representative library was picked for each tissue that contained large numbers of total tags. The library tag data was downloaded and analyzed in R (r-project.org). Markers with a tag frequency over 10000

TABLE 2.10: Thresholds used to calculate Cq values in PBCB pPCR runs.

Assay	STD Multiplier
BCR	19
CCDC80	15
EPCAM	22
ERBB2	12
KRT8	18
KRT19	14
LUM	42
SCGB	22
SLUG	18
SNAIL	17
TWIST	24

in breast cancer tissue and below 25 in WBCs were analyzed for total average tag frequency on the SAGE database. They were added to the prospective list if still promising after review of other libraries' expression. A preliminary list was chosen based on the ratio of tag frequency of neoplasm:WBCs and also scientific relevance. To further pare down the panel to a final 10, the preliminary markers were first tested on cancer cell lines and ultimately on breast cancer tissue samples and four normal control blood samples.

2.2.10.2 | Relative Gene Expression

Cq values were calculated in the LC480 software by using the Abs Quant/Fit Points method. The noise band/threshold was set using the STD Multiplier value for each assay and the same STD Multiplier was used between each target assay plate. The STD Multiplier values used for each assay are listed in Table 2.10. The Cq values were exported from the LC480 software as text files and imported into the R software program (r-project.org) for analysis of PBCB samples (R script in Appendix D). Data from preliminary experiments were analyzed in Excel.

The decision of treatment of non-detected values was based on potential bias of multiple options. For a non-detectable Cq value, there are three possibilities for what the non-detectable value represents:(1)low expression resulting in a Cq > 40, (2) an unexpressed transcript, or (3) a failure to detect a real Cq < 40 [111]. In this case, samples presenting with low expression ($>$ than 40) are not a concern, since the analysis is dependent on highly-expressed transcripts only (5 samples noticed with non-detect among 2 replicates with Cq 37). There were only a few samples that had exhibited a failure to amplify (n=2) and these were present as only one of three replicates. So these could be safely disregarded without great bias to the final mean Cq from the remaining replicates. Non-detects present among all 3 replicates, were considered as too low expressed or unexpressed and are represented in the data as NA.

BeadFind Args	justBeadFind –beadfind-minlivesnr 3 –region-size=216,224 –total-timeout 600
Analysis Args	Analysis –from-beadfind –clonal-filter-bkgmodel true –region-size=216,224 –bkg-bfmask-update false –gpuWorkLoad 1 –total-timeout 600 –gpt /opt/ion/config/gopt$_p$1.1.17$_a$*mpliseq*$_e$*xome.param.json*
Pre-BaseCaller Args for calibration	BaseCaller –barcode-filter 0.01 –barcode-filter-minreads 10 –phasing-residual-filter=2.0 –max-phasing-levels 2
Calibration Args	Calibration
BaseCaller Args	BaseCaller –barcode-filter 0.01 –barcode-filter-minreads 10 –phasing-residual-filter=2.0 –num-unfiltered 1000 –barcode-filter-postpone 1
Alignment Args	tmap mapall ... stage1 map4
IonStats Args	ionstats alignment
Analysis Parameters	default

FIGURE 2.5: Ion Torrent analysis parameters.

A threshold for data quality was not set. Outside of two sample replicate exclusions due to abnormal amplification, all data points were used in analysis. There were cases with some higher variance, and this was due to one replicate of three being divergent, but to avoid introducing any bias, the values were kept and the average of the triplicates were used.

The equation below was used to calculate the relative gene expression of the samples for each assay, according to the 2-ddCt method by Livak *et al.* [112]. With this equation, the samples were normalized to the reference/housekeeping gene, breakpoint cluster region *BCR*, and given as the fold change in expression compared to the calibrator cell. Relative expression of the control values were used as the threshold for determining CTC-positive patient samples. Control outliers were defined as any samples that were greater than 3 standard-deviations from the mean and were removed. Any patient samples that were greater than the maximum control expression for an assay was considered CTC-positive.

$$R = 2^{-(\Delta CP_{target(calibrator-sample)} - \Delta CP_{reference(calibrator-sample)})}$$

2.2.10.3 | Statistical Analysis

Statistical associations were evaluated between the clinicopathological characteristics of the patients and their CTC-status. The patient data was imported into R and analyzed using the "tableone" package (see Appendix D). The continuous variables were tested by the Kruskal-Wallis rank sum test and the categorical variable were tested by the Fisher's exact test.

2.2.10.4 | Next Generation Sequencing

Once the sequencing run was complete, the Torrent system performed its own analysis (Torrent Suite 5.0.3). The parameters of the Ion Proton analysis include those shown in Figure 2.5.

The run was inspected for quality and results. Individual mutations detected were inspected for validity and were compared between samples. Any discrepancies were investigated in the binary sequence alignment and variant call format (BAM and VCF) files in the Integrative Genomics Viewer (IGV).

Chapter 3

Results

A workflow of the methods used in this study are shown in Figure 3.1. These include tests that were required to ensure the validity of techniques, as well as selection of parameters for later methods.

3.1 | Validation of CTC enrichment by flow cytometry

The MINDEC method for CTC enrichment was previously developed with pancreatic cancer samples [87]. We wanted to validate the MINDEC method for the PBCB study by flow cytometry to determine recovery of CTCs and a spiking experiment was therefore performed. Five vials of blood were collected from a healthy volunteer for this purpose. One vial was set aside to be a whole blood reference for flow cytometry. PBMCs from the other four samples were isolated by density centrifugation as described in section 2.2.4. Harvested ZR-75-1 cells (section 2.2.1.4) were used to spike two of the PBMC samples with 10,000 cells. The other two samples were not spiked and used as negative controls. In addition, a positive control was created by adding the same spike volume to staining buffer (section 2.2.2). The PBMC samples were then negatively depleted of leukocytes by MINDEC strategy (section 2.2.4) and subsequently stained for flow cytometry analysis (section 2.2.2).

The samples were analyzed by flow cytometry (section 2.2.2). Selection of populations (gating) was done by comparison to the two control samples (whole blood and spike control). The averaged PBMC number from all four samples was compared to the original cell count and the cancer cell recovery in the two spiked samples was compared to the spike positive control (Figure 3.2). Of 10,000 ZR-75-1 cells spiked into sample, 9720 were recovered in the spike positive control and 7665±2.5 (n=2) were recovered in the enriched, spiked samples. This is a 78.85+-0.36% recovery when compared to the spike control. Of the starting PBMCs, 99.98% of the PBMCs were removed by the

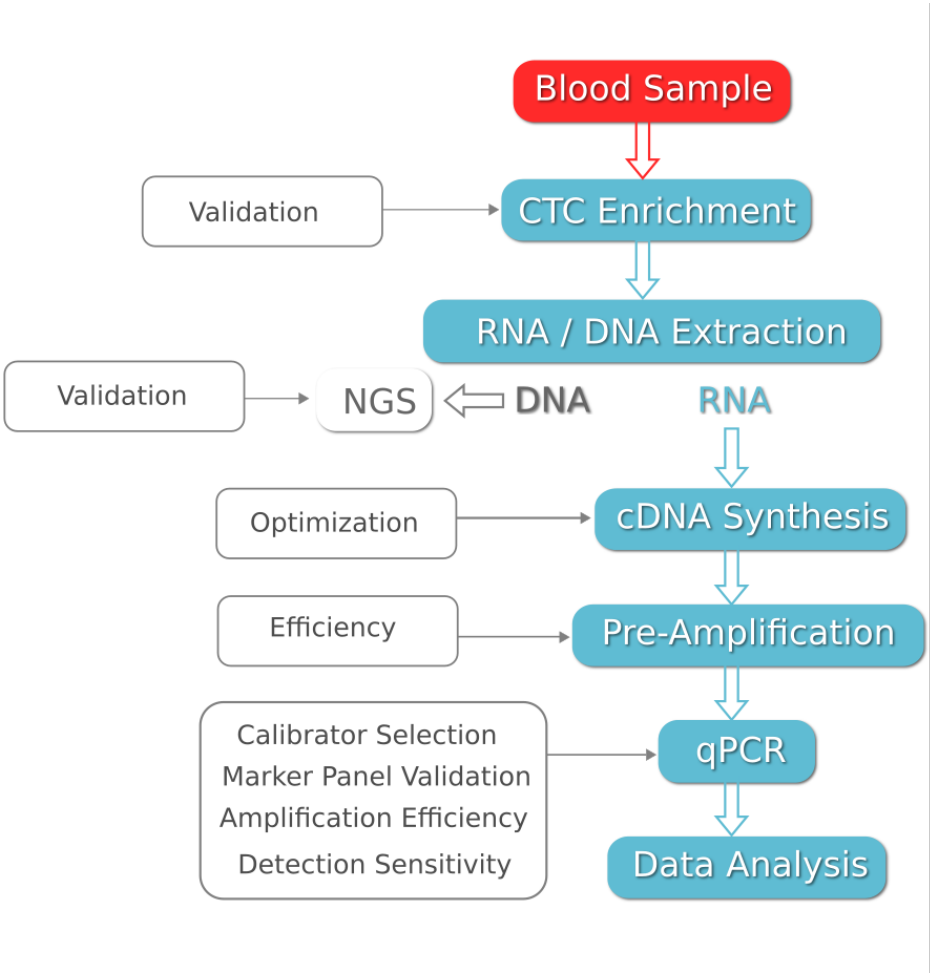

FIGURE 3.1: Methods workflow. The analysis of patient (PBCB) and control samples followed the main flow of the diagram from blood sample to data analysis and were the last experiments to take place. Additional and preliminary experimental tests are shown to the left in white boxes.

enrichment procedure. Starting PBMC number was 14,500,000 cells and 3500±122.93 PBMCs (n=4) were remaining in the enriched sample.

3.2 | Selection of candidate mRNA markers by SAGE analysis

As CTCs are a very heterogeneous population, we wanted to include a wide-coverage multimarker mRNA panel for indirect detection of these rare cells. For this purpose, SAGE analysis was used to select the best candidates. Several mRNA markers used in previous studies (see Appendix A) as well as differentially expressed tags were assessed across all SAGE libraries for WBC and breast tissue (see section 2.2.10.1). The most promising candidates that were analyzed are shown in Table 3.1 and include 25 markers that represent a variety of both epithelial and mesenchymal markers. A more restricted

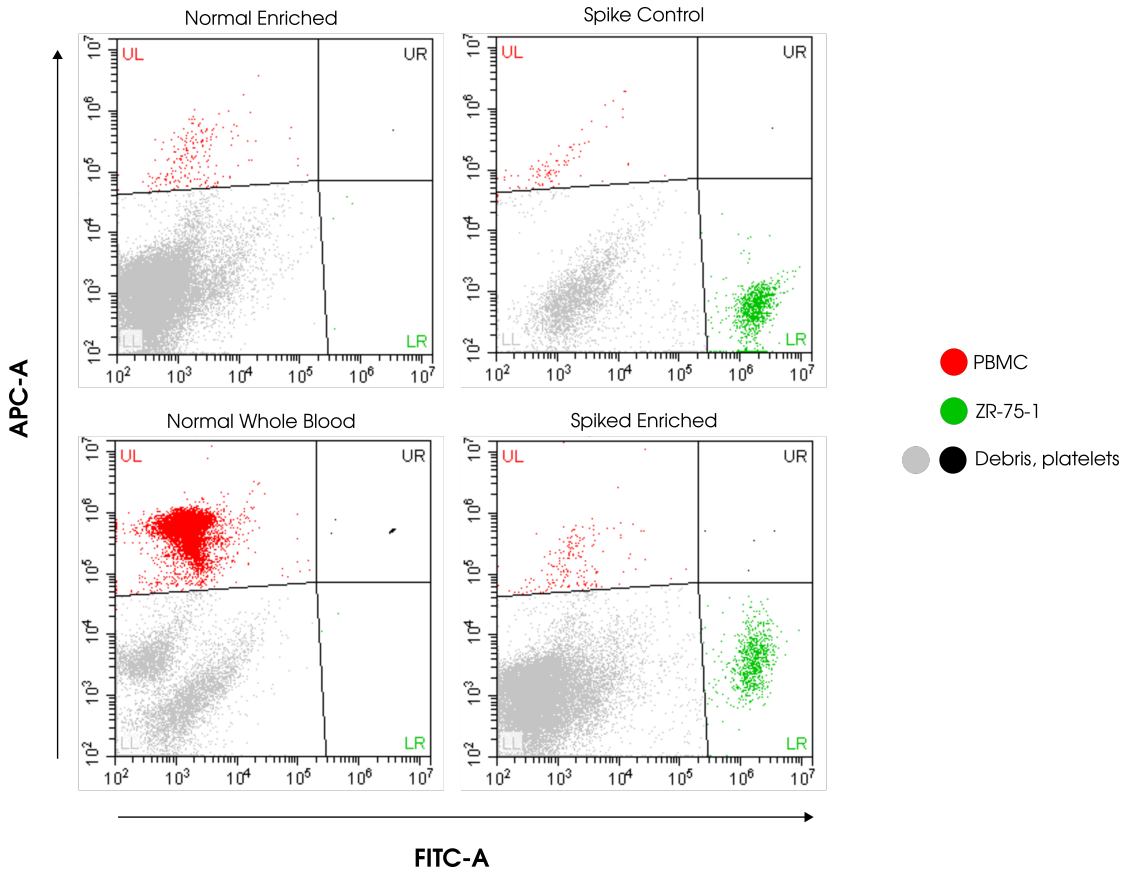

FIGURE 3.2: Flow Cytometry analysis of spiked samples and controls. APC-A shows the intensity of label bound to CD45. FITC-A shows the intensity of label bound to EpCAM. Colored populations are as follows: PBMCs, red; ZR-75-1 (cancer cells), green; Other particles/debris/platelets, black and gray. Normal whole blood is diluted blood. Normal enriched is an unspiked normal blood sample. Spike control is ZR-75-1 cells spiked into staining buffer. Spiked enriched sample is a normal blood sample spiked with ZR-75-1 cells. Gating position was based on the separation of WBCs in the whole blood sample and cancer cells in the spike control.

panel was chosen from these by comparing the ratio of SAGE tags in WBC to breast tissue. The selection of the best candidate markers was done based on a high expression in normal breast/breast cancer tissues and a low expression in WBCs, resulting in high ratios. All the epithelial marker candidates (*EPCAM*, *KRT16*, *KRT19*, KRT7, KRT8) were chosen for further analyses due to the low number of tags in WBCs (Table 3.1). EMT markers chosen were *ZEB1, SLUG, SNAIL, TWIST1*, and *TFF3*. Other markers included were *ERBB2, SCGB, HSP47, CCDC80,* and *LUM*. With the reference marker/housekeeping gene, breakpoint cluster region (*BCR*) included, a total of 16 markers were evaluated in preliminary experiments.

TABLE 3.1: Comparison of SAGE [110] tag counts in WBCs and breast tissue. Libraries: SAGE breast normal and SAGE breast carcinoma, SAGE White Blood Cells

		Average Tags/200,000 over all libraries			Ratio Breast:WBC	
Gene Tag	Marker Type	WBC	Breast – Normal	Breast – Neoplasia	Norm	Neoplasia
EPCAM	Epithelial	0.3	6.9	12.7	25	47
KRT16	Epithelial	0.0	18.6	4.5	-	-
KRT19	Epithelial	0.0	137.8	124.5	-	-
KRT7	Epithelial	0.0	166.5	66.0	-	-
KRT8	Epithelial	0.7	94.6	137.8	130	189
ZEB1	Mesenchymal	0.9	2.6	3.3	3	4
SLUG (SNAI2)	Mesenchymal	0.0	1.6	2.8	-	-
SNAIL (SNAI1)	Mesenchymal	1.2	1.0	0.9	1	1
TWIST (TWIST1)	Mesenchymal	0.2	1.1	1.8	-	-
FN1	Mesenchymal	14.6	7.7	49.1	1	3
ERBB2 (HER2)	Epithelial	0.9	3.8	57.1	4	63
hTERT	Mesenchymal	0.0	0.1	0.0	-	-
BIRC5	Mesenchymal	0.1	0.6	2.8	-	-
CDK4	Mesenchymal	2.6	3.6	8.9	1	3
TFF1	Mesenchymal	0.5	63.6	69.1	140	152
TFF3	Mesenchymal	0.0	13.3	269.9	-	-
HSPA5	Mesenchymal	44.4	115.2	207.8	3	5
HSP47	Mesenchymal	0.8	37.2	45.6	45	56
HSPA6	Mesenchymal	1.3	4.7	28.7	4	23
CCDC80	Novel	0.1	14.4	16.4	158	181
LUM	Novel	0.8	37.2	45.6	45	56
COL1A1	Mesenchymal	4.7	83.1	473.8	18	100
SPARC	Mesenchymal	11.5	114.0	246.6	10	22
GADD45A	Novel	0.5	29.0	27.6	53	51
SCGB (mammoglobin)	Epithelial	0.0	17.6	103.0	-	-

3.3 | Validation of candidate mRNA markers in cell lines & selection of calibrator

To be considered as a potential mRNA marker for indirect detection of CTCs the markers, had to be highly expressed in cell lines and breast cancer tissue, and have low or no background expression in control samples from healthy individuals. Based on these criteria the expression of all mRNAs were evaluated in four breast cancer cell lines, 10 breast cancer tissue samples and in 4 control samples.

3.3.1 | Cell Line Expression of Markers

The level of the candidate markers were measured in four cell lines as a preliminary test of marker potential and for calibrator cell line selection. The cell lines were chosen based on availability and their epithelial and mesenchymal characteristics. For this reason, a mesothelioma line was used in addition to breast cancer lines. The cell lines evaluated were ZR-75-1, MDA-MB-231, MCF-7, and SDM1032T (Table 3.2). We expected the candidate markers to be expressed in at least some of these cell lines.

The RNA was extracted from the harvested cell line lysates (section 2.2.1.4) and was then reverse transcribed (section 2.2.6.1). The cell line samples were not pre-amplified

TABLE 3.2: Cell line information from ECACC.

Cell Line	Primary Source	Phenotype
ZR-75-1	breast IDC, from malignant ascitic effusion	Epithelial
MCF-7	breast adenocarcinoma, from pleural effusion	Epithelial-like
MDA-MB-231	breast adenocarcinoma, from pleural effusion	Epithelial
SDM103T2	malignant pleural mesothelioma, from xenograft	Epithelial/ Mesenchymal

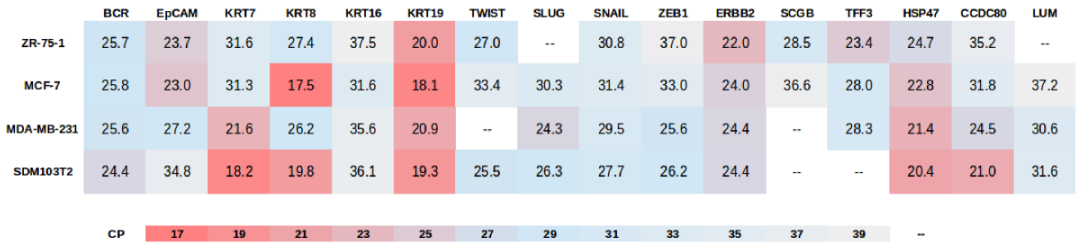

FIGURE 3.3: Expression of markers in cell lines. Given as Cq values (average over duplicates). No Cq value denotes a undetectable signal or a Cq greater than 39. A lower Cq value reflects higher expression of the marker

prior to quantification due to the abundant number of cells collected. Relative expression of the all genes of interest were then quantified by qPCR (section 2.2.7.2).

The average Cq values of each cell line amplified by each assay is shown in Figure 3.3. As expected, the reference gene (*BCR*) was evenly expressed between the cell lines with Cq values from 24.4 to 25.8 (highest in SDM103T2), and therefore the Cq values of the other markers were comparable. ZR-75-1 exhibited the highest expression for *ERBB2, SCGB,* and *TFF3* mRNA. MCF-7 was the best cell line for expression of *EPCAM, KRT8, KRT16*, and *KRT19*. The highest expressed markers in MBA-MB-231 were *SLUG, ZEB1,* and *LUM*. The remaining markers (*KRT7, TWIST1, SNAIL, HSP47*, and *CCDC80*) were highest expressed in SDM103T2. *KRT16* was excluded from further analysis due to low expression across all cell lines. Since there was not a single cell line with consistent expression over all markers, two cell lines were chosen to cover all markers for the calibrator cell. An 1:1 mix of ZR-75-1 and SDM103T2 cDNA was used to produce the calibrator cell (CC).

3.3.2 | Marker expression in breast tumor tissue and enriched controls

Breast tissue RNA samples from 10 different tumors were purchased from Asterand (section 2.1.2), available for analysis, and offered a diverse sample population for marker testing. In addition to the breast tissue samples, four normal control blood samples were taken from 3 healthy donors (2 samples came from one donor). The samples were enriched by our MINDEC processing protocol (section 2.2.4) and RNA was extracted (section 2.2.5). The RNA from these samples was reverse transcribed (section 2.2.6.1), with the maximum volume of sample was used as the RNA concentration of MINDEC samples were undetectable. As a consequence of this, the cDNA synthesized from all

TABLE 3.3: Average relative expression of breast tissues and normal blood controls. NB Ave with NA values denote there was no detectable expression in any of the samples. Ratio is Tumor Ave:NB Ave.

Assay	Tumor Ave	NB Ave	Ratio
ERBB2	1.29E+00	1.20E-02	107.6
SCGB	5.20E+02	NA	-
TFF3	2.10E+01	2.92E-01	71.9
HSP47	6.14E-01	5.70E-03	107.8
CCDC80	1.00E+00	2.76E-03	364.1
LUM	3.73E+03	1.74E+00	2139.4
SNAIL	1.49E+00	1.06E-01	14.0
TWIST	4.55E-01	1.05E-02	43.3
SLUG	2.06E+00	NA	-
ZEB1	3.67E+00	1.37E+00	2.7
EPCAM	1.08E+00	7.70E-03	140.2
KRT7	5.27E-02	NA	-
KRT8	3.91E-01	NA	-
KRT19	1.49E+00	NA	-

samples was also pre-amplified prior to qPCR (sections 2.2.7.1 and 2.2.7.2). The pre-amplified DNA was loaded on a 384-well plate and all samples (including the CC and NTC) were run in triplicate.

Cq values were exported from the LightCyler software and used to calculate relative gene expression (section 2.2.10.2). The expression profiles of the tissue samples varied, so the average expression of the tumor tissues (Tumor) and normal blood from controls (NB) were considered when selecting the final marker panel (Figure 3.4 and Table 3.3). The selection criteria were similar to previous, with high expression in tumor and low or no expression in normal blood being preferred resulting in a high ratio of expression. Detailed charts with relative expression for each tissue and control can be found in Appendix C. Expression was undetectable in all normal blood controls for the following markers: *KRT7, KRT8, KRT19, SCGB*, and *SLUG*. High background expression in the controls was present for the markers *HSP47, TFF3,* and *ZEB1*. For this reason, they were excluded from the final panel. *KRT7* was also excluded due to overall low expression in the breast tissue samples (average relative expression of 5.27E-02). The markers selected to comprise the final mRNA multimarker panel were the following: *CCDC80, EPCAM, ERBB2, KRT8, KRT19, LUM, SCGB, SLUG, SNAIL, TWIST*.

3.4 | Validation of quantitative PCR assays

3.4.1 | Amplification efficiency of assays

Taqman Gene Expression Assays are manufactured and tested by Applied Biosystems and are guaranteed to have deficiencies of 100±10%. To confirm this, the efficiency of all

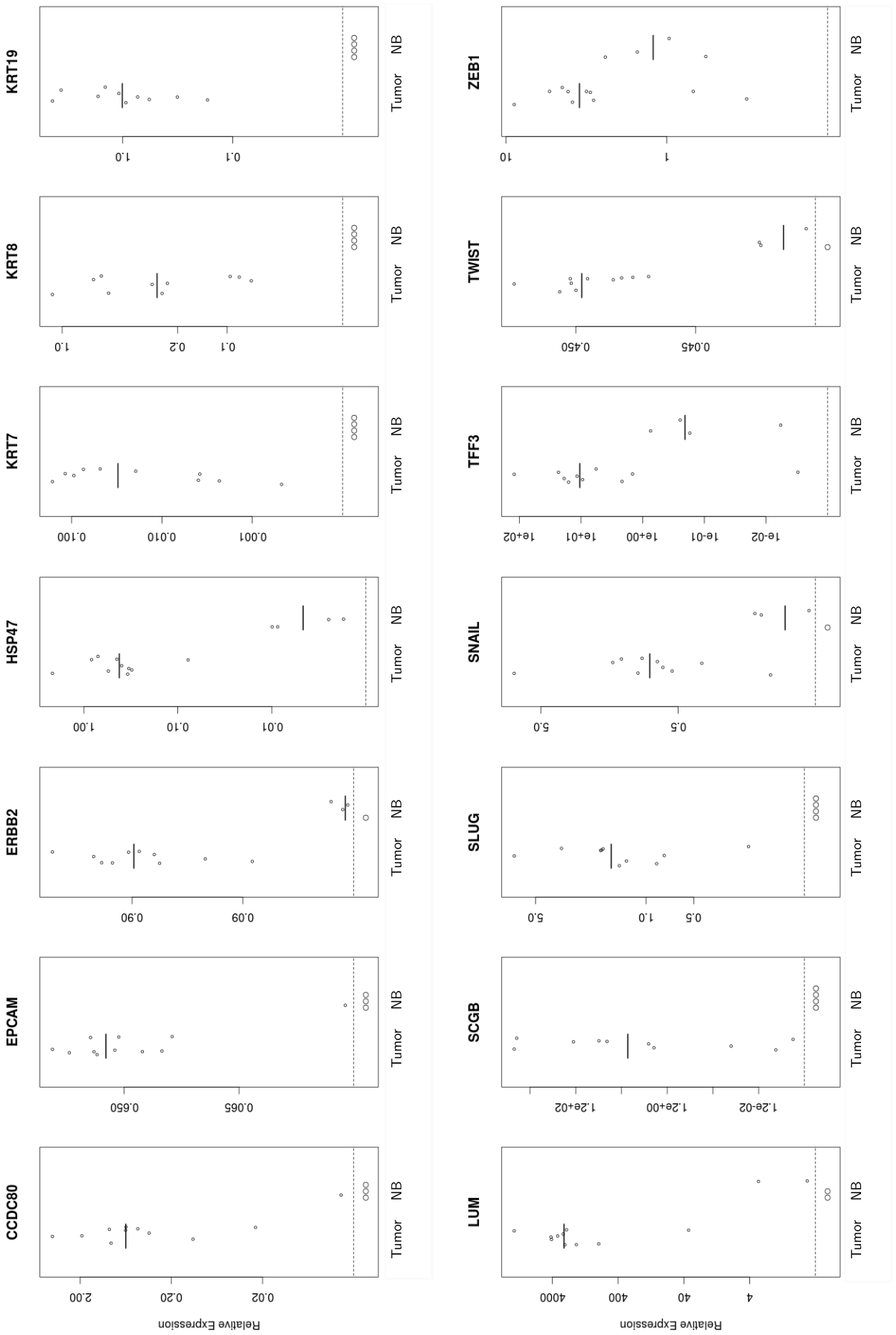

FIGURE 3.4: Relative level of the candidate markers in breast tumor (Tumor) and normal blood (NB) samples. Points below the lower dotted line represent samples with non-detectable expression (NA). The solid lines within the sample populations represent the median of that sample group.

TABLE 3.4: Amplification efficiency of assays (E) and coefficients of determination from linear regression of the standard curve (R^2). *LUM: 2 points only. **SNAIL: 4 points only.

Assay	E %	R2
BCR	96.0	0.997
CCDC80	96.8	0.998
HER2	101.7	0.996
EPCAM	97.4	0.995
HSP47	95.2	0.998
KRT7	93.4	0.999
KRT8	86.2	0.998
KRT19	96.2	0.999
*LUM**	98.0	0.985
SCGB	100.5	0.992
*SNAIL***	95.2	0.997
SLUG	88.2	0.994
TFF3	99.4	0.997
TWIST	89.5	0.994
ZEB1	93.7	0.998

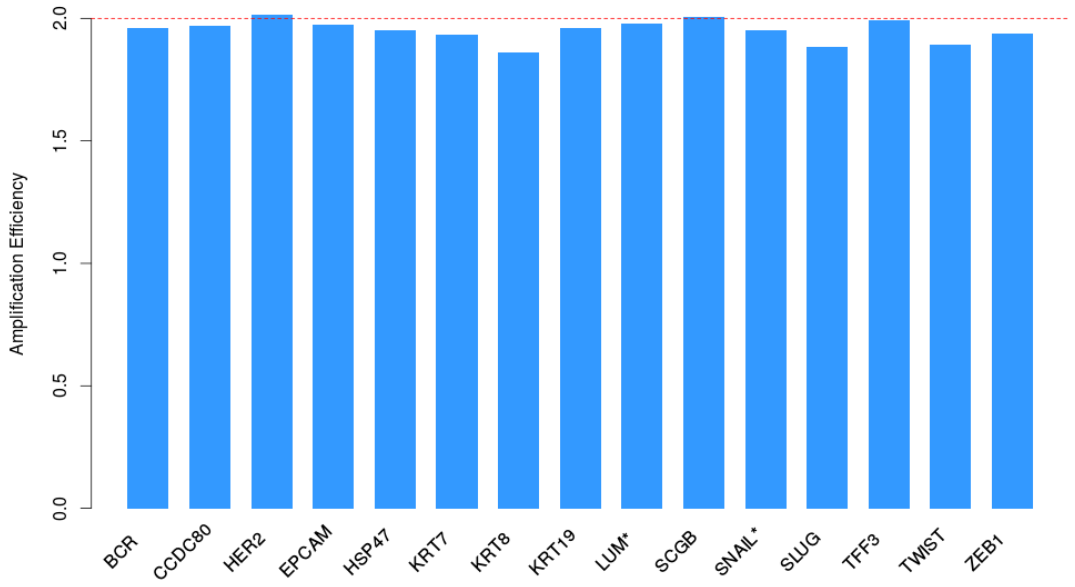

FIGURE 3.5: Amplification efficiency of assays. LUM: 2 points only. SNAIL: 4 points only.

assays were measured by standard curve analysis (Appendix B) using cDNA from the cell lines with highest expression of the marker tested as template in the qPCR reactions (section 2.2.8). The efficiencies measured were within this range with the exception of *KRT8* (86.2%), *SLUG* (88.2%), and *TWIST* (89.5%) (Table 3.4 and Figure 3.5).

3.4.1.1 | Optimization of reverse transcription method

During the qPCR validation, recently prepared cDNA samples were resulting in lower efficiencies than expected. When averaged across all experiments using the *KRT19* assay and MDA-MB-231 M-MLV reverse transcribed (section 2.2.6.1) cDNA, the efficiency was as low as 83.412.62% (n=8). This was in contrast to an average efficiency of 97.402.55% (n=5) for the *KRT19* assay when using cDNA from an older lung cancer calibrator sample as a template in the qPCR. After several experiments, it was concluded that the M-MLV reverse transcription method being used was incompatible with the Taqman Gene Expression Assays. This resulted in testing of different reverse transcription kits for compatibility: SuperScript IV FSS (SSIV) (section 2.2.6.3) and High Capacity cDNA Reverse Transcription (HCAB) (section 2.2.6.2). Using the same assay (*KRT19*), the SSIV kit resulted in an efficiency of 75.35% with MDA-MB-231 cDNA, while the HCAB kit yielded a 97.75% efficiency. Hence, the method that was found to be compatible with the downstream TaqMan Assays was the HCAB kit, recommended in the Taqman Gene Expression Assay protocol. The HCAB kit was therefore used to synthesize new cDNA from cell line RNA for the calibrator cells and efficiency measurements, and all PBCB samples.

3.4.2 | Amplification Efficiency of Template Pre-Amplification

Since the quantification of mRNA levels depends on the efficient pre-amplification of cDNA, the efficiency of this step was measured by similar standard curve analysis (Appendix B) to above (section 2.2.8). The amplification efficiency of the pre-amplification step was measured with 3 representative assays: *BCR, KRT19,* and *KRT8*. The efficiencies for the assays were found to be satisfactory at 96.3%, 93.2%, and 89.0%, respectively.

3.4.3 | Sensitivity

Spike experiments were performed in order to test the detection sensitivity of the enrichment and qPCR assays in combination. The sensitivity of the KRT19 assay was investigated by spiking normal blood samples with ZR-75-1 cells, and the sensitivity of the mesenchymal markers (*CCDC80, TWIST, SNAIL, SLUG, LUM*) by spiking with SDM-103T2 cells. The blood samples were collected from two healthy donors (n=5 from each). Four of the five samples were spiked with increasing numbers of cancer cells (10, 33, 100, 1000) (section 2.2.1.4). The remaining sample was unspiked and used as a negative and background expression control. The remainder of the protocol was carried out unchanged (sections 2.2.4, 2.2.5, 2.2.6.2, 2.2.7). Results are shown in Figure 3.6 and are averages from the two biological replicates used.

The *KRT19* assay was used as a representative for all the epithelial markers. There was no background expression detected in either unspiked sample. The sensitivity of

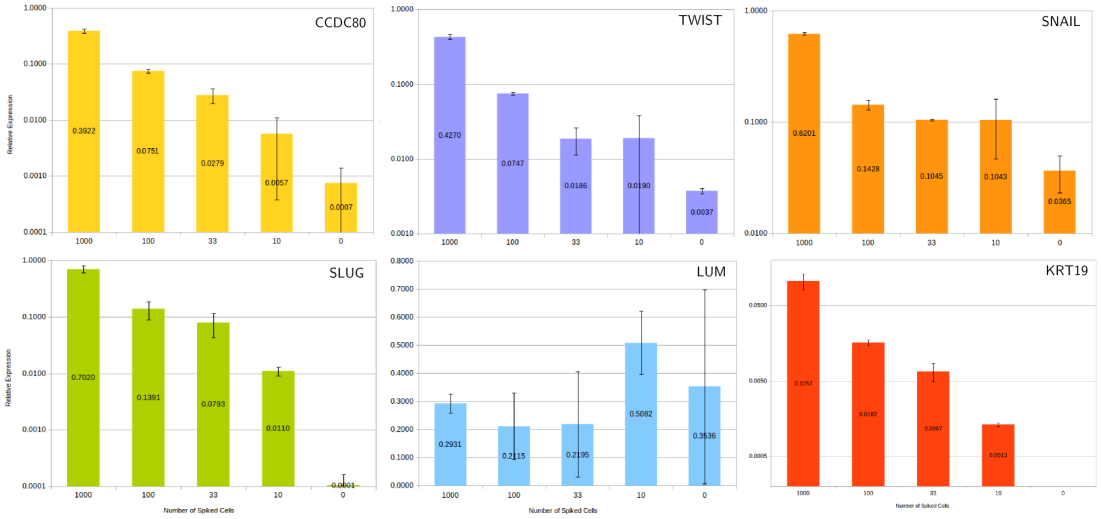

FIGURE 3.6: Sensitivity of enrichment and qPCR technique and detection of spiked cancer cells in normal enriched blood. Bar height represents the average over two biological replicates and the error bars display the standard deviation.

the assay was ≤10 cells. The expression across the spiked samples was linear (R^2 = 0.9978) and the average coefficient of variation between the two biological replicates was 16.24%. The biological replicates for the *CCDC80* assay were very divergent for the 10-cell and unspiked samples (CV of 93.30% and 88.08%, respectively). The sensitivity of this assay was ≤33 cells. The average expression over the spike series was linear (R^2 = 0.9922). The *LUM* assay did not exhibit linear expression in the spike series (R^2 = 0.0292) and was very inconsistent in general. Purely based on this cell line, the sensitivity is >1000 cells. The cell line expression of *LUM* was very low and not taken as representative of possible PBCB expression. Given the high expression of *LUM* in breast cancer tissue and low expression in healthy controls (Figure 3.4), it was still included in the multimarker panel. *SNAIL* both had considerable variation among the non-spiked and 10-cell spiked samples (36.88% and 55.45%). With the overlap of the samples, the sensitivity of the assay is ≤33 cells. *TWIST* varied by 99.61% on the 10-cell sample also causing the sensitivity to be ≤33 cells for this assay. *SLUG* expression was sensitive to ≤10 cells and also exhibited a fairly linear relationship between the spiked samples (R^2 = 0.9884). The 10-cell spike expression was over 100-fold higher than the background.

3.5 | CTC detection in PBCB Samples

PBCB patient and control blood samples were processed as described in section 2.2.4, pre-amplified (section 2.2.7.1), and marker levels quantified by qPCR (section 2.2.7.2). On each plate, the calibrator was run in duplicate with the target assay. On the first plate of each assay, the calibrator was also run in duplicate with *BCR* assay. Cq values were calculated in the LightCycler and the relative gene expression (RGE) was calculated (section 2.2.10.2). RNA quality of all samples was checked by *BCR* expression, used as

TABLE 3.5: Summary of relative gene expression and thresholds in control group. [a]Max value including all controls. [b]Outlier threshold of 3SD above the mean. [c]New control maximum and CTC-detection threshold after removal of outliers (only changed from [a] in *EPCAM, KRT8*, and *TWIST*). NA: denotes that the level was below the detection limit.

Assay	Mean	Max[a]	SD	Threshold[b]	Max[c]
CCDC80	4.87E-04	1.06E-03	2.62E-04	1.27E-03	1.06E-03
EPCAM	2.35E-03	1.18E-02	2.51E-03	9.89E-03	3.18E-03
ERBB2	1.87E-02	5.38E-02	1.22E-02	5.54E-02	4.11E-02
SCGB	NA	NA	NA	NA	NA
SNAIL	5.40E-02	1.79E-01	4.23E-02	1.81E-01	7.65E-02
TWIST	1.69E-02	7.47E-02	1.60E-02	6.48E-02	3.20E-02
KRT8	1.31E-04	5.02E-04	9.89E-05	4.28E-04	2.43E-04
LUM	3.29E-01	7.85E-01	2.46E-01	1.07E+00	7.85E-01
SLUG	1.00E-02	2.34E-02	8.49E-03	3.55E-02	2.34E-02
KRT19	5.12E-05	8.59E-05	5.12E-05	1.37E-04	8.59E-05

a reference gene. Only samples that were positive for *BCR* expression were further analyzed (none were excluded here). All RGE data are shown in Figure 3.7. Any PBCB sample with RGE greater than the maximum control was considered CTC-Positive.

3.5.1 | Healthy Controls

There were 30 healthy female control samples used in this project. The control RGE data was analyzed for each assay and this is summarized in Table 3.5. The mean relative gene expression (RGE) and standard deviation (SD) was calculated. Outliers in the control samples were considered as such if they were greater than 3 SD from the mean and removed from the final maximum calculation. This analysis yielded three outliers total among three assays: C59 in *EPCAM*, C56 in *KRT8*, and C58 in *TWIST*. With the outliers removed, there were no positive samples among the controls. There was however considerable background expression for most markers. The only exception was SCGB, with no control samples showing expression. Markers with the highest control expression were *ERBB2, SNAIL, CCDC80, KRT8*, and *TWIST*. *LUM* and *EPCAM* were in the middle with about half of the controls having background expression. Markers that had lower expression in controls were *KRT19, SLUG*, and *LUM*.

3.5.2 | Patient Samples

The relative gene expression of all samples (patients and controls) are shown in Figure 3.7. Any PBCB sample with RGE greater than the maximum control was considered CTC-Positive. Samples with CTC-positive status are shown in Table 3.6. In total, 37 samples were found to be CTC-positive (21.8% of all samples). This includes 2 sample timepoints from the same patient: 139 (V1 and V1.5) and 165 (V1 and V1.5). If patients

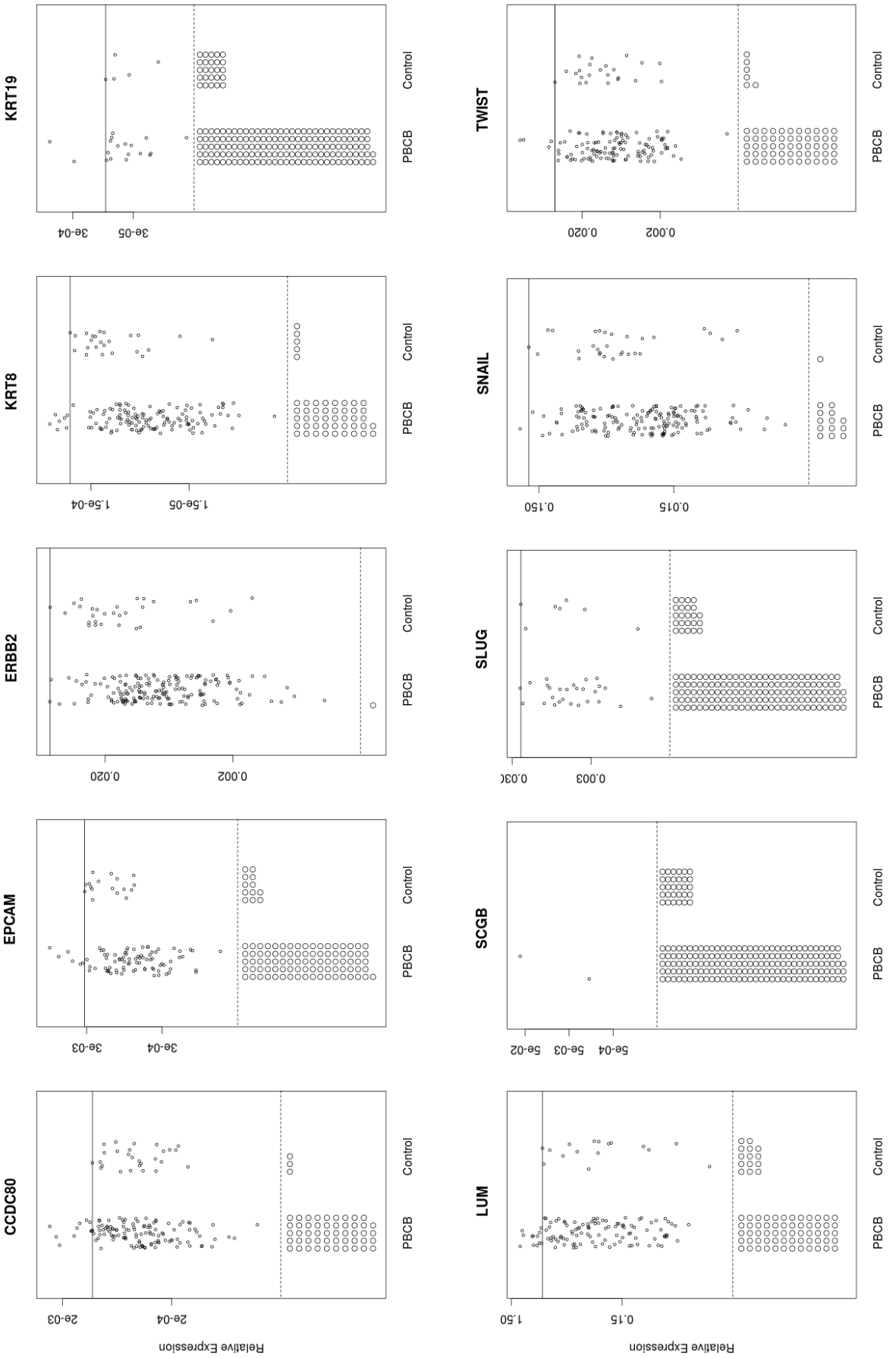

FIGURE 3.7: Relative Expression of Patients (labeled PBCB) and controls. Points displayed under the dotted line represent undetectable expression (no Cq value). The solid line marks the maximum control value and CTC-positive threshold for each marker.

TABLE 3.6: CTC-positive samples. Samples with gene expression over the maximum control were considered positive.

ID	Visit	CCDC80	EPCAM	ERBB2	KRT19	LUM	KRT8	SCGB	SLUG	SNAIL	TWIST
135	V1.5	•									
139	V1.5	•									
139	V2						•				
140	V1.5				•	•					
143	V1.5								•		
144	V2		•								
146	V2					•					
154	V1					•					
157	V1										•
158	V1	•									
165	V1	•	•								
165	V1.5		•								
170	V1.5					•	•				
175	V1					•					
176	V1		•				•				
177	V1		•								
186	V1	•									
188	V1.5	•									
189	V1.5	•									
190	V1.5		•								
196	V1					•					
197	V1					•					
198	V1					•					
209	V1				•		•	•			•
222	V1									•	
232	V1	•									
237	V1					•	•				
244	V1	•									
246	V1			•							
247	V1										•
253	V1						•	•			
256	V1					•					
260	V1										•
263	V1					•					
265	V1					•					
268	V1						•				
269	V1	•									
37		10	6	1	2	12	7	2	1	1	4

are counted as positive for having at least one positive sample, this is a total of 35 positive patients (26.3% of 133 patients evaluated). Seven patients (5.3%) were positive for more than one marker, of these 6 patients were positive for 2 and one patient for 4 markers. The markers with the most coverage were *CCDC80, LUM, EPCAM, KRT8,* and *TWIST* (Table 3.7). These four markers combined include 91.4% (32) of all the positive patients. The other markers (*ERBB2, KRT19, SCGB, SLUG*, and *SNAIL*) were detected at rates of less than 6% among the positive samples and a total of 7 patients.

If the markers are split into epithelial (*EPCAM, ERBB2, KRT19, KRT8, SCGB*) and EMT (*CCDC80, LUM, SLUG, SNAIL, TWIST*) categories, 24.3% (9) of positive samples exhibited an epithelial-only phenotype, 59.5% (22) were EMT-only, and 16.2% expressed both epithelial and EMT markers. One patient sample that was epithelial (EPCAM) and EMT positive (CCDC80) on the first visit, lost the EMT-positivity at

TABLE 3.7: Number of patients positive for each marker. *One sample not included in the EPCAM total since it was the second timepoint for the patient.

	Number of positive patients	**(% of all patients)**	**(% of positive patients)**
CCDC80	10	(7.5)	(28.6)
EPCAM	4*	(3.0)	(11.4)
ERBB2	1	(0.8)	(2.9)
KRT19	2	(1.5)	(5.7)
LUM	12	(9.0)	(34.3)
KRT8	7	(5.3)	(20.0)
SCGB	2	(1.5)	(5.7)
SLUG	1	(0.8)	(2.9)
SNAIL	1	(0.8)	(2.9)
TWIST	4	(3.0)	(11.4)
At least 1	35	(26.3)	(100)
At least 2	6	(4.5)	(17.1)
At least 4	1	(0.8)	(2.9)

the next visit (ID 165). In another case (ID 139), there was a switch from EMT to epithelial (*CCDC80* to *KRT8*). Only 2 samples were positive for *SCGB*, but in both of these, they were also positive for *KRT8*.

We analyzed CTC-status for association with clinicopathological parameters by the Fisher's exact test (categorical) and Kruskal-Wallis rank sum test (continuous) (section 2.2.10.3). However, CTC-status was not found to be a significantly associated with any of the clinicopathological features shown in Table 3.8 (stratified by markers in Appendix E). Metastases were already present in 28.4% of patients at diagnosis, but CTCs were not detected at a significantly different rate. Only 2 of the patients had detectable CTCs. There was a trend towards PR+ status in the CTC-positive group (not significant, p=0.167). Of 14 triple-negative patients, 3 were CTC-positive (3 for *LUM*, 1 also for *KRT8*). 3 DCIS patients were positive for CTCs (1 marker each: *EPCAM, SCGB,* and *TWIST*).

3.6 | Detection of CTCs by Targeted Sequencing

As a preliminary test, we wanted to investigate whether we were able to sequence to a detection level of 100 spiked cancer cells, in a background of leukocytes, using a commercial available kit for library preparation (Ion Ampliseq Cancer Hotspot Panel v2). This kit uses a single pool of primers to perform multiplex PCR for preparation of amplicon libraries from genomic "hot spot" regions that are frequently mutated in human cancer genes.

Three samples were sequenced in this pilot study on our Ion Proton instrument (section 2.2.9): a breast cancer cell line (ZR-75-1) and two normal, enriched blood samples spiked with 1000 and 100 ZR-75-1 cells. All three genomic DNA samples were amplified

TABLE 3.8: Patient clinicopathological characteristics stratified by CTC-status at first visit and all visits. Patients were counted as positive if they were positive for at least one marker. Fisher's exact test for categorical variables. *Kruskal-Wallis rank sum test for continuous variables.

		Visit 1 Only			All Visits		
	Overall	**neg**	**pos**	**p**	**neg**	**pos**	**p**
n	131	106	25		97	34	
Age (median [IQR])	60.00 [53.00, 65.50]	60.00 [52.25, 67.00]	62.00 [54.00, 64.00]	0.837*	60.00 [53.00, 66.00]	62.50 [53.00, 64.75]	0.703*
Diagnosis (%)				0.834			0.811
DCIS	17 (13.0)	15 (14.2)	2 (8.0)		14 (14.4)	3 (8.8)	
IDC	96 (73.3)	75 (70.8)	21 (84.0)		70 (72.2)	26 (76.5)	
ILC	8 (6.1)	7 (6.6)	1 (4.0)		6 (6.2)	2 (5.9)	
IMC	3 (2.3)	2 (1.9)	1 (4.0)		2 (2.1)	1 (2.9)	
IPC	1 (0.8)	1 (0.9)	0 (0.0)		1 (1.0)	0 (0.0)	
ITC	2 (1.5)	2 (1.9)	0 (0.0)		2 (2.1)	0 (0.0)	
other	4 (3.1)	4 (3.8)	0 (0.0)		2 (2.1)	2 (5.9)	
T Stage (%)				0.746			0.422
1	73 (55.7)	57 (53.8)	16 (64.0)		51 (52.6)	22 (64.7)	
2	40 (30.5)	33 (31.1)	7 (28.0)		31 (32.0)	9 (26.5)	
3	2 (1.5)	2 (1.9)	0 (0.0)		2 (2.1)	0 (0.0)	
is	7 (5.3)	7 (6.6)	0 (0.0)		7 (7.2)	0 (0.0)	
undetermined	9 (6.9)	7 (6.6)	2 (8.0)		6 (6.2)	3 (8.8)	
Tumor 1 Size(median [IQR])	16.00 [12.00, 26.75]	17.00 [12.00, 27.00]	15.00 [10.75, 23.50]	0.220*	17.00 [12.00, 29.25]	15.00 [11.75, 22.25]	0.427*
Multifocal (%)	16 (12.2)	13 (12.2)	3 (12.0)		13 (12.2)	3 (12.0)	
N Stage (%)				0.919			0.753
N0	89 (67.9)	71 (67.0)	18 (72.0)		65 (67.0)	24 (70.6)	
N1	23 (17.6)	18 (17.0)	5 (20.0)		16 (16.5)	7 (20.6)	
N2	5 (3.8)	5 (4.7)	0 (0.0)		5 (5.2)	0 (0.0)	
N3	1 (0.8)	1 (0.9)	0 (0.0)		1 (1.0)	0 (0.0)	
undetermined	13 (9.9)	11 (10.4)	2 (8.0)		10 (10.3)	3 (8.8)	
Metastasis (%)	19 (28.4)	18 (31.6)	1 (10.0)	0.260	17 (30.9)	2 (16.7)	0.485
Grade (%)				0.575			0.677
1	20 (15.3)	14 (13.2)	6 (24.0)		13 (13.4)	7 (20.6)	
2	47 (35.9)	38 (35.8)	9 (36.0)		34 (35.1)	13 (38.2)	
3	48 (36.6)	40 (37.7)	8 (32.0)		37 (38.1)	11 (32.4)	
DCIS	16 (12.2)	14 (13.2)	2 (8.0)		13 (13.4)	3 (8.8)	
ER Status (%)				0.603			0.626
neg	16 (12.2)	14 (13.2)	2 (8.0)		13 (13.4)	3 (8.8)	
pos	99 (75.6)	78 (73.6)	21 (84.0)		71 (73.2)	28 (82.4)	
undetermined	16 (12.2)	14 (13.2)	2 (8.0)		13 (13.4)	3 (8.8)	
PR Status (%)				0.407			0.167
neg	34 (26.0)	30 (28.3)	4 (16.0)		29 (29.9)	5 (14.7)	
pos	79 (60.3)	61 (57.5)	18 (72.0)		54 (55.7)	25 (73.5)	
undetermined	18 (13.7)	15 (14.2)	3 (12.0)		14 (14.4)	4 (11.8)	
HER2 Status(%)				0.552			0.874
neg	104 (79.4)	84 (79.2)	20 (80.0)		76 (78.4)	28 (82.4)	
pos	11 (8.4)	8 (7.5)	3 (12.0)		8 (8.2)	3 (8.8)	
undetermined	16 (12.2)	14 (13.2)	2 (8.0)		13 (13.4)	3 (8.8)	
Ki67(median [IQR])	31.00 [19.00, 44.00]	33.00 [20.00, 48.00]	24.00 [6.00, 38.50]	0.093*	31.00 [19.75, 48.25]	32.00 [12.50, 39.00]	0.332*
Lumpectomy (%)	101 (77.1)	80 (75.5)	21 (84.0)	0.438	73 (75.3)	28 (82.4)	0.482
Mastectomy(%)	35 (26.7)	29 (27.4)	6 (24.0)	0.807	26 (26.8)	9 (26.5)	1.000

by the primer pool included in the kit, which should result in 207 amplicons covering approximately 2,800 COSMIC mutations from 50 oncogenes and tumor suppressor genes.

Upon completion, the data was analyzed and saved on the Ion Torrent server. Analysis was done according to the run parameters and everything was automatically calculated (section 2.2.10.4). General information regarding the next-generation sequencing run are shown in Figure 3.8. Average ISP loading of the chip was 65%. From this, there were 40% usable reads and 35% empty wells. An enrichment of 100% enrichment was achieved meaning that no empty beads had been included in the sequencing run. However, 28% of ISPs were polyclonal, that is more than one template was attached to the bead, resulting in sequences with low quality. In total, 43% of sequences were of low quality. The average read length for the amplicons was 107 bp mean (109 bp median, 96 bp mode). The samples themselves ran successfully with a high amount of average reads across amplicons (Table 3.9). Both the spiked samples had a greater number reads and depth than the cell harvest. End-to-end reads, though, were similar.

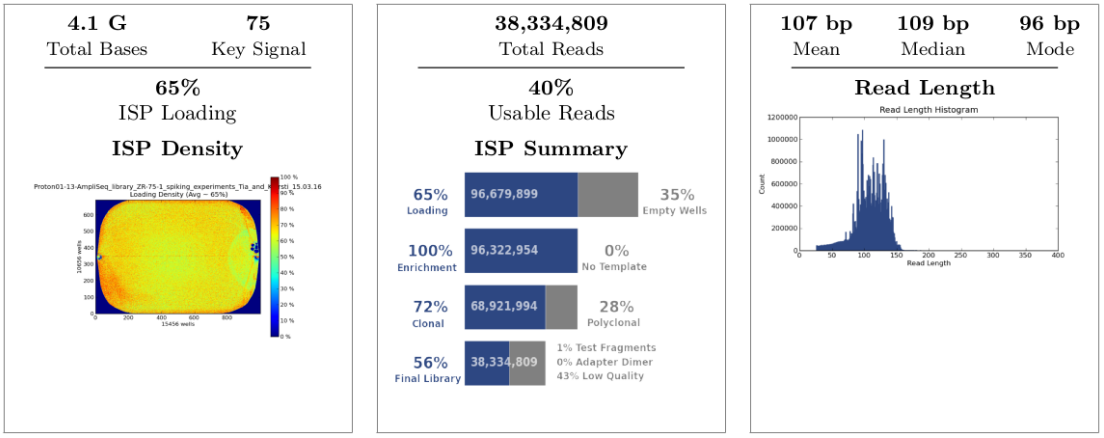

FIGURE 3.8: Summary of NGS Run.

TABLE 3.9: Summary of reads in each sequenced sample.

Sample	Reads	Mean Length	Amplicon Reads	End-to-End
ZR	10,251,920	113 bp	47,162	81.64%
1000	13,603,450	106 bp	64,631	80.68%
100	13,356,830	107 bp	63,547	81.64%

In the ZR-75-1 cell harvest and 1000-cell spike, 19 variant calls were made, while only 12 were called in the 100-cell sample (Table 3.10). The first 7 variants in the table are variants that are from the positive ZR-75-1 control sample that were also detectable in the 1000-cell spike. These variants were not detected in the 100-cell spike. The frequency of the variants in the 1000-cell spike were an average of 9.4% of the frequency found in the cell harvest sample. This would suggest that the total amount of cells in the sample is approximately ten times the amount of the spike, or 10,000 cells total. The most important variant is the PTEN variant on chromosome 10 (pos:89692839). This is a characteristic mutation of ZR-75-1 cells [113, 114] and was present in samples at a frequency of 98.8%, 14.0%, and undetected, respectively. This demonstrates that the cancer cells are detectable in the 1000-cell spike, but not the 100-cell spike. The 100-cell spike can be used as a comparison of leukocyte variants for the other samples, however. With this comparison, it can be seen that the first 9 variants in Table 3.10 are from the ZR-75-1 cells. The next 9 are shared among all samples, with some variation in frequency. Two variants were exclusive to the leukocytes.

Some mistakes were noted in the variant calling (a,b,c in Table 3.10). The variant at chr4:1806187 was not called in the 100-cell spike. Upon investigation of the BAM/VCF files, though it was also present at a frequency of 30%, but just not called as such. The variant on chr5:149433596 in the 100-cell spike (b) was called at 3.30% but after looking at the detailed variant information in IVG, it was seen that the software had misattributed variants from chr5:149433596 (all the T→A variants from TG→GA) as a separate call. This position had a lower read depth (22,267 reads) and was located near the end of the amplicon. This could cause the 3-4% of incorrect bases. All the samples

had about 5% frequency of a variant at position chr17:7578475. In IVG this particular base was at a lower frequency than both base insertions and deletions, thus showing low quality and should be disregarded. Other variants at non-standard frequencies (outside 50% and 100%) were also investigated in the BAM/VCF files, but there was nothing found to be the cause as they were all present at high coverage.

TABLE 3.10: Variant calling results from Cancer HotSpot gene panel on Ion Proton. Position: chromosome and location of variant. Ref: base present in reference sequence. Variant: variant base called that differs from the human genome reference (Hg19). Allele call: variant present on one (heterozygous) or both (homozygous) alleles. Gene: gene region in which variant is located. Frequency: detected rate of variant base in each sample. ZR: ZR-75-1. 1000: 1000-cell spike of normal blood with ZR-75-1 cells. 100: 100-cell spike of normal blood with ZR-75-1 cells. Normal variants: normal variants from the reference and corresponding frequencies in population [**?**]. Variant source: decided source of variant based on results.

					Frequency in each sample				
Position	Ref	Variant	Allele Call	Gene	ZR	1000	100	Normal Variants (%)	Variant Source
chr4:55972974	-	A	hetero	KDR	67.20%	-	-	NA	ZR-75-1
chr4:55152040	C	T	hetero	PDGFRA	23.10%	-	-	T (17.1217), A (0.0013)	ZR-75-1
chr2:212812097	T	C	hetero	ERBB4	42.50%	6.20%	-	C (32.0229)	ZR-75-1
chr3:178917005	A	G	hetero	PIK3CA	34.10%	3.90%	-	G (18.968)	ZR-75-1
chr4:55972974	T	A	hetero	KDR	63.20%	5.80%	-	A (21.4782)	ZR-75-1
chr4:1806131a	T	C	hetero	FGFR3	30.40%	4.60%	-	C (0.3188), A (0.0013)	ZR-75-1
chr10:89692839	T	G	homo	PTEN	98.80%	14.00%	-	no known variants	ZR-75-1
chr13:28602292	T	C	homo	FLT3	100.00%	7.90%	-	C (5.2367)	ZR-75-1
chr19:17945696	C	T	hetero	JAK3	58.40%	4.50%	-	T (0.8372)	ZR-75-1
chr4:1807894	G	A	homo	FGFR3	100.00%	100.00%	100.00%	A (91.8899)	Both
chr4:55141055	A	G	homo	PDGFRA	100.00%	100.00%	100.00%	G (93.6569)	Both
chr5:112175770	G	A	homo/hetero	APC	100.00%	62.10%	50.90%	A (62.2965)	Both
chr5:149433596	TG	GA	hetero/homo	CSF1R	52.60%	94.10%	96.70%	GA (2.8933)	Both
chr7:55249063	G	A	homo	EGFR	100.00%	100.00%	100.00%	A (50.4360)	Both
chr10:43613843	G	T	homo	RET	100.00%	100.00%	100.00%	T (70.7002), A(0.0019)	Both
chr13:28610183	A	G	homo	FLT3	100.00%	100.00%	100.00%	G (65.8885)	Both
chr17:7579472	G	C	homo	TP53	100.00%	100.00%	100.00%	C (62.2038), T (0.0006)	Both
chr4:1806187	C	A	hetero	FGFR3	29.50%	30.70%	-	A (0.0032)	Both
chr10:43615633	C	G	hetero	RET	-	30.90%	42.40%	G (19.1393), A (0.0032)	Leukocytes
chr11:108170506	A	C	hetero	ATM	-	49.40%	51.90%	C (0.1861)	Leukocytes
chr5:149433596	T	A	hetero	CSF1R	-	-	3.30%	G (68.0297)	False call
chr17:7578475	G	C	hetero	TP53	4.80%	5.10%	5.00%	A (0.0032)	False call

Chapter 4

Discussion

4.1 | Immunomagnetic enrichment

There are a variety of CTC enrichment methods relying on many different CTC qualities and are carried out in two major ways: positive selection of CTCs and negative depletion of leukocytes. Positive selection selects specifically for CTCs from within the complex blood environment while negative depletion relies on the removal of non-CTC cells from the same environment. Positive selection commonly relies on epithelial surface markers (*EPCAM*, *MUC1*, *ERBB2*) for capture of CTCs, such as CellSearch, ^{pos}CTCiChip, AdnaTest, and other general immunomagnetic selection. Most negative depletion methods rely on the selection of leukocytes by CD45, but some have expanded the panel by including antibodies for CD15 [56] and CD66b [81] (both granulocyte markers). Negative depletion of leukocytes was implemented here in the form of the MINDEC method. This method was developed by Lapin and colleagues in this lab and based on the principle of high-coverage multi-marker depletion of leukocytes by immunomagnetic beads [87]. The advantages to its use are both in minimal bias of CTC recovery by not only selecting for epithelial phenotypes and maximal depletion of PBMCs by using a multi-marker antibody max targeting many blood cell types.

Before use in patient samples, the recovery of the MINDEC method was tested using a cell line spike in normal control blood. The recovery of the MINDEC method measured here was 78.6±0.36%. Lapin et al. obtained a mean recovery of 82±10% for the same method, using different cell lines [87]. This is comparable and slightly better than other immunomagnetic negative enrichment methods. Similar negative enrichment has been done by others using only CD45 and resulted in recoveries of 74.8+-9.3%[115], 58 and 69% [77]. Conversely, the ^{neg}CTC-iChip recovery has been demonstrated as 97.0% by Ozkumur *et al.* [56]. Furthermore, positive immunomagnetic enrichment has been implemented by Nadal *et al.* by multi-KRT specific beads with a recovery of 53.3-73.3% [70], and Riethdorf *et al.* by CellSearch for a recovery of 80 and 82% [116]. Liu *et al.* [77]

also investigated the recovery of EpCAM-positive immunomagnetic selection alongside CD45 depletion and measured a recovery of 25%. With similar positive immunomagnetic enrichment but the addition microfluidic separation, Ozkumur had a recovery of 77.8-98.6% [56] with the ^{pos}CTC-iChip. For further comparison, the positive selection of CTCs based on size by microfiltration exhibited a spike recovery of 82 and 88%[64].

The differences in the MINDEC method compared to others may be due to the effect of the enrichment method itself, the detection/counting method, and cell line used for spiking. The length of the procedure is a potential factor in cell recovery with the MINDEC method. There is a significant amount liquid discard and collection during the enrichment and this lends itself to many opportunities for cell loss in the process. Other methods with greater recovery may be due to a simpler methodological design such as the automated procedures of CellSearch [71] and the CTC-iChips [56]. For quantification of recovered spiked recovered cells, fluorescent tagging and subsequent counting by flow cytometry was implemented in our study. This adds several more steps to the original enrichment method and can be subjective (gating strategy), adding its own error and chance of cell loss. The flow cytometry method relied on the detection of EpCAM (cancer cells) and CD45 (PBMCs) positive cells and the events counted were in populations gated by comparison to controls. Populations were not clearly separated due to noise present in the samples. To overcome this, the gating was set by quadrants to reduce any bias that could have been caused by manually drawing boundaries around groups. The noise present with lower signal for both APC and FITC is thought to be due to platelets and background labeling, especially since they are small in size. Populations in that area can be seen in the unenriched blood sample as well, but is more pronounced in the enriched samples. The cells from the spike recovery were slightly higher in CD45 than the spike control. The reason behind this is unknown but is likely from interactions within the PBMC enrichment sample. Detection and enumeration of spiked cells in other methods also involve fluorescent activated cell sorting (FACS) by EpCAM, plus keratins in Liu *et al.* [77], and microscopic examination by immunocytological staining [64]. The cells used in the MINDEC spike and recovery in our study were ZR-75-1 cells. These are not used in other recovery experiments which used CRC lines HCT116 (98.6% *EPCAM* expression) [115], SW620 (>99% *EPCAM* and KRT) [77], breast cancer line SKBR3 (24-fold *EPCAM* expression over IgG) [56], multiple lines: MCF-7, SKBB3, MDA-MB 231 and T47D by Nadal et el.[70], and NCI-H2030 by Desitter *et al.* [64]. In the upcoming manuscript about MINDEC [87], other cell lines were analyzed and did yield variable results so this can be a factor when comapring methods.

Our recovery assessment is a decent measurement of the performance of the method, but it cannot reflect *in vivo* use. CTC recovery *in vivo* is impossible to determine as the total number of CTCs in circulation is unknown in each patient. Real sample recovery will vary much more than a controlled experiment using one cell line, considering the discussed heterogeneity in CTC phenotypes when compared to cell lines. However, less variation in recovery should occur when using a negative enrichment method as it is

based on selection and removal of normal WBCs and they are more phenotypically predictable than cancer cells.

4.2 | Multi-marker detection method

The method used for detection of enriched CTCs varies considerably between studies. Many are based on surface markers or other cytological features if using flow cytometry, immunostaining, or FISH for detection (as with CellSpotter, Cytospin, and other custom techniques). Cells are then counted by cell sorting or microscopic analysis. Alternatively, many enrichments are followed by qPCR for detection of CTCs by mRNA quantifications and consist of many different markers as this is still a rapidly evolving field and consensus has not been reached. The advantage of this method is its potential to characterize many different CTCs while avoiding subjectivity in classification of a CTC due to observed cellular features. The drawback is that the cells are not counted in the process.

Here, we used a multi-marker mRNA panel comprised of epithelial and EMT markers to cover a variety of CTC-subtypes and to allow for further characterization of the CTC population. The specific markers were chosen based on the documented function of their transcripts and their performance in preliminary testing. *EPCAM*, keratins, breast cancer specific markers, EMT transcription factors, and novel markers were investigated.

To arrive at this marker panel, a preliminary list of potential markers was created (as described in Sections 2.2.10.1 and 3.2) that included markers investigated in other studies and markers found to be differentialy expressed in breast and WBCs in the SAGE database. Among 13 other common markers, two new markers (*LUM* and *CCDC80*) were selected, based on this analysis. The selected markers were first validated by measuring their levels in cancer cell lines. From this, only one marker (*KRT16*) was excluded. Of all the markers, it exhibited the lowest expression across all four cell lines tested. In addition, it was one of four keratins considered, so after exclusion there still remained three others for testing (*KRT7*, *KRT8* and *KRT19*). *LUM* was kept for further study in the tissues despite the average low expression in cell lines. The expression of the markers among the different cell lines was variable and seemed to follow the subtypes of breast cancer that they reflect. By transcriptional profiling, the cell lines fall into the following subtypes: MCF-7 as luminal A, ZR-75-1 as luminal B, and MDA-MB-231 as claudin-low [117]. It comes as no surprise then that the most agressive cell line type, MDA-MB-231 (claudin-low or triple-negative), expressed EMT markers at a higher level together with the mesothelioma line, SDM103T2. MDA-MB-231 was also presented by Holliday *et al.* to express E-cadherin at a lower level, further supporting an EMT phenotype [117].

For further validation, the breast tumor samples and control blood samples from healthy volunteers were analyzed. This provided essential data for the final determination of the

multi-marker panel as the cell lines did not reflect real tumor heterogeneity. The breast tumors demonstrated this heterogeneity, with wide variations of expression seen for each marker (Figure 3.4). The differential expression among tissues and between markers can be seen in more detail in Appendix C. *KRT7* had overall lower expression in the tissues compared to the other keratins and was thus excluded. While *SNAIL*, *SLUG*, and *TWIST* did not look very promising in the tumor samples, they were included due to their documented EMT marker potential. These solid tumor samples do have limited value as they may not share the same properties as CTCs with regard to loss of epithelial expression.

Additionally, the background expression of the markers in normal blood was a very important point to consider if CTC expression was to be detected over normal expression. For some markers, the control expression in normal blood cells was too high to include them in the final panel (*TFF3* and *ZEB1*). The normal control blood samples in this early validation experiment came from 3 healthy persons. In hindsight, this was not a large enough sample to reflect the variability of background expression and use the expression for initial selection. In future screening experiments, many more samples should be analyzed for preliminary validation of marker expression levels in PBMCs. To note, a larger cohort of 30 healthy volunteers was recruited for the PBCB patient analysis.

With the final 10 markers chosen for analysis, there were a wide-range of characteristics and functions covered (Table 4.1). Many of the markers used already have extensive use in the CTC field. Most prevalent of course being *EPCAM*. This is due to its important function as a cellular adhesion molecule in epithelial cells. *KRT19* and *KRT8* are also commonly used due to the prevalence of keratins as CTC markers. The role of these two are in maintenance of cell structure and integrity [118]. Many other keratins exist and have complex expression patterns in both solid tumors [119] and CTCs [42], but we were limited in scope and chose to have a equal or greater focus on EMT markers. Mammoglobin A (*SCGB2A2*) and *HER2* (*ERBB2*) were were chosen for their high expression in breast tumors and clinical relevance, respectively. Mammoglobin A, a secretoglobin, is only expressed in the mammary gland and is often over-expressed in breast cancer tissue and cell lines [120]. It serves as a useful marker, but its function is largely unknown [120]. Possible roles of the secretoglobin protein include signalling, immune response, chemotaxis, and steroid hormone transport [121]. Due to the exclusively epithelial source, it was categorized as an epithelial marker. Since *ERBB2* is involved in epithelial processes (Table 4.1), it was also considered an epithelial marker. The role of *SNAIL*, *SLUG*, and *TWIST* as EMT-initiating trasncription factors solidified their use as EMT markers [40, 50, 122, 123], as well as extensive evidence of their expression in both CTCs and DTCs [45, 68, 78, 79, 93, 124].

The two new markers (*LUM* and *CCDC80*) were retained in the panel from results of the preliminary tests. Both had higher expression in the more mesenchymal-like cell

TABLE 4.1: Marker Gene Ontology, NCBI.[118] *Inferred from Electronic Annotation (IEA). **Traceable Author Statement (TAS). ***Non-traceable Author Statement (NAS). ****Inferred from Physical Interaction (IPI).

	Function	Process	Component
CCDC80	Fibronectin & heparin binding	ECM organization, positive regulation of cell-substrate adhesion	basement membrane, insterstitial matrix
EPCAM	cadherin binding involved cell-cell adhesion, protein & protein complex binding	Cell-cell adhesion*, negative regulation of apoptotic process* & cell-cell adhesion mediated by cadherin, positive regulation of cell motility* & proliferation	plasma membrane, cell surface*, extracellular exosome
ERBB2	protein tyrosine kinase activity, transmembrane signaling receptor actvity, ErbB-3 class receptor binding	positive regulation of MAP kinase activity, cell adhesion, cell growth, epithelial cell proliferation, protein targeting to membrane	plasma membrane, receptor complex, nucleus
KRT8	protein binding, structural molecule activity*	extrinsic apoptotic signaling pathway*, response to hydrostatic pressure* & other organism,* sarcomere organization*	cytoplasm, extracellular exosome, intermediate filament, nucleus
KRT19	protein binding, structural constituent of cytoskeleton, structural constituent of muscle	reponse to estrogen, sarcomere organization, Notch signaling pathway*	cell periphery, costamere, extracelular exosome, plasma membrane
LUM	collagen binding, ECM constituent, protein binding	collagen fibril organization***, ECM organization**	extracellular exosome, ECM colocalization, extracellular space, fibrillar collagen trimer
SCGB	protein binding****	ND	ND
SLUG	protein binding, sequence-specific DNA binding, trascriptional repressor activity	Notch & Wnt signalling pathway, cellular respone to EGF stimulus, desmosome disassembly, EMT	nucleus, nuclear chromatin
SNAIL	kinase binding, protein binding, transcriptional repressor activity	EMT, negative regulation of DNA damage response & celldifferentiation, positive regulation of cell migration	cytoplasm, nucleus
TWIST	E-box binding, sequence-specific DNA binding, protein bdining, TF binding	cellular rrespone to hypoxia, negative regulation of DNA damage respone, cellular senescence, histone phosphorylationm double-strand break repair	nucleus

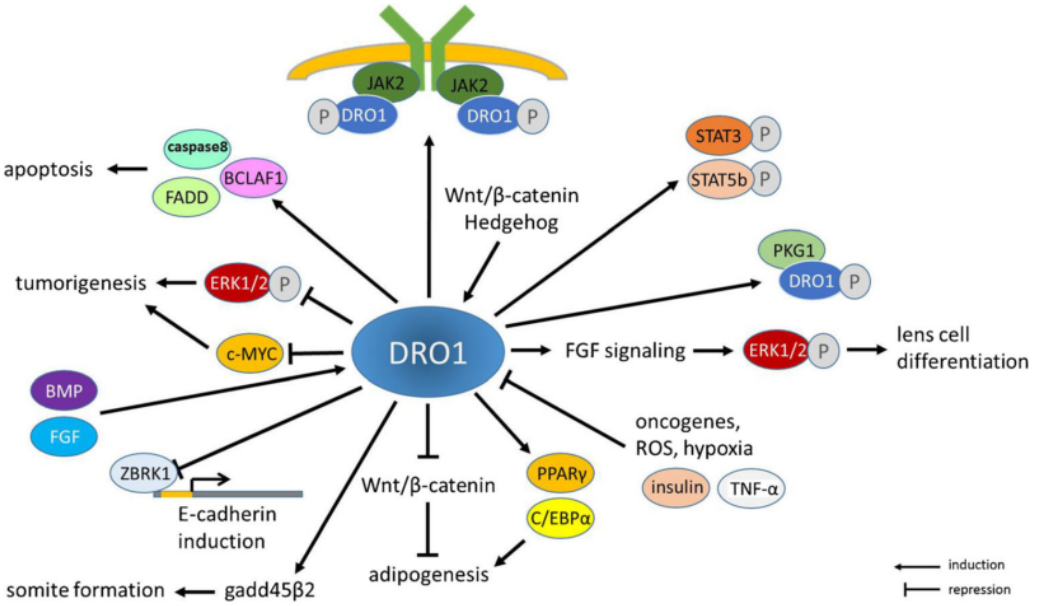

FIGURE 4.1: *CCDC80* (*DRO1*) molecular interactions. Reprinted with permission from Springer: Current Colorectal Cancer Reports, copyright 2015 [1]

lines and variable expression among the breast tumor samples, coupled with low control expression. They are novel in breast cancer CTCs with promising use as markers. The *CCDC80* (coiled-coil domain containing 80) gene codes for a presently uncharacterized protein involved in extra-cellular matrix (ECM) organization [118].

It has been implicated as a JAK2-binding protein[125], a downstream effector of the hedgehog pathway and fibronectin binding protein [126], involved in Wnt/β-catenin pathway [127] and adipogenesis [128] (Figure 4.2). It is considered by some to be a tumor suppressor as it has been found to promote cell adhesion, apoptosis, and E-cadherin expression in thyroid and colorectal cancer [1, 129]. On the other hand, *CCDC80* has also been shown to be over-expressed in response to estrogen with a potential carcinogenic role in the breast [130] and differentially expressed in single pancreatic CTCs along with other ECM genes (i.e. SPARC) [131]. The other ECM genes were investigated and were not found in the epithelial cells of the tumor, but expressed higher in CTCs and the stroma of tumors, colocalized with keratins at the epithelial-stromal border [131]. Lumican is encoded by *LUM* and is a protein that joins decorin, biglycan, fibromodulin, keratocan, epiphycan, and osteoglycin as a small, leucine-rich proteoglycan [118]. It has a well-established role in extracellular matrix assembly, specifically collagen fibril assembly and stability, and mediating cell-matrix interactions, cell migration, proliferation, tissue repair, and tumor growth [132–134]. The expression of lumican stromal cells has been associated with tumor invasiveness, progression, and shorter survival [134–139]. Inhibition of cancer growth by lumican has also been documented, however, in pancreatic cancer and melanoma [2, 140] and longer survival was documented with lumican-expressing cancer cells (versus stromal expression) [136]. The only investigation

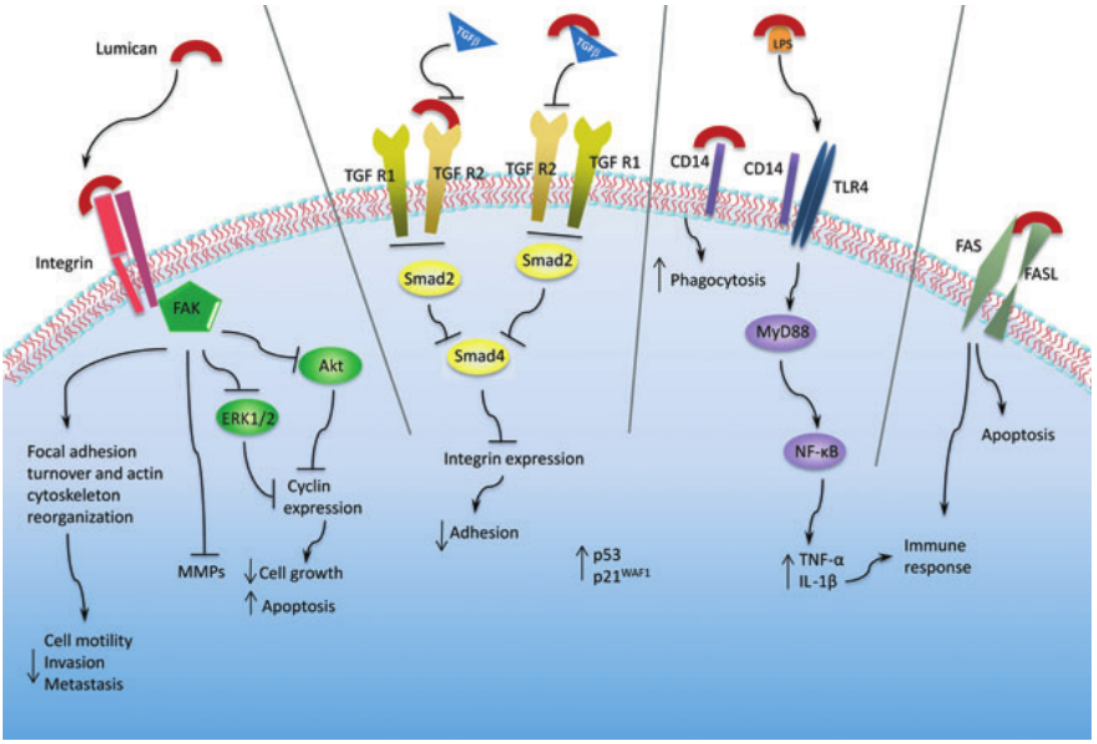

FIGURE 4.2: *LUM* molecular interactions. Reprinted with permission from John Wiley and Sons: FEBS Journal, copyright 2013[2]

into its role in breast cancer was by Panis *et al.* who found increased lumican expression in cells was associated with advanced disease [138].

Both genes serve functions in ECM organization and have lower expression in normal epithelial cells. Conversely, their higher expression in the stroma mean that excretion by fibroblasts or mesenchymal cells are their main source. This along with their associations with EMT-pathways and molecules (fibronectin, integrin, E-cadherin, Wnt/β-catenin, JAK/STAT3, TGF-β/AKT), and expression patterns in the cell lines, make them likely EMT markers. The research and findings are conflicting in both cases nonetheless and their use in this study is prospective. They may serve a role in EMT, but much more research is needed to substantiate that hypothesis and elucidate their role in breast cancer and CTCs specifically.

Multi-marker qPCR detection and characterization of CTCs is widespread in the field (Table 4.2) Multiple studies have shown the use of multiple markers is beneficial and confers greater sensitivity and detection over single marker analysis [60, 61, 76]. These panels have the potential to select for different CTCs, but can leave others undetected. Ideally, a set of markers would be found that could include all CTCs. As this is unverifiable, a panel with 100% sensitivity in regards to the metrics being measured would have to serve as proxy. These metrics could be based on basic cancer diagnosis or prognostic and response classifications.

TABLE 4.2: Genes used in other studies of multi-marker detection of CTCs. *Adna measures *EPCAM*, *MUC1*, *HER2*. **AdnaBC measures *GA733-2*, *MUC1*, *ERBB2*, β-actin. EMT, epithelial-mesenchymal transition. SC, stem cell.

Study	Marker Panel
Mikhitarian *et al.* 2008 [58]	*SCGB PIP CEA PSE KRT19 MUC1 EPCAM*
Aktas *et al.* 2009 [48]	Adna, EMT: *TWIST1*, *Akt2*, *PI3K*α; SC: *ALDH1*
Shen *et al.* 2009 [59]	Survivin, *hTERT* and *SCGB*
Obermayr *et al.* 2010 [60]	*CCNE2*, *DKFZp762E1312*, *EMP2*, *MAL2*, *PPIC* and *SLC6A8*; *EPCAM*, *SCGB*
Van der Auwera *et al.* 2010 [61]	Adna, *KRT19*, *MAM*
Markou *et al.* 2011 [66]	*KRT19*, *ERBB2*, *SCGB*, *MAGEA3*, *TWIST1*, *PBGB*
Molloy *et al.* 2011 [67]	*KRT19*, *p1B*, *EGP* and *SCGB*
Strati *et al.* 2011 [68]	*KRT19*, *MAGEA3*, *ERBB2*, *TWIST1*, *hTERT a+b+*, *SCGB*
De Albuquerque *et al.* 2011 [69]	*KRT19*, *SCGB*, *MUC1*, *EPCAM*, *BIRC5 ERBB2*
Giodrano *et al.* 2012 [78]	*TWIST1*, *SNAI1*, *ZEB1*, and *TG2*
Strati *et al.* 2013 [76]	*KRT19*, *ERBB2*, *MAGEA3*, *PBGB*, AdnaBC
Markiewicz *et al.* 2014 [79]	*KRT19*, *MGB1*, *VIM*, *TWIST1*, *SNAIL*, *SLUG*, *HER2*, *CXCR4* and *uPAR*
Vishnoi *et al.* 2015 [65]	83 genes in qPCR array
Kuniyoshi *et al.* 2015 [82]	*KRT19*, *ERBB2*, Oncotype genes

4.3 | Detection and characterization of CTCs in patient samples

4.3.1 | Detection rate of CTCs

The multi-marker panel decscribed above was used for the detection of CTCs in PBCB samples. A total of 133 patients were included in this study and CTCs were detected in 26.3% of patients (21.8% of all samples including multiple timepoints). This is a similar rate to other CTC studies in early breast cancer. Of the 27 selected studies described in the introduction, only 9 of them have included early breast cancer patients exclusively or in addition to metastatic disease. Of these, Molly *et al.* [67], Franken *et al.* [73], Lucci *et al.* [74], and Strati *et al.* [76] detected CTCs in 20%, 19%, 24%, and 14.2-22.8% (with three methods) of early breast cancer cases, respectively. The others obtained higher rates with Nadal *et al.*[70], Markiewicz *et al.* [79], and Kuniyoshi *et al.* [82] detecting CTCs in 46.9%, 41% and 69% (N0 and N+), and 55% and 77.6% (*KRT19* and *HER2* markers), respectively. These rates compare to detection rates ranging from 31% to over 80% in metastatic breast cancer [42, 48, 58, 60, 69, 75, 77]. A pooled analysis of studies studies including over 3000 patients done by Janni *et al.* [34] presented a detection rate of 20.2% in non-metastatic disease. In another analysis by Zhang *et al.* [52],

detection rates varied from 10-68% in nearly 7000 patients; this included both metastatic and non-metastatic disease and many different detection methods (ICC, RT-qPCR, and CellSearch). It is a representative sampling of the variability in CTC detection and includes a couple of studies mentioned here. Bidard *et al.* found 46.9% of metastatic patients had CTCs (≥5 per 7.5 mL blood) in an analysis of 20 studies [38].

Differences in detection rates can be attributed to many factors such as differential patient cohorts (early vs. progressed disease, low vs. high risk), sampling volume of blood, and both the enrichment and detection method used. Due to a larger source of tumor cells and also aggressiveness of disease, a higher detection rate is found in metastatic cancer. Because of the much lower amount of CTCs in ciriculation in early breast cancer, they are more difficult to detect. However, Fischer *et al.* have shown that even though they are are very rare, they may still be present in most patients and much higher blood volume is necessary for their capture [53]. Fischer *et al.* performed a comparison of CellSearch enumeration in leukaphoresis (LA) samples versus standard peripheral blood (PB) samples [53]. They detected CTCs in 91.7% of LA and 28% of PB samples. Terai *et al.* [102] found the femoral artery a better sampling site than the antecubital vein, finding CTCs in 100% of blood samples and in greater numbers, compared to 52.9% from the vein. However, this was in melonoma and may not be congruent in breast cancer.

It is difficult to compare results between different methods because they are measuring many different endpoints. This is clear when enumeration of CTCs is compared to other detection methods, but it also comes into play when you are comparing CTCs detected by cytological and gene expression methods. A CTC that expresses a specific protein may be detected by surface antigens and the mRNA expression of this gene may be investigated in another study. These results can be divergent because gene expression does not guarantee resulting protein expression. The differences between the gene and protein expression could be something to investigate in the future.

4.3.2 | CTC characteristics

The benefit of detection of CTCs by qPCR is the opportunity for further characterization of CTCs using the same data. With a larger marker panel, the information obtained has potential for interesting analysis of CTC features in each patient. Ideally, this can be correlated with clinical factors and outcomes. In this project, there were 35 patients positive for CTCs, with significant heterogeneity in marker expression. *LUM*, *CCDC80*, *KRT8* and *EPCAM* were the most prevalent markers with rates of 9.0%, 7.5%, 5.3% and 3.0% in all patients (34.3%, 27.0%, 20.0% and 11.4% among positive patients) The least represented among the group with only 1 or 2 samples were *ERBB2*, *KRT19*, *SCGB*, and *SNAIL*. With *CCDC80* and *LUM* included as EMT markers, 59.5% of samples were EMT-postive only. 24.3% of samples were epithelial-positive only, and 16.2% were

epithelial-EMT positive. Two samples that were positive for *CCDC80* on the first visit, lost that distinction upon the next visit, with one changing from *CCDC80* only to *KRT8* only, and the other from *CCDC80*/*EPCAM*+ to only *EPCAM*. For future analysis, it is useful to note that most positive samples could be characterized with just *CCDC80*, *LUM*, *EPCAM*, *KRT8*, and *TWIST*.

Several studies reporting the positivity of some of the same markers analyzed in our study report higher rates than we obtained. These rates ranged from about 25% for *EPCAM*, 25-46% in *KRT19*, 12.5-15.6% in *ERBB2*, 10-25% in SCGB2A2, and 31-42% in *TWIST1* [61, 66, 68, 69]. Similar findings to ours were presented by Obermayer *et al.* [60] where they detected *EPCAM* in just 5% of non-metastatic patients and Molloy *et al.* [67] finding 4.8% and 3.7% of early breast cancer patients positive for *KRT19* and *SCGB*, respectively. Obermayr used a density gradient (Oncoquick) for enrichment and analyzed the samples directly after by qPCR, whereas the other studies used positive selection of CTCs thus potentially imparting a bias to the collected population for the selection marker used (*EPCAM*, *KRT19*, *ERBB2*). Additionally, some of these other studies include metastatic patients which can increase the number of CTCs detected when compared to early cancer patients.

4.3.3 | Clinical associations

Current clinical potential lies in detection and enumeration of CTCs for improved treatment plans and outcomes in metastatic breast cancer. Large-scale pooled analyses of CTCs in breast cancer have found significant value in enumeration of CTCs in both metastatic and non-metastatic disease, with CTC-presence being an independent prognostic factor of progression-free survival and overall survival [34, 38, 52]. Prediction of survival was also improved by the addition of CTC-status at timepoints following treatment in the Bidard analysis [38]. With the known value then of CTC presence, more focus is now being given to the value of specific CTC-characteristics and their prognostic relevance. In Mikhitarian *et al.*, *MAM* positivity was associated with tumor grade, ER^- status, and high-risk patients [58]. Aktas *et al.* demonstrated the expression of EMT and stem cell markers in metstatic breast cancer, and that patients with these CTCs are also more likely to be non-responders to therapy [48]. The results using the three-maker panel by Shen *et al.* had a significant correlation with both TNM stage and lymph node metastasis [59]. The multi-marker panel by Molloy *et al.* was both significantly correlated and an independent predictor of relapse-free-survival [67]. Markiewicz *et al.* revealed that lymph node positive patients exhibited higher CTC number and specifically CTC expressing EMT markers [79].

In contrast, neither the presence of CTCs nor the specific CTC marker expression were significantly associated with differential clinical features in this study. However, while the data unfortunately yields no current predictive value for clinicopathological features,

survival analysis has yet to be made. More time following diagnosis needs to pass before that data is available. Also, given the low-risk cohort used here (55.7% Stage I, 30.5% Stage II), this lack of correlation is not surprising. Other studies of CTCs in breast cancer (that included early breast cancer cases) have also had the same lack of associations [66, 68, 70, 76, 82]. The absence of significant differences in patients exhibiting nodal spread and metastasis is also unsurprising given the lympatic versus hematogenous pathway to the nodes. In the Markiewicz study, this correlation seems to be marker specific since only CTCs positive for *MGB1* and *VIM* were independent predictors of nodal status [79].

The results in these studies may not transfer to CTC-characteristics in general, but may be specific to the marker and study design. The specific role and mechanism of each is largely unknown. A varied investigation into the characteristics of CTCs is an important step in this understanding. However, as more cells are detected and functionally characterized, most of them lack the aggressive profiles that would lead to invasion, proliferation, and metastasis [46]. It begs the question what specific roles these markers serve in circulating tumor cells and metastasis formation.

4.3.4 | Background expression and thresholds

The benefit of utilizing a negative enrichment method for CTCs is the analysis all CTCs regardless of surface markers. The disadvantage is that this leaves many PBMCs behind as well. Ozkumur found a log lower depletion of leukocytes in their negative enrichment chip compared to their positive enrichment chip due to reduced expression of CD45 in some leukocytes [56]. The challenege is then to detect CTCs among the normal cells. This is a problem in both cytological detection and by qPCR. Some of the markers considered here were excluded early due to control background expression. The larger control group used for PBCB analysis would have been more powerful for marker validation, as some markers used in the final panel did have higher background expression than expected. Only *SCGB* was 100% negative in the control group. To remedy this, the background expression of markers in PBMCs must be considered in the detection method.

This is a common problem faced across the field and one solution is to determine a threshold based on the control expression. In most studies using qPCR detection, a control set is also analyzed for marker expression, as has been done in this study. As with most other steps in the process, this a point of variability between studies as well. There is a balance when setting a threshold for obtaining high detection of real CTCs while avoiding false-positives. Here we used a threshold of the maximum control value, after removal of control outliers 3 standard deviations from the mean (99.7% confidence interval). The threshold set by the maximum control is common, but other have also used a 95% confidence interval and Molloy *et al.* implemented quadratic discriminant

analysis (QDA) to set a positive threshold (a statistical approach based on optimal separation of cohorts from a previous study) [67]. Some enrichment methods reduce the contaminating PBMCs so completely that positive expression for a marker is required for positive CTC-status.

Some markers in our multi-maker panel were very close to yielding no positives in the patient cohort (*SNAIL*, *SLUG*, *ERBB2*), and the background values of *EPCAM* and *KRT19* were also rather high. This was an unexpected result, since it is widely accepted that blood cells do not express eptithelial markers. While not found for *EPCAM*, other epithelial markers have been found expressed in PBMCs, however. Mikhitarian *et al.* found *KRT19*, *MUC1*, and *ERBB2* uninformative in their 2008 study due to competivitely high background levels [58]. Molloy *et al.* also reported no significant difference in QDA values between a control group and an early breast cancer cohort (from a 4-marker panel including *KRT19* and mammoglobin). The background *KRT19* levels could be due to illegitimate transcription by PBMCs or induction by cytokines as reviewed by Van der Auwera *et al.* [61] Obermayr *et al.* excluded *ERBB2* from their investigations due to the detectable levels in healthy controls [60]. You *et al.* has investigated the specific qualities of PBMCs that express epithelial markers (*HER2* and *KRT19*) at a low level and found that the main contributors to this expression were NK-cells/granulocutes (CD16) and lymphocytes (CD3/CD19), respectively [141]. They used antibodies specific to these cells for greater negative depletion of these popultions and found the depletion to increase in a dose-dependent manner with specific immunomagnetic bead addition. This could be useful for enhancement of the current MINDEC method. CD16 and CD19 antibodies are currently used in the immunomagnetic bead capture, but perhaps these amounts could be increased. It would also be interesting to investigate the properties of PBMC populations expressing *EPCAM* and other markers at low levels.

4.4 | CTC detection by sequencing

Sequencing of CTCs offers the ability to identify mutations present in the primary tumor as well as identify new and clinically relevant variants only present in CTCs. Similar genetic profiles in CTCs and primary tumors were documented by Alix-Panabieres *et al.* [101] and Heitzer *et al.* at the subclonal level [94]. Additionally, unique and cancer-relevant mutations in CTCs were found by Gold *et al.* [95] and Strauss *et al.* [57] Further divergence of CTCs from the primary tumor has been found with the presence of *HER2*+ CTCs in 27% of *HER2*- patients [69], which could point to mutational changes between the primary tumor and CTCs.

Patient samples were not analyzed here, but a methodological evaluation was carried out to measure the feasibility of the Ampliseq Cancer HotSpot Panel for detection of spiked cancer cells in normal blood. Two spiked normal samples (with 1000 and 100

cancer cells) were analyzed alongside a sample of only the spiking cells. The sequencing results demonstrate the limitation of the AmpliSeq method in CTC detection and characterization. Only 1000 cancer cells were detectable among an estimated 10,000 normal PBMCS (10%). A lower cell amount of 100 cancer cells fell just below the system's limit of 1% variant frequency [142] and seems as if it could be unreliable even as high as 3% frequency from the results presented here. This becomes an even larger problem when you consider trying to use this to detect only a few CTCs in one sample. Additionally, non-standard frequencies of variants (30-70%) were found in all samples with frequencies found outside the normal 100% and 50% for homo- and hetero-zygosity, respectively. Some of these may be due to sequencing quality due to location at the end of a fragment, but others were of higher quality and are perhaps explained by mosaicism, especially since it was more prevalent in the cancer cells. Cancer presents as an inherently diverse population, with a high propensity for mutation and clonal expansions [99]. The clonality of the cancer cell lines are prone to the same genomic changes [143]. The potential for mosaicism in healthy cells is less clear. However, the occurrence is becoming more and more noticeable with the increased use of sensitive methods like next generation sequencing [144, 145]. An error in variant calling is unlikely since the calling details were investigated in the BAM/VCF files, but there could be some other unidentified methodological reason behind the non-standard frequencies.

The Ampliseq method and HotSpot panel are limited firstly by their detection limits and secondly by their limited scope of targets. It is a panel made for use with many patient samples over many cancers and thus has a variety of targets to reflect that. It is not well-suited for investigation of novel mutations or more specific mutations in breast cancer. The commonly mutated regions in breast cancer included in the panel are in the *PIK3CA, PTEN, AKT1, TP53, CDH1, RB1* (6 out of 50) genes. This excludes *GATA3*, which with *PTEN* and *P53* are the only 3 mutations to occur in more than 10% of breast cancers [146], and *BRCA1/BRCA2*. More comprehensive and sensitive options for variant detection in CTCs exist that rely on unique identifiers for DNA template strands [147–149]. This is used for mutation-tracking on a strand by strand basis so errors not present on a whole uniquely identified template and its compliment can be excluded. Using this basis, duplex sequencing can achieve an error rate of less than 1 per 10^9 nucleotides and detection to $5x10^{-8}$, but is best used with short, targeted DNA regions (<1Mb) and ligation of the unique adapters has an efficiency of only 1-10% [147, 148]. The CAPP-Seq method is similar in that it also uses short and unique adapters for mutational analysis, but it is optimized for low amounts of template and also includes targeted library construction [149]. This limits the novel mutations to be found, but as long as the regions are well-selected, it should still yield many new findings. Further error in variant calling is reduced and detection can reach 0.01% by combining enhanced error suppression with the CAPP-Seq method for a complete integrated digital error suppression (iDES) technique [150]. This additional step is called polishing, and is

used to remove stereotypical errors in the sequencing by profiling background mutation errors across the analyzed locations [150].

These methods are described with use for cell-free DNA, but could be modified for use in CTCs. First, genomic DNA from CTCs must be extensively fragmented to acheive optimal lengths for library creation. Second, a detection method is needed to identify CTC-variants from the PBMC background. Detection of the mutations unique to CTCs should be easily distinguished from somatic mutations present in the PBMC pool just by frequency, as normal variants will occur at 50% or 100% frequencies and CTC variants will occur much lower. This could be filtered out during data analysis. Comparison with a matched leukocyte is another possibility, but CTCs missed during enrichment could be present in these samples. Another option is sequencing single CTCs to avoid background PBMCs, but there are advantages (specificity) and disadvantages (difficult collection, cost) there as well.

4.5 | Challenges and Future Perspectives

The primary challenges facing the detection of circulating tumor cells in cancer are the rarity of such cells in a sea of normal cells and their heterogenity, making it difficult to find a simple and effective way to capture all neoplastic cells in circulation. With the purpose of CTC diagnostics from a single vial of blood, this presents a problem in finding rare cells in such a small sample volume. Furthermore, when a CTC is detected, we are still unsure of what that means. It has been tied to worse prognosis and a link to metastasis is intuitive, but it is also known that there are patients in which CTCs are found and there is no resulting clinical effect.

This study was limited by some of the common challenges, such as a small sample volume, and detection of CTCs from a background of normal cells. To overcome this, an enhanced negative enrichment method (MINDEC) was used for unbaised and complete capture of CTCs and a multi-marker panel was used to detect the hetergeneous population. We do not know what cells were missed in the detection though, and the lack of clinical associations leave complete results to be determined when more data is obtained. More samples will be collected and analzyed as a part of the PBCB project (every 6 months for each patient) and the results from these later timepoints will be pooled for any further correlations and survival data.

The divergence of results here when compared to other studies highlight the problem of standardziation among CTC methods. Different methods in patient and control selection, collection and processing of samples, data interpretation, and CTC enrichment and detection all result in different outcomes. This is still a new and evolving field and best practices are yet to be decided. Until the research progresses and best practices and standards are reached, it will be difficult to translate CTCs into the clinical setting

[104]. Even with CellSearch, which has made it furthest in the translational journey, significant variation is present between settings and uses [52].

In the short-term, further progress can be made specific to this study. Additional work on a sequencing method for CTC characterization is necessary. The use of a molecular barcoding system of ultrasensitive variant detection is the most sensitive option. Work needs done in library optimization from the genomic CTC DNA, and how analysis could be combined with removal of leukocyte background and normal somatic mutations. Part of this could include collection of the PBMC fraction from the enrichment step for background comparison. This could also be used in relative expression analysis for validation of CTC expression in a selected group (as this is not realistic for every sample). CTCs may be present in these PBMC fractions, but likely not at a higher rate than the CTC fraction and if so, analysis of of both would be useful to understanding what kind of cells those are. Application of the multi-marker panel used here with other samples (DTCs in breast cancer, CTCs in pancreatic and colorectal cancers), with a focus on the new markers, will be important for further determination of their feasibility and clinical relevance in breast cancer and other cancer groups.

In the long-term, there are emerging options for the enhanced enrichment and detection of CTCs. Leukaphoresis could be used to overcome the challenge of both the small voume of typical bood samples and the removal of erythrocytes in one step. This has been shown to significantly increase the detection of CTCs in early breast cancer patients [53]. Additionally, an in vivo method for CTC collection by Gilupi (Gilupi CellCollector, www.gilupi.com) works by insertion into the arm similar to a normal venipuncture and left to isolate EpCAM-positive CTCs directly from the bloodstream for 30 minutes, acheiving 70% detection in early and late stage cancers. For further enrichment, negative selection is the best option considering the unbiased approach, but enhanced techniques could be utilized for the greater removal of leukocytes. A 5-antibody panel is used in this study to remove more leukocytes, but an investigation into more optimal ratios and what kind of WBCS are not captured would be useful for method optimization. Another avenue for enhanced enrichment is use of new nanotechnology. The use of microbeads for immunomagnetic collection and CTC-chips already benefits from nanotechnology integration, but these methods are improved and expounded upon in the newest research detailing graphene oxide films [151], gold nano slit microfulidic capture with simultaneous detection [152], and optimizing chip enrichment with microscale magnetic arrays [153]. In contrast, the enhancement of detection methods will not occur as much through methodological changes as much as the level of scale. Cytological assays are fairly standard and can not be enhanced greatly with the exception of identifying new surface markers. Gene expression analysis can benefit from better selection of markers to analyze, but the most promising way to enhance gene expression analysis is by high-throughput RNAseq. The cost of doing this is prohibitive to most labs, but this is the way forward to find new and better mRNA markers and other markers like noncoding RNAs (ncRNAs). ncRNAs include micro RNAs (miRNAs) and long intergenic

noncoding RNAs (lincRNAs) and can provide useful mutational profiles in cancer [25], possible therapeutic targets and diagnostic markers [101], and possible prognostic value [154]. qPCR analysis is limited with respect to number of transcripts analyzed at once, and this can be scaled up immenseley by large RNAseq panels or avoided completely by whole exome sequencing. With the addition of gene sequencing and even methylation analysis, whole patient cohorts can be grouped and characterized based on many CTC and tumor characteristics. Further power can be added by targeting single CTCs and clusters to find differential properties among CTCs themselves.

With the amount of information on CTC characteristics already accumulating, and the prospect of much more to come, an understanding of what it all means is necessary for estimating clinical utility. Relating CTC characterstisics to function is how to investigate the real impact of differential gene expression and mutations in CTCs *in vivo*. Establishment of CTC cell lines and xenograft assays are two ways CTCS can be tested for qualities such as drug sensitivities, response to drugs, and metastatic potential [37].

Enhanced prognostic and clinical power could be achieved by combination with other biomarkers such as cell-free or circulating tumor DNA (ctDNA). Circulating tumor DNA is found in the blood of cancer patients due to release as fragments from dead (necrotic and apoptotic) tumor cells. It is also referred to as cell-free DNA (cfDNA) since normal DNA fragments are present in the plasma as well. It only requires collection of blood plasma. For this reason, it has garnered parallel attention to CTCs and prognostic superiority over CTCs has been argued [155]. ctDNA has shown significant predictive value. [101, 156]. While ctDNA is thought to be only a reflection of the primary tumor due to their realted sequences, CTCs reflect qualities of the primary tumor and have also been found to diverge [93–95]. Additionally, ctDNA can only yield genetic information while CTCs offer the possibility of analysis of DNA, RNA, proteins, and functional assays. For this reason, a liquid biopsy of both ctDNA and CTCs could yield information about the primary tumor by both ctDNA and CTCs plus its potential for metastatic spread by CTC dissemination. ctDNA is also being collected from the plasma of every patient in the PBCB study and its investigation alongside CTCs should offer interesting biomarker comparison in future analysis. However, enhancement of enrichment and detection of both is necessary for the most clinical benefit.

Chapter 5

Conclusion

This study has established that early detection of CTCs in recently-diagnosed breast cancer patients is possible and that CTCs can exhibit both epithelial and EMT-related characteristics. Immunomagnetic depletion by the MINDEC method has a comparable recovery to other methods and allows for enrichment and detection of a heterogenous population of CTCs. Over a quarter of patients had detectable CTCs, with some expressing only EMT-related markers and only epithelial markers, and a few expressing both. This supports the idea that EPCAM or epithelial-only detection is severely limiting for CTC detection and characterization. Two novel markers were introduced (CCDC80 and LUM) and show great promise as breast cancer CTC markers with over half of the detected CTCs positive for those markers combined. They could be relevant markers in CTCs due to their functional characteristics, but their utility remains to be seen without further data and analysis. CTC-status was not significantly associated with clinicopathological features. Comparison with clinical follow-up data will allow for analysis on the prognostic relevance of our CTC measurements.

References

[1] J. I. Grill and F. T. Kolligs. DRO1/CCDC80 : a Novel Tumor Suppressor of Colorectal Carcinogenesis. pages 200–208, 2015. doi:10.1007/s11888-015-0276-3.

[2] S. Brézillon, K. Pietraszek, F. X. Maquart, and Y. Wegrowski. Lumican effects in the control of tumour progression and their links with metalloproteinases and integrins. *FEBS Journal*, 280(10):2369–2381, 2013. ISSN 1742464X. doi:10.1111/febs.12210.

[3] Breast Cancer Staging 7th Edition, 2009.

[4] T. Sorlie *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences of the United States of America*, 98(19):10869–74, 2001. ISSN 0027-8424. doi:10.1073/pnas.191367098. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=58566{&}tool=pmcentrez{&}rendertype=abstract.

[5] G. K. Malhotra, X. Zhao, H. Band, and V. Band. Histological, molecular and functional subtypes of breast cancers. *Cancer Biology and Therapy*, 10(10):955–960, 2010. ISSN 15384047. doi:10.4161/cbt.10.10.13879.

[6] ThermoFisher. Single Tube TaqMan® Gene Expression Assays, 2015. URL https://www.thermofisher.com/no/en/home/life-science/pcr/real-time-pcr/real-time-pcr-assays/taqman-gene-expression/single-tube-taqman-gene-expression-analysis.html.

[7] M. Ervik *et al.* Cancer Today, 2016. URL http://gco.iarc.fr/today.

[8] R. Siegel, K. Miller, and A. Jemal. Cancer statistics , 2015 . *CA Cancer J Clin*, 65(1):29, 2015. ISSN 1542-4863. doi:10.3322/caac.21254. URL http://onlinelibrary.wiley.com/doi/10.3322/caac.21254/pdf.

[9] I. K. Larsen, editor. *Cancer in Norway 2014 - Cancer incidence, mortality, survival and prevalence in Norway*. Oslo: Cancer Registry of Norway, 2015. ISBN 9785290343914. doi:10.1136/bmj.1.5178.1031-a. URL http://www.bmj.com/cgi/doi/10.1136/bmj.1.5178.1031-a.

[10] C. I. Szabo, M. C. King, and M. C. K. C I Szabo. Population genetics of BRCA1 and BRCA2. *American journal of human genetics*, 60(5):1013–20, 1997. ISSN 0002-9297. doi:papers2://publication/uuid/3E108D68-23A1-44A7-9F56-3D1C35A0FF2D. URL file:///Users/Claustrum/Dropbox/Papers2/Articles/1997/C{_}I{_}Szabo/C{_}I{_}Szabo{_}American{_}Journal{_}of{_}Human{_}Genetics{_}1.pdf$\delimiter"026E30F$nhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1712447{&}tool=pmcentrez{&}rendertype=abstract.

[11] K. J. Ruddy and E. P. Winer. Male breast cancer: risk factors, biology, diagnosis, treatment, and survivorship. *Annals of Oncology*, 24(6):1434–1443, 2013. ISSN 0923-7534, 1569-8041. doi:10.1093/annonc/mdt025.

[12] V. Kumar, A. K. Abbas, J. C. Aster, and S. Robbins. *Robbins basic pathology*. Philadelphia: Elsevier Saunders, 9th ed., 2013. ISBN 978-1-4377-1781-5.

[13] R. Janavičius. Founder BRCA1/2 mutations in the Europe: Implications for hereditary breast-ovarian cancer prevention and control. *EPMA Journal*, 1(3):397–412, 2010. ISSN 18785077. doi:10.1007/s13167-010-0037-y.

[14] W. Demark-wahnefried, E. V. Bandera, S. Gapstur, and A. V. Patel. American Cancer Society Guidelines on Nutrition and Physical Activity for Cancer Prevention Reducing the Risk of Cancer With Healthy Food Choices and Physical Activity. *CA: a cancer journal for clinicians*, 62:30–67, 2012. ISSN 1542-4863. doi:10.3322/caac.20140.Available.

[15] T. E. Robsahm *et al.* Nasjonalt handlingsprogram med retningslinjer for diagnostikk , behandling og oppfølging av maligne melanomer. 2011. URL http://www.helsedirektoratet.no/publikasjoner/nasjonalt-handlingsprogram-med-retningslinjer-for-diagnostikk-behandling-og-oppfolging-av-maligne-Publikasjoner/IS-1860{_}Maligne-melanomer.pdf.

[16] E. Marshall. Dare to Do Less. *Science*, 343(6178):1454–1456, 2014.

[17] N. C. Institute. Tumor Grade Fact Sheet. URL http://www.cancer.gov/about-cancer/diagnosis-staging/prognosis/tumor-grade-fact-sheet.

[18] T. Scholzen and J. Gerdes. The Ki-67 protein: From the known and the unknown. *Journal of Cellular Physiology*, 182(3):311–322, 2000. ISSN 00219541. doi:10.1002/(SICI)1097-4652(200003)182:3⟨311::AID-JCP1⟩3.0.CO;2-9.

[19] M. Gnant, N. Harbeck, and C. Thomssen. St. Gallen 2011: Summary of the consensus discussion. *Breast Care*, 6(2):136–141, 2011. ISSN 16613791. doi:10.1159/000328054.

[20] I. Gingras *et al.* The current use and attitudes towards tumor genome sequencing in breast cancer. *Scientific Reports*, 6(October 2015):22517, 2016. ISSN 2045-2322. doi:10.1038/srep22517. URL http://www.nature.com/articles/srep22517.

[21] D. F. Easton *et al.* Gene-Panel Sequencing and the Prediction of Breast-Cancer Risk. *The new england journal of medicine*, 342(23), 2015.

[22] A. W. Kurian *et al.* Clinical evaluation of a multiple-gene sequencing panel for hereditary cancer risk assessment. *Journal of Clinical Oncology*, 32(19):2001–2009, 2014. ISSN 15277755. doi:10.1200/JCO.2013.53.6607.

[23] W. Sieh, J. H. Rothstein, V. McGuire, and Alice S. Whittemore. The Role of Genome Sequencing in Personalized Breast Cancer Prevention. *Cancer Epidemiol Biomarkers Prev*, 23(11):2322–2327, 2014. ISSN 15378276. doi:10.1126/scisignal.2001449.Engineering.

[24] E. H. Lips *et al.* Next generation sequencing of triple negative breast cancer to find predictors for chemotherapy response. *Breast Cancer Research*, 17(1):134, 2015. ISSN 1465-542X. doi:10.1186/s13058-015-0642-8. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4592753{&}tool=pmcentrez{&}rendertype=abstract.

[25] S. Nik-Zainal *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, 534(7605):1–20, 2016. ISSN 0028-0836. doi:10.1038/nature17676. URL http://www.nature.com/doifinder/10.1038/nature17676.

[26] R. A. Weinberg. *The biology of cancer*. New York: Garland Science, 2nd ed., 2014. ISBN 978-0-8153-4220-5.

[27] E. Miller *et al.* Current treatment of early breast cancer: adjuvant and neoadjuvant therapy. *F1000Research*, 3(0):198, 2014. ISSN 2046-1402. doi:10.12688/f1000research.4340.1. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4224200{&}tool=pmcentrez{&}rendertype=abstract.

[28] F. B. De Abreu, G. N. Schwartz, W. A. Wells, and G. J. Tsongalis. Personalized therapy for breast cancer. *Clinical Genetics*, 86(1):62–67, 2014. ISSN 13990004. doi:10.1111/cge.12381.

[29] M. S. Wicha and D. F. Hayes. Circulating Tumor Cells: Not All Detected Cells Are Bad and Not All Bad Cells Are Detected. *Journal of Clinical Oncology*, 29(12):1506–1508, 2011. ISSN 0732183X. doi:10.1200/JCO.2010.34.0026.

[30] J. E. Talmadge and I. J. Fidler. AACR centennial series: The biology of cancer metastasis: Historical perspective. *Cancer Research*, 70(14):5649–5669, 2010. ISSN 00085472. doi:10.1158/0008-5472.CAN-10-1040.

[31] T. R. Ashworth. A case of cancer in which cells similar to those in the Tumours were seen in the blood after death. *Australian Med J*, 14(5):146–147, 1869.

[32] E. Pool and G. Dunlop. CANCER CELLS IN THE BLOOD STREAM. *Am J Cancer*, pages 99–103, 1934.

[33] PubMed. URL http://www.ncbi.nlm.nih.gov/pubmed/.

[34] W. Janni *et al.* Pooled Analysis of the Prognostic Relevance of Circulating Tumor Cells in Primary Breast Cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research*, pages 1–12, 2016. ISSN 1078-0432. doi:10.1158/1078-0432.CCR-15-1603. URL http://www.ncbi.nlm.nih.gov/pubmed/26733614.

[35] B. Rack *et al.* Circulating tumor cells predict survival in early average-to-high risk breast cancer patients. *Journal of the National Cancer Institute*, 106(5):1–11, 2014. ISSN 14602105. doi:10.1093/jnci/dju066.

[36] A. Toss, Z. Mu, S. Fernandez, and M. Cristofanilli. CTC enumeration and characterization : moving toward personalized medicine. 2(11), 2014. doi:10.3978/j.issn.2305-5839.2014.09.06.

[37] C. Alix-Panabières, K. Bartkowiak, and K. Pantel. Functional studies on circulating and disseminated tumor cells in carcinoma patients. *Molecular Oncology*, (January):1–7, 2016. ISSN 15747891. doi:10.1016/j.molonc.2016.01.004. URL http://linkinghub.elsevier.com/retrieve/pii/S1574789116000144.

[38] F.-C. Bidard *et al.* Clinical validity of circulating tumour cells in patients with metastatic breast cancer: a pooled analysis of individual patient data. *The lancet oncology*, 15(4):406–14, 2014. ISSN 1474-5488. doi:10.1016/S1470-2045(14)70069-5. URL http://www.ncbi.nlm.nih.gov/pubmed/24636208.

[39] J. Shaw Bagnall *et al.* Deformability of Tumor Cells versus Blood Cells. *Scientific reports*, 5(November):18542, 2015. ISSN 2045-2322. doi:10.1038/srep18542. URL http://www.nature.com/srep/2015/151218/srep18542/full/srep18542.html.

[40] S. MJ and G. puche JL. Circulating Tumor Cells (CTCs): From Detection to Dissection. *JBR Journal of Clinical Diagnosis and Research*, 03(01):1–3, 2015. ISSN 23760311. doi:10.4172/2376-0311.1000120. URL http://www.omicsonline.com/open-access/circulating-tumor-cells-ctcs-from-detection-to-dissection-2376-0311-1000120.php?aid=62550.

[41] D. Schilling *et al.* Isolated, disseminated and circulating tumour cells in prostate cancer. *Nature reviews. Urology*, 9(8):448–63, 2012. ISSN 1759-4820. doi:10.1038/nrurol.2012.136. URL http://dx.doi.org/10.1038/nrurol.2012.136.

[42] S. a. Joosse *et al.* Changes in keratin expression during metastatic progression of breast cancer: Impact on the detection of circulating tumor cells. *Clinical Cancer Research*, 18(4):993–1003, 2012. ISSN 10780432. doi:10.1158/1078-0432.CCR-11-2100.

[43] X. Ye and R. A. Weinberg. EpithelialMesenchymal Plasticity: A Central Regulator of Cancer Progression. *Trends in Cell Biology*, xx(x):1–12, 2015. ISSN 09628924. doi:10.1016/j.tcb.2015.07.012. URL http://linkinghub.elsevier.com/retrieve/pii/S0962892415001452.

[44] W. L. Tam and R. A. Weinberg. The epigenetics of epithelial-mesenchymal plasticity in cancer. *Nature medicine*, 19(11):1438–49, 2013. ISSN 1546-170X. doi:10.1038/nm.3336. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4190672{&}tool=pmcentrez{&}rendertype=abstract.

[45] N. Bednarz-Knoll, C. Alix-Panabières, and K. Pantel. Clinical relevance and biology of circulating tumor cells. *Breast Cancer Research*, 13(6):228, 2011. ISSN 1465-5411. doi:10.1186/bcr2940.

[46] I. Baccelli *et al.* Identification of a population of blood circulating tumor cells from breast cancer patients that initiates metastasis in a xenograft assay. *Nature biotechnology*, 31(6):539–44, 2013. URL http://www.nature.com.ez.srv.meduniwien.ac.at/nbt/journal/v31/n6/full/nbt.2576.html.

[47] H. Iinuma *et al.* Clinical significance of circulating tumor cells, including cancer stem-like cells, in peripheral blood for recurrence and prognosis in patients with dukes' stage B and C colorectal cancer. *Journal of Clinical Oncology*, 29(12):1547–1555, 2011. ISSN 0732183X. doi:10.1200/JCO.2010.30.5151.

[48] B. Aktas *et al.* Stem cell and epithelial-mesenchymal transition markers are frequently overexpressed in circulating tumor cells of metastatic breast cancer patients. *Breast cancer research : BCR*, 11(4):R46, 2009. ISSN 1465-542X. doi:10.1186/bcr2333. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2750105{&}tool=pmcentrez{&}rendertype=abstract.

[49] S. S. McAllister and R. a. Weinberg. The tumour-induced systemic environment as a critical regulator of cancer progression and metastasis. *Nature cell biology*, 16(8):717–27, 2014. ISSN 1476-4679. doi:10.1038/ncb3015. URL http://www.nature.com/doifinder/10.1038/ncb3015$\delimiter"026E30F$nhttp://www.ncbi.nlm.nih.gov/pubmed/25082194.

[50] M. G. Krebs *et al.* Molecular analysis of circulating tumour cells-biology and biomarkers. *Nature reviews. Clinical oncology*, 11(3):129–144, 2014. ISSN 1759-4774. doi:10.1038/nrclinonc.2013.253. URL http://www.ncbi.nlm.nih.gov/pubmed/24445517$\delimiter"026E30F$nhttp://www.nature.com/doifinder/10.1038/nrclinonc.2013.253.

[51] K. Pantel and M. R. Speicher. The biology of circulating tumor cells. *Oncogene*, (February):1–9, 2015. ISSN 1476-5594. doi:10.1038/onc.2015.192. URL http://www.ncbi.nlm.nih.gov/pubmed/26050619.

[52] L. Zhang *et al.* Meta-analysis of the prognostic value of circulating tumor cells in breast cancer. *Clinical Cancer Research*, 18(20):5701–5710, 2012. ISSN 10780432. doi:10.1158/1078-0432.CCR-12-1587.

[53] J. C. Fischer *et al.* Diagnostic leukapheresis enables reliable detection of circulating tumor cells of nonmetastatic cancer patients. *Proceedings of the National Academy of Sciences of the United States of America*, 110(41):16580–5, 2013. URL http://www.pnas.org/content/110/41/16580.long.

[54] R. A. Ghossein, S. Bhattacharya, and J. Rosai. Molecular detection of micrometastases and circulating tumor cells in solid tumors. *Clin.Cancer Res.*, 5(1078-0432 SB - IM):1950–1960, 1999.

[55] E. Racila *et al.* Detection and characterization of carcinoma cells in the blood. *Proceedings of the National Academy of Sciences of the United States of America*, 95(April):4589–94, 1998. ISSN 0027-8424. doi:10.1073/pnas.95.8.4589. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=22534{&}tool=pmcentrez{&}rendertype=abstract.

[56] E. Ozkumur *et al.* Inertial Focusing for Tumor Antigen-Dependent and -Independent Sorting of Rare Circulating Tumor Cells. *Science Translational Medicine*, 5(179):179ra47–179ra47, 2013. ISSN 1946-6234. doi:10.1126/scitranslmed.3005616. URL http://stm.sciencemag.org/cgi/doi/10.1126/scitranslmed.3005616.

[57] W. Strauss, J. Winer-Jones, L. Austin, P. Dempsey, and M. Cristofanilli. The LiquidBiopsy in metastatic breast cancer (MBC): A novel diagnostic platform for next generation sequencing (NGS) of circulating tumor cells (CTCs). *Cancer Research*, 75:abstract, 2014. doi:10.1158/1538-7445.SABCS14-P5-10-07.

[58] K. Mikhitarian *et al.* Detection of mammaglobin mRNA in peripheral blood is associated with high grade breast cancer: interim results of a prospective cohort study. *BMC cancer*, 8:55, 2008. ISSN 1471-2407. doi:10.1186/1471-2407-8-55.

[59] C. Shen, L. Hu, L. Xia, and Y. Li. The detection of circulating tumor cells of breast cancer patients by using multimarker (Survivin, hTERT and hMAM) quantitative real-time PCR. *Clinical biochemistry*, 42(3):194–200, 2009. ISSN 1873-2933. doi:10.1016/j.clinbiochem.2008.10.016. URL http://dx.doi.org/10.1016/j.clinbiochem.2008.10.016.

[60] E. Obermayr *et al.* Assessment of a six gene panel for the molecular detection of circulating tumor cells in the blood of female cancer patients. *BMC cancer*, 10(1):666, 2010. ISSN 1471-2407. doi:10.1186/1471-2407-10-666. URL http://www.biomedcentral.com/1471-2407/10/666.

[61] I. Van der Auwera *et al.* Circulating tumour cell detection: a direct comparison between the CellSearch System, the AdnaTest and CK-19/mammaglobin RT-PCR in patients with metastatic breast cancer. *British journal of cancer*, 102(2):276–284, 2010. ISSN 1532-1827. doi:10.1038/sj.bjc.6605472.

[62] F. Farace *et al.* A direct comparison of CellSearch and ISET for circulating tumour-cell detection in patients with metastatic carcinomas. *British journal of cancer*, 105(6):847–53, 2011. ISSN 1532-1827. doi:10.1038/bjc.2011.294. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3171010{&}tool=pmcentrez{&}rendertype=abstract.

[63] G. Hvichia *et al.* A novel microfluidic platform for size and deformability based separation and the subsequent molecular characterization of viable circulating tumor cells. *International Journal of Cancer*, 00:n/a–n/a, 2016. ISSN 00207136. doi:10.1002/ijc.30007. URL http://doi.wiley.com/10.1002/ijc.30007.

[64] I. Desitter *et al.* A new device for rapid isolation by size and characterization of rare circulating tumor cells. *Anticancer Research*, 31(2):427–441, 2011. ISSN 02507005. doi:31/2/427[pii].

[65] M. Vishnoi *et al.* The isolation and characterization of CTC subsets related to breast cancer dormancy. *Nature Publishing Group*, pages 1–14, 2015. doi:10.1038/srep17533. URL http://dx.doi.org/10.1038/srep17533.

[66] A. Markou, A. Strati, N. Malamos, V. Georgoulias, and E. S. Lianidou. Molecular characterization of circulating tumor cells in breast cancer by a liquid bead array hybridization assay. *Clinical Chemistry*, 57(3):421–430, 2011. ISSN 00099147. doi:10.1373/clinchem.2010.154328.

[67] T. J. Molloy *et al.* A multimarker QPCR-based platform for the detection of circulating tumour cells in patients with early-stage breast cancer. *British journal of cancer*, 104(12):1913–1919, 2011. ISSN 1532-1827. doi:10.1038/bjc.2011.164.

[68] A. Strati *et al.* Gene expression profile of circulating tumor cells in breast cancer by RT-qPCR. *BMC Cancer*, 11(1):422, 2011. ISSN 1471-2407. doi:10.1186/1471-2407-11-422. URL http://www.biomedcentral.com/1471-2407/11/422.

[69] A. De Albuquerque, S. Kaul, G. Breier, P. Krabisch, and N. Fersis. Multimarker analysis of circulating tumor cells in peripheral blood of metastatic breast cancer patients: A step forward in personalized medicine. *Breast Care*, 7(1):7–12, 2012. ISSN 16613791. doi:10.1159/000336548.

[70] R. M. Nadal *et al.* Biomarkers Characterization of Circulating Tumour Cells in Breast Cancer Patients. *Breast Cancer Research*, 14(3):R71, 2012. ISSN 1465-5411. doi:10.1186/bcr3180. URL http://breast-cancer-research.com/content/14/3/R71.

[71] M. Cristofanilli *et al.* Circulating tumor cells, disease progression, and survival in metastatic breast cancer. *The New England journal of medicine*, 351(8):781–791, 2004. ISSN 1533-4406. doi:10.1056/NEJMoa040766.

[72] D. F. Hayes. Circulating Tumor Cells at Each Follow-up Time Point during Therapy of Metastatic Breast Cancer Patients Predict Progression-Free and Overall Survival. *Clinical Cancer Research*, 12(14):4218–4224, 2006. ISSN 1078-0432. doi:10.1158/1078-0432.CCR-05-2821. URL http://clincancerres.aacrjournals.org/cgi/doi/10.1158/1078-0432.CCR-05-2821.

[73] B. Franken *et al.* Circulating tumor cells, disease recurrence and survival in newly diagnosed breast cancer. *Breast Cancer Research*, 14(5):R133, 2012. ISSN 1465-5411. doi:10.1186/bcr3333. URL http://breast-cancer-research.com/content/14/5/R133.

[74] A. Lucci *et al.* Circulating tumour cells in non-metastatic breast cancer: A prospective study. *The Lancet Oncology*, 13(7):688–695, 2012. ISSN 14702045. doi:10.1016/S1470-2045(12)70209-7. URL http://dx.doi.org/10.1016/S1470-2045(12)70209-7.

[75] Y. Shiomi-Mouri *et al.* Clinical significance of circulating tumor cells (CTCs) with respect to optimal cut-off value and tumor markers in advanced/metastatic breast cancer. *Breast Cancer*, 2014. ISSN 13406868. doi:10.1007/s12282-014-0539-x.

[76] A. Strati, S. Kazimir-Bauer, A. Markou, C. Parisi, and E. S. Lianidou. Comparison of three molecular assays for the detection and molecular characterization of circulating tumor cells in breast cancer. *Breast cancer research : BCR*, 15(2):R20, 2013. ISSN 1465-542X. doi:10.1186/bcr3395. URL http://www.ncbi.nlm.nih.gov/pubmed/23497487.

[77] Z. Liu *et al.* Negative enrichment by immunomagnetic nanobeads for unbiased characterization of circulating tumor cells from peripheral blood of cancer patients. *Journal of translational medicine*, 9(1):70, 2011. ISSN 1479-5876. doi:10.1186/1479-5876-9-70. URL http://www.translational-medicine.com/content/9/1/70.

[78] a. Giordano *et al.* Epithelial-Mesenchymal Transition and Stem Cell Markers in Patients with HER2-Positive Metastatic Breast Cancer. *Molecular Cancer Therapeutics*, 11(November):2526–2535, 2012. ISSN 1535-7163. doi:10.1158/1535-7163.MCT-12-0460.

[79] A. Markiewicz *et al.* Mesenchymal phenotype of CTC-enriched blood fraction and lymph node metastasis formation potential. *PLoS ONE*, 9(4), 2014. ISSN 19326203. doi:10.1371/journal.pone.0093901.

[80] M. Yu *et al.* Ex vivo culture of circulating breast tumor cells for individualized testing of Drug Susceptibility. *Science (New York, N.Y.)*, 345(6193):216–220, 2014. doi:10.1126/science.1253533. Ex. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4358808{&}tool=pmcentrez{&}rendertype=abstract.

[81] N. Aceto *et al.* Circulating Tumor Cell Clusters Are Oligoclonal Precursors of Breast Cancer Metastasis. *Cell*, 158(5):1110–1122, 2014. ISSN 00928674. doi:10.1016/j.cell.2014.07.013. URL http://www.cell.com/article/S0092867414009271/fulltext.

[82] R. K. Kuniyoshi *et al.* Gene profiling and circulating tumor cells as biomarker to prognostic of patients with locoregional breast cancer. *Tumor Biology*, 36(10):8075–8083, 2015. ISSN 14230380. doi:10.1007/s13277-015-3529-5.

[83] Evaluation of Automatic Class III Designation CellSearch Epithelial Cell Kit/Cell Spotter Analyzer: 21 CFR 866.6020, 2004.

[84] F.-C. Bidard *et al.* Clinical application of circulating tumor cells in breast cancer: overview of the current interventional trials. *Cancer metastasis reviews*, 32(1-2):179–88, 2013. ISSN 1573-7233. doi:10.1007/s10555-012-9398-0. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3655223{&}tool=pmcentrez{&}rendertype=abstract.

[85] P. K. Grover, a. G. Cummins, T. J. Price, I. C. Roberts-Thomson, and J. E. Hardingham. Circulating tumour cells: The evolving concept and the inadequacy of their enrichment by EpCAM-based methodology for basic and clinical cancer research. *Annals of Oncology*, 25(8):1506–1516, 2014. ISSN 15698041. doi:10.1093/annonc/mdu018.

[86] M. L. Hermiston, Z. Xu, and A. Weiss. CD45: a critical regulator of signaling thresholds in immune cells. *Annual review of immunology*, 21:107–137, 2003. ISSN 0732-0582. doi:10.1146/annurev.immunol.21.120601.140946.

[87] M. Lapin *et al.* An Enhanced Negative Depletion Strategy for Circulating Tumour Cell Enrichment. Manuscript under review in Scientific Reports. *Scientific Reports*.

[88] M. T. Gabriel, L. R. Calleja, A. Chalopin, B. Ory, and D. Heymann. Circulating Tumor Cells: A Review of Non-EpCAM-Based Approaches for Cell Enrichment and Isolation. *Clinical Chemistry*, 000, 2016. ISSN 0009-9147. doi:10.1373/clinchem.2015.249706. URL http://www.clinchem.org/cgi/doi/10.1373/clinchem.2015.249706.

[89] S. Grünert, M. Jechlinger, and H. Beug. Diverse cellular and molecular mechanisms contribute to epithelial plasticity and metastasis. *Nature reviews. Molecular cell biology*, 4(8):657–665, 2003. ISSN 1471-0072. doi:10.1038/nrm1175.

[90] C. Alix-panabieres. Minimal Residual Disease and Circulating Tumor Cells in Breast Cancer. *Recent Results in Cancer Research*, 195:69–76, 2012. ISSN 0043-5341. doi:10.1007/978-3-642-28160-0. URL http://www.springerlink.com/index/10.1007/978-3-642-28160-0.

[91] AdnaGen. How does the AdnaTest work? URL http://www.adnagen.com/cfscripts/main{_}technology{_}application.cfm?auswahl=01.25.10.

[92] S. A. Bustin *et al.* Special Report The MIQE Guidelines:. *Clinical Chemistry*, 55(4):611–622, 2009. doi:10.1373/clinchem.2008.112797.

[93] C. Alix-Panabières and K. Pantel. Challenges in circulating tumour cell research. *Nature reviews. Cancer*, 14(3):623, 2014. ISSN 1474-1768. doi:10.1038/nrc3686. URL http://www.ncbi.nlm.nih.gov/pubmed/24522844.

[94] E. Heitzer, M. Auer, P. Ulz, J. B. Geigl, and M. R. Speicher. Circulating tumor cells and DNA as liquid biopsies. *Genome Medicine 2013,*, 5(73):1–11, 2013.

[95] B. Gold, M. Cankovic, L. V. Furtado, F. Meier, and C. D. Gocke. Do circulating tumor cells, exosomes, and circulating tumor nucleic acids have clinical utility?: A report of the association for molecular pathology. *Journal of Molecular Diagnostics*, 17(3):209–224, 2015. ISSN 19437811. doi:10.1016/j.jmoldx.2015.02.001. URL http://dx.doi.org/10.1016/j.jmoldx.2015.02.001.

[96] Biomarkers In Risk Assessment: Validity And Validation (EHC 222, 2001), 2001. URL http://www.inchem.org/documents/ehc/ehc/ehc222.htm.

[97] L. R. Yates *et al.* Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nature medicine*, 21(7):751–759, 2015. ISSN 1546-170X. doi: 10.1038/nm.3886. URL http://www.nature.com/nm/journal/v21/n7/full/nm.3886.html?WT.ec{_}id=NM-201507{&}spMailingID=49045495{&}spUserID=MTEwMjUwNjE0MzgxS0{&}spJobID=720844904{&}spReportId=NzIwODQ0OTA0S0.

[98] H. Seol *et al.* Intratumoral heterogeneity of HER2 gene amplification in breast cancer: its clinicopathological significance. *Modern Pathology*, 25(7):938–948, 2012. ISSN 0893-3952. doi: 10.1038/modpathol.2012.36. URL http://dx.doi.org/10.1038/modpathol.2012.36.

[99] S. Nik-Zainal *et al.* The life history of 21 breast cancers. *Cell*, 149(5):994–1007, 2012. ISSN 00928674. doi:10.1016/j.cell.2012.04.023.

[100] J. Li *et al.* Targeted drug delivery to circulating tumor cells via platelet membrane-functionalized particles. *Biomaterials*, 2015. ISSN 01429612. doi:10.1016/j.biomaterials.2015.10.046. URL http://linkinghub.elsevier.com/retrieve/pii/S014296121500856X.

[101] C. Alix-Panabières and K. Pantel. Clinical Applications of Circulating Tumor Cells and Circulating Tumor DNA as Liquid Biopsy. *Cancer discovery*, 6(5):479–491, 2016. ISSN 2159-8290. doi: 10.1158/2159-8290.CD-15-1483. URL http://cancerdiscovery.aacrjournals.org/content/6/5/479.full.

[102] M. Terai *et al.* Arterial Blood, Rather Than Venous Blood, is a Better Source for Circulating Melanoma Cells. *Ebiom*, 2015. ISSN 2352-3964. doi:10.1016/j.ebiom.2015.09.019. URL http://dx.doi.org/10.1016/j.ebiom.2015.09.019.

[103] C. Raimondi, C. Nicolazzo, A. Gradilone, D. M. Molecolare, and S. Università. Circulating tumor cells isolation : the post-EpCAM era . 27(5):461–470, 2015. doi:10.3978/j.issn.1000-9604.2015.06.02.

[104] D. R. Parkinson *et al.* Considerations in the development of circulating tumor cell technology for clinical use. *Journal of Translational Medicine*, 10(1):138, 2012. ISSN 1479-5876. doi:10.1186/1479-5876-10-138. URL JournalofTranslationalMedicine.

[105] QIAGEN. AllPrep ® DNA/RNA/Protein Mini Handbook, 2014.

[106] ThermoScientific. NanoDrop 2000/2000c Spectrophotometer V1.0 User Manual. 2000.

[107] J. Vermeulen *et al.* RNA pre-amplification enables large-scale RT-qPCR gene-expression studies on limiting sample amounts. *BMC research notes*, 2:235, 2009. ISSN 1756-0500. doi:10.1186/1756-0500-2-235.

[108] ThermoFisher. Ion Proton Workflow. URL http://www.thermofisher.com/content/dam/LifeTech/migration/images/brands/ion.par.70403.image.740.360.1..gif?direct=1.

[109] Thermo Fischer Scientific. Ion AmpliSeq Cancer Hotspot Panel v2, 2015. URL https://www.thermofisher.com/order/catalog/product/4475346.

[110] K. Boon *et al.* An anatomy of normal and malignant gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 99(17):11287–92, 2002. ISSN 0027-8424. doi:10.1073/pnas.152324199. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=123249{&}tool=pmcentrez{&}rendertype=abstract.

[111] M. N. McCall, H. R. McMurray, H. Land, and A. Almudevar. On non-detects in qPCR data. *Bioinformatics*, 30(16):2310–2316, 2014. ISSN 14602059. doi:10.1093/bioinformatics/btu239.

[112] K. J. Livak and T. D. Schmittgen. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods (San Diego, Calif.)*, 25(4):402–408, 2001. ISSN 1046-2023. doi:10.1006/meth.2001.1262.

[113] B. Weigelt, P. H. Warne, and J. Downward. PIK3CA mutation, but not PTEN loss of function, determines the sensitivity of breast cancer cells to mTOR inhibitory drugs. *Oncogene*, 30(29):3222–3233, 2011. ISSN 0950-9232. doi:10.1038/onc.2011.42. URL http://dx.doi.org/10.1038/onc.2011.42.

[114] Q. B. She *et al.* Breast tumor cells with P13K mutation or HER2 amplification are selectively addicted to Akt signaling. *PLoS ONE*, 3(8), 2008. ISSN 19326203. doi:10.1371/journal.pone.0003065.

[115] Y. Lu *et al.* Isolation and characterization of living circulating tumor cells in patients by immunomagnetic negative enrichment coupled with flow cytometry. *Cancer*, 2015. ISSN 0008543X. doi:10.1002/cncr.29444.

[116] S. Riethdorf *et al.* Detection of circulating tumor cells in peripheral blood of patients with metastatic breast cancer: a validation study of the CellSearch system. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 13(3):920–928, 2007. ISSN 1078-0432. doi:10.1158/1078-0432.CCR-06-1695.

[117] D. L. Holliday and V. Speirs. Choosing the right cell line for breast cancer research. *Breast cancer research : BCR*, 13(4):215, 2011. ISSN 1465-542X. doi:10.1186/bcr2889. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3236329{&}tool=pmcentrez{&}rendertype=abstract.

[118] K. D. Pruitt *et al.* RefSeq: An update on mammalian reference sequences. *Nucleic Acids Research*, 42(D1):756–763, 2014. ISSN 03051048. doi:10.1093/nar/gkt1114.

[119] M.-M. Shao *et al.* Keratin expression in breast cancers. *Virchows Archiv : an international journal of pathology*, 461(3):313–22, 2012. ISSN 1432-2307. doi:10.1007/s00428-012-1289-9. URL http://www.ncbi.nlm.nih.gov/pubmed/22851038.

[120] M. a. Watson, C. Darrow, D. B. Zimonjic, N. C. Popescu, and T. P. Fleming. Structure and transcriptional regulation of the human mammaglobin gene, a breast cancer associated member of the uteroglobin gene family localized to chromosome 11q13. *Oncogene*, 16(6):817–824, 1998. ISSN 0950-9232. doi:10.1038/sj.onc.1201597.

[121] M. Lacroix. Significance, detection and markers of disseminated breast cancer cells. *Endocrine-Related Cancer*, 13(4):1033–1067, 2006. ISSN 13510088. doi:10.1677/ERC-06-0001.

[122] D. Hanahan and R. A. Weinberg. Hallmarks of cancer: The next generation. *Cell*, 144(5):646–674, 2011. ISSN 00928674. doi:10.1016/j.cell.2011.02.013. URL http://dx.doi.org/10.1016/j.cell.2011.02.013.

[123] S. Lamouille, J. Xu, and R. Derynck. Molecular mechanisms of epithelial-mesenchymal transition. *Nat Rev Mol Cell Biol*, 15(3):178–196, 2014. doi:10.1038/nrm3758.Molecular.

[124] K. Tjensvoll *et al.* Disseminated tumor cells in bone marrow assessed by TWIST1, cytokeratin 19, and mammaglobin A mRNA predict clinical outcome in operable breast cancer patients. *Clinical breast cancer*, 10(5):378–384, 2010. ISSN 1938-0666. doi:10.3816/CBC.2010.n.050. URL http://dx.doi.org/10.3816/CBC.2010.n.050.

[125] E. E. O'Leary *et al.* Identification of Steroid-Sensitive Gene-1/Ccdc80 as a JAK2-Binding Protein. *Molecular Endocrinology*, 27(4):619–634, 2013. ISSN 0888-8809. doi:10.1210/me.2011-1275. URL http://press.endocrine.org/doi/abs/10.1210/me.2011-1275.

[126] C. Brusegan *et al.* Ccdc80-l1 is involved in axon pathfinding of zebrafish motoneurons. *PLoS ONE*, 7(2), 2012. ISSN 19326203. doi:10.1371/journal.pone.0031851.

[127] E. M. Walczak *et al.* Wnt-Signaling Inhibits Adrenal Steroidogenesis by Cell-Autonomous and Non-Cell-Autonomous Mechanisms. *Molecular endocrinology (Baltimore, Md.)*, 28(August):me20141060, 2014. ISSN 1944-9917. doi:10.1210/me.2014-1060. URL http://www.ncbi.nlm.nih.gov/pubmed/25029241.

[128] F. Tremblay *et al.* Bidirectional modulation of adipogenesis by the secreted protein Ccdc80/DRO1/URB. *Journal of Biological Chemistry*, 284(12):8136–8147, 2009. ISSN 00219258. doi:10.1074/jbc.M809535200.

[129] A. Ferraro *et al.* Tumor suppressor role of the CL2/DRO1/CCDC80 gene in thyroid carcinogenesis. *Journal of Clinical Endocrinology and Metabolism*, 98(7):2834–2843, 2013. ISSN 0021972X. doi:10.1210/jc.2012-2926.

[130] D. Marcantonio, L. E. Chalifour, M. a. Alaoui-Jamali, L. Alpert, and H. T. Huynh. Cloning and characterization of a novel gene that is regulated by estrogen and is associated with mammary gland carcinogenesis. *Endocrinology*, 142(6):2409–18, 2001. ISSN 0013-7227. URL http://www.ncbi.nlm.nih.gov/pubmed/11356689.

[131] D. T. Ting *et al.* Single-cell RNA sequencing identifies extracellular matrix gene expression by pancreatic circulating tumor cells. *Cell Reports*, 8(6):1905–1918, 2014. ISSN 22111247. doi:10.1016/j.celrep.2014.08.029.

[132] P. J. Neame, C. J. Kay, D. J. McQuillan, M. P. Beales, and J. R. Hassell. Independent modulation of collagen fibrillogenesis by decorin and lumican. *Cellular and Molecular Life Sciences*, 57(5):859–863, 2000. ISSN 1420682X. doi:10.1007/s000180050048.

[133] S. Chen and D. E. Birk. The regulatory roles of small leucine-rich proteoglycans in extracellular assembly*. *FEBS Journal*, 280(10):2120–2137, 2010. doi:10.1111/febs.12136.

[134] R. V. Iozzo. The family of the small leucine-rich proteoglycans: key regulators of matrix assembly and cellular growth. *Critical reviews in biochemistry and molecular biology*, 32(2):141–74, 1997. ISSN 1040-9238. doi:10.3109/10409239709108551. URL http://www.ncbi.nlm.nih.gov/pubmed/9145286.

[135] E. Leygue, L. Snell, H. Dotzlaw, and E. Al. Expression of lumican in human breast carcinoma. *Cancer Research*, 58:1348–1352, 1998. ISSN 10441549. doi:10.1165/ajrcmb.19.4.2979.

[136] T. Ishiwata *et al.* Role of lumican in cancer cells and adjacent stromal tissues in human pancreatic cancer. *Oncology Reports*, 18(3):537–543, 2007. ISSN 1021335X.

[137] Z. Naito. Role of the small leucine-rich proteoglycan (SLRP) family in pathological lesions and cancer cell growth. *Journal of Nippon Medical School = Nippon Ika Daigaku zasshi*, 72(3):137–45, 2005. ISSN 1345-4676. doi:10.1272/jnms.72.137. URL http://www.ncbi.nlm.nih.gov/pubmed/16046829.

[138] C. Panis, L. Pizzatti, A. C. Herrera, R. Cecchini, and E. Abdelhay. Putative circulating markers of the early and advanced stages of breast cancer identified by high-resolution label-free proteomics. *Cancer Letters*, 330(1):57–66, 2013. ISSN 03043835. doi:10.1016/j.canlet.2012.11.020. URL http://dx.doi.org/10.1016/j.canlet.2012.11.020.

[139] A. Cho, V. M. Howell, and E. K. Colvin. The Extracellular Matrix in Epithelial Ovarian Cancer A Piece of a Puzzle. *Frontiers in Oncology*, 5(November):245, 2015. ISSN 2234-943X. doi:10.3389/fonc.2015.00245. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4629462{&}tool=pmcentrez{&}rendertype=abstract.

[140] X. Li *et al.* Extracellular lumican inhibits pancreatic cancer cell growth and is associated with prolonged survival after surgery. *Clinical Cancer Research*, 20(24):6529–6540, 2014. ISSN 1878-5832. doi::10.1158/1078-0432.CCR-14-0970.

[141] F. You *et al.* Low-level expression of HER2 and CK19 in normal peripheral blood mononuclear cells: relevance for detection of circulating tumor cells. *Journal of hematology & oncology*, 1:2, 2008. ISSN 1756-8722. doi:10.1186/1756-8722-1-2.

[142] J. Shendure and H. Ji. Next-generation DNA sequencing. *Nature Biotechnology*, 26(10):1135–1145, 2008. ISSN 1087-0156. doi:10.1038/nbt1486. URL http://www.nature.com/doifinder/10.1038/nbt1486.

[143] F. Kasai, N. Hirayama, M. Ozawa, M. Iemura, and A. Kohara. Changes of heterogeneous cell populations in the Ishikawa cell line during long-term culture: Proposal for an in vitro clonal evolution model of tumor cells. *Genomics*, 107(6):259–266, 2016. ISSN 08887543. doi:10.1016/j.ygeno.2016.04.003. URL http://linkinghub.elsevier.com/retrieve/pii/S0888754316300246.

[144] H. Holstege *et al.* Somatic mutations found in the healthy blood compartment of a 115-yr-old woman demonstrate oligoclonal hematopoiesis. *Genome Research*, 24(5):733–742, 2014. ISSN 15495469. doi:10.1101/gr.162131.113.

[145] M. Gajecka. Unrevealed mosaicism in the next-generation sequencing era. *Molecular Genetics and Genomics*, 291(2):513–530, 2015. ISSN 16174623. doi:10.1007/s00438-015-1130-7. URL "http://dx.doi.org/10.1007/s00438-015-1130-7.

[146] D. C. Koboldt *et al.* Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012. ISSN 0028-0836. doi:10.1038/nature11412.

[147] S. R. Kennedy *et al.* Detecting ultralow-frequency mutations by Duplex Sequencing. *Nat Protoc*, 4(164):2586–2606, 2011. doi:10.1038/nprot.2014.170.

[148] J. et al. Schmitt, M.W., Kennedy, S.R., Salk, M. W. Schmitt, J. et al. Schmitt, M.W., Kennedy, S.R., Salk, and M. W. Schmitt. Detection of ultra-rare mutations by next-generation sequencing Michael. *PNAS*, 109(36):14508–14513, 2007. ISSN 16136829. doi:10.1073/pnas.1208715109/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1208715109.

[149] A. M. Newman *et al.* An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nature medicine*, 20(5):548–54, 2014. ISSN 1546-170X. doi:10.1038/nm.3519. URL http://www.ncbi.nlm.nih.gov/pubmed/24705333.

[150] A. M. Newman *et al.* Integrated digital error suppression for improved detection of circulating tumor DNA. *Nat Biotechnol*, (October 2015), 2016. ISSN 1087-0156. doi:10.1038/nbt.3520. URL http://www.ncbi.nlm.nih.gov/pubmed/27018799$\delimiter"026E30F$nhttp://www.nature.com/nbt/journal/vaop/ncurrent/pdf/nbt.3520.pdf.

[151] Y. Li *et al.* Antibody-Modified Reduced Graphene Oxide Films with Extreme Sensitivity to Circulating Tumor Cells. *Advanced Materials*, pages n/a–n/a, 2015. ISSN 09359648. doi:10.1002/adma.201502615. URL http://doi.wiley.com/10.1002/adma.201502615.

[152] M. Z. Mousavi *et al.* Label-free detection of rare cell in human blood using gold nano slit surface plasmon resonance. *Biosensors*, 5(1):98–117, 2015. ISSN 20796374. doi:10.3390/bios5010098.

[153] P. Chen, Y.-Y. Huang, K. Hoshino, and J. X. J. Zhang. Microscale magnetic field modulation for enhanced capture and distribution of rare circulating tumor cells. *Scientific reports*, 5:8745, 2015. ISSN 2045-2322. doi:10.1038/srep08745. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4348664{&}tool=pmcentrez{&}rendertype=abstract.

[154] L. Xu *et al.* Comparison of the prognostic utility of the diverse molecular data among lncRNA, DNA methylation, microRNA, and mRNA across five human cancers. *PLoS ONE*, 10(11):1–17, 2015. ISSN 19326203. doi:10.1371/journal.pone.0142433.

[155] S.-J. Dawson *et al.* Analysis of Circulating Tumor DNA to Monitor Metastatic Breast Cancer. *New England Journal of Medicine*, 368(13):1199–1209, 2013. ISSN 0028-4793. doi:10.1056/NEJMoa1213261. URL http://www.nejm.org/doi/abs/10.1056/NEJMoa1213261.

[156] I. Garcia-Murillas. Mutation tracking in circulating tumor DNA predicts relapse in early breast cancer. *Science Translational Medicine*, 7(302):1–12, 2015.

[157] K. Tjensvoll *et al.* Persistent tumor cells in bone marrow of non-metastatic breast cancer patients after primary surgery are associated with inferior outcome, 2012. doi:10.1186/1471-2407-12-190.

[158] R. K. Farmen *et al.* Bone marrow cytokeratin 19 mRNA level is an independent predictor of relapse-free survival in operable breast cancer patients. *Breast Cancer Research and Treatment*, 108(2):251–258, 2008. ISSN 01676806. doi:10.1007/s10549-007-9592-x.

[159] B. Gilje *et al.* Comparison of molecular and immunocytochemical methods for detection of disseminated tumor cells in bone marrow from early breast cancer patients. *BMC cancer*, 14(1):514, 2014. ISSN 1471-2407. doi:10.1186/1471-2407-14-514. URL http://www.ncbi.nlm.nih.gov/pubmed/25023626.

[160] K. Tjensvoll *et al.* A small subgroup of operable breast cancer patients with poor prognosis identified by quantitative real-time RT-PCR detection of mammaglobin A and trefoil factor 1 mRNA expression in bone marrow. *Breast Cancer Research and Treatment*, 116(2):329–338, 2009. ISSN 01676806. doi:10.1007/s10549-008-0204-1.

[161] M. Yu *et al.* Circulating Breast Tumor Cells Exhibit Dynamic Changes in Epithelial and Mesenchymal Composition. *Science*, 339(6119):580–584, 2013. doi:10.1126/science.1228522.Circulating.

[162] C. L. Chen *et al.* Single-cell analysis of circulating tumor cells identifies cumulative expression patterns of EMT-related genes in metastatic prostate cancer. *Prostate*, 73(8):813–826, 2013. ISSN 02704137. doi:10.1002/pros.22625.

Appendix A

TABLE 1: Markers used for tumor cell detection in literature.

Paper	**Cancer Type**	**Markers**
Iinuma 2011[47]	colorectal cancer	*CEA, CK19, CK20, CD133*
Ozkumur 2013[56]	breast, prostate, melanoma	Broad Panel (n = 34)
Tjensvoll 2012[157]	metastatic breast cancer	MM panel: *CK19, hMAM, TWIST1*
Giordano 2012[78]	metastatic breast cancer	*TWIST1, SNAIL1, ZEB1, TG2, ERBB2* (CD24, CD44, CD133 measured by flow cyt)
Farmen 2008[158]	breast cancer	*CK19*
Gilje 2014[159]	early breast cancer	*KR19, TWIST, hMAM*
Liu 2011[77]	breast cancer (+ 7 others)	*EpCAM, CK7/8* *
Lu 2015[115]	colorectal cancer	*EpCAM, CK* panel *
Molloy 2011[67]	early breast cancer	*CK19, p1B, EGP, MmGl*
Nadal 2012[70]	breast cancer	ER, PR, EGFR (IF), *HER2, TOP2A*, CEP17 (FISH) *
Strati 2011[68]	breast cancer	*CK19, MAGE-A3, HER2, TWIST1, hTERT, MmGl*
Tjensvoll 2009[160]	breast cancer	*hMAM, TFF1, PDEF*
Tjensvoll 2010[124]	breast cancer	*TWIST1, CK19, MmGlA*
Yu 2013[161]	pancreatic cancer	Wnt2: *Etv4, Mycn, Fn1; CK, EpCAM*
Shen 2009[59]	breast cancer	survivin, *hTERT, hMAM*
Strati 2013[76]	early breast cancer	*HER2, MUC1, GA733-2, CK19, MAGE A3*
Chen 2013[162]	prostate cancer	broad panel (n=84)
Aktas 2009[48]	metastatic breast cancer	*Twist1, akt2, PI3K*α
Mikhitarian 2008[58]	breast cancer	*Mam, PIP, CEA, PSE, CK19, MUC1, EpCam*

Appendix B

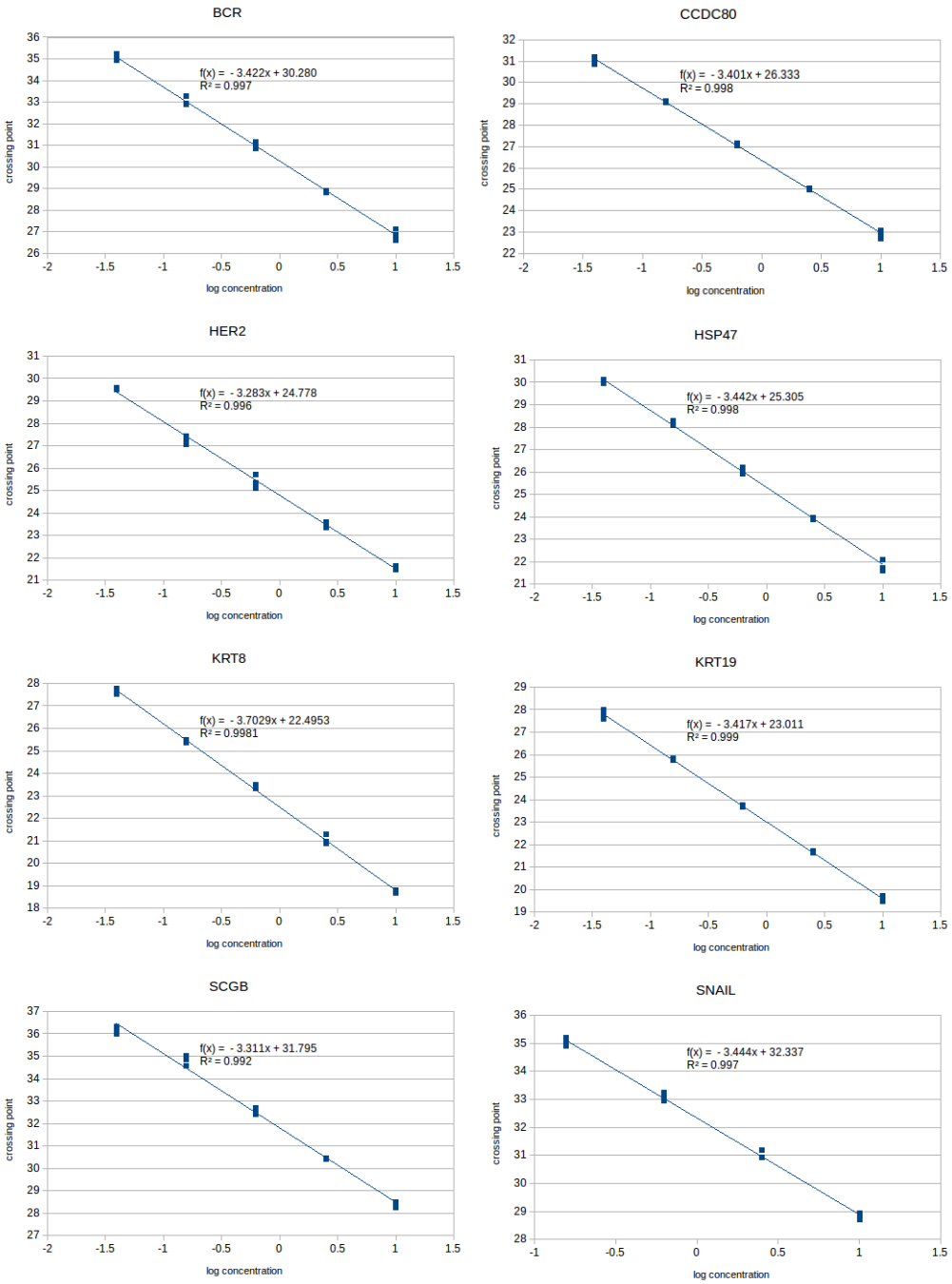

FIGURE 1: Standard curves used to calculate amplification efficiencies of assays (1)

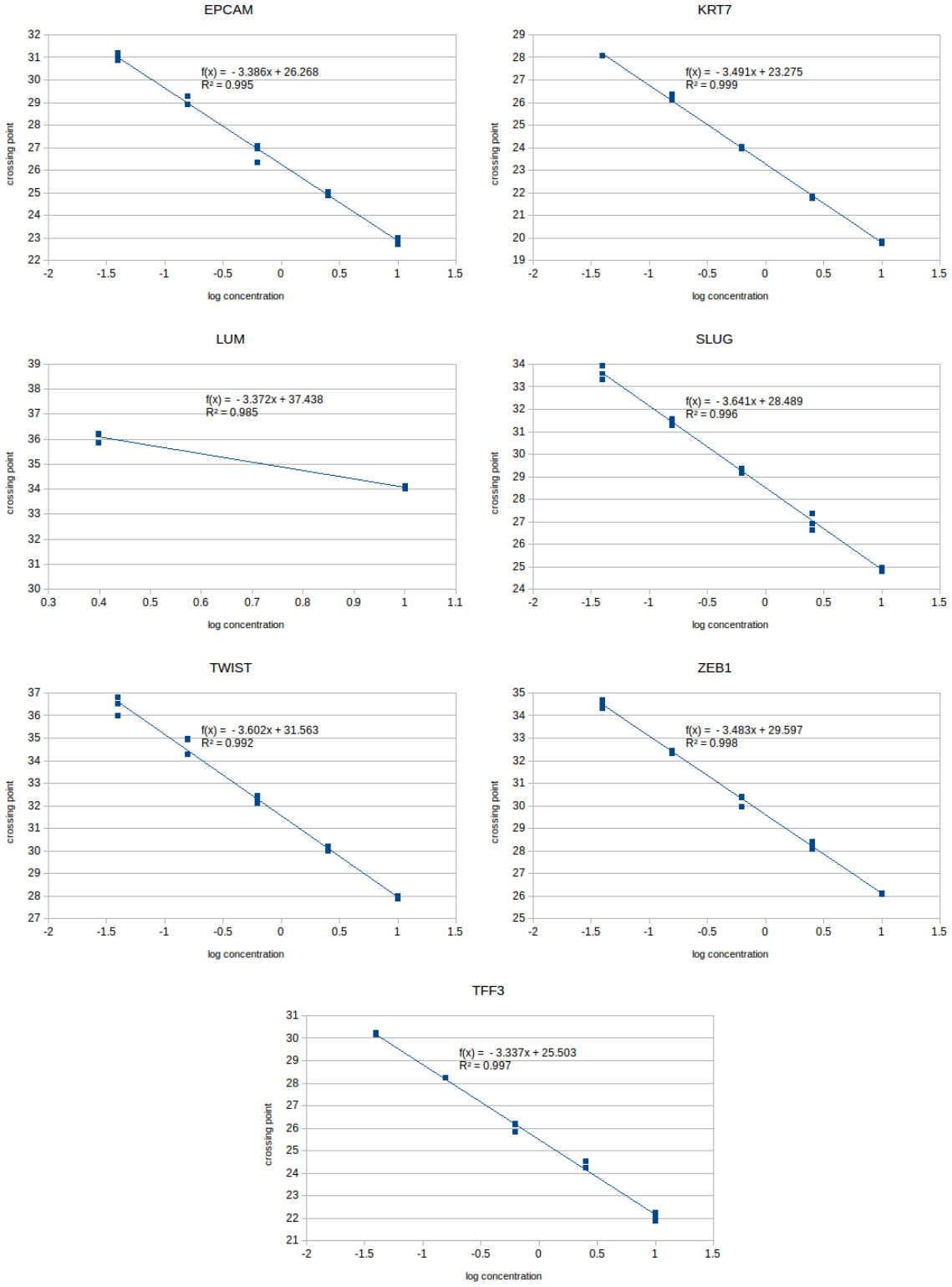

FIGURE 2: Standard curves used to calculate amplification efficiencies of assays (2)

Appendix C

FIGURE 4: Detailed gene expression data for each tumor sample and normal control.

Appendix D

.1 | PBCB data analysis

```
#packages needed

library("stringr")
library("reshape2")

#import data from results exported from LC480, clean up,
#and calculate relative expression

importLC480.pbcb <- function(x) {
  #import LC480-exported text file
  rawdata <- read.delim(x, header = FALSE)

  #keep only name and CP (col 4:5); remove header and column names (row 1:2)
  cp.data <- rawdata[-c(1,2), c(4,5)]

  #split name column into 3 to separate id, visit, and gene
  id.visit.gene <- str_split_fixed(cp.data$V4, " ", 3)

  #remove "'" from ids
  id <- gsub("'", "", paste(id.visit.gene[,1]))

  #ensure CP is numeric
  cp <- as.numeric(levels(cp.data$V5)[cp.data$V5])

  #combine into new data frame
  lcdata <- data.frame(ID = id, Visit = id.visit.gene[,2], CP = cp)

  return(lcdata)
}

#use function to import all plate data and then combine plates for each
#gene & order by ascending ID

bcr_1 <- importLC480.pbcb("bcr_1.txt")
bcr_2 <- importLC480.pbcb("bcr_2.txt")
bcr_bind <- rbind(bcr_1, bcr_2)
bcr <- bcr_bind[order(bcr_bind$ID),]

ccdc80_1 <- importLC480.pbcb("ccdc_1.txt")
ccdc80_2 <- importLC480.pbcb("ccdc_2.txt")
ccdc80_bind <- rbind(ccdc80_1, ccdc80_2)
ccdc80 <- ccdc80_bind[order(ccdc80_bind$ID),]
```

```
epcam_1 <- importLC480.pbcb("epcam_1.txt")
epcam_2 <- importLC480.pbcb("epcam_2.txt")
epcam_bind <- rbind(epcam_1, epcam_2)
epcam <- epcam_bind[order(epcam_bind$ID),]

erbb2_1 <- importLC480.pbcb("erbb2_1.txt")
erbb2_2 <- importLC480.pbcb("erbb2_2.txt")
erbb2_bind <- rbind(erbb2_1, erbb2_2)
erbb2 <- erbb2_bind[order(erbb2_bind$ID),]

scgb_1 <- importLC480.pbcb("scgb_1.txt")
scgb_2 <- importLC480.pbcb("scgb_2.txt")
scgb_bind <- rbind(scgb_1, scgb_2)
scgb <- scgb_bind[order(scgb_bind$ID),]

snail_1 <- importLC480.pbcb("snail_1.txt")
snail_2 <- importLC480.pbcb("snail_2.txt")
snail_bind <- rbind(snail_1, snail_2)
snail <- snail_bind[order(snail_bind$ID),]

twist_1 <- importLC480.pbcb("twist_1.txt")
twist_2 <- importLC480.pbcb("twist_2.txt")
twist_bind <- rbind(twist_1, twist_2)
twist <- twist_bind[order(twist_bind$ID),]

krt8_1 <- importLC480.pbcb("krt8_1.txt")
krt8_2 <- importLC480.pbcb("krt8_2.txt")
krt8_bind <- rbind(krt8_1, krt8_2)
krt8 <- krt8_bind[order(krt8_bind$ID),]

lum_1 <- importLC480.pbcb("lum_1.txt")
lum_2 <- importLC480.pbcb("lum_2.txt")
lum_bind <- rbind(lum_1, lum_2)
lum <- lum_bind[order(lum_bind$ID),]

slug_1 <- importLC480.pbcb("slug_1.txt")
slug_2 <- importLC480.pbcb("slug_2.txt")
slug_bind <- rbind(slug_1, slug_2)
slug <- slug_bind[order(slug_bind$ID),]

krt19_1 <- importLC480.pbcb("krt19_1.txt")
krt19_2 <- importLC480.pbcb("krt19_2.txt")
krt19_bind <- rbind(krt19_1, krt19_2)
krt19 <- krt19_bind[order(krt19_bind$ID),]

#combine all CPs into one data frame with gene names as colnames

all.data <- data.frame(ID = ccdc80$ID, Visit = ccdc80$Visit, BCR = bcr$CP, CCDC80
 = ccdc80$CP, EPCAM = epcam$CP, ERBB2 = erbb2$CP, SCGB = scgb$CP, SNAIL =
   snail$CP, TWIST = twist$CP, KRT8 = krt8$CP, LUM = lum$CP, SLUG = slug$CP,
 KRT19 = krt19$CP)

#store calibrator values in separate objects

bcr.cal <- c(all.data[c(430, 431, 432, 433), c(3)], all.data[c(432, 433), c(4)],
             all.data[c(432, 433), c(5)], all.data[c(432, 433), c(6)],
             all.data[c(432, 433), c(7)], all.data[c(432, 433), c(8)],
             all.data[c(432, 433), c(9)], all.data[c(432, 433), c(10)],
             all.data[c(432, 433), c(11)], all.data[c(432, 433), c(12)])
```

```
ccdc80.cal <- c(all.data[c(430, 431, 434, 435), c(4)])
epcam.cal <- c(all.data[c(430, 431, 434, 435), c(5)])
erbb2.cal <- c(all.data[c(430, 431, 434, 435), c(6)])
scgb.cal <- c(all.data[c(430, 431, 434, 435), c(7)])
snail.cal <- c(all.data[c(430, 431, 434, 435), c(8)])
twist.cal <- c(all.data[c(430, 431, 434, 435), c(9)])
krt8.cal <- c(all.data[c(430, 431, 434, 435), c(10)])
lum.cal <- c(all.data[c(430, 431, 434, 435), c(11)])
slug.cal <- c(all.data[c(430, 431, 434, 435), c(12)])
krt19.cal <- c(all.data[c(430, 431, 434, 435), c(13)])

#calculate calibrator means and sd and create data frame with values
#mean
mean.bcr.cal <- mean(bcr.cal)
mean.ccdc80.cal <- mean(ccdc80.cal)
mean.epcam.cal <- mean(epcam.cal)
mean.erbb2.cal <- mean(erbb2.cal)
mean.scgb.cal <- mean(scgb.cal)
mean.snail.cal <- mean(snail.cal)
mean.twist.cal <- mean(twist.cal)
mean.krt8.cal <- mean(krt8.cal)
mean.lum.cal <- mean(lum.cal)
mean.slug.cal <- mean(slug.cal)
mean.krt19.cal <- mean(krt19.cal)
#sd
sd.bcr.cal <- sd(bcr.cal)
sd.ccdc80.cal <- sd(ccdc80.cal)
sd.epcam.cal <- sd(epcam.cal)
sd.erbb2.cal <- sd(erbb2.cal)
sd.scgb.cal <- sd(scgb.cal)
sd.snail.cal <- sd(snail.cal)
sd.twist.cal <- sd(twist.cal)
sd.krt8.cal <- sd(krt8.cal)
sd.lum.cal <- sd(lum.cal)
sd.slug.cal <- sd(slug.cal)
sd.krt19.cal <- sd(krt19.cal)

gene.list <- c("BCR", "CCDC80", "EPCAM", "ERBB2", "SCGB", "SNAIL", "TWIST",
               "KRT8", "LUM", "SLUG", "KRT19")
mean.cal <- c(mean.bcr.cal, mean.ccdc80.cal, mean.epcam.cal, mean.erbb2.cal,
              mean.scgb.cal, mean.snail.cal, mean.twist.cal, mean.krt8.cal,
              mean.lum.cal, mean.slug.cal, mean.krt19.cal)
sd.cal <- c(sd.bcr.cal, sd.ccdc80.cal, sd.epcam.cal, sd.erbb2.cal, sd.scgb.cal,
            sd.snail.cal, sd.twist.cal, sd.krt8.cal, sd.lum.cal, sd.slug.cal,
sd.krt19.cal)

cal.data <- data.frame(Calibrator = gene.list, Mean = mean.cal, SD =
                         sd.cal)
cal.data <- transform(cal.data, CV = round((SD/Mean)*100, 1))

#remove CC, NTC, and empty wells from all.data

all.data.clean <- all.data[!(all.data$ID %in% c("CC", "NTC", "Sample")), ]

#simplify ID/Visit to use in frame of combined replicates

c.data <- all.data.clean[seq(1, nrow(all.data.clean), 3), 1:2]

#transform each gene CP row into matrices for analysis
```

```
mat.bcr <- matrix(all.data.clean$BCR, nrow = 3)
mat.ccdc80 <- matrix(all.data.clean$CCDC80, nrow = 3)
mat.epcam <- matrix(all.data.clean$EPCAM, nrow = 3)
mat.erbb2 <- matrix(all.data.clean$ERBB2, nrow = 3)
mat.scgb <- matrix(all.data.clean$SCGB, nrow = 3)
mat.snail <- matrix(all.data.clean$SNAIL, nrow = 3)
mat.twist <- matrix(all.data.clean$TWIST, nrow = 3)
mat.krt8 <- matrix(all.data.clean$KRT8, nrow = 3)
mat.lum <- matrix(all.data.clean$LUM, nrow = 3)
mat.slug <- matrix(all.data.clean$SLUG, nrow = 3)
mat.krt19 <- matrix(all.data.clean$KRT19, nrow = 3)

#take mean of each CP matrix and create new column for each gene

c.data$BCR <- apply(mat.bcr, 2, FUN=mean, na.rm=TRUE)
c.data$CCDC80 <- apply(mat.ccdc80, 2, FUN=mean, na.rm=TRUE)
c.data$EPCAM <- apply(mat.epcam, 2, FUN=mean, na.rm=TRUE)
c.data$ERBB2 <- apply(mat.erbb2, 2, FUN=mean, na.rm=TRUE)
c.data$SCGB <- apply(mat.scgb, 2, FUN=mean, na.rm=TRUE)
c.data$SNAIL <- apply(mat.snail, 2, FUN=mean, na.rm=TRUE)
c.data$TWIST <- apply(mat.twist, 2, FUN=mean, na.rm=TRUE)
c.data$KRT8 <- apply(mat.krt8, 2, FUN=mean, na.rm=TRUE)
c.data$LUM <- apply(mat.lum, 2, FUN=mean, na.rm=TRUE)
c.data$SLUG <- apply(mat.slug, 2, FUN=mean, na.rm=TRUE)
c.data$SLUG <- apply(mat.slug, 2, FUN=mean, na.rm=TRUE)
c.data$KRT19 <- apply(mat.krt19, 2, FUN=mean, na.rm=TRUE)

#calculate sd for each matrix and save in new data frame

c.data.sd <- c.data[, c(1,2)]
c.data.sd$BCR <- apply(mat.bcr, 2, FUN=sd, na.rm=TRUE)
c.data.sd$CCDC80 <- apply(mat.ccdc80, 2, FUN=sd, na.rm=TRUE)
c.data.sd$EPCAM <- apply(mat.epcam, 2, FUN=sd, na.rm=TRUE)
c.data.sd$ERBB2 <- apply(mat.erbb2, 2, FUN=sd, na.rm=TRUE)
c.data.sd$SCGB <- apply(mat.scgb, 2, FUN=sd, na.rm=TRUE)
c.data.sd$SNAIL <- apply(mat.snail, 2, FUN=sd, na.rm=TRUE)
c.data.sd$TWIST <- apply(mat.twist, 2, FUN=sd, na.rm=TRUE)
c.data.sd$KRT8 <- apply(mat.krt8, 2, FUN=sd, na.rm=TRUE)
c.data.sd$LUM <- apply(mat.lum, 2, FUN=sd, na.rm=TRUE)
c.data.sd$SLUG <- apply(mat.slug, 2, FUN=sd, na.rm=TRUE)
c.data.sd$KRT19 <- apply(mat.krt19, 2, FUN=sd, na.rm=TRUE)

#calculate CV

c.data.cv <- c.data[, c(1,2)]
c.data.cv$BCR <- (c.data.sd$BCR / c.data$BCR) *100
c.data.cv$CCDC80 <- (c.data.sd$CCDC80 / c.data$CCDC80) *100
c.data.cv$EPCAM <- (c.data.sd$EPCAM / c.data$EPCAM) *100
c.data.cv$ERBB2 <- (c.data.sd$ERBB2 / c.data$ERBB2) *100
c.data.cv$SCGB <- (c.data.sd$SCGB / c.data$SCGB) *100
c.data.cv$SNAIL <- (c.data.sd$SNAIL / c.data$SNAIL) *100
c.data.cv$TWIST <- (c.data.sd$TWIST / c.data$TWIST) *100
c.data.cv$KRT8 <- (c.data.sd$KRT8 / c.data$KRT8) *100
c.data.cv$LUM <- (c.data.sd$LUM / c.data$LUM) *100
c.data.cv$SLUG <- (c.data.sd$SLUG / c.data$SLUG) *100
c.data.cv$KRT19 <- (c.data.sd$KRT19 / c.data$KRT19) *100


#calculate dCt for each gene
```

```
dct.ccdc80 <- c.data$CCDC80 - mean.ccdc80.cal
dct.epcam <- c.data$EPCAM - mean.epcam.cal
dct.erbb2 <- c.data$ERBB2 - mean.erbb2.cal
dct.scgb <- c.data$SCGB - mean.scgb.cal
dct.snail <- c.data$SNAIL - mean.snail.cal
dct.twist <- c.data$TWIST - mean.twist.cal
dct.krt8 <- c.data$KRT8 - mean.krt8.cal
dct.lum <- c.data$LUM - mean.lum.cal
dct.slug <- c.data$SLUG - mean.slug.cal
dct.krt19 <- c.data$KRT19 - mean.krt19.cal
dct.bcr <- c.data$BCR - mean.bcr.cal

#calculate 2ddct for each gene

ddct.bcr <- 2^-(dct.bcr - dct.bcr)
ddct.ccdc80 <- 2^-(dct.ccdc80 - dct.bcr)
ddct.epcam <- 2^-(dct.epcam - dct.bcr)
ddct.erbb2 <- 2^-(dct.erbb2 - dct.bcr)
ddct.scgb <- 2^-(dct.scgb - dct.bcr)
ddct.snail <- 2^-(dct.snail - dct.bcr)
ddct.twist <- 2^-(dct.twist - dct.bcr)
ddct.krt8 <- 2^-(dct.krt8 - dct.bcr)
ddct.lum <- 2^-(dct.lum - dct.bcr)
ddct.slug <- 2^-(dct.slug - dct.bcr)
ddct.krt19 <- 2^-(dct.krt19 - dct.bcr)

#combine data for each gene in separate data frame and add sample grouping

sample.group <- paste0(c(rep("PBCB", 170), rep("Control", 30)))

bcr.data <- data.frame(c.data[,1:2], Group = sample.group, CP = c.data$BCR,
                       SD = c.data.sd$BCR, dCt = dct.bcr, ddCt = ddct.bcr)
ccdc80.data <- data.frame(c.data[,1:2], Group = sample.group, CP = c.data$CCDC80,
                      SD = c.data.sd$CCDC80, dCt = dct.ccdc80, ddCt = ddct.ccdc80)
epcam.data <- data.frame(c.data[,1:2], Group = sample.group, CP = c.data$EPCAM,
                   SD = c.data.sd$EPCAM, dCt = dct.epcam, ddCt = ddct.epcam)
erbb2.data <- data.frame(c.data[,1:2], Group = sample.group, CP = c.data$ERBB2,
                  SD = c.data.sd$ERBB2, dCt = dct.erbb2, ddCt = ddct.erbb2)
scgb.data <- data.frame(c.data[,1:2], Group = sample.group, CP = c.data$SCGB,
                 SD = c.data.sd$SCGB, dCt = dct.scgb, ddCt = ddct.scgb)
snail.data <- data.frame(c.data[,1:2], Group = sample.group, CP = c.data$SNAIL,
               SD = c.data.sd$SNAIL, dCt = dct.snail, ddCt = ddct.snail)
twist.data <- data.frame(c.data[,1:2], Group = sample.group, CP = c.data$TWIST,
                 SD = c.data.sd$TWIST, dCt = dct.twist, ddCt = ddct.twist)
krt8.data <- data.frame(c.data[,1:2], Group = sample.group, CP = c.data$KRT8,
                SD = c.data.sd$KRT8,  dCt = dct.krt8, ddCt = ddct.krt8)
lum.data <- data.frame(c.data[,1:2], Group = sample.group, CP = c.data$LUM,
              SD = c.data.sd$LUM,  dCt = dct.lum, ddCt = ddct.lum)
slug.data <- data.frame(c.data[,1:2], Group = sample.group, CP = c.data$SLUG,
           SD = c.data.sd$SLUG, dCt = dct.slug, ddCt = ddct.slug)
krt19.data <- data.frame(c.data[,1:2], Group = sample.group, CP = c.data$KRT19,
                       SD = c.data.sd$KRT19, dCt = dct.krt19, ddCt = ddct.krt19)

#combine all 2^ddct data in one data frame
rel.data <- data.frame(c.data[,1:2], Group = sample.group, CCDC80 = ddct.ccdc80,
                       EPCAM = ddct.epcam, ERBB2 = ddct.erbb2, KRT8 = ddct.krt8,
                       KRT19 = ddct.krt19, LUM = ddct.lum, SCGB = ddct.scgb,
                       SLUG = ddct.slug, SNAIL = ddct.snail, TWIST = ddct.twist)
```

```
#add groupings to c.data
c.data$Group <- sample.group


#calculate mean, max, and sd of control ddct for each gene
#subset
controls.ccdc80 <- ccdc80.data[(ccdc80.data$Group == "Control"), ]
controls.epcam <- epcam.data[(epcam.data$Group == "Control"), ]
controls.erbb2 <- erbb2.data[(erbb2.data$Group == "Control"), ]
controls.scgb <- scgb.data[(scgb.data$Group == "Control"), ]
controls.snail <- snail.data[(snail.data$Group == "Control"), ]
controls.twist <- twist.data[(twist.data$Group == "Control"), ]
controls.krt8 <- krt8.data[(krt8.data$Group == "Control"), ]
controls.lum <- lum.data[(lum.data$Group == "Control"), ]
controls.slug <- slug.data[(slug.data$Group == "Control"), ]
controls.krt19 <- krt19.data[(krt19.data$Group == "Control"), ]
#mean
ddct.mean.controls.ccdc80 <- mean(controls.ccdc80$ddCt, na.rm = TRUE)
ddct.mean.controls.epcam <- mean(controls.epcam$ddCt, na.rm = TRUE)
ddct.mean.controls.erbb2 <- mean(controls.erbb2$ddCt, na.rm = TRUE)
ddct.mean.controls.scgb <- mean(controls.scgb$ddCt, na.rm = TRUE)
ddct.mean.controls.snail <- mean(controls.snail$ddCt, na.rm = TRUE)
ddct.mean.controls.twist <- mean(controls.twist$ddCt, na.rm = TRUE)
ddct.mean.controls.krt8 <- mean(controls.krt8$ddCt, na.rm = TRUE)
ddct.mean.controls.lum <- mean(controls.lum$ddCt, na.rm = TRUE)
ddct.mean.controls.slug <- mean(controls.slug$ddCt, na.rm = TRUE)
ddct.mean.controls.krt19 <- mean(controls.krt19$ddCt, na.rm = TRUE)
#sd
ddct.sd.controls.ccdc80 <- sd(controls.ccdc80$ddCt, na.rm = TRUE)
ddct.sd.controls.epcam <- sd(controls.epcam$ddCt, na.rm = TRUE)
ddct.sd.controls.erbb2 <- sd(controls.erbb2$ddCt, na.rm = TRUE)
ddct.sd.controls.scgb <- sd(controls.scgb$ddCt, na.rm = TRUE)
ddct.sd.controls.snail <- sd(controls.snail$ddCt, na.rm = TRUE)
ddct.sd.controls.twist <- sd(controls.twist$ddCt, na.rm = TRUE)
ddct.sd.controls.krt8 <- sd(controls.krt8$ddCt, na.rm = TRUE)
ddct.sd.controls.lum <- sd(controls.lum$ddCt, na.rm = TRUE)
ddct.sd.controls.slug <- sd(controls.slug$ddCt, na.rm = TRUE)
ddct.sd.controls.krt19 <- sd(controls.krt19$ddCt, na.rm = TRUE)
#maximum
ddct.max.controls.ccdc80 <- max(controls.ccdc80$ddCt, na.rm = TRUE)
ddct.max.controls.epcam <- max(controls.epcam$ddCt, na.rm = TRUE)
ddct.max.controls.erbb2 <- max(controls.erbb2$ddCt, na.rm = TRUE)
ddct.max.controls.scgb <- max(controls.scgb$ddCt, na.rm = TRUE)
ddct.max.controls.snail <- max(controls.snail$ddCt, na.rm = TRUE)
ddct.max.controls.twist <- max(controls.twist$ddCt, na.rm = TRUE)
ddct.max.controls.krt8 <- max(controls.krt8$ddCt, na.rm = TRUE)
ddct.max.controls.lum <- max(controls.lum$ddCt, na.rm = TRUE)
ddct.max.controls.slug <- max(controls.slug$ddCt, na.rm = TRUE)
ddct.max.controls.krt19 <- max(controls.krt19$ddCt, na.rm = TRUE)

#calculate mean + 3SD

ccdc80.threshold.3sd <- ddct.mean.controls.ccdc80 + 3*ddct.sd.controls.ccdc80
epcam.threshold.3sd <- ddct.mean.controls.epcam + 3*ddct.sd.controls.epcam
erbb2.threshold.3sd <- ddct.mean.controls.erbb2 + 3*ddct.sd.controls.erbb2
scgb.threshold.3sd <- ddct.mean.controls.scgb + 3*ddct.sd.controls.scgb
snail.threshold.3sd <- ddct.mean.controls.snail + 3*ddct.sd.controls.snail
twist.threshold.3sd <- ddct.mean.controls.twist + 3*ddct.sd.controls.twist
krt8.threshold.3sd <- ddct.mean.controls.krt8 + 3*ddct.sd.controls.krt8
lum.threshold.3sd <- ddct.mean.controls.lum + 3*ddct.sd.controls.lum
```

```
slug.threshold.3sd <- ddct.mean.controls.slug + 3*ddct.sd.controls.slug
krt19.threshold.3sd <- ddct.mean.controls.krt19 + 3*ddct.sd.controls.krt19

#summarize control data in table
mean.controls <- c(ddct.mean.controls.ccdc80, ddct.mean.controls.epcam,
                   ddct.mean.controls.erbb2, ddct.mean.controls.scgb,
                   ddct.mean.controls.snail, ddct.mean.controls.twist,
                   ddct.mean.controls.krt8, ddct.mean.controls.lum,
                   ddct.mean.controls.slug, ddct.mean.controls.krt19)

sd.controls <- c(ddct.sd.controls.ccdc80, ddct.sd.controls.epcam,
                 ddct.sd.controls.erbb2, ddct.sd.controls.scgb,
                 ddct.sd.controls.snail, ddct.sd.controls.twist,
                 ddct.sd.controls.krt8, ddct.sd.controls.lum,
                 ddct.sd.controls.slug, ddct.mean.controls.krt19)

max.controls <- c(ddct.max.controls.ccdc80, ddct.max.controls.epcam,
                  ddct.max.controls.erbb2, ddct.max.controls.scgb,
                  ddct.max.controls.snail, ddct.max.controls.twist,
                  ddct.max.controls.krt8, ddct.max.controls.lum,
                  ddct.max.controls.slug, ddct.max.controls.krt19)

sd_threshold <- c( ccdc80.threshold.3sd, epcam.threshold.3sd,
                    erbb2.threshold.3sd, scgb.threshold.3sd,
                    snail.threshold.3sd, twist.threshold.3sd,
                    krt8.threshold.3sd, lum.threshold.3sd,
                    slug.threshold.3sd, krt19.threshold.3sd)


gene.list.1 <- c("CCDC80", "EPCAM", "ERBB2", "SCGB", "SNAIL", "TWIST",
               "KRT8", "LUM", "SLUG", "KRT19")

control.data <- data.frame(Controls = gene.list.1, Mean = mean.controls,
                           Max = max.controls, SD = sd.controls,
                           Threshold = sd_threshold)

#remove outliers from control sets over 3SD threshold
sub.ccdc80.control <- subset(controls.ccdc80, ddCt <= ccdc80.threshold.3sd)
sub.epcam.control <- subset(controls.epcam, ddCt <= epcam.threshold.3sd)
sub.erbb2.control <- subset(controls.erbb2, ddCt <= erbb2.threshold.3sd)
sub.scgb.control <- subset(controls.scgb, ddCt <= scgb.threshold.3sd)
sub.snail.control <- subset(controls.snail, ddCt <= snail.threshold.3sd)
sub.twist.control <- subset(controls.twist, ddCt <= twist.threshold.3sd)
sub.krt8.control <- subset(controls.krt8, ddCt <= krt8.threshold.3sd)
sub.lum.control <- subset(controls.lum, ddCt <= lum.threshold.3sd)
sub.krt19.control <- subset(controls.krt19, ddCt <= krt19.threshold.3sd)
sub.slug.control <- subset(controls.slug, ddCt <= slug.threshold.3sd)



ddct.max.control2.ccdc80 <- max(sub.ccdc80.control$ddCt)
ddct.max.control2.epcam <- max(sub.epcam.control$ddCt)
ddct.max.control2.erbb2 <- max(sub.erbb2.control$ddCt)
ddct.max.control2.scgb <- max(sub.scgb.control$ddCt)
ddct.max.control2.snail <- max(sub.snail.control$ddCt)
ddct.max.control2.twist <- max(sub.twist.control$ddCt)
ddct.max.control2.krt8 <- max(sub.krt8.control$ddCt)
ddct.max.control2.lum <- max(sub.lum.control$ddCt)
ddct.max.control2.slug <- max(sub.slug.control$ddCt)
ddct.max.control2.krt19 <- max(sub.krt19.control$ddCt)
```

```
max2.controls <- c(ddct.max.control2.ccdc80, ddct.max.control2.epcam,
                   ddct.max.control2.erbb2, ddct.max.control2.scgb,
                   ddct.max.control2.snail, ddct.max.control2.twist,
                   ddct.max.control2.krt8, ddct.max.control2.lum,
                   ddct.max.control2.slug, ddct.max.controls.krt19)

control.data$New.Max <- max2.controls


#subset data based on thresholds

#max2 -> max without 3SD outliers
pos.ccdc80.max2 <- subset(ccdc80.data, ddCt > ddct.max.control2.ccdc80)
pos.epcam.max2 <- subset(epcam.data, ddCt > ddct.max.control2.epcam)
pos.erbb2.max2 <- subset(erbb2.data, ddCt > ddct.max.control2.erbb2)
pos.scgb.max2 <- subset(scgb.data, ddCt > ddct.max.control2.scgb)
pos.snail.max2 <- subset(snail.data, ddCt > ddct.max.control2.snail)
pos.twist.max2 <- subset(twist.data, ddCt > ddct.max.control2.twist)
pos.krt8.max2 <- subset(krt8.data, ddCt > ddct.max.control2.krt8)
pos.lum.max2 <- subset(lum.data, ddCt > ddct.max.control2.lum)
pos.slug.max2 <- subset(slug.data, ddCt > ddct.max.control2.slug)
pos.krt19.max2 <- subset(krt19.data, ddCt > ddct.max.control2.krt19)


max2.pos <- rbind(pos.ccdc80.max2, pos.epcam.max2, pos.erbb2.max2, pos.krt19.max2,
                  pos.lum.max2, pos.krt8.max2,
                 pos.scgb.max2, pos.slug.max2, pos.snail.max2, pos.twist.max2)
max2.pos$Gene <- paste0(c(rep("CCDC80", 10), rep("EPCAM", 7), "ERBB2",
                          rep("KRT19", 2), rep("LUM", 12), rep("KRT8", 8),
                          rep("SCGB", 2), "SLUG", "SNAIL", rep("TWIST", 5)))



#transform data to long form
long.cdata <- melt(c.data, id.vars = c("ID", "Visit", "Group"), measure.vars = c("BCR",
                   "CCDC80", "EPCAM", "ERBB2", "SCGB", "SNAIL", "TWIST",
                   "KRT8", "LUM", "SLUG", "KRT19"), variable.name = "GENE", value.name
                   = "CP")

long.reldata <- melt(rel.data, id.vars = c("ID", "Visit", "Group"), measure.vars = c(
              "CCDC80", "EPCAM", "ERBB2", "SCGB", "SNAIL", "TWIST",
              "KRT8", "LUM", "SLUG", "KRT19"), variable.name = "GENE", value.name
                   = "RelExp")

#plot CP counts
#count number at each CP
# cp.count <- table(long.cdata$CP)
#
# cp.hist <- hist(long.cdata$CP, breaks = c(seq(19,40, by=1)), main =
#                   "Histogram of CP Values", ylab = "Count", xlab = "CP")
#
# bcr.hist <- hist(c.data$BCR, main = "Distribution of BCR CP", ylab = "count",
#                   xlab = "CP")
# ccdc80.hist <- hist(c.data$CCDC80, main = "Distribution of CCDC80 CP", ylab = "count",
#                   xlab = "CP")
# epcam.hist <- hist(c.data$EPCAM, main = "Distribution of EPCAM CP", ylab = "count",
#                     xlab = "CP")
# erbb2.hist <- hist(c.data$ERBB2, main = "Distribution of ERBB2 CP", ylab = "count",
```

```
#                     xlab = "CP")
# scgb.hist <- hist(c.data$SCGB, main = "Distribution of SCGB CP", ylab = "count",
#                     xlab = "CP")
# snail.hist <- hist(c.data$SNAIL, main = "Distribution of SNAIL CP", ylab = "count",
#                     xlab = "CP")
# twist.hist <- hist(c.data$TWIST, main = "Distribution of TWIST CP", ylab = "count",
#                     xlab = "CP")
# krt8.hist <- hist(c.data$KRT8, main = "Distribution of KRT8 CP", ylab = "count",
#                     xlab = "CP")
# lum.hist <- hist(c.data$LUM, main = "Distribution of LUM CP", ylab = "count",
#                     xlab = "CP")
# slug.hist <- hist(c.data$SLUG, main = "Distribution of SLUG CP", ylab = "count",
#                     xlab = "CP")
# krt19.hist <- hist(c.data$KRT19, main = "Distribution of KRT19 CP", ylab = "count",
#                     xlab = "CP")

#normality test
# shapiro.test(c.data$BCR)
# shapiro.test(c.data$CCDC80)
# shapiro.test(c.data$EPCAM)
# shapiro.test(c.data$ERBB2)
# shapiro.test(c.data$SCGB)
# shapiro.test(c.data$SNAIL)
# shapiro.test(c.data$TWIST)
# shapiro.test(c.data$KRT8)
# shapiro.test(c.data$LUM)
# shapiro.test(c.data$SLUG)
# shapiro.test(c.data$KRT19)


#export data tables
write.table(all.data, file = "all.data.pbcb.txt", row.names = FALSE, col.names
            = TRUE, sep = "\t" )

write.table(all.data.clean, file = "all.data.pbcb-lessextra.txt", row.names = FALSE,
col.names = TRUE, sep = "\t" )

write.table(c.data, file = "ave.data.pbcb.txt", row.names = FALSE, col.names
            = TRUE, sep = "\t" )

write.table(c.data.sd, file = "sd.data.pbcb.txt", row.names = FALSE, col.names
            = TRUE, sep = "\t" )

write.table(c.data.cv, file = "cv.data.pbcb.txt", row.names = FALSE, col.names
            = TRUE, sep = "\t" )

write.table(cal.data, file = "caldata.pbcb.txt", row.names = FALSE, col.names
            = TRUE, sep = "\t" )

write.table(long.cdata, file = "long.data.pbcb.txt", row.names = FALSE, col.names
            = TRUE, sep = "\t" )

write.table(bcr.data, file = "bcr.data.pbcb.txt", row.names = FALSE, col.names
            = TRUE, sep = "\t" )

write.table(epcam.data, file = "epcam.data.pbcb.txt", row.names = FALSE, col.names
            = TRUE, sep = "\t" )

write.table(erbb2.data, file = "erbb2.data.pbcb.txt", row.names = FALSE, col.names
            = TRUE, sep = "\t" )
```

```
write.table(scgb.data, file = "scgb.data.pbcb.txt", row.names = FALSE, col.names
            = TRUE, sep = "\t" )

write.table(snail.data, file = "snail.data.pbcb.txt", row.names = FALSE, col.names
            = TRUE, sep = "\t" )

write.table(ccdc80.data, file = "ccdc80.data.pbcb.txt", row.names = FALSE, col.names
            = TRUE, sep = "\t" )

write.table(twist.data, file = "twist.data.pbcb.txt", row.names = FALSE, col.names
            = TRUE, sep = "\t" )

write.table(krt8.data, file = "krt8.data.pbcb.txt", row.names = FALSE, col.names
            = TRUE, sep = "\t" )

write.table(lum.data, file = "lum.data.pbcb.txt", row.names = FALSE, col.names
            = TRUE, sep = "\t" )

write.table(slug.data, file = "slug.data.pbcb.txt", row.names = FALSE, col.names
            = TRUE, sep = "\t" )

write.table(slug.data, file = "krt19.data.pbcb.txt", row.names = FALSE, col.names
            = TRUE, sep = "\t" )

write.table(control.data, file = "control.data.pbcb.txt", row.names = FALSE, col.names
            = TRUE, sep = "\t" )

write.table(max.pos, file = "positive.samples.txt", row.names = FALSE, col.names
            = TRUE, sep = "\t" )

write.table(max2.pos, file = "positive2.samples.txt", row.names = FALSE, col.names
            = TRUE, sep = "\t" )

write.table(max3.pos, file = "positive3.samples.txt", row.names = FALSE, col.names
            = TRUE, sep = "\t" )


#plotting
#see limitplot3.R for jitter plots
```

.2 | Plotting the data: jitter plots

```
#limitplot + jitter function
#
limitplot3 <-function (..., lod, CI = 95, ratio = 1/25, shape = 1, size = 1,
    col = "black", main = "", xlab = "", ylab = "", names = "",
    axis = 5, logaxis = 1, stack = 5, jitterwidth = 0.2, jittershape = 1,
    jittersize = 1, jittercol = "black", log = "", blod = 1/2,plotm=0.25)
{
    if (log == "y") {
        CI_lod <- log(lod * blod)
        lod <- log(lod)
        ya <- log(c(...))
        xa <- rep(seq(1:length(list(...))), times = as.numeric(summary(list(...))
[1:length(list(...))]))
```

```
        pl <- data.frame(xi = xa + runif(length(c(...)), -jitterwidth,
            jitterwidth), yi = ya)
        plot(pl$xi[pl$yi >= lod], exp(pl$yi[pl$yi >= lod]), xlim = c(plotm,
            length(list(...)) + (1-plotm)), ylim = c(if (lod <= min(ya)) {
            exp(lod)
        } else {
            exp(lod - ((1 - max(summary(factor(xa[ya < lod])))%%stack/stack) +
                        max(summary(factor(xa[ya < lod])))/stack) * (max(ya) -
                lod)/(1/ratio))
        }, exp(max(ya))), yaxp = c(10*exp(lod), exp(max(ya)), n = logaxis),
            xaxt = "n", ylab = ylab, xlab = xlab, log = "y",
            pch = jittershape, cex = jittersize, col = jittercol,
            main = main)
        mtext(names, side = 1, line=1, at = seq(1:length(list(...))))
        segments(0, exp(lod), length(list(...)) + 1, exp(lod),
            lty = "dashed")
    }

  }
     for (i in 1:length(list(...))) {
            if (exp(lod) <= exp(median(c(ya[ya >= lod & xa == i],
                 seq(from = CI_lod, to = CI_lod, length.out = length(ya[ya <
                   lod & xa == i])))))) {
  segments(i - 0.25, exp(median(c(ya[ya >= lod &
  xa == i], seq(from = CI_lod, to = CI_lod, length.out = length(ya[ya <
                 lod & xa == i])))))), i + 0.25, exp(median(c(ya[ya >=
                lod & xa == i], seq(from = CI_lod, to = CI_lod,
               length.out = length(ya[ya < lod & xa == i])))))),lwd=2)


  }
  }

        if (lod > min(ya)) {
            for (i in 1:length(list(...))) {
                xp <- rep(seq(-0.2, 0.2, length.out = stack),
                  len = length(xa[xa == i & ya < lod])) + i
                yp <- rep(seq(1:(1 - max(summary(factor(xa[ya <
                  lod])))%%stack/stack) + max(summary(factor(xa[ya <
                  lod])))/stack), each = stack, len = length(xa[ya <
                  lod & xa == i]))
                points(xp, exp(lod - yp * (max(ya) - lod)/(1/ratio)),
                  pch = shape, cex = size, col = col)
            }
        }

    else {
        ya <- c(...)
        xa <- rep(seq(1:length(list(...))), times = as.numeric(summary(list(...))
[1:length(list(...))]))
        pl <- data.frame(xi = xa + runif(length(c(...)), -jitterwidth,
            jitterwidth), yi = ya)
        plot(pl$xi[pl$yi >= lod], pl$yi[pl$yi >= lod], xlim = c(plotm,
            length(list(...)) + (1-plotm)), ylim = c(if (lod <= min(ya)) {
            lod
        } else {
            lod - ((1 - max(summary(factor(xa[ya < lod])))%%stack/stack) +
                max(summary(factor(xa[ya < lod])))/stack) * (max(ya) -
                lod)/(1/ratio)
        }, max(ya)), yaxp = c(lod, max(ya), n = (axis - 1)),
```

```
            xaxt = "n", ylab = ylab, xlab = xlab, pch = jittershape,
            cex = jittersize, col = jittercol, main = main)
        mtext(names, side = 1, line=1, at = seq(1:length(list(...))))
        segments(0, lod, length(list(...)) + 1, lod, lty = "dashed")

      }
        for (i in 1:length(list(...))) {
            if (lod <= mean(c(ya[ya >= lod & xa == i], seq(from = lod *
                blod, to = lod * blod, length.out = length(ya[ya <
                lod & xa == i])))) {

            }
        }
        if (lod > min(ya)) {
            for (i in 1:length(list(...))) {
                xp <- rep(seq(-0.2, 0.2, length.out = stack),
                  len = length(xa[xa == i & ya < lod])) + i
                yp <- rep(seq(1:(1 - max(summary(factor(xa[ya <
                  lod])))%%stack/stack) + max(summary(factor(xa[ya <
                  lod]))))/stack), each = stack, len = length(xa[ya <
                  lod & xa == i]))
                points(xp, lod - yp * (max(ya) - lod)/(1/ratio),
                  pch = shape, cex = size, col = col)
            }
        }
    }
#end function
#

#find minimum value to set LOD

min.ccdc80.rel <- min(rel.data$CCDC80, na.rm=TRUE)
#3.27e-05
min.epcam.rel <- min(rel.data$EPCAM, na.rm=TRUE)
#5.9e-05
min.erbb2.rel <- min(rel.data$ERBB2, na.rm=TRUE)
#3.8e-04
min.krt8.rel <- min(rel.data$KRT8, na.rm=TRUE)
#2.03e-06
min.krt19.rel <- min(rel.data$KRT19, na.rm=TRUE)
#4.01e-06
min.lum.rel <- min(rel.data$LUM, na.rm=TRUE)
#2.4e-02
min.scgb.rel <- min(rel.data$SCGB, na.rm=TRUE)
#1.7e-03
min.slug.rel <- min(rel.data$SLUG, na.rm=TRUE)
#5.11e-04
min.snail.rel <- min(rel.data$SNAIL, na.rm=TRUE)
#2.2e-03
min.twist.rel <- min(rel.data$TWIST, na.rm=TRUE)
#2.8e-04

# group assay data for each individual plot and
# replace NA with value lower than minimum from above
#
ccdc80.pbcb <- rel.data[rel.data$Group == "PBCB", 4]
ccdc80.pbcb[is.na(ccdc80.pbcb)] <- 1e-05
ccdc80.control <- rel.data[rel.data$Group == "Control", 4]
ccdc80.control[is.na(ccdc80.control)] <- 1e-05
```

```
epcam.pbcb <- rel.data[rel.data$Group == "PBCB", 5]
epcam.pbcb[is.na(epcam.pbcb)] <- 1e-05
epcam.control <- rel.data[rel.data$Group == "Control", 5]
epcam.control[is.na(epcam.control)] <- 1e-05
#remove outlier from data calculated in pbcb.import.analysis.R
s.epcam.control <- epcam.control[(epcam.control <= epcam.threshold.3sd)]

erbb2.pbcb <- rel.data[rel.data$Group == "PBCB", 6]
erbb2.pbcb[is.na(erbb2.pbcb)] <- 1e-04
erbb2.control <- rel.data[rel.data$Group == "Control", 6]
erbb2.control[is.na(erbb2.control)] <- 1e-04

krt8.pbcb <- rel.data[rel.data$Group == "PBCB", 7]
krt8.pbcb[is.na(krt8.pbcb)] <- 1e-06
krt8.control <- rel.data[rel.data$Group == "Control", 7]
krt8.control[is.na(krt8.control)] <- 1e-06
#remove outlier from data calculated in pbcb.import.analysis.R
s.krt8.control <- krt8.control[(krt8.control <= krt8.threshold.3sd)]

krt19.pbcb <- rel.data[rel.data$Group == "PBCB", 8]
krt19.pbcb[is.na(krt19.pbcb)] <- 2e-06
krt19.control <- rel.data[rel.data$Group == "Control", 8]
krt19.control[is.na(krt19.control)] <- 2e-06

lum.pbcb <- rel.data[rel.data$Group == "PBCB", 9]
lum.pbcb[is.na(lum.pbcb)] <- 1e-02
lum.control <- rel.data[rel.data$Group == "Control", 9]
lum.control[is.na(lum.control)] <- 1e-02

scgb.pbcb <- rel.data[rel.data$Group == "PBCB", 10]
scgb.pbcb[is.na(scgb.pbcb)] <- 1e-05
scgb.control <- rel.data[rel.data$Group == "Control", 10]
scgb.control[is.na(scgb.control)] <- 1e-05

slug.pbcb <- rel.data[rel.data$Group == "PBCB", 11]
slug.pbcb[is.na(slug.pbcb)] <- 1e-04
slug.control <- rel.data[rel.data$Group == "Control", 11]
slug.control[is.na(slug.control)] <- 1e-04

snail.pbcb <- rel.data[rel.data$Group == "PBCB", 12]
snail.pbcb[is.na(snail.pbcb)] <- 1e-03
snail.control <- rel.data[rel.data$Group == "Control", 12]
snail.control[is.na(snail.control)] <- 1e-03

twist.pbcb <- rel.data[rel.data$Group == "PBCB", 13]
twist.pbcb[is.na(twist.pbcb)] <- 1e-04
twist.control <- rel.data[rel.data$Group == "Control", 13]
twist.control[is.na(twist.control)] <- 1e-04
#remove outlier from data calculated in pbcb.import.analysis.R
s.twist.control <- twist.control[(twist.control <= twist.threshold.3sd)]


#set LOD slightly higher than replaced NA values and plot

limitplot3(lum.pbcb,lum.control,
           lod = 1.5e-02, CI = 95, ratio = 1/25, shape = 1, size = 1,
           col = "black", main = "Relative Expression of LUM", xlab = "Lumican",
           ylab = "Relative Expression",
           names = c("PBCB", "Control"),
           axis = 5, logaxis = 1, stack = 5, jitterwidth = 0.2, jittershape = 1,
```

```
          jittersize = 0.5, jittercol = "black", log = "y")

limitplot3(ccdc80.pbcb, ccdc80.control,
          lod = 2e-05, CI = 95, ratio = 1/25, shape = 1, size = 1,
          col = "black", main = "Relative Expression of CCDC80", xlab = "CCDC80",
          ylab = "Relative Expression",
          names = c("PBCB", "Control"),
          axis = 5, logaxis = 1, stack = 5, jitterwidth = 0.2, jittershape = 1,
          jittersize = 0.5, jittercol = "black", log = "y")

limitplot3(epcam.pbcb, s.epcam.control,
          lod = 3e-05, CI = 95, ratio = 1/25, shape = 1, size = 1,
          col = "black", main = "Relative Expression of EPCAM", xlab = "EPCAM",
          ylab = "Relative Expression",
          names = c("PBCB", "Control"),
          axis = 5, logaxis = 1, stack = 5, jitterwidth = 0.2, jittershape = 1,
          jittersize = 0.5, jittercol = "black", log = "y")

limitplot3(erbb2.pbcb, erbb2.control,
          lod = 2e-04, CI = 95, ratio = 1/25, shape = 1, size = 1,
          col = "black", main = "Relative Expression of ERBB2", xlab = "ERBB2",
          ylab = "Relative Expression",
          names = c("PBCB", "Control"),
          axis = 5, logaxis = 1, stack = 5, jitterwidth = 0.2, jittershape = 1,
          jittersize = 0.5, jittercol = "black", log = "y")

limitplot3(krt8.pbcb, s.krt8.control,
          lod = 1.5e-06, CI = 95, ratio = 1/25, shape = 1, size = 1,
          col = "black", main = "Relative Expression of KRT8", xlab = "KRT8",
          ylab = "Relative Expression",
          names = c("PBCB", "Control"),
          axis = 5, logaxis = 1, stack = 5, jitterwidth = 0.2, jittershape = 1,
          jittersize = 0.5, jittercol = "black", log = "y")

limitplot3(krt19.pbcb, krt19.control,
          lod = 3e-06, CI = 95, ratio = 1/25, shape = 1, size = 1,
          col = "black", main = "Relative Expression of KRT19", xlab = "KRT19",
          ylab = "Relative Expression",
          names = c("PBCB", "Control"),
          axis = 5, logaxis = 1, stack = 5, jitterwidth = 0.2, jittershape = 1,
          jittersize = 0.5, jittercol = "black", log = "y")

limitplot3(scgb.pbcb, scgb.control,
          lod = 5e-05, CI = 95, ratio = 1/25, shape = 1, size = 1,
          col = "black", main = "Relative Expression of SCGB", xlab = "SCGB",
          ylab = "Relative Expression",
          names = c("PBCB", "Control"),
          axis = 5, logaxis = 1, stack = 5, jitterwidth = 0.2, jittershape = 1,
          jittersize = 0.5, jittercol = "black", log = "y")

limitplot3(slug.pbcb, slug.control,
          lod = 3e-04, CI = 95, ratio = 1/25, shape = 1, size = 1,
          col = "black", main = "Relative Expression of SLUG", xlab = "SLUG",
          ylab = "Relative Expression",
          names = c("PBCB", "Control"),
          axis = 5, logaxis = 1, stack = 5, jitterwidth = 0.2, jittershape = 1,
          jittersize = 0.5, jittercol = "black", log = "y")

limitplot3(snail.pbcb, snail.control,
          lod = 1.5e-03, CI = 95, ratio = 1/25, shape = 1, size = 1,
```

```
          col = "black", main = "Relative Expression of SNAIL", xlab = "SNAIL",
          ylab = "Relative Expression",
          names = c("PBCB", "Control"),
          axis = 5, logaxis = 1, stack = 5, jitterwidth = 0.2, jittershape = 1,
          jittersize = 0.5, jittercol = "black", log = "y")

limitplot3(twist.pbcb, s.twist.control,
          lod = 2e-04, CI = 95, ratio = 1/25, shape = 1, size = 1,
          col = "black", main = "Relative Expression of TWIST", xlab = "TWIST",
          ylab = "Relative Expression",
          names = c("PBCB", "Control"),
          axis = 5, logaxis = 1, stack = 5, jitterwidth = 0.2, jittershape = 1,
          jittersize = 0.5, jittercol = "black", log = "y")
```

.3 | Patient data analysis

```
library("tableone")
library("stringi")


#import data
patient.data <- read.csv("PBCB_ Database_020516.csv")

patient.data$Age <- 2016 - (patient.data$Birth.Year + 1900)

#add missing patient data- ID146
write.table(patient.data, file = "patient.data.0606.txt", row.names = FALSE, col.names
           = TRUE, sep = "\t" )
patient.data.0606 <- read.csv("patient.data.0606.csv")
patient.data <- patient.data.0606


#-------------------------------------------------------------------------
#cleanup data and make consistent
patient.data$T.Stage <- stri_replace_all(patient.data$T.Stage, "1", fixed = "1a")
patient.data$T.Stage <- stri_replace_all(patient.data$T.Stage, "1", fixed = "1b")
patient.data$T.Stage <- stri_replace_all(patient.data$T.Stage, "1", fixed = "1c")
patient.data$T.Stage <- gsub("^$", "undetermined", paste(patient.data$T.Stage))

patient.data$Lymph.Status <- stri_replace_all(patient.data$Lymph.Status, "N0",
fixed = "N0 ")
patient.data$Lymph.Status <- stri_replace_all(patient.data$Lymph.Status, "N0",
fixed = "pN0")
patient.data$Lymph.Status <- stri_replace_all(patient.data$Lymph.Status, "N1",
fixed = "pN1mic")
patient.data$Lymph.Status <- stri_replace_all(patient.data$Lymph.Status, "N1",
fixed = "N1mic ")
patient.data$Lymph.Status <- stri_replace_all(patient.data$Lymph.Status, "N1",
fixed = "pN1a")
patient.data$Lymph.Status <- stri_replace_all(patient.data$Lymph.Status, "N1",
fixed = "pN1")
patient.data$Lymph.Status <- stri_replace_all(patient.data$Lymph.Status, "N2",
fixed = "pN2a")
patient.data$Lymph.Status <- stri_replace_all(patient.data$Lymph.Status, "N2",
fixed = "pN2a ")
patient.data$Lymph.Status <- stri_replace_all(patient.data$Lymph.Status, "N3",
```

```
fixed = "pN3a")
patient.data$Lymph.Status <- stri_replace_all(patient.data$Lymph.Status,
"undetermined", fixed = "N0x ")
patient.data$Lymph.Status <- stri_replace_all(patient.data$Lymph.Status, "N0",
fixed = "pN")
patient.data$Lymph.Status <- stri_replace_all(patient.data$Lymph.Status, "N1",
fixed = "N1mic")
patient.data$Lymph.Status <- stri_replace_all(patient.data$Lymph.Status, "N1",
fixed = "SN1")
patient.data$Lymph.Status <- stri_replace_all(patient.data$Lymph.Status,
"undetermined", fixed = "N0x")
patient.data$Lymph.Status <- gsub("^$", "undetermined",
paste(patient.data$Lymph.Status))

patient.data$Diagnosis <- stri_replace_all(patient.data$Diagnosis, "other",
fixed = "IC")
patient.data$Diagnosis <- stri_replace_all(patient.data$Diagnosis, "IDC",
fixed = "IDC ")
patient.data$Diagnosis <- stri_replace_all(patient.data$Diagnosis, "ILC",
fixed = "ILC ")

patient.data$ER.status <- stri_replace_all(patient.data$ER.status, "pos",
fixed = "svak pos")
patient.data$ER.status <- gsub("^$", "undetermined", paste(patient.data$ER.status))

patient.data$PR.status <- stri_replace_all(patient.data$PR.status, "neg",
fixed = "neg ")
patient.data$PR.status <- stri_replace_all(patient.data$PR.status, "undetermined",
fixed = "neg/pos")
patient.data$PR.status <- stri_replace_all(patient.data$PR.status, "undetermined",
fixed = "pos/neg")
patient.data$PR.status <- gsub("^$", "undetermined", paste(patient.data$PR.status))

patient.data$HER2 <- stri_replace_all(patient.data$HER2, "neg",
fixed = "neg ")
patient.data$HER2 <- stri_replace_all(patient.data$HER2, "pos",
fixed = "pos(amp 5,3)")
patient.data$HER2 <- stri_replace_all(patient.data$HER2, "pos",
fixed = "pos(FISH amp 8.5")
patient.data$HER2 <- stri_replace_all(patient.data$HER2, "pos",
fixed = "Pos: grense, FISH 1.8")
patient.data$HER2 <- stri_replace_all(patient.data$HER2, "pos",
fixed = "FISH:lavamplifisert ratio 1,9")
patient.data$HER2 <- stri_replace_all(patient.data$HER2, "pos",
fixed = "pos)")
patient.data$HER2 <- gsub("^$", "undetermined", paste(patient.data$HER2))

patient.data$ID <- stri_replace_all(patient.data$ID, "", fixed = "s")

patient.data$Age.Group <- ifelse(patient.data$Age >= 55, "55.and.over", "Under.55")

patient.data$Ki67 <- ifelse(patient.data$Ki67.. <= 10, "low",
                            ifelse(patient.data$Ki67.. > 10 & patient.data$Ki67..
                                   <= 20, "interm", "high"))

patient.data$Ki67[is.na(patient.data$Ki67)] <- "undetermined"

patient.data$FEC[is.na(patient.data$FEC)] <- "0"
patient.data$EC[is.na(patient.data$EC)] <- "0"
patient.data$Taxotere[is.na(patient.data$Taxotere)] <- "0"
```

```
patient.data$chemo <- ifelse(patient.data$FEC == 1 | patient.data$EC == 1 |
                               patient.data$Taxotere == 1, "yes", "no")
patient.data$herceptin <- ifelse(patient.data$herceptin == 1, "yes", "no")
patient.data$endocrine <- ifelse(patient.data$adjuvant == 0, "no", "yes")

patient.data$Diagnosis <- stri_replace_all(patient.data$Diagnosis, "other",
fixed = "IMC")
patient.data$Diagnosis <- stri_replace_all(patient.data$Diagnosis, "other",
fixed = "ITC")
patient.data$Diagnosis <- stri_replace_all(patient.data$Diagnosis, "other",
fixed = "IPC")
patient.data$Diagnosis <- stri_replace_all(patient.data$Diagnosis, "other",
fixed = "MBC")

patient.data$TripNeg <- ifelse(patient.data$ER.status == "neg" & patient.data$PR.status
                               == "neg" & patient.data$HER2 == "neg", "yes", "no")


#-------------------------------------------------------------------------------
#subset positive patients to V1

#highthreshold (3SD outliers removed)
max2.posV1 <- subset(max2.pos, Visit == "V1")


#stratify by CTC-positivity
#CTC2 = V1 only
#CTC = all

patient.data$CTC2 <- ifelse(patient.data$ID %in% max2.posV1$ID, "pos", "neg")

patient.data$CTC <- ifelse(patient.data$ID %in% max2.pos$ID, "pos", "neg")

#-------------------------------------------------------------------------------
#create tableone

myVars <- c("Age", "Diagnosis", "T.Stage", "Tumor.1.Size", "Multifocal",
            "Lymph.Status", "Metastasis", "Grade", "ER.status", "PR.status",
            "HER2", "Ki67..", "lumpectomy", "mastectomy", "chemo", "herceptin",
            "endocrine", "Age.Group", "Ki67", "TripNeg")

catVars <- c("ID", "Diagnosis", "T.Stage", "Lymph.Status", "Metastasis", "Grade",
             "ER.status", "PR.status", "HER2", "lumpectomy", "mastectomy", "Multifocal",
             "chemo", "herceptin", "endocrine", "Age.Group", "Ki67", "TripNeg")


data.tab <- CreateTableOne(vars = myVars, data = patient.data, factorVars = catVars)

#print for copy/paste into excel
print(data.tab, nonnormal = c("Ki67..","Age", "Tumor.1.Size", "Tumor.2.Size"),
      quote = TRUE, noSpaces = TRUE)

#compare CTC status
data.tab.ctc <- CreateTableOne(vars = myVars, strata = "CTC", data = patient.data,
                               factorVars = catVars)

data.tab.ctc2 <- CreateTableOne(vars = myVars, strata = "CTC2", data = patient.data,
                                factorVars = catVars)

print(data.tab.ctc, nonnormal = c("Ki67..","Age", "Tumor.1.Size"),
```

```
      exact = c("ID", "Diagnosis", "T.Stage", "Lymph.Status", "Metastasis",
"Grade","ER.status", "PR.status", "HER2", "lumpectomy", "mastectomy",
"Multifocal", "chemo", "herceptin", "endocrine", "Age.Group", "Ki67"),
      quote = TRUE, noSpaces = TRUE)

print(data.tab.ctc2, nonnormal = c("Ki67..","Age", "Tumor.1.Size"),
      exact = c("Diagnosis", "T.Stage", "Lymph.Status", "Metastasis", "Grade",
                "ER.status", "PR.status", "HER2", "lumpectomy", "mastectomy",
                "Multifocal","chemo", "herceptin", "endocrine", "Age.Group", "Ki67"),
      quote = TRUE, noSpaces = TRUE)

#-------------------------------------------------------------------------------
#group lymph status for testing

patient.data.2 <- patient.data

patient.data.2$Lymph.Status <- stri_replace_all(patient.data.2$Lymph.Status, "N+",
fixed = "N1")
patient.data.2$Lymph.Status <- stri_replace_all(patient.data.2$Lymph.Status, "N+",
fixed = "N2")
patient.data.2$Lymph.Status <- stri_replace_all(patient.data.2$Lymph.Status, "N+",
fixed = "N3")


data.tab.ctc2 <- CreateTableOne(vars = myVars, strata = "CTC2", data = patient.data.2,
                                factorVars = catVars)

print(data.tab.ctc2, nonnormal = c("Ki67..","Age", "Tumor.1.Size", "Tumor.2.Size"),
      exact = c("Diagnosis", "T.Stage", "Lymph.Status", "Metastasis", "Grade",
                "ER.status", "PR.status", "HER2", "lumpectomy", "mastectomy"),
      quote = TRUE, noSpaces = TRUE)

data.tab.ctc <- CreateTableOne(vars = myVars, strata = "CTC", data = patient.data,
factorVars = catVars)

print(data.tab.ctc, nonnormal = c("Ki67..","Age", "Tumor.1.Size", "Tumor.2.Size"),
      exact = c("Diagnosis", "T.Stage", "Lymph.Status", "Metastasis", "Grade",
                "ER.status", "PR.status", "HER2", "lumpectomy", "mastectomy", "TripNeg"),
      quote = TRUE, noSpaces = TRUE)

#add stratification for indiv. markers

patient.data.2$CCDC80 <- ifelse(patient.data.2$ID %in% pos.ccdc80.max2$ID, "pos", "neg")
patient.data.2$EPCAM <- ifelse(patient.data.2$ID %in% pos.epcam.max2$ID, "pos", "neg")
patient.data.2$ERBB2 <- ifelse(patient.data.2$ID %in% pos.erbb2.max2$ID, "pos", "neg")
patient.data.2$KRT8 <- ifelse(patient.data.2$ID %in% pos.krt8.max2$ID, "pos", "neg")
patient.data.2$KRT19 <- ifelse(patient.data.2$ID %in% pos.krt19.max2$ID, "pos", "neg")
patient.data.2$LUM <- ifelse(patient.data.2$ID %in% pos.lum.max2$ID, "pos", "neg")
patient.data.2$SCGB <- ifelse(patient.data.2$ID %in% pos.scgb.max2$ID, "pos", "neg")
patient.data.2$SLUG <- ifelse(patient.data.2$ID %in% pos.slug.max2$ID, "pos", "neg")
patient.data.2$SNAIL <- ifelse(patient.data.2$ID %in% pos.snail.max2$ID, "pos", "neg")
patient.data.2$TWIST <- ifelse(patient.data.2$ID %in% pos.twist.max2$ID, "pos", "neg")

#add stratification for marker groups

pos.emt <- rbind(pos.slug.max2, pos.snail.max2, pos.twist.max2)
pos.new <- rbind(pos.ccdc80.max2, pos.lum.max2)
pos.emtnew <- rbind(pos.slug.max2, pos.snail.max2, pos.twist.max2, pos.ccdc80.max2,
                    pos.lum.max2)
pos.ep <- rbind(pos.epcam.max2, pos.krt8.max2, pos.krt19.max2, pos.erbb2.max2,
```

```
                pos.scgb.max2)
pos.epnew <- rbind(pos.epcam.max2, pos.krt8.max2, pos.krt19.max2, pos.erbb2.max2,
                   pos.scgb.max2, pos.ccdc80.max2, pos.lum.max2)

patient.data.2$EMT <- ifelse(patient.data.2$ID %in% pos.emt$ID, "pos", "neg")
patient.data.2$EP <- ifelse(patient.data.2$ID %in% pos.ep$ID, "pos", "neg")
patient.data.2$NEW <- ifelse(patient.data.2$ID %in% pos.new$ID, "pos", "neg")
patient.data.2$EMTNEW <- ifelse(patient.data.2$ID %in% pos.emtnew$ID, "pos", "neg")
patient.data.2$EPNEW <- ifelse(patient.data.2$ID %in% pos.epnew$ID, "pos", "neg")
patient.data.2$EMTEP <- ifelse(patient.data.2$ID %in% pos.emtnew$ID & patient.data.2$ID
                               %in% pos.ep$ID, "pos", "neg")


#
#analyze and print tables
#


data.tab.epcam <- CreateTableOne(vars = myVars, strata = "EPCAM", data = patient.data,
factorVars = catVars)

print(data.tab.epcam, nonnormal = c("Ki67..","Age", "Tumor.1.Size"),
      exact = c("ID", "Diagnosis", "T.Stage", "Lymph.Status", "Metastasis", "Grade",
                "ER.status", "PR.status", "HER2", "lumpectomy", "mastectomy",
                "Multifocal", "chemo", "herceptin", "endocrine", "Age.Group", "Ki67"),
      quote = TRUE, noSpaces = TRUE)


data.tab.erbb2 <- CreateTableOne(vars = myVars, strata = "ERBB2", data = patient.data,
factorVars = catVars)

print(data.tab.erbb2, nonnormal = c("Ki67..","Age", "Tumor.1.Size"),
      exact = c("ID", "Diagnosis", "T.Stage", "Lymph.Status", "Metastasis", "Grade",
                "ER.status", "PR.status", "HER2", "lumpectomy", "mastectomy",
                "Multifocal","chemo", "herceptin", "endocrine", "Age.Group", "Ki67"),
      quote = TRUE, noSpaces = TRUE)


data.tab.krt8 <- CreateTableOne(vars = myVars, strata = "KRT8", data = patient.data,
factorVars = catVars)

print(data.tab.krt8, nonnormal = c("Ki67..","Age", "Tumor.1.Size"),
      exact = c("ID", "Diagnosis", "T.Stage", "Lymph.Status", "Metastasis", "Grade",
                "ER.status", "PR.status", "HER2", "lumpectomy", "mastectomy",
               "Multifocal", "chemo", "herceptin", "endocrine", "Age.Group", "Ki67"),
      quote = TRUE, noSpaces = TRUE)


data.tab.krt19 <- CreateTableOne(vars = myVars, strata = "KRT19", data = patient.data,
factorVars = catVars)

print(data.tab.krt19, nonnormal = c("Ki67..","Age", "Tumor.1.Size"),
      exact = c("ID", "Diagnosis", "T.Stage", "Lymph.Status", "Metastasis", "Grade",
                "ER.status", "PR.status", "HER2", "lumpectomy", "mastectomy",
               "Multifocal", "chemo", "herceptin", "endocrine", "Age.Group", "Ki67"),
      quote = TRUE, noSpaces = TRUE)



data.tab.lum <- CreateTableOne(vars = myVars, strata = "LUM", data = patient.data,
```

```
factorVars = catVars)

print(data.tab.lum, nonnormal = c("Ki67..","Age", "Tumor.1.Size"),
      exact = c("ID", "Diagnosis", "T.Stage", "Lymph.Status", "Metastasis", "Grade",
                "ER.status", "PR.status", "HER2", "lumpectomy", "mastectomy",
               "Multifocal", "chemo", "herceptin", "endocrine", "Age.Group", "Ki67",
"TripNeg"),
      quote = TRUE, noSpaces = TRUE)



data.tab.scgb <- CreateTableOne(vars = myVars, strata = "SCGB", data = patient.data,
factorVars = catVars)

print(data.tab.scgb, nonnormal = c("Ki67..","Age", "Tumor.1.Size"),
      exact = c("ID", "Diagnosis", "T.Stage", "Lymph.Status", "Metastasis", "Grade",
                "ER.status", "PR.status", "HER2", "lumpectomy", "mastectomy",
               "Multifocal", "chemo", "herceptin", "endocrine", "Age.Group", "Ki67"),
      quote = TRUE, noSpaces = TRUE)


data.tab.slug <- CreateTableOne(vars = myVars, strata = "SLUG", data = patient.data,
factorVars = catVars)

print(data.tab.slug, nonnormal = c("Ki67..","Age", "Tumor.1.Size"),
      exact = c("ID", "Diagnosis", "T.Stage", "Lymph.Status", "Metastasis", "Grade",
                "ER.status", "PR.status", "HER2", "lumpectomy", "mastectomy", x
               "Multifocal", "chemo", "herceptin", "endocrine", "Age.Group", "Ki67"),
      quote = TRUE, noSpaces = TRUE)



data.tab.snail <- CreateTableOne(vars = myVars, strata = "SNAIL", data = patient.data,
factorVars = catVars)

print(data.tab.snail, nonnormal = c("Ki67..","Age", "Tumor.1.Size"),
      exact = c("ID", "Diagnosis", "T.Stage", "Lymph.Status", "Metastasis", "Grade",
                "ER.status", "PR.status", "HER2", "lumpectomy", "mastectomy",
               "Multifocal", "chemo", "herceptin", "endocrine", "Age.Group", "Ki67"),
      quote = TRUE, noSpaces = TRUE)


data.tab.twist <- CreateTableOne(vars = myVars, strata = "TWIST", data = patient.data,
factorVars = catVars)

print(data.tab.twist, nonnormal = c("Ki67..","Age", "Tumor.1.Size"),
      exact = c("ID", "Diagnosis", "T.Stage", "Lymph.Status", "Metastasis", "Grade",
                "ER.status", "PR.status", "HER2", "lumpectomy", "mastectomy",
               "Multifocal", "chemo", "herceptin", "endocrine", "Age.Group", "Ki67"),
      quote = TRUE, noSpaces = TRUE)


data.tab.emt <- CreateTableOne(vars = myVars, strata = "EMT", data = patient.data,
factorVars = catVars)

print(data.tab.emt, nonnormal = c("Ki67..","Age", "Tumor.1.Size"),
      exact = c("ID", "Diagnosis", "T.Stage", "Lymph.Status", "Metastasis", "Grade",
                "ER.status", "PR.status", "HER2", "lumpectomy", "mastectomy", x
              "Multifocal",  "chemo", "herceptin", "endocrine", "Age.Group", "Ki67"),
      quote = TRUE, noSpaces = TRUE)
```

```
data.tab.ep <- CreateTableOne(vars = myVars, strata = "EP", data = patient.data,
factorVars = catVars)

print(data.tab.ep, nonnormal = c("Ki67..","Age", "Tumor.1.Size"),
      exact = c("ID", "Diagnosis", "T.Stage", "Lymph.Status", "Metastasis", "Grade",
                "ER.status", "PR.status", "HER2", "lumpectomy", "mastectomy",
              "Multifocal",  "chemo", "herceptin", "endocrine", "Age.Group", "Ki67"),
      quote = TRUE, noSpaces = TRUE)


data.tab.new <- CreateTableOne(vars = myVars, strata = "NEW", data = patient.data,
factorVars = catVars)

print(data.tab.new, nonnormal = c("Ki67..","Age", "Tumor.1.Size"),
      exact = c("ID", "Diagnosis", "T.Stage", "Lymph.Status", "Metastasis", "Grade",
                "ER.status", "PR.status", "HER2", "lumpectomy", "mastectomy",
              "Multifocal",  "chemo", "herceptin", "endocrine", "Age.Group", "Ki67"),
      quote = TRUE, noSpaces = TRUE)


data.tab.emtnew <- CreateTableOne(vars = myVars, strata = "EMTNEW", data = patient.data,
factorVars = catVars)

print(data.tab.emtnew, nonnormal = c("Ki67..","Age", "Tumor.1.Size"),
      exact = c("ID", "Diagnosis", "T.Stage", "Lymph.Status", "Metastasis", "Grade",
                "ER.status", "PR.status", "HER2", "lumpectomy", "mastectomy",
              "Multifocal",  "chemo", "herceptin", "endocrine", "Age.Group", "Ki67"),
      quote = TRUE, noSpaces = TRUE)



data.tab.epnew <- CreateTableOne(vars = myVars, strata = "EPNEW", data = patient.data,
factorVars = catVars)

print(data.tab.epnew, nonnormal = c("Ki67..","Age", "Tumor.1.Size"),
      exact = c("ID", "Diagnosis", "T.Stage", "Lymph.Status", "Metastasis", "Grade",
                "ER.status", "PR.status", "HER2", "lumpectomy", "mastectomy",
                "Multifocal", "chemo", "herceptin", "endocrine", "Age.Group", "Ki67"),
      quote = TRUE, noSpaces = TRUE)


data.tab.emtep <- CreateTableOne(vars = myVars, strata = "EMTEP", data = patient.data,
factorVars = catVars)

print(data.tab.emtep, nonnormal = c("Ki67..","Age", "Tumor.1.Size"),
      exact = c("ID", "Diagnosis", "T.Stage", "Lymph.Status", "Metastasis", "Grade",
                "ER.status", "PR.status", "HER2", "lumpectomy", "mastectomy",
              "Multifocal",  "chemo", "herceptin", "endocrine", "Age.Group", "Ki67",
"TripNeg"),
      quote = TRUE, noSpaces = TRUE)
```

Appendix E

TABLE 2: Clinicopathological data stratified by *CCDC80+* CTCs

	neg	pos	p	test
n	121	10		
Age (median [IQR])	60.00 [53.00, 65.00]	63.00 [54.50, 67.00]	0.652	nonnorm
Diagnosis (%)			0.479	exact
DCIS	17 (14.0)	0 (0.0)		
IDC	88 (72.7)	8 (80.0)		
ILC	7 (5.8)	1 (10.0)		
other	9 (7.4)	1 (10.0)		
T.Stage (%)			1.000	exact
1	66 (54.5)	7 (70.0)		
2	37 (30.6)	3 (30.0)		
3	2 (1.7)	0 (0.0)		
is	7 (5.8)	0 (0.0)		
undetermined	9 (7.4)	0 (0.0)		
Tumor.1.Size (median [IQR])	16.00 [11.75, 26.25]	18.00 [15.00, 26.75]	0.383	nonnorm
Multifocal = 1 (%)	16 (100.0)	0 (NaN)	NA	exact
Lymph.Status (%)			0.586	exact
N+	26 (21.5)	3 (30.0)		
N0	82 (67.8)	7 (70.0)		
undetermined	13 (10.7)	0 (0.0)		
Metastasis = 1 (%)	18 (29.5)	1 (16.7)	0.667	exact
Grade (%)			0.261	exact
1	20 (16.5)	0 (0.0)		
2	41 (33.9)	6 (60.0)		
3	44 (36.4)	4 (40.0)		
DCIS	16 (13.2)	0 (0.0)		
ER.status (%)			0.650	exact
neg	15 (12.4)	1 (10.0)		
pos	90 (74.4)	9 (90.0)		
undetermined	16 (13.2)	0 (0.0)		
PR.status (%)			0.161	exact
neg	33 (27.3)	1 (10.0)		
pos	70 (57.9)	9 (90.0)		
undetermined	18 (14.9)	0 (0.0)		
HER2 (%)			0.583	exact
neg	95 (78.5)	9 (90.0)		
pos	10 (8.3)	1 (10.0)		
undetermined	16 (13.2)	0 (0.0)		
Ki67 (median [IQR])	31.00 [19.00, 44.00]	36.50 [18.00, 38.75]	0.858	nonnorm
lumpectomy = 1 (%)	93 (76.9)	8 (80.0)	1.000	exact
mastectomy = 1 (%)	31 (25.6)	4 (40.0)	0.456	exact

TABLE 3: Clinicopathological data stratified by *EPCAM+* CTCs

	neg	pos	p	test
n	126	5		
Age (median [IQR])	60.00 [53.00, 65.00]	64.00 [63.00, 70.00]	0.110	nonnorm
Diagnosis (%)			0.795	exact
DCIS	16 (12.7)	1 (20.0)		
IDC	92 (73.0)	4 (80.0)		
ILC	8 (6.3)	0 (0.0)		
other	10 (7.9)	0 (0.0)		
T.Stage (%)			0.309	exact
1	69 (54.8)	4 (80.0)		
2	40 (31.7)	0 (0.0)		
3	2 (1.6)	0 (0.0)		
is	7 (5.6)	0 (0.0)		
undetermined	8 (6.3)	1 (20.0)		
Tumor.1.Size (median [IQR])	16.50 [12.00, 27.00]	14.00 [11.00, 16.25]	0.277	nonnorm
Multifocal = 1 (%)	16 (100.0)	0 (NaN)	NA	exact
Lymph.Status (%)			0.607	exact
N+	28 (22.2)	1 (20.0)		
N0	86 (68.3)	3 (60.0)		
undetermined	12 (9.5)	1 (20.0)		
Metastasis = 1 (%)	19 (28.8)	0 (0.0)	1.000	exact
Grade (%)			0.140	exact
1	18 (14.3)	2 (40.0)		
2	45 (35.7)	2 (40.0)		
3	48 (38.1)	0 (0.0)		
DCIS	15 (11.9)	1 (20.0)		
ER.status (%)			0.760	exact
neg	16 (12.7)	0 (0.0)		
pos	95 (75.4)	4 (80.0)		
undetermined	15 (11.9)	1 (20.0)		
PR.status (%)			0.828	exact
neg	33 (26.2)	1 (20.0)		
pos	76 (60.3)	3 (60.0)		
undetermined	17 (13.5)	1 (20.0)		
HER2 (%)			0.691	exact
neg	100 (79.4)	4 (80.0)		
pos	11 (8.7)	0 (0.0)		
undetermined	15 (11.9)	1 (20.0)		
Ki67.. (median [IQR])	31.00 [19.00, 44.00]	28.00 [18.25, 33.00]	0.344	nonnorm
lumpectomy = 1 (%)	99 (78.6)	2 (40.0)	0.079	exact
mastectomy = 1 (%)	32 (25.4)	3 (60.0)	0.118	exact

Table 4: Clinicopathological data stratified by ERBB2+ CTCs

	neg	pos	p	test
n	130	1		
Age (median [IQR])	60.50 [53.00, 65.75]	53.00 [53.00, 53.00]	0.404	nonnorm
Diagnosis (%)			1.000	exact
DCIS	17 (13.1)	0 (0.0)		
IDC	95 (73.1)	1 (100.0)		
ILC	8 (6.2)	0 (0.0)		
other	10 (7.7)	0 (0.0)		
T.Stage (%)			1.000	exact
1	72 (55.4)	1 (100.0)		
2	40 (30.8)	0 (0.0)		
3	2 (1.5)	0 (0.0)		
is	7 (5.4)	0 (0.0)		
undetermined	9 (6.9)	0 (0.0)		
Tumor.1.Size (median [IQR])	16.00 [12.00, 27.00]	11.00 [11.00, 11.00]	0.302	nonnorm
Multifocal = 1 (%)	16 (100.0)	0 (NaN)	NA	exact
Lymph.Status (%)			1.000	exact
N+	29 (22.3)	0 (0.0)		
N0	88 (67.7)	1 (100.0)		
undetermined	13 (10.0)	0 (0.0)		
Metastasis = 1 (%)	19 (28.4)	0 (NaN)	1.000	exact
Grade (%)			0.275	exact
1	19 (14.6)	1 (100.0)		
2	47 (36.2)	0 (0.0)		
3	48 (36.9)	0 (0.0)		
DCIS	16 (12.3)	0 (0.0)		
ER.status (%)			1.000	exact
neg	16 (12.3)	0 (0.0)		
pos	98 (75.4)	1 (100.0)		
undetermined	16 (12.3)	0 (0.0)		
PR.status (%)			1.000	exact
neg	34 (26.2)	0 (0.0)		
pos	78 (60.0)	1 (100.0)		
undetermined	18 (13.8)	0 (0.0)		
HER2 (%)			1.000	exact
neg	103 (79.2)	1 (100.0)		
pos	11 (8.5)	0 (0.0)		
undetermined	16 (12.3)	0 (0.0)		
Ki67.. (median [IQR])	31.50 [19.25, 44.00]	9.00 [9.00, 9.00]	0.190	nonnorm
lumpectomy = 1 (%)	100 (76.9)	1 (100.0)	1.000	exact
mastectomy = 1 (%)	35 (26.9)	0 (0.0)	1.000	exact

Table 5: Clinicopathological data stratified by *KRT8+* CTCs

	neg	pos	p	test
n	124	7		
Age (median [IQR])	61.00 [53.00, 66.25]	53.00 [50.50, 61.50]	0.180	nonnorm
Diagnosis (%)			0.752	exact
DCIS	16 (12.9)	1 (14.3)		
IDC	91 (73.4)	5 (71.4)		
ILC	8 (6.5)	0 (0.0)		
other	9 (7.3)	1 (14.3)		
T.Stage (%)			0.799	exact
1	69 (55.6)	4 (57.1)		
2	38 (30.6)	2 (28.6)		
3	2 (1.6)	0 (0.0)		
is	7 (5.6)	0 (0.0)		
undetermined	8 (6.5)	1 (14.3)		
Tumor.1.Size (median [IQR])	17.00 [12.00, 28.00]	15.00 [8.70, 19.00]	0.294	nonnorm
Multifocal = 1 (%)	15 (100.0)	1 (100.0)	NA	exact
Lymph.Status (%)			0.362	exact
N+	29 (23.4)	0 (0.0)		
N0	83 (66.9)	6 (85.7)		
undetermined	12 (9.7)	1 (14.3)		
Metastasis = 1 (%)	19 (28.8)	0 (0.0)	1.000	exact
Grade (%)			0.115	exact
1	18 (14.5)	2 (28.6)		
2	47 (37.9)	0 (0.0)		
3	44 (35.5)	4 (57.1)		
DCIS	15 (12.1)	1 (14.3)		
ER.status (%)			1.000	exact
neg	15 (12.1)	1 (14.3)		
pos	94 (75.8)	5 (71.4)		
undetermined	15 (12.1)	1 (14.3)		
PR.status (%)			1.000	exact
neg	32 (25.8)	2 (28.6)		
pos	75 (60.5)	4 (57.1)		
undetermined	17 (13.7)	1 (14.3)		
HER2 (%)			0.443	exact
neg	99 (79.8)	5 (71.4)		
pos	10 (8.1)	1 (14.3)		
undetermined	15 (12.1)	1 (14.3)		
Ki67.. (median [IQR])	31.00 [19.00, 44.00]	36.50 [23.50, 39.00]	0.950	nonnorm
lumpectomy = 1 (%)	96 (77.4)	5 (71.4)	0.659	exact
mastectomy = 1 (%)	33 (26.6)	2 (28.6)	1.000	exact

Table 6: Clinicopathological data stratified by *KRT19+* CTCs

	neg	pos	p	test
n	129	2		
Age (median [IQR])	60.00 [53.00, 66.00]	64.50 [64.25, 64.75]	0.324	nonnorm
Diagnosis (%)			1.000	exact
DCIS	17 (13.2)	0 (0.0)		
IDC	94 (72.9)	2 (100.0)		
ILC	8 (6.2)	0 (0.0)		
other	10 (7.8)	0 (0.0)		
T.Stage (%)			1.000	exact
1	72 (55.8)	1 (50.0)		
2	39 (30.2)	1 (50.0)		
3	2 (1.6)	0 (0.0)		
is	7 (5.4)	0 (0.0)		
undetermined	9 (7.0)	0 (0.0)		
Tumor.1.Size (median [IQR])	16.00 [12.00, 27.00]	20.00 [17.50, 22.50]	0.747	nonnorm
Multifocal = 1 (%)	16 (100.0)	0 (NaN)	NA	exact
Lymph.Status (%)			1.000	exact
N+	29 (22.5)	0 (0.0)		
N0	87 (67.4)	2 (100.0)		
undetermined	13 (10.1)	0 (0.0)		
Metastasis = 1 (%)	19 (28.4)	0 (NaN)	1.000	exact
Grade (%)			1.000	exact
1	20 (15.5)	0 (0.0)		
2	46 (35.7)	1 (50.0)		
3	47 (36.4)	1 (50.0)		
DCIS	16 (12.4)	0 (0.0)		
ER.status (%)			1.000	exact
neg	16 (12.4)	0 (0.0)		
pos	97 (75.2)	2 (100.0)		
undetermined	16 (12.4)	0 (0.0)		
PR.status (%)			1.000	exact
neg	34 (26.4)	0 (0.0)		
pos	77 (59.7)	2 (100.0)		
undetermined	18 (14.0)	0 (0.0)		
HER2 (%)			1.000	exact
neg	102 (79.1)	2 (100.0)		
pos	11 (8.5)	0 (0.0)		
undetermined	16 (12.4)	0 (0.0)		
Ki67.. (median [IQR])	31.00 [19.00, 44.00]	34.50 [32.25, 36.75]	0.764	nonnorm
lumpectomy = 1 (%)	99 (76.7)	2 (100.0)	1.000	exact
mastectomy = 1 (%)	35 (27.1)	0 (0.0)	1.000	exact

TABLE 7: Clinicopathological data stratified by $LUM+$ CTCs

	neg	pos	p	test
n	120	11		
Age (median [IQR])	60.00 [53.00, 66.25]	61.00 [53.50, 64.00]	0.911	nonnorm
Diagnosis (%)			0.262	exact
DCIS	17 (14.2)	0 (0.0)		
IDC	87 (72.5)	9 (81.8)		
ILC	8 (6.7)	0 (0.0)		
other	8 (6.7)	2 (18.2)		
T.Stage (%)			0.863	exact
1	65 (54.2)	8 (72.7)		
2	37 (30.8)	3 (27.3)		
3	2 (1.7)	0 (0.0)		
is	7 (5.8)	0 (0.0)		
undetermined	9 (7.5)	0 (0.0)		
Tumor.1.Size (median [IQR])	17.00 [12.00, 27.00]	15.00 [10.50, 20.00]	0.328	nonnorm
Multifocal = 1 (%)	15 (100.0)	1 (100.0)	NA	exact
Lymph.Status (%)			0.779	exact
N+	27 (22.5)	2 (18.2)		
N0	80 (66.7)	9 (81.8)		
undetermined	13 (10.8)	0 (0.0)		
Metastasis = 1 (%)	19 (29.7)	0 (0.0)	0.553	exact
Grade (%)			0.202	exact
1	16 (13.3)	4 (36.4)		
2	44 (36.7)	3 (27.3)		
3	44 (36.7)	4 (36.4)		
DCIS	16 (13.3)	0 (0.0)		
ER.status (%)			0.428	exact
neg	14 (11.7)	2 (18.2)		
pos	90 (75.0)	9 (81.8)		
undetermined	16 (13.3)	0 (0.0)		
PR.status (%)			0.536	exact
neg	31 (25.8)	3 (27.3)		
pos	71 (59.2)	8 (72.7)		
undetermined	18 (15.0)	0 (0.0)		
HER2 (%)			0.505	exact
neg	94 (78.3)	10 (90.9)		
pos	10 (8.3)	1 (9.1)		
undetermined	16 (13.3)	0 (0.0)		
Ki67.. (median [IQR])	32.50 [19.00, 44.00]	24.00 [13.00, 63.00]	0.618	nonnorm
lumpectomy = 1 (%)	91 (75.8)	10 (90.9)	0.455	exact
mastectomy = 1 (%)	34 (28.3)	1 (9.1)	0.287	exact

TABLE 8: Clinicopathological data stratified by *SCGB+* CTCs

	neg	pos	p	test
n	129	2		
Age (median [IQR])	60.00 [53.00, 66.00]	58.50 [55.25, 61.75]	0.851	nonnorm
Diagnosis (%)			0.464	exact
DCIS	16 (12.4)	1 (50.0)		
IDC	95 (73.6)	1 (50.0)		
ILC	8 (6.2)	0 (0.0)		
other	10 (7.8)	0 (0.0)		
T.Stage (%)			0.120	exact
1	73 (56.6)	0 (0.0)		
2	39 (30.2)	1 (50.0)		
3	2 (1.6)	0 (0.0)		
is	7 (5.4)	0 (0.0)		
undetermined	8 (6.2)	1 (50.0)		
Tumor.1.Size (median [IQR])	16.00 [12.00, 27.00]	13.50 [7.75, 19.25]	0.488	nonnorm
Multifocal = 1 (%)	16 (100.0)	0 (NaN)	NA	exact
Lymph.Status (%)			0.300	exact
N0	88 (68.2)	1 (50.0)		
N1	23 (17.8)	0 (0.0)		
N2	5 (3.9)	0 (0.0)		
N3	1 (0.8)	0 (0.0)		
undetermined	12 (9.3)	1 (50.0)		
Metastasis = 1 (%)	19 (28.4)	0 (NaN)	1.000	exact
Grade (%)			0.252	exact
1	20 (15.5)	0 (0.0)		
2	47 (36.4)	0 (0.0)		
3	47 (36.4)	1 (50.0)		
DCIS	15 (11.6)	1 (50.0)		
ER.status (%)			0.430	exact
neg	16 (12.4)	0 (0.0)		
pos	98 (76.0)	1 (50.0)		
undetermined	15 (11.6)	1 (50.0)		
PR.status (%)			0.323	exact
neg	34 (26.4)	0 (0.0)		
pos	78 (60.5)	1 (50.0)		
undetermined	17 (13.2)	1 (50.0)		
HER2 (%)			0.371	exact
neg	103 (79.8)	1 (50.0)		
pos	11 (8.5)	0 (0.0)		
undetermined	15 (11.6)	1 (50.0)		
Ki67.. (median [IQR])	31.00 [19.00, 44.00]	39.00 [39.00, 39.00]	0.557	nonnorm
lumpectomy = 1 (%)	99 (76.7)	2 (100.0)	1.000	exact
mastectomy = 1 (%)	35 (27.1)	0 (0.0)	1.000	exact

TABLE 9: Clinicopathological data stratified by *SLUG*+ CTCs

	neg	pos	p	test
n	130	1		
Age (median [IQR])	60.00 [53.00, 65.00]	71.00 [71.00, 71.00]	0.186	nonnorm
Diagnosis (%)			1.000	exact
DCIS	17 (13.1)	0 (0.0)		
IDC	95 (73.1)	1 (100.0)		
ILC	8 (6.2)	0 (0.0)		
other	10 (7.7)	0 (0.0)		
T.Stage (%)			1.000	exact
1	72 (55.4)	1 (100.0)		
2	40 (30.8)	0 (0.0)		
3	2 (1.5)	0 (0.0)		
is	7 (5.4)	0 (0.0)		
undetermined	9 (6.9)	0 (0.0)		
Tumor.1.Size (median [IQR])	16.00 [12.00, 27.00]	14.00 [14.00, 14.00]	0.640	nonnorm
Multifocal = 1 (%)	16 (100.0)	0 (NaN)	NA	exact
Lymph.Status (%)			1.000	exact
N0	88 (67.7)	1 (100.0)		
N1	23 (17.7)	0 (0.0)		
N2	5 (3.8)	0 (0.0)		
N3	1 (0.8)	0 (0.0)		
undetermined	13 (10.0)	0 (0.0)		
Metastasis = 1 (%)	19 (28.4)	0 (NaN)	1.000	exact
Grade (%)			1.000	exact
1	20 (15.4)	0 (0.0)		
2	47 (36.2)	0 (0.0)		
3	47 (36.2)	1 (100.0)		
DCIS	16 (12.3)	0 (0.0)		
ER.status (%)			1.000	exact
neg	16 (12.3)	0 (0.0)		
pos	98 (75.4)	1 (100.0)		
undetermined	16 (12.3)	0 (0.0)		
PR.status (%)			1.000	exact
neg	34 (26.2)	0 (0.0)		
pos	78 (60.0)	1 (100.0)		
undetermined	18 (13.8)	0 (0.0)		
HER2 (%)			1.000	exact
neg	103 (79.2)	1 (100.0)		
pos	11 (8.5)	0 (0.0)		
undetermined	16 (12.3)	0 (0.0)		
Ki67.. (median [IQR])	31.00 [19.00, 44.00]	43.00 [43.00, 43.00]	0.433	nonnorm
lumpectomy = 1 (%)	100 (76.9)	1 (100.0)	1.000	exact
mastectomy = 1 (%)	35 (26.9)	0 (0.0)	1.000	exact

Table 10: Clinicopathological data stratified by *SNAIL*+ CTCs

	neg	pos	p	test
n	130	1		
Age (median [IQR])	60.50 [53.00, 65.75]	48.00 [48.00, 48.00]	0.173	nonnorm
Diagnosis (%)			1.000	exact
DCIS	17 (13.1)	0 (0.0)		
IDC	95 (73.1)	1 (100.0)		
ILC	8 (6.2)	0 (0.0)		
other	10 (7.7)	0 (0.0)		
T.Stage (%)			0.443	exact
1	73 (56.2)	0 (0.0)		
2	39 (30.0)	1 (100.0)		
3	2 (1.5)	0 (0.0)		
is	7 (5.4)	0 (0.0)		
undetermined	9 (6.9)	0 (0.0)		
Tumor.1.Size (median [IQR])	16.00 [12.00, 27.00]	23.00 [23.00, 23.00]	0.573	nonnorm
Multifocal = 1 (%)	15 (100.0)	1 (100.0)	NA	exact
Lymph.Status (%)			0.321	exact
N0	89 (68.5)	0 (0.0)		
N1	22 (16.9)	1 (100.0)		
N2	5 (3.8)	0 (0.0)		
N3	1 (0.8)	0 (0.0)		
undetermined	13 (10.0)	0 (0.0)		
Metastasis = 1 (%)	19 (28.4)	0 (NaN)	1.000	exact
Grade (%)			0.634	exact
1	20 (15.4)	0 (0.0)		
2	46 (35.4)	1 (100.0)		
3	48 (36.9)	0 (0.0)		
DCIS	16 (12.3)	0 (0.0)		
ER.status (%)			1.000	exact
neg	16 (12.3)	0 (0.0)		
pos	98 (75.4)	1 (100.0)		
undetermined	16 (12.3)	0 (0.0)		
PR.status (%)			0.137	exact
neg	34 (26.2)	0 (0.0)		
pos	79 (60.8)	0 (0.0)		
undetermined	17 (13.1)	1 (100.0)		
HER2 (%)			1.000	exact
neg	103 (79.2)	1 (100.0)		
pos	11 (8.5)	0 (0.0)		
undetermined	16 (12.3)	0 (0.0)		
Ki67.. (median [IQR])	31.00 [19.00, 44.00]	35.00 [35.00, 35.00]	0.810	nonnorm
lumpectomy = 1 (%)	101 (77.7)	0 (0.0)	0.229	exact
mastectomy = 1 (%)	34 (26.2)	1 (100.0)	0.267	exact

Table 11: Clinicopathological data stratified by *TWIST*+ CTCs

	neg	pos	p	test
n	127	4		
Age (median [IQR])	60.00 [53.00, 66.00]	59.00 [54.50, 63.50]	0.931	nonnorm
Diagnosis (%)			0.224	exact
DCIS	16 (12.6)	1 (25.0)		
IDC	94 (74.0)	2 (50.0)		
ILC	7 (5.5)	1 (25.0)		
other	10 (7.9)	0 (0.0)		
T.Stage (%)			0.267	exact
1	72 (56.7)	1 (25.0)		
2	38 (29.9)	2 (50.0)		
3	2 (1.6)	0 (0.0)		
is	7 (5.5)	0 (0.0)		
undetermined	8 (6.3)	1 (25.0)		
Tumor.1.Size (median [IQR])	16.00 [12.00, 26.50]	25.00 [17.00, 30.50]	0.642	nonnorm
Multifocal = 1 (%)	15 (100.0)	1 (100.0)	NA	exact
Lymph.Status (%)			0.569	exact
N0	86 (67.7)	3 (75.0)		
N1	23 (18.1)	0 (0.0)		
N2	5 (3.9)	0 (0.0)		
N3	1 (0.8)	0 (0.0)		
undetermined	12 (9.4)	1 (25.0)		
Metastasis = 1 (%)	18 (27.7)	1 (50.0)	0.490	exact
Grade (%)			0.644	exact
1	20 (15.7)	0 (0.0)		
2	45 (35.4)	2 (50.0)		
3	47 (37.0)	1 (25.0)		
DCIS	15 (11.8)	1 (25.0)		
ER.status (%)			0.679	exact
neg	16 (12.6)	0 (0.0)		
pos	96 (75.6)	3 (75.0)		
undetermined	15 (11.8)	1 (25.0)		
PR.status (%)			0.334	exact
neg	34 (26.8)	0 (0.0)		
pos	76 (59.8)	3 (75.0)		
undetermined	17 (13.4)	1 (25.0)		
HER2 (%)			0.608	exact
neg	101 (79.5)	3 (75.0)		
pos	11 (8.7)	0 (0.0)		
undetermined	15 (11.8)	1 (25.0)		
Ki67.. (median [IQR])	31.50 [19.00, 44.00]	30.00 [18.00, 34.50]	0.551	nonnorm
lumpectomy = 1 (%)	97 (76.4)	4 (100.0)	0.573	exact
mastectomy = 1 (%)	34 (26.8)	1 (25.0)	1.000	exact

TABLE 12: Clinicopathological data stratified by EMT+ only CTCs: (*LUM*, *CCDC80*, *SNAIL*, *SLUG*, *TWIST*)

	neg	pos	p	test
n	104	27		
Age (median [IQR])	59.50 [53.00, 66.25]	63.00 [53.00, 64.50]	0.862	nonnorm
Diagnosis (%)			0.317	exact
DCIS	16 (15.4)	1 (3.7)		
IDC	75 (72.1)	21 (77.8)		
ILC	6 (5.8)	2 (7.4)		
other	7 (6.7)	3 (11.1)		
T.Stage (%)			0.717	exact
1	56 (53.8)	17 (63.0)		
2	31 (29.8)	9 (33.3)		
3	2 (1.9)	0 (0.0)		
is	7 (6.7)	0 (0.0)		
undetermined	8 (7.7)	1 (3.7)		
Tumor.1.Size (median [IQR])	16.50 [12.00, 27.00]	15.50 [12.00, 24.50]	0.923	nonnorm
Multifocal = 1 (%)	12.5	11.1111111111	NA	exact
Lymph.Status (%)			0.572	exact
N0	69 (66.3)	20 (74.1)		
N1	17 (16.3)	6 (22.2)		
N2	5 (4.8)	0 (0.0)		
N3	1 (1.0)	0 (0.0)		
undetermined	12 (11.5)	1 (3.7)		
Metastasis = 1 (%)	17 (30.4)	2 (18.2)	0.715	exact
Grade (%)			0.448	exact
1	16 (15.4)	4 (14.8)		
2	35 (33.7)	12 (44.4)		
3	38 (36.5)	10 (37.0)		
DCIS	15 (14.4)	1 (3.7)		
ER.status (%)			0.357	exact
neg	13 (12.5)	3 (11.1)		
pos	76 (73.1)	23 (85.2)		
undetermined	15 (14.4)	1 (3.7)		
PR.status (%)			0.141	exact
neg	30 (28.8)	4 (14.8)		
pos	58 (55.8)	21 (77.8)		
undetermined	16 (15.4)	2 (7.4)		
HER2 (%)			0.349	exact
neg	80 (76.9)	24 (88.9)		
pos	9 (8.7)	2 (7.4)		
undetermined	15 (14.4)	1 (3.7)		
Ki67.. (median [IQR])	31.00 [19.00, 48.00]	32.50 [17.00, 42.00]	0.637	nonnorm
lumpectomy = 1 (%)	78 (75.0)	23 (85.2)	0.314	exact
mastectomy = 1 (%)	28 (26.9)	7 (25.9)	1.000	exact

Table 13: Clinicopathological data stratified by Epithelial+ CTCs: (*EPCAM*, *KRT8*, *KRT19*)

	neg	pos	p	test
n	118	13		
Age (median [IQR])	60.50 [53.00, 66.00]	59.00 [53.00, 64.00]	0.820	nonnorm
Diagnosis (%)			1.000	exact
DCIS	15 (12.7)	2 (15.4)		
IDC	86 (72.9)	10 (76.9)		
ILC	8 (6.8)	0 (0.0)		
other	9 (7.6)	1 (7.7)		
T.Stage (%)			0.402	exact
1	64 (54.2)	9 (69.2)		
2	38 (32.2)	2 (15.4)		
3	2 (1.7)	0 (0.0)		
is	7 (5.9)	0 (0.0)		
undetermined	7 (5.9)	2 (15.4)		
Tumor.1.Size (median [IQR])	17.00 [12.00, 29.75]	15.00 [10.25, 16.25]	0.065	nonnorm
Multifocal = 1 (%)	15 (100.0)	1 (100.0)	NA	exact
Lymph.Status (%)			0.710	exact
N0	79 (66.9)	10 (76.9)		
N1	22 (18.6)	1 (7.7)		
N2	5 (4.2)	0 (0.0)		
N3	1 (0.8)	0 (0.0)		
undetermined	11 (9.3)	2 (15.4)		
Metastasis = 1 (%)	19 (29.2)	0 (0.0)	1.000	exact
Grade (%)			0.347	exact
1	16 (13.6)	4 (30.8)		
2	44 (37.3)	3 (23.1)		
3	44 (37.3)	4 (30.8)		
DCIS	14 (11.9)	2 (15.4)		
ER.status (%)			0.889	exact
neg	15 (12.7)	1 (7.7)		
pos	89 (75.4)	10 (76.9)		
undetermined	14 (11.9)	2 (15.4)		
PR.status (%)			0.696	exact
neg	32 (27.1)	2 (15.4)		
pos	70 (59.3)	9 (69.2)		
undetermined	16 (13.6)	2 (15.4)		
HER2 (%)			0.865	exact
neg	94 (79.7)	10 (76.9)		
pos	10 (8.5)	1 (7.7)		
undetermined	14 (11.9)	2 (15.4)		
Ki67.. (median [IQR])	31.00 [18.50, 45.00]	32.00 [22.00, 37.50]	0.641	nonnorm
lumpectomy = 1 (%)	92 (78.0)	9 (69.2)	0.493	exact
mastectomy = 1 (%)	31 (26.3)	4 (30.8)	0.746	exact

Table 14: Clinicopathological data stratified by EMT+/Epithelial+ CTCs

	neg	pos	p	test
n	126	6		
Age (median [IQR])	60.00 [53.00, 66.00]	58.00 [50.00, 63.75]	0.447	nonnorm
Diagnosis (%)			0.682	exact
DCIS	17 (13.5)	0 (0.0)		
IDC	92 (73.0)	5 (83.3)		
ILC	8 (6.3)	0 (0.0)		
other	9 (7.1)	1 (16.7)		
T.Stage (%)			1	exact
1	69 (54.8)	4 (66.7)		
2	39 (31.0)	2 (33.3)		
3	2 (1.6)	0 (0.0)		
is	7 (5.6)	0 (0.0)		
undetermined	9 (7.1)	0 (0.0)		
Tumor.1.Size (median [IQR])	16.50 [12.00, 27.50]	15.00 [12.75, 20.25]	0.459	nonnorm
Multifocal = 1 (%)	15 (100.0)	1 (100.0)	NA	exact
Lymph.Status (%)			0.629	exact
N0	83 (65.9)	6 (100.0)		
N1	24 (19.0)	0 (0.0)		
N2	5 (4.0)	0 (0.0)		
N3	1 (0.8)	0 (0.0)		
undetermined	13 (10.3)	0 (0.0)		
Metastasis = 1 (%)	19 (28.4)	0 (0.0)	1	exact
Grade (%)			0.938	exact
1	19 (15.1)	1 (16.7)		
2	45 (35.7)	2 (33.3)		
3	46 (36.5)	3 (50.0)		
DCIS	16 (12.7)	0 (0.0)		
ER.status (%)			1	exact
neg	16 (12.7)	1 (16.7)		
pos	94 (74.6)	5 (83.3)		
undetermined	16 (12.7)	0 (0.0)		
PR.status (%)			0.59	exact
neg	34 (27.0)	1 (16.7)		
pos	74 (58.7)	5 (83.3)		
undetermined	18 (14.3)	0 (0.0)		
HER2 (%)			1	exact
neg	99 (78.6)	6 (100.0)		
pos	11 (8.7)	0 (0.0)		
undetermined	16 (12.7)	0 (0.0)		
Ki67.. (median [IQR])	31.50 [17.50, 44.00]	34.50 [25.50, 39.00]	0.694	nonnorm
lumpectomy = 1 (%)	97 (77.6)	4 (66.7)	0.62	exact
mastectomy = 1 (%)	33 (26.4)	2 (33.3)	0.658	exact

TABLE 15: Clinicopathological data stratified by *LUM*+ & *CCDC80*+ CTCs

	neg	pos	p	test
n	110	21		
Age (median [IQR])	59.50 [53.00, 65.75]	63.00 [53.00, 64.00]	0.809	nonnorm
Diagnosis (%)			0.116	exact
DCIS	17 (15.5)	0 (0.0)		
IDC	79 (71.8)	17 (81.0)		
ILC	7 (6.4)	1 (4.8)		
other	7 (6.4)	3 (14.3)		
T.Stage (%)			0.487	exact
1	58 (52.7)	15 (71.4)		
2	34 (30.9)	6 (28.6)		
3	2 (1.8)	0 (0.0)		
is	7 (6.4)	0 (0.0)		
undetermined	9 (8.2)	0 (0.0)		
Tumor.1.Size (median [IQR])	17.00 [12.00, 27.00]	15.00 [12.00, 22.00]	0.914	nonnorm
Multifocal = 1 (%)	15 (100.0)	1 (100.0)	NA	exact
Lymph.Status (%)			0.389	exact
N0	73 (66.4)	16 (76.2)		
N1	18 (16.4)	5 (23.8)		
N2	5 (4.5)	0 (0.0)		
N3	1 (0.9)	0 (0.0)		
undetermined	13 (11.8)	0 (0.0)		
Metastasis = 1 (%)	18 (31.0)	1 (11.1)	0.427	exact
Grade (%)			0.264	exact
1	16 (14.5)	4 (19.0)		
2	38 (34.5)	9 (42.9)		
3	40 (36.4)	8 (38.1)		
DCIS	16 (14.5)	0 (0.0)		
ER.status (%)			0.157	exact
neg	13 (11.8)	3 (14.3)		
pos	81 (73.6)	18 (85.7)		
undetermined	16 (14.5)	0 (0.0)		
PR.status (%)			0.051	exact
neg	30 (27.3)	4 (19.0)		
pos	62 (56.4)	17 (81.0)		
undetermined	18 (16.4)	0 (0.0)		
HER2 (%)			0.144	exact
neg	85 (77.3)	19 (90.5)		
pos	9 (8.2)	2 (9.5)		
undetermined	16 (14.5)	0 (0.0)		
Ki67.. (median [IQR])	31.50 [19.25, 44.00]	30.00 [16.00, 43.00]	0.610	nonnorm
lumpectomy = 1 (%)	83 (75.5)	18 (85.7)	0.403	exact
mastectomy = 1 (%)	30 (27.3)	5 (23.8)	1.000	exact